What is it like to be AlphaGo?

Jonathan Simon

April 30, 2021

1 Introduction

You are *phenomenally conscious*: you have subjective experiences of colors and sounds and joys and pains. Your brain, the likely culprit, is a network of neurons — an organic neural network. There are also artificial neural networks. They power some of the technology we use every day, like Facebook's auto-tagger and Google Translate. Are *they* phenomenally conscious?

There is plenty of debate about whether the robots will be conscious *one day* — once they become generally intelligent, and have wide enough repertoires to pass for human, like they do on the HBO show Westworld. But consciousness might be possible without general intelligence or a wide repertoire. Animal researchers have been arguing that octopuses, fish and even bees are conscious.[1] Fish and bees certainly lack general intelligence, and have limited repertoires. So what if some implementations of the artificial neural networks we've already built are conscious

[1]Franks et. al. (2018), Klein and Barron (2016), Godfrey-Smith (2016), Tye (2017).

right now, even though they lack general intelligence and have limited repertoires?

Call the thesis that some implementations of AI algorithms will be conscious within the next few years the **Sentience Now** thesis. I don't aim to convince you that this thesis is true: I am not convinced myself. But I think we can do more than just point to theories out there, like panpsychism, that entail it. There is a middle ground between trying to convince you (too hard), and observing that some philosophical theory or other entails it (too easy). I will argue here that **Sentience Now** is a *live possibility*.

What do I mean by a live possibility? I mean a possibility that we are obliged to take at least somewhat seriously in moral and practical deliberation. We can implicitly develop the idea with the following precautionary principle:

Precaution Suppose that if P is true (and you know it) then you ought to Φ. Then if P is a live possibility (for you, given your epistemic state), you ought to Φ insofar as Φ-ing is easy for you all things considered.

This states a minimal condition on what it is for a possibility to be live. I mean for 'easiness all things considered' to cover a lot. We can say that if Φ-ing conflicts with something else you are supposed to do, then it isn't easy all things considered. If Φ-ing conflicts with something that some other live possibility requires, then it isn't easy all things considered. And if P requires both Φ-ing and Ψ-ing, and each is easy on its own, but both are demanding together, then neither is easy all things considered. So **Precaution** does not offer much guidance in hard cases. Still, it is a substantive claim: there are lots of things that would be easy for us that we do not do, and if *we* are, say, a federal agency, these things can be quite significant. So

guidance in easy cases is important, too. And I don't say that **Precaution** exhausts the requirements that arise when possibilities go live: it is just a minimal constraint to give us a handle on the idea.[2]

The crucial point is that there is a level of justification beneath that required for belief which suffices to render a possibility salient in deliberation, by rendering it live.[3] How do we recognize when something has achieved this level? I propose that the following two conditions jointly suffice:

Reasonableness It enjoys a reasonable amount of evidential support.

No Dealbreakers There is nothing that decisively counts against it.

Jointly satisfying **Reasonableness** and **No Dealbreakers** is an accomplishment. Of course, this may depend on the context, our standards, and so on.[4] But we can take **Precaution** to ensure that the contextual standard is fairly high, because insofar as **Precaution** is salient, we tether the outcome of deliberation to practical consequences.

As such, we can see how various philosophical theories fall short of the standard. Consider a radical form of panpsychism according to which every physical being is

[2]Cf Sebo (2018).

[3]I note that in the setting of a probabilistic expected utility framework we don't need a special notion of live possibility to make this point: it then suffices to stress why **Sentience Now** should enjoy considerably more credence than most of us accord to, e.g., panpsychism. But we might put the notion of live possibility to use even in this framework: we might think of it as a sufficient condition for a possibilty to be allotted *enough* credence that its contributions to the overall utility calculus are *significant* (where the evaluation of the italicized notions may depend on the context).

[4]Context may be relevant to what level of support counts as reasonable, and what forms of opposition count as decisive. cf Karen Lewis on Hajek on conditionals. Note also Stalnaker on live options. Note finally that my notion is distinct from that of Isaac Levi ("serious possibility"). Levi's serious possibilities require a higher level of credence than my live possibilities.

conscious, including protozoa and protons. The theory is certainly easy to state. This does not mean that it is theoretically elegant, all things considered. But even if it were, theoretical elegance on its own does not amount to a reasonable amount of evidential support. So radical panpsychism does not satisfy **Reasonableness**.

Now consider an implementation-agnostic form of functionalism, which identifies some classical algorithm, A, as the algorithm for consciousness, and states that any implementation of A is conscious. This implies that nation-of-China implementations and tinker-toy implementations of A are conscious. But we have powerful, direct intuitions that nation-of-China and tinker-toy implementations of classical algorithms are not (ipso facto) conscious. So this implementation-agnostic form of functionalism does not satisfy **No Dealbreakers**.

