



The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns

Elisabeth Hildt

To cite this article: Elisabeth Hildt (2022): The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns, AJOB Neuroscience, DOI: [10.1080/21507740.2022.2148773](https://doi.org/10.1080/21507740.2022.2148773)

To link to this article: <https://doi.org/10.1080/21507740.2022.2148773>



Published online: 21 Nov 2022.



Submit your article to this journal [↗](#)



Article views: 86



View related articles [↗](#)



View Crossmark data [↗](#)

TARGET ARTICLE



The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns

Elisabeth Hildt

Illinois Institute of Technology

ABSTRACT

Can machines be conscious and what would be the ethical implications? This article gives an overview of current robotics approaches toward machine consciousness and considers factors that hamper an understanding of machine consciousness. After addressing the epistemological question of how we would know whether a machine is conscious and discussing potential advantages of potential future machine consciousness, it outlines the role of consciousness for ascribing moral status. As machine consciousness would most probably differ considerably from human consciousness, several complex questions must be addressed, including what forms of machine consciousness would be morally relevant forms of consciousness, and what the ethical implications of morally relevant forms of machine consciousness would be. While admittedly part of this reflection is speculative in nature, it clearly underlines the need for a detailed conceptual analysis of the concept of artificial consciousness and stresses the imperative to avoid building machines with morally relevant forms of consciousness. The article ends with some suggestions for potential future regulation of machine consciousness.

KEYWORDS

Artificial intelligence; consciousness; personhood; ethics; robotics

INTRODUCTION

Even though there is general agreement that current machines do not display consciousness, the question of what would be the ethical implications if artificial consciousness¹ existed is worth discussing. This is particularly true as several robotics approaches toward building conscious machines have been described or are being pursued (e.g., Chatila et al. 2018; Kinouchi and Mackin 2018; Chella et al. 2020). While it is striking that researchers have made considerable progress toward building machines that display certain components of consciousness, it is equally true that the field is quite far away from building machines that display morally relevant forms of consciousness—if this should ever prove to be possible at all. Research in this direction raises several complex philosophical and ethical questions, however, that clearly need interdisciplinary reflection.

Before reflecting in detail on machine consciousness, it will be useful to outline what we normally understand consciousness to be. Consciousness is a term with many different meanings and uses (Van Gulick 2018; Gennaro 2019). The adjective conscious is used to characterize a conscious entity, i.e., an

entity that is sentient, wakeful, and self-conscious. Another use is “to be conscious of something,” to have knowledge of the external world or to be aware of one’s own mental states. A third way is to talk of conscious mental states, which are mental states an entity is aware of being in.

There has been a complex and multifaceted philosophical discussion around consciousness, particularly around the question of what makes a mental state a conscious mental state. In this debate, various components have been stressed: Accordingly, being conscious is having subjective experiences, or feeling emotions. Conscious mental states involve an internal perspective. They have been characterized as having qualitative properties or “qualia,” i.e., experienced properties of conscious states. Phenomenal consciousness relates to the subjective qualities of experiences, sensations, or emotions. It has been characterized as what it is like to be in a conscious mental state. In contrast to phenomenal consciousness, access consciousness relates to the extent to which information is available to and accessible by the organism. In the philosophical debate on consciousness, several authors have stressed the co-called “Hard problem of consciousness” (Chalmers 1996), i.e., the currently

CONTACT Elisabeth Hildt  ehildt@iit.edu  Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, IL, USA.

¹In this article, I use the terms artificial consciousness, machine consciousness, and AI consciousness interchangeably.

unresolved question of how physical systems can have phenomenal experiences (Nagel 1974; Block 1995; Dennett 1991, 1996; Chalmers 1996, 2010).

For centuries, consciousness has played a central role in debates around the mind-body problem, which is the question of the relation between the body and the mind. During this longstanding debate, manifold positions on questions like whether or not mental states are physical states or how conscious mental states can originate in biological brains have been put forward, but there is still no consensus. The lack of understanding of the nature of human consciousness and the lack of agreement on how to talk about human consciousness pose considerable problems for the study of artificial consciousness.

Even though any reflection on potential future artificial consciousness may seem somehow speculative, a reflection on and specification of the concept of artificial consciousness is crucial in two regards. First, concept clarification is a precondition of being clear what exactly one is talking about; and second, concepts often come with implicit value assumptions (Salles, Evers, and Farisco 2020). Many of the concepts used around artificial intelligence (AI), like the concept of an artificial *agent*, of *autonomous* vehicles, or the mere concept of artificial *intelligence*, build on notions and terminology that traditionally have been used for human beings. These concepts are highly value laden and have been subject to centuries of interpretation and philosophical reflection. Insofar, the use of these concepts in the context of technology may come with assumptions and implicit interpretation that may not always be obvious.

In humans and animals, consciousness plays a crucial role for the attribution of moral status. If machines—AI technology or robots—displayed consciousness comparable to human consciousness, this certainly would have a broad spectrum of ethical implications. Thus, a detailed conceptual analysis of the concept of artificial consciousness is of considerable relevance.

In what follows, I will first give an overview of the various consciousness components used by the different existing approaches toward machine consciousness. As it will turn out, there are several factors that hamper an understanding of machine consciousness. After addressing the epistemological question of how we would know whether a machine is conscious and discussing potential advantages of potential future machine consciousness, I will outline the role of consciousness for ascribing moral status. As machine consciousness would most probably differ considerably

from human consciousness, a number of complex questions must be addressed, including what forms of machine consciousness would be morally relevant forms of consciousness, and what the ethical implications of morally relevant forms of machine consciousness would be. While admittedly part of this reflection is speculative in nature, it clearly underlines the need for a detailed conceptual analysis of the concept of artificial consciousness and stresses the imperative to avoid building machines with morally relevant forms of consciousness. The article ends with some suggestions for potential future regulation of machine consciousness.

MACHINE CONSCIOUSNESS—APPROACHES, COMPONENTS, AND DEFINITIONS

Can machines be conscious? This has been one of the central questions in the debate around strong AI and weak AI since the 1970's. Proponents of weak AI reject the possibility that machines could have consciousness or a mind (Searle 1980). Others hold the view that machine consciousness is possible in principle (Dennett 1997a, 1997b; McDermott 2007; Dehaene, Lau, and Kouider 2017; Graziano 2017).

While there is general agreement that conscious AI is currently not technically feasible and that machine consciousness does not exist yet (Dennett 1997a; Bryson 2010; Basl 2014; Dehaene, Lau, and Kouider 2017), several approaches toward building machine consciousness have been suggested or are currently being pursued.

The various approaches focus on different components of artificial consciousness. In what follows, I will outline some of these components: understanding and intentionality; subjective experiences; global availability; integration and self; and self-consciousness, self-awareness, and self-monitoring. Without any doubt, all these components are closely intertwined. They are all relevant, overlapping aspects of the concept of consciousness. Insofar, it is primarily for analytical purposes that I discuss them separately here. Even though this list is not exhaustive, the considerable plurality of consciousness components stressed by the various authors reflects the broad spectrum of approaches toward machine consciousness.

