On the Eve of Artificial Minds

Chris Eliasmith

I review recent technological, empirical, and theoretical developments related to building sophisticated cognitive machines. I suggest that rapid growth in robotics, brain-like computing, new theories of large-scale functional modeling, and financial resources directed at this goal means that there will soon be a significant increase in the abilities of artificial minds. I propose a specific timeline for this development over the next fifty years and argue for its plausibility. I highlight some barriers to the development of this kind of technology, and discuss the ethical and philosophical consequences of such a development. I conclude that researchers in this field, governments, and corporations must take care to be aware of, and willing to discuss, both the costs and benefits of pursuing the construction of artificial minds.

Keywords
Artificial cognition | Artificial intelligence | Brain modelling | Machine learning | Neuromorphic computing | Robotics | Singularity

Author

Chris Eliasmith
celiasmith@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Commentator

Daniela Hill
daniela.hill@gmx.net
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

Prediction is difficult, especially about the future
– Danish Proverb

1 Introduction

The prediction game is a dangerous one, but that, of course, is what makes it fun. The pitfalls are many: some technologies change exponentially but some don't; completely new inventions, or fundamental limits, might appear at any time; and it can be difficult to say something informative without simply stating the obvious. In short, it's easy to be wrong if you're specific. (Although, it is easy to be right if you're Nostradamus.) Regardless, the purpose of this essay is to play this game. As a consequence, I won't be pursuing technical discussion on the finer points of what a mind is, or how to build one, but rather attempting to paint an abstract portrait of the state of research in fields related to machine intelligence broadly construed. I think the risks of undertaking this kind of prognostication are justified because of the enormous potential impact of a new kind of technology that lies just around the corner. It is a technology we have been dreaming about—and dreading—for hundreds of years. I believe we are on the eve of artificial minds.

In 1958 Herbert Simon & Allen Newell claimed that "there are now in the world machines that think" and predicted that it would take ten years for a computer to become world chess champion and write beautiful music (1958, p. 8). Becoming world chess champion took longer, and we still don't have a digital Debussy. More importantly, even when a computer became world chess champion it was not generally seen as the success that Simon and

Newell had expected. This is because the way in which Deep Blue beat Gary Kasparov did not strike many as advancing our understanding of cognition. Instead, it showed that brute force computation, and a lot of careful tweaking by expert chess players, could surpass human performance in a specific, highly circumscribed environment.

Excitement about AI grew again in the 1980s, but was followed by funding cuts and general skepticism in the "AI winter" of the 1990s (Newquist 1994). Maybe we are just stuck in a thirty-year cycle of excitement followed by disappointment, and I am simply expressing the beginning of the next temporary uptick. However, I don't think this is the case. Instead, I believe that there are qualitative changes in methods, computational platforms, and financial resources that place us in a historically unique position to develop artificial minds. I will discuss each of these in more detail in subsequent sections, but here is a brief overview.

Statistical and brain-like modeling methods are far more mature than they have ever been before. Systems with millions (Garis et al. 2010; Eliasmith et al. 2012) and even tens of millions (Fox 2009) of simulated neurons are suddenly becoming common, and the scale of models is increasing at a rapid rate. In addition, the challenges of controlling a sophisticated, nonlinear body are being met by recent advances in robotics (Cheah et al. 2006; Schaal et al. 2007). These kinds of methodological advances represent a significant shift away from classical approaches to AI (which were largely responsible for the previously unfullfilled promises of AI) to more neurally inspired, and brain-like ones. I believe this change in focus will allow us to succeed where we haven't before. In short, the conceptual tools and technical methods being developed for studying what I call "biological cognition" (Eliasmith 2013), will make a fundamental difference to our likelihood of success.

Second, there have been closely allied and important advances in the kinds of computational platforms that can be exploited to run these models. So-called "neuromorphic" computing—hardware platforms that perform brain-style computation—has been rapidly scaling up, with several current projects expected to hit millions (Choudhary et al. 2012) and billions (Khan et al. 2008) of neurons running in real time within the next three to four years. These hardware advances are critical for performing efficient computation capable of realizing brain-like functions embedded in and controlling physical, robotic bodies.

Finally, unprecedented financial resources have been allocated by both public and private groups focusing on basic science and industrial applications. For instance, in February 2013 the European Union announced one billion euros in funding for the Human Brain Project, which focuses on developing a large scale brain model as well as neuromorphic and robotic platforms. A month later, the Obama BRAIN initiative was announced in the United States. This initiative devotes the same level of funding to experimental, technological, and theoretical advances in neuroscience. More recently, there has been a huge amount of private investment:

Google purchased eight robotics and AI companies between Dec 2013 and Jan 2014, including industry leader Boston Dynamics Stunt (2014).

Qualcomm has introduced the Zeroth processor, which is modeled after how a human brain works (Kumar 2013). They demonstrated an Field-Programmable Gate Array (FPGA) mock-up of the chip performing a reinforcement learning task on a robot.

Amazon has recently expressed a desire to provide the Amazon Prime Air service, which will use robotic quadcopters to deliver goods within thirty minutes of their having been ordered (Amazon 2013).

IBM has launched a product based on Watson, which famously beat the best human Jeopardy players (http://ibm.com/innovation/us/watson/). The product will provide confidence based responses to natural language queries. It has been opened up to allow developers to use it in a wide variety of applications. They are also developing a neuromorphic platform (Esser et al. 2013).

In addition, there are a growing number of startups that work on brain-inspired computing including Numenta, the Brain Corporation, Vi-

carious, DeepMind (recently purchased by Google for $400 million) and Applied Brain Research, among many others. In short, I believe there are more dollars being directed at the problem than ever before.

It is primarily these three forces that I believe will allow us to build *convincing* examples of artificial minds in the next fifty years. And, I believe we can do this without necessarily defining what it is that makes a "mind"—even an artificial one. As with many subtle concepts—such as "game," to use Wittgenstein's example, or "pornography," to use Supreme Court Justice Potter Stewart's example—I suspect we will avoid definitions and rely instead on our sophisticated, but poorly understood, methods of classifying the world around us. In the case of "minds," these methods will be partly behavioural, partly theoretical, and partly based on judgments of similarity to the familiar. In any case, I do not propose to provide a definition here, but rather to point to reasons why the artifacts we continue to build will become more and more like the natural minds around us. In doing so, I survey recent technological, theoretical, and empirical developments that are important for supporting our progress on this front. I then suggest a timeline over which I expect these developments to take place. Finally, I conclude with what I expect to be the major philosophical and societal impacts on our being able to build artificial minds. As a reminder, I am adopting a somewhat high-level perspective on the behavioural sciences and related technologies in order to make clear where my broad (and likely wrong) predictions are coming from. In addition, if I'm not entirely wrong, I suspect that the practical implications of such developments will prove salient to a broad audience, and so, as researchers in the area, we should consider the consequences of our research.

2 Technological developments

Because I take it that brain-based approaches provide the "difference that makes a difference" between current approaches and traditional AI, here I focus on developments in neuromorphic and robotic technology. Notably, all of the developments in neuromorphic hardware that I discuss below are inspired by some basic features of neural computation. For instance, all of the neuromorphic approaches use spiking neural networks SNNs to encode and process information. In addition, there is unanimous agreement that biological computation is in orders of magnitude more power efficient than digital computation (Hasler & Marr 2013). Consequently, a central motivation behind exploring these hardware technologies is that they might allow for sophisticated information processing using small amounts of power. This is critical for applications that require the processing to be near the data, such as in robotics and remote sensing. In what follows I begin by providing a sample of several major projects in neuromorphic computing that span the space of current work in the area. I then briefly discuss the current state of high-performance computing and robotics, to identify the roles of the most relevant technologies for developing artificial minds.

To complement its cognitively focused Watson project, IBM has been developing a neuromorphic architecture, a digital model of individual neurons, and a method for programming this architecture (Esser et al. 2013). The architecture itself is called TrueNorth. They argue that the "low-precision, synthetic, simultaneous, pattern-based metaphor of TrueNorth is a fitting complement to the high-precision, analytical, sequential, logic-based metaphor of today's of von Neumann computers" (Esser et al. 2013, p. 1). TrueNorth has neurons organized into 256 neuron blocks, in which each neuron can receive input from 256 axons. To assist with programming this hardware, IBM has introduced the notion of a "corelet," which is an abstraction that encapsulates local connectivity in small networks. These act like small programs that can be composed in order to build up more complex functions. To date the demonstrations of the approach have focused on simple, largely feed-forward, standard applications, though across a wide range of methods, including Restricted Boltzmann Machines (RBMs), liquid state machines, Hidden Markov Model (HMMs), and so on. It should be noted that the proposed chip does not yet exist, and

current demonstrations are on detailed simulations of the architecture. However, because it is a digital chip the simulations are highly accurate.

A direct competitor to IBM's approach is the Zeroth neuromorphic chip from Qualcomm. Like IBM, Qualcomm believes that constructing brain-inspired hardware will provide a new paradigm for exploiting the efficiencies of neural computation, targeted at the kind of information processing at which brains excel, but which is extremely challenging for von Neumann approaches. The main difference between these two approaches is that Qualcomm has committed to allowing online learning to take place on the hardware. Consequently, they announced their processor by demonstrating its application in a reinforcement learning paradigm on a real-world robot. They have released videos of the robot maneuvering in an environment and learning to only visit one kind of stimulus (white boxes: http://www.youtube.com/watch?v=8c1Noq2K96c). It should again be noted that this is an FPGA simulation of a digital chip that does not yet exist. However, the simulation, like IBM's, is highly accurate.

In the academic sphere, the Spinnaker project at Manchester University has not focused on designing new kinds of chips, but has instead focused on using low-power ARM processors on a massive scale to allow large-scale brain simulations (Khan et al. 2008). As a result, the focus has been on designing approaches to routing that allow for the high bandwidth communication, which underwrites much of the brain's information processing. Simulations on the Spinnaker hardware typically employ spiking neurons, like IBM and Qualcomm, and occasionally allow for learning (Davies et al. 2013), as with Qualcomm's approach. However, even with low power conventional chips, the energy usage is projected to be higher on the Spinnaker platform per neuron. Nevertheless, Spinnaker boards have been used in a wider variety of larger-scale embodied and non-embodied applications. These include simulating place cells, path integration, simple sensory-guided movements, and item classification.

