Synthese (2024) 203:173
https://doi.org/10.1007/s11229-024-04524-z

ORIGINAL RESEARCH

Does the no miracles argument apply to AI?

Darrell P. Rowbottom[1] · William Peden[1] · André Curtis-Trudel[1]

Received: 9 September 2023 / Accepted: 8 February 2024

Abstract

According to the standard no miracles argument, science's predictive success is best explained by the approximate truth of its theories. In contemporary science, however, machine learning systems, such as AlphaFold2, are also remarkably predictively successful. Thus, we might ask what best explains such successes. Might these AIs accurately represent critical aspects of their targets in the world? And if so, does a variant of the no miracles argument apply to these AIs? We argue for an affirmative answer to these questions. We conclude that if the standard no miracles argument is sound, an AI-specific no miracles argument is also sound.

Keywords Artificial intelligence · Scientific realism · Machine learning · Scientific progress · Scientific representation · AlphaFold

1 Introduction and context

Machine learning (ML) techniques are now a standard part of scientists' toolkits in many areas. However, the models constructed using ML are data-driven: they are constructed using statistical methods—embodied in an ML algorithm—applied to "training" data. This contrasts with traditional scientific modelling, where explicit theories and models are used to construct equations to represent systems (Knüsel & Baumberger, 2020: p. 47). So do the empirical successes of AI systems using ML indicate that they contain the approximate truth concerning their targets?

✉ Darrell P. Rowbottom
darrellrowbottom@ln.edu.hk

William Peden
williampeden@ln.edu.hk

André Curtis-Trudel
andre.curtistrudel@ln.edu.hk

[1] Department of Philosophy, Lingnan University, 8 Castle Peak Road, Tuen Mun, Hong Kong

Published online: 13 May 2024

If so, there would be important implications. First, AI methods are employed in local scientific scenarios where scientists have instrumentalist-friendly goals: predicting phenomena in an accurate, simple, and computationally tractable way. If a realist attitude is nonetheless appropriate in this context, it would be a striking example of the distinction between what a scientist *aims* to do and what their practice cognitively *achieves*. This distinction is a significant part of the scientific realism debates, and concerns related issues such as the epistemic role of theoretical virtues in scientific practice. Second, the opacity of how AI systems learn and infer is of great contemporary concern (Biddle, 2020; Creel, 2020; Sullivan, 2022). If the empirical successes of such systems are indicative of 'hidden' accurate hypotheses or models therein, then that would be a reason to pursue a way to access or (even partially) grasp them.

Furthermore, scientific realism is a popular view, and many realists endorse the NMA (in variants as presented by Dawid & Hartmann, 2018; Fahrbach, 2011; Henderson, 2017; Menke, 2014; and Sprenger, 2016). But if one endorses this, should one also commit to the view that empirically successful AI systems contain accurate representations of their targets? To what extent do the two views stand or fall together?[1]

One reason for thinking that many scientific realists will find an AI-based NMA appealing is that scientists often talk about AI systems accurately representing various aspects of the world. Consider AlphaFold 2 (AF2), which predicts the structure of proteins on the basis of their amino acid sequences. Rubiera (2021) writes that its:

> Evoformer … is able to efficiently extract information from a multiple sequence alignment and build an accurate representation of the parts of the protein in close contact …[2]

Scientific realists of a traditional stripe (as distinct from selective realists)—and hence, advocates of the NMA—are sympathetic to taking such claims at 'face value'. That's because scientific realism typically involves theses such as 'claims about scientific objects, events, processes, properties, and relations … whether they be observable or unobservable, should be construed *literally as having truth values, whether true or false*' (Chakravartty 2017) and 'the ampliative-abductive methods employed by scientists to arrive at their theoretical beliefs are reliable: they tend to generate approximately true beliefs' (Psillos, 1999: p. xxi). Hence, realists will tend to think that Rubiera's claims about representations in AF2 are approximately true.

The paper proceeds as follows. In the next section, we present the two main versions of the standard NMA and illustrate how one of these is defective. As a result, we consider only the surviving frequency-based variant in the remainder of the paper. In section three, we identify and address three objections to the notion that the standard frequency-based variant of NMA may be extended to AI systems. We then present our AI-based NMA; we do not argue that it is sound, but only that it is as good as the standard NMA. In the penultimate section, we illustrate how this NMA works for AF2. We finish by presenting corollaries of our findings.

[1] Scientific realists who do not think the NMA is a good argument in the normal case are liable to think the same in the AI case.

[2] As we will later explain, the evoformer is only the central part of AF2. Jumper et al. (2021) describe this not only as having two kinds of representations—MSA and pair—as inputs and outputs, but also as exchanging information between these different kinds of representation.

2 The standard no miracles argument(s)

The NMA appears in a short passage by Putnam (1975) and is rooted in the work of Smart (1968). The rudimentary idea is that the empirical successes of science would be miraculous if scientific realism were not true. However, the original NMA is imprecisely formulated. As a result, it has been understood in various ways.

The vagueness has several facets. First, 'scientific realism' is a label for a cluster of views, some of which differ significantly from others. For instance, according to van Fraassen (1980) scientific realism concerns the aim of science. In contrast, it has an epistemic core according to Psillos (1999).[3] Second, 'empirical success' may be construed in several ways. For some, like Musgrave (1988), only novel predictions contribute to success, whereas for others, such as Keynes (1921: pp. 305–306), evidence can be provided by accommodations of old data.[4] Third, the NMA's scope is a matter of interpretation; for instance, different variants of the NMA pertain to isolated theories, larger groups of theories, and scientific theories as a whole.

We will do three things to limit the need to go into detail concerning these matters. First, we will restrict our attention to a claim which underpins all extant versions of the NMA:

> (E) Empirical successes in science enabled by scientific theories are non-miraculous because such theories are typically or probably approximately true.[5]

Second, we will only consider cases involving numerous successful novel predictions (i.e., where the presence of empirical success is uncontroversial). Third, we will not consider NMAs from an individual theory's empirical success to its probable approximate truth.

This third decision is motivated by the fact that NMAs concerning individual theories have a fatal flaw when expressed in a Bayesian framework: they commit the base rate fallacy (Howson, 2013; Magnus and Callender, 2004). It is easy to show this. Let S represent 'T is empirically successful' and ≈T represent 'T is approximately true'. The individual-theory variant of the NMA is:

(a) $P(S|\approx T)$ is high.
(b) $P(S|\neg\approx T)$ is low.
Therefore, (c) $P(\approx T|S)$ is high.[6]

[3] Such differences exist between almost every account, although they are often more subtle. See Rowbottom (2019a; 2019b: appendix) for a comparison of the views of Boyd, Musgrave, Psillos, van Fraassen, and others.

