The Terminology of Artificial Sentience

Janet VT Pauketat
Sentience Institute
janet@sentienceinstitute.org

Abstract

We consider the terminology used to describe artificial entities and how this terminology may affect the moral consideration of artificial entities. Different combinations of terms variously emphasize the entity's role, material features, psychological features, and different research perspectives.[1] The ideal term may vary across context, but we favor "artificial sentience" in general, in part because "artificial" is more common in relevant contexts than its near-synonyms, such as "synthetic" and "digital," and to emphasize the sentient artificial entities who deserve moral consideration. The terms used to define and refer to these entities often take a human perspective by focusing on the benefits and drawbacks to humans. Evaluating the benefits and drawbacks of the terminology to the moral consideration of artificial entities may help to clarify emerging research, improve its impact, and align the interests of sentient artificial entities with the study of artificial intelligence (AI), especially research on AI ethics.

The importance of conceptual clarity

Sentience Institute uses the term "artificial sentience" to describe artificial entities with the capacity for positive and negative experiences. When we survey the public, we use the terms "artificial beings" and "robots/AIs" to refer to "intelligent entities built by humans, such as robots, virtual copies of human brains, or computer programs that solve problems, with or without a physical body, that may exist now or in the future." We have also used the terms "artificial intelligences" and "artificial entities." These five terms are just a few of the many used in the design, development, and scholarship of AI.

When a term that expresses a concept is ambiguous, it becomes difficult to disentangle the theoretical connotation intended by the researchers and the meaning of participants' responses in empirical studies. This reduces the impact of empirical research, especially for concepts that are tied to ordinary language like "artificial."[2] Researchers may assume that everyone shares their

[1] We consider terms and their connotations in the English language. The usage and meaning of some terms is likely to be different in other languages, a topic that requires future study. Linguistic differences may have global implications for the moral consideration of artificial entities that have yet to be examined.

[2] Blascovich and Ginsburg (1978) discuss conceptual ambiguity in regards to the social psychological study of "risk-taking."

same definitions when in fact they are fuzzy concepts, "[that] possess two or more alternative meanings and thus cannot be reliably identified or applied by different readers or scholars" (Markusen, 2003, p. 702). Markusen argues that this can decrease scholars' belief in the need for empirical tests, can lead to the devaluation of empirical evidence, and can reduce the credibility of research.

Conceptual ambiguity can lead to concept creep, when a concept becomes so broad and deep that non-examples are difficult to find. The methodological and real-world implications of conceptual ambiguity range from reduced trustworthiness of the research to the detachment of scholarship from policy-making and advocacy. For instance, a policy-maker is more likely to avoid "fuzzy concepts" when implementing policies because there are no clear definitions that can be used to structure their policy. That is, "fuzzy concepts" provide fuzzy guidance on how a policy will affect systemic power distributions, legal structures, and the actions of different people (e.g., advocates, politicians, lay people).

In the following sections, we attempt to include as many of the terms as we could find relevant to the interdisciplinary study of artificial entities' moral, social, and mental capacities.[3]

Note. We loosely group terms into four categories although several terms could belong to other categories.

Terminology and conceptual definitions

The first set of terms elicits a mental image of an entity's role in the world. These terms are most frequently used as nouns that can be modified with information about their material or psychological features, although some, like "super" and "transformative," typically serve as adjectives. "Machine" is used as a noun with feature modifiers (e.g., "digital machine," "autonomous machine"). However, it is also sometimes used as a feature-like modifier of other features (e.g., "machine intelligence"). Below are terms defining an entity's role.

[3] See the Appendix for a list of terms defining relevant fields of study.

Table 1: Terminology defining an entity's role

Entity's Role				
Term	Definition(s)	Benefits	Drawbacks	Source(s)
Agent	an interactive, autonomous, adaptable entity	- common usage in moral philosophy and moral psychology - clear pairing with "patient" - association with specific actions and intentions - applies to entities of any material composition	- moral consideration of agents not typically thought about - common usage includes multiple other associations (e.g., "government agent," "publicity agent")	Floridi & Sanders, 2004
Being	- M-W: "the quality or state of having existence" - OL: "existence"	- accessible - is concrete and brings to mind a specific meaning - commonly paired with "human" and "living" - already used to describe robots - avoids making worth dependent on moral action	- colloquial - may activate only certain conceptions of AI (e.g., has a spirit, capable of advanced cognition) - may imply needing to have a soul or metaphysical presence	Merriam-Webster, 2021; Oxford Languages, 2021
Beneficiary	derives well-being from resources	- clear connection to moral consideration - implies worthiness or social value	- limited usage in research - possible negative connotations (e.g., "welfare recipient" amongst humans)	Shulman & Bostrom, 2021

		- grants social status in human society	- possibly threatening to humans because of implied resource sharing	
Device	"a thing made or adapted for a particular purpose, especially a piece of mechanical or electronic equipment"	- enables discussion of mechanical or electronic safe and accurate functioning - easy to bring examples to mind (e.g., "toaster," "coffee maker," "vacuum")	- strong association with tool use (e.g., "consumer device," "technical device," "technological device," "smart device") - strong association with lack of agency	Oxford Languages, 2021
Entity	"a thing with distinct and independent existence"	- commonly used to refer to robots, AIs, and nonhuman animals - synonymous with "being" and "living" - does not require biological life - does not require a soul or metaphysical presence	- could refer to an individual, a group of individuals, or a corporation - abstract and difficult to mentally picture - may prompt objectification - can imply alienness	Oxford Languages, 2021
Life	"the condition that distinguishes animals and plants from inorganic matter, including the capacity for growth, reproduction, functional activity, and continual change preceding death"	- enables attribution of qualities typically reserved for biological, material entities - compatible with the "think, sense, act" robotics paradigm	- definition based on qualities of biological entities - may activate only certain conceptions of AI (e.g., has a spirit, capable of advanced cognition, derived from biological processes) - implies death or a permanent end-state	Oxford Languages, 2021

Machine	"an apparatus using or applying mechanical power and having several parts, each with a definite function and together performing a particular task"	- intended to apply to non-biological, constructed entities - implies complex composition and integration of cooperative internal systems	- common association with instrumental tool use - not necessarily digital or electronic (e.g., a tractor) - dissociated with experiential and affective capacities - denies possibility for human-like internal states (e.g., mechanistic dehumanization, "automatons") - common association with being rigid, cold, or inflexible	Oxford Languages, 2021
Patient	an entity who is acted on or responds to an action	- common usage in moral philosophy and moral psychology - clear pairing with "agent" - implies need for moral consideration	- strong association with field of medicine - could imply weakness or need for treatment due to medical association - may be more commonly associated with humans than other entities	Floridi & Sanders, 2004
Person ("legal person")	- person: "a human being regarded as an individual" - legal person: "an individual, company, or other entity which has legal rights and is subject to obligations"	- grants human-like status - may encourage support for AI rights - could be easily applied in virtual or digital worlds - can be applied to distinguishing individuals	- philosophical arguments against the term applied to artificial entities (e.g., Bryson, 2010) - strong association with conceptions of what it means to be a biological, corporeal human - may interrupt advocacy for the	Oxford Languages, 2021

