Artificial Consciousness: Where Does Science Fact Break from Science Fiction, and How Do We Know?

Theron E. Fairchild, Kanagawa University, Japan

The Asian Conference on Film & Documentary 2014
Official Conference Proceedings

Abstract
This paper explores what has been termed artificial consciousness (AC) (a.k.a., synthetic consciousness or artificial sentience). Related to its companion, artificial intelligence (AI), the subject might sound more like science fiction or fantasy than possibility. Though humans have been speculating about nonhuman consciousness for centuries, it was in the 1960s when computer science promised the rise of machines with human-level abilities. While the real-world challenges persist, we went ahead and built autonomous, fictional entities like HAL 9000 and the Terminator. This task has been relatively simple for standard narrative, by merely placing anthropomorphic character over a machine. In reality, constructing the human platform, through silicon or otherwise, is more than a matter of physical engineering or reshuffling human qualities. In fact, a truly artificial agent has very little need to replicate human intelligence or other capabilities. Consequently, the potential emergence of real-world AC could have less to do with our machines than with the success or failure of our minds to comprehend it. Given the typical portrayal of AI in fiction, as talking bipedal robots and doomsday machines, and our centuries of misunderstanding organic life forms, including our own, we might simply be incapable of imagining where the future is headed.

Keywords: artificial intelligence, artificial consciousness, machine consciousness, science fiction, synthetic phenomenology

Introduction

This paper explores the possibility of what has been termed synthetic or artificial consciousness (AC). To some, this might sound more like science fiction than fact, but to those engaged in the dialog, the potentiality of AC is a serious prospect. In the 1960s, artificial intelligence and computer science promised the rise of machines with human-level intelligence and other capabilities. In fiction, this has been comparatively easy to engineer. Simply take the human character and assign it machine qualities, or merely do what we have been doing for centuries with animals, which is anthropomorphize them with human qualities. On the film screen, this has worked rather well. Our autonomous, sentient agents of tomorrow have come in humanoid forms such as Roy Batty, Major Kusanagi, and Lt. Commander Data. By comparison, clustering forms of artificial entities have been known as HAL9000, Skynet, and the Matrix.

Multiple Definitions, Levels of Description, and the Unique Human

Much of the subtext in this paper concerns the intersecting vectors of fact and fiction. This relates to a kind fictionalism, a philosophical discourse useful for conveying ideas, but nonetheless must not be interpreted as literal truth (Eklund, 2011). Despite the portrayal of AI and robots in film and in popular press, subsequently nested within a framework of Western archetypes utilizing the typical good-bad narrative dichotomy, the potential evolution of machine consciousness is a much stickier, less well-defined proposition than fiction suggests. As such, artificial or synthetic consciousness could be interpreted a number of ways, including the rise of consciousness from information. Such a view calls attention to the psycho-philosophical constructs of intelligence, cognition, affect, and human ethics. In turn, each must be examined on the road to machine intelligence and nonhuman ethics. Hence, before making significant headway on some kind of definition for artificial consciousness, human efforts to the effect will need to recognize our indefinite understanding of what might give rise to AC itself.

Western science and even art, particularly in the American vein, have been greatly influenced by products of philosophy and behavior rather than processes of an internal nature (see Allen, 2011): in other words, philosophizing and doing take precedence over experiencing (except when it comes to nonhumans, then we conveniently play it the other way, guaranteeing an argument of human specialness in both directions). Although the neuroscientific revolution of the past decade has done much to change the attitude toward nonhuman consciousness (for a review, see Edelman & Seth, 2009), the influence of behaviorism still dominates many of the social sciences, from behavioral psychology to behavioral economics. It also fuels the divisions in AI, between an industrial model of narrow AI, focused on function and application in commercial and political economy, and an evolutionary-inspired model of general AI, which, in part, hopes to generate the autonomous entities of popular imagination. At present, compared to the success of the former camp, the latter, more-general view of these two fundamental directions in AI, has been considered a failure. But the underlying problem itself is nothing new; in fact, it is as old as our ability to process information, form definitions, and manage ambiguity, particularly when we consider multiple levels of description.

