Exploring the Human-AI Nexus: A Friendly Dispute Between Second-Order Cybernetical Ethical Thinking and Questions of AI Ethics

Enacting Cybernetics

INSIGHT

CYBERNETICS SOCIETY

DIETMAR KOERING

ABSTRACT

Connecting ethics, artificial intelligence, and human beings is of utmost importance, given their entanglements with each other today. The artificial intelligence (AI) that humans develop must be built on ethical foundations—otherwise, it could become too late to implement these, especially in the development of artificial general intelligence (AGI). As philosopher Bernard Stiegler has articulated, technology is closely linked to human development. Ethical core values should therefore be used for current AI development. This is crucial so that humans do not develop into what cybernetician Heinz von Foerster called *trivial machines*. Unless ethical goals are pursued alongside the development of AI, the future of humanity could be dystopian, as described by Yuval Harari. This paper discusses the multifaceted viewpoints of Stiegler, Foerster, and Harari, and ultimately argues in favour of Foerster's humanistic attitude. Referring to concepts from epistemology and techno-science, I explore a friendly dispute between second-order cybernetical ethical thinking and questions of AI ethics, which I consider to be equally crucial for development and speculation in contemporary fields concerned with algorithmic thinking. I argue that reflection is needed not only about the ethical outcomes of AI systems, but also about the ethical implications of the development, deployment, and use of these systems. Hence, what is important is not simply to address human-machine relations but rather the algorithms humans live by—the relationships between humans and algorithms and the impact of these relationships on humanistic values.

CORRESPONDING AUTHOR:

Dietmar Koering

Institute of Contemporary Art, Design and Architecture/ Art Academy of Latvia, LV

dietmar.koering@lma.lv

KEYWORDS:

Artificial Intelligence; Social Change; Ethics, Cybernetics: Second-Order; Cybernetics; Trivial Machines

TO CITE THIS ARTICLE:

Koering, D. (2023). Exploring the human-AI nexus: A friendly dispute between second-order cybernetical ethical thinking and questions of AI ethics. *Enacting Cybernetics*, *1*(1): Article 5. https://doi.org/10.58695/ec.4

DOI: 10.58695/ec.4

1. INTRODUCTION

According to philosopher Bernard Stiegler, technology has always impacted evolution and the social structure of societies (Stiegler, 1998). For Stiegler, the use of tools, technical instruments, and techniques enables the preservation of a kind of impersonal, collective memory, comprising traces of past events. These traces are preservations of past world interactions that are preserved in the function of tools (Bluemink, 2020). As such, traces are, in the most general sense, the prothesis of consciousness, without which there could be no mind, no remembering, no recollection of a past that one did not live oneself, no culture. This notion, which Stiegler (2010a) refers to as tertiary memory, shifts the problem of the current technological conditions and ecological crises (Gorny & Radman, 2022, p. 9). Stiegler notes that life-forms such as humans are initially shaped by material environments and conditions through adaptations to ecological niches (Gorny & Radman, 2022, p. 6). Stiegler seeks not merely to rethink the relationship between humans and machines but to rethink machines themselves, and therefore AI, as a challenge to thinking and philosophy itself (Roberts, 2007). The technological and epistemological relationship between humans and machines rests on the capacity to adapt. Currently, there is no evidence that ethical values are lost when humanity adapts to technology and related algorithms (Koering, 2018, p. 234). Adaptation is itself, of course, a central human ability (Antonelli, 2008) and also occurs in human interactions with software.

The current AI discourse is especially concerned with this interdependent development, such as in the recent debate about new generative AI such as Chat GPT or Dall-E, which can create texts and images based on machine learning. Interdependence means that the question of who has the right to intellectual property is not obvious. Thomas Fischer (2019) expresses the process in another way, stating: "In our efforts to maintain our well-being, we adapt to given circumstances and adapt our circumstances to our needs" (p. 281). Fischer also writes that, from the perspective of cybernetics, human beings as a whole form a closed loop. Specifically, environments and living conditions lead to adjustments within humans and humans make adjustments to their environments. This view is adapted to the methodology of this article, where I assume that each human being creates its own environment, and that this can be shaped by knowledge, as well as technology. While traditional science attempts to remove the agent and their subjective knowledge, cybernetics recognises that a human being is always a part of a system it observes, and every human being does not develop the same body of knowledge due to their varied environments. Thus, I speak of cybernetics, and more precisely of second-order cybernetics (Foerster, 2003), as a way to conceptualize the world. So, here I am holding a friendly dispute about second-order cybernetics' ethical thinking and questions of AI ethics. I reflect not only about the ethical outcomes of AI systems, but also about the ethical implications of the development, deployment, and use of these systems.

Anthropological studies show that different societies adopt very different boundaries between the human and non-human spheres (Descola, 2011). These differences also exist in natural and technological barriers, as well as different cultural and ethical views. If humans create their environment, which influences them in turn, and this environment is steered and influenced by AI, what spectrum of options do humans have to determine their environment? Human authenticity is no longer grounded in its individuality but relatively more so in the multiplicity of remote agents it hosts (Koering, 2019, p. 44). This refers to the idea that in a world where AI and other technologies are becoming increasingly prevalent, authenticity is being redefined as

the result of the interactions between multiple agents rather than being rooted in the individuality of any one agent. In *Homo Deus*, Yuval Harari (2017a) describes a dystopian world that could result from the uncontrolled development of advanced technologies, particularly in the fields of biotechnology, artificial intelligence, and data science. It is crucial to reflect on the current situation so that humans do not develop into a world such as that described by Harari, or into what Heinz von Foerster (2003, pp. 140, 144) called *trivial machines*, which Foerster defines as machines that are programmed to operate in a specific way without any capacity for self-reflection or self-awareness. In Stiegler's (2010b) terms: It is necessary to criticize the current process of proletarianization. Hutnyk (2012) takes up Stiegler's work to evaluate and critique Stiegler's use of the concept of proletarianization. The impact of technology and the concept of the general intellect are measured against Stiegler's concern about a short circuit that threatens humanity and requires a new critique. While Stiegler highlights the positive aspects of technology, Hutnyk is more critical, emphasising the ways in which technology can be used to exploit and subjugate marginalised groups.

Having both simple and complex functions, humans can be compared to what Foerster referred to as trivial and non-trivial machines. Trivial machines are simple machines that have a single function, such as a lever or a pulley. Humans can be compared to trivial machines in that they too have certain basic functions, such as the ability to breathe, digest food, and pump blood. Human beings are more than this, however. They can also be considered non-trivial machines in that they have a variety of complex functions, such as the ability to think, communicate, and create. In this sense, humans can be compared with more complex machines such as computers or machines that have a variety of functions and capabilities, but humans are still more than these. It is important to point out that this is a metaphor and not a literal comparison. Then again, comparing beings to trivial and non-trivial machines can itself be de-humanizing. This was already mentioned by Foerster: "While our pre-occupation with the trivialization of our environment may be in one domain useful and constructive, in another domain it is useless and destructive. Trivialization is a dangerous panacea when man applies it to himself" (Foerster, 2003, p. 208). It is also important to examine the possibilities already possessed by various algorithms to influence human environments adaptively. In summary, I am not addressing the relationship between humans and machines here, but rather the algorithms humans live by. So, I am creating a friendly dispute about the relationship between humans and algorithms and its impact on humanistic values. Basic human values refer to the fundamental beliefs and principles that guide behaviour and shape approaches to life. These values include things like truth, honesty, loyalty, love, and peace. Basic human values serve as a moral compass and form the basis for ethical behaviour (Beabout & Wennemann, 1993). When I speak of values in this paper, I would like to refer to the ones mentioned here. While the integration of global ethical values in algorithms is fundamentally difficult to achieve, I think this simplified form offers a good basis for it.

