mark.sprevak@ed.ac.uk

Final version due to appear in R. Sun (Ed.) *Cambridge Handbook on Computational Psychology* , Cambridge: Cambridge University Press

Philosophical issues in computational cognitive sciences

Mark Sprevak
University of Edinburgh

19 May 2021

1 Introduction

In 1962, Wilfred Sellars wrote: 'The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term' (Sellars, 1962, p. 35). Sellars was concerned to explain the relationship between philosophy and the sciences. In my view, he got it exactly right. Philosophical issues are marked out not by having some uniquely philosophical subject matter, but in terms of the overall scope of the enquiry. When one turns to philosophical issues, what one is doing is taking a step back from some of the details and considering broadly how matters hang together relative to the ambitions and goals that motivated the scientific enquiry in the first place. In the case of the computational cognitive sciences, this may involve asking such questions as: Are there aspects of cognition or behaviour that are not amenable to computational modelling? How do distinct computational models of cognition and behaviour fit together to tell a coherent story about cognition and behaviour? What exactly does a specific computational model tell (or fail to tell) us about cognition and behaviour? What distinguishes computational models from alternative approaches to modelling cognition and behaviour? How does a computational model connect to, and help to answer, our pre-theoretical questions about what minds are and how they work?

Progress in answering these questions may come from any or all sides. It would be a mistake to think the philosophical issues are somehow only the domain of

philosophers. Anyone who takes computational modelling seriously as an attempt to study cognition is likely to want to know the answer to these questions and is liable to be able to contribute to the project of answering them. What philosophers bring to the project is a set of conceptual tools and approaches that have been developed in other domains to address structurally similar issues. They also have the luxury of being allowed to think and write about the big questions.

Sellars had a relatively narrow conception of what it meant to understand how things hang together. He interpreted this as an attempt to reconcile two distinct images of the world: the *scientific image* (which describes relationships between the posits of the sciences – e.g. cells, molecules, atoms, forces) and the *manifest image* (which describes relationships between the posits of human common-sense understanding of the world – e.g. persons, thoughts, feelings, ideas) (Sellars, 1962). This chapter adopts a somewhat looser and broader interpretation of the project. Models in the computational cognitive sciences are often partial, provisional, and one of many possible alternatives that consistent with the data. It would be misleading to assume that current computational cognitive science contains a single, coherent account or embodies 'the' scientific image of cognition. Similar concerns could be raised about our manifest image in light of work on cross-cultural variation in human folk understanding and conceptualisations of the world (Barrett, 2020; Henrich, Heine and Norenzayan, 2010; Nisbett, 2003). The view adopted in this chapter is that one's goal should be to understand how the many, diverse current approaches to computational modelling of cognition hang together, both with each other, with work in the other sciences (including neuroscience, cellular biology, evolutionary biology, and the social sciences), and with the various pre-theoretical folk questions and insights we have regarding the mind. Here, there is no prior commitment to a single, well-defined scientific image or manifest image, but rather the ambition to understand how the various perspectives we have on cognition and behaviour cohere and allow us to understand what minds are and how they work (for more on this point, see Sprevak, 2016).

1.1 Overview of chapter

Researchers generally aim to build a computational model of some limited domain within cognition or behaviour (e.g. face recognition, cheater detection, word segmentation, or depth perception). Splitting up human cognition and behaviour into various smaller domains raises questions about *how* we should do it. In the language of philosophers, this is about how we should *individuate* our cognitive capacities and behaviour (M. L. Anderson, 2014; Barrett and Kurzban, 2006; Machery, forthcoming). Modelling cognition and behaviour in this way also raises questions about how the models of individual capacities we hope to obtain will eventually be woven

together to create a coherent, integrated model of cognition. This concerns the issue of how we *unify* models of distinct cognitive domains (Colombo and Hartmann, 2017; Danks, 2014; Eliasmith, 2013).

This chapter focuses on a separate set of foundational issues that are related, but posterior to, the two just mentioned. These concern possible *gaps* left by this strategy for modelling cognition. If the strategy were completed, would there be any cognitive capacities that would be missing from the final picture? Are there any aspects of cognition for which we should expect *not* to obtain a computational model? Are there certain cognitive domains that are, for some reason, 'no go' areas for computational modelling? The chapter examines three possible candidates: *semantic content* (Section 2), *phenomenal consciousness* (Section 3), and *central reasoning* (Section 4). In each case, philosophers have argued that there are good reasons to believe that we cannot obtain an adequate computational model of the domain in question.

2 Semantic content – Searle's Chinese room argument

John Searle's Chinese room argument is one of the oldest and most notorious 'no go' arguments for computational modelling of cognition. The precise extent of its target has been liable to shift between different presentations of the argument. Searle has claimed in various contexts that the argument shows that *understanding*, *semantic content*, *intentionality*, and *consciousness* cannot adequately be captured by a computational model (according to Searle, all these properties are linked, see Searle, 1992, pp. 127–197). In his original formulation, Searle's target was *understanding*, and specifically, our ability to understand simple stories. He considered whether a computational model would adequately be able to account for this cognitive capacity. More precisely, he considered whether such a model would be able to explain the difference between understanding and not understanding a simple story (Searle, 1980; cf. models of understanding in Schank and Abelson, 1977; Winograd, 1972).

2.1 The Chinese room argument

Searle's argument consisted in a thought experiment concerning implementation of the computation. Imagine a monolingual English speaker inside a room with a rule-book and sheets of paper. The rule-book contains instructions in English on what to do if presented with Chinese symbols. The instructions might take the form: 'If you see Chinese symbol X on one sheet of paper and Chinese symbol Y on another, then write down Chinese symbol Z on a third sheet of paper'. Pieces of paper with Chinese writing are passed into the room and the person inside follows the rules and passes pieces of paper out. Chinese speakers outside the room label

the sheets that are passed in 'story' and 'questions' respectively, and the sheets that come out 'answers to questions'. Imagine that the rule-book is as sophisticated as you like, and certainly sophisticated enough that the responses that the person gives are indistinguishable from those of a native Chinese speaker. Does the person inside the room thereby understand Chinese? Searle claims that they do not (for discussion of the reliability of his intuition here, see Block, 1980; Maudlin, 1989; Wakefield, 2003).

Searle observes that the Chinese room is a computer, and he identifies the rule-book with the (symbolic) computation that it performs. He then reminds us that the thought experiment does not depend on the particular rule-book used: it does not matter how sophisticated the rule-book, the person inside the room would still be shuffling Chinese symbols without understanding what they mean. Since any symbolic computational process can be described by some rule-book, the thought experiment shows that the person inside the Chinese room will not understand the meaning of the Chinese expressions they manipulate no matter which symbolic computation they perform. Therefore, we can conclude that the performance of a symbolic computation is insufficient, by itself, to account for the difference between the system performing the computation understanding and not understanding what the Chinese expressions mean. Searle infers from this that any attempt to model understanding purely in terms of a formal, symbolic computation is doomed to failure. According to Searle, the reason why is that a formal computational model cannot induce *semantic* properties, which are essential to accounting for a semantically laden cognitive process like understanding (Searle, 1980, p. 422).

2.2 The problem of semantic content

Many objections have been raised to Searle's Chinese room argument (for a summary, see Cole, 2020). However, it is notable that despite the argument's many defects, the basic conclusion that Searle drew has been left largely unchallenged by subsequent attacks. This is that *manipulation of formal symbols* is insufficient to generate the semantic properties associated with cognitive processes like understanding. In Searle's terms, the Chinese room thought experiment is an illustration of a general principle that 'syntax is not sufficient for semantics' (Searle, 1984). Note that 'syntax' here does not refer to the static grammatical properties of symbols or well-formedness of linguistic expressions, but to the algorithmic rules by which symbolic expressions are manipulated or transformed during a computation. 'Semantics' refers specifically to the denotational aspects of the meaning associated with symbolic expressions – their intentional properties, or what they refer to in the world.

