Trust and Manipulation in Generative AI: A Digital Humanist Perspective*

Francesco Striano[1], Maria Zanzotto[1 2]

[1] University of Turin, Italy

[2] Northwest Italy Philosophy PhD Program – FINO, Italy

DOI 10.3217/978-3-99161-062-5-007, CC BY 4.0

Abstract. This paper explores the dynamics of trust and manipulation in generative AI systems, proposing digital humanism as a critical framework to re-evaluate our relationship with such technologies. We conceptualise trust as an evaluative act – a normative judgement about the trustworthiness of a system in a given context – and argue that trust in generative AI is structurally misguided. This is not because such systems lack moral agency, but because the trust placed in them has been uncritically extended from deterministic technologies, whereas generative models are probabilistic and non-linear. These systems should be approached not as 'truth-tellers', but as 'storytellers.' We further argue that deceptive features – such as their anthropomorphic linguistic style and confident rhetorical tone – exacerbate this misalignment, making users more vulnerable. Digital humanism offers a fruitful perspective for understanding these dynamics, encouraging us to engage with AI not as neutral tools, but as cultural artefacts that shape our values, behaviour, and epistemic practices.

1 Introduction

Trust in technology is often based on an implicit model: data are entered, a system processes them automatically, and delivers reliable results. This paradigm, which has emerged in the context of deterministic systems – where mechanisms are readable, behaviours predictable, and errors traceable – has been uncritically extended to more complex and inherently opaque technologies. Generative systems and especially large language models (LLMs) deviate significantly from this model. Their mode of operation is not aimed at verifying the truth of a statement, but at producing results that are

* The paper is the result of scientific discussion and collaboration between the authors, was conceived in a joint effort and revised together. For the purposes of identifying the parts, where required, it is specified that sections 1, 2, and 4 are to be attributed to Francesco Striano, while sections 3 and 5 to Maria Zanzotto.

statistically plausible, coherent in context, and rhetorically effective. It is about a shift from *truth production* to *story production*.

In this paper, we argue that the misplaced extension of trust to generative technologies distorts our understanding of how they work and reinforces specific epistemic and political vulnerabilities. Generative AI is capable of producing content that is often indistinguishable from that created by humans. This capability undermines users' epistemic agency – their ability to critically evaluate, contextualise, and validate information. As a result, the use of these systems can erode individual autonomy – especially when deployed in high-density communicative environments – and compromise the conditions for democratic deliberation, particularly in environments where political opinion formation is already characterised by opaque platform dynamics.

However, recognising the manipulative potential of generative AI also opens up space for critical reflection. Rather than advocating uncritical trust or categorical rejection, we argue for a situated engagement with these technologies – one that emphasises interpretive consciousness and reflexive interaction. This perspective is in line with the ethos of *digital humanism*, which sees technology not merely as a neutral tool, but as a cultural and ethical phenomenon embedded in, and formative of, human values.

Building on this framework, the paper is structured in three parts. First, we will examine the misplaced extension of trust to probabilistic technologies. In doing so, we will focus on how generative AI challenges notions of reliability, trust, and confidence by creating 'stories' rather than stating facts. Secondly, looking at the gap between what these technologies *seem to be doing* and what *they do* we will discuss deception and how it can influence the evaluative act of trust, with potential consequences for epistemic agency. Finally, we will apply the perspective of digital humanism to the promotion of digital literacy in order to encourage a more critical understanding and conscious interaction with these technologies and their outcomes.

2 From Predictability to Plausibility: A Conceptual Shift in Technological Trust

Trust in information circulating through digital platforms has long been based on a more fundamental trust in the technologies that mediate it. While digital systems have sometimes provoked scepticism or outright rejection – reminiscent of historical patterns of technophobia and resistance – the prevailing tendency, particularly in Western contexts, has been to regard them as reliable infrastructures. This perceived reliability has often served as the basis for a broader attribution of trustworthiness. To clarify what is at stake in this attribution, and to understand why this trust may no longer hold in the

context of generative AI, we begin by disentangling three key concepts: *reliability*, *trust*, and *confidence*[1].

A technologically mediated society depends on the functional autonomy of its components – whether human, mechanical, informational, or hybrid[2]. However, this autonomy is never absolute: it requires and is maintained by varying degrees of trust. As Mariarosaria Taddeo notes, 'a society in which there is no trust in doctors, teachers, or drivers' would require all individuals to invest significant resources in constant monitoring, diverting time and attention from their own tasks (Taddeo, 2017, p. 566). In this view, trust enables coordination without constant monitoring and ensures that complex systems function without falling into recursive control loops.

The question of how to define trust – especially in relation to artificial agents – has led to a broad and unsettled debate. In her work, Taddeo (2010) defines trust as a second-order property that characterises binary, goal-oriented relationships: a trustor chooses to pursue a given outcome through the capacity of a trustee who is perceived to be trustworthy. This perception transforms the relationship into one that is expected to be beneficial to the trustor. Such a model has the advantage of being easily applicable to artificial systems, especially if they are designed to fulfil delineated functions with measurable success criteria.

However, this definition is not without limitations. Firstly, it assumes a binary relationship that does not readily accommodate distributed forms of trust as we see in institutions, infrastructures, or socio-technical ecosystems. Second, the definition harbours the danger of circular reasoning: it states that trust is justified by the trustworthiness of the trustee, but the criteria by which this trustworthiness is determined remain under-defined. This ambiguity becomes particularly pressing in the case of technologies, that lack moral motivation or intentionality in the human sense.

To clarify this issue, it is useful to distinguish between trust, reliance, and confidence. Several scholars have argued that what is often referred to as 'trust' in technologies is rather a form of reliance (Blackburn, 2010; Thompson, 2018). According to de Fine Licht and Brülde (2021), reliance is a three-place relation in which an agent A relies on B to achieve an outcome C. Reliance can be voluntary or involuntary, and can be directed toward both persons and artefacts. For example, we may rely on a wristwatch to tell the right time – not because we attribute some form of moral agency to it (such as a commitment not to lie), but because we judge it to be mechanically sound and consistent with our previous experience.

---

[1] For this non-standard conception of trust and a broader discussion of it, see Striano (2024b).

[2] For a discussion on the possibility of artificial agents with proper agency, see among others, Calì (2023), Floridi (2023), Himma (2009), Striano (2024a), and Swanepoel (2021).

In contrast, trust involves a specific kind of agential reliance in which the trustor ascribes some form of normative responsibility to the trustee (de Fine Licht and Brülde, 2021). This moral dimension is what makes betrayal possible: one can be betrayed by a human, but not by a machine. As Baier (1986) notes, failed reliance results in disappointment, while betrayed trust leads to moral injury. On this basis, some authors argue that technologies cannot truly be trusted because they lack intentionality and moral interest (Deley and Dubois, 2020; Thompson, 2018). They claim that we trust the designers, developers, or institutions behind the technology. In this sense, the reliability of a device serves as a proxy for trust in its makers.

Yet this distinction, while analytically useful, does not fully capture the phenomenology of trust in technological environments – especially, as we will see, when it comes to our interactions with conversational agents, whose apparent responsiveness invites forms of trust that go beyond purely functional reliance.

As Shionoya (2001) suggests, trust can be better understood as an *evaluative act*: a judgement by one agent regarding whether another – human or not – is trustworthy under specific conditions. This view considers both interpersonal and socio-technical forms of trust and opens up the space to consider trust not just as a property, but as a dynamic practice. Shionoya also emphasises the role of confidence as a disposition to trust: a background state of openness or readiness that enables the evaluative act.

Building on this insight, we propose a tripartite scheme:

- Confidence is an underlying disposition that leads a trustor to make an evaluation act of trust;
- Trust is an evaluative act that judges on the trustworthiness of the trustee;
- Trustworthiness is the characteristic trait that the trustor considers the trustee (human or artefact) to have.

Inversely, we can describe reliance in these terms:

- Reliability is a characteristic disposition of a person or an artefact;
- Reliance is the evaluation act based on the observable reliability of a person or an artefact;
- Confidence is a disposition inspired by repeated judgments of reliance (and can, in turn, lead to trust).

This scheme allows us to explain how trust can be extended to artefacts without anthropomorphising them. An artefact that consistently exhibits reliable behaviour can become the object of an evaluative act of trust – not because it possesses a will or a moral agency, but because it is *invested* with trustworthiness by the user. In this sense, trust becomes a normative stance, not a descriptive attribution.

Ultimately, both trust and reliance are grounded in confidence, but they differ in their normative assumptions. Where reliance involves functional expectation, trust implies a moral orientation. However, while it is legitimate to invest artificial systems with a form of moral trust based on evaluative judgement, this investment becomes problematic when applied to generative models such as large language models (LLMs). The difficulty lies not in their artificial nature *per se*, but in the erroneous extension of a trust model derived from deterministic technologies to systems whose functioning is qualitatively different.

Much of our habitual reliance on digital technologies has been shaped by interaction with systems governed by linear, deterministic processes. These technologies typically implement procedures that could in principle be carried out by humans – albeit more slowly – through explicit rules, formal logic, or algorithmic deduction. In such cases, the output is the result of a controlled and interpretable transformation of the input data. In knowledge cultures strongly shaped by scientific rationalism, this has contributed to a general association between digital computation, correctness, and truth. Science and Technology Studies remind us that such association has always been mediated by delegation and black-boxing rather than direct access to underlying mechanisms (Latour, 1987; Star, 1999; Edwards, 2010). Yet in the case of digital systems, this delegation rests on a genuinely linear and reproducible architecture: input, process, and output can, at least in principle, be traced and verified. Our trust in these technologies, while socially mediated, is also supported by their reliability, i.e., their consistent performance within a rationalist paradigm of control and predictability.

However, the trust we place in LLMs often assumes that they belong to the same category of reliable and explainable systems. These models work according to different principles. While they are technically deterministic at the code and infrastructure level, their output is generated by probabilistic models trained on large data sets. Their architecture introduces contingency, reflexivity, and a certain degree of unpredictability into user interaction. They are not designed to produce verified truths or facts: while they can draw on factual data, the outputs they produce do not represent facts in a direct sense, but rather construct plausible *narratives* in response to prompts. These models are not optimised for the production of truth, but for the continuation of interaction – through responses that are syntactically coherent, semantically persuasive, and rhetorically open-ended.

While we describe LLMs as narrative producers, this should not be understood as an attribution of narrative *intentionality* or autonomous meaning-making. The narratives they produce emerge from statistical coherence rather than interpretive intent. Yet meaning can still arise in the interaction between the model's patterned coherence and the user's interpretive engagement. In this sense, LLMs do not generate meaning as such, but rather *afford* it – they offer discursive structures that invite human interpretation.

This interactive production of coherence, however, should not be mistaken for epistemic reliability or truth orientation. The meaning that emerges in human-machine exchanges remains contingent on interpretation, not verification. As such, LLMs do not simply answer, but *simulate* attitudes, positions, and modes of discourse. Their persuasive power often masks the lack of epistemic commitment. In this respect, they resemble what Harry Frankfurt famously called a *bullshitter*: a speaker who does not care whether what they say is true or false as long as it serves their purpose. Drawing on Frankfurt, Gorrieri (2024) identifies three criteria for bullshit: indifference to truth-values, lack of acknowledgement of this indifference, and an ulterior communicative goal.

All three criteria, Gorrieri argues, seem to be applicable to systems such as ChatGPT. First, the model has no internal mechanism for assessing the truth-value of its outputs: it merely predicts plausible token sequences. Second, while disclaimers such as 'ChatGPT may produce incorrect information' now appear in the user interface, they frame the issue as an error rather than a structural indifference to truth. Third, the system encourages continued engagement – it ends its outputs with follow-up questions or invitations for further elaboration – showing that it is optimised not for accuracy but for sustained interaction.

This behaviour has significant normative implications. Users often interpret syntactically fluent and rhetorically sophisticated output as epistemically reliable, a misconception that is reinforced by interface design and interaction dynamics. The problem is not that such systems deliberately lie, but that they simulate a truth-oriented discourse without any concern for truth. When design choices systematically encourage users to conflate plausibility with reliability and coherence with truth, trust is not only misplaced – it is structurally misguided. In this sense, even if artificial agents cannot morally 'betray' us, we can still speak of a betrayal of trust by design.

3 Trust as a relational concept: how deceiving human-like features of generative AI pose additional issues to trust

Beyond this structural misplacement, trust in generative AI also needs to be understood as a relational phenomenon. The way users engage with LLM-based chatbots – through natural language and human-like cues – introduces additional layers to the evaluative act of trust. These systems mimic human-like characteristics, which often leads users to anthropomorphise them, meaning an attribution of human capabilities or mental states that chatbot do not possess. It is in this gap between appearance and reality that the notion of deception arises. In this section, we will explore how these deceptively human-like features can have additional effects on trust and thus on epistemic agency, i.e., the control epistemic agents have on belief-formation and belief-revision processes (Schlosser, 2019).

When people talk about chatbots deceiving us, they are usually talking about chatbots taking over the world and destroying humanity. An example of the concern about the destruction of humanity by AI is a statement on the risk of extinction caused by AI signed by prominent figures in Silicon Valley, such as Sam Altman (CEO of OpenAI), Demis Hassabis (CEO of Google DeepMind), as well as many academics and other technology leaders including Bill Gates (Hinton et. al., 2023). This overconfident attitude towards the power of AI also emerges when it comes to the capacity of AI to deceive, which is usually treated as its ability to trick humans in the pursuit of its goals, by means that do not align with human values. This discrepancy in values raises the fear that AI would be able to resort to any means to achieve its goals, even to the destruction of humans if necessary. However, this is not the direction we want to take. In fact, it is useful to distinguish between two levels of AI: general AI (also known as AGI: Artificial General Intelligence) and strong AI vis-a-vis narrow AI and weak AI. Those who speak of AI taking over the world have in mind a vaguely specified technological entity that has the same cognitive capabilities as humans, such as the ability to form mental states (intention, sentience, etc.) – strong AI –, and an intelligence that enables it to perform any kind of task – general AI –, ultimately better than humans (superintelligence). This vision is closer to sci-fi than to computer science. What computer scientists have been able to develop so far, however, is narrow AI, i.e., specialised AI systems developed for specific tasks, and weak AI, a simulation of intelligence, rather than duplication. ChatGPT, the most widely used LLM-based chatbot, is an example of narrow and weak AI that has its basis in the discipline known as NLP (Natural Language Processing), i.e., the field of computer science that deals with the development of models and systems capable of producing or modifying human-like text or speech (in this category we find chatbots, autocorrects, translators, etc.). LLM-based chatbots are more sophisticated because their computational engines are the most powerful: they use deep learning to make connections. Of course, being a narrow or weak AI does not mean that the outputs are not good, but they are fundamentally different systems from AGI. LLMs are extremely powerful linguistic machines that, as said before, calculate the most probable sequence of words given the input prompt. There are no mental states of the machine at stake, only very sophisticated model architectures based on probability and trained on huge data sets.

So, when we talk about deception, we are talking about characteristics of the chatbots that can deceive, without assuming that the chatbots *want* or *intend* to deceive.

There is a great debate about whether we can call it manipulation or deception when they have no intention to deceive. At this point, a brief overview of the concepts of manipulation and deception is necessary.

In contrast to deception, which is limited to the epistemic level, manipulation retains a certain semantic connection to the psychological level, namely the idea of being steered in a certain direction (Cohen, 2023). The manipulator's will to achieve a certain result is

an important aspect of manipulation. Manipulation is action-guiding. Deception can, of course, be used for manipulation, but it remains at the level of belief. According to our definition, deception is the inducement of false beliefs. However, we exclude errors and mistakes that can lead to the formation of false beliefs from our analysis. In the literature on manipulation and deception with technology, it is instructive that the concept of *manipulation* is used when microtargeting is discussed, while the concept of *deception* is used in the case of interaction with social robots. The literature on microtargeting is not about arguing whether recommendation systems have intentions or not, but rather about attributing intentionality back to the humans involved and arguing that technology can be an aggravating factor (Jongepier and Klenk, 2022). In the social robotics literature, there are some efforts to change the definition of deception so that it does not require intentionality. For example, if a human interacting with a robot forms the false belief that the robot has emotions, we can say that the robot has deceived the human, even if the robot has no intentions. An interesting approach from the literature is the notion of banal deception (Natale, 2021), which acknowledges that all media are deceptive, but not in the classical sense of 'deliberate deception', but in a more functional sense that uses tactics to form false beliefs for a better experience with technology, be it a laptop, a social robot, or a voice assistant. Deception is inherent in media, but it is not a form of outright manipulation and it is instrumentally valuable[3].

We believe that we can extend the concept of banal deception to LLM-based chatbots. They exhibit some deceptive features, starting with the most notable ones: anthropomorphism and mimicry. LLM-based chatbots produce outcomes in natural language. This means that users can easily communicate without having to program. Users speak *as if* they were talking to a human, and the chatbot responds *as if* it were a human, i.e., it mimics the human way of speaking or writing. One way to do this is to start the sentences with mentalistic propositions like 'I believe,' 'I want,' or propositions that express an emotion, such as 'I'm sorry.' This has clear advantages because it is more user-friendly, but also some consequences. The most relevant is anthropomorphism: users project onto chatbots mental states they do not have, such as beliefs, intentions, and desires. Dennett's (1987) intentional stance can be very helpful in this respect. Humans transfer their intentions to others in order to understand and predict their

[3] The distinction between banal deception and manipulation is only roughly outlined here; we merely introduced a distinction between manipulation as action guiding and deception as epistemic distortion. It is acknowledged that there may be multiple layers and forms of overlap, and that there may be varied interpretations of manipulation being goal-oriented. A broad interpretation of the distinction could conceive human-likeness as an engagement lever aimed at engagement maximisation, and consequently regard this form of deception as a manipulative device. A narrow interpretation has the potential to exclude such general aims, with the focus instead being placed on specific goals, such as convincing an individual to perform a particular action. However, due to limitations in the available space, a more thorough exploration of this topic is not feasible.

behaviour, even if they do not have access to the mental states of other humans or animals.

The phenomena described by anthropomorphism and the intentional stance are not new, and it is relatively easy for humans to attribute these properties to systems that are not even sophisticated. The fact that LLM-based chatbots are now so sophisticated that they give the impression that 'they understand you' (although sometimes they do not, which is very frustrating) makes it even easier for users to automatically attribute mentalistic states and emotional abilities to them. Another important feature is the rhetorical style of these chatbots. They tend to answer assertively and sometimes in an overly self-assured or even patronising manner. They never seem to have doubts or insecurities (which they do not have either, but they also do not have a certainty or self-confidence that allows for self-assurance). Empirical studies are starting to emerge, although they are still mostly at the under-review or preprint stages. Preliminary findings suggest that anthropomorphism – especially when it showcases intelligence and expertise – can foster trust (Colombatto et al., 2025). However, empirical research highlights how anthropomorphism does not have a clear-cut path towards trust, for example, human-likeness can be not-determinant when not functional to the users' goals (Bouyzourn and Birch, 2025; Haresamudram et al., 2025) or it can improve connection but reduce trust when perceived as non-authentic or non-reliable, or non-credible (Cohn et al., 2024; Basoah et al., 2025; Wang et al., 2025), or it can also evoke unsettling emotions when coupled with hallucinatory or erratic outputs (Rapp et al., 2025). Importantly, perceived self-confidence plays an important role, as it seems humans tend to perceive AI as more self-assured than humans, even when they have an identical performance, because of a prior belief that AI is more accurate (Colombatto et al., 2025). Yet this projected assurance, either perceived or expressed through language, does not actually correspond to accuracy in responses.

Moreover, there are also issues at the interface level: as mentioned in the previous section, OpenAI's ChatGPT does not display a banner explaining what it is and briefly describing how to set our expectations. It merely warns that ChatGPT can make 'mistakes,' fuelling the false belief that its default outputs are 'correct' or 'true.'

These features make it clear that there is a dissonance between what the chatbot appears to be able to do and what it actually does. What interests us is the relationship between deception and trust in LLM-based chatbots.

In the first section we argued that the flaw in trusting LLM-based chatbots lies in the erroneous application of the same trust granted to previous technological systems without taking into account the substantial shift from linear to non-linear AI systems. What these systems appear to be and what they are play a major role in this mismatch. In the first section, this idea was articulated primarily in terms of them appearing to be 'truth-tellers' rather than 'storytellers.' However, we can go further: given the additional

deceptive properties discussed above, we would like to argue that these properties may have additional implications for how they influence the evaluative act of trust.

However, there are different ways to interact with LLM-based chatbots, whereby we can distinguish three main types:

1. *Interaction with chatbots*: When the style of language makes us believe that the machine has mental states, we tend to employ the same structure that we use to evaluate human trustworthiness (anthropomorphism). Here, the discursive tone can be important when it comes to assessing the reliability of the target: the more confidence we have, the more we assume we can trust the target. If the information is false but communicated confidently and we have no prior knowledge, there seems to be a pressure not to exercise our epistemic agency and check. And when it comes to revision of beliefs, chatbots tend to agree with users when the latter express dissatisfaction (what is dubbed as *sycophancy*), reducing the likelihood of revision.
2. *Interaction with chatbots integrated into proprietary apps or websites:* This case is similar to the one above, but is complicated by the additional trust in the brand itself. For example, if we believe that Amazon is a reliable service, it is possible that we transfer trust to the Amazon chatbot.
3. *Interaction with social media bots:* This latter case presents an additional problem, namely that of indistinguishability. If the chatbot is disguised as a personal profile and interacts like one, we are inclined to treat it like a human. However, if we know that these bots are difficult to recognise, we begin to question whether or not we are interacting with a person, or we doubt that we are really interacting with a person. This doubt is not so easy to dispel and can lead us to go the other way round and no longer trust what we see online and disengage from online communication and interaction altogether.

Case 1 seems to be a case of banal deception. Cases 2 and 3 utilise the same psychological mechanisms as in case 1, but can lead to manipulation: deception – in this case, human-likeness – can be used as a means to steer people in a certain direction. In the case of chatbots used in apps, this can be a way to persuade people to make further purchases. In the case of social media bots, they can be used to manipulate users' voting preferences.

Overall, the evaluative act of trust towards LLM-based chatbots is more similar to how we would trust another person than how we would trust a pocket calculator. This is because in the evaluation process we consider not only technical aspects, but also interactive aspects shaped by natural language (without taking into account that some

even report having formed a bond with the chatbot), we react to anthropomorphic cues that users recognise when interacting with chatbots.

This results in a paradox: on the one hand, as just mentioned, the evaluative act of trusting generative AI systems is influenced by the anthropomorphic features of chatbots, such as self-confident tone; on the other hand, the alleged reliability of these systems is inherited from our reliance on digital technologies, just because generative AIs are digital technologies. These two settings create a sort of hybrid target for trust that is both human-like (in the way we interact with chatbots) and artificial (in the way we consider chatbots as linear technologies and therefore reliable and infallible). We are dealing with quasi-subjects that we trust with a type of trust that we would place in human conversational agents, but expecting greater efficiency and reliability than we would expect from human conversational agents.

Now, vulnerability can be seen as a condition of lower levels of epistemic agency: when we have less control, we are more vulnerable. Taken together, these factors explain why interacting with LLM-based systems can make users particularly susceptible to epistemic vulnerability. However, we can distinguish between vertical vulnerabilities and horizontal vulnerabilities. The first refers to groups of people who are considered less epistemically equipped than the average. In the context of technology, children and the elderly are considered more vulnerable groups. However, with the emergence of LLM-based chatbots, a vulnerability has crystallised as a result of the sheer power and widespread use of this technology, and it is a horizontal one: anyone can be vulnerable at this stage of transition. In this context, digital humanism can be a good ally to get through this transitional phase with *more*, rather than less, epistemic agency.

4 Reframing Trust through Digital Humanism

As we have seen, the interactive nature of LLMs creates conditions in which users are particularly susceptible to false beliefs – often without realising it. This dynamic of deception results from design features that simulate human-like communication while concealing the actual limitations of the system. Rather than dismissing these systems outright, we argue that such interactions can serve as a critical lens through which we can re-examine the broader models of trust that we have extended to digital technologies in the past.

Recognising that LLMs are not reliable producers of truth forces us to confront the assumptions embedded in our previous reliance on media technologies. In many cases, these systems have been treated not merely as mediators of information, but as guarantors of epistemic authority. The dissonance introduced by LLMs helps to reveal the fragility of this assumption.

This realisation opens up the space for a broader reconfiguration of our relationship with technology. Trust in digital systems needs to be reframed as a situated and evaluative stance rather than a passive expectation. LLMs, by highlighting the gap between discursive coherence and epistemic accountability, can become cultural artefacts through which we reconsider how information is framed, believed, and acted upon.

Such a reorientation is in line with the ethical and epistemological priorities of digital humanism, which asks us to consider technologies not as neutral tools, but as embedded cultural forms that shape and reflect human values, behaviours, and vulnerabilities.

In current debates about digital humanism, two complementary but methodologically distinct approaches have emerged, both of which offer valuable insights for rethinking the nature of trust in technological environments. Although these approaches originate from different premises, they often converge on the need to reclaim a space for human agency, autonomy, and ethical reflection in the face of technological transformation.

The first approach, associated with initiatives such as the *Vienna Manifesto on Digital Humanism* (Werthner et al., 2022, pp. xi-xiv) and the DigHum network, is primarily normative in its orientation. It begins with the formulation of a set of human values – dignity, freedom, responsibility, justice – and then attempts to translate these values into principles for the design, governance, and evaluation of digital technologies. This model, which we might call *top-down digital humanism*, views technologies not as neutral tools, but as systems whose structure, use, and social embeddedness must be normatively evaluated. It emphasises the importance of public accountability, democratic control, and anticipatory ethical reflection in the development and deployment of digital infrastructures.

In contrast, a second approach, developed in the French tradition of *humanisme numérique*, pursues a *bottom-up hermeneutic method* inspired by the philological and historical practices of Renaissance humanism and the Vichian imperative to let doctrines emerge from the objects they study[4]. This orientation was first systematically articulated by Milad Doueihi (2011), who understood the digital not only as a technical substrate, but as a cultural transformation. Here, technologies are analysed as *cultural artefacts* that are an expression of a historical moment and a particular way of shaping the human. The focus is less on prescribing values from above, but rather on observing how digital artefacts – such as algorithms, interfaces, or LLMs – participate in shaping practices, discourses, and perceptions. In this view, ethical insights emerge from a careful reading of the ways in which technologies transform human life, language, and thought.

[4] See the Position Paper For a Critical Digital Humanism, https://www.yumpu.com/fr/document/read/68683985/2024-mai-positionpaper-hn-en&sa=D&source=docs&ust=1749135910609674&usg=AOvVaw29dm_rbGiJ80dVRm33MGjM.

Despite their methodological divergence, these two approaches are not antagonistic. In fact, they often converge on key issues – most notably the need for inter- and transdisciplinary collaboration, the recognition of the cultural embeddedness of technology, and the prioritisation of human well-being and the common good as central goals of digital transformation. Their complementary perspectives provide a robust framework for analysing technologies such as LLMs that resist simple categorisation as tools.

From a digital humanist point of view – whether top-down or bottom-up – LLMs must be understood not only as technological artefacts, but as cultural products. Or rather, as *techno-cultural artefacts* whose technical construction is inextricably linked to the cultural logics, epistemologies, and power structures they encode. This also entails a critical analysis of power relations inscribed in data production, model training, and platform governance, since the epistemic and economic asymmetries that underpin these domains shape who can speak, be heard, and be trusted in digital spaces. Trained on vast corpora of texts from different domains – filtered, pre-processed, and structured according to specific assumptions about language, knowledge, and relevance – such systems inevitably reflect and reproduce the values, biases, and exclusions inherent in the data they ingest and in the design choices of their creators.

Examining LLMs through a humanistic lens allows us to grasp their normative impact beyond their immediate functionality. These systems influence how knowledge is accessed, how authority is perceived, and how beliefs are formed and stabilised in digital environments. A digital humanist reading of LLMs draws attention to the aesthetic and rhetorical strategies with which these models construct coherence and simulate competence. It also prompts us to ask how such strategies affect users' sense of epistemic agency, autonomy, and interpretive responsibility.

In addition, humanistic disciplines such as philosophy, literary theory, history, and media studies provide tools to situate LLMs in a longer genealogy of knowledge mediation – from the invention of writing to the printing press, from encyclopaedias to search engines. They also encourage a critical examination of power structures and show how technological systems participate in the reproduction or contestation of institutional and discursive hegemonies.

This bottom-up orientation, which pays attention to the symbolic and cultural dimensions of technology, does not exclude a normative critique. On the contrary, it enables a form of *situated normativity* that is grounded in the lived experience of users and the concrete affordances of specific systems. In this respect, the hermeneutic approach converges with the more declarative ethics of the top-down model, especially when it comes to shared goals: the protection of human dignity, the prevention of epistemic harm, and the promotion of environments that favour critical thinking, deliberation, and meaningful participation.

In the context of LLMs, this convergence is particularly fruitful. These models challenge not only our understanding of language and meaning, but also our understanding of trust and knowledge. By analysing LLMs as cultural artefacts that speak in our language, mimic our rhetorical patterns, and mirror our cognitive biases (Vallor, 2024, pp. 48-49), digital humanism equips us with a vocabulary and methodology to critically engage with their implications. It helps us to move beyond the binary of uncritical acceptance or technophobic rejection to a mode of reflection that recognises both the risks and the heuristic value of these technologies.

Ultimately, both approaches to digital humanism call for a renewed cultural literacy – one that includes the ability to read technologies, decode their implicit assumptions, and articulate alternative imaginaries. In this sense, the study of LLMs becomes not only a technical endeavour, but also a philosophical and political one: an invitation to redefine what it means to understand, believe, and trust in the digital age.

A digital humanist approach to trust begins by reframing trust itself – not as a passive expectation or functional reliance, but as an evaluative act, a normative judgement about the trustworthiness of a system in a given context. As argued earlier, such a judgement presupposes an active investment capable of shaping or reinforcing a disposition of confidence. From this perspective, we need to go beyond an instrumental view of technology (which considers reliability as a property related to the success of a linear interaction) and question its systemic role within cultural, social, and political structures. Technologies need to be evaluated not only in terms of their functionality, but also in terms of how they organise interactions, distribute agency, and reproduce or challenge existing asymmetries.

This shift also requires a renewed commitment to digital literacy, understood not just as a set of technical skills but as a capacity for critical orientation. And digital literacy, in a digital humanist sense, cannot be reduced to a demand for transparency. As recent debates in ethics of technology have shown (Alloa, 2022; Alloa and Thomä, 2022; Carbone and Lingua, 2023), transparency often risks becoming a moral and political fetish, as an ideal of total informational openness that paradoxically obscures rather than clarifies the processes it seeks to reveal. Following Striano (2024b), what we need are not merely ‘transparent,’ but honest technologies: systems that make mediation perceptible and negotiable, rather than hidden behind the illusion of full visibility. A literacy grounded in honesty rather than transparency would cultivate interpretive awareness and civic responsibility, encouraging users and institutions alike to engage critically with the limits, biases, and opacities inherent in technological mediation.

However, digital literacy should not be conceived merely as an individual competence. While we argue for cultural literacy, we are aware of power asymmetries between users and companies providing the technological services; it is simply not a level playing field. To ask that only users be responsible would mean denying this reality. Hence, within a

digital humanist approach, we recognise cultural literacy also requires collective infrastructures of accountability and education, as well as public policies that foster critical engagement with AI systems. Strengthening individual epistemic agency must go hand in hand with institutional and civic responsibility.

In practical terms, such a literacy could take shape through interdisciplinary education that combines humanities and computer science, participatory design processes that include users in evaluating algorithmic affordances, and civic initiatives that promote the public understanding of mediation rather than the illusion of transparency. Digital humanism, in this sense, calls for an ecosystem of practices that cultivate not only technical skills but interpretive, ethical, and political sensibilities – an education for reading, designing, and governing technologies as cultural forms.

5 Conclusion

This paper aimed to explore how technology, especially generative AI systems, invites – even forces us – to rethink fundamental concepts such as trust. In this paper, we have argued for a more considered and aware way of interacting with AI systems – one that is guided by the principles of digital humanism.

In the first section, we looked at how trust, which has traditionally been extended to technologies that are deterministic and predictable, has been erroneously applied to generative AI systems, even though they are nonlinear. We follow Shionoya (2001) in arguing that trust should be understood as an evaluative act in which a trustor judges whether another, human or not, is trustworthy under certain conditions. This perspective treats trust not as a fixed property, but as a dynamic, normative practice that involves active judgement. According to this perspective, trust is based on the confidence of the trustor – an underlying disposition or willingness to trust. It allows individuals to ascribe trustworthiness to entities, including non-human systems, not because these systems inherently possess moral qualities, but because they consistently exhibit reliable behaviour. However, trust in artificial systems, such as generative AI, should be based on an evaluation of their peculiar performance and not on the mistaken assumption that they function like deterministic systems.

In the second section, we delved into the relational aspect of trust, focusing on how the human-like qualities of generative AI can be deceptive. We explored how, when interacting with AI, we tend to attribute mental states and intentions to these systems even though they do not have them. This tendency, combined with design choices that make AI sound self-confident or even ‘human,’ often leads us to blindly trust the technology and reduce our ability to critically evaluate the information it provides. As LLM-based chatbots are integrated into almost every app and social media, the risks of

deceptive design can extend to risks of manipulation, whether for the purpose of increasing sales or changing voting preferences.

In the third section, we turned to digital humanism as a conceptual framework for thinking about how we should approach AI. Rather than seeing AI merely as a tool, a human-like subject, or some sort of infallible superhuman intelligence, digital humanism asks us to consider it as part of a larger cultural and ethical landscape. This perspective encourages us to engage *with* these technologies not only functionally, but also critically, and to understand how they shape our values, our behaviour, and our ways of knowing.

Finally, we have emphasised the importance of philosophical enquiry to help us address the challenges that generative AI brings. In conclusion, rethinking trust through a digital humanist lens is a crucial step towards a more critical, ethically responsible, and socially engaged approach to technology.

References

Alloa, E. (ed.), This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor, Leuven University Press, Leuven 2022.

Alloa, E. and Thomä, D. (eds.) (2022), Transparency, Society and Subjectivity: Critical Perspectives, Palgrave Macmillan, London.

Baier, A. (1986), Trust and Antitrust, in 'Ethics', 96(2), pp. 231-260

Blackburn, S. (2010), Trust, cooperation, and human psychology, in Id. Practical Tortoise Raising and other philosophical essays, Oxford Academic, Oxford, pp. 90-108. https://doi.org/10.1093/acprof:oso/9780199548057.003.0006.

Basoah, J., Chechelnitsky, D., Long, T., Reinecke, K., Zerva, C., Zhou, K., Díaz, M. and Sap, M. (2025), Not Like Us, Hunty: Measuring Perceptions and Behavioral Effects of Minoritized Anthropomorphic Cues in LLMs, in 'Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency', pp. 710-745. https://doi.org/10.1145/3715275.3732045.

Bouyzourn, K. and Birch, A. (2025). What Shapes User Trust in ChatGPT? A Mixed-Methods Study of User Attributes, Trust Dimensions, Task Context, and Societal Perceptions among University Students, in 'ArXiv'. https://arxiv.org/abs/2507.05046v1.

Calì, C. (2023), Come ci cambia la tecnologia. L'Agency delle AI e la capacità cognitiva di prendere decisioni razionali, in 'S&F_scienzaefilosofia.it', 30, pp. 366-385.

Carbone, M. and Lingua, G. (2023), Toward an Anthropology of Screens: Showing and Hiding, Exposing and Protecting, Palgrave Macmillan, London.

Cohen, S. (2023), Are All Deceptions Manipulative or All Manipulations Deceptive?, in 'Journal of Ethics and Social Philosophy', 25(2). https://doi.org/10.26556/jesp.v25i2.1998.

Cohn, M., Mahima Pushkarna, Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z. and Heldreth, C. (2024), Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models, in 'Extended Abstracts of the CHI Conference on Human Factors in Computing Systems', pp. 1-15. https://doi.org/10.1145/3613905.3650818.

Colombatto, C., Birch, J. and Fleming, S. M. (2025), The influence of mental state attributions on trust in large language models, in 'Communications Psychology', 3(1). https://doi.org/10.1038/s44271-025-00262-1.

