Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03849-x OPEN

Extended human agency: towards a teleological account of AI

Jörg Noller [1]✉

This paper analyzes human-machine interrelation concerning artificial neuronal networks (ANNs) from a teleological point of view. The paper argues that AI cannot be understood adequately in terms of subjectivity or objectivity but rather as a new kind of teleological relationship that holds between human and artificial performances of intelligence. Thereby, AI is understood as an enactivist extension of human agency, both in instrumental and moral terms. This hybrid account will be distinguished from four alternative accounts of human-machine relations: (i) the simulation account, according to which AI simulates human rationality; (ii) the instrumentalist account, according to which AI is just a tool; (iii) the anthropomorphic account, according to which AI is human-like; and (iv) the indifference account, according to which AI will merge with human rationality due to technological progress. Against these four accounts, the paper argues for a teleological account of AI as extended human agency that is part of the human lifeworld. By focusing on the teleological interrelation of socially grounded databases and algorithms, the paper finally develops an account of responsible AI that considers its specific relatedness with human actions, purposes, and intentions by means of language. Understanding human-machine relations in terms of extended agency finally allows to tackle the question of how to avoid the problems of AI bias and opacity.

[1] LMU Munich, Munich, Germany. ✉email: joerg.noller@lrz.uni-muenchen.de

Extending human agency

The phenomenon of artificial intelligence (AI), especially artificial neural networks (ANNs) and machine learning (ML), has become the focus of philosophical interest more than almost any other topic concerning digitalization. Whereas many studies are dedicated to the ethical status in general (Dubber et al. 2020; Floridi 2023) and the question of responsible and trustworthy AI in particular (Agarwal and Mishra 2021; Voenecky et al. 2022), this paper relates AI's ethical status to its epistemological status (Buckner 2018; Buckner 2024) and, more generally, to its relation to the human lifeworld (Floridi 2014; Floridi 2015; Stalder 2018; Durt 2022; Noller 2022) or "*Umwelt*" (Froese and Ziemke 2009), generally understood in terms of our everyday lives and practices that are not only restricted to scientific practices. The concept of lifeworld was introduced by Edmund Husserl (1970) into the philosophical discourse, and I will discuss his conception with regard to AI in chapter 2. This lifeworld-relationship is essential for evaluating AI as its ethical status crucially depends on what AI actually is and 'does,' and how to conceptualize it. For example, it is unclear in which sense AI can be called "artificial" and "intelligent." Our lifeworld interpretation of AI depends on how to understand these concepts exactly.

Shifting from the question of trustworthy AI to a conception of extended human agency allows me to focus on human AI autonomy, understood as the responsible use and implementation of AI into everyday practices. In this paper, I will focus on ANNs to discuss the question of artificial intelligence's relation to the human lifeworld and extended human agency – both from an individual and collective perspective. The reason for focusing on ANNs is twofold: first, ANNs may be considered one of the most advanced AI technologies, having become part of our everyday lives and practices – not only as a technological phenomenon, but also as something that has become part of science and culture. Therefore, AI is not only becoming more and more part of our everyday cognitions (Smart 2018) but also actions. This raises the question of how to understand AI from a practical point of view. There are many ways to conceive of AI in terms of action. One prominent view is to attribute practical or even moral properties to AI systems as such as if they acted in a moral way. Recently, however, the attribution of moral qualities to AI systems has been criticized for being a way of anthropomorphizing them. Instead, philosophers have argued that we should shift our attention to those who program AI systems (Agarwal and Mishra 2021, ix–x.) that is, to understand moral responsibility in relation to AI from the privileged expert – that is, creator and designer – point of view. However, our everyday use of AI demands a third conception beyond anthropomorphization and expertism, which I will call the *extended action account*. A reference point for this alternative account is the role of databases on which ANNs operate from a quasi-empirical point of view (Buckner 2018), as well as the algorithms by which ANNs operate on databases from a conceptual point of view. Databases are not epistemologically neutral given that they may transport empirical stereotypes or misrepresent opinions or facts. This phenomenon has generally been described as an "implicit bias" (Brownstein 2019) towards non-intentional distortions. However, there may even be forms of explicit bias – for instance, when we intentionally distort the databases on which AI systems operate by contributing distorted data to the database.

