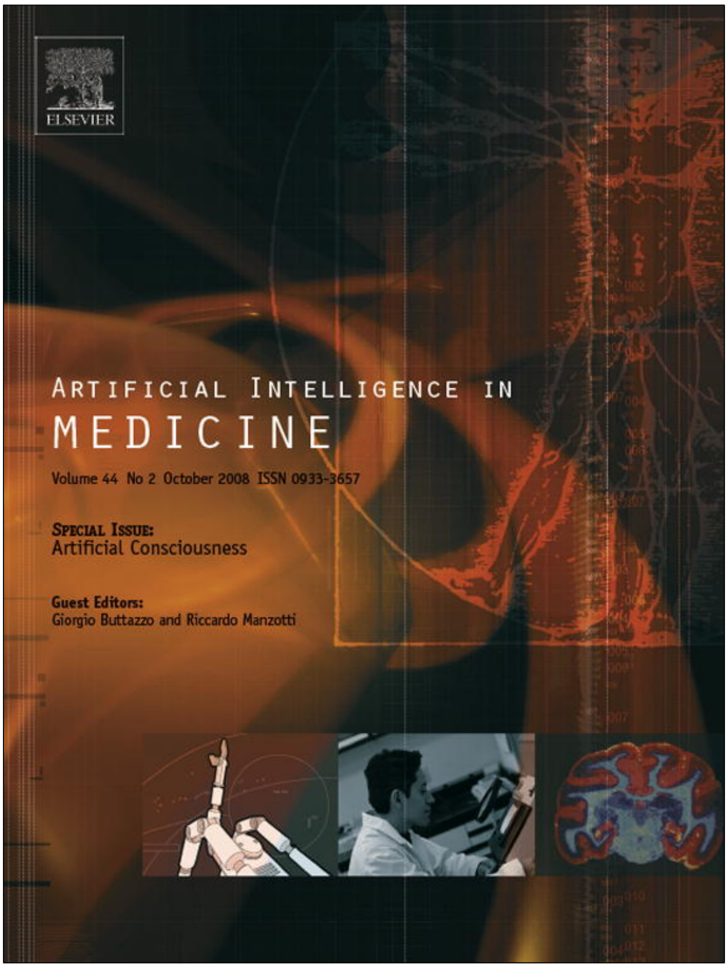
ELSEVIER
ARTIFICIAL INTELLIGENCE IN
MEDICINE
Volume 44 No 2 October 2008 ISSN 0933-3657
SPECIAL ISSUE:
Artificial Consciousness
Guest Editors:
Giorgio Buttazzo and Riccardo Manzotti

Artificial Intelligence in Medicine (2008) 44, 119—137

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Philosophical foundations of artificial consciousness

Ron Chrisley[*]

Centre for Research in Cognitive Science and Department of Informatics, University of Sussex, Brighton BN1 9QH, United Kingdom

Received 2 May 2008; received in revised form 28 July 2008; accepted 29 July 2008

KEYWORDS
Artificial consciousness;
Machine consciousness;
Prosthetic artificial intelligence;
Synthetic phenomenology;
Interactive empiricism;
Heterophenomenology

Summary

Objective: Consciousness is often thought to be that aspect of mind that is least amenable to being understood or replicated by artificial intelligence (AI). The first-personal, subjective, what-it-is-like-to-be-something nature of consciousness is thought to be untouchable by the computations, algorithms, processing and functions of AI method. Since AI is the most promising avenue toward artificial consciousness (AC), the conclusion many draw is that AC is even more doomed than AI supposedly is. The objective of this paper is to evaluate the soundness of this inference.
Methods: The results are achieved by means of conceptual analysis and argumentation.
Results and conclusions: It is shown that pessimism concerning the theoretical possibility of artificial consciousness is unfounded, based as it is on misunderstandings of AI, and a lack of awareness of the possible roles AI might play in accounting for or reproducing consciousness. This is done by making some foundational distinctions relevant to AC, and using them to show that some common reasons given for AC scepticism do not touch some of the (usually neglected) possibilities for AC, such as *prosthetic*, *discriminative*, *practically necessary*, and *lagom* (necessary-but-not-sufficient) AC. Along the way three strands of the author's work in AC — *interactive empiricism*, *synthetic phenomenology*, and *ontologically conservative heterophenomenology* — are used to illustrate and motivate the distinctions and the defences of AC they make possible.

1. Introduction

Consciousness is often thought to be that aspect of mind that is least amenable to being understood or replicated by artificial intelligence (AI). The first-personal, subjective, what-it-is-like-to-be-something nature of consciousness is thought to be untouchable by the computations, algorithms, processing and functions of AI method. Since AI is the most promising avenue toward artificial consciousness (AC), the conclusion many draw is that AC is

[*] Tel.: +44 1273 678581; fax: +44 1273 877873.
E-mail address: ronc@sussex.ac.uk.

0933-3657/$ — see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2008.07.011

even more doomed than AI supposedly is. In what follows I hope to show that this pessimism is unfounded, based as it is on misunderstandings of AI, and a lack of awareness of the possible roles AI might play in accounting for or reproducing consciousness. I aim to do this by making some foundational distinctions relevant to AC, and using them to show that some common reasons given for AC scepticism do not touch some of the (usually neglected) possibilities for AC, such as *prosthetic*, *discriminative*, *practically necessary*, and *lagom* (necessary-but-not-sufficient) AC. Along the way I will use three strands of my own work in AC — *interactive empiricism*, *synthetic phenomenology*, and *ontologically conservative heterophenomenology* — to illustrate and motivate the distinctions and the defences of AC they make possible.

Prima facie, it might seem easy to distinguish AI from AC: AI is the attempt to create artefacts that are intelligent,[1] and AC is the attempt to create artefacts that are conscious. But things are more complicated than that, for two reasons. First, consciousness and intelligence are not so clearly distinguishable. For example, in most cases where we would say that a task requires intelligence, we would also say that it requires consciousness. Second, the field of AI, in its broadest sense, is poorly served by the name "artificial intelligence". This term hides the fact that despite an early emphasis on problem solving, the field has always had more than just intelligence in its sights. That is, AI is the attempt to create artefacts that have mental properties, or exhibit characteristic aspects of systems that have such properties, and such properties include not just intelligence, but also those having to do with, e.g., perception, action, emotion, creativity, and consciousness. In this sense, AC is a sub-field of AI. But there is reason to believe that it is a sub-field that, because of the very nature of its topic, will have to use at least some methods that are substantial departures from those typically used in AI.

The combination of the preceding two points has the following upshot: In that most or all of mentality and intelligence involves consciousness, most or all of AI would seem to fall under the AC rubric: if AC is a subset of AI, it is a nearly exhaustive one. But even when concerned with mental capacities that usually involve consciousness in humans, AI typically proceeds in a way that downplays the distinctly phenomenological aspects of those capacities. True AC, on the other hand, is as much concerned with explaining phenomenological reports, and even the limitations of consciousness and attention, as it is with getting systems to perform cognitive tasks. AI typically focuses on those aspects of mentality that do not require one to confront the problems of consciousness head-on, while AC bravely confronts the consciousness-rich residue. Except that for AC researchers, consciousness is not as peripheral as the term "residue" suggests; it is central to mentality.

There is no firm consensus as to how to characterise the specifically conscious aspects of mentality that consequently distinguishes AC. An indicative example, however, can be based on [1] (see also [2]):

"[A] successful explanation of phenomenal consciousness... should (1) explain how phenomenally conscious states have a subjective dimension; how they have feel; why there is something which it is like to undergo them; (2) why the properties involved in phenomenal consciousness should seem to their subjects to be intrinsic and non-relationally individuated; (3) why the properties distinctive of phenomenal consciousness can seem to their subjects to be ineffable or indescribable; (4) why those properties can seem in some way private to their possessors; and (5) how it can seem to subjects that we have infallible (as opposed to merely privileged) knowledge of phenomenally conscious properties."

(See also Aleksander's "axioms" and Metzinger's "constraints", discussed by Clowes and Seth, this issue.)

In addition to these quite general *explananda*, AC researchers often aim to explain particular data, introspective or more traditionally behavioural, that have to do with what it is like to perform some cognitive achievement. For example, some AC researchers in robot navigation and planning (e.g., Tom Ziemke, Dan-Anders Jirenhed and Germund Hesslow; cf [3—5]) are as concerned with exploring the extent to which the processes the robot employs can be usefully viewed as instances of imagination and the existence of an "inner world", as they are in actually getting the robot to avoid collisions, find its way to a goal location, etc. Other AC research explicitly concerned with the phenomenological aspects of imaginative and counterfactual reasoning includes [6—11].

[1] Although the last half-century or so has seen the introduction of various new approaches to AI, including connectionism/neural networks, dynamical systems engineering, embodied/situated robotics, and artificial life, the term "artificial intelligence" is often used more narrowly, to refer to approaches that emphasize symbolic computation. Indeed, it was this particular approach that was dominant among those who first used the term "AI" to describe their work (as is well-known, John McCarthy coined the term in 1956), a situation that arguably continues to this day. To avoid confusion in what follows, the term "symbolic artificial intelligence" (or "symbolic AI") will be used to refer to this specific approach, and "artificial intelligence" (or "AI") to the general endeavour.

2. Varieties of artificial consciousness

To understand the full range of possibilities for AC, and to make clear the connections with, and differences from, AI, we can make some distinctions for AC that parallel distinctions made for AI. To highlight the fundamental issues at stake, however, the distinctions will not be the usual ones based on the various models, theories, and technologies used by different AI approaches (e.g., symbolic vs. connectionist vs. situated robotics vs. ...). Rather, the distinctions will be more generic, emphasizing differences in goals and aims, rather than technical means. How these distinctions inter-relate will be summarized in a table at the end of the Section.

