PRE-PRINT (25/05/2018)
PUBLISHED VERSION IN V.C. MÜLLER (ED.), *PHILOSOPHY AND THEORY OF ARTIFICIAL INTELLIGENCE*. SAPERE VOLUME 44. BERLIN: SPRINGER NATURE. PAGES 3-18.

Artificial Consciousness: From Impossibility to Multiplicity

Chuanfei Chin

Department of Philosophy, National University of Singapore, Singapore 117570
phiccf@nus.edu.sg

Abstract. How has multiplicity superseded impossibility in philosophical challenges to artificial consciousness? I assess a trajectory in recent debates on artificial consciousness, in which metaphysical and explanatory challenges to the possibility of building conscious machines lead to epistemological concerns about the multiplicity underlying 'what it is like' to be a conscious creature or be in a conscious state. First, I analyse earlier challenges which claim that phenomenal consciousness cannot arise, or cannot be built, in machines. These are based on Block's Chinese Nation and Chalmers' Hard Problem. To defuse such challenges, theorists of artificial consciousness can appeal to empirical methods and models of explanation. Second, I explain why this naturalistic approach produces an epistemological puzzle on the role of biological properties in phenomenal consciousness. Neither behavioural tests nor theoretical inferences seem to settle whether our machines are conscious. Third, I evaluate whether the new challenge can be managed through a more fine-grained taxonomy of conscious states. This strategy is supported by the development of similar taxonomies for biological species and animal consciousness. Although it makes sense of some current models of artificial consciousness, it raises questions about their subjective and moral significance.

Keywords: artificial consciousness, machine consciousness, phenomenal consciousness, scientific taxonomy, subjectivity.

1 Introduction

I want to trace a trajectory in recent philosophical debates on artificial consciousness. In this trajectory, metaphysical and explanatory challenges to the possibility of building conscious machines are supplanted by epistemological concerns about the multiplicity underlying 'what it is like' to be a conscious creature or be in a conscious state. Here *artificial consciousness* refers, primarily, to phenomenal consciousness in machines built from non-organic materials. Like most of the philosophers and scientists whom I discuss, I will follow Block (1995) in using the concept of phenomenal consciousness to refer subjective experience. By Block's definition, the sum of a state's phenomenal properties is what it is like to be in that conscious state, and the sum of a creature's phenomenal states is what it is like to be that conscious creature. The paradigms of such conscious states include having sensations, feelings, and perceptions.

Many surveys on artificial consciousness stress that this sub-field in artificial intelligence research has multiple interests (Gamez 2008; Holland and Gamez 2009; Reggia 2013; Scheutz 2014). Its research programmes aim to build machines which mimic behaviour associated with consciousness, machines with the cognitive structure of consciousness, or machines with conscious states. Often a distinction is drawn between *strong* artificial consciousness, which aims for conscious machines, and *weak* artificial consciousness, which builds machines that simulate some significant correlates of consciousness. Of course, a research programme may nurture interests in both strong and weak artificial consciousness; and the same model may be used to investigate both strong and weak artificial consciousness.

I shall focus on philosophical challenges to strong artificial consciousness. First, in the next section, I will analyse two earlier challenges which claim that phenomenal consciousness cannot arise, or cannot be built, in machines. These are based on Block's Chinese Nation and Chalmers' Hard Problem. To defuse such challenges, we can appeal to empirical methods and models of explanation. Second, I will explain why this naturalistic approach leads to an epistemological puzzle on the role of biological properties in phenomenal consciousness. Neither behavioural tests nor theoretical inferences seem to settle whether our machines are conscious. Third, I will evaluate whether the new challenge can be handled by a more fine-grained taxonomy of conscious states. This strategy is supported by the development of more fine-grained taxonomies for biological species and animal consciousness. Although it makes sense of some current models of artificial consciousness, it raises questions about their subjective meaning and moral status.

2 The impossibility of artificial consciousness

The literature on artificial consciousness contains several philosophical challenges to the possibility of building conscious machines (Bishop 2009; Gamez 2008; McDermott 2007; Prinz 2003; Reggia 2013; Scheutz 2014). Such challenges draw on philosophical arguments about the nature of consciousness and our access to it. One set of challenges is against the *metaphysical possibility* of artificial consciousness. These are based on the provocative thought experiments in Block (1978), Searle (1980), and Maudlin (1989), which suggest that machines, however sophisticated in functional or computational terms, cannot be conscious. Another set of challenges is directed at the *practical possibility* of building conscious machines. They are based on philosophical claims, made by McGinn (1991), Levine (1983), and Chalmers (1995), about our ignorance of how conscious states arise from physical states. According to these challenges, we can hardly expect to produce consciousness in machines if we cannot explain it in human brains.

Most theorists of artificial consciousness are not troubled by such challenges. In his survey, Scheutz (2014) describes two attitudes that support this stance. Here is how I understand them. First, some theorists hold a *pragmatic attitude* towards the concept of consciousness. They define this concept in an operational way, in terms of the processes and principles which psychologists take to underlie consciousness. Their aim is to use these processes and principles to improve performance in machines. They do not want

to replicate consciousness, so they need not worry if consciousness can arise, or be produced, in machines. This attitude particularly suits those whose research lies in weak artificial consciousness. Second, other theorists hold a *revisionary attitude*. They want to refine or replace the concept of consciousness through their empirical investigation of the underlying processes and principles identified by psychologists. In doing so, they wish to contribute to both psychology and philosophy. For instance, their models of the relevant processes and principles may enable new psychological experiments and produce new theories of consciousness. These may, in turn, influence philosophical intuitions and views about consciousness.

