

REPORT

Trustworthy AI through the lens of Digital Humanism

A theoretical and practical
perspective

Erich Prem, Klaus Staudacher,
Steven Umbrello, Daniela Vacek

August 2026

Introduction

Recent debate in AI ethics has said a lot about trustworthy AI, which also remains an important building block of core guidelines and regulations (Reinhardt 2023), most notably the European Commission's High-Level Expert Group's *Ethics Guidelines for Trustworthy AI*. Some of the views in the debate remain pessimistic about the prospects of AI being trustworthy (Bryson 2018, Metzinger 2019, Freiman 2023), at least in a sense any stronger than reliability (Ryan 2020). Some of the recent optimistic views disagree, defending the concept in its literal sense, or at least a sense that goes beyond mere reliability (Ryan 2025; Zanotti et al. 2024).

But is trustworthy AI a theoretical or a practical problem? These two perspectives are often intermixed. To illustrate, Zanotti et al. (2024, 2695) oscillate between conceptual questions (such as whether the notion of trustworthy AI is a categorical error or conceptual nonsense) that largely motivate the philosophical debate, and highly practical questions, such as what advantages the notion of trustworthy AI should bring and what role it plays in the *Ethics Guidelines for Trustworthy AI*. Freiman (2023, 1356) in turn attempts to switch the focus from the theoretical question to a very practical one, relying on the unlikelihood of abandoning the notion of trustworthy AI: "Trying to correct others who are unaware of the philosophical underpinning of the concept of 'trust', arguably amounts to a waste of limited resources. It is reasonable to say that such efforts would most likely be ignored". Freiman points out that the notion is already too ubiquitous to be abandoned. But does this make sense of the conceptual nonsense "trustworthy AI", as the title of his paper seems to be promising to achieve? It does not, in our opinion, despite the fact that his observations and practical recommendations might be worthwhile. But the two questions remain different ones, and while an answer to one can inform the answer to the other, perhaps via conceptual engineering, we believe that more clarity would be gained if they were first tackled separately.

Digital Humanism is both a theoretical position and an ethics for the digital age. As such, it possesses means for tackling both of the above perspectives. From a theoretical point of view, it rejects positions such as the mechanistic paradigm ("humans are machines") and the animistic paradigm ("machines are (like) humans") that question or

blur the line between humans and machines (for the rejection of these paradigms from the perspective of Digital Humanism, see Nida-Rümelin and Staudacher 2024, Prem 2024). From a practical point of view, it formulates a human-centred approach with principles that put people first by placing human dignity, democratic governance, social cohesion, and environmental sustainability at the centre of technological development.

In line with this, the present paper pulls these two perspectives apart, analysing the issue of trustworthy AI first as a *theoretical* problem. From the perspective of Digital Humanism, the resulting analysis can then be seen as a theoretical position. The paper reinterprets the resulting analysis as a *practical* position of Digital Humanism. This is to be understood as an ethical, or more broadly, a normative position. This distinction is crucial given that the conceptual question concerns whether AI systems can be trustable entities at all, whereas the practical question concerns which socio-technical arrangements Digital Humanism can endorse, regardless of whether “trustworthy AI” is taken literally or only indirectly.

The question of the trustability of digital technologies

From a theoretical perspective, Digital Humanism raises the following question with regard to the concept of trustworthy AI: Does the notion of “trustworthy AI” already represent an anthropomorphisation of AI (Nyholm 2020)? Or, to put it in another way: Are AI systems, by their very design, entities with which we as human *trustors* can have a relationship of trust in a conceptually coherent way? Can they, as they are constituted, even possess the necessary characteristics for being *trustees*, and being *trustable*? (on the concept of *trustability*, see Jannel & Tallant 2026).

Terms belonging to the word family “trust,” such as “to trust someone” or “trustworthiness,” are used in very different ways. One distinguishing feature is obviously who serves as the trustee; for example, and without claiming to be exhaustive: persons (‘In order to climb at the limit, a climber must have complete trust in their belay partner’), institutions and authorities (‘Alan has complete confidence in the independence of his country’s judiciary’), but also animals (‘The blind man trusted his dog completely’) and digital technologies (e.g. in: ‘I trust my new laptop much more than the previous model’ or in the aforementioned *Ethics Guidelines for Trustworthy AI*).