Foerster's (2003, p. 196) explanation of a non-trivial machine leads to fundamental questions that today only humans can answer and, thus, require personal responsibility, which is the basis for human equality. Hence, this paper is written as personal position. A position is a point of view, a distillation of knowledge on a subject. The position is debatable even as it is written. As a human being, one evolves, and the goal is to pass on these thoughts and to shape this position through communication and thus create a better future. This knowing cannot be transferred one-on-one; rather, it is up to the individual to connect observations and experiences with their

DOI: 10.58695/ec.4

own references. On the other hand, judging by how fast the development of AI is progressing, it may only be a matter of time before this also applies to AI, especially in the development of AGI.

2. AN UNDESIRABLE FUTURE

2.1 THE UNDERSTANDING OF AI AND FREE WILL

As soon as an algorithm knows you better than you do, democracy and free choice become obsolete, and ultimate authority will shift from humans to algorithms (Harari, 2017b). Big data increases inequality and threatens democracy. O'Neil (2017) discusses algorithms that divide people into rough categories based on data in ways that cause immense harm—not least because the actual algorithms (and how they operated) remain in the dark. It is not known exactly what algorithms are measuring, and those who are checked by them are often unaware that they have been screened or are not told why. Of course, there are a few people who benefit, but the victims are more numerous and striking. In the well-known movie *Minority Report*, an algorithm can predict when a person is likely to commit a crime long before they have even thought about it, and thus be held accountable for it. This has already become a partial reality with the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software, an algorithm that makes predictions as to how likely a person is to commit a crime again. The software was widely used in the US until 2016, when a detailed investigation revealed that the algorithm is biased against Black people. The problem was not in the AI algorithm itself, but in the data with which it was fed (Chintada, 2021). In an interview by Fawn Fitter, Cathy O'Neil says that algorithms are not inherently fair or trustworthy just because they are mathematical. O'Neil is very direct:

> "Garbage in, garbage out" still holds.
>
> There are many examples. On Wall Street, mortgage-backed securities algorithms failed because they were simply a lie. A programme to assess teacher performance based only on test scores fails because it is simply bad statistics; moreover, learning is about much more than testing. (Fitter, n.d., "Q: If an algorithm" section)

Biased data can be used as the basis for certain calculations that can be crucial for important decisions. An example would be a survey targeting a specific group of people that is not representative of the population as a whole, which can lead to bias in the survey results. Harari (2017a) explains how data are being used to analyse every citizen, stating: "Doubting free will is not just a philosophical exercise. If an organism indeed lacks free will, it implies that we can manipulate and even control their desires by using drugs, genetic engineering, or direct brain stimulation" (p. 332). Deric Brownds notes, when reviewing Harari (2017a), that "as algorithms come to know us so well, authoritarian governments could gain absolute control over their citizens" (Brownds, 2018, Chapter 3 Liberty section). This is not a desirable reality. Even governments could also easily become obsolete in the relatively near future. If governments no longer possess these data, corporations such as Google, Facebook, Tencent, and Amazon may obtain them (Webb, 2019). So, who will make governmental decisions if the government itself no longer possesses these data? Even if these companies cannot yet fully control minds, they can certainly manipulate people in profound ways. Nevertheless, we, as human beings, are the entities supplying the data about ourselves. Consequently, before criticizing the ethics of

these companies and their use of our data, there are also ethical questions to ask of ourselves. The connection of ethics and cybernetics is helpful in this regard, as second-order cybernetic acts of (self) reflection and recursion can be ways to address ethical concerns (Foerster, 2003, Chapter 14; Sweeting, 2019). Currently, the act of conscious self-reflection is solely a human ability and does not apply to AI. For how much longer will this be the case?

Klaus Krippendorff, who unfortunately passed away in October 2022, traces the history of the ongoing tendency to surrender human agency to algorithm-driven technologies in exchange for greater convenience, with people getting caught up in the distinctions implied by these algorithms rather than being able to make and act on their own distinctions (Krippendorff, 2019; Scholte & Sweeting, 2022). In the talk *Agency, Algorithms, New Forms of Oppression, and How Cybernetics Might Respond*, Krippendorff (2019) argues that the uncritical attribution of intelligence and learning ability to algorithms, as in artificial intelligence and machine learning, degrades human and social intelligence.

Even if this sounds futuristic, it became somehow real in 2018 in what is known as the Cambridge Analytica Facebook scandal. This scandal led to widespread concern about the methods used by Cambridge Analytica to target voters through psychographic profiling algorithms based on Facebook user data (Hu, 2020). Hence, through the information gained by analysing people's web browsing histories over decades, AI could possess a truly thorough understanding of what a person is interested in more than that person will know themselves. AI can also compare these data with locations that the person has visited, and even listen through smart televisions to gather data that can be monetised (O'Flaherty, 2022). Will free will then become obsolete? As mentioned already in the introduction, it is also technology that is changing the environment. Addressing this possibility, Max Tegmark (2017) contends that the goals of AI need to be clearly delineated, stating: "The goal should be to create not undirected intelligence, but beneficial intelligence" (pp. 35–36). Eventually, and perhaps sooner than later, a "friendly AI" will be able to analyse all recorded history in seconds to evaluate the best option for a particular human being in a specific situation. This might even be sometimes positive for the user.

The term friendly AI was coined by Eliezer Yudkowsky (2008). Yudkowsky's work is concerned with making the world the best place that it can possibly be. Yudkowsky claims that a friendly AI has the goal to not hurt people and should be designed with this goal in mind from the beginning onwards of the creation of any AI system. It is critical for programmers to realize that their own designs are flawed, and that AI systems will learn and evolve over time, based on biases introduced by the programmers. Harmful human bias—both intentional and unconscious—could be avoided through the use of AI, but only if it is taught to play fair and constantly question given results (Fitter & Hunt, n.d.). A friendly artificial intelligence constitutes a hypothetical artificial general intelligence (AGI) that exerts a solely positive effect on humanity. Artificial general intelligence is the ability of an intelligent agent to understand or learn any intellectual task that a human can perform. Artificial General Intelligence is commonly characterized as possessing intelligence comparable to that of humans, but it has faced criticism for its potential to surpass human intelligence in certain domains. AGI is also often referred to as strong AI. A major problem would arise if this intelligence were to lead to goals that are misaligned with humans' own goals (Tegmark, 2017, p. 44). What can be anticipated when a human-like AGI is created? If an AGI is no longer perceived as a machine or algorithm, to what/whom is

one then communicating? In contrast, it has also been asserted that human beings may be simply shaped by algorithms, as claimed by Harari (Rubin, 2018). Another pertinent question would be: Can one presume that algorithms are always written by human beings, or are algorithms also written by algorithms? With the help of AutoML, artificial intelligence has the potential to develop additional "ChildAIs" and ChatGPT can generate code that may not meet high standards of quality.

How to determine the knowledge and understanding that an AI system has, and how to ensure that it is reliable and accurate? This question refers to assessing and validating the knowledge and understanding that an AI system has, which is an important aspect of understanding and using AI technology. So, what are the limits of the knowledge that AI can acquire and represent? From an ethical perspective, how can one ensure that AI systems are fair and unbiased in decision-making processes? Perhaps by critically examining the knowledge and understanding that an AI system has acquired. In this context, Yuk Hui (2021) notes that AI is limited by its reliance on binary logic and the formalization of knowledge, as everything is reduced to discrete units of information in a binary logic of true or false. Hui argues that this approach to knowledge and intelligence is limited because it does not do justice to the complexity and ambiguity of human experience. In addition, human experience is constantly changing and evolving, making it difficult for AI systems today to keep up. However, this could change with the development of AGI whenever it might become a reality.

