Synthese
DOI 10.1007/s11229-015-0731-3

Explanatory completeness and idealization in large brain simulations: a mechanistic perspective

Marcin Miłkowski[1]

Received: 12 January 2014 / Accepted: 16 March 2015

Abstract The claim defended in the paper is that the mechanistic account of explanation can easily embrace idealization in big-scale brain simulations, and that only causally relevant detail should be present in explanatory models. The claim is illustrated with two methodologically different models: (1) Blue Brain, used for particular simulations of the cortical column in hybrid models, and (2) Eliasmith's SPAUN model that is both biologically realistic and able to explain eight different tasks. By drawing on the mechanistic theory of computational explanation, I argue that large-scale simulations require that the explanandum phenomenon is identified; otherwise, the explanatory value of such explanations is difficult to establish, and testing the model empirically by comparing its behavior with the explanandum remains practically impossible. The completeness of the explanation, and hence of the explanatory value of the explanatory model, is to be assessed vis-à-vis the explanandum phenomenon, which is not to be conflated with raw observational data and may be idealized. I argue that idealizations, which include building models of a single phenomenon displayed by multi-functional mechanisms, lumping together multiple factors in a single causal variable, simplifying the causal structure of the mechanisms, and multi-model integration, are indispensable for complex systems such as brains; otherwise, the model may be as complex as the explanandum phenomenon, which would make it prone to so-called Bonini paradox. I conclude by enumerating dimensions of empirical validation of explanatory models according to new mechanism, which are given in a form of a "checklist" for a modeler.

✉ Marcin Miłkowski
marcin.milkowski@gmail.com

[1] Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland

Published online: 01 April 2015

Keywords Human brain project · Blue brain · SPAUN · Mechanistic explanation · Idealization · Bonini's paradox

1 Introduction

Computer simulation is an essential tool in neuroscience and serves various purposes. In this paper, I focus on the explanatory uses of computer simulations in neuroscience and argue that they are explanatory insofar as they are models of brain mechanisms. To do so, I assume a neo-mechanistic[1] approach to explanation (Bechtel 1994; Craver 2007; Kaplan 2011; Machamer et al. 2000; Miłkowski 2013) and briefly account for computer simulations of the brain in mechanistic terms. The account suggests that to serve their explanatory purposes, brain models in general, and computer simulations in particular, may and indeed should be idealized. Complex mechanisms are best elucidated by idealized explanatory models.

Several large-scale brain simulations exist (De Garis et al. 2010) but only some of them aim at biological realism. They also vary with respect to the number of spatiotemporal scales included in the model. For example, the Blue Brain project offers an unprecedented level of detail, describing a part of the somatosensory cortex in 14-day-old rat, and Markram claims that the Blue Brain simulations are meant to "aid our understanding of brain function and dysfunction" (Markram 2006, p. 153). Just like the Blue Brain, most other extant large-scale brain simulations do not aim at modeling intelligent behaviors, which occur at temporal scales of minutes to hours—in part because we do not yet know the intermediate-scale structure of the brain, so we are unable to encode it into simulations (De Garis et al. 2010). Some, however, are trying to fill the gap. For example, Semantic pointer architecture unified network (Spaun), built in the lab of Chris Eliasmith (Eliasmith et al. 2012), is a recent 2.5-million spiking neuron simulation of the brain, and is the largest simulation of this kind. The purpose of the model is to find out how functional capacities arise in biological brains, but the level of biological detail is much lower than in Markram's Blue Brain. The Blue Brain excludes the psychological evidence; Spaun abstracts away from molecular detail.

Correct mechanistic explanations need to satisfy several norms, completeness being one of the most important: the causal model of the mechanism that displays an *explanandum* phenomenon needs to be complete in order to qualify the explanation as complete. Complete explanatory texts "represent all and only the relevant portions of the causal structure of the world" (Craver 2007, p. 27). It might therefore seem that mechanists need to defend Markram's quest for accuracy, and that they would view Eliasmith's models as essentially incomplete; arguably, some might even criticize both models for excluding some temporal scales. Indeed, one may read mechanists saying

[1] Throughout the paper, I speak of "mechanists" (and "the mechanistic framework") to refer to theorists who are committed to (some versions of) the framework defended in the seminal paper of Machamer et al. (2000) and its later improvements. The *neo*-mechanistic framework does not assume, as the old-style mechanical philosophy of the 17th century did, a limited number of admissible interactions produced by the shape, motion, and contact between the parts of the mechanism.

that the principle in computational neuroscience is "the more detail, the better" (Kaplan 2011, p. 347). But the mechanistic norm of completeness should not be confused with the attempt to include all possible detail. Could the exclusions in both simulations be therefore justified?

Mechanists stress that there is a need to precisely specify the *explanandum* phenomenon, which decides what is relevant to the explanation, so not just any detail counts, and Kaplan naturally does not claim that causally irrelevant detail is explanatory. I do not agree with Kaplan and other mechanists, however, about the role of idealization in neuroscience. While they allow idealization for practical reasons and because of technological limitations, I think idealization is required in principle in explanations of sufficiently complex mechanisms. Contrary to appearances, idealization need not imply violation of the mechanistic norms of explanation, in particular the completeness norm. Below, I defend the claim that mechanistic explanation via idealized models is justified by relevance considerations, not only by technological limitations such as tractability. Namely, the aim of most idealizations is to make the essence of the phenomenon the focus of the model, and abstract away from irrelevant detail. Idealizations involve building models of a single phenomenon displayed by multi-functional mechanisms (mechanistic explanatory norms do not require a single model to explain everything), lumping together multiple factors in a single causal variable, simplifying the causal structure of the mechanisms, and multi-model integration. Phenomena explained by mechanistic models are often also idealized and as long mechanistic models represent all and only causal factors relevant to their phenomena, they do not violate the completeness norm. Were the completeness norm incompatible with idealization, the mechanistic framework would be descriptively inaccurate, since idealized explanatory texts are prevalent in the fields of study that commonly use mechanistic explanations.

The structure of the paper is as follows. In the first section, I sketch the mechanistic account of the simulation-based explanation. Then I describe the Blue Brain and Spaun in greater detail, and apply the mechanistic framework in order to analyze both, and show that both are idealized in various ways. I stress that mechanistic models may require further mathematical processing, and that hybrid mathematical–mechanistic explanations can be accounted for in the mechanistic framework for explanation. I conclude by enumerating dimensions of empirical validation of explanatory models as a "checklist" for a modeler.

2 Mechanistic account of simulation-based explanation

According to the mechanistic account of explanation, to explain a phenomenon φ is to elucidate the causal structure of the mechanism that gives rise to φ. While mechanisms are defined variously, the core idea is that they are organized systems, comprising causally relevant component parts and operations (or activities) thereof (for a recent review, see, e.g., Illari and Williamson 2011). Component parts of the mechanism interact, and their organized operation contributes to the capacity of the mechanism to exhibit φ; in that, mechanistic explanation can be understood as closely related to (but

not reducible to) functional analysis (Cummins 1975, 2000).[2] Mechanists recognize the importance of mechanistic explanations in sciences other than fundamental physics (in particular, life sciences, neuroscience, and cognitive sciences). Many hope that an adequate description of the principles implied in explanations, and generally accepted as sound, will help clarify the distinction between good explanations and bad (Craver 2007). In other words, the aim of the mechanistic theory of explanation is to be both descriptively and normatively adequate.