[4] One way of encapsulating these alternative views is to link high confirmation value with (probable) approximate truth. The underlying disagreement then concerns how theories come to be confirmed. On the history of this debate, see Musgrave (1977), Douglas and Magnus (2013), and Barnes (2022). The controversy continues; see, for instance, Dellsén (Forthcoming).

[5] This claim is used to support 'the success-to-truth inference', according to which 'where science is sufficiently successful—makes accurate predictions and/or exhibits significant explanatory power—the relevant theoretical hypotheses are (probably and/or approximately) true' (Vickers 2019: p. 571). As Vickers notes, this is at the heart of scientific realism.

[6] What exactly counts as high or low is open to some interpretation, in so far as an informal argument is being formalised. However, it is clear, for instance, that a high probability is greater than 0.5. It is also plausible that it is greater than 0.75.

However, Bayes's theorem entails that:

$$P(\approx T|S) = \frac{P(S|\approx T)P(\approx T)}{P(S|\approx T)P(\approx T) + P(S|\neg\approx T)P(\neg\approx T)} \tag{1}$$

So the argument is invalid because the premises are silent on P(≈T), which must be above a significant threshold for (c) to be true when (a) and (b) are true. Moreover, to add this assumption would generate a *petitio principii*, because anti-realists do not accept it.

However, an NMA with greater scope need not have the same flaw, as Menke (2014) and Henderson (2017) note. Dawid and Hartmann (2018) show this with a formal treatment, which we summarise below. We follow their approach for illustrative purposes, as it elegantly makes manifest the means by which empirical data from past science is relevant in such expanded NMAs. We acknowledge that some advocates of the NMA object to this Bayesian variant.[7] But the obstacles to extending the NMA to AI systems that we discuss subsequently go for all extant NMA variants. Indeed, our discussion would remain pertinent even if the individual-theory NMA were to be reincarnated.

Consider a domain of theories, only some of which have been accurately classified as empirically successful or unsuccessful. For example, the domain might be the theories of mechanics generated by scientists in the actual world; the proper subset of classified theories would include most current and historical mechanical theories, but no future and unconceived theories. Let R be 'The relative frequency of empirically successful theories, classified in the domain, is *r*'. R is evidence for the relative frequency of empirically successful theories in the entire domain. Moreover, R is empirically determinable in principle. Analogously, finding a 0.8 relative frequency of heads results, after flipping a coin numerous times, would lead one to conclude that the flipping process was biased towards a heads result.

The frequency-based variant of NMA, which incorporates R, is:

(i) P(S|≈T&R) is high.
(ii) P(S|¬≈T&R) is low.
Therefore, (iii) P(≈T|S&R) is high.

Bayes's theorem relates (i), (ii), and (iii) as follows:

$$P(\approx T|S\&R) = \frac{P(S|\approx T\&R)P(\approx T|R)}{P(S|\approx T\&R)P(\approx T|R) + P(S|\neg\approx T\&R)P(\neg\approx T|R)} \tag{2}$$

[7] Psillos (2009: p. 195) does so, in writing: 'Bayesian reasoning does not have rules of acceptance. On a strict Bayesian approach, we can never detach the probability of the conclusion of a probabilistic argument, no matter how high this probability might be. So, strictly speaking, we are never licensed to accept a hypothesis on the basis of the evidence.' We won't delve into this issue. Middle ground is possible, however. For instance, one could add acceptance rules to the probabilistic framework presented here.

P(≈T|R), which appears in place of P(≈T) in the individual-theory NMA, must exceed a significant threshold for (iii) to be true when (i) and (ii) are true.[8] But (i) and (ii) are not silent on the value of P(≈T|R), as we now show.

It follows from the axioms of probability that:

$$\mathrm{P(S|R) = P(S| \approx T\&R)P(\approx T|R) + P(S|\neg\approx T\&R)P(\neg \approx T|R)} \tag{3}$$

And from this and $\mathrm{P(\approx T|R) + P(\neg\approx T|R) = 1}$, it follows that:

$$\mathrm{P(S|R) = P(S|\approx T\&R)P(\approx T|R) + P(S|\neg\approx T\&R)\{1 - P(\approx T|R)\}}$$

$$\mathrm{= P(S|\approx T\&R)P(\approx T|R) - P(S|\neg\approx T\&R)P(\approx T|R) + P(S|\neg\approx T\&R)}$$

$$\mathrm{= P(\approx T|R)\{P(S|\approx T\&R) - P(S|\neg\approx T\&R)\} + P(S|\neg\approx T\&R)}$$

Therefore,

$$\mathrm{P(\approx T|R) = \frac{P(S|R) - P(S|\neg\approx T\&R)}{P(S|\approx T\&R) - P(S|\neg\approx T\&R)}} \tag{4}$$

This entails that P(≈T|R) is highly sensitive to P(S|R), given our initial assumptions. Substituting (i) and (ii) into (4), we find:

$$\mathrm{P(\approx T|R) = \frac{P(S|R) - Low}{High - Low}} \tag{5}$$

P(S|R) can be empirically determined, given some reasonable additional assumptions. For instance, imagine the theory under consideration, T, is randomly picked from all theories in the pertinent domain. Then it is natural to think that P(S|R) is the relative frequency of empirically successful theories, r. The crucial point is that scientists can gather evidence about the relative frequency of successful theories, which bears on whether (iii) given (i) and (ii).[9] This possibility is not present in the individual-theory NMA.

Dawid and Hartmann (2018: p. §7) also discuss the value that P(S|R) must take for the NMA to be plausible. For reasons of economy, we will not reconstruct their general result. However, here is a numerical illustration of conditions under which the

[8] Assuming (i) and (ii) and (2):

$$\mathrm{P(\approx T|S\&R) = \frac{High \times P(\approx T|R)}{High \times P(\approx T|R) + Low \times \{1 - P(\approx T|R)\}}}$$

[9] Although a Bayesian treatment of this step is possible, we follow Dawid and Hartmann (2018) by simplifying via a non-Bayesian estimation of P(S|R). A Bayesian treatment would involve conditioning on each theory individually. With suitable priors, the step would still go through, but it would be complex and given a sufficiently large sample of theories it would not significantly differ from estimating from r.

