To cite this article:

Yu, K., & Robinson, M. M. (in press). Prediction, risk, and illusion. *Computational Brain & Behavior*.

Prediction, Risk, and Illusion

Kenny Yu[1]*, Maria M. Robinson[2]

[1]Quantitative Psychology and Individual Differences, KU Leuven, Belgium
[2]Department of Psychology, University of Warwick, United Kingdom

Abstract

Shiffrin et al. (2026) identify several forms of illusory understanding in science. We focus on one they emphasize: the illusion that predictive success implies causal insight. Their diagnosis, we argue, lacks a criterion for when incomplete understanding becomes illusory, conflating predictions that warrant confidence with those that do not. We propose that what separates warranted confidence from illusion is not causal completeness but the risk structure of the prediction: whether the model could have failed had the hypothesis been wrong. The question shifts from "how complete is the causal story?" to "how severe was the test?"

Shiffrin et al. (2026) argue that predictive success gives scientists an illusion of causal understanding. Models that predict well are easily mistaken for models that explain well (cf. Meehl, 1978; Shmueli, 2010). Their demonstration that even linear regression harbors unsuspected depths is a valuable corrective. We think, however, that the diagnosis is weakened by an underspecified benchmark. The authors acknowledge that complete causal understanding is "rarely or possibly never" attained (p. 1), yet their analysis leaves open when incomplete understanding becomes illusory. Absent such a criterion, the unreachable ideal risks serving as a de facto diagnostic threshold, rendering all scientific understanding perpetually illusory by definition. We propose the *risk structure* of the predictive test as such a criterion.

Let e denote a predictive success and H a causal hypothesis. The evidential value of e for H is captured by the likelihood ratio $\mathcal{L} = P(e \mid H) \,/\, P(e \mid \neg H)$. In low-risk prediction, a flexible model accommodates many outcomes, making $P(e \mid \neg H)$ approach $P(e \mid H)$ and driving $\mathcal{L}$ toward 1; the prediction succeeds but is uninformative. In high-risk prediction, commitment to a specific outcome under novel conditions ensures $P(e \mid \neg H) \ll P(e \mid H)$, yielding a large $\mathcal{L}$. Risk is the complement of flexibility: a prediction is risky to the extent it would have differed had the hypothesis been false (Mayo, 2018), not merely that failure was possible, but that it was plausible (Vanpaemel, 2020; Van Dongen et al., 2023). This connects to the paradoxes the authors discuss. In Simpson's and Lord's paradoxes, the core error is treating observational regularities as licensing interventional conclusions (Pearl, 2009); the observed association is as likely under rival causal structures as under the preferred one, leaving $\mathcal{L}$ close to 1 even as the conclusions drawn assume it is large.

When prediction creates illusion. The authors observe that "judgment of a good prediction tends to stop a search for a better model" (p. 23). This is most acute for within-sample fit, where a flexible

*Corresponding author: kenny.yu@kuleuven.be

model is adjusted until it reproduces observed data. Such fit provides weak evidence for the model's causal claims, because the model might have fit equally well under different data-generating processes (Roberts and Pashler, 2000): formally, $\mathcal{L} \approx 1$. Yet low-risk prediction is still psychologically potent; the close match between expectation and observation mimics genuine insight. This is the authors' fifth illusion type: the belief "that one understands the cause of phenomena because there exists a model or procedure that predicts well" (p. 4). The illusion is not that prediction occurred, but that prediction was informative when the model was never at serious risk of failure.

When prediction corrects illusion. Not all predictions have this structure. When Einstein's general relativity, developed to resolve Mercury's anomalous perihelion advance, successfully predicted the bending of starlight observed during the 1919 solar eclipse, $P(e \mid \neg H)$ was low: a gravitational theory could straightforwardly have failed when extended to light propagation. Success under these conditions provides provisional evidence that the posited mechanism is stable across the tested contexts (cf. Bühlmann, 2020), not a complete causal account but understanding adequate within an explicitly stated scope.