The various uses of the predicate *trust* – which have themselves been the subject of extensive philosophical analysis¹ – might initially appear to indicate that trust is not a purely interpersonal phenomenon and that, accordingly, not only persons but also other entities, including digital technologies, may serve as trustees without conceptual difficulty.

Yet irrespective of these different categories of putative trustees, and regardless of how the terms “to trust” and “trustworthy” are interpreted in such contexts, there are good reasons to maintain that the concept of trust is fundamentally rooted in an interpersonal

¹ See the references cited in McLeod (2023) for a brief list of philosophical contributions that do not regard trust as a purely interpersonal phenomenon.

relation – one in which both trustor and trustee are persons (see the *Internet Encyclopedia of Philosophy* entry on trust, Carter and Simion 2020, (1a)). At the very least, the dominant paradigm of trust remains interpersonal, and most authors working on the concept of trust would agree that only those forms of trust are coherent “that share important features of (i.e., can be modelled on) interpersonal trust” (see the *The Stanford Encyclopedia of Philosophy* entry on trust, McLeod 2023).

Accordingly, before assessing whether the notion of “trustworthy AI” is conceptually appropriate, it is necessary to clarify what it means to trust other human beings and what it is for a person to be regarded as trustworthy. That the dominant paradigm of trust is interpersonal does not, of course, in itself imply that relations of trust cannot also be extended to non-personal trustees. However, we argue that there are at least *some* uses of the concept of trust in which the trustor presupposes certain personal properties in the trustee, and that such uses of the trust predicate should therefore not be extended to AI systems as long as these systems lack the relevant personal properties.²

The philosophical discussion on the concept of interpersonal trust shows great differences in detail and can only be outlined here. On the first approximation, trust can be described as “an attitude or a combination of attitudes (e.g., optimism, hope, belief, etc.) toward a trust recipient, which involves a certain (not negligible) vulnerability to a possible breach of trust on the part of the trusting party” (see the *Internet Encyclopedia of Philosophy* entry on trust, Carter and Simion 2020, 1).

According to Baier the “difference between trusting others and merely relying on them [...] seems to be reliance on their good will toward one, as distinct from their dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives, not directed on one at all” (Baier 1986, 234);³ emotional reactions also differ when the person on whom one relied or in whom one trusted behaved differently than expected or

² We therefore do not aim, in what follows, to offer a comprehensive analysis of the concept of trust; rather, we restrict our discussion to those cases of an interpersonal trust-relation in which certain personal properties are ascribed to the trustee.

³ Budnik furthermore emphasises the *reciprocity* of a disinterested good will as a decisive feature of trust conceived as a temporally stable relationship (cf. Budnik 2023, 23). Kemmerling likewise stresses the reciprocity of the trust relation and also points to parallels with the relations of *friendship* and *love*. Just as there can be love “at first sight,” so too there can be trust “at first sight”; and just as friendship, trust too ‘seeks to be reciprocated’ (cf. Kemmerling 2011, 146 f.).

hoped for: in such a situation, in the case of mere reliance, one is merely disappointed, whereas in the case of trust a feeling of betrayal arises (Baier 1986).

Katherine Hawley (2014, 2019) also sees the feeling of betrayal as an appropriate reaction when the person one has trusted does not behave as hoped; unlike Baier, however, she identifies in the trustee not good will, but rather commitment as the relevant distinguishing feature of trust as opposed to mere reliance.

Whatever position one takes here, there is some evidence to suggest that the reason why person P1 behaves in the way hoped for by another person P2 is relevant to whether we say that P2 (merely) relies on or trusts P1 to behave in this way. For example, it may be that one does not consider a contractual partner to be particularly trustworthy in terms of contract compliance but nevertheless relies on the contractual partner to comply with the contract in view of the penalty they would otherwise face (see Kemmerling 2011, 134). In this example, there is at least a divergence between the degree to which one trusts the contractual partner on the one hand and the degree to which one relies on them on the other.

This type of divergence points to yet another characteristic of trust, namely that it seems to inherently involve “a certain degree of *epistemic uncertainty*” (Budnik 2023, 47; translated by the authors). One could also put it this way: trust only becomes necessary when I cannot be sure that the other person will behave in the way I hope they will. Despite this, it remains controversial whether epistemic uncertainty is a necessary condition for trust.