NMA is sound. In line with (i) and (ii), let P(S|≈T&R) be 0.9 and P(S|¬≈T&R) be 0.1. Then, from (2),

$$P(\approx T|S\&R) = \frac{0.9 \times P(\approx T|R)}{0.9 \times P(\approx T|R) + 0.1 \times \{1 - P(\approx T|R)\}} \tag{6}$$

Now let P(S|R) = 0.3. Then, from (4),

$$P(\approx T|R) = \frac{0.3 - 0.1}{0.9 - 0.1} = 0.25 \tag{7}$$

Finally, by substitution from (7) into (6),

$$P(\approx T|S\&R) = \frac{0.9 \times 0.25}{0.9 \times 0.25 + 0.1 \times 0.75} = 0.75 \tag{8}$$

Thus, P(S|R) need not be high for the NMA to be sound, despite any first appearances to the contrary. Although P(S|R) is merely 0.3, (iii) holds in our illustration.

We should mention one further ambiguity concerning the NMA's scope before we continue, which we have already addressed implicitly in the formal treatment. The NMA is not concerned with explaining *how* we obtain successful theories. It is concerned with explaining *why* successful theories survive the selection process. So van Fraassen's (1980: pp. 39–40) 'Darwinist' hypothesis that 'only the successful theories survive' is not a competing proposal to ascribing approximate truth to those theories. Analogously, in the words of Leplin (1997: p. 9):

> To explain why the *Wimbledon finalists* are so great, it is perfectly appropriate to cite the stringency of the selection procedures for entry into the tournament ... It is hardly surprising that the finalists are great players, considering what they had to go through to get there. However, none of this explains why *these particular individuals*, who happen to be the finalists ... are so great. On the contrary, it is their being great that explains their having managed to survive the rigors of selection...[10]

To amplify this point, we might imagine players who could become Wimbledon finalists but do not enter the tournament for political reasons. What explains their capacity for success? Our answer might mention athleticism, technique, decision-making, composure, focus, and so forth. Analogously, some unconceived theories would be empirically successful if we discovered them. NMA-style arguments posit properties of these theories, like accuracy, consistency, simplicity, and truth.

This concludes our presentation of the standard frequency-based NMA. *We will not argue or even assume that it is sound*. Rather, we will argue that a highly similar NMA, which is just as plausible, pertains to a significant class of AI systems. Hence, *if* one believes that the standard NMA is sound, *then* one should also think that there is a

[10] As a similar criticism of van Fraassen's proposal, Kitcher (1993: p. 156) writes: 'Darwinists want to know … what kind of organism-environment relationships confer reproductive success.' Van Fraassen's evolutionary account of theories' empirical success fails to provide an analogous explanation for empirical success.

sound NMA involving some AI systems. Or so we contend. This might not be a boon for scientific realism because the AI-based NMA could be the basis for an anti-realist *reductio ad absurdum* of the standard NMA.

In the next section, we consider potential disanalogies between theories and AI systems, and address these in order to develop an AI-specific NMA. In the section thereafter, we discuss how this argument applies to AlphaFold2 (AF2), which is a striking example of an empirically successful ML system.

3 Developing an AI-specific NMA

There are three main reasons why the standard frequency-based NMA might not straightforwardly apply to AI systems, such as ML networks like AF2. First, it is dubious that truth-bearers—and hence, bearers of approximate truth—are present in such systems. No linguistic entities are stored therein, for instance, despite some aspects of the AF2 programme being motivated by putative physical laws.[11] Second, it is questionable whether an AI can contain representations, because scientific representation is often taken to be intentional. Third, even assuming representations are present, no theories are straightforwardly identifiable in these systems.

We will now deal with each of these issues in turn. First, we will argue that accuracy may serve as a substitute for approximate truth. Second, we will show that AIs may represent aspects of the world in the same way that human brains do, on a naturalistic theory of representation. Third, we will argue that an AI-specific NMA may cover more than just theories. It can cover all the representative machinery used to generate outputs from inputs.

3.1 Accuracy as a substitute for approximate truth

Because theories are linguistically expressible, they are evidently capable of being true or false, and approximately true by extension. However, the information inside a predictively successful AI, such as AF2, is non-linguistic. So how could an NMA apply to such a system? In answering this, we will assume a non-epistemic theory of truth, such as a correspondence theory, which accords with scientific realism.

First, a proposition is true precisely when it *faithfully represents* something, such as a state of affairs. As Psillos (2004: p. 143) puts it: ‘Truth gives us purchase on the world. It connects our thoughts and beliefs to some external reality, thereby giving them [faithful] representational content.’ Or as Jubien (2001: p. 50) puts it: ‘Propositions represent the world as being one way or another. If they did not represent in this way, it would be utterly implausible to view them as the ultimate bearers of truth values.’ Thus, truth is just a *special case* of faithful representation, which is typically understood to

[11] Rowbottom (2022) argues that some non-sentential/non-propositional entities are approximately true. However, his argument pertains only to entities that have sentential/propositional parts (and which can be converted into sentential/propositional form by simple operations, such as deletion of some parts).

apply to propositions, and derivatively entities that contain or express them (e.g., sentences and beliefs). Or, at least, this is the dominant view of propositions.[12]

Second, although it often passes unremarked on, the semantic view of theories, which is now prevalent, precludes theories from being true, or approximately so, in any straightforward linguistic sense. Indeed, as Ruyant (2020: p. 7966) notes:

> There is a straightforward tension between semantic realism and the semantic conception of theories, insofar as one of the main purposes of the latter was to acknowledge that scientific representation is not linguistic and thus to get rid of problematic issues falling under the scope of philosophy of language. A model, contrarily to a linguistic statement, is not generally said to be true or false: instead it is said to be good or bad, or accurate or inaccurate.

Strictly speaking, then, 'approximate truth' should not appear in the standard NMA (in so far as this is not intended to preclude a semantic conception of theories). However, philosophers of science often use 'approximate truth' broadly, and explicitly take it not to presuppose a syntactic view of theories. For example, Chakravartty (2010: p. 49) writes:

> [I]n the sciences, approximate truth is best understood as a virtue multiply realized by means of different kinds of representational relationships between scientific products such as theories and models on the one hand, and target systems in the world on the other.

We could follow Chakravartty's lead.[13] We prefer, however, to use 'representational accuracy' in developing an AI-specific NMA. The primary reason, as noted above, is that representational accuracy is a more general notion than truth.

