Human Arenas
https://doi.org/10.1007/s42087-024-00437-y

RESEARCH

Modelling Thought Versus Modelling the Brain

Orly Shenker[1]

Received: 26 January 2024 / Revised: 5 August 2024 / Accepted: 7 August 2024

Abstract
What is the connection between modelling thought and modelling the brain? In a model (as understood here), we strip away from the modelled system some non-essential features and retain some essential ones. What are the essential features of thought that are to be retained in the model, and conversely, what are its inessential features, that may be stripped away in the model? According to a prevalent view in contemporary science and philosophy, thought is a computation, and therefore its essential features are its computational features. A necessary part of the computational view of thought is the idea that the same computation can be realised by, or implemented in, physically heterogeneous systems, an idea known as "Multiple Realizability" of the computational features or properties by the physical ones. I will describe why the very idea of Multiple Realizability, especially in the case of mental computation, entails mind-body dualism, and explore some implications of this conclusion concerning the question of which are the essential features of thought to be retained in modeling it.

Keywords Physicalism · Non-reductive physicalism · Models · Reduction · Multiple realizability · Computation · Computational theory of mind · Dualism

Paper

What is the connection between modelling thought and modelling the brain? There are various theories in the philosophy of science about what models are (see overview in Frigg and Hartmann 2020). According to the *minimal model* idea (see Batterman and Rice 2014, Weisberg, 2013) we strip away from the modelled system some non-essential features and retain some essential ones. By exploring the essential features in this way, we can learn about them, without the unnecessary interference of the inessential ones, that may introduce noise or other obstacles. For example, in maps we care less about the size and more about the spa-

✉ Orly Shenker
orly.shenker@mail.huji.ac.il

[1] The Hebrew University of Jerusalem, Jerusalem, Israel

Published online: 14 August 2024

tial relations between geographical elements; in billiard balls models of gases, we care more about their relative positions and velocities and less about their quantum mechanical nature; in computerised models of hurricanes, we leave out their destructive features, while keeping in some features that we find helpful in predicting their behaviour in the real world. All these are examples of stripping away the inessential while retaining the essential features of the modelled system in the modelling system.

The question is then what makes something a model of *thought*. To answer this question one needs, of course, to say what thought is. But as we shall immediately see, answering this question may come hand in hand with answering the question of what makes something a model of thought; and, given the state of art in contemporary science and philosophy, it may indeed be easier to start with the latter. To see what a model of thought may be, we need to answer the questions of what are the essential features of thought that are to be retained in the model, and conversely, what are its inessential features, that may be stripped away in the model.

According to a prevalent view in contemporary science and philosophy, thought is a computation, and therefore its essential features are its computational features. Some include representational features and perhaps others, but I shall not go into these debates. On this prevalent view, whatever is involved in computational features is to be retained in the model (see, e.g., Shagrir, 2022). According to this same prevalent view, among the inessential features of thought that need not be retained in the model – and even are better kept out of it – are the features of the hardware that implements the computation making up the thought. In particular, the fact that in humans the hardware is the brain (or some aspect or part or some other feature of this bodily organ) is an inessential feature of thought, and therefore can and even should be kept out of a good model of thought. This is an implication of the assumption that is *a necessary part* of the computational view of thought and is famously called "Multiple Realisation" (or, at least, "Multiple Realizability") of the computational features or properties by the physical ones.[1] In other terms, the same computation can be realised by, or implemented in, physically heterogeneous systems. Of course, the fact that in us the computation of thought is implemented by the brain, makes the brain an especially important system. But if we want to model thought as such, we need to leave out of the model the fact that it happens to be implemented, in us, by the brain.

