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

20 October 2021

What counts as a philosophical issue in computational cognitive science? This chapter briefly reviews possible answers before focusing on a specific subset of philosophical issues. These surround challenges that have been raised by philosophers regarding the scope of computational models of cognition. The arguments suggest that there are aspects of human cognition that may, for various reasons, resist explanation or description in terms of computation. The primary targets of these 'no go' arguments have been *semantic content*, *phenomenal consciousness*, and *central reasoning*. This chapter reviews the arguments and considers possible replies. It concludes by highlighting the differences between the arguments, their limitations, and how they might contribute to the wider project of estimating the value of ongoing research programmes in computational cognitive science.

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). On this view, 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 of the science and considering how matters hang together relative to the broad 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 that philosophical issues are somehow only within the purview of academic philosophers. Anyone who takes computational modelling seriously as an attempt to study cognition is likely to want to know the answers to these questions and is also liable to be able to contribute to the project of answering them. What philosophers bring to this joint 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 separate images that we have of how the world works: the *scientific image* (which describes the posits of the natural sciences – cells, molecules, atoms, forces, etc.) and the *manifest image* (which describes the posits of human common-sense understanding of the world – persons, thoughts, feelings, ideas, etc.) (Sellars, 1962). This chapter adopts a somewhat looser interpretation of the project. Models in the computational cognitive sciences are often partial, provisional, and selected from many possible alternatives that are also consistent with the data. It would be misleading to think that current computational cognitive science contains a single, coherent account that is 'the' scientific image of cognition. Similar concerns could also be raised about our manifest image of the world in light of observations of cross-cultural differences 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 the philosopher's goal is to understand how the many (and varied) 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 our various pre-theoretical folk questions and insights regarding the mind. There is no prior commitment here 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 (Sprevak, 2016).

Under this broad heading, there is a huge range of work. This includes consideration of how to interpret the terms of specific computational models – about which parameters one should be a 'realist' or an 'instrumentalist' (Colombo and Seriès, 2012; Rescorla, 2016); how to make sense of theoretical concepts that appear across multiple models, like the notion of a cognitive 'module' (Carruthers, 2006; Samuels, 1998); analysis and formalisation of general features of experimental methodology in computational neuroscience (Glymour, 2001; Machery, 2013); identification of differences between computational approaches and rival approaches to modelling cognition (Eliasmith, 2003; van Gelder, 1995); consideration of how techniques in machine learning and AI might inform work in computational neuroscience (Buckner, 2021; Sullivan, 2019); interpretation of experimental results that function as evidence for specific computational models (Apperly and Butterfill, 2009; Block, 2007; Shea and Bayne, 2010); and consideration of how computational models of cognition connect to wider questions about the nature of the human mind, its subjective experiences, its evolutionary history, and the kinds of social and technological structures that it builds (Clark, 2016; Dennett, 2017; Godfrey-Smith, 2016; Sterelny, 2003).

The primary focus here will, by necessity, be narrower than the full extent of issues within this diverse intellectual landscape. This chapter focuses on challenges raised to computational modelling that arise from philosophical work on the nature of cognition and consciousness.

1.1 Overview of chapter

When building a computational model in the cognitive sciences, researchers generally aim to build a model of some prescribed subdomain within cognition or behaviour (e.g. of face recognition, cheater detection, word segmentation, or depth perception). Splitting up human cognition into various smaller domains raises questions about *how* one should do this. This is the problem of how one should *individuate* our cognitive capacities and overt behaviour (M. L. Anderson, 2014; Barrett and Kurzban, 2006; Machery, forthcoming). It also raises questions about how the separate models of individual cognitive subdomains that one hopes to obtain will subsequently be woven together to create a coherent, integrated understanding of cognition. This concerns the issue of how one should *unify* models of distinct aspects of cognition (Colombo and Hartmann, 2017; Danks, 2014; Eliasmith, 2013).

This chapter focuses on a set of issues that are related, but posterior, to the two just mentioned. These concern possible *gaps* left by this strategy for modelling cognition. If this strategy were in an ideal world to run to completion, would there be any aspects of cognition or behaviour that would be missing from the final picture? Are there any aspects of cognition for which we should *not* expect to obtain

a computational model? Are there 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.

These 'no go' arguments may be subdivided further into *in principle* and *in practice* arguments. In principle arguments aim to show that it is *impossible* for any computational model to account for the cognitive capacity in question. In practice arguments are weaker. They aim only to show that, given our current state of knowledge, we should not expect to discover such a model – an adequate model *might* exist, but we should not expect to find it, at least in the foreseeable future.

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 concerning computational modelling of cognition. The precise nature of its intended 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 him, 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 inside the room 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 him, 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 main 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, whatever its specific shortcomings, is an illustration of a valid 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 refers 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 (i.e. what they refer to in the world).

Searle is not alone in making this claim. Putnam (1981) argued that manipulating 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 physical environment might undergo exactly the same formal transitions, but have different meaning attached to its symbolic expressions based on its relationship to different environmental properties. Fodor (1978) described two physically identical devices that undergo the same symbol-shuffling processes, one of which runs a simulation of the Six-Day War (with its symbols referring to tank divisions, jet planes, and infantry units) and the other runs a simulation of a chess game (with its 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 might 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 causal, non-formal relationships to external objects in the environment in perception or motor control).

