Minimal models of consciousness: Understanding consciousness in human and non-human systems

Wanja Wiese (wanja.wiese@rub.de)

Abstract

Should models of consciousness be detailed *mechanistic* models of particular types of systems, or should they be *minimal* models that abstract away from the underlying mechanistic details and provide generalisations?

Detailed mechanistic models may afford a complete and precise account of consciousness in human beings and other, physiologically similar mammals. But they do not provide a good model of consciousness in other animals, such as non-vertebrates, let alone artificial systems. Minimal models can be applicable to a wide range of different conscious systems. But do they provide genuine explanations that are autonomous from explanations by detailed mechanistic models?

This paper provides a taxonomy of minimal models and measures of consciousness, clarifies their relation to detailed mechanistic models, and highlights benefits of different minimalist approaches to consciousness.

1 Introduction

Two goals, simple but overwhelmingly strong, govern the development and assessment of many models of consciousness: generality and completeness.[1] On the one hand, an ideal model of consciousness applies not just to some types of conscious states in human beings, but to all forms of consciousness, in all conscious species, and, potentially, even artificial systems. On the other hand, a model should be mechanistic and detailed enough to enable precise, empirically testable predictions about particular conscious systems.

---

[1] What exactly does completeness entail? One might believe that a model of some phenomenon is complete only if it represents as many details as accurately as possible. Upon closer inspection, this characterization is too crude: more details are not always better. A useful explanatory model should therefore capture only explanatorily relevant factors. In particular, a complete mechanistic causal model should capture all relevant causal factors that are relevant for the target phenomenon (Craver & Kaplan, 2020). See also section 2.1 below.

Attempting to satisfy both goals at once, however, is almost certainly doomed to fail.[2] Should models that aim for generality, and those that prioritise completeness, be regarded as ships that set off in opposite directions, are blown by distinct great winds, and eventually land in different havens?[3]

In this paper, I provide a meta-theoretical contribution to this question, by focusing on minimal models and measures of consciousness (Metzinger, 2020; Wiese, 2020). I argue that a genuine minimal-model explanation of consciousness (that is distinct from mechanistic explanation) is conceivable, but that currently, explanations of consciousness that draw on minimal models will to some extent be dependent on detailed mechanistic models: the latter are at least needed to *justify* explanations provided by minimal models of consciousness (Rusanen & Lappi, 2016), even if minimal models pick out the *explanatorily relevant* properties (Klein et al., 2020; Miracchi, 2017) that are required for a general understanding of consciousness. Nevertheless, some claims entailed by minimal models of consciousness can be autonomous from implementational details if they provide (non-causal) mathematical explanations (Mancosu, 2018) or *potential* explanations that afford (non-explanatory) objectual understanding (Lipton, 2019).

The paper is structured as follows. In section 2, I clarify the scope of models and measures of consciousness and briefly review different notions of scientific understanding. In section 3, I provide a taxonomy of minimalist approaches to consciousness, review their respective benefits, and clarify their relationship to mechanistic models. Furthermore, I discuss whether minimal models of consciousness should be expected to yield *minimal-model explanations* in the sense of Batterman & Rice (2014).

In section 4, I discuss the recently-proposed conscious Turing machine (Blum & Blum, 2022; Blum & Blum, 2021) as a minimal model of consciousness. In the concluding section, I highlight some open questions and point to promising avenues for future research on minimal models of consciousness.

---

[2] Historically, much of science and philosophy displays what Glennan (2017), following Wittgenstein, calls a *craving for generality*. According to Glennan, a characteristic feature of mechanistic approaches is that they abandon the quest for generalisations and laws, because "[t]he generalizations we sometimes call laws are heuristic; they do not reflect the deep reality of things" (Glennan, 2017, p. 3).

[3] My apologies to Bertrand Russell for adapting the first lines of his autobiography.

2 Models and measures of consciousness: General considerations

2.1 The scope of models and measures of consciousness

The scope of models and measures of consciousness can differ in terms of the types of *global states of consciousness* (Bayne et al., 2016; Mckilliam, 2020) that they target, or in terms of the types of *conscious systems* to which they are applicable, or in terms of the types of *features of consciousness* that are modelled.

Relevant features of consciousness include phenomenal and functional properties. Many models of consciousness focus on functional properties associated with consciousness, such as global availability (Dehaene et al., 2017) or meta-cognitive features (Fleming, 2020; Lau, 2022). Other models explicitly target states of *phenomenal consciousness* that are characterised by phenomenal properties. Phenomenal properties include properties like the experienced redness of a conscious percept of a tomato or the sense of mental agency that we typically have over conscious thoughts. Of most interest are properties that are instantiated during most (if not all) conscious episodes, such as phenomenal unity (Bayne, 2010; Mason, 2021; Wiese, 2016, 2018a, 2022). Understanding such 'structural properties', as they are sometimes called (Seth, 2009), is particularly relevant, because such properties need to be taken into account by all models of consciousness that strive for completeness. For this reason, such properties are also useful for comparing and assessing theories of consciousness (Del Pin et al., 2021; Seth & Bayne, 2022).

In what follows, it will be useful to make a broad distinction between two goals that may be pursued by different approaches to consciousness. One goal consists in understanding consciousness in systems with a known capacity for consciousness (e.g., human beings). The other goal consists in understanding consciousness in non-human systems, including systems for which it is unknown whether they have a capacity for consciousness in the first place (e.g., robots and other artificial systems). It is useful to distinguish between these two goals because they put an emphasis on different norms for explaining and understanding, viz. on *completeness* and *generality*, respectively.

Models of consciousness in human beings and other creatures with a known capacity for consciousness should ideally be *complete*. This does not mean that they should be as detailed as possible (Craver & Kaplan, 2020). In most cases, however, they will benefit from taking at least some implementational details into account. For instance, a computational model of some features of human consciousness may highlight high-level properties (Blum & Blum, 2022, p. 8), but this will

only yield a complete explanation if it is also shown how the computations specified by the model are implemented in the brain (Piccinini, 2020, p. 159). More generally, models that abstract away from mechanistic details may pick out the *explanatorily relevant* properties (Klein et al., 2020; Miracchi, 2017) that are required to understand consciousness, but more detailed models are at least needed to *justify* explanations provided by more abstract models (Rusanen & Lappi, 2016).

When it comes to understanding consciousness in non-human systems, a focus on *generality* is desirable. That is, models of consciousness should be pitched at a level of abstraction that makes them applicable to non-human systems (including other animals and artificial systems). On the one hand, this means that models should not refer to medium-dependent properties of human consciousness (e.g., neural mechanisms that form part of the neural basis of human consciousness). On the other hand, this means that models should not refer to types of conscious experience that may be distinctive for (some) human beings (e.g., feelings of *Weltschmerz*, or cognitive phenomenology associated with *diagram chasing* in homological algebra). In particular, general models of consciousness must abstract away from human-specific properties of consciousness and from human-specific types of conscious states.

Note that the two norms are not mutually exclusive. However, the two goals mentioned above (understanding human consciousness vs. understanding consciousness in any kind of system) suggest different emphases. Pursuing these goals will also benefit in different ways from different types of minimality (which is the topic of section 3 below).

2.2 Models of consciousness and scientific understanding

Before discussing the potential contribution of *minimal* models of consciousness, it is useful to distinguish different forms of scientific understanding. Of particular relevance are *objectual understanding* (e.g., *understanding consciousness*)[4] and *explanatory understanding* (such as different ways of *interrogative* understanding, e.g., understanding *why* certain neural activity correlates with certain conscious experiences, or understanding *how* anaesthetics lead to a loss of consciousness).

Objectual understanding can have different targets, including theories and models (e.g., understanding a mathematical theory of consciousness, or understanding a computational models of

[4] Objectual understanding is sometimes also called "holistic understanding" (Hannon, 2021, p. 282). Apart from explanatory and objectual understanding (which are forms of *theoretical* understanding), there is also *practical* understanding (Bengson, 2017, p. 15), i.e., understanding *how to do something* (Hannon, 2021, p. 283).

binocular rivalry), or understanding phenomena (e.g., understanding blindsight or understanding conscious attention).

According to some accounts, understanding consists in grasping logical or semantic relations between the contents of beliefs (Kim, 1994), e.g., grasping that integrated information theory does not entail panpsychism. Achieving a deeper understanding in this sense typically results in having a more coherent set of beliefs about a particular domain (Kvanvig, 2018). Understanding may also target (dependency) relations between objects in the world (Kim, 1974; Lipton, 1991/2004), e.g., grasping that certain brain lesions *cause* a loss of consciousness, or grasping that consciously perceiving a face is *correlated* with activity in a particular brain area. These ways of conceiving the objects of understanding are not mutually exclusive: by grasping relations between contents of beliefs, we may also grasp relations between phenomena in the world, which are targeted by beliefs (Strevens, 2011).