The question arises then about what kinds of systems can implement a given computation, and the answer to this question involves complications pertaining to the well-known multiple computations theorem and its implications known as the Individuation Problem and the Triviality Problem. Let us leave those to the side at the moment and assume, for the time being, that it is meaningful to ask what systems can implement a computation, in particular the one which constitutes thought, and that we can think of some candidates for this. We do not need details for our present discussion. However, as I shall now show, in

[1] Let me emphasize that in my view the thesis that thought (or cognition, or whichever way one wishes to call or characterize it) is computation, is committed to multiple realizability in the strongest sense: the computational view entails multiple realizability, and is therefore committed to whatever the idea of multiple realizability is committed to, including – as I discuss here and elsewhere – mind-body or other forms of dualism. Some contemporary thinkers challenge this view, saying for example that these fields of research (that in my view are to be seen as research programs) are compatible with physicalism, especially if one assumes that many physical models are couched in abstract terms. I disagree, and the reason is that in physicalism there is no room for abstract entities, and so models (including so-called mathematical or computational ones) are always only physical facts (systems, objects, properties, processes, etc.). An expansion on this point will take us outside the scope of this paper.

discussing modelling of thought, within the framework of the idea that thought is a computation, it turns out that we cannot leave out all talk of the implementing material, since the model itself is always material. Even a mathematical model is always implemented in some matter, and this point – which we justifiably normally ignore when discussing mathematics in general, and mathematical models in particular – cannot be ignored here.

The computational view of the mind, with its underlying Multiple Realizability thesis, guides major research programs in contemporary science: in particular, the grand programs of computational neuroscience and cognitive science. The idea of Multiple Realizability, which underlies the computational research programs in science, is dominant in contemporary literature on the philosophy of mind and philosophy of science, especially as a part of ideas called "non-reductive physicalism", of which "functionalism" and especially "computational functionalism" are important cases. The vast majority of the literature in both science and philosophy supports the computational view and varieties of the Multiple Realizability thesis, and therefore it is unnecessary for me to present further support for it in this paper, and I dedicate the space to presenting the possibility that this popular view is misguieded (see overview and references in Bickle, 2020 and Stoljar, 2024). Indeed, the prevalence and dominace of the computational view of the mind, and its underlying Multiple Realizability thesis, should not mislead us: prevalence is not in and of itself a justification of any scientific view. In this case, it is not a justification for endorsing the idea of Multiple Realizability; on the contrary, the prevalence of this idea calls for a re-examination of its justification. Its lure, since the 1960s, can be understood as it prima-facie offers a way to accept a materialist view which is in line of the natural sciences, and yet account for the intuition that there is something about the mental (and other features of reality) that is more abstract. It is taken by many to have empirical support, but whether the cases studied are indeed ones of multiple realization is debated (see e.g. Polger & Shapiro, 2016, Maimon & Hemmo, 2022). For this reason the debate on whether multiple realization, and more generally multiple realizability, holds in our world, is open. In my view, the very idea of Multiple Realizability entails dualism. Below I will present some reaons for this view. This idea has important implications for understanding how thought can be modelled: such modeling needs to avoid multiple realization, for otherwise it will entail mind-body dualism and, in this sense, will not be scientific.

Let me present very briefly why I think that the very idea of Multiple Realizability, which underlies the computational view of the mind, entails dualism. Having done so, we will be able to continue exploring our central topic of modelling thought. I will present my argument with the help of an abstract example (which is, itself if you like, a model in the above sense). Consider the physical state of some system of interest, according to whatever physics tells us is the case. (Here, "physics" means contemporary physics, but this is not essential to the argument). Normally, in science, and especially in the branches of science involved in studying thought, we are not interested in all the minute details pertaining to the fundamental making of the world; it suffices for us to have a partial description of the physical state, pertaining to some aspect of the physical state. For instance, relevant aspects of a physical system, such as a particular brain, are the electric field in its periphery, or the concentration of some kind of molecules in some part of it. The aspect of interest is a physical property of the system, and we can group states of the same system or of different systems into the same Physical Kind, according to whether or not they share this physical aspect. Let us consider various physical kinds, or physical properties, and the way in which they

belong to or fall under various mental kinds. For example, suppose that the physical kinds P1, P2, and P3 fall under the mental kind M1; the physical kinds P4, P5, and P6 fall under the mental kind M2; and the physical kind P7 falls under the mental kind M3. The fact that different physical kinds belong to the same mental kind is the idea of Multiple Realizability. The cases that fall under the different Pi or the different Mi may be kinds related to the brains of different animals or even inanimate systems that are believed to implement mental kinds and in particular thought.

