Final draft – please cite published version in Tony Cheng, Ryoji Sato, and Jakob Hohwy (eds.) *Expected Experiences: The Predictive Mind in an Uncertain World*. Routledge: forthcoming.

Realism and instrumentalism in Bayesian cognitive science

Danielle Williams and Zoe Drayson

0. Introduction

There are two distinct approaches to Bayesian modelling in cognitive science. *Black-box* approaches use Bayesian theory to model the relationship between the inputs and outputs of a cognitive system without reference to the mediating causal processes; while *mechanistic* approaches make claims about the neural mechanisms which generate the outputs from the inputs. This paper concerns the relationship between these two approaches. We argue that the dominant trend in the philosophical literature, which characterizes the relationship between black-box and mechanistic approaches to Bayesian cognitive science in terms of the dichotomy between instrumentalism and realism, is misguided. We propose that the two distinctions are orthogonal: black-box and mechanistic approaches to Bayesian modelling can each be given either an instrumentalist or a realist interpretation. We argue that the current tendency to conflate black-box approaches with instrumentalism and mechanistic approaches with realism stems from unwarranted assumptions about the nature of scientific explanation, the ontological commitments of scientific theories, and the role of abstraction and idealization in scientific models. We challenge each of these assumptions to reframe the debates over Bayesian modelling in cognitive science.

This paper proceeds as follows. In Section 1 we introduce Bayesian cognitive science and highlight the widespread tendency among philosophers to assume that all black-box approaches are instrumentalist and that all mechanistic approaches are realist. In Section 2, we outline the distinction between realism and instrumentalism in philosophy of science and argue that scientific realism is compatible with a wider range of explanatory practices than some philosophers would have us believe. We use these findings in Section 3 to demonstrate that the distinction between black-box and mechanistic approaches to Bayesian cognitive science does not map neatly onto the distinction between instrumentalist and realist interpretations of Bayesian models, and we show why the two issues should not be conflated. In Section 4, we identify and explore three sources of the problematic conflation relating to ideas about mechanistic explanation, Marr's levels of analysis, and the role of representation in Bayesian computation.

Final draft – please cite published version in Tony Cheng, Ryoji Sato, and Jakob Hohwy (eds.) *Expected Experiences: The Predictive Mind in an Uncertain World*. Routledge: forthcoming.

1. Bayesian cognitive science

1.1 Bayesian inference and cognitive science

Bayesian inference is a method of statistical inference on which probability is understood as measuring degrees of belief in a hypothesis. Hypotheses are updated in light of new evidence or information according to Bayes' rule of conditionalization, which specifies how to calculate the posterior probability of a hypothesis based on its prior probability, the evidence, and the likelihood of the evidence given the prior probability.[1]

Models of Bayesian inference have been successfully applied to a wide variety of domains: to make stock-market predictions, to analyze differential gene expression, to monitor water quality conditions, and to measure the diagnostic accuracy of medical tests, for example. In these cases, Bayesian models are used to characterize the relationship between the inputs and outputs of a formal system. While we often rely on physical computers to perform the complex likelihood calculations on large datasets, there is no suggestion that the mechanisms of the stock market, genetic expression, water quality, or medical testing are themselves physical machines performing Bayesian computations. We are instead taking a 'black-box' approach, on which we apply Bayesian theorizing to the inputs and outputs of a system without making any claims about the nature of the mechanisms which mediate between the inputs and outputs. In cognitive science, black-box approaches to Bayesian models are exemplified by the project of *rational analysis*. Rational analysis uses Bayesian models of conditional probabilities to calculate the optimal input-output function for cognitive tasks, ranging from low-level sensorimotor tasks to high-level reasoning. Bayesian rational analysis models are computational in the sense that they characterize input-output functions, but they make no assumptions that cognizers are themselves physical computers which perform the calculations between input and output.[2]

[1] The precise details of Bayes' theorem are not relevant to our arguments here. For a thorough introduction to Bayes' theorem, see Joyce (2008).
[2] For more on the theoretical framework of rational analysis, see Anderson (1991) and Chater and Oaksford (1999).

Cognitive science has an interest in computational models, however, which is not restricted to input-output functions. The brain itself can be characterized as a physical computer: a machine which performs these computational functions by transforming the inputs into outputs according to an algorithmic process. This way of modelling cognition takes a mechanistic approach rather than a black-box approach, targeting the computational processes which causally mediate between the inputs and outputs. In the case of Bayesian cognitive science, the mechanistic approach suggests that the nervous system implements Bayesian computational functions: it carries and updates information in a way which approximates Bayesian models of probabilistic inference. Some cognitive scientists apply Bayesian models only to particular cognitive functions (e.g. sensory processing, language learning) while other take the "Bayesian brain hypothesis" to provide a unified account of all cognition, perception and action. Mechanistic approaches to Bayesian modelling in cognitive science can also differ in their details: whether they apply a single Bayesian model or a hierarchy of many Bayesian models, for example, and whether they involve prediction error minimization and data compression strategies.