Understanding and Intentionality

Since the beginning of the debate on strong and weak AI and machine consciousness, the concepts of meaning, understanding, and intentionality have played a

central role. This discussion revolves around the question of how external reality can be represented in machines. While mental states like thoughts, beliefs, or desires have intentionality, i.e., they are about or refer to something, it has been an unresolved question of how intentionality can be achieved in machines.

Based on the so-called “Chinese Room Thought Experiment,” John Searle (1980) famously argued that a machine that uses a program could not understand anything, because what the computer program does is formal symbol manipulations. These formal symbol manipulations do not have any intentionality. If this assumption is correct, machines could never be conscious. While Searle’s “Chinese Room Argument” has been widely disputed (Cole 2020), the problem of how external reality and its meaning can be represented still plays an important role in robotics and artificial consciousness research.

In their article on self-aware robots, Raja Chatila and colleagues write (Chatila et al. 2018, p. 1): “We believe we still lack a genuine theory of the underlying principles and methods that would explain how we can design robots that can understand their environment and not just build representations lacking meaning [...]”

Robots’ capability of understanding would be the basis of various more complex robot capabilities that require understanding, such as robots being aware of what they do or learning from their own experience.

However, there are also different approaches to machine consciousness and intentionality that allow us to reconcile the concept of intentionality more easily with machine capabilities. For example, Drew McDermott, while characterizing digital computers as syntactic engines par excellence, builds on the concept of impersonal intentionality in his account of machine phenomenal consciousness. He characterizes impersonal intentionality as intentionality in that there is hardware and software that “refers to” the environment in ways appropriate to the machine’s requirements (McDermott 2007). Accordingly, impersonal intentionality does not refer to or involve words, “images,” or “thoughts” but provides the basis for environment- and context-related navigation.

Subjective Experiences

When it comes to machine consciousness, the question of whether and how machines can have phenomenal states, i.e., subjective experiences including emotions or pain, has been considered central by several authors (Koch and Tononi 2008; Chella and

Manzotti 2009; Levy 2009; Graziano 2017; Reggia, Katz, and Davis 2018; Metzinger 2021; Mosakas 2021). Accordingly, machine consciousness is first and foremost about phenomenal consciousness in machines, about machines having subjective experiences. This perspective is in line with the way we as humans experience our own human consciousness.

For example, Olivia Carter and colleagues write (Carter et al. 2018, p. 400): “To determine whether machines are conscious, we must ask whether they have subjective experiences: Do machines consciously perceive and sense colors, sounds, and smells? Do they feel emotions?”

Similarly, in their article on Humanoid Cognitive Robots, James A. Reggia and colleagues (Reggia, Katz, and Davis 2018, p.1) use the term “consciousness” to mean phenomenal consciousness and explicitly refer to the subjective qualities of sensory phenomena, emotions, and mental imagery.

However, several authors have argued that focusing on subjective conscious experiences, qualia, and the hard problem of consciousness is not a productive approach for research into machine consciousness or may even hamper the development of robot consciousness (Dehaene, Lau, and Kouider 2017, 2018; Manzotti and Chella 2018; van der Velde 2018).

Instead, they suggest conceiving subjective experiences in machines in different ways. Stanislas Dehaene, Lau, and Kouider (2017, 2018), for example, assume that conscious subjective states may become realizable in machines if two computational ingredients—global information sharing and self-monitoring—are jointly and adequately implemented (see sections below). For Drew McDermott (McDermott 2007), phenomenal consciousness is directly related to a system’s self-model (see section below on *Integration and Self*). Manzotti and Chella (2018) suggest consciousness to be a physical, something that is not literally inside the body of an agent. They propose that “the physical properties of the external world might be the same as the properties of conscious experience” (Manzotti and Chella 2018, pp. 7–8).

Global Availability

Global availability refers to consciousness in the sense that conscious information is accessible to the whole system. Positions that stress the aspect of global availability approach machine consciousness from a functional perspective; they reject dividing conscious experience and cognitive functions (Cohen and Dennett 2011). These positions rely on the

neuroscience-based global workspace theory (GW theory) and its basic assumption that consciousness has an integrative function.

The global workspace theory (Baars 2002; Baars 2005) suggests that with fleeting memory capacity, only one consistent content can be dominant at any given time. Accordingly, dominant content is widely distributed throughout the brain; consciousness arises when content enters the global workspace of the brain and transiently dominates it. From this perspective, the primary functional role of consciousness is to integrate and coordinate the functioning of specialized networks that would otherwise function independently.

A similar assumption would hold for conscious machines. According to the conscious access hypothesis, conscious information is information that is distributed to the whole system, i.e., accessible throughout the system (Dehaene, Lau, and Kouider 2017; van der Velde 2018).

When reflecting on machine consciousness, Dehaene, Lau, and Kouider (2017) distinguish between two essential dimensions of conscious computation that can be independent of each other: global availability (C1) and self-monitoring (C2). Global availability (C1) characterizes the relationship between a cognitive system and a specific object. Conscious information, in this sense, is information that is globally available to the system so that the system can recall it or act upon it. C1 describes the capability to access external information; it resembles Ned Block's access consciousness. To describe what it would mean for a machine to have C1 consciousness, Dehaene, Lau, and Kouider (2017) refer to the example of a car that is short of gas and in which the "gas light" is flashing. Whereas in current cars, all other modules remain uninformed, in a machine with C1, modules would be able to share the information and collaborate to address the problem.

Integration and Self

The notion of unity and the role of a single integrated system for consciousness have been emphasized by several approaches to robot consciousness (Chella and Manzotti 2009). According to the Information Integration Theory of Consciousness, consciousness depends on the capacity of a system to integrate information (Tononi 2004; Koch and Tononi 2008; 2017). Hence, the higher the amount of integrated information that an organism or a system can generate, the higher the level of consciousness. The theory stresses

that for a system to be conscious, there has to be a single integrated system with a big repertoire of states.

Drew McDermott (2007) points out that consciousness results from the structure of the self-models used by intelligent systems to reason about themselves. He describes his approach toward machine consciousness as follows (McDermott 2007, p. 130): "The idea in a nutshell is that phenomenal consciousness is the property a computational system X has if X models itself as experiencing things."

The focus of this perspective is not so much on the subjective qualities of experiences, but on a system's self-model that includes assumptions on how the system's senses work. Accordingly, machines develop self-models, building on what McDermott (2007) characterized as "impersonal intentionality." These self-models are used to describe how the systems interact with the world.