The definitional problem begins right away, with the very concept of *information*. The word *data* is often employed synonymously these days, but it certainly does not appease a majority. In fact, the father of information theory, Claude Shannon (1916-2001), considered the word information as merely useful for discussion, while a single concept of information could hardly account for any ultimate definition of what information was or what it could do (Floridi, 2010). In other words, Shannon avoided the common pitfall of reifying a construct, which is the mistake, throughout history, of assigning physical or event-meaning to an abstraction of language.

The problem is much more apparent with the constructs of consciousness, intelligence, and emotion, whereby we tend to discuss each in terms of how much or how many, or of strictly having or not having. It is mainly problematic because consciousness, intelligence, and even emotions do not actually exist (on emotion, see Griffiths, 2008); they are language constructs or abstractions that enable us to have conversations about the phenomena they purport to represent. As such, it is not possible to have actual quantities or on/off experiences of any of them. Parisi (2007) considered this a fundamental flaw with the philosophical approach to consciousness, where the question itself, What is consciousness? remains particularly useless for any scientific attempt to untangle the possibility of machine consciousness. This might seem counterintuitive in the context of natural conversation, but only because we have come to think of humans (and all entities, biological or otherwise) in terms of "being in possession of" some well-defined characteristic. This is one of the core difficulties plaguing many pursuits, including the use of interpretation and comprehension among storytellers and the public alike, where mixing (distorting) conventions and levels of description are not only acceptable but considered creative.

It also presents a paradox that probably most are not quick to realize. Throughout history and philosophy, and commonly today in psychology, popular media, and elsewhere, we often talk about "human nature," which we eagerly employ in our struggle to maintain segregated uniqueness in relation to nonhuman species and machines (on interspecies relations, see Corbey & Lanjouw, 2014). In this struggle, the constructs of intelligence, emotion, and consciousness are gold standards for how we think, how we feel, and how we experience, thus giving us our unique nature. The problem, as mentioned above, and critiqued at length by Ashworth (2000), is that this so-called nature has mostly been drafted from the theoretical abstractions of our own minds. Without biological or empirical bases, they cannot count as definitions of nature in any physical or material sense. It is in our nature to breathe; this is readily obvious. It is in our nature to eat. But we would not use these two premises as defining characteristics of Homo sapiens. When we actually look at biology and evidence, we find that what we consider to be emotions, as with breathing and eating, have prehistoric neurological origins across species, while intelligent behavior, as we understand it, is common to at least mammals (Johnson, 2010; Rumbaugh & Washburn, 2003). The real difference, then, given the more-advanced cortical development of the human brain, is that we can perform higher levels of intelligent behaviors, which includes making abstract terminology. This advanced intelligence, coincidentally, fuels the trend, in both science and popular culture, of us downplaying our primitive emotionality in favor of our new and unique intelligence.

Problems with Organic Consciousness

To the issue of multiple levels of description, Albert Einstein wrote, "Body and soul are not two different things, but only two different ways of perceiving the same thing. Similarly, physics and psychology are only different attempts to link our perceptions together by way of systematic thought" (as cited in Levy, 2010, p. 14).

The point is important because we humans tend to define all things based on our limited or even single perceptions of the world and its various phenomena. Subsequently, artificial consciousness can be a particularly difficult concept to grasp, due, in no small part, to our hard enough time grappling with human consciousness, let alone the thorny topic of nonhuman or animal consciousness. Allen (2011) pointed out that most animals clearly have ordinary consciousness, meaning they demonstrate states of wakefulness or awareness of immediate environment. At least mammals, too, experience their own forms of affect, based on their ability to sense, and seemingly process, experiences of pain and suffering in themselves and in others (Johnson, 2010).

More controversial, though, are the rather subjective and experiential forms of consciousness, as well as self-consciousness, and what Block (1995) called access consciousness: the ability to represent conscious information to others, typically through the use of language. This all ties into what is known as the multiple realizability of mind, a longstanding antireductionist argument in philosophy that, essentially, defines the mind as a confluence of abstractions leading to functions, behaviors, or characteristics of what we would recognize as a mind itself (e.g., see Bickle, 2013). This has been a major position in psychological science, which has been readily adapted to AI theory concerning emotion and consciousness (Scheutz, 2014). Other theories also abound for machine consciousness, including that of Manzotti (2007), whereby consciousness results from various process schemes manifesting from an architecture. By comparison, Morasso (2007) advanced the approach that consciousness is dependent on a nervous system, thus it must be more than a mere mental phenomenon. This latter argument runs throughout neuroscientific studies (Cvetkovic & Cosic, 2011), on humans and even nonhumans (Boly et al., 2013), which cannot be covered in this current paper. Safe to say, however, that a single approach, a single interpretation, or even a single definition of the phenomenon we label consciousness, would not be a fair approach, regardless of popular, literary, or other academic ruminations to the opposite.