Searle is not alone in making this kind of claim. Putnam (1981) argued that ma-

nipulating symbols (mere 'syntactic play') cannot determine what a computation's symbols refer to, or whether they carry any referential semantic content at all (pp. 10–11). Burge (1986), building on earlier work by Putnam and himself on referring terms in natural language, noted that a physical duplicate of a computer placed in a different environment may undergo exactly the same formal transitions, but have entirely different meaning attached to its symbolic expressions based on its causal relationship to different environmental properties. Fodor (1978) described two physically identical devices that undergo the same symbol-manipulation procedures, one of which runs a simulation of the Six-Day War (with symbols referring to tank divisions, jet planes, and infantry units) and the other runs a simulation of a chess game (with symbols referring to knights, bishops, and pawns). Harnad (1990) argued that all computational models based on symbol processing face a 'symbol grounding' problem: although some of their symbols may have their semantic content determined by their formal relationship to other symbols, that sort of process has to bottom out somewhere with symbolic expressions that have their meaning determined in some other way (e.g. by their non-formal relationship to external objects in the environment in perception or motor control).

These considerations are also not confined to symbolic computational models of cognition. Similar arguments could be made for computational models that are defined over numerical values or probabilities. Consider artificial neural network models. These models consist in collections of formal nodes and connections that chain together a long sequence of mathematical operations on numerical activation values or connection weights (adding, multiplying, thresholding values). What do these numerical activation values or connection weights mean? How do are they relate to distal properties and objects in the environment? As outside observers, we might *interpret* the numerical values inside an artificial neural network as referring to certain things (just as outside observers we might interpret certain symbolic expressions in a classical, symbolic computation as referring to certain things). Independent of our interpreting attitudes, however, the mathematical rules that define a connectionist model do not fix this semantic content. The rules associated with an artificial neural network describe how numerical values are transformed during a computation (during inference or learning), but they do not say what those numbers (either individually or taken in combination) represent in the world. The numerical rules that define an artificial neural network no more imbue it with semantic content than the symbolic rules that operate over expressions do for a classical, symbolic computation. Models that operate over probabilities or probability distributions face a similar kind of problem. These computational models are typically defined in terms of numerical operations on probability distributions (understood as ensembles of values that satisfy the requirements for a measure of probability). These distributions might be interpreted by us, as external observers, as

probabilities of certain distal events occurring, but the mathematical rules governing the transformation of these distributions do not usually, by themselves, determine what those distal events are.

It is worth emphasising that there is no suggestion here that computation and semantic content are entirely independent factors in human cognition. Arguably, some symbolic expressions can get their meaning fixed through their formal computational role (most plausibly, this is the case for expressions that represent the logical connectives like AND and OR). However, even the most enthusiastic proponents of conceptual role semantics or procedural semantics do not, or at least do not normally, suggest that *all* semantic content is determined in this way. An adequate account of semantic content will need to include, not only formal relationships among computational states, but also non-formal relationships between those computational states and distal states in the external environment (for discussion of this point, see Block, 1986; Harman, 1987; Johnson-Laird, 1978).

2.3 Theories of content

A lesson that philosophers have absorbed from this is that a computational model of cognition will need to be supplemented by another kind of model in order to adequately account for cognition's semantic properties. Modelling cognition should therefore be seen as a project with at least two branches. One branch consists in describing the formal computational transitions or functions associated with a cognitive or behavioural capacity. The other branch connects the abstract symbols or numerical values described in the first branch to distal objects in the environment via semantic relations (see Chalmers, 2012, pp. 334–335). This two-pronged approach is perhaps best represented by the research programme of Jerry Fodor. Fodor argued that one should sharply distinguish between one's *computational theory* (which describes the dynamics of abstract computational vehicles in cognition) and one's *theory of content* (which describes how those vehicles get associated with specific distal representational content). It would be a mistake to think that one's computational theory can determine one's semantic properties (or vice versa) (for a helpful summary of this research programme, see Fodor, 1998, pp. 9–12). (Fodor makes exactly this point in response to the Chinese room argument, see Fodor (1980)).

What does a theory of content look like? Fodor argued that it should aim to answer two questions: (S1) How do computational states get their semantic properties? (S2) Which specific semantic content do they have? Fodor suggested any answers suitable to cognitive science should be *naturalistic*. By this, he meant that answers to questions S1 and S2 should appeal to properties and relations that are not themselves semantic or intentional. They should explain how semantic content arises, and how

specific semantic contents get paired with computational states, in terms of the kinds of non-semantic properties that are typically invoked in the natural sciences (e.g. physical relations between the brain and its environment). A theory of content should not, for example, attempt to answer S1 or S2 by appeal to further semantic or mental properties, such as the interpreting attitudes of us as external observers or the intentional mental states of the subject themself (Fodor, 1990, p. 32; Loewer, 2017).

Fodor developed his own naturalistic theory of content, called the 'asymmetric dependency theory'. This theory of content claimed that semantic content is determined by certain law-like relationships between symbols in the cognitive agent and current environmental stimuli (Fodor, 1990). Teleological theories of content attempt to naturalise content by appeal, not to conditions involving current environmental stimuli, but to conditions that were rewarded during past learning, or that were selected for in the cognitive agent's evolutionary history (Dretske, 1995; Millikan, 2004; Papineau, 1987; Ryder, 2004). Use-based theories of content attempt to naturalise content by appeal to structural isomorphisms between multiple computational states in the cognitive agent and states of the world, claiming that their structural correspondence accounts for how the computational states represent (Ramsey, 2007; Shagrir, 2012; Swoyer, 1991). Information-theoretic theories of content attempt to naturalise content by appeal to Shannon information (Dretske, 1981); recent variants propose that semantic content is determined by whatever distal state of affairs maximises a measure of mutual Shannon information between that distal content and the computational state (Isaac, 2019; Skyrms, 2010; Usher, 2001) – this echoes statistical methods used in cognitive neuroscience by external observers to assign content to neural responses in the sensory or motor systems (Eliasmith, 2005; Rolls and Treves, 2011; Usher, 2001). Shea (2018) provides a powerful new naturalistic theory of content which weaves together elements of all the approaches above, arguing that semantic content is determined by different sorts of naturalistic conditions in different contexts.

Naturalising semantic content is an aspiration rather than a completed goal and serious difficulties confront all current approaches. Common challenges that face current theories include allowing for the possibility of misrepresentation; avoiding unacceptably large amounts of indeterminacy in the ascribed semantic content; and providing a sufficiently general account that covers not only representations involved in perception and motor control, but also non-sensorimotor representations like DEMOCRACY, TIMETABLE, or QUARK (for an overview of the problems, see Adams and Aizawa, 2021; Neander and Schulte, 2021; Shea, 2013).

Some philosophers have suggested that we adopt an alternative approach to accounting for semantic content. Egan (2014) argues that we should assume, as a working

hypothesis, that the semantic content associated with cognition will never be naturalised. This is not because of any spookiness associated with semantic content, but because ascription of semantic content is an inherently messy matter that is influenced by endless, unsystematisable pragmatic concerns (Chomsky, 1995; Egan, 2003). There is unlikely to be a natural science of semantic content determination for similar reasons as there is unlikely to be a natural science of rude jokes. Egan suggests that ascriptions of semantic content, although they fall outside the domain of natural science, nevertheless play an auxiliary role in scientific explanation by functioning as an 'intentional gloss' that relates formal computational models to our informal, non-scientific descriptions of behavioural success and failure.