In philosophy of science, there is a debate over the relation between explanation and understanding. More specifically, there are at least two issues. One is whether explanatory understanding (e.g., understanding *why*) can be reduced to knowledge of an explanation. According to reductionist accounts, knowing an explanation gives you understanding (Achinstein, 1983; Khalifa, 2017; Lipton, 2019; Salmon, 1984).[5] According to non-reductionistic accounts, understanding requires more than merely knowing an explanation, or is otherwise different from knowledge (Elgin, 2007; Kvanvig, 2003).

The other issue is whether some instances of *objectual* understanding (e.g., understanding *consciousness*) are significantly different from *explanatory* understanding: are some instances of objectual understanding non-explanatory ways of understanding (Lipton, 2019)? Do some cases of objectual understanding facilitate explanatory understanding (Carter & Gordon, 2014)? Or can objectual understanding be reduced to explanatory understanding (Khalifa, 2017)?

A further debate concerns the role of idealisation in scientific understanding: to what extent does understanding by a scientific model require that the model incorporates true assumptions? According to some, idealisation is at most pragmatically useful (i.e., some deviations from the truth are acceptable, but not in themselves epistemically valuable, Kvanvig, 2003, ch. 8), according to others,

[5] Apart from this, there can also be *feelings of understanding*. The phenomenology of understanding must be distinguished from understanding as an epistemic state (Regt, 2004). What is at issue in this debate is whether understanding as an epistemic state can be reduced to knowledge of an explanation.

idealisations have direct epistemic value, in that they enable understanding (Elgin, 2007; Potochnik, 2017).

These debates are relevant to clarifying the role of minimal models in a scientific understanding of consciousness, in so far as types of minimal models may differ with respect to how idealised (distorted) they are, and with respect to whether they make an explanatory contribution. Different forms of scientific understanding are summarised in table 1 below.

Table 1	
Forms of understanding	**Examples in the science of consciousness**
Explanatory understanding	
Understanding *why* (causal or non-causal[6])	Understanding why only activity in certain neural areas (such as the thalamo-cortical system) correlates with consciousness, whereas activity in other areas (such as the cerebellum) does not. Such an understanding may be provided by "explanatory correlates of consciousness" (Seth, 2009).
Understanding *how*	Understanding how activity in pyramidal neurons can enable or disable conscious processing (Aru et al., 2020).
Understanding by *unification*	Understanding that different phenomena can be explained with the same or similar mechanisms (Danks, 2014, p. 171; Miłkowski, 2016, p. 24) .
Objectual understanding	
Understanding *X*	Understanding consciousness.
Understanding *theories*	Understanding how different empirical theories of consciousness are related to one another and to metaphysical theories; understanding theoretical and empirical implications of different theories.
Understanding by *how-possibly* explanation (knowledge of a *possibly true* explanation, Lipton, 2019)	(i) Understanding that a global neuronal workspace could be an implementation of the conscious Turing machine (Blum & Blum, 2021). (ii) Understanding that all conscious systems systems whose consciousness arises from a global workspace are subject to the same complexity constraints, limiting the capacity of the global workspace (Blum & Blum, 2022). (See section 4 for details on the conscious Turing machine.)

[6] It is debated to what extent there are non-causal (e.g., mathematical) explanations within empirical science (see Ross & Woodward, 2023, sec. 5).

3 What are minimal models and measures of consciousness and why are they useful?

The term "minimal model of consciousness" can be interpreted in at least two ways. A minimal model of consciousness can be a model of a minimal *form of consciousness*[7], or it can be a minimal *model* of some type or feature of consciousness. Here, I shall focus on the latter. That is, the minimality in question is not the minimality of the target phenomenon (i.e., it is not about the simplest form of consciousness), but it is about the minimality of the model itself.

A somewhat related minimalist approach is exemplified by the search for markers of consciousness, such as clusters of cognitive capacities that are associated with consciousness (Birch, 2022; Ginsburg & Jablonka, 2019). Such a "theory-light strategy" constitutes a *minimalist* approach because it makes little theoretical assumptions about consciousness (however, it rules out theories like cognitive epiphenomenalism, Birch, 2022, p. 141). A marker can be regarded as a 'minimal measure' of consciousness (see table 2 and the next section for further clarification). For this reason, the discussion in this paper will not be restricted to minimal *models* proper, but will include models *and measures* of consciousness.[8]

Table 2			
Type of minimality:	**Highly coarse-grained**	**Highly idealised (distorted)**	**Highly selective**
Goal: Understanding consciousness in systems with a known capacity for	**Example:** Computational characterisation of human consciousness and its features.	**Example:** Perturbational complexity index as an approximation to a measure of integrated information.	**Example:** Specifying building blocks (like metacognition) for theories of consciousness.

[7] It is common to distinguish different concepts of consciousness (e.g., phenomenal consciousness and access consciousness, see Block, 1995), as well as different types or global states of consciousness, such as alert wakefulness, dreaming, or disorders of consciousness (Bayne et al., 2016; Mckilliam, 2020). One can then ask: is there a "simplest form of conscious experience" (Metzinger, 2020, p. 2)? Here, simplicity can, e.g., be understood in terms of the experienced contents, or in terms of the minimally sufficient conditions for any experience at all to occur. Windt (2015) introduced the term "minimal phenomenal experience" (MPE) for this type of conscious experience. Crucially, MPE must be distinguished from what is called the "minimally conscious state" (Giacino et al., 2002); this term refers to states in which subjects show some behavioural evidence for consciousness (unlike in coma or unresponsive wakefulness syndrome, Laureys et al., 2010). By contrast, MPE is not defined in terms of evidence, but in phenomenological terms (describing phenomenal properties or contents of conscious experiences). That is, MPE involves "phenomenological 'minimality'" (Gamma & Metzinger, 2021, p. 3).

[8] I am grateful to Andrew Y. Lee for suggesting the label "models and measures".

consciousness (Ideally: *complete* understanding)	**Benefit:** Deeper understanding by focus on most relevant properties.	**Benefit:** Making a measure of consciousness computationally tractable.	**Benefit:** developing (new) theories of consciousness.
Goal: Understanding consciousness in non-human systems (Ideally: *general* understanding.)	**Example:** Computational characterisation of general properties of consciousness. **Benefit:** Applicable to different types of system.	**Example:** Candidates for non-necessary (ideally: sufficient) conditions for consciousness, like various learning abilities. **Benefit:** Providing a practically useful, (defeasible) marker of consciousness.	**Example:** A model that only specifies some necessary (but not sufficient) conditions for consciousness. **Benefit:** We can *rule out* that a system is conscious if it fails to fulfill a necessary condition.

3.1 Types of minimality and their benefits

In general, the minimality of a minimal model or measure of consciousness can be defined in terms of:

(i) its level of grain ("minimal" in the sense of "highly coarse-grained" or "capturing only high-level properties"),

(ii) its fidelity ("minimal" in the sense of "highly idealised/distorted"), or in terms of

(iii) the number of factors contributing to the phenomenon that are captured by the model ("minimal" in the sense of "highly selective").[9]

The benefits of these three types of minimality are to some extent dependent on the goals pursued. Recall from section 2 that two relevant goals are the goal of understanding consciousness in human beings (i.e., in a system that is already known to be conscious, or that has a known capacity for consciousness), and non-human systems, including systems for which it is unknown or uncertain whether they are conscious. For instance, we would like to have a better understanding of how different states of consciousness are realised in the human brain. Apart from that, it would also be relevant to know which (if any) artificial systems can be conscious. I shall now review a few examples of the three kinds of minimality and also consider their respective benefits.

[9] For a more general account of these features, see Weisberg (2007), who uses "idealisation" as a label for all three types. The focus in this paper is much narrower; for instance, not every idealised model that asserts some falsehoods is a minimal model. I am grateful to Andrew Y. Lee for feedback on the labels I use in this section.

(i) A high-level model (e.g., a computational model) of consciousness can be regarded as minimal in the sense that it abstracts away from the underlying neural mechanisms. In general, if a model captures only the core causal factors contributing to a phenomenon, it affords a deeper understanding by highlighting "the factors that really make a difference" (Weisberg, 2007, p. 652). In the science of consciousness, Klein et al. (2020) have recently applied this idea by arguing that, rather than searching for the neural *correlates* of consciousness, a focus should be on finding the *difference makers* of consciousness (see also Miracchi, 2017). If the model targets general properties of consciousness (not just of human consciousness), it may also help determine what properties non-human conscious systems need to have. Adding complexity to a high-level/coarse-grained minimal model will typically spoil its benefits—e.g., because this will introduce factors that are irrelevant to understanding the target phenomenon.