Aside from databases, we also need to consider algorithms when it comes to the extended action account. An algorithm, understood as "a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose, under given provisions" (Hill 2016, 48), can lead to what Eli Pariser (2011) has called a "filter bubble," which epistemically and practically distorts our orientation in the digital lifeworld. Filter bubbles pretend to represent the objective reality, yet in actuality manipulate digital users mainly for economic reasons by only presenting information that might be of (commercial) interest for those institutions that have created the algorithmic system that leads the user to stay in it. In any case, AI is transforming our everyday space of action, and, therefore, also our ethical concepts, which makes an ethical discussion of its lifeworld relation pressing (Powers and Ganascia 2020, 28).

This paper aims to determine the relationship of AI's human-machine interaction from the perspective of human action and social lifeworld. Recently, the focus on responsible AI has shifted from machines as merely technological objects to contexts such as "socio-technical systems" (Dignum 2019, 52). In order to develop my account of extended human agency, I shall draw on Don Ihde's phenomenology of human-technology relations and develop an account of AI-extended human agency, thereby also drawing on the enactivist AI research program (Smart 2018; Froese and Ziemke 2009). This will allow us to understand AI in such a way that instead of being logically situated within the framework of a subject-object schema, as suggested by the term "human-machine interaction," it is understood as part of our everyday lifeworld, in which human cognitive and practical capacities as well as the performance of AI are deeply interwoven. To analyze these epistemic and ethical relationships, I shall refer to conceptions of enactive artificial intelligence (Froese and Ziemke 2009; Smart 2018). Thereby, I will apply these enactive conceptions not primarily to human cognition but to human action. Recently, these phenomena of human-machine interaction have been described as "cyberbilities" and "hybrid agency" (Essmann and Mueller 2022). However, the relationship between databases, algorithms, and human action still needs to be further developed with regard to ANNs and their lifeworld integration. Only based on such an understanding of the broader lifeworld framework can ethical questions regarding human-machine interaction and AI be discussed appropriately.

Methodologically, I will approach the phenomenon of AI from a critical middle position, and distinguish my extended action account from the following four positions: (i) the simulation account, according to which AI simulates human intelligence; (ii) the instrumentalist account, according to which AI is just a tool; (iii) the anthropomorphic account, according to which AI is human-like; and (iv) the indifference account, according to which AI will merge with human intelligence due to technological progress, as represented in Ray Kurzweil's singularity thesis (Kurzweil 2005, 9), which has been criticized as "Silicon Valley ideology" (Nida-Rümelin and Weidenfeld 2022, 4).

I shall develop my extended action account of AI by understanding it not as a specific technology of digitalization but as a phenomenon deeply interwoven with our everyday actions. Thereby, teleological structures such as a purpose or an intention will serve as the level of abstraction (LoA) from which I will discuss ANNs.[1] I shall argue that it is this LoA that allows us to tackle the problems of opacity and bias and to develop an account of responsible AI. In doing so, it is essential to avoid dualistic subject-object divisions between humans on the one hand and machines on the other. To this end, I will reflect on the interactive dimension between humans and ANNs by focusing on the interrelation of human purposes, intentions, socially grounded databases, and algorithms. I shall call this the hybrid view of AI. Thereby, I shall focus on the role of language as a 'transporter' of human purposes and cognitions that can be extended by means of AI. The role of language has been emphasized from the perspective of extended mind theorists (Clark and Chalmers 1998, 11). I shall draw on this theory and extend it to human agency.

AI, teleology, and the digital lifeworld

What is a "digital lifeworld"? The concept of "life-world" was first introduced from a phenomenological perspective into the philosophical discourse by Edmund Husserl. Recently, it has received increasing attention with regard to the philosophical significance of AI (Durt 2002, 68). Husserl understands the lifeworld in contrast to what he calls the "practical world" and "science," which he calls "purposeful structures" (Husserl 1970, 382). However, the lifeworld, as Husserl defines it, is not something that lacks any purpose, but is rather the very *condition* of purposive structures: "all setting of ends, all projecting, presupposes something worldly; the wherewith, i.e., the life-world, is given prior to all ends." (Husserl 1970, 138). In what follows, I shall argue that AI and other digital technology are not opposed to the lifeworld but rather part of it, and constitute a structure that I shall call the digital lifeworld.[2] Along these lines, Luciano Floridi has coined the term "onlife," which describes a structure of the digital lifeworld. By "onlife" he means the fact that "[t]he digital online world is spilling over into the analogue-offline world and merging with it" (Floridi 2014, 43). With regard to the digital lifeworld, Floridi (2015, 2) has argued that we find ourselves in a situation characterized by "the shift from the primacy of stand-alone things, properties, and binary relations, to the primacy of interactions, processes and networks." What is special about the digital lifeworld, however, is that we are not only consumers of digital information but also their producers. We can therefore use the term "prosumer," to use a label introduced by Alvin Toffler, originally meant to characterize "people who consumed what they themselves produced" (Toffler 1980, 282), in order to describe the agential role of members of the digital lifeworld. The digital prosumer, however, does not consume what she produces, but rather consumes and produces by being part of the digital lifeworld and existing "onlife."

