Three Strategies for Salvaging Epistemic Value in Deep Neural Network Modeling*

Philippe Verreault-Julien[†]

February 5, 2023

Abstract

Some how-possibly explanations have epistemic value because they are epistemically possible; we cannot rule out their truth. One paradoxical implication of that proposal is that epistemic value may be obtained from mere ignorance. For the less we know, then the more is epistemically possible. This chapter examines a particular class of problematic epistemically possible how-possibly explanations, viz. *epistemically opaque* how-possibly explanations. Those are how-possibly explanations justified by an epistemically opaque process. How could epistemically opaque how-possibly explanations have epistemic value if they result from a process about which we lack knowledge or understanding? This chapter proposes three different strategies to salvage epistemic value from epistemic opacity, namely salvaging value from 1) functional transparency, 2) modal operator interpretation, and 3) pursuitworthiness. It illustrates using cases from deep neural network modeling.

Contents

[*] In preparation for *Modeling the Possible*, edited by Tarja Knuuttila, Till Grüne-Yanoff, Rami Koskinen, and Ylwa Sjölin Wirling, Routledge.

[†] Eindhoven University of Technology. Email: p.verreault-julien@tue.nl. Website: https://pvjulien.net.

1 Introduction

What epistemic role highly idealized models may play has recently received a lot of attention. One prevalent proposal that has emerged is that these models provide *how-possibly explanations* (HPEs) (e.g. Bokulich 2014; Forber 2010; Grüne-Yanoff 2013; Reutlinger, Hangleiter, and Hartmann 2018; Rohwer and Rice 2013; Verreault-Julien 2019; Ylikoski and Aydinonat 2014). One alleged type of epistemically valuable HPEs are those considered as *epistemically possible* (Brainard 2020; Grüne-Yanoff 2013; Grüne-Yanoff and Verreault-Julien 2021; Sjölin Wirling and Grüne-Yanoff 2021; Verreault-Julien 2019). In a nutshell, epistemically possible HPEs are possible explanations that are not ruled out by our knowledge. Although potentially fruitful, this approach to epistemic value also has a paradoxical implication; the less we know, the more epistemically possible HPEs we have, which implies that ignorance itself would be a driver of epistemic value.

In this paper, I examine a particular class of puzzling epistemically possible HPEs that I call *epistemically opaque* HPEs (EO-HPEs). EO-HPEs are HPEs obtained by an epistemically opaque process such as computational simulation or deep neural network (DNN) models. In short, a process is epistemically opaque when an agent lacks knowledge or understanding of why the process yields the results that it does (e.g. Beisbart 2021; Durán and Formanek 2018; Humphreys 2009). The notion of EO-HPEs aims to capture a particular reason why we lack justification for the HPE, viz. the very process used to establish it. Contrary to HPEs acquired via a transparent process (e.g. an analytical model), EO-HPEs seem to face different justificatory and validation challenges. The problem EO-HPEs raise is the following: How could EO-HPEs have epistemic value if they result from a process about which we lack knowledge or understanding?

I argue that, in practice, the process's opacity is not always an obstacle to EO-HPEs' epistemic value. More specifically, I present three ways EO-HPEs may have epistemic value despite their opacity. First, some EO-HPEs result from a process which is functionally transparent: we have some understanding of how the algorithm works. Second, some EO-HPEs are only opaque according to some interpretations of the modal operator: again, this implies having some knowledge of the process's capacities. Third, some EO-HPEs are pursuitworthy even if they result from an opaque process; they may be promising despite a lack of justification. I illustrate using cases from DNN models.

This chapter makes two chief contributions. First, it elaborates on a recent and promising account of the value of HPEs as epistemically possible explanations. In particular, it identifies an obstacle to that account, namely the epistemic opacity of some modeling processes. Second, it present different ways to salvage the epistemic value of these models despite their opacity.

2 The Epistemic Value of HPEs

Highly idealized models often seem or actually fall short of faithfully representing the world. As a result, what sort of epistemic contribution they may make and why it is valuable is a source of contention. One proposal that has gained a lot of ground recently is that these models provide HPEs (e.g. Bokulich 2014; Forber 2010; Grüne-Yanoff 2013; Reutlinger, Hangleiter, and Hartmann 2018; Rohwer and Rice 2013; Verreault-Julien 2019; Ylikoski and Aydinonat 2014). There are various accounts of HPEs (e.g. Brainard 2020; Bokulich 2014; Brandon 1990; Dray 1968; Forber 2010; Hempel 1965; Verreault-Julien 2019), but arguably all of them emphasize that HPEs have modal features that differentiates them from how-actually explanations (HAEs). For my purposes, I will follow Verreault-Julien's (2019) account. According to him, explanations are sets of propositions (see also Strevens 2013). An explanation contains two subsets of propositions; the explanans, the propositions that do the explaining, and the explanandum, the propositions that describe what is explained. Explanations have to satisfy internal and external conditions of adequacy (see also Strevens 2013). The former refer to the form or structure of the explanation, the latter to the ontological match. For instance, a deductive-nomological (DN) explanation (Hempel and Oppenheim 1948) has to have the form of a deductive argument (internal conditions) and must have a true explanans and explanandum (external conditions).

In Verreault-Julien's terminology, HPEs have the general form '$\diamond$(p because q)' where p is the explanandum, q the explanans (e.g. generalization plus initial and auxiliary conditions), and $\diamond$ denotes a modal operator meaning 'it is possible that' according to a given an interpretation of the operator. HAEs are simply propositions of the form 'p because q'. The key difference lies in the introduction of a modal operator $\diamond$ in front of the explanation. In a nutshell, whereas HAEs are *actual* explanations, HPEs are *possible* explanations. The modal operator is in front of the whole 'p because q' to reflect that either the explanans, the explanandum, or the explanatory relation can be possible. For instance, sometimes scientist use actual causes and initial conditions to derive a possible explanandum, other times they start with an actual explanandum and try to generate it with possible initial conditions, etc.[1]

One crucial feature of that characterization of HPEs is that they have a truth value: 'It is possible that (p because q)' can be true or false.[2] Another key feature is that the modal operator can be interpreted in different ways. For instance, an explanation can be logically, mathematically, nomologically, causally, etc. possible. To assess whether an HPE is true or false, we thus need to know the interpretation of the modal operator.

