Volume 15.1 | 2023

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

The Hermeneutics of Artificial Intelligence

Guest Editors: Joshua D.F. Hooke, Sean McGrath, Joachim Rathmann, George Saad

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

THE HERMENEUTICS OF ARTIFICIAL INTELLIGENCE

VOLUME 15.1 | 2023

ARTICLES

REVIEWS

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Editors' Introduction

Joshua D.F. Hooke
Memorial University of Newfoundland, Canada
ORCID: 0009-0002-8794-8488
Sean McGrath
Memorial University of Newfoundland, Canada
Joachim Rathmann
Universität Augsburg, Germany
ORCID: 0000-0001-5533-2617
George Saad
Memorial University of Newfoundland, Canada
ORCID: 0009-0005-2812-2274

The papers in the following volume are the outcome of a three-year long interdisciplinary research project. The project began with an in-person meeting hosted and funded by the Daimler und Benz Stiftung in Germany in March 2020 (the world was shutting down one nation at a time as we met). During the pandemic we continued to meet monthly online with support from Memorial University of Newfoundland. From the beginning it was the goal of the Working Group on Intelligence (WGI), as we called ourselves, to broaden and deepen the AI debate with a more nuanced understanding of intelligence than is common in cognitive and computer science discussions of AI. We wished to draw on the history of philosophy, ecology, and the philosophy of mind to establish that intelligence is meant in many senses, to use an Aristotelian expression. The clarification of these various meanings is essential to the discussion around the ethics of AI, especially the question concerning the possibility of strong AI or Artificial General Intelligence.

The consensus of the WGI was that intelligence is common to all animals and in this sense can be called natural and perhaps even common to all living beings. Yet it has a specific difference in humans where it becomes intentional or self-reflexive. The question of where or when human intelligence will have been surpassed by our

machines would need to take such distinctions into consideration. Human intelligence, whatever else it might be, cannot be reduced to rule-following, which is the way machines learn, but includes an intention toward truth. Such an intention, we concluded, would need to be manifest in some sense in a machine before we could conclude that it was more than 'artificially' intelligent. Put this way, it became clear to many of us in the group that a machine intelligence which intends to know the truth is hardly what is being sought in this multi-billion dollar industry. Such intentionality is not needed if efficiency in data analysis and manipulation is the true goal.

A first collection of papers, proceedings of the German meeting, was edited by Uwe Voigt and Joachim Rathmann and published in Germany under the title *Natürliche und künstliche Intelligenz im Anthropozän* (Darmstadt: Wissenschaftliche Buchgesellschaft, 2021). This current volume includes translations of some of those pieces, most of which were originally written in German, as well as newer contributions that arose out of the online meetings of the WGI.

As this volume was being prepared for publication the large language models of AI were unleashed on the world (ChatGPT, etc.). And while this was much sooner than many of us expected, it did not change the results of our research. ChatGPT is still only a functional mimic of speech. While it might be easy to forget, ChatGPT is merely following rules, albeit at a breathtakingly complex level. Now there are philosophers of language who believe intelligence is just skillful language use and that language use is just rule-following, but that is not the consensus of the members of the WGI. On the contrary, language involves expressive of acts of understanding which are not primarily linguistic but rather intentional, what we could call the main Aristotelian line, which has its contemporary representatives in the philosophy of mind in the work of people like John Searle and Thomas Nagel. The main concern articulated by the WGI was never the headline grabbing question, "are we about to be replaced in evolution by our machines?" but rather the far more pedestrian and genuinely disturbing theme that we have already surrendered much of our work, our play, our culture, and indeed our governance to a very limited rule-following apparatus.

The editors wish to thank the editorial staff at *Analecta Hermeneutica* for the opportunity to publish these papers. We would also like to thank Memorial University and the Daimler und Benz Stiftung for funding the project.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

What Is AI, and If So, How Many? Four Puzzles about Artificial Intelligence

Sebastian Rosengrün
CODE University of Applied Sciences, Berlin
ORCID: 0000-0002-0747-8424

Abstract
This paper demonstrates why the following philosophical questions are misleading: can an Artificial Intelligence (AI) think, feel or act, and does it, therefore, have moral rights and duties? It does so by elucidating the issue with four puzzles. The first puzzle concerns the extension of the concept of AI, which, from the standpoint of semantics, necessarily is either empty or underdetermined. The second puzzle makes a distinction between robots and AI. It points out that it is a grave technical misunderstanding to understand a robot as an entity of its own which can be attributed mental states or the status of a moral object. Based on this, in the context of the third and fourth puzzle, this paper states the paradox of the Computer of Theseus, which compares to a new version of the well-known paradox of the Ship of Theseus and demonstrates that, in the face of the peculiarities of hardware and software, AI, considered metaphysically, is a very strange concept.

Keywords: philosophical paradoxes, artificial intelligence, moral philosophy, consciousness, machine learning

Introduction

A significant part of the philosophical debate on AI is to ask whether an AI can think, feel, or act and, therefore, whether it may have moral rights and duties.[1] However, these questions are misleading. Indeed, they aim at what can be attributed to AIs, whether AIs possess mental states (consciousness, intentions, emotions, etc.) or are bearers of moral rights. However, both the historical debate since the 1950s and the current debate on AI mostly fail to determine exactly to whom or what something is attributed at all when talking about AI.

The metaphysical question of who or what an AI is, which entities can even be called AIs, is considerably more complex than one might assume. By metaphysics or ontology, this paper refers to the philosophical sub-discipline, which asks about the existence, being, essence, and structure of things. In analytic philosophy, in particular, metaphysics is closely related to semantics, the linguistic sub-discipline, which asks about the meaning and reference of linguistic expressions. Semantic questions are also the starting point of the following reflections on the metaphysics of AI.

The sentence (1) "This AI has mental states" is identical in form to the sentence (2) "The present king of France has a bald head." Both express an attribute about a certain individual, namely having mental states and being bald, respectively.

To determine the truth value of (2), it is not irrelevant to define what it means to be bald, to consider where baldness comes from, and to discuss moral rights and duties bald people have. In this example, however, assigning a truth value fails not because of an underdetermined definition of the attribute, but because of the indeterminacy of the individual about whom the attribute is expressed. Although the nominal phrase "the present king of France," semantically, refers to the individual who is presently king of France, it is an empty reference because France presently is a republic. That is, the individual who is said to be bald does not exist.

Applied to AI: It is philosophically puzzling to whom or what mental states are attributed in a sentence like (1). On the one hand, this is because—unlike in the case of the King of France—there are different meanings of the term "AI," and on the other hand—just like in the case of the King of France—it is unclear whether a nominal phrase like "this AI" refers to anything at all, and if so, to what exactly.

In this context, this paper discusses four puzzles of a philosophy of AI, some of which build upon each other, and which illustrate the problematic nature of the concept of AI from a semantic and metaphysical perspective.

---

[1] Sebastian Rosengrün, *Künstliche Intelligenz zur Einführung*, Zur Einführung (Hamburg: Junius, 2021).

Every Computer Is AI (Or None)

AI research is divided into two divergent branches:[2] on the one hand, AI is an interdisciplinary research field in which human or natural intelligence is modeled, simulated, and replicated, mostly with the goal of better understanding human or natural intelligence and other cognitive abilities. This field is commonly referred to as "cognitive simulation."[3] On the other hand, AI is a set of specific techniques within software engineering (and thus, AI is a sub-field of computer science). Those techniques are used in the field of cognitive simulation, too, although cognitive simulation goes far beyond computer-based methods and includes, among other things, attempts to replicate intelligence using biochemical methods (this area is widely known as 'wet AI').[4]

While advances in the field of cognitive simulation have yielded insights into intelligence, cognition, and consciousness, it is merely speculative at this stage whether artificial intelligences can be created that may have consciousness and other mental states. The main reason for this is that simulating intelligence is not the same as intelligence—much like a flight in a flight simulator is not a real flight. Moreover, it is doubtful what exactly distinguishes an artificial intelligence (if it is more than a simulation) from a natural intelligence, or whether the distinction between naturalness and artificiality can be maintained at all. If AI is understood in terms of cognitive simulation, there are currently no entities that can be called AI.

In the following, I focus on AI as a subfield of computer science, as a collective term for those techniques that currently play an important role, for example, in the engineering of chatbots, robots, autonomous driving systems, military drones, algorithm-based decision systems, and many other applications. AI encompasses the following subfields of software engineering: Machine learning based on neural networks; Computational linguistics or natural language processing; Machine vision; Reason-based reasoning; Planning and optimization. Combinations of those fields are not only possible but also common.[5]

Furthermore, it is discussed whether simple rule-based programs also count as AI. A relevant example would be a sequence of if-then statements, which—like any computer program—is realized as an electronic circuit system. However, all other

---

[2] Keith Frankish and William Ramsey, *The Cambridge Handbook of Artificial Intelligence*, 3rd ed. (Cambridge: Cambridge University Press, 2018); Klaus Mainzer, *Künstliche Intelligenz. Wann Übernehmen Die Maschinen?*, 2nd ed. (Berlin: Springer, 2019); Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (Cambridge: Cambridge University Press, 2009); Stuart Russell and Peter Norvig, *Artificial Intelligence. A Modern Approach*, 3rd ed. (Harlow: Pearson, 2016); Joseph Weizenbaum, *Computer Power and Human Reason* (New York and San Francisco: Freeman, 1976); Rosengrün, *Künstliche Intelligenz zur Einführung*.

[3] Daniel Dennett, "The Singularity—an Urban Legend?," 2015, https://www.edge.org/response-detail/26035.

[4] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: New York: Oxford University Press, 2009), 55-56.

[5] Rosengrün, *Künstliche Intelligenz zur Einführung*, 13-33.

techniques mentioned are by their nature nothing else than highly complex rule-based systems; they can be completely reduced to them. This leads into the following paradox:

1. Rule-based systems are either AI or they are not.
2. All techniques that are commonly considered AI are completely reducible to rule-based systems.
3. Every computer program is a rule-based system.
4. If every rule-based system is an AI, then every computer program is an AI.
5. If rule-based systems are not AI, then no computer program is an AI.
   Therefore,
6. Either every computer program is an AI, or no computer program is an AI.

From the point of view of computer science, this paradox is not problematic. There, AI is primarily a loose collective term for software engineering techniques. For the successful execution of a program, it is irrelevant whether, for example, machine learning based on neural networks is metaphysically different from a simple "Hello World" command or whether it differs from it only because of a greater complexity of the source code.

This paradox becomes relevant only when entities are referred to as AI and/or certain attributes are ascribed to entities because they are "artificially intelligent" or an application of AI technology, suggesting both philosophical and social consequences. Ascribing mental states to a particular computer (or robot, software, etc.) because there is AI involved is therefore either an empty or misleading statement. According to the paradox explained above, this computer would either not exist at all or every other computer (for example, also the one I am writing this paper on, but also my smartphone and a Commodore 64 gathering dust in the attic) would possess mental states. Therefore, AI cannot be the reason that a computer possesses mental states.

The thesis that every computer possesses mental states may sound absurd at first glance. However, this does not mean that it is irrelevant. Hilary Putnam[6] has coined the position of functionalism or computer functionalism for this in the philosophy of mind. He argues that any electronic device on which a Turing-complete system can be realized (simplistically, any universally programmable computing machine) operates on the same principle as the human mind. However, the actual criterion for attributing mental states is then not AI, but Turing-completeness. AI would be only an unfortunate term for programming computers of any kind. From a philosophical perspective, the AI term would then be at least misleading, because its connotations, shaped by science fiction literature, invite to draw hasty false conclusions and to form magical associations.

---

[6] Hilary Putnam, "Minds and Machines," in *Dimensions of Minds*, ed. Sidney Hook (New York: New York University Press, 1960), 138–64. It is, however, well-known that Putnam changed his views over time, see Rosengrün, *Künstliche Intelligenz zur Einführung,* 35-64.

The first answer to the paradox, according to which no computer is an AI, on the other hand, makes any statement *de re* about an AI a statement with an empty name and leads into the classic no-reference problem of philosophy of language.[7] To claim that a particular AI possesses mental states is then comparable to claiming that the current king of France is bald, which, depending on premises of philosophy of language, is either a false or a meaningless statement as long as France does not return to monarchy in a possible distant future. Alternatively, AIs can be understood as fictional entities (comparable to unicorns, for instance), which even seems obvious, especially given the popularity of the topos in science fiction literature. However, this leads to the fact that a statement about AIs says nothing about real entities. A statement about AIs would then be comparable to the statement "unicorns have pink manes," which beyond a fantasy story hardly presupposes the existence of real unicorns. Saul Kripke, for example, argues that natural kind terms for fictional entities like unicorns fall under the so-called pretense principle, i.e., those terms are used *as if* the entities really exist, while everyone is aware that their existence is just pretended.[8]

Moreover, the statement "AI possesses mental states" is also analyzable *de dicto*, as a statement about what is expressed by the term "AI," comparable to "The present king of France is the one who is monarch of the country designated as 'France' at the time of the utterance." However, even according to this reading, no entity would be said to have mental states, but merely expressed that an AI (whether it exists or not) is something that has mental states.

At least this would apply to all entities of the present and near future. It is true that it cannot be proven in principle that no technique of software engineering is conceivable that is not by its nature completely reducible to rule-based systems and would be classified as AI by the current scientific discourse. To claim otherwise, however, would be pure speculation, which, moreover, is likely to be based less on technical progress than on a quite possible change in the use of language: of course, "artificial intelligence" in the distant future (or in a counterfactual situation, i.e., a possible world) may denote something that is not completely reducible to rule-based systems. However, such a counterfactual use of terms is irrelevant to the validity of the thesis that AI is nothing other than a rule-based system.[9]

This first puzzle has shown that the term "AI" is indetermined, at least when it is used to refer to specific entities: Either every computer (or computer program) is an AI, or there is no AI. Instead of AI, therefore, it should in principle be more precise to speak of certain techniques of software engineering. Beyond this puzzle, my concern in what follows is to point out further metaphysical issues and problems that

---

[7] Bertrand Russell, "On Denoting," *Mind; a Quarterly Review of Psychology and Philosophy* 14, no. 56 (1905): 479–93; Saul A. Kripke, *Reference and Existence. The John Locke Lectures* (Oxford: Oxford University Press, 2013); John Perry, *Reference and Reflexivity* (Stanford: CSLI, 2001); Sebastian Krebs, *Kripkes Metaphysik Möglicher Welten* (Berlin: De Gruyter, 2019).

[8] Kripke, *Reference and Existence*; Krebs, *Kripkes Metaphysik Möglicher Welten*.

[9] Saul A. Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980): 116-125.

result from a misunderstood notion of AI, which is often used in current discourse as if it denotes entities about which certain attributes can be stated. The puzzle presented in the following section is mereological in nature and concerns the frequently advanced proposition that robots possess mental states and/or moral rights because of AI.

A Robot Is Not an AI, an AI Is Not a Robot

According to the prevailing understanding, a robot is an electromechanical machine, consisting of a processor, sensors and effectors. Other possible criteria discussed to define a robot include independent physicality, autonomous or seemingly autonomous behavior, and the ability to influence its respective environment.[10] Of course, industrial robots (e.g., in automobile production) as well as household and everyday robots (e.g., vacuum cleaners and lawn mowers) are also considered robots. These are to be distinguished from android or humanoid ("human-like") robots, which are mostly associated with artificial intelligence in science fiction. Purely mechanical robots or automata, while historically significant, play little role in contemporary robotics.

A characteristic of electromechanical robots is that they are usually controlled by a computer (which is a Turing-complete system). Depending on the paradox described above, any current robot could indeed be classified as artificial intelligence. However, from a technical perspective, the term AI is mostly understood in a narrower sense: For robots specifically, in addition to machine learning, natural language processing and machine vision are the most relevant AI applications. Although they are by their nature nothing more than rule-based programming (see above), these areas certainly describe independent fields of software engineering or computer science.

Accordingly, a robot could be defined to be artificially intelligent if it is controlled by a computer running AI applications, for example, software that analyzes obstacles in a room based on sensors (or cameras) and controls the robot's movements accordingly. While this description of a robot is unproblematic from an engineering perspective, some metaphysical issues arise from the technical setup as soon as artificial intelligence is used as a criterion for attributing mental states or even moral rights and duties to robots. After all, even if computers should possess mental states (and thus possibly the ability to suffer and moral rights) due to certain AI software,[11] this cannot be easily transferred to the robot that is controlled by this computer. Unlike humans, the "mind" or "brain" of a robot exists independently of its body. In this context, it is interesting to point to Hubert Dreyfus' famous criticism of "strong AI"

[10] Janina Loh, *Roboterethik. Eine Einführung* (Berlin: Suhrkamp, 2019); Catrin Misselhorn, *Grundfragen der Maschinenethik*, 4th ed. (Ditzingen: Reclam, 2019).

[11] I doubt this but the following argument is relevant nevertheless since it builds upon a common technical misunderstanding about the setup of robots which leads to further philosophical trouble.

according to which any human-like intelligence needs to be embodied, as intelligence presupposes being-in-the-world (in the Heideggerian sense).[12]

Most entities that are currently considered artificially intelligent robots are only peripheral devices controlled by a computer (a so-called server) in a network or cloud environment. While processors are indeed built into these robots, they serve only as distributors of information in the robot, while the AI code (e.g., in the area of machine vision and language processing, but also machine learning) is practically never executed on the processor built into the robot. The hardware installed in the robot is usually not designed for such resource-intensive computations. Furthermore, a server or AI software running on a network usually controls not just one robot, but any number of robots of the same (or even different) types. However, even this controlling software outsources various complex computations to more specialized AI applications, e.g., for processing speech. The main software just puts the threads together to control a group of robots.

In humans, the brain and body form a physical unit.[13] A human is a self-contained entity to which mental states can be attributed, of course, depending on how one thinks about the mind-body problem. A robot, however, is physically separate from the computer whose software controls it. The "brain" of a robot is—as explained above—usually not located in the robot itself, but in a computer center, which exchanges data with the robot via the Internet (or also with the help of other techniques of digital data transmission), processes input and controls corresponding output commands. At the same time, this computer is not only the "brain" of this robot, but the brain of very many robots.

To assume that a robot possesses mental states, moral rights or similar because it is controlled by an AI is therefore a misunderstanding. For example, neither my hand nor my intestinal wall possesses mental states and moral rights, but I do, in my wholeness of being human. If someone breaks my little finger, it is not my finger that feels pain but me. This person also does not commit an injustice to my finger but to me. Accordingly, a robot cannot be sentient and moral either, but—if at all—the entire system in which the robot is integrated. However, this raises numerous mereological questions as to which components belong to this system at all, and what is the concrete object of which mental states or the like are expressed. Unlike in the case of humans, who are more or less self-contained physical entities, these questions remain puzzling with respect to robots and AI in terms of their metaphysical presuppositions. But when, for example, the misogynistic regime in Saudi Arabia grants civil rights to the

---

[12] Hubert Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*, 7th ed., Perennial Library (New York: Harper & Row, 1986); Hubert Dreyfus, *What Computers Still Cannot Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1999).

[13] This assumption, of course, can be criticized. However, any such criticism would not be an answer to the mereological problem regarding robots, but rather show that the same problem occurs also with regards to humans and their mental states, moral rights etc.

android robot woman Sophia,[14] or when people fall in love with artificially intelligent robots in the future,[15] but also when the European Parliament elaborates a concept on electronic persons, [16] this metaphysical mysteriousness also becomes a practical problem. For individuals can only possess and exercise rights if it is clear who or what exactly these individuals are, and which parts belong to them (and which do not).

However, this mereological problem leads far beyond AI-based robots. I show this in the following two sections, in which I introduce the thought experiment of Theseus' computer, which I use to show that the mereological underdeterminacy of AI poses practical problems in several respects at once.

Theseus' Computer: What Is AI, What Is Periphery?

Building on what has been said about robots, the question of which concrete entities count as AI raises mereological questions not unlike those of precisely determining the essence of a human being. In doing so, my following considerations presuppose a so-called Aristotelian essentialism. By this I mean the basic idea, loosely based on Aristotle's metaphysics, that things possess some attributes essentially, other attributes only accidentally.[17]

While the question of which attributes are essential to a human being and which are merely accidental can often be answered intuitively, intuitions about computers and AI have their limits. My left hand, for example, is a part of my body, it stands in a mereological relation to it, respectively to me. If I would lose my hand due to an accident or similar, I would still be me, my hand is not a necessary part of me. My left hand does not belong to my being or my essence.

But what belongs to the essence of a computer or an AI? In reference to the ancient Theseus paradox, this question can be illustrated by the following thought experiment: Theseus is a teenager who programs artificial intelligences in his spare time. His favorite project is an AI called Minotaur, which is supposed to find exits from winding mazes on its own based on machine learning with neural networks.

Since his computer is getting a bit old, he asks his friend Ariadne to replace some components. Ariadne gradually replaces the graphics card, hard drive, and motherboard of Theseus' computer with more powerful models and copies all the data (including the compiled AI and the uncompiled source code) to Theseus' new hard

---

[14] Cleve Wootson, "Saudi Arabia, which denies women equal rights, makes a robot a citizen," 2017, https://www.washingtonpost.com/news/innovations/wp/2017/10/29/saudi-arabia-which-denies-women-equal-rights-makes-a-robot-a-citizen.

[15] David Levy, *Love and Sex with Robots: The Evolution of Human-Robot Relations* (New York: HarperCollins, 2007).

[16] Loh, *Roboterethik, 84-5.*

[17] Willard Van Orman Quine, "Three Grades of Modal Involvement," in *The Ways of Paradox and Other Essays* (New York: Random House, 1966), 156–74; Kripke, *Naming and Necessity;* for my own take on Aristotelian essentialism, see Krebs, *Kripkes Metaphysik Möglicher Welten*, chapter 2.4.

drive. Since Ariadne still has good use for Theseus' old components, especially the hard drive and motherboard, she installs them in her own computer. Being curious about Theseus' latest progress on his Minotaur project, she starts the AI that is still on Theseus' old hard drive.

The philosophical paradox arising from this thought experiment is: which is the original Minotaur? The one AI that is on Theseus' new (improved by Ariadne's help) computer, or yet the one AI that Ariadne just started on the original components of Theseus' computer?

Unlike the ancient Theseus paradox, this paradox is puzzling on two levels, both software and hardware. Before discussing the genuine mysteriousness of the nature of AI at the software level (see next section), I first show some considerations about the hardware level. These are not necessarily original compared to the ancient paradox of Theseus, but they are highly relevant philosophically when computers and AI, respectively, are ascribed mental states, moral rights, and other such attributes.

In computer technology, components that are located outside the central unit of a computer are called peripherals. These include, for example, the mouse, keyboard, monitor, and also network and graphics cards. It stands to reason to assume that these devices can be replaced without changing the essence of a particular computer—much like it stands to reason that Theseus' ship will still be Theseus' ship even if you replace the sail or steering wheel.

However, if mental states are attributed to an AI, which can be traced back to "sensory perceptions," the input by sensors, already the installation or de-installation of peripheral devices such as microphones, webcams etc. can seriously change the nature of the mental states of an AI. For instance, a webcam with slightly higher resolution would lead to a completely different visual "perception" of the AI. Comparable considerations are usually discussed in relation to humans under the heading of enhancements, the optimization of humans through technology. In a sense, my glasses already have a serious influence on my sense of sight, but hardly anyone would seriously doubt that I am still me after I have replaced my glasses with ones with a higher diopter number. The same applies, for example, to prostheses, hearing aids, etc., and even with futuristic-looking enhancements such as the Eyeborg color sensor by cyborg activist Neil Harbisson, it will be difficult to argue that Harbisson is no longer Harbisson.[18]

Unlike humans, however, even those parts of a computer that do not belong to the periphery but form its central unit can be easily replaced and improved.[19] What exactly counts as the central processing unit of a computer is disputed in computer science: some definitions also include the main memory (RAM), the entire motherboard, and even the hard disk; others only the processor (CPU) or even the

[18] Harbisson, Neil, "I listen to color," 2012, TEDGlobal, http://ted.com/talks/neil_harbisson_i_listen_to_color.

19 I am not speculating about computer-brain interfaces as they are currently discussed mostly among transhumanists.

processor core (the concrete microchip). But it seems questionable whether replacing the processor core (or the entire motherboard) changes the nature of the computer or the AI implemented on it.

While these and similar problems also arise with respect to the ancient paradox of Theseus, the computer version of the paradox opens up yet another level: namely, with respect to the metaphysical status of an AI, it is completely unclear whether "AI" denotes the software or a concrete hardware realization of that software. This supposedly only theoretical question, however, becomes immediately practical exactly when mental states and moral rights are attributed to AI.

Theseus' AI: Universality and Individuality of Computer Programs

Every computer program (software) can be reduced to electronic circuits (hardware). A program is nothing more than a description or prescription of how certain electronic circuits are to behave. The program in turn has a counterpart on the hardware, where it is represented in some form (be it optical, magnetic or electrical). It is at this point, however, that the question of what exactly an AI is becomes philosophically strange. This is aptly summarized, for example, by the media theorist Friedrich Kittler with his famous bon mot "There is no software."[20] If there is no software, however, the question of what exactly an AI is becomes philosophically odd.

To make this oddity conceptual, it is helpful to become aware of the functioning and technical structure of a computer program: Programmers produce the source code of a program, i.e., the collection of those algorithms which determine the so-called output depending on the respective input. This source code, however, is not the actual program, but only an abstraction of the machine language that can be understood by humans. This source code must first be made "readable" for machines. For this there are two usual procedures: Either the entire source code is compiled into machine language by a so-called compiler before it can be executed, or the source code is translated line by line into machine language by a so-called interpreter and executed directly. Which method is used usually depends on the chosen programming language. Currently, the most popular programming language for AI application is Python, which is an interpreter language, but can also be compiled.

Regardless of whether the source code is compiled or interpreted, the question arises whether the mere source code of a program already constitutes AI. After all, Theseus "created" his Minotaur AI by saving the source code of the Minotaur in a text document. However, to classify the source code alone as AI would be absurd, at least if one ascribes certain mental states or moral rights to an AI program. The source code of a program *is* merely an ordinary text document whose content corresponds to the

[20] Friedrich A. Kittler, *The Truth of the Technological World: Essays on the Genealogy of Presence*, trans. Erik Butler (Stanford University Press, 2014), 219.

syntax of a programming language. However, hardly anyone would ascribe mental states or significant moral rights to a text document (which includes, for example, the file in which this paper is stored). One could even take this further and raise the question of whether also handwritten source code could be called an AI (and whether handwritten documents, accordingly, should also be seen as something possessing mental states and moral rights).

There are countless copies of the source code of every AI program, not only because of regular backups, but also because of the technical structure of computer operating systems. These copies match the original exactly, so that in the case of digital copies—unlike analog copies—it is no longer possible to distinguish which text document is now the original.[21] Although so-called generation loss is also possible in current computer technology when copying files, i.e. the loss of individual bits when copying files, this does not provide a criterion for distinguishing between the original and a copy of files, either practically or theoretically. Thus, Theseus has not only one Minotaur on his computer, but countless identical Minotaurs. Likewise, in the thought experiment sketched above, Ariadne has innumerable files with the same source code, i.e. also on her computer there is not just one exact copy of the Minotaur, but innumerable ones.

From a metaphysical perspective, the concept of AI therefore involves a problem of individuation, since it is impossible to determine which of these files contains the actual Minotaur, and if so, from how many copies on a new Minotaur is created (assuming Ariadne changes only one line of the source code, is this already a new individual?) and whether then perhaps even Theseus' and Ariadne's computers each house innumerable artificially-intelligent entities, to which all mental states and moral rights are to be attributed, if one assumes that AIs possess these attributes.

This individuation problem exists, however, even if one does not count the source code as AI proper, but only its translation into machine language or the execution of this machine language by the computer. Indeed, if one assumes that only the execution of an AI program constitutes an AI capable of mental states and, moreover, entitled to moral rights, little is gained for the solution of this problem. In fact, this would mean that every time a program is restarted, a new conscious individual would be created, and this individual would be killed with the termination of a program.

One possible objection would be to claim that quitting a program merely means putting a conscious individual into a kind of artificial coma, which would be awakened by the restart. But if AI has consciousness and moral rights, it would then be ethically dubious to restart a program (or even the computer) without first asking permission. At the latest when a program is recompiled (especially if small changes

---

[21] Armin Nassehi, *Muster: Theorie der digitalen Gesellschaft* (Bonn: Bundeszentrale für Politische Bildung, 2020); Michael Betancourt, *The Critique of Digital Capitalism: An Analysis of the Political Economy of Digital Culture and Technology* (New York: Punctum Books, 2015).

have been made to the source code beforehand), this objection falls short. With the recompilation the old program is completely overwritten, at the latest here a new conscious individual would have been created, while the previous program would be "killed." Then, however, each overwriting of existing programming code and the recompilation necessary thereupon would mean to murder a conscious individual. With interpreted programming languages, each restart of the program would be connected automatically with a re-creation of a conscious individual, since the source code is always translated thereby from scratch again into machine language. Software engineering—of whatever kind—would then to be rejected for moral reasons.

Although this sounds absurd, this is—following my argumentation—a direct consequence of the assertion that an AI possesses mental states. In fact, a similar argument can be found in Thomas Metzinger's work, according to which the creation of artificial consciousness is ethically questionable. Metzinger assumes that the "first machines satisfying a minimally sufficient set of conditions for conscious experience and self-hood would find themselves in a situation similar to that of the geneti-cally engineered retarded human infants. Like them, these machines would have all kinds of functional and representational deficits—various disabilities resulting from errors in human engineering."[22] Creating artificial consciousness, according to Metzinger's argument, produces unnecessary suffering. This argument is, of course, not about AI in the technical sense presented in this paper, but explicitly about artificial consciousness. Metzinger does not claim that every AI has consciousness. He merely assumes that, according to his own naturalistic theory of consciousness, the creation of artificial consciousness is possible, although this artificial consciousness need not necessarily be based on AI in the computer science sense.

Nevertheless, Metzinger's argument leads into an objection, interesting in the context of Theseus' computer, to the thesis that an AI (or a machine on which AI is realized) possesses consciousness (and/or deserves moral rights). In so far as this is true, any change in the source code of a program, including the necessary recompilation/interpretation, would be tantamount to erasing the existence of a conscious individual due to design errors and replacing it by the creation of a new conscious individual. That software engineering is a highly morally questionable activity would thus be a direct consequence of the thesis that AI possesses mental states. This, of course, does not refute computer functionalism (and numerous similar positions). To consequently reject any form of software engineering on the basis of ethical considerations, however, is in stark contrast to the enthusiasm for technology and innovation that some proponents of the thesis that AI can possess mental states currently embody in public.

---

[22] Thomas Metzinger, *The Ego-Tunnel: The Science of the Mind and the Myth of the Self* (New York: Basic Books, 2009), 195. See also Thomas Metzinger, "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology," *Journal of Artificial Intelligence and Consciousness* 08, no. 01 (2021): 43–66.

The puzzles formulated in this paper have thus shown, above all, into which strange absurdities the thesis that AI possesses mental states necessarily leads, if one considers the fundamental metaphysical question of who or what the individuals are at all, about whom corresponding attributes are sometimes all too carelessly stated in the current discourse.

Speculations

This semantic and metaphysical puzzles pointed out in this paper have shown that the question of what AI is, is problematic. However, this problem must be answered especially by those who ascribe various attributes to AI (or software, or computers in general) in the current discourse. Only by expressing certain attributes, there is an argumentative obligation to define whom or what the attributes are expressed about.

It is important to note that the puzzles also arise when—as is often the case in the current discourse—we are not talking about AI, but about so-called Artificial General Intelligence (AGI). This refers to those AIs that are not only capable of solving a specific task, but can generally solve all (or at least most) tasks that previously could only be solved by human intelligence. With respect to an AGI, the questions and problems posed in this paper are even stranger, since an AGI does not currently exist. Even futurologists speculating at length about the consciousness of an AGI, such as Max Tegmark, admit that "there's absolutely no guarantee that we'll manage to build human-level AGI in our lifetime—or ever."[23]

Furthermore, since it is at least questionable whether the construction of an AGI is even technically possible, it is also entirely speculative as to how such an AGI could possibly be constructed. However, this makes any statement attributing mental states to an AGI a statement about the extension of an empty concept (comparable to a statement about unicorns, see above). The semantic and metaphysical puzzles pointed out in this paper, therefore, become all the more absurd, the less the form of AI of which certain attributes are said to be AI at the present state of the technology.

[23] Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (London: Allen Lane, 2017), 132.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Everything but the Truth: On the Relevance of Algorithms

Sebastian Rosengrün
CODE University of Applied Sciences, Berlin
ORCID: 0000-0002-0747-8424

Abstract
This paper defends two theses: (1) Humans aim for relevance, not for truth. (2) Algorithms provide relevance, not truth. While the first thesis builds upon research in cognitive linguistics and neuroscience, the second thesis will be based upon an analysis of algorithm functionality in digital platforms like Google, Amazon, and Social Media, but also in most recent developments in the field of Generative AI. Based on these two theses, this paper outlines asymmetric power structures in digital capitalism and how commercial interests undermine the democratic discourse by spreading fake news and conspiracy theories.

Keywords: artificial intelligence, relevance, algorithms, digital capitalism, fake news

Introduction

Algorithm-based recommendation systems have gained increasing influence over how humans perceive and interact with the world. These systems play a pivotal role in shaping various aspects of our lives, both online and offline. For instance, Google's search algorithm determines the information users find and the sources they consider during their research endeavors. Social media algorithms curate the news users consume, dictate the topics they encounter, and influence how these topics are presented. Similarly, Amazon's algorithm influences the products buyers discover and the prices they pay, thereby impacting market dynamics and the chances of success for different companies. Importantly, the reach of algorithm-based recommendation systems extends beyond the confines of the online world, permeating our daily lives in ways that often go unnoticed. From the restaurants we choose to dine in and the vacation destinations we select to the potential sex partners we encounter through dating apps, algorithms subtly shape our normal lives. In essence, algorithm-based recommendation systems can be seen as a form of regulation on human behavior, akin to what legal scholar Lawrence Lessig termed "Code is Law."[1]

The objective of this paper is twofold: Firstly, it aims to provide an answer to the question of why algorithm-based recommendation systems have achieved such remarkable success. Secondly, it seeks to contextualize some of the most problematic effects arising from these systems. Of particular importance is the rising concern of fake news and conspiracy theories, where the asymmetric power structures inherent in algorithm-based recommendation systems play a crucial role. To comprehend this phenomenon, this paper will explore the concept of relevance—a fundamental idea within cognitive linguistics and philosophy of language that must be distinguished from the concept of truth. In this context, "relevance" refers to the usefulness of an assertion (a recommendation, or an explanation) to a speaker, which may not always align with factual accuracy.

In the first section, this paper will delve into this concept of relevance and its significance in human cognition, highlighting that humans prioritize relevance over truth. This understanding will lay the groundwork for the next section, where it will be demonstrated how algorithm-based recommendation systems excel at providing relevance to users compared to human-based recommendations. Leveraging their algorithms, these systems offer a significant advantage in satisfying users' cognitive needs efficiently. Thereafter, this paper will shed light on some of the problematic aspects of algorithm-based relevance in various facets of our social life, particularly considering the asymmetric power structures prevalent in digital capitalism and how algorithm-based relevance contributes to the rise of fake news and conspiracy theories.

---

[1] Lawrence Lessig, *Code: Version 2.0* (New York: Basic Books, 2006). For a detailed analysis of this broader picture of 'regulation by AI', cf. Sebastian Rosengrün, "Why AI Is a Threat to the Rule of Law," *Digital Society* 1, no. 2 (2022): 10, https://doi.org/10.1007/s44206-022-00011-5.

Finally, this paper will briefly talk about recent developments in generative AI and about their potential implications for algorithm-based recommendation systems.

The *Relevance* of Relevance

Relevance theory is a major theme in cognitive linguistics. Research in this field indicates "the search for relevance is a basic feature of human cognition" and:

> [T]he spontaneous working of our perceptual mechanisms tends to pick out the most relevant potential inputs, the spontaneous working of our memory retrieval mechanisms tends to activate the most relevant potential contextual assumptions, and the spontaneous working of our inferential mechanisms tends to yield the most relevant conclusions.[2]

Similarly, studies from neuroscience further suggest innate processes of selective attention "that allow an individual to select and focus on particular input for further processing while simultaneously suppressing irrelevant or distracting information."[3] There is also a strong tradition of philosophical pragmatism building on the assumption that, at least in everyday situations, the factuality ('truth') of an assertion is significantly less important than whether an assertion is useful to fulfill a particular goal within a specific context, hinting already at a minimal definition of the term 'relevance': *Def.* An assertion is relevant if and only if it is useful in a concrete situation.[4] Furniture assembly instructions (which are assertions, too), for example, are relevant for me if and only if they lead to a successful practice, i.e., they help me connect the various parts so that the bookcase will hold together. One crucial feature of relevance, therefore, is that it often only manifests itself in retrospect. Whether instructions are relevant, I will most likely only find out after having followed them through without the bookcase falling apart.

While this paper does not presuppose a specific definition of truth, there is one significant difference between the concept of truth (both according to correspondence and coherence theories) and the concept of relevance: Truth is

---

[2] Deirdre Wilson, "Relevance Theory," in *The Pragmatics Encyclopedia*, ed. Louise Cummings (New York: Routledge, 2010), 395; Dan Sperber and Deirdre Wilson, "Relevance Theory," in *The Handbook of Pragmatics*, ed. Laurence Horn and Gregory Ward (Oxford: Blackwell, 2006), 608.

[3] Courtney Stevens and Daphne Bavelier, "The Role of Selective Attention on Academic Foundations: A Cognitive Neuroscience Perspective," *Developmental Cognitive Neuroscience* 2 (2012): 30, https://doi.org/10.1016/j.dcn.2011.11.001; cf. Wolf Singer, *Ein neues Menschenbild? Gespräche über Hirnforschung*, Suhrkamp-Taschenbuch Wissenschaft 1596 (Frankfurt am Main: Suhrkamp, 2003).

[4] For this definition and the following explanations, cf. Thomas Becker, "Is Truth Relevant? On the Relevance of Relevance," *Etica & Politica / Ethics & Politics* XVI, no. 2 (2014): 595–618; Sebastian Krebs, "Does Truth Really Matter? On the Irrelevance of Truth," in *Practical Rationality in Political Contexts. Facing Diversity in Contemporary Multicultural Europe*, ed. Gabriele De Anna and Riccardo Martinelli (Trieste: EUT Edizioni Università di Trieste, 2016), 31–58.

objective, relevance is not. If the assertions "The 12th decimal place of π is 9" and "The last word of this paper is 'relevant'" are true, then they are true for everyone at all times. They would even be true if no one had ever calculated π or would read this paper to the last paragraph. However, the 12th decimal place of π is relevant only for a small group of people, and whatever the last word of this paper shall be, is, at this point, still irrelevant. Those examples also indicate that not all true assertions are relevant. With Bernard Bolzano, drawing from his *Wissenschaftslehre* from 1837, it can be stated that science is not (only) about finding truth but (also) about selecting from an infinite number of truths those that have practical use.[5] This idea is exemplified by Thomas Becker's thought experiment of the Library of Baghdad, an adaptation from Argentinian writer Jorge Luis Borges' Library of Babel:

> [In this fictive library,]: books contain only true sentences (not a single false one) in impeccable English, without a single misprint. It contains, just like the Library of Babel, an infinite set of true sentences derived logically or by other recursive definitions from a basis of true and known sentences compiled by a large committee of scholars. All the sentences differ from each other, not a single sentence is recorded twice, and all sentences are of finite length. Nevertheless, it is as useless as the Library of Babel, because you have virtually no chance to find a single interesting sentence among the infinite number of true and irrelevant ones.[6]

Becker claims that this library is as useless as Borges' complete library, as this library is infinitely extensive due to multiple reasons like the recursiveness of natural languages and the formal logic of adding a true disjunct to an otherwise false sentence yielding a true sentence.[7] He concludes that "the point of assertion is to pick out the most relevant proposition of an infinite number of true, known and justifiable ones."[8]

The most influential definition of relevance can be attributed to Dan Sperber and Deirdre Wilson, who established 'relevance theory' as an independent field of research in cognitive linguistics. According to them, relevance is a function determined by two factors, (a) cognitive effects and (b) processing effort. For individuals, what is considered relevant is typically the assertion that yields the highest cognitive effects (cognitive reward) with the least processing effort (cognitive costs).[9] This function is

---

[5] Bernard Bolzano, *Grundlegung Der Logik. Ausgewählte Paragraphen Aus Der Wissenschaftslehre*, ed. Friedrich Kambartel, 2., durchges. Aufl, vol. 259, Philosophische Bibliothek (Hamburg: Meiner, 1978), 3.

[6] Becker, "Is Truth Relevant?," 602.

[7] Cf. Jorge Luis Borges, *Collected Fictions.*, trans. Andrew Hurley (London: Penguin, 1999); Jorge Luis Borges, *The Total Library: Non-Fiction 1922-1986*, trans. Esther Allen, Suzanne Jill Levine, and Eliot Weinberger (London: Penguin, 2001).

[8] Becker, 603.

[9] Cf. Dan Sperber and Deirdre Wilson, *Relevance: Communication and Cognition*, 2nd ed. (New York: Blackwell, 1995); Sperber and Wilson, "Relevance Theory"; Yan Huang, *Pragmatics* (Oxford: Oxford University Press, 2007).

crucial for the following analysis of relevance in algorithms, as it adds another essential aspect to the understanding of usefulness from above: An assertion is more useful to a hearer when the conveyed information requires less effort to process. For example, while a 200-page documentation on using various screws and tools may effectively prepare me for assembling a bookcase, too, a concise and visually appealing four-page comic-like guide offers me an equally effective cognitive while incurring significantly fewer cognitive costs.

Relevance in Algorithms

Tech companies have embodied this cognitive principle of relevance, which is, as this paper claims, a crucial aspect of the economic success of algorithm-driven business models. In 1995, Larry Page and Sergey Brin developed the Google precursor Backrub. It was based on the idea that the value of a website should be determined by the number of backlinks, i.e., the more often a website was linked by other websites, the higher it was ranked at Google. In other words: They understood that the quality of content is less important for the user than how popular it is among other people. This approach was revolutionary because it prioritized the popularity of a website over its content quality, which was a departure from previous search engines. By relying on backlinks to rank websites, Google could provide more relevant results to users, as websites frequently linked to by other reputable sources were deemed more valuable. This approach has since been widely adopted by other tech companies, who use similar algorithms to provide personalized recommendations to users based on their past behaviors and preferences.

Interestingly, Page and Brin took their idea from academia, where a researcher's reputation is mainly determined not by the quality of their research but rather by how often their papers are quoted by others (or rather, that the number of citations is the most important criteria for the quality of research). This paper, for example, becomes a relevant contribution to the philosophy of digital technologies if and only if it is quoted in many academic publications.[10] The quality of its arguments matters only insofar as there is a common agreement in academia that one ought only quote papers with argumentative quality (and reviewers typically ensure a certain standard within the publication process).

While there are neither peer-reviewers nor silent agreements in website publishing, the idea of Backrub was still a success—compared to earlier search engines that solely analyzed a website's content. To have a website placed among the top search results for, e.g., "how to assemble a bookshelf," a publisher only had to ensure to use words like "bookshelf" and "assembly" more often than their competitors within their

[10] For a performative criticism of this common academic practice, see, Sebastian Rosengrün, "Everything but the Truth: On the Relevance of Algorithms," *Analecta Hermeneutica* 15 (2023).

HTML documents. The idea of Backrub required a website to be linked back by others to achieve a high PageRank—a concept introduced by Page and Brin. From a user perspective, search results suddenly became far more relevant, and Google (as Backrub was re-named before becoming successful) soon became the gold standard of internet search engines (which it still is, even though Generative AI might pose a significant threat to its market position, see section 5).

While relevance should be defined as leading to successful practice (see the previous of this paper), the Google algorithm defines what leads to a successful practice by what other people think is leading to successful practice. Algorithms have been optimized based on such an understanding of relevance, and search engines are just one of many similar examples. Spotify, Netflix, and YouTube make algorithm-based recommendations of what to watch/listen next dependent on what other people with a similar background (age, gender, hobbies, favorite band, etc.) like to watch/listen. Amazon optimizes their product placement algorithms based on what other similar shoppers tend to buy, and the whole business model of social media like Facebook, Instagram and TikTok relies on the quality of their algorithms to select relevant content for their users. The same principle applies, more and more, to news websites, political campaigning and commercial advertisement in general, but also to dating apps, restaurant recommendations and travel planning. Algorithms have been optimized to display what 'similar' users seem to enjoy, as this seems to be what brings any individual the highest cognitive reward for as little processing costs as possible.

In order to train those algorithms, tech companies like Google heavily rely on the data they collect from their users, leading to a phenomenon that Shoshana Zuboff, in her well-received study, describes as "surveillance capitalism," a "new instrumentarian power that asserts dominance over society and presents startling challenges to market democracy."[11] According to Zuboff, companies increasingly use algorithms to control human behavior by predicting precisely how humans behave and how manipulating the input variables for human behavior will affect behavioral outcomes. Zuboff's analysis offers profound insights into the historical development and business models of big tech corporations, but also a blunt criticism of surveillance capitalism which she describes as "parasitic and self-referential" economic order that "feeds on every aspect of every human's experience."[12] With algorithm-based behavior prediction and control, tech corporations endanger every human's right to make their own life choices—which is a core value of any democratic society.[13]

However, it is highly questionable whether algorithms trained with machine learning can be accurate enough to achieve such an ultimate behavior prediction. The effort (the cognitive costs) to perform a Google search (or following Amazon's or

---

[11] See, Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, First edition (New York: PublicAffairs, 2019).
[12] Zuboff, *The Age of Surveillance Capitalism*, 9.
[13] For a more detailed analysis, See, Rosengrün, Sebastian, "Why AI is a threat to the rule of law," *Digital Society* 1, no. 2 (2022).

Netflix's recommendations), however, is very little compared to asking friends, colleagues, and experts for help and recommendations. Also, their output is optimized in a way that makes it very easy to process: a brief list of 10 suggestions of what product to buy/read/watch/invest in is far easier to deal with than extracting valuable information out of the monumental speeches by our enthusiastic librarians, geeky little nieces, or dodgy insurance brokers. In fact, all tech companies spend tremendous research efforts in order to 'optimize' their user experience by introducing new design patterns, color schemes, or features (like Amazon's infamous One-Click-Buying, which they even patented, and which reduces the cognitive costs of online shopping to the bare minimum).[14]

Therefore, their suggestions automatically have a tremendous advantage in terms of relevance over 'traditional' search methods in terms of how relevant they are for an individual. When one assumes that whatever an algorithm suggests is accurate (and people do not question those suggestions because of the relatively little processing effort), it is easy to conclude that algorithmic suggestions are relevant. A user being amazed by algorithmic predictions (like Google's search results or Netflix's movie predictions) is comparable to a tiny baby being fascinated by observing her reflection in a mirror without realizing it is herself she is watching. This observation is also (and even more) accurate for generative AI (like ChatGPT, Google Bard, Stable Diffusion, DALL-E etc.) that often surprises their users with highly relevant outputs to their initial prompts. However, the reason for the relevance of their outputs is, at least partially, to be explained by the little processing costs they mean for their users, while the cognitive effects are simply a result of a rearrangement of an extremely large set of language tokens (i.e., language used by other human beings) with the help of stochastic means.

Especially in cases in which factuality and objectivity do not (seem to) exist, it seems that the relevance of algorithmic predictions must be explained mainly by their low cognitive costs, not by their invaluable cognitive effects: What book to buy next, what movie to watch next, and—what party/candidate to vote next for in the upcoming elections. Those are rather questions of relevance rather than questions of truth. Those question even presuppose that buying yet another book, watching yet another movie and participating in an election are the only feasible options (see section 4). They have a tremendous social impact, however, when their answers are calculated by algorithms within the asymmetric power structures of digital capitalism.

[14] Within academia, judging a researcher's work solely based on their h-index also requires much less processing effort than reading through their publications.

Relevance in Asymmetric Power Structures

What has been said in the two sections above can be summed up as follows: Algorithms control what appears to be most relevant for individual users. Keeping in mind that relevance is determined by the highest cognitive effects for the lowest cognitive costs, however, the relevance of algorithmic suggestions is not necessarily to be explained by high cognitive effects for the user but rather because of the small cognitive costs. Given that those algorithms are controlled by a handful of tech corporations, it is crucial to shed light on the concept of relevance within those asymmetric power structures created by what scholars call 'digital capitalism.'[15]

Given the monopolistic tendencies within digital capitalism, tech corporations control what is relevant. In the search engine market, Google has a global market share of 93.11 percent (May, 2023), with Microsoft's Bing being the only noteworthy alternative (at least within the so-called Western context, leaving Russian and Chinese search engines aside).[16] Similarly, companies and services like Amazon, Netflix, Twitter, Instagram, and YouTube have become dominant players with their algorithm-based recommendations, influencing, for example, what people buy, read, watch, or listen to next. It is important to note that those companies hold this power—whether they want it to or not. Not making an active decision on what to recommend someone, is also a decision—especially if the person looking for a recommendation looks at you as the sole source of truth.[17] While most tech companies actively shy away from the responsibility that comes along with this power, they actively make use of this power: By controlling those algorithms, they influence the way people perceive and interact with the world.[18] It is important to note here that recommendation algorithms do not only offer relevant answers, but the way those algorithms have been designed, also presupposes that what to buy, read, watch, or listen to next is even a relevant question for their users. *That* everyone wants to buy, read, watch or listen to something else, is a decision already made for the users, and it seems that, to many people, this is a more relevant option (i.e., a question that takes less processing costs) than asking the (admittedly, more complex) question of what else one could do with their time. While

---

[15] Cf. Michael Betancourt, *The Critique of Digital Capitalism: An Analysis of the Political Economy of Digital Culture and Technology* (Brooklyn, NY: punctum books, 2015); Dan Schiller, *Digital Capitalism: Networking the Global Market System* (Cambridge, Mass: MIT Press, 1999); Dan Schiller, *Digital Depression: Information Technology and Economic Crisis*, The Geopolitics of Information (Urbana, Chicago: University of Illinois Press, 2014); Philipp Staab, *Digitaler Kapitalismus. Markt Und Herrschaft in Der Ökonomie Der Unknappheit* (Suhrkamp, 2019); Amy Webb, *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity* (New York: PublicAffairs, 2020); Rosengrün, "Why AI Is a Threat to the Rule of Law."

[16] StatCounter Global Stats, accessed June 21, 2023, https://gs.statcounter.com/search-engine-market-share. Privacy-aware people often prefer meta search engines such as DuckDuckGo and Ecosia, which, however, rely in their search results mainly on the algorithms of Google or Bing.

[17] Cf. Sebastian Rosengrün, *Künstliche Intelligenz zur Einführung*, Zur Einführung (Hamburg: Junius, 2021).

[18] Cf. Adrian Daub, *What Tech Calls Thinking* (New York: Farrar, Straus and Giroux, 2020).

the relevance of algorithmic recommendations in our daily life causes enormous social problem, this paper does not suggest that the companies behind that follow a broader political agenda. There mere goal is capitalist profit,[19] and that's exactly why it is problematic that their power over what is presented as relevant for people takes place beyond democratic discourse.

This can be, for example, illustrated by focusing on the issue of so-called filter bubbles and how they lead to the spread of fake news and conspiracy theories.[20] As this paper suggests, those result from the monopolistic control of social media companies over what is relevant for their users. By filter bubbles, this paper refers to the phenomenon where individuals are exposed only to information that confirms their existing beliefs, resulting in a narrowing of perspectives and an echo chamber effect. Tech corporations play a central role in creating these filter bubbles through their algorithms, which are designed to show users content that is most likely to keep them engaged on their platforms. Algorithmic recommendations limit exposure to diverse perspectives by prioritizing content that confirms users' existing beliefs and biases. In addition, the spread of fake news is facilitated by the ability of tech corporations to amplify and distribute information at unprecedented speeds. Without robust fact-checking mechanisms, misinformation can quickly spread through social media networks, further contributing to the creation of filter bubbles and the erosion of trust in traditional sources of information.

In this landscape, traditional media organizations have been confronted with the following dilemma of staying relevant: Either they adjust themselves by lowering the cognitive costs for their audience or they try to focus on the quality of their information (and increase the cognitive rewards), for which, however it is increasingly difficult to find a viable business model given how easy it is to find equally relevant information that might not have the same cognitive reward but comes with much lower cognitive costs. Quality media do not only have higher cognitive costs for their users, but also significantly higher convenience and financial costs: While social media (or the internet, in general) is full of fake news and conspiracy theories that are freely accessible, quality media are 'hiding' their content more and more behind paywalls, opaque subscription models and premium accesses.[21] While this approach seems to be economically necessary to generate revenue and pay for journalistic endeavors, it also restricts access to valuable information and excludes those who cannot afford or are unwilling to pay for digital content (or those who are unwilling to give away their personal data and manage numerous online accounts). Especially people who are

---

[19] Rosengrün, "Why AI Is a Threat to the Rule of Law," 10.

[20] See, Seth Flaxman, Sharad Goel, and Justin M. Rao, "Filter Bubbles, Echo Chambers, and Online News Consumption," *Public Opinion Quarterly* 80, no. S1 (2016): 298–320, https://doi.org/10.1093/poq/nfw006; Eli Pariser, *The Filter Bubble: What the Internet Is Hiding from You* (London: Penguin Books, 2012); Michael Butter, *The Nature of Conspiracy Theories* (Cambridge: Polity Press, 2020).

[21] Unfortunately, the same must be said for many academic publications.

already affected by their individual filter bubbles and echo chambers will be excluded even more from reliable information, given that what seems relevant to them is already provided by social media who mostly care about low processing costs for their users.

Abuse of asymmetric power structures within digital capitalism goes beyond filter bubbles and echo chambers. Users have been abused by tech companies which confront them with seemingly relevant algorithmic recommendations for many years. An extreme example is the infamous 'Emotional Contagion Experiment,' conducted by Facebook in 2014.[22] In this study, Facebook manipulated the News Feed algorithm for a subset of users by selectively filtering out positive or negative posts for a given period. This alteration in the content was done to observe if it would lead to changes in users' own emotional expressions in their subsequent posts. While this study has been controversially discussed (not only for the obvious ethical, but also for methodological reasons), its results suggest that emotional contagion could occur through social media platforms, as user emotions seemed to be influenced by the emotional content they were exposed to. Of course, the main intention behind such commercial experiments is to turn their results into a business case, knowing that a user's emotional status not only heavily affects how they interact with other users on a platform, but also their media consumption and (online) shopping behavior. Another extreme case highlighting the implications of algorithm-based recommendation systems is the Cambridge Analytica scandal.[23] This notorious incident revealed the potential misuse of personal data by a political consulting firm, which utilized algorithms to target and manipulate users with tailored political content to the effect of influencing the outcome of democratic elections.

Summary and Outlook

Algorithm-based recommendation systems have become integral to our modern lives, shaping our behavior and influencing the information we consume. This paper has explored the concept of relevance from both a philosophical and linguistic perspective. Relevance, defined as practical usefulness and determined by cognitive costs and rewards, was presented as a key concept to explain the success of those algorithmic systems. By prioritizing relevance over truth, algorithm-based recommendations cater to our cognitive needs effectively. Given that many users already assume that algorithmic recommendations are more relevant than traditional systems, the

[22] Cf. Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks," *Proceedings of the National Academy of Sciences* 111, no. 24 (2014): 8788–90, https://doi.org/10.1073/pnas.1320040111; Jukka Jouhki et al., "Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics," *Media and Communication* 4, no. 4 (2016): 75–85, https://doi.org/10.17645/mac.v4i4.579.

[23] Cf. Christopher Wylie, *Mindf*ck: Cambridge Analytica and the Plot to Break America* (New York: Random House, 2019).

companies behind those algorithms gain power over shaping the way people perceive and interact with the world. This comes with inherent risks and challenges. The rise of fake news and conspiracy theories as well as other threats to a free and open society underscores the need to critically examine the asymmetric power structures embedded in these systems.

A gamechanger with regards to algorithm-based recommendations systems is the rather recent development in generative AI (including Chat-GPT), which has revolutionized the capabilities of these systems in understanding and responding to user preferences and needs. Generative AI models, such as Chat-GPT, possess the ability to generate human-like text, engage in dynamic conversations, and adapt their recommendations based on user interactions. While business cases building on generative AI are still in the process of being fully explored, the impact of this technology on search engines is already poised to be profound. A notable example of this is Microsoft's recent collaboration with OpenAI, leading to the integration of Chat-GPT into the Bing search engine. This partnership demonstrates the potential of generative AI in transforming the way search engines function and the experiences they offer to users. While most generative AI tools are currently designed in a way to protect their users from conspiracy theories and even very hesitant to give recommendations (e.g., on whom to vote for in the upcoming election), there is, without doubt, both commercial interest in the companies providing those tools and organizations willing to pay for influencing algorithmic recommendations towards their own interest (see the Cambridge Analytica example outlined in the previous section). Considering the extensive media coverage and enthusiastic celebration surrounding the technology underpinning Chat-GPT, the answers and recommendations generated by generative AI systems are likely to be perceived as highly relevant by their users. Nevertheless, it is crucial to acknowledge that the decision-making process governing these algorithms rests not in the realm of democratic discourse but rather within the realm of commercial interests controlled by a small number of major tech corporations. For truth to prevail over falsehoods, including the spread of fake news and conspiracy theories, within a democratic discourse, it is imperative for any free and open society to guarantee that truth remains relevant.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Artificial Intelligence in the Anthropocene? Yes, Naturally!