One of the critical requirements of the mechanistic explanation is that the *explanandum* phenomenon be specified. All mechanisms posited in explanations have an explanatory purpose, and for this reason their specification is related to an epistemic interest. For the same reason, the spatiotemporal boundaries of the mechanism, though not entirely arbitrary, can be carved in different ways depending on what one wishes to explain (Craver 2009; Pöyhönen 2013). There are no mechanisms per se; there are only mechanisms *of* phenomena: they display phenomena that can be explained causally, by taking into account their organization. This principle is usually dubbed "Glennan's law"; Glennan (1996) strongly argued that mechanisms are individuated functionally by their capacities, or by the ability to display phenomena. There need not be one-to-one correspondence between phenomena and mechanisms. A single mechanism may display multiple phenomena, as is usual in biological systems.

The *explanandum* phenomenon has to be specified for a given mechanistic model[3]; otherwise the model's use and value will be unclear. The specification of the phenomenon is not to be confused with raw, unrefined observation, or with common-sense intuition about the capacity under consideration. The specification of the capacity may be (and usually is) improved during the modeling process. An early hypothesis about the capacity may well be misleading to the modeler, and only further research might show that a phenomenon is not what it was initially supposed to be (Dennett 1998, p. 314). For example, numerous explanations were offered for the results in the famous Wason task in reasoning (Wason 1966). However, the phenomenon may just as well be an artifact of averaging the results over subjects, and there is no single phenomenon to be explained but a number of different phenomena resulting from the ambiguous wording of the task (Stenning and Lambalgen 2001, 2008).

Hence, explananda are not raw observational data but models of data (Suppes 1962). [Suppes' point has been rediscovered independently in the distinction between data and phenomena introduced by Bogen and Woodward (1988).] For example, models of language production usually presuppose that a user's productivity is the *explanandum* phenomenon, even though it is impossible to empirically observe a language user producing an infinite set of sentences. Because of excellent theoretical reasons for believing that language users have this capacity, productivity will be described in a

[2] There is an ongoing controversy whether mechanistic explanations are law-based or not (Andersen 2011); the argument in this paper is logically independent from the answer to the latter question.

[3] For Craver, the *explanans* is the mechanism, as he defends so-called ontic account of explanation; however, with Wright (2012), I adopt the representational account of explanation here: it is the models of mechanisms that do the explanatory work. I will talk of models of mechanisms throughout this paper. Simulations, analyzed later, are just one kind of models; I leave other kinds aside in this paper (for example, I do not analyze animal models) but I believe that the points about idealization apply also for such models.

model of data. This point is important in the context of computational neuroscience; some computational descriptions of capacities of neural subsystems will fall out naturally in this account as specifications of phenomena. For example, what Chirimuuta (2014) calls "canonical neural computations" is a specification of the explanandum phenomenon rather than an explanation per se. The very act of classifying of a phenomenon this way or another may have an immense theoretical value, as Chomsky's stress on productivity, and in this sense, contributes to a better understanding of the phenomenon. However, it seems more natural to see classifications as descriptions, since merely saying that language is productive does not answer any *why* or *how* questions about productivity.[4] By the same token, a mere description of an effect in psychology is not by itself explanatory (Cummins 2000).

Several general norms of mechanistic explanation are related to how one specifies the capacity of the mechanism; in short, the phenomenon should exist (so the specification needs to be true), and it should be characterized correctly and fully (Craver 2007, pp. 123–128). For example, if Stenning and Lambalgen are correct, many explanations of the Wason task are spurious just because the specification of "the" phenomenon to be explained is false; instead, many different phenomena, though all real, were lumped together and there is no common underlying mechanism.

Craver (2007) distinguishes two kinds of incomplete mechanistic models: mechanism sketches and mechanism schemata. Sketches usually contain gaps and placeholder terms (such as "processing"). Schemata contain placeholders (or black boxes) to be filled when evidence is available. Only mechanistic explanations that do not contain any gaps or placeholders are complete. By "completeness" here, notably, Craver means completeness *relative* to the explanandum phenomenon: only components and activities *causally relevant* to the phenomenon should be included, not just any spatiotemporal parts. The completeness of the mechanistic model is to be understood as specifying the whole causal model; to specify the causal model, one needs to know all and only the relevant variables and their connections in the graph that describes it. What's important, the mechanistic account needs to appeal to a theory of causal relevance that does not make causal relevance equivalent to explanatory relevance (on pain of inducing a vicious circle)[5]; one such theory is the interventionist theory of causation that offers axioms and formal semantics for statements about causal relevance (Galles and Pearl 1997). So for example, a mechanistic explanation of a mouse trap does not need to specify the maker of the metal parts in the trap (unless it is somehow relevant to the functioning of the trap). Changing the maker of such parts would not affect the capacity of the trap to catch mice.

Adding more detail is not guaranteed to make the explanation better; only causally relevant detail matters. Here, mechanistic explanation is similar to Cummins's (1975)

[4] One may still insist that *what*-questions can prompt explanatory answers; however, *what*-questions that are not reducible to *how*- and *why*-questions yield different kind of answers. For example, Burge (2010) sees one task of philosophy in deepening knowledge and understanding of constitutive conditions of something's being what it is; such conditions ground explanations of something's nature. In this sense, models of data would be explanatory as well, but that seems to be stretching the term "explanation" beyond its useful scope in philosophy of science, as it would largely overlap with "description".

[5] This worry has been voiced by one of the anonymous referees of this paper.

functional analysis in stressing that explanations of organized systems should include only relevant components. At the same time, for mechanists, the notion of organization in Cummins's account is underspecified: for Cummins, simple box-and-arrow diagrams are enough to spell it out. The mechanism requires that organization is anchored in spatiotemporal entities and processes (Craver 2007, p. 138). [For a more detailed comparison of mechanistic and functional explanation, see Miłkowski (2013), Chapter 3 and Piccinini and Craver (2011).]

The practice of modeling of complex systems shows also that modelers strategically specify their explananda to exclude causal interactions of certain subsystems and external factors. The resulting omissions are not considered to be explanatory parts of models (if a model does not reflect some factors, it need not mean that the model is inaccurate); and it's best to consider them as idealizations of explananda. This point is compatible with the claim that mechanistic models aim at explaining phenomena, or models of empirical data. Phenomena occurring in complex systems are usually idealized, and idealizations are introduced not just for tractability but for explanatory purposes. Simply, complexity of causal interactions in a complex physical mechanism cannot be fully reflected in a model on pain of making the model totally explanatorily obscure. Yet a model of an idealized phenomenon need not violate the mechanistic norm of completeness, as long as it does include all and only causal interactions relevant to the phenomenon. I will return to this point, after introducing the methodology of both simulations studied in Sect. 3.

Computer simulations can be understood as mechanistic models specifying the functioning of a given mechanism (Miłkowski 2013). A special case of computer simulation is a simulation of computational processes, for example a simulation that explains how a given physical computer works. Obviously, one can also model a mouse trap on a computer without assuming that the trap is a computer at all. For my purposes here, however, it is irrelevant whether computational neuroscience treats brain computation realistically or not, as it makes no difference for questions of how to assess accuracy and relevance of detail in the model.

Even if the simulation can be run on a computer, this need not mean that the explanatory model of phenomenon is mechanistically complete. Completeness is to be assessed vis-à-vis the explanandum phenomenon, and computer simulations usually contain ad hoc additions needed to run them: these are the decisions made by the modeler without any empirical evidence (Frijda 1967). It is notoriously hard to disentangle such ad hoc additions from the rest of the model. If we were to assess completeness simply by checking whether, say, the software can be run without problems, then including a number of ad hoc parts in the model would be enough to make it complete. Yet this would not be explanatory completeness.

