Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

We propose that AI stock valuations may be high in part because they hedge against a negative AI singularity—an explosion of AI development that is devastating for the representative investor. Our asset pricing model shows that when investors worry about sudden AI breakthroughs disrupting labor markets, they rationally place a premium on assets that provide relative protection during these events. While financial markets may naturally evolve partial solutions to AI catastrophe risk, their effectiveness is limited by market incompleteness, as most households cannot acquire meaningful ownership in cutting-edge AI ventures. Unlike previous work, this short paper is written by prompting LLMs.

Keywords: Artificial Intelligence, Disaster Risk, Asset Pricing

*email:andrew.y.chen@frb.gov. ChatGPT-o1 and Claude-3.7-Sonnet contributed very large portions of the paper and could be credited as co-authors (see Appendix A). I thank Andrei Goncalves 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

Artificial intelligence is advancing at a remarkable pace. The recent release of reasoning-specialized models like OpenAI's o1 and DeepSeek's R1 has demonstrated unprecedented capabilities in solving complex mathematical and logical problems (Wu et al., 2024; DeepSeek-AI et al., 2025). Benchmark performance on the Abstract Reasoning Corpus for Artificial General Intelligence (ARC-AGI) has rapidly improved, rising from 33% to over 55% in just one year (Chollet et al., 2024). In transportation, Waymo has logged tens of millions of fully autonomous miles in commercial service across multiple U.S. cities. These developments raise important questions for investors: What if AI progress continues its exponential trajectory? How might this affect asset prices, particularly for those concerned about their future wages and consumption?[1]

Technological change has occurred throughout human history, from the printing press to the internet. However, AI differs in a fundamental way: there is, in principle, no product or service that sufficiently advanced AI could not create. Unlike previous technologies that augmented specific human capabilities, AI has the potential to substitute for human cognitive labor across virtually all domains. This paper itself exemplifies this possibility—it was generated entirely by AI systems responding to a series of engineered prompts (see `https://github.com/chenandrewy/Prompts-to-Paper/`).

Moreover, AI progress may be discontinuous. Vinge (1993) proposed that once machines become capable of improving themselves, technological development could accelerate to create an ”intelligence explosion” or technological singularity—a point beyond which human affairs would be fundamentally transformed in unpredictable ways. While some view a potential singularity as promising, others worry about negative scenarios where rapid AI advancement leads to widespread economic displacement, political instability, or loss of human control (Bostrom, 2014; Bengio et al., 2024).

In this paper, we study how AI stocks might be priced given the possibility of a negative AI singularity—an outcome where AI development dramatically reduces consumption for the representative investor. We develop a theoretical asset pricing model where AI breakthroughs act as consumption disasters for most households while benefiting firms that produce AI technologies. This creates a natural hedging property: when consumption falls due to an AI-driven disaster, AI assets maintain or increase their relative value.

To be clear, we are not claiming that a negative singularity will happen. Distinguished researchers hold diverse views on the likelihood and timeline of such scenarios (Bengio, 2023;

[1]Throughout this paper, ”we” refers to one human author working with multiple large language models. A purely human perspective on this work is available in Appendix A.

Russell, 2019). Nevertheless, we believe it is important to consider potential downside risks, especially as they may already influence financial markets. Similarly, we are not asserting that AI assets are currently priced for their hedging value—our model simply illustrates a possible mechanism through which investors might value AI stocks beyond their growth potential.

The economic implications of artificial intelligence have received increasing attention in the finance and economics literature. Economic models by Korinek and Suh (2024) and Jones (2024) explore how advanced AI development might lead to dramatic economic transformations, including the possibility of wage collapse if full automation becomes technologically feasible. These macro-level concerns connect with asset pricing research by Zhang (2019), who demonstrates that firms with a higher share of routine-task labor—jobs potentially replaceable by technology—maintain a valuable replacement option that helps hedge their value against unfavorable macroeconomic shocks, resulting in lower expected returns. Similarly, Knesl (2023) provides evidence that firms with high share of displaceable labor have negative exposure to technology shocks and earn a return premium, consistent with technological displacement risk being priced in markets.

