An International Journal in Philosophy, Religion, Politics, and the Arts

Existenz

Volume 17, No 1, Spring 2022 ISSN 1932-1066

Robots, Emotions, and Interobjectivity

Jörg Noller
Ludwig-Maximilians-University Munich, Germany
joerg.noller@lrz.uni-muenchen.de

Abstract: Catrin Misselhorn develops a non-reductive account of artificial intelligence that has become an aspect of the human lifeworld and her approach takes also into account the importance of emotions for artificial intelligence systems. Thereby Misselhorn argues for a middle ground between a form of weak technological reductionism and a strong transhumanist utopianism. While I am sympathetic to this general approach, I shall propose an alternative life-worldly interpretation of artificial intelligence systems. To this end I develop an interobjective rationale regarding artificial intelligence that does not merely refer to robots as being individual entities but one that refers to the way as to how artificial intelligence interferes with human activities in everyday life. Hence, I argue that artificial intelligence should be conceived of not in terms of robotics or devices but rather in terms of situated processes and capacities that function from within virtual realities. According to this alternative view, artificial intelligence is neither an object nor a subject but a self-reflective process that has the potential to shape human interactions with reality and society, and of enhancing or restricting individual and collective freedoms.

Keywords: Misselhorn, Catrin; artificial intelligence; robots; emotions; lifeworld; transhumanism; virtual reality; interobjectivity; freedom.

Beyond Reductionism and Transhumanism

Catrin Misselhorn's two recent books, *Grundfragen der Maschinenethik* and *Künstliche Intelligenz und Empathie* build upon each other. While the former lays the ground for a machine ethics in general, the latter focuses specifically on the complexities of emotionally informed artificial intelligence. Misselhorn points out—in my opinion correctly so—that machine ethics needs to be distinguished from other forms of ethics of technology since, as she puts it, the former addresses the development of an ethics that informs machines, whereas the latter is concerned with the ethics of handling and utilizing machines.[1] Misselhorn is mainly concerned with the question as to

> whether and how to construct machines that can make and implement moral decisions themselves, and whether one should do so. [*GM* 8]

She criticizes the transhumanist singularity thesis, which assumes the development of a super intelligence displaying self-consciousness and freedom of the will. Instead, Misselhorn argues in favor of striking a balance between technological reductionism and transhumanism, that is, endorsing the feasibility of

> equipping machines with the ability to make moral decisions and to act according to them in a functional sense, [*GM* 15]

while at the same time being opposed to the transhumanist view that human abilities such as

[1] Catrin Misselhorn, *Grundfragen der Maschinenethik*, Ditzingen, DE: Reclam 2018, p. 8. [Henceforth cited as *GM*, all translations are mine]

consciousness and freedom of the will can be fully reproduced and duplicated or even outperformed by way of artificial intelligence systems.

In my view, Misselhorn's approach is convincing for she argues also for a third way that is beyond weak simulation and strong duplication of human abilities by artificial intelligence systems. This third way would be necessary if one were to agree with Misselhorn that morally acting machines

> bring fundamental changes to the way we conceive of ourselves and to the way of how we live together in society. [*GM* 15-6]

However, I suggest to conceive of this third way between weak simulation and strong duplication of human intelligence in a different manner than Misselhorn does.

Methodologically, I will borrow from the position by Julian Nida-Rümelin and Natalie Weidenfeld and distinguish my approach from the following four positions:

* the ideologization of artificial intelligence systems in the context of what Nida-Rümelin and Weidenfeld coined a transhumanist-futurological "Silicon Valley ideology";[2]
* the associated anthropomorphization of artificial intelligence in the sense of a "modern animism" that understands artificial intelligence systems as being living persons [*DH* 3];
* the banalization of artificial intelligence as a mere tool, that is, regarding it as a technological reduction,
* the dramatization of artificial intelligence as a danger for modernity, that is, as being an ideological counter position to the Silicon Valley ideology; this amounts to a conflict between apocalyptic fears and euphoric expectations [*DH* 1].

