Consciousness and Cognition 45 (2016) 210–225

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Review article

Artificial consciousness and the consciousness-attention dissociation

Harry Haroutioun Haladjian [a,*], Carlos Montemayor [b]

[a] *Laboratoire Psychologie de la Perception, CNRS (UMR 8242), Université Paris Descartes, Centre Biomédical des Saints-Pères, 45 rue des Saints-Pères, 75006 Paris, France*
[b] *San Francisco State University, Philosophy Department, 1600 Holloway Avenue, San Francisco, CA 94132 USA*

ARTICLE INFO

Article history:
Received 6 July 2016
Accepted 12 August 2016

Keywords:
Artificial intelligence
Artificial consciousness
Consciousness
Visual attention
Phenomenology
Emotions
Empathy

ABSTRACT

Artificial Intelligence is at a turning point, with a substantial increase in projects aiming to implement sophisticated forms of human intelligence in machines. This research attempts to model specific forms of intelligence through brute-force search heuristics and also reproduce features of human perception and cognition, including emotions. Such goals have implications for artificial consciousness, with some arguing that it will be achievable once we overcome short-term engineering challenges. We believe, however, that phenomenal consciousness cannot be implemented in machines. This becomes clear when considering emotions and examining the dissociation between consciousness and attention in humans. While we may be able to program ethical behavior based on rules and machine learning, we will never be able to reproduce emotions or empathy by programming such control systems—these will be merely simulations. Arguments in favor of this claim include considerations about evolution, the neuropsychological aspects of emotions, and the dissociation between attention and consciousness found in humans. Ultimately, we are far from achieving artificial consciousness.

Contents

* Corresponding author.
E-mail addresses: haroutioun@gmail.com (H.H. Haladjian), cmontema@sfsu.edu (C. Montemayor).
URLs: http://www.haroutioun.com (H.H. Haladjian), http://www.carlosmontemayor.org (C. Montemayor).

http://dx.doi.org/10.1016/j.concog.2016.08.011
1053-8100/© 2016 Elsevier Inc. All rights reserved.

1. Introduction

One of the most compelling topics currently debated is how it may be possible to develop consciousness in machines. While related questions have been discussed academically in the cognitive sciences for some time now, the idea of artificial consciousness has received more attention in popular culture recently. There is a growing number of articles in magazines and newspapers that discuss related advances in Artificial Intelligence (AI), from self-driving cars to the "internet of things" where common household objects can be intelligently connected through a centralized control system. Most recently, Google's DeepMind group developed an artificial agent capable of winning the game of Go against humans, which is considered to be a huge accomplishment in AI since it goes beyond brute-force search heuristics and uses deep learning models (Silver et al., 2016). This advance promises more impressive innovations in the future.

Along with these advancements is a growing fear that we may be creating intelligent systems that will harm us (Rinesi, 2015). This topic has been addressed in many settings, from international conferences to popular books (e.g., Bostrom, 2014; Brooks, Gupta, McAfee, & Thompson, 2015). Films and television shows, such as *Ex Machina*, *Her*, and *Battlestar Galactica*, present scenarios where AI systems go rogue and threaten humans. While these fictional accounts remain unachievable with today's technology, they are beginning to feel more and more possible given how fast computers are "evolving".

The increase of these discussions in the mainstream media is telling, and one could say that Artificial Intelligence is truly at a turning point. Some think that the so-called 'singularity' (the moment in which AI surpasses human intelligence) is near. Others say there is now a Cambrian explosion in robotics (Pratt, 2015). Indeed, there is a surge in AI research across the board, looking for breakthroughs to model not only specific forms of intelligence through brute-force search heuristics but by truly reproducing human intelligent features, including the capacity for emotional intelligence and learning. Graziano (2015), for example, has recently claimed that artificial consciousness may simply be an engineering problem—once we overcome some technical challenges we will be able to see consciousness in AI. Even without accomplishing this lofty goal of machine sentience, it still is easy to see many examples of human-like rational intelligence implemented in computers with the aim of completing tasks more efficiently.

What we will argue, however, is that *phenomenal consciousness*, which is associated with the first-person perspective and subjectivity, cannot be reproduced in machines, especially in relation to emotions. This presents a serious challenge to AI's recent ambitions because of the deep relation between emotion and cognition in human intelligence, particularly social intelligence. While we may be able to program AI with aspects of human conscious cognitive abilities, such as forms of ethical reasoning (e.g., "do not cause bodily harm", "do not steal", "do not deceive"), we will not be able to create actual emotions by programming certain monitoring and control systems—at best, these will always be *merely simulations*. Since human moral reasoning is based on emotional intelligence and empathy, this is a substantial obstacle to AI that has not been discussed thoroughly.

Before proceeding, however, it is crucial to distinguish the present criticism of AI from a very influential one made by Searle (Searle, 1980, 1998). Searle has famously criticized AI because of its incapacity to account for *intentionality* (i.e., the feature of the mental states that makes them about something, essentially relating them to semantic contents), which he takes to be exclusively a characteristic of conscious beings. He also argues that a consequence of this criticism is that phenomenal consciousness (i.e., what it is like to have an experience) is necessarily a biological phenomenon. Searle, therefore, takes the limitations of AI to be principled ones that will not change, regardless of how much scientific progress there might be.

Critics have argued that the intuition that only biological beings can have intentional minds may be defeated (e.g., see Block, 1995a) and that cyborg systems or an adequate account of how the brain computes information could refute the Chinese room thought experiment (Churchland & Churchland, 1990; Pylyshyn, 1980). These criticisms have merit, and we largely agree with them, but only with respect to the kind of consciousness that Block (1995b) calls 'access consciousness'. Thus, we believe there is a very important ambiguity in this debate. While we agree with Searle that phenomenal consciousness is essentially a biological process and that AI is severely limited with respect to simulating it, we agree with his critics when they claim that AI may be capable of simulating and truly achieving *access consciousness*. This is why the consciousness and attention dissociation is crucial for our purposes, because it states that attention is essentially related to accessing information (see Montemayor & Haladjian, 2015).

Our criticism of AI, therefore, is more nuanced than Searle's in three important respects. First, we limit our criticism exclusively to the type of consciousness that is characteristic of feelings and emotions, independently of how they are related to semantic contents or conceptual categories (i.e., phenomenal consciousness). Second, the limitations of AI with respect to simulating phenomenal consciousness will be independent of considerations about understanding the meaning of sentences. The limitations of AI that we outline will extend to other species, which do not manifest the capacity for language but which very likely have phenomenal consciousness. Thus, our criticism of AI is more truly based on biological considerations than Searle's. Third, and quite importantly, we grant that AI may simulate intelligence, rationality, and linguistic behavior successfully, and that AI agents may become intelligent and competent speakers just like us; however, we challenge the idea that they will experience feelings or emotions in the same way as humans. This has the interesting implication that AI agents lack moral standing, assuming that experiencing emotions and feelings is a necessary condition for moral standing.

Our criticism of AI assumes views that are not entirely uncontroversial. For example, some would object to the distinction between access and phenomenal consciousness, or like Searle, to separating intentionality from phenomenality. But we hope to show that our assumptions and criticism have several advantages over other views, including Searle's. One advantage is avoiding the ambiguity aforementioned. Another advantage is the emphasis on empathy. Empathy and the intensity of emotions have not been considered as central when challenging AI. This is a puzzling situation, given the importance of phenomenal consciousness for empathy, moral standing, and moral behavior. A contribution of this paper is to improve this situation by taking the intrinsic moral value of consciousness as fundamental.

To fully appreciate how the intrinsic moral value of phenomenal consciousness matters to AI's limitations, consider the fact that although semantic information can be easily copied, and programs with syntactic features may be reproduced and copied several times, it seems that the way a subject experiences the intensity of an emotion cannot be replicated. This non-semantic uniqueness might be the most important aspect of phenomenal consciousness. It certainly seems to be more important than the fact that the mind relates to semantic contents—however way those contents are defined (e.g., see Aaronson, 2016). Having made these clarifications, we now turn to our criticism of AI, not based on semantics, but on the importance of emotions and their intrinsic normative value.

While the idea that phenomenal consciousness cannot be realized in machines seems like an obvious conclusion to make, there are reasons to explore this issue further and more carefully. Advances in AI are quickening in pace, and as software and hardware technologies continue to progress there will be increased accessibility to more powerful machines that can perform more sophisticated computing. In the field of biocomputers, there are even developments of using enzymes to create "genetic logic gates" (Bonnet, Yin, Ortiz, Subsoontorn, & Endy, 2013) that could be used to build biological microprocessors for potentially controlling biological systems (Moe-Behrens, 2013). A related fear is that if we use living materials to build and run software, how are we certain that such organic-based technologies are not going to be conscious eventually? Of course, this is a compelling topic in science fiction that is not likely to be realized any time soon, but a proper discussion of this potential situation is important at this stage.

A key issue is that AI has expanded its goals beyond the original Turing test for intelligence and now tries to include more complex functions such as perception and emotion. This is a natural progression, given that perception and emotion modulate and give rise to many forms of cognitive activities associated with human intelligence (Pessoa, 2013). One may think that if this project succeeds and artificial agents pass not only Turing intelligence tests but also *emotional* Turing tests (Picard, 1997; Reichardt, 2007), artificial agents may achieve a level of conscious awareness similar to human beings. In fact, according to the most optimistic interpretation of AI research (e.g., Kurzweil, 1999), artificial agents may become sources of ethical and rational flourishing because they would not be subject to the biological constraints that humans inherit from their genetic lineage, thereby enhancing the possibilities for improvement in ways that are impossible for us mortals.