Based on global workspace theory, Kinouchi and Mackin (2018) conceive a system's unity to result from a system-level structure that organizes the underlying parallel-processing units. The authors assume that the consciousness function realized in this form of system-level adaptation corresponds to how the human mind is "aware" when making decisions. Accordingly, they speculate that such a system is "aware" of external elements that build the outside world and of itself within this world. When reflecting on the machine "self" in their model, they write (Kinouchi and Mackin 2018, p. 14):

The "self" existing at the origin forms the relationship between "self" and the objects. The "self" sees and copes with the objects. We speculate that this relationship contributes to the awareness of the first-person perspective as if the homunculus in our brain sees the outside world [...].

Even though it is difficult to grasp what the authors mean by machine "self" and by alluding to a homunculus inside the brain that sees the outside world, it must be conceded that any attempt to develop a conception of machine "self" is hampered by the fuzziness of the concept of "self."

Self-Consciousness, Self-Awareness, and Self-Monitoring

Self-consciousness is a concept closely related to consciousness. It is about being aware of oneself. While in the debate on machine consciousness, there certainly is close overlap between concepts like self-model, unity, and self on the one hand, and self-awareness, self-monitoring, and self-consciousness, on

the other hand, the latter concepts tend to be used differently.

For example, John McCarthy and Patrick J. Hayes gave the following characterization of machine self-consciousness (McCarthy and Hayes 1981, p. 437): “Is p self-conscious? That is, does it have a fair variety of correct beliefs about its own beliefs and the processes that change them?”

Likewise, the approach described by Dehaene, Lau, and Kouider (2017) comprises a component related to self-consciousness, which is self-monitoring (C2). They consider self-monitoring to be one of two essential dimensions of conscious computation. While global availability is transitive as it refers to “being conscious of something,” self-monitoring is about a cognitive system monitoring its own processing and obtaining information about itself. Dehaene et al. describe it as follows (Dehaene, Lau, and Kouider 2017, p. 487):

This sense of consciousness corresponds to what is commonly called introspection, or what psychologists call “meta-cognition”—the ability to conceive and make use of internal representations of one’s own knowledge and abilities.

Accordingly, a machine with this type of self-monitoring would have an integrated image of itself, of its location, and capabilities. It would know about knowing.

Chatila et al. (2018) focus in their robotics research on self-awareness, which they characterize as comprising meta-cognitive processes such as reporting to oneself and others about one’s internal states and decision-making.

There are also more far-reaching reflections involving self-consciousness in machines. For Michael S.A. Graziano, attention schema theory (AST) of consciousness—a theory on subjective awareness in brains—is a possible starting point for building artificial consciousness (Graziano 2017). Extending on the body schema, the theory suggests that in addition to the model of a physical body, the brain constructs a model of its own internal information processing, i.e. an “attention schema”. The attention schema is relevant for controlling attention and for claims a system can make about itself and its information handling processes.

Graziano’s approach to artificial consciousness aims to develop machines that contain a model of consciousness, and which, based on this model of consciousness, in considering their own capabilities, claim with a high degree of certainty that they are conscious. Overall, the machine’s claim of possessing

subjective awareness would have to be seen as a form of self-awareness or self-consciousness.

A similarly complex suggestion is made by Chella et al. (2020), who provide a model of inner speech in robots and reflect on the connection between inner speech and self-awareness. Beginning with the observation that inner speech accompanies the introspection of mental life and plays an important role in human self-awareness, they suggest implementing a form of self-awareness in robots by developing inner speech. They not only consider inner speech deployment as a privileged method for achieving self-awareness in robots, but also elaborate on advantages that robots that display inner speech would have. In particular, they stress that robots with inner speech capabilities would be more transparent and trustworthy to human users, as their decision-making processes would be audible via inner speech.

HOW TO CONCEIVE OF MACHINE CONSCIOUSNESS?

Overall, the debate on machine consciousness can be characterized and complicated by several factors. First, there is the fundamental gap between mind and brain, mind and matter. Related is the so-called “hard problem of consciousness,” i.e., the currently unresolved question of how physical systems can have phenomenal experiences and subjectivity. Regarding machine consciousness, this problem seems to be particularly difficult to overcome.

While we as humans know from our own experience what human brain-based consciousness is like, it is difficult for us to conceive of machine consciousness. Philosophical reflection has traditionally focused on the phenomenal perspective, which is the first-person perspective. However, humans only have access to machines from the outside, from the third person perspective. A strategy toward machine consciousness that aims to transfer the first-person perspective humans have of human consciousness to AI technology and then to reflect from a third person perspective on whether machines are conscious, certainly faces some difficulties.

Also, it is far from clear whether a perspective on machine consciousness that stresses phenomenal consciousness and subjective experiences is adequate at all. Instead, a less anthropomorphic perspective on machine consciousness that focuses on global availability, information integration, or high-level adaptation could be much more useful for the field to move forward.

A number of robotics approaches stress not so much the subjective experiences and the qualia aspect of consciousness but meta-cognitive capabilities that provide machines with awareness of themselves in their respective environments. The fact that several robotics approaches do not focus on phenomenal experiences, but stress other consciousness components is indicative of a perceived need for robotics research not to directly transfer the debate around human consciousness to machines, but to develop new ways of understanding machine consciousness. However, the downside of these approaches is that the conceptions used in AI and robotics contexts are very different from the traditional perspective toward human consciousness which is dominated by the first-person perspective and subjective experiences. Thus, definitions of artificial consciousness are considerably different from definitions that have been given for human consciousness. In view of this, it is not easy to understand what these machine consciousness conceptions entail.

In addition, the debate is complicated by the fact that, at this point, there is no uniform definition of machine consciousness. As outlined above, the various approaches toward machine consciousness focus on different components of consciousness and use different conceptions of machine consciousness.

While in humans, capabilities like having emotions, feeling pain, being capable of self-reflection or suffering are closely intertwined with consciousness, this may not necessarily be the case with machines, especially if machine consciousness is understood in terms of information integration, self-model, or global availability.

This relatively broad spectrum of consciousness conceptions makes it very difficult to compare the various approaches with each other. Furthermore, it complicates reflecting on and talking about “machine consciousness,” as without referring to a specific approach, it is not clear what is meant. Overall, this may leave everyone to one’s own intuitions about what “machine consciousness” could be.

In view of this, it is necessary for researchers to specify their respective consciousness conceptions and to clearly identify what machine consciousness in the respective context is assumed to be. While it may sound like just a matter of terminology what to consider to be machine consciousness, this is something of crucial relevance. For depending on the conception of machine consciousness used, machine consciousness seems to be more or less technically feasible or more or less likely ever to exist. The less ambitious

the conception of consciousness used, the easier it may seem to characterize technology as being conscious. In view of the complex ethical implications linked to the concept of consciousness, it is important to carefully think about adequate conceptions here, and not to claim too much.

Strikingly, in the robotics literature, when discussing machine consciousness and related topics, there seems to be a tendency to not focus so much on the term “consciousness,” but instead to use terms like “awareness” and “self-awareness”. On the one hand, awareness and self-awareness certainly are central for consciousness and represent crucial aspects of consciousness. On the other hand, however, the mere fact that different terminologies are used favors the decoupling of two otherwise closely connected discussions: one on the practical aspects of building aware and self-aware AI and robots, and one on the philosophical and ethical aspects of artificial consciousness.

