2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events
(PerCom Workshops)

# “The Trust in AI-Mix: Technical, Legal and Ethical Requirements as well as Psychology”

Jan Spilski
*Center for Cognitive Science*
*University of Kaiserslautern-Landau*
*(RPTU)*
Kaiserslautern, Germany
0000-0003-4105-4585

Anja Spilski
*Business School Pforzheim*
*Pforzheim University*
*(HS PF)*
Pforzheim, Germany
0009-0007-3485-6340

Thomas Lachmann
*Centro de Investigación Nebrija en*
*Cognición (CINC)*
*Universidad Nebrija*
Madrid, Spain
0000-0002-6901-5935

***Abstract*——The increasing spread of artificial intelligence (AI) applications underscores the necessity for an interdisciplinary initiative to ensure the trustworthiness of AI and its alignment to human values and security. A distinction is made between trustworthy AI and trust in AI. The former is defined by adherence to ethical, technical, and legal standards and describes an implementation perspective. The latter reflects a psychological perspective and focuses on the perception of users. For the implementation of trustworthy AI, ethical guidelines from institutions and companies, technical challenges in terms of explainability, robustness, and bias avoidance, and legal requirements are discussed. The user perspective is considered in two respects; first drawing from implementation guidelines provided by the human-centered AI approach. Second, while focusing on the perceptual perspective of trust in AI, findings of developmental psychology, cognitive science, and human–computer interaction are discussed. By connecting technical implementation, ethical and legal requirements, as well as psychological insights, we propose a comprehensive and holistic "trust-in-AI mix" to guide the development of AI systems that foster user trust in AI.**

***Keywords— trust, trustworthy, artificial intelligence, technical, legal, ethical, requirements, psychology***

## I. Introduction

Artificial Intelligence (AI) has exhibited significant advancement in a multitude of tasks across diverse domains [1] and has demonstrated the potential to become a dominant technology in the coming years. AI can be defined as the simulation of human intelligence processes by machines, particularly computer systems [2]. These processes include learning (the acquisition of information and its utilization), logical reasoning (inferences) and self-correction. Interactions with AI applications have become an integral part of everyday life for many people, for example when using facial recognition to unlock a mobile phone or when speaking to a voice assistant [3]. The development of generative AI (e.g., ChatGPT) represents a profound advance in AI capability.

Despite the rapid spread of AI, concerns on the protection of privacy or the prevention of bias, discrimination, hallucination and misinformation have been raised [4, 5, 6]. There is a consensus that any AI-based system must be compatible with human values, work in the interest of human safety and continue to give humans the possibility to influence the AI system [7]. Consequently, there is an increasing necessity for AI that is both trustworthy and responsive in terms of social and ethical considerations [8].

The issue of trust in AI systems, therefore, is a research issue in diverse domains. From a technical perspective, initial solutions for the technical challenges associated with building trustworthy AI systems exist [1]. An example is GAIA-X [9] in the European Union (EU), which has developed a technical framework for trustworthy AI. From the ethical and legal perspectives, countries and organizations have formulated guidelines and regulations to give orientation in the process of developing trustworthy AI. Examples are the "Ethics Guidelines for Trustworthy AI" [10] and the "AI Act" [11] in the EU, the program "Explainable Artificial Intelligence" (XAI) [12] in the U.S.A., or the "Governance Principles for a New Generation of Artificial Intelligence: Developing Responsible Artificial Intelligence" in China [13].

A further perspective focuses on the users of AI systems and forms a human-centered approach to artificial intelligence (HCAI) that complements the aforementioned perspectives. As a design framework for developing future AI solutions [14], HCAI can be defined as follows: "*Human-Centered Artificial Intelligence utilizes data to empower and enable its human users, while revealing its underlying values, biases, limitations, and the ethics of its data gathering and algorithms to foster ethical, interactive, and contestable use*" [15]. Consequently, recent research has provided a comprehensive framework for HCAI with three fundamental dimensions – Technology, Ethics, User – and their interactions [5], as well as a corresponding methodological framework [14, p. 3] (shown in Figure 1).

