Electronic Markets
https://doi.org/10.1007/s12525-022-00605-4

PREFACE

Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities

Roman Lukyanenko[1] · **Wolfgang Maass**[2] · **Veda C. Storey**[3]

© The Author(s), under exclusive licence to Institute of Applied Informatics at University of Leipzig 2022

Abstract
With the rise of artificial intelligence (AI), the issue of trust in AI emerges as a paramount societal concern. Despite increased attention of researchers, the topic remains fragmented without a common conceptual and theoretical foundation. To facilitate systematic research on this topic, we develop a Foundational Trust Framework to provide a conceptual, theoretical, and methodological foundation for trust research in general. The framework positions trust in general and trust in AI specifically as a problem of interaction among systems and applies systems thinking and general systems theory to trust and trust in AI. The Foundational Trust Framework is then used to gain a deeper understanding of the nature of trust in AI. From doing so, a research agenda emerges that proposes significant questions to facilitate further advances in empirical, theoretical, and design research on trust in AI.

Keywords Artificial intelligence (AI) · Trust · Foundational Trust Framework · Trust in AI · Explainable AI · Transparency · Systems

JEL Classification L63 · L64 · L86 · C80 · D11 · C71 · C72 · C73 · J00

Introduction

Few technological developments rival the explosive growth of artificial intelligence (AI). AI is estimated to contribute $15 trillion to global GDP by 2030 (Rao & Verweij, 2017). In fact, it has been argued that the country-leader in AI is to become the world's preeminent power of the future (Gill, 2020). Some call AI "the pinnacle of [human] ingenuity" (Filippouli, 2017). With so many expectations vested into AI, recent Gartner's hype cycles are dominated by AI-based technologies (e.g., robots, chatbots) and the variants of AI itself (e.g., causal AI).[1]

Whereas traditionally AI focused on logic-based models, the growth of data, coupled with advances in computational power, shifted the focus almost exclusively to data-intensive AI. Machine learning, where computers are trained to extract useful patterns from data, is now the dominant form of AI (Agrawal et al., 2018; Cerf, 2019). In addition, techniques, such as natural language processing (extraction and processing of natural human language) and computer vision (extraction of meaning from images and video), are also prominent (Eisenstein, 2019; McAfee & Brynjolfsson, 2017).

The successes of AI are mounting. AI is transforming businesses and entire industries, such as manufacturing, transportation, and finance. For example, electronic marketplaces, including Amazon and Alibaba, are using AI technologies to provide smart services to consumers, optimize logistics, analyze consumer behavior, and derive innovative product and service designs (Chamorro-Premuzic et al.,

This article is part of the Topical Collection on Trust in artificial intelligence

✉ Roman Lukyanenko
romanl@virginia.edu

Wolfgang Maass
wolfgang.maass@iss.uni-saarland.de

Veda C. Storey
VStorey@gsu.edu

[1] University of Virginia, Charlottesville, VA, USA

[2] Saarland University and German Research Center for Artificial Intelligence (DFKI), Saarbruecken, Germany

[3] Georgia State University, Atlanta, GA, USA

[1] See, for example, the 2022 report: https://www.gartner.com/en/articles/what-s-new-in-the-2022-gartner-hype-cycle-for-emerging-technologies.

Published online: 28 November 2022

2019; Jia et al., 2018; Kiron & Schrage, 2019). Although not fully autonomous, vehicles supported by AI are now common-place on roads and highways (Kirkpatrick, 2022; Waldrop, 2015). Medical diagnoses are being routinely performed by AI (Davenport & Kalakota, 2019; Langlotz, 2019; D. Lee & Yoon, 2021). Specific activities, such as market segmentation, sentiment analysis, spam detection, high-frequency stock trading, are nearly universally conducted using AI. AI, such as the GPT-3, LaMDA and DALL-E 2 systems, is now capable of generating realistic scientific papers,[2] writing poetry,[3] composing music and creating art.[4]

With the rise of AI, the issue of trust in this technology emerges as a paramount societal concern. Applications such as AI-based surgery and medical diagnoses, driverless cars, jail and parole, automated job applications screening, wealth investment, and AI-based military weapons, raise numerous ethical and existential questions and result in fear and anxiety. Many avant-garde scientists (e.g., Stephen Hawking), and business leaders (e.g., Elon Musk, Bill Gates) consider there to be major threats to society from sophisticated AI solutions (Bostrom, 2014; Harari, 2016; Marr, 2018; Yudkowsky, 2008).

Responding to these challenges is a growing chorus of research on trust in AI (including papers accepted for this Special Issue). These studies capitalize on an already established foundation on trust in social settings and trust toward technology. This literature, however, remains fragmented, without a common foundation that could integrate the results. The coverage of trust in AI, thus far, has also been uneven with much emphasis on specific topics, potentially at the expense of others.

We develop a *Foundational Trust Framework*. The framework provides a conceptual, theoretical, and methodological foundation for trust research in general, and trust in AI, specifically. The framework positions trust in AI as a problem of interaction among systems and applies systems thinking and general systems theory to trust. The paper synthesizes works of Luhmann (1995, 2018) with other theories of systems (Ackoff, 1971; Bunge, 2003b; von Bertalanffy, 1968) to develop a formalized foundation for trust research resulting in the Foundational Trust Framework.

The Foundational Trust Framework is then applied to trust in AI. Emerging from this application is an agenda for research on trust in AI, which identifies unexplored or under-explored, emerging opportunities. The agenda poses important questions to facilitate further advances in empirical, theoretical, and design research.

This preface is organized as follows. Section "Background: Trust in AI" provides a background on trust in AI, followed by a review of the literature in Section "Existing literature on trust in AI". Section "Foundational Trust Framework" develops the Foundational Trust Framework, which is followed by a proposed agenda for research on trust in AI in Section "Trust in AI and trust in AI research agenda". Section "Discussion and conclusions" discusses the contributions of the framework and our proposed research agenda. Section "Special issue on "Trust in AI" in Electronic Markets" highlights the papers that appear in this special issue.

Background: Trust in AI

Trust is generally regarded as a psychological mechanism for reducing uncertainty and increasing the likelihood of a successful (e.g., safe, pleasant, satisfactory) interaction with entities in the environment. When we trust someone, we expend less cognitive, physiological, and economic resources dealing with this entity. Trust has been evolutionarily beneficial for humans (Yamagishi, 2011) and is argued to be a prerequisite for any social interaction (Luhmann, 2018). Table 1 provides a variety of definitions of trust in diverse disciplines. These definitions demonstrate the wide range of conceptualizations of trust (and trust in AI). They also reveal the lack of consensus on understanding the nature of trust, leading to the need to develop the Foundational Trust Framework presented later in this preface.

Trust is a critical aspect of AI adoption and usage. Trust becomes an important factor for overcoming a substantial uncertainty which pervades the development and deployment of AI. The uncertainly and ambiguity leads to much caution, skepticism, and distrust.

In many ways, distrust in AI is well-grounded. Notwithstanding the spectacular successes, many existing AI-based technologies have failed dramatically. The failures may be due to biases in AI algorithms, resulting in discriminatory practices at massive scale. A canonical example is the failure of the tool COMPASS designed to aid judges in release and detention decisions. The AI-based tool, upon further investigation, was found to be biased towards African-Americans (Mehrabi et al., 2021).[5]

The failures may be rooted in errors when training AI. An example in the sensitive medical context is the failure of the famous AI system, IBM Watson (Davenport & Ronanki, 2018). As IBM engineers trained the software on hypothetical cancer patients, rather than real ones, medical specialists identified unsafe treatment recommendations, such as to

[2] https://www.scientificamerican.com/article/we-asked-gpt-3-to-write-an-academic-paper-about-itself-mdash-then-we-tried-to-get-it-published/

[3] https://thewalrus.ca/ai-poetry/

[4] https://www.nytimes.com/2022/08/24/technology/ai-technology-progress.html

[5] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Table 1 Select definitions of trust from different domains

Study	Definition	Object of trust
Glikson and Woolley (2020)	tendency to take a meaningful risk while believing in a high chance of positive outcome	Artificial intelligence (virtual agents and robots)
Jacovi et al. (2021)	directional transaction between two parties: if A believes that B will act in A's best interest, and accepts vulnerability to B's actions, then A trusts B. Interpersonal trust. Human-AI trust. If H (human) perceives that M (AI model) is trustworthy to contract C, and accepts vulnerability to M's actions, then H trusts M contractually to C	Humans, Artificial intelligence (virtual agents and robots)
Gillath et al. (2021)	affective route to boost trust is defined as an increase in the faith in the trustworthy intentions of others, or the confidence people place in others based on how they feel about them	Artificial intelligence
Kożuch and Sienkiewicz-Małyjurek (2022)	social capital based on mutual relations between people and organizations, increasing reciprocity and commitment	Public safety networks
Mayer et al. (1995)	willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party	Organizational settings
Wan et al. (2022)	subjective willingness and strength of both parties to implement an agreement	Blockchain
Rousseau et al. (1998)	psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another	Organizational settings
Sabel (1993)	mutual confidence that no party involved in an exchange will exploit the other's vulnerability	Economy
Boon and Holmes (1991)	state involving confident positive expectations about another's motives with respect to oneself in situations entailing risk	Social relations
Gefen et al. (2003)	set of specific beliefs that deal with integrity, benevolence, ability, and predictability	E-commerce settings

give a cancer patient with severe bleeding a drug that could worsen it (Storey et al., 2022). The IBM Watson was discontinued by its early adopter, the MD Anderson clinic, after sinking $62 million in its failed realization (Lohr, 2021). As a result of design flaws or human operating errors, AI-based driverless cars ran over and killed pedestrians (Scanlon et al., 2021; Wakabayashi, 2018). Despite years of development and progress in driverless technology, modern roads are still dominated by the imperfect human drivers.

The growing list of nefarious actions perpetrated with the aid of AI are also affecting the trusting beliefs in this technology. For example, hackers use AI-based approaches to increase sophistication and scale of their attacks (Sadiku et al., 2020). The constant fight against such nefarious AI compels Taddeo and colleagues (Taddeo, 2021; Taddeo, McCutcheon, & Floridi, 2019) to argue that trust may never be fully achievable in the context of cybersecurity.

Many obstacles stand in the way of robust and reliable AI. The quality of AI depends on the quality of the data used for training AI models (Sambasivan et al., 2021), which may be rooted in murky and ill-understood organizational routines (Storey et al., 2022). The systems based on AI may be developed by inexperienced teams who unwittingly may introduce errors and biases (Mehrabi et al., 2021). Controlling the quality of data used to train AI can be exceedingly difficult, especially if some or all of the training data comes from data collection online, such as social media or crowdsourcing (Allahbakhsh et al., 2013; Kosmala et al., 2016; Lukyanenko & Parsons, 2018; Salk et al., 2015).

Furthermore, the option to "look" inside the models of AI, such as deep learning neural networks, remains limited due to the great complexity of these, and other, powerful AI models (Castelvecchi, 2016; Domingos, 2012; David Gunning & Aha, 2019). The research on explainable AI (XAI) is rapidly progressing, but, despite substantial progress (Adadi & Berrada, 2018; Dosilović et al., 2018; Mueller et al., 2019; Rai, 2020), even leaders in the field, such as Google, admit to not fully knowing how their models work (Storey et al., 2022). The research also suffers from a notable gap: "most of the existing literature on XAI methods is based on the developer's intuition rather than [on the needs of] the intended users" (Adadi & Berrada, 2018, p. 52153). Hence, when accessed by non-technical audiences, many explanations themselves require explanation (Adadi & Berrada, 2018; Lukyanenko, Castellanos, et al., 2021a; Miller et al., 2017). The need for intuitive explanations is especially pronounced in sensitive domains, such as healthcare (Lötsch et al., 2021).

Despite high-profile failures, the spectacular successes of AI are equally impressive. These range from such highly publicized events as winning the popular quiz show *Jeopardy!* (Ferrucci, 2010) and beating the reigning Go champion (Holcomb et al., 2018) to driverless cars traversing the real roads (Waldrop, 2015). There are even more less publicly

visible, but highly impactful achievements in diverse applications, such as fraud detection, micro-targeted advertisements, medical diagnoses, and manufacturing automation (Agrawal et al., 2018; Brynjolfsson & McAfee, 2014). These successes increase trust in specific applications of AI and the AI industry as a whole (Glikson & Woolley, 2020).

Ironically, the successes of AI may also contribute to distrust, as AI technology is also falling victim of its own success. AI can be seen as "the fundamental technology that underlies "*Surveillance Capitalism*," defined as an economic system centered on the commodification of personal data with the core purpose of profit-making" (Vardi, 2022, p. 5). AI supports such controversial practices as extremely granular analysis of personal data, resulting in the eerie feeling that an AI knows you better than you know yourself (Thompson, 2018), or dynamic pricing, when service or product offerings are hyper-optimized to our willingness or even ability to pay (Haenlein et al., 2022; Shartsis, 2019). AI also underlies government or employer surveillance of individuals (e.g., via facial recognition technologies). These uses of AI re-enforce the fear that humans are being reduced to AI's inputs (Harari, 2016; Leidner & Tona, 2021).

The relentless expansion of AI brings about concerns about the future of work (Adamczyk et al., 2021; Park & Kim, 2022; Petersen et al., 2022). According to some reports, an estimated 50% of the current occupations may be displaced due to automation (Frey & Osborne, 2017; Petersen et al., 2022). Other estimates are even higher (Shaturaev, 2022). While not all job losses result in ultimate unemployment (as new careers become possible as a result of AI) (Belchik, 2022; Park & Kim, 2022), the economics of AI is a contributing factor to its distrust, especially by those who have already lost employment opportunities or fear being left behind.

Another source of distrust is rooted in concerns over the long-term consequences of progress in AI. Current efforts to expand the capabilities of AI are considered by some thinkers to be a steppingstone toward the ultimate end of humanity (Alfonseca et al., 2021). AI is feared to be a precursor to superintelligence. A superintelligence is any intellect that vastly outperforms the best human abilities in nearly all domains and contexts, including creativity, common sense, and social skills (Bostrom, 1998; Yampolskiy, 2015). If, and when, such technology is attained, it may not be "just another technology." Rather, it may be a turning point in human civilization, and potentially, the entire universe, because it would unleash possibilities that are beyond current comprehension (Bostrom, 1998; Harari, 2016).

Superintelligence may threaten the very survival of humans. Reasonable questions to ask are: Would an all-powerful super-intelligent being find any use for humans? Would our dismal historic track record of wars, violence, and discrimination be viewed by the super-intelligent being as a reason to remove humans from existence? Would we be seen as a defunct and fundamentally flawed branch of cosmic evolution?

The relentless progress in AI paves the way for this superintelligence possibility (Floridi, 2019; Ramge, 2019).[6] Voices of fear, skepticism, and concern for a super-intelligent future is a backdrop to the problem of trust in *existing* and *near future* AI-based technologies. The more human activities are touched, affected by, transformed, or automated by AI, the more concerns about safety, reliability, predictability, transparency, dependency on these technologies, emerge. These concerns lead to the following societal question: *Can we as individuals, collectives, institutions, countries, and humanity as a whole, trust artificial intelligence?* As IBM proclaimed: "What's next for AI? – Building Trust."[7]

Thus, the issue of trust (and distrust) of AI is obviously complex, multilayered, deeply intertwined with economic, social, political, and psychological factors, in addition to the technology itself.

Existing literature on trust in AI

In response to the growing importance of AI, trust in AI emerges as a major research area, resulting in a rapidly expanding body of literature. As evident from the complex issues surrounding AI, trust in AI, fundamentally, is a multidisciplinary research topic. Among the areas that actively contribute to this discussion are artificial intelligence and computer science, human computer interaction, organizational science, philosophy, psychology, sociology, marketing, software engineering, information systems, medicine, political science, and economics. Within these disciplines, distinct (although often overlapping) conceptualizations, approaches, and solutions to trust in AI are being developed.