I take this last point to mean that empirical research into strong artificial consciousness need not be halted by the intuitions and views current in philosophy. To demonstrate this, I will show that theorists of artificial consciousness can appeal to empirical methods and models of explanation to defuse some philosophical challenges. In particular, I will look at how we can respond to two challenges to the possibility of building conscious machines – one based on Block's Chinese Nation thought experiment, the other on Chalmers' Hard Problem of consciousness.[1] Even those theorists who are less inclined to take philosophical challenges seriously can clarify their methodological commitments by considering these responses. Moreover, in the next section, I will show why the commitments underlying these responses lead to an epistemological puzzle which should interest all theorists of artificial consciousness.

(a) The first challenge centres on the nature of consciousness. It suggests that conscious machines cannot be built since machines cannot be conscious. More precisely, it suggests that the functional properties realisable by machines are not sufficient for consciousness. In Block's thought experiment, a billion people in China are instructed to duplicate the functional organisation of mental states in a human mind. Through radio connections and satellite displays, they control an artificial body just as neurons control a human body. They respond to various sensory inputs into the body with appropriate behavioural outputs. But, according to Block (1978), we are loath to attribute consciousness to this system: 'there is prima facie doubt whether it has any mental states at all – especially whether it has what philosophers have variously called "qualitative states," "raw feels," or "immediate phenomenological qualities"' (73). If our intuition about the Chinese Nation is sound, then consciousness requires more than the functional properties discovered in psychology. If so, the machines that realise only these functional properties cannot be conscious.

I do not think that we need to defer to this intuition about the Chinese Nation. Rather we should use empirical methods to uncover more about the nature of consciousness. Our best research – in psychology, neuroscience, and artificial consciousness – may determine that functional properties at a coarse-grained psychological level are sufficient for consciousness. Or it may determine that functional properties at a more fine-grained neurological level are necessary too. Whether the relevant properties are realisable in our machines is a further question, also to be determined by empirical investigation. None of this research should be pre-empted *a priori* by what our intuition

[1] I learnt especially from the responses offered in Prinz (2003) and Gamez (2008). I have put aside challenges based on Searle's Chinese Room thought experiment: they are analysed exhaustively in the literature on artificial consciousness, with what looks to be diminishing returns. One response to these challenges can be modelled after my response in (a).

says in a thought experiment and what that supposedly implies about the possibility of conscious machines.

Even Block would agree on this methodological point. He notes that, intuitively, the human brain also does not seem to be the right kind of system to have what he calls 'qualia', the subjective aspect of experience. So our intuition, on its own, cannot be relied on to judge which system does or does not have qualia. According to Block, we can overrule intuition if we have independent reason to believe that a system has qualia, and if we can explain away the apparent absurdity of believing this. Here his qualm about a system like the Chinese Nation rests mainly on our lack of a theoretical ground to believe that it has qualia. No psychological theory that he considers seems to explain qualia. That is why he insists of the system: 'any doubt that it has qualia is a doubt that qualia are in the domain of psychology' (84). To assuage this qualm, we need to build an empirical theory of consciousness which explains qualia and evaluates whether Chinese Nations, machines, and other systems have them.

(b) The second challenge directly addresses our explanation of consciousness. It suggests that we cannot build machines to be conscious even if machines can be conscious. According to Chalmers (1995), the Hard Problem we face is to explain how conscious experiences arise from physical processes and mechanisms in the brain. He distinguishes this from easy problems which require us to explain various psychological functions and behaviours in terms of computational or neural mechanisms. We have yet to solve the Hard Problem because we do not know how consciousness is produced in the human brain. But, until we do so, we cannot produce consciousness in a machine except by accident. Here is how Gamez (2008) sums up this line of reasoning based on our ignorance: "if we don't understand how human consciousness is produced, then it makes little sense to attempt to make a robot phenomenally conscious' (892).

I find two related reasons to reject this challenge. First, the production of consciousness may not require its explanation. Through empirical investigation, we may be able to produce consciousness without explaining it in terms of physical processes and mechanisms in the brain. If so, it suffices for us to create in machines the conditions which give rise to consciousness in humans; we need not understand, in philosophically satisfying terms, how the conditions do this. Our research to produce consciousness in machines may then help our research to explain consciousness in humans. This cross-fertilisation between research programmes would be in keeping with the revisionary attitude that Scheutz highlights.

Second, even if we need some kind of explanation to enable production, the explanation of consciousness in empirical terms may not require a solution to the Hard Problem. Through their empirical theories, scientists do not aim to explain, in some metaphysically intelligible way, how the properties of a phenomenon 'arise from' other properties at lower levels. Instead, they aim to establish a theoretical identity for the phenomenon in terms of its underlying properties (Block & Stalnaker 1999; McLaughlin 2003; Prinz 2003; Shea & Bayne 2010). (I say more about how this applies to consciousness science in the next section.) To build their theories, scientists draw correlations between levels, tying together some higher-level and lower-level properties. In the biological and psychological sciences, what requires this kind of explanation between levels depends on context: it is often determined by which properties, at higher or lower levels, appear anomalous (Wimsatt 1976; Craver 2009, ch. 6; Prinz 2012, 287-8). These practices suggest that an empirically successful theory

of consciousness need not fill in the gap between phenomenal and physical properties – at least, not in the terms defined by Chalmers' Hard Problem.