An extremely important feature of our example illustrates the following well-known idea, essential for the scientific study of the connection between mind and brain: given the physical kind, we can deduce the mental kind, and in this sense, the physical (logically) determines or fixes the mental. In the philosophy of science, we say that this is ensured by the fact that the mental kinds *supervene* on the physical kinds; there cannot be a change of a mental kind that is not accompanied by a change of the physical kind. (A violation of supervenience, in which the mental state changes without a change in the physical state, is considered the hallmark of mind-body dualism). Thus, in our example, there is a combination of supervenience and multiple realisation, and this is the case people have in mind when they think about the computational theory of mind. But what makes it the case that these particular physical kinds fall under these particular mental kinds? Why isn't the partition a different one? For example, why is P4 a member of the M2 mental kind rather than the M1 mental kind? Such a case would still retain the supervenience relation, after all. So supervenience is a feature of the partition, but doesn't fix the partition. What does fix it then?

What determines the association between mental and physical kinds? Here are four families of possible answers to the question of what fixes which mental kinds are instantiated by which physical kind; specifically, why are P1, P2, and P3 members of the M1 mental kind while P4 is a member of the M2 mental kind? I think these answers cover all possible or relevant ones, and certainly everything that can be found in the scientific and philosophical literature on the subject.

Option 1: Reductive type-type identity physicalism[2] Physical kinds that fall under a given mental kind *share some physical feature*. In our example, P1, P2, and P3 are members of the M1 mental kind because they share some physical aspect, and P4 belongs to a different mental kind M2 because it doesn't have the physical aspect that P1, P2, and P3 share. But on this explanation, P1, P2, and P3 form a physical kind: since these three physical kinds share a physical feature, they are sub-kinds of the same physical super-kind, characterised by the physical feature that they share (a paradigmatic example is a more coarse-grained kind). However, in this case, the M1 mental kind is not multiply realised, but is rather realised by a single physical super-kind! In the philosophical literature we say that for Multiple Realizability to hold, the physical kinds falling under the mental kinds must be physically heterogeneous. The idea is that physically heterogeneous physical kinds realise the same mental kind. And so, to explain why are P1, P2, and P3 elements of the M1 mental kind while P4 is a member of the M2 mental kind we assume that P1, P2, and P3 share some physical aspect that P4 doesn't share with them, **we do not have Multiple Realisation**, **and hence we are not in the framework of the computational theory of mind**. Instead, what we have is a case of Materialist or Physicalist Reductive Type Identity theory. But since we want to

[2] An example for such a view is Flat Physicalism, see Hemmo and Shenker (2022a, b, 2023). This is a non-eliminativist version of physicalism.

explore the computational theory of mind, which entails Multiple Realisation, Option 1 is not suitable for our purpose, and in this sense is not available to us. Let us explore alternative options for explaining why P1, P2, and P3 are elements of the M1 mental kind while M4 is a member of the M2 mental kind.

Option 2. Contextual reductive type-type identity physicalism On this option, the fact that some physically heterogeneouos types (or tokens) fall under the same a mental kind is explained by the fact that they take place within a certain *shared context*. For example, they may somehow interact with environmental features that react to them ending up in a similar state, and in the sense they somehow register the system's state or properties. It is convenient to think of these environmental features in terms of an *observer* or at least a *measuring device* (even if the environmental features in question are not an observer nor a measring device in the intuitive sense of the term).[3] Suppose, then, that there is an observer or a measuring device that measures all of the physically heterogeneous facts that fall under the same mental kind. Let's call all those different kinds of contexts "an observer". And so, it may happen that the observer is in the same state when, for example, P1, P2, or P3 obtains, but is in a different state when P4 obtains. One may say that relative to that observer, P1, P2, and P3 "look" the same, but P4 "looks" different.