These considerations are also not confined to symbolic computational models of cognition. Similar observations could be made about computational models that are defined over numerical values or over probabilities. Consider artificial neural networks. These computational models consist in collections of abstract nodes and connections that chain together long sequences 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 they relate to distal properties or objects in the environment? As outside observers, we might *interpret* numerical values inside an artificial neural network as referring to certain things (just as, in a similar fashion, 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 an artificial neural network 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. Numerical rules no more imbue an artificial neural network with semantic content than do the symbolic rules that operate over expressions for a classical, symbolic computation (cf. Searle, 1990). Computational models that operate over probabilities or probability distributions face a similar kind of problem. These models are normally defined in terms of operations on probability distributions (understood as ensembles of numerical values that satisfy the requirements for a measure of probability). These distributions might be interpreted by us as external observers as probabilities of certain 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 computational and semantic aspects of cognition are wholly independent. It is likely that some symbolic expressions get their meaning fixed via their formal computational role (plausibly, this is the case for expressions that represent logical connectives like AND and OR). What is being claimed is that not *all* semantic content can be determined in this way, by formal computational role. An adequate account of semantic aspects of cognition 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 in relation to procedural semantics or conceptual-role semantics, 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 will need to be supplemented by another kind of model in order to adequately account for cognition's semantic properties. The project of modelling cognition should correspondingly be seen as possessing at least two distinct branches. One branch consists in describing the formal computational transitions or functions associated with a cognitive process. 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 most clearly laid out in the writings of Jerry Fodor. Fodor argued that one should sharply distinguish between one's *computational theory* (which describes the dynamics of abstract computational vehicles) 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 semantic properties or vice versa (see Fodor, 1998, pp. 9–12). (Fodor (1980) makes this observation in his response to the Chinese room argument, essentially conceding that Searle's conclusion about pure syntax is correct but obvious.)

What does a theory of content look like? Fodor argued that a good theory of content should try to answer two questions about human cognition: (S1) How do its computational states get their semantic properties? (S2) Which specific semantic contents do they have? Fodor also suggested that a theory of content suitable for fulfilling the explanatory ambitions of computational cognitive science should be *naturalistic*. What this last condition means is that the answers a theory of content gives to questions S1 or S2 should not employ semantic or intentional concepts. A theory of content should explain how semantic content in cognition arises, and how

specific semantic contents get determined, in terms of the kinds of non-semantic properties and processes that typically feature in the natural sciences (e.g. physical, causal processes that occur inside the brain or the environment). A theory of content should not attempt to answer S1 or S2 by, for example, appealing to the semantic or mental properties of external observers or the intentional mental states of the subject themselves (Fodor, 1990, p. 32; Loewer, 2017).

Fodor developed his own naturalistic theory of content, which he called the 'asymmetric dependency theory'. This theory claimed that semantic content in cognition is determined by a complex series of law-like relationships obtaining between current environmental stimuli and formal symbols inside the cognitive agent (Fodor, 1990). In contrast, teleological theories of content attempt to naturalise content by appeal 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 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 of this approach propose that semantic content is determined by whichever distal states maximise mutual information with an internal computational state (Isaac, 2019; Skyrms, 2010; Usher, 2001) – this echoes methods used by external observers in cognitive neuroscience to assign representational content to neural responses in the sensory or motor systems (Eliasmith, 2005; Rolls and Treves, 2011; Usher, 2001). Shea (2018) provides a powerful naturalistic theory of content that weaves together elements of all the approaches above and suggests that naturalistic semantic content is determined by different types of condition in different contexts.

No naturalistic theory of content has yet proved entirely adequate, and naturalising content remains more of an aspiration than an attained solution. Among the challenges faced by current approaches are allowing for the possibility of misrepresentation; avoiding introducing unacceptably large amounts of indeterminacy in cognitive semantic content; and providing a sufficiently general theory of cognitive semantic content that will cover not only the representations involved in perception and motor control but also more abstract representations like DEMOCRACY, TIMETABLE, and QUARK (see Adams and Aizawa, 2021; Neander and Schulte, 2021; Shea, 2013).

Some philosophers have suggested the need for a different approach to explaining semantic content in the computational cognitive sciences. Egan (2014) argues that we should assume, at least as a working hypothesis, that cognitive semantic con-

tent *cannot* be naturalised. This is not because the semantic content in question is determined by some magical, non-naturalistic means, but because the way in which we ascribe semantic content to formal computational models is an inherently messy matter that is influenced by endless, unsystematisable pragmatic concerns (Chomsky, 1995; Egan, 2003). Semantic content determination is just not the sort of subject matter that lends itself to description by any concise non-intentional theory – one is unlikely to find a naturalistic theory of semantic content for similar reasons that one is unlikely to find a concise non-intentional theory of jokes, excuses, or anecdotes. Egan suggests that pragmatic ascription of semantic content to computational models nevertheless plays a residual 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 (Egan, 2010).