One important interpretation of the operator is in terms of *epistemic possibility* (see Brainard 2020; Grüne-Yanoff 2013; Grüne-Yanoff and Verreault-Julien 2021; Sjölin

[1] See Grüne-Yanoff (2013) for various concrete examples.

[2] This is an important contrast with accounts that view HPEs as not satisfying any external conditions (e.g. Hempel 1965).

Wirling and Grüne-Yanoff 2021; Verreault-Julien 2019). Epistemic possibility tells us ways things might be relative to a given body of knowledge. In a nutshell, an HPE is epistemically possible iff our knowledge doesn't rule it out.[3] Epistemically possible HPEs are often those scientists submit when considering the set of possible explanations for a phenomenon. For instance, there might be a multitude of epistemically possible explanations for 'Why *p*?'. *p* might be because *q*, *r*, or another explanans. These possible explanations can be incompatible with one another, but since our knowledge cannot rule them out, they are all epistemically possible. To eliminate some epistemically possible explanations, we need additional evidence to update our knowledge and be able to rule out some of the possible explanations.[4]

To give a more concrete example, consider the phenomenon of people developing unusual blood clots following an injection of the COVID-19 Vaxzevria (AstraZeneca) vaccine in winter and spring 2021. Initially, the scientific community considered unlikely that the vaccine might be responsible for these clotting events. However, they could not rule it out.

> But the finding leaves researchers wrestling with a medical mystery: why would a vaccine trigger such an unusual condition? "Of course, there are hypotheses: maybe it's something with the vector, maybe it's an additive in the vaccine, maybe it's something in the production process ... I don't know," says Sabine Eichinger, a haematologist at the Medical University of Vienna. "It could be any of these things." (Ledford 2021, 334)

Further data collection and analysis supported the hypothesis that the vaccine was the cause (e.g. Whiteley et al. 2022). Modeling and experimental evidence, in turn, identified the adenovirus vector as the likely suspect (Baker et al. 2021), thus contributing to rule out other causes, such as the production process or the additives.

In winter and spring 2021, explaining the clotting by citing the vaccine was an epistemically possible HPE. An inasmuch as the HPE was true, it was valuable. However, this account of the epistemic value of HPEs has paradoxical features. Epistemic possibility is constrained by a body of knowledge. Thus, the less we know, the less epistemic possibilities are constrained. And the less epistemic possibilities are constrained, the more true epistemically possible HPEs we will have. This seems to be an infelicitous result. How could epistemic value derive from mere ignorance? Consider the hematologist in the quote above listing possible explanations and saying "I don't know". Her knowledge ruled out some explanations, but not others. Now, suppose a layperson would have been asked about possible explanations for the clotting. Presumably, that person's knowledge would have ruled out even less possible explanations. Should we conclude that the

[3] The semantics of epistemic modals is fraught with difficulties, which are outside the scope of this chapter. See Egan and Weatherson (2011) for an overview of key issues.

[4] A proponent of inference to the best explanation may also add that theoretical virtues can provide a basis for elimination (see e.g. Lipton 2004).

layperson's candidate explanations have epistemic value, especially compared to those of the expert scientist? We may want to resist that conclusion. But what if ignorance was built in the process of obtaining some epistemically possible HPEs?

3 Epistemically Opaque HPEs

In this section, I want to draw attention to the fact that the process for justifying HPEs is sometimes *epistemically opaque*. Epistemic opacity refers to the general idea that we do not always know or understand all the epistemically relevant features of a process (e.g. Beisbart 2021; Creel 2020; Durán and Formanek 2018; Humphreys 2009). What are these epistemically relevant elements? Various proposals have been made. Creel (2020) distinguishes between functional, structural, and run transparency, transparency here understood as the reverse of opacity.[5] Durán and Formanek (2018) identify the relevant elements with the justificatory steps; a process is opaque for an agent if she does not have access and cannot survey all the steps. Zednik (2021) argues that what these relevant elements are depends on the interests of particular stakeholders. Beisbart (2021, 11644) considers that the application of a method is opaque if it has a "disposition to resist epistemic access".

That being said, most accounts of epistemic opacity agree on the following two features. First, epistemic opacity can be agent-relative or process-relative. Agent-relative opacity depends on the cognitive capacities of agents or epistemic communities. An otherwise transparent model may be epistemically opaque for an agent who lacks the required skills or knowledge to grasp the epistemically relevant features. Process-relative opacity, also sometimes called 'essential opacity' (Alvarado 2021; Humphreys 2009), depends on the nature of the process itself. To make an analogy, one can fail to see through a window because of one's myopia (agent-relative) or because a film is applied to it (process-relative), making it opaque. Similarly, DNN models are often considered to have features that make them opaque, irrespective of the agents involved. In the rest of this chapter, I will only be concerned with process-relative opacity.

Second, process-relative opacity affects the justification we have for a process's results. One reason for this is because we cannot as easily assess the process's reliability. For instance, consider the now classic case of deep learning classifier for images of wolves (Ribeiro, Singh, and Guestrin 2016). Given an image as an input, the model outputs whether it contains a wolf or a husky. The model is accurate in the test data, but unbeknownst to the user, the model picks out wolves from the presence of snow in the background. In that sense, the process is not reliable because its classification doesn't depend on the features that wolves possess. It would therefore likely fail when presented with a picture of a wolf in a grassy environment. The field of explainable artificial intel-

[5] I will discuss Creel's account in more detail in section 4.1.

ligence (XAI) aims to make some features of deep learning models more transparent in order to help assess a process's reliability.