The characteristic of epistemic uncertainty is also related to a specific quality that we attribute to someone, at least implicitly, by trusting them, namely the quality of *agency*, i.e. the capacity to act.⁴ This connection between trust, epistemic uncertainty and agency will be briefly discussed now.⁵

Actions in the sense relevant for the present paper differ from mere behaviour in several ways. In addition to a purely spatial-temporal behavioural component, such actions are characterised by intentionality: If, during a bus ride, the passenger P1 loses her balance

⁴ Cf. Schlosser 2019: “In very general terms, an agent is a being with the capacity to act, and ‘agency’ denotes the exercise or manifestation of this capacity.”

⁵ In this and subsequent paragraph, we are following Nida-Rümelin & Staudacher 2024, 17-30.

as a result of emergency braking in such a way that she falls on passenger P2 and the latter is injured as a result, this is described and evaluated differently than if P1 drops on P2 and P2 suffers the same kind of injury; it's only in the second case that we attribute intentions to P1 and that we would call her role in the incident an action; in the first case, on the other hand, we would say it was an unintentional, involuntary behaviour not at all guided by her intentions.⁶ Another property of such actions is that they are reason-guided, i.e., that the acting person always has a reason or reasons for his action;⁷ actions are constituted by reasons; not necessarily good reasons, but they are never performed without reason. And it is because of their being constituted by reasons that actions are, one could say, "capable of rationality"; for how reasonable or unreasonable an action is depends on how good or bad the reasons are that speak for or against it, how reasonable or unreasonable an action is depends on how good or bad the reasons for or against carrying it out are. Despite this, we leave it open whether all actions have these features. For instance, one can object that intentional action is distinct from action *simpliciter* (Hyman 2015) or list examples of legitimate mundane actions that are completely mindless (Arpaly 2003).

Still, for the purposes of this exploration, a thin notion of action is insufficient for trustability. Even if one grants that some actions need not be fully reason-guided in order to count as actions, it does not follow that every agent capable of action in this thin sense is a suitable object of trust. Trustworthiness appears to presuppose not merely the capacity to do something intentionally, but a stronger capacity for reflective control over one's conduct. In this respect, entities such as Frankfurtian wantons may still count as agents in a minimal sense, yet they are not paradigm candidates for trustworthiness, precisely because they lack the kind of self-governing control that makes it appropriate to hold someone answerable in the robust interpersonal sense relevant here.

The reasons that guide our actions are typically the result of a (sometimes very short) deliberation process, in which the different reasons are weighed up against each other and which, when it is

⁶ When, in everyday life, the term "behaviour" is used to describe actions (e.g., in formulations such as "Explain your strange behaviour from last night!"), it – correctly – refers to intentional behaviour.

⁷ As a rule, the acting person can also state the reason when asked. Even if the reason(s) should have slipped his mind –e.g., due to a loss of memory as a result of an accident – she had this/these reason(s) at the time of the act.

completed (and only then!), leads to a decision which is then realised by an action. In short, therefore, we can say: “No action without decision” (Nida-Rümelin 2005). Bernard Lonergan’s formulation is helpful here. He contends that possible courses of action are grasped via practical insight, motivated by reflection, and executed by decision, and, as such, the measure of the freedom with which such an act occurs is also the measure of the subject’s responsibility for it (Lonergan 1992). A decision whose content is already determined before it is made is just not a decision!⁸ If that is right, then trustability presupposes not merely behaviour, and not even mere freedom of action, but a form of agency sufficiently integrated with reflection, decision, and responsibility.⁹

The rationality of actions and their freedom, based on the openness of the preceding process of consideration and decision-making, are precisely the characteristics that make us responsible for our actions (we do not thereby deny that there are different, more capacious notions of responsibility).¹⁰

It is true that we can often predict another person’s actions quite well (e.g., because we know them well and can therefore estimate how they will behave in certain situations, or because there are typical actions for certain situations that most people perform regularly); but this is by no means always the case!

If actions are free in the sense that they arise from prior decisions without thereby determining how a person will behave in a given situation, epistemic uncertainty may arise concerning how that person

⁸ These remarks on the freedom of decisions and actions are not intended as a refutation of the deterministic thesis that everything that happens in the world is predetermined in such a way that all our decisions and actions are not free. The aim here is merely to clarify what is meant by decisions and actions in purely linguistic terms (and in the sense relevant for the present paper). If the deterministic thesis were true (which is by no means certain), then, according to the present conception of decisions and actions, strictly speaking there would be no decisions and actions, and our self-assessment that we as humans can make decisions and perform actions would be false.