This distinction between black-box approaches and mechanistic approaches to Bayesian cognitive science is widely acknowledged in the literature under a variety of different labels. Jones and Love (2011), for example, use the label 'Bayesian Fundamentalism' for the black-box approach and the label 'Bayesian Enlightenment' for the mechanistic approach. In much of the philosophical literature on Bayesian cognitive science, however, there is a tendency to frame the distinction between mechanistic and black-box approaches to Bayesian models as a version of the distinction between realism and instrumentalism about scientific theories. Once we have provided evidence of this tendency, we will argue that a clearer understanding of scientific realism demonstrates that the realism/instrumentalism distinction is orthogonal to the distinction between mechanistic and black-box approaches to cognition.

1.2 Philosophical interpretations of Bayesian cognitive science

Philosophical discussions of approaches to Bayesian cognitive science often liken the mechanistic approach to scientific realism, and the black-box approach to instrumentalism. Sprevak, for example, proposes that the mechanistic 'Bayesian brain' approach is realist, on the grounds that it interprets the central terms of Bayesian models as "picking out real (and as yet unobserved) entities and processes in the human brain" (Sprevak 2016, p.94). He contrasts the mechanistic

approach with black-box approaches such as rational analysis, which he takes to be instrumentalist because they are "formal devices" which do not refer to neural entities and processes (Sprevak 2016, p.94).[3] Rescorla (2019) explicitly defends a realist interpretation of Bayesian cognitive science by appealing to the mechanistic approach to Bayesian models, on which causal structures implementing Bayesian inferences mediate between input-output mappings. He contrasts the "realism" of his mechanistic approach with the "instrumentalism" of black-box approaches on which Bayesian model are useful fictions: "predictively useful devices that do not accurately depict psychological reality" (Rescorla 2019, p. 57). Conversely, Block (2018) argues that Bayesian cognitive science is not committed to the sorts of physically-implemented internal representations associated with mechanistic approaches, and he uses this to justify taking an instrumentalist interpretation of Bayesian cognitive models. The literature thus seems to assume that only mechanistic approaches to Bayesian models, with their commitment to concrete neural entities and causal processes, are realist: black-box approaches to Bayesian models, which model the formal relationship between the inputs and outputs, are taken to be instrumentalist.

We will now argue that these assumptions in the literature conflate several different dimensions of theory interpretation, with problematic consequences. We will first explore the debate between realism and instrumentalism more generally, before looking at how it applies to Bayesian cognitive science.

2. Scientific realism

Scientific realism is the position that our scientific theories and models provide us with knowledge of the mind-independent world.[4] Most scientific realists make the following related claims: the semantic claim that scientific theories should be taken at face value as making truth-evaluable claims; the epistemological claim that accepting a theory involves believing that it is true; and the metaphysical claim that a theory is ontologically committed to the entities that it posits, whether

[3] Danks (2014) also notes this tendency to interpret rational analysis approaches to Bayesian modeling as instrumentalist; like us, however, he thinks this conflation should be avoided. We discuss this further in Section 3.

[4] We will remain largely neutral with respect to the relationship between scientific theories and scientific models. For a more nuanced discussion, see Frigg and Hartmann (2020).

observable or unobservable.[5] Scientific antirealism can take a number of different forms, depending on which of these commitments it rejects. Most prominent is instrumentalism, which claims that our best scientific theories do not provide us with knowledge of the unobservable world, and instead are merely useful tools or instruments for practicing science.[6]

Scientific realism proposes that scientific explanations provide us with knowledge of the objective world, ontologically committing us to the entities which do explanatory work in a scientific theory or model. Following Psillos, we can call this the 'explanatory criterion' on reality: "something is real if its positing plays an indispensable role in the explanation of well-founded phenomena" (Psillos 2005, p. 389). It is important to understand that the explanatory criterion itself is a permissive one, which does not place any restrictions on the kinds of things which are real, beyond their explanatory role. In particular, the explanatory criterion does not require that real entities are concrete entities.[7]

Some scientific realists add further constraints to the explanatory criterion, proposing that scientific explanations must be causal explanations, and that only concrete entities can figure in causal explanations. But these further constraints require additional argument and should not be mistaken for necessary conditions on scientific realism itself. It is widely accepted that science makes use of non-causal explanation in addition to causal explanation: Saatsi (2021), for example, argues that physics features a 'menagerie' of non-causal explanations which appeal to geometry, symmetry, and intertheoretic relations (see also Reutlinger and Saatsi 2018 and Lange 2016). Even if we focus specifically on causal explanations, it is unclear that we must be committed only to concrete entities: it might be suggested that abstract entities can figure in causal explanations (e.g. Kersten 2020) or be explanatorily relevant without being causally relevant (e.g. Pincock 2015). As Psillos (2005) emphasizes, the explanatory criterion on scientific realism should not be confused with a causal criterion.