Overview of trust and AI literature

Computer science, and its subfield of artificial intelligence, investigate the nature of trust in AI from the point of view of computation and algorithm development. As discussed, such efforts include ways to progress AI systems to become more transparent and explainable (Abdul et al., 2018; Adadi & Berrada, 2018; David Gunning & Aha, 2019; Storey et al., 2022). They also actively investigate the problem of

[6] https://www.cnbc.com/2021/08/24/elon-musk-warned-of-ai-apocalypsenow-hes-building-a-tesla-robot.html

[7] https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html

machine-learning biases (Mehrabi et al., 2021), which is a key source of AI failure that engenders distrust in specific AI systems and the AI industry as a whole.

Human computer interaction (HCI) investigates the design and psychological mechanisms that impact users' trusting perceptions in AI systems and their subsequent use behaviors (Lee & See, 2004; Robert Jr et al., 2020; Söllner et al., 2012). The HCI studies advocate for greater transparency, systematicity, level of control, structuring, and rigor in the development of AI systems (Hoff & Bashir, 2015; Lee & See, 2004). Thus, one of the thought leaders, Gary Marcus proclaims: "Don't trust AI until we build systems that earn trust."[8] Such systems should be based on solid engineering principles, such as designing for failure, having failsafe measures, explicit maintenance protocols, redundancy, and design process transparency (Marcus & Davis, 2019).

Design process transparency has recently become a topic of interest to the conceptual modeling community (Fettke, 2020; Lukyanenko et al., 2020; Lukyanenko, Castellanos, et al., 2019a; Maass & Storey, 2021; Reimer et al., 2020). Conceptual models, such as entity relationship diagrams or UML class diagrams, are commonly used to design databases (Davies et al., 2006; Dobing & Parsons, 2006; Fettke, 2009; Storey & Goldstein, 1993; Teorey et al., 1986). They are also used as tools of structuring, diagnosing and documenting the construction of IT and business processes (Hvalshagen et al., 2023; Mylopoulos, 1998; Recker et al., 2021; Wand & Weber, 2002). Extensive prior research has investigated what makes conceptual models easy to comprehend, including by non-expert users (Bodart et al., 2001; Castellanos et al., 2020; Eriksson et al., 2019; Khatri et al., 2006; Lukyanenko, Parsons, & Samuel, 2019b; Moody, 2009; Samuel et al., 2018; Shanks et al., 2008).

Building on these foundations, the benefits of conceptual modeling are now being extended to AI. Thus, research shows that the carefully-crafted by human experts conceptual models can improve the transparency and explainability of AI models (Lukyanenko et al., 2020; Maass et al., 2021, 2022a, b). Conceptual modeling can thus facilitate greater trust in AI technologies. In general, there is a growing movement to add more domain knowledge into data-driven AI.[9] This is reminiscent of the symbolic AI tradition (Crevier, 1993; Domingos, 2015; Minsky, 1974), but with the recognition of the expanding ability of modern AI algorithms (e.g., backpropagation) to extract complex patterns in large datasets.

Related to transparency is the perception or belief in control. People tend to trust entities or processes over which they have control, even when the control is illusory (Komiak & Benbasat, 2008; McKnight et al., 1998). Indeed, predictability of AI is a key trust antecedent (Brashear et al., 2003). Hence, building control mechanisms in AI is not only important for safety reasons (Alfonseca et al., 2021), but also to enhance trust. Therefore, human autonomy, the right or the power to have control of own decision and choices, is one of the most common principles of ethical AI (Floridi & Cowls, 2021).

Psychology, especially social psychology, has much to contribute to the topic of trust in AI because it provides concepts and theories to understand the nature of trust (Rotenberg, 2019; Schul et al., 2008; Simpson, 2007), including trust in technology. Computer science and artificial intelligence have historically benefitted from insights in psychology, as human anatomy is used both as a metaphor, as well as a reference, for how to develop and improve AI (Samuel, 1959; von Neumann, 1958). Among the notable insights from psychology are dispositional and cultural factors impacting trust. Hence, it is estimated that over 60% of people may have an aversion bias toward algorithmic decision making (Stackpole, 2019). Another insight is that similarity in shared values is among the strongest psychological antecedents of trust (Garcia-Retamero et al., 2012; Siegrist & Zingg, 2014). Trust also appears to be partially culturally determined. For example, Americans tend to trust people primarily based on whether they share category memberships; in contrast, Japanese tend to trust others based on direct or indirect interpersonal links (Yuki et al., 2005).

Drawing on foundations in psychology, information systems, software engineering and computer science disciplines have been investigating issues related to trust and technology, and more recently, trust in AI. Research in psychology demonstrates that agreeable people tend to be more trusting (Mooradian et al., 2006); a finding which generalizes to robots (Chien et al., 2016; Oksanen et al., 2020). Likewise, consistent with findings in psychology, trust strategies differ across IT user age groups (Hoff & Bashir, 2015; Steinke et al., 2012). Among other cross-disciplinary insights are the models of trust in technology adoption and use, differential impact of cognitive and emotional elements of trust, and the impact of anthropomorphism and user-technology likeness on technology adoption and use (Benbasat & Wang, 2005; Dimoka, 2010; Gefen et al., 2003; Komiak & Benbasat, 2006; Sanders et al., 2011).

An interdisciplinary area of AI ethics is emerging (Haenlein et al., 2022; Leidner & Tona, 2021; Robert Jr et al., 2020). One of its objectives is to provide guidance for developing AI. A promising direction is development of ethical codes of conduct, and protocols and methods to be followed by AI developers and organizations voluntarily, as

[8] https://www.economist.com/open-future/2019/12/18/dont-trust-ai-until-we-build-systems-that-earn-trust

[9] https://venturebeat.com/ai/andrew-ng-predicts-the-next-10-years-in-ai/

industry-wide norms (Crawford & Calo, 2016). Hence, IBM developed an "AI FactSheet" – a voluntary, but increasingly popular, checklist that captures various aspects of AI systems aimed at increasing its trustworthiness (Arnold et al., 2019). Some advocate a "buddy system" in which AI project development teams include behavioral scientists so to provide the needed expertise in trust psychology (Stackpole, 2019). This recommendation is supported by other scholars (Storey et al., 2022).

An alternative to self-regulation is legal mandate and enforcement. Here, many research issues must be addressed. Examples include: how to define AI to ensure the right technology is regulated, while not stifling development of other technologies; how to create fundamentally safer software (Ellul, 2022); whether regulations be applied only to sensitive cases or any AI irrespective of use (Haenlein et al., 2022); and whether AI can be regulated as a component of software or if the entire AI systems must be subject to such actions (Ellul, 2022).

There are debates on the very possibility of instilling ethics in AI. Bostrom (2014, p. 227) argues that, ultimately, human values "bottom out in terms that appear in the AI's programming language, and ultimately in primitives such as mathematical operations and addresses pointing to the contents of individual memory registers." Others take an opposite view: "developing an understanding of ethics as contemporary humans understand is actually one of the easier problems facing AI" (E. Davis, 2015, p. 122). Much work remains on reconciling these divergent positions.

Important contributions to the ethics debate originate from philosophy, which builds on its historic foundations in epistemology, axiology, ethics, philosophy of life, wellness and happiness (Rescher, 2013; Sturt, 1903). From these, the foundations of philosophy of trust emerging (Faulkner & Simpson, 2017; Scheman, 2015; Whyte & Crease, 2010).

Organizational studies contribute to trust in AI with a unique organizational focus. These works extend the foundations in organizational trust to AI because trust is a key element of social interactions (Mayer et al., 1995; McAllister, 1995). For example, research considers organizational culture, norms and dynamics as predictors of trust and adoption of AI-based technologies (Glikson & Woolley, 2020). Economics and organizational perspectives provide insights into the types of occupations most likely to be transformed by AI-based automation (Bickley et al., 2022; Brynjolfsson & McAfee, 2014; Frey & Osborne, 2017; M.-H. Huang & Rust, 2018), which could explain the disposition to distrust AI by those potentially (or already) affected (Agrawal et al., 2018; Faraj et al., 2018; Glikson & Woolley, 2020). These studies further investigate the dispositional factors that result in greater or lesser trust in general technology, automation, and AI. Another notable contribution of organizational studies is the focus on a non-individual level of analysis, such as groups or organizations, in the formation of organizational trusting beliefs toward AI (Jarvenpaa et al., 1998; Jarvenpaa & Leidner, 1999; Li et al., 2021).

AI has been a disruptive technology for organizations. Among the key issues related to AI trust is the development of organizational policies dealing with AI ethics and trust. However, these efforts fail to establish a consensus among the guidelines or resolve internal contradictions (Thiebes et al., 2021).

Trust is an active research area in economics, where it is a basis for much economic activity as a form of social capital. As Akerlof (1978, p. 500) states: trust-based "unwritten guarantees are preconditions for trade and production." Since buyers and sellers do not have perfect information about one another, within the context of information asymmetry, trust fills this void, making many risky transactions possible. Under these assumptions, game theoretic approaches have been widely used in economics to investigate trust, including when dealing with AI (e.g., Boero et al., 2009; Keser, 2003; Schniter et al., 2020). Among the findings of such studies is that users may equally trust fellow humans and robots when similar payoffs are expected (Schniter et al., 2020).

The target application domains of AI, such as medicine, engineering, finance, transportation, or military investigate trust and AI in specific contexts. In healthcare settings, for example, some issues are how to: increase trust and facilitate adoption of AI-based systems in hospitals, by patients and healthcare workers (Asan et al., 2020; Paré et al., 2020); increase transparency of AI-based systems; and reduce bias in medical applications (Starke et al., 2022; Vokinger et al., 2021; Wang & Siau, 2018). One notable insight from such sensitive and mission-critical domains, is the value of using human-in-the-loop in AI (Holzinger, 2016; Paré et al., 2020). This occurs, for example, when AI delegates a classification decision to a human if it lacks confidence for a given case. The human-in-the-loop approach in medicine promises to mitigate bias, improve transparency, and increase trust in AI-based systems (Holzinger, 2016; Holzinger et al., 2019). Furthermore, trust in healthcare and other critical contexts appears to be more sensitive to structural assurances and influences from other people (e.g., doctors, spiritual leaders, family members) (Jermutus et al., 2022).

Finally, taking stock of the ever-expanding debate, general frameworks, conceptual and theoretical models on trust in AI have been developed. The academic literature developed a number of general theoretical models, focusing on the nomological network of trust in AI; that is, antecedents of trust in AI and its consequences (Asan et al., 2020; Jacovi et al., 2021; Lansing & Sunyaev, 2016; Siau & Wang, 2018; Söllner & Leimeister, 2013). A data-centric approach, which considers data inputs and outputs to AI as a factor in trust, has been proposed by Thiebes et al. (2021). Frameworks of dimensions of trust in AI are increasingly developed (Gulati

et al., 2017; Siau & Wang, 2018; Starke et al., 2022). These commonly extend established trust dimensions from organizational and psychology literature (Gefen et al., 2003; Mayer et al., 1995; McAllister, 1995).

The industry and policy makers proposed a number of "trust in AI" frameworks. The aim is to guide ethical design and use of technology by formulating principles of trustworthy AI (Floridi & Cowls, 2021; Heer, 2018; Jobin et al., 2019; Rossi, 2018; Saif & Ammanath, 2020; Thiebes et al., 2021). Thus, the European Union's 2019 "Ethics Guidelines for Trustworthy AI",[10] establish four principles of trustworthy AI: respect for human autonomy, prevention of harm, fairness and explainability (Smuha, 2019). The OECD's "Tools for trustworthy AI", provide a framework to compare implementation tools for trustworthy AI systems.[11] Industry leaders in AI develop own trustworthy AI principles, frameworks, and policies, such as those by IBM[12] or Google.[13]

Limitations of existing approaches to AI trust

While there has been much progress on trust in artificial intelligence, there are notable limitations of the approaches taken to address this topic.

First, efforts to date have been narrowly focused, reflecting disciplinary traditions and objects of interest. For example, in much of the "technical" literature dealing with trust, the focus has been on algorithmic transparency, accountability, explainability and privacy. Hagendorff (2020, p. 103) argues, these measures dominate computer science and artificial intelligence literature because they are "easily operationalized mathematically and thus tend to be implemented in terms of technical solutions." Likewise, economic approaches favor easily quantifiable solutions rooted in information asymmetry. This resulted in the prevalence of "trust games" (e.g., Boero et al., 2009; Keser, 2003) and the wide-spread application of game theory (Kuipers, 2018). Some work in applied disciplines, such as software engineering, computational linguistics and conceptual modeling, proposed general models of trust which are typically motivated by specific challenges in these disciplines (Amaral et al., 2020; Amaral et al., 2019; Dokoohaki & Matskin, 2008; Golbeck et al., 2003; J. Huang & Fox, 2006). These models have benefits of exactness, internal consistency, and formality. However, it is unclear whether these models hold for all trust cases or only those which correspond to the pragmatic assumptions embedded in these models. Hence, the need exists to lay out a domain-invariant foundational framework for trust, which can also be used to evaluate the models of trust proposed in different disciplines.

Second, many of the existing guidelines associated with trust and ethical development of AI are in conflict with one another (Thiebes et al., 2021). For example, transparency is in the trade off with algorithmic performance, such as classification accuracy of machine learning models (Knight, 2017). There is also a tension between privacy and performance. More powerful AI models may be achieved with the collection and usage of more granular personal data (Harari, 2016; Thiebes et al., 2021). Considering these tensions, it is difficult to rely on input from industry leaders on how to enhance trust in AI. It would be analogous to inmates running their own prison, suggesting that more input is needed from beyond the industry itself; from those without vested interest.

Finally, surprisingly, little guidance on the matter of trust in AI comes from the foundational reference disciplines: philosophy, psychology, and sociology. Naturally, philosophy, psychology, sociology have been referenced extensively in trust-oriented research in applied disciplines, such as organizational studies, information systems, human computer interaction. However, these disciplines tackle specific problems, such as trust when adopting e-commerce applications (Gefen et al., 2003), recommender technologies (Komiak & Benbasat, 2006) or economic dealings with robots (Schniter et al., 2020). Relative to research on applied areas of trust, fundamental theoretical work has been scarce.

The paucity of theoretical work is well-understood in these reference disciplines. One of the seminal scholars on psychology of trust, Simpson expresses "surprise", that despite the "centrality of trust in relationships" the topic did not receive "widespread theoretical and empirical attention" in psychology (Simpson, 2007, p. 264). Similarly, trust is "surprisingly" ignored in moral philosophy, an obvious discipline for philosophical examinations of trust (Baier, 1986, p. 232). Luhmann (2000, p. 94) makes the same observation in his discipline: "trust has never been a topic of mainstream sociology." In particular, we continue to lack a unified Foundational Trust Framework grounded in basic theoretical notions.

Considering the limitations of existing approaches to trust in AI, it is not surprising to observe persistent criticisms of the AI industry for insufficient trust-building measures related to ethical behavior when developing and implementing AI (Vardi, 2022). Academics, policy makers, and thought leaders widely recognize the need to develop more effective approaches to trust in AI systems, which would be actionable and acceptable to both the AI industry and the public at large (Financial Times, 2021; Hagendorff, 2020; Vardi, 2022).

[10] https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

[11] https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm

[12] https://www.ibm.com/watson/trustworthy-ai

[13] https://ai.google/principles/

Foundational Trust Framework

Research to date has taken a variety of perspectives on trust in AI. These include technical, psychological, economic, organizational and philosophical-ethics approaches. While these domains overlap, little attempt has been made to integrate them into a unified approach.

A more effective approach to tackling trust in AI begins with a better understanding of the foundations of this complex issue. We need to establish the basics and fundamentals to have a solid foundation for debate and development of the solutions. This was the original intent of Luhmann (2018), who offered perhaps the most extensive theory of trust. We are motivated by this effort and further formalize and extend Luhmann to build grounded, rigorous, and fruitful foundation for future studies on trust and trust in AI.