A different approach to Egan's suggests that ascriptions of semantic content to computational models should be treated as a kind of idealisation or fiction within computational cognitive science (Chirimuuta, forthcoming; Coelho Mollo, 2021; Sprevak, 2013). This builds on a broader trend of work in philosophy of science that emphasises the value of idealisations and fictions in all domains of scientific modelling, from particle physics to climate science. Idealisations and fictions should be understood not necessarily as defects in a model, but as potentially valuable compromises that provide benefits with respect to understanding, prediction, and control that would be unavailable from a scientific model that is restricted to literal truth telling (Elgin, 2017; Morrison, 2014; Potochnik, 2017).

While philosophers do not agree about how to answer S1 and S2, there is near consensus that a *purely* computational theory would not be adequate. A computational model of cognition must be supplemented by something else – a naturalistic theory of content, an intentional gloss, or a reinterpretation of scientific practice – that explains how the (symbolic or numerical) states subject to computational rules gain their semantic content. Moreover, this is widely assumed to be an *in principle* limitation to what a computational model of cognition can provide. It is not a shortcoming that can be remedied by moving to a new computational model or one with more sophisticated formal rules.

2.4 Content and physical computation

The preceding discussion operated under the assumption that a computational model is defined *exclusively* in terms of formal rules (whether those be symbolic or numerical). This fits with one way in which computational models are discussed in the sciences. Mathematicians, formal linguists, and theoretical computer scientists often define a computational model as a purely abstract, notional entity (e.g. a set-

theoretic construction such as a Turing machine, Boolos, Burgess and Jeffrey, 2002). However, researchers in the applied sciences and in engineering often talk about their computational models in a different way. In these contexts, a computational model is often also tied to its implementation in a particular physical system. Part of a researcher's intention in proposing such a model is to suggest that the formal transitions in question are implemented in that specific physical system. In the case of the computational cognitive sciences, formal transitions are normally assumed to be implemented (at some spatiotemporal scale) in the cognitive agent's physical behaviour or neural responses.

If a formal computation is physically implemented, the physical states that are manipulated will necessarily stand in some non-formal relations to distal entities in the world. Physically implemented computations cannot help but stand in law-like causal relations to objects in their environment, or 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 implementation relates to semantic content 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 Coelho Mollo, 2018; Dewhurst, 2018; Lee, 2018; Piccinini, 2015, pp. 26–50; Rescorla, 2013; Shagrir, 2020; Sprevak, 2010). At the moment, there is no consensus among philosophers about whether, and to what extent, physical implementation constrains the semantics of a computation's states. Consequently, it is worth bearing in mind that Searle's observation that 'syntax is not sufficient for semantics', even if true for the purely formal computations that he had in mind, may not apply to the physically implemented computations proposed in many areas of the computational cognitive sciences (see Boden, 1989; Chalmers, 1996, pp. 326–327; Dennett, 1987, pp. 323–326)

3 Phenomenal consciousness – The hard problem

'Consciousness' may refer to many different kinds of mental 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 a 'no go' argument concerning *phenomenal consciousness*. 'Phenomenal consciousness' refers to the subjective, qualitative feelings that accompany some aspects of cognition. When you touch a piece of silk, taste a raspberry, or hear the song of a blackbird, over and above any processes of classification, judgement, report, attentional shift, control of behaviour, and planning, you also undergo subjective sensations. There is something it *feels like* to do these things. Some philosophers

reserve the term 'qualia' to refer to these feelings (Tye, 2018). The *hard problem* of consciousness is to explain why phenomenal feelings accompany certain aspects of 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 against physicalism is a 'no go' argument phrased in terms of the conceivability of a philosophical zombie. A philosophical zombie is a hypothetical being who is a physical duplicate of a human and who lives in a world that is a physical duplicate of our universe – a world with the same physical laws and the same instances of physical properties. The difference between our world and the zombie world is that the agents in the zombie world either lack conscious experience or have a different distribution of phenomenal experiences across their mental life from our own. A zombie's cognitive processes occur 'in the dark' or they are accompanied by different phenomenal experiences from our own (e.g. it might experience the qualitative feeling we associate with tasting raspberries when it tastes blueberries and vice versa).

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

The next step in the conceivability argument is to say that our ability to conceive of a scenario is a reliable guide to whether it is possible. If a world in which zombies exist is conceivable, then we should believe, pending evidence to the contrary, that it corresponds to a genuine possibility. However, if a zombie world is possible, then the distribution of physical properties and physical laws could be exactly as it is in our world and the beings of that world either lack phenomenal experience or have different phenomenal experiences from our own. That means that in *our* world there must be some additional ingredient, over and above the physical facts, that is responsible for the existence and distribution of our phenomenal experiences. Something other than the physical laws and physical properties must explain the difference between our world and a zombie world. Our phenomenal consciousness

cannot be determined only by the physical facts because those facts also hold in the zombie world. Advocates of the conceivability argument conclude that a theory of consciousness that appeals exclusively to physical facts is unable to explain the existence and distribution of our phenomenal experiences (Chalmers, 1996, pp. 93–171; Chalmers, 2010d).