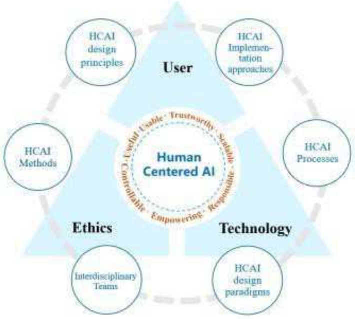

Fig. 1. A comprehensive HCAI framework (taken from Xu et al. [14, p. 3])

However, the extant guidelines to develop trustworthy AI systems do not adequately address the question of how end users actually build trust in AI systems [8]. Instead, the field seems to be predominantly driven by technical considerations [16], although there is consensus across research, industry, and policy regarding the necessity of human-centered aspects. However, rather abstract guidelines without a deeper understanding of how users develop trust in AI may limit the utility of the guidelines [17]. In addition, recent research [17] on the extent to which the EU design principles have been implemented through design rules in the major tech companies demonstrated only a partial fulfilment. A notable gap exists in terms of key human-centered requirements. Before this background, not only the human-

2766-8576/25/$31.00 ©2025 IEEE
DOI 10.1109/PerComWorkshops65533.2025.00068

centered *development* of AI but also the question of human *perceptions* of trust in AI systems is of special importance.

Therefore, the aim of the present article is to reinforce the human-centered perspective on AI in two respects. First, with respect to the HCAI idea of integrating technical, ethical/legal and user-centered aspects, we give an overview on the relevant AI design principles from these dimensions. Second, we add the perceptual (i.e. psychological) perspective with findings from the fields of cognitive science, developmental psychology, and human-computer interaction in relation to AI. Finally, our discussion will consider the technical, ethical, legal, and psychological perspectives as a “trust-in-AI mix”.

## II. Trustworthy AI and Trust in AI

We distinguish trustworthy AI and perceived trust in AI. The term "trustworthiness” is an attribute of an AI system [18] that has been designed in accordance with technical, legal, and ethical *requirements*, and may be objectively considered a trustworthy AI. The implementation of these requirements ensures that the expectations of stakeholders are met in a verifiable manner [19].

In contrast, the term "trust in AI" refers to the subjective perception of trust by individuals. We suggest that individuals may build trust even in an untrustworthy AI system or, vice versa, they may distrust an objectively trustworthy AI system. The reasons for these potential discrepancies may be explained by psychological insights related to AI. Consequently, in the course of this article we will consider the *psychological perspective* and discuss findings on how individuals build trust in AI systems.

The term "trust" is defined in different disciplines, including sociology, psychology, and economics [20]. In a systematic review on user trust in AI-supported technologies [21], the definition by Mayer et al. [22] was identified the most frequently used. Accordingly, trust is *"the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party"* [22, p. 712). Consequently, trust in technology is a socio-technical construct that incorporates a reciprocal relationship between people and technology [2]. Trust in AI is dynamic and subject to change over time [23]. The formation of "initial trust" is contingent on an individual's disposition or external influences [24] and serves as an important factor in the adoption of novel technologies [25]. A different question is how trust, once established, can be maintained over time [23].

## III. Implementation Perspective: Developing Trustworthy AI Systems

As indicated previously, the development process of trustworthy AI systems should be based on the consideration of requirements related to *ethical* and *technical* issues that are complemented by the *user* perspective (HCAI approach). Since ethical and technical issues are usually reflected in laws, we add the consideration of *legal* issues.

### A. Ethical Issues

From the ethical perspective, the term "trustworthy AI" encompasses not only the trustworthiness of the AI system itself, but also the trustworthiness of all processes and actors that are part of the system's lifecycle. A number of institutions and organizations have already developed guidelines to inform the development of ethically trustworthy AI. Two broad categories can be distinguished: general design guidelines and action-oriented design principles. Examples for general design guidelines are the "Guidelines for Ethically Oriented Design" [26] provided by the IEEE, or the EU "Ethics Guidelines for Trustworthy AI" [10]). Action-oriented design principles usually stem from initiatives of technology companies and their research divisions who have introduced design principles for practical implementation.

With regard to general design guidelines, the EU "Ethics Guideline for Trustworthy AI" includes four ethical principles, (1) respect for human autonomy, (2) prevention of harm, (3) fairness, and (4) explainability [10], that should be considered at each stage of the development process (proposal, implementation, operation of AI services). To guide the implementation of these principles, the following seven requirements for trustworthy AI have been formulated: *Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental well-being, Accountability.*

As mentioned above, a number of technology companies formulated action-oriented design principles to guide practical implementation. The "People + AI Guidebook" by Google presents methods, best practices, and illustrative examples for the design of human-centered AI [27]. The document presents 23 design patterns that address commonly encountered inquiries of developers. The "HAX-Toolkit" by Microsoft [28] concentrates on the interaction between humans and AI, with the objective of creating a positive user experience using 18 design principles that encompass the AI-user interaction process.

