AI and Ethics (2025) 5:19–28
https://doi.org/10.1007/s43681-024-00626-z

ORIGINAL RESEARCH

From human-system interaction to human-system co-action and back: ethical assessment of generative AI and mutual theory of mind

Florian Richter[1]

Received: 27 June 2024 / Accepted: 18 November 2024 / Published online: 4 December 2024
© The Author(s) 2024

Abstract
Human-machine ethics has emerged as a rapidly growing research field in recent years. However, it seems that Generative Artificial Intelligence (AI) leads to a paradigm shift from human-machine interaction to co-action. The ethical assessment of such relationships is still in the making and needs further scrutiny. First, studies about the influence of technology in human-system interactions and manipulation are reviewed. Second, the "mutual theory of mind" approach is critically examined to identify its shortcomings. Third, creating user models is reconstrued to demonstrate the strategies of systems. Finally, use cases are discussed and assessed to outline ethical implications.

Keywords Generative AI · Mutual theory of mind · Human-machine ethics

1 Introduction

Technology ethics has become a multifaceted research area ranging from normative to empirical investigations. Specifically, in human-machine ethics, it is paramount to investigate technology's mediating role and ethical implications. The conceptual level has been investigated, for instance, in postphenomenology [1–3]. Furthermore, in recent years, several studies have been conducted to shed light on the influence of technology on human behavior in human-machine ethics, from the influence of chatbots [4] to the ethics of autonomous vehicles [5, 6], hybrid traffic [7], and the ethical implications of care [8]. However, immersion in virtual reality and real-time feedback from systems to subjects via Generative Artificial Intelligence (AI) is still in the making. It is discussed under the concept of co-action. The ethics of human-system interaction, which AI technologies have already disrupted, faces new challenges in developing ethical assessment tools for human-system co-action.[1]

In human-system interactions, the developers and users try to align their expectations about how the artifact/system should function and how it should be designed. This is done via studies in UX Design, acceptance research, etc. To a certain degree, the *practice* is more or less transparent. In human-system co-action, the user *and* the (semi-) autonomous system produce an effect or output. Examples are text or picture generation via prompts or brain-computer chips. However, it is opaque to the user how the system functions to produce a specific effect and who is responsible in the end. How is it possible to reflect on the role of the system in creating the output? How can the user be involved in the developmental process? And how can such systems be assessed ethically. It is hypothesized that if the system functioning is made explicit to yield a more transparent practice, like in human-system interaction, and the user is included, then an ethical assessment of co-actions and the use of Generative AI is possible. Therefore, we must move from human-system interaction to human-system co-action *and back* [9, pp. 140–144]!

Aligning expectations of systems such as conversational agents (CA) and humans has been discussed under a "mutual theory of mind" [10, 11]. It should allow "smoother human-CA conversations" by adapting the mutual theories or models of the mind [10]. However, it is unclear if such models of the mind reflect either the system's functioning or the human agent well. Instead, the system ascribes roles, creates relevance rankings, establishes routines for the user,

[1] I am grateful for comments from Matthias Uhl and many discussions with him about the topic and an earlier draft of the paper.

✉ Florian Richter
florian.richter@thi.de

[1] Technische Hochschule Ingolstadt, Esplanade 10,
85049 Ingolstadt, Germany

and based on such an analysis, the system offers recommendations, explanations, or options for the user. By having an actual theory of mind, the user can also, to a certain extent, exploit the system's strategic functioning to reach a specific goal or objective. This could be extended by assuming that the system or the virtual conversational agent has adaptational strategies, too, to react to these strategies. The strategic "exploitation" could be captured via game theoretic models. Nevertheless, it presupposes that the user and the system "understand" each other, i.e., can interact on a strategic level. However, in the case of co-action, where an effect is produced by the user and a (semi-)autonomous system [9, p. 140], like in the generation of pictures or texts via prompts or in the use of brain-computer interfaces, difficulties arise to align the expectations of the user with the system strategies due to the opaqueness of the action environment and context for the user. This has various ethical implications regarding how trust can be developed and how the system strategies influence the users without manipulating them.

The paper makes four contributions. First, studies about the influence of technology in human-system interactions are reviewed to assess and summarize the ethical implications of the mediating role of technology. Second, the approach of a "mutual theory of mind" is discussed. Its shortcomings are identified in opening the search space for more adequate tools to assess human-system co-action. Third, creating user models is analyzed to demonstrate the strategies of systems. However, and that is the final aspect of the paper, such strategies can, on the one hand, be exploited by users but, on the other hand, be used to control or manipulate the user. Here, use cases are discussed and assessed to outline ethical implications.

2 The mediating role of technology

2.1 Philosophical background

Technology is sometimes defined as a collection or system of connected artifacts [12, p. 42, 13, p. 2]. This usually entails an instrumental reduction of artifacts serving as tools to achieve goals set by humans [14]. To address these reductionist views, various approaches have been proposed in recent decades, such as ideas that stem from Wittgenstein's philosophy of language [15], postphenomenological approaches (e.g., Don Ihde and Peter-Paul Verbeek), dialectical approaches [16], or lines of thought that emerged from Technology Assessment [17]. All approaches have in common that technology mediates our access to the world and us in a certain way. Technologies are not "opposed to humans, or mere extensions of us, they need to be seen as media for our connections with the world" [1, p. 29]. (This presumably explains also the intricate relation of the philosophy of technology with anthropological statements like, e.g., that we are "deficient beings" (Arnold Gehlen).)