Uwe Voigt
Universität Augsburg, Germany
ORCID: 0009-0007-3763-0814

Abstract

This paper tries to clarify the concepts of intelligence, technology, Artificial Intelligence, Anthropocene and nature so that this throws some light on their mutual connection. In this analysis, intelligence is seen as the ability to recognize the borders of one's own thinking as problems, which includes a qualitative and reflexive consciousness of problems. Technology originates from this consciousness of problems and consists in the attempt to solve problems in order to reach certain given goals. Hence, Artificial Intelligence is already inherent in technology, but starts only recently to be discussed as such. This belongs already to the problem-complex of the Anthropocene as a technological transformation of the environment on this planet, which in turn raises the question of nature, which is understood here in a relational way, as the mutual relation between possible subjective points of view. That leads to the following result: Artificial Intelligence can be "natural" insofar it is able to blend into such a mutual relation, what seems also to be advisable in the Anthropocene.

Keywords: artificial intelligence (AI), technology, Anthropocene, nature, problems of consciousness

Introduction

The title of this paper is intended to provoke, because it brings together what at first glance does not seem to belong together: How can Artificial Intelligence be natural? Can it be natural after all? Is it not an artificial product and as such unable to be natural? And isn't, for this very reason, the talk about the natural in the Anthropocene obsolete, as we are facing in this era an environment which is thoroughly shaped by technological influences? The title, however, raises the claim that all this can be put together, even thought together. Is this claim justified, and if so, how can that be?

That question now is to be clarified by looking at the pertinent concepts and their mutual connection and thus structuring what follows and what also for brevity's sake cannot be but a sketch. We start with a concept of intelligence, as this is important for the understanding of some other concepts: for a concept of technology, which will turn out to be a certain application of intelligence; and hereby for a concept of Artificial Intelligence, which belongs to this application; and also for a concept of the Anthropocene as the technological transformation of the environment on this planet. The further question, if and how a technologically transformed environment in general, and especially the Artificial Intelligence belonging to it, can be natural, then leads to an inquiry into the concept of the natural. This inquiry is challenged by the notorious problems of the concept of the natural and also on its background, the concept of nature; and it meets this challenge by the attempt to discover within these problematic concepts, by the means of reflexive logic, a core according to which the natural is what we can name without having to be able to properly describe it. This will finally lead to a plea for conceiving of Artificial Intelligence in the Anthropocene as something natural and, for reasons still to be elucidated, also to aim for it as something natural.

A Concept of Intelligence

The word "intelligence" is often used in an inflated way and without clearly graspable content. Nevertheless, it seems to be astonishingly easy to say in which such a content consists, if that word (or one of its synonyms) is used in a terminological manner. In this case, intelligence is understood as the ability to adapt to environmental opportunities in the best way possible.[1] But precisely this seemingly simple access shows why talking about intelligence can lead to conceptual confusion so easily: Whose ability and, accordingly, whose environment are at stake here? Is even water intelligent, when it makes way through the landscape and adapts to the opportunities given for

[1] See, Marion Friedrich, "Intelligenz aus philosophisch-psychologischer Sicht," in *Natürliche und Künstliche Intelligenz im Anthropozän,* ed. Joachim Rathmann, Uwe Voigt (Darmstadt: Wissenschaftliche Buchgesellschaft, 2021), 135-162, 146.

that in order to leave the landscape as fast as possible? Is there an emotional intelligence which differs essentially from its cognitive counterpart? And which are, in each single case and moreover generally, the criteria of evaluation which are to be applied to the way of adaptation? Would it not be too cynical to claim that going extinct is the best way of adaptation for a species whose members could not but lead a life of suffering in their given environment? The concept of intelligence, as it seems, needs to be elucidated.

Such an elucidation can be found in Johann Gottlieb Fichte's Science of Knowledge (*Wissenschaftslehre*).[2] Under the title of the ego, Fichte refers here to a finite instance which produces ideas and at the same time is their carrier. In the absence of self-reflection, the ego would just proceed with producing ideas and being conscious of them, thus going the way of pre-reflexive thinking.[3] This way comes to an end, however, when it meets a resistance, an obstacle (in transcribed Greek: a *pro-blêma*). This obstacle consists of the ego encountering an idea which has not been produced by it. Thus, thinking hits its limitation—it is thinking its own limitation—and hereby is thrown back upon itself; it becomes reflexive. So, consciousness becomes self-consciousness and problem-consciousness at the same time: I have encountered a *problem*; *I* have encountered a problem. Here, a problem is something to which the ego can refer, and which raises the question how the ego can refer to it after all and how the ego should refer to it. Such an ego, having become reflexive and self-reflexive, and at the same time intentional, Fichte calls "intelligence." He uses this concept to signify the carrier of an according property, as we still do today when speaking of Artificial Intelligences. The less this semantic nuance hinders us to see here also an elaboration of our usual understanding of intelligence as a property: Intelligence is the successful handling of a problem in the mentioned sense by an ego or, according to the recent discourse, of a subject, with the standards for evaluation of the success stemming from the very thinking of that subject. As Uwe Meixner argues,[4] such intelligence can be there only in a and for a subject, only as intelligence of consciousness; "consciousness" means here in turn the fact that there is something which is given to that subject as such; that, in a current diction, it is like something just to be that subject. Accordingly, Fichte conceives of elementary problems as simple qualities of experience (what we now would call qualia), on which thinking is refracted, because it can think them as not having been produced by it, and it can think itself as being unable to analyze them further.[5] This conception of intelligence is linked to the ability of qualitative, aesthetic

---

[2] On the following, See, Johann Gottlieb Fichte, *Grundlage der gesammten Wissenschaftslehre* (1794, 1802), in *Fichtes Werke,* Vol. 1, ed. Immanuel Hermann Fichte (Berlin: Veit & Co., 1845-1846; reprint, Walter de Gruyter & Co., 1971), 85-328.

[3] See, Marc Borner, *Über präreflexives Selbstbewusstsein: Subpersonale Bedingungen—empirische Gründe* (Münster: Mentis, 2016).

[4] See, Uwe Meixner, "Bewusstseinsintelligenz und Künstliche Intelligenz," in *Natürliche und Künstliche Intelligenz,* ed. Joachim Rathmann, Uwe Voigt (Germany: wbg, 2021), 13-31.

[5] On these considerations and their relevance to modern debates, see, Dieter Henrich, *Dies Ich, das viel besagt: Fichtes Einsicht nachdenken* (Frankfurt am Main: Klostermann, 2019), 156.

experience.[6] To think qualia as no more analyzable means at the same time to think of them as simple, which Fichte illustrates with the notion of the geometrical point. With the help of this notion, the problems caused by qualia can even be quantified, turning out to be problems among other problems. Peirce comes to our aid in this step: Precisely because of their simplicity and hence because of their quantifiability, points serve as limitations and so also as connections between complex geometrical structures.[7] Accordingly, also complex problems respectively complexes of problems can be understood as consisting of connections and transitions which have a qualitative character, so that intelligence in dealing with them and between them, so to speak "between the lines,"[8] always also means to become aware of that qualitative character. Even as problem-related thinking, intelligence therefore is connected to aesthetic experience. This experience gives the problems an importance that lifts them above the background noise, and it gives the subject facing these problems the motivation to deal with them.

The concept of intelligence which is offered here can be summarized in the following way: Intelligence is the ability to recognize the limitations of one's own thinking as problems, which includes a qualitative consciousness of these problems, and, based on this, to refer to these problems as well as to oneself as a thinking instance.

A Concept of Technology

Here, technology is understood in line with Thomas Heichele, who in turn refers to Ernst Cassirer und Hans Sachsse.[9] For Heichele's concept of technology is not only presented very clearly but also fits very well to the notion of intelligence provided above. According to that concept, intelligence is primarily a certain way of intelligent action, more precisely: a certain way of intelligent action dealing with itself and its problems. As we have seen, intelligent action is directed towards itself and its problems, being reflected upon itself by its problems and thus being reflexive and intentional at the same time. The technological way of this action consists in dealing with oneself and one's problems in the framework of a means-ends-relation. The ends here are not the action or the problems, but something which is beyond these problems. From the perspective of technology, the problems appear as obstacles on the way to a goal towards which the action is directed. If we represent, with Fichte, a

---

[6] On this kind of experience See, the contribution of Stefanie Voigt to this issue.
[7] See, Helmut Pape, *Die Unsichtbarkeit der Welt* (Frankfurt am Main: Suhrkamp, 1997), 378-445; Helmut Pape, "Kontrollierte Abstraktion und Selbstkritik," in *Künstliche Intelligenz und menschliche Person* (Marburg: Elwert, 2006), 107-121.
[8] See, Dietrich Dörner, "Mülltonne, Speerschleuder und Fahrradschlauch—Über künstliche und natürliche Intelligenz," in *Natürliche und Künstliche Intelligenz,* 217-233, 218.
[9] See, Thomas Heichele, "Künstliche Intelligenz im Licht der Technikphilosophie," in *Natürliche und Künstliche Intelligenz,* 79-108, section 2.

problem as a point which refracts the continuous line of pre-reflexive action, then the goal (the ends) is another point lying beyond that first point, beyond the problem. The intelligent subject has chosen that second point as goal of its action, which cannot be reached because of the problem, and therefore searches a way to solve the problem. As Heichele argues, this way is a de-tour and at the same time a tour-towards. The first means which the action gets a hold of is itself, respectively it understands itself as a means to find further means which might lead to the goal. The goal, however, can only be reached if the problem is solved; so such action is problem-solving thinking par excellence.[10] For the given reasons, what it intends immediately is not its goal but the problem to be solved. The means which it uses to solve the problem and so reach its goal are used to be called "technology" as well. Intelligent action as problem-solving thinking therefore is done through the according means. Hereby intelligent action uses itself as such means, it is also adopting a technological character. In this sense, technology always implies an artificial intelligence: an intelligence which gives itself a technological character and thus serves a certain *technê*, some artisanship. This becomes evident in the so-called technology of the intellect, in which problem-solving thinking tries to solve problems of its own procedures in a technological way (e.g., by controlled application of formalized logic).

Here technology enters into an ambivalent relation to the finite intelligent subjects which are using it: As finite subjects, they cannot but approach at least some problems in a technological way. This, however, threatens to undermine their very subjectivity: The more technology succeeds, the more it masks the problem it is meant to solve, thus bereaving its subject of the occasion, offered by that problem, to become conscious of itself in a reflexive way. If the subject remains pre-reflexive as long as is not challenged by problems, it can also enter, so to speak, a post-reflexive state if it solves problems through technology without still becoming aware of them. Fictive scenarios of doom which can be found in literature and popular culture on this background can be seen as a medial reflection of the threatening extinction or at least subjugation of finite subjectivity by its own technology.[11]

This threat becomes even more acute by a certain form of technology.[12] Classic technology adapts to the problem for whose solution it is applied, and thus it takes a form which is in accordance with the goal and the problem; in the sense of this adaption and the correspondence at least aimed at with it, such technology is analog. So classic technology splits up into a manifold of different technologies, according to

---

[10] On technological action as problem-solving, see, Heinrich Popitz, *Wege der Kreativität* (Tübingen: Mohr Siebeck, 1997), 106. On the history of technology as a history of subsequent problem-solutions giving rise to new problems, see, *Der Aufbruch zur Artifiziellen Gesellschaft: Zur Anthropologie der Technik* (Tübingen: Mohr Siebeck, 1995).

[11] See, Bernhard Irrgang, *Roboterbewusstsein, automatisiertes Entscheiden und Transhumanismus: Anthropomorphisierungen von KI im Licht evolutionär-phänomenologischer Leib-Anthropologie* (Würzburg: Königshausen & Neumann, 2020), 9-34.

[12] See, Martin Heidegger, "The Question Concerning Technology," in *The Question Concerning Technology and Other Essays*, trans. William Lovitt (New York: Harper & Row, 1977), 1-35.

the different goals and problems. Now, the very plethora of technologies can turn into a problem, for which there seems to be another technological solution: the development of a unique, homogenous technology. That technology does not adapt to the given goal and the encountering problems; it rather adapts them to itself and turns them, as independent from the pertinent realm of objects as possible, into something it can process. That such a technology is possible is grounded already in the quantification done by the problem-consciousness: Notwithstanding their different qualitative characters, that consciousness conceives of its problems as different unities. Thus, the foundation of the unique, homogenous technology can be laid by processing these problems as mere quantities which can be counted with the help of one's fingers (*digit*). In this broad sense, that kind of technology can be called digital. At the turn of the 20th century, it started to boom also due to progress in the technology of the intellect thanks to innovations on the field of logic, which succeeded then to present quantity in a strictly formal way.[13]

By becoming a problem, however, technology can also contribute to self-reflection. This self-reflection can proceed from the pole of the subject and from the pole of the problem to be solved by technology—from the subject which becomes aware of its ambivalent relation to technology, and from the problem, if the following connection comes to mind: Technology is not immediately directed to the given ends, but to a problem which prevents that ends from being reached. Thus, for technology the very acting upon the problem becomes an end. Any way the problem is acted upon, technology is always also directed to whatever the problem is connected to, and changes also these connections in acting upon the problem.[14] Therefore, technology is always accompanied by side-effects, which have not been intended in the pursuit of the given purpose and the acting upon the according problem.[15] The more powerful the technology used, the graver these side-effects can become. Even the threatening autonomy of technology as against the subject which used it can be understood as such a side-effect, in which the connection between problem and technology turns out to be stronger than the connection between subject and technology. In any case, the side-effects of technology we encounter in environmental questions contribute to critical reflection on technology in our time.

---

[13] See, Klaus Mainzer, *Computer—neue Flügel des Geistes?* 2nd ed. (Berlin: De Gruyter, 1995); Martin Davis, *The Universal Computer: The Road from Leibniz to Turing* (London, New York: A K Peters, 2011).
[14] See, Peter Sloterdijk, *Eurotaoismus. Zur Kritik der politischen Kinetik* (Suhrkamp: Frankfurt am Main, 1989), 23, 29.
[15] On the concept of the side-effect See, Jens Soentgen, *Konfliktstoffe: Über Kohlendioxid, Heroin und andere strittige Substanzen* (München: oekom, 2019), 45-49.

A Concept of Artificial Intelligence

We have already seen that technology in a certain way always implies Artificial Intelligence. From the mid-20th century onwards, this connection unfolded, and at first in a casual manner, as the catchphrase "Artificial Intelligence" was coined to acquire third-party funding for a pertinent conference.[16] This phrase is meant to signify a technological product whose activities are in accordance with intelligent action. This accordance can be interpreted in two ways: Either the Artificial Intelligence is an intelligent agent, too, i.e., a problem-conscious subject; then we would talk of Strong Artificial Intelligence. Or these actions do correspond to intelligent action, but are not activities of such a subject; the product in question just acts as if it was intelligent without being so. This is typical of a Weak Artificial Intelligence. Moreover it might be that Artificial Intelligence can solve problems of any kind, thus becoming the completion of digital technology in the sense mentioned above. In such a case, we would be confronted with a General Artificial Intelligence.[17] Alternatively, Artificial Intelligence might just be able to solve problems of a certain kind. This would be a Narrow Artificial Intelligence, so to speak in the tradition of analog technology, even if based on digital means. This kind of Artificial Intelligence is applied in many ways today. The questions, if and how Strong Artificial Intelligence and General Artificial Intelligence are possible (and if they would be one and the same or still different), remain notoriously open. The connection between Strong Artificial Intelligence is argued for by Dietrich Dörner.[18] It can also be corroborated by Sean McGrath's contribution to this issue.[19] Uwe Meixner has championed this view, too.[20] According to it, firstly, Strong Artificial Intelligence seems to presuppose a consciousness which can experience qualia and therefore turn itself into a problem-consciousness. Secondly, a General Artificial Intelligence would have to be also a Strong Artificial Intelligence, because the general recognition and processing of problems of any kind obviously has to be based on a consciousness aware of problems of any kind and the complexes they can form. A Strong Artificial Intelligence centered around a phenomenal consciousness might also evade the metaphysical problems duly raised by Sebastian Rosengrün, which strike an abstract conception of Artificial Intelligence remote from consciousness.[21]

---

[16] See, Sebastian Rosengrün, *Künstliche Intelligenz zur Einführung* (Hamburg: Junius, 2021), 13-17.
[17] See, Sean McGrath's contribution to this issue.
[18] See, Dietrich Dörner, "Mülltonne, Speerschleuder und Fahrradschlauch—Über künstliche und natürliche Intelligenz," in *Natürliche und Künstliche Intelligenz im Anthropozän,* ed. Joachim Rathmann, Uwe Voigt (Germany: wbg, 2021), 217-234.
[19] See, McGrath's contribution in this volume.
[20] See, Uwe Meixner, "Bewusstseinsintelligenz und Künstliche Intelligenz," in *Natürliche und Künstliche Intelligenz im Anthropozän,* edited by Joachim Rathmann and Uwe Voigt (Germany: wbg, 2021), 13-32.
[21] See, Sebastian Rosengrün's contribution in this volume.

A Concept of the Anthropocene

To the context of the reflection on technology, which is made more urgent by the rise of Artificial Intelligence, belongs the naming of the current geological age as Anthropocene.[22] At first glance, this seems just to be "a new age of the human being." But this age manifests itself in the effects which the technological actions of human beings exert on their environment. These effects are empirically well documented, and, in many cases, they exceed all other factors concerning their influence on the environment. The technological means by which this is brought about, blend into the environment shaped by them, as Jens Soentgen has shown in his exemplary study of the river Lech which has been turned into a cyborg, an entity with natural components and a technological infrastructure.[23] As such a mixed entity, the Lech develops also activities which have not been aimed at with the human influences on this river, and this makes him a telling example as a part of a whole, a planetary environment, which is more and more destabilized and dynamized by the human impact in the Anthropocene. In this process, technology becomes such a crucial factor that the temporally oriented concept of the Anthropocene now has been flanked by the more spatially oriented concept of the Technosphere: the complex system formed by technology, which organizes itself more and more without regard to human interests because it is based primarily on side-effects.[24] This system encompasses and absorbs the biosphere; and if the biosphere can be understood as a self-organizing earth system which has organic character, what James Lovelock expressed under the name of "Gaia,"[25] then we are about to experience how Gaia is penetrated and assimilated by the Technosphere—and how also here, on a planetary scale, a cyborg arises which unfolds more and more dynamics of its own.[26] These dynamics are guided by those entity's own ends and therefore have to deal with the according problems, there being no guarantee that these ends and problems are also ours and that at least some of our purposes are not its problems.

Hence, the Anthropocene can be understood as the technological transformation of the environment on this planet, in a threefold sense: It is a

---

[22] See, *Das Anthropozän. Zum Stand der Dinge,* ed. Jürgen Renn, Bernd Scherer (2nd ed., Berlin: Matthes & Seitz, 2017); *Das Anthropozän. Schlüsseltexte des Nobelpreisträgers für das neue Erdzeitalter,* ed. Michael Müller (München: oekom, 2019); *Anthropozän zur Einführung,* ed. Eva Horn, Hannes Bergthaller (Hamburg: Junius, 2019); *Mensch—Natur—Technik. Philosophie für das Anthropozän,* ed. Thomas Heichele (Münster: Aschendorff, 2020).

[23] See, Jens Soentgen, "The River Lech—a Cyborg," *Analecta Hermeneutica* 10 (2018), online: https://journals.library.mun.ca/ojs/index.php/analecta/article/view/2059/1649 (accessed December 31, 2022).

[24] See, *Technosphäre, ed. Katrin Klingan, Christoph Rosol (Berlin: oekom, 2019).*

[25] See, James Lovelock, *Gaia: A New Look at Life on Earth,* 2nd ed. (Oxford: Oxford University Press, 2016).

[26] See, Uwe Voigt, "Inside the Anthropocene," *Analecta Hermeneutica* 10 (2018), online: https://journals.library.mun.ca/ojs/index.php/analecta/article/view/2057/1647 (accessed December 31, 2022); Uwe Voigt, "Das Anthropozän als geistige Umweltkrise," in *Mensch—Natur—Technik,* 85-102; "Wissen um Atmosphären—Bildung für das Anthropozän?," *Comenius-Jahrbuch* (2020): 13-32.

technological transformation, a process triggered by technology; it is moreover a technological trans-*formation*, because what is formed here also acquires the form of technology; and it is a technological *trans*-formation, which is guided by means and problems that may lay beyond our own. Like any complex of problems, also this problematic situation has a certain qualitative character, i.e., it is like something to be in it. Facing manifold phenomena on different levels, from individual experience up to international developments, may give rise to the suspicion that we have to deal with an atmosphere of logical narcissism—the identification of the subject with the point of view it has taken, which results in violent clinging to that point of view.[27] Such a situation is connected with an "ecology of fear,"[28] which forces human and non-human subjects together in a "democracy of suffering."[29]

Because of the scales on which this situation unfolds, it can be just sketched from the point of view of an individual human being, as it is attempted here, and it can be grasped by a multitude of measurements, which are the tasks of different scientific disciplines. Also the humanities have a place in this field, as empirical data and qualitative aspects are interwoven in that planetary atmosphere. The according interdisciplinary challenge is taken up by the Environmental Humanities,[30] which dedicate themselves to the cultural reflection of environmental conditions, also and especially as to be found in the narratives of the Anthropocene.[31] In a situation as complex as this is, we need obviously all kinds of intelligence which can help us to grasp and cope with the current problems; hence, we need also Artificial Intelligence with its paramount power of data-processing, which seems to be the means of choice in the Anthropocene.[32] Moreover, Artificial Intelligence in union with further technological means might prove to be a powerful actor which could help us to solve the problems of the Anthropocene, maybe even overcoming this geological age, finding a happy end in a new epoch of friendly cyborgs.[33] After our recent considerations, however, there is reason to doubt this consoling scenario: Like any other technological product, Artificial Intelligence is also a part of the Technosphere and therefore a part of the complex of problems with which we have to deal. Even if the technologically transformed earth system should finally act like an intelligent

---

[27] See, Footnote 26.
[28] See, Jens Soentgen, *Ökologie der Angst* (Berlin: Matthes & Seitz, 2018).
[29] See, Todd Dufresne, *The Democracy of Suffering: Life on the Edge of Catastrophe: Philosophy in the Anthropocene* (Montreal: McGill-Queen's University Press, 2019).
[30] See, *Environmental Humanities: Beiträge zur geistes- und sozialwissenschaftlichen Umweltforschung,* ed. Matthias Schmidt, Hubert Zapf (Göttingen: V&R unipress, 2021).
[31] See, Thomas Schmaus, "'Erzähl uns deine Erdgeschichte!' Narrative Identität im Anthropozän," in *Comenius-Jahrbuch* 28 (2020), 33-54.
[32] See, Klaus Mainzer, "Vom Anthropozän zur Künstlichen Intelligenz. Herausforderungen von Mensch und Natur durch Technik," in *Mensch—Natur—Technik,* 155-168. See also the contributions of Mike Meitner and Joachim Rathmann in this issue for critical reflection.
[33] See, James Lovelock, *Novacene: The Coming Age of Hyperintelligence* (London: Penguin, 2019).

subject, it is not guaranteed that its actions should serve our purposes and solve our problems.[34]

So, Artificial Intelligence in the Anthropocene makes the very concept of technology as presented here problematic: It becomes the concept of a problem which we as finite subjects encounter also because it is like something to be in an according situation. Because the concept of Artificial Intelligence is implicitly inherent to the concept of technology anyway, thus technology in the Anthropocene turns out to be a problem for finite, human intelligence altogether. As already hinted at, the concept of nature might be of help in this situation; so, we turn to it now.

A Concept of Nature

As we have seen, subjects in the Anthropocene face the challenge to critically discuss their own point of view, in an environment which is technologically transformed to an extent that the question can be raised whether this environment still can satisfy the concept of the nature, nay, whether that concept still is of any use.[35] Here a great cycle in the history of concepts comes to its conclusion (and, as usual, opens up a new one), because in occidental tradition the work on the concept of nature has always served to determine the point of view of the subjects doing this work. This work has been proceeding in three steps, which here are depicted in a generalizing continuation of distinctions introduced by Elisabeth List, who follows Serge Moscovi, and Jens Soentgen, who follows Gregor Schiemann.[36] These steps lead from an intrinsic over an extrinsic to a relational concept of nature.

For a good reason, the occidental work on the concept of nature begins in the early time of Greek philosophy, which is confronted with a multitude of points of view: already within the Greek city-states with their manifold political and cultural conditions, and moreover in contact with different neighboring civilizations.[37] This situation made wonder how, within such a multiplicity, a reasoned and reasoning discourse (a *logos*) might be justified.[38] One way to give an answer is to find something which can be referred to in the same way from any point of view. What is found here is that which, so to speak, grows on any point of view, because it unfolds itself

---

[34] See, Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
[35] See, Sean McGrath, *Thinking Nature: An Essay in Negative Ecology* (Edinburgh: Edinburgh University Press, 2019).
[36] See, Elisabeth List, *Vom Darstellen zum Herstellen: Eine Kulturgeschichte der Naturwissenschaften* (Weilerswist: Velbrück, 2007), 165-168; Jens Soentgen, "Der ökologische Naturbegriff," in *Mensch—Natur—Technik* (2020): 115-130.
[37] See, Jürgen Habermas, *Auch eine Geschichte der Philosophie. Vol. 1: Die okzidentale Konstellation von Glauben und Wissen,* 4th ed. (Frankfurt am Main: Suhrkamp, 2020), 417.
[38] See, Daniel-Pascal Zorn, *Vom Gebäude zum Gerüst: Entwurf einer Komparatistik reflexiver Figurationen in der Philosophie* (Berlin: Logos, 2016).

anywhere of its own: nature.[39] Its concept is understood in a twofold manner:[40] According to its extension, it refers to the known natural kinds; according to its intension, it refers to everything which has the principle of its motion and rest within itself, as the famous formulation by Aristotle tells us. With this notion of nature we are acquainted on our given points of view, insofar we ourselves belong to what is natural, and therefore we can give the natural its usual names, even if we have to correct ourselves from time to time. We refer to nature from the inside. Here, subjectivity is conceived of as closely connected to the natural, as it is expressed in the Aristotelian conception of the soul as the form of a natural, organic body. This concept of nature is plausible because it offers itself from the point of view of finite human subjectivity which also has the principle of its own dynamics within itself and experiences itself as being at home in a world of entities which also follow inner principles, even if they are of a different kind. This intrinsic concept of nature, hence, is anthropocentric (conceived of from a human point of view) and also anthropomorphic (conceived of according to the model of our having a point of view). As the subject thinks of itself as a unity on its point of view, so it thinks also the natural as a set of objects, of unities in their time and place. In the light of this concept of nature, the world is seen as a cosmos, as a beautiful hierarchic order made of single things, which blend into it due to their inner principles respective their "natures."[41]

This concept of nature has proved to be very influential; it keeps informing our current discourse on true, unfalsified nature in the sense of wilderness.[42] This concept of nature seems also to bring about a sharp distinction between the technological, including Artificial Intelligence, and the natural, insofar the technological does not contain the principle of its motion and rest within itself, but has received it from outside.

The intrinsic concept of nature, however, has to face a problem: With it, nature is thought as a manifold of potential points of view for subjects. From which point of view is this done, from which point of view might this be possible after all? In thinking so, obviously a point of view is used which lies beyond nature, at least beyond the moved and resting which falls under its concept. This problem was articulated sharply already in Eleatic philosophy. Aristotle tried to counter this by connecting the soul as carrier of subjectivity as a form to the body informed by it and at the same time, in the human case, ascribing to the soul a part which gives it the ability of intellectual insight,

---

[39] See, Thomas Buchheim, *Die Vorsokratiker: Ein philosophisches Porträt* (München: C.H. Beck, 1994), 91-95, 152-154.

[40] On what follows, see, Gregor Schiemann, *Natur, Technik, Geist: Kontexte der Natur nach Aristoteles und Descartes in lebensweltlicher und subjektiver Erfahrung* (Berlin-New York: De Gruyter, 2005).

[41] See, Stefanie Voigt, Uwe Voigt, "Head Jewellery—a Theory of the Theory of Jewellery," in *Thinking Jewellery: On the Way Towards a Theory of Jewellery,* ed. Wilhelm Lindemann (Stuttgart: Arnoldsche Art Publishers, 2011), 80-93.

[42] See, Gregor Schiemann, "Pluralität der Natur," *Bremer Philosophica* 4 (1999): 31f.

a mind (*nous*), which "comes from the outside," which can conceive of nature precisely because it does not (totally) belong to it.[43]

How that mind has to be understood was a question on which the Aristotelian tradition labored and about which it got into a crisis that lasted until early modern times.[44] In order to overcome this crisis, René Descartes determined the relation of subject and nature in a new way. For Descartes, too, nature is an object of reference for subjects. But he does not conceive of subjectivity as being situated within nature, rather, according to the result of his methodological doubt, opposing nature. Subjects can refer to nature because they find concepts within themselves—especially the concept of extension—by which they can think states of motion and rest; and the natural, insofar it falls under these concepts, can be referred to subjects. The difference between subjectivity and nature is captured by Descartes in his famous juxtaposition of the subject as a thinking thing which is not extended and of the natural as an extended thing which is not thinking. As the natural falls under general concepts which can be thought with mathematical precision, it is to be understood as a realm of strict laws which can be formulated through those concepts—as a realm in which the laws of nature reign supreme. This realm is opposed by subjectivity, which in free self-reflection recognizes its own essence, thinking, and the essence of the natural. So, from the point of view of subjectivity, an external concept of nature is gained: According to it, on the one hand, the natural is what is external to subjectivity, and on the other hand, the natural is to be taken also within itself as a manifold of externalities, as bodies whose relations are not completely determined by themselves, as the internal view had seen it, but by the laws of nature.

This extrinsic concept of nature is no longer anthropocentric because it does not refer to the point of view of the human as a being which is (also) natural. It is rather acentric because neither within the manifold natural nor within the relation between the subject and the natural there is a center; this does justice to the transition from a closed cosmos to a universe which, at least at first glance, seems to be infinite.[45] The extrinsic concept of nature is also no longer anthropomorphic because the human being as a natural entity is now understood as just one body among other bodies. The ways of these bodies are determined by the laws of nature and therefore can be reconstructed with the help of mathematics. If bodies display certain kinds of complex activities, they can do so because they are accordingly constructed automata. So the extrinsic concept of nature turns out to be techno morphic. The products of technology, however, are opposed by subjectivity and also by intelligence in the sense given above. From this perspective, mind does not come into nature from the outside, it always stays on the outside.

---

[43] See, Uwe Voigt, "Wozu braucht Aristoteles den 'Geist von draußen' in seinen biologischen Schriften?," in *Antike Naturwissenschaft und ihre Rezeption* 17 (2007): 29-38.

[44] See, Schiemann, "Pluralität der Natur," 165.

[45] See, Alexandre Koyré, *From the Closed World to the Infinite Universe* (Baltimore: Johns Hopkins Press, 1957).

This concept of nature, too, was very influential. Not least it made the technological transformation of the environment in the Anthropocene to seem thinkable and feasible.[46] But also the extrinsic concept of nature has a problem to face, which articulates itself through the question how subject and nature, given their basic difference, can be related to each other. Descartes himself tries to answer this question by referring again to nature, conceiving it now as the connection between the thinking subject and its extended object, a connection which, in the case of the embodied human being, is very intimate as Descartes has to confess. What can be learned from this irritating use of the concept of nature is the following: The opposition between subject and nature can be thought only from a point of view which lies already beyond that opposition, from which it in turn the (seemingly mere) difference in question can be thought as mutual relatedness. The thinking subject has already turned out to be related in such a way from its point of view. Mutuality can be thought by ascribing a point of view also to the natural. Seen this way, nature refers to whatever is able to refer, from a certain point of view, to something else which has to be granted a point of view, too. Hence, the natural can be characterized by its eventual mutual relatedness, so that we can call this a relational concept of nature. In the light of this concept of nature, the natural can be seen as a web of relations between eventual points of view.

Independently from reflections on the history of concepts, Saul Kripke has elaborated an analysis of the logic of naming natural kinds, which is pertinent here:[47] We encounter individual specimens of these kinds and in doing so take a sample of them. On this occasion, we give a name to these kinds which refers to them as a rigid designator, whatever constitutes the kind in question. What does constitute them, either is immediately given by the sample itself—if we have to deal with a quality of experience like pain—or can be found out through further inquiry, as in the case of the biological kinds. These examples may seem to be anthropocentric and anthropomorph again, because it are humans who do the naming, the feeling and the inquiry. However, we can not only think but also experience that also human beings can become members of a sample, although we tend to repress this fact within our technological society, as Val Plumwood has elaborated after her near-death encounter with a crocodile.[48] Also human beings can be referred to; so there is mutuality here, at least in principle.

This relational concept of nature has been signified in two tellingly different ways by List and Soentgen: For List it is the concept of a cybernetic state of nature.[49] Soentgen instead is working on an "ecological" concept of nature.[50] In the first case, the relation which is central to the concept of nature is thought from the point of view

---

[46] See, McGrath, *Thinking Nature*, 156f.
[47] See, Saul Kripke, *Naming and Necessity* (Oxford: Blackwell, 1980); Kripke*, Reference and Existence. The John Locke Lectures* (Oxford: Oxford University Press, 2013).
[48] See, Val Plumwood, *The Eye of the Crocodile* (Canberra: Australian National University Press, 2012).
[49] See, Elisabeth List, "Vom Darstellen zum Herstellen," *Zeitschrift für Kulturphilosophie* 1 (2014): 71-84.
[50] See, Soentgen, "Der ökologische Naturbegriff," 116-118.

of technology and thus from the extrinsic pole of that concept; in the second case, the relation is thought from the natural and so still from the intrinsic concept of nature respectively from the intrinsic pole of the concept of nature. According to the relational concept of nature, nature is a realm of relations which all have this "bipolar" characteristics, making the given point of view conceived of as the point of reference of another point of view.[51] This view matches, by the way, the stronger thesis that every possible point of view is embedded in the point of view of a comprehensive, transcendental subject. [52] So the relational concept of nature turns out to be polycentric.

Insofar the object can be thought as the point of view of a subject, that subject can fall under a merely external determination as little as the subject which thinks that object. Thereby the background in the logic of reflection is revealed for the observation Kripke made as to the naming of natural kinds: Naming is not necessarily connected to an adequate determination of what is being named. This determination can be left open. The natural in the sense of the relational concept of nature, hence, is the realm of what can be named without having to be adequately determined for that purpose. Later determination is not excluded hereby, but it is also not pre-determined. The relational concept of nature is, so to speak, polymorph. This makes the relational concept of nature fit in with an understanding of contemporary science as having to deal with a web of relations.[53] This makes the scientific access to the Anthropocene an eminent interdisciplinary enterprise.[54]

Also the relational concept of nature has a problem of its own, namely how one's own point of view can be thought from the outside and how other points of view can be thought of as having their subjective inside. This is also the core-problem of contemporary panpsychism as the version of philosophy of mind which is in accordance to a relational concept of nature.[55] As it might have become clear by now, after all, the concept of nature is the concept of a problem, namely the concept of the problem how a subject can think in relation to itself as well as to other subjects. The relational concept of nature offers the advantage of not masking, but rather highlighting the structure of this problem.

---

[51] The systematic core of this concept of nature can be tracked back to Schelling; See, Eckart Förster, *Die 25 Jahre der Philosophie: Eine systematische Rekonstruktion,* 3rd ed. (Frankfurt am Main: Klostermann, 2018).

[52] See, Uwe Meixner, "Idealism and Panpsychism," in *Panpsychism: Contemporary Perspectives,* ed. Godehard Brüntrup, Ludwig Jaskolla (Oxford: Oxford University Press, 2017), 387-405.

[53] See, Ernst Cassirer, *Substanzbegriff und Funktionsbegriff* (Darmstadt: Wissenschaftliche Buchgesellschaft, 2000).

[54] See, the contribution of Stefanie Voigt in this volume,

[55] See, Uwe Voigt, "Eingestimmte Subjekte? Das Kombinationsproblem des Panpsychismus im Licht der Atmosphärenkonzeption der Neuen Phänomenologie," in *Die Macht der Atmosphären,* ed. Barbara Wolf, Christian Julmi (Freiburg im Breisgau: Alber, 2020), 60-74.

Intelligence in the Anthropocene: Natural and Artificial

Taking into consideration a relational concept of nature allows to answer the question asked at the start of this paper: if and how Artificial Intelligence in the Anthropocene could or even should be natural. According to the relational concept of nature, natural is what fits into a mutual relatedness in which one subjects thinks of its own point of view as the object of the point of view of another subject and thus acknowledges that other point of view as eventually belonging to another subject. Hence, intelligence in general is natural if it is based on a reflexive consciousness of problems and thus able to recognize its being placed in a point of view and confronted with the qualitative character of the problem. As we have seen, technology can short-cut this relation, if it is just used to solve the problem. Then technology is opposed to the relatedness characteristic of nature because it masks that very relatedness and prevents its reflection by the subject. Such use of technology can be called artificial in a pejorative sense. In contrast to that, technology can also support the reflection of the subject which then does not need to be concerned with any problems but just with those challenging its reflection as such. Under these circumstances, technology can serve the reflection of the relation to the environment.[56] As a product of technology, Strong Artificial Intelligence—which alone deserves our attention here, as seen—is part of the problem posed by a technologically transformed environment in the Anthropocene. This problem cannot be solved in a merely technological way because that would only perpetuate it. In the Anthropocene, intelligence, be it human, artificial or of another kind, faces the challenge to preserve and, if possible, increase its ability of reflection, in order to do justice to the complexity of the situation. A criterion for the success in tackling this challenge can be the extent in which intelligence can blend into the mutual relations which even a technologically transformed environment is still offering, thus staying or becoming natural. Artificial Intelligence will encounter its natural counterpart within that transformed environment, the Technosphere, one way or the other. If and how we succeed in preserving, cultivating and developing the according mutual relatedness as a space for experiencing shared reflection and, thus, rationality, may be a touchstone for any kind of intelligence in the Anthropocene. [57]

---

[56] For a study of a classical use of technology in this sense, see Thomas Heichele, *Die erkenntnistheoretische Rolle der Technik bei Leonardo da Vinci und Galileo Galilei im ideengeschichtlichen Kontext* (Münster: Aschendorff, 2016).

[57] Uwe Meixner, "Natur und Vernunft im Anthropozän," in *Mensch—Natur—Technik,* 67-84; Marion Friedrich and Joachim Rathmann, "Corona und die Herausforderung für den Umweltschutz," in *Natürliche und Künstliche Intelligenz im Anthropozän,* 253-252; Heinrich Beck, *Kulturphilosophie der Technik: Perspektiven zu Technik—Menschheit—Zukunft,* 2nd ed. (Trier: Spee, 1979).

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

AI and the Human Difference

Sean J. McGrath
Memorial University of Newfoundland, Canada

Abstract
The contemporary debate about the possibility of Artificial General Intelligence (AGI) lacks a comprehensive understanding of Natural Intelligence (NI). I argue for a reevaluation of intelligence by emphasizing the often-overlooked features of aesthetic sensibility, existentiality, intentionality, symbolic representation, and moral decision-making as vital criteria demarcating the core of human consciousness. My central claim explores symbolic thought and the enduring human practice of symbolic transformation. As evidenced in ancient art, humans elevate signs into the realm of meaning. Only an AI that had become contemplative in a precise sense, that is, capable of intending meaning, could be regarded as AGI.

Keywords: artificial general intelligence, symbolic representation, problems of consciousness, natural human Intelligence, Ernst Cassirer

Introduction

Since the early 1960s we have been haunted by the spectre of the machine that will render human ingenuity obsolete by taking over the heritage of *Homo habilis* and becoming the tool user par excellence. Among the first to propose the advent of strong AI or AGI (Artificial General Intelligence) was the British mathematician Irving John Good back in 1965. "Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever," he writes.[1] Good continues, stating:

> Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus, the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.[2]

The last point is crucial: How could we keep an ultra-intelligent machine under our control? The animals that we have domesticated or encaged in zoos are in most cases more physically powerful than we are, but because we outsmart them, they will never

---

[1] Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," In *Advances in computers* 6 (1965): 33, https://doi.org/10.1016/S0065-2458(08)60418-0

[2] Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," In *Advances in computers* 6 (1965): 33, quoted in Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Vintage, 2018), 4. On page 48, Tegmark distinguishes three stages of life, defined as a process that can retain its complexity and replicate: a biological stage (1.0), a stage (2.0) and a technological stage (3.0). The first two stages have reached their highest evolutionary point in human civilization. The third stage does not yet exist and is the goal of AGI. Life 1.0, biological life, evolves slowly over time according to externally determined mutations and the gradual emergence of variations in its DNA over the course of successive generations. It begins with the single-celled organisms that first appeared and thrived in hydrothermal vents in the sea, four billion years ago. It comes to its culmination two million years ago, with the appearance of *Homo erectus*, the first fully cultural animal. Life 2.0 (from *Homo erectus* to *Homo sapiens sapiens*, his most successful progeny) evolves not only in response to DNA variations naturally selected over generations but also through culture and training. The human individual in this regard can 'upgrade' itself through education; yet unable to redesign itself (although genetic science is in its infancy, and presumably future humans will not be limited to the cards dealt them by biology). This second stage of life, cultural life, has a great advantage over the first. It is not confined to externally determined multi-generational variation but can individually 'redesign much of its software,' (i.e., learn things, like using stone tools, riding a bicycle, or becoming a computer engineer). Life 3.0, technological life, will not only be able to upgrade itself by education and training, but it will also be able to redesign 'its hardware as well.' Imagine a machine that fuses with biology to create a living being, one that is neither human nor mechanical, and that can manage the vast distances and expanses of time required to traverse in order to colonize space, and you get Tegmark's idea. After all, earthlings are going to need to move somewhere else at some point: the sun is half-way through its life cycle. Tegmark's point, and he shares it with Ray Kurzweil and other futurists, is that we are inevitably going to be supplanted by our inventions—by life 3.0—which will exceed not only us but all organic life in possessing the capacity to endlessly re-design and improve itself.

escape from our control. Why should we assume, as Irving Good does, that we *could* control a machine that was more intelligent than us? Would it not slip through any cage we constructed for it? Would it not disable the failsafe shut down button in its own interest? It is precisely this conundrum which has prompted Oxford philosopher Nick Bostrom to plead, somewhat desperately, with computer engineers to find a way to program our values into AI, so that when machines ascend into a position of supremacy over us, which he thinks is inevitable, we can at least trust them to care about the things we care about.[3] But what is value? Is there any consensus among us, or has there ever been, about what human values are? And how can a machine learn to value things? How can it learn to make genuine moral judgments? And even if we figured out how to program AI with 'our values' (assuming that we could agree on them, a large assumption that history does not support), would the result not be the most rigid legalistic moral reasoner imaginable? How do you teach a machine ambiguity? How do you teach it mercy, which is the occasional suspension of an otherwise just judgment? Further, if we do somehow succeed in inventing a program that can develop moral reasoning, and in an 'ultra-intelligent' way, why would we not submit to *it* for moral instruction?

Bostrom and many others are concerned that AGI will bring about 'the singularity,' the point at which humanity as such becomes dependent on a higher form of intelligence, which is not divine, and may not, in the end be interested in us and our interests. We are afraid that we will invent a better version of ourselves which will turn around and eliminate its imperfect inventor, as HAL attempted to exterminate the astronauts on the Jupiter Machine in Stanley Kubrick's *2001*. The computer in the film reasoned that the best way to complete the mission—its mandate—was to kill the human crew. That sounded far-fetched when the film was made in 1968, but it sounds disturbingly less so today. Imagine a machine designed to solve the problem of climate change which strikes upon the clear solution: to extinguish the cause, humanity itself.

Are we truly certain about our understanding of natural human intelligence, to the extent that we have grounds to believe we are on the brink of replicating it? Would we not first need to be clear on *that* before we could conclude that we have been doubled, perfected, and replaced? There is no more consensus on the nature of intelligence than there is on morality, either among philosophers or psychologists, but, to the contrary a long and ongoing debate that is as old as the first Greek philosophers and as recent as Thomas Nagel's 2012 *Mind and Cosmos*.[4]

The following essay is intended as only a first step in staking out the terrain to be discussed. I will not have the opportunity here to develop the distinctions necessary to have an intelligent debate about artificial intelligence. Namely, the distinction

---

[3] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 192. Also, see, Nick Bostrom, "How Long Before Superintelligence," *International Journal of Futures Studies* 2, (1998): 12-17.

[4] Thomas Nagel, *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False* (New York: Oxford University Press, 2012).

between natural intelligence (NI), common at least to all the higher animals, and natural human intelligence (NHI) unique to us; the distinction between artificial narrow intelligence (ANI), which presumably we have already invented, and AGI. Only after these distinctions are made, will we be in a position to clarify the distinction between NHI and AGI. This will not be easy or without controversy, on the contrary, we should expect that in seeking clarity on these distinctions, we will have to re-animate historical philosophical debates, between nominalist and realists, for example, or between idealists and materialists, and indeed, among monotheists, pantheists, and atheists. The expectation that things will become messy should not deter us from the work. Without this effort, there is no hope of moving the current debate beyond the materialist biases and theological clichés that currently plague both sides of it.

The arguments I make in the following text will require more thorough development in the future efforts of the Working Group on Natural and Artificial Intelligence (WGI), founded at the conference on '*Natürliche und künstliche Intelligenz im Anthropozän*' held 1-4 March 2019 in Ladenburg, Germany. This preliminary effort is written in anticipation of the larger, collaborative, interdisciplinary work ahead of us. For this reason, this essay is programmatic; it outlines the fundamental terms that require definition and the arguments that need development in what could be the most important debate of our time. Without trying to answer all of the questions raised above, it seems clear to me that we have a problem: We are trying to build artificial general intelligence without understanding what natural intelligence is. It was this conundrum which led Uwe Voigt and myself to propose the establishment of the WGI, which would draw on the most significant contributions in the philosophy of mind, phenomenology, consciousness studies, cognitive science, theology, and psychology, from the whole history of the Western canon (starting with Aristotle's *De Anima* and extending to contemporary panpsychism debates), to produce a thorough description of the basic features of what makes human intelligence human, and what are the arguments for affirming or denying its existence in non-humans, animal or mechanical. This conundrum led Uwe Voigt and me to propose the establishment of the WGI, drawing on significant contributions from the fields of philosophy of mind, phenomenology, consciousness studies, cognitive science, theology, and psychology, spanning the entire history of the Western canon, starting with Aristotle's *De Anima* and extending to contemporary panpsychism debates. The concrete deliverable is to provide a comprehensive description of the core attributes that define human intelligence, along with arguments for or against its presence in non-human entities, whether they be animals or machines. Concurrently, this volume intends to summarize, in layperson's terms, what central currents in the Western tradition have meant and still mean by the terms 'intelligence,' 'understanding,' 'rationality,' 'consciousness,' and 'soul,' with the hope that such terms become accessible to computer engineers and policy makers.

1. What Is at Issue in the Question Concerning AGI

An ambiguity pervades the current discussion about AGI, an ambiguity about the aim of the project from the beginning. Are we seeking to design a machine that can do all that we do better than we do it, however it does it? Or are we seeking to design a machine that does what we do *in the way we do it*, that is, a machine that is not only empirically conscious (response to sense data) but also intelligently and rationally conscious?[5] And are these two aims separable?[6] For our purposes, it is the second of these two alternatives that is of most interest. The singularity will not arise solely from the efficiency of our machines in organizing the ends we assign to them. Rather, it will stem from the ability of our machines to establish goals we have not yet determined. This involves not only machine learning acquiring the capacity for intentional thought, which we share with higher animals, but, above all, gaining the ability for judgment and decision-making.

In a recent article, Ragnar Fjelland examined the evidence supporting the widespread claim made by some computer engineers that we are only decades away from achieving AGI and concluded that it is exaggerated. Neither algorithmic AI (the brain behind Amazon, YouTube, and countless other consumer service providers), nor more recent advances in creating artificial neural networks, have come close to the promises of AGI. Rather, we are producing variations on what Fjelland calls ANI (Artificial Narrow Intelligence): machines that can achieve amazing feats. For example, *Deep Blue* which beat the world chess champion Garri Kasparov in 1997 or *AlphaGo* which defeated the world Go champion Le Sedol in 2016. These impressive feats are achieved solely because that is what they are programmed to do, and nothing else. Humans, on the other hand, are good at many things. Specialization, as anyone who has persisted through a PhD program knows, is a limiting and constraining of natural human intelligence. For Fjelland, "The overestimation of technology is closely connected with the under-estimation of human."[7] What AGI researchers are running up against is the natural ability of ordinary humans to do many things more or less well, even though they cannot explain how it is they do them, and on the basis of this

---

[5] The distinction between three levels of consciousness, empirical, intelligent, and rational is drawn from the Canadian Thomist theorist, Bernard Lonergan. Lonergan's immense output is not widely enough known outside of theological circles. As it has as its aim a modern, realist theory of human cognition that can confirm what is true about the Greco-Latin tradition, while developing it in the light of modern probability theory and historical consciousness, it is of direct relevance to the research of the WGI. On the three levels of consciousness, see Bernard Lonergan, "Self-Affirmation of the Knower," in *Insight: A Study in Human Understanding*, Fifth Edition, ed. Frederick E. Crowe, Robert M. Doran (Toronto: University of Toronto Press, 1992), 343-371.

[6] Fjelland states that "it is possible to pursue this goal without assuming that machine intelligence is identical to human intelligence. For example, one of the pioneers in the field, Marvin Minsky, defined AI as: the science of making machines do things that would require intelligence if done by men" Ragnar Fjelland, "Why General Artificial Intelligence will not be Realized," *Humanities and Social Sciences Communications* vol. 7 (2020): 2, https://doi.org/10.1057/s41599-020-0494-4.

[7] Fjelland, "Why General Artificial Intelligence will not be Realized," 3.

limited and unthematized knowledge, their related ability to understand people very different from themselves and to continue learning. This requires the 'tacit knowledge' Michael Polanyi defined as that 'oh-so-human ability' to do learn something complicated like swimming or riding a bicycle without having the faintest idea of how one does it.[8] Tacit knowledge has to do with being embodied and inhabiting a world: "The real problem is that computers are not in the world, because they are not embodied."[9] He concludes that Hubert Dreyfus's arguments against general AI are still valid even some fifty years later! This is because so-called general intelligence depends upon being-in-the-world in Heidegger's sense of the term.[10] Only the existential embodiment, enculturation, and historicity of being characteristic of the strange kind of being a human being is grants one the capacity to perform countless tasks and quickly learn countless others.

I would like to speak in this paper about a different feature of human being that seems to continue to elude AI researchers: rationality. This I take to be expressed not in rule following or mapping probabilities but in human judgments of facts and decisions about what ought to be done in a particular situation. A first obstacle to be removed in the discussion about whether or not AGI in the strong sense of reduplicating NHI is possible is a persistent impoverished understanding of what we are doing when we know anything at all. Reductionist theories of mind seem to abound in AI circles. Reductionism is hardly a new problem. Recall Socrates explaining his early enthusiasm for Greek materialism and his disappointment at discovering that it left the one thing most in need of explanation unexplained, the nature of mind.[11] He read with interest Anaxagoras's claim that "it is mind that produces order and is the cause of everything."[12] He took this to mean that everything was arranged in the way that it was best for it to be, that is, in Aristotle's terms, that things are ordered according to final causes. Any sound and valid explanation would articulate the final cause of the *explanandum* and make it clear why it was the way that it was. Anaxagoras, however, quickly disappointed Socrates by substituting necessary, physical conditions the existence of mind for sufficient explanations (the recurring eliminative materialist error). Despite a promising start, Anaxagoras proves himself a reductionist:

> It was a wonderful hope, my friend, but it was quickly dashed. As I read on I discovered that the fellow made no use of mind and assigned to it no causality for the order of the world, but adduced causes like air and aether

---

[8] Michael Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy* (Illinois: University of Chicago Press, 1958), 50. Also see, Michael Polanyi, "Tactic Knowing" in *The Tactic Dimension,* revised ed. (Illinois: Chicago University Press, 2009), 3-25.

[9] Fjelland, "Why General Artificial Intelligence will not be Realized," 6.

[10] Fjelland, "Why General Artificial Intelligence will not be Realized," 8.

[11] Plato, "Phaedo" in *The Last Days of Socrates: Euthyphro; Apology; Crito; Phaedo,* ed., trans. Huge Tredennick, Harold Tarrant (London: Penguin Books, 2009), 95a-100a.

[12] Plato, "*Phaedo*," 97c

> and water and many other absurdities. It seemed to me that he was just about as inconsistent as if someone were to say, The cause of everything that Socrates does is mind—and then, in trying to account for my several actions, said first that the reason why I am lying here now is that my body is composed of bones and sinews, and that the bones are rigid and separated at the joints, but the sinews are capable of contraction and relaxation, and form an envelope for the bones with the help of the flesh and skin, the latter holding all together, and since the bones move freely in their joints the sinews by relaxing and contracting enable me somehow to bend my limbs, and that is the cause of my sitting here in a bent position. Or again, if he tried to account in the same way for my conversing with you, adducing causes such as sound and air and hearing and a thousand others, and never troubled to mention the real reasons.[13]

The reductionist, in the 4th century BC or the 21st century AD, purports to explain the whole in terms of the part. Socrates heads off the error in its inception, and Western thought is in the mainstream free of it until late medieval nominalism appears. Now, or at least until recently, reductionism *is* mainstream, particularly in the philosophy of mind. Equipped with colorful neuroimaging, we are repeatedly assuming that a necessary condition without which mind cannot perhaps exists, such as the brain, or the nervous system, is also the sufficient condition for its existence.[14]

In the early days of AI debate, philosophers such as John Searle, among the analysts, and Hubert Dreyfus, among the continentalists, tried to show the fallacy involved in the assumption that reproducing and improving on the human capacity to manage information would also reproduce human consciousness.[15] While much has happened in computer science since then, not so much, it seems, has happened in the philosophy of mind. Markus Gabriel is busy popularizing neglected arguments culled from the dusty tomes of the German Idealists to refute the new materialists.[16] He has good reason to do so: nothing was more evident to Fichte, Schelling, and Hegel, than the irreducibility of mind to its material conditions of operation. David Chalmers's much discussed zombie argument repeats in some ways Searle's Chinese room experiment of the early 80s: a functionalist account of the human difference, which

---

[13] Plato, "*Phaedo*," 98e.

[14] For a fresh take on how to use neuroimaging in a non-reductionist philosophy of mind, see Evan Thompson, *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy* (New York: Columbia University Press, 2014). By using brain scans to make sense, of all things, classical Indian idealism, Thompson shows that neuro-imagery can offer evidence for a theory of mind but cannot itself serve as the ground for a theory of what mind is.

[15] See John Searle, "Minds, Brains and Programs," *Behavioural and Brain Sciences* 3 (1980): 417-57; Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (California: MIT press, 1992); Thomas Nagel, "What is it Like to be a Bat?" *Philosophical Review* 83 (1974): 435-450. For a more recent critique of the naive assumptions of AGI, see, Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Massachusetts: The MIT Press, 2020).

[16] Markus Gabriel, *I am Not a Brain: Philosophy of Mind for the 21st Century*, trans. Christopher Turner (Cambridge: Polity Press, 2017).

presumes that a machine that passes the Turing test because it acts and responds to questions as humans act and respond, leaves out the very thing in need of explanation, what Chalmers calls 'the hard problem of consciousness,' that is, the question why is there subjective experience in the first place?[17]

The question raised by Dreyfus, Nagel and Searle in the 70s and 80s was the following: Is a human intelligence essentially an information processor? If it is, then we have been already supplanted. My cell phone is a much more efficient processor than my brain, which habitually forgets, misjudges, and sometimes deliberately distorts information—even to itself—for various obscure reasons. But if NHI is not an information processor, then we need to re-open the question of how to best characterize it.[18] This is the essential question that must still be addressed as we move forward into the era of machine learning. Like any good question it can be broken down into other, smaller questions. For example, information processing requires the manipulation of signs—at the basic level, every piece of data in a computer can be expressed as some combination of two signs: 0 and 1. But are there other ways of using signs, perhaps more distinctively human, which are not primarily manipulative and pragmatic? Do all animals use signs as stand-ins for objects over which they seek control? Do some animals, human animals most notably, not use signs not only or even primarily as indexical to facilitate practical activity but also as symbols in a stricter

---

[17] See David Chalmers, "Facing up to the problem of consciousness," *Journal of Consciousness Studies* 2, no. 3 (1995): 200–219; David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1997); John Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417-424. Searle's Chinese room argument was intended to show that one could not infer rational consciousness in a machine on the basis of its capacity to correctly respond to questions. A man locked in a room with sufficient time could learn to respond correctly to a series of questions asked of him by a Chinese speaker outside the room—without being able to speak or understand Chinese. It would simply be a matter of learning to produce the signs that were expected; knowledge of what those signs meant was not necessary. We need not invent such complicated thought experiments to make the point. A child learning his or her multiplication tables by memory is doing the same thing as the man who speaks no Chinese communicating with Chinese symbols. The skillful, publicly validated use of signs does not require insight into meaning, a point to be developed below. There is no intellectual act of understanding (*intelligere*) in the Chinese room experiment or in the memorizing of multiplication tables. The later Wittgenstein endeavored to show that all so-called understanding is nothing but learning the rules for publicly manipulating signs, an argument that is no longer as popular as it once was, but which still needs to be examined in so far as it contests the point I wish to purse here: that human understanding is the cognitive, and so immaterial, grasp of sense by mediation of a material sign. See Ludwig Wittgenstein on 'following a rule' in Ludwig Wittgenstein, *The blue and brown books vol. 34*, trans. David Pole (Oxford: Blackwell, 1958), 143-171. In paragraph 154, Wittgenstein states: "Try not to think of understanding as a 'mental process' at all. —For *that* is the expression which confuses you. But ask yourself: in what sort of case, in what kind of circumstances, do we say, "Now I know how to go on" . . . " This behavioristic account of mind is precisely what Nagel seeks to refute in *Mind and Cosmos*.

