Science in the Age of Algorithms

Sendhil Mullainathan Ashesh Rambachan*

October 10, 2025

Abstract

Algorithms will not simply contribute to science; they will reorganize it. We sketch how science will look in the near future. Notably, algorithms will formalize crucial parts of science that currently happen off screen — such as new idea generation or intuitions about what theories matter. They will enable fundamentally different modes of theorizing. These changes will most affect what we call the "patchwork sciences," which includes fields such as economics, chemistry, biology, medicine and psychology. In short, we argue that the methods of science will look very different in a world where algorithms can process data, reason and form their own models of the world.

1 Introduction

At long enough time scales, the pace of human progress is dictated by the pace of scientific progress. So, if AI is to accelerate growth, it needs to accelerate science. There are reasons to believe it will: algorithms already show facility with both data and theory. They have been effective in applying known theories — from AlphaFold's protein structure predictions (Jumper et al., 2021) to solving partial differential equations in fluid dynamics (Kochkov et al., 2021). They can find patterns in data overlooked by people (Mullainathan and Spiess, 2017; Athey, 2018). They even show promise in formal reasoning, so much so that automated proof and conjecture generation is spreading through mathematics research (e.g., Davies et al., 2021; Fawzi et al., 2022; Trinh et al., 2024; Feldman and Karbasi, 2025). The question, therefore, is not whether algorithms will accelerate science, but how.

To answer that question, we look to a historical analogy. Electricity once promised to reshape manufacturing, and it spread quickly. Manufacturers desiring cheaper and more reliable power swapped out steam engines for electric motors. Yet rapid adoption did not yield immediate benefits. The real transformation came later, with a crucial realization: electricity did not merely power machines — it removed a constraint. Existing factory layouts were constrained by the steam engine. Electricity-powered factories could lay out machinery to optimize production, and once

*Mullainathan: Massachusetts Institute of Technology and NBER (sendhil@mit.edu). Rambachan: Massachusetts Institute of Technology (asheshr@mit.edu). We thank Ajay Agrawal for helpful comments, and we are especially grateful to our discussant John McHale.

new factories were built, the anticipated productivity boom materialized (Devine, 1983; David, 1990). The biggest impact of electricity was to change the factory floor.[1] The biggest impact of AI will be to change the factory floor of science.

This chapter sketches a vision for that new factory floor. Our vision rests on an observation about current scientific practice. Thinking about the new factory reveals a problem with the old factory floor — or, more specifically, a problem with our implicit presumptions about what science is.

We might say science is the careful testing of and adjudication between theories, and that those theories are self-contained, precise, and parsimonious. The hope is that this iterative process over time converges to the correct theory. Along that convergence path, a currently-best theory provides leverage over the world. Scientists discover truths that others — engineers, for instance — can apply to practical problems. The development of Newtonian mechanics perhaps best exemplifies this view.

The practice of science often doesn't match the conceit of science. Take a specific example from economics: how people make choices over time. A natural and initial theory was exponential discounting. A major breakthrough was the recognition that people have self-control problems and that a concise theory — (quasi-)hyperbolic discounting — could model them. Though substantial data supported this new idea, it did not "dethrone" exponential discounting. Both theories were used; sometimes the same researcher would use exponential discounting for one problem and choose hyperbolic discounting for another. In fact researchers today draw on a portfolio of theories: dual-self models, projection bias, cognitive models and so on. They pick and choose between models through some combination of intuition and past experiences (some of which are empirical). You hear phrases like "self control is not a first order problem in this application."[2] Perhaps most important, we are not done — there is more to be understood about intertemporal choice.

The stylized view of sciences doesn't quite fit this example. While it focuses on the testing new theories (e.g. hyperbolic discounting), it excludes the *generation* of such theories. That is left as an informal, creative human activity. It also does not account for the simultaneous persistence of multiple *inconsistent* theories. Hyperbolic discounting did not displace exponential discounting; it just joined the pantheon. Importantly, it does not account for the *human* intuition (not precise testing) that adjudicates which theories to apply where.

These activities need not be outside of science. New ideas do not emerge from thin air. Hyperbolic discounting came from noting behaviors inconsistent with exponential discounting. Similarly,

[1] David and Wright (2006) argue that general purpose technologies" yield benefits only after extensive co-investments in complementary organizational and business practices. Brynjolfsson, Rock and Syverson (2019, 2021) apply this to AI, proposing that the broader economy may be on a productivity J-curve" where productivity gains await complementary investments.

[2] Researchers might also decline to include hyperbolic discounting due to tractability concerns—the model becomes significantly more complex. But this too is implicitly a judgment that the mechanism is not important enough in the setting to justify the added complexity. The decision reflects a belief about what is first-order, not merely a computational constraint.

our intuitions about which theories to apply where simply capture past experiences applying these theories. These activities all combine empiricism and formalism — the hallmarks of science.

Algorithms allow us to bring these key elements on stage, and make them a formal part of science. For example, we can formalize the idea of hypothesis and theory generation; and build tools that automate parts of that activity. So, our vision for the new factory floor is a marriage of two forces: the new capacities enabled by algorithms and the pre-existing mismatches between the conceit and practice of science.

The new factory floor applies to a broad swath of sciences, such as psychology, medicine, chemistry and biology. We will argue that all these sciences — surprisingly — have more in common with economics than they do with Newtonian mechanics. All these sciences maintain multiple conflicting theories. All have pockets of understanding against a background of mystery. Even in disciplines where the "underlying" laws are known (e.g. quantum mechanics and chemistry), at the phenomenological level (e.g. molecules) they struggle in the same way as fields where underlying laws are unknown. There is so much science left to do precisely because so much remains to be understood. We call these the "patchwork sciences," and they are our focus.

The theme of this book is to imagine transformative AI — systems approaching or exceeding human expert capabilities across domains. Our assumptions are more modest. We focus on algorithms that *already exist* or could emerge through ordinary engineering advances, rather than extrapolating scaling laws or assuming AGI breakthroughs. Our decidedly conservative stance belies a radical claim. AI will transform science even before "transformative" AI is achieved (and even whether or not it is achieved).

Finally, a note on scope: this chapter draws extensively on our own research. That reflects bounded imagination rather than exaggerated self-importance. Re-imagining the factory floor requires concrete examples of algorithms; and those we have built are simply more tangible to us.[3] It is hard to see the future through eyes other than our own.