Each form of mediation entails a certain ethical viewpoint on technology. From an instrumentalist perspective, artifacts appear just as means and are value-neutral or -free [14]. From a postphenomenological perspective, technology seems to be shaping our access to the world. Consequently, technology is imbued with values [3]. Despite the fierce defense of the value-neutrality thesis of technology [14], it can be doubted that it can serve as a basis for technology ethics and adequately grasp its challenges. Technological systems can be much more complex than tools like guns or hammers. Some systems are highly automated and act partly autonomously (like self-driving cars). Hence, they could be understood as displaying, to a certain degree, moral agency (compare the four levels of moral agency by James H. Moore [18]). Additionally, some systems have a high depth of engagement into the human (e.g., chips implanted into the brain). For instance, self-driving cars operate within a complex system and interact with humans (pedestrians, human drivers).

2.2 Empirical studies in human-machine ethics

Therefore, based on the theoretical assumption that technology is value-laden, shapes our practices, and forms who we are as humans, several empirical studies have been conducted to shed light on the influence of technology on human behavior in human-machine ethics. Some approaches stem from behavioral economics or psychology [4–8]. Although empirical studies about the influence of moral advisors are still scarce [19], some studies show that users follow moral advice from systems or bots even if it is clear to them that they are interacting with an AI-based system [4, 20–22] Furthermore, it seems that people sometimes overtrust robots (for instance, in emergency situations) [23]. Hence, it seems evident that technological systems can influence the moral behavior of humans.

Under the assumption that a clear moral framework could be provided, it would, therefore, seem that a certain kind of nudging via technological systems could also be in the interest of its users [24]. The system would influence the users positively, helping them to make morally correct choices or to live in a morally right way. Although Thaler and Sunstein proclaim that nudging people is just a "paternalism of means" [25, p. 7 and 312], it is, in fact, also a paternalism of ends because some kind of common sense moral framework is supposed that neither is questioned nor does it include stakeholders to assess which ends they consider acceptable.

Klincewicz explicitly considers moral AI machines as "means" to influence behavior and to somehow remove immoral behavior. He writes, "it does give a promising alternative to pharmaceuticals and neural intervention for individualized moral enhancement." Furthermore, such a moral enhancement would not be paternalistic, according to him, because it persuades reason-responsive beings and is thus not coercive. Nevertheless, the problematic issue is that Klincewicz already knows what the morally correct behavior is, at least in some specific cases that he discusses. In his example, a racist would and should be persuaded by a moral machine to report an abuse of a Chinese person by a police officer [26]. Yet, this is a moral problem and not an ethical issue where different ethical stances, values, and principles must be weighed against each other. To extend a moral machine that solely functions as a persuasion tool to these ethical cases would then not only be a paternalism of means and nudge people to do the morally right thing, but it would also be a paternalism of ends.

The example from Klincewicz above does not need specific technology ethics. General moral considerations would be sufficient [27, p. 16]. However, it is crucial to investigate how AI-based systems influence their users and co-agents, i.e., how they shape the space of possibilities to (co-)act. In this process, two problematic issues arise:

1. It is essential to distinguish between mere influence that can be considered ethically acceptable (like most advertisements) and manipulation (trickery like grandparent scam). Here, clear normative criteria need to be established.
2. Biases and discrimination are essential parts of the debate in the ethics of algorithms [28]. However, biases cannot be eliminated completely to reach an unbiased original position.– Technology (but also morality) would be value-neutral.– Any relevance ranking already discriminates in the sense of distinguishing between something; otherwise, we could only use meaningless tautologies, such as "Manipulation is manipulation." or "A value is a value."

2.3 Technology between influence and manipulation

While the former issue will be examined here shortly, the latter will be discussed in section four. Although "it has been suggested that manipulation sits on a continuum of influence between rational persuasion and coercion" [29, p. 4], it is rather difficult to determine clearly what manipulation is. Some claim that to "covertly influence" someone is manipulation, i.e., "to intentionally alter their decision-making process without their conscious awareness" [30]. However, manipulation does not need to be hidden; for instance, a real estate agent might offer freshly baked cookies in a house, while the potential buyers might also be aware of the fact that they are manipulated [29]. Another attempt is to define manipulation as "bypassing the chooser's deliberative capacities" [31, p. 86]. Nevertheless, one might also be manipulated by responding to reasons. For instance, one might have a strong sense of fairness, and a manipulator might appeal to it to get a better share. Furthermore, manipulation is characterized as trickery that presupposes an intention on the side of the manipulator [32]. However, it can also be the case that the person is a pathologic manipulator and manipulates without being consciously aware of it. Additionally, despite not having intentions, technological systems could be characterized as manipulative. Even though such systems might be used solely as tools by a human agent to manipulate someone, some systems can also be manipulative without being intentionally designed to do so [29, p. 8]. Manipulation does not need to be done intentionally. As long as it is goal-oriented (where the goals are either set by humans or systems), it could be manipulative [33]; otherwise, "influence that is purely accidental" would count as manipulation [29, p. 9].