My argument is based upon understanding artificial intelligence not so much as a particular technique of digitalization, but rather as a phenomenon within a lifeworldly framework of digitality that is opposed to a merely technological framework of digitalization. A philosophical analysis of artificial intelligence needs to reflect upon its technological meaning, and it also needs to analyze its lifeworld meaning. With "lifeworld meaning" I refer to everyday practices that seamlessly involve artificial intelligence and allow to virtualize previously analogue and physical processes. I find it to be especially important to avoid dualistic subject-object divisions between human beings on the one hand and machines on the other hand, as they underlie an instrumental and therefore reductionist understanding of artificial intelligence programming.

Artificial Intelligence and Digitality

With the rise of complex media applications, for example the internet, in addition to carrying messages, media themselves become ontological factors of the human lifeworld. Hence, in order to emphasize the lifeworldly significance of digitization, one does not need to focus exclusively on the technology of digitization; rather one needs to focus on what is being called "digitality." Digitality indicates that the product derived from digitization—what emerges from it—is itself something meaningful and qualitative that can no longer be reduced to being a product of a purely technical or medium nature. Digitality is thus, in short, the qualitative, lifeworldly side of digitization and is in this respect not subjected to a general criticism of media technology.

There are three phenomena that can be considered paradigms of digitality; these include successfully distributing and connecting Internet content worldwide, the advancements in artificial intelligence research, and the systematic integration of computer games and virtual reality in everyday life. As phenomena of digitality, these three paradigms are interconnected and deeply interwoven. By way of connecting digital contents and targeting human neurophysiological and emotional reactions, such a media has the capacity to change one's way of perceiving the world, thereby creating collectively shared perceptions of new realities. The intense and widespread impact of these new realities calls for a genuinely philosophical analysis, for they are not just marginal technical phenomena with a negligible impact on human life, but they are deeply immersed into daily life and transform the human lifeworld. Moreover, the engineering designs for these media enact ontological factors of a reality in which a common distinction between sender and receiver is increasingly diffused.

When digitization is no longer viewed to be merely a technical development, but itself becomes a

[2] Julian Nida-Rümelin and Natalie Weidenfeld, *Digital Humanismus: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age*, Cham, CH: Springer Nature 2022, p. 4. [Henceforth cited as *DH*]

part, even a structure of the human lifeworld, it takes on the function of digitality. Hence, a stipulation of concepts is being made for allowing content designers and end users alike to grasp the digital lifeworld and its ontology. A philosophy of digitality needs to be equally distinct from sociological, psychological, technological, and economic approaches regarding the phenomenon of digitalization. The philosophy of digitality, however, is not merely an additional marketing approach for diverse stakeholders, but it aims to substantiate the other approaches by critically reflecting upon the underlying concepts used in this genre, for example those of simulation, fiction, virtuality, or reality. Artificial intelligence in particular must be discussed in light of these concepts.

Artificial Intelligence and Lifeworld

The lifeworld of human beings can be analyzed by way of pattern recognition. In her book *Künstliche Intelligenz und Empathie*, Misselhorn argues that artificial intelligence systems can be utilized for analyzing patterns of morality, of health, of finance, of art, and even of emotions.[3] However, her approach does not address the question as to how exactly, that is on the ontological level, artificial intelligence is being integrated into the human lifeworld. I think that one can better avoid the transhumanist conception of strong artificial intelligence, that is, the hypothesis that artificial intelligence is a duplication of human subjectivity, if one does not conceive of it in terms of individual quasi-subjects such as robots but rather in terms of processes that guide these machines. Misselhorn rightly points out that there are machines of great ethical importance, such as war robots and fully automated self-driving vehicles. Here the question arises whether these machines can be considered as acting in an autonomous way or not. I argue that one cannot call them autonomous in the lexical understanding of the term, yet they could be called hypothetically autonomous machines. They lack autonomy, insofar as they do not set themselves original, freely chosen purposes and goals and insofar as these are being given to them from the outside, namely by programmers or users. Albeit, I argue that while reflective power of judgment cannot be assigned to artificial intelligence, it can be granted a kind of determining power of judgment.

