"Trust" and "Trustworthiness" in the AI Act

Lucia Franke, Benjamin Müller

Abstract *In this paper, we examine the use of "trust" and "trustworthiness" within the AI Act and supporting EU documents on AI. We argue that the underlying concept is limited to reliability, which misconstrues trust as a calculable object and therefore neglects its fundamental meaning. In contrast to this techno-centric view we briefly sketch the idea of trust to demonstrate the essential social and interpersonal aspects of trust, which cannot be neglected in this manner. Our final remarks reflect these considerations within the broader context of the EU's ethical framework.*

> "In a context of rapid technological change, we believe it is essential that trust remains the bedrock of societies, communities, economies and sustainable development. We therefore identify **Trustworthy AI as our foundational ambition**, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place." (EG: 4)

In 2024, the Council of the European Union pioneered the world's first comprehensive official AI regulation, titled the Artificial Intelligence Act (AI Act). It may surprise some that one of its key aspects is the stipulation that Artificial Intelligence (AI) shall be "trustworthy". What does this mean? How can AI be trustworthy? And what concept of trust underlies these requirements?

To answer these questions, this paper first examines the use of "*trust*" and "*trustworthiness*" within the AI Act and the supporting Ethics Guidelines, while also briefly outlining the general EU framework. Subsequently, it interprets the conception of trustworthiness found in these texts primarily as reliability and then criticizes it on that basis. Third, it presents an alternative understanding of trust to highlight a perspective that may be missing from the EU framework. Finally, it summarizes the main considerations and offers a brief outlook on possible further steps.

1. The EU's Usage of Trust and Trustworthiness in AI Documents

With the EU AI Act, the EU Commission aimed to achieve two key goals: creating legal certainty for providers and promoting trust[1] among users, including customers and national public administrations, regarding the use of AI. This aim reflects the Act's explicit acknowledgement of "the need to build trust" (Recital 6 AI Act; Impact Assessment Executive Summary: 2). The legislator seeks to achieve this, in part, by promoting "human-centric and trustworthy artificial intelligence" (Art. 1 (1) AI Act; Recital 1 AI Act). Despite this emphasis, the Act itself provides no explicit definition or further explanation of "trustworthy AI". What do "human-centric" and "trustworthy" mean in this context?

Neither the definitions in Art. 3 nor the subsequent normative provisions offer a substantive clarification of these key terms. The recitals suggest a definition rooted in overarching principles (Recital 7 AI Act) and alignment with broader EU values and fundamental rights (see e.g., Recital 27 AI Act), yet they ultimately fall short of providing a precise conceptual framework.[2]

To gain a clearer conceptual understanding of this use of "trust" – or, more precisely, the predominantly used term "trustworthiness" – it is helpful to examine the earlier policy initiatives and guidance documents that have informed EU legislation on artificial intelligence.[3] Among these, the *Ethics Guidelines for Trustworthy AI* (hereinafter "EG"), drafted by the *EU High-Level Expert Group on Artificial Intelligence* (AI

1 Interestingly, the noun "trust" itself appears only three times within the Recitals of the AI Act. In contrast, the adjective "trustworthy" is used more frequently.

2 They suggest that trustworthiness depends, at a minimum, on risk mitigation (Recital 64 AI Act), safety, transparency, institutional oversight, and privacy (Recital 68 AI Act), and adherence to ethical AI principles (Recital 165 AI Act).

3 While the European Commission, as early as April 2018, when it presented its plan for 'Artificial Intelligence for Europe,' had already determined that the development and application of AI in the EU required 'an environment of trust and accountability around the development and use of AI', COM(2018) 237 final, p. 13; in order to build and strengthen this trust, measures for data protection and IT security, as well as for the explainability of AI systems – the argument being that those who do not understand how AI works cannot trust it – and through effective legal remedies for those harmed by AI were already considered necessary at that time, COM(2018) 237 final, pp. 14 et seq) , this notion did not translate into concrete regulatory language. In the White Paper on Artificial Intelligence published in February 2020, the Commission presented a comprehensive policy concept for AI regulation, which was based on the two pillars of an 'ecosystem of trust' and an 'ecosystem of excellence,' thereby strongly emphasizing the importance of 'trust,' yet again without providing a clear definition of the term or engaging with the concept in a substantive manner (White Paper on Artificial Intelligence – A European approach to excellence and trust, Brussels, 19.02.2020, COM(2020) 65 final, pp. 1, 3). Only the Impact Assessment, which served as the basis for the initial draft of the AI Act, includes in its glossary the definition of 'Trustworthy AI' derived from the EG.

