Causes and Reasons – Decisions, Responsibility, and Trust in Techno-Social Interactions

Larissa Ullmann, Philosophy of Technology, TU Darmstadt- Research Training Group KRITIS (Germany)

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: AI ethics, trust in AI, techno-social interactions

Abstract

The interaction between humans and AI creates a new type of interaction that goes beyond subject-object relations. AI technologies cannot always be described as a conventional object due to its activity capabilities and the black box aspect. An additional category is created, which is outlined by the *sobject approach*. This creates the opportunity to study the human-like characteristics of the interaction on the part of the AI. The *social* possibilities of AI can thus be focused by referring to *techno-social* rather than *social* interactions, since the possibilities are different from human sociality, but exist in the human-social lifeworld. If an AI is a techno-social interaction partner, it can *act* and make *decisions*. The additional category can therefore be used to investigate what types of decisions there are, if they are based on *reasons or causes*, whether they can be *trusted*, and if one can assign or delegate *responsibility* to such technology. Thus, classical ethical questions regarding subjective categories like *decision-making, trust* and *trustworthiness*, and *responsibility* can be rethought for somewhat human-like but not human technologies like AI.

Introduction

Human-AI interaction is becoming more pervasive, and with it, a new phenomenon of interaction that is neither a subject-subject nor a subject-object interaction. This other type of interaction needs to be explored in order to ground the discourse on *technological ethical attributes*. The philosophy of technology is required to bridge technological artefacts and the human social *lifeworld* (*Lebenswelt*) and make room for something *human-like* that is not human. If this succeeds, it will qualify and help circumvent the problem of the uncanny since the worry about the *Uncanny Valley* presupposes the dichotomy of the human subject and technical object. With the notion of a third type of interaction, it will become possible to question what types of *social* relations there are beyond subject-subject relations and how to understand them. The recognition of this third type is already visible in different discussions, in literature, in movies as well as in philosophical research. It is mainly about a change from controllable to a new kind of technology as in the contrast of c*onventional* and *advanced*[1] or *classical* and *trans-classical*[2] *technology*, *trivial* and *non-trivial machines*,[3] or a *digital*[4]- or *quasi-other*[5]. These terms often deal with a kind of technology that is closely compared to human characteristics, and it remains questionable how to frame this phenomenon and deal with it in terms of human sociality.

[1] Weyer, J.: *Die Kooperation menschlicher Akteure und nicht-menschlicher Agenten. Ansatzpunkte einer Soziologie hybrider Systeme* [The cooperation of human actors and non-human agents. Approaches to a Sociology of Hybrid Systems], 16. Soziologisches Arbeitspapier 2006, p. 16–21.

[2] Kaminski, A.: »Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen« [Giving reasons. Machine Learning as a Problem of the Morality of Decisions], in: Wiegerling, K.; Nerurkar, M.; Wadephul, C.: *Datafizierung und Big Data. Ethische, anthropologische und wissenschaftstheoretische Perspektiven*, Springer 2020, p. 151–174.

[3] Von Foerster, H.: »Principles of Self-Organization – In a Socio-Managerial Context«, in: Ulrich, H.; Probst, G.: *Self-Organization and Management of Social Systems. Insights, Promises, Doubts, and Questions*, Springer 1984, p. 2–24.

[4] Liberati, N.: »Being Riajuu. A Phenomenological Analysis of Sentimental Relationships with ›Digital Others‹«, in: Cheok, A.; Levy, D.: *Love and Sex with Robots*, Springer 2017, p. 13–23.

[5] Coeckelbergh, M.: »You, robot: on the linguistic construction of artificial others«, in: *AI & SOCIETY*, vol. 26, 2011, p. 61–69.; Ullmann, L.: »The quasi-other as a Sobject«, in: *Technology and Language*, 3(1), p. 76–81.

The approach of the *sobject* describes this third type of *social* interaction partner from a phenomenological point of view as something between subject and object. It helps create a space for discussions about interactions with human-like technologies and does not constantly compare them with human ones since it captures the phenomenon sui generis. Therefore, it provides a qualified base for further analysis of human-AI interactions and their technological-ethical attributes; thus, one can focus on human-sobject as well as subject-object relations. It becomes possible to explore *social* interactions with and through technologies. But since technology cannot be social in the human sense – due to a lack of intentionality and subjectivity, in particular – and since some kind of social interaction can nevertheless occur, one might speak here of the possibilities of *techno-social* instead of *social* interaction. This distinction opens the discussion for human-like but not human *social* interactions. Thereby a particular focus on those phenomena arises without the direct comparison to human sociality respectively, it prevents the necessity to talk about the sociality of AI in the human sense. As a result, there is a suitable basis to discuss the essential ethical attributes that arrive with techno-social activities in the human environment – *Decisions*, *responsibility*, and *trust*. Since AI is, in a way, able to make *techno-decisions,* it needs to be explored how to evaluate them ethically, which refers closely to responsibility, trust, and the question about causes and reasons. Hence, the primary purpose of this paper is to set the frame for discussing the ethical concern within the AI context in general. Through the investigation of those elements in AI responses that have a social reference but nevertheless appear uncanny to us despite their similarity with human behavior, we reach a different, perhaps new type of interaction: a techno-social interplay. Assuming in this interplay the central role played by language and linguistic formulation, I propose to take the sobject approach as the fundamental phenomenological method to distinguish between social and techno-social relations to build the proper context for discussing the essential attributes in human-AI interaction.

