Could a Robot Be Conscious? Some Lessons from Philosophy

Markus Gabriel

Contents

Why There Could Not Be Any Conscious Robots	58
The Meaning of Existence	58
The Nature of Consciousness	59
Kinds of Possibility	61
Intelligent Robots?	63
The Human Context	65
Neo-Existentialism	65
Values and the Humanities	66
Concluding Remarks	67
References	67

Abstract

In this chapter, the question whether robots could be conscious is evaluated from a philosophical perspective. The position taken is that the human being is the indispensable locus of ethical discovery. Questions concerning what we ought to do as morally equipped agents subject to normative guidance largely depend on our synchronically and diachronically varying answers to the question of “who we are.” It is argued here, that robots are not conscious and could not be conscious, where consciousness is understood as a systemic feature of the animal-environment relationship. It is suggested, that ethical reflection yields the result that we ought not to produce cerebral organoids implanted in a robotic “body.”

Keywords

Artificial intelligence · Robot · Consciousness · Philosophy · Existentialism · Ethics

Could a Robot be Conscious? The shortest answer to the question posed in my title is: “No.” In what follows, I will lay out some reasons for why we should endorse the shortest answer. At the same time, I will argue that the right way of looking at the issues at stake has significant consequences for our relationship to the digital landscape we inhabit today.

Robots and A.I.-systems created by machine learning experts and research teams play a central role in our “infosphere” (Floridi 2014). Yet, in order to understand that role, it is crucial to update our conception of ourselves, the human being. For, as I will argue, the human being is the indispensable locus of ethical discovery. Questions concerning what we ought to do as morally equipped agents subject to normative guidance largely depend on our synchronically and diachronically varying answers to the question of who we are.

M. Gabriel (✉)
Epistemology, Modern, and Contemporary Philosophy, Bonn
University, Bonn, Germany
e-mail: gabrielm@uni-bonn.de

My paper has two parts. In the first part, I argue that robots (1) are not conscious and (2) could not be conscious if consciousness is what I take it to be: a systemic feature of the animal-environment relationship.¹ In the second part, I will sketch an updated argument for the age-old idea (versions of which can be found in Plato, Aristotle, Kant, Hegel and beyond) that human sociality and, therefore, morality hinges on our capacity to think of ourselves as animals located in a context inhabited by humans and non-human animals alike. This context is grounded in inanimate nature, which presents us with necessary, but not sufficient conditions for consciousness.

Why There Could Not Be Any Conscious Robots

The Meaning of Existence

Ontology is the systematic investigation into the meaning of “existence”. If successful, it leads to knowledge of existence, i.e. the property or properties constitutive of some object’s being there. In a series of books, I have defended the view that to exist means “to appear in a field of sense” (Gabriel 2015a, b). To summarize the outcome of the arguments in this context: there is no system such that every system (except for itself) is a subsystem of that very system. It is impossible for there be a single, all-encompassing field of sense such that every object is part of it. This entails that, necessarily, every object belongs to a specific domain (or field of sense, as I call it), which conditions the field-relative properties that put it in touch with other objects in the same field.

For instance, the Vatican is a legal entity with an impressive history. The Vatican appears in the field of sense of history. There are other legal entities in that field: Europe, Italy, the University of Bonn, refugees, passports, taxes, airports etc. The number 5 appears in a different field of sense, such as the series of natural numbers or that of prime numbers etc. The number 5 does not belong to history; nor is history a mathematical object.

Any view of the form that there is just one overall domain of entities (such as the material-energetic layer of the universe or what have you) is incoherent, as it relies on the inconsistent and paradox-generating notion that there is an all of reality which encompasses everything there is.

Given that there cannot be an all-encompassing domain of objects subject to one set of laws or principles, we are entitled to reject “naturalism” in the sense of a view of the form that all knowledge is natural-scientific knowledge. Natural-scientific knowledge only deals with one domain of objects,

namely the kinds of objects than are part of the material-energetic system of the universe with which we can causally interact.

As a short-cut to this result, one could, of course, simply point out that the very idea of such an all-encompassing domain of natural-scientific enquiry is quite evidently incompatible with mathematical logic and meta-mathematics. We actually know that there cannot be a scientific model (or any other model for that matter) such that absolutely everything (from the early universe to transfinite sets to Angela Merkel) falls within the range of its explanatory power. What is more, there is no such thing as “science” or “natural science” in the sense of a unified theoretical project whose singular terms (such as “boson”, “dark matter”, “molecule”, “neuron”, “glia cell”, “galaxy”, “the universe”, “robot”, “A.I.” or what have you) each and all refer to well-defined entities in a single domain (“reality”, “the universe”, “being”, “the world” or what have you).

Locating an entity in a domain of investigation presupposes a stable insight into what it is. What an entity is, depends on the field(s) of sense in which it makes an appearance. If we drop the field-parameter in our description of how certain entities relate to each other (such as humans and robots, minds and bodies, consciousness and brain-tissue, numbers and countable objects etc.), we will wind up with what Gilbert Ryle famously dubbed a category mistake.²

The idea of “conscious robots” potentially rests on a category mistake. The reason why it is easy to be misled is the following. Humans produce artifacts out of biological and non-biological matter. We build cars, tables, houses, guns, subway systems, smartphones, servers, high-performance computers, statues, etc. Throughout the recorded history of human behavior, we find that humans have produced artifacts, some of which resembled humans and other animals, including cave paintings, statues, etc.

What is equally remarkable is the fact that we find a long-standing desire to produce an artefact in our own image, i.e. something that resembles the feature that we still deem central to human beings: Logos.³ The most recent and somewhat amplified version of this tendency is the idea that we

¹The notion of consciousness is steeped in paradox and, therefore, highly questionable. For more on this see Gabriel (2019, 2017).

²What often goes unnoticed is that the paradigmatic category mistake according to Ryle is precisely a mistake in ontology in the sense deployed here: “It is perfectly proper to say, in one logical tone of voice, that there exist minds and to say, in another logical tone of voice, that there exist bodies. But these expressions do not indicate two different species of existence, for “existence” is not a generic word like “coloured” or “sexed.” They indicate two different senses of “exist,” somewhat as “rising” has different senses in “the tide is rising”, “hopes are rising”, and “the average age of death is rising”. A man would be thought to be making a poor joke who said that three things are now rising, namely the tide, hopes and the average age of death. It would be just as good or bad a joke to say that there exist prime numbers and Wednesdays and public opinions and navies; or that there exist both minds and bodies” (Ryle 1949, p. 23).

³For a critique of this narrative see Smith (2019) and Gabriel (2020a).

might be able to produce intelligent agents, which resemble humans not only in intellectual capacity, but even in shape and movement.