3 The multiplicity in phenomenal consciousness

I have shown how empirical methods and models of explanation can defuse philosophical challenges to the possibility of artificial consciousness. They allow us to counter intuitions drawn from thought experiments on the nature of consciousness, and to undercut arguments derived from our ignorance of how conscious states arise from physical states. By appealing to these empirical methods and models, we adopt a naturalistic approach to the study of artificial consciousness. We use empirical methods, as far as possible, to answer questions about the nature of consciousness and our access to it. We thereby allow empirical discoveries about phenomenal consciousness to inform our conceptual understanding of artificial consciousness. But that naturalistic approach produces a different philosophical challenge, arising from what we discover to be the multiplicity underlying consciousness. This new challenge to artificial consciousness is epistemological: it suggests that, even if we can build conscious machines, we cannot tell that the machines are conscious.

The challenge rests on our difficulty in determining the role of biological properties in phenomenal consciousness. Unless we determine their role, we cannot discover whether our machines, lacking at least some of these properties, are conscious. Several philosophers analyse this difficulty (Block 2002; Papineau 2002, ch. 7; Prinz 2003, 2005; Tye 2016, ch. 10). Yet their arguments are largely ignored by theorists of artificial consciousness. I will focus on Prinz's arguments – since they arise naturally from his work on an empirical theory of consciousness and are addressed directly to theorists of artificial consciousness.

Prinz begins by analysing, at the psychological level, the *contents* of our conscious states and the *conditions* under which they become conscious. Following Nagel, he considers having a perspective to be fundamental to consciousness: 'We cannot have a conscious experience of a view from nowhere' (2003, 118). In his analysis, humans experience the world, through our senses, 'from a particular vantage point'. So the contents of our consciousness are both perceptual and perspectival. These contents become conscious when we are paying attention. When these contents become available for our deliberation and deliberate control of action, they enable our flexible responses to the world. Putting together these hypotheses, Prinz proposes that consciousness arises in humans when we attend to phenomena such that our perspectival perceptual states become available for deliberation and deliberate control of action.

Next, by drawing on empirical studies, Prinz maps these contents and conditions of conscious states onto the computational and neural levels. In information processing, the contents of consciousness seem to lie at the intermediate level. Our intermediate-level representations are 'vantage-point specific and coherent' (2003, 119). They are distinct from higher-level representations which are too abstract to preserve perspective, and lower-level representations which are too local to be coherent. In computational models of cognition, attention is a process that filters representations onto the next stage, while deliberate control is handled by working memory, a short-

term storage capacity with executive abilities. In the human brain, these computational processes are implemented by a neural circuit between perceptual centres in the temporal cortex, attentional centres in the parietal cortex, and working memory centres in the frontal cortex (2003, 119; 2005, 388). Prinz (2012) cites several lines of evidence indicating that gamma vectorwaves play the crucial role in these brain regions. So, according to his latest theory, consciousness arises in us 'when and only when vectorwaves that realize intermediate-level representations fire in the gamma range, and thereby become available to working memory' (293). That is, in empirical terms, a good candidate for the neurofunctional basis of consciousness in humans.

Despite this progress, Prinz (2003, 2005) highlights an epistemological limitation, which is independent of whatever empirical theory of consciousness we settle on. He argues that we cannot determine if our biological properties are constitutive of consciousness. So we cannot discover if our machines, which will lack at least some of these properties, are conscious. This is the basis of his pessimism about research in strong artificial consciousness: 'It simply isn't the case that scientific investigations into the nature of consciousness will make questions of machine consciousness disappear. Even if scientific theories of consciousness succeed by their own standards, we must remain agnostic about artificial experience' (117).

Like others who share his pessimism, Prinz cites the in-principle failure of behavioural tests to settle these questions (Prinz 2003, IV; Block 2002; Papineau 2002, ch. 7, 2003). How do we find constitutive properties of consciousness? The standard method is to test for what Prinz calls 'difference-makers' (121). It involves changing processes at a tested level while keeping constant processes at other levels. If this change makes a difference to conscious behaviour in humans, then some properties at this tested level are constitutive of consciousness. Suppose that it is technically possible to substitute silicon chips for neurons in the human brain. And suppose that it is nomologically possible to do so while keeping constant the relevant processes at the psychological and computational levels.[2] This surgically altered person will become a 'functional duplicate of a normal person with a normal brain' (123). By design, the functional duplicate will behave exactly as conscious humans do – reporting pain, showing signs of anger, apparently 'seeing sunsets and smelling roses'. Yet our current tests for consciousness centre on behaviour. So we do not have a genuine test for consciousness in the duplicate. We cannot, by these tests, tell if our properties at the biological level are constitutive of consciousness.