With that said, life can give rise to not only phenomenological consciousness, as evidenced by its existence in humans, but also to a plurality of conscious states and experiences. In neurology, for example, consciousness involving the fully formed human brain is not discussed as an on/off phenomenon, but as a function of vigilance and awareness, in which the complete and simultaneous engagement of both represents the state of being fully rested, awake, and attentive. Many other states also exist, including the various forms of normal sleep, as well as states experienced as a result of brain trauma. The age of the particular individual is also significant, specifically the amount of natural neurological development or degradation during the lifespan. In cross-species comparisons, the same complex considerations also hold true. An example would be the adult chimpanzee, which has demonstrated several of

the neurocognitive and behavioral characteristics commonly associated with the average human toddler (Matsuzawa, 2013; Rumbaugh & Washburn, 2003).

Artificial, Artifact, Animal

While mainstream AI still remains fairly disinterested in the possibility of conscious machines, other scholars, typically working across disciplines, have attempted to lay a foundation for its scientific study. Chella and Manzotti (2007), for example, have argued for a newer, more-sophisticated modeling of consciousness, which partly makes use of how it arises in the human brain. Others have put forth the concept of synthetic phenomenology (Aleksander & Morton, 2007; Chrisley, 2009) as a method for examining the potential consciousness of artifacts: an approach that states a machine or artifact does not necessarily need to be conscious for it to contribute to the study of machine consciousness.

The argument goes back to at least Alan Turing (1950), who articulated the impossibility of getting inside a machine, to evaluate any potential feelings or thoughts, and considered the idea of doing so an expression of mere human solipsism. A more contemporary, even stronger critique was put forth by Pollack (2006), who questioned whether human-level intelligence is even the standard by which all intelligence should be measured. Pollack refuted whether a mind is even necessary for intelligent behavior, considering that evolution itself, often through symbiotic composition (Watson & Pollack, 2003), has produced a myriad of intelligent-behaving species that lack minds or even sophisticated brains. Others (Buttazzo, 2001; Haikonen, 2012; Reggia, 2013) have questioned whether consciousness, too, even requires our classic definition that brains or minds must exist in some prescribed manner before consciousness can occur. The suggestion here is that organized patterns in neural networks are simply enough.

Regardless of the approach to understanding nonhuman consciousness, asserting that a machine or artifact cannot be conscious is a bit like saying a rabbit cannot have manners. This analogy might seem a bit removed from the issue at hand, but it calls attention to the important role that evaluation plays in constructing or defining reality based on abstractions. In this case, manners are a subjective human construct that take no account of what might constitute politeness in a rabbit society. Given the needs of living as a rabbit, essentially as prey not predator, rabbits might be far more courteous than any other species, including humans. Accordingly, their inability to express this in human terms does not make them any less capable of having manners. Ultimately, stating that rabbits cannot have manners, in the same way that nonhuman agents cannot have consciousness, is a meaningless claim; and it does not draw attention to a difference in species as much as it does to problems associated with applications of language. The only fact we actually glean for certain, which adds to the pointlessness, is that rabbits cannot be humans.

Taking a case in the reverse direction, most dolphins are in control of a very sophisticated echolocation system, giving them a heightened awareness of underwater frequencies and sounds that goes unmatched by most other aquatic creatures (Berta, Sumich, & Kovacs, 2006). This phenomenon is a matter of biology, which has no equivalent among primates. The bottlenose dolphin also sleeps one hemisphere at a time, allowing it to maintain levels of awareness during slumber, a characteristic

completely outside the human experience. Given these facts, filtered through standard definitions of consciousness, dolphins must be in possession of higher levels of consciousness than humans, at least ordinary consciousness in their native ocean environments. In fact, based on the same logic commonly employed about the uniqueness of human consciousness, dolphins would be fully conscious and special, while humans would be lesser creatures with limited awareness and capabilities, merely existing in the world while dolphins were excelling in it. In addition, given the human inability to speak a dolphin code, or even register sound frequencies that dolphins process quite naturally, humans must actually be only partly conscious. For them to thrive as a species, therefore, they will need to evolve dolphin-level capabilities.