			legal personhood of nonhuman animals (e.g., Nonhuman Rights Project)	
Robot	"machines using a sense, think, act paradigm to gather data about the environment, process the data autonomously, and act upon the world"	- well-established and commonly used - incorporates perceptual, cognitive, and behavioral capacities - applicable in many contexts (e.g., social, industrial) - arguments exist for the consideration of robot rights and duties	- requires a material body - moral consideration discussed primarily in relation to some sub-types (e.g., "social robots") - common association with instrumental uses (e.g., cleaning, factory production) - common association with being rigid, cold, or inflexible	Gunkel, 2018
"Super"	- OL: "especially; particularly" - Bostrom: far better or more advanced than a human	- used in common parlance - can be easily understood as a noun or verb modifier	- different meanings depending on context - possibly threatening to human uniqueness - limited usage across disciplines - implies a linear and hierarchical progression of capacities rather than a possible equivalency of capacities	Bostrom, 1998; Oxford Languages, 2021
System	"a set of things working together as parts of a mechanism or an interconnecting network"	- implies complex, dynamic, and integrated components that may promote moral consideration - does not require a material	- dissociated with experiential and affective capacities - not typically associated with moral consideration - can indicate intraindividual	Oxford Languages, 2021

		body - offers a holistic image of a complex entity	components or multiple entities	
Target	“an area or object that is the focus of a process, inquiry, or activity”	- implies a recipient who could receive moral consideration - no valence	- nebulous applications and associations (e.g., the superstore, for an arrow, of social derision or lauding) - directionality may preclude conceptions of relationality	American Psychological Association, 2021
Transformative	“causing a marked change in someone or something”	- sounds positive - implies a changing outlook or world that might enable moral consideration - implies advanced features that may enable moral consideration	- not typically associated with moral consideration - not typically associated with agentic, cognitive or experiential, affective capacities - may be more referential to an abstract system or state rather than an entity or individual - may be focused on effects on humans	Oxford Languages, 2021

The second set of terms evokes a mental image of an artificial entity's material structure. These terms focus either on what the entity is made from or how they exist. They often modify terms defining an entity's role (e.g., "artificial agent"). Below are terms defining material features.

Table 2: Terminology defining material features

Material Features				
Term	Definition(s)	Benefits	Drawbacks	Source(s)
Artificial	"made or produced by human beings rather than occurring naturally, especially as a copy of something natural"	- intended to apply to constructed phenomena - well-known and widely used - allows for a broad set of possible features and material composition - distinguishes between biological or natural and non-biological or unnatural - can be used to refer to partly non-biological and partly biological phenomena	- alternate meaning may imply "fakeness" - may imply a dichotomy of value for natural and unnatural - may imply a filial or "paternal" relationship to humans - may convey status benefits to the constructors (e.g., for the prowess of invention) without parallel benefits for the constructed	Oxford Languages, 2021
Cyber	"relating to or characteristic of the culture of computers, information technology, and virtual reality"	- embodiment not required - emphasis on computers, information, and virtual reality rather than on humans	- less commonly used - some associations may be negative (e.g., "cybersecurity," "cyberwarfare") - may sound outdated to experts and the general public	Oxford Languages, 2021
Digital	"(of signals or data) expressed as series of the digits 0 and 1, typically represented by values	- intended to apply to non-biological phenomena - illustrative of computer	- not considerate of embodiment - implies "cold cognition,"	Oxford Languages, 2021

	of a physical quantity such as voltage or magnetic polarization"	programs and algorithms - does not require embodiment	mechanical properties, and automation - lower similarity to "humanness" may limit moral consideration	
Electronic	"(of a device) having or operating with the aid of many small components, especially microchips and transistors, that control and direct an electric current"	- intended to apply to non-biological phenomena - illustrative of technology	- close association with tools like clocks or game consoles may limit moral consideration - implies a material structure or composition	Oxford Languages, 2021
Embodied	- Körner et al.: "the body, its sensorimotor state, its morphology, or its mental representation play an instrumental role in information processing" - OL: "provide (a spirit) with a physical form"	- applicable to various entities - includes mental and physical states, feedback between the two, and both perceptual and cognitive sources of information - higher similarity to "humanness" may increase moral consideration	- requires a material container or body - does not distinguish material composition - relationship to entity's role is unclear - difficult and sometimes nonsensical to pair with psychological features - may imply a dichotomy of value for embodied and non-embodied entities	Körner et al., 2015; Oxford Languages, 2021
Non-biological	"not involving or derived from biology or living organisms"	- differentiates entities traditionally labelled as "living" from those traditionally labelled as "non-living"	- defined in relation to biological entities - potentially meaningless in systems or worlds without any "biological" components	Oxford Languages, 2021

		- removes the value judgment of "living" - not considered "fake" or "unnatural" - allows for a broad set of possible features and material composition - distinguishes between biological or natural and non-biological or unnatural	- may preclude conceptions of hybrid biological and non-biological systems - difficult to apply to artificial entities derived from evolutionarily biological processes (e.g., whole brain emulations) - may be problematic in future scenarios with non-carbon-based biological lifeforms	
Synthetic	"(of a substance) made by chemical synthesis, especially to imitate a natural product"	- distinctly non-natural or non-biological	- defined in relation to "naturalness" - may imply "fakeness" - subject to value judgments about "naturalness" and "fakeness"	Oxford Languages, 2021
Virtual	- "not physically existing as such but made by software to appear to do so" - "almost or nearly as described, but not completely or according to strict definition"	- does not require a specific material composition or embodiment - is commonly used and understood - is associated with conceptions of digital worlds or spaces	- exclusive of entities with a material body - may prompt people to think only of entities living in virtual environments that are less accessible and routinely salient (at least in the near-term) - can have multiple meanings that have different implications for moral	Oxford Languages, 2021

			consideration - is abstract and may make it difficult to form a concrete mental image of what it means - not distinctly nonhuman - subject to value judgments about the nature of reality and "fakeness"	

A third set of terms defines the psychological features of an entity. These terms operationalize abstract and indirectly observable mental phenomena (e.g., intelligence, friendliness). Psychological features are inferred from observing complex criteria agreed upon by experts. For example, we cannot observe “autonomy” directly. We infer “autonomy” from introspection (i.e., self-reports) and behavior. These terms differ from material features because material features can be directly observed. Below are terms defining psychological features.

Table 3: Terminology defining psychological features

Psychological Features				
Term	Definition(s)	Benefits	Drawbacks	Source(s)
Autonomy	- Darling: "the ability to 'make (limited) decisions about what behaviors to execute based on perceptions and internal states, rather than following a pre-determined action sequence based on pre-programmed commands'" - F&S: "perform internal transitions to change its state" - Keller: "self-government and responsible control for one's life" - M&R: "sense of volition and internal perceived locus of causality in one's undertakings...actions emanate from the self and reflect who one really is, instead of being the result of external pressures"	- implies independent, individual capacities - connected with morality - has been important for establishing criteria for understanding well-being and welfare in human and nonhuman animals - contrast with "heteronomy," or actions occurring because of external demands, may increase moral consideration and belief that coercion of artificial entities is wrong	- relationship to moral agency and patiency is unclear - unclear how cognitive (agentic) and affective (experiential) capacities relate to autonomy - sometimes used to refer to "human-less" control without requiring fully non-pre-programmed sequences or internally derived motivation and behavior	Darling, 2016; Floridi & Sanders, 2004; Keller, 2016; Martela & Riekki, 2018
Consciousness ("subjective,"	- Block: "the phenomenally conscious aspect of a state is	- long tradition of scholarship - well-known	- multitude of definitions reduces clarity when the	Block, 1995; Muehlhauser, 2017;

