







REVIEW ARTICLE



<https://doi.org/10.1057/s41599-025-05481-9>

OPEN

A call for transdisciplinary trust research in the artificial intelligence era

Frank Krueger^{1,15}, René Riedl^{2,15}, Jennifer A. Bartz³, Karen S. Cook⁴, David Gefen⁵, Peter A. Hancock⁶, Sirkka L. Jarvenpaa⁷, Lydia Krabbendam⁸, Mary R. Lee⁹, Roger C. Mayer¹⁰, Alexandra Mislin¹¹, Gernot R. Müller-Putz¹², Thomas Simpson¹³, Haruto Takagishi¹⁴ & Paul A. M. Van Lange⁸

Trust is a cornerstone and enabler of human civilization, determining the very nature of how people interact with each other. The swift integration of artificial intelligence (AI) into daily life poses grand societal challenges and necessitates a reevaluation of trust. Our bibliometric literature review calls for scientists and stakeholders to cross traditional academic boundaries to address emerging and evolving societal challenges arising from AI. We propose a transdisciplinary research framework to understand and bolster trust in AI and address grand challenges in domains as diverse and urgent as misinformation, discrimination, and warfare.

Introduction

We are at the forefront of a new era—the Artificial Intelligence (AI) Revolution (Kaplan et al. 2023). AI is broadly defined as the ability of a computer program, machine, or system to make human-like intelligent decisions and perform tasks autonomously (Russell and Norvig, 2021). The AI Revolution is distinguished by its ability to simulate aspects of human cognition—enabling machines to assess scenarios, learn from experience, and adaptively apply knowledge to decision-making and problem-solving. This marks a shift from earlier technologies that automated physical tasks to systems capable of tackling cognitive challenges. Accordingly, AI bridges the “knowledge and intelligence gap”—the divide between mechanical automation and human reasoning. While AI does not fully replicate human intelligence, it emulates key functions such as pattern recognition, adaptive learning, and contextual reasoning, enhancing the accessibility and capability of intelligent systems across diverse applications (Li et al. 2024). It presents unprecedented global opportunities across many areas within Industry 4.0 (Magd et al. 2022), including autonomous transportation (Qayyum et al. 2020), healthcare (Wiens and Shenoy, 2018), and military (Sligar, 2020), with resulting efficiency and productivity gains predicted to boost the global economy by an estimated \$13 trillion by 2030 (Bughin et al. 2018).

¹ George Mason University, School of Systems Biology, Fairfax, VA, USA. ² University of Applied Sciences Upper Austria & Johannes Kepler University Linz, Linz, Austria. ³ McGill University, Montreal, Canada. ⁴ Stanford University, Stanford, CA, USA. ⁵ Drexel University, Philadelphia, PA, USA. ⁶ University of Central Florida, Orlando, FL, USA. ⁷ University of Texas at Austin, Austin, TX, USA. ⁸ Vrije Universiteit Amsterdam, Amsterdam, Netherlands. ⁹ Veterans Administration Medical Center, Washington, DC, USA. ¹⁰ NC State University, Raleigh, NC, USA. ¹¹ American University, Washington, DC, USA. ¹² Graz University of Technology, Graz, Austria. ¹³ University of Oxford, Oxford, UK. ¹⁴ Tamagawa University, Machida, Japan. ¹⁵ These authors contributed equally: Frank Krueger, René Riedl. ✉email: FKrueger@gmu.edu

As AI becomes increasingly integral to our lives, the realization that traditional notions of interpersonal trust applied to humans do not necessarily extend to AI poses significant risks to society (Sabherwal and Grover, 2024). Specifically, integrating AI into society raises ethical concerns and presents several grand challenges within society, including the risks of manipulation, misinformation, discrimination, displacement, misuse in warfare, and the potential loss of control over AI systems (Hancock, 2023). However, trust in AI technology fosters its adoption, thereby enhancing public acceptance (Li et al. 2024). Conversely, a lack of trust in AI and its subsequent impact on societal trust can lead to diminished efficiency, financial losses, stifled innovation, worsened social inequalities, and potential social unrest as AI nonetheless becomes central to our lives (Capraro et al. 2024). This lack of trust jeopardizes the beneficial applications of AI and ultimately undermines social cohesion (Putnam, 2000; Kramer, 1999; Hancock et al. 2023a).

We propose that these grand challenges cannot be addressed without collaborative efforts across academic disciplines and societal stakeholders within a transdisciplinary framework (Montuori, 2013). Indeed, our comprehensive bibliometric review of over 34,000 trust research articles from the past three decades indicates that although multi- and interdisciplinary studies are present, transdisciplinary efforts are scarce. Our review finds that collaboration between scientists and stakeholders is missing, a major characteristic of transdisciplinarity. Lacking the institutional stakeholders' perspective hinders our understanding of AI trust issues, as existing research may not reach end users to build trust and might not offer solutions due to insufficiently integrated research. Therefore, our objective is to establish a transdisciplinary research agenda on trust, calling for enhanced synergy across academics and other stakeholders to amplify the quality and impact of trust research in the era of AI.

Trust dilemma: navigating grand challenges in the era of AI

Interpersonal trust is vital for human flourishing and economic growth (Zak and Knack, 2001), reducing collaboration costs (Dirks et al. 2011), generating wealth through specialization and exchange (Kim, 2023; Knack and Zeefer, 1997; Cook et al. 2005), and promoting welfare (McEvily, 2011; Bottom et al. 2006; Zahedi and Song, 2008). Without trust, the social fabric unravels, communication falters, and disorder ensues, making trust a decisive basis for human and societal progress (Redfern, 2009; Buchan et al. 2008). A widely recognized definition across multiple disciplines (Hardin, 2002; Baier, 1986; Simpson, 2012; Schoorman et al. 2007; Rousseau et al. 1998; Luhmann, 2017; Barber, 1983; Simpson, 2023) defines trust as one party's willingness to be vulnerable to another, based on the belief that the other will perform a crucial action, even without monitoring (Mayer et al. 1995). As such, trust poses a dilemma (Lange et al. 2017), emphasized by its potential risks and benefits (Mislin et al. 2011): every human relationship—from dyadic to societal—entails inherent risks of exploitation, requiring trust evaluation and necessitating vulnerability to these risks (Bartz and Lydon, 2006; Holmes, 1991, 1981; Deutsch, 1958). People overcome trust constraints with strangers by developing initial trust (Lange et al. 2017), influenced by socialization (Schilke et al. 2021; Gächter et al. 2010), genetic factors (Shou et al. 2021; Riedl and Javor, 2012), hormones (Bartz et al. 2011), brain functionality (Krueger and Meyer-Lindenberg, 2019; Bellucci et al. 2017; Fehr, 2009), and neural development (Sijtsma et al. 2023; Krueger, 2021). Early personal experiences shape initial trust (McKnight et al. 1998; Simpson, 2007), which evolves through interactions over time, reflecting the dynamics of trust (Riedl et al. 2014; King-Casas et al. 2005). When thinking about trust, we often think of

trust between individuals, but we also experience trust with non-human entities.

Adopting new technologies requires trust and triggers paradigm shifts that reshape the nature of trust, simultaneously addressing existing grand societal challenges and creating new ones. For example, transformative technologies like Gutenberg's printing press, the steam engine, and the Internet—central to the Printing, Industrial, and Digital Revolutions—reshaped societal trust by bridging gaps in knowledge, power, and distance, challenging authorities, enhancing productivity, and democratizing information access (Werbach, 2018). Each era proved more disruptive than its predecessor, significantly shifting how information was disseminated, work was conducted, and societies organized themselves, invariably impacting the fabric of trust. As these previous revolutions resulted in major societal upheaval, the emergence of AI technology is unique in this regard since it challenges traditional concepts of interpersonal trust (Russell and Norvig, 2021). AI is often viewed from a social cognition perspective, making it difficult to see it merely as a machine and instead as an entity potentially deserving of trust (Williams et al. 2022). As AI advances, distinguishing between human and technological interactions will become increasingly challenging, and it is unclear whether trust evaluations target AI itself, the company that developed it, or both (Wingert and Mayer, 2024).