2.1. Scientific vs. engineering

As with AI, a distinction can be made between two related, but distinct, goals in pursuing AC. *Engineering AC* is primarily concerned with creating artefacts that can do things that previously only naturally conscious agents could do; whether or not such artificial systems perform those functions or behaviours in the way that natural systems do is not considered a matter of primary importance. Of central (sole?) concern are the functional capabilities of the developed technology: what functional benefits can accrue from making a system behave more like a conscious organism? Whether or not the system so developed is *really* conscious is not an issue. Scientific *AC*, however, is primarily concerned with understanding the processes underlying consciousness, and the technologies provided by engineering AC, however impressive, are only considered of theoretical relevance to the extent that they resemble or otherwise illuminate the processes underlying consciousness.

2.2. Strong vs. weak (vs. *lagom*)

Within scientific AC, further distinctions can be made concerning the relation that is believed to hold between the technology involved in an AC system and consciousness. Adapting terminology from [12], *weak* AC is any approach that makes no claim of a relation (e.g., necessary and sufficient conditions) between the technology and consciousness. This would be a use of technology for understanding consciousness in a way similar to the use of computational simulations of hurricanes in meteorology: understanding can be facilitated by such simulations, but no one supposes that this is because hurricanes are themselves computational in any substantive sense. At the other extreme, *strong* AC is any approach whose ultimate goal is the design of systems that, when implemented, are thereby instantiations of (are sufficient for) consciousness. For example, a symbolic AI approach to strong AC maintains that an appropriately programmed computer actually is aware, is conscious, has experiences, etc.

Between these two extremes is a neglected zone of possibility that might be termed *lagom* AC. "*Lagom*" is a Swedish word with no direct English equivalent, which means something like "perfection through moderation" (and is thus reminiscent of Aristotle's golden mean). The *lagom* AC view, unlike weak AC, claims that the modelling relation holds as a result of deeper, explanatory properties being shared by the technology and conscious mental phenomena. However, unlike strong AC, *lagom* AC does not go on to claim that instantiating these common properties is alone sufficient for instantiating consciousness—something else might be required. It would be natural to say that while strong AC aims to discover sufficient conditions for consciousness, *lagom* AC only aims to discover (some of the) necessary conditions for it. But that's not quite right. It would be if there were only one way of being a conscious thing. But although that may be the case, it may instead be that there is more than one way to be conscious. If so, then what is important, even necessary, for being conscious in one of these ways might not be necessary for being conscious in another one of these ways. So *lagom* AC in general can succeed even if it only finds some of the necessary conditions for some way of being conscious. But let's put things in perspective: this kind of success would itself be a momentous achievement, and would contribute greatly to understanding how consciousness can be part of the natural world, even if it were not the whole story, and even if it were not part of the story of how specifically human consciousness is part of the natural world. That said, it should be stressed that there is a particular kind of *lagom* AC that does try to discover the necessary computational conditions for human consciousness, and it is this subspecies that is of most interest to the majority of the people who think about the issues being considered in this essay.

The point of raising the possibility of *lagom* AC is that many of the arguments against AC have the implicit form:

(1) AC is either strong or weak.
(2) Weak AC isn't interesting.
(3) Strong AC is false because of X (for different values of "X").
(4) Thus, no interesting form of AC is possible.

The acknowledgement of the possibility of *lagom* AC invalidates these arguments by denying the first premise. Of course, little would change if it turned out that *lagom* AC could be dismissed as uninteresting in the same way that weak AC can. This depends, obviously, on what is meant by ''interesting''. In the case of arguments concerning AC, it usually means the same thing as is meant in similar debates concerning AI: an AI (or AC) position is interesting in the relevant sense if its truth implies that the mind is computational. The claim that the mind is computational itself comes in varying strengths, but to keep things brief, the upshot can be put as follows. The claim that *some* computational concepts are *part* of the explanation of *some* mental phenomena, while much weaker (and thus harder to defeat) than the claims that opponents of AI usually argue against, is still interesting in the relevant sense: it implies that computation has a role to play in understanding the mind, and not just in the weak way that computational technology plays a role in understanding hurricanes. The analogous claim for an AI approach to AC would be: *some* computational concepts are *part* of the explanation of *some* conscious phenomena. To argue against this kind of AC, one has to show that *no* computational concepts are *ever* a useful *part* of explaining *any* kind of consciousness. Doing this is a tall order, and to my knowledge, no one has done so. For example, to refute this form of AC it is not enough to argue that instantiating the computational processes C employed in the AC explanation of some conscious phenomenon P are not sufficient for P (e.g., that there is an ''explanatory gap'' [13] between C and P). This form of AC already concedes that C might not be sufficient for P; it does not follow that C is not (partially) explanatory of P.

2.3. Constitutively necessary vs. practically necessary

In a sense, weak AI (and weak AC) has been given short shrift in philosophical discussions. The plausibility of *lagom* AC lies in it restricting its ambitions to necessary conditions for achieving artificial consciousness; weak AC may also aspire to necessary conditions for achieving artificial consciousness, albeit of a different sort. Even if they do not model the processes underlying human intelligence or consciousness, certain artificial intelligence technologies may be a practical requirement for achieving AC, be it of the engineering or the scientific variety. Unlike *lagom* AC, the necessity involved is not a constitutive matter of what properties the *artificial agent* must have for it to be conscious, but rather a practical matter of what tools and concepts *we* must have to be able to build it. These practical requirements are themselves of two sorts: causal and conceptual.

- Causal requirements have to do with the kinds of software, hardware, user interfaces, etc. that we will need to help us achieve AC. No doubt sophisticated, ''intelligent'', computational technology not yet invented will be needed to help us collect, mine and process the enormous quantities of data we can anticipate to acquire over the next decades (centuries?) concerning the operations of the brain and body that underlie consciousness. Similar advances in technological AI will also likely be needed to assist in the design of any system complex enough to be a candidate for an artificial consciousness.
- Conceptual requirements have to do with the kinds of systems we will need to have experience of building, and the kinds of learning/creativity/ perceptual/performance enhancing technologies we will need to develop, in order to get ourselves into the right conceptual/knowledge state for achieving AC (cf [14]).

It is likely that both sorts of practically necessary technologies will have an articulated trajectory. That is, there will likely be many technologies that are not themselves part of an AC we will build in some distant future, nor even part of the technologies we will need at that time to build that AC, but are part of a long chain of antecedent practical requirements for getting to that final stage.

This might seem very ''uninteresting'', in the sense discussed above; as practically necessary some forms of weak AC may be, by the definition of weak AC, their properties have no bearing on the nature of mind. In particular, even if such supporting technologies were thoroughly computational, this would not imply that consciousness itself is computational. Although this is so, the practical upshot, in terms of what technologies and concepts researchers (and thus students!) in AC, or perhaps even consciousness studies in general, need to have a mastery of, remains the same. It makes no difference whether the understanding, designing and building of computational systems turns out to be a constitutive or a practical requirement for advances in a science of consciousness (if it does); it would be a requirement nonetheless. Facility with non-constitutive weak AC may be as fundamental to achieving *lagom* or strong AC as skill at building telescopes and grinding lenses was to the development of modern astronomy.

2.4. Constitutive vs. discriminative

A distinction related to the points made concerning *lagom* scientific AC has to do with the form of the explanatory ambitions involved. The explanatory form that gets the most attention with respect to consciousness is the *constitutive form*: an account of what makes something conscious, as opposed to not being conscious. An alternative explanatory form is the *discriminative form*: for something that is already known (or presumed) to be conscious, what makes it the case that it is in *this* kind of conscious state as opposed to *that* kind? (cf. [10,11]).

Once one makes this distinction, it should be clear that a particular scientific AC research programme might contribute to discriminative explanations of consciousness, even if it does not contribute to a constitutive explanation of consciousness. Indeed, it may do so even if it is *impossible* for scientific AC in general to provide on its own, or even contribute to, a constitutive explanation of consciousness. This kind of discriminative AC is a kind of *lagom* scientific AC: it does not provide sufficient conditions for consciousness, but nonetheless contributes to explanations of consciousness by offering mechanisms that allow a subject to meet some necessary conditions for being in some kinds of experiential state. An example of such *lagom*, discriminative AC work will be given in Section 3.2.3.

2.5. Autonomous vs. prosthetic

Most AC research is *autonomous*, in that it aims to create a self-standing, individual artificial consciousness. Much less frequently discussed is the possibility of *prosthetic* AC: "artificial consciousness" as a phrase parallel in construction to "artificial leg" rather than "artificial light". Prosthetic AC would seek to extend, alter or augment already existing consciousness rather than create it anew.[2] It is a misunderstanding to think that creating or discovering more instances of a phenomenon to be explained only increases the problem; rather, the consideration of new instances is a way to increase the robustness of one's models and theories. Of course, we can generate novel experiences in relatively mundane ways: travel to a foreign location, listen to a new kind of music, taste an exotic dish. These are mundane in that they alter experience by altering the objects of experience (the environment). Despite being mundane, the method of systematic variation of the environment and observation of the concomitant changes in experience is an important part of consciousness science. However, prosthetic AC aims to make a distinct contribution. What is special about prosthetic AC is that it creates new experiences not by altering the objects of experience, but by altering the (transparent[3]) agent-based processes that enable experience. In this sense, it shares its aims, if not its methods, with a diverse set of activities: meditation, certain rituals, some psychoactive drug use, some forms of biomorphic art (e.g., the work of Stelarc), etc. (In fact, some AI-based AC practitioners find the phrase "artificial consciousness" so evocative of these alternative means of achieving altered states of consciousness that they prefer to use the term "machine consciousness" to describe what they themselves do.) Prosthetic AC, on the other hand, would more likely involve implants and complex sensory—motor interfaces than it would drugs or chanting. That said, scientific prosthetic AC would be distinct from the other means of altering consciousness mentioned above not just in the technology involved, but in that the goal of producing these alternative forms of experience would be subsidiary to the goal of understanding consciousness in a scientific way. The idea is that AC might contribute to our understanding of consciousness as much by systematically altering or extending it as by replicating it.