I agree with Prinz (2003) that this thought experiment highlights a 'serious epistemological problem' (130). Indeed, I believe that he and others understate the depth of the problem. They focus on the failure of behavioural tests to discover if biological properties make a difference to conscious states. Prinz claims that this 'method of difference-makers seems to be the only way to find out what levels matter' (130). Yet, like other philosophers, he also recommends that we use inference to the

[2] This is a common idealisation in the thought experiment. In reality, we will find more than one psychological level and more than one computational one (Prinz 2003, 120-1). During chip replacement, we are more likely to keep constant processes at less fine-grained psychological and computational levels. The epistemological difficulty with testing remains, though it is made more complicated. Elsewhere, in Chin (2016), I analyse more complicated versions of the multiple-kinds problem in consciousness science; see also Irvine (2013), ch. 6.

best explanation to establish a theoretical identity for consciousness (116).[3] He does not explain why this theoretical inference cannot clarify the role of biological properties in consciousness and, thereby, improve the current tests for consciousness.

Let me make these connections explicit through the multiple-kinds problem shown in Figure 1. As the thought experiment suggests, we will discover at least two functional structures responsible for conscious behaviour in humans. One is a neurofunctional structure, such as that identified in Prinz's theory. Another is a functional structure that abstracts away from some biological mechanisms in the neurofunctional structure. Therefore, the kind defined by the neurofunctional structure ($kind_2$) is nested within the kind defined by the more abstract functional structure ($kind_1$). $Kind_1$ includes conscious humans and our functional duplicates, while $kind_2$ excludes the functional duplicates. So which is *the* structure of consciousness? Which structure defines a kind formed by all and only conscious beings?

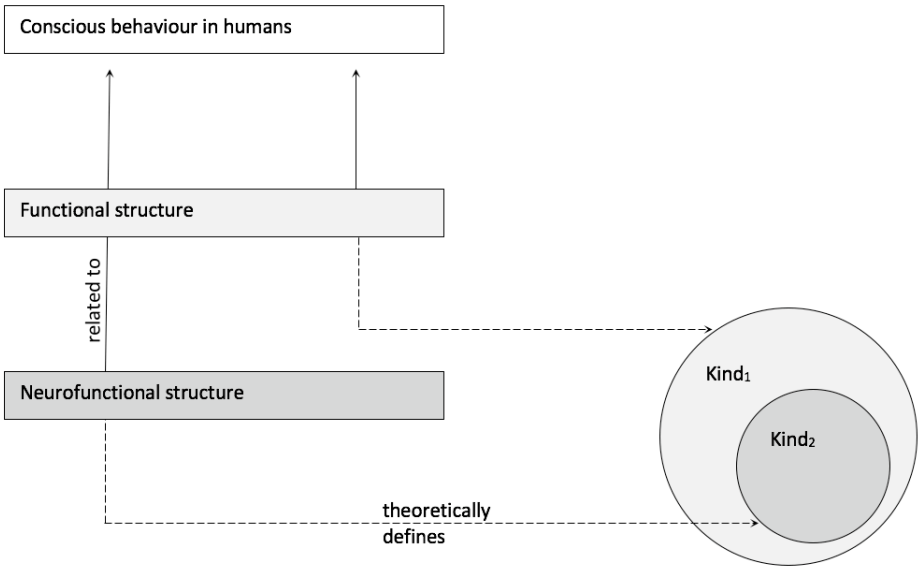

Fig. 1. The multiple kinds in phenomenal consciousness

Prinz's argument shows that current tests, based on behaviour, cannot solve this multiple-kinds problem. I want to extend this argument, to show why inference to the best explanation does not help. Both the neurofunctional structure and the more abstract structure are correlated with consciousness in humans. Both are also systematically related to conscious behaviour in humans. By focusing on the systematic relations between the neurofunctional structure, consciousness in humans, and their conscious behaviour, we can support an identity between consciousness and the neurofunctional structure. But this move is *ad hoc*, classifying our functional duplicates by fiat as not conscious. On the other hand, by focusing on the equally systematic relations between

[3] Other philosophers include Block and Stalnaker (1999), McLaughlin (2003), Shea and Bayne (2010), and Allen and Trestman (2016), §4.3.

the more abstract functional structure, consciousness in humans, and their conscious behaviour, we can support an identity between consciousness and that structure. Yet this is equally *ad hoc*, re-classifying the duplicates by fiat as conscious.

Neither hypothesis offers a simpler explanation. Whether we identify consciousness with the neurofunctional structure or the more abstract structure, we must invoke both structures to account for the total explananda. If we identify consciousness with the neurofunctional structure, then we must use the more abstract structure to explain why the duplicates share the same behaviour as humans even though the duplicates do not have human brains. If we identify consciousness with the more abstract structure, then we must use the neurofunctional structure to explain how the more abstract structure is implemented differently in conscious humans and their duplicates. The first hypothesis interprets consciousness as only one implementation of the more abstract structure, while the second interprets the neurofunctional structure as only one implementation of consciousness. So the familiar norms of explanatory simplicity do not help to choose between these hypotheses. That is why the multiple-kinds problem seems intractable. If we cannot solve this problem, then we cannot tell whether the biological properties that our machines lack are constitutive of consciousness. And, therefore, we cannot tell whether our machines are conscious.

4 The development of scientific taxonomies

I have shown why the naturalistic approach that defuses earlier philosophical challenges on artificial consciousness produces an epistemological puzzle on the role of biological properties in consciousness. Through empirical investigation, we will discover multiple functional structures underlying consciousness in humans. Neither behavioural tests nor theoretical inferences are able to pick out one structure from among them, in order to define a kind formed by all and only conscious beings. Unless we solve this multiple-kinds problem, we cannot determine whether the biological properties that our machines lack are constitutive of consciousness. In this section, I want to examine how other scientists develop more fine-grained taxonomies to manage their multiple kinds. Then I will evaluate how theorists of artificial consciousness can use this taxonomic strategy.