Manipulation is non-intentional [29] and is thus different from concepts like nudges, dark patterns, and persuasive technology.[2] In the latter concepts, some kind of intentional and deliberative decision on the side of the designer or developer is made to influence the user, whether in a good or bad way. Nudges are, for instance, based on a choice architecture that is deliberately constructed to influence the user to achieve a goal that supposedly– according to the designers or developers– is good for her (for an overview of the topic, see [34, 35]). Dark patterns, on the other hand, "are deliberately constructed designs with the intent of inducing users to take actions they might not willingly choose if they were fully informed or aware of alternatives" [36, p. 207, 37]. Representants of libertarian paternalism consider nudges ethically acceptable [24], while dark patterns are considered ethically unacceptable. Persuasion could be considered acceptable if it is based on (good) reasons, while it can be considered unacceptable in the sense of coax or cajole. Of course, such a distinction is very rough, and there is also a "tension between persuasive UX practices and manipulative designs" [38]. Nevertheless, it seems that normative criteria are established to a certain degree in these cases within the field of human-computer interaction. However, manipulation is rather difficult to determine ethically.

[2] I am thankful to an anonymous reviewer for pointing out that these concepts play a crucial role in human-computer interaction. Nevertheless, the main background of the paper is philosophical and ethical and thus rooted more in the metaethical discussion.

Klenk adds what he calls an “indifference criterion”: the manipulator is “indifferent to revealing reasons to their victims in their choice of the means of influence that they employ.” Furthermore, it “can be interpreted non-intentionally by thinking about the function of a chosen means of influence,” like in a movie recommender system [29, p. 8/9]. Nevertheless, this seems to presuppose some strategic planning on the side of the manipulator, which means it should be chosen to achieve a goal most efficiently and effectively. The problem is here that almost any kind of influence could count as manipulative. A driver assistant system that influences the driver on an operational level would also be manipulative as long as it does not reveal reasons why it chooses certain means to protect the driver. Usually, there is also no time for some deliberative process or exchange with the driver to give her reasons, which means should be chosen regarding the situation.

Yet, the systems were tested and developed accompanied by a deliberative process. The debate’s primary focus is evaluating means-end schemata that individuals could employ. Ultimately, the ethical task would be to assess the manipulative side-effects of goal-oriented behavior and their risks. Also, Klenk already sees the problematic issue that systems can be manipulative and that questions of manipulation must be incorporated into the technological systems design process to prevent manipulation [29]. However, the system perspective must be more clearly elaborated by investigating the system strategies employed to model users and how users can be included in shaping the systems. This will also have the advantage that worries where influence “undermines the target’s autonomy” because “first, it can lead them to act toward ends they haven’t chosen, and second, it can lead them to act for reasons not authentically their own” [30, p. 9], could be disregarded. The stakeholders can commit themselves to system strategies that serve them and agree to them in specific contexts, i.e., they consider them (ethically) acceptable.

The advantage is that criteria for the acceptability of influence can be established. At the same time, manipulation can be used for specific moral cases, such as manipulating someone to act against her values. The former is an ethical issue from a system perspective, whereas the latter is a moral one and within personal ethics. The next step will be to examine an attempt to understand system strategies (“mutual theory of mind”). However, it will be shown that such an approach is insufficient as insofar as the systems are understood as intentional agents.

3 Aligning expectations

3.1 Mutual theory of mind in human-computer interaction

The ontological question of whether machines can possess consciousness is rather speculative. Nevertheless, it might be that users perceive machines as conscious. A study demonstrated “that people already ascribe a degree of consciousness to existing technologies” [39] Furthermore, in a Theory of Mind task, where participants had to describe something to a listener, here a virtual agent, it has been found that the participants adapted “their response by 30% when conducting the Theory of Mind task with a virtual agent compared to doing the task alone” [40, p. 6]. Nevertheless, it is rather one-sided ascription of consciousness than a mutual theory of mind.

Aligning expectations of systems like CA and humans has been discussed by Wang et al. under a “mutual theory of mind.” It should allow “smoother human-CA conversations” by adapting the mutual theories or models of the mind [10]. However, it is not so clear what the goal of Wang et al. is. It seems like they want to confront the high expectations of users so that they adjust (and maybe lower) their expectations of the system functioning [9] because, as studies suggest, users have high expectations of CAs [41, 42] and this can lead to disappointments or to stop using such systems. It would thus be essential to design systems that use affordance cues to clarify the system’s capabilities [43]. According to the study of Wang et al., a personalized and better-adapted functioning (in their study: better answers) seems to correlate with an assessment of the system as presenting a more human-like and intelligent picture of the system [10]. Of course, better answers can come along with more benevolent attitudes. Furthermore, they infer that someone with a more verbose language might “confuse” the CA with complex inputs, so it might fail to deliver “supportive or efficacious responses, leading to undesirable CA perception” [10]. However, someone with a broader vocabulary would probably not consider a CA an adequate interlocutor and thus have an undesirable perception of the CA. Assuming that either the CA has a theory of mind for the human user or the other way around seems inadequate because it is mainly based on stereotypes (heuristics) that each one creates.

Interestingly, such heuristics can also be biased, as described by Eicher et al., who start with the assumption that the student “has a theory of mind for how the computer operates.” However, common “misconceptions,” for instance, in variable assignments in programming, can be seen as systematic departures or biases about the functioning of the computer. They identify a set of common misconceptions to develop a “tool […] to give the computer the ability to develop a theory of mind about the student and to use that in collaboration with the student to correct misconceptions" [11]. It might be doubted whether something like this can be interpreted as a (mutual) theory of mind where the tool develops a model for the student, and the student has a (sometimes biased) model of the computer. Instead, it is the assignment of a stereotype to the user/student and offering a recommendation or helpful explanation. By having an actual theory of mind, the student can also, to a certain extent, exploit the strategic functioning of the program to obtain instant feedback without a more extensive quarrel with her mistakes.

