Rethinking Trust in Responsible AI

Marina Tropmann-Frick[1,*], Michael Gille[1], Susanne Draheim[1], Philine Pommerencke[1], Maximilian Kiener[2] and Jonas Bozenhard[2]

[1]*Hamburg University of Applied Sciences, Berliner Tor 7, 20099 Hamburg, Germany*

[2]*Hamburg University of Technology, Institute for Ethics in Technology, am Schwarzenberg-Campus 3, 21073 Hamburg, Germany*

Abstract

Trust is widely recognized as a core principle of Responsible AI, yet its interpretation varies significantly across disciplines. This paper examines how computer science, sociology, philosophy, and law conceptualize trust in AI systems, highlighting both tensions and complementarities. From a computer science perspective, trust is often approached as a set of system-level properties that should be formalized and evaluated with metrics. In contrast, the social sciences and humanities emphasize its relational, normative, and institutional dimensions. We argue that trust cannot be reduced to a single system property or technical measure, as it emerges from social-technical interactions involving users, developers, legal norms, and social expectations. To support interdisciplinary dialogue, we propose treating trust as a boundary concept that enables cooperation across epistemic communities acknowledging conceptual differences. Trust is widely recognized as a core principle of Responsible AI, yet its interpretation varies significantly across disciplines. This paper examines how computer science, sociology, philosophy, and law conceptualize trust in AI systems, highlighting both tensions and complementarities. From a computer science perspective, trust is often approached as a set of system-level properties that should be formalized and evaluated with metrics. In contrast, the social sciences and humanities emphasize its relational, normative, and institutional dimensions. We argue that trust cannot be reduced to a single system property or technical measure, as it emerges from social-technical interactions involving users, developers, legal norms, and social expectations. To support interdisciplinary dialogue, we propose treating trust as a boundary concept that enables cooperation across epistemic communities acknowledging conceptual differences.

Keywords

Responsible AI, trust, trustworthy AI, AI governance, boundary concept, interdisciplinarity

1. Introduction

Trust is widely invoked as an anchor concept of Responsible AI by regulators, researchers, and developers alike. Yet across disciplines, trust is conceptualized in strikingly different ways: as a technical attribute and measurable property, as a normative stance, or as a social relationship, sometimes with economic connotations. Conceptual equivocality handicaps efforts to design, assess, and govern AI systems responsibly. We contend that the absence of a shared conceptual foundation and a lack of reflective awareness of differences impedes interdisciplinary collaboration and, ultimately, weakens the governance of AI systems. This paper aims to highlight diverse epistemic meanings of the term trust. By doing so, we take an initial step toward clarifying how trust is understood and employed across Responsible AI discourses. Rather than proposing a single, unified definition of trust, we suggest treating it as a *boundary concept* [1], which acknowledges the various conceptualizations across disciplines and thereby supports innovation and mutual learning among researchers from different fields. Our goal is to lay the groundwork for a more integrated and reflexive approach to trust in AI contexts. We focus on interdisciplinary perspectives to reconceptualize trust as a boundary concept, i.e. one that can facilitate dialogue across disciplines without erasing conceptual differences.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

[*] Corresponding author.

marina.tropmann-frick@haw-hamburg.de (M. Tropmann-Frick); michael.gille@haw-hamburg.de (M. Gille); susanne.draheim@haw-hamburg.de (S. Draheim); philine.pommerencke@haw-hamburg.de (P. Pommerencke); maximilian.kiener@tuhh.de (M. Kiener); jonas.bozenhard@tuhh.de (J. Bozenhard)

0000-0003-1623-5309 (M. Tropmann-Frick); 0000-0001-9978-0817 (M. Gille); 0000-0001-7515-7473 (S. Draheim); 0009-0004-4062-5694 (P. Pommerencke); 0000-0001-8454-8265 (M. Kiener); 00000-0003-3346-0305 (J. Bozenhard)