The above line-of-thought is important because an underlying theme, throughout the evolution of AI, has been whether an artifact or synthetic entity can or will develop human-like capabilities. This has certainly been the case among the artificial general intelligence (AGI) community, which has set its sites on that very goal. More realistically, and perhaps more disturbingly, depending on one's outlook, is that AI has no inherent need to be like humans, in the same way that a human has no need to evolve into a dolphin, any more than our rabbits need to develop human-like manners. What is much more likely is that an evolving AI could skip past humans altogether, as it has already done computationally, and proceed to super-intelligent entity with full autonomy.

Because we cannot be sure of the consequences from this trajectory, it remains potentially disturbing for people, and has likewise inspired a whole genre of dystopic literature and film where machines take over the world. It has also inspired formal critiques of the problem within AI itself, ranging from legitimate concerns about control and coding (Armstrong, 2014), cooperative morality and reciprocal altruism (Fox & Shulman, 2010), and anthropomorphic bias and designed friendliness (Bostrom, 2014; Yudkowsky, 2008). The problem has been simultaneously met by an opposite, evolutionary argument (Pinker, 2007; Swayne, 2013), that super-intelligence, by definition, means the ability to solve problems in adaptive, non-threatening ways; in other words, just because humans have evolved emotionally aggressive and destructive tendencies it does not mean that a machine, without any of the same biological architecture, will follow the same course.

Crafting the Fiction

From the first page of his book entitled *Between Literature and Science*, Swirski (2000) argued that in order to make any future accessible, the writer of science fiction and fantasy cannot delve too far into a truly plausible future, one that would merely appear incomprehensible to most readers. The criticism concerns the pragmatic side of aesthetics, to which the writer must ground a story in familiarities associated with past and present. Such narrative strategy, quite common in fiction development, remains effective because it exploits the tendency, documented throughout psychology, for people to rely on heuristics, biases, and cognitive illusions for everyday judgments and existence. In one influential work on the familiarity bias, Kahneman and Tversky (1973) documented that people tend to construct representations of reality based on what they already know, as opposed to new evidence or statistical logic, a habit leading to fallacious intuitions and erroneous

predictions. A related term is the assimilation bias, whereby we tend to conform new data to match our existing beliefs, rather than the other way around. Thomas (2013) addressed the dilemma in terms of mental imagery, which we depend on tremendously for our daily functionality, a habit that nonetheless posits numerous problems for perception, memory, and meaning formation. In short, regardless of what we see with our eyes, our internal images guide many of our objective decisions, images that are entirely subjective phenomena.

Given such predicaments, and predisposed to excessive reliance on vivid but not necessarily appropriate information (i.e., the availability bias), we are faced with the hard challenge of trying to discuss the future without falling into self-deception, either scientifically or fictionally. As a consequence, Swirski (2000) claimed that the gulf between the two cultures of science and literature is not as vast as purported. Instead, the situation should be seen as more of a relationship, complex and interesting, in which science fiction writers in particular have built "epistemic bridges" (p. x) between the two worlds. Broderick (2000) stated it perhaps more curiously. "Quite a few writers in and out of science fiction have been eddying in the slipstream of science toward a gnarly attractor in narrative space …" (p. 3).

In the study of creativity in arts and science, Root-Bernstein and Root-Bernstein (2004) noted many similarities between the two worlds as well. For example, artists and scientists, including mathematicians, tend to share psychological testing profiles. All three, despite differences in product, also speak about their processes in similar ways, processes that include complex pattern recognition, managing abstraction, and the intentional employment of imagination. The particular importance of imagination has been discussed by many (e.g., Gendler, 2013; Markman, Klein, & Suhr, 2009; Taylor, 2011), not simply as a tool for creativity and storytelling, but as a requirement for all manner of affect and learning, including social understanding, empathy, moral reasoning, and simulated projection into other times and events, which aids critical thought.