How does the multiple-kinds problem arise elsewhere? One prominent instance is what biologists call the 'species problem'.[4] When biologists try to classify organisms into species, they discover multiple structures underlying biodiversity. These structures centre on interbreeding, genetic or phenotypic similarity, ecological niche, evolutionary tendency, or phylogeny. They lead to conflicting definitions of what a species is. Different structures define overlapping kinds, consisting of different populations of organisms. According to the biologists Coyne and Orr (2004), at least nine species definitions remain 'serious competitors'. Three of them are often mentioned in the

[4] This problem is analysed by both biologists and philosophers: see the surveys in Coyne and Orr (2004); Cracraft (2000); Ereshefsky (2010, 2017); and Richards (2010). I also learnt from the analysis in LaPorte (2004), though we come to different conclusions. Richards (2010) argues that the problem goes back to pre-Darwinian times: Darwin himself was confronted by 'a multiplicity of species concepts' (75).

philosophical literature: the Biological Species Concept (BSC), the Phylogenetic Species Concept (PSC), and the Ecological Species Concept (ESC).[5] They focus, respectively, on three primary processes involved in evolution: sexual reproduction, descent from common ancestry, and environmental selection pressures. Of the three, which defines the nature of species?

Proponents of the BSC, the PSC, and the ESC sometimes claim that their definition of species is the 'best'.[6] But, in practice, biologists choose between these definitions according to their empirical interests. As de Queiroz (1999) explains, 'they differ with regard to the properties of lineage segments that they consider most important, which is reflected in their preferences concerning species criteria' (65). Their choice of the BSC, the PSC, or the ESC allows them to investigate the wider explanatory structures associated, respectively, with sexual reproduction, descent from common ancestry, or ecological niche. For instance, those who are interested in the history of life prefer the PSC over the BSC because they believe that reproductive isolation is 'largely irrelevant to reconstructing history' (Coyne and Orr 2004, 281). Those who are interested in the explanation of biodiversity reject the PSC because they see phylogeny as 'largely irrelevant to understanding the discreteness of nature'. Instead they use the BSC to study populations that sexually reproduce or use the ESC to study adaptive zones in ecology.

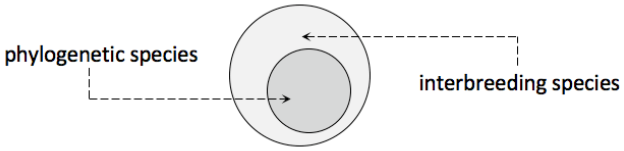

Fig. 2. Two overlapping kinds of biological species

The result is a more fine-grained taxonomy of species, which can be used to manage the multiple kinds found within biodiversity. Biologists now distinguish between species which arise from interbreeding, species which arise from phylogenetic connection, and species which arise from environmental selection (Ereshefsky 2010). As Figure 2 shows, the BSC and the PSC tend to define overlapping kinds of populations. When genealogically distinct populations can reproduce with each other, the populations of a phylogenetic species are nested within the populations of an interbreeding species. Through their taxonomy, biologists can clarify the relations between these kinds and demarcate the explanatory structures involving these kinds.

---

[5] The BSC defines species as 'groups of interbreeding natural populations that are reproductively isolated from other such groups' (Mayr 1969). The PSC defines them as the 'smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent' (Cracraft 1983). The ESC defines them as 'a lineage (or a closely related set of lineages) which occupies an adaptive zone minimally different from that of any other lineage in its range and which evolves separate from all lineages outside its range' (Van Valen 1976).

[6] As Cracraft (2000) warns, 'the notion of "best" is always relative' (10). He urges us to 'look hard at the context of what *best* might mean', including how general in application a definition is meant to be, and whether a more general definition is always more useful.

With the more fine-grained taxonomy in place, what matters to biological explanation is not whether the BSC or the PSC offers the 'best' definition of species. Rather biologists have to ensure that those who are interested in interbreeding species not confuse classifications with those who are interested in phylogenetic species. In a context with shared interests, such confusion is unlikely to arise. For instance, most biologists interested in sexual reproduction and its effects focus on interbreeding species. Their interests already pick out these relevant kinds from the overlapping ones associated with sexual reproduction, descent from ancestry, and environmental selection pressures. In a context with competing interests, biologists can avoid misunderstanding by making explicit reference to either interbreeding species, phylogenetic species, or ecological species. However, in some general contexts, biologists need not specify the kinds to which they refer. They may be keen to make generalisations across different branches of biology (Brigandt 2003). So their claims apply uniformly to interbreeding species, phylogenetic species, and ecological species.

The multiple-kinds problem also afflicts debates on animal consciousness. Here it lies closer to our epistemological puzzle on artificial consciousness. For animal consciousness, the problem arises because we discover at least two cognitive structures underlying consciousness in humans. Both structures are responsible, in different ways, for conscious behaviour in humans. I will follow how Godfrey-Smith (2016a, b) distinguishes these structures. The first involves simple modes of information processing associated with pain and other primitive bodily feelings, such as thirst and feeling short of breath. This structure enables us to respond to actual and potential injury with flexible non-reflexive behaviour. The second structure involves more sophisticated modes of information processing which integrate information from different senses and bodily feelings, through the use of memory, attention, and executive control. According to some theories of cognition, this structure allows us to model the world before responding to it.