Foundations of trust based on ideas of Niklas Luhmann

From social perspective, fundamental contributions to trust have been made by Niklas Luhmann (Luhmann, 1995, 2000, 2018; Luhmann & Gilgen, 2012). Luhmann understood trust very broadly as confidence in one's own expectations viewing trust as an elementary and indispensable fact of social life. He viewed trust is a mechanism for reducing social complexity that works by generalization of expectations and gives order to an individual's inner understanding of complex outer environments. The concept of trust in Luhmann's works spans over *personal trust* towards individuals, as well as *system trust* towards social systems (Luhmann, 2018). System trust is realized if a system's behavior is reliable over time. Indeed, slight changes in input might cause large changes in the behavior of AI-based systems, potentially causing a decline in system trust. It requires permanent feedbacks. The expectation of explainable AI is that it supports sustainable system trust.

Luhmann (2018) proposes that trust relations are based on three structural components: (1) substitution by inner order, (2) need to learn, and (3) symbolic control. Trust increases order by reducing the complexity of an "outer" world by "inner" representations of a subject. It reduces complexity by elimination of action alternatives. Trust is not solely anchored in experience but built by generalizations of "trust judgments" so that trusting intentions can be transferred to similar cases. Luhmann calls this "symbolic fixation" of events in environments. Furthermore, trust is conditional and depends on feedback loops. In essence, trust is a means of making the "non-bypassable risk" of complexity tolerable. As Luhmann poignantly claims: "a complete absence of trust would prevent even getting up in the morning" (Luhmann, 2018, p. 4).

Luhmann (2018) highlights that learning of trust relations is barely understood but considers its foundation in phases, analogous to child development when building complex trust relations with other subjects, by expanding a subject's self to other subjects. In complex social orders, "trust in systems" supports our ability to connect with the decisions taken by others in that social system. Hence, trust abstracts concepts, such as money, truth, and power. Technical systems, including AI-based information systems, are considered part of complex social systems.

Luhmann (2018) argues that the social importance of trust lies in the ability to engage in social interactions in the absence of full transparency. The sheer impossibility of obtaining full transparency, even in simple matters, is vividly depicted in an anecdotal autobiographical account of Albert Szent-Györgyi, a Nobel laureate, first to isolate vitamin C (cited in von Bertalanffy, 1968, p. 5):

> [When I joined the The Institute for Advanced Study in Princeton], I did this in the hope that by rubbing elbows with those great atomic physicists and mathematicians I would learn something about living matters. But as soon as I revealed that in any living system there are more than two electrons, the physicists would not speak to me. With all their computers they could not say what the third electron might do.

Copious examples from diverse domains testify that humans can exhibit trusting behaviors, while lacking the understanding of the inner workings of the systems. When shopping online via an SSL protocol, driving over a suspension bridge, boarding an airplane, taking a train, driving a car, walking inside a building, or swiping a credit card, many (or most) humans lack detailed and nuanced knowledge of the technological mechanisms that underlie these systems and make them safe to use. The same goes for trust toward humans. For example, trust in a personal trainer occurs in the absence of the full knowledge of the innerworkings of the trainer's brain.

Considering the impossibility of attaining full knowledge of reality, even sometimes in simple matters, the question becomes: how can trust in AI be established? As Luhmann (2018) asserts, trust in fellow humans, or in technology, occurs and develops inside broader systems. Both AI technology, as well as the humans and who use it, are embedded and part of broader systems that shape and impact trusting beliefs. We hence adopt the systemist approach as a *foundation of trust in AI* to put the search for trust in AI on a solid theoretical footing.

Understanding the fundamental nature of trust: Foundational Trust Framework

We use systems theory to develop a general formalized understanding of the nature of trust in AI or *Foundational Trust Framework*. Systems theory is an umbrella term for several closely related and overlapping theories which deal with the nature of systems, their interactions and uses. Luhmann (Luhmann, 1995; Luhmann, 2018; Luhmann & Gilgen, 2012), developed a theory of trust and proposed to conceptualize trust as a mechanism to interact with social systems. However, Luhmann's theory lacks a number of focal constructs present in other systems theories, such as that of von Bertalanffy (1968), Ackoff (1971) and Bunge (1979, 2018). Likewise, Luhmann's ideas are not well-systematized and, at times, appear contradictory and confusing (Kroeger, 2019; Morgner, 2018). We, therefore, extend Luhmann's theory and synthesize it with other relevant propositions of systems theory. The result is a formalized set of propositions that establish trust as a process occurring in systems. These propositions form the Foundational Trust Framework, which we then used to understand the nature of trust in artificial intelligence.[14]

A system is a primitive and basic scientific and social concept (Ackoff, 1971; Bunge, 2003a; Luhmann, 1995, 2018; Luhmann & Gilgen, 2012; von Bertalanffy, 1968). To define system, we adopt an established definition by Ackoff (1971, p. 662): a system is "an entity which is composed of at least two elements ... each of a system's elements is connected to every other element, directly or indirectly. No subset of elements is unrelated to any other subset." This definition is not incompatible with that of Bunge, who defined systems as a "complex object every part or component of which is connected with other parts of the same object in such a manner that the whole possesses some features that its components lack – that is, emergent properties" (Bunge, 1996, p. 20). Effectively, the first definition focuses on the structural aspect of systems, whereas the second is on the consequences of having a systemic structure, namely, emergence.

Systems are often argued to be the principal building blocks of the universe. For example, Hawking and Mlodinow (2010) in writing that "a 'system' … could be a particle, a set of particles, or even the entire universe" imply everything is a system. Some make such claim explicit. Hence, Bunge, responding to recent advances in particle physics, became convinced that there are no structureless entities. The world, according to Bunge (2003a, p. 25) is "made up of interconnected systems." Bunge (2017) explains (p. 174, emphasis added):

> By calling all existents "concrete systems" we tacitly commit ourselves in tune with a growing suspicion in all scientific quarters - that there are no simple, structureless entities.

Bunge continues, that reality is made of systems "is a programmatic hypothesis found fertile in the past, because it has stimulated the search for complexities hidden under simple appearances" (Bunge, 2017, p. 174; see further discussion in: Lukyanenko, Storey, & Pastor, 2021b). Ludwig von Bertalanffy, the founder of systems theory, puts forward a similar argument: contemporary advances in sciences can be in part attributed to the adoption of the systems view of phenomena (von Bertalanffy, 1968).

Whether all entities are complex remains debatable; however, trust is a mechanism for the reduction of complexity (Luhmann, 2018). Given that complexity is the interactions of multiple components of a whole resulting in emergent properties or behavior (Bunge, 2003a; Johnson, 2002), trust as a concept presupposes existence of systems. Furthermore, we assume that only complex entities, namely, humans, potentially other animals (Griffin & Speck, 2004), and intelligent machines, are capable of exhibiting trust. Accordingly, we assume the trustees and trustors of all kinds be systems.

Thus, the adoption of a systemist perspective on trust is appropriate and could be uniquely suitable to study the fundamentals of this psychological and social mechanism. More formally:

> Proposition 1: All objects and subjects of trust are systems.

There are potentially as many systems as objects of thought or action (see on the notion of object in relationship to system in: Bunge, 2003b; Lukyanenko & Weber, 2022). Therefore, there is no such thing as a single notion of trust. Potentially, there are as many notions of trust as there are known systems. Indeed, Luhmann's (Luhmann, 1995; Luhmann, 2018) notion of *personal trust* under this view is also a systemic notion: it is a trust of one system (i.e., human) towards another system (e.g., another human or technology).

We can distinguish different kinds of systems. Luhmann (1995) focused on social systems; however, these systems are based on other, more fundamental kinds. Bunge (1996) distinguishes the following levels of systems: physical, chemical, biological, social, and technical. These systems emerge from another (e.g., social from biological) with higher level systems being made of components of systems of lower level. Hence, chemical molecules are made of physical atoms, whereas atoms are made of subatomic particles. Humans are biological systems (i.e., made of cells which

[14] Our analysis focuses on trust. However, the same arguments can be applied to distrust. As Luhmann (1995, 2018) pointed out, distrust mirrors trust and acts as a way to reduce complexity, except resulting in avoidance, rather than closeness.

are based on organic chemical components, such as amino acids). Organizations, such as corporations or universities, are social systems composed of humans and other systems (e.g., human artifacts).

Depending on the goal, other ways to classify systems exist (Ackoff & Gharajedaghi, 1996). Here, we note two distinctions among systems.

First, some systems are conceptual systems – specific kinds of systems that exist in the minds of humans.[15] Some contents of human mind can be conceptualized as conceptual systems (Bunge, 1979); that is, interconnected ideas, thoughts, propositions, and theories. Conceptual systems emerge from the biochemical operations of the human brain (Bunge, 2006).

Second, some systems are purposeful, in that they "produce (1) the same functionally defined outcome in different ways in the same environment, and (2) functionally different outcomes in the same and different environments." (Ackoff & Gharajedaghi, 1996, p. 13). These systems set and pursue their own goals and interact with their environment accordingly. Humans are purposeful systems and imbue their goals and aspirations in the systems they create, such as social and technological systems.

These distinctions become important when dealing with trust. The next proposition captures the levels and kinds of systems:

> Proposition 2: There exist systems at different levels and of different kinds.

The social world is conceptualized as a multilayered interconnected web of systems arising from human deeds and speech acts – externalizations of conceptual systems of humans (Searle, 1995, 2010). For example, a notion of a new conference dealing with trust and AI is a conceptual system inside the mind of one of the authors of this paper. A proposal about such conference expressed verbally or written in an email, is a speech act which may create such conference – a social system. The higher order entities, such as conferences, universities, corporations, governments, countries, are all social systems (Bailey, 1991; Buckley, 1967; Bunge, 1996; Dubin, 1978; Searle, 1995).

Systems have two kinds of properties (Bunge, 1979, 2018): properties of parts (termed intrinsic and *hereditary)* and properties of the systems themselves (termed *systemic*). Hereditary properties are properties of the components (which a system inheres from these components). For example, a charge of an electron is a hereditary property, which is an intrinsic property of the electron, a component of the broader physical system, atom. The income of a family is a hereditary (and intrinsic) property of the individual family members.

Some of the hereditary properties have direct and additive impact on the properties of the system. These systemic properties are directly derivable from the hereditary properties and are called *aggregate properties*. For example, the gross domestic product of all nations can be added up to the *global gross domestic product*, which is an aggregate property. A mass of the computer is the sum of the mass of its hardware components.

In contrast, *emergent properties* are those properties that the system components lack (Bedau, 1997; Bedau & Humphreys, 2008; Bunge, 2003a; von Bertalanffy, 1968). These properties emerge when the components become part of the whole and begin interacting with one another in a certain way. (See the notion of *mechanism* below). The emergent properties, unlike aggregate properties, are not directly derivable from the knowledge of the properties of the components. For example, *swarming* is an emergent property of some animal and artificial communities, including certain species of fish, bird, ants, bees and even robots (Hunt, 2019). Social cohesion is an emergent property of a social group.

Emergent properties shape emergent behavior of systems, or the changes in the properties over time. In biology, a swarm, for example, may overwhelm its prey or fend off a predator. These are emergent behaviors, rooted in the corresponding emergent properties, which an individual member of the swarming community does not possess.

As Luhmann suggests, to develop trust, humans acquire properties of the systems in question and share them with others (Luhmann, 2018, p. 42). Indeed, this sharing is one reason why social communications are valuable. They permit efficient reduction of uncertainty about the world (Luhmann, 1995). Combing the notions of hereditary and emergent properties with Luhmann's theory of trust, we obtain:

> Proposition 3: Trust in systems is a function of knowledge of properties of systems, both hereditary and systemic (aggregate and emergent).

Systems change in the virtue of the changes to their properties. Following Bunge (1977, 1996), we call these changes *events*. These changes may be random or form patterns. The strongest, most enduring of these patterns are known as *laws*. Laws are stable patterns which hold "independently of human knowledge or will" (Bunge, 1996, p. 27). For example, the law of gravity is an enduring universal pattern formed at the earliest moments of the Big Bang (Hawking & Mlodinow, 2010). Weaker patterns are social norms or cultural customs. These patterns are human-dependent and

[15] Our focus is human trust in AI. However, in principle other sentient beings, such as non-human animals, may exhibit trust. Furthermore, computational trust, whereby trust-like behavior is built into algorithms can also be conceptualized under a broad umbrella of trust.

change when humans who adhere to these patterns stop following them.

Multiple events form *processes*. A process is "a sequence, ordered in time, of events such that every member of the sequence takes part in the determination of the succeeding member" (Bunge, 2017, p. 172). For example, corrosion of metal due to oxidation (loss of electrons as part of a chemical reaction), voting in elections and booking a flight are different processes. The more stable the patterns which underlie the events and processes of the system, the more predictable is the behavior of the system. Hence, rust forming on an exposed iron rivet of a metal bridge is more predictable than the outcome of an election in a democratic society.

Predictability of a process is a function of our knowledge of its inner-workings, or its *mechanism*. In most cases observing processes directly and understanding their underlying mechanisms is impossible, as much of reality is inaccessible to our direct observation (Archer, 1995; Bhaskar, 1978). Instead, we resort to forming hypotheses and theories about unobservable mechanisms of the systems of interest.

As humans, we mainly construct inferences about these underlying mechanisms based on the imperfect knowledge we have, by considering the past performance of a given system, or the trust in this system by other people, whose opinions we respect. Consequently, even opaque systems can be trusted. Accordingly, we propose:

> Proposition 4: Trust in systems is a function of the stability and predictability of its events and processes; some events and processes are inherently more stable and, hence, predictable, than others.

Since not all events and processes are known or can be observed directly and fully, we further stipulate:

> Proposition 4a: Trust in systems is a function of the knowledge of the stability and predictability of its events and processes.

We assume that all systems interact with other systems. That is, there are no closed systems (Bunge, 2006; Lukyanenko et al., 2022). Arguably, all systems are open systems, because even tightly controlled laboratory experiments do not occur in complete isolation from the environment (Bhaskar, 1978). According to modern quantum theory, all systems may potentially interact with one another (Hawking & Mlodinow, 2010). Another version of this idea is the famous butterfly effect – that the flap of a butterfly's wings in Brazil can set off a tornado in Texas (Abraham & Ueda, 2000; Lorenz, 1972).

When systems interact with one another, the result may be: an alteration, acquisition, loss of properties of systems, or creation or destruction of systems. Hence, when two molecules encounter one another, they may fuse, creating a new chemical compound. Likewise, when two people meet and like each other, they may decide to get married, thereby creating a new social system – family.

The interactions happen not only between systems, but also within systems. Systems at lower levels impact systems at higher levels and vice versa. For example, individual voters (components of the social system, their country) may change the direction of the country due to their votes. This is *micro-to-macro* direction of an interaction. In contrast, the political program adopted by a country (e.g., isolationism), may impact how the citizens behave. This is *macro-to-micro* direction of systemic interaction.

While all systems interact with other systems, humans (or other agents of trust), may not be aware of all systemic interactions. Proximal chains – where one system directly impacts another - are what commonly get noticed. In other words, pragmatically, the systems in direct contact with one another are the ones that "make a difference to each other" (Rosemann & Green, 2002, p. 82).

For example, voting "in the context" of an ongoing armed conflict influences the voter turnout, and, in most cases, has an impact on how people vote. A model of the interaction will be incomplete if it did not account for the obvious, commonly proximal, systems interacting with the focal system. In the example of the armed conflict and voting, the systems would include the voters and the belligerent parties, who could coerce people to vote a certain way or abstain from voting.

Since the systems by interacting alter the properties of each other, they may affect trust. This is captured in an old Roman saying: If you lie down with dogs, you get up with fleas, or in Latin, *qui* cum *canibus concumbunt* cum *pulicibus surgent*. However, the point is not only about bad influence upon others, but, rather, more general. Since trust reduces uncertainty about the world, knowing which systems a focal system interacts, enables a better understanding of the nature and behavior of the focal system. More formally:

> Proposition 5: Trust in systems is a function of the knowledge of the interaction of this system with other systems.

Each system in the world is unique in some way. No two systems are identical. Even artificial systems, such as two seemingly identical pencils created by a standardized manufacturing process, are different systems, in that they occupy different spaces in the universe, and contain slightly different histories (e.g., one was created before another). Indeed, strictly speaking, the composition of the two pencils would also differ, as precise control over the arrangement of subatomic particles, is beyond current ability of humanity.