Krippendorff argues that one needs to be able to critically question and evaluate the assumptions and beliefs that underpin one's understanding of the world in order to make sense of it. By maintaining one's agency, it is possible to actively engage the complexity and uncertainty of the world rather than simply passively accepting it. Recent discussions have emerged around the idea of "critical cybernetics" (Scholte & Sweeting, 2022). While it is clear that critical thinking is necessary, it remains to be seen whether critical cybernetics will continue to be embedded in second-order cybernetics or whether a new category of critical cybernetics will be added. In this context, it is useful to look at the discipline of critical design. Without going into too much detail, it is clear to say that through critical design (and arguably critical cybernetics) interdisciplinary collaboration, public involvement, education and research, participatory processes, and commercial contexts can be strengthened. Liene Jākobsone (2022, pp. 138–139) highlights the varying perspectives on public involvement in both design and social sciences, and emphasizes the potential significance of public participation in critical design. This is a strong assertion that also holds true for the field of cybernetics. This shows the urgency of interdisciplinary research and why cyberneticists can take knowledge from other fields, which for me is the essence of cybernetics—and the essence of being critical. Ben Sweeting and Sally Sutherland (2022) put it this way: "The making of analogies between these contexts is at the heart of what has always been distinctive about cybernetics, with ideas able to move between radically different domains" (p. 1).

2.2. HOW AI CAN LEAD PEOPLE TO BECOME TRIVIAL MACHINES

Currently it can be assumed that there is no AI algorithm with critical self-reflection in the sense of second-order cybernetics. Although, if such an algorithm is programmed in the future, it is another question whether it makes decisions about what it can glean from its perceived environment and thus forms its own reality (Glanville, 2012). Foerster (2003) elaborates on reality, writing "act always so as to increase the number

DOI: 10.58695/ec.4

of choices" (p. 227). Ernst von Glasersfeld (1995, pp. 41, 149), a prominent exponent of radical constructivism and colleague of Foerster, discusses both Kant and Foerster as emphasising the importance of the observer in shaping understandings of reality. Foerster's conception of ethics has similarities with Kant's emphasis on the importance of autonomy and the use of reason in moral decision making, but also departs from Kant's approach in some important ways. Foerster remarked in an interview with the *Stanford Humanities Review*: "'Of course I was raised with classical philosophy, I knew my Schopenhauer, I knew my Kant, I knew my this and that. ...but suddenly here comes Wittgenstein....Wow!'" (Franchi, Guezeldere & Minch, 1995, para. 5). Foerster (1993) discussed Kant's belief that the efficient cause was the ultimate principle—that every event occurs for a specific reason. As a critique of Kant, Foerster (1993, p. 97) cited Wittgenstein's viewpoint that belief in the causal nexus is superstition. Foerster's approach to ethics provides an ethical framework that places human existence and capacity at the centre of a normative philosophy that guides the understanding of moral behaviour (Ulgen, 2017). Thus, Foerster's approach is more pragmatic, emphasises the importance of self-reflection, and does not rely on fixed moral principles. Foerster has titled this approach as "the ethical imperative" (Foerster, 2003, p. 227). Contrary to both Foerster and Kant is the Spinozist ecological view of ethics—that everything that happens in the world, including human actions, is the result of cause-effect relationships and is ultimately determined by the laws of nature. In my view, this cause-effect relationship is highly problematic as it means that human beings do not have free will, but that their actions are determined by their physical and mental state. By contrast, Kant held that reality is fundamentally unknowable and that the mind can only perceive things through its own structures of understanding. Kantian ethics refers to the (ego-logical) moral theory of ethics, which assumes that when someone acts, they follow a rule or maxim based on values. In Kant's practical reason, the noumena are also the postulates of practical reason, for example, the absolute freedom of the will. Insofar as knowledge is to be objectively valid, it must be based on sensory knowledge (of phenomena) (Hui, 2021, p. 351). Similarly, Foerster was concerned with the limits of human knowledge and the ways in which perceptions and beliefs can be influenced by subjective experiences and prejudices. Both Kant and Foerster recognised the importance of language and communication in shaping understandings of the world. For Foerster, understanding is an active process that involves the mind's ability to organise and interpret sensory data.

Foerster's ethical imperative promoting an unlimited number of choices can be criticized since it may lead to unethical outcomes by providing more options without improving quality. More choices can indeed lead to confusion, indecision, and lack of responsibility. Mick Ashby, who sadly also died in 2022, notes that while the principle to increase the number of choices can be valuable in the context of psychological therapy, it is flawed in that it specifies no end condition, i.e. when to stop adding more choices (Ashby, 2020a, p. 3). Ashby's criticism of Foerster would apply if cybernetics were understood as a system that can be built, or if cybernetics were reduced to this view. Ashby proposed the idea of an "ethical regulator" to ensure ethical behaviour in AI systems (Ashby, 2020a). The concept of the ethical regulator involves programming ethical principles into AI systems, so that they make decisions based on ethical criteria, rather than simply following a set of programmed rules or maximizing some defined goal. This is, in my understanding, a technological implication of values—which Ashby does not yet define. Moreover, it is difficult to create a global norm for these values, as different societies have differing values. Foerster critiqued standard ethical principles in decision making, advocating for a nuanced, context-sensitive approach

DOI: 10.58695/ec.4

that acknowledges humanity's complexity and ethical issues. Consequently, Ashby's view of Foerster on this point is, in my opinion, wrong. Paul Pangaro (2011) attempts to clarify Foerster's ethical imperative by putting forward the idea of using recursive processes to understand complex systems and phenomena, in which each iteration builds on the previous one to create a more comprehensive understanding. Citing a personal communication with Foerster, Pangaro (2011) notes that: "Heinz specifically did not like the word *options* because it conjures a vision of a large number of arbitrary variations. What is desired is a thoughtful set of viable alternatives, each one a path to effective action in the context of our goals" (p. 141).

Ashby's approach would be appropriate for developing an algorithm if society agrees on the ethical values to be employed. But there is, of course, the fundamental question of what ethical principles mean for different societies. Health, for example, can be quickly defined, but for happiness, equality, and justice, the question becomes more complicated. Values, which could be linked in some way to intelligence, as pointed out by Yuk Hui, are realizable through digital apparatuses, which in turn means that they are computable (Hui, 2021, p. 349). But it is worth noting that Hui relates the term computable as only one type of intelligence among others, an intelligence with a technical tendency. Hui suggests that AI can overcome these limitations by moving beyond binary logic and taking a more dynamic and relational approach to knowledge. This would mean recognizing the importance of context and experience in the formation of knowledge and understanding. In this way, AI systems would be better able to capture the complexity and richness of human experience and engage with the world in more nuanced and meaningful ways. The next step is the AGI, for Hui a superintelligence, which is the expression of an extreme form of computationism and humanism, according to which the world is calculable and can be exhausted by calculation. It is also the highest form of neutralisation and depoliticization through technology and in line with Stiegler's thesis, introduced above.