These concerns are familiar from Quine's 'indispensability argument', which was originally used to argue that the abstract mathematical structures which are essential to so much scientific theorizing

[5] Structural realism is a form of scientific realism which reconsider the ontological claim to suggest that we should be committed not to entities but only to the structural content of our theories. We will set aside structural realism for the rest of this paper, but see Ladyman (2014) for an overview.
[6] There are different routes to this conclusion: see Stanford (2016) for further discussion.
[7] As Psillos puts it, the explanatory criterion "does not dictate the status of entities that are explanatorily indispensable; in particular it does not disallow abstract entities from being real" (Psillos 2005, 389).

are real. Versions of the indispensability argument have been used to argue that we should be ontologically committed to other non-concrete entities where they play an essential explanatory role in scientific theorizing. Psillos (2011), for example, proposes that if non-concrete entities such as frictionless planes, ideal gases, perfectly spherical objects, and mass-points play an indispensable role in our best scientific theories, then such entities are real.

The permissiveness of the explanatory criterion on scientific realism also allows that at least some forms of abstraction and idealization are compatible with scientific realism. There is a sense in which all scientific models involve a process of abstraction: we use models to theorize about real-world phenomena because models are simpler and easier to manipulate than the phenomena themselves, allowing us to focus on particular entities, properties and relations at the expense of others.[8] Leaving out details does not entail saying anything false or inaccurate, and thus abstraction alone does not seem to pose any challenges to scientific realism. Some scientific models, however, also involve a process of idealization: they distort the nature of certain parameters, deliberately mispresenting the world. While some forms of idealization are doubtless incompatible with realism, scientific realists can allow for idealizations insofar as they maintain *approximate truth*.[9] As Eliot-Graves and Weisberg point out, "[r]ealists can argue that judicious idealizations are sensitive to the way that the world really is" (Eliot-Graves and Weisberg 2014, p.183). We propose that questions about abstraction and idealization are largely orthogonal to questions about realism and instrumentalism. Following Danks (2014), we suggest that the debate between realists and instrumentalists concerns how to interpret the commitments of theory or model, while questions about abstraction and idealization concern the dimension of approximation: what falls within the scope of a theory and what is excluded?[10]

[8] Determining the target system of a scientific model is a matter of identifying the domain of study and determining which parameters to focus on and which to omit: see Eliot-Graves (2020) on target systems, and the importance of deciding the level of grain at which to partition the domain. See also Frigg and Nguyen (2017).

[9] Cashing out what approximate truth might be is a further challenge which we will not address here. See Chakravarrty (2011) for further discussion of both formal and informal explications of the concept.

[10] See Chapter 2 of Danks (2014) for further discussion. A similar point is made by Weiskopf (2011), who distinguishes between the level of precision or 'grain' of a theory and its correctness. Eliot-Graves and Weisberg (2014) propose that neither the realist nor the anti-realist can appeal to idealization to make their case.

In this section, we have focused on the explanatory criterion for scientific realism and suggested that scientific realism *per se* is compatible with non-causal explanation and non-concrete entities, as well as some forms of abstraction and idealization. In the following section, we will apply these considerations to the debate over Bayesian models in cognitive science to demonstrate that black-box approaches to Bayesian inference can be given a realist interpretation rather than a merely instrumentalist interpretation.

3. Reconsidering Bayesian realism

As we saw in Section 1.2, there is a tendency for philosophers to interpret black-box approaches to Bayesian cognitive science (such as rational analysis) as instrumentalist: these approaches are treated merely as predictive tools, rather than as explanatory theories with ontological commitments. We propose here that black-box approaches to Bayesian models in cognitive science can be genuinely explanatory, and therefore open to a realist interpretation.

First, notice that there is nothing essentially non-causal about black-box approaches in general: they can be characterizing a causal relationship between inputs and outputs even where they are abstracting away from (or "screening off") the mediating mechanisms. In Bayesian cognitive science, however, black-box approaches are usually proposed as formal models, which involve abstracting from the causal relations to focus on formal relations. Even if we take causal explanations to be the norm in the physical sciences, this is less obviously the case in the special sciences: psychological explanations, as Weiskopf (2011) emphasizes, seem to come in causal and non-causal varieties. Once we accept that the explanatory criterion on scientific realism is not necessarily a causal criterion, then there is a *prima facie* case to be made that formal models are genuinely explanatory and not merely predictive.[11]

There is, however, a further motivation to give an instrumentalist interpretation of black-box approaches to Bayesian cognitive models. Several philosophers (e.g. Colombo and Series 2012,

[11] Bechtel and Shagrir, for example, take proponents of rational analysis to be offering probabilistic models of cognition which "provide *explanatory* mathematical theories of a cognitive capacity without referring to specific psychological and neural mechanisms" [Bechtel and Shagrir 2015, p.314, our italics]. Reijula similarly claims that "[r]ational analysis is an account of how probabilistic modeling can be used to construct non-mechanistic but self-standing *explanatory* models of the mind" (Reijula 2017, p.2975, our italics).