Here, the question arises of how to understand AI in such a way that it can be part of our digital lifeworld. In his paper "A Phenomenology of Technics," Don Ihde distinguishes between various forms of human-technology relations. In what follows, I shall discuss these relations with regard to the role that AI plays in the digital lifeworld. "Embodiment relations" (Ihde 2009, 77), such as "visual technics" as glasses, change the way we experience our body and transform our perception. Here, technology "is the symbiosis of artifact and user within a human action" (Ihde 2009, 77). "Hermeneutic relations," on the other side, mediate between me and the world. An example for this relation is a thermometer, whose temperature number is something we need to understand in order to establish a relationship to the world and to understand whether it is warm or cold. "Alterity relations" are relations in which technology appears to be something other than me. Ihde argues that computer technologies, in particular, constitute alterity relations (Ihde 2009, 92). Finally, Ihde proposes what he calls "background relations." These relations describe the fact that in our lifeworld, many technologies are working in the background and are not becoming visible as objects but rather produce "background noise" (Ihde 2009, 95).

Here, the question arises as to what kind of relationship AI constitutes according to Ihde's distinctions. Conceiving of AI in terms of a part of the human lifeworld, as shall be undertaken in this paper, does not allow us to conceive of AI in terms of an "alterity relation." Floridi has convincingly argued that we are experiencing a kind of "blurring of the distinction between human, machine and nature" (Floridi 2015, 2). Likewise, conceiving of AI as part of our lifeworld does not allow us to understand it in terms of a "background relation," since this would underestimate its importance. What comes close to the digital lifeworld, however, is Ihde's "embodiment relation." This relation has recently been discussed in terms of "embodied AI" (Froese and Ziemke 2009). However, in contrast to glasses, AI is not an "artifact" but rather a structure and process in which we are involved. Likewise, our relationship to AI is not adequately understood in terms of a "hermeneutic relationship," since we are not only using AI as a tool but rather engage with it in a much deeper way insofar as we are part of its processing, e.g., by contributing to its database. Therefore, I shall argue that AI constitutes a new relationship that I shall call an "extension relationship."

To better understand this AI extension relationship, however, we must first determine what is meant by "intelligence". To be sure, the notion of "intelligence" can be understood in various ways. On a very basic level of abstraction, "intelligence" can be understood in terms of the capacity and operation of information processing. Forms of information processing are comparing and distinguishing information, which includes detecting logical relationships such as identity, implication, or opposition, but also abstraction. Concerning the capacity of abstraction, Buckner (2018, 5341) has argued that ANNs have the capacity to abstract from individual empirical data certain patterns, which leads to a process of increasing representational content, a process that he calls transformation and "categorical abstraction." By means of such a process of abstraction, ANNs transform empirical and social data into patterns that can be understood in terms of concepts.

What is common to most conceptions of intelligence is the capacity to realize purposes – be they non-moral, moral, or even biological in the case of an organism (Froese and Ziemke 2009, 479). For example, a subject can be called "intelligent," insofar it realizes a purpose that either has been given to it from something other than itself (we may call this "heteronomous intelligence"), or that has been given to it by itself (we may call this "autonomous intelligence"). The link between human and artificial intelligence, as it is currently mostly discussed regarding ANNs, is the concept of a *pattern*. Both human and artificial intelligence can be understood in terms of pattern recognition and acting according to these patterns. Furthermore, intelligence can be realized by means of *following rules* in order to realize a particular goal or purpose. As such, the concept of an algorithm becomes important when discussing ANNs, as it allows integrating AI into human actions. Hill (2016, 36 and 48) has generally defined an algorithm as a means for "problem solving," and problems presuppose a "given purpose." Both the concept of a pattern as well as the concept of an algorithm allow connecting the concept of ANNs with the concept of human action that is situated in a lifeworld.