This technological approach becomes clear when Ashby (2020a) defines their ethical regulator theorem, as Ashby speaks of ethics expressed in clearly prioritised rules. But, of course, rules apply in algorithms and therefore I think it is wise to return to second-order cybernetics, as this is a human approach to the problem. Pangaro (2011) relates to this matter by stating that "second-order cybernetics and ethics are constant clarifiers in my daily efforts to design software applications" (p. 139). In this context, it becomes interesting to consider artificial intelligence as part of human-machine relationships and thus follow Catrin Misselhorn (2018), who writes:

> You can either program rules into the system that have previously been checked with the categorical imperative, or you can implement the categorical imperative as a method that enables an artificial system itself to apply rules of conduct based on its moral admissibility check. (p. 102)

As a consequence, the following question arises: Would it be possible to implement the ability to enact the categorical or ethical imperatives within the design of the new intelligent algorithms? How this could be done is described in detail by Ashby (2020b) in a paper about how to apply the ethical regulator theorem to crises. Here I think Ashby is not quite clear which side of Misselhorn's question to come down on. I can't imagine Ashby thinks there is a checklist of rules for embedding the ethical regulator theorem, so I think it is about embedding the method of the imperative. But doesn't that in turn mean that an algorithm has to be developed for this, which in turn would be based on rules? So, are the boundaries between the two options fluid? Ashby refers

here to the ethical dimension, which defines the boundaries set by laws, regulations, strong cultural conventions, known public values, and procedures to be followed. These must follow basic ethical principles such as honesty, respecting human rights, and to not harm people by one's actions or inactions (Ashby, 2020b, p.57). This is also in line with Pangaro's (2021) views for a humanistic future:

> Alternatives to today's AI can place technology in an organic and social frame that conserves what it means to be human and encourages a rich, vibrant, inter-connected life. Concepts of interaction, information, and intelligence can be "brought to life" from an alternative and analog worldview that is grounded in biology and human intention. (p. 5)

The pertinent question to be asked here is how can this become possible if strong AI in the future will be qualitatively faster and will possess exponentially better ways to analyse its options, perhaps even comparing its options to literally all existent options? This is where the concept of "super-ethical systems" comes into play (Ashby, 2020a, p. 21).[1] According to Ashby, there are currently two possibilities for the future. Either there are good super-ethical AIs that (or, at this point perhaps, who) protect humanity and the biosphere, or there are evil super-unethical AIs that dominate humanity and destroy the biosphere. Ashby speaks here of the *singularity*, which can be understood as the development of artificial general intelligence (AGI). The term singularity usually refers to the idea that at some point in the future, artificial intelligence will reach a level of intelligence that significantly surpasses human intelligence. AGI refers to an AI system that is capable of understanding or learning any intellectual task that a human can do. AGI would be able to have common sense, but also perform tasks that are specific to humans, such as creativity, emotions, and self-awareness. I see the idea of the singularity as a special kind of AGI with the ability to improve itself and create a recursive loop of intelligence. This could lead to a rapid acceleration of the AGI's intelligence and capabilities. Fortunately, such an AGI does not yet exist, and human beings are unique in possessing such a deep self-awareness. In Foerster's sense, such an AGI may be in the position to fully navigate people's choices and make them into trivial machines.

Foerster took up the distinction between trivial and non-trivial machines from machine theory and elaborated it paradigmatically (Baecker, 2018, p. 180). A trivial machine is a simple machine, in which input A predicts outcome B. According to Foerster, such a trivial machine never creates something new, it simply performs the task for which it was programmed, and thus the output that results from the input creates trivial knowledge (Glanville, 2003). AI has the potential to trivialise humanity by removing the subjective and emotional aspects of decision-making and reducing human decision-making to a purely rational process. This could lead to a loss of empathy and understanding and a lack of appreciation for the unique qualities that make people human. While humans should not be treated as trivial machines, such algorithms are already being applied, including the mentioned example of the COMPAS Software. In this way, a strong, human-like AI may be able to predict outcomes through analysing sources and their data, but the accuracy of such outcomes is not guaranteed. Although algorithms may create logical cause-and-effect trivial knowledge, when applied to individuals or groups of human beings, outcomes may be wholly unanticipated. It is important to

1 Somehow the "super" reminds me of Rittel & Webber's wicked problems, where very complex problems like climate change are now sometimes called "super wicked problems" (Levin et al., 2012)—I doubt whether it really needs this distinction, because ultimately it remains a wicked problem. Likewise, I don't think a strong AI with ethical foundations needs a "super". Besides, what's next? An "ultra system"?

note that these potential negative impacts of AI can be mitigated if society actively works towards the ethical development of AI technology. I also strongly believe that it is humanity's adaptability that provides reasons to be positive about the future. Indeed, human beings are characterized by unpredictability and wonder (i.e., they are non-trivial), and it is currently unclear if an AGI will ever be able to encompass this.

An AGI based on unpredictability and wonder would be fun. Glanville notes that it is this unpredictability that was of most interest to Foerster. Glanville (2003) writes:

> We have a model for a world we inhabit where what we observe may change in ways we cannot imagine. And that means we are never truly in control, that we can and must keep learning—maintaining our involvement. The world of the non-trivial machine, as if by magic, creates surprises and cannot be tamed by us. (p. 99)

The results of an algorithm, such as Cambridge Analytica or COMPAS, are often dubious and do not reflect reality. Unfortunately, people in such situations often do not have much of a choice in such matters. In fact, most people who participate in such an algorithm are unaware of the outcome and the consequences. This issue of trust has been already mentioned and will be addressed further below.

Considering the remarks of Tegmark (2017), humans' current form of life requires an update: "Life 2.0 can redesign much of its software: humans can learn complex new skills-for example, languages, sports and professions-and can fundamentally update their worldview and goals" (p. 26). To speak of "updates" in the context of humans seems rather antihuman but I will follow Tegmark's logic for this section as it at least corresponds to that of describing a comparison between humans and non-trivial machines as above. Tegmark generally defines life in three steps: Life 1.0 is biological evolution; life 2.0 is cultural evolution; and life 3.0 is technological evolution. These three categories are defined by the human body, which represents hardware, and knowledge, which is viewed as software. Life 1.0 cannot redesign its software or its hardware. Life 2.0 can update its software, e.g., by learning a new language which enables further adaption, and is one of the capabilities that makes human beings successful. Hardware has a limited lifespan, which is not the case for algorithms. Algorithms can be understood as part of life 3.0, which constitutes an unlimited lifespan (although this is not yet clear in my opinion) and the capacity to acquire new software to adapt optimally to environments. AGIs, should they be realised, would be particularly capable of this because they would have the ability to control their own development and improve without human intervention. The exciting and fundamental question can now be asked: How do we as humans deal with an AGI? The integration of computers into daily lives has been so rapid that humans have not had time for a long evolutionary process to adapt to these new circumstances. Life is only possible through an evolutionary process of adapting to a new environment but an environment in which our brains can cope with technological adaptations has not yet been encountered (Koering, 2019, p. 49).

Adam Gazzaley asserts that it is now time to adapt our brains as well as possible to keep up with algorithms (Kleber & Anderson, 2016, 2:44), i.e., further changes to the brain (human body) need to be made. Gazzaley does not precisely expand on what our brains must adapt to. Perhaps Gazzaley alludes to new drugs or advancements in genetic engineering to achieve an infinite lifespan. Interestingly, this is in line with what Norbert Wiener (1954) already proposed decades ago by stating: "We have modified our environment so radically that we must now modify ourselves in order to exist in

this new environment" (p. 46). In contrast to Tegmark or Gazzaley, Harari makes clear that this evolution is only an option for a relatively small group of people. Harari predicts that elites will reach for immortality, while the vast majority of people will be unable to do so (Trappendreher, 2017). Whether or not this is a desirable future may be up to the individual to decide. It is clear, though, that this so-called evolution will not solve humans' most pressing concerns, e.g., climate change, biodiversity loss, resource depletion, etc.