Accuracy comes in degrees. For example, the damped periodically-forced pendulum is a more accurate representation of the longcase clock's driving mechanism than the simple pendulum. So when we write of 'accurate representation', we mean representation with a high degree of accuracy. We leave open exactly what threshold is involved; we simply allow that it falls somewhat short of completely faithful (or veridical) representation. Analogously, 'approximate truth' reflects a degree of partial truth that passes a threshold falling short of truth (about the pertinent subject matter). But most users of 'approximate truth', including advocates of the standard NMA, do not specify a threshold.

We hold that AIs can contain propositions, in a particular mode of presentation, and hence approximately true components. But we also think that AIs can contain other kinds of accurate representation. Thus, we hold that the representations in AIs can be accurate overall, despite only a proper subset of those representations being propositional. We will say more about this in the next two subsections.[14] For the moment, however, note that propositions need not be encoded or expressed in a *natural* language. Indeed, Floridi (2005: p. 366) writes, in developing his account of information,

[12] See Brown (2021), who does not share this view, for further discussion.

[13] However, the syntactic view of theories may also have been unfairly dismissed, as argued by Lutz (2014; 2017).

[14] See also Rowbottom, Curtis-Trudel and Peden (2023) for our stance on how such representations can constitute evidence.

that: ‘It is preferable to speak of “truthful data” rather than “true data” because the data in question may not be linguistic (a map, for example, is truthful rather than true)…’.

The analogy with a map is useful. A map is not a typical linguistic token like an utterance or an inscribed sentence on a page. Yet a map can provide information; a map has propositional content. Thus, insofar as this propositional content is true, the map can be truthful, even though the map as such is not true or false. It might show, for instance, that ‘The Nile is 4100 miles long’. More generally, the map has accuracy conditions: situations under which, for a given interpretation, it provides an accurate representation of some feature it depicts. Thus, a map may encode true and approximately true propositions without (on the standard view of truth-bearers) having truth conditions.

3.2 Accurate representation in AI systems

This brings us to the pressing question of whether AIs—e.g., ML networks such as AF2—can create, embody, or employ scientific representations. Several influential accounts of scientific representation explicitly rule out this possibility. For example, Giere (2010: p. 269) advocates ‘an intentional conception of representation in science that requires bringing scientific agents and their intentions into the picture’ and Suárez (2004: p. 773) defends an inferential conception of representation on which ‘A represents B only if … A allows competent and informed agents to draw specific inferences regarding B’.[15] Some philosophers have also explicitly ruled out the possibility of representation in AI systems on the grounds that such systems lack the requisite intentions. For instance, Boge (2021: p. 50) argues that, absent a human interpreter, talk of internal representations in an ML network is at best metaphorical.

It is illuminating, however, to consider why Giere (2010) appeals to scientists’ intentions. He writes:

> [R]epresentation with models cannot just be a matter of similarity between a model and the thing modeled. There are two major reasons why this is so. First, we need to know which similarities matter. That there will always be some similarities is vacuously true. Second … similarity is a symmetrical relation while representation is asymmetrical … If we add the intensions [sic] of an agent or agents, both of these problems disappear … agents specify which similarities are intended, and for what purpose. (Giere, 2010: p. 274)[16]

For argument’s sake, accept that models depend, for their empirical (and other) successes, on being similar to their targets in some respects. Accept that if X represents Y, then some elements of X are similar to some elements of Y. Must we introduce agents and their intentions to specify when an X comes to represent a Y? The answer is no. We only need *intentionality*, or for X to be *about* Y. We do not need intentions.

[15] As Schlosser (2019) notes, ‘Usually … the term “agency” is used … to denote the performance of intentional actions’. So ‘agents’ usually refers to entities that can perform such actions. AIs cannot.

[16] See also Giere (2004: p. 747): ‘Anything is similar to anything else in countless respects, but not anything represents anything else. It is not the model that is doing the representing; it is the scientist using the model who is doing the representing’.

When X is about Y, it does not follow that Y is about X; asymmetry holds. Moreover, when X is about a part of Y, it does not follow that X is about all of Y; some similarities between X and Y may not 'matter'. Hence introducing agency is unnecessary to address Giere's concerns, provided that intentionality is possible without agency.

We will shortly proceed to show that intentionality (and thus representation) is possible without agency, on the dominant naturalistic view concerning mental representations (or mental content).[17] But before we do so, we would emphasise that many of the key ideas behind agent-based accounts of scientific representation may continue to hold when agents are removed from the picture. For example, one need not deny that 'There is no representation except in the sense that some things are used, made, or taken, to represent some things as thus or so' (van Fraassen, 2008: p. 23). One might hold that AIs, such as ML networks, can make and use 'some things' to represent other things, just as human scientists can.

Consider also the following passage from Suárez (2004: p. 778):

> [S]cientific representation is, unlike linguistic reference, not a matter of arbitrary stipulation by an agent, but requires the correct application of functional cognitive powers (valid reasoning) by means that are objectively appropriate for the tasks at hand (i.e., by models that are inferentially suited to their targets).

By substituting 'computational powers' for 'cognitive powers (valid reasoning)', space is made for AIs to represent without precluding human agents from intentionally doing the same. According to the prevalent computational theory of mind, human brains may form sub-personal representations in the same way.

This brings us to naturalistic theories of mental content, which have three main strands: causal (Dretske, 1981; Fodor, 1987, 1990; Usher, 2001), structuralist (Block, 1986; Cummins, 1996) and teleosemantic/functional (Millikan, 1984, 1989; Papineau, 1987; Neander, 1991). Some contemporary theories of content, such as Shea's (2018) 'varitel' semantics, involve all three elements. We will not explore the differences between these approaches or contrast them with non-naturalistic approaches that introduce intentions. Rather, we will illustrate that these naturalistic theories are compatible with ML networks like AF2 representing aspects of the world. For present purposes, this strategy is appropriate because these theories are dominant in psychology and philosophy of mind, and because we are exploring the prospects for an extension to the standard NMA, which is a realist argument. Realists will tend to take empirically successful contemporary theories in psychology to be approximately true, provided they endorse the standard NMA. Moreover, many prominent scientific realists—such as Boyd (1980), Psillos (1999), and Papineau (2010)—endorse naturalism.

