From Hallucination to Reliability: Generative Modeling and the Structure of Scientific Inference

Charles Rathkopf

January 2026

> "But in the practice of science, knowledge is an affair of *making* sure, not of grasping antecedently given sureties."
>
> John Dewey[1]

[1] [Dewey, 1958, p. 154]. [Emphasis in original.]

Abstract

Generative AI increasingly supports scientific inference, from protein structure prediction to weather forecasting. Yet its distinctive failure mode, *hallucination*, raises epistemic alarm bells. I argue that this failure mode can be addressed by shifting from data-centric to phenomenon-centric assessment. Through case studies of AlphaFold and GenCast, I show how scientific workflows discipline generative models through theory-guided training and confidence-based error screening. These strategies convert hallucination from an unmanageable epistemic threat into bounded risk. When embedded in such workflows, generative models support reliable inference despite opacity, provided they operate in theoretically mature domains.

Word count: 8,331 **Keywords:** generative AI, hallucination in AI, scientific AI, AI reliability, reliabilist epistemology, AI epistemology, protein structure prediction, weather forecasting

1 Hallucination as a threat to reliability

In recent years, generative AI has become deeply embedded in scientific practice. It is now used to synthesize data for climate models [Kadow et al., 2020], to map phase transitions in novel materials [Arnold et al., 2024], and to predict molecular interactions for drug discovery [Sidhom et al., 2022]. Unlike classificatory AI models, generative AI models produce outputs that are highly detailed and informationally rich. That richness makes them epistemically valuable, but also leaves them susceptible to a new kind of error that has come to be known as "hallucination" [Ji et al., 2023; Sun et al., 2024].

As a first pass, hallucinations[2] can be characterized as errors that are not merely inherited from the training data, but are, in some sense, produced by the model itself. This claim is substantive: not all errors count as hallucinations. A corrupted measurement or a mislabeled datapoint is an error, but not a hallucination. I return to a more precise analysis in Section 3, but even this initial sketch helps explain why the phenomenon demands attention. It is natural to worry that any model prone to hallucination may not be trustworthy—and indeed, the epistemic risks are serious. AlphaFold 3, among the most celebrated generative models in science, has been shown to produce detailed molecular structures where none exist [Abramson et al., 2024]. GANs used in medical imaging have introduced phantom anomalies—a fracture-like line in an unbroken bone, or a lesion in healthy tissue [Shin et al., 2021].

These are not just rounding errors. Undetected, hallucinations can lead researchers and clinicians toward serious mistakes in inference and decision-making. In fact, the epistemic challenges posed by hallucinations run deeper than these examples suggest. There is reason to think that hallucinations are *inevitable* byproducts of the mechanisms of generative inference. The intuition behind this claim is that training such models involves a fundamental tradeoff between novelty and reliability [Sajjadi et al., 2018; Sinha et al., 2023; Xu et al., 2024]. A model constrained to strictly mirror its training data may be reliable but incapable of generating novel insights. Allowing a model to extrapolate, by contrast, enables novelty but invites fabrication.

Another reason that hallucinations threaten reliability is that they are sometimes difficult to detect [Ji et al., 2023; Bubeck et al., 2023]. This is not always the case. When we have thorough background knowledge of the target phenomenon, hallucinations can be easy to spot. For example, earlier versions of DALL-E and Stable Diffusion often

[2] The term invites a misleading comparison to human perceptual experience. Generative AI models do not consciously perceive the world, let alone misperceive it. Nevertheless, since the term is already widely used in the technical literature, insisting on a replacement would only introduce a cognitively costly neologism into an already difficult discussion.

generated images of human hands with six fingers [Wang et al., 2024]. But scientific AI operates at the frontiers of human knowledge, where error detection is intrinsically more difficult. Where our background knowledge is weakest, errors are most likely to go undetected. And the longer they remain undetected, the more they threaten to derail any decision-making processes based on model outputs.

Drawing these observations together, it seems that hallucinations are, in at least some cases, *substantive*, *inevitable*, and *difficult to detect*. Worse still, deep learning models are epistemically opaque [Humphreys, 2009; Creel, 2020]. In traditional closed-form models, evidence of reliability is often grounded in knowledge of how parameters relate to the properties of the target system. When errors arise, they can be traced to specific parameters and corrected. In DNNs, by contrast, it is unclear whether individual parameters represent anything at all.

This combination of opacity and model-generated error creates what I will call *the diagnostic problem*: once we discover a hallucination downstream, how can we use that knowledge to systematically improve the model? In normal scientific practice, error correction is iterative. When a model produces questionable results, researchers trace the error back to specific parameters, adjust them based on their representational role, and thereby reduce the probability of similar errors in future applications. Opacity blocks precisely this diagnostic workflow. Without knowing which parameters are responsible for an error, we cannot learn from our mistakes in the systematic way science typically demands. If we cannot check the representational fidelity of individual parameters, how might we justify using these models at all? Reliabilist epistemology [Goldman, 1979; Lyons, 2019] offers a straightforward answer: *we observe its track record.* Instead of explaining why a model succeeds, we infer its reliability from past performance. This approach, which Duede [2023] calls *brute inductivism*, reduces scientific epistemology to an accounting exercise.

Suppose a model achieves high accuracy on benchmarks or aligns well with historical data. Researchers then infer—perhaps naively—that the model will be reliable in future applications. But as Duede's unflattering label suggests, brute inductivism is an inherently precarious strategy. Past success offers no guarantee of future performance, particularly in novel settings [Grote et al., 2024]. Nevertheless, these models have already demonstrated their ability to outperform traditional approaches in all sorts of important predictive tasks. The real challenge, then, is not *whether* generative AI should be used in science, but how it can be used responsibly.

Addressing this challenge requires greater clarity about what counts as a hallucination. Existing definitions, whether formal or informal, tend to evaluate model outputs primarily by their relationship to training data. But, as I will argue in

Section 3, this *data-centric* approach forecloses the very solutions that make successful applications possible. If we assess hallucinations by their deviation from training data, we are led toward filtering strategies that would eliminate precisely those outputs where genuine scientific discovery occurs. What matters for scientific reliability is not whether outputs deviate from training data, but whether they misrepresent the target phenomena we aim to understand. Shifting to this *phenomenon-centric* view reveals that many outputs flagged as hallucinations under data-centric definitions pose no genuine epistemic threat, while others—those that cannot be reliably detected or filtered—demand more careful management.

To illustrate how scientists address the diagnostic problem, I examine two case studies: AlphaFold 3, which predicts molecular structures, and GenCast, which generates probabilistic weather forecasts. These models operate in entirely different scientific domains—one at the scale of molecules, the other at the scale of planetary weather systems. Nevertheless, both mitigate hallucinations by embedding theoretical constraints and uncertainty management strategies directly into their modeling architectures. These principles do not eliminate model-generated error entirely, but they show how, despite the distinctive challenges posed by generative AI, such errors can be effectively managed. Crucially, these design principles are neither automatic nor inevitable. They emerge from carefully managed scientific workflows, and their effectiveness depends on deliberate design and maintenance. By articulating the rationale behind these strategies, I aim to clarify how generative AI can be integrated into scientific practice without unduly compromising reliability.

2 On the inevitability of hallucination in generative AI

2.1 What is generative AI?

The term *generative AI* is sometimes taken to refer to any AI system that mimics the cultural products of human creativity. While many generative models do exactly that, mimicking human output is just one of many ways these architectures can be deployed. They are also used to produce numerical, physical, and scientific data of all kinds. Here is a definition that is sufficiently abstract to capture this broader scope:

> A *generative AI model* is a machine learning system trained to produce complex data structures that adhere to patterns learned from training data, while generalizing beyond the exact instances in that data.