2 Rethinking the Current Factory Floor

We begin with the conventional view of scientific progress — what we might call the "reigning champion" view. In this view, science maintains a current best theory. When a new challenger arrives, we run empirical tests: derive predictions where the theories conflict, gather evidence, and crown a winner. If the new theory better explains the evidence, it supplants the incumbent. This narrative appears throughout textbooks and popular accounts of scientific progress — for example, Kepler's model dethroning the Ptolemaic model.

[3]Moreover, we cannot survey the full landscape of AI for science with the depth this argument demands; for broader surveys, see Raghu and Schmidt (2020), Krenn et al. (2022) and Wang et al. (2023) as well as Liang (2025) in economics.

No doubt the reigning champion view captures an essential aspect of science. Yet there are large gaps between it and actual scientific practice. Even if we have recognized the gaps between idealized and actual scientific practice, we have lacked tools to address them, filling them instead with informal human processes. This section maps these gaps — not as failures, but as opportunities for where algorithms can transform science.

2.1 How Do New Theories Emerge?

Testing the reigning champion against new entrants begs the question: where do challengers come from? For example, expected utility theory reigned as the dominant model of decision-making under risk after Von Neumann and Morgenstern formalized it in 1944. Then came challengers — prospect theory, rank-dependent utility, salience theory. How did they arise?

The answer reveals a structured process we call "anomaly generation." In 1953, Allais constructed two choice menus (which we now know as the Allais paradox) in which reasonable choices violate expected utility theory's independence axiom (Allais, 1953). Then came the common ratio effect — another minimal example illustrating a violation. Then came Kahneman and Tversky's catalog of anomalies, which led directly to prospect theory (Kahneman and Tversky, 1979). The process continues today: new researchers proposing new anomalies.

By "anomaly" we mean carefully curated *minimal* examples that a theory, however stretched, cannot explain. Their power lies in their simplicity: Allais needed only two choice menus — two carefully constructed data points — to challenge the reigning champion. These minimal constructions recur throughout science. In physics, the double-slit experiment uses light passing through one slit, then two, to produce the smallest possible demonstration that light behaves as both particle and wave. In other parts of economics, game theory advanced through carefully constructed counterexamples that exposed gaps in bargaining theory and strategic reasoning. Anomalies have driven scientific progress across domains.

How were these famed anomalies constructed? Someone observed the world and produced an empirical intuition. They contrasted that empirical intuition with their understanding of the theory. When they find a mismatch, they carefully curated a minimal example that distilled the problem. Every aspect of this process is currently human. Allais relied on intuition about choices, knowledge of expected utility's axioms, and creativity to construct revealing menus. Yet every aspect could, in principle, be algorithmic.

In recent work, we built algorithms that automatically generate candidate anomalies (Mullainathan and Rambachan, 2025). The inputs are simple: a dataset and a formal theory. The outputs are minimal examples where the algorithm's predictions conflict with the theory. The central idea is straightforward: train a black-box model on the data to capture empirical patterns, and then adversarially search for minimal datasets where the black-box and theory diverge sharply.

Applied to expected utility theory, the algorithm rediscovers anomalies like the Allais paradox and the common ratio effect. It also generates novel anomalies absent from existing work on risky choice. When we test these in incentivized experiments, people violate expected utility at rates comparable to canonical anomalies.

This work brings a critical piece of scientific practice from off screen to on-stage. Probing the reigning champion for instructive failures need not be an act of human creativity. Because it involves data and search, it is exactly something algorithms can play a huge role in. The work of science can expand from "Is this theory right?" to "Where and how does it fail?"

Anomaly generation is only one form of off-screen activity in science. Hypothesis generation is another — the deeply intuitive leaps that precede formal theory construction and entirely absent from our formal scientific machinery. Ludwig and Mullainathan (2024) bring this activity into the limelight by building an algorithmic approach to hypothesis generation. In studying judicial decisions, a model trained on mugshot pixels predicts pretrial release decisions far better than models using structured defendant characteristics. Algorithmically generated examples are produced to elicit novel hypotheses from human subjects: judges are influenced by facial traits like "well-groomed" and "heavy-faced" — hypotheses absent from prior work.

Taken together, anomaly generation and hypothesis generation reveal a pattern: much of the work that produces new competitors and scientific progress is itself structured work, currently happening off-screen in informal human processes. The reigning champion view focuses on meticulously testing theories once they exist. Algorithms can now participate in the earlier stage: generating the challengers themselves.

2.2 Do We Converge on a Single Theory?

The next omission from the reigning champion view can be seen in another concrete economics example. Consider modeling health insurance markets. Do you focus on asymmetric information? Behavioral confusion? Strategic interactions? Expected utility or prospect theory? Exponential or hyperbolic discounting? Each choice yields a different model — none obviously wrong.

Instead of a reigning champion, we have a portfolio of theories. These theories each focus on a different aspect of the problem. They may even make inconsistent assumptions. Behavioral economics illustrates this starkly. The field offers probability weighting, hyperbolic discounting, loss aversion, models of limited attention. Yet we deploy them selectively. In many contexts, we still reach for expected utility or exponential discounting, despite knowing they fail in specific ways. This is not temporary – it will not be resolved once we "figure out" the right theory. Instead we have an arsenal.

Why do we accumulate an arsenal of theories? The reason is that no single model achieves high "completeness" across the settings we study. Fudenberg et al. (2022) formalize this by comparing

theories to machine learning benchmarks — proxies for the best possible predictor given available features. The results vary dramatically, both across theories and across contexts. Some theories explain over 90% of predictable variation in a specific setting; whereas others explain barely a fraction. This is not because we lack cleverness. The world we confront is incredibly rich, and any tractable theory must abstract away most of that complexity.

2.3 How Are Theories Applied?

The reigning champion view imagines science as a tournament. If we accept that we are in a patchwork science, then what does it mean to crown a winner? We are not accumulating a single victorious theory but rather tools, each with strengths and limitations, each applicable in some contexts and not others. This matches the old maxim - "all models are wrong, some are useful." But how do we decide which ones are useful where?

Return to the health insurance example. Imagine a researcher presenting a health insurance model: asymmetric information, expected utility, Bertrand competition. The audience nods along. But if the researcher also modeled gym membership as a signaling device, the audience would say: 'That mechanism doesn't seem first-order here.'