This could be extended by assuming that the system or the virtual CA has adaptational strategies, too. Hence, future studies in the field of ethics of Generative AI should also consider game theory to investigate how humans and systems employ strategies to get what they want, need, and seek depending on their motivations or programmed structures. This point also depends on the design of the chosen architectures.

Some literature about strategic behavior and a theory of mind for technological systems is located within the fields of human-robot or robot-robot interaction, where, for instance, one robot predicts the movement of another robot [44]. One study is based on a version of the prisoner's dilemma where humans interact with robots [45]. Furthermore, it has been investigated how a shared plan between a robot and a human can be executed without too many disturbances by the robot. Such interruptions by the robot could potentially lead to annoyance on the side of the human. For that purpose, the robot monitors the human agent and builds some kind of theory of mind [46]. However, such attempts to suppose a theory of mind remain within operational or strategic behavior. So, although the theory of mind covers the ability to "understand" shared plans, strategies, or predicting movements, it is far away from a theory of mind. Different topics are usually discussed in the philosophy of mind and are connected with the concept of mind, like, for instance, intentions, feelings, consciousness, critical judgment, etc. These abilities or characteristics are usually ascribed to human agents but not to technological systems. Scott et al. also state, "Theory of Mind is a specialist term." In their survey of 100 people about "conceptualizations of or attitudes towards machine minds," they found that "not one participant referred to Theory of Mind" [39, p. 5]. Furthermore, human agency is not limited to strategic or reward hacking behavior to find shortcuts. Human agents can commit themselves to a certain behavior or propositional content, understand their behavior's (moral) consequences, and give reasons for it. They are able to play a normative-deontic game [47].

Large language models are also used to assess human interpretation of robot behavior [48]. Deshpande and Magerko try to use "social cognition" and the "Observable Creative Sensemaking (OCSM) framework" to improve human-system interactions [49, p. 4]. They even use Generative AI, particularly image generation, as a "speculative case study". They state that the integration of such a framework "into Midjourney" would enable "AI to participate actively in the creative process. The model becomes more adept at interpreting creative prompts, engaging in a dynamic dialogue with users, and producing images that are not only contextually relevant but also aligned with the user's evolving creative journey" [49, p. 5]. Through "dialogue with users", "user feedback", and the generation of "multiple image options" a better understanding of human interactions with systems could be achieved [49, p. 5]. However, Deshpande and Magerko do not reflect the ethical dimension of all these issues and how, already in the developmental phase, human-system interaction could be made more transparent. Thus, feedback while using the system remains more within the co-action paradigm and might still be rather opaque for the user.

3.2 Aligning expectations in cognitive psychology

Ascribing systems a theory of mind (either to have a mind or to understand the mind of humans) is thus just a metaphorical way of speaking, and so is to speak of cooperation between humans and technological systems. How humans align and develop to have expectations has been investigated empirically in cognitive psychology. We can switch "perspectives as needed indefinitely," and we have an "ability" that is "involved in many aspects of shared intentionality": "recursive mindreading or recursive intention-reading"[3] [50, p. 96/97]. However, there is also a normative layer that must be reflected to fully understand expectations:

This "recursive intention reading transforms everything: turning helping and sharing into mutual expectations or even norms of cooperation." Through these mutual expectations, we bind ourselves to norms. That is why it is "turning practical reasoning into cooperative reasoning" [50, p. 106].

In sociology, Niklas Luhmann introduced the term expectations of expectations (*Erwartungserwartungen*) [51]. This recursivity is integral to understanding us as intentional and cooperative agents. A technological system cannot have

[3] "We thus proposed that the basic cognitive skill of shared intentionality– recursive mindreading– arose as an adaptation for collaborative activity specifically (given an initial adaptation in the direction of tolerance and generosity with food), leading to the creation of joint attention and common ground. The combination of helpfulness and recursive mindreading led to mutual expectations of helpfulness and the Gricean communicative intention as a guide to relevance inferences" [50, p. 217/218].

expectations. Certain autonomous systems might be interpreted as having intentions, but they cannot be understood as cooperative agents. They have certain dispositional features that let them react in a certain way [51]. The system might be *able* to choose means and ends, which could be described as reactions of dispositions they possess. There are specific possible means or ends available and form a realm or space of possibilities that can be expressed in counterfactual statements: "If systems S would have chosen means M, it would have resulted in output O." Dispositions of intelligent systems express if they are reliable and make "good" discriminations or inferences, "*ranges of counterfactual robustness*" [52, p. 103]. We humans are discursive and normative beings. We make commitments, and we can challenge them or give reasons for them in the form of, for instance, principles, theories, and factual statements [47, 52]. We operate in a "space of reasons"[4] [53].