So how does this play into the world of AI or AC? The future tends to beget fear, particularly in an age devoted to the nonstop broadcasting of global unrest and impending doom. AC is a topic of the future, and will become more relevant as more of that future passes into present. Compounding that anxiety has been such franchise films as *The Terminator* (Hurd, 1984) and *The Matrix* (Silver, 1999), in addition to the ultra-realistic genre defining *2001: A Space Odyssey* (Kubrick, 1968). In each case, the main characters are faced with a powerful central-AI bent on human annihilation. In the near 15-year spacing between each film, that AI depiction evolved like a religion, from the unassuming psychotic agent of HAL9000, to the wrathful and biblical god-like entity known as the Matrix. Also, in each of those films, the humans retained no way of determining whether each AI were conscious, merely the sense that they were because of the destructive behavior they emanated.

Crafting the Fact

The fictions of the preceding paragraph addressed problems for AC, ones not without parallel in the real world. The centralized processing AI approach, throughout the history of AI development, has thus far been a predominately Anglo-American phenomenon, one that readily stokes the fires of Orwell's (1949) warnings of Big

Brother. On the surface, IBM's Watson computer, which proved victorious on the game show *Jeopardy*, a feat previously reserved only for humans, has a public persona visually bathed in the light of cool blue, a color that psychology has shown to be relaxing and non-threatening to humans. The visual choice is important because HAL 9000 was swimming in alarming red. With the exception of this visual, and the possibility that HAL was probably self-aware while Watson most probably is not, the two systems are architecturally similar. In other words, both are, essentially, physically amorphous clustering systems of multiple processors, focused around a central hub that expresses intelligent behavior. This has been the fundamental direction of large-scale AI development since its inception over 60 years ago, and its fictional portrayal in *2001* is, if anything, a testament to the visionary outlook of director Stanley Kubrick and his project consultant, the scientist and science-fiction author Arthur C. Clarke.

While the game-show champion Watson shares parallels with HAL 9000, the immense computational clustering and algorithmic AI power of a company like Google more resembles Skynet from the *Terminator* series. In addition, according to clues garnered from trends in spending and activity, the National Security Agency (NSA) (Bamford, 2009; Global Research, 2013) and the U.S. Defense Department (Elkus, 2014) are massively investing in AI development, with their combined efforts possibly accounting for the largest AI spending globally. Like in the films presented above, there really is no way of determining if any real-world agents in the Google, Pentagon, and especially NSA scenarios can attain, or have already attained, something resembling a conscious state, particularly since each platform is unavailable for public assessment. In particular, if the Defense Department and the NSA are developing AI similar to the more-publicly understood Google architecture, this would mean narrow, task-specific functionality where any kind of consciousness or self-awareness, recognizable by humans, is not going to be detectable.

The truth is, as the above-mentioned films have implied, and as the arguments from earlier sections in this paper have suggested, such consciousness might actually be outside our innate human capability to ever detect. Given our track record with other nonhuman species to date, even if we managed some form of detection, all probability indicates that AC will be incomprehensible (e.g., see Heaven, 2013). So incomprehensible, in fact, that some sophisticated AI designs might already have a primitive form of consciousness, which makes our biological inability to experientially conceive or even appreciate dolphin echolocation seem like an elementary-school problem. At least dolphins are formed of the same organic matter as humans, with dual-hemispheric brains responsive to a central nervous system, as all mammals, birds, and most other macro-level species are. But how does a silicon-based artifact, with a nervous system of binary 0s and 1s, potentially interpret, understand, and maybe even imagine itself within an environment? What will be the case when machines move beyond silicon and binary, to the barely understood world of quantum computing, which is now in its infancy?

In his pivotal 1950 paper, Turing proposed a test for machine intelligence, one that has since become known simply as the Turing Test. In almost direct rebuttal 30 years later, Searle (1980) proposed his Chinese room argument, against the possibility of any kind of computer mind or consciousness. Both of these have been discussed extensively for decades, taking on near-religious significance in AI debate, so they

will not be addressed here. But the truth is, neither is up to the task outlined in the preceding paragraph, of assessing the literally alien form that machine consciousness is likely to take or has already started taking. As such, given the current limits of our perceptions, languages, and biases against fathoming future conditions, AC might as well be from another galaxy.

