

# How deep learning can justify pursuit, and why it matters

André Curtis-Trudel<sup>1</sup>, Niall Roe<sup>2</sup>, Konstaninos Voudouris<sup>3, 4</sup>

<sup>1</sup>Department of Philosophy, University of Cincinnati

<sup>2</sup>Department of History and Philosophy of Science, University of Cambridge

<sup>3</sup>Institute for Human-Centered AI, Helmholtz Munich

<sup>4</sup>Leverhulme Center for the Future of Intelligence, University of Cambridge

Draft of 5/9/25

## Abstract

We are in the midst of a deep learning revolution in science – or so many scientific and philosophical commentators would suggest. This article addresses a relatively underexplored aspect of this putative revolution, concerning how deep learning models (DLMs) impact the decision to rationally pursue a scientific idea. First, we develop an economic model of pursuitworthiness and use this model to analyze two ways that DLMs can justify pursuit: (i) by increasing the expected epistemic value of pursuit, and (ii) by decreasing the expected practical cost of pursuit. Then, we put this analysis to work. We argue (i) that it clarifies the sense in which DLMs may be said to be revolutionizing science, by radically impacting the economics of scientific activity, and (ii) that it brings into sharper focus certain scientific risks – what we call ‘illusions of pursuitworthiness’ – incurred by the shift toward DLM-driven science.

**Word count:** ~10,800

## 1. Introduction

In an article posted to Google’s DeepMind website in November 2024, DeepMind team members suggest that a “quiet revolution” is afoot (Griffin et al. 2024). A new generation of artificial intelligence technologies, driven by advances in deep learning, are transforming scientific discovery. Deep learning models (DLMs) are supplementing—and in some cases supplanting—more traditional observational, experimental, and modeling methods, accelerating the production of scientific outputs, and driving frontier research. According to DeepMind, we are entering a “new golden age” of scientific discovery, and autonomous, end-to-end DLM-driven discovery workflows are on the horizon (Choudhary et al. 2023; King & Zenil, 2023).

Philosophers of science have only begun to grapple with the impact of deep learning on scientific discovery. Extant work has addressed, among other things, whether DLM-driven discovery methods are methodologically novel (Nickles 2020, 2022); whether DLMs should lead us to rethink the prospects for and nature of scientific discovery (Boge 2022; Clark & Khosrowi 2022; Champion forthcoming); and whether the nascent field of ‘interpretable’ machine learning might facilitate discovery by allowing scientists to probe DLMs for their insights (Freiesleben et al. 2022; Zednik & Boelsen 2022; Barman et al. 2024).<sup>1</sup>

In this article, we address a comparatively underexplored question: how can DLMs justify the *pursuit* of a scientific idea? The decision to pursue an idea typically precedes the decision to regard that idea as a discovery (if it eventually comes to be accepted). And as Duede (2023, p. 1097) notes, DLMs frequently act “as guides for exploring promising avenues of pursuit,” orienting scientists towards novel insights and discoveries. What is less clear is when and why this guidance is justified. Indeed, to the extent that this question has been addressed, commentators have simply *assumed* that DLMs can justify pursuit without explaining why. Yet such an explanation is needed, not only to understand how DLMs can rationally guide scientists towards discoveries, but also to bring out much of what is distinctive – and perhaps even revolutionary – about the shift towards DLM-laden science.

Our aim in this article is to outline such an explanation. We begin in Section 2 by sketching a general decision-theoretic framework for evaluating how any methodological innovation, including DLMs, can rationally impact pursuitworthiness calculations. With this

---

<sup>1</sup> There is, of course, an extremely well-known and influential body of work on automated scientific discovery predating the current deep learning wave. See, e.g., Langley et al. (1987), Simon (1988).

framework in hand, we use it in Section 3 to highlight two main ways DLMs can justify pursuit: (i) by increasing the expected epistemic gain of pursuing certain lines of inquiry, and (ii) by decreasing the practical costs of pursuit. Although these do not exhaust how DLMs might impact pursuitworthiness calculations, they arguably reflect their most significant and widespread impacts given the current state of DLM technology.

Then, we draw out some broader philosophical implications of this analysis. In Section 4, we build on our framework to clarify how DLMs might be said to ‘revolutionize’ science. On our view, much of deep learning’s potential lies in its ability to impact the economics of scientific inquiry. Then, we argue that viewing the shift towards DLM-laden science through the lens of pursuitworthiness brings a novel family of scientific risks into sharper focus. Just as DLMs might promote “illusions of understanding” (Messeri & Crockett 2024) by leading scientists to overestimate the extent of their explanatory knowledge, we suggest that the ease and speed with which DLMs can be deployed can encourage “illusions of pursuitworthiness” by leading scientists to regard some ideas as more pursuitworthy than they in fact are. Understanding these risks will be a key item on the agenda for philosophers of science going forward, or so we we’ll suggest.

## **2. A framework for pursuit**

Before we can say how deep learning justifies pursuit, we need to say something about what makes an idea pursuitworthy in general. In this section, we sketch an idealized framework for regimenting pursuitworthiness judgments. This framework builds on economic accounts which construe reasoning about pursuit as a kind of cost-benefit analysis (Peirce 1876/1998; Rescher 1976, 1989; Wible 1998; Nyrup 2015; Duerr & Fischer 2025). On this approach, roughly speaking, it is rational to pursue an idea to the extent that pursuit would yield a favourable balance of epistemic benefits against practical costs.<sup>2</sup> We will regiment these judgments in a simple decision-theoretic model and show how to use it to make comparisons between different *methods* – including deep learning methods – that scientists might adopt.<sup>3</sup>

---

<sup>2</sup> Ethical considerations arguably matter too (Douglas 2009), but we will bracket them for simplicity’s sake. For a recent defense of economic approaches against some alternatives, see Duerr and Fischer (2025).

<sup>3</sup> In regimenting pursuitworthiness judgments this way, we do not assume that there is a ‘logic’ of pursuit, at least if this is understood as the claim that it is possible to articulate a set of proprietary inference rules or ‘minimal’ criteria for pursuit (Shaw 2022). We do assume that reasoning about pursuit is a kind of practical (means/end) reasoning.

## 2.1 Preliminaries

A few preliminary remarks before we get going. First, we're going to focus on pursuit of 'mid-sized' scientific entities like (families of) hypotheses, models, and theories – what we'll generically call *ideas*. Sometimes, philosophical discussion of pursuit focuses on pursuit of 'large-scale' entities like paradigms (Kuhn 1962), research programmes (Lakatos 1999) or research traditions (Laudan 1978). We're open to the idea that deep learning impacts pursuit of such entities, but they won't be our focus. Deep learning's most distinctive impacts to date are relatively local, affecting the decision to pursue ideas rather than larger entities. This makes ideas a natural place to start.<sup>4</sup>

Second, our framework is not intended to ground empirical claims about how scientists in fact make pursuitworthiness judgments. In practice, such decisions will be a matter of informed judgement – Duhem's 'bons sens'. Instead, we will use the framework to regiment judgments about when DLMs 'justify' the decision to pursue an idea. All we mean by this is that the choice to use a DLM can increase the pursuitworthiness of an idea, *ceteris paribus*. Obviously, DLMs can justify pursuit in this sense without providing all-things-considered justification for pursuit (for one thing, as noted, we're ignoring moral considerations). As we'll see, the framework is useful in large part because it provides a particularly clean way to tease apart different ways DLMs can justify pursuit.

Finally, readers will observe that our framework is entirely general, in that one could use it to understand how *any* methodological innovation can justify pursuit. By appealing to such a general account, we mean to show that much of deep learning's impact can be understood through familiar economic mechanisms rather than relying on special assumptions about the nature of deep learning specifically.

## 2.2 Epistemic dimensions of pursuit

It is now widely recognized that the epistemic considerations bearing on rational pursuit systematically differ from those bearing on rational acceptance, and that the epistemic

---

The considerations outlined in this section are best construed as defeasible, *ceteris paribus* elements of such reasoning.

