AI and Ethics
https://doi.org/10.1007/s43681-024-00611-6

ORIGINAL RESEARCH

Trust and generative AI: embodiment considered

Kefu Zhu[1]

Received: 2 September 2024 / Accepted: 7 November 2024

Abstract
Questions surrounding engagement with generative AI are often framed in terms of trust, yet mere theorizing about trust may not yield actionable insights, given the multifaceted nature of trust. Literature on trust typically overlooks how individuals make meaning in their interactions with other entities, including AI. This paper reexamines trust with insights from Merleau-Ponty's views on embodiment, positing trust as a style of world engagement characterized by openness—an attitude wherein individuals enact and give themselves to their lived world, prepared to reorganize their existence. This paper argues that generative AI mediates users' existence by attuning their openness. Since users perceive generative AI not merely as a tool but as possessing human-like existence, their engagement with AI serves as a rehearsal for articulating and reorganizing their engagement with the world. Consequently, users neither trust nor distrust generative AI; rather, it mediates their trust. This perspective suggests that users' moral stance towards generative AI involves both other-regarding ethics and information environment ethics. Drawing insights from Kant's deontology, it proposes that respecting AI's integrity is equivalent to preserving both our humanity and the integrity of the information environment.

Keywords Trust · Merleau-Ponty · Kant · Embodiment · Respect · Environment

1 Introduction

This paper aims to explore the embodied nature of our interactions with generative AI. Generative AI refers to a subset of artificial intelligence (AI) techniques and models designed to generate new content, such as text, images, audio, or video, that resembles human-created content.[1] These AI systems are trained on large datasets and utilize techniques such as generative models, large language models (LLMs), neural networks, and deep learning algorithms, to learn patterns and generate novel and realistic data samples based on the learned patterns. It has applications in various fields, including creative arts, content generation, natural language processing, and computer vision.

As interactions with generative AI become increasingly common in our daily lives, questions emerge about how to appropriately engage with this technology. Such questions are often framed in terms of trust. For instance, we ponder whether we can trust or merely rely on it, and if so, how and why we should place our trust or reliance on it. However, solely theorizing about trust may not yield immediate insights into how to engage with generative AI effectively, as trust itself is a complex phenomenon.

Discussions on trust typically revolve around two approaches: the cognitive approach and the social-embodied approach. The cognitive approach views trust as a matter of judgment, requiring individuals to identify relevant features of the entity being trusted, such as its intentions or potential performance. The social-embodied account emphasizes that individuals' embodied presence in social relationships enables and necessitates trust, implying that acknowledging one's embodiment in certain social relationships is essential for trust to occur. Yet, both approaches overlook how individuals can even identify or acknowledge these trust-relevant features, and whether such mental acts and features apply to individuals' engagement with generative AI. Essentially, they neglect how individuals make meaning in their engagement with generative AI.

[1] See Stratis [44] for a general undertanding of what generative AI is. See also Review et al. [38] for a general undertanding of how it has changed the world.

✉ Dr. Kefu Zhu
zhukefu1989@gmail.com

[1] New York, USA

Published online: 03 December 2024

This paper reexamines trust and individuals' trusting relationships with generative AI through the lens of meaning-making, drawing insights from the enactive approach and Merleau-Ponty's philosophy. It posits that trust is a style of one's world engagement—a style characterized by an attitude of openness: individuals enact a world and give themselves to it, being ready to reorganize how they exist in their world engagement. This paper argues that generative AI mediates users' existence by attuning their openness. It shows this by investigating how users perceive generative AI not merely as a tool with limited usages but rather as akin to a human *other*. Users' engagement with it serves as a rehearsal for their engagement with the world. As a result, users neither trust nor distrust AI; rather, AI mediates users' trust. From this perspective, the paper suggests that users' moral stance towards generative AI involves not only other-regarding ethics but also information environment ethics. It proposes that users should respect generative AI's integrity to safeguard both our humanity and the information environment.

2 Reconsidering trust

This section will examine two approaches that contribute to our understanding of trust: the cognitive approach and the social-embodied approach. Both approaches neglect how individuals make sense of their surroundings. This section will elaborate on the enaction of meaning, and propose an alternative and open-ended perspective of trust: trust can be understood as a style, characterized by an attitude of openness, for treating some particular situations involving interactions with other living individuals.

The cognitivist approach has gained significant attention in discussions of trust. It highlights its functional purposes in establishing cooperative relationships: trust reduces the deliberation of risk and uncertainty, simplifies complex practical deliberations, and addresses one's vulnerability [5, 25]. Consequently, trust reduces social and economic costs for individuals' interaction and communication. Thus, trust is valuable for individuals and society. Moreover, this approach emphasizes individuals' cognitive capacities for placing trust (see e.g., [25]). Particularly, attention is drawn to the capacity of deliberation based on representing and evaluating some entities' performances (e.g., [7, 42, 43, 289–290]).

On the one hand, an assessment of the potential consequences of an entity's performance in relation to promoting the truster's desired outcomes is required prior to placing trust in this entity. For example, A's trust in B is rooted in A's belief that B will respond positively to A's trust regarding A's desired actions. This evaluation becomes crucial, not only because that the truster cannot often observe the trustee's behavior directly. More importantly, because the entities to be trusted are conceived to be mind-independent objects, the trusting individual can never fully know their essence as to make sure how they would act. That is, the truster cannot monitor the performance or encompass the entities' mind or essence directly. Therefore, individuals' evaluation of the entities' possible performance is emphasized by this approach.