The phrase 'first-order' appears constantly in applied work, loosely meaning: this mechanism matters; that one does not. But we have no formal criteria. Our decisions rest on tacit judgments about which mechanisms matter in each context. Ask someone why a mechanism is first-order in one setting but not another, and you might get appeals to intuition or past experience.

While this may sound imprecise, it would be foolish to dismiss these insights. The discipline has accumulated real knowledge about what matters where, passed down through papers, seminars, and advising relationships. The problem is not that these judgments are worthless — it is that they happen entirely off-screen, never formalized as part of the scientific process.

This reveals a central challenge. We are accumulating knowledge not only through controlled experiments, but through application: using theories to make sense of the world, observing what works, building intuition about when each tool applies. In this sense, application is not a second-class activity that happens after theory is built and tested. Application is how we learn the pivotal piece of knowledge in patchwork sciences: not which theory is right, but which theory is useful where.

Yet, across a wide variety of domains, machine learning algorithms outperform expert judgment at prediction (e.g., Kleinberg et al., 2018; Mullainathan and Obermeyer, 2021; Rambachan, 2024; Mullainathan, 2025). If algorithms outperform us at prediction, why would our intuitive judgments about "first-order" mechanisms be superior? For algorithms to play a larger role in this essential activity of deployment, they must somehow engage with our tacit knowledge. They must learn not just to predict, but to organize understanding in ways that respect the structure we believe matters — and do so in a form that allows us to evaluate when that structure is appropriate and

when it needs revision.

3 The Patchwork Sciences

The questions we have posed do not apply universally. Classical mechanics at human scales — modeling planetary motion or projectile trajectories — requires only a compact set of laws from which we derive precise predictions. The theory is complete at the scale of measurement. Quantum mechanics at atomic scales is similar: computationally intensive, but not fundamentally incomplete at the scales where it applies.

Where, then, does our discussion apply? It applies to fields lacking a small set of laws that govern outcomes at the scale we measure. As alluded to earlier, we call these the *patchwork sciences*. They include not only economics and psychology, but also, surprisingly, several "hard" sciences as well.

3.1 The Obvious Cases

For the social sciences, this characterization is perhaps obvious. Economics aspired to physics-like unification; though, that unification never arrived. Instead, we have a library of models deployed selectively by context: expected utility theory for some decisions, hyperbolic discounting for intertemporal choice, loss aversion for framing effects, limited attention for information processing. No single model explains behavior across contexts.

Psychology operates similarly. We have distinct theories of memory (working memory, long-term memory, episodic versus semantic), attention (selective, divided, sustained), and learning (classical conditioning, operant conditioning, observational learning). These are not chapters in a unified theory of mind — they are separate frameworks applied at different levels of analysis in different contexts.

One might object: surely if we understood the underlying neuroscience, we could derive these phenomena from first principles. Model every neuron and out pops behavior. But this misses the point. The models we use are not approximations of some deeper neuroscience. The models we use operate at the level where we measure: choices, not neurons; markets, not molecules. The arsenal exists because no tractable theory captures all behavior at the scale where we work.

3.2 The Surprising Cases

It is for this exact reason that many of the "hard sciences" are also patchworks.

Chemistry provides a striking example. The Schrödinger equation governs molecular behavior completely, yet no chemist uses it to predict reactions. The level of quantum description doesn't match the level where chemistry operates: quarks versus compounds

Consider a very tangible problem, perhaps the most basic problem in chemistry: given a

reaction predict its yield (what compounds it will produce). Chemists reason through what they call "object" and "process" rules — localized theories about molecular structure and reactivity. Predicting whether and how a reaction occurs means reasoning about how electrons move and bonds break, invoking concepts like electronegativity and steric hindrance and applying rules for electrophilic aromatic substitution or retrosynthetic analysis. These are not derived from quantum mechanics — they are empirically grounded and communicated through diagrams and reasoning. The chemist deploys them selectively, guided by experience about which mechanism dominates in each context. Chemists may not only disagree on which mechanisms are first-order, but some chemists are proven to better at this activity than others. Once again, a patchwork of theories is applied by expert scientists.

Medicine exhibits the same structure. Genes encode proteins, proteins drive cellular processes, cellular dysfunction causes disease. Yet no physician diagnoses or treats patients by reasoning up from molecular mechanisms. The causal chain is too long, too context-dependent, and too incompletely understood. Clinical medicine instead operates through frameworks that apply at the level where doctors actually work. An electrocardiogram approximates the heart as a simple dipole, ignoring the complex three-dimensional propagation of electrical signals through heterogeneous tissue. Diagnostic pathways classify chest pain into cardiac versus non-cardiac based on symptom clusters, not cellular pathophysiology. Treatment protocols for sepsis follow empirically validated bundles of interventions, not first-principles immunology. These frameworks are deployed selectively, adapted to the clinical context at hand.

Rutherford supposedly declared: "all science is either physics or stamp collecting" — chemistry is applied physics, biology applied chemistry, and so on. But this reductionism is wrongheaded. Even if we could simulate every quantum interaction to predict labor markets, the predictions would be no better than appropriately chosen models at the working scale. The patchwork is not temporary; it is the structure of science at working scales.

3.3 The Contrast: Closed Sciences

It is useful to contrast patchwork sciences with what we might call "closed" sciences — domains governed by a small set of fixed rules where the central challenge is derivation within that system. Mathematics is the paradigmatic example. The rules of logic and set theory are fixed. The challenge is computational: how do you search over the vast space of possible proofs to find one that establishes the desired result?

The problems facing AI in mathematics are thus quite different from those in patchwork sciences. In mathematics, the challenge is more akin to designing AI systems to solve chess or Go. The rules are known, the objective is clear, and the difficulty is navigating a combinatorially large search space efficiently. Automated theorem proving exemplifies this approach: the algorithm

searches for or generates proofs within a well-defined formal system (Fawzi et al., 2022; Trinh et al., 2024; Romera-Paredes et al., 2024).

Other closed sciences include areas of theoretical physics where fundamental equations are established, or formal systems in computer science where correctness is defined with respect to a specification. In each case, discovery means finding new derivations within an existing formal structure.

This reveals a fundamental divide: closed sciences operate within fixed formal systems where discovery means search and derivation. Patchwork sciences work across incomplete theories at scales where foundational laws provide no leverage. How algorithms transform science depends on which side of this divide the science falls.

3.4 Exploit versus Expand Understanding

The key aspect of a patchwork science is that there is much left to discover. That realization makes clear two different way we can use algorithms.