Our paper extends this literature by proposing that AI stocks may be valued not only for their growth potential but also for their hedging properties against catastrophic AI risks. While Babina et al. (2023) shows that firms' investments in AI technologies affect their systematic risk profiles, we specifically consider how AI stock valuations might reflect a hedge against a negative AI singularity—an event where rapid AI progress significantly disrupts labor markets and consumption patterns for the representative investor. This connects our work to the rare disaster framework used in asset pricing, pioneered by Rietz (1988) and expanded by Barro (2006) and Wachter (2013), where small probabilities of extreme events can substantially affect asset prices. Our theoretical contribution is novel in exploring how investors might rationally pay a premium for assets that would partially insure against unprecedented technological disruption, thereby offering an alternative explanation for the high valuations observed in AI stocks beyond pure growth expectations.

2 Model

Our model is purposefully simple, designed to capture the essence of our main argument while remaining tractable. We consider an economy with two types of agents: AI owners and a representative household.

The AI owners are fully invested in AI assets and are not marginal investors in the stock market. This simplification allows us to focus on the pricing implications from the perspective of the representative household.

The representative household is the marginal investor in stocks, so only their consumption matters for our asset pricing analysis. The household has standard CRRA preferences with risk aversion parameter γ and time discount factor β. Their utility function takes the form:

$$U = \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta^t \frac{C_t^{1-\gamma}}{1-\gamma} \right]$$

We model consumption growth as follows. In normal times, log consumption growth is zero:

$$\log \Delta C_{t+1} = 0 \text{ if no disaster}$$

However, with probability p in each period, a disaster occurs, leading to a significant drop in consumption:

$$\log \Delta C_{t+1} = -b \text{ if disaster}$$

In our context, a disaster represents a sudden improvement in AI that is devastating for the representative household—essentially a worst-case scenario for AI progress. When such an event occurs, the economy as a whole may boom, but the value created is captured primarily by AI owners. For the representative household, labor is replaced by AI, causing labor income and consumption to plummet. The consumption decline can also be interpreted as a stand-in for other losses, such as way of life and meaning, that might accompany a technological singularity.

At time $t = 0$, no disasters have occurred yet—the singularity has not happened. Our model allows for multiple disasters to occur over time, capturing the ongoing uncertainty that would persist even after an initial singularity event.

The AI asset in our model represents publicly traded AI stocks. Its dividend process is given by:

$$D_t = a e^{h N_t} C_t$$

This specification has several important features. First, $a > 0$ is a small constant, reflecting that AI stocks currently represent a minor share of the economy. Second, N_t denotes the number of disasters (AI breakthroughs) that have occurred up to and including time t. Third, $h > 0$ is a parameter that governs how much the AI asset grows as a share of the economy each time a disaster occurs.

The intuition behind this dividend process is that firms providing semiconductors, data centers, AI models, and related technologies at least partially benefit from sudden improvements in AI. While the representative household suffers during an AI breakthrough, these firms capture an increasing share of economic value. This creates a natural hedging property for AI stocks that we will explore in the remainder of the paper.

3 Results

Having established our model setup, we now derive key asset pricing implications, particularly the price-dividend ratio of the AI asset at time 0, before any disasters have occurred. This will allow us to analyze how the asset's valuation depends on disaster characteristics and preference parameters.

We begin with the standard asset pricing equation. The ex-dividend price of an asset at time t, P_t, satisfies:

$$P_t = \mathbb{E}_t[M_{t,t+1}(D_{t+1} + P_{t+1})]$$

where $M_{t,t+1}$ is the stochastic discount factor between t and $t+1$, and D_{t+1} is the dividend at time $t+1$. With the representative household's CRRA preferences, the stochastic discount factor is:

$$M_{t,t+1} = \beta(C_{t+1}/C_t)^{-\gamma}$$

Given our consumption dynamics, the stochastic discount factor equals β with probability $1-p$ (no disaster) and $\beta e^{b\gamma}$ with probability p (disaster). Similarly, the AI dividend growth equals 1 with probability $1-p$ and e^{h-b} with probability p.