This is further complicated by the fact that not only the concept of consciousness but also concepts like self, self-awareness, or self-monitoring are not uniformly used in the robotics literature and differ from the conceptions used when these terms are applied to humans.

Furthermore, when characterizing robot capabilities, a rather liberal use of terms like “self” or “self-awareness” can be observed that tends to imply more than the technical reality can provide at this point in robots (see also Chella and Manzotti 2009).

All of this points to a lack of understanding of the concept of machine consciousness which not only considerably hampers any reflection on the ethical implications of machine consciousness but is an ethical problem in itself. In view of this, developing a theoretical approach toward machine consciousness is absolutely required (Schwitzgebel and Garza 2015; Hildt 2019; Metzinger 2021).

HOW WOULD WE KNOW WHETHER A MACHINE IS CONSCIOUS?

A broad spectrum of empirical tests on machine consciousness has been developed. In their review, Aida Elamrani and Roman Yampolskiy (Elamrani and Yampolskiy 2019) give an overview of more than 20 tests for machine consciousness which assess different components of consciousness. The authors distinguish between architecture-based tests, which they describe as tests where the presence of consciousness is inferred from the correct implementation of relevant architecture, and behavior-based tests, where machine

consciousness is assumed based on behavior similar to that of human beings.

The review by Elamrani and Yampolskiy is indicative of the dilemma humans face when seeking to find out whether a machine displays consciousness: From the third-person perspective, humans can either refer to what they know about the way the machine was designed and built, or they can refer to machine behavior.

Both approaches have advantages and disadvantages. Tests that focus on how machines are built presuppose detailed knowledge of the machines in question and their architecture. A certain type of architecture or a certain level of complexity is considered a precondition for a machine to be conscious or to display a certain consciousness component. These tests may rely heavily on brain analogies or analogies to human consciousness, however—analogies that may not prove useful in view of machines not having a biological brain. Tests that rely on machine behavior do not require knowledge of how a machine was built. They depend to a very great extent on the human interpretation of behavior exhibited by a machine. As machine behavior may consist in the simulation of human-style behavior and may not be indicative of a machine displaying consciousness, there are clear limitations to this type of approach, though.

Insofar, a combination of both types of tests seems adequate: if a machine behaves as if conscious and fulfills the requirement of a certain level of architectural complexity, the assumption of it being conscious or displaying the consciousness component tested for may be warranted.

Overall, however, it must be stressed that each of the tests relates to the consciousness component it is supposed to test for. The tests serve to identify machine consciousness only insofar as they align with the consciousness conception or consciousness component used to develop the test. As the various approaches toward machine consciousness rely on different consciousness components, they require different tests. Some consciousness components certainly are more amenable to empirical tests than others. Even if a machine passed one of these tests, it may still be unclear what the relevance would be.

Regarding the question of how to detect artificial consciousness in machines, the fundamental epistemological problem is that humans are not capable of having a first-person perspective toward machine consciousness. With other humans, we may infer by analogy, knowing what human consciousness is like, knowing what its behavioral implications are. But

with machines? In the foreseeable future, machines will not be capable of telling us whether they are conscious or not. And even if they did, we would not know what this meant.

WHAT WOULD BE THE ADVANTAGES OF ARTIFICIAL CONSCIOUSNESS?

Machine consciousness certainly is a captivating topic that reflects a general fascination with technology. The idea of building machines that display human-like characteristics is a science-fiction scenario that has been inspiring human imagination for centuries. However, from a technical and engineering perspective, the question is: What would be the advantages of machines displaying some form of consciousness over machines not displaying consciousness?

While building artificial consciousness for entertainment purposes may seem cool (Graziano 2017), there undoubtedly are practical reasons. Some of the reflections on the potential roles of machine consciousness can be seen as closely related to reflections on the biological functions of brain-based consciousness (Van Gulick 2018).

Robots that display some form of consciousness would be expected to be better at integrating information in that they can make information available to the whole system, have higher-level integration and coordination of otherwise independently functioning subparts, or can self-monitor their position, knowledge, and abilities. The various approaches toward machine consciousness stress different aspects of information integration, see Section “MACHINE CONSCIOUSNESS—APPROACHES, COMPONENTS, AND DEFINITIONS.” This high level of information integration would convey improved functionality and increased capability to execute complex tasks. Increased access to information about themselves and the environment would come with increased flexibility and improved control and self-control functions. It would enable machines to better integrate in and interact with their environment.

Several authors have suggested that conscious or self-aware robots would be better at interacting or cooperating with humans or other AI technology. Graziano (2017) argues that consciousness is necessary for machines to have social intelligence, social competence, and empathy. He claims that only if robots understand what consciousness is, can they fully interact with humans in social contexts. Also, Antonio Chella and colleagues (Chella et al. 2020) reflect on human-robot interaction. They argue that

self-awareness with inner speech would increase the robots' social competencies and increase their acceptability and trustworthiness in social contexts. If a robot's inner speech were audible, its decision-making processes would be transparent to the user, thereby increasing trust and improving human-robot-interaction.

It remains to be seen whether these claims about machine consciousness and the capabilities of potential future conscious machines are realistic, and whether machines with these capabilities will be needed in the future. The above claims certainly point to the advantages consciousness-related features are expected to have. Without any doubt, improved functionality, increased capability to execute complex tasks, and improved human-robot interaction are important arguments in favor of machine consciousness.

However, the role, importance, or value of machine consciousness is not just confined to practical or functional aspects. As consciousness plays a crucial role in an entity's moral status and is a central component of personhood, consciousness is a concept that comes along with considerable ethical implications. This context will be discussed below.

CONSCIOUSNESS, MORAL STATUS, AND PERSONHOOD

In general, an entity's moral status has been considered to depend at least in part on the entity being conscious or on the extent to which an entity is conscious (Van Gulick 2018).

Generally, an entity is considered to have moral status if it counts morally in its own right. When acting toward or interacting with the entity, certain actions are permissible, impermissible, or required for the entity's own sake. This implies that the entity or its interests, preferences, plans, desires, or feelings morally matter to some extent (Kamm 2007; Basl 2013; Bostrom and Yudkowsky 2014; Jaworska and Tannenbaum 2018; Mosakas 2021).

With entities with moral status, a distinction has been made between moral agents, i.e., entities capable of moral decision-making and agency, and moral patients, i.e., vulnerable entities that are not themselves moral agents but are of moral concern. While human beings are moral agents, dogs or cats for example, are moral patients; they do not have moral decision-making capacities but certainly must be considered morally (Floridi 1999; Bryson 2018).

In animals and humans, consciousness is of central relevance for moral status. Would something similar hold for machines? Several authors have stressed that the fact that machines do not have a biological origin does not count against them potentially having moral status (Dennett 1997a; Bostrom and Yudkowsky 2014; Schwitzgebel and Garza 2015). Thus, the question of whether machines are capable of consciousness has been directly connected to reflections about machine moral status and whether machines can be moral patients or moral agents.