Consider a system that contains representations *ex hypothesi*, such as a human brain. On the naturalistic view of content, it is possible, in principle, to provide a complete causal account of how that system behaves—of how its inputs lead to its outputs—without appeal to semantic properties. Thus, content is not strictly needed to predict how the system will behave or to explain how it has behaved, independently of its environment. However, ascribing content does explain how that system *interacts*

[17] Following our earlier discussion of propositions, 'A state with content is a state that *represents* some part or aspect of the world; its content is the way it represents the world as being' (Brown 2022).

with its environment, and hence helps to predict how it will respond to changes in that environment. As Shea (2013: pp. 498, 499) explains, the interactions of a system depend on both the system's environment *and* its internal algorithms. For example, covariance of the system with its environment that is purely a matter of chance is not the sort of input–output relationship that explains the interactions of the system with its environment. Instead, to explain these interactions, we need to refer to the internal algorithms, and hence the inner semantic contents in the system.[18]

Take AF2, which we will look at in greater depth later, as an illustration. It is *designed* to perform a specific task: to determine the structure of proteins from their amino acid sequences. It is supplied with information relevant to performing that task, from scientists' past successes in making traditional template-based predictions of protein structure. AF2 has subsequently *learned*, on the basis of this information and trial and error, to perform its task better. For instance, it has learned to disregard some protein templates. Thus, it is reasonable to conclude that it contains representations of protein shapes and so forth, because its interactions with its environment would be mysterious if it lacked this content. We might also go one level 'deeper', in so far as different modules in AF2 have specific tasks. For example, the Evoformer determines which elements of the target protein are effectively 'in contact', given its embedding, and this is independent of the Evoformer being a part of AF2. To see this, we need only note that a human scientist could use its outputs to delimit the possible structures of a protein, especially in simple cases. It is natural to add that if the representations provided by the Evoformer were systematically inaccurate, this would spell trouble for the predictions of protein structure by AF2.

Note that in the neural network components of AF2, any representations must be construed as distributed; none of the nodes can 'code' for any particular symbol, so the representation is sub-symbolic. But information processing in the brain is plausibly the same, as it is a non-artificial neural network.[19]

3.3 Theories versus laws, models and other representative machinery

The standard NMA concerns empirically successful theories. However, it is well-known that empirical success does not consist in making correct predictions from an isolated theory. As Duhem (1954: p. 183) stated, 'an experiment…can never condemn an isolated hypothesis but only a whole theoretical group… [so] a "crucial experiment" is impossible'.[20] Thus, it is essential to consider how empirical success *actually* comes about. We shall argue that pen-and-paper cases of prediction are sufficiently similar to those involving AI for an AI-based NMA to be unproblematic.

[18] Sebastián (2021) also makes the case that AIs can represent in a similar way.

[19] See Tamir & Shech (2023) and Duede (2023) for recent discussion of representation in AI systems. The hypothesis that information processing in the brain is achieved by a non-artificial neural network was first proposed by McCulloch and Pitts (1943). It continues to be a prominent position in cognitive science (Buckner and Garson 2019). This hypothesis potentially has some consequences for the idea that there is a language of thought ('mentalese')—and a correlate in the case of AF2—but this matter remains unsettled. See, for instance, Smolensky (1990), Chalmers (1990), Fodor (1997), and Shea (2007).

[20] Duhem restricted his thesis to physics, but philosophers have typically extended its scope to all of science.

Le Verrier's work on planetary motion provides an excellent illustration of how complicated prediction (or retrodiction) can be. In his second memoir of 1846, he 'demonstrated… a formal incompatibility between the observations of Uranus and the hypothesis that this planet is subject only to the actions of the sun and of other planets acting in accordance with the principle of universal gravitation' (quoted by Hanson, 1962: p. 361). But how did he do this? At his disposal, he had numerous observations of the paths of the known planets, estimates of the masses of those planets and the sun, Newton's law of gravitation, and the laws of classical mechanics. However, this was insufficient to generate predictions. First, Le Verrier had to infer probable planetary paths from the observations or data points concerning positions at various times. Second, he needed to use an abstract model of the solar system, which replaced planets with idealised entities, such as point masses. Third, he had to apply various mathematical approximations—concerning, for instance, perturbations—to 'animate' that model according to the pertinent laws.

Imagine momentarily, for convenience, that Le Verrier's prediction at this juncture was novel and true; imagine he 'saved' the orbit of Uranus for the first time. (We will dispense with this artifice shortly.) Let's now pause to consider what would have been empirically successful *in a sense relevant to the standard NMA*, and how so, in this case. Would it have been the law of gravitation, which counts as a theory on, say, Popper's (1959) view? There are two worries if we think of 'empirical success' in such a narrow way when considering the NMA. On the one hand, it is unclear why the approximate truth of Newton's law of gravitation would go towards explaining the result *unless the other assumptions and machinery involved in generating it were also accurate to a considerable degree*. One could, perhaps, rephrase the NMA's conclusion as 'what best explains the success of science, on the assumption that the other representative elements therein are accurate to a high degree, is the approximate truth of its theories'. However, this limits its scope considerably; indeed, an anti-realist might accept the conclusion but deny the assumption. On the other hand, if one can run the argument for Newton's law of gravitation in isolation, then one can run the argument for the first law of classical mechanics, the second law of classical mechanics, and so forth. But that seems misguided. Moreover, having preferred the frequency-based NMA over its individual-theory counterpart, there is no obstacle to considering all the representative machinery—especially laws and models—used in such scenarios. We conclude that it is implicit in the standard frequency-based NMA that the success of science is best explained by the fact that the representations therein have a significant degree of accuracy *on the whole*. Ultimately, one's construal of theories—e.g., as bundles of models or as universal/statistical generalisations—is of little import in assessing it.

Let us now return to Le Verrier's story, which is also useful in illustrating how 'empirical success' need not involve merely taking a predictive apparatus, plugging in true inputs, and deriving novel (or previously unexplained) outputs. Le Verrier's prediction of the existence of Neptune was a remarkable feat because he had to 'work backwards' and explain what accounted for the aberrant orbit of Uranus. As Hanson (1962: p. 362) explains:

> If one knows a planet’s mass and its orbital elements, the disturbance it produces in another body is easily determined. This is the classical problem of perturbations. Leverrier’s problem … consists in describing the disturbances in Uranus, from which he then infers the mass and orbital elements of the disturbing planet. This is … “the inverse perturbation problem”; it is considerably more intricate to resolve than the classical problem.

Hanson (1962) suggests that the problem involved eight unknowns, but Lequeux (2013) claims that it involved twelve. The disagreement arises because Hanson (1962) counts the unknowns remaining *after* Le Verrier made several key assumptions. One such assumption was that the unseen planet’s orbit would lie in the same plane as the ecliptic. Another involved using a putative empirical law—the Titius-Bode law—to determine that the major axis of the unseen planet would be around double Uranus’s. That there was only one unseen body to be found was an assumption even before ‘the twelve’.