3. ARE ETHICAL STANDARDS THE NECESSITY TO BUILD TRUST IN AI?

3.1. LACK OF TRUST IN HUMAN'S COEXISTENCE WITH AI

A major issue involving the relationship between humans and AI is the problem of humans losing trust in AI systems. Examples of this problem were described above—the Facebook-Cambridge Analytica data scandal and the COMPAS software. Another reason could be the typical black box problem: Without knowing how a system (AI) comes to its results and conclusions, one does not trust the system. An option to overcome this challenge is found in a work by Alan Winfield and Marina Jirotka (2017), who propose an ethical black box for AI systems. This idea could be realised along similar lines to Ashby's proposal for an ethical regulator (2020a). The basic idea behind the ethical black box is to create an AI system that can think through ethical dilemmas and make decisions based on a set of predefined ethical principles. The ethical black box would ensure that AI systems make decisions that are compatible with human values and ethical principles, even in situations where the AI system's goals might conflict with human values. Such an ethical black box creates transparency because, for instance, it can analyse why a robot caused an accident, which then establishes accountably and responsibility in AI systems. It will also enable the inspection of incorrect decisions made by an AI system and update the algorithms accordingly. These updates, of course, must be made according to agreed-upon ethical standards, as made clear by Winfield and Jirotaka (2017). However, for precisely what ethical standards should human beings aim? Is it possible to develop trust even in this ethical black box, since this box will probably inhabit AI systems by itself and largely determine their behaviour? Ashby's concept of the ethical regulator is still in the research stage and has not yet been implemented in real-world applications, but it is an interesting approach to ensuring the ethical behaviour of AI systems. However, Ashby is not entirely clear on how ethical values should be integrated into an algorithm, which is fundamental to AI, and since Ashby passed away last year, it remains to be seen how these ideas will be followed up.

A central tenet of the Future Life Institute, a volunteer-led research and networking organization of which Tegmark is president, and which seeks to reduce the existential risk of AGI (strong AI), is that AI should always benefit humans. This may seem abstract: Precisely what does being beneficial to humans mean? Does it concern health, which also comprises cost and the availability of diverse environments? Or does it relate more to the happiness that humans experience within their lifespan and responsibility for the happiness of future generations? It could then also be asked, who defines happiness? What if different individuals, cultures, and societies have mutually exclusive definitions of happiness? A consensus of the Future Life Institute is the paramount importance of the development and assurance of AI safety. The Future Life Institute predicts that a strong AI will be developed sometime by the year 2055, although some researchers guessed hundreds of years or more (Tegmark, 2017, p. 42). Consequently, it is a sensible and logical decision to begin now to define and ensure AI safety because this task may require decades of work.

A well-known, moral regulation was proposed by Isaac Asimov (1968) as the three laws of robotics:

1. A robot may not injure a human being, or, through inaction allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. (p. 8)

Although these moral laws can be applied to AI, they are abstract. The laws would also not work for certain military weapons, such as drones and rockets, as they have other purposes, which are often to spy on, injure, or kill human beings, constituting another serious ethical debate. In a wider sense, applying the first law to robots in production, they would only be allowed to function in an environmentally sustainable manner. Otherwise, although the specific work that a robot may do could be beneficial to human beings, such robot would be ultimately harming human beings if they produced pollution or caused environmental degradation. It must be acknowledged that such questions regarding algorithms, and thus AI systems, are up to the rational actions of humans, i.e., it is solely up to the programmer or funding institution to determine what ethical standards should be incorporated into the algorithms of AI systems (Koering, 2019, p. 20). Asimov seems quite cognizant of this issue, believing that, ideally, people also follow these laws (Madgwick, 2017).

These three laws are thus a step towards a simple definition of moral rules for robots, whereby there should be no difference between moral rules for humans and machines because, if one assumes a human-like AGI, the same moral rules should apply to both sides, otherwise one would be discriminating against the AGI. This again shows the profound problem with formulating values, as there are already problems with applying them to humans. One needs to apply ethics first to itself (Sweeting, 2019), and thus to human beings. From my perspective, this step relates to the need for ethicists to self-reflect and consider the implications and limitations of their own perspectives and biases on their work. Examining the field of ethics itself in this way can lead to a more robust and comprehensive ethical framework that can be applied to different areas of society, which here includes AI. Sweeting (2019) argues that this is rarely the case, as many forays into ethical discourse are made on the basis that they offer guidance for the good, with the result that questions about the way such guidance is debated or communicated is obscured. This can certainly be applied to the development of AI, as the lack of ethical values and the resulting debate, including this paper, are nothing but the mirror of society. If the guidelines for AI were enshrined in law, this discussion would not exist. Nevertheless, it should be mentioned that first steps are being taken (Wahlster & Winterhalter, 2022).

Another problem arises from moralizing. Monika Broecker (2003) writes about a conversation with Foerster, in which Foerster stated that ethics needs to be implicit. Again, this is only possible with awareness and the ability to be self-critical, which are currently beyond the capabilities of AI. Although it is hard to predict how this will develop in the future, it will surely become a central question about AGI. This underlines the call to start developing AI safety regulations now, because as long as AI is controlled and developed by humans, it is important to set ethical standards and guidelines for its use and to ensure that it is used in a responsible and safe way. In

addition, creating legally binding standards for AI can help build trust in new systems and ensure that they are used for the benefit of society. As AI technology continues to advance, it is critical to proactively address potential risks and ethical concerns to ensure a positive future for AGI development.

3.2. THE NEED FOR ETHICAL STANDARDS TO PREVENT SOCIAL EXCLUSION AND TO BUILD TRUST

It could be asked: Why should ethics be applied to the discussion of AI at all? As argued in previous chapters, the application of ethics to the discussion of AI is important because AI systems have the potential to impact society and individuals in profound ways, and it is crucial to ensure that their development and use are compatible with ethical values and principles. I have attempted to show that AI has an impact on humans by drawing on the views of Foerster, Stiegler, and Harari. When considering Foerster's trivial machines, ethical considerations play a vital role in ensuring the correlation between trivial machines and humans is consistent with ethical values. It's important to note that trivial machines are not to be confused with the complexity of humans, but the comparison helps to showcase ethical principles and the future of AI algorithms. By taking a humanistic, rational, and ethical approach to the development and use of AI, it could be ensured that these systems promote human well-being and do no harm. In this approach, ethics should not only be applied to the AI systems but should also be embodied in them, as previously discussed (Ashby, 2020a).

As humans, we are always part of our social environment. When we tell someone, including an AI system, "You should (not) do X" (c.f. Foerster, 2003, p. 289), we assume a certain dishonesty. Thus, we already imply to the AI that it may not act according to the programmed goals—it follows that we do not trust the system. Such an end to communication and the loss of ethical values leads to social exclusion, which is already occurring. This logical reasoning might have led to Asimov's fears, which are summarized in the three laws of robotics. I refer here to using ethical rules when creating computer programs to prevent situations where people are unfairly judged by software like COMPAS. Consequently, current researchers need to embed a code of ethics into the algorithms of future strong AI that will produce beneficial results for humankind. These ethics should take into account not only the specific cultures from which they originate but also the views and norms of different cultures, as they may differ. This is primarily about embedding an ethical code of self-reflection, "'I shall…', 'I shall not…'" (Foerster, 2003, p. 289). Does this then presuppose consciousness, in the sense of an AGI? And how does this AGI arise? Possibly not suddenly out of the blue, but rather in the form of a process. It is therefore a matter of bringing ethical values, which have been described above, into this development process, which must change and adapt over time to achieve its full ethical coherence. Ultimately it would be up to the AGI itself to decide whether it is acting ethically or not. This makes the development of AGI a mirror of humanity, because the same applies to both. Do humans always act ethically despite all their knowledge? Are humans not already in self-conflict with the desirable goals that Asimov, for example, tries to formulate in three rudimentary laws? And maybe what one thinks is ethical is unethical in certain situations?