A different approach suggests that assignments of semantic content should be treated as a special kind of scientific fiction or idealisation in computational cognitive science (Chirimuuta, forthcoming; Mollo, forthcoming; Sprevak, 2013). This builds on recent philosophical work that emphasises the critical role of idealisations, fictions, and felicitous falsehoods in science, and that these features are often not undesirable and impossible to disentangle from literal truth telling (Elgin, 2017; Morrison, 2014; Potochnik, 2017).

Philosophers disagree on many points regarding how to model semantic content, but there is a broad consensus that a computational theory will need to be supplemented by something else – whether that be a naturalistic theory of content, an intentional gloss, or a reinterpretation of scientific practice – that accounts for how the states subject to computational rules gain their semantic content. This completes our description of one aspect of cognition for which a purely computational model is unlikely to be sufficient.

2.4 Content and physical computation

An important wrinkle has not yet been mentioned. The preceding discussion adopted the assumption that a computational model is defined exclusively in terms of formal rules (whether those be symbolic or numerical rules). This fits with how mathematicians and theoretical computer scientists talk about their computational models. In these contexts, a computational model is a purely abstract, mathematically definable entity (e.g. a set-theoretic entity). However, it does not reflect how many researchers in science and engineering talk about computational models. In these contexts, a computational model is often tied to its embodiment in a physical system. Part of one's goal in proposing a computational model is to suggest that the formal transitions are implemented in some specific physical system. In the case of computational cognitive science, these formal transitions are normally assumed to be implemented, at some temporal scale, in the cognitive agent's physical behaviour or neural responses.

When a formal computation is physically implemented, the physical states that are manipulated will stand in some non-formal relations to distal entities in the world. Physically implemented computations are inherently connected to the rest of the world: their computational states stand in law-like causal relations to objects in their environment, they have a history (and one that might involve past learning and evolution). Given this, it is by no means obvious that a physically implemented computation, unlike a purely formal abstract computation, is silent about, or does not determine, assignment of semantic content. Understanding whether and how physical computations and semantic content are related is a substantial question and one that is distinct from those considered above (for various proposals about the relationship between physically implemented computation and semantic content, see Dewhurst, 2018; Lee, 2018; Mollo, 2018; Piccinini, 2015, pp. 26–50; Rescorla, 2013; Shagrir, 2018; Sprevak, 2010). At the moment, there is no consensus about whether, and to what extent, physical implementation constrains the facts about a computation's semantics. In light of this, Searle's observation that 'syntax is not sufficient for semantics', even if valid for the formal cases he had in mind, may have questionable lasting significance for computational cognitive science (Boden, 1989; Chalmers, 1996, pp. 326–327; Dennett, 1987, pp. 323–326)

3 Consciousness – The hard problem

'Consciousness' may refer to many different kinds of cognitive phenomena, including sleep and wakefulness, self-consciousness, reportability, information integration, and allocation of attention (see van Gulick, 2018, for a survey). This section focuses exclusively on 'no go' arguments concerning *phenomenal consciousness*. 'Phenomenal consciousness' refers to the subjective, qualitative feelings that accompany many aspects of cognition. When you touch a piece of silk, taste a raspberry, or hear the song of a blackbird, over and above various processes of classification, judgement, report, attentional shift, control of behaviour, and planning, you also undergo subjective sensations. There is something it subjectively *feels like* to do these things. Some philosophers reserve the term 'qualia' to refer to these qualitative feelings (Tye, 2018). The 'hard problem' of consciousness is to explain why these feelings accompany cognition and to account for their distribution across our cognitive life (Chalmers, 1996, pp. 3–31; Chalmers, 2010b).

3.1 The conceivability argument against physicalism

The conceivability argument is a 'no go' argument against any physicalist theory of phenomenal consciousness. It is usually phrased in terms of the conceivability of zombies. A zombie is a hypothetical being who is a physical duplicate of a human and who lives in a universe that is a physical duplicate of our universe –

a world with exactly the same physical laws and exactly the same distribution of physical properties. The only difference between our world and a zombie world is that the relevant entities in the zombie world either lack conscious experience or have a different distribution of phenomenal experiences across their cognitive life from our own. A zombie's cognitive processing either occurs 'in the dark' or it is accompanied by different phenomenal experiences from our own (e.g. it might experience the qualitative feeling we associate with tasting raspberries when it sees tastes blueberries and vice versa).

It does not matter to the conceivability argument whether such a zombie could come into existence in our world, has ever existed, or is ever likely to exist. What matters for the argument is only whether one can rationally *conceive* of such a being. Can one imagine a physical duplicate who lacks phenomenal consciousness, or who has a different distribution of phenomenal experiences from our own? Many philosophers have argued that zombies of both kinds are easily conceivable (Chalmers, 1996, pp. 96–97; Kripke, 1980, pp. 144–155; Nagel, 1974). By this, they don't meant that zombies could exist in our own world, or that we should entertain doubts about whether other humans are zombies. Indeed, for many dualists it is impossible for a zombie to exist in our world. What they mean is that the *idea* of a zombie is a coherent one – it does not contain a contradiction, like the idea of a married bachelor or the highest prime number.

The next step in the argument relies on the assumption that rational conceivability is a reliable guide to possibility. If a world with zombies is conceivable, then we should believe, barring evidence to the contrary, that it corresponds to a genuine possibility. But if a zombie world is possible, then the physical laws and physical states could be the same as they are in our world and the relevant beings in that world either lack phenomenal experience or have different phenomenal experiences. That means that in our *actual* world, there must be some further ingredient, over and above the physical facts, that is responsible for our phenomenal experience. Something beyond the physical facts must account for the difference between our world and a zombie world. The existence of phenomenal consciousness and its distribution across our cognitive life cannot rest on the physical facts alone, because those facts could have been the same and the phenomenal experiences vary. Advocates of the conceivability argument conclude that a purely physicalist theory is unable to answer the hard problem of consciousness (Chalmers, 1996, pp. 93–171; Chalmers, 2010d). (This conclusion is shared by Jackson (1982)'s knowledge argument, based around a thought experiment with Mary, a neuroscientist who has never seen colour; the reasoning behind the knowledge argument is closely related to that of the conceivability argument; for the connection, see Chalmers (2010d), pp. 192–196).

According to the conceivability argument, no physicalist theory can answer two

important questions about phenomenal consciousness: (C1) How does phenomenal conscious experience arise at all? (C2) Why are our phenomenal conscious experiences distributed in the way that they are across our cognitive life? No matter which physical facts one cites about the brain or environment in response to these questions, none (either singly or jointly) entail that conscious experience occurs – for it is possible that the same physical facts could have obtained and those conscious experiences been absent or different (as they are in a zombie world). This raises the question of what, over and above the physical facts, is responsible for the existence and distribution of our phenomenal experiences. Advocates of the conceivability argument propose various remedies at this point, all of which involve expanding or revising our current physicalist scientific ontology. Our focus will not be on those remedies, but only on the negative point that phenomenal consciousness is a 'no go' area for physicalist theories (see Chalmers, 2010a, pp. 126–137, for a survey of non-physicalist proposals).

3.2 The conceivability argument against computational functionalism

The conceivability argument against physicalism has been adapted to create a 'no go' argument against computational theories of phenomenal consciousness.