There are two ways in which mechanistic computer simulations may correspond to their targets. First, they may be *weakly equivalent* to the target, in that they only describe the initial and termination conditions (in the case of cyclical mechanisms, initial and termination conditions may pertain to a single cycle or to a series of cycles). Second, they may be *strongly equivalent*, when they also correspond to the process that generates the termination condition. These notions have been used in methodology of computer simulation since 1960s (Fodor 1968, Chapter 4). Similar notions have been introduced by Bernard Zeigler (1976) in his theory of modeling and simulation:

a model is said to be *replicatively* valid if it can generate output data from known input data; it is *predictively* valid when its output corresponds to the new data, and *structurally* valid when the structure of the model corresponds to the operations of the real system being modeled. Zeigler's predictive validity is equivalent to Fodor's weak equivalence, and his structural validity to Fodor's strong equivalence [Weisberg (2013) makes the same distinction but uses somewhat confusing terminology, by talking about *dynamical* and *representational* fidelity.]. Only strongly equivalent models are explanatory, according to the mechanistic account.

One particularly important kind of mechanistic explanation, constitutive explanation, requires multi-level investigation of the organization. Mechanistic levels are not levels of abstraction; they are levels of *composition* or organization (Craver 2007).[6] Such levels are constituted by whole–part relationships. This means that a lower level in a mechanism is a proper part of its higher level. Constitutive explanation includes *at least* three such levels: the *bottom* (−1) level, which is the lowest level in the given analysis and describes the internals of mechanism parts and their interactions; an *isolated* (0) level, at which the parts of the mechanism are specified along with their interactions (activities or operations); and the *contextual* (+1) level, at which the function of the mechanism is seen in a broader context. Depending on the shared scientific practice, the bottom and the uppermost level in the explanation will vary (Machamer et al. 2000). One can easily introduce a further level if needed. Let's take an explanation that accounts for reproduction of bacteria. The reproduction of bacteria in a given environment (contextual level) is explained in terms of division (isolated level that ignores the environment), and division in terms of cellular mechanisms (the bottom level). The cellular-level mechanism can be further explained by its molecular parts, which would introduce a fourth level in this explanation. The possibility of adding further levels is essential in computational neuroscience, which needs to draw evidence from biophysical observation and from behavioral studies. In addition, there is no consensus regarding the bottom level in computational neuroscience, which is particularly salient in the cases considered in Sect. 3 of this paper. The mechanistic framework can help decide some of such questions by showing that some levels are relevant to explanations, while others are not. If they are relevant, the completeness norm requires them.

Mechanistic models need to specify causally relevant variables at all and only levels of organization considered relevant to the explanation. This applies also to mechanistically understood computer simulations, which need to conform to general modeling principles at the same time. Hence, the empirical adequacy of the simulation can be tested by checking whether it is strongly equivalent to the *explanandum* phenomenon. In neuroscience, usual structural validation methods apply, including chronometry (Posner 2005), various kinds of experimental and natural interventions (Craver 2007), brain slicing microscopy (Seung 2012), optogenetics (Deisseroth et al. 2006), brain imaging—though with usual caveats (Trout 2008)—and task decomposition (Newell and Simon 1972). All in all, the more independent observables are tested, the more robust the model. Mere phenomenological validation modeled after

[6] I set Craver's account of levels of realization aside here, as it is not relevant to the issue of completeness and idealization.

the Turing test (Turing 1950) cannot establish the model's empirical adequacy. On the contrary, what is important is simply evidence about component parts, operations, and overall organization at all levels of the mechanism. Chronometry relates to psychological time-reaction studies, which are usually located at higher levels with respect to neuronal mechanism; optogenetics usually to bottom levels of the mechanism, and various interventions are used at different levels, if possible. Therefore, one of the norms of the constitutive mechanistic explanation is to build a tightly coordinated model at all levels of organization (which is implied by the completeness norm), and to make it possible to integrate it with higher and lower levels, if possible. For this reason, the neo-mechanistic framework is particularly sensitive to interfield research (Darden and Maull 1977).

Another ideal of mechanistic explanation is producing how-actually explanations. A *how-actually* causal explanation elucidates how the actual explananda are caused in contrast to a *how-possibly* explanation that elucidates possible ways of causing the explananda; a *how-plausibly* explanation is "more or less consistent with the known constraints on the components, their activities, and their organization" (Craver 2007, pp. 112–113). How-possibly explanations risk positing entities that are not actual causal factors, hence they may violate the completeness norm. The mechanistic framework ranks how-possibly explanations with their possibly causally irrelevant variables lower than how-plausibly explanations, as the latter ones include only entities and activities that seem plausible in light of our current theories, so they are less likely to be causally irrelevant. Because of mere complexity of neural systems and ethical, technological, and theoretical difficulties in brain research we simply lack necessary evidence to fully validate mechanistic models. Thus, we can only hope for how-plausible explanations in contemporary computational neuroscience.

To return to the constitutive account of mechanistic explanation that I introduced on the previous page, I will illustrate it with a cash register in a supermarket. The explanandum phenomenon is the capacity to add prices of individual items and determine the overall sum to be paid by a customer. At the contextual level, one describes the cash register as playing a certain role in the supermarket, by allowing easy calculation of the sum to be paid. This includes a bar-code scanner, a conveyor belt, etc. At the isolated level, a dedicated computer using special software is described. The constraints such as commutativity or associativity of addition are included in the description of the software. Yet without describing the machine that can run the software, this level of description is incomplete. Some failures of the cash register can be explained not only in terms of the software bugs but also as hardware failures. Also, the particular display configuration, which can be related to user preferences at the contextual level, is usually not described fully in the software specification. It is the isolated level where one describes the physical machine that can display the product name for the cashier clerk and, more fundamentally, can run code by reading it from external memory. The formal description, usually in terms of the programming language or diagrams, is put into correspondence with the machine.[7] At the bottom level, the operations

[7] In this paper, I abstract away from a complex issue of the structure of computational mechanistic models. They usually contain both a formal computational model and a mechanistic model, which are put into correspondence. For more detail, see Miłkowski (2011, 2014).

of the electronic parts of the machine are explained by reference to their properties, relationships, and organization. Just because vast differences between different types of registers are possible, exact explanations will differ. Also, self-checkout machines will have the capacity to collect cash automatically, which needs to be explained as well (the explanandum will be different), and so forth.

The completeness norm itself requires the multi-leveled structure of explanation, as the capacity of the cash register can be fully explained only in the context of the supermarket. A cash register in an art gallery may as well function as a work of art, and its capacity to add prices would not make any difference if the cash register is just on display there. In addition, the activities of a user of the cash register are causally relevant for its performing the function of addition. This is why the contextual level is relevant for this explanation. Moreover, one needs to introduce the bottom level to understand how the function of addition is realized; depending on how exactly one conceives the capacity of the cash register, the level of detail on the bottom detail will vary. In other words, the bottom level is related to our epistemic interest and specification of the explanandum phenomenon. For example, a service engineer could want to explain and predict the breakdown patterns in the cash register. This is impossible without knowing how exactly the cash register performs the calculation, and that requires the detail on the isolated and the bottom level where wear and tear can occur. But a philosopher may be interested only in a capacity framed solely in mathematical terms, ignoring the need to fix broken cash registers (although the explanans needs to include spatiotemporal entities and activities).

3 Blue Brain meets Spaun

The Blue Brain is one of the most detailed simulations of the brain available, and certainly one of the most widely-known. It promises a "quantum leap" in computational neuroscience by building accurate models of the mammalian brain from the first principles. The bottom level of the model is cellular rather than the genetic or molecular (or, indeed, quantum!). The phenomenon modeled is a "2-week-old rat somatosensory neocortex corresponding to the dimensions of a neocortical column (NCC) as defined by the dendritic arborizations of the layer 5 pyramidal neurons" (Markram 2006, p. 155). There is a vast amount of quantitative data for the model from around 15,000 experiments (including recordings of multi-neuron patch-clamps and microscopy on brain slices). These allow systematic quantification of molecular, morphological, and electrical properties of the neurons and synaptic pathways.