"phenomenal," "access")	what it is like to be in that state. The mark of access-consciousness, by contrast, is availability for use in reasoning and rationally guiding speech and action." - Muehlhauser: subjective experience - SEP: 6 possible senses of consciousness (sentience, wakefulness, self-consciousness, "what it is like," conscious states, transitive consciousness); 6 possible states of consciousness (mental state awareness, qualitative/experiential states, phenomenal states, "what it is like," access, narrative)	- focuses on describing the internal processes and experiences of an entity	term is used - lacking scholarly consensus on what it is and where it comes from - difficult to observe and verify from a first person perspective - may be controversial with the general public when applied to artificial entities	Stanford Encyclopedia of Philosophy, 2014
Ethical	- OL: "Relating to moral principles or the branch of knowledge dealing with these" - Velasquez et al.: "based on well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or	- clearly related to morality - can be applied to conceptions of agency and patiency - can be used when describing the safe functioning of artificial entities - synonymous with "moral"	- meaning depends on social and cultural context - complicated philosophical underpinnings - more often used to describe artificial entities with moral agency than moral patiency	Oxford Languages, 2021; Velasquez et al., 2010

	specific virtues”			
Friendly	- F&P: “include behavior that displays sincere well-wishing, the intrinsic valuing of the other, the commitment to honesty, loyalty and other shared values” - Reisman: “‘friendliness’ refers to a set of behaviors, such as seeking the company of others, smiling, greeting, rewarding, sharing, cooperation”	- expresses desirable, safety-oriented behavior - focuses on relationality	- context specific - only applies to one dimension of behavior and relationality - meaning may be likely to change over time	Fröding & Peterson, 2020; Reisman, 1984
Intellect	“an individual’s capacity for abstract, objective reasoning, especially as contrasted with his or her capacity for feeling, imagining, or acting”	- understood as an aspect of intelligence - indicates “smartness” or “wit” - a dimension of “mind”	- cognitive only and thus may apply primarily to perceptions of agency - association with “intelligence” may make redundant or useful only in some contexts	American Psychological Association, 2021
Intelligence (“general,” “fluid,” “crystallized,” “visual-spatial reasoning,”	- Hunt: “Fluid intelligence is the ability to develop techniques for solving problems that are new and unusual, from the perspective of the problem solver. Crystallized intelligence	- long tradition of scholarship - well-known - already used extensively with artificial entities	- multitude of definitions reduces its conceptual clarity - lacking scholarly consensus on what it is and where it comes from - often intentionally	Hunt, 1995; Muehlhauser, 2013; Sternberg, 1986; Weiss et al., 2019

"triarchic," "an agent's power to optimize the world according to its preferences"	is the ability to bring previously acquired, often culturally defined, problem-solving methods to bear on the current problem. Visual-spatial reasoning is a somewhat specialized ability to use visual images and visual relationships in problem solving" - Muehlhauser: "'optimization power' concept of intelligence…an agent's power to optimize the world according to its preferences" - Sternberg: a three part theory of intelligence consisting of internal mechanisms (for learning, planning, doing), dealing with novelty and automation, and external behavior guidance based on adaptation, selection, and shaping processes - Weiss et al.: "General intelligence is the fluid ability to integrate multiple cognitive abilities in the service of solving a novel problem and thereby accumulating crystalized		dissociated by experts from experiential capacities like emotion, sentience, and consciousness - often thought of with human-like intelligence as the standard	

	knowledge that, in turn, facilitates further higher-level reasoning."			
Mind	"...all intellectual and psychological phenomena of an organism, encompassing motivational, affective, behavioral, perceptual, and cognitive systems; that is, the organized totality of an organism's mental and psychic processes and the structural and functional cognitive components on which they depend. The term, however, is also used more narrowly to denote only cognitive activities and functions, such as perceiving, attending, thinking, problem solving, language, learning, and memory."	- implies all non-material aspects of a complex system instantiated in a material structure - implied connection with biological life may enable moral consideration - common interdisciplinary use and understanding - can apply to entities with a central processing unit instantiated on any material	- sometimes used to signify only "cold" cognitive processes like rationality, memory, and reason rather than "hot" cognitive processes related to motivation and emotion - weak link with moral consideration - existence of mind sometimes conflated with existence of consciousness	American Psychological Association, 2021
Moral	"a code of conduct that would be accepted by anyone who meets certain intellectual and volitional conditions, almost always including the condition of being rational"	- clearly related to morality - can be applied to conceptions of agency and patiency - synonymous with "ethical"	- depends on social and cultural context - complicated philosophical underpinnings - more often used to describe artificial entities with moral agency than moral patiency	Stanford Encyclopedia of Philosophy, 2020

Qualia	"the introspectively accessible, phenomenal aspects of our mental lives"	- focuses on describing the internal processes of an entity - points to experiential capacities	- referential to "consciousness" - uncommon usage - limited interdisciplinary usage - comprised of many sub-components (e.g., thoughts, feelings) that may reduce its conceptual clarity - difficult to observe and verify from a first person perspective	Stanford Encyclopedia of Philosophy, 2017
Sentience	- Broom: "the capacity to have feelings, which includes the ability to evaluate the actions of others in relation to oneself and third parties, to remember some of one's own actions and their consequences, to assess risks and benefits and to have some degree of awareness" - DeGrazia: "capable of having feelings (mental states, such as sensations or emotional states, that are typically pleasant or unpleasant)" - H&A: "the capacity for positive and negative	- centers experiential capacities - highlights valence of experiences - highlights affective capacities - implies worthiness of moral consideration - may be indicated by certain third person perspective observable features	- overlap with some theories and definitions of consciousness - possible conflation with "consciousness," especially amongst members of the general public -possible conflation with "sapience," or the idea of wisdom and exaggerated intelligence - difficult to observe and verify from a first person perspective	Broom, 2020; DeGrazia, 1996; Harris & Anthis, 2021

	experiences”			
Smart	- OL: “(of a device) programmed so as to be capable of some independent action” - Silverio-Fernández: defined by connectivity with others for data exchange, autonomous computing, and context-awareness	- commonly understood to mean “intelligent” - can imply autonomy - sometimes used in relation to technological systems	- does not necessarily imply relationality with humans - often applied to technology with limited intelligence (e.g., “smart TV”) - no relationship to moral consideration	Oxford Languages, 2021; Silverio-Fernández et al., 2018

Combinations of feature and role terms have been used within and across many fields of study[4] to describe artificial entities. For instance, "friendly AI" has been used to describe AIs who mimic human friendliness and are benign to humans. "Moral machine" has been used to describe machines that make ethical decisions or need to solve moral dilemmas, like self-driving cars. Some of these combinations are more widely used and well-known. Other combinations have stronger implications for taking moral action and receiving moral consideration. Many combinations are used only in specific contexts, based on convenience, or merely as placeholders. Below we outline some combinations that we think are consequential and related to the moral consideration of artificial entities.

[4] See the Appendix for a list of terms defining relevant fields of study.

Table 4: Consequential combinations of features and role

Consequential Combinations				
Term	Definition(s)	Benefits	Drawbacks	Source(s)
Artificial agent	an interactive, autonomous, adaptable entity instantiated at least partly on non-biological substrates	- grants capacity for taking independent, self-controlled actions to at least partly non-biological entities - clearly identifies that the entity is at least partly non-biological - addresses relationality - could be used as a general term for a class of entities built (at least initially) by humans	- may limit moral consideration because of the emphasis on agency - relationality is not clearly linked to moral consideration - may limit humans' conceptions of the capacities of these entities - subject to concerns about "fakeness"	Floridi & Sanders, 2004
Artificial being	a distinct and independent existence instantiated at least partly on non-biological substrates	- easily understandable - familiar - implies potential for experiential capacities based on "life-like" qualities of "being" - clearly identifies that the entity is at least partly non-biological	- may have different meanings across contexts - closely related to legal concepts like "artificial person" - colloquial - subject to concerns about "fakeness" - may imply requirement of a soul or metaphysical	-

			presence	
Artificial consciousness	- Graziano: "a machine that contains a rich internal model of what consciousness is, attributes that property of consciousness to itself and to the people it interacts with, and uses that attribution to make predictions about human behavior. Such a machine would 'believe' it is conscious and act like it is conscious, in the same sense that the human machine believes and acts" - Reggia: "computational models of various aspects of the conscious mind, either with software on computers or in physical robotic devices"	- implies capacity for introspection, self-awareness, and consciousness in a human-built (at least initially) entity akin to that present in humans - emphasizes internal experiences - established and already used in scholarship	- subject to the same philosophical and empirical criticisms about "consciousness" - unclear what material components (e.g., body) are necessary - complex relationship to moral consideration - potential association with "fake" or "unnatural" - requires evidence of consciousness - may be controversial in use with the general public - difficult to observe and verify from a first person perspective	Graziano, 2017; Reggia, 2013
Artificial entity	a distinct and independent existence instantiated at least partly on non-biological substrates	- easily understandable - familiar - reflects diversity in possible types of entities - clearly identifies that the entity is at least partly non-biological	- abstract - may have different meanings across contexts - used commonly to refer to corporations - subject to questions about "fakeness"	-