However, as already mentioned, reality contains a number of regularities – reflecting the common fundamental forces that act upon matter. Consequentially, all systems are similar

to one another in one or more ways (Goodman, 1972). Systems with "one or more" properties in common, form *classes* or *kinds* (Bunge, 2006, p. 13).

The notion of classes or kinds is important. Similar systems – or systems of the same class -- exhibit similar behavior. Hence, knowing properties of one system makes it possible to infer properties of another system of the same kind (Medin & Schaffer, 1978; Murphy, 2004; Parsons & Wand, 2008a, 2008b; Rosch, 1977). The more coherent the classes (i.e., the more similar or nearly identical its members are), the stronger and more justified the inferences regarding the properties of its members. For example, the more similar some birds are, the more justified we are in believing they will eat the same food. Related to this point, Luhmann argues, we form trust (or distrust) toward systems *in general*, such as toward bureaucratic organizations (Luhmann, 1995, p. 385). More formally, we state that trust can be directed towards *classes of systems*.

Discovering which systems are alike (members of the same class), or developing such systems (e.g., creating artifacts of the same class, such as of Tesla Y model, or of a deep learning neural network), allow humans to transfer trusting beliefs across different *individual* systems and across different *classes* of systems. The more similar the classes or its members, the greater our confidence that trust in a particular system can be applied to trust to all members of the system of that class. These arguments result in the following propositions:

> Proposition 6: Trusting beliefs can be transferred from particular systems to classes of systems, and vice versa.
> Proposition 6a: Trusting beliefs can be transferred from one class of systems to another class of systems.
> Proposition 6b: The more similar systems or classes of systems to one another and to other classes, the more trusting beliefs are transferred from one system or class of systems to another system or class of systems.

Open systems have internal components and boundaries. A system's boundary are those subsystems that directly interact with the environment, whereas those subsystems that only interact with other subsystems of its parent system are the internal components. For example, a public relations office of a company is part of its boundary, whereas its quality control department is an internal component. Since we assume that all systems are open systems, system openness is not of a kind, but of a degree. A secluded and self-sufficient monastery or the *Jarawas* of the Andaman Islands are less open than a roadside vegetable stall or a Shanghai Stock Exchange. Similarly, some artifacts are in constant interaction with other systems (e.g., social networking platforms, news aggregates), whereas others interact with other systems less frequently (e.g., forgotten JPEG file on a Windows computer).

Open systems are characterized by *equifinality* -- the same final state can be reached in open systems from different initial conditions and in different ways (von Bertalanffy, 1968). This means that, when dealing with open systems, it is not enough to know what the inputs are; rather, the knowledge of the internal components and mechanism is required to predict the behavior of such systems. For example, internal operations of a company may be hidden from its customers; yet these operations control how goods and services are delivered. Likewise, the logic of a machine learning model may not be accessible to its users. These issues partially explain the difficulty in building trust in complex systems, including those with AI components. Considering the notions of internal components and boundary together with Proposition 4 (trust in systems is a function of knowledge of properties of systems), we obtain a corollary proposition to Propositions 4:

> Proposition 7: Trust in systems is a function of knowledge of properties of internal and boundary components of systems and its mechanisms.

We now turn to where trust is formed; that is, in agents, which can be humans and others capable of exhibiting trust. We assume only humans are presently capable of experiencing the subjective feeling and conceptualization of trust. At the same time, modern intelligent machines can be programmed to act in a trusting manner. This means they can put themselves in a position of vulnerability toward another system – that is, allow another system to access and alter its internal components - based on an assumption that the other system may not take advantage of their exposure. As an example, consider intrusion-tolerant systems. These AI-based systems are designed to leverage the learned knowledge of the properties of other systems to estimate the probability of these systems to be malicious and harmful (Verissimo et al., 2009). Such systems can be called trusting, under a behavioral interpretation of trust. Henceforth, we use the notion of an *agent of trust* (or agents) to refer to humans and machines generally and use the specific concepts of humans vs. trusting machines when the distinction is necessary.

An agent that exhibits trusting behavior is itself a system. Consequently, we can apply the systemic notions to that system. Notions of hereditary and systemic properties, composition, environment, structure, and mechanism of systems become relevant. Specifically, what is known as *dispositional factors* in the psychology and social science trust literature, are under our framework *trust-related properties of agents*. They can be as either hereditary or systemic properties. For example, the patterns of brain activity corresponding to trusting behaviors (Dimoka, 2010) are hereditary properties, whereas conscious feeling of trusting someone is emergent.

The trust-related properties of agents develop in the context of systems interacting with other systems. Indeed, trust research has long recognized, for example, that human trust is affected by influences from friends on a social network, especially if the influencers have more direct knowledge or experience with the objects of trust (Jermutus et al., 2022; Kubiszewski et al., 2011). Similar influences may occur in the context of AI-based systems. For example, deep learning neural networks may embed other (pre-trained) neural networks (a process known as transfer learning) (Bengio, 2012; Pratt, 1992), which can transfer the trust knowledge from one system to another. We capture these ideas in a general proposition:

Proposition 8: Trust is a function of properties of trusting agents.

The agents of trust are purposeful systems in that they pursue certain objectives when interacting with the environment. Consequentially, trust is contingent on the purpose of the interaction. For example, when the purpose is to invest one's life's savings, trust in an investment broker is of paramount concern. Conversely, if the purpose is to casually inquire on whether the broker experiences increased foot traffic in the location, trust in the same broker is less consequential.

Luhmann (2018) suggests that trust is always required for social exchange. Even in the cases of casual interaction, trust happens in the background, barely below conscious awareness; nonetheless it remains critical. When asking someone about how they feel, we trust that this person will not harm us as part of this innocent verbal exchange. Hence, under the assumption that trust is always present, we can view the purpose of the interaction as a moderator upon trust. Specifically, the purpose determines how many, and the degree of detail, of the properties of the system under consideration that the agents of trust should analyze and consider when developing trust.

When the attainment of the goal is of no significant value, fewer properties of the system in question need to be evaluated. When the purpose is concerned with a mission-critical or life-threatening objective, a more exhaustive evaluation of the system is expected. This leads to the following proposition:

Proposition 9: The purpose of interaction moderates the formation of trust by focusing on specific properties of the target system.

Finally, we now define the concept of trust from the systemist perspective. We first define of concept of *human trust*, followed by the general trust definition.

Based on Luhmann (2018), trust is a mechanism for reducing complexity in the real-world. Trust allows agents of trust to act in the world in the absence of full information about all the relevant systems (their properties, history, etc.). Specifically, from systems theory perspective, to know a system fully is to know its properties, composition, environment, structure, and mechanism as well as its history. Clearly, such possibility rarely, if ever, exists. As in reality, the lack of full information is the norm, Luhmann famously asserted, that without trust humans would be paralyzed – completely unable to function in the world.

Trust acts as a bridge from imperfect knowledge to (level of trust-dependent) confident action. Hence, trust is a mental mechanism in humans, and a type of heuristic. Heuristics are mental mechanisms that are sufficient in most situations for a successful outcomes, but are not guaranteed to result in optimal solutions (Gigerenzer & Todd, 1999; Tversky et al., 1990). For example, to understand a problem, an effective heuristic may be to sketch it on paper as a model or diagram (Polya, 2004; Woo, 2011).

What kind of heuristic is trust? Given the propositions above, trust is a mental heuristic that focuses the limited attentive, cognitive, and affective resources of a human on those properties of a system under consideration, which, from the general knowledge and beliefs of the human, make the interaction with this system safe, comfortable, predictable or beneficial. Lacking the ability to ascertain these aspects of systems fully, trust is a mechanism for attending to, perceiving, conceptualizing, and memorizing those things that matter for safe and beneficial interactions. Effectively, trust is a filter upon reality. Furthermore, interaction with systems is continuous. In the process of interaction both the agents and the systems may change. New information may also become available. Hence, the filter needs to be constantly updated. In sum, trust is a *dynamic filter* upon the properties, composition, environment, structure, and mechanism, as well as history of systems (collectively referred to as *properties of systems*).[16]

The filtering of reality does not occur instantaneously. Indeed, trust develops over time, and may increase or decrease as more properties of the system under

[16] Much of prior research can be considered as investigating the specific properties or behaviors (i.e., changes of states of systems, events, processes) resulting in building trust or distrust. Common among these properties are benevolence, integrity, honesty, abilities, trustworthiness, credibility, adherence to ethical norms, predictability, goodwill, and character (Gefen et al., 2003; Glikson & Woolley, 2020; Jarvenpaa et al., 1998; Jarvenpaa et al., 2000; Mayer et al., 1995; McAllister, 1995). Likewise, predictability (e.g., McKnight et al., 1998) is the conclusion drawn from the knowledge of properties of systems, the laws that bind them, and the history that shows the stability of the system in adhering to these laws. Character (e.g., Giffin, 1967) is a summation of the properties of the system that represents its essence. In the same vein, for Fukuyama (1996), trust is built through expectations of regular, honest and cooperative behavior; this again, is a reference to properties of systems. Furthermore, consistent with research in social sciences, trust is cognitive, but also has an emotional or affective component (Glikson & Woolley, 2020; Komiak & Benbasat, 2006; McAllister, 1995; Schniter et al., 2020).

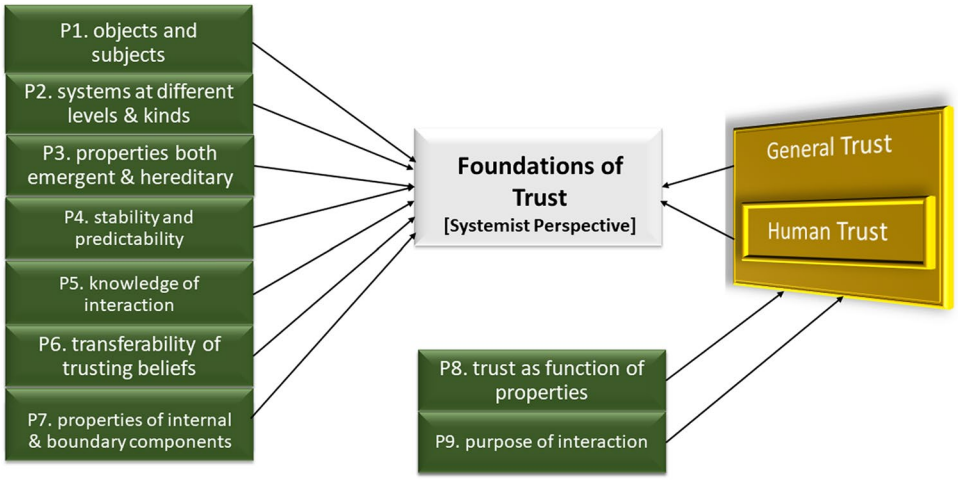

Fig. 1 Foundational Trust Framework

consideration become known (or as the system under consideration evolves). As humans (or agents) accumulate more and more consistent information about a target entity, the trust in this entity may increase or diminish (Luhmann, 2018; Rempel et al., 1985; Weber et al., 2004). Consistently with these arguments, rather than considering trust as a static set of beliefs, we view trust as a *process* occurring in humans.

Finally, as discussed, all systems impact other systems, directly or indirectly (via other systems). Indeed, avoidance is a form of interaction. It is an active stance that requires the alteration of behavior. Rather than conceptualizing trust as a binary: to interact or not interact, we suggest that trust determines the *extent* and the *parameters* of the interaction.

First, trust determines the *extent* of the interaction. We understand extent of the interaction in terms of system boundaries and interaction frequency. When we trust someone, we are more open to the object of trust. In systemic terms, we extend our *boundary* towards another system. In other words, more of our internal components become external. We may also interact with that system more often. Here, following Luhmann (2018), we assume that each interaction carries a risk, so the more often we interact, the greater risk we incur by interacting more often.

Second, trust determines the *parameters* of the interaction. The parameters are conditions of the interaction. For example, when we suspect someone is not trustworthy, we may require evidence in support of their claims. Notably, this is the stance of modern science: scientific claims before they can be published in reputable journals and conferences require evidence; the grander the claims, the greater the evidence required (Cronbach & Meehl, 1955; Larsen et al., 2020; Newton & Shaw, 2014). In contrast, full trust means unconditional acceptance without pre-conditions or evidence. Hence, the process of trust involves both mental and physical changes, where mental states shape the physical reactions, and physical sensory inputs shape the mental states. Combining our arguments about the nature of human trust, we define human trust as follows:

Definition 1: Human Trust. Human trust is a process within humans (mental, physiological) that considers the properties of another system to control the extent and parameters of the interaction with this system.

By generalizing human mental and physiological mechanisms to agents of trust, such as artificial intelligent agents, we obtain a general definition of trust:

> Definition 2: General Trust. Trust is an information-processing and behavioral process within a trusting agent that considers the properties of another system to control the extent and parameters of the interaction with this system.

The Definition 2 accounts for the growing number of cases where technologies are interacting with humans and other technologies directly, such as an Internet of Things device, or autonomous stock trading algorithm. In these technologies we can understand trust as designed procedures that controls how to interact with other systems (computers or humans) based on the consideration of their properties.

Summary of the Foundational Trust Framework

Propositions 1–9 and the definition of trust form the theoretical basis of the Foundational Trust Framework are shown in Fig. 1, where trust is considered from a systems perspective.

Table 2 summarizes the propositions and definitions of the Framework. These general propositions are broadly applicable to any context. They also become the basis for a deeper and more rigorous understanding the nature of trust in artificial intelligence.

Table 2 Propositions and definitions of the Foundational Trust Framework

Propositions and definitions	Definition
Proposition 1	All objects and subjects of trust are systems
Proposition 2	There exist systems at different levels and of different kinds
Proposition 3	Trust in systems is a function of knowledge of properties of systems, both emergent and hereditary
Proposition 4	Trust in systems is a function of the stability and predictability of its events and processes; some events and processes are inherently more stable and, hence, predictable, than others A) Trust in systems is a function of the knowledge of the stability and predictability of its events and processes
Proposition 5	Trust in systems is a function of the knowledge of the interaction of this system with other systems
Proposition 6	Trusting beliefs can be transferred from particular systems to classes of systems, and vice versa A) Trusting beliefs can be transferred from one class of systems to another class of systems B) The more similar systems or classes of systems to one another and to other classes, the more trusting beliefs are transferred from one system or class of systems to another system or class of systems
Proposition 7	Trust in systems is a function of knowledge of properties of internal and boundary components of systems and its mechanism
Proposition 8	Trust is a function of properties of trusting agents
Proposition 9	The purpose of interaction moderates the formation of trust by focusing on specific properties of the target system
Definition 1 Human Trust	Human trust is a process within humans (mental, physiological) that considers the properties of another system to control the extent and parameters of the interaction with this system
Definition 2 General Trust	Trust is an information-processing and behavioral process within a trusting agent that considers the properties of another system to control the extent and parameters of the interaction with this system

Trust in AI and trust in AI research agenda

We now apply the systems notions captured in the *Foundational Trust Framework* to the context of trust in artificial intelligence. This application permits rethinking the nature of trust in AI and suggests opportunities for future research.

Human-AI systems

A basic implication of the Foundational Trust Framework is that the AI technology, organizations, and users of AI are fundamentally systems. Collectively, we call them *human-AI systems*, to underscore the pivotal role of humans and AI, while not discounting the important contribution of other involved systems, such as organizations. Formally, we define human-AI systems as:

> Definition 3: Human-AI Systems. Human-AI systems are socio-technical systems composed of AI-based technologies, humans, and other systems that interact with, or are potentially affected by, the AI-based technologies.

Humans are the key components of Human-AI Systems. They harbor trusting processes (Definition 1). Humans are complex systems and are parts of other systems. Some examples of these, latter systems, include families, companies, professional and social networks (Barabási, 2003; Lazer et al., 2009).