There are also a number of neuromorphic projects that use analog instead of digital implementations of neurons. Analog approaches tend to be several orders of magnitude more power efficient (Hasler & Marr 2013), though also more noisy, unreliable, and subject to process variation (i.e., variations in the hardware due to variability in the size of components on the manufactured chip). These projects include work on the Neurogrid chip at Stanford University (Choudhary et al. 2012), and on a chip at ETH Zürich (Corradi, Eliasmith & Indiveri 2014). The Neurogrid chip has demonstrated larger numbers of simulated neurons—up to a million—while the ETH Zürich chip allows for online learning. More recently, the Neurogrid chip has been used to control a nonlinear, six degree of freedom robotic arm, exhibiting perhaps the most sophisticated information processing from an analog chip to date.

In addition to the above neuromorphic projects, which are focused on cortical simulation, there have been several specialized neuromorphic chips that mimic the information processing of different perceptual systems. For example, the dynamic vision sensor (DVS) artificial retina developed at ETH Zürich performs real-time vision processing that results in a stream of neuron-like spikes (Lichtsteiner et al. 2008). Similarly, an artificial cochlea called AEREAR2 has been developed that generates spikes in response to auditory signals (Li et al. 2012). The development of these and other neuronal sensors makes it possible to build fully embodied spiking neuromorphic systems (Galluppi et al. 2014).

There have also been developments in traditional computing platforms that are important for supporting the construction of models that run on neuromorphic hardware. Testing and debugging large-scale neural models is often much easier with traditional computational platforms such as Graphics Processing Unit (GPUs) and supercomputers. In addition, the development of neuromorphic hardware often relies on simulation of the designs before manufacture. For example, IBM has been testing their TrueNorth architecture with very large-scale simulations that have run up to 500 billion

neurons. These kinds of simulations allow for designs to be stress-tested and fine-tuned before costly production is undertaken. In short, the development of traditional hardware is also an important technological advance that supports the rapid development of more biologically-based approaches to constructing artificial cognitive systems.

A third area of rapid technological development that is critical for successfully realizing artificial minds is the field of robotics. The success of recent methods in robotics have entered public awareness with the creation of the Google car. This self-driving vehicle has successfully navigated hundreds of thousands of miles of urban and rural roadways. Many of the technologies in the car were developed out of DARPA's Grand Challenge to build an autonomous vehicle that would be tested in both urban and rural settings. Due to the success of the first three iterations of the Grand Challenge, DARPA is now funding a challenge to build robots that can be deployed in emergency situations, such as a nuclear meltdown or other disaster.

One of the most impressive humanoid robots to be built for this challenge is the Atlas, constructed by Boston Dynamics. It has twenty-eight degrees of freedom, covering two arms, two legs, a torso, and a head. The robot has been demonstrated walking bipedally, even in extremely challenging environments in which it must use its hands to help navigate and steady itself (http://www.youtube.com/watch?v=zkBnFPBV3f0). Several teams in this most recent Grand Challenge have been awarded a copy of Atlas, and have been proceeding to competitively design algorithms to improve its performance.

In fact, there have been a wide variety of significant advances in robotic control algorithms, enabling robots—including quadcopters, wheeled platforms, and humanoid robots—to perform tasks more accurately and more quickly than had previously been possible. This has resulted in one of the first human versus robot dexterity competitions being recently announced. Just as IBM pitted Watson against human Jeopardy champions, Kuka has pitted its high-speed arm against the human ping-pong champion Timo Boll (http://www.youtube.com/watch?v=_mbdtupCbc4). Despite the somewhat disappointing outcome, this kind of competition would not have been thought possible a mere five years ago (Ackerman 2014).

These three areas of technological development—neuromorphics, high-performance conventional computing, and robotics—are progressing at an incredibly rapid pace. And, more importantly, their convergence will allow a new class of artificial agents to be built. That is, agents that can begin processing information at very similar speeds and support very similar skills to those we observe in the animal kingdom. It is perhaps important to emphasize that my purpose here is predictive. I am not claiming that current technologies are sufficient for building a new kind of artificial mind, but rather that they lay the foundations, and are progressing at a sufficient rate to make it reasonable to expect that the sophistication, adaptability, flexibility, and robustness of artificial minds will rapidly approach those of the human mind. We might again worry that it will be difficult to measure such progress, but I would suggest that progress will be made along many dimensions simultaneously, so picking nearly any of dimensions will result in some measurable improvement. In general, multi-dimensional similarity judgements are likely to result in "I'll know it when I see it" kinds of reactions to classifying complicated examples. This may be derided by some as "hand-wavy", but it might also be a simple acknowledgement that "mindedness" is complex. I would like to be clear that my claims about approaching human mindful behaviour are to be taken as applying to the vast majority of the many measures we use for identifying minds.

3 Theoretical developments

Along with these technological developments there have been a series of theoretical developments that are critical for building large-scale artificial agents. Some have argued that theoretical developments are not that important: suggesting that standard back propagation at a sufficiently large scale is enough to capture

complex perceptual processing (Krizhevsky et al. 2012). That is, building brain-like models is more a matter of getting a sufficiently large computer with enough parameters and neurons than it is of discovering some new principles about how brains function. If this is true, then the technological developments that I pointed to in the previous section may be sufficient for scaling to sophisticated cognitive agents. However, I am not convinced that this is the case.

As a result, I think that theoretical developments in deep learning, nonlinear adaptive control, high dimensional brain-like computing, and biological cognition *combined* will be important to support continued advances in understanding how the mind works. For instance, deep networks continue to achieve state-of-the-art results in a wide variety of perception-like processing challenges (http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html#43494641522d3130). And while deep networks have traditionally been used for static processing, such as an image classification or document classification, there has been a recent, concerted move to use them to model more dynamic perceptual tasks as well (Graves et al. 2013). In essence, deep networks are one among many techniques for modeling the statistics of time varying signals, a skill central to animal cognition.

However, animals are also incredibly adept at *controlling* nonlinear dynamical systems, including their bodies. That is, biological brains can *generate* time varying signals that allow successful and sophisticated interactions with their environment through their body. Critically, there have been a variety of important theoretical advances in nonlinear and adaptive control theory as well. New methods for solving difficult optimal control problems have been discovered through careful study of biological motor control (Schaal et al. 2007; Todorov 2008). In addition, advances in hierarchical control allow for real-time computation of difficult inverse kinematics problems on a laptop (Khatib 1987). And, finally, important advances in adaptive control allow for the automatic learning of both kinematic and dynamic models even in highly nonlinear and high dimensional control spaces (Cheah et al. 2006).

Concurrently with these more abstract characterizations of brain function there have been theoretical developments in neuroscience that have deepened our understanding of how biological neural networks may perform sophisticated information processing. Work using the Neural Engineering Framework (NEF) has resulted in a wide variety of spiking neural models that mirror data recorded from biological systems (Eliasmith & Anderson 1999, 2003). In addition, the closely related liquid computing (Maass et al. 2002) and FORCE learning (Sussillo & Abbott 2009) paradigms have been successfully exploited by a number of researchers to generate interesting dynamical systems that often closely mirror biological data. Together these kinds of methods provide quantitative characterizations of the computational power available in biologically plausible neural networks. Such developments are crucial for exploiting neuromorphic approaches to building brain-like hardware. And they suggest ways of testing some of the more abstract perceptual and control ideas in real-world, brain-like implementations.

Interestingly, several authors have suggested that difficult perceptual and control problems are in fact mathematical duals of one another (Todorov 2009; Eliasmith 2013). This means that there are deep theoretical connections between perception and motor control. This realization points to a need to think hard about how diverse aspects of brain function can be integrated into single, large-scale models. This has been a major focus of research in my lab recently. One result of this focus is Spaun, currently the world's largest functional brain model. This model incorporates deep networks, recent control methods, and the NEF to perform eight different perceptual, motor, and cognitive tasks (Eliasmith et al. 2012). Importantly, this is not a one-off model, but rather a single example among many that employs a general architecture intended to directly address integrated biological cognition (Eliasmith 2013). Currently, the most challenging constraints for running models like Spaun are technological—computers are not fast enough. However, the

neuromorphic technologies mentioned previously should soon remove these constraints. So, in some sense, theory currently outstrips application: we have individually tested several critical assumptions of the model and shown that they scale well (Crawford et al. 2013), but we are not yet able to integrate full-scale versions of the components due to limitations in current computational resources.

Taken together, I believe that these recent theoretical developments demonstrate that we have a roadmap for how to approach the problem of building sophisticated models of biological cognition. No doubt not all of the methods we need are currently available, but it is not evident that there are any major conceptual roadblocks to building a cognitive system that rivals the flexibility, adaptability, and robustness of those found in nature. I believe this is a unique historical position. In the heyday of the symbolic approach to AI there were detractors who said that the perceptual problems solved easily by biological systems would be a challenge for the symbolic approach (Norman 1986; Rumelhart 1989). They were correct. In the heyday of connectionism there were detractors who said that standard approaches to artificial neural networks would not be able to solve difficult planning or syntactic processing problems (Pinker & Prince 1988; Fodor & Pylyshyn 1988; Jackendoff 2002). They were correct. In the heyday of statistical machine learning approaches (a heyday we are still in) there are detractors who say that mountains of data are not sufficient for solving the kinds of problems faced by biological cognitive systems (Marcus 2013). They are probably correct. However, as many of the insights of these various approaches are combined with control theory, integrated into models able to do efficient syntactic and semantic processing with neural networks, and, in general, become conceptually *unified* (Eliasmith 2013), it is less and less obvious what might be missing from our characterization of biological cognition.

4 Empirical developments

One thing that might be missing is, simply, knowledge. We have many questions about how real biological systems work that remain unanswered. Of course, complete knowledge of natural systems is not a prerequisite for building nearly functionally equivalent systems (see e.g., flight). However, I believe our understanding of natural cognitive systems will continue to play an important role in deciding what kinds of algorithms are worth pursuing as we build more sophisticated artificial agents.