The word “complex” is carrying a lot of weight. The complexity of model outputs plays a central role in understanding both why hallucinations are inevitable and why they pose a distinctive epistemic threat in scientific applications. Here, “complexity” refers to high dimensionality: model outputs are structured, multi-component entities rather than scalar values or discrete labels. In many generative models, output dimensionality is proportional—either strictly or approximately—to that of the exemplars in the training data. In some cases, such as *GenCast*, input and output have equal and fixed dimensionality by design, since both represent meteorological fields over a grid on the Earth’s surface. In others, such as autoregressive language models, outputs may exceed the length or complexity of the inputs (e.g., “Write me an essay about the history of AI”). Even then, they remain bounded by architectural constraints, such as context windows and maximum token length, and shaped by the complexity and scale of the training data.

In both kinds of case, the generative task involves producing plausible outputs in a high-dimensional space whose structure is incompletely determined by the training distribution.

Two contrasts help clarify what makes generative AI distinctive. First, unlike classification models, which learn to represent the conditional distribution $P(Y \mid X)$ over discrete labels Y, generative models aim to learn a representation of the full distribution $P(X)$, enabling them to produce novel samples that extend the distribution in coherent ways [Kingma and Welling, 2013; Goodfellow et al., 2014; Buckner, 2024]. Second, unlike classical generative statistical models such as Poisson processes or Markov chains, which generate data from predefined parametric distributions [Grimmet and Sterzaker, 1992], generative AI models learn latent representations that capture complex, often idiosyncratic statistical structure [Rezende et al., 2014; Yang et al., 2023]. This capacity makes them uniquely valuable for domains where explicit theory remains incomplete, such as materials science or drug discovery.

2.2 Inevitability arguments

There is a growing literature on the inevitability of hallucination in large language models. Xu et al. [2025] appeal to no-free-lunch theorems, Banerjee and Jacob [2024] draw an analogy to Gödel’s first incompleteness theorem, and Kalai and Vempala [2024] provide an information-theoretic lower bound on hallucination frequency. But these arguments focus on autoregressive architectures and do not transfer straightforwardly to the scientific models addressed in this paper.

Unlike autoregressive architectures, which are well suited to sequential data such as text, many scientific generative models are designed to preserve global coherence across high-dimensional structures. Diffusion models, in particular, generate outputs through a process of *global iterative refinement*, progressively denoising a sample over multiple steps [Song et al., 2021]. Variational autoencoders (VAEs) and generative adversarial networks (GANs), though architecturally distinct, pursue the same end: to reconstruct complex global structure from sparse data. These models are typically applied in domains where long-range dependencies span multiple spatial or structural dimensions—protein folding, weather dynamics, material synthesis. I focus on these architectures for two reasons. First, they have figured centrally in some of the most celebrated successes of generative AI in the natural sciences. Second, because their outputs are not merely large but genuinely high-dimensional—structured across space, geometry, or topology—the detection of hallucination poses distinct challenges. Unlike language models, which produce long sequences of discrete tokens, these models (often) generate high-dimensional outputs, in which hallucination detection is intrinsically more difficult. In what follows, therefore, I borrow some ideas from inevitability arguments developed for language models to this broader class of scientific models.

One kind of argument is broadly information-theoretic. The idea is that generative AI models do not contain enough information to represent complex empirical distributions accurately. To see this, consider the size of the output space relative to the model's internal parameter space. Generative models operate in high-dimensional output spaces, with far more possible configurations than any dataset can sample faithfully. For example, a 12-megapixel image with 256 intensity levels per channel has $10^{86,000,000}$ possible configurations. A 100-amino acid protein has 10^{130} possible sequences, not counting conformational variants [Dryden et al., 2008]. Meteorological models, for example, must track millions of degrees of freedom—combinations of pressure, temperature, humidity, and other variables across thousands of spatial and temporal points. Even the largest training sets cover only a vanishing fraction of these spaces. Moreover, models compress these sparse samples into relatively small parameter sets. A protein diffusion model may train on a few hundred thousand examples, but must generalize across 10^{100} possible sequences and conformations. This compression all but ensures that many outputs will be generated in regions where the training data provides little guidance. And where the training data provides little guidance, hallucination is inevitable.

A second argument is geometric. Generative models learn a mapping from high-dimensional data to latent representations and generate new outputs by sampling and decoding from this space. But in high-dimensional settings, geometric properties become unintuitive. As Arjovsky et al. [2017] note, real data typically lie on low-dimensional manifolds within a much larger ambient space. When generative models are trained on multiple such manifolds, interpolation in latent space can result in outputs that fall *between* those manifolds. These inter-manifold regions are unsupported by the training distribution. When a model samples from these regions, the result is a hallucination.

Both arguments suggest that generative models are destined to produce outputs that are, in some sense, wrong. But they say nothing about another property that is both commonly associated with hallucination, and important in thinking about generative AI in scientific contexts: superficial plausibility. Even if hallucinations are inevitable, they would not pose much of a threat if they were easy to detect. Unfortunately, when scientific models are operating at the frontier of human knowledge, they are not. Here is one way to think about why.

Generative models tend to capture short-range dependencies more faithfully than long-range ones. This reflects a basic statistical fact: the nearer the elements, the clearer the pattern. Local structures such as bond angles in molecules or temperature gradients in weather fields recur with high signal and low variation. Long-range dependencies, by contrast, are more easily obscured by noise. They often involve more subtle or indirect interactions, and models may lack both the capacity to represent them and the training data to learn them reliably. As a result, generative models, whether built on diffusion processes or transformers, often produce outputs that are locally plausible but globally flawed. A protein may contain chemically sound fragments yet fold into an unstable conformation. A weather forecast may model regional dynamics with precision while violating large-scale conservation laws. Large language models exhibit a parallel tendency: they produce coherent sentences and paragraphs that fail to cohere at the level of extended argument. In each case, local plausibility masks more distributed structural flaws.

These considerations motivate a general conclusion: any generative model that aims to produce complex, structured data will sometimes produce hallucinations. Moreover, contrary to what the recent success of AI scaling laws might suggest, even massive increases in the size of the training data will not make hallucinations of this kind go away.

2.3 Hallucination in diffusion models, and a proposed solution

This conclusion is reinforced by more targeted empirical work on diffusion models. A recent study by Aithal et al. [2025] provides the first detailed characterization of hallucination in these models. Their analysis of the problem, along with their proposed mitigation strategy, offers a useful point of contrast with the account I will develop.

First, a word about diffusion models themselves. These models are trained by corrupting data through a forward process that gradually adds Gaussian noise over many steps, until the data is nearly indistinguishable from pure noise. The model then learns to reverse this process by denoising: at each step, it estimates how the noisy data point should be adjusted to make it more likely under the original data distribution. This adjustment is governed by the *score function*, defined as the gradient of the log-density of the data distribution with respect to the input. Rather than learning the data distribution directly, diffusion models are trained to approximate this score function. But neural networks tend to learn *smooth* approximations of it, even when the true function contains sharp discontinuities. As Aithal et al. emphasize, this smoothness leads to interpolations across low-density regions which, in turn, leads to hallucinations.[3]

Aithal et al. train diffusion models on synthetic datasets specifically designed to make the structure of the data manifold transparent. In one experiment, the training data is sampled from a mixture of eight well-separated Gaussians, with each point drawn from a single mode. In another, they use binary 10×10 grids, constrained so that exactly half the cells are activated according to simple structural rules. The purpose of these setups is to ensure that the generative principles underlying the training data are fully known. This allows them to test whether a diffusion model can learn those principles without producing spurious outputs.

