Routledge
Taylor & Francis Group

Australasian Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rajp20

Really Real Patterns

Tyler Millhouse

To cite this article: Tyler Millhouse (2021): Really Real Patterns, Australasian Journal of Philosophy, DOI: 10.1080/00048402.2021.1941153

To link to this article: https://doi.org/10.1080/00048402.2021.1941153

AUSTRALASIAN JOURNAL OF PHILOSOPHY
https://doi.org/10.1080/00048402.2021.1941153

Really Real Patterns

Tyler Millhouse

Santa Fe Institute

ABSTRACT
Dennett [1991] proposes a novel ontological account of the propositional attitudes—real patterns. Despite its name, the degree to which this account is committed to realism remains unclear. In this paper, I propose an alternative criterion of pattern instantiation, one that assesses the difficultly of faithfully interpreting a physical system as instantiating a particular pattern. Drawing on formal measures of simplicity and similarity, I argue that, for well-instantiated patterns, our interpretation will be computable by using a short program. This approach preserves the flexibility of Dennett's original, while offering substantially stronger realist commitments.

ARTICLE HISTORY Received 23 May 2020; Revised 14 May 2021

KEYWORDS patterns; models; information; similarity; realism; ontology

1. Introduction

Dennett's account of *real patterns* proposes a kind of 'mild realism' about the propositional attitudes. He outlines a *via media* between Fodor's 'industrial strength' realism and the Churchlands' eliminative materialism [Dennett 1991: 30]. Despite his focus on the philosophy of mind, the idea of real patterns has recently enjoyed a renaissance in the philosophy of science. For example, Ladyman and Ross [2007, 2013] and Wallace [2012] have put the idea to work as (at least) a general account of special science ontology. For application to particular domains, see Ross [1995], Ross and Spurrett [2004], Anderson [2017], or Burnston [2017]. For additional criticism and discussion, see Millhouse [2021]. Drawing on information theory, Dennett's key insight is to exploit the connection between regularity and compression. This allows him to craft a compelling account of real patterns as objective regularities in physical phenomena. Nevertheless, the view occupies an uneasy position between realism and instrumentalism, as Dennett acknowledges [1991].

My primary aim here is not to settle the question of whether Dennett is a realist or an instrumentalist. Rather, my aim is to argue for a new and more demanding criterion for the reality of patterns. This criterion is inspired by real patterns, both in its original form and in later interpretations [D. Wallace 2012], but it also builds on algorithmic information theory and on similarity criteria of model fidelity [Weisberg 2016]. While Dennett relies on the connection between compression and regularity, I rely on the connection between compression and similarity. The resulting account preserves all of the key insights of real patterns (graded instantiation, objective evaluability, and the connection between patterns and compressibility), while satisfying key realist

desiderata (preserving multiple realizability and explaining scientific scepticism about model instantiation).

The paper is organized as follows: In section 2, I review real patterns as proposed by Dennett [1991], and I explain and defend the concern that real patterns courts instrumentalism. In section 3, I propose a novel criterion of pattern instantiation, motivated by information-theoretic criteria of similarity and by recent philosophical work on physical computation and scientific modelling. In section 4, I respond to the concerns of a contemporary defender of real patterns [D. Wallace 2012], and I suggest a strategy for deploying this criterion as a highly general tool for assessing scientific models. Finally, in section 5, I briefly review the proposed criterion and highlight directions for future research.

2. Real Patterns

In data science, it is generally recognized that compression is achieved by exploiting regularities (or *patterns*) in data [C. Wallace 2005]. In algorithmic information theory, the number of bits required to encode a particular message is construed as the length of the shortest program (on a reference universal Turing machine) that halts with the message on the tape. This is known as the *Kolmogorov complexity* of the message [Sipser 2013].[1] A string of 10,000 bits chosen at random will probably require a program about as long as the string itself, since there is probably no pattern that a program could exploit in generating the string. In contrast, a string of 10,000 zeros could be represented by a short program that simply prints '0' 10,000 times [Li and Vitányi 2008].[2] Fortunately, the details of algorithmic information theory need not concern us here. Adverting to a formal criterion of complexity is primarily useful for focusing our attention on what such a criterion must deliver and for illustrating how a precisification of our ordinary notion of simplicity might proceed.

To understand how Dennett [1991] makes philosophical use of the connection between regularity and compression, it is worth revisiting one of his central examples—image compression. This case illustrates several of Dennett's key philosophical insights. The first is that *the presence of a pattern in data is a matter of degree*. Dennett considers a simple checkerboard image to which various amounts of random noise have been added. By conducting a simple experiment, one can see that the addition of random noise to an image increases the size of the image file when a compressed file format is used but not when an uncompressed format is used (see Fig. 1).

Moreover, the checkerboard pattern remains recognizable for some time, as more and more noise is added. The fact that the pattern is still present *to some degree* can be confirmed by comparing a noisy (but recognizable) checkerboard image to an image consisting *entirely* of random noise (again see Fig. 1). The former will require a smaller file, demonstrating that a pattern is present in the noisy checkerboard image. Of course, if we continue adding random noise to the checkerboard image, its file size will increase gradually until the pattern is effaced.