<sup>4</sup> Moreover, insofar as the decision to pursue a paradigm, programme, or tradition involves the decision to pursue certain ideas, an account of the pursuitworthiness of the former requires an account of the pursuitworthiness of the latter. So an account of the pursuitworthiness of ideas will arguably be a necessary of an account of paradigms, programmes, or traditions.

requirements for pursuit are ordinarily regarded as less demanding than for acceptance (Achinstein 1993, McMullin 1976, Laudan 1978, Whitt 1992, Nickles 2006, Šešelja & Straßer 2014, Shaw 2022). An idea is accepted – treated as a bona fide item of scientific knowledge – after having already been pursued, and usually because such pursuit demonstrated a sufficiently high degree of positive epistemic value. The epistemic demands on pursuit are comparatively minimal.

While there is little agreement on what these demands come to in detail (Shaw 2022), two points stand out. First, while the fact that we have evidence for an idea can be an epistemic reason to pursue it, it is often enough to show that an idea is minimally plausible or perhaps consistent with an ambient background research programme or paradigm, all else equal. Sometimes, it is rational to pursue an idea even if it is partially or already confirmed, or even it is known to be implausible or false (Peirce 1932, 1.120, Nyrup 2015). This can be so if, for instance, pursuit might help scientists refine a modeling technique, improve other, more plausible models, or eliminate hypotheses from contention if it would be easy to do so.

Second, and perhaps more important, is an idea's epistemic promise or potential, in that the fact that an idea is promising can be a reason – perhaps even in some cases an overwhelming reason – to pursue it. Here we consider the following sorts of questions

How likely is [an idea] to give rise to interesting extensions? Does it show promise of being able to handle the outstanding problems (inconsistencies, anomalies, etc.) in the field? Is it likely to unify hitherto diverse areas, or perhaps open up entirely new territory? (McMullin 1976, p. 423).

More recent work along these lines has focused specifically on an idea's potential theoretical virtues (Nyrup 2015, Duerr & Fischer 2025). Nyrup (2015), for instance, argues that (potential) explanatoriness is a *ceteris paribus* reason to pursue one hypothesis over another: if  $H_1$  is (or is expected to be) more explanatory than  $H_2$ , then  $H_1$  is more pursuitworthy than  $H_2$  when other factors are held fixed. In a similar vein, Duerr and Fischer (2025) argue that an idea is pursuitworthy to the extent that it would favourably balance a range of Kuhn virtues – accuracy, unificatory power, simplicity, and so forth (Kuhn 1977). Other suggestions along these lines are clearly possible, and there is clearly considerable room for disagreement about which

virtues matter and how to weigh them. As we'll see, however, our analysis of how DLMs justify pursuit is compatible with a range of views on this mark.

To help clarify how these points bear on the pursuitworthiness of an idea, it is useful to regiment them in a simple causal decision-theoretic model.<sup>5</sup> Our starting point is Nyrup's (2015) model of the expected epistemic value (EEV) of an idea. The expected epistemic value of an idea concerns that idea's ability to satisfy our epistemic goals or aims. For Nyrup, there are three possible outcomes of the decision to pursue an idea H: we might accept H, we might reject H, or we might remain agnostic. Each can serve our aims, albeit in different circumstances. Acceptance (rejection) is apt if we wish to believe truths (avoid believing falsehoods), while agnosticism is apt if our evidence is equivocal, for instance. If we assume that these are the three main outcomes of pursuit, then the expected epistemic value of H can be construed as follows:

$$\begin{aligned}
 EEV(p(H)) &= EV(a(H))Pr(a(H)|p(H)) \\
 &\quad + EV(r(H))Pr(r(H)|p(H)) \\
 &\quad + EV(ag(H))Pr(ag(H)|p(H))
 \end{aligned}
 \tag{1}$$

Where  $p(H)$ ,  $a(H)$ ,  $r(H)$ , and  $ag(H)$  reflect the decisions to pursue, accept, reject, or remain agnostic about H, respectively, and  $Pr(a(H)|p(H))$ ,  $Pr(r(H)|p(H))$ , and  $Pr(ag(H)|p(H))$  are the probability of acceptance, rejection, or agnosticism given the choice to pursue H.<sup>6</sup>

Note that this formula is neutral about how to determine  $EV(a(H))$ ,  $EV(r(H))$ , or  $EV(ag(H))$ . This is a feature, not a bug. The epistemic value of a particular outcome will typically be highly context-sensitive, depending on the goals and interests of the researchers in question, their research programme, and the state of the relevant scientific knowledge. In some contexts, it may be valuable to accept an idea which is not explanatory but which yields greater predictive power, while in other contexts the reverse may hold. We'll say a bit more about these tradeoffs in Section 3.

---

<sup>5</sup> Causal decision theory is not the only way to regiment these ideas. For others, see, e.g., (Kitcher 1995, Ch. 8; Peirce 1876 / 1998; Wible 1998; Duerr & Fischer 2025).

<sup>6</sup> We remain agnostic about the interpretation of these probabilities (Joyce 1999), although we note that a subjectivist interpretation is particularly natural in this context. Nothing in our analysis of deep learning's impact on pursuit hinges on this choice, however.

We should also note that it is possible to complicate this model in various ways.<sup>7</sup> We won't worry about these complications, except one. (1) ignores practical considerations like the cost of pursuit. Yet as we will see next, such considerations are essential for understanding when an epistemically valuable hypothesis is *worth* pursuing.

### 2.3 Practical dimensions of pursuit

Practical considerations are relevant to pursuit because pursuit requires an investment of time, labour, and other resources. This investment must be balanced against the expected epistemic value of pursuit. This dimension of pursuit plays a central role in C.S. Peirce's account of the economics of science:

Proposals for hypotheses inundate us in an overwhelming flood, while the process of verification to which each one must be subjected before it can count as at all an item, even of likely knowledge, is so very costly in time, energy, and money—and consequently in ideas which might have been had for that time, energy, and money, that Economy would override every other consideration even if there were any other serious considerations. In fact there are no others. (Peirce 1932–58, 5.602)

We needn't accept Peirce's suggestion that 'economy' is the *sole* consideration to appreciate the point that economic considerations factor heavily into the decision to pursue an idea. In some cases, they may override 'purely' epistemic concerns.<sup>8</sup>

What sorts of practical costs bear on the pursuitworthiness of an idea? Duerr and Fischer suggest that the relevant practical costs are cognitive, concerning "the mental efforts of the ideal scientist" (2025, p. 17). Their thought is that some ideas are harder to pursue than others because they are complex, poorly understood, or involve significant departures from more familiar ideas. But while cognitive costs are undoubtedly part of the story, it would be an oversight to ignore time, money, and physical resources – material costs of doing science. It is possible to boost the pursuitworthiness of an idea by *decreasing* the (material) cost of pursuing it. This can be

---

<sup>7</sup> For instance, we could conditionalize on the truth of H at each line. This would reflect the fact that it is epistemically valuable to accept (reject) H when it is true (false) and costly to accept (reject) H when it is false (true) (see Nyrup 2015).

<sup>8</sup> See e.g., (Peirce 1992, ch. 31; 1998, ch. 12).

accomplished by deploying methods which are cheaper, faster, or otherwise less resource intensive while holding epistemic value and cognitive labour fixed.

As with the epistemic considerations surveyed above, it is possible to regiment the practical dimensions of pursuit in an idealized decision-theoretic framework. One straightforward model associates a fixed practical cost,  $PC(p(H))$ , with the pursuit of an hypothesis.<sup>9</sup> However, it is possible that different investments of time, energy, and other resources might be needed to arrive at different attitudes towards an idea. An idea could be easily confirmed if true, but hard to disconfirm if false, for instance. For this reason, it is useful to distinguish the resource investment associated with coming to accept an idea,  $PC(a(H))$ , from the investment associated with coming to reject it,  $PC(r(H))$ . Additionally, we should factor in the cost of *failed* investigations into an idea yielding agnosticism,  $PC(ag(H))$ . These factors are naturally captured by the *expected* practical cost of pursuit:

$$\begin{aligned}
 EPC(p(H)) &= PC(a(H))Pr(a(H)|p(H)) \\
 &\quad + PC(r(H))Pr(r(H)|p(H)) \\
 &\quad + PC(ag(H))Pr(ag(H)|p(H))
 \end{aligned}
 \tag{2}$$

Note that while the model allows for the possibility that  $a(H) \neq r(H)$ , if  $a(H) = r(H)$ , it simplifies to  $PC(\sim ag(H))Pr(\sim ag(H)|p(H)) + PC(ag(H))Pr(ag(H)|ag(H))$ .