Given that the context, with the user, is added to the computer, to form an extended system, this is an explanation **within reductive type-type identity physicalism** for why different devices can serve as computers that implement the same computations *for us as users*, as follows. According to the "**multiple computations theorem**" (Putnam, 1967, Shagrir 2012) each device implements numerous computations by its numerous physical properties. Why then do we, as users, see our laptop implementing one computation and not another? (This is known as the "individuation problem".) The answer is that the engineers single out a certain property of this physical system to which we, the *users*, with our specific sense organs and specific brains and specific cultures, are sensitive (via the input and output devices), and then construct the device so that it will have a certain dynamics such that this particular property will evolve in a way that parallels the states of the desired Turing machine. Similarly, the engineers can pick out another aspect of another (physically heterogeneous!) system, to which we, the users, are also sensitive, and build for it suitable dynamics, so that the aspect plus dynamics in the second system will have the same harmony as the aspect plus dynamics in the first system. Both will seem *to us* – with our sense organs and brains and cultures, via the suitable input and output devices – as implementing the same Turing machine. The computation in each system is selected only relative to the user, and different systems are seen to implement the same computation only relative to the user. So, **this is a good explanation of how different devices implement the same computation for us, but – if the context is seen as an extended system - it does so within the framework of reductive type-type identity physicalism, which is reducible to option 1 above.**

But we are not interested (in this discussion) in devices built to serve as computers for us; we are interested in *mental* computations, in the computational theory of mind. And here

[3] I avoid going here into the question of what consists an observer or a measuring device. In particular, since the physical kinds or tokens are heterogeneous, the observer or measuring device does not strictly-speaking measure a shared feature of them; but it nevertheless somehow enters the same state and in that sense registers these physical kinds or tokens as belonging to the same mental kind.

there is no external intended user: and so this line of thinking cannot solve the individuation problem of the *mental* computation implemented in our brains (which are the cognition, according to the computational theory of mind), nor can it explain the (alleged) multiple realisations of mental computations implemented in the brain and (allegedly) in other hardwares, and in this sense, it cannot explain *us* (or our cognition) in terms of computations. The reason is that – as we have just seen – the only way to ascribe a certain computation to a brain, and to see physically heterogeneous systems as implementing the same computations, is to do so relative to a user. But here there is no longer a user: we don't *use* our brains as computers; rather, we *are* those computations. Some other user needs to take the role of preferring a computation, and of recognizing the brain as implementing the same computation as other hardware. Who or what is that user? Who or what is the observer that looks at our brains to pick out that particular computation? Who or what is the observer relative to which our brain and other hardware fall under the same mental or computational kind? To explain mental computation, Option 2 requires that the fact that my mental state at the moment is M1 rather than M2 would be determined by some external observer; and somehow, I have epistemic access to that extra observer's state, when I know (by introspection) my own thought. This is a version of the famous Homunculus fallacy. And as in this fallacy, this idea leads to infinite regress, since the state of that external observer is itself one out of its many physical aspects, and to prefer it over others we need to bring in another observer, a measuring device of a measuring device, and so *ad infinitum*.

So if this line of thinking is to work as an explanation of mental computation (without a user) and not fall into infinite regress, we must assume something non-physical, and it becomes a form of dualism.

Option 3. Functionalism On this option, the fact that explains why P1, P2, and P3 are elements of the M1 mental kind while P4 is a member of the M2 mental kind is that the physical kinds P1, P2, and P3, despite being physically heterogeneous, *share a functional role* in the computational process or the causal network. This is the famous and prevalent idea of *Functionalism*. But this option, despite its popularity, *doesn't solve the problem but repeats it*. Instead of searching for the common physical feature in the states or in the physical kinds themselves, we are now looking for the shared facts that make it the case that different sequences of states (or the different systems that undergo them) implement the same function. The above considerations apply, *mutatis mutandis*.