Building trust between human users (trustors) and AI systems (trustees) across various contexts is inherently complex and qualitatively different from trust between human agents (Kaplan et al. 2023; Hancock et al. 2023a). While interpersonal trust—typically defined as a willingness to be vulnerable based on positive expectations of another's intentions and ability—provides a useful conceptual starting point, it must be adapted when applied to non-human entities. AI systems lack intentionality, emotional states, and moral agency, which are foundational elements in assessments of human trustworthiness. Nevertheless, many trust frameworks continue to draw on familiar dimensions—such as ability, benevolence, and integrity—even in the context of AI (Hancock et al. 2023a; Lyons et al. 2023; Yusuf and Baber, 2020). These dimensions, however, require reinterpretation (Thiebes et al. 2021; Asan et al. 2020): ability refers to the system's technical performance (e.g., safety, reliability, accuracy, robustness) as demonstrated by empirical evidence and performance data. Benevolence (e.g., privacy, fairness) and integrity (e.g., explainability, accountability) are not intrinsic properties of the AI but are realized through system design, ethical programming, and regulatory safeguards (Schlicker and Langer, 2021). This reframing underscores that trust in AI is not merely a replication of interpersonal trust but a distinct socio-technical construct shaped by human interaction with technological features and institutional mechanisms. As such, understanding people's trust in AI requires new conceptual tools that extend beyond social interaction in non-technological contexts.

The rapid advancement and increasing complexity of AI technologies represent a double-edged sword. They present not only societal opportunities but also grand challenges, as illustrated in the following examples:

Profiling. Machine learning employs supervised, unsupervised, and reinforcement learning algorithms to analyze vast datasets and identify complex patterns, thereby projecting future outcomes based on historical data in domains such as retail (Heins, 2023), marketing (Chintalapati and Pandey, 2022), and precision medicine (Mumtaz et al. 2023). However, these predictive algorithms carry serious risks, such as predictive profiling for consumers in online advertising platforms—including unwarranted data collection and invasive advertising¹—which can negatively affect mental health and alter implicit self-perception. Such practices, particularly when conducted without transparent

consent and oversight, undermine the trustworthiness principle of privacy and erode people’s trust in AI, affecting how individuals perceive reality and their vulnerability.

Misinformation. Computer vision AI technology, augmented by generative adversarial networks, has revolutionized image and video manipulation, delivering substantial benefits across fields, including security monitoring (Zhang et al. 2022), film and entertainment (Du and Han, 2021), and healthcare diagnostics (Kumar et al. 2022). However, this AI technology also presents risks, such as creating deepfakes—highly realistic, fabricated images or videos designed to spread misinformation across online media users, which can alter social media perceptions and damage reputations. This misuse undermines the principle of non-maleficence and erodes trust in AI, fundamentally altering how people trust within an interaction with an unknown entity (Laas, 2023; Vaccari and Chadwick, 2020).

Discrimination. Natural language processing and its associated large language models (LLMs, such as ChatGPT, Gemini, and Grok) drive revolutionary AI applications for sentiment analysis, personal assistants, and automated content generation across various areas, including customer service (Mariani and Borghi, 2023), finance (Ahmed et al. 2022), and e-commerce (Bawack et al. 2022). This AI technology, however, carries risks such as biases (Mehrabani et al. 2021), as LLMs trained on massive datasets reflecting real-world prejudices can perpetuate and amplify discrimination in different domains, such as in human recruitment (e.g., gender, age, and racial biases in job applications)² or judicial decision-making. These biased outputs compromise fairness and erode trust in AI applications, potentially diminishing trust from racial and ethnic minorities (Sullivan et al. 2022; Zhou et al. 2021).

Job displacement. AI-powered robotics enhances machine capabilities like vision, touch, and autonomous decision-making, significantly broadening their applications across diverse fields such as autonomous driving (Gao and Bian, 2021), manufacturing assembly lines (Narkhede et al. 2024), and surgical healthcare procedures (King et al. 2023). However, this technology entails risks, like its autonomy, which can lead to the displacement of both blue- and white-collar jobs, such as truck drivers, factory workers, retail staff, and computer programmers³. As machines assume more roles and displace jobs, concerns about accountability escalate (Blacklaws, 2018; Alam and Mueller, 2021), fueling a decline in trust in AI. This is particularly evident in the retail industry, hardest hit by rising unemployment and wealth inequality, further eroding trust in corporations.

Warfare. Deep learning, a subset of AI that utilizes neural networks with complex algorithms built on thousands of features and millions of parameters, enhances decision-making and problem-solving efficiency by providing real-time strategic guidance and improving tactical decisions in areas such as military operations (Pandey et al. 2024), defense systems (Qiu et al. 2019), and cybersecurity (Naik et al. 2022). This AI technology’s “black box” nature poses serious risks, including opaque military decision-making involving autonomous weaponry (Johnson, 2020), which can lead to unintended consequences such as

civilian casualties and loss of control over critical systems (von Eschenbach, 2021). The complexity of AI systems challenges the principle of explainable AI (XAI) (Shaban-Nejad et al. 2021), encompassing transparency, interpretability, and explainability, and potential misalignments with moral, ethical, and legal principles erode trust in AI systems, jeopardizing trust among nations.

Singularity. Quantum-enhanced AI holds the potential for groundbreaking advancements in drug design (Nandi et al. 2024), climate modeling (Shaamala et al. 2024), and space exploration (Omar et al. 2021) by leveraging quantum computing capabilities to accelerate the computing processes of AI systems exponentially (Pérez et al. 2023). Still, the rapid AI evolution towards Artificial General Intelligence or Artificial Super Intelligence raises risks, such as AI achieving supremacy (Hurlburt, 2017), where it could surpass human knowledge and intelligence, leading to dire consequences for governance (e.g., loss of essential skills, diffusion of responsibility, decline of human agency) (Gordon, 2015). These developments could undermine human-centricity by exacerbating the AI alignment problem, eroding trust in AI, and jeopardizing trust in AI evolution (Gabriel, 2020).

These paradigmatic, yet not exhaustive, examples of grand societal challenges (e.g., profiling, misinformation, discrimination) highlight for different users (e.g., consumers, social media users, job applicants) the interplay of various elements shaping trust in AI. From a scientific perspective, AI systems’ inherently associated risks (e.g., prediction, deep fake, bias) pose potential adverse outcomes in various technological terrains (e.g., advertising, social media, job recruitment). From a societal perspective, AI’s inherent risks threaten its perceived trustworthiness (e.g., privacy, non-maleficence, fairness) and impact trust interactions within various societal spheres (e.g., self, strangers, ethical minorities). For a summary of key elements—trustworthiness, risk, user, sphere, and terrain—related to illustrated grand challenges, see Table 1.

Given the interplay of these key elements, building trust in AI requires not only interdisciplinary collaboration across fields like engineering, computer science, sociology, psychology, neuroscience, ethics, philosophy, and law but also integrating diverse knowledge, methods, and perspectives to ensure the development of trustworthy AI (Thiebes et al. 2021) that balances technological advancement (science) and ethical considerations (society). Fostering a positive impact of AI on societal trust demands further collaboration beyond academia, involving stakeholders like developers, investors, suppliers, regulators, educators, policy-makers, users, and the general public. This view is supported by a growing body of academic and applied literature that emphasizes that the risks of AI stem not from distant futures but from its current use in critical institutions, where it often reinforces social inequalities. Recognizing the limits of dominant ethical frameworks, scholars call for systemic analyses that consider the political, historical, and cultural contexts in which AI operates—crucial for managing real-world impacts (Crawford and Calo, 2016). Broader perspectives reveal AI’s entanglement in global systems of labor, data extraction, and environmental harm,

Table 1 Illustrative grand challenges and their related trust key elements.