#### 2.4 Putting it all together

When deciding whether to pursue a scientific hypothesis, the economics of inquiry require us to perform a cost-benefit analysis. An idea is pursuitworthy to the extent that its benefits justify its cost. There are various ways to model this tradeoff, but one straightforward approach construes the pursuitworthiness of an idea as the expected epistemic value of pursuit minus the expected practical cost:

$$P(H) = EEV(p(H)) - EPC(p(H))$$

---

<sup>9</sup> In an unpublished manuscript, Nyruup (ms) sketches such a model; see also Duerr and Fischer's (2025) 'virtue-economic' model.

$P(H)$  reflects the pursuitworthiness of an idea simpliciter. When  $EEV(p(H))$  is at all positive,  $H$  is potentially pursuitworthy, subject to applying a sufficiently affordable method.<sup>10</sup> These analyses are not made in a vacuum. Pursuitworthiness judgments are often comparative: they are sensitive to the expected value of different possibilities open to the scientist at a given time. Here, there are two main dimensions to consider. The first concerns an idea's standing relative to its peers, against which it competes for resources. At any given time, the scientist is confronted with a range of possible ideas and must pick from them a proper subset to pursue (or continue to pursue) (cf. Laudan 1978, p. 71). The second concerns the *methods* available to a scientist at a given time. This dimension has received less explicit attention in recent discussions of pursuitworthiness, but as we will see it is important to understanding deep learning's impact on pursuit. Given a fixed idea, a scientist may pick from a range of methods available for pursuing that idea, and different methods may come with different benefits and costs. The scientist's challenge is to apportion resources over these possible ideas, by deploying available methods, to maximize overall epistemic benefit while using resources efficiently. Next, we will see how deep learning impacts how scientists address this challenge.

### 3. Three ways DLMS can justify pursuit

We take it to be more or less uncontroversial that DLMS *can* and *do* guide scientific pursuit (Duede 2023; Finn & Khosrowi 2024). It is less clear, though, why this guidance is justified. In this section, we'll use our framework to map out some of the more significant ways that DLMS can justify pursuit.

We should manage expectations. There are too many applications of deep learning in science to catalogue here (Choudhary et al. 2024). Instead, to keep the project manageable, we'll use AlphaFold (Jumper et al., 2021) and other DLM-based protein-structure prediction systems like RoseTTAFold (Baek et al., 2021) as running examples. These are perhaps the most successful applications of deep learning to emerge in recent years, and, as we'll suggest, they exemplify the most important respects in which DLMS can justify pursuit. Indeed, their impact

---

<sup>10</sup>  $EEV(p(H))$  could be negative, for example when  $H$  is extremely appealing but known to be wrong. We consider this among other pursuit risks in the penultimate section.

on pursuit has been explicitly noted by scientists reflecting on the impact of these systems on structural biology. In a recent comparison of AlphaFold predictions and experimental structure determination methods, Terwiliger and colleagues suggest that AlphaFold predictions are best construed as “exceptionally useful hypotheses” which can guide pursuit:

AlphaFold predictions are already changing the way that hypotheses about protein structures are generated and tested ... they provide plausible hypotheses that can suggest mechanisms of action and allow designing of experiments with specific expected outcomes. Using these predictions as starting hypotheses can also greatly accelerate the process of experimental structure determination. (Terwiliger et al. 2023, p. 4).

Although they do not use the language of pursuitworthiness, it seems clear that these remarks point towards AlphaFold’s role in guiding pursuit. In what follows, we’ll unpack their suggestion with reference to our economic framework.

Eyeballing equation (3), we can see that to increase the pursuitworthiness of an idea we can either increase EEV or decrease EPC. And indeed Terwiliger et al.’s remarks seem to suggest that AlphaFold makes both sorts of contribution. First, we’ll highlight two ways DLMs like AlphaFold can justify pursuit by boosting expected epistemic value, by suggesting ideas which are plausible, or which are expected to be particularly fruitful. Then, we’ll focus on the idea that DLMs often justify pursuit not by improving an idea’s epistemic prospects but rather by decreasing practical costs.<sup>11</sup>

### *3.1 Increasing expected epistemic value I: plausibility and coherence*

Let’s begin with Terwiliger et al.’s remark that AlphaFold suggests *plausible* hypotheses. By this they mean that its outputs are, at least for a wide range of cases, roughly what one would expect to see given available theoretical background. This background includes both general principles about how proteins tend to fold as well as more local knowledge about how proteins of the sort in question behave. By virtue of providing hypotheses which are plausible in this sense,

---

<sup>11</sup> Two caveats. First, we treat epistemic and practical factors separately, but practice a given DLM will ordinarily have multiple overlapping impacts. It will not always be clear where the epistemic factors shade off and the practical factors begin. Second, we aren’t saying that these are the only ways DLMs can justify pursuit. We claim only that these are a few of the most significant ways, given the current state of technology.

AlphaFold provides epistemic justification for pursuit – for instance, for pursuit of potential mechanisms of action or new experiments; by contrast, if these outputs were not plausible, it’s hard to see how they would justify these decisions.<sup>12</sup>

Note that this is a straightforward consequence of our framework. One way to bring this out is as the relative expected epistemic value of an hypothesis given the choice to pursue it with the help of a DLM versus the choice to pursue it some other way. Let  $p_{AF}(H)$  be the decision to pursue  $H$  with the help of AlphaFold. And let  $p_{\sim AF}(H)$  reflect the decision to pursue  $H$  some other, like a more traditional experimental or simulation method. Arguably, the fact that the AlphaFold reveals  $H$  to be plausible bears on the likelihood of acceptance, rejection, or agnosticism, given pursuit. If, for example,  $\Pr(a(H) | p_{AF}(H)) > \Pr(a(H) | p_{\sim AF}(H))$ , we’ll have  $EEV(p_{AF}(H)) > EEV(p_{\sim AF}(H))$ , when all else is equal. In other words, the expected epistemic value of pursuit will increase if using AlphaFold leads to likelihood of acceptance when compared to other methods. And a parallel story can be told for when AlphaFold recommends *rejection* of an hypothesis too, when rejection would be epistemically valuable.

We submit that, in this respect, AlphaFold and other protein structure predictors are representative of a wider class of cases. Many ‘vanilla’ scientific applications use DLMs to supplement, and in sometimes replace, more traditional predictive, observational, and classificatory methods. In such cases, the fact that they produce a (minimally) plausible idea  $H$  provides reason to pursue it, in something like the way that experiment or observation can make an idea more pursuitworthy by rendering it minimally plausible or coherent with background knowledge (Whitt 1992).<sup>13</sup>

Of course, not every AlphaFold output can justify pursuit this way. What grounds the judgment that AlphaFold hypotheses are plausible (when they are)? More abstractly, what does it take for a DLM to justify pursuit by making an idea plausible? We already noted that the epistemic demands on pursuit are less demanding than acceptance. Any attempt to answer this

---

<sup>12</sup> AlphaFold2 *does* produce implausible hypotheses for some families of proteins, like those with ligand binding. It also does not account for environmental interactions, among much else. Such cases are well-known, and, unsurprisingly, structural biologists do *not* regard AlphaFold outputs as plausible guides to pursuit in such cases (Terwiliger et al. 2024).