Going beyond Husserl's phenomenological approach, I shall interpret the concept of lifeworld in terms of everyday practice. Therefore, by "lifeworld," I refer to a context consisting of various kinds of patterns, be they economic, emotional, acoustic, visual, or linguistic, as well as certain purposes that need to be realized. Consisting of patterns that are of practical relevance, the lifeworld becomes an object of ANNs. We can use ANNs for almost every situation in everyday life, as long as it has to do with patterns. For example, we can employ AI to translate texts into any language we want (e.g., by means of a deep learning-based translator such as DeepL), we can use AI to create pictures according to our voice commands (e.g., by means of a text-to-image model generator such as Dall-E2), we can use AI to detect diseases through image recognition, and so on. The common link between all these uses of AI is the fact that they are used for a specific *purpose*; or, in other words, they share a teleological structure. I understand a teleological structure as a structure that consists of purposes that are to be accomplished for various reasons. In general, actions are those types of entities that allow to realize purposes. As AI is becoming more and more part of our lifeworld and our actions, it

is also becoming more and more a phenomenon of teleological significance. In what follows, I shall understand a teleological structure as a context of *individual or collective purposes*. These purposes can be realized and extended by means of intelligence; that is, by means of pattern recognition and algorithms.

Within the process of ANNs, the concept of a pattern is closely related to the concept of a database. According to my teleological account of AI, the databases are social in nature; that is, they transport semantic meaning, intentions, and human purposes, mostly but not exclusively by means of language. For example, I am contributing to the linguistic database of an AI by publishing texts on the internet. These texts may become part of an AI's training. AI usually detects patterns within large databases by means of algorithms, which humans cannot process due to the sheer amount of information. In regard to social databases, AI's pattern recognition is not just an abstract calculation but enables and enhances teleological orientation. This orientation is realized by means of algorithms. In what follows, I shall understand algorithms in terms of abstract teleological *instructions for actions*. Accordingly, responsible AI consists of adequately applying algorithms to social data to avoid any form of bias. Based on social data, pattern recognition becomes a social practice because patterns can serve as concepts and, therefore, as reasons for action.

The human lifeworld is fundamentally structured by human purposes and intentions. As I will show in what follows, AI can be integrated into the human lifeworld by extending, relating, abstracting, and realizing these purposes in terms of digital structures that correspond to it. The digital lifeworld is, accordingly, structured by human purposes and interests that are mediated and transformed by means of digital phenomena such as AI. Therefore, an ethical evaluation of AI's performance depends on whether and how we conceptualize those intimate relationships in terms of teleological processes.

I shall understand AI not in the sense of algorithms and processes that operate in the background of human culture and lifeworld, and which just reduce "big data" to "small data" (Stalder 2018, 59). Rather, I shall interpret them in the sense of complex processes that are currently subsumed under the term "machine learning" (ML), which enters into a much closer relationship with us in the social lifeworld – it "gets closer to our skin than other technologies" (Müller 2023), because it deals with social databases, such as linguistic ones, and patterns that can be integrated into human teleological processes. Therefore, the distinction between "human" and "machine" becomes increasingly less important from the perspective of extended human agency.

AI beyond simulation and duplication

Talking about the "artificiality" of intelligence is ambiguous. According to the weak interpretation, "artificiality" specifies the intelligence of technical systems in the sense that they *simulate* human intelligence. According to the strong interpretation, however, machines *realize* human intelligence. These two interpretations roughly correspond to the distinction between weak and strong AI, which has become commonplace.[3] Conceptions of weak AI often understand it as a mere technical property of objects – as an "alterity relation" or merely "background relation," according to Ihde's terminology – and thus tend to trivialize AI. In contrast, conceptions of strong AI understand intelligence as a property of a subject and thus tend to anthropomorphize it. By focusing less on the subject or object of AI and more on the process of intelligence as such, these problems can be avoided. The hybrid account of AI that I will argue for in this paper is rather concerned with the interrelation of human subjectivity (intentions, purposes, actions) and AI technology. It aims to relate the main components of ANNs – the databases and the algorithms – to human subjectivity as part of a teleological framework. As such, AI can be connected to and integrated into the human lifeworld such that the respective performances interfere with, extend, and complement those of human subjects.