[1] There might be more than one shortest program, but I have retained the definite article to avoid the awkwardness of using the indefinite article with a superlative in English.

[2] There are analogous results in Shannon information theory [Shannon 1948], but the algorithmic approach is more natural here since programs make explicit the patterns exploited in compression. For an overview of the relationship between Shannon information and Kolmogorov complexity, see Grünwald and Vitányi [2004].

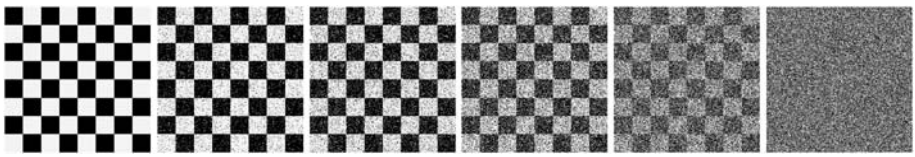

Figure 1: six images with increasing noise levels. The uncompressed file size for every image is 65 KB, while the compressed sizes are as follows (from left to right): 1 KB, 19 KB, 30 KB, 46 KB, 58 KB, and 65 KB.

Dennett's second key insight is that *patterns in data are objective*. Whether a compressed file is smaller than an uncompressed file is simply a matter of counting bits. To put the point in terms of algorithmic information theory, it is trivial to determine whether one program (represented as a string of ones and zeros) is shorter than another.[3] As Dennett argues, 'A pattern exists in some data—is real—if there is a description of the data that is more efficient than the bit map, whether or not anyone can concoct it' [ibid.: 34]. In the real world, these data may describe anything from baseball statistics to astronomical measurements. What is essential is that we have some raw data that we can recover from a more economical description. The fidelity and economy of this description tells us to what extent a pattern is present in the data, while the details of that description tell us the nature of the pattern itself.

This seems like an appropriate way to think about patterns in static data, but it remains unclear what compression has to say about predictive models of physical systems whose state evolves over time. I will say more about this in a subsequent section, but for now I will focus on Dennett's approach. Essentially, a model of a physical system (like a human brain) can be cast at different levels of abstraction. To describe the operation of the brain (and hence to account for its behaviour) at the microphysical level would require a vastly complicated model. Fortunately, we can also adopt a simpler, high-level model—for example, one that posits a belief-desire psychology. Such models have remarkable accuracy, given their simplicity *relative* to low-level models. Dennett argues that this economy, consistently with compression in general, depends on regularities in human behaviour and, indirectly, on the cognitive processes that generate it.[4]

Under this compressibility criterion, models are favoured to the extent that they accurately and economically represent the relationships between variables of interest. These variables, Dennett argues, are fixed via an 'interpretation scheme' that identifies such variables for subsequent modelling [ibid.: 41]. For example, Dennett considers an intentional systems model of a chess-playing computer. In this case, the interpretation scheme relies on an 'ontology of chess-board positions, possible chess moves, and the grounds for evaluating them' [ibid.]. Once this interpretation is fixed, argues Dennett, we can develop a simple and predictive model of the computer that looks at the state of the board, posits intentional states (including beliefs about the state of the board, instrumentally useful moves, or what its opponent might do), and predicts chess

[3] The choice of encoding scheme and reference universal Turing machine complicates matters, but there are convergence guarantees for different choices that suffice for practical purposes [Sipser 2013]. See Millhouse [2019] for a philosophical discussion of reference machine selection.

[4] The idea that modelling can be viewed as a kind of compression has also been proposed in data science/information theory [C. Wallace 2005; Rathmanner and Hutter 2011]. However, I argue [2021] that these proposals differ in important ways from Dennett's.

moves. For Dennett, the simplicity and predictive power of this model depends on the fact that the model captures a real pattern in the target system (that is, the computer).

Nevertheless, two models can agree in their simplicity and predictive power, yet differ in their ontology (that is, the things that they propose). Neither complexity nor predictive accuracy could distinguish between such models. One might wish to say that such models are only superficially different and that they offer fundamentally consistent descriptions of underlying physical processes. Unfortunately, this position does not bear much scrutiny.

Following Pylyshyn [1984], consider the fact that there are often several different algorithms that solve the same problem, including a host of algorithms for sorting lists. Further, these algorithms often do not differ significantly in program length [Cormen et al. 2009].[5] Computer scientists have even developed conventions for representing the algorithmic differences between programs. This 'pseudocode' is not a programming language, but rather an informal set of conventions for describing algorithms that abstracts away from the finer points of implementation [ibid.: 3]. It is hard to see how Dennett's criterion could capture this element of scientific practice without addressing both (i) what function a program computes and (ii) how it computes that function via a series of computational steps.[6] For analogous reasons, it seems reasonable to ask that real patterns address both chess playing behaviour and the cognitive states and processes proposed to explain such behaviour.