They build on the same basic intuition as the preceding arguments: hallucinations arise when a model generates samples in low-density regions of the learned distribution. To make this idea precise, they introduce a threshold-based definition of hallucination:

$$H_\epsilon(q) = \{x \mid q(x) \leq \epsilon\} \tag{1}$$

Here, $q(x)$ is the model's estimated probability density at output x, and ϵ is a small threshold. Samples that fall in regions of low density, such as the space between well-supported modes, are flagged as hallucinations.

Aithal et al. also offer a proposed solution to the problem of hallucination, and it is underwritten by their formal definition. They introduce a distance metric (operationalized via the variance of the model's prediction $\hat{x}_0$ as a proxy for density) that quantifies how far a generated output strays from known high-density regions in the training data. Any sample falling below the threshold—i.e., in the set $H_\epsilon(q)$—is discarded. By adjusting ϵ, they aim to eliminate

[3] The score function of a distribution $q(x)$ is defined as the gradient of its log-density: $\nabla_x \log q(x)$. This function reflects how the probability density changes near a given point. In many real-world distributions—especially those with multiple distinct modes—the log-density may change abruptly between regions, resulting in sharp transitions or discontinuities in the score function. But neural networks tend to approximate this function in a smooth and continuous way, which causes them to interpolate across gaps between modes. For further explanation, see Luo [2022] or Song et al. [2021].

hallucinations while retaining most legitimate outputs. According to their evaluation, this method removes 95% of hallucinatory samples while preserving the vast majority of in-distribution outputs.

This filtering solution is entirely reasonable when applied to synthetic datasets where the data-generating process is both simple and fully specified, and designed to test model behavior rather than to represent an external phenomenon of scientific interest. However, in a scientific setting, our primary goal is to acquire information about the nature of the target. Specifically, the aim is to discover hidden structure that is not explicitly represented in the training data. But here, we encounter a fundamental difficulty with the conception of hallucination we have employed thus far: a generated output that falls between known modes may signal *either* a modeling error *or* a discovery—an instance of structure that the training data failed to make explicit. In the context of scientific inquiry, then, Aithal et al. [2025]'s solution is too conservative: it eliminates precisely those cases where the most interesting knowledge might emerge. If hallucination were simply a matter of deviation from training data, then every output that deserves to be called a genuine discovery would, *ipso facto*, count as a hallucination, and nearly all of those would get filtered out.

Scientific generative models are effectively engaged in inductive inference. And, as the logicians say, inductive inference is *ampliative*. Yet this ampliative capacity leaves us with a problem: when does generalization constitute genuine scientific discovery, and when does it constitute hallucination? Answering this requires shifting attention from the training data toward the empirical target phenomena. So we need an account of hallucination centered on how model outputs inform (or mislead us about) the target system itself.

3 Rethinking hallucination

Before developing an improved analysis of hallucination, I want to zoom out briefly and draw a parallel between the data-centric attitude that seems prevalent in AI today and a similar attitude that prevailed in 20th-century philosophy of science. The logical empiricists (especially Carnap [1928] in the *Aufbau*) viewed science as the reconstruction of observational data. That view withered under criticism, but the underlying idea that theories earn legitimacy only by recovering or predicting patterns in data of some sort seems to have been widely accepted well past the middle part of the twentieth century. But as But as Bogen and Woodward [1988] forcefully argued, this outlook fails to account for the constructed nature of data. Because most of the data sets that scientists work with are shaped by the contingencies of measurement techniques and experimental design, scientific reasoning necessarily involves questions about how data can sometimes mislead us about the nature of the target phenomenon. To put this thought in slogan form, the data are a means to an end, rather than an end in themselves.

Once we accept that the goal of science is not fidelity to the data but fidelity to the phenomenon, we arrive at a different picture of how generative AI models ought to be assessed. A model's training data does not define the limits of its validity. What matters is whether its outputs illuminate the target. This idea is visualized in Figure 1. The relationship between a model's output and the training data is one of statistical resemblance; the relationship between the output and the target phenomenon is representational. In the introduction I said that a hallucination is an *error* that is produced by the model (rather than one that is inherited from the training data.) An error is a deviation from some standard, and the picture above makes it clear that the target phenomenon, rather than the training data, is the relevant standard. But now we should also ask: what kind of deviations count as errors? That question has no straightforward answer because what counts as an error depends in part on the interpretive practices of the relevant scientific community.

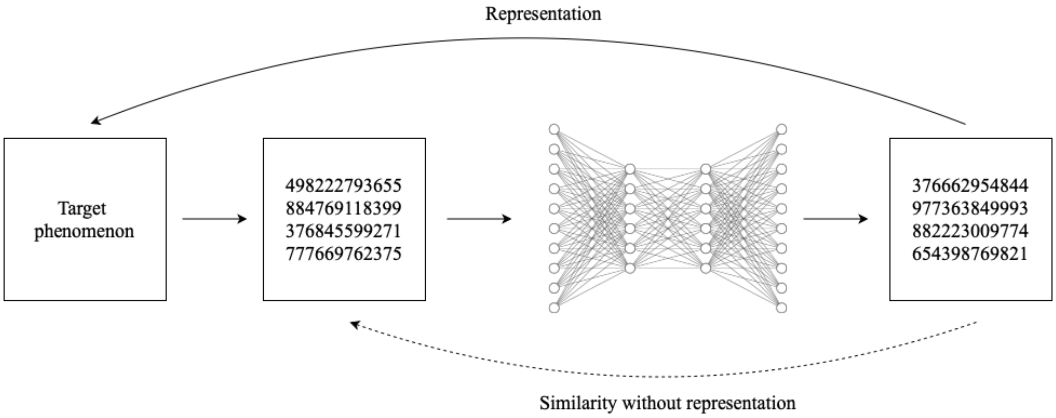

Figure 1: Diagram illustrating the relationship between a target phenomenon, a dataset constructed from observations of the target (second box), a generative deep neural network (DNN), and the DNN's output. The DNN produces outputs that resemble samples from the training data but do not represent them. Whether an output functions as a representation depends on our inferential practices, and in scientific contexts, these practices are aimed at understanding the target phenomenon—not merely reconstructing the training data. The backward arrow ("similarity without representation") indicates that while the model output may exhibit statistical similarity to training data, it is not used as a representation of the training data itself. Rightward arrows indicate causal rather than representational relations.

To see why interpretation matters, consider a non-scientific example. The website `thesecatsdonotexist.com` produces realistic images of cats using a StyleGAN model trained on photographs of real cats. Now suppose you were learning about cats from a sequence of images that included both real photographs and outputs from this model. According to the canonical interpretive scheme for photographs, according to which they depict particular, spatiotemporally located individuals, the StyleGAN images count as misrepresentations. However, if you can reliably identify which images come from the StyleGAN, these misrepresentations need not stand in the way of the acquisition of new knowledge. The synthetic images still provide accurate information about the statistical properties of cat-like appearances, even though they fail as photographic representations of particular animals.

The crucial point is that an output can be a misrepresentation according to the standard interpretive scheme for its domain, and nevertheless remain epistemically benign as long as users can detect it and adjust their interpretive stance accordingly. When mixed with genuinely photographic images, the StyleGAN images are hallucinations, but, even if they differ systematically from genuine photos in ways that are hard to see, they need not cause any false beliefs about cats.[4] In scientific contexts, similarly, an output might misrepresent the target phenomenon according to standard domain conventions, and yet remain benign or even useful if researchers can identify it and interpret it appropriately.