The relevant consideration here that a zombie who is our *computational* duplicate is conceivable – an entity who performs the same computations as we do but who either lacks conscious experience entirely, or who has a different distribution of conscious experiences. Similar motivations apply as in the case of the original conceivability argument against physicalism. One might imagine a system implementing any computational process, or computing any function, but for this to fail to be accompanied by a phenomenal experience, or for it to be accompanied by a phenomenal experience different to our own. No matter how complicated the computation, nothing about performing a computation seems to *entail* anything about the existence or distribution of subjective experiences. One might imagine an exact computational duplicate of a human – a system who undergoes the same computational transitions – but whose cognitive life remains 'all dark' inside, or who has different subjective experiences (for detailed examples of such thought experiments, see Block, 1978; Dennett, 1978; Maudlin, 1989). In this respect, phenomenal consciousness appears different from other cognitive capacities. Cognitive capacities like classification, planning, and motor control, which do lend themselves to computational modelling, are defined by their *function* – by what they do, and how they contribute to behaviour. Phenomenal consciousness is not defined by its function, but by its *experiential feel* (Chalmers, 2010b, pp. 6–9). As with the original argument, it does not matter whether a computational zombie could exist in our world. What matters is only whether this type of zombie is rationally conceivable.

A separate consideration is that the original conceivability argument against physicalism entails a 'no go' claim concerning any attempt to solve the hard problem by appeal to physical computations (Chalmers, 1996, p. 95). Plausibly, any world that is a physical duplicate is a world that is a duplicate in terms of the physical computations that are performed. The physical facts about a world – e.g. about brains, their environment, and the laws that relate them – should be sufficient to determine which physical computations are implemented. If one accepts the conceivability argument against physicalism, then it is possible for a physical duplicate of our own world to exist in which phenomenal consciousness is absent or distributed differently. Therefore, it is possible for a world that is a duplicate in terms of the physical computations performed to exist in which phenomenal consciousness is absent or differently distributed. Hence, in our actual world there must be some additional ingredient, above and beyond the physically implemented computations, that accounts for phenomenal consciousness. No theory that appeals to physical computations alone can be sufficient to explain existence and distribution of our phenomenal experiences.

3.3 Naturalistic dualism

Advocates of the conceivability argument are careful about the scope of their 'no go' claim. What is claimed is that *solving the hard problem* by appeal to physical or computational facts is not possible. No physical or computational theory can answer C1 or C2. This does not mean, however, that a physical or computational theory cannot answer other important questions about phenomenal consciousness.

Chalmers (2010b; 2010c) suggests that a computational or physical theory can tell us a great deal about *correlations* between the physical/computational facts and phenomenal facts. The conceivability argument does not deny that such correlations exist or that they hold reliably. The important point, however, is that any information that a physical/computational model of phenomenal consciousness provides is at best correlational. Such a model cannot solve the hard problem of consciousness, because it is possible for a world to be physically identical and for those correlations to fail to obtain.

An analogy might help to clarify this point. Suppose that one were to engage in a study of the phenomena of lightning and thunder that is purely correlational. One might build a computational model of the phenomena that describes instances of each and correlations between them. In a similar fashion, one might embark on a study of physical states and phenomenally conscious states and attempt to describe the correlations between them. In both cases, what would be missing is a full understanding of how the relevant variables are linked. Lightning regularly co-occurs with thunder, but no pattern of lightning occurrences entails an occurrence

of thunder, and vice versa. In the case of lightning and thunder, this deficiency in a purely correlational model can be remedied by introducing extra physical variables – e.g. distributions of electrical charges in the air, ground potential, measurements of air density. In a suitably enlarged model – one that contains more physical variables and charts the relationships between them – it would be possible to see why the observed correlations between lightning and thunder obtain and exactly how and why they fail – i.e. how changes to one physical variable *necessitate* changes to the other. In the case of phenomenal consciousness, the conceivability argument aims to show that this is not possible. One must go entirely *outside* the realm of physical variables to account for the correlation between physical states and phenomenal states. Filling in the gap between brain processes and phenomenal experience cannot be done by introducing extra physical variables into one's model (or by introducing more complex physical relationships between variables). No matter how many physical variables one introduces, none necessitates phenomenal experiences – for all those physical variables could be the same and the consciousness experience absent or different. A physical/computational model of consciousness can only provide us with the common physical/computational correlates of phenomenal conscious, not an explanation of how or why that experience occurs.

3.4 Eliminativism and related replies

Not all philosophers accept the reasoning behind the conceivability argument. Dennett argues that human conceptual and imaginative capacities have not evolved to draw reliable conclusions about hypothetical zombies. For all we know, zombie thought experiments work on our imaginations a little like viewing an M.C. Escher drawing: we appear to see something remarkable, but only because we have failed to spot some contradiction hidden in the picture. Dennett suggests that the correct inference to make is not that we have established a hard a priori limit to what a physical/computational model of phenomenal consciousness can achieve, but that the entire project of trying to set a hard a priori limit on what a physical/computational model can achieve is deeply misconceived (Dennett, 2013). It could be that a truly thorough, mature conceptualisation of a physical and computational duplicate of ourselves, imagined down to the smallest detail, would rule out the possibility that it could be a zombie (Dennett, 1995; Dennett, 2001).

Dennett's doubts about the reliability of our intuitions in zombie thought experiments may temper enthusiasm for the proposition that zombies are conceivable, but this by itself does not rescue physicalism. In order to do this, Dennett also commits to the a speculative, positive claim that *if* we were to successfully wrap our heads around the right computational (and/or physical) relationships that correlate with consciousness, then we would see that they *must* bring all aspects of consciousness

along with them. Advocates of the conceivability argument, while typically open to the idea that zombie intuitions are defeasible (we might be deluding ourselves about the conceivability of a zombie), tend to pour scorn on this latter contention. No matter how complex a computational model is they say, it simply isn't clear how its operation could necessitate conscious experiences (Strawson, 2010). The idea that somewhere out there, in the landscape of all possible computational models, a series of operations exists that magically necessitates conscious experience is pure moonshine or physicalist dogma (Strawson, 2018).

A position one might be driven towards, and which Dennett defends in other work, is that certain aspects of consciousness – the irreducibly private, first-person aspects targeted by zombie thought experiments – are not real. This amounts to a form of eliminativism about phenomenal consciousness (Irvine and Sprevak, 2020). Such positions face a heavy intuitive burden. Our subjective feelings are among the things we are most certain exist. Denying their reality may strike one as unacceptable. Nevertheless, the predictive and explanatory benefits offered by past scientific theories have prompted us to abandon other seemingly secure assumptions about the world. If it can be shown that we are somehow labouring under an illusion concerning phenomenal experience, then many of the difficulties posed by the conceivability argument to computational modelling would evaporate. If there were no such thing as phenomenal consciousness, then there would be nothing for a computational model to explain.

However, in addition the intuitive burden just mentioned, a further difficulty faces eliminativist approaches. This is to explain how the claimed illusion associated with phenomenal consciousness works. This is called the 'illusion problem' (Frankish, 2016). Some eliminativists claim that the illusion problem can be solved by appeal to some physical/computational story about the mechanisms of our internal information processing and self report (Clark, 2000; Dennett, 1991; Frankish, 2016; Graziano, 2016). However, while such an account might be able to explain why we *believe* or *act* as if we had phenomenal consciousness, it is not clear how it can generate the felt first-person illusion of consciousness (Chalmers, 1996, pp. 184–191). In general, it is not clear how any physical or computational process can generate the felt first-person illusion of conscious experience, any more than it can generate the felt first-person reality of conscious experience. The illusion problem may in the end be no easier for a physicalist/computational theory to solve than the original hard problem of consciousness (Prinz, 2016).

