Author version— forthcoming in *Synthese*

# Strong Novelty Regained: High-Impact Outcomes of Machine Learning for Science

Heather Champion (Western University Canada)

hchampi2@uwo.ca

**Abstract**

A general class of presupposition arguments holds that the background knowledge and theory required to design, develop, and interpret a machine learning (ML) system imply a strong upper limit to ML's impact on science. I consider two proposals for how to assess the scientific impact of ML predictions, and I argue that while these accounts prioritize conceptual change, the presuppositions they take to be disqualifying for strong novelty are too restrictive. I characterize a general form of their arguments I call the Concept-free Design Argument: that strong novelty is curtailed by utilizing prior conceptualizations of target phenomena in model design. However, I argue that if ML design choices (such as ground-truth labels for supervised ML and inductive biases) are based on prior conceptualizations of phenomena, it need not impede conceptual change. Furthermore, while their accounts focus narrowly on conceptual change, a variety of learning outcomes also contribute to strong scientific change. Thus, I present a variety of types of strong novelty from philosophy of creativity, epistemology, and philosophy of science that paint a more varied picture of how ML advances science. One of these is a form of *local* theory-independent learning from data that signals an aim to substantially revise existing theory, but it is not easily undermined by prior assumptions about target phenomena. Furthermore, generating surprise, reducing utility blindness, and eliminating deep ignorance also indicate high impact to scientific knowledge or research direction. I illustrate these types of strong novelty with several cases of scientific discovery with algorithms. My taxonomy clarifies several desiderata for machine-based exploration and should inform choices in designing for scientific change.

**Keywords**: Novelty; Creativity; Scientific Discovery; Machine Learning.

## 1. Introduction

A general class of presupposition arguments holds that the background knowledge and theory required to design, develop, and interpret a machine learning (ML) system imply a strong upper limit to ML's impact on science. For instance, some philosophers argue that existing theoretical frameworks impose hard constraints on the possibility of ML generating strong change to biological knowledge (Ratti 2020) or that prior paradigm commitments imply humans still play

the "leading role" in guiding conceptual revision in psychiatry (Genin et al. 2024). Others worry that a "theory-free" ideal perpetuates a false conception of scientific objectivity and the view that ML will profoundly change scientific knowledge by minimizing or eliminating theory (Andrews 2024). Furthermore, some suggest ML's potential to enable scientific progress might be analyzed by the idea of "use novelty," which is highly sensitive to what prior theory was used for model definition and training (Boge 2022). Yet, as these authors demonstrate, a plethora of prior assumptions regarding phenomena inform the design of an ML system—and so the prospects of ML generating *strong* novelty seem to have been lost.

However, as I will argue, these accounts tend to overstate the presuppositions that are disqualifying for strong novelty. Also, they each focus narrowly on one or two types of high impact to scientific knowledge (e.g. conceptual change), neglecting the variety of outcomes that are significant for scientific advancement. Often, they also conflate the goals of automating science with advancing science: they focus primarily on the novelty generated by an ML system rather than by a human-machine collective.[1] Finally, they tend to fault *presuppositions* for a gap between some preferred concept of strong novelty and the limitations of ML. But this neglects other contributing factors such as theory's openness to revision or if the concept of novelty matches the level of generality of an ML problem (e.g., does ML function as an appraiser when strong novelty means deciding between research paradigms?).

This paper will engage with several of the shortcomings of presupposition-based accounts, aiming to reorient the discourse regarding novelty and ML towards a more pluralistic,

[1] Clark and Khosrowi (2022) argue convincingly that AI entities deserve consideration as members of a discovering collective and suggest that some contextualized form of novelty or originality might be used to analyze their contribution.

prudently permissive perspective. I present a variety of types of strong novelty that paint a more varied picture of how ML advances science. One of these is a form of *local* theory-independent learning from data that signals an aim to substantially revise existing theory but is not easily undermined by prior assumptions about target phenomena (Section 3.5). Like presupposition-based accounts, I focus on the novelty of ML outcomes rather than the novelty of ML processes (such as creative processes; Halina 2021). Unlike existing accounts, I identify types of strong epistemic impact that might be applied to any learning system, whether human, machine, or human-machine (revealing co-learning outcomes). This approach is appropriate for designing new research projects with ML: in learning scenarios, some form of local epistemic change is necessary for wider historical impact. Indeed, an outcome that is "psychologically" novel with respect to some learning system is necessary to generate a change that has never occurred in history before ("historical novelty").[2] Therefore, system learning outcomes suggest relevant desiderata when designing for scientific change. Moreover, I present cases that show how these outcomes culminate in historical novelty for science.

The paper will proceed as follows. I first consider two existing proposals for how to assess the scientific impact of ML predictions, and I argue that while these accounts prioritize conceptual change, the presuppositions they take to be disqualifying for strong novelty are too restrictive. Specifically, I argue that a general form of their arguments I call the Concept-free Design Argument relies on too narrow a concept of strong novelty and too broad an account of the presuppositions that limit it (i.e. provisional assumptions and theoretically non-salient aspects, Section 2). I then introduce my taxonomy, drawing on literature from the philosophy of

[2] Boden (2004, 2009) makes this point for creative outcomes: historically creative ideas (that have never occurred in history before) are special cases of psychologically creative ideas (that are new to the person who generates them).

creativity, epistemology, and the philosophy of science (Section 3). I begin with types of unexpectedness that signal deep, local scientific impact (surprise, blindness reduction, deep ignorance elimination). I then revisit the significance of conceptual change—a wide form of impact—and I argue that a variety of ML outputs might stimulate conceptual change, while deep learning (DL) might directly generate novel conceptualizations. Finally, I argue that the degree of independence of *local* theory that demarks or explains phenomena helpfully discriminates strong from weak learning outcomes. I end by considering the significance of my taxonomy in helping to design new research projects and agendas, as well as the role it might play in dialectic with process-centered accounts (Section 4).

## 2. Existing philosophical accounts of strongly novel ML outcomes

Philosophers often conduct case studies that focus on the role of theory in science; thus, their accounts of ML's impact tend to be narrow and limited to a particular domain. Nonetheless, here I consider general features of two accounts that distinguish strong and weak types of novelty according to certain types of presuppositions. I first introduce their projects along with the view that strong novelty is conceptual change (Section 2.1), and then I identify and challenge a general form of their arguments I call the Concept-free Design Argument: that strong novelty is curtailed by utilizing prior conceptualizations of target phenomena in model design (2.2–2.4). Finally, I consider what other intuitions these accounts capture that should be retained (2.5).

### 2.1. Conceptual Change

Ratti (2020) considers how ML models might impact theory in the domains of molecular biology and genomics. He identifies three ordinal kinds of novelty: the weakest kind occurs if an ML model indicates that a new biological model might be added to a family of models, allowing

existing theory to be explored in a new direction ("N1," pp. 88–89). A stronger, revisionary, kind of novelty would occur if an ML model represented patterns in data that suggested a modification were needed to a family of biological models, such as a mechanism ("N2"). He characterizes the strongest kind of novelty as a change to the background theory that informs an investigation, which he understands as conceptualizations of phenomena ("N3"). He concludes that it is not possible for ML to generate N3 novelty: existing biological concepts must be used to label examples for training the model (supervised ML) or to interpret novel classes (unsupervised ML). Meanwhile, N2 is only possible in principle but not in practice since researchers rely on existing biological theory to evaluate the reliability of any unexpected algorithmic outputs. Thus, he argues ML only generates N1 novelty in biology (pp. 91–92).[3]

Boge (2022) focuses on the relationship between discoveries made with DL algorithms and scientific understanding. He argues that since DL models are content instrumental (their formal elements need not be assigned any meaning for them to have predictive power) and opaque with respect to what complex, abstract features of the data they utilize, understanding mechanisms that govern a target phenomenon is possible only under certain conditions. Particularly, it requires a prior conceptualization of input data and output predictions, as well as

[3] Ratti (2020) argues that since the structure of theory varies in each scientific discipline, articulating types of novelty must be done on a domain-specific basis (2020, p. 86). Thus, his conclusion applies to molecular biology and genomics. Nonetheless, he expects that it might extend to other disciplines that involve a mechanistic background interpreted in a qualitative way (p. 95). However, it is unclear why qualitative theoretical structure might curtail strong novelty, except perhaps as fully automated conceptual change. Moreover, I challenge the programmatic nature of his claim—the idea that domain-specific theory should act as a normative constraint for investigating novelty and ML (p. 95). Indeed, philosophers have offered domain-general accounts of exploratory activities (e.g. Elliott 2007), and most ML algorithms and explainable AI techniques are amenable to this level of analysis as they are designed to be domain-general (e.g. Zednik & Boelsen, 2022).

background knowledge for connecting an interpretation of what the algorithm has learned to the target phenomenon (p. 55). He thus concludes that DL models might play a role in stimulating radical conceptual shifts, but in the most exploratory contexts where there is little background theory to guide model interpretation, DL is unlikely to generate the scientific understanding needed for conceptual change unless the "right set of geniuses, with the necessary 'exotic' ideas, are around" to make the connection (p. 72).