However, research on whether these more practical principles fulfil the general EU design guidelines found discrepancies [17], particularly for the requirements related to diversity, non-discrimination, fairness, and environmental and social wellbeing. Researchers conclude that there is still a dearth of pragmatic approaches for implementing human-centered AI design [29], which represents a substantial obstacle to the practical deployment of HCAI [15, 30].

### B. Technical Issues

The term "technically trustworthy AI" is the desired result from decisions on the proposal and development of an AI system and its progress by modifications. Technical challenges include the explainability of the processes, the technical robustness [1], the avoidance of hallucination in large language models (LLMs), and the prevention of bias.

*Explainability* is the extent to which the internal mechanics of the AI system can be understood in human terms, i.e., explaining "what it has done, what it is doing now, and what will happen next; and disclose the salient information that it is acting on" [31]. However, there is an inverse correlation between the learning performance of the AI system and the explainability of these capabilities [1]. Chander et al. [1] provided an overview on the diverse tools to approach explainability.

*Technical Robustness* can be defined as the capacity to withstand adverse conditions and digital security threats [1]. Adversarial attacks are deliberate perturbations conducted to mislead AI models into incorrect predictions. These attacks exploit vulnerabilities by introducing subtle changes that are imperceptible to humans. Robustness pertains to the capability of a system or algorithm to address the potential for errors, missing values, erroneous and previously unseen data [18]. Consequently, controlled and certified attack methods like the Fast Gradient Sign Method or the LBFGS attack [32] can be used to train the systems in order to enhance their resilience [33]. Tools like CleverHans [34] and the Adversarial Robustness Toolbox [35] can be used to test and improve the defense mechanisms of models.

The combined consideration of explainability and robustness can foster a trustworthy AI system. However, trade-offs may appear between current techniques for improving both criteria [36].

*Avoidance of hallucination in LLMs:* Although LLMs (e.g., ChatGPT) demonstrate coherent responses and evident lines of reasoning, their results are based on statistical (on the basis of probability) patterns in word embeddings rather than true cognitive processes [37, 38]. Hallucination and misinformation may and do occur. Because hallucination refers to false but convincing outputs, a system may be perceived trustworthy without a foundation. To point this out strikingly, Hicks et al. [39] even chose the statement "ChatGPT is bullshit" as the title of their article. Methods like "chain-of-thought" prompting can enhance reliability by breaking tasks into granular steps. However, because LLMs do not integrate true reasoning processes [40], outputs can also be manipulated through tailored prompts, making them susceptible to misinformation. Efforts to improve robustness (e.g., GPT-4o) mitigate but do not eliminate these risks. There is still the black-box nature of LLMs which impedes recognizing faults in their outputs. Without a cognitive grounding in first-order facts, LLMs cannot be fully trusted [41]. Zhou et al. [41] posit that future progress will be primarily achieved by combining generative capabilities with structured reasoning.

*Prevention of bias:* AI development should be fair. Tools like AI Fairness360 detect and mitigate biases, offering 70 bias metrics and 10 algorithms for comprehensive bias evaluation[42].

### C. Legal Issues

The protection of data that is used for the AI systems is one aspect of the relevant legal requirements and reflected in data protection regulations in several countries. In addition to these regulations that govern the general handling of data, legal requirements have also been created in more specific areas of AI. These include, for example, AI regulations in the EU [11], China [13] and some federal laws in the U.S. that address AI [43]. In terms of the contents of regulations, the "EU AI Act" highlights four key aspects, (1) an AI classification according to related risk, (2) that obligations are mainly related to developers and providers of high-risk AI systems, (3) that considered users are natural or legal persons that deploy an AI system in a professional capacity, not affected end-users, and (4) the handling of "general-purpose AI". Four risk classifications are differentiated: unacceptable, high, limited, and minimal risk. An example of an unacceptable risk are social scoring systems and manipulative AI. In contrast, minimal risk is unregulated, which includes AI-based video games or spam filters. However, it is essential to conduct continuous reassessments of risk classification, particularly in light of developments such as generative AI [11]. The obligations put on the developers and providers increase in accordance with the risk classification.

Assessments are available on whether AI systems are in line with the "EU AI Act" in order to provide organizations with pragmatic direction to the translation of the rather abstract tenets into verifiable criteria [44]. Constantinides et al. [45] present a semi-automatic method for assessing the risks of AI in mobile and wearable applications. They used LLM prompts to generate realistic AI use cases and classified each use case concerning the EU AI Act risk classification. Applied to 138 generated AI applications and their risk classification, the method demonstrated an accuracy rate of 85% compared to a manual validation process [45].