Fortunately, on this front there have been two announcements of significant resources dedicated to improving our knowledge of the brain, which I mentioned in the introduction. One is from the EU and the other from the US. Each are investing over $1 billion in generating the kind of data needed to fill gaps in our understanding of how brains function. The EU's Human Brain Project (HBP) includes two central subprojects aimed at gathering mouse and human brain data to complement the large-scale models being built within the project. These subprojects will focus on genetic, cellular, vascular, and overall organizational data to complement the large-scale projects of this type already available (such as the Allen Brain Atlas, http://www.brain-map.org/). One central goal of these subprojects is to clarify the relationship between the mouse (which is highly experimentally accessible) and human subjects.

The American "brain research through advancing innovative neurotechnologies" (BRAIN) initiative is even more directly focused on large-scale gathering of neural data. Its purpose is to accelerate technologies to provide large-scale dynamic information about the brain that demonstrates how both single runs and larger neural circuits operate. Its explicit goal is to "fill major gaps in our current knowledge" (http://www.nih.gov/science/brain/). It is a natural complement to the human connectome project, which has been mapping the structure of the human brain on a large-scale (http://www.humanconnectomeproject.org/). Even though it is not yet clear exactly what information will be provided by the BRAIN intiative, it is clear that significant resources are being put into developing technologies that draw on nanoscience, informatics, engineering, and other fields to measure the brain at a level of detail and scale not previously possible.

Table 1

Within (years)	Von Neumann (real-time neurons)	Neuromorphic (real-time neurons)	Behaviours
5	10^8	10^7	constrained environment navigation; uncluttered vision/audition recognition; simple language understanding; slow, robust motor control
10	10^9	10^8	good open world large-scale navigation; learn simple few-step cognitive tasks; rapid single limb motor control; general, shallow semantics and syntax
15	10^{10}	10^9	literal natural language (4-5 year old equivalent); learn new many step tasks; near human quality perceptual skills (categorization and recognition in arbitrary, moving scenes)
25	10^{11}	10^{11}	arbitrary autonomous navigation; highly tunable cognitive system for task specialization, generally exceeding human ability on gross manual tasks, basic problem solving, etc.; human quality perception
50	10^{13}	10^{14}	human level full body agility; above average IQ; nuanced natural language; better than human perception

While both of these projects are just over a year old, they have both garnered international attention and been rewarded with sufficient funding to ensure a good measure of success. Consequently, it is likely that as we build more sophisticated models of brain function, and as we discover where our greatest areas of ignorance lay, we will be able to turn to the methods developed by these projects to rapidly gain critical information and continue improving our models. In short, I believe that there is a confluence of technological, theoretical, and empirical developments that will allow for bootstrapping detailed functional models of the brain. It is precisely these kinds of models that I expect will lead to the most convincing embodiments of artificial cognition that we have ever seen—I am even willing to suggest that their sophistication will rival those of natural cognitive systems.

5 A future timeline

Until this point I have been mustering evidence that there will soon be significant improvements in our ability to construct artificial cognitive agents. However, I have not been very specific about timing. The purpose of this section is to provide more quantification on the speed of development in the field.

In Table 1, the first column specifies the timeframe, the second suggests the number of neurons that will be simulatable in real-time on standard hardware, the third suggests the number of neurons that will be simulatable in real-time on neuromorphic hardware, and the last identifies relevant achievable behaviours within that timeframe.

I believe that several of the computational technologies I have mentioned, as well as empir-

ical methods for gathering evidence, are on an exponential trajectory by relevant measures (e.g., number of neurons per chip, number of neurons recorded Stevenson & Kording 2011). On the technological side, if we assume a doubling every eighteen months, this is equivalent to an increase of about one order of magnitude every five years. I should also note that I am assuming that real-time simulation of neurons will be embedded in an interactive, real-world environment, and that the neuron count is for the whole system (not a single chip). For context, it is worth remembering that the human brain has about 10^{11} neurons, though they are more computationally sophisticated than those typically simulated in hardware.

Another caveat is that it is likely that large-scale simulations on a digital Von Neumann architecture will hit a power barrier, which makes it likely that the suggested scaling could be achieved, but will be cost-prohibitive in fifty years. Consequently, a neuromorphic alternative is most likely to be the standard implementational substrate of artificial agents.

Finally, the behavioural characterizations I am giving are with a view to functions necessary for creating a convincing artificial mind in an artificial body. Consequently, my comments generally address perceptual, motor, and cognitive skills relevant to reproducing human-like abilities.

6 Consequences for philosophy

So suppose that, fifty years hence, we have developed an understanding of cognitive systems that allows us to build artificial systems that are on par with, or, if we see fit, surpass the abilities of an average person. Suppose, that is, that we can build artificial agents that can move, react, adapt, and think much like human beings. What consequences, if any, would this have for our theoretical questions about cognition? I take these questions to largely be in the domain of philosophy of mind. In this section I consider several central issues in philosophy of mind and discuss what sorts of consequences I take building a human-like artificial agent to have for them.

Being a philosopher, I am certain that, for any contemporary problem we consider, at least some subset of those who have a committed opinion about that problem will not admit that any amount of technical advance can "solve" it. I suspect, however, that their opinions may end up carrying about as much weight as a modern-day vitalist. To take one easy example, let us think for a moment about contemporary dualism. Some contemporary dualists hold that even if we had a complete understanding of how the brain functions, we would be no closer to solving the "hard problem" of consciousness (Chalmers 1996). The "hard problem" is the problem of explaining how subjective experience comes from neural activity. That is, how the phenomenal experiences we know from a first-person perspective can be accounted for by third-person physicalist approaches to understanding the mind. If indeed we have constructed artificial agents that behave much like people, share a wide variety of internal states with people, are fully empirically accessible, and report experiences like people, it is not obvious to what extent this problem will not have been solved. Philosophers who are committed to the notion that no amount of empirical knowledge will solve the problem will of course dismiss such an accomplishment on the strength of their intuitions. I suspect, however, that when most people are actually confronted with such an agent—one they can interrogate to their heart's content and one about which they can have complete knowledge of its functioning—it will seem odd indeed to suppose that we cannot explain how its subjective experience is generated. I suspect it will seem as odd as someone nowadays claiming that we cannot expect to explain how life is generated despite our current understanding of biochemistry. Another way to put this is that the "strong intuitions" of contemporary dualists will hold little plausibility in the face of actually existing, convincing artificial agents, and so, I suspect, they will become even more of a rarity.

I refer to this example as "easy" because the central reasons for rejecting dualism are only strengthened, not *generated*, by the existence of sophisticated artificial minds. That is,

good arguments against the dualist view are more or less independent of the current state of constructing agents (although the existence of such agents will likely sway intuitions). However, other philosophical conundrums, like Searle's famous Chinese room (1980), have responses that depend fairly explicitly on our ability to construct artificial agents. In particular, the "systems reply" suggests that a sufficiently complex system will have the same intentional states as a biological cognitive system. For those who think that this is a good rejection of Searle's strong intentionalist views, having systems that meet all the requirements of their currently hypothetical agents would provide strong empirical evidence consistent with their position. Of course, the existence of such artificial agents is unlikely to convince those, like Searle, who believe that there is some fundamental property of biology that allows intentionality to gain a foothold. But it does make such a position seem that much more tenuous if every means of measuring intentionality produces similar measurements across non-biological and biological agents. In any case, the realization of the systems reply does ultimately depend on our ability to construct sufficiently sophisticated artificial agents. And I am suggesting that such agents are likely to be available in the next fifty years.

More immediately, I suspect we will be able to make significant headway on several problems that have been traditionally considered philosophical *before* we reach the fifty-year mark. For example, the frame problem—i.e., the problem of knowing what representational states to update in a dynamic environment—is one that contemporary methods, like control theory and machine learning, struggle with much less than classical methods. Because the dynamics of the environment are explicitly included in the world-model being exploited by such control theoretic and statistical methods, updating state representations naturally includes the kinds of expectations that caused such problems for symbolic approaches.

Similarly, explicit quantitative solutions are suggested for the symbol-grounding problem through integrated models that incorporate aspects of both statistical perceptual processing and syntactic manipulation. Even in simple models, like Spaun, it is clear how the symbols for digits that are syntactically manipulated are related to inputs coming from the external world (Eliasmith 2013). And it is clear how those same symbols can play a role in driving the model's body to express its knowledge about those representations. As a result, the tasks that Spaun can undertake demonstrate both conceptual knowledge, through the symbol-like relationships between numbers (e.g., in the counting task), and perceptual knowledge, through categorization and the ability to drive its motor system to reproduce visual properties (e.g., in the copy-drawing task).

In some cases, rather than resolving philosophical debates, the advent of sophisticated artificial agents is likely to make these debates far more empirically grounded. These include debates about the nature of concepts, conceptual change, and functionalism, among others. However these debates turn out, it seems clear that having an engineered, working system that can generate behaviour as sophisticated as that that gave rise to these theoretical ideas in the first place will allow a systematic investigation of their appropriate application. After all, there are few, if any, limits on the empirical information we can garner from such constructed systems. In addition, our having built the system explicitly makes it unlikely that we would be unaware of some "critical element" essential in generating the observed behaviours.

Even without such a working system, I believe that there are already hints as to how these debates are likely to be resolved, given the theoretical approaches I highlighted earlier. For instance, I suspect that we will find that concepts are explained by a combination of vector space representations and a restricted class of dynamic processes defined over those spaces (Eliasmith 2013). Similarly, quantifying the adaptive nature of those representations and processes will indicate the nature of mechanisms of conceptual change in individuals (Thagard 2014). In addition, functionalism will probably seem too crude a hypothesis given a detailed understanding of how to build a wide variety of

artificial minds. Perhaps a kind of "functionalism with error bars" will take its place, providing a useful means of talking about degrees of functional similarity and allowing a quantification of functional characterizations of complex systems. Consequently, suggestions about which functions are or are not necessary for "mindedness" can be empirically tested through explicit implementation and experimentation. This will not solve the problem of mapping experimental results to conceptual claims (a problem we currently face when considering non-human and even some human subjects), but it will make functionalism as empirically accessible as seems plausible.

In addition to these philosophical issues that may undergo reconceptualization with the construction of artificial minds, there are others that are bound to become more vexing. For example, the breadth of application of ethical theory may, for the first time, reach to engineered devices. If, after all, we have built artificial minds capable of understanding their place in the universe, it seems likely we will have to worry about the possibility of their suffering (Metzinger 2013). It does not seem that understanding how such devices work, or having explicitly built them, will be sufficient for dismissing them as having no moral status. While current theories of non-human ethics have been developed, it is not clear how much or little theories of non-biological ethics will be able to borrow from them.