4 User modeling

User expectations can be obtained by collecting data from users, tracking them, gathering information from online reviews [54], and (automatically) inferring their expectations from semantic contexts [55]. Such implicit feedback is located more in a behavioral framing. This is also the case for "behavioral evaluations such as A/B tests measuring user activity instead of surveys or ethnographic analyses that capture users' experience and perception of the system" [56, p. 221]. Explicit feedback can also be obtained via participatory user studies [57] or surveys. Furthermore, user-system negotiations via dialogues with (Generative) AI might present new ways to engage with users [49]. Technological systems can also create models of users, for instance, through generalization. In doing so, the user expectations can be met or disappointed. A movie recommendation system, for example, can meet expectations by being optimized via implicit or explicit feedback [58, 59]. The systems (based on a neural network, for instance) cannot establish expectations but only (weighted) distributions of probabilities. Consequently, in human-system co-actions, no expectations can be aligned like it is possible between humans. However, expectations play a crucial role in the difference between the expected outcome (inputted prompt) and the realized outcome (generated picture, text).

Someone who paints a landscape picture has an idea in her head what she wants to paint.[5] Nevertheless, it depends on the acquired abilities to paint how close one comes to the imagined picture. Generative AI offers here new forms of amplifying or substituting human abilities. Painting a picture can also be accompanied by disappointment. Still, an experienced painter or someone else (a teacher) could at least offer advice to improve the picture and advise on which abilities need to be trained to paint a better picture. However, when using Generative AI, it is mostly not clear what would lead to disappointment; for instance, the disappointment of the expectation could be based on (1) systemic strategies (determined by the developers), for example, for predictions ("synthetic users" [60], user personas), preset rules (affirmative action policies), etc., (2) the coordination via third parties ("stereotypes" [61], "anonymous communitization") [60, 62, p. 153], (3) learning or adaptive algorithms (machine learning, neural networks), (4) an inaccurate use [27], (5) implicit or (6) explicit feedback [58, 59]. A system that generates pictures might have an implemented affirmative action policy that outputs for the prompt "nurse" 50% male nurses and 50% female nurses, even though the system was trained mainly with pictures from female nurses. The clustering of users by some characteristics that they share is an example of the second form, where it is mostly not transparent for the user how the system clusters and which features are crucial to, for example, recommend a specific movie. An inaccurate use might be an incomplete or unspecific prompt. Implicit feedback is obtained by monitoring the user, and explicit feedback is received by likes from the user or, more elaborately, by research (for instance, acceptance studies).[6]

Assessing the systems might now seem hopeless because the algorithms are mostly black boxes in Generative AI models. However, the alignment of prompted input and realized output could also be systematically examined to understand how the system works. Text or pictures must be generated and categorized preliminary into the six types above to examine the system. This needs to be further investigated empirically to describe human-system co-action systematically. If the expectations are disappointed, the types can only be inferred abductively from the effects (output). Empirical research is needed to assess the outputs and to understand under which type the specific output is realized. Some use cases and examples will be shortly discussed in the next section as a first step toward more systematic empirical work.

[4] This is also why choosing adequate means [29, p. 8/9], as dispositional systems do, cannot be manipulative but only if a normative commitment is incorporated.

[5] Generative AI can also help generate ideas in a co-creation process [71]. I am thankful to an anonymous reviewer for indicating this issue.

[6] Generative AI can also be used to improve system transparency by introducing new ways of negotiations between users and systems. (I am thankful to an anonymous reviewer for pointing this out.) This issue is also discussed in [48, 49]. I believe that this is certainly a way to approach the ethical issues surrounding Generative AI. Nevertheless, the aim is here not so much to enhance system transparency but rather to make the practice of developing Generative AI more transparent.

5 Ethical implications– use cases

The use of Generative AI poses ethical challenges: For instance, a prompt like “CEO of a big company” outputs mainly pictures of male managers and “nurse” only female nurses (webpages used: davinci.ai/ and craiyon.com). The conception of fairness is not reflected. This might lead to technology resembling reality (as unjust as it might be). Of course, which measurements should be taken in these cases is also unclear. Should the output be randomized, should an affirmative action policy via a fixed rule be implemented (e.g., to output 60% female managers), or should the output mirror an (unjust) reality? Yet, conceptions of fairness, such as affirmative action policies, are also limited in such contexts. They can lead to the generation of absurd pictures, as has been done recently by the Google program Gemini that outputted indigenous people in German World War Two uniforms [63]. The ethical principles behind such technologies must be made explicit and discussed, not only in the offices of tech companies but also in exchange with the broader public.

In the health sector, prioritizing patients for referral to specialists is crucial. This is partly done by using natural language processing (NLP). However, Generative AI also has innovation potential, where a chatbot might function as an assistant to the physician [64, 65]. The prioritization is based on increasing efficiency and effectiveness. Nevertheless, such systems can also systematically discriminate against minorities, as has been discovered in studies [66]. The normative frame, i.e., here, conceptions of fairness, must be discussed and operationalized; otherwise, the conception of prioritization solely relies on the technological categories of efficiency and effectiveness.

Values such as fairness are abstract in the sense of meaningless if they are not connected with other notions like moral desert, equal shares, and randomization. These connections are commitments that have conceptual content that can be challenged. Awad et al. discuss the example of prioritization for kidney transplants. According to them, prioritization could be formalized in descriptive ethics via studies of the preferences or intuitions of laypeople regarding the case to extract features that matter. Furthermore, they claim that the means of prioritization could be based on the notion of the “general health” of the patients [67, p. 391]. However, the value of fairness is conceptualized via a notion of moral desert, i.e., one deserves a kidney over another person based on a specific feature (may it be acquired through effort (consistent dietary) or purely accidentally, such as age). Other ways of conceptualizing fairness, such as implementing a rule (waiting list (“first come, first served”)) or randomization [68], are not discussed. The ethical task is discussing different strategies that can be used and how they are justified. The decision on which strategy is used is, in the end, not ethical anymore but pragmatic-moral within the context of a specific political system and culture. Awad et al. propose a computational approach to ethics that presupposes already a specific conceptual content for the value of fairness [67]. Algorithmizing or formalizing normative ethics means either playing a formal game with well-defined formulas where no justification can be sought or presupposing a specific notion that is not questioned.