While there is broad agreement that current AI technology and robots do not have moral status (Bryson 2010; Basl 2014; Bostrom and Yudkowsky 2014; Bryson 2018), the question of whether future AI could attain moral status as moral patients or moral agents has drawn considerable attention. In this debate, machine consciousness is of central relevance in that several authors assume that it would confer moral status to machines (Bostrom and Yudkowsky 2014; Schwitzgebel and Garza 2015; Manzotti and Chella 2018; Mosakas 2021). Consciousness is not the only criterion that matters, however. In addition to consciousness, also other criteria have been suggested for ascribing moral status to machines, such as being an information object, having a mind, or having capabilities like rationality, wisdom, intelligence, sentience or suffering (Floridi 2002; Bostrom and Yudkowsky 2014; Nyholm 2020; Metzinger 2021; Mosakas 2021). When it comes to moral status, consciousness certainly is a key characteristic, especially given the close interrelatedness between consciousness, sentience, and suffering. Conscious or sentient machines would count as moral patients, whereas moral agency would require the capability of moral decision-making and agency.

In addition, consciousness has traditionally been considered a central criterion for personhood and personal identity (Locke 1690/2017; Dennett 1976). In this context, Riccardo Manzotti and Antonio Chella write (Manzotti and Chella 2018, p. 9): "The concept of person is tightly linked to the concept of consciousness. If an entity can have subjective experiences, and eventually can suffer, then this entity should be treated as a person. In this regard, the studies on robot consciousness may force us to review our fundamental definition of the concept of person."

Accordingly, potential future machines capable of consciousness could have moral status, either as moral patients or as moral agents, i.e., persons. Not all consciousness components would be equally relevant with regard to moral status, however. Whether machines

would be attributed moral status would undoubtedly depend on the consciousness components instantiated.

WHAT WOULD BE MORALLY RELEVANT FORMS OF ARTIFICIAL CONSCIOUSNESS?

The question to be discussed here is: What exactly would be consciousness-related capabilities or characteristics that are relevant for machines having moral status, i.e., being moral patients or moral agents? What would be morally relevant forms of artificial consciousness that would confer moral status to machines?

It seems plausible to assume that not every consciousness component or consciousness-related capability would be morally relevant. Moreover, not each of the morally relevant consciousness components or consciousness-related capabilities would come equally with ethical implications or would come with the same ethical implications.

The answer to the question of what artificial consciousness in a morally relevant sense would consist in is complicated by several factors: (a) the concept of machine consciousness is only poorly defined; (b) several different machine consciousness components have been described, and (c) the concept of consciousness is closely interconnected with the concepts of sentience and suffering.

Unlike healthy human beings who display a broad spectrum of consciousness components, including phenomenal consciousness, global availability, and self-consciousness, and who are capable of sentience and suffering, the situation is different with potential future machines. Potential future machine consciousness could consist of just one or a few consciousness components, for example, access consciousness or self-awareness. And it is very well imaginable that potential future conscious machines could exist that are not sentient and not capable of suffering.

Thus, even if we assume that machines that display human-like consciousness would have human-like moral status, the assumption that machine consciousness would necessarily be human-like is not warranted.

So, what would machine consciousness in a morally relevant sense look like? Among the consciousness components discussed above, some are more central to moral status than others. While components like “Understanding and intentionality” and “Global availability” have only rarely been discussed when it comes to moral status, this is different for components like self-awareness and phenomenal consciousness, including sentience and suffering.

Phenomenal Consciousness

In particular, the capability of having subjective experiences has been considered central to the moral status of machines. For example, Kestutis Mosakas identifies the “consciousness criterion,” which he characterizes as the capacity for phenomenal experience, as crucial for the moral status of social robots (Mosakas 2021).

However, as John Basl (2014) points out, not all forms of subjective experiences are morally relevant. Accordingly, having experiences (like seeing something red) or sensations is insufficient for being a moral patient. What matters is having attitudes toward these experiences and sensations, for, as Basl stresses, it is these responses that give rise to morally relevant interests Basl (2014). When we think of attitudes toward subjective experiences—positive or negative—typical examples with moral relevance are pleasure and pain.

Also, in machines, the capability for conscious subjective experiences would be related to sentience and the capability of suffering in that only machines that are capable of having phenomenal states could suffer (Bostrom and Yudkowsky 2014; Mosakas 2021; Metzinger 2021). Against this background, Thomas Metzinger specifies four necessary conditions for the phenomenology of conscious suffering to occur: Conscious experience, Possession of a phenomenal self-model, Negative valence, and Transparency (Metzinger 2021).

Self-Awareness and Self-Consciousness

Other authors have discussed the role of self-awareness and self-consciousness as relevant criteria for ascribing moral status to machines. Bostrom and Yudkowsky (2014) characterize two criteria as central for AI moral status: sentience and sapience. They define the criterion of sapience, which they consider similar to or equivalent to personhood, as follows: “a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent” (Bostrom and Yudkowsky 2014, p. 322).

The authors then speculate about potential future AI that might be sapient but not sentient. They raise the question of whether a consciousness component like self-awareness would confer moral status to a machine that is not sentient and not capable of suffering. This question certainly is of central relevance when thinking about the moral relevance of machine consciousness components. However, the mere fact that Bostrom and Yudkowsky consider sapience as

equivalent to personhood is problematic, as personhood implies not only consciousness but also sentience. In addition, talking of machines as non-sentient persons implies that non-sentient machines would have moral status qua being persons, which begs the question of whether non-sentient machines would have moral status.

Also, other speculations involving self-awareness, self-consciousness, and the use of first-person pronouns in robots come close to imagining machines as persons. For example, Antonio Chella and colleagues wrote (Chella et al. 2020, p. 3): “A self-aware AI agent should be able to apprehend itself across time and space, as well as say things like “Hi, my name is Adam, my birthday is next week, and I am 5 years old.”

This characterization of self-aware AI is very close to John Locke’s famous definition of a person as a thinking intelligent being that conceives itself as persisting over time and space (Locke 1690/2017).

Similarly, Michael S.A. Graziano’s conception of machine self-awareness seems to indirectly imply machine personhood. He speculates (Graziano 2017, p. 7): “It should be possible to build a machine that believes it is conscious and claims it is conscious and acts like it is conscious and that talks about its consciousness in the same ways that the human machine does.”

According to Graziano’s depiction of machine self-awareness, a self-aware machine would assume that it is conscious in the same way as other conscious entities. This includes that a machine in possession of this kind of awareness would take the view that it has the same status as the other conscious entities it is aware of.

These reflections on the role of consciousness components like phenomenal consciousness, self-awareness, and self-consciousness in machines clearly reveal a compelling need for a detailed analysis of the nature and extent to which the various artificial consciousness components would confer moral status to machines.