Two common examples of epistemically opaque processes are computational simulations and deep learning models. They are often considered to be 'black-boxes'. Yet, they are sometimes used for explanatory purposes. I call HPEs justified by an epistemically opaque process epistemically opaque HPEs (EO-HPEs).[6] EO-HPEs are prima facie problematic because we lack justification for the process that provides evidence for the HPEs. As we have seen, epistemically possible HPEs need to be suitably constrained if they are to have value. Otherwise, ignorance would increase value, which is an undesirable implication. However, it seems this is precisely what epistemically opaque processes do: they generate results via a process about which we lack knowledge. Therefore, our knowledge cannot suitably rule out epistemic possibilities and thus run the risk of *over*generating HPEs. For instance, assume we want to explain phenomenon *p*. We can simulate how different aspects of the phenomenon interact with each other. If we can generate the phenomenon's features, we may be tempted to infer that we have properly identified the phenomenon's explanans. However, because of the simulation's opacity, we do not (always) know if the results are due to the explanans or some other aspects of the process. For instance, a computational artefact may have caused the results. Or, we might just be unable to identify the explanans within the simulation process and thus lack an understanding of why we obtain particular results. How possibly can EO-HPEs have epistemic value if they are justified by an epistemically opaque process?

4 Salvaging Epistemic Value

In this section, I argue that despite the opacity of the process, we may have reasons to attribute value to the resulting EO-HPEs. I examine three cases from DNN models and propose three different strategies for salvaging epistemic value from the resulting EO-HPEs. They consist in salvaging value from 1) functional transparency, 2) modal operator interpretation, and 3) pursuitworthiness. These strategies are not mutually exclusive in two senses. First, two strategies may be available for the same EO-HPE, e.g. we may want to consider the EO-HPE functionally transparent *and* pursuitworthy. Second, the strategies may not always be logically independent. For instance, one might consider functional transparency necessary for selecting which modal operator should apply.[7] Nonetheless, I believe they constitute sufficiently different approaches to assessing the epistemic value of EO-HPEs.

[6] I borrow the terminology from Šešelja (2022), who uses it in a different way.

[7] Räz and Beisbart (2022) argue that understanding the model is necessary for explanatory understanding of phenomena. Here I remain agnostic regarding that claim.

4.1 Functional Transparency

There are many scientific questions related to animal coloration (Cuthill et al. 2017). What are the best colors to avoid detection in particular environments? Do the colors depend on the observer's visual system? Why, for instance, is tiger's fur orange? From an evolutionary perspective, we might want to know why species have evolved the color processing they have or why they have the colors that they do.

To make progress on these questions, Fennell et al. (2019) put a DNN to work to help identify which colors optimize or minimize detectability. This is crucial to understand the fitness effects of some phenotypes which, in turn, may explain why they were selected. In theory, it is possible to test empirically every single possible color on human subjects. In practice, it is impractical because color spaces are very large. For instance, testing the whole RGB gamut of 16,777,216 different colors would be a costly and time-consuming endeavor. Fennell et al. thus proposed to use a neural network to predict detection time on empirically untested colors. First, the researchers collected training data by carrying out an experiment with human subjects. They observed how much time it took humans to detect a randomly colored target in two simulated environments, a temperate forest and a semi-arid desert. They also processed images in order to simulate detection time for dichromats, i.e. species that perceive color via only two channels. Humans are trichromats and perceive colors through three channels, but most non-human mammals are dichromats and are effectively red-green color blind; red appears green to them. Then, the researchers trained a DNN to interpolate between experimented inputs and predict detection time. Suppose we have experimental data on magenta and cyan objects, but not on blue ones. The neural network interpolates between magenta and cyan to create the blue color and then estimate a detection time. By doing this for every shade, the researchers obtained predicted detection times for the whole RGB color space. As a result, the DNN allowed to identify the best and worst colors for detection.

Later in the article, Fennell et al. suggest that the results may help explain why some predators, e.g. tigers, are not green despite the optimal concealment it would provide. Consider the following question: 'Why is tiger's fur orange and not green?' The HPE they submit can be formulated as follows: 'It is epistemically possible that tiger's orange fur was selected for because it provides excellent concealment from dichromats'. Or, put slightly differently, there is little evolutionary pressure for tigers to evolve a green coat insofar as orange appears green for their preys.

This HPE relies on the hypothesis that the shade of green dichromats see in place of orange is actually hard to detect for them. Although seemingly obvious, that dichromacy enhances detection ability is also a serious hypothesis (e.g. Melin et al. 2007). More importantly, it relies on the prediction that shades close to the "dark olive" optimum are actually difficult to detect for dichromats in a temperate forest environment. However, if we do not understand why the model made the predictions that it did, how can we be sure it identified actual optima and minima? Identifying actual optima and minimal doesn't imply that we would have an HAE because other factors may be responsible for coloration. But it makes the HPE a more serious candidate.

Here, it is useful to differentiate ways a process can be opaque. Creel (2020) distinguishes between functional, structural, and run transparency. *Functional* transparency consists in knowing the functioning of the algorithm. By 'functioning', Creel doesn't mean knowing *how* the computational system instantiates the algorithm, but simply knowing *what* algorithm it instantiates. Knowing how the code produces the algorithm is *structural* transparency. *Run* transparency consists in knowing how a computational system was run in a particular instance, including the hardware implementation and how the program interacts with data. According to Creel, these different types of transparency are logically independent. One may not know how a program was run in a particular occasion (run transparency), yet know the algorithm's functioning (functional transparency).[8]

This taxonomy suggests one first line of defense for the value of EO-HPEs. We may say that an EO-HPE results from a process that lacks structural or run transparency, but which is functionally transparent. In the context of explanation, functional transparency is important since it allows us to identify the difference-makers a model captures (Räz and Beisbart 2022). And if an algorithm is functionally transparent, the process's opacity is less of a problem than whether the model provides a valid representation of the target (Sullivan 2022).