HLEG) – a body explicitly established by the European Commission – stands out.[4] While not legally binding, these Guidelines were referenced during the preparatory process for the legislative proposal, within the text of the AI Act itself (Art. 95 (2) (a) AI Act) and several times in its recitals (Commission, SWD(2021) 84 final, Part 1/2: 2, 10, 38, 41). They describe their understanding of trustworthy AI in terms of three major components:

> "(1) it should be lawful, ensuring compliance with all applicable laws and regulations, (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since to ensure that, even with good intentions, AI systems do not cause any unintentional harm. Each component is necessary but not sufficient to achieve Trustworthy AI. Ideally, all three components work in harmony and overlap in their operation. Where tensions arise, we should endeavor to align them." (EG, Conclusion: 35)

To elaborate on these three abstract keywords (*lawful*, *ethical* and *robust*), a comprehensive framework is also established. Proceeding from fundamental rights and four Ethical Principles, the group develops seven key requirements for trustworthy AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability.[5] None of these requirements alone, nor all of them collectively, can ensure trustworthiness; rather, they are intended to form the foundation for possible trust.

In addition, these seven key requirements constitute the foundation of an assessment list for trustworthy AI. This list underwent a piloting process with over 350 stakeholders. The final edition was published on July 17, 2020, with the following promotion: "Through the Assessment List for Trustworthy AI (ALTAI), AI principles are translated into an accessible and dynamic checklist that guides developers and deployers of AI in implementing such principles in practice." (ALTAI) The website also advertises with the slogan: "Measure if your organization's AI is trustworthy". Although the authors of the EG are indeed aware that trustworthiness does

4 The lack of explicit definitions or conceptual clarifications within the AI Act itself is consistent with earlier stages of EU regulation development on artificial intelligence. Already the Commission's initial Communication on "Artificial Intelligence for Europe" (COM(2018)237 final, April 2018) and later the White Paper on Artificial Intelligence (February 2020) broadly emphasized the critical importance of trust without concretely defining or operationalizing the term. The Ethics Guidelines developed by AI HLEG therefore constitute an essential interpretative resource in this respect.

5 The AI Act interprets these as ethical principles. See Recital 27 AI Act. The wording of the EG itself isn't strict, they also refer to this list as principles. See EG: 14, Fn 35.

not arise merely from rules and law, and that it cannot be guaranteed, produced, or even definitively verified. They even remind readers explicitly "that such assessment list will *never be exhaustive*. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying requirements, evaluating solutions, and ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders therein" (EG: 31).

Thus, it is a complicated and ongoing process for a technology to gain trustworthiness, one with no definitive endpoint. Instead, trustworthy AI requires continuous consideration and implementation within the broader intellectual approach. "Beyond developing a set of rules, ensuring Trustworthy AI requires us to build and maintain an ethical culture and mind-set through public debate, education and practical learning." (EG: 9) These two aspects – cultivating an ethical culture and fostering an appropriate mindset – are central to this endeavor, creating an environment in which trust can grow and new technologies like AI can gain trustworthiness. The recitals of the AI Act and the Ethical Guidelines clearly state that both are based on the fundamental rights of EU Treaties and the EU Charter, which are rooted in human dignity. Therefore, the entire approach of this framework is labeled as "human-centric" (See AI Act R 1, 6 and 27 as well as EG: 9 et seq.).[6]

Concerning the trustworthiness of AI, the overall message conveyed by these documents appears to be that AI is trustworthy if it is embedded in the general framework of the EU, its value system, and its set of beliefs. Instead of defining trust, these texts focus on the necessary circumstances for trust.
Compliance with these framework conditions is thus presented as a mechanism that confers trustworthiness.

