Book Review

Jeff Buechner, *Gödel, Putnam, and Functionalism: A New Reading of Representation and Reality*, MIT 2008

Witold M. Hensel[1] and Marcin Miłkowski[2]

[1]*Faculty of History and Sociology, University of Białystok*
whensel@poczta.onet.pl
[2]*Institute of Philosophy and Sociology, Polish Academy of Sciences*
mmilkows@ifispan.waw.pl

The paper is a critical review of the book *Gödel, Putnam, and Functionalism: A New Reading of Representation and Reality* by Jeff Buechner, which is a defense of computational functionalism against arguments formulated by Putnam, Searle, Fodor, Lucas and others. Buechner, after having meticulously analyzed these arguments, concludes that all of them fail to show that computational functionalism is not a viable strategy to model the mind in cognitive science. As such, it is a defense of a mathematically-informed version of computational functionalism. We discuss Beuchner's strategy in quite a bit of detail and make some comments.

Key words: *computationalism, functionalism, Gödel, Putnam*

The primary aim of Jeff Buechner's *Gödel, Putnam, and Functionalism* is to defend computational functionalism (henceforth functionalism) against a sustained attack mounted on it by Hilary Putnam in *Representation and Reality*. But *Gödel, Putnam, and Functionalism* is more than a critical discussion of a single philosophical work. It is a thorough examination of some of the most influential lines of argument taken against functionalism and cognitive science not only by Putnam, but also by Penrose, Searle, Fodor and others.

Chapter 1 opens with a discussion of anti-mechanist arguments appealing

Journal of Cognitive Science 15: 391-402, 2014

to Gödel's incompleteness theorems. Buechner divides those arguments into metaphysical and epistemic ones. A metaphysical argument (henceforth MGM) seeks to demonstrate that the human mind is not a finitary machine. The best-known MGM, presented by Lucas (1961), appeals to Gödel's first incompleteness theorem, but Buechner focuses on its modified version, based on Gödel's second theorem (unfortunately, he clarifies this in an end note, which is easy to miss, so a reader who is familiar with the Lucas argument is likely to wonder why the argument is being "misrepresented").

Suppose that machine M is proposed as a true description of the human mind. Plausibly, M is both capable of doing arithmetic and consistent. According to an MGM, a human agent presented with *any* such M will be able to prove that M is consistent. But if someone is able to prove the consistency of M then M cannot be a true description of her mind. This is because, by Gödel's second incompleteness theorem, if M is consistent then M cannot prove it. It follows that, for any consistent M capable of doing arithmetic, M is not a true computational description of the human mind.

MGMs attempt to establish the existence of a metaphysical divide between minds and machines by showing that human cognition surpasses the capacities of any computer program. However, as Putnam (1995) observed, MGMs cannot secure their conclusion unless showing the consistency of an arbitrary machine is humanly possible. This is very important, since Gödel's second incompleteness theorem only applies to consistent systems: given that anything follows from a contradiction, if M is inconsistent, the formula expressing M's consistency *is* derivable by M. The rub is that if M's description is too long for anyone to survey (and there is no reason to expect a machine equivalent to the human mind to have a short description) then no human agent will be able to prove M's consistency and the anti-functionalist's argument won't get off the ground. This is so whether or not our minds are subject to Gödel's theorem.

An epistemic argument (henceforth EGM) makes a weaker claim than an MGM. It is also silent on how humans may differ from machines. What it aims to show is that, even if the mind *is* a finitary machine, we could never know whether any proposed computational description of it is consistent. In rough outline, the argument goes as follows. Let C be a computer program describing human cognition. If humans can prove truths of Peano

arithmetic, so can C. But if that is the case then C is Gödel-susceptible: by Gödel's second incompleteness theorem, C cannot prove its own consistency. Therefore, cognitive science will never be able to prove the consistency of any proposed computational description of human cognition.

Buechner makes an important observation, which applies to both MGMs and EGMs. The second incompleteness theorem implies that the consistency statement of a Gödel-susceptible formal system cannot be proved *mathematically with mathematical certainty*. It is therefore an open question whether it can be proved mathematically with less than mathematical certainty (say, by some statistical reasoning) or in another epistemic modality (i.e., not within mathematics). This creates a gap in the anti-functionalist's reasoning: the Gödel theorem puts no genuine limits on a cognitive science that does not aim at mathematical certainty.

In order to close that gap, the anti-functionalist must now demonstrate that each method for establishing the consistency of C with less than mathematical certainty or in another epistemic modality is either Gödel-insusceptible or unwarranted. Buechner argues that this cannot be done, for, plausibly, the number of such methods is infinite (cf. pp. 49-57).