Let us call these kinds of objects *anthropoids*. An *anthropoid* is a robot run by the kind of software nowadays subsumed under the heading of “A.I.” Both robots and machine learning techniques are progressing at a rate that makes it possible for us to imagine robots moving in ways strikingly similar to humans (and other animals). To the extent to which they perform functions we classify as “intelligent” in humans (and other animals), we are likely to be prone to think of them as potential candidates for membership in the “kingdom of ends” (Kant 2016, AA IV, 439),⁴ i.e. in the domain of autonomous and, therefore, moral agents.

However, this is a mistake, as I now want to argue. There is nothing we owe to our artefacts directly. What we owe to our robots is at most a function of what we owe to each other as proprietors of technology. If you destroy my garden robot, you harm me, but you cannot harm my robot in a morally relevant manner, just as you cannot harm a beach by picking up a handful of sand. A beach is a bunch of stones arranged in a certain way due to causal, geological pressures including the behavior of animals in its vicinity. A beach does not have the right kind of organization to be the direct object of moral concern.

Ontologically speaking, robots are like a beach and not like a human, a dog, a bee etc. Robots are just not alive at all: they might be at most (and in the distant future) zombies in the philosophical sense of entities hardly distinguishable from humans on the level of their observable behavior. Analogously, A.I. is not actually intelligent, but only seems to be intelligent in virtue of the projection of human intelligence onto the human-machine interface.

Here is a series of arguments for this view.

The Nature of Consciousness

Let us begin with the troublesome concept of *consciousness*. Consciousness is a process we can know only in virtue of having or rather being it. I know that I am a conscious thinker in virtue of being one. To be conscious is to be a state that is potentially self-transparent insofar as its existence is concerned.⁵ This does not mean that we can know everything about consciousness simply in virtue of being conscious. This is obviously false and, by the way, has never been maintained

by anyone, not even Descartes who is often discredited in this regard for no good reason.⁶ Both right now, as I am awake, and in certain dream states, I am aware of the fact that I am aware of something. This feature of self-awareness is consciousness of consciousness, i.e. self-consciousness. Some element of consciousness or other always goes unnoticed. As I am right now conscious of my consciousness, I can focus, for instance, on the structure of my subjective visual field only to realize that many processes in my environment are only subliminally available to conscious processing, which means that I am conscious of them without thereby being conscious of that very consciousness.

Trivially, not all consciousness is self-consciousness, as this leads into a vicious infinite regress. If all consciousness were self-consciousness, then either self-consciousness is consciousness or it is not. If self-consciousness is consciousness, there is a consciousness of self-consciousness and so on ad infinitum. We know from our own case that we are not conscious of anything without a suitable nervous system embedded in an organism. We know from neurobiology and human physiology that we are finite animals such that it is evidently impossible for us to be in infinitary states where each token of consciousness is infinitely many tokens of self-consciousness. There simply is not enough space in my organism for that many actual operations of self-reference.⁷

As a matter of fact, we do not know what, if anything, is the minimal neural correlate of consciousness (the MNCC). Recent contenders for a model designed to begin to answer this question despite the complexity involved in the neuroscientific endeavor to pinpoint such a correlate include “global workspace theory” (GWT) and “integrated information theory” (IIT).⁸

Whatever the right answer to the question concerning the MNCC will turn out to be (if it is even a meaningful

⁴Kant’s famous phrase for the domain within which moral agents move. See AA IV, 439.

⁵On this well-known aspect of consciousness famously highlighted (but by no means first discovered) by Descartes, see the concept of “intrinsic existence” in the framework of Integrated Information Theory, one of the currently proposed neuroscientific models for the neural correlate of consciousness. See the exposition of the theory in Koch (2019).

⁶For a detailed historical argument according to which Descartes does not even have the concept of consciousness that he is often criticized for introducing, see Hennig (2007). For an outstanding philosophical reconstruction of Descartes’ account of human mindedness that takes into account that his views are actually incompatible with the dualism typically ascribed to him see Rometsch (2018).

⁷The situation is different if we have a rational soul in the traditional sense of the term introduced by Plato and Aristotle and handed down to us via medieval philosophy. We should not naively discard this traditional option on the ground that we mistakenly pride ourselves for knowing that we are animals, because no one in the tradition denied this! The widespread idea that we began to realize that humans are animals in the wake of Darwin is unscientific and ignorant story-telling. Short proof: ζῷον λόγον ἔχον, *animal* rationale. It should be obvious that Plato and Aristotle, the inventors of logics, were able to accept the following syllogism: (P1) All humans are rational animals. (P2) All rational animals are animals. (C) All humans are animals.

⁸For a recent overview over the main contenders for the right model of the neural correlate of consciousness see Block (2019). For a critique of the concept of consciousness deployed by Dehaene et al. (2017), see Gabriel (2020b, §§9 and 15).

question), it has to respect the following *indispensability thesis*: the reality of the human standpoint is indispensable for any scientific account of consciousness.⁹ There is no third-person point of view, no purely observational stance, such that we can distinguish between conscious and non-conscious entities/processes in the universe. Scientific knowledge-acquisition at some point or other presupposes a conscious knowledge-claim maintained and defended by a human thinker or group of human thinkers. For, scientific knowledge is a paradigm case of truth-apt justified belief. To make a (scientific) knowledge claim concerning a bit of reality means that one has reasons to believe that the bit under investigation has some of the central properties ascribed to it by a model. (Scientific) knowledge claims are not blind guesses. They are highly methodologically controlled outcomes of human activity which will always make use of some technology or other (including pencil and paper; conferences; fMRI; the LHC etc.). (Scientific) knowledge claims do not simply emerge from anonymous activity, they are high-level performances of human beings, often making use of non-human machinery.

Arguably, there are various potential philosophical confusions built into the very idea of searching for the MNCC, as there are many historically shifting meanings of the word “consciousness” (which differ, of course, between various natural languages dealing with the kind of phenomena grouped together by candidate meanings of “consciousness” in contemporary Anglophone philosophy and mind science).

To begin with we ought not to lose track of the distinction between “narrow” and “wide contents” of consciousness.¹⁰ Consciousness has both an object and a content. The *object of consciousness* is that which it concerns, for instance, the Eiffel tower, if I look at it, or some of my internal states, when I feel hungry, say. Actually, whenever I am conscious of anything in my environment (such as the Eiffel tower), I am at the same time also conscious of some internal states of my organism. Typically, we are never conscious of just one

thing alone.¹¹ Consciousness is of objects in a field.¹² Consciousness itself is a field which encompasses subfields. I can be conscious of a recently deceased friend, say, which means that I can have visual or olfactory memories of his presence. In this scenario, I am conscious of a past conscious episode relating to my friend and not directly of my friend. Memories are not perceptions despite the fact that they involve percepts.