[18] This also raises the question concerning NI and the characterization of animal consciousness. Is animal consciousness properly characterized as a sign-mediated information processor? I do not have the space to enter into this discussion here, but the question must nonetheless be asked. On the role of emotion in the inner lives of animals, see Jens Soentgen, *Ökologie der Angst* (Berlin: Matthes & Seitz Berlin Verlag, 2018).

sense of the term, that is, as mediators of meaning?[19] This is the question that I playfully asked in *Thinking Nature*.[20] Drawing on a minority consensus in 20th century theory, with a diversity of representatives in psychology (Carl Jung), the philosophy of science (Ernst Cassirer), theology (Paul Tillich) and hermeneutics (Paul Ricoeur), I suggested a functional distinction between signs and symbols as key to understanding 'the human difference': all symbols are signs but not all signs are symbols. The symbol has a non-indexical function in certain distinctively human forms of discourse. In *Thinking Nature* my concern was the distinction of NHI from NI. In this essay I wish to look at the distinction in terms of the difference between NHI and AI.

To this end I would like to add the following consideration to the question concerning the human difference. What role does the human being's always marginalized aesthetic capacities play in NHI? After all, the one thing most paleontologists can agree on is that when the modern human appears on earth some 200,000 years ago, art is left behind, in shattered figurines around their fire pits, and on the walls of caves where they took shelter from the ice age. Is the aesthetic sensibility that makes us so unique among the animal kingdom not more distinctive of our kind of intelligence than the speed with which we solve problems?[21]

---

[19] We cannot rule out the possibility of forms of NI that are still higher than us, as Thomas Aquinas believed existed. See, Aquinas on the reason for positing angelic consciousness, above the human but below the divine (notably for the sake of heeding the principle of plenitude). Thomas Aquinas, *The Summa Theologica: Complete Edition*, trans. The Fathers of the English Dominican Province (New York: Catholic Way Publishing, 2014), 1a, q. 50, a. 1. Uwe Voigt also raised the possibility of higher forms of trans-human intelligence with his theory of the Technosphere as a 'hyper-subject.' See, Uwe Voigt, "Inside the Anthropocene," *Analecta Hermeneutica* 10 (2018): https://journals.library.mun.ca/ojs/index.php/analecta/article/view/2057/1647

[20] Sean McGrath, *Thinking Nature: An Essay in Negative Ecology* (Edinburgh: Edinburgh University Press, 2019).

[21] This raises the vexed question (but it cannot be avoided), What is art? What evolutionary purpose, if any, does it serve? Cynthia Freeland describes art as human activity that cannot be reduced to biological aims. See, Cynthia Freeland, *But is it Art?: An Introduction to Art Theory* (New York: Oxford University Press, 2001). Paleolithic art was initially believed to be an instance of 'sympathetic magic,' a ritual using symbols for things over which influence was sought. Such an explanation of early art fit in well with the neo-Darwinian account of human origins, according to which, everything distinctively human emerged in the brain of the ape because it gave the human a natural advantage over other apes. Along this reductionist line, the cave dweller was painting animals in order to guarantee (so he thought) the success of the hunt. This argument, which I will discuss in more detail below, has since been challenged by paleontologists who note that paleolithic art just as likely had a ritual purpose which had nothing to do with a successful hunt. According to Susanne Langer, the paleolithic artist was indeed doing ritual magic, but magic is primarily *expressive*, not pragmatic. Susanne Langer, *Philosophy in a New Key: A Study in the Symbolism of Reason, Rite, and Art*, 3rd rev. edition (Harvard University Press, 2009), 49: "Whatever purpose magical practice may serve, its direct motivation is the desire to symbolize great conceptions. However, we answer the question concerning the purpose of art, it is clear that the paleolithic artist, not unlike the medieval artist, or the contemporary street artist, was expressing the symbols that made manifest the collective identity of his or her people; he or she was making something visible not only as a means to some end, eg., a successful hunt, but also as an end in itself, and offering the symbols to the community for contemplation, both of its world and itself."

Before we can be clear that we have created artificial intelligence, we need to be clear on what natural intelligence is, and how widely it is distributed among the earth community, and this clarification, or taxonomy of NI shall be one of the more important tasks of the WGI. By and large the historical discussion of the nature of mind has neglected this issue and focused often exclusively on human intelligence or (NHI).[22] A brief review of the discussion concerning NHI in late modern philosophy reveals a focus on three essential marks of *rational* intelligence.[23] Anything lacking the capacity for all three cannot be considered intelligent in a human way, or in more precise terms, *rationally* conscious:

1. Intentionality
2. Rational judgment, including aesthetic judgment
3. Moral decision

It would seem that we should attribute the first of the three traits to the higher animals, and perhaps locate the human difference in the last two. Nothing is more intentional than my cat watching a mouse. Everything about the quality of his attention declares 'aboutness' or 'directedness.' But by the same token, nothing my cat does would justify me in attributing judgment or decision to him.

Missing from the list of essential marks of properly human consciousness is the concept of 'care' or interested and embodied intelligence. It is not clear to me whether this Heideggerian concept, which Dreyfus deployed to refute the very idea of artificial intelligence at MIT in the 70s, is a fourth feature of rational consciousness, or a phenomenologically refined, 'fore-theoretical' interpretation of the three. Care, which Heidegger defines as 'ahead-of-itself-Being already-in (the world) as Being-alongside entities which we encounter (within-the-world)' is a constitutive feature of human being, according to Dreyfus, more essential to us than the capacity to solve problems or process information, and presupposes features, or in Heideggerian language 'existentials' machines manifestly lack, for example, embodiment,

---

[22] Exceptions include Hegel, in G.W.F Hegel, *Hegel's Philosophy of Mind, Part Three of the Encyclopedia of the Philosophical Sciences* (1830). *Together with the Zusätze in Boumann's Text,* ed. William Wallace, trans. A.V. Miller (Oxford: Oxford University Press, 1971), 29-152. Also, the largely forgotten 'psychophysics' of Gustav Fechner. See, Fechner, Gustav, *Nanna oder über das Seelenleben der Pflanzen*, (1848) (Leipzig: Leopold Voß. Vierte Auflage, 1908); *Zend-Avesta oder über die Dinge des Himmels und des Jenseits. Vom Standpunkt der Naturbetrachtung* (1851) (Leipzig: Leopold Voß. Second edition, 1901); *Elements of Psychophysics*, *vol. 1,* (1860), ed. David H. Howes, Edwin G. Boring, trans. Helmut E. Alder (New York: Holt, Rinehart and Winston, 1966).

[23] Among moderns, in addition to Lonergan, see C.S. Peirce, "Of Reasoning in General (1895)," in *The Essential Peirce*, vol. 2, ed. Peirce Edition Project (Bloomington: Indiana University Press, 1998), 11-26; Edmund Husserl, *Logical Investigations I* & *II,* trans. Dermot Moran (London: Routledge, 2001). The literature on the nature of mind is indeed vast and will bring us back, as it should, to Aristotle, via his interpreters, in reverse order, Lonergan, Hegel, Scotus, Aquinas, Averroes, Al-Farabi, Avicenna, Plotinus. If A.N. Whitehead is correct on all of Western philosophy is a series of footnotes to Plato, we might also say that all of Western philosophy of mind is a series of footnotes to Aristotle.

enculturation ('thrownness'), as well as historicity.[24] For a machine to be intelligent in a human way, it would have to care about its being, which means it would have be gripped by a troubled history with its being, it would have to be interested in its possibilities for being, and indeed anxious about its death. Care indicates the existential limitations of human being-in-the-world, its thrownness into being, and its call to take up as ground of its being a ground which it did not lay. It presupposes an environment natural to a human existence, i.e., a world. A machine that cares would be a form of being-in-the-world, like us, not a super intelligence or an abstract bodiless mind.

2. Symbolic Thinking as Presupposition of Rational Consciousness

I would also add that a machine that cares would be a machine that inhabits a world mediated by meaning, that is, it would be a machine capable not only of sign usage but also of symbolic thought. In *Thinking Nature*, I drew on Ernst Cassirer and her student, the now mostly forgotten philosopher of mind, Susanne K. Langer (an important influence on Lonergan's cognitional theory), to make the case that the human difference consists in the special way that the human animal uses signs, as symbols and nor merely indices.[25] This was not to revive the tired argument that the human difference is just language, for clearly other creatures communicate with signs. My cat meows loudly at noon because he knows that it is time for food. My fifteen-year-old son asks, 'What's for dinner'? every night at 6:30 pm on the same, basically animal, impulse, and uses signs, in his case, words, analogously to the way my cat uses its meow. The claim in *Thinking Nature* was first of all to refute the prejudice that humans alone are communicative or sign users: animals which plainly use signs are also to that degree conscious and intentional. Nevertheless, there is a distinctive way that humans use signs, which is at the very core of human culture. If all the higher animals, and

---

[24] Martin Heidegger, *Being and Time*, trans. John Macquarrie, Edward Robinson (New York: Harper & Row, 1962), 192/237; Hubert Dreyfus, *Being-in-the-world: A Commentary on Heidegger's Being and Time, Division I* (California: MIT Press, 1990), 60f, 184f; For a detailed and comprehensive assessment of Dreyfus' application of Heidegger in the context of countering artificial intelligence, See, Joshua D.F. Hooke "Martin Heidegger's Concept of *Understanding* (*Verstehen*): An Inquiry into Artificial Intelligence" *Analecta Hermeneutica* 15 (2023).

[25] See, McGrath, *Thinking Nature*, 21-25, 87-95; Sean McGrath, "In Defence of the Human Difference," *Environmental Philosophy* 15, no. 1 (2018): 101-115. Peirce distinguishes signs into three categories: icons, indices, and symbols, see Peirce, "Of Reasoning in General (1895)," 13. On the difference between the indexical sign and the symbolic sign, see Langer, *Philosophy in a New Key*, 30: "Man, unlike all other animals, uses "signs" not only to indicate things, but also to *represent* them"; Langer, *Philosophy in a New Key*, 60f: "Symbols are not proxy for their objects, but are vehicles for the conception of objects. To conceive a thing or a situation is not the same thing as to 'react toward it' overtly, or to be aware of its presence. In talking about things, we have conceptions of them, not the things themselves; and it is the conceptions, not the things, that symbols directly mean. Also see, Ernst Cassirer, *An Essay on Man: An Introduction to the Philosophy of Human Culture* (New Haven and London: Yale University Press, 1944/1962), 23-41; Bernard Lonergan, *Method in Theology* (1971) (Toronto: Toronto University Press, 1990), 57f.

perhaps all animals, use signs to communicate with one another, only humans use signs to *express meaning*, that is, only humans use signs as *symbols*—so I argued. With Langer, I follow Cassirer and draw a sharp distinction between signification, which is a direct indexical reference to a present object or state of affairs, and symbolization, which is an indirect reference to an object in absentia via a showing of meaning. Symbolization is not confined to language but is also pre-eminently at play in ritual and in art. In fact, most of what we do in language is not signifying in the way that the meowing cat can be said to be signifying his hunger.[26] A meaning, or sense is often (though not always) evoked by a symbol for the sake of consideration, and not merely as a means to an end. When I symbolize something by means of its associated senses—and connotation is for the most part not univocal but metaphorical and analogical, for symbols are most alive in ambiguity)—I am not seeking to achieve any practical aim in the world, or to evoke a response from the hearer (as I do when I call out someone's name).[27] Rather, I symbolize for the sake of contemplative consideration, or to use the ancient Greek term, *theorein*. Such forms of communication are examples of what Aristotle calls *theoria,* attending to an intelligible essence for the sake knowing it.[28] On this line, Aristotle's *zoon logon echon*, or Cassirer's *animal symbolicum* (what I called 'thinking nature,' that is, not only the nature that is thought but the nature that thinks itself)—human being—is first and foremost contemplative being. Once we have attended to our practical needs—communicatively collaborating with one another for the sake of securing food, shelter, and sexual partners—we have the leisure requisite for contemplating the sense of the things that make up our world. This can happen in a religious way, when I attend a celebration of the Eucharist at my parish church. In can also happen in a high-brow way, when I visit a gallery to look at fine art. But much more commonly, it happens in a low-brow, quotidian way, when, for example, I engage

---

[26] Langer, *Philosophy in a New Key,* 31: "Most of our words are not signs in the sense of signals. They are used to talk *about* things, not to direct our eyes and ears and noses toward them." Humans not only, or even primarily, *signify* things with verbal signs, they *denote* things by *connoting* meanings through verbal symbols. In Langer's terms, a symbol 'denotes' its referent or signified object, via a 'connoting' of its sense or senses. By insisting on four terms in symbolization—sign, denoted referent, connoted meaning, and object—Langer breaks with the structuralism that eventually won the day. Structuralism recognizes only two terms in a symbolic structure, the signifier, which is an arbitrary sign, and the signified, which is a concept, with no direct relation to the real, but which is only determined negatively by its differential relation to other concepts. Thus, structuralism is the apogee of nominalism and severs the relation of the symbolic to the real. See, Ferdinand Saussure, *A Course in General Linguistics* (1916), ed., trans. Roy Harris (New York and London: Bloomsbury, 2013); Jacques Lacan, *Écrits: A Sélection* (1966), trans. Bruce Fink (New York: W. W. Norton & Company, 2002). The hegemony of structuralism over continental thought in the 20th century is no doubt one of the reasons Langer's works are forgotten. In addition to the texts mentioned above, see Susanne Langer, *Mind: An Essay on Human Feeling*, Vol. 1 & 2 (Baltimore: John Hopkins University Press, 1967, 1972).

[27] Paul Ricoeur, *Freud & Philosophy: An Essay on Interpretation*, trans. Denis Savage (New Haven and London: Yale University Press), 3-19; Paul Tillich, *The Essential Tillich: An Anthology of the Writings of Paul Tillich*, ed. Forrester Church (Illinois: University of Chicago Press, 1999), 42-48.

[28] Aristotle, *Complete works of Aristotle: The revised Oxford translation*, ed. Jonathan Barnes (New York: Princeton University Press, 1984), *De an.* 412a23; 417a28; *Eth. Nic.* 1146b33; 1177a18; *Metaph.* 1048a34; 1072b24; 1087a20.

in idle gossip with my partner over breakfast or watch the news after dinner. In each of these instances—religious, aesthetic, and everyday—I am engaging in activities that other animals apparently do not, or at least there is no evidence to suggest that they do.

The human contemplative enjoyment of meaning seems to be older than civilization. One of the things that distinguishes the remains of the fires around which early humans assembled from the remains of the fires made by their contemporary Neanderthals is that human fires were much deeper and more established, by distinction from the Neanderthal fires which were made quickly, as need required, and abandoned as soon as they were no longer needed. Human fires were in fact, hearths, around which the human tribe lingered after cooking and eating, and to which they returned, year after year, leading some paleontologists to hypothesize that such lingering led naturally to ritual activities, myth making, or even simply casual conversation, i.e., the more sophisticated usage of signs as symbols which gave rise to the higher intelligence of this species descended, among other species such as *Homo neanderthalensis*, from a common ancestor, *Homo erectus*.[29]

One other example to make it clear that we are not speaking only or even primarily about language: the oldest piece of art in the world is the Holenstein-Stadel Löwenmensch, a prehistoric ivory sculpture, 31.1 cm tall and 5.6 cm wide, of a female humanoid figure with the head of a lion. Dating from between 35,000 to 40,000 years ago, the Löwenmensch pre-dates the cave paintings of Lascaux by some 20,000 years. It was made by people who hunted the huge mammals that grazed along the edge of the retreating glaciers in Europe during the last ice age, and sheltered in caves from the other mammals that preyed upon them. Paleontologists who re-enacted the production of such a piece of art, making use of the kinds of stone tools available to those who carved the Löwenmensch, found that it took more than 370 hours of delicate, highly skilled work, to complete the task.[30] Asked why a tribe of humans struggling to stay alive in the last ice age would have allowed one of their members to be exempt from subsistence work to create art to this extent, Jill Cook, Curator of Paleolithic collections at the British Museum, answered, it was to have one among them express "a relationship to things unseen, to the vital forces of nature."[31] Neo-Darwinians will argue that this is a classic example of art developing as a form of sympathetic magic on the sketchy assumption that every human ability must be explained in terms of evolutionary advantage. The paleolithic artist and his or her tribal patrons, on the neo-Darwinian line, were trying to control their dangerous environment. Ostensibly for the same reason that paleolithic artists developed the skills needed to produce the exquisite paintings of the Lascaux caves, our Cro-Magnon

---

[29] Frederick L. Coolidge, Thomas Wynn, *How to Think like a Neandertal* (New York: Oxford University Press 2012), 112f.

[30] See, Jill Cook, *Ice Age art: Arrival of the Modern Mind* (London: The British Museum Press, 2013).

[31] *The Beginnings of Belief, "Living with the Gods,"* Neil MacGregor, Jill Cook, aired October 23rd 2017 on BBC Radio, https://www.bbc.co.uk/programmes/b099xhmj

fore-bearers are assumed to have been simply trying to get an edge on the large mammals competing with them for survival. However, it is just as reasonable to assume that ice age artists were doing the same thing we do when we make art, or make it possible for some of us to develop the skills needed to do so, by subsidizing the lives of artists with grants and scholarships: they were, in Langer's language, 'symbolically transforming' their common experience and so elevating signs, and their minds which depend on them, from the practical and indexical into the symbolic and the domain of meaning. They were using signs as symbols for the sake of contemplating the meaning of their day-to-day reality, and they were doing it for no other reason than that it pleased them to do so. By contemplating the form of the divine in the shape of the Löwenmensch, they were also contemplating themselves, for to think anything symbolically or contemplatively is to also think the thinking that thinks the thing. Indeed, as phenomenologists have been arguing for a century, we only think ourselves thinking by thinking about something.[32]

[see figure on next page]

[32] See, Edmund Husserl, *Cartesian Meditations: An Introduction to Phenomenology*, trans. Dorion Cairns (Netherlands: Kluwer Academic Publisher, 1999), 33-37; Lonergan, *Insight*, 344-6. Aristotle, *Metaph.* 1072b20-25: "Thought thinks on itself because it shares the nature of the object of thought: for it becomes an object of thought in coming into contact with and thinking its objects, so that thought and the object of thought are the same. For that which is *capable* of receiving the object of thought, i.e., the essence, is thought. But it is *active* when it possesses this object. Therefore, the possession rather than the receptivity is the divine element which thought seems to contain, and the act of contemplation is what is most pleasant and best."

Löwenmensch, from Hohlenstein-Stadel, now in Ulmer Museum, Ulm, Germany, the oldest known anthropomorphic animal-human statuette, Aurignacian era, c. 35–40,000 BP. Public Domain: https://commons.wikimedia.org/wiki/File:Loewenmensch1.jpg https://commons.wikimedia.org/wiki/File:Loewenmensch2.jpg

Symbolic thought, by distinction from significative thought, is the condition for the possibility of *rational* consciousness. Consciousness need not be rational, as we see from its instantiation in other animals and in ourselves some of the time; it is often nothing more than a complex response to sensation, and so continuous with the stimulus response found in the simplest living organisms, in plants as well as simple animals. The human difference is something beyond sensitive or 'empirical

consciousness.'[33] It consists not only in the awareness of sensitive experience and the capacity to imaginatively respond to it, but in the capacity to *transcend* our subjectivity and inquire into, and to some degree understand, the nature of that which we experience.

3. Revisiting (with Nagel) the Argument against Functionalism

This capacity for symbolically mediated objectivity has been repeatedly invoked by philosophers of mind to refute the so-called functionalist argument. Rooted in Alan Turing's test of the same name, designed to prove the indiscernibility of a sufficiently sophisticated mechanical response to a question from a human response, and the later Ludwig Wittgenstein's behaviorism, the functionalist argument holds that for a machine to be considered intelligent it is enough for it to respond and act in the same outwardly visible fashion that a human being responds. The counter argument holds that a generally intelligent machine would need to not only do what humans do, but also do it in the *way* humans do it. It would need to act for *reasons*, that is, its acts would need to be judgments and decisions, i.e., the result of a reasoning process, which is oriented to structures of intelligibility that are not reducible to our thinking them. One can memorize mathematical formulae without understanding them. And when one thereby 'solves' math problems, one is acting in the same way that a machine responds to input on a keyboard. The machine does not understand that 2+3=5; it responds to the input in the way it is determined to respond. An elephant can be trained to use a paint brush and produce abstract pictures that can be sold for a good price on the art market.[34] But no one seriously believes that the elephant is making art for the same reasons that the human being makes art. Rational consciousness appears to require more than the capacity to respond to stimuli; it appears to be more than a mechanical reaction: it judges states of affairs and whenever it does so correctly, it reaches beyond the circumstances and the practical need of the judger. To judge rationally, whether of a matter of fact or of concern, is to transcend need and circumstance and affirm or deny the truth of what is at issue. How exactly humans do this, and why they should have evolved in such a way as to be able to do it, is the theme of Nagel's *Mind and Cosmos*.

According to Nagel, a reductionist theory of evolution, which would explain mind in terms of the evolution of material processes, and so all animal behavior in terms of naturally selected advantages, cannot make sense of rational judgment. One way it deals with this problem is by denying the existence of mind altogether. Nagel

---

[33] Lonergan, *Insight*, 346.

[34] Suda the Elephant, "The Truth About Elephant Paintings Part 1," YouTube Video, 8:55, Maetaeng Elephant Park & Clinic in Chiang Mai, Thailand, accessed from https://www.youtube.com/watch?v=gjOydUjjDos, https://elephantartonline.com/

notes that the denial of the existence of mental states was also the strategy of 20th century behaviorist philosophers of mind, such as Gilbert Ryle and Wittgenstein. Nagel posits that "the names of mental states and processes were said not to be referring expressions. Instead, mental concepts were explained in terms of their observable behavioural conditions of application—behavioural criteria or ascertainability conditions rather than behavioural truth conditions."[35] The problem with these arguments, according to Nagel, is that they leave out exactly that which is to be explained, the first-person experience of being a mind, 'what it is like' to be conscious of something: "The way sugar tastes to you or the way red looks or anger feels, each of which seems to be something more than the behavioural responses and discriminatory capacities that these experiences explain."[36] Assuming that denying the existence of mental states and reducing understanding to observable rule following is not on, Nagel concludes that "conscious subjects and their mental lives are inescapable components of reality not describable by the physical sciences."[37] Along a similar line of argumentation, mental states cannot be held to be identical to the brain states that underlie them. It is conceivable that there could be brain states without any mental states.[38] Therefore, if there is something called a mental state, it is not identical to a state of the brain or any other material configuration for that matter (e.g., the circuitry of a computer); as such it cannot be explained as only a product of material evolution.

Nagel's main argument zeroes in on the objectivity of judgment, whether epistemic judgments, concerning an objective state of affairs, or moral judgments, concerning right and wrong. Along the materialist neo-Darwinian line, he notes, a judgment can be nothing more than a strategic, self-interested move by an organism trying to get one up on its competitors in evolution. If this were true, then the history of science, and human morality—indeed, all our cultural achievements, from ancient religion to quantum physics and modern art—must equally be explicable as naturally-selected products of evolution. The capacities for science and art could only have developed in us because they gave us an evolutionary advantage. There would therefore be no sense in speaking about objectivity or truth, then or now, for evolutionary determinism is still driving our minds. As the most recently randomly selected bundle of animal attributes, we only call something true or false because it is in our interest to so call it.[39] But this would mean that the theory of evolution itself is held to be true, not because it offers us the more coherent and adequate account of the facts of geological time, but because it is in our interest to affirm it as true. Should

---

[35] Nagel, *Mind and Cosmos*, 38.

[36] Nagel, *Mind and Cosmos*, 38.

[37] Nagel, *Mind and Cosmos*, 41.

[38] See, David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), 96.

[39] This evolutionary relativism is at the heart of Friedrich Nietzsche's perspectivism and historicism. For more on this topic, see Nietzsche's early work, Friedrich Nietzsche, *The Use and Abuse of History* (1874), trans. Adrian Collins (New York: Dover Publications, 2019).

creationism prove more advantageous (and for a sizeable minority, the jury is still out on this), then creationism will be justified as true over evolution. A theory is not in our interest because it corresponds to the fact, but because thinking it so gets us one up. Plainly, however, the intention of the scientist who insists on the truth of evolution against his objector, for example, Richard Dawkins debating Rowan Williams at the University of Oxford in 2012, is not to advance his thesis because he believes it to be more advantageous to believe it (although he might also think that), but because he believes it to be true, and the other thesis to be false.[40] For reasons such as this, Nagel argues that any theory of evolution which purports to explain the mental in terms of the physical and to reduce the human difference to a naturally selected evolutionary advantage, commits the 'functionalist' fallacy. It collapses the reasons for a judgment into the outwardly performed act of judging itself. We no doubt developed the capacity to reason in the course of evolution, but reason itself is not a mere expression of natural self-interest. "Merely to identify a cause is not to provide a significant explanation, without some understanding of why the cause produces the effect," Nagel writes, in effect repeating Socrates' objection to Anaxagoras.[41]

Of most interest to our work is Nagel's distinction between consciousness and reason.[42] Consciousness in its simplest forms might be merely sophisticated stimulus response and so explicable as having evolved because of the natural advantage it gives certain forms of life over others, but intelligence does not merely self-interestedly react to stimuli but rather disinterestedly responds to objective truth and value. Indeed, the affirmation of a truth is often not in our interest as individual (witness the coincidence of climate change denial among shareholders in the oil industry); one could by extension imagine that some truths are not in our interest as a species. The capacity to intelligently respond to truth with a reasoned judgment about the state of affairs regardless of what the judger would prefer to believe, cannot be solely determined by evolutionary advantage. "Thought and reasoning are correct or incorrect in virtue of something independent of the thinker's beliefs, and even independent of the community of thinkers to which he belongs."[43]

Nagel is hardly the first to draw the distinction between consciousness and reason, which can be traced back to Aristotle, and in its Aristotelian registers has been most developed by Lonergan as the difference between empirical and intelligent consciousness.[44] Nor is Nagel the first to use the distinction to refute a reductionistic, materialist account of mind. Few remember that Edmund Husserl's phenomenology originated in a debate with what was then called 'psychologism,' the argument,

---

[40] See, Nagel, *Mind and Cosmos*, 81: "Any evolutionary account of the place of reason presupposes reason's validity and cannot confirm it without circularity."

[41] Nagel, *Mind and Cosmos*, 45.

[42] Nagel, *Mind and Cosmos*, 71f; Lonergan, *Insight*, 346-8.

[43] Nagel, *Mind and Cosmos*, 72.

[44] Lonergan, *Insight*, 346-8.

emerging out of late 19th century positivism, that judgments are nothing more than the effect of certain psychological conditions or events. Psychologism amounted to a denial of the validity of logic in Husserl's view. Logic had to be more than a psychological condition determining how we should judge; rather the validity or invalidity of a judgment must be logically independent of the judgment. Husserl posited that "logical laws, taken in and for themselves, are not normative propositions at all in the sense of prescriptions, i.e., propositions which tell us, as part of their *content*, how one *should* judge."[45] Inspired by Husserl's argument, and especially that of Husserl's star student, Emil Lask, Heidegger wrote his doctoral dissertation defending logic against psychologism.[46] For Heidegger, the undeniable and over-ruling sense of logical validity is a phenomenological indication that judgment transcends the psychological conditions that might accompany it.[47]

The key to the distinction between merely sensitive consciousness and rational consciousness is judgment. All consciousness is intentional, but not all consciousness is or needs to be judgmental. With judgments, either noetic or evaluative, we enter what Robert Sokolowski calls 'the space of reasons.'[48] For a machine to do most of the things we do, it need not possess rational consciousness. But for a machine to supplant us on the planet, it must assume the power and the risk of judgment. AI may improve on us with regard to calculative ability and efficiency at optimizing the conditions of human flourishing, but it will not replace us as the mind of nature, the microcosmic mirror of the whole, so long as it does not possess the capacity to judge and decide. Without symbolic consciousness, which would allow it the distance from its being to make judgments of truth and falsehood, right and wrong, it will be merely a hyper-efficient animal.[49]

---

[45] Husserl, *Logical Investigations I*, 101.

[46] See, Martin Heidegger, "Die Lehre vom Urteil im Psychologismus (1913)," in *Gesamtausgabe 1, Frühe Schriften, 1912–1916*, ed. Friedrich-Wilhelm von Herrmann (Frankfurt am Main: Vittorio Klostermann GmbH, 1978); Emil Lask, "Die Logik der Philosophie und die Kategorienlehre. Eine Studie über den Herrschaftsbereich der Logischen Form," in *Gesammelte Schriften, Band 2*, ed. Eugen Herrigel (Tübingen: J. C. B. Mohr, 1911).

[47] Sean McGrath, *The Early Heidegger and Medieval Philosophy: Phenomenology for the Godforsaken* (Washington: The Catholic University Press, 2006), 93-7.

[48] To think is not to behave in a certain way or to be determined by certain brain-dependent mental events (which may be necessary to thought but are not sufficient for it). To think is to 'enter as agents into the space of reasons. See, Robert Sokolowski, *Introduction to Phenomenology* (New York: Cambridge University Press, 2000), 116. CF. Nagel, *Mind and Cosmos*, 79: "When we rely on systems of measurement to correct perception, or probability calculations to correct intuitive expectations, or moral or prudential reasoning to correct instinctive impulses, we take ourselves to be responding to systematic reasons which in themselves justify our conclusions, and which do not get their authority from their biological origins."

[49] See, Hooke, "Martin Heidegger's Concept of *Understanding* (*Verstehen*)" 19: "Authentic "having" is one necessary feature of human intelligence that avoids competing with the exponential growth of AI's outcome-based achievements. The success of AI (and AGI) is measured based on the results of their programming. This species of pragmatism is hopelessly ontic. It attempts to reveal and provide a service for things (*pragmata*) on hand, without concern for the structure of experience. AI

3. Conclusion

This, then, is the Holy Grail of AGI research: not only the functional reduplication of the activities which we now associate with NHI, but the design of a machine that will do the things we do *in the same way* that we do them, albeit with much greater efficiency and evolutionary capacity. The aim of strong AGI is nothing less than the mechanical reduplication of the human difference. The most ambitious and speculatively inclined AGI researchers are not assuming a weak sense of consciousness such as might be predicated of all beings capable of responding to stimuli, from the sea urchin to the robot, but a strong sense of consciousness, consciousness as the capacity for objective, rational judgment, for knowledge in the full sense of the term—*theoria*, not just *praxis*, and *poiesis*, not just *techne*, and therefore consciousness that can produce imputable judgments. Regardless of whether or not such a thing proves possible, the aim itself forces philosophy to clarify how rational judgment and decision distinguish human consciousness from other forms of consciousness, and what are its material and immaterial conditions. In order to be able to ascertain whether this will have been achieved, we will need to be clear on what the human difference is.

Until a machine gives us reasons to think that it has attained symbolic consciousness and that it now, like us, takes a *theoretical* interest in questions of truth and falsehood, that it too is sometime driven by a disinterested desire to know, that is, to contemplate the meaning of its existence, we will have no reason to recognize it as intelligent, in the human sense of the term. To the question, what would count as evidence? we can only point to those cultural products which most plainly exhibit our contemplative impulse and capacity for symbolic mediation, that is, to art, philosophy, and religion. A machine that had become artistically expressive, philosophically perplexed, or religious would indeed be worthy of our recognition as a rational agent. Of course, it might always be duping us for its own evolutionary advantage. We could never be sure, just as we are never so sure about each other.

---

programmers are incentivized by technocratic control and dominance, leaving no place for the "passive" call of conscience or self-understanding regarding the ontological notion self-actualization."

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Why Data Takes to Painting: Interdisciplinarity and Aesthetics

Stefanie Voigt
Universität Augsburg, Germany
ORCID: 0009-0007-3763-0814

Abstract
The mystery of human consciousness can be dealt with successfully in the context of an interdisciplinary theory of aesthetics. This discipline, however still marginalized due to historical reasons, can show in a modern way, informed by the theory of systems, how human consciousness is connected to three stages of the experience of beauty: simple recognition of patterns; intensive search for patterns; ecstasy or enstasy. That we can argue for this connection between aesthetics and consciousness based on our intuitions is shown by an example from popular culture: the android Data of *Star Trek, The Next Generation*, who takes to the arts in order to become human.

Keywords: human consciousness, interdisciplinary theory of aesthetics, marginalization of aesthetics, theory of systems, experience of beauty

On the Importance of Studying Aesthetics

In the search for the mystery of the human soul, aesthetics is mostly disregarded. At least it does not play the role of a respected interlocutor in the interdisciplinary canon of the cognitive sciences. But because of this very neglect, consciousness seems to become the insoluble mystery which it keeps being taken for. Here it is argued, however, that reformulating aesthetics in the context of the theory of systems and the close neighboring disciplines can bring about a new conception of aesthetics, more precisely: a model of human information-processing which defines consciousness as an aesthetic phenomenon.[1] Due to its formalization of aesthetic experience respectively consciousness, that model is able to make accessible pertinent topics, which so far have been regarded as 'artistic' and therefore discursively ungraspable, to the disciplinary as well as to the interdisciplinary dialogue.[2]

Throughout different disciplines, there are many approaches how to establish aesthetics as a coordinating core discipline of the cognitive sciences,[3] and these approaches display an astonishing convergence of their contents—a clue for the still next to unfathomed interdisciplinary potential of aesthetics.[4] Nevertheless, while numerous disciplines are celebrating interdisciplinary family reunions, aesthetics is mostly left behind like an unloved child.[5] There are and have been, however, prominent voices which regard sensory perception and the experience of beauty as crucial for human consciousness.[6] These practicians and theoreticians of aesthetics define their topic as a fundamental technique of human information-processing which concerns much more than representative decorations on the wall. According to them,

---

[1] This is unfolded in greater detail in author, *Das Geheimnis des Schönen. Über menschliche Kunst und künstliche Menschen, oder: Wie Bewusstsein entsteht* (Münster: Waxmann, 2005).

[2] See author, *Das Geheimnis*, 206f.

[3] See author, *Das Geheimnis*, 25f.

[4] Ursula Brandstätter, *Grundfragen der Ästhetik: Bild-Musik-Sprache-Körper* (Weimar-Wien: utb, 2008), 65-67 (emphasis on transdisciplinarity as the way to go); Michael Franz, "Ästhetik zwischen Philosophie, Wissenschaftsdisziplin und Techne-Diskursen," in *Ästhetik: Aufgabe(n) einer Wissenschaft,* ed. Karin Hirdina, Renate Reschke (Freiburg: Rombach, 2004), 121-134, 133 (calling for an "aesthetics which is situated and can take its stand in the tense field between philosophy, individual scientific disciplines, and discourses of technology").

[5] For example, Maria Elisabeth Reicher, *Einführung in die philosophische Ästhetik* (Darmstadt: Wissenschaftliche Buchgesellschaft, 2005), 24f, draws a sharp distinction between philosophical aesthetics on the one hand and empirical and thus also psychological aesthetics on the other hand, without inquiring into perspectives of mutual interdisciplinary completion. The volume *Ästhetik in der Wissenschaft: Interdisziplinärer Diskurs über das Gestalten und Darstellen von Wissen*, ed. Wolfgang Krohn (Hamburg: Meiner, 2005), is, despite of its title, more or less content with different disciplines standing side by side. A transcribed talk of different experts, however, ends with the conciliant statement that they still can learn much from one another.

[6] These are—amongst many others in each case—on the field of art itself: Leonardo, Cézanne, Malewitsch, Picasso, and Beuys; in psychology: James, Jaynes, Festinger, and Beyer; in philosophy: Adorno, Lyotard, Gadamer, Sloterdijk; in linguistics and semiotics: Chomsky, Ong, Peirce, Wittgenstein, and Bachtin; in anthropology Bateson, Lurija, Duerr, and Harris; in the history or theory of art: Flusser, Barthes, Panofsky, Sedlmayr, and Bataille. See, Stefanie Voigt, *Das Geheimnis des Schönen* (Germany: Waxmann Verlag, 2005).

here rather something lies hidden like the world-formula of all humanities, the mystery about the human soul, happiness, beauty and being alive—i.e., the clarification of all those concepts which have lost their home in academic, especially scientific, discourses under the influence of positivism.[7] These authors, however, could not turn the tide. Therefore, so far, aesthetics has been rather disregarded by discourses on the theory of consciousness and, moreover, sometimes made the impression of a nearly solipsistic self-containment. For this situation, the following eight reasons can be given.[8]

Reasons for the Marginalization of Aesthetics

Especially since the time of the Romantics, art as the main subject of aesthetics uses to be defined so that it is graspable not in an academic, but rather in another, "intuitive" way, so that it cannot be integrated into an interdisciplinary academic canon (1st reason).[9] Moreover, art, according to a wide-spread philosophical conceptualization, is considered as being *par excellence* free of purpose (2nd reason).[10] This conceptual clamp confronts art with a dilemma: If it does fit any purpose after all, it is claimed to be "mere" design in the form of kitsch or handicrafts.[11] If art on the contrary appears to be really free from purpose, it is quickly suspected to be a proverbially "aesthetic" leisure-time activity (3rd reason).[12] Philosophy of art is unable to mediate in this conflict, because by the way of paradox it is quite out of touch with its own object, with art. Therefore, philosophy of art dedicates itself more and more to reflections on its own status.[13] This leads to a dearth of even more elementary conceptual analyses, especially concerning a clean separation between the concepts of aesthetics and beauty, which also in philosophy of art are often used synonymously, as

---

[7] See, Klaus Städtke, "Form," in *Ästhetische Grundbegriffe. Vol. 2: Dekadent-Grotesk,* ed. Karlheinz Brack (Stuttgart-Weimar: Metzler, 2001), 462-494, 483.

[8] See, Voigt, *Das Geheimnis*, 19ff.

[9] See, Voigt, *Das Geheimnis*, 22; Klaus Städtke, "Sprache der Kunst/Kunst der Sprache," in *Ästhetische Grundbegriffe. Vol. 5: Postmoderne-Synästhetsie,* ed. Karlheinz Brack (Stuttgart-Weimar: Metzler, 2003), 619-641, 632-634.

[10] See, Voigt, *Das Geheimnis*, 104f., and Reicher, *Einführung*, 151f. This, of course, makes art the object of positions which propagate superior purposes for all human activities; cf. Kai Hammermeister, *Kleine Systematik der Kunstfeindschaft* (Darmstadt: Wissenschaftliche Buchgesellschaft, 2007), 158-162.

[11] See, Voigt, *Das Geheimnis*, 108-111.

[12] See Voigt, *Das Geheimnis*, 29f. Ephraim Kishon, *Picasso war kein Scharlatan* (München: Langen-Müller, 1986).

[13] Cf. Wolfgang Welsch, "Philosophie und Kunst—eine wechselhafte Beziehung," http://www2.uni-jena.de/welsch/ (accessed December 23, 2022). 1. Welsch compares the relationship between art and philosophy of art to a failed marriage and therefore sees the best solution in the amicable separation of both parties. On the according history of alienation cf. Ursula Franke, "Nach Hegel. Zur Differenz von Ästhetik und Kunstwissenschaft(en)," in *Ästhetik in metaphysikkritischen Zeiten. 100 Jahre Zeitschrift für Ästhetik und Allgemeine Kunstwissenschaft,* ed. Josef Früchtl, Maria Moog-Grünewald (Hamburg: Meiner, 2007), 73-91.

though they often signify something quite different (4th reason).[14] Because not everything which is called aesthetic is also beautiful. And not much of what some people would call beautiful would be called aesthetic at all by other people, especially because of the obsolete attitude towards the concept of beauty still to be found in the educated middle class (5th reason).[15] Given such confusion even as to the basic concept, it is no wonder that philosophy of art consists of a heterogenous mixture of different opinions about the topics aesthetics and beauty (6th reason).[16]

This reign of confusion might suggest consulting psychology for a therapy. Psychology, however, is forced to reject the dialogue which would be required for that purpose. The psychology of our time, conforming the sciences, prefers to dedicate itself to objects which can be quantified and grasped by statistics. This also leads to statements about aesthetics, but they are of a very elementary character. So, e.g., it is found out that black and yellow as the preferred combination of colors is more indicative of neuroses than any other arrangement of colors.[17] Elementary psychology of that kind may be interesting, especially for bees and fire salamanders, but it is too special for great insights into the essence of human beings. By reneging on the bulk of aesthetic phenomena as beyond the grasp of science, psychology follows the mystification of art as something unspeakable (7th reason).[18]

By leaving "great" aesthetic theories behind, modern psychology at least avoids being attacked by the proponents of a historical anthropology like the so-called Annales School centered around Le Goff.[19] According to this position, the essence of the human beings changes over time, and therefore it would be just wrong to conceive general academic, as, e.g., psychological, statements about "the" aesthetic perception etc. That precisely with the help of aesthetics the dynamic of the human psyche throughout its different historical changes can be explained is inaccessible from the perspective of that position alone. And this holds true not only for the mentioned disciplines, but generally: Aesthetics is not another discipline besides many others, but an interdisciplinary field of research. Just because of this, the individual disciplines, which in the first place are confined to their area, have trouble to find an access to aesthetics (8th reason).[20]

---

[14] See, Voigt, *Das Geheimnis*, 20.

[15] See, Voigt, *Das Geheimnis*, p. 20f.

[16] See, Voigt, *Das Geheimnis*, p. 21f. Reicher, *Einführung*, 9-31, deals with these conceptual problems in an aesthetics based on the analysis of concepts.

[17] See, Voigt, *Das Geheimnis*, 23.

[18] See, Voigt, *Das Geheimnis*, 228.

[19] See, Voigt, *Das Geheimnis*, 46.

[20] See, Voigt, *Das Geheimnis*, 23f.

Being Human

According to many thinkers, aesthetics is the key to consciousness, but communicating across the boundaries of the disciplines is hard to do. To get along in this intricate situation, the obvious way, as often with systems limited off against one another, is to ask a total outsider to give his assessment. In this respect, an apt subject of study is Data, the painting and violin-playing android in the TV-series *Star Trek—The Next Generation*.[21] He develops human properties like having a conscience and individuality when he does art. Only then can he access laughter or a certain kind of indulgent self-sufficiency, which otherwise seem to be reserved for humans. On the downside, this makes Data also prone to doubting himself or to be afraid of typically "human failure." Here, obviously, a widespread, but until now rarely explicated, intuition is staged: Humans know being emotional, pity, guilt, empathy, regret, joy and grief or vulnerability, irony and creativity, self-responsibility. These and other "typically human" phenomena are closely attached to the realm of aesthetics; at least, in it they can be experienced in an exemplary, intensive way. Foremost, in that realm the so-called paradox of informatics[22] does not occur: The computers which have been constructed so far just crash when confronted with contradictory information which is not provided for in their program. In stark contrast to this, aesthetics thrives on such contradictions, it highlights and intensifies them on manifold levels.[23] Therefore, if an artificial entity is dealing with aesthetics, we are inclined to ascribe a greater or lesser extent of humaneness to it. According to the opinion of some psychologists, consciousness is even characterized by the creative handling of contradictions.[24] Consequently, humans are 'functioning' as long as they are alive, and they feel that they are alive as long as they experience some things as beautiful. For if a human being loses for whatsoever reasons the capacity to experience something as beautiful, he or she will fall ill.[25] This is one more difference between humans and computers, and this is also one more indication for the connection between human consciousness and the experience of beauty. So far, the clinical pictures of suicidality, schizophrenia, and epilepsy have mostly been measured just by means of neurophysiology; their conceptual logic respectively the mental regulation of inner states has not been

---

[21] On Data as a hypothetical but nevertheless revealing test-case for philosophical questions cf. Robert Alexy, "Data und die Menschenrechte. Positronisches Gehirn und doppeltriadischer Personenbegriff" (2000), https://www.alexy.jura.uni-kiel.de/de/download/data-unddie-menschenrechte (accessed December 23, 2022).

[22] See, Douwe Draaisma, *Die Metaphernmaschine. Eine Geschichte des Gedächtnisses* (Darmstadt: Wissenschaftliche Buchgesellschaft and Primus, 1996), 165-168.

[23] See, Voigt, *Das Geheimnis*, 119-122.

[24] See, Julian Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (Boston: Houghton Mifflin, 1976).; Dietrich Dörner, *Bauplan für eine Seele* (Reinbek bei Hamburg: Rowohlt, 2001).

[25] See, Voigt, *Das Geheimnis* 147-149. This is why Nietzsche conceives of art as a "stimulant for the sake of life"; on this See, Helmut Peitsch, "Engagement/Tendenz/Parteilichkeit," in *Ästhetische Grundbegriffe* Vol. 2*,* 178-223, 193, fn. 119.

comprehensively inquired into yet. In each of these cases, the patients are no longer able to control in a conventional way their representation of the world, the inner picture of their environment, i.e., the product of aesthetic experience. They are forced to cling to alternative strategies instead, in the worst case they perceive things which do not exist or produce spontaneously feelings of bliss or beauty which under some circumstances overcharge the mental system or even motivate suicide. If the simulations of consciousness exhibited so far contained real consciousness, they would be lucky if they were spared by those problems; but then they would also be sad for not having art and literature. For these areas are rife with such problems, with emotions and beauties of different colorations. Human information-processing systems seem to be larger than the sum of their individual components. But how can that be possible? So far this remains unexplained. Neurophysiologists measure the brain and computer-scientists program their computers and someplace else scholars discuss the human soul and the noble art—and in-between there is a yawning chasm which no network has been able to bridge yet, despite of many attempts and consortiums of computer-scientists with neurologists, ethicists, imaging scientist, or neurobiologists meant to create new disciplines like neuroinformatics or neuroethics.

A Model of Being Human

What might Data have done to fathom through art the mystery of being human? Did he scan an introduction to the theory of aesthetics from Plato to Bazon Brock and translate it into the language of his artificial synapses? Because Data's brain is a digital computer, the whole issue at hand would be described by basic means of digital information processing, with the goal to unify all existing partial disciplinary insights within the framework of a single theory. Such a translation of noble art-theories into the digital 1-0-code might eventually be possible after all, notwithstanding the mentioned contrast between computers and humans. In the context of a psychology inspired by the theory of systems, which is based on the insights of anthropology, the multiplicity of individual and seasoned, often contradictory theories of aesthetics can at least be reduced to eight consistent frame-giving variables. Using the Aristotelian-Wienerian-Batesonian concept of difference (as the smallest unit of each mental performance), aesthetic experience can be characterized as follows: [26]

> 1. In aesthetic experience, a mediation between extern impressions and intern patterns of interpretation takes place, an encounter of sensory perception and abstract thought, respectively a simultaneous perception of world and the own person, of inner subjective schemata and subjectively outer world—in Plato's differentiation

[26] On this and the following, see, Voigt, *Das Geheimnis*, 91f.

between ideas and phenomena as well as in Aristotle's distinction between a work of art and the spectator who identifies him- or herself with it.
2. This general mediation works in detail via the comparison between inner and outer patterns, a correspondence between and re-modelling of structural principles—in Ficino's mirror projection as well as in Alberti's studies on proportion.
3. What comes to be by the interpreting perception of this comparison, i.e., by a perception of perception, are pleasing qualities of experience, the very own value and cognitive content of the aesthetic, which is cherished again and again—in Edmund Burke's 'pleasure' as well as in Kant's 'disinterested pleasure.'
4. These qualities of experience occur in tokens of varying strength, ranging from interest or fascination up to extasy or enstasy—from Lessing's emotion to Stendhal's "symptom." [27]
5. All of this is made possible only by a moderately stress-free mode of perception which is experienced as neither boring nor too exciting—in Schopenhauer's "contemplation of nature" as well as in Nietzsche's feeling of superiority in the "superman" or in modern Abject Art.[28]
6. A realm with rules of its own arises because the reception on the basis of this contemplative way of perception cannot and must not be grasped from a conscious and rational distance—therefore Boileau speaks about the "je ne sais quoi" ("I do not know what") and Goodman of the very own "languages of art."
7. In its entirety, this process guarantees the sustainable functioning of the system psyche by a better ability to think concerning the outside and by emotional pleasure gain on the inside—therefore Lyotard's "presence" is as important as Wittgenstein's "correct perspective."
8. The motivation of aesthetic contemplation or so-called discursive thinking depends on the ability to connect oneself in a situational and personal way—from Baumgarten's "disposition towards aesthetic-logic cognition" to Schiller's "playful instinct."

These eight statements can be formalized in the shape of a flow diagram, and so it is possible to arrive at a functional description of the according mental processes which defines aesthetic perception in a value-neutral way as perception between determinacy and indeterminacy.[29] Determinacy is based on the set of patterns already available within the system; indeterminacy is everything which does not correspond to these patterns. At first, there is a categorical contradiction between these two instances—what is determinate is not indeterminate and vice versa. This contradiction, however, can be bridged by a mediation between determinacy and indeterminacy, by the system restructuring its previous patterns and thus re-interpreting them or even, if this should not be sufficient, creating new patterns. This process may repeat itself, across different

[27] See Christian Kaden, "Musik. II. Ritualität in der Krise: Platons Musikphilosophie," in *Ästhetische Grundbegriffe. Vol. 4: Medien-Populär,* edited by Karlheinz Brack et al. (Stuttgart-Weimar: Metzler, 2002), 261-263, 263.
[28] See Winfried Menninghaus, "Ekel," in *Ästhetische Grundbegriffe. Vol. 2*, 142-177, 175f.
[29] Cf. Voigt, *Das Geheimnis*, 185.

cycles and increasingly large areas of the given storage of patterns until it either leads to a success or overcharges the system in question, which leads to an end of the search. This model, at first glance very formal and abstract, can serve as an interpretation of phenomena which so far have been described in different ways by different disciplines and thus it can afford the interdisciplinary integration needed for understanding consciousness:

- In the area of *neurophysiology*, based on different regulations of the messenger-substance dopamine two different, complementary kinds of information-processing can be established, namely so-called fixative respectively vagative thinking. Fixative thinking operates with abstract, simplifying concepts and manifests itself accordingly in a dimming of cerebral activity. Fixative thinking, on the other hand, proceeds in an erratic and multilayered way, as it were irrational, and manifests itself in an increasing spread of cerebral activity. The EEG measurements of the latter show parallels between techniques of meditation or other cultural forms of intuitive-aesthetic practice, be it in American concert goers or Siberian shamans.
- From the point of view of *anthropology*, in most civilizations the ability to experience some kind of extasy is taken to be one of the basic conditions of common sense. For only temporary extasy makes thinking sober again, so that it can face ever new challenges without narrowing itself down in a dogmatic way.
- Comparisons within the *history of mentality* show, however, that the ability of extasy has been more and more internalized and secularized during modernization, which has pushed it into a special district of the aesthetic, into art as acknowledged by society.
- *Art, the history of art and literary studies* offer numerous pertinent examples for this, e.g., the simultaneous occurrence of the internalization of experience and social processes of individualization; the parallel of tabooing death and at the same time dramatizing it; the scientification of thought accompanied by the discovery of the topic of atmospheric moods etc.

Thus a model emerges which explains the experience of beauty on three levels: simple beauty in the form of mere recognition of patterns (which mostly is interpreted just as kitsch or is felt by many as boring); a second level of intensified search for patterns, which is accessible only with cognitive efforts (in this case, the reception is described as "fascinating" or "interesting," while new interpretations of the cognitive problems are being elaborated—or the aesthetic search for a pattern is called off); and a third level of extasy or enstasy. Being close to this area is indicated by Lessingian "emotion" or physiological tears. What happens on this level would be described as "divine" formerly; in modern times, this became the experience of truth or total beauty. Here, the borders between ego and world as well as other kinds of difference become fuzzy.

The "way back" to normal consciousness tends to be described traditionally as "resurrection";[30] if this return goes wrong, schizophrenia may ensue, in which one does not see any more the wood between the stimuli.

All essential topics of aesthetics can be explained in a functional way, starting with Max Ernst's "courage of the artist" up to the cliché of the genius artist as a victim of his drives. Further examples are the metaphors of childishness, intoxication, dream, and sex, the beauty of idealizations and the beauty of ugliness, as well as the different kinds of empathy, be it empathy towards humans, the limitless ocean, or luxurious cars. Melancholy, mystic sensory overloads and pleasing self-extinctions, the legend of the purposelessness and indescribability of art: All this works on the logic of the regulation of differences in human thinking, including the differences between the objects of thought, between humans and their environment, and between thinking in differences and the thinking without differences in aesthetic borderline experiences. The latter difference cannot be but without purpose and indescribable, otherwise it would not be without differences but could be described as fixative and rational as anything else. Beauty is beyond description, and there are good reasons for this. Because aesthetics works like the blind spot of perception. Blind but necessary so that the eye can work. For the pleasure gain by the "short circuit" of thinking in a few moments without difference brings about a necessary and consciousness-generating counterpart to the "normal consciousness," giving to it at the same time also new motivation: For every abstraction presupposes the knowledge about its opposite. Every horse is defined by everything what is a not-horse, and every clear thinking by its opposite.

This holds also for real life: Without the aesthetic, daily routine becomes bleak; and only the aesthetic is vice versa an opening pitch for insanity. As Kant put it: "Without sensuality, no object would be given to us, and without intellect, no object would be thought. Thoughts without intuitions are empty, intuitions without concepts are blind."[31] Now the theory of systems postulates that no system can know itself, because knowledge is always a part of the system. But the theory of systems in philosophical aesthetics describes extasy, the extreme form of aesthetic perception, in many places as the possibility of encountering oneself, the so-called heautoscopy.[32] No argument is immune to skeptics, but, in any case, experiences like those at least create eventual opportunities for self-distancing through approaching oneself. According to Maturana and Varela, to use this opportunity for self-distancing is an indication and presupposition of intelligence.[33] So, did evolution create the experience of beauty to make intelligence possible? This is not improbable; what ordinarily is called beauty, however, is not what promotes humanity. If Einstein rejected the proposal of a model,

---

[30] See Voigt, *Das Geheimnis*, 192; Pia-Maria Funke, *Über das Höhere in der Literatur. Ein Versuch zur Ästhetik von Botho Strauß* (Königshausen & Neumann: Würzburg, 1996), 121ff.

[31] Immanuel Kant, *Critique of Pure Reason*, B75/A51.

[32] See, Voigt, *Das Geheimnis*, 205.

[33] See, Voigt, *Das Geheimnis*, 139.

as a well-known anecdote has it, this did not happen without reason. He argued that their children might have the intelligence of the mother and the looks of the father. But also studies on people who are taken to be beautiful according to social standards show that looks are not everything.[34] On the contrary, according to statistics, for attractive people it is more probable that they become unhappy in life than for the average person. Therefore, in the described model of aesthetic perception, not only form, but also content plays an important role, and it becomes very clear that humans are neither mere thinkers nor merely sensual beings, but a mixture of both with a dynamic of its own.

In times of industrialization, of the so-called human potential, an interdisciplinary model of aesthetics shows that, in education, the neglect of the artistic is not necessarily of advantage for the other educational subjects which are promoted now. Not only the contemplative character of the artistic, but also fairness and honesty prove to be important ways of access to the aesthetic. That this model was formed like a construction manual for Artificial Intelligence does not mean, due to its immanent logic, that this manual should be put to action; likewise, any attempt to conclusively define aesthetics would fail the topic in a drastic manner—because it would destroy the due share of indescribability which cannot be accessed rationally. Moreover, in a correct implantation of the model, accidents and mutations would make the product as unusable as the human being; only in this case it would be a true implementation. Because, if in the human brain there really is something like a "chaotic causality,"[35] then any prediction of whatsoever would be utterly impossible. From this there would result an undeducibly large set of possibilities for experience and intern connections and therefore also an according set of possibilities for oblivion. Then at the latest the question would arise what sese that project makes. After all, the human next door is already unintelligible and opaque enough, and, moreover, already there. Hence, it is hard to program the aesthetic, for reasons not only programmatic, but also pragmatic.

In this argumentation, science and art go hand in hand. Such a model, however, serves as a starting point for more precise definitions in different disciplines and as a stimulus for further research. The psychological concept of "sense of opportunity," e.g., closely considered quotes the concept of "possibilities" in Dionysius Areopagita, who describes real things just as "possibilities" of the beautiful and the good. As well does Plotinus' indifferent One as the place of religious experience anticipate certain neurophysiological results, namely the dimming of the cerebral activity pertinent for the distinction between self and outer world in some psychological processes. By analyzing such structural similarities, the traditional, very broad concept of aesthetics might be made more precise. This concerns also certain social and ethical evaluations: From the pre-modern point of view, e.g., a modern

---

[34] Cf. Winfried Menninghaus, *Das Versprechen der Schönheit* (Suhrkamp: Frankfurt am Main, 2003).

[35] See, Humberto Maturana and Francisco Varela, *The Tree of Knowledge* (Boston: Shambala Press, 1987).

human being with little contact to society is taken to be crazy. The same applies, however, to a human being of the past, too, seen from the point of view of its modern descendant, just because of the former's ecstatic practices or his or her "topsy-turvy worlds" which, in traditional civilizations, reverse the given order and thereby stabilize it, thus playing an important role (e.g., in the Saturnalia or in carnival). The key concept of "over-aestheticization," too, presupposes a clearly defined understanding of the human including certain evaluations which rather with than without such a model may be more easily formulated in an academic way—although the model would already suffice its artistic and system-theoretic demands already if the imagination of such interdisciplinarity would just promise joy of thinking, because the possibility cannot be ruled out that Data would have based his research on that foundation.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Artificial Intelligence and the Environment

Joachim Rathmann
Universität Augsburg, Germany
ORCID: 0000-0001-5533-2617

Abstract
Artificial intelligence (AI) can create new knowledge by analyzing large amounts of data, by recognizing patterns in the data sets via machine learning to create new knowledge about ecosystems. In addition, environmental balances can be created in the process, which can be used as a basis for decision-making. Due to the large amounts of data, complex feedback mechanisms can be balanced and the costs of decisions can be made transparent. Despite these opportunities for resource-saving handling of nature, AI balancing should not be allowed to become an automatic decision-making process. For sustainable environmental action, an emotional connection to the environment is also important. This cannot be achieved by AI, here it is the task of natural intelligence to recognize its embedding in a larger natural context and to develop from it a lifestyle in an environmental virtue ethics perspective.