In chapter 2, Buechner discusses the possibility that the anti-functionalist would be back in the game if she could prove that all epistemically justified methods for establishing the consistency of C with less than mathematical certainty or in another epistemic modality are Gödel-susceptible. This is precisely what Putnam tries to accomplish in 'Reflexive Reflections' (Putnam 1994).

Since a detailed discussion of Putnam's proof would take us too far afield, we must confine ourselves to bare essentials (for details see Buechner, pp. 59-62). The important thing is that the proof, which appeals to Gödel's second incompleteness theorem, hinges on the possibility of arithmetizing the notion of justification. Buechner considers several methods of formalizing the notion in a way that would preserve Putnam's inference, but none of them appears viable.

In *Representation and Reality*, Putnam makes an even stronger claim than in 'Reflexive Reflections': that all our methods of inquiry into the world, such as demonstrative inference, inductive inference, rational interpretation, reasonable reasoning and general intelligence, are susceptible to Gödel's

theorems. If this is the case then, by Gödel's second incompleteness theorem, we cannot establish their consistency with *any* degree of credibility.

However, according to Buechner, Putnam's claim engenders a paradox. Suppose Putnam is right: we cannot establish the consistency of the totality of our methods of inquiry into the world with any degree of credibility. In order to rationally accept Putnam's claim, we would have to know that its proof is correct. However, if all our methods of inquiry are Gödel-susceptible then we have no reason to expect the methods used in the proof to be consistent. Therefore, Putnam's claim is unwarranted. But if we cannot establish Gödel-susceptibility of all our methods of inquiry then it is possible that not all such methods are Gödel-susceptible. If that is the case, we can justifiably use them in the proof of Putnam's theorem. However, if the proof succeeds, we will have to acknowledge, again, that the claim is unjustified. Briefly: Putnam's proof succeeds if and only if Putnam's proof fails.

The next three chapters focus on so-called triviality arguments, which purport to show that the computational theory of mind is trivial because everything can be said to implement some computation. This version of the triviality argument wouldn't be as disastrous to functionalism as a version of the argument showing that any computation whatsoever is implemented by any physical system. Indeed, this is supposed to be one of the most important results of *Representation and Reality*, where Putnam even offers a mathematical proof that this – in his opinion – is indeed the case. To be exact, he claims to have proven that any ordinary open physical system implements *every* finite-state automaton. Computational functionalism, and computational modeling in cognitive science as well, are doomed: there are no interesting predictions about any physical systems if these predictions are made in terms of computation, as they are true of anything. In other words, they are only trivially true, devoid of any informative value.

If the proof is correct then, even if Gödelian arguments fail, functionalism cannot be sustained. For this reason, Buechner is definitely right to scrutinize Putnam's theorem in chapter 4. Before doing that, he analyzes, in chapter 3, a simpler version of a triviality argument, presented by Edward Stabler (1987) in his analysis of Kripke's attempt to refute functionalism.

Buechner shows that Stabler's triviality argument (called "S-triviality") fails (cf. pp. 104-111). The argument is based on a rather loose definition of physical computation that requires there to be a mapping between a physical system and the steps of the computation. As there are no restrictions on the mapping function, we can pick any one we like, and a trivialist may pick one that will furnish us with a complete mapping of a computational system in any other system. And the argument will follow: indeed, if to physically compute for a physical system is to have a mapping function between physical states and the steps of the computation (or between the input/output of the computation), then the triviality argument is cogent. Useful replies to such arguments are based on the fact that the mapping does not furnish counterfactual-supporting predictions about the physical system (in a stronger version, one may say that a trivial mapping is not a causal model of the system). Buechner offers a different argument, based on complexity theory. Namely, he shows that the time complexity of the mapping function, or the number of steps in the computation as related to the number of input values to the algorithm, will be much lower than that of an actual computation; computing any mathematical function will take the same time, which is directly inconsistent with the results of complexity theory. Hence, S-triviality must be false.