The *content of consciousness* is the way in which the object appears to me. Sensory modality is, therefore, part of the content, as is perspective. In general, consciousness has an ego-central index: I am here and now conscious of a scenario (a dynamic field of processes and objects) in a bunch of sensory modalities (Burge 2013). Notice that I cannot be conscious of any scenario without myself being part of that scenario. I am here right now as the entire animal I am. It is impossible to perceive anything without literally being a part of the same field(s) as the objects, including the various physical fields whose interaction is required as a medium for the production of mental content.

Narrow content emerges in the context of internal self-awareness of my organism. It deals with states I am in. Pain, for instance, or the color sensations I experience when I close my eyes, have narrow content. They deal with internal states of my organism. Narrow content is a series of perspectives of the organism upon itself. Narrow content is, as it were, an internal window onto processes within the animal I am.

Wide content emerges in a systemic context which includes states of affairs beyond my ectodermic limitations. If I see a table, the table is nowhere to be found in my organism. My organism is evidently too small to encompass all objects of perception. St. Peter’s Basilica is bigger than my organism. If I see it, it cannot be “in me”. It is simply nonsensical to believe that there is no external world in which St. Peter’s can be found on the dubious ground that it is allegedly impossible to directly perceive reality. Wide perceptual content deals directly with external reality. Yet, it does so in a species- and individual-relative way. St. Peter’s looks different to you and

⁹This important methodological and ontological principle has recently been violated, for instance, by Hoffman (2019). If our perceptual system were constitutively out of touch with reality, how could we know this by deploying scientific methods which presuppose that our modes of information-processing in a laboratory are not only contingently reliable detectors of an unknowable thing in itself, but rather the correct instruments to figure out how things really are? The scientist who tries to make a case that all perception as such is a form of illusion cannot coherently establish this result, as she will make use of her perceptual apparatus in making her blatantly self-contradictory statement.

¹⁰On this distinction see the classical paper by Block (1986).

¹¹Another important exception here are mystical experiences such as the unity with the One described by Plotinus or the *visio beatifica* known from the Christian tradition. Similar accounts can be found in any of the other major world religions. Again, we should not simply discard these accounts, which would make us incredibly ignorant vis-à-vis the very genealogy of the idea of “consciousness” which (like it or not) originates from that tradition.

¹²Consciousness is one field of sense among others. Not all fields of sense are conscious or related to consciousness. The consciousness-field and its objects are arguably entangled. In any event, there is a strong correlation between consciousness and its objects which does not entail without further premises that the objects of consciousness are necessarily correlated with consciousness. I reject the kind of premises that typically lead to the position that the objects of consciousness would not have existed, had we not been conscious of them. For a recent exchange on this issue, see Meillassoux (2006) and the highly sophisticated response from the point of view of philosophy (of quantum mechanics) by Bitbol (2019).

me and be it for the trivial reason that we will never strictly speaking occupy the same spatio-temporal location.¹³

Both wide and narrow content owe their specific structure in human (and non-human) conscious animals to, among other things, evolutionary parameters. This is a fact about animals. Consciousness is part of our adaption to our ecological niche, which in turn is causally shaped by our self-conscious adaption to that adaption guided by scientific and technological progress. If anything, we therefore have very strong evidence for the general hypothesis that consciousness is a biological phenomenon.

Kinds of Possibility

The next step in my argument consists in delivering the premise that we simply cannot reproduce the natural, necessary conditions for the existence of consciousness in the sense of anything like a full ego-centric index. One reason for this is that we are astronomically far away from knowing the necessary (and jointly sufficient) conditions for human (and non-human) consciousness in the full sense, where “the full sense” encompasses both wide and narrow content. We do not know enough about the causal architecture of consciousness as a natural phenomenon in order to even begin constructing potentially conscious robots. Therefore, if any actually existing robot is anywhere near consciousness, this would be a sheer coincidence. It is as rational to believe that any actually existing robot or computer is conscious as that the Milky Way or a sandstorm in the Atacama Desert is conscious.¹⁴

To be sure, it is apparently at least logically possible that a given robot is conscious. But this does not entail that we have any actual reason to believe that a given robot is conscious (i.e. if it’s being logically possible just means we cannot say we know it is not the case). It is, thus, irrational and unscientific to believe that currently existing robots are conscious.

At this point, someone might wonder if this train of thought rules out that robots could be conscious. So far, I seem not to have established that no robot *could* ever be conscious. At this stage, we need to be careful so that our modal “intuitions” do not idle. The central modalities are: actuality, possibility, contingency, and necessity. If we ask the question “could robots be conscious?” we are after a possibility. Is it possible that robots are conscious? So far, I have argued that

there is no reason to think that any currently existing robot is conscious. A robot is an artifact of human industry and so far, no relevant robot in that sense is conscious.¹⁵ If we want to consider the possibility of robot consciousness, we therefore need either some evidence or argument that supports the rationality of the belief that *future* robots could meet the relevant threshold of consciousness. Otherwise we wind up with what I would like to call *extremely weak possibility*.

A possibility is extremely weak if (1) nothing we actually know supports it and (2) we have no reason to believe that the possibility amounts to a logical contradiction. Conscious robots are currently extremely weakly possible. What is more, I now want to argue that they are currently at most extremely weakly possible. The argument relies on the notion of “biological externalism,” which I have proposed in my recent book *The Meaning of Thought* (Gabriel 2020a).¹⁶

In philosophy, *semantic externalism* is roughly the view that some terms refer to some natural kinds in such a way that a competent user of those terms need not know all essential feature of the kinds in order to count as a competent speaker.¹⁷ For instance, I am a competent user of the term “fermion,” but not an expert user. A professional nuclear physicist will use the term “fermion” in contexts where I would be likely to make nonsensical utterances. One strength of semantic externalism in general is that it gives an account of the fact that we often speak about things in a competent way without thereby knowing their essence. The standard example in philosophy is use of the term “water”. When Thales or Aristotle referred to the stuff in the Aegean Sea as ὕδωρ, they were thereby referring to something that essentially involves H₂O molecules. However, if we asked them about ὕδωρ, they could not even entertain the thought that it essentially involves H₂O molecules, because there was no such thought available to them in their linguistic community. From the epistemic standpoint of the Ancient Greeks, it would have looked possible that water could consist

¹³For a paradigmatic exposition of a direct realism which takes objective looks as relations between a perceiver and the perceived environment into account see the discussion in Campbell and Cassam (2014).

¹⁴On the recent resurgence of panpsychism in the sense of the view that basically everything is (or might be) conscious, see Harris (2019) and Goff (2019). I assume that it is reason to reject a given account of the meaning of “consciousness” if it either entails the truth of or significantly raises the probability of panpsychism.