I suspect that the complexities introduced to ethical theory will go beyond adding a new category of potential application. Because artificial minds will be designed, they may be designed to make what have traditionally been morally objectionable inter-mind relationships seem less problematic. Consider, for instance, a robot that is designed to gain maximal self-fulfillment out of providing service to people. That is, unlike any biological species of which we are aware, these robots place service to humans above all else. Is a slave-like relationship between humans and these minds still wrong in such an instance? Whatever our analysis of why slavery is wrong, it seems likely that we will be able to design artificial minds that bypass that analysis. This is a unique quandary because while it is currently possible for certain individuals to claim to have such slave-aligned goals, it is always possible to argue that they are simply mistaken in their personal psychological analysis. In the case of minds whose psychology is designed in a known manner, however, the having of such goals will at least seem much more genuine. This is only one among many new kinds of challenges that ethical theory will face with the development of sophisticated artificial agents (Metzinger 2013).

I do not take this surely unreasonably brief discussion of any of these subtle philosophical issues to do justice to them. My main purpose here is to provide a few example instances of how the technological developments discussed earlier are likely to affect our theoretical inquiry. On some occasions such developments will lead to strengthening already common intuitions; on others they may provide deep empirical access to closely related issues; and on still other occasions these developments will serve to make complex issues even more so.

7 The good and the bad

As with the development of many technologies —cars, electricity, nuclear power—the construction of artificial minds is likely to have both negative and positive impacts. However, there is a sense in which building *minds* is much more fraught than these other technologies. We may, after all, build agents that are themselves capable of immorality. Presumably we would much prefer to build Commander Data than to build HAL or the Terminator. But how to do this is by no means obvious. There have been several interesting suggestions as to how this might be accomplished, perhaps most notably from Isaac Asimov in his entertaining and thought-provoking exploration of the three laws of robotics. For my purposes, however, I will sidestep this issue —not because it is not important, but because more immediate concerns arise from considering the development of these agents from a technological perspective. Let me then focus on the more immediately pressing consequences of constructing intelligent machines.

The rapid development of technologies related to artificial intelligence has not escaped the notice of governments around the world. One of the primary concerns for governments is the potentially massive changes in the nature of the economy that may result from an increase in automatization. It has recently been suggested that almost half of the jobs in the United States are likely to be computerized in the next twenty years (Rutkin 2013). The US Bureau of Labor and Statistics regularly publishes articles on the significant consequence of automation for the labour force in their journal Monthly Labor Review (Goodman 1996; Plewes 1990). This work suggests that greater automatization of jobs may cause standard measures of productivity and output to increase, while still increasing unemployment.

Similar interest in the economic and social impacts of automatization is evident in many other countries. For instance, Policy Horizons Canada is a think-tank that works for the Canadian government, which has published work on the effects of increasing automatization and the future of the economy (Arshad 2012). Soon after the publication of our recent work on Spaun, I was contacted by this group to discuss the impact of Spaun and related technologies. It was clear from our discussion that machine learning, automated control, robotics, and so on are of great interest to those who have to plan for the future, namely our governments and policy makers (Padbury et al. 2014).

This is not surprising. A recent McKinsey report suggests that these highly disruptive technologies are likely to have an economic value of about $18 trillion by 2025 (Manyika et al. 2013). It is also clear from the majority of analyses, that lower-paid jobs will be the first affected, and that the benefits will accrue to those who can afford what will initially be expensive technologies. Every expectation, then, is that automatization will exacerbate the already large and growing divide between rich and poor (Malone 2014; "The Future of Jobs: The Onrushing Wave" 2014). Being armed with this knowledge now means that individuals, governments, and corporations can support progessive policies to mitigate these kinds of potentially problematic societal shifts (Padbury et al. 2014).

Indeed, many of the benefits of automatization may help alleviate the potential downsides. Automatization has already had significant impact on the growth of new technology, both speeding up the process of development and making new technology cheaper. The human genome project was a success largely because of the automatization of the sequencing process. Similarly, many aspects of drug discovery can be automatized by using advanced computational techniques (Leung et al. 2013). Automatization of more intelligent behaviour than simply generating and sifting through data is likely to have an even greater impact on the advancement of science and engineering. This may lead more quickly to cleaner and cheaper energy, advances in manufacturing, decreases in the cost and access to advanced technologies, and other societal benefits.

As a consequence, manufacturing is likely to become safer—a trend already seen in areas of manufacturing that employ large numbers of robots (Robertson et al. 2005). At the same time, additional safety considerations come into play as robotic and human workspaces themselves begin to interact. This concern has resulted in a significant focus in robotics on compliant robots. Compliant robots are those that have "soft" environmental interactions, often implemented by including real or virtual springs on the robotic platform. As a result, control becomes more difficult, but interactions become much safer, since the robotic system does not rigidly go to a target position even if there is an unexpected obstacle (e.g., a person) in the way.

As the workplace continues to become one where human and automated systems cooperate, additional concerns may arise as to what kinds of human-machine relationships employers should be permitted to demand. Will employees have the right not to work with certain kinds of technology? Will employers still have to provide jobs to employees who refuse certain work situations? These questions touch on many of the same subjects highlighted in the previous section regarding

the ethical challenges that will be raised as we develop more and more sophisticated artificial minds.

Finally, much has been made of the possibility that the automatization of technological advancement could eventually result in machines designing themselves more effectively than humans can. This idea has captured the public imagination, and the point in time where this occurs is now broadly known as "The Singularity," a term first introduced by von Neumann (Ulam 1958). Given the vast variety of functions that machines are built to perform, it seems highly unlikely that there will be anything analogous to a mathematical singularity—a clearly defined, discontinuous point—after which machines will be superior to humans. As with most things, such a shift, if it occurs, is likely to be gradual. Indeed, the earlier timeline is one suggestion for how such a gradual shift might occur. Machines are already used in many aspects of design, for performing optimizations that would not be possible without them. Machines are also already much better at many functions than people: most obviously mechanical functions, but more recently cognitive ones, like playing chess and answering trivia questions in certain circumstances.

Because the advancement of intelligent machines is likely to continue to be a smooth, continuous one (even if exponential at times), we will likely remain in a position to make informed decisions about what they are permitted to do. As with members of a strictly human society, we do not tolerate arbitrary behaviour simply because such behaviour is possible. If anything, we will be in a *better* position to specify appropriate behaviour in machines than we are in the case of our human peers. Perhaps we will need laws and other societal controls for determining forbidden or tolerable behaviour. Perhaps some people and machines will choose to ignore those laws. But, as a society, it is likely that we will enforce these behavioural constraints the same way we do now—with publically sanctioned agencies that act on behalf of society. In short, the dystopian predictions we often see that revolve around the development of intelligent robots seem no more or less likely because of the robots. Challenges to societal stability are nothing new: war, hunger, poverty, weather are constant destabilizing forces. Artificial minds are likely to introduce another force, but one that may be just as likely to be stabilizing as problematic.

Unsurprisingly, like many other technological changes, the development of artificial minds will bring with it both costs and benefits. It may even be the case that deciding what is a cost and what is a benefit is not straightforward. If indeed many jobs become automated, it would be unsurprising if the average working week becomes shorter. As a result, a large number of people may have much more recreational time than has been typical in recent history. This may seem like a clear benefit, as many of us look forward to holidays and time off work. However, it has been argued that fulfilling work is a central to human happiness (Thagard 2010). Consequently, overly limited or unchallenging work may end up being a significant cost of automation.

As good evidence for costs and benefits becomes available, decision-makers will be faced with the challenge of determining what the appropriate roles of artificial minds should be. These roles will no doubt evolve as technologies change, but there is little reason to presume that unmanageable upheavals or "inflection points" will be the result of artificial minds being developed. While we, as a society, must be aware of, and prepared for, being faced with new kinds of ethical dilemmas, this has been a regular occurrence during the technological developments of the last several hundred years. Perhaps the greatest challenges will arise because of the significant wealth imbalances that may be exacerbated by limited access to more intelligent machines.

8 Conclusion

I have argued that we are at a unique point in the development of technologies that are critical to the realization of artificial minds. I have even gone so far as to predict that human-level intelligence and physical ability will be achieved in about fifty years. I suspect that for many famil-

iar with the history of artificial intelligence such predictions will be easily dismissed. Did we not have such predictions over fifty years ago? Some have suggested that the singularity will occur by 2030 (Vinge 1993), others by 2045 (Kurzweil 2005). There were suggestions and significant financial speculation that AI would change the world economy in the 1990s, but this never happened. Why would we expect anything to be different this time around?

In short, my answer is encapsulated by the specific technological, theoretical, and empirical developments I have described above. I believe that they address the central limitations of previous approaches to artificial cognition, and are significantly more mature than is generally appreciated. In addition, the limitations they address—such as power consumption, computational scaling, control of nonlinear dynamics, and integrating large-scale neural systems—have been more central to prior failures than many have realized. Furthermore, the financial resources being directed towards the challenge of building artificial minds is unprecedented. High-tech companies, including Google, IBM, and Qualcomm have invested billions of dollars in machine intelligence. In addition, funding agencies including DARPA (Defense Advanced Research Projects Agency), EU-IST (European Union—Information Society Technologies), IARPA (Intelligence Advanced Research Projects Agency), ONR (Office of Naval Research), and AFOSR (Air Force Office of Scientific Research) have contributed a similar or greater amount of financial support across a wide range of projects focused on brain-inspired computing. And the two special billion dollar initiatives from the US and EU will serve to further deepen our understanding of biological cognition, which has, and will continue, to inspire builders of artificial minds.

While I believe that the alignment of these forces will serve to underpin unprecedented advances in our understanding of biological cognition, there are several challenges to achieving the timeline I suggest above. For one, robotic actuators are still far behind the efficiency and speeds found in nature. There will no doubt be advances in materials science that will help overcome these limitations, but how long that will take is not yet clear. Similarly, sensors on the scale and precision of those available from nature are not yet available. This is less true for vision and audition, but definitely the case for proprioception and touch. The latter are essential for fluid, rapid motion control. It also remains to be seen how well our theoretical methods for integrating complex systems will scale. This will only become clear as we attempt to construct more and more sophisticated systems. This is perhaps the most fragile aspect of my prediction: expecting to solve difficult algorithmic and integration problems. And, of course, there are myriad other possible ways in which I may have underestimated the complexity of biological cognition: maybe glial cells are performing critical computations; maybe we need to describe genetic transcription processes in detail to capture learning; maybe we need to delve to the quantum level to get the explanations we need—but I am doubtful (Litt et al. 2006).