The AI-based technology is a technical system; that is, it is comprised of hardware and software components. At a bare minimum, an AI can be a simple decision model composed of rules learned from previously supplied examples, along with the procedures for applying these rules to unforeseen cases. The rules themselves can be conceptualized as subsystems, composed of variables and operations. From here, AI systems only become more complex, as components become more nuanced and elaborate. Hence, an IBM Watson is an AI made of natural language processing, information retrieval, knowledge representation, automated reasoning, and machine learning modules.[17]

To effect change in the world, AI needs to be executed via programming code on some hardware. This could be an ordinary laptop or a supercomputer. The hardware of IBM Watson at the time of winning Jeopardy! in 2011, was composed of[18]:

> a cluster of ninety IBM Power 750 servers (plus additional I/O, network and cluster controller nodes in 10 racks) with a total of 2880 POWER7 processor cores and 16 Terabytes of RAM. Each Power 750 server uses a 3.5 GHz POWER7 eight core processor, with four threads per core.

Many AI systems interact with, or are part of, other technical systems. For example, AI is a core component of driverless cars (Kirkpatrick, 2022; J. D. Lee & Kolodge, 2020;

[17] https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/
[18] https://www.csee.umbc.edu/2011/02/is-watson-the-smartest-machine-on-earth/

Waldrop, 2015). The driverless cars, in turn, interact with a variety of other systems (e.g., roads, pedestrians, traffic signs).

Since AI permeates more and more facets of human life, trust in AI involves the consideration of larger and larger physical, biological, conceptual, social, and technical systems, as components or containers for human-AI systems. The main implication of viewing human-AI entities as systems is the potential utility of applying systems theory to better understand trust in these systems.

Research opportunity Investigate the antecedents, processes, and outcomes of trust in human-AI systems by adopting the systems theoretical perspective of the Foundational Trust Framework.

Definition of trust in AI

Using the Foundational Trust Framework enables us to provide a systems-based definition of trust in AI. For this, we use Definition 1 of human trust, because our focus is on human-AI systems, and combine it with the propositions of the framework that deal with properties of systems and classes of systems (Propositions 3–7). From this synthesis, we define human trust in AI as:

Definition 4: Human Trust in AI. Human trust in AI is a human mental and physiological process that considers the properties of a specific AI-based system, a class of such systems or other systems in which it is embedded or with which it interacts, to control the extent and parameters of the interaction with these systems.

There are several properties of this definition that set it apart from common work on trust and trust in AI (cf. definitions in Table 1). First, the definition does not specify the typical mental states of trust, such as benevolence, or integrity. By abstracting from the specific mental states, the definition facilitates additional research on the relevant mental states involved in trust. After all, no canonical states of trust have been established, with substantial debates on this topic. Our definition both circumvents this debate, and encourages the exploration of mental states and physical processes not adequately studied (e.g., those dealing with emotions) (McAllister, 1995).

Second, the focus of the definition is not on static states, but rather on the *dynamics of trust* – its process and mechanism. This is consistent with more recent work on trust and trust formation that investigates how trust develops over time (Glikson & Woolley, 2020; Komiak & Benbasat, 2008; Lumineau & Schilke, 2018; Weber et al., 2004). However, much remains unknown about this process (Stackpole, 2019). Thus, the process of trust formation and evolution in AI remains a major open research gap, as captured in the questions: *How does trust in AI emerge, evolve, and dissolve? What psychological mechanisms underlie the development of trust in AI?*

At the same time, the definition continues to underscore the importance of specific mental states. Recall that processes, from a systemic perspective, are sequences of states. Hence, the definition also supports research that considers trust as a mental state (Rousseau et al., 1998), while underscoring a wider research gap on trust as a process.

Third, the definition specifically deals with trust not only in a particular AI technology, but also in a class of technologies. Thus far, research on trust had predominantly focused on classes of technologies (e.g., anthropomorphic recommender agents). In prior studies, users interacted with a specific AI, with the arguments and conclusions always drawn with respect to a class of a technologies, such as recommender agents, driverless cars, virtual agents, e-vendors, robots, or chatbots (e.g., Glikson & Woolley, 2020; Lansing & Sunyaev, 2016; Renner et al., 2022). Our definition underscores the importance of examining both specific trust, as well as trust in a class of technologies. A research question that has not been well addressed so far is: *How does trust transfer from a particular AI technology to a class of technologies?*

Fourth, trust in the broader systems in which AI is embedded due to AI being a component of the system, has, so far, escaped much research. This is a *micro-to-macro* direction of trust. Most research has examined the relationship in the opposite direction. Specifically, studies have examined the influence of trust in a broader sociotechnical system as a mechanism that builds trust in a particular technology that harbors the system in question (Gefen et al., 2003; Renner et al., 2022). However, with the ubiquity of AI, humans many also transfer trust from AI to the social or other technical systems within which AI is embedded. Trust in laptops, for example, was enhanced by placing a sticker "Intel inside," thereby transferring trust from a well-known and reputable component, the central processing unit produced by Intel, to the broader technical system, that is the entire laptop (Anati et al., 2013; Davis, 2002). By analogy, we can hypothesize that some successful and high-profile AI may give credence to the organizations that adopt it.

Furthermore, trusting beliefs (and hence, behavioral intentions) may also transfer from other systems with which an AI interacts – an equally unchartered territory for trust in AI research. More work is hence needed on *micro-to-micro* transfer of trust. Likewise, it might be possible to trust one part of the broader system in which different AI technologies are embedded, while distrusting other parts. This corresponds to *part-of-to-part-of* analysis of trust.

Finally, little research to-date has considered how trust in the *entire, global human-AI system* affects trust in a specific AI system (or *macro-to-micro* trust transfer). When it comes to the macro-micro transfer of trust, two opposing

currents collide. On the one hand, many successes of AI, especially in high profile, sensitive domains, and customer-facing applications (e.g., Siri finally providing the information you request), should add to the overall trust in AI as a technology. This may help users develop trusting beliefs toward a specific AI (e.g., an AI-powered medical diagnostic tool). At the same time, the many publicized failures, as well as concerns about the repercussions of AI being too invasive, add to its distrust and make it more difficult for trusting beliefs to develop. These concerns include the role of AI in surveillance capitalism, government control over individuals, threats to employment, and the potential for adversarial AI, among others (Bostrom, 2014; Park & Kim, 2022; Petersen et al., 2022; Vardi, 2022). A systematic investigation of these concerns is needed, as well as the ability to measure and model them.

Fifth, as in the Framework, the definition of human trust in AI does not make a claim that trust is required for safe or enjoyable or comfortable interaction. Indeed, it makes a more fundamental claim that trust is a prerequisite for interaction. This generalization makes it possible to explore various kinds of reasons for building trust in AI technology, such as ensuring safety, comfort, social harmony, as well as profitability, and economic utility. It also encourages the pursuit of broader dependent variables of human-AI trust (such as social harmony, human happiness and well-being).

Consistent with these properties of the human trust in AI definition, we encourage future studies to pursue the following research directions:

Research opportunity. Investigate relevant mental states and mechanisms that underlie the development of human trust in AI

Research opportunity. Investigate trust in specific AI technologies, and the manner in which trust is transferred from a specific technology to a class of technologies.

Research opportunity. Investigate a broad spectrum of dependent variables of human trust in AI.

Indirect, passive users and others affected by AI

Based on the definition of human-AI systems (Definition 3), *humans*, a key component of human-AI systems, are viewed broadly to include the users, potential users, policy makers, and others who interact or are potentially affected by AI-based technologies. In contrast, extant research on trust in AI predominantly focuses on humans who are in direct contact with AI; that is, AI users (Gefen et al., 2003; Glikson & Woolley, 2020; Mcknight et al., 2011; Renner et al., 2022). The latter work is of undisputed importance, as we continue to discover new facts about the nature of direct human-AI use. At the same time, an overlooked opportunity exists.

Our definition views humans who are part of human-AI systems broadly. Consistent with systems theory, which suggests that systems interact with other systems directly and indirectly (Propositions 5 and 7), we suggest that human participants of the human-AI systems include not only immediate users. Participants can be potential users, social influencers (such as friends or family), policymakers, developers of these systems, project managers, or other organizational and extra organizational actors, such as policy activists, policy makers or lawyers.

In addition, the AI users should also encompass passive and indirect users, such as patients, pedestrians, inmates; that is, those people who do not directly use the technology, but who are affected by AI's actions. Indeed, those people who do not directly use the AI technology, such as hospital patients, still interact with the broader socio-technical system; for example, an AI-powered hospital (Crawford & Calo, 2016). These cases are mostly ignored by current research on trust in AI. This leads to the following important research opportunity:

Research opportunity Investigate trust for indirect, passive users as well as for others affected by AI technology.

Complexity of human-AI systems

One of the implications of adopting a systems perspective is the heightened focus on the complexity of human-AI systems. Actual complexity in systems can be understood as the number of component-parts along with the way in which these parts are structured and interact with one another and with other systems (Bunge, 2003b; K. Li & Wieringa, 2000; Lukyanenko et al., 2022). In contrast, perceived complexity is human's interpretation and conceptualization of a system as being complex (K. Li & Wieringa, 2000; Schlindwein & Ison, 2004). Generally, the greater the actual complexity, the greater its perceived complexity. However, experience with systems may reduce perceived complexity (via such psychological heuristics as chunking).

Many human-AI systems are very complex. Progress in storage, data transmission, and computational power permit the realization of more complex AI algorithms (such as long short-term memory deep learning neural networks). We expect complexity of AI technology to continue increasing.

Likewise, as organizational theory suggests, organizations also progress by the way of increasing internal complexity (e.g., growing from one-dimensional functional structures to customer-oriented matrices) (Galbraith, 2014). This progression is further enabled by the progress in information technologies, including AI, permitting more nuanced and personalized product and service offerings. Complexity of human social, political and economic systems is expected to

increase as human development marches on (Harari, 2016; Lukyanenko et al., 2022).

The key implication of the complexity of human-AI interactions is the ever-growing importance of trust. As Luhmann, 2018 argues and the Foundational Trust Framework (human trust) suggests, trust becomes an indispensable mechanism for handling complexity. The greater the complexity, the more important are the issues of trust, thus motivating increased research on trust in AI. This also implies the need for a proactive stance on the part of the involved community, including scientists, industry, and policymakers. More attention and resources need to be dedicated to building and communicating trust in AI, and to ensure safe and beneficial interactions with this technology.

With the expectation of increasing complexity, trust in AI offers a fertile ground to test Luhmann's fundamental hypothesis of the role of trust in managing complexity. Likewise, the measures to increase trust, such as greater AI transparency, should be more important for those components of human-AI systems, which are more complex. Considering the growing complexity, the search for ways to make AI more transparent is only going to become more challenging over time. These considerations underlie two related research opportunities: one dealing with the understanding of the nature of trust; the other, with the way to leverage trust in a proactive manner in order to facilitate human-AI interactions.

Research opportunity: Investigating the contribution of trust to the development, adoption, and use of AI systems at various levels of complexity

Research opportunity: Investigating design principles that leverage trust to mitigate AI systems complexity

We further expect mechanisms for building trust to differ based on the kinds of systems involved. For example, Glikson and Woolley (2020) show that trust in robots vs virtual AI (e.g., recommender agents) develops differently. Humans tend to begin with lower trust in robotic AI but develop greater trust over time. This trajectory is reversed for virtual AI. Such findings have actionable implications for the design of AI, such that more trust-building measures may be needed at the onset of the use of robots, whereas more trust maintenance can be valuable for supporting virtual agents. Similarly, Lansing and Sunyaev (2016) build a trust-aware taxonomy of cloud services. These findings motivate research that considers trust in different kinds of AI. Ultimately, it would be useful to develop a taxonomy of AI with respect to trust.

Research opportunity: Investigate trust in different kinds of AI and develop an AI-trust taxonomy

System openness

All human-AI systems are *open systems*. This means they interact with other systems in their environment. As indicated in Proposition 7, trust in open systems is a function of knowledge of properties of internal and boundary components of systems and its mechanism (due to equifinality).

Indeed, both AI as well as users and organizations have internal components that might not be visible or even whose existence may not be known to their partner systems. This makes it more important to understand the inner workings of AI, humans, and organizations. This suggests significant research opportunities dealing with explainable and transparent AI. The current AI industry is dominated by complex machine learning models, such as those based on *deep learning*, which occurs when there are types of artificial neural networks that are composed of multiple layers to progressively extract higher-level features from the raw input (Goodfellow et al., 2016). The opacity of such models undermines the ability to understand and explain how and why such models make their decisions (Adadi & Berrada, 2018; Castelvecchi, 2016; D Gunning, 2016; Mueller et al., 2019; Storey et al., 2022). Likewise, more research is needed in psychology, neuroscience, cognitive science, human-computer interaction on understanding the inner workings of human mind, and trust formation, especially within the context of artificial intelligence (consistent with Proposition 8).

As AI becomes ubiquitous, the need for increasing transparency grows, as recognized by governments and policy makers. Under the European Union's "General Data Protection Regulation 2016/679," companies need to provide their customers with "meaningful information about the logic involved" in their computer programs (Article 13.2(f)). The right to explanation is being considered as the next basic human right (Selbst & Powles, 2018; Wachter et al., 2017).

However, as the Foundational Trust Framework shows, when dealing with complex open systems, attaining full explanation is not a realistic goal. Indeed, even the developers of AI systems do not fully understand the inner workings of their technologies (Storey et al., 2022). We also expect the sophistication of machine learning algorithms to outpace the efforts to make them more transparent. Hence, a research question becomes: *What are the minimally viable requirements for explanation to satisfy and meet the essential societal needs?* Or:

Research question: What are the requirements and boundaries of the right to explanation?

Another open, and ill-understood challenge is the impact of the degree of systems' openness on trust. Here, system openness can be measured as the number and intensity of interactions between components of the system and other systems in the environment. The more components interact with other systems, the less predictable the behavior of the system

becomes, so the impacts of these interactions need to be understood. This follows from our intuitive understanding of equifinality, as a property of all open systems (or Proposition 7).

We can posit that the more open the AI system is, the more challenging it is to trust it (and establish trust in it). Hence, a driverless car, whose mission is to transport people in an enclosed warehouse (a semi-closed social system), would be more trusted than the same car tasked with transporting people on the road. Indeed, high-profile failures of driverless cars have been attributed to the unpredictability of real-world road conditions where, alongside typical pedestrians, we can expect idiosyncratic behavior of people and wildlife. In one such episode, an autonomous Toyota hit a visually impaired athlete, to which the apologetic CEO of the company stated: "the incident showed that autonomous vehicles are 'not yet realistic for normal roads'" (Reuters, 2021). This statement implies different degrees of trust and distrust due to the openness of human-AI systems.

Despite the many consequential and life-threatening incidents, little research has been conducted on the nature of systems openness, including the AI technology itself, as well as the systems in which the technology is embedded. Note the importance of distrust. If we know the system to be extremely open and difficult to fully predict, distrust toward the system is an appropriate coping mechanism. Indeed, distrust is just as important in ensuring that the complexity of social interactions is reduced to manageable levels (Luhmann (2018)). Unfortunately, relatively little research has dealt with distrust (Benamati et al., 2006; Dimoka, 2010; Hsiao, 2003; Komiak & Benbasat, 2008). It is ill-understood when distrust toward AI is healthy and appropriate. Our framework suggests that the degree of openness should be a factor when considering the antecedents of distrust that future studies could investigate. We, thus, suggest the following research opportunities:

Research opportunity: Investigate the impact of different degrees of system openness on establishing and maintaining trust in AI.
Research opportunity: Investigate when cultivating healthy distrust toward AI is appropriate.

Nested and varied systems

As Proposition 2 asserts, all systems are composed of other systems. Likely, trust towards the exact same AI system would differ depending upon which broader system it is a component of. Indeed, there was very little consideration of trust when IBM Watson was playing in the game of *Jeopardy!* (Ferrucci, 2010). The issue of trust gained immediate prominence when clinics, such as MD Anderson, began adopting IBM Watson to diagnose and treat cancer (Davenport & Ronanki, 2018).[19]

[19] This example can also be relevant for the analysis of trust formation based on the purpose of the interaction (or Proposition 9).