¹⁵Of course, humans (and some non-human animals) are artifacts of human activity. As Aristotle nicely puts it: ἄνθρωπος ἄνθρωπον γεννᾷ (Aristotle, *Metaphysics*, VII 71032a 25 et passim). However, we do not classify humans as robots. If humans count as robots, machines, A.I.s or computers, it would be trivially true that robots, machines, A.I.s and computers are conscious because we are. The question “could robots be conscious?” deals exclusively with the concept of a robot as a non-biological (possibly anthropoid) artifact of human industry.

¹⁶This is a revised version of Gabriel (2018a). See also Gabriel (2020b) §§6–11, where I defend “mental realism,” i.e. the view that mental terms (including: thinking, intelligence, consciousness) refer to processes which have necessary biological preconditions. Notice that the argument does not preclude the existence of future conscious robots controlled by cerebral organoids. Such hybrid entities are more likely to be possible than in silico conscious robots.

¹⁷For an overview see Rowlands (2014).

of various arrangements of elementary particles (atoms) in their sense. It would not have been irrational for Aristotle to assert that water consists of some Platonic atomic figure or other, as laid out in the *Timaios*. For him, that could be the case. Yet, Aristotle would have made a mistake, had he endorsed a view about water that rules out that water essentially involves H₂O molecules.

As far as consciousness is concerned, we are in a similar epistemic situation. We do not know which natural kinds, if any, are correlated with consciousness. However, we must not forget that the mainstream of contemporary scientifically minded theorists of consciousness tends to believe that consciousness has some necessary natural prerequisites or other. Unfortunately, this misleads far too many theorists into some version of “functionalism”. It is certainly true that the biological prerequisites for conscious states significantly vary across species, across individuals within a species and even within an individual across time. In this sense, the physical underpinning of consciousness is multiply realizable in different structures. Yet, this factual variation in the (neural?) support of consciousness, of course, does not per se support the stronger claim that consciousness is realizable in inanimate matter or any matter that preserves the functional architecture correlated with consciousness in animals.

In this context, let *functionalism* be the view that consciousness is identical to a process which consists in the realization of a role in a system of sensory inputs and behavioral outputs. Most functionalists are committed to *multiple realizability*. According to this concept, the functional role of consciousness can be realized in different materials. We humans realize it in (neural?) tissue, other creatures might realize it in some other tissue. If consciousness is multiply realizable, it seems to be possible to produce it out of material other than biological tissue. This possibility is stronger than the extremely weak possibility that there could simply be robots. Functionalism to some extent supports the hypothesis that robots could be conscious. However, functionalism combined with multiple realizability is in serious trouble, as is well-known in the philosophical community, but often ignored by the interested bystander.¹⁸ The major weakness is a consequence of the fact that we currently do not even know the MNNC. For all we know, if consciousness is identical to a functional role, this role could be performed by the universe as a whole or by some surprising subsystem of it (such as a galaxy cluster or a beach). This explains the presence of panpsychism in contemporary philosophy, i.e. the (misguided) notion that consciousness might be everywhere.¹⁹

¹⁸See especially: Block (1978) and Searle (1992). An example of a cautionary tale of philosophical confusions concerning the alleged substrate-independence of life and intelligence is Tegmark (2017).

¹⁹Notice that IIT entails that panpsychism is false. See Tononi and Koch (2014). IIT’s model for consciousness provides us with a defeasible criterion for the presence of consciousness in a system. In its current

Functionalism tends to lead to acceptance of the notion that consciousness could emerge out (or rather: be a property) of basically any system whose parts are arranged in such a way that we can describe their operations in terms of sensorial input and behavioral output.

According to biological externalism, “consciousness” and cognate terms in our mentalistic vocabulary refer to processes which have necessary biological prerequisites. Thus, there could be no conscious robot produced entirely out of inanimate matter. Support for biological externalism can be derived from our current knowledge base concerning consciousness. There should be no doubt that ever since humans have been in the business of thinking about their own mental states and those of other animals, they have been thinking about something with some biological underpinning. At least, this is not what is contentious between the functionalist and her opponent.²⁰ Thus, as far as we know, consciousness has essentially biological preconditions. This does not mean that consciousness is a purely biological product, as I will now argue.

Biological externalism alone does not suffice to rule out future robots controlled by cerebral organoids. However, I surmise that such robots are biologically impossible. This claim is empirical, but partially grounded in a further maneuver of philosophical theorizing. Above, I mentioned an indispensability thesis. As far as consciousness is concerned, the indispensability of consciousness consists in the fact that we cannot circumvent it in any complete account of human mindedness. Scientists are conscious and, therefore, consciously interested in consciousness for various reasons (including ethical considerations, because we believe that we ought to care for conscious creatures more than for entirely non-conscious matter). However, the level of indispensability is located on a scale which differs from that occupied by neural tissue alone. Any theorist we have encountered so far, has been a full-blooded human being with animal parts that include many cell types other than neurons. Human animals are not neural tissue implanted in an organism, as it were. My skin is not just a bag containing a central nervous system hidden from view by a more complex organization. Simply put: I am not a brain.²¹ Human neural tissue is produced by a human organism out of stem cells in complex biological processes. It develops over the course of pregnancy in such

state, it rules out that beaches could be conscious. It also rules out that computers and robots could be conscious.

²⁰There is the additional difficulty that humans have been thinking about divine thinking for long stretches of our history harking back as far as the first written documents of humanity. This supports the notion that there could be divine non-biological thought and consciousness. However, it does not back up the idea of finite, conscious robots produced by human industry out of non-biological matter.

²¹For a detailed defense of a non-reductive account of human thinkers see Gabriel (2017).

a way that at some (as yet unknown) point a human fetus becomes conscious. No known neural tissue outside of an organism has the property of consciousness. This is probably not a coincidence, as consciousness is a product of processes that can be studied by evolutionary biology. All cell types came into existence in this way. Neural tissue comes into existence in causal contexts that produce organisms. Let us call the structure of this process *systemic organization*.²² As far as I know, no sample of neural tissue that even comes close to being a candidate for a correlate of anything mental has been found outside of systemic organization.

The absolute majority of actual conscious states we know of has both objects and content. Consciousness is usually about something in some particular mode of presentation or other. Without integration into an organism, it is quite senseless to think of neural tissue as performing any function that correlates with consciousness, as we know it. Should it be possible to produce cerebral organoids that are in conscious states, those states would at most resemble a tiny fraction of a proper subset of our conscious states. No organized heap of neural tissue will perceive anything in its environment without proper sense organs. To perceive our environment is precisely not a kind of hallucination triggered by otherwise unknowable external causes. Such a view—which I attack under the heading of “constructivism”—is profoundly incoherent, as it amounts to the idea that we cannot ever really know anything about our environment as it really is. This makes it impossible to know that there is neural tissue sealed off from an external environment so that the view that we are literally “brains in a vat,” i.e. neural tissue hidden under our skulls, is utterly incoherent (Putnam 1981). External reality as a whole cannot be a kind of hallucination or user illusion produced by a subsystem of the central nervous system. For, if it were, we could not know this alleged fact by studying the central nervous system, because the central nervous system itself belongs to external reality. Thus, the central nervous system is not a hallucination by the central nervous system. If we know anything about ourselves as animals capable of perception, we thereby know that we can know (parts of) external reality.