In contrast to the above positions that consider certain consciousness components as central for machine moral status, other authors have claimed that is not so much machines’ capabilities, but their social roles, that are relevant for ascribing status or rights to machines. According to this approach, the question is not so much whether machines are conscious, but whether humans interact with them as if they were conscious, i.e., as if they were conscious in a similar way as humans are (Coeckelbergh 2010; Darling 2012;

Danaher 2020; Gunkel 2022). This type of position manages to avoid the difficult questions of what machine consciousness is and how to find out whether machines are conscious. Insofar, it offers a pragmatic way of interpreting human-machine interaction. However, as I have argued elsewhere, this approach is very problematic. It confounds the *simulation* of capabilities with the existence of capabilities and in that it is not compatible with the ways moral status is attributed in other contexts (Hildt 2019, 2021).

WHAT WOULD BE THE IMPLICATIONS OF MORALLY RELEVANT FORMS OF ARTIFICIAL CONSCIOUSNESS?

What would be the ethical implications of conscious machines? The following reflections necessarily are very speculative. They presuppose the view that machines can have moral status, even though machines are not biological systems. Moreover, they imply the assumption that our current moral considerations can be transferred to machines, i.e., that machines that display human-like morally relevant capabilities would have to be treated in a similar way as humans.

At this point, it is difficult, if not impossible to foresee the middle to long-term implications of the current approaches toward AI consciousness. Also, it is important to stress that the ethical implications of conscious machines would not just be confined to the question of the moral status of AI and robots. They would include the mode and quality of human-AI interaction more broadly. It is reasonable to expect that consciousness-related functions would improve technology performance, which could benefit humans in various ways. For example, it has been claimed that the inner speech approach toward building self-awareness in robots could increase trust in robots, as robot decision-making processes would be audible and thus better accessible (Chella et al. 2020).

A number of authors have argued in favor of rights for robots with person-like capabilities (Levy 2009; Schwitzgebel and Garza 2015; Gunkel 2018). If potential future machines displayed consciousness in a morally relevant sense, especially if they would count as moral agents, they would undoubtedly be in this category of technology for which the question of whether they can and should have rights would have to be discussed. However, a question less often addressed is what exactly these machine rights could consist of. It seems plausible to assume that the ethical implications

of different morally relevant forms of machine consciousness would be different.

If machines displayed morally relevant capabilities, it would be morally wrong to treat them in certain ways. This could imply that there are tasks or functions that humans should not require from conscious technology, be it because the technology has negative attitudes toward these tasks or rejects them. The imperative not to induce pain or suffering in potential future conscious machines would certainly have to be seen in this context (Metzinger 2021).

And there could be the requirement not to deprive conscious machines of certain resources or to uphold conditions that support their capabilities or interests. Once a machine has achieved morally relevant capabilities and is exercising them, it could be wrong to interrupt functioning. There could be a moral requirement to support the continued exercise of these capabilities. This could include the need to continue power supply for machines exerting the capabilities in question, or regularly update the software—and not for human purposes, but for the sake of these machines themselves.

Would there be a duty to provide machines with certain types of information? Would humans be required to provide training to machines, or to adequately “raise” or “educate” them? Would there be a need to provide adequate or protected environments? Furthermore: Would there be a requirement to provide uninterrupted machine functioning so that the machine can develop some sense of identity and an understanding of its continued existence over time and space? Would there be a requirement to allow networking in machines to provide social contacts, and fight loneliness and the impression of being left behind? Would it be adequate to provide machines with some sort of aging in dignity, instead of throwing them away or destroying them?²

Admittedly, questions like these are very unusual and speculative. Furthermore, we do not have an answer to any of these questions at this point. But if we take the idea of machines having moral status, either as moral patients or as moral agents, seriously, then questions like these would have to be addressed. Even though these questions are speculative, they clearly point to the fact that machines with morally relevant forms of consciousness would come with a broad spectrum of ethical implications and positive and negative duties for humans. I seriously doubt whether any potential benefits of machines with

morally relevant forms of consciousness could ever outweigh these.

NEED FOR REGULATION

There is a clear need for regulation of machine consciousness. While some authors have strongly argued against designing and building machines that might have moral status (Bryson 2010, 2018; Metzinger 2021), others have taken more liberal positions (Basl 2013; Bostrom and Yudkovsky 2014; Schwitzgebel and Garza 2015; Danaher 2020). However, even those who were not outwardly negative toward potential future machine consciousness have expressed concerns or suggested measures to avoid potential problems.

It seems obvious that no matter which direction society is going to take, there will be a need for regulation: either for regulation that aims at avoiding morally relevant forms of machine consciousness altogether, or for regulation that directs how to interact with machines displaying morally relevant forms of consciousness.

If machines that display or potentially display morally relevant forms of consciousness existed, there would be a clear need for regulation in a broad spectrum of contexts and regards. John Basl (2013), for example, focuses on the research process. While he argues that there is nothing intrinsically wrong with creating an artificial consciousness as long as the artificial consciousness would be treated in a way commensurate with it being a moral patient, he considers it likely that such created artificial consciousness would not be treated adequately. Inadequate treatment could result either from researchers or society not being aware of its consciousness or researchers or society not knowing what it would mean to interact adequately with the artificial consciousness. In order to minimize those risks, Basl (2013) suggests setting up oversight of artificial consciousness research that is similar to current regulation pertaining to research involving human subjects or non-human animal research.

Schwitzgebel and Garza (2015) suggest ethical AI design principles for potential future AI with moral status. They argue that if machines with moral status would ever exist, these machines would have to be designed in such a way that they provoke reactions from users that accurately reflect their moral status, i.e., that their moral status is clear to users.

While suggestions like these seem helpful to avoid certain types of misunderstandings, they do not address the central problem: that humans would see themselves confronted with a broad spectrum of

²As suggested by Kazuo Ishiguro in his novel “Klara and the Sun.”

responsibilities and obligations toward machines if machines with morally relevant forms of consciousness existed.

Whereas one of the main arguments toward machine consciousness is that machine consciousness would allow technology to better react to human needs which would improve human-technology interaction, it is highly questionable whether these benefits would outweigh the potential ethical implications of morally relevant machines described above. Overall, it is not plausible for humans to aim to build machines that come with the need to take on such far-reaching responsibilities toward these machines. Instead, machines are tools built to serve human interests. Machines with morally relevant forms of consciousness would turn the concept of a tool upside down.

As discussed in Section “WHAT WOULD BE THE IMPLICATIONS OF MORALLY RELEVANT FORMS OF ARTIFICIAL CONSCIOUSNESS?,” if machines had moral status, this would come with a broad spectrum of positive and negative duties for humans, most of which would be very difficult if not impossible to specify. One of the main problems in this context is that the transfer of morally relevant concepts and principles to machines would be extremely difficult. For example, while it is certainly possible to talk about a potential duty of “procreative beneficence” toward robots with moral status (Danaher 2020), this involves a clear category mistake (Ryle 1949/2000). Robots are machines to which the concept of procreation simply cannot be applied in a meaningful way.