<sup>13</sup> This raises additional questions which we defer for future work. For instance, what kinds of evidence do DLMs provide, and in virtue of what does this evidence guide pursuit? For some preliminary discussion see Finn & Khosrowi (2024).

question should reflect this lower standard.<sup>14</sup> One promising answer is due to Duede, who suggests that “the mere inductive support DLMs provide is epistemically sufficient to guide pursuit” (2023, p. 1094). By ‘inductive support’ Duede means a model’s performance on holdout training data. Although Duede does not use the language of plausibility, one way to put the idea is that if a model performs suitably well on a suitably specified holdout test, then we have a reason to regard that model’s outputs as ‘plausible’ – plausible enough, at any rate, to consider pursuing them.<sup>15</sup>

We think this is headed in the right direction, but it is arguably not the whole story (nor, to be clear, does Duede claim that it is). One well-known issue is that performance on holdout data is no guarantee of a model’s deployment performance (Zhang et al. 2021; Freiesleben & Grote 2024). Shortcut learning and distribution shifts can undermine a model’s deployment performance in ways that are difficult to detect by inspecting holdout loss (Geirhos et al. 2020). If a model is known to rely on shortcuts or surface statistics, or if the deployment distribution shifts, it may not be sensible to pursue its outputs even if it exhibits low holdout loss.

In fact, performance on holdout tests is just one piece of information scientists bring to bear in determining whether DLM outputs are plausible enough to pursue. If an output is consistent with relevant domain-specific background knowledge or coheres with that knowledge in the right way (cf. Šešelja & Straßer 2014), a scientist might have justification ‘epistemically sufficient’ for pursuit even if the model overall performs equivocally on holdout tests. For instance, AlphaFold performs reasonably well on holdout data and a small set of unseen competition structures (but see Akdel 2022 for limitations). Yet there is little hope of validating its outputs wholesale against experimentally determined holdout structures. Instead, as noted earlier, decisions about whether to pursue an AlphaFold hypothesis are supplemented by general theoretical considerations and background knowledge. When hypotheses appear plausible by the lights of this background, they are apt for pursuit, *ceteris paribus*. Here, then, we see that plausibility need *not* reduce to questions of DLM inductive support or reliability.

---

<sup>14</sup> Indeed, considered scientific judgments about AlphaFold’s potential role reflects this – even though its hypotheses are regarded as plausible starting points for further inquiry, structural biologists generally regards them as needing further (typically experimental) validation (Akdel et al. 2022; but see Zhakarova ms).

<sup>15</sup> This suggestion aligns with reliabilist approach to the question of when we’re justified in believing the outputs of a machine learning system (e.g., Durán & Formanek 2018, Duede 2022, Dúran 2025). On this kind of approach, we would say that an idea produced by a DLM is plausible if the system is suitably reliable (in relevant respects).

### 3.2 Boosting expected epistemic value II: fruitfulness

Plausibility is only part of the story. Let's go back to Terwiliger et al.'s remark that AlphaFold outputs suggest mechanisms of action, like binding mechanisms or other interactions, or possible experiments. It's natural to interpret them as saying that these outputs are pursuitworthy (when they are) not just because they are plausible, but also because they help biochemists accomplish other epistemic goals like discovering mechanisms and experiments. In other words, it's sometimes rational for biochemists to pursue AlphaFold-generated hypotheses because they are potentially *fruitful* (McMullin 1976). Notice that this is conceptually distinct from the fact that AlphaFold outputs are plausible. Rather, in this case, what matters is that AlphaFold outputs can potentially contribute to other epistemic goals in structural biology.

That AlphaFold can justify pursuit this way is a straightforward consequence of our framework. Once again this is easiest to see when we consider comparative pursuitworthiness judgments. Suppose that AlphaFold generates a prediction  $H$  which suggests a possible mechanism of action, and suppose that it is epistemically valuable, given the goals and interests of biochemists, to accept hypotheses which underwrite or furnish possible mechanisms. And consider an alternative  $H^*$  which does not suggest a possible mechanism, or which suggests an idea which is less clear about the relevant mechanism than  $H$ . Plausibly, we have  $EV(a(H)) > EV(a(H^*))$ , from which it follows that  $EEV(p(H)) > EEV(p(H^*))$ , when everything else is held fixed (cf. Nyrup 2015).

It should be evident that roughly the same kind of story could be told for any epistemic good which might be attached to the output of a DLM. There is, of course, room for disagreement about *which* epistemic goods this includes. Recent philosophical discussion tends to construe the fruitfulness of an idea as its potential to exhibit, or contribute to, theoretical virtues like understanding and explanation, among others (Šešelja & Straßer 2014, Ivani 2019, Duerr & Fischer 2025). In the case of DLMs, it is exceedingly controversial whether they *can* provide understanding (Chirumuuta 2020, Sullivan 2022, Rätz & Beisbart 2022) or explanation (Gross 2024, Matthiesen forthcoming). We won't try to settle *that* debate here. What's important for our purposes is the observation that *if* a DLM output potentially has or promotes virtues like understanding or explanation, then it can straightforwardly provide justification for pursuit along roughly these lines.

In the case of AlphaFold, an idea is judged fruitful by practitioners after a kind of post hoc analysis, given relevant background knowledge. Yet there is an increasingly large body of work which attempts to design DLMs which themselves *identify* or *produce* fruitful ideas in the first place. This would provide a rather different justification for pursuit. Here, what would justify pursuit is not that an idea has passed a post hoc analysis of its promise, but rather that it is flagged by a system explicitly designed to identify or produce fruitful ideas in the first place. Technical work in this vein is ongoing, and notable examples include rediscovering known physical concepts in toy settings (Iten et al. 2020, Vervoort et al. 2023), discovering novel theories of human decision-making (Peterson et al. 2021, Binz et al. 2024), identifying high-impact research proposals (Sikimić & Radovanović 2022), and building semi-autonomous research assistants, capable of suggesting novel, tailor-made ideas for individual scientists or research groups (Krenn et al. 2022; Gu & Krenn 2025a, 2025b), among others.

We're not going to speculate about the potential success of these efforts.<sup>16</sup> What's interesting, for our purposes, is the question of how the outputs of such systems can justify pursuit. To some extent, the story will parallel the first sort of case: they can justify pursuit by positively impacting the epistemic value of acceptance (rejection, agnosticism). In this case however there is an additional and somewhat more subtle possibility. Recall that pursuitworthiness judgments are often comparative. Part of what's going on in these cases is that DLMs are being used to impact the range of (plausible) ideas under consideration. In particular, the expectation guiding these systems is that they might furnish scientists with *new* potentially fruitful ideas. This can justify the decision to pursue an idea not only by directly impacting its expected epistemic value but also by shifting its standing relative to its peers.

To bring this out, consider that at any given time, the scientist is confronted with a range of possible ideas and must pick from them a proper subset to pursue (or continue to pursue) (Laudan 1978, p. 71). The challenge is to apportion resources over this subset in a way that maximizes overall epistemic benefit while using resources efficiently. We want to suggest it's useful to think of these cases – in which a DLM tries to identify fruitful ideas – as involving

---

<sup>16</sup> See Khosrowi (2025) for relevant discussion. While much of this work is in its infancy, initial results are striking. For instance, Gu and Krenn's (2025b) model was trained to identify potentially fruitful links between scientific concepts. To gauge the potential impact of novel links identified by their network, they surveyed researchers on 4,451 novel research ideas produced by the system. Nearly 25% were ranked as 'interesting' or 'very interesting' by domain experts. Not awful for a machine.

expansion or contraction of this subset. By introducing additional projects – some of which may have significant promise, as flagged by the system – scientists must reconsider how to redistribute resources. There is no *a priori* guarantee that an optimal distribution for one set of projects will resemble, even approximately, the optimal distribution over a different, expanded set. Thus, the introduction of novel, potentially fruitful ideas can justify pursuit not just by increasing the pursuitworthiness of an idea considered in isolation, but also by encouraging scientists to apportion finite resources over the set of plausible ideas in new ways.