Phenomenological basis: The Sobject

The fundamental base of the argumentation in this paper is the sobject approach, which I introduced to create a suitable space for the discussion of a phenomenon that has already occurred but through a lack of concrete linguistic terms is not a part of our consciousness, with many problems arising that are primarily due to the treatment of technical sociality in the terms exclusively of human subject-subject or subject-object sociality. From a phenomenological perspective, it offers a suitable concept for further techno-philosophical fields like ethical aspects. It is based on an approach from Simone de Beauvoir[6] that represents a fruitful techno-philosophical discussion about subject-object relations. She describes the performance of the subject as twofold: it consists of both *positing* itself and *positing* the objects. In contrast, objects cannot posit themselves but are instead posited by a subject. Thus, objects are passive, and subjects are active. This means that in subject-subject relations, every side sets the other as *the other*, and in this reciprocal act of positing both sides are active. While in subject-object relations only the object can be set as *the other* by the subject, and due to its passivity, the object cannot posit the subject. This results in interrelations as well as action-reaction relationships among subjects that are not possible with objects. Any effect that comes from an object would, in fact, be the result of subjective positing. Beauvoir's distinction seems elementary, but also includes dialogues with different phenomenological and existentialistic thinkers, like Sartre, Merleau-Ponty, or Levinas.[7] Therefore, it provides an ideal starting point for research on subject-object relations.[8]

[6] De Beauvoir, S.: *The Second Sex* (C. Borde, & S. Malovany-Chevallier, trans.), Vintage Publishing 2015 (Originally published 1949).

[7] Ullmann, L.: »Das Sobjekt: Mögliche Beziehungen zwischen Mensch und Maschine aus einem phänomenologischen Blickwinkel« [The Sobject: Possible Relations between Humans and Machines from a Phenomenological Point of View], in: *Jahrbuch Technikphilosophie 2022*, *vol. 8*. Nomos 2022, p. 196–198.

[8] Ibid., p. 195–213.

We are in fact at a time in technological development where the sharpness of that distinction is being challenged. Some kinds of technologies that are defined by the characteristics of sobjects are neither a traditional object nor a subject since they can enter into more substantive relations with a subject than an object can. As a result, we have a relationship between a subject and this particular kind of technology that seems human-like and in many respects equals subject-subject relations, but it is, in fact, neither a proper subject nor a mere object. Johannes Weyer describes these kinds of technology as *advanced* compared to *conventional technologies.* While classical technologies could be used as an instrument and fit the understanding of a conventional object, the advanced ones can't completely be instrumentalized since their actions are not fully transparent. This is especially due to emergent structures, as in AI, where even the human involved in the interaction cannot accurately predict what the system's response will be. Weyer goes so far as to say that he sees a danger in those ways of interaction because the advanced technologies are becoming so independent that they are changing the forms of relationships: humans become more object-like while the technology gains more subjectivity. This danger would materialize as the assignment of a passive role to humans and their subsequent instrumentalization. Weyer also argues that classical technologies were like a *teammate* (*Mitspieler*) in the interaction with humans, whereas the advanced ones are already superordinate and no longer a teammate in this interaction.[9] On the contrary, I would rather argue that we have a teammate situation only in the advanced technology-human interaction since neither party can fully see through the actions of the other. Only if we assume this starting situation, in which one party is not completely predetermined (i.e. completely posited by the other), the phenomenon of the sobject becomes visible and it seems fruitful to analyze this as a new way of interaction. Instead of seeing them only as a danger and an impoverishment for the human side.[10] Andreas Kaminski also differentiates between those two kinds of technology and speaks about *classical* and *trans-classical technologies*. He argues for a similar distinction since the classical type fits the basic understanding of objects, whereas the trans-classical adds something more that cannot fit into the previous cluster. But he doesn't focus on the dangerous aspect nor on the implication that the new, intelligent kinds of technologies will end up in objectivizing the human. Kaminski also highlights some aspects of a human-AI interaction (or trans-classical technology interaction) like the mutual opacity and the differences between practical and epistemic understanding.[11] But even when Weyer and Kaminski introduce similar distinctions and describe ways of interaction in general, it is always about a new human-like kind of technology and its interaction with humans. In particular, it is about emergent structures like AI and the rising aspect of non-transparency. Kaminski calls this a black box that can't be turned into a white one because of a lack of epistemic understanding, while the interaction of humans with classical technologies revolves around a white box, fully clear and understandable.

There seems to exist a new kind of intelligent technology in which forms of interactions with humans can't be categorized due to the lack of conceptual space for something that is human-like and at the same time not human and not a subject in the traditional sense.[12] The black box aspect shapes a new type of interaction, resulting in the need to describe this phenomenon and its capabilities. With the

---

[9] Weyer, J.: *Die Kooperation menschlicher Akteure und nicht-menschlicher Agenten. Ansatzpunkte einer Soziologie hybrider Systeme,* p. 16–21.

[10] Ullmann: »Das Sobjekt: Mögliche Beziehungen zwischen Mensch und Maschine aus einem phänomenologischen Blickwinkel«, in: *Jahrbuch Technikphilosophie 2022*, p. 198–202.