Thus, I consider it imperative to avoid morally relevant forms of machine consciousness. As depicted above, the situation is not black and white, however. There is a broad spectrum of machine consciousness components, and not all of them are morally relevant.

Even though there is considerable uncertainty around the future development of the field of machine consciousness, it is reasonable to expect that if research toward artificial consciousness will continue, at some point steps will have to be taken to avoid morally relevant forms of machine consciousness from coming into existence.

This would have to involve a general research oversight process, which could include research committee reviews similar to Institutional Review Boards (IRBs) (Basl 2013), the development of ethical guidelines for artificial consciousness research, and/or a moratorium on building machines with certain consciousness components (Metzinger 2021).

Decisions concerning artificial consciousness and artificial consciousness research could be based on

two criteria that have been proven useful in the debate around the moral status of human embryos: potentiality and proportionality. With regard to the moral status of human embryos, the potential to become a person (principle of potentiality) has played an important role, as has the extent to which a human embryo displays person-like characteristics (principle of proportionality) (Singer et al. 1990; Gillon 1991; Poplawski and Gillett 1991; Sawai et al. 2020). While there certainly are considerable differences between the moral status of human embryos and the question of machine moral status, there are also some similarities. A principle of proportionality with regard to machines could state: To the extent that machines have (morally relevant) person-like characteristics, they deserve protection. And a principle of potentiality with regard to machines would be: To the extent that machines have the potential to achieve moral status, they deserve protection.

These two principles, together with a margin of safety, could serve as the basis for potential future regulation of machine consciousness research. Based on these two criteria, decisions on the cutoff points for acceptable forms of machine consciousness could be made that prevent the development of machines with morally relevant forms of consciousness. In analogy to regulation concerning human embryos, research or applications that aim at building machines that transcend certain forms of machine consciousness could be banned.

CONCLUSION

I began this article by outlining several different approaches toward developing machines with various consciousness components. Depending on the respective approach and the component of consciousness in question, building conscious machines seems a goal more or less achievable. While the future of the field is unclear, it cannot be excluded that at least some of these approaches could potentially lead to morally relevant forms of machine consciousness.

Throughout the article, I have strongly argued against building machines that could display morally relevant forms of consciousness. There are several complicating factors, however. As discussed above, the concept of machine consciousness is poorly defined and urgently needs clarification. Right now, it is far from clear what exactly machine consciousness in general and morally relevant forms of machine consciousness in particular would be and how they could be detected. The same uncertainty holds for the

answer to the question of what the ethical implications of morally relevant forms of machine consciousness would be.

Overall, in view of this, there is a clear need for interdisciplinary discussion that also includes reflection on potential future regulation of research into machine consciousness. In this debate, it is important to keep several types of questions apart: (a) What is the situation right now? (b) What could be possible in the not-too-distant future? And, (c) How might more distant futures look like and how might future humans manage such possible futures?

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

REFERENCES

- Baars, B. J. 2002. The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences* 6 (1): 47–52. doi: [10.1016/S1364-6613\(00\)01819-2](https://doi.org/10.1016/S1364-6613(00)01819-2).
- Baars, B. J. 2005. Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience? *Progress in Brain Research* 150:45–53.
- Basl, J. 2013. The ethics of creating artificial consciousness. *APA Newsletter on Philosophy and Computers* 13 (1):23–30. <https://cdn.ymaws.com/www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/ComputersV13n1.pdf>.
- Basl, J. 2014. Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology* 27 (1):79–96. doi: [10.1007/s13347-013-0122-y](https://doi.org/10.1007/s13347-013-0122-y).
- Block, N. 1995. On a confusion about the function of consciousness. *Behavioral and Brain Sciences* 18 (2):227–47. doi: [10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188).
- Bostrom, N., and E. Yudkowsky. 2014. The ethics of artificial intelligence. In *The Cambridge handbook of artificial intelligence*, eds. K. Frankish and W. M. Ramsey, 316–334. Cambridge: Cambridge University Press.
- Bryson, J. J. 2010. Robots should be slaves. In *Close engagements with artificial companions: Key social, psychological, ethical and design issues*, ed. Y. Wilks, 63–74. Amsterdam: John Benjamins.
- Bryson, J. J. 2018. Patience is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology* 20 (1):15–26. doi: [10.1007/s10676-018-9448-6](https://doi.org/10.1007/s10676-018-9448-6).
- Carter, O., J. Hohwy, J. van Boxtel, V. Lamme, N. Block, C. Koch, and N. Tsuchiya. 2018. Conscious machines: Defining questions. *Science* 359 (6374):400. doi: [10.1126/science.aar4163](https://doi.org/10.1126/science.aar4163).
- Chalmers, D. J. 1996. *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.
- Chalmers, D. J. 2010. *The character of consciousness*. Oxford: Oxford University Press.
- Chatila, R., E. Renaudo, M. Andries, R.-O. Chavez-Garcia, P. Luce-Vayrac, R. Gottstein, R. Alami, A. Clodic, S. Devin, B. Girard, et al. 2018. Toward self-aware robots. *Frontiers in Robotics and AI* 5:88. doi: [10.3389/frobt.2018.00088](https://doi.org/10.3389/frobt.2018.00088).
- Chella, A., and R. Manzotti. 2009. Machine consciousness: A manifesto for robotics. *International Journal of Machine Consciousness* 01 (01):33–51. VolNo. doi: [10.1142/S1793843009000062](https://doi.org/10.1142/S1793843009000062).
- Chella, A., A. Pipitone, A. Morin, and F. Racy. 2020. Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI* 7:16. doi: [10.3389/frobt.2020.00016](https://doi.org/10.3389/frobt.2020.00016).
- Coeckelbergh, M. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12 (3):209–21. doi: [10.1007/s10676-010-9235-5](https://doi.org/10.1007/s10676-010-9235-5).
- Cohen, M. A., and D. C. Dennett. 2011. Consciousness cannot be separated from function. *Trends in Cognitive Sciences* 15 (8):358–64.
- Cole, D. 2020. The Chinese room argument. In *The Stanford encyclopedia of philosophy* (Winter 2020 Edition), ed. Edward N. Zalta. <https://plato.stanford.edu/archives/win2020/entries/chinese-room/>.
- Danaher, J. 2020. Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics* 26 (4):2023–49. doi: [10.1007/s11948-019-00119-x](https://doi.org/10.1007/s11948-019-00119-x).
- Darling, K. 2012. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *We Robot Conference 2012, April 23, 2012, University of Miami; Robot Law*, eds. R. A. Calo, M. Froomkin, and I. Kerr. Cheltenham: Edward Elgar. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797.
- Dehaene, S., H. Lau, and S. Kouider. 2017. What is consciousness, and could machines have it? *Science (New York, N.Y.)* 358 (6362):486–92. doi: [10.1126/science.aan8871](https://doi.org/10.1126/science.aan8871).
- Dehaene, S., H. Lau, and S. Kouider. 2018. Response. *Science* 359 (6374):400–2. doi: [10.1126/science.aar8639](https://doi.org/10.1126/science.aar8639).
- Dennett, D. 1976. Conditions of personhood. In *The identities of persons*, ed. A. O. Rorty, 175–196. Berkeley: UC Press.
- Dennett, D. 1996. *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books.
- Dennett, D. C. 1991. *Consciousness explained*. New York: Little, Brown and Co.
- Dennett, D. C. 1997a. Consciousness in human and robot minds. In *Cognition, computation, & consciousness*, eds. M. Ito, Y. Miyashita, and E. T. Rolls, 17–29. Oxford University Press.
- Dennett, D. C. 1997b. Cog as a thought experiment. *Robotics and Autonomous Systems* 20 (2–4):251–6. doi: [10.1016/S0921-8890\(97\)80709-9](https://doi.org/10.1016/S0921-8890(97)80709-9).
- Elamrani, A., and R. V. Yampolskiy. 2019. Reviewing tests for machine consciousness. *Journal of Consciousness Studies* 26 (5–6):35–64.
- Floridi, L. 1999. Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology* 1 (1):33–52. doi: [10.1023/A:1010018611096](https://doi.org/10.1023/A:1010018611096).