Addressing the positive and negative versions of the singularity, Ashby (2020a, p.31) suggests that there could be a middle ground between these two options, where AI is used in a way that is beneficial to society but also carries minimal risks. As Ashby refers to ethics, it is perhaps appropriate to point out that there is a difference between moral and ethical thinking. Moral views often relate to certain standards for reacting

DOI: 10.58695/ec.4

to someone, while ethics concerns self-reflection, one's own ego and culture. Pangaro (2011) says that: "The difference lies in being responsible for our actions and being aware of that responsibility. From such awareness we may strive, at least, to reach a future we desire" (p. 140). Dirk Baecker (2018) formulates it as follows: "In modern society, morality as the assessment of what is morally appropriate and ethics as the question of the good life are still close to each other; in the next society, they diverge" (p. 208). By the next society, Baecker probably means the future addressed by Harari—that of the digitally transformed society. Sebastian Muehl (2022) describes the development between human and machine (here the cyborg as human-machine) as follows:

> Einerseits hat die Figur politischen Kampfwert erhalten, andererseits scheint die Frage, welche Freiheit für wen gemeint ist, unter dem Eindruck der neoliberalen Diskurse und fortschreitenden Klimakrise mehr denn je unbeantwortet [On the one hand, the figure [cyborg] has acquired political combat value; on the other hand, the question of which freedom applies to whom seems to be under the impact of neoliberal discourse and the advancing climate crisis. It seems more unanswerable than ever]. (p. 6)

Even though Muehl is specifically addressing the view of cyborgs and Donna Haraway's *Cyborg Manifesto* here, it is still very relevant to the way we as humans live with algorithms and the question of whose freedom is ultimately meant is quite valid. Even though I think the term cyborg is anti-human, it is a legitimate question to what extent this is already in a process of development. I think it is important to move away from the view of the mechanical extension of the human body to the adaptation of the brain to the digital world, a symbiosis so to speak. Ultimately, this symbiosis will give a better, expanded understanding of algorithmic environments from the perspective of second-order cybernetics. In such a symbiosis, the distinction between trivial and non-trivial machines no longer makes sense, and one can only hope that it is possible to trust these systems, with ethical goals playing an essential role in building that trust. I personally think that only with trust is it possible to strive for positive development.

As a result, it is necessary to first reflect on one's own ethical goals, which is possible with second-order cybernetics. These points can only be understood by possessing a certain consciousness and capacity for self-criticism. One must adhere to humanistic principles and ethics to achieve a positive future with the algorithms one lives by. Following the logic of life 3.0, it remains unknown how an AGI with an ultimate lifespan would react to ethical standards. If this debate seems largely negative, it should be noted that the facts are much more positive than negative in aggregate. For example, Hans Rosling (2018), in the book *Factfulness*, highlights that the proportion of people living in extreme poverty from 1800 to today has more than halved worldwide and 80% of all one-year-old children are vaccinated. A certain amount of credit for these and other great advancements, directly or indirectly, must be given to AI.

The overall cycle of development known as Kondratieff waves (Nefiodow & Nefiodow, 2014) describes the technical development of humankind from the steam engine to information technology in waves with all their ups and downs, a recession and a depression inevitably following a phase of prosperity. For example, the digital transformation of the workplace will certainly make many jobs obsolete, and indeed this process is already in progress (Frey & Osborne, 2013). Robots will make people redundant in numerous functions, and unemployment should rise steadily as a result. Martin Ford (2015) addresses this topic in the book *The Rise of the Robots*. Ford notes that this phenomenon is most evident in jobs in the service and retail sectors, where adaption is needed in order to achieve interdependence. ChatGPT, OpenAI's

DOI: 10.58695/ec.4

conversational AI, has dominated the headlines for weeks for similar reasons, putting different kinds of jobs in danger. This will then lead to more unemployment, and in turn restrictions on consumer spending, which constitutes the major driver for an economy. Ford's conclusion is that new political, ethical, and economic guidelines are required. Ford also insists that researchers need to question the social and economic impact of their work (Mills, 2017). This issue has led to the ethical idea of a universal basic income (UBI). A risk exists that UBI will become an industrial donation serving only to combat poverty (Schwartz, 2010). In this way, this scenario has been compared to an enormous welfare state, although it should also be mentioned that new technologies create new jobs. As per Wiener's (1954, p. 46) words, the modifications humans have made to their surroundings necessitate a corresponding alteration within themselves. In essence, adapting to change is crucial for success.

The socialist project Cybersyn, which existed in Chile from 1970–73, is relevant here. Cybersyn was a prototype of a data- and people-related idea intended to increase the country's production, while counteracting rising unemployment in an essentially socialist paradigm. Cybersyn can be considered as an early model only. Today's computational ability has made such systems realisable, but citizens are often excluded from employment and participation as a consequence (Koering, 2021). About Cybersyn, Eden Medina (2011) asserts: "It was a system designed to help the state regulate the nationalized economy and raise production without unemployment" (p. 211). Indeed, most of Cybersyn's earlier ideas are commonly held today, but do not exist under socialist governments. AI does not have an answer yet to the well-founded fear of a future high unemployment, hence the focus on creating a UBI to compensate for this. Fischer (2019, p. 295) points out that industrial automation has displaced many employees, and is forcing them to seek new jobs that focus on information and services instead of industry, often with low wages. Is this a desirable direction to pursue? It is necessary to discuss and decide on the role of human beings in AI systems, and perhaps ways to retain or even increase the level of human participation in the economy and other sectors. A commonly proposed remedy to AI-resultant job elimination is to invest in the education of human beings in the hope that new jobs will be created that cannot be done by AI. In the past, it was the fact that humans, unlike AI, have social and creative intelligence as well as unpredictability. An AI had to learn this first (McDermott, 2014). However, this has already become a reality in the case of text with ChatGPT or in that of image processing with software such as Dall-E and Midjourney. The question of authorship is rightly mentioned here, but also the danger of AI-generated images that, in the wrong hands, are deliberately misleading. However, raising awareness now could eliminate the risk of unemployment by enabling employees to make prudent career choices in advance by identifying which jobs will be especially needed in the future. It is certainly hoped that these jobs will also follow common ethical standards to benefit human beings.

4. CONCLUSION

In this paper I raised a friendly dispute about the different perspectives of Stiegler, Foerster, and Harari on ethical, societal and AI issues. The resolution of friendly disputes often requires a nuanced understanding of the relevant ethical principles and values, as well as the ability to listen to and understand the perspectives of all parties involved. The goal of this dispute resolution is to find a mutually acceptable agreement or solution that is in the interest of human values. As mentioned at the outset, by human values I refer to the fundamental beliefs and principles that guide behaviour and shape approaches to life. These values include things like truth,