⁹ This is also why mere freedom of action is too weak a criterion for trustability. If freedom is reduced to the absence of external impediment, then compulsive or wanton conduct may count as free in a thin sense, while still lacking the reflective control needed for robust trustworthiness. Compare Lonergan’s distinction between essential and effective freedom, where freedom is linked to practical insight, reflection, and decision (see Lonergan 1992, 632–646).

¹⁰ The conceptual connection between responsibility, freedom, and reason (as understood in the present paper) can be formulated as follows: Because or insofar as we are rational, i.e., we have the capacity for deliberation, we are, by exercising this capacity to deliberate, free, and only because and to the extent that we are free, we can be responsible (see for this connection in detail: Nida-Rümelin & Staudacher 2024 19–23).

will act.¹¹ This uncertainty does not always occur, however. In many cases we can predict another person's behaviour with reasonable reliability – for instance because we know the person well or because certain situations tend to elicit characteristic responses from most people.

Notably, we treat agency-based epistemic uncertainty concerning human behaviour differently from (to some extent) structurally similar uncertainty concerning machines. In the former case, trust may be justified. Whether it is justified will typically depend on features of the trusted person such as goodwill, a sense of commitment, or other attitudes relating to the welfare or interests of the trusting party. In the case of machines, by contrast, such considerations appear inapplicable.

This asymmetry becomes clear when we ask: “Why do you trust that person P will act in accordance with your trust?” and “Why do you trust that machine M will do so?” In the first case, the answer will plausibly appeal to attitudes attributed to P that justify the trust. In the second case, however, only statistical predictions about the machine's behaviour are available. Even where such predictions exist, they merely describe the uncertainty involved – indeed, they presuppose it! – and therefore do not justify trusting the machine despite that uncertainty. For this reason, the very idea of trusting a machine under such conditions already appears misplaced.

Within this framework, the relationship between trust and agency can be understood as follows: trusting someone to act in accordance with trust entails relying on the expectation that they will behave from an attitude constitutive of the trust relationship (depending on the theoretical account, for example an attitude of goodwill or of commitment), *even though*, as agents, they could act otherwise. In a word, the relevant agency that we are exploring here is not merely the capacity to produce outputs or even intentional performances, but more precisely the capacity to answer for one's conduct under conditions of reflective self-control.

¹¹ Within this uncertainty, the following three categories can be distinguished with respect to a person's behaviour: (i) risk (only probabilistic assumptions can be made regarding the possible courses of action available in a given situation); (ii) ignorance (the possible courses of action available to the person in a situation are known, but not even rough probability assumptions can be made about them); and (iii) complete ignorance (it is not even clear which courses of action the person is considering).

How, against the background of this characterization of the relationship of trust between persons, should it be interpreted when people say that they trust an AI system? In our view, the following interpretations can be distinguished. First of all, we can read the phrase “trustworthy AI” indirectly: AI is trustworthy iff the stakeholders involved in developing, manufacturing, distributing, deploying, owning or using AI are trustworthy (this parallels indirect readings of “responsible AI”, cf. Vacek 2026). In this usage of the predicate “trustworthy,” AI systems are not ascribed any personal properties. Rather, human trustors use the term ‘trust’ in relation to AI systems metaphorically and/or for reasons of “linguistic economy” (i.e., to avoid cumbersome formulations such as “I don’t trust the AI system, but rather the designers”, and the like). In other words, they do not really trust the AI system and would confirm this if asked. This use of the term “trust” is therefore not problematic from the perspective of Digital Humanism. Digital Humanism as a theoretical position of trustworthy AI can thus well be spelled this way in a coherent manner.

Secondly, we can read the phrase “trustworthy AI” directly: AI is trustworthy iff it is fitting to trust AI, where trust in question is identical to the human-human relation of trust (this parallels accounts of “responsible AI” that allow for responsibility of AI systems, such as Sullins 2011 or Altehenger et al 2024). On this interpretation, human trustors misattribute personal properties such as (intentional, reason-guided) agency to AI systems. With regard to this erroneous attribution of personal properties to AI, three case scenarios can be distinguished.