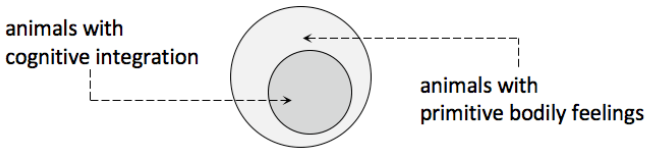

Fig. 3. Two overlapping kinds of animals

Figure 3 shows that these two cognitive structures define two overlapping kinds of animals. The kind of animals with cognitive integration is nested within the kind with primitive bodily feelings, because cognitive integration requires more machinery, such as memory, attention, and executive control. So which is *the* cognitive structure of consciousness? Which structure defines a kind formed by all and only conscious animals? If cognitive integration is necessary for consciousness, then only animals with memory, attention, and executive control count as conscious. But if primitive bodily feelings are sufficient for consciousness, many more animals count as conscious, so long as they have the sensorimotor capacities associated with primitive bodily feelings.

Faced with this multiple-kinds problem, Godfrey-Smith (2016a, b) proposes a more fine-grained taxonomy of subjective experiences in animals. There are at least two kinds of subjective experiences. The basic kind, which evolved first, consists of experiences of pain and other primitive bodily feelings; the complex kind, which evolved later, consists of experiences which integrate information from different senses and bodily feelings. Both kinds of subjective experiences are found in conscious humans: 'Much human experience does involve the integration of different senses, integration of the senses with memory, and so on, but there is also an ongoing role for what seems to be old forms of experience that appear as intrusions into more organized kinds of processing' (2016b, 500). Through his taxonomy, we can clarify the relations between both kinds of experiences and demarcate the explanatory structures involving both kinds.

With the more fine-grained taxonomy in place, we can see that what matters in the explanation of animal behaviour is not whether the basic or complex kind of subjective experiences counts as conscious. Rather theorists of animal consciousness can focus on either kind of experiences according to their empirical interests, so long as their terminology does not obscure the differences between both kinds. For instance, Godfrey-Smith classifies only experiences with cognitive integration as conscious: '"Consciousness" is something beyond mere subjective experience, something richer or more sophisticated' (2016a, 53). Animals which experience pain and other primitive bodily feelings have qualia; it feels like something to be them. But, without cognitive integration, they do not count for him as conscious: 'I wonder whether squid feel pain, whether damage feels like anything to them, but I do not see this as wondering whether squid are conscious' (2016b, 484). As he acknowledges, other theorists with different interests tend to equate qualia with phenomenal consciousness: 'If there is something it *feels like to be* a system, then the system is said to have a kind of consciousness' (483-4). In turn, these theorists have to distinguish phenomenal consciousness from other, more sophisticated, kinds of consciousness that require cognitive integration.

How might this taxonomic strategy address the epistemological puzzle on artificial consciousness? We can develop a more fine-grained taxonomy of conscious states, in order to manage the multiplicity that troubles theorists of artificial consciousness. If Prinz is right, then we need to distinguish at least two kinds of states. The first consists of neurofunctional states, such as those specified in his theory of consciousness. Our functional duplicates do not have this kind of states. The second consists of functional states that abstract away from some biological mechanisms in the neurofunctional states; both humans and the duplicates share this kind of states. With this taxonomy, we can clarify the relations between the neurofunctional and functional states, then demarcate the explanatory structures involving both kinds of states. What matters in explaining humans and duplicates is not whether the neurofunctional or functional states count as conscious. Rather theorists of consciousness can focus on either kind of states according to their empirical interests, so long as their terminology does not obscure the differences between both kinds of states. Those who classify only the neurofunctional states as conscious still need to acknowledge the role of the functional states, which explain why the duplicates behave in ways that indicate consciousness in humans. Those who classify the functional states as conscious still need to acknowledge the role of the neurofunctional states; they explain how the functional states are realised in humans.

This analysis brings out an epistemological difference between the case of biological species and that of artificial consciousness. Biologists are now confident that interbreeding species, phylogenetic species, and ecological species play significant explanatory roles. They know that the kinds associated with the BSC, the PSC, and the ESC are involved in different explanatory structures associated with sexual reproduction, ancestral descent, and ecological niche. In contrast, we do not yet know, in any precise terms, the states that will play significant explanatory roles in research on artificial consciousness. However, this difference does not invalidate our use of the taxonomic strategy. We need only begin with a provisional taxonomy of conscious states to explore the different explanatory structures that interest us. As we discover more about these explanatory structures, we can refine the taxonomy so that it reflects, in more precise terms, the computational and biological processes cited in our explanations. That is similar to how biologists developed their taxonomy for species.