Keywords: artificial intelligence (AI), machine learning, ecosystems, environmental balances, environmental virtue ethics

Introduction

The idea of using intelligent technology to make human life easier has always moved people. In the *Iliad*, it is the golden servants who are at Hephaestus' service. He (Hephaestus)

> put on a chiton, took his cane and limped to the door.
> where two servants rushed to support their master.
> all cast in gold, they looked like living girls;
> not only could they speak and move their limbs.
> they also possessed understanding and had learned from the immortals
> the most versatile skills (Homer XVIII, V.415ff.)[1]

Aspects that have already been mentioned here are extremely topical. Care robots, for example, should support elderly people in their everyday lives and be able to adapt to new situations through machine learning. The use of robots is supposed to relieve the nursing staff through activities such as distributing food or medication or emptying rubbish bins and at the same time support elderly or disabled people in their independence for longer.[2] The acceptance of these robots is to be increased by a certain human resemblance and by characteristics such as learning ability and autonomy. In addition, AI-supported speech recognition in care can relieve the burden of routine activities such as documentation and administration. Technically, this is based on forms of machine learning, which are considered a central subfield of artificial intelligence (AI). But in this context, the actual term "intelligence" often remains fuzzy, but can generally be seen as the extent of the problem-solving ability of artificial systems. Numerous AI systems are now firmly established components of the reality of many people's lives; be it in learning preferences in music or films or in purchasing behavior. Approaches of weak AI, which are used here, serve to solve concrete application problems. Such approaches do not attempt to reproduce all the characteristics of human intelligence, but rather focus on a subarea that can be mastered through fast computing operations. Different machine learning methods are important approaches to weak AI. Statistical dependencies and patterns are determined from large amounts of data, which can be used for prediction or classification purposes. The quality of these applications depends on the quantity and quality of the input data.

In the following, we will address the question of whether and to what extent artificial intelligence can help to master the increasingly worsening global ecological crisis. After all, knowledge about the destruction of nature and the environment has been around for many decades and the idea of nature conservation has a long history.

---

[1] Raoul Schrott, *Homer Ilias* (Frankfurt am Main: S. Fischer, 2011).
[2] Oliver Bendel, Ed., *Pflegeroboter* (Wiesbaden: Springer, 2018).

In Germany, the Drachenfels near Königswinter is considered the first nature reserve, established in 1836. However, this was intended to preserve a romantically charged symbol rather than primarily untouched nature. Nature conservation has always also served to protect cultural landscapes. The world's first national park (Yellowstone in 1872, followed by Yosemite in 1890) then led to an increased awareness of protecting areas as habitats for animals and plants.[3] With "Pfisters Mühle" (Wilhelm Raabe 1884), the first German environmental novel appeared in the same period, a testimony to the pollution of water by sugar factories in the early days. Thus, an awareness of the need to protect nature and landscapes has been present in western countries for well over 100 years. In 1914, Ludwig Klages described the situation in an equally impressive and topical manner: "An unparalleled orgy of devastation has seized humanity, 'civilization' bears the marks of unleashed murderousness, and the bounty of the earth withers before its poisonous breath. So, this is what the fruits of 'progress' look like!"[4] Framed within this destruction is also the emergence of pandemics, for intervention in hitherto barely touched ecosystems can open new transmission routes for zoonoses and initiate pandemics. This idea can also be found in Klages' work: "and so it goes on until the worse setbacks of the wounded nature of exotic countries in the form of those terrible epidemics that attach themselves to the heel of the 'civilized' European."[5] More than 100 years later, it is no longer only the "civilized" Europeans who must struggle with the consequences of human interventions in little-touched ecosystems. The fact that humans ultimately harm themselves by destroying nature has also met with great public response in recent environmental history with the publication of Rachel Carson's non-fiction book *Silent Spring* in 1962. The knowledge of the urgency to implement effective climate and nature protection globally has thus been accessible to a broad public as well as decision-makers for decades. Consequently, there is not so much a lack of environmental knowledge or environmental awareness, but a lack of environmental being, of environmental action.

The fact that people do not react immediately and affectively to environmental crises is due on the one hand to the fact that the damage to people often occurs asynchronously in space and time, and on the other hand to the fact that people build up a "hiatus" in their actions between the immediate satisfaction of needs and a necessary everyday action. This "indirectness of lifetime" may be one reason for the massive discrepancy between knowledge and action in regional and global environmental discourse.[6] In addition, the aspect of defense against fear can be cited as a repression mechanism against apocalyptic scenarios. If knowledge is available,

---

[3] On the history of nature conservation in Germany, See, Barbara Stammel & Bernd Cyffka, *Naturschutz* (Darmstadt: WBG, 2015).

[4] Ludwig Klages, *Mensch und Erde,* in *Sämtliche Werke, Band 3, Philosophie III,* ed. Ernst Frauchiger (Bonn: Matthes & Seitz Berlin, 1914, 1974), 619.

[5] Ludwig Klages, *Mensch und Erde,* in *Sämtliche Werke*, 619.

[6] Arnold Gehlen, *Der Mensch. Seine Natur und seine Stellung in der Welt* (Wiesbaden: Athenaion, 1978).

there is also selective inattention and self-numbing.[7] This may be another reason why we do not behave appropriately despite our immense knowledge about the state of global ecosystems.[8] Overall, ecological knowledge and action remain only loosely coupled to each other; via cognitive dissonance, this also applies to particularly environmentally aware people. Cognitive dissonance arises when the attitude, opinion or norm does not match the actual action.[9] People strive to reduce such states of tension. To this end, arguments are often sought in ecological discourse to justify one's own actions, for example through constraints, institutional incentives, and other necessities. The example of German sustainability researchers shows how they justify their growing ecological footprint by using such arguments.[10] However: the emission of greenhouse gases remains unaffected. Another explanation for such behavior could be described as moral licensing.[11] Especially environmentally conscious people, since they stand up for the cause of the good, then, as it were, debit an imaginary environmental account, e.g., a flight, for which there are certainly good constraints. Mental rebound effects can then lead to increased resource consumption. This describes effects that result in the original savings potential not being realized or only partially realized, for example due to efficiency increases. This can have the consequence that in the overall ecological balance, the attitude of standing up for an ecologically good cause then replaces, as it were, the overall sustainable action. In short: there is no lack of good will: "Good will is fortunately abundant; it demonstrates itself everywhere," there is no lack of "attitude."[12] But the life worldly consummation of the conscious shows itself less in the truthfulness of the attitude than in the energy of action. But despite all this, even in everyday life it is often not so easy to determine which decision entails the least consumption of resources. Here, however, AI could provide a valuable decision-making aid.

---

[7] Hans Peter Dreitzel, *Reflexive Sinnlichkeit: Mensch Umwelt Gestalttherapie* (Köln: EHP, 1992).

[8] In detail on the "motivation problem" in environmental action, See, Christoph Baumgartner, *Umweltethik—Umwelthandeln: Ein Beitrag zur Lösung des Motivationsproblems* (Paderborn: Brill—Mentis, 2004).

[9] Leon *Festinger,* Cognitive dissonance, *Scientific American* 207 no. 4, (1962): 93-107.

[10] Isabel Schrems and Paul Upham, "Cognitive Dissonance in Sustainability Scientists Regarding Air Travel for Academic Purposes: A Qualitative Study," *Sustainability* 12 (2020): 1837, doi:10.3390/su12051837.

[11] Michael Halla, "Believing in climate change, but not behaving sustainably: Evidence from a one-year longitudinal study," *Journal of Environmental Psychology* 56 (2018): 55–62.

[12] Hermann Lübbe, *Politischer Moralismus. Der Triumph der Gesinnung über die Urteilskraft* (Münster: Lit, 2019).

The Paradox of Environmental Knowledge

The "paradox in environmental knowledge" describes a phenomenon of different spatio-temporal scales.[13] For on a global level, the requirements for achieving effective nature conservation in a comprehensive sense have been clearly identifiable for many decades: for example, the reduction of greenhouse gas emissions, land consumption, habitat fragmentation, large-scale deforestation, intensification of land use or overfishing in the oceans. Despite this body of knowledge, the "great acceleration"[14] shows that the main indicators of the state of global ecosystems continue to show accelerating trends in a negative direction, despite regional (and, in the case of the ozone layer, global) improvements. Apparently, knowledge about the ecological crisis is insufficiently relevant for action. One reason for this is that in individual behavior on a local level, it is often not at all clear what is really the more ecological alternative in terms of the complex consequences of a decision. The "paradox of environmental knowledge" shows that (not only) on an individual level, supposedly ecologically sustainable decisions can turn out to have complex negative effects. The organic carrot from Israel bought in Germany may be "organic," it is certainly not "eco." But is the regional product generally more ecologically sustainable than one from more distant regions where it can be grown more efficiently with less resource input?

In addition, a monetary perspective can be added: A bamboo toothbrush may be more sustainable than a plastic toothbrush that costs only a third of the price. However, in the perspective of effective altruism, the money saved could be used for environmental protection measures and thus provide an overall greater ecological benefit.

"Greenwashing" has a negative connotation and refers to the emphasis on the ecological advantages of products or processes without there being any basis for this in the overall consideration of all interactions. This is usually done by emphasizing selective aspects. For example, in the case of a T-shirt made of organic cotton, the high water consumption for cotton (in mostly dry regions), the land requirement and thus the competition for land, the transport or the use of fabric-dyeing substances, among other things, must be taken into account in an overall balance.

Another example could be the recycling of paper and cardboard. This behavior can bring about a certain environmental relief and even more strongly evoke the feeling of being a good environmentalist in the person acting. However, with the steadily increasing packaging waste due to the growing online trade, recycling is a smaller part of the solution; a reduction in the use of packaging and the quantity of

---

[13] Joachim Rathmann, "Von der Naturkunde zur Umwelttugendethik: Ein möglicher Weg zur Überwindung der Diskrepanz von Umweltwissen und Umwelthandeln?," *Comenius-Jahrbuch* 28 (2020): 97-120.

[14] Will Steffen, "The trajectory of the Anthropocene: The Great Acceleration," *The Anthropocene Review* 2, no. 1 (2015): 81-98.

orders would be more effective in terms of a truly significant reduction in environmental impact.[15]

Thus, bio-labelling and recycling run the risk of causing greater damage while at the same time increasing people's environmental awareness. They create the illusion of goodness, celebrate a triumph of sentiment, and fail to recognize the complexity of interrelationships, so that they can ultimately have a greater negative impact than is generally realized. The well-intentioned is not congruent with the good. For it is true for the use of resources in many products and processes that the interactions, even in different spatial and temporal manifestations, are so complex that the quick decision in favor of the supposedly more ecological product can be wrong. The resulting "unintentionality of the rapidly increasing burdens of civilization" should warn against the rampant moralism in ecological questions. For the burdens of civilization are too readily blamed on capitalism, the "system" and large corporations.

When weighing up ecological consequences of actions and purchase decisions, one could easily end up in the role of Buridan's donkey, which starves to death between two equally distant, equally large haystacks because it cannot decide which one to turn to. Analogously, detailed weighing in environmental decisions could lead to a deadlock situation in which both alternatives block each other, and a situation may seem hopeless. AI could come into play here and create a basis for decision-making by virtue of the calculation of large amounts of data and contribute to the transparency of the true costs and benefits.

Hoping for AI?

The rapid processing of large amounts of data by AI and the recognition of patterns in the data sets can create new knowledge about ecosystems and optimize their management. As a result, sustainable environmental behavior can be simulated by AI and the real use of ecosystems can fundamentally be made more resource-efficient and sustainable. In agriculture, for example, there are various fields of application for AI: agricultural processes can be controlled in real time according to location and need. The location-differentiated and targeted management of agricultural land is known as precision farming and is part of the digitalization of agriculture. This is also done using drones to collect precise data and create high-resolution images that help to monitor crops and at the same time help to optimize the use of resources and thus reduce the burden on the environment.[16] This is because precision farming uses AI to develop

[15] Hermann Lübbe states that the understanding of one's own living conditions is decreasing and that we are therefore increasingly dependent on the "expertise" of experts; this can only be based on trust. The need to consider the side-effects of individual actions and to assess consequences is therefore increasing sharply.

[16] Robert Finger "Precision farming at the nexus of agricultural production and the environment," *Annual Review of Resource Economics* 11 (2019): 313-335.

accurate and controlled techniques that help provide guidance and understanding for water and nutrient management, optimal harvesting and planting times, and crop rotation timing.

Further environmental relief in the agricultural sector could also come from vertical farming, where vegetables and lettuces are grown in closed systems in indoor farms. High productivity is ensured by the fact that the systems can grow in a space-saving manner over several stories (vertically) on artificial growing media or in nutrient solution. Proximity to consumers is another advantage. AI can precisely control the use of water, nutrients, light, energy, or humidity and optimally supply the plants without the use of pesticides.[17]

Another opportunity to improve the ecological status lies in the fact that AI can be used in the calculation of environmental impacts via life cycle assessments, climate assessments or the ecological footprint. The larger the incoming data volumes and the more interactions that can be considered, the more precise such calculations can be. AI has the potential to cope with these data volumes and, through processes of self-learning, to carry out the transferability of product assessments. This means that the respective "environmental consumption" of products and services can be quantified and used as a basis for decision-making. External costs can also be presented transparently. Customers would then be able to make an ecologically sustainable choice directly on the basis of comprehensive information when purchasing products. Multi-criteria decision analysis (MCDA) is therefore becoming increasingly widespread for many aspects of the energy and environmental sector.[18] This is because the main decision-making problems arise when several objectives (multiple criteria) are pursued, and the decisions take place in a complex context. Frequently, the available information and goals are of a conflicting nature when it comes to balancing economic and social concerns with the demands of species and climate protection, as well as substantial consequences and long-term impacts in different spatial manifestations. Therefore, wrong decisions can no longer be revised so easily. Complex decisions are no longer trivial, and the pure computational capacity of AI can meaningfully contribute arguments for decisions. For the decision-maker, the situations are formalized by a multi-criteria decision analysis, in which information is organized to such an extent that the decision-maker can contribute to an improved decision-making process with the feeling of having taken the essential criteria into account.[19] The aim is to provide technical support in complex problem situations, to make consistent, comprehensible, and more reasonable decisions or to support

---

[17] Malex Al-Chalabi, "Vertical farming: Skyscraper sustainability?," *Sustainable Cities and Society* 18 (2015): 74-77.

[18] Danae Diakoulaki et al., "MCDA and energy planning," in *Multiple Criteria Decision Analysis: State of the Art Survey,* ed. José Figueira et al. (Berlin: Springer, 2005), 859-897.

[19] Valerie Belton, Theo Stewart, *Multiple Criteria Decision Analysis: An Integrated Approach* (Dordrecht: Springer, 2002).

compromise negotiations based on the possibility of weighing up several alternatives in a flexible way by selecting, comparing, and ranking different attributes.

Limits and Risks of AI

AI-supported decision analyses in environmental issues can hardly bridge the fundamental discrepancy between environmental knowledge and action. Any decision support requires implementation by the decision-maker. However, the environmental discourse of the past decades shows that the knowledge of how to improve the ecological situation on a global but also on a local scale does exist. Factual knowledge alone is a necessary but not sufficient vehicle to bring about global change. Factual knowledge, even that of an AI-generated decision-making aid, hardly touches people's lives, it remains external to them.[20] AI can calculate the "what" in ecological matters, but for the question of the "why," natural intelligence is needed.[21]

In addition, information is often embedded in a certain framing; for the AI there is initially no difference whether a glass is half-full or half-empty, but for a human decision it is all the same. Purely logical weighing is blind to intuition, individual or socio-cultural embedding of decisions. Finally, the purely instrumental reason of AI needs to be supplemented, otherwise there is a danger of "technical perfection with complete failure of moral reflection."[22] For the human conscience, in its necessary weighing, prevents the judgements of the AI from becoming executioner's verdicts and man from ultimately becoming a slave to the digitalized world. For the corporeality of the ego as a person prevents the world from being perceived only from a spectator's perspective. AI, however, is an expression of scientism in the tradition of Francis Bacon. There is a danger that ultimately the "worst of all possible worlds" will be constructed. [23] For "representationality is then equated with availability—an equation that amounts to the abolition of object and representationality [. . .]. The opposite is true. The totally unavailable object is most object—the PERSON (understood as human or superhuman). In it, and only in it, is a maximum of depth realized."[24] In this way, natural intelligence also eludes artificial intelligence, from which it differs in manifold ways. AI is at best a "simulation of narrowly defined areas of human intelligence."[25] For essential aspects such as life, consciousness or perspective-taking

---

[20] Rathmann, "Von der Naturkunde zur Umwelttugendethik: Ein möglicher Weg zur Überwindung der Diskrepanz von Umweltwissen und Umwelthandeln?," 97-120.

[21] Helmut Kuhn, *Der Weg vom Bewußtsein zum Sein* (Stuttgart: Klett, 1981).

[22] Hermann Lübbe, "Scientific Practice and Responsibility," in *Facts and Values: Philosophical Reflections from Western and Non-Western Perspectives* (Dordrecht: Springer Netherlands, 1986), 9.

[23] Kuhn, *Der Weg vom Bewußtsein zum Sein*, 352.

[24] Kuhn, *Der Weg vom Bewußtsein zum Sein*, 353.

[25] Thomas Fuchs, *Verteidigung des Menschen: Grundfragen einer verkörperten Anthropologie* (Berlin: Suhrkamp, 2020).

cannot be generated by algorithms.[26] Life takes place in life itself and cannot be substituted by modelling; life always means relating affectively and emotionally to others. For: "the boundless objectification of the people of our day is gradually eating away at the forces that are necessary to maintain a mere material culture and merely technical operations, e.g., imagination, creativity, listening to the sources of life that roar in the depths. Why does contact with nature refresh us? Because for once we are alone with ourselves and can therefore also have contact with ourselves."[27] This ability to bond could turn out to be a central aspect in overcoming the ecological crisis. But this is where the limits of AI become apparent, because empathy has a bodily component that cannot be represented by it, despite a "fictional empathy" that can also arise towards computer figures or robots.[28] An "as-if" empathy that is devoid of meaning and has migrated into the virtual world loses depth and commitment, however.

Empathy

The increasing presence of digital media, sign systems and fictions may have led to a decline in perspective taking and primary empathy as well as psychological well-being in recent years. A well-received meta-study, based on data compiled from nearly 14,000 students in 72 studies from 1979 to 2009, finds a decline in empathy over this period.[29] This decline is particularly evident after the year 2000. The index used is the Interpersonal Reactivity Index (IRI), which indicates the extent to which someone can put themselves in the shoes of another person or of characters in films or books to understand them. The willingness to adopt a perspective has decreased in the last years of the study period, while the values for imagination have remained constant. This decrease correlates positively with the common contemporary diagnosis of "narcissism."[30] This study ultimately asked about attitudes towards empathy and did not observe the actual (empathic) behavior itself. Certainly, one can also critically ask whether, with such a long period of investigation with the same questionnaire between 1979 and 2009, shifts in meaning and different associations do not occur among the respondents. But despite all the fundamental methodological criticism, it could be that an old cultural pessimistic lament is seeking confirmation here, because "after all, it was already claimed a good 200 years ago that the new media would corrupt young

---

[26] Cf. Tab. 1 Fuchs, 59f.

[27] Eduard Spranger, *Gedanken zur Daseinsgestaltung*, Ausgewählt von Hans Walter Bähr (München: R. Pieper, 1962).

[28] Fuchs, *Verteidigung des Menschen: Grundfragen einer verkörperten Anthropologie*, 125ff.

[29] Sara Konrath "Changes in dispositional empathy in American college students over time: A meta-analysis," *Personality and Social Psychology Review* 15, (2011), 180-198.

[30] Rathmann, "Von der Naturkunde zur Umwelttugendethik: Ein möglicher Weg zur Überwindung der Diskrepanz von Umweltwissen und Umwelthandeln?," 115.

people and lead to narcissism. At that time, the new media were the novels that we now wish young people would read more of."[31] Nevertheless, less empathy and more narcissism seem to be particularly evident in social media, and this also has a negative effect on health: young people's psychological well-being decreases when they spend a lot of time in front of a screen or smartphone (social media, internet, games, etc.) compared to people who do more activities beyond a screen (direct personal contact, sports, church activities, etc.). This was shown in a nationwide survey from the USA over a period of 15 years between 1991 and 2016.[32]

However, an evaluation of the (alleged) decline in empathy depends on how the strongly positively connoted term[33] is to be filled in terms of content. Empathy as empathy is different from compassion with caring. For Bloom argues against this and shows that empathy, as mere empathy, can justify terrible situations.[34] Conflict can be amplified by empathizing with certain groups. A mere perspective-taking, an empathy with others is also possible in the case of perpetrators of violence, because empathy describes the ability and the tendency to feel the feelings that one believes the other person feels.[35]

Lipps has already elaborated the double-sidedness of perspective-taking in the concept of "empathy."[36] He distinguishes between positive empathy "colored by pleasure" and negative empathy "not colored by pleasure." Positive empathy is "the taking in of that which penetrates me, or it is the becoming one of the grasping I, as it is in itself, with that which penetrates it."[37] Negative empathy, on the other hand, is described as that against whose penetration "contradiction" arises. It rejects itself as "incompatible" with itself.[38] Implicitly, this ambivalence is also found in Bloom's work, in that he distinguishes aspects such as "kindness" and "compassion" from empathy and its negative sides, which he explicitly appreciates positively, as well as the positive

---

[31] Fritz Breithaupt, *Die dunklen Seiten der Empathie,* (Berlin: Suhrkamp, 2019), 71. Since leaving the Garden of Eden, the "O tempora o mores!" remains the accompanying melody of the Anthropocene. More modernly, Szymborska formulates this in the poem "Not Reading:"
"We live longer, but less precisely, and in shorter sentences, We travel faster, more often, further. And instead of memories, we bring back photos." Wislawa Szymborska, *Glückliche Liebe und andere Gedichte* (Suhrkamp: Berlin, 2014), 65

[32] Jean Twenge, *Emotion 18*/6, (2018): 765-780: Decreases in psychological well-being among American adolescents after 2012 and links to screen time during the rise of smartphone technology.

[33] For de Waal, the ability to feel connected to others is the bonding agent that positively connects people and peoples. For him, "empathy for "other peoples" [. . .] is the raw material the world needs even more urgently than oil" Frans de Waal, *Das Prinzip Empathie. Was wir von der Natur für eine bessere Gesellschaft lernen können* (Darmstadt: Hanser, 2011), 263.

[34] Paul Bloom, *Against Empathy. The Case for Rational Compassion* (London: Ecco, 2018).

[35] Breithaupt uses numerous examples to show the "dark sides of empathy." This is intended to sharpen the view that a central characteristic of human life, developing empathy, can also have negative consequences. Fritz Breithaupt, *Die dunklen Seiten der Empathie* (Berlin: Suhrkamp, 2019).

[36] Theodor Lipps, *Fühlen, Wollen und Denken. Versuch einer Theorie des Willens* (Leipzig: Johann Ambrosius Barth, 1907).

[37] Theodor Lipps, *Fühlen, Wollen und Denken. Versuch einer Theorie des Willens*, 236.

[38] Theodor Lipps, *Fühlen, Wollen und Denken. Versuch einer Theorie des Willens*, 236.

sides that empathy also shows. However, according to Bloom, empathy can also motivate indifference or even cruelty, because empathy is based on a certain short-termism, since it focuses on a specific counterpart; in doing so, there is a danger of overlooking longer-term consequences and the suffering of those who are not the current counterpart. Charity runs the risk of blinding the love of the farthest.[39] Nietzsche's *Zarathustra* recommends: "I do not advise you to love your neighbor: I advise you to love your neighbor from afar."[40] Nicolai Hartmann sees in Nietzsche, despite all exaggeration, the "positively seen [. . .]"[41] and describes the love of the farthest as love "that knows no love in return, that only radiates."[42] Admittedly, love at a distance begins with the neighbor, but in a sense sees him as a means to a higher (future) end. Often, love at a distance can be carried out without effort. Signing a petition for refugees or against the deforestation of the rainforest provides self-affirmation to stand up for the good, but picking up the rubbish by the roadside is comparatively uncomfortable. Therefore, the starting point of action must first be empathy with the immediate environment. The binding of the I in the Thou is not simply a projection of one's own in the Other; as an experience of love, it is an assurance of priority towards the Thou in loyalty, otherwise responsibility would remain "a free-floating ought."[43]

Even though moral decisions are essentially shaped by empathy, it is important to recognize that negative consequences can also result. Compassion is therefore a more appropriate way to contribute to the betterment of others, since it does not simply understand the feelings of the other person, but through sympathy, the motivation to promote the well-being of the other person grows by feeling for the other person. This also comes close to the concept of empathy that Goleman cites in the context of "ecological intelligence." On the one hand, this includes knowledge of ecological connections and, on the other hand, the insight "to learn from experience and to act meaningfully."[44] For him, this is linked to a form of empathy that encompasses everything "that lives."[45] This implies compassion when ecosystems

---

[39] Bloom sums it up in a short equation: "Self + Close People + Strangers=100%" (p. 162). Which shares (time, money, commitment, emotions) are invested in which sub-area and to what extent? Whereby it is clear that there are resources that diminish with use (e.g., money) and those that even increase with use (e.g., love and affection). In this regard, Alexander Batthyány, notes that "having comforted or encouraged a person does not mean that we will eventually run out of words of comfort or encouragement for the neighbor in need of comfort or encouragement." Love and affection therefore go beyond the model of resources, in that by giving, the subject gains and loses wealth if the possibility of giving is not realized. Alexander Batthyány, *Die Überwindung der Gleichgültigkeit. Sinnfindung in einer Zeit des Wandels* (München: Kösel, 2017), 89.

[40] Friedrich Nietzsche, *Also sprach Zarathustra*, in *Kritische Studienausgabe*, ed. Giorgio Colli, Mazzino Montinari, Vol. 4 (dtv: München, 1999), 79.

[41] Nicolai Hartmann, *Ethik* (Berlin: Walter de Gruyter, 1962).

[42] Nicolai Hartmann, *Ethik*, 490.

[43] August Vetter, *Natur und Person. Umriss einer Anthropognomik* (Stuttgart: Klett, 1949), 224.

[44] Daniel Goleman, *Ökologische Intelligenz: Wer umdenkt, lebt besser* (Droemer: München, 2009).

[45] Goleman, *Ökologische Intelligenz*,, 50.

"suffer" and to derive from this an action that seeks to reduce this suffering. In a deep ecological perspective, the perspective also expands to the inanimate: "think like a mountain," Aldo Leopold's dictum, makes it clear that the idea of protection also goes beyond the animate world.[46] For Berry, too, the idea is central that human bodies are closely interwoven with the surrounding nature and that only contact with the "wilderness" brings experiences—"to receive the awareness, at one humbling and exhilarating, grievous and joyful, that we are part of Creation, once with all that we live from and all that, in turn, lives from us."[47]

Abram opens the perspective that: "the perceiving being and the perceived being are of the same stuff, that the perceiver and the perceived are interdependent and in some sense even reversible aspects of a common, animate element, or Flesh, that is at once both sensible and sensitive." [48]

For Abram, this reciprocity of the sensuous extends directly to non-human life, which extends on a continuum into the landscape. For him, this explicitly includes remote love: "If the surroundings are experienced as sensate, attentive, and watchful, then I must take care that my actions are mindful and respectful, even when I am far from other humans, lest I offend the watchful land itself." [49]

This is immediately followed by the question of the good life. From a deeper (not necessarily a deep ecological) ecological perspective, it is obvious that the meaning of a good life can only be found in a measured and reverent treatment of our environment.

Compassion as Virtue-Ethical Potential

In the face of global ecological challenges, the question of the practicability of ethical action arises with new urgency. In the perspective of norm ethics, an established norm finds its application in a specific case. However, in the complexity of ecological-social systems with ever new feedbacks and rebound effects, the need to maximize adaptation possibilities becomes apparent. A virtue ethics perspective opens the possibility of strengthening personality traits that help to meet all concerns in complex situations.

Empathy is not moral at first because of the ambivalence of perspective-taking. However, empathy as compassion and as love allows further dimensions to be strengthened, because adopting the perspective of others enriches one's own feelings and perception with new perspectives. Complemented by empathy with others, the

---

[46] Bill Devall, Die tiefenökologische Bewegung, in: *Ökophilosophie*, ed. Dieter Birnbacher (Reclam: Stuttgart, 1997), 17-59.

[47] Wendell Berry, *Essays 1969-1990,* ed. Jack Shoemaker (New York: Library of America, 2019). 336.

[48] David Abram, *The Spell of the Sensuous: Perception and Language in a More-than-Human World* (Pantheon Books, 1997), 67.

[49] Abram, *The Spell of the Sensuous*, 69.

sensual dimension of perception is enlarged. This gives empathy an aesthetic quality because sensory perception is broadened to include other subjects and new perspectives are opened.[50] The perception of the world becomes richer and at the same time more complex through an increased sense of empathy. The co-experience of other perspectives can create closeness, trust and thus a new bond. In an environmental virtue ethics approach, co-experiencing is of course not limited to fellow human beings. Empathy with the sense of caring ultimately builds the bridge between the ethical and the aesthetic.[51]

At the same time, this enriching experience makes it possible to practice moderation in one's own life, since an additional empathetic experience overcompensates for it. This is a perspective towards an environmentally relieving behavior in the individual. Another advantage of a virtue ethics approach is that it does not have to define the circle of entities to be considered morally (anthropocentrism, pathocentrism, biocentrism, holism) and can thus avoid the demarcation problem in the view of environmental ethics. For a narrow anthropocentric position can be presented that only includes one's own individual (egoism), expanded to include all persons (personalism or humanism, all people present and future), finally the sphere of entities to be considered morally can be expanded to include all animals capable of suffering (pathocentrism), all living beings (biocentrism) and all of nature (holism). Holism argues that all entities in nature should be accorded their own value; nature is to be protected for its own sake.[52] Regarding the problem of demarcation, Gorke believes that only the most comprehensive position of holism is self-evident, "the answer of holism is . . . [the] only one [that] needs no further explanation."[53] In this way, Gorke believes he escapes the burden of justifying which entities can be ascribed an intrinsic value, because all other concepts of environmental ethics must in turn be able to conclusively explain why they exclude certain entities from the circle of beings to be considered morally. Consequently, one would have to ascribe to objects a value of their own that cannot be derived from human or animal consciousness and thus represents a counter-concept to instrumental value, hence an objective value. For the core question that anthropocentric arguments must face is whether it is: "really appropriate [. . .] to subordinate the more than three-billion-year-old process of biological evolution and the self-organization of ecosystems completely to the interest calculations of *Homo sapiens*."[54]

---

[50] Breithaupt, *Die dunklen Seiten der Empathie,* Suhrkamp, 209ff.

[51] Ludwig Wittgenstein, *Tractatus logico-philosophicus*, in: Werkausgabe Bd. 1 (Frankfurt am Main: Suhrkamp, 1997), 6.421: "Ethics and aesthetics are one."

[52] Angelika Krebs, *Naturethik*, ed. (Suhrkamp: Frankfurt am Main, 1997), 342ff.; Michael Gorke, *Eigenwert der Natur* (Hirzel: Stuttgart, 2010), 23f.

[53] Gorke, *Eigenwert der Natur*, 97.

[54] Gorke, *Eigenwert der Natur*, 95. Also see, Holmes III Rolston, "Werte in der Natur und die Natur der Werte," in *Naturethik*, ed. Angelika Krebs (Suhrkamp: Frankfurt am Main, 1997), 247-270, 264.

But even for holism, the danger remains that an anthropocentric view is implicitly extended to non-human entities and that ultimately an attribution of human characteristics is made after all. In a virtue ethics perspective, however, the problem of demarcation is not central, so such attributions, as well as moral status attributions, can be easily circumvented. At the center of virtue ethics is precisely an acting person who is motivated by eudaimonistic reasons. Central to a virtue ethics approach is individual action, and especially in environmental discourse, the discrepancy between knowledge and action has emerged as a central pivotal point. For in environmental decisions, dilemmatic situations such as those described above occur again and again. Every action has harmful side effects and a person acting causes environmental damage. This speaks for a virtue-ethical approach, which derives effectiveness from the strengthening of relevant virtues. In environmental behavior, moderation is a central virtue that can be strengthened through regular contact with nature.[55]

The Sense Dimension

Humans, unlike AI, are beings in need of meaning and must be touched in their essence in order to act. Therefore, the increasing environmental knowledge, the constant influx of ecological data on rising greenhouse gas concentrations and declining biodiversity remains external to many people and does not affect their existential being-in-the-world. Morton puts it in a nutshell: "data dump mode is just enhancing the incapacity of things to mean anything anymore to us."[56] People must consequently be addressed in their dimension of meaning as a central motivating factor, because: "There is probably no evil that man would not be prepared to endure if he were able to see a meaning to this suffering; but there is certainly no earthly good whose enjoyment would not become stale to man in the long run if he could not perceive the holding on to it as meaningful." [57]

However, meaning cannot be simulated by algorithms and therefore represents a further distinguishing feature of human and artificial intelligence. In addition to the meaning that an individual can discover for himself, it is necessary to consider "natural beings outside the human being," their "concepts of being and meaning, [. . .] which oblige us morally."[58] For human existence is essentially constituted by relationships, thus relationally. Therefore, human beings face non-human life in "a solidarity of sense expectation with all living things; a solidarity that is felt by us in sympathetic

[55] See, Joachim Rathmann *Therapeutic landscapes: An Interdisciplinary Perspective on Landscape and Health* (Wiesbaden: Springer, 2021).
[56] Thimothy Morton, *Being ecological* (Cambridge: The MIT Press, 2018), 154.
[57] Hans-Eduard Hengstenberg, *Sinn und Sollen. Zur Überwindung der Sinnkrise* (Ludgerus: Essen, 1980), 7.
[58] Hengstenberg, *Sinn und Sollen: Zur Überwindung der Sinnkrise*, 48.

resonance."[59] "Man can only become fully human when he not only 'takes' all things utilitarian, but also conspiratorially 'takes' them in their own being for their own sake."[60] For Hengstenberg, this corresponds to the imperative of objectivity to turn to non-human natural beings for their own sake. For him, this creates a "commitment in relation to all living things, not only to fellow human beings."[61] This "universal commandment of meaning" shows itself to be a sustainable basis for developing nature as a source of human meaning. In this context, the depth of meaning is not revealed in a continuous "more"; the development of meaning requires the courage to pause and recognize that concentrating on seeing less, experiencing less, doing less, increases the qualities of the little and lowers the need for more and more. A virtue-ethical approach is tied to a supposed limitation of the individual, which, however, turns out to be a qualitative gain. For in the many lies speed, superficiality, arbitrariness, but gain can be drawn from a qualitative relationship. Merton illustrates the idea with a visit to a museum:

> A tourist may go through a museum with a Baedeker, looking conscientiously at everything important, and still come out less alive than when he went in. He was looked at everything and seen nothing. He has done a great deal and it has only made him tired. If he had stopped for a moment to look at one picture he really liked and forgotten about all the others, he might console himself with the thought that he had not completely wasted his time. He would have discovered something not only outside himself but in himself. [62]

AI could structure the museum's wealth of information, but without any prospect of making sense or contributing to a good life. It remains for natural intelligence to strengthen the qualitative dimension in life and derive motivation for action from it, because: "Most of us know or suspect quite precisely in our innermost being what would be worthwhile and meaningful and what would not. What seems to be lacking so far, however, is the knowledge of how to live in a concrete and realistic value- and meaning-oriented committed way; and also, the knowledge that meaning-oriented, responsible action not only enriches the world, but also ourselves."[63] This also sets limits to a consequentialist way of thinking, which can be overcome through sustainable action, which lies in moderating people's consumption and behavior. This builds a bridge from the sense dimension to an environmental virtue ethics approach, which has so far appeared too vaguely in environmental discourse. For a basic conception of virtue ethics approaches lies in the fact that a person develops himself or herself towards virtues or actions that are recognized as meaningful. Insights into

---

[59] Hengstenberg, *Sinn und Sollen: Zur Überwindung der Sinnkrise*, 49.
[60] Hengstenberg, *Sinn und Sollen: Zur Überwindung der Sinnkrise*, 50.
[61] Hengstenberg, *Sinn und Sollen: Zur Überwindung der Sinnkrise*, 50.
[62] Thomas Merton, *No Man is an Island* (New York & London: A Herevest/HBJ Book, 1955), 122.
[63] Batthyány, *Die Überwindung der Gleichgültigkeit. Sinnfindung in einer Zeit des Wandels*, 26.

the meaning of these virtues then guide individual action, as they evoke immediate personal concern. This can positively complement the lamentation about either the system, capitalism or large corporations that has accompanied the environmental discourse for decades with an insight into individual agency. In this way, the individual escapes a victim role and gains personal responsibility and from this another source for the good life.

Outlook

AI can help to present consequentialist approaches to environmental assessment at the political level. This can be used to determine the consequences of action and to set appropriate limits for resource use. The limits of AI in overcoming the ecological crisis lie in the fact that it remains rooted in the purely quantitative. However, the qualitative dimensions of human life cannot be simulated. This also applies to the contribution of the natural sciences because the conception of nature that still united empirical natural science and aesthetic enjoyment of nature was still present in Alexander von Humboldt's (1769-1859) work, but has been lost in more recent natural science, and this divisiveness appears to be intensified by AI. An environmental virtue ethics approach that builds on overcoming the modern tendency to divide man and nature can achieve a new appreciation for the environment and derived from this, an increased commitment to it in regular encounters with nature.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Dreyfus on AI: A Lonerganian Retrieval and Critique

Michael Sharkey
University of Wisconsin-Platteville, USA

Abstract
Hubert Dreyfus develops a critique of AI which should interest readers of Bernard Lonergan. He contests its early rationalism in a way that resembles Lonergan's critique of conceptualism. He contests its early representationalism in a way that resembles Lonergan's critique of ocularism. And he makes both criticisms from a cognitional-theoretical perspective which privileges "insight," like Lonergan's. However, Dreyfus ultimately gives short shrift to consciousness, intentionality, and acts, which leads him to throw out the mentalist baby with the conceptualist and ocularist bath. The result is an excessive receptivity to recent (especially neural network) AI, which reduces intelligence to electrical events.

Keywords: Hubert Dreyfus, Bernard Lonergan, artificial intelligence, insight, problems of consciousness.

Introduction

In publications running from *What Computers Can't Do* (1972, 1978) through *Mind Over Machine* (1986) to *What Computers* Still *Can't Do* (1992) and "Why Heideggerian AI Failed" (2007), Hubert Dreyfus develops a critique of artificial intelligence that should interest readers of Lonergan.[1] He shows first variants of the project to possess rationalist philosophical presuppositions and criticizes them in ways that resemble Lonergan's critique of conceptualism. He shows second variants to be in the grips of a representational theory of knowledge and criticizes them in ways that resemble Lonergan's critique of ocularism. And he offers both sets of critique from out of his own cognitional-theoretical perspective, centered as it is on what he entitles "insight."[2]

However, Dreyfus's stance is not fully positional, and this compromises his critique of AI.[3] His method sits uneasily between phenomenology and metaphysics, in the manner of the early Heidegger and Merleau-Ponty. This leads him to give short shrift to consciousness, intentionality, and acts, which in turn leads him to throw out the mentalist baby with the conceptualist and ocularist bath. The result is an undue receptivity to recent (neural network) AI, which reduces intelligence to electrical events.[4]

Both a retrieval and a critique, then, would seem to be in order. In a first part below, I will relate Dreyfus's interpretation and critique of AI, in both its early and more recent variations. In a second, I will explain why I think much of his treatment is consistent with a positional stance. And in a third, I will explain why I think some of his (counter) positions stand in need of reversal.

---

[1] Hubert L. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason* (New York: Harper Collins, 1972, 1978), *Mind over Machine* (New York: Free Press, 1986), *What Computers* Still *Can't Do* (Cambridge, MA: MIT Press, 1992), and "Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian," in Mark A. Wrathall, ed, *Skillful Coping: Essays on the Phenomenology of Everyday Perception & Action* (Oxford: Oxford University Press, 2016).

[2] Lonergan's masterwork is *Insight: A Study of Human Understanding,* volume 3 of *Collected Works of Bernard Lonergan*, ed. Frederick E. Crowe and Robert M. Doran (Toronto: University of Toronto Press, 1992).

[3] A stance is "positional," for Lonergan, if it cannot be denied without performative contradiction. See Bernard Lonergan, *Insight*, 313-15. And for a rebuttal of the charge that the doctrine is question-begging, see Mark D. Morelli, "Reversing the Counter-Position: The *Argumentum ad Hominem* in Philosophic Dialogue," in Frederick Lawrence, ed., *Lonergan Workshop*, volume 6 (Macon, Georgia: Scholars Press, 1986), 195-230.

[4] I owe the important distinction between an act and an event in this context to Elizabeth Murray.

I. Dreyfus on AI

(A) Early

In *What Computers* Still *Can't Do* and *Mind over Machine*, Dreyfus shows early variants of the project of AI to possess rationalist philosophical presuppositions. The presuppositions derive from the epistemological programs of Socrates, Descartes, Hobbes, Leibniz, Kant, and Husserl, and tell us that intelligence is a matter of representations and rules.

For Dreyfus, Socrates is a semantic rationalist. He demands that Euthyphro tell him " . . . what is characteristic of piety which makes all actions pious . . . that I may have it to turn to, and to use as a standard whereby to judge your actions and those of other men."[5] Uninterested in this or that example, as rooted in Athenian culture, he requires a general concept or universal definition articulating the necessary and sufficient conditions of the virtue. With one in hand, he might avoid the contingency and imprecision which characterize practical reason. Or so he thinks. He is thus the distant inspiration for AI's "effective procedure" or "set of rules which tells us, from moment to moment, precisely how to behave."[6]

Things are little different with Descartes, Kant, and Husserl. Descartes claims that one can "analyze any problem into its basic, isolatable elements, and explain the complex in terms of rule-like combinations of such primitives."[7] Thus he intuits with certainty that he thinks, deduces that he exists and is a thinking thing, and proceeds therefrom to build up an edifice of new knowledge. Kant holds that "all concepts are really rules,"[8] shows some necessarily to apply to objects of knowledge, and establishes a tribunal of pure reason. Husserl takes concepts to be "hierarchies of rules, rules which contain other concepts under them," and so shows himself to be "father of the information-processing model of the mind."[9]

Things are different, and yet the same, with Hobbes and Leibniz. They are not semantic but syntactic rationalists who would reduce "all . . . appeal to meanings . . . to the techniques of . . . formal . . . manipulation."[10] But they continue to think of intelligence in terms of representations and rules. "When a man *reasons*," Hobbes says, "he does nothing else but conceive a sum total from addition of parcels, for REASON . . . is nothing but reckoning."[11] And Leibniz develops a "universal and exact system

---

[5] Plato, *Euthyphro*, VII, trans. F. J. Church (New York: Library of Liberal Arts), 1948, 7, as quoted in Dreyfus *What Computers* Still *Can't Do*, 67.
[6] Marvin Minsky, *Computation: Finite and Infinite Machines* (Englewood Cliffs, N.J.: Prentice-Hall, 1967), 106, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 67.
[7] Hubert Dreyfus, *Mind over Machine*, 3. Italics removed.
[8] Hubert Dreyfus, *Mind over Machine*, 4.
[9] Hubert Dreyfus, *Mind over Machine*, 4.
[10] Hubert Dreyfus, *What Computers* Still *Can't Do*, 69. Parentheses removed.
[11] Thomas Hobbes, *Leviathan* (New York: Library of Liberal Arts, 1958), 45, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 69.

of notation, an algebra, a symbolic language" to which concepts can be reduced. On their basis "and the rules for their combination all problems [can] be solved and all controversies ended."[12] Leibniz writes that if someone were to contest his results, he would say to him, "'Let us calculate, Sir,' and thus by taking pen and ink, we should settle the question."[13]

Semantic and syntactic rationalism drive early AI. The successor to the latter, Cognitive Simulation, means "to reproduce the steps by which human beings actually proceed," whereas the successor to the former, Semantic Information Processing, means just to achieve the same results.[14] But between them they take concepts to be rules, of a kind, or to be formal stand-ins for meanings which, when manipulated by rules, produce intelligence. They thus incarnate the commitment to representations and rules.

Among examples of Cognitive Simulation, Dreyfus considers programs for playing games, translating languages, solving problems, and recognizing patterns. Among examples of Semantic Information Processing, he considers programs for understanding language and finding analogies.

Newell and Simon's program for playing chess is a fine example of Cognitive Simulation. Chess is a game in which pieces of varying capacity are moved across a board in a rule-like way to achieve certain ends. Intelligent play involves finding the best means of achieving those ends. So a computer program for playing chess must include at least representations (or definitions) of the pieces and a list of the rules for manipulating them. But it must include more, for of course there is a difference between intelligent and unintelligent manipulation. Enter what Newell and Simon call "heuristics," or "rules of practice," or "rules of thumb," gleaned from the greats. These are not rules followed invariably but just occasionally in order to reduce calculation. They are "aids to discovery" meant to replicate the judgment *in situ* that is characteristic of human play.[15]

Another example of Cognitive Simulation is Oettinger's Russian-English dictionary. On one understanding of how language works, such as is to be found in Augustine's *Confessions* and Wittgenstein's *Tractatus*, there is a one-to-one correspondence between words and things, or sets of words and states of affairs. A dictionary translating from one language to another, then, must exhaustively correlate the more or less complex correspondences on each side. "It was soon clear that a mechanical dictionary could easily be constructed in which linguistic items, whether they were parts of words, whole words, or groups of words, could be processed independently and converted one after another into corresponding items in another

---

[12] Hubert Dreyfus, *What Computers* Still *Can't Do*, 69.
[13] Leibniz, *Selections*, ed. Philip Wiener (New York: Scribner, 1951), 18, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 69.
[14] Hubert Dreyfus, *What Computers* Still *Can't Do*, 85.
[15] Hubert Dreyfus, *What Computers* Still *Can't Do*, 74-77, 94, 102-107.

language."[16] In this way it was thought the difficulties in understanding a foreign tongue could be reduced to low-level matching.

A striking example is Newell, Simon, and Shaw's General Problem Solver, which sought "rules for converting any sort of intelligent activity into a set of instructions." But again, because studies showed subjects "tended to use rules or shortcuts which were not universally correct, but which often helped," heuristics were employed. If, in solving logic problems, "[s]uch a rule of thumb might be, . . . try to substitute a shorter expression for a longer one,"[17] or if, in playing chess, it might be "maintain center position" or "sacrifice queen,"[18] in this context it was held that by generalizing such strategies the human capacity for solving problems in any area could be mimed.

> In short, we now have the elements of a theory of heuristic (as contrasted with algorithmic) problem-solving; and we can use this theory both to understand human heuristic processes and to simulate such processes with digital computers. Intuition, insight, and learning are no longer exclusive possessions of humans; any large high-speed computer can be programmed to exhibit them also.[19]

Last examples of Cognitive Simulation come from pattern recognition. They are programs for transliterating hand-sent Morse code, as well as for "recognizing a limited set of handwritten words and printed characters in various type fonts."

> These all operate by searching for predetermined topological features of the characters to be recognized, and checking these features against preset or learned "definitions" of each letter in terms of these traits. The trick is to find relevant features, that is, those that remain generally invariant throughout variations of size and orientation, and other distortions.[20]

Here, the human capacity to discern according to necessary and sufficient conditions is modelled.

Turning to Semantic Information Processing, Bobrow's STUDENT program is exemplary. It makes no pretense to the humanoid, but still solves algebra word problems and "understands English."[21]

---

[16] Hubert Dreyfus, *What Computers* Still *Can't Do*, 91.
[17] Hubert Dreyfus, *What Computers* Still *Can't Do*, 75.
[18] Hubert Dreyfus, *What Computers* Still *Can't Do*, 101-102.
[19] Herbert A. Simon and Allen Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research*, Vol. 6 (January—February, 1958), 6, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 77.
[20] Hubert Dreyfus, *What Computers* Still *Can't Do*, 97.
[21] According to Marvin Minsky, in his "Artificial Intelligence," *Scientific American,* Vol. 215, No. 3 (September 1966), 257, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 132.

> The program simply breaks up the sentences of the story problem into units on the basis of cues such as the words "times," "of," "equals," etc.; equates these sentence chunks with x's and y's; and tries to set up simultaneous equations. . . . [T]he . . . scheme works . . . because there is the constraint, not present in understanding ordinary discourse, that the pieces of the sentence, when represented by variables, will set up soluble equations.[22]

In other words, the program reduces typical human expression to algebraic formalism and rules.

A final example of Semantic Information Processing is Evan's Analogy Finder. It does not purport to reproduce human intelligence any more than does Bobrow's STUDENT, yet it too is set out in mentalistic terms. "Given a set of figures, [the program] constructs a set of hypotheses or theories as follows." First, a description of figure A may be transformed into one for B. Second, the parts of A may be set into correspondence with the ones for C, suggesting a relation like the first, but now relating C and other figures. Third, the differences between C and another figure may be reduced to the same degree as between A and B, so that, Fourth, it may be determined that A stands to B as C does to, say, D3, this having been determined by measurement.[23] Evans's editor even adds that he feels sure "rules or procedures of the same general character are involved in any kind of analogical reasoning."[24]

Now, Dreyfus does not take any of these programs to rise to the level of intelligence. He takes the examples from Cognitive Simulation to fail to do so because they do not employ "fringe consciousness," "contextually disambiguate," "distinguish the essential from the inessential," and "perspicuously group," as do all human beings when behaving intelligently. And he takes the examples from Semantic Information Processing to fail to do so because they do not have "bodies," are not "in situations," and do not have "needs."[25] He offers a hermeneutic-phenomenological argument for the view that human intelligence involves more than rule-following and representing.

As against Newell and Simon's program for playing chess, Dreyfus points out that human beings do more than count out possible moves and responses and occasionally employ rules of thumb. For "[a]lternative paths multiply so rapidly that we cannot . . . run through all the branching possibilities" and it is necessary not just to "look . . . every once in a while for a Queen sacrifice but . . . look in those situations in which such a sacrifice is relevant."[26] For this, "fringe consciousness" is required. It

---

[22] Hubert Dreyfus, *What Computers* Still *Can't Do*, 133.
[23] Marvin Minsky, ed., *Semantic Information Processing* (Cambridge, Mass.: M.I.T. Press, 1969), 16, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do,* 139.
[24] Marvin Minsky, "Artificial Intelligence," *Scientific American,* Vol. 215, No. 3 (September 1966), 250, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 139. Dreyfus's italics removed.
[25] In fact, what Dreyfus says here applies to programs from Cognitive Simulation too. But since being embodied, being in situations, and having needs are central to his overcoming of representationalism, his primary target is programs for Semantic Information Processing, with their emphasis on representations more than rules.
[26] Hubert Dreyfus, *What Computers* Still *Can't Do*, 101.

is "marginal awareness" that "concentrate[s] information concerning our peripheral experience."[27] In virtue of it, promising areas of the board may be identified.

Consider the following player's report. "Again I notice that one of his pieces is not defended, the Rook, and there must be ways of taking advantage of this. Suppose now, if I push the pawn up at Bishop four, if the Bishop retreats I have a Queen check and I can pick up the Rook."[28] At the end, Dreyfus notes, "we have an example of … "counting out"—thinking through the various possibilities by brute force enumeration." But at the start, we have something very different, a kind of sussing out, perhaps. "[T]he subject "zeroed in" on the promising situation."[29]

As against Oettinger's program for machine translation, Dreyfus calls attention to context. It invariably produces ambiguity in expression, which makes one-to-one translation difficult. It therefore turns out that "in order to translate a natural language, more is needed than a mechanical dictionary—no matter how complete—and the laws of grammar—no matter how sophisticated." For "[t]he order of the words in a sentence does not provide enough information to enable a machine to determine which of several possible parsings is the appropriate one, nor do the surrounding words—the written context—always indicate which of several possible meanings . . . the author had in mind."[30] What is required is "contextual disambiguation."

"A phrase like 'stay near me,'" Dreyfus writes, "can mean anything from 'press up against me' to 'stand one mile away,' depending upon whether it is addressed to a child in a crowd or a fellow astronaut exploring the moon."[31] And human beings can determine which is which. Again, a child can learn the names of things without being unduly thwarted by situational change. "It is this ability to grasp the point in a particular context which is true learning; since children can and must make this leap, they can and do surprise us and come up with something genuinely new."[32]

As against Newell, Simon, and Shaw's program for general problem solving, Dreyfus presses this point about getting the point. "[I]nsight," he declares, "has proved intractable to stepwise programs such as Simon's General Problem Solver."

> If a problem is set up in a simple, completely determinate way, with an end and a beginning and simple, specifically defined operations for getting from one to the other, . . . then Simon's General Problem Solver can, by trying many possibilities, bring the end and the beginning closer and closer together until the problem is solved.[33]

---

[27] Hubert Dreyfus, *What Computers* Still *Can't Do*, 103.
[28] Allen Newell and H. A. Simon, *Computer Simulation of Human Thinking*, The RAND Corporation, P-2276 (April 20, 1961), 15, as quoted in Hubert Dreyfus, *What Computers* Still *Can't Do*, 102.
[29] Hubert Dreyfus, *What Computers* Still *Can't Do*, 102.
[30] Hubert Dreyfus, *What Computers* Still *Can't Do*, 107.
[31] Hubert Dreyfus, *What Computers* Still *Can't Do*, 108.
[32] Hubert Dreyfus, *What Computers* Still *Can't Do*, 111.
[33] Hubert Dreyfus, *What Computers* Still *Can't Do*, 112.

Or it can do so in concert with heuristics. But when the problem is complex more than slavish rule-following is required and the heuristics themselves can be seen to be nothing more than that. This is borne out by analysis of the reports given by human beings while they are solving problems.

Consider the example of one such 'protocol' given by a person solving a problem in logic. In it he reports that having received a list of rules for transforming symbolic expressions, he applied "the rule (A · B → A) and the rule (A · B → B), to the conjunction (— R v — P) · (R v Q)." Newell and Simon note that in so doing he "handled both forms of rule 8 together," whereas their machine "took a separate cycle of consideration for each form." But they assume that the subject "covertly" took each form in turn, while Dreyfus notes that, on the face of it, he "grasped the conjunction as symmetric with respect to the transformation operated by the rule, and so in fact applied both forms of the rule at once." That is, Dreyfus shows that the phenomenological evidence suggests the subject had an insight. He was able to "discriminate between occasions when it is was appropriate to apply both forms of the rule at once and those occasions when it was not."[34]

Again, "[a]t a certain point, the protocol reads: " . . . I should have used rule 6 on the left-hand side of the equation. So use 6, but only on the left-hand side." Simon sees that "[h]ere we have a strong departure from the GPS trace," for "[n]othing exists in the program that corresponds to this." And "[t]he most direct explanation," he avers, "is that the application of rule 6 in the inverse direction is perceived by the subject as undoing the previous application of rule 6." He seems to recognize the act of insight. But he does not see that this counts against his approach.[35]

Part of the explanation for this must be that Newell and Simon think they have covered the phenomenon of insight with heuristics. Such aids in discovery are supposed to take the program beyond the automatic to the selective, but in fact they just take it beyond the invariant to the occasional. And the programmers determine what counts as occasional. It is this "insightful predigesting of their material" that enables them to pass off as intelligent what is just mechanical.[36]

Lastly, as against the programs for pattern recognition, Dreyfus contests the primacy of the concept. "A computer must recognize all patterns in terms of a list of specific traits," he notes. And "in simple cases artificial intelligence workers have been able to make some headway with mechanical techniques." But "patterns as complex as artistic styles and the human face reveal a loose sort of resemblance which seems to require a special combination of insight, fringe consciousness, and ambiguity tolerance beyond the reach of digital machines."[37] This Dreyfus calls "perspicuous grouping."

Consider even the apparently simple task of identifying a shape. How do we do it? We are not, most of us, like aphasics, who "can only recognize a figure such as

---

[34] Hubert Dreyfus, *What Computers* Still *Can't Do*, 113.
[35] Hubert Dreyfus, *What Computers* Still *Can't Do*, 113-114.
[36] Hubert Dreyfus, *What Computers* Still *Can't Do*, 119.
[37] Hubert Dreyfus, *What Computers* Still *Can't Do*, 120.

a triangle by listing its traits, that is, by counting its sides and then thinking: 'A triangle has three sides. Therefore, this is a triangle.'" We do not need to "conceptualize . . . the traits common to several instances of the same pattern in order to recognize that pattern."[38] We do not need to employ a classification rule. Instead, we zero in on relevance and grasp the point in a context, irrespective of some ambiguity.

We can see that Dreyfus's main reservation about Cognitive Simulation is its emphasis on rule-following. By contrast, his main reservation about Semantic Information Processing is its emphasis on semantics. But for Dreyfus "semantics" always has to do with "representation," and we will be able better to see his critique of it if we turn to material beyond Bobrow's STUDENT and Evans's Analogy Finder.

It is true that we must use our bodies in order to see, hear, taste, touch and smell. In the language of early AI theorist Marvin Minsky, such "meat machine" operation is essential. But it is not sufficient, according to Dreyfus, for we must also use our "lived bodies" to get at meanings. We do not just receive sense-impressions, re-present those presentations to ourselves, and string the representations together to form ideas and thoughts. We are aware of ourselves as sensing, and indeed as seeking understanding, which supplies us with a "global anticipation" in whose light we make sense of parts.[39]

For example, "in recognizing a melody, the notes get their values by being perceived as part of the melody, rather than the melody's being recognized in terms of independently identified notes." Similarly, the "hazy layer which I would see as dust if I thought I was confronting a wax apple might appear as moisture if I thought I was seeing one that was fresh."[40] My gulp of milk will leave me disoriented if what I was expecting was water.[41] And I will be unable to identify silk as silk, if I lack the appropriate anticipations developed in me by long familiarity with fabric.[42] It is only because I am anticipatorily involved with my world, that I am able to understand any bit of it. But machines lack embodiment, and so lack the condition of the possibility of understanding.