According to the conceivability argument, a physicalist theory cannot answer the following questions: (C1) How does our phenomenal conscious experience arise at all? (C2) Why are our phenomenal conscious experiences distributed in the way they are across our mental life? No matter which physical facts one cites, none adequately answer C1 or C2 because the same physical facts could have obtained and those conscious experiences be absent or different, as they are in a zombie world. This raises the question of what – if not the totality of physical facts – is responsible for the existence and distribution of our phenomenal experiences. Advocates of the conceivability argument have various suggestions at this point, all of which involve expanding or revising our current scientific ontology. The focus of this chapter will not be on those options, but only on the negative point that phenomenal consciousness is somehow out of bounds for current approaches to modelling cognition (see Chalmers, 2010a, pp. 126–137, for a survey of non-physicalist options).

3.2 The conceivability argument against computational functionalism

The conceivability argument against physicalism may be modified to generate a 'no go' argument against computational accounts of phenomenal consciousness.

The primary consideration here is that a hypothetical zombie who is our *computational* duplicate seems to be conceivable. This is a being who performs exactly the same computation as we do but who either lacks conscious experience or has a different distribution of conscious experiences from our own. Similar reasoning to justify both the conceivability and possibility of such a being applies as in the case of the original conceivability argument against physicalism. It seems possible to imagine a being implementing any computation one chooses, or computing any function, and for this to fail to be accompanied by a phenomenal experience, or for it to be accompanied by a phenomenal experience different from our own. No matter how complex the rules of a computation, nothing about it seems to *necessitate* the existence or distribution of specific subjective experiences. One might imagine a silicon or clockwork device functioning as a computational duplicate of a human – undergoing the same computational transitions – but its cognitive life remaining 'all dark' inside, or being accompanied by different subjective experiences from our own (for analysis of such thought experiments, see Block, 1978; Dennett, 1978; Maudlin, 1989). As with the original conceivability argument, it does not matter whether a computational zombie could exist in our world; what matters is only whether a

world with such a being is conceivable.

A separate consideration is that the original conceivability argument appears to entail a 'no go' conclusion concerning any computational model of consciousness that has a physical implementation (Chalmers, 1996, p. 95). Plausibly, any world that is a physical duplicate of our world is a world that is also a duplicate in terms of the physical computations that are performed. It seems reasonable to assume that the physical facts about a world fix which physical computations occur in that world. According to the original conceivability argument, a world that is a physical duplicate of ours could be one in which there is no consciousness or consciousness is distributed differently. Putting these two claims together, a world that is a duplicate of ours in terms of the physical computations performed could be one in which phenomenal consciousness is absent or differently distributed. Hence, in our world there must be some extra factor, over and above any physical computations, that explains the existence and distribution of our phenomenal experiences. A scientific model that appeals only to physical computations – which are shared with our zombie counterparts – would be unable to explain the existence and distribution of our phenomenal experiences.

It is worth stressing that the conceivability argument places no barrier against a computational or physical model explaining access consciousness. 'Access consciousness' refers to the aspects of consciousness associated with reportability and information sharing: storage of information in working memory, information sharing across various processes of planning, reporting, control of action, decision making, and so on (Block, 1990; Block, 2007). Baars (1988) proposed Global Workspace Theory (GWT) as a way in which information from different cognitive processes comes together. Dehaene and colleagues developed GWT and provided a possible neural implementation (Dehaene and Changeux, 2004; Dehaene and Changeux, 2011; Dehaene, Changeux et al., 2006). A theory of this kind might be able to account for how and why certain pieces of information get shared and play a greater role in driving thought, action, and report. However, advocates of the conceivability argument claim that a model of access consciousness cannot explain phenomenal consciousness. Following similar reasoning to that described in the previous section, they argue that one can conceive of a system having access consciousness, but it still lacking phenomenal consciousness or having a different distribution of phenomenal experience to our own. Access consciousness does not necessitate the occurrence of phenomenal feelings (for a contrary view, see Cohen and Dennett, 2011). For these thinkers, explaining access consciousness is classified under the heading of an 'easy problem' of consciousness (Chalmers, 2010b).

3.3 Naturalistic dualism

It is important to understand the extent of the intended 'no go' claim about phenomenal consciousness. What is claimed is that *solving the hard problem* is beyond the ability of a physical or computational model of consciousness. This does not mean, however, that a physical or computational account can tell us nothing about phenomenal consciousness. Chalmers (2010b; 2010c) argues that a computational or physical model can, for example, tell us a great deal about *correlations* between physical/computational states and our phenomenal experiences. The conceivability argument does not deny that such correlations exist, and measurement of brain activity shows ample evidence of correlations between brain states and phenomenal experience. Describing and systematising these correlations may have considerable value to science in terms of allowing us to categorise, predict, and control our phenomenal states. Such a model cannot, however, explain why phenomenal experience occurs, for it cannot rule out the possibility that the same physical or computational states could occur without any conscious accompaniment.

