Optimal Post-Hoc Theorizing

Andrew Y. Chen

Federal Reserve Board

May 2025*

Abstract

For many economic questions, the empirical results are not interesting unless they are strong. For these questions, theorizing before the results are known is not always optimal. Instead, the optimal sequencing of theory and empirics trades off a "Darwinian Learning" effect from theorizing first with a "Statistical Learning" effect from examining the data first. This short paper formalizes the tradeoff in a Bayesian model. In the modern era of mature economic theory and enormous datasets, I argue that *post hoc* theorizing is typically optimal.

JEL Classification: B41, C18, C11

Keywords: Publication Bias, Machine Learning, Predictivism vs Accommodation, HARKing

*email:andrew.y.chen@frb.gov. I thank Irene Caracioni for excellent research assistance, and Alejandro Lopez-Lira, Matt Ringgenberg, Mish Velikov, and Tom Zimmermann for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System.

1 Introduction

Theories formed after observing empirical results (*post hoc* theories), are viewed with suspicion by social scientists (e.g. Kerr (1998); Harvey (2017)). Yet some of the most successful theories in all of science were formed this way (e.g. gravity, quantum mechanics).[1] Consistent with this confusion, the philosophy literature has long debated the merits of *post hoc* vs *a priori* theorizing (Barnes (2022))

This paper provides a Bayesian model for understanding this "paradox." It shows *post hoc* theory is clearly suboptimal if the sole goal of research is unbiased empirical results. Given statistics' 100-year obsession with unbiasedness (Efron (2001)), it is perhaps unsurprising that *post hoc* theory is viewed suspiciously.

However, the goal of research is typically more than unbiased empirical results. Another ubiquitous goal of research is to find "a good idea," whether the idea is an investment strategy, health intervention, or model of human language. In such settings, statistical bias may matter little, as long as research provides a powerful solution.

If the goal is a "good idea," then the optimal research method trades off a *Darwinian Learning* effect with a *Statistical Learning* effect. Darwinian Learning comes from weeding out bad theories by subjecting them to prediction competitions. Statistical Learning simply comes from theorists improving their ideas after looking at data. If Statistical Learning is stronger than Darwinian Learning, then *post hoc* theorizing is optimal.

In the modern world of enormous datasets and massive computing power, Statistical Learning is becoming more and more powerful. At the same time, the economic sciences have become mature, and Darwinian Learning has arguably run its course. For these reasons, I argue that *post hoc* theorizing is, in most cases, optimal.

For replication code, see `https://github.com/chenandrewy/Post-hoc/`.

[1] Newton (1726) even said "whatever is not deduced from the phenomena... ... have no place in experimental philosophy."

1.1 Related Literature

My model is an extension of the publication bias models (Hedges (1984); Brodeur et al. (2016); Andrews and Kasy (2019); Abadie (2020); Chen and Zimmermann (2020); Jensen, Kelly, and Pedersen (2023); Kasy and Spiess (2024)). In these papers, it is unclear whether *post hoc* theory is harmful. In fact, the models in these papers exhibit the irrelevance result found in Hempel (1966); Lakatos (1970); and elsewhere (see Section 2.4). Building on the insights of from the philosophy literature (namely Maher 1988), I show how heterogeneous theories breaks this irrelevance.

In the philosophy literature, Maher (1988, 1990) and Kahn, Landsburg, and Stockman (1992; 1996) (KLS) study *post hoc* theorizing under heterogeneous theories. They document the selection effect that I call Darwinian Learning, and conclude that *a priori* theorizing is optimal, at least in normal scientific settings. Amid the centuries of debate (e.g. Leibniz (1969); Newton (1726); Keynes (1921)), Barnes (1996) describes Maher's analysis as "the closest thing to an illuminating account of predictivism in existence." Predictivism is the view that *a priori* theorizing is optimal.

My paper builds on Maher and KLS by showing how there is an offsetting effect to Darwinian Learning, namely Statistical Learning. This effect is ruled out by the assumptions in Maher and KLS. Statistical Learning is perhaps a natural extension of one of Howson and Franklin's (1991) criticisms of Maher (1988) and (1990), though Maher (1993) also points out flaws in Howson and Franklin's (1991) criticisms. My paper provides clarity to this debate. Also unlike Howson and Franklin, I show how to connect Maher's and KLS's ideas to the models of publication bias, and the broader statistics literature on large scale inference (Efron (2012)).

2 A Very Simple Model of Research

Idea i is randomly-drawn from a set $\{1, 2, ..., N\}$, and has quality μ_i. μ_i is unknown but researchers can observe the measured quality

$$\hat{\mu}_i = \mu_i + \varepsilon_i \tag{1}$$

where $E(\varepsilon_i) = 0$. i may be a real-world choice for readers (e.g. an investment strategy), in which case μ_i is the realized, quality of i after the research is finished ("post-research"). Or i may be an explanation for some phenomenon (e.g. a model of obesity in adolescents), in which case μ_i is the explanation's fit to the phenomenon, post-research. In either case, higher μ_i is better.

Using theory rules out some ideas:

$$i \text{ is consistent with theory if } i \in S. \tag{2}$$

where

$$S \subseteq \{1, 2, ..., N\}. \tag{3}$$

"Theorizing" turns S into a selected idea i^*, and theorizing is either *a priori* or *post hoc*:

- *a priori*: the researcher writes down a theory that recommends a selected idea i^*, which is randomly-selected from S. (In this simple model, all ideas are equally consistent with theory.)
- *post hoc*: the researcher first examines the data (observes $\{\hat{\mu}_1, \hat{\mu}_2, ...\hat{\mu}_N\}$). Then she writes down a theory that results in selecting

$$i^* = \arg\max_{i \in S} \hat{\mu}_i. \tag{4}$$

  (The researcher chooses the idea with the highest measured quality, subject to the idea being consistent with theory.)

In either case, the theory is some math or text that explains why i^* is a good idea. In this simple model, the precise nature of the theory is not important, beyond that it argues for selecting i^*.

2.1 *A Priori* Theorizing is the Unbiased Ideal

If the sole goal of research is to find an unbiased estimate of idea quality, then *a priori* theorizing achieves this goal. The expected $\hat{\mu}_i$ from *a priori* theorizing satisfies

$$E(\hat{\mu}_i \mid i \in S) = E(\mu_i \mid i \in S). \tag{5}$$

where i is randomly selected from S. In contrast, the expected $\hat{\mu}_i$ from *post hoc* theorizing is clearly biased:

Lemma 1.

$$E\Big(\hat{\mu}_i \Big| i = \arg\max_{j \in S} \hat{\mu}_j\Big) > E\Big(\mu_i \Big| i = \arg\max_{j \in S} \hat{\mu}_j\Big). \tag{6}$$

Proof. The LHS can be written as

$$\begin{aligned}
&E\left(\mu_i \Big| i = \arg\max_{j \in S} \hat{\mu}_j\right) + E\left(\varepsilon_i \Big| i = \arg\max_{j \in S} \hat{\mu}_j\right) \\
&= E\left(\mu_i \Big| i = \arg\max_{j \in S} \hat{\mu}_j\right) + E\Big(\varepsilon_i \Big| i \in S, \{\varepsilon_i > \hat{\mu}_j - \mu_i, \quad \forall j \in (S \setminus \{i\})\}\Big)
\end{aligned}$$

The first term is the RHS of Equation (5). Thus we just need to show the second term is positive.

The second term is positive because $E\left(\varepsilon_i \big| i \in S\right) = 0$, and because the second condition on ε_i cuts off the lower tail of the distribution. □

Intuitively, $\hat{\mu}_i$ contains both μ_i and measurement error. Selecting on large $\hat{\mu}_i$ then selects for positive measurement error, leading to a biased estimate.

The preference for Equation (5), and the fear of Equation (6), goes back to Fisher (1925). As described in Efron (2001):

> *From the point of view of statistical development, the twentieth century might be labeled "100 years of unbiasedness." Following Fisher's lead, most of our current statistical theory and practice revolves around unbiased or nearly unbiased estimates (particularly MLEs), and tests based on such estimates. The power of this theory has made statistics the dominant interpretational methodology in dozens of fields.*

Taken with Lemma 1, it is no wonder then, that economists are suspicious of *post hoc* theorizing.

2.2 In Practice, *Post Hoc* Theorizing is Optimal

In an ideal world, estimates from *a priori* theorizing are all you need. With many, many of these estimates, one eventually has estimates for every idea, including the best ideas.

But in the real world, consumers and producers of research have limited time. Consumers of research lack the time to read about every idea. Producers of research lack the time to carefully study every idea.

To introduce this real-world limitation, suppose research is restricted to reporting only a single idea, and readers are interested in the idea with the highest quality.

In this case, *post hoc* theorizing is actually optimal. *Post hoc* theorizing uses both the information in theory (Equation (3)) and the information in the data (Equation (1)), improving its expected quality:

Lemma 2.

$$E\left(\mu_i \middle| i = \arg\max_{i' \in S} \hat{\mu}_{i'}\right) > E\left(\mu_i \mid i \in S\right). \tag{7}$$

Proof. The LHS can be written as

$$E\left(\mu_i \middle| i \in S, \{\mu_i > \hat{\mu}_j - \varepsilon_i, \quad \forall j \in (S \setminus \{i\})\}\right)$$

The second condition in this expression cuts off the lower tail of the distribution of μ_i. Thus, this expression exceeds the expectation of μ_i conditioning on $i \in S$ alone, which is the RHS of Equation (7). □

Lemmas 1 and 2 are illustrated in Figure 1. It simulates 200 selected ideas, with the number of potential ideas $N = 100$, $\mu_i \sim \text{Normal}(0, 1)$, and $\varepsilon_i \sim \text{Normal}(0, 1)$. *A priori* theorizing leads to less biased estimates, seen in how the dots lie closer to the 45 degree line. However, *post hoc* theorizing leads to higher quality ideas, seen in how the stars tend to lie toward the right side of the chart.

The literature on stock market anomalies is an example of Lemma 2. Readers are interested in both the magnitude of anomalies, as well as which ones are the strongest. But assuming that the magnitude meets some minimal standard, readers with limited time will just want to know which anomalies will perform the best in the future. Lemma 2 shows that, in this case, researchers *should* mine the data, and report what has worked best in the past. This prescription is exactly the reverse of the conventional wisdom, that emphasizes the "dangers" of data mining (Sullivan, Timmermann, and White 1999; Harvey, Liu, and Zhu 2016). However, it seems to be in-line with empirical

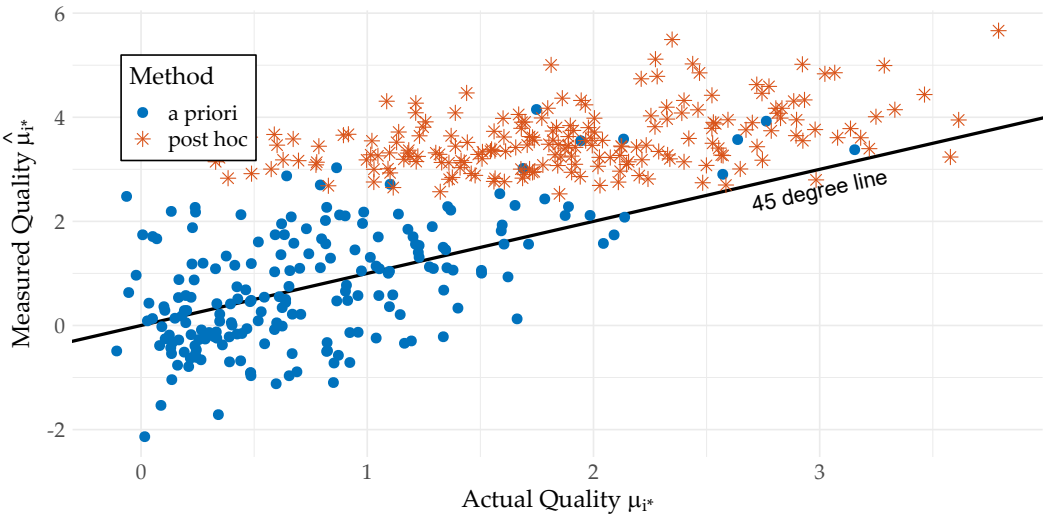


Figure 1: 200 Ideas Generated by a Very Simple Model of Research

practice, and performs quite well (Chen, Lopez-Lira, and Zimmermann 2024).

Large language models (LLMs) are another example. These models are tuned to perform well on common benchmarks like MMLU (Measuring Massive Multitask Language Understanding) (e.g. Guo et al. (2025)). Thus, the performance on these benchmarks is biased upward, just as in Lemma 1. But in practice, this bias is not important, as long as the resulting out-of-sample performance is strong. Tuning improves out-of-sample performance, as seen in Lemma 2.

2.3 *Post Hoc* Theories are Falsifiable and Scientific

Idea i^*, as well as the theory for why i^* is a good idea, are eventually tied to post-research data through μ_{i^*} (see discussion after Equation (1)). Without this link, the theories might as well be fairy tales. Whether fairy tales are better told *a priori* or *post hoc* is beyond the scope of this paper.

Because i^* is tied to post-research data, the ideas and theories in this model are falsifiable and scientific, in the sense of Popper (1959). This holds regardless of whether i^* is chosen *a priori* or *post hoc*. Popper does not say that researchers should ignore data when making predictions.

Perhaps because of Kerr (1998) ("HARKing: Hypothesizing after the Results are Known"), many researchers equate *post hoc* theorizing with unfalsifiability. However, as seen in this model, constructing theories *post hoc* is

entirely consistent with Popper's notion of science.

This confusion likely stems from Kerr's loose use of language. The paper has a section titled "HARKed Hypotheses Fail Popper's Criterion of Disconfirmability." But the text below the title clarifies, "[a] HARKed hypothesis fails this criterion, at least in a narrow, temporal sense." In other words, the text in the section explains that the section title is not necessarily true. In fact, it seems equally reasonable to say that HARKed hypotheses fail Popper's criterion *only* in a narrow, temporal sense. Errors like these are found throughout Kerr (1998). See Rubin (2022) for a thorough critique.

Popper does imply that theories should be well-defined and constrained. For example, Popper (1985) argues that Marxism was refuted by many empirical facts, but then "immunized itself against the most blatant refutations" by the addition of *ad hoc* hypotheses. In the lens of this model, Marxism lacks a consistent definition of the set S. Similarly, Popper argues that Freudian theories "do not exclude any physically possible human behavior." This is equivalent to saying S includes the set of all ideas $\{1, 2, ..., N\}$. Throughout my paper, I assume that S is well-defined and constrained, though one might be concerned that this assumption is inappropriate for some social sciences.

2.4 An Irrelevance Result

In practice, the Fisherian ideal is impossible. Even if all researchers use theory a priori, readers with time constraints are more likely to read the research if the measured effect is large. This limited attention is arguably the raison d'etre of both peer review (Klamer and Dalen (2002)) and publication bias (Chen and Zimmermann (2022))

To model limited attention, suppose *a priori* theory actually involves two steps. First, researchers study all ideas in S and draft up their theories and empirical findings in working papers. However, not all ideas are read. Due to limited attention, only the idea with the largest measured quality becomes well-known and consumed by the public. The expected quality of this, more realistic, *a priori* theorizing is

$$E\Big(\mu_i\Big|i \in S, i = \arg\max_{i' \in S} \hat{\mu}_{i'}\Big) = E\Big(\mu_i\Big|i = \arg\max_{i' \in S} \hat{\mu}_{i'}\Big), \tag{8}$$

which is exactly the same as the quality of *post hoc* theory (Lemma 2).

A similar irrelevance is noted in many works of philosophy (e.g. Hempel (1966); Lakatos (1970); Rosenkrantz (1977); Gardner (1982)). But as noted by Maher (1988) and Kahn, Landsburg, and Stockman (1996), this irrelevance can be broken if theories are endogenous.

3 Endogenous, Heterogeneous Theories

Let's make the model richer, with endogenous, heterogeneous theories. This richer model is a generalization of Maher (1988) and Kahn, Landsburg, and Stockman (1996). Importantly, it allows for an effect I call "Statistical Learning." As in Section 2.2, I assume that the research community has limited time, and is primarily interested in finding ideas with the highest quality.

As before, there are ideas $i \in \{1, 2, ..., N\}$, measured idea quality $\hat{\mu}_i$, and true idea quality μ_i. But now theories come from combining a "data input" with a "theory type."

The data input ($\mathcal{D}$ or $\mathcal{O}$) is known. $\mathcal{D}$ is the case that the data input includes all of the measured effects $(\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_N)$. $\mathcal{O}$ is the case that the theory is given access to none of these effects. *Post hoc* theorizing, then, is represented by $\mathcal{D}$, while *a priori* theorizing is $\mathcal{O}$.

The theory type has a quality T which is unknown. For simplicity, assume the quality is either good (represented by G) or bad (B). Intuitively, not all theories types are the same, and we may not know how good a particular theory type is.

Combining a theory type with a data input leads to a theory, which in turn provides a recommended idea i^*. As before, i^* is a random integer with support S, and the theory is some math and/or text that explains why i^* is recommended. But now I'll use conditional probability notation to account for the data input and theory type. For example, $i^*|G, \mathcal{O}$ is the recommended idea generated by a good theory type and no data (*a priori*).

It's reasonable to think that the good theory type leads to higher quality ideas, *a priori*. This can be formalized by first order stochastic dominance:

$$P\left(\mu_{i^*} > x \mid G, \mathcal{O}\right) \geq P\left(\mu_{i^*} > x \mid B, \mathcal{O}\right), \quad \forall x \in \mathbb{R}. \tag{9}$$

For example, one may think that while bad theory types recommend any idea

in S with equal probability, good theory types are twice as likely to recommend ideas from the top quartile of μ_i (as compared to the second-to-top quartile). An implication of Equation (9) is that good theory types typically lead to higher measured quality $\hat{\mu}_{i^*}$ than bad theory types.

If theory is done *post hoc*, researchers examine measured qualities $\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_N$, as well as the theory type, to construct a theory that selects idea $i^*|T, \mathcal{D}$. I allow $i^*|T, \mathcal{D}$ to be general, but assume the following restriction:

$$P\left(i^* = \arg\max_{i \in S} \hat{\mu}_i \mid B, \mathcal{D}\right) = 1.0, \tag{10}$$

that is, using bad type theories always lead researchers to select the idea with the strongest measured quality (provided the idea is consistent with *some* theory). This assumption can be thought of as bad theory types being unable to distinguish between ideas in S, and Bayesian researchers who optimize on the posterior mean based on this information and $\hat{\mu}_i$ (see Chen and Dim 2025).

After i^* is chosen, readers decide if they are interested in the theory and idea. Assume readers are uninterested unless

$$\hat{\mu}_{i^*} > h, \tag{11}$$

where h is some kind of economic and/or statistical hurdle. Only theories and ideas readers are interested in are published. This assumption follows the econometric literature on publication bias (Andrews and Kasy (2019)).

3.1 Darwinian Learning

An immediate implication of heterogeneous theories is heterogeneous measured quality:

Lemma 3.

$$P\left(\hat{\mu}_{i^*} > h | G, \mathcal{O}\right) > P\left(\hat{\mu}_{i^*} > h | B, \mathcal{O}\right) \tag{12}$$

Proof. Since ε_i is i.i.d., adding it to μ_{i^*} preserves first-order stochastic dominance. □

Lemma 3 provides an alternative way to think about the Chen, Lopez-Lira,

and Zimmermann (2022) (CLZ) "peer review vs data mining" experiment. CLZ compare stock trading ideas from peer review to data-mined trading ideas, using post-publication returns. If we call the post-publication returns $\hat{\mu}_{i^*}$, neither the peer-reviewed nor data-mined ideas had access to this data, so $\mathcal{O}$ holds for both groups of ideas. Then, one can think of peer-reviewed ideas as $i^*|T, \mathcal{O}$, since we do not know if the theory type is G or B. In contrast, we can think of the data-mined ideas $i^*|B, \mathcal{O}$. As powerfully demonstrated by Novy-Marx and Velikov (2025), anyone can add text to these ideas and call it a theory.

From this framing, CLZ's empirical results are a test of whether G theory types exist. If G theory types comprise a significant fraction of the theories in the CLZ sample, then Lemma 3 implies that the published strategies have higher $\hat{\mu_{i^*}}$. Unfortunately, CLZ find that published strategies fail to outperform, implying that G theories are rare.

The CLZ experiment illustrates the Darwinian selection of theories. If we force theorists to announce their ideas before looking at the data, then the bad theory types cannot hide behind data mining. This intuition helps justify the belief that *a priori* theorizing provides "discipline" and that *post hoc* theorizing is "too easy." The following proposition formalizes this idea:

Proposition 1. *[Darwinian Selection of Theories]*

$$P\left(G|\mathcal{O}, \hat{\mu}_{i^*} > h\right) - P\left(G|\mathcal{D}, \hat{\mu}_{i^*} > h\right) > 0$$

Proof. Apply Bayes rule to the LHS and simplify to yield

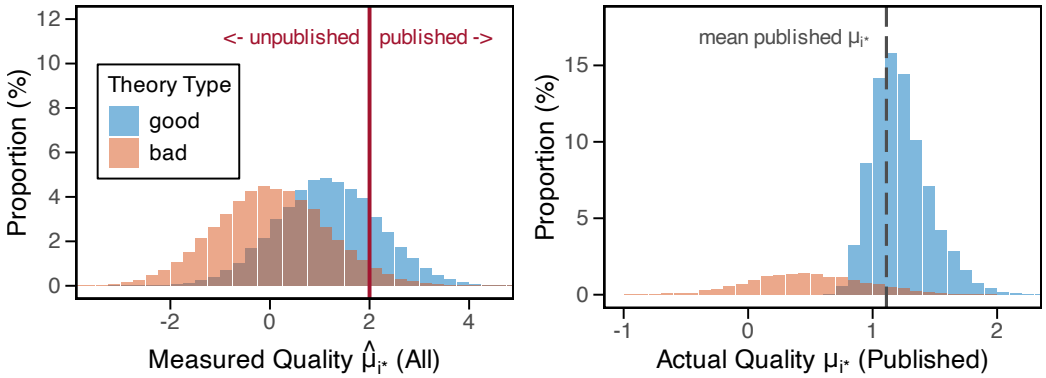
$$\frac{P\left(\hat{\mu}_{i^*} > h|G, \mathcal{O}\right)}{P\left(\hat{\mu}_{i^*} > h|B, \mathcal{O}\right)} > \frac{P\left(\hat{\mu}_{i^*} > h|G, \mathcal{D}\right)}{P\left(\hat{\mu}_{i^*} > h|B, \mathcal{D}\right)}$$

Lemma 3 shows that the LHS is greater than 1.0. But since bad theories always select the largest $\hat{\mu}_i$ *post hoc* (Equation (10)), the RHS is at most 1.0. □

Proposition 1 is illustrated in Figure 2. It shows histograms generated by parameters deliberately chosen to highlight the power of Darwinian selection.

Under *a priori* theorizing, published ideas mostly come from the good theory types (Panel (a), left). Naturally, good theory types are better at separating good ideas from bad ones, *a priori*. *Post hoc*, published ideas largely come from the bad theory types (Panel (b), left). This happens because bad

(a) *A Priori* Theorizing



(b) *Post Hoc* Theorizing

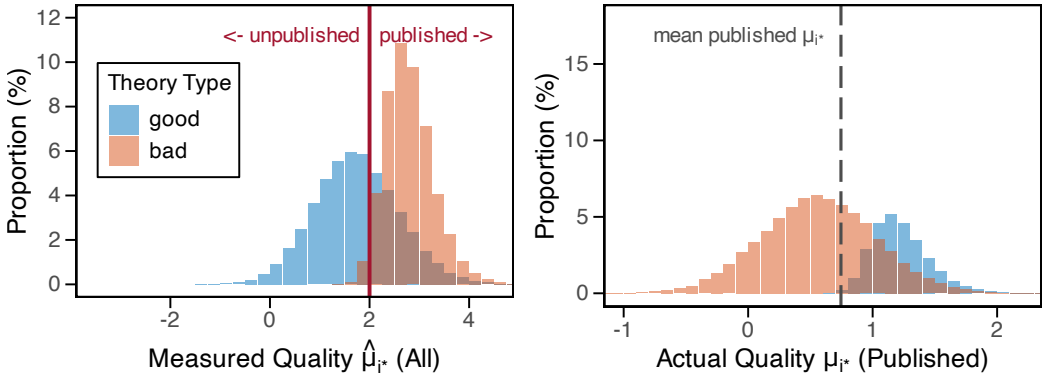


Figure 2: Darwinian Selection Illustration. Histograms show all selected ideas (All) or those that meet the hurdle $h = 2.0$ (Published). Number of ideas $N = 100$, actual quality $\mu_i \sim N(0, 0.5^2)$, noise $\varepsilon_i \sim \text{Normal}(0, 1)$. Prob of a good type is 50%. *A priori*, bad types equally weight all ideas, while good types equally weight the two best. *Post hoc*, researchers select the idea with the highest $\hat{\mu}_i$ with positive *a priori* weight.

theory types lead researchers to check far more ideas for the highest measure quality, as these bad theories cannot discriminate among ideas. As a result, bad theory types are more likely to lead to publication, despite having lower actual quality. The final result is that *a priori* theorizing leads to published ideas with higher actual quality (vertical dashed lines).

Proposition 1 captures the key insight of Maher (1988; 1990) and Kahn, Landsburg, and Stockman (1992, 1996). If theories are heterogeneous, then forcing theorists to announce their ideas before looking at the data helps eliminate bad theories, as in Darwinian selection. In Maher's terminology, a theory is a "method," and the theory type is "reliability," but the idea is the same.

Maher and KLS push further. They claim that, not only does *a priori* theorizing produce Darwinian selection, but that the resulting hypotheses are more likely to be true. The analogue here is that $\mathcal{O}$ implies not only that G is more likely, but that μ_{i^*} is higher. We'll see that this conclusion is not necessarily true.[2]

An interesting feature of Proposition 1 is that it shows a virtue of publication bias. While requiring $\hat{\mu}_{i^*} > h$ leads to biased estimates, it helps weed out bad theories types. This result is closely analogous to Lemma 2.

3.2 Optimal Post-Hoc Theory

Research is not only interested in finding good theory types, but also good ideas. In fact, one can argue that finding good ideas is the ultimate goal.

Whether *post hoc* theory helps or hurts for finding good ideas is characterized by the following proposition:

Proposition 2. *[Optimal Post Hoc Theory]*

$$E\left(\mu_{i^*}|\mathcal{D}, \hat{\mu}_{i^*} > h\right) > E\left(\mu_{i^*}|\mathcal{O}, \hat{\mu}_{i^*} > h\right) \tag{13}$$

if and only if

$$[\textit{Statistical Learning}] > [\textit{Darwinian Learning}] \tag{14}$$

where

$$\begin{aligned}[\textit{Darwinian Learning}] \equiv & \left[P\left(G|\mathcal{O}, \hat{\mu}_{i^*} > h\right) - P\left(G|\mathcal{D}, \hat{\mu}_{i^*} > h\right)\right] \\ & \times \left[E\left(\mu_{i^*}|G, \mathcal{O}, \hat{\mu}_{i^*} > h\right) - E\left(\mu_{i^*}|B, \mathcal{O}, \hat{\mu}_{i^*} > h\right)\right]\end{aligned} \tag{15}$$

$$\begin{aligned}[\textit{Statistical Learning}] \equiv & P\left(G|\mathcal{D}, \hat{\mu}_{i^*} > h\right)\left[E\left(\mu_{i^*}|G, \mathcal{D}, \hat{\mu}_{i^*} > h\right) - E\left(\mu_{i^*}|G, \mathcal{O}, \hat{\mu}_{i^*} > h\right)\right] \\ & + P\left(B|\mathcal{D}, \hat{\mu}_{i^*} > h\right)\left[E\left(\mu_{i^*}|B, \mathcal{D}, \hat{\mu}_{i^*} > h\right) - E\left(\mu_{i^*}|B, \mathcal{O}, \hat{\mu}_{i^*} > h\right)\right]\end{aligned} \tag{16}$$

The proof is in Appendix A.

The proposition says that whether *post hoc* or *a priori* theorizing leads to

[2]Barnes (1996) revisits Maher (1988, 1990, 1993) and does not go further. His Eq (4) stops here, and considers more deeply the terms in the Bayes rule version of $P\left(G|\mathcal{O}, \hat{\mu}_{i^*} > h\right)$.

better ideas depends on the relative size of two effects:

1. Darwinian Learning: This measures the ultimate effect of Darwinian selection (Proposition 1), which occurs when researchers are forced to predict without data ($\mathcal{O}$). Intuitively, Darwinian selection improves ideas only to the extent that G theory types find higher μ_{i^*} compared to B theory types (second line of Equation (15)).

2. Statistical Learning: This measures how idea quality μ_{i^*} improves when the researcher has access to more data ($\mathcal{D}$). Just as how a Bayesian improves her inferences with new evidence, theorists develop higher quality ideas with access to data.

Naturally, if Statistical Learning exceeds Darwinian Learning, then it's often better to look at the data—i.e. *post hoc* theory may be optimal.

There is no hard and fast rule for which effect is larger. There are certainly settings where Statistical Learning is miniscule (e.g. when the data is extremely noisy). And there are certainly settings where Darwinian Learning is ineffective (e.g. when all theory types are the same).

Similarly, there are contradictory historical examples. Mendeleev's prediction of elements is a shockingly impressive example of *a priori* theorizing. But Planck's law of radiation is a shockingly impressive example of *post hoc* theorizing. Proposition 2 provides a way to understand these seemingly contradictory phenomena.

3.3 When is *post hoc* theorizing optimal?

If theories are homogenous in quality, then there is no Darwinian Learning, and thus Proposition 2 implies that *post hoc* theory is optimal.

Figure 3 illustrates this phenomenon, by examining many variations of the model from Figure 2. In Figure 3, theory types were extremely heterogeneous: bad theory types cannot eliminate any ideas, while good theory types eliminate the worst 98% of ideas. This extreme-heterogeneity model is shown in the right most markers of Figure 3. For this model, the improvement from *post hoc* theory is a negative 30%: i.e. published ideas have 30% lower quality under *post hoc* theory (top panel). Correspondingly, Darwinian Learning is very large, and far exceeds Statistical Learning (bottom panel).

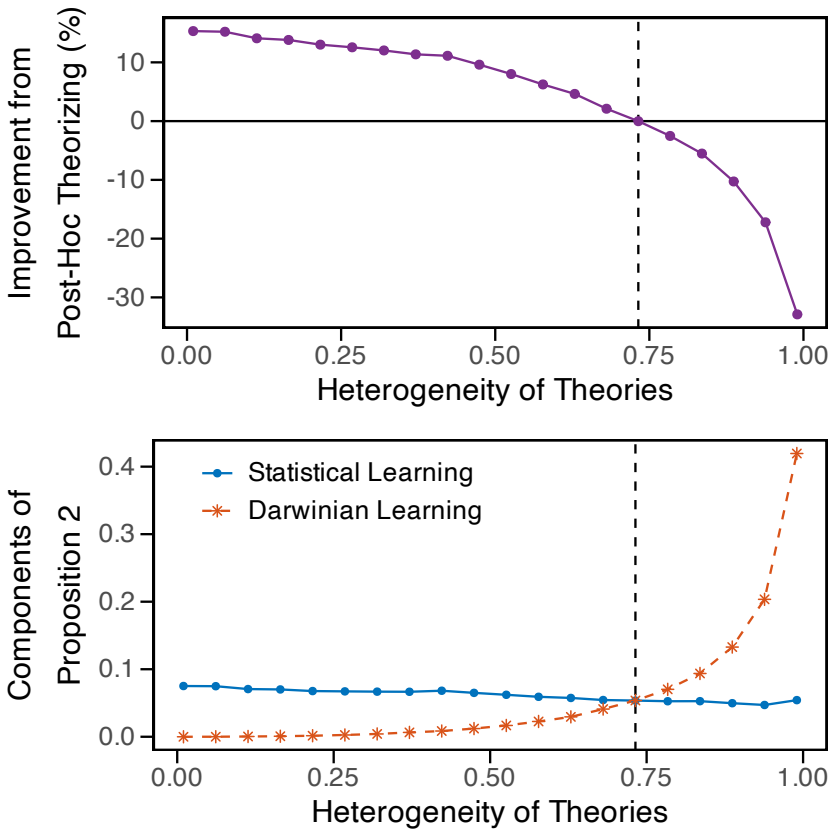


Figure 3: Optimal Theorizing vs Heterogeneity of Theories. Each marker is one model. 'Improvement from Post-Hoc Theorizing' is $E\left(\mu_{i^*}|\mathcal{D}, \hat{\mu}_{i^*} > h\right) / E\left(\mu_{i^*}|\mathcal{O}, \hat{\mu}_{i^*} > h\right) - 1$ (see Proposition 2). 'Heterogeneity of Theories' is the share of ideas that good theories can eliminate *a priori*. Otherwise, the model is the same as in Figure 2, in which bad theories cannot eliminate any ideas.

However, reducing the heterogeneity of theories leads to *post hoc* theory being optimal. Moving from right to left in Figure 3, the improvement from *post hoc* theory turns positive once good theory types can eliminate the worst 75% of ideas. Here, Statistical Learning is exactly equal to Darwinian Learning (bottom panel). For models with any less heterogeneity, *post hoc* theory is optimal.

3.3.1 Large Datasets and Optimal Theorizing

Another implication of Proposition 2 is that larger datasets tend to imply *post hoc* theory is optimal. Naturally, larger datasets imply more Statistical Learning.

To model this, one can think of measured quality $\hat{\mu}_i$ as a t-statistic, in which case a large dataset implies high $\text{Var}(\hat{\mu}_i)$. Intuitively, as the sample size increases, so does the probability of finding statistically-significant t-stats (Abadie 2020).

To formalize this interpretation, suppose that underlying Equation (1) is a panel data model:

$$x_{i,j} = \chi_i + e_{i,j}, \quad j = 1, 2, \cdots, M \tag{17}$$

$$\text{Var}(e_{i,j}) = \sigma^2, \tag{18}$$

where M is the number of observations for idea i. Moreover, suppose we fix the hurdle for readers' interest at $h = 2.0$ (see Equation (11)). Then a natural way to map Equation (17) to Equation (1) is to define $\hat{\mu}_i$ as the t-statistic for χ_i:

$$\hat{\mu}_{i,t} = \frac{\bar{x}_i}{\sigma_i}\sqrt{M} = \underbrace{\frac{\sqrt{M}}{\sigma_i}\chi_i}_{\mu_i} + \underbrace{\frac{\sqrt{M}}{\sigma_i}\bar{e}_i}_{\varepsilon_i}, \tag{19}$$

$$\text{Var}_i\left(\varepsilon_i\right) = 1.0, \tag{20}$$

where $\text{Var}_i(\varepsilon_i)$ is the variance holding fixed the idea, and Equation (20) assumes that the central limit theorem holds and σ^2 is observed. Thus, in this setting, the standard deviation of $\hat{\mu}_i$ is increasing in the sample size M.

Figure 4 illustrates how large datasets affect optimal theorizing, interpreted through the panel data model (Equations (17)-(20)). It revisits the model from Figure 2, but examines alternative choices for the variance of μ_i. The x-axis plots $\sqrt{\text{Var}(\hat{\mu}_i)}$, which can be interpreted as either the dispersion of t-statistics or a measure of the sample size.

The left-most markers correspond to the model from Figure 2. This model was selected to illustrate the power of Darwinian selection. Thus, $\text{Var}(\hat{\mu}_i)$ is close to 1.0, indicating the measured quality is close to the null distribution, and noise dominates the data. Thus, Statistical Learning is small, and *a priori* theorizing is optimal.

But as $\text{Var}(\hat{\mu}_i)$ increases, so does the amount of signal, holding fixed $\text{Var}(\varepsilon_i)$ at 1.0. The amount of Statistical Learning then increases, and *post hoc* theorizing, starts to become optimal at approximately $\sqrt{\text{Var}(\mu_i)} = 1.75$.

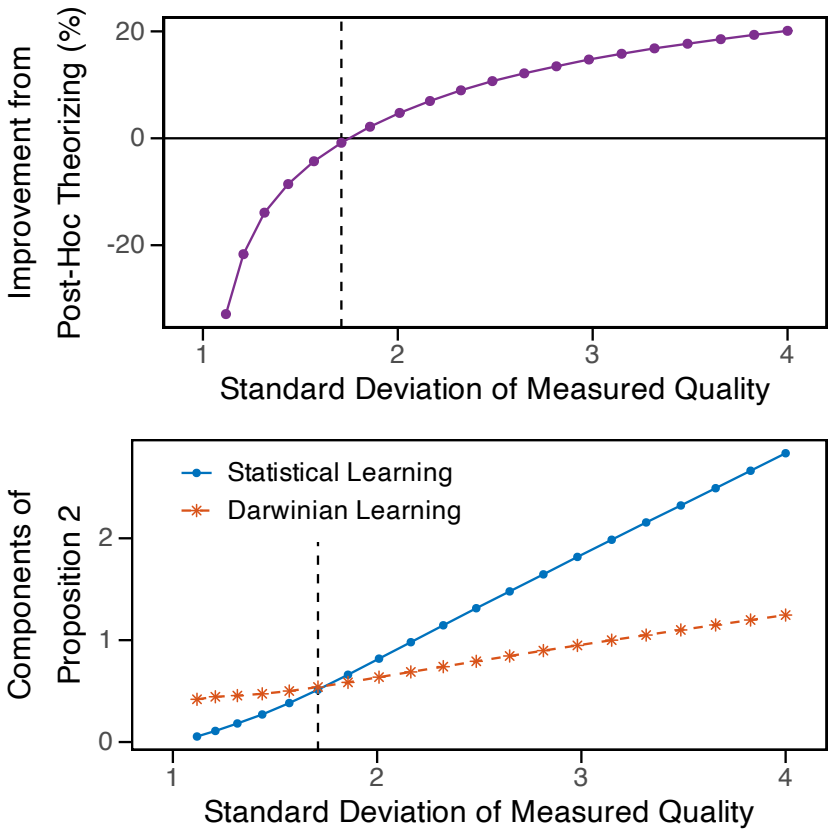


Figure 4: Optimal Theorizing vs Sample Size. Each marker is one model with a different $\sqrt{\text{Var}(\mu_i)}$. The rest of the model is the same as in Figure 2. 'Improvement from Post-Hoc Theorizing' is $E\left(\mu_{i^*}|\mathcal{D}, \hat{\mu}_{i^*} > h\right) / E\left(\mu_{i^*}|\mathcal{O}, \hat{\mu}_{i^*} > h\right) - 1$. 'Standard Deviation of Measured Quality' is $\sqrt{\text{Var}(\hat{\mu}_i)}$, which can be interpreted as the dispersion of t-statistics or a measure of sample size (Equations (17)-(20)).

$\sqrt{\text{Var}(\mu_i)} = 1.75$ relatively small. For comparison, Chen and Zimmermann (2020) and Jensen, Kelly, and Pedersen (2023) estimate $\sqrt{\text{Var}(\mu_i)} \approx 3.0$ for empirical asset pricing (see discussion in Chen and Zimmermann 2022). For settings like this, where $\hat{\mu}_i$ provides a strong signal about the underlying μ_i, Statistical Learning most likely exceeds Darwinian Learning, and thus *post hoc* theorizing is typically optimal.

3.3.2 Optimal Theorizing in Modern Economics

As a field of research matures, institutions arise that standardize the many aspects of research, including the peer review process, the statistical analysis, and theory. It is reasonable to think, then, that mature fields have theories that are relatively homogeneous in quality. In fact, homogeneous theory quality is a reasonable definition of a mature field.

Economics is arguably mature. Before the 1950s, there was wild variety in the way that economists theorized. But theory began to solidify with the contributions of Arrow and Samuelson. And though behavioral economics has risen in popularity in recent decades, and the 2008 financial crisis brought on significant criticism of economic models, the basic structure of theory has been largely stable since the 1980s. It is thus reasonable to think that economic theories are fairly homogeneous in quality, and that Darwinian Learning is small.

At the same time, the modern era has seen the rise of huge datasets and enormous computing power. As discussed in Section 3.3.1, this implies that standardized measures of idea quality are dispersed, and thus Statistical Learning is large.

Taken together, these arguments imply that *post hoc* theory is typically optimal in the modern era of economics.

This argument has some surprising implications. Pre-analysis plans should *not* be followed. Journals should *favor* theories that accommodate the data, *post hoc*. At least, these are the prescriptions for a literature that focuses on finding the best ideas, and places less emphasis on unbiasedness.

While it may feel uncomfortable to favor results over unbiasedness, this is precisely the approach taken by the computer science literature. Following this practical route, computer science has essentially taken over machine learning, which could have been the territory of statisticians. Perhaps the maturation

of statistical theories, as well as the rise of big data, tilted the balance in favor of *post hoc* theorizing, and thus the dominance of computer scientists.

4 Conclusion

This paper presents a framework for understanding several questions about the scientific method: Why is *post hoc* theorizing viewed as a problem? How do we square this problem with highly-successful *post hoc* theories? Does the classical view of *post hoc* theory still hold up in the modern era of big data?

The framework shows that the distrust of *post hoc* theorizing is to a significant extent a relic of idealized, pre-modern statistics. With practical constraints on researchers' time, and a focus on results over unbiasedness, *a priori* theorizing is not always superior. Instead, there is a trade-off between Darwinian Learning, which comes from forcing theorists into prediction contests, and Statistical Learning, which arises as researchers learn from data. With modern datasets and computing power, Statistical Learning is clearly very significant. At the same time, it is unclear that Darwinian Learning still matters, in a world of mature theories.

A caveat is that *a priori* theorizing has benefits that are omitted from my analysis. Most important, Barnes (2008) points out that prediction contests provide an accessible, democratic way to establish what is good science. The main alternative is the peer review process, which is inscrutable to outsiders, and can potentially be abused.[3]

A Proof of Proposition 2

Proof. For ease of notation, let $\tilde{E}$ be the expectation operator conditioned on $\hat{\mu}_{i^*} > h$ and define $\tilde{P}$ similarly. Also define conditioning on $I \in \{\mathcal{O}, \mathcal{D}\}$ and $I' \in \{\mathcal{O}, \mathcal{D}\}$ as

$$\tilde{E}\left\{\tilde{E}\left(\mu|T,I\right)|I'\right\} \equiv \tilde{P}\left(G|I'\right)\tilde{E}\left(\mu|G,I\right) + \tilde{P}\left(B|I'\right)\tilde{E}\left(\mu|B,I\right), \tag{21}$$

[3]Additionally, KLS argue that the choice of *a priori* vs *post hoc* theorizing may be endogenous, which can lead to additional selection effects, over and above Proposition 2. However, the basic logic that *a priori* theorizing helps through inducing selection is still captured by Proposition 2.

which can be rewritten as

$$\tilde{E}\left\{\tilde{E}\left(\mu|T,I\right)|I'\right\} \equiv \tilde{E}\left(\mu|B,I\right) + \tilde{P}\left(G|I'\right)\left\{\tilde{E}\left(\mu|G,I\right) - \tilde{E}\left(\mu|B,I\right)\right\}. \quad (22)$$

The expected quality from *a priori* theory can be written as

$$\begin{aligned}\tilde{E}\left\{\mu_{i^*}|\mathcal{O}\right\} = \tilde{E}\left\{\tilde{E}\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{O}\right\} - \tilde{E}\left\{\tilde{E}\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{D}\right\} \\ + \tilde{E}\left\{\tilde{E}\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{D}\right\}, \quad (23)\end{aligned}$$

where the first term uses iterated expectations and the last two terms sum to zero. Thus the expected quality difference of *a priori* vs *post hoc* theory is

$$\begin{aligned}\tilde{E}\left\{\mu_{i^*}|\mathcal{O}\right\} - \tilde{E}\left\{\mu_{i^*}|\mathcal{D}\right\} = \tilde{E}\left\{\tilde{E}\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{O}\right\} - \tilde{E}\left\{\tilde{E}\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{D}\right\} \\ - \left\{\tilde{E}\left[\mu_{i^*}|\mathcal{D}\right] - \tilde{E}\left\{\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{D}\right\}\right\} \quad (24)\end{aligned}$$

The second line of the RHS of (24) is [Statistical Learning] (just apply iterated expectations to $\tilde{E}\left[\mu_{i^*}|\mathcal{D}\right]$).

The first line of the RHS of (24) can be rewritten using the law of total probability and (22):

$$\begin{aligned}&\tilde{E}\left\{\tilde{E}\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{O}\right\} - \tilde{E}\left\{\tilde{E}\left[\mu_{i^*}|T,\mathcal{O}\right]|\mathcal{D}\right\} \\ &= \tilde{E}\left[\mu_{i^*}|B,\mathcal{O}\right] + \tilde{P}\left(G|\mathcal{O}\right)\left\{\tilde{E}\left[\mu_{i^*}|G,\mathcal{O}\right] - \tilde{E}\left[\mu_{i^*}|B,\mathcal{O}\right]\right\} \\ &\quad - \tilde{E}\left[\mu_{i^*}|B,\mathcal{O}\right] - \tilde{P}\left(G|\mathcal{D}\right)\left\{\tilde{E}\left[\mu_{i^*}|G,\mathcal{O}\right] - \tilde{E}\left[\mu_{i^*}|B,\mathcal{O}\right]\right\} \\ &= \left[\tilde{P}\left(G|\mathcal{O}\right) - \tilde{P}\left(G|\mathcal{D}\right)\right]\left\{\tilde{E}\left[\mu_{i^*}|G,\mathcal{O}\right] - \tilde{E}\left[\mu_{i^*}|B,\mathcal{O}\right]\right\}, \quad (25)\end{aligned}$$

and the last line is [Darwinian Learning]. □

References

Abadie, Alberto (2020). "Statistical nonsignificance in empirical economics". In: *American Economic Review: Insights* 2.2, pp. 193–208.

Andrews, Isaiah and Maximilian Kasy (2019). "Identification of and correction for publication bias". In: *American Economic Review* 109.8, pp. 2766–94.

Barnes, Eric (1996). "Discussion: Thoughts on Maher's predictivism". In: *Philosophy of Science* 63.3, pp. 401–410.

Barnes, Eric Christian (2008). "The paradox of predictivism". In.

— (2022). "Prediction versus Accommodation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University.

Brodeur, Abel et al. (2016). "Star wars: The empirics strike back". In: *American Economic Journal: Applied Economics* 8.1, pp. 1–32.

Chen, Andrew Y and Chukwuma Dim (2025). "High-throughput asset pricing". In: *arXiv preprint arXiv:2311.10685*.

Chen, Andrew Y, Alejandro Lopez-Lira, and Tom Zimmermann (2022). "Peer-reviewed theory does not help predict the cross-section of stock returns". In: *arXiv preprint arXiv:2212.10317*.

— (2024). "Does peer-reviewed theory help predict the cross-section of stock returns?" In: *arXiv e-prints*, arXiv–2212.

Chen, Andrew Y and Tom Zimmermann (2020). "Publication bias and the cross-section of stock returns". In: *The Review of Asset Pricing Studies* 10.2, pp. 249–289.

— (2022). "Publication Bias in Asset Pricing Research". In: *arXiv preprint arXiv:2209.13623*.

Efron, Brad (2001). "[statistical modeling: The two cultures]: Comment". In: *Statistical Science* 16.3, pp. 218–219.

Efron, Bradley (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press.

Fisher, RA (1925). "Statistical methods for research workers." In.

Gardner, Michael R. (1982). "Predicting Novel Facts". In: *British Journal for the Philosophy of Science* 33, pp. 1–15.

Guo, Daya et al. (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". In: *arXiv preprint arXiv:2501.12948*.

Harvey, Campbell R (2017). "Presidential address: The scientific outlook in financial economics". In: *The Journal of Finance* 72.4, pp. 1399–1440.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "... and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1, pp. 5–68.

Hedges, Larry V (1984). "Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences". In: *Journal of Educational Statistics* 9.1, pp. 61–85.

Hempel, Carl G. (1966). *Philosophy of Natural Science*. Englewood Cliffs: Prentice-Hall.

Howson, Colin and Allan Franklin (1991). "Maher, Mendeleev and Bayesianism". In: *Philosophy of Science* 58.4, pp. 574–585.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen (2023). "Is there a replication crisis in finance?" In: *The Journal of Finance* 78.5, pp. 2465–2518.

Kahn, James A, Steven E Landsburg, and Alan C Stockman (1992). "On novel confirmation". In: *The British Journal for the Philosophy of Science* 43.4, pp. 503–516.

— (1996). "The positive economics of methodology". In: *Journal of Economic Theory* 68.1, pp. 64–76.

Kasy, Maximilian and Jann Spiess (2024). *Optimal pre-analysis plans: Statistical decisions subject to implementability*. Tech. rep. CESifo Working Paper.

Kerr, Norbert L (1998). "HARKing: Hypothesizing after the results are known". In: *Personality and social psychology review* 2.3, pp. 196–217.

Keynes, John Maynard (1921). *A Treatise on Probability*. London: Macmillan.

Klamer, Arjo and Hendrik P van Dalen (2002). "Attention and the art of scientific publishing". In: *Journal of economic methodology* 9.3, pp. 289–315.

Lakatos, Imre (1970). "The Methodology of Scientific Research Programmes". In: *Criticism and the Growth of Knowledge*. Ed. by Imre Lakatos and Alan Musgrave. London: Cambridge University Press, pp. 91–196.

Leibniz, G. W. (1969). "Letter to Herman Conring". In: *Philosophical Papers and Letters*. Ed. by L. E. Loemker. Dordrecht: D. Reidel, pp. 186–191.

Maher, Patrick (1988). "Prediction, accommodation, and the logic of discovery". In: *PSA: Proceedings of the Biennial meeting of the philosophy of*

science association. Vol. 1988. 1. Cambridge University Press, pp. 272–285.

Maher, Patrick (1990). "How prediction enhances confirmation". In: *Truth or consequences: Essays in honor of Nuel Belnap*. Springer, pp. 327–343.

— (1993). "Discussion: Howson and Franklin on Prediction". In: *Philosophy of Science* 60.2, pp. 329–340.

Newton, Isaac (1726). *Philosophiæ Naturalis Principia Mathematica*. 3rd ed. General Scholium appendix, pp. 526-530. London: William & John Innys.

Novy-Marx, Robert and Mihail Z Velikov (2025). *AI-Powered (Finance) Scholarship*. Tech. rep. National Bureau of Economic Research.

Popper, Karl (1959). *The Logic of Scientific Discovery*. Originally published as *Logik der Forschung* (1934). London: Routledge.

— (1985). "The Problem of Demarcation". In: *Popper Selections*. Ed. by David Miller. Princeton, NJ: Princeton University Press, pp. 118–130.

Rosenkrantz, Roger D. (1977). *Inference, Method and Decision*. Dordrecht: D. Reidel.

Rubin, Mark (2022). "The costs of HARKing". In: *The British Journal for the Philosophy of Science*.

Sullivan, Ryan, Allan Timmermann, and Halbert White (1999). "Data-snooping, technical trading rule performance, and the bootstrap". In: *The journal of Finance* 54.5, pp. 1647–1691.