There are two sub-kinds of prosthetic AC: *conservative* and *radical*. The former seeks to create alternative material bases for extant kinds of experience; the latter seeks to create alternative material bases that result in novel kinds of experience. To elaborate, let us first consider radical prosthetic AC, as it is easier to see the value of its contributions. The controlled, systematic generation and observation of novel ways of producing familiar phenomena can allow one better to see the

[2] This parallels a similar distinction in AI, although prosthetic AI had more prominence in that field's early years than it does now. For example, Ross Ashby, a venerated pioneer in the field of dynamical AI, proposed a "design for an intelligence amplifier" [15].

[3] Even in the special case of auto-cerebroscopes and the like, the enabling processes are still transparent in the sense that although they happen to be the objects of consciousness in such cases, they are not so *qua* enabling processes; they are not so by virtue of being enabling processes. So a silicon visual implant would continue to be a kind of prosthetic AC even if it were fitted with a monitor that somehow let the owner of the implant to inspect the states of the implant, thus rendering it an object of experience. On the other hand, its status as prosthesis as characterised here *would* be in question if the use of said monitor were essential for the implant to provide its functionality. An aside: it is for this reason, as I argue in a paper entitled "Cognitive Provenance and the Extended Mind: Why Otto's Beliefs are not in his Notebook" (to appear in Gallagher, S. and Menary, R. (editors), *Cognition: Embodied, Embedded, Enactive, and Extended* (New York: Palgrave-Macmillan)), that the case of Otto and his notebook described in [16] is not a case of prosthesis, is not a case of Otto's mind extending into his notebook; unlike the cases under consideration here, Otto can only use the notebook for information retrieval by virtue of it being an object of perception for him, at least orginally.

underlying structure of how said phenomena are produced in the normal case. Radical prosthetic AC, however, seeks to confer on the subject experiences of a fundamentally different kind than the subject had enjoyed before, such as a new sense modality, or that of using a new limb. One need not look far to find examples of such prostheses. On some accounts (e.g. [17]) sensory substitutions devices (e.g., TVSS [18]) do not, per their name and conventional wisdom, substitute experiences in one sense modality for experiences in another (e.g., vision for touch); rather, they confer on the subject *sui generis*, novel forms of sensory experience. If this is right, such devices make possible contributions that go well beyond passive collection of new data on extant modalities, by allowing active exploration of the space of possible conscious experience via experimental intervention. Such active variation and consideration of new kinds of phenomena is the heart of the experimental scientific method.

Conservative prosthetic AC seeks to reproduce in us experiences similar to those that we already enjoy, just with a different material base. Technically, hearing aids, glasses, a blind person's cane, or even a hammer would count as (simplistic) forms of conservative prosthetic AC. And consideration of such prostheses can and has informed more than one theory of perceptual experience (e.g. [16,19] p 67; [20] p 143; [21]). But more interesting possibilities include those offered by the aforementioned sensory substitution devices, even on the conservative assumption that such devices merely allow one standard sensory modality to be substituted for another (e.g., vision for touch, in the case of TVSS).

A digression into the philosophy of science will help illustrate the value of conservative prosthetic AC. Often in science we perform experimental investigations in terms of a set of independent variables (''input''), over which we have direct control; and dependent, observational variables (''output''), over which we only have indirect control, via the independent variables. A standard way of increasing one's understanding of a phenomenon is to fix most of the independent variables, and manipulate the remainder of them in such a way as to create a large variation in the dependent variables. We could call this ''depth-first exploration in input space''. A complementary way to explore the relation between the independent and dependent variables is ''depth-first exploration in output space''. That is, take an observed (dependent variable) phenomenon O that is produced by all of the independent variables being in a particular state S. Put the independent variables into a state S' that differs from S in the value of, say, half of all constituent variables. Keeping these differing values fixed, then attempt to manipulate the rest of the variables into a state S″ such that S″ produces O. In short, if you can get a balsa wood wing to do roughly the same thing as a paper wing, then perhaps you have learned something about aerodynamics that you would not easily have learned just by generating different kinds of paper wing. So also, then, for prosthetic AC: in getting the same experiences to be realised in different hardware, one might acquire insights not easily gained by trying to generate radically different experiences prosthetically.

Perhaps most exciting is that prosthetic AC could be truly first-personal; if the theorist is also the person whose experiential states are being technologically modified, then some fundamental worries about the ability of AC to account for the subjective aspects of experience can be finessed. In particular, prosthetic AC seems well suited to be a kind of discriminative AC, as just discussed in Section 2.4. In this sense, prosthetic AC has similarities with synthetic phenomenology (Section 3.2.3).

2.6. Human (or biological) vs. general

An important distinction within AC has to do with whether one is attempting to reproduce/explain human (or biological) consciousness in particular, or whether one is attempting to reproduce/explain consciousness more generally. The quest for generality can be either more ambitious or more modest. For example, in the case of constitutive scientific AC, generality would be more ambitious if one were attempting to explain, for any possible conscious agent, why that agent is conscious. A more modest form of generality would be merely (!) attempting to explain how it is that some particular non-human (non-biological) physical thing is also a conscious thing.

Even if one's ultimate goal is to explain human consciousness, the thought goes, understanding how some particular artificial system is capable of being conscious might assist completing that task. As Fodor states [22], ''Nobody has the slightest idea how anything material could be conscious. Nobody even knows what it would be *like* to have the slightest idea about how anything material could be conscious.'' An understood, non-human AC would give us much more than a slight idea of how something material could be conscious. In fact (some would say), it might even be a mistake to spend much time investigating the specifics of the human case (e.g., human neurophysiology) at this current, inchoate stage of consciousness science. Doing so would be like trying, e.g., to understand flight by looking at bird's feathers under a microscope. Instead, the analogy (cf. [23—26]) continues, we only came to understand natural flight by achieving artificial flight, and we only succeeded in doing that once we stopped trying to slavishly copy the superficial characteristics of biological flyers, and instead sought to create artefacts that allowed systematic isolation and identification of the relevant aerodynamic forces and factors. As for flight, so also for consciousness, according to this view.

Similar points can be made for discriminative accounts of consciousness. For an artificial system that is believed to be conscious, even if one could not explain why it is so, a discriminative account of its various experiential states would be of value, even if one did not thereby have a discriminative account that applied to the human case.

Although the above discussion has focussed on the autonomous case, the distinction also applies to prosthetic scientific AC: one may ambitiously believe one is investigating consciousness in general when one does such work, or one may take such work to only reveal insights into how biological or human (or one's own!) conscious states depend on the material world.

2.7. Putting it all together

A graphical representation of the relations between the varieties of AC just described is presented in Table 1.

The main, crosscutting distinctions are scientific vs. engineering AC, and autonomous vs. prosthetic AC. Both main kinds of scientific AC have further sub-kinds, depending on the strength of the relation that is meant to hold between the AC technology and consciousness: sufficiency (strong), constitutive necessity (*lagom*), practical necessity or no relation (weak). However, *lagom* AC is not a distinguishable sub-kind of prosthetic AC, which renders the sufficiency/necessity distinction moot. Similarly, while strong and *lagom* autonomous AC can be further distinguished by whether they aim to explain what it is to be conscious (constitutive) or whether they aim to explain which experiential state an agent assumed to be conscious is in (discriminative), strong prosthetic AC, in that it presupposes the consciousness being augmented, may only play a discriminative explanatory role.

Table 1 Varieties of artificial consciousness research goals

	Scientific	Engineering
Autonomous		
Strong	Constitutive	Discriminative
Lagom	Constitutive	Discriminative
Weak	Practically necessary (Causally or Conceptually)	Non-necessary
Prosthetic (Conservative or Radical)		
Strong	Discriminative	
Weak	Practically necessary (Causally or Conceptually)	Non-necessary

Not shown: the "Human vs. General" distinction.

Obviously, trying to express such complex relations in a two-dimensional diagram will introduce distortions and hide inter-relations. For example, the diagram makes many of the distinctions seem dichotomous, whereas in reality the very same research can be, e.g., both scientific and engineering AC. An example of some inter-relations that are not depicted concerns weak prosthetic AC. Such work may be practically necessary in the construction of autonomous, strong AC. Or it may be that autonomous, strong AC is impossible, but weak prosthetic AC may be practically necessary for producing a non-AC explanation of consciousness. Yet these possibilities are not displayed in the table. Despite these limitations, it is hoped that the figure makes clear some gross features of the distinctions made in this section. These distinctions will be used in the following sections to discuss the motivations, difficulties, and prospects for different kinds of AC research.

3. Motivations for AC

Why might one think that AC is achievable? The variety of AC research goals documented in the previous section suggests that there will be no single answer to this question. This section, therefore, details specific motivations for engineering AC, general scientific AC, and two specific kinds of scientific AC (strong autonomous AC and weak conceptually necessary AC).

3.1. Motivations for engineering autonomous AC

The plausibility of engineering prosthetic AC lies in the fact that we have already succeeded in making some conservative achievements. This makes further conservative prostheses, and more radical ones, appear achievable. Motivations for scientific prosthetic AC have already been given in Section 2.5. The plausibility of engineering autonomous AC cannot derive from extant success in such a direct way. One general motivation, then, comes from a belief in the universality of natural laws. Whatever laws are responsible for the natural occurrence of some phenomenon must in principle be exploitable by humans to make an artificial occurrence of that phenomenon.[4] Conscious humans, are, after all, physical things: if one kind of physical thing can be conscious, why not another? If we can produce new conscious beings through childbirth, why not through some other means that affords us more control over the outcome? That the resource cost (e.g., materials, time and energy) of doing so may be practically prohibitive will, however, make the more cautious engineer seek firmer foundations.