In the next chapter, the focus is on Putnam's argument, which is more complex than S-triviality. Before presenting the argument, Buechner delves into a detailed discussion of the principles assumed by Putnam, and stresses again, although in a slightly obscure way, that a theory of error in computation is needed. Unless one knows which function the system is *supposed* to compute, one cannot say whether the system miscomputed the function (the author discusses this under the guise of the Kripke-Wittgenstein problem, though the discussion would be much more accessible in terms of functions, for the debate over the purported normative nature of function is very much relevant to the topic and, there, the notion of error has been discussed explicitly; cf. Dretske (1986)). These principles, the continuity principle and the principle of noncyclical behavior, are shown to be problematic from the physical point of view. What's worse, their use in Putnam's argument leads to a dilemma:

> First horn: If the Principle of Noncyclical Behavior has the modal status of a physical law, then either the ordinary macroscopic open systems to which it applies are a small fraction of all macroscopic open systems (and so it might exclude the human brain) or else it includes all macroscopic open systems (and so includes the human brain), but contradicts mathematical facts (such as the indeterminacy of classical Newtonian mechanics). Second horn: If the Principle of Noncyclical Behavior does not have the modal status of a physical law, then the counterfactuals defined in constructing the state of affairs necessary for the contradiction with the Principle of Continuity in the triviality lemma can turn out false. (Buechner 2008, p. 148)

Again, the triviality argument succumbs to replies appealing to modal considerations, as trivial ascriptions do not support the right counterfactuals. Buechner presents also another argument for triviality, called "step-function triviality", and shows that it cannot be sustained on pain of contradicting Kolmogorov complexity theory.

Chapter 5 discusses well-known arguments by John Searle for the subjective nature of computation, illustrated by Searle with an example of his wall supposedly computing (trivially) the Wordstar word-processing application. Buechner argues that if Searle is right that computation is observer-relative then Searle's argument is also observer-relative, and there might be two observers that ascribe vastly different content to such arguments. Certainly, Searle cannot think that his arguments are *not* expressed in any syntax; because they are, they cannot be intrinsic to the physical (despite Searle's protestations that they are intrinsically meaningful, their *public* meaning hinges upon an observer, so it's observer-relative). According to Buechner, Searle's relativism about computation leads to unrestricted relativism about anything, including his own arguments. Indeed, triviality trivialized. He goes on to show some bizarre consequences of Searle's view as well, e.g., ones derived from the double recursion theorem from computability theory (cf. pp. 173-176).

Chapter 6 focuses on multiple realization (MR), which is usually endorsed by functionalists, including the early Putnam. MR in its functionalist version asserts that there are infinitely many physical realizations of an

arbitrary functional state, which purportedly blocks reduction of the mental to the physical. In *Representation and Reality*, Putnam claims that the same kind of argument undermines functionalism, as there are infinitely many computational realizations of an arbitrary intentional state. Computational MR can spring from two sources, which Buechner calls *computational growth* and *content growth*.

Computational growth has to do with the fact that there are various models of computation used in cognitive science, a wide range of possible computer architectures and an indefinite number of programming languages; moreover, even if all those are fixed, different algorithms can be used to compute a single function.

Content growth is somewhat more complicated. To see how it arises, assume that there is an indefinite number N of English speakers and that there is a set of beliefs about cats common to all the speakers. Suppose further that each speaker also has at least one belief about cats that she shares with no other speaker. The upshot is that, for each speaker, if you encode all her beliefs about cats into a single computational state, no speaker will be in the same computational state as any other speaker. Accordingly, the computational realization of the intentional state expressed by the utterance "I see a cat", which we assume to be synonymous for all the speakers, is a disjunction of N computational states. Since this reasoning is easily applicable to any word, the number of such disjuncts is practically infinite.

As Buechner observes, this argument rests on the assumption that no computational model of human intentional states can capture the difference between ordinary beliefs and beliefs constitutive of meaning. Although Putnam tries to make this assumption plausible, the only model he considers is one according to which sentences of a language of thought are assigned utilities and degrees of confirmation that change according to Bayes's theorem. Given the model's paucity, it is no wonder that it lacks the resources to adequately represent meaning relations. Given its logical positivist pedigree, it is also no wonder that all theories of meaning formulated in terms of it turn out to be holistic. That, however, does not establish the stronger claim that *all* computational models are equally deficient.

However, there is, according to Buechner, a general argument for this stronger claim buried in several of Putnam's writings. The argument says that no computational model of human intentional states can distinguish between meaning-constitutive and auxiliary beliefs, because it is impossible to distinguish between them without appealing to the principles of rational interpretation. And the procedure of rational interpretation cannot be formalized, which means that it cannot appear in any computational model.

In fact, Putnam supports this general argument with an even more general piece of reasoning. He argues that functionalism is committed to a non-holistic conception of meaning (mental states are individuated by virtue of their content) and to metaphysical realism. Alas, according to Putnam, metaphysical realism entails indeterminacy of meaning, so the two commitments are mutually incompatible. Apparently, the only way to escape indeterminacy of meaning is to jettison the idea of truth as correspondence and adopt epistemic semantics.