2. Which Meaning of Trust?

Having examined the presentation of "Trust" and "Trustworthiness" in the AI Act and the EG, the question of which specific understanding of trust underlies these

6 The glossary of the EG additionally sums up the idea of Human-Centric AI: "The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoy[s] a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come." EG: 37.

requirements arises. The AI Act and EG provide no explication of "trust" or "trustworthiness". A simplistic or reductive definition would, in any case, likely undermine the subtle and complex social nature of this phenomenon. Nevertheless, these documents must presuppose a shared, perhaps common, understanding of "trust" and "trustworthiness". The terms themselves, however, allow for a range of interpretations. A framework alone does not guarantee an appropriate understanding of trust. Without further explanation or specification, it remains unclear which particular conception is used. What kind of trust, then, is meant?

Before analyzing the specific understanding of "trust" and "trustworthiness" in the AI Act and EG, it is important to distinguish between these two notions. These terms are too often used interchangeably in general discourse, and, it seems, at times also in the EU documents themselves.

Trust usually refers to a subjective, relational stance or attitude of a person or community towards another entity (be it a person, institution, or system). It involves vulnerability, expectation, and unpredictability.

Trustworthiness, conversely, refers to the objective qualities, characteristics, or adherence to norms possessed by an entity that make it worthy of being trusted. It is a property independent of whether anyone actually decides to trust.

While these notions are connected, neither is a necessary or sufficient condition for each other. When someone is trustworthy, it does not mean that everybody automatically trusts him. Conversely, one can be trusted yet not be trustworthy at all. Worthiness alone is no guarantee for anything; yet, it is still a desirable quality.

The analysis in the following subsections will argue that the EU framework primarily focuses on establishing conditions for AI's trustworthiness, largely interpreting this concept through the lens of technical reliability and compliance (see Section 2.1). It is possible that the AI Act, at least in parts, uses the term "trustworthiness" in ways specific to IT security law (see Recital 27 CRA) or a political agenda (e.g., Ribeiro et al., 2016, among others).[7] The incorporation of technical "*trustworthiness levels*" for risk categorization and certification into the Act demonstrates the partial integration of such a technical understanding of trust into the regulatory framework.[8]

---

7 Recital 123 AI Act, which stipulates that a conformity assessment is required to ensure a high level of trustworthiness, can be interpreted as meaning that trustworthiness is achieved solely through this assessment, which itself incorporates the concept of trustworthiness levels from Regulation 2018/881.

8 According to ENISA, cybersecurity and trustworthiness are closely intertwined: "the requirements of trustworthiness complement and sometimes overlap with those of AI cybersecurity in ensuring proper functioning. […] Hence, trustworthiness features such as robustness, oversight, accuracy, traceability, explainability and transparency inherently support and complement cybersecurity" (ENISA, *Cybersecurity of AI and Standardisation*, 2023: 9–10). However, ENISA also points out that trustworthiness depends not only on cybersecurity, but also on regulatory oversight and control (ENISA, *Cybersecurity of AI and Standardisation*, 2023: 23).

Furthermore, "trustworthy AI" serves as a political objective and guiding slogan prominent in the policy discourse preceding the formal AI Act proposal. Foundational documents, such as the European Commission's White Paper on Artificial Intelligence, emphasized "trust" as a prerequisite for public acceptance and the successful economic development and deployment of AI technologies within the Union (Proposal: 1, 30; White Paper: 1, 3). In this context, "*trustworthiness*" was framed as an economic necessity – based on the rationale that only trustworthy AI would ultimately succeed in the market (Proposal: 30) – rather than as a precisely defined technical or ethical attribute.

2.1. Trustworthiness as Reliability

Nevertheless, the underlying concept of trust in the EU documents seems to be orientated towards economical and mechanical modes of thought, which filter through the wording of the texts under consideration, even though the practice and culture of trust are also emphasized.