This puts us in position to distinguish hallucinations from another familiar category of scientific misrepresentation: idealizations. Scientific models include systematic distortions—treating gases as point particles or assuming frictionless planes—to render phenomena tractable [Weisberg, 2007; Strevens, 2016]. These are *strategic* distortions, deliberately introduced to facilitate inference. Idealizations involve misrepresentation, but they operate within a well-understood interpretive framework where the distortions are controlled and their effects on downstream inference are anticipated. Hallucinations, by contrast, are *non-strategic* misrepresentations—unintentional artifacts of the generative process that must be either filtered out or managed through interpretive reframing.

With these distinctions in place, we can now define hallucinations for scientific applications of generative AI. A hallucination is a generative AI model output that satisfies three conditions:

1. It counts as a misrepresentation of the target system according to the canonical interpretive scheme for outputs of that kind.

2. It is non-strategic: an unintended artifact of the generative process rather than a deliberate idealization.

3. Its misrepresentational status was produced by the model's generative activity, rather than having been inherited from the training data.

[4] This is why we can coherently talk about detecting, flagging, and filtering hallucinations without contradiction. Hallucination is defined by misrepresentation according to canonical interpretive schemes, not by capacity to produce false belief. Successful detection simply prevents epistemically harmful misrepresentations from propagating through the workflow.

[5]

This definition is intended to pick out the class of hallucinations more accurately than existing alternatives, but it is also structured to help analyze how scientists address the diagnostic problem introduced earlier. The three-part definition directs our attention to the right questions: (1) What interpretive scheme governs scientific practice in this domain? (2) How do workflows distinguish strategic from non-strategic misrepresentations? (3) What mechanisms prevent model-generated errors from propagating? Our case studies show that when these questions are answered carefully, the diagnostic problem becomes tractable. Recall that the problem arises from the combination of opacity and model-generated error: once we discover a hallucination downstream, how can we use that knowledge to improve reliability? The traditional approach—localizing the guilty parameter and adjusting it—is unavailable. But as we will see, scientists work around this limitation not by making models transparent, but by embedding them in workflows that manage error at the level of outputs rather than parameters.

4 AlphaFold and the neutralization of hallucination

AlphaFold, DeepMind's protein structure prediction system, represents one of the most significant recent achievements in scientific AI. The second model in the AlphaFold series, AlphaFold 2, solved the long-standing protein folding problem and led to the 2024 Nobel Prize in Chemistry, awarded to Demis Hassabis, John Jumper, and David Baker. The latest iteration, AlphaFold 3, builds on this foundation but significantly expands the model's capabilities. It goes beyond folding to predict interactions between proteins and small molecules, including ions, nucleotides, and drug-like compounds. This expansion is enabled by a core architectural shift: AlphaFold 3 incorporates a diffusion module to generate plausible molecular structures across a broader range of biological targets. But that flexibility also increases the risk of hallucination. This risk is explicitly acknowledged in the paper that introduces the model:

> The use of a generative diffusion approach comes with some technical challenges that we needed to address. The biggest issue is that generative models are prone to hallucination, whereby the model may invent plausible-looking structure even in unstructured regions [Abramson et al., 2024, p. 496].

[5] Because this definition is restricted to generative AI models, it implicitly incorporates the high-dimensional, structured nature of their outputs. As discussed in Section 2.1, this complexity is central to understanding both why hallucinations are inevitable and why they pose distinctive epistemic challenges in scientific contexts.

This admission makes clear that hallucination is not a marginal failure mode, but a central epistemic challenge for scientific AI. So how does the AlphaFold 3 team mitigate hallucination? The answer lies in two main strategies: (i) the use of theoretical knowledge to guide training, and (ii) the use of confidence-based error screening to guide the interpretation of model output.

4.1 Theory-guided training

Unlike large language models,[6] which learn statistical structure from vast, heterogeneous, and poorly organized datasets, AlphaFold 3 is trained on the Protein Data Bank (PDB), a highly curated repository of experimentally validated molecular structures. Moreover, the training procedures encode well-established physical and biochemical constraints through carefully designed *violation loss functions*. Candidate outputs are penalized if they exhibit steric clashes, implausible bond lengths, or physically unrealistic torsional angles—the rotational angles around chemical bonds that determine backbone geometry.

The necessity of these constraints is demonstrated empirically. The AlphaFold team reports that without violation loss terms, “the network is observed to frequently violate the chain constraint during the application of the structure module” [Jumper et al., 2021]. That is, the model produces structures with impossible bond geometries—steric clashes where atoms are represented closer together than the van der Waals radius permits, peptide bonds at wrong angles, and so on. The violation loss actively suppresses such physically incoherent predictions by shaping the diffusion model’s learned score function to reward outputs that adhere to molecular physics. Even with these penalties in place, AlphaFold’s raw outputs still require a final refinement step using molecular dynamics simulations (Amber force field) to perfectly enforce physical constraints. This post-processing step indicates just how difficult it is to satisfy these constraints through neural network training alone.

The epistemic force of AlphaFold’s constraint-based design stems not only from the content of the physical laws it encodes, but from the fact that the evidential basis for those laws is largely independent of the training distribution. Confirmation is strengthened when distinct bodies of empirical knowledge, each grounded in a different measurement techniques, converge on a common target [Sober, 1989; Schupbach, 2018]. The structural regularities distilled into

[6] Some generative protein folding models, such as Meta’s ESM protein model, are described by their authors as “language models”, despite being trained on biochemical data rather than natural language. When I use the term “language model”, I am referring to models trained on natural language.

the PDB and the theoretical constraints operationalized in the loss function arise from separate physical processes and measurement paradigms. Their convergence in AlphaFold's architecture transforms an inductive generalization into a theoretically disciplined scientific inference.

Another training-phase technique is *cross-distillation*, in which AlphaFold is retrained using the outputs of other models with simpler and more interpretable error profiles. Comparing these models is another way that systematic bias can be exposed. For instance, recent work [Brotzakis et al., 2025] retrained AlphaFold on coarse-grained structural approximations, increasing the model's caution in structurally ambiguous regions where hallucinations tend to arise.

This analysis also speaks to concerns about whether deep learning models can acquire causal knowledge. Some critics worry that DNNs merely find statistical patterns without learning the causal structure that underlies them [Pearl, 2018; Marcus and Davis, 2019]. AlphaFold's design suggests a more nuanced picture. While the model may not represent causal mechanisms in a form that supports arbitrary counterfactual reasoning, its training is nevertheless disciplined by causal knowledge. Theoretical constraints grounded in physics and chemistry actively shape the optimization process. Moreover, the Protein Data Bank is not a random sample of molecular configurations but a theoretically curated archive of structures inferred through techniques like X-ray crystallography, cryo-EM, and NMR spectroscopy. These techniques themselves depend on causal models of how electromagnetic radiation interacts with molecular structure. So although AlphaFold may not contain an explicit, manipulable causal model à la Pearl, its learned representations implicitly encode causal constraints from molecular physics.

4.2 Confidence-based error screening

Hallucinations that cannot be eliminated may still be rendered epistemically harmless, as long as we have a method for singling them out. That is the role of *confidence-based error screening*. In AlphaFold, this is achieved (in part) by means of residue-level reliability scores that help scientists distinguish between outputs that support inference and those that warrant caution.[7]

The central tool here is the Predicted Local Distance Difference Test (pLDDT). Rather than measuring proximity to training examples, pLDDT estimates the local reliability of a predicted structure based on internal consistency cues.

[7] This functionality is sometimes grouped under the heading of "uncertainty quantification," but that term often refers to formal confidence intervals in the context of statistical testing. In contrast, AlphaFold's scores are learned by the model and serve a primarily to enable scientists to screen for unreliable outputs. "Confidence-based error screening" is my own term, which I think more accurately reflects the epistemic function of the relevant techniques.