(ii) A measure of consciousness can also be minimal in the sense that it provides a simplified and distorted approximation to a measure of the target phenomenon. An example is given by the perturbational complexity index (PCI) (Sarasso et al., 2015, 2021), which approximates the measure of integrated information provided by integrated information theory (Albantakis et al., 2022; Oizumi et al., 2014). This has pragmatic benefits, because directly computing integrated information is not currently feasible for large systems.

The PCI provides a way of quantifying the capacity of neural systems (in different global states) to generate integrated and differentiated activity. In line with predictions of integrated information theory, there is converging empirical evidence to the effect that the PCI tracks levels of arousal, and can thereby distinguish between, e.g., alert wakefulness, anaesthesia, and disorders of consciousness (see Sarasso et al., 2021). The PCI can thus help in understanding (and measuring) consciousness in human beings. Applying it to, e.g., artificial systems, however, would mean going beyond the available evidence (because we would first have to establish which artificial systems *could* be conscious in the first place).

More generally, behavioural and cognitive markers can be regarded as simplified measures of consciousness. In principle, such markers can be applicable to non-human systems, including artificial systems, but since the evidence provided by such markers tends to be defeasible, one has to rule out potential defeaters (Schneider, 2019; Shevlin, 2020; Tomasik, 2014; Tye, 2016).

(iii) A minimal model might also target only some aspects of consciousness. In contrast to the first type of minimality, which aims to highlight the *most relevant* properties, perhaps under a high-level,

coarse-grained description, the third type of minimality may ignore factors that would be relevant to understanding the target phenomenon. Furthermore, highly selective minimal models may still be relatively detailed and provide a fine-grained description of the selected properties of the target phenomenon.

When it comes to understanding consciousness in human beings, a focus on some features of consciousness may be especially beneficial. Such *building blocks*, like, for instance, metacognition or meta-representation, can be used to augment existing theories. An example is Michael Graziano's attention schema theory, which purports to combine global workspace theory with higher-order thought theory (Graziano et al., 2019)—i.e., it purports to combine two building blocks of consciousness.[10] A model of a single building block can help in this project, even though it does not by itself provide a complete understanding of consciousness. Similarly, models of consciousness may focus on necessary (but not sufficient) conditions for consciousness.

A minimal *unifying* model (Wiese, 2020) is a particular type of selective minimal model: it highlights core assumptions that different theories of consciousness have in common; in doing so, it focuses on central, general features of a global state of consciousness (e.g., ordinary waking states), and specifies necessary (but not sufficient) conditions for such states of consciousness. Candidates for necessary features that can be used to construct a minimal unifying model include *information generation* (in the sense of Kanai et al., 2019), *complexity* (Rorot, 2021), and *temporality* (Singhal & Srinivasan, 2021). These are all candidates for necessary features that are shared by large classes of conscious experiences. Some of them figure prominently in many theories of consciousness (Wiese, 2020).[11]

One could argue that a strictly *minimal* model of consciousness that specifies a necessary condition for consciousness must specify a *minimally necessary condition* (just as neural correlates of consciousness are supposed to be *minimally sufficient*, Chalmers, 2000). A neural network model of consciousness with a single unit could count as specifying a minimally necessary condition. However, such a truly minimal necessary condition is not informative. Hence, there is a trade-off between generality and

---

[10] In joint work with Azenet Lopez, I am currently investigating such "building block approaches" to theories of consciousness more generally.

[11] The definition of a minimal unifying model provided in Wiese (2020) leaves open at what level of analysis necessary conditions for consciousness are described. For instance, information generation (Kanai et al., 2019) is most readily characterised in terms of information theory. Other cadidates for necessary conditions, such as temporality, can also be described in terms of the phenomenal properties instantiated by temporal experience. A further candidate for a necessary condition is *recurrent processing* (which in turn may require a bifurcation dynamics, Sergent & Dehaene, 2004), a feature of neuronal information processing.

informativity. The challenge, then, is to find a sweet spot between candidates for necessary conditions that are too general (which would render them uninformative), and candidates that are too specific (in which case they are less likely to be necessary). This also shows that the term "minimal" should not be taken literally in this context. The aim is not to find a strictly minimal approach, but, e.g., a minimal model in the class of models that specify informative (or heuristically useful) candidates for necessary conditions.

It should also be noted that the trade-off between generality and informativity is shaped by the intended *scope* of descriptions of necessary conditions. For instance, necessary conditions for conscious processing *in the human brain* will be much more specific than necessary conditions that are fulfilled by any conscious system. Similarly, necessary conditions for particular types of conscious experience (such as seeing red or tasting a lemon) can also be expected to be more specific than conditions that are necessary for any conscious experience at all to occur (Bayne, 2007). It may seem most useful to find conditions that are universally necessary (i.e., for any form of consciousness in any type of system), but, as Ron Chrisley has pointed out, a substantial contribution to understanding consciousness can also be made by finding "some of the necessary conditions for some way of being conscious" (Chrisley, 2008, p. 121).

A benefit of models that only specify necessary conditions is that they could be used to *rule out* that a system is conscious if it fails to satisfy a necessary condition. For instance, perhaps consciousness requires a particular type of fine-grained causal structure, which would entail that computers with a classical hardware cannot be conscious (Tononi & Koch, 2015). Of course, asserting that some condition is universally necessary for consciousness (i.e., not just in human beings, but in all kinds of system) is an extraordinarily strong claim. Hence, it also requires a strong justification. The project of finding universally necessary conditions for consciousness is therefore more ambitious than the project of finding markers of consciousness, and less ambitious than the project of finding universally necessary and sufficient conditions for consciousness (which some theories of consciousness, like integrated information theory, try to achieve).

Conversely, minimal models can also be used to argue *against* theoretical assumptions shared by different approaches, or against certain theories. For instance, Herzog et al. (2007) argue that conditions proposed by many theories of consciousness can be fulfilled by simple, non-conscious systems: "for each model of consciousness there exists a minimal model, i.e., a small neural network,

that fulfills the respective criteria, but to which one would not like to assign consciousness" (Herzog et al., 2007, p. 1055).[12]

Minimal models can not only be useful to assess theories of consciousness, but also general assumptions about consciousness (which may be shared by different theories). In a series of papers, Johannes Kleiner (2020; Kleiner & Hartmann, 2021; Kleiner & Hoel, 2021) has recently proposed and explored a general framework for mathematical models of consciousness, which allows one to formulate *minimal mathematical models* of consciousness (although the framework is not restricted to minimal models). This is particularly interesting, because it enables the derivation of conditions that are, within Kleiner's framework, *mathematically necessary* for adequate models of consciousness (e.g., Kleiner, 2020, p. 29).

In sum, different minimal models and measures of consciousness can have a variety of benefits. How does the contribution that minimal models can make to scientific understanding differ from the contribution by detailed mechanistic models?

3.2 Minimal vs. mechanistic models

In debates about minimal models in neuroscience, it is sometimes stressed that at least some minimal models are non-mechanistic (Chirimuuta, 2014; Holmes, 2021; Ross, 2015). If correct, some *minimal-model explanations* would constitute a form of explanation that is categorically distinct from mechanistic explanations. This also raises the question to what extent minimal models are explanatory, or whether they provide non-explanatory understanding.

In this section, I clarify how minimal models differ from mechanistic models. I argue that, under a relatively broad conception of mechanistic models, there is only a gradual difference between minimal and mechanistic models. If, however, mechanistic models are construed as models of particular instances of a type of phenomenon, one can draw a sharp boundary between mechanistic and minimal models. According to this construal, minimal and mechanistic models are distinct. Nevertheless, one can argue that some minimal-model *explanations* are dependent on mechanistic explanations, even if minimal models are not forms of mechanistic models. More specifically, one

[12] Such models could be described as models that are highly general, but not informative enough to count as useful minimal models of consciousness—which they are not intended to be anyway. Perhaps one could turn this into a criterion for assessing the adequacy of theories of consciousness: are minimal models that satisfies the core tenets of the theory still informative?

can argue that minimal models are logically, but not epistemically autonomous from mechanistic models – just as computational explanations are logically, but not epistemically, autonomous from knowledge of the implementing mechanisms (Rusanen & Lappi, 2016).