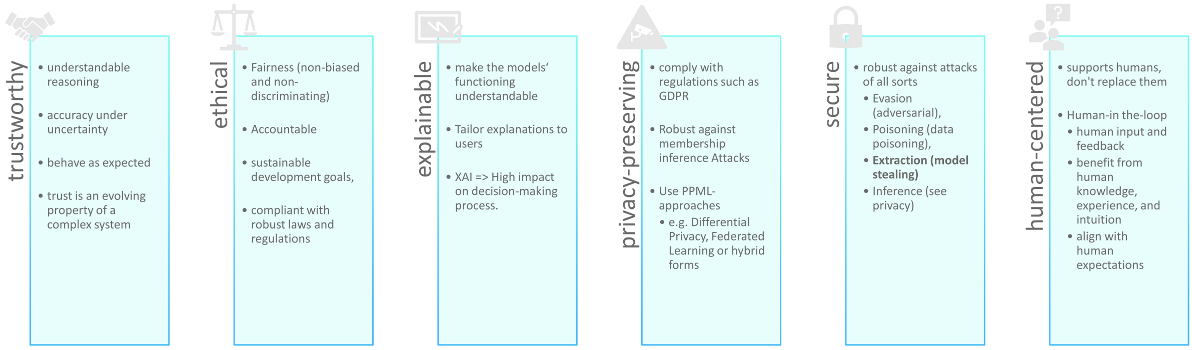
trustworthy
• understandable reasoning
• accuracy under uncertainty
• behave as expected
• trust is an evolving property of a complex system
ethical
• Fairness (non-biased and non-discriminating)
• Accountable
• sustainable development goals,
• compliant with robust laws and regulations
explainable
• make the models' functioning understandable
• Tailor explanations to users
• XAI => High impact on decision-making process.
privacy-preserving
• comply with regulations such as GDPR
• Robust against membership inference Attacks
• Use PPML-approaches
• e.g. Differential Privacy, Federated Learning or hybrid forms
secure
• robust against attacks of all sorts
• Evasion (adversarial),
• Poisoning (data poisoning),
• Extraction (model stealing)
• Inference (see privacy)
human-centered
• supports humans, don't replace them
• Human-in the-loop
• human input and feedback
• benefit from human knowledge, experience, and intuition
• align with human expectations

Figure 1: Conceptual Dimensions of Responsible AI.

Trust is often treated as if it were an integrative concept. We contrast this view by sketching its role in divergent epistemic frameworks, opening space for interdisciplinary discourse and critique and, at the same time, making the misalignment between (quantitative) indicators/metrics and the plural meanings of trust transparent. The meaning of trust depends on which discipline gets to define it, with considerable intradisciplinary differences. We want to frame trust as a site of negotiation, emphasizing process over consensus, and challenging the reductionist use of metrics. Our contribution is guided by the question, how trust can be understood and operationalized as a boundary concept that enables communication across disciplinary approaches to AI governance.

To provide a structured basis for interdisciplinary analysis, we use the VERIFAI framework of Responsible AI as a shared reference point [2, 3]. While trust itself may not be reducible to any single property, several system-level dimensions are commonly associated with trustworthiness and responsibility in both technical and non-technical contexts. As illustrated in Figure 1, we identify six dimensions of Responsible AI: Trustworthy system behavior, referring to predictability, reliability, and stability under uncertainty; ethical and legal alignment, including fairness, accountability, and compliance with laws and sustainability goals; explainability, ensuring that models and decisions are understandable to relevant users; privacy preservation, through legal compliance and technical protection against inference or leakage; security, addressing resilience to adversarial, poisoning, and extraction attacks; human-centeredness, promoting meaningful human oversight, agency, and alignment with user expectations.

Responsible AI refers in this context to the development, deployment, and governance of AI systems in a manner that is in line with ethical values, legal norms, and societal expectations. While definitions vary a lot across different scientific communities, we adopt the following as a provisional framing of Responsible AI [2, 4]: Responsible AI is human-centered and ensures users' trust through ethical ways of decision making. The decision-making must be fair, accountable, not biased, with good intentions, non-discriminating, and consistent with societal laws and norms. Responsible AI ensures that automated decisions are explainable to users while always preserving users privacy through a secure implementation.