More rigorous reasons for believing that specifically computation-based engineering AC is possible lie in results from computability theory that predate AI and computer technology itself. Turing's introduction of a set of formal models of digital, algorithmic computation is the key notion here. Now called Turing machines, these automata are given a symbol string (usually interpreted as an integer) as an input, and produce a symbol string (also usually interpreted as an integer) as the output for that input. In this way such machines can be understood to be computing functions over the integers. Turing proved that there exist Universal Turing machines that can simulate the action of any other Turing machine. This, coupled with the assumption that if any function can be computed at all, it can be computed by a Turing machine (an assumption roughly equivalent to the ''Church-Turing Thesis''), yields the result that a Universal machine can, in a sense, compute anything that is computable at all. In particular, many take this to establish that a Universal machine can compute any function a human can compute. If one then adds the assumption that human behaviour, or at least the mental processes that give rise to it, can be conceptualized as mathematical functions of the relevant sort, then it follows that any Universal machine can, in principle, be programmed to exhibit behaviour functionally equivalent to that of any human being—including behaviour we would normally say requires consciousness.

These mathematical considerations have apparently borne fruit in everyday life, resulting in a general technological optimism concerning the assimilation of the mental to the computational. More and more behaviours that were thought to require mind — planning, proving theorems, playing chess, diagnosing blood diseases, handwriting recognition, speech recognition, etc. — have been replicated by computers. So why not all of mentality? Why not consciousness, or at least the behaviours normally thought to require it? For it is only the latter with which engineering AC is explicitly concerned.

3.2. Motivations for scientific AC

In general, motivating scientific AC is twice as hard as motivating engineering AC. First, in a way similar to the case of engineering AC, one has to motivate the possibility of the technology being able to implement some competence associated with or indicative of consciousness. But beyond that, one has to give some reason to believe that this will illuminate consciousness in some way; that such replication can also serve as an explanation. Before looking at specific motivations for specific kinds of scientific AC, it will be useful to see how AC-based explanation is meant to proceed.

A central component of AC-based explanation, like that of AI-based explanation in general, is that of computational modelling. Modelling, in particular computational modelling, confers several benefits on scientific investigation [28], including allowing one to:

(1) State theories precisely;
(2) Test theories rigorously;
(3) Encounter unforeseen consequences of one's theories;
(4) Construct detailed causal explanations of actual events.

Thus even at its weakest, AI methodology can offer these benefits to the understanding of mentality (compare Clowes and Seth, this issue). Moving from mere (weak) simulation toward strong AI/AC presumably multiplies these benefits, especially 2 and 3 (cf. [29]). Some might wonder how 2 and 3 can apply, particularly in the case of autonomous AC. Given the subjective nature of consciousness, how could one test a theory of it in an AC? What would count as confirmation (or falsification)? This is a common, general worry about AC: ''how would you ever know if you succeeded?'' Although the overall structure of this essay dictates that this point should be addressed in Section 4, responding to it brings up some points about the nature of AC explanation that are best mentioned here.

The first thing to note is that the question is only a problem for some kinds of AC. Pragmatically minded engineering AC will not care whether a duck is

[4] Although this belief is, in modern times, usually accompanied with a belief in materialism (the denial of a separate mental world or substance), I have elsewhere questioned whether materialism *per se* makes AI or AC any more plausible than it is under dualism ([27], vol. 1, p 12).

''really'' conscious or not as long as it looks, swims and quacks like a conscious duck. Neither is the question a problem for prosthetic scientific AC, since the recipient of the experiential prosthesis will be in a position to confirm its success (or failure).

Autonomous scientific AC, on the other hand, does have to face the question. But so does any third-personal naturalistic explanation of consciousness. Any such account faces the problem of confirming or falsifying the underlying theory's claims concerning the presence (constitutive theories) or kind (discriminative theories) of consciousness associated with a given physical system. So this problem cannot serve as a critique of AC in particular.

But more can be said in defence of autonomous scientific AC against this criticism, which usually assumes a simplistic vision of how an AI explanation of consciousness might proceed. In fact, there are at least three ways AI methodology can be applied to explaining consciousness such that the question ''how would you know if you succeeded?'' can be answered:

- One can attempt to model the physical system underlying consciousness, at various levels of abstraction (e.g., a connectionist model is pitched at a lower, more hardware-dependent level of abstraction than is a typical symbolic model).
- One can attempt to model conscious processes directly, e.g. by using introspection to note their causal structure, and them implementing this structure in an AI system (usually symbolic).
- One can attempt to model the behaviour of a system known or believed to be conscious, without any direct knowledge of the underlying physical or phenomenological structure, in the hope that reproducing both actual and counterfactual behaviour is sufficient to ensure that the same consciousness-producing causal structure is thereby implemented.

These models can then be used to predict and explain phenomena in the usual way.

Some will still insist that for any AI system that is supposedly conscious, one can always imagine it being built and behaving the same way and yet not being conscious. So what exactly has been explained? In Section 3.2.2 it will be argued that such ''modal intuitions'' may be the result of an unhelpful, faulty conception of consciousness that engaging in AC itself may help to repair. In the meantime, one constraint can be noted: in arguing that we would not ever be able to know that an AI system is conscious, one should be careful not to set the epistemological bar so high that one calls into question one's knowledge that other humans are conscious.

3.2.1. Motivating strong autonomous AC

One could transform the computability-based motivations for engineering AC into motivations for scientific AC with some behaviourist assumptions about the nature of mind. For example, if one assumes that all there is to mind is a capacity to produce certain forms of behaviour, then explaining how a system has a capacity for such behaviour would be an explanation of why it is conscious. But, despite distractions such as the so-called ''Turing Test'', most AI (and AC) researchers are not tempted by behaviourism; quite the contrary. AC has its roots in AI, which (at least the symbolic variety) in turn has its roots in anti-behaviourist cognitivism. This is the view that cognition (more narrowly: thinking; more broadly: all mentality) actually is (or at least involves; cf. the discussion of *lagom* AC in Section 2.2) a kind of (symbolic) computation. A primary motivation for this view is a belief in the multiple-realisability of mental states, itself motivated by, e.g., thought experiments concerning creatures (''Martians'') that behave just like humans, but have a very different physiology. Since it would be politically incorrect in the extreme to deny mental states to these Martians, it must be a mistake to think that mental states can only be implemented in biological states such as those of humans and animals on Earth. The question then arises: what *do* Martians and Earthlings have in common by virtue of which they both enjoy mental lives? The cognitivist answer is not behaviour, but abstract causal organisation, which is describable using computational formalisms such as Turing machines. The belief that it is this abstract causal organisation that constitutes or individuates mental states is called *functionalism*. A strong version of functionalism would maintain that instantiating a particular causal organisation is sufficient for instantiating that mental state, while a *lagom* version would only make it a necessary condition. It follows from strong functionalism that mentality, especially thinking, is at root a formal activity: unlike, say, a rainstorm, if you computationally simulate thinking, you actually recreate it. The upshot of this is that, according to either version of functionalism, showing how a physical system instantiates the causal organisation characteristic of a conscious state would count as an explanation, if only a partial one, of how it is that the physical system is in that conscious state.

A compelling reason for believing that this form of explanation is possible for at least some mental states is that it is the way that we have understood complex computational systems for more than half a century. We find it natural to explain the behaviour of computers in terms of such mental-like concepts as *representation*, *memory, information*, and *recall*; even *learns*, *believes*, *knows*, and *wants*. What's more, the behaviour of computers is ''reliably and voluminously'' predictable and explicable in terms of these concepts ([30], p 15), in stark contrast with the limited payoff one gets in animistic attributions of mentality to the sun, rivers, thermostats or one's car. But the important point is that we do not find the fact that computers are explicable in these terms mysterious; there is no mind—body problem for computers. Someone armed with the proper computational concepts can understand why a machine built this way can also be an agent usefully interpretable as knowing the rules of chess and wanting to win. The suggestion for functionalist AC is that if we knew how to apply the same or similar concepts to a conscious system, we would thereby de-mystify its phenomenality.

This functionalist approach assumes that one can, in practice, provide an analysis of mental states in terms of their abstract causal structure. For the case of thinking, this assumption, at least to those cognitivistically inclined, is not problematic. It seems reasonable to assume that the steps in an episode of thought are accessible to the thinker; thus, it is in general possible for a thinker to write those steps down in a finite list. The Church-Turing thesis then implies that this algorithm can be turned into a set of instructions for a computer, and thereby create a system that reproduces, or at least models, that particular episode of thinking. The success of a translation of this account from thinking to consciousness is mixed. On the one hand, the assumption that one is aware of one's own conscious states is, if anything, more secure than the equivalent assumption about thought. But even some of those who would agree that the essence of thinking is its introspectable causal structure will baulk at the same suggestion with regard to consciousness. In any case, the truth of functionalist accounts of consciousness only requires that there be a characteristic causal structure, not that it be introspectable.

3.2.2. Motivating weak conceptually practically necessary AC: Interactive empiricism[5]

Not all limitations on our understanding, even scientific understanding, of the world are merely a matter of not having enough data. In particular, our problems in trying to understand consciousness are conceptual; the obstacles we face in understanding what it is for a physical thing to also be an experiencing thing are not just a matter of not having enough knowledge of the nervous system. Even if we knew much more about the nervous system than we do now, we would still have some fundamental puzzling questions. We have a naturalist intuition that consciousness, like anything else, is, in some sense, physical at root. On the other hand, we have another intuition — Dennett's ''zombic hunch'' [31] — that it is possible for there to be a creature — a zombie — that is physically just like us, but for which ''there's no one home'': it is not conscious. At least some people believe that it is not inconceivable that there could be something that is physically (and thus behaviourally) identical to you and yet different from you with respect to its experiential properties, even to the point of not having any experiences at all: a zombie-you. Those two intuitions are in direct conflict: the naturalist intuition is that if you fix the physical you fix everything else, whereas the zombic hunch is that even if you fix in the physical you still have not fixed the experiential. The presence of both of these intuitions produces an unsatisfactory cognitive dissonance.