### *3.3 Lowering expected practical costs*

In broaching the idea that the scientist’s challenge is to apportion resources over a set of ideas, we reencounter the notion that practical considerations bear on the justification of pursuit. There are many respects in which DLMs clearly lower practical costs: it is often faster and cheaper to run than alternative observational or experimental methods, it does not tire, and does not need to break for lunch or play, among much else. Work on automated structure prediction systems is motivated in large part by the significant practical costs of experimental and simulation-based methods; whereas these methods can take weeks or months, DL-based systems can produce accurate predictions in seconds (Akdal et al. 2022; Jones & Thornton, 2023). AlphaFold has been used to generate the Protein Structure Database, containing over 214 million structural predictions (Varadi et al., 2024). The significant practical costs associated with solving even a single structure experimentally make it practically impossible to generate this database without the help of DL-based methods. Given an amino acid sequence whose tertiary structure is unknown or only partially known, researchers can circumvent otherwise time- and labour-intensive experimental methods by querying this database to obtain a preliminary structural prediction, compare the target sequence to related sequences, and so forth. Even relatively low-quality predictions can be used as reference points for later modeling and experimental work by pointing researchers away from dead ends.

While automated protein-structure prediction systems are perhaps the most impressive and well-known illustration of this point, similar (albeit less dramatic) dynamics can be observed in many other areas of science (Choudhary et al. 2024). Indeed, many ‘vanilla’ scientific applications of DLMs involve automating routine tasks which would be otherwise extremely time- or labour-intensive.

One might wonder how lowering practical costs can justify pursuit, but this is a straightforward consequence of our decision-theoretic model. Once again, let  $p_{\text{DLM}}(H)$  be the decision to pursue an idea  $H$  with the help of a DLM, and let  $p_{\sim\text{DLM}}(H)$  be the decision to pursue  $H$  without the help of a DLM. If  $\text{EPC}(p_{\sim\text{DLM}}(H)) > \text{EPC}(p_{\text{DLM}}(H))$ , then equation (3) tells us that all else equal,  $P(p_{\sim\text{DLM}}(H)) < P(p_{\text{DLM}}(H))$ . That is, the fact that we can now pursue  $H$  in a way that is faster, more efficient, etc., provides *ceteris paribus* justification to pursue it. Note that there are different ways to arrive at the claim that  $\text{EPC}(p_{\sim\text{DLM}}(H)) > \text{EPC}(p_{\text{DLM}}(H))$ . Most obviously, a DLM might reduce the resources of time, energy, etc., needed to arrive at the decision to accept or reject an idea – i.e., reduce  $\text{PC}(a(H))$  or  $\text{PC}(r(H))$  – but without changing the likelihood of either outcome. This is a natural way to model what’s going on with AlphaFold. Alternatively, it could be that the efficiency of an DLM increases the likelihood of acceptance or rejection and decreases the likelihood of failed inquiry (that is, it decreases the likelihood that pursuit will yield agnosticism).

In these cases, we’ve considered how using a DLM can increase the pursuitworthiness of a single hypothesis. However, as we’ve suggested, one of the most important ways DLMs can impact pursuitworthiness calculations is by impacting the distribution of expected costs and benefits over the set of available ideas. Here is a simple illustration of one way this can happen, inspired by AlphaFold’s impact on structural biology. Suppose we are initially confronted with a range of ideas  $H_1, H_2, \dots, H_{100}$ . Let’s suppose that  $\text{EEV}(p(H_1)) > \text{EEV}(p(H_i))$  and  $\text{EPC}(p(H_1)) = \text{EPC}(p(H_i))$ , for  $i > 1$ . In this case,  $H_1$  is intuitively the most pursuitworthy of the bunch (and our model agrees). Now suppose that a methodological innovation – such as AlphaFold – dramatically lowers the practical cost of pursuing all but  $H_1$ . That is, we now have  $\text{EPC}(p(H_i)) \ll \text{EPC}(p(H_1))$ . In this circumstance, it may be reasonable to pursue the  $H_i$  (or a subset thereof) rather than  $H_1$ , for the simple reason that this would yield a greater overall return on investment given available resources. And this is so even if the  $H_i$  are less epistemically valuable considered in isolation. This is a stark illustration of how lowering the expected practical costs of pursuit can dramatically alter the range of hypotheses scientists have reason to pursue.

#### 4. Why it matters

The purpose of this exercise was to make good on the idea that DLMs *can* justify pursuit. This is important because it helps explain how DLMs can help scientists meet their epistemic goals. In this section, we turn to some broader philosophical issues. We will argue that pursuitworthiness

provides a useful lens for understanding the role of DLMs in contemporary science. One advantage of this lens is that it yields a plausible account of the sense in which DLMs can be said to be ‘revolutionizing’ science. Another is that it brings into sharper focus a largely unexplored family of scientific risks. Our aim in this section is to spell out these advantages and highlight some new questions that emerge from this vantage point.

#### *4.1 Same old stew? Understanding the deep learning revolution*

Much philosophical reflection on DLMs has revolved, in one way or another, around the question of whether DLM-based science introduces genuine novelties or whether the widespread adoption of DLMs might even constitute a ‘revolution’ in science. This reflection is animated in part by a concern, not always fully articulated, that the shift towards DLM-based science does not introduce genuinely novel philosophical issues – that, as far as philosophy of science goes, the philosophical issues raised by deep learning are the ‘same old stew’ (cf. Frigg & Reiss 2009).

In response to this sort of concern, some philosophers seek to locate the revolutionary potential of DLMs in their ability to learn novel kinds of scientific representations. To mention one recent example, Boge (2023) suggests that DLMs’ ability to learn analogues scientific concepts, what Boge calls ‘functional concept proxies’ (FCPs), underlies much of their interest and revolutionary potential:<sup>17</sup>

FCPs ... give us a fairly clear sense of what is special about DNNs in science ... there is a good case that FCPs are responsible, not only for Clever Hans behavior, but also the super-human (or, more generally: unrivaled) performance of DNNs that we are currently witnessing in several domains. Thus, with the current revolution in AI as predominantly brought about by present-day DNNs, we may also be witnessing a shift in the way we do science, as they take us a step away from traditional procedures such as the formation of concepts on which we base theories and models to generate successful explanations and predictions. (Boge 2023, p. 35)

For Boge, the potential for DLMs to learn such analogues is particularly significant in domains where more traditional theoretical and modeling approaches fall short, and in other

---

<sup>17</sup> Boge construes FCPs as vectors in a DLM’s latent space which preferentially activate for certain inputs.

work he offers as an example the search for novel particles in high energy physics (Boge 2022). The expectation behind this story is that if DLMs *do* learn novel concepts – for instance, corresponding to a new particle type – we might be able to extract them and put them to work. For our purposes, the central point to note is that on this sort of approach, the revolutionary potential of DLM-laden science is to a large extent *doctrinal*, providing new ways to discover analogues of novel scientific concepts or other representations – including, perhaps worryingly, representations largely inscrutable to human scientists.

While we don't want to rule out the possibility of a story along these lines, it is not clear that this kind of story can do everything we might want it to. For one thing, it is controversial whether DLMs *do* learn analogues of scientific concepts or other representations (Buckner 2023, Rowbottom et al. 2024). And even if they do, it is questionable whether we can extract them and put them to work in the manner required by Boge's story. As Boge (2024, p. 32) notes, the few cases where scientists seem to have successfully identified (analogues of) concepts learned by DLMs involve antecedently well-understood toy phenomena which we *already* knew how to describe. But perhaps most important obstacle is that the question whether DLMs have learned analogues of scientific concepts is simply *irrelevant* for many of the most important ways DLMs are currently being used. AlphaFold has revolutionized structural biology, and this is so whether or not it has learned (analogues of) novel biochemical concepts. Thus, it seems that we need a story about DLMs' revolutionary potential (or at least one important aspect of that potential) that does *not* revolve around their ability to learn or generate novel concepts or other kinds of representations.

Unsurprisingly, we think that DLMs' ability to radically impact pursuitworthiness calculations must be a central part of this story. Scientists are increasingly confronting problems whose scale and complexity make them ill-suited to more traditional methods (Berens et al. 2023). And decisions about which ideas to pursue (or continue to pursue) emerge at nearly every stage of scientific practice, from day-to-day activities like observation, experimentation, and model-building, to the generation of scientific ideas, to decisions about which research should be published and which projects funded, among much else (Van Noorden & Perkel, 2023; Duede et al. 2024). Given that DLMs are poised to play a role in all these activities, they are fundamentally reshaping the economics of scientific pursuit. On this sort of story, what is distinctive about the impact of DLMs is not (or not merely) something about their internal

workings, but rather that they operate at a speed and scale which can dramatically impact which ideas scientists are justified in pursuing. From this perspective, then, the most distinctive aspects of shift towards DLM-laden science are methodological rather than doctrinal.