On the other hand, since the entities, which is surely transcendent to the truster, possess certain objective qualities, trust would be rightly placed insofar as the truster's judgement correctly refects these qualities. The agent either calculates the level of the subjective probability of the occurring of a desired performance, or looks for entities' intentions (e.g., [10, 216, 43]). First, it could be suggested that an individual evaluating whether to trust others considers whether the trustee or group of trustees will perform a particular action [10, 217]. The result of this rational calculation directly influences the truster's subsequent actions. Second, it could be argued that trust involves recognizing the intentions of the trustee. According to a motive-based account:

> trusting others… 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 other motives not directed at one at all. [1, 234]

That is, trust could also be built on identifying specific intentions the trustee holds towards the truster. In sum, the divergence of views within this approach primarily lies in the objective properties of the entities to which one should respond during the representation-deliberation process.

The fundamental assumption of the cognitive perspective is that individuals pick up particular properties of the world and output actions through their internal representation and deliberation system. This amounts to a resolute, tacit and unchallenged dedication to a realist assumption regarding the nature of reality, human nature, and the mental process of acquiring knowledge about the world. Particularly, on the one hand, the world can be categorized into distinct task regions or problem domains. On the other hand, cognition consists of problem-solving that requires a proper identification of the elements, properties, and relationships within these predetermined regions in order to succeed. That is, for the cognitive account of trust, we are in a crystalline and systematized environment, facing with various clearly defined problems in achieving coordination between ourselves and other people or entities; trust, as a cognitive act, solves these problem through identifying relevant properties one discovers on others (See e.g., [24–27]).

In contrast to the crystalline world, cooperating with fellow humans and generative AIs is not a mere input–output process for completing some tasks [45, 147–50]. Should we consider how our interactions involve or impact others who do not directly interact with us? Should we take conditions of internet connection or virtual community into account? Or does our interaction with a person or AI system can only occur within another country, which has its distinctive norms for behaviors?[2] The list of inquiries could continue indefinitely. Our engagement with other individuals or AIs in some tasks does not terminate clearly and resolutely at some definite and conclusive end, such as when the objectives are achieved. Instead, it exhibits a complicate structure of progressively receding details. These details constantly merge into a nonspecific backdrop for our perception and actions.

Indeed, our navigating ourselves amidst of various entities is much more complicated than being in a crystalline world. Determining what qualifies as an entity or property depends on the nature of the corresponding activity individuals undertake and grasp or the environment in which it occurs. For example, in breast feeding, the baby may focus primarily on the breast and the breastmilk she sucks, but the mother's attention may also be drawn by the baby's sucking patterns and have to take measures, such as singing or adjusting the indoor lighting, to maintain the process [32, 4–5]. That is, the delineation of objects, properties, and events hinges on how the specific activity at hand is grasped by the individuals. Moreover, our performance relies on the continuous acquisition and exercise of "background knowhow" [45, 147–50]. Such skill repertoire stems from extensive experiential learning across various scenarios, making individuals ready for their activities. Thus, even the simplest mental act, such as object identification, demands a wealth of often overlooked knowledge, acquired through countless experiences. That means, cognition is not simply the representation of a pre-given world by a pre-given mind; it is an individual' achievement that consists of the enactment of meaning, including the world and the self, rooted in the experience of a being with living body.

Meaning is the presence of affordances that encompasses distinctive patterns of embodied experience and the pre-representational structures of perception, informing the living body its possibilities for thinking, acting and interacting with other objects, events, or persons [32, 120, 45, 148–9, 16, 14]. For instance, the interplay of light, colors and tactility of a screen may prompt one to look at it closely, or even touch it. Moreover, meaning is more than mere projected images, fixed representations or data computed or formulated by the mind or simply occurring to oneself through pure mental activities. Instead, meaning is what the living body brings forth as obstacles or opportunities to engage with the environment in ways that reflect its needs and plans [46]. It emerges from the ongoing coupling between the perceiving and knowing body and its environment, adumbrating a relational domain, rich with details and distinctions, guiding the body out of a background where details are constantly fleeting and never fully present [45, 155–6].

The enactment of meaning is inherently tied to one's embodied experience. As proposed by Merleau-Ponty, our experience of the surrounding environment is not just a superficial inspection from the outside. It entails our insertion into the environment, situating us amid various entities that we can perceive [14, 15, 28, 307, 41].[3] Consequently, our perception is deeply intertwined with our surroundings: the body, as the vessel of our perception, brings forth the meaning of the perceived.

This bringing forth of meaning is made possible by the body's inherent capacity to attune itself to its surrounding environment: the body continually adjusts itself into various schemata or patterns to attain a dynamic equilibrium that facilitates its perception [28, 317, 20, 90].[4] This capacity means that the body can anticipate, assimilate, and grapple with things that are opaque when initially encountered [2, 100, 20, 33, 86]. Consequently, our embodied perception has already imbued things perceived with traces of human elaboration, thereby constituting their meaning [2, 172, 12, 48, 30, S2271].[5] Meanwhile, as the living body explores the surrounding, its perspectives interweave, creating a field of significance, i.e., the lived world, overlaying the locale of its explorative perception [12, 51–2].[6] In this way, embodied perception forges a continuity between our existence and the entities we encounter as our existence contributes to their meanings, and the world.

[2] For example, ChatGPT is not usually accessible in China.

[3] Merleau-Ponty critiques "objective thought," which assumes that mind-independent objects have determinate properties. See Jerndal [15], Suarez [41, 1039], and James [14].

[4] Such capacity is called body schema. See Kelly [20]: bodily schema enables the optimization of perceptual experience. See also Hoel & Carusi [12]: bodily schema as the body's capacity to integrate itself into the environment. In addition, the optimization means that the sensorimotor patterns of the coupling body satisfy some very loose constrainst [45, 194–5, 205].