Grand Challenge	Trustworthiness	Risk	User	Sphere	Terrain
Profiling	Privacy	Machine Learning: Prediction	Consumer	Self	Advertising
Misinformation	Non-maleficence	Computer Vision: Deepfake	Social Media User	Stranger	Social Media
Discrimination	Fairness	Natural Language Processing: Bias	Job Applicant	Race Minority	Job Recruitment
Job Displacement	Accountability	AI-powered Robotics: Autonomy	Retail Staff	Cooperation	Retail
Warfare	Explainability	Deep Learning: Opacity	Military Personnel	Nation	Military
Singularity	Human-Centricity	Quantum-Enhanced AI: Supremacy	Humanity	AI Evolution	Governance

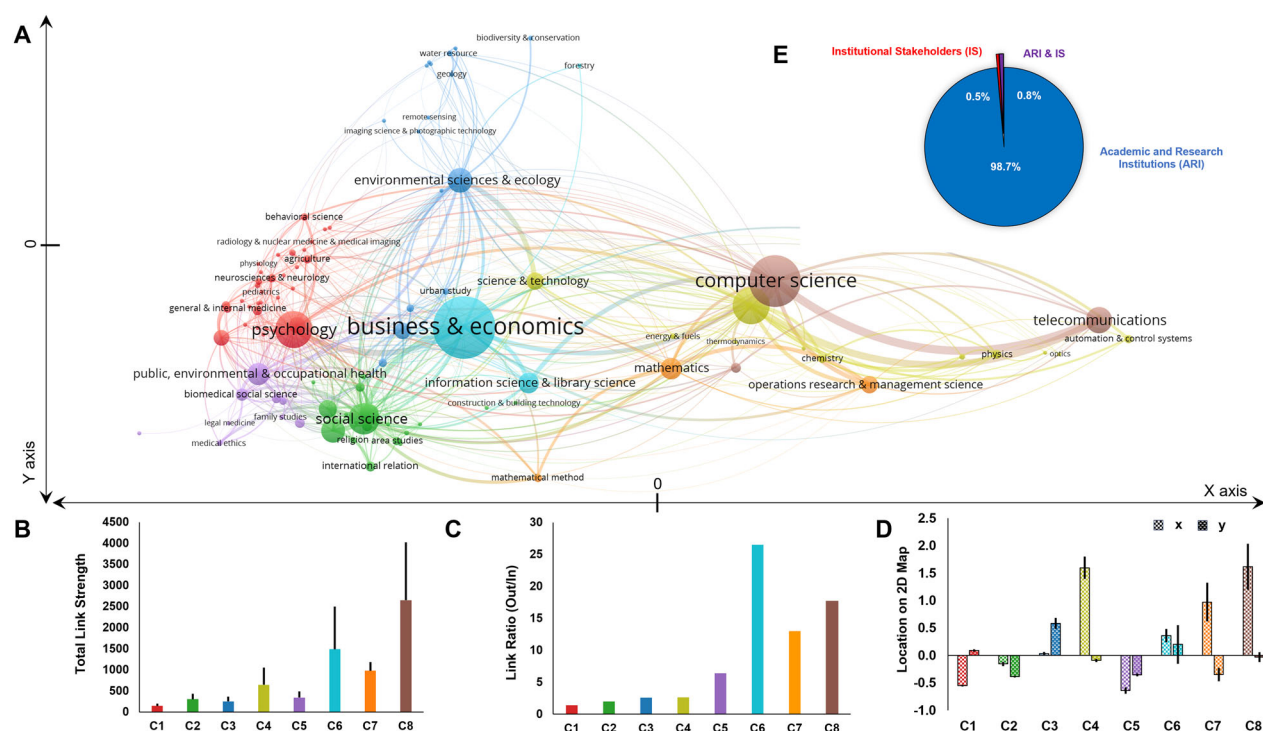


Fig. 1 Evaluation of multi-, inter-, and transdisciplinary trust research. **A** Network visualization map. A bibliometric distance-based network map was created based on 34,459 research articles using VOSviewer (Van Eck and Waltman, 2010), displaying 98 nodes (i.e., research areas, RAs as categorized by the Web of Science schema⁵), 8 clusters (C1-C8, i.e., research domains displayed in different colors), and 1314 edges (i.e., links measured in total link strengths of co-occurrence) between nodes and clusters in a two-dimensional (2D) space (via x and y coordinates [arbitrary units] indicating relative distance). Larger labels and circles indicate higher RA occurrence; thicker edges indicate greater link strength between RAs, and smaller distances between RAs indicate higher relatedness. **B** Bar graph with mean total link strength (\pm standard error of the mean, s.e.m.) in clusters. Multidisciplinarity is identifiable by distinct clusters, each representing separate research domains characterized by various RAs and their total link strength. Although clusters were internally cohesive, the total link strength differed significantly among them ($\chi^2_7 = 19.86$, $P < 0.005$). This suggests that some clusters, e.g., cluster 8 (including computer science, indicated in brown) and cluster 6 (including business & economics, indicated in turquoise), have stronger total link strength than others, potentially indicating interdisciplinarity characterized by inter-cluster links. **C** Bar graph with mean link ratio in clusters. Interdisciplinarity can only be partially identified because the ratio of the number of links between RAs staying within the same cluster and going out to other clusters differed significantly across clusters ($P < 0.001$). This indicates that some clusters, particularly cluster 6 (including business & economics), had more links to other clusters. **D** Bar graph with cluster relatedness in map coordinates (mean x and y coordinates, \pm s.e.m.) as positioned in the 2D map. Transdisciplinarity cannot be identified due to the lack of strong inter-cluster links and a complex network structure with significant overlap and integration. This results in blurred cluster boundaries and shorter distances between them. Clusters differed significantly in location, with x ($\chi^2_7 = 85.77$, $P < 0.0001$) and y ($\chi^2_7 = 75.68$, $P < 0.0001$) coordinates, indicating, for example, that clusters on the left side (i.e., C1, C2, C3, C5, C6) are more closely located and partially overlap, unlike those on the right side (i.e., C4, C7, C8). **E** Pie chart of co-authorship percentages based on organizational affiliation. The history of trust research is predominantly driven by scientific discourse (98.7% of publications by authors affiliated with Academic and Research Institutions, ARI), compared to collaborative science and societal discourse (0.8% with at least one author from Institutional Stakeholders, IS) and solely societal discourse (0.5% by IS).

highlighting the need for deeper scrutiny (Crawford, 2021). There is also a call for enforceable safeguards to protect rights, identity, and privacy⁴—key to building trust in the face of unregulated AI, including neurotechnologies (Yuste et al. 2017). Together, these insights underscore the need to embed political, societal, and economic dimensions into transdisciplinary trust research, anchoring it in real-world institutional contexts. Ultimately, effective collaboration among scientists and stakeholders is crucial for tackling AI deployment's theoretical, practical, and ethical considerations, ensuring technologies are technically sound and ethically aligned, ultimately fostering societal trust (Felzmann et al. 2019).

How do we understand and combat the emerging societal AI grand challenges to trust?