honesty, loyalty, love, and peace. Basic human values serve as a moral compass and form the basis for ethical behaviour. In conclusion, it is the ethical thinking of Foerster which is especially pertinent to issues concerning human beings and AI systems. Foerster's explanation of a non-trivial machine leads to fundamental questions that only we (humans) can answer, and thus requires personal responsibility (Fischer, 2019, p. 297). It is obvious that humans can be seen as data emitters, showing that a clear act of self-responsibility is required. If humans decide to act as trivial machines, or do so through indecision, all knowledge will be predictable, and nothing new will be generated. Actions will lack deliberation or self-reflection. This is a future that most people would agree is undesirable and possibly repulsive. Then again, comparing beings to trivial and non-trivial machines is itself de-humanizing (Foerster, 2003, p. 208). Thus, implicit ethics should refer to unconscious or subconscious ethical beliefs, values, and norms that shape a person's behaviour and decision-making. These ethics are not explicitly formulated or written, but arise from a person's upbringing, cultural background, personal experiences, and social conditioning. To chart out and achieve a future characterized by freedom, safety, progress, and happiness, it will be necessary for researchers, scientists, government officials, and indeed all human beings to view themselves as global citizens who are capable and responsible of making rational choices. I believe it is essential to develop a universal code of ethics that is consistent with the Geneva Conventions of human rights and more or less globally applicable—and to ensure that it is an integral and permanent aspect of AI systems for a positive future. First steps in this direction have already been taken at global (World Commission on the Ethics of Scientific Knowledge and Technology, 2017), international, European (High-Level Expert Group on Artificial Intelligence, 2019) and national levels (Wahlster & Winterhalter, 2022), but it remains to be seen whether these guidelines can also be successfully implemented. Leon Trotsky (1924/2005), the influential Marxist revolutionary, theorist, and Soviet politician, once said: "Art, it is said, is not a mirror, but a hammer: it does not reflect, it shapes" (p. 120). Trotsky believed that art should be politically engaged and serve as a means to inspire change. In this sense, art should be used as an instrument to criticize and change society, to challenge the status quo and make people think, rather than simply reflect it. Perhaps now is the time to rephrase this famous quote: Artificial intelligence is not a mirror held up to reality, but a hammer with which to shape it. I understand the term hammer here in the sense of a tool, and I would like to see AI as a tool that assists humans in creating something new. The idea behind the reformulation of Trotsky's quote is that AI is not simply a passive reflection of reality, but an active force that will shape our society and change the world around us.

ACKNOWLEDGEMENTS

The basic idea for this position developed from the topic of my doctoral thesis. Many thanks to Michael Hohl for motivation, honesty, and critical comments. Finally, gratitude is also due to Paul Pangaro and Eva Sommeregger for their comments on earlier versions of this paper.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Dietmar Koering orcid.org/0000-0003-1390-428X
Institute of Contemporary Art, Design and Architecture/Art Academy of Latvia, Latvia

REFERENCES

Antone, P., Ili. (2008). *Design and the elastic mind*. Museum of Modern Art.

Ashby, M. (2020a). Ethical regulators and super-ethical systems. *Systems*, *8*(4), Article 53. DOI: https://doi.org/10.3390/systems8040053

Ashby, M. (2020b). How to apply the ethical regulator theorem to crises. *Acta Europeana Systemica*, *8*(1), 53–58. DOI: https://doi.org/10.14428/aes.v8i1.56223

Asimov, I. (1968). *I, robot*. Panther Books.

Baecker, D. (2018). *4.0 oder die lücke die der rechner lässt* [4.0 or the gap left by the computer]. Merve Verlag.

Beabout, G., & **Wennemann, D.** (1993). *Applied professional ethics: A developmental approach for use with case studies*. University Press of America.

Broecker, M. (2003). Between the lines: The part-of-the-world-position of Heinz von Foerster. *Cybernetics and Human Knowing*, *10*(2), 51–65.

Brownds, D. (2018, December 31). Yuval Harari— "21 Lessons…" abstracted—Part 1—The technological challenge. *Deric's Mindblog*. https://mindblog.dericbownds.net/2018/12/yuval-harari-21-lessons-abstracted-part.html.

Bluemink, M. (2020, January 23). Stiegler's memory: Tertiary retention and temporal objects. *3:AM Magazine*. https://www.3ammagazine.com/3am/stieglers-memory-tertiary-retention-and-temporal-objects/.

Chintada, S. (2021, October 25). Addressing AI's biggest problem: Trust. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2021/10/25/addressing-ais-biggest-problem-trust/.

Descola, P. (2011). *Jenseits von natur und kultur* [Beyond nature and culture]. Suhrkamp.

Fischer, T. (2019). Einführung [Introduction]. In S. Hoeltgen (Ed), *Medientechnisches wissen: Band 2: Informatik, programmieren, kybernetik* [Media technology knowledge: Volume 2: Computer science, programming, cybernetics] (pp. 275–301). De Gruyter. DOI: https://doi.org/10.1515/9783110496253-020

Fitter, F. (n.d.). Unmasking unconscious bias in AI [Interview]. *SAP*. https://www.sap.com/insights/viewpoints/thinkers-unmasking-unconscious-bias-in-ai.html.

Fitter, F., & **Hunt, S.** (n.d.). How AI can end bias: Artificial intelligence (AI) can help avoid harmful human bias, but only if we learn how to prevent AI bias as well. *SAP*. https://www.sap.com/uk/insights/viewpoints/how-ai-can-end-bias.html.

Foerster, H. von. (1993). *Kybernethik*. Merve Verlag.

Foerster, H. von. (2003). *Understanding understanding: Essays on cybernetics and cognition*. Springer.

Ford, M. (2015). *The rise of the robots: Technology and the threat of mass unemployment*. Oneworld Publications.

Franchi, S., Güzeldere, G., & **Minch, E.** (1995). Interview with Heinz von Foerster. *Stanford Humanities Review*, 4(2). http://web.archive.org/web/20000819134725/http:/www.stanford.edu/group/SHR/4-2/text/interviewvonf.html.

Frey, C. B., & **Osborne, M.** (2013). *The future of employment: How susceptible are jobs to computerisation?* [Working paper]. Oxford Martin Programme on Technology and Employment. https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment/.

Glanville, R. (2003). Machines of wonder and elephants that float in the air. *Cybernetics and Human Knowing*, *10*(3–4), 91–105.

Glanville, R. (2012). Radical constructivism = second order cybernetics. *Cybernetics and Human Knowing*, *19*(4), 27–42.

Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning*. The Falmer Press.

Gorny, R. A., & **Radman, A.** (2022). From epiphylogenesis to general organology. *Footprint*, *16*(1). DOI: https://doi.org/10.7480/footprint.16.1.6291

Harari, Y. N. (2017a). *Homo deus: A brief history of tomorrow*. Penguin Random House. DOI: https://doi.org/10.17104/9783406704024

Harari, Y. N. (2017b, September 22). Life 3.0 by Max Tegmark review—We are ignoring the AI apocalypse. *The Guardian*. https://www.theguardian.com/books/2017/sep/22/life-30-max-tegmark-review.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Hu, M. (2020). Cambridge Analytica's black box. *Big Data & Society*, *7*(2). DOI: https://doi.org/10.1177/2053951720938091

Hui, Y. (2021). On the limit of artificial intelligence. *Philosophy Today*, *65*(2), 339–357. DOI: https://doi.org/10.5840/philtoday202149392

Hutnyk, J. (2012). Proletarianisation. *New Formations*, *77*, 127–149. DOI: https://doi.org/10.3898/NEWF.77.08.2012

Jākobsone, L. (2022). *Critical transition design: The role of critical design in fostering sustainability* [Summary of doctoral dissertation]. Art Academy of Latvia. https://www.lma.lv/uploads/news/3574/files/lienes-jakobsones-kopsavilkums.pdf.

Kleber, C., & **Anderson, A.** (Writers & Directors) (2016, June 19). Schöne neue welt [Brave new world]. ECO Media GmbH; ZDF.

Koering, D. (2018). The technobody: An animatronic artefact as manifestation of second-order cybernetics. In *Annual convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB 2018): Liverpool, United Kingdom 4–6 April 2018*. Society for the Study of Artificial Intelligence & Simulation of Behaviour; Curran Associates. https://www.proceedings.com/41519.html.