3. Really Real Patterns

3.1 Desiderata for a Realist Theory

Just as Dennett aims to distinguish 'real' patterns from 'bogus' patterns, my aim is to provide a criterion for determining when (and to what degree) a proposed pattern is *really* instantiated by a physical system [1991: 29]. As discussed above, this criterion should consider both (i) the power of a pattern to predict a particular phenomenon and (ii) what it entails about how the target system gives rise to that phenomenon. Here I follow Weisberg [2016] in distinguishing between *dynamical fidelity criteria* and *representational fidelity criteria*. According to him, 'Dynamical fidelity criteria tell us how close the output of the model—the predictions it makes about the values of dependent variables given some independent variables—must be to the output of the real-world phenomenon' [ibid.: 267]. Representational fidelity criteria tell us 'whether the structure of the model *maps well* onto the target system of interest' [ibid., emphasis mine].

This distinction is warranted in light of scepticism *within science* about whether predictively successful models comport with reality (see Maddy [1992]). It is also warranted by the fact that one aim of scientists is to capture the structure of their target systems [Godfrey-Smith 2006]. A realist account of pattern instantiation should

[5] Given a richer set of variables (e.g. one that includes processing time), we might be able to distinguish between these algorithms via predictive economy. However, Dennett gives no account of when a data set is sufficiently rich to support ontological conclusions about a model's posits, and the information-theoretic justification for real patterns does not (by my lights) suggest one.

[6] To be clear, nothing about Dennett's view precludes our applying his criterion to a model whose data are computational state histories rather than input/output pairs. In this case, a simple and predictive model *would* indicate a particular pattern in those states. However, in this case, the computational states are not an explanatory posit of the model; they are (presumably) a pre-established feature of the target system.

make sense of such concerns by explaining how a predictively successful model might fail to capture the structure of its target system.

A realist criterion must also preserve multiple realizability and avoid chauvinism about particular realizers. Despite the problems outlined in the previous section, there is something deeply right about Dennett's 'maximal neutrality' about pattern realization [2009: 346]. For example, multiple realizability is a key feature of functionalist accounts of the mind, and a revised account of real patterns should preserve this feature. To do so, it must be clear how the structure proposed by a model can be realized by physically dissimilar systems.

In developing my criterion, I will focus on models that propose an *explanation* of some phenomenon. Following Maudlin, explanatory models are those that *at least* answer two questions—'What is there?' and 'How does it behave?' [2015: 64]. A model's answers to these questions tell us about its ontology and its dynamics, respectively. For example, dynamical models are paradigmatically explanatory, since they posit a phase space as well as an evolution rule describing the time evolution of points in that space. In other words, they describe the possible states of a system and provide laws governing the evolution of that system over time. Explanatory models are of interest here because they raise the kind of ontological questions that my account is meant to answer. When evaluating an explanatory model, we can ask whether it offers an accurate picture of what there is and how it behaves over and above its ability to predict certain variables in terms of others. For models that offer no such picture (like linear models), this question cannot arise.

3.2 Compressibility and Similarity

One family of views proposes that model accuracy amounts to a kind of structural similarity between models and the target systems that they model (for example, Giere [1988], Godfrey-Smith [2006], and Weisberg [2016]). As it happens, there are deep connections between compressibility and similarity—just as there are deep connections between compressibility and regularity. The intuition here is fairly straightforward. Suppose that we want to quantify the similarity of two objects, A and B. One simple strategy would be to count the differences between A and B. While the basic approach seems sensible, the strategy itself is highly underspecified. How are differences individuated? How do we handle different degrees of difference?

Earlier, we considered the complexity of binary strings (section 2). Now we'll consider measures of similarity between binary strings. As it happens, there are multiple ways to formalize the notion of similarity between strings. To appreciate how such a formalization might go, consider the concept of *edit distance*—in particular, *Levenshtein distance* [Levenshtein 1966]. The Levenshtein distance between two strings is the minimum number of single character edits (insertion, deletion, and substitution) required to convert one string into another. Consider the following strings:

(A) 01111111111111111111
(B) 00000000000000000000

The Levenshtein distance between A and B is 19, since nineteen substitutions must be made. However, what is the rationale for limiting ourselves to single character edits? If these strings continued in the same fashion for hundreds of bits, it might become more

economical to use instructions like this: go to digit 2; loop *n* times: substitute 0; and move right. Compare that to this one: go to digit 2; substitute 0; go to digit 3; substitute 0 … ; and so on. A more limited set of allowed operations makes sense when, say, one is trying to guess intended words from misspelled words. Spelling errors are often the result of single character errors, like substituting a wrong letter, omitting a correct letter, or adding an incorrect letter [Wagner and Fischer 1974].

However, what if we want a domain general measure of similarity? *Algorithmic information distance* is just such a measure. The algorithmic information distance between two strings is the length of the shortest program on a reference universal Turing machine that generates one string, given another, and *vice versa* [Bennett et al. 1998]. What algorithmic information distance and edit distance have in common is that they rely on the complexity of a specific procedure in order to establish similarity. However, by appealing to the computationally complete capacities of a universal Turing machine rather than a limited set of edit operations, algorithmic information distance provides a highly general measure of similarity [ibid.].