Consider AlphaFold. It is clearly a major breakthrough. But it is a particular kind of breakthrough: it lets us better *apply* our understanding of organic chemistry and biology. Which is obviously of immense value. At the same time, it does not expand our own understanding. For all of its value, no one (to our knowledge) has extracted a new theory from AlphaFold — it predicts structures without revealing the underlying mechanisms or rules of protein folding.

That distinction - exploiting vs expanding theories - is central to our approach to the new factory floor. Much of the existing work on AI/ML for science focuses on exploitation. For the patchwork sciences, though, our existing theories are incomplete. Exploiting them has short-run gains but in the long run we must expand them. As a result, our rendition of the new factory floor is focused on new discoveries.

4 The New Factory Floor for Patchwork Sciences

How should the factory floor be reorganized for patchwork sciences? We established that these sciences work differently than closed sciences. The problem is not searching through well-defined proof spaces. The space itself is incomplete. Progress therefore requires three interconnected activities: innovating on understanding by discovering what current theories miss, applying understanding by learning which theories work where, and accumulating understanding by expanding the theoretical toolkit.

To design algorithms that participate meaningfully in each of these activities, we need more precision about what theories are and how they function in the patchwork sciences. We therefore begin with a simple framework to clarify what we mean by a "theory" and how theories relate to purely predictive algorithms. This framework is deliberately minimal, but it serves a crucial function: it lets us see where and how algorithms can contribute to the three core activities. We

then turn to concrete examples—consumer choice and mathematics education—that illustrate how these activities might operate in practice.

4.1 A Simple Framework for Patchwork Sciences

Return to the health insurance example from earlier. A researcher must choose: model asymmetric information or behavioral confusion? Expected utility or prospect theory? Each choice yields different predictions, none obviously wrong. Why do we maintain this arsenal of theories rather than converging on a single best model?

The answer lies in a fundamental tradeoff that characterizes patchwork sciences today: theories gain transferability across contexts by sacrificing flexibility in any one context. Even though prospect theory better captures the underlying psychology, expected utility's parsimony makes it more portable to contexts where much remains unmeasured. To see this clearly, we introduce a simple framework that distinguishes theories from purely predictive algorithms in the patchwork sciences.

Suppose some outcome y results from measured features x and unmeasured features z, generated according to $y = g^*(x,z)$. What is measured is not fixed: new instruments can convert unmeasured z into measured x. The measured and unmeasured features vary across contexts c — such as different populations, environments or time periods — each with its own distribution $P_c(\cdot)$ over (X,Z). We observe data from many contexts $\{(X_i,Z_i,Y_i)\}_{i=1}^{n_c}$; some contexts are data-rich and others are data-sparse. Our goal is to learn structure that allows us to model the outcome well across contexts.

In this setup, a theory is a restrictive set of "allowable functions" $\mathcal{F}^T$, where each $f \in \mathcal{F}^T$ maps measured features x to outcomes y (Mullainathan and Rambachan, 2025). For instance, expected utility theory corresponds to the set of choice functions consistent with its axioms. Prospect theory corresponds to a different (larger) set of allowable functions that relax independence and incorporate probability weighting and reference dependence. A theory's functions are interpretable and simple — they can be fit in any context with relatively little data. We select a theory's specific function in context c by minimizing prediction error:

$$\widehat{f}_c = \arg\min_{f \in \mathcal{F}^T} \frac{1}{n_c} \sum_{i=1}^{n_c} \ell(Y_i, f(X_i)).$$

By contrast, a black-box predictive algorithm is unconstrained. Given the measured features, it can approximate the best (mean-square error) predictor $\widetilde{g}_c(x) := \mathbb{E}_{P_c}[g^*(X,Z) \,|\, X = x]$ in any context more flexibly. In data-rich contexts, black-box methods often achieve better predictions than any theory precisely because they flexibly adapt to each context rather than ex-ante committing to restrictive structure.

Why Performance Varies Across Contexts: Even though the true model g^* is shared across all contexts, a theory's performance will vary across contexts because the distribution of unmeasured features z changes. What we call "noise" is actually misspecification that arises from these unmeasured features. In context c, any measure feature x exhibits variation in outcomes y due to unmeasured variation in z. Even if we knew g^* and predicted using $\widetilde{g}_c(x) := \mathbb{E}_{P_c}[g^*(X,Z) \,|\, X = x]$, errors would still arise from unmeasured context-specific variation. This means a theory's performance — given what we measure — inherently depends on the context. As a consequence, given our existing arsenal, the best-fitting theory may also vary across contexts.

This variation is permanent. It persists not because our theories are crude, but because the world is complex. We do not measure everything that might matter. Tractable theories therefore make tradeoffs: abstract from the unmeasured, impose structure on the observed. Different theories make different tradeoffs, and those tradeoffs succeed in different contexts as the landscape of unmeasured variation shifts.

Even more surprisingly, an incorrect theory (in the sense that $g^* \notin \mathcal{F}^T$) could outperform the correct model g^* given what we measure. Return to our insurance example: prospect theory may better capture the true psychology, but if many important features remain unmeasured, expected utility's simpler structure might yield better predictions in practice. The theory's constraint – its limited flexibility – could become an advantage in the face of limited data and measurements.

The Tradeoff Today: This framework reveals why the patchwork sciences maintain arsenals rather than converging on single theories. Theories connect low-data to high-data environments by positing reusable structure. They commit to using the same simple functional forms across contexts. This makes them learnable from sparse data but potentially inferior to black-box methods when data is abundant. Black-box algorithms, by contrast, better approximate $\widetilde{g}_c$ in each context precisely because they are not constrained by any theory's structure. They sacrifice transferability for flexibility.

In this sense, theories are like "foundation models" with particularly constrained representations — they trade flexibility for transferability across contexts. In the current factory floor, we accept this tradeoff. Researchers choose theories based on intuition about which mechanisms are "first-order" in each context — tacit judgments about which constraints will transfer. But these judgments happen informally, off-stage, guided by experience rather than systematic evidence.

Implications for the New Factory Floor: This framework clarifies what algorithms need to do to transform patchwork sciences. First, they must help us *innovate* by discovering where current theories fail — in any context, identifying where the theory's constraints are too restrictive (i.e., anomaly generation) or suggesting entirely new features z to measure. Second, they must help us *apply* theories systematically by learning which theories work where—formalizing the tacit judgments about "first-order" mechanisms. Third, they must help us *accumulate* understanding by building better constrained function classes $\mathcal{F}^T$ — or, more radically, by changing the form

in which we store understanding altogether. We now turn to how each of these activities might operate on the new factory floor.