Specifically, AlphaFold generates multiple structure predictions through a stochastic sampling procedure, and pLDDT scores reflect the degree of local agreement among these samples. Where predictions converge tightly, the model assigns high confidence; where they diverge—often due to physical indeterminacy or lack of constraint—it flags the output as unreliable. The underlying idea is simple but powerful: hallucinations are not uniform across stochastic samples. By generating multiple outputs with different random seeds, AlphaFold can identify regions of disagreement and treat them as signals of uncertainty. Idiosyncratic errors tend to cancel out in the aggregate, allowing the model to screen for instability without requiring access to ground truth.[8]

This mechanism is particularly effective in identifying intrinsically disordered regions (IDRs), whose structures are environmentally contingent and cannot be predicted with high fidelity. Rather than hallucinating a confident structure, AlphaFold returns low-confidence, flexible representations, rendered in a distinctive "noodle-like" visual form that contrast sharply with the well-folded, compact forms nearby. This visual convention reinforces the model's confidence scores and functions as a cue to practicing scientists that the structure is not to be over-interpreted. Brotzakis et al. [2025] show that AlphaFold 3 outperforms specialized tools in detecting such regions, despite not being explicitly trained for this purpose.

One might object that if a model's representations are inaccurate, one shouldn't put much stock in its internal confidence scores either. This worry is not misplaced: confidence-based screening is not epistemically infallible. But its value does not depend on access to ground truth at inference time. What matters is that these scores correlate robustly with empirical reliability across a wide range of cases. In AlphaFold's case, high pLDDT values have been shown to track subsequent experimental validation with remarkable consistency. The metric does not guarantee correctness, but it provides a calibrated signal of when the model's outputs can be used for inference, and when they should not be. This is enough both to shift hallucinations from epistemic threats to manageable uncertainties, and to give scientists license to treat the high-confidence outputs as serious candidates for belief.

Crucially, AlphaFold achieves this level of reliability despite the opacity of its internal representations. Its trustworthiness does not depend on understanding what individual parameters represent, but rather on how the model is embedded in a carefully designed workflow. Theory-guided training shapes the model's outputs through physical

[8] pLDDT scores are produced by a head in AlphaFold's architecture that is trained to predict the expected deviation between predicted and true interatomic distances for each residue. During training, this head is supervised using experimentally validated structures, allowing the model to calibrate its internal confidence estimates. At inference time, however, the score is computed purely from the model's own internal representations; no comparison to ground truth is made.

constraints, while confidence-based screening provides systematic signals about when those outputs can support reliable inference. This is the core insight of *computational reliabilism* [Durán and Formanek, 2018; Duran, 2023]: models can support reliable inference even when they are not transparent, so long as they are embedded in well-designed error-screening workflows.

It is worth contrasting this approach with filtering-based methods like those proposed by Aithal et al., which define hallucinations as outputs that deviate from the training distribution. As argued in Section 2, such deviations are inevitable in high-dimensional generative models. But AlphaFold's confidence-based error screening does not treat deviation from training data as a defect per se. Unlike distributional filters that discard statistically anomalous samples, pLDDT permits substantial departures from the training distribution as long as they are robust across the model's internal ensemble. This allows AlphaFold to support meaningful extrapolation, while still flagging outputs that are likely to be unreliable.

4.3 Scope and Limitations

The strategies demonstrated here succeed because structural biology is a theoretically mature domain. The Protein Data Bank encodes decades of experimental work, yielding a training corpus of over 200,000 structures. The violation loss functions operationalize physical laws that have been validated across countless experiments: van der Waals radii, bond angle constraints, torsional potentials. Confidence scores can be calibrated against ground truth because experimental validation techniques like X-ray crystallography, cryo-EM, and NMR spectroscopy are well-established and widely available.

In domains where such resources are absent, comparable reliability may be unattainable. Theory-guided training requires robust theory to guide it. Confidence-based screening requires extensive validation data to calibrate against. AlphaFold's success demonstrates what is possible when generative AI is embedded in a mature scientific framework, but it does not show that similar workflows can be constructed anywhere.

5 From molecules to meteorology

If AlphaFold demonstrates how generative AI can support inference in molecular biology, GenCast shows how the same epistemic principles extend to large-scale dynamical systems. Unlike protein folding, which targets a stable conformational structure, meteorological forecasting concerns a chaotic, evolving system in which small errors in initial conditions can rapidly amplify [Lorenz, 1995]. It also lacks a clearly defined end state. Traditional numerical weather prediction (NWP) systems, such as those developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), generate forecasts by numerically solving fluid dynamics equations. These methods are physically grounded and benefit from interpretable parameters, but they are computationally intensive. The challenge is especially acute for low-frequency, high-impact events—such as floods or wildfires—which lie in the tails of the distribution. Capturing them reliably requires extremely large ensembles, and computational costs rise steeply with event rarity.

GenCast offers a relatively computationally efficient alternative. It is a new, diffusion-based generative model trained on ERA5, a reanalysis dataset[9] produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 combines physics-based simulations with extensive observational data, including satellite, radar, and ground measurements, and uses advanced data assimilation techniques. The result is not a record of direct observations. Rather, it is a model-informed reconstruction that aims to balance empirical accuracy with physical coherence.

Like AlphaFold, GenCast learns the statistical structure of valid trajectories and generates plausible forecasts via a generative process. Though it lacks an explicit representation of fluid dynamics, it matches or exceeds the performance of traditional simulation-based systems on several standard forecasting metrics [Price et al., 2024]. Earlier models such as Pangu Weather [Bi et al., 2023] achieved comparable accuracy, but GenCast was the first to combine this with a systematic probabilistic evaluation framework. This probabilistic framework enables systematic assessment of when and where the model's predictions are reliable.

As with AlphaFold, GenCast's reliability stems from its integration into a theory-informed workflow. Both systems are trained not on raw, unstructured data but on theoretically curated datasets: AlphaFold on the Protein Data Bank, which encodes experimentally validated molecular structures, and GenCast on ERA5, which integrates physical models with observational data to reconstruct coherent atmospheric states. In each case, domain knowledge shapes the training corpus itself, ensuring that the model learns from inputs already disciplined by physical theory. Moreover, like AlphaFold, GenCast incorporates physically grounded loss functions that penalize violations of general physical laws such as conservation of mass, momentum, and energy [Kashinath et al., 2021]. Earlier machine learning models could

[9] In meteorology, a *reanalysis* is a dataset created by assimilating diverse historical observations into numerical weather prediction models, producing a spatially and temporally coherent reconstruction of past atmospheric states. Although physically constrained, reanalysis outputs are model-dependent and may reflect biases or limitations of the underlying data assimilation systems [McGovern et al., 2024].

generate forecasts that appeared locally plausible but violated global coherence. For example, they might predict a negative humidity value or a physically unrealistic temperature gradient [Watt-Meyer et al., 2021]. GenCast avoids such failures by incorporating these physical constraints during training and by relying on architectures that tend to preserve them at inference time. These practices reduce the risk that model outputs will be epistemically disruptive, in the sense defined in Section 3.

These theoretical constraints help ensure that GenCast's outputs are physically plausible—but plausibility alone does not guarantee reliability. As with AlphaFold, GenCast also implements a strategy for detecting and managing residual errors. In place of AlphaFold's explicit confidence scores, GenCast addresses hallucination through a form of ensemble-based uncertainty estimation. Rather than assigning confidence values to individual predictions, it introduces stochastic variation at inference time, generating an ensemble of plausible forecasts from different random seeds. Given the chaotic nature of weather systems, each trajectory varies in local details, but the ensemble as a whole preserves coherent global structure—reflecting past variability rather than hand-tuned perturbations [Lessig et al., 2023]. The epistemic value of this approach lies in how this variability reveals where inference is likely to be unreliable. Because hallucinations differ across stochastic runs, their dispersion serves as a signal of epistemic instability. Unstable predictions appear as outliers, while robust features emerge as recurring patterns. In this way, GenCast provides a confidence signal. Although it is not a separately computed output, as it is in AlphaFold, it is a reliable statistical pattern that expert scientists can leverage.