The nesting character of trust has been subject of research, albeit generally not from the systems point of view. Gefen et al. (2003) identified a positive relationship between "structural assurances" and trust in e-vendors. Structural assurances include legal recourse, guarantees, and regulations "such as the Better Business Bureau's BBBOnline Reliability seal (www.bbb.com), the TRUSTe seal of the eTrust (www.etrust.com), or a 1–800 number" (Gefen et al., 2003, p. 65). From a systems point of view, these are *properties of other systems* with which a focal system (e-vendor) interacts (i.e., the Propositions 6 and 7 of our Framework). However, the importance of these properties and their respective systems have been treated in an incidental manner. Thus far, research mainly focused on the types of AI-based technologies, such as robots vs chatbots (Glikson & Woolley, 2020), but we lack the understanding of the nature of *other technological and socio-technological systems* involved in building trust.

Research opportunity: Investigate the impact of other systems, which interact with or embed AI, when researching trust in AI.

The landscape of human-AI systems is vast, with many kinds of technologies, communities, and people involved (Jobin et al., 2019). Our framework supports investigations on the contribution of other technological and socio-technological systems involved in building trust by, among other claims, asserting that trust can be transferred from systems of similar kinds and from classes of systems to instances, and vice versa (as per Propositions 6 and 7). This suggests the need to develop ontologies and taxonomies of systems, which would group together systems that have a similar impact on trust in AI.

Research opportunity: Design ontologies and taxonomies of systems involved in human-AI systems.

Summary

While the literature on trust in technology, including AI, is extensive, and continues to expand, notable gaps exist. The application of the Foundational Trust Framework reveals a vast unchartered territory of research opportunities. For example, as per this analysis, we continue to lack an understanding of how different components of human-AI systems interact. As AI technologies are invariably embedded in broader systems, trust transfer between these systems remains an ill-understood process. At the same time, very little research has been conducted on distal users and others potentially affected by AI. These, and many other research opportunities, indicate the benefits of adopting a

Table 3 Research opportunities and propositions of the Foundational Trust Framework

Propositions and definitions	Propositions
Research opportunity. Investigate the antecedents, processes, and outcomes of trust in human-AI systems by adopting the systems theoretical perspective of the Foundational Trust Framework	1–9
Research opportunity. Investigate relevant mental states and mechanism which underlie the development of human trust in AI	3–7
Research opportunity. Investigate trust in specific AI technologies, and the way trust is transferred from a particular technology to a class of technologies, from AI systems to broader social systems, and vise-versa.	3–7
Research opportunity. Investigate a broad spectrum of dependent variables of human trust in AI.	3–7
Research opportunity. Investigate trust for indirect, passive users as well as others affected by AI technology	5, 7
Research opportunity: Investigate the contribution of trust to the development, adoption, and use of AI systems at various levels of complexity	1–9
Research opportunity: Investigate design principles which leverage trust to mitigate AI systems complexity	1–9
Research opportunity: Investigate trust in different kinds of AI and develop an AI-trust taxonomy	6
Research opportunity: Investigate the impact of different degrees of system openness on establishing and maintaining trust in AI	7, 8
Research opportunity: Investigate when cultivating healthy distrust toward AI is appropriate	7,8
Research opportunity: Investigate the impact of other systems, which interact with or embed AI, on trust in AI	2
Research opportunity: Design ontologies and taxonomies of systems involved in human-AI systems	6,7
Research opportunity: Adapt and extend the Foundational Trust Framework into specific scenarios and domains of focus	N/A

systems perspective on trust in AI, which is our general recommendation.

Table 3 summarizes the research directions suggested here and the propositions of the Foundational Trust Framework upon which they are based. It shows that the research agenda we outlined refers to all aspects of the framework. This, however, does not mean the framework is comprehensive. Future research can continue drawing upon the Foundational Trust Framework to motivate other studies we did not explicitly consider here, such as the impact of the types of tasks involved on trust in AI (J. D. Lee & Kolodge, 2020; J. D. Lee & See, 2004).

There are also opportunities to extend the Foundational Trust Framework itself. The aim of the framework is to be general and unifying. Still, as any model, it has limitations, some unintentional, some, by design. For example, the Framework assumes a non-reciprocal interaction between an agent of trust and the target system. Future work can extend the framework by considering the reciprocal links between the two or more parties of trust. Despite the lack of explicit representation of this scenario in the Framework, it still indirectly supports modeling such interactions. The Framework treats all parties as systems (Proposition 1); hence, systems analysis can be applied recursively to each involved party. However, the Framework lacks the constructs for modeling some pertinent properties of reciprocal relationships, such as information asymmetry, dependency, and contingency. Similarly, by design, the Framework does not represent the value of the exchange, except, indirectly, as a property of the task, or as a property of the agent of trust (as a perception or belief). This too can be extended by a more direct modeling of values. The Foundational Trust Framework is domain-agnostic. However, it can be adapted to specific domains – indeed, we attempted just this task by applying it to trust in AI. Future studies can further adapt the framework for other domains (e.g., e-commerce, customer relationship, social media) or evaluate the domain-specific computational models of trust based on the general propositions of the Framework. Consistently, we propose the research opportunity:

> **Research opportunity** Adapt and extend the Foundational Trust Framework into specific scenarios and domains of focus.

Discussion and conclusions

With the rise of AI, often dubbed the pinnacle technology (Bostrom, 2014; Filippouli, 2017), the issue of trust in this technology emerges as a paramount concern. This stimulates a growing volume of trust in AI literature. This literature, however, remains fragmented, without a common foundation, which could integrate the different studies. The coverage of trust in AI so far has also been uneven. Topics, such as trust in robots or trust in medical AI systems, have received substantial scrutiny. In contrast, topics such as the mechanisms by which beliefs in a particular technology get transferred to a class of technologies, have not been actively pursued.

This research develops a *Foundational Trust Framework*. The Framework provides a conceptual, theoretical, and methodological foundation for trust research in general, and trust in AI, in particular. The framework positions trust in AI as a problem of interactions among systems. Doing so, permits the application of systems thinking and general systems theory. Guided by the seminal arguments of systems theory, the framework advances systems-grounded

propositions about the nature of systems and trust. It also offers a general definition of trust. The framework has the potential to develop a common foundation for varied trust research.

We illustrate the potential of the Foundational Trust Framework by applying it to trust in AI. This application yields the key focal object of research – human-AI systems. It also paves the way to an inclusive definition of human trust in AI. The application of the framework to trust in AI motivated several research opportunities. These research directions are derived from the propositions of the framework, while also extending issues related to trust in AI into unchartered territories. These research opportunities surface new questions that can facilitate further advances in empirical, theoretical and design research on trust in AI.

In addition to the research opportunities identified in the paper, the Foundational Trust Framework explicitly provides for other research opportunities related to trust and artificial intelligence. First, by conceptualizing the objects of trust in AI as general systems, the framework paves the way for studies where trust originates in nonhuman agents. Definition 2 explicitly supports this research.

Second, an intriguing possibility is the reversal of the relationship between AI and humans. With the continued progress in AI, a futuristic, but the already plausible question is: *What is the nature of AI's trust in humans?* In fact, avantgarde thinkers have already posited this question (Bostrom, 1998; Harari, 2016). It has already entered the public discourse in the form of movies, such as *The Matrix* or *Terminator*. Our framework facilitates and encourages reconfigurations in the relationship among systems, upon which the future of humanity could very well be based.

Regardless of the form from which the perspective of trust is taken, a foundation of trust based on systems is expected to be enduring. Likewise, trust in AI will continue to be of much societal concern and an important topic of future research.

Special issue on "Trust in AI" in Electronic Markets

This special issue sought contributions on trust in artificial intelligence. Below, we use our Foundational Trust Framework to briefly highlight the accepted papers.

- René Riedl, in the paper "Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions" presents a literature review that examines general and specific psychological characteristics of a user (universal and specific personality traits) in relation to technologies and AI technologies in particular. In many cases, AI-based information systems are used by individual users. The personality of the user plays a decisive role in the adoption of such information systems. Some users are fundamentally open to new technologies, whereas others are more skeptical. Thus, Riedl's contribution refers to the characteristics of a trusting agent (proposition 8 of the Foundational Trust Framework) (Riedl, 2022).
- Rongbin Yang and Santoso Wibowo in their paper "User trust in artificial intelligence: A comprehensive conceptual framework," likewise conduct a systematic literature review on user trust in artificial intelligence (AI) from different perspectives. The authors identify the various components, influencing factors, and outcomes of users' trust in AI. A comprehensive conceptual framework is proposed for a better understanding of users' trust in AI. The framework helps AI-supported service providers comprehend the concept of user trust from different perspectives. The findings highlight the importance of building trust based on different facets to facilitate positive cognitive, affective, and behavioral changes among the users. The authors' framework is specific to AI. The Foundational Trust Framework can be used in conjunction with this paper to better understand the nature of trust in AI (Yang & Wibowo, 2022).
- The paper "The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study" by Jonas Wanner, Lukas-Valentin Herm, Kai Heinrich, and Christian Janiesch reports on a study related to trust in AI. The authors show how contemporary decision support systems are increasingly relying on artificial intelligence technology that display human-like decision capacities that resemble a black box. Their research develops a theoretical model that explains end-user adoption of such intelligent systems. Their model is tested in an industrial maintenance workplace with the results suggesting that acceptance is performance-driven at first sight, but that transparency plays an important indirect role in regulating trust and the perception of performance. The paper underscores the importance of transparency, captured in the propositions 3–7 of the Foundational Trust Framework (Wanner et al., 2022).

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3173574.3174156

Abraham, R., & Ueda, Y. (2000). *The chaos avant-garde: Memories of the early days of chaos theory*. World Scientific. Retrieved September 10, 2022, from https://books.google.ca/books?id=olJqDQAAQBAJ

Ackoff, R. L. (1971). Towards a system of systems concepts. *Management Science, 17*(11), 661–671. https://doi.org/10.1287/mnsc.17.11.661

Ackoff, R. L., & Gharajedaghi, J. (1996). Reflections on systems and their models. *Systems Research, 13*(1), 13–23. https://doi.org/10.1002/(SICI)1099-1735(199603)13:13.0.CO;2-O

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adamczyk, W. B., Monasterio, L., & Fochezatto, A. (2021). Automation in the future of public sector employment: The case of Brazilian Federal Government. *Technology in Society, 67*, 101722. https://doi.org/10.1016/j.techsoc.2021.101722

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235–251). Elsevier. https://doi.org/10.1016/B978-0-12-214850-7.50022-X

Alfonseca, M., Cebrian, M., Anta, A. F., Coviello, L., Abeliuk, A., & Rahwan, I. (2021). Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research, 70*, 65–76. https://doi.org/10.1613/jair.1.12202

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing, 2*, 76–81. https://doi.org/10.1109/MIC.2013.20

Amaral, G., Guizzardi, R., Guizzardi, G., & Mylopoulos, J. (2020). Ontology-based modeling and analysis of trustworthiness requirements: Preliminary results. In G. Dobbie, U. Frank, G. Kappel, S.W. Liddle, & H. C. Mayr (Eds.), *Conceptual Modeling. ER 2020. Lecture Notes in Computer Science* (vol. 12400, pp. 342–352). Springer, Cham. https://doi.org/10.1007/978-3-030-62522-1_25

Amaral, G., Sales, T. P., Guizzardi, G., & Porello, D. (2019). Towards a reference ontology of trust. In H. Panetto, C. Debruyne, M. Hepp, D. Lewis, C. Ardagna, & R. Meersman (Eds.), *On the move to meaningful internet systems: OTM 2019 Conferences. OTM 2019. Lecture Notes in Computer Science* (vol .11877, pp. 3–21). Springer, Cham. https://doi.org/10.1007/978-3-030-33246-4_1

Anati, I., Gueron, S., Johnson, S., & Scarlata, V. (2013). Innovative technology for CPU based attestation and sealing. In *Proceedings of the 2nd international workshop on hardware and architectural support for security and privacy* (vol. 13, no. 7). ACM New York.

Archer, M. S. (1995). *Realist social theory: The morphogenetic approach*. Cambridge University Press.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Natesan Ramamurthy, K., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay J., & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through Supplier's declarations of conformity. *IBM Journal of Research and Development, 63*(4/5). https://doi.org/10.1147/JRD.2019.2942288

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research, 22*(6), e15154. https://doi.org/10.2196/15154

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

Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science, 252*(5013), 1668–1675. https://doi.org/10.1126/science.2047876

Barabási, A.-L. (2003). *Linked: The new science of networks*. Basic Books.

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives, 11*, 375–399.

Bedau, M. A., & Humphreys, P. E. (2008). *Emergence: Contemporary readings in philosophy and science*. MIT Press. https://doi.org/10.7551/mitpress/9780262026215.001.0001

Belchik, T. A. (2022). Artificial intelligence as a factor in labor productivity. In A. V. Bogoviz, A. E. Suglobov, A. N. Maloletko, & O. V. Kaurova (Eds.), *Cooperation and sustainable development. Lecture Notes in Networks and Systems* (vol. 245). Springer. https://doi.org/10.1007/978-3-030-77000-6_62

Benamati, J., Serva, M. A., & Fuller, M. A. (2006). Are trust and distrust distinct constructs? An empirical study of the effects of trust and distrust among online banking users. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 06*, 121.2. IEEE Computer Society. https://doi.org/10.1109/HICSS.2006.63.

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems, 6*(3), 72–101. https://doi.org/10.17705/1jais.00065

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 17–36.

Bhaskar, R. (1978). *A realist theory of science*. Harvester Press.

Bickley, S. J., Chan, H. F., & Torgler, B. (2022). Artificial intelligence in the field of economics. *Scientometrics*, 1–30. https://doi.org/10.1007/s11192-022-04294-w

Bodart, F., Patel, A., Sim, M., & Weber, R. (2001). Should optional properties be used in conceptual modelling? A theory and three empirical tests. *Information Systems Research, 12*(4), 384–405. https://doi.org/10.1287/isre.12.4.384.9702

Boero, R., Bravo, G., Castellani, M., & Squazzoni, F. (2009). Reputational cues in repeated trust games. *The Journal of Socio-Economics, 38*(6), 871–877. https://doi.org/10.1016/j.socec.2009.05.004

Boon, S. D., & Holmes, J. G. (1991). The dynamics of interpersonal trust: Resolving uncertainty in the face of risk. *Cooperation and Prosocial Behavior*, 190–211.

Bostrom, N. (1998). How long before superintelligence? *International Journal of Futures Studies, 2*(1), 1–9.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. Retrieved September 10, 2022, from https://books.google.ca/books?id=7_H8AwAAQBAJ

Brashear, T. G., Boles, J. S., Bellenger, D. N., & Brooks, C. M. (2003). An empirical test of trust-building processes and outcomes in sales manager-salesperson relationships. *Journal of the Academy of Marketing Science, 31*(2), 189–200. https://doi.org/10.1177/0092070302250902

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Buckley, W. (1967). *Sociology and modern systems theory.* Englewood Cliffs, New Jersey: Prentice-Hall.

Bunge, M. A. (1977). *Treatise on basic philosophy: Ontology I: The furniture of the world*. Reidel.

Bunge, M. A. (1979). *Treatise on basic philosophy: Ontology II: A world of systems*. Reidel Publishing Company.

Bunge, M. A. (1996). *Finding philosophy in social science*. Yale University Press.

Bunge, M. A. (2003a). *Emergence and convergence: Qualitative novelty and the unity of knowledge*. University of Toronto Press.

Bunge, M. A. (2003b). *Philosophical dictionary*. Prometheus Books.

Bunge, M. A. (2006). *Chasing reality: Strife over realism*. University of Toronto Press.

Bunge, M. A. (2017). *Philosophy of science: Volume 2, from explanation to justification*. Routledge. Retrieved September 10, 2022, from https://books.google.ca/books?id=NtwzDwAAQBAJ

Bunge, M. A. (2018). Systems everywhere. In *Cybernetics and applied systems* (pp. 23–41). CRC Press.

Castellanos, A., Tremblay, M., Lukyanenko, R., & Samuel, B. M. (2020). Basic classes in conceptual modeling: Theory and practical guidelines. *Journal of the Association for Information Systems, 21*(4), 1001–1044. https://doi.org/10.17705/1jais.00627

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News, 538*(7623), 20.