So it is far from trivial that a given possibility comes to count as live. In the remainder of this paper, I will argue that **Sentience Now** is a live possibility. There is a further question of what moral or practical obligations this entails. I will comment on this further question but will not seek to adjudicate it here.[5]

I will proceed as follows. In §2 I argue that **Sentience Now** enjoys a reasonable amount of evidential support. In §2.1 I note that there are several well-motivated theories of consciousness that credit consciousness to an organizational structure that is present in some animals with limited repertoires, including fish and bees. In §2.2 I argue that standard AI architectures exhibit different fragments of this organizational structure, and newer architectures increasingly put these fragments together. I take this to show that **Sentience Now** enjoys a reasonable amount of

[5]Compare Schwitzgebel and Basl 2019, *Aeon*

evidential support, even if its level of support falls short of being convincing. Then in §§3-5, I consider different candidate "deal-breaker" objections, and argue that none of them are decisive. In §3 I consider potential dealbreakers at the algorithmic level: features of the algorithms implemented by contemporary AI that might seem to show that these algorithms do not suffice for subjective consciousness. I will argue that while some of these features may show that contemporary AI algorithms do not suffice for general intelligence, none show that they are insufficient for subjective consciousness. In §4 I consider potential dealbreakers at the implementation level: features of the implementations of contemporary AI algorithms that might seem to show that they cannot support subjective consciousness. I argue that none of these considerations shows that these implementations cannot support subjective consciousness. In §5 I consider the objection that there is simply no fact of the matter about whether near-future AIs are conscious. In reply I summarize my argument, given elsewhere, that where subjective consciousness is concerned there is always a fact of the matter, even if it is hard to know what it is. Finally in §6 I discuss the question of what moral or practical implications the live possibility of **Sentience Now** might have.

2 From Fish and Bees to AlphaGo

2.1 Subcortical Integration May Suffice for Consciousness

Many leading theories of consciousness construe it as a specific form of integration. Global workspace theories equate consciousness with the specific sort of integration associated with short-term working memory, which in humans is a form of broadcasting hub in the pre-frontal cortex which uses attention to select salient inputs from across the brain and broadcast them for use by many systems including action planning, language and metacognition.

The more attention is paid to the details of the human implementation of such a mechanism, the more it appears to require sophisticated cortical functionalities of the sort that we associate with general intelligence.[6] However, the more we abstract away from those details, the more permissive our criteria becomes. Tye (2002) for example defends the PANIC theory of consciousness on which any content is conscious which is poised (ready for use), abstract, non-conceptual, and intentional, and he argues that creatures such as bees may in fact have states or contents that fit this description.

A distinct conception of consciousness as integration is Merker (2007). Here the focus is on a more specialized form of integration located in the midbrain (in humans) whose function is to allow the organism to distinguish perceptual changes owing to environmental change from those owing to the agent's own movements (re-afference) and to integrate information about an organism's status, needs, and relation to its environment in order to select for action (see Klein and Barron 2016 for summary).

[6] Carruthers (2019).

Klein and Barron (id) argue that these functionalities can be found in insects such as bees, concluding that the latter are likely to exhibit some form of consciousness.

Also worth mentioning is Tononi et. al. (2016)'s Integrated Information Theory (IIT). According to this theory the integration of information as such, without any reference to its specific function, suffices for consciousness. This theory entails that most causally unified systems exhibit some degree of consciouness.

For our purposes what matters is that at least one of the more permissive integrativist theories on the table is a live possibility, in the above sense of **Reasonableness** and **No Dealbreakers**. This must be assessed case by case. IIT's formal or mathematical successes may perhaps suffice for **Reasonableness** (see for example Tegmark 2014). However, it seems to entail that CD players are conscious (Aaronson 2015) which arguably amounts to a dealbreaker.

Permissive global workspace and brainstem integration theories may fare better. Both identify mechanisms whose existence and importance is supported by substantive evidence. And it is also established that in both cases, interfering with these mechanisms interfere with functions associated with consciousness in humans. So in both cases **Reasonableness** seems to be satisfied.

What about dealbreakers? Permissive global workspace theory says that the basis of consciousness is the global workspace, irrespective of whether systems such as metacognition and language are consumers of the global workspace's content. It might seem that a dealbreaker here is that on this theory, the interruption of metacognition, language and so on should not interfere with consciousness. But this would be a fallacy: it is one thing to say that a global workspace, together with

its consumer systems in a given organism, suffice for consciousness, and another to say that an organism might be conscious even if the normal consumer systems of its global workspace are damaged.

Similarly, an apparent dealbreaker for the brainstem integration theory is that it seems to deny that specific cortical structures are essential for specific modalities of experience. Visual experience, for example, appears to depend on visual cortical areas such as V1, V4 and so on. But a cinematic metaphor may help here. Integrative activity in the midbrain may serve to turn the movie projector on, while more specific perceptual functionalities, realized in the cortices in the human case, may determine what is playing. So I provisionally conclude that there are no dealbreakers here either.

2.2 Subcortical Style Integration as a development of Deep Reinforcement Learning

In this section, I'll argue that if bees can do it, AIs can do it too. In particular, the conception of consciousness associated with permissive global workspace theory and brainstem integration theory have analogues in the space of architectures of current or near future artificial neural networks. This entails that the hypothesis that such systems are conscious satisfies **Reasonableness**, provided that the very suggestion that artificial systems are consciousness is not itself a dealbreaker. But to insist on that latter point is just to insist on an austere, Searlian conception of the biological nature of consciousness. This is a difficult line to maintain. What is it about biological systems per se that privileges them? One issue here is that while

biological theories are typically contrasted with functional ones, in fact the difference between, say, carbon and silicon is ultimate a functional difference: a question of how the same cast of sub-atomic characters interact with one another. Why should one such underlying pattern of interaction (that associated with carbon molecules) turn out to be crucial even though it does not matter to the functions associated with consciousness as we generally understand them? Intuitions such as those behind Searle's Chinese room and Block's nation of China thought experiments certainly tell us something about which implementations are appropriate, but the dealbreaker intuition does not seem to be at the computational (rather than the algorithmic or implementational) levels.[7]

Deep reinforcement learning already fosters us with models that fit the profile of integration implemented in the human brainstem. A game-playing system such as AlphaStar, for example, takes in pixels as inputs but also maintains information about the avatar, integrating the two in a manner that facilitates action planning. (THIS IS A STUB, ELABORATE)

Attentional-mechanisms in transformers and RNNs point in the direction of what a permissive global workspace might look like in an artificial neural network. GPT-3 might not be conscious, but the capacity to consider all of the elements of a sequence and their relations to each other simultaneously, and for this to be regulated by dynamically modified, learned attention, suggest a crucial step in the direction of an artificially implemented global workspace. Bengio (2019) investigates modifications to such systems that may bring us even closer to the functionality of the human

[7]Chalmers (2016)'s dancing and fading qualia arguments present further serious difficulties for the claim that only carbon-based systems can be conscious.

global workspace. (THIS IS A STUB, ELABORATE)

I this section I have presented the hypothesis that current systems like AlphaStar or near future ones like (hypothetical) transformers that meet the constraints described in Bengio (2019) are conscious. I conclude that this hypothesis satisfies **Reasonableness**, and that it does not encounter dealbreakers at the computational level. I turn now to the question of whether it encounters dealbreakers at either the algorithmic and implementational level.

3 Algorithms

I now begin my survey of possible dealbreakers: arguments or intuitions that might decisively defeat any argument in favor of **Sentience Now**. I stress that my focus here is on the question of whether some consideration arises which is decisive in its own right, in the manner of a basic intuition, rather than the question of whether there is a theory that some accept that tells against **Sentience Now**. Our question is whether anything prevents **Sentience Now** from being a live possibility: the fact that other conflicting theories are also live possibilities does not do so.

I take it for granted that there is no decisive refutation of the claim that AI systems will one day be conscious, and I have argued just above that there is no decisive refutation of the claim that creatures with limited repertoires, far short of general intelligence, are or may be conscious. I therefore take it that, in terms of Marr (1982)'s distinction between computational, algorithmic and implementational questions, there are no decisive challenges at the computational level to the prospect

that a current or near future AI system is conscious. It is a live possibility that there are computational tasks that are executed by current or near future AI systems, that do not require general intelligence or highly varied repertoires, that can lead to consciousness if executed in the right way. The question is whether something about the way these AI systems do it disqualifies them, either at the algorithmic or the implementational level.[8]

I will begin by looking at algorithms — the high-level instruction sets or recipes that specify how a computational system solves a problem. I will then look at implementations — the concrete processes of actually carrying out the instructions specified by algorithms.

Algorithms: Different Learning Styles

In this section, I'll talk about some of the more dramatic differences between the algorithms that leading Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformers implement and the analogous ones we do. These mainly have to do with learning style.

To begin with, supervised learners (including image classifying CNNs like AlexNet and ResNet, and language processing RNNs like Google's NMT) learn by being given a cheatsheet, and then going back and revising their work. In contrast, human infants must learn some things without the benefit of such explicit feedback.

One can also sharply distinguish the training phase of an artificial neural network

[8]I use Marr's distinction for convenience and don't mean to exploit its subtleties. For my purposes a tripartite distinction between abstract functional specification, software and hardware serves just as well.

from its 'trained' phase. The training phase involves repeated, computationally intensive applications of backpropagation (or a similar algorithm) to hone in on the network's optimal settings of weights and biases, while the trained phase, using such a network to make predictions, in simple classification cases merely involves computing a feedforward function with those optimized weights and biases. In contrast, in our own case learning and predicting seem to be inseparably linked, and it is hard to imagine how we could consciously do one without the other.

For another thing, most artificial neural networks require scores of examples to learn the rule, and have difficulty transferring what they have learned to related domains. In contrast, children can learn from few examples, and can transfer their knowledge easily.[9]

Artificial neural nets can also process many inputs at once, and given suitable hardware they can do so in parallel as quickly as they would process a single input.[10] For example, you might train an image classifier in batches of 100 or 1000 or 10000 images at a time. That's not how we do it.

—

In reply, first, we do learn things in a supervised way *sometimes* and we are no less conscious of those things. It is also worth noting that the divide between supervised and unsupervised learning may not be as deep as it first appears. Much of unsupervised learning consists of methods for systems to develop their own labels for data (e.g., autoencoders, generative adversarial models) and (in effect or in fact) training

[9]Marcus (2018).

[10]Typically, this is just the question of the dimensionality of certain matrices or tensors to be multiplied. But on a GPU or TPU matrix or tensor multiplication is carried out in parallel.

on those labels.

Also, many artificial neural networks are designed for online learning — that is, continuing to learn from examples while running online. And while there are some things that we humans can learn in a few shots, for others we need to put in our 10,000 hours. It took Lee Sedol far longer to acquire his expertise at Go than it took AlphaGo to learn to beat him, but his game is no less conscious because of it.[11] And the ability to transfer knowledge across a wide variety of domains cannot be a prerequisite for consciousness if bees are conscious, while at the same time systems like AlphaZero can transfer knowledge, e.g. from Go to Chess.[12]

Finally, don't forget that you, too, process lots of inputs at once. You probably can't process as many inputs of precisely the same type as an artificial neural network can, but why should an inability of yours count against the artificial neural network's chances of being conscious?

One point of phenomenology arises here. Many argue that consciousness is necessarily *unified*: if there is something it is like for you to experience A (at time t) and there is something it is like for you to experience B (at time t) then there is something it is like for you to experience A-and-B (at time t).[13] You might argue that nothing can possibly have *unified* experience of 1000 cat pictures at once, and so, if consciousness must be unified, whatever a CNN is doing with all of those cat pictures, it can't be experiencing them.

[11] Of course the computer needs to run through many more games than Sedol ever played. But this might not be a fair comparison, since we aren't tallying the games Sedol mentally simulates, as well as the many other things he had to learn to get to a Go board.

[12] Silver et. al. (2017b).

[13] E.g., Bayne and Chalmers (2003).

But this is too quick. First of all, of course you can experience 1000 cat pictures all at once if they are laid out in the right way, eg, adjacent to one another. Maybe that is what it looks like to the CNN fed 1000 input images. Alternatively, try to imagine one cat while you're looking at another. If you can do that, you're capable of co-experiencing two incompatible visual experiences. So why can't a CNN co-experience many?

But say you are dead certain that a CNN can't be co-conscious of 1000 cat images at once. One option then is to think of a CNN as constituting a collection of experiencers rather than just one. By analogy, many argue that split brain patients must be constituted by multiple experiencers insofar as they involve distinct centers of consciousness. Another option is to deny, as some philosophers do, that all consciousness is unified. Then the lesson may be that CNN architecture can yield novel forms of disunified conscious experience.

Algorithms: Can the Brain do Backpropagation?

Another question concerns the learning algorithm itself. The state of a neuron in a typical artificial neural networks is a real number given to eight or sixteen bits, while in contrast the state of a real neuron is either on (spiking) or off (not spiking). This suggests at a glance that real brains do not implement algorithms based on anything like gradient-descent (such as backpropagation, the algorithm most artificial neural networks use), because that kind of algorithm relies on differentiability, and a binary spike/no-spike activation function is non-differentiable.

However, there are various ways that the brain might approximate backpropaga-

tion.[14] Some of these have been explored in the context of artificial neural networks. A *spiking neural network* is an artificial neural network whose activation function is binary, and which represents continuous parameters via the temporal interval between spikes (and which can be implemented on 'neuromorphic' hardware like IBM's TrueNorth). There is evidence that mammalian brains also represent continuous parameters via temporal coding. For example, Di Lorenzo et. al. (2009) found that rats represent tastants (sweet, sour, etc) by means of a neural morse code in midbrain structures. Bohte et. al. (2002) develop an analogue of backpropagation suitable for a spiking neural network.[15] Scellier and Bengio (2016) explore equilibrium propagation, another artificial learning framework that may be implementable by brains.

I haven't conceded that the initial difference here is decisive: it isn't obvious that learning with backpropagation is somehow a handicap where consciousness is concerned. But even supposing somehow that it is, this is not decisive against **Sentience Now**, because alternative learning algorithms exist which allow current and near-future AIs to learn using methods that more closely parallel those used by the brain.

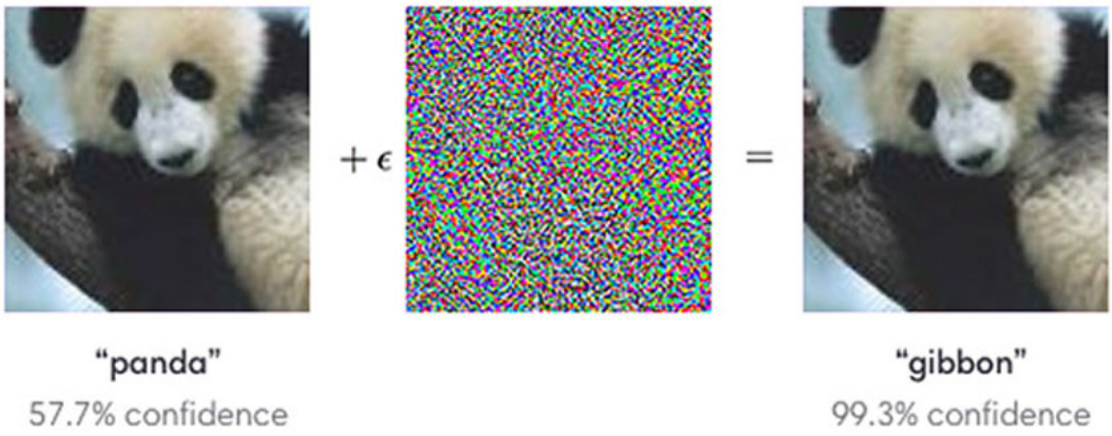
[14]For discussion watch Hinton https://www.youtube.com/watch?v=VIRCybGgHts

[15]At the time of writing, spiking neural networks cannot rival leading continuous-activation-valued artificial neural networks in image classification or natural processing tasks, and the search for an optimal learning algorithm for spiking neural networks is ongoing. For present purposes a proof of concept suffices.

Algorithms: Adversarial Examples

There is a peculiar kind of 'hack' to which artificial neural networks are susceptible: adversarial examples. Take a well-trained image classifier that is good at recognizing pandas, and take a specific picture of a panda that it recognizes. There is a way of adding a layer of visual 'noise' to obtain a new picture that looks just like the original to humans, but that fools the image classifier (fig.1).

Figure 1: Panda or Gibbon? (Goodfellow et. al. (2015)



It certainly seems bizarre that artificial neural networks can be fooled in this way. If you can't see that both pictures in fig. 1 are pictures of pandas, are you ever really seeing pandas? To aggravate matters, it turns out that there is nothing special about gibbons here: for any number of categories that a standard state of the art image classifier can classify, you can find a perturbation of a picture of a panda that will fool the classifier into thinking it is in that category.

In the words of Goodfellow et. al. (2015), "These results suggest that classifiers based on modern machine learning techniques, even those that obtain excellent

performance on the test set, are not learning the true underlying concepts that determine the correct output label. Instead, these algorithms have built a Potemkin village that works well on naturally occuring data, but is exposed as a fake when one visits points in space that do not have high probability in the data distribution."

Note that this is not simply a form of overfitting. The phenomenon is robust even against models that perform competitively on ImageNet competitions and against leading autonomous vehicle algorithms in deployment.[16]

But while humans might not be vulnerable to precisely this kind of illusion, we are certainly vulnerable to many others. Indeed, Elsayed et. al. (2018) show how to generate adversarial examples that humans misclassify (albeit under highly constrained conditions). In signal detection theoretic terms, there will always be cases of ambiguity between what to read as signal and what to read as noise, and in such cases a minor perturbation of the source may tip the threshold.[17] So let she who is immune to illusion cast the first stone.[18]

[16]Eykholt et. al. (2018)

[17]For a dramatic example consider the 'yanni/laurel' effect: an audio sample that sounds like 'laurel' to some and 'yanni' to others. You might clearly hear 'laurel' where I hear 'yanni', just as you might clearly see a panda where the CNN sees a gibbon. But it turns out to be a question of whether one's perceptual system foregrounds the higher or lower frequencies of the sample. See for demonstration https://www.nytimes.com/interactive/2018/05/16/upshot/audio-clip-yanny-laurel-debate.html.

[18]The discussion in Goodfellow et. al . (2015) continues : "This is particularly disappointing because a popular approach in computer vision is to use convolutional network features as a space where Euclidean distance approximates perceptual distance. This resemblance is clearly flawed if images that have an immeasurably small perceptual distance correspond to completely different classes in the network's representation." Perhaps the question here is whether the same may be said of human vision.

4 Implementation: Hardware Matters

Even if we grant that some contemporary artificial neural networks run algorithms that can engender consciousness, it is a further question whether they run them in the right way. After all, provided that these algorithms are expressed in a way that abstracts sufficiently from implementational details, they will be implementable by a Turing machine, and a Turing machine can be made out of Legos and punchcards. But intuitively, nothing made of Legos and punchcards can come to consciousness simply because of the sequence in which it punches the cards. Here, we consider several potential examples of dealbreakers that might show that the hardware of contemporary AIs cannot lead to consciousness even if it runs the right algorithms.

Implementation: The CPU Bottleneck

There is a theory of consciousness called the integrated information theory (IIT).[19] According to IIT, a system must be more causally interconnected than all subsumed, overlapping or nearby systems in order to be conscious. This implies that in general, systems running on standard CPUs or GPUs will not be conscious, because typical CPUs and GPUs will exhibit a lower level of causal interconnectivity than some of their (microscopic) constituents.[20]

But integrated information theory is a very contentious theory, not the least because it implies that if your CPU is not conscious, some of its constituents are.[21] I do not think it is established that IIT is a live option: it is theoretically elegant in

[19] Tononi et. al. (2016).
[20] Id.
[21] Id.

certain ways, but it is not clear that this rises to the level of reasonable evidential support, and it has implications that we may have decisive intuitions against, such as the implication that (given the right implementation) a CD player is conscious.[22]

Even if IIT were a live option, as I have already noted, we are looking here for deal-breakers: decisive intuitions or arguments that rule out that a target theory is a live option. It does not suffice for this that some other theory, conflicting with our target, is a live option. Finally, even if we take IIT's integrational constraint to be decisive, this still leaves open the prospect that suitably neuromorphic implementations like spiking neural networks are conscious.

Implementation: Cloudy Consciousness?

Most contemporary neural network algorithms are implemented in distributed fashion, spread out over a potentially vast array of servers "on the cloud". This makes for some potentially substantive differences between us and them. A cloud implementation of such an algorithm might span cities and in principle could span continents. But how could a conscious being span cities or continents?

It is worth setting aside two less compelling concerns from a more compelling one. You might think there is an absolute cap on how big a conscious being can be. But that is not very compelling. Indeed, we can imagine a future in which human cognitive function is outsourced to the cloud (and some would argue that it already is).[23] A slightly more tempting thought is that conscious beings have to experience the world from a *perspective* and it is hard to see how something as big

[22] Aaronson (2015).

[23] See e.g. Clark and Chalmers (2003).

(and faceless) as a decentralized cloud of servers could have a perspective. But we must be wary of homuncular thinking. Which part of your brain is *its* center? Where perspective matters in experience, it arises from the contents of our experiences or how we process them, not from the shapes of our heads.

A more serious issue is that it is hard to count the implementations of an algorithm,[24] all the more so when these are distributed. Consider Google's neural translation algorithm. Its computing power is backed by the cloud, but it interfaces with every personal computing device that uses it. Are there a legion of overlapping implementations here, or is there just one, doing a lot of multitasking?[25]

In some contexts one might be tempted to treat the question of how to count implementations as merely verbal. But where consciousness is concerned this is difficult. Consider again the question of the unity of consciousness discussed above. Presumably, if Google Translate is a single consciousness, it is disunified, but if it is many consciousnesses, each of them might be unified. This does not seem to be merely verbal, any more than it is a merely verbal issue whether a split brain patient is really a single disunified consciousness or multiple unified ones. And that issue has clear ethical and epistemological consequences, which almost certainly are not merely verbal.[26]

But this is not to say that the entire project is hopeless. The point is that because of the split brain case we already confront variants on this problem. So we might

[24]Bostrom (2006), Klein, Maudlin.

[25]Note there there are two multiplicative factors here: both the multiplicity owing to redundancy and that owing to different interfaces i.e. different total realizers.

[26]For related ethical and epistemological issues see Olson (2002), Unger (2004), Bostrom (2006), Briggs and Nolan (2015), Simon (2018).

draw on philosophical answers to the puzzles posed by split brain cases to inform our answers here. Must the consciousness of an AI indeed be unified? If so, then a single multitasker might not be suited for consciousness: instead, we might do better to think of its many (overlapping) interfaces with individual clients (e.g., personal computers) as the conscious beings.

Neither option is obviously untenable. In particular, there is no injunction against overlapping experiencers. Octopuses may feature a center of consciousness in each tentacle.[27] Many defend the view that split brain patients host multiple minds — or indeed, a fractional number.[28] And there is another puzzle about human consciousness, the problem of the many minds, suggesting that for every human mind there is a swarm of distinct but overlapping minds: consider the set of neurons that make up your mind, then subtract a neuron. Shouldn't that set also make up a mind? If so, we have a precedent for countenancing multiple overlapping experiencers here, and if not, then we probably have some way of avoiding that conclusion here.[29]

5 No Fact of the Matter?

There is one possible objection that merits further discussion. Thus far I have been assuming that there is a fact of the matter about whether current AIs are conscious: its just hard for us to figure out what it is. Sometimes, however, when we cannot answer a question, it is because there is no answer — i.e., no fact of the matter (in

[27]Godfrey-Smith (2016).

[28]Nagel (1971), cf Bostrom (2006).

[29]Unger (2004), Simon (2018) and others have argued that the choice in essence is between accepting multiple overlapping experiencers and appealing to some form of emergentism. But the relevant forms of emergentism could also be applied here.

philosophy jargon: it is *indeterminate*).

Might that be what is going on here? After all, doesn't consciousness come in degrees? Surely, some see with more intensity than others, consciousness fades away as one falls asleep, and we feel some pains more than others. So perhaps the answer is that current AIs are neither fully conscious, nor fully non-conscious — much as we might say that someone with thin, patchy hair is neither bald nor non-bald.

But grant that consciousness comes in degrees. Consider the volume on a radio. It comes in degrees of intensity, but it is either on or off. If its value is greater than zero, it is on, otherwise it is off.[30] There is no quick inference from 'degrees' to 'no fact of the matter'.

Also, there's a wrinkle called the problem of higher order indeterminacy.[31] Basically, you don't do yourself any favors by labelling something as indeterminate if you're trying to avoid classifying it, because now you have two problems: finding the boundary between non-C and indeterminately-C, and also finding the boundary between indeterminately-C and C.

Finally, usually when there is indeterminacy it is transparent to competent speakers of the language. You can just tell (I hope) that it is folly to argue at length about whether a 5' 11" European male is tall. This is because the rules of language don't settle it, and you grasp these rules well enough to tell. It is not as though you can perfectly well imagine such a man being short, and equally well imagine such a man

[30]This calls for qualification at the quantum level, but we can find more esoteric illustrations of the same point there. Consider, for example, the question of 'how present' a particle is in some region. This is measured by the amplidute of its wave function there, which comes in degrees. But provided that it has positive measure, the particle is there to some degree.

[31]Williamson (1994).

being tall, with these being two different ways the world might be for all you know. In contrast, there is a special *explanatory gap* surrounding consciousness. Given any physical system — say, an implementation of a neural network, you can perfectly well imagine it being conscious, and also imagine it not being conscious, with these being two different ways the world might be for all you know. This suggests at minimum that if there is indeterminacy here it is not the usual kind.[32]

I conclude that there is no decisive objection to **Sentience Now** from the prospect that the sentience of current and near future AIs is indeterminate. Indeed, in light of the foregoing, it looks more likely that there is no indeterminacy here, except for the epistemic kind.

6 So What?

If the foregoing is correct, it is a live possibility that some current or near-future implementations of artificial neural network algorithms are conscious. So what?

One route to practical consequences is to find a further principle which is both (a) a live possibility when taken in conjunction with **Sentience Now**, and (b) such that it entails that some or all of the AI systems that **Sentience Now** speaks of will be moral patients.

Note that, as with most any possibility operator, P and Q might both be live possibilities without P∧Q being one (for example, set Q = ¬P). The conjunction of two live possibilties is only itself a live possibility when the truth of one conjunct does not foster any new decisive defeaters for the other, or undercut the reasonable

[32]Simon (2012), (2017).

evidence for it.

To frame the issue, consider the following principle:

Sentience Suffices Any being that is phenomenally conscious is *ipso facto* a moral patient.

This principle certainly satisfies condition (b). What about condition (a)? It depends. Perhaps you take **Sentience Suffices** to be a live possibility, but only insofar as no beings without general intelligence are conscious. Maybe for this reason, the implication that a contemporary AI has moral status would be a dealbreaker for you. Then you might take **Sentience Now** to be a live possibility and take **Sentience Suffices** to be one, but deny that their conjunction is.

However, while I can't speak for you, I certainly don't have the intuition that general intelligence is necessary for moral patiency. If you can convince me that near-future AI undergo the kinds of states that we often call 'valenced': states such as pain and pleasure, yearning or satisfaction, you could probably convince me that they have moral patiency. This recommends a modification of **Sentience Suffices**:

Valenced Sentience Suffices Any being that is phenomenally conscious and experiences valenced states like pain and pleasure, yearning or satisfaction, is *ipso facto* a moral patient.

Now we have a principle that probably satisfies (a) for most of us, except perhaps for those of us who are antecedently convinced that nothing lacking general intelligence is a moral patient. But what about (b)?

This brings us to the titular question of this paper, a question I have thus far done nothing to address. I have argued that there may well be something it is like to be AlphaGo or AlphaStar but I have said little about what it is.

Let's focus on AlphaStar, who comes closer to meeting the criteria I outline in §2.2. So, given that there is something it is like to be AlphaStar, what is it like? Is it just a matter of experiencing in some visual or abstract way the structural layout of the game it plays, or does AlphaStar actually *want* to win, feel satisfaction when it does and frustration when it does not? If so, **Valenced Sentience Suffices** satisfies both a) and b), and **Precaution** entails that we should treat implementations of AlphaStar as moral patients, at least when it is easy to do so.

It is tempting to think of reinforcement learners as desiring reward and perhaps feeling satisfaction or pleasure upon receiving it, much as we do. But while it may be necessary for those feelings that some kind of description in reinforcement-theoretic terms is available, it is doubtful that it is sufficient. After all, supervised learning may be considered a special case of reinforcement learning, but it is not plausible that a numerical digit classifier really experiences desire to name the right numbers.

Clearly a sustained look at reinforcement learning and decision making architectures is required to address this matter. At the moment, I am undecided as to whether the possibility in question here enjoys the level of support needed to merit being considered as live (in conjunction with **Sentience Now** and **Valenced Sentience Suffices**). I hope to address this in future work.[33]

Suppose though that it is. What follows? Well, if every instance of AlphaStar is

[33]See also Muehlhauser (2017) and Bostrom and Shulman (forthcoming).

a moral patient, then it suddenly becomes important to settle how many instances there are. Above I mentioned that there are various puzzles about how to count implementations of algorithms, and how to count minds, and the normative upshot.[34] These become even more challenging here.

A special concern here is that we have a new, feasible path to creating utility monsters. Suppose we end up creating millions of new instances of AlphaStar, happily gaming away, in the process of training a new version (in distributed fashion, on the cloud). Now we seem to be under some obligation to leave them all running, and a challenge arises of how to say when this stops being 'easy', owing say to the energy costs. A related issue: are we now under some obligation to train systems in such a way that they enjoy more positive reward and less negative reward, as long as this is relatively easy?

Many further questions arise that require us to get fairly nuanced about how to understand **Precaution**'s 'easy all things considered' clause, and possibly to look beyond it to further principles to help us adjudicate. If P requires both Φ-ing and Ψ-ing, and each is easy on its own but together they become demanding, should we still do one, and if so which one? If P and Q are both live possibilities and P requires Φ-ing while Q requires Ψ-ing , but Φ-ing and Ψ-ing together are demanding (or impossible), ought we still do one, and if so which one? I hope to return to these questions in future work.[35]

[34]See again Bostrom (2006), Briggs and Nolan (2015), Johnston (2016) and Simon (2018).

[35]Thanks to David Chalmers, Martin Gibert, Dominic Martin, Jocelyn Maclure and the audience at the Université de Montréal for helpful feedback.

References

Aaronson, S. (2015) "Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)." Shtetl-Optimized. (Weblog) url: www.scottaaronson.com. Retrieved 2015-11-23.

Bayne, T. and Chalmers, D. (2003). "What is the unity of consciousness?" In Axel Cleeremans (ed.), *The Unity of Consciousness*. Oxford University Press.

Bohte, S., Kok, J. and La Poutre, H. (2002). "Error-backpropagation in temporally encoded networks of spiking neurons." *Neurocomputing* 48: 17-37.

Bostrom, N. (2006). "Quantity of Experience: Brain Duplication and Degrees of Consciousness." *Mind Mach* 16: 185-200.

Bostrom, N. and Shulman, C. (forthcoming). "Digital Minds".

Di Lorenzo,P., Leshchinskiy, S., Moroney, D., and Ozdoba, J. (2009) Making Time Count: Functional Evidence for Temporal Coding of Taste Sensation. Behav Neurosci. 2009 February ; 123(1): 14-25.

Elsayed, G.F., Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, Jascha Sohl-Dickstein (2018). "Adversarial Examples that Fool both Computer Vision and Time-Limited Humans." arXiv:1802.08195 [cs.LG].

Eykholt, K., Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. (2018). "Robust Physical-World Attacks on Deep Learning Visual Classification". arXiv:1707.08945v5 [cs.CR].

Godfrey-Smith, P. (2016). *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. Farrar, Straus and Giroux; New York.

Goodfellow, I., Shlens, J. & Szegedy, C.. (2015). "Explaining and Harvesting Adversarial Examples." *Proceedings of ICLR 2015.*

He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV].

Klein, C. and Barron, A.. (2016) "Insects have the capacity for subjective experience." *Animal Sentience* 2016.100.

Krizhevsky, A.,Sutskever, I., and Hinton, G. (2012). "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, 1097-1105.

Marcus, G. (2018). "Deep Learning: A Critical Appraisal.". arXiv:1801.00631 [cs.AI]

Marr, D. (1982). *Vision*. MIT Press. Cambridge.

Meuhlhauser, L. (2017). Open Philanthropy Report on Artificial Consciousness

Mnih, V., Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg & Demis Hassabis. (2015). "Human-level control through deep reinforcement learning." *Nature* vol. 518, pages 529-533.

Nagel, T. (1971). "Brain Bisection and the Unity of Consciousness." *Synthese* 22 (May):396-413.

Sebo, Jeff. (2018) "The Moral Problem of Other Minds." *Harvard Review of Philosophy*

Silver, David; Huang, Aja; Maddison, Chris J.; Guez, Arthur; Sifre, Laurent; Driessche, George van den; Schrittwieser, Julian; Antonoglou, Ioannis; Panneershelvam, Veda; Lanctot, Marc; Dieleman, Sander; Grewe, Dominik; Nham, John; Kalchbrenner, Nal; Sutskever, Ilya; Lillicrap, Timothy; Leach, Madeleine; Kavukcuoglu, Koray; Graepel, Thore; Hassabis, Demis (2016). "Mastering the game of Go with deep neural networks and tree search." Nature. 529 (7587): 484–489.

Silver, D., Schrittwieser, J., Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis. (2017a). "Mastering the Game of Go Without Human Knowledge." *Nature* Vol. 550, pages 354-359.

Silver, D.,Thomas Hubert, T., Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Si-

monyan, Demis Hassabis. (2017b). "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." arXiv:1712.01815 [cs.AI].

Simon, J. (2012). *The Sharp Contour of Consciousness*. Doctoral Dissertation: NYU.

Simon, J. (2017) "Vagueness and Zombies." *Philosophical Studies* 174(8): 2105-2123.

Simon, J. (2018). "The Hard Problem of the Many". *Philosophical Perspectives* 31:449-468.

Tononi, Giulio, Boly, Melanie, Massimini, Marcello & Koch, Christof. (2016) "Integrated Information Theory: from consciousness to its physical substrate." *Nature Reviews Neuroscience* 17(7): 450-461.

Williamson, Timothy (1994). *Vagueness*. Routledge.

Wu, Y., Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean. (2016) "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". arXiv:1609.08144 [cs.CL].