The key point is that a program for recovering A from B and *vice versa* must contain just enough bits to reproduce those elements of A that B lacks (and *vice versa*). Thus, the *shortest possible* program that maps A to B and B to A will contain only as many bits as are required to encode the differences between the strings. In this sense, algorithmic information distance measures the degree of similarity between A and B, with shorter programs indicating greater similarity.

3.3 A Compressibility Criterion for Pattern Instantiation

As I mentioned earlier, one way to think about model accuracy is as a kind of structural similarity between a model and the world. For present purposes, I will assume this view of model accuracy and offer a similarity criterion for pattern instantiation. This criterion construes pattern instantiation as a relation between an instantiated model and an instantiating system. The pattern described by an explanatory model is its account of a system's behaviour defined over a particular representation of that system (such as the rule-governed evolution of points in a phase space). This pattern is well-instantiated to the extent that the model is well-instantiated. A model is well-instantiated to the extent that it bears the correct relation to the instantiating system—structural similarity.

In formalizing this criterion, I will assume that we are interested in determining whether a model is well-instantiated by a physical system as described by an accurate fundamental physical model.[7] With minor modifications, however, the criterion will allow us to determine whether a high-level model is well-instantiated by a physical system as described by an accurate *intermediate*-level model. For simplicity, I will also assume that both of the relevant models are dynamical systems models.

When we think about model accuracy as a kind of similarity between models, we are presumably interested in the similarity of their ontology and dynamics. In the case of dynamical systems models, this means that we are interested in assessing what each model says about (i) the states that a system can be in (the phase space of the

[7] A discussion of the accuracy conditions for fundamental physical models would require a separate treatment. Fortunately, even if we assume an accurate low-level model of a system, deep questions remain about the instantiation of higher-level models.

model) and (ii) how those states evolve over time (its evolution rule). In the philosophical literature on computer instantiation, one strategy has been to evaluate computer instantiation in terms of a counterfactual-supporting mapping between states of a physical system and states of an abstract machine [Chalmers 1996]. That is, a physical system instantiates a particular abstract machine *iff* states of the system can be mapped to states of the machine such that the system *under this mapping* obeys the machine's transition function. Unfortunately, this criterion permits a flood of spurious mappings that threaten to trivialize physical computation. Taking a cue from politics, these mappings are often described as 'gerrymandered,' since they ascribe implausible computations to a system without regard for the features of its physical states. They map very different physical states to the same computational state or map very similar physical states to different computational states.

To remedy this problem, I have proposed [2019] a simplicity constraint on mappings, and I motivated this constraint by appealing to program length as a measure of similarity. Earlier, we considered the shortest program for computing a function that accepts one string and returns another *and vice versa*. This two-way program is appropriate, since similarity is a symmetric relation. Instantiation, however, is an asymmetric relation. Fortunately, we can also consider the length of the shortest *one-way* program (that is, the program that simply accepts one string and returns another). These program lengths (or *conditional Kolmogorov complexities*) are closely related to algorithmic information distance and tell us about a specific aspect of similarity between strings [Bennett et al. 1998]. In particular, they tell us how much additional information is required to produce one string, given free access to another.

While conditional Kolmogorov complexity considers the length of the shortest program that accepts one string and returns another, we can just as easily consider the length of the shortest program that accepts physical states and returns computational states. This measure—analogous but not identical to conditional Kolmogorov complexity—tells us how much information is required to determine the computational state of a system on the basis of its physical state. I argue that requiring a short program precludes gerrymandering and secures the related (and highly plausible) aim of ensuring that our mapping of physical states to computational states reflects structural features of the physical system. My original argument [ibid.] focused more on naturalness than on similarity, and so the connection between compressibility and similarity was not entirely clear. For that reason, I will reconstruct and clarify that argument here, focusing on similarity.

The key point (for present purposes) is that mapping complexity increases whenever (i) physically different states must be treated as a single computational state or (ii) physically similar states must be treated as distinct computational states. By 'physically different states', I do not mean 'states that are merely dissimilar', but instead 'states that share no concise set of features that distinguish them from other states'. For example, if I randomly sort baseball cards into two stacks, there will probably be no feature that the cards in one stack share and the cards in the other stack lack. In contrast, suppose that I sort the cards into two piles, according to whether the cards were issued before or after 1960. In the former case, it would be difficult to offer a brief description of which cards were in which pile. In the latter case, a brief description (such as the one that I gave just now) would be easy to construct. Given this, recall that what a mapping between physical and computational states must do

is to tell us which physical states belong to which computational states. In other words, there will be sets of physical states that belong to a single computational state. As with the baseball cards, it will be easier to say which physical states belong to a particular computational state if there is a concise set of physical features that those states share and that other physical states lack.

By 'physically similar states,' I do not mean 'states that are merely similar'; I mean 'states that exhibit no concise set of features that distinguish them from each other'. For example, suppose that I have 100 copies of a particular baseball card, and a collector offers to buy whichever 50 of the cards are in the best condition. I deliver this set of cards, but she promptly asks me to describe my selection criteria. If half of the cards had obvious creases, my answer could be quite brief. However, if all of the cards were in very similar condition, I might have to talk about subtle measures of fading, staining, or fraying of the edges. The point is that, when the physical states belonging to different computational states are very similar, additional features must be introduced or else specified with greater precision in order to distinguish them from one another.