Block 2018) have proposed that the *idealizations* involved in Bayesian modelling motivate an instrumentalist interpretation of Bayesian models in cognitive science. Do black-box approaches to Bayesian cognitive models involve the sort of distortions which would make them incompatible with scientific realism? Rational analysis models, for example, seem to rely on the notion of *optimal* or *ideal* reasoning: cognitive processes which minimize expected cost with respect to a specific cost function.[12] But as we suggested in Section 2, at least some forms of idealization are compatible with scientific realism. Optimality explanations are widely accepted in biological sciences, for example, as genuinely explanatory.[13] If frictionless planes and ideal gases can be posits of scientific theories without leading to anti-realism, as Psillos (2011) suggests, then why think that positing ideal reasoners is any more problematic? A second sort of idealization associated with Bayesian models concerns their computational *intractability*. Block (2018) and Mandelbaum (2019) suggest that where the processes involved in calculating Bayesian likelihoods are computationally intractable, a Bayesian model cannot be given a realist interpretation. But the appeals to approximate truth (considered in Section 2) which are common throughout scientific realism would seem to address this concern: where our psychological models approximate idealized Bayesian inference through tractable computations, there is no need to resort to instrumentalism.[14] We thus follow Danks in concluding that "the close tie between rational analyses and instrumentalist theories is unwarranted" (Danks 2008, p.67).[15]

In this section, we have suggested that black-box approaches to Bayesian modelling in cognitive science need not be understood as merely predictive: formal Bayesian models like rational analysis can be genuinely explanatory, therefore deserving of a realist interpretation rather than an instrumentalist one. The fact that a black-box approach abstracts away from the mediating mechanisms does not entail that it lacks ontological commitments. We propose that the onus is on

[12] Notice that there is nothing about black-box approaches to Bayesian cognitive models which demand their optimality: while all optimal inference is Bayesian, it is not the case that all Bayesian inference is optimal (Ma 2012). Insofar as rational analysis models require optimality, however, concerns about idealization will arise.
[13] The best explanation of the life-cycles of cicada populations, for example, refers to the evolutionary optimality of mathematically prime periods for minimizing intersection with other creatures' life-cycles. Rice (2012) considers both causal and non-causal interpretations of optimality explanations.
[14] See Rescorla (2019) for further discussion. A similar point is made by Kirchoff et al (forthcoming).
[15] Two further concerns about idealization are sometimes levelled at Bayesian cognitive models. The first draws on the connection between Bayesian inference and rational normativity to suggest that Bayesian cognitive science does not offer descriptive scientific theories (see, e.g., Mandelbaum 2019); for a response, see Rescorla (2016). A second concern appeals to the competence/performance distinction in cognitive psychology to suggest that Bayesian theories idealize away from performance limitations (see Franks 1995); Patterson (1998) provides a response.

the instrumentalist to establish that the kinds of idealization involved in Bayesian cognitive science are any more problematic than the sorts of idealization involved in scientific models of ideal gases and frictionless planes.

Conversely, we propose that mechanistic approaches to Bayesian cognitive science do not have to be given a realist interpretation. Unlike black-box approaches, mechanistic approaches to Bayesian cognition focus on modeling the information-processing which mediates between inputs and outputs of the cognitive system. While these models are often given a realist interpretation, it is also possible to construe them instrumentally such that we are merely talking *as if* there are neural representations and unconscious inference: it is possible to give instrumentalist, fictionalist, and eliminativist interpretations of mechanistic information-processing models.[16]

The upshot of this is that the distinction between realist and instrumentalist interpretations of a Bayesian theory is logically independent from the distinction between black-box and mechanistic approaches to Bayesian modelling. In the following section we consider *why* the Bayesian debate in cognitive science between black-box and mechanistic approaches has become misleadingly characterized in terms of instrumentalism and realism. We propose that there are three main reasons: the first related to recent work on mechanistic explanation, the second related to Marr's levels of analysis, and the third related to ideas about representation.