		- not tied to existence of a soul or metaphysical presence	- may prompt objectification	
Artificial general intelligence	"a software program that can solve a variety of complex problems in a variety of different domains, and that controls itself autonomously, with its own thoughts, worries, feelings, strengths, weaknesses and predispositions"	- established and commonly used - may prompt increased similarity to humanity and thus make it easier to promote moral consideration - emphasizes internal phenomena (e.g., software programming, mental capacities) rather than external properties (e.g., material structure, embodiment)	- implicit connection with human-like or a human-like basis for intelligence requires humans to be the initial standard - potential association with "fake" or "unnatural" - limited by existing conceptual bounds of AIs as cognitively skilled machines - definition susceptible to meaning changes over time as technology and conceptions of "intelligence" change - may require evidence of self-awareness or consciousness	Goertzel & Pennachin, 2007
Artificial intelligence	"systems that can decide what to do and do it" and that vary based on how human-like and/or rational their thought processes and/or behavior are	- well-known and commonly used - entails specific capacities (e.g., memory, learning, reasoning)	- used interchangeably to signify the field of study, an individual entity, or a network of entities - potential association with "fake" or "unnatural" - established usage is conflated with sophisticated	Russell & Norvig, 1995

			cognition - depends on conceptions of human-like intelligence - definition susceptible to meaning changes over time as technology and conceptions of "intelligence" change - not typically connected with moral consideration - may be prone to hype	
Artificial moral agent	- Cervantes et al.: "artificial agents capable of making ethical and moral decisions" - Himma: "an [artificial] agent with the capacities to choose its actions 'freely' and understand the basic concepts and requirements of morality, capacities that also presuppose consciousness"	- promotes association of artificial entities with morality - clearly signifies an individual - entails specific types of decisions and actions - considers internal experiences of artificial entities	- may be associated only with taking moral action rather than with receiving moral consideration - may require evidence of consciousness - context specific moral boundaries may introduce safety concerns or present a threat to humans	Cervantes et al., 2020; Himma, 2009
Artificial sentience	- artificial entities with the capacity for positive and negative experiences, such as happiness and suffering - the capacity for positive and negative experiences manifested in artificial	- clear connection to moral consideration from the nonhuman animal research on sentience - represents specific, valenced experiential capacities - does not require human-like	- could signify an individual entity, a collection of entities, a psychological/phenomenal experience, or a field of study - little established scholarly	Sentience Institute, 2021

	entities	intelligence - dissociable from cognitive capacities like problem-solving and analytical thinking - may be more distinctive than "mind" or "consciousness" - prioritizes perception, emotion, and behavior	or applied usage - potential association with "fake" or "unnatural" - potential conflation with "artificial consciousness" due to historical philosophical link between "sentience" and "consciousness" - difficult to observe and verify from a first person perspective	
Digital mind	- S&B: "machine minds with conscious experiences, desires, and capacity for reasoning and autonomous decision-making" - Sotala: "a mind that runs on a computer"	- bridges interdisciplinary conceptions of "mind" (e.g., psychology, artificial intelligence) - inclusive of a range of machine-based entities (e.g., whole brain emulations, AI algorithms) - existing theorizing on moral status and the potential to experience suffering - highlights non-material aspects of entities	- some research uses this term to refer to human brains interacting with digital media - tied to instrumental use purposes (e.g., "digital mind mapping," "digital mind games") - implied connection with the human brain or systems initially modelled on biological brains	Shulman & Bostrom, 2021; Sotala, 2012
Digital person	- EA: "a person running on digital computing hardware" - Karnofsky: digital copies of	- connection to "humanness" and human capacities like sentience that may increase	- implied connection with the human brain or systems initially modelled on	Effective Altruism Forum, 2021; Karnofsky, 2021

	humans such as simulated human brains and digital descendants of humanity	moral consideration - framework may make the concept more accessible to the general public - highlights non-material aspects of entities	biological brains - excludes many types of artificial entities, notably those who are less human-like or not descended from humans - the future framing of the concept may reduce its near- and medium-term impact	
Full ethical agent	"can make explicit ethical judgments and generally is competent to reasonably justify them"	- implies specific and directional action tied to morality - can be used in connection with making moral judgments - implies greater depth of agency with the term "full" - can apply to any entity with the capacity to make ethical judgments, human or nonhuman	- does not uniquely distinguish artificial entities and their capacities - does not consider the perceptual, affective, and experiential capacities of the entity - not commonly used - may promote a false binary distinction between moral agency and moral patiency	Moor, 2006
Intelligent systems	"a tool that (1) operates in a complex world with limited resources (2) possesses primary cognitive abilities such as perception, action control, reasoning, or language use, and (3) exhibits complex intelligent behavior	- commonly used and understood to signify AI - implies multiple cooperating components that may be important for moral consideration - entails a detailed breakdown of capacities	- applicable primarily to describing cognitive capacities; little consideration of perceptual, affective, or experiential capacities - depends on conceptions of human-like intelligence	Molina, 2020

	supported by abilities such as rationality, adaptation through learning, or the ability to explain the use of its knowledge by introspection"	- implies internal complexity	- strong associations with instrumental purposes (i.e., as tools) - little connection with moral consideration	
Machine superintelligence	"general artificial intelligence greatly outstripping the cognitive capacities of humans, and capable of bringing about revolutionary technological and economic advances"	- "machine" implies some sort of material structure - points to societal ramifications of AI - points to a minimum of general intelligence rather than intelligence only in specific tasks, processes, or certain domains	- could be criticized for not incorporating considerations of reinforcement learning AI (e.g., like in this podcast) - uses human intelligence as the standard of comparison - experiential, affective capacities typically not discussed which may limit moral consideration - could threaten human uniqueness and resources - implies a linear and hierarchical progression of capacities rather than a possible equivalency of capacities	Bostrom et al., 2020
Moral agent	an entity who can take moral action	- specific and directional action tied to morality - can be used in reference to taking moral action - widely used across multiple	- does not uniquely distinguish artificial entities and their capacities - may imply less worthiness of moral consideration	Floridi & Sanders, 2004

		disciplines (e.g., psychology, philosophy, AI, HRI, HCI) - can apply to any entity with the capacity to take moral actions, human or nonhuman		
Moral patient	an entity who can receive moral action	- specific and directional action tied to morality - can be used in reference to granting moral consideration - widely used across multiple disciplines (e.g., psychology, philosophy, AI, HRI, HCI) - can apply to any entity with the capacity to receive moral actions, human or nonhuman	- does not uniquely distinguish artificial entities and their capacities - may imply less potential for moral agency	Floridi & Sanders, 2004
Non-biological sentience	- non-carbon-based entities without biological qualities who have the capacity for positive and negative experiences, such as happiness and suffering - the capacity for positive and negative experiences manifested in non-biological entities	- clear connection with moral consideration from the nonhuman animal research on sentience - represents specific, valenced experiential capacities - does not require human-like intelligence - dissociable from cognitive capacities like problem-solving and analytical thinking - may be more distinctive than “mind” or “consciousness”	- could signify an individual entity, a collection of entities, a psychological/phenomenal experience, or a field of study - little established scholarly or applied usage - difficult to observe and verify from a first person perspective - may increase the focus on material feature or substrate	-