Cerf, V. G. (2019). AI is not an excuse! *Communications of the ACM, 62*(10), 7–7. https://doi.org/10.1145/3359332

Chamorro-Premuzic, T., Polli, F., & Dattner, B. (2019). Building ethical AI for talent management. *Harvard Business Review, 21*.

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). *Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits*. *60*(1), 841–845. SAGE Publications Sage CA: Los Angeles, CA.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*(7625), 311–313. https://doi.org/10.1038/538311a

Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence*. Basic Books.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281. https://doi.org/10.1037/h0040957

Davenport, T. H., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94. https://doi.org/10.7861/futurehosp.6-2-94

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review, 96*(1), 108–116. https://hbr.org/2018/01/artificial-intelligence-for-the-real-world

Davies, I., Green, P., Rosemann, M., Indulska, M., & Gallo, S. (2006). How do practitioners use conceptual modeling in practice? *Data & Knowledge Engineering, 58*(3), 358–380. https://doi.org/10.1016/j.datak.2005.07.007

Davis, E. (2015). Ethical guidelines for a superintelligence. *Artificial Intelligence, 220*, 121–124. https://doi.org/10.1016/j.artint.2014.12.003

Davis, S. (2002). Brand asset management: How businesses can profit from the power of brand. *Journal of Consumer Marketing, 19*(4), 351–358. https://doi.org/10.1108/07363760210433654

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly, 34*(2), 373–3A7. https://doi.org/10.2307/20721433

Dobing, B., & Parsons, J. (2006). How UML is used. *Communications of the ACM, 49*(5), 109–113. https://doi.org/10.1145/1125944.1125949

Dokoohaki, N., & Matskin, M. (2008). Effective design of trust ontologies for improvement in the structure of socio-semantic trust networks. *International Journal On Advances in Intelligent Systems, 1*(1942–2679), 23–42.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books. Retrieved September 10, 2022, from https://books.google.com/books?id=WpTSDQAAQBAJ

Dosilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE. https://doi.org/10.23919/MIPRO.2018.8400040

Dubin, R. (1978). *Theory building*. Free Press. Retrieved September 10, 2022, from http://books.google.ca/books?id=a0NqAAAAMAAJ

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. Cambridge, MA: MIT Press. Retrieved September 10, 2022, from https://books.google.ca/books?id=72yuDwAAQBAJ

Ellul, J. (2022). Should we regulate artificial intelligence or some uses of software? *Discover Artificial Intelligence, 2*(1), 1–6. https://doi.org/10.1007/s44163-022-00021-9

Eriksson, O., Johannesson, P., & Bergholtz, M. (2019). The case for classes and instances-a response to representing instances: The case for reengineering conceptual modelling grammars. *European Journal of Information Systems, 28*(6), 681–693. https://doi.org/10.1080/0960085X.2019.1673672

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

Faulkner, P., & Simpson, T. (2017). *The philosophy of trust*. Oxford, England: Oxford University Press. Retrieved September 10, 2022, from https://books.google.ca/books?id=YIgLDgAAQBAJ

Ferrucci, D. (2010). Build watson: An overview of DeepQA for the Jeopardy! challenge. *Proceedings of the 19th International Conference on Parallel Architectures and Compilastion Techniques*, 1–2. https://doi.org/10.1145/1854273.1854275

Fettke, P. (2009). How conceptual modeling is used. *Communications of the Association for Information Systems, 25*(1), 43. https://doi.org/10.17705/1CAIS.02543

Fettke, P. (2020). Conceptual modelling and artificial intelligence: Overview and research challenges from the perspective of predictive business process management. In *Joint Proceedings of Modellierung 2020 Short, Workshop and Tools & Demo PapersWorkshop on Models in AI* (pp. 157–164).

Filippouli, E. (2017). *AI: The pinnacle of our ingenuity*. Retrieved September 28, 2022, from Global Thinkers Forum website: https://www.globalthinkersforum.org/news-and-resources/news/ai-the-pinnacle-of-our-ingenuity

Financial Times. (2021). *Building trust in AI systems is essential*. Financial Times. Retrieved September 10, 2022, from https://www.ft.com/content/85b0882e-3e93-42e7-8411-54f4e24c7f87

Floridi, L. (2019). Should we be afraid of AI? *Aeon Magazine*.

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In *Ethics, governance, and policies in artificial intelligence* (pp. 5–17). Springer. https://doi.org/10.1002/9781119815075.ch45

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

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

Galbraith, J. R. (2014). *Designing organizations: Strategy, structure, and process at the business unit and enterprise levels*. Wiley. Retrieved September 10, 2022, from https://books.google.ca/books?id=KVd5AgAAQBAJ

Garcia-Retamero, R., Müller, S. M., & Rousseau, D. L. (2012). The impact of value similarity and power on the perception of threat. *Political Psychology, 33*(2), 179–193. https://doi.org/10.1111/j.1467-9221.2012.00869.x

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly, 27*(1), 51–90. https://doi.org/10.2307/30036519

Giffin, K. (1967). The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological Bulletin, 68*(2), 104. https://doi.org/10.1037/h0024833

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press USA.

Gill. (2020). *Whoever leads in artificial intelligence in 2030 will rule the world until 2100*. Brookings. Retrieved September 25, 2021 from https://www.brookings.edu/blog/future-development/2020/01/17/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100/

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. https://doi.org/10.1016/j.chb.2020.106607

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Golbeck, J., Parsia, B., & Hendler, J. (2003). Trust networks on the semantic web. In *Cooperative information agents VII* (Vol. 2782, pp. 238–249). Springer. https://doi.org/10.1007/978-3-540-45217-1_18

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 437–447).

Griffin, D. R., & Speck, G. B. (2004). New evidence of animal consciousness. *Animal Cognition, 7*(1), 5–18. https://doi.org/10.1007/s10071-003-0203-x

Gulati, S., Sousa, S., & Lamas, D. (2017). Modelling trust: An empirical assessment. In R. Bernhaupt, G. Dalvi, A. K. Joshi, D. Balkrishan, J. O'Neill, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2017. INTERACT 2017. Lecture Notes in Computer Science* (vol 10516). Springer. https://doi.org/10.1007/978-3-319-68059-0_3

Gunning, D. (2016). Explainable artificial intelligence (XAI). Defense advanced research projects agency. *Defense Advanced Research Projects Agency (DARPA), nd Web, 2*.

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine, 40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Haenlein, M., Huang, M.-H., & Kaplan, A. (2022). Guest editorial: Business ethics in the era of artificial intelligence. *Journal of Business Ethics, 44*(1), 1–3. https://doi.org/10.1007/s10551-022-05060-x

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Random House.

Hawking, S., & Mlodinow, L. (2010). *The grand design*. Random House Digital, Inc.

Heer, J. (2018). The partnership on AI. *AI Matters, 4*(3), 25–26. https://doi.org/10.1145/3284751.3284760

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

Holcomb, S. D., Porter, W. K., Ault, S. V., Mao, G., & Wang, J. (2018). Overview on deepmind and its alphago zero AI. In *Proceedings of the 2018 International Conference on Big Data and Education* (pp. 67–71). https://doi.org/10.1145/3206157.3206174

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics, 3*(2), 119–131. https://doi.org/10.1007/s40708-016-0042-6

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., & Palade, V. (2019). Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Applied Intelligence, 49*(7), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5

Hsiao, R.-L. (2003). Technology fears: Distrust and cultural persistence in electronic marketplace adoption. *The Journal of Strategic Information Systems, 12*(3), 169–199. https://doi.org/10.1016/S0963-8687(03)00034-9

Huang, J., & Fox, M. S. (2006). An ontology of trust: Formal semantics and transitivity. In *International Conference on Electronic Commerce* (pp. 259–270). https://doi.org/10.1145/1151454.1151499

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

Hunt, E. (2019). *The social animals that are inspiring new behaviours for robot swarms*. The Conversation. Retrieved September 10, 2022, from http://theconversation.com/the-social-animals-that-are-inspiring-new-behaviours-for-robot-swarms-113584

Hvalshagen, M., Lukyanenko, R., & Samuel, B. M. (2023). Empowering users with narratives: Examining the efficacy of narratives for understanding data-oriented conceptual models. *Information Systems Research*, 1–38. https://doi.org/10.1287/isre.2022.1141

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624–635). https://doi.org/10.1145/3442188.3445923

Jarvenpaa, S. L., Knoll, K., & Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Information Systems, 14*(4), 29–64. https://doi.org/10.1080/07421222.1998.11518185

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791–815. https://doi.org/10.1287/orsc.10.6.791

Jarvenpaa, S. L., Tractinsky, N., & Vitale, M. (2000). Consumer trust in an internet store. *Information Technology and Management, 1*(1), 45–71. https://doi.org/10.1023/A:1019104520776

Jermutus, E., Kneale, D., Thomas, J., & Michie, S. (2022). Influences on user trust in healthcare artificial intelligence: A systematic review. *Wellcome Open Research, 7*, 65. https://doi.org/10.12688/wellcomeopenres.17550.1

Jia, K., Kenney, M., Mattila, J., & Seppala, T. (2018). *The application of artificial intelligence at Chinese digital platform giants: Baidu* (p. 81). ETLA Reports.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, S. (2002). *Emergence: The connected lives of ants, brains, cities, and software*. Simon and Schuster.

Keser, C. (2003). Experimental games for the design of reputation management systems. *IBM Systems Journal, 42*(3), 498–506. https://doi.org/10.1147/sj.423.0498

Khatri, V., Vessey, I., Ramesh, V., Clay, P., & Park, S.-J. (2006). Understanding conceptual schemas: Exploring the role of application and IS domain knowledge. *Information Systems Research, 17*(1), 81–99. https://doi.org/10.1287/isre.1060.0081

Kirkpatrick, K. (2022). Still waiting for self-driving cars. *Communications of the ACM, 65*(4), 12–14. https://doi.org/10.1145/3516517

Kiron, D., & Schrage, M. (2019). Strategy for and with AI. *MIT Sloan Management Review, 60*(4), 30–35.

Knight, W. (2017). DARPA is funding projects that will try to open up AI's black boxes. *MIT Technology Review*.

Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly, 30*(4), 941–960. https://doi.org/10.2307/25148760

Komiak, S. Y. X., & Benbasat, I. (2008). A two-process view of trust and distrust building in recommendation agents: A process-tracing study. *Journal of the Association for Information Systems, 9*(12), 727–747. https://doi.org/10.17705/1jais.00180

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment, 14*(10), 551–560. https://doi.org/10.1002/fee.1436

Kożuch, B., & Sienkiewicz-Małyjurek, K. (2022). Building collaborative trust in public safety networks. *Safety Science, 152*, 105785. https://doi.org/10.1016/j.ssci.2022.105785

Kroeger, F. (2019). Unlocking the treasure trove: How can Luhmann's theory of trust enrich trust research? *Journal of Trust Research, 9*(1), 110–124. https://doi.org/10.1080/21515581.2018.1552592

Kubiszewski, I., Noordewier, T., & Costanza, R. (2011). Perceived credibility of internet encyclopedias. *Computers & Education, 56*(3), 659–667. https://doi.org/10.1016/j.compedu.2010.10.008

Kuipers, B. (2018). How can we trust a robot? *Communications of the ACM, 61*(3), 86–95. https://doi.org/10.1145/3173087

Langlotz, C. P. (2019). Will artificial intelligence replace radiologists? *Radiology. Artificial Intelligence, 1*(3), e190058. https://doi.org/10.1148/ryai.2019190058

Lansing, J., & Sunyaev, A. (2016). Trust in cloud computing: Conceptual typology and trust-building antecedents. *ACM Sigmis Database: The Database for Advances in Information Systems, 47*(2), 58–96. https://doi.org/10.1145/2963175.2963179

Larsen, K. R., Lukyanenko, R., Muller, R., Storey, V. C., Vander Meer, D., Parsons, J., & Hovorka, D. S. (2020). Validity in design science research. In *International Conference on Design Science Research in Information Systems and Technology* (pp. 1–15). Springer Berlin/Heidelberg.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science (New York, N.Y.), 323*(5915), 721. https://doi.org/10.1126/science.1167742

Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International Journal of Environmental Research and Public Health, 18*(1), 271. https://doi.org/10.3390/ijerph18010271

Lee, J. D., & Kolodge, K. (2020). Exploring trust in self-driving vehicles through text analysis. *Human Factors, 62*(2), 260–277. https://doi.org/10.1177/0018720819872672

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

Leidner, D. E., & Tona, O. (2021). The CARE theory of dignity amid personal data digitalization. *MIS Quarterly, 45*(1), 343–370. https://doi.org/10.25300/MISQ/2021/15941

Li, J., Zhou, Y., Yao, J., & Liu, X. (2021). An empirical investigation of trust in AI in a Chinese petrochemical enterprise based on institutional theory. *Scientific Reports, 11*(1), 1–12. https://doi.org/10.1038/s41598-021-92904-7

Li, K., & Wieringa, P. A. (2000). Understanding perceived complexity in human supervisory control. *Cognition, Technology & Work, 2*(2), 75–88. https://doi.org/10.1007/s101110050029

Lohr, S. (2021). *What ever happened to IBM's Watson?* The New York Times. Retrieved September 10, 2022, from https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html

Lorenz, E. (1972). Predictability. *139th AAAS Meeting*, 1–6.

Lötsch, J., Kringel, D., & Ultsch, A. (2021). Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics, 2*(1), 1–17. https://doi.org/10.3390/biomedinformatics2010001

Luhmann, N. (1995). *Social systems*. Stanford University Press. Retrieved September 10, 2022, from https://books.google.ca/books?id=zVZQW4gxXk4C

Luhmann, N. (2000). Familiarity, confidence, trust: Problems and alternatives. *Trust: Making and Breaking Cooperative Relations, 6*(1), 94–107.

Luhmann, N. (2018). *Trust and power*. John Wiley & Sons.

Luhmann, N., & Gilgen, P. (2012). *Introduction to systems theory*. Wiley. Retrieved September 10, 2022, from https://books.google.ca/books?id=3mnUSAAACAAJ

Lukyanenko, R., Castellanos, A., Parsons, J., Chiarini Tremblay, M., & Storey, V. C. (2019a). Using conceptual modeling to support machine learning. In C. Cappiello & M. Ruiz (Eds.), *Information systems engineering in responsible information systems* (pp. 170–181). Springer International Publishing. https://doi.org/10.1007/978-3-030-21297-1_15

Lukyanenko, R., Castellanos, A., Samuel, B. M., Tremblay, M., & Maass, W. (2021a). Research agenda for basic explainable AI. *Americas Conference on Information Systems* (pp. 1–8).

Lukyanenko, R., Castellanos, A., Storey, V. C., Castillo, A., Tremblay, M. C., & Parsons, J. (2020). Superimposition: Augmenting machine learning outputs with conceptual models for explainable AI. In *1st international workshop on conceptual modeling meets artificial intelligence and data-driven decision making* (pp. 1–12). Springer.

Lukyanenko, R., & Parsons, J. (2018). Beyond Micro-tasks: Research opportunities in observational crowdsourcing. *Journal of Database Management (JDM), 29*(1), 1–22. https://doi.org/10.4018/JDM.2018010101

Lukyanenko, R., Parsons, J., & Samuel, B. M. (2019b). Representing instances: The case for reengineering conceptual modeling grammars. *European Journal of Information Systems, 28*(1), 68–90. https://doi.org/10.1080/0960085X.2018.1488567

Lukyanenko, R., Storey, V. C., & Pastor, O. (2021b). Foundations of information technology based on Bunge's systemist philosophy of reality. *Software and Systems Modeling, 20*(1), 921–938. https://doi.org/10.1007/s10270-021-00862-5

Lukyanenko, R., Storey, V. C., & Pastor, O. (2022). System: A Core conceptual modeling construct for capturing complexity. *Data & Knowledge Engineering, 141*, 1–29. https://doi.org/10.1016/j.datak.2022.102062

Lukyanenko, R., & Weber, R. (2022). A realist ontology of digital objects and digitalized systems. In *"Digital first" era — A joint AIS SIGSAND/SIGPrag workshop* (pp. 1–5). Virtual Workshop.