The recitals of the AI Act already allude to a functional, measure-based understanding of trustworthiness (e.g. Recital 64, 123 AI Act). References to specific measures, such as mandatory risk management systems, conformity assessments, and technical solutions like content watermarking, exemplify the measure-based dimension; the functional aspect is highlighted through principles implying operational characteristics like technical robustness and transparency, and by tying requirements to the system's defined purpose and context (Laux et. al. 2024: 3, 6; otherwise Ho and Caals 2024: 358).

Robustness is also one of the three key components for trustworthy AI in the EG. The EG elaborate on the need for robust AI as follows:

> "Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. Such systems should perform in a safe, secure and reliable manner, and safeguards should be foreseen to prevent any unintended adverse impacts. It is therefore important to ensure that AI systems are robust. This is needed both from a technical perspective (ensuring the system's technical robustness as appropriate in a given context, such as the application domain or life cycle phase), and from a social perspective (in due consideration of the context and environment in which the system operates). Eth-

ENISA thus posits that the AI Act's requirements (data and data governance, record-keeping, transparency and provision of information to users, human oversight, risk management system, quality management, conformity assessment, robustness) aim to establish a trustworthy AI ecosystem, although they often coincide with cybersecurity requirements (ENISA, *Cybersecurity of AI and Standardisation*, 2023: 19–20).

> ical and robust AI are hence closely intertwined and complement each other." (EG: 7)

This quotation provides significant insight into the underlying understanding of trustworthiness and is merely one example among others. Systems should be safe, secure, and reliable, potential harm needs to be prevented. However, the call to "*prevent any unintended adverse impacts*" is the antithesis of trust: It is a call for control. Meanwhile, robustness itself is a technical term, arguably ill-suited and imprecise for social issues. What "trustworthy" appears to mean here is *reliability* (see also Costa 2024: 39).

Accordingly, the other two components – "*lawful*" and "*ethical*" – would then simply mean reliably conforming to law and ethics. This interpretation is also supported by the phrasing in the EG's conclusion: AI "should be lawful, ensuring compliance with all applicable laws and regulations, (2) it should be ethical, ensuring adherence to ethical principles and values" (EG, p. 35). Compliance and adherence leave no room for deviation and thus no opportunity to genuinely trust.

Reliability is indeed a core component of trustworthiness. An entity that consistently fails to perform as expected – such as a person who regularly breaks promises or a doctor whose diagnoses are mostly wrong – is justifiably deemed not trustworthy. However, while reliability is central to trustworthiness, the fundamental issue with reducing the concept of trustworthiness solely to technical reliability, or with equating trustworthiness (even if encompassing reliability) with trust itself, is the neglect of the freedom aspect and the relational, often uncertain, dynamics inherent in genuine trust. Trust is not simply inherent in an entity's demonstrated reliability or trustworthiness. The main issue with limiting trust or trustworthiness to reliability is the neglect of the necessary aspect of freedom in trust and, consequently, of all social and interpersonal meanings of trust. Doing so makes trust or trustworthiness seem like a calculable object, one that can be measured, operationalized, or even produced, as the assessment list in the Guidelines implies (EG: 24 et seq. and ALTAI).

Unlike technical reliability (and robustness), a person's trust cannot be engineered, calculated, verified, manufactured, or produced. It does not even have a definitive status. Trust is a special relation between two free beings, possessing its own rules rather than adhering to formal laws. It always remains vague and uncertain and is arguably never robust by any means.

The Guidelines acknowledge in their glossary that "trust is usually not a property ascribed to machines" (p. 38). However, the authors of the Guidelines evidently have no problem with calling AI systems potentially trustworthy; furthermore, they emphasize the need to ensure that these systems – along with all associated individuals and processes – are trustworthy.

As Kaminski (in this volume) emphasize, genuine trust in technological artifacts has always been critically dependent on the institutional frameworks and contexts that create, regulate, oversee, and assume responsibility for these technologies. Institutionalized trust emerges from mechanisms of accountability, transparency, certification, liability structures, political accountability, and ongoing societal negotiation processes. Thus, when we speak of trustworthy AI, we are often referring to the trustworthiness of the institutional and human processes surrounding AI – those that develop, deploy, regulate, monitor, and take responsibility for AI systems (Ho and Caals, 2024: 368; Costa 2024: 37). To meaningfully speak of trustworthy AI, therefore, political and regulatory frameworks must move beyond mere technical standards.