Perhaps it goes without saying that, all things considered, I believe the timeline I propose is a plausible one.[1] This, of course, is predicated on there being the societal and political will to allow the development of artificial minds to proceed. No doubt researchers in this field need to be responsive to public concerns about the specific uses to which such technology might be put. It will be important to remain open, self-critical, and self-regulating as artificial minds become more and more capable. We must usher in these technologies with care, fully cogniscent of, and willing to discuss, both their costs and their benefits.

Acknowledgements

I wish to express special thanks to two anonymous reviewers for their helpful feedback. Many of the ideas given here were developed in discussion with members of the CNRG Lab, participants at the Telluride workshops, and my collaborators on ONR grant N000141310419 (PIs: Kwabena Boahen and Rajit Manohar). This work was also funded by AFOSR grant FA8655-13-1-3084, Canada Research Chairs, and NSERC Discovery grant 261453.

1 It is quite different from that proposed by the HBP, for example. For further discussion of the differences in perspective between the HBP and my lab's work, see Eliasmith & Trujillo (2013).

References

Ackerman, E. (2014). Robots playing ping pong: What's real, what's not? *IEEE Spectrum*

Amazon (2013). Amazon Prime Air.

Arshad, I. (2012). People and machines: Competitors or collaborators in the emerging world of work?

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory.* Oxford, UK: Oxford University Press.

Cheah, C. C., Liu, C. & Slotine, J. J. E. (2006). Adaptive tracking control for robots with unknown kinematic and dynamic properties. *The International Journal of Robotics Research*, *25* (3), 283-296. SAGE Publications.

Choudhary, S., Sloan, S., Fok, S., Neckar, A., Trautmann, E., Gao, P., Stewart, T., Eliasmith, C. & Boahen, K. (2012). Silicon neurons that compute. In A. E. P. Villa, W. Duch, P. Érdi, F. Masulli and G. Palm (Eds.) *Artificial neural networks and machine learning - ICANN 2012* (pp. 121-128). Berlin, GER: Springer. 10.1007/978-3-642-33269-2_16

Corradi, F., Eliasmith, C. & Indiveri, G. (2014). Mapping arbitrary mathematical functions and dynamical systems to neuromorphic VLSI circuits for spike-based neural computation. *IEEE International Symposium on Circuits and Systems (ISCAS) 2014* (pp. 269-272). IEEE. 10.1109/ISCAS.2014.6865117

Crawford, E., Gingerich, M. & Eliasmith, C. (2013). Biologically plausible, human-scale knowledge representation. In M. Knauff, M. Pauen, N. Sebanz and I. Wachsmuth (Eds.) *35th Annual Conference of the Cognitive Science Society* (pp. 412-417). Austin, TX: Cognitive Science Society.

Davies, S., Stewart, T., Eliasmith, C. & Furber, S. (2013). Spike-based learning of transfer functions with the SpiNNaker neuromimetic simulator. *The 2013 International Joint Conference on Neural Networks (IJCNN)* (pp. 1832-1839). IEEE. 10.1109/IJCNN.2013.6706962

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition.* New York, NY: Oxford University Press.

Eliasmith, C. & Anderson, C.H. (1999). Developing and applying a toolkit from a general neurocomputational framework. *Neurocomputing*, *26-27* (0), 1013-1018. 10.1016/S0925-2312(99)00098-3

——— (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems.* Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y. & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338* (6111), 1202-1205. 10.1126/science.1225266

Eliasmith, C. & Trujillo, O. (2013). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology*, *25*, 1-6. 10.1016/j.conb.2013.09.009

Esser, S. K., Andreopoulos, A., Appuswamy, R., Datta, P., Barch, D., Amir, A., Arthur, J., Cassidy, A., Flickner, M., Merolla, P., Chandra, S., Basilico, N., Carpin, S., Zimmerman, T., Zee, F., Alvarez-Icaza, R., Kusnitz, J. A., Wong, T. M., Risk, W. P., McQuinn, E., Nayak, T. K., Singh, R. & Modha, D. S. (2013). Cognitive computing systems: Algorithms and applications for networks of neurosynaptic cores. *The 2013 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-10). IEEE. 10.1109/IJCNN.2013.6706746

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28* (1-2), 3-71. 10.1016/0010-0277(88)90031-5

Fox, D. (2009). IBM reveals the biggest artificial brain of all time. *Popular Mechanics*

Galluppi, F., Denk, C., Meiner, M. C., Stewart, T., Plana, L. A., Eliasmith, C., Furber, S. & Conradt, J. (2014). Event-based neural computing on an autonomous mobile platform. *Proceedings of IEEE international conference on robotics and automation (ICRA)*. IEEE.

De Garis, H., Shuo, C., Goertzel, B. & Ruiting, L. (2010). A world survey of artificial brain projects, Part I: Large-scale brain simulations. *Neurocomputing*, *74* (1-3), 3-29. 10.1016/j.neucom.2010.08.004

Goodman, W. C. (1996). Software and engineering industries: Threatened by technological change? *Monthly Labor Review*, *119* (8), 37-45. Bureau of Labor Statistics, U.S. Department of Labor. 10.2307/41844604

Graves, A., Mohamed, A. & Hinton, G. E. (2013). speech recognition with deep recurrent neural networks. *IEEE international conference on acoustic speech and signal processing (ICASSP)* (pp. 6645-6649). Vancouver, Canada: IEEE.

Hasler, J. & Marr, H. B. (2013). Finding a roadmap to achieve large neuromorphic hardware systems. *Frontiers in Neuroscience*, *7* (118), 1-29. 10.3389/fnins.2013.00118

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* Oxford, UK: Oxford University Press.

Khan, M. M., Lester, D. R., Plana, L. A., Rast, A., Jin, X., Painkras, E. & Furber, S. B. (2008). *SpiNNaker:*

Mapping neural networks onto a massively-parallel chip multiprocessor. IEEE. (pp. 2849-2856). 10.1109/IJCNN.2008.4634199

Khatib, O. (1987). A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal of Robotics and Automation*, *3* (1), 43-53.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems 25* (p. 4).

Kumar, S. (2013). Introducing qualcomm zeroth processors: Brain-inspired computing. http://www.qualcomm.com/media/blog/2013/10/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing

Kurzweil, R. (2005). *The singularity is near.* New York, NY: Penguin Books.

Leung, E. L., Cao, Z., Jiang, Z., Zhou, H. & Liu, L. (2013). Network-based drug discovery by integrating systems biology and computational technologies. *Briefings in bioinformatics*, *14* (4), 491-505. Oxford, UK: Oxford University Press.

Lichtsteiner, P., Posch, C. & Delbruck, T. (2008). Temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, *43* (2), 566-576.

Li, C., Delbruck, T. & Liu, S. (2012). Real-time speaker identification using the AEREAR2 event-based silicon cochlea. *2012 IEEE International Symposium on Circuits and Systems* (pp. 1159-1162). 10.1109/ISCAS.2012.6271438

Litt, A., Eliasmith, C., Kroon, F., Weinstein, S. & Thagard, P. (2006). Is the brain a quantum computer? *Cognitive Science*, *30* (3), 593-603.

Maass, W., Natschläger, T. & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, *14* (11), 2531-2560. Cambridge, MA: MIT Press.

Malone, P. (2014). Wealthy need to share the spoils of automation. http://www.canberratimes.com.au/comment/wealthy-need-to-share-the-spoils-of-automation-20140301-33sp6.html

Manyika, J., Chui, M., Bughin, J. & Dobbs, R. (2013). Disruptive technologies: Advances that will transform life, business, and the global economy. McKinsey Global Institute.

Marcus, G. (2013). Is "Deep Learning" a revolution in artificial intelligence? *The New Yorker.*

Metzinger, T. (2013). Two principles for robot ethics. In E. Hilgendorf and J. P. Günther (Eds.) *Robotik und Gesetzgebung* (pp. 263-302). Baden-Baden, GER: Nomos.

Newquist, H. P. (1994). *The brain makers.* Indianapolis, IN: Sams Publishing.

Norman, D. A. (1986). Reflection on cognition and parallel distributed processing. In J. L. McClelland and D. E. Rumelhart (Eds.) *Parallel distributed processing: Exploration inthe microstructure of cognition* (p. 531). Cambridge, MA: MIT Press.

Padbury, P., Christensen, S., Wilburn, G., Kunz, J. & Cass-Beggs, D. (2014). MetaScan 3: Emerging technologies. *MetaScan 3: Emerging technologies* (p. 45). Cambridge, MA: Policy Horizons Canada. http://www.horizons.gc.ca/eng/content/metascan-3-emerging-technologies-0

Pinker, S. & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28* (1-2), 73-193. 10.1016/0010-0277(88)90032-7

Plewes, T. J. (1990). Labor force data in the next century. *Monthly Labor Review*, *113* (4). Bureau of Labor Statistics, U.S. Department of Labor.

Print Edition (2014). The future of jobs: The onrushing wave. *The Economist.*

Robertson, J., Sheppard, T. & Sarnes, S. (2005). Workers compensation claim frequency continues to decline, particularly for smaller claims. *NCCI Research Brief*, *2*

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. I. Poser (Ed.) *The architecture of mind: A connectionist approach* (pp. 133-159). Cambridge, MA: MIT Press.

Rutkin, A. H. (2013). Report suggests nearly half of U.S. jobs are vulnerable to computerization. *MIT Technology Review.*

Schaal, S., Mohajerian, P. & Ijspeert, A. (2007). Dynamics systems vs. optimal control: A unifying view. *Progress in brain research*, *165* (1), 425-45. 10.1016/S0079-6123(06)65027-9

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3* (03), 417-424. Cambridge, UK: Cambridge University Press. 10.1017/S0140525X00005756

Simon, H. A. & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, *6* (1), 1-10.