In the first case, the error is based on a lack of knowledge about how algorithmically functioning AI works. And if the trustors were *informed* that algorithmically functioning AI cannot have personal properties, they would no longer say that they trust an AI system or that there can be such thing as trustworthy AI. Thus, the “information test” changes the language used by the “trustors” and thus refutes the thesis that they actually trust the AI system in a personal sense.

In the second case, the trustors have complete knowledge of how AI works (well, at least to the extent that is standardly possible; we do not deny the “black box” nature of many current AI systems), but that does not hinder them from attributing personal qualities to AI, because they think what is called personal qualities in human beings can completely be translated in a reductionist way to a purely mechanist account of human beings; if asked, they would argue that, since human beings – despite being nothing more than complex machines –

possess personal properties such as (intentional and reason-guided) agency, other complex machines, for instance AI systems, may likewise exhibit such personal properties.

In such a use of the concept of trust, we are therefore dealing with trustors who adhere to the mechanistic paradigm mentioned above. Since this usage is motivated by the mechanistic paradigm, it is, from the perspective of Digital Humanism, just as inadequate as the paradigm itself. Furthermore, if the mechanistic paradigm were correct, human beings would not genuinely possess personal properties such as (intentional, reason-guided) agency; instead, such properties would be merely illusory.

In the third case as well, human trustors attribute personal properties to AI despite having full (or at least extensive) knowledge of how AI systems function. Unlike in the second case, however, they remain agnostic about the question of whether human beings themselves might ultimately be nothing more than complex machines. On the basis of a functionalist or emergentist approach, they nevertheless assume that even present-day AI systems possess personal properties despite their purely mechanical character. In this respect, they ultimately represent proponents of the animistic paradigm mentioned above. Since this use of the concept of trust is motivated by the animistic paradigm, it is – like the paradigm itself – equally inadequate from the perspective of Digital Humanism.

However, Digital Humanism is not a speciesist position. It therefore does not in principle exclude the possibility that, in the future, AI systems might exist that possess the personal properties required for trustability. That conceptual openness, however, does not affect the assessment of contemporary AI systems, which do not exhibit reflective self-awareness, intelligent inquiry, rational judgment, or moral deliberation in the relevant sense.

Finally, apart from the obvious possibility of seeing “trustworthy AI” (much like, on many occasions, “responsible AI”) as a mere check-box exercise that does not tell us much about the nature of trust(worthiness) and includes an ad hoc list of “good things” we would like AI systems to satisfy, there is one further alternative, a direct but weak reading of trustworthy AI: AI is trustworthy iff it is fitting to trust AI, where trust is construed as a human-AI relation distinct from the human-human relation of trust (much like direct weak reading of AI role responsibility in Vacek 2026). On this interpretation, the trust relation is no longer interpersonal; however, no personal

properties are ascribed to AI systems. This use of the concept of trust is therefore unproblematic from the perspective of Digital Humanism. If the trusters were informed that the AI system does not have personal properties, this would not change their use of language in relation to AI systems. The “information test”, therefore, does not change the trustors use of language, but rather confirms that they use a new/different concept of trust. We note that this reading leaves many important questions open and would have to be properly analysed to see it in full detail (this, however, is beyond the scope of the present paper).

In summary, aside from the last reading, linguistic usage does not support the claim that the concept of trust has undergone a conceptual shift that would allow AI systems to be regarded as trustable entities independently of whether they possess personal properties.¹² When AI systems are nevertheless treated as trustable entities, this is because personal properties are attributed to them. Such attributions either rest on a corrigible error resulting from a misunderstanding of the algorithmic functioning of AI, or they arise from the adoption of a mechanistic or animistic paradigm – both of which are rejected by Digital Humanism.

By contrast, the uses of the notion of “trust” in readings that do not involve the attribution of personal properties to the trustee, are unproblematic from the perspective of Digital Humanism.

Finally, it should be emphasised that expressions of trust in contemporary AI systems do not justify the conclusion that such systems possess personal properties. Positive findings concerning such expressions of trust should therefore not lead us to attribute personal properties to AI systems – let alone to deploy them in roles that presuppose such properties.