An analogy might help to clarify this point. Suppose that one were to begin a correlational study of the phenomena of lightning and thunder. One might build a statistical model that captures the relationship between observations of the two phenomena. In a similar fashion, one might engage in a correlational study of brain states and phenomenally conscious states and attempt to capture their relationship. In both cases, something would be missing from the model that is produced. What would be missing is an understanding of how and why the two variables are linked. Lightning typically co-occurs with thunder, but not always, and no pattern of lightning *necessitates* an observation of thunder (atmospheric conditions might cause sound waves to be refracted or deadened before they reach the observer). This gap in the model can be rectified by introducing further physical variables (e.g. distributions of electrical charges in the air, measurements of air density and temperature). In an enlarged, more detailed, physical model, it should be possible to explain why observations of lightning are correlated with observations of thunder, and how and why such correlations might fail to obtain. In the case of phenomenal consciousness, the conceivability argument claims to show that this kind of remedy is not available. The 'explanatory gap' between the two variables cannot be filled by introducing extra physical variables into one's model. No matter how many physical variables one adds, the model will still not entail the occurrence of phenomenal experiences – for, according to the conceivability argument, all these physical variables could be the same and the consciousness experience be absent or different. A physical/computational model of consciousness can provide us with a description of the correlates of consciousness, but it cannot provide an explanation of why those correlates are accompanied by phenomenal experience.

3.4 Eliminativism and related replies

Not all philosophers accept the reasoning behind the conceivability argument. Dennett argues that one can easily be misled by 'intuition pumps' like zombie thought experiments. These can work on our imagination like viewing a picture by M.C. Escher: we appear to see something new and remarkable, but only because certain considerations have been omitted or played up and we have failed to spot some hidden inconsistency in the imagined scenario. Dennett suggests that a more reasonable conclusion to draw is not that phenomenal consciousness is a 'no go' domain for computational modelling of cognition but that the project of trying, from the armchair, to set a limit on what a physical/computational model can and cannot explain is deeply misconceived (Dennett, 2013). For all we know, a truly thorough, mature conceptualisation of a physical or computational duplicate of our world, imagined down to the smallest detail, would rule out the possibility that there could be zombies (Dennett, 1995; Dennett, 2001).

Dennett's remarks about the reliability of our intuitions about zombies may dampen one's confidence in the 'no go' argument. However, this by itself does not block the argument. In order to do this, Dennett also commits to the more speculative, positive claim that *if* we were to successfully wrap our heads around some future correct computational model of consciousness, then we would see that it *must* bring all aspects of consciousness along with it. Advocates of the conceivability argument, while typically open to the idea that zombie intuitions are not apodictically certain (we might be deluding ourselves about the conceivability of a zombie world), tend to pour scorn on this latter contention. No matter how complex a computational model is, they say, it simply is not clear how it could entail that specific conscious experiences occur (Strawson, 2010). The idea that, somewhere in the space of all possible computational models, some model exists that entails conscious experience is, according to these critics, pure moonshine or physicalist dogma (Strawson, 2018).

A position one might be driven towards, and which Dennett defends in *Consciousness Explained* (1991), is that certain aspects of consciousness – namely, the first-person felt 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. The existence and character of our feelings of phenomenal consciousness seem to be among the things about which we are most certain. Denying these subjective 'data', which are accessible to anyone via introspection, may strike one as unacceptable. Nevertheless, past scientific theories have prompted us to abandon other seemingly secure assumptions about the world. If it can be shown that when we introspect on our experience we are mistaken, then perhaps eliminativism can be defended. The potential benefits of eliminativism about phenomenal consciousness are considerable: the hard

problem of consciousness and the challenge posed by the conceivability argument would dissolve. If there is no phenomenal consciousness, then there is nothing for a computational model to explain.

Unfortunately, in addition to the difficulty just mentioned, a further problem faces eliminativist accounts. This is to explain how the (false) data we have about the existence and character of our phenomenal consciousness arise in the first place. This is the so-called *illusion problem* (Frankish, 2016). Some researchers claim that our impression that we have phenomenal consciousness is caused by misfiring of mechanisms of our internal information processing and self report (Clark, 2000; Dennett, 1991; Frankish, 2016; Graziano, 2016). However, such accounts tend to explain only why we *believe* or *act as if* we have phenomenal consciousness. It is not clear how the hypothesised mechanisms generate the *felt* first-person illusion of consciousness (Chalmers, 1996, pp. 184–191). In other words, it is not clear how unreliable introspective mechanisms could generate the false impression of phenomenal consciousness, any more than reliable introspective mechanisms could generate the true impression of phenomenal consciousness. The challenge that an eliminativist faces is to show that the illusion problem is easier to solve by computational or physical means than the hard problem of consciousness (see 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 over a large and open-ended set of representations, including our common-sense understanding of the world. Modelling human-level central reasoning is closely tied to the problem of creating a machine with artificial general intelligence. Current AI systems tend to function only within relatively constrained problem domains (e.g. detecting credit card fraud, recognising faces, winning at Go). They generally perform poorly, or not at all, if the nature of their problem changes, or if relevant contextual or background assumptions change (Lake et al., 2017; Marcus and Davis, 2019). In contrast, humans are relatively robust and flexible general-purpose reasoners. They can rapidly switch between different tasks without significant interference or relearning, they can deploy relevant information across tasks, and they tend to be aware of how their reasoning should be adjusted when background assumptions and context change.