Here, we can use a famous thought-experiment by Donald Davidson as a conceptual magnifying glass. In a forthcoming paper I have co-authored with the cosmologist George F. R. Ellis, we use this thought-experiment in order to illustrate the concept of top-down causation²³ (Gabriel and Ellis 2020, forthcoming). Davidson asks us to imagine that lightning

strikes a dead tree in a swamp while Davidson is standing nearby. As Davidson’s body dissolves due to the causal circumstances, the tree’s molecules by coincidence turn into a replica of his body which begins to behave like Davidson, moves into his house, writes articles in his name, etc. (Davidson 1987). We maintain that Swampman is physically impossible. No molecule by molecule duplicate of a person could arise spontaneously from inanimate matter. The evolutionary pre-history and the adaptation of an organism to its causally local environment (its niche) are essential for the organism’s existence. To the extent to which we could possibly recreate the conditions of survival for neural tissue complex enough to be a candidate for a token of the MNNC, our social activity of producing those conditions and artificially maintaining them in existence would at best replace the structure of organization. Thus, any robot actually capable of being run by “mental,” i.e. actually conscious software would have to have the relevant biological hardware embedded in a context which plays the role of an organism.

The organism controls local patterns of causation in a top-down way. The organism is thus ontologically prior to the causal order of its elements (Noble 2016). If we randomly copied the order of an organism’s elements, we would still not have copied the organism. To be more precise, we would have to copy the causal order of an organism’s elements in the right way in order for a Swampman to be alive, which means that the contextual, social constraints on his production set the conditions for the lower-level elements to realize Swampman. Random physical structure is not enough for Swampman to be so much as alive for any amount of time. Hence, there could not be a Swampman replica of Davidson. Our use of the material of the thought experiment is supposed to illustrate that evolutionary causal history, including an organism’s niche construction and its social contexts, is essential for the causal constitution of conscious life and thought. Even if we could replicate human organisms by some hitherto unavailable procedure, this would not be evidence for a bottom-up process, as the relevant causal context would, of course, include us and the technical apparatus needed in order to achieve the feat of bringing organic matter into shape.

Intelligent Robots?

One line of argument for the possibility of conscious robots draws on the notion of artificial consciousness and assimilates this discussion to that of AI. Yet, this is a red herring, as the term “intelligence” generates confusions similar to those associated with “consciousness”.²⁴ In general, *intelligence*

²²For an ecological account of consciousness see Fuchs (2018).

²³M. Gabriel/G. Ellis, “Physical, Logical, and Mental Top-Down Effects,” in: M. Gabriel/J. Voosholz (eds.), *Top Down Causation and Emergence*. 2020 (forthcoming). The thought experiment can be found in D. Davidson, “Knowing One’s Own Mind”, in: *Proceedings and Addresses of the American Philosophical Association* 60 (1987), pp. 441–458.

²⁴On some of the confusions in artificial consciousness debates see Schneider (2019).

can be seen as the capacity to solve a given problem in a finite amount of time. Let us call this concept “undemanding intelligence”. It can be measured by constraining the exercise of the capacity to time parameters. In this light, a system S^* is more intelligent than a System S if it is more efficacious in solving a problem, i.e. if it finds a solution quicker. *Learning* is the process of replacing a given first-order *object-problem* by another higher-order *meta-problem* in such a way that finding solutions to the object-problem enhances the capacity of finding solutions to the meta-problem. A *standard artifact* is a non-biological product of human industry. Usually, a standard artifact is associated with a human goal-structure. In the case of modern technology, the human goal-structure is essentially tied to a division of labor. The division of labor of modern technology is too complex for any individual to know how each and every participating individual contributes to the production of the outcome. The management structure of our production of material goods (including the hardware required for any actually functioning robot) functions by producing meta-problems handed down in the form of object-problems to agents on the ground-floor of production. Without this socially immensely complex structure of the production of systems capable of justifying scientific knowledge-claims, there would be no robots.

More specifically, no AI-system has the property of intelligence outside of the top-down context realized at the human-machine-interface. No standard artifact (which includes software qua program-structure) has any degree of intelligence outside of a human use-context. AI essentially differs from human, animal intelligence for a simple reason: the parameters of our goal-structure are fundamentally set by our survival form. Intelligence first and foremost arises in the context of solving the central maintenance (survival) problem of human, animal organization. Animals have interests which in turn serve the goal of maintaining them in existence. This goal-structure came into being in the universe as a consequence of as yet not fully understood processes. We have no final answer to the question of the origin of life. Yet, whatever the actual causal context for the emergence of life, it is the breeding ground of intelligence.

The term “intelligence” derives from the Latin *intelligere/intelligentia* which means “understanding”. We should distinguish intelligence in the traditional sense from undemanding intelligence. Accordingly, we can introduce the notion of *demanding intelligence* (D.I.). In his classic, *The Emperor’s New Mind*, Sir Roger Penrose has shown that D.I. is not a matter of explicable rule-following (Penrose 1989). D.I. consists in finding a new solution space to an inherited problem by discovering an entirely new meta-problem. This requires the capacity for understanding oneself

as a creative thinker engaged in an activity of thinking that cannot be formalized at all. In this context, I have recently argued that our AI/machine learning programs amount at best to thought-models (Gabriel 2020a). Thought-models can be very powerful tools. Think of everyday modern technological products such as search engines, which serve the function of mining data by means of a formal representation of a mode of organizing potential material for thought. The internet significantly contributes to our cognitive enhancement in that it provides us with quick solutions to given problems so that we can use our mental time more efficiently. By deploying thought-models as instruments in our own struggle for survival and progress, we, humans, become more intelligent in that we create the potential for new modes of thinking. If anything, our digital technology produces conditions of emergence for intelligent human behavior. We can make intelligent use of our technology, and we should begin to realize that this does not at all entail that our technology is intelligent by itself.