The subsequent discovery of Neptune was a monumental success for science—Hanson (1962: p. 363, 364) remarks that ‘Leverrier had carried Newtonian mechanics into the brightest heaven of scientific achievement’—despite being based on so much supposition and an element of luck. The Titius-Bode law, for instance, is far from being approximately true, although it works well for some planets. In principle, however, a ‘holistic’ frequency-based NMA can better accommodate such cases than an individual-theory NMA. Indeed, the episode also constituted a success for the Titius-Bode law—albeit perhaps a lesser one—although Hanson (1962) fails to remark on this. But this does not require that said law was approximately true considered in isolation.

Moreover, a model may appear to be representationally inaccurate because its target has been misidentified. Take the simple model of the pendulum as a case in point. It is not an accurate representation of long case clock pendulums of standard construction. However, if one takes its target to be how factors such as gravitational field strength, pendulum length, frequency, and period directionally interrelate, then it appears representationally accurate. It correctly captures the fact that increasing field strength decreases swing period, that increasing length decreases swing frequency, and so on. In the case of AI systems, this issue is especially interesting because the ‘target’—or what the system has learned to represent—is not always obvious. (To achieve its final task, a system may need to perform a variety of sub-tasks.) In the case of AF2, as we will see, it is relatively uncontroversial that some things, like heavy atom positions, are represented. But what are the representations used at a ‘lower’, or more fundamental, level (e.g., in the process of refining position hypotheses)?[21] For instance, are there many different statistical laws, governing how structural elements interrelate, which are implemented by the system? It is reasonable to think so, but the opacity inherent in the system is an obstacle to knowing. Ascribing accuracy to such representative machinery overall does not, however, require being able to achieve greater transparency than we already have (in cases such as AF2).

[21] Scientists are also content to write of hypotheses being formed and refined in AF2. For instance, Jumper et al, (2021: p. 285) state: ‘a concrete structural hypothesis arises early within the Evoformer blocks and is continuously refined’.

3.4 An AI-specific NMA

We are now in a position to state an AI-specific NMA. Let A be an arbitrary system in the domain of predictive AI systems utilised in science. For example, it could be an ML network. Let S_A be 'A is empirically successful' and $\approx$A be 'A accurately represents its task-related empirical target(s)'. Finally, let F be 'The relative frequency of empirically successful AIs classified in the domain is *f*.'

Here is the AI-specific NMA:

(a) $P(S_A | \approx A \& F)$ is high.
(b) $P(S_A | \neg \approx A \& F)$ is low.
Therefore, (c) $P(\approx A | S_A \& F)$ is high.

This argument is an instance of the inferential pattern in the NMA; for a given standard of 'high' and 'low', it is remarkably general. It differs notably from the traditional NMA only in so far as 'accurate representation' replaces 'approximate truth' and it involves no explicit mention of theories.

3.5 Summary

In the AI-specific NMA, a claim concerning the representational accuracy of the information in a system appears in place of the claim about the approximate truth of a theory that features in the traditional NMA. In this section, we have shown why this is so. First, approximate truth is normally taken to be a special kind of veridical representation. Second, AI-systems, such as AF2, may represent in the same way that human brains can. Third, the representations in such systems need not merely be correlates of theories or hypotheses; they might be correlates of initial conditions, models, and other machinery used for predictive purposes, too.

4 AlphaFold2 as a case study

AF2 (Jumper et al., 2021) is the most recent in a series of AI systems for predicting a protein's three-dimensional structure from its amino acid sequence. It tackles one of the most important problems in structural biology, because a protein's structure is indicative of its potential functions and mechanistic interactions. Although AI-inspired approaches to this problem have become increasingly common in recent years, AF2 is the first system to achieve near-atomic accuracy when compared to experimental methods. AF2 debuted at CASP14, the industry-standard biennial structure prediction competition, where results from structure prediction systems are compared to experimentally determined structures. Its median accuracy was within 0.96 Å for backbone protein structure, with similarly impressive performance for side chain structures. The runner up achieved backbone accuracy only within 2.8 Å (Jumper et al., 2021: p. 584).[22]

[22] For more recent assessments of AF2's impact, see AlQuraishi (2021) and Jones & Thornton (2022).

AF2 is trained primarily on the Protein Data Bank (PDB), an extensive database of solved protein structures. However, it does not merely memorize potential structures from this database and attempt to match input sequences to them. Rather, AF2 works by generating and iteratively improving a hypothesis about the most likely structure, given an input sequence. As its designers characterize it, AF2 works by as identifying a 'concrete structural hypothesis' early in the processing which is then 'continuously refined' during computation (Jumper et al., 2021: p. 585).

Although a full explanation of AF2's technical details are beyond the scope of this paper—the supplementary information to Jumper et al. (2021) alone runs to sixty-two pages—a few points are worth mentioning. First, AF2 processes information in two main stages. The first stage, dubbed the 'Evoformer', takes an amino acid sequence as input and outputs the aforementioned structural hypothesis. The structural hypothesis consists of two matrices, called the single and pair representations, respectively. The single representation contains information about the multiple sequence alignment (MSA) for the input sequence, and captures information about evolutionarily related residues in the input sequence. The pair representation contains information about the likely 3D structure encoded by the input sequence, given that MSA. This structural hypothesis is then passed to AF2's second main component, the structure prediction module, which takes the hypothesis and infers backbone and side chain structure through an iterative process.

Second, each of these components is informed by domain-specific knowledge. For instance, the MSA on which the single representation is based is motivated by the insight that evolutionarily related proteins fold in similar ways, and so finding an evolutionarily similar sequence provides information about the target sequence's likely final structure. Similarly, the pair representation is based on a template drawn from a database of known structures. Here the idea is that most proteins approximate known template structures, and so identifying a good template can significantly reduce the overall search space. And both the Evoformer and structure modules incorporate geometric constraints, such as the triangle inequality, which any physically possible structure much satisfy.

Despite its impressive performance, much is unknown about how AF2 works. In particular, little is known about the specific structural hypotheses AF2 learns. This is due in large part to AF2's opacity. Although AF2's high-level architecture is well-understood, it is far from clear how to translate its learned parameter values into humanly-graspable principles for identifying structure from sequence. Research in this area is ongoing.[23]

With all of this in mind, let's return to the AI-specific NMA. Although we will not attempt to pin down specific values for $P(S_A|\approx A\&F)$ or $P(S_A|\neg\approx A\&F)$, it is natural to think that the former is high and the latter low for many predictively successful AI systems. This is illustrated by AF2. The fact that AF2 accurately represents its target domain—e.g., by learning structural hypotheses relating sequence to 3D structure—-makes it highly probable that it will be empirically successful. By contrast, imagine

[23] For instance, Jumper et al., (2021: Supplementary Methods, §1.13) carry out a series of ablation studies, which consist in knocking out certain components of AF2 and assessing whether its performance degrades. This provides indirect evidence for the contribution made by different modules, although it is too coarse-grained to provide much understanding of the hypotheses AF2 learns.

that AF2 systematically misrepresented its target domain, perhaps by violating certain geometrical constraints in its structural hypotheses. If this were so, it is far less plausible that AF2 would enjoy the same degree of empirical success.

A few clarificatory remarks are in order at this point. First, it is important to remember that, just as the explanandum in the NMA is not how we obtain successful theories, the explanandum in the AI-specific NMA is not how we obtain successful AI systems. This matters because one might be tempted to account for AF2's accuracy by appealing to its training history, as opposed to its representational capacities. However, what needs to be explained is not how AF2 came to perform as it does. Rather, what needs to be explained is why AF2, given its current (trained) state, is predictively successful. The reason is that it accurately represents its target domain—or so the modified NMA suggests. Moreover, this explanation would hold even if AF2 had a different training history or no such history at all. Even if it popped into existence fully formed on DeepMind's servers, AF2 would still perform impressively.

Second, if sound, the modified NMA demonstrates that AF2 accurately represents its target domain. However, it does not tell us how AF2 accurately represents that domain. Indeed, as mentioned earlier, AF2's opacity prevents us from straightforwardly understanding how it does what it does.

Finally, we have only considered the modified NMA for a specific kind of deep learning system. But the modified argument applies much more widely than this. Not every AI system, or every machine learning system, uses deep learning. Consider a symbolic rule-based system such as a decision tree (e.g., Mitchell, 1997: ch. 3). In such a system, learned decision rules are encoded as conditional statements linking variables. When such a system is predictively successful, it is plausible that these conditionals accurately capture certain relationships between variables in the target domain. It is natural to think that the explanation for the system's success is that these if–then statements capture genuine relationships in the target domain, and it would not be hard to construct an NMA which concludes as much. Nevertheless, we have argued that a modified NMA applies even to opaque systems such as AF2, which do not wear their representational features on their sleeves.

By way of closing this section, we note a recent discussion of AF2 in which Skolnick et al. (2021: p. 4827) remark that AF2 seems to work 'by magic': 'Input a protein sequence and by "magic" the protein's three-dimensional structure appears.' Yet if the AI-specific NMA is sound, there is no magic here. There are no miracles, either. Indeed, Skolnick et al. (ibid.) note as much, continuing: 'Actually, AF2 figures out the complex interrelationships of the protein's residues that dictate what structure that protein sequence adopts.' We agree, although if the modified NMA is sound, AF2 probably achieves something more substantial. AF2 does not merely 'figure out' these complex interrelationships. It represents them and relies upon those representations when it infers structure from sequence.

5 Conclusion

We have argued that considerations in favour of standard NMA-style arguments apply, *mutatis mutandis*, to NMAs concerning empirically successful AI systems. If the former kind of argument is sound, so is the latter. They stand or fall together.[24] We will conclude with a few corollaries.

First, the success of an AI-specific NMA would suggest adequacy criteria for certain projects in explainable AI (‘XAI’). One goal of XAI research is to render opaque AI systems ‘interpretable’ or ‘explainable’ to human users (Burrell, 2016; Ribeiro et al., 2016). Plausibly, however, what interpretability involves depends on how a system is used in a specific context (Páez, 2019; Zednik, 2021). In light of a salient AI-specific NMA, one natural requirement in scientific contexts is that an XAI method should uncover the representations actually employed by an AI system. Yet not every XAI method is apt for this task.

Second, as we noted at the outset, several philosophers have suggested that AI systems might help scientists to understand their target domains. The existence of a salient AI-specific NMA might support this project, insofar as it would buttress the claim that the system accurately represents its target. One route to understanding via such a system would be to grasp these representations. Thus our argument should be congenial to those who think that AI systems may provide understanding and that understanding is factive (or perhaps quasi-factive).

Finally, our findings support the potential for various non-standard NMAs—e.g., NMAs for selective realisms—to be rendered AI-specific. Consider, for example, the idea that the empirical successes of theories are best explained by structure mapping rather than by approximate truth (Worrall, 1989). For an AI-specific equivalent to work, the information inside an AI, such as AF2, must be capable of representing structures in the world. We have argued that AIs are capable of this (although not that the empirical successes thereof indicate the presence of accurate structural representations).

Acknowledgements The work described in this paper was fully supported by a Senior Research Fellowship award from the Research Grants Council of the Hong Kong SAR, China (‘Philosophy of Contemporary and Future Science’, Project no. SRFS2122-3H01). We are grateful for feedback on earlier versions of the paper from audiences at Cambridge University, Kyoto University, Technical University Berlin, and Zhejiang University.

Funding Open Access Publishing Support Fund provided by Lingnan University. This work was funded by Research Grants Council, University Grants Committee, SRFS2122-3H01.

Declarations

Conflict of interest There are no other conflicts of interest to declare.

[24] To reiterate, however, we have *not* argued that either argument is sound.

References

AlQuraishi, M. (2021). Protein-structure prediction revolutionized. *Nature, 596*(7873), 487–488.

Barnes, E. (2022). Prediction versus accommodation. In E. N. Zalta (ed.), *Stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/archives/win2022/entries/prediction-accommodation/

Biddle, J. (2020). On predicting recidivism: Epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy, 52*, 321–341.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy, 10*, 615–678.

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines, 32*, 43–75.

Boyd, R. (1980). Scientific realism and naturalistic epistemology. *PSA: Proceedings of the biennial meeting of the philosophy of science association* (pp. 613–62).

Brown, T. D. (2021). Propositions are not representational. *Synthese, 199*, 5045–5060.

Brown, C. (2022). Narrow mental content. In E. N. Zalta (ed.), *Stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/archives/sum2022/entries/content-narrow/

Buckner, C. and Garson, J. (2019). Connectionism. In E. N. Zalta (ed.), *Stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/archives/fall2019/entries/connectionism

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society, 3*, 1–12.

Chakravartty, A. (2010). Truth and representation in science: Two inspirations from art. In R. Frigg & M. Hunter (Eds), *Beyond mimesis and convention: Representation in art and science* (pp. 33–50). Springer.

Chakravartty, A. (2017). Scientific realism. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/

Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science, 2*, 53–62.

Creel, K. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*, 568–589.

Cummins, R. (1996). *Representations, targets, and attitudes*. MIT Press.

Dawid, R., & Hartmann, S. (2018). The no miracles argument without the base rate fallacy. *Synthese, 195*, 4063–4079.

Dellsén, F. Forthcoming. An epistemic advantage of accommodation over prediction. *Philosophers' Imprint*.

Douglas, H., & Magnus, P. (2013). State of the field: Why novel prediction matters. *Studies in History and Philosophy of Science, 44*, 580–589.

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.

Duede, E. (2023). The representational status of deep learning models (arXiv:2303.12032). arXiv. http://arxiv.org/abs/2303.12032

Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton University Press.

Fahrbach, L. (2011). How the growth of science ends theory change. *Synthese, 180*, 139–155.

Floridi, L. (2005). Is semantic information meaningful data? *Philosophy and Phenomenological Research, 70*, 351–370.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Fodor, J. (1990). *A theory of content and other essays*. MIT Press.

Fodor, J. (1997). Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition, 62*, 109–119.

Giere, R. (2004). How models are used to represent reality. *Philosophy of Science, 71*, 742–752.

Giere, R. (2010). An agent-based conception of models and scientific representation. *Synthese, 172*, 269–281.

Hanson, N. R. (1962). Leverrier: The zenith and nadir of Newtonian mechanics. *Isis, 53*, 359–378.

Henderson, L. (2017). The no miracles argument and the base rate fallacy. *Synthese, 194*, 1295–1302.

Howson, C. (2013). Exhuming the no-miracles argument. *Analysis, 73*, 205–211.

Jones, D. T., & Thornton, J. M. (2022). The impact of AlphaFold2 one year on. *Nature Methods, 19*(1), 15–20.

Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589.

Keynes, J. M. (1921). *A treatise on probability*. Macmillan.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press.

Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science, 84*, 46–56.

Leplin, J. (1997). *A novel defense of scientific realism*. Oxford University Press.

Lequeux, J. (2013). *Le Verrier—magnificent and detestable astronomer*. Springer.

Lutz, S. (2014). What's right with a syntactic approach to theories and models? *Erkenntnis, 79*, 1475–1492.

Lutz, S. (2017). What was the syntax-semantics debate in the philosophy of science about? *Philosophy and Phenomenological Research, 95*, 319–352.

Magnus, P. D., & Callender, C. (2004). Realist ennui and the base rate fallacy. *Philosophy of Science, 71*, 320–338.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 7*, 115–133.

Menke, C. (2014). Does the miracles argument embody a base rate fallacy*?*. *Studies in History and Philosophy of Science Part A, 45*, 103–108.

Millikan, R. G. (1984). *Language*. MIT Press.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy, 86*, 281–297.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Musgrave, A. (1977). Logical versus historical theories of confirmation. *British Journal for the Philosophy of Science, 25*, 1–23.

Musgrave, A. (1988). The ultimate argument for scientific realism. In R. Nola (Ed.), *Relativism and realism in science* (pp. 229–252). Kluwer.

Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science, 58*, 168–184.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*, 441–459.

Papineau, D. (1987). *Reality and representation*. Blackwell.

Papineau, D. (2010). Realism, Ramsey sentences and the pessimistic meta-induction. *Studies in History and Philosophy of Science, 41*, 375–385.

Popper, K. R. (1959). *The logic of scientific discovery*. Basic Books.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. Routledge.

Psillos, S. (2004). Truth as a value. In L. G. Christophorou & G. Contopoulos (Eds), *Universal values* (pp. 143–146). Academy of Athens.

Psillos, S. (2009). *Knowing the structure of nature: Essays on realism and explanation*. Springer.

Putnam, H. (1975). What is mathematical truth? In H. Putnam (Ed.), *Mathematics, matter and method, collected papers.* (Vol. 2). Cambridge University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *ArXiv, 1602*, 04938v3.

Rowbottom, D. P. (2019a). Scientific realism: What it is, the contemporary debate, and new directions. *Synthese, 196*, 451–484.

Rowbottom, D. P. (2019b). *The instrument of science: Scientific anti-realism revitalised*. Routledge.

Rowbottom, D. P. (2022). Can meaningless statements be approximately true? On relaxing the semantic component of scientific realism. *Philosophy of Science, 89*, 879–888.

Rowbottom, D. P., Curtis-Trudel, A., & Peden, W. (2023). Evidence, computation and AI: Why evidence is not just in the head. *Asian Journal of Philosophy, 2*, 11.

Rubiera, C. O. (2021). AF2 is here: What's behind the structure prediction miracle. Oxford protein informatics group. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

Ruyant, Q. (2020). Semantic realism in the semantic conception of theories. *Synthese, 198*, 7965–7983.

Schlosser, M. (2019). Agency. In: E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy.* Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Sebastián, M. A. (2021). First-person representations and responsible agency in AI. *Synthese, 199*, 7061–7079.

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind and Language, 22*, 246–269.

Shea, N. (2013). Naturalising representational content. *Philosophy Compass, 8*, 496–509.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Skolnick, J., Gao, M., Zhou, H., & Singh, S. (2021). AlphaFold 2: Why it works and its implications for understanding the relationships of protein sequence, structure, and function. *Journal of Chemical Information and Modelling, 61*, 4827–4831.

Smart, J. J. C. (1968). *Between science and philosophy*. Random House.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*, 159–216.

Sprenger, J. (2016). The probabilistic no miracles argument. *European Journal for Philosophy of Science, 6*, 173–89.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71*, 767–779.

Sullivan, E. (2022). Inductive risk, understanding, and opaque machine learning models. *Philosophy of Science, 89*(5), 1065–1074.

Tamir, M., & Shech, E. (2023). Machine understanding and deep learning representation. *Synthese, 201*, 51. https://doi.org/10.1007/s11229-022-03999-y

Usher, M. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language, 16*, 331–334.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.

Vickers, P. (2019). Towards a realistic success-to-truth inference for scientific realism. *Synthese, 196*, 571–585.

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica, 43*, 99–124.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy and Technology, 34*, 265–288.