Small fragments of human-level 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-level

central reasoning in full – in particular, accounting for its flexibility, reliability, and deep common-sense knowledge base – remains an unsolved problem. Philosophers have attempted to argue that this lacuna is no accident, but arises because central reasoning is in a certain respect a 'no go' area for computational accounts of cognition.

4.1 The frame problem

Philosophers often describe their 'no go' arguments about central reasoning as instances of the frame problem in AI. This can be misleading as 'the frame problem' refers to a more narrowly defined problem specific to logic-based approaches to reasoning in AI. The frame problem in AI concerns how a logic-based reasoner should represent the effects of actions without having to represent all of an action's non-effects (McCarthy and Hayes, 1969). Few actions change every property in the world – eating a sandwich does not (normally) change the location of Australia. However, the information that *Eat(Sandwich)* does not change *Position(Australia)* is not a logical truth but something that needs to be encoded somehow, either explicitly or implicitly, in the system's knowledge base. Introducing this kind of 'no change' information in the form of extra axioms that state every non-effect of every action – 'frame axioms' – is unworkable. As the number of actions (N) and properties (M) increases, the system would need to store approximately NM axioms. The frame problem in AI concerns how to encode this 'no change' information more efficiently. The challenge is normally interpreted as the problem of formalising a general inference rule that an action does not change a property unless the reasoning system has evidence to the contrary. Formalising this rule poses numerous technical hurdles, and it has stimulated important developments in non-monotonic logics, but it is widely regarded as a solved issue within logic-based AI (Lifschitz, 2015; Shanahan, 1997; Shanahan, 2016).

A number of philosophers, inspired by the original frame problem, have suggested that there are broader and more fundamental difficulties with explaining human-level central reasoning with computation. They do not, however, agree about the precise nature of these difficulties, their scope, or their severity. A number of proposals – confusingly also called the 'frame problem' – can be found in Pylyshyn (1987) and Ford and Pylyshyn (1996). Useful critical reflections on this work are found in Chow (2013), Samuels (2010), Shanahan (2016), and Wheeler (2008). The remainder of this section summarises two attempts by philosophers to pinpoint the problem with modelling human-level central reasoning.

4.2 Dreyfus's argument

The first argument was developed by Hubert Dreyfus (1972; 1992). Dreyfus initially targeted classical, symbolic computational approaches to central reasoning. The sort of computational model he had in mind was exemplified by Douglas Lenat's Cyc project. This project aimed to encode all of human common-sense knowledge in a giant symbolic database of representations over which a logic-based system could run queries to produce general-purpose reasoning (Lenat and Feigenbaum, 1991). Dreyfus argued that no model of this kind could capture human-level general-purpose reasoning. This was for two main reasons.

First, it would be impossible to encode all of human common-sense knowledge with a single symbolic database. Drawing on ideas from Heidegger, Merleau-Ponty, and the later Wittgenstein, Dreyfus suggested that any attempt to formalise human common-sense knowledge will fail to capture a background of implicit assumptions, significances, and skills that are required in order for that formalisation to be used effectively. These philosophers defended the idea that our common-sense knowledge presupposes a rich background of implicit know-how. Fragments of this know-how can be explicitly articulated in a set of symbolic rules, but not all of it at once. Attempts to formalise all of human common-sense knowledge in one symbolic system will, for various reasons, leave gaps, and attempts to fill those gaps will introduce further gaps elsewhere. The goal of formalising the entirety of human common-sense knowledge in symbolic terms will run into the same kinds of problems that caused Husserl's twentieth-century phenomenological attempt to describe explicitly 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 regarding what he calls the 'Background' in Searle, 1992, pp. 175–196.)

Second, even if human common-sense knowledge could be encoded in a single symbolic database, the computational system would find itself unable to use that information efficiently. Potentially, any piece of information from the database could be relevant to any task. Without knowledge about the specific problems the system was facing, there would be no way to screen off any piece of knowledge as irrelevant. Because the database would be so large, the system would not be able to consider every piece of information it had in turn and explore all its potential implications. How, then, would it select which symbolic representations were relevant to a specific problem at hand? In order to answer this, it would need to know which specific problem it was facing – about its context and which background assumptions it was safe to make. But how would it know this? Unless the programmer had told it the answer, the only way would seem to be to deploy its database of common-sense knowledge to infer the type of situation it was in and the nature of the problem it now faced. But that leads one back to the original question of how it was to

use information in that database efficiently. In order to deploy its vast database efficiently, the system would have to know which pieces of knowledge were relevant to the problem at hand. In order to know that, it would have to know what that problem was. But in order to know this, it would need to be able to use its database of knowledge efficiently, which it cannot do because it would not know which pieces of knowledge were relevant. Dreyfus concludes that any computational model that attempts to perform central reasoning would 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 classical, symbolic computational attempt to model human-level general-purpose reasoning. In later work, Dreyfus attempted to extend his 'no go' argument to other kinds of computational model – connectionist networks trained under supervised learning and reinforcement learning. He cautiously concluded that although these models might avoid the first problem (connectionist networks are not committed to formalising knowledge with symbolic representations), they are still affected by something similar to the second problem. Our current methods for training connectionist networks and reinforcement-learning systems tend to tune these models to relatively narrow problem domains. Such systems have not shown the flexibility to reproduce human-level general-purpose central reasoning; they tend to be relatively brittle (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 in principle 'no go' (it is *impossible* for any classical, symbolic computational model to account for central reasoning) to more of a hedged prediction based on what has been achieved by machine-learning methods to date (we do not – yet – know of a method to train a connectionist network to exhibit human-level flexibility in general-purpose reasoning).