4 Central reasoning – The frame problem

A third major target for philosophical 'no go' arguments is *central reasoning*. This concerns our ability to engage in reliable, general-purpose reasoning on a large and open-ended set of representations, including our common-sense understanding of the world. Modelling central reasoning with computation is closely tied to the problem of building a form of general artificial intelligence (AGI). Our current AI systems function effectively only within carefully constrained problem domains (e.g. detecting credit-card fraud, recognising faces, winning at Go). They perform poorly, or not at all, if the nature of their problem changes, or if contextual or background assumptions change (Lake et al., 2017; Marcus and Davis, 2019). In contrast, humans are relatively robust and flexible general-purpose problem solvers. They can rapidly switch between different kinds of task without significant interference or relearning, import relevant information across tasks, and they are aware of how their reasoning should be adapted as background assumptions and context change. What is more, they do this in a way that is informed by a vast database of common-sense knowledge about how the physical and social world works.

Small fragments of human central reasoning have been computationally modelled using various logics, heuristics, and other formalisms (e.g. J. R. Anderson, 2007; Davis and Morgenstern, 2004; Gigerenzer, Todd and the ABC Research Group, 1999; McCarthy, 1990; Newell and Simon, 1972). However, modelling human general-purpose reasoning in full – and in particular, accounting for its flexibility, reliability, and common-sense knowledge base – remains an unsolved problem. This is evinced by our failure to build a computational model that exhibits anything like the levels of reliability, flexibility, and context-sensitivity demonstrated by humans. Philosophers have attempted to show that this lacuna in computational modelling is no accident, but arises because central reasoning is a 'no go' area for computational approaches to cognition.

4.1 The frame problem

'No go' arguments about central reasoning are often treated as examples of the frame problem in AI. This is misleading, as in AI the frame problem refers to a more narrowly defined problem specific to logic-based approaches. The frame problem is about how a logic-based reasoning system should represent the effects of an action without having to explicitly represent an action's many non-effects (McCarthy and Hayes, 1969). Most actions leave most properties unchanged – eating a sandwich does not (normally) change the location of Antarctica. However, that the action *Eat(Sandwich)* does not change the property *Position(Antarctica)* is not a logical truth, but something that needs to be somehow encoded in the system's

knowledge base. Introducing this kind of information in the form of extra axioms that enumerate the non-effects of every action – 'frame axioms' – is unworkable as a general solution. As the number of actions and properties increases, the system would suffer an explosion in the number of frame axioms. The frame problem is how to encode this 'no change' information in a more efficient way. It is normally interpreted as the problem of formalising the rule that an action does not change a property unless there is evidence to the contrary. Formalising this principle poses numerous technical hurdles, and it has stimulated important developments in non-monotonic logics, but it is widely regarded as a solved issue for logic-based approaches (Lifschitz, 2015; Shanahan, 1997; Shanahan, 2016).

A number of philosophers, inspired by the frame problem, have argued that there are broader and more fundamental difficulties with modelling central reasoning with computation. They do not always agree about the precise nature of these problems, or about their exact scope or severity. A number of these problems – confusingly also labelled the 'frame problem' – can be found in the essays inside Pylyshyn (1987) and Ford and Pylyshyn (1996), useful critical reflections on which are provided by Chow (2013), Samuels (2010), Shanahan (2016), and Wheeler (2008). The next two sections describe attempts by philosophers to identify these 'no go' barriers to computational accounts of human central reasoning.

4.2 Dreyfus's argument

The first comes from Hubert Dreyfus (1972; 1992). Dreyfus initially focused on classical, symbolic computational approaches to central reasoning. The kind of model he had in mind is exemplified by Douglas Lenat's Cyc project. This aimed to encode the entirety of human common-sense knowledge in a database of explicit symbolic representations over which a logic-based inference system could run queries (Lenat and Feigenbaum, 1991). Dreyfus argued that any such project for modelling human central reasoning faced two insuperable problems.

First, the goal of encoding all human common-sense knowledge in symbolic form was unattainable. Drawing on the work of Heidegger, Merleau-Ponty, and the later Wittgenstein, Dreyfus argued that any attempt to formalise common sense with symbolic representations will fail to capture a background of implicit assumptions, significances, and skills that are required in order for that formalisation to be used effectively. There is no way to make our common-sense knowledge explicit and symbolically encoded without presupposing a rich, implicit background of know-how. Fragments of common-sense reasoning can be formalised, but attempts to formalise it all will leave gaps, and attempts to fill those gaps will introduce gaps elsewhere. According to Dreyfus, the goal of formalising common-sense reasoning will run into the same problems that caused Husserl's twentieth-century phenomenological

attempt to describe all the principles and beliefs that underlie human intelligent behaviour to fail (H. L. Dreyfus, 1991; H. L. Dreyfus and S. E. Dreyfus, 1988). (Searle makes a similar point about what he calls the 'Background' in Searle (1992), pp. 175–196.)

Second, even if human common-sense knowledge could be formalised in symbolic rules, another problem arises concerning how a system would be able to use that information. Potentially, any piece of information could be relevant to any problem – there is no way to screen off, a priori, any piece of knowledge as irrelevant. However, given the size of the common-sense knowledge base, the system cannot consider every piece of information it has in turn and explore all their potential implications. How then does it select which representations are relevant to its current problem? In order to do this, the system needs to know a considerable amount about the nature of the current problem – about its current context and which background assumptions it is safe to make. How does it know this? Unless the external designers cheat and tell it the answer, the only way seems to be to deploy its vast database of common-sense knowledge to determine the type of situation it is in. But that leads back to the original problem of how it uses that information and how it selects which pieces of information are relevant. In order to deploy its vast database of knowledge, the system has to know which pieces of knowledge are relevant to the current context; in order to know this, it has to know what the current context is; but in order to know that, it needs to be able to deploy its knowledge effectively, which it can't do because it doesn't know which pieces of knowledge are relevant. Dreyfus concludes that any computational model that attempts to perform flexible, context-sensitive central reasoning will be trapped in an endless loop of attempting to determine context and relevance (H. L. Dreyfus, 1992, pp. 206–224).

Dreyfus claimed that these two problems affect any attempt to account for central reasoning with a classical, symbolic computational model. In later work, Dreyfus attempted to extend his argument to other kinds of computational model. He considered connectionist networks trained under supervised learning and reinforcement learning. He cautiously concluded that although these approaches might get around the first problem (because they are not committed to formalising knowledge with symbolic representations), they are still plagued by something akin to the second problem. Current machine learning regimes tend to tune models to relatively specific, narrow problems domains and after training networks have not (yet) shown the flexibility to reproduce general-purpose central reasoning (H. L. Dreyfus, 1992, pp. xxxiii–xliii; H. L. Dreyfus, 2007). It is worth noting that the character of Dreyfus's argument changes here from that of an unqualified 'no go' claim (it is *impossible* for a computational model to account for central reasoning), to a more nuanced prediction based on what has been achieved by computational models to date (we do not know how to train a connectionist network to be flexible enough to

provide a model of central reasoning).

Dreyfus proposed that central reasoning should be modelled with a dynamical, embodied approach that falls under the heading of 'Heideggerian AI'. The details of such a view are unclear, but broadly speaking the idea is that our inferential skills and embodied knowledge are coordinated and arranged such that they are solicited by the external situation and current context to bring certain knowledge to the fore. The resources needed to determine relevance and context therefore do not lie in a computation inside our heads, but are somehow encoded in the dynamical relationship between ourselves and the external world (Haugeland, 1998). Wheeler (2005; 2008) develops a version of Heideggerian AI that takes inspiration from the situated robotics movement (Brooks, 1991). H. L. Dreyfus (2007) argues instead for an approach based around the neurodynamics work of Freeman (2000). Neither has yet produced a working model that performs appreciably better than conventional computational alternatives.