The modelers justify their choice of the young sensory column by claiming that it is evolutionarily one of the simplest available for experimental research. This means that it may serve as a simplified model of more complex mammalian brains. The significance of the NCC is, according to Markram, immense: the NCC is a 10,000 neuron microcircuit that is repeated numerous times in the cortex of mammalian brains. According to him, the only effective difference between the human cortex and the rat's is the number of NCCs involved. If we had a working simulation of the NCC, we might build bigger cortical simulations. At the same time, it is unclear whether the NCC has any well-defined biological function (Horton and Adams 2005); for this

reason, NCC might not be a mechanism at all, as that would mean that there is no phenomenon that it stably displays. Similarly, an arbitrarily chosen spatiotemporal part of the digestive tract is not a mechanism as such, even if it can be anatomically delineated, unless there is a well-defined capacity it is responsible for.

Around 10,000 neurons are modeled using the computational template called the Blue Column. This includes different types of neurons in layer 1, multiple subtypes of pyramidal neurons in layers 2–5, spiny stellate neurons in layer 4, and more than 30 anatomical–electrical types of interneuron with variations in each of layers 2–6. To run the simulation, care is taken to include statistically-plausible variations in the model. The plausible variations include also the pattern of the neural connectivity (the "connectome"). The Blue Brain follows the so-called Peters' Rule, which states that connectivity is random (Seung 2012), as a connectome has not yet been discovered for the rat. This, however, means that the organization of the mechanism is also random, and even if it is plausible that the creation and elimination of neural connections is to some extent stochastic, violations of Peters' Rule are known. For this reason, the accuracy of the simulation is limited to component parts and operations of the mechanism; it is stipulated that its orchestrated operation should be based on random connectivity. Seung stresses that the model might therefore follow the known principle of computer science: Garbage In, Garbage Out. On the other hand, the random pattern of connectivity used in the Blue Brain is estimated to be around 74 % accurate (Hill et al. 2012). For this reason, Seung's claim may be overstated. But the randomly established connectivity is definitely a placeholder that can be replaced by the proper connectome, which makes the Blue Brain an incomplete model. In other words, it is a schema in the sense defined earlier in the paper, and is poised to provide only how-plausibly explanations.

What is crucial in understanding the purpose of any computer simulation is the way it is tested, or validated, to use the technical term adopted by the modeling community. The standard way of proving that the model is valid at this level is to quantify its divergence from the phenomenon with respect either to the input and the output of the model for weakly equivalent models, or to the underlying process and the input/output for strongly equivalent ones.

The Blue Brain has not been tested to show how much it diverges from available evidence from 15,000 experiments, so its validity is unclear; or at least the results of such tests are not included in any publicly available publications. Papers describing the results of the project talk, for example, about building models automatically from the data and optimizing them using multiple objectives (Druckmann et al. 2007). Another result of the project is a novel method of reliably linking the microscopic membrane ion channels to the macroscopic electrical behavior of neurons (Druckmann et al. 2011). This method can be used to test different computational models of neurons but was not applied to the whole model of the rat's NCC. Of course, the method is relatively new, while the project started a decade ago.

One of the goals in the Blue Brain project seems to be integrating multiple sources of data in a single model; the data is usually partial and sometimes inconsistent, so building large simulations is simply a way of creating a more reliable database about the phenomenon. Creating a new process that simulates and calibrates as well as systematically analyzes the biological accuracy and consistency of each modification

of their NCC model was a milestone of the project. Two other milestones achieved in 2007 were (a) developing a technique to automatically build microcircuits from biological data, and (b) developing a cellular-level model of NCC, which was used to "stimulate research into the simulation of neural micro-circuitry".

As such, the Blue Brain does not seem to have a clear explanandum phenomenon (even if the NCC were to have a clear biological function). The model merely describes the rat's NCC rather than explains some specific capacity mediated by the NCC. However, it would be too quick to dismiss the model as non-explanatory, as "the Blue Brain" is actually not a label for a particular model but for a *family* of models that can be run using the data about the NCC. Rather, it should be considered an *environment for explanation* rather than a single *explanatory model.* For this reason, it can be used to produce explanatory models.

For example, a model was created (Reimann et al. 2013)[8] to investigate the origin of the local field potentials (LFPs), which are crucial to investigating the dynamical properties of the information processing in the brain (for a current review, see Buzsáki et al. 2012). The model was used to undermine the traditional assumption that LFPs reflect synaptic and passive return currents, and to this purpose, single synapses and single-neuron physiology were reproduced. In other words, the target phenomenon of the simulation is not LFP but the physiology of a single neuron, which is then used to *infer* system-wide properties such as LFPs. These properties are not directly observable; one cannot directly measure the electrical activity of thousands of neurons and the resulting brain waves. But using the computational simulation, the traditional hypothesis that LFPs reflect synaptic and passive conductance was overthrown: It is the active currents that dominate the generation of LFPs. The model created some testable predictions as well. For example, it predicts that 150 Hz bandwidths are heavily contaminated with spiking, which was already confirmed (Schomburg et al. 2012).

The high level of detail of this model does not mean that the simulation is necessarily mechanistically complete: It does not include glial nor astrocytic processes, and the nonmyelinated presynaptic axonal compartments, which probably only minimally contribute to LFPs, were excluded as well (Reimann et al. 2013, p. 387). LFPs are *inferred* from the output of the simulation, not generated directly in the model, and that inference usually needs several approximations. In other words, they are not immediately modeled, so the NCC is not completely reconstructed in the model, but this was not the modelers' intention.

The two-step modeling procedure is typical in LFP models (Lindén et al. 2013), and it shows a particular hybrid nature of the model in question. First, morphologically reconstructed neurons are simulated using NEURON software (Carnevale 2007) to provide transmembrane currents; then extracellular potentials are calculated based on these (Nunez and Srinivasan 2006). The Blue Brain LFP model provides the first part, and LFP properties have to be computed separately. It is an interesting case of model interoperability; the result from a computer simulation can be used for further mathematical and discursive operations, which will create a *hybrid* model, containing a mechanistic model of the brain and the mathematical (completely formalized or

[8] I owe the reference to this work to the anonymous referee of the previous version of the paper.

not) model of the LFPs, based on relatively well-understood biophysics of the process responsible for creating LFPs. There already exist software environments such as LFPy (Lindén et al. 2013) that link mechanistic neural simulations with external computations, or, simply, mechanisms with equations.

Although the modelers include low-level biophysical detail in the Blue Brain framework, there is not enough focus on the orchestrated operation of the NCC as a multi-level mechanism given the intended goals of the project. The overarching goal of Blue Brain, according to the project website (The Blue Brain Project EFPL 2011), is to "reverse engineer the mammalian brain", and that includes building models at "*different* scales" and discovery of basic principles governing the structure and function of the brain. Even more, it is to model the "*complete* human brain" (emphasis added). Although one can run simulations useful for research on brain waves, it is unclear whether the Blue Brain could include any high-level constraints, whether connectivity patterns, lateralization, or neuropsychological results. Sporns notes:

> Rather than designing an architecture that incorporates patterns from all scales that are experimentally accessible, including the large scale of neural populations and brain regions, bottom-up approaches [such as the Blue Brain] attempt to construct the brain by brute force, neuron by neuron and synapse and synapse. What is lacking are the important constraints provided by empirical and theoretical research on principles of brain organization and architecture (Sporns 2012, p. 168).