As mentioned, to clarify the potential for AI systems, it is helpful to frame this issue by considering how human consciousness and attention are dissociated, or what we call CAD for the "consciousness-attention dissociation" (Montemayor & Haladjian, 2015). Since human visual attention is now increasingly used to examine the nature of conscious experience, it is critical to understand how they are related. When you examine this relationship you find that there is a strong case for dissociation between attention and consciousness in humans. That is, the basic forms of attention do not require consciousness to operate successfully. Perception is supported by many mechanisms that operate outside of phenomenal consciousness, such as attention routines (Cavanagh, 2004). AI systems, like those associated with computer vision, can be said to possess forms of intelligence related to attention-based selective information processing for monitoring or search routines. According to CAD, such attention routines would not entail conscious awareness in humans. This means that even if AI reached similar or superior levels of intelligence based on attention routines, machines would still lack consciousness (since consciousness is unnecessary for these functions). Moreover, even if consciousness could be identified unambiguously in machines—which is not an easy task—there is the possibility that there are different types of phenomenal consciousness (Kriegel, 2015), perhaps only some of which could be susceptible of AI reductive computing. These would be related to how attention occurs without phenomenal consciousness.

In support of a dissociation between consciousness and attention, consider that the sort of phenomenal consciousness that is experienced by humans must be a more recent advancement in evolutionary terms—although it is related to visual attention, the two are generally separate processes as the more basic forms of attention developed prior to those associated with conscious attention (Haladjian & Montemayor, 2015). Abilities related to the selective processing of visual information, such as color, shape, and motion, are basic abilities found in animals and humans. These can be thought of as modules of perception that can be activated based on the environmental and task demands (Pylyshyn, 1999), and can be described as attentional routines (Cavanagh, 2004). From a computer science perspective, the halting problem (i.e., the termination of a function when its purpose is complete) is not an issue for abilities such as these basic forms of attention. There are computer programs to execute shape detection, object tracking, and face recognition. In contrast to these attention routines, phenomenally conscious experience does face the halting problem since consciousness is not something that clearly ends once a task is executed—it is an integrated unified experience that runs at varying degrees of activation (though its activation can be reduced when asleep or under anesthesia).

Another point related to evolution is that dexterous complex actions, which were genetically designed from millennia of evolution, are notoriously difficult for AI and machines to simulate. Using a familiar example, one can program a computer to beat any human in the game of chess, but it is very difficult to program a robot that could dexterously move the pieces of the chessboard like a human. Thus, even the attention routines associated with unconscious or implicit motor control will not be easy to reproduce, let alone their integration with conceptual knowledge and the kind of conscious contents that humans manifest in language.

This idea is related to Moravec's paradox, which is the problem that while abstract and complex thought is easy to compute, basic motor skills are very hard to model computationally. Hans Moravec (1988) explained this puzzling asymmetry precisely by appealing to evolution. Our species had millions of years to develop finely tuned skills, which operate unconsciously or automatically, while complex rational thought is a recent addition to our cognitive abilities. This line of reasoning must be carefully considered. One critical consequence of developing this point is that conceptual conscious attention must have evolved later than basic perceptual attention (Haladjian & Montemayor, 2015).

Yet, one does not need to accept such evolutionary arguments to appreciate how CAD makes the unqualified AI proposal problematic. The main issue to consider is that *while simulated intelligence may be intelligence*, *simulated emotion cannot be emotion* (Turkle, 2005/1984). This is because intelligence is basically computation and is not necessarily dependent on phenomenal consciousness, but human feelings are dependent on it—a distinction that has not been properly appreciated in the AI literature. This issue, as mentioned, is likely related to the problem that while programs can be easily copied, the phenomenal experience of a subject cannot. There is a divide between people that say simulated thinking will be enough to produce a human mind, while others argue that although simulated thinking can be considered thinking, simulated emotion can never be considered to truly be emotion (e.g., these ideas are presented in an episode of the Radiolab podcast, WNYC Studios, 2012). In essence, this is the distinction between access and phenomenal consciousness as discussed in the philosophical literature. In addition, human intelligence has an evolving purpose with ever-changing goals, something that AI has not yet achieved on its own, since it is always reliant on its programmer. These motivation and goal-setting abilities also face the problem of when to stop operating (e.g., a halting function), which is certainly not well-defined in human consciousness. Regardless of what one thinks about evolution and Moravec's paradox, these problems confront future research. Ultimately, we believe that there is no possibility to create an artificial phenomenal consciousness because of the empirically grounded implications of CAD for AI.

In the next sections, we will outline the challenges for creating an artificial intelligence system that has any sort of conscious experience. Henceforth, we simply refer to phenomenal consciousness as 'consciousness'. This discussion will require an examination of what AI systems are capable of now, how they are related to human abilities (particularly attention, emotions, and motivation), and the implications of the human consciousness-attention dissociation for AI.

2. The art of human intelligence

Since the age of digital computing, the human brain has been compared to the computer in attempts to better understand how it works. The field of cognitive science grew out of this tradition (e.g., see Pylyshyn, 1984). Scientists continue to explore the relationship between the brain and computers, which has generally taken the form of computational models of brain processes and has led to a better understanding of perception and cognition (e.g., see Reggia, 2013). On the flip side, this analogy of mind and machine also drives innovations in technology that aim to achieve human-like performance. Fundamentally, these advances in computing require an understanding of how the selective information processing mechanisms of attention works in humans.

2.1. Attention

The human mind is capable of performing many impressive actions and calculations, from basic information processing, to highly dexterous movements, reasoning, and the production of aesthetic experiences. If we are to understand how machines might achieve any form of consciousness, we must first outline our understanding of human consciousness and the related brain functions, such as visual attention and working memory. Attention is particularly important because it is how the brain selectively processes information obtained from sensory and memory sources. This ability is critical for an organism's survival—attention supports basic skills such as navigation, identifying sources of sustenance, identifying predators or conspecifics, and more complex tasks such as tool usage and social interactions.

What has been particularly informative is the research on the visual system, where attention has been studied extensively. Several forms of visual attention have been identified, including feature-based, spatial, and object-based. These forms of attention can occur automatically (e.g., from exogenous stimuli) or be more willfully directed (e.g., from endogenous sources). What they all have in common is that they can work toward the goal of selectively processing information in relevant ways to allow an organism to interact with its environment.

Feature-based attention is a more primitive information selection mechanism related to low-level perceptual processes. These information processing systems are organized according to specialized brain regions responsible for registering specific types of visual information, such as color, motion, or segment orientation (for a review, see Maunsell & Treue, 2006). Feature-based attention interacts with these low-level systems to select information in a typically automatic manner, but this selection process can be biased by higher-level signals based on task demands, including motivations. Different signals are involved in how information reaches various areas of the brain, some of them with different evolutionary origins and purposes.

Another form of attention, spatial attention, filters information based on spatial coordinates and is another important mechanism that evolved early in order to aid navigation and goal-directed motor behaviors. This attention "spotlight" can be focused on a specific region and shifted around as needed (Posner, Snyder, & Davidson, 1980), and can be made more diffused or a more focused "zoom lens" (Eriksen & Yeh, 1985). A distributed attention can capture quickly a statistical summary representation of the information outside of the focus of attention (e.g., Alvarez & Oliva, 2008), which also helps compute the overall properties of a visual scene. Computer vision has done a decent job of simulating both feature-based attention and spatial attention, though not yet as efficiently as the human visual system (e.g., Yang, Shao, Zheng, Wang, & Song, 2011).

Attention can also operate on things in the world that display object-like properties, such as cohesion, symmetry, and common fate (for reviews, see Chen, 2012; Scholl, 2001). Object-based attention requires a two-stage process that begins with the (generally automatic) individuation of objects (Pylyshyn, 2000). Selective attention operates upon these "indexed" items in order to bind object features, which are made available through feature maps (Treisman & Gelade, 1980), resulting in sustained object-based mental representations that allow object identification. Together, individuation and identification contribute to the experience of attending to specific items in the world and supports abilities like enumerating sets of items, tracking multiple objects, or attending to a single item in detail. Selective attention plays a crucial role in forming persisting object representations by allowing features from a visual scene to build and maintain a coherent representation incrementally in visual memory (Treisman, 1998, 2006). These mid-level "object file" representations (Kahneman, Treisman, & Gibbs, 1992) are generally considered the product of object-based attention. Another version of an object file is an "event file", which incorporates both features and motor commands (Hommel, 2004, 2007; Zmigrod, Spapé, & Hommel, 2009). This richer notion of object representations can combine cross-modal sensory representations and also integrates action-planning information.

These forms of attention can be automatic and often produced without any conscious awareness (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Mudrik, Faivre, & Koch, 2014; Mulckhuyse & Theeuwes, 2010; van Boxtel, Tsuchiya, & Koch, 2010). There are many examples of attention processing that do not reach awareness but yet still allow individuals to perform actions successfully, as in the case of blindsight (Kentridge, 2012; Kentridge, Nijboer, & Heywood, 2008). In principle, AI could develop search routines robust enough to simulate these forms of attention, although Moravec's paradox remains a caveat for the implementation of the mechanisms supporting these forms of attention.

Additionally, there is conceptual attention, which can also occur automatically in humans, but which requires a semantics that would be hard for AI systems to emulate (this is basically the focus of Searle's criticism). Exactly at what moment conceptual attention plays a role in object-based attention is a subject of debate, but it clearly plays a critical role in higher-level human visual attention and involves some level of understanding and judgment making. Machine learning programs, such as those implemented in object recognition, already implement some form of implicit conceptual attention, although required caveats are needed here as well. Thus, the prospect of AI forms of conceptual attention that resemble or even outperform human conceptual attention do no seem as remote as 30 years ago, roughly when the Chinese room debate was being discussed.

With respect to agency and motivation, while some complex tasks do require an engaged and sustained effortful attention, like for example when learning a new skill or in cases of perceptual learning (Meuwese, Post, Scholte, & Lamme, 2013), there are also some complex attentional processes that can be so engrossing that they begin to feel effortless (Bruya, 2010). This effortless attention is particularly relevant for emotional engagement because it seems to be related to expertise and is suggestive of how memory systems can interact with attention to influence the conscious perception of effort and time. Effortless attention, in this sense, is a more controversial form of attention (compared to other higher-level attentional processes) and has only recently been discussed in the literature. Similarly, the feeling of "flow" is related to effortless attention, where one's focus is on the mechanics of a physical activity with very little effort or attention to other forms of stimulation (see Csikszentmihalyi, 1997). This kind of attention, therefore, cannot be reduced to search routines and selection processes, and requires levels of cognitive integration that may prove too challenging for AI. The feeling of effort, however, would presumably be irrelevant for such AI simulations. Therefore, the real challenge is how to understand a *subject's motivation*, which is fundamental to attention routines, in AI systems. Solving this problem in AI may not be insurmountable if AIs learn complex forms of decision-making. But, as noted, what they would lack is any experience of effort or effortlessness.