Both Ratti and Boge suggest that an outcome that induces the need to change the background theory used to guide an exploratory activity is a very strong kind of scientific impact. For Boge, background theory refers to the general structure of mechanisms or state spaces being explored in a concrete setting; he follows Franklin's (2005) distinction between background theory and local theory. Ratti defines "N3" novelty in terms of change to the theoretical "store," meaning "theory when it is used to conceptualize phenomena" (p. 89). Although these are slightly different concepts,[4] in this section, I will focus on conceptual change, which aligns with both of their accounts and captures the idea that background theory provides the structure, rather than the particular "ingredients" for a mechanism or state space (Boge, p. 69). I agree that conceptual change regarding phenomena is an important dimension of strong novelty. However, I disagree that utilizing existing concepts in model design generally hinders it.

[4] Ratti's (2020) characterization of the biological "store" stems from Douglas and Magnus' (2013) concept of theoretical framework, but he distinguishes between ways of using a biological theory in contrast to understanding theory and framework as separate conceptual entities (p. 88). Notably, Douglas and Magnus' theoretical framework includes a variety of concepts and commitments at all levels of scientific description, such as auxiliary hypotheses (p. 88). Thus, change to background theory might in general indicate a different idea than change to conceptualization of phenomena. Furthermore, Boge (2022) might mean theory at a lower level of description: he mentions that quantum field theory is the background theory of particle physics (p. 69). Nonetheless, I focus on conceptual change as it is relevant to a general form of their argument that I consider in the remainder of Section 2.

Furthermore, I take a broader view of the significance of conceptual change, and I propose a variety of ways that ML can contribute to it (see Section 3.4).

### 2.2. The Concept-free Design Argument

Not only do Boge and Ratti share a similar definition of strong novelty, but they also take a similar view of what design choices curtail it. Generalizing their accounts, I call the argument that ML does not generate strong novelty if prior conceptualizations of target phenomena are used to design a model the **Concept-free Design Argument**:

> $Pr_1$ Strong novelty is change to conceptualizations of phenomena.
>
> $Pr_2$ If prior conceptualizations of phenomena are used to design a model, the model does not stimulate/generate significant conceptual change.
>
> $\rightarrow$ ML design choices that operationalize prior conceptualizations of phenomena curtail strong novelty.

I have already explicated Boge and Ratti's views of strong novelty ($Pr_1$): Ratti sees strong novelty as change to the theoretical framework used to conceptualize biological phenomena ("N3"), Boge sees strong novelty as change to background theory, which includes conceptual change. Their reasons supporting the second premise ($Pr_2$) differ: Ratti emphasizes the use of ML in biology is part of a theory-informed practice that involves an end-to-end commitment to a certain explanatory model from data acquisition to post-hoc interpretation of model outputs (p. 89). Boge, on the other hand, seems concerned with how much empirical progress might be possible if prior conceptualizations help to generate ML predictions: he suggests a form of "use novelty" might be applied to ML outputs to signal previously unobserved phenomena (p. 47, see Section 2.4 below).

However, I argue that the Concept-free Design Argument faces two major limitations: (1) it overfits to a single type of strong novelty, while exploratory activities generate a variety of strong impacts to scientific knowledge and research direction, and (2) it assumes that prior conceptualizations are theoretically "interesting" in that they suitably demark or explain phenomena.[5] However, I will show next that implementing prior conceptualizations of phenomena via two ML design choices (i.e. ground-truth labels for supervised ML, inductive biases) need not curtail strong novelty, even when understood in the sense of $Pr_1$. The remainder of the paper will address both limitations, serving to reorient the discourse regarding novelty and ML: I introduce a wide variety of concepts of strong novelty (Section 3), I highlight cases and uses of supervised ML that show its high impact on scientific knowledge and research (see Section 3.2–3.4), and I introduce a concept of theory-independence that better nuances how design choices might curtail strong novelty (Section 3.5).

### 2.3. What's in a label?

Both Boge (2022) and Ratti (2020) take the view that supervised ML does not produce strongly novel scientific outcomes since, in the case of classification, training a model involves a ground-truth label that represents a preconception of how to identify a known phenomenon. Ratti emphasizes the use of supervised ML in biology is "pervaded by *non-novelty*"; it simply automates the identification of biological entities that are well characterized and formally defined (p. 90). But Boge also claims that if token predictions merely correspond to "the recognition of the presence of a type of phenomenon of interest" they are generally weaker than predictions of novel types (p. 48). Thus, they claim that *unsupervised* ML affords the possibility of generating

[5] Thanks to an anonymous reviewer for the helpful suggestion to frame my response to Ratti and Boge by identifying a general form of their argument along these lines.

stronger novelty since it groups unlabeled data points into clusters that can be used to define novel types of phenomena.

The idea that utilizing a class identifier undercuts strong novelty risks overly constraining exploratory science. Even when strong novelty is defined as conceptual change (in the sense of $Pr_1$), identifying a phenomenon does not imply it is adequately understood or explained. For example, the theoretically salient aspect of predicting the locations of earthquake aftershocks is how to conceptualize the nature of the relationship between mainshocks and aftershocks, not whether a particular terrestrial location is classified as mainshock, aftershock, or no shock (I also discuss how this case demonstrates blindness reduction, another form of strong novelty, in Section 3.2).

Moreover, identifying some known aspects of a phenomenon does not inhibit significant conceptual change. This assumption relies on too sharp a distinction between identification and refinement of target phenomena, but exploratory work might iterate between these. For example, Yao (2023) argues that some (token) astronomical events act as "Rosetta Stone" clues for unlocking the keys to articulating general astrophysical types (as do "traces" in the historical sciences, p. 1389). Boge does recognize that ML might be supervised with less theory: if tokens are merely labeled as "background" or "anomaly" during training, the latter can be analyzed for traces of new types of particles (p. 67). But the labels used to train a supervised ML model might correspond to varying *degrees* of scientific understanding and play a more or less established role in a given theory.[6]

[6] Granted, if labels are manually annotated by humans, the resources required seem to imply some degree of conceptual stability. Nonetheless, labels are not always generated in this way: they might merely represent a known outcome corresponding to an input sample.

In sum, characterizing predictions as strong or weak according to whether they are of token or type does not generally align with scientific impact, even for novel conceptualizations of phenomena. It is also too coarse for describing the degrees of scientific understanding that might be used in developing an ML model. Implementing prior conceptualizations of phenomena via ground-truth labels need not undermine strong novelty.

### 2.4. The Inductive Bias Worry

I have framed the Concept-free Design Argument in terms of the model design choices that might limit conceptual/theoretical progress, but it is closely related to the Inductive Bias Worry:

> **Inductive Bias Worry**: If prior conceptualizations of target phenomena are used to choose inductive biases for model training, ML predictions might not enable significant empirical progress, and hence little conceptual change.