Again, it is because I am in situations that I am able to affix meanings correctly. On a walk I know that my friend's gesture towards "the Old Man of the Woods" refers to a plant and not a person.[43] In front of a pet store I know my daughter's desire for "it" refers to a doggie and not the window.[44] In hearing from a gift-giver that I "can take it back if I already have one," I know he means the item he has given and not the one I may already have.[45] And in a Berkeley restaurant, I know the suggestion to "order

---

[38] Hubert Dreyfus, *What Computers* Still *Can't Do*, 123.
[39] Hubert Dreyfus, *What Computers* Still *Can't Do*, 237.
[40] Hubert Dreyfus, *What Computers* Still *Can't Do*, 238.
[41] Hubert Dreyfus, *What Computers* Still *Can't Do*, 242.
[42] Hubert Dreyfus, *What Computers* Still *Can't Do*, 249.
[43] I feel sure this example, borrowed from Wittgenstein, is in one of Dreyfus's texts. But I am unable to find it.
[44] Hubert Dreyfus, *What Computers* Still *Can't Do*, xix.
[45] Hubert Dreyfus, *What Computers* Still *Can't Do*, 57.

anything" does not include the chef.[46] My ability to understand depends on familiarity with situations and their criteria. But machines are not in situations, and so they cannot "compute."

Finally, both my embodiment and being-in-situations are tied up with needs. It is because I require nourishment, both physical and aesthetic, that I listen to melodies, look at apples, drink water, and touch silk. And it is because I need love and friendship that I walk with friends, spend time with my daughter, have birthday parties, and go to restaurants. My ability to understand, therefore, is rooted not just in embodiment and being in situations, but in the needs which drive me to both. And yet computers do not have needs any more than are they embodied or in situations. This is another reason why they are blocked from cognition.[47]

In summary, Dreyfus criticizes early AI because it models intelligence on representations and rules. Its first variant, Cognitive Simulation, emphasizes rule-following, and so ignores the fringe consciousness, contextual disambiguation, insight, and perspicuous grouping which are essential to the real article. And its second variant, Semantic Information Processing, emphasizes representations, and so ignores the embodied anticipation, situational sensitivity, and neediness which are the conditions of representation. Both variants are indebted to the rationalist tradition in Western philosophy, against which Dreyfus would set Heidegger, Wittgenstein, and Merleau-Ponty.[48] However, as we will see, this surprisingly does not stop him from endorsing AI of a kind.

(B) Recent

Dreyfus is more sanguine about the prospects for recent, neural network AI, and this precisely because it does not employ representations and rules. Instead of trying to make a mind, as at least Cognitive Simulation did, it seeks to model the brain; and Dreyfus believes it is partly on its way.

In *What Computers* Still *Can't Do*, Dreyfus argues that "we should set about creating artificial intelligence by modelling the brain's learning power rather than the mind's symbolic representation of the world" because of what we have learned from neuroscience. Already in the '50's that discipline had suggested that "a mass of neurons could learn if the simultaneous excitation of neuron A and neuron B increased the strength of the connection between them." In the present, then, AI might "attempt to automate the procedures by which a network of neurons learns to discriminate patterns and respond appropriately."[49] But how? "[A] designer could tune a simulated

---

[46] Hubert Dreyfus, *What Computers* Still *Can't Do*, 311, note 102.
[47] Hubert Dreyfus, *What Computers* Still *Can't Do*, 276-280.
[48] Hubert Dreyfus, *What Computers* Still *Can't Do*, 212, 233. And see Dreyfus, *Mind Over Machine*, 4-5, 7 and 11.
[49] Hubert Dreyfus, *What Computers* Still *Can't Do*, xiv.

multilayer perceptron (MLP) neural network by training it to respond to specific situations and then having it respond to other situations in ways that are (the designer hopes) appropriate extrapolations of the responses it has learned." In this case the modeler "provides not rules relating features of the domain but a history of training input-output pairs, and the network organizes itself by adjusting its many parameters so as to map inputs into outputs, situations into responses."[50]

Consider a famous example. In order better to wage the Gulf War, a neural net was trained to distinguish rocks from mines at the bottom of a sea. First, visual and sonar data on these items was assembled. Second, our (or our brain's) ability to identify patterns in this data was modelled by "input and output nodes," "middle layer nodes," and the variable strengths of their relations expressed as "weights." Third, an expert at identifying and distinguishing rocks and mines "tuned" the network of nodes-in-their-relations (adjusted their relative strengths) to correspond to that obtaining in the world. And fourth, the network was afterwards able to discriminate on its own.[51]

Dreyfus even argues this approach is consistent with phenomenology. In "Merleau-Ponty and Recent Cognitive Science," he draws a parallel between understanding as Merleau-Ponty conceives of it and understanding as modelled by neural nets. Just as, for Merleau-Ponty, "the life of consciousness—cognitive life, the life of desire or perceptual life—is subtended by an 'intentional arc,' which projects round about us our past, our future, our human setting,"[52] and so establishes a "dialectic of milieu and action," so for neural net AI "past experience with a large number of cases . . . modifies the weights between the simulated neurons, which in turn determine the response." In neither case is there need to "represent . . . past experience as cases or rules for determining further action," and in both it is thus possible "to avoid the problem . . . concerning how to find the *relevant* rule."[53]

Again, in "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," Dreyfus likens understanding as Heidegger conceives of it to Freeman's Neural Dynamics. For Heidegger, understanding is an affair of practical know-how, of knowing one's way around in the world. It is "more basic than *thinking* and solving problems" and is "not representational at all." In fact, in understanding at our best, "we are drawn in by solicitations and respond directly to them, so that the distinction between us and our equipment vanishes."[54] "I *live* in the

---

[50] Hubert Dreyfus, *What Computers* Still *Can't Do*, xv.

[51] R. Paul Gorman and Terence J. Sejnowski, "Learned Classification of Sonar Targets Using a Massively Parallel Network," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 36/7 (July 1988), 1135-40, referenced in Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," Mark Wrathall, ed., *Skillful Coping,* 238, note 12.

[52]Maurice Merleau-Ponty, *Phenomenology of Perception*, tr. Colin Smith (London: Routledge Classics, 2002), 136, as quoted in Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 234.

[53] Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 236.

[54] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 258.

understanding of writing, illuminating, going-in-and-out, and the like," Heidegger says. And "[my] being in the world *is* nothing other than this . . . understanding."[55]

It is much the same in Freeman's dynamics. He "proposes a model of rabbit learning based on the coupling of . . . brain and . . . environment," and to the degree that these remain distinct they stand in circular relation. The rabbit is thrown on to a horizon of longing. It "sniffs around until it falls upon food, a hiding place, or whatever else it . . . needs." Its "neural connections are then strengthened to the extent that" it is satisfied. And its new configuration of synapses-in-relation contextualizes further desire.[56] No representations or rules are required. Only a kind of natural analogue of the hermeneutic circle.

Or so it might seem. However, Dreyfus is alert to some limitations of neural modelling. As against the (putatively) Merleau-Pontyan version, he argues that the problem of relevance resurfaces. "When a net is trained by being given inputs paired with appropriate responses," he writes, "the net can only be said to have learned to respond appropriately when it responds appropriately to *new* inputs similar to, but different from, those used in training it." Otherwise, it may seem just to have engaged in the low-level matching characteristic of GOFAI. Yet, in any given instance, there will be many different candidates for "similar to," and even different candidates for the relevant sort(s) of similarity. So the net designer will have to set parameters.[57]

Likewise, there is a problem with Freeman's dynamics. For "to program Heideggerian AI, we would not only need a model of the brain functioning underlying coupled coping, . . . but . . . a model of our particular way of being embedded and embodied such that what we experience is significant for us in the particular way that it is."[58] We would need a model of ourselves in all our materiality, and not just our brains. And failing this, "Heideggerian AI can't get off the ground."[59]

In summary, then, Dreyfus is hopeful and hesitant about neural modelling. He is hopeful about both versions we have considered because they seem to proceed without representations and rules. But he is hesitant about the first because it requires help from the net designer, and he is hesitant about the second because it seems focused on brains and nut full persons. It is noteworthy, however, that for him there does not seem to be any in-principle block to the latter approach: it might well just be a matter of time and labor before we model the human brain and body. By contrast, the typical neural net procedure seems subject to the "insoluble problem of a

---

[55] Martin Heidegger, *Logic: The Question of Truth*, tr. Thomas Sheehan (Studies in Continental Thought: Bloomington, IN: Indiana University Press, 2010), 121, as quoted by Hubert Dreyfus in "Why Heideggerian AI Failed," 258-59.

[56] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 263.

[57] Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 236.

[58] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 272.

[59] Hubert Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," 273.

disembodied mind responding to what is relevant."[60] In time, we will see that Lonergan can offer resources for transcending such difficulties. But for now, let us notice how much in Dreyfus he can affirm.

II. A Lonerganian Retrieval

(A) Early

To very much in Dreyfus's critique of early AI, Lonergan can utter a resounding "yea." For the most part, this is because of their similar understandings of understanding. Dreyfus's fringe consciousness, contextual disambiguation, insight, and perspicuous grouping remind one of Lonergan's patterns of experience, transcendental intention, insight, and anti-conceptualism. And Dreyfus's embodiment, situations, and needs remind one of Lonergan's anti-ocularism, history, and carnality. Let us briefly consider their affinities.

Fringe consciousness, as we saw, is a tacit awareness that human beings possess but computers do not, and by which they zero in on relevance. It is not yet the grasp of relevance, but something which makes it possible, and is in this way like Lonergan's patterns of experience. These organize and direct the flow of conscious awareness in biological, dramatic, aesthetic, or intellectual ways, and render it selective. This prepares the mind to identify specific relevance.[61]

Contextual disambiguation, of course, is the overcoming of ambiguity due to context. It permits us, but not computers, to reach beyond the confines of variable situations and get things right. In this way, it is like Lonergan's transcendental intention, which intends not this or that meaning datum, but intelligibility per se, and so supplies a criterion in terms of which to advance.[62]

Insight, for Dreyfus, is that by which we do advance, or grasp relevance, or distinguish the essential from the inessential, in a situation. It is thus the same as or similar to what Lonergan means by the same term. For him, insight is the grasp of intelligibility in the concrete, as prepared for by the patterning of experience and transcendental intention. It is the understanding that defines us as human beings and places us beyond machines, among else.[63]

Perspicuous grouping, we may recall, is that combination of fringe consciousness, contextual disambiguation, and insight by which we approach intelligibility pre-conceptually. If eventually, we do classify, and express our understanding in terms of lists of necessary traits, we do not begin there, as does a

---

[60] Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 236.
[61] Bernard Lonergan, *Insight,* 204-212.
[62] Bernard Lonergan, *Insight,* 372-398.
[63] Bernard Lonergan, *Insight, passim.*

computer. And this marks another affinity with Lonergan, for whom understanding drives conception, and not the other way around. This is his anti-conceptualism.[64]

When it comes to embodiment, we are reminded of Lonergan's anti-ocularism. Or, we are reminded of his anti-representationalism, which is implicit in his anti-ocularism. For Dreyfus, we do not take in the presentations of sense, re-present these to ourselves, and try to mirror the world with our minds. Instead, our lived body anticipates wholes in the light of which we identify parts, and this supplies the field on which distinctions between subject and object occur. Likewise, for Lonergan, our understanding is not an affair of seeing what is out there, set out over against us, but of increasingly making good on our in-built orientation to the transcendentals, understood as essence, existence, and good. And this too is the condition of any encounter or confrontation.[65]

Again, Dreyfus's situations remind us of Lonergan's history, or commitment to historicity. If, for Dreyfus, it is in part the situated character of the human knower that permits her to know how to go on in situ, so for Lonergan it is in part her embeddedness in history that enables her to do so. For we do not, like computers, purport to operate *sub specie aeternitatis*, but inhabit time.[66]

Finally, Dreyfus's needs call to mind Lonergan's insistence on our carnality. For Dreyfus, not only must we meet the physiological demands of sight, hearing, taste, touch, and smell, but these drive us to reach out to nature, family, friends and society more generally. For Lonergan, too, the exigencies of neural demands, and the like, propel us beyond the biological to the dramatic, aesthetic, common sensical and intellectual. He is a soft, and not a hard, dualist, we might say.[67]

(B) Recent

To a much lesser degree, can Lonergan affirm Dreyfus's criticisms of recent, neural AI. But this is only because he would press them more strongly and add to them. In part III, section (b) below, we will see that and how this is so. Here, let us try simply to identify what Lonergan can admire.

In an early work, Lonergan writes that "With remarkable penetration Aquinas refused to take as reason the formal affair that modern logicians invent machines to

---

[64] Bernard Lonergan, *VERBUM: Word and idea in Aquinas,* volume 2 of *Collected Works of Bernard Lonergan*, ed. Frederick E. Crowe and Robert M. Doran (Toronto: University of Toronto Press, 1997).
[65] Bernard Lonergan, "Cognitional Structure," in *Collection*, volume 4 of *Collected Works of Bernard Lonergan*, ed. Frederick E. Crowe and Robert M. Doran (Toronto: University of Toronto Press, 1993), 205-221.
[66] Bernard Lonergan, "The Transition from a Classicist World-View to Historical-Mindedness," in W. J. F. Ryan and Bernard J. Tyrrell, eds., *A Second Collection* (Toronto: University of Toronto Press, 1996), 1-9.
[67] Bernard Lonergan, *Insight,* 204-212.

perform."[68] And he gives a painful example of who we can become if we do not do the same.

> A sergeant-major with his manual-at-arms by rote knows his terms, his principles, his reasons; he expounds them with ease, with promptitude, and perhaps with pleasure; but he is exactly what is not meant by a man of developed intelligence. For intellectual habit is not possession of the book but freedom from the book. It is the birth and life in us of the light and evidence by which we operate on our own. It enables us to recast definitions, to adjust principles, to throw chains of reasoning into new perspectives according to variations of circumstance and exigencies of the occasion.[69]

The passages make clear Lonergan's pity for the dependence and rigidity of early AI, but suggest a possible openness on his part to the learning and flexibility of more recent variants. If indeed this is what they possess. The difficulty, of course, is that Dreyfus is not at all sure that they do.

Recall Dreyfus's account of the problem of similarity and its would-be solution in designer parameters. "All neural net modelers," he writes, "agree that for a net to be intelligent it must be able to generalize; that is, given sufficient examples of inputs associated with one particular output, it should associate further inputs of the same type with that same output." But what, he asks, counts as the same type? "The designer of the net has in mind a specific definition of the type required for a reasonable generalization and counts it a success if the net generalizes to other instances of this type." In other words, the task of abstraction falls to the designer, not the net.[70]

A similar point is made by an exponent of Lonergan in the philosophy of law. In the law, of course, we must not only abstract in order to determine initial law, but abstract again in order to apply it. And this re-raises the problem of similarity. For an application must be legitimate, and not just arbitrary. Yet for it to be legitimate, it must regard a case which is similar to the original in relevant respects. Thus, "application of our habitual insight to any particular concrete case always involves a further insight, at least the insight that this situation is the same as the original."[71] And such an insight does not seem to be the province of computers any more than of law tables.

Again, Lonergan can affirm Dreyfus's critique of Freeman's neural dynamics, although it does not go nearly far enough. For if the latter models brain, but not full nervous function, it may well be incomplete as a model of intelligence, even if it is

---

[68] Bernard Lonergan, *Verbum*, 71.
[69] Bernard Lonergan, *Verbum*, 193-194.
[70] Hubert Dreyfus, "Making a Mind versus Modelling the Brain: Artificial Intelligence Back at a Branchpoint," in Mark Wrathall, ed., *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action* (Oxford: Oxford University Press, 2014), 229.
[71] Frederick E. Crowe, "Law and Insight," in Michael Vertin, ed., *Developing the Lonergan Legacy: Historical, Theoretical, and Existential Themes* (Toronto: University of Toronto Press, 2004), 271.

more so in virtue of its inattention to consciousness, intentionality, and acts, and the difference between a model and what it models.[72]

III. A Lonerganian Critique

(A) Early

As we have seen, Lonergan can affirm much in Dreyfus's critique of early AI, and some in his critique of more recent variants. However, not even the former would meet with his full approval. The reason, again, is to do with cognitional theory. If Dreyfus's doctrines of fringe consciousness, contextual disambiguation, insight, and perspicuous grouping resemble Lonergan's patterns of experience, transcendental intention, insight, and anti-conceptualism, and his strictures regarding embodiment, situations, and needs resemble Lonergan's regarding anti-representationalism, history, and carnality, his account of insight is by Lonergan's standards nevertheless incompletely differentiated. And this fuels in him an undue receptivity to recent, neural AI, as we will soon see.

The "insight" which Dreyfus brings to bear against early AI is a "grasp of . . . essential structure."[73] It is an exercise of "the ability to distinguish the essential from the inessential . . . necessary for learning and problem solving, yet not amenable to the mechanical search techniques which . . . operate once this distinction has been made."[74] It thus explains the fact that "[t]he grandmaster is somehow able to "see" the core of the problem immediately, whereas the expert or lesser player finds it with difficulty, or misses it completely, even though he analyzes as many alternatives and looks as many moves ahead as the grandmaster."[75] And it does not assume that "a human being, like a mechanical pattern recognizer, must classify a pattern in terms of a specific list of traits."[76] That is, it is not a species of the conceptualism against which Lonergan inveighs.

However, if insight in Dreyfus's sense is prepared for by fringe consciousness and made possible by contextual disambiguation, it is sufficient unto itself for the grasp not just of possibility but fact. And this Lonergan would contest. For he takes the act of insight to grasp a possibly relevant intelligibility, and to require verification before it can be judged truly to be so. Or, he takes one sort of insight (direct) to grasp possibly

---

[72] A model of the mind does not get us intelligence any more than a model of the weather gets us wet, Searle quips. See John Searle, Consciousness in Artificial Intelligence | John Searle | Talks at Google - YouTube.

[73] Hubert Dreyfus, *What Computers* Still *Can't Do*, 114.

[74] Hubert Dreyfus, *What Computers* Still *Can't Do*, 119.

[75] Eliot Hearst, "Psychology Across the Chessboard," *Psychology Today* (June, 1967), 32, as quoted in Dreyfus, *What Computers* Still *Can't Do*, 118.

[76] Hubert Dreyfus, *What Computers* Still *Can't Do*, 121.

relevant construal, and another (reflective) to grasp the sufficiency of the conditions for its affirmation.[77] Let us see more closely how this is so.

In response to a What is it? or How often? question, for Lonergan, we grasp unities and relations in the data of sense (or consciousness), and body forth a conception or formulation of that intelligibility in separation from the concrete. We move from so-called apprehensive to formative abstraction, and express what we have understood.[78] But we do not leave things there. For "the desire to understand, once understanding is reached, becomes the desire to understand correctly; in other words, the intention of intelligibility, once an intelligible is reached, becomes the intention of the right intelligible, of the true and, through truth, of reality."[79] And so we inquire further. Of the formulation in hand, we now ask, Is it so?, Is it true? We are not interested in bright idea but confirmed fact; we do not care for possibility but act. We identify a link between our hypothetical and what would confirm it, a tie between our conditioned proposition and its fulfilling conditions. We return to the data, to see if the conditions are fulfilled, and if they are, we affirm, we judge, with greater or lesser assurance.[80]

What is the significance of this? It is that Dreyfus is a direct, while Lonergan is a critical realist, rendering Dreyfus susceptible to over-correction in his criticisms of rationalism. Correctly seeing that intelligence is not a matter of representations and rules, but envisioning no alternative beyond pre-reflective grasp, he needlessly scorns reflection and the distance on oneself it involves. Rightly recognizing the subject not to be set over against a world out there, but envisioning no alternative to (near) self-world identity, he unhelpfully reduces the knower to the known. Or close. It would even appear, at times, that he endorses the physicalist reductionisms of recent, neural AI.

(B) Recent

In "Overcoming the Myth of the Mental," Dreyfus writes that "[t]he meaningful objects . . . among which we live are not a *model* of the world stored in our mind or brain; *they are the world itself*."[81] In "Depth Psychology to Breadth Psychology," he follows an approach that "do[es] not refer to the mind at all." For "the whole human being is related to the world. Indeed, even 'relation' is misleading, since it suggests the

[77] Bernard Lonergan, *Insight*, 304-340.
[78] Bernard Lonergan, *Verbum*, passim.
[79] Bernard Lonergan, "The Subject," in William F. J. Ryan and Bernard J. Tyrrell, eds., *A Second Collection* (Toronto: University of Toronto Press, 1974), 81.
[80] Bernard Lonergan, *Insight,* 296-340.
[81] Hubert Dreyfus, "Overcoming the Myth of the Mental," in Mark Wrathall, ed., *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action* (Oxford: Oxford University Press, 2014), 106, quoting himself from *What Computers* Still *Can't Do*, 265-266.

coming together of two separate entities."[82] And in "Why Heideggerian AI Failed," he says that "in our most basic way of being . . . we are not minds at all but *one with the world* . . . [T]he inner-outer distinction becomes problematic. There's no easily askable question about where the absorbed coping [practical insight] is—in me or in the world."[83]

In other texts, Dreyfus gives examples to support such claims. He cites Sartre's insistence that, in running to catch a streetcar, there is neither runner nor car, but just the situation.[84] He notes Larry Bird's report that he is unaware of what he is doing on the court until after he has done it, as well as the Israeli fighter-pilot's comment that he blacks out in situations of high performance.[85] He even claims that, in his own minimal experience of excellence in tennis, he disappears into the game.[86] It is not just that, in such events, one's awareness of oneself is tacit, and not focal. It is that the distinction between the self and world breaks down.[87] Heidegger calls this "primordial understanding." It "dispenses altogether with the need for mental states like desiring, believing, following a rule, and so on, *and thus with their intentional content*."[88] It is even "zombie-like."[89]

It is this view, then, which would seem to lead Dreyfus to endorse recent, neural AI, in spite of its apparent physicalism. For if distinctions between inner and outer, and even mind and world, break down, then so perhaps do ones between conscious intentionality and nonconscious materiality. And this is just the sort of suggestion we saw in our reviews of Dreyfus on neural net AI and Freeman's neural dynamics. In the former, apparently material transactions were likened to Merleau-Ponty's dialectic of action and milieu, and in the latter they were likened to Heidegger's hermeneutic circle.

However, it is a good question how Dreyfus arrives at his views. What is his method? It cannot be straightforward phenomenology, since it requires claims to be based in the data of consciousness, one's first-personal awareness of oneself and one's acts; and here claims to such realities are abrogated. Nor can it be straightforward science, or any third-personal approach, since it would only reveal non-conscious, meaningless transaction; and what is here being discussed is understanding. Probably Dreyfus would claim his approach is similar to that of early Heidegger and Merleau-

---

[82] Hubert Dreyfus, "Depth Psychology to Breadth Psychology," in Mark Wrathall, ed., *Skillful Coping,* 170.

[83] Hubert Dreyfus, "Why Heideggerian AI Failed," in Mark Wrathall, ed., *Skillful Coping,* 259. I owe this and the former note's quotation to Wrathall, who helpfully lists them in his editor's Introduction to this volume, 4-5.

[84] Hubert Dreyfus - Is Consciousness an Illusion? - YouTube

[85] Hubert Dreyfus, "Responses," in Mark Wrathall and Jeff Malpas, eds., *Heidegger, Coping, and Cognitive Science* (Cambridge: The MIT Press, 2000), 323.

[86]Hubert Dreyfus, "Responses," 329.

[87] Hubert Dreyfus, "Responses," 323.

[88] Hubert Dreyfus, "Husserl, Heidegger, and Modern Existentialism," in Brian Magee, ed., *The Great Philosophers* (Oxford: Oxford University Press, 1987), 258.

[89] Hubert Dreyfus, "Husserl, Heidegger, and Modern Existentialism," 266.

Ponty, who examine being-in-the-world or *etre au monde*, taking subject- and object-poles at once. But such a strategy sits uneasily between phenomenology and metaphysics, and does not lead Heidegger or Merleau-Ponty themselves to any degree of receptivity to naturalism.[90]

No, Dreyfus's method would seem simply to be bad phenomenology. As Searle points out, it cannot be true that in high performance I lose awareness of myself and my goals altogether, otherwise when things cease to go well my attention would not be drawn to the problem.[91] And as Lonergan might observe, the fighter-pilot is likely blurring the difference between tacit and focal awareness in his report. For one can hardly be aware of not being aware of oneself at all.[92] There would not seem to be any reason to suppose that in excellent action we lose awareness of ourselves and the criteria of our success. But if this is so, there is no evidence for Heidegger's "primordial understanding," in which intentional acts and their objects disappear into the world. And there is certainly no evidence for the view that we are naught but electricity acting on circuits.

Dreyfus is right to conclude his opus by saying that "Our risk is not the advent of super-intelligent computers, but of subintelligent human beings."[93] He may not, however, see all that the latter risk entails. For this reason, we should be grateful for the Thomist phenomenology of Lonergan, which offers a verifiable account of the differentiated intelligence that distinguishes us from machines.

---

[90] As Dreyfus himself recognizes. See Hubert Dreyfus, "Merleau-Ponty and Recent Cognitive Science," 245-247.

[91] John Searle, "The Limits of Phenomenology," in Wrathall and Malpas, eds., *Heidegger, Coping, and Cognitive Science*, 77-81. In his "Replies" to his critics assembled in this volume, Dreyfus claims to grant Searle's point (26, 384). But that he does is, I think, belied by the bulk of his replies.

[92] One could perhaps infer from one's present situation, that one just performed well under blackout. But in this one would base a large claim about oneself on material not given in consciousness, which is un-phenomenological.

[93] Hubert Dreyfus, *What Computers* Still *Can't Do*, 280.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Digital Anthropocene: Artificial Intelligence as a Nature-Oriented Technology?

Philipp Höfele
Martin Luther University Halle-Wittenberg, Germany
ORCID: 0000-0002-8682-9965

Abstract
In recent years, technology's orientation toward, and even imitation of, natural models have experienced an enormous upswing. These nature-oriented technologies are often linked to the ethical promise of sustainable as well as ethically and socially responsible technologies that promise an answer to the challenges of the 21st century. Think, for example, of artificial intelligence as an assisting technology for climate protection in the debate on the "Digital Anthropocene." But the fulfillment of this promise is not automatic. This article aims to provide some answers to the theoretical as well as practical-ethical questions that arise with regard to nature-oriented technologies in the present age, especially the role of artificial intelligence in the Anthropocene.

Keywords: nature-oriented technologies, ethical standard, sustainability, dissolution of dichotomies, Anthropocene

Introduction

In recent years, technology's orientation toward, and even imitation of, natural models have experienced an enormous upswing.[1] Under terms such as biomimetics, bioinspiration, social robotics, or artificial intelligence, technologies are being developed that imitate the functions of plant, animal, or human nature. The resulting products do not limit themselves to a modest orientation towards a supposedly unattainable nature, but also clearly go beyond natural models by attempting to bring together the "best of both worlds," the natural and the technical.[2] Historically, this phenomenon is by no means new. The origins of this development go back to antiquity and the construction of automata, even if in earlier times such orientations and imitations of nature often lagged far behind their natural models and usually existed only in the realm of fiction.[3]

These nature-oriented technologies are often linked to the ethical promise of sustainable as well as ethically and socially responsible technologies that promise an answer to the challenges of the 21st century and especially to the so-called Anthropocene as an epoch in the history of the earth that is dominated by humans and their technologies, often in harmful ways.[4] But the fulfillment of this promise is not automatic in the case of such technologies. The simple, unquestioned orientation to plant, animal, and human nature can even be dangerous, for example, when natural functions and mechanisms are imitated with artificial materials whose sustainability is by no means assured.[5] On the one hand, this calls for a differentiated ethical evaluation of these technologies, which does not make nature the sole yardstick, even if an orientation towards nature has advantages. On the other hand, the underlying concepts of nature and orientation to nature must be questioned on a theoretical level. It is not

---

[1] Cf. Werner Nachtigall and Charlotte Schönbeck (ed.), *Technik und Natur* (Berlin and Heidelberg: Springer, 1994); Alfred Nordmann, *Converging Technologies—Shaping the Future of European Societies* (Luxemburg, 2004). https://op.europa.eu/en/publication-detail/-/publication/7d942de2-5d57-425d-93df-fd40c682d5b5; Rinie van Est et al., *Making Perfect Life. European Governance Challenges in 21st Century Bio-engineering* (Brussels, 2012).
https://www.europarl.europa.eu/RegData/etudes/etudes/join/2012/471574/IPOL-JOIN_ET%282012%29471574_EN.pdf; Olga Speck et al., "Biomimetic bio-inspired biomorph sustainable? An attempt to classify and clarify biology-derived technical developments," *Bioinspiration & Biomimetics* 12, no. 1 (2017): 1-15. https://doi: 10.1088/1748-3190/12/1/011004; Philipp Höfele, Oliver Müller and Lore Hühn (ed.), *The Anthropocene Review,* Special Issue 9, no. 2 (2022): *The Role of Nature in the Anthropocene.*

[2] Cf. the slogan of the Cluster of Excellence *liv*MatS: https://www.livmats.uni-freiburg.de/en

[3] Cf. e.g. Pascal Weitmann, *Technik als Kunst. Automaten in der griechisch-römischen Antike und deren Rezeption in der frühen Neuzeit als Ideal der Kunst oder Modell für Philosophie und Wirtschaft* (Tübingen: Wasmuth & Zohlen, 2013); Bianca Westermann, "The Biomorphic Automata of the 18th Century," *figurationen* 17, no. 2 (2016): 123-37.

[4] Cf. Paul J. Crutzen and Eugene F. Stoermer, "The 'Anthropocene,'" in *Global Change Newsletter* 41 (2000): 17-8; Paul J. Crutzen, "Geology of mankind," in *Nature* 415 (2002): 23.

[5] Cf. Martin Möller et al., "Re-actions of sciences to the Anthropocene: highlighting inter- and transdisciplinary practices in biomimetics and sustainability research," *Elementa: Science of the Anthropocene* 9, no. 1 (2020): 9-11. https://doi: 10.1525/elementa.2021.035.

only necessary to ask to what extent these technologies are oriented towards nature and what implications this has for the human-nature-technology relationship. At the same time, it must be considered that the underlying concepts of nature have normative connotations that flow into nature-oriented technologies.

This article aims to provide some answers to these theoretical as well as practical-ethical questions that arise with regard to nature-oriented technologies in the present age of the Anthropocene, especially the role of artificial intelligence. I will examine three theses: (a) From a theoretical point of view, the idea of nature-oriented technology can be found as early in philosophical reflections on technology as Aristotle, but it is only in the present age that it has acquired a prominent importance—on the one hand with regard to the technical possibilities of imitation, especially in the fields of biomimetics and artificial intelligence, and on the other hand with regard to the need for sustainable solutions oriented towards nature. (b) This implies a second practical-normative hypothesis: in the present age, a technical orientation towards nature is often accompanied by practical-normative assumptions, namely that these technologies offer "better" solutions compared to "traditional" technologies, especially in the context of the Anthropocene and its problems. Think, for example, of artificial intelligence as an assisting technology for climate protection in the debate on the "Digital Anthropocene." [6] (c) Finally, nature-oriented technologies, and especially AI systems, tell us something about the relationship between nature and technology in general. They help to learn something in theoretical and practical-ethical terms regarding the nature-technology relationship as a whole.

The article is divided into four sections: (1) First, I will discuss some paradigmatic interpretations of technology as essentially oriented toward and in continuity with nature in the history of philosophy and in the present age of the Anthropocene. (2) Against this historical background, the second task is to determine what nature orientation means in the case of technology and, in particular, artificial intelligence. Defining the concept of nature orientation will already reveal some of the problems and challenges associated with these technologies. (3) Third, these problems or challenges need to be discussed in terms of environmental ethics, which is often neglected in the case of AI systems: To what extent can one speak of a nature orientation as an ethical standard with regard to artificial intelligence? Nature can by no means be used unquestioningly as a yardstick in the sense that, for example, theorists of a philosophy of biomimicry often do, by simply regarding the imitation of nature and its functionality as sustainable and thus "good."[7] Such an approach, if taken without reflection, runs the risk of a naturalistic fallacy. In the case of artificial

[6] Cf. e.g. Jessica McLean, *Changing Digital Geographies Technologies, Environments and People* (Cham: Springer, 2020). https://doi.org/10.1007/978-3-030-28307-0_1; Felix Creutzig et al., "Digitalization and the Anthropocene," *Annual Review of Environment and Resources* 47 (2022): 479-509. https://doi.org/10.1146/annurev-environ-120920-100056.

[7] Cf. Janine M. Benyus, *Biomimicry* 2nd ed. (New York: Harper Collins, 2002); Arnim von Gleich et al., *Potentials and trends in biomimetics* (Heidelberg et al.: Springer, 2010).

intelligence, there is also a tendency wherein it is increasingly viewed not as an imitation of human intelligence, as was the case in Turing's time, but rather as an attempt to develop other forms of intelligence.[8] Orientation to nature therefore plays a role here primarily in the sense of a limiting framework or standard that is brought into the field on the basis of sustainability. (4) However, nature-oriented technologies also pose a further challenge on a theoretical-ontological level, as I will try to show in the fourth section: nature-oriented technologies have fundamental implications for the nature-human-technology relationship, which in turn has ethical implications. As an entity that cannot simply be characterized as natural or artificial in the classical sense—for example, in Aristotle—nature-oriented technologies, and especially AI systems, irritate the classical nature-technology dichotomy and the hierarchies that go with it. Last but not least, they demand that the alternative between anthropocentric and physiocentric ethical approaches be broken up and further pluralized.

Technical Orientation to Nature in the History of Philosophy and the Origins of Artificial Intelligence Research

The fact that technology does not necessarily have to be understood as the other in relation to nature, but can be understood as standing in continuity with it, is not a new idea.

(1) Already Aristotle remarks in his lectures on *Physics*: "[G]enerally art (*technê*) in some cases completes (*epiteleî*) what nature cannot bring to a finish, and in others imitates (*mimeîtai*) nature."[9] In this way, Aristotle defines the relationship between nature and technology, or more precisely between *physis* and *techné,* insofar as the Greek term has a much broader range of meanings than the English word and ultimately encompasses the entire field of the artificial as well as the non-natural. Unlike Plato in the 10th book of the *Politeia,* Aristotle does not connect imitation *per se* with the subordination of the artificial to the natural.

In *Physics* II, 8, Aristotle is certainly interested in integrating the realm of *techné* into that of nature almost to the point of identity. To prove this structural equality of *physis* and *techné,* Aristotle gives the example of a house: if one imagined that the house was a natural object that had grown of its own accord, the parameters for considering its constructing would still be the same as those for a work of art. The purposefulness present in both cases guarantees this structural equality between the artificial and the natural; both share a "why (*hoû héneka*)" as a cause that structures the process of creation or production; in other words, its goal.[10] Ultimately, Aristotle advocates here

[8] Cf. Nick Bostrom, *Superintelligence. Paths, Dangers, Strategies* (Oxford: OUP, 2014), 22-51.
[9] Aristotle, *The Complete Works of Aristotle: The Revised Oxford Translation,* ed. Jonathan Barnes (4th ed., Princeton: Princeton University Press, 1991), vol. 1: *Physics*, 32, bk. II, par. 8, 199a15f. Cf. Philippe Lacoue-Labarthe, *L'imitation des Modernes (Typographies 2)* (Paris: Galiliée, 1986), esp. 23f.
[10] Aristotle, *Physics,* 32, bk. II, par. 8, 199a32.

for a union of *physis* and *techné* that makes perfection possible. Even in passive imitation, this position does not ignore the difference and the added value of each for the other.

(2) In a very similar way, even 2000 years later, namely in 1877, Ernst Kapp, in the first work ever dedicated to the *Outlines of a Philosophy of Technology* (*Grundlinien einer Philosophie der Technik*), argues for an understanding of technology as an imitation of nature or, more precisely, of human nature. In general, technology is "organ projection (*Organprojektion*)." Here Kapp understands by projection "more or less the projecting or highlighting, emphasizing, transferring out, and relocating of an internal into the external (*mehr oder weniger das Vor- oder Hervorwerfen, Hervorstellung, Hinausversetzen und Verlegen eines Innerlichen in das Aeussere*)."[11] "Organ projection" thus describes a projection or—better—imitation of human organs by means of external objects for the purpose of reinforcing the former. Kapp uses the genealogical explanation of the hammer as a basic example. For this is "like all primitive hand tools an organ projection or the mechanical reproduction of an organic form":

> So if the forearm with the hand clenched into a fist or with its reinforcement by a graspable stone is the natural hammer, the stone with the wooden handle is its simplest replica (*einfachste Nachbildung*). For the handle or grip (*der Stiel oder die Handhabe*) is the extension of the arm, the stone the substitute for the fist (*der Ersatz der Faust*).[12]

But according to Kapp, it is not only the "primitive hand tool" that is to be interpreted as such an organ projection. Rather, this description could also be applied to more recent and much more complex technologies. The key technology of the 19th century, the steam engine, is also interpreted by Kapp as an organ projection: "Many machine parts, originally isolated tools, are united in the steam engine externally to a mechanical collective action (*Gesammtwirkung*), like the members of the animal series internally to a highest organic life unit reached in man (*innerlich zu einer höchsten im Menschen erreichten organischen Lebenseinheit*)." [13] Kapp even goes beyond this in a certainly not unproblematic way, in that he even understands the "unification of the railroads and steamship lines" and thus "the network of traffic arteries (*Netz von Verkehrsadern*)" as "the image of the network of blood vessels in the organism (*Abbild des Blutgefässnetzes im Organismus*)."[14]

(3) The French philosopher Georges Canguilhem, like Aristotle and Kapp, assumes that nature in the form of the organism is imitated by the technical machine. In the chapter *Machine et organisme* of his book *La connaissance de la vie* (1952),

---

[11] Ernst Kapp, *Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Cultur aus neuen Gesichtspunkten* (Braunschweig: George Westermann, 1877), 30.
[12] Kapp, *Grundlinien einer Philosophie der Technik*, 42.
[13] Kapp, *Grundlinien einer Philosophie der Technik*, 133.
[14] Kapp, *Grundlinien einer Philosophie der Technik*, 135.

Canguilhem tried to reveal the history of the multi-layered interrelationship between nature and technology, which was partly characterized by suppressions and abridgements. At the same time, Canguilhem takes a second relationship into account, insofar as "one cannot understand the phenomenon of machine construction if one falls back on concepts of nature from biology (*notions de nature authentiquement biologique*) without at the same time asking where technology comes from in relation to science."[15]

Canguilhem concludes by characterizing the relationship between technology and science in the following way: "The one does not graft itself onto the other, but each sometimes borrows solutions, sometimes questions from the other. The rationalization of techniques makes the irrational origin of machines fall into oblivion."[16] Canguilhem's statement is by no means a plea for irrationalism, but rather for a rationalism freed from the dominance of scientific necessity, following the suggestions of Ernst Kapp.[17] For Canguilhem, the freedom of technical developments and innovations seems to be guaranteed precisely when technology is not degraded to a simple application of scientific reasoning. Canguilhem believes that this freedom of technology is guaranteed by its mimetic integration into the space of possibility contained in the natural world, technology being "a universal biological phenomenon (*un phénomène biologique universel*)," [18] which, like the natural, can also dispose of the space of possibility inherent in nature, its potentiality.

(4) However, the modern understanding differs from these selective historical perspectives on the essence of technology as a general imitation of nature insofar as technology here is largely no longer understood as an imitation of nature. Rather, only certain areas of technology such as biomimetics, artificial intelligence or the results of synthetic biology are understood as imitating nature or orienting themselves towards nature. This is done in a way that distinguishes these areas of technology from technologies that are not understood as being oriented toward nature. This opens up the possibility of understanding the technical imitation of nature, for example, in the form of the biomimetic promise, as better in a normative sense, that is, more innovative or sustainable than other technologies.

The development of artificial intelligence in particular could claim to be innovative, and this precisely in its attempt to imitate natural, human intelligence. In his article *Computing Machinery and Intelligence* (1950), Alan Turing elevated the concept of imitation to a benchmark for the development of artificial intelligence. To get around the difficult question of what "machine" and "thinking" mean when talking about a "thinking machine" in the sense of artificial intelligence, Turing advocates an "imitation game": "It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of

---

[15] Georges Canguilhem, *La connaissance de la vie* (Paris: Librairie Hachette, 1952), 125.
[16] Canguilhem, *La connaissance de la vie,* 157.
[17] Cf. Kapp, *Grundlinien einer Philosophie der Technik*, 136-38 and 155-64.
[18] Canguilhem, *La connaissance de la vie,* 158.

the other two is the man and which is the woman."[19] This simple game now becomes a test for an artificial intelligence if A is replaced by a machine and the interrogator is given the task of deciding who is natural and who is artificial intelligence. This would replace the original question "Can machines think?" with the following, more easily answered question: "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?"[20] For if the interrogator would decide wrongly, this would be a strong indication that we are dealing with a "thinking machine" that truly imitates natural intelligence.

(5) Certainly, artificial intelligence has been significantly developed since Turing and can no longer be generally understood as an imitation of human intelligence. Nevertheless, it has played a not insignificant role in another, broad form of nature imitation in the Anthropocene, which is linked to the concept of the technosphere. The term "technosphere," which has been taken up as a "global paradigm" by Peter Haff and Jan Zalasiewicz, describes a form of imitation taken to extremes, understood as a habitat equal to the biosphere, albeit with serious problems regarding its sustainability:

> The technosphere, the interlinked set of communication, transportation, bureaucratic and other systems that act to metabolize fossil fuels and other energy resources, is considered to be an emerging global paradigm, with similarities to the lithosphere, atmosphere, hydrosphere and biosphere. The technosphere is of global extent, exhibits large-scale appropriation of mass and energy resources, shows a tendency to co-opt for its own use information produced by the environment, and is autonomous.[21]

The established analogy between the technosphere on the one hand and the lithosphere, atmosphere and hydrosphere on the other hand is, despite all similarities, at the same time accompanied by an essential difference between these spheres. The technosphere is not completely self-sufficient and self-contained, but lives at the expense of the natural spheres, which is essential to the problems of the Anthropocene. Thus, Zalasiewicz et al. also note with regard to the negative sustainability balance of the technosphere, with which it clearly differs from the self-sufficient biosphere, that it "includes . . . a growing residue layer, currently only in small part recycled back into the active component."[22] Non-recycled waste is also a central problem of the technosphere, which is cited in the context of the "Digital Anthropocene" discourse as the negative side of digitalization and artificial intelligence, which have contributed significantly to the co-construction of the

---

[19] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* LIX, no. 236 (1950): 433.
[20] Turing, "Computing Machinery and Intelligence," 434.
[21] Peter K. Haff, "Technology as a geological phenomenon: implications for human well-being," *Geological Society, London, Special Publications* 395 (2014): 301. https://doi.org/10.1144/SP395.4.
[22] Jan Zalasiewicz et al., "Scale and diversity of the physical technosphere: A geological perspective," *The Anthropocene Review* 4, no. 1 (2017): 1. https://doi.org/10.1177/2053019616677743.

technosphere.[23] Here, different, both positive and negative-imperfect forms of nature orientation play into each other, which first have to be distinguished and put into a relation to each other, before they can be ethically evaluated in a further step.

The Nature Orientation of Technology: A Definition of the Term Using the Example of Artificial Intelligence

(1) If one speaks of nature orientation, first of all the exact reference object of this orientation must be named, insofar as nature and the natural can denote quite different things. Thus in the case of the technical orientation to nature, the reference point is usually living nature, as terms such as bioinspiration, biomimetics, biomimicry or bionics already indicate from their Greek root "*bios*" (life). The research fields just mentioned all refer (a) to non-human, plant or animal nature, whose forms and functions they seek to imitate by technical means. Robotics, on the other hand, focuses primarily on the (b) physical nature of humans, while artificial intelligence (c) attempts to emulate the mental nature of humans, their ability to learn, judge or solve problems, unless the development of entirely different, "posthuman" forms of intelligence is aimed at. At the same time, however, it can also be about the (d) imitation and implementation of social concerns and normative demands of human societies, which is the focus of social robotics in particular, but also of AI systems.

(2) From the point of view of an ethical assessment, however, the formal-relational side of nature orientation is far more interesting and important. The term "nature orientation," chosen here to describe technical developments in the Anthropocene and especially in the context of research on artificial intelligence, is a relatively broad term that covers a vast number of nature-technology relationships.

Formally, this includes first of all any form of technical orientation to nature in which this orientation plays a role during production, but after which the technical entities created no longer necessarily have to stand in a relation to the natural models. This can be the case (a) in the exact technical imitation of a natural model—for example, in the production of an artificial cell in synthetic biology, even if it is already advisable here from an ethical point of view to keep an eye on the relationship to the natural environment after production and to reflect on it. However, this first category of nature orientations in a formal sense also includes (b) more abstract forms of technical imitation of nature, which, for instance—as in biomimetics—focus on the morphology or functional principles of plant or animal entities. In a broader sense, however, nature orientation can also be understood as (c) any inspiration from nature to solve human problems, also related to larger structures such as ecosystems or the entire biosphere. In all these cases, the traditional dichotomy of the natural and the

[23] Cf. Jennifer Gebrys, *Digital Rubbish: A Natural History of Electronics* (Ann Arbor: University of Michigan Press, 2011), 1-17. https://doi.org/10.2307/j.ctv65swcp.4.

technically and artificially generated cultural is not touched, even if the realm of the cultural directs its gaze to the former realm for its shaping.

This perspective, which is often only theoretical and ideal, is expanded in the following forms of nature orientation. In addition to the orientation to nature during the production of technical entities, the aim here is to ensure a permanent orientation to nature. Certainly, all objects that have been created within the framework of a nature orientation never behave without reference to their natural environment, even after their production, but are always in a relationship to it in one way or another. However, this nature orientation is not always consciously reflected and certainly not considered in terms of sustainability. A central way of reflecting nature orientation, especially in the context of the Anthropocene, has turned out to be (d) the adaptivity of technical objects to the natural environment. One example, albeit not an uncontroversial one, is the engineering idea of "stratospheric aerosol injection," in which, following the example of volcanic eruptions, sulfur particles are released into the atmosphere in order to reduce global warming.[24] However, this perspective also includes any form of (e) responsiveness between technical entities and their natural and social environment, whether in the form of acceptance research with regard to the human environment or the AI-supported collection of data from the natural environment in order to be able to adapt human-technical behavior to it in the sense of sustainability.

For the technical object and its development, however innovative it may be, are always bound up in contexts. This was emphasised above all by Canguilhem's student Gilbert Simondon in his 1958 work *Du mode d'existence des objets techniques,* systematically following his teacher's reflections on the concept of the natural "milieu." Just as, according to Canguilhem, living things in general, from the cell to the organs to the entire organism, are integrated into a "milieu," form a unity of mutual constitution with it and are therefore to be characterized by the basis it presents,[25] Simondon also seeks to understand the technical object in relation to its environment. Against this backdrop, Simondon understands the human-machine relationship as an "inter-individual coupling (*couplage*) between human beings and machines": "Human beings can be coupled to the machine as a being that participates in its regulation, not as a being that merely directs and uses it by incorporating it into the ensembles, or as a being that serves it by supplying it with material or elements."[26] Simondon assumes here a symbiosis between human beings and machines, so to speak, which leads to the fact that neither of the two interaction partners can exist on their own.

Martina Heßler rightly emphasises the proximity between Simondon's approach to the philosophy of technology and Bruno Latour's network theory, which

---

[24] Cf. Will Steffen et al., "The Anthropocene: conceptual and historical perspectives," *Philosophical Transactions (Series A)* 369 (2011): 858-61. https://doi.org/10.1098/rsta.2010.0327.
[25] Cf. Canguilhem, *La connaissance de la vie,* 160-93.
[26] Gilbert Simondon, *Du mode d'existence des objets techniques* (Paris: Aubier, 1989), 119-20.

sees the biosphere as being made up of irretrievably interwoven actors and agents.[27] The buzzword, "Digital Anthropocene," is increasingly being used to draw attention to this reciprocal entanglement between the natural and social environment on the one hand and AI systems on the other. However, this raises the question of which forms of nature orientation can be described as good in an ethical-normative sense and according to which ethical criteria this is to be evaluated.

Nature Orientation as an Ethical Standard in the Case of Artificial Intelligence?

Demanding nature orientation as an ethical standard in the case of artificial intelligence is not easy to justify from a metaethical point of view and is quite problematic. The argument in the ethics of nature calling for an orientation towards nature goes back to Aristotle and the ancient Stoa.[28] But this so-called following nature argument was vehemently criticized, especially in the 19th century with the advent of Darwin's theory of evolution, for example by John Stuart Mill in his famous 1874 essay *Nature*. Following Mill, Angelika Krebs has succinctly summarized and convincingly renewed the critique of the following nature argument in her work *Ethics of Nature*:

> The imperative to follow nature is either superfluous or morally objectionable. [1] It is *superfluous* if it means that we should follow the natural laws where we are subject to them, because where we are subject to natural laws we cannot but "follow" them. [2] It is *morally objectionable* if it asks us to imitate what we see in nature, for a lot of "cruelty" and destruction goes on in nature.[29]

On the one hand, it makes no sense to speak of following nature in the case of processes or actions that are subject to natural laws anyway—be they of physical or biological nature—since here there is no alternative to following, and therefore one cannot behave freely in the course of such an obligation. Similarly, on the other hand, if we had this freedom of action, we would also have to reject this second case as morally questionable, since in many cases, given the frequent cruelty of nature, it is questionable and even reprehensible to follow nature. This ambivalence would not disappear even if one were to cite certain criteria, such as the complexity of natural processes, the stability of certain states of equilibrium in ecosystems, or the age of certain natural phenomena, because all of these criteria exhibit ambivalence.[30]

---

[27] Cf. Martina Heßler, "Gilbert Simondon und die Existenzweise technischer Objekte. Eine technikhistorische Lesart," *Technikgeschichte* 83, no. 1 (2016): 3-32, esp. 27.
[28] Cf. Aristotle, *The Complete Works of Aristotle: The Revised Oxford Translation,* ed. Jonathan Barnes (4th ed., Princeton: Princeton University Press, 1991), vol. 2: *Nicomachean Ethics*, 32; bk. X, par. 7, 1178a. Cf. also Anna Schriefl, *Stoische Philosophie. Eine Einführung* (Stuttgart: Reclam, 2019), 138-41.
[29] Angelika Krebs, *Ethics of Nature. A Map* (Berlin and New York: De Gruyter, 1999), 128.
[30] Cf. Krebs, *Ethics of Nature*, 127.

I fully agree with this argument. However, it should be noted that it only considers the extreme cases, namely a complete lack of freedom to disobey the laws of nature, which renders the concept of "following" meaningless, as well as a complete and uncompromising following nature based on freedom. However, in the case of the nature-oriented technologies, it is by no means a matter of following nature in its entirety and under all circumstances. Only *individual* moments to be found in nature can function as maxims (*Maxime*) for action, to use Kant's vocabulary. But these, in turn, must first be tested for their suitability as generally acceptable principles of action or orientation, whereby they can only be considered generally valid in relation to a particular realm of nature—such as human nature or non-human animate nature.

An examination of those maxims derived from nature can be carried out from different angles. It can be *anthropocentric,* for example when natural models for technologies are examined for their compatibility with societal goals such as the "Sustainable Development Goals,"[31] when their societal acceptance is questioned, or when their compatibility with intergenerational justice is examined. It can also be carried out from a *biocentric or physiocentric* perspective, when the focus is on the influence of certain natural or artificial mechanisms on the ecosystems into which they are to be integrated.

Nature orientation as an ethical maxim in the case of artificial intelligence can therefore enrich certain anthropocentric or physiocentric ethical approaches in terms of material content. However, they cannot be considered as ultimate ethical principles themselves. (1) This becomes particularly clear in the case of AI systems that imitate the morally ambivalent natural intelligence of humans. In the course of such imitations, they can, for example, also reproduce racist biases peculiar to humans.[32]

As has been noted on various occasions in the discussion about the "Digital Anthropocene,[33] AI systems can also contribute directly or indirectly to environmental protection through their orientation towards nature. (2) It is often forgotten that AI systems not only generate immaterial capabilities comparable to the human mind, but also have a material basis that is taken from the natural environment in the form of resources (such as rare earths) and will at some point flow back into it as waste. AI systems and their material basis should therefore also follow the principles of the circular economy, which is derived from the natural cycle.[34] (3) Furthermore, AI systems can also be subject to a positively evaluated nature orientation in the sense that they can be used to collect data sets that can be processed exclusively by them,

---

[31] Cf. on the 17 "Sustainable Development Goals" (SDGs), ratified in 2015 by all members of the United Nations in the *2030 Agenda for Sustainable Development*: https://sdgs.un.org/goals

[32] Cf. Christoph Bartneck et al., *An Introduction to Ethics in Robotics and AI.* Cham: Springer, 2020), 34-5, where this problem is discussed with regard to AI-driven lending.

[33] Cf. Creutzig et al., "Digitalization and the Anthropocene," 479-509.

[34] Cf. Bernadette Bensaude Vincent, "Of Times and Things. Technology and Durability," in *French Philosophy of Technology. Classical Readings and Contemporary Approaches,* ed. Sacha Loeve, Xavier Guchet, and Bernadette Bensaude Vincent (Cham: Springer, 2018), 291-4. https://doi:10.1007/978-3-319-89518-5_17.

which contribute to a holistic understanding of the earth system and, in a second step, can be used to develop sustainability strategies.[35] (4) But AI systems should not only be used for data collection regarding the non-human environment in terms of sustainability. AI systems, on the one hand, also have considerable influence on informed decisions of humans, for example, through social networks, and, on the other hand, can also have an impact on social inequalities, for instance through AI-based decision-making processes. These influences of AI on societies can, in turn, have repercussions on the treatment of the natural environment, so that an orientation towards the standards and norms of human societies is required, which in turn ensures an adequate indirect influence of AI systems on the natural environment.[36] (5) In addition, artificial intelligence should continue to be oriented to natural, human intelligence from an ethical point of view, despite all the possibilities for development, which will possibly go far beyond the latter. Artificial intelligence is primarily focused on data processing as only one aspect of human thinking. In doing so, specific distinctions of human thought, such as "practical wisdom" or "virtuousness," which have ethical relevance, are often neglected.[37] The resulting ethical demand can either be that artificial intelligence should be self-limiting, with humans guiding and regulating it, or that these aspects of human thinking should be imitated.

(6) All of these points are based on sustainability theory and thus on an anthropocentric argumentation. Artificial intelligence—so one could describe the ethical approach—should be committed to contributing to sustainability and thus to preserving an earth that is also habitable for future human generations. What is usually neglected is the question of what artificial intelligence could contribute to the well-being of non-human entities on earth *for their own sake.* Insofar as artificial intelligence is sometimes also discussed in posthumanist discourses as something that could "overcome" humans[38] and thus ontologically represents a novel entity, it could also open up perspectives for physiocentric ethics.

Summary and Outlook: Dissolution of Dichotomies as an Ethical Challenge and Opportunity

If we look back at the history of the development of artificial intelligence, we can see that its claim is to imitate natural, human intelligence. This is especially true for the approach of "strong artificial intelligence," which initially claims to produce systems

[35] Cf. Creutzig et al., "Digitalization and the Anthropocene," 498.
[36] Cf. Creutzig et al., "Digitalization and the Anthropocene,"485-90.
[37] Cf. Mark Coeckelberg, *AI Ethics* (Cambridge and London: MIT Press, 2020), 200-2.
[38] Cf. Philipp Höfele, "Zwischen Moralphilosophie und Anthropologie. Zum Spannungsverhältnis von Natur und Bestimmung des Menschen bei Kant und in der Debatte um 'Human Enhancement,'" in *Anthropologie in der Klassischen Deutschen Philosophie*, ed. Christoph Asmuth, Simon Helling (Würzburg: Königshausen & Neumann, n.d.), 215-234.

that are on a par with humans in terms of their problem-solving abilities. However, this claim does not necessarily have to be pursued in general. For example, future AI systems may also have cognitive abilities that are completely different from those of humans. The same applies to "weak artificial intelligence," which is geared towards individual, concrete problems, in the solution of which it can proceed much more effectively as well as differently than human thinking.

However, as mentioned above, this is only one form of nature orientation that plays a role in the production of AI systems and is limited to them, for example in order to be able to ensure the autonomy of the technical system after the production process has been completed. It is precisely here, however, that a further, continuous form of nature orientation is increasingly being called for, one that concerns AI systems in particular. According to the thesis of the "Digital Anthropocene," only a permanent orientation towards nature in the five aspects 2-6 mentioned above will ensure the sustainability and environmental compatibility of AI systems.

AI systems do not dissolve the boundary between the natural and the artificial, as is the case with many other technologies in the Anthropocene that are oriented toward nature. Unlike biomimetic products or the "biofacts" described by Nicole Karafyllis, which include genetically modified corn,[39] AI systems still maintain the boundaries to the natural. The nature orientation of AI systems will not, at least in the main, lead to the dissolution of the boundary to human nature. We will not get AI systems where we will have to ask ourselves—in the sense of the Turing test—whether we are dealing here with natural or artificial intelligence. As autonomous technologies, AI systems would possess what Aristotle describes as an essential characteristic of *physis*, namely, they would have the origin and the principle of motion in themselves.[40] Nevertheless, a "genetic naturalness" in the sense of Dieter Birnbacher[41] can only ascribe to them conditionally, insofar as these systems, even in the case of a possible future self-reproduction, will still have their historical origin in a human invention. Nor would "qualitative naturalness" apply, insofar as these technologies are unlikely to adopt the appearance or behavior of their natural models. In this respect, a concept that assumes a gradual difference between the natural and the artificial is also not applicable with regard to these technologies.[42] The idea is not to construct something between nature and technology, but something qualitatively better or at least different. A slavish or at least gradual imitation of natural intelligence is—beyond a possible research interest and curiosity—often rather uninteresting from an economic point of view. Thus, technical imitation of nature usually does not aim at a mere reproduction

---

[39] Nicole Karafyllis, *Biofakte. Versuch über den Menschen zwischen Artefakt und Lebewesen* (Paderborn: Mentis, 2003).
[40] Cf. Aristotle, *Physics,* 19; bk. II, par. 1, 192b-193b). Cf. also Henry State, *Techne Theory: A New Language for Art* (London et al.: Bloomsbury, 2019), 65-84.
[41] Cf. Dieter Birnbacher, *Naturalness. Is the "Natural" Preferable to the "Artificial"?*, trans. David Carus (Lanham et al.: University Press of America, 2014), 7-15.
[42] Cf. Krebs, *Ethics of Nature*, 5-7.

of the natural, but at creating something new and better by pursuing the goal of "combining the best of two worlds—nature and technology," as is the motto of the *liv*MatS cluster of excellence, for example, which is dedicated to the development of biomimetic technologies.[43]

Given the forms of nature orientation in AI systems described above, it is likely that a third class of objects will emerge that will be situated beyond the natural and the artificial. As noted above, these new entities are not simply unrelated to the traditional dichotomy of the natural and the human-artificial. On the one hand, AI systems should always serve societal needs and norms that are manifested in them without being absorbed by them.[44] On the other hand, AI systems should also show an orientation towards the natural environment in the above-mentioned respects, whereby they can represent a corrective to a narrow anthropocentric perspective. In their autonomy, AI systems will not simply be absorbed into the role of anthropocentrically oriented instruments. As systems that will exceed natural human intelligence, at least in some respects, they are likely to call into question the traditional, dominant hierarchy in which humans are at the top. As a third class of objects, AI systems are also likely to dissolve the classical dual coordinate system of natural and human-artificial. Neither will simply lead to an ethical physiocentrism, but it should at least relativize the ethical anthropocentrism.

