Failure Modes of Backtesting

By Adam Dhillon, PhD, LongTail Alpha Portfolio Research

May 6th, 2026

Executive Summary

- Misleading backtests are easily constructed, and researchers are often incentivized to build misleading backtests.
- There are many ways in which a backtest can be made misleading, intentionally or otherwise.
- It is often possible to spot deceptive forces in a backtest. Less frequently, it is possible to prevent these from affecting backtesting results.

Reliance on backtesting can expose an investor to a variety of errors. These errors are often exacerbated by conflicting incentives: a marketing department may engineer backtests to look especially tempting despite a lack of true forward-looking utility. However, they can also be generated by well-meaning, but ultimately careless researchers in a hunt for alpha. Regardless of incentives, there are a few failure modes in backtesting that should be kept in mind, and in some cases, some clear indicators of poor scientific reasoning that ought to raise suspicion.

The Fundamental Problems of Backtesting and Frequentist Statistics

While it is not the primary purpose of this discussion, it is important to acknowledge two fundamental problems. One, in backtesting, a researcher is often attempting to gather evidence about future market behavior. However, financial markets are a game played by innumerable intelligent agents in a randomly evolving universe, and so the market behavior of yesterday may have little to do with the behavior of tomorrow, and implementing ideas through trading may itself change the behavior of the market.

Two, backtesting results are frequently reported using frequentist statistics. This puts another degree of separation between the results of a backtest and the truth of probable future behavior: a p-value is the probability of getting a result “as extreme or more” from a hypothetical distribution—it is not, in fact, the probability of a particular hypothesis being true given the evidence of an experiment. Ultimately, this issue can be addressed by instead using Bayesian

statistics, which work directly with the probability of a hypothesis. With the exception of a couple asides, the purpose of this discussion is not to pontificate on the virtues of Bayesian statistics, and such discussions can indeed be found elsewhere (Yudkowsky 2003, 2005). Instead, we will show that backtesting using frequentist statistics can fail in ways related or unrelated to frequentism, and we will discuss some of the ways one can spot these failures.

Adversarial Incentives

Failures of backtesting to convey truthful information are driven in part by adversarial incentives. A marketing department seeking to attract clients might prioritize constructing strategies with great historical performance without as much concern with out-of-sample returns.

It's not enough to be skeptical of external marketing material: if internal researchers are incentivized to produce strategies that look good to internal portfolio managers and traders, similar deception could infect their backtests, even unintentionally. Looking to academics for answers is also an imperfect answer, as the scientific reproducibility crisis has also become a debated concern in the world of academic finance (Hou, Xue, Zhang, 2017), (Jensen, Kelly, & Pedersen, 2023), (de Prado & Fabbozi, 2026). After all, academia has its own incentives that fail to perfectly align with unadulterated truth-seeking.

Hindsight

Any layman can lament, "I wish I had bought Bitcoin the first time I had heard about it; surely I would have sold at $100,000!" Similarly, any strategy one wants to backtest is likely to have systematic hindsight bias: an investor who has paid attention to economic history is unlikely to forget, at least subconsciously, the important factors in market movements of the past, and proposed strategies are therefore more likely than otherwise to be tailored to history, overselling their value for true forecasting.

Inflated Hypothesis Spaces: P-hacking, Overfitting

There is a large class of backtesting failures that can be grouped into a simpler umbrella category of "too many choices." In machine learning, overfitting happens when an algorithm trains "too well" on a data set, resulting in poor out-of-sample performance. One cause, though not the only, is that the machine learning architecture is sufficiently complex such that it can represent the training set in excessive detail. In these cases, merely simplifying the model can lead to improvements in out-of-sample performance (Dietterich, 1995). Note that this is an issue of relative complexity, not absolute: Deepseek-V4-Pro has well over a trillion parameters, but it has been trained on tens of trillions of tokens (Deepseek-AI, 2026). Importantly, this is not exclusively a machine learning problem, and arguably not even primarily a machine learning problem. In

financial problem domains with limited data, overparameterization can result in quite misleading results.

Mathematically, the idea here is simple. Suppose we're using p-values of some sort to evaluate strategies. Then, we're considering the probability of obtaining "as good or better returns" conditioned on the null hypothesis "the strategy has no edge," for some definition of each. If we obtain experimental returns translating to a p-value of 0.05, we are saying that 1 in 20 random strategies from some reference class ought to provide as good or better returns. If we considered just 5 random strategies—that is, independent strategies with no edge—we ought to get at least one that provides as good or better returns with probability $1 - (0.95)^5 = 0.22$. Then, it would be quite misleading to consider five configurations of a strategy, one of which succeeds with a p-value of 0.05, and promote that particular configuration as a strategy that achieves a p-value of 0.05 without its context as one of several attempts. Such a claim would be a form of P-hacking; an honest researcher would report the other configurations and, if inclined to report frequentist statistics on the results, would adjust the p-values to control for false positives, and potentially follow those adjustments downstream to Sharpe or *t*-ratios (Harvey & Liu, 2015). Note that the Bayesian approach confines these problems to the "prior probability," separate from the "likelihood ratio" derived from the data.

At LongTail Alpha, we're keenly aware of the historical bleed of a typical left tail hedge, whose value does not come from a strong CAGR, but rather from smoothing out left tail events in a larger portfolio. However, it is not particularly difficult to put together a strategy with a historically costless convex payoff profile and negative equity market beta. To construct a strategy that is costless during non-events, we will start with a premium-neutral put spread ladder, buying and selling 12 month puts on SPX Index every 3 months. Playing around with strikes reveals that very tight spreads reduce the intensity of drawdowns, and drawing the strikes close to the money also improves returns. Finally, we select just the March expiries for our strategy as they outperform the other expiries. Now, we have historical support for our engineered left-tail strategy, which repeatedly delivers convex payouts in times of crisis without bleeding premium. See Figure 1 for an apparently performant strategy.

But how many strategies did we consider here? We only want to consider tradeable out-of-the-money put spreads that have strikes 2 percent apart or more (which avoids non-trades) and the farthest we will go out-of-the-money is 40 percent. Under these conditions, choosing only odd percent strikes, there are 19 choose 2 or 171 combinations of strikes available to us. We also considered all nontrivial subsets of March, June, September, and December yearly expirations. This multiplies our choices by $2^4 - 1$ or 15, for a total of 2,565 choices. While we did not compare returns for every combination, we explored the hypothesis space sufficiently to find a compelling

return series—while there was no guarantee we could find such a strategy in this space, the excessive parameter flexibility made locating a historically strong strategy quite easy.

Figure 1: A "Marketable" Left-Tail Strategy

Unrealistic Simplifications

However, simplicity is not a universal good. Another way that backtesting can fail is by forgetting to consider real-world constraints like transaction costs and market impact at size.

A critical subcategory of such simplifications arises from not thinking carefully about technological and informational history (Challet & Ayed, 2013). For example, a trader with a Bloomberg terminal and Claude Code can relatively easily stand up a systematic intraday trading strategy that performs quite well historically. However, the trader may have failed to consider differences in access to information ("Was this data reliably available at the time of each trade?"), computation ("Could I realistically have run this algorithm in real-time using a tiny fraction of the computational power of my cell phone?"), liquidity ("Could I actually have found counterparties for the trades I'm simulating?"), among others.

Regime Shift

A truism: The global economy may change in some significant way, rendering the high-alpha strategies of last year no longer profitable. It is possible to hunt alpha with scientific rigor and still find oneself at odds with potentially unforeseen changes in market dynamics.

Avoiding the Problem

Some of these problems can have detectable markers. Adversarial incentive structures are fairly visible and quite common as many researchers have incentives not always aligned with truth-seeking. Even internal researchers might be encouraged to convince portfolio managers to allocate capital to their ideas rather than communicate clearly null results. Importantly, a strong scientific culture with appropriate incentives can fight these tendencies, and a strong peer review process can aid in filtering even external results.

While an adversarially-incentivized researcher will not tell you how many strategies they tried, or how many parameters they adjusted, or how many parameter values they considered "in-scope," it is often not difficult to get a feel for the degree of overparameterization. A useful heuristic is to think about how much information is required to explain the strategy. Larger strategy hypothesis spaces demand more information to accurately communicate the chosen strategy. For example, in perfect communication, to distinguish between 2 strategies, one needs only provide one bit of information, whereas to identify one of 2^{100} strategies in a hypothesis space, one would need 100 bits of information. As such, wariness of detail can protect an investor from these too-large hypothesis spaces. Toward this end, a strategy with more symmetry—matching or opposite behavior or numbers—requires less information in explanation than one with asymmetry ("perform X trade with March expiration and Y trade with September expiration" is harder to communicate than "perform X trade for both March and September expirations," especially as X and Y increase in complexity). Similarly, round numbers are an indicator of a less detailed strategy (9 percent out-of-the-money is more likely to be cherry-picked than 10 percent out-of-the-money). Additionally, "canonical" choices can reduce information complexity (choosing a contract based on some date scheme is a strategy more likely to be carefully chosen out of a larger pool than just using the active contract). Note that the wisdom of aversion to detail is not an absolute, as deeply complex ideas are often critical for obtaining alpha. However, complex strategies can be pulled from small sets of hypotheses using deep economic, financial, or mathematical ideas. As alluded to before, in a Bayesian setting, we can incorporate the complexity of a strategy into a "prior probability" of it having an edge, which directly factors into the posterior probability—the "final result" of a Bayesian analysis.

Evaluating the use of accurate transaction costs is relatively simple and can be done in replication if a publication isn't sufficiently detailed. Similarly, a seasoned investor ought to know which instruments have sufficient liquidity for minimal market impact with size appropriate for their portfolio.

Given that an artificially performant strategy obtains its performance from ahistorical modeling, it should disproportionately perform during the period of ahistoricity. For example, if presented

with a return series for a computationally intensive strategy with standout performance in the 1970s and 80s and decidedly mediocre returns since 2010, one can reasonably increase the probability that the returns were technically impossible.

These are certainly not the only failure modes of a backtest, nor should one dismiss the value of backtesting in predicting the future. Care is needed to ensure this tool is used responsibly for assisting in the investor's goal of forecasting market behavior.

References

Arnott, R. D., Harvey, C. R., & Markowitz, H. (2018). A backtesting protocol in the era of machine learning. *Available at SSRN 3275654*.

Challet, D., & Ayed, A. B. H. (2013). Predicting financial markets with Google Trends and not so random keywords. *arXiv preprint arXiv:1307.4643*.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, *27*(3), 326-327.

DeepSeek-AI. (2026, April 24). "DeepSeek-V4 Preview Release." DeepSeek API Docs. api-docs.deepseek.com/news/news260424.

Harvey, C. R., & Liu, Y. (2015). Backtesting. *Available at SSRN 2345489*.

Hou, K., Xue, C., & Zhang, L. (2017). *Replicating anomalies* (No. w23394). National Bureau of Economic Research.

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is there a replication crisis in finance?. *The Journal of Finance*, *78*(5), 2465-2518.

de Prado, Marcos Lopez, Fabbozi, F., April 2026, "The False Discovery Rate in Finance: Identification Failure and Search-Adjusted Estimation." *ADIA Lab Research Paper Series*, No. 24.

Yudkowsky, E. (2005). *A Technical Explanation of Technical Explanation*. https://intelligence.org/files/TechnicalExplanation.pdf.

Yudkowsky, E. (2003). *An Intuitive Explanation of Bayes' Theorem.* http://yudkowsky.net/rational/bayes.

Important Disclosures

Adam Dhillon, Ph.D. is a Research Associate at LongTail Alpha, LLC, an SEC-registered investment adviser and a CFTC-registered CTA and CPO. Any opinions or views expressed by Dr. Dhillon are solely those of Dr. Dhillon and do not necessarily reflect the opinions or views of LongTail Alpha, LLC or any of its affiliates (collectively, "LongTail Alpha"), or any associated persons of LongTail Alpha. You should not treat any opinion expressed by Dr. Dhillon as investment advice or as a recommendation to make an investment in any particular investment strategy or investment product. Dr. Dhillon's opinions and commentaries are based upon information he considers credible, but which may not constitute research by LongTail Alpha. Dr. Dhillon does not warrant the completeness or accuracy of the information upon which his opinions or commentaries are based. This publication is for illustrative and informational purposes only and does not represent an offer or solicitation with respect to the purchase or sale of any particular security, strategy or investment product. Past performance is not indicative of future results. Different types of investments involve varying degrees of risk, including possible loss of the principal amount invested. Therefore, it should not be assumed that future performance of any specific investment or investment strategy, or any non-investment related content, will be profitable or prove successful. Nothing contained herein is intended to predict the performance of any investment.