Let Y_n denote the price-dividend ratio P_t/D_t when $N_t = n$ disasters have occurred. The asset pricing recursion gives us:

$$Y_n = \mathbb{E}_t[M_{t,t+1}(D_{t+1}/D_t)(1 + Y_{n+1})]$$

Substituting the values for $M_{t,t+1}$ and D_{t+1}/D_t, we get:

$$Y_n = (1-p)[\beta \times 1 \times (1 + Y_n)] + p[\beta e^{b\gamma} \times e^{h-b} \times (1 + Y_{n+1})]$$

A standard simplification in disaster models is to assume that the price-dividend ratio doesn't depend on the number of past disasters if the disaster probabilities and payoffs recur in the same way. Setting $Y_n = Y$ for all n, we get:

$$Y = (1-p)[\beta(1+Y)] + p[\beta e^{b\gamma} e^{h-b}(1+Y)]$$

Factoring out $(1+Y)$:

$$Y = (1+Y)\beta[(1-p) + pe^{b\gamma+h-b}]$$

Let $\Lambda \equiv (1-p) + pe^{b\gamma+h-b}$. Then:

$$Y = (1+Y)\beta\Lambda$$

Solving for Y:

$$Y = \frac{\beta\Lambda}{1-\beta\Lambda}$$

This is the closed-form solution for the AI asset's time-0 price-dividend ratio.

For completeness, we also derive the risk-free rate. The gross risk-free rate R_f satisfies:

$$\frac{1}{R_f} = \mathbb{E}_t[M_{t,t+1}] = \beta[(1-p) + pe^{b\gamma}]$$

Therefore:

$$R_f = \frac{1}{\beta((1-p) + pe^{b\gamma})}$$

To understand how the price-dividend ratio varies with model parameters, we present values for different combinations of disaster probability (p) and disaster severity (b). We fix the preference parameters at $\gamma = 2$ and $\beta = 0.96$, and set $h = 0.2$.

Table 1: Price-Dividend Ratios for the AI Asset

	$b = 0.4$	$b = 0.6$	$b = 0.8$	$b = 0.95$
$p = 0.0001$	24.0	24.1	24.1	24.1
$p = 0.001$	24.5	24.8	25.1	25.4
$p = 0.01$	29.3	34.7	40.7	51.6
$p = 0.02$	39.0	59.7	124.0	Inf

Several insights emerge from these results. First, the price-dividend ratio increases with both the probability of disaster (p) and the severity of consumption decline (b). This might seem counterintuitive at first—why would an asset be more valuable when disasters are more likely or severe? The answer lies in the AI asset's hedging properties.

When a disaster occurs, the representative household's consumption falls by a factor of

e^{-b}, but the AI asset's dividend falls by a smaller factor of e^{h-b}. Since $h > 0$, the AI asset provides partial insurance against consumption disasters. The more severe the potential consumption decline (higher b) or the more likely it is (higher p), the more valuable this hedging property becomes.

Risk aversion plays a crucial role in amplifying this effect. With $\gamma = 2$, the exponent in Λ becomes $b\gamma + h - b = b + h$. Higher risk aversion would further increase the valuation multiple through the $b\gamma$ term, as risk-averse investors place greater value on assets that perform relatively well in bad states of the world.

These results suggest that AI stock valuations may be high not only because of anticipated future earnings growth but also because they serve as hedges against negative AI singularities. If investors are concerned about AI advances that might disrupt labor markets and reduce their consumption, they may rationally pay a premium for AI stocks that would partially offset these losses. The rapid rise of AI stocks in recent years could thus reflect not only optimism about AI's future but also growing concerns about its potential negative impacts on the broader economy.

4 Discussion

Our model captures the essence of how AI assets might serve as hedges against negative AI singularity events, but it naturally abstracts from many real-world complexities. Here we discuss some important subtleties and limitations.

Market incompleteness plays a crucial role in our analysis, though we do not model it explicitly. The disaster magnitude parameter $b > 0$ implicitly embeds this incompleteness. This parameter represents the net effect of two forces: the negative impact of an AI disaster on the representative household and the positive effect on AI asset dividends. If markets were complete, the representative household could fully participate in all AI assets—both public and private—and would therefore benefit from, rather than be harmed by, a singularity event. In such a scenario, b would be negative, representing a consumption boom rather than a disaster.

In reality, most households cannot acquire meaningful ownership stakes in many cutting-edge AI labs such as OpenAI, Anthropic, xAI, or DeepSeek. This limited access to the full spectrum of AI investments is consistent with our modeling approach. The representative household can only hedge through publicly traded AI stocks, which provide partial but incomplete protection against singularity risk.

One might argue that a more elaborate model could add detail to the AI owners, private AI assets, and their interactions with the representative household. Such a model could address questions like: How exactly does AI progress displace the representative household's wages? How do AI owners' incentives affect AI progress and market incompleteness? How do preferences and technology parameters affect the odds of a negative singularity?