One way of responding to this is to diagnose the cognitive dissonance as the result of a flaw in our concept of consciousness. If our concept of consciousness has paradoxical implications, perhaps we should try to develop a new concept of consciousness that does not. Perhaps we should look to conceptual change as a way to resolve this problem of the conflict between our naturalistic inclination and the zombic hunch. This suggestion is at once both facile and mysterious. It seems like a content-free response one can make to any situation, and yet it also prompts the question: how could we make such a conceptual change? One constraint is this: we do not want to change the subject. We want to change our concept of consciousness in a way that ensures that in employing the new concept, it is consciousness that we are still talking and thinking about; it is just that we are doing so in way that no longer produces the problematic cognitive dissonance. How can this change of concept without change of topic be achieved?

One answer comes from causal theories of reference [32]. These theories can be used to explain our intuition that a term can express a different concept even though its reference remains the same. On such a view, when we think about gold using our concept of gold, we think about the same thing that the ancients thought about when they thought about gold using their concept of gold, because

[5] This section is an edited version of a discussion in [14].

the same stuff — gold — that was the cause of their thoughts about gold is also the cause of our thoughts about gold (to put it roughly). But we are not employing the same concept they did. We now have a better concept of gold; we know what the essence of gold is — having an atomic number of 76. Although the ancients were confused and had many false beliefs about gold, it did not mean they were not thinking about gold. In fact, we can only make sense of the idea that some of their beliefs, such as "gold is a compound", were false, by understanding them as predicating of gold a property ("being a compound") that gold does not in fact have.

The proposal, then, is that we can do the same for consciousness: develop a new concept of consciousness that refers to the same phenomenon that our current concept does, but which is different enough to allow us to solve some of the conceptual problems we face. But it seems unlikely that the kind of conceptual change required can itself come about through merely propositional processes: such processes as adding propositions to, or subtracting propositions from, one's stock of beliefs (whether it be by learning some more facts about consciousness or about the brain, or by engaging in philosophical argumentation); or creating a new concept out of logical combinations of the concepts one already possesses. Such methods seem unable to surmount the impasse we have reached. If the recent history of discussions and disputes in the areas of consciousness studies and the philosophy of consciousness is any indication, there is no rational way to convince somebody who has the zombic hunch not to have the zombic hunch, and vice versa [31]. It is possible that the kind of conceptual change required cannot be achieved simply by reading journal articles about consciousness and engaging in other purely linguistic, propositional modes of inquiry. Of course these modes are extremely useful in developing our understanding of many things, including consciousness. But there is reason to believe that they are not enough to resolve certain intractable conceptual problems, especially in the case of consciousness.

Rather, if we are going to change our concept of consciousness so that we can make scientific sense of consciousness and understand the place of consciousness in the natural world, we might have to have our concept of consciousness undergo what you might call a non-propositional change, a non-propositional development of our concept of consciousness: changes to a concept that are not simply a case of adding propositions to, or subtracting propositions from, one's stock of beliefs, nor a case of creating a new concept out of logical combinations of the concepts one already possesses. But not just any kind of non-propositional change to one's concept of consciousness will be of use here. Getting hit on the head, for instance, might change what you think or what you think you think about consciousness, as might undergoing neurosurgery, or perhaps taking certain kinds of psychoactive drugs. Perhaps you will have an 'aha!' experience concerning the nature of consciousness if you do some of those things; perhaps some of them will result in conceptual change. But these are not the kinds of non-propositional change relevant to scientific explanations of consciousness. Like the examples just given, the relevant kinds of non-propositional change are not achieved by hearing a philosophical argument, nor by reading a passage of text. But unlike those examples, the scientifically relevant changes are still rational changes to which norms of justification apply; in particular, they are based on an experience of the subject matter. Unlike the examples above, where the change in your conceptual state is non-propositional but random, we seek changes to our concepts that track reality in some way, even if they cannot be summarised or transmitted propositionally (e.g., through text). This sounds impossible, but is, in fact, a pervasive phenomenon.

To see why this is so, we need to get a clearer view of what concepts are. Consider the famous duck—rabbit figure, or the Necker cube. Wittgenstein argued that what underlies being able to move between the different ways of seeing the Necker cube or the duck—rabbit is the "mastery of a technique" [33]. This is exactly the kind of ability we are looking for in the case of conceptual change concerning consciousness: the capacity to see consciousness in a new way. But it is also something more: the capacity to see how what you see in a new way is the very same thing as what you saw in the old way. To be able to see how a thing that is appreciable from the consciousness perspective is also the very same thing that is understandable from the physical perspective, to be able to see how an experience thought about using the successor concept of consciousness is the very same thing that was being thought about using the predecessor concept of consciousness, to move seamlessly between those two viewpoints, is a skill. If so, then we now know that what we need in order to resolve the conceptual problems of consciousness is a skill: Acquiring the right concept of consciousness is a matter of acquiring a skill.

Note that skills are usually such that they cannot be transmitted by text alone. You cannot acquire the ability to ride a bicycle solely by reading journal articles on the vestibular system, or by having a philosophical argument on the knowing that/knowing how distinction. Rather, skill in a domain typically requires experience of that domain. For instance, some symbolic, linguistic, propositional, information (the advice of a friend) helped me learn to juggle. But it was not sufficient for me to acquire that skill. It is true that I tried to acquire the skill without that conceptual knowledge and did not do very well; it was only when my friend helped me, through language, to draw attention to certain aspects of my experience, I was able to juggle, in my own feeble way. But the advice alone did not give me the skill; I had to attempt to juggle in order for the advice to be of any use. So also, I am maintaining here, for the case of understanding consciousness. This is not to say that there is no role for argumentation, thinking, reading journal articles, etc., but rather that these are not enough, and we need something in addition we need a skill that cannot be acquired purely propositionally.

The idea that skill acquisition is a way to achieve non-propositional conceptual change is an important part of a view that I call ''interactive empiricism''. This is distinct from empiricism *simpliciter*, the view that all concepts must be grounded in sense experience. Interactive empiricism is not a species of empiricism in that sense. Rather, it is the view that the possession of some concepts requires having a particular kind of experience, a kind usually not emphasized in traditional empiricism. The ''interactive'' in ''interactive empiricism'' indicates what this particular kind of experience must be: the experience of interacting with the subject matter of the concept, the stuff the concept is *of*. To have the kind of concept that will solve our conceptual problems, one must master a technique of understanding how one's perspective on the subject matter — in this case consciousness itself — will change in the light of one's different ways of intervening in that subject matter. Both the acquisition and application of this skill requires having experiences of interacting with the subject matter of consciousness. Riding a bicycle is not merely a passive reception of input of the kind the traditional empiricists were thinking of when they talked about grounding ideas in experience. The experience is interactive, the result of a dynamic engagement with the world. One acquires the skill of riding a bicycle because one experiences sensory feedback in response to one's actions; the kind of experience that goes beyond a mere a one-way input from the world to your ideas. So also, I argue, for the skills that will constitute our successor concepts of consciousness.

This view is consonant with discoveries in cognitive science that interaction is essential to understanding cognition. For example, perception essentially involves interaction on some views, such as O'Regan and Noë's sensory—motor contingency theory [21]. On that account you can only be perceiving the world if you have some capacity to interact with it, or if you are actually interacting with it. Consciousness essentially involves interaction on views such as Susan Hurley's, as the title of her book, *Consciousness in Action* [34], testifies. Cognition in general is thought to essentially involves interaction on views such as Mark Bickhard's; hence the title of his ''Interactivist Manifesto'' [35]. But a helpful illustration of a concrete way in which interaction is crucial to certain kinds of cognitive development, in this case visual perception, comes not from such recent work, but rather the classic study by Held and Hein [36]. They placed neonate kittens with undeveloped visual systems in an apparatus consisting of a circular room with a bar suspended from the ceiling, able to rotate about its midpoint. At the end of each bar was a harness for a kitten. The harnesses were such that one kitten was touching the ground and was able to move around relatively freely, but the second kitten was suspended in a way that its movements did not change its position in the world at all; rather, its position was determined by the movements of the first kitten. For the first kitten, there was a very natural interdependence between its actions and the visual input it received. For the second kitten, there was little or no correlation between its muscle movements and the visual input it received, because the visual input was largely determined by the first kitten. No matter what the second kitten did with its limbs, it did not, in the main, affect the input it received. The result of this study was striking; after the developmental period; the first kitten had more or less normal vision, while the second kitten was more or less blind. This suggests that having the right kind of interaction, engaging in action and having sense experience that is appropriately related to those actions, is crucial to certain kinds of development.

Although it is only an analogy, a striking suggestion can be made: perhaps our conceptual development shares this property with visual development in kittens. If interaction is a general cognitive principle that governs our conceptual development as well, then some developments in our concepts, for instance our concept of consciousness, may also require us to intervene in a subject matter and then receive reciprocal experiences that are appropriately related to those interventions. In the case we are considering, the interventions will be in the phenomenon of consciousness itself. To put the point another way: a general science of human cognition should apply to individual cognizers; specifically, it should apply to AC researchers, be they cognitive scientists, philosophers, or engineers. And if one's cognitive science says that cognition in general, and conceptual development in particular, is interactive, it may also be that making philosophical advances via conceptual development will necessarily involve engaged, experiential activity.

This, finally, is the motivation for weak conceptually necessary AC. The kind of interaction that is relevant to understanding consciousness, cognitive systems, artificial consciousness, etc., is the design and construction of actual working systems (either autonomous or prosthetic) that model or exhibit consciousness-related phenomena. These kinds of interactions may be the kind that are required for the requisite conceptual development, that will allow us to acquire the skills that constitute a conceptual advance with respect to consciousness. There might be other kinds of interaction that could assist in conceptual development with respect to consciousness. For instance, in interacting with each other or interacting with subjects in the experimental laboratory, if we were not only interacting in a more or less normal way, but also had access to real-time scans of each other's brains, this might also be a way of having the kind of interactive experience required to develop our concept of consciousness. But this is a much less plausible idea than the engineering-based AC approach. First, there are the ethical issues: true interaction would require intervention on the lowest, neural level: directly altering another's neural state. But there are non-ethical problems as well. Compare trying to understand, say, how a computer works in a similar way. That is, suppose that while you are using a computer you have an oscilloscope and you can see what is going on at the hardware level of the computer. In theory perhaps you could get some great insights, and acquire some skill that would constitute a better concept of computation, but it seems unlikely; it is just too much of a jump from the lowest to the highest level to expect you to see some interesting correlations rather, it seems likely that a step-by-step, level-by-level, structured approach is required for interaction to have an effect on our concept of consciousness. So also, then, with the brain scan suggestion. By contrast, designing and building cognitive systems can and does provide this mediated structuring of activity, and is thus more likely to be the kind of interaction that is going to yield the right kind of conceptual change.