2.2. A problem of expertise?

A possible explanation for this strong emphasis on technical reliability in the understanding of Trustworthy AI could lie in the composition of the expert group. To borrow an insight from the group itself: One potential problem might be the lack of diverse skills and competencies in the team developing ethical guidelines (cf. EG: 25, see also pp. 18, 23). Although "ethics" is in the title (and is described in its own glossary as a philosophical discipline), the "high level expert group" includes only three professional philosophers among its 52 members.[9] To be fair, It is indeed an expert group on AI. But why should an expert on AI also be an expert in AI ethics? Without questioning their expertise on AI, one can question their expertise in ethics and their awareness of the subtle distinctions and challenges inherent in human thought and social practices, including trust. For this subject, perhaps a different group composition would have been more appropriate.

Another indication of this issue is that the EG's definition of "Trust" in its glossary is simply taken "from the literature" (EG: 38) – a business paper by an information scientist and his student, rather than from any humanities scholar – and it does not substantially address ethical aspects. This borrowed definition, however, seems more appropriate than the conception of trust developed in the main body of the EG. Amusingly, the cited paper itself states: "Ethics and governance of artificial intelligence are areas that need more attention." (Siau and Wang 2018: 52) Ironically, subsequent attention was then drawn to the very paper that had called for it.

None of this is necessarily a problem, although it is, at a minimum, questionable. On the one hand, there is an issue of expertise; on the other, potentially stemming

9 These are Coeckelbergh, Floridi and Metzinger. You get six professional philosophers if you add one ethicist (Van Wynsberghe) and two legal scholars, who also work on Philosophy of law (Hilgendorf and Yeung).

from this, there may be a deficient understanding of the phenomenon of trust itself, particularly when it is reduced to reliability.

2.3 Structural Deficits of the AI Act: Insufficient Context Sensitivity and Top-Down Approach

These conceptual shortcomings, notably the inclination to equate trustworthiness with reliability (as discussed in Section 2.1), may also be reflected in the AI Act's structural approach, which struggles with the inherent context-dependency of trustworthiness (the focus of this section). The inherent context-dependency of trustworthiness challenges attempts to systematically enumerate conditions for "*trustworthy AI*". The AI Act contains a relatively rigid evaluative framework in which obligations significantly rely on a static risk classification (Laux et. al. 2024: 7; Nasr-Azadani 2024: 19). While this classification primarily addresses potential harms, this focus on harm prevention aligns only partially with the broader measures needed to establish trustworthiness. Consequently, this framework conflicts with the dynamic nature of trust, limiting adaptive, context-sensitive assessment (in essence, the measures dictated by the harm-centric classification may significantly exceed the requirements for trustworthiness in certain cases while potentially falling short in others).

Underlying this structural rigidity is the Act's predominantly top-down approach to defining and enforcing trustworthiness (exception in case of "*special trust domains*", Ho and Caals, 2024: 360). Specifically, the AI Act's legislative approach primarily views trustworthiness as a top-down condition to be imposed externally, based on adherence to prescribed criteria, rather than as a socially negotiated and collectively emergent phenomenon (Laux et al. 2024: 4, 7). While these criteria may not be explicitly intended as the sole guarantee of trust, this approach poses the risk of "*compliance theatre*" in practice – that is, focusing on actions that create the appearance of meeting regulatory requirements without necessarily achieving substantive goals, such as genuine trustworthiness (Costa 2024: 39). Companies deploying or providing AI might primarily view the regulatory requirements as a checklist to be completed, leading to the misleading assumption that formal compliance equates to being fully trustworthy. By emphasizing measurable criteria over process and potentially fostering this "*tick-box*" mentality, the AI Act may overlook or insufficiently recognize the social dimension of trust-building. Consequently, it risks omitting or even hindering processes essential for authentic trust formation.