Buechner replies to this by arguing that functionalism need not lead to Quinean indeterminacy because, *pace* Putnam, it is not committed to metaphysical realism. The upshot of chapter 6 is that, unless it is shown that rational interpretation cannot be formalized, computational growth is the more plausible source of computational MR than content growth.

However, establishing computational MR is not enough to refute computationalism, since the functionalist can insist that the infinite number of computational realizations of an intentional state can be reduced to a small number of physical-functional properties that all such realizations have in common. Chapter 7 introduces Putnam's EQUIVALENCE argument to the effect that this is impossible. More technically: according to EQUIVALENCE, there is no non-trivial and psychologically realistic computable partitioning of the infinitely large set of computational realizations of an arbitrary intentional state into a small set of equivalence classes. The reason is that such a reduction could only be accomplished by appeal to rational interpretation, and any algorithm for rational interpretation would have to be infinitary.

A large part of chapter 7 is devoted to assessing the conclusion of EQUIVALENCE, especially, though not exclusively, as applied to MR generated by computational growth. Buechner cites a number of

mathematical results that may be relevant in this connection, arguing that, in light of those results, Putnam's diagnosis is rash. Generally speaking, the trouble with MR generated by computational growth is that, though easy to establish, it is also easily reducible to a small number of equivalence classes (in which case EQUIVALENCE fails). The problem with MR induced by content growth is that it is not easy to establish at all, so, in its case, EQUIVALENCE may not even get off the ground.

In chapter 8, Buechner analyzes the assumptions Putnam uses to argue for EQUIVALENCE. These include the principles of rational interpretation, and in particular, the principles of synonymy determination. The overall strategy for establishing EQUIVALENCE is to claim that it is logically impossible to algorithmize rational interpretation and synonymy determination because the latter requires making coreferentiality decisions, which in turn require a vast amount of information about the speaker's environment. This includes information about all possible theories of the universe, since the speaker may inhabit any possible environment. Such requirements indeed make it difficult to formalize a coreferentiality algorithm.

Buechner shows, however, that Putnam's assumptions are implausible. Putnam claims that a formal theory of coreferentiality should 'anticipate' all future theories of the universe. But human beings do not fulfill this requirement, as they are not infinitary (and the set of all possible theories, Buechner argues, is not recursively enumerable). And similar points can be made about human rationality. Briefly, there is no reason to think that the coreferentiality algorithm should surpass the capacities of human beings. Moreover, Buechner shows that Putnam's use of 'rational interpretation' is ridden with equivocation, which, if removed, does not help the argument for EQUIVALENCE.

Buechner also shows that, contra Putnam, there are restrictions on the kinds of languages for which rational interpretation (in one of its meanings) is supposed to work. To do this, he uses an interesting argument based on Goodman's work on similarity. Goodman's (1951) claim, which has since acquired rigorous mathematical proof in the form of the so-called Ugly Duckling Theorems, asserts that, without restrictions on logical primitives, anything can be shown to be similar to anything else. Human cognitive systems have to be biased in some way. As Buechner argues, if we refuse

to put some natural constraints on human cognition, then cognitive science may well be impossible.

In chapter 9, Buechner comes back to rational interpretation and synonymy determination, though his arguments have already shown that Putnam's assumptions are implausible and they work only on pain of assuming an unrealistic account of human cognition. He discusses eight arguments for non-formalizability of rational interpretation; one based on incompleteness theorems; another on the purported non-formalizability of non-demonstrative reasoning; third based on Quinean indeterminacy; fourth based on Vico's view that interpretation requires formalizing a conception of what it is to be a human being; fifth based on radical Quinean holism; sixth based on necessary globality of rational interpretation; seventh based on Twin-Earth considerations; and the last one based on the fact that synonymy determination is essentially context-sensitive. The discussion is sometimes brief but clearly shows that these arguments are not successful. In the Appendix, Buechner discusses the form of an algorithm for rational interpretation as required for EQUIVALENCE.

All in all, Buechner offers an impressively detailed discussion of Putnam's arguments against functionalism and their assumptions, which alone would make the book under review worth reading. But he also makes some excellent points by drawing on current work in mathematics and philosophy of mathematics. That being said, the book is not without shortcomings. It ignores many discussions of computationalism and functionalism in the contemporary philosophy of science, where notions such as multiple realization, structural realism, natural kind, explanation and reduction have all received critical attention in the last couple of decades. As a result, Buechner's treatment of these notions is either shallow or unconvincing.