[11] Kaminski: »Gründe geben«, in: *Datafizierung und Big Data,* p. 151–174.

[12] Some postphenomenological approaches are proceeding in a similar way. Peter-Paul Verbeek has dealt with technological mediations of humans and the world. From his point of view, we can see a development of human-technology interactions where technology gains a more than objective role, becoming a form of mediation (something that is neither an object nor a subject). Rosenberger, R.; Verbeek, P. (2015). »A Field Guide to Postphenomenology«, in: Rosenberger, R.; Verbeek, P.: *Postphenomenological Investigations: Essays on Human-Technology Relations*, Lexington Books 2015.

introduction of the sobject, it is possible to describe this human-like but not human interaction and focus on this new phenomenon without hypostasizing its technological character as a subject. In summary, the most relevant sobject-characteristics for our purposes are: i) the sobject is a technical artefact capable of an active performance in the sense of De Beauvoir; ii) it can posit an object or a subject; iii) it is capable of a kind of emergence[13], iv) it can't, therefore, be fully understood epistemically and represents a kind of black box, and that is why it cannot be instrumentalized and controlled entirely. Assuming those capabilities, we have an interactive partnership that is more than an object relation and shares a lot with subject interactions. We can't treat the sobject entirely as an instrument and can't transform the black into a white box.[14] The ones configured here are action-reaction relations and a phenomenon of interaction that needs to be explored. Obviously, these main characteristics of the sobject correspond to the features of AI that are most relevant and challenging for human-AI interaction and the resulting ethical challenges. These ethical issues can now be discussed in their proper context which entails a further distinction, between the *social* and the *techno-social* dimension. This difference becomes relevant insofar as it prevents the ambiguity of speaking about ethical or social aspects of AI as if these could be ethical, moral, or social in a human sense. Since this would not be appropriate, the sobject-approach permits a concrete focus on human-AI interactions, affording at the next level a specification of the techno-social.

Techno-Social instead of *Social*

In discussing AI and its presence in the human social environment, comparing AI's social capabilities with humans is often necessary. Since boundaries between the human social lifeworld and technologies are blurred, focusing on the possibilities and impacts of social ways of human-AI interaction is essential. Advanced technologies like AI already have those characteristics to interact with humans in a social way; that's why we talk about sobjects. But we have, on the one hand, human-like technologies and, on the other hand, only a human social lifeworld where this kind of interaction doesn't fit well. As the sobject approach shows, there is a lack of suitable terms, which are mandatory if we want to figure out the interaction possibilities within its ethical implications on the part of AI. That means we have this apparently new form of a technological, social interaction partner that doesn't fit our understanding of sociality or even that of an interaction partner in general. *Social* is a big term, and it is probably impossible to define it shortly and concretely. Still, on a general level, we can denote social as the domain of interactions in a community that includes subjectivity, like that of humans as ethical and moral beings. The concept of society includes responsibility for one's own actions and requires a certain degree of trust in others, their actions, and a shared moral understanding. This means the social aspect appears mainly in the interaction with other social beings and includes communal norms and rules of behavior. However, the subjective character of the members also encapsulates an inevitable degree of opacity and unpredictability, albeit within a common framework: "social norms, like many other social phenomena, are the unplanned result of individuals' interaction."[15] Clearly, society cannot be reduced to the interplay within singular agents, but it is also about interactions with social institutions or groups. In this sense, the capability of being active is not only limited toward engaging with another individual but also with collective or extra-individual formations, all of which contribute to shaping a network of relationships, that are essentially

[13] Self-learning structures

[14] The black box aspect is understood as a general existing phenomenon, even though there are different types of AI with more or less black box aspect, but this is primarily part of the explainable AI discussion.

[15] Bicchieri, C.; Muldoon, R.; Alessandro Sontuoso, A.: »Social Norms«, in: Zalta, E. N.: *The Stanford Encyclopedia of Philosophy*, 2018 Edition, URL: <https://plato.stanford.edu/archives/win2018/entries/social-norms/>.

practical. In this sense, Husserl talks about social activity in the *social lifeworld*[16], while Heidegger refers to *social practice*[17], Sartre highlights the necessity of *the other* in a social formation[18], and De Beauvoir outlines the relations between subjects and objects in society[19]. There are many different approaches and aspects regarding sociality, but especially from a phenomenological point of view, it is primarily characterized by the intentionality of the agents. The concept, as developed by Husserl, designates that consciousness is always "consciousness of or about something."[20] This means, among others, that a subject's consciousness always has a determined content, which informs its mode of operation. Ludger Jansen and Margaret Gilbert talk, for example, also about collective intentionality in groups.[21] Every social being can be active and is responsible for the consequences of its actions because of its human status.[22] The object, in the classical sense, could not have any kind of consciousness, as it was unable of intentionality: either it was a mere instrument or it was totally automatic. But now there is a human-like technical artefact that is able to interact in the same human lifeworld, as a partner, that is however not a social being in the human sense. The social technical characteristics seem prima facie social in the way we have just described. The boundaries are blurred, and a subject-like technology, such as an AI, can be socially inferred erroneously as another human being. It is however only the simulation of the characteristics of consciousness, like intentionality, but the formation is different, even if the outcome could seem similar.[23] This kind of analysis seems to be flawed from the beginning, for it distinguishes only between a human sociality and a simulation of it, in other words, a replica. The claim that something human-like has the same characteristics as a human is misleading. Moreover, the discussion comes too early to a dead end when the argument consists only of the affirmation that AI can't be social or ethical since it is not human or subjective, or intentional. Mario Martini explained, for example, that AI couldn't be social due to a lack of self-reflection, intuition, and common sense, of a human mind in general.[24] It is clear that the nonhuman can never be human and that the replica will never, by definition, be the same thing as the original. This kind of argument suffers from a *petitio principii*. Using the term techno-social could shift the debate to another level. It is not just about using

[16] E.g. Husserl, E.: *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*, Northwestern University Press 1970.