What is a mechanism in the first place? A mechanism is a collection of entities that are responsible for some phenomenon, by virtue of their activities or interactions. A mechanistic model of a target phenomenon specifies how the entities or parts within a mechanism are organised in such a way that their activities and interactions bring about the phenomenon (for various characterisations of mechanisms, which emphasise slightly different aspects, see Craver & Tabery, 2017).

Mechanistic explanation involves a mapping from a model to a mechanism. More specifically, variables in the model must correspond to features of the target system, and dependencies between variables must correspond to causal relations[13] between components of the target system. Kaplan & Craver (2011, p. 611) call this requirement "model-to-mechanism mapping" (3M). 3M leaves it open in what detail a model must capture the components, activities, and properties of a mechanism. Furthermore, one may wonder whether *all* variables in a model must correspond to features of a mechanism.

Fur such reasons, Craver & Kaplan (2020) provide a more nuanced formulation of 3M. According to this revised version, 3M*, at least *some* details are necessary for a successful explanation, but only *some* variables in a model must correspond to features of the target system (Craver & Kaplan, 2020, p. 297). Furthermore, they point out that more details are not always better, and not even more *relevant* details (Craver & Kaplan, 2020, pp. 303, 310).

Note that Craver and Kaplan propose a very wide characterisation of mechanistic explanation: a model can be explanatory, even if it does not capture a mechanism's properties in detail, but only captures some of its core properties.[14] In particular, a multiply-realisable model can be explanatory, because it satisfies 3M*.

---

[13] It is not obvious that all dependencies between variables must correspond to *causal* relations. Some mechanistic explanations might contain elements of structural explanations (Mancosu, 2018), and it is debatable to what extent structural (or topological) explanations are distinct from mechanistic explanations (see Kostić, 2018).

[14] Other proponents of the *New Mechanical Philosophy* are less liberal, stressing that mechanistic explanations refer to properties of particular systems and capture details that distinguish one instance of a type of phenomenon from others, because "the source of their causal powers lies in those particular instances" (Glennan, 2017, p. 3). In line with this, one can argue that more liberal approaches "dilute the mechanistic approach nearly beyond recognition" (Carrillo & Knuuttila, 2023, p. 2).

Lyre (2018) uses *dynamical models* as examples to show that explanations referring to particular mechanisms, on the one hand, and explanations referring to general (multiply-realisable) structures, on the other, can be regarded as complementary. What is more, such explanations are not even completely distinct, because dynamical models refer to the structural organisation of their mechanistic realisers. This is so in spite of the fact that dynamical and mechanistic models can be associated with different types of explanation, which Lyre calls "vertical" and "horizontal", respectively. A 'vertical' mechanistic explanation focuses on a particular instance of a phenomenon, highlighting spatiotemporal and causal relations between nested parts in a multi-level hierarchy. A 'horizontal' dynamical explanation, by contrast, focuses on high-level, generalisable properties that are shared by different instances of the phenomenon. This contrast notwithstanding, there is, as Lyre (2018, p. 5154) puts it, an "intersection point" at which the two forms of explanation meet: because some of the properties picked out by the dynamical explanation must be realised by the mechanism.

To clarify: the point is not that dynamical explanations are just incomplete or partial mechanistic explanations. That is, the point is not that dynamical explanations refer to high-level properties, and that mechanistic explanations refer to both high-level *and* low-level properties. Rather, dynamical explanations go beyond mechanistic explanations, in that they specify dynamical laws that are satisfied by a wide class of systems. At the same time, mechanistic explanations go beyond dynamical explanations, because they specify how properties of a particular mechanism, at different levels of organisation, are related. The intersection point is constituted by the properties that "directly correspond to the organizational structure of the underlying realizing mechanisms" (Lyre, 2018, p. 5142). One way to further spell this out is to say that dynamical explanations, although generalizable, also refer to some low-level properties of the underlying mechanisms; another way to spell this out is to say that the high-level properties picked out by dynamical explanations are *aspects* (or mereological parts) of their low-level realisers (Piccinini, 2020, p. 26).

Can this idea inform the relation between minimal and mechanistic models, more generally? Above, in section 3.1, I characterised minimal models in terms of three features: (i) their level of grain, (ii) their fidelity, (iii) and the number of factors that they capture. On the one hand, this characterisation is compatible with the position that minimal and mechanistic models differ only gradually, and that minimal models, to the extent that they *are* explanatory, do not provide a distinct form of explanation. On the other hand, the dynamical models referred to by Lyre (2018) are not only minimal (especially in terms of their level of grain), but also offer a 'horizontal' form of explanation, instead of a mechanistic, 'vertical' form of explanation. In general, it is therefore useful to distinguish

minimal from mechanistic models. At the same time, it should be kept in mind that minimal models are not completely *autonomous* from mechanistic explanations.

This also holds for computational models. Such models are especially relevant in the science of consciousness, because they are often considered to provide a "bridge" between accounts of the contents of consciousness and accounts of their (neural) realisers (Grush, 2006; Madary, 2016; Ramstead et al., 2020, 2022; Vilas et al., 2021; Wiese, 2018b; Williford, 2017; Yoshimi, 2014)—which also indicates that explanations involving computational models are to some extent dependent on accounts of their (neural) realisers.

To illustrate generalizable insights that computational models can provide, I review the conscious Turing machine as an example (in section 4). Before that, I shall briefly discuss whether consciousness science should strive to achieve explanations in the sense of Batterman's (2001, 2002) minimal-model explanations (as suggested by Gamma & Metzinger, 2021, p. 3; Metzinger, 2020, pp. 3–5).[15]

3.3 B-Minimal models

In this section, I briefly review Robert Batterman's work on minimal model explanations (Batterman, 2001, 2002). Following Chirimuuta (2014), I shall call Batterman-style models *B-minimal models*. A focus will be on whether this type of model (which was originally proposed to account for phenomena outside of consciousness research) is applicable to consciousness and what form of understanding it may provide.

B-minimal model explanations share with other types of minimalist approaches a commitment to the following two assumptions (see Elliott-Graves & Weisberg, 2014, p. 178): (1) the goal is to capture only the core factors[16] that give rise to a phenomenon; (2) this enables a better understanding than more detailed models. Notwithstanding these common assumptions, Batterman's account of minimal model explanation differs in a crucial respect from other minimalist approaches.

---

[15] Batterman himself does not discuss this type of explanation in the context of research on consciousness.

[16] According to Weisberg (2007, p. 643), these are causal factors (see also Weisberg, 2012, p. 103). Batterman & Rice (2014, p. 361), by contrast, stress that the common features shared by different instances of a phenomenon are not necessarily causal. Lyre (2014) argues, in discussing Batterman's (2003) interpretation of the Berry phase in quantum theory, that also minimal mathematical models admit of a realist interpretation, according to which such models refer to causal-mechanistic relations.

According to Batterman & Rice (2014), minimal model explanations require more than identifying core factors that different systems have in common. In addition, one must also specify *why* these factors are relevant *and why all other factors are irrelevant* (Batterman & Rice, 2014, p. 365).[17]

Similarly to the dynamical explanations discussed by Lyre (2018, see section 3.2 above), B-minimal models are not just highly coarse-grained mechanistic models, but provide a kind of "horizontal" explanation: according to Batterman & Rice (2014, p. 361) merely citing the core properties that are most relevant to understanding a phenomenon is not enough. The decisive explanatory work is not done by including such core properties in a model, but also involves explaining why these properties are shared by different instances of the phenomenon (Batterman & Rice, 2014, p. 370).

What would be required to successfully apply this type of explanation to minimal models of consciousness? Assume that the aim is a general understanding of consciousness (in different kinds of systems, including human beings, other animals, and artificial systems). Assume, furthermore, we have a B-minimal model of a particular global state (in different kinds of systems), such as ordinary wakefulness. In order to be explanatory (in the sense of Batterman and Rice), one would have to identify features that are relevant for this particular global state of consciousness to occur. In addition, one would have to:

1) explain why these features are relevant for this particular global state of consciousness;
2) explain why remaining details are irrelevant for the occurrence of this global state of consciousness;
3) explain why different conscious systems have these features in common (if they are in the same global state of consciousness).

To some extent, the search for neural correlates of consciousness (construed as the minimally sufficient neural basis of consciousness) can be seen as an attempt to identify the features that a B-minimal model of consciousness would specify. A neural correlate of a state of consciousness is, according to the classic definition by David Chalmers, "a minimal neural system whose state is sufficient for the corresponding conscious state" (Chalmers, 2000, p. 25). An NCC is minimal in the sense that no state of its proper parts is sufficient for the corresponding conscious state.