The authors' airplane wing example (pp. 5–6) illustrates the distinction between low- and high-risk prediction. Wing designers "are unwilling to rely solely on the equations; they test designs in wind tunnels and tank tests". Presented as evidence of incomplete understanding, this is equally evidence of high-risk prediction in action. A design can fail spectacularly, narrowing the space of viable models. Their Newton vignette (Stigler, 2006) makes the converse point. Newton computed the correct probability for a dice problem but offered a verbal rationale that would fail under slight variations. The prediction was correct; the illusion lay in the rationale, which no further test was designed to challenge.

Opacity and the structure of risk. The authors' strongest case concerns opaque models such as Large Language Models. An LLM trained on vast data can generate fluent output for nearly any input, so $\mathcal{L} \approx 1$ for any given output. When such flexibility is combined with opacity, the illusion is at its most powerful. The response is to design higher-risk tests: targeted perturbations that isolate specific hypotheses and could straightforwardly fail (Woodward, 2004). For LLMs, we are far from this: behavioral tests can probe what a model does, but opacity prevents formulating mechanistic hypotheses, leaving causal claims untested even when predictions succeed. Interpretability research that exposes internal structure is one step toward making severe tests feasible; the framework identifies this as a benchmark to pursue rather than a ready solution. Where such tests cannot yet be designed, the difficulty itself is diagnostic: it marks where the illusions the authors describe will be hardest to dislodge.

Implications. A natural objection is that even high-risk prediction yields incomplete understanding, which is precisely the illusion the authors diagnose. But this conflates two distinct epistemic states. Incomplete understanding is universal, and saying so is near-trivial. The productive question is whether understanding is adequate to its intended scope. Illusory understanding arises when confidence exceeds what the evidence warrants for that scope: when scientists act as though $\mathcal{L} \gg 1$ on the basis of tests where $\mathcal{L} \approx 1$, whether because the theory accommodates too much, the test discriminates too little, or the claim outstrips the tested scope. High-risk prediction reduces this gap between confidence and warrant, though understanding remains incomplete. A framework unable to distinguish within-sample fit from severe out-of-sample testing lacks diagnostic power. We focus on prediction-based illusions; other illusion types may need dif-

ferent diagnostics, but severity offers a common thread: illusion is calibrated against the risk of the test, not against an unreachable ideal of completeness.

The risk framework suggests a practical reorientation. Rather than treating prediction as suspect, it shifts the question from "how complete is the causal story?" to "how well calibrated are claims to their tested scope?" This means distinguishing observational from interventional predictions, pre-committing to novel tests before seeing data, and treating generalization performance as provisional evidence for scope (cf. Busemeyer and Wang, 2000). None of this is easy: delineating theoretical scope and separating core from auxiliary assumptions remain genuine challenges, which partly explains why the illusions the authors describe are so persistent. But a framework that makes these demands explicit is more productive than one that measures understanding against an unreachable ideal. The solution lies not in distrusting prediction but in demanding that predictions be severe enough to warrant the confidence they carry.

Acknowledgments

K.Y. is supported by a Research Foundation – Flanders (FWO) project (G079520N) and in part by the Research Fund of KU Leuven (C14/23/062).

References

Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science*, 35(3):404–426.

Busemeyer, J. R. and Wang, Y.-M. (2000). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology*, 44(1):171–189.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, Cambridge.

Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology*, 46(4):806–834.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition.

Roberts, S. and Pashler, H. (2000). How Persuasive Is a Good Fit? A Comment on Theory Testing. *Psychological Review*, 107(2):358–367.

Shiffrin, R. M., Stigler, S., and Keil, F. (2026). Illusions of understanding in the sciences. *PsyArXiv*. Preprint; target article for Computational Brain & Behavior special issue.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3):289–310.

Stigler, S. M. (2006). Isaac Newton as a Probabilist. *Statistical Science*, 21(3):400–403.

Van Dongen, N., Sprenger, J., and Wagenmakers, E.-J. (2023). A Bayesian perspective on severity: Risky predictions and specific hypotheses. *Psychonomic Bulletin & Review*, 30(2):516–533.

Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review*, 127(1):136–145.

Woodward, J. (2004). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.