Another aspect of attention to consider is the ability to attend to different mental states. For example, attention to emotions related to features of the environment depends on an older neural network that recruits systems for immediate action and arousal. Attention to features, such as color, can occur in unison with attention to emotions, but the neural correlates of these different networks can be distinguished as independent from each other (Pauers, Kuchenbecker, Neitz, & Neitz, 2012). This strongly suggests that integrating the vast forms of attention networks in humans depends on more than mere search algorithms. In fact, the best way to understand how these different forms of attention are integrated is in terms of the motivations and goals of an agent, and this is precisely what AI systems seem to lack (for an account of how attention necessitates agency, see Wu, 2011, 2013). Therefore, just from considerations regarding attention, one can argue for a much more pessimistic view of AI emotional intelligence because of the lack of clear cognitive motivations. The research on consciousness provides further insight on how this integration between motivation and experiences becomes more complicated, given the consciousness-attention dissociation. Nonetheless, problems stemming from motivation and attention are not as serious as those regarding consciousness. (For instance, AIs may be able to recognize color better than humans and make complex decisions based on such recognition, but it is unlikely that they will feel the emotions that humans feel when looking at a beautiful painting.)

2.2. Consciousness

In the philosophical literature, it is now standard to distinguish at least two senses of 'consciousness' that have critical implications for AI: phenomenal and access consciousness. Because of its importance, we shall restate a few key points concerning this distinction. Phenomenal consciousness is what most people talk about when they use the term 'consciousness'—it is the subjective experience of being aware of a certain thought or feeling. It is what you are experiencing as you read this sentence and includes all the sensations that you are experiencing, like the sound of the refrigerator in the background and the chair against your back and the slight feeling of hunger in your stomach. Access consciousness is a little more complicated to describe. It is a form of consciousness that allows representational content to be made available for use by various cognitive systems for thought and action (Block, 1995b). In this sense, access consciousness relates to the basic forms of attention, since attention serves the role of processing information within the brain in a task-relevant manner to guide action or thought, which are also distinctive features of access consciousness. For our purposes, access consciousness is not distinct from attention since these basic forms of selective information processing can be programmed into machines (e.g., attention routines) without requiring an accompanying subjective phenomenal experience. What is important to elucidate is how phenomenal consciousness might be possible to implement in machines. While we can have AI with elements of access consciousness, is it possible that they would ever feel anything related to that content?

Another form of consciousness relevant for our discussion is self-awareness, which allows self-reflection and is tied closely to subjective phenomenal consciousness. Whether or not animals possess self-awareness remains debatable, but some have proposed ways in which basic consciousness might be identified in animals (Edelman, Baars, & Seth, 2005; Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). For example, Bayne's (2007) theory of "creature consciousness" specifies whether or not an organism can be said to be phenomenally conscious by requiring mechanisms that generate the "phenomenal field" (possibly related to activity in the thalamus), and neural inputs from the different cortical areas responsible for processing sensory and memory-related information. Only after integrative processes occur can consciousness be considered present in animals, but such claims require empirical support. Currently, problem-solving behaviors (e.g., tool usage) in animals provide the best indication of the possible presence of conscious attention in animals (for a review on animal consciousness, see Griffin & Speck, 2004). It is likely that such behaviors, however, could depend just on access consciousness. In any case, reflective self-consciousness can be distinguished from other forms of phenomenal consciousness, and further complicates the issue of agency and motivation for consciousness and attention as it appears to require a phenomenal component.

There are several empirical studies on consciousness that describe the likely structures that support it. In general, the brain requires some level of recurrent processing and not just the feedforward movement of information (Di Lollo, Enns, & Rensink, 2000; Seth & Baars, 2005; Tononi & Koch, 2008). Complex neural networks with recurrent processing, especially those that have signals originating from the frontal cortex, are considered later adaptations. More deliberate forms of conscious attention also are associated with activations in the "newer" brain areas like the prefrontal cortex, and are supported by working memory systems (but there is no conclusive evidence that the frontal cortex is necessary for conscious awareness). Similarly, consciousness can be described as having different levels of activation, with some events remaining "preconscious" and others entering full awareness (Dehaene et al., 2006). These ideas are presented under models such as the "global workspace" or "broadcast" views of consciousness (Baars, 2005; Dehaene & Naccache, 2001). Under these accounts, the main adaptive purpose of conscious attention is to "broadcast" contents that are computed in a uniform format (presumably conceptual). This plays an important epistemic role since it allows access to contents across different modalities for the purpose of providing information to support goal-driven actions and beliefs. This may be one of the crucial roles of consciousness—to integrate complex information from different modalities (but some argue that integration can occur outside of awareness, e.g., Mudrik et al., 2014). In fact, conscious attention, where attention and consciousness clearly overlap, might be the only aspect of consciousness that has a functional role for complex organisms (Haladjian & Montemayor, 2015).

Although we are far from fully understanding how the brain supports consciousness or what its evolutionary purpose may be, there has been a substantial advancement in what we know about it. While consciousness is closely tied to attentional processes, consciousness and attention cannot be simply reduced to one another. Attention often operates outside of consciousness (Koch & Tsuchiya, 2007) and most likely appeared before consciousness (Haladjian & Montemayor, 2015). The consciousness-attention dissociation is important to emphasize, since it has profound implications for artificial consciousness. For example, even if technology develops to the point of exactly simulating human attention, goal-oriented motivation, and global access, being able to achieve phenomenal consciousness, including that for emotions, would still be beyond reach.

2.3. Emotions

Emotions play an important role in human life. They regulate mental states in order to produce certain behaviors and guide the organism in a contextually relevant way (e.g., producing positive emotional states in rewarding situations like mating or eating, or anxiety in response to fear). Emotions also influence changes in physiology and bodily states, such as the quickening of the heartbeat, pupil dilation, and tensing of muscles. Thus, emotions are closely related to the neurophysiological state of the brain and body through the nervous system, and can be critical in influencing the ability and the manner in which we act. Emotions also are a large phenomenal component of conscious experience and subjectivity.

Distinctions between three possible types of consciousness illustrate three ways of cognizing emotions: by *experiencing* them (phenomenal consciousness), by *accessing information* through them (access consciousness), and by *attributing them* to oneself and others through a judgment (self consciousness). It is the *experiencing* way of cognizing emotions through phenomenal consciousness—arguably the most fundamental aspect—that presents serious challenges for AI. It is also the experiencing way of cognizing emotions that make our conscious lives subjectively valuable and interesting (e.g., listening to a moving piece of music, admiring a sunset, tasting a complex wine).

To further clarify what counts as an emotion, it includes basic "primary" emotional responses such as fear and anger (Ekman, 1992) and more complex evaluations of situations and empathy, which can have moral implications (Decety & Cowell, 2014). The circuitry that contributes to primary emotions is evolutionarily older than the circuitry of long-term and sequential planning, and is present in animals that can display basic emotional responses such as fear responses (LeDoux, 2000, 2012). While animals do exhibit basic emotions, whether or not they have the same subjective "feeling" that we do is difficult to determine. Nevertheless, it is worth pointing out that animals do have similar physiological reactions as humans, at least within the context of basic emotional responses that rely on similar brain circuitry, and therefore, may be conscious of those emotions.

An important feature of emotions in humans is that they include the functional aspect of having an "emotional state", which facilitates relevant actions within the context of the current environment, as well as experiencing "feelings", which are the self-reflective conscious experiences of emotions (Tsuchiya & Adolphs, 2007). Human feelings introduce the variability associated with the subjective interpretation of emotional states, something that seems impossible to replicate in a computer. The systems that support emotional processing in humans are closely tied to those responsible for cognition (Pessoa, 2008), so emotional feelings generally can be considered a more advanced form of mental activity, especially when it concerns moral and aesthetic judgments. Feelings also seem to be fundamental to the sense of self and self-awareness (Damasio, 1994), which further suggests that the presence of feelings requires higher-level interactions than basic emotions.

To better understand emotion's role in cognition, it is important to examine the related research from the neuropsychological perspective (for good reviews, see Pessoa, 2008, 2012). The neurophysiology of emotions can be considered a basic evolutionary advance for organisms. For example, the limbic system (particularly the amygdala) is closely associated with generating emotional responses and seems to be the only one that can "hijack" the cortex and take over the computational, rational processing of the brain (LeDoux, 2000, 2003). Fear responses seem to serve the purpose of engaging an organism to attend to critical aspects of the environment and takes over cognitive activity. Much research has shown that stimuli with some sort of emotional content tends to activate more extensive cortical areas of the visual system related to attention (Pessoa, 2013; Pessoa, Kastner, & Ungerleider, 2002), and may even be detected outside of conscious awareness, or at least has a very low threshold for detection with minimal awareness (for reviews, see Mitchell & Greening, 2012; Pessoa, 2005).

Emotional systems are found in both animals and humans (LeDoux, 2012), which presents the question of whether or not emotions rely on consciousness. The fear response your housecat is able to display produces a threatening stance towards the aggressor and is also a way of engaging the animal to be extremely alert to make appropriate responses. Conversely, the seemingly blissful state of a purring feline also appears to indicate a certain brain state that suggests it is not agitated (since we really cannot know what cats think, that is the best we can deduce). These basic emotional states must serve some purpose in preparing an organism for relevant actions while also communicating to other organisms. There must be something it is like for the cat to undergo these states (phenomenal consciousness) and there is also information that is being used for immediate action and decision making (access consciousness); the experience may be distinct from the information it conveys, which could even be conceptual, as in the case of humans.

Although emotions seem to be an important part of our conscious experience, there is research that suggests some emotional cues can be processed non-consciously (Tamietto & de Gelder, 2010), which calls into question the overall role of awareness for functional responses to emotions (Pessoa, 2005). Along these lines, a distinction has been made between emotions and feelings and the role of consciousness in these, with conscious experience being a central signature of feelings (Tsuchiya & Adolphs, 2007). Thus, the distinctions between consciousness and types of attention become fundamental. There might be access to emotional information without consciously feeling an emotion and there may also be unconscious attention routines that process such information and affect behavior without producing any type of awareness.

A more dramatic example of the dissociation between a basic emotional response and the subjective feeling associated with it is seen in people who are born with a congenital insensitivity to pain (see Heckert, 2012). This neurological condition prevents the individual from subjectively feeling the stimulation that typically would cause pain, as well as the associated emotions of fear and aversion. In such cases, it is impossible for the individual to understand what pain is and what others might feel when experiencing pain. Thus, it seems that feelings rely on the detection of the neurobiological signals that produce emotional responses and distinct phenomenal experiences like pain or pleasure—all of which are crucial for producing empathy.

The question to ask about emotion is whether or not this type of physiological response, based on instinct, experience, and even reasoning, can be not only implemented or simulated, but also actualized in AI. In terms of an emotional Turing test, how many subroutines do you need to include in an AI system so that you get the same kind of complex emotional responses seen in humans? A distinction between what is achievable and what needs to be achieved becomes obvious. A key argument to consider is that all such functions and routines for emotion simulations in AI would depend on halting thresholds (i.e., the point where a computer program should stop), essential to the very notion of computability. But by definition, phenomenally conscious states are not reducible to such routines. The experience of regret, for instance, is not simply the end point of running an attention routine. On the contrary, it is a complex state that cannot be reduced to any simple halting function. The essence of such experiences is to engage the subject as a whole, not just for reaching a specific inferential conclusion, but essentially to affect and shape the subject's entire conscious awareness.

2.4. Empathy and moral reasoning

Empathy, the capacity to feel or understand the subjective perspective of conspecifics, is fundamentally related to emotional responses. This requires the agent to have some type of theory of mind and an understanding that similar agents will possess analogous emotional states. We cannot empathize without knowing this relationship between our self and another, which appears developmentally in a child's second year of life (Zahn-Waxler, Radke-Yarrow, Wagner, & Chapman, 1992). The ability to empathize relies on self-consciousness and self-recognition, which also develops around two years of age (Rochat, 2003; Rochat & Striano, 2000). This form of consciousness seems to develop with experience, particularly of the social kind. Some would argue that this ability for social perception is the basis for consciousness in general (Graziano & Kastner, 2011). While machine learning in AI can achieve many things, it is questionable whether or not it can ever develop empathy (e.g., see Miner et al., 2016).

Moral reasoning is at least partly based on this ability to empathize. This has an important consequence: ethics and morality seem to necessitate phenomenality in typical human moral reasoning, but displaying intelligence does not necessarily require phenomenal experiences—they happen in tandem in humans, but they can be, and generally are, dissociated. This, we propose, is a critical point for understanding the challenges that confront AI. It is one thing to be capable of *detecting* emotions and running rule-based algorithms to reach a conclusion at a halting threshold. It is an entirely different achievement to be able to *empathize* with others based on how we feel. Aesthetic judgments are related to moral judgments, and also require phenomenality. Although it may be controversial to claim that moral and aesthetic evaluation necessitate conscious experience, we take it to be assumed implicitly or explicitly in most theories of moral and aesthetic judgment.[1]

When considering ethics and morality, we are presented with another problem that faces the implementation of emotions in AI systems (e.g., Picard's affective computing account). Because of considerations concerning utility and value (see Kahneman & Thaler, 2006), it seems plausible to conclude that emotional background is narratively structured and not exclusively utility structured. Research shows that the more utility-based a person is, the less inclined she will be to attend and respond quickly to morally relevant stimuli (Haidt, 2007). One should not take this evidence from neuroscience and psychology as confirmation of an ethical or moral perspective—psychology *describes* phenomena while morality *prescribes* actions. But there is undoubtedly a fundamental and even constitutive relation between phenomenal experiences associated with moral feelings of approval or condemnation and any conceivable moral theory (Carter & McBride, 2013).

If the dissociation between consciousness and attention, which we explain below in more detail, is taken into consideration, one can easily see that AI agents will be able to reason their way through the inferences of a developed ethical theory, but still lack any significant form of morally appropriate experiences. Having an ethical theory is not the main obstacle for the implementation of morality in AI. The real obstacle is the subjective, empathic nature of moral experiences. Having such a theory, therefore, is not what is most distinctively *human* about morality. Rather, human morality is comprised of our strong and biologically rooted reactions to the pain and suffering of others. Our reactions express who we are, and they need not be mere knee jerk-like reflexes. On the contrary, they are constitutive of our responsiveness to others, and ultimately, of the expectations we typically assume as morally adequate (Strawson, 1962).[2] Ethical theory and search algorithms for detecting emotions are good for simulating ethical behavior. But if CAD is right, they are incapable of truly producing it: for that you need phenomenal consciousness.

2.5. The consciousness-attention dissociation

The relationship between consciousness and attention is one that has been explored with more scrutiny in recent years, particularly in attempts to understand human consciousness via empirical research on attention. While we will not go into great detail about this dissociation here, it is generally agreed that the two are largely independent—we call this the consciousness-attention dissociation, or CAD (see Montemayor & Haladjian, 2015).

There are several reasons why phenomenal consciousness and attention can be considered dissociated. First of all, research on the neural relationship between the two has found them to have overlapping networks but generally to be distinct brain processes (Koch & Tsuchiya, 2007, 2012; Lamme, 2004). For example, some studies have identified that additional neural networks are activated for conscious report than for simple attentive processing (see Dehaene & Naccache, 2001). The case for dissociation is also made through studies on blindsight (Kentridge, 2011, 2012; Weiskrantz, 2009) and visual illusions where the information that enters conscious perception does not match the more accurate information sent to

[1] Consider the classic utilitarian principle that one must reduce the amount of pain and maximize happiness or well-being, or the Kantian principle that human life is intrinsically and categorically valuable, not just instrumentally valuable.

[2] A similar point about aesthetic judgment can be traced back to at least Burke (1757).

execute motor actions, such as making eye movements (Lisi & Cavanagh, 2015; Spering & Carrasco, 2015). There is also evidence suggesting that one can attend to the emotional content of stimuli that are not consciously seen, which is considered an affective blindsight (Celeghin, de Gelder, & Tamietto, 2015). Moreover, it has been shown that unconscious processing can influence decision making (Newell & Shanks, 2014).

Another point to consider is that there may be forms of phenomenal conscious emotion that are not particularly guided by voluntary or routine-based attention. In fact, the most interesting forms of conscious experience for emotion may be of this kind. For example, what makes an aesthetic experience powerful is not that one wants it to be powerful or that one wants simply to detect the colors of the sunset over the harbor in front of them. This would be a rather poor and inaccurate description of aesthetic experiences typically experienced by humans. Such experiences are a combination of the results of attention routines plus a subjective experience that combines conceptual information as well as emotional states in the form of feelings.

Color vision presents another example of dissociation. Researchers have found that trichromatic color vision depends on a single genetic addition of a third light-sensitive protein called opsin, rather than neural modifications during cognitive development (Mancuso et al., 2009). With the single introduction of a new gene, a brain that was not habituated to respond to a whole range of color gains the ability to distinguish it. We shall call this ability *recognitional* color vision. Pauers et al. (2012) identified a different cognitive network depending on melanopsin that independently regulates emotional reactions and circadian regulation involving color, which is evolutionarily older. We will call this capacity *emotional color vision*. They argue that melanopsin can influence the circadian system, which consequently affects emotional regulation (Tucker, Feuerstein, Mende-Siedlecki, Ochsner, & Stern, 2012), even when the cones and rods in the eyes are "disabled", for example, when there is natural degeneration of photoreceptors or in laboratory conditions when there is a controlled exposure to a constant, bright light. This indicates that there are separate neural systems that send information for experiencing a color and for experiencing a related emotion to that color. Crucially, the visual system communicates with the limbic system through a different network from the one it uses for color *detection*. This research by Neitz and colleagues (e.g., Mancuso et al., 2009; Pauers et al., 2012) has the remarkable implication that while our capacities to distinguish specific shades of color may differ from person to person, our emotional responses are independent of such capacities and seem to be much more uniform across individuals. What we intuitively classify as our conscious experience of the colors we see depends on the cones rather than on the older circadian circuitry, which explains why our emotional reactions are much more uniform than the colors we subjectively experience. Thus, genetics and neuroscience justify a distinction between recognitional and emotional color vision.

Here is how CAD generates a concrete problem for AI vision. Even assuming that color recognition in AI could exactly simulate cone-color detection (which is not an assumption that should be easily granted), such "experiences" of color would be unlike those experienced by humans, where they are also integrated with experienced emotions. The CAD framework entails that experiencing an emotion will not just be a matter of recognizing it through attention, but will require a deeper empathic component. That is, attention to color does not necessarily entail emotion. Since the experience of a typical human is unified—it is not the mere addition of color detection plus emotion detection—the colors that AI agents will detect will never correspond to typical human color experiences.

Additionally, cross-modal integration for emotions, decision-making, and attention to social cues may be semantically integrated, in a way that mere simulation may not capture (e.g., attending to a sardonic versus honest smile). Sarcasm for example, is notoriously difficult to be detected by AI systems, let alone be truly understood or experienced (Joshi, Bhattacharyya, & Carman, 2016). Conscious attention, understood phenomenally, provides a stable unity of information for the sole purpose of producing intense immediate engagement and empathy, in an automatic way, for both detection of emotions and also for empathically understanding them. This could be seen as a functional role that consciousness plays, since it enables a system to engage with its immediate surroundings, including social contexts, by diverting important cognitive resources to relevant objects or events.

Although not much research has been done on the evolution of attention and consciousness (but see Edelman et al., 2005), there is growing interest in exploring what an evolutionary approach can tell us about human attention and consciousness (Cosmides & Tooby, 2013; Graziano, 2014; Haladjian & Montemayor, 2015). The main conclusion that becomes clear is that attention is a basic mechanism that evolved early to help organisms interact with their environment. As interactions grew more sophisticated (e.g., social behavior), more complex forms of attention were required. Phenomenal consciousness, while its purpose is still debated, had to evolve after the basic forms of attention for detection of features. Therefore, one must agree that there is some level of dissociation between consciousness and attention based on the evolution of the nervous system, because attention's primary and earliest purpose is to generate action schemas that respond to features, rather than generate conscious awareness and socially engaging experiences.

Attending to and recognizing emotions, as mentioned, shall not be identified with experiencing them. Emotions and feelings also have a deeply social dimension. The strong tendency to reduce these features to mechanistic algorithmic routines may work for a vast amount of attention routines, but not for feelings and agency (at least moral agency). The conclusion that it is impossible to implement consciousness in AI is not based on some humanistic type of fervor, dogmatic adherence to the "hard problem of consciousness", or intuitions about semantics and the "Chinese room"; rather, it is based on empirical evidence, considerations about evolution, and the sociobiological functions of emotions. It is also based on our contemporary understanding of computation and AI systems.

Purpose and motivation are essential to many attention routines, and to the extent that AI systems lack intrinsic motivations (which they lack absolutely as things stand now), they cannot be considered agents. Even if only attention is considered, AI systems are quite limited. CAD complicates the picture in a more fundamental and principled way: *even if* AI systems managed to have motivations, those would be motivations to detect and encode in an instrumental way, rather than intrinsic motivations based on non-instrumental and empathic ways of understanding how it feels to experience a concrete emotion.

In sum, when examining the crucial selective information-processing abilities of attention, it becomes clearly dissociated from conscious awareness in many instances. For the case of emotion, it is more complicated. Emotion can operate on the level of basic responses to environmental cues, like those that generate a fear response. But it can also be more involved and be experienced as a feeling within our conscious awareness. When an intense emotion is felt, whether a basic emotion like anger or a complex emotion like shame, it is a very distinctive subjective experience that can have profound effects on our thoughts and actions. It is unlikely that it will be possible for AI systems to have such an experience, for the reasons examined above. Any such "experiences" in AI would be simulations and not genuine subjective experiences.

3. The art of artificial intelligence

The challenge of modeling emotions in AI and the dissociation between consciousness and attention do not mean that AI systems will not be incredibly transformative and useful. On the contrary, artificial intelligence, in terms of storing facts and performing rule-based reasoning, has already changed our world through sophisticated computing systems. Following advancements in the mid-twentieth century, which include Alan Turing's proposal for a general computing machine (Turing, 1950), technology has been advancing at an impressive rate. Computing ability has been growing exponentially for the past forty years, increasing at a pace that follows Moore's Law, and is expected to continue to do so, at least in the near future (Bauer, Veira, & Weig, 2013). Another factor influencing technology is the vast storage of knowledge and the global exchange of information via the internet, which also has grown exponentially and is expected to nearly triple in volume between 2014 and 2019 (Cisco Systems Inc., 2015). Along with these abilities for the computing and the transmission of information, it is inevitable that the next frontier for attention-grabbing breakthroughs will be in the area of artificial intelligence.

3.1. Recent developments

The recent work in AI is impressive and includes intelligent systems that accomplish many complex human-like tasks. When looking at computer vision, for example, abilities like object tracking and facial recognition were designed on principles based on the human visual and memory systems and have allowed advances that support innovations like self-driving cars and complex surveillance systems. The next expected advancements are even more impressive, with the potential for developing general purpose AI systems that can learn without guidance and even exhibit signs of creativity.

Artificial intelligence has become ubiquitous. There are many commonplace instances of AI that range from machines for industrial production to everyday household objects like smartphones. Although not yet perfected, smartphone applications playing the role of personal assistants that answer questions based on voice commands are the clearest examples of basic AI reaching a widespread audience, with almost two-thirds of adults in the United States owning a smartphone in 2015 (Pew Research Center, 2015). We are on the verge of having self-driving cars on the roads, as well as other types of automated transportation, which inherently require considering "ethical behavior", for example, programming what actions should be made when such automated machines are presented with situations of unavoidable accidents. Other applications that have the potential to reach a wider group of people include the use of robotic assistants in medical settings (Shibata, 2004; Turkle, Taggart, Kidd, & Dasté, 2006), which may be particularly useful in elder care (Robinson, MacDonald, & Broadbent, 2014). While some aspects of medical care can be replaced with robots, it seems there will be something missing when diagnosed by a machine and not a human, partly because of the lack of empathy that we have been emphasizing.

Some advocate the potential of the "internet of things", where household appliances and other devices are connected to a centralized computing system, promising a futuristic home setting once only envisioned in cartoons. This has the power of making life more automated and efficient (e.g., your grocery list is updated when your refrigerator senses that you have run out of eggs). It also has the power to know a lot about people's behaviors, with implications for consumer possibilities that include marketing relevant products for purchase, helping reduce the waste of energy resources, or identifying critical biometrics that require medical attention. Naturally, such gathering of information generates privacy concerns and this ethical issue is part of the current debate. Such innovations show how the connectivity of devices and services, which gather much information about the user, can inform predictive algorithms to tailor a user's daily experience beyond the current tailoring that happens when browsing or shopping on the internet.

Advances in machine learning play a large part in how much artificial intelligence can be improved. The ultimate goal is to develop some sort of AI system that is capable of autonomous learning from raw data, which has already been achieved in terms of playing video games (see Mnih et al., 2015). The significance of a computer program recently winning the game of Go (Silver et al., 2016) is that the possible moves for this game are so vast that they cannot be computed through brute-force calculations—it required a level of deep machine learning (and learned decision-making) that stresses the importance of general purpose algorithms. This type of artificial intelligence is truly impressive and at the forefront of current research and advances.

Along with these advances in AI, the field has moved beyond the design of intelligence in computing systems and into the realm of reproducing more complex human abilities like perception and emotion (Li, 2014). These more complex human abilities do in fact modulate activities related to intelligence in humans, so it is only logical that computing advancements would lead to attempts to model and integrate more complex systems into AI. Indeed, one can find many similarities between advances in technology and biological evolution (Wagner & Rosen, 2014).

In computer vision, for example, there are now programs that perform object tracking and face recognition, as well as object recognition to aid visually-impaired people. Such complex abilities for object recognition evolved more recently in humans, with some arguing that consciousness is required for holistic face processing (Axelrod & Rees, 2014). When a machine recognizes a face, however, it is based on stored information about images and uses comparison algorithms to determine the level of similarity between two images—a relatively slow and imperfect procedure. It operates as rudimentary forms of feature-based attention routines. No consciousness is required, which some may argue is possibly the missing piece that needs to be included in order to improve computer vision. In any case, these are all basic attentional capacities, and if CAD is taken into account, they need not entail any form of conscious experience.

There are more focused attempts at incorporating principles from the human emotional system in product design and computing systems (Ahn & Picard, 2014a, 2014b). The commercial use of such technologies include the tailoring of user experiences based on emotion recognition (Bradshaw, 2016; Weintrauboct, 2012). The Kismet robot, for example, is programmed to detect emotions in facial expressions and respond accordingly, which often gives the impression of interacting with a living being even though it is simply a mechanical object with a set of programmed responses to detected emotional cues (Breazeal & Scassellati, 1999, 2002). Nevertheless, the experience of interacting with this robot is quite compelling. What some argue is that with these mechanical instantiations of emotion, we are in danger of losing a sense of authenticity; that is, we may be substituting authentic human relations with simulated ones, which could have negative implications for society (Turkle, 2007). Additionally, being able to relate better to AI through an emotional connection may encourage us to protect these machines like we do our family or pets, further complicating the social role of machines. Nevertheless, the implementation of human-like emotions into AI systems is an attempt to improve human-computer interactions in the sense that they may be able to elicit more willingness for humans to interact with them and provide more contextually relevant responses by computers.

A recent line of research—Rosalind Picard's affective computing program (Picard, 1997)—has assumed that artificial agents may be designed to pass the emotional Turing test, and in fact must do so to be truly intelligent. Technology has achieved much in terms of recognizing emotions and measuring physiological changes (e.g., due to stress or frustration), and this information can be used to provide personalized feedback or adjust a machine's performance (Picard, 2002a; Picard, Vyzas, & Healey, 2001). The development of this computational account of emotion suggests that emotions can be understood by machines somewhat reliably and thus can be reduced to algorithms to some degree (Picard, 2002b, 2007). This ability to understand human emotion has clear implications for product development and marketing (e.g., Ahn & Picard, 2014a), but also for making human-computer interaction more relevant and meaningful. While Picard acknowledges that computers may not achieve the level of conscious awareness that humans have, she argues that computers can achieve a minimal sense of conscious awareness, including self-awareness. For this proposal to make sense, of course, one must know what is meant by 'conscious awareness'. Picard proposes a very flexible and general account; however, when using more rigorous definitions, the scientific prospects of emotional AI are bleak. This is exemplified by the clear dissociation between forms of attention and forms of consciousness. Here is where we clearly encounter limitations in AI.

3.2. Limitations of artificial consciousness

We shall restate the limitations previously mentioned in order to emphasize that we are not promoting fear or anger about AI. We do not believe there is a "dooms day" scenario about to happen, where humans become instruments of incredibly clever machines. Like many others writing on AI, we prefer to take a more productive and moderate approach; however, we believe that proponents of AI have been overoptimistic about the prospects of artificial consciousness. Perhaps the most significant obstacles to any form of artificial consciousness are versions of the halting problem described above. These are strictly technical or computational problems that become even more problematic if one considers CAD.

One version of the halting problem is a *motivational halting* problem. Humans and animals initiate attention routines based on purposes and goals. It is not clear how to even model human-like motivation in an AI system, and this concerns just simple forms of attention and not consciousness. This means that it is difficult to even understand basic agency in AI systems. A related but much more difficult problem concerns consciousness. AI systems are defined in terms of halting functions. By definition, however, the qualitative character of experiences is not the result of any halting function. This *experiential independence* of consciousness from halting functions is deeply related to the unity of human consciousness and to the guidance and immediate relevance of how we experience emotions. It is also deeply related to the fact that unlike programs, consciousness does not seem susceptible of copying or reproducing. A function halts, like an attentional routine halts, but consciousness never halts nor does it depend on the conclusion of a program. It is plausible to think of CAD as delineating halting functions for attention routines and that conscious phenomena are not susceptible to such a computational approach.

There are also other problems that are important to revisit in the light of these technical problems. Going back to Moravec's paradox, Turing computability was a wonderful achievement that nevertheless lacked the sophistication of complex biological agents like humans, partly because dexterous performance requires precise motor control and motivation. It is clear that the success of computational systems concerns broadly epistemic functions, such as the use of conditionals and specified routines. Computational systems are present throughout nature and can be considered to be forms of intelligence, but what human intelligence possesses that goes beyond other forms of intelligence is a sense of purpose that is biologically, historically, and culturally determined (see Wolfram, 2016). Furthermore, feelings, and their interface with rationality, are narrative-dependent. And like the basic biological and motor-control functions, basic emotions belong to an older evolutionary repertoire of cognitive skills (that likely existed before human-like phenomenal consciousness developed).

Additionally, the ability to empathize highlights how difficult it will be to achieve emotional intelligence in machines. Empathy is fully dependent on phenomenality—it seems impossible to imagine empathy in others without an associated phenomenal experience. Simply reflecting on that sentence proves that point. One can be in full agreement with Picard and others that emotions are rational and computationally implementable, but that is not very informative, specifically when it comes to the computability of feelings. Furthermore, awareness of motivation seems to essentially require empathy when one considers paradigmatic cases of emotional engagement.

This complexity helps explain why much of AI focuses on step-by-step reasoning and simple motor skills. Intelligent behaviors like playing chess or simple movements through the environment (e.g., like a robotic vacuum cleaner) are relatively simple to program, especially if the mechanics to physically execute the behaviors are basic and have dedicated mechanisms to execute these functions (e.g., using sensors to avoid obstacles, which can be considered a rudimentary form of spatial attention). When you introduce high levels of variability, like those present in dynamic environments such as busy city streets, it becomes much more of a challenge. Take self-driving cars, for example. Although they can navigate along highways and avoid obstacles pretty reliably, there are instances where they fail. A somewhat comical example is when multiple cars arrive simultaneously at an intersection with stop signs, an autonomous car often freezes and cannot move through the intersection because the programming is rigid and must wait for the other cars to stop completely, which humans rarely do (Richtel & Dougherty, 2015). Here is where social cues are important, like when one driver is slightly more bold than another and would assert herself when faced with a hesitating driver.

Being able to recognize and interact with other objects, which object-based attention in humans does quite easily, still needs to be improved in computer vision. Interestingly, these computer programs are also faced with philosophical "moral dilemmas" in cases of unavoidable accidents—sometimes an AI system must decide whether or not it should sacrifice itself, and its owner, in order to cause the least amount of damage or loss of life (Heikkila, 2016). How to deal with these situations is particularly challenging—should these AI systems choose self-preservation, which includes the owner's benefit, or be at the service of the greater good? Notice however, that the difficulty avoided by implementing utilitarian rules does not solve the problem of what could *motivate* the AI system to do anything in the first place.

Another example of how such reasoning can be implemented considers social evolution and Robert Axelrod's research on game theory (Axelrod, 1984). According to this approach, AI systems are just part of a game that operates by manipulating us as pieces in the game. One certainly can model such games on a computer and in AI, and this would lack any sort of phenomenal experience or even a clear form of integrated attention. Some think these models are important for understanding where consciousness might exist in AI monitoring and control systems. Graziano, for example, argues that these models are similar to how the mind might model information processed by attention that reaches conscious awareness (Graziano & Webb, 2014). By representing these models, consciousness occurs when one is aware of these representations. Nevertheless, even if mental computational models of attention support consciousness in humans, this theory still does not address the phenomenal question adequately, that is, how a mechanical system can become consciously aware. It also faces the homunculus problem without a clear explanation of how one avoids the infinite regress problem of how a system achieves awareness of mental representations.

Human intelligence also includes empathy, and as we have argued, this creates a much more complicated problem for AI. Even if the motor skills and step-by-step reasoning can be fully implemented in robotics, we would have no understanding of empathy in the AI sense. At best, AI will produce sophisticated epistemic agents that can accomplish specific goals. AI cannot produce morally complex emotional agents with phenomenal experiences, motivations, or emotions. All simulacra can be based on attentional-functional routines, but simulacra cannot reproduce the grip, motivation, and immediate urgency of phenomenal consciousness that contains emotional content. In terms of morality in AI, without empathy there is no clear path toward a sense of morality other than what can be programmed, which then makes ethical considerations reliant on the creators of the computing systems (Rinesi, 2015). Ethical aspects of AI become increasingly important as they are implemented in more objects and these objects become connected and controllable via centralized computing systems. Although some computer programs can learn a form of normative morality by reading text and then begin to behave in ethical ways based on the content of the text (Riedl & Harrison, 2015), they cannot truly empathize and understand social interactions, and this area needs to be considered more carefully. This lack of empathy becomes especially important when assessing the utility of AI programs that become a more integrated part of society (e.g., see Miner et al., 2016).

Considering emotions from a neurophysiological perspective, they are basic systems that evolved before these emotional states were "felt" and long before the ability to empathize with others over these states. So it becomes even more apparent that while it may be possible to program attentional systems and even emotion-like responses, these artificial systems will not necessarily possess a phenomenal subjective experience. Even if cognitive scientists were certain about what produces phenomenal consciousness, we are far from knowing how to implement it in machines. This has an important consequence:

moral standing generally depends on phenomenal consciousness. We protect human life and the life of animals on this basis. If AI systems lack it, there is no reason to protect them (as is depicted in the film *Ex Machina*). This is certainly an important consequence of CAD and the arguments presented above that indicate how organisms (including humans) can interact with attentional processes while lacking phenomenal experience.

Another aspect of AI that is lacking concerns endogenous, spontaneous motivation. In living organisms, there are various motivating factors that shape the evolution of the organism, primarily factors surrounding the sustenance and propagating of life—a basic drive that underlies the evolution of species. From the drive to seek food, to basic fear responses, to social cooperation, living organisms (particularly more complex ones like humans) have adapted in ways that generally facilitate the continuation of a species. Indeed, an important aspect of human interactions is the level of empathic motivation that occurs with different degrees of intensity across social groups, which also cannot be simulated by machines (Winczewski, Bowen, & Collins, 2016). Feelings, being part of a core emotional neural system that must have evolved before higher forms of conceptual cognition, are not going to be reproduced by any kind of purely informational search algorithm or even a large collection of algorithms. At best, feelings can only be simulated, as seen in research that presents computational accounts of emotions.

Artificial intelligence has come a long way in the last few decades. Yet, it still is far from possessing human qualities, such as having general learning abilities or self-awareness, and more obviously, feelings and empathy. There are advances that point toward better learning capabilities, but these are still generally at a very simple level if one considers phenomenal consciousness rather than the access to information.

4. Conclusion

As Sherry Turkle argues, simulated thinking is (may be) thinking, but simulated feeling is not (can never be) feeling (Turkle, 2005/1984). Intelligence is essentially computation—something that machines were designed to do. Emotions that generate feelings, on the other hand, must involve phenomenal subjective experience. Machines may be able to self-reference and reason based on computational procedures, thus performing intelligently without conscious awareness. But to have empathy and social reasoning is beyond their ability.

To make better sense of this, one can examine the dissociation between consciousness and attention (CAD). According to the CAD framework, the type of consciousness that manifests in emotional arousal, experiences of moral approval or rejection, and experiences of the sublime and the beautiful are independent from the forms of intelligence and rationality associated with the attentional processing of features, objects, and events. Vivid experiences related to empathy have a cognitive foundation in the distinct role consciousness plays, which differs from attention to contents. There is also a technical or computational distinction: attention routines are programmable in machines in terms of functions that halt at a certain point, but consciousness does not seem susceptible of such treatment. It is also uncertain whether or not a large collection of routines could produce empathy and the required phenomenal consciousness accompanying it. These are two different senses of cognition and rule-guidance: the "should" of rationality and the "should" of empathy. For these reasons, it is difficult to argue for any AI gaining the ability to pass an emotional Turing test.

One question that arises is whether or not emotional responses in humans are all learned responses. This would suggest that simulated responses, which are programmed, are sufficient for an AI to pass the emotional Turing test at a very basic level. Support for this possibility could be seen if humans can learn to "turn off" emotions or emotional responses to stimuli, which seems like something one is capable of doing in extreme situations. Yet, this still does not address the *feeling* of emotion—there is an undeniable phenomenal experience that is present in our moral reasoning and this is based on primary emotions and independent from the associated behavioral responses, whether they are learned or not. It is unlikely that feelings, and any associated phenomenology, can be implemented in machines.

This brings us back to our main argument: simulated intelligence can be considered intelligence but simulated emotion cannot be experienced emotion. Intelligent behavior is computation, but feelings involve more than just computation. Because attention is largely independent of consciousness in humans, the forms of information processing systems related to attention are relatively straightforward to program and implement in machines. One important argument for this dissociation in humans is the earlier evolution of attention, before any conscious forms of attention. To achieve conscious experience in machines, and those that include emotional content, one must understand how attention and consciousness are dissociated in humans and what purpose conscious awareness may serve. For example, consciousness may be related to more sophisticated abilities such as multi-modal integration or empathy, which facilitates social cooperation in more "advanced" species like humans. These abilities seem to be related to intrinsic motivations to propagate the species and also include a social dimension—motivations that will not arise spontaneously in machines. Artificial intelligence may include simulated motivation in its programming, but it still will lack phenomenal content and self-awareness, which are at the core of empathy. This key point prevents AI from becoming moral agents or the conscious and malicious machines of science fiction.

Acknowledgements

We would like to thank an anonymous reviewer for suggestions that significantly improved this paper. HHH received postdoctoral research support from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement No. AG324070 awarded to Patrick Cavanagh.

References

Aaronson, S. (2016). Can computers become conscious? Retrieved from; http://www.scottaaronson.com/blog/.

Ahn, H.-I., & Picard, R. W. (2014b). Modeling subjective experience-based learning under uncertainty and frames. *Paper presented at the twenty-eighth AAAI conference on Artificial Intelligence, Québec City, Canada*. <http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8436>.