3. The idea of trust

Following the critique of current approaches, it seems helpful to sketch, at least in broad strokes, the idea of trust. This perspective is offered not as a comprehensive trust theory but rather as an indication of what might be missing when "trustworthiness" in contexts like the EU AI Act and EG is predominantly framed in technical terms and limited to reliability.

Trusting always involves the freedom of another person and the inherent risk that accompanies it. When we trust, we, in a sense, expect the unexpected: We hold a positive expectation about an outcome that we ultimately cannot fully control, precisely because it relies on free action. Consider trusting a friend with a secret: We depend on their discretion despite their freedom to choose otherwise. Trust becomes essential precisely where direct control ends – whether because we are overwhelmed by the situation or because other persons are involved, whose actions are also free and therefore always pose a certain risk; one can never be sure what another person will do next. Freedom, ironically, is the prerequisite for trust and, at the same time, the reason it is needed in the first place. Speaking of trust is meaningful only in connection with freedom. Without a free being involved, there is no need for trust (and, in a strict sense, it is not even possible).

Beyond this lies the miracle of trust: that it can provide actual certainty regarding the unpredictable. "When I trust someone, his certainty of himself is for me the certainty of myself; I recognize my own Being-for-myself in him, I recognize that he recognizes my Being-for-myself and that it is for him purpose and essence." (Hegel 2018: 219). Trusting means that I am certain that my purpose and essence are also his. This marvellous connection of trust would require an extensive examination.

This idea contrasts sharply with our relations to traditional machines, which operate deterministically according to their design or programming. It makes little sense to speak of "trusting" them in an interpersonal way. This distinction arguably extends to so-called artificial intelligence, which is neither free nor genuinely intelligent; it operates statistically and simply performs calculations based on algorithms (Laux et al. 2024: 4). This fundamental characterization applies across the spectrum, from large language models to other types of machine learning models or systems, even if their specific capabilities and how we interact with them differ markedly.

Considering trust through this lens uncovers another paradox in the quest for "*trustworthy AI*": Systems are required to be more predictable and reliable (see "*Reliability and Reproducibility*," EG: 17) to be deemed trustworthy – a demand that runs counter to the acceptance of freedom and unpredictability inherent in the general idea of trust.

Perhaps the common shift in discourse from "trusting AI" to "trustworthy AI" implicitly acknowledges this contradiction. The question is then no longer: "*Can you trust AI?*" but is, instead, reversed: "*Is the AI at least worthy of trust?*" The concept of

trustworthiness avoids the critical difficulty of directly trusting an AI or other machine. Instead, it describes the relation between a free person and a non-free object, whose properties (such as reliability and safety – meaning that it can be used without posing a danger) are assessed.

In many cases, especially with complex AI systems, users or even regulators cannot independently verify the system's reliability or adherence to all specified requirements. Therefore, forming a judgment about an AI system's actual reliability or compliance usually necessitates placing trust in others – in the developers, certifiers, regulators, or third-party auditors who conduct assessments and provide assurances. This introduces an often-overlooked layer: Establishing trust in the AI system's operation is not merely a matter of verifying its properties but frequently requires prior trust in the institutions and individuals evaluating and overseeing that system's reliability and compliance. Trust is not a technical problem but a social issue.

This sketch also raises a concluding philosophical question regarding regulatory efforts in general: Can the quality that invites authentic trust truly be established or guaranteed through regulation and technical compliance alone? Or is such an endeavor inherently contradictory, risking a fundamental misunderstanding of trust itself? While ensuring AI systems are reliable and safe is undeniably crucial, this alternative perspective suggests that equating technical "trustworthiness" with a richer, freedom-based understanding of trust may be insufficient.

4. Final remarks

In any case, a deeper understanding is still needed of what kind of tool AI actually is, how to use it responsibly, and for what specific purposes. Simultaneously, we must elaborate on the circumstances under which its developers can be genuinely trusted or deemed trustworthy.

If central concepts remain abstract and disconnected from concrete contexts, there is a danger that they will degenerate into empty keywords or buzzwords in debates. This risk is heightened when such a keyword, like "trust", is prominently highlighted, as seen in the EU's AI Act. Unfortunately, even the best concepts can be misused (one might recall the French Revolution's descent into the Reign of Terror).