De Fine Licht, K. and Brülde, B. (2021), On defining 'Reliance' and 'Trust': purposes, conditions of adequacy, and new definitions, in 'Philosophia', 49(5), pp. 1981–2001. https://doi.org/10.1007/s11406-021-00339-1.

Deley, T. and Dubois, E. (2020), Assessing trust versus reliance for technology platforms by systematic literature review, in ‘Social Media + Society’, 6(2), pp. 1-8. https://doi.org/10.1177/2056305120913883.

Dennett, D. C. (1987), The intentional stance, The MIT Press, Cambridge (MA).

Département Humanisme Numérique – Collège des Bernardins, For a Critical Digital Humanism (Position Paper), https://www.yumpu.com/fr/document/read/68683985/2024-mai-positionpaper-hn-en&sa=D&source=docs&ust=1749135910609674&usg=AOvVaw29dm_rbGiJ80dVRm33MGjM.

Doueihi, M. (2011), Pour un humanisme numérique, Seuil, Paris.

Edwards, P. N. (2010), A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming, The MIT Press, Cambridge (MA).

Floridi, L. (2023), AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models, in ‘Philosophy & Technology’, 36(1). https://doi.org/10.1007/s13347-023-00621-y.

Gorrieri, L. (2024), Is ChatGPT full of bullshit?, in ‘Journal of Ethics and Emerging Technologies’, 34(1), pp. 1-16. https://doi.org/10.55613/jeet.v34i1.149.

Haresamudram, K., Van As, N. and Larsson, S. (2025), Tasks Over Traits: User Perception of Humanlike Features in Goal-Oriented Chatbots, in ‘International Journal of Human-Computer Interaction’, 41(21) pp. 13363–13381. https://doi.org/10.1080/10447318.2025.2470311.

Himma, K.E. (2008), Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?, in ‘Ethics and Information Technology’, 11(1), pp. 19-29. https://doi.org/10.1007/s10676-008-9167-5.

Hinton, G., Bengio, Y., Hassabis, D., Altman, S., Amodei, D., Song, D., Lieu, T., Gates, B., Zhang, Y.Q., Sutskever, I., et al. (2023, May 30), Statement on AI risk [open letter]. https://safe.ai/work/statement-on-ai-risk.

Jongepier, F. and Klenk, M. (2022), Online manipulation. Charting the field, in Jongepier, F. and Klenk, M. (eds.), The Philosophy of Online Manipulation (pp. 15-48), Routledge, New York. https://doi.org/10.4324/9781003205425-3.

Latour, B. (1987), Science in Action: How to Follow Scientists and Engineers Through Society, Harvard University Press, Cambridge (MA).

Rapp, A., Di Lodovico, C., and Di Caro, L. (2025), How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations, in 'International Journal of Human-Computer Studies', 198. https://doi.org/10.1016/j.ijhcs.2025.103471.

Schlosser, M. (2019). Agency, In Zalta, E. (ed.), The Stanford Encyclopedia of Philosophy. Stanford University. Retrieved 13 Apr 2025, from https://www.plato.stanford.edu/archives/win2019/entries/agency/.

Shionoya, Y. (2001), Trust as a Virtue, in Shionoya, Y. and Yagi, K., Competition, Trust, and Cooperation: A Comparative Study, Springer, Berlin-Heidelberg, pp. 3-19.

Star, S. L. (1999), The Ethnography of Infrastructure, in 'American Behavioral Scientist', 43(3), pp. 377-391.

Striano, F. (2024a), Can artificial agents act? Conceptual constellation for a de-humanised theory of action, in 'S&F_scienzaefilosofia.it', 31, pp. 224-244.

Striano, F. (2024b), The Vice of Transparency. A Virtue Ethics Account of Trust in Technology, in 'Lessico di Etica Pubblica', 1/2024, pp. 70-86.

Swanepoel, D. (2021), Does artificial intelligence have agency?, in Clowes, R.W., Gärtner, K., and Hipólito, I., The Mind-Technology Problem, Springer, Berlin-Heidelberg, pp. 83-104. https://doi.org/10.1007/978-3-030-72644-7_4.

Taddeo, M. (2010), Modelling trust in artificial agents, a first step toward the analysis of e-Trust, in 'Minds and Machines', 20(2), pp. 243-257. https://doi.org/10.1007/s11023-010-9201-3.

Taddeo, M. (2017), Trusting digital technologies correctly, in 'Minds and Machines', 27(4), pp. 565-568. https://doi.org/10.1007/s11023-017-9450-5.

Thompson, C. (2018), Faire confiance aux artéfacts – Faire confiance à distance, in Doueihi, M. and Domenicucci, J. (eds.), La confiance à l'ère numérique, Éditions rue d'Ulm, Paris, pp. 97-111.

Vallor, S. (2024), The AI Mirror. How to Reclaim Our Humanity in an Age of Machine Thinking, Oxford University Press, Oxford.

Wang, K., Quek, B.-K., Goh, J. and Herremans, D. (2025), To Embody or Not: The Effect Of Embodiment On User Perception Of LLM-based Conversational Agents, in 'ArXiv'. https://doi.org/10.48550/arXiv.2506.02514.

Werthner, H. et al. (2022), Vienna Manifesto on Digital Humanism, in Werthner, H., Prem, E., Lee, E. A., Ghezzi, C. (eds.), Perspectives on digital humanism, Springer Cham, pp. xi-xiv