Ahn, H.-I., & Picard, R. W. (2014a). Measuring affective-cognitive experience and predicting market success. *IEEE Transactions on Affective Computing, 5*(2), 173–186. http://dx.doi.org/10.1109/TAFFC.2014.2330614.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392–398. http://dx.doi.org/10.1111/j.1467-9280.2008.02098.x.

Axelrod, R. M. (1984). *The evolution of cooperation*. New York: Basic Books.

Axelrod, V., & Rees, G. (2014). Conscious awareness is required for holistic face processing. *Consciousness and Cognition, 27*, 233–245. http://dx.doi.org/10.1016/j.concog.2014.05.004.

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research, 150*, 45–53. http://dx.doi.org/10.1016/S0079-6123(05)50004-9.

Bauer, H., Veira, J., & Weig, F. (2013). Moore's law: Repeal or renewal? Retrieved from McKinsey & Company. <http://www.mckinsey.com/insights/high_tech_telecoms_internet/moores_law_repeal_or_renewal>.

Bayne, T. (2007). Conscious states and conscious creatures: Explanation in the scientific study of consciousness. *Philosophical Perspectives, 21*(1), 1–22. http://dx.doi.org/10.1111/j.1520-8583.2007.00118.x.

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

Block, N. (1995a). The mind as the software of the brain. In D. Osherson, L. Gleitman, S. M. Kosslyn, S. Smith, & S. Sternberg (Eds.), *An invitation to cognitive science* (pp. 170–185). Cambridge, MA: MIT Press.

Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P., & Endy, D. (2013). Amplifying genetic logic gates. *Science, 340*, 599–603. http://dx.doi.org/10.1126/science.1232758.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (1st ed.). Oxford: Oxford University Press.

Bradshaw, T. (2016). Apple buys emotion-detecting AI start-up. *The financial times*. <http://www.ft.com/cms/s/0/2b915242-b571-11e5-8358-9a82b43f6b2f.html> (January 7).

Breazeal, C., & Scassellati, B. (1999). A context-dependent attention system for a social robot. *Paper presented at the proceedings of the sixteenth international joint conference on artificial intelligence*.

Breazeal, C., & Scassellati, B. (2002). Robots that imitate humans. *Trends in Cognitive Sciences, 6*(11), 481–487. http://dx.doi.org/10.1016/S1364-6613(02)02016-8.

Brooks, R., Gupta, A., McAfee, A., & Thompson, N. (2015) Artificial intelligence and the future of humans and robots in the economy. *The Malcolm and Carolyn Wiener annual lecture on science and technology, council on foreign relations*. <http://www.cfr.org/technology-and-science/artificial-intelligence-future-humans-robots-economy/p36197> (February 27, 2015).

Bruya, B. (2010). *Effortless attention: A new perspective in the cognitive science of attention and action*. Cambridge, MA: MIT Press.

Burke, E. (1757). *A philosophical enquiry into the origin of our ideas of the sublime and beautiful*. London: R. and J. Dodsley.

Carter, S., & McBride, M. (2013). Experienced utility versus decision utility: Putting the 'S' in satisfaction. *The Journal of Socio-Economics, 42*, 13–23. http://dx.doi.org/10.1016/j.socec.2012.11.009.

Cavanagh, P. (2004). Attention routines and the architecture of selection. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 13–28). New York: Guilford Press.

Celeghin, A., de Gelder, B., & Tamietto, M. (2015). From affective blindsight to emotional consciousness. *Consciousness and Cognition, 36*, 414–425. http://dx.doi.org/10.1016/j.concog.2015.05.007.

Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics, 74*(5), 784–802. http://dx.doi.org/10.3758/s13414-012-0322-z.

Churchland, P. S., & Churchland, P. M. (1990). Could a machine think? *Scientific American, 262*(1), 32–37.

Cisco Systems Inc. (2015). The Zettabyte Era—Trends and analysis Retrieved from; <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html>.

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology, 64*, 201–229. http://dx.doi.org/10.1146/annurev.psych.121208.131628.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life* (1st ed.). New York: Basic Books.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: G.P. Putnam.

Decety, J., & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in Cognitive Sciences, 18*(7), 337–339. http://dx.doi.org/10.1016/j.tics.2014.04.008.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204–211. http://dx.doi.org/10.1016/j.tics.2006.03.007.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*(1–2), 1–37. http://dx.doi.org/10.1016/S0010-0277(00)00123-2.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General, 129*(4), 481–507. http://dx.doi.org/10.1037/0096-3445.129.4.481.

Edelman, D. B., Baars, B. J., & Seth, A. K. (2005). Identifying hallmarks of consciousness in non-mammalian species. *Consciousness and Cognition, 14*(1), 169–187. http://dx.doi.org/10.1016/j.concog.2004.09.001.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200. http://dx.doi.org/10.1080/02699939208411068.

Eriksen, C. W., & Yeh, Y.-Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance, 11*(5), 583–597. http://dx.doi.org/10.1037/0096-1523.11.5.583.

Graziano, M. S. A. (2014). Speculations on the evolution of awareness. *Journal of Cognitive Neuroscience, 26*(6), 1300–1304. http://dx.doi.org/10.1162/jocn_a_00623.

Graziano, M. S. A. (2015). Build-a-brain: We could build an artificial brain that believes itself to be conscious. Does that mean we have solved the hard problem? Retrieved from; https://aeon.co/essays/can-we-make-consciousness-into-an-engineering-problem.

Graziano, M. S. A., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience, 2*(2), 98–113. http://dx.doi.org/10.1080/17588928.2011.565121.

Graziano, M. S. A., & Webb, T. W. (2014). A mechanistic theory of consciousness. *International Journal of Machine Consciousness, 6*(2), 1–14. http://dx.doi.org/10.1142/S1793843014001316.

Griffin, D. R., & Speck, G. B. (2004). New evidence of animal consciousness. *Animal Cognition, 7*(1), 5–18. http://dx.doi.org/10.1007/s10071-003-0203-x.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002. http://dx.doi.org/10.1126/science.1137651.

Haladjian, H. H., & Montemayor, C. (2015). On the evolution of conscious attention. *Psychonomic Bulletin & Review, 22*(3), 595–613. http://dx.doi.org/10.3758/s13423-014-0718-y.

Heckert, J. (2012). The hazards of growing up painlessly. *The New York times magazine*. <http://www.nytimes.com/2012/11/18/magazine/ashlyn-blocker-feels-no-pain.html> (November 15).

Heikkila, A. (2016). Self-driving cars and the Kobayashi Maru. *Techcrunch*. <http://techcrunch.com/2016/02/27/self-driving-cars-and-the-kobayashi-maru/>.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*(11), 494–500. http://dx.doi.org/10.1016/j.tics.2004.08.007.

Hommel, B. (2007). Feature integration across perception and action: Event files affect response choice. *Psychological Research Psychologische Forschung, 71*(1), 42–63. http://dx.doi.org/10.1007/s00426-005-0035-1.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016). Automatic sarcasm detection: A survey. *arXiv preprint arXiv:1602.03426*.

Kahneman, D., & Thaler, R. H. (2006). Anomalies: Utility maximization and experienced utility. *The Journal of Economic Perspectives, 20*(1), 221–234. http://dx.doi.org/10.1257/089533006776526076.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*(2), 175–219. http://dx.doi.org/10.1016/0010-0285(92)90007-O.

Kentridge, R. W. (2012). Blindsight: Spontaneous scanning of complex scenes. *Current Biology, 22*(15), R605–606. http://dx.doi.org/10.1016/j.cub.2012.06.011.

Kentridge, R. W. (2011). Attention without awareness: A brief review. In C. Mole, D. Smithies, & W. Wu (Eds.), *Attention: Philosophical and psychological essays* (pp. 228–246). Oxford: Oxford University Press.

Kentridge, R. W., Nijboer, T. C. W., & Heywood, C. A. (2008). Attended but unseen: Visual attention is not sufficient for visual awareness. *Neuropsychologia, 46*(3), 864–869. http://dx.doi.org/10.1016/j.neuropsychologia.2007.11.036.

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences, 11*(1), 16–22. http://dx.doi.org/10.1016/j.tics.2006.10.012.

Koch, C., & Tsuchiya, N. (2012). Attention and consciousness: Related yet different. *Trends in Cognitive Sciences, 16*(2), 103–105. http://dx.doi.org/10.1016/j.tics.2011.11.012.

Kriegel, U. (2015). *The varieties of consciousness*. New York: Oxford University Press.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. New York: Viking.

Lamme, V. A. F. (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Networks, 17*(5–6), 861–872. http://dx.doi.org/10.1016/j.neunet.2004.02.005.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155–184. http://dx.doi.org/10.1146/annurev.neuro.23.1.155.

LeDoux, J. E. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology, 23*(4–5), 727–738. http://dx.doi.org/10.1023/A:1025048802629.

LeDoux, J. E. (2012). Rethinking the emotional brain. *Neuron, 73*(4), 653–676. http://dx.doi.org/10.1016/j.neuron.2012.02.004.

Li, F.-F. (2014). The digital sensory system: A quest for visual intelligence in computers. *Paper presented at the stanford engineering's EngX: The digital sensory system* (May 20).

Lisi, M., & Cavanagh, P. (2015). Dissociation between the perceptual and saccadic localization of moving objects. *Current Biology, 25*(19), 2535–2540. http://dx.doi.org/10.1016/j.cub.2015.08.021.

Mancuso, K., Hauswirth, W. W., Li, Q., Connor, T. B., Kuchenbecker, J. A., Mauck, M. C., ... Neitz, M. (2009). Gene therapy for red–green colour blindness in adult primates. *Nature, 461*(7265), 784–787. http://dx.doi.org/10.1038/nature08401.

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences, 29*(6), 317–322. http://dx.doi.org/10.1016/j.tins.2006.04.001.

Meuwese, J. D. I., Post, R. A. G., Scholte, H. S., & Lamme, V. A. F. (2013). Does perceptual learning require consciousness or attention? *Journal of Cognitive Neuroscience, 25*(10), 1579–1596. http://dx.doi.org/10.1162/jocn_a_00424.

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine, 176*(5), 619–625. http://dx.doi.org/10.1001/jamainternmed.2016.0400.