3.2.3. Motivating discriminative *lagom* AC: synthetic phenomenology

As with any science, a science of consciousness requires an ability to specify its *explananda* (facts, events, etc. to be explained) and its *explanantia* (states, facts, events, properties, laws, etc. that do the explaining). Conscious states may be expected to play both of those roles. A science of consciousness, then, has a double need for a way to specify experiences. At least part of what is essential to most, if not all, experiences is their *content*. The content of an experience is the way the experience presents the world as being. How can we specify the content of particular experiences?

The standard way of specifying the content of mental states is by use of 'that' clauses. For example, 'Sue believes that the dog is running' ascribes to Sue a belief, the content of which is the same as, and is therefore specified by, the phrase following the word 'that': i.e., 'the dog is running'. Although "that" clauses work for specifying the content of linguistically and conceptually structured mental states (such as those involved in explicit reasoning, logical thought, etc.), there is reason to believe that some aspects of mentality (e.g., some aspects of visual experience) have content that is not conceptually structured [37—40]. Insofar as language carries only conceptual content, "that" clauses will not be able to specify the non-conceptual content of experience. An alternative means is needed.

I have previously suggested [40] that we might instead use the states of a robotic model of consciousness to act as specifications of the contents of the modeled experiences. Doing so would allow us, in principle, to systematically and canonically communicate to each other precise experiential contents not otherwise specifiable with conceptual language. This idea, called "synthetic phenomenology" (see also [41]), has been developed for the case of specifying the non-conceptual content of visual experiences in the SEER-3 project [10,11]. Specifications using SEER-3 rely on a discriminative theory of visual experience based on the notion of enactive expectations (expectations the robot has to receive a particular input were it to move in a particular way). Depictions of the changing expectational state of the robot can be constructed in real time, depictions that require the viewer to themselves deploy sensory—motor skills of the very kind that the theory takes to be essential to individuating the specified content. Thus, the viewer comes to know the discriminating characteristics of the content in an intuitive way (in contrast to, say, reading a list of formal statements each referring to one of the millions of expectations the robotic system has).

The case for synthetic phenomenology, then, is a case for discriminative *lagom* AC. It is not a case of strong AC, since, as is pointed out in the SEER-3 reports referenced above, no assumption of sufficiency for the replication of consciousness is made, nor is such needed, for the robotic-assisted specifications to perform their task. It is not weak AC since it is being claimed that the patterns of expectations in the model are the very features that, in the system being modeled, differentiate one experiential content from another. On the other hand, it may turn out to be a practically (either causally or conceptually) necessary technology for scientific AC, (in which case Table 1 is wrong to suggest that only weak AC can have these characteristics).

Consideration of the case of synthetic phenomenology raises another issue concerning the framework of AC distinctions depicted in Table 1: could there ever be a case of discriminative AC that is not *lagom*, that is, a case of strong discriminative AC? The point of the constitutive vs. discriminative distinction was to allow for AC models that do not attempt to explain why something *is* conscious, but instead explain why it is in this conscious state rather than that one. But a strong AC model, it seems, would exclude the modesty of the discriminative approach, since it would, presumably, attempt to explain not just which conscious state a system is in, but why it is conscious at all. Although there indeed seem to be interdependencies between the two distinctions that challenge the simple lines of Table 1, the present confusion can be eliminated by understanding strong discriminative AC to be a model that is *metaphysically* sufficient for the presence of consciousness, but *explanatorily* sufficient only for distinctions between conscious states.

4. Difficulties for artificial consciousness

There are several reasons why one might think that either AC cannot be achieved in the engineering sense, or that it cannot contribute to an understanding of consciousness. However, armed with the distinctions in possible ambitions of AC from Section 2, it may be possible to delineate the reach of these objections with greater precision than has been done before, in a way that reveals at least some substantial AC goals to be unobjectionable. Before considering specific problems concerning the Chinese room, and qualia, some more general difficulties will be discussed, and set aside.

4.1. General difficulties

4.1.1. Objections not specific to AI

It would be inappropriate to consider here objections to scientific AC that are not specific to it; that is, objections that non-AC approaches face as well. This is not to demean the seriousness of these objections, only to point out that this is not the proper venue for their consideration.

First, there are the problems shared by all scientific approaches to understanding consciousness. For example, it is usually agreed that phenomenal states can be observed directly only by the subject of those states, yet objective or at least inter-subjective observation and verification is thought to be at the very heart of scientific method (cf., e.g. [42,43]). Similarly, Searle [44] points out that reductive science has proceeded by carving off the appearances from the reality of the phenomenon:

> "But in the case of consciousness, its reality is the appearance; hence, the point of the reduction would be lost if we tried to carve off the appearance and simply defined consciousness in terms of the underlying physical reality... Reductions that leave out the epistemic bases, the appearances, cannot work for the epistemic bases themselves. In such cases, the appearance is the reality." (p 122)[6] (For a response to this form of objection that questions the notion of objectivity on which it relies, see [45].)

Another objection that all scientific approaches to consciousness face in some form or other, the "how would you ever know if you succeeded?" objection, was already discussed in Section 3.2.

Next to set aside are the difficulties that all naturalistic (viz., non-dualistic) approaches to understanding consciousness face, such as Jackson's knowledge argument [46], the explanatory gap [13], or the "hard problem" [47]. According to this sort of objection, the (knowledge of) non-phenomenal properties of a system cannot explain, or at least do not entail (knowledge of) the phenomenal properties of that system.

However, a response to Jackson can be given based on the picture developed in Section 3.2.2. Once one accepts, as urged in that section, that science is itself an interactive activity that involves interventional experience of the world, the whole premise of Jackson's argument is revealed to be contradictory. Jackson asks us to imagine that Mary knows everything science has to say about colour vision but has never seen red. On the view of science presented above, this is revealed to be a contradiction. As Alter [48] has independently observed, Jackson assumes that all the knowledge of physical science can be written down and acquired by Mary through reading. If the gist of Section 3.2.2 is right,

[6] It is true that Searle then goes on to say that this impossibility of a reductive science of consciousness need not trouble us: "Once the existence of (subjective, qualitative) consciousness is granted, then there is nothing strange, wonderful or mysterious about its irreducibility." But on the other hand, Searle admits that there are virtues to reduction: "To get a greater understanding and control of reality, we want to know how it works causally, and we want our concepts to fit nature at its causal joints."

then this assumption of Jackson's is false. Science in general involves experiential activity, and colour science in particular involves the having of colour experiences in a way that systematically depends on our interventions. So for Mary to truly know everything that ''science knows'' about colour vision, she would have to have all the experiences that scientists have needed to have about colour vision, in particular the experience of seeing red. To suppose this and to suppose also that Mary has not had the experience of seeing red is a contradiction. An underestimation of the extent of the social store of scientific knowledge is often the result of an overly individualistic view of science.

Also note that *lagom* AC is not at odds with the existence of an explanatory gap, or the ''hard problem'', since it aspires to contribute to an explanation of consciousness without attempting to provide sufficient conditions for it. Similar exemptions can be made for prosthetic and discriminative AC.

4.1.2. Objections not specific to consciousness

Just as it would be wide of the mark to consider here objections not specific to AI, so also it would be to consider objections that do not specifically target AI's ability to produce or explain consciousness, as opposed to mental states in general. These objections typically vary depending on the AI approach being criticised; indeed, the development of some AI approaches can be seen as attempts to overcome the general limitations of another (usually the symbolic) approach. For example, it is argued that symbolic AI cannot provide an accurate model of human cognition, since in such AI systems, millions of serially dependent operations must be performed to, say, recognise an object, but this is done in the brain in fewer than one hundred such steps [49]. Connectionist AI is then offered as an approach that does not suffer from this problem. Another example of a purported general obstacle to symbolic AI in particular is the frame problem [50,51].

The diagonal argument against symbolic AI is harder to classify. It dates back to Gödel and Turing, but was developed philosophically by Lucas (cf. [52] for a retrospect) and, more recently, Penrose [53]. It can be shown that no Turing machine can compute the non-halting function. Enumerate all the Turing machines. Now consider this function: ''For all n, halt if and only if the nth Turing machine does not halt when given input n''. (Here ''halt'' just means: gives a well-formed answer). No Turing machine that is sound (i.e., never gives a well-formed answer that is incorrect) with respect to this function can halt when given its own enumeration number k as an argument (it can halt on k only if it does not halt on k, on pain of being unsound). Furthermore, I just proved to you that any such sound Turing machine k does not halt when given input k. Thus you and I can answer the question (and presumably compute its characteristic function!) correctly for all n, while no Turing machine can. So we can do more than Turing machines: No Turing machine can model what we are doing when we are considering such mathematical questions. Thus, a symbolic AI employing only sound algorithms cannot even reproduce our behaviour, let alone explain it.

What makes the argument difficult to classify is the fact that not only does it tell against engineering goals as much as scientific ones, it appears to be against artificial mentality in general (no characteristics specific to consciousness, such as the ones in Carruthers' list of desiderata from Section 1, are mentioned in the argument). Yet Penrose maintains that consciousness is central to the argument's success. It is that we are conscious, Penrose claims, that explains why we are able to jump out of the algorithmic ''system'', see patterns that are not classically computable, etc. Once again, there are too many replies and counter-replies to consider them here (although see [54]). But some observations can be made; specifically, *lagom* AC (and AI) is again immune to this argument. Even if there are aspects of consciousness that are non-computational, this does not show that computation of some sort is not necessary/explanatory for those aspects of consciousness. Nor does it show that all aspects or instances of consciousness have non-algorithmic components. Furthermore, the argument does not apply in general to alternative approaches to AI that do not place sound algorithms centre-stage. And one may wonder: even if computers cannot recreate human consciousness, is halting-problem-defying human consciousness the only kind of consciousness possible in this universe? If not, then AI (even sound algorithmic symbolic AI) explanations of these other kinds of consciousness have not been shown to be impossible (cf. the human vs. general distinction, Section 2.6). Finally, it is not clear that the argument has anything to say against prosthetic AC.