Douglas Engelbart coined the term "augmented human intellect" (Engelbart 1962) for this understanding of AI. He understands this as a "systematic approach to improving the intellectual effectiveness of the individual human." (Engelbart 1962, ii) However, this augmentation of human intelligence should not be understood as "isolated clever tricks that help in particular situations." Rather, according to Engelbart, "extension" denotes a holistic and systemic phenomenon; it concerns "a way of life in an integrated domain where hunches, cut-and-try intangibles, and the human 'feel for situation' usefully coexist with powerful concepts, streamlined terminology and notation, sophisticated methods, and high powered electronic aids." (Engelbart 1962, 1)

Despite Engelbart's holistic approach, which understands human-machine interaction as a "set of interacting components rather than by considering the components in isolation" (Engelbart 1962, 2), this "systematic approach" is still too focused on quantitative performance enhancement in the sense of "increasing human intellectual effectiveness" (Engelbart, 1962, 3). Through these instrumentally understood extensions as "augmentation means" (Engelbart 1962, 9), the qualitative dimensions of teleological human-machine interaction are lost sight of. A qualitative and not purely quantitative-instrumental interference model of AI that extends our autonomy suggests that it should not be understood in terms of "human-machine interaction" but rather further analyzed in terms of AI-extended human agency in which not only human intelligence but also other human capacities – such as our will or our emotions – are augmented by AI. These capacities have recently been called "cyberbilities" (Essmann and Mueller 2022, 428). These extended and even transformed human capacities need not only belong to individual subjects but may concern societies and institutions as well (Essmann and Mueller 2022, 428).

The question of whether and to what extent machines can be called intelligent was already addressed by Alan Turing in his classic essay "Computing Machinery and Intelligence" (Turing 1950). Turing proposes a simulation account of AI and argues that whether and to what extent machines can think cannot be answered directly – it is "too meaningless to deserve discussion" (Turing 1950, 422). Interestingly, however, Turing argues that this question would no longer be considered meaningless in the year 2000 – the very year in which, according to Stalder, "a new cultural constellation" (Stalder 2018, 4) – the "digital lifeworld," as it were, – began: "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted." (Turing 1950, 422) Instead of the question, "Can machines think?," Turing suggests replacing it with the question, "Are there imaginable digital computers which would do well in the imitation game?" (Turing 1950, 442). Turing avoids a direct definition of the term artificial intelligence and replaces it with a thought scenario in which the conditions are developed for a machine to be described as (artificially) intelligent with good reason.

Turing developed his simulation account of AI through what he called the "imitation game" (Turing 1950, 433). This experiment, also known as the "Turing test," consists of three instances – a woman (A), a man (B), and a person (C), who is supposed to find out the gender of the two persons unknown to her by asking clever questions, who in turn try to keep C in the dark about their identity. Now, either the woman or the man is replaced by a computer, so the following question arises as a condition for the

success of the attribution of artificial intelligence to the computer: "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" (Turing 1950, 434)

Central to this is Turing's conception of the test as a *game* between three instances. The computer's intelligence is measured by the degree to which it can *imitate* a natural person. This imitation performance must be linguistic, argumentative, and dialogical, i.e., highly social and context-sensitive. This imitation performance, like the artificiality of intelligence, can be understood in two different ways. On the one hand, "imitation" can mean something like simulation. Still, on the other hand, it can also be understood in the sense of a duplication, i.e., the realization of human intelligence in terms of an enactivist account. John Searle, in particular, has pointed out this difference (Searle 1984). He argues that computers are only able to simulate human cognitive performance, but not to duplicate it – a difference he calls the "key distinction" (Searle 1984, 37–38) Searle links the "key distinction" between mere simulation and actual duplication with his distinction between syntax and semantics (Searle 1984, 34). He argues that computers are only ever capable of syntactic operations, but that these can never lead to semantic content, which, according to Searle, only mental states possess (Searle 1984, 36).

However, the distinction between simulation and duplication of human intelligence is only relevant as long as we attribute it to a machine conceived as a subject. If we look at AI not from the subject's point of view but as a process in terms of intelligence performance and operation, that is, as an extension of a human's cognitive and agential faculties and their purposes, this distinction becomes less and less significant. By conceiving of AI neither as an object – in terms of an "alterity relation" –, nor as a subject but as an interrelated process in terms of an "extension relation," its performance can be embedded in various lifeworld contexts and systematically linked to human agency. From this teleological point of view, AI does not appear as abstract algorithms and a mere calculation of numbers, but as a meaningful process that deeply concerns human practices.