Stevenson, I. H. & Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature neuroscience*, *14* (2), 139-42. 10.1038/nn.2731

Stunt, V. (2014). Why Google is buying a seemingly crazy collection of companies. *CBC News.*

Sussillo, D., Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, *63* (4), 544-57. 10.1016/j.neuron.2009.07.018

Thagard, P. (2010). *The brain and the meaning of life.* Princeton, NJ: Princeton University Press.

——— (2014). Explanatory identities and conceptual change. *Science & Education*, *23* (7), 1531-1548. 10.1007/s11191-014-9682-1

Todorov, E. (2008). Optimal control theory. In K. Doya (Ed.) *Bayesian brain: Probabilistic approaches to neural coding* (pp. 269-298). Cambridge, MA: MIT Press.

——— (2009). Parallels between sensory and motor information processing. In M. S. Gazzaniga (Ed.) *The cognitive neurosciences* (pp. 613-624). Cambridge, MA: MIT Press.

Ulam, S. (1958). Tribute to John von Neumann. *Bulletin of the American Mathematical Society*, *64* (3), 5.

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* (pp. 11-22). Westlake, OH: NASA Conference Publication 10129.

Future Games

A Commentary on Chris Eliasmith

Daniela Hill

In this commentary, the future of artificial minds as it is presented by the target article will be reconstructed. I shall suggest two readings of Eliasmith's claims: one regards them as a thought experiment, the other as a formal argument. While the latter reading is at odds with Eliasmith's own remarks throughout the paper, it is nonetheless useful because it helps to reveal the implicit background assumptions underlying his reasoning. For this reason, I begin by "virtually reconstructing" his claims as an argument—that is, by formalizing his implicit premises and conclusion. This leads to my second claim, namely that more than technological equipment and biologically inspired hardware will be needed to build artificial minds. I then raise the question of whether we will produce *minds* at all, or rather functionally differentiated, fragmented derivates which might turn out not to be notably relevant for philosophy (e.g., from an ethical perspective). As a potential alternative to artificial minds, I present the notion of postbiotic systems. These two scenarios call for adjustments of ethical theories, as well as some caution in the development of already-existing artificial systems.

Keywords
Artificial minds | Artificial systems ethics | Biological cognition | Mindedness | Postbiotic system

Commentator

Daniela Hill
daniela.hill@gmx.net
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Chris Eliasmith
celiasmith@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

1 Introduction

This commentary has two main aims: First, it aims to reconstruct the major important predictions and claims Eliasmith presents in his target article as well as his reasons for endorsing them. Second, it plays its own version of "future games"—the "argumentation game"—by taking some suggestions presented by Eliasmith maximally seriously and then highlighting problems that might arise as a consequence. Of course, these consequences are of a hypothetical nature. Still, they are theoretically relevant for the question of what will be needed to build full-fledged artificial cognitive agents.

Chris Eliasmith discusses recent technological, theoretical, and empirical progress in research on Artificial Intelligence and robotics. His position is that current theories on cognition, along with highly sophisticated technology and the necessary financial support, will lead to the construction of sophisticated-minded machines within the coming five decades (Eliasmith this collection, p. 2). And also vice versa: artificial minds will inform theories on biological cognition as well. Since these artificial agents are likely to transcend humans' cognitive performance, theoretical (i.e., philosophical and ethical) as well as pragmatic (e.g., legal and cultural laws etc.) consequences have to be considered throughout the process of developing and constructing such machines.

Hill, D. (2015). Future Games - A Commentary on Chris Eliasmith.
In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 12(C). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570610

The ideas Eliasmith presents are derived from developments in three areas: technology, theory, and funding; and I will demonstrate the background assumptions underlying these. In this way, I want to demonstrate that if we read Eliasmith as defending a formal argument (rather than a thought experiment), this argument has the form of a *petitio principii*. To illustrate this very clearly, a formal reconstruction of the (not explicitly endorsed, but implicitly assumed) arguments will be conducted. I then argue that even though they are constructed as arguments, and Eliasmith's claims fail, his suggestions provide an insightful contribution to the philosophical debate on artificial systems and the near future of related research. I further want to stress that we should perhaps confine ourselves to talking about less radical alternatives that do not necessarily include the mindedness of artificial agents, but have some element of biological cognition (architecture or software) in them. A number of subordinate questions have to be looked at in order to arrive at a point where a justified statement about the possibility of phenomenologically convincing artificial minds can be made. These considerations include more possibilities than simply the dichotomy of human-like vs. artificial. This is due to our *having* to think about possibilities that lie between or beyond these two extremes, such as fragmented minds and postbiotic systems, since they might soon emerge in the real world. The way in which these will be relevant to philosophy will be largely a question of their psychological make-up—most notably, their ability to suffer.

To start with, the following two sections will present some relevant aspects of the position expressed in the target article. They will summarize, and highlight some of the article's many informative and noteworthy suggestions. I shall also bring in some additional thoughts that I consider important. Afterwards, I will play a kind of future game of my own: I take Eliasmith's predictions very seriously and point at some of the problems that might arise if we were to take his suggestions as arguments. To be fair, Eliasmith himself says that what he presents are "likely wrong" predictions (this collection, p. 3). So on a more charitable reading, his claims are not intended to be arguments at all. Yet the attempt to reconstruct them as a formal argument has the advantage of showing that his claims are based on a reasoning that is itself problematic.

2 Are artificial minds just around the corner?

Eliasmith's perspective on the architecture of minds is a functionalist one (this collection, p. 2, p. 6, pp. 6–7, pp. 9–11, p. 13). The thread running through his paper is his interest in "understanding how the brain functions" and realizing "detailed functional models of the brain" (ibid., p. 9). The basic idea is that if we construct artificial minds and endow them with certain functions (such as natural language and human-like perceptual abilities), we can examine empirically, in a process comparable to reverse engineering, what it is that constitutes so-called mindedness (ibid., p. 11). But in their striving to unearth the nature of mindedness, it is not the task of artificial intelligence research or biology to deliver comprehensive and full-fledged theories on biological cognition in general and human cognition in particular. Rather, a very interesting reciprocal relationship between the two parties, in which one learns from the other, is what will propel forward our understanding of biological cognitive systems. In the following I give an overview of the most relevant points that are presented in the target article. They will be divided up into the original sections (technical, theoretical, and empirical).

First, in the technical area and according to Eliasmith, we are fairly far advanced—although there are certain hindrances to successfully implementing theories on this technology. The main obstacle is the size of artificial neuronal systems and, connected to that, their power consumption. Even though neuromorphic chips are being improved steadily, the number of neurons that can be reproduced artificially is still much lower than the number of neurons a human brain has. Thus, the processing of information is significantly slower than in natural cognitive systems (Eliasmith this collection, p.

14). Consequently, what can be realized in the field is still far from the complexity displayed by natural, biological cognition. However, as Eliasmith argues, since we are already in possession of the theoretical groundwork, the main barrier to overcome are technological advances (ibid., p. 9). Throughout the paper, Eliasmith informs the reader that in case we had the technologies needed, artificial minds would immediately be created (ibid., e.g., p. 7, p. 9, p. 11). However, where Eliasmith emphasizes technological barriers, I would like to point out that *theoretical* obstacles exist as well. These mainly revolve around the fact that a system of ethics has to be created *before* we encounter artificial agents. Eliasmith also comments on the consequences for philosophy, arguing that some major positions in the philosophy of mind, such as functionalism, will receive more empirical grounding (ibid., p. 11).

It seems as if the tacit understanding that Eliasmith has of the function of artificial minds is that they serve as shared research objects of biology and artificial research science in order to gain a better understanding of biological cognition (this collection, p. 9). That is of course only true if indeed the functional architecture of the artificial agent produces convincing behavior, similar to that of biological cognitive systems (humans and animals alike). To illustrate possible problems, one can think of the fact that in research, we learn from animal experiments, even though these animals are quite different from us in many ways. They are, however, similar or at least comparable in one epistemically relevant and specific aspect, i.e., the one that is to be examined, for example in certain aspects of metabolism used to test whether a new drug causes liver failure in humans (Shanks et al. 2009, p. 5). It is the same with artificial agents: they are similar to us in their behavior and thus a worthwhile research object. As such, we could formulate the underlying reasoning as a variant of analytical behaviorism. Analytical behaviorists suppose that intrinsic states of a system are mirrored in certain kinds of behavior. Two systems displaying identical behavior on the outside can be investigated in order to detect whether they do so on the inside as well (Graham 2010). This means that we could gain insight on the origin of mental states from a functionally isomorphic system, i.e., an artificially constructed system that is identical in organization and behavior to the natural system copied.

Last, since it seems that it will be possible in the future, given the required hardware, to design artificial agents according to our needs, it does not appear far-fetched to assume that the quality of human life might consequently be improved to a great extent (Eliasmith this collection, p. 11). This requires, however, that we make up our own minds about how to interact with such agents, which rights to grant and which to deny them. And also the opposite case may not be disregarded: it is imaginable that the artificial agents will at some point turn the tables and be the ones to decide on *our* rights (cf. Metzinger 2012). In highlighting aspects from different areas to be considered, Eliasmith reminds us of the possibilities that lie ahead of us, but also of the challenges that might show up and have to be faced. I want to suggest that we also take into consideration alternative outcomes that are not minds in the biological sense, but rather derivates of minds. I will therefore put the notion of postbiotic systems into play as a way of escaping the dichotomy "human-like" vs. "artificial" (Metzinger 2013). The philosophical point here is that the conceptual distinction between "natural" and "artificial" may well turn out to be non-exhaustive and non-exclusive: there might well, as Metzinger points out, be future systems that are neither artificial nor biological. By no means do I intend to argue against the use of scientific models, since they are what good research needs. Rather, I wish to draw attention to the possible emergence of intermediate systems, rather than only the extremes (i.e., human-like vs. artificial agents), or classes of systems that go beyond our traditional distinctions, but which nevertheless count as "minded". As mentioned above, this is due to these intermediate or postbiotic systems being possible much earlier—probably preceding full-blown minded agents.

I will end this section by drawing attention to some of the author's thoughts on the crucial elements of artificially-minded systems. According

to Eliasmith, three types of skills are vital in building artificial minds: cognitive, perceptual, and motor skills have to be combined to create a certain behavior of the minded artificial agent. This behavior will then serve as the basis for us humans to judge whether we perceive the artificial agent as "convincing" or not (Eliasmith this collection, p. 9). Unfortunately, no closer specification of what it is to be "convincing" is given in the target article. No theoretical demarcation criterion is offered. What we can say with great certainty, however, is that in the end our subjective *perception* of the artificial agents will be the decisive criterion. One could speculate on whether it is merely an impression, or even an illusion, that leads us to concluding that we are facing a *minded* agent. According to Eliasmith, any system that produces a robust social hallucination in human observers will count as possessing a mind.