This suggests that mapping simplicity depends on a kind of structural similarity between the instantiating system and instantiated model (in this case, a physical system and model of computation, respectively). In particular, it suggests that mappings will be simpler when the physical states corresponding to each computational state are (i) similar to each other and (ii) different from the physical states corresponding to other computational states. In other words, mappings will be simpler when distinctions drawn by the model reflect patterns of similarity and difference present in the instantiating system.

Fortunately, we can easily generalize the simplicity criterion for computer instantiation to the instantiation of dynamical systems models *tout court*. We can formalize this generalized criterion as follows:

(1) Let M be a dynamical model and P be a physical system. For P to instantiate M, there must be some function, I, that maps the phase space of P to the phase space of M. (*Mapping Constraint*)
(2) For any state of M, m_i, at any time, t_i, if the evolution rule of M requires that m_i evolve to m_j at time t_j, then if P is in some state p_i at t_i such that $I(p_i) = m_i$, then P will evolve to some state p_j at t_j such that $I(p_j) = m_j$. (*Counterfactual Constraint*)
(3) Let $\hat{I}$ be the simplest function that satisfies (1) and (2) for P and M. P instantiates M *well* to the extent that $\hat{I}$ is simple. The simplicity of $\hat{I}$ is construed as the Kolmogorov complexity of $\hat{I}$ or $K(\hat{I})$. (*Simplicity Criterion*)

In other words, for a physical system to instantiate a particular explanatory model, there must be a relatively simple mapping from states of the physical system to states of the model under which the system can be understood to reliably obey the evolution rule of the model. As we will see, it will be necessary to qualify and elaborate this criterion in several ways, but this formulation captures its central proposal.

To see how this criterion addresses a weakness of Dennett's view, consider the aim laid out in section 2—to distinguish between explanatory models that invoke different ways of computing the same function. Consider two finite-state machines that compute the same function over binary strings (such as identifying strings that

contain only zeros) but differ in the number of states they have and in their rules for transitioning between them. These machines compute the same function in different ways. As we saw, Dennett's criterion struggled to discriminate between computational models in these cases because of its focus on predicting the target phenomenon (such as a machine accepting strings of zeros and rejecting strings that contain a '1').

In contrast, the criterion here directly evaluates the structural similarity between the states/state transitions exhibited by the target system and those described by each model. It also retains the original's concern about predictive power, because satisfying the counterfactual constraint (ultimately) implies that the function computed by the model is computed by the system under our mapping. The reason for this is that obeying the transition function of the model is sufficient (but not necessary) for computing the relevant function. The exact details of our mapping would differ from case to case, but the key point here is that the simplicity criterion directly considers those elements of a model relevant to evaluating its account of how a target phenomenon is produced—what it says about the states of a system and about how those states evolve over time.

3.4 Further Considerations

The simplicity criterion proposes that models requiring simpler mappings more faithfully represent the target system than do models requiring more complex mappings, all else being equal. Naturally, a number of other factors might be relevant in determining when all else is equal, and difficult questions remain about how to weigh the factors considered above. For example, we must weigh the relative importance of mapping simplicity and the reliability with which the physical system evolves, as predicted by our model under that mapping. After all, counterfactual reliability is itself a graded notion. Probabilistic models can exhibit different degrees of specificity and different degrees of accuracy with respect to observed distributions. Even deterministic models might predict the correct outcome more or less often. It is also important to consider the simplicity of the model itself. This will probably be an important factor, especially when adjudicating between models that are similarly accurate and similarly reliable.

Identifying all relevant factors, developing conventions for quantifying them, and assessing their relative importance is an important (if demanding) project. My main concern here, however, is to explain the connection between simplicity and similarity and to establish the relevance of this connection to pattern instantiation. As such, I do not think that we should be too committed (yet) to any particular formalization of this criterion, and I am open to any compelling alternatives. Regardless, it will usually suffice to (i) show how a project of precisifying our informal interpretation might proceed, (ii) address any apparent barriers to this project, and (iii) provide reasonably strong reasons to think that the resulting mapping will be substantially simpler for the proposed model than for relevant competitors. Again, the key point of my argument is that there is a genuine connection not only between compression and regularity, but *between compression and similarity*. While one may disagree about the precise formalization of the criterion or the proper weighting of relevant factors, the connection between compression and similarity will remain.

3.5 Realist Desiderata Revisited

Earlier, I argued that a realist account of pattern instantiation must (i) explain scientific scepticism about the accuracy of predictively successful models and (ii) remain flexible enough to accommodate cases of multiple realizability. With respect to the first *desideratum*, I follow prior work in the philosophy of science which observes that scientists are interested in 'whether the structure of the model maps well onto the target system of interest', where 'maps well' is interpreted as an abstract similarity relation [Weisberg 2016: 267]. The simplicity criterion, in turn, quantifies the degree of similarity in specific cases. On this view, a scientist might be sceptical about whether the ontology and dynamics of a predictively successful model are sufficiently similar to the ontology and dynamics of the target physical system.