2. Disciplinary Perspectives on Trust in AI

The concept of trust, embedded in the discourse surrounding Responsible AI, is interpreted and implemented very differently across disciplines. To unpack this complexity, we examine how trust is conceptualized and approached within four disciplinary domains central to AI governance: computer and data science, sociology, philosophy, and law. Although the terminology of trustworthy AI has been institutionalized in EU policy discourse (see, e.g. [5] and Recital 1 AI Act), this paper deliberately employs the broader notion of responsible AI. The reason is twofold: first, trustworthiness risks suggesting that trust is a measurable property of technology, whereas interdisciplinary scholarship underscores its relational and context-dependent nature; second, responsibility better captures the dynamic socio-technical, legal, and ethical practices through which legitimacy and accountability are constructed. The choice of terminology thus aims to extend the debate beyond compliance-driven checklists towards a more reflexive, process-oriented understanding. This section outlines the divergent yet complementary perspectives, highlighting where they converge, where they differ, and how they might inform a more integrated interdisciplinary understanding of trust in AI. Each of the subsections is authored by a domain expert, representing their respective disciplinary perspective.

2.1. Computer/Data Science

The Computer and Data Science field addresses trust in AI as a set of properties that can be formalized, measured, evaluated and optimized [6, 7]. This perspective allows for a division of the monolithic concept of trust into a composite of technically observable and preferably formally specified quantitative metrics. Metrics are not neutral, but metric choice, thresholding, evaluation and interpretation depend on the AI model, data usage and analytical purpose. Responsible AI properties encode assumptions about acceptable trade-offs, are often multi-objective and cannot be optimized simultaneously without conflict. As such, metric-based evaluation must be integrated with formal verification methods and domain-specific constraints.

We group the metrics according to core dimensions of Responsible AI (Figure 1). Direct formalizations and technical metrics are not readily available for dimensions such as trustworthiness and human-centeredness, which remain only partially measurable [3]. Accordingly, computer science research tends to focus on the four inner dimensions where quantitative evaluation is more established. Fairness metrics [8, 9] aim to quantify whether model outputs exhibit statistical parity across groups or individuals. Group-level metrics typically evaluate the statistical distribution of predictions with respect to sensitive attributes and ground truth labels. Individual-level fairness, by contrast, relies on consistency scores or counterfactual analysis, which estimate whether similar individuals receive similar outcomes under controlled perturbations of non-permissible features. Explainability and transparency [10, 11] are evaluated by assessing the interpretability of model behavior. Local explanation techniques such as SHAP and LIME approximate the marginal influence of input features on specific model outputs. Metrics such as faithfulness and stability assess whether these approximations capture actual model behavior under perturbation. In complex models such as deep neural networks, explanation quality becomes sensitive to architecture and gradient behavior, and is often evaluated through post hoc attribution or diagnostic probing. Privacy and robustness form another group of metrics [12, 13]. Differential privacy provides a formal privacy notion that bounds the information gain about individuals in the training data. Robustness, on the other hand, is measured via attack success rates under adversarial perturbations or via out-of-distribution generalization errors. Both require simulation of worst-case or bounded adversarial scenarios to quantify system reliability under non-standard input conditions.

Some of these metrics are partially incompatible: improving group fairness may reduce individual fairness; increasing robustness may reduce accuracy; adding privacy guarantees may impact explainability. Metric design therefore implicitly encodes value trade-offs and requires interpretation via domain experts.

Trust, while often linked to a system behavior, cannot be reduced to a single quantifiable or formalizable system property. Although technical dimensions such as fairness, robustness and explainability contribute to the dimension of trustworthiness, they capture only partial aspects of a much broader concept. Trust is context-dependent and conditioned by user expectations, domain-specific risks and boundaries, and environmental settings. From the computer science perspective this creates a fundamental limitation. Trust cannot be fully captured by technical metrics, it requires input from multiple disciplines.