3 Playing the "argumentation game"

In the following I will play the "argumentation game" and for a moment assume that what Eliasmith presents us with actually is argumentation. The goal of this section is not to claim that Eliasmith really *argues* for the emergence of artificial minds in the classical way. Rather, I wish to highlight that possibly more than technological equipment and biologically inspired hardware need to be taken into account before research can present us with a mind, as outlined by Eliasmith. If we deconstruct his line of reasoning and virtually formalize the *argument*, we don't find valid argumentation but rather a set of highly educated—and certainly informative—claims about the future, which doubtlessly help us prepare for a future not too far ahead of us. I will utilize the terms "argumentation", "argument", "premise", and "conclusion" in the following, but it should always be remembered that these terms are only "virtually" or hypothetically. So let us see how Eliasmith proceeds:

If we play the argumentation game, a first result is that Eliasmith's virtual argument becomes problematic at the moment he starts elaborating on theoretical developments that have been made and that will propel forward the development of "brain-like models" (this collection, p. 6). From the perspective of an incautious reader, the entire section "Theoretical developments" could be seen as resulting in a claim that can be traced back to a *petitio principii*. This means that the conclusion drawn at the end of the argumentative line is identical with at least one of the implicit premises. The implicit argumentation is made up of three relevant parts and unfolds as follows: first, building brain-like models is not only a matter of the available technological equipment (ibid., first paragraph; cf. premise 1). Instead, if we face a convincing artificially-minded agent, it is characterized by both sophisticated technological equipment and by our discovery of principles of how the brain functions, such as learning or motor control (ibid.; cf. premise 2). And so, in conclusion, it follows that if biological understanding and technological equipment come together, we will be able to build brain-like models and implement them in highly sophisticated cognitive agents (ibid.).

The incautious reader would now have to believe that Eliasmith is confusing necessary and sufficient conditions. Let us look at this assumed argument in some more detail. Formulated as a complete argument we would get: "If it is not the case that technological equipment *alone* leads to the building of brain-like models for artificial cognitive agents, but we face a good artificial minded agent which is endowed with certain technology as well as biologically inspired hardware, we have to conclude that this certain technology and biologically inspired hardware are not only necessary, but also sufficient for building brain-like models for artificial cognitive agents."

The formal expression of this argument would be the following:

T: We have developed sophisticated technological equipment.

B: We have developed biologically-inspired hardware.

M: We can build brain-like models which can be implemented in artificial cognitive agents.

$$\frac{\neg(T \rightarrow M) \qquad M \rightarrow (T \,\&\, B)}{(T \,\&\, B) \rightarrow M}$$

As is obvious from how the argument is constructed, it is invalid. So, what we can say at this point is that the combination of both technical features and biologically-inspired neuromorphic hardware very likely does get us some way, but we might have to consider which elements are missing so that we really end up building what will be perceived as minds. I shall propose some possibilities in the following section. The author even supposes that we will be able to build artificial agents ready to rival humans in cognitive ability (Eliasmith this collection, p. 9). I am convinced that it is not cognitive artificial agents that will be the crucial hurdle, but rather their mindedness. I am also convinced that the huge amount of money spent on certain research projects will most likely result in improved models of the brain, as suggested by Eliasmith (ibid., p. 8), but it is not obvious to me how investing a vast amount of money necessarily results in relevant findings. It is also possible that no real progress will be made. Stating the opposite, which Eliasmith does not, resembles a claim based on expertise as bulletproof evidence. Sure enough, monetary sources are needed to make progress, but they are no *guarantee*. So possibly technology, biological theories on the brain's functioning, and money, essentially, might not lead to sophisticated cognitive agents being built (ibid.). The point is not that we should not invest money unless a positive outcome is guaranteed. Rather, we need a theoretical criterion for mindedness that is philosophically convincing—and not only robust, but epistemically unjustified social hallucinations. This theoretical criterion is what we lack.

4 What could artificial minds be?

In this section, I intend to sketch some important issues and questions for the future debate on artificial minds. I shall examine whether predictions on the concept of *artificial minds* can be made at the present state of the debate and based on the empirical data we currently have. This involves knowledge about what a mind is, and knowledge about how an *artificial* mind is characterized. In reconstructing Eliasmith's understanding of what a mind is, we may find the following statement informative: he relies on behavioral, theoretical, and similarity-based methods (this collection, p. 3). The possible problem with this approach is that the characterization of the methods is very limited. To point to some relevant questions: what is the behavior of a mind? What about the fact that *mind* is not even close to being well understood theoretically? How do similarity-based methods avoid drawing problematic conclusions from analogies (cf. Wild 2012)? Importantly, at this point we are only talking about natural, biologically-grounded minds. Answers as to what an *artificial* mind is supposed to be might exceed the concept of mind in ways we are unable to tell at the present moment.

Let us see how Eliasmith characterizes artificial *minds*. One can see this as a judgment based on the similarity of behavior originating from two types of agents: humans and artificial. Functions need to be developed that are necessary for building an artificial mind. These functions lead to a certain kind of behavior. This behavior is achieved by perceptive, motor, and cognitive skills, which are needed to make the behavior seem human-like. Thus, the functions implemented on sophisticated kinds of technology will, in the end, lead to human-like behavior (Eliasmith this collection, p. 9). The reason why the argumentative step from cognition, perception, and motor skills to mindedness can be made is the underlying assumption that the behavior resulting from these three types of skills is *convincing* behavior in our eyes (Eliasmith this collection, p. 10). Similarity judgments, so Eliasmith argues, might appear "hand-wavy". Still, he uses them to reduce the complexity that mindedness brings with it (ibid., pp. 5–6), and he certainly succeeds in drawing attention to a whole range of important issues. However, it could well be that the reduction to human-like behavior as the benchmark for assessing mindedness is too simple. After all, analytical behaviorism today counts as a failed philosophical research program. There could be much more to mindedness than behavior. We just do not know what this is yet. As a possible candidate we might consider the previously

mentioned psychological make-up of artificial agents, such as their being endowed with internal states like ours. One might think of robust first-person perspectives, but also about emotions like pain, disappointment, happiness, fear, and the ability to react to these. Other options include interoceptive awareness or the ability to interact socially—and much more.

5 What should we brace ourselves for?

Given the complexity of mindedness and our very limited understanding of what constitutes it, what else can we talk about? We could consider further possibilities of artificial systems that might arise, thereby enlarging the set of constraints that has to be satisfied. Some of them seem much more likely than artificial minds, and they might precede minds chronologically. I would like to focus on the idea of *fragmented minds* on the one hand and of *postbiotic systems* on the other, as two versions of artificial systems. An artificially-constructed fragmented mind is characterized by only partial satisfaction of the constraints fulfilled by a human mind. It could thus, very much like autistic persons with savant syndrome (i.e., more than average competence in a certain domain, e.g., language learning or music), and possess only some of our cognitive functions, but be strikingly better at them than normal humans are and ever could be, given their biological endowment.[1] Postbiotic minds, on the other hand, could satisfy additional constraints that are not yet apparent presently. I will conclude with some reflections on the new kind of ethics that will have to be created in order to approach new kinds of cognitive agents. As pointed out above, I assume that cognitive agents will be possible much earlier than truly minded agents. Learning, remembering, and other cognitive functions can already be recreated in artificial systems like *Spaun*. Still, human cognition is very versatile and complex. A fully minded agent, in contrast to a merely cognitive agent, might also be able to experience herself as a cognitive agent.

1 In that case, the variable **B** from above (biologically inspired hardware) would not be a necessary condition for finding out more about mindedness.

Therefore, I propose that cognitive systems could be created that do not yet qualify as a copy of our cognitive facilities, but which cover only parts of our cognitive setup. I call these *fragmented minds*. Importantly, the word *minds* does not refer here to the artificiality of the system at all. There are human beings with fragmented minds, too, such as babies, who do not yet display the cognitive abilities we ascribe to adult humans in general, or the aforementioned autistic humans with savant syndrome. Fragmented minds are contrasted with what we experience as normal human minds. *Fragmented* means that the created system possesses only part of the abilities that our mind displays. The term *mind* delineates the—historically contingent—point of reference that is human beings. How are fragmented minds further characterized? Eliasmith himself gives us an example: we could design a robot (an artificial mind) that gains fulfillment from serving humans (this collection, p. 11). This would only be possible if aspects of our own minds were not part of the mental landscape of this robot. We could roughly formulate such an aspect, such as the will to design one's own life. Folk psychology would most likely regard this robot as lacking a free will, which is in conformity with the idea of slavery that Eliasmith acknowledges (ibid.). So a fragmented mind is an artificial system that possesses part of a biological cognitive system's abilities instead of the rich landscape most higher animals (e.g., some fishes and birds, certainly mammals), as well as humans, display.

Related to the aspect of fragmented minds is the idea that we could refrain from creating minds that might cause us a lot of moral and practical trouble, and instead focus on building sophisticated robots designed to carry out specific kinds of tasks. Why do we need to create artificial *minds*? What is the additional value gained? If these robots are not mindful, we will circumvent the vast majority of conceptual and ethical problems, such as legal questions (What is their legal status compared to ours?) or ethical considerations (If I am not sure whether an artificial agent can perceive pain, how should I treat it in order to not cause harm?). In which case, they

would only be more capable technology than what we know at present, and most likely be of no major concern for the philosophy of mind. However, if they *are* mindful, we doubtlessly have to think about new ways of approaching them ethically.