4.3 Fodor's argument

According to Jerry Fodor, two problems conspire to prevent a computational model being able to account for central reasoning: the 'globality' problem and the 'relevance' problem (Fodor, 1983; Fodor, 2000; Fodor, 2008). Like Dreyfus, Fodor focused primarily on how these problems affect classical, symbolic models. Fodor thought that non-symbolic (e.g. connectionist) models faced other problems related to their ability to reproduce the systematicity and compositionality of human thought that render them unsuitable as models of central reasoning (Fodor, 2008; Fodor and Lepore, 1992; Fodor and Pylyshyn, 1988). (For a review of other connectionist approaches to central reasoning, see Samuels (2010), pp. 289–290.)

The globality problem concerns how a system computes certain properties that are relevant to central reasoning: simplicity, centrality, and conservativeness of beliefs. Fodor suggested that these epistemic properties are 'global', by which he meant that they may depend on any number of the system's beliefs; they are not features that supervene exclusively on local properties of the individual belief of which they are predicated. A belief might count as simple in one context (i.e. relative to one set of surrounding beliefs), but complex in another. The simplicity of a belief is not an intrinsic property of that belief. Therefore, simplicity cannot depend solely on a belief's intrinsic computational properties. Fodor suggests that classical computational processes are sensitive *only* to intrinsic, local computational properties of the representations they manipulate. Therefore, any reasoning that requires sensitivity to global properties cannot be a classical computational process.

Fodor's globality argument has been roundly criticised (Ludwig and Schneider, 2008;

Samuels, 2010; Schneider, 2011). Critics point out that computations may be sensitive, not only to the local properties of individual representations, but also to *relations* between representations: how a belief's local computational properties relate to the properties of other representations and how these relate to the system's general rules of syntactic processing. The failure of a global property, such as simplicity, to supervene on a belief's local computational properties has no bearing on whether simplicity can be tracked by a computational process. Simplicity may supervene on, and be reliably computationally tracked by following, syntactic relationships between multiple representations. Fodor anticipates this response, however – in Fodor (2000) he labels it M(CTM). He argues that solving the globality problem in this way runs into his second problem.

The second problem – the relevance problem – arises when a reasoning system needs to make an inference based on a large number of beliefs, any number of which may be relevant to the problem at hand. Typically, only a few of these beliefs will be relevant – for example, to computation of a global property like simplicity. Only a fraction of the system's representations need to be considered. The relevance problem is to determine the membership of this fraction. Humans tend to employ representations that are relevant to their current inference or planning task – they hone in on contextually appropriate beliefs and goals. How does their cognitive system know which representations are relevant without doing an exhaustive search through all its beliefs and goals to check each and their consequences for relevance? How does it determine which representations to consider when computing a global property like simplicity? Echoing the worries raised by Dreyfus, Fodor says we do not know of a computational method to solve this problem – one that is able to pare down the set of all the system's representations to a subset relevant to the current task, when that task may change rapidly and about which no a priori assumptions can be made. This *relevance* problem lies at the heart of both Fodor's and Dreyfus's objections to computational accounts of central reasoning (Chow, 2013, pp. 312–313).

4.4 Responses to the problem

Many philosophers have suggested that humans solve the relevance problem using some set of heuristics. They point to heuristic methods used by Internet search engines, which approximately determine relevance, and the fact that humans often fail to spot relevant information or deploy irrelevant information when they engage in general-purpose reasoning (Carruthers, 2006; Clark, 2002; Lormand, 1990; Samuels, 2005; Samuels, 2010). While considerations like these might increase our confidence that certain aspects of the relevance problem can be solved by computational means, they do not cut much ice unless one can say in detail how the observed success rate of humans is produced. Heuristics might at some level inform

human central reasoning, but unless one can say which heuristics are used and how these produce the impressive flexibility and reliability seen in human reasoning, it is hard to say that one has solved the relevance problem (see Chow, 2013, pp. 315–321).

Shanahan and Baars (2005) and Schneider (2011) suggest that the relevance problem is solved by the Global Workspace Theory (GWT) of consciousness. GWT is a proposed large-scale architecture for human cognition in which multiple 'specialist' cognitive processes compete for access to a global workspace where central reasoning takes place. Access to the global workspace is controlled by 'attention-like' processes (Baars, 1988). Mashour et al. (2020) and Dehaene and Changeux (2004) describe a possible neural basis for GWT. Goyal et al. (2021) suggest GWT as a way to enable special-purpose AI systems to share information and coordinate actions. GWT is a promising architecture, but it does not answer the worries raised by Dreyfus and Fodor. The model does not explain the mechanism by which information from specialists is regulated to guarantee relevance to the context and the contents of the central workspace. Baars and Franklin (2003) suggest there is an interplay between 'executive functions', 'specialist networks', and 'attention codelets' that control access to the global workspace, but exactly how these components function to track relevance is left unclear. As with the suggestions about heuristics, GWT is not (or not yet) a worked-out solution to the relevance problem (see Sprevak, 2019, pp. 557–558).

A notable feature of current philosophical 'no go' arguments about central reasoning is that, unlike the arguments of Sections 2 and 3, they do not directly apply to all computational models. Both Dreyfus's and Fodor's arguments consist in pointing out problems with *past* and *current* computational approaches to central reasoning. The persuasive force of this against untried future computational approaches is questionable. Sceptics might see the consideration they raise as evidence that central reasoning is likely to never yield to a computational approach – Dreyfus and Fodor suggest that, pending evidence to the contrary, this is the rational conclusion to draw. Fans of computational modelling might respond that no one thought that modelling central reasoning would be anything other than a hard research problem; it is not surprising that it has not been solved yet, and the landscape of untried computational models is very large (Samuels, 2010, pp. 288–292).

5 Conclusion

This chapter describes a small sample of philosophical issues that have received attention in the computational cognitive sciences. Its focus has been on 'no go' arguments regarding three aspects of human cognition: semantic content, phenomenal consciousness, and central reasoning. A 'no go' argument aims to identify a possible

or expected gap in the model of cognition we hope to eventually obtain from the computational cognitive sciences. One might worry that predicting such gaps now is rash or impractical given the relatively early state of development of the cognitive sciences. This is not the case. The project bears directly on questions about the estimated plausibility of proposed computational approaches, the motivations for pursuing them, and the rationale for investing in future research on computational or non-computational alternatives. Such judgements cannot be avoided and are made not uncommonly on the basis of hunches about the likely future success of research programmes. Philosophical work in this area can help to systematise evidence and enable decision makers give a reason-based appraisal of what is and is not likely to be achieved.

Bibliography

Adams, F. and K. Aizawa (2021). "Causal theories of mental content". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2021. URL: https://plato.stanford.edu/archives/spr2021/entries/content-causal/.

Anderson, J. R. (2007). *How Can the Human Mind Occur in a Physical Universe?* Oxford: Oxford University Press.

Anderson, M. L. (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. and S. Franklin (2003). "How conscious experience and working memory interact". In: *Trends in Cognitive Sciences* 7, pp. 166–172.

Barrett, H. C. (2020). "Towards a cognitive science of the human: Cross-cultural approaches and their urgency". In: *Trends in Cognitive Sciences* 24, pp. 620–638.

Barrett, H. C. and R. Kurzban (2006). "Modularity in cognition: Framing the debate". In: *Psychological Review* 113, pp. 628–647.