4. The sources of the conflation

4.1 Mechanistic misunderstandings

There is a recent trend in philosophy of science to focus on the role of mechanisms in scientific explanation. A mechanism, for these purposes, is a concrete system composed of causal entities organized in such a way that their activities and interactions produce a scientific phenomenon of interest. The proponents of this 'new mechanist' approach focus on a particular subset of causal explanation: scientific discovery and explanation are taken to be the discovery and explanation of

[16] For examples of instrumentalist, fictionalist, and eliminativist interpretations of neural information processing, see Sprevak (2013) and Drayson (2022).

causal mechanisms (Machamer, Darden and Craver, 2000).[17] Some proponents of the 'new mechanist' approach go so far as to suggest that all scientific explanations are mechanistic, or that a scientific theory is explanatory in virtue of its mechanistic nature.[18] According to this view, formal models are not genuinely explanatory: they merely provide a framework or schema which does not become genuinely explanatory until it is cashed out with mechanistic details. Applied to Bayesian cognitive science, this would suggest that rational analysis models are not explanatory unless they are accompanied by 'Bayesian brain' models of neural mechanisms, and thus that we cannot give a realist interpretation of rational analysis models alone. We acknowledge the importance of mechanistic explanations in science but reject the claim that only mechanistic models explain, for reasons already discussed in Section 2. Cognitive science, in particular, has a history of embracing both mechanistic and non-mechanistic explanations.[19]

4.2 Marrian misunderstandings

Some philosophers propose that black-box approaches to Bayesian cognitive science must be given an instrumentalist interpretation on the grounds that they focus on what Marr (1982) calls the 'computational level' of analysis. We propose that once we get clear about realism and instrumentalism concerning physical computation and the correct understanding of Marr's levels of analysis, it should be obvious that there is nothing essentially instrumentalist about Marr's computational level.

Consider how the distinction between realism and instrumentalism can be applied specifically to physical computation. A realist about physical computation proposes that when we describe a

[17] Mechanistic explanation goes beyond mere causal explanation of entities and their activities: it must "further describe how those entities and activities are organized (e.g., spatially and temporally) into a mechanism" (Craver 2006, p.373).
[18] Craver allows that perhaps not all explanations are mechanistic, but proposes that "in many cases [...] the distinction between explanatory and non-explanatory models seems to be that the latter, and not the former, describe mechanisms" (Craver 2006, p.367).
[19] Levy and Bechtel (2013) make a similar point when they argue, in direct contrast to Craver, that abstract models play a role in explaining the behavior of particular systems: the process of abstraction both identifies the relevant causal organization and facilitates generalization. Weiskopf (2011) argues that we should not be misled into thinking that cognitive models are mechanistic even where their structure resembles that of mechanistic models.

physical system as performing a computational function, we are making a claim about the mind-independent world:

> "realism about [physical] computation [...] is the view that whether or not a particular physical system is performing or implementing a particular computation is at least sometimes a fact that obtains independently of human beliefs, desires and intentions." (Ladyman 2009, p. 377)

An instrumentalist about physical computation can thus be characterized as claiming that when we describe a physical system as performing a computational function, we are making a claim which is in some sense relative to our own interests, goals, or background assumptions. Hardcastle (1995) proposes such a view:

> "whether a physical system is actually computing [...] depend[s] upon the interests and aims of the people involved in the investigation. [...] whether the assignment of a function to a physical system counts as an explanation depends upon the contingent interests of the relevant community." (Hardcastle 1995, p.314)

How does this distinction between realism and instrumentalism about physical computation relate to Marr's levels of computational analysis? Marr (1982) proposed that physical computers or information processing systems can be described and analyzed at three distinct levels. We can ask what formal function the system is performing (the computational level), what specific algorithms or programs it is using to perform the function (the algorithmic level), and which specific hardware is implementing these programs (the implementation level). An important feature of physical computation is that one computational function can be performed by many different algorithms, which in turn can be implemented by many distinct kinds of hardware. When we specify an information-processing systems at Marr's computational level, therefore, facts about the algorithmic and implementation levels remain underdetermined. But this underdetermination should not be confused with an agnosticism or skepticism about *whether there is* a physical implementation of the computational model. As a result, there is nothing essential instrumentalist about Marr's computational level.

In the literature on Bayesian cognitive science, arguments for instrumentalism sometimes proceed via claims about Marr's computational level.[20] Such arguments tend to start from the claim that black-box approaches to Bayesian inference (rational analysis, in particular) have a special connection to Marr's computational level.[21] We are focused here on the next step of these arguments: the claim that Marr's computational level deserves an instrumentalist rather than a realist interpretation. We reject this move for the reasons articulated above. Each of Marr's levels of computational analysis can be given a realist or an instrumentalist interpretation: even if Bayesian rational analysis has a special connection to Marr's computational level, there is no straightforward argument for an instrumentalist interpretation of these Bayesian models.[22]

Marr's levels of analysis provide a methodological tool which allows us to focus on different ways to understand computational systems in cognitive science. Both the realist and the instrumentalist about physical computation can adopt Marr's framework and make claims at each of the three levels of analysis, because Marr's framework is largely neutral with respect to these questions about theory interpretation.[23] It is therefore a mistake to think that computational-level claims must be given an instrumentalist interpretation.