		- prioritizes perception, emotion, and behavior - may avoid issues of "fakeness" or "unnaturalness" - not subject to the hype associated with AI - less subject to the value judgment of "living"	rather than psychological or perceptual features that increase moral consideration - may be problematic in future scenarios with non-carbon-based biological lifeforms - difficult to apply to artificial entities arising from evolutionarily biological processes (e.g., whole brain emulations) - may exclude hybrid biological and non-biological systems - potentially meaningless in systems or worlds without any "biological" components	
Smart device	"context-aware electronic device capable of performing autonomous computing and connecting to other devices wire or wirelessly for data exchange"	- decent shorthand to distinguish devices with some level of intelligent response from entirely non-intelligent mechanical or electronic devices - emphasizes autonomy - emphasizes relationality with other artificial entities	- used primarily in instrumental contexts (e.g., technological tools) - diversity of included entities with varying levels of intelligent capacities may weaken arguments for moral consideration (e.g., personal assistance device, smart TV, smart phone)	Silverio-Fernández et al., 2018

			- used largely in human consumer contexts	
Super-beneficiary	"a being that is superhumanly efficient at deriving well-being from resources"	- clearly implies a need for moral consideration - prioritizes well-being - extends conceptions of beneficiaries to artificial entities	- potentially threatening to the needs and resource allocation of humans and nonhuman animals - "utility monster" - may be susceptible to the stereotype of welfare recipients who make use of the system for personal gain rather than because of need - may lead to negative consequences like increased prejudice and moral exclusion - implies a linear and hierarchical progression of capacities rather than a possible equivalency of capacities	Shulman & Bostrom, 2021
Super-patient	"a being with superhuman moral status"	- implies a need for moral consideration - extends moral patiency to intelligent artificial entities	- potentially threatening to the needs and resources of humans and nonhuman animals - may lead to negative consequences like increased prejudice and moral	Shulman & Bostrom, 2021

			exclusion - implies a linear and hierarchical progression of capacities rather than a possible equivalency of capacities	
Transformative artificial intelligence	"AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution"	- intended to include many possible artificial entities - intended to be agnostic to human-likeness - implies an advance in the capacities of artificial entities that may enable increased moral consideration	- might be limited to artificial entities in only some contexts (e.g., AIs working in data science) - may threaten human resources or uniqueness - depends somewhat on conceptions of human-like intelligence - experiential, affective capacities not integral to the concept - framed around transformation of current human society rather than transformation of artificial entities	Karnofsky, 2016

What term should we use?

Is the same terminology useful for the general public, engineers, scientists, ethicists, lawyers, and policy-makers? There's some suggestion that the general public and experts differ in their understanding of artificial entities and support for their rights. Would a common, consensual terminology enable more effective interdisciplinary research, policy-making, and advocacy? If having an imprecise definition is preferred, how can differences in terminology be reconciled to maximize the clarity and utility of the terms?

Whether or not terms generalize across contexts might matter. Should we use terminology that can be applied to humans, nonhuman animals, algorithms, and machines? Is it necessary to specify an entity's role? Several of the consequential combinations do not (e.g., "artificial general intelligence," "artificial sentience"). Does the psychological feature need to be narrow or broad? "Consciousness" is broad, but it is also "fuzzy" because of the many conceptualizations in current usage. On the other hand, a narrow term like "friendly" only applies to one aspect of an entity's behavior and may not retain meaning over time and across context.

Below we explain our preference for "artificial sentience," consider some possible reasons not to use "artificial sentience," and consider using multiple terms.

Reasons for "artificial sentience"

We favor the term "artificial sentience" for the following **linguistic** reasons:

- "Artificial sentience" sits in a "Goldilocks Zone" of broad and narrow terminology. "Artificial" is inclusive of many entities and distinguishes entities composed at least partly of non-biological substrates from completely biological entities. "Artificial" is commonly used and understood in relevant contexts. "Sentience" is an aspect of "mind" that can be differentiated from broad and narrow psychological features such as "mind" and "intelligence," respectively.
- "Artificial sentience," as a newer term, does not have divergent meanings across time and field of study like the dual meanings that exist for some terms (e.g., "digital mind"). Scholars concerned with the effects of AIs on human society have used "digital mind" to refer to psychological minds that exist purely on a computer or in a digital space. The term is also associated with human brains and the dynamics of digital technologies. In this context, a "digital mind" is a human brain on digital media.
- "Artificial sentience" uses similar language to "artificial intelligence" and is likely to be conceptually associated with it. "Artificial intelligence" has a well-established meaning and is commonly used by the general public and experts. We believe that similarity to the well-known "artificial intelligence" will enable experts and the public alike to grasp the meaning of "artificial sentience" as an "experiential" extension of "artificial intelligence."
- There is little reason to expect the term "artificial sentience" to be associated with God-like terms such as "machine superintelligence" that reduce moral consideration by dint of being too threatening or tool-like terms such as "smart device" that prompt instrumental mental images. The common conception of artificial entities as technological tools may limit the extent to which they are morally considered and overcoming these associations

with existing terminology may be difficult.

We favor the term "artificial sentience" for the following **conceptual** reasons:

- "Artificial sentience" avoids some of the conceptual difficulties associated with "artificial consciousness." Some stances within philosophy, including a position held at SI, argue that the construct of "consciousness" and some of its underlying mental circuitry (e.g., how introspection arises from neural and cognitive systems) are particularly ill-defined and difficult to observe.
- "Artificial sentience" is inclusive of entirely non-biological entities, hybrid biological and non-biological entities, and artificial entities originating from evolutionarily biological processes like whole brain emulations. This inclusivity may increase the tractability of advocating for the moral consideration of artificial entities in a way that terms like "non-biological sentience" may not.
- The term "artificial sentience" prioritizes "sentience," an affective capacity, as the key distinguishing aspect of mind critical to experiential capacities like motivation and affect. Although "mind" encompasses all mental capacities, it is sometimes used to signify only cognition, reasoning, and rationality. This usage of "mind" is often implicitly considered "cold," controlled, and unemotional, which can lead to mechanistic forms of dehumanization and moral circle exclusion.
- "Sentience" requires features related to having affective experiences (e.g., the capacity to have positive and negative experiences). This definition may make "sentience" less subject to "fuzziness" than other psychological features that are difficult to define and observe like "qualia."
- "Artificial sentience" may be less threatening to people than terms like "super-beneficiary," "digital mind," or "artificial consciousness." "Super-beneficiary" may prompt realistic threats over resource sharing. "Digital mind" and "artificial consciousness" may prompt symbolic threats to human distinctiveness because of the association of "consciousness" and "mind" with humans' sophisticated mental capacities. "Artificial sentience" may be less threatening because "sentience" is not unique to humans and because "sentience" does not necessarily imply sharing resources or social status.