Hence, I suggest understanding artificial intelligence not in terms of individual machines that possess certain abilities in a functional sense, as Misselhorn argues it (*KIE* 15), but instead in terms of modular functions that do not exist independently from human subjects that conceive and coordinate these functions by autonomously setting goals for them. These goals can be of merely instrumental pertinence or of moral relevance. Borrowing from Yuk Hui's terminology, I shall call this modular functionality an "interobjective" account of artificial intelligence. However, I contrast my use of "interobjective" from Hui's use by applying it to the human intersubjective lifeworld, while Hui uses it solely in a technical sense and contrasts it to the human intersubjective lifeworld.[4] Yet my use of an interobjective account is closer to Luciano Floridi's position than to the one by Hui. Floridi asserts that the digital processes of artificial intelligence can be integrated seamlessly into the human lifeworld, which he calls a "frictionless infosphere" and "data superconductivity."[5] Whereas Floridi's account is neutral concerning the moral status of such a lifeworldly integration of digitalization, which raises the question whether it is compatible with totalitarian uses of individual preferences, I suggest to interpret it in terms of what Immanuel Kant has called a "public use of reason" as opposed to a merely "private use of reason."[6] For Kant argues that one's self-inflicted immaturity is finally due to one's incapacity to use one's own reasoning, which means to subject one's thought and action to "statutes and formulas," that is, to let oneself be determined by laws that are different from the moral law, a state of affairs which Kant calls heteronomy. Kant speaks of these statutes and formulas in terms of "mechanical tools", which he considers as being "the leg cuffs of a perpetual

[3] Catrin Misselhorn, *Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co*, Ditzingen, DE: Reclam, 2021. [Henceforth cited as *KIE*, all translations are mine]

[4] Yuk Hui, *On the Existence of Digital Objects*, Minneapolis, MN: University of Minnesota Press 2016, pp. 158-60

[5] Luciano Floridi, *The 4th Revolution: How the Infosphere is Reshaping Human Reality*, New York, NY: Oxford University Press 2014, p. 42.

[6] Immanuel Kant, "Beantwortung der Frage: Was ist Aufklärung," *Berlinische Monatsschrift, Zwölftes Stück* (December 1784), 481-494, here pp. 484-5, my translation. [Henceforth cited as *WA*]

immaturity" (*WA* 493). I propose to transfer Kant's insight of immaturity to the utilization of artificial intelligence insofar as it is imperative that its use must avoid being subject to algorithmic statutes and formulas as a mere means to achieve, for instance, economic ends.

This means that human-machine-interaction will be oriented toward human ends and will not be used to instrumentalize or objectify human beings. My lifeworldly perspective regarding the use of artificial intelligence is incompatible with totalitarian uses of individual preferences, for it is rooted in a beneficial advancement of humanity. In this way, artificial intelligence data processing does not function as a subject or object that is different from the human subject but, by analogy, it functions in terms of what Marshall McLuhan envisions as being an "extension of man."[7] Artificial intelligence utilization is profoundly linked to its users who provide the data base for the respective algorithmic operations. Unlike traditional tools, such as, for instance, a hammer, artificial intelligence gathers data by incorporating the users' preferences as a kind of feedback loop. If the users do not adequately contribute to the data base or infosphere, or if intentional manipulation of the infosphere takes place, this would ultimately lead to moral malfunctioning of guidance by way of artificial intelligence and may harm its users. Hence, data ethics and artificial intelligence usage are deeply linked. Artificial intelligence computing can assist its user to extend subjectivity, not in terms of a super-intelligent reproduction, but in terms of what I call a transsubjective context in which the subjective states of humans, such as one's interests, aims, knowledge, or emotions are not exclusively bound to one specific human being but rather are shared, communicated, and linked with other users. Ideally, this can lead to a Kantian public use of reason.

Speaking of the artificiality of intelligence that is made possible by recent technological development is by itself already an ambiguous use of language. Stuart Russell and Peter Norvig, in 1995, argued to differentiate between "strong AI" and "weak AI."[8] In the context of a weak definition, artificiality further specifies the intelligence of technical systems in the sense that they simulate human intelligence. A strong reading, however, suggests that machines realize intelligence akin to humans. Interpretations of weak AI often regard it as a mere technical property of objects and thus tend toward a banalization or instrumental reduction of it, while interpretations of strong AI understand intelligence as being an integral property of a subject and thus tend toward ideologization, anthropomorphizing, or dramatization, which, however, are less philosophical than speculative in character. By focusing less on the question regarding a subject-based or object-based interpretation of artificial intelligence and more on the actual intelligence performance as such, these problems can be avoided. This allows for artificial intelligence to be more easily integrated and included in the lifeworld in such a way that its performances interfere with, extend or restrict, and complement or hinder the performances of human beings.