The results of the Blue Brain modeling can be used to compute properties related to LFPs but higher-level constraints and interventions are not included either in the model directly or in the principles of modeling. From the mechanistic point of view, the contextual level considerations, in particular precipitating and inhibiting conditions, are simply missing. In other words, there is a large gap between the official goals of the project and the actual methodology of simulations; it remains unclear how the architecture, structure and function of the *whole* mammalian brain could be uncovered by running *only* models of the NCC.

In Spaun, neural populations are the bottom level of the model, unlike in the Blue Brain. Spaun's principles used to build the model are based on the neural engineering framework (Eliasmith and Anderson 2003). The neural network is composed of biologically-plausible spiking neurons (though much simpler and less heterogeneous than in the Blue Brain). The main difference between the model of the NCC and Spaun is that the latter is intended to model high-level behavioral capacities and is able to perform eight diverse tasks (without modifying the model); the main purpose of the model, however, is not to offer explanations for all these tasks but "to propose a unified set of neural mechanisms able to perform them all" (Eliasmith et al. 2012, p. 1204). In other words, Spaun is both a large-scale simulation and a cognitive architecture (for an extended treatment of the architecture and approach to the modeling, see Eliasmith 2013). Spaun has five subcomponents, which deal with (1) information encoding; (2) transformation calculation; (3) reward evaluation; (4) information decoding; and (5) motor processing. Visual information is fed to the system, and it controls a physically modeled arm.

The central notion of the framework underlying Spaun is that of a semantic pointer; it is used to create higher-level representations that efficiently encode (or compress) lower-level neural information. Thanks to this, the model can be both biologically realistic and able to perform cognitive tasks. Spaun integrates several approaches to cognitive systems. First, the modelers use control theory to model dynamics of control (such as involved in controlling the arm); second, it is computational as far as it transforms information (though without implementing any Mentalese); third, it is inspired by insights from systems neuroscience into the function of brain parts (the building blocks of Spaun, omitted here for brevity, correspond to anatomically defined brain areas).

The tasks performed by Spaun are:

1. *Copy drawing* Given a randomly chosen handwritten digit, Spaun produces the same digit written in the same style as the handwriting.
2. *Image recognition* Given a randomly chosen handwritten digit, Spaun produces the same digit written in its default writing.
3. *Reinforcement learning* Spaun performs a three-armed bandit task, in which it must determine which of three possible choices generates the greatest stochastically generated reward. Reward contingencies can change from trial to trial.
4. *Serial working memory* Given a list of any length, Spaun reproduces it.
5. *Counting* Given a starting value and a count value, Spaun writes the final value (i.e., the sum).
6. *Question answering* Given a list of numbers, Spaun answers either one of two possible questions: (i) what is in a given position in the list? or (ii) given a kind of number, at what position is this number in the list?
7. *Rapid variable creation* Given example syntactic input/output patterns (e.g., 0074→74; 0024→24; etc.), Spaun completes a novel pattern given only the input (e.g., 0014 → ?).
8. *Fluid reasoning* Spaun performs a syntactic or semantic reasoning task that is isomorphic to the induction problems from the Raven's Progressive Matrices test for fluid intelligence (Raven 1993).

The list of tasks might be in itself impressive (the model's predictive validity regarding biological behavior is around 90 % in all of them), but the model's main feature is to flexibly switch between these tasks. The main phenomenon of the mechanism is rapid behavioral flexibility. In addition, it is possible to extend Spaun by adding more tasks, because the switching mechanism—which embodies a hypothesis about the function of the basal ganglia—is scalable.

The difference between the Spaun and the Blue Brain is not only the amount of biological detail; Spaun includes detailed hypotheses about the function of neural populations localized in several brain regions (rather than about the function of micro-circuitry, whose detailed operation is considered explanatorily irrelevant). For this reason, it qualifies as a mechanistic model (and not a framework; the framework here is the Neural Engineering Framework), with an explicit *explanandum* phenomenon. Clearly, as long as the model is empirically structurally validated, it fulfills mechanistic criteria for explanations. In some cases (question answering), the model has some predictions that have yet to be tested. Obviously, there is no available neuroscientific

data that would make it possible to directly test whether the overall architecture of Spaun is correct or not.

That Spaun has less detail about individual neurons than Blue Brain does not make it non-mechanistic. Mechanistic models may be quite abstract, and the level of detail is to be decided by the modeler (Levy and Bechtel 2013). Including all kinds of information in the model may be even detrimental to its purpose and is not required by the completeness norm. Here, a law of diminishing returns applies: more is not always better. Adding more detail may lead to so-called Bonini's Paradox, i.e., to the result that the model is as difficult to understand as the phenomenon under modeling, and for complex artificial networks simulating the brain, the paradox looms large (Dawson 1998, p. 17). In addition, if there are parts of the mechanism that do not contribute to the *explanandum* phenomenon, these do *not* qualify as component parts of the mechanism. They are not relevant to explanation, and can be considered useless noise. A similar point is true of introducing further (upper or lower) levels in the constitutive explanation: if they don't make predictions or explanations more precise and accurate, they should be left out. For example, the hat left in the car, even if spatiotemporally included in it, is not a component part of the car, as it does not contribute to its transportation capacity in any way. The art of modeling is therefore one not only of inclusion but also of exclusion.

Spaun may be considered an instance of Galilean idealization, which is the practice of introducing distortions into theories with the goal of simplifying theories or focusing only on essential features of the phenomenon (Nowak 2000; Weisberg 2007).[9] According to most theories of idealization, the distortions have to be removed in order to apply the theory or the model to actual phenomena. But such idealizations—contra Weisberg (2013)—do not require that all possible detail be given, as that would lead to Bonini's paradox. Weisberg claims that the ultimate goal of the Galilean idealization is complete representation (Weisberg 2013, p. 111). In his opinion, this means that each property of the target "must be included in the model," and "anything external to the phenomenon that gives rise to its properties must also be included". He continues: "Finally, structural and causal relationships within the target phenomenon must be reflected in the structure of the model (…) the best model is one that represents every aspect of the target system and its exogenous causes with an arbitrarily high degree of precision and accuracy" (Weisberg 2013, p. 106).

Yet for mechanism, removing distortions and approximating truth is not equivalent to Weisberg's ideal of complete representation, which is not, and indeed should *never* be, fulfilled for any model on pain of Bonini's paradox. The mechanistic completeness norm requires only relevant detail. Nowak, one of the primary defenders of Galilean idealization, claims that "it consists in focusing on what is essential in a phenomenon and in separating the essence from the appearance of the phenomenon" (Nowak 2000, p. 110). With complex systems such as brains, idealization may be the key way to avoid the Bonini's paradox, and there is no reason why the mechanists should exorcise idealization from computational neuroscience. On the contrary, idealizing is not dictated merely by technological limitations but required for explanations of well-defined phe-

[9] Weisberg, *contra* Nowak, does not see idealization as the focus on the essential features of the phenomenon.

nomena in biological systems. These phenomena co-occur with multiple phenomena at the same time (since biological systems are usually highly multi-functional), so there is a strong need to abstract away and sometimes strategically distort the organization of the mechanism in its model. For that reason, Spaun only includes the architecture relevant to the tasks executed by the model. Even the Blue Brain model of LFP, with its stress on low-level detail, does not include the processes considered to be only minimally relevant to the phenomenon of brain waves.