One way to assess how well GenCast's internal uncertainty estimates align with forecasting performance is through the spread–skill ratio, which compares the ensemble's internal variance (the spread) with its actual forecast error (the skill).[10]A spread–skill ratio near 1 indicates that the model's uncertainty estimates are well-matched to its performance—neither overconfident nor needlessly conservative. In the GenCast evaluation, this ratio remained close to 1 across a range of forecast horizons (i.e., the time intervals into the future for which predictions are made), confirming

[10] The spread–skill ratio (SSR) compares an ensemble's internal variance (spread) to its forecast error (root-mean-square error, RMSE). It is given by:

$$\text{SSR} = \frac{\text{Spread}}{\text{RMSE}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - \bar{f})^2}}{\sqrt{\frac{1}{M}\sum_{j=1}^{M}(\bar{f}_j - o_j)^2}}$$

where f_i is the forecast from ensemble member i, $\bar{f}$ is the ensemble mean forecast, o_j is the observation at verification case j, N is the number of ensemble members, and M is the number of forecast–observation pairs. An SSR near 1 indicates well-calibrated uncertainty; values significantly greater or less than 1 suggest overdispersion or underdispersion, respectively [Fortin et al., 2014]. Variations on this definition exist, and there is ongoing debate about which formulation is most appropriate in different settings.

that the ensemble's internal dispersion reliably mirrors the inherent empirical uncertainty in chaotic systems [Price et al., 2024].

5.1 Scope and Limitations

GenCast inherits its reliability from meteorology's theoretical and observational infrastructure. Atmospheric dynamics obey quantitative physical laws (Navier-Stokes equations, thermodynamic principles, conservation laws) that can be encoded in loss functions. The ERA5 reanalysis dataset integrates decades of satellite, radar, and ground observations with physics-based models, providing a training corpus that is both extensive and theoretically constrained. Forecast verification is rapid: predictions can be checked against actual weather within days or weeks, enabling continuous calibration of ensemble-based uncertainty estimates through metrics like the spread-skill ratio. This combination of quantitative theory, dense observational networks, and fast empirical feedback is not universal. In domains lacking these features—where theories are qualitative, data are sparse, or validation timescales are long—ensemble-based uncertainty estimation may be less reliable and physical constraints harder to operationalize. For example, climate modeling on decadal timescales faces validation challenges that weather forecasting does not, since we cannot quickly verify 30-year projections. Similarly, in fields where physical laws are poorly understood or observational coverage is patchy, the confidence signals that GenCast provides may not emerge reliably. GenCast demonstrates how theory-rich workflows can discipline generative inference under favorable conditions, but these conditions cannot be assumed to hold everywhere.

GenCast's success shows how hallucinations can be made epistemically tractable. Rather than causing scientists to adopt false beliefs, errors are converted into expected, bounded deviations that the broader workflow is designed to absorb. This undermines the suspicion that opaque generative models leave us with no alternative but what Duede memorably called *brute inductivism.* On that view, the lack of interpretability precludes the possibility of theory-guided inference, and users are left to trust outputs solely on the basis of observed empirical correlations. But as the GenCast case makes clear, we have more to rely on here than the naked predictive track record. Ensemble-based uncertainty modeling converts errors into calibrated signals about where predictions are reliable. By combining theory-informed training with explicit error detection practices, GenCast enables a mode of inference that is neither brute nor blind. It underscores the same lesson we saw in AlphaFold: generative models become reliable not by virtue of their internal transparency, but by being embedded in theory-rich, uncertainty-aware practices that help scientists anticipate and manage error.

6 Discovery and justification

I have argued that the threat of hallucination does not undermine the reliability of scientific AI because generative models can be embedded in epistemically robust workflows. Yet one might object that my emphasis on strategies for reliability misses what ought to be the centerpiece of any response to concerns about hallucination: the main epistemic safeguard for scientific AI is post-hoc empirical validation. We trust AlphaFold primarily because we can experimentally test its predictions, and we trust GenCast because we can wait two weeks and see whether it rains.

This objection is inspired by Duede's [2023] argument that concerns about AI reliability often reflect a misunderstanding of its scientific role. Duede claims that AI is fundamentally a tool for *discovery*, not for justification. On this view, my concern about hallucination wrongly presupposes that AI is in the business of delivering justification for model outputs. Duede might argue that AI merely offers heuristic guidance: it narrows the space of relevant hypotheses, but those hypotheses only become candidates for belief once they have been subjected to empirical test. From a thoroughly empiricist standpoint, the strategies scientists use to mitigate hallucination appear secondary—or even unnecessary.

This response echoes Karl Popper's [1959] influential distinction between the context of discovery and the context of justification. Popper famously argued that the epistemology of science should concern itself solely with justification through rational reconstruction, since the processes of discovery are guided by intuition, creativity, and other factors beyond rational control.

Yet the historical turn in the philosophy of science has cast doubt on the sharpness of this division. Popper's distinction, while conceptually useful, is ultimately artificial. In practice, discovery and justification are often intertwined. Scientific heuristics are not arbitrary guesses; they are evaluated by their empirical traction and shaped by theoretical expectations. This is especially evident in the development and funding of generative AI systems like AlphaFold. The model was not funded merely for its capacity to generate intriguing hypotheses, but because, prior to large-scale empirical validation efforts, its developers demonstrated that it could reliably predict biologically plausible protein structures. Its ability to infer accurate 3D conformations from amino acid sequences had clear implications for understanding biological function and disease. This predictive success led to rapid adoption across the life sciences, where researchers now use its outputs to guide experimentation and hypothesis formation. AlphaFold is not treated as a tool for blind exploration, but as a theory-informed model capable of supporting novel inferences. When scientists take one of its outputs to be approximately true, their belief enjoys a measure of epistemic justification.

The same reasoning applies to weather models such as GenCast. Its probabilistic forecasts are not treated as mere exploratory hypotheses awaiting eventual testing; rather, its reliability is continuously assessed through rigorous calibration against both theory and observational data. GenCast is already being used operationally, for example in forecasting applications like *OpenSnow*, which helps backcountry skiers assess conditions and make daily decisions.[11] Such practical use makes sense only if GenCast provides a justified basis for action in advance, rather than merely suggesting hypotheses for empirical investigation.

More generally, this suggests a modest lesson: generative models are most effective where background knowledge is sufficiently extensive to constrain their outputs and structure their use. In domains like protein folding or meteorology, theory provides a framework that helps identify and account for error. Where such knowledge is lacking, errors are harder to interpret and more likely to mislead. Generative models can accelerate discovery, but they do so most reliably where prior understanding already runs deep.

Duede rightly challenges overly skeptical views that demand too much of generative AI. But by relegating AI entirely to the context of discovery, he leans too heavily on a distinction that, in practice, is difficult to sustain. Discovery and justification are deeply intertwined in scientific inquiry. Principled strategies for managing hallucination are not epistemically superfluous; they are essential to the responsible integration of AI into scientific practice.