4.2 Key Aspects of the New Factory Floor

What would it take for algorithms to participate meaningfully in innovation, application, and accumulation rather than merely assist at the margins? We examine each in turn.

4.2.1 Innovating: New Ideas and Discoveries

Algorithms could drive innovation along three dimensions. First, they could identify where current theories are too restrictive — in any context, finding examples where a theory's constrained function class $\mathcal{F}^T$ fails to capture patterns that black-box predictive algorithms detect. This is anomaly generation, as we discussed earlier, and algorithms can already contribute to this activity.

Second, leveraging their ability to predict well in data-rich contexts, algorithms could go further and propose better function classes—new ways to model the measured features x that maintain simplicity and interpretability while reducing misspecification. For example, recent work on symbolic regression (Udrescu and Tegmark, 2020; Cranmer, 2023) asks whether we can approximate black-box algorithms using more interpretable functional forms. This offers one path (of course there may be many others) from anomaly generation to theory refinement: when black-boxes outperform theories, algorithms might reveal which additional structure – still interpretable and transferable – closes the gap.

The third dimension is perhaps the most important in patchwork sciences. In domains where existing theories explain little, new discovery often lies not in improved fit (e.g., R^2) given what we currently measure, but in suggesting entirely new quantities to measure. Mendel's work in genetics did not begin by improving the R-squared of plant trait models. Its core insight was that observable variation in traits might be driven by unmeasured variables—genotypes—that could not yet be directly observed. The theory's power came from suggesting a new direction of measurement: tangible z that should be measured and incorporated into x. The same applies in psychology (e.g., working memory), epidemiology (e.g., asymptomatic transmission), and economics (e.g., expectations, inattention). Innovation is not merely a search over theories that explain what we already measure. Rather, it redefines what is observable.

Yet these creative innovations remain segregated in the current factory floor: flashes of human creativity suggest measurements, then we test them rigorously. This is a key limitation of the current factory floor that algorithms could address. Algorithms already assist in hypothesis and anomaly generation. But they could go further—operating over raw text, running simulations, proposing alternative variable definitions, or identifying latent inconsistencies across literatures. Algorithms could explore conceptual and measurement spaces, generating directions for new

models rather than just optimizing existing ones.

4.2.2 Applying: The Codification of Intuition

Researchers use "first-order" as shorthand: this mechanism matters, that one doesn't. These judgments — for example, whether to include hyperbolic discounting, loss aversion, or signaling — rest on tacit knowledge about what matters where.

We can formalize this activity. If we have a portfolio of theories T_i, $i = 1,...,K$, then what researchers do is implicitly construct a function $\phi(c)$ that maps each context c to the important or most important theories for that context. Our intuitions about "first-order" mechanisms are our informal implementations of $\phi(c)$. These judgments encode real knowledge: the discipline has learned, through application, that certain mechanisms matter more in certain settings. But the knowledge remains off-stage — inarticulable and context-dependent. Indeed, large language models trained on scientific corpora may already capture some of this tacit knowledge implicitly, having learned from patterns in how researchers deploy theories across papers.

Algorithms could help formalize $\phi(c)$ more systematically. As theories are applied across contexts, we generate data about which theories work where. That data could train a different kind of machine learning model — not one that predicts outcomes directly, but one that predicts which theories will perform well in which contexts. Notice the distinction: this is not using algorithms to replace theories, but to systematically learn the mapping from contexts to theories. It is meta-prediction: tracking what models work best and where, then building algorithms to predict those patterns. This prediction problem need not be solved literally or directly. The point is that this is the functionality we want in a patchwork: systems that can reason about which theories apply where.

Such models would codify our tacit judgments. Rather than relying on individual researchers' intuitions about what's first-order, we could systematically accumulate evidence about theory performance across contexts. When entering a new domain, instead of asking "what does my experience suggest?" we could query: which theories have performed well in structurally similar contexts? Such models would bring an essential piece of scientific practice from backstage intuition into formal machinery.

4.2.3 Accumulating: The Language of Theory

We have discussed how algorithms could help us innovate—discovering where theories fail—and apply—learning which theories work where. But what about accumulation: building better constrained function classes and storing them for future use? This raises a more fundamental question: what form should theories take?

The framework clarified that theories are constrained function classes that trade flexibility for

transferability. But it did not specify how those constraints should be represented. The default answer, at least in economics and much of science, is mathematical. We introduce notation, articulate ideas and derive implications through rigorous deduction. A theory's power comes from its mathematical structure

Yet consider a small survey. Ask economists their favorite theoretical idea. The answers would be consistent: opportunity cost, prisoner's dilemma, adverse selection, signaling. These share a striking feature: they're not technically challenging. Opportunity cost requires no mathematics; the prisoner's dilemma is a 2-by-2 game; the insight of the "market for lemons" is in its name. What makes them powerful is not derivational depth, but breadth: they help you notice aspects of the world you'd otherwise overlook. Once you understand adverse selection, you see it everywhere—in insurance markets, labor markets, online platforms. These are "sprawling theories:" simple to state, powerful in range.

"Burrowing theories" derive deep implications from precise formalisms — they dig vertically into a single context. "Sprawling theories" offer insights that bind to many contexts — they spread horizontally. In terms of our framework, sprawling theories achieve high transferability across diverse contexts c. Burrowing theories may fit one context exquisitely but transfer poorly.

The current factory floor privileges burrowing theories because we train researchers to formalize, axiomatize, and prove. This emphasis traces back to what (Wigner, 1960) called "the unreasonable effectiveness of mathematics" — the mysterious alignment between mathematical structure and physical reality. But this alignment may not extend to all domains. Wigner was explaining physics, where fundamental laws are parsimoniously expressed in mathematical language. The patchwork sciences are different.

Perhaps sprawling theories — theories expressed in language, in diagrams, in qualitative frameworks like we teach in introductory economics — have been undervalued precisely because they do not fit the mathematical template. This is not unique to economics. Even in the physical sciences, there are domains where theory is not expressed mathematically. Again consider how chemists reason about reactions. They draw molecular structures with arrows showing how electrons move, how bonds break and form. Each step invokes theory — ideas about electronegativity, steric hindrance, resonance stabilization. But these are not mathematical derivations. They are communicated through diagrams that apply intuition and encode existing understanding.