4.3 Representational misunderstandings

Bayesian cognitive models are sometimes considered to be instrumentalist if they are not ontologically committed to the existence of representational vehicles (e.g. neurons) which explicitly encode probabilities: Block, for example, proposes an instrumentalist interpretation of Bayesian

[20] Block, for example, claims that neural processes "can be considered Bayesian but only on an instrumentalist interpretation pitched at Marr's computational level rather than the algorithmic level" (Block 2018, p.8).
[21] See, for example, Griffiths et al. (2012); Tenenbaum, Griffiths, & Kemp (2006, p.206), Jones and Love (2011), Oaksford and Chater (2007), and Icard (2018).
[22] While we focus on rejecting the link between Marr's computational level and instrumentalism, there may also be reason to question the link between rational analysis and Marr's computational level: see Kitcher (1988) and Bechtel and Shagrir (2015). Danks also suggests that instrumentalist interpretations of rational analysis derive "largely from the connection with the computational level of Marr's trichotomy" which he proposes is "neither necessary nor desirable" (Danks 2008, p.67).
[23] Similar points are made by Danks and Egan. Danks proposes that questions of theory interpretation (e.g. realism, scope, optimality) are not settled by Marr's trichotomy or by computational models in cognitive science more generally: instead, "we need to do some philosophy to really understand what the cognitive science means" (Danks 2014, p.16). Egan points out (also in relation to Marr's framework) that if theory interpretation could be read so easily off our computational models, "much of the philosophy of science would be out of business" (Egan 1995, p.186).

cognitive models which "are not committed to the representation in real visual systems of priors or likelihoods or their multiplication within the system" (Block 2018, p.8). We propose that a realist interpretation of Bayesian models does not require the explicit encoding of priors and likelihoods. As Ma (2012) points out, we can distinguish Bayesian inference from computing with probability distributions: the fact that a brain performs Bayesian inference (and even does so optimally) does not imply that neurons encode probabilities. A similar point is made by Rescorla, who argues that realism only requires approximate conformation to Bayesian norms and concludes that "[i]n rejecting explicit enumeration of credences, Block is not rejecting realism" (Rescorla 2019, p. 59).

As we have suggested above, to be a realist about physical computation is to think that whether a physical system implements an abstract computation is a mind-independent matter. Philosophers have widely differing views on what it is to physically implement a computation, however, and each of these views will result in a different variety of realism.[24] Some philosophers propose that realism about Bayesian cognitive models does not commit one to any claims about representation, for example. Orlandi, for example, argues that one can be a "fairly robust realist" about Bayesian models of perception without thinking that such models posit representations (Orlandi 2016, p.342): Bayesian priors and likelihoods might merely be functional features or biases operating over non-representational causal states.[25] Anderson (2017) proposes that our brains could implement Bayesian computation by reconfiguring or guiding the parameters of a control system, rather than by updating internal representations. These non-representational versions of Bayesian realism are, however, controversial: if we assume that inference is a semantically evaluable process, then Bayesian inference would seem to require some form of representation. But even if one commits to a representational view of Bayesian computational models, there is still room for debate about the nature of those representations, depending on how inflationary or deflationary a notion

[24] As Ladyman argues, "unless we have a precise account of implementation it will not be possible to decide whether or not realism is correct just because it will not be clear what 'computation' means" (Ladyman 2009, p. 377). Williams (forthcoming) further explores the role that implementation theories play with respect to realism about physical computation.

[25] In a later paper, Orlandi (2018) proposes that where a system updates according to Bayes' theorem, this suggests a representational picture.

of representation we adopt.[26] Some deflationary notions of representation can allow that hypotheses or credences are implicitly represented rather than explicitly encoded.

4.4 Summary

In Section 4, we have explored some of the motivations which drive certain philosophers to conflate black-box Bayesian approaches with instrumentalism, or to conflate mechanistic Bayesian approaches with realism. Some philosophers assume that scientific explanations must be mechanistic, and thus that non-mechanistic approaches can only be instrumental; some assume that non-mechanistic theories are at Marr's computational level and that Marr's computational level must be given an instrumentalist interpretation; and some assume that realism about physical computation is only compatible with a particularly strong claim about the sorts of representations involved in Bayesian computational models. Each of these claims requires further argumentation, and does not follow from the fact that both black-box and mechanistic approaches to Bayesian cognitive science exist.

5. Conclusion

This paper has explored ways of understanding black-box and mechanistic approaches to Bayesian models of cognitive science. As we highlighted in Section 1, the philosophical literature has exhibited a tendency to map these two different approaches onto the distinction in philosophy of science between realist and instrumentalist interpretations of a theory or model. Our main contention in this paper is that this tendency should be resisted, because there are two separate issues in play. One issue concerns the relationship between input-output models of computational functions and mechanistic models of physical computation, while a second issue concerns the ontological, semantic and epistemological interpretations of these models. We have argued that both black-box and mechanistic approaches to Bayesian cognitive science can be given realist or

[26] See Ramsey (2021) for a discussion of how Chomsky's deflationary view of representation and Egan's quasi-deflationary view of representation compare to more robust notions of representation in cognitive science.