---

[17] Batterman and Rice distinguish their own account from what they call "common features accounts". According to such accounts, a model is explanatory by virtue of accurately capturing relevant features of the target system. The main difference between their account of B-minimal model explanation and "common features accounts" is *not* that there are no common features in B-minimal models (*pace* Mancosu, 2018). The main difference is that the explanatory benefits of minimal models do not depend on accurately representing features of the target system, but on capturing all features that are relevant for the occurrence of the phenomenon that is to be explained, and being able to specify *why* they are relevant.

Put differently, NCCs of conscious states are relevant because they correlate with consciousness, and they exclude all irrelevant properties (otherwise they would not be minimal). However, they are not explanatory. They do not explain *why* certain features are relevant to consciousness, while others are not (Seth, 2009).

Even if an account of neural correlates were explanatory (perhaps by specifying *difference makers*, Klein et al., 2020; Miracchi, 2017), there would still be a significant difference between such an explanation and a B-minimal model explanation. The latter explains why a phenomenon is displayed by a variety of different systems (Thompson, 2021). But this requires two conditions. First, the relevant properties must be described in a way that applies to different systems, not just human neural mechanisms. Second, we must know which other kinds of system are conscious, before we can explain what relevant features, if any, are shared by conscious human beings, other conscious animals, and potentially conscious artificial systems. But the question which animals are conscious is still to some extent controversial, and we are largely ignorant when it comes to artificial systems. In other words, we currently do not know which common features are generally relevant for consciousness; *a fortiori*, we cannot explain *why* these features (and not others) are relevant.

Hence, it seems that trying to achieve a B-minimal model explanation of consciousness would currently be too ambitious. In particular, the relevant features that such an explanation would have to pick out may need to be necessary and sufficient features of consciousness. Although some strands of current research may provide insight into sufficient conditions for consciousness (e.g., research on neural correlates of consciousness), merely sufficient conditions (which may not be necessary for consciousness in most animals and artificial systems) will not get us far. For B-minimalist explanations, in particular, one would also have to identify features that are shared by all conscious systems (and these features may turn out to be necessary features of consciousness).

However, minimal models of consciousness that propose necessary conditions for consciousness can be seen as partial B-minimal models. Recall that a B-minimal model explanation of consciousness would also specify *why* proposed necessary conditions are necessary for consciousness. If there are functions that are entailed by consciousness, part of such an explanation might involve showing that the proposed conditions are necessary for functions entailed by consciousness. Although a complete B-minimal model explanation would require specifying all relevant necessary conditions for

consciousness, we can make first steps in this direction by identifying some relevant necessary conditions.[18]

4 An example: The conscious Turing machine as a minimal model of consciousness

The conscious Turing machine (CTM) (Blum & Blum, 2022; Blum & Blum, 2021) characterises features of consciousness in terms of multiply realisable high-level properties. Approaches like this are desirable, because they promise to be applicable to human *and non-human* systems. They may provide a general understanding of consciousness.

The CTM can be seen as an abstract mechanistic model (like a mechanism sketch), or a non-mechanistic minimal model. Depending on this, one can expect it to provide different types of explanation or understanding. As a mechanism sketch, the CTM can be expected to provide how-possibly explanations. As a non-mechanistic minimal model, it can be expected to provide understanding that is somewhat independent of the mechanistic implementation. In particular, I shall argue that the CTM can provide non-causal mathematical explanations of some features of consciousness.

4.1 A brief overview of the conscious Turing machine

The CTM is an abstract model of consciousness, just as the Turing machine is an abstract model of computation. It is inspired by the global workspace theory (Baars, 1988). According to this theory, unconscious processing is localised processing within cognitive modules. For instance, most sensory signals that are used to control the movement of the legs while walking are not consciously experienced, because they are only processed by the motor system. Conscious processing, by contrast, involves globally available information in a *workspace* to which different cognitive modules have access. For instance, when you are carefully descending a steep slope, the relevant sensory

[18] It should also be noted that B-minimal models fit well with approaches in consciousness science that focus on features of consciousness—e.g., Seth's 'real problem' of consciousness (Seth, 2016), or multi-level constraint approaches (Metzinger, 2003/2004; Revonsuo, 1998; Wiese, 2018b) that analyse features of consciousness on different levels of description. The third why-question (why do different conscious systems have these features in common?) may complement such approaches in a particularly useful way: rather than asking why certain processes or mechanisms give rise to consciousness (the hard problem), the question asks why different global states of consciousness, or different kinds of conscious creatures share these features.

information is not just processed by your motor system, but is also available for verbal report ("I almost slipped. Next time I'll take the stairs."), it is available for voluntary attention (enabling you to focus on proprioceptive and haptic signals that are most relevant to the task at hand), and available to other cognitive subsystems.

In short, global workspace theory claims that consciousness requires orchestrating the activity of a multitude of unconscious processors. Chunks of information that are processed by individual processors are not consciously experienced, unless the information enters the global workspace and thereby becomes available to all processors. This involves a competition between chunks for access to the workspace.

The CTM provides a high-level, theoretical computer science perspective on the processes that are postulated by global workspace theory. In particular, the CTM formally defines a *chunk* and specifies an *algorithm* through which the competition for access to the workspace is resolved. The workspace is modelled as a single Short-Term Memory (STM), whereas the unconscious processors constitute the system's Long-Term Memory (LTM). The processors produce *chunks*, which are formally defined as six-tuples that contain, among others, a *gist* (the content represented by the chunk), the address of the processor that produced the chunk, the time it was produced, and a *weight* representing an estimate of the gist's importance as well as of its valence (positive or negative) (Blum & Blum, 2021, pp. 6–7).

Chunks participate in a well-defined probabilistic competition for access to the workspace (STM) by entering an *Up-Tree*, which can be considered a tournament with elimination: at each stage, chunks compete one-on-one, and only the winning chunks enter the next stage, until a single chunk reaches the top of the tree and thereby enters the workspace. This chunk is broadcast to all processors in LTM via a *Down-Tree* (Blum & Blum, 2021, pp. 8–9). The algorithm used by the CTM guarantees that the chances that chunks enter the STM are independent of their location (i.e., independent of which processors send them to the Up-Tree to enter the competition); a chunk's probability of winning the competition depends on its *mood* and *intensity*, which are initially set to the chunk's *weight* and the *absolute value* of the chunk's *weight*, respectively, but get updated during the competition (Blum & Blum, 2021, pp. 9, 23–24).

In sum, the CTM is a high-level description of processes that might be realised in conscious human beings and other animals. Since it is not a model of brain function, it does not specify how it may be implemented. It might even be that the CTM entails false claims about the processes underlying

consciousness. It is thus not obvious how the CTM can explain or provide an understanding of consciousness. The purpose of the following section is to examine these issues.

4.2 What forms of understanding are afforded by the conscious Turing machine?

The CTM is a computational model. As Rusanen & Lappi (2016) point out, computational models of capacities in physical systems are not epistemically autonomous from the implementational details. Hence, one can argue that CTM provides at most "how possibly" explanations: investigating the underlying neural mechanisms is required to determine whether one has found the *actual* explanation of the target phenomenon (Piccinini, 2020, p. 159). Still, a "how possibly" explanation may provide (non-explanatory) *understanding*, even if the explanation is merely a (false) *potential explanation* (Lipton, 2019).

The aim of this section is to further clarify what forms of understanding CTM affords. In particular, I will probe to what extent aspects of CTM *are* epistemically autonomous and to what extent CTM can provide *explanations*. I approach this by discussing two challenges for CTM, which I call the *justification challenge* and the *explanation challenge*. The justification challenge consists in showing how claims entailed by CTM can be justified; this is required to provide more than mere *how-possibly* explanations. The explanation challenge consists in specifying the explanatory contribution of CTM.

Regarding the justification challenge, recall that the CTM is inspired by the global workspace theory (Baars, 1988). To the extent that this theory, or the global *neuronal* workspace theory, is supported by empirical evidence (Mashour et al., 2020), there is also evidential support for a model that specifies an architecture and an algorithm for implementing a global workspace. In other words, if we have reason to believe that global workspace theories capture truths about consciousness, then we have reason to believe that the CTM also captures truths about consciousness. This brings us to the explanation challenge: what, if anything, does the CTM add to the explanatory understanding afforded by global workspace theories?

In general, there are at least three ways to respond to the explanation challenge. Ultimately, these strategies are complementary, so there is no need to favour one and reject the others.