This raises a radical possibility in the age of algorithms: theories need not be mathematical. Mathematics became theory's language because it is interpretable to humans — a shared, precise language we can manipulate and verify. But this constraint comes from human cognition, not from the structure of understanding itself. Algorithms do not face this constraint. Neural networks can store representations far more complex than any human can manipulate or write down.

If the goal is to accumulate understanding that enables prediction, application, and innovation

across contexts, why must theories remain equations? They might instead become learned representations: computational objects that sprawl across contexts, adapt to new data, and support flexible application. A learned representation would not look like expected utility theory's axioms. But if it organized understanding in ways that transferred across contexts — if it achieved high completeness in many settings, not just one — it would serve the essential function of a theory.

This is not replacing understanding with black-boxes. It is changing the form in which we store understanding. The constraint—the commitment to structure that transfers—remains. What changes is the requirement that this structure be expressible in mathematical notation that humans can parse. The new factory floor could accumulate understanding in forms native to algorithms: representations that sprawl across contexts, encode transferable structure, and enable the kind of systematic innovation and application that patchwork sciences require.

4.3 Some Imagined Examples of the New Factory Floor

What would these three activities — innovation, application, accumulation — look like when integrated rather than segregated? When algorithms participate systematically rather than assist at the margins? To make this concrete, we reimagine two familiar domains. These descriptions are deliberately speculative, but grounded in capabilities that already exist or require only modest engineering advances. They illustrate not what algorithms can do today, but what becomes possible when we reorganize the factory floor around them.

4.3.1 Consumer Choice in the Age of Algorithms

Consider a canonical question in economics: how do consumers make choices among competing products on a large online retailer? The traditional approach begins with a tightly parametrized structural model: specify a utility function, estimate parameters, test predictions. The researcher observes purchases, clicks, and prices from browsing sessions, then asks: How price-sensitive are consumers? What features drive choice? How would demand shift if products changed?

Now imagine a different factory floor. Each consumer leaves rich behavioral traces: purchases, searches, support messages, dwell times on pages. Products are not described by hand-coded attributes but by everything they project: images, reviews, price histories, compatibility metadata, influencer commentary. The researcher's goal remains prediction: as new consumers and products arrive, what will be chosen?

The core artifact is no longer a "best-in-class" equation estimated once. Instead, it is an algorithm that learns a joint representation of consumers and products, continuously refined by new observations. That algorithm importantly blurs the line between *applying* theories and *testing* theories. Each act of application becomes an act of learning. As it predicts across product categories, customer segments, and time periods, it internalizes which features drive attention in

which contexts. Perhaps younger consumers weight aesthetics heavily for certain products but not others. Perhaps consideration sets narrow when prices cross specific thresholds. The model does not start knowing these patterns. It discovers them by applying across diverse contexts and observing where its predictions succeed or fail.

How does a researcher interact with this system? Not by estimating parameters, but by querying the representation. Probe it with counterfactuals: how would this consumer respond to a price change? Which product features would most increase demand in this segment? The model's responses reveal what it has learned about behavior — not through equations, but through the structure embedded in its representations. Interpretability tools might extract human-readable patterns: customer types defined by behavioral clusters, product features that matter differentially across contexts, phase transitions in how competition operates.

This is not a pipe dream. This is a "foundation model" for consumer behavior (e.g., Bommasani et al., 2022). It is a scientific object encoding our best current understanding in learned representations. It accumulates knowledge through application, adapting as it deploys across consumers and products. The factory floor has been reorganized: understanding is no longer prior to application but emerges from it.

4.3.2 Education in the Age of Algorithms

Mathematics pedagogy has traditionally been a craft, not a science. Teachers accumulate intuitions about student misconceptions: why students overgeneralize fraction rules, which visual representations clarify proportional reasoning. These insights exist as tacit knowledge — a patchwork of local insights deployed selectively by skilled teachers.

This could change. Online platforms now capture detailed logs of how students solve problems: each step taken, each error made, how understanding evolves over time. Thousands of students working through varied mathematical content generate precisely the data needed to systematize the patchwork. An algorithm trained on this data learns patterns in how students understand mathematics — not a single theory, but regularities across concepts and populations. Certain confusions may recur systematically (conflating intensive and extensive quantities, overgeneralizing procedures). Students may cluster into distinct modes of mathematical reasoning. The model discovers what matters when through deployment: predicting student responses across fractions, algebra, and geometry.

With this data, an algorithm could systematize the patchwork of mathematical pedagogy. The algorithm learns a representation of how students understand mathematics — not a single theory, but patterns that hold across concepts and populations. Perhaps certain conceptual confusions recur systematically across topics (conflating intensive and extensive quantities, overgeneralizing algorithmic procedures). Perhaps students cluster into distinct modes of mathematical reasoning. The model discovers these patterns through deployment: as it predicts student responses across

contexts—fractions, algebra, geometry—it internalizes what matters when.

Such an algorithm would transform both how we use theories and the kind of knowledge we accumulate over time. Teachers query the model to personalize instruction, identifying which explanation clarifies a specific confusion. Researchers probe its representations to extract general principles and discover systematic misconceptions invisible to any individual teacher.

In this example, algorithms allow us to introduce the factory floor of science to mathematics pedagogy. Understanding accumulates through continuous deployment rather than remaining siloed in individual practitioners. The model applies current knowledge to improve learning while revealing its gaps — misconceptions we didn't know existed, patterns invisible without aggregation across contexts.

5 What is Needed to Build the New Factory Floor

Re-imagining the factory floor of science is not only a conceptual task — it is a technical one. Patchwork sciences require three interconnected activities: accumulating understanding by expanding the theoretical toolkit, applying understanding by learning which theories work where, and innovating on understanding by discovering what current theories miss.

For algorithms to participate meaningfully in each activity — not merely assist at the margins — we highlight four principles that must underlie building AI systems for science: world modeling, rigorous design, modular design, and cross-mode communication. These are not features to be tacked on, but foundational principles that steer algorithms toward truly accelerating scientific discovery in patchwork sciences.

5.1 World Models, not Pattern Matchers

To contribute meaningfully to science, algorithms must move beyond next-token prediction to world modeling: building latent spaces that encode mechanisms, causal structure, and shared abstractions across tasks. The distinction is fundamental: pattern matchers exploit correlations in training data; world modelers internalize the generative structure producing those patterns, enabling the principled transfer required in patchwork sciences (e.g., Richens and Everitt, 2024; Richens et al., 2025).