Another closely related objection is worth addressing. One might worry that the strategies I have described all have a negative cast: they are concerned primarily with screening for error and adjusting our inferences accordingly. This process of error detection and adaptation, one might argue, is categorically distinct from the accumulation of positive evidence for the truth of a model's output. But from a reliabilist perspective, that distinction breaks down. The reliability of an inferential process increases whenever potential errors are filtered out or otherwise managed. And since reliability is, for the reliabilist, the key property that transforms true belief into knowledge, the task of identifying and managing error is directly relevant to the epistemic status of model-supported beliefs.

[11] GenCast is sufficiently new that real-world applications are only now emerging. It is likely that higher-stakes applications will soon follow.

7 Error mitigation and the construction of reliable science

Generative AI is increasingly central to scientific inquiry, yet the specter of hallucination has raised legitimate doubts about its reliability. I have argued that while hallucination presents a novel epistemic threat, it is not a reason to adopt wholesale skepticism about the use of generative AI in science. The challenge can be addressed by shifting from a data-centric to a phenomenon-centric conception of hallucination. Rather than assessing model outputs by their deviation from training data, we must evaluate them by their correspondence with target phenomena. This shift opens space for alternative reliability strategies, such as theory-guided training and confidence-based screening, that establish systematic output-to-target connections despite opacity. As the cases of AlphaFold and GenCast demonstrate, these methods do not eliminate error, but they make errors anticipatable and manageable.

Although these strategies do not rely on aligning individual model parameters with interpretable features of the world, they are nevertheless more sophisticated than the kind of *brute inductivism* that Duede has criticized. The mechanisms by which these systems manage error are not ad hoc. They draw on independently supported theoretical knowledge about target phenomena and integrate that knowledge into scientific workflows during model development and in the interpretation of results.

Earlier I invoked Durán's *computational reliabilism* to explain how AlphaFold achieves reliability through workflow design despite opacity. Durán and Formanek (2018) originally developed this framework for opaque computer simulations, and Durán [2025] has extended it to algorithms more broadly. Computational reliabilism emphasizes that reliability assessment should focus on the *process* rather than the algorithm in isolation, where "process" encompasses the broader socio-techno-scientific context in which algorithms are designed, used, and maintained.

My contribution extends this framework to address hallucinations in generative AI for science. The distinctive challenge is that hallucinations must be assessed by comparing outputs to target phenomena rather than training data. This comparison is operationalizable only where we possess sophisticated theoretical knowledge about the target system. Knowing that proteins fold according to thermodynamic principles, that weather obeys conservation laws, or that molecular bonds have characteristic geometries is what enables the two reliability strategies examined here. Theory-guided training works by embedding constraints derived from our understanding of target phenomena into the model's architecture and loss functions. Confidence-based screening works because we can calibrate uncertainty metrics against theoretical expectations about when predictions should and shouldn't be reliable. Without mature domain theory, we lack the resources to establish systematic connections between model outputs and target phenomena.

This explains why these strategies succeed in structural biology and meteorology but cannot be assumed to work universally. Both domains benefit from decades of theoretical development: protein folding thermodynamics, structural biochemistry, atmospheric physics, fluid dynamics. This theoretical maturity provides the scaffolding needed to discipline generative models. Where such theory is absent or immature, comparable reliability may be unattainable regardless of computational resources or data availability. Understanding reliability requires assessing the workflow as a whole, not just the model in isolation. Consider Isomorphic Labs, one of the most prominent companies building drug discovery pipelines around AlphaFold-style models. In a recent interview, Rebecca Paul, the company's head of medicinal chemistry, explained that AlphaFold predictions with binding probability scores below 0.7 are systematically filtered out before any synthesis occurs [Paul and Jaderberg, 2025]. Predictions above this threshold are then validated experimentally through X-ray crystallography, with results feeding back into model refinement. The model output is just one step in a larger process that includes filtering based on confidence thresholds, experimental validation, and iterative improvement. Reliable inference emerges from this extended workflow, not from the model's computation alone.

This observation connects to longstanding debates in reliabilist epistemology about process individuation [Goldman, 1986]. What counts as "the process" that produces belief? Processes can be individuated more narrowly (the model's computation) or more broadly (the theory-informed workflow connecting outputs to phenomena). In traditional scientific contexts, we assess experimental procedures as wholes, not individual instruments in isolation. The same principle applies here: the appropriate unit of assessment is the theory-informed, uncertainty-aware, iteratively validated workflow—not the model's internal mechanisms. This is why opacity, though epistemically significant, does not preclude reliable inference. The workflow, infused with theoretical knowledge at multiple points, is what must be reliable.

These conclusions about workflows and process individuation apply clearly to the current generation of generative AI systems. But given how rapidly AI architectures are evolving, I want to make two additional claims about the intended scope of the arguments above.

First, about the workflows: The strategies examined here (theory-guided training and confidence-based screening) depend fundamentally on sophisticated theoretical knowledge about target phenomena. Establishing reliable output-to-target connections requires mature domain theory, regardless of the model architecture employed. Theoretical understanding of protein physics, atmospheric dynamics, and molecular chemistry is what enables the embedding of constraints during training, the calibration of uncertainty metrics, and the interpretation of systematic failure patterns. Moreover, this theoretical knowledge enters the reliability-supporting workflow at multiple stages, not only during model training but also in filtering outputs, designing validation experiments, and interpreting results. Where such theory is available, as in structural biology and meteorology, generative AI can support reliable inference. Where it is not, alternative approaches will be needed.

Second, about the models themselves: A closely related question is whether the models must be domain-specific, or whether general-purpose architectures could eventually achieve similar reliability when embedded in appropriate workflows. On this question I remain agnostic. While domain-specific models currently outperform general-purpose models (largely because domain-specific architectures naturally align with domain-specific workflows), there are reasons to take seriously the possibility that general-purpose models might eventually achieve comparable reliability. Multi-modal models such as GPT-4 and Gemini 2.5 exhibit capacities that cannot be replicated by chaining together narrow tools. Their integration of language, image, and video appears to produce synergies within the model's latent space that surpass what modular composition can deliver. Whether chemical or physical representations could be treated as additional modalities in this sense, and whether such integration, when combined with domain-specific theoretical scaffolding at the workflow level, would yield genuinely new epistemic benefits, remains an open question worth exploring.

The central argument of this paper reflects the Deweyan insight quoted in the epigraph: scientific knowledge is not a matter of grasping antecedently given certainties, but of developing ways to *make sure*—methods for identifying error and managing uncertainty. Generative AI calls for new methods of doing that work, and simultaneously makes that work more difficult.[12] Nevertheless, as AlphaFold and GenCast demonstrate, there is reason to hope that we will be equal to the task.

[12] This dependency on foundational science suggests a cautionary note about research funding. While enthusiasm for AI-driven science grows, there is a risk of reallocating funding away from traditional theory-building and experimentation. Yet generative AI models achieve reliability precisely because they are embedded in theoretical frameworks developed through decades of foundational research. The epistemic scaffolding that AI-based inference depends on requires continued investment. Neglecting these foundations would be self-defeating.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pages 1–3, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w.

Sumukh K. Aithal, Pratyush Maini, Zachary Lipton, and J. Zico Kolter. Understanding Hallucinations in Diffusion Models through Mode Interpolation. *Advances in Neural Information Processing Systems*, 37:134614–134644, January 2025.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *ArXiv*, January 2017.

Julian Arnold, Frank Schäfer, Alan Edelman, and Christoph Bruder. Mapping Out Phase Diagrams with Generative Classifiers. *Physical Review Letters*, 132(20):207301, May 2024. doi: 10.1103/PhysRevLett.132.207301.

S. Banerjee and A. M. Jacob. LLMs will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*, 2024.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3.

James Bogen and James Woodward. Saving the Phenomena. *Philosophical Review*, 97(3):303–352, 1988. doi: 10.2307/2185445.