The criterion is also flexible enough to preserve multiple realizability and reject chauvinism. The criterion does not require that we specify what it means for *any* physical system to instantiate a particular model. Each mapping is specific to the physical system in question and does not require a general theory of what it is, *physically speaking*, to instantiate a particular model. We need only say what it means for *this or that* physical system to instantiate a particular model. This explains how the criterion is consistent with the multiple realizability of patterns and the rejection of chauvinism —that is, by permitting different mappings for systems that differ physically.

3.6 A Case Study: The Blockhead

Before moving on, it will be worthwhile to examine a case where the proposed criterion delivers plausible judgments about model instantiation. It will be particularly helpful to consider a case that supports realist concerns about real patterns. Recall that Dennett [1991] offers real patterns as an ontological account of intentional mental states identified via his *intentional stance* [Dennett 1989]. Block [1981] asks us to imagine that a vast look-up table has been programmed to hold realistic conversations of some arbitrary finite length. Dispensing with the usual trappings of natural language processing, this 'blockhead' substitutes rote memorization for genuine linguistic competence. The blockhead records the conversation up to the present moment, and every possible conversation history (up to some arbitrary length) has an entry in its look-up table. Each entry in the lookup table is matched to a conversationally appropriate response that the blockhead dutifully produces.

While Block intended this case to expose the limitations of behaviourism in psychology, the blockhead also threatens Dennett's real patterns approach to intentional systems. After all, as Dennett argues, 'Anything that is usefully and voluminously predictable from the intentional stance is, by definition, an intentional system' [2009: 339]. *Ex hypothesi*, the blockhead is just such a system. Building on some of Dennett's other work, several philosophers of cognitive science have pursued externalism about mental states as a strategy for addressing these (and other) concerns about the intentional stance. For example, it is far from clear that the blockhead has (or could have) the right kind of causal history to qualify as a true believer. While this is a promising strategy, I am considering the blockhead in the context of concerns about whether a system, *narrowly considered*, genuinely exhibits certain patterns of activity or organization, as represented by a model [D. Wallace 2012; Weisberg 2016]. Fortunately, addressing

these concerns will be complementary to any externalist account that make at least some demands on the internal structure of cognitive systems.

The blockhead is such a compelling case because the blockhead so clearly differs from ordinary humans. The main problem with the blockhead seems to be that its propositional attitudes are more a matter of our exegetical efforts than of *any* belief-like or desire-like internal structure. After all, if we opened up the blockhead and examined the present state of its software, we would gain no insights into its present intentional states beyond those possessed by any attentive interlocutor. We would simply find an accurate record of what both parties have said. Suppose that we wanted to detect deception, resolve a semantic ambiguity, or infer an unstated belief. Nothing new would be revealed by examining the conversation history stored inside the blockhead.

A consequence of this is that if we want to interpret the blockhead as an intentional system, we are forced to deploy our own folk psychological competence to craft this interpretation. In contrast, imagine a highly idealized human that tokens a unique symbol string in her head for every belief/desire that she has, where each string is unique to a belief/desire with a particular propositional content. Examining the internal state of such a human would make attributing plausible beliefs and desires to her a (relatively) trivial matter of parsing strings. These attributions would require little or no competence with intentional stance reasoning.

Returning to the criterion offered earlier, consider the length of the shortest program for computing (for each system) a mapping from its physical states to its intentional states. These state attributions must ultimately facilitate accurate behavioural predictions, and so they must at least be plausible in light of the systems' observed behaviour; random or arbitrary mappings will not do. While the mapping program in both the idealized human case and the blockhead case must have the competence required for detecting and parsing symbol strings, only the latter must solve the problem of interpreting the verbal behaviour of its subject. Recall that the physical states of the idealized human token symbols that correspond directly to her intentional states. The blockhead, on the other hand, tokens symbols representing its actual (and potential) verbal behaviour. Hence, to ascribe predictively useful intentional states to the blockhead on the basis of its physical states requires our mapping program to interpret the blockhead's behaviour in much the same way as any attentive interlocutor.

Obviously, incorporating this folk psychological competence would substantially lengthen our mapping program for the blockhead. This illustrates how program length varies with the amount of outside information that must be supplied to interpret one system as another—in this case, a look-up table as an intentional system. Since there is a straightforward correspondence between the physical states of the human and her intentional states, less information must be brought in to determine her intentional state from her physical state. This is analogous to the formal result discussed earlier—that a program for mapping one string to another (or *vice versa*) must contain at least enough bits to represent any information present in one string but not the other.

In this way, the simplicity criterion confirms our intuitive sense that the 'beliefs' ascribed to the blockhead are more a matter of our exegetical efforts than of any facts about the functional organization of the blockhead. After all, even when we have perfect access to the internal states of the blockhead, our program must replicate

all of the work of someone without such access in order to ascribe beliefs to the blockhead. Actual humans, of course, will not afford as simple a mapping as the idealized case, but the extreme practical difficulties of realizing a blockhead suggest that no actual system would have anything like its 'cognitive' architecture. With that said, this is not the place to settle questions about the status or realization of actual human beliefs, but I think we have every reason to suspect that humans will be decidedly closer to the idealized case than they are to the blockhead—even if they ultimately fail to satisfy the exacting standards of 'industrial strength' realists [Dennett 1991: 30].