¹² We do not claim that this finding extends to other non-human trustee entities. However, it seems plausible that similar categories of readings could also be formulated for such entities.

Trust and Digital Humanism: the ethical position

As the considerations in the previous section have shown, AI systems (and digital technologies more generally) cannot themselves be *trustworthy* (in the sense humans can), since they lack trustability; rather, only the stakeholders involved in developing, manufacturing, distributing, deploying, owning, or using AI can properly be regarded as trustable and therefore potentially as trustworthy. This raises the question of what ethical significance the notion of *trust* can have at all from the perspective of Digital Humanism, and whether *trustworthiness* in relation to AI systems should instead be interpreted as *reliability*, or whether the concept of “Trustworthy AI” ought to be replaced by that of “Reliable AI”.

However, it must be borne in mind that in technology – and in computing in particular – *reliability* is a technical term typically defined in a purpose-relative manner as “the probability that a system, component, or item will perform its intended function without failure for a specified period of time under stated operating conditions. It is a quantitative measure of dependability, focusing on failure-free operation within defined parameters”¹³. Under this interpretation, a conception of *trustworthy AI* understood as *reliable AI* would amount merely to a concept expressing the suitability of an AI system with respect to the purpose for which it was designed.

Yet the purposes in question may be of very different kinds and may themselves be ethically highly problematic. In this purely functional sense, for example, even autonomous weapon systems – such as so-called “killer robots” – could be regarded as trustworthy if they reliably fulfil the objectives for which they were created. The goals pursued by such systems, however, are evidently difficult to reconcile with the human-centred approach characteristic of Digital Humanism. More generally, the same difficulty arises with regard to the “ethical blindness” of this reliability-based conception of trustworthiness, which asks exclusively whether a technology fulfils the purposes for

¹³ Cf. What is Reliability? <https://asq.org/quality-resources/reliability>

which it was designed, regardless of what those purposes may in fact be.

Since digital technologies themselves cannot be trustworthy due to their lack of trustability, and since interpreting trustworthiness merely as reliability results in an ethically blind and therefore questionable conception, *trustworthiness* should not, from the perspective of Digital Humanism, be regarded as a decisive concept for the ethical evaluation of AI systems or digital technologies.

This finding, however, does not imply that, from the perspective of Digital Humanism, *trust* with regard to digital technologies is irrelevant. Rather, the interpretation referred to as the indirect reading of “trustworthy AI” does possess significance (for the time being, we leave behind the second theoretical alternative, namely the option of human-AI trust, which remains conceptually too underdeveloped to offer clear path forward in developing Digital Humanism as a practical position on trustworthy AI). From the standpoint of Digital Humanism, digital technologies must be developed, manufactured, distributed, deployed and used in such a way that trust in the actors responsible for them as stakeholders does not lead to violations of the core principles of Digital Humanism.

These principles concern the *impact of technology on human beings and their co-evolution with technological systems*; they include the *mandate that technology must protect both people and the environment, the demand that technology should strengthen democracy and society, and the assertion that technologies are malleable* (that is, that the use and application of digital technology can be shaped and directed, and that responsibility ultimately remains with human beings rather than with machines); see Prem 2024.

In what follows, we aim to show that blind trust in the stakeholders of digital technologies can lead to violations of these principles and that, for this reason, vigilance is required. The point is not that the principles of Digital Humanism themselves define trustability. Rather, once the conceptual question of trustworthiness has been clarified, these principles specify the conditions under which trust in the relevant human stakeholders is normatively appropriate and, thus, able to be operationalised. Digital Humanism, therefore, shifts the evaluative focus away from the question of whether a technology is itself trustworthy and toward the question of whether the socio-technical arrangement surrounding it is worthy of justified trust. In this respect, trust is not an end in itself but a relational and institutional

achievement that must remain subordinated to the values we have been referencing, in particular, to human dignity, democratic legitimacy, as well as responsibility.

First, if technology must protect persons and the environment, then trust in those who design and deploy digital systems is appropriate only where such trust does not expose users or affected communities to foreseeable harms that could have been mitigated. In this sense, blind trust is objectionable not simply because it may turn out to be disappointed, but instead given that it may facilitate forms of manipulation/deception, dependency, opacity, or environmental externalisation that are incompatible with human dignity and self-determination. What matters here is not whether the system performs its intended function efficiently. The crucial dimension is whether the actors responsible for it have designed, deployed, monitored, and governed it in ways that protect the interests and standing of those affected by it.

Secondly, if technology should strengthen democracy and society, then trust must not be arranged in ways that displace public accountability or normalise the outsourcing of human judgment to artefacts. Excessive trust in digital systems can weaken democratic culture when it obscures the human decisions sedimented in design, substitutes human reasonable judgment, diffuses responsibility across opaque technical infrastructures, or encourages acceptance of outputs that ought instead to remain contestable. From a Digital Humanist perspective, the central question is therefore not how to maximise trust in AI systems, but more comprehensively how to organise trustworthy institutional relations among developers, deployers, regulators, and users such that contestability and public oversight remain intact.

Thirdly, the claim that technologies are malleable has a direct practical implication on patterns of warranted trust and misplaced trust. These are not fixed properties of artefacts but outcomes of design and governance choices. This is where Value Sensitive Design becomes useful, provided it is interpreted correctly. On a strong Digital Humanist reading, VSD should not be understood as attributing ethical agency to artefacts or as “building morality into” machines. Instead, the stronger reading offers a higher-resolution perspective, best understood as an accountability practice for human stakeholders: identifying direct and indirect stakeholders, eliciting and ranking the relevant values, translating those values into concrete design requirements, and revising systems over time as contexts and impacts

change. In that sense, VSD operationalises the indirect reading of trustworthy AI, because it keeps responsibility with human actors while providing a structured way to shape socio-technical systems so that reliance becomes more appropriate and less hazardous. Concretely, this means moving from abstract commitments such as dignity, fairness, transparency, or accessibility to participatory co-design, explicit design requirements, stakeholder procedures, and post-deployment monitoring practices.

Conclusion

This paper has argued that “trustworthy AI” must be disentangled into a theoretical and a practical question. The theoretical question concerns whether AI systems can be trustable entities in the robust interpersonal sense. From the perspective of Digital Humanism, contemporary AI systems do not satisfy the relevant conditions, because trustability presupposes a kind of agency integrated with reflection, decision, and responsibility that current AI does not possess. The practical question concerns how digital technologies ought to be designed, deployed, audited, governed, and used. Here, Digital Humanism does not recommend abandoning the language of trust altogether, but it does relocate its proper object: not the artefact itself, but the human and institutional actors responsible for the socio-technical arrangement. For this reason, trustworthiness should not be treated as the decisive concept for the ethical evaluation of digital technologies. More fundamental are the demands that technologies protect persons and the environment, strengthen democracy and society, and remain open to responsible human shaping. Under those conditions, the most defensible sense of “trustworthy AI” is an indirect one, oriented toward accountable human governance rather than anthropomorphic trust in machines.

Acknowledgements

We gratefully thank Julian Nida-Rümelin for authorising the publication of this report, which will appear in shortened form in *Philosophical foundations and implications of digital humanism* (forthcoming).

Prem E., Staudacher K., Umbrello S., Vacek D. (2026/2027). The concept of "Trustworthy AI" – A Case Study of Problematic Anthropomorphization? In: Nida-Rümelin J., Weidenfeld N. (to appear) *Philosophical foundations and implications of digital humanism*.

References

- Altehenger, H., Menges, L., & Schulte, P. (2024). How AI systems can be blameworthy. *Philosophia*, 52(4), 1083–1106.
- Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. New York: Oxford University Press.
- Baier, A. (1986). Trust and Antitrust. *Ethics*, 96(2), 235.
<https://doi.org/10.1086/292745>
- Bryson, J. J. (2018). *AI & Global Governance: No One Should Trust AI*. November 13, 2018, United Nations University, Centre for Policy Research. <https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html>
- Budnik, Ch. (2023). *Vertrauensbeziehungen. Normativität und Dynamik eines interpersonalen Phänomens*, Berlin.
- Carter, J. A., & Simion, M. (2020). The ethics and epistemology of trust. *Internet Encyclopedia of Philosophy*.
<https://iep.utm.edu/trust/>
- Chen, Y., Wen, B., & Zulkernine, F. (2025). A Multiagent Summarization and Auto-Evaluation Framework for Medical Text: Development and Evaluation Study. *Jmir Ai*, 4, e75932–e75932.
<https://doi.org/10.2196/75932>
- Eggmann, F., Weiger, R., Zitzmann, N. U., & Blatz, M. B. (2023). Implications of large language models such as ChatGPT for dental medicine. *Journal of Esthetic and Restorative Dentistry*, 35(7), 1098–1102. <https://doi.org/10.1111/jerd.13046>
- Freiman, O. (2023). Making sense of the conceptual nonsense "trustworthy AI". *AI and Ethics*, 3(4), 1351–1360.
- Friedman, B. & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA: MIT Press.
- Hawley, K. (2014). Trust, Distrust and Commitment. *Noûs* 48 (1), 1–20.
- Hawley, K. (2019). *How to Be Trustworthy*. Oxford University Press, USA.