Another example of using Generative AI is the personalization of learning environments to generate tailor-made learning paths and exercises for students [69]. In particular, the system strategies for personalization, but also the normative background of the concept, would have to be further questioned since, on the one hand, it would have to be shown in a technically transparent way under which strategies the systems are adapted to the users and, on the other hand, to what extent personalization in education makes sense. Even though it promises a fairer education due to adaption to specific learning needs, personalization– and technology in general [70]– reduce resistances that are essential for learning and are actually needed for the development of competencies or skills.– In all examples, a clear reflection and discussion of the ethical and normative framework is lacking. Participatory research is essential here in establishing how values should be conceptualized and which conceptualization stakeholders consider (ethically) acceptable. Such investigations of the acceptability of technological systems can help to prevent such systems from being manipulative. Although they might influence the users, a deliberative process would have been established to find and choose acceptable conceptualizations.

6 Conclusion

Seemingly, Generative AI could help to make technological systems more transparent because it can provide feedback or interact with users via dialogues. For this reason, it has been proposed to implement a theory of mind in systems. However, it has been shown that this conceptualization has several shortcomings. Already in the developmental phase, it is crucial to align the expectations of developers and users because the user is influenced or manipulated through the mediating role of technology. Even if the developers tried not to intentionally deceive the user via, for instance, dark patterns, some kind of non-intentional manipulation might be the consequence of the system implementation. Hence, in the process of human-system co-action, this kind of influence or manipulation might already be taking place and, furthermore, be opaque to the users and, even worse, not in their interest.

Systems cannot form expectations to conform to users' interests, goals, or values. In the end, it is only possible to align expectations between humans: developers and designers on the one side and users on the other side. The discussed use cases showed that it is crucial to understand, for instance, which conception of fairness is and can be implemented in a system. For this reason, the preferences and values of the stakeholders need to be explored. Which conception of fairness do they favor? And how can we democratize Generative AI?

Much work remains to be done before a full understanding of the ethics of Generative AI is established. In terms of future research, it would be crucial to extend the present discussions by examining the operationalization of ethical values, such as fairness, privacy, etc., in technological systems. Furthermore, it needs to be investigated how systems strategies can be made transparent to users so that users also can find counter-strategies or align their expectations. If it is possible to implement different conceptions of values, then there is a need for empirical research that explores the preferences of the stakeholders and which conception they would favor. This implies converting or transforming human-system co-action to a more transparent human-system interaction framework.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Verbeek, P.-P.: Beyond interaction: a short introduction to mediation theory. Interactions **22**(3), 26–31 (2015)
2. Rosenberger, R., Verbeek, P.-P.: A Postphenomenological field guide. In: Rosenberger, R., Verbeek, P. (eds.) Postphenomenological Investigations: Essays on Human-Technology Relations, pp. 9–41. Hrsg., New York/London (2015)
3. van de Poel, I., Verbeek, P.-P.: Can technology embody values?. In: Kroes, P., Verbeek, P. (eds.) The Moral Status of Technical Artefacts. Philosophy of Engineering and Technology, vol. 17, pp. 103–124, Heidelberg/New York/London (2014)
4. Krügel, S., Ostermaier, A., Uhl, M.: ChatGPT's inconsistent moral advice influences users' judgment. Sci. Rep. **13**, 4569 (2023)
5. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I.: The moral machine experiment. Nature **563**(7729), 59–64 (2018)
6. Krügel, S., Uhl, M.: Autonomous vehicles and moral judgments under risk. Transp. Res. Part A: Policy Pract. **155**, 1–10 (2022)
7. Karpus, J., Krüger, A., Tovar Verba, J., Bahrami, B., Deroy, O.: Algorithm exploitation: humans are keen to exploit benevolent AI. iScience **24**(6), 1–16 (2021)
8. Schönmann, M., Bodenschatz, A., Uhl, M., Walkowitz, G.: The care-dependent are less averse to care robots: an empirical comparison of attitudes. 15, (2023). Int. J. Soc. Robot. **15**, 1007–1024 (2023)
9. Hubig, C.: Die Kunst des Möglichen III: Macht der Technik, Bielefeld (2015)
10. Wang, Q., Saha, K., Gregori, E., Joyner, D., Goel, A.: Towards mutual theory of mind in human-AI interaction: how language reflects what students perceive about a virtual teaching assistant. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp. 1–14 (2021)
11. Eicher, B., Cunningham, K., Gonzales, S.P.M., Goel, A.: Toward mutual theory of mind as a foundation for co-creation. In: International Conference on Computational Creativity, Co-Creation Workshop (2017)
12. Ropohl, G.: Technikbegriffe zwischen Äquivokation und Reflexion. In: Banse, G., Grunwald A. (eds.) Technik und Kultur: bedingungs- und Beeinflussungsverhältnisse, pp. 41–54, Hrsg., Karlsruhe, (2010)
13. VDI, Richtlinie 3780: Technikbewertung, Begriffe und Grundlagen, Berlin (2000)
14. Pitt, J.C.: "Guns Don't Kill, People Kill"—values in and/or around technologies. In: Peter, K., Verbeek, P. (eds.) The Moral Status of Technical Artefacts. Philosophy of Engineering and Technology, vol. 17, pp. 89–101, Dordrecht (2014)
15. Coeckelbergh, M.: Moved by Machines: Performance Metaphors and Philosophy of Technology, New York (2019)
16. Hubig, C.: Die Kunst des Möglichen I: Technikphilosophie als Reflexion der Medialität, Bielefeld (2006)
17. Grunwald, A., Julliard, Y.: Technik als Reflexionsbegriff: Überlegungen zur semantischen Struktur des Redens über Technik. Philos. Nat. **42**, 127–157 (2005)
18. Moore, J.: Four kinds of ethical robots. Philos. Now **72**, 12–14 (2009)
19. Köbis, N., Bonnefon, J., Rahwan, I.: Bad machines corrupt good morals. Nat. Hum. Behav.Behav. **5**(6), 679–685 (2021)
20 Krügel, S., Ostermaier, A., Uhl, M.: Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. Philos. Technol. **35**, 17 (2022)
21 Krügel, S., Ostermeier, A., Uhl, M.: Algorithms as partners in crime: a lesson in ethics by design. Comput. Hum. Behav.. Hum. Behav. **138**, 107483 (2023)
22. Leib, M., Köbis, N., Rilke, R., Hagens, M., Irlenbusch, B.: Corrupted by algorithms? how ai-generated and human-written advice shape (dis)honesty. Econ. J. **134**(658), 766–784 (2024)
23. Robinette, P., Li, W., Allen, R., Howard, A., Wagner, A.: Overtrust of robots in emergency evacuation scenarios. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 101–108 (2016).