2.2. Sociology

The lack of a uniform conceptualization of the term trust across different scientific disciplines also applies to sociology itself [14]. Introduced by Georg Simmel (1858–1918), the concept of trust has since been explored by numerous sociologists [15]. To examine all these conceptualizations in detail would exceed the scope of this paper. Therefore, two theoretical approaches that are widely recognized as classics in the sociology of trust, will be shortly summarized in the following and then applied to the field of trust in AI. The first approach comes from Niklas Luhmann, who was one of the first to systematically examine and analyse the concept of trust introduced by Simmel. The second approach summarized is from Anthony Giddens, who further elaborated Luhmann's conceptualization of trust.

Following the definition of Luhmann [16], trust can be seen as a mechanism to reduce complexity and enable people to make decisions and take actions in a complex (social) environment which does not give the individual enough knowledge to be secure about the future and consequences of their decision. As far as the degree of available information is concerned, trust cannot arise out of nothing, nor is it necessary when the situation is completely clear and certain. It builds upon past experiences and social interactions, which show some kind of continuity that suggests that the future is, yet uncertain, predictable in some way. Against this background, Luhmann [16] makes it clear that trust can be seen as a conscious act of will in which the individual decides to not try to seek more information to gain total security about the outcome, but to trust that the trustee will behave as expected. With this decision, the trustor takes a personal risk, as they make themself vulnerable against a breach of trust from the trustee.

Considering the complexity of our modern world, in which people do not only interact in a personal context, but more often are part or representative of a larger social system or institution, Luhmann [16] differentiates between interpersonal trust and institutional trust. While the former is directed at another individual, which is personally known by the trustor, the latter is directed towards an organization or abstract institution, such as law, money or science. Anthony Giddens [17] further elaborates Luhmann's concept and notes that interpersonal trust and systemic trust can also be related to each other, as persons can act as representatives of a system. Trust in another individual can be either interpersonal (meaning that the trust relies on past interactions between trustor and trustee and is directed towards the person), or systemic (i.e., trust is based on the trustor's general trust in the organisation or system to which the trustee belongs and is then transferred to the person). It can also be the other way round: One might not trust a system (e.g. a hospital), but one special person belonging to it (e.g. physician).

Applying the sociological theory of trust to the question what trust towards AI means, it is necessary to first determine the perspective. It is too unspecific to generally talk about "Trust towards AI" without specifying the context. For example, one could look at the trust a user has towards an LLM-based chatbot, applying the concept of interpersonal trust, or we could talk about trust in the organization that produced that chatbot, which would be the systemic trust. Therefore, a multifaceted approach is imperative, necessitating the consideration of a range of factors and context conditions. The question of whether to place trust in the provider or manufacturer of the AI is a salient one. The question of whether to place trust in the competence of the developers who designed and trained the system is a salient one. The question of whether to place trust in the specific AI tool in use is a salient one. The reliability and representativeness of the training data must be established. Alternatively, should we place our trust in the respective output generated in response to a specific prompt? [18]

This differentiated approach is particularly necessary because, in practice, a general attitude of over- or under-trust towards AI systems is often observed. Such sweeping generalizations are frequently influenced by individual prior knowledge, experience or narratives conveyed by the media. Instead, the necessity lies in the establishment of 'calibrated trust', which can be defined as a situation-specific evaluation of the reliability of an AI system within a designated application context, alongside the determination of the extent to which efforts should be invested in the critical monitoring of the respective outcomes [18].

From this perspective, the formation of trust in a technical system, such as AI, is inherently embedded in socio-technical interactions. This is an ongoing negotiation process between humans and machines, in which questions of (human) identity and control are redefined through interaction [19, 20]. A fundamental design objective is thus to devise and operationalize a socio-technical negotiating process between humans and machines, in addition to learning from data and its documentation. In this understanding, trust is not a prerequisite, but rather arises in the course of interaction – through the coupling and co-evolution of human and AI actors [21].

2.3. Philosophy

In the philosophy of trust, the most widely accepted definition invokes three conditions for trusting: The trustor relies on the trustee to be (a) competent and (b) willing or motivated to do what we trust them to do (based on shared moral norms) and is (c) exposed to some level of risk or vulnerability by doing so [22]. Trust relates in interesting ways, among others, to the concepts of reliance, transparency, and responsibility.