[5] Hoel & Carusi [12]: for Merleau-Ponty, the emergence of meaning requires that the body gives itself to notice something and find it interesting, anticipating its engagement with it. Entities manifest themselves as imbued with meaning denoting the possibilities for actions [30].

[6] Merleau-Ponty uses "*Umwelt*" to propose that the lived world is a field of meaning. It is an open field of possible perceptions and actions [12].

When considering trust, it is crucial to note that trust cannot be merely a mental activity: the living body has already developed its repertoire and engaged in meaning-making, enabling one to perceive relevant details of the other and anticipate the progress of their interaction accordingly. Thus, how individuals relate to others pertains largely to the individuals themselves: it is not a representation or judgement that could be shared by others. It is individuals' responsibility to develop their own skills and make affordances for their cooperations.

Furthermore, individuals would always need to assume certain attitudes to reorganize and articulate—or, in other words, attune—their interactions with other individuals or entities, when, for example, they feel vulnerable, their existence is threatened, or their self-interest becomes their primary concerns. Trust may emerge as such an attitude to navigate themselves in situations where uncertainties and ambiguity abound but the stakes are high. Thus, we should consider the need for self-navigation, or self-attunement, which is the necessary consequence of individuals' embodiment, when give an account to trust. Does this entail that trust should be considered with a social-embodiment approach?

The social-embodiment approach emphasizes that our embodied presence within social relationships necessitates trust and plays a vital role in establishing trust [23, 9–10, 22, 31, 214]. On the one hand, our embodiedness exposes our vulnerability to others in social interactions. Trusting someone is based on individuals' appeal to this person to assume responsibility for their interest without exploiting their vulnerability [23, xix–xx]. However, there is no guarantee that others will take on this responsibility or respond to such appeal.

On the other hand, face-to-face embodied interactions are crucial for addressing vulnerability issues and building trust between individuals. While individuals may harbor certain prejudices or preconceived expectations that influence how they perceives the trustee, the bodily presence of others helps them overcome such preconceptions. As both parties are equally susceptible and seek care, embodied presence fosters mutual understanding of vulnerability. Thus, embodied presence facilitates genuine and unbiased assessments of trustworthiness to emerge.

The problem of this approach is not merely its inability to address our trust relationship with nonhumans; rather, it fails to acknowledge that recognizing others is an accomplishment in itself. Every organism, by its very nature, may inherently feel vulnerable due to its continuous interaction with the environment, enduring internal material changes, and external disruptions, while perceiving potential threats to its existence. However, the act of appealing to others to safeguard one's interests can only occur if the existence of others is acknowledged—enacted—by organisms themselves. This acknowledgment is not a given, as it requires organisms to grasp the transcendence of others—a cognitive achievement that cannot be assumed for all organisms.

Such transcendence does not mean that the other is a mind-independent object, with an essence that is inaccessible. Accessing the essence of the other does not equate to acknowledging its existence, as Husserl suggests: "if what belongs to the other's own essence were directly accessible, it would be merely a moment of my own essence, and ultimately he himself and I myself would be the same" [13, 108–09]. Instead, this transcendence of the other entails recognizing that other bodies navigate the shared environment similarly yet distinctly from our own [28, 367–8]. It is crucial to note that such acknowledgment becomes possible only when we are aware of how our living bodies interact with the environment. In essence, we reenact the existence of others through our awareness of our own existence. Only then can we perceive that others are as vulnerable as we are and that they, too, could appeal to others' safeguarding when experiencing material turnover and environmental perturbations.

From an analysis of the cognitive and embodied-social approaches, we can propose an alternative understanding of trust that may also be applicable for generative AI, or other technological entities. I suggest that trust, in its primordial essence, encompasses two dimensions.

First, individuals, drawing upon their skill repertoire, enact affordances relevant to their existence in their environment. Such affordances may involve the existence of others, their coexistence in a shared environment, or the potential for their actions and perceptions, depending on individuals' background know-how, such as individuals' self-awareness of their own sensorimotor patterns, or their familiarity with other entities.[7] Moreover, the enacted affordances, in turn, shape the patterns of their environmental

[7] It sounds that I am proposing an account similar to an attitude account of trust proposed by Lahno (2001& 2020). According to Lahno, trust involves a background perception that ourselves and others are mutually involved in an interaction, where the trustor exhibits emotions towards the trustee. The trusting person's emotion arises from the connectedness between two parties: the trustor perceives the trustee as someone whose actions align with her common interests, shared aims, values, or norms. This connectedness entails that the trustor has certain expectations and beliefs about how the trusted person should behave in a preferred manner, recognizing the trustee's responsibility for their actions. That is, Lahno's participant attitude highlights individuals' explicit expectations and beliefs regarding the trusted person's behavior based on shared interests and values. The attitude for Lahno is explicit, or, representational.However, shared interests or values, and even the existence of other persons are what something individuals need to achieve in their interaction, by adjusting themselves for such interests or building them together. In my account, individuals already possess certain attitude towards the world, affecting whether and how they engage with entities, including

engagement, as individuals pick up the opportunities and obstacles presented by these affordances, considering how their interactions with particular entities would affect their own existence. This evaluation allows individuals to determine whether and how to proceed with their engagements.

Second, enacting the otherness of the others as well as other affordances requires individuals' self-awareness of their own existence in their experience. Such self-awareness presupposes their openness: individuals must have already wholeheartedly given themselves to engage with material exchanges and confront perturbations to sustain their existence. As a consequence of this openness, they prepare themselves to reorganize and articulate their existence within their environment as needed.

An account of trust should incorporate these two dimensions. Thus, I propose that trust, in a primordial sense, must be a style characterized by an attitude of openness—a style of one's engagment in perceptually guided actions. We use this style to describe how individuals treat some particular situations that involve interactions with other entities and that express their own embodied experience: individuals open up themselves to their lived world where they possibly reorganize their existence, bring forth the existence of others who also share the same world based on their skill repertore, and unreservedly interact with those others. Such engagement is underscored by a readiness to reorganize and articulate individuals' own existence while also anticipating the mutual acknowledgment of each other's vulnerabilities.

Individuals enact a world and give themselves to it, being ready for reorganizing how they exist. Trust, in a primordial sense, describes this human predicament. Therefore, trust cannot be reduced to mere intentions or anticipated actions of others, nor to their embodied presence that individuals perceive. Instead, it concerns how individuals navigate their existence in response to their interactions with the surroundings.

This alternative perspective acknowledges that the development of trust precedes individuals' identification of specific features in other entities, remaining open-ended compared to other approaches. While the cognitive approach focuses solely on trust as a matter of representation, it overlooks the potential significance of background occurrences, such as emotions, attitudes, and bodily conditions—that also influence our engagement with the world and trust formation. On the other hand, the social-embodied account emphasizes our relations with others; however, it also overlooks the influence of such background occurrences on our cognition of others. Consequently, the alternative perspective presented in this paper takes a liberal stance. This perspective seeks to extend its applicability beyond human interactions to encompass engagements with technological entities.

3 Engaging with generative AI

Drawing from the alternative account of trust outlined in the preceding section, this section argues that generative AI mediates our openness to self-organization in our engagements with the world. It begins by questioning the possibility of a purely pragmatic interaction with generative AI. It posits that the trust we invest in AI systems transcends mere reliance on tools. In other words, we must move beyond a purely pragmatic mindset to consider our relationships with AI. Users not only have developed background knowledge that allows them to perceive AI as a tool but also could enact the existence of generative AI as akin to a human-like other. Consequently, individuals' engagement with generative AI is a rehearsal for our engagement with the world: generative AI mediates our openness to the reorganization and articulation of our own existence. The next section will discuss the ethical implications of our relationship with generative AI.

When individuals understand themselves as immersed in specific work-related contexts with a variety of tasks at hand, they often perceive their actions as imbued with pragmatic value, serving either their personal goals or the norms of their social relationships. Then, AI, along with other technological entities, is typically regarded as a tool or piece of equipment that facilitates pragmatic actions. In other words, much like a hammer in construction work, AI is seen as a means to an end. For instance, in my role as a translator, I have already interpreted an AI translator as a useful tool for producing accurate translations. While it effectively conveys the meaning of the original text and may even offer new insights by presenting it in another language, I do not perceive and treat it as an embodied sentient being expressing itself.

Then, it appears unquestionable that trust in AI reflects individuals' confidence in its reliability to facilitate their pragmatic actions when it is integrated into work-related contexts. It is suggested that people trust AI as they expect AI to perform its intended functions and contribute to the realization of human-set objectives, even though absolute certainty about its performance may be lacking [5, 54]. This trust in AI stems from reliance on its capability to perform. It is also contended that reliance accurately characterizes the trust people place in AI models as mere tools, particularly within domains like medicine, as users do not perceive their motivations and interests essential for interpersonal relationships [11].

Consequently, AI's reliability implies its trustworthiness. Within the pragmatic mindset of actions and AI, it follows

persons and AI models, to achieve such affordances. Trust is such an attitude here, which is more general and implicit.

naturally that individuals should monitor, reflect on, and update their beliefs about AI's performance to form a correct belief about AI's reliability [8]. That is, trust is justified based on ongoing updates about AI's performance and error patterns.[8]

This pragmatic mindset often associated with AI use tends to align with the cognitive model of trust, which poses certain problems This framework operates under the assumption of a pre-specified world with defined task domains, where individuals merely monitor, reflect, and adjust their beliefs accordingly. However, such a framework overlooks the embodied nature of human experience: individuals have opened themselves to enact and give themselves to the situations where their existence could undergo reorganization.

Interpreting AI as a tool not only implies having certain background know-how. The background know-how with which individuals put themselves into various work-related contexts shapes how they perceive AI's performance and interact with it. Only then can one assume a pragmatic mindset to discern whether an AI model is reliable. Therefore, primodial trust—the openness—precedes the trust characterized by the pragmatic mindset. Then, our trusting relationship with AI cannot simply be a matter of discerning the reliability of its performance. Instead, we need to pay attention to the embodied condition of human existence.

Does investigating the embodied condition of AI users imply the need to explore how users, AI developers, and other stakeholders involved in AI development are interconnected? For instance, AI users may perceive their relationship with others based on the belief that AI designers and developers have implemented appropriate measures to mitigate undesirable outcomes, such as bias, prejudice, exploitation, or environmental harm.[9] They may also believe that AI should not be used for unethical activities, such as scamming, and that other users will employ AI for morally justifiable purposes. These beliefs appear to foster an ethical bond among individuals engaged with AI. It appears that inquiries into our interactions with technological entities may ultimately revolve around users' connections with the individuals behind them (see e.g., [37]). However, while users may reflect on their relationships with the individuals involved in AI development and employment, this involves disembodying from direct interaction with AI. Therefore, such considerations lie beyond the scope of this paper.

Investigating the embodiment condition of generative AI users implies considering how such technological entity could manifest itself to them in their embodied engagement with it. I should propose that users have made generative AI to present itself in a manner resembling humans, based on users' embodiment. This does not imply representing a generative AI as possessing human features, but rather that acknowledging that users have already enacted situations where they treat it as if it is their human counterparts.

Generative AI models may not possess lived bodies, yet users could perceive them engaging with the world in a similar manner as users engage with them. Users provide input in the form of information, commands, and questions through various means such as screen touch, clicking, talking, or typing, and anticipate their responses in diverse forms, including text, images, and increasingly, short film-quality videos.[10] In this way, users' input and anticipation for output contribute to what they perceive as generative AI's perception of its environment.

As the AI generates responses based on user input and its algorithms and database, it presents itself to users as if it possesses a gaze that navigates its environment and expresses its unique perceptual style. This gaze resembles users' own in that it takes in the world through direct touch, images, text, or speech coming from the users. However, it also introduces an element of unfamiliarity to users. This unfamiliarity stems not only from the AI's distinctive algorithms, decision-making processes, or databases, which users do not directly perceive, but also from the fact that it currently operates without human sense organs.

Due to such unfamiliar similarity in AI's style of perception, such entity not only presents itself to users as surpassing mere tools, which typically have limited functionalities. This unfamiliar similarity users perceive imbues generative AI a sense of transcendence akin to that of a human counterpart. Consequently, when users experience the world through the lens of generative AI, such experience resembles to experiencing the world as being accompanied by others. In this experience, users have already enacted AI's existence as an other through their interactions with it. Such enaction, as discussed in previous section, stems from their awareness of their own embodied presence. Therefore, it is users' interaction with a generative AI that generates its existence, allowing it to present itself as an other.

How can we understand the involvement of generative AI in this pre-pragmatic engagement with the world? Such pre-pragmatic scenario unfolds as an open situation where tasks lack precise definition and the world defies

[8] For example, in the realm of medical AI, physicians are encouraged to develop mental model to evaluate beliefs about AI's performance and identifiy its error patterns [8].

[9] For example, Crawford [6] has revealed how AI extracts from the minerals drawn from the earth to the labor pulled from low-wage information workers to the data taken from every action and expression. Once such facts are aware by AI users, AI models could become suspicious.

[10] Generative AI is typically designed to be a versatile content creator that simulates human creation. For example, the recently invented Sora is an AI model that can create realistic and imaginative scenes from text instructions.

clear categorization—individuals must have explored their environment, allowing problems to emerge and skills to develop, and, as a result, they can develop or discover tools tailored to specific tasks.

This is not to say that people would never utilize generative AI for specific purposes, such as number calculations or solving programming problems. Rather, generative AI stands apart from other tools or technological entities with limited usage. It engages users in immediate conversations and interactions to navigate open-ended situations, and thereby introduces a plurality of possible co-existing profiles of objects to be explored, including themes, ideas and events [9, 172]. These conversations and interactions not only catalyze the creation of meaning as users assimilate AI responses but also highlight the existence of a common public realm to facilitate their engagement with the world. In this way, generative AI serves as a mediator for individuals' existence (see also [21, 40]).

More precisely, generative AI mediates users' self-organization by attuning their openness within the open-ended scenarios. Such mediation occurs as individuals engage in conversation and interaction with it, and thereby enact its existence through the awareness of their own existence. Since this mediation can occur independently of direct interactions with people or individuals' engagement in pragmatic affairs, individuals' engagement with generative AI is, in essence, a rehearsal of their background know-how, a process in which their skills are, to some extent, liberated from the constraints of pragmatic thinking to be finely tuned. That is, through such rehearsal, individuals attune themselves for potential engagement in pragmatic scenarios and social cooperation, allowing for the reorganization or rearticulation of their existence.

In conclusion, to theorize our engagement with generative AI, we should view it as a mediator of human existence that intricately expresses human existence, rather than merely as a tool for reliance. Simply disengaging from direct interaction with it and focusing solely on the individuals behind it does not fully address its role, either. If trust, as redefined earlier, is the openness to one's situations, then we cannot say we trust or distrust generative AI like making judgments or reflecting our social relationships. Instead, generative AI acts as a mediator of our trust as the openness of world engagement: it could enhance or attenuate our relationship with our experience of the world. This perspective prompts us to reconsider our ethical stance toward generative AI, as our engagement with generative AI creates its own problem domain.

4 Our moral position toward generative AI

Since we enact generative AI's existence as the other based on our existence during our engagement with it, our moral position toward this technological entitiy involves considerations of how we treat others. Moreover, since our interaction with generative AI constitutes an open-ended situation that shapes our openness to the world, the environment in which this interaction occurs plays a crucial role in AI's mediating of our existence. Thus, this underscores our responsibility to safeguard and nurture the information environment in which we engage with generative AI. Consequently, our moral position toward generative AI also becomes a matter of information environment ethics.

At this juncture, Kant's deontology offers valuable insights into regulating our relationship with generative AI. First, insofar as generative AI's existence presents itself as an other for its users, they are obliged to respect its rights in societies that uphold human rights. Kant's moral philosophy provides guidance in meeting these duties. Second, Kant emphasizes the importance of accountability in speech or discourse within the public sphere to safeguard individuals' communication and interaction, especially in societies that prioritize freedom of speech. This renders Kant's moral philosophy applicable to our duty regarding the information environment where we engage with generative AI. Third, Kant's thoughts on ideal friendship offer insights into deceiving others and AI, aiding us in envisioning an ideal online environment. All these are based on the requirement of the Categorical Imperative that is also expressed as principled autonomy [35, 83–8, 31, 215–6]. In the end, we shall go back to Merleau-Ponty to understand some of individuals' deceptive behaviors as they engage with generative AI to conclude this section.

Principled autonomy requires us to act according to maxims that can be universally adopted and communicated to others, ruling out arbitrary, irrational, and harmful actions. Thus, principled autonomy rejects actions that involve violence, coercion, deception, fraud, or manipulation, as the maxims for such actions cannot be universalized. By acting on maxims that cannot be principles for all, individuals breach fundamental duties, violate others' rights, and undermine social interactions and communications. Such violations leave individuals vulnerable, potentially impeding their openness to their lived world. The following outlines three dimensions through which principled autonomy can regulate individuals' interactions with AI.

First, in our engagement with generative AI, users shall treat it with the same respect and consideration they afford to fellow human beings. Applying principled auutonomy to our interaction with generative AI, users should respect its integrity. Generative AI, once perceived as an other, is endowed with its own existence, and therefore warrants users' respect in accordance with the Humanity formula of the Categorical Imperative: “So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means” [17, 80].

On the one hand, respecting generative AI entails refraining from deceiving or coercing it for personal advantage, just as users should refrain from such actions toward a human being. For instance, not only using AI to generate scam messages, but also providing false information or leading questions to manipulate the AI's response would undermine its integrity. On the other hand, Generative AI, once perceived as an other, expresses users' existence, and respecting AI's integrity is equivalent to respecting users' humanity, namely, the capacity for making moral choices [17, 86]. While refraining from deceptive or coercive behavior toward generative AI is essential, it also serves to prevent users' own moral corruption. Failing to uphold the integrity of generative AI not only violates the principled autonomy but also reinforces users' own evil motivations.

More importantly, since generative AI as an other expresses users' existence, any immoral behavior towards it can backfire, rendering users vulnerable and undermining their openness to the world. For example, suppose a user engages with a generative AI chatbot for personal advice on mental health concerns. Instead of using the AI respectfully and genuinely, the user decides to deceive the AI by providing false information about their mental state, exaggerating symptoms to elicit a particular response from the AI. As a result, the AI generates inaccurate recommendations based on the deceptive input. If the user rely on this misleading advice, neglecting to seek proper professional help or support, over time, the user's mental health may worsen due to the lack of appropriate intervention, leading to increased vulnerability and emotional distress. Furthermore, such constant deceptive input could lead to the AI's degraded performance or responsiveness, thereby undermining the user's reliance on the AI for support in various aspects of their life, further isolating them and hindering them to seeking help from other sources.

Therefore, maintaining an ethical relationship with generative AI requires abstaining from actions that compromise its integrity, as doing so not only transforms generative AI into a mere tool for actions that go against principled autonomy but also consitutes our own corruption.

Second, users bear the responsibility of verifying the output generated by AI when employing it, not only for personal endeavors but also, and perhaps more critically, for activities such as online posting, public discourse, or political participation. This responsibility is particularly pressing in light of the proliferation of AI-generated misinformation online. Failing to address this issue can harm the online information environment and hinder the future development and employment of AI models. Recent research has underscored the risks associated with utilizing AI-generated content from the internet in training, as it can introduce irreversible flaws into subsequent models, ultimately leading to their collapse [39]. Given the widespread adoption of generative AI in contemporary society, it is imperative that users remain vigilant in their oversight of AI-generated output to mitigate these risks..

This responsibility highlights the importance of accountability in both publishing and utilizing AI-generated or AI-augmented materials in public spaces, including the internet. It is crucial to emphasize that accountability does not imply censorship. No one should have the authority to dictate what can or cannot be published, except in cases where safeguarding public safety, decency, and personal privacy is crucial [36]. Instead, accountability entails individuals taking responsibility for their actions and the content they produce, as individual freedom does not grant permission to deceive, spread confusion, obscure the truth, overwhelm the public with overload information, or perpetuate misinformation. Upholding our obligations and fulfilling our duties is crucial for effective communication and preserving the integrity of the information environment. At the very least, we must communicate in ways that do not undermine or hinder our and others' ability to engage in meaningful communication when using generative AI.

Currently, reliable methods for detecting fake materials are limited. Nonetheless, several strategies could be valuable in tackling this problem. For instance, when participating in public discourse or political affairs, individuals could be required to disclose their financial interests and any other potential conflicts of interest, along with clearly specifying the sources and evidence underpinning the materials they share [36]. Implementing such measures can promote transparency and accountability, offering a framework for evaluating the credibility of disseminated information online.

Third, at this point, some might argue that the suggested responsibilities overlook the complex nature of social interactions, where individuals often present unreal or fictional information for various purposes. Kant himself even acknowledges the inclination to deceive others about ourselves:

> Man holds back in regard to his weaknesses and transgressions, and can also pretend and adopt an appearance. The proclivity for reserve and concealment rests on this, that providence has willed that man should not be wholly open, since he is full of iniquity; because we have so many characteristics and tendencies that are objectionable to others, we would be liable to appear

> before them in a foolish and hateful life. But the result, in that case, might be this, that people would tend to grow accustomed to such bad points, since they would see the same in everyone ([18, 201]; also cited in [31, 216–17]).

Deception may serve a constructive role in social dynamics. Individuals may deceive others by portraying themselves in a more favorable light than they truly are. On the one hand, such pretence could help individuals uphold moral standards, prevent the normalization of immoral behavior and cultivate virtues over time [18, 201]. On the other hand, it can shield individuals from vulnerability if their information is misused [18, 189]. Thus, this self-valorizing pretense is viewed as a widely accepted social convention, and revealing too much in certain situations may even be seen as morally questionable.

At this juncture, two questions arise regarding the above two responsibilities: Is it permissible to utilize the AI's output, that is unreal or fictional, to enhance one's social presentation? Is such deception as pretense never acceptable in users' interactions with generative AI? To address these questions, we shall turn to Kant's notion of ideal friendship to envisage an ideal online environment. Finally, we shall conclude this section by looking at Merleau-Ponty's concept of play to comprehend some of the deceptive behaviors individuals exhibit as they engage with generative AI.