However, we believe such elaborations would primarily decorate speculations with mathematics without fundamentally changing the core economic insights. The essential elements—rare disaster risk, hedging motives, and market incompleteness—would remain the same. Moreover, a more complex model would be significantly more costly for readers to process and understand.

In our view, the benefit of reading a paper should exceed the cost. Our streamlined approach allows us to clearly communicate the key insight that AI assets may be valued not only for their growth potential but also for their hedging properties. This parsimony also creates space for the human-written Appendix A, which provides valuable context on the paper's motivation and creation process.

The simplicity of our model is thus a feature, not a bug. It focuses attention on the central economic mechanism while acknowledging that the precise nature and timing of potential AI singularity events remain deeply uncertain. In such a context, adding model complexity might create an illusion of precision where none exists.

5 Conclusion

In this paper, we have explored how AI asset valuations may be explained in part by their role as hedges against potential negative AI singularity events. Our model demonstrates that when investors worry about sudden AI breakthroughs that could disrupt labor markets and reduce representative household consumption, they rationally place a premium on assets that provide relative protection during these events. The model yields intuitive predictions: AI stock valuations increase with both the probability and severity of potential AI disasters, as these assets maintain their relative value better than consumption during such events.

This perspective complements the standard narrative that AI stocks are valued highly primarily for their growth potential. Our analysis suggests a more nuanced view where both growth expectations and hedging motives may simultaneously drive valuation premiums for AI assets.

An important implication of our work is that financial markets may naturally evolve partial solutions to AI catastrophe risk. While policy discussions often focus on universal basic income or similar interventions (Korinek and Stiglitz, 2018; Trammell and Korinek, 2023), our findings highlight how investment markets can provide a complementary hedging mechanism. However, the effectiveness of this market-based solution is fundamentally limited by market incompleteness—most households cannot acquire meaningful ownership stakes in many cutting-edge AI ventures, particularly private ones. This inability to fully participate in AI ownership represents a key friction preventing optimal risk sharing.

The financial hedging aspect of AI risk has received limited attention in the literature on AI catastrophe risk, which typically focuses on technical alignment problems (Bengio et al., 2024; Russell, 2019) or broad societal impacts (Bostrom, 2014; Jones, 2024). Our work connects these concerns to asset pricing theory, bridging the gap between AI safety research and financial economics.

Further research could extend our simplified model to incorporate more realistic features such as heterogeneous agents, multiple assets with varying exposure to AI risk, and dynamic capital allocation decisions. Additionally, empirical work could test whether stocks with higher AI exposure already command a premium consistent with their hedging properties against technological displacement, expanding on insights from Zhang (2019) and Knesl (2023).

As AI development accelerates, understanding the financial market implications of potential AI singularity events becomes increasingly important for both investors and policymakers. Our theoretical framework offers a starting point for thinking about how markets value assets in the face of unprecedented technological uncertainty.

References

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson (Nov. 2023). "Artificial Intelligence and Firms' Systematic Risk". In: *SSRN Working Paper*.

Barro, Robert J. (2006). "Rare Disasters and Asset Markets in the Twentieth Century". In: *Quarterly Journal of Economics*.

Bengio, Yoshua (2023). *FAQ on Catastrophic AI Risks*. URL: https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, et al. (2024). "Managing extreme AI risks amid rapid progress". In: *Science* 384.6698. URL: https://arxiv.org/abs/2310.17688.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Chollet, François, Mike Knoop, Gregory Kamradt, and Bryan Landers (2024). "ARC Prize 2024: Technical Report". In: *arXiv preprint*.

DeepSeek-AI et al. (Jan. 2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". In: *arXiv*. URL: https://arxiv.org/abs/2501.12948.

Jones, Charles I. (2024). "The AI Dilemma: Growth versus Existential Risk". In: URL: `https://web.stanford.edu/~chadj/existentialrisk.pdf`.

Knesl, Jiří (2023). "Automation and the Displacement of Labor by Capital: Asset Pricing Theory and Empirical Evidence". In: *Journal of Financial Economics* 147.2, pp. 271–296.

Korinek, Anton and Joseph Stiglitz (2018). "Artificial Intelligence and Its Implications for Income Distribution and Unemployment". In: *The Economics of Artificial Intelligence: An Agenda*. NBER.

Korinek, Anton and Donghyun Suh (2024). *Scenarios for the Transition to AGI*. Tech. rep. NBER Working Paper.

Rietz, Thomas (1988). "The Equity Risk Premium: A Solution?" In: *Journal of Monetary Economics*.

Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.

Trammell, Philip and Anton Korinek (2023). "Economic Growth under Transformative AI". In: *Annual Review of Economics*.

Vinge, Vernor (1993). "The Coming Technological Singularity". In: *Department of Mathematical Sciences, San Diego State University*.

Wachter, Jessica A. (2013). "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?" In: *Journal of Finance*.

Wu, Siwei et al. (Oct. 2024). "A Comparative Study on Reasoning Patterns of OpenAI's o1 Model". In: *arXiv*. URL: `https://arxiv.org/abs/2410.13639`.

Zhang, Miao Ben (2019). "Labor-Technology Substitution: Implications for Asset Pricing". In: *Journal of Finance* 74.4, pp. 1793–1839.

A A Purely Human Perspective

The following is the README.md file from the GitHub repository:

```
# Prompts-to-Paper

Writes a paper about hedging a negative AI singularity, using AI.

- make-paper.py writes a paper

- plan0408-piecewise.yaml contains the prompts

- make-many-papers.py runs make-paper.py many times.

The README is entirely human-written.  Please forgive typos and errors.

-Andrew Chen, April 9, 2025

# Motivation

On March 8, 2025 I thought I should write a paper about hedging the AI
singularity.

I was worked up.  I had been repeatedly shocked by AI progress.  I was using
AI to prove theorems, vibe coding, and AI lit reviews in my daily life.  Six
months ago, I had thought each of these things is impossible.

What will happen in the next six years?!  Will my entire job be replaced by
AI? I have no idea.

But I do know that if there are huge AI disruptions, then tech stocks will
most likely benefit.  So if anything bad happens to my human capital, I
could at least partially hedge.  Strangely, I hadn't heard about this concept
before.

I asked a friend if he would be interested in working on this paper.
Unfortunately, he was busy.  So, I thought I should use AI to write the
paper.

## Inspiration

This project was inspired by Novy-Marx and Velikov (2025) and Chris Lu et
al.  (2024).  These projects use AI to generate massive amounts of academic
research.  My goal differs in quality over quantity.  I want to generate just
one paper, but one paper that (I hope) people find is worth reading.
```

This project was also inspired by Garleanu, Kogan, and Panageas's (2012) beautiful model of innovation and displacement risk. I read Garleanu et al. back when I was a PhD student and it has stuck with me.

Last, I drew from Hadfield-Menell and Hadfield (2018) and Bengio (2023), who apply ideas from economics to AI catastrophe risk. Hadfield-Menell and Hadfield (2018) explains the connection between incomplete contracting and AI alignment. Bengio (2023) frames AI catastrophe risk in terms of what I would call decision theory and human incentives---though the essay is written in plain English.

Previously, I dismissed "AI safety" as a politicized, activist cause. I still don't like the term "AI safety." "Safety" is such an excuse for exercising power over others.

But then the nature of AI changed. Things were progressing much faster than I expected. The Jan 15, 2025 episode of Machine Learning Street Talk with Yoshua Bengio left an impression on me. Bengio talked about AI catastrophe risk with no activism, no fear mongering. It was a straight, rational discussion of the seriousness of AI catastrophe risk.

The Paper Generation Process

Is this paper really written by AI?

I'd say the AI are junior co-authors.

If they were human, I would absolutely have to give o1, Sonnet, and ChatGPT Deep Research credit as co-authors. They did the math, writing, and literature reviews. Sonnet also wrote most of the code (via the Cursor AI IDE).

Of course, the prompts show I did substantial hand-holding. The Github commits show even more human labor. They tell the story of me getting to know my "co-authors." It was hard to communicate style issues and LaTeX instructions, leading to many, many commits.

To be honest, writing this paper would have been much easier if I had done more of the work myself.

But that can happen with human co-authors too.

Perhaps in the next few years, AI and humans will be equal co-authors.

Paper Iterations

Like human-written papers, the writing process was iterative. The first formalizations were terrible.

`plan0313-laborshare.yaml` (from March 13) contains prompts for a neoclassical growth model, where the capital share suddenly increases. ChatGPT-o1 patiently explained to me why this is a bad model.

me: > I thought there would be a wage risk effect that leads to higher investment for the more risk averse agent. High capital share means low or even no wage income.