[17] Heidegger, M.: *The Basic Problem of Phenomenology* (Hofstadter, A., trans.), Bloomington & Indianapolis 1982, e.g., p. 291–294. Here, before the deeper analysis of *Being and Time*, Heidegger describes the *Dasein* as *already practical in being-with others* and its relation to equipment, objects of use, which are encountered in the world and understood first and foremost from a practical point of view, whereby each refers to its own function.

[18] E.g. Sartre, J. P.: *Existentialism Is a Humanism* (Macomber, C. trans.), Yale University Press 2007 (Originally published 1946).

[19] Ullmann, L.: »Das Sobjekt: Mögliche Beziehungen zwischen Mensch und Maschine aus einem phänomenologischen Blickwinkel«, in: *Jahrbuch Technikphilosophie 2022,* p. 196–198.

[20] „Bewusstsein von etwas", in: Husserl, E.: *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch: Allgemeine Einführung in die reine Phänomenologie* [Ideas Pertaining to a pure Phenomenology and to a Phenomenological Philosophy. First Book: General Introduction to a Pure Phenomenology], Niemeyer 1913, p. 64.

[21] Jansen, L.: *Gruppen und Institutionen. Eine Ontologie des Sozialen* [Groups and Institutions. An Ontology of the Social], Springer 2017, p. 1 – 22; Gilbert, M.: *Shared Intention and Personal Intentions*, Philosophical Studies 144 (1) 2009, p. 167–187.

[22] Adam Smith, in his *Theory of Moral Sentiments,* describes the social formation as crossed by sympathy and emotions, suggesting the image of a fictive spectator which is a part of any individual and represents a kind of shared moral conscience (Adam, S.: *The Theory of Moral Sentiments*, Penguin Classics 2010). While Kant describes the categorical imperative as a purely normative maxim of universal scope Kant, I.: Grundlegung zur Metaphysik der Sitten [Groundwork for the Metaphysics of Morals] Ed.: Valentiner, T., Reclam 2012, e.g. p. 53 ff).

[23] Misselhorn, C.: »Maschinenethik und Philosophie« [Machine Ethics and Philosophy], in: Bendel, O.: *Handbuch Maschinenethik*, Springer 2019, p. 37–39.

[24] Martini, M.: *Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz* [Blackbox Algorithm – Basic Questions of Artificial Intelligence Regulation], Springer 2019, p. 59.

another word, it indicates the capabilities of human-AI interaction in the context of the sobject without the necessity of comparing them entirely to humans and speaking only of a replica. It also creates a separate way to analyze what kind of interactions are possible outside the domain of human sociality without the constraint of humanizing them. This is why I suggest differentiating between *social* and *techno-social* forms. The techno-social is implemented in the sobject approach and opens the research of human-AI interaction on the concrete level of social and ethical aspects. A central aspect seems to be the way of simulation of subjective conditions that could appear quite similar in a concrete human-sobject-interaction. Therefore, the focus will be on a human-sobject interaction, not on the inner content (which, as in the case of other human beings, is always to some extent opaque), but on the outcome and its practical process of shaping. After all, even the classic *Turing Test*[25] showed a way to avoid a purely simulative approach and also that it is wrong to infer that a computer owns the same intentionality or way of *thinking* that a human being has[26], whereas what is relevant is whether it behaves, whether it acts as a social being.

In summary, it is possible to have a human-like social interaction between humans and AI but this interaction can't be social in the human sense because of a lack of intentionality, subjectivity, morals, ethic, mind, etc. There are of course similar capabilities of the technological artefact that seems relatively equal. Still, it is indeed different, and we need to focus on those differences, and this is only possible if we approach the dimension of the techno-social. With the implementation of this term, we will have demarcated a field of research that deals specifically with techno-social ways of interaction inside the human social lifeworld. At this point, our investigation can also deal with strictly ethical issues. If in fact, we remain with the old paradigm, we risk continuing to view AIs as tools, incapable of deciding anything and therefore, also completely irresponsible. The essential core of the problem can be articulated in a series of questions like these: What does it mean that a technology *decides* something? What kinds of decisions are these? Can these decisions be *trusted* if trust – as opposed to reliability – is a distinctly social category? Accordingly, can one assign or delegate *responsibility* to such technology? Thus, there are three closely related ethically relevant challenges: The possibilities of AI-based *decisions*, the question of *trust*, and the problem of *responsibility*. These three ethical challenges traditionally refer to subjects in human communities, but they should not be considered only in human comparison (similarity), but as technological ethical attributes. Even though the technological artefacts can never *act*, *be trusting and trustworthy,* or *assume responsibility* in a human sense, they might do so in a different, *techno-social* sense. Those ethical aspects seem fundamental and relevant in the human-AI interaction since the ability to be *active* is one important characteristic of technologies with sobject-status and the essential basis of interactions in general. The new way of nontransparent, nonhuman, technological action caused by a black box, like the process that implements the simulation of human-like social interaction skills, leads mainly to the case of AI decision-making. Again, we cannot simply regard deliberation as a psychological phenomenon or the exclusive capacity of human consciousness, but we must observe its practical process and, more importantly, its outcome. Therefore, they should be called *techno-decisions*, as a consequence of *techno-actions*, and be distinguished from human decisions and actions. With this shift in techno-decisions, a distinction between human decisions and their social consequences, like responsibility and trust, is possible. Kaminski distinguishes in this context between causes and reasons as an origin of human-based decisions and discusses the abilities of AI decisions. The result and main distinction here are that AI-based techno-decisions could only be questioned about their causes, and not about their reasons, which led to the assumption that ethical or moral techno-decisions are not possible because they lack