Dreyfus proposed that central reasoning should be modelled using a dynamical, embodied approach to cognition that has come to be known as 'Heideggerian AI'. The details of such a view are unclear, but broadly speaking the idea is that the relevant inferential skills and embodied knowledge for general-purpose reasoning are coordinated and arranged such that they are solicited by the external situation and current context to bring certain subsets of knowledge to the fore. The resources needed to determine relevance 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 for an alternative approach based around the neurodynamics work of Freeman (2000). Neither has yet produced a working model that performs appreciably better at modelling human-like context sensitivity than more conventional computational alternatives.

4.3 Fodor's argument

Jerry Fodor argued that two related problems prevent a computational model from being able to account for human-level central reasoning. He called these 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 of central reasoning. Fodor believed that a non-symbolic model (e.g. a connectionist system) would be unsuited to modelling human-level central reasoning because it cannot account for the systematicity and compositionality that he considered necessary features of human thought (for that argument, see Fodor, 2008; Fodor and Lepore, 1992; Fodor and Pylyshyn, 1988). (For discussion of connectionist approaches to central reasoning, see Samuels, 2010, pp. 289–290.)

The globality problem concerns how a reasoning system computes certain epistemic properties that are relevant to general-purpose reasoning: simplicity, centrality, and conservativeness of representations. Fodor suggested that these properties are 'global', by which he meant that they may depend on any number of the system's other representations. They are not features that supervene exclusively on intrinsic properties of the individual representation of which they are predicated. A representation might count as simple in one context – for example, relative to one set of surrounding beliefs – but complex in another. The simplicity of a representation is not an intrinsic property of a representation. Hence, its simplicity cannot depend solely on a representation's intrinsic, local syntactic properties. Fodor claimed that a classical computational process is sensitive *only* to the intrinsic, local syntactic properties of the representations it manipulates. Therefore, any central reasoning that requires sensitivity to global properties cannot be a classical computational process.

Fodor's globality argument has been roundly criticised (e.g. by Ludwig and Schneider, 2008; Samuels, 2010; Schneider, 2011). Critics point out that computations may be sensitive not only to the intrinsic properties of individual representations, but also to syntactic relationships between representations: for example, how a representation's local syntactic properties relate to the local syntactic properties of other representations and how they relate to the system's rules of syntactic processing. The failure of an epistemic property like simplicity to supervene on a representation's intrinsic, local syntactic properties does not mean that simplicity cannot be tracked or evaluated by a computational process. Simplicity may supervene on, and be reliably tracked by following, the syntactic relationships between representations. Fodor (2000) anticipates this response, however – he labels it M(CTM). He argues that solving the globality problem in this way runs into his second problem.

The second problem arises when a reasoning system needs to make an inference based on a large number of representations, any combination of which may be relevant to the problem at hand. Typically, only a tiny fraction of these representations will be relevant to the inference. The relevance problem is to determine the membership of this fraction. Humans tend to be good at focusing in on only those representations from their entire belief set that are relevant to their current context or task. But we do not know how they do this. Echoing the worries raised by Dreyfus, Fodor says we do not know of a computational method that is able to pare down the set of all the system's representations to only those relevant to the current task.

4.4 Responses to the problems

Some philosophers have responded to these problems by emphasising the role of heuristics in relevance determination. They point to the computational methods used by Internet search engines, which, although far from perfect, often do a decent job of returning relevant results from very large datasets. They also stress that humans sometimes fail to deploy relevant information or that they use irrelevant information when reasoning (Carruthers, 2006; Clark, 2002; Lormand, 1990; Samuels, 2005; Samuels, 2010). These two considerations might increase our confidence that human-level central reasoning – and in particular the relevance problem – might be tackled by computational means. However, it does not cut much ice unless one can say which heuristics are used and how the observed success rate of humans is produced. Heuristics might, at some level, inform human central reasoning, but unless one can say precisely how they do this – and ideally produce a working computational model that exhibits levels of flexibility and reliability similar to those seen in human reasoning – it is hard to say that one has solved the problem (see Chow, 2013, pp. 315–321).

Shanahan and Baars (2005) and Schneider (2011) suggest that the issues that Dreyfus and Fodor raise can be resolved within GWT. GWT is a proposed large-scale computational architecture 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 several special-purpose AI systems to share information and coordinate decision making. GWT is a promising architecture, but it is unclear whether it can function as a response to the arguments of Dreyfus and Fodor. The model does not explain the mechanism by which information from specialist processes is regulated so as to be relevant to the current 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 work to track relevance is left unclear. As with the suggestion about heuristics, GWT is not (or not yet) a worked-out solution to the relevance-determination problem (see Sprevak, 2019, pp. 557–558).