Mechanistic models, as with all models, may lump together many components and operations for simplicity and tractability. Such models are called *integrated* by Zeigler (1976). For example, two causal variables in a causal model may be replaced by a single variable, and if there is no change in overall behavior at the input/output of the model, this may be justified. But there is a price to pay: the integrated model will have a lower ranking in terms of structural validity. There may, however, be additional ways to make inferences about the underlying causal structure. For example, the two lumped variables may be related in some regular fashion; in this way, the loss of validity is relatively minor. All in all, all higher-level models qualify as integrated mechanistic models in this sense; and so does Spaun.

However, if the focus of the explanation is the whole human brain with all its currently known capacities related to the eight tasks in question, Spaun fails to answer several important questions about its functioning. For example, it has a very minimal long-term memory mechanism (i.e., only in the striatum) and only a minimal procedural memory. In this respect, the model is incomplete for behaviors involving long-term memory consolidation. Similarly, it has no flexible attention and the eye position remains the same, which is definitely not the case for human vision (the eye position is then an instance of Galilean idealization). For this reason, Spaun would be incomplete as a general model of the human brain with all capacities related to the eight cognitive tasks performed by Spaun—in such a case, it is a mechanism sketch at best, as it heavily idealizes away from capacities that the brain has. If we take Spaun to be a model of flexible task-switching in the brain, then its fidelity is higher, as the evaluation is always relative to the model's *intended* scope. Just because the aim of the model is given in a slightly ambiguous manner—for validation purposes, the data about eight different tasks is used, but in general descriptions it is stressed that flexible task switching is the *explanandum* phenomenon—there are two ways of assessing the completeness of the model. Also, for the overall task (fulfillment of the eight individual tasks), no formal description is given. Hence, flexible task-switching remains underspecified as a capacity. From the mechanistic perspective, this means that we cannot precisely state the explanatory value relative to the phenomenon of task switching.

Because of the lack of experimental data, building a how-actually model of task switching is now impossible—there is simply not enough data on the explanandum phenomenon. What Spaun might realistically achieve, given the current performance limitations of standard computers (it is not executed on a supercomputer but on large computer clusters), is only a highly idealized how-plausibly explanation. However, there is already evidence about behavior that is not reproduced by Spaun; for example, it is difficult for humans to switch from tasks that were strongly activated (Mayr and Keele 2000). Of course, it is possible that Spaun was not supposed to answer

questions about inhibitions related to moving between tasks, but given the general underspecification, a critic of the model is justified in expecting that general properties of human performance would be represented. To wit, idealizations in the specification of the explanandum phenomenon need to be made explicit if one wants to evaluate the completeness of the model.

One interesting feature of both models is that they idealize phenomena, and hence restrict the amount of relevant detail to be included in their model. However, they are also incremental, as they remain partially incomplete because of the lack of some relevant empirical evidence; for example, the Blue Brain lacks the empirical evidence about the connections, and Spaun has only minimal longer-term memory mechanisms. This means that some detail is simply lacking. Yet other detail is also strategically excluded, not just for technological and practical reasons; in the Blue Brain, one can hypothesize that it is the level of the cognitive architecture; in Spaun, it is the molecular level. These levels can be causally relevant to the explananda, even if minimally so.

While the Blue Brain is supposed to become even more low-level as the molecular level of the NCC is added (De Garis et al. 2010, p. 7), Spaun purportedly abstracts from such detail and achieves correspondence with behavior. At the same time, the Blue Brain tries to account for all available low-level evidence to link biophysics with neuronal computation. One way of defending the latter approach would be to say that the result of the project is a better understanding of extant experimental evidence that has to be made consistent: the database of results will be reusable for other projects. The modeling approach of the Blue Brain is definitely structural, typical for life sciences, where detailed research into structure may precede function ascription (Seung 2012). Knowing the structure of micro-circuitry will be useful in specifying the exact *explanandum* phenomenon in all explanations of the brain cortex, and the Blue Brain environment is poised to offer exact specifications of the NCC structure. The structure of the model is designed to make it easy to automatically integrate multiple sources of low-level information, though Blue Brain ignores higher levels of organization of the brain. Nevertheless, evidence from higher levels of organization is idealized away in the NCC model, which means that the model cannot be used to generate new hypotheses about the higher-level capacities of the NCC, as they have to be inferred by modelers from the output of the model by using external resources. They are not reconstructed in the Blue Brain.

The Blue Brain, and similar simulations, such as those proposed under the forthcoming Human Brain Project (Kandel et al. 2013), may help us better understand the underlying organization of the NCC. At the same time, the Blue Brain does not integrate higher-level constraints, and it is unclear whether the NCC is a mechanism just because its biological function is not well-defined. For this reason, the Blue Brain data need to be supplemented with additional assumptions, like in the LFP model, and it can offer low-level constraints on the adequacy of the specification of the explanandum phenomenon. The Blue Brain is a simulation *environment* and for this reason, it is not explanatory in itself, in contrast to models such as Spaun, even if a precise assessment of the explanatory power of the latter is difficult.

It is only to be hoped that the Human Brain Project (HBP), a large EU program (total funding is to exceed €1190 million), which is a follow-up to the Blue Brain project, will eventually include multiple scales in their framework, as higher-level

data integration is also crucial in brain modeling at higher spatiotemporal scales. One subproject is cognitive architectures, but Markram has already commented that Spaun is not a model of the brain (Eliasmith and Trujillo 2014), so he seems to underestimate the value of building cognitive architectures anyway; recently, this part of the project has become the focus of a major controversy (The Lancet Neurology 2014). The stress on biophysical detail in HBP is related to the intended use of the simulation in medical research, which may indeed require accurate modeling of diverse conditions related to neurodegenerative diseases. These simulations can be embraced by mechanists as long as the detail given in the simulation turns out to be causally relevant to the phenomenon displayed (or inhibited by a medical condition).

4 Conclusion: evaluating and integrating large-scale simulations

Large-scale simulations of the brain may serve various uses. Simulation allows testing of interventions that would be otherwise unethical; interventions in simulations may be much cleaner (even ideal), while experimental techniques in vivo are less precise; it can be used to integrate data from many sources and to find inconsistencies or gaps; it is often used to prove the feasibility or scalability of a certain framework; running a computer explanation may show that a computer model explains capacities of the brain, and so on.

An assumption that all simulations are supposed to be explanatory has led to spurious controversies: Henry Markram of the Blue Brain project accused Dharmendra Modha of scientific fraud (Adee 2010), claiming that Modha never simulated a cat brain. Indeed, he did not; all Modha created was a cat-*scale* simulation, which was an achievement from a computer science point of view, but not an achievement for neuroscience. Modha's aim is to produce hardware and software for running massive parallel computations on neurally-inspired computers, and biological fidelity has almost no bearing on this project (his team has already managed to build a simulation with 500 billion neurons, which is larger than a human brain; see Wong et al. 2012). Consequently, Modha's non-explanatory models need not be models of real brain mechanisms.

Some brain simulations are paradigmatic examples of the use of Big Data in science, where results of thousands of experiments are integrated and analyzed to find regularities and patterns. The Blue Brain project and the Human Brain Project belong in this category, offering a way of creating data-driven, detailed simulations as based on the Blue Brain environment. But sometimes it is individual simulations that explain and predict highly complex phenomena. Such is the case with Spaun, and slightly less so with the model of the LFP produced using the Blue Brain, as the model itself does not represent LFPs directly, only data that can be used to compute them.

There are several dimensions of empirical validity of explanatory computer simulations used in neuroscience. I already mentioned that according to mechanism, they need to be complete by covering the relevant causal factors. The completeness norm requires both structural validity and how-actually explanations, or at least how-plausibly explanations. It also requires that all levels of the mechanism are included, which usually demands also a behavioral match if the capacities of the complete brain

are to be modeled, so evidence from psychological experiments needs to be included in the validation. Quite obviously, simulations should also have biological relevance, and their relevance is greater if they are general.