One important aim of the field of explainable artificial intelligence (XAI) is to increase transparency in one or the other of these senses. XAI methods (e.g. Lundberg and Lee 2017; Mordvintsev, Olah, and Tyka 2015; Ribeiro, Singh, and Guestrin 2016) can increase functional transparency by telling us why the algorithm made the decision it did on a particular or multiple inputs.[9] In turn, XAI methods can help uncover HPEs (Zednik and Boelsen 2022). Whether some methods will provide the required functional knowledge of the algorithm ultimately depends on the context (Zednik 2021). Some systems may be more difficult to interpret than others. In other cases, the amount of information we need might be minimal.

Do we have functional transparency in the case of the DNN for color detection? The algorithm is to some extent functionally transparent for two reasons. First, as Fennell et al. note, the problem the DNN needs to solve in that context is relatively low dimensional as only the color of the spheres changes between images of a particular environment. High dimensionality makes systems less transparent (Domingos 2012), but this is not the case here.[10] Second, Fennell et al. did carry out a limited validation experiment

[8] In a slightly different context, Sullivan (2022) also argues that "implementation black-boxes" may not prevent understanding the higher-level functioning of an algorithm.

[9] To what extent XAI methods can make a process functionally transparent is open to debate (e.g. Babic et al. 2021; Rudin 2019). My goal is not to settle it. Instead, my aim is to point out that this is an available strategy.

[10] It should be noted that the researchers believe their approach could also be useful for studying color de-

in which they tested with human subjects detection times of twenty-five 'easy', 'intermediate', and 'hard' colors. They found the predictions consistent with the experimental results for both the dichromat and trichromat conditions. The validation experiment plays a role akin to that of explainability techniques, viz. it helped make transparent that the DNN did pick out actual features that increase or decrease detection time.[11]

Despite the process's opacity in some respects, we do have knowledge of some of its epistemically relevant parts, viz. the functioning of the algorithm. In turn, this sort of transparency improves the justification we have in the process and indicates that the EO-HPEs we obtain is not a mere product of ignorance. This strategy might not be available for all DNNs. Despite our best XAI efforts, the model might remain functionally opaque. In this case, it may be better to justify the value of the EO-HPE differently, for instance by using the other lines of defense I propose below.

4.2 Modal Operator Interpretation

How the brain works remains for all practical purposes a mystery. It has been suggested that artificial neural networks (ANNs), especially deep convolutional neural networks (DCNNs), may provide candidate explanations of how the brain computes inputs into outputs (see e.g. Hassabis et al. 2017; Kriegeskorte 2015; Yamins and DiCarlo 2016). DCNN models seem to replicate, among others, how the brain processes visual sensory inputs using a hierarchy of representations that lead to object recognition. In particular, they are relatively good at predicting neurological data. Empirical results tend to show that artificial computer vision systems with an architecture that resembles that of biological organisms outperform those that do not. According to Kriegeskorte (2015, 431), "[t]his observation affirms the intuition that computer vision can learn from biological vision. Conversely, biological vision science can look to engineering for candidate computational theories". In short, the idea is that if models based on the architecture of the brain perform as well as better than biological systems, then these same models may explain how the biological systems work. However, the opacity of DCNNs is an obstacle to their explanatoriness. Indeed, since we do not understand how exactly the models build the representations and transform them, how could they provide an (etiological) explanation of brain sensory information processing? Insofar as we do not understand all the epistemically relevant features of these DCNSS, they provide EO-HPEs of neurological phenomena.

EO-HPEs are justified by an epistemically opaque process. As the previous section showed, having an EO-HPE does not imply that we are ignorant of *all* the relevant epistemic aspects. Sometimes, a DNN might be functionally transparent. Here, I would like to apply a slightly similar strategy and show that there are valuable things we know de-

tection in higher dimensionality spaces (see also Fennell et al. 2021; Talas et al. 2020).

[11] Here, I am bracketing the issue of whether testing on human subjects is a good proxy for other non-human species.

spite being in the presence of an EO-HPE.

In section 2, we have seen that HPEs have the general form '$\diamond$(p because q)'. In this formulation, the modal operator $\diamond$ can receive different interpretations. As a result, we may reach different conclusions of epistemic possibility depending on which interpretation we adopt. For instance, one HPE may be only logically possible whereas another may be nomologically possible. In the case of DNNs, even though we lack justification for the process's results, this lack of justification might only concern some interpretations of the modal operator. In particular, I want to suggest that we may be justified in the *mathematical* results, but not, e.g., the causal ones.

Although we lack a complete and full understanding of the capacities of DNNs (see e.g. Zhang et al. 2021), we do have some understanding of their mathematical properties. In particular, we know that they are so-called universal function approximators (see e.g. Cybenko 1989; Goodfellow, Bengio, and Courville 2016, sec. 6.4.1; Hornik, Stinchcombe, and White 1989; Zhou 2020). What do universal approximation theorems imply for the epistemic value of derived EO-HPEs? Universal approximator theorems are proofs that any network with at least one hidden layer and a sufficiently large number of hidden units can approximate any function between inputs and output. This mathematical result is important because "it takes off the table the question of whether any particular function is computable using a neural network. The answer to that question is always 'yes'. So the right question to ask is not whether any particular function is computable, but rather what's a *good* way to compute the function" (Nielsen 2015, ch. 4, emphasis in original). Assuming that biological neural activity can be represented by a function, universal approximation theorems prove that DNNs can approximate it. In other words, if a problem can be expressed by a mathematical function, then a DNN can solve it.