Also ethically relevant are intermediate systems, systems that are not clearly either natural or artificial. These systems have been called *postbiotic systems* (Metzinger 2012, p. 268). What characterizes postbiotic systems is the fact that they are made up of both natural and artificial parts, thus belonging to neither of the exhaustive categories "natural" or "artificial". In that way a natural system, e.g., an animal, could be controlled by artificially-constructed hardware (as in hybrid bio-robotics); or, in the opposite case, artificial hardware could be equipped with biologically-inspired software, which works in very much the same way as neuronal computation (Metzinger 2012, pp. 268–270; Metzinger 2013, p. 4). Perhaps Eliasmith's own brain-like model *Spaun* is a postbiotic system in this sense, too. In what way would these systems become ethically relevant? Although the postbiotic systems in existence today do not have the ability to subjectively experience themselves and the world around them, they might have it in the future. In being able to subjectively experience their surroundings, they are probably also able to experience the state of suffering (Metzinger 2013, p. 4). Everything that is able to consciously experience a frustration of preferences as a frustration of its *own* preferences automatically becomes an object of ethical consideration, according to this principle. For such cases, we have to think of ethical guidelines *before* we are confronted with a suffering postbiotic mind, which could be much earlier than we expect. Before thinking about how to implement something as complex and unpredictable as an artificial *mind*, one should consider what one does *not* want to generate. This could, for example, be the ability to suffer, the inability to judge and act according to ethical premises, or the possibility of developing itself further in a way that is not controllable by and potentially dangerous for humans.

6 Conclusion

In this commentary, I have played the "argumentation game" as my own version of Eliasmith's "future game". The intention behind this was to demonstrate that we very likely need more than sophisticated technology and biologically-inspired hardware to build brain-like models ready to be applied in artificial cognitive agents. As such, I playfully took Eliasmith's considerations on the future of artificial minds as arguments, and demonstrated that they would result in a *petitio principii*. In so doing, I highlighted that necessary conditions do not have to be sufficient as well. While this is common philosophical currency, it is instructive to spell this out in the case of artificial agents. So in the present case, what constitutes artificial cognitive systems and what is needed to gain a deeper understanding of how the mind works might include more factors than the two crucial ones Eliasmith outlines, namely biological understanding and its implementation in highly-sophisticated technology. I proposed some possibilities that might turn out to be informative for future considerations on what constitutes an artificial mind. In particular, I mentioned experiential aspects, such as the perception of emotions and reactions to them, as well as internal perceptions like interoceptive awareness. In general, this means that we need theoretical criteria that are convincing for philosophy in order to overcome referring to robust yet convincing social hallucinations. Further, to illustrate that the distinction between natural and artificial systems might not be exhaustive, I pointed to the notions of fragmented minds and postbiotic systems as possible developments for the nearer future. They have to be considered, in particular with respect to their ethical implications, before they are developed and implemented in practice.

Even though we lack a more fine-grained, deeper understanding of what constitutes minds, Eliasmith shows us that it is worth thinking about what we already *do* have at hand for constructing artificially-minded systems. He demonstrates vividly that two factors —technology and biology—are of major import-

ance on the route to artificially-cognitive, if not minded, agents. And he brings into discussion a number of far-reaching consequences that will apply in case we do succeed in building artificial minds within the next five decades. These will inform the development of these artificial systems as well as philosophical debate, both on an ethical, as well as theoretical level. In this way, Eliasmith's contribution has to be regarded as significant in terms of preparing us for the decades to come.

Acknowledgements

First and foremost, I am grateful to Thomas Metzinger and Jennifer M. Windt for letting me be part of this project, thus providing me a unique opportunity to gain valuable experience. Further special thanks go to the two anonymous reviewers, as well as the editorial reviewers for their insightful comments on earlier versions of this paper. Lastly, I wish to express my gratitude to Anne-Kathrin Koch for sharing her expertise with me.

References

Eliasmith, C. (2015). On the eve of artificial minds. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Graham, G. (Ed.) (2010). Behaviorism. *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/behaviorism/

Metzinger, T. (2012). *Der Ego-Tunnel: Eine neue Philosophie des Selbst: Von der Hirnforschung zur Bewusstseinsforschung*. Berlin, GER: Bloomsbury.

——— (2013). Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.) *Robotik und Gesetzgebung* (pp. 263-302). Baden-Baden, GER: Nomos.

Shanks, N., Greek, R. & Greek, J. (2009). Are animal models predictive for humans? *Philosophy, Ethics, and Humanities in Medicine*, *4* (2), 1-20. 10.1186/1747-5341-4-2

Wild, M. (2012). *Fische. Kognition, Bewusstsein und Schmerz: Beiträge zur Ethik und Biotechnologie*. Bern, CH: Bundesamt für Bauten und Logistik BBL.

Mind Games

A Reply to Daniela Hill

Chris Eliasmith

In her discussion of my original article, Hill reconstructs an argument I may have been making, argues that the distinction between natural and artificial minds is not exclusive, and suggests that my reliance on behaviour as a determiner of "mindedness" is a dangerous slip back to philosophical behaviourism. In reply, I note that the logical fallacy of which I'm accused (circular reasoning) is not the one present in the reconstruction of my argument (besides the point), and offer a non-fallacious reconstruction. More importantly, I note that logical analysis does not seem appropriate for the discussion in the target article. I then agree that natural and artificial minds do not make up two discrete categories for mindedness. Finally, I note that my research program belies any behaviourist motivations, and reiterate that even though behaviour is typically important for identifying minds, I do not suggest that it is a substitute for theory. However, the target article is not about such theory, but about the near-term likelihood of sophisticated artificial minds

Keywords
Artificial minds | Behaviourism | Logical analysis | Minds

Author

Chris Eliasmith
celiasmith@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Commentator

Daniela Hill
dhill@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

I think Hill is right to wonder aloud about my methodology in the target article. After all, I just ignored the hard philosophical issue of saying what minds really are. I pretended (somewhat self-consciously) that we all know what minds are, and so that we will simply be able to tell when someone has created one, if they ever do. But, I did that for a reason. The reason was this: I did not want to get lost in the minutiae of metaphysics when my focus was on a technological revolution—one with significant philosophical consequences (which is also not to say I don't like such minutiae in their proper time and place).

2 A failure of logic

However, Hill was also not especially taken by the reasons I provided for expecting such developments either. Hill's suggestion is that the best reasonable argument you could construct from my original considerations was fallacious. Though, Hill is quick to point out that I didn't take myself to be constructing an argument: "... not to claim that Eliasmith really argues for the emergence of artificial minds" (this collection, p. 4).

Nevertheless, her analysis is that what I have provided is best understood as a *petitio principii* (aka circular argument): "this means that the conclusion drawn at the end of the ar-

Eliasmith, C. (2015). Mind Games - A Reply to Daniela Hill.
In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 12(R). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570788

gumentative line is identical with at least one of the implicit premises" (Hill this collection, p. 4). Unfortunately, the technical analysis offered (p. 5) is a *non sequitur* (i.e., there is no logical connection between the premises and conclusion). Regardless, one fallacy is as embarassing as the other.

However, I'd like to suggest that if we wanted to recast the original paper as a logical argument, then a simple *modus ponens* will do: if we have a good theory and the technological innovations necessary to implement the theory, then we can build a minded agent. We have good (and improving) theory and will have the proper technological innovations (in the next 50 years), therefore we will be able to build a minded agent (in the next 50 years). Indeed, most of the paper is arguing for the plausibility of these premises.

More to the point, however, I think that we can take this as an object lesson for when logical inference is really just the wrong kind of analysis of a paper. Instead of trying to provide a logical argument from which the conclusion necessarily follows from the premises, I am providing series of considerations that I believe make the conclusion likely given both the current state of affairs, and expected changes. In short, I think of the original paper as providing something more like a series of inferences to the best explanation: all of which are, technically, fallacious; and all of which are directed at establishing premises.

3 Back to minds

Despite disagreeing with the analysis of the logical structure of the paper, I do appreciate the emphasis that Hill has placed on philosophical and ethical aspects of our attempts to construct minds. In the original article, I only very briefly touch on those issues. However, I would be quick to point out that I do not think, and never intended to suggest, that the distinction between "natural" and "artificial minds" was an absolute, "exhaustive," or "exclusive" one (Hill this collection, p. 3). Like most interesting and complex features, possession of 'mindedness' no doubt comes in degrees. In fact, I think that our attempts to construct artificial minds will provide a much better sense of the dimensions along which such a continuum is best defined.

Finally, I must admit that I find it somewhat alarming that I'm being characterized as a behaviourist in Hill's article—*that* has definitely never happened before: "Let us see how Eliasmith characterizes artificial minds. One can see this as a judgment based on the similarity of behaviour originating from two types of agents: humans and artificial" (this collection, p. 5). Hence, I was espousing "analytical behaviorism… a failed philosophical research program" (Hill this collection, p. 6). Indeed, I, like all behavioural scientists, believe that behaviour is one important metric for characterizing the systems of interest. However, the reason I focus on internal mechanisms in my own research – all the way down to the neural – is that I believe those mechanisms give us critical additional constraints for identifying the right class of algorithms that give rise to behaviours. Consequently, I wholeheartedly agree with Hill that "There could be much more to mindedness than behaviour" (this collection, p. 6). So, for the record, I believe that our best theories for how to build minds are going to be highly informed by low-level mechanisms.

That being said, I also think that most people's judgments of whether or not something counts as being minded is going to come down largely to their being convinced of the naturalness, or "cognitiveness" of the behaviour that is exhibited by agents we construct. Notice that there is a difference between a claim of how people will judge mindedness, and a claim about theories of mindedness or how we ought to best achieve that judgment. Turing was, after all, onto *something* with his test.

4 Conclusion

I noted in the original article (Eliasmith this collection) that I was attempting to avoid becoming mired in tangential debates regarding what it is to have a mind by simplifying the criteria for mindedness (for the purposes of that article). Exactly the kinds of debates I was attempting to avoid are raised in Hill's comment-

ary. For example, I don't think we know if there is a clean contrast between a "fully minded agent" and a "merely cognitive agent" (Hill this collection, p. 6). Perhaps there is, and perhaps it is that a fully minded agent can "experience herself as a cognitive agent", (Hill this collection, p. 6) but perhaps not. This does not strike me as a decidable question at present.

So, perhaps my unwillingness to venture into the murky waters of necessary and sufficient conditions for having a mind came off as making me look like a behaviourist. But in truth, my purpose was rather to focus on providing a variety of evidence that I think suggests that artificial minds are not as far away as some have assumed. There is, I believe, a historically unique confluence of theory, technology, and capital happening as we speak.

References

Eliasmith, C. (2015). On the eve of artificial minds. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Hill, D. (2015). Future games - A commentary on Chris Eliasmith. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.