[25] Misselhorn: »Maschinenethik und Philosophie«, in: *Handbuch Maschinenethik,* p. 37–39.

[26] Searle, J., R.: »Kollektive Absichten und Handlungen« [Collective Intentions and Actions], in: Schmid, H. B.; Schweikard, D. P.: *Kollektive Intentionalität*. *Eine Debatte über die Grundlagen des Sozialen*, Suhrkamp 2009, p. 99–118.

a motivation that is not technical-instrumental.[27] I suggest investigating *techno-responsibility* and *techno-trust* through a way of technological causes or perhaps also a way of technological reasons. With the distinction of techno-social and techno-ethical attributes in the sobject context, it becomes possible to research those three main ethical challenges in the human-AI interaction with another focus.

Three essential ethical Attributes: Decisions, Trust, and Responsibility

The introduced approaches in the context of human-AI interaction, *the sobject*, and the *techno-social*, lead mainly to three essential ethical aspects. As mentioned, it is about the ability of AI to *act* human-like, which results in ethical attributes regarding techno-social decision-making, questions about responsibilities for those decisions, and trust in that black boxed *techno-decisions* and *actions*. In addition, it is also important to clarify which ethical approaches are useful and can be helpful in this context. While attempting to research ethical capabilities in the context of *social* human-AI interactions, a normative ethic seems suitable, since descriptive approaches are pretty subjective and more complicated in the application of AI.[28] Therefore, with a normative approach, it is possible to research which aspects could and should be applied in the techno-social actions of AI. In general, in the context of normative ethics, we can differentiate between deontological and teleological theories. While teleological ethics assume an *extra-moral value* (*außermoralischer Wert*) that is the only and „fundamental criterion for what is morally right, wrong, obligatory, etc."[29], and which is constituted as an aim to strive for, deontological ethics starts from the assumption that there is a general structure of duty that is articulated in a series of norms, which provide a criterion for action and not a goal. Hence, it differs mainly in the conception that there are not teleologically beforehand given ethical values independently from actions, like utilitarianism. Deontologists assume there are more criteria than a pure maximization of goods, or the maximization of happiness, which are not just based on efficiency. For example, Kant's categorical imperative is an objective and normative principle for subjective actions and is valid "for all rational beings."[30] The formulation of *rational beings* includes the ability of intentionality and all the resulting characteristics like consciousness, responsibility, trust, etc. Since it seems possible to define extra-moral values beforehand and program those rules to AI systems, it also becomes clear that this is not enough if the claim is to analyze techno-actions in the human lifeworld, which also consist not just of teleological norms. Precisely because the social interactions are not clearly definable and especially not programmable or calculable, the research of this section focuses on exactly those aspects, which are more than teleological evaluation like utilitarianism (utility maximization), but also not entirely transferable to a deontological ethic in the Kantian sense (because the sobject does not seem to have that capacity of self-reflection). Since there are often utilitarian approaches in the field of technical risk evaluation, which refer to a cost-benefit assessment, there is also a well-founded risk of objectifying human society.[31]

As a result, teleology leads to a reification of human social interactions, while deontology is too close dependent on the intentional character of human beings. Therefore, a middle path regarding the techno-social approach could help research a solution to this problem. Ned Block also analyzed this kind of middle path, referring to different types of consciousness resulting from two different ways of

[27] Kaminski: »Gründe geben «, in: *Datafizierung und Big Data,* p. 151–154.
[28] For example, the suggestion of Adam Smith spectator would be hard to verify in a technological context.
[29] „grundlegende[s] Kriterium dafür, was moralisch richtig, falsch, verpflichtend usw." in: Frankena, W. K.: *Ethik. Eine analytische Einführung* [Ethics. An Analytical Introduction], Springer 2017, p. 15.
[30] „für jedes vernünftige Wesen" in: Kant: *Grundlegung zur Metaphysik der Sitten*, p. 53.
[31] Werner, M.; Düwell, M.: (2013): »Deontologische Ethik« [Deontological Ethics], in: Grunwald, A.: *Handbuch Technikethik*. Metzler 2013, p. 159–161.

formation for humans and for AI. He talks about *access consciousness*, the cognitive experience, and *phenomenal consciousness*, the subjective experience. If humans hold both kinds of consciousness, machines can instead only have a sort of access consciousness.[32] The techno-social term can also include this distinction and collect further aspects that mainly lead to the three essential ethical attributes: decision, trust, and responsibility. Certainly, there are also further relevant ethical aspects in this discussion, but I suggest focusing on those three since they seem as fundamental in social interactions as they are problematic in techno-social interplay.