One question that arises is whether more general philosophical accounts of scientific change can help us understand this shift. Some commentators suggest that it amounts to a paradigm shift (e.g., Jin et al. 2025), but for reasons already touched on it seems clear that this isn't quite right. Domains in which DLMs have found wide application do not for that reason tend to involve theoretical or conceptual changes, as we noted, yet such doctrinal differences are central to traditional accounts of large-scale scientific structure and change (Kuhn 1962). For this reason, such accounts appear to be ill-equipped to help us understand the move towards DLM-laden science.

We suggest that a more helpful tool in this context is the concept of a scientific *repertoire* (Ankeny & Leonelli 2016). A repertoire, roughly speaking, is a set of strategies for managing the “conceptual, material, and social components” of scientific research in a way that allows them to be “effectively used to acquire the resources, capacities, and expertise needed to pursue an inquiry” (ibid, p. 20). Although adopting a repertoire sometimes involves doctrinal commitments like adopting a model or a theory, it also involves adoption of certain methods and technologies. The strategies in question concern how to coordinate these methods and technologies to meet scientists' epistemic goals, given the material (social, institutional, etc.) forces at play. Repertoires thus foreground the material elements of scientific inquiry in a way that dovetails nicely with our emphasis on the economics of pursuit.

Repertoires offer several advantages for understanding the impact of DLMs on science, and we will note a couple of the more important ones here, deferring a more comprehensive appraisal for future work. First, by foregrounding strategies for coordinating material and technological elements of scientific inquiry, repertoires focus our attention on precisely those aspects of scientific practice where we've argued that DLMs have profound effects: on the allocation of time, resources, and effort. From this angle, the shift to DLM-laden science is transformative to a significant extent because it involves a shift in scientific repertoire – in the strategies scientists use to pursue their ends – rather than a shift in doctrine.

Second, repertoires also help explain why DLMs have enjoyed such rapid uptake in certain parts of science. DLMs are particularly well-positioned to justify pursuit in domains with

an abundance of training data, where the cost of alternative approaches is high, and for problems whose solutions can be characterized in terms of a suitable loss function (Choudhary et al. 2024, Griffin et al. 2024). It is natural to think of the adoption of DLMs in such domains as involving the adoption of new repertoires which coordinate how DLMs slot into preexisting workflows. The repertoire emerging in structural biology surrounding AlphaFold is once again a particularly nice example. Technologies like the AlphaFold Protein Database the AlphaFold Server present new strategies for pursuing hypotheses about protein structure and function. These strategies must coordinate with existing workflows and methods in ways that we are only beginning to understand (Matthiessen forthcoming). Of course, different repertoires will emerge in different cases, and we should expect significant variation depending on local conditions. Much philosophical work remains to be done mapping out these conditions and how they bear on repertoire development.

#### *4.2 Anticipating risks of pursuitworthiness*

Understanding the transformation imposed by a move towards DLMs in science is important not just for understanding the gross structure and operation of DLM-laden science, but also for anticipating a distinctive class of scientific risks that emerge from this transformation – what we call risks of pursuitworthiness. We will start by distinguishing these kinds of risks from a few others before outlining two of the main risks of pursuitworthiness posed by DLMs in science.

On the one hand, the legal and moral risks associated with DLMs are by now reasonably well understood (e.g., Liao 2020). There is also a growing body of work addressing distinctively epistemic risks raised by DLMs – where epistemic risks, roughly speaking, are risks of getting things wrong (Biddle 2016, p. 2020). For instance, in a recent important discussion, Messeri and Crockett argue that DLMs may promote ‘illusions of understanding’, because “[t]here is evidence for considerable overoptimism in scientific claims that are based on machine learning model performance” (2024, p. 53). For them, the concern is that scientists might be misled into thinking that they have more explanatory knowledge or understanding than they actually do, given the simplicity and user-friendliness of certain DLM technologies (see also Boge 2022, Krenn et al. 2022). Other commentators highlight epistemic risks associated with hallucination (Birhane et al. 2023), and the emergence of epistemic monocultures (Andrews et al. 2024), among other things.

However not every scientific risk posed by DLMs is a risk of getting things wrong – of accepting something false. Sometimes, DLMs pose a risk because they encourage scientists to explore avenues which they do not have good reasons to explore. Such *risks of pursuitworthiness* – or pursuit risks, as we’ll sometimes call them – arise when scientists are led to believe that some ideas are more pursuitworthy than they in fact are.

To illustrate the sort of risk we have in mind, consider the phenomenon of ‘state of the art’ (SOTA)-chasing in large language model (LLM) design (Millière, 2024). One of the main ways to measure the capacities of an LLM is to compare its performance against a benchmark – roughly speaking, a set of expected responses for certain inputs. Standard benchmarks measure things like a model’s ability to solve math problems or solve abstract reasoning puzzles (e.g., Chollet 2019). The idea behind this line of work is that a model will perform well on a certain benchmark only if, and to the extent that, it possesses the capacity in question (say, a capacity for abstract reasoning). One challenge is that it is very difficult to determine whether good performance on a benchmark *is* indicative of the capacity in question, and a significant amount of effort is devoted to tuning models to perform well on benchmarks even if the benchmark’s connection to the target capacity is uncertain.<sup>18</sup> Here, then, we may have a pursuit risk: while one might think it worthwhile to pursue increased performance on a given benchmark, this may lead scientists away from their epistemic goals – for instance, if the epistemic benefits of increased performance are in fact minimal – and may waste resources of time, energy, or labour power that could be used more effectively for other ends.

This is, of course, just one example. To provide a more systematic account of how DLMs can contribute to pursuit risks, it will be useful to return to our framework for understanding how DLMs can justify pursuit. We will highlight three ways DLMs can lead to pursuit risks, mirroring the three ways we claim DLMs can justify pursuit, although we do not take these to be exhaustive.

First, consider the claim (Section 3.1) that DLMs can justify pursuit by suggesting ideas which are plausible. The flip side of this is a risk of *spurious plausibility* – of a DLM leading researchers to think that ideas are more plausible – and hence more pursuitworthy – than they in fact are. Uneven coverage in training data, data contamination, unknown deployment distribution

---

<sup>18</sup> A nice example of Goodhart’s ‘law’, that when measures become targets, they cease to be good measures (Manheim & Garrabrant, 2018).

shifts, and undetected shortcut learning can all contribute to this risk, insofar as they can lead scientists to regard DLM outputs as more plausible or better supported than they ought to. A nice example of this risk in practice comes from cases in which AlphaFold is *overconfident*, for instance by representing a region of a protein as having a certain stable configuration when in fact that region is intrinsically disordered (e.g., Pratt et al. 2025). In such cases, a spuriously plausible prediction can make it harder for researchers to reach their epistemic goals and can waste practical resources, for instance by suggesting unhelpful or dead-end experiments or mechanisms.

Second, consider that DLMs can justify pursuit by suggesting fruitful ideas (Section 3.2). Here, we have a corresponding pursuit risk of *illusory fertility*, wherein DLMs lead scientists to regard certain ideas as more fruitful than they in fact are. Given that DLM-, and especially LLM-driven idea generators are in their infancy, this risk is admittedly more speculative than the last one. Nevertheless, given that these systems have been touted as a promising way to identify novel research ideas, this risk is potentially severe. And there is some evidence to suggest that the ideas identified by these systems are not as fruitful as they might initially appear. For instance, Si et al. (2025) found that LLM-generated ideas in natural language processing were systematically regarded as less fruitful than expert-generated ideas even when implemented by expert researchers. Here, too, the concern is that illusory fertility can lead scientists further away from their epistemic goals rather than towards them. In these cases, we risk more than just the practical costs of pursuit. At their worst, DLMs could encourage pursuing ideas that not only come to nothing, but in fact prove harmful – projecting a positive  $EEV(p(H))$  when  $EV(H)$  proves negative.