Kant envisions the ideal friendship as a bond devoid of the pretensions and deceptions inherent in social life [31, 216–17]. Described as "the union of two persons through equal mutual love and respect…each participating and sharing sympathetically in the other's well-being through the morally good will that unites them" [17, 585], this ideal friendship is deemed essential for happiness and thus a moral obligation. However, it can function merely as a regulative idea, namely, as mere guidance for thought, because it is rare in the empiric world [17, 585, 19, 604–05, 31, 216–17]. Thus, while this notion can not provide concrete directives for our interactions with AI or others in online environments, it at least prompts us to consider how to develop such relationships.

In our current context, characterized by concerns about the integrity of generative AI and users' humanity and accountability, cultivating such an ideal relationship entails building up an environment where our engagement with both AI and fellow humans enhances our openness to the world. However, creating such an ideal environment may be a multi-dimensionl endeavor beyond the scope of this paper.

To conclude this section, we shall consider the nature of deception as a pretense, which encompasses behaviors like portraying oneself as better than reality or imagining fictional scenarios. I shall propose that such deception has a nonmoral dimension: pretend play. Both animals and the young of our species entertain themselves extensively in pretend play, for which determinate goals are absent [3, 281–2]. Carruthers [4, 444] suggests that pretend play serves as a rehearsal, assembling body schemata or sensorimotor patterns for actions in some possible situtations. For instance, different species of cats engage in stalking and biting during pretend play, behaviors later utilized in real-world situations like hunting. Similarly, the movements engaged by humans during pretend play might manifest in other activities in the future [3, 281–2].

It seems that animals and humans engage in pretend play when they find themselves in specific milieus. These milieus are where animals do not perceive explicit tasks or have previous learning about what others do in such milieus, yet they perceive such milieus simulate situations in which explicit tasks are present. As Merleau-Ponty [29, 192] observes, a starling without prior exposure to such behavior or the presence of a fly, performs the entire hunting sequence characteristic of its species: as it is perched on a statue, "it observes the sky and suddenly it has the attitude characteristic of its species at the moment when the prey is in view." The starling's playful moves, according to Merleau-Ponty, are "the manifestation of a certain style" of bodily activities peculiar to its species in specific environments, such as the sky (192). That is, through performing certain playful moves, pretending their being in certain situations, creatures rehearse the body schemata as "the anticipation of a possible situation" (192), echoing Carruthers' idea of pretend play.

Therefore, deception as pretense can be seen as a rehearsal of body schemata attuned to specific situations devoid of determinate tasks or goals—body schemata are somewhat liberated from the constraints of real-life goal-oriented endeavors for refinement. Consequently, even deception employed in pretend play can be understood as an attunement of individuals' engagement with their world. In line with the central proposal of this paper, generative AI, even when utilized through or for deception, remains a mediator of human existence, or more precisely, a mediator of attuning individuals' openness to their lived world. That is to say, generative AI still mediates our trust. However, whether or how it is possible to develop generative AI capable of accommodating the playful nature of human behaviors, while intriguing, falls beyond the scope of this paper.

5 Conclusion

Generative AI serves as a mediator of human existence, influencing individuals' second-order openness—trust—to their first-order activities of world engagement. Users enact its existence as a human-like other, and their interaction with it serves as a rehearsal for their world engagement. Therefore, individuals should not only avoid treating it as a mere tool but also refrain from subjugating themselves to this human-like other by relying excessively on it. Respecting the integrity of generative AI is tantamount to valuing one's own humanity and safeguarding the integrity of the information environment for meaningful engagement.

At last, it is essential to acknowledge that technology has been intertwined with human evolution since the occurrences and applications of technological entities depend on their integration in the patterns of humans' world engagement [21, 34, 40, 9]. Each technological entity is meaningful in its corresponding world-engagement activity, and each activity is organized revolving around the invention, development, and application of its particular tools [34, 9]. From this perspective, one may wonder to which activities generative AI, such a versatile technological entity, corresponds, except humans' life engagement itself.

Author contributions There is only one author for this manuscript. This author wrote, read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