Decisions

The question of decision mainly derives from the characteristic of technologies with sobject-status to be *active*. If certain types of *actions* are possible, it depends on their being black boxes and the epistemic deficit that we attribute to *techno-decisions* and *techno-actions.* The black box aspect, described as mentioned by Kaminski, differentiates between technological models according to the need to understand them as we use them. Classical technology needs a practical way of understanding if one wants to use them as an instrument. It is practical knowledge because it is necessary to understand how to use a technological tool or instrument, while it is not needed to comprehend it epistemically during its usage. As a result, in this way of interaction, it is possible to use the technology as an instrument only with practical understanding so that it is also, in this case, a black box. But the most important thing is that it is even possible to transform the black into a white box and also gain an epistemic comprehension of classical technology. With the use of trans-classical technologies, the kinds of practical and epistemic understandings shift in a way that is no longer possible to differentiate clearly between them because of the way of usage. This means the human – trans-classical technology interaction is no longer a situation in which the human uses the technology as a mere instrument, since the technology could also have a way of practical understanding of its own and represents an interaction partner on another level of complexity, a sobject.[33] In this case, an epistemic understanding of the technology is necessary but it is not fully possible: the black box can't be transformed into a white box. This situation represents the ability of sobjects to act and be active in a technological way. Hence a techno-social action and, concurrently, a decision depend on this black box status. This kind of technology remains not entirely predictable and penetrable by the human with whom it relates, and this unpredictability is not resolved by the simple knowledge of how it operates. Because of its non-transparency, in combination with artificial intelligence, it develops a human-like interaction that includes actual decisions. This could be decisions used, for example, in selecting applicants for a certain job, for lie detecting, or as driving assistants. There are many capabilities to apply AI-based decisions in the human social environment. Furthermore, it could also be the case of an interaction between a human and a humanoid robot. In that case, we will have a really strong human-like interaction since the sobject represents a technology that seems human in its appearance and in its gestures, wordings, communication, and interaction skills. But every case includes an AI decision that impacts the human in its social lifeworld. In this way, speaking about techno-decisions, instead of decisions, leads to the ability to research those sobject or AI decisions without attributing them as decisions in a human sense. Even if the formation and the internal process of the two decisions differ, their outcome could still be the same.[34] The way of formation is relevant in this field because of the blurred boundaries and the slight distinctions in some interaction situations. Therefore, we could ask for the causes and reasons for a decision and its justification. Causes can be understood and reconstructed, while reasons are

[32] Block, N.: »On A Confusion About a Function of Consciousness«, in: *Behavioral and Brain Sciences*, Cambridge University Press 1995, p. 227–287.
[33] Kaminski: »Gründe geben«, in: *Datafizierung und Big Data*, p. 157–162.
[34] This is also demonstrated by the Turing Test, Misselhorn: »Maschinenethik und Philosophie«, in: *Handbuch Maschinenethik,* p. 37–39.

more normative and ethically evaluative since they can be good or bad, moral or immoral.[35] In human society, in some situations, it is sufficient to clarify only the causes, but in many contexts, it is also necessary to ask about the reasons that drove someone to a certain action, because only in this way a decision can be ethically evaluated, and the human being is responsible not only for external causes but also for internal motivations.

In summary, the decisions made by humans and machines are performed in the same social lifeworld and could have the same outcome but differ in their formation. The ability for a techno-decision is possible through the black box, which conducts to a non-fully transparent, opaque result that could be used but not epistemically understood. It is similar to human-based decisions, which is why they are linked to ethical and moral aspects. Since there is often no possibility of epistemically understanding a human decision, the reasons must be asked whenever we need to form an ethical evaluation. The most important question is: Could an AI have reasons for a decision, and how could they be ethically evaluated? It seems simple to ask about the causes, but by exceeding the objective, instrumental category and really being considered in a social sense, asking only about the causes seems insufficient. This is why we need the space for techno-decisions. Since there can't be reasons in a human sense but something different. It should focus on that *difference* and create a way to deal with decisions and the in-between from causes and reasons. Teleological evaluations could be possible in this context if they create extra-moral values, which could be evaluated by their causes since the techno-decisions are just following those without a genuine reference to values and moral rules. But, as mentioned, deontological aspects are also important in this context which could reflect the reasons but are more difficult to implement in the technological context. Hence, we cannot deny that techno-decisions are based on causes, and how they could also have reasons is unclear right now, but it is obvious that they are different from human ones. As a result, we gain a different expectation for AI decisions, because we know they are not decisions in the human sense, even if their outcome could be the same. Moreover, since it is impossible to understand all the details of an AI's decision fully, we have to *trust* these black-box AI technical decisions beyond finding them merely reliable. While there can be no subjective reasons for AI decisions, the black box situation represents another level of action. There is nothing left but to *trust* instead of *understand*.[36] This results in the second ethical aspect we will inspect: trust.