(i) The first way to deal with the explanation challenge is perhaps the most conservative. It consists in showing how CTM deepens the understanding provided by empirical global workspace theories. According to this approach, CTM does not provide additional explanatory

insights.[19] However, by abstracting away from implementational details, the CTM highlights explanatorily relevant properties that can be realised by different types of system. This provides a way to *generalise* the explanations provided by empirical global workspace theories.

(ii) The second way to respond to the challenge is to argue that CTM provides *non-explanatory* understanding that goes beyond the understanding provided by global workspace theories. This move does not solve the challenge, but highlights that the CTM can also have non-explanatory benefits.

(iii) The third strategy is to point to forms of explanation that are not afforded by global workspace theories, by highlighting non-causal forms of explanatory understanding (e.g., *mathematical* explanation).

Let us consider these three strategies in more detail.

4.2.1 Generalised explanations in the CTM

Piccinini (2020) argues that functional analysis, as well as high-level computational models, only provide partial explanations. Although explanatory models are typically idealised to some extent, and need to abstract away from any irrelevant details, models that specify neural mechanisms afford a deeper understanding than high-level computational models (such as CTM), according to Piccinini:

> It's the kind of understanding that allows us to take the system apart, put it back together, or build another one like it. It's the kind of understanding that allows us to break the system in selective ways and fix it when it's broken. Nothing less than a mechanistic explanation gives us this depth of understanding" (Piccinini, 2020, p. 160)

Indeed, if we want to understand, for instance, how certain brain lesions affect consciousness, then a high-level model of consciousness won't help. But then again, one can reply that developers of minimal models "are not looking for a model of the brain but for a simple model of consciousness" (Blum & Blum, 2022, p. 3). In other words, although minimal models (such as CTM) leave many

[19] To be fair, one should note that there are some significant differences between the CTM and Baars's global workspace model. With respect to architecture, the CTM has no central executive but is a more distributed system. The Blums predict (personal communication) that a central executive is not necessary for consciousness. Other differences include CTM's rich internal multi-modal language (Brainish) for inter-processor communication and several important LTM processors, particularly the Model of the World processor, which play a significant role in CTM's "feeling of consciousness" (see Blum & Blum, 2021, sec. 4).

questions about *human* consciousness unanswered (e.g., about specific effects of certain lesions), they can answer more general questions about consciousness and associated phenomena (such as blindsight or inattentional blindness). Furthermore, because of the generality of models such as CTM, these explanations will also apply to non-human conscious systems.[20]

Here is an example from the CTM. The model can explain why the capacity of the workspace is limited; in the CTM, only one 'chunk' at a time can be globally broadcast, and the size of the chunks that enter the workspace (and its components) must be limited in size. The reason for this is, perhaps unsurprisingly, complexity. The explanatory value provided by CTM is that it makes it possible to *quantify and prove* these complexity constraints (Blum & Blum, 2022, p. 4). Because of the generality of the CTM, the results hold not just for the human brain, but (if a global workspace theory is true), for all conscious systems systems whose consciousness arises from a global workspace. In particular, the limitations of the human global workspace are not (just) driven by neural constraints, but also by universal complexity constraints.

In short, minimal models of consciousness like CTM may provide a *general* explanatory understanding, provided they are consistent with neural mechanistic details. In fact, high-level computational models may even be *required* to get a handle on consciousness in artificial systems, because the underlying mechanism in synthetic entities can be expected to bear little resemblance with the neural underpinnings of human consciousness.

4.2.2 Non-explanatory understanding in the CTM

According to the first strategy, high-level explanations of conscious phenomena are only partially autonomous, because they still depend on a bottom-up justification (Rusanen & Lappi, 2016). Is this assumption of merely partial autonomy necessary? One might attempt to defend a stronger claim, according to which a model can have explanatory value by specifying an algorithm, regardless of whether that algorithm is implemented in the brain. Related suggestions are made by Blum & Blum (2022) in the appendix[21] to their main paper:

---

[20] It has even been suggested that CTM could be developed into a 'blueprint' for artificial consciousness (Oliveira, 2022).

[21] Available at https://www.pnas.org/action/downloadSupplement?doi=10.1073%2Fpnas.2115934119&file=pnas.2115934119.sapp.pdf (visited on 20 July 2023).

> The CTM is defined by the 7-tuple < **STM, LTM, Up-Tree, Down-Tree, Links, Input, Output** >. *Other formats may be just as good*. Some format had to be chosen. […]
>
> *The binary* ***Up-Tree*** *could more generally be a* ***k****-ary tree* for some small **k**, **k** much less than **N** (the number of processors). […] [T]he **Up-Tree** is made binary because binary is both simple and sufficient, and because the choice between 2 chunks at a node is slightly simpler to describe. (p. 3; italics added);

These statements suggest that some details of the computational architecture and the algorithms to which Blum & Blum (2022) refer are to some extent arbitrary and explanatorily irrelevant. If true, this suggests that showing which specific algorithm is implemented by the brain is all the more explanatorily irrelevant. If versions of the CTM with different architectural and algorithmic details could be "just as good", then knowing that the brain implements *this* rather than *that* algorithm does not provide a deeper understanding.

But now it may almost seem as if there are no constraints at all on high-level computational models. This raises the question: what distinguishes a model that specifies a possible algorithm for the computations underpinning consciousness from an arbitrary, fictional description of consciousness? Given that no explanatory value can be expected from the explanatorily irrelevant details of the model (e.g., a binary vs. a **k**-ary tree, with $\mathbf{k} > 2$), what new insights about consciousness *does* the model provide?

To reply to this, it is useful to focus on those details of CTM that go beyond the basic tenets of the global workspace theory. Even if there is evidence that the brain implements a global workspace, there is no knowledge about which algorithm is implemented by the brain. In this situation, specifying a possible algorithm, i.e., giving a "how possibly" explanation *has* value and provides (non-explanatory) understanding. In the appendix, Blum & Blum (2022) write: "No other **GWT**-related theory gives a substantive idea how processors might decide among themselves what information to send to the stage." (p. 6). In other words, CTM addresses the question how the competition between different processors (that try to send signals to the global workspace) *could* be resolved.

Again, this is not an explanatory insight, but still provides understanding. What exactly does this understanding consist in, given that (for all we know) the brain might *not* implement the algorithm specified by CTM? In a landmark paper, Peter Lipton suggests that merely possible explanations could provide at least two forms of non-explanatory understanding (Lipton, 2019, pp. 51–52): (a) If we find out that a proposed possible explanation of why some phenomenon occurred is false, we

gain knowledge: we know that this is not why the phenomenon occurred. This, however, cannot be the understanding CTM provides by specifying a possible algorithm, because we do not know (yet) that this algorithm is not implemented by the brain. (b) Lipton makes a further suggestion, which may be applicable to the CTM: a merely possible explanation may point to invariant features, which it shares with the actual explanation, thereby "showing a degree of necessity to the actual explanation" (Lipton, 2019, p. 51). Applied to the CTM: if certain details of the computational architecture and the algorithms specified by the CTM are arbitrary, this shows that the non-arbitrary, invariant features of the algorithm and architecture are robust (if not necessary), because they are compatible with a variety of versions of the CTM. Whatever the actually true version of CTM is (if indeed there is one), we gain a deeper understanding by learning that some of its properties are not contingent on a particular version of CTM. This also speaks to the first strategy (i), according to which CTM may provide a more general understanding.

4.2.3 Non-causal explanations in the CTM

Some aspects of this more general understanding might actually be afforded by non-causal *mathematical* explanations. For instance, this may apply to the fact that a chunk's chances of entering the workspace is independent of its location (Blum & Blum, 2021, p. 23). Of course, even if the brain implements a global workspace, it might implement a version that does not have this property. But if it did, CTM would show that this property would not have to be justified by reference to causal mechanistic details; it would follow from the mathematical properties of the algorithm. In other words, part of the justification challenge would be solved mathematically. One would still need empirical support to justify *that* the brain implements an algorithm that indeed has the properties described by CTM; but the mathematical explanation *why* the algorithm has these properties would be independent of the underlying mechanistic details.

To sum up, CTM may provide different forms of understanding. It may provide a deepened understanding by generalising explanations provided by empirical global workspace theories. This has two aspects. On the one hand, CTM can highlight explanatorily relevant difference makers (high-level, multiply realisable properties). On the other hand, this also shows that some explanations generalise to systems that may have different implementing mechanisms. Furthermore, CTM may provide mathematical explanations of algorithmic properties. The claims entailed by CTM are logically, but not epistemically autonomous from knowledge of the implementing mechanisms (Rusanen & Lappi, 2016). However, at least mathematical explanations of *why* the algorithm specified

by CTM has certain properties are epistemically autonomous (even though the claim *that* the brain implements an algorithm with these properties is *not* epistemically autonomous).