In general, simulations are usually useful when they actually run. Running a complex simulation may lead to results that are too difficult for the modelers to predict without the use of a computer. Hence, implemented computational models are thought to be of higher value, as within them unexpected properties of models are easier to find (Farrell and Lewandowsky 2010). The evaluation of the goodness of fit for such models is also straightforward, by contrast to evaluation of their verbal descriptions, which always remain underspecified. If the modeler makes the model code generally available, others can replicate the results and assess the fit to phenomena under simulation. In addition, one may 'probe' the model by tweaking its properties and checking the resultant behavior. This way, for example, one might tweak the connectivity pattern in neural networks to see how it affects their behavior.

Neither is higher fidelity of the model necessarily a virtue from the mechanistic point of view, contrary to appearances. What is essential is whether the *explanandum* phenomenon is elucidated by the overall functioning of the model and that all components and operations *relevant* to the phenomenon are included in the model. But the phenomenon may be idealized to avoid Bonini's paradox and to include only its essential features. Without a precise specification of the phenomenon, however, it is not possible to determine the degree of structural validity of the model. This pertains to Spaun, with its slightly generic specification of the *explanandum* phenomenon (task switching). It is difficult to say what exactly should be the result of the working simulation. Should we see task inhibition between activated tasks, for example?

The NCC model, as many other models in computational neuroscience, draws data from databases of experiments in vivo. As I stressed, the simulation used to hint at the origin of the LFPs in the brain did not reconstruct LFPs in themselves; it was only used to infer their properties. This means that multiple models need to be integrated to build complete models of the brain, including also models of high-level behavior. The mechanistic framework requires that the modeler looks down at the low-level causes of phenomena, but also around and up, to use Bechtel's (2009) phrase. High-level constraints on organization are crucial in explaining neural mechanisms, as is the influence of other mechanisms at the same level of organization. Looking down, around and up can be achieved either by building interoperable models that share assumptions and can be simply chained together, or by performing a special kind of idealization, called *multiple-models idealization* (MMI). It consists in "the practice of building multiple related but incompatible models, each of which makes distinct claims about the nature and causal structure giving rise to a phenomenon" (Levins 1966; Weisberg 2007, p. 645). In MMI, it is hoped that the truth is in the intersection of lies in multiple models. Less metaphorically, it means that independently built simulations, all empirically validated, are compared to see their common predictions and structural patterns, which are hypothesized to reflect robust components and operations in the models (for more on the notion of robustness, see Wimsatt 2007).

However, creating a correct MMI meta-model is non-trivial, as common patterns and predictions may stem from the same set of ad hoc assumptions made by modelers, for example; in general, the more independent assumptions of models in MMI, the

better, as they will be more statistically independent as a whole, so it will be less likely that they produce the same results by coincidence. Automatically building different models may also help compensate the confirmation bias, or the tendency to seek evidence that confirms (rather than disconfirms) the modeler's hypothesis (Farrell and Lewandowsky 2010). At the same time, technological limitations make MMI expensive: computational neuroscience requires time-consuming computations, and creating multiple simulations with the same level of detail would simply need more time (though the time complexity will grow by a constant factor k, where $k$is the number of models plus overhead of comparing the results; constant factors are negligibly low from the point of view of computational complexity theory).

These requirements can be summarized in the form of a simple checklist for a modeler.

1. *Is it clear what the explanandum phenomenon is*? The phenomenon to be explained should be clearly identified.
2. *Is the explanandum phenomenon analyzed and well-understood*? The dynamical structure of a phenomenon (e.g., precipitating and inhibiting conditions on multiple time-scales) is needed for explanations to be complete.
3. *Is explanation general and does it predict previously unobserved behavior*? If not, the simulation might just be an instance of overfitting.
4. *Is the model implemented*? Verbal descriptions or even formal specifications of models are not as methodologically valuable as complete implementations.
5. *Does the model fulfill the criteria of structural validity and is it empirically adequate vis-à-vis actual phenomena*? *Are all components and operations completely specified on multiple levels*? Idealized explanations (mechanism sketches) are also valuable, but it should be possible, in principle, to fill in the gaps in the sketch.
6. *Is the model interoperable*? In neuroscience, it is necessary to draw evidence from different sources. Interoperable models that can be "plugged in" to other models as their inputs are therefore more valuable. Alternatively, multiple different models with different assumptions can be created to see which components and operations are robust.

My goal in this paper was to make clear that the completeness norm does not require including all possible detail in simulations understood as mechanistic models, as well as to show how simulations function as idealized models of mechanisms. Idealization is not detrimental to mechanistic modeling; idealized models perform fine if they focus on the essence of the phenomenon to be explained. Similarly, an idealized model may explain a complex mechanism with multiple capacities accurately as long as it represents the essential factors accurately. More is not always better; and sometimes more is even worse. For this reason, idealization in models of complex mechanisms is unavoidable not only for technological reasons, but primarily for explanatory reasons, and it is explanatory relevance to the phenomenon at hand that justifies the use of idealization. Moreover, models are representations, and as soon as they become too complex for their users (be that human beings or other models that interface them), they are no longer performing their explanatory function.

As the recent controversy over the Human Brain Project shows (The Lancet Neurology 2014), the neuroscience community is not unanimous about the bottom-up

approach to modeling and the exclusion of the higher-level spatiotemporal scales. Given the ambitious goals of the project, it simply seems premature to promise huge progress in modeling the complete mammalian brain or discovery of new therapies for neurodegenerative diseases without a systematic methodology that would link multiple levels of organization in a single framework. And this is exactly what the Human Brain Project is currently lacking.

Acknowledgments Work on this paper was financed by the Polish National Science Centre OPUS Grant, under the decision DEC-2011/03/B/HS1/04563. Previous versions of this paper were presented before audiences at the Jagiellonian University in Cracow, Polish–Japanese Academy of Information Technology in Warsaw, and Maria Curie Skłodowska University in Lublin during the conference "Algorytm a heurystyka w badaniach nad umysłem". The author wishes to thank Costas Anastassiou, Chris Eliasmith, Przemysław Nowakowski, and several anonymous referees of this journal for their very helpful comments and suggestions.

References

Adee, S. (2010). Cat-brain fever. *IEEE Spectrum*, *47*(1), 16–17. doi:10.1109/MSPEC.2010.5372479.

Andersen, H. (2011). The case for regularity in mechanistic causal explanation. *Synthese*, *189*(3), 415–432. doi:10.1007/s11229-011-9965-x.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines*, *4*(1), 1–25. doi:10.1007/BF00974201.

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*(5), 543–564. doi:10.1080/09515080903238948.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, *97*(3), 303–352.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature Reviews. Neuroscience*, *13*(6), 407–420. doi:10.1038/nrn3241.

Carnevale, T. (2007). Neuron simulation environment. *Scholarpedia*, *2*(6), 1378. doi:10.4249/scholarpedia.1378.

Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153. doi:10.1007/s11229-013-0369-y.

Craver, C. F. (2007). *Explaining the brain. Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575–594. doi:10.1080/09515080903238930.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy*, *72*(20), 741–765.

Cummins, R. (2000). "How does it work" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–145). Cambridge, MA: MIT Press.

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science*, *44*(1), 43–64.

Dawson, M. (1998). *Understanding cognitive science*. Malden, MA: Blackwell.