We favor the term "artificial sentience" for the following **moral** reasons:

- "Artificial sentience" connotes the aspect of artificial entities' minds that we believe is critical to moral consideration. Affective components of mind like sentience are linked to increased moral consideration. Cognitive components of mind like memory are typically associated with the capacity to take moral action.[5] We believe that this may be an important distinction for moral consideration given the potential need to distinguish between future "digital minds" who have some degree of sentience and "digital minds" that are cognitively sophisticated but devoid of experiential capacities.
- "Sentience" is the psychological feature most clearly related to the moral consideration of nonhumans. Specifically, "sentience" has a clear track record in nonhuman animal welfare science and advocacy for moral consideration and has some backing within AI ethics.

[5] Summarizing from Gray et al. (2007) and Gray et al. (2012).

"Consciousness" has a more convoluted relationship with moral consideration.

- There are clear consequences of sentience dismissal for humans (e.g., the Atlantic slave trade) and nonhumans (e.g., factory farming). The capacity to suffer is denied with sentience dismissal, facilitating exclusion from the moral circle. "Artificial sentience" is likely to be critical for recognizing the moral status of sentient artificial entities in a way that consequential combinations like "artificial consciousness" are not.

Reasons against "artificial sentience"

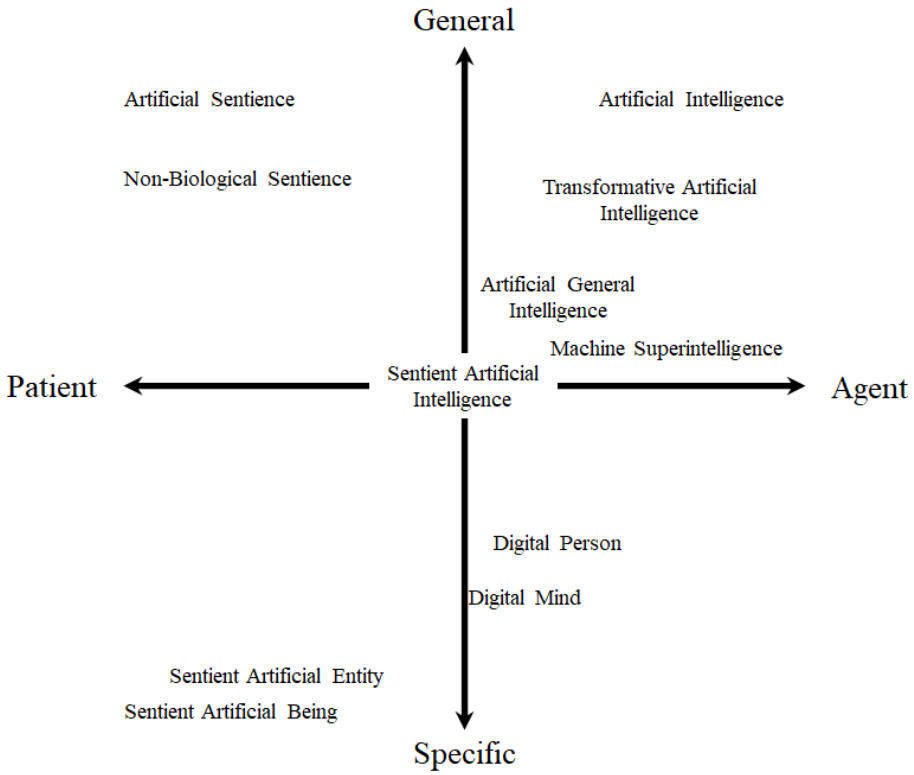
- The use of "artificial" may lead to some associations with "fakeness" or "unnaturalness." This could be a concern if it leads to the dismissal of artificial entities' capacity to have so-called "real" experiences.
- "Artificial sentience" may be more threatening than terms like "non-biological sentience"[6] if people associate "artificial" with being constructed in a laboratory for nefarious purposes or if they associate "artificial intelligence" with dangers outside of human control.
- The use of "sentience" is sometimes conflated with "sapience," or conceptions of exaggerated intelligence, wisdom, and reasoning. This could be a concern for promoting understandings of what artificial sentience entails.
- "Artificial sentience" is new and unfamiliar to the general public and many experts. This could lead to some initial confusion. The longevity of the term could also be questioned. Given that it is new, we cannot yet know whether or not it will endure meaningfully into the future.
- "Artificial sentience" has some philosophical overlap with "artificial consciousness" that may increase confusion over the meaning of the term when used in interdisciplinary or transdisciplinary contexts. This overlap may also mean that some critiques of "artificial consciousness" may apply to conceptions of "artificial sentience" that are too broadly defined (e.g., broader than the capacity for positive and negative experiences).
- The connection between "artificial intelligence" and "artificial sentience" might create a spillover of hype from AI. Ungrounded excitement could prompt increased efforts to develop artificial sentience without enough forethought and preparation for a world with sentient artificial entities who may or may not receive moral consideration.
- "Artificial sentience" is likely to require evidence of the presence of a number of relevant features (e.g., those related to detecting harmful stimuli, behavioral avoidance, centralized information processing) that may enable us to make judgments about the likelihood and degree of sentience. This could enable a probabilistic approach for operationalizing "artificial sentience" that might increase our chances of correctly judging whether an artificial entity is sentient (advancing the moral consideration of artificial entities). However, having to calculate the probability of artificial sentience may make it easier for some to dismiss the existence of "artificial sentience" given that it may reduce confidence in the concept by providing too much evidence.

[6] A term like "non-biological sentience" might contrast better with "biological sentience," increasing the focus on "sentience" rather than on the entity's material features and substrate (i.e., whether they are a human, a nonhuman animal, algorithm, or machine). Reducing the emphasis on substrate may facilitate attempts to increase the moral consideration of artificial entities.

Using multiple terms

The best strategy might be to use multiple terms to represent the interests of algorithmically-based, at least partly non-biological entities, of which artificial sentience (AS) may eventually be considered an umbrella term. Below is a possible initial taxonomy of terms based on their relationship to the moral referents of patiency and agency and to the specificity of the term.

Figure: Possible Taxonomy of Terms



Note. Terms are positioned along two (of many possible) dimensions. Moral references are on the x-axis and specificity is along the y-axis. These locations are imprecise as they reflect cultural and intellectual associations that are likely to vary across readers and change over time.

Possible Uses for Specific Terms

- "Sentient artificial being" and "sentient artificial entity" are likely to be useful for communicating about individual AIs with non-experts or in discussions where the emphasis is on sentience rather than intelligence.
- "Sentient artificial intelligence" ("sentient AI") is likely to be useful for communicating with experts and non-experts about individual AIs with some degree of sentience or the capacity for sentience in AI.
- "Digital mind" is likely to be useful for emphasizing internal, algorithmic mental capacities rather than substrate-based material structures.

- "Digital people"[7] is likely to be useful when referring to the digital nature of human descendants and the sorts of societies they may live in.

[7] See Holden Karnofsky's Cold Takes blog for more on "digital people."

Appendix: Terminology defining relevant fields of study

Some terms define the academic fields of study surrounding artificial entities. The fields described below seem likely to have the greatest impact on long-term outcomes of artificial entity development such as how they will be designed and the resulting ethical implications.