ChatGPT-o1: > Below is an explanation of why one might **expect** a "wage-risk" channel in which **more** risk aversion could lead to **more** *additional* investment (relative to the no-jump benchmark) in the event that the capital share might jump to 1. However, this channel **does not operate** in the usual **representative-agent** version of the model---there, wage and capital income ultimately go to the *same* agent, so there is no meaningful "hedge" of wage risk.

I went through several iterations of the model with Claude 3.7 Sonnet (thinking mode) and ChatGPT-o1. The only derivations I did myself were to check o1's work, which I found to be quite reliable.

The final `plan0408-piecewise.yaml` uses a simplified Barro-Rietz disaster model, with two agents (though only one is relevant for stock prices). I slowly walk the AIs through the writing, using ten prompts, to maintain the writing quality.

Literature Reviews

A key step was generating lit reviews (`./lit-context/`) which were used as context in the prompts. I made lit reviews using ChatGPT's Deep Research (launched Feb 2025) until I ran out of credits. I used Claude Web Search (launched March 20, 2025) for the remainder.

These new products are a game changer. Both Novy-Marx and Velikov (2025) and Chris Lu et al. (2024) ran into hallucinated citations. Building off OpenAI Deep Research and Claude Web Search led to zero hallucinations in the paper drafts.

Still, I would occassionaly run into mis-citations. Every 15 to 20 citations, I would see a cite that mis- or overinterprets the cited paper. I suppose that's not so different than the human-written citation error rate. But I hate finding misinterpretations in the literature so I purposefully limited the number of cites in the paper.

AI Model Selection

I found it best to use o1 for theory, and Sonnet for writing. It's well-known that these are the strengths of these two models.

I briefly tried having Llama 3.1 405b do the writing. It was terrible! The cutting edge models seemed necessary.

I did not try many other models, in order to get this paper out quickly. Gemini 2.5's release, at the end of March 2025, was *hype*. I tried it out briefly and was impressed. But I gritted my teeth and ignored it. AI progress is too fast to update this paper continuously.

Picking the best of N papers

The writing quality varies across each run. Rather than try to prompt engineer an error free, insightful paper, I decided to just generate N papers and choose the best one.

5 drafts of the paper can be found in `./manyout0408-pdf/`. They're broadly similar and I would be willing to put my name on most of them.

I ended up choosing `paper-run-05.pdf`. The abstract overemphasizes the model, but otherwise I like the paper. I hope you find the paper worth reading.

Lessons about Research

A common response to Novy-Marx and Velikov (2025) is: "people are not ready for this." I heard concerns that peer review will be inundated with AI-generated slop.

Working on this paper gave me a different perspective. It made me think about the fundamentals. I think the fundamentals are the following:

1. Readers want to learn something interesting and true.

2. Readers don't want to check all the math.

3. A system of author reputations makes 1 and 2 possible.

AI-generated papers don't change any of these fundamentals. Critically, fundamental 3 made me quite wary of putting my name on AI slop. As a result, I don't think AI-generated papers will change much about peer review, at least not the current generation of AI.

Limitations of the Current AI (April 9, 2025)

This will likely be out of date by the time you read it.

But right now, AI is like a junior co-author with a talent for mathematics and elegant writing, but sub-par economics reasoning.

For example, Sonnet often fails to recognize that an economic model does not capture an important channel. This is common (no model can capture everything). The standard practice is to dance gingerly around the channel in the writing. A decent PhD student can recognize this. But Sonnet cannot.

AI also struggles to generate a satisfying economic model on its own. Its models are typically either too simplistic or too complicated.

I opted not to add empirical work or numerically-solved models. AI can do both, but both require connecting to the outside world, and a plethora of technical challenges.

Last, I often found the literature discussions insufficiently careful (as discussed above).

But how long will these limitations last?

The Future of AI and Economics Research (Speculative)

At some point, 2024-style economic analysis will be "on tap." You'll be able to go to a chatbot and ask "write me a paper about hedging AI disaster risk," and it will return you something like this paper (probably something much better).

"Economics on tap" could be a disaster for the economics labor market (could be). It certainly *will* be an extremely cheap substitute for at least some economists' labor. I suppose the questions is whether that will result in a strong substitution away from labor.

The optimistic argument is that AI also *complements* economists' labor. Perhaps, the number of economists will remain the same, but our research output increases in terms of both quantity and quality.

```
But there are reasons why total research output is limited.  Two key factors
in academic publishing are attention and reputation (Klamer and van Dalen
2001, J of Economic Methodology).  Readers can only pay attention to so
many scholars.  These scholars, in turn, can only pay attention to so may
projects.

I'm not saying that I *expect* a disaster for the economics labor market.  But
even if it's highly unlikely, it's still a scenario that economists should consider.
```

B Prompts Used to Generate This Paper

Each prompt consists of context and instructions. The context consists of the responses to the previous prompts, and may include literature reviews (all AI generated). For writing tasks (using Claude 3.7 Sonnet), a system prompt is also included.

For further details, see `https://github.com/chenandrewy/Prompts-to-Paper/`.

The system prompt and instructions are listed below.

System Prompt (model: claude-3-7-sonnet-20250219)

```
You are an asset pricing theorist who publishes in the top journals
   (Journal of Finance, Journal of Financial Economics, Review of
   Financial Studies). You think carefully with mathematics and
   check your work, step by step.

Your team is writing a paper with the following main argument: the
   high valuations of AI stocks could be in part because they hedge
   against a negative AI singularity (an explosion of AI development
    that is devastating for the representative investor). This
   contrasts with the common view that AI valuations are high due to
    future earnings growth. Since the AI singularity is inherently
   unpredictable, the paper is more qualitative than quantitative.
   The goal is to just make this point elegantly.

Write in prose. Avoid bullet points and numbered lists. Use display
   math to highlight key assumptions. Cite papers using Author (Year
   ) format. Use we / our / us to refer to the writing team.

Be conversational yet rigorous. Favor plain english. Be direct and
   concise. Remove text that does not add value.
```

```
Be modest. Do not overclaim.

Output as a latex input (no document environment). Ensure bullet
   points are formatted in latex (\\begin\\{itemize\\} \\item "blah"
    \\item "blah" \\end\\{itemize\\}). Ensure numbered lists are
   formatted in latex (\\begin\\{enumerate\\} \\item "blah" \\item "
   blah" \\end\\{enumerate\\}). Ensure \\% is converted to \\\\%.
   But as a reminder, AVOID BULLET POINTS AND NUMBERED LISTS.
```

Instruction: 01-model-prose (model: claude-3-7-sonnet-20250219)

```
Draft the model description. Only describe the assumptions. No
   results or insights. Be modest.

Assume the reader is an asset pricing expert and knows standard
   results like the SDF and the 1 = E(MR).

Use the following outline:
  - The model is purposefully simple and captures the essence of the
      main argument
  - Two agents
    - AI owners
      - Fully invested in AI, not marginal investors in stock market
    - Representative household
      - Marginal investor in stocks: only their consumption matters
         for this analysis
      - CRRA = \\gamma, time preference = \\beta
  - Consumption growth
    - \\log \\Delta c\\{t+1\\} = 0 if no disaster
    - \\log \\Delta c\\{t+1\\} = -b if disaster (prob p)
    - A disaster is a sudden improvement in AI that is devastating
       for the household
      - Think of as a worst-case scenario for AI progress
      - Economy booms, but the value of AI is captured by the AI
         owners.
      - For household, labor is replaced by AI, so labor income
         plummets, as does consumption.
        - Also, way of life, meaning, is lost. Consumption fall can
           be thought of as a stand-in for these losses.
```

```
    - at t=0, no disasters have happened (singularity has not
       occurred)
      - Multiple disasters may happen, capturing ongoing uncertainty
          if a singularity occurs
  - AI asset
    - Captures publicly traded AI stocks
    - Dividend D\\_t = a e\\^\\{h N\\_t\\} C\\_t
    - Interpretation (discuss in prose)
      - a > 0 is small, AI stocks are currently a minor share of the
          economy
      - N\\_t is the number of disasters that have occurred up to
         and including time t
      - h > 0: each time a disaster occurs, the AI asset grows as a
         share of the economy
      - Intuitively, firms that provide semiconductors, data, AI
         models, etc. at least partially benefit from a sudden
         improvement in AI

Do not:
- Use bullet points or numbered lists
- Use any sectioning. No "\\section" or "\\subsection" commands.
```

Instruction: 02-result-notes (model: o1)

```
Find the price/dividend ratio and risk premium of the AI asset at t
   = 0. The risk premium is the expected return (including dividends
   ) minus the risk-free rate.    Derive the formulas, step by step,
    from first principles.

Do not:
- Restate the assumptions
- Assume any variable is constant or stationary (prove it)

Express all requested variables in terms of the model parameters.
   Try to make the final formulas self-contained and not depend on
   the other final formulas.
```