Notice here, that the very notion of "function" is not a physical one, and therefore the very notion that, on the functionalist view, explains the shared mental kind, is not a feature of the states or kinds or processes or systems in and of themselves; "functions" exist only relative to the interest of an observer (and then infinite regress looms). Importantly: not only relative to the observation capabilities of the observer, but to its *interest*. For example, "survival" of a biological kind is not a feature of any material feature of the world throughout history. "Survival" (importantly, survival *so far*) amounts to the conjunction "this feature obtained, and the kind evolved and survived so far". But biological function is sometimes *defined* in terms of contribution to survival, in a way that renders the above idea *circular*. (Importantly in this context, evolutionary arguments do not support "if-then" forms, since kinds may not survive even if all their organs work perfectly, if a natural disaster occurs). So evolutionary arguments do not and cannot pick out certain facts as functions.

Notice also that evolutionary biology is not an aspect of context or a potential context in the sense of option 2 above, for the following reasons. Evolution may explain (to some extent, given the role arbitrariness may have in it) how we came to be the way that we are; but it does not explain how the way that we are gives rise to the mental or cognitive. Evolutionary theory can support (to some extent) a claim that it would be good to respond to the environment in the way that we do, but this does not select a computation; the environmental approaches to the individuation problem that follows from the multiple computaions theorem do not work. As physics can easily show, the physical making of every organism interacts with a multitude of magnitudes in the physical environment, and arguably many of them can be seen as computations, following the multiple computations theorem. Since only some of these magnitudes are taken to select certain features of the organism as implementing the compution that are associated with cognition, and since there is no physical feature that selects these environemtnal features as prefered, selecting these environment features only in order to prefer "the" computation ascribed to the organism in question is patently circular, and hence non explanatory.

And so if this line of thinking is to work as an explanation of mental computation, it must either reduce to type-type identity physicalism which is option 1 above, or assume something non-physical, and becomes a form of dualism.

Option 4. Outright dualism This option is prevalent as well in the philosophical literature and says that the fact that these physical kinds fall under these mental or computational kinds is explained by some *brute fact*. It is a brute fact, a primitive fact, that P1, P2, and P3 but not P4 fall under M1. Importantly: *even if the mental kinds supervene* on the physical kinds, information about which physical kind obtains is *insufficient* to infer which mental kind obtains in each individual case, since to know that one needs to know how the physical kinds are partitioned to mental kinds (in a way that may satisfy supervenience). The partitioning into mental kinds, i.e. the very fact that a particular physical kind or token falls under a particular mental kind, is – on the "brute facts" view – not determined by the physical kind or token itself, but rather, by the "brute fact". Since this is merely a matter of "brute fact", the same physical kind or token could have fallen under a different mental kind. (If not, we are back to option 1, i.e., reductive physicalism.) And so, in order for me to know whether at this moment I feel pain of some sort at a certain moment, or entertain a thought with some content at a certain moment, I need to have information conerning which mental kind the current physical case falls under, which means that I must necessarily have epistemic access to the brute facts (even, to repeat, if the mental kinds supervene on physical kinds). And since these brute facts are not physical facts, they are non-physical facts, and hence on this view we have epistemic access to non-physical facts, all the time. This is outright dualism.

From this, it follows that the very idea of Multiple Realizability entails dualism, and since the former is a necessary element of all forms of functionalism including the computational theory of mind, which is at the heart of cognitive science and computational neuroscience, all these scientific endeavours are based – unavoidably! – on dualism. There is no way to endorse both the computational theory of mind and a physicalist worldview as those are contradictory ideas. The only way to endorse a physicalist worldview is to endorse physicalism, or, under the name it is known in contemporary philosophical literature, Reductive Type Identity Physicalism, in which all the facts are physical, including properties and

whatever gives rise to them. An example of such a theory is Flat Physicalism (Hemmo and Shenker 2022a, 2023a, 2023b).