Although most critiques of AI and AC implicitly assume a target technology (e.g., digital computation for the diagonalisation argument; von Neumann machines running programs for the Chinese room), there is at least one argument against strong scientific AI that applies independently of approach. This argument claims that artificial intelligence is an oxymoron, or a contradiction in terms ([27], pp 3—4; [55], p 418). It maintains there is an incompatibility between the possibility of having a mind and at the same time being an artefact in any interesting sense. To be considered artificial, an AI system would have to be not just the result of our labour (children are that), but also designed by us (if not designed by us, can it really be considered artificial in any interesting sense?). But, it can be argued, this means that any purpose, meaning or intentionality in the system is not its own, but rather derivative from our design of it. Yet mindedness is autonomous, exhibiting original, non-derivative intentionality.

As with so many other objections to AC, this argument does not apply to prosthetic AC, since the recipient of the prosthesis presumably already possesses the autonomy required for real mindedness. Nor does it apply to *lagom* AC. It may still be necessary for conscious agents that they have a particular abstract causal organisation, even if another condition must be met as well: that they not have that organisation as the result of intelligent design. But this reveals the interesting point that at least some AC approaches are not really concerned with *artificial* intelligence per se; they could still make their contribution as to the necessary conditions for consciousness even if it were the case that the only systems that could ever be conscious were natural, non-artificial ones. In this respect, "machine consciousness" is a more inclusive term for the field (cf. the parenthetical remark in Section 2.5).

However, it is just these sorts of considerations that raise the suspicion that the argument is *too strong*. It would imply that I might not be, and might never have been, conscious, given that it is possible that I was created by intelligent design (be it divine or mundane). Yet (as Descartes made vivid) surely I cannot accept that it is possible that I am not, nor ever have been, conscious! On the other hand, it would be odd indeed if the details of my origins, usually believed to be an empirical matter, could be known by me *a priori* in this way.

With these "wide-of-the-mark" objections dealt with or set aside, we can proceed to consider a few objections that are clearly specific to AI accounts of consciousness.

4.2. The Chinese room

Perhaps the most well known of the specific objections to symbolic AI accounts of consciousness is Searle's Chinese room argument [12]. Searle argues against the claim that a computer could understand solely by virtue of running the right program. To do this, he exploits the subjective, conscious nature of understanding, and the formal, implementation-independent nature of symbolic computation and programs. Since Searle can himself implement any purported Chinese-understanding-endowing program, and presumably would not come to understand Chinese thereby, he apparently refutes the strong AI claim he targets. Again, it is not possible to rehearse here the various replies and counter-replies that have been given. But it is worth noting that *lagom* AC is immune to this argument. The Chinese room may show that computation is not *sufficient* for conscious understanding, but it does not show that it is not *necessary* for it, nor does it show that computation cannot play an explanatory role with respect to consciousness. And of course the argument does not apply in general to alternative approaches to AC that do not place as much emphasis on implementing formal programs.

Determining whether the Chinese room argument applies to prosthetic AC is tricky. On the one hand, there is the point (made in response to several objections to AC) that since, in the case of prosthesis, we already know that the basic sufficiency conditions for consciousness are met, there is little room for denying that they have been met for the prosthetic experiences in question. On the other hand, one might think that this is exactly the kind of move that Searle's argument is against. The thought experiment can be seen as involving an elaborate and implausible prosthesis, in which the basic sufficiency conditions for consciousness have been met (in Searle). The added technology that would, from a behavioural or functional point of view, be sufficient for introducing new kinds of experience into the system, in fact are not, if the intuitions employed by the thought experiment are to be trusted.

4.3. Qualia and ontologically conservative heterophenomenology

In Section 3.2.1 we saw how functionalism is one of the primary motivations for optimism about scientific AC. One of the most well known objections to functionalism is its inability to deal with *qualia*. Qualia are the "raw feels" of experience, the elements that make up what it is like to be in a particular conscious state. Dennett [56] diagnoses the concept of qualia as the concept of mental particulars that have the four properties of being "ineffable, intrinsic, private, and directly or immediately apprehensible in consciousness". Arguments against functionalism aim to show that sameness of functional state does not imply sameness of qualia; thought experiments involving absent or inverted qualia seem to show that explanations that appeal only to the abstract causal organisation of a system will always leave something out: how it feels, in terms of qualia, to be in that state.

Again, there is no space here to review these arguments in detail, still less the replies and counter-replies they give rise to. But a few points can be made that arise out of the framework given in Section

2. First, insofar as qualia-based objections are against the sufficiency of computation for consciousness, they count only against strong scientific AC. This has been readily acknowledged, but it is usually assumed that this leaves only weak scientific AC, and it can be dismissed as "uninteresting", in the sense of Section 2.2. But as was shown in that section, *lagom* scientific AC is interesting in the pertinent sense (it aims to explain constitutive, *necessary* aspects of consciousness) without making the claims of sufficiency that render strong scientific AC vulnerable to the qualia-based attack. Only if one went further, to a wildly implausible, hyper-qualia position that maintains that there are not even any necessary functional conditions for being in a particular qualitative state, would *lagom* scientific AC be under threat as well. That is, rather than the standard zombie arguments, which focus on the possibility that there might be physically indistinguishable agents that *differ* in their phenomenal properties, one would need to embrace a hyper-qualia position that claims that two systems with little or nothing in common (me, and an electron, say), might have exactly the *same* phenomenal state.

Similarly, prosthetic scientific AC is not touched, since it is clear in such cases that the sufficiency conditions for qualia *are* met. Since variations in qualia can be generated by making systematically related variations in the functional structure of the prosthetic technology, claims that such structure has no explanatory role to play in an account of such experiential states is unmotivated, to say the least.

One defence of strong scientific AC against qualia-based arguments that will be considered here is based on Dan Dennett's notion of *heterophenomenology* (defined below). The defence proceeds in two stages.

First, it is admitted that functionalism cannot account for qualia as characterised by the aforementioned four properties of ineffability, intrinsicness, privacy, and immediacy; but this is not taken to be a problem for functionalist explanations of consciousness because it is denied, on Dennett's view, that there are any aspects of consciousness that have these four properties. Again, the arguments and "intuition pumps" for this claim are too various and detailed to recount here, but the overall point is that these conditions together make qualia something unsuitable either for explaining or for being explained. Anything that meets the four conditions is, by definition, a difference that can make no difference, something which is beyond all possible verification, and thus meaningless. In a move that is usually taken to follow from what has just been explicated (but see below), the existence of qualia is then denied.

The second step of the defence then proceeds as a kind of error theory: Even though there is nothing that can be the referent of the term "qualia", a complete theory of consciousness should explain why it is that people use the term (or terms like it) in the ways that they do, why they are disposed (if they are) to believe that there are some aspects of consciousness that are ineffable, intrinsic, private, and immediate. Rather, one takes subjects' phenomenological reports seriously in the sense that the existence of the reports and the beliefs they express are data to be explained, but one does not take them at face value by assuming that the beliefs must be true. This is the method that Dennett has labelled "heterophenomenology" [57—59].

Once this data is in, theories, including AC-based ones, can be constructed to explain it. These theories and models would attempt to answer questions such as: If belief in qualia is universal or near universal, what could explain this? Is there some way in which (falsely, given step one) believing such things about oneself and one's experience might provide some information processing, meta-management, or communicative advantage? (Carruthers' five desiderata from Section 1 are particularly applicable here.) Once these facts are explained, it is claimed, those who now feel that the heterophenomenological method is leaving something out by denying the existence of qualia will no longer feel that way. Consciousness will have been explained, including (importantly) why it previously seemed that it could not be so explained.

The impact of heterophenomenology on consciousness studies, however, has been limited. For many, the eliminativism of qualia that this strategy entails is unacceptable. On the other hand, Dennett would say that the elimination of qualia is precisely the point of the strategy: to make sense of our qualia talk without thereby committing oneself to the existence of entities that have the combination of properties which Dennett (rightly) finds so objectionable. But there is a middle way, one that uses heterophenomenology in an AC context to achieve Dennett's goal, without thereby incurring a commitment to eliminating qualia, which so many find counter-intuitive. The approach, originally introduced in [60] (pp 31—35), and which we can call "ontologically conservative heterophenomenology", relies on the causal theory of reference (cf. Section 3.2.2) to make sense of the following idea: even though nothing has the four properties believed to be characteristic of qualia, it does not immediately follow that qualia do not exist, that the term "qualia" does not refer. This is because it may be that the term "qualia", like "gold", has its reference fixed not by description (the four properties), but by causal relations between the term/concept and whatever caused the term/concept to be introduced in the first place, and/or whatever conditions resulted in it being a useful term. With such a theory in place, we can now see that there is a gap between showing that nothing has the four properties believed to be characteristic of qualia, and showing that qualia do not exist, just as showing that the beliefs the ancients had about gold were false did not amount to showing that there is no such thing as gold.

The advantage of being neutral about the ontological status of qualia is that it allows a continuity (viz., sameness of term/reference) between the current, problematic qualia-based view and any improved, paradox-free successor account. Dennett's eliminativism foregoes such continuity, and it is therefore susceptible (rightly or wrongly) to charges of ''changing the subject'' and ''not addressing the problem''.

