Abduction, Judgment, and Algorithms: A New Factory Floor for Science? Discussion of Mullainathan and Rambachan[1]

Ajay Agrawal
University of Toronto and NBER

John McHale
University of Galway

Alexander Oettl
Georgia Tech and NBER

September 24, 2025

Most discussions of artificial intelligence (AI) place it in the lineage of the telescope, microscope, and computer: a powerful new tool poised to significantly influence the rate and direction of inventive activity. Like its predecessors, AI can detect patterns and phenomena beyond human sensory and cognitive limits, and it can automate calculation, data processing, and elements of experimental design. However, Mullainathan and Rambachan have written a paper that offers a more provocative claim. It frames AI not merely as the next great scientific instrument, but as potentially more transformative than any prior tool—comparable instead to the creation of the scientific method itself. In their vision, rather than AI presenting a powerful new tool to further enhance productivity on the factory floor of science, they advance a thesis that AI could redesign the factory floor itself, reshaping the foundational organization of scientific work. This view prompts a profound reconsideration of the division of labor between humans and machines, not only in the discovery process but in the very act of understanding itself, potentially extending into the creative, judgment-driven work of abductive inference that underlies hypothesis generation. This shift would place new pressures on what we call "research judgment," the tacit and context-dependent capacity to generate fruitful questions and frame problems, raising the question of whether, and how, such judgment can be replicated, augmented, or redefined in an era of algorithmic discovery.

What distinguishes Mullainathan and Rambachan's contributions is how they combine pioneering work developing useful machine learning algorithms for doing science with reflections on the philosophical implications of those algorithms for the scientific method itself (see, e.g., Fudenberg et al. 2022; Ludwig and Mullainathan 2024; Mullainathan and Rambachan 2024). Their chapter in

[1] This discussion essay is written for the NBER Economics of Transformative AI Workshop. Contact: ajay.agrawal@rotman.utoronto.ca, john.mchale@nuigalway.ie, and alex.oettl@scheller.gatech.edu

this volume gives them the opportunity to focus on the more philosophical (and speculative) questions, and in particular to set out a vision for how the "factory floor of science" could evolve. The result is both an insightful analysis of how the current factory floor operates and a manifesto of sorts of what science could look like in the age of algorithms.

The chapter, then, considers two main questions:

- What does the arrival of a new class of intelligent agents reveal about the *existing* organization of the factory floor of science?
- How might (should?) the factory floor be *re-organized* given this new class of intelligent agents?

We consider each of these questions in turn, drawing in part on our own recent work on AI in science (Agrawal et al., 2023, 2024, 2025).

What does the arrival of a new class of intelligent agents reveal about the existing organization of the factory floor of science?

Using the illustrative example of how consumers make choices, they begin by setting out what they see as the standard view of the factory floor of science. They capture this as a three-stage process (or "assembly line") involving: theory, test, and use. They then draw on their own work building algorithms for science to reflect on what this new class of intelligent agents tells us about the *real* workings of the factory floor.

There is the *expected* – that algorithms have great potential to improve science; and, most importantly for the present paper, the *unexpected* – the things that the rise of this new class of intelligent agents reveal about what is going on "off-stage" from the standard view of the factory floor. Among the revelations they highlight: the importance of pattern and anomaly detection to generating new research questions and hypotheses (Ludwig and Mullainathan, 2024; Mullainathan and Rambachan, 2024); the importance of tacit knowledge in binding the symbols of our theories to actual meanings in the world; and the often poor performance of our theories and derived empirical models in predicting variation in our dependent variables compared to what the algorithms reveal is predictable from the data (Fudenberg et al., 2022).

In our own recent work (Agrawal et al., 2025), we highlight something complementary that we think is underemphasized in standard accounts of the factory floor – the importance of abductive

inference. Abduction has many of the features that Sendhil and Ashesh highlight: an emphasis on hypothesis generation; a stress on the importance of tacit knowledge of context in the generation process; and a highlighting of the joint importance of predictability and understanding. However, identifying the "dark matter" of the factory floor as abduction has the advantage of connecting what is missed in standard accounts of the scientific process to a concept that has a long history in both the philosophy of science and cognitive psychology.

To see the central role of abductive inference in science, it is useful to recall Richard Feynman's celebrated description of the factory floor in physics:

> In general we look for a new [physical] law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guess is right. Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science (Feynman, 1992, p. 156).

The initial guess – i.e., generation of a hypothesis – is the outcome of abductive inference. The idea of abduction was developed by the 19th and early 20th century American Philosopher Charles Sanders Peirce. Peirce pithily described the process as:

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence there is reason to suppose that A is true (Peirce, 1998, p. 216).

In the familiar context of economic modeling, deduction can usefully be viewed as the process of going from a set of assumptions or axioms to valid conclusions; induction as the generalization from specific observations; we can then think of abduction as the process of choosing the set of assumptions that underpin an economic model in the first place – a process of search through a vast combinatorial space of possible sets of assumptions and implied models. Working with deduction, abduction is then the origin of our guesses or hypotheses, which then can be tested through induction.[2] Along with the judgment used in question generation, the use of abductive inference in hypothesis generation might be collectively called "research judgment."

[2] Note the commonalities between Peirce's focus on "surprise" and their focus on anomaly detection in generating new questions and hypotheses.

How might (should?) the factory floor be re-organized given this new class of intelligent agents?