24. Thaler, R.H., Sunstein, C.R.: Nudge: The Final Edition, New Haven/London (2021)
25. Thaler, R.H., Sunstein, C.R.: Nudge: The Final Edition, Penguin Books (2021)
26. Klincewicz, M.: Artificial intelligence as a means to moral enhancement. Stud. Log. Gr. Rhetor. **48**(1), 171–187 (2016)
27. Hubig, C.: Die Kunst des Möglichen II: Ethik der Technik als provisorische Moral, Bielefeld (2007)
28 Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L.: The ethics of algorithms: key problems and solutions. In: Floridi, L. (ed.) Ethics, Governance, and Policies in Artificial Intelligence, pp. 97–123. Springer (2021)
29 Klenk, M.: Ethics of generative AI and manipulation: a design-oriented research agenda. Ethics Inf. Technol. **26**, 9 (2024)
30 Susser, D., Roessler, B., Nissenbaum, H.: Technology, autonomy, and manipulation. Internet Policy Rev. (2019). https://doi.org/10.14763/2019.2.1410
31 Sunstein, C.: The Ethics of Influence: Government in the Age of Behavioral Science. Cambridge University Press (2016)
32. Noggle, R.: Pressure, trickery, and a unified account of manipulation. Am. Philos. Q. **57**(3), 241–252 (2020)
33. Noogle, R.: The ethics of manipulation. In: The Stanford Encyclopedia of Philosophy, Summer 2022 Edition
34. Bergram, K., Djokovic, M., Bezençon, V., Holzer, A.: The digital landscape of nudging: a systematic literature review of empirical research on digital nudges. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI'22), pp. 1–16 (2022)
35. Valta, M., Menzel, J., Maier, C., Pflügner, K., Meier, M., Weitzel, T.: Digital nudging: a systematic literature review and future research directions. In: Proceedings of the 2022 Computers and People Research Conference (SIGMIS-CPR'22), pp. 1–10 (2023)
36 Nie, L., Zhao, Y., Chenglin Li, C., Lu, X., Liu, Y.: Shadows in the interface: a comprehensive study on dark patterns. Proc. ACM Softw. Eng. **1**(FSE), 204–225 (2024)
37 Gray, C.M., Chen, J., Chivukula, S.S., Qu, L.: End user accounts of dark patterns as felt manipulation. Proc. ACM Hum.-Comput. Interact **5**(CSCW2), 1–25 (2021)
38. Sánchez Chamorro, L., Bongard-Blanchy, K., Koenig, V.: Ethical tensions in UX design practice: exploring the fine line between persuasion and manipulation in online interfaces. In: Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS'23), pp. 2408–2422 (2023)
39. Scott, A.E., Neumann, D., Niess, J., Woźniak., P.W.: Do you mind? User perceptions of machine consciousness. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23), pp. 1–19 (2023)
40. Heyselaar, E., Bosse, T.: Linking theory of mind in human-agent interactions to validated evaluations: Can explicit questionnaires measure implicit behaviour?. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA'21), pp. 120–127 (2021)
41. Luger, E., Sellen, A.: "Like having a really bad PA": the gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16). Association for Computing Machinery, pp. 5286–5297 (2016)
42. Zamora, J.: "I'm sorry, dave, I'm afraid I can't do that: chatbot perception and expectations. In: Proceedings of the 5th International Conference on Human Agent Interaction (HAI'17). Association for Computing Machinery, pp. 253–260 (2017)
43. Liao, Q.V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M.P., Wang, D., Muller, M., Shami, N.S., Geyer, W.: All work and no play?. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18). Association for Computing Machinery, pp. 1–13 (2018)
44. Kim, K.-J., Lipson, H.: Towards a simple robotic theory of mind. In: Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems (PerMIS'09), pp. 131–138 (2009)
45. Hegel, F., Krach, S., Kircher, T., Wrede, B., Sagerer, G.: Theory of mind (ToM) on robots: a functional neuroimaging study. In: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI'08), pp. 335–342 (2008)
46. Devin, S., Alami, R.: An implemented theory of mind to improve human-robot shared plans execution. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI'16), pp. 319–326 (2016)
47 Brandom, R.: Making It Explicit: Reasoning, Representing, and Discursive Commitment. Harvard University Press (1994)
48. Verma, M., Bhambri, S., Kambhampati, S.: Mudit Verma, Sidhant Bhambri, and Subbarao Kambhampati. theory of mind abilities of large language models in human-robot interaction: an illusion?. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI'24), pp. 36–45 (2024)
49. Deshpande, M., Magerko, B.: Embracing embodied social cognition in AI: moving away from computational theory of mind. In: Extended abstracts of the 2024 CHI conference on human factors in computing systems (CHI EA'24), pp. 1–7 (2024)
50 Tomasello, M.: Origins of Human Communication. The MIT Press (2008)
51. Hubig, C.: Verschmelzung von Technik und Leben? Begriffserklärungen an der Schnittstelle von Mensch und technischem System. In: Herzberg, S., Watzka, H. (Eds.) Transhumanismus: Über die Grenzen technischer Selbstverbesserung, pp. 145–160, Hrsg., Berlin/Boston (2020)
52 Brandom, R.: Between Saying and Doing. Oxford University Press (2008)
53. Sellars, W.: Empiricism and the Philosophy of Mind, 4 Hrsg., Cambridge (Mass.)/London (2003)
54. Svikhnushina, E., Placinta, A., Pu, P.: User expectations of conversational chatbots based on online reviews. In: Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS'21), pp. 1481–1491 (2021)
55. Hotzkow, J.: Automatically inferring and enforcing user expectations. In: Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2017), pp. 420–423 (2017)
56. Ekstrand, M.D., Willemsen, M.C.: Behaviorism is not enough: better recommendations through listening to users. In: Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16), pp. 221–224 (2016)
57. Park, S., Lim, Y.-K.: Investigating user expectations on the roles of family-shared AI speakers. In: Proceedings of the 2020 CHI Conference on human factors in computing systems (CHI'20), pp. 1–13 (2020)
58. Jawaheer, G., Szomszor, M., Kostko, P.: Comparison of implicit and explicit feedback from an online music recommendation service. In: Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec'10). Association for Computing Machinery, New York, pp. 47–51 (2010)
59. Zhao, Q., Harper, F.M., Adomaviciu, G., Konstan, J.A.: Explicit or implicit feedback? engagement or satisfaction? a field experiment on machine-learning-based recommender systems. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC'18). Association for Computing Machinery, New York, pp. 1331–1340 (2018)
60. Johnson, A., Taatgen, N.: User Modeling. In: Handbook of human factors in Web design, Lawrence Erlbaum Associates, pp. 4244–39 (2005)