However, as the quote above alludes to, knowing that the network represents a logically possible function still leaves several questions unanswered. First, it does not tell us whether a given DNN will be able to learn the function. Second, we may not even know what function the DNN instantiates.[12] Third and relatedly, if we do not know what function the DNN instantiates, then we can hardly assess whether it is a good approximation of the real function. Therefore, we do not know whether this is actually the function computed by the brain nor how the brain actually computes the function. But, crucially, we know it is *mathematically possible* for DNNs to represent and approximate this function.

Although sciences are often more interested in HPEs that are causally possible, knowing that a problem has a possible mathematical solution is a valuable, albeit first step, in explaining a phenomenon. All that is epistemically causally possible is also logically possible. But not all that is logically possible is also causally possible. Determining the logical possibility of an explanation does some minimal headway into its causal possi-

[12] Perhaps one way of understanding this is by decomposing Creel's (2020) functional transparency in multiple components. Here, we may want to say that we know how the algorithm works at a very high mathematical level, but not at a lower representational or semantic one.

bility. For instance, in the context of economics, Verreault-Julien (2017) argues that the mathematical proof of the existence of a general equilibrium contributed to providing a mathematical HPE. So sometimes our ignorance will concern the causally possible, and not the logically possible, as is with ANNs of neurological computation.

4.3 Pursuitworthiness

One important aspect of the 'protein folding problem' (e.g. Dill and MacCallum 2012) concerns the ability to predict the three-dimensional shape of proteins — their structure — from its amino acid sequence. AlphaFold, a neural-network developed by DeepMind (Jumper et al. 2021), made a breakthrough contribution to solving that problem. It surpassed by a wide margin other models in the 14th Critical Assessment of protein Structure Prediction (CASP14), a biennial competition pitting different prediction methods against each other. AlphaFold is trained on the Protein Data Bank (PDB), a database of experimentally verified protein structures. The structure of proteins can be determined experimentally using, for instance, X-ray crystallography or cryo-electron microscopy. However, it is a difficult and expensive process. Since we know the amino sequence of many more proteins than we do of their structure, it is useful to predict structures from sequences. Theoretically, since the biological function of a protein depends on its structure, this holds the promise of improving our understanding of protein function. Practically, knowing protein structure may, among others, significantly speed up the development of new drugs.

For all its success at predicting, we do not have a full understanding of how or why AlphaFold works so well.

> Last, and perhaps the more immediate problem, AlphaFold2 models cannot be explained or externally validated. From our human perspective, it's essentially 'alien' technology that is currently beyond our understanding, so 'asking' why it predicted something in a particular conformation is clearly not feasible. (Jones and Thornton 2022, 18)

The model architecture is constrained by some scientific knowledge (Jumper et al. 2021), but the associations it establishes between known structures and sequences is opaque to users. It states the confidence it has in its predictions, which provides some information with respect to their potential reliability. Yet, albeit useful, this information doesn't make the model transparent in any of Creel's (2020) senses. We are thus in the presence of a model that is relatively good at predicting the structure of proteins, but of which we are ignorant of how it arrives at those predictions.

Although it is still early to assess the epistemic contribution of AlphaFold and similar models (e.g. Baek et al. 2021; Yang et al. 2020), to say that the scientific community was enthusiastic about the potential uses of the model would be an understatement (e.g.

Callaway 2022; Thornton, Laskowski, and Borkakoti 2021). To illustrate, consider the case of the SARS-CoV-2 virus responsible for the COVID-19 pandemic. The virus's proteins determine how it interacts with other biological systems, like humans. The function of a protein depends on its structure. Thus, knowing the structure can contribute to answering explanation-seeking questions such as 'Why protein *p* has function *f*?', 'Why are some variants more infective or virulent than others?', or 'Why is drug *d* effective against COVID-19?'. Unfortunately, we do not have experimentally validated structural models of all proteins, which limits our capacity to answer such questions (see e.g. Yan et al. 2022).

In March 2020, DeepMind (2020) released structural models for the SARS-CoV-2 membrane protein, Nsp2, Nsp4, Nsp6, and the Papain-like proteinase.[13] These structural models have then served as the basis of possible explanations of phenomena related to the virus. One notable example is due to Sadek, Zaha, and Ahmed (2021), who investigated the higher infectivity of the Omicron variant using AlphaFold and without relying on further experimental results. SARS-CoV-2 enters the host via the so-called spike protein. There are experimentally validated structures of the spike protein. However, how the many mutations translated into structure changes was unknown. Sadek et al. used AlphaFold to predict how these mutations would impact the structure. They concluded the following.

> Our study suggests that the higher infectivity of the Omicron variant can be explained in part by on the significant mutations in the RBD and the postfusion enhancement of the FP. Importantly, these results require further validation by X-ray crystallography and/or cryo-EM of the Omicron variant S-protein. (Sadek, Zaha, and Ahmed 2021, 5)

Another study relied on the model of protein Nsp6, which is involved in the infection process. One way it does it is by interacting with sigma receptors, which themselves are linked to the endoplasmic reticulum stress response. It was suggested that drugs that target the sigma receptors might reduce the reproduction of the virus. Two drugs, haloperidol and dextromethorphan, target the sigma receptors. Researchers remarked that haloperidol seemed to reduce viral production, while dextromethorphan increased it (Gordon et al. 2020). However, there was no explanation for this difference. Pandey et al. (2020) used AlphaFold's Nsp6 structural model to simulate how it interacts with the drugs. They concluded Nsp6 binds differently with the drugs, which may explain their differential effect on viral reproduction.