- Hyman, J. (2015). *Action, knowledge, and will*. Oxford: Oxford University Press.
- Jannel, R., & Tallant, J. (2026). Trustability and trustworthiness: conceptual foundations and the case of AI. *AI Ethics* 6, 13. <https://doi.org/10.1007/s43681-025-00839-w>
- Kemmerling, A. (2011). Vertrauen und Verlass. In: *Sonntag, Karlheinz, Vertrauen*, Heidelberg, 129–148.
- Lonergan, B. (1992). *Insight: A study of human understanding* (F. E. Crowe & R. M. Doran, Eds.; 5th ed., rev. and augmented; Collected Works of Bernard Lonergan, Vol. 3). University of Toronto Press.
- Metzinger, T. (2019). *Ethics washing made in Europe*. Der Tagesspiegel. <https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html>
- McLeod, C. (2023). Trust. *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/trust/>
- Natale, S. (2021). *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*. Oxford: Oxford University Press.
- Nida-Rümelin, J. (2005). *Über menschliche Freiheit*, Stuttgart: Reclam, 45–60.
- Nida-Rümelin, J., & Staudacher, K. (2024). Philosophical Foundations of Digital Humanism. In *Introduction to Digital Humanism: A Textbook* (pp. 17–30). Cham: Springer Nature Switzerland.
- Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. London: Rowman and Littlefield.
- Prem, E. (2024). Principles of digital humanism: A critical post-humanist view. *Journal of Responsible Technology*, 17, 100075.
- Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics*, 3(3), 735–744.
- Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5), 2749–2767.
- Ryan, S. (2025). Is Trustworthy AI a Conceptual Nonsense?. In *Artificial Intelligence and the Future of Human Relations: Eastern*

and Western Perspectives (pp. 75–90). Singapore: Springer Nature.

Schlosser, M. (2019). Agency. *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/entries/agency>

Sullins, J. P. (2011). When is a robot a moral agent. *Machine ethics*, 6(2001), 151–161.

Umbrello, S. (2024). *Technology Ethics: Responsible Innovation and Design Strategies*. Cambridge, UK: Polity

Vacek, D. (2026). Responsible artificial intelligence?. *AI & Society* 41, 43–55.

Zanotti, G., Petrolo, M., Chiffi, D., & Schiaffonati, V. (2024). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society*, 39(6), 2691–2702.

Colophon

TITLE

Trustworthy AI through the lens of Digital Humanism. A theoretical and practical perspective

AUTHORS

Erich Prem, Klaus Staudacher, Steven Umbrello, Daniela Vacek

TYPE

Report

PUBLISHER

The European Digital Humanism Initiative (EUDHIT)

DATE

August 2026

LICENSE

Creative Commons Attribution 4.0 International (CC BY 4.0)

CONTACT

eudhit.eu

office@eudhit.eu

VERSION

Version 1, 2026

DOI

10.5281/zenodo.22011874

CITATION RECOMMENDATION

Prem, E., Staudacher, K., Umbrello, S., Vacek, D. 2026. "Trustworthy AI through the lens of Digital Humanism. A theoretical and practical perspective". EUDHIT Report, August 2026. DOI: 10.5281/zenodo.22011874. <https://eudhit.eu>

About EUDHIT

The European Digital Humanism Initiative (EUDHIT) supports strengthening societies through Digital Humanism, shaping technology for democracy and inclusion and believes in human values at the core of digital futures.

eudhit.eu



**Funded by
the European Union**

The EUDHIT project is funded by the European Union's Horizon Europe research and innovation programme under the grant agreement ID 101212890 (EUDHIT). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.