Is That the Only Model?

While the Anglo-American world remains transfixed on the potentiality of cluster systems, AI in Japan is closely bound up with advances in robotics. Specifically, advances in autonomous humanoid robotics have inspired an array of architectures and possibilities, including the internationally famed Asimo system by the Honda Corporation. Humanoid robots, and their typically negative portrayal in Western cinema, are greeted in nearly opposite fashion in Japanese society (Katsuno, 2011). One reason is purely pragmatic. The Japanese population is rapidly aging and there will be no one to care for them (Iida, 2013; Ryall, 2013). As such, while Western analysts and the public debate the anxiety caused by intelligence (e.g., Barrat, 2013) and human-like physical appearances (Bar-Cohen & Hanson, 2009) of machines, affect for human interfacing has become a topic of commercial and social development throughout East Asia. A big question, then, posited throughout the history of AI, is whether machines can emote. But as argued earlier in this paper, this line of question remains somewhat depthless.

Psychology has clearly outlined that affect is a construction, and constructions require both a speaker and a listener to agree on their meaning in order for such constructions to carry accurate substance or implication. Hence, all affective states, when at least two parties are involved, require both an emoter and a perceiver in order for any individual emotion to carry a definition. Take, for example, something as simple as a smile. How can we tell what the smiler is actually feeling, or whether our interpretation of that smile is even accurate? Is the person happy, nervous, or perhaps masking an altogether different feeling? Likewise, does the smiler even know why he or she is smiling? Maybe the smile is not masking an emotion from the external viewer, but masking it from the emoter. We are not always in command of, or in touch with, how we feel, sense, or react at any given moment.

The truth is, affect is as much, if not more, context driven and externally interactive as it is internally processed, which implies that nonhuman agents are capable of emoting even if they cannot feel anything on the inside. To an elderly person confined to a wheelchair or nursing facility, an emoting robot might very well be the most affectively rewarding experience of that person's present life. That very theme was played out in the 2012 small-budget film, *Robot and Frank* (Acord et al.), in which Robot itself was inspired by Honda's Asimo design. In that narrative, as with the argument above, regarding the constructed nature of affect and consciousness, one would be hard-pressed to deny that an agent such as Robot lacked a form of self-awareness. Perhaps more importantly, from a functional, phenomenological, and even humanistic perspective, if an argument against an affective or conscious humanoid agent were analytically successful, would it make any difference to the individual who believed otherwise? As noted earlier in this paper, on heuristics and bias as a cornerstone of judgment, the answer is probably no.

Where Does This Leave Us?

In following traditions such as psychology, mathematics, and most sciences, not understanding what one is up against is no barrier to studying it. AI, particularly the type advocating human-like general intelligence, has become a multidisciplinary field, one that has turned around and challenged the very parameters it has borrowed, not the least of which includes the study of information, evolution, intelligence, and consciousness. In this light, regardless of hopes or fears, several AI researchers (e.g., Bostrom, 2013; Goertzel, 2014; Muehlhauser, 2013) have assumed the inevitability of super-intelligence and synthetic consciousness, and moved directly to a proactive stage of trying to provide some assurances, from the outset, that future AI or AC will not end up becoming the apocalyptic legend of Western film lore. The attitude is controversial, displeasing many in the more-traditional AI and philosophy communities alike. But from a dynamical chaos perspective, where sensitivity to initial conditions holds unseen deterministic reverberations throughout the future of the system, making an attempt at getting things right from the beginning might not be a bad approach.

In the meantime, as for whether actual consciousness in artificial agents is even plausible, at least a kind that we can understand, or will remain an artifact for science fiction, only time can say. In Stanislaw Lem's 1961 novel *Solaris*, subsequently made into films of the same title by Andrei Tarkovsky (1972) and Steven Soderbergh (2002), the author's central theme addressed the ultimate inadequacy of communication between humans and nonhuman species. Such a message will likely remain paramount for some years to come.

References

Acord, L., Bisbee, S., Kelman-Bisbee, J., & Niederhoffer, G. (Producers), & Schreier, J. (Director). (2012). *Robot & Frank* [Motion picture]. United States: Park Pictures.