A notable feature of the 'no go' arguments that target human-level central reasoning is that, unlike the 'no go' arguments of Sections 2 and 3, they do not straightforwardly generalise across the space of all computational models. Both Dreyfus's and Fodor's arguments consist in pointing out problems with specific computational approaches to central reasoning – primarily, with classical, symbolic models and current connectionist and reinforcement-learning approaches. The persuasive force of what they say against untried or as-yet unexplored computational approaches is unclear. Sceptics might see in their arguments evidence that central reasoning is unlikely to ever yield to a computational approach – Dreyfus and Fodor both suggest that the track record of failure of computational models should lead one to infer that no future computational model will succeed. Fans of computational modelling might respond that explaining central reasoning is an extremely hard research problem and it should not be surprising if it has not yet been solved by computational methods. The landscape of as-yet untried computational methods is very large and, pending evidence to the contrary, we should not presume that central reasoning cannot yield to a computational model (Samuels, 2010, pp. 288–292).

5 Conclusion

This chapter describes a small sample of philosophical issues in the computational cognitive sciences. Its focus has been 'no go' arguments regarding three distinct aspects of human cognition: semantic content, phenomenal consciousness, and central reasoning. One might worry that the project of placing limits on what the computational cognitive sciences can achieve is rash given their relatively early state of development. But this would be to misinterpret how the 'no go' arguments function. These arguments attempt to formalise objections – of different types and different strengths – to the assumption that every aspect of cognition can be adequately explained with computation. This need not shut down debate on the topic, but can serve as an opening move and a potentially helpful spur. The project bears directly on questions about the estimated plausibility of future research programmes within the cognitive sciences, the motivations for pursuing them, and the rationale for devoting resources to computational versus non-computational approaches. Such judgements cannot be avoided; they are made regularly within the cognitive sciences. They are also best made on a considered basis, with reasons marshalled and assessed. Philosophical work in this area can help to systematise

evidence and provide decision makers with reason-based considerations about what challenges the computational cognitive sciences are likely to face.

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.

Apperly, I. A. and S. A. Butterfill (2009). "Do humans have two systems to track belief and belief-like states?" In: *Psychological Review* 116, pp. 953–970.

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.

— (1990). "Consciousness and accessibility". In: *Behavioral and Brain Sciences* 13, pp. 596–598.

— (2007). "Consciousness, accessibility, and the mesh between psychology and neuroscience". In: *Behavioral and Brain Sciences* 30, pp. 481–548.

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

Boolos, G., J. P. Burgess and R. C. Jeffrey (2002). *Computability and Logic*. 4th ed. Cambridge: Cambridge University Press.

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

Buckner, C. (2021). "Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour". In: *The British Journal for the Philosophy of Science*. DOI: 10.1086/714960.

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.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

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

— (2021). "Deflationary realism: Representation and idealisation in cognitive science". In: *Mind and Language*, pp. 1–19. DOI: 10.1111/mila.12364.

Cohen, M. A. and D. C. Dennett (2011). "Consciousness cannot be separated from function". In: *Trends in Cognitive Sciences* 15, pp. 358–364.

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.

Colombo, M. and P. Seriès (2012). "Bayes on the brain—On Bayesian modelling in neuroscience". In: *The British Journal for the Philosophy of Science* 63, pp. 697–723.

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.

— (2011). "Experimental and theoretical approaches to conscious processing". In: *Neuron* 70, pp. 200–227.

Dehaene, S., J.-P. Changeux, L. Naccache, J. Sackur and C. Sergent (2006). "Conscious, preconscious, and subliminal processing: A testable taxonomy". In: *Trends in Cognitive Sciences* 10, pp. 204–211.

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.

Dennett, D. C. (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.

— (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. 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.

— (2010). "Computational models: a modest role for content". In: *Studies in History and Philosophy of Science* 41, pp. 253–259.

— (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. (2003). "Moving beyond metaphors: Understanding the mind for what it is". In: *The Journal of Philosophy* 10, pp. 493–520.

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.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Godfrey-Smith, P. (2016). "Mind, matter, and metabolism". In: *The Journal of Philosophy* 113, pp. 481–506.

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.

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. (2013). "In defense of reverse inference". In: *The British Journal for the Philosophy of Science* 65, pp. 251–267.

— (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.

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.

— (2016). "Bayesian sensorimotor psychology". In: *Mind and Language* 31, pp. 3–36.

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. (1998). "Evolutionary psychology and the massive modularity hypothesis". In: *The British Journal for the Philosophy of Science* 49, pp. 575–602.

— (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.

— (1990). "Is the brain's mind a computer program?" In: *Scientific American* 262, pp. 20–25.

— (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.

— (2020). "In defense of the semantic view of computation". In: *Synthese* 197, pp. 4083–4108.

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.

Shea, N. and T. Bayne (2010). "The vegetative state and the science of consciousness". In: *The British Journal for the Philosophy of Science* 61, pp. 459–484.

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.

Sterelny, K. (2003). *Thought In A Hostile World*. Oxford: Blackwell.

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

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

Sullivan, E. (2019). "Understanding from machine learning models". In: *The British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axz035.

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 Gelder, T. (1995). "What might cognition be, if not computation?" In: *The Journal of Philosophy* 91, pp. 345–381.

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.