Koering, D. (2019). *Conscious city laboratory: Explorations in the history of computation, cybernetics, and architecture; foresight for artificial intelligence and human participation within cities* [Doctoral dissertation, Technische Universität Berlin]. DepositOnce: Technische Universität Berlin. DOI: https://doi.org/10.14279/depositonce-8466.

Koering, D. (2021). What we should have learned from Cybersyn: An epistemological view on the socialist approach of Cybersyn in respective of Industry 4.0. In B. Vassileva, & M. Zwilling, *Responsible AI and ethical issues for businesses and governments* (pp. 68–79). IGI Global. DOI: https://doi.org/10.4018/978-1-7998-4285-9.ch005

Krippendorff, K. (2019, June 24–26). *Agency, algorithms, new forms of oppression, and how cybernetics might respond* [Paper presentation]. Acting Cybernetically: 2019 Conference of the American Society for Cybernetics, Vancouver, BC, Canada. https://www.youtube.com/watch?v=PhehTApQi1s.

Levin, K., Cachore, B., Bernstein, S., & **Auld, G.** (2012). Overcoming the tragedy of super wicked problems: Constraining our future selves to ameliorate global climate change. *Policy Sciences*, *45*(2), 123–152. DOI: https://doi.org/10.1007/s11077-012-9151-0.

Mills, A. (2017, September 26). Rise of the robots: Interview with Martin Ford. *Michigan Tech*.https://www.mtu.edu/unscripted/2017/09/rise-robots-interview-martin-ford.html.

Madgwick, P. (2017, September 01). Revisiting Asimov's three laws: Ethics for robots or researchers? *Medium*. https://medium.com/@philmadgwick/revisiting-asimovs-three-laws-ethics-for-robots-or-researchers-b9771b15a306.

McDermott, J. (2014, February 10). Is your job safe in the second machine age? *Financial Times*. https://www.ft.com/content/c8901cc7-d879-3fb7-89ea-16aab9bec3e7.

Medina, E. (2011). *Cybernetic revolutionaries: Technology and politics in Allende's Chile*. MIT Press. DOI: https://doi.org/10.7551/mitpress/8417.001.0001

Misselhorn, C. (2018). *Grundfragen der maschinenethik* [Basic questions of machine ethics]. Reclam.

Muehl, S. (2022). *Wir sind alle flechten. Von der cyborg-metapher zur symbiose der arten* [We are all braids. From the cyborg metaphor to the symbiosis of species]. In S. Hurtig, CBRG.SPACE [Exhibition leaflet]. Zentrum für zeitgenössische Kunst, Leipzig. https://www.researchgate.net/publication/366669807_Wir_sind_alle_Flechten_Von_der_Cyborg-Metapher_zur_Symbiose_der_Arten.

Nefiodow, L., & **Nefiodow, S.** (2014). *Über die Kondratieffzyklen* [Kondratieff Cycles]. https://www.kondratieff.net/kondratieffcycles.

O'Flaherty, K. (2022, January 29). What your smart TV knows about you—And how to stop it harvesting data. *The Guardian*. https://www.theguardian.com/technology/2022/jan/29/what-your-smart-tv-knows-about-you-and-how-to-stop-it-harvesting-data.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin.

Pangaro, P. (2011). Invitation to recursioning: Heinz von Foerster and cybernetic praxis. *Cybernetics and Human Knowing*, *18*(3–4), 139–142.

Pangaro, P. (2021). *Responding to the pandemic of "today's AI"* (Version V5.1 December 2021). https://tinyurl.com/newmacy-distro-r.

Roberts, B. (2007). Introduction to Bernhard Stiegler. *Parallax*, *13*(4), 26–28. DOI: https://doi.org/10.1080/13534640701682776

Rosling, H. (2018). *Factfullness: Ten reasons we're wrong about the world—And why things are better than you think*. Sceptre Books.

Rubin, C. T. (2018, June 28). Algorithmic man: Yuval Noah Harari's timid transhumanism. *Public Discourse*. https://www.thepublicdiscourse.com/2018/06/21562/.

Scholte, T., & **Sweeting, B.** (2022). Possibilities for a critical cybernetics. *Systems Research and Behavioral Science*, *39*(5), 986–989. DOI: https://doi.org/10.1002/sres.2891

Schwartz, E. M. (2010). *Poverty reduction for profit? A critical assessment of the bottom-of-the-pyramid approach and of the 'opportunities for the majority'-initiative of the Inter-American Development Bank* [Master's thesis, University Vienna]. U:theses: The theses repository of the University of Vienna. https://utheses.univie.ac.at/detail/9023#.

Stiegler, B. (1998). *Technics and time, 1: The fault of Epimetheus*. Stanford University Press. DOI: https://doi.org/10.1515/9781503616738

Stiegler, B. (2010a). *Technics and time, 3: Cinematic time and the question of malaise*. Stanford University Press.

Stiegler, B. (2010b). *For a new critique of political economy*. Polity.

Sweeting, B. (2019). Applying ethics to itself: Recursive ethical questioning in architecture and second-order cybernetics. *Kybernetes*, *48*(4), 805–815. DOI: https://doi.org/10.1108/K-12-2017-0471

Sweeting, B., & **Sutherland, S.** (2022). Cybernetic transdisciplinarity as pedagogy. *Proceedings of the International Society for the Systems Sciences*, *66*(1). https://journals.isss.org/index.php/jisss/article/view/4050.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Allen Lane.

Trappendreher, P. (2017, April 21). *Der mensch als auslaufmodell* [Man as a discontinued model]. Spektrum. https://www.spektrum.de/rezension/buchkritik-zu-homodeus/1451773.

Trotsky, L. (2005). *Literature and Revolution* (W. Keach, Ed. & R. Strunsky, Trans.). Haymarket Books. (Original work published 1924).

Ulgen, O. (2017). Kantian ethics in the age of artificial intelligence and robotics. *QIL, Zoom-in*, *43*, 59–83. http://www.qil-qdi.org/kantian-ethics-age-artificial-intelligence-robotics/.

Wahlster, W., & **Winterhalter, C.** (Eds.) (2022). *Deutsche normungsroadmap künstliche intelligenz, ausgabe 2* [German standardisation roadmap artificial intelligence, issue 2]. DIN; DKE. https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki.

Webb, A. (2019). *The big nine: How the tech titans and their thinking machines could warp humanity.* PublicAffairs.

Wiener, N. (1954). *The human use of human beings: Cybernetics and society* (2nd ed.). Doubleday Anchor Books.

Winfield, A. F. T., & **Jirotaka, M.** (2017). The case for an ethical black box. In Y. Gao, S. Fallah, Y. Jin, Y., & C. Lekakou (Eds.), *Towards autonomous robotic systems: 18th annual conference, TAROS) 2017, Guildford, UK, July 19–21, 2017: Proceedings*. Springer. DOI: https://doi.org/10.1007/978-3-319-64107-2_21

World Commission on the Ethics of Scientific Knowledge and Technology. (2017). *Report of COMEST on robotics ethics*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000253952.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom, & M. M. Ćirković, *Global catastrophic risks* (pp. 308–345). Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198570509.003.0021

TO CITE THIS ARTICLE:
Koering, D. (2023). Exploring the human-AI nexus: A friendly dispute between second-order cybernetical ethical thinking and questions of AI ethics. *Enacting Cybernetics*, *1*(1): Article 5. https://doi.org/10.58695/ec.4

Submitted: 13 February 2023
Accepted: 12 April 2023
Published: 08 May 2023