Aleksander, I., & Morton, H. (2007). Depictive architectures for synthetic phenomenology. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness*. Exeter: Imprint Academic.

Allen, C. (2011). Animal consciousness. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Winter 2011 ed.). Retrieved from http://plato.stanford.edu/archives/win2011/entries/consciousness-animal/

Armstrong, S. (2014). *Smarter than us: The rise of machine intelligence*. Berkeley, CA: Machine Intelligence Research Institute.

Ashworth, P. D. (2000). *Psychology and "human nature."* East Sussex: Psychology Press.

Bamford, J. (2009, January 1). The new thought police: The NSA wants to know how you think—maybe even what you think. *NOVA*. Retrieved from http://www.pbs.org/wgbh/nova/military/nsa-police.html

Bar-Cohen, Y., & Hanson, D. (2009). *The coming robot revolution: Expectations and fears about emerging intelligent, humanlike machines*. New York: Springer.

Barrat, J. (2013). *Our final invention: Artificial intelligence and the end of the human era*. Burlington, MA: Academic Press.

Berta, A., Sumich, J. L., & Kovacs, K. M. (2006). *Marine mammals: Evolutionary biology* (2nd ed.). Burlington, MA: Academic Press.

Bickle, J. (2013). Multiple realizability. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Spring 2013 ed.). Retrieved from http://plato.stanford.edu/archives/spr2013/entries/multiple-realizability/

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*, 227-247.

Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., Edelman, D. B., & Tsuchiya, N. (2013). Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in Psychology, 4* (625), 1-20. http://dx.doi.org/10.3389/fpsyg.2013.00625

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford: Oxford University Press.

Buttazzo, G. (2001). Artificial consciousness: Utopia or real possibility? *Computer, 34*(7), 24-30.

Chella, A., & Manzotti, R. (Eds.). (2007). *Artificial consciousness*. Exeter: Imprint Academic.

Chrisley, R. (2009). Synthetic phenomenology. *International Journal of Machine Consciousness, 1*(1), 53-70.

Corbey, R., & Lanjouw, A. (Eds.). (2014). *The politics of species: Reshaping our relationships with other animals*. Cambridge: Cambridge University Press.

Cvetkovic, D., & Cosic, I. (Eds.). (2011). *States of consciousness: Experimental insights into meditation, waking, sleep and dreams*. New York: Springer.

Edelman, D. B., & Seth, A. K. (2009). Animal consciousness: A synthetic approach. *Trends in Neurosciences, 32*(9), 476-484. doi:10.1016/j.tins.2009.05.008

Ekland, M. (2011). Fictionalism. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Fall 2011 ed.). Retrieved from http://plato.stanford.edu/archives/fall2011/entries/fictionalism/

Elkus, A. (2014, March 6). Artificial intelligence: War's new grammar. *War on the rocks*. Retrieved from http://warontherocks.com/2014/03/artificial-intelligence-wars-new-grammar/

Floridi, L. (2010). *Information: A very short introduction*. Oxford: Oxford University Press.

Fox, J., & Schulman, C. (2010). *Superintelligence does not imply benevolence*. Paper presented at EACAP10: European Conference on Computing and Philosophy (K. Mainzer, Ed.), Munich, Germany. Retrieved from https://intelligence.org/files/SuperintelligenceBenevolence.pdf

Gendler, T. (2013). Imagination. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Fall 2013 ed.). Retrieved from http://plato.stanford.edu/archives/fall2013/entries/imagination/

Global Research (2013, June 16). The next NSA spying shoe to drop: "Pre-crime" artificial intelligence [Blog]. Retrieved from http://www.globalresearch.ca/the-next-nsa-spying-shoe-to-drop-pre-crime-artificial-intelligence/5339360

Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence, 5*(1), 1-46. doi:10.2478/jagi-2014-0001

Griffiths, P. S. (2008). *What emotions really are: The problem of psychological categories*. Chicago: University of Chicago Press.

Haikonen, P. O. (2012). *Consciousness and robot sentience*. Singapore: World Scientific.

Heaven, D. (2013). Not like us: Artificial minds we can't understand. *New Scientist, 219*(2929), 32-35. doi:10.1016/S0262-4079(13)61996-X

Hurd, G. A. (Producer), & Cameron, J. (Director). (1984). *The terminator* [Motion picture]. United States: Hemdale Communications.