Based on his ten-years in-depth study under the joint sponsorship of the Stanford Research Institute and the Directorate of Information Sciences of the Air Force Office of Scientific Research, Douglas Engelbart has coined the phrase "augmenting human intellect" in an attempt of increasing human capabilities, including

> more-rapid comprehension, better comprehension, the possibility of gaining a useful degree of comprehension in a situation that previously was too complex, speedier solutions, better solutions, and the possibility of finding solutions to problems that before seemed insoluble.[9]

Engelbart understands this also to mean a "systematic approach to improving the intellectual effectiveness of the individual human being" (*AHI* ii). This augmentation of human intelligence is not to be understood in the sense of applying "isolated clever tricks that help in particular situations" (*AHI* 1). Rather, he considers enhancement to be a holistic and systemic phenomenon that concerns

> a way of life in an integrated domain where hunches, cut-and-try, intangibles, and the human "feel for a situation" usefully co-exist with powerful concepts, streamlined terminology and notation, sophisticated methods, and high-powered electronic aids. [*AHI* 1]

[7] Marshall McLuhan, *Understanding Media: The Extensions of Man*, New York, NY: McGraw-Hill, 1964.

[8] Stuart Russell and Peter Norvig, *Artificial intelligence: A Modern Approach*, Englewood Cliffs, NJ: Prentice-Hall, 1995.

[9] Douglas C. Engelbart, *Augmenting Human Intellect: A Conceptual Framework*, Menlo Park, CA: Stanford Research Institute 1962, p. 1. [Henceforth cited as *AHI*]

Thomas Ramge has therefore very aptly pleaded for AI to be understood in this networked lifeworld sense:

> The English abbreviation AI could then soon no longer stand for Artificial Intelligence, but...for Augmented Intelligence, that is, not for artificiality, but for augmentation.[10]

Engelbart's "system approach to human intellectual effectiveness," understands human-machine interaction as a "set of interacting components [and opposes] considering the components in isolation" (*AHI* 2). I argue that this system approach is still too focused on quantitative performance enhancement for the purpose of increasing human intellectual effectiveness. This effective enhancement performance is conceived as being a mere continuity, equal to a holistic and systemic prosthetics. I think that with Engelbart's instrumentally understood extensions of human capabilities as "augmentation means" (*AHI* 9), the qualitative dimensions that characterize virtual reality in the sense of digitality and that go hand in hand with a holistic integration of intelligence performances move out of sight. A qualitative and not purely quantitative-instrumental interference model of artificial intelligence, which I shall propose, suggests not to grasp it in the sense of a human-machine-interaction, but to analyze it further in the sense of the concept of interobjectivity, in which human intelligence is extended by artificial intelligence in terms of virtual reality. The distinction between strong and weak AI is therefore only significant insofar as it is understood in the context of a subject-object relationship of human-machine-interaction. It becomes meaningless when artificial intelligence is understood within the framework of a concept of interobjectivity, transsubjectivity, and digitality, as an integral virtual aspect of a global lifeworld.

Artificial Intelligence and Emotions

In *Künstliche Intelligenz und Empathie*, Misselhorn has argued for a non-reductive account of rationality, thereby considering the important role of emotions. I do agree with her position that the role of emotions in conceptualizing rationality in general, and regarding deep learning models in particular must be taken seriously. However, I differ from Misselhorn in my interpretation of emotions. I do not conceive of emotions as being a sub-class of affective phenomena, next to moods (*Stimmungen*), sentiments (*Empfindungen*) and character traits (*Charakterzüge*). Rather, I propose to sharply distinguish emotions from mere affections. In my view, emotions are an integral expression of human subjectivity for they are being connected with self-consciousness and self-reflection, different from mere feelings such as hunger, and they are of robust extension, that is, they are not merely of a momentary duration such as affects like anger. Emotions are distinguished phenomena, both in terms of their rationality and quality of feeling. Misselhorn explains:

> Emotional artificial intelligence endangers not only privacy, but also intimacy, the realm of our most personal thoughts and feelings, including our sex live, that constitute us as persons in a very fundamental manner. Access to this realm jeopardizes self-determination in a whole new way and on a yet unprecedented scale. [*KIE* 37]

Interobjectivity and Transsubjectivity

In an attempt of fully integrating artificial intelligence systems into the lifeworld of humans albeit without proclaiming technological alienation, I would like to shift the focus from machines as being mere subjects, and humans as being mere objects to these machines (or *vice versa*) toward a larger context, in which both interlocutors interact with one another in terms of interobjectivity. This holds especially true in the context of machine learning where human creativity directly feeds the algorithmic landscape of deep learning. Hence, I propose a conception of emotional artificial intelligence in which data gathering takes place in virtue of the processes of emotional pattern recognition that are aspects of everyday life. When artificial intelligence is seen as being an extension of oneself, emotional expressions that are mediated by deep learning become expressions of oneself, and ultimately will be understood as a kind of mediated emotional self-awareness. This complexity would shift the focus from the machine to oneself and to one's unique personality traits, of which emotions are pivotal.

Surely, artificial neural networks are essentially built upon pattern recognition which has both a formal (algorithmic) and a material (empirical) side. In this respect, deep learning cannot easily be reduced to algorithmic processes alone. Rather, there

[10] Thomas Ramge, *Augmented Intelligence: Wie wir mit Daten und KI besser entscheiden*, Ditzingen, DE: Reclam 2020, p. 45, my translation.

is a complex interrelationship between algorithm and database, which can be understood in the sense of the relationship between what Kant has called concepts and intuition. I suggest to understand it in the sense of Felix Stalder's concept of algorithmicity,[11] that is, the algorithm permeates the database and makes it manageable as such. This brings Kant's words to mind:

> Without sensibility no object would be given to us, and without understanding none would be thought. Thoughts without content are empty, intuitions without concepts are blind.[12]

When they are applied in this context, I take this to mean that data-based intuitions without algorithmic concepts are blind, and algorithmic concepts without data-based intuition are empty.

As artificial intelligence is dependent upon a database, these data can be gathered from real life or from fictional life. The decisive factor here is that humans themselves enter the database—for example, in the case of linguistic corpora—and corporate stakeholders write algorithms for it. In this respect, one's relationship to the process of artificial intelligence is always a virtual self-relationship, albeit a strongly mediated one. Artificial intelligence, as manifested in the example of machine learning, is essentially pattern recognition. Patterns can be understood in various ways and they structure human life in different domains. In economics they concern growth patterns and constellations, in medicine disease patterns (visual, acoustic, olfactory), in language speech patterns (visual and acoustic), in ethics behavior patterns, and in aesthetics composition patterns. Through its paradigmatic structure of digitality, algorithmic deep learning is not only an instrument, but is increasingly becoming part of the modern lifeworld in the sense that it has become an extension of thinking and acting. For the end of gaining conceptual clarity, it is vital to ask whether artificial intelligence is a medium, a simulation, a subject, or a virtual reality. Furhtermore, it must also be determined more precisely in which sense the designation artificial intelligence can be meaningfully called artificial and intelligent.

By conceptualizing artificial intelligence as a process rather than a subject, the performance of intelligence can be embedded in various lifeworldly contexts and linked to human interactions, both individually and collectively. This connectivity, in turn, can be further defined in the larger context of digitality. Provided that the artificiality of deep learning is understood not only in terms of a simulation of natural intelligence, but as a new mode of utilizing intelligent performance that is distinct from human intelligence, it is plausible to discuss this capability in the context of virtual reality. This is so because virtual reality is to be strictly distinguished from simulation, even though conceptually a transition from simulation to virtual reality can be described, which then can be called virtualization. Virtual reality begins with mere simulation, namely, the model-like orientation to a given, natural reality, but then increasingly emancipates itself from it in such a way that the simulative aspect increasingly recedes in favor of a generative or duplicating aspect. Finally, there is no structural ontological analogy at all in a virtualized process, but solely an analogy of purpose and causality.