5 Conclusion

I have provided a taxonomy of minimalist approaches to consciousness and discussed how they can promote the science of consciousness. Minimal models and measures can be minimal with respect to (i) their level of grain, (ii) their fidelity, or (iii) the number of contributing factors captured by them (see also Weisberg, 2007). Minimal models can be distinguished from mechanistic models, but some minimal models may only gradually differ from abstract mechanistic models.

In general, minimal models of consciousness are not autonomous from implementational details: even if claims about explanatorily relevant factors are independent from their mechanistic implemenation, they are not epistemically autonomous (Rusanen & Lappi, 2016), because they need to be justified by showing how high-level properties are realised by the underlying mechanisms (Piccinini, 2020). Nevertheless, claims entailed by minimal models of consciousness *can* be autonomous from implementational details if they provide potential explanations that afford non-explanatory understanding or non-causal mathematical explanations.

Achieving a genuine minimal-model explanation (in the sense of Batterman & Rice, 2014) is a goal that current consciousness science can approximate in certain ways, even if not all of its aspects are currently attainable. A particular challenge is to find answers to the distribution question (which non-human systems are conscious?), since minimal-model explanations specifically explain *why* some phenomenon is displayed by a variety of different systems (Thompson, 2021).

The conscious Turing machine (Blum & Blum, 2022; Blum & Blum, 2021) can be regarded as a minimal model of consciousness that may provide different forms of explanatory and non-explanatory understanding. The conscious Turing machine is inspired by global workspace theory. For future research, it will be particularly interesting to explore minimal models that are inspired by other theories, or to investigate to what extent minimalist assumptions play a role in theories of consciousness (such as the *principle of minimal existence* in integrated information theory, see Albantakis et al., 2022, p. 11).

Acknowledgments: Many thanks to Azenet Lopez, Holger Lyre, Alfredo Vernazzani, and Lenore and Manuel Blum for a number of very helpful comments on a draft of this paper. I am also grateful

to (in alphabetical order) Ricarda Haeseler, Christoph Hausdorf, Ramón Imort, Andrew Y. Lee, Nicolas Loerbroks, and to all people who attended presentations of earlier versions of this paper at ASSC26 in New York, at MoC3 in Stanford, and at GAP11 in Berlin.

References

Achinstein, P. (1983). *The nature of explanation*. Oxford University Press.

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2022). *Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. arXiv:2212.14787*. http://arxiv.org/abs/2212.14787

Aru, J., Suzuki, M., & Larkum, M. E. (2020). Cellular mechanisms of conscious processing. *Trends in Cognitive Sciences*, *24*(10), 814–825. https://doi.org/10.1016/j.tics.2020.07.006

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Batterman, R. W. (2001). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press. https://doi.org/10.1093/0195146476.001.0001

Batterman, R. W. (2002). Asymptotics and the role of minimal models. *The British Journal for the Philosophy of Science*, *53*(1), 21–38.

Batterman, R. W. (2003). Falling cats, parallel parking, and polarized light. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *34*(4), 527–557. https://doi.org/10.1016/S1355-2198(03)00062-5

Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, *81*, 349–376. https://www.jstor.org/stable/10.1086/676677

Bayne, T. (2007). Conscious states and conscious creatures: Explanation in the scientific study of consciousness. *Philosophical Perspectives*, *21*(1), 1–22. https://doi.org/10.1111/j.1520-8583.2007.00118.x

Bayne, T. (2010). *The unity of consciousness*. Oxford University Press.

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are there levels of consciousness? *Trends in Cognitive Sciences*, *20*(6), 405–413. https://doi.org/10.1016/j.tics.2016.03.009

Bengson, J. (2017). The unity of understanding. In S. R. Grimm (Ed.), *Making sense of the world: New essays on the philosophy of understanding* (pp. 14–53). Oxford University Press.

Birch, J. (2022). The search for invertebrate consciousness. *Noûs*, *56*(1), 133–153. https://doi.org/10.1111/nous.12351

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247.

Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. *Proceedings of the National Academy of Sciences*, *119*(21), e2115934119. https://doi.org/10.1073/pnas.2115934119

Blum, M., & Blum, L. (2021). A theoretical computer science perspective on consciousness. *Journal of Artificial Intelligence and Consciousness*, *08*(01), 1–42. https://doi.org/10.1142/S2705078521500028

Carrillo, N., & Knuuttila, T. (2023). Mechanisms and the problem of abstract models. *European Journal for Philosophy of Science*, *13*(3), 27. https://doi.org/10.1007/s13194-023-00530-z

Carter, J. A., & Gordon, E. C. (2014). Objectual understanding and the value problem. *American Philosophical Quarterly*, *51*(1), 1–13.

Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions* (pp. 17–40). MIT Press.

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

Chrisley, R. (2008). Philosophical foundations of artificial consciousness. *Artificial Intelligence in Medicine*, *44*(2), 119–137. https://doi.org/10.1016/j.artmed.2008.07.011

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*, *71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Craver, C., & Tabery, J. (2017). Mechanisms in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017). https://plato.stanford.edu/archives/spr2017/entries/science-mechanisms/; Metaphysics Research Lab, Stanford University.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Mit Press.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*, 486–492.

Del Pin, S. H., Skóra, Z., Sandberg, K., Overgaard, M., & Wierzchoń, M. (2021). Comparing theories of consciousness: Why it matters and how to do it. *Neuroscience of Consciousness*, *2021*(2). https://doi.org/10.1093/nc/niab019

Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, *132*(1), 33–42. https://doi.org/10.1007/s11098-006-9054-z

Elliott-Graves, A., & Weisberg, M. (2014). Idealization. *Philosophy Compass*, *9*(3), 176–185. https://doi.org/10.1111/phc3.12109

Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, *2020*(1), niz020. https://doi.org/10.1093/nc/niz020

Gamma, A., & Metzinger, T. (2021). The minimal phenomenal experience questionnaire (MPE-92M): Towards a phenomenological profile of "pure awareness" experiences in meditators. *PLOS ONE*, *16*(7), e0253694. https://doi.org/10.1371/journal.pone.0253694

Giacino, J. T., Ashwal, S., Childs, N., Cranford, R., Jennett, B., Katz, D. I., Kelly, J. P., Rosenberg, J. H., Whyte, J., Zafonte, R. D., & Zasler, N. D. (2002). The minimally conscious state: Definition and diagnostic criteria. *Neurology*, *58*(3), 349–353. https://doi.org/10.1212/WNL.58.3.349

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. MIT Press.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Graziano, M. S. A., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2019). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cogn Neuropsychol*, 1–18. https://doi.org/10.1080/02643294.2019.1670630

Grush, R. (2006). How to, and how not to, bridge computational cognitive neuroscience and Husserlian phenomenology of time consciousness. *Synthese*, *153*(3), 417–450.

Hannon, M. (2021). Recent work in the epistemology of understanding. *American Philosophical Quarterly*, *58*(3), 269–290. https://doi.org/10.2307/48616060

Herzog, M. H., Esfeld, M., & Gerstner, W. (2007). Consciousness & the small network argument. *Neural Networks*, *20*(9), 1054–1056. https://doi.org/10.1016/j.neunet.2007.09.001

Holmes, T. (2021). Cognitive dynamical models as minimal models. *Synthese*, *199*(1), 2353–2373. https://doi.org/10.1007/s11229-020-02888-6

Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M., & Guttenberg, N. (2019). Information generation as a functional basis of consciousness. *Neuroscience of Consciousness*, *2019*(1). https://doi.org/10.1093/nc/niz016

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627. https://doi.org/10.1086/661755

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108164276

Kim, J. (1974). Noncausal connections. *Noûs*, *8*(1), 41–52. https://doi.org/10.2307/2214644

Kim, J. (1994). Explanatory knowledge and metaphysical dependence. *Philosophical Issues*, *5*, 51–69. https://doi.org/10.2307/1522873

Klein, C., Hohwy, J., & Bayne, T. (2020). Explanation in the science of consciousness: From the neural correlates of consciousness (NCCs) to the difference makers of consciousness (DMCs). *Philosophy and the Mind Sciences*, *1*(IIII). https://doi.org/10.33735/phimisci.2020.II.60

Kleiner, J. (2020). Mathematical models of consciousness. *Entropy*, *22*(66), 609. https://doi.org/10.3390/e22060609

Kleiner, J., & Hartmann, S. (2021). The closure of the physical is unscientific. *arXiv:2110.03518 [q-Bio]*. http://arxiv.org/abs/2110.03518

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(niab001). https://doi.org/10.1093/nc/niab001

Kostić, D. (2018). Mechanistic and topological explanations: An introduction. *Synthese*, *195*(1), 1–10. https://doi.org/10.1007/s11229-016-1257-z

Kvanvig, J. L. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge University Press.