IIda, M. (2013, June 19). Robot niche expands in senior care. *The Japan Times*. Retrieved from http://www.japantimes.co.jp/news/2013/06/19/national/social-issues/robot-niche-expands-in-senior-care/#.VD4L1NSUd8w

Johnson, C. M. (2010). Observing cognitive complexity in primates and cetaceans. *International Journal of Comparative Psychology, 23*, 587-624.

Katsuno, H. (2011). The robot's heart: Tinkering with humanity and intimacy in robot-building. *Japanese studies, 31*(1), 93-109. http://dx.doi.org/10.1080/10371397.2011.560259

Kubrick, S. (Producer & Director). (1968). *2001: A space odyssey* [Motion picture]. United Kingdom: Metro-Goldwyn-Mayer.

Lem, S. (1961/2011). *Solaris* [Kindle DX version]. Retrieved from http://www.amazon.com

Levy, D. (2010). *Tools of critical thinking: Metathoughts for psychology*. Long Grove, Illinois: Waveland Press.

Manzotti, R. (2007). From artificial intelligence to artificial consciousness. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness*. Exeter: Imprint Academic.

Markman, K. D., Klein, W. M. P., & Suhr, J. A. (Eds.). (2009). *Handbook of imagination and mental simulation*. New York: Psychology Press.

Matsuzawa, T. (2013). Evolution of the brain and social behavior in chimpanzees. *Current Opinion in Neurobiology, 23*, 443-449. http://dx.doi.org/10.1016/j.conb.2013.01.012

Morasso, P. (2007). The crucial role of haptic perception: Consciousness as the emergent property of the interaction between brain, body and environment. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness*. Exeter: Imprint Academic.

Muehlhauser, L. (2013). *Facing the intelligence explosion* [Kindle DX version]. Retrieved from http://www.amazon.com

Orwell, G. (1949). *Nineteen Eighty-Four*. London: Signet Classic.

Parisi, D. (2007). Mental robotics. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness*. Exeter: Imprint Academic.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. New York: Viking.

Pollack, J. B. (2006). Mindless intelligence. *IEEE Intelligent Systems, 21*(3), 50-56. http://dx.doi.org/10.1109/MIS.2006.55

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

Root-Bernstein, R., & Root-Bernstein, M. (2004). Artistic scientists and scientific artists: The link between polymath and creativity. In R. J. Sternberg, E. L. Grigorenko, & J. L. Singer (Eds.), *Creativity: From potential to realization* (pp. 127-151). Washington, DC: American Psychological Association.

Rumbaugh, D. M., & Washburn, D. A. (2003). *Intelligence of apes and other rational beings*. New Haven, CT: Yale University Press.

Ryall, J. (2013, April 12). Japan turns to robots as population declines. *Deutsche Welle*. Retrieved from http://www.dw.de/japan-turns-to-robots-as-population-declines/a-17270786

Scheutz, M. (2014). Artificial emotions and machine intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 247-266). Cambridge, U.K.: Cambridge University Press.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-457.

Silver, J. (Producer), & Wachowski, L., & Wachowski, A. (Directors). (1999). *The matrix* [Motion picture]. United States & Australia: Silver Pictures & Village Roadshow Pictures.

Swayne, M. (2013, October 8). Get happy: Why superintelligent AI will probably be superfriendly. *Humanity-Plus Magazine*. Retrieved from http://hplusmagazine.com/

Taylor, M. (2011). Imagination. In M. A. Runco & S. Pritzker (Eds.), *Encyclopedia of creativity* (2nd ed.) [Online version]. http://dx.doi.org/10.1016/B978-0-12-375038-9.00118-7

Thomas, N. J. T. (2013). Mental imagery. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Fall 2013 ed.). Retrieved from http://plato.stanford.edu/archives/win2011/entries/mental-imagery/

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 49*, 433-460.

Watson, R. A., & Pollack, J. B. (2003). A computational model of symbiotic composition in evolutionary transitions. *BioSystems, 69*, 187-209. http://dx.doi.org/10.1016/S0303-2647(02)00135-1

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. N. Ćirković (Eds.), *Global catastrophic risks* (pp. 308-345). New York: Oxford University Press.