Recent evidence suggests that even our most sophisticated foundation models fall short. Vafa et al. (2024) test whether LLMs internalize coherent world models by training them on environments with known state spaces—navigation, logic puzzles, board games. While these models excel at next-token prediction, their internal representations fail to distinguish underlying states, leading to brittle generalization. Vafa et al. (2025) extend this finding: even when trained on structured data like planetary trajectories or board games, models do not encode the true generative mechanisms. Fine-tuning a model on orbital mechanics to predict forces reveals it has not learned Newton's laws. Instead, models rely on task-specific heuristics that collapse outside the training domain.

The conclusion is stark: today's foundation models are pattern matchers, not world modelers.

There exists a gap between algorithms today and what is needed on the factory floor of science. We do not yet have algorithms that go beyond pattern matching to build coherent world models — structures that capture mechanisms which can be reused, refined, and probed. If science is to advance via automated model-building, we need systems that learn the structure of worlds, not just sequences.

5.2 Rigorous Evaluation

Many popular assessments of AI's contribution to science rely on first-hand experiences. A researcher might report that a language model assisted in proving a theorem or suggested a productive research direction. But how much did the algorithm contribute versus the human who curated the inputs, reformulated suggestions, and verified outputs? Without rigorous evaluation, we cannot distinguish genuine algorithmic insight from sophisticated pattern matching—or worse, lucky guesses among many failed attempts.

This is not a new problem. Early machine learning research faced similar challenges until the field adopted common task frameworks: standardized datasets and metrics that enabled systematic comparison across methods. ImageNet transformed computer vision not by solving vision completely, but by establishing shared infrastructure — a fixed dataset with clear evaluation metrics — for measuring progress (Donoho, 2024). Science needs equivalent frameworks.

As examples, evaluating algorithmic contributions to science would require providing specific answers to how we would measure: What fraction of the intellectual work did the algorithm perform versus the human? Is the algorithm contributing genuine insight, or accelerating search through a space already defined by humans? How do we distinguish novel discovery from retrieval or recombination of existing knowledge?

Without systematic evaluation frameworks, we cannot improve these systems. Further accelerating progress requires moving from anecdotal reports to rigorous evaluation.

5.3 Modular Architecture

Science works through division of labor: experimentalists design measurements, theorists build formal models, statisticians handle inference. This modularity enables targeted improvement: we can identify which component is the bottleneck and focus effort there.

Current large language models resist this modularity. Consider a thought experiment: suppose we prompted an LLM to generate anomalies for expected utility theory or produce novel hypotheses about facial characteristics in judicial decisions. If the LLM failed to produce useful outputs, what would we do next? Change the prompt? Add examples? Fine-tune the model? Each intervention affects the entire system unpredictably. Perhaps we are better off waiting until a new base model

is released.

By contrast, our procedures are decomposable: (i) collect data on a behavior; (ii) train a predictor; (iii) use that predictor to generate contrasts via optimization. If an initial design fails, we can diagnose which component is responsible. Perhaps we collected the wrong data, perhaps our predictor is insufficiently accurate, perhaps our optimization procedure is inadequate. This decomposition clarifies how to improve.

Importantly, our description of the factory floor suggests a natural decomposition. The broader goal — an AI system that conducts science — can be separated into sub-components that accumulate understanding, systems that apply, and systems that innovate on it. Each activity can be developed independently, provided they connect through well-defined interfaces.

This is the "hourglass architecture" that has proven successful throughout computing: a narrow waist of shared abstractions enables innovation on both sides independently. Just as the Internet protocol allows diverse applications above and diverse physical networks below to evolve separately, clear interfaces between scientific activities would allow different research communities to specialize — for example, some building better world models and others designing evaluation frameworks. The monolithic approach collapses this structure, forcing every advance to redesign the entire system.

5.4 Human-Algorithm Communication

In a world of incomplete understanding, multiple perspectives are valuable. Rather than privileging one mode – algorithms, human intuition, or formal theories – we should enable communication and composition across them. Even if algorithms become better than humans at both prediction and interpretation, there may still be complementary gains from specialization. Humans might focus on identifying which measurements matter, while algorithms excel at organizing patterns within those measurements. Formal theories might excel at specifying constraints and invariances, while neural networks capture the messy deviations from those idealized structures.

Some promising approaches already exist in this direction. Neurosymbolic programming composes neural networks with formal symbolic reasoning, enabling systems that combine the pattern recognition capabilities of deep learning with the logical rigor of formal methods (Lake, Salakhutdinov and Tenenbaum, 2015; Wong et al., 2023, 2025). Interpretability techniques such as sparse autoencoders (Movva et al., 2025; Peng et al., 2025) extract human-interpretable structure from learned representations, enabling people to inspect and understand what algorithms have discovered.

Yet these approaches remain limited. Most interpretability methods are post-hoc: they explain what a model learned after training, but don't enable real-time collaboration during learning. Most neurosymbolic systems require extensive manual engineering to specify how components interact. Communication remains bolted on rather than native.

How might native communication reshape the factory floor? As an example from the near

future, an algorithm might identify an anomaly in consumer choice data, a human might articulate why this pattern matters for welfare analysis, and a formal model might specify the mechanism — with each contribution legible to the others. When applying theories across contexts, humans might specify which mechanisms seem first-order, algorithms might test these intuitions systematically, and formal theories might clarify when such intuitions should transfer. This iterative exchange across modes of understanding enables effective accumulation, application, and innovation require.

5.5 Articulating the Challenges

These four functionalities don't build themselves. Computer science advances by solving hard problems: benchmarks that seem impossible, datasets that expose gaps, challenge tasks that force new approaches. Progress requires problems worth solving.

Here are some problems worth solving in the patchwork sciences. In health insurance markets, create an algorithm — one that embodies both existing behavioral theories (loss aversion, probability weighting, inertia) and black-box machine learning — that predicts behavior accurately enough to improve choices and provide advice but also generates behavioral anomalies for researchers. Build a system that discovers genuinely novel chemical reaction mechanisms and reveals to chemists where existing understanding of object and process rules break down.

These are not vague aspirations. They are concrete technical challenges that, if solved, would reshape how patchwork sciences operate. The factory floor will not reorganize itself. It requires deliberate engineering — and engineers need clear targets.