Other researchers (Gupta et al. 2021) used AlphaFold's predicted structure for Nsp2 to guide and validate their experimental cryo-electron microscopy data. The result was a complete structure of the Nsp2 protein. Analyzes of that structure suggested various

[13] They released updated versions in April and August 2020.

possible explanations involving the interaction between the host and Nsp2 for why some variants of the virus were more virulent.

Although epistemically opaque, AlphaFold is used to generate EO-HPEs. One reason for this, I submit, is because AlphaFold's results/models are *pursuitworthy*.[14] A scientific hypothesis or theory deserving attention is often said to be fruitful or pursuitworthy.[15] The idea of pursuitworthiness allows to demarcate between the hypotheses that should be pursued from those that should not. That HPEs should be fruitful or pursuitworthy is not a novel idea.[16] Notably, Resnik (1991, 142) argues that "a how-possibly explanation is better than a pseudo-explanation, since it has other important explanatory virtues, such as simplicity, testability, precision, fruitfulness, and the like".

Of course, what exactly pursuitworthiness entails is contentious (e.g. Šešelja and Straßer 2014; Shaw 2022) and I do not aim to settle this here. For my purposes, it suffices to note that pursuitworthiness becomes relevant when we lack sufficient epistemic support to otherwise discriminate between hypotheses. Indeed, if we *knew* that a given HPE was epistemically superior, we could just cease considering the other possibilities. But EO-HPEs, by virtue of being epistemically possible, are equally epistemically justified; they are all consistent with our knowledge.

Despite AlphaFold's opacity, as the immense interest surrounding it testifies, the scientific community clearly considers its results to be epistemically significant. Moreover, scientists would not engage in costly and time-consuming experiments if they did not believe the predicted models were useless. AlphaFold opens up new areas of investigations and suggests possibilities researchers had not and could not have contemplated before. In short, its results are pursuitworthy. Of course, this does not mean that *all* of its results are equally valuable. AlphaFold seems to fare better in some areas than others, although its predicting ability is also surprising in others. But some of the EO-HPEs we obtain with AlphaFold's assistance manifestly deserve our attention.

5 Conclusion

Many HPEs have epistemic value because they are epistemically possible, viz. we cannot rule out their truth. Epistemically possible HPEs play a central role in scientific progress and reasoning since they are often the precursors to how-actually explanations. When scientists want to explain a phenomenon, they submit a list of explanations consistent with what they know and then try to rule them out.

Although attractive, this picture of the value of HPEs also has an undesirable feature; more ignorance leads to more epistemically possible HPEs. Surely those HPEs cannot be as valuable, if they are at all valuable? In this chapter, I have examined a class of po-

[14] To be clear, I am not claiming this is the only reason or perhaps even the best one.

[15] See Kuhn (1977) and Laudan (1977) for early and influential accounts.

[16] See, e.g., Hempel (1942) on 'explanation sketches'.

tentially problematic HPEs, namely HPEs that originate from an epistemically opaque process. Looking at different cases of such HPEs stemming from DNN models, I have proposed three different strategies to salvage value in the face of opacity, namely salvaging value from 1) functional transparency, 2) modal operator interpretation, and 3) pursuitworthiness. All these strategies provide a rationale for attributing value to the HPE even though some ignorance is involved in how we obtain them.

Interestingly, not all strategies have an equally obvious connection to truth. Salvaging value from functional transparency or the modal operator interpretation have one; opacity is not a major obstacle because we do have *some* knowledge of epistemically relevant parts. However, salvaging value from pursuitworthiness has a more elusive relationship with truth. Perhaps this makes this strategy worthy of future attention.

References

Alvarado, Ramón. 2021. "Explaining Epistemic Opacity." Preprint. http://philsci-archive.pitt.edu/id/eprint/19384.

Babic, Boris, Sara Gerke, Theodoros Evgeniou, and I. Glenn Cohen. 2021. "Beware Explanations from AI in Health Care." *Science* 373 (6552): 284–86. https://doi.org/10.1126/science.abg1834.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, et al. 2021. "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network." *Science* 373 (6557): 871–76. https://doi.org/10.1126/science.abj8754.

Baker, Alexander T., Ryan J. Boyd, Daipayan Sarkar, Alicia Teijeira-Crespo, Chun Kit Chan, Emily Bates, Kasim Waraich, et al. 2021. "ChAdOx1 Interacts with CAR and PF4 with Implications for Thrombosis with Thrombocytopenia Syndrome." *Science Advances* 7 (49): eabl8213. https://doi.org/10.1126/sciadv.abl8213.

Beisbart, Claus. 2021. "Opacity Thought Through: On the Intransparency of Computer Simulations." *Synthese*, July, 1–24. https://doi.org/10.1007/s11229-021-03305-2.

Bokulich, Alisa. 2014. "How the Tiger Bush Got Its Stripes: 'How Possibly' Vs. 'How Actually' Model Explanations." *The Monist* 97 (3): 321–38. https://doi.org/10.5840/monist201497321.

Brainard, Lindsay. 2020. "How to Explain How-Possibly." *Philosopher's Imprint* 20 (13): 1–23.

Brandon, Robert N. 1990. *Adaptation and Environment*. Princeton: Princeton University Press.

Callaway, Ewen. 2022. "What's Next for AlphaFold and the AI Protein-Folding Revolution." *Nature* 604 (7905): 234–38. https://doi.org/10.1038/d41586-022-00997-5.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87 (4): 568–89. https://doi.org/10.1086/709729.

Cuthill, Innes C., William L. Allen, Kevin Arbuckle, Barbara Caspers, George Chaplin, Mark E. Hauber, Geoffrey E. Hill, et al. 2017. "The Biology of Color." *Science* 357 (6350): eaan0221. https://doi.org/10.1126/science.aan0221.