Z. Faidon Brotzakis, Shengyu Zhang, Mhd Hussein Murtada, and Michele Vendruscolo. AlphaFold prediction of structural ensembles of disordered proteins. *Nature Communications*, 16(1):1632, February 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-56572-9.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, April 2023.

Cameron J. Buckner. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press, Oxford, New York, February 2024. ISBN 978-0-19-765330-2.

Rudolf Carnap. *Der logische Aufbau der Welt*. Felix Meiner Verlag, Hamburg, 3 edition, 1928.

Kathleen A. Creel. Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4):568–589, 2020. doi: 10.1086/709729.

John Dewey. *Experience and Nature*. Dover Publications, New York, NY, USA, 1958.

David T.F Dryden, Andrew R Thomson, and John H White. How much of protein sequence space has been explored by life on Earth? *Journal of The Royal Society Interface*, 5(25):953–956, April 2008. doi: 10.1098/rsif.2008.0085.

Eamon Duede. Deep Learning Opacity in Scientific Discovery. *Philosophy of Science*, 90(5):1089–1099, December 2023. ISSN 0031-8248, 1539-767X. doi: 10.1017/psa.2023.8.

Juan M. Durán. In Defense of Reliabilist Epistemology of Algorithms. *European Journal for Philosophy of Science*, 15 (37):1–20, 2025. doi: 10.1007/s13194-025-00664-2.

Juan M. Durán and Nico Formanek. Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, 28(4):645–666, December 2018. ISSN 1572-8641. doi: 10.1007/s11023-018-9481-6.

Juan Manuel Duran. Machine learning, justification, and computational reliabilism. https://philsci-archive.pitt.edu/22726/, 2023.

V. Fortin, M. Abaza, F. Anctil, and R. Turcotte. Why Should Ensemble Spread Match the RMSE of the Ensemble

Mean? *Journal of Hydrometeorology*, 15(4):1708–1713, August 2014. ISSN 1525-7541, 1525-755X. doi: 10.1175/JHM-D-14-0008.1.

Alvin I. Goldman. What is Justified Belief? In George Pappas, editor, *Justification and Knowledge: New Studies in Epistemology*, pages 1–25. D. Reidel, 1979.

Alvin I. Goldman. *Epistemology and Cognition*. Harvard University Press, Cambridge, 1986.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

G. R. Grimmet and D. R. Sterzaker. Probability and Random Processes. Oxford: Oxford Sc. Publ. 1992.

Thomas Grote, Konstantin Genin, and Emily Sullivan. Reliability in Machine Learning. *Philosophy Compass*, 19(5): e12974, 2024. ISSN 1747-9991. doi: 10.1111/phc3.12974.

Paul Humphreys. The Philosophical Novelty of Computer Simulation Methods. *Synthese*, 169(3):615–626, 2009. doi: 10.1007/s11229-008-9435-2.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, March 2023. ISSN 0360-0300. doi: 10.1145/3571730.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

Christopher Kadow, David Matthew Hall, and Uwe Ulbrich. Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, 13(6):408–413, June 2020. ISSN 1752-0908. doi: 10.1038/s41561-020-0582-5.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated Language Models Must Hallucinate, March 2024.

K. Kashinath, M. Mustafa, A. Albert, J-L. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, A. Manepalli, D. Chirila, R. Yu, R. Walters, B. White, H. Xiao, H. A. Tchelepi, P. Marcus, A. Anandkumar, P. Hassanzadeh, and null Prabhat. Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200093, February 2021. doi: 10.1098/rsta.2020.0093.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 2013.

Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadtler, and Martin Schultz. AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning, September 2023.

Edward Lorenz. *The Essence of Chaos*. The University of Washington Press, 1995.

Calvin Luo. Understanding Diffusion Models: A Unified Perspective, August 2022.

Jack C. Lyons. Algorithm and Parameters: Solving the Generality Problem for Reliabilism. *Philosophical Review*, 128 (4):463–509, 2019. doi: 10.1215/00318108-7697876.

Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon, 2019.

Amy McGovern, Ann Bostrom, Marie McGraw, Randy J. Chase, David John Gagne, Imme Ebert-Uphoff, Kate D. Musgrave, and Andrea Schumacher. Identifying and Categorizing Bias in AI/ML for Earth Sciences. *Bulletin of the American Meteorological Society*, 105(3):E567–E583, March 2024. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-23-0196.1.

Rebecca Paul and Max Jaderberg. A quest for a cure: AI drug design | Isomorphic Labs, June 2025.

Judea Pearl. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018.

Karl Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1959.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, pages 1–7, December 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08252-9.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286. PMLR, June 2014.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 5234–5243, Red Hook, NY, USA, December 2018. Curran Associates Inc.

Jonah N. Schupbach. Robustness Analysis as Explanatory Reasoning. *The British Journal for the Philosophy of Science*, 69(1):275–300, March 2018. ISSN 0007-0882. doi: 10.1093/bjps/axw008.

YiRang Shin, Jaemoon Yang, and Young Han Lee. Deep Generative Adversarial Networks: Applications in Musculoskeletal Imaging. *Radiology: Artificial Intelligence*, 3(3):e200157, March 2021. ISSN 2638-6100. doi: 10.1148/ryai.2021200157.

John-William Sidhom, Giacomo Oliveira, Petra Ross-MacDonald, Megan Wind-Rotolo, Catherine J. Wu, Drew M. Pardoll, and Alexander S. Baras. Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. *Science Advances*, 8(37):eabq5089, September 2022. doi: 10.1126/sciadv.abq5089.

Ritwik Sinha, Zhao Song, and Tianyi Zhou. A Mathematical Abstraction for Balancing the Trade-off Between Creativity and Reality in Large Language Models, June 2023.

Elliott Sober. Independent Evidence about a Common Cause. *Philosophy of Science*, 56(2):275–287, 1989. ISSN 0031-8248.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Michael Strevens. How Idealizations Provide Understanding. In Stephen Robert Grimm, Christoph Baumberger, and Sabine Ammon, editors, *Explaining Understanding: New Perspectives From Epistemology and Philosophy of Science*. Routledge, 2016.

Yujie Sun, Dongfang Sheng, Zihan Zhou, and Yifei Wu. AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, 11(1):1–14, September 2024. ISSN 2662-9992. doi: 10.1057/s41599-024-03811-x.

Yuxuan Wang, Tianwei Cao, Kongming Liang, Zhongjiang He, Hao Sun, Yongxiang Li, and Zhanyu Ma. Mixture-of-hand-experts: Repainting the deformed hand images generated by diffusion models. In *Pattern Recognition and Computer Vision: 7th Chinese Conference, PRCV 2024, Urumqi, China, October 18–20, 2024, Proceedings, Part V*, pages 143–157, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-981-97-8619-0. doi: 10.1007/978-981-97-8620-6_10.

Oliver Watt-Meyer, Noah D. Brenowitz, Spencer K. Clark, Brian Henn, Anna Kwa, Jeremy McGibbon, W. Andre Perkins, and Christopher S. Bretherton. Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations. *Geophysical Research Letters*, 48(15):e2021GL092555, 2021. ISSN 1944-8007. doi: 10.1029/2021GL092555.

Michael Weisberg. Three Kinds of Idealization. *Journal of Philosophy*, 104(12):639–659, 2007. doi: 10.5840/jphil20071041240.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is Inevitable: An Innate Limitation of Large Language Models, January 2024.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is Inevitable: An Innate Limitation of Large Language Models, February 2025.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.*, 56(4):105:1–105:39, November 2023. ISSN 0360-0300. doi: 10.1145/3626235.