Of course, it may turn out that ''qualia'' *is* like the term ''phlogiston'': it so poorly maps to anything actually going on that there will be no place even for a successor concept of qualia in a future science of the mind. But the point is that this should be decided empirically/experientially; it is not the place of philosophy to prejudge the issue (and thereby needlessly deny the heterophenomenological approach of many potential adherents). If it turns out that building AC models of qualia does underwrite their existence, albeit one which is more accurately captured with a distinct, successor concept of qualia, it would counts as an instance of the interactive empiricism described in Section 3.2.2: a (practically necessary?) change in our concept of consciousness brought about by the experiential activity of model construction.

Acknowledgements

I would like to thank the members of PAICS (the Philosophy of Artificial Intelligence and Cognitive Science research group in the Department of Informatics at the University of Sussex) and other members of the audience at the E-intentionality seminar of 24 January 2008 for their helpful comments. Thanks also to Oliver Sharpe and Berkan Eskikaya.

References

[1] Carruthers P. Précis of Phenomenal consciousness, http://lgxserver.uniba.it/lei/mind/forums/002_0002.htm; 2001 [accessed 1.5.08].

[2] Carruthers P. Phenomenal consciousness. Cambridge: Cambridge University Press; 2000.

[3] Hesslow G. Conscious thought as simulation of behaviour and perception. Trends in Cognitive Sciences 2002;6:242—7.

[4] Hesslow G, Jirenhed D-A. The inner world of a simple robot. Journal of Consciousness Studies 2007;14:85—96.

[5] Hesslow G, Jirenhed D-A. Must machines be zombies? Internal simulation as a mechanism for machine consciousness. In: Proceedings of the AAAI fall symposium on machine consciousness. Washington: AAAI Press; 2007. p. 78—83.

[6] Aleksander I. How to build a mind: toward machines with imagination. London: Weidenfeld and Nicolson; 2000.

[7] Holland O, Goodman R. Robots with internal models: a route to machine consciousness? Journal of Consciousness Studies (Special Issue on Machine Consciousness) 2003;10(4—5): 77—109.

[8] Haikonen PO. You only live twice: imagination in conscious machines. In: Chrisley R, Clowes R, Torrance S, editors. Proceedings of the AISB05 symposium on machine consciousness. Hatfield: AISB Press; 2005. p. 19—25.

[9] Shanahan M. A cognitive architecture that combines internal simulation with a global workspace. Consciousness and Cognition 2006;15:433—49.

[10] Chrisley R, Parthemore J. Synthetic phenomenology: exploiting embodiment to specify the non-conceptual content of visual experience. Journal of Consciousness Studies 2007;14(7):44—58.

[11] Chrisley R, Parthemore J. Robotic specification of the non-conceptual content of visual experience. In: Proceedings of the AAAI fall symposium on consciousness and artificial intelligence: theoretical foundations and current approaches; 2007. p. 36—42.

[12] Searle J. Minds, brains and programs. Behavioral and Brain Sciences 1980;3:417—57.

[13] Levine J. Materialism and qualia: the explanatory gap. Pacific Philosophical Quarterly 1983;64:354—61.

[14] Chrisley R. Interactive empiricism: the philosopher in the machine. In: McCarthy, N, editor. Philosophy of Engineering: proceedings of a series of seminars held at The Royal Academy of Engineering. London: Royal Academy of Engineering; in press.

[15] Ashby R. Design for an intelligence-amplifier. In: Chrisley R, editor. Artificial intelligence: critical concepts, vol. III. London: Routledge; 2000. p. 191—209. Originally appeared in Shannon CE, McCarthy J, editors. Automata studies. Princeton: Princeton University Press; 1956. p. 215—34.

[16] Clark A, Chalmers D. The extended mind. Analysis 1998;58: 7—19.

[17] Auvray M, Myin E. Perception with compensatory devices: from sensory substitution to sensorimotor extension. Cognitive Science; in press.

[18] Bach-Y-Rita P. Brain mechanisms in sensory substitution. New York and London: Academic Press; 1972.

[19] Heidegger M. Being and time a translation of Sein and Zeit. Albany, NY: State University of New York Press; 1996.

[20] Merleau-Ponty M. Phenomenology of perception. London: Routledge; 1962.

[21] O'Regan K, Nöe A. A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences 2001;24(5):939—73.

[22] Fodor J. The big idea, can there be a science of mind? Times Literary Supplement 1992;3:5—7.

[23] Armer P. Attitudes toward intelligent machines. In: Chrisley R, editor. Artificial intelligence: critical concepts. London: Routledge; 2000. p. 325—42. Originally appeared in Feigenbaum E, Feldman J, editors. Computers and thought. New York: McGraw-Hill; 1963. p. 389—405.

[24] Whitby B, Yazdani M. Artificial intelligence: building birds out of beer cans. Robotica 1987;5:89—92.

[25] Chrisley R. Embodied artificial intelligence. Artificial Intelligence 2003;149:131—50.

[26] Sloman A, Chrisley R. More things than are dreamt of in your biology: information processing in biologically-inspired robots. Cognitive Systems Research 2005;6(2): 145—74.

[27] Chrisley R, editor. Artificial intelligence: critical concepts. London: Routledge; 2000.

[28] Sloman A. The computer revolution in philosophy. Atlantic Highlands: Harvester Press; 1978.

[29] Brooks R. Intelligence without reason. In: Chrisley R, editor. Artificial intelligence: critical concepts, vol. III. London: Routledge; 2000. p. 107—63. Originally appeared as MIT AI Lab Memo 1293; 1991.

[30] Dennett DC. The intentional stance. Cambridge: MIT Press; 1987.

[31] Dennett DC. Sweet dreams: philosophical obstacles to a science of consciousness. Cambridge: MIT Press; 2005.

[32] Kripke S. Naming and necessity. Cambridge: Harvard University Press; 1980.

[33] Wittgenstein L. Philosophical investigations. Oxford: Blackwell; 1972.

[34] Hurley S. Consciousness in action. Cambridge: Harvard University Press; 1998.

[35] Bickhard M. Interactivism: A Manifesto. In: Campbell RL, Ó Nualláin S, Bickhard, MH, editors. The study of mind: toward inter- and intra-disciplinary cooperation. Available at http://www.lehigh.edu/∼mhb0/InteractivismManifesto.pdf; in press [accessed 11.11.07].

[36] Held R, Hein A. Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology 1963;56(5):872—6.

[37] Evans G. The varieties of reference. Oxford: Oxford University Press; 1982.

[38] Cussins A. The connectionist construction of concepts. In: Boden M, editor. The Philosophy of Artificial Intelligence. Oxford: Oxford University Press; 1990. p. 368—440.

[39] Peacocke C. A study of concepts. Cambridge: MIT Press; 1992

[40] Chrisley R. Taking embodiment seriously: non-conceptual content and robotics. In: Ford K, Glymour C, Hayes PJ, editors. Android epistemology. Cambridge: MIT Press; 1995. p. 141—66.

[41] Aleksander I, Morton H. Depictive architectures for synthetic phenomenology. In: Chella A, Manzotti R, editors. Artificial consciousness. Exeter: Imprint Academic; 2007. p. 30—45.

[42] Nagel T. What is it like to be a bat? In: Block N, editor. Readings in the philosophy of psychology, volume one. Cambridge: Harvard University Press; 1980. p. 159—70.

[43] Nagel T. The view from nowhere. Oxford: OUP; 1986.

[44] Searle J. The rediscovery of the mind. Cambridge: MIT Press; 1992.

[45] Chrisley R. A view from anywhere: prospects for an objective understanding of consciousness. In: Pylkkänen P, Vadén T, editors. Dimensions of conscious experience (Advances in consciousness research 37). Amsterdam/Philadelphia: John Benjamins Publishing; 2001. p. 3—13.

[46] Jackson F. Epiphenomenal qualia. Philosophical Quarterly 1982;32:127—36.

[47] Chalmers DJ. The conscious mind: in search of a fundamental theory. Oxford: OUP; 1996.

[48] Alter T. A limited defence of the knowledge argument. Philosophical Studies 1998;90:35—56.

[49] Feldman J, Ballard D. Connectionist models and their properties. Cognitive Science 1982;6:205—54.

[50] McCarthy J, Hayes PJ. Some philosophical problems from the standpoint of artificial intelligence. In: Michie D, Meltzer B, editors. Machine intelligence 4. Edinburgh: Edinburgh University Press; 1969. p. 463—502.

[51] Pylyshyn Z, editor. The robot's dilemma: the frame problem in artificial intelligence. Norwood: Ablex Publishing; 1987.

[52] Lucas J. Minds, machines and gödel: a retrospect. In: Chrisley R, editor. Artificial intelligence: critical concepts, vol. III. London: Routledge; 2000. p. 359—76. Originally appeared in Millican, PJR, Clark A, editors. Machines and thought: the legacy of alan turing. Princeton: Princeton University Press; 1996. 103—24.

[53] Penrose R. The shadows of the mind: a search for the missing science of consciousness. Oxford: Oxford University Press; 1994.

[54] Chrisley R. Transparent computationalism. In: Scheutz M, editor. New computationalism: conceptus-studien 14. Sankt Augustin: Academia Verlag; 2000. p. 105—21.

[55] Boden M. Artificial intelligence and natural man. Atlantic Heights: Harvester Press; 1977.

[56] Dennett DC. "Quining Qualia". In: Marcel A, Bisiach E, editors. Consciousness in contemporary science. New York: Oxford University Press; 1988. p. 42—77. Reprinted in Lycan W, editor. Mind and cognition: a reader. Cambridge: MIT Press; 1990. 519—47; and in Goldman A, editor. Readings in philosophy and cognitive science. Cambridge: MIT Press; 1993. 381—414.

[57] Dennett DC. How to study consciousness empirically, or nothing comes to mind. Synthese 1982;59:159—80.

[58] Dennett DC. Consciousness explained. Boston: Little Brown; 1991.

[59] Dennett DC. Who's on first? Heterophenomenology explained. Journal of Consciousness Studies 2003;10(9—10):19—30.

[60] Sloman A, Chrisley R. Virtual machines and consciousness. Journal of Consciousness Studies 2003;10(4—5):113—72.