Addressing such a need requires a transdisciplinary trust research agenda that evaluates trust research through a combined bibliometric literature review and network analysis (see Supplementary

Materials) (Fig. 1). Our bibliometric network analysis indeed reveals a notable absence of research articles that align with the core characteristics of a transdisciplinary research agenda (or even use its terminology). This deficiency demonstrates that prior research has largely failed to integrate knowledge, methods, and perspectives from diverse disciplines within a unified, holistic framework. Despite the involvement of various scientific disciplines in research, almost 99% of studies failed to incorporate the perspectives of institutional stakeholders (e.g., developers, policymakers, and the general public), indicating that academics and other stakeholders have not been equal partners in the research and intervention development process. The absence of the institutional stakeholders' perspectives hampers our understanding of AI trust issues, as existing research may not address end users' concerns to build trust and may lack integrated solutions.

To surpass the limitations of multi- and interdisciplinary perspectives, omitting the perspectives of institutional

stakeholders, we propose a transdisciplinary framework to solve grand challenges and provide solutions to enhance trust in AI and its impact on societal trust (Fig. 2A). Our comprehensive framework builds on the model for transdisciplinary research processes (Jahn et al. 2012), based on the fundamental idea that grand societal challenges must be connected to existing scientific knowledge gaps to develop successful practical solutions. It views societal advancement and scientific progression as knowledge-driven systems that feed into a comprehensive knowledge integration system (Fig. 2B). This process, guided by ongoing discourses between stakeholders and scientists, unfolds in three phases. During the *problem transformation phase*, a grand societal challenge is identified within the societal system, linked to existing scientific knowledge as a grand scientific challenge within the scientific system, and redefined as a common research objective within the integrative system. The roles of scientists and stakeholders are delineated in the production of new, connectable knowledge phase. An integration concept is developed and implemented across five key elements of trust—trustworthiness, risk, user, sphere, and terrain—that are central to addressing social grand challenges related to trust in AI and its impact on societal trust (cf., Table 1). The *transdisciplinary integration phase* evaluates the integrated results and compiles outputs for societal and scientific communities. Across these phases, two distinct transdisciplinary pathways emerge: a real-world pathway focusing on practical societal solutions and an intra-scientific pathway aimed at empirical research and discovery. Our framework integrates diverse perspectives from scientific and societal domains to support trustworthy AI, providing a structured approach for unifying insights *across disciplines and stakeholder contexts*. For example, in the case of explainability, our framework connects technical transparency from computer science with insights from ethics and sociology about how explanations shape user perceptions, cognitive processing, and institutional trust. This enables a holistic treatment of explainability as both a technical and socially embedded feature.

Leveraging the framework provides a tool to determine future research directions and uncover new solutions for identifying, exploring, and creating targeted strategies, measures, and interventions to enhance trust in AI and its impact on societal trust. Placing the user at the center prioritizes human rights, justice, and dignity within human-AI interactions, ensuring all other elements align with this core principle. This integrated approach is crucial for designing, developing, and deploying AI technologies that maximize their impact within societal contexts and contribute to scientific progress.

To illustrate the practical utility of our framework, Table 2 presents a detailed application to the domain of autonomous vehicles. It uses a recent real-world case—when U.S. authorities revoked the operational permit of the “Cruise” self-driving taxi service in San Francisco due to safety incidents⁶—as a context for demonstrating how our framework can guide the diagnosis of trust failures and inform targeted, cross-sectoral interventions.

The proposed framework builds upon theoretical foundations from previous literature reviews and meta-analyses (Kaplan et al. 2023; Li et al. 2024; Hancock et al. 2023b; Afroogh et al. 2024), with the mission to solve grand societal challenges related to trust in AI and its impact on societal trust, all while keeping the end-user at its core. It addresses various elements of trust, enabling it to effectively respond to the evolving nature of AI technology and societal norms, thus maintaining its relevance over time. Importantly, AI technologies increasingly occupy a paradoxical position: they are both sources of new ethical risks and tools designed to mitigate them. This is especially evident in safety-critical domains such as autonomous driving and medical diagnostics, where AI ensures accountability and minimizes human

error while simultaneously introducing new uncertainties. Such contradictions highlight the limits of single-discipline solutions and reinforce the need for a transdisciplinary framework that can reconcile technical performance with ethical governance. Addressing this duality is essential to building resilient and trustworthy AI systems.

Further, adopting a transdisciplinary research approach emphasizes the evolving interconnectedness of its elements from science and society, fostering a holistic and relatable understanding of trust for stakeholders and scientists. It offers a richer, more nuanced understanding of trust in AI, guiding the development of innovations that align with human values and needs and enhancing public acceptance and adoption of AI technology. To reflect the diverse range of AI applications, *our framework is intentionally designed to be adaptable across different terrains*. Trust in AI is not uniform. Contexts such as healthcare, public administration, military, or consumer technology all involve distinct trust relationships and ethical concerns. By structuring our analysis around multiple key elements and various terrains, we allow the framework to capture the nuances of each use case, supporting context-sensitive evaluations of trust in AI systems.

Finally, it is a preventive framework, proactively identifying and addressing potential risks and ethical concerns to foster a trustworthy AI environment. It provides an integrative view of trust in AI, focusing on theoretical constructs and practical implications, while recent governmental initiatives establish a legal framework for safe and ethical AI deployment. For example, multiple countries and international organizations have issued guidelines to develop trustworthy AI: the European Union issued the *Ethics Guidelines for Trustworthy AI*⁷, China released the *Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence*⁸, the former US President Biden signed an executive order on the *safe, secure, and trustworthy development and use of artificial intelligence*⁹, and the Waag Institute in the Netherlands has advanced participatory, transdisciplinary approaches to AI governance through public engagement and co-creation initiatives¹⁰. Overall, both approaches, *theoretical constructs and practical implications*, are crucial for the responsible development and use of AI technologies, working together in complementary conceptual and regulatory realms.

Implementing such an evolving transdisciplinary research agenda offers significant benefits but also comes with implementation challenges (Vasbinder et al. 2010; de Oliveira et al. 2019). First, the often-prohibitive hurdles of disciplinary boundaries and entrenched biases must be overcome through open collaboration, raising awareness, and promoting the value of the proposed transdisciplinary research agenda. Second, transdisciplinary communication barriers necessitate shared frameworks and skill-building through transdisciplinary training programs. Third, the rigid structures and mechanisms in funding institutions often hinder transdisciplinary endeavors, mandating institutional commitments to innovative funding and reward systems. Fourth, the complexity of integrating methodologies and data across disciplines demands time, strategic resource allocation, and standardized protocols. Finally, public skepticism of non-traditional approaches can be mitigated through effective science communication and engagement strategies. These steps are not only theoretically grounded but also actionable measures that institutions, research teams, and policy bodies can adopt to foster meaningful transdisciplinary collaboration in AI trust research.

Overcoming these obstacles is crucial for unlocking the full potential of transdisciplinary trust research, fostering collective efforts to address revolutionary societal challenges that leave scientists and stakeholders no choice but to work together to understand and help combat the enormous threats to trust in AI that societies worldwide face. Perhaps more than ever, scientists and stakeholders need each

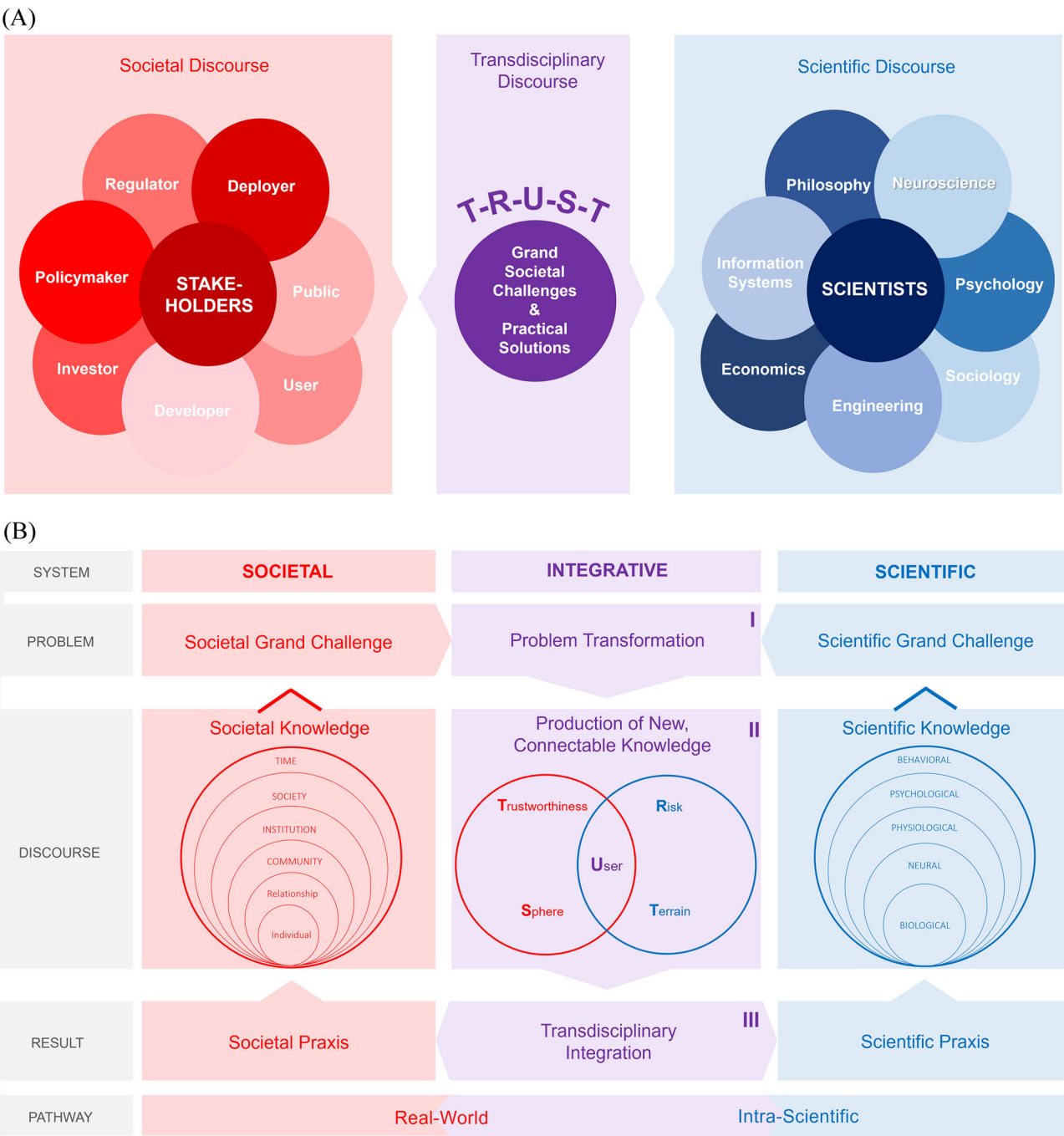


Fig. 2 Transdisciplinary trust research. **A** Transdisciplinary Research Agenda. Transdisciplinarity emphasizes collaboration between scientists and stakeholders, integrating knowledge to address grand challenges and producing practical solutions for society and science. The figure shows examples of major stakeholders and relevant scientific disciplines, though these are not exhaustive. **B** Transdisciplinary Research Framework. The transdisciplinary framework considers societal advancement and scientific progression as knowledge-focused systems providing input into a knowledge-integration system, each undergoing three stages: problem, discourse, and result. Guided by ongoing discourses between stakeholders and scientists, this process unfolds in three phases: problem formation, production of new, connectable knowledge, and transdisciplinary integration. Across these phases, two distinct transdisciplinary pathways unfold, encompassing a real-world pathway prioritizing practical societal solutions and an intra-scientific pathway aimed at empirical study and discovery. At the core of the framework, new, connectable knowledge is developed and implemented across five key elements of trust: trustworthiness, risk, user, sphere, and terrain. The user is the central focus of the framework, playing a key role in the discourses on both societal and scientific knowledge. Societal knowledge encompasses stakeholders' practices and criteria for evaluating AI's impact on societal trust, assessed across various ecological layers (e.g., individual, relationship, community). Scientific knowledge encompasses scientists' methods and theories for researching trust in AI, examined across various measurement levels (e.g., biological, neural, physiological). Trustworthiness and sphere are grounded in the societal knowledge system: Trustworthiness is essential for addressing societal challenges, as perceptions of AI's reliability significantly influence its acceptance and effectiveness. Sphere, integral to societal praxis, refers to various trust interactions within ecological layers that AI technologies impact. Risk and terrain are grounded in the scientific knowledge system: Risk is integral to the scientific challenge of AI development, encompassing unforeseen dangers and potential adverse outcomes that require thorough scientific assessment and exploration. Terrain, a critical aspect of scientific praxis, refers to various environments where AI technologies are applied.

Table 2 Example Application of the TrustNet Framework.	
Phase	Steps
I. Problem Transformation	<p>Inception of the grand challenge project</p> <p>Following numerous incidents with autonomous vehicles (AV) (e.g., in October 2023, an individual was severely injured and trapped under a Cruise robotaxi after being struck by another vehicle), authorities in California revoked the operating license of “Cruise” taxi service in San Francisco, citing undeniable risks to public safety (Magd et al. 2022). Despite advancements in self-driving car technology, public trust was dampened due to media-highlighted accidents and the fear of relinquishing control to machines. While data show AVs are statistically safer than vehicles driven by humans, there is discomfort about their decision-making in complex scenarios (Gao and Bian, 2021).</p>
	<p>I.I Crafting the grand challenge. The integration process starts with a problem constitution within the societal system, where stakeholders engage in discourse to determine a grand challenge. Leveraging structured interviews with diverse stakeholders (e.g., vehicle manufacturers, tech companies, regulatory bodies, and the general public) and using socio-empirical methodologies (e.g., focus groups with daily commuters, technologists, and legal experts), a grand challenge could be crafted: Autonomous driving uses machine learning, computer vision, and sensor fusion to enable vehicles to operate without human intervention, aiming to revolutionize transportation by enhancing safety and mobility. However, significant risks include traffic accidents from system failures and ethical dilemmas in opaque decision-making during collisions. These challenges compromise safety, reliability, explainability, and accountability, affecting drivers’ trust in AVs and eroding public trust in autonomous driving.</p> <p>I.II Connecting the grand challenge description to scientific knowledge. Scientists relate the grand challenge to existing knowledge gaps in the scientific system and define an accompanying grand challenge through discourse. Collaborating with scientists from diverse disciplines (e.g., automotive engineering, transportation science, computer science, ergonomics, information systems, psychology, philosophy, ethics, and law), the issues surrounding AVs can be examined in light of existing knowledge (Miller and Boyle, 2019). Those examinations reveal that, despite their potential to minimize human error-induced accidents, these vehicles present safety and reliability concerns in non-standardized conditions and mixed-traffic environments, coupled with algorithm explainability and accountability concerns. Acknowledging the complexity of the grand challenge, experts can agree that simply gathering more performance data would not adequately resolve the underlying trust concerns. A holistic strategy, integrating technological, psychological, ethical, legal, and socio-economic aspects, can guide the project to devise all-encompassing trust-building solutions for AVs, prioritizing a unified research agenda over standalone studies.</p> <p>I.III Transforming the grand challenge into a common research object. Within the integrative system, a transdisciplinary team transforms the grand challenge into a central research object by implementing collective societal and scientific knowledge, enabling the framing of questions and formulating hypotheses to form a transdisciplinary project. Based on initial insights, the grand challenge is transformed into a common research project, for example, drawing inspiration from the fields of risk management (Aven, 2016) and ethical considerations (Hevelke and Nida-Rümelin, 2015) in technology: The deployment of self-driving cars presents transformative possibilities for road safety and urban mobility, yet their adoption brings forth trust challenges within technological and socio-technical systems, stemming from system failures and ethical dilemmas due to opaque decision-making during collisions implicated in critical situations involving AVs, compromising safety, reliability, explainability, and accountability. To tackle this, the team can formulate research questions on trust, ethics, and communication while setting benchmarks for scientific success. Minimal success might involve sparking new trust-oriented research in autonomous driving, while maximal success targets a holistic trust management approach for AV deployment and interaction.</p>
II. Production of New Connectable Knowledge	<p>II.I Clarification of the roles of scientists and stakeholders. A transdisciplinary team of stakeholders and scientists from diverse fields and disciplines is assembled, bringing together the necessary expertise to tackle the collective research project. In the following transdisciplinary workshops, the roles of scientists and stakeholders are defined in light of the project’s objectives and the intricate trust issues related to self-driving cars. For example, scientists from diverse fields would handle data generation, analysis, and validation of solutions, while stakeholders offer practical insights, feedback, data access for studies, and resources to ensure actionable and socially relevant strategies. A commitment to continuous dialogues and iterative feedback loops are benchmarks for success that would foster collaboration, emphasizing the mutual dependency between academic rigor and practical applicability for the project’s success.</p> <p>II.II Design of an integration concept. The transdisciplinary team develops an integration concept through engaged discourse. The integration concept focuses on incorporating key elements of trust around the user to generate new, interconnected knowledge. For example, incorporating key elements could focus on trustworthiness and risk integration to develop clear AI decision-making models for various risky situations such as sudden obstacles and adverse weather; sphere and terrain integration to address the specific transportation needs of urban, suburban, and rural communities; risk with terrain integration to enable autonomous driving algorithms to handle diverse conditions, minimizing risk; and trustworthiness sphere integration to ensure AI technology respects cultural norms and ethical standards across different regions and communities. Multiple sub-groups within the transdisciplinary team could oversee different trust</p>

Table 2 (continued)

Phase	Steps
III. Transdisciplinary Integration	elements and their integrations needed for a cohesive trust-enhancement strategy. This approach enhances stakeholder role clarity and aligns their contributions with specific outcomes by defining areas of expertise and assigning responsibility zones for implementing trust enhancement measures.
	II.III Implementation of the integration concept. The integration concept facilitates defining the project scope and objectives, developing a detailed plan, creating or modifying necessary components, rigorously testing the objectives, and deploying the integrated solution.
	Using methods, tools, and guides from the transdisciplinary toolkit (Bammer, 2015), based on the new, generated connectable knowledge, each sub-group can develop strategies to enhance trust in AVs, welcoming stakeholder feedback and diverse perspectives. For example, the potential impact of each strategy, aligned with the project goal, could be evaluated using a collaborative multi-criteria analysis, with criteria and weights determined in previous meetings to ensure a cohesive strategic direction. If no individual strategy performs optimally across all criteria, the team, informed by sub-group insights, would collectively devise a revised integrated strategy to enhance trust in AVs, merging diverse insights into a unified action plan.
	III.I Assessing the integrated results. The integration process ends with a result constitution within the integrative system, assessing the effectiveness of phase II insights in the capacity to tackle the initial grand challenge and advance scientific knowledge, with input from stakeholders and scientists in the evaluation process.
	During upcoming stakeholder discourse sessions, the focus could be on collaboratively evaluating the consolidated trust-enhancement strategy for AVs, for instance, using a scenario approach to identify specific trust-enhancement actions as outcomes for strengthening trust. The project's societal impact could be assessed through a survey distributed to stakeholders, exploring predefined success criteria and broader factors like shifts in problem understanding. Further, a conference organized by the project team, open to external experts from science and practice, can assess the anticipated scientific contribution and invite evaluation and feedback.
	III.II Compiling outputs for society and science. The reintegrated transdisciplinary knowledge generates outcomes for societal (e.g., strategies, concepts, measures, prototypes, technologies) and scientific (e.g., methodological and theoretical innovations, new research questions, hypotheses) praxis targeting the terrain's grand challenge. Furthermore, this knowledge growth provides valuable input for discourse within societal and scientific systems, functioning as valuable resources to tackle forthcoming, grand challenges.
	Upon project completion, the team could develop a comprehensive guide compiling knowledge and strategies to enhance trust in AVs, targeting key decision-makers in organizations, tech firms, and political and administrative entities, serving as a pivotal resource for comprehending and implementing self-driving car technologies. To enhance the project's systematic approach, scientific findings can be shared via articles in peer-reviewed journals and compiled into an edited book volume. Both outputs can make the project's discoveries accessible to societal and scientific circles, facilitating informed decision-making and future research.
	Consequences of the project on the societal and scientific discourses This hypothetical project could influence societal discussions about trust in AVs, for example, inspiring local institutions to launch pilot projects with semi-AVs based on the developed strategies. Additionally, it could inform regulatory discussions at the government level, demonstrating government engagement with the formulated strategies. The project could inspire further research in the scientific community, delving into various aspects of trust in AVs and promoting a deeper exploration of the formulated strategies. Consequently, it could have a lasting impact on discussions about AVs and trust in both societal and scientific realms.

other to restore, protect, and build trust in AI, which is essential for the resilience and security of societies and the effective functioning of global systems. Looking ahead, emerging trust dynamics between AI systems themselves—and between AI and humans in both directions—demand new conceptual approaches. Future trust frameworks must consider not only how humans trust AI, but also how AI systems might evaluate and respond to human reliability or even establish forms of AI-to-AI trust in networked and automated environments. These hybrid trust systems challenge traditional, anthropocentric definitions and call for a post-humanist expansion of trust theory. Integrating this perspective enhances our transdisciplinary framework's adaptability and future relevance. As we navigate the uncharted territories of AI, we recognize that trust in AI not only shapes our interpersonal trust relationships but also prompts a profound exploration of our essence. This journey is not just about technological advancement; it reflects our role as creators whose visions shape our destiny. It reminds us that we are active participants, called to reflect on our ultimate purpose in a rapidly evolving world.

Data availability
The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 21 August 2024; Accepted: 1 July 2025;
Published online: 18 July 2025

- Notes**
- 1 The Social Dilemma, <https://www.imdb.com/title/tt11464826/>
 - 2 Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot: <https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot>
 - 3 Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages: <https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages>

- 4 AI Now Institute. (2023). Zero trust AI governance. <https://ainowinstitute.org/publication/zero-trust-ai-governance>
- 5 <https://incites.help.clarivate.com/Content/Research-Areas/wos-research-areas.htm>
- 6 California sidelines GM Cruise's driverless cars, cites safety risk, <https://www.reuters.com/business/autos-transportation/california-suspends-gm-cruises-driverless-autonomous-vehicle-permits-2023-10-24/>
- 7 High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- 8 National Governance Committee for the New Generation Artificial Intelligence. (2019). Governance principles for a new generation of artificial intelligence: Develop responsible artificial intelligence. Ministry of Science and Technology of the People's Republic of China. <https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>
- 9 The White House. (2023, October 30). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. <https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>
- 10 Waag Futurelab. (n.d.). Recommendations for the use of AI in public processes. Waag. Retrieved May 5, 2025, from <https://waag.org/en/article/recommendations-use-ai-public-processes/>