Mullainathan and Rambachan's offer an enticing vision of what the factory floor could look like in the age of AI. The vision is a dramatically improved ability to explain the world, combining an ability to predict outcomes and an ability to (at least partially) understand how those outcomes have come about. The "world model" underpinning science moves substantially from the head of the scientist to the algorithms. It is important that even with this transition they see critical ongoing roles for human scientists, but these roles will change:

> [S]cientists see what is not in the data. They design new measurement instruments, imagine hidden constructs, and ask questions that representations have not yet captured. Interpretability helps widen the space of use – but depth may no longer require it. People define the measurement boundaries; models help organize what fits inside (Mullainathan and Rambachan, 2025, p. 12).

Although Sendhil and Ashesh don't frame it this way in the paper, one way to think about their vision for the new factory floor is the blending of the approaches to the $\hat{y}$ and $\hat{\beta}$ problems distinguished in Mullainathan and Speiss (2017). $\hat{y}$ problems are about prediction – effectively predicting the variation in some dependent variable, y, that is of interest. $\hat{\beta}$ problems are about credibly identifying the causal impact of a change in some independent variable of interest on the dependent variable. Traditionally, $\hat{\beta}$ is estimated using some parametric model, where the causal effect of interest is derived from some mathematically specified model. The underlying model will often be critical to understanding the nature of the causal processes that are being hypothesized. Fudenberg et al. (2022) argue that many economic theories are highly *incomplete* in the sense of explaining a relatively small fraction of the predicable variation in the dependent variable of interest, where the predictable variation is determined using a machine-learning algorithm – i.e., there is poor $\hat{y}$ performance. The new factory floor would develop more complete and integrated theories (not necessarily expressed in mathematics), but the underlying AI models would, in sharp contrast with the black boxes of today's machine-learning algorithms, be interpretable. It is hoped that the blend would allow for a much better combination of predictability and understanding.

How *feasible* is this new vision of the factory floor? It would be foolhardy to dismiss the vision given the dizzying rate of progress in AI. The potential for combining foundation models as a base on

top of which is added some reasoning/search process suggests a promising path towards simulating human abductive-inference capabilities. However, other recent work by teams that Sendhil and Ashesh are part of suggests that significant additional breakthroughs will be necessary to achieve this vision (Vafa et al., 2024, 2025). While current foundation models exhibit impressive predictive capabilities, analysis of the models indicates that they often fail to capture underlying causal structures.

Limitations in the creative capabilities of LLMs may explain their poor performance on François Chollet's ARC challenge, which requires solving visual grid-based puzzles using just a few examples as an archetypal test of abductive inference—forming and testing hypotheses about transformations from basic operations. Traditional LLMs fare badly compared to humans, though reasoning models like OpenAI's o3 show substantial improvement, with ongoing research evaluating their efficacy on tougher versions. Questions remain about generalizing these techniques to other abductive problems, suggesting that simulating such reasoning could bottleneck AI-driven world modeling for the new factory floor of science.

Granting feasibility, is it obvious that the new factory floor is desirable? In public lectures Sendhil has been an articulate advocate for AI that serves as a tool for humans – "bicycles for the mind." But would a factory floor for science that moves much of the development of world models from human minds to AI models really fit this image? How much would it be a "bicycle for the mind" as opposed to something approaching a "self-driving bicycle"? The huge upsides of an AI for science that allows for possibly radically more complete models of structure in the world are clear enough – new medicines for hitherto intractable diseases, new materials to generate and store energy, new understandings of how consumers make choices. But there are potential downsides. Although they see a significant ongoing role for human scientists, the envisioned factory floor, where much of the understanding is contained within the AI models, does seem to imply a significant amount of *outsourcing* of what is currently done by human scientists to AI. It is debatable whether such AIs would mainly serve as tools used by scientists rather than replacements for those scientists. The scientific labor market may not be immune from the kind of automation forces that are affecting labor markets more generally.

The overall model of science could move from a combination of formal mathematical models and tacit knowledge held in scientists' minds to a regime in which the entire model resides within the AI. In such a world, scientists would attempt to interpret and understand the model learned by the AI and succeed only partially, with portions of the model possibly remaining inaccessible to human

understanding. It is not obvious how this asymmetry will affect the rate and direction of scientific progress—or how it will reshape the factory floor of science—when the complete model lives in the AI and the scientist can only comprehend fragments.

References

Agrawal, A., McHale, J., and Oettl, A. (2023). Superhuman science: How artificial intelligence may impact innovation. *Journal of Evolutionary Economics*, 33:1473–1517.

Agrawal, A., McHale, J., and Oettl, A. (2024). Artificial intelligence and scientific discovery: a model of prioritized search. *Research Policy*, 53(5):1–16.

Agrawal, A., McHale, J., and Oettl, A. (2025). The economics of artificial intelligence and science. Draft chapter for forthcoming NBER volume, *Economics of Science*.

Feynman, R. (1992). *The Character of Physical Law*. Penguin Books, London. Originally published 1965.

Fudenberg, D., Kleinberg, J., Liang, A., and Mullainathan, S. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, 130(4):956–990.

Kumar, A., Clune, J., Lehman, J., and Stanley, K. O. (2025). Questioning representational optimism in deep learning: The fractured entangled representation hypothesis.

Ludwig, J. and Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

Mullainathan, S. and Rambachan, A. (2024). From predictive algorithms to automatic generation of anomalies. Working Paper 32422, NBER.

Mullainathan, S. and Rambachan, A. (2025). Science in the age of algorithms.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Peirce, C. S. (1998). *The Essential Peirce: Selected Philosophical Writings Volume 2 (1983-1913)*. Indiana University Press, Bloomington and Indianapolis. Originally from 1903.

Vafa, K., Chen, J., Rambachan, A., Kleinberg, J., and Mullainathan, S. (2024). Evaluating the world model implicit in a generative model. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Vafa, K., Chang, P., Rambachan, A., and Mullainathan, S. (2025). What has a foundation model found? using inductive bias to probe for world models. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada.