Crude, Cheesy, Second-Rate Consciousness

Joanna J. Bryson

University of Bath, Bath BA2 7AY, United Kingdom
email: j.j.bryson@bath.ac.uk

Abstract. If we aren't sure what consciousness is, how can we be sure we haven't already built it? In this article I speak from the perspective of someone who routinely builds small-scale machine intelligence. I begin by discussing the difficulty in finding the functional utility for a convincing analog of consciousness when considering the capabilities of modern computational systems. I then move to considering several animal models for consciousness, or at least for behaviours humans report as conscious. I use these to propose a clean and simple definition of consciousness. I use this definition to suggest which existing artificial intelligent systems we might call conscious. I then contrast my theory with related literature before concluding.

1 INTRODUCTION

> "If the best the roboticists can hope for is the creation of some crude, cheesy, second-rate, artificial consciousness, they still win." — Daniel Dennett (1994), *The Practical Requirements for Making a Conscious Robot*

While leading a group building a humanoid robot in the 1990s, Rodney Brooks complained about the term *robot brain* [5]. You can have or even sell a robot hand or arm or eye or even face, and they will be nothing like a human hand, arm, eye or face. But as soon as you say you refer to a robot brain, people say "That's not a brain." The aim of this article is to make you (the reader) look at some existing artificially-intelligent systems and say "You know, maybe that *is* robot consciousness."

From experience, I know this is hard to do. I've sat in a Cambridge, Massachusetts diner with other postdocs after a Dennett seminar, listening to them assert that science would solve consciousness, but not in their lifetimes — not in the next hundred years. Their justification for this statement was that we knew nothing about the topic. Even if we accept this statement as fact (which I don't), they conceded that in the previous ten years there we had come to understand many things well that we previously hadn't known about.

While trying to understand why my colleagues were certain we were so far away from a science of consciousness, I challenged them about how a computer could prove itself conscious. Almost anyone who owns a computer can make it type or even say "I am conscious." Dennett [20] implies that our own empathy should be used to judge the achievement of artificial consciousness. But teddy bears and pet rocks trigger empathy with no intelligence at all, while sadly human history is full of people mischaracterising other people as objects. We need a better criteria.

My colleagues the postdocs said that consciousness was a special sort of self-knowledge, being aware of what you are thinking. But computer programs have perfect access to all their internal states. If you set up a program correctly, you can ask it exactly what line of code — what instruction — it is executing at any time, and precisely what values are in its memory. This is in fact the job of software for debugging computer code, such as an Interactive Development Environment (IDE). IDEs are a very common type of computer program which are not generally considered even to be AI, let alone to be conscious [9].

If consciousness is just perfect memory and recall, then video recorders are conscious. If consciousness requires access to process as well as memory, computers have that access so they are conscious. But in this article I will not focus on phenomenological theories of consciousness. I will look instead at a recent functionalist theory from philosophy, and relate that theory to what is known about the impact of consciousness on expressed behaviour. From this I will propose a new version of the an old theory — that conscious experience correlates perfectly with a particular sort of search for appropriate action selection.

My theory is that consciousness is a limited-capacity system for learning about potential connections between context and action. We direct it primarily to situations that are uncertain and immediate. This allows us to optimise our use of this resource in building our expertise in our current environments. Cognition in general is computation constrained by being produced on demand, in real time. It is characterised by tradeoffs made by the fact action and computation take time, and that actions are expressed in a dynamic world. Consciousness is one part of cognition, a sequential search among possible candidate actions or interpretations, which lasts for a period determined by our own estimated uncertainty.

2 MULTIPLE DRAFTS AND CONCURRENCY

One well-known functionalist theory of consciousness is Dennett's multiple drafts theory. This theory starts from the observation that brains have many things going on in them simultaneously. Consciousness is a constrained narrative thread that runs between these things in some order, not necessarily a temporal one [21, 25]. In Dennett's more recent work, consciousness is described as a spotlight that shines on no more than one of these things at a time, at least it only shines brightly on one [24].

But why is the brain doing so many things at once? The reason is because if many processors run at the same time, more can get done quickly. In computer science, this is called *concurrency* [46]. In animals, each neuron is a processor, and at the same time collections of them can be organised into processing units. Many things can be done simultaneously without a problem — for example, all of the visual field can be monitored for motion in some area, or various limbs can be controlled in a rhythmic way to produce gaits.

Concurrency is a great strategy for problems that can be taken apart into pieces. But the 'hard problem' in concurrency comes when you need to combine all or even some of the answers you find back together again. This is called the problem of *coordination*. For an example, think of bees. A colony of bees can explore a large space around their hive to find flowers by having each bee fly in a random direction. They will explore even more space by using simple rules each bee can know, like “don’t fly near another bee”. But how much would it help the colony if only one bee finds a small patch of very productive flowers? When a bee communicates its discoveries by the waggle dance, a lot of other bees have to stop what they are doing to be involved, and one bee has to spend a *lot* of time and energy dancing [49]. When you consider not only the cost to the bees currently engaged in the communicative task, but also the complexity of this behaviour and the time it took to evolve, you realize the dance must represent a substantial adaptive advantage to the bee colony. Some individuals sacrifice time and energy, and the result is that on average each of these highly-related individuals has a better chance of finding food and bringing it home [30].

How does this relate to consciousness? Perhaps self awareness only seems a significant part of consciousness because there is a significant portion of the self of which we are *not* aware. One of the key attributes of consciousness is that it is a “bottleneck” or constraint — a limit that takes an otherwise uniform whole and makes some subpart of it special. In the bee case, that limiting process is the communication to others when a good source of food has been found by one bee — the recruitment of others to a single location[1]. This same sort of communicating role has also been suggested for consciousness — that it might be some kind of global workspace that all other parts of mind reference when there is important work to be done [1, 17, 44]. I will propose a markedly different and less powerful role in the following sections — that consciousness is just a process of learning about possible actions. However, because consciousness utilises specialised resources (such as close visual attention), it slows or arrests much other possible behaviour.

To conclude the present discussion of concurrency, I want to return the discussion briefly to consciousness-like elements in extant AI systems. Some approaches to artificial intelligence also have concurrent processes which normally operate more or less independently. In AI as in other disciplines such as Psychology or EvoDevo, this decomposition of the whole into some specialised subparts is called *modularity* [7, 10]. Just as in Psychology and EvoDevo, the utility of modularity in AI is that more complicated systems can be developed more simply and operate more quickly [29, 42].

The problem of coordination in AI is called *action selection* [7]. This problem emerges whenever multiple modules are contending for a single resource [3]. An example of a “resource” in this sense can be as simple as physical location. I cannot stand and give a talk at a meeting at the same time as I enjoy myself in a café, so if I want to do both I have to find some sequential ordering for my actions. Another such resource is speech — we can only say one word at a time, so words must be sequenced. And, critically for the Dennett [24] description of his attentional spotlight theory, memory. Apparently, episodic memory is a constrained resource, and only some of the things we are thinking about or perceiving will wind up in it.

[1] The foraging bee knows this is worth doing if there are relatively few other successful bees, which it can measure by the time taken for another bee to come store the food she has gathered [43].

So, this is the beginning of what a conscious AI system must look like. It must be modular or otherwise concurrent, and it must have some fixed resources which can only be allocated to one module or other unit of control at a time.

3 A FUNCTIONALIST HYPOTHESIS OF CONSCIOUSNESS

Dennett [24] currently claims that the only common characteristic of the conscious parts of our intelligence is “the historical property of having won a temporally local competition with sufficient decisiveness to linger long enough to enable recollection at some later time”. But the question of course is, competition for what? As Dennett points out in the same essay, one contestable resource in humans is *public expression*. If your current thoughts made it so far as to become verbalised, they are now a part of the public awareness. In this case the “local competition” is not only internal but also external — with other speakers. Further, the episodic memory is not only your own but also that of any other hearers, or readers. I return to this point below.

Most theory of mind, however, focuses on individual consciousness. Why should an individual be conscious of only one of their own actions at a time, if they are carrying out many? Perhaps the sequential phenomenological experience that we call “consciousness” emerged as a side effect of the system humans use for action selection — that is, for sequencing behaviour in those special situations where sequencing is needed. As I mentioned earlier, one fundamental sequencing problem is navigation — you can only be in one place at a time. The same neurological modules that in humans have been shown key to forming episodic memory have been shown in rats to be critical to learning new patterns of motion. This is the hippocampal system, and again I will discuss this point further below.

Of course, consciousness itself may be a scarce resource that can only be focussed on one task at a time, and thus forces sequencing. Norman and Shallice [34] propose that consciousness is a set of specialised extra resources which are brought to the problem of sequencing behaviour when the brain is either uncertain about the correct sequence — as when in a new context or when working on a new task — or when such sequencing is particularly important, for example when performing a delicate operation.

The Norman and Shallice theory is very similar to the one I propose here, except that in the present theory I emphasise uncertainty, not heightened control. Norman and Shallice (like many authors) are somewhat unspecific about what the “special resources” consciousness brings to such difficult situations might be. Here I make a specific proposal, although it won’t be fully justified until later in the article.

My proposal is that consciousness and episodic memory are the parts of a process for adaptable action selection. This process consists of:

1. fixing attention on an aspect of a behaviour context, and
2. allowing the brains natural priming mechanisms to search for potential actions that might be best suited to this context.

This sort of action selection is exceptional — most aspects of behaviour are predicted directly by their context and do not need such a process of search.

Because human behaviour is unusually plastic, we spend quite a lot of our time doing this sort of thing, even when the next action is not particularly difficult or pressing. Perhaps due to the tools and concepts provided by language and culture, we can even use consciousness to reason about abstract problems with no immediate sensory correlates. Thus we might think about our work when driving home when the road itself does not demand conscious attention.

Further, this process of conscious attention is also critical to episodic memory formation. Thus conscious attention not only allows real-time search, but records a "teachable moment" which the brain seems to use off-line to set future expectations [26, 32], and which are apparently available for future conscious recall.

3.1 Time and Search

The model I have just described of interacting attention and action I derived from a model developed by researchers in human vision. Wolfe et al. [50] primarily focus on demonstrating that when performing a new task, one doesn't learn from that performance when one can use vision rather than memory to guide the behaviour. But the model described above derives from an incidental model Wolfe et al. develop as part of their account. This model accounts for the difference in the time it takes to find some visual stimuli compared to others.

Studies that measure the time it takes to perform a task are called *reaction time* (RT) studies. In vision, if you have a field of dots where most are red but one is blue, you will find the blue one very quickly, and your RT will not depend on how many red dots there are. Similarly, if there are a number of **T**s on a screen and one **L**, you will not have trouble finding the one L, and you will find it quickly no matter how many **T**s there are. However, if the screen has many **T**s and many **L**s, and **T**s are both red and blue, but only one **L** is blue, it will take you a relatively long time to find the one blue **L**. Further, your RT will depend on the number of distracting objects there are — the more red **L**s or blue **T**s there are, the longer it will take you to find the blue **L**.

Treisman and Gelade [47] discovered this phenomena, and referred to the properties like colour and shape which required time invariant to the number of elements as *pop-out properties*. The fact that these searches can be achieved in constant time indicate that they are being performed concurrently — essentially, all the cells in the back of your eyes can be looking for the blue spot at the same time, and sets of cells in your visual cortex do the same for objects like **T**. But apparently identifying a conjunction of primitive pop-outs — e.g. that something is both blue *and* a **T** — cannot be done this way.

Wolfe and his colleagues proposed a relatively simple explanation for what happens in this case. The subject just randomly looks at items that have popped out with one of the traits, and checks if they also have the other trait. Eventually they will look at the right one[2]. So for example, you might just look at anything blue in the field (perhaps

[2] There is an older, less parsimonious account for this, where the visual system builds a "return inhibition map" once a potential target is recognised as inadequate. Wolfe et al point out this extra mechanism is unnecessary so long as the sampling among the candidates is sufficiently random.

returning multiple times to some objects) and eventually you will either see that one is also a **T** or give up. Thus the process of recognising and visually targeting blueness or **T**ness is not very conscious, but the process of finding a conjunction, saying "is that *both* blue *and* a **T**" apparently must be.

In order to support the definition of consciousness I have introduced, I now describe two more results from experimental psychology. I will then return to the question of conscious machines. Both of the experimental psychology examples concern something Dennett [24] describes as "imponderable" — consciousness in non-human species.

4 ANIMAL MODELS OF CONSCIOUSNESS

4.1 'Declarative' Memory in Rats

My first scientific interest in animal consciousness came when a colleague made passing reference to declarative memory in a rat. Whether or not rats are aware, I was quite certain they didn't declare anything, which is the definition I'd learned for that term. But there is reasonably good evidence that rats have explicit episodic memory. We know this from their behaviour, and from its analogues to human behaviour in similar situations. The humans we can ask about their conscious experience.

In this case, the person who was being asked was Henry Gustav Molaison, then known as patient HM. HM had both of his hippocampuses removed to treat his severe epilepsy, and as a result lost the ability to form new episodic memories. When I was a psychology undergraduate in the 1980s, we were taught that HM had lost the ability to *consolidate* short-term memories into long-term memories, but this theory proved false. At that time it was believed that when rats had their hippocampuses lesioned (destroyed) they could still consolidate their memory, but they had certain problems with navigation. Thus apparently hippocampuses were for navigation in rats but memory consolidation in humans. This turned out to be wrong — the right answer is both more parsimonious and more interesting.

What HM can't do is that he can't remember an episode after that episode finishes. He had enough working memory that he could be taught a task which he would perform successfully, but then if he was distracted (e.g. by the introduction of a new task), he would no longer remember being taught the first task, or even recognise the person who taught him. Yet although he had his surgery in the 1950s, HM started acquiring semantic knowledge 1960s culture, such as John F. Kennedy and rock music [18].

Eventually, experimenters stopped only asking HM what he remembered, and instead gave him the same sort of task the lesioned rats were successfully learning. HM was presented with an apparatus and instructed to turn push a button on it when the light turns on. When he did so he was rewarded with a penny. After several successful repetitions, the experimenters distracted him by asking him to count his pennies. After such a distraction, HM reported that he didn't recognise the apparatus or know what it was for. But when the light went on, he pushed the button, just as a rat would have. When they asked him why he did that, he said "I don't know" [16].

Given then that rats and humans were less different than once thought, let us return the question of rat episodic memories. One of the "navigational" tasks the rats had problems with was the radial arm maze — a maze with eight arms coming out from a centre. The trick with this maze is to remember which three arms the scientists put food in, and to go to each of them and not the others because you only have a little time in the maze. Also, you can't learn to go to the three arms in a particular order, because little doors slide up and down randomly, preventing access at irregular times. The rat thus has to remember which of the three arms you've already been down *today* to make sure to go down each of them once. When the rats had no hippocampuses, they could still learn which three arms had the food day after day, just like HM could tell you about the Beatles. But on any particular day, they didn't efficiently go down those three arms once each, like a normal rat would. Rather, they acted like they couldn't remember what they'd just been doing. Just like HM. This is what my colleague had referred to as "declarative memory". The ordinary rats (the ones that still had their hippocampuses) were showing they had it by going down the three arms each once [16].

4.2 Absent-Mindedness in Macaques

The previous example demonstrates that animals as far away from us phylogenetically as rats express at least some of the characteristics we characterise as conscious, and that they use these attributes for remembering things and determining their actions. Of course, rat awareness is probably quite different from primate awareness. In a controversial set of experiments, Rolls [40] found evidence that while rats occupy their hippocampuses primarily with information about their present location, macaque monkeys have more representations of the location they are *looking* at. Perhaps rats are *only* self-conscious, while a monkey can consider things at other locations.

The final experimental psychology study I present to justify my model of consciousness concerns attention, learning, and the effect of aging in macaques. Task learning is studied using standard tasks. One of the most widely studied tasks in both human and animal cognition is called *transitive inference*. The logical property of transitive inference (TI) is familiar from mathematics: if $A > B$ and $B > C$, then $A > C$. This property holds for some domains (e.g. real numbers) but not all — for example, primate dominance 'hierarchies' sometimes include cycles, where animal A can displace animal B, $B > C$, but $C > A$ [51]. Although originally seen as a good test of concrete reasoning [35], a series of experiments showed that pre-concrete-operational children, then monkeys, rats and even pigeons were capable of performing TI [51, for a review]. In fact, the $A > C$ inference seems oddly automatic *provided* that a subject can learn the two premise pairs $A > B, B > C$. However, it is very difficult to learn two contradictory premises involving a single stimuli such as B. Both animals and young children require a great deal of training to memorise the original, adjacent pairs. Without well-structured training, many of them fail; with such training, some still fail [14].

The TI experiment I am describing here again concerns reaction time. There are a number of characteristic effects that happen when animals (including humans) learn a long sequence of transitive pairs, such as: $A > B; B > C; C > D; D > E; E > F$. One characteristic is that the further apart two stimuli are from each other in that chain, the *faster* the animal is at making their choice. This is called the *Symbolic Distance Effect* (SDE) [6]. The SDE predicts for example that within a single subject, the reaction time for answering $B?E$ is on average shorter than that for answering $B?D$.

As described earlier, reaction times have traditionally been considered to be indicative of a cognitive process. Cognitive scientists have therefore worked to determine what computation animals might be utilising which both performs transitive inference and yet goes faster as a chain gets longer [6, 45]. But the theory of consciousness I presented above provides a different explanation. My theory predicts that the more uncertain animals are about their next action, the longer they hesitate. This allows their brain to search for a better, more certain solution, using a process like I described above for vision.

The motivation for abandoning the standard cognitive account of the SDE is simple: Rapp et al. [37] have shown that elderly rhesus macaques don't exhibit the SDE, but perform TI just as well as younger monkeys that do. The elderly macaques perform at the same speed no matter how far apart the pairs are, *and* that RT is considerably faster than the one for macaques that do show the SDE. Clearly then, the SDE does not reflect computing TI. We might have guessed this earlier, since in fact the SDE is only an aggregate effect that does not apply to every individual McGonigle and Chalmers [33]. But the Rapp et al. results are particularly striking.

If the SDE is not correlated to performance in transitive inference, then what is it for? Decision time is significant for group-living scramble feeders like rhesus macaques — slower monkeys will get less food and therefore be less fit [28]. So what benefit compensates for this costly loss of time? Fortunately, by chance, Rapp et al. found another difference between the elderly macaques and those exhibiting SDE. Due to an error in their experimental design, Rapp et al. started rewarding all their monkeys on the pair $B?D$ at chance at an intermediate level of testing. Most subjects consequently stopped preferring $B > D$ and rather went to chance on choosing between B and D when they were presented as a pair. But the elderly monkeys, the same ones that hadn't been hesitating, also didn't appear to notice the change in reward. They continued choosing B from that pair for the entire experimental procedure.

What does account for the SDE? This returns us to my theory of consciousness. I believe that subjects hesitate before action for a duration which is proportional to their uncertainty that the next action is correct. During that period, the subject searches for other clues, other possible solutions that would increase their certainty. In this learning-enhancing state, they are more likely to associate rewards with actions. Bryson [12] demonstrates that the uncertainty endemic in learning the initial pairs for TI can indeed account for the SDE in aggregate across a pool of subjects.

My theory implies that the older lab monkeys are more likely to go into "auto-pilot" mode on a simple lab task like TI. This could be adaptive for them, since if they'd lived that long in the wild they probably already know how to perform most tasks. Further, they might be losing scramble competitions since in general older animals are slower than younger ones. Thus further learning on established tasks may not be worth the time it takes for elderly monkeys. Of course, with macaques we cannot be certain that they are performing their transitive inference decisions without conscious awareness, because we can't ask them directly about their memory. Consequently, we need to find a way to extend this research into human subjects.

5 DO WE HAVE CONSCIOUS MACHINES YET?

For the purpose of this article though, I will now assume that my hypothesis is good enough to at least start exploring the the philosophical consequences. To return to the point in the first paragraph — is there anything already present in robotics that is as much like consciousness as robot hands are like human hands, or robot legs are like animal legs?

We do not require the full rich human pageantry of verbal narrative with qualia, meta-reasoning and self-concept. But just the "crude, cheesy, second-rate artificial consciousness" alluded to by Dennett [22, p. 137]. What I have proposed above (taken all together) is that consciousness has several criteria:

1. There must be multiple, concurrent candidate processes for conscious attention.
2. There must be some special process applied to a selected one of these processes.
3. This special process must achieve some function, probably concerning sequencing actions. And,
4. as a side effect, the object of this attention will normally be recorded in episodic memory, at least for a while.

Do any machines meet these criteria? I think probably yes. As pathetic as they are compared to humans or our science fiction, I think many of the humanoid robot systems which engage in dialog with human users and attempt to select objects from table tops can probably be thought of as meeting all these criteria in a crude, cheesy sort of way. Such robots are at MIT, Georgia Tech and the University of Birmingham, to name just a few [4, 27, 41].

If you think on a larger, Chinese-room sort of scale for a cognitive system, we might also see AI playing a part in other kinds of consciousness. For example, the Internet employs massive concurrency to create a world-wide database of useful information. If someone wants to act on a piece of that information, they employ a search engine to limit their view of all that data to say ten URLs with context on a single web-page. Under the definition of consciousness above, a page enters the consciousness of the system as a whole at the same time it enters the consciousness of the human being who is doing the final selection of the page to be viewed.

In other words, there are a hierarchy of conscious systems all exploiting a single agent's attention and action (the human.) The browser or search-engine on their own would *not* be conscious, because both require the human to do the actual sequencing. However, the human, the browser (e.g. Firefox) and the search engine (e.g. Google) all retain explicit memory of the selected Internet item and some summary details about the context of its selection, at least for some time. The browser will use this memory to suggest that page to the person again; the search company will use this memory to make it more likely this page is shown to other people who search, and the human will use the information for whatever they originally intended (or possibly something else). Thus a single action selection mechanism is used concurrently by three different sorts of cognitive systems.

Dennett [24] points out that modern humans have a sort of super-consciousness, what he calls *the publication competence*. That is, if a person comments on something they are aware of, then that object of their attention enters not only their own episodic memory, but also that of anyone else around them (at least, to the extent that they communicate successfully.) This is sort of a bigger, brighter spotlight that has even more probability of future recall and impact on actions, because now the process is public rather than private. Bryson [11] suggests this same public analogy for the attentional spotlight, arguing that well-known intellectuals help society find good ideas, serving as cultural-evolutionary selective pressure [c.f. 38]. Notice in my example of the previous paragraph, that the cognitive system that includes the AI search engine also has the publication competence. Once a search engine has determined something is salient, it will draw the attention of many others to that something. Thus search engines serve at least some of the same role as academics and journalists in contributing to our culture's steady increase in holding rich, useful knowledge [13].

6 WHAT THIS THEORY IS NOT

The theory I've presented here is entirely agnostic about qualia, self knowledge and so forth. If we assume that my definition of consciousness is the essential one, than it is a property that humans share with many other animals — mammals at least, and there is also a hypothesised equivalent to the hippocampus in birds. From this starting place, the self-localisation phenomena described by Lenggenhager et al. [31] for example could well be a consequence of the conscious search process I describe being applied over the sorts of information frequently used by modern humans as a part of our action selection. Note that I say "modern humans", by this I mean not so much anatomically- as culturally-modern. I believe that if there were a number of children raised by wolves, each would be relatively unlikely to be entirely self-conscious. The concept of self is something we have developed and / or evolved culturally, and that we spend a great deal of time communicating to each other. Not only children but even adults sometimes need to be reminded that others feel nearly exactly the same way they do when placed in the same situation.

My theory of consciousness is related to but not identical with the currently-popular Global Workspace Theory (GWT) [1, 44]. As I said earlier, while my theory does relate to some coordinated effort between brain systems, the same could be said of any mental process. I do not believe that *any* process in the brain is completely global, for simple reasons of combinatorics [8]. Processes like those described by Shanahan [44] could well determine the highest-level task- or goal-selection algorithms in autonomous systems, systems that in animals largely correlate to chemical / hormonal regulation systems such as emotions and drives [15, 39]. This is an important part of action selection, and also one that may be combinatorially accessible, if we assume that we have relatively fewer high-level goals compared to our complete array of possible actions. Goal selection is not the same as detailed, dextrous motion control. Much AI experimentation with spreading-activation systems of action selection has shown that these systems do not scale to any sort of complex action selection such as is displayed by mammals [19, 48].

This is not to say I dislike all or even most of the content of the current GWT as described by [2]. My theory covers a far smaller range of the conscious phenomena, but also an aspect which Baars does not concentrate on. The main purpose for consciousness to Baars is to integrate a large variety of information sources. The main purpose of consciousness for me is to allocate an appropriate amount of time to learning about and searching for the next action. These theories may be perfectly compatible. Baars' mechanisms could well be seen as the *how* of consciousness, and the *why does it feel like that?* Here my theory has focussed on primarily on the *when* and the *what is it really for?*

7 CONCLUSION

The basic goal of this article is to argue that there already is something we could call robot consciousness, at least to the same extent that there are already robot hands and robot legs. I make this argument not so much for sensationalism, but to try to get to the root question of what consciousness really is. This is interesting not only for abstract scientific reasons, but also because so many people identify consciousness as a critical attribute when they think about ethics.

Part of the reason we have trouble understanding consciousness is because the term has origins in folk-psychology and as such covers a large range of phenomena [23]. Some of these phenomena are in all probability not actually particularly related in any causal or mechanistic sense. What I have done here is concentrate on two criteria for consciousness Dennett [24] identifies:

1. that it is something that happens to one candidate process among many, and
2. that it creates a lasting impression in something like episodic memory.

From this I have proposed that consciousness is part of a particular process of action selection — one that is triggered by uncertainty and allows for the exploration and association of new actions in a particular context. This is in contrast to the majority of action selection, which is more-or-less reducible to stimulus-response, produced within a framework of simple automatic arbitration between high-level goals [36]. From my definition of consciousness, I have been able to argue that we can find evidence of consciousness not only in animals but also in *existing* AI systems.

None of my arguments are meant to belittle consciousness in any way, although they are intended to demystify it. I am not claiming consciousness is emergent or epiphenomenal, nor am I in anyway otherwise being anti-realist about it. Rather, consciousness is a central process to the part of intelligent behaviour I am most happy to call "cognitive".

Explaining how something works is by no means the same as explaining it away. Similarly, by disassociating consciousness from mystic ideas of soul I do not deny the central role of a concept of self in current human morality, nor the critical importance of moral behaviour to any social species. Even the crude, cheesy, second-rate artificial consciousness I have described are not I think belittled by that description — anything but. I think clarifying our concepts on cognition can help us appreciate the progress we have already made in AI as well as improve our approaches. Hopefully as we develop more informed perspectives on intelligence, we will begin building more useful — and more conscious — cognitive systems.

Acknowledgements

This article was originally commissioned for and presented to The Second Vienna Conference on Consciousness (September 2008) as a discussion of Daniel Dennett's contribution to the same meeting. Thanks to John Dittami for the invitation; to Wayne Christiansen and Alejandro Rosas Lopez for comments on an earlier draft; and to Dittami and Dennett for their comments at the meeting. This research was supported by a fellowship from The Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria. Thanks to the University of Bath for providing me the sabbatical leave to pursue the fellowship. Thanks also to the attendees of The Second AISB Symposium Computing and Philosophy, particularly to Mark Bishop, for their comments on a presentation of this material.

References

[1] Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA.

[2] Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. In Laureys, S., editor, *The Boundaries of Consciousness: Neurobiology and Neuropathology*, volume 150, chapter 4, pages 45–53. Elsevier.

[3] Blumberg, B. M. (1996). *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT. Media Laboratory, Learning and Common Sense Section.

[4] Breazeal, C., Berlin, M., Brooks, A., Gray, J., and Thomaz, A. L. (2006). Using perspective taking to learn from ambiguous demonstrations. *Robotics and Autonomous Systems*, 54(5):385–393.

[5] Brooks, R. A. and Stein, L. A. (1994). Building brains for bodies. *Autonomous Robots*, 1(1):7–25.

[6] Bryant, P. E. and Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, 232:456–458.

[7] Bryson, J. J. (2000). Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(2):165–190.

[8] Bryson, J. J. (2002). Language isn't quite *that* special. *Brain and Behavioral Sciences*, 25(6):679–680. commentary on Carruthers,"The Cognitive Functions of Language", same volume.

[9] Bryson, J. J. (2004). Consciousness is easy but learning is hard. *The Philosophers' Magazine*, (28):70–72.

[10] Bryson, J. J. (2005). Modular representations of cognitive phenomena in AI, psychology and neuroscience. In Davis, D. N., editor, *Visions of Mind: Architectures for Cognition and Affect*, pages 66–89. Idea Group.

[11] Bryson, J. J. (2006). The attentional spotlight. *Minds and Machines*, 16(1):21–28.

[12] Bryson, J. J. (2009). Age-related inhibition and learning effects: Evidence from transitive performance. In *The* 31^{st} *Annual Meeting of the Cognitive Science Society (CogSci 2009)*, pages 3040–3045, Amsterdam. Lawrence Erlbaum Associates.

[13] Bryson, J. J. (2010). Cultural ratcheting results primarily from semantic compression. In *Evolution of Language 8*, Utrecht. in press.

[14] Bryson, J. J. and Leong, J. C. S. (2007). Primate errors in transitive ‘inference’: A two-tier learning model. *Animal Cognition*, 10(1):1–15.

[15] Bryson, J. J. and Tanguy, E. A. R. (2010). Simplifying the design of human-like behaviour: Emotions as durative dynamic state for action selection. *International Journal of Synthetic Emotions*, 1(1):30–50.

[16] Carlson, N. R. (2000). *Physiology of Behavior*. Allyn and Bacon, Boston, seventh edition.

[17] Carruthers, P. (2003). The cognitive functions of language. *Brain and Behavioral Sciences*, 25(6):657–674.

[18] Corkin, S. (2002). What’s new with the amnesic patient H.M.? *Nature Reviews Neuroscience*, 3(2):153–160.

[19] Davelaar, E. J. (2007). Sequential retrieval and inhibition of parallel (re)activated representations: A neurocomputational comparison of competitive queuing and resampling models. *Adaptive Behavior*, 15(1):51—71.

[20] Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press, Cambridge, MA.

[21] Dennett, D. C. (1991). *Consciousness Explained*. Little Brown & Co., Boston.

[22] Dennett, D. C. (1994). The practical requirements for making a conscious robot. *Philosophical Transactions: Physical Sciences and Engineering*, 349(1689):133–146.