Elsewhere I have argued that the concept of virtuality has received increased attention less regarding simulation than regarding reality.[13] There is an increasing willingness to recognize virtual reality as a reality of its own kind and to distinguish it from mere simulation. Taking this distinction between simulation and virtual reality into account, the artificiality of convolutional neural network architecture can be understood as an aspect of virtuality as it realizes intelligence in its own causal way. This stipulation of intelligence is particularly present in process AI, which is initially modeled on human brain functionality and learning, by way of representing simulations yet with increasing development it is unfolding its own logic, which can no longer be understood solely in a structurally analogous way of mere simulation.

The question now arises to what extent the operational functionality of artificial intelligence systems can be described as learning, as intelligence, and as knowledge. A purely behaviorist description of machine learning falls short for it does not consider the process that leads to the specific results whose epistemic significance is in question. In this context it seems to be

[11] Felix Stalder, *Kultur der Digitalität*, Stuttgart, DE: Suhrkamp 2015, pp. 145-76.

[12] Immanuel Kant, *Critique of Pure Reason*, transl. and ed. Paul Guyer and Allen W. Wood, New York, NY: Cambridge University Press 1998, pp. 193-4, B 75.

[13] Jörg Noller, Digitalität: *Zur Philosophie der digitalen Lebenswelt*, Basel, CH: Schwabe 2022, pp. 26-44.

functionally appropriate to speak of a learning process. Upon processing large amounts of information, some researchers argue that a learning effect of the artificial neural network can be achieved, which is measured by way of adequate responses to a given task.

Conclusion: Artificial Intelligence as Virtual Reality

If one takes artificial intelligence as being an aspect or structural moment of the human lifeworld, questions regarding its moral-philosophical status as a subject or as an agent no longer arise. Rather, the focus shifts to the normative context of reasons that is initiated by a human being or a community and in which its utilization flows without any friction. The ethical challenge of artificial intelligence utilization consists above all in integrating its technology into the human lifeworld in a virtualizing manner so that it interacts with humans in their lifeworld and enables augmented forms of reality. Such interference is not only to be understood in the sense of a quantitative increase of human intelligence, but as an extension of other capabilities, such as imposing one's will upon others or enhancing one's power of judgment. This expansion of human capacities is to be understood not so much as a technology that affects individuals, but as a qualitative virtual space that regulates interobjective and transsubjective communal action. The virtual space of human action can be enhanced by way of integrating it into an infrastructure that utilizes artificial intelligence. This, of course, does not exclude that a disingenuous use of this infrastructure can lead to a restriction of virtual spaces of action (for example, authoritative claims regarding spreading misinformation), or that humans are being transferred into mere illusory spaces, which are then mistakenly believed to be reality (for example, agents of globalist organizations are legitimately immune to civil and criminal prosecution). The challenge with respect to implementing a frictionless lifeworld for artificial intelligence is the human coexistence with robotics, when these machines are understood as being instrumental objects or autonomous or semi-autonomous subjects.

Once artificial intelligence is being integrated into the human lifeworld, the question regarding its moral-philosophical status as a subject or as an agent no longer arises. Rather, my interest is in the normative context of reasons which is initiated by selective crowd sourcing and respective data centers in which human performances flow into artificial networks without any friction. Data collection is always grounded in a heteronomous way: it requires a teleological initiation that artificial intelligence action receives from outside. The normative problem of artificial intelligence utilization thus does not arise so much from its mode of operation, but from the fact that it may not be integrated into the human lifeworld and that it confronts humans as if it were a technological, even technocratic object that cannot be communicated and thereby entrenching a subject-object divide, which leads to a technological alienation. Seen in this light, artificial intelligence is neither an object nor a subject but an enhancement of human life with the aim of extending or instrumentalizing its approach to reality and society, and to enhance or restrict individual and collective freedom. Which one of these outcomes will prevail will depend upon thoughtful legislative rigor and cultural guidance.