What is essential to the AI-extended account of human agency is the interrelatedness of algorithms, databases, and teleological structures. According to this account, algorithms function as extensions of human purposes in that they can be used to realize them methodologically. Likewise, the databases on which algorithms operate are deeply linked to human epistemic practices, like the linguistic and semantic contents of the internet. The chatbot ChatGPT, for example, has been trained based on several hundreds of billions of tokens of ultimately human origin.[4] AI and human intelligence are interrelated in such a way that the former's algorithms are oriented and trained towards the teleology and structure of the human lifeworld. This teleology can be understood in terms of instrumental reason but must not be restricted to it. Rather, in being trained on various purposes, AI is part of a theoretical and practical rationality that is generally oriented towards realizing purposes. The AI-extended account of human agency is therefore distinguished from conceptions of hybrid moral responsibility or extended agency theory (Hanson 2009), according to which "responsibility is shared between humans and machines." (Berber 2023, 1900) Against conceiving of AI as an object into which morality needs to be implemented (Verbeek 2011), the AI-extended account of human agency conceives of AI as a deep interrelation between technology and human purposes. It is not the technological object that needs to be moralized but rather the relationship and processes that emerge between AI and human subjects – be they individual or collective.

The extension of human agency manifests in the training of an ANN, that is, using specific algorithms and methods that result in outputs corresponding to human purposes. This practice has been described as "a fine art": "Success with backpropagation and other connectionist learning methods may depend on quite subtle adjustment of the algorithm and the training set." (Buckner and Garson 2019) Thereby, human responsibility concerns both the database and the training procedure by means of algorithms. The former needs to be as objective and neutral as possible, the latter must adequately represent human purposes. It is the unity of (empirical) database and (conceptual) algorithms that enable responsible AI-extended agency.

Conceiving AI-extended objectivity in terms of a unity of social data and conceptual algorithms allows reducing both opacity and bias of AI systems, which are considered the "central issues" (Müller 2023) of current data ethics. Solving the opacity problem is only possible if AI performances are understood in terms of representations and extensions of human interactions. Solving the bias problem of AI requires interpreting social databases such as the internet in terms of individual or collective human products. Objectifying AI, however, renders these solutions impossible since it categorically separates AI performances from those of humans.

AI-extended action

In the current debate, Luciano Floridi has further defined the lifeworld significance of AI as part of his theory of the *infosphere* and understood it in terms of "artificial agents." He thus extends the concept of moral agents by including AI under the concept of "moral agents" and "moral patients" (Floridi, 2013, 134). Floridi develops a concept of "mindless morality," (Floridi 2013, 135) which he considers applicable to AI. This term implies that AI cannot be attributed mental characteristics such as intelligence and free will. Nevertheless, Floridi argues that artificial agents can be ascribed *accountability*, even though they are not ascribed *responsibility*, which only concerns those persons who created and conceived the artificial agent (Floridi 2013, 135). Floridi relates this distinction between *accountability* and *responsibility* to the relationship between parent and child. While parents are jointly *responsible* for their child's behavior, they cannot be held legally *accountable* for their actions once it has reached a certain age (Floridi 2013, 135).

According to Floridi, the decisive factor for AI's moral and ontological definition is the descriptive framework and the level of abstraction on which it is based. Applied to AI, this means that the level of abstraction of information ("informational level of abstraction") and not that of substances must be chosen to describe its reality. The ontological level of information differs from Newton's ontology, for example, which is based on material objects and substances (Floridi 2013, 27). Floridi argues that the choice of the level of description and abstraction is not to be understood relativistically but is motivated by the purpose of explanation and a teleology of reasons (Floridi 2013, 146–147).

Floridi agrees that AI is increasingly becoming part of our everyday lives, and cannot be adequately understood within an ontology of substance. However, the concept of AI as (moral) subjects is problematic insofar as he chooses an informational (onto)logical level of description that is not appropriate to the digital lifeworld. Rather, it is essential to choose a concept of reality in such a way that it can integrate both human persons and AI into a comprehensive lifeworld without changing the level of abstraction. As already shown in the previous sections, the "extension relation" allows conceiving the level of description of AI-extended agency not in terms of the concept of information but rather of that of a purpose, understood as a teleological structure such as a goal or an intention.

Clark and Chalmers (1998) have argued for an extended mind thesis according to which cognition is extended outside the human mind, and according to which the environment plays an "active role in driving cognitive processes" (Clark and Chalmers 1998, 7). This extension of our cognition in terms of language can be applied to AI. ANNs operate on vast language databases originating from the internet and our digital lifeworld, thereby extending our cognitions. However, by means of operating on linguistic databases, ANNs do not only extend our cognitions but also our purposes and therefore actions. Hence, an account of AI-extended human agency needs to primarily focus on how language transports individual as well as collective intentionality and purposes, and how these subjective states can be externalized in such a way that human autonomy is rather extended than restricted. Referring to the enactivist cognitive account by Clark and Chalmers (1998), I shall argue that the lifeworld of language not only extends cognition but also human actions and autonomy.