Indeed, this taxonomic strategy can already make sense of some current models of artificial consciousness. Some theorists suggest that building the right computational processes into machines is sufficient to make them conscious. For instance, Dehaene, Lau, and Kouider (2017) propose that machines are conscious if they can select information for global broadcasting, making it flexibly available for computations, and if they can self-monitor those computations. To support their proposal, they claim that a machine with both computational processes will behave 'as though it were conscious' (492). They also cite evidence suggesting that subjective experience in humans 'appears to cohere with' global broadcasting and self-monitoring (492). Other theorists believe that building the right biological processes into machines is necessary to make them conscious. Haladjian and Montemayor (2016) connect consciousness to biological processes in humans that endow them with emotion and empathy. So, in their view, machines designed purely to compute with artificial intelligence will not have subjective experiences. According to Godfrey-Smith (2016b), machines can have subjective experiences only if they have some functional properties associated with 'living activity' (505). For him, these properties include the robustness and adaptability typical of complex biological systems in humans.

From our perspective, these models of artificial consciousness need not come into conflict. Rather we can see them as jointly clarifying the more fine-grained taxonomy of conscious states needed in research on artificial consciousness. On one hand, Dehaene, Lau, and Kouider (2017) are investigating the kind of states which are defined purely in computational terms without reference to biological mechanisms; in particular they are interested in the explanatory structures associated with global broadcasting and self-monitoring. On the other hand, Haladjian and Montemayor (2016) and Godfrey-Smith (2016b) are interested in another kind of states, defined partly in biological terms; they raise different difficulties for realising such states in machines.

5 Conclusion

In this paper, I assessed a trajectory in which multiplicity superseded impossibility in philosophical challenges to artificial consciousness. First, I tackled two earlier challenges which claim that phenomenal consciousness cannot arise, or cannot be built, in machines. The first challenge, from the nature of consciousness, is based on Block's

Chinese Nation thought experiment. The second challenge, from the explanation of consciousness, is based on Chalmers' Hard Problem. I showed how a naturalistic approach, appealing to empirical methods and models of explanation, can defuse these challenges. To discover if machines can be conscious, we should rely on theories of consciousness developed through empirical methods, rather than the intuitions about consciousness provoked by thought experiments. To explain consciousness in empirical terms, we need not supply a philosophically satisfying account of how phenomenal properties arise from physical ones.

Second, I explained why this naturalistic approach leads to an epistemological puzzle on the role of biological properties in phenomenal consciousness. Through empirical investigation, we will discover multiple functional structures underlying consciousness in humans. As several philosophers argued, behavioural tests cannot pick out one structure from among them, in order to define a kind formed by all and only conscious beings. I argued that inference to the best explanation cannot help too. If we cannot solve this multiple-kinds problem, then we cannot determine whether the biological properties that our machines lack are constitutive of consciousness. We also cannot determine whether these machines are conscious.

Third, I evaluated whether a taxonomic strategy used in other sciences can address this new challenge. To manage the overlapping kinds which they cite in explanations, theorists of biological species and animal consciousness develop more fine-grained taxonomies. I argued that, similarly, theorists of artificial consciousness can develop a more fine-grained taxonomy of conscious states, which distinguishes between the neurofunctional states specified in an empirical theory of consciousness and the functional states that abstract away from some biological mechanisms in the neurofunctional states. Such a taxonomy enables us to clarify the relations between both kinds of states and demarcate the explanatory structures involving both kinds. In addition, I argued that this taxonomic strategy helps to make sense of current models of artificial consciousness, including those which require only computational states and those which require partly biological states. We can interpret them as models for investigating different kinds of conscious states.

This strategy presents us with three related challenges, on the explanatory, subjective, and moral significance of the kinds in any new taxonomy. First, we need to establish that these kinds of states play significant explanatory roles in research on artificial consciousness. This is primarily an empirical challenge, depending on theorists of artificial consciousness to explore different explanatory structures that interest us. Second, we need to examine the subjective significance of these kinds of states. Thus far, we have construed a conscious state's phenomenal properties as capturing 'what it is like to be' in that state. But this construal does not help to discriminate what the multiple kinds mean in subjective terms. We may do so by investigating the capacities and interactions made possible by the underlying structures that define these kinds. For instance, some basic structures may support what it is like to be an artificial patient, while others may support what it is like to be an artificial agent. Third, we need to explore the moral significance of these kinds of states. In what ways do the artificial patients count as moral patients whose suffering we must ameliorate? In what ways do the artificial agents count as moral agents whose lives we must attend to?

Acknowledgements

I thank Arzu Gokmen, Michael Prinzing, and Kaine Yeo for their suggestions. Abhishek Mishra, Susan Schneider, Paul Schweitzer, and Alexandra Serrenti commented on the talk. This research was supported by an NUS Early Career Award.

References

Allen, C., & Trestman, M. (2016). Animal Consciousness. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Winter 2016. https://plato.stanford.edu/archives/win2016/entries/consciousness-animal/. Accessed 10 Jan 2018.

Bishop, J.M. (2009). Why Computers Can't Feel Pain. *Minds and Machines* 19(4): 507-516.

Block, N. (1978). Troubles with Functionalism. In N. Block, *Consciousness, Function, and Representation: Collected Papers, Volume 1* (2007), 63-101. Cambridge, MA: MIT Press.

Block, N. (1995). On a Confusion about the Function of Consciousness. In N. Block, *Consciousness, Function, and Representation: Collected Papers, Volume 1* (2007), 159-213. Cambridge, MA: MIT Press.

Block, N. (2002). "The Harder Problem of Consciousness." In N. Block, *Consciousness, Function, and Representation: Collected Papers, Volume 1* (2007), 397-433. Cambridge, MA: MIT Press.

Block, N., & Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *Philosophical Review* 108(1): 1-46.