[23] Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition*, 79:221–237.

[24] Dennett, D. C. (2009). Can we really close the cartesian theater? Is there a homunculus in our brain? In Dittami, J., editor, *Proceedings of the Vienna Conferences on Consciousness*. University of Vienna Press. in press, available from the Web.

[25] Dennett, D. C. and Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Brain and Behavioral Sciences*, 15:183–247.

[26] Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., and Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences*, 104(18):7723.

[27] Hawes, N., Sloman, A., Wyatt, J., Zillich, M., Jacobsson, H., Kruijff, G.-J., Brenner, M., Berginc, G., and Skočaj, D. (2007). Towards an integrated robot with multiple cognitive functions. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, pages 1548–1553.

[28] Isbell, L. A. (1991). Contest and scramble competition: patterns of female aggression and ranging behavior among primates. *Behavioral Ecology*, 2(2):143–155.

[29] Kirschner, M. W., Gerhart, J. C., and Norton, J. (2006). *The Plausibility of Life*. Yale University Press, New Haven, CT.

[30] Leadbeater, E. and Chittka, L. (2007). Social learning in insects — from miniature brains to consensus building. *Current Biology*, 17(16):703–713.

[31] Lenggenhager, B., Tadi, T., Metzinger, T., and Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317(5841):1096–1099.

[32] McClelland, J. L., McNaughton, B. L., and O’Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

[33] McGonigle, B. O. and Chalmers, M. (1992). Monkeys are rational! *The Quarterly Journal of Experimental Psychology*, 45B(3):189–228.

[34] Norman, D. A. and Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In Davidson, R., Schwartz, G., and Shapiro, D., editors, *Consciousness and Self Regulation: Advances in Research and Theory*, volume 4, pages 1–18. Plenum, New York.

[35] Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books, NYC.

[36] Prescott, T. J. (2007). Forced moves or good tricks in design space? Landmarks in the evolution of neural mechanisms for action selection. *Adaptive Behavior*, 15(1):9–31.

[37] Rapp, P. R., Kansky, M. T., and Eichenbaum, H. (1996). Learning and memory for hierarchical relationships in the monkey: Effects of aging. *Behavioral Neuroscience*, 110(5):887–897.

[38] Richerson, P. J. and Boyd, R. (2005). *Not By Genes Alone: How Culture Transformed Human Evolution*. University Of Chicago Press.

[39] Rohlfshagen, P. and Bryson, J. J. (2008). Improved animal-like maintenance of homeostatic goals via flexible latching. In Samsonovich, A. V., editor, *Proceedings of the AAAI Fall Symposium on Biologically Inspired Cognitive Architectures*, pages 153–160, Arlington, VA. AAAI Press.

[40] Rolls, E. T. (1999). Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, 9:467–480.

[41] Roy, D. K. and Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

[42] Samuels, R. (2005). The complexity of cognition: Tractability arguments for massive modularity. In Carruthers, P., Laurence, S., and Stich, S., editors, *The Innate Mind: Structure and Contents*, pages 107–121. Oxford University Press.

[43] Schmickl, T. and Crailsheim, K. (2004). Costs of Environmental Fluctuations and Benefits of Dynamic Decentralized Foraging Decisions in Honey Bees. *Adaptive Behavior*, 12(3-4):263–277.

[44] Shanahan, M. P. (2005). Global access, embodiment, and the conscious subject. *Journal of Consciousness Studies*, 12(12):46–66.

[45] Shultz, T. R. and Vogel, A. (2004). A connectionist model of the development of transitivity. In *The 26^{th} Annual Meeting of the Cognitive Science Society (CogSci 2004)*, pages 1243–1248, Chicago. Lawrence Erlbaum Associates.

[46] Sipser, M. (2005). *Introduction to the Theory of Computation*. PWS, Thompson, Boston, MA, second edition.

[47] Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

[48] Tyrrell, T. (1994). An evaluation of Maes's bottom-up mechanism for behavior selection. *Adaptive Behavior*, 2(4):307–348.

[49] von Frisch, K. (1967). *The Dance Language and Orientation of Bees*. Harvard University Press, Cambridge, MA.

[50] Wolfe, J. M., Klempen, N., and Dahlen, K. (2000). Postattentive vision. *The Journal of Experimental Psychology: Human Perception and Performance*, 26(2):293–716.

[51] Wright, B. C. (2001). Reconceptualizing the transitive inference ability: A framework for existing and future research. *Developmental Review*, 21(4):375–422.