D.I. is the capacity to change a problem space in virtue of an account of our activity of creating and changing problem spaces. In classical philosophical parlance, D.I. is self-consciousness or self-awareness: we, human beings, become aware of the fact that we are intelligent animals. In the context of exercises of that awareness we can produce thought-models designed to re-produce elements of our thought-activity in a simplified way. It is in the nature of a model of something to reduce the complexity of a target system. Models are modes of abstraction. They distinguish between an essential and an inessential feature of a target system relative to a goal-structure. A scientific model, such as the contemporary standard model of particle physics, is not a copy of physical reality, but a mode of abstracting away from levels of the universe we inhabit. It is crucial for the standard model that it does not mention the scientists who produced it in the quest for understanding the universe, precisely because scientists and their actual thoughts do not appear in the standard model. Scientists are not a bunch of elementary particles. The idea that scientists are ultimately reducible, i.e. logically replaceable by a bunch of elementary particles arranged in the right way, is a terrible confusion of model and reality. For more on this see Ellis (2016).

Analogously, the notion that human thinking is a rule-governed process exactly like that to be found in a Turing machine (or any other model of information-processing) conflates a model with the reality it is designed to make more intelligible to human thinkers. If we abstract away from the context we actually occupy as human thinkers, it should not be a surprise that we cannot recover our own minds from observing the behavior of our artifacts.

The Human Context

Neo-Existentialism

Human beings are sapient creatures. When Linnaeus suggested “homo sapiens” as the name for our species, he was fully aware of the fact that human beings fundamentally relate to themselves in a specific way. This is why he defines the human being in terms of our capacity for self-knowledge: “nosce te ipsum” (Linnaeus 1792). In this context, humans produce models of themselves. The German word for this is “Menschenbild”, which means “image of humankind”. A conception of man is an image, a model of what we think we are. Evidently, there is a variety of such images. Some believe that they have an immortal soul which is the locus of their humanity and dignity. Others think of themselves as sophisticated killer apes whose goal is to spread their genes. Whatever the right answer to the question of who or what we are as human beings, it must consider the remarkable fact that there is a range of answers to that question in the first place.

In this context, I propose a framework for the study of the human context I call “Neo-Existentialism” (Gabriel 2018b). Neo-Existentialism offers an account of what it is to be human, an account of humanity. *On this account, to be human is to instantiate the capacity to think of oneself as an agent of a certain kind and to (sometimes) act in light of that conception.* We can think of this as *higher-order anthropology*. The capacity to instantiate humanity in ways that differ synchronically and diachronically across individuals and populations does not itself differ synchronically and diachronically across individuals and populations.

Neo-Existentialism differs from many forms of classical existentialism in that it draws on a potentially unified conception of humans as both objects of natural science, medicine, etc. and subjects of truth-apt, historically variable self-conceptions that are the target of the humanities and social sciences. It thus bridges the perceived gap between the natural sciences and the humanities by locating scientific knowledge-acquisition in the human context.

Among other things, it has the advantage of offering a solution to the so-called mind-body problem that is designed to bring all academic disciplines to the same table, the one we all sit at in virtue of our humanity. In the philosophy of mind, neo-existentialism argues that there is no single phenomenon or reality corresponding to the ultimately very messy umbrella term “the mind”. Rather, the phenomena typically grouped together under this heading are located on a spectrum ranging from the (by now) obviously physical to the non-existing. However, what does unify the various phenomena subsumed under the messy concept of “the mind” is that they result from the attempt of the human being to distinguish itself both from the purely physical universe and from the rest of the animal kingdom. In so doing, our self-

portrait as specifically minded creatures evolved in light of our equally varying accounts of what it is for non-human being to exist.

Being a German-speaking philosopher, I suggest that we replace the messy term “mind” by what we call “Geist” in my neck of the woods. *Geist* is not a natural kind or complicated structure of natural kinds, but precisely something that does not exist independently of the specific descriptions used in order to point out phenomena whose very existence depends on mutual ascriptions of tokens of mental states such that their accuracy-conditions presuppose anchoring both in the external natural world and a linguistic division of labor. “Geist” is what you get under conditions of mutual action explanation in a context where you cannot delegate the vocabulary by which you conceive of yourself and others as human to a neutral, natural-scientific standpoint.

To look at reality from the standpoint of a human being (*Geist*), means that we produce thought-models in the context of the human life form. There is no way to circumvent this. This is why Neo-Existentialism rejects Daniel Dennett’s influential distinction between the *physical*, the *design* and the *intentional* stance (Dennett 1987, 2017). There is no physical stance capable of dealing with the human being if this stance abstracts from the fact that the scientist is a human being endowed with a mind suitable for making knowledge-claims etc. The physical stance simply evaporates if we try to think of it entirely independently from the intentional stance. Dennett’s mistake consists in thinking of the intentional stance as a kind of model or theory of human agency which serves the function of a useful illusion. According to contemporary philosophical classification systems, his view is a form of “mental fictionalism” according to which attributing mental states such as “consciousness” to an entity such as a human animal is literally false, but useful.²⁵

The starting point of Neo-Existentialism’s framework is the observation that human agents (sometimes) act in light of a concept of who/what they are. A striking example for this would be the difference between someone who does what she does in virtue of her belief that she has an immortal soul whose morality is tested during her earthly life by a transcendent God on the one hand and, on the other, someone who believes that she is a sophisticated killer ape without any homuncular control center; a complex biological cellular network whose goal is maintenance in the form of survival and the spreading of her genes via procreation. There are many other forms of actually being human, or realizing one’s own humanity by acting in light of a shared conception of what humanity is. The humanities, such as anthropology, religious studies, the sociology of knowledge, etc. investigate such

²⁵For a discussion of “mental fictionalism” see the various contributions in *The Monist* Vol. 96, no. 4 (2013).

ways of being human in their specific mode of institutional, historical, etc. manifestation.

In this context, Neo-Existentialism distinguishes between *two kinds of error*: error about a natural kind vs. error about oneself as an agent. From the standpoint of *the epistemic conception of reality*,²⁶ we can define a “natural kind” as a type of object that it is exactly what it is regardless of the truth or falsity of our attitude. Electrons or supernovae are candidates for natural kinds, as they have their properties no matter what anyone believes about them. At some level or other, the natural sciences discover properties of natural kinds even though they cannot be reduced to this feature, because they sometimes create new objects and, therefore, bring properties into being that are a function of their theories. This is what happens in material science or in a particle accelerator which is capable of generating new particles. Science is not just a list of natural kinds and their properties, a table of elements.

We can distinguish “natural kinds” from what Ian Hacking has helpfully labelled “interactive kinds” (Hacking 1999, 2002). Specifically, humans are interactive kinds in virtue of the fact that it matters for who we are how we conceive of ourselves. My favorite example is someone who thinks that he is a talented Tango dancer, but in reality, can hardly dance at all. This person might lead a deluded life to the extent to which his integration into a group can severely suffer from his wrong beliefs about himself. Wrong beliefs about myself—including paradigmatically: wrong beliefs about my Self—can change my properties. If I have wrong beliefs, I am in a different state than if I have true beliefs. Thus, beliefs matter for who and what we are. Our wrong beliefs can guide our actions. The deluded tango dancer acts in light of a misguided (partially false) conception of himself so that a feedback loop between wrong beliefs and action comes into existence.