Brigandt, I. (2003). Species Pluralism Does Not Imply Species Eliminativism. *Philosophy of Science* 70(5): 1305-1316.

Chalmers, D.J. (1995). Facing up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3): 200-219.

Chin, C. (2016). *Borderline Consciousness, Phenomenal Consciousness, and Artificial Consciousness: A Unified Approach*. University of Oxford DPhil thesis.

Coyne, J.A., & Orr, H.A. (2004). Speciation: A Catalogue and Critique of Species Concepts. In A. Rosenberg & R. Arp (eds.), *Philosophy of Biology: An Anthology*, 272-92. Oxford: Wiley-Blackwell.

Cracraft, J. (1983). Species Concepts and Speciation Analysis. In R.F. Johnston (ed.), *Current Ornithology*, 159-87. New York: Springer.

Cracraft, J. (2000). Species Concepts in Theoretical and Applied Biology: A Systematic Debate with Consequences. In Q.D. Wheeler & R. Meier, *Species Concepts and Phylogenetic Theory: A Debate*, 3-14. New York: Columbia University Press.

Craver, C.F. (2009). *Explaining the Brain*. Oxford: Oxford University Press.

Dehaene, S., Lau, H., Kouider, S. (2017). What Is Consciousness, and Could Machines Have It? *Science* 358(6362): 486-92.

Ereshefsky, M. (2010). Species, Taxonomy, and Systematics. In A. Rosenberg & R. Arp (eds.), *Philosophy of Biology: An Anthology*, 255-71. Oxford: Wiley-Blackwell.

Ereshefsky, M. (2017). Species. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2017. https://plato.stanford.edu/archives/fall2017/entries/species/. Accessed 10 Jan 2018.

Gamez, D. (2008). Progress in Machine Consciousness. *Consciousness and Cognition* 17(3): 887-910.

Godfrey-Smith, P. (2016a). Animal Evolution and the Origins of Experience. In D.L. Smith (ed.), *How Biology Shapes Philosophy: New Foundations for Naturalism*, 51-71. Cambridge: Cambridge University Press.

Godfrey-Smith, P. (2016b). Mind, Matter, and Metabolism. *Journal of Philosophy* 113(10): 481-506.

Haladjian, H.H., & Montemayor, C. (2016). Artificial Consciousness and the Consciousness-Attention Dissociation. *Consciousness and Cognition* 45(October): 210-25.

Holland, O., & Gamez, D. (2009). Artificial Intelligence and Consciousness. In W.P. Banks (ed.), *Encyclopedia of Consciousness*, 37-45. Oxford: Academic Press.

Irvine, E. (2013). *Consciousness as a Scientific Concept: A Philosophy of Science Perspective*. Dordrecht: Springer.

LaPorte, J. 2004. *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64(October): 354-61.

Maudlin, T. (1989). Computation and Consciousness. *Journal of Philosophy* 86(8): 407-32.

McDermott, D. (2007). Artificial Intelligence and Consciousness. In P.D. Zelazo, M. Moscovitch, & E. Thompson (eds.), *The Cambridge Handbook of Consciousness*, 117-50. Cambridge: Cambridge University Press.

McGinn, C. (1991). *The Problem of Consciousness: Essays Towards a Resolution*. Oxford: Blackwell.

McLaughlin, B.P. (2003). A Naturalist-Phenomenal Realist Response to Block's Harder Problem. *Philosophical Issues* 13(1): 163-204.

Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Clarendon Press.

Prinz, J. (2003). Level-Headed Mysterianism and Artificial Experience. In O. Holland (ed.), *Machine Consciousness*, 111-32. Exeter: Imprint Academic.

Prinz, J. (2005). A Neurofunctional Theory of Consciousness. In A. Brook & K. Akins (eds.), *Cognition and the Brain: The Philosophy and Neuroscience Movement*, 381-96. Cambridge: Cambridge University Press.

Prinz, J. (2012). *The Conscious Brain: How Attention Engenders Experience*. Oxford: Oxford University Press.

Queiroz, K. de. (1999). The General Lineage Concept of Species and the Defining Properties of the Species Category. In R.A. Wilson (ed.), *Species: New Interdisciplinary Essays*, 49-89. MIT Press.

Reggia, J.A. (2013). The Rise of Machine Consciousness: Studying Consciousness with Computational Models. *Neural Networks* 44(August): 112-31.

Richards, R.A. (2010). *The Species Problem: A Philosophical Analysis*. Cambridge: Cambridge University Press.

Scheutz, M. (2014). Artificial Emotions and Machine Consciousness. In K. Frankish & W.M. Ramsey (eds.), *The Cambridge Handbook of Artificial Intelligence*, 247-66. Cambridge: Cambridge University Press.

Searle, J.R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3(3): 417-57.

Shea, N., & Bayne, T. (2010). The Vegetative State and the Science of Consciousness. *British Journal for the Philosophy of Science* 61 (3): 459-84.

Tye, M. (2016). *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* New York: Oxford University Press.

Valen, L.V. (1976). Ecological Species, Multispecies, and Oaks. *Taxon* 25(2/3): 233-39.

Wimsatt, W.C. (1976). Reductionism, Levels of Organization, and the Mind-Body Problem. In G.G. Globus, G. Maxwell, & I. Savodnik (eds.), *Consciousness and the Brain*, 205-67. Dordrecht: Springer.