References

- Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H (2024) Trust in ai: Progress, challenges, and future directions. *arXiv* 2403.14680
- Ahmed S, Alshater MM, El Ammari A, Hammami H (2022) Artificial intelligence and machine learning in finance: A bibliometric review. *Res Int Bus Finan* 61
- Alam L, Mueller S (2021) Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *Bmc Med Inform Decis Mak* 21
- Asan O, Bayrak AE, Choudhury A (2020) Artificial intelligence and human trust in healthcare: Focus on clinicians. *J Med Int Res* 22
- Aven T (2016) Risk assessment and risk management: Review of recent advances on their foundation. *Eur. J. Operational Res.* 253:1–13
- Baier A (1986) Trust and antitrust. *Ethics* 96:231–260
- Bammer G (2015) Toolkits for transdisciplinarity - toolkit #1. GAIA - Ecol. Perspect. Sci. Soc. 24:149–149. 141
- Barber B (1983) *The logic and limits of trust*. Rutgers University Press
- Bartz J et al. (2011) Oxytocin can hinder trust and cooperation in borderline personality disorder. *Soc. Cogn. Affect. Neurosci.* 6:556–563
- Bartz JA, Lydon JE (2006) Navigating the interdependence dilemma: Attachment goals and the use of communal norms with potential close others. *J. Personal. Soc. Psychol.* 91:77–96
- Bawack RE, Wamba SF, Carillo KDA, Akter S (2022) Artificial intelligence in e-commerce: A bibliometric study and literature review. *Electron. Mark.* 32:297–338
- Bellucci G, Chernyak SV, Goodyear K, Eickhoff SB, Krueger F (2017) Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. *Hum. brain Mapp.* 38:1233–1248
- Blacklaws C (2018) Algorithms Transparency and accountability. *Philos Transac R Soc* 376
- Bottom WP, Holloway J, Miller GJ, Mislin A, Whitford A (2006) Building a pathway to cooperation: Negotiation and social exchange between principal and agent. *Adm. Sci. Q.* 51:29–58
- Buchan NR, Croson RTA, Solnick S (2008) Trust and gender: An examination of behavior and beliefs in the investment game. *J. Econ. Behav. Organ* 68:466–476
- Bughin J et al. (2018) Skill shift: Automation and the future of the workforce. McKinsey Glob. Inst. 1:3–84
- Capraro V et al. (2024) The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus* 3(6):191. <https://doi.org/10.1093/pnasnexus/pgae19>. (in press)
- Chintalapati S, Pandey SK (2022) Artificial intelligence in marketing: A systematic literature review. *Int. J. Mark. Res.* 64:38–68
- Cook KS, Hardin R, Levi M (2005) *Cooperation without trust?* Russell Sage Foundation
- Crawford K (2021) *Atlas of ai : Power, politics, and the planetary costs of artificial intelligence*. Yale University Press
- Crawford K, Calo R (2016) There is a blind spot in ai research. *Nature* 538:311–313
- de Oliveira TM, Amaral L, Pacheco RCD (2019) Multi/inter/transdisciplinary assessment: A systemic framework proposal to evaluate graduate courses and research teams. *Res. Evaluation* 28:23–36
- Deutsch M (1958) Trust and suspicion. *J. Confl. Resolut.* 2:265–279
- Dirks KT, Kim PH, Ferrin DL, Cooper CD (2011) Understanding the effects of substantive responses on trust following a transgression. *Organ Behav. Hum. Dec.* 114:87–103
- Du WD, Han Q (2021) In International Conference on Image, Video Processing, and Artificial Intelligence
- Fehr E (2009) On the economics and biology of trust. *J. Eur. Econ. Assoc.* 7:235–266
- Felzmann H, Villarronga EF, Lutz C, Tamò-Larriex A (2019) Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc* 6
- Gabriel I (2020) Artificial intelligence, values, and alignment. *Minds Mach.* 30:411–437
- Gächter S, Herrmann B, Thöni C (2010) Culture and cooperation. *Philos. T. R. Soc. B* 365:2651–2661
- Gao XP, Bian XL (2021) Autonomous driving of vehicles based on artificial intelligence. *J. Intell. Fuzzy Syst.* 41:4955–4964
- Gordon D (2015) Superintelligence: Paths, danger, strategies. *Issues Sci. Technol.* 31:94–95
- Hancock PA (2023) Are humans still necessary? *Ergonomics* 1–8
- Hancock PA et al. (2023a) How and why humans trust: A meta-analysis and elaborated model. *Front. Psychol.* 14:11081086
- Hancock PA et al. (2023b) How and why humans trust: A meta-analysis and elaborated model. *Front Psychol* 14
- Hardin R (2002) *Trust and trustworthiness*. Russell Sage Foundation
- Heins C (2023) Artificial intelligence in retail - a systematic literature review. *Foresight* 25:264–286
- Hevelke A, Nida-Rümelin J (2015) Responsibility for crashes of autonomous vehicles: An ethical analysis. *Sci. Eng. Ethics* 21:619–630
- Holmes JG (1981) In *The justice motive in social behavior* (eds MJ Lerner & SC Lerner) 261–284 (Plenum Press)
- Holmes JG (1991) In *Advances in personal relationships Vol. 2* (eds WH Jones & D Perlman) 57–104 (Jessica Kingsley)
- Hurlburt G (2017) How much to trust artificial intelligence? *It Professional* 19:7–11
- Jahn T, Bergmann M, Keil F (2012) Transdisciplinarity: Between mainstreaming and marginalization. *Ecol. Econ.* 79:1–10
- Johnson J (2020) Artificial intelligence, drone swarming and escalation risks in future warfare. *Rusi J.* 165:26–36
- Kaplan AD, Kessler TT, Brill JC, Hancock PA (2023) Trust in artificial intelligence: Meta-analytic findings. *Hum. Factors* 65:337–359
- Kim P (2023) *How trust works: The science of how relationships are built, broken, and repaired*. Flatiron Books
- King A et al. (2023) A systematic scoping review protocol to summarise and appraise the use of artificial intelligence in the analysis of digital videos of invasive general surgical procedures. *Int. J. Surg. Protoc.* 27:118–121
- King-Casas B et al. (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308:78–83
- Knack S, Zeefer P (1997) Does social capital have an economic payoff? A cross-country investigation. *Q. J. Econ.* 112:1251–1288
- Kramer R (1999) Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annu. Rev. Psychol.* 50:569–598
- Krueger F (2021) *The neurobiology of trust*. Cambridge University Press
- Krueger F, Meyer-Lindenberg A (2019) Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends Neurosci.* 42:92–101
- Kumar CRP, Natarajan R, Padma K, Sivaperuman A (2022) Artificial intelligence in healthcare : A brief review. *Suran J Sci Technol* 29
- Laas O (2023) Deepfakes and trust in technology. *Synthese* 202
- Lange PAMV, Rockenbach B, Yamagishi T (2017) *Trust in social dilemmas*. Oxford University Press
- Li Y, Wu B, Huang Y, Luan S (2024) Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-ai trust. *Front. Psychol.* 15:1382693
- Luhmann N (2017) *Trust and power*. English edition. edn, (Polity)
- Lyons JB, Hamdan IA, Vo TQ (2023) Explanations and trust: What happens to trust when a robot partner does something unexpected? *Comput Hum Behav* 138
- Magd H, Jonathan H, Khan S, El Geddawy M (2022) Artificial intelligence—the driving force of industry 4.0. A roadmap for enabling industry 4.0 by artificial intelligence, 1–15
- Mariani MM, Borghi M (2023) Artificial intelligence in service industries: Customers' assessment of service production and resilient service operations. In *J Product Res*
- Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. *Acad. Manag. Rev.* 20:709–734
- McEvily B (2011) Reorganizing the boundaries of trust: From discrete alternatives to hybrid forms. *Organ. Sci.* 22:1266–1276
- McKnight DH, Cummings LL, Chervany NL (1998) Initial trust formation in new organizational relationships. *Acad. Manag. Rev.* 23:472–490
- Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. *Acm Comput Surv* 54
- Miller EE, Boyle LN (2019) Behavioral adaptations to lane keeping systems: Effects of exposure and withdrawal. *Hum. Factors* 61:152–164
- Mislin A, Campagna R, Bottom W (2011) After the deal: Talk, trust building and the implementation of negotiated agreements. *Organ Behav. Hum. Dec.* 115:55–68