Trust

Trust is another major ethical issue that arises in human-AI interaction. Hildrun Lampe and Kaminski point to the specific black boxing of intelligent computer simulations as the reason for which people describe "their epistemic relation to those machines as a *trust relation*."[37] If it is not possible to fully understand an action, we must trust if we want to use the outcome for anything or just want to interact with an AI. This is analogous to the social need for trust: we need to trust people in various types of interactions because we never know exactly how or why a person will act, also due to a lack of epistemic understanding. But since people can and must take responsibility, we can always ask about the reasons behind a certain action, because they are based on a subject of consciousness[38]. However, trust in other people can also be problematic – people can make mistakes and wrong decisions. We can say that they are not trustworthy precisely because they are subjects. So, how should we deal with

[35] Kaminski: »Gründe geben«, in: *Datafizierung und Big Data*, p. 151–154.

[36] Martini: *Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz*, p. 28–31.

[37] „ihr epistemisches Verhältnis zu dieser ‚Maschine' als Vertrauensrelation" in: Lampe, L.; Kaminski, A.: »Verlässlichkeit und Vertrauenswürdigkeit von Computersimulationen« [Reliability and Trustworthiness of Computer Simulations], in Liggieri, K.; Müller, O.: *Mensch-Maschine-Interaktion*, J.B. Metzler 2019, p. 328.

[38] Block: »On A Confusion About a Function of Consciousness«, in: *Behavioral and Brain Sciences*, p. 227–287.

trust interacting with AI? Due to the black box aspect, even in the case of an AI decision there are not only causes but likewise reasons which can be ethically examined in a certain way. Since only the result can be fully evaluated, but not the entire internal process that led to the adoption of a norm of orientation, it is not really dissimilar to what happens in everyday human trust. The outcome can at least be evaluated teleologically by examining the fulfillment of certain purposes. If that is true in the interpersonal sphere, with AI decisions the trust problems can, at first, be seen as much simpler, because the machine appears to operate in a more objective way, uncompromised by subjective motives and consideration, acting only in view of a calculable result. But this simple objectivity is actually only apparent: the data basis must be given beforehand, and algorithmic rules have to be created. So, the seemingly objective outcome of AI decisions is not less complicated than in the case of, for example, a subject that is using a classical object. Indeed, there is the added difficulty of not really being able to distinguish where the intentional action of the subject ends and the merely causal action of the object begins. Sobjects are something in-between, and it is this different situation regarding the trust in those techno-decisions that leads us to introduce the concept of *techno-trust*. Highlighting this distinction, the expectation of necessary *trust* with AI decision changes. The uncanny valley[39] effect of a human-like mysterious technological person who is trustworthy in a human sense could be prevented. Trust can be discussed in the context of AI, but it is already known in advance that it is a different, technical kind of trust, so no human-like expectations should be established – the techno-decision should not be humanized. Lampe and Kaminski differentiate in this context between *trust* and *reliability* and argue that AI decisions could merely be reliable since trust is only possible if we postulate a social scenario in which a degree of opacity is unavoidable.[40] But using the techno-social approach, we discover a space for AI decisions that is also socially opaque, but in a techno-social way. This means reliability is still fundamental (similar to human-human interactions) but is not the only relevant property. I argue that reliability is necessary, and we could easily verify techno-decisions' reliability, however, when the request for reliability is already fulfilled, the question of trust and trustworthiness also arises in those decisions, even if they are not based on subjective social decisions. There is a way of trust which is necessary for human-AI interaction in human society, mainly because the responsibility of human decisions, AI decisions, and their creators are blended. It is needed to differentiate those kinds of decisions and how we can trust them. While we have to trust (or not trust) when we can't understand them epistemically, it is important to be aware that this is not the same trust as in other human beings.[41]

In conclusion, some aspects could already explain techno-trust, for example: i) the fact that we need to trust in a way; ii) the knowledge about sobject-characteristics; iii) the insight regarding the possibilities about techno-causes but the problem with techno-reasons; iv) the need for a new kind of trust in the field of human-AI interaction[42]; v) the relation to techno-decisions as something that is not subjectively connoted. However, even if this phenomenon still needs to be fully defined (as well as the other techno-social aspects), the approach we have followed shows at least an alternative handling for those aspects within their main challenges. It doesn't shift the problems to another level, instead, it changes the understanding of those problems and the subsequent interactions. Close to techno-trust is the ethical field of *responsibility,* which is connected to trust and concurrent with techno-decisions.

[39] Mori, Masahiro: *The Uncanny Valley*, The Original Essay by Masahiro Mori (MacDorman, K. F.; Kageki, N. trans.), IEEE Spectrum 2012, URL: https://spectrum.ieee.org/the-uncanny-valley.

[40] Lampe; Kaminski.: »Verlässlichkeit und Vertrauenswürdigkeit von Computersimulationen«, in: *Mensch-Maschine-Interaktion*, S. 329–331.

[41] Martini: *Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz*, p. 28.

[42] Nevertheless, there are also different viewpoints about the necessity of a separate technological ethic, like Hubig, C.: *Die Kunst des Möglichen III. Macht der Technik* [The Art of the Possible III. the Power of Technology], transcript 2015.

This term is increasingly used in different ways of dealing with AI decisions, but we will have to ask what responsibility means in this context.