Extending human agency refers to the networking of AI capabilities with interests and purposes set by autonomous people. As such, the ethical significance of AI must not be understood in a vertical sense, namely in the sense that we are confronted with a new kind of power that is superior to us humans in terms of intelligence. Rather, we must think of the significance of AI in a horizontal sense: AI is increasingly being integrated into our lifeworld and networked with it, so that it is perceived less and less as a subject or object opposing us, but rather as an open teleological process that can be networked with human processes. Both the algorithms and the databases that constitute ANNs are ultimately of human origin and centered around human purposes. Therefore, responsible AI demands that we become aware of this teleological human-machine dependence beyond mere subjectivity and objectivity.

Conclusion

If we understand AI as a part or structural moment of our lifeworld, the question of its moral-philosophical status as a subject or as an actor no longer primarily arises. Rather, we are interested in an enactivist setting and the normative context of reasons, which are initiated by human individuals and collectives, and which can be extended by the integration of AI performances. However, AI performance as such is always heteronomously interested and motivated: it requires a teleological initiation that it receives from outside, that is, from autonomous human beings. The normative problem of AI, therefore, arises not so much from the way it functions but from the fact that it may not be properly integrated into our lifeworld, and that it confronts us as a technological object in terms of an "alterity relation" or "background relation," cementing a subject-object divide, and even promoting our dependence on technologically privileged experts and institutions. From an ethical perspective, a transition from an "oppositional approach," according to which AI represents a potential danger to us, to a "systemic approach," according to which AI is viewed as "a set of technologies that are embedded in a system of human agents, other artificial agents, laws, non-intelligent infrastructures, and social norms", is therefore necessary (Powers and Ganascia 2020, 48–49).

Hence, the ethical challenge of AI is primarily to integrate its technology into our lifeworld so that it interacts with us in the lifeworld context and enables new forms of rationality. Such interaction is not only to be understood in the sense of a quantitative increase in our intelligence but also as an expansion of other faculties, such as our will or our power of judgment. This expansion of our faculties must not be understood as a technology that affects us as individuals but as an extension of individual and collective agency. Of course, this does not exclude the possibility that the improper use of AI leads to a restriction of human agency instead of its extension. Therefore, a responsible use of AI entails extending our space of action and integrating AI into our lifeworld such that it becomes transparent to our original and autonomously given purposes, for which we are ultimately responsible.

Data availability

This is not required for this paper, since no data set was used for the research.

Received: 11 February 2024; Accepted: 24 September 2024;
Published online: 04 October 2024

Notes

1 For the notion of the level of abstraction, see Floridi (2008).
2 Durt (2022, 68) has argued that AI "integrates into the lifeworld in a way not known from previous technology." However, Durt does not consider the practical and agential role of AI as I do in my AI-extended action view.
3 Russell and Norvig (2022), 1020: "[T]he assertion that machines could act as if they were intelligent is called the weak AI hypothesis by philosophers, and the assertion that machines that do so are actually thinking (not just simulating thinking) is called the strong AI hypothesis."
4 See https://en.wikipedia.org/wiki/GPT-3#Training_and_capabilities.