For example, he simply assumes that computationalism implies MR, without taking into account the recent critical discussion (Bechtel & Mundale 1999). Yet it is by no means obvious under what conditions two instantiations of the same computational system would actually count as different physical realizations of it. The crux of the problem is that, for MR to occur, the capacity realized in two systems has to stay the same, while the realizations must differ in a way relevant to the capacity. In that,

realization of a property is different from its instantiation. But it is not clear when two human brains, assuming that they have exactly the same computational capacity at some level of abstraction, exhibit a sufficient number of relevant physical differences to count as *different* realizations of that capacity. Buechner misses the point that functionalists, such as Block and Fodor (1972), have never plausibly shown that human psychological capacities are actually multiply realized if these capacities are precisely characterized (Polger 2008). In other words, one should ask whether extreme anti-reductionism, backed up by MR arguments, is actually the default assumption in computational modeling. Perhaps the actual practice of cognitive science presupposes a heuristic identity theory (Bechtel & McCauley 1999)?

In brief, if assumptions about MR, critically examined by philosophers of science in the last two decades or so, are not as simple as Buechner supposes them to be, then the whole theory of realization assumed by Putnam in his triviality proof and anti-reductionist arguments might be deficient, as it conflates instantiation with realization and, thus, cannot distinguish between relevantly different realizations and numerically different idealizations. Alas, Buechner is completely silent on alternative suggestions regarding the notion of physical realization of computation.

The author also overlooks a general problem with many variants of structuralism, including functionalism: namely, that a simple mapping between structure and reality cannot avoid indeterminacy. The problem was first noticed by M.H.A. Newman (1928), who showed that Russell's causal theory of perception fails precisely because one cannot define the relationship between the structure of the world and the world itself without avoiding triviality or incoherence. Although seemingly out-dated, the problem raised by Newman still appears in many guises, from triviality arguments to indeterminacy of meaning, to Goodman's troubles with similarity (cf., e.g., a discussion of the relevance of Newman's challenge to structural realism in Psillos 1999, pp. 63-69). And Buechner does not offer any answer to the challenge, really.

The basic conceptual connections assumed by Buechner, such as the link between functionalism and cognitive science, have also been put under serious pressure in the last decade. Indeed, there are philosophers of

cognitive science, such as Bill Bechtel, who do not espouse functionalism and yet defend a version of computationalism (e.g., computational mechanism). So, even if Buechner is successful in showing that functionalism is not in such enormous trouble as Putnam contended, this might still be fairly uninteresting news for cognitive scientists, including those who are engaged, as the majority are, in computational modeling. Having said that, there is definitely a lot one can learn from Buechner, precisely because he reads Putnam from the mathematical and purely philosophical point of view.

References

Bechtel, W., & Mundale, J. 1999. Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science* 66(2), 175–207.

Bechtel, W., & McCauley, R. N. 1999. Heuristic identity theory (or back to the future): The mind-body problem against the background of research strategies in cognitive neuroscience. In *Proceedings of the 21st Annual Meeting of the Cognitive Science Society* (pp. 67–72). Mahwah, NJ: Erlbaum.

Block, Ned, & Jerry A. Fodor. 1972. What Psychological States Are Not. *The Philosophical Review* 81 (2) (April 1): 159–181.

Dretske, F. I. 1986. Misrepresentation. In R. Bogdan (Ed.), *Belief: form, content, and function* (pp. 17–37). Oxford: Clarendon Press.

Goodman, N. 1951. *The Structure of Appearance*. Cambridge, MA: Harvard UP.

Lucas, J. R. 1961. Minds, machines and Gödel. *Philosophy* 36, 112-127.

Newman, M.H.A. 1928. Mr. Russell's 'causal theory of perception.' *Mind* 37(146), 137-148.

Polger, Thomas W. 2008. Evaluating the Evidence for Multiple Realization. *Synthese* 167 (3) (August 27): 457–472. doi:10.1007/s11229-008-9386-7.

Psillos, S. 1999. *Scientific Realism: How Science Tracks Truth.* London – New York: Routledge.

Putnam, H. 1995. Review of Shadows of the Mind by Roger Penrose. *Bulletin of the American Mathematical Society* 32, 370–373.

Putnam, H. 1988. *Representation and Reality*. Cambridge, MA – London: MIT Press.

Putnam, H. 1994. Reflexive reflections. In J. Conant (Ed.), *H. Putnam, Words and Life*. Cambridge, MA – London: Harvard UP, 416–427.

Stabler, E. 1987. Kripke on functionalism and automata, *Synthese* 70, 1-22.