In such a theory, accepting the contemporary scientific central idea that the mental, in particular thought, is strongly connected with the neuronal system in the brain (perhaps together with some other bodily organs), the hypothesis would be that thought is (identical with-, nothing but-) a feature of the brain, which is given by a partial description of the brain. (Terminological and metaphysical point: single realisation is not identity. See discussion in Polger & Shapiro, 2016.) Thought is not *realised* by the brain, because there is no multiple realizability, but is (identical with-, nothing but-) a material aspect of the material brain. The brain does not give rise to thought, because brain and thought are one and the same thing. Brain, or an aspect of the brain, and thought are one and the same thing that has two names. This has the following implications concerning modelling thought. In the computation theory of mind, we said that the essential features of thought, to be retained in the model, are computational ones, and the inessential features of the brain, that are better left out of the model of thought, are features of the material brain which happens to implement this computation in humans. This picture changes radically when we endorse the materialist theory of mind, where the brain is the essential feature of thought, to be retained in the model. More generally, the essential features of thought to be retained in a model of it are certain (material) features of the (material) brain, and the inessential features of thought, that can be (or are better) left out of the model, are other (material) features of the (material) brain. We need to *copy* that essential (material) feature of the (material) brain into the model.

The prevalent view in contemporary literature does not request an exact copy, and sometimes requires, at most, that the model represents the brain. Since this is a central conclusion of this paper let me clarify this point. (I will not discuss here the notion of representation which is known to be problematic and hard to explain in naturalistic terms.) Here I employ – as stated at the opening of this paper – *the minimal model view*, according to which a model is a distilled essential feature of the modelled system. The essential feature of a "thinking" physical system, in a context in which its "thinking" is what we want to model, is precisly the physical aspect of that system which is (identical to) this mental feature. To emphasise: in a reductive type-type identity thoery, the exact same physical aspect of the material system, and only that exact same physical aspect, is the mental state in question, and none other will do. We have here a convergence from two directions: the *reductive identity theory of the mind* selects a physical aspect of the system as the one which I "thought"; and the *minimal model view* says that his aspect is the one to be distilled in the model.

By the way, the minimal model in question can be a mathematical model, which will capture the physical magnitudes of thought in the same way that mathematical models of stars and galaxies, hurricanes, digestion, and photosynthesis capture the essential features of those. It can also be a computerised model, again in the same sense that we have computerised models of stars and galaxies, hurricanes, digestion, and photosynthesis. Notice that a computerised model is not a mathematical model, but a material model, in which the modelling system is the electric circuit, which is everything that there is in a computer. (In my view this would be the case for every so-called mathematical model, but I will not go into it here). Thus, a computerised model of thought has to copy – *duplicate*, in the material sense of the term – the modelled system, here the relevant (material) feature of the brain.

This raises the question of what are the (material) features of the (material) brain that are (identical with) thought, those that we need to retain in the model if it is to be a model of

thought. In terms of modelling, it leads us to the following question. If we aim at modelling thought, we retain (copy, duplicate!) in the model the essential features of the brain that are essential for thought to occur. By definition though, if we retain (copy, duplicate!) all of the features that are essential for thought, then, instead of merely *modelling* thought, we create thought itself. In the case of maps that was clear: we retained the relations between geographical elements and left out the size and the third dimension; in the case of hurricanes that was clear as well: we left out the destructive features. But what about the brain? Which features of the brain are we to duplicate and which to leave out, if we want to have a model of thought, but not thought itself? What (material) features of the (material) brain are in fact essential and interesting, and perhaps necessary for thought but not sufficient for it to take place? How to model thought without (re)creating thought itself? These are still open questions in contemporary science and we know so little about them that modelling thought is as of now impossible.