References

Agarwal S, Mishra S (2021) Responsible AI: Implementing Ethical and Unbiased Algorithms. Springer, Cham
Berber A (2023) When something goes wrong: Who is responsible for errors in ML decision-making? AI Soc 39:1891–1903. https://doi.org/10.1007/s00146-023-01640-1
Brownstein M (2019) Implicit Bias. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Fall 2019 Edition). https://plato.stanford.edu/archives/fall2019/entries/implicit-bias
Buckner C (2018) Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese 195:5339–5372. https://doi.org/10.1007/s11229-018-01949-1
Buckner C (2024) From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence. Oxford University Press, New York
Buckner C, Garson J (2019) Connectionism. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Fall 2019 Edition). https://plato.stanford.edu/archives/fall2019/entries/connectionism
Clark A, Chalmers D (1998) The extended mind. Analysis 58(19):7–19
Dignum V (2019) Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, Cham
Dubber MD, Pasquale F, Das S (eds) (2020) The Oxford Handbook of Ethics of AI. Oxford University Press, New York
Durt C (2022) Artificial Intelligence and Its Integration into the Human Lifeworld. In: Voenecky S, Kellmeyer P, Mueller O, Burgard W (eds) The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives. Cambridge University Press, Cambridge, p 67–82. https://doi.org/10.1017/9781009207898.007
Engelbart D (1962) Augmenting Human Intellect: A Conceptual Framework. Stanford Research Institute, Menlo Park
Essmann B, Mueller O (2022) AI-Supported Brain–Computer Interfaces and the Emergence of 'Cyberbilities'. In: Voenecky S, Kellmeyer P, Mueller O, Burgard W (eds) The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives. Cambridge University Press, Cambridge, p 427–443. https://doi.org/10.1017/9781009207898.033
Floridi L (2008) The Method of Levels of Abstraction. Minds Mach 18:303–329. https://doi.org/10.1007/s11023-008-9113-7
Floridi L (2013) The Ethics of Information. Oxford University Press, Oxford
Floridi L (2014) The 4th Revolution: How the Infosphere is Reshaping Human Reality. Oxford University Press, Oxford
Floridi L (2015) The Onlife-Manifesto: Being Human in a Hyperconnected Era. Springer, Cham
Floridi, L (2023) The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities, Oxford University Press, Oxford

Froese T, Ziemke T (2009) Enactive artificial intelligence: Investigating the systemic organization of life and mind. Artif Intell 173:466–500. https://doi.org/10.1016/j.artint.2008.12.001
Hanson FA (2009) Beyond the skin bag: on the moral responsibility of extended agencies. Ethics Inf Technol 11:91–99. https://doi.org/10.1007/s10676-009-9184-z
Hill RK (2016) What an Algorithm Is. Philos Technol 29(1):35–59. https://doi.org/10.1007/s13347-014-0184-5
Husserl E (1970) The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy (trans: Carr D). Northwestern University Press, Evanston
Ihde D (2009) A Phenomenology of Technics. In: Kaplan D (ed) Readings in the Philosophy of Technology. Rowman & Littlefield, Lanham, p 76–97
Kurzweil R (2005) The Singularity is Near. Viking Penguin, New York
Müller, VC, (2023) Ethics of Artificial Intelligence and Robotics. In: Zalta EN, Nodelman U (eds) The Stanford Encyclopedia of Philosophy (Fall 2023 Edition). https://plato.stanford.edu/archives/fall2023/entries/ethics-ai
Nida-Rümelin J, Weidenfeld N (2022) Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age. Springer, Cham
Noller J (2022) Digitalität: Zur Philosophie der digitalen Lebenswelt. Schwabe, Basel
Pariser E (2011) The Filter Bubble: What the Internet Is Hiding from You. Penguin, New York
Powers TM, Ganascia J-G (2020) The Ethics of the Ethics of AI. In: Dubber MD, et al., (eds) The Oxford Handbook of Ethics of AI. Oxford University Press, New York, p 27–51. https://doi.org/10.1093/oxfordhb/9780190067397.013.2
Russell S, Norvig P (2022) Artificial Intelligence: A Modern Approach. Pearson, Boston, et al
Searle J (1984) Minds, Brains and Science. British Broadcasting Corporation, London
Smart PK (2018) Human-extended machine cognition. Cogn Sys Res 49:9–23. https://doi.org/10.1016/j.cogsys.2017.11.001
Stalder F (2018) The Digital Condition (trans: Pakis VA). Wiley, Cambridge and Medford
Toffler A (1980) The Third Wave. Bantam Books, New York
Turing A (1950) Computing Machinery and Intelligence. Mind 59:433–460. https://doi.org/10.1093/mind/LIX.236.433
Verbeek PP (2011) Moralizing Technology: Understanding and Designing the Morality of Things. The University of Chicago Press, Chicago and London
Voenecky S, Kellmeyer P, Mueller O, Burgard W (eds) (2022) The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives. Cambridge University Press, Cambridge

Acknowledgements

The research is funded by the German Research Foundation (DFG), Scientific Network "Philosophy of Digitality" (project number 541433337).

Author contributions

The author of this paper is the only author.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The author declares no competing interests.

Ethical approval

Ethical approval was not required as the research did not involve human participants.

Informed consent

Informed consent was not required as the research did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Jörg Noller.

Reprints and permission information is available at http://www.nature.com/reprints

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