61. Rich, E.: Stereotypes and User Modeling. User Models in Dialog Systems. Symbolic Computation (1989)
62. Hubig, C.: Virtualisierung der Technik—Virtualisierung der Lebenswelt. In: Gethmann, C.F.(Ed.) Lebenswelt und Wissenschaft: XXI. Deutscher Kongreß für Philosophie, pp. 146–159, Hrsg., Hamburg (2011)
63. Google pauses AI-generated images of people after ethnicity criticism. The Guardian (2024)
64 Saqib, M., Iftikhar, M., Neha, F., Karishma, F., Mumtaz, H.: Artificial intelligence in critical illness and its impact on patient care: a comprehensive review. Front. Med. (Lausanne) **1**, 1176192 (2023)
65. Pessach, I., Shaked, O., Lipsky, A., Zeevi, A., Lilly, C., Blum, J.: Focusing advanced clinicians on the more critically ill patients using artificial intelligence. Crit. Care Med. **48**(1), 177 (2020)
66. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
67. Awad, E., Levine, S., Anderson, M., Anderson, S.L., Conitzer, V., Chrockett, M., Everett, J.A., Evgeniou, T., Gopnik, A., Jamison, J.C., Kim, T.W., Liao, S.M., Meyer, M.N., Mikhail, J., Opoku-Agyemang, K., Schaich Borg, J., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., Tenenbaum, J.B.: Computational ethics. Trends Cogn. Sci.Cogn. Sci. **26**(5), 388–405 (2022)
68 Broome, J.: Fairness. In: *Proceedings of the Aristotelian Society, New Series*, vol. 91, pp. 87–101 (1990/1991)
69. Freier, C., Bocklet, T., Helte, A.-K., Hoffmann, F., Hunger, M., Kovács, L., Richter, F., Riedhammer, K., Schmohl, T., Simon, C.: Wie kann videogestütztes Lernen die Erwartungen Studierender und Dozierender erfüllen? Soziale Passagen **15**(2), 631–635 (2023)
70. Ortega y Gasset, J.: Betrachtungen über die Technik. In: Gesammelte Werke, vol. IV, Stuttgart, pp. 7–69 (1978)
71 Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B., Lu, Z.: It felt like having a second mind: investigating human-AI co-creativity in prewriting with large language models. Proc. ACM Hum. Comput. Interact. **8**(CSCW1), 1–6 (2024)

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