4. The Varieties of Instantiation

As framed above, the simplicity constraint considers a mapping between the phase spaces of two dynamical systems models—one representing our target physical system and one being our proposed model of that system. Of course, there are other kinds of models that we might like to evaluate (for example, an intentional systems model), and it would be worth clarifying how the criterion might be applied to a wider range of scientific models and explanatory projects. Fully extending the criterion to different types of models merits a separate discussion, but the basic recipe for carrying out this extension seems clear.

Since explanatory models tell us what there is and how it behaves, a natural way to extend the criterion is to require mappings that involve the ontologies of the instantiating and instantiated models (the 'what' in 'What is there?' and the 'it' in 'How does it behave?') [Maudlin 2015: 64]. For an intentional systems model, this is a set of beliefs and desires. For dynamical systems, this is a space of possible states. The same is true for standard Markov models. For n^{th}-order Markov models, this is state histories of length n. For agent-based models, this is a set of states for agents. Hence, if we wanted to see how well a dynamical system realizes an agent-based model, we might require a mapping from states of the dynamical system to states of the relevant agents under which the states of those agents evolve over time according to the rules of the agent-based model. This flexible approach allows us to target the posits that figure in the evolution rule of a model—whatever those posits might be.

With that said, D. Wallace [2012] offers compelling reasons to think that we are often interested in assessing models more abstractly. Even the notion of 'structural similarity' (discussed above) does not specify at which level of abstraction similarity should be assessed. For example, when asking whether a physical system instantiates a computer, we might actually want to know how the physical states of a machine correspond to the computational states of a computer. However, in assessing whether a physical system realizes an economic model, our standards may differ. Such a model might accurately predict economic trends, even if it proposes idealized agents that have no direct correspondence (either numerically or qualitatively) with agents in the world. In light of cases like these, Wallace argues that we should assess model instantiation *at the level of histories*.

More specifically, Wallace's criterion requires a mapping from histories in the target system to histories in the model [ibid.: 54]. This criterion raises the same instrumentalist worries as Dennett's because it does not say which features of these histories should form the basis of our mapping. For example, the low-level histories of a computer will include details relevant to both (i) the input/output behaviour of an algorithm and (ii) the internal state transitions that accomplish this behaviour. However,

Wallace's criterion does not require us to construct a mapping that is actually sensitive to (ii). Wallace emphasizes this consequence himself, but regards it as essential to accommodating different standards for model instantiation (such as those at play in the instantiation of computation and economic models, respectively) [ibid.: 56–7]).

To be clear, I entirely agree with Wallace that we may often be satisfied by less restrictive standards of instantiation, but there is no reason to endorse a one-size-fits-all account in order to accommodate these cases. Earlier, I suggested that our mappings should consider the ontology of the instantiating system and the ontology of the instantiated model. This suggestion can be understood as the maximally realist approach to assessing instantiation in any given case. Of course, nothing requires that we insist on the most realist approach irrespective of context. This stringent realism can be relaxed by allowing our mappings to consider *abstract* features of the model's ontology.

Consider a model of population ecology. Such a model might treat a population as a collection of individuals, but, like an economic model, there might be no simple correspondence between individuals in the target environment and individuals in the model. With that said, the high-level variable *total population* depends on the individuals present in the population, but it is also invariant with respect to their individual fates. If one animal dies and another is born, the population remains the same. If we are interested in how faithfully a model represents changes in an animal population over time, we might consider the simplicity of a mapping between states of the target environment and total population values in the model. If our mapping is simple and counterfactual-supporting, then we might say that the target environment realizes the model *at a high level* (the population level). We might also say that the model bears a high-level similarity to the target environment. On the contrary, if our mapping requires a large number of caveats and qualifications, we would doubt that the model bears even a high-level similarity to the target environment.

Of course, adopting a paradigmatically realist approach to assessing instantiation does not settle the question of whether an instantiation is genuine. It is merely the decision to evaluate the instantiation of a model at the ground level (that is, in terms of the things that figure directly in its ontology and dynamics). Higher-level approaches are certainly permitted, but, even in the best case of mapping simplicity and counterfactual reliability, we cannot say that the low-level postulates of the model have been vindicated; only the high-level postulates have been. The level at which we construct our mapping is the level at which the simplicity criterion assesses similarity. The higher this level, the more we are concerned with high-level similarity. The lower this level, the more we are concerned with low-level similarity and (ultimately) the details of implementation. As such, if we are primarily interested in high-level similarity (as in the cases discussed above), a high-level approach is entirely justified.[8]

5. Conclusion

In this paper, I have defended a novel criterion for the reality of patterns. This criterion is inspired by real patterns but resolves several serious concerns about the original's

[8] In some cases, we might be interested in assessing only some aspects of our model (e.g. what it says about particular high-level variables). To do so, we might consider a mapping between the states of a physical system and a proper subset of the variables employed by our model.

commitment to realism. It does so by combining insights about compression and similarity with the idea that model fidelity amounts to a kind of structural similarity between a model and its target system. On my view, structural similarity is assessed by considering a mapping between the physical states of the system and the states of the system as represented by our model. To the extent that there exists a simple counterfactual-supporting mapping of this kind, our model and the pattern that it describes is well-instantiated. Of course, many important questions remain to be answered. Not the least of these asks how to balance other *desiderata* (such as model simplicity) with mapping simplicity and counterfactual support. Nevertheless, I hope that this paper is a constructive step towards a satisfactory criterion for the reality of patterns.[9]

Disclosure Statement

No potential conflict of interest was reported by the author.