De Garis, H., Shuo, C., Goertzel, B., & Ruiting, L. (2010). A world survey of artificial brain projects, part I: Large-scale brain simulations. *Neurocomputing*, *74*(1–3), 3–29. doi:10.1016/j.neucom.2010.08.004.

Deisseroth, K., Feng, G., Majewska, A. K., Miesenböck, G., Ting, A., & Schnitzer, M. J. (2006). Next-generation optical technologies for illuminating genetically targeted brain circuits. *The Journal of Neuroscience (the official journal of the Society for Neuroscience)*, *26*(41), 10380–10386. doi:10.1523/JNEUROSCI.3863-06.2006.

Dennett, D. C. (1998). *Brainchildren. Essays on designing minds*. Cambridge, MA: MIT Press.

Druckmann, S., Banitt, Y., Gidon, A., Schürmann, F., Markram, H., & Segev, I. (2007). A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data. *Frontiers in Neuroscience*, *1*(1), 7–18. doi:10.3389/neuro.01.1.1.001.2007.

Druckmann, S., Berger, T. K., Schürmann, F., Hill, S., Markram, H., & Segev, I. (2011). Effective stimuli for constructing reliable neuron models. *PLoS Computational Biology*, *7*(8), e1002133. doi:10.1371/journal.pcbi.1002133.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering. Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. *Science (New York)*, *338*(6111), 1202–1205. doi:10.1126/science.1225266.

Eliasmith, C. (2013). *How to build the brain: A neural architecture for biological cognition*. New York: Oxford University Press.

Eliasmith, C., & Trujillo, O. (2014). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology*, *25*, 1–6. doi:10.1016/j.conb.2013.09.009.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*(5), 329–335. doi:10.1177/0963721410386677.

Fodor, J. A. (1968). *Psychological explanation: An introduction to the philosophy of psychology*. New York: Random House.

Frijda, N. H. (1967). Problems of computer simulation. *Behavioral Science*, *12*(1), 59–67. doi:10.1002/bs.3830120109.

Galles, D., & Pearl, J. (1997). Axioms of causal relevance. *Artificial Intelligence*, *97*(1–2), 9–43. doi:10.1016/S0004-3702(97)00047-7.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.

Hill, S. L., Wang, Y., Riachi, I., Schürmann, F., & Markram, H. (2012). Statistical connectivity provides a sufficient foundation for specific functional connectivity in neocortical neural microcircuits. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(42), E2885–E2894. doi:10.1073/pnas.1202128109.

Horton, J. C., & Adams, D. L. (2005). The cortical column: A structure without a function. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *360*(1456), 837–862. doi:10.1098/rstb.2005.1623.

Illari, P. M., & Williamson, J. (2011). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119–135. doi:10.1007/s13194-011-0038-2.

Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R., & Koch, C. (2013). Neuroscience thinks big (and collaboratively). *Nature Reviews. Neuroscience*, *14*(9), 659–664. doi:10.1038/nrn3578.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*(3), 339–373. doi:10.1007/s11229-011-9970-0.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*(4), 421–431.

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science*, *80*(2), 241–261. doi:10.1086/670300.

Lindén, H., Hagen, E., Łęski, S., Norheim, E. S., Pettersen, K. H., & Einevoll, G. T. (2013). LFPy: A tool for biophysical simulation of extracellular potentials generated by detailed model neurons. *Frontiers in Neuroinformatics*, *7*, 41. doi:10.3389/fninf.2013.00041.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Markram, H. (2006). The blue brain project. *Nature Reviews. Neuroscience*, *7*(2), 153–160. doi:10.1038/nrn1848.

Mayr, U., & Keele, S. W. (2000). Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General*, *129*(1), 4–26. doi:10.1037/0096-3445.129.1.4.

Miłkowski, M. (2011). Beyond formal structure: A mechanistic perspective on computation and implementation. *Journal of Cognitive Science*, *12*(4), 359–379.

Miłkowski, M. (2013). *Explaining the computational mind*. Cambridge, MA: MIT Press.

Miłkowski, M. (2014). Computational mechanisms and models of computation. *Philosophia Scientiæ*, *18*(3), 215–228.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nowak, L. (2000). The idealizational approach to science: A new survey. In L. Nowak & I. Nowakowa (Eds.), *Idealization X: Richness of idealization* (pp. 109–184). Amsterdam: Rodopi.

Nunez, P., & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of EEG*. Oxford: Oxford University Press.

Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. doi:10.1007/s11229-011-9898-4.

Posner, M. I. (2005). Timing the brain: Mental chronometry as a tool in neuroscience. *PLoS Biology*, *3*(2), e51. doi:10.1371/journal.pbio.0030051.

Pöyhönen, S. (2013). Carving the mind by its joints: Culture-bound psychiatric disorders as natural kinds. In M. Miłkowski & K. Talmont-Kaminski (Eds.), *Regarding the mind, naturally: Naturalist approaches to the sciences of the mental* (pp. 30–48). Newcastle upon Tyne: Cambridge Scholars Publishing.

Raven, J. (1993). *Manual for Raven's progressive matrices and vocabulary scales*. Oxford: Oxford Psychologists.

Reimann, M. W., Anastassiou, C. A., Perin, R., Hill, S. L., Markram, H., & Koch, C. (2013). A biophysically detailed model of neocortical local field potentials predicts the critical role of active membrane currents. *Neuron*, *79*(2), 375–390. doi:10.1016/j.neuron.2013.05.023.

Schomburg, E. W., Anastassiou, C. A., Buzsáki, G., & Koch, C. (2012). The spiking component of oscillatory extracellular potentials in the rat hippocampus. *The Journal of Neuroscience (the official journal of the Society for Neuroscience)*, *32*(34), 11798–11811. doi:10.1523/JNEUROSCI.0656-12.2012.

Seung, S. (2012). *Connectome: How the brain's wiring makes us who we are*. Boston: Houghton Mifflin Harcourt.

Sporns, O. (2012). From simple graphs to the connectome: Networks in neuroimaging. *NeuroImage*, *62*(2), 881–886. doi:10.1016/j.neuroimage.2011.08.085.

Stenning, K., & Lambalgen, M. Van. (2001). Semantics as a foundation for psychology: A case study of Wason's selection task. *Journal of Logic, Language and Information*, *10*(3), 273–317. doi:10.1023/A:1011211207884.

Stenning, K., & Lambalgen, M. Van. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 international congress* (pp. 252–261). Stanford: Stanford University Press.

The Blue Brain Project EFPL. (2011). Goals. Retrieved August 27, 2014 from http://jahia-prod.epfl.ch/site/bluebrain/op/edit/page-58109.html.

The Lancet Neurology. (2014). The Human Brain Project: Mutiny on the flagship. *The Lancet. Neurology*, *13*(9), 855. doi:10.1016/S1474-4422(14)70181-4.

Trout, J. D. (2008). Seduction without cause: Uncovering explanatory neurophilia. *Trends in Cognitive Sciences*, *12*(8), 281–282. doi:10.1016/j.tics.2008.05.004.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. doi:10.1093/mind/LIX.236.433.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmondsworth: Penguin.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy*, *104*(12), 639–659.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. New York: Oxford University Press.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Wong, T. M., Preissl, R., Datta, P., Flickner, M., Singh, R., Esser, S. K., et al. (2012). IBM Research Technical Report No. RJ10502 (ALM1211-004)) (Vol. 10502). IBM Research, Almaden, San Jose.

Wright, C.D. (2012). Mechanistic explanation without the ontic conception. *European Journal for Philosophy of Science*. doi:10.1007/s13194-012-0048-8.

Zeigler, B. P. (1976). *Theory of modelling and simulation*. New York: Wiley.