To sum up, let me repeat what we can and cannot do when modelling thought. If we want to avoid mind-body dualism, what we cannot do when studying thought with computers is: (1) try to model thought by retaining computational features and leaving out the material details of the brain; (2) run some computation on a computer and take it to be "a cognitive computation" that is also implemented in the human brain. The latter would be an expression of the computational theory of mind and assume Multiple Realizability, hence assuming that there are non-material facts in the world, some kinds of magical Homunculi, and other ideas we mistakenly thought we already got rid of. I emphasise that, even though this practice may be prevalent in contemporary science, there is no way around this dualist result, no way to continue considering computations to be the essential features of the mind and at the same time remain physicalists. We should be strongly aware of this in our work in cognitive science and computational neuroscience. I cannot overstate the importance of this point and shall illustrate it. Some contemporary scientists, endorsing the (dualist) computational theory of mind, came up with an idea that is a natural conclusion of this (dualist) theory: we need, they suggest, seek better hardware and upload our mind to it, thus obtaining eternal mental life or duplicate ourselves in as many copies as we wish. This is not modelling thought but creating thought itself. It is a natural and almost trivial implication of the computational theory of mind.

When it comes to studying thought with computerised models, the only legitimate and useful physicalist way is to use them for modelling the physical, physiological, chemical, or thermodynamic processes that take place in the brain, in the same way, that we do for stars and galaxies, photosynthesis, and digestion. Searching for those material features of the brain is the task of future brain and cognitive science. Of course, in a materialist picture, we can also (re)create thought itself, and perhaps obtain eternal mental life, by replicating all the (material) features of the (material) brain that together suffice for thought. But can we distinguish between essential and inessential features of the (material) brain, so that we can model thought without (re)creating thought itself? This is an open question.

Acknowledgements This paper is based on research carried out jointly with Meir Hemmo, University of Haifa, Israel. This research was supported by ISF grant number 690/21. I am grateful to the useful comments of two anonymous referees.

Author Contributions This paper is based on research carried out jointly with Meir Hemmo, University of Haifa, Israel.

Funding Israel Science Foundation, grant number 690/21.
Open access funding provided by Hebrew University of Jerusalem.

Data Availability Not applicable.

Declarations

Competing Interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable.

References

Batterman, R. W., & Collin, C. R. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349–376. https://doi.org/10.1086/676677

Bickle, J. (2020). Multiple Realizability, The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (Ed.), https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability/

Frigg, & Roman and Stephan Hartmann. (2020). Models in Science, The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (Ed.), https://plato.stanford.edu/archives/spr2020/entries/models-science/

Hemmo, M., & Shenker, O. (2022a). Flat Physicalism. Theoria 2022, 1–22. https://doi.org/10.1111/theo.12396

Hemmo, M., & Shenker, O. (2022b). Why Functionalism Is a Form of 'Token-Dualism' in: Hemmo, Meir, Ioannidis Stavros, Shenker, Orly, Vishne Gal (2021). Levels of Reality in Science and Philosophy: Re-examining the multi-level structure of reality. Springer. http://philsciarchive.pitt.edu/18073/

Hemmo, M., & Shenker, O. (2023). Is the mind in the brain in contemporary computational neuroscience? Studies in history and. *Philosophy of Science*, *100*, 64–80. https://doi.org/10.1016/j.shpsa.2023.05.007

Maimon, A., & Hemmo, M. (2022). Does Neuroplasticity Support the Hypothesis of Multiple Realizability? *Philosophy of Science* (2022), 89, 107–127 https://doi.org/10.1017/psa.2021.16

Polger, T., & Shapiro, L. (2016). *The multiple-realization book*. Oxford University Press.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan, & D. D. Merrill (Eds.), *Art, mind and religion* (pp. 37–440). University of Pittsburgh.

Shagrir, O. (2012). Can a Brain Posess two minds? Journal of Cognitive Science. 13, 145–165.

Shagrir, O. (2022). *The nature of physical computation*. Oxford University Press.

Stoljar, D. (2024). Physicalism, The Stanford Encyclopedia of Philosophy (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (Eds.), https://plato.stanford.edu/archives/spr2024/entries/physicalism/

Weisberg, M. (2013). *Simulation and Similarity: Using models to understand the World*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199933662.001.0001