Cybenko, G. 1989. "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals and Systems* 2 (4): 303–14. https://doi.org/10.1007/BF0255 1274.

DeepMind. 2020. "Computational Predictions of Protein Structures Associated with COVID-19." https://www.deepmind.com/open-source/computational-predictions-of-protein-structures-associated-with-covid-19.

Dill, Ken A., and Justin L. MacCallum. 2012. "The Protein-Folding Problem, 50 Years On." *Science* 338 (6110): 1042–46. https://doi.org/10.1126/science.1219021.

Domingos, Pedro. 2012. "A Few Useful Things to Know about Machine Learning." *Communications of the ACM* 55 (10): 78–87. https://doi.org/10.1145/2347736.2347755.

Dray, William H. 1968. "On Explaining How-Possibly." *The Monist* 52 (3): 390–407.

Durán, Juan M., and Nico Formanek. 2018. "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28 (4): 645–66. https://doi.org/10.1007/s11023-018-9481-6.

Egan, Andy, and Brian Weatherson. 2011. *Epistemic Modality*. Oxford: Oxford University Press.

Fennell, John G., Laszlo Talas, Roland J. Baddeley, Innes C. Cuthill, and Nicholas E. Scott-Samuel. 2019. "Optimizing Colour for Camouflage and Visibility Using Deep Learning: The Effects of the Environment and the Observer's Visual System." *Journal of The Royal Society Interface* 16 (154): 20190183. https://doi.org/10.1098/rsif.2019.0183.

———. 2021. "The Camouflage Machine: Optimizing Protective Coloration Using Deep Learning with Genetic Algorithms." *Evolution* 75 (3): 614–24. https://doi.org/10.1111/evo.14162.

Forber, Patrick. 2010. "Confirmation and Explaining How Possible." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 41 (1): 32–40. https://doi.org/10.1016/j.shpsc.2009.12.006.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Gordon, David E., Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M. White, Matthew J. O'Meara, et al. 2020. "A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing." *Nature* 583 (7816): 459–68. https://doi.org/10.1038/s41586-020-2286-9.

Grüne-Yanoff, Till. 2013. "Appraising Models Nonrepresentationally." *Philosophy of Science* 80 (5): 850–61. https://doi.org/10.1086/673893.

Grüne-Yanoff, Till, and Philippe Verreault-Julien. 2021. "How-Possibly Explanations in Economics: Anything Goes?" *Journal of Economic Methodology* 28 (1): 114–23. https://doi.org/10.1080/1350178X.2020.1868779.

Gupta, Meghna, Caleigh M. Azumaya, Michelle Moritz, Sergei Pourmal, Amy Diallo, Gregory E. Merz, Gwendolyn Jang, et al. 2021. “CryoEM and AI Reveal a Structure of SARS-CoV-2 Nsp2, a Multifunctional Protein Involved in Key Host Processes.” bioRxiv. https://doi.org/10.1101/2021.05.10.443524.

Hassabis, Demis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 2017. “Neuroscience-Inspired Artificial Intelligence.” *Neuron* 95 (2): 245–58. https://doi.org/10.1016/j.neuron.2017.06.011.

Hempel, Carl G. 1942. “The Function of General Laws in History.” *The Journal of Philosophy* 39 (2): 35–48. https://doi.org/10.2307/2017635.

———. 1965. “Aspects of Scientific Explanation.” In *Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science*, 331–496. New York: Free Press.

Hempel, Carl G., and Paul Oppenheim. 1948. “Studies in the Logic of Explanation.” *Philosophy of Science* 15 (2): 135–75.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. “Multilayer Feedforward Networks Are Universal Approximators.” *Neural Networks* 2 (5): 359–66. https://doi.org/10.1016/0893-6080(89)90020-8.

Humphreys, Paul. 2009. “The Philosophical Novelty of Computer Simulation Methods.” *Synthese* 169 (3): 615–26.

Jones, David T., and Janet M. Thornton. 2022. “The Impact of AlphaFold2 One Year On.” *Nature Methods* 19 (1): 15–20. https://doi.org/10.1038/s41592-021-01365-3.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kriegeskorte, Nikolaus. 2015. “Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing.” *Annual Review of Vision Science* 1 (1): 417–46. https://doi.org/10.1146/annurev-vision-082114-035447.

Kuhn, T. 1977. “Objectivity, Value Judgment, and Theory Choice.” In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, 320–39. Chicago: University of Chicago Press.

Laudan, Larry. 1977. *Progress and Its Problems: Towards a Theory of Scientific Growth*. Berkeley and Los Angeles: University of California Press.

Ledford, Heidi. 2021. “How Could a COVID Vaccine Cause Blood Clots? Scientists Race to Investigate.” *Nature*, April. https://doi.org/10.1038/d41586-021-00940-0.

Lipton, Peter. 2004. *Inference to the Best Explanation*. Second. London: Routledge.

Lundberg, Scott M, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Melin, Amanda D., Linda M. Fedigan, Chihiro Hiramatsu, Courtney L. Sendall, and Shoji Kawamura. 2007. “Effects of Colour Vision Phenotype on Insect Capture by a Free-Ranging Population of White-Faced Capuchins, Cebus Capucinus.” *Animal Behaviour*

73 (1): 205–14. https://doi.org/10.1016/j.anbehav.2006.07.003.

Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. 2015. “Inceptionism: Going Deeper into Neural Networks.” *Google AI Blog*.

Nielsen, Michael A. 2015. *Neural Networks and Deep Learning*. Determination Press.

Pandey, Preeti, Kartikay Prasad, Amresh Prakash, and Vijay Kumar. 2020. “Insights into the Biased Activity of Dextromethorphan and Haloperidol Towards SARS-CoV-2 NSP6: In Silico Binding Mechanistic Analysis.” *Journal of Molecular Medicine* 98 (12): 1659–73. https://doi.org/10.1007/s00109-020-01980-1.