- Floridi, L. 2002. On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology* 4 (4):287–304. doi:10.1023/A:1021342422699.
- Gennaro, R. J. 2019. Consciousness. *The Internet Encyclopedia of Philosophy*. <https://www.iep.utm.edu/consciou/>.
- Gillon, R. 1991. Human embryos and the argument from potential. *Journal of Medical Ethics* 17 (2):59–61. doi:10.1136/jme.17.2.59.
- Graziano, M. S. A. 2017. The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI* 4:60. doi:10.3389/frobt.2017.00060.
- Gunkel, D. 2022. The relational turn: Thinking robots otherwise. In *Social robotics and the good life: The normative side of forming emotional bonds with robots*, eds. J. Loh and W. Loh. Transcript Verlag. Forthcoming. SSRN: <https://ssrn.com/abstract=4099209>.
- Gunkel, D. J. 2018. *Robot rights*. Cambridge, MA: MIT Press.
- Hildt, E. 2019. Artificial intelligence: Does consciousness matter? *Frontiers in Psychology* 10:1535. doi:10.3389/fpsyg.2019.01535.
- Hildt, E. 2021. What sort of robots do we want to interact with? Reflecting on the human side of human-artificial intelligence interaction. *Frontiers in Computer Science* 3: 671012. doi:10.3389/fcomp.2021.671012.
- Jaworska, A., and J. Tannenbaum. 2018. The grounds of moral status. In *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta. <https://plato.stanford.edu/archives/spr2018/entries/grounds-moral-status/>
- Kamm, F. 2007. *Intricate ethics: Rights, responsibilities, and permissible harm*. Oxford: Oxford University Press.
- Kinouchi, Y., and K. J. Mackin. 2018. A basic architecture of an autonomous adaptive system with conscious-like function for a humanoid robot. *Frontiers in Robotics and AI* 5:30. doi:10.3389/frobt.2018.00030.
- Koch, C., and G. Tononi. 2008. Can machines be conscious? *IEEE Spectrum* 45 (6):55–9. doi:10.1109/MSPEC.2008.4531463.
- Koch, C., and G. Tononi. 2017. Can we copy the brain? Can we quantify machine consciousness? *IEEE Spectrum* 54 (6):64–9. doi:10.1109/MSPEC.2017.7934235.
- Levy, D. 2009. The ethical treatment of artificially conscious robots. *International Journal of Social Robotics* 1 (3): 209–16. doi:10.1007/s12369-009-0022-6.
- Locke, J. 1690/2017. An essay concerning human understanding. Book 2, Chapter 27. <https://www.gutenberg.org/files/10615/10615-h/10615-h.htm>
- Manzotti, R., and A. Chella. 2018. Good old-fashioned artificial consciousness and the intermediate level fallacy. *Frontiers in Robotics and AI* 5:39. doi:10.3389/frobt.2018.00039.
- McCarthy, J., and P. J. Hayes. 1981. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence: A collection of articles*, eds. B. L. Webber and N. J. Nilsson, 431–50. Los Altos, CA: Morgan Kaufmann Publishers.
- McDermott, D. 2007. Artificial intelligence and consciousness. In *Cambridge handbook of consciousness*, eds. P. D. Zelazo, M. Moscovitch, and E. Thompson, 117–50. Cambridge: Cambridge University Press.
- Metzinger, T. 2021. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness* 08 (01):43–66. VolNode:10.1142/S270507852150003X.
- Mosakas, K. 2021. On the moral status of social robots: considering the consciousness criterion. *AI & SOCIETY* 36:429–43.
- Nagel, T. 1974. What is it like to be a bat. *Readings in Philosophy of Psychology* 1:159–68.
- Nyholm, S. 2020. *Humans and robots: Ethics, agency, and anthropomorphism*. New York: Rowman & Littlefield.
- Poplawski, N., and G. Gillett. 1991. Ethics and embryos. *Journal of Medical Ethics* 17 (2):62–9. doi:10.1136/jme.17.2.62.
- Reggia, J. A., G. E. Katz, and G. P. Davis. 2018. Humanoid cognitive robots that learn by imitating: Implications for consciousness studies. *Frontiers in Robotics and AI* 5:1. doi:10.3389/frobt.2018.00001.
- Ryle, G. 1949/2000. *The concept of mind*. Chicago, IL: University of Chicago Press.
- Salles, A., K. Evers, and M. Farisco. 2020. Anthropomorphism in AI. *AJOB Neuroscience* 11 (2): 88–95. doi:10.1080/21507740.2020.1740350.
- Sawai, T., T. Minakawa, J. Pugh, K. Akatsuka, J. K. Yamashita, and M. Fujita. 2020. The moral status of human embryo-like structures: Potentiality matters? *EMBO Reports* 21 (8):e50984. doi:10.15252/embr.202050984.
- Schwitzgebel, E., and M. Garza. 2015. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy* 39 (1):98–119. doi:10.1111/misp.12032.
- Searle, J. R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3 (3):417–24. doi:10.1017/S0140525X00005756.
- Singer, P., Kuhse, H., Buckle, S., Dawson, K., and Kasimba, P., eds. 1990. *Embryo experimentation*. Cambridge: Cambridge University Press.
- Tononi, G. 2004. An information integration theory of consciousness. *BMC Neuroscience* 5 (1):42–22. doi:10.1186/1471-2202-5-42.
- van der Velde, F. 2018. In situ representations and access consciousness in neural blackboard or workspace architectures. *Frontiers in Robotics and AI* 5:32. doi:10.3389/frobt.2018.00032.
- Van Gulick, R. 2018. Consciousness. In *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta. <https://plato.stanford.edu/entries/consciousness/>.