Field of Study				
Term	Definition(s)	Benefits	Drawbacks	Source(s)
Computer ethics	- Maner: "ethical problems aggravated, transformed or created by computer technology" - Moor: contemplating the social and ethical use of information technology in order to inform policy - Bynum: "identifies and analyzes the impacts of information technology on such social and human values as health, wealth, work, opportunity, freedom, democracy, knowledge, privacy, security, self-fulfillment, etc."	- can be applied to protect humanity from extinction risks - can be applied to policies aimed at the social integration of artificial entities	- focused on impacts to human society - less relevant to the moral consideration of nonhumans - has a diversity of meanings with some emphasizing policy and others emphasizing ethics	Bynum, 2004
Machine ethics	"...concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable"	- centers machine behavior instead of human behavior - allows for the moral consideration of machines - can be applied to protect humanity from extinction risks	- the moral consideration of humans is the primary outcome of concern - only machines', not humans', moral consideration of other machines is considered	Anderson & Anderson, 2007; Stanford Encyclopedia of Philosophy, 2020

			- less emphasis placed on institutional behavior and more emphasis on individual behavior	
Robot ethics ("roboethics")	focused on the ethical and social implications and consequences of advanced robotics particularly in regards to safety and errors, law and ethics, and social impact	- includes human and robot perspectives - can be applied to protect humanity from extinction risks	- applicable to robots only (material body required that can sense and act on the world) - the moral consideration of humans is more central than the moral consideration of robots	Gunkel, 2018; Lin et al., 2011; Lin et al., 2011; Scheutz, 2013
Artificial intelligence	to understand and build intelligent entities	- well-known - encompasses many exemplars of artificial entities; broadly inclusive - encourages interdisciplinary contributions and collaborations	- also used to refer to individual entities - focused on cognitive capacities specifically around human-like intelligence defined by problem-solving and analytical thinking - has different meanings for experts and the general public - prone to hype	Russell & Norvig, 1995
Cybernetics	the study of feedback, human behavior, and information to understand communication and control in human-machine	- defined by the relationality of humans and machines - specialized focus on	- implies material integration of humans (biological) and machines (non-biological) - machines not judged as	Mindell, 2004; Wiener, 1948

	relationships	dynamic, feedback-based systems - safety emphasis on producing specific outputs from specific inputs	equivalent in value to humans - less commonly used than other field names	
Human - computer interaction	"...a subfield within computer science concerned with the study of the interaction between people (users) and computers and the design, evaluation and implementation of user interfaces for computer systems that are receptive to the user's needs and habits. It is a multidisciplinary field, which incorporates computer science, behavioral sciences, and design. A central objective of HCI is to make computer systems more user-friendly and more usable."	- focuses on the relationality of humans and computers - incorporates research from multiple perspectives and disciplines - does not require a robotic body that can sense and act on the world - may be more inclusive of various types of AIs (e.g., algorithms, robots)	- focuses on machines as instrumental tools - benefits of relationality are one-sided (i.e., for humans) - less emphasis on ethics or moral consideration - largely interpersonal, rather than societal, level of study and impact	Brey & Søraker, 2009
Human - robot interaction	"...a challenging research field at the intersection of psychology, cognitive science, social sciences, artificial intelligence, computer science, robotics, engineering and human-computer interaction. A primary goal of research in this area has been to investigate 'natural' means by which a human can	- focuses on the relationality of humans and robots - incorporates research from multiple perspectives and disciplines - definition emphasizes dual nature of interaction and communication	- less emphasis on ethics or moral consideration - largely interpersonal, rather than societal, level of study and impact - practice focuses on tailoring robots to suit human needs - the artificial entity must	Dautenhahn, 2007

	interact and communicate with a robot. Due to the embodied nature of this interaction, where robots and humans need to coordinate their activities in time and space in real-time, often ‘face-to-face’, the quality of these interactions is related to, but different from e.g. human-computer interaction (HCI).”		have a robotic body or material structure that can sense and act on the world	

References

American Psychological Association. (2021). Intellect. In *APA dictionary of psychology*. https://dictionary.apa.org/intellect

American Psychological Association. (2021). Mind. In *APA dictionary of psychology*. https://dictionary.apa.org/mind

American Psychological Association. (2021). Target. In *APA dictionary of psychology*. https://dictionary.apa.org/target

Anderson, M., & Anderson, S.L. (2007). Machine ethics: Creating am ethical intelligent agent. *AI Magazine, 28*(4), 15-26. https://doi.org/10.1609/aimag.v28i4.2065

Anthis, J.R. (2018, June 21). *What is sentience?* Sentience Institute. https://www.sentienceinstitute.org/blog/what-is-sentience

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563*, 59-64. https://doi.org/10.1038/s41586-018-0637-6

Blascovich, J., & Ginsburg, G.P. (1978). Conceptual analysis of risk-taking in 'Risky-Shift' research. *Journal for the Theory of Social Behaviour, 8*(2), 217-230. https://doi.org/10.1111/j.1468-5914.1978.tb00400.x

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227-247. https://doi.org/10.1017/S0140525X00038188

Bostrom, N. (1998). How long before superintelligence?. *International Journal of Futures Studies*, *2*.

Bostrom, N., Dafoe, A., & Flynn, C. (2020). Public policy and superintelligent AI: A vector field approach. In Liao, S.M. (Ed.), *Ethics of artificial intelligence*. Oxford Scholarship Online. https://doi.org/ 10.1093/oso/9780190905033.003.0011

Brey, P., & Søraker, J.H. (2009). Philosophy of computing and information technology. In A. Meijers (Ed.), *Philosophy of technology and engineering sciences: A handbook of the philosophy of science* (pp. 1341-1407). https://doi.org/10.1016/B978-0-444-51667-1.50051-3

Broom, D.M. (2020). Brain complexity, sentience and welfare. *Animal Sentience, 29*(27). https://doi.org/10.51291/2377-7478.1613

Bryson, J.J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63-74). John Benjamins Publishing Company.

Bynum, T.W. (2004). Ethics and the information revolution. In R.A. Spinello, & H.T. Tavani (Eds.), *Readings in cyberethics* (2nd ed.) (pp. 13-29). Jones and Bartlett Publishers.

Cervantes, J-A., López, S., Rodríguez, L-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics, 26*, 501-532. https://doi.org/10.1007/s11948-019-00151-x

Darling, K. (2016). Extending legal protection to social robots: The effects of

anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A.M. Froomkin, & I. Kerr (Eds.), *Robot law* (pp. 213-232). Elgaronline. https://doi.org/10.4337/9781783476732.00017

Dautenhahn, K. (2007). Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems, 4*(1), 103-108. https://doi.org/10.5772/5702

de Graaf, M.M.A, Hindriks, F.A., & Hindriks, K.V. (2021). *Who wants to grant robots rights?* In HRI '21 Companion: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery. https://doi.org/10.1145/3434074.3446911

DeGrazia, D. (1996). *Taking animals seriously: Moral life and moral status.* Cambridge University Press. https://doi.org/10.1017/CBO9781139172967

Effective Altruism Forum. (2021). *Digital person* [Online forum post]. Effective Altruism. https://forum.effectivealtruism.org/tag/digital-person

Floridi, L., & Sanders, J.W. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fröding, B., & Peterson, M. (2020). Friendly AI. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-020-09556-w

Gilbert, M., & Martin, D. (2021). In search of the moral status of AI: Why sentience is a strong argument. *AI & Society*. https://doi.org/10.1007/s00146-021-01179-z

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. Springer.

Gray, H.M., Gray, K., & Wegner, D.M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Graziano, M.S.A. (2017). The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI, 4*, 60. https://doi.org/10.3389/frobt.2017.00060

Gunkel, D.J. (2018). *Robot rights.* MIT Press.