6 Conclusion

Bertrand Russell once seemingly confessed to the weakness of philosophy (Russell, 1912):

> *If you ask a mathematician, a mineralogist, a historian, or any other person of learning, what definite body of truths has been ascertained by his science, his answer will last as long as you are willing to listen. But if you put the same question to a philosopher, he will, if he is candid, have to confess that his study has not achieved positive results such as have been achieved by other sciences.*

This is, of course, a setup. He goes on to say:

> *It is true that this is partly accounted for by the fact that, as soon as definite knowledge concerning any subject becomes possible, this subject ceases to be called philosophy, and becomes a separate science.*

Astronomy, and physics more broadly, was once called "natural philosophy." The same is true for psychology, economics, biology and so on. Russell's argument is that: when a field is wrestling

with vague, hard ideas it is philosophy; when progress allows us to make those ideas tangible and precise, it gets "spun out" into its own discipline.

Much of this chapter has been philosophy in disguise. We flirted with concepts such as induction and abduction. We probed metaphysical issues such as what does "truth" of a theory mean in a patchwork world. We even brushed against epistemics by asking with the right "form factor" of theories could be, if not math. Yet this quiet dialogue with philosophy was quiet exactly because our proposals were tangible and specific: rather than speculation, we were able to talk about specific tools. To apply Bertrand Russell's maxim: in the age of algorithms, philosophy of science will transition from philosophy to science.

References

Allais, Maurice. 1953. "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine." *Econometrica*, 21(4): 503–546.

Athey, Susan. 2018. "The Impact of Machine Learning on Economics." *The Economics of Artificial Intelligence: An Agenda*, 507–547. University of Chicago Press.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. "On the Opportunities and Risks of Foundation Models."

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2019. "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics." In *The Economics of Artificial Intelligence: An Agenda*. , ed. Ajay Agrawal, Joshua Gans and Avi Goldfarb, 23–57. University of Chicago Press.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2021. "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies." *American Economic Journal: Macroeconomics*, 13(1): 333–72.

Cranmer, Miles. 2023. "Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl."

David, Paul A. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *The American Economic Review*, 80(2): 355–361.

David, Paul A, and Gavin Wright. 2006. "General purpose technologies and surges in productivity: Historical reflections on the future of the ICT revolution." In *The Economic Future in Historical Perspective*. , ed. Paul A David and Mark Thomas, 135–166. Oxford University Press/British Academy.

Davies, Alex, Petar Velickovic, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. 2021. "Advancing mathematics by guiding human intuition with AI." *Nature*, 600(7887): 70–74.

Devine, Warren D. 1983. "From shafts to wires: Historical perspective on electrification." *Journal of Economic History*, 43(2): 347–372.

Donoho, David. 2024. "Data Science at the Singularity." *Harvard Data Science Review*, 6(1). https://hdsr.mitpress.mit.edu/pub/g9mau4m0.

Fawzi, Alhussein, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, Demis Hassabis, and Pushmeet Kohli. 2022. "Discovering faster matrix multiplication algorithms with reinforcement learning." *Nature*, 610(7930): 47–53.

Feldman, Moran, and Amin Karbasi. 2025. "Gödel Test: Can Large Language Models Solve Easy Conjectures?"

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy*, 130(4): 956–990.

Jumper, John, 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. 2021. "Highly accurate protein structure prediction with AlphaFold." *Nature*, 596(7873): 583–589.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263–291.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Kochkov, Dmitrii, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. 2021. "Machine learning–accelerated computational fluid dynamics." *Proceedings of the National Academy of Sciences*, 118(21): e2101784118.

Krenn, Mario, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, Zhenpeng Yao, and Alán Aspuru-Guzik. 2022. "On scientific understanding with artificial intelligence." *Nature Reviews Physics*, 4(12): 761–769.

Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. "Human-level concept learning through probabilistic program induction." *Science*, 350(6266): 1332–1338.

Liang, Annie. 2025. "Using Machine Learning to Generate, Clarify, and Improve Economic Models."

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation*." *The Quarterly Journal of Economics*, 139(2): 751–827.

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. "Sparse Autoencoders for Hypothesis Generation."

Mullainathan, Sendhil. 2025. "Economics in the Age of Algorithms." *AEA Papers and Proceedings*, 115: 1–23.

Mullainathan, Sendhil, and Ashesh Rambachan. 2025. "From Predictive Algorithms to Automatic Generation of Anomalies."

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*, 31(2): 87–106.

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics*, 137(2): 679–727.

Peng, Kenny, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. 2025. "Use Sparse Autoencoders to Discover Unknown Concepts, Not to Act on Known Concepts."

Raghu, Maithra, and Eric Schmidt. 2020. "A Survey of Deep Learning for Scientific Discovery."

Rambachan, Ashesh. 2024. "Identifying Prediction Mistakes in Observational Data." *The Quarterly Journal of Economics*, 139(3): 1665–1711.

Richens, Jonathan, and Tom Everitt. 2024. "Robust agents learn causal world models."

Richens, Jonathan, David Abel, Alexis Bellot, and Tom Everitt. 2025. "General agents contain world models."

Romera-Paredes, Bernardino, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. "Mathematical discoveries from program search with large language models." *Nature*, 625(7995): 468–475.

Russell, Bertrand. 1912. *The Problems of Philosophy.* London:Williams and Norgate.

Trinh, Trieu H., Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. "Solving olympiad geometry without human demonstrations." *Nature*, 625(7995): 476–482.

Udrescu, Silviu-Marian, and Max Tegmark. 2020. "AI Feynman: A physics-inspired method for symbolic regression." *Science Advances*, 6(16): eaay2631.

Vafa, Keyon, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. 2024. "Evaluating the World Model Implicit in a Generative Model."

Vafa, Keyon, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. 2025. "What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models."

Wang, Hanchen, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. 2023. "Scientific discovery in the age of artificial intelligence." *Nature*, 620(7972): 47–60.

Wigner, Eugene P. 1960. "The unreasonable effectiveness of mathematics in the natural sciences." *Communications on Pure and Applied Mathematics*, 13(1): 1–14.

Wong, Lionel, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. 2023. "From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought."

Wong, Lionel, Katherine M. Collins, Lance Ying, Cedegao E. Zhang, Adrian Weller, Tobias Gerstenberg, Timothy O'Donnell, Alexander K. Lew, Jacob D. Andreas, Joshua B. Tenenbaum, and Tyler Brooke-Wilson. 2025. "Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models."