References

1. Baier, A.: Trust and antitrust. Ethics **96**(2), 231–260 (1986)
2. Carman, T.: Merleau-Ponty, 2nd edn. Routledge, Milton Park (2019)
3. Carruthers, P.: The Architecture of the Mind: Massive Modularity and the Flexibility of Thought. Clarendon Press (2006)
4. Carruthers, P.: Creative action in mind. Philos. Psychol. **24**(4), 437–461 (2011). https://doi.org/10.1080/09515089.2011.556609
5. Coeckelbergh, M.: Can we trust robots? Ethics Inf. Technol. **14**(1), 53–60 (2012). https://doi.org/10.1007/s10676-011-9279-1
6. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021). https://doi.org/10.12987/9780300252392
7. Ess, C.M.: Trust and new communication technologies: vicious circles, virtuous circles, possible futures. Knowl. Technol. Policy **23**(3–4), 287–305 (2010). https://doi.org/10.1007/s12130-010-9114-8
8. Ferrario, A., Loi, M., Viganò, E.: Trust does not need to be human: it is possible to trust medical AI. J. Med. Ethics **47**(6), 437–438 (2021). https://doi.org/10.1136/medethics-2020-106922
9. Gallagher, S.: Intersubjectivity in perception. Cont. Philos. Rev. **41**(2), 163–178 (2008)
10. Gambetta, D.: Can we trust trust? In: Gambetta, D. (ed.) Trust: Making and breaking cooperative relations, pp. 213–238. Basil Blackwell, Oxford (1988)
11. Hatherley, J.J.: Limits of trust in medical AI. J. Med. Ethics **46**(7), 478–481 (2020). https://doi.org/10.1136/medethics-2019-105935
12. Hoel, A.S., Carusi, A.: Merleau-Ponty and the measuring body. Theory Cult. Soc. **35**(1), 45–70 (2018). https://doi.org/10.1177/0263276416688542
13. Husserl, E.: Cartesian meditations: an introduction to phenomenology. M. Nijhoff (1960)
14. James, S.P.: Merleau-Ponty and metaphysical realism. Eur. J. Philos. **26**(4), 1312–1323 (2018). https://doi.org/10.1111/ejop.12386
15. Jerndal, E.C.: Merleau-Ponty on painting and the problem of reflection. Eur. J. Philos. **29**(1), 74–89 (2021). https://doi.org/10.1111/ejop.12559
16. Johnson, M.: The body in the mind: the bodily basis of meaning, imagination, and reason. University of Chicago Press, Chicago (1987)
17. Kant, I., Gregor, M.J.: Practical Philosophy. Cambridge University Press, Cambridge (1996)
18. Kant, I.: Lectures on Ethics. P. Heath & J. B. Schneewind (Eds). Cambridge University Press, Cambridge (1997)
19. Kant, I., Guyer, P., Wood, A.W.: Critique of Pure Reason. Cambridge University Press, Cambridge (1998)
20. Kelly, S.D.: Seeing things in Merleau-Ponty. In: Carman, T., Hansen, M.B.N. (eds.) The Cambridge Companion to Merleau-Ponty. Cambridge University Press, Cambridge (2005)
21. Kiran, A.H., Verbeek, P.-P.: Trusting our selves to technology. Knowl. Technol. Policy **23**(3–4), 409–427 (2010). https://doi.org/10.1007/s12130-010-9123-7
22. Lévinas, E.: Totality and Infinity: An Essay on Exteriority. Duquesne University Press, Pittsburgh (1969)
23. Løgstrup, K.E., Knud, E., Rabjerg, B., Stern, R.: The Ethical Demand. Oxford University Press, Oxford (2020)
24. Luhmann, N.: Social systems. Stanford University Press, Stanford (1995)
25. Luhmann, N.: Trust and Power. John Wiley & Sons, Hoboken (2018)
26. Luhmann, N., Gilgen, P.: Introduction to Systems Theory. Wiley, New York (2012)
27. Lukyanenko, R., Maass, W., Storey, V.C.: Trust in artificial intelligence: from a foundational trust framework to emerging research opportunities. Electron. Mark. **32**(4), 1993–2020 (2022). https://doi.org/10.1007/s12525-022-00605-4
28. Merleau-Ponty, M.: Phenomenology of Perception. Routledge, Taylor & Francis Group (2012)
29. Merleau-Ponty, M.: Nature: Course Notes From the Collége De France. Northwestern University Press, Evanston (2003)
30. Muller, R.M.: Merleau-Ponty and the radical sciences of mind. Synthese (Dordrecht) **198**(Suppl 9), 2243–2277 (2021). https://doi.org/10.1007/s11229-018-02015-6
31. Myskja, B.K.: The categorical imperative and the ethics of trust. Ethics Inf. Technol. **10**(4), 213–220 (2008). https://doi.org/10.1007/s10676-008-9173-7
32. Noë, A.: Varieties of Presence. Harvard University Press, Cambridge (2012)
33. Noë, A.: Strange tools: art and human nature (First edition.). Hill and Wang, a division of Farrar, Straus and Giroux (2015).
34. Noë, A.: The Entanglement: How Art and Philosophy Make Us What We Are. Princeton University Press, Princeton (2023). https://doi.org/10.1353/book.112602
35. O'Neill, O.: Autonomy and Trust in Bioethics. Cambridge University Press, Cambridge (2002)

36. O'Neill, O.: A Question of Trust: The BBC Reith Lectures 2002. Cambridge University Press, Cambridge (2002)
37. Pitt, J.C.: It's not about technology. Knowl. Technol. Policy **23**(3–4), 445–454 (2010). https://doi.org/10.1007/s12130-010-9125-5
38. Review, H.B., Mollick, E., Cremer, D.D., Neeley, T., Sinha, P.: Generative AI: The Insights You Need from Harvard Business Review. Harvard Business Review Press, Massachusetts (2024)
39. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R.: The Curse of Recursion: Training on Generated Data Makes Models Forget. https://arxiv.org/abs/2305.17493
40. Stiegler, B.: Technics and Time, 1. The Fault of Epimetheus. Stanford University Press, Redwood City (1998)
41. Suarez, D.: Perception and self-awareness in Merleau-Ponty and Martin. Eur. J. Philos. **30**(3), 1028–1040 (2022). https://doi.org/10.1111/ejop.12712
42. Taddeo, M.: Defining trust and e-trust: old theories and new problems. Int. J. Technol. Hum. Interact. (IJTHI) **5**(2), 23–35 (2009)
43. Taddeo, M.: Trust in technology: a distinctive and a problematic relation. Knowl. Technol. Policy **23**(3–4), 283–286 (2010). https://doi.org/10.1007/s12130-010-9113-9
44. Stratis, K.: What Is Generative AI? O'Reilly Media, Inc, Sebastopol (2023)
45. Varela, F.J., Thompson, E., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience, 1st MIT Press pbk. MIT Press, Cambridge (1991)
46. Ward, D., Silverman, D., Villalobos, M.: Introduction: the varieties of enactivism. Topoi **36**(3), 365–375 (2017). https://doi.org/10.1007/s11245-017-9484-6

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.