First, while trust involves a form of reliance, it is not reducible to it. Trust is more than predicting that someone will act a certain way, it is a morally loaded stance. As Baier [23] notes, "trusting can be betrayed, or at least let down, and not just disappointed." This distinction is crucial: reliance leads to disappointment when expectations are not met; trust leads to betrayal when shared moral expectations are violated. Disappointment reflects a failed prediction; betrayal reflects a broken moral commitment. Thus, whereas reliance is descriptive and predictive, trust is fundamentally normative: rooted in shared values and obligations.

Second, trust requires at least some level of transparency to be warranted. The more one knows about the trustee's competence and motivations, the better one can assess their trustworthiness and the risks involved in placing trust. However, trust conceptually still requires some degree of uncertainty too (e.g. about the competency and will of the trustee and external factors that might prevent the trustee from fulfilling what one has entrusted to them) and, thus, a leap of faith, as it were. Therefore, full transparency eliminates the need for trust, reducing it to a matter of risk calculation [22].

Third, trust entails a distinctive form of responsibility grounded in *answerability*. When we trust someone, we do not merely expect outcomes, we expect that they can be called to answer for their actions in light of shared norms [24, 25]. This means the trusted party is not just expected to act reliably, but to explain or *justify* their conduct if questioned. Unlike mere reliance, which does not presume moral engagement, trust invokes a relational obligation: the trustee is answerable to the trustor, and moral answerability is an important part of what gives trust its normative depth.

Overall, then, we argue that, from a philosophical perspective, applying the concept of trust to AI can be misleading. AI systems are not full-fledged moral agents with a will, and it is therefore inaccurate to say that an AI is "willing" or "motivated" to do what we trust it to do. When an AI fails to meet our expectations, we may be disappointed by its malfunctioning, but we would not say that it has *betrayed* us or violated moral norms. For this reason, it is more appropriate to apply the categories of *reliance* or *reliability* to AI, rather than the concept of trust. However, we can still apply trust to the people behind AI, namely, the designers, developers, and deployers of these systems. Accordingly, the notion of "trustworthy AI" must not conflate two distinct dimensions: first, the technical task of building reliable, transparent systems that minimize the risk of disappointment: what we might call *reliable* AI; and second, the moral and political responsibility of the human agents involved: those who can be held answerable for failures, biases, or harms.

2.4. Law

Among recent legislative attempts to address AI, the EU's AI Act stands out as the regulatory approach most comprehensively anchored in notions of trust [26, 27]. The AI Act´s legal basis, Article (Art.) 114 of the Treaty on the Functioning of the EU, aims to ensure the smooth functioning of the internal market. This leads the EU to lean on trust considerations as both a legal and economic condition of integrating its digital single market [28]. In this context, trust is not merely a matter of consumer confidence or technical compliance; it functions as a foundational principle for enabling cross-border exchanges of AI technologies under conditions of perceived legitimacy and shared risk tolerance. Establishing such a harmonized regulatory framework (Art. 1 AI Act; Recital 1), the EU seeks to preclude regulatory fragmentation among its member states and attempts to embed trustworthiness as a structural precondition for market participation, transforming trust from a diffuse socio-political and economic expectation into a legal construct with intended market-shaping effects.

The notion of trust is deployed by the European Commission to guide the design and operation of AI systems, and to promote societal acceptance of these technologies across EU member states [29]. In its ambition to establish "*a legal framework for trustworthy AI*", the European Commission envisages the development of a human-centric "*ecosystem of trust*" [30]. In line with this statement, the EU AI Act elevates trust (and trustworthiness) to the status of a guiding principle and conceptualizes it not merely as an ethical aspiration but a structural element of legitimacy that underpins the EU's entire regulatory strategy for AI. The concept of trust is applied to market proponents and AI products as well as to the (EU internal) market for AI as a whole.