Lumineau, F., & Schilke, O. (2018). Trust development across levels of analysis: An embedded-agency perspective. *Journal of Trust Research, 8*(2), 238–248. https://doi.org/10.1080/21515581.2018.1531766

Maass, W., Castellanos, A., Tremblay, M. C., Lukyanenko, R., & Storey, V. C. (2022a). Concept Superimposition: Using conceptual modeling method for explainable AI. In *AAAI Spring Symposium: MAKE 2022* (pp. 1–6). Palm Springs.

Maass, W., Castellanos, A., Tremblay, M., & Lukyanenko, R. (2022b). AI Explainability: A conceptual model embedding. In *International Conference on Information Systems* (pp. 1–8).

Maass, W., & Storey, V. C. (2021). Pairing conceptual modeling with machine learning. *Data & Knowledge Engineering*, 101–123. https://doi.org/10.1016/j.datak.2021.101909

Maass, W., Storey, V. C., & Lukyanenko, R. (2021). From mental models to machine learning models via conceptual models. In *Exploring Modeling Methods for Systems Analysis and Development (EMMSAD 2021)* (pp. 1–8). Melbourne, Australia.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.

Marr, B. (2018). *Is artificial intelligence dangerous? 6 AI risks everyone should know about*. Forbes. Retrieved May 13,

2022, from https://www.forbes.com/sites/bernardmarr/2018/11/19/is-artificial-intelligence-dangerous-6-ai-risks-everyone-should-know-about/

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

McAfee, A., & Brynjolfsson, E. (2017). *Machine, platform, crowd: Harnessing our digital future*. WW Norton & Company.

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

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS), 2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review, 23*(3), 473–490. https://doi.org/10.5465/amr.1998.926622

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35. https://doi.org/10.1145/3457607

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *ArXiv Preprint*.https://doi.org/10.48550/arXiv.1712.00547

Minsky, M. (1974). *A framework for representing knowledge*. M.I.T. A.I.Laboratory.

Moody, D. L. (2009). The "physics" of notations: Toward a scientific basis for constructing visual notations in software engineering. *Software Engineering, IEEE Transactions On, 35*(6), 756–779. https://doi.org/10.1109/TSE.2009.67

Mooradian, T., Renzl, B., & Matzler, K. (2006). Who trusts? Personality, trust and knowledge sharing. *Management Learning, 37*(4), 523–540. https://doi.org/10.1177/1350507606073424

Morgner, C. (2018). Trust and society: Suggestions for further development of Niklas Luhmann's theory of trust. *Canadian Review of Sociology/Revue Canadienne de Sociologie, 55*(2), 232–256. https://doi.org/10.1111/cars.12191

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.1902.01876

Murphy, G. (2004). *The big book of concepts*. MIT Press.

Mylopoulos, J. (1998). Information modeling in the time of the revolution. *Information Systems, 23*(3–4), 127–155. https://doi.org/10.1016/S0306-4379(98)00005-2

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Sage.

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology, 3336*. https://doi.org/10.3389/fpsyg.2020.568256

Paré, G., Marsan, J., Jaana, M., Tamim, H., & Lukyanenko, R. (2020). IT vendors' legitimation strategies and market share: The case of EMR systems. *Information & Management, 57*(5), 103291.https://doi.org/10.1016/j.im.2020.103291

Park, J., & Kim, J. (2022). A data-driven exploration of the race between human labor and machines in the 21st century. *Communications of the ACM, 65*(5), 79–87. https://doi.org/10.1145/3488376

Parsons, J., & Wand, Y. (2008a). A question of class. *Nature, 455*(7216), 1040–1041. https://doi.org/10.1038/4551040a

Parsons, J., & Wand, Y. (2008b). Using cognitive principles to guide classification in information systems modeling. *MIS Quarterly, 32*(4), 839–868. https://doi.org/10.2307/25148874

Petersen, B. K., Chowhan, J., Cooke, G. B., Gosine, R., & Warrian, P. J. (2022). Automation and the future of work: An intersectional study of the role of human capital, income, gender and visible minority status. *Economic and Industrial Democracy, 0143831X221088301*. https://doi.org/10.1177/0143831X221088301

Polya, G. (2004). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Pratt, L. Y. (1992). Discriminability-based transfer between neural networks. *Advances in Neural Information Processing Systems*, 5.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Ramge, T. (2019). *Who's afraid of AI?: Fear and promise in the age of thinking machines*. The Experiment.

Rao, A. S., & Verweij, G. (2017). *Sizing the prize: What's the real value of AI for your business and how can you capitalise* (pp. 1–30). PwC Publication.

Recker, J., Lukyanenko, R., Sabegh, M. A., Samuel, B. M., & Castellanos, A. (2021). From representation to mediation: A new agenda for conceptual modeling research in a digital world. *MIS Quarterly, 45*(1), 269–300. https://doi.org/10.25300/MISQ/2021/16027

Riedl, R. (2022) Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electronic Markets, 32*(4). https://doi.org/10.1007/s12525-022-00594-4

Reimer, U., Bork, D., Fettke, P., & Tropmann-Frick, M. (2020). *Preface of the first Workshop "Models in AI"* (pp. 128–129). http://ceur-ws.org/Vol-2542/MOD-KI-preface.pdf

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

Renner, M., Lins, S., Söllner, M., Thiebes, S., & Sunyaev, A. (2022). Understanding the necessary conditions of multi-source trust transfer in artificial intelligence. In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS2022)* (pp. 1–10). http://hdl.handle.net/10125/80057

Rescher, N. (2013). *Value matters: Studies in axiology* (Vol. 8). Walter de Gruyter.

Reuters, T. (2021). *Toyota halts use of self-driving vehicle at Paralympic village after collision with visually impaired athlete*. CBC Sports. Retrieved April 5, 2022, from https://www.cbc.ca/sports/paralympics/toyota-halts-self-driving-vehicles-use-after-olympic-village-accident-1.6157569

Robert Jr., L. P., Bansal, G., Melville, N., & Stafford, T. (2020). Introduction to the special issue on AI fairness, trust, and ethics. *AIS Transactions on Human-Computer Interaction, 12*(4), 172–178. https://doi.org/10.17705/1thci.00134

Rosch, E. (1977). *Classification of real-world objects: Origins and representations in cognition* (pp. 212–222). In P. N. Johnson-Laird & P. C. Wason, (Eds.). Cambridge University Press.

Rosemann, M., & Green, P. (2002). Developing a meta model for the Bunge–Wand–Weber ontological constructs. *Information Systems, 27*(2), 75–91. https://doi.org/10.1016/S0306-4379(01)00048-5

Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs, 72*(1), 127–134. https://jia.sipa.columbia.edu/building-trust-artificial-intelligence

Rotenberg, K. J. (2019). *The psychology of interpersonal trust: Theory and research*. Routledge.

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

Sabel, C. F. (1993). Studied trust: Building new forms of cooperation in a volatile economy. *Human Relations, 46*(9), 1133–1170. https://doi.org/10.1177/001872679304600907

Sadiku, M. N., Fagbohungbe, O. I., & Musa, S. M. (2020). Artificial intelligence in cyber security. *International Journal of Engineering Research and Advanced Technology, 6*(05), 01–07. https://doi.org/10.31695/IJERAT.2020.3670

Saif, I., & Ammanath, B. (2020). Trustworthy AI is a framework to help manage unique risk. *MIT Technology Review*, 1–5. https://www.technologyreview.com/2020/03/25/950291/trustworthy-ai-is-a-framework-to-help-manage-unique-risk/

Salk, C. F., Sturn, T., See, L., Fritz, S., & Perger, C. (2015). Assessing quality of volunteer crowdsourcing contributions: Lessons from the cropland capture game. *International Journal of Digital Earth, 1*, 1–17. https://doi.org/10.1080/17538947.2015.1039609

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 39,* 1–15. https://doi.org/10.1145/3411764.3445518

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1147/rd.116.0601

Samuel, B. M., Khatri, V., & Ramesh, V. (2018). Exploring the effects of extensional versus intentional representations on domain understanding. *MIS Quarterly, 42*(4), 1187–1209. https://doi.org/10.25300/MISQ/2018/13255

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y., & Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. In P*roceedings of the Human Factors and Ergonomics Society Annual Meeting* (vol. 55, no. 1, pp. 1432–1436). SAGE Publications. https://doi.org/10.1177/1071181311551298

Scanlon, J. M., Kusano, K. D., Daniel, T., Alderson, C., Ogle, A., & Victor, T. (2021). Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. *Accident Analysis & Prevention, 163*(1), 106–154. https://doi.org/10.1016/j.aap.2021.106454

Scheman, N. (2015). Epistemology resuscitated: Objectivity as trustworthiness. In *Shifting Ground: Knowledge and Reality, Transgression and Trustworthiness*. Oxford University Press.

Schlindwein, S. L., & Ison, R. (2004). Human knowing and perceived complexity: Implications for systems practice. *Emergence: Complexity and Organization, 6*(3), 27–32.

Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology, 78*, 102253. https://doi.org/10.1016/j.joep.2020.102253

Schul, Y., Mayo, R., & Burnstein, E. (2008). The value of distrust. *Journal of Experimental Social Psychology, 44*(5), 1293–1302. https://doi.org/10.1016/j.jesp.2008.05.003

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Searle, J. R. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Selbst, A., & Powles, J. (2018). *"Meaningful information" and the right to explanation* (pp. 48–48). PMLR.

Shanks, G., Tansley, E., Nuredini, J., Tobin, D., & Weber, R. (2008). Representing part-whole relations in conceptual modeling: An empirical evaluation. *MIS Quarterly, 32*(3), 553–573. https://doi.org/10.2307/25148856

Shartsis, A. (2019). *Council post: Dynamic pricing: The secret weapon used by the world's most successful companies*. Forbes. Retrieved April 8, 2022 from https://www.forbes.com/sites/forbestechcouncil/2019/01/08/dynamic-pricing-the-secret-weapon-used-by-the-worlds-most-successful-companies/

Shaturaev, J. (2022). Economies and management as a result of the fourth industrial revolution: An education perspective. *Indonesian Journal of Educational Research and Technology, 3*(1), 51–58.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*(2), 47–53.

Siegrist, M., & Zingg, A. (2014). The role of public trust during pandemics. *European Psychologist, 19*(1), 23–32. https://doi.org/10.1027/1016-9040/a000169

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science, 16*(5), 264–268. https://doi.org/10.1111/j.1467-8721.2007.00517.x

Smuha, N. A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International, 20*(4), 97–106. https://doi.org/10.9785/cri-2019-200402

Söllner, M., Hoffmann, A., Hoffmann, H., & Leimeister, J. M. (2012). How to use behavioral research insights on trust for HCI system design. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (pp. 1703–1708). https://doi.org/10.1145/2212776.2223696

Söllner, M., & Leimeister, J. M. (2013). What we really know about antecedents of trust: A critical review of the empirical information systems literature on trust. In *Psychology of Trust: New Research, D. Gefen*. Verlag/Publisher: Nova Science Publishers. Available at SSRN: https://ssrn.com/absract=2475385

Stackpole, B. (2019). AI ain't for everyone—Who trusts bots, and why. *MIT Sloan*, 1–2. https://mitsloan.mit.edu/ideas-made-to-matter/ai-aint-everyone-who-trusts-bots-and-why

Starke, G., van den Brule, R., Elger, B. S., & Haselager, P. (2022). Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics, 36*(2), 154–161. https://doi.org/10.1111/bioe.12891

Steinke, F., Fritsch, T., & Silbermann, L. (2012). Trust in ambient assisted living (AAL)-a systematic review of trust in automation and assistance systems. *International Journal on Advances in Life Sciences, 4*(3–4).

Storey, V. C., & Goldstein, R. C. (1993). Knowledge-based approaches to database design. *MIS Quarterly, 17*(1), 25–46. https://doi.org/10.2307/249508

Storey, V. C., Lukyanenko, R., Parsons, J., & Maass, W. (2022). Explainable AI: Opening the black box or Pandora's box? *Communications of the ACM*, *65*(4), 27–29. https://doi.org/10.1145/3490699

Sturt, H. (1903). Happiness. *The international. The Journal of Ethics, 13*(2), 207–221.

Taddeo, M. (2021). On the risks of trusting artificial intelligence: The case of cybersecurity. In: J. Cowls, & J. Morley (Eds.), *The 2020 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook* (pp. 97–108). Springer. https://doi.org/10.1007/978-3-030-80083-3_10

Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature*

Machine Intelligence, 1(12), 557–560. https://doi.org/10.1038/s42256-019-0109-1

Teorey, T. J., Yang, D., & Fry, J. P. (1986). A logical design methodology for relational databases using the extended entity-relationship model. *ACM Computing Surveys, 18*(2), 197–222. https://doi.org/10.1145/7474.7475

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Thompson, N. (2018). When tech knows you better than you know yourself. *WIRED*, 1–4.

Tversky, A., Kahneman, D., & Moser, P. (1990). Judgment under uncertainty: Heuristics and biases. *Rationality in Action: Contemporary Approaches*, 171–188.

Vardi, M. Y. (2022). ACM, ethics, and corporate behavior. *Communications of the ACM, 65*(3), 5–5. https://doi.org/10.1145/3516423

Verissimo, P., Correia, M., Neves, N. F., & Sousa, P. (2009). Intrusion-resilient middleware design and validation. In *Information Assurance, Security and Privacy Services* (Vol. 4, pp. 615–678).

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Communications Medicine, 1*(1), 1–3. https://doi.org/10.1038/s43856-021-00028-w

von Bertalanffy, L. (1968). General system theory: Foundations, development, applications. Braziller.

von Neumann, J. (1958). *The computer and the brain*. New Haven, CT: Yale University Press.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76–99.

Wakabayashi, D. (2018). *Self-driving Uber car kills pedestrian in Arizona, where robots roam*. The New York Times. Retrieved September 10, 2022, from https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

Waldrop, M. M. (2015). No drivers required. *Nature, 518*(7537), 20. https://doi.org/10.1038/518020a

Wan, Y., Gao, Y., & Hu, Y. (2022). Blockchain application and collaborative innovation in the manufacturing industry: Based on the perspective of social trust. *Technological Forecasting and Social Change, 177*, 121540. https://doi.org/10.1016/j.techfore.2022.121540

Wand, Y., & Weber, R. (2002). Research commentary: Information systems and conceptual modeling—A research agenda. *Information Systems Research, 13*(4), 363–376. https://doi.org/10.1287/isre.13.4.363.69

Wang, W., & Siau, K. (2018). Trusting artificial intelligence in healthcare. *Americas Conference on Information Systems* (pp. 1–1).

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets, 32*(4). https://doi.org/10.1007/s12525-022-00593-5

Weber, J. M., Malhotra, D., & Murnighan, J. K. (2004). Normal acts of irrational trust: Motivated attributions and the trust development process. *Research in Organizational Behavior, 26*, 75–101. https://doi.org/10.1016/S0191-3085(04)26003-8

Whyte, K. P., & Crease, R. P. (2010). Trust, expertise, and the philosophy of science. *Synthese, 177*(3), 411–425. https://doi.org/10.1007/s11229-010-9786-3

Woo, C. (2011). The role of conceptual modeling in managing and changing the business. *International Conference on Conceptual Modeling* (pp. 1–12). Springer.

Yamagishi, T. (2011). *Trust: The evolutionary game of mind and society*. Springer.

Yampolskiy, R. V. (2015). *Artificial superintelligence: A futuristic approach*. cRc Press.

Yang, R. & Wibowo, S. (2022) User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets, 32*(4). https://doi.org/10.1007/s12525-022-00592-6

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. *Global Catastrophic Risks, 1*(303), 184.

Yuki, M., Maddux, W. W., Brewer, M. B., & Takemura, K. (2005). Cross-cultural differences in relationship-and group-based trust. *Personality and Social Psychology Bulletin, 31*(1), 48–62. https://doi.org/10.1177/0146167204271305

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.