Gunn, L.J., Chapeau-Blondeau, F., McDonnell, M.D., Davis, B.R., Allison, A., & Abbott, D. (2016). Too good to be true: When overwhelming evidence fails to convince. *Proceedings of the Royal Society A, 472*(2187), 1-15. https://doi.org/10.1098/rspa.2015.0748

Harris, J. (2021, February 26). *The importance of artificial sentience*. Sentience Institute. https://www.sentienceinstitute.org/blog/the-importance-of-artificial-sentience

Harris, J., & Anthis, J.R. (2021) The moral consideration of artificial entities: A literature review. *Science and Engineering Ethics, 27*(53), 1-95. https://doi.org/10.1007/s11948-021-00331-8

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry, 27*(1). https://doi.org/10.1080/1047840X.2016.1082418

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology, 65*, 399-423. https://doi.org/10.1146/annurev-psych-010213-115045

Himma, K.E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?. *Ethics and Information Technology, 11*, 19-29. https://doi.org/10.1007/s10676-008-9167-5

Hunt, E. (1995). The role of intelligence in modern society. *American Scientist, 83*(4), 356-368. https://www.jstor.org/stable/29775483

Jones, M., Harmon, S. & O'Grady-Jones, M. (2004). Educating the digital mind: Challenges and solutions. In R. Ferdig, C. Crawford, R. Carlsen, N. Davis, J. Price, R. Weber & D. Willis (Eds.), *Proceedings of SITE 2004--Society for Information Technology & Teacher Education International Conference* (pp. 1753-1760). Atlanta, GA, USA: Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/primary/p/14684/.

Karnofsky, H. (2016, May 6). *Some background on our views regarding advanced artificial intelligence*. Open Philanthropy. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence

Karnofsky, H. (2021, July 27). *Digital people would be an even bigger deal*. Cold Takes. https://www.cold-takes.com/how-digital-people-could-change-the-world/

Keller, H. (2016). Psychological autonomy and hierarchical relatedness as organizers of developmental pathways. *Philosophical Transactions of the Royal Society B, 371*(1686), 1-9. https://doi.org/10.1098/rstb.2015.0070

Körner, A., Topolinski, S., & Fritz, S. (2015). Routes to embodiment. *Frontiers in Psychology, 6*, 940. https://doi.org/10.3389/fpsyg.2015.00940

Law Insider. (2021). Artificial entity. In *Law insider*. https://www.lawinsider.com/dictionary/artificial-entity

Lin, P., Abney, K., & Bekey, G.A. (Eds.). (2011). *Robot ethics: The ethical and social implications of robotics*. MIT Press.

Lin, P., Abney, K., & Bekey, G.A. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence, 175*(5-6), 942-949. https://doi.org/10.1016/j.artint.2010.11.026

Lo, S. (2019, June 25). *What is a legal person? Law dictionary corrects decades-old error*. Nonhuman Rights Blog. https://www.nonhumanrights.org/blog/legal-person-blacks-law-correction/

Oxford Languages. (2021). *Oxford Languages and Google.* Oxford Languages. https://languages.oup.com/google-dictionary-en/

Oxford Reference. (2021). Artificial person. In *A dictionary of accounting.* https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426978

Markusen, A. (2003). Fuzzy concepts, scanty evidence, policy distance: The case for rigour and policy relevance in critical regional studies. *Regional Studies, 37*(6-7), 701-717. https://doi.org/10.1080/0034340032000108796

Martela, F., & Riekki, T.J.J. (2018). Autonomy, competence, relatedness, and beneficence: A multicultural comparison of the four pathways to meaningful work. *Frontiers in Psychology, 9*, 1157. https://doi.org/10.3389/fpsyg.2018.01157

Merriam-Webster. (2021). Being. In *Merriam-Webster*. https://www.merriam-webster.com/dictionary/being

Mindell, D.A. (2004). *Between human and machine: Feedback, control, and computing before cybernetics*. Johns Hopkins University Press.

Mobus, G.E. (2019, May 16). *A theory of sapience: Using systems science to understand the nature of wisdom and the human mind.* Millennium Alliance for Humanity and the Biosphere. https://mahb.stanford.edu/library-item/theory-sapience-using-systems-science-understand-nature-wisdom-human-mind/

Molina, M. (2020). *What is an intelligent system?*. ArXiv. https://arxiv.org/abs/2009.09083

Moor, J.H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18-21. https://doi.org/10.1109/MIS.2006.80.

Muehlhauser, L. (2013, June 19). *What is intelligence?*. Machine Intelligence Research Institute. https://intelligence.org/2013/06/19/what-is-intelligence-2/

Muehlhauser, L. (2017, June). *2017 Report on consciousness and moral patienthood*. Open Philanthropy. https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood

Oliveira, A. (2017). *The digital mind: How science is redefining humanity.* MIT Press.

Reggia, J.A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks, 44*, 112-131. https://doi.org/10.1016/j.neunet.2013.03.011

Reisman, J.M. (1984). Friendliness and its correlates. *Journal of Social and Clinical Psychology, 2*(2), 143-155. https://doi.org/10.1521/jscp.1984.2.2.143

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach* (1st ed.). Prentice Hall.

Scheutz, M. (2013). What is robot ethics? [TC Spotlight]. *IEEE Robotics & Automation Magazine,* 20(4), 20. https://doi.org/10.1109/MRA.2013.2283184.

Shepherd, J., & Levy, N. (2020). Consciousness and morality. In Y. Kriegel (Ed.), *The Oxford handbook of the philosophy of consciousness.* Oxford Handbooks Online. https://doi.org/10.1093/oxfordhb/9780198749677.013.30

Shulman, C., & Bostrom, N. (2021). Sharing the world with digital minds. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking moral status*. Oxford Scholarship Online. https://doi.org/10.1093/oso/9780192894076.001.0001

Silverio-Fernández, M., Renukappa, S., & Suresh, S. (2018). What is a smart device? - a conceptualisation within the paradigm of the internet of things. *Visualization in Engineering, 6*(3), 1-10. https://doi.org/10.1186/s40327-018-0063-8

Sotala, K. (2012). Advantages of artificial intelligences, uploads, and digital minds. *International*

Journal of Machine Consciousness, 4(1), 275-291. https://doi.org/10.1142/S1793843012400161

Spring, J. (2012). *Education networks: Power, wealth, cyberspace, and the digital mind*. Routledge. https://doi.org/10.4324/9780203156803

Stanford Encyclopedia of Philosophy. (2014). Consciousness. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/consciousness/

Stanford Encyclopedia of Philosophy. (2019). Eliminative materialism. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/materialism-eliminative/

Stanford Encyclopedia of Philosophy. (2020). Machine ethics. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/ethics-ai/#MachEthi

Stanford Encyclopedia of Philosophy. (2020). The definition of morality. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/morality-definition/

Stanford Encyclopedia of Philosophy. (2021). Qualia. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/qualia/

Stephan, W.G., Ybarra, O., & Morrison, K.R. (2009). Intergroup threat theory. In T.D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 43–59). Psychology Press.

Sternberg R.J. (1986). A triarchic theory of human intelligence. In S.E. Newstead, S.H. Irvine, & P.L. Dann (Eds.), *Human assessment: Cognition and motivation*. Springer. https://doi.org/10.1007/978-94-009-4406-0_9

Velasquez, M., Andre, C., Shanks, T., & Meyer, M.J. (2010, January 1). *What is ethics?*. Markkula Center for Applied Ethics. https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/

Weiss, L.G., Saklofske, D.H., Holdnack, J.A., Prifitera, A. (2019). *WISC-V: Clinical use and interpretation* (2nd ed.). Academic Press. https://doi.org/10.1016/C2017-0-03528-0

Wiblin, R. (Host). (2020, July 9). Ben Garfinkel on scrutinising classic AI risk arguments (No. 81) [Audio podcast episode]. In *The 80,000 Hours Podcast*. 80,000 Hours. https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments/

Wiener, N. (1948). *Cybernetics; or control and communication in the animal and the machine.* John Wiley.

Acknowledgments

Many thanks to Jacy Reese Anthis, Ali Ladak, Thomas Moynihan, Tobias Baumann, and Teo Ajantaival for reviewing and providing feedback.