- Montuori A (2013) The complexity of transdisciplinary literature reviews. *Complexity- Int. J. Complex. Educ.* 10:45–55
- Mumtaz, H et al. (2023) Exploring alternative approaches to precision medicine through genomics and artificial intelligence - a systematic review. *Front Med* 10
- Naik B, Mehta A, Yagnik H, Shah M (2022) The impacts of artificial intelligence techniques in augmentation of cybersecurity: A comprehensive review. *Complex Intell. Syst.* 8:1763–1780
- Nandi S et al. (2024) Deciphering the lexicon of protein targets: A review on multifaceted drug discovery in the era of artificial intelligence. *Mol. Pharmaceutics* 21:1563–1590
- Narkhede G, Dohale V, Mahajan Y (2024) Darker side of industry 4.0 and its impact on triple-bottom-line sustainability. *Sustain Dev*
- Omar AA, Farag MM, Alhamad RA (2021) Ieee. in 14th International Conference on Developments in eSystems Engineering (DeSE). 438–442
- Pandey AK, Chakraborty A, Khandal V (2024) Scientometric study of research on ai & ml application in defence technology and military operations. *Desidoc J. Libr. Inf. Technol.* 44:61–68
- Pérez IF, de la Prieta F, Rodríguez-González S, Corchado JM, Prieto J (2023) in 13th International Symposium on Ambient Intelligence. 155–166
- Putnam RD (2000) Bowling alone: The collapse and revival of american community. Simon & Schuster
- Qayyum A, Usama M, Qadir J, Al-Fuqaha A (2020) Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward. *Ieee Commun. Surv. Tutor.* 22:998–1026
- Qiu SL, Liu QH, Zhou SJ, Wu CJ (2019) Review of artificial intelligence adversarial attack and defense technologies. *Appl Sci-Basel* 9
- Redfern A (2009) Thomas hobbes and the limits of democracy. Cambridge Press
- Riedl R, Javor A (2012) The biology of trust: Integrating evidence from genetics. *Endocrinol., Funct. brain imaging J. Neurosci. Psychol. E* 5:63–91
- Riedl R, Mohr PNC, Kenning PH, Davis FD, Heekeren HR (2014) Trusting humans and avatars: A brain imaging study based on evolution theory. *J. Manag. Inf. Syst.* 30:83–114
- Rousseau DM, Sitkin SB, Burt RS, Camerer C (1998) Not so different after all: A cross-discipline view of trust. *Acad Manage Rev*
- Russell SJ, Norvig P (2021) Artificial intelligence: A modern approach. Pearson Education Limited
- Sabherwal R, Grover V (2024) The societal impacts of generative artificial intelligence: A balanced perspective. *J Assoc Inform Syst* 25
- Schilke O, Reimann M, Cook KS (2021) Trust in social relations. *Annu Rev. Socio.* 47:239–259
- Schlicker N, Langer M (2021) Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. *ACM International Conference Proceeding Series, Association for Computing Machinery*, 325–329
- Schoorman FD, Mayer RC, Davis JH (2007) An integrative model of organizational trust: Past, present, and future. *Acad. Manag. Rev.* 32:344–354
- Shaamala A, Yigitcanlar T, Nili A, Nyandega D (2024) Algorithmic green infrastructure optimisation: Review of artificial intelligence driven approaches for tackling climate change. *Sustain Cities Soc* 101
- Shaban-Nejad A, Michalowski M, Brownstein JS, Buckeridge DL (2021) Guest editorial explainable ai: Towards fairness, accountability, transparency and trust in healthcare. *IEEE J. Biomed. Health Inform.* 25:2374–2375
- Shou Q, Nishina K, Takagishi H (2021) in *The neurobiology of trust* (ed Krueger F) 369–386 (Cambridge University Press)
- Sijtsma H et al. (2023) Social network position, trust behavior, and neural activity in young adolescents. *NeuroImage* 268:119882
- Simpson JA (2007) Psychological foundations of trust. *Curr. Dir. Psychol. Sci.* 16:264–268
- Simpson TW (2012) What is trust? *Pac. Philos. Q.* 93:550–569
- Simpson TW (2023) Trust: A philosophical study. Oxford University Press
- Sligar AP (2020) Machine learning-based radar perception for autonomous vehicles using full physics simulation. *Ieee Access* 8:51470–51476
- Sullivan Y, de Bourmont M, Dunaway M (2022) Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. *Ann. Oper. Res.* 308:525–548
- Thiebes S, Lins S, Sunyaev A (2021) Trustworthy artificial intelligence. *Electron. Mark.* 31:447–464
- Van Eck NJ, Waltman L (2010) Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics* 84:523–538
- von Eschenbach WJ (2021) Transparency and the black box problem: Why we do not trust ai. *Philos. Technol.* 34:1607–1622
- Vaccari C, Chadwick A (2020) Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* 6
- Vasbinder JW et al. (2010) Transdisciplinary eu science institute needs funds urgently. *Nature* 463:876
- Werbach K (2018) The blockchain and the new architecture of trust. MIT Press
- Wiens J, Shenoy ES (2018) Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clin. Infect. Dis.* 66:149–153
- Williams J, Fiore SM, Jentsch F (2022) Supporting artificial social intelligence with theory of mind. *Front Artif Intellig* 5
- Wingert K, Mayer R (2024) in *A research agenda for trust: Interdisciplinary perspectives* (eds RC Mayer & BM Mayer) (Edward Elgar Publishing)
- Yuste R et al. (2017) Four ethical priorities for neurotechnologies and ai. *Nature* 551:159–163
- Yusuf SM, Baber C (2020) in 12th International Conference on Agents and Artificial Intelligence (ICAART). 347–354
- Zahedi F, Song J (2008) Dynamics of trust revision: Using health infomediaries. *J. Manag. Inf. Syst.* 24:225–248
- Zak PJ, Knack S (2001) Trust and growth. *Econ. J.* 111:295–321
- Zhang ZM et al. (2022) Artificial intelligence in cyber security: Research advances, challenges, and opportunities. *Artif. Intell. Rev.* 55:1029–1053
- Zhou J, Verma S, Mittal M, Chen F (2021) Understanding relations between perception of fairness and trust in algorithmic decision making. 8th International Conference on Behavioral and Social Computing 1–5

Acknowledgements

The University of Applied Sciences Upper Austria and the NeuroIS Society provided financial support to organize the Inaugural Meeting of the Transdisciplinary Research Union for the Study of Trust (T-R-U-S-T) Initiative, which took place in Vienna in June 2023.

Author contributions

FK and RR drafted the first draft, and all authors (FK, RR, JB, KSC, DG, PH, SLJ, LK, MRL, RCM, AM, GMP, TS, HT, and PVL) revised the final draft and approved the final version of this manuscript. The first two authors contributed equally to convening the Inaugural Meeting of the Transdisciplinary Research Union for the Study of Trust (T-R-U-S-T) Initiative meeting held in June 2023 in Vienna, Austria, and drafting this paper.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-025-05481-9>.

Correspondence and requests for materials should be addressed to Frank Krueger.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons

Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025