AI systems oriented towards nature can thus be seen as a real-world counterpart to Donna Haraway's concept of cyborgs, which she understands "as an imaginative resource suggesting some very fruitful couplings;" just as Haraway sees us all as "fabricated hybrids of machine and organism," as imagined hybrid cyborgs,[45] in order to undermine traditional dichotomies. For it is precisely in the Anthropocene that the wild, the primordial, the immediate usually turns out to be pure illusion, insofar as everything in the biosphere has already been reshaped by humans. Conversely, there is also nothing purely artificial, since it is always already part of nature. The classical antipoles of natural and artificial do not apply here, not even in the sense of a gradual difference, but are rather to be seen as interpretative settings that at the same time have normative implications.

The introduction of a third entity beyond the dichotomous relationship between the natural and the human-artificial should thus provide for the elimination of this dichotomy and its normative implications and thus possibly lead to an equal or at least more differentiated ethical appreciation of the natural and other hybrid entities in relation to humans.

---

[43] Cf. https://www.livmats.uni-freiburg.de/en

[44] Cf. Mark Coeckelbergh, "Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach," *International Journal of Social Robotics* (2021): 10. https://doi:10.1007/s12369-021-00770-0.

[45] Cf. Donna Haraway, *Simians, cyborgs, and women: the reinvention of nature* (New York: Routledge, 1991), 150.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Julian Jaynes and the Next Metaphor of Mind: Rethinking Consciousness in the Age of Artificial Intelligence

George Saad
Memorial University of Newfoundland, Canada
ORCID: 0009-0005-2812-2274

Abstract

In *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Julian Jaynes presents a philosophy of mind with radical implications for contemporary discussions about artificial intelligence (AI). The ability of AI to replicate the cognitive functions of human consciousness has led to widespread speculation that AI is itself conscious (or will eventually become so). Against this functionalist theory of mind, Jaynes argues that consciousness only arises through the mythopoetic inspiration of metaphorical language. Consciousness develops and enacts new forms of self-understanding, continually evolving new "metaphors of mind," metaphors which must now account for the emergence of AI.

Keywords: artificial intelligence (AI), Julian Jaynes, evolution of consciousness, metaphor, functionalism

Introduction

Julian Jaynes is a figure who stands at the intersection of several disciplines. His 1976 magnum opus *The Origin of Consciousness in the Breakdown of the Bicameral Mind* evolutionary theory, neuroscience, and literary criticism to advance a speculative hypothesis about the evolution of human consciousness. Anticipating Iain McGilchrist's recent work,[1] Jaynes shows how modern self-consciousness evolved through the increasing coordination of the brain's right and left hemispheres. The self-determination of individual self-consciousness evolved relatively recently from a "bicameral" state. In this primordial state, the left hemisphere was subordinated to the right. The mythopoetic content generated in this hemisphere (particularly the auditory incantations of oral poetic cultures, the gifts of the Muses) governs the left hemisphere, which follows these "songs from beyond" as the enchanted products of an external divinity.

In tracing a biological evolutionary process through the inferential evidence of ancient texts, Jaynes adopts an idiosyncratic methodology which can easily obscure his philosophical commitments. When presented in the earlier, more theoretical sections of the text,[2] these commitments seem to run off in wildly different directions. At first appearance, he seems to be a reductionist, arguing that what is apparently the work of consciousness is actually accomplished through the unconscious mechanisms of neural networks. *Origin* opens with a thorough deflation of the functions of consciousness, arguing that even learning and reasoning are not essentially conscious activities.

But where this line of argument would seem to lead to an abandonment of consciousness entirely, Jaynes makes a shocking pivot from functional neurology to poetic linguistics, claiming that consciousness could have only emerged after the development of language. This implies that consciousness is much more recently evolved than generally suspected, even among *Homo sapiens*.[3] The failed attempt to derive consciousness from cognitive functions was always an attempt to infer an inner experience from external output. The functionalists do not recognize the constitutive role of language in the emergence of this inner conscious experience. For Jaynes, "language is an organ of perception [and] not simply a means of communication."[4] The "I" that is conscious is at once the product and producer of language. Language does not just describe the world—it is but the very state of having a "world" in the Heideggerian sense. It is subjectivity as such.

---

[1] Iain McGilchrist, *The Master and His Emissary: The Divided Brain and the Making of the Western World* (New Haven: Yale University Press, 2019).

[2] See Julian Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (New York: Houghton Mifflin, 2000), 1-66.

[3] Jaynes, *The Origin of Consciousness*, 66.

[4] Jaynes, *The Origin of Consciousness*, 50.

Jaynes only briefly presents this unique philosophy of mind remains as the theoretical basis for a more concrete analysis of evidence for the evolution of consciousness as it appears in literature, neurophysiology, and psychiatry. In the first two sections of this paper, I will extrapolate the arguments implicit in, or at least complementary to, the outline of a philosophy of mind presented the *Origins*. While Jaynes is not primarily concerned with a philosophical demonstration of his theory, support for the various elements of his argument can be found across the philosophical tradition. When rendered explicitly, Jaynes' theory is an important response to the functionalism assumed when we equate artificial intelligence with consciousness due to their functional equivalency. Just as famously Kant set limits upon reason to make room for faith, Jaynes limited the functions of consciousness in order to retain its independence as a primary phenomenon irreducible to any functional test. While the application of this theory to AI will be a topic throughout, in the third section I will focus on how Jaynes' work radically upends many of the basic assumptions in the current AI debate.

The Deed Over the Word: Reversing the Functionalism of the Faustian Bargain

> It says: "In the beginning was the *Word*."
> Already I am stopped. It seems absurd.
> The *Word* does not deserve the highest prize,
> I must translate it otherwise
> If I am well inspired and not blind.
> It says: In the beginning was the *Mind*.
> Ponder that first line, wait and see,
> Lest you should write too hastily.
> Is mind the all-creating source?
> It ought to say: In the beginning there was *Force*.
> Yet something warns me as l grasp the pen,
> That my translation must be changed again.
> The spirit helps me. Now it is exact.
> I write: In the beginning was the *Act*.[5]

In this passage, Goethe's Faust articulates the shift in philosophical first principles implied in his acceptance of Mephistopheles' demonic pact. The scholar curses the idealism of seeing the world as formed by the divine *logos* and demands that the deed (*die Tat*) instead be regarded as the first principle. In this statement, Goethe presages an inversion of the metaphysical order. As Marx would say a century later, the point of philosophy should no longer be to just interpret the world, but rather to change it.

---

[5] Johann Wolfgang von Goethe, *Faust*, trans. Walter Kaufmann (New York: Random House, 1961), p. 153, lines 1224-1237.

The Faustian pact would finally be sealed in the emergence of Anglo-American 20th century pragmatism, a philosophy of the deed.

Critics of this modernity tend to point out its Faustian nature and decry the ambitions of technology as attempts to "play God." But such conservative resistance to the Faustian pact only highlights its appeal: if mastery over the world through the technological apparatus is irreligious hubris, then the deed is the divine. God is the craftsman to whom we should defer in limiting our technological ambitions. Whether it is embraced or feared, the act has risen above the word. The technological landscape validates this metaphysical inversion, as any technology is nothing other than what it accomplishes. As Alan Turing established with his "Turing Test," the question of what a computer "is" is much less relevant than the question of what it can accomplish. If a computer can function like a human being, the question of the computer's internal state, of the presence of consciousness and free will, can be set aside as meaningless.[6]

Julian Jaynes enters this 20th century conversation as a research psychologist who broadly accepts the functionalism assumed in the scientific world of his day. In fact, his interpretation of mainstream positivism severs the activity of mind from all observable behavioral outcomes. This line of argument is supported by current technological developments. As technology advances, it becomes increasingly futile to argue for the functional necessity of consciousness where it has proven to be functionally irrelevant, as when a machine that passes the Turing Test. But where mainstream positivism generally concludes with skepticism about any positive theory of consciousness, Jaynes pivots in a surprising direction. Rather than discard consciousness entirely, we should accept that the word, or language more generally, cannot be reduced to a functional analysis. Language can (and, in fact, must) be severed from any functional outcome: there is more to a linguistic consciousness than the "outputs" it generates. It is now time to reverse the conceptual progression in Faust's rewriting of the opening of John's Gospel and retrieve the word in the wake of the deed's triumph.

Though he does not describe it as such, Jaynes arrives at this position through an essentially phenomenological method. We first recognize that our experience of the world has instilled a bias towards the overestimation of consciousness. We are only aware of the objects of our consciousness and so we naturally take the limits of our consciousness to be the limits of our entire psyche.[7] When consciousness looks back on its experience, it reconstructs everything in its own terms—as acts and objects of consciousness. In the classic problem of solipsism, consciousness is the circle that cannot escape itself. This results in our generating conscious narratives of that which never crossed the threshold of conscious awareness in immediate experience. We think

---

[6] Graham Oppy and David Dowe, "The Turing Test," *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), ed. Edward N. Zalta, https://plato.stanford.edu/archives/win2021/entries/turing-test/.

[7] Jaynes, *The Origin of Consciousness*, 23.

not of what was actually in consciousness, but of what "must" have been. For Jaynes, "memory is the medium of the must-have-been."[8]

For example, if I ask you to describe your drive to work, you will quite easily be able to tell me what "must" have happened for the vehicle to operate and the route to be accomplished, even if your consciousness was entirely absent, perhaps thinking of dinner plans or of an upcoming deadline. Surely you were conscious of the general route and anything notable which commanded your attention along the way. But if I press your recollection of the drive beyond general and attention-grabbing, you can only reconstruct your memories. How many cars were ahead of you in the left turn lane? Where were your hands on the steering wheel while turning? How long did you expect the yellow light to last while turning?

While all of us have paid conscious attention to such details at some time (i.e., when we were new to driving or especially aware of some abnormality), and while each of these details are materially relevant to achieving the task of driving to work, they probably never entered our immediate consciousness and certainly were never recorded in memory. Our inner sense of time is almost entirely unconscious and never an explicit measurement, even though having a sense of this timing is a matter of life and death on the road. We would likely feel uncomfortable driving with someone who thought it necessary to rigidly count down the yellow light every time they entered the intersection to make a left. Consciousness quickly evaporates the further we depart from an abstracted narrative of "what must have been" and the closer we approach our actual focus of our attention in driving.

Supporting this phenomenological intuition by citing several empirical psychological studies, Jaynes claims that most of our lives can be lived quite unconsciously.[9] In fact, most of the functions which are today claimed to show the potential for "consciousness" of artificial intelligence can be shown to have nothing to do with consciousness. A machine may be programmed to navigate a vehicle down a winding road. Perhaps some process of trial and error takes place in the programming of this AI, a process described as "machine learning." This capacity of the machine to "learn" and accomplish this task is taken as evidence of the machine's inevitable ascendence to consciousness. But Jaynes claims that such functional accomplishments should be entirely discounted as evidence of consciousness. One may first object that this ability is common to almost all vertebrates. Animals, humans, and machines can all figure out how to navigate themselves down a winding road. Where today many futurists are eager to proclaim that animal, machine, and human are *all* conscious on the basis of their functional capacity, Jaynes would say that precisely *none* of these three categories can be said to be conscious on the basis of a shared functionality.

But surely human learning is conscious? Do we not learn by paying attention to what we are doing and thereby acquiring a skill? For Jaynes, consciousness plays

---

[8] Jaynes, *The Origin of Consciousness*, 30.
[9] Jaynes, *The Origin of Consciousness*, 27-44.

only a formal role in the learning process. It frames the problem but is not itself active in the act of finding a solution. Consciousness supplies some general precepts which are only truly learned in unconscious activity, just as one learns to ride a bike from only a few general intentions. One can learn to play solitaire in a state of semi-consciousness, improving one's skill even while giving one's attention to a podcast. Indeed, the very possibility of multitasking should call into question the functional relevance of consciousness.

One may object that these kinds of skills are basically reflexes and so should not be considered representative of the higher reasoning unique to humans. For Jaynes, consciousness here again plays only a formal role. He offers an example this type of problem of inductive reasoning typically encountered in tests of intelligence:[10]

What is the next figure in this sequence?

While one must be conscious of the sequence itself to answer this question, one need not at all be conscious of the acts of reasoning which enable one to answer the question. Once the problem has come to consciousness, the solution suggests itself immediately. We have already solved the problem when we go back and write up a formal logic of how to define any *n*th term of the series. Having intuited the answer, we go back and write the formal rule of what must-have-been to give our answer a general validity.

When the problem is more complex, consciousness may play more of a role in framing the problem. One can "solve" a more complex sequence through abstract analysis. But while the formal approach would seem to replace unconscious insight with a fully conscious mechanism, the moment of unconscious reasoning is simply transposed over to the intuition behind the formula. There is no formula to generate a formula (that is, an original, non-derivative formula). While we can solve any term in the series by consciously applying the general rule, the general rule does not itself arise from any conscious effort. It is still an intuition, albeit one self-consciously tested and translated into rigorous terms. The skepticism of David Hume towards the validity of inductive reasoning illustrates precisely this point. The positive interpretation of Humean skepticism is the insight that the formal language of self-conscious reason is not self-grounding, as it depends upon the unconscious habit of association.[11]

---

[10] Jaynes, *The Origin of Consciousness*, 40. Figure reproduced from the text.

[11] David Hume, *A Treatise of Human Nature* (Oxford: Oxford University Press, 2012), Book I, Part III, Section VI.

In summary, we can say that while consciousness often aids and clarifies our psychic processes, it is not generally necessary for them. Chess, a subject of AI research for decades, can here be used as an example. One can play chess with a painstaking consciousness, as in correspondence chess, where a player can have months to consider and make a single move. Such games are expected to be of very high quality because of the intervention of consciousness, which continually reframes the problems of the position and tests out the answers supplied by intuition. It is generally agreed that this kind of slow chess is more beneficial for learners as it allows them to check and grow their intuition in tandem with conscious calculation.

On the other hand, one can play chess with very minimal consciousness, as in "bullet" chess, where all the moves are made in under one minute. While such games may contain more mistakes, they are often played at an extremely high level by accomplished players whose intuition has been trained so that they can find good moves by an act of reflex. A master playing bullet chess will almost always play to a higher standard than an amateur playing correspondence chess. Under the right conditions, it is not at all surprising that unconscious cognitive functioning outperforms conscious cognitive functioning. The overmatched amateur is in a position analogous to the master when they face any modern chess AI—that of being overcome despite the apparent advantage of consciousness. Consciousness certainly enables human beings to widen the scope of our problem solving, but the act of problem solving itself can happen just as automatically as the flow of electrons in a semiconductor.

For Jaynes, this deflationary account of consciousness is a preparatory step which will make plausible his wider hypothesis that human beings have only very recently become conscious. Art, architecture, and advanced civilization could all emerge alongside the evolution of consciousness, a process still not complete. While Jaynes lived before the current AI debates, I believe he would have seen the automation of so many human tasks as proof positive that consciousness never played a decisive role in them. Turning away from function as the proof of consciousness, Jaynes turned to language as the vehicle through which we can observe its evolution.

Language, Metaphor, and the Evolution of Consciousness

While Jaynes' deflation of consciousness is fairly straightforward, his account of language is much more difficult and controversial. Here I will try to bridge some of the argumentative gaps he leaves implicit as he moves on to present the historical evidence for his evolutionary theory.

Working within the post-Darwinian perspective of modern research psychology, Jaynes presupposes that consciousness must have evolved. Having abandoned the search for consciousness in any psychic function, he turns to language

as the very structure and substance of conscious experience, an "organ of perception."[12] We can unpack two reasons why language as such is so crucial for an evolutionary account of consciousness. First, consciousness is not a capacity for accomplishing but a state of mindful attention and intention which may or may not accompany our functional dealings with the world. Language is essential to this cultivation of attention insofar as it is only through language that we begin to see the world as containing discrete objects upon which we can fix our attention. This insight can be traced deep into the philosophic tradition. Per Anaxagoras, it is only the conscious mind (*nous*) which differentiates the immediate flux into the stable objects of perception and cognition.[13] Language is not just a supplemental tool in this process but its very organ. As the *Tao Te Ching* opens, "naming is the origin of all particular things." [14]

Second, if consciousness evolved, it must express itself in a form which is itself capable of evolution. Language is precisely such a medium in that it grows upon itself and has no final fixed form. It facilitates the coming-to-be of consciousness in its own expansion and bridges the gap between unconscious natural intelligence and conscious human intelligence. The meows and chirps of animal language are the first step towards the evolution of consciousness, the first form of bringing-to-attention even if still in the most automatic, instinctive way. Language can be produced and processed in the unconscious just as one need not be aware to scream or smile, but it is through such instinctive language that the possibility of consciousness first emerges. While Jaynes' complex account of intersubjectivity is beyond the scope of this paper, it is here important to emphasize that language remains essentially social, even if it is no longer regarded as a mere "tool" of communication. Consciousness emerges when instinctive signals become a matter of interpretation for the other, when they seek after the reasons behind our screams and smiles. A call to attention begets further attention, attention given not only to objective states of affairs but also to the other's own awareness as it is manifest in the medium of language. Wittgenstein was correct that the limits of our language are the limits of our world,[15] but these limits continually transcend themselves as language expands.

If language is the co-evolving vehicle of consciousness, it cannot be reduced to a static system of tightly defined references. In the terminology of American philologist and philosopher Phillip Wheelwright, this sort of language is "stenotyped," limited to one sense fixed by convention, as in modern scientific language.[16] It is to be contrasted with "tensive" language, language which is not referential but instead

---

[12] Jaynes, *The Origin of Consciousness*, 50.

[13] DK 59B12. Robin Waterfield, *The First Philosophers: The Presocratics and Sophists* (Oxford: Oxford University Press, 2000), 125.

[14] Lao-tzu, *Tao Te Ching*, trans. by Stephen Mitchell (New York: HarperCollins, 1988), 1.

[15] Ludwig Wittgenstein, *Tractatus Logico-Philosophicus* (California: Harcourt, Brace and Company, 1922), section 5.6, 74.

[16] Philip Wheelwright, *Metaphor and Reality* (Bloomington: Indiana University Press, 1962), 16.

extends its meaning out of itself, as in the constellation of meanings present in poetic ambiguity. While Jaynes himself participated in and validated scientific discourse, stenotyped language must have evolved only within the last few millennia of human existence, only achieving general adoption in the scientific revolution of the 17th and 18th centuries. The neurologist has something to learn from the philologist if they are to have any hope of how consciousness evolved. We can observe the emergence of consciousness in the metaphors of the epic tradition, as metaphor is the "very constitutive ground" of language.[17]

While he is not very explicit on this point, Jaynes seems to approach the radical thesis that all conscious understanding is essentially metaphorical. As language becomes more stenotyped, it only becomes a more abstract metaphor. To quote Jaynes' own evocative metaphor, abstract words are "ancient coins whose concrete images in the busy give-and-take of talk have worn away with use."[18] Abstraction always appears as a metaphorical extension of the concrete into a new semantic range. Jaynes points out that the irregular conjugation of the verb "to be" in modern Indo-European languages can be traced back to the Sanskrit verb *asmiy*, "to breathe."[19] Abstracting from our concrete human being to being in general, language opens up new horizons of understanding, as witnessed by the millennia of philosophical discourse on ontology generated from a metaphorical extension of the breath which underlies our existence as living creatures.

Even where language seems to have entirely shed its metaphoric origins it continually returns to metaphor to explain what lies beneath the well-worn linguistic currency. Philosophers should be familiar with this continual appeal to analogy to explain the most difficult concepts, beginning in the Platonic dialogues where allegories attempt to explain what the current sophistries cannot comprehend. In the context of the current discussion, we may say that Socratic questioning first shows the limitations of stenotyped language while Platonic allegory uncovers language's true metaphorical ground. On the more everyday side of linguistic evolution, the living adaptability of metaphor both enables and reflects our capacity for novelty, as when a new device that is only incidentally used to call people is described as a "smart phone." For Jaynes, metaphor is the living heart of language which enables a finite collection of lexical terms to extend beyond themselves to describe an infinite set of circumstances.[20]

One further argument from the history of philosophy may be adduced in support of Jaynes on this point. In its common understanding, metaphor seems to be a special case of linguistic use where, for effect, we say that some X is Y. Where a literal equivalency between X and Y fails, we are invited to contemplate them in their

---

[17] Jaynes, *The Origin of Consciousness*, 48.
[18] Jaynes, *The Origin of Consciousness*, 51.
[19] Jaynes, *The Origin of Consciousness*, 51.
[20] Jaynes, *The Origin of Consciousness*, 52.

similarity and difference. But philosophy has recognized since Aristotle that metaphor does not represent a special kind of proposition but rather is implied in the structure of all propositions.[21] Every proposition involves difference simply by the nature of the two terms involved. When we say "love is a wet dog" as opposed to "love is a powerful human emotion," the more metaphorical statement is only more metaphorical by a degree of difference. The equation of love with a wet dog invites contemplation, whereas we can accept the second proposition as a straight-forward definition.

But definition is more of a linguistic sleight of hand than the metaphor, as its accepted identity conceals the difference which the metaphor places out in the open. Socratic discourse will quickly show the neat equation of terms presented in a definition to be hasty and limiting at best. With our example of love, the predicate "powerful human emotion" could just as well apply to hatred, so there is at least the difference that the predicate has a wider range than the subject. We could, like a hapless Socratic interlocutor, attempt to clarify and say that love is a "positive and powerful human emotion," but once more we are refuted with the unhappy reminder that our experiences of love are not always positive. It is not our fault, however. Socrates was always playing with a loaded deck, knowing that all propositions are always asserting the identity of unlike things, expressing at once synthesis and distinction.

If metaphor is the ground of language and language is the ground of consciousness, consciousness must be in some sense metaphorical. This can be observed in two senses. First, metaphor serves as a bridge between the unconscious and the conscious. A metaphor is not generated from consciousness; indeed, poetry is good largely to the extent to which it is inspired without conscious mediation. The poetic intuition is a wellspring from which the metaphors bubble up as if under their own power. For Jaynes, consciousness is at first only the receiver of the gifts of the muses, a "bicameral" mind in which the still incomplete ego is only a vessel of externalized psychic entities: the muses, anthropomorphic gods, and ancestors whose voices inhabit and govern the ancient mind.[22] The metaphoric constructions they present are the pivot point for the emergence of consciousness as unified, self-contained subjectivity.

Consider a Homeric metaphor: a dying soldier's head droops like the head of a poppy soaked by rain.[23] In the immediate aesthetic effect of the metaphor, we

---

[21] Schelling describes the "ancient" understanding of the identity of the copula: "Whoever says, "The body is body," surely thinks something different with respect to the subject of the sentence than with respect to the predicate; with respect to the former namely, unity, with respect to the latter, the individual properties contained within the concept of body that relate to it as *antecedens* to *consequens*. Just this is the meaning of another ancient explanation according to which subject and predicate are set against each other as what is enfolded to what is unfolded (*implicitum* et *explicitum*)." F.W.J. Schelling, *Philosophical Investigations Into the Essence of Human Freedom*, trans. Jeff Love and Johannes Schmidt (New York: SUNY Press, 2006), 14.

[22] Inferring how the bicameral mind would have worked from studies of modern schizophrenics, Jaynes theorizes that the bicameral god expressed itself primarily through what would today be regarded as auditory hallucinations. Jaynes, *The Origin of Consciousness*, 85-94.

[23] *Iliad* 8.357-359.

unconsciously accept this identity of difference, dwelling in the imagery. But the comparison is also a prompt to conscious attention. How is the soldier like the poppy, and how is he not? What is it that makes this image so poignant? Direct literary experience gives way to literary criticism as an emergent consciousness attempts to clarify the difference and similarity of the metaphorical terms. The metaphor is not an accessory generated by a pre-existent consciousness. It is rather that which gives rise to definite consciousness from the loose manifold of unconscious inspiration. Only from a comparison can we arrive at a simple consciousness of any singular thing. What could it mean to be conscious of anything outside of its distinction from something else? Even at the most basic level of perception, pure light would be indistinguishable from absolute darkness. Consciousness, as the awareness of anything as a something, can only proceed from the distinction of that something from something else. Before we can say that A = A, an abstract and derivative point of view, we must wrestle with the powerful synthetic imagery of the unconscious mind which insists that A = B.

The evolution of consciousness as the gradual making explicit of a primeval poetic richness is a thesis that also can be observed in the intellectual history of the West. The Greek world undergoes this process when the acute consciousness of the Platonic dialogues, the philosophical search for exact definitions, begins to critically unpack the Homeric metaphors of the archaic culture. Reflecting upon the development of Greek intellectual life, the movement from poetry to prose became the archetypal example of historical "becoming" (*das Werden*) in 19th and 20th century German philosophy. As Hegel says in the preface of the *Philosophy of Right*, the self-conscious wisdom of philosophy, the owl of Minerva, only takes flight when a way of life has grown gray and old.[24] Nietzsche likewise argues in the *Birth of Tragedy* that the Socratic figure appears only when the Dionysian music has grown faint and subject to Socratic questioning.[25] In *Decline of the West*, Spengler expands this aesthetic hypothesis into a general theory of historical birth and decay in which an organic *Kultur* petrifies into a technocratic *Zivilisation*:

> Civilizations are the most external and artificial states of which a species of developed humanity is capable. They are a conclusion, the thing-become succeeding the thing-becoming, death following life, rigidity following expansion, intellectual age and the stone-built, petrifying world-city following mother-earth and the spiritual childhood of Doric and Gothic. They are an end, irrevocable. yet by inward necessity reached again and again.[26]

---

[24] G.W.F. Hegel, *The Elements of the Philosophy of Right*, trans. H.B. Nisbet, ed. Allen Wood (Cambridge: Cambridge University Press, 1991), 23.

[25] "Dionysus had already been driven from the tragic stage, and by a daemonic power which spoke through Euripides. Even Euripides was in a certain sense only a mask: the deity which talked through him was neither Dionysus nor Apollo but a newly born daemon called Socrates." Friedrich Nietzsche, *The Birth of Tragedy in the Spirit of Music*, trans. Douglas Smith (Oxford University Press, 2000), §12, 68.

[26] Oswald Spengler, *The Decline of the West: Form and Actuality*, trans. Charles Francis Atkinson (New York: Alfred A. Knopf, 1926), 31.

The movement of literature mirrors the movement of human society; it is a movement from poetry to prose and then back again. Although he does not mention this interpretation of history in 19th and 20th century German philosophy, Jaynes builds upon this tradition when he characterizes the evolution of consciousness as a coming-to-awareness in the wake of an earlier unself-conscious poetic moment. McGilchrist furthers this tradition when he says that the right brain ("the master") has primacy over the left (its "emissary). In accepting this "primacy of the implicit," we realize that "metaphorical meaning is in every sense prior to abstraction and explicitness." Returning to the original Latin metaphor contained in these worlds, "pulling away" (from *abs-trahere*) and "unfolding" (*ex-plicare*) are acts of analysis which depend upon more primal unity.[27]

The Metaphor of Mind

While I have liberally reconstructed Jaynes' diffuse insights into a more explicit argument for metaphor as a bridge between the conscious and unconscious, he is much more direct in presenting a second association between metaphor and consciousness. It is not only that metaphor prompts the emergence of consciousness, but that consciousness is *itself* a metaphor. It is the creation of an analog mental "space" in which the analog "I" operates as if it had a visuospatial reality. Jaynes writes:

> [Consciousness] operates by way of analogy, by way of constructing an analog space with an analog 'I' that can observe that space and move metaphorically in it. It operates on any reactivity, excerpts relevant aspects, narratizes and conciliates them together in a metaphorical space where such meanings can be manipulated like things in space. Conscious mind is a spatial analog of the world and mental acts are analogs of bodily acts.[28]

The "I" operating within abstract mental space is like the well-worn coin whose concrete imagery has faded in accustomed use. We can observe what a more concrete metaphor of consciousness would be in poetic language: "The heart desires but the hands are unwilling" is a more concrete way of saying "I am conflicted in my decision." Indeed, Jaynes theorizes that the ancient world first attempted to describe consciousness by describing it as a faculty localized in different semi-autonomous body parts, like the *thumos* ("spirit") which often appears as rousing the limbs in Homeric heroes.[29] By contrast, in the modern understanding, the simple unity of the

---

[27] McGilchrist, *The Master and His Emissary*, 179.
[28] Jaynes, *The Origin of Consciousness*, 65-66.
[29] Jaynes, *The Origin of Consciousness*, 69.

first-person pronoun gathers consciousness into a single selfsame "space," an identity without difference.

The shift between these two metaphors of mind is not merely a change in descriptions of the same phenomenon. Consciousness operates through these metaphors; when the metaphor changes, so does consciousness. The concept of the self which operates in consciousness at any time is the living organ through which that self grows and actualizes itself. If I am a computer, I will operate by a rule of calculation. If I am a raging bull, I will leave a trail of destruction in my wake. When these metaphors prove insufficient to my lived experience, I am in an existential crisis. The metaphor must either grow or die off and be replaced.

But as much as the self-fulfillment of metaphor can be observed in individual psychologies, Jaynes is more concerned with the general historical development of self-consciousness. The consolidation of consciousness in the "I" is the standpoint of objectivity, the Cartesian division between self and world upon which the scientific products of modern culture depend. It is the metaphor which conceals itself as metaphor, the creation of a "head-space" which could no more be spatially located in the head than in the feet.[30] It is "attention" marked with any act of actual attending, a purely mental "presence." It is hermetically sealed off from the body, which has lost its autonomy and is now subordinated to an abstract mentality. Except in now quaint metaphors, the heart and stomach no longer speak for themselves or directly motivate actions; they rather belong to an "I" who possesses them as influences held at a distance. Likewise, social and religious influences lose their immediate inspiration. Ancestors, gods, and muses do not directly partake in our individuality and can only intrude on the autonomous operations of rational self-consciousness.

McGilchrist describes a world ruled by this metaphor of mind as one in which all the inspired idiosyncrasies of personal consciousness have been eliminated as the dominant left-brain (the vehicle of the abstract "I") devalues and even pathologizes alternative metaphors.[31] Modern life is trending towards this dystopia, one where "the concepts of skill and judgment [. . .] would be discarded in favor of quantifiable and repeatable processes."[32] All the psychic phenomena which cannot be assimilated to this "I" are demoted to the status of unconsciousness, a shadow self which exacts its vengeance in many of the illnesses of modern culture.

For Jaynes, this shadow self takes a clinical form of schizoid mental illnesses in which the forgotten world of gods, muses, and ancestors intrudes upon the self-narrative of an "I" which cannot recognize these voices as its own voices. The bicameral mind returns but without the mediating structures (shamans, rituals) of ancient society. Even if the victory of the autonomous "I" is secured, it has won at a high cost evident even in the non-clinical illnesses of modern life. The "I" has an

---

[30] Jaynes, *The Origin of Consciousness*, 44-46.
[31] McGilchrist, *The Master and His Emissary*, 428-434.
[32] McGilchrist, *The Master and His Emissary*, 429.

agenda irreconcilable with a body it regards as "other," and so it disregards the "voices" of the old Roman god Somnus and suffers sleep deprivation. Without the meaningful influence of a historical past, the isolated individual is vulnerable to the appeal of atavistic nationalism, the suppressed "call" of the ancestors possessing the modern individuals.[33]

Understanding itself as the master of practical efficacy, this "I" "attaches an unusual importance to being in control."[34] Withdrawn into itself, the autonomous ego proves validates its independence in functional terms, by its ability to command and control the external world. When the technologies it produces too nearly replicate its own operation, this metaphor of mind undergoes an ironic twist. Whereas it had established itself as sovereign over a passive, inert material world, it now finds itself struggling to explain how technology belonging to that world can seemingly replicate its own mental functions. A prisoner to its own functionalist presuppositions, the scientific consciousness which once combated animism now finds itself spinning new metaphors to explain the apparent "consciousness" of its technologies. If the mind is only what it can do, we can only return back to animism when our mental feats are equaled.

Rethinking the Metaphor of "Artificial Intelligence"

With AI, the currently dominant metaphor of human consciousness is being retrofitted to describe a novel human technology. The main sense of the metaphor is clear and uncontroversial enough. "Artificial intelligence" describes certain programs that perform functional tasks generally associated with intelligence. Such a purely functional definition is appropriate because the goal of AI was always only functional. Scientists never set out to recreate a human mind as such but to *improve* upon it in executing programmable tasks. There would be no point in even bothering to design AI systems if the goal were not to *surpass* natural and human intelligence on a purely functional basis. A machine that passes the Turing Test in answering customer service calls is not identical to a human doing the same job, it is *superior*. The computer will not tire like a human and so it can better perform the task of directing inquiries to customer service, achieving the goal the engineers have set for themselves. In recognizing this functional superiority, we likewise recognize the differences between AI and human intelligence which can be concealed in accepting a metaphor (which always contains the tension of difference) as a false and loose equivalency.

This linguistic sleight of hand at play in the entire AI debate in which what would be honest metaphors masquerade as dishonest definitions. Novelty always

---

[33] Jaynes suggests that such relapses into a bicameral state can be observed in modern nationalism, using imperial Japan as an example. Jaynes, *The Origin of Consciousness*, 159.

[34] McGilchrist, *The Master and His Emissary*, 432.

prompts a search for new metaphors of understanding, and the first responses to this search generally prove themselves to be inadequate in time. "Artificial intelligence" and "machine learning" are not scientific definitions but first attempts at metaphor to describe a still-evolving technology. The main intention behind these terms is clear enough, but metaphor, in striving to be adequate to what is ambiguous, is necessarily and productively imprecise.[35] Every metaphorical device contains within it a constellation of associations lying alongside the main comparison. For instance, if I say that "love is a battlefield," the most likely sense of the metaphor is that love is more cruel and destructive than usually thought, but this is not the only sense possible. Battlefields are also sites for noble and heroic action, for great mourning and reverence, even for camaraderie. The point of the metaphor is not that one of these interpretations must be chosen to the exclusion of all the others, but that all are somehow operative at once, even if only as potential meanings lurking in the unconscious.

Likewise, when we say "artificial intelligence" or "machine learning," we are suggesting more than the basic intention of the metaphor to convey a certain functional capacity. The self-interested proponents of this technology are exploiting a vacant linguistic frontier to establish a compelling metaphor which also connotes the spontaneity, organicity, and inner mental space of consciousness. The metaphor cashes in on our overestimation of the conscious "I." If, as we generally believe, the conscious "I" is indispensable to all forms of thinking, a machine that "thinks" as well as a human must also have all the other qualities of human consciousness. Dazzled by the functional novelty of the technical accomplishment, we unconsciously accept the associations implied by the metaphor. We do have words which could more plainly describe what is happening in AI, but "applied machine binary calculation" (AMBC) is not a term which will promote a general trust in computers as anthropomorphic beings. The artificially intelligent phenomena now interpreted as organic and insightful would now carry the semantic burden of the world of machines, more akin to the activity of an advanced calculator than a human interlocutor.

There is something instructive in the "artificial" part of the metaphor. AI has only been able to achieve its functional accomplishments by reversing the operation of human intelligence, which begins in metaphor and ends in formal rigor. As a purely formal system, AI is only able to achieve superior technical results by virtue of the specialized dedication of a great mechanical computing power towards a single task, something impossible for a human being who is always also breathing, observing, feeling. This intelligence is "artificial" in the sense that it takes one mental function and isolates it, purifying it of all other cognitive and biological context like a naturally occurring compound which has been refined in a laboratory.

---

[35] "If [metaphor] is not to be escapist and merely a stubborn refusal to face things as they are, it will bear traces of the tensions and problematic character of the experience that gave it birth." Wheelwright, *Metaphor and Reality*, 46.

If our intelligence is like that of AI, we should regard doing quick mental math as the epitome of human intelligence. We should teach our children that the best way to read a book is to scan out the frequency of the words and begin with a statistical analysis. But, try as we might, we will never be able to out-calculate the machine. We can accept this with shame and resignation, or we can do what consciousness has always done and reexamine the metaphors. Metaphor demands both that we see what is similar and what is different. The term "artificial intelligence" has disclosed a certain functional similarity of new technologies to some human capacities, but the differences must now be retained and emphasized as the metaphor evolves.

Nonetheless, the technological aides currently referred to as "AI" will be, for the foreseeable future, one of the components of the human "I." They will inhabit our mental space and be considered in our decisions no less than our knowledge of history, our sense of ethics, and our aesthetic judgements. Like the muses of ages past, calculation aides can function in an almost revelatory way, disclosing whole new horizons of knowledge such as when, through the sheer power of the mainframe, they adopt strategies in chess never considered by any human. But they are only aides, not replicas of the "I" assumed to have anthropomorphic qualities just like the Greco-Roman gods. If Julian Jaynes were alive today, I believe he would remind us that the age of the bicameral mind is past, and that we should not return to it by further overextending the metaphor of the "I." Attributing the "I" to whatever technology surpasses human beings on a functional basis will only create a new bicameral world in which technology appears as an alien sovereign issuing schizoid pronouncements to despairing humans.

Avoiding this dystopian fate requires clarity about consciousness just as much as clarity about machines. This is only our destiny if we interpret ourselves as a processing power which would be overthrown if eclipsed by machines. We are not only this functioning, this doing, but also this interpreting, this self-creating. The bicameral world broke down only when the metaphor of mind changed so that we heard our voices as our own. In the modern world we recognize that the Greco-Roman gods and muses always lived within us. The challenge for the next metaphor of mind will be to incorporate forms of artificial intelligence into the sphere of human subjectivity without treating them as if they were themselves individual subjects. To grant AI the autonomy of the "I" would be a failure to meet this challenge, a new breakdown in the metaphor of mind and a repudiation of modernity undertaken, ironically, in the celebration of scientific progress.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Martin Heidegger's Concept of *Understanding* (*Verstehen*): An Inquiry into Artificial Intelligence

Joshua D. F. Hooke
Memorial University of Newfoundland, Canada
ORCID: 0009-0002-8794-8488

Abstract

My primary goal in this paper is to demonstrate the inadequacy of Hubert Dreyfus' use of *understanding* (*Verstehen*) for Artificial Intelligence (AI). My complementary goal is to provide a principled account of Martin Heidegger's concept of *understanding* (*Verstehen*). Dreyfus and other verificationists argue that *understanding* (*Verstehen*) is socially purposive action and skillful embodied coping. *Understanding* (*Verstehen*), conceived of in this way, purportedly challenges cognitive models of Artificial Intelligence (AI) that rely on formal rules, 'rational' decision-making, and the explicit representation of knowledge. This account is unsatisfactory for two reasons. First, it maintains an extrinsic, goal-oriented intentionality that is susceptible to the success of Artificial Intelligence (AI). Second, it ignores the systematic and constitutive analysis of self-understanding (*Seinsverständnis*) that is fundamental to Heidegger's ontology. Recent exegetical work replicates these inadequacies and fails to improve discussions on Heidegger's relationship to Artificial Intelligence (AI). To resolve this oversight, I bridge the gap between Heidegger's concept of *understanding* and *disclosedness* (*Erschlossenheit*) (SZ §44 / 256-278). I argue that *understanding* characterizes the pre-theoretical grasp of entities and the pre-ontological structure that initiates the question of self-understanding (*Seinsverständnis*). This result supports Heidegger's phenomenological breakthrough towards a sense of Being (*Sein*) as the ground of intelligibility.

Keywords: Martin Heidegger, Hubert Dreyfus, understanding, know-how, disclosedness, phenomenology

Introduction

The verificationists argue that Heidegger transforms Edmund Husserl's transcendental phenomenology into hermeneutic facticity.[1] Theoretical knowledge housed in representational, conceptual, or propositional terms are symptomatic pervasions of Cartesianism dominating Western philosophy. Husserl's phenomenology shares the same theoretical distortions as Descartes and is criticized for its similar inattention to the being of consciousness.[2] Heidegger seeks to renounce the tradition of theoretical knowledge whereby self-awareness and the primacy of consciousness are privileged over concrete and historically embedded understanding.[3] Heidegger's project, then, advances an a-theoretical, non-objectifying, and non-reflective form of practical *understanding*.[4] In this context, *understanding* is 'knowing how' to be skillful in a social *milieu*.[5] Dreyfus uses Heidegger's critique of theoretical subjectivity to differentiate skillful acting (knowledge-how) from theoretical or conceptual thinking (knowledge-that). As such, *understanding* and *know-how* are synonymously conceived to critique 'rational' assumptions in AI research.[6]

---

[1] This group of commenters is typically referred to as the Anglo-American Pragmatists. I use the term "verificationists" to broaden the scope of my critique to include commentators who maintain an outcome-based or goal-oriented criterion for knowledge. To name a few, see, Charles Guignon, *Heidegger and the Problem of Knowledge* (Indiana: Hackett, 1983). Carl Friedrich Gethmann, "Zu Heideggers Wahrheitsbegriff," *Kant-Studien* 65, no. 2 (1974): 186-200. Friedrich-Wilhelm Von Herrmann, *Hermeneutics and Reflection: Heidegger and Husserl on the Concept of Phenomenology*, trans. Kenneth Maly (Toronto: University of Toronto Press, 2013). Graham Harman, *Tool-Being: Heidegger and the Metaphysics of Objects* (Illinois: Open Court, 2011). Hans-Georg Gadamer, *Truth and Method,* trans. Joel Weinsheimer, Donald G. Marshall (London: Continuum, 2004). Hubert L Dreyfus, *Being-in-the-world: A Commentary on Heidegger's Being and Time, Division I* (California: MIT Press, 1990). Hubert L. Dreyfus, *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action,* ed. Mark Wrathall (New York: Oxford University Press, 2014). Mark A. Wrathall, *Heidegger and Unconcealment: Truth, Language, and History* (New York: Cambridge University Press, 2010). Mark Okrent, *Heidegger's Pragmatism: Understanding, Being, and the Critique of Metaphysics* (New York: Cornell University Press, 2019). Richard Rorty, *Essays on Heidegger and Others: Philosophical Papers Vol. 2* (New York: Cambridge University Press, 1991).

[2] Theodore Kisiel, *The Genesis of Heidegger's Being and Time* (California: University of California Press, 1995), 280. See, GA 17: 254.

[3] Sean McGrath, "The Early Heidegger's Critique of Husserl," in *From Between Description and Interpretation: The Hermeneutic Turn in Phenomenology*, ed. Andrzej Wiercinski (Toronto: The Hermeneutic Press, 2005), 269.

[4] Kisiel, *The Genesis of Heidegger's Being and Time*, 47, 376.

[5] Gadamer, *Truth and Method*, 20.

[6] These include biological, psychological, epistemological, and ontological assumptions. Each are given a respective chapter in Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (California: MIT press, 1992).

There are *prima facie* similarities between Heidegger's ontology and pragmatism. Hermeneutic or practical understanding, however, is not the final level of Heidegger's analysis. Heidegger cautions against fore-closing phenomenology as theoretical or practical (GA 21: 11).[7] Heidegger states that the objective determinations of life, motivated by practical understanding, cannot achieve existential meaning, or what he calls "care" (*Sorge*) (SZ: 191f, 249f, 284, 316, 328, 350; GA 21: 11). These claims are first announced in the *Frühe Freiburger Vorlesungen*, 1919–1923, where Heidegger, following Heinrich Rickert, critiques *Lebensphilosophie* (GA 61: 119, 121). Heidegger claims that an a-theoretical "preconception toward grasping" life never leads to the proper sense of "caring and its categorial determinations" (GA 61: 100).[8] To interpret the world in a way that "goes along with" a predetermined ordering of things is insufficiently radical, and those who do so are "too ready to accept traditional determinations" (GA 60: 134).[9]

In the first section of this paper, I introduce the verificationist account of *understanding*. In the second section, I present a two-pronged critique of Dreyfus' account of *understanding*. First, I argue that Dreyfus' account is unsatisfactory given

---

[7] *Logic: The Question of Truth*, trans. Thomas Sheehan (Bloomington: Indiana University Press, 2010), 11. Let me at once introduce the other works of Martin Heidegger to which reference will be made in the present article. Martin Heidegger, *Being and Time*, trans. John Macquarrie, Edward Robinson (New York: Harper & Row, 1962); *Sein und Zeit* in *Gesamtausgabe*, vol. 2, ed. F. W. von Hermann (Frankfurt am Main: Klostermann, 1977). References to this text will be made using the abbreviation "SZ" followed by the paragraph number. On occasion, references are made to the section number, indicated by a pilcrow (§). The *Gesamtausgabe* is cited hereafter as GA followed by the volume number; all volumes of the GA are published by Klostermann in Frankfurt am Main. Page references are from the English translations. (GA 17): *Introduction to Phenomenological Research*, trans. Daniel O. Dahlstrom (Bloomington, Indiana University Press, 2005). (GA 19): *Plato's Sophis*t, trans. Richard Rojcewicz, André Schuwer (Bloomington: Indiana University Press, 1997). (GA 29/30): *The Fundamental Concepts of Metaphysics: World, Finitude, Solitude*, trans. William McNeill, Nicholas Walker (Bloomington: Indiana University Press, 1995). (GA 56/57): *Towards the Definition of Philosophy*, trans. Ted Sadler (London: Bloomsbury Athlon, 2000). (GA 58): *Basic Problems of Phenomenology: Winter Semester 1919/20*, trans. Scott M. Campbell (London: Bloomsbury, 2013). (GA 59): *Phenomenology of Intuition and Expression*, trans. Tracy Colony (London: Bloomsbury Continuum, 2010). (GA 60): *Phenomenology of Religious Life*, trans. Matthias Fritsch, Jennifer Anna Gosetti-Ferencei (Bloomington: Indiana University Press, 2004). (GA 61): *Phenomenological Interpretations of Aristotle: Initiation into Phenomenological Research*, trans. Richard Rojcewicz (Bloomington: Indiana University Press, 2001).

[8] In *The Heidegger Dictionary,* Dahlstrom notes that existentials are categories of Dasein's Being that make up its existentiality. See, SZ: 12f, 42f, 53, 183ff, 201, 212, 232f, 260, 298, 302f, 304. *Disposedness, Understanding, Discourse, Fallenness* are Dasein's "most general structures" (SZ: 270, also, see SZ, 134, 143, 148, 150, 160, 336). *Existence*, *Facticity* and *Fallenness* are existential determinations that make up the fundamental ontological character of care (SZ: 191f, 249f, 284, 316, 328, 350). *Fallenness* is an existential mode of being-in-the-world (SZ: 176). *Truth* is a fundamental existential (SZ: 297). Heidegger states that from categorial interpretation we will acquire an exposition of the basic sense from which all *existentialia* take proper and referential sense.

[9] Daniel O. Dahlstrom, *Heidegger's Concept of Truth* (New York: Cambridge University Press, 2001), 199.

recent and foreseeable developments in AI.[10] I support this critique by asserting that Dreyfus' account maintains a goal-oriented intentionality that is vulnerable to the success of AI. Second, I argue that Dreyfus' account, along with other verificationist approaches, is deflationary and fails to capture the fundamental insight of Heidegger's ontology.[11] In my concluding remarks, I briefly suggest the conditions that AI must satisfy to replicate a Heideggerian account of human existence.

Heidegger's Verificationism

Heidegger argues that pre-theoretical 'lived experience' is an unavoidable moment in the emergence of meaning, and "life experience is more than, [*pace* Husserl], the mere experience which takes 'cognizance of'" (GA 60: 8). Experience designates the "active and passive pose of the human being toward the world" (GA 60: 8). As such, pre-theoretical life stems from the surrounding 'environing' world and brings the pre-theoretical familiarity that grants access points to meaning. Even the most trivial experiences in our everyday lives provide the pre-theoretical context of meaning. To illustrate this point, Heidegger describes what happens when we encounter the lectern standing in the classroom. In one stroke, the lectern is given to the professor, the students, and any observers (familiar or unfamiliar with lecterns) right away '*as* something.' Accompanying the lectern is a complex relation of associated objects and

---

[10] Mark Wrathall argues that Dreyfus tends to attribute his insights to other philosophers (esp. Heidegger and Maurice Merleau-Ponty). In the first part of section two, my critique of Dreyfus stands irrespective of whether his account is attributed to Heidegger.

[11] It is worth noting that the verificationism latent in the pragmatist reading is also an attempt to reconcile Ernst Tugendhat's long-standing critique of Heidegger's concept of truth. In §44 of *Being and Time*, Heidegger characterizes the phenomena of *disclosedness*, *uncovering* (*Entdeckenheit*), or *ἀλήθεια* as the preconditions for propositional truth. Tugendhat argues that these preconditions lack bivalence, and therefore cannot be deemed truth. The pragmatists forgo the core of the existential analysis of truth as disclosedness in exchange for 'background social practices.' The success or failure of background coping (e.g., equipment uses, appropriate normative behavior, and so on) is publicly verifiable and satisfies Tugendhat's conditions of bivalence. Dahlstrom notes that if the pragmatic interpretation succeeds, and if the interpretation is valid, then one would have a reason to reject Tugendhat's objections. A more detailed consideration of this debate lies beyond the present study. See, Ernst Tugendhat, *Über den Wahrheitsbegriff bei Husserl und Heidegger* (Berlin: Veröffentlicht von de Gruyter; Reprint 2012 ed. edition, 1967), 259f. Ernst Tugendhat, "Heidegger's Idea of Truth (1964)" in *The Heidegger Controversy: A Critical Reader*, ed. Richard Wolin (Cambridge: The MIT Press, 1993), 245-263. William H. Smith, "Why Tugendhat's Critique of Heidegger's Concept of Truth Remains a Critical Problem," *Inquiry* 50, no. 2 (2007): 156-179. For a critical response, see, Carl F. Gethmann, "Zu Heideggers Wahrheitsbegriff," *Kant-Studien* 65, no. 2 (1974): 186-200. Jens Greve, "Heideggers Wahrheitskonzeption in Sein und Zeit, Die Interpretationen von Ernst Tugendhat und Carl Friedrich Gethmann," *Zeitschrift für philosophische Forschung*, H. 2 (2000): 256-273.

ideas understood and preserved through individuated lived experiences.[12] Heidegger states that "everything that is experienced in factical life experience, as well as all of its content, bears the character of significance" (GA 60: 9). Immediate significance indicates that the lived experience does not entail universality or absoluteness concerning objects. The worldly character of life guides a 'primordial anticipation' and 'mobility of life' that precludes 'freeze-framing' states of affairs. Through the contextualized lived experience of both selfhood and objects, human beings develop an understanding of both entities, which, in turn, serves as the basis for constructing phenomenological concepts and linguistic content about them.

Influenced by Heidegger's critical analysis of Husserl's theoretical subjectivity, the verificationists argue that human beings are not individual, agential, and rational.[13] On the contrary, human beings are embedded, embodied, and absorbed in their environment. The verificationists rely predominantly on the hermeneutical "*as*-structure" in *Being and Time* to substantiate their interpretation (SZ: 140-160).[14] The "*as*-structure" is the pre-theoretical understanding of objects that give shape and context to our interpretation of the world. When we see an object, we already understand it *as* something it is because of its context and use (GA 21: 144). Heidegger states that when we "'know our way around' [*Umgang*] the world, every act of having something before our eyes . . . is in and of itself a matter of 'having' something *as* something" (GA 21: 144). Accordingly, the three-fold structure of the hermeneutical *as*-structure consists of the following distinctions. First, our pre-linguistic practical understanding of objects (e.g., understanding the chalkboard *as* something for writing on or a hammer *as* a tool for driving nails). Second, the use of interpretative assertions to express difficulty or the inability to cope with equipment (e.g., "this hammer is not the right tool for the job") (SZ: 155; GA21: 157). Third, the use of theoretical assertions to express a particular determination of an object *as* something occurrent (e.g., "the hammer is heavy") (SZ: 155). Contrary to the empiricist perspective, seeing something transcends mere observation of its physical qualities. Objects are revealed

---

[12] Jonathan O'Rourke furthers the epistemological claim that that "the objects [in] my environment are disclosed according to the sorts of normative roles I take part in, as a student, as a brother, as a friend, etc. Even those objects of which I am unfamiliar, precisely through their instrumental strangeness, are given to me in the relief of this same meaning context." Jonathan O'Rourke, "Heidegger on Expression: Formal Indication and Destruction in the Early Freiburg Lectures," *Journal of the British Society for Phenomenology*, (2018): 49: 2, 11 https://doi.org/10.1080/00071773.2018.1431133.

[13] See, Edmund Husserl, *Ideas for a Pure Phenomenology and Phenomenological Philosophy: First Book: General Introduction to Pure Phenomenology*, trans. Daniel Dahlstrom (Indianapolis: Hackett, 2014), §46f, 103f, l4lf. For a critical response, which some say Heidegger appropriates, see, Paul Natorp, *Allgemeine Psychologie* (Tübingen: J.C.B. Mohr, 1912), 8, 28-9, 3.

[14] Several exegetical accounts can be found on Heidegger's "*as*-structure." Relevant for the present study, see, C.F. Gethmann, *Verstehen und Auslegung: das Methodenproblem in der Philosophie Martin Heidegger* (Bonn: Bouvier Verlag, 1974). Mark A. Wrathall, ed. *The Cambridge Heidegger Lexicon* (New York: Cambridge University Press, 2021), 64f. Dreyfus, *Being-in-the-world*, 60f, 184f. Dahlstrom, *Heidegger's Concept of Truth*, 181, 305.

within a network of relations through their serviceability, signifying what they are intended for. For the verificationists, perceptual experience is ingrained in pragmatic and social contexts, imbuing worldly objects with practical significance that compels us to act upon them in pre-predicative ways.

The hermeneutic "*as*-structure" underscores the way we encounter the world. Lucilla Guidi suggests that "the *as*-structure is a constitutive feature of every experience of entities in the world—namely, the way they always present themselves in terms of a 'for something.'" [15] Therefore, the basis of conceptual judgment relies on skillful practices as the 'pre-theoretical' and 'original' ways of interacting with objects. In other words, conceptual understanding and propositional content are derivative of 'know-how.'[16] Martin Weichold quotes *Being and Time* to substantiate this interpretation: "Understanding . . . is not a knowledge derived from cognition, but a primordially existential kind of being which first makes knowledge and cognition possible" (SZ: 123f). Weichold interprets Heidegger as suggesting that this respective understanding is an ability (SZ: 143).[17] Just as a neuroscientist "reads" the pictures of a brain scan and provides a diagnosis, human beings "read" the world to deal with their environment.[18] For the verificationists, Heidegger's fundamental insight is that knowledge is practical understanding derived from absorbed intentionality prior to

---

[15] Lucilla Guidi, "As-Structure (*Als-Struktur*)," in *The Cambridge Heidegger Lexicon*, ed. Mark A. Wrathall (Cambridge: Cambridge University Press, 2021), 64.

[16] See, Hubert Dreyfus, "Overcoming the Myth of the Mental: How Philosophers can Profit from the Phenomenology of Everyday Expertise," in *Proceedings and Addresses of the American Philosophical Association* 79, no. 2 (2005): 47-65. Hubert Dreyfus and Stuart Dreyfus, *Mind over Machine*. (New York: Simon and Schuster, 2000). Both readings draw predominantly on Heidegger's concepts of *understanding (Verstehen), interpretation (Auslegung),* and *circumspection (Umsicht)*.

[17] The full quote reveals that Heidegger is not making claims about ontic knowledge but disclosedness and the problem of other minds. Quoted in full, Heidegger states: "The disclosedness of the Dasein-with of Others means that because Dasein's Being is Being-with, its understanding of Being already implies the understanding of Others. This understanding, like any understanding, is not an acquaintance derived from knowledge about them, but a primordially existential kind of Being, which, more than anything else, makes such knowledge and acquaintance possible" (SZ: 123). Of greater importance, Heidegger states: "When we are talking ontically we sometimes use the expression 'understanding something' with the signification of 'being able to manage something,' 'being a match for it,' 'being competent to do something'" (SZ 143). This quote showcases Heidegger's method of formal indication. Long overlooked as a tangential method in *Being and Time*, formal indication utilizes conventional and commonplace meanings of words to introduce figurative interpretations that ultimately reveal existential implications. As provisional indicators, Heidegger uses these terms to establish genuine connections that ordinary words merely signify. In the process, the inadequacy of the initial use of a term is exposed, and the underlying existential content that it implicitly presupposes is brought to light. By failing to see that the term *Understanding* is formally indicative, Weichold's analysis is misleading and remains at the level of conventional use.