Kvanvig, J. L. (2018). Knowledge, understanding, and reasons for belief. In D. Starr (Ed.), *The Oxford handbook of reasons and normativity* (pp. 685–705). Oxford University Press.

Lau, H. (2022). *In consciousness we trust: The cognitive neuroscience of subjective experience*. Oxford University Press.

Laureys, S., Celesia, G. G., Cohadon, F., Lavrijsen, J., León-Carrión, J., Sannita, W. G., Sazbon, L., Schmutzhard, E., Wild, K. R. von, Zeman, A., Dolce, G., & European Task Force on Disorders of Consciousness, the. (2010). Unresponsive wakefulness syndrome: A new name for the vegetative state or apallic syndrome. *BMC Medicine*, *8*(1), 68. https://doi.org/10.1186/1741-7015-8-68

Lipton, P. (2004). *Inference to the best explanation* (2. ed.). Routledge. (Original work published 1991)

Lipton, P. (2019). Understanding without explanation. In H. W. de Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 43–63). University of Pittsburgh Press.

Lyre, H. (2014). Berry phase and quantum structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *48*, 45–51. https://doi.org/10.1016/j.shpsb.2014.08.013

Lyre, H. (2018). Structures, dynamics and mechanisms in neuroscience: An integrative account. *Synthese*, *195*(12), 5141–5158. https://doi.org/10.1007/s11229-017-1616-4

Madary, M. (2016). *Visual phenomenology*. MIT Press.

Mancosu, P. (2018). Explanation in mathematics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/mathematics-explanation/

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Mason, J. W. D. (2021). Model unity and the unity of consciousness: Developments in expected float entropy minimisation. *Entropy*, *23*(1111), 1444. https://doi.org/10.3390/e23111444

Mckilliam, A. K. (2020). What is a global state of consciousness? *Philosophy and the Mind Sciences*, *1*(IIII). https://doi.org/10.33735/phimisci.2020.II.58

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity* (2nd ed.). MIT Press. (Original work published 2003)

Metzinger, T. (2020). Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of "pure" consciousness. *Philosophy and the Mind Sciences*, *1*(I), 7. https://doi.org/10.33735/phimisci.2020.I.46

Miłkowski, M. (2016). Unification strategies in cognitive science. *Studies in Logic, Grammar and Rhetoric*, *48*(1), 13–33. https://doi.org/10.1515/slgr-2016-0053

Miracchi, L. (2017). Generative explanation in cognitive science and the hard problem of consciousness. *Philosophical Perspectives*, *31*(1), 267–291. https://doi.org/https://doi.org/10.1111/phpe.12095

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Oliveira, A. L. (2022). A blueprint for conscious machines. *Proceedings of the National Academy of Sciences*, *119*(23), e2205971119. https://doi.org/10.1073/pnas.2205971119

Piccinini, G. (2020). *Neurocognitive mechanisms*. Oxford University Press.

Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.

Ramstead, M. J. D., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., Pagnoni, G., Smith, R., Dumas, G., Lutz, A., Friston, K., & Constant, A. (2022). From generative models to generative passages: A computational approach to (neuro) phenomenology. *Review of Philosophy and Psychology*, *13*(4), 829–857. https://doi.org/10.1007/s13164-021-00604-y

Ramstead, M. J. D., Wiese, W., Miller, M., & Friston, K. J. (2020). *Deep neurophenomenology: An active inference account of some features of conscious experience and of their disturbance in major depressive disorder*. http://philsci-archive.pitt.edu/18377/

Regt, H. W. de. (2004). Discussion note: Making sense of understanding. *Philosophy of Science*, *71*(1), 98–109. https://doi.org/10.1086/381415

Revonsuo, A. (1998). How to take consciousness seriously in cognitive neuroscience. *Communication and Cognition: An Interdisciplinary Quarterly Journal*, *30*(3-4), 185–205.

Rorot, W. (2021). Bayesian theories of consciousness: A review in search for a minimal unifying model. *Neuroscience of Consciousness*, *2021*(2), niab038. https://doi.org/10.1093/nc/niab038

Ross, L. N. (2015). Dynamical models and explanation in neuroscience. *Philosophy of Science*, *82*(1), 32–54. https://doi.org/10.1086/679038

Ross, L., & Woodward, J. (2023). Causal approaches to scientific explanation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2023). https://plato.stanford.edu/archives/spr2023/entries/causal-explanation-science/; Metaphysics Research Lab, Stanford University.

Rusanen, A.-M., & Lappi, O. (2016). On computational explanations. *Synthese*, *193*(12), 3931–3949. https://doi.org/10.1007/s11229-016-1101-5

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, A. G., Brichant, J.-F., Boveroux, P., et al. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, *25*(23), 3099–3105. https://doi.org/10.1016/j.cub.2015.10.014

Sarasso, S., Casali, A. G., Casarotto, S., Rosanova, M., Sinigaglia, C., & Massimini, M. (2021). Consciousness and complexity: A consilience of evidence. *Neuroscience of Consciousness*, *niab023*. https://doi.org/10.1093/nc/niab023

Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon?: Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, *15*(11), 720–728. https://doi.org/10.1111/j.0956-7976.2004.00748.x

Seth, A. K. (2009). Explanatory correlates of consciousness: Theoretical and computational challenges. *Cognitive Computation*, *1*(1), 50–63. https://doi.org/10.1007/s12559-009-9007-x

Seth, A. K. (2016). The real problem. *Aeon*. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, 1–14. https://doi.org/10.1038/s41583-022-00587-4

Shevlin, H. (2020). General intelligence: An ecumenical heuristic for artificial consciousness research? *Journal of Artificial Intelligence and Consciousness*, *07*(02), 245–256. https://doi.org/10.1142/S2705078520500149

Singhal, I., & Srinivasan, N. (2021). Time and time again: A multi-scale hierarchical framework for time-consciousness and timing of cognition. *Neuroscience of Consciousness*, *2021*(2), niab020. https://doi.org/10.1093/nc/niab020

Strevens, M. (2011). *Depth: An account of scientific explanation*. Harvard University Press.

Thompson, J. A. F. (2021). Forms of explanation and understanding for neuroscience and artificial intelligence. *Journal of Neurophysiology*, *126*(6), 1860–1874. https://doi.org/10.1152/jn.00195.2021

Tomasik, B. (2014). *Do artificial reinforcement-learning agents matter morally? arXiv:1410.8233*. http://arxiv.org/abs/1410.8233

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B*, *370*, 20140167. https://doi.org/10.1098/rstb.2014.0167

Tye, M. (2016). *Tense bees and shell-shocked crabs*. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780190278014.001.0001/acprof-9780190278014

Vilas, M. G., Auksztulewicz, R., & Melloni, L. (2021). Active inference as a computational framework for consciousness. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00579-w

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, *104*(12), 639–659. https://doi.org/10.5840/jphil20071041240

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Wiese, W. (2016). How to solve the problem of phenomenal unity: Finding alternatives to the single state conception. *Phenomenology and the Cognitive Sciences*, 1–26. https://doi.org/10.1007/s11097-016-9478-7

Wiese, W. (2018a). *Experienced wholeness. Integrating insights from Gestalt theory, cognitive neuroscience, and predictive processing*. MIT Press.

Wiese, W. (2018b). Toward a mature science of consciousness. *Frontiers in Psychology*, *9*, 693. https://doi.org/10.3389/fpsyg.2018.00693

Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness*, *2020*(1). https://doi.org/10.1093/nc/niaa013

Wiese, W. (2022). Attentional structure and phenomenal unity. *Open Philosophy*, *5*(1), 254–264. https://doi.org/10.1515/opphil-2022-0197

Williford, K. (2017). A brief on Husserl and Bayesian perceptual updating. *Axiomathes*, *27*(5), 503–519. https://doi.org/10.1007/s10516-017-9342-6

Windt, J. M. (2015). Just in time. Dreamless sleep experience as pure subjective temporality. In T. K. Metzinger & J. M. Windt (Eds.), *Open MIND*. MIND Group. https://doi.org/10.15502/9783958571174

Yoshimi, J. (2014). Narrowing the explanatory gap with bridge metaphors [Conference Proceedings]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 3143–3148.