Some proper subset or other of our mentalistic vocabulary is such that it comprises elements that do not refer to natural kinds. This does not rule out a priori that there is another proper subset of the same vocabulary that happens to pick out natural kinds. Vigilance, various urges we consciously experience, and maybe phenomenal consciousness (what-it-is-likeness) as a whole belong to this category. As things stand, our mentalistic vocabulary is differentiated both diachronically and synchronically over different natural languages and specialized idiolects. It is not unified in any specific way

beyond the fact that we typically invoke it in contexts where action explanation, including activities such as predicting or regulating future behavior, matters. But this is precisely a manifestation of “Geist”. As long as humans interact in an institutionalized form of any kind, the game of mutual action-explanation and attitude adjustment to the observed and extrapolated actions of others will go on and produce new vocabularies and situations. Monarchies, right- and left-wing populism, neurosis, credit cards, fear of the Gods, love of wisdom, class struggle, ideology, moral righteousness, and indefinitely many other facets of human reality will never be replaced by a unified, centralized committee of neuroscientific Newspeak headed by some eliminative materialist or other.

Neo-Existentialism is not a relapse into metaphysical dualism according to which there are exactly two kinds of objects in the universe: material and mental substances. That would only lead us back to the unsolvable mystery of their interaction. Mental causation is real in that tokens of the types picked out by the relevant proper subset of our mentalistic vocabulary that makes us *geistig*, are integrated into a meshwork of necessary and jointly sufficient conditions. This meshwork essentially involves natural conditions, such as nervous systems embedded in healthy organisms etc. There is no overall privileged grounding relation running through all token meshworks. Any actual token of the meshwork, any of its states, can take any form out of a huge, epistemically indefinite and historically open set of possible ways of being human. We continue to generate new ways of being human without there being any a priori catalogue. This is the sense in which humans do not have an essence: there is no surveyable totality of modes of realizing humanity.

Values and the Humanities

(Moral) values are grounded in the universal form of being human. The universal form of being human consists in our capacity to lead a life in light of a conception of the human being and its place in animate and inanimate nature. Our anthropological self-conception cannot be exhaustively studied by the natural sciences. The kinds of complexity involved in high-level human social systems, the dynamics of historical developments, the plurality and history of human languages, art forms, religion etc. cannot seriously be reduced to the models of explanation characteristic of natural-scientific knowledge-acquisition.

The humanities remind us of our own humanity. This is one of their crucial roles in modernity. Natural science will not survive the materialist attacks on the humanities that, among other things, are outcomes of a misguided scientific worldview according to which all (real) knowledge is natural-scientific knowledge. It is simply false that all there is the

²⁶The epistemic conception of reality is the notion that to be real means to be the target of fallible belief. This conception of reality is broad enough to encompass immaterial objects, such as mathematical objects, (possibly) consciousness and laws of nature and to think of them as real. Reality cannot be reduced to the material-energetic layer of the physical universe. As a matter of fact, this is a lesson from modern physics itself, an insight we owe to quantum theory, which has ultimately superseded the naïve conception of “matter” and “causation” as a series of “micro-bangings” of “atoms in the void”. For more on this see Ladyman and Ross (2007), Ellis (2016), Ismael (2016), and Falkenburg (2007, 2012).

material-energetic layer of the physical universe. We know that this is false from the various humanities and social sciences, which clearly study objects and processes that are by no means identical to objects and processes studied by any combination of actually existing disciplines from the range of the natural sciences.

The materialist version of the scientific worldview is an ideological distortion of scientific activity easily exploited by groups whose interest lies in pinning down an alleged specific essence of the human being, such as the false idea that the human self is identical to the brain or some other subsystem of the nervous system (Gabriel 2018c). If we are sophisticated killer apes, it does not make sense to resist Chinese or North-Korean style full-blown cyberdictatorship, say. There is no normativity inherent in the concept of a culturally sophisticated primate, let alone in that of a bunch of neural tissue to be found in an organism. If we were identical to one of those models of the human being, we would lose the very concept of human dignity underpinning the value system of the democratic rule of law.

Natural science as such is the value-free discovery of natural facts. This is why science has not only contributed to human health, security, and flourishing, but at the same time turned out to be the biggest killing machine humanity has ever witnessed. Millions of people were killed in the wake of scientific progress, and humanity is currently on the brink of self-extinction as a consequence of the misguided idea that human progress can be replaced by natural-scientific and technological progress. This ideology is quite literally a dead-end, based on flawed metaphysics.

Concluding Remarks

What we are currently witnessing on a global scale in our digital age is a struggle among different conceptions of the human. We rid ourselves of the very capacity to describe our situation, to make it transparent and, thereby, to defend the kinds of universal values that guarantee socio-economically mediated access to humanity's invariant core, if we march for science without marching for the humanities. That humanity should not destroy itself by ruining the only planet we will ever thrive on, cannot be deduced at all from natural scientific and technological knowledge. As long as we do not grant the humanities and all other academic disciplines equal epistemological standing, natural science too will be easy prey for those who do not care about the facts, but really are interested only in maximizing the reach of their will to power. Thinking that scientific knowledge is valuable is simply not a piece of natural-scientific knowledge. To think otherwise, is to be deluded and to fall in the target domain of the humanities and social sciences which, among other things, ought to study the delusions of the so-called scientific worldview.

For all the reasons sketched in my paper, we are entitled to reject the very idea of conscious robots. Let me conclude by pointing out that even if (*per impossibile*) there could be conscious robots, this very possibility does not entail the desirability of their actual existence. Rather, I suggest by way of a conclusion that ethical reflection yields the result that we ought not to produce cerebral organoids implanted in a robotic "body."

This is no argument against technological or medical progress. Rather, it is a reminder of the fact that scientific discovery is subject to the value-system of the human life form, the only system we can know of as humans. Whether or not it is somehow backed up by a transcendent God does not matter for ethics, as morality takes care of itself: what we ought to do cannot merely be a consequence of the fact that the Almighty dictates what we ought to do anyhow. The fact that a certain kind of action is good or evil, i.e. ethics, cannot be derived from the mere decree of any kind of will. If there is good and evil (morally recommended and morally prohibited action), God himself does not create it.²⁷

References

- Bitbol, M. (2019). *Maintenant la finitude. Peut-on penser l'absolu?* Paris: Flammarion.
- Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology* (Minnesota studies in the philosophy of science) (Vol. 9, pp. 261–325). Minneapolis: University of Minnesota Press.
- Block, N. (1986). Advertisement for a semantics of psychology. *Midwest Studies in Philosophy*, 10(1), 615–678. <https://doi.org/10.1111/j.1475-4975.1987.tb00558.x>.
- Block, N. (2019). What is wrong with the no-report paradigm and how to fix it. *Trends in Cognitive Sciences*, 23(12), 1003–1013. <https://doi.org/10.1016/j.tics.2019.10.001>.
- Burge, T. (2013). Self and self-understanding: The Dewey lectures (2007, 2011). In T. Burge (Ed.), *Cognition through understanding* (Self-knowledge, interlocution, reasoning, reflection: Philosophical essays) (Vol. 3, pp. 140–226). Oxford: Oxford University Press.
- Campbell, J., & Cassam, Q. (2014). *Berkeley's puzzle: What does experience teach us?* Oxford: Oxford University Press.