Finally, we saw that DLMs can justify pursuit by driving down practical costs when all else is held fixed (Section 3.3). In this case, we might call the corresponding risk one of *efficiency traps*, in that the perceived speed and availability of DLMs can also introduce perverse incentives into decisions about what to pursue or how to apportion scarce scientific resources.<sup>19</sup> One aspect of this is the concern that the ease with which these systems can be deployed and the speed with which they can generate ideas may lead to situations in which convenience

---

<sup>19</sup> Here we bracket the moral aspects of practical costs, and so do not discuss a further problem, that of *illusory efficiency*. While a scientific community may be able to use a DLM cheaply, our current formulation does not consider costs external to this community required for the creation and maintenance of the DLM, costs paid by other segments of society.

overwhelms epistemic value: scientists may be led to pursue low-quality or useless ideas, just because they can. A second, somewhat subtler concern is that decisions to pursue ideas which are efficient in the short term may lead to greater inefficiencies down the line. For instance, both concerns seem to underlie renewed calls in structural biology to better understand folding mechanisms rather than resting content with powerful predictive methods (Chen et al. 2023). Here, the concern is that if we do not continue to devote resources towards understanding folding mechanisms, future structural biology will be impoverished with fewer and worse theoretical resources, and that it can be practically difficult to escape from such a situation.

Evidently, much more could be said about each of these risks, but we think this is enough to get a feel for how they work. And while we don't want to overstate these risks – we don't want to be alarmist – we nevertheless think that these risks are potentially serious. And although much philosophical and metascientific work remains to be done to gauge their impact and the extent to which these arise in practice, highlighting them at this early stage is important for considering ways they might be ameliorated going forward.

## **5. Conclusion**

What, if anything, is distinctive about the shift towards DLM-laden science? In this article, we've built on recent discussions which address the impact of DLMs on scientific pursuit (Duede 2023, Finn & Khosrowi 2024). First, we've explained on principled, general grounds how DLMs can justify pursuit, either by increasing an idea's expected epistemic value or decreasing the practical costs of pursuing it. Second, we've argued that much of the revolutionary potential of DLMs should be not be understood principally in terms of doctrinal shifts, but rather in terms of their potential to impact the economics of pursuit, by introducing scientific repertoires which are fast, flexible, and domain-neutral. Third, we've argued that addressing the DLM revolution in science through the lens of pursuitworthiness brings a new class of scientific risks into sharper focus. While we've seen that it's possible for DLMs to rationally guide and accelerate scientific inquiry, they also introduce *risks of pursuitworthiness*, in which DLMs promote pursuit of low-quality, bogus, or even pseudoscientific ideas. Understanding these risks – and how to ameliorate them – will be a key item on the agenda for theorists going forward.

## References

- Achinstein, P. (2014). Evidence. In S. Psillos & M. Curd (Eds.), *The Routledge companion to philosophy of science* (pp. 381–392). New York: Routledge.
- Akdel et al. (2022). A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, 29(11), 1056–1067.  
<https://doi.org/10.1038/s41594-022-00849-w>
- Andrews et al. (2024). The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, 5(9). <https://doi.org/10.1016/j.patter.2024.101027>
- Baek et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871–876. <https://doi.org/10.1126/science.abj8754>
- Barman, Kristian G., Sascha Caron, Tom Claassen, and Henk de Regt. 2024. “Towards a Benchmark for Scientific Understanding in Humans and Machines”. *Minds & Machines* 34:6. <https://doi.org/10.1007/s11023-024-09657-1>
- Berens, P., Cranmer, K., Lawrence, N. D., Luxburg, U. von, & Montgomery, J. (2023). AI for Science: An Emerging Agenda (arXiv:2303.04217). arXiv.  
<https://doi.org/10.48550/arXiv.2303.04217>
- Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5), 277–280.  
<https://doi.org/10.1038/s42254-023-00581-4>
- Biddle, J. B. (2016). Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease. *Perspectives on Science*, 24(2), 192–205. [https://doi.org/10.1162/POSC\\_a\\_00200](https://doi.org/10.1162/POSC_a_00200)
- Binz et al. (2024). Centaur: a foundation model of human cognition. arXiv preprint arXiv:2410.20268.
- Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32, 43–75. <https://doi.org/10.1007/s11023-021-09569-4>
- Boge, F. J. (2023). Functional Concept Proxies and the Actually Smart Hans Problem: What’s Special About Deep Neural Networks in Science. *Synthese*, 203(1), 16.  
<https://doi.org/10.1007/s11229-023-04440-8>
- Buckner, C. J. (2023). *From Deep Learning to Rational Machines*. Oxford University Press.
- Champion, H. (forthcoming). Strong Novelty Regained: High-Impact Outcomes of Machine Learning for Science. *Synthese*.

- Chen et al. (2023). Protein folds vs. protein folding: Differing questions, different challenges. *Proceedings of the National Academy of Sciences*, 120(1), e2214423119.  
<https://doi.org/10.1073/pnas.2214423119>
- Chirimuuta, M. (2020). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*. <https://doi.org/10.1007/s11229-020-02713-0>
- Chollet, F. (2019). On the Measure of Intelligence (arXiv:1911.01547). arXiv.  
<https://doi.org/10.48550/arXiv.1911.01547>
- Clark, E., & Khosrowi, D. (2022). Decentring the discoverer: How AI helps us rethink scientific discovery. *Synthese*, 200(6), 463. <https://doi.org/10.1007/s11229-022-03902-9>
- Djorgovski, D. G., Mahabal, A. A., Krone-Martins, A., & Graham, M. J. (2023). Applications of AI in Astronomy. In A. N. Choudhary, G. C. Fox, & A. J. G. Hey (Eds.), *Artificial intelligence for science: A deep learning revolution* (pp. 81–93). World Scientific.
- Duede, E. (2022). Instruments, agents, and artificial intelligence: Novel epistemic categories of reliability. *Synthese*, 200(6), 491. <https://doi.org/10.1007/s11229-022-03975-6>
- Duede, E. (2023). Deep Learning Opacity in Scientific Discovery. *Philosophy of Science*, 90(5), 1089–1099. <https://doi.org/10.1017/psa.2023.8>
- Duede, E., Dolan, W., Bauer, A., Foster, I., & Lakhani, K. (2024). Oil & Water? Diffusion of AI Within and Across Scientific Fields (arXiv:2405.15828). arXiv.  
<https://doi.org/10.48550/arXiv.2405.15828>
- Duerr, P. M., & Fischer, E. (2025). Rationally warranted promise: The virtue-economic account of pursuit-worthiness. arXiv. <https://doi.org/10.48550/ARXIV.2501.05142>
- Durán, J. M. (2025). Beyond transparency: Computational reliabilism as an externalist epistemology of algorithms (arXiv:2502.20402). arXiv.  
<https://doi.org/10.48550/arXiv.2502.20402>
- Durán, J. M., & Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, 28(4), 645–666.  
<https://doi.org/10.1007/s11023-018-9481-6>
- Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, 202(4), 109. <https://doi.org/10.1007/s11229-023-04334-9>
- Frigg, R., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese*, 169(3), 593–613.

- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673. <https://doi.org/10.1038/s42256-020-00257-z>
- Gross, F. (2024). The Explanatory Role of Machine Learning in Molecular Biology. *Erkenntnis*. <https://doi.org/10.1007/s10670-023-00772-6>
- Gu, X., & Krenn, M. (2025a). Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *Machine Learning: Science and Technology*. <https://doi.org/10.1088/2632-2153/add6ef>
- Gu, X., & Krenn, M. (2025b). Interesting Scientific Idea Generation using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders (arXiv:2405.17044). arXiv. <https://doi.org/10.48550/arXiv.2405.17044>
- Iten, R., Metger, T., Wilming, H., del Rio, L., & Renner, R. (2020). Discovering physical concepts with neural networks. *Physical Review Letters*, 124(1), 010508. <https://doi.org/10.1103/PhysRevLett.124.010508>
- Jin, li, & et al. (2025). AI for Science 2025. Retrieved October 8, 2025, from <https://www.nature.com/articles/d42473-025-00161-3>
- Jones, David. T., & Thornton, J. (2023). AlphaFold—The End of the Protein Folding Problem or the Start of Something Bigger? In A. N. Choudhary, G. C. Fox, & A. J. G. Hey (Eds.), *Artificial intelligence for science: A deep learning revolution* (pp. 67–80). World Scientific.
- Jumper et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Khosrowi, D. (2025). Automating Pursuitworthiness: Four Concerns About The Proper Roles for Machine Learning in Scientific Discovery [Preprint]. <https://philsci-archive.pitt.edu/26889/>
- Khosrowi, D., & Finn, F. (2024). Can Generative AI Produce Novel Evidence? <https://philsci-archive.pitt.edu/24201/>
- Kitcher, P. (1995). *The Advancement of Science: Science Without Legend, Objectivity Without Illusions* (1st ed.). Oxford University Press New York. <https://doi.org/10.1093/0195096533.001.0001>

- Kovalevskiy, O., Mateos-Garcia, J., & Tunyasuvunakool, K. (2024). AlphaFold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, 121(34), e2315002121. <https://doi.org/10.1073/pnas.2315002121>
- Krenn et al. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12), 761–769. <https://doi.org/10.1038/s42254-022-00518-3>
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. Univ. of Chicago Press.
- Ivani, S. (2019). What we (should) talk about when we talk about fruitfulness. *European Journal for Philosophy of Science*, 9(1), 4. <https://doi.org/10.1007/s13194-018-0231-7>
- Langley, P. W., Simon, H. A., Bradshaw, G., & Zytkow, J. M. (1987). *Scientific Discovery: Computational Explorations of the Creative Process*. The MIT Press. <https://doi.org/10.7551/mitpress/6090.001.0001>
- Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth* (1st paperback print). Univ. of Calif. Press.
- Lakatos, I. (1999). *The methodology of scientific research programmes* (J. Worrall & G. Currie, Eds.). Cambridge University Press.
- Liao, S. M. (Ed.). (2020). *Ethics of artificial intelligence*. Oxford University Press. <https://doi.org/10.1093/oso/9780190905033.001.0001>
- Manheim, D., & Garrabrant, S. (2019). Categorizing Variants of Goodhart’s Law (arXiv:1803.04585). arXiv. <https://doi.org/10.48550/arXiv.1803.04585>
- Matthiesen, D. (forthcoming). “Understanding Phenomena Through Opaque Models: Machine Learning and Integrative Practices in Structural Biology” in D. Rowbottom, A. Curtis-Trudel, and D. Barack (eds), *The Role of AI in Science: Epistemological and Methodological Studies*, Routledge.
- Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002), 49–58. <https://doi.org/10.1038/s41586-024-07146-0>
- Millière, R. (2024). Philosophy of cognitive science in the age of deep learning. *WIREs Cognitive Science*, 15(5), e1684. <https://doi.org/10.1002/wcs.1684>
- Nickles, T. (2006). Heuristic Appraisal: Context of Discovery or Justification? In J. Schickore & F. Steinle (Eds.), *Revisiting Discovery and Justification* (Vol. 14, pp. 159–182). Kluwer Academic Publishers. [https://doi.org/10.1007/1-4020-4251-5\\_10](https://doi.org/10.1007/1-4020-4251-5_10)

- Nickles, T. (2020). Alien Reasoning: Is a Major Change in Scientific Research Underway? *Topoi*, 39(4), 901–914. <https://doi.org/10.1007/s11245-018-9557-1>
- Nickles, T. (2022). Whatever Happened to the Logic of Discovery? From Transparent Logic to Alien Reasoning. In W. J. Gonzalez (Ed.), *Current Trends in Philosophy of Science* (Vol. 462, pp. 81–102). Springer International Publishing. [https://doi.org/10.1007/978-3-031-01315-7\\_5](https://doi.org/10.1007/978-3-031-01315-7_5)
- Nyrup, R. (2015). How Explanatory Reasoning Justifies Pursuit: A Peircean View of IBE. *Philosophy of Science*, 82(5), 749–760. <https://doi.org/10.1086/683262>
- Nyrup, R. (2020). ‘Of Water Drops and Atomic Nuclei: Analogies and Pursuit Worthiness in Science.’ *The British Journal for the Philosophy of Science*, 71(3), 881–903. <https://doi.org/10.1093/bjps/axy036>
- Nyrup, R. (ms). A Pursuit Worthiness Account of Analogies in Science. <https://philsci-archive.pitt.edu/12577/>
- Peirce, C. S. (1932). *Collected Papers of Charles Sanders Peirce, 1931-1935* Edited by Charles Hartshorne and Paul Weiss. Vol. 1–6. 8 vols.
- Peirce, C. S. (1992). *The Essential Peirce: Selected Philosophical Writings. Vol. 2: 1893-1913*. Edited by Nathan Houser, Christian J. W. Kloesel, and Peirce Edition Project. Bloomington: Indiana University Press.
- Peirce, C. S. (1998). *Writings of Charles S. Peirce. Vol. 4: 1879 - 1884*. Ed. Christian J. W. Kloesel. Bloomington: Indiana Univ. Press.
- Peirce, C. S. (1999). *Writings of Charles S. Peirce. Vol. 1: 1857 - 1866*. Ed. Max Harold Fisch. Bloomington: Indiana Univ. Press.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547), 1209–1214. <https://doi.org/10.1126/science.abe2629>
- Pratt et al. (2025). AlphaFold 2, but not AlphaFold 3, predicts confident but unrealistic  $\beta$ -solenoid structures for repeat proteins. *Computational and Structural Biotechnology Journal*, 27, 467–477. <https://doi.org/10.1016/j.csbj.2025.01.016>
- Räz, T., & Beisbart, C. (2022). The Importance of Understanding Deep Learning. *Erkenntnis*. <https://doi.org/10.1007/s10670-022-00605-y>

- Rescher, N. (1989). *Cognitive Economy: The Economic Dimension of the Theory of Knowledge*. University of Pittsburgh Press.
- Rowbottom, D. P., Peden, W., & Curtis-Trudel, A. (2024). Does the no miracles argument apply to AI? *Synthese*, 203(5), 173. <https://doi.org/10.1007/s11229-024-04524-z>
- Šešelja, D., & Straßer, C. (2014). Epistemic justification in the context of pursuit: A coherentist approach. *Synthese*, 191(13), 3111–3141. <https://doi.org/10.1007/s11229-014-0476-4>
- Si, C., Hashimoto, T., & Yang, D. (2025). The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas (arXiv:2506.20803; Version 1). arXiv. <https://doi.org/10.48550/arXiv.2506.20803>
- Sikimić, V., & Radovanović, S. (2022). Machine learning in scientific grant review: Algorithmically predicting project efficiency in high energy physics. *European Journal for Philosophy of Science*, 12(3), 50. <https://doi.org/10.1007/s13194-022-00478-6>
- Simon, H. A. (1988). *Models of Discovery*. Reidel.
- Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1), 109–133. <https://doi.org/10.1093/bjps/axz035>
- Terwilliger et al. (2023). AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*. <https://doi.org/10.1038/s41592-023-02087-4>
- Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature*, 621(7980), 672–675. <https://doi.org/10.1038/d41586-023-02980-0>
- Varadi et al. (2024). AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1), D368–D375. <https://doi.org/10.1093/nar/gkad1011>
- Vervoort, L., Shevlin, H., Melnikov, A. A., & Alodjants, A. (2023). Deep Learning Applied to Scientific Discovery: A Hot Interface with Philosophy of Science. *Journal for General Philosophy of Science*, 54(2), 339–351. <https://doi.org/10.1007/s10838-022-09625-2>
- Wible, J. R. (1994). Charles Sanders Peirce's economy of research. *Journal of Economic Methodology*, 1(1), 135–160. <https://doi.org/10.1080/13501789400000009>
- Wible, J. R. (1998). *The Economics of Science*. Routledge.
- Zakharova, D. (ms). *The Epistemology of AI-driven Science: The Case of AlphaFold*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 107–115.  
<https://doi.org/10.1145/3446776>