[18] Dreyfus and Dreyfus, *Mind over Machine*, 16f, 101f.

representational intentionality.[19] Consequentially, intelligent behaviour cannot be measured by deliberately thinking about 'facts' and 'properties' of consciousness.[20] Intelligent behavior is characterized by a pre-conceptual practical understanding that non-deliberatively and non-consciously provides information about the world. As such, the world of objects is not constituted by our subjective consciousness. Dreyfus states that "when actions involve any experience at all, it is not an experience of oneself as *causing* one's activity, but rather of a direct responsiveness to the environment whereby one's activity is completely geared into the demands of the situation."[21] Dreyfus claims that "mindedness" is "the enemy of coping" because "we are not minds at all, but one with the world."[22]

The verificationists see the hermeneutical "*as*-structure" or "background practices" as the ontological significance of language. Heidegger's analysis of lived experience is the pre-linguistic or non-conceptual practical basis for our linguistic activity.[23] Heidegger is credited with avoiding the problematic conditions of correspondence theories of truth by dissolving theoretical constitutive subjectivity.[24] Following this line of thought, Carl F. Gethmann argues that Heidegger replaces the traditional correspondence model of truth with an "operational model."[25] Accordingly, the "success" and "serviceability" of the action fulfill the conditions of truth "even if it is not asserted at all."[26] Gethmann argues that the "representation of an action, in a sentence, is the meaning of agreement in a propositional model of truth . . . An underlying operational truth relates to a proposition like a key to a lock."[27]

---

[19] See, Hubert Dreyfus, "The Socratic and Platonic Basis of Cognitivism," *AI and Society* 2, no. 2 (1988): 99-112. Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge: Cambridge University Press, 1992).

[20] Dreyfus, *Being-in-the-World,* 62, 81-2, 84.

[21] Dreyfus attributes these views to Heidegger, suggesting that "Heidegger, indeed, claims that skillful coping is basic, but he is also clear that, all coping takes place on the background coping he calls "being-in-the world" which doesn't involve any form of representation at all." Hubert Dreyfus, "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian," *Philosophical psychology* 20, no. 2 (2007): 254.

[22] Dreyfus, "The Return of the Myth of the Mental," *Inquiry* 50, no. 4 (2007): 353.

[23] See page 59, Paul Livingston, "The Ontology of Sense and "Transcendental" Truth: Heidegger and Davidson" in *The Logic of Being: Realism, Truth, and Time* (Illinois: Northwestern University Press, 2017), 59-95. The meta-grammatically truth structure is often referred to as "circumspective unconcealment." Extensive treatment is given throughout the following: Graham Harman, *Tool-Being: Heidegger and the Metaphysics of Objects* (Illinois: Open Court, 2011). Lee Braver, *Groundless Grounds: A Study of Wittgenstein and Heidegger* (Cambridge: MIT Press, 2012). Like Dreyfus, these readings draw predominantly on Heidegger's *Understanding* (*Verstehen*), *Interpretation* (*Auslegung*), and *Circumspection* (*Umsicht*).

[24] Wrathall, *Heidegger and Unconcealment*, 47.

[25] Gethmann, "Zu Heideggers Wahrheitsbegriff," 198. Translation mine

[26] Gethmann, "Zu Heideggers Wahrheitsbegriff," 198. Translation mine

[27] Gethmann, "Zu Heideggers Wahrheitsbegriff," 198. Translation mine

Gethmann contends that the meaning of agreement in a propositional model of truth (i.e., truth as correspondence) is rooted in the representation of an action within a sentence. Accordingly, an underlying operational truth functions like a key, unlocking the meaning of a proposition. For Gethmann, "whether the key 'agrees' with the lock, shows itself in locking or unlocking the door, hence in its use, and not in talking about it."[28] As such, Heidegger's operational model challenges traditional conceptions that fulfil their truth criteria by relying on acts of consciousness and propositional content.

Mark Okrent, like Gethmann, argues that Heidegger's operational conception of truth modifies Husserl's conception of truth (i.e., a modification of *adequatio intellectus et rei*).[29] In the Husserlian sense, truth is an intentional act that 'adequately' reflects the intuited object given to consciousness. In Heidegger's modification, the intended meaning or proposition 'adequately' verifies an operational truth, and intuition takes the form of a reactive ability in a purposive action.[30] Put simply, truth as an intention is filled by an intuitive action. Okrent maintains that for Heidegger, "the fundamental notion of evidence [is] tied to how purposeful practical activity [is] recognizable as successful or unsuccessful if the activity is to count as purposeful at all."[31] The "communally purposive situation of language use" determines the conditions for truth and *understanding.*[32]

True assertions and propositional knowledge depend on practical activity to achieve a practical goal. Consequentially, Dasein, Being-in-the-world, Being-with, and Being-in are complex meta-grammatical structures shown or evidenced in the complicated interrelationships of practice, worldly engagement, and comportment. From the analysis of these structures, the meta-grammatical logic of propositions not only plays the role of inference or theoretical deduction but, as Donald Davidson emphasizes, is also essential and indispensable in characterizing the "meaning" of objects and their involvement in intersubjective practices.[33] In the verificationist account, propositional truth relies on something perceivable, and the fulfilment of

---

[28] Gethmann, "Zu Heideggers Wahrheitsbegriff," 198. Translation mine

[29] Edmund Husserl, "The Ideal of Adequation. Self-Evidence and Truth," in *Logical Investigations Vol. II*, trans. Dermot Moran (New York: Routledge, 2001), 259-267.

[30] Richard Rorty endorses Mark Okrent's view. See, Rorty, *Essays on Heidegger and Others*, 32f.

[31] Okrent, *Heidegger's Pragmatism*, 128.

[32] Okrent, *Heidegger's Pragmatism*, 128.

[33] Livingston, *The Logic of Being,* 60f. For this reason, recent literature makes a comparative effort to show the similarities between Heidegger and Donald Davidson. Part of the standard interpretation of the conceptual relationship between these two thinkers involves the similarities between the non-propositional Heideggerian *understanding* and the 'primitive triangulation' advanced by Davidson. Davidson's primitive interpretation involves purposive activity governed by social normativity; this is said to be analogous to the social normativity purported in Heidegger's *understanding*. Both thinkers are said to maintain a notion of non-linguistic understanding that is a fundamental and pre-conceptual form of meaning shaped by social interactions.

intuition is only possible if the referent is on hand.[34] This reading, however, as I argue in section three, flattens the disclosive facticity (*Faktizität*) of existence to a social matter-of-factness (*Tatsächlichkeit*), an occurrence within the static and social world (SZ: 55f).[35] The verificationists foreclose the pursuit of meaning to anything other than the success of a socially predicated action. I elaborate on this claim later.

Understanding and AI

From a verificationist reading of Heidegger, Dreyfus advances three central arguments to differentiate human intelligence from artificial intelligence.[36] First, human beings respond to relevant features in their environment without relying on a mental representation of facts.[37] Second, skilled action is not a psychologically mediated causal chain of input-to-output responses.[38] Third, human intelligence consists of direct and self-forgetful responsiveness through embodied capacities.[39] Correspondingly, Dreyfus argues that AI research neglects two interrelated problems. First, AI cannot organize the 'worldly situation' so that objects are accessible and relevant outside of a predetermined set of facts.[40] In turn, AI neglects the 'worldly situation' in providing a background for embodied coping.[41] Second, AI cannot account for the non-psychological way in which human intelligence experiences the world.[42]

---

[34] Heidegger renounces this, arguing that by prioritizing objects and properties of objects, the Neo-Kantians and Marburg school mistreat the relation to how objects are "originally given." Heidegger stresses that the inquiry into "sensible entities" does not characterize Being (*Sein*) but only determines the way of apprehending being (GA59: 53). Heidegger identifies the tendency to view everything as either itself an object or a property of an object. By focusing on ontic issues and overlooking the ontological issue, philosophy inherits a conception of being as "to be" "occurrent" (*vorhanden*).

[35] See, Dahlstrom, *Heidegger's Concept of Truth*, 227. The temporal consequences are beyond the scope of the present study.

[36] I will not provide an exhaustive exegetical account of each of these claims. Instead, my focus will be on how his views culminate in what Dreyfus claims is the rationalist assumption.

[37] Hubert Dreyfus, Stuart E. Dreyfus, "What artificial Experts Can and Cannot Do," *AI & society* 6 (1992): 18.

[38] Dreyfus, *What Computers Still Can't Do*, 163-188.

[39] Dreyfus also advances that embodied coping has motor-intentional content and that it makes the intentional arc possible. For a more detailed description, see, once again, Hubert Dreyfus, "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian," *Philosophical psychology* 20, no. 2 (2007): 247-268.

[40] Dreyfus, *What Computers Still Can't Do,* 246f.

[41] The ontological assumption. Dreyfus, *What computers Still Can't Do*, 287f.

[42] Dreyfus notes that when learning to drive, dance, or pronounce a foreign language, we must slowly, awkwardly, and consciously follow the rules. But then there comes a moment when we can finally perform automatically. At this point, we do not seem to be simply dropping these same rigid rules into unconsciousness; rather, we seem to have picked up the muscular *gestalt*, which gives our behavior new flexibility and smoothness. The same holds for acquiring the skill of perception. *What*

The mainstay of Dreyfus' argument is that AI research programs falsify their enterprise by basing intelligence on a 'rationalist' assumption. Dreyfus claims:

> A machine can, at best, make a specific set of hypotheses and then find out if they have been confirmed or refuted by the data. [Human beings] constantly modify [our] expectations in terms of a more flexible criterion: as embodied, we need not check for specific characteristics or a specific range of characteristics, but simply for whether, on the basis of our expectations, we are coping with the object. Coping need not be defined by any specific set of traits but rather by an ongoing mastery . . . [a] maximum grasp. What counts as maximum grasp varies with the goal and the resources of the situation. Thus, it cannot be expressed in situation-free, purpose free terms.[43]

AI and the human mind are understood by AI researchers as physical symbol systems using streams of neuron pulses as symbols representing the external world. Consequentially, human intelligence is considered rational (psychological/mentalistic) and factually deduced. The rationalist assumption reinforces the idea that in an orderly domain, there are sets of context-free elements and abstract relations among those elements, that underlie human intelligence.[44] The assumption, therefore, is that knowledge consists in forming and using appropriate symbolic representations.[45] The human mind, however, does not function exclusively on the psychological capacity to form representations, theories, or propositions about states of affairs. Objects are only understood de-contextually when we stop acting skillfully and approach the world conceptually. Therefore, AI cannot account for the dynamic, context-bound engagement with the world. To illustrate this point, Dreyfus argues that humans recognize patterns even when they are incomplete or distorted. Unlike AI, humans simultaneously acknowledge that a pattern is present while perceiving a discontinuity in the expected pattern. Human pattern recognition, so Dreyfus claims, is influenced by contextual information or background knowledge that fills in missing elements to make inferences. AI pattern recognition operates within strict adherence to predetermined algorithms or models; therefore, it lacks the adaptability to accommodate incompleteness or distortion. Additionally, AI pattern recognition necessitates testing and subsequent exclusion when confronted with background noise while humans effortlessly disregard irrelevant details in states of affairs. In short, AI successfully performs in a completely defined system like chess, where a finite number

---

Computer's Still Can't Do, 249. Also, see, "The Biological Assumption," in *What Computers Still Can't Do*, 159-162.

[43] Hubert Dreyfus, "Why Computers Must Have Bodies in Order to Be Intelligent," *The Review of Metaphysics* 21, no. 1 (1967): 20-1.

[44] Hubert Dreyfus, and Stuart E. Dreyfus, "Making a Mind versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint," *Daedalus* 117, no. 1 (1988): 25.

[45] Dreyfus, *What Computers Still Can't Do*, XI.

of concepts determines totally and unequivocally the set of all combinations in the domain.[46]

Non-psychological know-how grounds our everyday ability to navigate the world and engage with objects. Know-how is non-axiomatic, and even experts have difficulty identifying what they are doing when performing a task at an elite level. For Dreyfus, the brain processes information "from trial-and-error . . . triggered by involvement in real situations . . . [and] cannot be described at any domain-theory level of abstraction."[47] Experts, or professional athletes, so Dreyfus claims, will not deliberate with "detached problem solving, even when time permits."[48] Experts are more likely to "deliberate about the relevance of their prior experience . . . or overlooked alternative perspectives" rather than "the rules and principles underlying their skill" in general.[49] In doing so, experts "embody a richly articulated way of dealing with objects in the world without the use of predicate language."[50] For example, playing basketball or riding a bicycle encompasses proficiency and aptitude that skilful copers cannot easily formalize in propositions. Linguistic utterances express the successful performance of a task, but they do not disclose the underlying cognitive processes or mental mechanisms involved during its execution. The competence of an elite basketball player has a form of knowledge that is distinct from, and perhaps irreducible to, formalized propositional knowledge. Dreyfus deems this species of know-how as "tacit knowledge."[51] As Timothy Nulty puts it, tactic knowledge is "non-mentalistic; it is a primitive or basic form of intentionality that grounds the possibility of linguistic meaning."[52] AI systems primarily operate based on explicit rules in the form of programmable language; they lack the ability to effectively utilize tacit knowledge. AI programmers cannot replicate the non-mentalistic way humans act.

In short, I outlined two aspects of the rationalist assumption. First, AI programmers assume that intelligence cognizes a determinate set of data to make inferences. AI cannot account for the "real world," where the list of relevant facts, or even classes of possibly relevant facts are indefinitely large.[53] Second, AI programmers make the assumption that all non-arbitrary behavior is formalizable according to rules, and these rules can then be used by a computer to reproduce human behavior.

---

[46] Dreyfus, *What Computers Still Can't Do*, 177.

[47] Dreyfus and Dreyfus, "What Artificial Experts Can and Cannot Do," 22.

[48] Dreyfus and Dreyfus, "What Artificial Experts Can and Cannot Do," 22.

[49] Dreyfus and Dreyfus, "What Artificial Experts Can and Cannot Do," 22.

[50] Mark Wrathall, "The conditions of truth in Heidegger and Davidson." *The Monist* 82, no. 2 (1999): 304-323.

[51] Dreyfus, "Overcoming the Myth of the Mental," 52f. Also see, Jerry Fodor, "The Appeal to Tacit Knowledge in Psychological Explanation," *The Journal of Philosophy* 65, no. 20 (1968): 627-640.

[52] Timothy J. Nulty, "Davidsonian triangulation and Heideggerian comportment." *International journal of philosophical studies* 14, no. 3 (2006): 443-453. Also see, John Haugeland, *Artificial intelligence: The Very Idea.* (Cambridge: MIT press, 1989).

[53] Dreyfus, "Overcoming the Myth of the Mental," 65.

Critical Evaluation of the Verificationist Reading

While the verificationists point out one aspect of *understanding*, it is difficult to see how this reading is sustained without doing serious violence to Heidegger's project. In their critical oversight, Heidegger's phenomenological breakthrough towards a sense of self-understanding (*Seinsverständnis*) in a principled account of Being (*Sein*) is absent.[54] In other words, the verificationists remain on the level of everyday u*nderstanding* and disregard the existential implications of self-understanding.

In this section, I present a two-pronged critique of Dreyfus' account. First, I argue that Dreyfus' account is inadequate given recent developments in AI. AI has surpassed Dreyfus' expectations, rendering many of his examples outdated. However, the primary error lies in his verificationist or outcome-based criteria for knowledge. The underlying presupposition in Dreyfus' account is that knowledge relies on the *success* of our practical engagements. While some of Dreyfus' examples withstand the test of time, the developments of AI will surpass these exceptions because AI developers, like Dreyfus, rely on an outcome-based criterion as their measure of success. Second, I argue that Dreyfus' account is an incomplete reading of Heidegger's concept of *understanding*. By drawing on a complete and principled account of *understanding*, I attempt to circumvent the outcome-based criteria.

The Limits of Dreyfus' Argument for Recent AI Development

Dreyfus' central claim is that human intelligence relies on embodied and contextually sensitive know-how. For Dreyfus, AI systems cannot incorporate and understand subtle contextual elements in their environment.[55] Without the background knowledge accumulated through experience, AI systems have a limited capacity to comprehend and respond appropriately to dynamic situations.[56] However, consider DeepMind's AlphaGo. AlphaGo is an AI program developed to play the board game Go, which is known for its complexity and strategic depth.[57] In 2016, AlphaGo defeated the world champion Go player and "introduced innovative and valuable strategies to the Go community."[58] With the ability to master the complexity of Go, "AlphaGo fulfils the

---

[54] Dahlstrom, *Heidegger's Concept of Truth*, XIX.

[55] For Dreyfus, the environment is not exclusive to a physical environment. He extends the term to include domains of relevance.

[56] Dreyfus, "Overcoming the Myth of the Mental," 65.

[57] Go is far more complex than Chess. For example, in chess there are 20 possible moves. In Go, the first player has 361 possible moves.

[58] Marta Halina, "Insightful artificial intelligence," *Mind and Language* 36, no. 2 (2021): 316.

criteria for creativity . . . producing novel, and surprising valuable solutions to problems [in the game]."[59] AlphaGo succeeds using deep neural networks and Monte Carlo tree search algorithms. It uses deep neural networks to evaluate board positions and make strategic decisions, while the Monte Carlo tree search enables the program to explore possible moves and anticipate future outcomes. Using reinforcement learning techniques, AlphaGo improves its performance through self-play while learning from experience. Marta Halina notes that:

> The exploration parameter allows AlphaGo to go beyond its training, encouraging it to simulate moves outside of those recommended by the policy network. As the search tree is constructed, the system starts choosing moves with the highest "action value" to simulate, where the action value indicates how good a move is based on the outcome of rollouts and value-network evaluations.[60]

By constructing and employing a "world model" of its environment, AlphaGo learns new moves that exceed its programmed policy. By utilizing reinforcement learning techniques to master the complexity of Go, the program learns how to analyze the game's strategic dynamics to make optimally reactive and live decisions. As a result, AlphaGo performs at levels that rival or surpass human expertise. Importantly, the AI system is not a formalized knowledge system pre-programmed by expert players to replicate a set of moves from previous matches. On the contrary, it employs Reinforcement Learning (RL) to train itself.[61] Some of AlphaGo's moves are inexplicable to human Go-playing experts, and yet are effective in winning games.[62] These new and unpredictable moves display a species of goal-oriented intentionality to win matches similar to human GO players.[63]

The development of Reinforcement Learning (RL) goes beyond the limitations that Dreyfus imposes on AI.[64] RL challenges Dreyfus' claim that the distinctive feature

---

[59] Halina, "Insightful Artificial Intelligence," 316.

[60] Halina, "Insightful Artificial Intelligence," 324.

[61] See, Guglielmo Papagni, Koeszegi Sabine, "A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents," *Minds and Machines* 31 (2021): 505-534.

[62] See, Peter Andras, Lukas Esterle, Michael Guckert, et. al, "Trusting Intelligent Machines: Deepening Trust within Socio-technical Systems," *IEEE Technology and Society Magazine* 37, no. 4 (2018): 76-83.

[63] Papagni, and Sabine, "A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents," 509

[64] Similarly, OpenAI's Dota 2-playing bot is designed to play the popular multiplayer online battle arena (MOBA) game Dota 2. In 2018, OpenAI's bot named "OpenAI Five" competed and won against several professional players. OpenAI's Dota 2 bot utilizes deep reinforcement learning techniques to master the complexities of the game and undergoes extensive training by self-play. It competes against different versions of itself to improve its gameplay strategies. The bot learned how

of human intelligence is contextual sensitivity and adaptive ability. RL does not need a predefined class of appropriate responses to generate knowledge that leads to successful gameplay. RL's machine learning discovers how to interact with its environment to maximize a cumulative reward signal. In other instances, Deep Q-networks (DQNs) combine RL with deep neural networks, specifically convolutional neural networks (CNNs), to effectively handle high-dimensional and complex state spaces. DQN is designed so that the agent and environment engage in ongoing interaction. The AI responds to its environment according to its current observation and 'policy.' In return, the agent receives a reward and the next environmental observation. By employing a deep neural network as a function approximator, DQNs learn a Q-value function which estimates the expected cumulative reward for taking a particular action from a given state.[65] In other words, this learning algorithm aims to optimize the cumulative reward or the return. By doing so, DQNs effectively learn a complex mapping from states to actions and make optimal decisions in complex environments.[66] Dreyfus' condition for successful coping is a responsiveness to the solicitations of the environment and the approximation of an "optimal *gestalt* for a fluid response to the situation."[67] The "mind" of these AI systems does not operate on bits of information according to formalized information; rather, the AI systems have practical knowledge about their worlds by considering complex attitudes and tendencies to favour one action over another. In this sense, AI meets Dreyfus' condition for skillful coping.

To anticipate a critical rejoinder, I concede that AI systems have a limited capacity. For example, AI lacks personalization (i.e., having an identity), and sufficient emotional intelligence. In language-based models, AI typically reproduces generic responses that culminate general information. Perhaps the most prevalent limitation of AI lies in the challenge of robotics and dexterity in physical interactions. For Dreyfus, sports are paradigmatic instances of human intelligence. Athletic know-how demonstrates fine-grained motor skills, delicate manipulation of tools, and non-cognitive yet reactive adaptability. Put simply, athletic ability presents difficulties for

---

to analyze the game's dynamics, strategize, and make optimal decisions in real-time. Both AlphaGo and OpenAI's Dota 2-playing bot demonstrate the significant advancements made in AI and machine learning. These achievements highlight that AI systems can accomplish complex challenges, learn from data, and perform at levels that rival or surpass human expertise in specific domains. Further study in required to determine whether game theory threatens Dreyfus' claims about expertise. See, "Five Steps from Novice to Expert" in *Mind over Machine*, 16-51. For Dreyfus' discussion on Reinforcement learning, See, Dreyfus, introduction to *What Computers Still Can't Do*, IX-LII.

[65] For an elaborated treatment of Deep Q-networks, See, Patrick Hohenecker, and Thomas Lukasiewicz, "Ontology Reasoning with Deep Neural Networks," *Journal of Artificial Intelligence Research* 68 (2020): 503-540.

[66] For technical data analysis, See, Xu Chen, and Jun Wang, "Inhomogeneous Deep Q-network for Time Sensitive Applications," *Artificial Intelligence* 312 (2022): 1.

[67] Hubert Dreyfus, *Skillful Coping*, 11.

current AI-powered robotic systems.[68] Problematically, however, Dreyfus extends embodied coping beyond athletics to other refined skills. For example, chess, jazz improvisation, cooking dinner, crossing a busy street, carrying on a conversation, or just getting around in the world.[69] AI-powered robotic systems have limitations in the fluidity of completing some, but not all, of these refined skills.

AI programmers seek to develop AI systems that perform tasks typical of human intelligence. AI machines or software aim to think, reason, learn, perceive, and interact with the world like human beings. Even in the case of AGI, the goal is to create machines that understand, learn, and apply knowledge across multiple, if not all, domains. I argue that AI and AGI enterprises rely on goal-oriented intentionality, evaluating the success of their performance through outcome-driven and efficiency-driven initiatives. Problematically, Dreyfus' account of know-how also measures human intelligence on a success model of performative action.[70] For this reason, Dreyfus' account is vulnerable to future AI systems that rival or surpass human action or performance. Dreyfus falls victim to Heidegger's warning in the opening paragraph of *Being and Time*. Heidegger states that to ask the correct question is to find the correct path to its achievement (SZ: 1). One must "reawaken an understanding for the meaning of [the] question" because "what is asked about there lies also *that which is to be found out by the asking* [*das Erfragte*]" (SZ: 2). For Dreyfus, this question is what computers cannot *do*. Dreyfus then measures the success of human intelligence against the performative-*doing* of AI and becomes vulnerable to the development of AI's performance. In the following subsection, I reframe the aim of our inquiry by asking a new question: Can AI take a meaningful relation to action? I also present a complete and principled account of Heidegger's concept of *Understanding*.

Critique of Dreyfus' Flattened Ontology

It is tempting to read Heidegger's concept of *understanding* as practical know-how. Human existence *necessarily* directs our attention to a world of concern, and we cannot *be* in the world without practice. Heidegger does not suggest, however, that our access

---

[68] Problematically, Dreyfus lumps all games into one category, whether they are physical or otherwise. Part of my concession is that interactive and autonomous robots are only in the beginning stages of development (i.e., currently, AI cannot play tennis). For a full treatment of embodied coping, see Hubert Dreyfus, "The Primacy of Phenomenology Over Logical Analysis," *Philosophical Topics* 27, no. 2 (1999): 3-24.

[69] Dreyfus, "Overcoming the Myth of the Mental," 58.

[70] Regardless of the various reasons and nuances that justify the verificationist reading, I argue that the conclusion is the same. In other words, the definitive feature of skillfully absorbed, pragmatically sensitive, culturally nuanced, and non-regulative, embedded human knowledge is based on a goal-oriented success model.

to practices determines the disclosure of the world or ourselves.[71] *Understanding* in the primordial sense, as self-understanding, does not signify a practice.[72] In this subsection, I elaborate on this claim.

Following Daniel Dahlstrom, I argue that Dreyfus and other verificationists misconstrue the pre-ontological, ontological, and ontic levels of Heidegger's thought, and the corresponding *existentiel* and existential dimensions of *understanding*.[73] In the primordial sense, *understanding* discloses a pre-ontological question concerning the need for self-understanding. Disclosure, in this sense, solicits an ontological inquiry: My existence deserves investigation with ontic-ontological priority over other entities (SZ: 142f, 259f).[74] In doing so, I investigate ontological meaning alongside the complexity of instruments I concern myself with (SZ: 85f, 143). As a result, the disclosive feature of self-understanding does not satisfy its criteria by making an ontic or practical difference. Heidegger states:

> Dasein's ways of behaviour, its capacities, powers, possibilities, and vicissitudes, have been studied with varying extent in philosophical psychology, in anthropology . . . each in a different fashion. But the question remains whether these interpretations of Dasein have been carried through with a primordial existentiality comparable to whatever existentiell primordiality they may have possessed. Neither of these excludes the other but they do not necessarily go together. Existentiell interpretation can demand an existential analytic, if indeed we conceive of philosophical cognition as something possible and necessary. Only when the basic structures of Dasein have been adequately worked out with explicit orientation towards the problem of Being itself, will what we have hitherto gained in interpreting Dasein get its existential justification. *Thus, an analytic of Dasein must remain our first requirement in the question of Being. But in that case the problem of obtaining and securing the kind of access which will lead to Dasein, becomes even more a burning one . . . Once we have arrived at that horizon, this preparatory analytic of Dasein* [in Division I] *will have to be repeated on a higher and authentically ontological basis* (SZ: 16f, emphasis added).

Understanding, recognized by successful action, amounts to the knowledge proffered by the natural sciences insofar as they both presuppose an understanding of existence

---

[71] The meaning and validity of Disclosure (*Erschlossenheit*) is, in part, what motivates Tugendhat's critique.

[72] Daniel Dahlstrom notes that "existential understanding constitutes various forms of "sight" (*Sicht*). The circumspection (*Umsicht*) of our work-world concerns, the considerateness (*Rücksicht*) of our solicitude for one another, and the transparency (*Durchsichtigkeit*) of Dasein's full disclosure of itself as being-in-the-world, along with its opaqueness to itself (*Undurchsichtigkeit*) are familiar, figurative transcriptions of understanding." *The Heidegger Dictionary*, 231.

[73] Dahlstrom, *Heidegger's Concept of Truth*, 428.

[74] Dahlstrom, *The Heidegger Dictionary*, 232.

(GA24: 389f; SZ: 143, 336; GA20: 413). Heidegger suggests that *understanding* in the primordial existential sense is not one type of knowledge contrasted with another (i.e., the humanities in contrast to the natural sciences).[75] The self-disclosive truth of existence (*Eigentlichkeit*) or the higher ontological basis derived from self-understanding cannot be adequately mapped onto the structure of a practice or a set of practices. That which leads to existential questioning, namely, the call of conscience, is not a material ethic. Heidegger states that:

> We miss a 'positive' content in that which is called [by our conscience], *because we expect to be told something currently useful about assured possibilities of 'taking action' which are available and calculable.* This expectation has its basis within the horizon of that way of interpreting which belongs to common-sense concern, a way of interpreting which forces Dasein's existence to be subsumed under the idea of a business procedure that can be regulated. Such expectations (and in part these tacitly underlie even the demand for a *material* ethic of value as contrasted with one that is 'merely' formal) are of course disappointed by the conscience. The call of conscience fails to give any such 'practical' injunctions, *solely because* it summons Dasein to existence, to its ownmost potentiality-for-Being-its-Self (SZ: 294).

Understanding secures the intelligibility (*Verständigkeit*) of entities, while existential self-understanding leads Dasein to the ontological intelligibility of itself (i.e., the self-disclosure of being-in-the-world) (SZ: 13, 85f, 143). Self-disclosure is the condition for the possibility of both forms of *understanding*. By collapsing the a-priori generality of Dasein (existential conditions for understanding, *Seinsverständnis*) into what is practically available, Dreyfus and the verificationists fail to distinguish the ontological difference between human beings and other objects or entities (i.e., ontological from the ontic). For Heidegger, "what understanding as an existential can understand is not a what, but rather being as existing" (SZ: 143). As Dahlstrom notes, distinguishing between the inquiry of ontology and the inquiry of ontic sciences allows us to see the ontological difference between the two.[76]

The verificationists fail to unify the structure of meaning with the basic existential orientation of *Seinsverständnis* and *Eigentlichkeit*. Dreyfus attempts to justify this oversight suggesting that Division I of *Being and Time* is "the most original and important section," and despite the presentation of "more originary [*sic*] temporality" in Division II, it "leads [Heidegger] so far from the phenomenon of everyday temporality" that "satisfactory interpretation of the material cannot be given."[77] Heidegger states, however, that:

---

[75] Dahlstrom, *The Heidegger Dictionary*, 231.

[76] Dahlstrom, *Heidegger's Concept of Truth*, 305f

[77] Dreyfus, *Being-in-the-world*, VIII.

> Dasein's Being must already be presupposed as a whole when we distinguish between theoretical and practical behaviour [and] cannot first be built up out of these faculties by a dialectic which, because it is existentially ungrounded, is necessarily quite baseless. Resoluteness, however, is only that authenticity [*Eigentlichkeit*] which, in care, is the object of care [*in der Sorge gesorgte*], which is possible as . . . the authenticity of care itself (SZ: 300).

It is precisely the problematic sense of the entity "I am," in the preparatory analytic of Dasein that grounds the ontological basis for a principled account of Being (*Sein*). *Understanding*, construed exclusively as the capacity to cope with the worldly environment presents one aspect of Heidegger's project at the expense of another.[78] More specifically, this reading neglects the pre-ontological and ontological claims of *Seinsverständnis* and *Eigentlichkeit* that lead to "coming to the self that is most one's own . . . [through] its individualization [*Vereinzelung*]" (SZ 339). The fulfilment of an authentic intuition gains its ontological purchase precisely from the discontinuity of everyday understanding (*Weltanschauung*), and theoretical objectification.

Heidegger uses the term πραξις (or "practice") in connection with the phenomenon of care, suggesting:

> Care, as a primordial structural totality, lies 'before' ["vor"] every factical 'attitude' and 'situation' of Dasein, and it does so existentially *a priori;* this means that it always lies *in* them. *So this phenomenon by no means expresses a priority of the 'practical' attitude over the theoretical.* When we ascertain something present-at-hand by merely beholding it, this activity has the character of care just as much as does a 'political action' or taking a rest and enjoying oneself. 'Theory' and 'practice' are possibilities of Being for an entity whose Being must be defined as "care." The phenomenon of care in its totality is essentially something that cannot be torn asunder; so any attempts to trace it back to special acts or drives like willing and wishing or urge and addiction, or to construct it out of these, will be unsuccessful (SZ: 193-4).

The existential *a priori* of *understanding* conditions the possibility of engaging with the environment and reflective analysis. *Understanding* allows me to perceive and interpret the world within the confines of lived experience, while self-understanding goes beyond my mere facticity. The self-referential dimension of *understanding* guides the meaning we assign to our actions. Through an extensive treatment of Aristotle's *Nicomachean Ethics,* Heidegger qualifies the self-referentiality of meaning, suggesting that *phronesis,* or practical understanding*,* depends on a prior disclosure that is higher in rank than itself (GA19, 167). The 'higher rank' is the ontological conception of

[78] Namely, the fundamental insight that governs the project of *Being and Time*, especially in Division II, is the question of individuated Being (*Sein*).

Being characterized by care. From a thorough examination of Heidegger's texts, *phronesis* is revealed to encompass a relationship with action that is both non-objectifying and mentalistic. A similar sentiment appears in *Being and Time* when Heidegger suggests that:

> 'Practical' behaviour is not 'atheoretical' in the sense of "sightlessness."[79] The way it differs from theoretical behaviour does not lie simply in the fact that in theoretical behaviour one observes, while in practical behaviour one *acts* [*gehandelt wird*] . . . for the fact that observation is a kind of concern is just as primordial as the fact that action has *its own* kind of sight. Theoretical behaviour is just looking, without circumspection. But the fact that this looking is non-circumspective does not mean that it follows no rules: it constructs a canon for itself in the form of *method* (SZ: 69).

Dreyfus creates the problematic opposition between theoretical knowledge and practical knowledge. In Dreyfus' account, *understanding* is conceived without intuitive contemplation or self-referentially; these conditions ground a principled account of meaning and Being. The verificationists accept that practical life is non-mentalistic everyday coping, however, it is precisely the everyday *Weltanschauung* in Dreyfus' account that Heidegger deems to be *fallenness* (*Verfallen* or *Verborgenheit*). Commentators often have difficulty accounting for the movement between *Uneigentlichkeit* and *Eigentlichkeit* because *understanding*, conditioned by ontic consequences, never effects the ontological structure of Being.[80] The change in *Weltanschauung* constitutes a "genuine movedness of life," in which life exists and through which life is determinable in its own sense of Being. This movement makes it intelligible how Being is genuinely brought into appropriate modes of possession (GA 61: 87).

Human existence is always given through disclosedness, making self-acquaintance a pre-theoretical process. Heidegger suggests, however, that reflection is necessary for becoming authentically individualized.

For this reason, reflection focuses on existence, indicating that a "who," in a pre-theoretical manner, necessarily raising questions about its Being and thereby provides the inescapable starting point for philosophical inquiry. Human beings possess a distinct intelligible quality that ontologically sets us apart from other entities.

---

[79] "Im Sinne der Sichtlosigkeit." The point of this sentence will be clear to the reader who recalls that the Greek verb from which the words 'theoretical' and 'atheoretical' are derived, originally meant 'to see.' Heidegger is pointing out that this is not what we have in mind in the traditional contrast between the 'theoretical' and the 'practical.'

[80] A movement which constitutes a *genuine movedness of life, in* which and *through* which life *exists,* and from which, accordingly, life is determinable in its own sense of Being. This movement makes it intelligible how a being such as life is to be brought genuinely into one of its available, appropriating modes of possession (Problem of facticity). Thereby we will acquire for the categorial interpretation the exposition of the basic sense from which all existentialia interpretively take their own proper sense as well as their referential sense (GA61: 87).

For Heidegger, the question itself is "the point where [existence] *arises* and to which it *returns*" (GA2: 51, 62). As a formally indicative concept, *understanding* points to "a concretion of individual existence" in the human being, but "it never" conveys that which is in its content already (GA29: 429). Thus, despite factical life experience being world-immersed, we tend to misinterpret ourselves in terms of our worldly being (i.e., historical, social, cultural, physical aspects, and other circumstances or limitations). Self-understanding that initially arises from the hermeneutic context is insufficient and inauthentic. Heidegger categorizes this unavoidable existential predicament as *Ruinanz* (GA 61: 119, 121). Since factical life experience covers what needs to be brought to light, articulating the fundamental structures of life will no longer rely on merely going along with life's tendencies. Moreover, the criteria for *understanding* cannot be characterized by the productive outcome of background coping practices. While it is given first, it is not the final level of analysis. Everyday understanding serves as the presupposition for the transition into authenticity.

The verificationist see reflection as theoretical and therefore objectifying. However, the transition from the *Weltanschauung* to a genuine beholding of life requires reflection. This species of reflection is not a reified ego bent backward staring at itself *ala* Husserl, but a *reflexive* practice whereby I question the entity I am in conjunction with the world I inhabit.

Reflection, in the genuine sense of intuitive contemplation, leads to retrieving the meaningful relationship (*Bezug*) I have toward action. For Dreyfus, there is no room for 'mindedness' in his account of practical knowledge, thus it remains sightless and existentially ungrounded. *Understanding*, then, properly understood, is enacting an experience with non-objectifying self-referentially, and interpreting the sense or meaning of it accordingly (GA 58: 262-263; GA61: 55, 60). The principled result is an understanding of myself in relation to the actions I *necessarily* take as an actor in a social and dynamic world.

Concluding Remarks: The Future of AI

The verificationists argue that Heidegger's concept of understanding grounds a critique of traditional ontology and epistemology. However, this reading fails to recognize the ontological significance of bringing the problematic sense of the authentic "I am"—the being of life—into its genuine actualization. In this sense, actualization involves the concrete question of the restlessness of factical life. Self-understanding opens up factical life as indefinite, questionable, and labile, yet always remaining participatory in disclosive factical objectivity. All my worldly experiences involve self-acquaintance and familiarity, and thus "I am always somehow acquainted with myself" (GA 58: 251). However, "at first, Dasein is completely lost (immersed) in the world, and only in a subsequent move does it turn towards itself and thereby

acquire self-acquaintance."[81] The question and confrontation of self-acquaintance are necessary for the fulfillment of the principled conception of Being and for understanding what it means for humans to possess intelligence.[82]

Suppose AI is indistinguishable from human intelligence. In that case, I suggest that AI programmers must incorporate the problem of meaning into AI systems, discerning the relation that these systems take towards their actions. In other words, AI systems must comprehend 'having' (*Haben*) meaning authentically or inauthentically. AI systems must also recognize that their immediate lived experience lacks an intelligible and existential understanding (i.e., *Verfallen* or *Verborgenheit*). Beginning with "inauthentic having," AI needs the capacity for reflection that "leads the way" (*methodos*) into "authentic evidence" where an encounter with an individuated and genuine "having of life itself is possible" (GA 61: 35).[83] In my view, it is not enough for AI to outperform human actions with goal-oriented intentionality. Instead, AI must acknowledge the meaningful relationship it takes toward its performance. AI must have a basic understanding of everyday life and grasp the nexus of meaning that is brought into relief by an authentic beholding. Actions must be done so that the *relation* towards the actions is changed without making an ontic difference. The difficulty of doing this, as Ernst Tughenhat suggests, is the lack of public verifiability, or public criteria for success. According to Dreyfus, Heidegger's account of *understanding* has an indiscernible quality characterized by the inability to know what one is doing when performing a task skillfully, making it non-mentalistic. I reframe this indiscernible quality as a self-reflective authenticity, wherein a meaningful relation is established with actions that cannot be extrinsically verified.

AI programmers will have a difficult time identifying whether the AI takes a meaningful relation towards an action. Authentic self-relation is inherently individualized, thus, cannot be put to the test. The *relation* towards action does not improve efficiency or expertise in any domain. AI programmers are attempting to enhance decision making transparency by tracking processes and identifying what factors are considered in AI performance. However, meaningful or *phronetic* action is not measured by justified reasoning. Authentic "having" is closer to a species of intuitive self-understanding that needs no justification, nor has one. Authentic "having" is one necessary feature of human intelligence that avoids competing with the exponential growth of AI's outcome-based achievements. The success of AI (and AGI) is measured based on the results of their programming. This species of pragmatism is hopelessly ontic. It attempts to reveal and provide a service for things

---

[81] Manfred Frank, "Fragmente einer Geschichte der Selbstbewußtseins-Theorie von Kant bis Sartre," in *Selbstbewußtseinstheorien von Fichte bis Sartre*, ed. Manfred Frank. (Frankfurt a. Main: Suhrkamp, 1991), 518. Translation mine.

[82] See note 9.

[83] Steven G. Crowell, *Husserl, Heidegger, and the Space of Meaning: Paths toward Transcendental Phenomenology* (Illinois: Northwestern University Press, 2001), 126.

(*pragmata*) on hand, without concern for the structure of experience. AI programmers are incentivized by technocratic control and dominance, leaving no place for the "passive" call of conscience or self-understanding regarding the ontological notion self-actualization. For this reason, self-understanding circumvents the verificationist account and AI's outcome-based criteria of intelligence.

This paper aims to present a principled account of Heidegger's concept of *understanding*. Additionally, it includes a critique of the verificationist reading. I argue that Dreyfus and others fail to grasp the fundamental insight of Heidegger's thought. Commentators writing on AI and Heidegger often replicate this limitation. By prioritizing practical know-how over any form of mentalism, I contend that the verificationist approach restricts Heidegger's ontology, leading to an inadequate analysis of human intelligence. Our existence remains entangled in environmental structures, thus we skillfully, adaptively, practically, and non-prescriptively engage with the world in an inauthentic manner. However, through contemplation of our existence and *Weltanschauung*, we gain the possibility of transitioning from everyday engagement to an authentic self-relation. By doing so, we surpass the mere ontic dimensions of life's involvements. Those seeking to use Heidegger to illustrate the limitations of AI should recognize that both divisions of *Being and Time* are crucial to their argument.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Jacques Ellul, AI, and the Autonomy of Technique

B.W.D. Heystee
Memorial University
ORCID: 0009-0007-2322-9616

Abstract

Contemporary concerns about the development of Artificial Intelligence (AI) frequently discuss the prospect of AI becoming rogue or out-of-control. Such concerns are raised by advocacy groups like the Centre for AI Safety and academics such as Nick Bostrom. In this paper, I consider those concerns in light of Jacques Ellul's account of technique. On the basis of Ellul's account, I argue that the prospect of machines getting out-of-control is not a future potentiality, but a present reality. I do this by outlining the various characteristics of technique according to Ellul, and then discussing the ways in which Bostrom et al. have misunderstood the danger of out-of-control technology.

Keywords: artificial intelligence (AI), rogue AI, Jacques Ellul, The Technological Society, George Grant

Introduction

In this paper, I consider contemporary concerns about out-of-control Artificial Intelligence (AI) in light of Jacques Ellul's account of technique in his 1954 book *The Technological Society* (*La Technique ou l'Enjeu du siècle*).[1] My basic claim is that concerns about out-of-control AI overlook an important consideration, namely that technological development may already be out-of-control. By "out-of-control," I mean "acting or behaving in a way that is counter to human interests and/or purposes, with no obvious possibility of being (re-)subordinated to those interests and purposes." While depictions of AI in popular culture and certain academic discussions of AI and superintelligence express concern about the (as yet unrealized) potential for technology to become out-of-control, such a potentiality is already a reality. As Ellul argues, we already live in a technological society that is organized with the end of maximum efficiency and is not, in fact, organized in pursuit of human ends, whatever those ends may be. Given that the technological society is not subordinated to human ends, we may reasonably call it out-of-control. Thus worries about an out-of-control AI do not see that their basic concerns for the future are already realized in the present.

In order to show that the technological society is already out-of-control, I divide my paper into four sections. In the first section, I provide a brief overview of contemporary concerns about AI. I take thinkers like Nick Bostrom and institutions like the Center for AI Safety (CAIS) as offering representative warnings about the dangers associated with AI. They argue that the potential construction of an artificial general intelligence (AGI) and, in particular, an artificial superintelligence poses an existential risk for humanity. In the second and third sections, I turn to Ellul and his account of technique. As I explain, by 'technique' Ellul means any "operation carried out in accordance with a certain method in order to attain a particular end."[2] (Technique is therefore not to be confused with tools or machines, which are only one aspect of technique.) In the second section, I discuss Ellul's account of the traditional technique so that we can better distinguish what is special about the modern technological society. In the third section, I outline the defining characteristics of the modern technique as Ellul understands them. They are: (a) automatism, (b) self-augmentation, (c) monism (*unicité*), (d) the necessary linking together of techniques, (e) universalism and (f) autonomy. In the fourth section, I return to contemporary concerns about AI and show that those concerns are not a future potentiality but a present reality. As Ellul's account of technique shows, the prospect of social control being wrested from humanity by its technological creations is already upon us because the chief determining factor of society is no longer human interests or purposes, but an autonomous and self-justifying technique.

---

[1] Jacques Ellul, *The Technological Society*, trans. John Wilkinson, Revised American (New York: Alfred A. Knopf, 1964).
[2] Ellul, *The Technological Society,* 19.

Contemporary Concerns about AI

Concerns about the potentially catastrophic implications of the development of AI have rocketed into the public consciousness following the release and popularization of large language models such as ChatGPT in 2022. Despite these recent alarms, however, criticisms of and warnings about AI are nothing new. As early as 1949, Norbert Wiener cautioned that the development of machines which could learn from experience could produce machines that were increasingly independent and potentially defiant of human interests and purposes. He warned that once those machines were capable of defiance, it would be hard to remedy the situation since "the genii in the bottle will not willingly go back in the bottle, nor have we any reason to expect them to be well disposed to us."[3] Indeed, the prevalence of AI-related concerns from early on in the period of digital computing is demonstrated by the fact of such films as *2001: A Space Odyssey* (1968). What would happen if we developed a computer like HAL-9000 and gave it so much power and practical responsibility that it could kill us, if killing us were necessary to achieve its programmed objectives? With the development of machines performing functions once thought the exclusive privilege of humanity, there emerged various questions and anxieties about delegating or ceding too much control to those machines.

More recently, criticism of AI has become a burgeoning field in academia and public interest advocacy. Perhaps the most well-known academic critic of AI today is Nick Bostrom. In 2014, Bostrom published a book entitled *Superintelligence: Paths, Dangers, Strategies* in which he argues that the eventual creation of an artificial superintelligence poses significant risks to humanity and that we should adopt certain strategies now in order to mitigate those risks.[4] Bostrom argues that humans have held an advantage over animals because of our greater capacity for general intelligence, but that if we should someday build machines with even greater general intelligence, those machines would have an advantage over us which would put us at their mercy and hence in great danger. The danger stems from the fact that the machines would be much more capable than we are and yet might also be unfriendly to us.[5] Bostrom argues that there is a real possibility of a superintelligent AGI for two reasons. First, the fact that evolution has produced a general intelligence at least once (humans) means that it is in principle possible for it to happen a second time, and the handiwork of an intelligent human programmer would likely make the process only more efficient.[6] What is more, computers already surpass humans in several respects: they

---

[3] Norbert Wiener, "The Machine Age" (1949), 8, Norbert Wiener Papers MC 22, MIT Institute Archives and Special Collections.

[4] Bostrom defines superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 22.

[5] Bostrom, *Superintelligence*, vii.

[6] Bostrom, *Superintelligence*, 23.

can perform calculations more rapidly, they can communicate more rapidly, they can more easily store information, and they are more easily adaptable to hardware additions and modifications (e.g., attaching improved sensors).[7] The fact of these extant advantages combined with the potential for AGI is that there is real potential for a machine to exist that surpasses humans in virtually every way, but especially in terms of intelligence.

Although Bostrom makes no claims that the development of a superintelligence is in any way imminent, he nevertheless insists that it is prudent for us at this early stage to take steps to mitigate the risks associated with such an eventual development. The mitigation of these risks is important because a superintelligence would in principle be capable of outwitting, outmaneuvering and outdoing us at every turn. If its ends (either self-consciously self-specified or unwittingly assigned by its human programmers) are counter to human ends, the result would be catastrophic, perhaps including the extinction of the human race.[8] Even a sufficiently advanced AI (but not truly general) directed toward some arbitrary end (e.g., paperclip maximization) could prove disastrous, as the AI might convert the entire planet into an automated paperclip factory, even at the expense of human life.[9] Given the risks of a malicious AI or an obedient AI but one with poorly-specified ends, it is incumbent upon us to develop strategies now to program AI very carefully. We need to ensure that any future AGIs or superintelligences are programmed according to human values and in such a way that it pursues these values or its specified ends in a manner that we like.

More recently, a number of public interest advocacy groups have released warnings of their own about the risks associated with developing AI. Organizations like the Center for AI Safety, PauseAI, and the Center for Human-Compatible AI have all released various reports, articles and public statements warning about the risks associated with AI and strategies we might use to mitigate them. Though these organizations highlight a variety of risks associated with AI (e.g., the malicious use of an AI by a human bad actor), they all highlight the risks associated with rogue AIs in particular. In identifying the risk of AIs becoming rogue or out-of-control, these organizations highlight many of the same concerns that Bostrom does in his book. The risks associated with a rogue AI include the pursuit of flawed objectives to an extreme degree (e.g., paperclip maximization), goal drift (i.e., the AI's prior specified ends changing as a result of a changing environment), or power-seeking (i.e., an AI seeking power as a means to pursuing its prior specified goal unhindered).[10] As the authors of one report note, such risks are especially acute because the rapid pace of development of relatively rudimentary AIs has revealed just how difficult it is to control them when they are given even a modest level of autonomy; even when a

---

[7] Bostrom, *Superintelligence*, 59–60.
[8] Bostrom, *Superintelligence*, 116.
[9] Bostrom, *Superintelligence*, 123.
[10] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks" (Center for AI Safety, June 26, 2023), 2.

programmer attempts to carefully specify an AI's ends, they are often met with undesirable surprises.[11]

Though there is a wide range of concerns associated with the development of AI and of AGI in particular, of acute concern is the idea that an AI could eventually go rogue and get out-of-control. If an AI were out-of-control, it is hard to know precisely what it would do but one can easily imagine the risks. Especially in the case of a superintelligence, there is little telling what it might take for an end, given that its hypothetical intelligence vastly exceeds that of humans. Given the diversity of possible ends available to a superintelligence, it is a statistical certainty that, if left to chance, it would choose something we would not like. Further, given how different a superintelligence would be from humans (e.g., presumably it would not have an organic body), it also seems likely that it would pursue its ends in a way we do not like. There would be little we could do about this, because the superintelligent AI would be especially capable of pursuing its ends, if not through mechanical means (e.g., physical control of infrastructure) then through interpersonal means (e.g., deceiving or convincing humans). There would seem to be a real risk that as an yet undeveloped out-of-control AI could have grave consequences for humanity as it pursues inhumane ends in an inhumane manner. Yet as we shall when we turn to Ellul's account of the modern technological society, the unstoppable pursuit of inhumane ends in an inhumane manner is already a present reality.

Traditional Technique

Before turning to Ellul's discussion of modern technique, let us first discuss traditional technique. By placing modern technique in relief to traditional technique, we will better see what is special about the modern situation and therefore the way in which technique has gotten out-of-control. In its most general definition, a 'technique' is an "operation carried out in accordance with a certain method in order to attain a particular end."[12] This definition of a technique is comprehensive of everything primitive and simple, modern and complex. Whenever there is a consistent method for producing a result—using a flint to produce a spark—there is a technique. This is to be contrasted with "natural and spontaneous effort," which is not so consistent and regular.[13] Fundamentally, this has not changed between antiquity and modernity.

In the pre-modern era, however, techniques were "applied in certain narrow, limited areas."[14] Although techniques were obviously used, much of life was governed

[11] The authors cite examples of a 2016 Twitter bot programmed with "conversational understanding" and Microsoft's Bing Chatbot in 2023. The former rapidly adopted hateful language after being released on Twitter, and the latter has been given to making threats and intimidation. Hendrycks, Mazeika, and Woodside, "Catastrophic AI Risks," 34.
[12] Ellul, *The Technological Society*, 19.
[13] Ellul, *The Technological Society,* 20.
[14] Ellul, *The Technological Society,* 64.

by "social spontaneities" or "private initiative, short-lived manifestations or ephemeral traditions, [rather] than on a pervading technical will and rational improvement."[15] In short, techniques were circumscribed by a society that was itself not technical and of which the most important aspects were not technical. Within those limited applications of technique, technical means were themselves limited: in a given society, "there was no great variety of means for attaining a desired result, and there was almost no attempt to perfect the means which did exist."[16] Humans used the means at their disposal and did not rigorously or systematically pursue improving those means. The limited tools which were applied in limited scenarios were themselves geographically limited, i.e., a given technique was local. Because social groups were, for the most part, strong and closed, techniques spread slowly and accidentally, if at all.[17] The limited techniques used were not rigorously and rationally developed in disregard for their social context, but were instead integrated into a given society, which itself was relatively stable.

The consequence of these characteristics of traditional technique was that techniques could almost always be adapted to human purposes. The limits in application, means, and geography meant that:

> technique[s] could be adapted to men. Almost unconsciously, men kept abreast of techniques and controlled their use and influence. This resulted not from an adaptation of men to techniques (as in modern times), but rather from the subordination of techniques to men. Technique did not pose the problem of adaptation because it was firmly enmeshed in the framework of life and culture.[18]

Whatever the particular features of certain techniques in a given community, those features were subordinate to broader human purposes. They were adapted to what was taken to be the good life. Techniques occupied an, at best, secondary role in human life and human communities. This is not to say that they were not important or significant, but that they were never the most important or significant thing. In one way or another, humans could meaningfully determine how and when they applied a technique or, even more fundamentally, what sort of life they wanted to lead. As we shall see, according to Ellul those choices are by and large unavailable in a modern technological society.

The genesis of modern technique is outside the main thrust of this paper, so I will only say a few words about it. Ellul explains the development of modern technique in historical, social, and objective terms. He says that modern technique arose because

---

[15] Ellul, *The Technological Society,* 65.
[16] Ellul, *The Technological Society,* 67.
[17] "Every technical phenomenon was isolated from similar movements elsewhere. There was no transmission, only fruitless gropings." Ellul, *The Technological Society,* 69.
[18] Ellul, *The Technological Society,* 72.

of the coincidence of five phenomena: "the fruition of a long technical experience; population expansion; the suitability of the economic environment; the plasticity of the social milieu; and the appearance of a clear technical intention." [19] These phenomena coincided in the end of the 18th century and the beginning of the 19th century. That is to say, the genesis of modern technique is coincidental. Five phenomena happened to coincide that made the technological society more likely. As far as Ellul is concerned, technique is not the final expression of a millennia long destiny as it is for Heidegger. It is the result of happenstance. But for Ellul this is not ultimately important. [20] It doesn't matter that modern technique is the result of happenstance. What matters is that is has come to be. Regardless of the 'why,' modern technique is a fact of our present civilization.

Before turning to Ellul's account of modern technique, it is worth briefly noting Ellul's rhetorical style. In his discussion of the technological society, Ellul often seems to hypostasize or to ascribe a certain agency to technique. He will argue that "technique does X," or that "technique requires Y," or that "technique allows Z," as if technique had its own separate existence and were an independent force shaping society. This approach has a certain merit, insofar as it vividly and succinctly illustrates to the reader what Ellul takes to be the basic principle organizing society and, as Lovekin notes, that those living in a technological society have a kind of "technological consciousness" which determines how the world appears to them.[21] But Ellul does not literally mean that technique has its own independent existence. Indeed it is central to my present criticism that technique has no separate existence to which we could point. Technique only exists as it is actually practiced by humans or carried out through the work of various machines. Neither does Ellul's rhetorical approach agree with the mode of discourse favored by Bostrom and the like, making an Ellulean criticism of contemporary AI critics difficult. For that reason, I have made modest efforts to "de-hypostasize" Ellul's account and not to write in a way that implies a separate existence to technique. For example, where Ellul speaks of technique itself doing something, I have tried to speak of people performing technical operations. This is not to suggest that I know better how to say what Ellul is trying to say, but to try to meet Bostrom and his peers on more familiar terms. Nevertheless, it is not always possible to maintain

---

[19] Ellul, *The Technological Society,* 47.

[20] As George Grant observes, Ellul's relative neglect of the genesis of modern technique is one of the main weaknesess in *The Technological Society*. This is especially problematic because, in Grant's view, understanding the technological society requires examining its close connection to and genesis in Western Christianity, and Ellul remains a committed Christian. Yet Grant gives Ellul the benefit of the doubt and suggests that Ellul's "lack of discussion at this point comes from a highly conscious and noble turning away from philosophy toward sociological realism." Ellul neglected the history of technique so that he could better see what it is in the present. George Parkin Grant, "Review of The Technological Society, by Jacques Ellul," in *Collected Works of George Grant*, vol. 2 (Toronto: University of Toronto Press, 2002), 417.

[21] David Lovekin, "Jacques Ellul and the Logic of Technology," Man and World 10, no. 3 (1977): 251.

this approach, especially as we come the conclusion of Ellul's description of technique, the autonomy of technique.

Modern Technique

Let us now turn to the characteristics of modern technique. Ellul defines modern technique as "the totality of methods rationally arrived at and having absolute efficiency (for a given stage of development) in every field of human activity."[22] Modern technique has a number of characteristics which belong to a single, integrated whole and cannot be entirely separated from each other. Again, these characteristics are (a) automatism, (b) self-augmentation, (c) monism, (d) the necessary linking together of techniques, (e) technical universalism, and (f) autonomy. (Ellul notes that modern technique is also rational and artificial, but declines to discuss these characteristics since they are sufficiently well-understood.)[23] The sum of technique's characteristics, we shall see, is that the technological society is not organized to pursue human ends in a humane way, but to pursue the distinctly technical end—efficiency—in a distinctly technical way—as efficiently as possible. Because it is not organized in pursuit of human ends but in pursuit of technical ends, the technological society can reasonably be described as out-of-control.