### D. User Issues

The HCAI adds the user perspective when thinking about designing AI solutions. Aiming to "maximize the benefits of AI technology to humans and avoid its potential adverse effects" [14, p. 1), the approach brings forward the ideas of the ethical and technical perspectives and connects them with key human factors. The three dimensions (technology, ethics, user) were broken down to implementation approaches, connected with key human factors and lead to the following seven design goals: *Trustworthy AI, Scalable AI, Useful AI, Usable AI, Empowering humans, Responsible AI*, *Human controllable AI* [14]. The HCAI as a comprehensive and human-integrating approach provides a lot of guidance for developers of AI systems. However, its focus is on implementation. The guidelines suggest how to design AI systems in a way that users *potentially* trust, but do not consider further individual factors that explain when and how users actually build trust in AI systems. In order to obtain a complete picture, we consider perceptual insights and turn to a psychological perspective on AI.

## IV. Psychological Perspective: How Individuals Build Trust in AI

When asking when and how individual build trust in AI systems, we can draw from various subdisciplines of psychological research and human-computer interaction.

### A. Insights from Developmental Psychology

Developmental psychologists understand the concept of trust as a foundational emotional state that emerges predominantly during the early stages of childhood [46]. An influencing factor is the concept of "attachment style" which has been extensively researched [47]. Attachment style emerges from early social interactions and has been demonstrated to exert a significant influence on human behavior not only in childhood but also through adulthood [48]. In a series of three experimental studies, Gillath et al. [2] found a relation between attachment security and trust in AI.

### B. Cognitive and Emotional Routes to Trust in AI

In terms of perception, two primary routes to building trust in AI have been identified [49], the cognitive (based on rational thinking) and the emotional route. Both routes could have a significant impact on human trust [50]. For the cognitive route, five key factors explain the building of trust by an individual in the domain of virtual AI, embedded AI and robotic AI systems. These are *Tangibility, Transparency, Reliability, Task Characteristics* and *Immediacy Behaviors*. For the emotional route, the following key factors lead to higher trust in AI: *Tangibility, Anthropomorphism, and*

*Immediacy Behaviors* [49]. The relationships between the identified factors and trust in AI are summarized in Table I.

TABLE I. COGNITIVE & EMOTIONAL ROUTES TO TRUST IN AI

| | | |
|---|---|---|
| Cognitive Routes | Tangibility | Physical presence and visual presence enhance trust. |
| | Transparency | Trust in AI increases with transparency about its reliability and algorithmic processes, particularly where clarity on system functionality is crucial for complex managerial applications. |
| | Reliability | Low reliability leads to a decrease of trust in AI. It difficult to rebuild trust and it takes a lot of time. However, robots that are perceived as highly intelligent tend to maintain user compliance even when they malfunction. |
| | Task Characteristics | Trust in AI is higher for technical and data-driven tasks, but lower for tasks requiring social intelligence, where human competence is preferred. |
| | Immediacy Behaviors | Responsiveness, adaptability, pro-social actions, personalization and persuasive strategies enhance trust. Constant employee surveillance undermines it. |
| Emotional Routes | Tangibility | Physical presence can increase likeability, but can also trigger fear. A "persona" increases emotional trust, while unawareness of AI involvement can provoke anger. Positive emotions are often associated with the good reputation of the developing company. |
| | Anthropomorphism | Human-like characteristics promote positive emotions and increase trust, but can also cause discomfort and create unrealistic expectations of AI capabilities. Trust is further enhanced by attractiveness and user-related personalization (e.g. similar ethnicity or facial features). |
| | Immediacy Behaviors | Human-like actions increase emotional trust and likeability, with flawed robots often preferred to flawless ones. The impact of human-like behavior on trust varies according to user predispositions. |

*Note.* The table based on the explanations from Glikson & Woolley [49, p. 632, 643].

### C. *Insights from Research on Human-Computer Interaction*

A systematic review in the area of human-computer interaction (HCI) [21] identified three prominent influencing factors on user trust in AI systems: user characteristics, socio-ethical factors, and technical and design features.

#### 1. *User Characteristics*

The *personality traits of users* (in particular the Big Five, [51]), exert an influence on the level of trust that individuals place in AI-supported systems. Zhou et al. [52] pointed out that a user interface should be designed to collect and display user personality traits. This would enable users to become aware of their personality traits and understand how these influence their decision-making when interacting with AI.

*Prior positive experiences and repeated exposure* of the user with a provider or manufacturer facilitates the transfer of trust to other systems from the same provider or manufacturer [53, 54]. Repeated interactions with AI systems enhance user understanding and increase perceptions of system integrity. This may be attributed to the greater availability of cognitive resources. Studies on embodied conversational agents revealed that trust increased with duration of use, irrespective of initial trust levels [54, 55].