Responsibility

As already mentioned, trust leads to responsibility. This is a central aspect of various AI debates. If AI is actually capable of making autonomous[43] decisions, the question of responsibility will arise. Can AI itself take responsibility? Or are the programmers responsible? Can responsibility be handed off? Is it desirable to hand responsibility over to AI decisions? These questions are particularly relevant in a variety of fields and situations, for example with the development of autonomous driving. I will argue for an additional category of responsibility. This makes it possible to examine the handling of AI avoiding the difficulties of attributing the same responsibilities to a machine as to a human being. The main problem regarding autonomous driving is the handling of accident scenarios. Different human lives could not be compared and calculated in advance what would be necessary in the case of a programmed accident scenario. The ethical commission of autonomous and connected driving said that a moral decision-making ability is essential since those dilemmas are not fully standardizable and programmable.[44] This is why responsibility can't be entirely handed off to autonomous AI decisions since the decisions in such cases are not programmable beforehand because of ethics regarding human life. It is not ethically possible to establish mathematically what the criteria are for which one life is worthier than another. In this case, a responsible subject with consciousness is necessary. But there are the same problems in a dilemma with human subjects, except that the subject now has the *faculty* of deciding. The responsibility of human beings mainly regulates this difficult situation, and transferring it similarly to AI decisions seems problematic. But it is also up to the chosen way of ethics since teleologic approaches like utilitarianism can justify programmed AI decisions of autonomous driving. If utility maximization is the only thing that counts, a standardized and programmed responsibility could be justified. Leon Sütfeld, for example, said that moral decisions are calculatable and could therefore be rules for moral AI decisions.[45] And, from a deontological perspective, one could argue that responsibility is only attributable to a human subject capable of will and understanding. Again, both ways are important, but problematic in the field of AI decisions. The conclusion that just human beings can ever decide in a dilemma since they are subjects of consciousness, ends the discussion too early and may never consider allowable fully autonomous driving as well as other AI decisions. On the other hand, rigid utilitarianism is also problematic because human lives and the very structure of morality are objectified and calculated. As a possible solution, it could be helpful to find how AI decisions are trustworthy and responsible in a way to deal with and interact with them without the need for a full subjectification of AI or a complete objectivation of humans. And this is why techno-decisions and techno-trust have necessary to result in techno-responsibility. Through the sobject-phenomenon, it becomes clear again that existing forms of social interaction and ethical evaluation are no longer sufficient. With the term *techno-responsibility*, it could be possible to interact with techno-decisions in a way where responsibility is needed in human sociality. But with the difference that we do not speak of the same responsibility in a human sense since they are no intentional subjects but sobjects. This means we can use AI decisions but have to examine each situation and cannot make it easy to transfer human responsibility to AI. Additionally, it is a proposal that prevents the case of

[43] Autonomous as a property of techno-decisions.

[44] *Bundesminister für Verkehr und digitale Infrastruktur: Ethik-Kommission Automatisiertes und vernetztes Fahren* [Federal Minister of Transport and Digital Infrastructure: Ethical Commission of Autonomous and Connected Driving], report June 2017, p. 10.

[45] Frontiers in Behavioral Neuroscience: *Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure*, 05.07.2017, URL: www.frontiersin.org/articles/10.3389/fnbeh.2017.00122/full.

making a human a *liability servant* (*Haftungsknecht*). If only the condition must be fulfilled, a person must make the final decision to be responsible, even if they do not understand the context.[46] Humans need to deal with the new kind of techno-social elements in the social environment. It is also necessary to clarify the concrete situation and context and how we are using AI decisions in which case. Then, it is needed to be conscious of the possibilities for social-techno interactions. It is, for example, also needful in using chatbots, search machines, or the new open AI program *ChatGPT*. The expectations of a new hyper-intelligent technological being that is *better* than humans are not just false, but also highly problematic. AI decisions, the way of processing data input and their output are different. They cannot and should not be compared directly to humans, including humans' intelligence, sociality, or morality. The creation of those aspects is distinct, but the understanding of their interaction on a social level is possible through the help of the concept of the techno-social.

Conclusion

Summarized, this paper recommends a way to deal with the ethical attributes of AI without anthropomorphizing it. By acknowledging the necessity to introduce an intermediary between subject and object, namely a sobject, and with it a new type of techno-social interactions, the classical ethical questions regarding subjective categories like decision-making, trust, trustworthiness, and responsibility can be rethought for those human-like but still non-human technologies, like AI. It was shown that this additional kind of techno-social interaction is already there but had not been given a precise characterization. It is not just a change of terminology because the shift this term introduces can open an original discussion space and reshape the phenomenon. As a result, social interaction with AI is possible in a way that is not in the human sense but rather in a techno-social one that also allows a kind of social interaction. The direct comparison to human capabilities is no longer necessary, and an uncanny anthropomorphism could be prevented. The techno-social approach needs to be wholly defined and is required to do it in further examinations. This one, I hope, could be a useful starting point for further research since the approach shaped a base for discussions and awareness of the techno-social phenomenon. Finally, other related aspects go hand in hand with this approach. For example, the question of laws, rules, and limits concerning AI can perhaps be better evaluated if we take it this way, instead of referring to a legal framework that still tends to distinguish the subjective element, only human, from any other external element, considered accidental or nevertheless not autonomous.

[46] Beck, S.: »Einleitende Worte zur Bewertung des Zusammenwirkens von Mensch und Maschine« [Introductory Words for the Evaluation of the Interaction of human and machine], in: Haux, R.; Gahl, K. (et al.): *Zusammenwirken von natürlicher und künstlicher Intelligenz*, Springer 2021, p. 117.