instrumentalist interpretations. To be a realist about either a mechanistic or a non-mechanistic model is to think that the model provides explanations of the objective world which are ontologically committed to the existence of mind-independent entities. Scientific realism alone does not logically entail that all entities are concrete or that all explanations describe causal mechanisms; many self-professed scientific realists reject one or both of these constraints. To be an instrumentalist about either a black-box or a mechanistic model is to think that the model is not ontologically committed to mind-independent entities: perhaps because the model fails to refer, perhaps because we are not justified in forming beliefs about the world on the basis of these models, or perhaps because the entities involved are relative to our interests. The mere fact that a model is mechanistic or non-mechanistic does not tell us whether it should be given a realist or instrumentalist interpretation.

We argued for this conclusion by first considering scientific realism and instrumentalism more generally, and then applying these considerations to Bayesian cognitive science. Considerations of the explanatory criterion and indispensability arguments suggest that a realist Bayesian model can posit abstract as well as concrete entities, provide explanations which go beyond descriptions of causal mechanisms, and incorporate processes of abstraction and idealization without sacrificing its realist credentials. None of these claims are particularly controversial in the literature on scientific realism: notice that they are compatible with acknowledging that some abstract entities are not scientifically explanatory, that causal explanations can have benefits that non-causal explanations lack, and that some forms of idealization are in tension with scientific realism. So why do Bayesian approaches to cognitive science take a narrower view of scientific realism, on which realism is aligned with descriptions of concrete mechanisms, and anything else is considered instrumentalist? We have proposed that this unnecessarily narrow construal of realism is motivated by one or more misunderstandings about cognitive science. It is a mistake, we argued, to think that explanations in cognitive science must be wholly mechanistic, or that the instrumentalism of non-mechanistic computational descriptions is written into Marr's framework, or that realism about physical computation dictates specific requirements about the nature of representation.

Debates about scientific realism and instrumentalism are philosophical debates about the metaphysical, epistemological and semantic interpretations of scientific theories and models. But there is more to theory interpretation than the realist/instrumentalist dimension: questions about

optimality, approximation and abstraction, for example, are not answered simply by labelling a theory as realist or instrumentalist. This point is nicely articulated by Danks, who argues that the claims of cognitive science do not themselves dictate the complex continuum of commitments that interest us when interpreting the theories in question:

> "The specification of a cognitive theory — whether framework, architecture, or model — almost never (in isolation) commits one to any particular picture of the world, or constrains the ways that theory could be implemented, or determines how we could confirm or learn about the truth of that theory." (Danks 2014, p. 16)

6. References

Anderson, John R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences* 14 (3):471-485.

Anderson, M.L. (2017) Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. *Philosophy and Predictive Processing*. Brain and Mind Institute Researchers' Publications. 53.

Bechtel, William & Shagrir, Oron (2015). The Non-Redundant Contributions of Marr's Three Levels of Analysis for Explaining Information Processing Mechanisms. *Topics in Cognitive Science* 7 (2): 312-322.

Block, Ned (2018). If perception is probabilistic, why doesn't it seem probabilistic? *Philosophical Transactions of the Royal Society B* 373 (1755).

Chakravartty, Anjan (2011). Scientific Realism. *Stanford Encyclopedia of Philosophy*.

Chater, Nick & Oaksford, Mike (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences* 3 (2):57-65.

Colombo, Matteo & Seriès, Peggy (2012). Bayes in the Brain—On Bayesian Modelling in Neuroscience. *British Journal for the Philosophy of Science* 63 (3):697-723.

Craver, Carl F. (2006). When mechanistic models explain. *Synthese* 153 (3):355-376.

Danks, David (2008). Rational analyses, instrumentalism, and implementations. In Nick Chater & Mike Oaksford (eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford University Press. pp. 59-75.

Danks, David (2014). *Unifying the Mind: cognitive representations as graphical models*. MIT Press.

Drayson, Zoe (2022). What we talk about when we talk about mental states. In Tamas Demeter, T. Parent & Adam Toon (eds.), *Mental Fictionalism: Philosophical Explorations*. Routledge.

Egan, Frances (1995). Computation and content. Philosophical Review 104 (2):181-203.

Elliott-Graves, Alkistis (2020). What is a Target System? *Biology and Philosophy* 35 (2):1-22.

Elliott-Graves, Alkistis & Weisberg, Michael (2014). Idealization. *Philosophy Compass* 9 (3):176-185.

Franks, Bradley (1995). On explanation in cognitive science: Competence, idealization, and the failure of the classical cascade. *British Journal for the Philosophy of Science* 46 (4):475-502.

Frigg, Roman & Hartmann, Stephan (2020). Models in Science. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Stanford.

Frigg, Roman & Nguyen, James (2017). Models and representation. In Lorenzo Magnani & Tommaso Bertolotti (eds.), *Springer Handbook of Model-Based Science*. pp. 49-102.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138*(3), 415–422.