The term trustworthiness, despite being so central, is only sparsely embedded in the binding provisions of the Act, appearing explicitly only in Art. 1 (statement of purpose) and Art. 95 (voluntary codes of conduct). Its formal legal status is, therefore, somewhat diffuse. To determine the trustworthiness of AI systems, metrics are recommended, intended to promote trust [5]. Internal metrics, while not binding, enable firms to demonstrate measurable progress and accountability, a possible factor in gaining trust. The AI Act operates on the premise that trust emerges where risk is either absent or noticeably mitigated (Art. 1 (1) AI Act; recitals 65, 164). With this risk-based approach to regulating innovative technology (AI Act, recital 5), the EU therefore determines the absence or mitigation of risks and harm as a requisite for trustworthiness (AI Act, recital 25-27), in line with the discussion beyond the AI Act [31]. Nevertheless, it is questionable whether legality automatically establishes or strengthens trust [26]. The same holds for the acceptability of risks, which will not always lead to trustworthiness of AI systems, at best being a necessary condition, not a sufficient one. On closer inspection, conceptual ambiguity, legal uncertainty and practical difficulties impede consistent trust governance [32], which is further exacerbated by the central importance of trust concepts, albeit with different emphases, in, e.g., the General Data Protection Regulation, the Digital Markets Act and the Digital Services Act, which are outside the scope of this contribution.

3. Inter- and Cross-disciplinary Discussion and Outlook

Trust, while central to Responsible AI discourse, is interpreted and operationalized differently across disciplines. From a computer science perspective, trust is primarily linked to measurable system properties such as fairness, robustness, and explainability. Sociology, in contrast, frames trust as an evolving relationship embedded in context, shaped by values, practices, and institutional arrangements. Philosophical perspective emphasizes trust as an epistemic stance, based on expectations, intentions and moral responsibility. Legal perspectives link trust to accountability and institutional frameworks, especially in the context of the European AI regulation.

These divergent conceptualizations emphasize different objects and/or perspectives (like technical systems, human actors or institutions and processes), rely on different levels of assumptions (like formal, interpretive or normative), and pursue different aims. Such differences in the notion of trust make it nearly impossible to develop a unified theoretical definition, but combined they offer complementary insights that are essential for understanding trust in AI. Therefore, we propose conceptualizing trust as a boundary concept that is sufficiently flexible to establish interdisciplinary collaboration without a specific definitional consensus. This boundary concept approach is particularly well suited to be enriched by additional perspectives from, e.g., psychology and economics, which would provide further insights into individual and systemic dimensions of trust.

We see potential in the development of hybrid approaches that systematically combine formal, technical and contextual analyses. Such approaches may involve the co-design of evaluation frameworks, interdisciplinary review mechanisms and participatory methods incorporating diverse stakeholder perspectives. This objective can be seen as a driver to enable a calibration of trust to the level of user confidence in an AI system proportional to the capabilities, limitations, application context and social environment.

This analysis is part of an ongoing effort to develop a deeply interdisciplinary approach not only for trust in AI, but also for related complex concepts such as human-centeredness, accountability, and social acceptability in socio-technical AI systems. We aim to move beyond discipline-specific definitions and instead build a shared framework with respect to conceptual differences enabling joint system design and evaluation.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and Deepl translator for grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. Löwy, The strength of loose concepts—boundary concepts, federative experimental strategies and disciplinary growth: The case of immunology, History of science 30 (1992) 371–396.

[2] S. Göllner, M. Tropmann-Frick, B. Brumen, Towards a definition of a responsible artificial intelligence, in: Information modelling and knowledge bases XXXV, IOS Press, 2024, pp. 40–56.

[3] S. Göllner, M. Tropmann-Frick, Bridging the gap between theory and practice: Towards responsible ai evaluation., in: CHAI@ KI, 2023, pp. 68–76.

[4] S. Goellner, M. Tropmann-Frick, B. Brumen, Responsible artificial intelligence: A structured literature review, arXiv preprint arXiv:2403.06910 (2024).

[5] A. HLEG, High-level expert group on artificial intelligence.(2019). ethics guidelines for trustworthy ai, European Commission. Available at: https://ec. europa. eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019).

[6] A. Gittens, B. Yener, M. Yung, An adversarial perspective on accuracy, robustness, fairness, and privacy: multilateral-tradeoffs in trustworthy ml, IEEE Access 10 (2022) 120850–120865.

[7] P. Hu, Y. Lu, et al., Dual humanness and trust in conversational ai: A person-centered approach, Computers in Human Behavior 119 (2021) 106727.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.

[9] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar, A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices, in: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 2239–2248.

[10] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, C.-C. Tu, Leveraging latent features for local explanations, in: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, 2021, pp. 1139–1149.

[11] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, Advances in neural information processing systems 31 (2018).

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).

[13] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE symposium on security and privacy (SP), IEEE, 2017, pp. 3–18.

[14] J. Evers, Vertrauen eine soziologische betrachtung, in: Vertrauen und Wandel sozialer Dienstleistungsorganisationen: Eine figurationssoziologische Analyse, Springer, 2017, pp. 37–55.

[15] G. Möllering, The nature of trust: From georg simmel to a theory of expectation, interpretation and suspension, Sociology 35 (2001) 403–420.

[16] N. Luhmann, Vertrauen: Ein mechanismus der reduktion sozialer komplexität, UTB, 2000.

[17] A. Giddens, J. Schulte, Konsequenzen der moderne, Suhrkamp, 1995.

[18] U. Schmid, Trustworthy artificial intelligence: comprehensible, transparent and correctable, Hannes Werthner· Carlo Ghezzi· Jeff Kramer· Julian Nida-Rümelin· Bashar Nuseibeh· Erich Prem· (2024) 151.

[19] H. C. White, Identity and control: How social formations emerge, Princeton university press, 2008.

[20] R. Häußling, C. Härpfer, M. Schmitt, Soziologie der Künstlichen Intelligenz: Perspektiven der Relationalen Soziologie und Netzwerkforschung, transcript Verlag, 2024.

[21] C. Härper, Von der kunst des lernens: Einige bemerkungen zur intentionalität von in-und output (2024).

[22] C. McLeod, E. N. Zalta, Trust in stanford encyclopedia of philosophy, Metaphysics Research Lab, Stanford University (2006).

[23] A. Baier, Trust and antitrust, ethics 96 (1986) 231–260.

[24] M. Kiener, Varieties of answerability, in: The Routledge Handbook of Philosophy of Responsibility, Routledge, 2023, pp. 204–216.

[25] M. Kiener, Strict moral answerability, Ethics 134 (2024) 360–386.

[26] A. Tamò-Larrieux, C. Guitton, S. Mayer, C. Lutz, Regulating for trust: Can law establish trust in artificial intelligence?, Regulation & Governance 18 (2024) 780–801.

[27] B. Lund, Z. Orhan, N. R. Mannuru, R. V. K. Bevara, B. Porter, M. K. Vinaih, P. Bhaskara, Standards, frameworks, and legislation for artificial intelligence (ai) transparency, AI and Ethics (2025) 1–17.

[28] A. Engel, Licence to regulate: Article 114 tfeu as choice of legal basis in the digital single market, in: New Directions in Digitalisation: Perspectives from EU Competition Law and the Charter of Fundamental Rights, Springer Nature Switzerland Cham, 2024, pp. 13–28.

[29] B. Beckert, The european way of doing artificial intelligence: The state of play implementing trustworthy ai, in: 2021 60th FITCE communication days congress for ICT professionals: Industrial data–cloud, low latency and privacy (FITCE), IEEE, 2021, pp. 1–8.

[30] E. Union, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, COM/2021/206final (2021) 1–107.

[31] J. Newman, A taxonomy of trustworthiness for artificial intelligence, CLTC: North Charleston, SC, USA 1 (2023).

[32] M. Kattnig, A. Angerschmid, T. Reichel, R. Kern, Assessing trustworthy ai: Technical and legal perspectives of fairness in ai, Computer Law & Security Review 55 (2024) 106053.