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.