Block, N. (1978). "Troubles with Functionalism". In: *Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science*. Ed. by C. W. Savage. Vol. 9. Minneapolis: University of Minnesota Press, pp. 261–325.

— (1980). "What intuitions about homunculi don't show". In: *Behavioral and Brain Sciences* 3, pp. 425–426.

— (1986). "Advertisement for a Semantics for Psychology". In: *Midwest Studies in Philosophy* 10, pp. 615–678.

Boden, M. A. (1989). "Escaping from the Chinese Room". In: *Artificial Intelligence in Psychology*. Cambridge, MA: MIT Press, pp. 82–100.

Brooks, R. A. (1991). "Intelligence without representation". In: *Artificial Intelligence* 47, pp. 139–159.

Burge, T. (1986). "Individualism and psychology". In: *Philosophical Review* 95, pp. 3–45.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

— (2010a). "Consciousness and its place in nature". In: *The Character of Consciousness*. Oxford University Press, pp. 103–139.

— (2010b). "Facing up to the problem of consciousness". In: *The Character of Consciousness*. Oxford University Press, pp. 3–34.

— (2010c). "How can we construct a science of consciousness?" In: *The Character of Consciousness*. Oxford University Press, pp. 37–58.

— (2010d). "The two-dimensional argument against materialism". In: *The Character of Consciousness*. Oxford University Press, pp. 141–205.

— (2012). "A computational foundation for the study of cognition". In: *Journal of Cognitive Science* 12, pp. 323–357.

Chirimuuta, M. (forthcoming). *How to Simplify the Brain*. Cambridge, MA: MIT Press.

Chomsky, N. (1995). "Language and nature". In: *Mind* 104, pp. 1–61.

Chow, S. J. (2013). "What's the problem with the frame problem?" In: *Review of Philosophy and Psychology* 4, pp. 309–331.

Clark, A. (2000). "A case where access implies qualia?" In: *Analysis* 60, pp. 30–38.

— (2002). "Global abductive inference and authoritative sources, or, how search engines can save cognitive science". In: *Cognitive Science Quarterly* 2, pp. 115–140.

Cole, D. (2020). "The Chinese Room Argument". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2020. URL: https://plato.stanford.edu/archives/win2020/entries/chinese-room/.

Colombo, M. and S. Hartmann (2017). "Bayesian cognitive science, unification, and explanation". In: *The British Journal for the Philosophy of Science* 68, pp. 451–484.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge, MA: MIT Press.

Davis, E. and L. Morgenstern (2004). "Introduction: Progress in formal commonsense reasoning". In: *Artificial Intelligence* 153, pp. 1–12.

Dehaene, S. and J.-P. Changeux (2004). "Neural mechanisms for access to consciousness". In: *The Cognitive Neurosciences, III*. Ed. by M. Gazzaniga. Cambridge, MA: MIT Press, pp. 1145–1157.

Dennett, D. C. (1978). "Why you can't make a computer that feels pain". In: *Synthese* 38, pp. 415–456.

— (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

— (1991). *Consciousness Explained*. Boston, MA: Little, Brown & Company.

— (1995). "The unimagined preposterousness of zombies". In: *Journal of Consciousness Studies* 2, pp. 322–326.

— (2001). "The zombic hunch: Extinction of an intuition?" In: *Royal Institute of Philosophy Supplement* 48, pp. 27–43.

— (2013). *Intuition Pumps And Other Tools for Thinking*. New York, NY: W. W. Norton and Company.

Dewhurst, J. (2018). "Individuation without representation". In: *The British Journal for the Philosophy of Science* 69, pp. 103–116.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

— (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (1972). *What Computers Can't Do*. New York, NY: Harper & Row.

— (1991). *Being-in-the-world: A Commentary on Heidegger's* Being and Time*, Division I*. Cambridge, MA: MIT Press.

— (1992). *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

— (2007). "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian". In: *Artificial Intelligence* 171, pp. 1137–1160.

Dreyfus, H. L. and S. E. Dreyfus (1988). "Making a mind versus modeling the brain: Artificial intelligence back at a branchpoint". In: *Daedalus* 117, pp. 15–44.

Egan, F. (2003). "Naturalistic inquiry: Where does mental representation fit in?" In: *Chomsky and his Critics*. Ed. by L. M. Antony and N. Hornstein. Oxford: Blackwell. Chap. 4.

Egan, F. (2014). "How to think about mental content". In: *Philosophical Studies* 170, pp. 115–135.

Elgin, C. Z. (2017). *True Enough*. Cambridge, MA: MIT Press.

Eliasmith, C. (2005). "Neurosemantics and categories". In: *Handbook of Categorization in Cognitive Science*. Ed. by H. Cohen and C. Lefebvre. Amsterdam: Elsevier, pp. 1035–1055.

— (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

Fodor, J. A. (1978). "Tom Swift and his procedural grandmother". In: *Cognition* 6, pp. 229–247.

— (1980). "Searle on what only brains can do". In: *Behavioral and Brain Sciences* 3, pp. 431–432.

— (1983). *The Modularity of Mind*. MIT Press.

— (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

— (1998). *Concepts*. Oxford: Blackwell.

— (2000). *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

— (2008). *LOT2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fodor, J. A. and E. Lepore (1992). *Holism: A Shopper's Guide*. Oxford: Blackwell.

Fodor, J. A. and Z. W. Pylyshyn (1988). "Connectionism and cognitive architecture". In: *Cognition* 28, pp. 3–71.

Ford, K. M. and Z. W. Pylyshyn, eds. (1996). *The Robot's Dilemma Revisited*. Norwood, NJ: Ablex.

Frankish, K. (2016). "Illusionism as a theory of consciousness". In: *Journal of Consciousness Studies* 23, pp. 11–39.

Freeman, W. J. (2000). *How Brains Make Up Their Minds*. New York, NY: Columbia University Press.

Gigerenzer, G., P. M. Todd and the ABC Research Group, eds. (1999). *Simple Heuristics that Make Us Smart*. New York, NY: Oxford University Press.

Goyal, A., A. Didolkar, A. Lamb, K. Badola, N. R. Ke, N. Rahaman, J. Binas, C. Blundell, M. Mozer and Y. Bengio (2021). "Coordination among neural modules through a shared global workspace". arXiv:2103.01197.

Graziano, M. S. A. (2016). "Consciousness engineered". In: *Journal of Consciousness Studies* 23, pp. 98–115.

Harman, G. (1987). "(Nonsolipsistic) conceptual role semantics". In: *New Directions in Semantics*. Ed. by E. Lepore. London: Academic Press, pp. 55–81.

Harnad, S. (1990). "The symbol grounding problem". In: *Physica D* 42, pp. 335–346.

Haugeland, J. (1998). "Mind embodied and embedded". In: *Having Thought: Essays in the Metaphysics of Mind*. Ed. by J. Haugeland. Cambridge, MA: Harvard University Press, pp. 207–240.

Henrich, J., S. J. Heine and A. Norenzayan (2010). "The weirdest people in the world?" In: *Behavioral and Brain Sciences* 33, pp. 61–135.

Irvine, E. and M. Sprevak (2020). "Eliminativism about consciousness". In: *The Oxford Handbook of the Philosophy of Consciousness*. Ed. by U. Kriegel. Oxford: Oxford University Press, pp. 348–370.

Isaac, A. M. C. (2019). "The semantics latent in Shannon information". In: *The British Journal for the Philosophy of Science* 70, pp. 103–125.

Jackson, F. (1982). "Epiphenomenal qualia". In: *Philosophical Quarterly* 32, pp. 127–136.