²⁷In my book *Moralischer Fortschritt in dunklen Zeiten* (Berlin: Ullstein 2020) I defend a brand of "universal moral realism", according to which moral value (the good, the neutral, the bad/evil) is not in the eye of any beholder, but rather an objectively existing property of the action assessed by a morally trained participant in a morally relevant practice. Like all other forms of truth-apt judgment, moral judgment is fallible. Moral properties are relational: they essentially express relationships between human beings and, indirectly, between different organic life forms. Thus, there is neither a direct nor an indirect set of duties with respect to the inanimate layer of the physical universe. What we owe to our inanimate environment is always a function of what we owe to each other and to the rest of the animal kingdom. Conscious cerebral organoids implanted in a robot would be organisms to whom we owe something. In particular, we owe it to them not to produce them in the first place.

- Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*, 60, 441–458. <https://doi.org/10.2307/3131782>.
- Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358, 486–492. <https://doi.org/10.1126/science.aan8871>.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge: MIT Press.
- Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. New York: W.W. Norton and Company.
- Ellis, G. (2016). *How can physics underlie the mind? Top-down causation in the human context*. Berlin/Heidelberg: Springer.
- Falkenburg, B. (2007). *Particle metaphysics: A critical account of subatomic reality*. Berlin/Heidelberg: Springer.
- Falkenburg, B. (2012). *Mythos Determinismus: Wieviel erklärt uns die Hirnforschung?* Berlin/Heidelberg: Springer.
- Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford: Oxford University Press.
- Fuchs, T. (2018). *Ecology of the brain: The phenomenology and biology of the human mind*. Oxford: Oxford University Press.
- Gabriel, M. (2015a). *Why the world does not exist*. Cambridge: Polity.
- Gabriel, M. (2015b). *Fields of sense: A new realist ontology*. Edinburgh: Edinburgh University Press.
- Gabriel, M. (2017). *I am not a brain: Philosophy of mind for the 21st century*. Cambridge: Polity.
- Gabriel, M. (2018a). *Der Sinn des Denkens*. Berlin: Ullstein Buchverlage.
- Gabriel, M. (2018b). *Neo-existentialism*. Cambridge: Polity.
- Gabriel, M. (2018c). Review of Owen Flanagan and Gregg Caruso, eds., *Neuroexistentialism: Meaning, morals, and purpose in the age of neuroscience*. Available via Notre dame philosophical reviews. Retrieved February 12, 2020, from <https://ndpr.nd.edu/news/neuroexistentialism-meaning-morals-and-purpose-in-the-age-of-neuroscience/>.
- Gabriel, M. (2019). The paradox of self-consciousness: A conversation with Markus Gabriel. Available via edge. Retrieved February 12, 2020, from https://www.edge.org/conversation/markus_gabriel-the-paradox-of-self-consciousness.
- Gabriel, M. (2020a). *The meaning of thought*. Cambridge: Polity.
- Gabriel, M. (2020b). *Fiktionen*. Berlin: Suhrkamp.
- Gabriel, M., & Ellis, G. (2020). Physical, logical, and mental top-down effects. In M. Gabriel & J. Voosholz (Eds.), *Top-down causation and emergence*. Dordrecht: Springer, Forthcoming.
- Goff, P. (2019). *Galileo's error: Foundations for a new science of consciousness*. New York: Pantheon Books.
- Hacking, I. (1999). *The social construction of what?* Cambridge/London: Harvard University Press.
- Hacking, I. (2002). Historical ontology. In P. Gärdenfors, J. Wolenski, & K. Kijania-Placek (Eds.), *In the scope of logic, methodology and philosophy of science* (Vol. II, pp. 583–600). Dordrecht: Springer.
- Harris, A. (2019). *Conscious: A brief guide to the fundamental mystery of the mind*. New York: HarperCollins.
- Hennig, B. (2007). Cartesian conscientia. *British Journal for the History of Philosophy*, 15(3), 455–484. <https://doi.org/10.1080/09608780701444915>.
- Hoffman, D. (2019). *The case against reality: Why evolution hid the truth from our eyes*. New York: W. W Norton & Company.
- Ismael, J. T. (2016). *How physics makes us free*. Oxford: Oxford University Press.
- Kant, I. (2016). *Grundlegung zur Metaphysik der Sitten*. Riga: Hartknoch.
- Koch, C. (2019). *The feeling of life itself: Why consciousness is widespread but can't be computed*. Cambridge: MIT Press.
- Ladyman, J., & Ross, D. (2007). *Every thing must go. Metaphysics naturalized*. Oxford/New York: Oxford University Press.
- Linnaeus, C. (1792). *The animal kingdom or zoological system* (trans: Kerr R) (pp. 44–53). Edinburgh: Creech.
- Meillassoux, Q. (2006). *Après la finitude. Essai sur la nécessité de la contingence*. Paris: Éditions du Seuil.
- Noble, D. (2016). *Dance to the tune of life: Biological relativity*. Cambridge: Cambridge University Press.
- Penrose, P. (1989). *The emperor's new mind. Concerning computers, minds, and the laws of physics*. Oxford: Oxford University Press.
- Putnam, H. (1981). Brains in a vat. In H. Putnam (Ed.), *Reason, truth and history* (pp. 1–21). Cambridge: Cambridge University Press.
- Rometsch, J. (2018). *Freiheit zur Wahrheit: Grundlagen der Erkenntnis am Beispiel von Descartes und Locke*. Frankfurt am Main: Klostermann.
- Rowlands, M. (2014). *Externalism: Putting mind and world back together again*. New York/Oxford: Routledge.
- Ryle, G. (1949). *The concept of mind*. Chicago: University of Chicago Press.
- Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton: Princeton University Press.
- Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge: MIT Press.
- Smith, B. C. (2019). *The promise of artificial intelligence: Reckoning and judgment*. Cambridge: MIT Press.
- Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. New York: Alfred A. Knopf.
- Tononi, G., & Koch, C. (2014). Consciousness: Here, there but not everywhere. Available via Cornell University. Retrieved February 12, 2020, from <https://arxiv.org/ftp/arxiv/papers/1405/1405.7089.pdf>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.