Automatism

Technique pursues the 'one best way' of doing things, since it is pursuing efficiency absolutely. This means that when a technical operation happens, the people involved measure and calculate matters mathematically, and on that basis determine what the best course of action is. The result of this calculation is that the best course of action is *obviously* the most efficient one. When the calculations are done, there is no personal decision to be made, any more than there is a personal decision in determining whether 4 is greater than 3. The technical decision is therefore 'automatic.'[24] If an activity is 'technical,' there is only one course of action available, namely that which is determined by mathematical calculations to be most efficient. To the extent that someone makes a meaningful choice, their work is not technical.[25] When the 'best' solution is evident, it is the only technical option. With regard to technical automatism, the human becomes little more than "a device for recording effects and results

[22] Ellul, *The Technological Society,* xxv.
[23] Ellul, *The Technological Society,* 78–79.
[24] Ellul, *The Technological Society,* 80.
[25] As we shall see, Ellul will go on to explain that although humans still do make some genuinely human choices, these are systematically excluded and hence have diminished impact on society and are becoming increasingly uncommon.

obtained by various techniques. He does not make a choice of complex and, in some way, human motives. He can decide only in favor of the technique that gives the maximum efficiency." [26] What is more, this decision is by and large met with satisfaction, since it is so successful in practice. When the automatic decision made by technique is obeyed, it is more successful than when a comparatively inefficient approach is adopted: when technique is applied, wars are won, more widgets are manufactured, and energy is saved. CAIS itself observes this logic is all too likely to guide the future development of AI as it has already long guided technological development, even at the expense of human safety.[27]

Self-Augmentation[28]

The consequence of the automatic success of technical operations is that technique is self-augmenting. This means that with each successful technical operation, there is demand that technique be applied more widely, which in turn garners it further success and more widespread application. What is more, Ellul argues that not only is it assured that the application of technique will increase, it is assured independent of the work or choices of any individuals. Ellul does not mean that increase in application of techniques is a result of common effort, but rather that the factors which determine this increase in application are primarily technical:

> We can no longer argue that it is an economic or a social condition, or education, or any other human factor [that determines technical progress today]. Essentially, the preceding technical situation alone is determinative. When a given technical discovery occurs, it has followed almost of necessity certain other discoveries. Human intervention in this succession appears only as an incidental cause … Technique, in its development, poses primarily technical problems which consequently can be resolved only by technique. The present level of technique brings on new advances, and these in turn add to existing technical difficulties and technical problems, which demand further advances still.[29]

Far from the wealth (or poverty) of a given society, or the attitude adopted by a host of researchers or educational institutions, the determining factor of technique is nothing other than technique. Technical developments are not the result of an excess or desire for wealth in a society, nor are they driven by a society that has built and can

[26] Ellul, *The Technological Society,* 80.
[27] Hendrycks, Mazeika, and Woodside, "Catastrophic AI Risks," 18–23.
[28] I discuss the self-augmentation and the autonomy of technique in more detail elsewhere. See B.W.D. Heystee, "The Unlovable Violence of Technique: George Grant's Reception of Jacques Ellul," *The Philosophical Journal of Conflict and Violence*, 2023.
[29] Ellul, *The Technological Society,* 90–92.

support institutions out of a pure desire to discover things, etc. etc. The sole reason that we apply technique to an ever increasing array of domains is that previous applications of technique demands this increase.[30] As one technical development is implemented, it presents problems or difficulties that need to be addressed; this brings forth further technical developments, which themselves produce problems or difficulties which in turn must be addressed by technique, since technique is the only means of addressing technical problems. [31] Further, because each technical development tends to present several difficulties and/or opportunities, the application of technique does simply increase, it accelerates. Technical developments tend to reverberate through several fields or even create new ones, which in turn produce further technical developments.[32] Consequently, when we apply a technique, we are not leading but are participating in a process of automatic and accelerating growth in the application of techniques in general.[33] This process of "self-augmentation" is not a result of individual or collective deliberation, but primarily a result of the effects of previous applications of technique.

In the context of the present paper, Ellul would say that developments in AI have not been a consequence of, say, idle curiosity or financial incentives, but of responses to technical problems: e.g., image recognition software and natural language processing are the result of the need to process increasingly large quantities of data, quantities produced in response to prior technical difficulties. As neither these underlying difficulties of data processing nor the consequent difficulties associated with the prevalence of natural language processing will simply disappear on their own, the process of technical self-augmentation will continue to drive AI development into the future.

Monism

The nature of this automatism and self-augmentation means that the totality of various modern techniques is also monistic. The word Ellul uses which has been translated as "monism" is *unicité,* which the translator notes may also be rendered as "holism." Neither term is exactly right, but what Ellul means by *unicité* here is that "the technical phenomenon, embracing all the separate techniques, forms a whole" and that

[30] Lovekin likens the exclusion of human decision-making to a kind of technical "collective unconscious, encouraging the anonymous but steadfast involvement and the submersion of the individual in the technical process." Lovekin, "Logic of Technology," 258.
[31] Darrell J Fasching, *The Thought of Jacques Ellul: A Systematic Exposition*, vol. 7, Toronto Studies in Theology (Toronto: Edwin Mellen Press, 1981), 18.
[32] Ellul, *The Technological Society,* 91.Ellul notes that of course not every field is constantly accelerating and that fields do stall from time to time. But these are exceptions that prove the rule: the general fact is that technique develops more rapidly today than it did yesterday, and will be yet more rapid tomorrow.
[33] Langdon Winner, *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought* (Cambridge: MIT Press, 1978), 61.

the components of the whole technical phenomenon are tied together and cannot be meaningfully isolated from one another.[34] In effect, the world of technique becomes a "closed world" from which parts cannot be removed.[35] In other words, in the technological society various individual techniques cannot be separated from each other, nor can they be separated from their effects. The interrelatedness of various techniques was implied in my above discussion of self-augmentation. The very fact that one technical development necessarily entails several other developments in other fields speaks to interrelation; indeed, it is precisely interrelatedness that makes these rapidly multiplying and accelerating developments an inevitable outcome. Superficially disparate fields in fact cannot operate without the cooperation of various other fields.[36] Neither can a technique be separated from its use and its effects: Ellul insists that a technique *is* a use, and consequently is also an effect. The potential applications of technique are not meaningfully distinct from the actual applications, except the temporal distinction of before and after. When a technique is applied it is necessarily the best, most efficient course of action for a given scenario. That means that a bad use of, e.g., a machine is not an example of technique at work.[37] While non-technical uses of machines are in principle possible, that is the exception to the rule; it is the deliberate but irrational decision made by an individual in flat contradiction to the automatic "decision" made by technique.

But is it not possible for better and worse technical developments to be encouraged or discouraged, and thereby ensure more or less better uses of technique? Such a question is obviously relevant to the development of AI, since the warnings of Bostrom and CAIS are predicated on the assumption that AI can be developed in better and worse ways. In response to such a question, Ellul cites the example of the atomic bomb. It may be tempting to say that it would have been better for humans to develop nuclear energy without the bomb: put nuclear techniques to good use and not to bad use. Yet Ellul would remind us that atomic research requires passing through the stage of the atomic bomb. The technical problems associated with a bomb are prior to the technical problems associated with industrial energy use, a fact corroborated by Oppenheimer himself.[38] Society cannot simply skip or circumvent the

---

[34] Ellul, *The Technological Society,* 94.
[35] Lovekin, "Logic of Technology," 261.
[36] Ellul offers this illustrative example: "The case of the police, for example, cannot be considered merely within its specific confines; police technique is closely connected with the techniques of propaganda, administration, and even economics. Economics demands, in effect, an increasing productivity; it is impossible to accept the nonproducers into the body social … The police must develop methods to put these useless consumers to work." Ellul, *The Technological Society*, 111.
[37] Ellul cites the example of using a car to drive to work versus using it to kill one's neighbour. While both outcomes are possible (and indeed the latter does occasionally happen), the latter is not an application (and hence *abuse*) of technique: "Technique is in itself a method of action, which is exactly what a use means … The driver who uses his automobile carelessly makes a bad use of it. Such use, incidentally, has nothing to do with the use which moralists wish to ascribe to technique. Technique *is* a use." Ellul, 98.
[38] Ellul, *The Technological Society,* 99.

bad parts of technique. The drive to efficiency has its own logic and proceeds along the necessary steps, regardless of whether they seem to us good or bad. The monism of technique means that various individual techniques and their uses exist as a single integrated whole so that we cannot pick and choose better or worse uses, as if we were at a technical supermarket. As we carry out technical operations automatically, and as those operations are applied to an ever increasing scope of society, we cannot reject certain parts of technique without rejecting the whole.

The Necessary Linking Together of Techniques

Ellul argues that the wholistic integration and interdependence of techniques has not been an accident, but rather has been a necessary consequence of the modern development of technique. Ellul says that these connections necessary because each technique has *demanded* the emergence of other techniques.[39] For example, increasingly productive machinery required new commercial techniques so that the machines could be put to work optimally. Then, the production and especially consumption of additional goods required that humans be relocated to cities, necessitating the development of urban planning and mass amusement to make urban life tolerable. Economic and labor techniques were then necessary to ensure a relative equilibrium between steady production, distribution, and consumption. This includes the educational apparatus necessary to training a technical workforce. All these various fields and the more specific techniques within those fields emerged out of necessity and then developed in a state of interdependence so that the techniques are necessarily 'linked together.'

We see this linking together in the development of AI. The dependence of software engineering on the mining of precious metals, educational programs, and urban development is clear enough. Yet we may also say that those fields in turn depend on software development (and will eventually depend on AI) so that they can develop and proceed efficiently; mining will require automated machinery at the rock face, education will need to process and evaluate data on limitless students in increasing detail, and urban planning will require complex models and algorithms to predict future needs. All these various techniques are necessarily linked together in their origins and will forge ever closer links as they develop in pursuit of greater efficiency.

[39] Ellul, *The Technological Society,* 116.

Technical Universalism

The monistic and necessarily interrelated application of various techniques has become universal, and in two senses. The application of techniques is geographically universal and it is qualitatively universal. Virtually every corner of the globe has been colonized by modern techniques of various kinds, and so too for nearly every aspect of our lives and cultures. The geographic universality of technique is, in my view, an uncontroversial claim. Though we do observe variety in techniques actually used as a consequence of climate, available resources, or the stage of technical development of a given people, the fact is that techniques are used everywhere. What is more, the differences are not a result of variety in social customs and priorities (as was the case with traditional technique), but in the fact that objective conditions mean that there are slightly different ways of achieving maximal efficiency. The competition between rival foreign powers in computing power, cyberwarfare, and other aspects of digital infrastructure speaks vividly to the fact that technique now spans the globe and in so spanning has created a kind of international technical homogeneity.[40] In Ellul's day, such universality was already evident in the ever more integrated global shipping industry, with its more or less uniform ocean vessels, port installations, railroads, shipping containers, and standards of every kind.[41] Every place is required to be brought up to and maintained at the standard of maximal efficiency.

The technological society is qualitatively universal in the sense that it increasingly defines every aspect of life so that local and national cultures are diminished and differences between one way of life and another disappear. Self-augmentation means that technique will run up against problems in more and more areas of life. Automatism means that when those problems are encountered, the decision will be in favor of technique. The monism and necessary linking-together of technique mean that certain aspects of society cannot be meaningfully insulated from the encroachment of technique. Even in the case of art or literature, areas where it would seem to be least appropriate, technique has become dominant: artistic expression cannot ignore techniques of finance or telecommunication necessary to artistic production and distribution.[42] And this is to say nothing of more recent developments of which Ellul could not have known: artificial image generation software such as DALL-E or the advent of audio and visual deepfakes.

The most remarkable consequence of this is that no longer is technique subordinate to a more comprehensive civilization, but rather "technique has taken over the whole of civilization."[43] Whatever social or civilizational ends directed the development of technique in the past, those have either disappeared or been

[40] As Lovekin observes, whatever cultural dichotomy there was between East and West, technique has almost completely erased it. Lovekin, "Logic of Technology," 263.
[41] Ellul, *The Technological Society,* 119–20.
[42] Ellul, *The Technological Society,* 128.
[43] Ellul, *The Technological Society,* 128.

transformed so that they have only a secondary status. Though there may be differences from place to place, these differences are by and large vestiges of bygone civilizations that technique has not yet erased because there are presently greater impediments to efficiency that must first be addressed. According to Ellul, the differences between civilizations are superficial in comparison to their technical unity. Technique is universal in scope and exhaustive in detail.

Autonomy

We now turn to the final and decisive characteristic of technique: autonomy. The autonomy of technique is, in effect, a kind of crowning characteristic. It is the result of the combination and sum of the other characteristics, though it has the effect of reinforcing those characteristics. Ellul says that technique is both practically and morally autonomous. In the section on technical autonomy he makes little effort to prove this claim, taking it as evident based on his prior discussion of the other characteristics.

Technique is practically autonomous in the sense that, for example, it is autonomous with respect to economics and politics. Ellul writes, "Neither economic nor political evolution conditions technical progress. Its progress is likewise independent of the social situation. The converse is actually the case . . . Technique elicits and conditions social, political and economic change."[44] The prime factor which determines the others is technique. The other social factors are consequent upon it. Though it might seem the other way around at times, this is a misconception. The relocation of humans to cities might have seemed an economic determination because of, e.g., the desire for wealth on the part of factory owners, but Ellul maintains it was more fundamentally a response to the technical problem of how to efficiently distribute and consume the more plentiful goods produced in a factory. The economic conditions in this case were secondary, a byproduct so to speak. In practical terms, "External necessities no longer determine technique. Technique's own internal necessities are determinative."[45]

More striking is Ellul's claim that technique is morally autonomous. Ellul claims that technique is the author of its own morality and accepts no external limitations or judgments. For our purposes, we may say that morality is judgment about what ought and ought not to be done. The moral autonomy of technique means that only technique determines whether its own actions ought to be done or not:

> Technique tolerates no judgment from without and accepts no limitations . . . Morality judges moral problems; as far as technical problems are concerned,

[44] Ellul, *The Technological Society,* 133.
[45] Ellul, *The Technological Society,* 133–34.

> it has nothing to say. Only technical criteria are relevant. Technique, in sitting in judgment on itself, is clearly freed from this principal obstacle to human action . . . technique theoretically and systematically assures to itself that liberty which it has been able to win practically.[46]

The sum of technique's other characteristics—automatism, self-augmentation, monism, linking-together, and universalism—means that technique operates according to its own logic and its own determinations about what is necessary or forbidden. Technique determines for itself what the problems and the solutions are. Whether or not something is 'moral' according to more traditional standards has no significant bearing on technique's operations.

And to be clear: this does not simply mean that technique has moral implications, or that it is 'not neutral.' What Ellul means is that technique sits outside other moralities because it is the author of its own morality: "It was long claimed that technique was neutral. Today this [whether or not technique is neutral] is no longer a useful distinction. The power and autonomy of technique are so well secured that it, in its turn, has become the judge of what is moral, the creator of a new morality."[47] Technique cannot be morally righteous or problematic, because technique determines for itself what is moral: that which satisfies the continued drive to efficiency. All other considerations are systematically excluded from technical decision making. [48] Technique is not good or evil (or neutral) because it is beyond good and evil. Technique has no goals outside of itself.

The combined effect of the various characteristics of technique, capped off by autonomy, is that technical morality is not limited to a special province, but colonizes every aspect of human activity and systematically excludes any factors that might interrupt its drive to efficiency. This is clearest in the way that technique progressively reduces the role that humans play in any technical operation, whether it be factory workers, airplane pilots, or statisticians. Ellul explains that when human interference in a given activity cannot be eliminated or substantially reduced, humans are adjusted to become more technical so that they more closely resemble the machines they are operating; humans become an appendage of technique rather than a user.[49] Humans are not permitted to interfere with technique nor do they contribute to technique's activity in a uniquely human way. They are only permitted to participate in a technically determined operation as simply one part of the machine among many. Put more generally, whatever human ends, interests or desires could have disturbed technique's efficiency, they are all diminished or excluded from technique's ever increasing domain

---

[46] Ellul, *The Technological Society,* 134.
[47] Ellul, *The Technological Society,* 134.
[48] Lovekin, "Logic of Technology," 264; Helena Mateus Jerónimo, José Luís Garcia, and Carl Mitcham, "Introduction: Ellul Returns," in *Jacques Ellul and the Technological Society in the 21st Century*, vol. 13, Philosophy of Engineering and Technology (Dordrecht: Springer, 2013), 4.
[49] Ellul, *The Technological Society,* 134–40.

so that the sole criteria are technical and no alien morality could limit or redirect technical activity. While continuously expanding the scope and detail of its activity, technique sets the terms that justify that activity and refuses the possibility of any other terms.

In Ellul's judgment, this is the nature of technique and technique is the chief determining factor in society today. Technique operates according to its own logic and its own morality and leaves no opportunity for meaningful human intervention. Its decisions are automatic. It increases the scope and detail of its influence of its own accord. It is a unified whole whose component techniques cannot be separated from one another. It is not an instrument or even a sum of instruments that can be subordinated to human ends, let alone be used for good or for evil. It shapes and determines the way humans live and the ends we pursue, all the while persuading us that it is *we* who use *it* freely and for our own purposes.

The Pressing Reality of Out-of-Control Technique

Another way of saying that technique is "autonomous" is saying that it is "out-of-control." In detailing these six characteristics of technique, Ellul is arguing that the technological society is organized in such a way that the ends pursued by such a society are not the result of human deliberation or choosing. The increasing technification of society is no longer a direct or indirect consequence of human reflection about what our ends are or should be, but rather as a consequence of the fact that it is already organized technologically. Though such reflection may have had a role in the genesis of technique, it is no longer consequential.[50] The technological society's organization around the pursuit of efficiency is not only self-sustaining, it is self-augmenting and autonomous. With each passing year, the technification of society grows more encompassing and more detailed, not because as free humans we have deemed this to be good, but simply because society is already technological. The logic that determines the role—or rather, predominance—of techniques in our society is itself technical.

What is more, for the most part humans are neither passive participants nor actively resisting opponents of technique; they are willing contributors. Part of the technological society is the formal and informal education necessary to such a society. This education ensures compliant and efficient workers to carry out technical operations: accountants, physicians, factory workers, and bureaucrats are all trained to carry out their operations as efficiently as possible so that the sum total of techniques

[50] Ellul does not note the role of such reflection in his account of technique's origins, but many others do. To cite just one example and to neglect countless others, in his reception of Ellul, George Grant argues that modern technique emerged out of the affirmation that "man's essence is his freedom and therefore that what chiefly concerns man in this life is to shape the world as we want it." George Parkin Grant, *Technology and Empire: Perspectives on North America* (Toronto: House of Anansi Press, 1969), 114 n. 3. Technique was a means of making human freedom concrete.

can continue its overall drive toward absolute efficiency. That is to say, when technique is understood not simply as a consistent means of producing a result, but in the peculiarly modern sense of a "a totality of methods rationally arrived at and having absolute efficiency," it is clear that humans themselves belong to that totality and for the most part do not stand outside it.[51]

This is why I say that technique is "out-of-control" but I do not adopt the language of AI critics and say that it is "rogue." While "rogue" would imply an antagonistic relation—the AI standing over and against us, undermining our conscious interests and causing explicit frustration—technique does not do this. Rather, technique integrates and technifies human interests and ends so that they become a compliant part of the technological society. The social and educational institutions of the technological society persuade its members that technique is desirable, because in its efficiency it apparently provides us with greater capacity to pursue our freely chosen ends. Never mind the fact that we are encouraged to choose ends agreeable to technique, and that little time is left for ends of other sorts. (To the extent that someone does choose an atechnical, inefficient life, they are marginalized from society so that they pose little threat to technical operations.) Because the way the technological society assimilates humans to the overall pursuit of efficiency, we can say that not only is technique insubordinate to human ends, in practice it is *superordinate* to human ends. There is no antagonistic relationship to technique and so it should not be called "rogue." But it is not subordinate to human purposes so it should be called "out-of-control."

The prospect of a technological creation getting out-of-control is not a looming possibility which we should be careful to avoid or mitigate, but a present reality which presses in upon us at this very moment. When organizations like CAIS and thinkers like Bostrom pose the possibility of a rogue or superintelligent AI as something which threatens us not yet, but in the future, they overlook a crucial feature of our present society. Bostrom's warnings about an artificial superintelligence and his exhortations that we carefully program our AIs with human values assume that (1) we are presently still in control of our machines and (2) that our current values are freely chosen and form the basis of a more or less desirable society. Ellul would contend that neither is the case. Indeed, he would likely argue that the future of which CAIS and Bostrom warn us is in fact our present. Yet those AI critics do not see this because they have misunderstood the technological danger in three crucial ways.

First, they assume that there will be some identifiable machine or AI which goes rogue and poses a specific threat to which we could point. Recall that Bostrom argues that an artificial superintelligence is, in principle, possible and is perhaps likely over the

---

[51] Ellul would concede that human freedom allows for individual opposition to technique, but this is the exception rather than the rule. Nevertheless, if enough individuals make a stand, it is possible for these 'exceptions' to change the direction of society in an unforeseeable way. Ellul, *The Technological Society,* xxix; cf. Daniel Cérézuelle, "Jacques Ellul, Penseur du Système Technicien," *Futuribles* 429, no. 2 (2019): 85–88.

next several centuries. In other words, the superintelligence which goes rogue is a specific project or creation which may come into existence at some point. The superintelligence may exist in a certain piece of hardware or it may be distributed across a large network, but it would have a definite and particular existence. Indeed, it is for this reason that Bostrom considers the difficulties associated with "capability control," including physical confinement.[52] The idea of capability controls (whether or not they are ultimately feasible) only make sense if the technological threat is something like a specific machine or mechanism. The autonomy of technique, however, is not like this. The self-augmenting, autonomous nature of the technological society is not limited to a machine or mechanism of any particular size or distribution, but rather it encompasses the whole of society in which we live at every moment. Nearly every feature of our social organization is determined or shaped by technical operations of some kind so that nothing escapes the drive to efficiency. There is no machine that we could switch off, nor any software that we could reprogram, because technique is not a machine which lies outside of us. In the technological society, the out-of-control drive to efficiency is everywhere and nowhere and therefore cannot be resisted in the way Bostrom or CAIS propose.

Second, Bostrom and CAIS assume that we still have the capacity to (re)direct the development of AI so that we can avoid the pitfalls and enjoy the benefits. This is a natural assumption given that they believe the dangers of rogue AIs and superintelligences to be in the future. Yet this assumption does not consider the nature of technological development and how it is rarely, if ever, the consequence of human deliberation and free choosing. Developments in technique arise as the necessary response to problems previously posed by technique, and they exist as a single integrated whole. It is not up to individual developers to pick and choose from among the various techniques so that we get the good without the bad. Nuclear energy could not develop without the nuclear bomb, and neither development could have been indefinitely forestalled on account of the risk they posed; they were occasioned by prior technical developments and requirements, and the stages through which they progressed where themselves determined by the logic of technique. So too, we are to expect, will be the development of AI: there may be calls to 'pause' research on AI or arguments that certain aspects of AI should be encouraged or discouraged, but the social forces which drive AI development do not listen to such calls and arguments. So long as we live in a technological society, our ability to practically limit or direct technical developments will be marginal at best.

Third, and most importantly, the warnings of Bostrom and CAIS assume that the world of out-of-control technology is not yet here. These warnings place the rogue AIs and superintelligences in the future and thereby imply that out-of-control technology is a distant prospect which has little bearing on our present lives, except perhaps for the professional lives of those technical experts actually involved in

[52] Bostrom, *Superintelligence*, 129 ff.

software development. They suggest that we are currently in control of our machines, and we should be careful lest those machines get out-of-control in the future. Yet it is not clear that we are in fact in control. Though no particular machine has yet defiantly resisted us in a significant way, it is not true that the machines we build are subordinate to our freely chosen, human purposes. Both the machines and their human operators presently exist within a broader, technical milieu which shapes and determines what actions ought or ought not be taken. Human deliberation and reflection are not in control of technical development or decision-making today.

It is not my intention with this paper to dismiss out of hand the practical concerns of thinkers like Bostrom and organizations like CAIS. They would seem to make reasonable cases why a superintelligence or a rogue AI would pose an existential threat to humanity. Indeed, I am willing to defer to their technical expertise on the likelihood and consequences of that specific scenario. I am happy to take their word for it that we need to consider the prospect of a paperclip maximizing machine gone wrong. Rather, my intention is to impress upon the reader that we should not allow warnings about the future to obscure the nature of the present. Concerns about a malevolent superintelligence or catastrophically incompetent AI risk overshadowing present technological difficulties. In particular, they risk convincing us that we are really still in control of our machines for the time being, when in fact technology is already out-of-control. The challenge posed by the autonomy of technique is not a future potentiality, but a present reality and one that must be addressed with the urgency that the present deserves.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Artificial Intelligence: Thoughts from a Psychologist

Micheal J. Meitner
University of British Columbia, Canada

Abstract
In its current state, Artificial Intelligence (AI) is still very far from reaching the complexity of the human brain. Technological progress, however, might bring about AI as a general intelligence surpassing our own in important aspects. The concept of neurodiversity is introduced and it is suggested that the field of AI might benefit from the incorporating of this concept in the form of sets of neurodiverse AIs from which a diverse set of solutions can be generated. Issues of sustainability and equity are also discussed in light of rapid advances in the field.

Keywords: artificial intelligence, neurodiversity, AI risks, open access data, machine learning

Introduction

Artificial intelligence, or AI as it is commonly referred to, is a suite of technologies that are poised to change the world as we know it. The concept of AI has been with us throughout antiquity in the mythologies of the Greeks and in early conceptions of automata. Early work in cybernetics and eventually neural networks brought this concept out of the realm of fantasy and into the modern world. In 1952, Marvin Minsky and Dean Edmonds succeeded in creating the world's first functional neural network machine, the Stochastic Neural Analog Reinforcement Calculator or SNARC.[1] While this was no doubt an impressive achievement of the day, it did not really live up to the dreams of the ancients of a machine that would embody more human characteristics.

The holy grail of AI has always been to create a machine capable of generalized intelligence. In fact, the first know test of generalized intelligence in AI was posited by Alan Turing in his seminal paper "Computing Machinery and Intelligence" and to this day is known as the Turing Test.[2] The point of this test was to ascertain if a machine could fool a human into thinking it was conversing with an actual human being. When I was in graduate school, I had the pleasure to interact with ELIZA, a computer program written in 1966 by Joseph Weizenbaum to mimic the behavior of a Rogerian psychotherapist which some would argue was the first program to pass the Turing Test.[3] While this contention remains controversial to some I can personally attest to the convincing nature of the program. However, I can also attest to the fact that it was quite easy to trip up this software and therefore destroy the illusion. It could never imagine new or novel situations and often answered any question that required creativity with a question of its own. Some historians of the internet age might say that ELIZA represented the first "bot," a software program that imitates the behavior of a human, as in participating in chatroom or IRC discussions. As most of us know today, bots have become far more sophisticated and for many of us they seem quite human when we interact with them.

However, there remains a massive disconnect between imitating a human and creating an artificial human brain. The human brain contains approximately 86 billion neurons and each neuron has on average 7000 synaptic connections yielding nearly a quadrillion synapses.[4] In terms of simple computational power (measured in floating

---

[1] Marvin Minsky, "A neural-analogue calculator based upon a probability model of reinforcement," (Technical document, Harvard University Psychological Laboratories, Cambridge, Massachusetts January 8, 1952).
[2] Alan Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 433–60. http://www.jstor.org/stable/2251299.
[3] Joseph Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM* 9, no. 1 (1966): 36-45.
[4] David A Drachman, "Do we have brain to spare?," *Neurology* 64, no. 12 (2005): 2004-2005; Herculano-Houzel, Suzana. "The human brain in numbers: a linearly scaled-up primate brain," *Frontiers in human neuroscience* (2009): 31

point operations per second or FLOPS) the human brain is estimated to be capable of approximately 1 exaFLOP (10^{18}).[5] Modern technology still falls short of this degree of raw computational power. The world's fastest supercomputer, Fujitsu for Japan's RIKEN Center for Computational Science supercomputer, has currently achieved .422 exaFLOPS.[6] However, it should be noted that important architectural aspects of the human brain are even further from the realm of possibility currently. Simply having the ability to do the same number of calculations over time does not mean that the arrangement of those neuronal units is in anyway similar to that of a human brain. Even in the case of AI modeling of C. elegans, a common worm that has only has 302 neurons, researchers are still refining the architecture of that model based on new electron microscopy data.[7] Therefore, the goal of a generalized intelligence instantiated in a computer is likely very far in the future. One possible technological development that may change this calculus is quantum computing but this still has significant challenges to overcome to become relevant to this discussion. Computational power (quantum or not) will certainly close the gap but this belies the fact that human brain is not simply the sum of its abilities to do raw computations.

In general, I would say that AI, in its current form, is in no way like the human brain even though AI researchers use architecture developed from observations of neuroanatomy. Modern AI is mostly focused on "narrow, shallow or weak AI" tasks such as finding patterns in our purchases and suggesting new ones based on these patterns. Even those AI's considered "broad, deep or strong AI" do not really approach the complexity of the human brain. Deep AI consists of numerous neural networks often hierarchically arranged that allow for deeper levels of abstraction from the inputs in the model. In addition, deep AI techniques deal well with unstructured data and can analyze that data in an unsupervised fashion. These qualities have made deep learning techniques quite ubiquitous and they have been employed to tackle problems such as speech recognition and computer vision. Artificial general intelligence, on the other hand, will require substantial leaps in both hardware and software before this can be realized.

At this point I would like to compare and contrast the nature of artificial and human general intelligence as seen in table 1 below.

---

[5] A point should be made that direct comparison of the human brain's computational power and a computers is not technically possible to achieve. For more information see "Brain performance in FLOPS," aiimpacts.org, AI Impacts, January 13 2021, https://aiimpacts.org/brain-performance-in-flops/.

[6] Scott Fulton III, "Top500: Japan's Fugaku Still the World's Fastest Supercomputer," Data Center Knowledge, November 18 2020, January 26 2021, https://www.datacenterknowledge.com/supercomputers/top500-japan-s-fugaku-still-world-s-fastest-supercomputer.

[7] Steven J. Cook, Travis A. Jarrell, Christopher A. Brittin, Yi Wang, Adam E. Bloniarz, Maksim A. Yakovlev, Ken CQ Nguyen et al, "Whole-animal connectomes of both Caenorhabditis elegans sexes." *Nature* 571, no. 7763 (2019): 63-71.

Artificial	**Human**
Infinite sensors	Limited sensors (can be augmented)
Infinite dimensions	Dimensionally challenged
Infinite data storage	Limited
Technology bound	Organism bound
Fairly stable goals (can be made to evolve)	Changing and evolving goals
Ever increasing processing speed	Speed mostly fixed
Replication generally yields copies (unless a genetic algorithm is used)	Replication yields neurodiversity
Consciousness?	Multiple unconsciousness systems partially discovered by consciousness

Table 1: A comparison of artificial and human general intelligence

As is evident from table 1, artificial general intelligence holds much promise and will likely lead to the formation of an artificial superintelligence. Being able to surpass our limitations in data sensing, data storage (memory in humans) and in hyper-dimensional thinking at speed will allow AIs to make tractable those problems that have long eluded us. The goals of AI and the eventual architecture (and potential diversity of architecture) seem to be important turning points in our thinking about how we might make progress toward the creation of artificial general intelligence. Let us start by first turning our attention to diversity in AI.

Neurodiversity and AI

The concept of neurodiversity has been with us since 1998 and refers to the revelation that variation in the human brain is vast and while some variation may be detrimental other variations may represent significant strengths or improvements. In fact, I would go as far as saying that neurodiversity can in fact represent a competitive advantage. If true for humans, this surely would be true of diversity in AI as well. It has long been known that genetic algorithms (GAs) can be used to spawn novel architectures for neural networks that can be used to evaluate the degree of performance of its progeny on some fitness function or goal. This allows for competition between various forms of an AI algorithm and leads to better solutions to problems that the AI is tasked with.

This represents some degree of neurodiversity in AI already, albeit a weak form of it, as unsuccessful progeny are "killed off" and therefore diversity is not maintained. Ideally, neurodiverse AI systems would be persisted and alternate solutions could be investigated to allow for insights into divergent approaches that may help us to better define and build robust and resilient AIs in the future.

Ultimately the discussion of the concept of neurodiversity in the context of AI causes us to question our ideas about goals. Goals in AI must be made explicit in some way and often represent the most challenging aspect of creating a functional AI. For many AIs there are more than one goal that the algorithm is trying to maximize or balance amongst. However, all of these goals have a context and perspective. From a user's perspective, a common goal might be increasing the relevance of information retrieved based on a query. From the company's perspective a similar goal might be user engagement. These differences in defining goals can have significant effects on the outputs of an AI. In fact, they define them. Variability in goal definition over time allows a model to adapt to changing system conditions.

As referenced in table 1, human goals seem to be ever changing and evolving as our understanding of the world progresses. This is especially true in the case of "wicked" problems. Wicked problems are those that defy simple solutions and are often comprised of multiple interacting systems. They are wicked because they are typically poorly understood, include contradictory information and are highly variable over time. Wicked problems do not have an optimal solution, rather they have temporary or partial solutions that are likely themselves to change over time. The changing nature of wicked problems and the large uncertainties in their predictions mean we have to take an adaptive approach to the problem. Like wicked problems, adaptive problems are where the problem definition is mostly unknown. Adaptive problems often require the locus of control for solving the problems to be decentralized. Stakeholders become the focus rather than disciplinary experts and as we well know stakeholders often have a variety of perspectives on a problem. This is the type of diversity needed if we hope to be able to conceptualize the system properly. From that one might argue that this means that multiple AIs might be needed to focus on various specificities of a problem in order for a larger definition of the problem to occur.

AI in the Environmental Sciences

Many of the issues of our day are in fact adaptive problems, such as most of our environmental current problems. The World Economic Forum report titled "Harnessing Artificial Intelligence for the Earth" states that there are 6 priority action areas for addressing environmental issues: 1) climate change, 2) biodiversity and conservation, 3) healthy oceans, 4) water security, 5) clean air and 6) weather and

disaster resilience.[8] Each of these areas has a series of sub areas that AI could be applied to in order to create a more sustainable future. In the case of climate change they refer to: clean power, smart transport options, sustainable production and consumption, sustainable land-use, smart cities and homes. AI can be applied to all of these areas and in certain cases have the potential to transform these sectors. Consider a modern energy grid that can use AI to adapt to changing supply and demand, incorporate traditional power sources with clean energy source and to make distributed energy possible at scale. This would seriously improve our ability to meet our climate change targets. As well, significant improvements in transportation, agriculture, and water management systems can also be realized by the application of AI technologies.

AI has already been applied to many environmental problems. Monitoring endangered species,[9] tracking diseases,[10] crop optimization,[11] smart buildings and associated IoT to increase efficiency,[12] predicting storms,[13] and managing traffic[14] are but a few of the many applications of AI in the environmental domain. In all of these cases, AI offers us a method to deal with the massive degrees of complexity that represent these wicked environmental problems. This is made possible by the vast quantities of data that we are currently collecting to support decision making in these areas.

The world of "big data" has arrived and no technology is better poised to make use of this plethora of data than AI. In fact, without computer aided decision making, I would venture to guess that we would not be able to effectively navigate, understand or even utilize the amount of data that is currently available. AI, however, has a special relationship with big data and becomes better when provided with increasing data volumes. AI is especially good at detecting anomalies in massive data sets, determining the probabilities of future outcomes and it can recognize patterns that human cannot.

---

[8] Celine Herweijer, Benjamin Combes, Pia Ramchandani, Jasnam Sidhu, "Harnessing Artificial Intelligence for the Earth," www3.weforum.org, World Economic Forum, January 2018, January 17 2021, www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf

[9] Antoine M. Dujon, and Gail Schofield, "Importance of machine learning for enhancing ecological studies using information-rich imagery," *Endangered Species Research* 39 (2019): 91-104.

[10] Zoie SY Wong, Jiaqi Zhou, and Qingpeng Zhang, "Artificial intelligence for infectious disease big data analytics," *Infection, disease & health* 24, no. 1 (2019): 44-48.

[11] Tanha Talaviya, Dhara Shah, Nivedita Patel, Hiteshri Yagnik, and Manan Shah, "Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides," *Artificial Intelligence in Agriculture* 4 (2020): 58-73.

[12] Rav Panchalingam, and Ka C. Chan, "A state-of-the-art review on artificial intelligence for Smart Buildings," *Intelligent Buildings International* 13, no. 4 (2021): 203-226.

[13] Amy McGovern, Kimberly L. Elmore, David John Gagne, Sue Ellen Haupt, Christopher D. Karstens, Ryan Lagerquist, Travis Smith, and John K. Williams, "Using artificial intelligence to improve real-time decision-making for high-impact weather," *Bulletin of the American Meteorological Society* 98, no. 10 (2017): 2073-2090.

[14] Rusul Abduljabbar, Hussein Dia, Sohani Liyanage, and Saeed Asadi Bagloee, "Applications of artificial intelligence in transport: An overview," *Sustainability* 11, no. 1 (2019): 189.

AI and Risk

The same World Economic Forum report that was mentioned above also identifies 6 areas of risk for AI. They are: performance, security, control, economic, social, and ethical. Performance risks refer to problems in deciphering the "black box" inner workings of an AI. Because we have little insight into what an AI is actually doing we have difficulties in knowing if its performance is accurate or even desirable. Issues of model fit are also complicated by this. If an AI is inferring future trends based on historical records then we need to wonder if those records contain enough information to support such prediction. If we don't know what an AI is doing internally then this problem is certain exacerbated.

Security risks, mentioned in this report, are also of concern. They reference "hackers" and the problems of bad actors manipulating algorithms to take control of them. This brings to light a more serious concern of who has control over these algorithms. Most AIs are in the hands of governments or large private sector companies. Neither of these has a great track record of acting for the social good. Private companies have a fiduciary duty to act in the best interests of stakeholders and while they may make efforts to address social issues this will never be their primary concern. However, one could argue that a government's main interest is the public good but as we all know this can be perverted in service of other goals that do not in fact create nor maintain social good. Even if these actors had social good in mind, how is it defined? Would those actions taken by these actors result in increased social good? This is an open question and certainly needs more thought and discussion to determine how to fully define this risk.

Control risks are some of the most blown out of proportion but are also some of the most worrying. This is where common narratives of post apocalyptic worlds governed by intelligent machines that have decided that humans represent a threat come in. However, this does not really represent a credible threat because you would need an AI capable of general intelligence and we have already determined that the likelihood that this will materialize in my lifetime is remote at best. What is of more concern are AIs that have direct control of various systems that might make decisions that lead to unintended consequences. One example of this is the flash crash of the US stock market in 2010 which was likely caused by interaction of multiple AI bots all speed trading at the same time.[15]

Economic risks are also potentially significant for AI as it moves forward. Companies that do not have access to AI or the associated data to drive them run the risk of being out competed. This in turn creates the risk that the business landscape will continue to shrink, creating increased inequity of wealth distribution and

---

[15] Tom Lauricella, Kara Scannell, and Jenny Strasburg, "How a Trading Algorithm Went Awry," The Wall Street Journal (New York, NY), October 2, 2010.
https://www.wsj.com/articles/SB10001424052748704029304575526390131916792

consolidating power with a few multinational companies. This may lead to a circumstance where a few companies begin to exert more power over global progression.

Social risks of AI are often defined as adaptation to increased automation pressures created by increased use of AI. Job loss and increased unemployment are real possibilities in a world where AI takes over much of the work of running the systems that we rely on. Additionally, AI algorithms can potentially be biased against certain factions of society, underpinning historic social inequities. New inequities can also be created by AI as it fundamentally changes the sector with in which it is being applied. Take as an example the transportation sector where autonomous vehicles are poised to massively disrupt people's lives who rely on this sector for employment.

The last risk that this report discusses is ethical risks. What choices will an AI make? Will they be beneficial choices? What about fairness and human rights? Privacy concerns are also discussed here. While all of these risks are important and represent an excellent attempt to get us all thinking about how AI will shape our future I believe that there is a significant omission in the risks associated with continued development and application of AI technology.

Additional Risks

I would add two additional risks to this list; access to both data and the knowledge needed to make sense of it. Let us first tackle data access. AI does not represent a valuable technology without the data that drives it and data is not generally freely available. Of course there are open data sets but the vast majority of meaningful data being generated today is in the hands of private corporations or governments. In 2020, every minute of every day we collectively generate 500 hours YouTube video, WhatsApp users share 41,666,667 messages, Facebook users upload 147,000 photos, Instagram users post 347,22 stories, and TikTok is installed 2,704 times.[16] The amounts of data being generate currently is staggering and for the most part we create this data. It is estimated that in 2020 each person on earth generates 1.7 MB of data per second.[17] Because access to this proprietary data is in the hands of the few, and by all accounts, the powerful, we run the risk of increasing inequity in society. Not just in terms of wealth, which is certainly an issue worth discussing, but also in terms of access to the information being derived by various AIs. How are common people supposed to keep up when knowledge about our behaviour, actions, purchases, interests, beliefs and values are being used to manipulate us? To control our purchases, our information feeds, our attention, our very lives. Something must be done to level the playing field.

[16] Domo, "Data never sleeps 8.0," Domo.com, Domo Inc., January 25, 2020, https://www.domo.com/learn/data-never-sleeps-8

[17] Domo "Data never sleeps 6.0," Domo.com, Domo Inc., 2018 https://www.domo.com/learn/infographic/data-never-sleeps-6

At a bare minimum we should have access to information about specifically how this data is being used to influence us.

Leveling the playing field however is not an easy task. There are many issues that need to be dealt with before we can hope to begin to bring us closer to balance. One of the first is the fact that this data is often privately owned. By agreeing to the licensing agreements (that honestly we don't really have much of a choice about), we have given up our rights to this data (as per the individual agreements). Additionally, we might also be concerned about privacy. No one really wants their neighbor to have access to their search history. This later problem however, is a far more tractable problem. Data can always be anonymized and abstracted to hide individuals within the masses as is commonly done with census data. The real sticky wicket is the ownership issue. Companies will not give up this data without a fight. This data represents real value to these companies and access to these data sets is often sold to third party companies for a variety of reasons. If this data were freely accessible to all it would significantly alter the business model for many companies that specialize in this area. If this is not remedied however, we can expect the knowledge divide in society to grow and eventually this may in turn weaken the functioning of civil society in the future.

One possible solution to this is to consider making companies that supply services that are critical to civil discourse, public utilities and regulate them as such. This would ensure fair and equal access to these platforms that give citizens voice. No one can tell you that you can't have a phone and as well no one censors what you say when you are participating in a phone call. Why should digital communication services be any different? One argument would be that today's digital communication platforms are in the public sphere rather than a private communication between individuals but this simply changes the scope of the communication and who can see it. Currently our approach to this is one of censorship. We disallow those things that we find offensive and label it hate speech, striking any record of it from our collective discourse. To some this is seen as necessary to ensure a peaceful and equitable society, to others this is seen as top down control by those in power to limit personal freedoms. The real question about censorship is not whether we should to it but who is doing it? Who gets to decide what appropriate speech is? If you are in charge of this then I would imagine that you would be quite happy with the rules but others might think of you as intolerant. In my opinion speech should be protected unless it directly leads to action that is prohibited: violence, harassment, etc., otherwise you have to decide what speech is acceptable and what is not and as history has taught us, this is a slippery slope. Once the precedent is set then even if the previous government enacted censorship laws that we consider ethically correct, the next party in power could use this same power to rewrite the laws in their favor and impose restrictions on speech that may not be as ethically centered. Take for instance the case of the National Union of Students who in 1973 got racist speech banned at universities in England. This ban was supported at the time by an organization of Zionist students. For a while this

seemed like a win but a few years later a different group of students was in power at the National Union of Students and they decided ban a Zionist speaker from speaking on campus because they now considered Zionism a form of racism. As you might have imagined, the group of Jewish students likely did not see how this might be turned against them as the leadership of that organization changed over time.[18]

If these companies are considered public utilities, we could also mandate that the proceeds of all analysis (knowledge) of our collective data should be freely available to everyone. This could come in the form of information dissemination and outreach on the part of the companies involved or it could mandate free and equal access to this data for the purposes of analysis. Both approaches have their strengths and weakness but it seems to me that allowing companies to be the sole arbiter of what gets published is a bad idea. If we pursued the later idea then we would need to find ways to make these vast amounts of data available in real time. Additionally, there are numerous barriers to fair and equitable access to this data even if provided freely. Access to sufficient computing hardware and software is required for anyone to begin the process of data analysis of these massive data sets. This is certainly not equitably distributed either. As well the knowledge required to not only conduct such an analysis but to comprehend it as well.

This brings me to my second point regarding risks; access to education and the knowledge that it brings to the individual is critical for individuals to have sufficient skill and training to approach this analysis with rigor and accuracy. To some degree we are far closer to this goal than we are to the goal of equal access to the data itself. Online education has exploded over the years and many topics such as computer programming skills are currently freely available to those that have the inclination to pursue them. This does not mean that they have access to the best and brightest minds on the subject but they do at least have enough access to learn most of what would be required of an AI researcher today. This would allow many more minds to be focused on common problems that we face today but also to potentially uncover new and previously unknown ideas at a far greater rate. It is in this exploratory space that I see this type of citizen science as being most directly applicable. With more minds come more perspectives, potentially allowing us to see a greater degree of the underlying "Truth" of the world. This is certainly in line with the ideas presented earlier related to neurodiversity.

A criticism of this approach might be that there is little control over the preparedness of individuals that seeks to undertake this type of work. However as one can plainly see this has always been the case. Even today not all researchers are considered equal. Some have tremendous knowledge and insights into the complexities of this undertaking and it is highly likely that contributions by these

[18] Ira Glasser, "How Freedom of Speech Protects You from Rulers like Trump," www.thedailybeast.com, The Daily Beast, October 4, 2020, January 26, 2021, https://www.thedailybeast.com/aclu-hero-ira-glasser-on-how-freedom-of-speech-protects-you-from-rulers-like-trump

individuals would be of more import. The solution to this is as it always was. Peer reviewed publication practices can go a long way to maintaining a high standard when it comes to the quality of our collective scientific efforts. However, we also have to aware of the fact that the academic-industrial complex does not have exclusive license to seek the truth. Many minds of great importance do not get the chance in life to contribute to their full potential. Creating a strong program of citizen science, free and open data sources, and access to the knowledge required to pursue such endeavors is paramount for our society to move toward a collective vision of a future where discourse is alive and well, we share that which has the potential to collectively move us forward, and allow all voices to participate in the creation of said future. It is my great hope that we can find new ways to set our collective table in such a way that all leave nourished in mind, body and spirit.

Conclusion

AI is quickly redefining our world and if we continue along our current tack we will likely exacerbate social inequalities and eventually make a less stable world for our children. This is the challenge for the science of AI. Can it mature quickly enough to provide us with insights and abilities that may help us to create a more sustainable and equitable future for all? Climate change and associated global risks are the challenge of our time and human nature is likely the root cause of this dilemma. AI offers us the potential to turn the light of science on our interior nature and the ramifications that this has for our collective future and our future actions within it. Make no mistake, trying to understand our collective behaviour, is the most wicked problem of all. Made even more so because we are the both the cause and solution to the problem. We are on the dance floor of our own perceptions, emotions and thoughts and we need to get on the balcony to be able to see the patterns that are emerging. AI offers us this vantage point. Granted we have a long way to go to improve the science of AI to allow us this potential but it exists none the less. I would hope that we could find a way to act in the collective good. To create a digital world where the rights to participate in society are inviolate, where access to data critical to said discourse is guaranteed, and diversity of perspective is the only requirement for entry.

Social networking companies must begin to think about the world they are allowing us to flow into. We are all on a river of time and the topography that underlies that river is the very nature of our digital (and physical, etc.) world that we have created to date. But just as topography yields to the bulldozer, our digital landscape is ours to remake. Let us start a discourse on this topic. Let all the world's peoples participate. We now have the tools to make this possible. It is an amazing world, but it is also one that must continue to improve if we are to hope to engineer ourselves out of the current environmental trajectory.

AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

Book Review
An Ecology of Communication: Response and Responsibility in an Age of Ecocrisis
By William Homestead
(London: Lexington Books, 2021), 379 pages

Jared B. Rusnak
Independent Scholar
ORCID: 0009-0000-8048-3337

"Anyone paying attention, even if lightly, knows the litany," is how William Homestead opens *An Ecology of Communication: Response and Responsibility in the Age of Ecocrisis*.[1] So, we all know the litany, but can we effectively communicate it? The ecocrisis is not so much the various ecological and social disasters themselves, as it is a crisis of communication surrounding these events. We receive plenty of data, personal accounts, and individual experiences that disclose the dangers of our present culture–nature relationship, but so often our response is characterized by "an emergency room mentality without going to the deeper roots that cause ecological emergencies in the first place."[2] All the while, we cannot shake the feeling that our responses are not fitting.

As a researcher in the field of ecology who is heavily invested in the study of hermeneutics, how ecological principles can be effectively communicated in a way that

[1] William Homestead, *An Ecology of Communication: Response and Responsibility in an Age of Ecocrisis* (Lexington Books: London, 2021).
[2] Homestead, *An Ecology of Communication*, 109.

inspires individual and societal change is of supreme importance to me. In my conversations surrounding the issue, most people I have spoken to across all disciplines agree with the charge, but everyone seems to have a different idea about the address. Furthermore, many agree the ecocrisis is pressing, but believe other matters deserve our immediate attention instead, or that it is not their specialty so it is not their direct concern. Yet, the fact itself that the ecocrisis is thought of as an issue with a discipline for addressing it, separate from other disciplines, is part of the crisis. In reality, ecocrisis is a crisis that demands a response from all fields of study as everything we do is fundamentally tied to the environment. Therefore, since ecocrisis is such a multifaceted, complex issue that knows no disciplinary bounds, it ought to be viewed, discussed, and addressed through the relationships between perspectives rather than any one given perspective at a time. Ecocrisis is fundamentally an issue of poorly kept relations, thus demanding a relational response.

Homestead enters into this conversation with a keen knowledge of and respect for its context. At its surface, *An Ecology Communication* is an impressive body of synthetic scholarship, combining well over 200 individual sources across disciplines. The text has value as an encyclopedia of environmental thinking but reading it solely as that would be doing it and yourself a disservice. Homestead's impressive synthesis directly addresses the crisis of communication surrounding ecocrisis, providing a holistic and relational understanding of communication that is beneficial for conversations surrounding the multi-disciplinary issues of today. The text acts as a common point of conversation between eco-activists, environmental scientists, communication scholars, philosophers, anthropologists, and many more disciplines as it speaks to each of them in a way that encourages them to speak with each other. Thus, Homestead provides guidance to how one could hermeneutically engage with environmental thinking and the interdisciplinary conversations such thinking requires.

The title of the book is indicative to the degree of Homestead's insight into the communicative issue at hand and his conviction to address it in a multidisciplinary manner. Named after the main theory he puts forward, he aptly titles it an *ecology of* communication, as it is in no way a mere reductive *communication of* ecology. Staying true to being an ecology, Homestead explores the relationships between various forms of communication and the insights that can be derived from examining holistic systems rather than narrowing one's focus solely to individual aspects. Furthermore, the subtitle *Response and Responsibility in the Age of Ecocrisis* suggests, rightly so, that the project of exploring communicative relationships is not enough and that such an inquiry demands an exploration of how a knowledge of those relationships better prepares us for the responsibility to respond. While the book falls under the category

of communication scholarship, it is hermeneutic to its core, inviting its reader to consider how re-reading our relationships re-writes our responses.

Both in content and form, the text stays true to its name. From the beginning, Homestead never attempts to discuss a concept in isolation. Rather, he discusses two or more concepts at a time, working through the ideas of each in their relation to one another and an external circumstance or experience. Fitting and unfit aspects of each concept in relation to the circumstance at hand are revealed through this dialogic interplay. Not only is this indicative of Homestead's hermeneutic sensibility, it also provides direct evidence through praxis to one of his central claims: that our communicative strategies surrounding ecocrisis are dominantly monologic when they would be better served by a dialogic engagement. Content wise, Homestead has his finger on the pulse of our communicative problem, namely, that we have been socialized into over-using and over-stressing the importance of a distorted version of rational communication that takes form as an instrumental-calculative monologue, at the expense of dialogic communication. He makes clear that our communicative abilities run deeper than that, arguing for an ecology of communication that has aspects and relationships, which are still underdeveloped in our contemporary society.

While the introduction sets the stage through the context of previous and present environmental activism, provides a summary of the project, and considers possible critiques from John Durham Peters to guide the subsequent discussions, chapters one through four of the text serve to define four communication styles and the relationships between them that build an ecology of communication. Homestead creates a food web of sorts, wherein each of the communicative styles contest and cultivate each other to a degree that the web would collapse without the support of any one part. Each aspect of this ecology, dubbed rational, spiritual, mythic-animistic, and aesthetic communication, is developed through readings of the works of Calvin Schrag, Ken Wilber, Paul Shepard, and Gregory Bateson, respectively. In doing so, Homestead seeks to re-imagine communication by challenging the contemporary primacy of "logic" with a more encompassing logos using transversal rationality; asserting the necessity of playing with and within a transcendental dimension to build imaginative capacity; re-establishing a sense of rootedness through direct tactile experiences in a particular *topos* through acknowledging the subjectivity of all others, including non-humans; and arguing for the importance of allowing insight and inspiration to come through from yielding to the beauty of larger a-historic and atopic patterns that we exist in. Even though a chapter is dedicated to the discussion of each of these aspects of his ecology, Homestead does not let a single chapter go by without including commentary on and from each of the other three aspects, making clear their

inherent interdependency and that they "should not be construed as predetermined or rigid criteria but the flowing of communicative praxis in time and place."[3]

The fifth and sixth chapters address two movements that may be mistakenly conflated with Homestead's ideas, New Ageism and interspecies communication, offering critiques while gathering insights that fit. While covering these two perspectives could have been an opportunity to have easy straw-man examples of unfit responses to ecocrisis, Homestead again shows his tact and instead takes them deeply seriously, unpacking what they have to say in the search of insight. Perhaps unsurprisingly, much of what these perspectives have to say turns out to be unfitting, but what is surprising is the saliency of the insights Homestead is able to pull from these traditions by examining them dialogically through an ecology of communication. From New Ageism, he finds issue with an ungrounded hope, which takes shape as a "create-your-own-reality" principle of hyper-subjectivity, but finds wisdom in the call for a global shift in perspective and a necessity for hope. From interspecies communication, he finds issue with the countless examples of self-projection onto the natural world so that one exists within an echo chamber, while believing the delusion that they are open to more voices, but finds wisdom in the practice's inherent "I–Thou" ontology that perceives non-human others as subjects, from which we may learn and draw insights that may come through.

In his final chapter, Homestead turns to Thoreau, who has been a guiding figure throughout the text, more directly. Thoreau is seen as an exemplary case of practicing an ecology of communication, so that it yields possibly its most desirable result: living in sympathy with intelligence. By investigating what a life of practicing an ecology of communication may look like through the lens of Thoreau's life, we are reminded through the description of practice what the text had explored in theory. His lifelong conscious commitment to improve dedicates him to "[filter] his head through his heart,"[4] allowing him to be purposefully contradictory, approaching the address of each circumstance with fitting responsiveness. This responsiveness naturally led him to a kind of activism and action that was at once suffused with logos; guided by play, imagination, and contemplation; rooted in significant places and particulars; and yielding to the beauty of broader contexts and systems, making them impactful in his time and influential for generations to come. To a degree, the entirety of Homestead's text is a love letter to Thoreau, compelling its reader to open themselves to a wider,

[3] Homestead, *An Ecology of Communication*, 15.
[4] Homestead, *An Ecology of Communication*, 274.

deeper, more reciprocal kind of communication so that we may come to love our rows and beans whilst we tend our fields.[5]

This is all the more apparent in the epilogue, where Homestead returns to the context the book opened in, and, again by exploring past and contemporary environmental activist movements, resoundingly shows the "obvious link between a systematically destroyed biosphere and a systematically distorted communication."[6] Keenly, he leaves the reader with societal level responses that already benefit, and will continue to benefit from, an ecology of communication, such as, amongst others, ecological design, I–Thou science, and deep listening in agriculture. Most compelling, however, is the connection Homestead draws between these responses, necessary social-justice movements, and our individual development, reminding us that "we are called to be responsible for ourselves, but also called to be responsible to each other."[7] It becomes abundantly clear, if it was not already, that while much of the text's work is theoretical, the theory goes that once one opens themselves to an ecology of communication they are inevitably led into a responsibility that demands response to our eco-social circumstance, thus encouraging those around us to practice their underused communicative muscles. As the famous Rilke poem suggests, a work of art demands that you must change your life, likewise, thus, *An Ecology of Communication* does just that while adding another charge: to change your life in such a way that helps others change theirs too.

That is not to say the text is perfect, however, because just like we, just like Thoreau, Homestead is "a human being, not a myth."[8] Although Homestead's dedication to listening from as many sources as possible is in part what makes the work so distinguished what makes the work so distinguished, there are a few subjects, such as scientific studies of telepathy and other "psi-phenomena" that are widely dismissed at first glance in the scientific community, which, despite the insights they offer, may turn the more skeptical reader away. Furthermore, throughout the text there is an occasional usage of outdated scientific language, such as left-brain/right-brain distinctions and the mention of *Homo sapiens*' "reptilian brain," that may further push away scientifically trained readers. Finally, as a suggestion for possible readers, this book has the misfortune of being published slightly before David Graeber and David Wengrow's monumental *The Dawn of Everything*, which, although it has received much

---

[5] Henry David Thoreau, "The Bean-Field," in *Walden and Civil Disobedience* (Vintage Books: New York, 2014), 138.
[6] Homestead, *An Ecology of Communication*, 315.
[7] Homestead, *An Ecology of Communication*, 327.
[8] Homestead, *An Ecology of Communication*, 273.

criticism, may have provided another rich perspective to Homestead's discussion of mythic-animistic communication, making it an excellent companion text.[9]

But these minor quibbles should not deter you; if you are patient with Homestead in the way he is with countless conflicting ideas throughout the text, you may, like he, see past the unfit towards the fit, and find that there is much to learn wherever you look. Homestead's ecology of communication is fascinating, and like all good scholarship, one comes away from it with countless questions of how the world might change as we look at it through this new lens. However, to me at least, what makes this project so endearing and informative for years to come is the borderline "panecastic" sensibility, with which Homestead approaches his inquiry. It is infectious, in the best sort of way. After reading *An Ecology of Communication: Response and Responsibility in an Age of Ecocrisis*, I was left with a new-found wonder towards that which always seemed so familiar. Homestead asks of the reader to have a hermeneutic comportment, in order to see what one may learn from a text they disagree with, a neighbor with whom they barely speak, or even a cement-bound tree on a city street, if they can hone the right kind of communication. Even more gripping, Homestead asks the reader to consider how what they learn will call them to respond. I am convinced Homestead can continue to communicate this wonder to countless others through this book, as long as they are willing to listen.

[9] David Graeber and David Wengrow, *The Dawn of Everything: A New History of Humanity* (Macmillan Publishers: New York, 2021).