Johnson-Laird, P. N. (1978). "What's wrong with Grandma's guide to procedural semantics: A reply to Jerry Fodor". In: *Cognition* 6, pp. 249–261.

Kripke, S. A. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lake, B. M., T. D. Ullman, J. B. Tenenbaum and S. J. Gershman (2017). "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40, e253.

Lee, J. (2018). "Mechanisms, wide functions and content: Towards a computational pluralism". In: *The British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axy061.

Lenat, D. B. and E. A. Feigenbaum (1991). "On the thresholds of knowledge". In: *Artificial Intelligence* 47, pp. 185–250.

Lifschitz, V. (2015). "The dramatic true story of the frame default". In: *Journal of Philosophical Logic* 44, pp. 163–196.

Loewer, B. (2017). "A guide to naturalizing semantics". In: *Companion to the Philosophy of Language*. Ed. by B. Hale, C. Wright and A. Miller. 2nd ed. New York, NY: John Wiley & Sons, pp. 174–196.

Lormand, E. (1990). "Framing the frame problem". In: *Synthese* 82, pp. 353–374.

Ludwig, K. and S. Schneider (2008). "Fodor's challenge to the classical computational theory of mind". In: *Mind and Language* 23.3, pp. 123–143.

Machery, E. (forthcoming). "Discovery and confirmation in evolutionary psychology". In: *The Oxford Handbook of Philosophy of Psychology*. Ed. by J. Prinz. Oxford University Press.

Marcus, G. and E. Davis (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY: Penguin Books.

Mashour, G. A., P. R. Roelfsema, J.-P. Changeux and S. Dehaene (2020). "Conscious processing and the Global Neuronal Workspace hypothesis". In: *Neuron* 105, pp. 776–798.

Maudlin, T. (1989). "Computation and consciousness". In: *The Journal of Philosophy* 86, pp. 407–432.

McCarthy, J. (1990). *Formalizing Common Sense: Papers by John McCarthy*. Ed. by V. L. Lifschitz. Norwood, NJ: Ablex.

McCarthy, J. and P. J. Hayes (1969). "Some philosophical problems from the standpoint of artificial intelligence". In: *Machine Intelligence 4*. Ed. by B. Meltzer and D. Michie. Edinburgh: Edinburgh University Press, pp. 463–502.

Millikan, R. G. (2004). *The Varieties of Meaning*. Cambridge, MA: MIT Press.

Mollo, D. C. (2018). "Functional individuation, mechanistic implementation: The proper way of seeing the mechanistic view of concrete computation". In: *Synthese* 195, pp. 3477–3497.

— (forthcoming). "Deflationary realism: Representation and idealisation in cognitive science". In: *Mind and Language*. URL: http://philsci-archive.pitt.edu/17591/.

Morrison, M. (2014). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.

Nagel, T. (1974). "What is it like to be a Bat?" In: *Philosophical Review* 83, pp. 435–450.

Neander, K. and P. Schulte (2021). "Teleological theories of mental content". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2021. URL: https://plato.stanford.edu/archives/spr2021/entries/content-teleological/.

Newell, A. and H. A. Simon (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E. (2003). *The Geography of Thought*. New York, NY: The Free Press.

Papineau, D. (1987). *Reality and Representation*. Oxford: Blackwell.

Piccinini, G. (2015). *The Nature of Computation*. Oxford: Oxford University Press.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago, IL: University of Chicago Press.

Prinz, J. (2016). "Against illusionism". In: *Journal of Consciousness Studies* 23, pp. 186–196.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Pylyshyn, Z. W., ed. (1987). *The Robot's Dilemma*. Norwood, NJ: Ablex.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rescorla, M. (2013). "Against structuralist theories of computational implementation". In: *The British Journal for the Philosophy of Science* 64, pp. 681–707.

Rolls, E. T. and A. Treves (2011). "The neural encoding of information in the brain". In: *Progress in Neurobiology* 95, pp. 448–490.

Ryder, D. (2004). "SINBAD neurosemantics: A theory of mental representation". In: *Mind and Language* 19, pp. 211–240.

Samuels, R. (2005). "The complexity of cognition: Tractability arguments for massive modularity". In: *The Innate Mind: Vol. I, Structure and Contents*. Ed. by P. Carruthers, S. Laurence and S. P. Stich. Oxford: Oxford University Press, pp. 107–121.

— (2010). "Classical computationalism and the many problems of cognitive relevance". In: *Studies in History and Philosophy of Science* 41, pp. 280–293.

Schank, R. C. and R. P. Abelson (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*. Cambridge, MA: MIT Press.

Searle, J. R. (1980). "Minds, brains, and programs". In: *Behavioral and Brain Sciences* 3, pp. 417–424.

— (1984). *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

— (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Sellars, W. (1962). "Philosophy and the scientific image of man". In: *Frontiers of Science and Philosophy*. Ed. by R. Colodny. Pittsburgh, PA: University of Pittsburgh Press, pp. 35–78.

Shagrir, O. (2012). "Structural representations and the brain". In: *The British Journal for the Philosophy of Science* 63, pp. 519–545.

— (2018). "In defense of the semantic view of computation". In: *Synthese*.

Shanahan, M. (1997). *Solving the Frame Problem*. Cambridge, MA: Bradford Books, MIT Press.

— (2016). "The frame problem". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2016. URL: https://plato.stanford.edu/archives/spr2016/entries/frame-problem/.

Shanahan, M. and B. Baars (2005). "Applying global workspace theory to the frame problem". In: *Cognition* 98, pp. 157–176.

Shea, N. (2013). "Naturalising representational content". In: *Philosophy Compass* 8, pp. 496–509.

— (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.

Skyrms, B. (2010). *Signals*. Oxford: Oxford University Press.

Sprevak, M. (2010). "Computation, individuation, and the received view on representation". In: *Studies in History and Philosophy of Science* 41, pp. 260–270.

— (2013). "Fictionalism about neural representations". In: *The Monist* 96, pp. 539–560.

— (2016). "Philosophy of the psychological and cognitive sciences". In: *Oxford Handbook for the Philosophy of Science*. Ed. by P. Humphreys. Oxford: Oxford University Press, pp. 92–114.

— (2019). "Review of Susan Schneider, *The Language of Thought: A New Philosophical Direction*". In: *Mind* 128, pp. 555–564.

Strawson, G. (2010). *Mental Reality*. 2nd ed. Cambridge, MA: MIT Press.

— (Mar. 2018). "The consciousness deniers". In: *The New York Review of Books*.

Swoyer, C. (1991). "Structural representation and surrogative reasoning". In: *Synthese* 87, pp. 449–508.

Tye, M. (2018). "Qualia". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2018. URL: https://plato.stanford.edu/archives/sum2018/entries/qualia/.

Usher, M. (2001). "A statistical referential theory of content: Using information theory to account for misrepresentation". In: *Mind and Language* 16, pp. 311–334.

Van Gulick, R. (2018). "Consciousness". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2018. URL: https://plato.stanford.edu/archives/spr2018/entries/consciousness/.

Wakefield, J. C. (2003). "The Chinese room argument reconsidered: Essentialism, indeterminacy, and Strong AI". In: *Minds and Machines* 13, pp. 285–319.

Wheeler, M. (2005). *Reconstructing the Cognitive World*. Cambridge, MA: MIT Press.

— (2008). "Cognition in context: Phenomenology, situated robotics and the frame problem". In: *International Journal of Philosophical Studies* 16, pp. 323–349.

Winograd, T. (1972). "Understanding natural language". In: *Cognitive Psychology* 3, pp. 1–191.