Drawing from the analysis in the preceding sections, the EU's approach to trustworthiness appears problematic. As discussed, the term "trust" seems either to lack a distinct, clearly articulated concept – leaving "trust" open to abuse, misleading practices, and contributions to chaotic, inefficient discussions – or to be implicitly limited to mere technical reliability (Laux et al. 2024: 4). This second possibility, potentially suggested by assessment lists implying that trust can be calculated or even

produced, is itself misleading (as indicated, for example, by the promotion of the assessment list).

Thus, both potential paths are problematic and challenge the EU's own ethical framework. Even if the concept of trust cannot be defined with absolute precision, it must be related to everyday practices and challenges. The analysis presented here underscores the need for a more profound engagement with the underlying conception of trust. The framework claims to be human-centric, yet the concept of trust it employs seems predominantly tech-centric, lacking social dimensions such as freedom and vulnerability that typically accompany issues of trust (Costa 2024: 31).

Understanding the intended concept of trust requires familiarity with the broader framework, not just the AI Act alone. While embedding AI in general frameworks is the right direction, further steps are clearly needed. To effectively cultivate trust and move beyond purely technical solutions, a comprehensive governance strategy is necessary. This strategy must move beyond the technical perspective of reliability (as discussed in Section 2.1) and purely regulatory top-down mandates (as discussed in Section 2.3) and incorporate the relational and social dimensions of trust outlined in Section 3. Such a strategy could incorporate elements such as institutionalizing mechanisms for transparent ethical deliberation to address complex, domain-specific value conflicts; establishing clear accountability structures through defined institutional responsibilities and robust liability regulations; or implementing continuous monitoring processes to track real-world performance and impacts, thereby enabling dynamic and adaptive trust. Despite such potential practical steps, the entire issue demands deeper conceptual thought – which, along with political considerations, is fundamentally a task for philosophy and law.

References

Costa, Maria I. (2024): "Building on the EU's Unique Strategy for Artificial Intelligence (AI). Can an Ethical Foundation Be Successfully Integrated into Its Design and Deployment?", in: UNIO – EU Law Journal, 10(1), pp. 30–41.

Hegel, G. W. F. (2018): The Phenomenology of Spirit, Translated by M. Inwood, Oxford University Press.

Ho, Calvin W. L. and Caals, Karel (2024): "How the EU AI Act Seeks to Establish an Epistemic Environment of Trust", in: Asian Bioethics Review, 14(1), pp. 345–372.

Laux, Johann, Wachter, Sandra and Mittelstadt, Brent (2024): "Trustworthy Artificial Intelligence and the European Union AI Act. On the Conflation of Trustworthiness and Acceptability of Risk", in: Regulation & Governance, 18(1), pp. 3–32.

Nasr-Azadani, Mohamad. M. and Chatelain, Jean. L. (2024): "The Journey to Trustworthy AI-Part 1. Pursuit of Pragmatic Frameworks, arXiv:2403.15457.

Ribeiro, Marco T., Singh, Sameer and Guestrin, Carlos (2016): "Why Should I Trust You? Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Siau, Keng and Wang, Weiyu (2018): "Building Trust in Artificial Intelligence, Machine Learning, and Robotics", in: Cutter Business Technology Journal, 31(7), pp. 6–13.

Legal Sources

Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (2020), https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, last access: July 21, 2025.

Commission Staff Working Document Executive Summary of the Impact Assessment Report (2021): Brussels, 21.4.2021, SWD(2021) 85 final, 2021/0106(COD). [Impact Assessment Executive Summary]

Commission staff working document impact assessment accompanying the 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 (2021), SWD/2021/84 final, Part 1. [Proposal]

Commission (2020), White Paper on Artificial Intelligence – A European approach to excellence and trust, Brussels, 19.02.2020, COM (2020) 65 final, pp. 1, 3. [White Paper]

High-Level Expert Group on Artificial Intelligence (2019): Ethics Guidelines for Trustworthy AI [EG]

Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (2020), https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, last access: July 21, 2025.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act)