Users differ in terms of their *technical proficiency regarding AI use, user uncertainty and intention to use*. Those with low technical competence and low intention to use believe less in a successful collaboration [56].

The early involvement of users, provision of training, enhancement of the user experience, and empowerment are regarded as potential solutions for achieving acceptance and willingness to use, and consequently trust in AI. In order to reduce user uncertainty, it is necessary to improve factors that influence user comprehension, the feeling of control, and information accuracy [57]. The divergence between user expectations and experiences was identified as a potential threat to user trust, particularly in instances where users rely heavily on specific AI [58]. Furthermore, reducing cognitive load enhances trust by enabling more effective system comprehension and user motivation [52].

#### 2. *Socio-ethical Factors*

How an AI-supported system is integrated into its surrounding environment influences the initial trust that users place in such systems [58]. This concerns a number of factors including visible means for data protection, high-quality of user interactions [59], performance-oriented technical support [60], maintenance of open communication channels, and the collection of continuous feedback.

#### 3. *Technical and Design Features*

Technical and/or design attributes for virtual agents (e.g., chatbots, avatars) can be employed to increase trust [21]. These include anthropomorphism and human-like characteristics, benevolent characteristics like smiling and showing interest in the user, using an AI agent design that fosters the perception of a physical and psychological closeness to the user [53], the social presence of the AI-supported system [49], or the integrity of the AI system (i.e., repeated satisfactory task performance) [53]. Sharma [61] demonstrated that the presence of multimedia features, security certificates or logos, contact details, and a social network logo on AI-based systems utilizing a website were significant factors influencing user trust.

No single factor is sufficient to determine user trust. A multiplicity of factors influences trust, thereby underscoring the necessity of tailoring system features to the attributes of the target audience. Technical and design elements that affect user trust should inform the system design strategy, delineating which features to prioritize in accordance with the context and system objectives. An understanding of the characteristics of the user can facilitate the optimization of AI systems for specific user types, thereby enabling adjustments that enhance system usability and trust [21].

## V. DISCUSSION & CONCLUSSION

"AI is [only] a collection of software techniques which boosts the machines to learn and do the computing tasks. So, AI is not a trusted one, it needs increase in trustworthiness, which is still far away" [1, p. 44]. In this article, we have consequently distinguished *trustworthy AI* and *trust in AI*. The former is defined by adherence to ethical, technical, and legal standards, the latter reflects a psychological construct which is influenced by factors such as user characteristics, socio-ethical factors, and technical and design features [21] and which is inherently dynamic. Findings from the sub-disciplines of psychology can elucidate the processes through which individuals perceive, evaluate, and adapt to AI systems. For instance, initial trust is frequently contingent upon external cues such as endorsements or user reviews [24, 25], whereas sustained trust necessitates consistent performance and demonstrable reliability [23]. Errors, such as AI failures or hallucination in LLMs, can erode trust despite technical robustness [37, 41].

While the trade-off between explainability and performance remains a central technical challenge, integrative approaches, such as the combination of explainable AI (XAI) techniques with robust AI against adversarial attacks, present potential avenues for the enhancement of both transparency, reliability, and hence, trust in AI. Nevertheless, these solutions require precise calibration to prevent any potential compromise to system efficiency.

This highlights the necessity of designing systems that not only perform well but also communicate limitations effectively to manage user expectations. Despite an objectively trustworthy design, users may still exhibit mistrust due to cognitive biases, personality, previous experiences, a lack of transparency, or misunderstandings about the system's capabilities [21]. These potential discrepancies require a dual approach that not only develops systems that meet robust standards but also fosters user trust, for example, by education and transparent communication.

The HCAI approach suggests the integration of technical and ethical frameworks and relevant laws in a comprehensive and human-oriented approach that provides substantial guidance for developers of AI systems. Nevertheless, its guidelines offer suggestions for the design of AI systems that users may perceive as trustworthy, without consideration of the additional individual factors that influence when and how users actually establish trust in as trustworthy designed AI systems. Therefore, we call for a stronger consideration of the psychological perspective on AI. The trust-in-AI mix, as outlined in the title, should therefore consider the technical, ethical, legal, and psychological perspectives in order to effectively build trustworthy AI systems that are perceived as reliable by the general public. In order to implement this approach, the following requirements must be met:

Researchers from a range of disciplines should work together to advance this field of study. It is crucial to facilitate discourse between technologists, ethicists, psychologists, and legal experts to advance the development of comprehensive AI systems. It is essential to conduct systematic and continuous evaluation throughout the entire AI product life cycle. Initial approaches in the area of compliance with regulatory and ethical requirements have been demonstrated with the use of semi-automatic evaluations [45]. Furthermore, it is necessary to integrate user-centered metrics into the standards and, where necessary, to develop and validate them according to psychometric quality standards. This will enable the use of these empirical metrics to quantify user trust dynamics and to integrate feedback loops to facilitate iterative improvements in AI designs.