Hardcastle, Valerie Gray (1995). Computationalism. *Synthese* 105 (3):303-17.

Icard, Thomas F. (2018). Bayes, Bounds, and Rational Analysis. *Philosophy of Science* 85 (1):79-101.

Jones, Matt & Love, Bradley C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences* 34 (4):169-188.

Joyce, James (2008). Bayes' theorem. *Stanford Encyclopedia of Philosophy*.

Kersten, Luke (2020). How to be concrete: mechanistic computation and the abstraction problem. *Philosophical Explorations* 23 (3):251-266.

Kirchhoff, Michael, Kiverstein, Julian and Robertson, Ian. (forthcoming) The Literalist Fallacy & the Free Energy Principle: Model-building, Scientific Realism and Instrumentalism. British Journal for the Philosophy of Science.

Kitcher, Patricia (1988). Marr's Computational Theory of Vision. *Philosophy of Science* 55 (March):1-24.

Ladyman, James (2009). What Does it Mean to Say a Physical System is Implements a Computation? *Theoretical Computer Science* 410 (4-5).

Ladyman, James (2014). Structural Realism. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University.

Lange, Marc (2016). *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. Oxford, England: Oxford University Press USA.

Levy, Arnon & Bechtel, William (2013). Abstraction and the Organization of Mechanisms. *Philosophy of Science* 80 (2):241-261.

Ma, Wei Ji. (2012) Organizing probabilistic models of perception. *Trends in cognitive sciences* 16.10: 511-518.

Machamer, Peter, Darden, Lindley & Craver, Carl F. (2000). Thinking about mechanisms. *Philosophy of Science* 67 (1):1-25.

Mandelbaum, Eric (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind and Language* 34 (2):141-157.

Marr, David (1982). *Vison*. W. H. Freeman.

Oaksford, Mike & Chater, Nick (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford University Press.

Orlandi, Nico (2016). Bayesian Perception Is Ecological Perception. *Philosophical Topics* 44 (2):327-351.

Orlandi, Nico (2018). Predictive perceptual systems. *Synthese* 195 (6):2367-2386.

Patterson, Sarah (1998). Competence and the Classical Cascade: A Reply to Franks. *British Journal for the Philosophy of Science* 49 (4):625-636.

Pincock, Christopher (2015). Abstract Explanations in Science. *British Journal for the Philosophy of Science* 66 (4):857-882.

Psillos, Stathis (2005). Scientific realism and metaphysics. *Ratio* 18 (4):385–404.

Psillos, Stathis (2011). Living with the abstract: realism and models. *Synthese* 180 (1):3-17.

Ramsey, W. (2021). Defending Representation Realism. In Joulia Smortchkova, Krzysztof Dołęga, and Tobias Schlicht (eds.) *What are Mental Representations?* 54-78. Oxford University Press.

Reijula, Samuli (2017). "How could a rational analysis model explain?" COGSCI 2017, Proceedings of the 39th Annual Conference of the Cognitive Science Society, 2975- 2980.

Rescorla, Michael (2016). Bayesian Sensorimotor Psychology. *Mind and Language* 31 (1):3-36.

Rescorla, Michael (2019). A realist perspective on Bayesian cognitive science. In Anders Nes and Timothy Chan (eds.) *Inference and Consciousness*. Routledge NY.

Reutlinger, Alexander & Saatsi, Juha (eds.) (2018). *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*. Oxford University Press.

Rice, Collin C. (2012). Optimality explanations: a plea for an alternative approach. *Biology and Philosophy* 27 (5):685-703.

Saatsi, Juha (2021). Non-causal explanations in physics. In Eleanor Knox & Alastair Wilson (eds.), *The Routledge Companion to Philosophy of Physics*. Routledge.

Sprevak, Mark (2013). Fictionalism about Neural Representations. *The Monist* 96 (4):539-560.

Sprevak, Mark (2016). Philosophy of the psychological and cognitive sciences. In Humphreys, Paul (ed.) *The Oxford Handbook of Philosophy of Science*. Oxford University Press.

Stanford, P Kyle (2016). Instrumentalism: Global, local, scientific. In Humphreys, Paul (ed.) *The Oxford Handbook of Philosophy of Science*. Oxford University Press.

Tenenbaum, Joshua B.,Griffiths, Thomas L. & Kemp, Charles (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences* 10 (7): 309-318.

Weiskopf, Daniel A. (2011). Models and mechanisms in psychological explanation. *Synthese* 183 (3):313-338.

Williams, Danielle (forthcoming) Markov blankets: realism and our ontological commitments. Commentary on Jelle Bruineberg, Krzysztof Dolega, Joe Dewhurst, and Manuel Baltieri, 'The Emperor's New Markov Blankets'. *Brain and Behavioral Sciences*.