Mitchell, D. G. V., & Greening, S. G. (2012). Conscious perception of emotional stimuli: Brain mechanisms. *The Neuroscientist, 18*(4), 386–398.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. http://dx.doi.org/10.1038/nature14236.

Moe-Behrens, G. H. G. (2013). The biological microprocessor, or how to build a computer with biological parts. *Computational and Structural Biotechnology Journal, 7*(8), 1–18. http://dx.doi.org/10.5936/csbj.201304003.

Montemayor, C., & Haladjian, H. H. (2015). *Consciousness, attention, and conscious attention*. Cambridge, MA: The MIT Press.

Moravec, H. P. (1988). *Mind children: The future of robot and human intelligence*. Cambridge, MA: Harvard University Press.

Mudrik, L., Faivre, N., & Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences, 18*(9), 488–496. http://dx.doi.org/10.1016/j.tics.2014.04.009.

Mulckhuyse, M., & Theeuwes, J. (2010). Unconscious attentional orienting to exogenous cues: A review of the literature. *Acta Psychologica, 134*(3), 299–309. http://dx.doi.org/10.1016/j.actpsy.2010.03.002.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences, 37*(01), 1–19. http://dx.doi.org/10.1017/S0140525X12003214.

Pauers, M. J., Kuchenbecker, J. A., Neitz, M., & Neitz, J. (2012). Changes in the colour of light cue circadian activity. *Animal Behaviour, 83*(5), 1143–1151. http://dx.doi.org/10.1016/j.anbehav.2012.01.035.

Pessoa, L. (2005). To what extent are emotional visual stimuli processed without attention and awareness? *Current Opinion in Neurobiology, 15*(2), 188–196. http://dx.doi.org/10.1016/j.conb.2005.03.002.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148–158. http://dx.doi.org/10.1038/nrn2317.

Pessoa, L. (2012). Beyond brain regions: Network perspective of cognition-emotion interactions. *Behavioral and Brain Sciences, 35*(3), 158–159. http://dx.doi.org/10.1017/S0140525x11001567.

Pessoa, L. (2013). *The cognitive-emotional brain: From interactions to integration*. Cambridge, MA: The MIT Press.

Pessoa, L., Kastner, S., & Ungerleider, L. G. (2002). Attentional control of the processing of neutral and emotional stimuli. *Cognitive Brain Research, 15*(1), 31–45. http://dx.doi.org/10.1016/S0926-6410(02)00214-8.

Pew Research Center (2015). The smartphone difference Retrieved from; http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/.

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W. (2002a). Affective medicine: Technology with emotional intelligence. *Studies in Health Technology and Informatics, 80*, 69–83. http://dx.doi.org/10.3233/978-1-60750-924-0-69.

Picard, R. W. (2007). Toward machines with emotional intelligence. In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *The science of emotional intelligence: Knowns and unknowns* (pp. 396–418). New York: Oxford University Press.

Picard, R. W. (2002b). What does it mean for a computer to "have" emotions? In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 213–236). Cambridge, MA: MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(10), 1175–1191. http://dx.doi.org/10.1109/34.954607.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology, 109*(2), 160–174. http://dx.doi.org/10.1037/0096-3445.109.2.160.

Pratt, G. A. (2015). Is a Cambrian explosion coming for robotics? *The Journal of Economic Perspectives, 29*(3), 51–60. http://dx.doi.org/10.1257/jep.29.3.51.

Pylyshyn, Z. W. (1980). The 'causal power' of machines. *Behavioral and Brain Sciences, 3*(3), 442–444. http://dx.doi.org/10.1017/S0140525X0000594X.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22*(3), 341–365 (discussion 366–423).

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences, 4*(5), 197–207. http://dx.doi.org/10.1016/S1364-6613(00)01477-7.

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.

Reichardt, D. M. (2007). A definition approach for an ''Emotional Turing Test''. In A. C. R. Paiva, R. Prada, & R. W. Picard (Eds.). *Affective computing and intelligent interaction* (Vol. 4738, pp. 716–717). Berlin, Heidelberg: Springer.

Richtel, M., & Dougherty, C. (2015). Google's driverless cars run into problem: Cars with drivers. *The New York times* (p. A1). <http://nyti.ms/1LRy9MF> (September 2).

Riedl, M. O., & Harrison, B. (2015). Using stories to teach human values to artificial agents. *Paper presented at the 2nd international workshop on AI, ethics, and society*. <http://www.aaai.org>.

Rinesi, M. (2015). The price of the Internet of Things will be a vague dread of a malicious world Retrieved from IEET.org website. <http://ieet.org/index.php/IEET/more/rinesi20150925>.

Robinson, H., MacDonald, B., & Broadbent, E. (2014). The role of healthcare robots for older people at home: A review. *International Journal of Social Robotics, 6*(4), 575–591. http://dx.doi.org/10.1007/s12369-014-0242-2.

Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition, 12*(4), 717–731. http://dx.doi.org/10.1016/S1053-8100(03)00081-3.

Rochat, P., & Striano, T. (2000). Perceived self in infancy. *Infant Behavior and Development, 23*(3–4), 513–530. http://dx.doi.org/10.1016/S0163-6383(01)00055-8.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1–2), 1–46. http://dx.doi.org/10.1016/S0010-0277(00)00152-9.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. http://dx.doi.org/10.1017/S0140525X00005756.

Searle, J. R. (1998). *Mind, language, and society: Philosophy in the real world* (1st ed.). New York, NY: Basic Books.

Seth, A. K., & Baars, B. J. (2005). Neural Darwinism and consciousness. *Consciousness and Cognition, 14*(1), 140–168. http://dx.doi.org/10.1016/j.concog.2004.08.008.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences, 12*(8), 314–321. http://dx.doi.org/10.1016/j.tics.2008.04.008.

Shibata, T. (2004). An overview of human interactive robots for psychological enrichment. *Proceedings of the IEEE, 92*(11), 1749–1758. http://dx.doi.org/10.1109/JPROC.2004.835383.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. http://dx.doi.org/10.1038/nature16961.

Spering, M., & Carrasco, M. (2015). Acting without seeing: Eye movements reveal visual processing without awareness. *Trends in Neurosciences, 38*(4), 247–258. http://dx.doi.org/10.1016/j.tins.2015.02.002.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 1–25.

Tamietto, M., & de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience, 11*(10), 697–709. http://dx.doi.org/10.1038/nrn2889.

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Sciences, 1124*, 239–261. http://dx.doi.org/10.1196/annals.1440.004.

Treisman, A. (1998). Feature binding, attention, and object perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 353*(1373), 1295–1306. http://dx.doi.org/10.1098/rstb.1998.0284.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*(4–8), 411–443. http://dx.doi.org/10.1080/13506280500195250.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. http://dx.doi.org/10.1016/0010-0285(80)90005-5.

Tsuchiya, N., & Adolphs, R. (2007). Emotion and consciousness. *Trends in Cognitive Sciences, 11*(4), 158–167. http://dx.doi.org/10.1016/j.tics.2007.01.005.

Tucker, A. M., Feuerstein, R., Mende-Siedlecki, P., Ochsner, K. N., & Stern, Y. (2012). Double dissociation: Circadian off-peak times increase emotional reactivity; aging impairs emotion regulation via reappraisal. *Emotion, 12*(5), 869–874. http://dx.doi.org/10.1037/a0028207.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 443–460.

Turkle, S. (2005/1984). *The second self: Computers and the human spirit* (20th anniversary ed.). Cambridge, MA: MIT Press.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies, 8*(3), 501–517. http://dx.doi.org/10.1075/is.8.3.11tur.

Turkle, S., Taggart, W., Kidd, C. D., & Dasté, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. *Connection Science, 18*(4), 347–361. http://dx.doi.org/10.1080/09540090600868912.

van Boxtel, J. J. A., Tsuchiya, N., & Koch, C. (2010). Consciousness and attention: On sufficiency and necessity. *Frontiers in Psychology, 1*(217). http://dx.doi.org/10.3389/fpsyg.2010.00217.

Wagner, A., & Rosen, W. (2014). Spaces of the possible: Universal Darwinism and the wall between technological and biological innovation. *Journal of The Royal Society Interface, 11*(97). http://dx.doi.org/10.1098/rsif.2013.1190.

Weintraubоct, K. (2012). But how do you really feel? Someday the computer may know. *The New York Times*. <http://www.nytimes.com/2012/10/16/science/affective-programming-grows-in-effort-to-read-faces.html> (October 15).

Weiskrantz, L. (2009). *Blindsight: A case study spanning 35 years and new developments* (2nd ed.). Oxford: Oxford University Press.

Winczewski, L. A., Bowen, J. D., & Collins, N. L. (2016). Is empathic accuracy enough to facilitate responsive behavior in dyadic interaction? Distinguishing ability from motivation. *Psychological Science, 27*(3), 394–404. http://dx.doi.org/10.1177/0956797615624491.

WNYC Studios (Producer). (2012). *Talking to machines*. <http://www.radiolab.org/story/137407-talking-to-machines/> (January).

Wolfram, S. (2016). *AI & the future of civilization/interviewer: E. Regis & N. G. Carr*. Edge Conversation: Technology, Edge Foundation Inc. (2016, March 1).

Wu, W. (2013). Mental action and the threat of automaticity. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the will* (pp. 244–261). Oxford: Oxford University Press.

Wu, W. (2011). Attention as selection for action. In C. Mole, D. Smithies, & W. Wu (Eds.), *Attention: Philosophical and psychological essays* (pp. 97–116). Oxford: Oxford University Press.

Yang, H., Shao, L., Zheng, F., Wang, L., & Song, Z. (2011). Recent advances and trends in visual tracking: A review. *Neurocomputing, 74*(18), 3823–3831. http://dx.doi.org/10.1016/j.neucom.2011.07.024.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology, 28*(1), 126–136. http://dx.doi.org/10.1037/0012-1649.28.1.126.

Zmigrod, S., Spapé, M., & Hommel, B. (2009). Intermodal event files: Integrating features across vision, audition, taction, and action. *Psychological Research Psychologische Forschung, 73*(5), 674–684. http://dx.doi.org/10.1007/s00426-008-0163-5.