Boge (2022) argues the use of deep neural network models might produce a large gap between empirical discovery and scientific understanding. But since he also suggests that some strong predictions are "use novel" (Worrall, 1985) relative to model design choices, he seems to worry that choosing inductive biases (introduced below) based on prior conceptualizations of target phenomena might impede empirical scientific progress with ML. Indeed, he suggests that use novelty might be significant after describing two roles for novel predictions: to judge how empirically successful a theory is relative to a competitor (Lakatos, 1970) and to assess the reliability of a discovery method (Maher, 1988).

Philosophers originally proposed the concept of use novelty to capture the idea that some evidence plays a special role in supporting a theory: if a theory entails a certain empirical fact that the theorist did not use to construct the theory, the fact seems to act as a reason for accepting

the theory (see Barnes 2022). Gardner (1982, p. 3) calls this heuristic idea "use novelty," and Worrall (1978, 1985) also develops it. The idea improves upon a purely temporal concept of how novelty contributes to theory confirmation, where only historically novel consequences of a theory are epistemically privileged. The heuristic version captures that if a theory offers new explanations of known facts that it was not specifically designed to accommodate, these *use novel* facts might also offer strong support for the theory. Nonetheless, the heuristic conception still faces the historical difficulty of analyzing what facts a theorist used to construct a theory, which are not usually comprehensively described in the published record. Thus, Gardner (1982) argues for a knowledge-based version that mitigates this difficulty: it only requires analyzing what facts a theorist *knew* when constructing a theory, such that any unknown consequences might contribute to theory acceptance (i.e. if a theorist did not know a fact, presumably, they did not use it for theory construction).

Boge's remarks on DL discovery best align with the knowledge-based concept of use novelty, which is both temporal and heuristic: he argues that a strong prediction is finding a previously unobserved phenomenon and use novelty might signal "that information about that phenomenon was not included in training and model-definition" (p. 47). While it is somewhat difficult to analyze what facts a DL algorithm "used" to generate a prediction, it is somewhat easier to judge what information it "knew"; for instance, inductive biases encode prior information about the preferred types of solutions to direct the learning algorithm. They include design choices such as the initial values for the weights of a neural network or the model architecture that best suits a prediction task. These choices might be based on presuppositions about target system dynamics or on general features of the data.

However, in ML (particularly DL) it is largely an empirical question how inductive biases contribute to successful (i.e. accurate or explainable) predictions. Researchers study the impact of various inductive biases on model generalization, explainability, and problem tractability. Some approaches impose strong, explicit theoretical constraints on the learning algorithm (see Kashinath et al. 2021), while others rely on implicit background knowledge (see Thuemmel et al. 2024). Indeed, it seems to be an emerging scientific question which approach is best for discovery, perhaps with no general answer (Iten et al. 2020, Cranmer et al. 2020, McCabe et al. 2023). Thus, it is far from clear that the Inductive Bias Worry is warranted.

Furthermore, recall that known features of phenomena may not include the theoretically salient aspects (as I argued in Section 2.3 above), and they also do not imply an adequate empirical characterization—utilizing prior concepts does not imply there is little left to learn. It is also somewhat curious that theoretical choices involved in generating data should get a "free ride" in characterizing use novelty: Boge's examples of use novelty include the discovery of novel thermoelectric materials from the texts of prior scientific papers, where data is highly mediated by theory (p. 47). Granted, as he notes, use novelty might also play a role in establishing the reliability of a discovery method (Maher 1988). But assessing reliability in this way does not gauge the scientific impact of a prediction.

Instead, more nuanced distinctions are needed regarding the type of theoretical presuppositions that might impede empirical progress, limiting conceptual change ($Pr_1$). For example, I argue that independence of "local" theory that demarks or explains phenomena affords strong novelty with ML (see Section 3.5).

### 2.5. Beyond conceptual change

In addition to identifying strong novelty as conceptual change, Boge (2022) and Ratti's (2020) accounts capture the more general intuition that *major revisions to theory or existing scientific knowledge* are strongly novel. Also, Boge's distinction between discovery and understanding suggests that a strong outcome of scientific exploration is one that makes *a large impact on the direction of subsequent research*, especially research that targets the understanding that might be lacking. Finally, their accounts suggest that predictions that are *free of certain kinds of theoretical bias* count as novel in a special way. (However, using prior conceptualizations of phenomena to design labels or inductive biases need not curtail strong novelty.) Next, I use these general intuitions to suggest several new dimensions of strong novelty.

## 3. Dimensions of strong novelty for scientific exploration

In this section, I introduce concepts of strong novelty from various philosophical domains that capture the intuitions discussed above that a strongly novel outcome of scientific exploration is one that has great impact on scientific knowledge or research direction and is achieved without a certain kind of theoretical bias. I argue that these concepts make precise what "strong" novelty is for scientific exploration with ML. I do not take any of these individual dimensions to be necessary for strong novelty. Indeed, the first three concepts are forms of "unexpectedness" identified by philosophers of creativity and epistemologists that are largely incompatible, such that if an outcome generates one to a significant degree, it does not generate the others.[7]

[7] I leave open the possibility that these types may occur as a series of outcomes in an exploratory project, and I do not attempt to identify conditions for discretizing outcomes.

Meanwhile, the last two concepts (conceptual change, local theory-independent learning from data) are independent: they can co-occur with the others but are not fully captured by them. My aim is to provide a useful taxonomy for characterizing how ML (and other computational techniques) can contribute to strongly novel scientific outcomes, but I expect and welcome amendments to it based on further study and ML's expanding set of uses in science.

While the philosophy of creativity is relevant in several ways to the analysis of ML-enabled science (creativity can be ascribed to a person, process, or product), I focus on the analysis of outcomes. Nonetheless, creative process and creative outcome are closely linked: most philosophers agree that some sort of process condition is required to identify products that are creative, but they disagree on what it should be. Some argue that a creative outcome requires an agential process (Paul & Stokes 2018). However, in keeping with my focus on strong impact and collective discovery, I will consider accounts that make a nominal commitment to individual process. Particularly, Simonton (2012) argues that Campbell's (1960) theory of creative process explains why creative outcomes differ from obvious solutions: their utility is initially unknown and must be tested, which requires some process of "blind" variation of ideas and selective retention. Notably, this suggests that a wide variety of algorithms might contribute to creative outcomes. Furthermore, this prior blindness suggests the first two ways unexpected outcomes might be generated—by surprise, and by blindness reduction.

### 3.1. Surprise

In the analysis of creative outcome, surprise indicates *a change to the content of belief regarding an idea's utility* (Tsao et al. 2019). For example, surprise occurs if an idea that was first thought to be useless turns out to be useful (Tsao et al. call this "implausible utility"). By Simonton (2012, 2022) and Boden's (2004) criteria, a novel outcome that is useful and surprising

is also *creative*. Alternatively, surprise also occurs when belief that an idea is useful is disconfirmed; for instance, when an anomaly challenges existing theory and prompts a field to enter a highly exploratory phase (Simonton calls this "problem finding" and connects it to Kuhn's revolutionary science; 2016). In either case, highly surprising outcomes significantly change scientific knowledge or the direction of research.

Surprise varies in degrees according to the magnitude of change in expected utility.[8] I take scientific utility to be context dependent: it includes true or approximately true claims, as well as scientific concepts that do not have truth values but might be judged by other normative means, such as their stability and generative power (see end of this section and Section 3.4). Also, although this concept of surprise comes from Simonton's psychological analysis of creative outcome, I take the relevant context of evaluation to be the community of experts engaged in researching a scientific problem (Simonton affirms a similar distinction between "objective" and "subjective" estimates of creative criteria; 2012). At the same time, I do not limit my account to human changes in expectation; it may be relevant to consider how outcomes change what machine "scientists" find plausible (Guimerà et al. 2020).