References

Andersen, H.K. 2017. Patterns, Information, and Causation, *The Journal of Philosophy* 114/11: 592–622.
Bennett, C.H., P. Gács, M. Li, P.M. Vitányi, and W.H. Zurek 1998. Information Distance, *IEEE Transactions on Information Theory* 44/4: 1407–23.
Block, N. 1981. Psychologism and Behaviorism, *The Philosophical Review* 90/1: 5–43.
Burnston, D. 2017. Real Patterns in Biological Explanation, *Philosophy of Science* 84/5: 879–91.
Cormen, T.H., C.E. Leiserson, R.L. Rivest, and C. Stein 2009. *Introduction to Algorithms, 3rd edn*, Cambridge, MA: The MIT Press.
Chalmers, D.J. 1996. Does a Rock Implement Every Finite-State Automaton? *Synthese* 108/3: 309–33.
Dennett, D.C. 1989. *The Intentional Stance*, Cambridge, MA: The MIT Press.
Dennett, D.C. 1991. Real Patterns, *The Journal of Philosophy* 88/1: 27–51.
Dennett, D.C. 2009. Intentional Systems Theory, in *The Oxford Handbook of Philosophy of Mind*, ed. B. McLaughlin, A. Beckermann, and S. Walter, Oxford: Clarendon Press: 339–49.
Giere, R. 1988. *Explaining Science: A Cognitive Approach*, Chicago: University of Chicago Press.
Godfrey-Smith, P. 2006. The Strategy of Model-Based Science. *Biology and Philosophy* 21/5: 725–40.
Grünwald, P. and P.M. Vitányi 2004. Shannon Information and Kolmogorov Complexity, URL = https://arxiv.org/abs/cs/0410002
Ladyman, J. and D. Ross 2007. *Every Thing Must Go: Metaphysics Naturalized*, Oxford: Clarendon Press.
Ladyman, J. and D. Ross 2013. The World in the Data, in *Scientific Metaphysics*, ed. D. Ross, J. Ladyman, and H. Kincaid, Oxford: Oxford University Press: 108–50.
Levenshtein, V.I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Soviet Physics Doklady* 10/8: 707–10.
Li, M. and P.M. Vitányi 2008. *An Introduction to Kolmogorov Complexity and its Applications, 3rd edn*, New York: Springer.
Maddy, P. 1992. Indispensability and Practice, *The Journal of Philosophy*, 89/6: 275–89.
Maudlin, T. 2015. Physics, Philosophy, and the Nature of Reality, *Annals of the New York Academy of Sciences* 1361/1: 63–8.
Millhouse, T. 2019. A Simplicity Criterion for Physical Computation, *The British Journal for the Philosophy of Science* 70/1: 153–78.
Millhouse, T. 2021. Compressibility and the Reality of Patterns, *Philosophy of Science* 88/1: 22–43.
Pylyshyn, Z.W. 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, MA: The MIT Press.

[9] I am indebted to Shaun Nichols, Daniel Dennett, Jenann Ismael, Richard Healey, David Wallace, Kathleen Creel, Brandon Ashby, Nathaniel Oakes, Jeri Millhouse, and Amanda Romaine for their comments, advice, and encouragement. Any remaining errors or omissions are entirely my own.

Rathmanner, S. and M. Hutter 2011. A Philosophical Treatise of Universal Induction, *Entropy* 13/6: 1076–136.
Ross, D. 1995. Real Patterns and the Ontological Foundations of Microeconomics, *Economics and Philosophy* 11/1: 113–36.
Ross, D. and D. Spurrett 2004. What to Say to a Skeptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists, *Behavioral and Brain Sciences* 27/5: 603–64.
Shannon, C.E. 1948. A Mathematical Theory of Communication, *The Bell System Technical Journal* 27/3: 379–423.
Sipser, M. 2013. *Introduction to the Theory of Computation, 3rd edn*, Boston: Cengage Learning.
Wagner, R.A. and M.J. Fischer 1974. The String-to-String Correction Problem, *Journal of the ACM*, 21/1: 168–73.
Wallace, C.S. 2005. *Statistical and Inductive Inference by Minimum Message Length*, New York: Springer.
Wallace, D. 2012. *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*, Oxford: Oxford University Press.
Weisberg, M. 2016. Modeling, in *The Oxford Handbook of Philosophical Methodology*, ed. H. Cappelen, T. Szabó Gendler, and J. Hawthorne, Oxford: Oxford University Press: 262–84.