Räz, Tim, and Claus Beisbart. 2022. “The Importance of Understanding Deep Learning.” *Erkenntnis*, August. https://doi.org/10.1007/s10670-022-00605-y.

Resnik, David B. 1991. “How-Possibly Explanations in Biology.” *Acta Biotheoretica* 39 (2): 141–49. https://doi.org/10.1007/BF00046596.

Reutlinger, Alexander, Dominik Hangleiter, and Stephan Hartmann. 2018. “Understanding (with) Toy Models.” *The British Journal for the Philosophy of Science* 69 (4): 1069–99. https://doi.org/10.1093/bjps/axx005.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. KDD ’16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Rohwer, Yasha, and Collin Rice. 2013. “Hypothetical Pattern Idealization and Explanatory Models.” *Philosophy of Science* 80 (3): 334–55. https://doi.org/10.1086/671399.

Rudin, Cynthia. 2019. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Sadek, Ali, David Zaha, and Mahmoud Salama Ahmed. 2021. “Structural Insights of SARS-CoV-2 Spike Protein from Delta and Omicron Variants.” bioRxiv. https://doi.org/10.1101/2021.12.08.471777.

Šešelja, Dunja. 2022. “What Kind of Explanations Do We Get from Agent-Based Models of Scientific Inquiry?” Preprint. http://philsci-archive.pitt.edu/20532/.

Šešelja, Dunja, and Christian Straßer. 2014. “Epistemic Justification in the Context of Pursuit: A Coherentist Approach.” *Synthese* 191 (13): 3111–41. https://doi.org/10.1007/s11229-014-0476-4.

Shaw, Jamie. 2022. “On the Very Idea of Pursuitworthiness.” *Studies in History and Philosophy of Science* 91 (February): 103–12. https://doi.org/10.1016/j.shpsa.2021.11.016.

Sjölin Wirling, Ylwa, and Till Grüne-Yanoff. 2021. “Epistemic and Objective Possibility in Science.” *The British Journal for the Philosophy of Science*, August, 1–31. https://doi.org/10.1086/716925.

Strevens, Michael. 2013. “No Understanding Without Explanation.” *Studies in History and Philosophy of Science Part A* 44 (3): 510–15. https://doi.org/10.1016/j.shpsa.2012.1

2.005.

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* 73 (1): 109–33. https://doi.org/10.1093/bjps/axz035.

Talas, Laszlo, John G. Fennell, Karin Kjernsmo, Innes C. Cuthill, Nicholas E. Scott-Samuel, and Roland J. Baddeley. 2020. "CamoGAN: Evolving Optimum Camouflage with Generative Adversarial Networks." *Methods in Ecology and Evolution* 11 (2): 240–47. https://doi.org/10.1111/2041-210X.13334.

Thornton, Janet M., Roman A. Laskowski, and Neera Borkakoti. 2021. "AlphaFold Heralds a Data-Driven Revolution in Biology and Medicine." *Nature Medicine* 27 (10): 1666–69. https://doi.org/10.1038/s41591-021-01533-0.

Verreault-Julien, Philippe. 2017. "Non-Causal Understanding with Economic Models: The Case of General Equilibrium." *Journal of Economic Methodology* 24 (3): 297–317. https://doi.org/10.1080/1350178X.2017.1335424.

———. 2019. "How Could Models Possibly Provide How-Possibly Explanations?" *Studies in History and Philosophy of Science Part A* 73: 22–33. https://doi.org/10.1016/j.shpsa.2018.06.008.

Whiteley, William N., Samantha Ip, Jennifer A. Cooper, Thomas Bolton, Spencer Keene, Venexia Walker, Rachel Denholm, et al. 2022. "Association of COVID-19 Vaccines ChAdOx1 and BNT162b2 with Major Venous, Arterial, or Thrombocytopenic Events: A Population-Based Cohort Study of 46 Million Adults in England." *PLOS Medicine* 19 (2): e1003926. https://doi.org/10.1371/journal.pmed.1003926.

Yamins, Daniel L. K., and James J. DiCarlo. 2016. "Using Goal-Driven Deep Learning Models to Understand Sensory Cortex." *Nature Neuroscience* 19 (3): 356–65. https://doi.org/10.1038/nn.4244.

Yan, Weizhu, Yanhui Zheng, Xiaotao Zeng, Bin He, and Wei Cheng. 2022. "Structural Biology of SARS-CoV-2: Open the Door for Novel Therapies." *Signal Transduction and Targeted Therapy* 7 (1): 1–28. https://doi.org/10.1038/s41392-022-00884-5.

Yang, Jianyi, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. 2020. "Improved Protein Structure Prediction Using Predicted Interresidue Orientations." *Proceedings of the National Academy of Sciences* 117 (3): 1496–1503. https://doi.org/10.1073/pnas.1914677117.

Ylikoski, Petri, and N. Emrah Aydinonat. 2014. "Understanding with Theoretical Models." *Journal of Economic Methodology* 21 (1): 19–36. https://doi.org/10.1080/1350178X.2014.886470.

Zednik, Carlos. 2021. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34 (2): 265–88. https://doi.org/10.1007/s13347-019-00382-7.

Zednik, Carlos, and Hannes Boelsen. 2022. "Scientific Exploration and Explainable Artificial Intelligence." *Minds and Machines* 32 (1): 219–39. https://doi.org/10.1007/s11023-021-09583-6.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Communications of the ACM* 64 (3): 107–15. https://doi.org/10.1145/3446776.

Zhou, Ding-Xuan. 2020. "Universality of Deep Convolutional Neural Networks." *Applied and Computational Harmonic Analysis* 48 (2): 787–94. https://doi.org/10.1016/j.acha.2019.06.004.