The consideration of trust in AI systems requires valid methods for measuring human trust. Trust as a multifaceted and latent psychological construct cannot be directly observed [48]. A multitude of methodologies, such as self-reported measures, metrics such as compliance rates and decision times, physiological measures including eye tracking or brain imaging techniques have been employed to elucidate the cognitive and emotional processes underlying trust. Although these methods are promising, they rely on subjective scale correlations and are resource-intensive [8].

Finally, technical and design features play a pivotal role in shaping user perceptions. Elements like multimedia integration, anthropomorphism, and consistent task performance foster user trust. However, the effectiveness of these features depends on their alignment with the system's context and the target audience's characteristics.

To conclude, a multidisciplinary approach is necessary both to design and to evaluate AI systems. Insights from developmental psychology, cognitive and emotional psychology, and HCI reveal critical insights. Not just one single factor determines trust. Understanding the cognitive, emotional, and socio-technical dimensions of trust is essential for building trustworthy AI that evoke trust in AI.

## ACKNOWLEDGMENT

The project was funded by the German Federal Ministry for Economic Affairs and Climate Action (research project GAIA-X Merlot, reference 68GX21008#).

## REFERENCES

[1] B. Chander, C. John, L. Warrier, and K. Gopalakrishnan, "Toward Trustworthy Artificial Intelligence (TAI) in the Context of Explainability and Robustness," *ACM Comput. Surv.*, 2024, doi: 10.1145/3675392.

[2] O. Gillath, T. Ai, M. S. Branicky, S. Keshmiri, R. B. Davison, and R. Spaulding, "Attachment and trust in artificial intelligence," *Computers in Human Behavior*, vol. 115, p. 106607, 2021.

[3] H. Liu et al., "Trustworthy AI: A Computational Perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, pp. 1–59, 2023.

[4] J. Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*", in K. Martin, Ed., Ethics of data and analytics. Concepts and cases, pp. 1–4. [Online]. Available: https://www.taylorfrancis.com/books/9781003278290

[5] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI," *ArXiv*, abs/2105.05424, 2021.

[6] R. Yang and S. Wibowo, "User trust in artificial intelligence: A comprehensive conceptual framework," *Electron Markets*, vol. 32, no. 4, pp. 2053–2077, 2022, doi: 10.1007/s12525-022-00592-6.

[7] O. Ozmen Garibay *et al.,* "Six Human-Centered Artificial Intelligence Grand Challenges," *International Journal of Human–Computer Interaction*, vol. 39, no. 3, pp. 391–437, 2023.

[8] Y. Li, B. Wu, Y. Huang, and S. Luan, "Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust," *Frontiers in psychology*, vol. 15, p. 1382693, 2024, doi: 10.3389/fpsyg.2024.1382693.

[9] GAIA-X European Association for Data and Cloud, *Gaia-X Trust Framework - 22.10 Release.* [Online]. Available: https://docs.gaia-x.eu/policy-rules-committee/trust-framework/22.10/(accessed: 2024)

[10] European Commission and Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*: Publications Office, 2019. Accessed: Nov. 29 2024. [Online]. Available: https://data.europa.eu/doi/10.2759/346720

[11] *Artificial Intelligence Act: AI Act*, 2024. Accessed: Nov. 29 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj

[12] D. Gunning and D. W. Aha, "DARPA's Explainable Artificial Intelligence Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[13] National Governance Committee for the New Generation Artificial Intelligence, *Governance Principles for the New Generation Artificial Intelligence: Developing Responsible Artificial Intelligence.* Available: https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html (accessed: Nov. 29 2024).

[14] W. Xu, Z. Gao, and M. Dainoff, "An HCAI Methodological Framework: Putting It Into Action to Enable Human-Centered AI," Nov. 2023. [Online]. Available: http://arxiv.org/pdf/2311.16027v3

[15] T. Capel and M. Brereton, "What is Human-Centered about Human-Centered AI? A Map of the Research Landscape," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany, 2023, pp. 1–23. Accessed: Nov. 29 2024.

[16] J. Lee, *Industrial AI.* Singapore: Springer Singapore, 2020. [Online]. Available: https://doi.org/10.1007/978-981-15-2144-7

[17] J. Kutz, J. Neuhüttler, J. Spilski, and T. Lachmann, "AI-based Services-

Design Principles to Meet the Requirements of a Trustworthy AI," in *The Human Side of Service Engineering*, 2023, pp. 57–66.

[18] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy Artificial Intelligence: A Review," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–38, 2023, doi: 10.1145/3491209.

[19] Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence, ISO/IEC TR 24028:2020, International Organization for Standardization, 2020 - 05.

[20] O. Schilke, M. Reimann, and K. S. Cook, "Trust in Social Relations," *Annu. Rev. Sociol.*, vol. 47, no. 1, pp. 239–259, 2021.

[21] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective," *International Journal of Human–Computer Interaction*, vol. 40, no. 5, pp. 1251–1266, 2024.

[22] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *The Academy of Management Review*, vol. 20, no. 3, p. 709, 1995, doi: 10.2307/258792

[23] Siau, Ken, Wang, Weiyu, "Building Trust in Artificial Intelligence, Machine Learning, and Robotics.," *Cutter Business Technology Journal*, 2018, pp. 47–53, 2018.

[24] D. H. McKnight, L. L. Cummings, and N. L. Chervany, "Initial Trust Formation in New Organizational Relationships," *The Academy of Management Review*, vol. 23, no. 3, p. 473, 1998, doi: 10.2307/259290.

[25] X. Li, T. J. Hess, and J. S. Valacich, "Why do we trust new technology? A study of initial trust formation with organizational information systems," *The Journal of Strategic Information Systems*, vol. 17, no. 1, pp. 39–71, 2008, doi: 10.1016/j.jsis.2008.01.001.

[26] R. Chatila and J. C. Havens, "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," in *Intelligent Systems, Control and Automation: Science and Engineering, Robotics and Well-Being*, M. I. A. Ferreira, J. S. Sequeira, G. S. Virk, M. O. Tokhi, and E. E. Kadar, Eds., Springer International Publishing, 2019, pp. 11–16.

[27] Google PAIR, *People + AI Guidebook: Designing human-centered AI products.* [Online]. Available: https://pair.withgoogle.com/guidebook/ (accessed: Nov. 29 2024).

[28] Microsoft, *Guidelines for Human-AI Interaction: Microsoft HAX Toolkit.* [Online]. Available: https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/ (accessed: Nov. 29 2024).

[29] M. Hartikainen, K. Väänänen, A. Lehtiö, S. Ala-Luopa, and T. Olsson, "Human-Centered AI Design in Reality: A Study of Developer Companies' Practices," in *Nordic Human-Computer Interaction Conference*, Aarhus Denmark, 2022, pp. 1–11.

[30] W. J. Bingley *et al.,* "Where is the human in human-centered AI? Insights from developer priorities and user experiences," *Computers in Human Behavior*, vol. 141, p. 107617, 2023.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

[32] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," Aug. 2016. [Online]. Available: http://arxiv.org/pdf/1608.04644v2

[33] S. Kotyan and D. V. Vargas, "Adversarial robustness assessment: Why in evaluation both L0 and L∞attacks are necessary," *PloS one*, vol. 17, no. 4, e0265723, 2022, doi: 10.1371/journal.pone.0265723.

[34] N. Papernot *et al.,* "Technical Report on the CleverHans v2.1.0 Adversarial Examples Library," Jun. 2018. [Online]. Available: http://arxiv.org/pdf/1610.00768v6

[35] M.-I. Nicolae *et al.,* "Adversarial Robustness Toolbox v1.0.0," Nov. 2019. [Online]. Available: http://arxiv.org/pdf/1807.01069v4

[36] A. Datta, M. Fredrikson, K. Leino, K. Lu, S. Sen, and Z. Wang, "Machine Learning Explainability and Robustness," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Virtual Event Singapore, 2021, pp. 4035–4036.

[37] H. Ying *et al.,* "InternLM-Math: Open Math Large Language Models Toward Verifiable Reasoning," May. 2024. [Online]. Available: http://arxiv.org/pdf/2402.06332v2

[38] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," Available: http://arxiv.org/pdf/2303.08896v3

[39] M. T. Hicks, J. Humphries, and J. Slater, "ChatGPT is bullshit," *Ethics Inf Technol*, vol. 26, no. 2, 2024, doi: 10.1007/s10676-024-09775-5.

[40] J. Wei *et al.,* "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[41] B. Zhou, D. Geißler, and P. Lukowicz, "Misinforming LLMs: vulnerabilities, challenges and opportunities," Aug. 2024. [Online]. Available: http://arxiv.org/pdf/2408.01168v1

[42] M. Arnold *et al.,* "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM J. Res. & Dev.*, vol. 63, 4/5, 6:1-6:13, 2019, doi: 10.1147/JRD.2019.2942288.

[43] White & Case LLP, *AI Watch: Global regulatory tracker - United States.* [Online]. Available: https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united

[44] L. Floridi, M. Holweg, M. Taddeo, J. Amaya Silva, J. Mökander, and Y. Wen, "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act," *SSRN Journal*, 2022, doi: 10.2139/ssrn.4064091.

[45] M. Constantinides, E. P. Bogucka, S. Scepanovic, and D. Quercia, "Good Intentions, Risky Inventions: A Method for Assessing the Risks and Benefits of AI in Mobile and Wearable Uses," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, MHCI, pp. 1–28, 2024.

[46] L. Markson and Y. Luo, "Trust in early childhood," in *Advances in Child Development and Behavior*: Elsevier, 2020, pp. 137–162.

[47] J. Bowlby, "Attachment and loss: retrospect and prospect," *The American journal of orthopsychiatry*, vol. 52, no. 4, pp. 664–678, 1982, doi: 10.1111/j.1939-0025.1982.tb01456.x.

[48] J. A. Simpson and G. Vieth, "Trust and Psychology: Psychological Theories and Principles Underlying Interpersonal Trust," in *The neurobiology of trust*, F. Krüger, Ed., Cambridge, NY, Port Melbourne, New Delhi, Singapore: Cambridge University Press, 2022, pp. 15–35.

[49] E. Glikson and A. W. Woolley, "Human Trust in AI: Review of Empirical Research," *ANNALS*, vol. 14, no. 2, pp. 627–660, 2020.

[50] K. A. Hoff and M. Bashir, "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2014, doi: 10.1177/0018720814547570.

[51] R. R. McCrae and P. T. Costa, "Comparison of EPI and psychoticism scales with measures of the five-factor model of personality," *Personality and Individual Differences*, vol. 6, pp. 587–597, 1985.

[52] J. Zhou, S. Luo, and F. Chen, "Effects of personality traits on user trust in human–machine collaborations," *J Multimodal User Interfaces*, vol. 14, no. 4, pp. 387–400, 2020, doi: 10.1007/s12193-020-00329-9.

[53] J. Foehr and C. C. Germelmann, "Alexa, Can I Trust You? Exploring Consumer Paths to Trust in Smart Voice-Interaction Technologies," *Journal of the Association for Consumer Research*, vol. 5, no. 2, pp. 181–205, 2020, doi: 10.1086/707731.

[54] Z. Yan, Y. Dong, V. Niemi, and G. Yu, "Exploring trust of mobile applications based on user behaviors: an empirical study," *J Applied Social Pyschol*, vol. 43, no. 3, pp. 638–659, 2013.

[55] A. C. Elkins and D. C. Derrick, "The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents," *Group Decis Negot*, vol. 22, no. 5, pp. 897–913, 2013, doi: 10.1007/s10726-012-9339-x.

[56] M. Klumpp and H. Zijm, "Logistics Innovation and Social Sustainability: How to Prevent an Artificial Divide in Human–Computer Interaction," *J of Business Logistics*, vol. 40, no. 3, pp. 265–278, 2019, doi: 10.1111/jbl.12198.

[57] H. Hoffmann and M. Söllner, "Incorporating behavioral trust theory into system development for ubiquitous applications," *Pers Ubiquit Comput*, vol. 18, no. 1, pp. 117–128, 2014.

[58] M. Lee, L. Frank, and W. IJsselsteijn, "Brokerbot: A Cryptocurrency Chatbot in the Social-technical Gap of Trust," *Comput Supported Coop Work*, vol. 30, no. 1, pp. 79–117, 2021.

[59] X. Lin, X. Wang, and N. Hajli, "Building E-Commerce Satisfaction and Boosting Sales: The Role of Social Commerce Trust and Its Antecedents," *International Journal of Electronic Commerce*, vol. 23, no. 3, pp. 328–363, 2019, doi: 10.1080/10864415.2019.1619907.

[60] M. T. Thielsch, S. M. Meeßen, and G. Hertel, "Trust and distrust in information systems at the workplace," *PeerJ*, vol. 6, e5483, 2018.

[61] V. M. Sharma, "A comparison of consumer perception of trust-triggering appearance features on Indian group buying websites.," *Indian Journal of Economics and Business*, vol. 14, no. 2, pp. 163–177, 2015.