An example that illustrates that computational methods can make a very strong impact on scientific knowledge at even the basic level of evidence-gathering procedures, and without very sophisticated processes of algorithmic abstraction, is Krenn et al.'s (2017) discovery of a new design for a quantum optic experiment using a topological search algorithm. Their objective was to find high-dimensional, multipartite entangled quantum states by varying the possible elements

[8] In analyzing creative outcome, Simonton (2016) employs a psychological response parameter representing prior knowledge of an idea's utility. He notes that it regards *justification* for belief (footnote 3, p. 196), but Tsao et al. (2019) make a useful Bayesian distinction between expectation and uncertainty (see also Section 3.2 on blindness reduction).

of an experimental apparatus (e.g. crystals, beam splitters, prisms). They used discrete topological search since quantum states are discrete (and are not well-suited for gradient-based optimization). Krenn et al. (2020) explain that when they found a ten-dimensional entangled state, it was unexpected because the search only allowed using two crystals that were thought to each generate three-dimensionally entangled photon pairs, so they expected that the maximum dimension of entangled states would be 3 x 3 = 9 dimensional. As they investigated the anomalous design further, they found that it involved an extra event where the two crystals fire together and the paths of the entangled photons are coherently superposed. Although they found the design relied on a technique that has been explored in other contexts (the Wang, Zou, and Mandel technique employed in quantum spectroscopy, quantum imaging, and more; 1991), it had not been used for high-dimensional, multipartite quantum entanglement generation before (Krenn et al. 2017). The idea that a very high-dimensional entangled state could be generated by two crystals was *implausible* with respect to their current knowledge of the possibility space in quantum optics, but it turned out to be quite *useful* for experimental design. Thus, their topological search algorithm generated surprise.[9]

*Relationship to other types of strong novelty:*

Notably, surprise may or may not occur along with conceptual change (see Section 2.1), while changes to concepts need not be surprising. Thus, conceptual change is independent of surprise. Nonetheless, as conceptual change has a broader impact than local belief revision (see

[9] Krenn et al. (2020) also argue that the design relies on a new *concept* they call "Entanglement by Path Identity," and they were able to generalize this concept to other experimental setups. See Section 3.4 for a discussion of the role of computational techniques in generating conceptual change.

Section 3.4), surprise regarding novel concepts suggests a deep *and* wide impact. Still, surprise requires a sufficient degree of prior belief in an idea's utility to signal a meaningful change. Without this, the concepts of epistemic change I propose next might instead indicate strong novelty: reducing utility "blindness" in a domain of widespread uncertainty or eliminating deep ignorance regarding a proposition that a researcher is unaware of.

### 3.2. Blindness reduction

Another form of unexpectedness that constitutes strong novelty in scientific exploration is what Tsao et al. (2019) call "blindness reduction." Blindness reduction refers to *a decrease in uncertainty regarding an expectation of how useful an idea is*. In Bayesian terms, blindness reduction narrows the probability distribution over possible utilities. (By contrast, surprise changes the expected mean of the distribution). In some exploratory contexts, there is widespread and extreme uncertainty regarding the utility of many ideas; there, blindness reduction constitutes a strongly novel outcome, steering scientific research towards promising ideas.

An example that illustrates the potential of blindness reduction to make a profound impact on research direction is the case that Duede (2023) considers of using supervised ML to predict the locations of earthquake aftershocks from mainshock locations (DeVries et al. 2018). He argues that this case exemplifies a general exploratory strategy that is not negatively impacted by DL opacity: it involves postulating that a functional relationship exists between a certain selected dependent variable of a dataset and other independent variables and then training a DL model to test this idea. If the model's predictive accuracy is higher than chance it narrows the uncertainty that the relationship exists as postulated. The DL aftershock model achieved much greater accuracy than existing theoretical models (an AUC of 0.849 in contrast to 0.583, p.

632), which gave researchers support for the idea that theory could be significantly improved.[10] This outcome showcases blindness reduction that steers research in a context of high prior uncertainty: leading theoretical models did not do well at explaining the data much better than chance, and it was unclear which (if any) geophysical properties might improve theory. Blindness reduction prompted subsequent exploration: researchers used the spatial distribution of the DL model's aftershock predictions to identify three salient geophysical parameters that explain nearly all its variance in predictions.

*Relationship to other types of strong novelty:*

Although Tsao et al. (2019) identify surprise as the strong driver of scientific change (as with Kuhn's paradigm shifts) and blindness reduction as a weaker form of learning (Kuhn's "normal science," p. 284), this dichotomy neglects that when prior uncertainty regarding one or more ideas is extremely high, surprise is unlikely since it requires a significant degree of belief in an

[10] Notably, this case is not about reducing what Sullivan (2022) calls "link uncertainty." She takes the view that it is not opacity that threatens an ML model's ability to generate scientific understanding of target phenomena, but rather the lack of supporting scientific and empirical evidence connecting the model to the target. Thus, reducing link uncertainty should improve understanding, and it may convert "how-possibly" questions to "how-actually" ones (if such a distinction is warranted; Boge 2022, Bokulich 2014). However, reducing utility blindness is independent of reducing link uncertainty: it is possible to alleviate the former without decreasing the latter or vice versa. Duede's (2023) examples do not seem to require low link uncertainty but still demonstrate blindness reduction, and it is not clear that decreasing link uncertainty would directly contribute to this. Duede emphasizes that researchers did not make any attempt to interpret *how* the model made predictions by identifying salient features of the data.

idea's utility. In that case, learning is more sensitive to blindness reduction.[11] Therefore, although surprise and blindness reduction can co-occur in general, when blindness reduction signals high-impact learning, it is largely incompatible with surprise. Still, both surprise and blindness reduction require a state of awareness of a proposition such that it is meaningful to characterize a belief or uncertainty regarding it. ML might also play a prior role in generating this state of awareness, offering another kind of strong impact to exploratory research: eliminating deep ignorance.

### 3.3. Deep ignorance elimination

Being deeply ignorant means being in a state of unawareness of a true proposition. Eliminating deep ignorance occurs by *generating awareness of a claim that is true* (whether or not it also generates a belief in that claim's truth value). Outcomes that eliminate deep ignorance are likely to constitute strong novelty for science because they present a reason to pursue research in a novel direction (assuming they meet some relevant criteria of reliability): either to assess a claim's truth value or to attain warrant for belief in a claim's truth value.[12] Although this concept of epistemic change does not vary in degrees, the scientific impact still does (i.e. the "strength" of the novelty).

[11] Their Bayesian formulation of learning an idea's utility contains separable terms for surprise and blindness reduction and is generally more sensitive to surprise. However, they normalize their term for posterior surprise to the prior uncertainty; thus, if prior uncertainty is high, learning is driven by blindness reduction.

[12] Granted, other epistemic and practical reasons may supersede this one.

I take this concept from a propositional account of ignorance (Peels 2014). I propose it is relevant to characterizing outcomes of rational scientific activities.[13] Peels explains: "S is deeply ignorant that p iff (i) it is true that p, and (ii) S neither believes that p, nor disbelieves that p, nor suspends belief on p" (p. 485). Here, let "p" be a true claim that a certain idea is useful (taking again a context-specific notion of scientific utility). Then eliminating deep ignorance means changing unawareness of p to either belief, disbelief, or suspension of belief on p.

Examples of how ML contributes to eliminating deep ignorance in science are prolific; the size and complexity of datasets used in many scientific domains mean that researchers are often unaware of informational patterns that the data contain and what empirical regularities they might provide evidence for.[14] Krenn et al. (2022) argue that one way computational methods aid scientific understanding is to act as a "computational microscope" for viewing patterns that scientists would not otherwise be able to see. Khosrowi and Finn (2025) consider whether the pattern recognition abilities of generative AI suggest their outputs should count as novel synthetic evidence, playing a similar epistemic role to material evidence and expert judgment.

---

[13] But note that the analysis of ignorance concerns much broader issues: it includes concepts of ignorance as actively upheld false outlooks or substantive epistemic practice; see El Kassar (2018).

[14] Deep ignorance elimination is compatible with Ratti's (2020) weakest sense of novelty, which regards enlarging the scope of existing theory ("N1," pp. 88-89). However, eliminating deep ignorance concerns the exploratory aspect of gaining knowledge of a possibility space, not necessarily the acceptance of novel hypotheses. This is significant because it has broader implications. For example, Spelda and Stritecky (2021) argue that generative AI can reverse the problem of unconceived alternatives for scientific realism by helping to ascertain more of the consequences of so-far empirically equivalent theories by producing data from the "left out" regions of the possibility spaces of available evidence. This means acquiring useful modal knowledge of phenomena that might be discovered in the future, and in contrast to N1 novelty, it might lead to *revision* of theoretical commitments. Furthermore, my account differs from Ratti in that I do not view enlargement of existing theory as inherently weaker than conceptual change (see also end of Section 3.5). Thanks to an anonymous reviewer for raising this connection.

They remark that the impact would be to "provide genuinely new knowledge to agents who lack the ability to make those same inferences" (p. 2). Nonetheless, an in-principal argument that some computational tools provide epistemic access to patterns that some agents would not otherwise be able to obtain is *not* required to show deep ignorance elimination: all that is required is a change from unawareness of a useful proposition to awareness of it.

To illustrate, Ludwig and Mullainathan (2024) demonstrate the success of an approach for extracting scientific hypotheses that have never been considered before from ML models on a problem in economics. The task is to discover novel factors that have not already been identified in previous research that predict judicial decisions regarding pretrial detention. They build a supervised ML model that "fuses" decision trees and convolutional neural networks, leveraging both structured data and facial images of defendants, but their analysis focuses on the latter since it explains much of the variation in the fusion model's predictions (pp. 775–777).[15] To describe the novel features, they recruit non-expert annotators and present them with pairs of synthetic mug shots that are as similar as possible except for the difference between model's predicted detention probability, which has been maximally increased by morphing the images (pp. 756–757). The annotators identify being "well-groomed" and "heavy-faced" as correlated with the model's prediction that a defendant is more likely to be released. (They confirm this by tasking a different group of annotators with coding real mug shots with the new features.) As with blindness reduction, eliminating deep ignorance provides a new direction for future research: the

[15] The data selected includes more than fifty thousand arrests made over three years in Mecklenburg County, North Carolina (Ludwig and Mullainathan 2024, pp. 767–775).

authors note the novel features do not seem to be simple proxies for factors like substance abuse, mental health, or socioeconomic status, and so they plan to perform a causal investigation.[16]

This case is also noteworthy for how the researchers isolate the contribution of the learning algorithm to eliminating deep ignorance (Ludwig and Mullainathan, 2024). They use non-expert annotators in order to bound the creativity involved in interpreting the images: they do not want domain expertise to add to what the algorithm has already learned (p. 759). Also, they analyze whether the algorithm has detected novel sources of signal by fitting a simple linear regression (1) of actual judicial decisions (dependent variable) to the model's predictions (independent variable), and they find that it better fits the variance in judicial decisions than a regression (2) including all previously known visual traits (i.e. demographic and psychological, e.g. "attractiveness," "competence," etc., pp. 785–786). Meanwhile, a regression (3) that includes both the model's predictions and all known visual and demographic traits does not much alter the coefficient representing the model's contribution compared to (1).[17] Thus, they conclude that the algorithm has discovered genuinely novel sources of signal since controlling for all previously known factors (even a variable that incorporates existing tacit knowledge; pp. 775, 784) only modestly diminishes the predictive power of the model.

*Relationship to other types of strong novelty:*

Deep ignorance elimination does not reduce to the concepts of belief revision I have introduced previously. Instead, these can be related to Peels' (2014) other types of ignorance:

[16] This success also gives the authors confidence that their approach will be useful in a wide range of scientific contexts where the aim is to automatically generate hypotheses that humans have never considered before (Ludwig and Mullainathan, 2024).

[17] See also Table III, Ludwig and Mullainathan (2024, pp. 783–783) where (1) is column 1, (2) is column 5, (3) is column 7.

surprise might change disbelief that p to belief that p (eliminating disbelieving ignorance), and blindness reduction narrows uncertainty regarding p such that it might change suspension of belief that p to belief that p (eliminating suspending ignorance) or warrantless belief that p to warranted belief that p (eliminating warrantless ignorance). Furthermore, deep ignorance elimination is not likely to co-occur with these other concepts as a single exploratory outcome: if a researcher disbelieves or suspends belief that p once becoming aware of it, it seems to require further investigation to generate surprise or blindness reduction. Thus, the concepts of strong novelty I have discussed so far in this section are largely incompatible outcomes. They are useful for analyzing local epistemic change, but an ML outcome might also make wider impact to knowledge structure if it generates *conceptual* change.

### 3.4. Conceptual change, revisited

New concepts are generally regarded as having a strong impact on scientific knowledge and research direction: concepts are used to design experiments, identify phenomena, articulate theory that explains phenomena, and construct theories that unify other theories. In this section, I revisit the significance of conceptual change as a dimension of strong novelty, as it includes a wide variety of targets at various levels of scientific description (c.f. Section 2.1). Furthermore, I take the perspective that, like the other concepts of strong novelty I introduce, conceptual change varies in degrees and thus may not always contribute the strongest impact. Here, I will not attempt to defend any particular positive notion of what concepts are, but I will make use of Thagard's (1990) argument that conceptual change does not amount to merely local belief revision (e.g. addition or removal of conceptual features). This negative perspective affirms that concepts are not just sums of beliefs in necessary and sufficient definitional criteria or simple symbolic frames that host features and prototypical exemplars (e.g. Minsky frames; 1975). I will

also not attempt to identify conditions for distinguishing when a new concept can be clearly differentiated from prior concepts. I will argue that a variety of ML outputs might generate conceptual change, and I will consider whether it indicates a different kind of strong novelty if an ML algorithm merely prompts novel human conceptualizations, or if it directly generates novel concepts.

The negative perspective that concepts are not merely characterized by a set of necessary and sufficient definitional features suggests that it is not only unsupervised ML algorithms, which cluster data points into novel groupings of features, that might generate conceptual change. For example, if concepts are mental representations of complex structures that include kind relations, part-whole relations, instances, and rules of inference in which a concept figures such as Thagard (1990, p. 266) proposes, then a variety of ML outputs might count as candidates for the constitutive elements of concepts (e.g. symbolic equations, token exemplars, decision rules).[18] This further problematizes the simple token/type account of weak and strong novelty (see Section 2.3). Alternatively, a view of concepts as distributed patterns of neural activation at least allows that various ML outputs might act as stimulants for human changes to concepts. Moreover, it raises the questions of (1) whether some ML algorithms such as DL also form conceptualizations useful for tasks, and (2) if so, whether the outcomes constitute stronger novelty for science.

[18] Although Thagard (1990) arranges various kinds of conceptual changes in a rough order of increasing strength, where novel instances and rules figure lower, and hierarchical reorganizations of structure higher (p. 268), I highlighted that novel instances play a special role in some scientific disciplines such as astronomy (see Section 2.3). Thus, I do not attempt to generalize how various kinds of conceptual change might align with scientific impact.

An affirmative answer to (1) is plausible: some philosophers propose that DL algorithms learn category representations that count as candidates for concepts.[19] For example, López-Rubio (2021) argues that convolutional neural networks (CNNs) and generative artificial networks (GANs) build internal states that can be mapped to the kinds of complex visual categories humans use in cognitive processes (such as high-level visual stimuli, e.g. "chair" or "tree"). Also, Buckner (2018) proposes a mechanism of abstraction by which CNNs form category representations and (2020) explores whether their failures on seemingly insignificant changes to inputs (i.e. "adversarial examples") might in some cases indicate they are representing patterns in data in an "alien" way (p. 734).

An affirmative answer to (2) is more tenuous: DL-learned categories might represent data in a way that is unfamiliar to humans and thus afford the possibility of greater conceptual change. However, as López-Rubio (2021) explains, network dissection methods that associate individual neurons with a (visual) category still rely on human interpretation of objects and their contexts. This raises the problem that for exploratory activities that aim to develop new concepts based on neuron-to-world associations, there may be a large gap between DL-based discovery and understanding (aligning with Boge's account, 2022). Furthermore, these methods may be misguided as current evidence that deep neural networks represent concepts in individual units is somewhat weak (Freiesleben 2024). Nonetheless, investigating how DL abstraction might be used for conceptual change suggests a promising line of research for understanding ML's scientific impact (in contrast to negative appraisals of disruption based solely on the use of presuppositions). Notably, in contrast to the Concept-free Design Argument (Section 2.2), it is

[19] These accounts do not defend a particular constitutive account of concepts that would suggest objective criteria for the suitability of candidates.

far from clear that conceptual change need be diminished by design choices that utilize existing human concepts; instead, approaches that enforce the learning of human concepts might improve model interpretability (see Freiesleben 2024), thereby aiding conceptual change.

*Relationship to other types of strong novelty:*

Conceptual change is an important dimension of strong novelty that does not reduce to other concepts of epistemic change (Sections 3.1–3.3). Thagard (1990) identifies ten roles that concepts play in human reasoning and language and argues that local belief revisions are inadequate to explain most of them; rather, they require accounting for conceptual structure (pp. 258–259). Meanwhile, he notes the simple belief-revision approach never addresses the origin of concepts, but a structural approach suggests concepts can be generated by example and by combining previous concepts (pp. 259–260). A neural perspective also problematizes the simple belief revision account—López-Rubio (2021) shows how GANs learn constraints along with similarities based on instances: inserting an object into a visual scene also modifies the area around it (p. 10023).

While I have argued that conceptual change need not be impeded by prior conceptualizations of target phenomena (Section 2), intuitively, how and what prior information is utilized to generate predictions seems relevant to assessing their novelty and impact. As it is still unclear what kind of theoretical bias might reduce the potential for scientific change, I next propose a final concept of strong novelty that directly addresses this.

### 3.5. Local theory-independent learning from data

I argue that outcomes that are *learned from data* with some *independence of local theory* regarding a target phenomenon are good candidates for strong novelty. Local theory means

theory that demarks or explains a phenomenon that is a target of investigation. Independence selects outcomes that are not strongly determined by local theory. Learning in this "bottom-up" way constitutes a form of strong novelty for science because it signals an *aim* to find a new research direction, often by relying on a different set of cognitive tools for analyzing complex systems. Thus, although this concept places a minimal process condition on an outcome, such that it is not assessed by impact only but by whether the outcome is arrived at in the "right sort of way," it helpfully signals scientific contexts in which existing theory is open for substantial enlargement or revision.

I derive this concept of strong novelty from the philosophical literature on exploratory experimentation. In general, exploratory experiments are activities that do not aim to test particular hypotheses and that involve extensive variation of parameters (Elliott 2007). Franklin (2005) describes a type of mapping activity where theory plays a background role in guiding researchers to relevant properties, but it does not function "locally" in the sense of describing causal hypotheses regarding map constituents (p. 893). She highlights a case that demonstrates the strong impact of maps on subsequent research: a DNA microarray allows biologists to collect "wide" data regarding how the levels of all kinds of mRNA vary over the cell cycle, and it enabled the Spellman group (1998) to eventually formulate hypotheses that identified and explained various similarities in the mapped gene behaviors.

While exploratory science encompasses a wide range of activities, I add the data-driven learning condition to this concept of strong novelty to select activities that employ methods designed to characterize data in a novel way (e.g. unsupervised ML, supervised representation learning). For example, Chattopadhyay et al. (2019) aim to generate a novel galaxy classification scheme directly from observational data (i.e. data that has not been reduced in the usual way; for

example, to typical ratios of spectral line emissions known to be useful for distinguishing phenomena). They employ linear independent component analysis (ICA) on a set of 49 observable attributes covering a range of physical characteristics, followed by unsupervised K-means clustering to generate ten novel galaxy classes. These steps demonstrate local theory-independent learning from data: the ICA representation learning technique omits theoretical assumptions about which components are best for representing the most robust regularities, and K-means clustering proceeds without prior assumptions about how to demark particular classes in the ICA space. Nonetheless, inferences involving linear ICA still rely on the assumption that relevant signals can be identified by linear combinations of observable features, which might be theoretically motivated or a simplifying assumption. If the former, it highlights that this concept of strong novelty varies in degrees. If the latter, it does not diminish the potential impact to subsequent research.

In contrast, Parker et al.'s (2024) foundation model for galaxy objects might achieve a greater degree of independence of local theory. They align image and spectral data in a novel representation space with contrastive (self-supervised) learning. As contrastive learning strategies do not assume linear independence, these would involve fewer assumptions about the data-generating process (which might be based on local theory). I leave a detailed consideration of these cases and the kind(s) of novelty introduced by foundation models to future work, but I emphasize here that local theory-independent learning from data is characterized by an aim to suspend existing theory regarding target phenomena to find a novel direction for subsequent research that attempts to demark or explain them.

*Relationship to other types of novelty*

Local theory-independent learning from data is distinct from the other types of strong novelty in my taxonomy by how it accounts for the operationalization of existing theory. As it is sensitive to *how* outcomes are generated, it might co-occur with the concepts that are not, suggesting further ways to differentiate types of strong impact.

Moreover, this concept of strong novelty is better suited for analysis in the context of scientific exploration than existing presupposition-based accounts. Particularly, it clarifies that *local theory* is the kind of prior information that might diminish the scientific impact of a prediction. Only when local theory (that demarks or explains phenomena) is highly fixed is the degree of learning highly constrained. However, local presuppositions might be used in a provisional, exploratory manner (e.g. iterating between identification and refinement of target phenomena, c.f. Section 2.3). Also, learning outcomes achieved with independence from local theory signal *deep* impact (according to the degree of independence, as do the concepts of local epistemic change in Sections 3.1–3.3), while conceptual change signals *wide* impact (which may not necessarily be major). Thus, in contrast to Ratti's (2020) account, I do not take enlargement of existing theory ("N1") to be inherently "weaker" than conceptual change ("N3"). Furthermore, my account is more general than his, and some exploratory activities might aim to generate novelty along *both* dimensions in scientific domains without stabilized concepts.

In addition, while independence of local theory is less sensitive to prior knowledge and presuppositions about phenomena than use novelty (see Section 2.4), it is *more* sensitive to theoretical choices regarding data and model interpretation. It is not curtailed merely by design choices to use labels for supervised learning or inductive biases, if these are independent of fixed, local theory. However, it captures that theory might constrain novelty at various stages of an ML pipeline. Furthermore, it may be more amenable to historical analysis than use novelty:

the aim to break with existing theory in order to substantially recharacterize a target phenomenon is often stated as a project aim in the published record.[20]

## 4. Discussion

A major aim of identifying types of strong novelty for science is to help design new research projects and agendas. My taxonomy should facilitate cross-disciplinary engagement as these novelty desiderata might be relevant to a wide range of scientific fields. For example, the general strategy for reducing utility blindness with neural networks (Section 3.2) can be applied to problems from a variety of domains and is useful for assessing the pursuit-worthiness of a new idea. Similarly, ML is likely to make a large scientific impact when the aim is to eliminate deep ignorance, particularly in data-intense projects where computational tools help to overcome practical and epistemic barriers to discovery.

Also, my taxonomy should help to nuance claims of ML's impact on science by attending to a wide range of its uses (perhaps restructuring the divide between scientific optimism and philosophical pessimism, Duede 2023). This might include tracking the success of various algorithms in contributing to scientific change. While statistical learning theory provides guarantees for the generalization performance of some types of models (Luxburg & Schölkopf 2011), it does not address how to ensure that various scientific desiderata are met (but see Freiesleben et al. for an account of scientific inference with interpretable ML, 2022). Assigning credit to ML systems for their past contributions to discovery is likely to play at least some normative role in justifying the choice of a computational approach for a given exploratory aim.

[20] On the other hand, some versions of use novelty, such as the knowledge-based concept discussed in Section 2.4, require making additional historical inferences regarding the epistemic states of the scientists.

At the same time, designing new research projects requires a careful examination of the capacities of ML algorithms. My top-down analysis complements these assessments but would also benefit from future engagement with bottom-up perspectives since each approach has its limitations; top-down requires connecting desiderata to design criteria, which may be unavailable or without guarantee of success, while beginning with the novelty of processes (e.g. their creativity or goal-directedness) risks ultimate misalignment between means and ends. Finally, bi-directional engagement is likely to be fruitful because while the reliability of processes is relevant to designing for novelty, novelty also plays a role in the reliability of predictive processes. For example, reliable statistical inference methods such as extrapolation seem significant both for the learning outcomes they generate (perhaps affording epistemic access to a novel domain) and for how their outcomes contribute to the robustness of predictive processes (consider Cranmer et al. 2021, Freiesleben & Grote 2023, Grote et al. 2024). I expect my taxonomy to add clarity to the dialectic between outcome and process-centered accounts of strong novelty.

## 5. Conclusion

I have presented concepts of strong novelty that suggest a variety of ways that ML makes high impact to science. Conceptual change indicates a broad form of scientific impact that is not reducible to local notions of epistemic change. However, to fully appreciate the ways that ML advances science, philosophical consideration of novelty and ML must move beyond conceptual change. Surprising outcomes change scientific beliefs, while reducing utility blindness and eliminating deep ignorance play a key progressive role in exploratory contexts. Meanwhile, the concept of local theory-independent learning from data is a better starting point for future

reflection on what design choices are likely to achieve strong novelty than use novelty or mere token/type distinctions: it places a minimal constraint on the theory incorporated into an ML project.

## References


Andrews, Mel. "The Immortal Science of ML: Machine Learning & the Theory-Free Ideal," 2024. Pre-print at https://philsci-archive.pitt.edu/23840/.

Barnes, Eric Christian. "Prediction versus Accommodation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Winter 2022. Metaphysics Research Lab, Stanford University, 2022. https://plato.stanford.edu/archives/win2022/entries/prediction-accommodation/.

Boden, Margaret A. "Computer Models of Creativity." *AI Magazine* 30, no. 3 (2009): 23–34. https://doi.org/10.1609/aimag.v30i3.2254.

———. *The Creative Mind: Myths and Mechanisms*. 2nd Ed. London: Routledge, 2004. https://doi.org/10.4324/9780203508527.

Boge, Florian J. "Two Dimensions of Opacity and the Deep Learning Predicament." *Minds and Machines* 32 (2022): 43–75. https://doi.org/10.1007/s11023-021-09569-4.

Bokulich, Alisa. "How the Tiger Bush Got Its Stripes: 'How Possibly' vs. 'How Actually' Model Explanations." *The Monist* 97, no. 3 (2014): 321–38. https://doi.org/10.5840/monist201497321.

Buckner, Cameron. "Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks." *Synthese* 195, no. 12 (2018): 5339–72. https://doi.org/10.1007/s11229-018-01949-1.

———. "Understanding Adversarial Examples Requires a Theory of Artefacts for Deep Learning." *Nature Machine Intelligence* 2, no. 12 (2020): 731–36. https://doi.org/10.1038/s42256-020-00266-y.

Campbell, Donald T. "Blind Variation and Selective Retentions in Creative Thought as in Other Knowledge Processes." *Psychological Review* 67, no. 6 (1960): 380–400. https://doi.org/10.1037/h0040373.

Chattopadhyay, Tanuka, Didier Fraix-Burnet, and Saptarshi Mondal. "Unsupervised Classification of Galaxies. I. Independent Component Analysis Feature Selection." *Publications of the Astronomical Society of the Pacific* 131, no. 1004 (2019): 108010. https://doi.org/10.1088/1538-3873/aaf7c6.

Clark, Elinor, and Donal Khosrowi. "Decentring the Discoverer: How AI Helps Us Rethink Scientific Discovery." *Synthese* 200, no. 6 (2022): 463. https://doi.org/10.1007/s11229-022-03902-9.

Cranmer, Miles, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. "Discovering Symbolic Models from Deep Learning with Inductive Biases." In *Advances in Neural Information Processing Systems*, 33:17429–42. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper/2020/hash/c9f2f917078bd2db12f23c3b413d9cba-Abstract.html.

Cranmer, Miles, Daniel Tamayo, Hanno Rein, Peter Battaglia, Samuel Hadden, Philip J. Armitage, Shirley Ho, and David N. Spergel. "A Bayesian Neural Network Predicts the Dissolution of Compact Planetary Systems." *Proceedings of the National Academy of Sciences* 118, no. 40 (2021): e2026053118. https://doi.org/10.1073/pnas.2026053118.

DeVries, Phoebe M. R., Fernanda Viégas, Martin Wattenberg, and Brendan J. Meade. "Deep Learning of Aftershock Patterns Following Large Earthquakes." *Nature* 560, no. 7720 (2018): 632–34. https://doi.org/10.1038/s41586-018-0438-y.

Douglas, Heather, and P. D. Magnus. "State of the Field: Why Novel Prediction Matters." *Studies in History and Philosophy of Science Part A* 44, no. 4 (2013): 580–89. https://doi.org/10.1016/j.shpsa.2013.04.001.

Duede, Eamon. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science* 90, no. 5 (2023): 1089–99. https://doi.org/10.1017/psa.2023.8.

El Kassar, Nadja. "What Ignorance Really Is. Examining the Foundations of Epistemology of Ignorance." *Social Epistemology* 32, no. 5 (2018): 300–310. https://doi.org/10.1080/02691728.2018.1518498.

Elliott, Kevin C. "Varieties of Exploratory Experimentation in Nanotoxicology." *History and Philosophy of the Life Sciences* 29, no. 3 (2007): 313–36.

Franklin, L. R. "Exploratory Experiments." *Philosophy of Science* 72, no. 5 (2005): 888–99. https://doi.org/10.1086/508117.

Freiesleben, Timo. "Artificial Neural Nets and the Representation of Human Concepts." arXiv, March 26, 2024. https://doi.org/10.48550/arXiv.2312.05337.

Freiesleben, Timo, and Thomas Grote. "Beyond Generalization: A Theory of Robustness in Machine Learning." *Synthese* 202, no. 4 (2023): 109. https://doi.org/10.1007/s11229-023-04334-9.

Freiesleben, Timo, Gunnar König, Christoph Molnar, and Alvaro Tejero-Cantero. "Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena." arXiv, 2022. https://doi.org/10.48550/arXiv.2206.05487.

Gardner, Michael R. "Predicting Novel Facts." *The British Journal for the Philosophy of Science* 33, no. 1 (1982): 1–15. https://doi.org/10.1093/bjps/33.1.1.

Genin, Konstantin, Thomas Grote, and Thomas Wolfers. "Computational Psychiatry and the Evolving Concept of a Mental Disorder." *Synthese* 204, no. 3 (2024): 88. https://doi.org/10.1007/s11229-024-04741-6.

Grote, Thomas, Konstantin Genin, and Emily Sullivan. "Reliability in Machine Learning." *Philosophy Compass* 19, no. 5 (2024): e12974. https://doi.org/10.1111/phc3.12974.

Guimerà, Roger, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel Miranda, Jordi Pallarès, and Marta Sales-Pardo. "A Bayesian Machine Scientist to Aid in the Solution of Challenging Scientific Problems." *Science Advances* 6, no. 5 (2020): eaav6971. https://doi.org/10.1126/sciadv.aav6971.

Halina, Marta. "Insightful Artificial Intelligence." *Mind & Language* 36, no. 2 (2021): 315–29. https://doi.org/10.1111/mila.12321.

Iten, Raban, Tony Metger, Henrik Wilming, Lidia del Rio, and Renato Renner. "Discovering Physical Concepts with Neural Networks." *Physical Review Letters* 124, no. 1 (2020): 010508. https://doi.org/10.1103/PhysRevLett.124.010508.

Kashinath, K., M. Mustafa, A. Albert, J-L. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, et al. "Physics-Informed Machine Learning: Case Studies for Weather and Climate Modelling." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379, no. 2194 (February 15, 2021): 20200093. https://doi.org/10.1098/rsta.2020.0093.

Khosrowi, Donal, and Finola Finn. "Can Generative AI Produce Novel Evidence?" *Philosophy of Science* (2025). https://philsci-archive.pitt.edu/24775/.

Krenn, Mario, Manuel Erhard, and Anton Zeilinger. "Computer-Inspired Quantum Experiments." *Nature Reviews Physics* 2, no. 11 (2020): 649–61. https://doi.org/10.1038/s42254-020-0230-4.

Krenn, Mario, Armin Hochrainer, Mayukh Lahiri, and Anton Zeilinger. "Entanglement by Path Identity." *Physical Review Letters* 118, no. 8 (2017): 080401. https://doi.org/10.1103/PhysRevLett.118.080401.

Krenn, Mario, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, et al. "On Scientific Understanding with Artificial Intelligence." *Nature Reviews Physics* 4, no. 12 (2022): 761–69. https://doi.org/10.1038/s42254-022-00518-3.

Lakatos, I. "Falsification and the Methodology of Scientific Research Programmes." In *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*, edited by Alan Musgrave and Imre Lakatos, 91–196. Cambridge: Cambridge University Press, 1970. https://doi.org/10.1017/CBO9781139171434.009.

López-Rubio, Ezequiel. "Throwing Light on Black Boxes: Emergence of Visual Categories from Deep Learning." *Synthese* 198 (2021): 10021–41. https://doi.org/10.1007/s11229-020-02700-5.

Ludwig, Jens, and Sendhil Mullainathan. "Machine Learning as a Tool for Hypothesis Generation." *The Quarterly Journal of Economics* 139, no. 2 (2024): 751–827. https://doi.org/10.1093/qje/qjad055.

Luxburg, Ulrike von, and Bernhard Schölkopf. "Statistical Learning Theory: Models, Concepts, and Results." In *Handbook of the History of Logic*, edited by Dov M. Gabbay, Stephan Hartmann, and John Woods, 10:651–706. Inductive Logic. North-Holland, 2011. https://doi.org/10.1016/B978-0-444-52936-7.50016-1.

Maher, Patrick. "Prediction, Accommodation, and the Logic of Discovery." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1988, no. 1 (January 1988): 272–85. https://doi.org/10.1086/psaprocbienmeetp.1988.1.192994.

McCabe, Michael, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, et al. "Multiple Physics Pretraining for Physical Surrogate Models." arXiv, October 4, 2023. https://doi.org/10.48550/arXiv.2310.02994.

Minsky, Marvin. "A Framework for Representing Knowledge." In *Mind Design II: Philosophy, Psychology, and Artificial Intelligence, The MIT Press, 1997*, edited by John Haugeland, 111–42, 1975. https://doi.org/10.7551/mitpress/4626.003.0005.

Parker, Liam, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, et al. "AstroCLIP: A Cross-Modal Foundation Model for Galaxies." *Monthly Notices of the Royal Astronomical Society* 531, no. 4 (2024): 4990–5011. https://doi.org/10.1093/mnras/stae1450.

Paul, Elliot Samuel, and Dustin Stokes. "Attributing Creativity." In *Creativity and Philosophy*, edited by Berys Gaut and Matthew Kieran. Routledge, 2018.

Peels, R. "What Kind of Ignorance Excuses? Two Neglected Issues." *The Philosophical Quarterly* 64, no. 256 (2014): 478–96. https://doi.org/10.1093/pq/pqu013.

Ratti, Emanuele. "What Kind of Novelties Can Machine Learning Possibly Generate? The Case of Genomics." *Studies in History and Philosophy of Science Part A* 83 (2020): 86–96. https://doi.org/10.1016/j.shpsa.2020.04.001.

Simonton, Dean Keith. "Creativity, Automaticity, Irrationality, Fortuity, Fantasy, and Other Contingencies: An Eightfold Response Typology." *Review of General Psychology* 20, no. 2 (2016): 194–204. https://doi.org/10.1037/gpr0000075.

———. "Taking the U.S. Patent Office Criteria Seriously: A Quantitative Three-Criterion Creativity Definition and Its Implications." *Creativity Research Journal* 24, no. 2–3 (2012): 97–106. https://doi.org/10.1080/10400419.2012.676974.

———. "The Blind-Variation and Selective-Retention Theory of Creativity: Recent Developments and Current Status of BVSR." *Creativity Research Journal*, 2022, 1–20. https://doi.org/10.1080/10400419.2022.2059919.

Spelda, Petr, and Vit Stritecky. "What Can Artificial Intelligence Do for Scientific Realism?" *Axiomathes* 31, no. 1 (2021): 85–104. https://doi.org/10.1007/s10516-020-09480-0.

Spellman, Paul T., Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. "Comprehensive Identification of Cell Cycle–Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization." *Molecular Biology of the Cell* 9, no. 12 (1998): 3273–97. https://doi.org/10.1091/mbc.9.12.3273.

Sullivan, Emily. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* 73, no. 1 (2022). https://doi.org/10.1093/bjps/axz035.

Thagard, Paul. "Concepts and Conceptual Change." *Synthese* 82, no. 2 (1990): 255–74.

Thuemmel, Jannik, Matthias Karlbauer, Sebastian Otte, Christiane Zarfl, Georg Martius, Nicole Ludwig, Thomas Scholten, et al. "Inductive Biases in Deep Learning Models for Weather Prediction." arXiv, April 30, 2024. http://arxiv.org/abs/2304.04664.

Tsao, J. Y., C. L. Ting, and C. M. Johnson. "Creative Outcome as Implausible Utility." *Review of General Psychology* 23, no. 3 (2019): 279–92. https://doi.org/10.1177/1089268019857929.

Wang, L. J., X. Y. Zou, and L. Mandel. "Induced Coherence without Induced Emission." *Physical Review A* 44, no. 7 (1991): 4614–22. https://doi.org/10.1103/PhysRevA.44.4614.

Worrall, John. "Scientific Discovery and Theory-Confirmation." In *Change and Progress in Modern Science*, with Joseph C. Pitt. The University of Western Ontario Series in Philosophy of Science. Springer Netherlands, 1985. https://doi.org/10.1007/978-94-009-6525-6_11.

Worrall, John. "The Ways in Which the Methodology of Scientific Research Programmes Improves on Popper's Methodology." In *Progress and Rationality in Science*, with Gerard Radnitzky and Gunnar Andersson. Boston Studies in the Philosophy of Science. Springer Netherlands, 1978. https://doi.org/10.1007/978-94-009-9866-7_3.

Yao, Siyu. "Excavation in the Sky: Historical Inference in Astronomy." *Philosophy of Science* 90, no. 5 (2023): 1385–95. https://doi.org/10.1017/psa.2023.22.

Zednik, Carlos, and Hannes Boelsen. "Scientific Exploration and Explainable Artificial Intelligence." *Minds and Machines* 32, no. 1 (2022): 219–39. https://doi.org/10.1007/s11023-021-09583-6.

## Acknowledgements

I would like to thank Donal Khosrowi for insightful discussions regarding the significance of recent cases of scientific discovery with machine learning and the role of inductive biases, as well as for encouraging me to consider pertinent literature regarding the epistemology of ignorance. I would also like to thank Thomas Grote and Chris Smeenk for helpful discussions and feedback on this paper, which helped to refine it. I am also grateful for conversations with

others at the University of Tübingen, Leibniz University Hannover, and Kristian Gonzalez Barman.

**Funding Information**

This paper draws on research supported by (1) the Social Sciences and Humanities Research Council of Canada, (2) the German Academic Exchange Service (Deutscher Akademischer Austauschdienst), and (3) the German Science Foundation (Excellence Cluster 2064 "Machine Learning – New Perspectives for Science," project number 390727645).

(1) Ce article s'appuie sur des recherches financées par le Conseil de recherches en sciences humaines du Canada.