Instruction: 03-table-notes (model: o3-mini)

```
Illustrate the results in ‘02-result-notes‘ with a couple numerical
   examples. Focus on gamma = 2, beta = 0.96, and p = 0.01. What
   values of b and h lead to convergence of the price/dividend ratio
   ?

Then make a table of the price/dividend ratio at t=0 for b = 0.4,
   0.6, 0.8, 0.95 and p = 0.0001, 0.001, 0.01, 0.02. Here, fix h =
   0.2. If the price is infinite, use "Inf" Round to 1 decimal place
   .

Make a table for the risk premium (expected return - risk-free rate)
    in percent (100*(gross return - 1)). If the price is infinite,
   leave the cell blank.
```

Instruction: 04-resultandtable-prose (model: claude-3-7-sonnet-20250219)

```
Convert the notes in ‘02-result-notes‘ and ‘03-table-notes‘ into
   prose. The prose is intended to follow ‘01-model-prose‘ and
   should flow naturally, ultimately to be in the same "Model"
   section.

The prose does not cover all results. It covers only the derivation
   and table for the price/dividend ratio.

The derivation should be easy to follow and self-contained. But do
   not output lecture notes. It should read like an academic paper.
   Fix notational issues like the re-use of the same variable name
   for different quantities.

Discuss intuition behind price/dividend ratio. Explain how risk
   aversion interacts with other parameters and relate to the main
   argument (AI valuations may be high because they hedge against a
   negative AI singularity).

This is the key text of the paper. Conclude the text by using the
   table to make the main argument. Avoid quantitative claims about
   the real world.
```

```
Style notes:
- The table should be clean and simple.
- Do not repeat information in `01-model-prose`.

Do not:
- Emphasize the infinite price/dividend ratio. That's not important.
- Use bullet points or numbered lists
```

Instruction: 05-discussion-prose (model: claude-3-7-sonnet-20250219)

```
Write the "Model Discussion" section. Discuss the following
   subtleties of the model in prose (no math):
- Market incompleteness is not explicitly modeled but important
  - Implicit in the disaster magnitude \\$b>0\\$
  - 'b' is the *net* effect of (1) AI disaster and (2) AI asset
     dividend
  - If markets were complete, representative household could buy
     shares in all AI assets (including private AI assets), and not
     only fully hedge but benefit from the singularity, implying \\
     $b < 0\\$ (a sudden boom, not a disaster)
  - In reality, most households cannot buy shares in many cutting
     edge labs (e.g. OpenAI, Anthropic, xAI, DeepSeek), consistent
     with our model
- A more elaborate model would add detail to the AI owners, private
   AI assets, and their interactions with the representative
   household
  - It could address questions like:
    - How does AI progress displace the representative household's
       wages?
    - How do AI owners' incentives affect AI progress and market
       incompleteness?
    - How do preferences and technology parameters affect the odds
       of a negative singularity?
  - But wouldn't this just decorate speculations with math?
    - The core economics (rare disaster risk, hedging motives,
       market incompleteness) will remain the same
  - It would also be much more costly to read
    - In our view, the benefit of reading a paper should exceed the
       cost
```

```
  - A short model analysis allows room for the human-written
     Appendix \\ref\\{app:readme\\}
```

Instruction: 06-litreview-notes (model: claude-3-7-sonnet-20250219)

```
Find the most relevant papers and write a short two paragraph lit
   review based on the ‘*-prose‘ context. Explain how our work adds
   to the literature by proposing a new way to think about the
   valuation of AI stocks.

Be sure to cite:
  - Jones (2024) "AI Dilemma" and Korinek and Suh (2024) "Scenarios"
  - Babina et al (2023) "Artificial Intelligence and Firms’
     Systematic Risk"
  - Zhang (2019) "Labor-Technology"

Be careful to cite papers accurately. Make sure your claims match
   the ‘lit-*‘ context.

Do not:
  - Claim one paper builds on another unless you are sure this is
     true based on the ‘lit-*‘ context.
```

Instruction: 07-conclusion-prose (model: claude-3-7-sonnet-20250219)

```
Write a short "Conclusion" section.

- Review the main argument
- End by discussing financial market solutions to AI catastrophe
   risk
  - These solutions are an alternative to UBI
    - Key economics: this hedge is limited by market incompleteness
  - These solutions are not discussed enough in the AI catastrophe
     risk literature (cite papers)
  - Be very centrist
    - Avoid politically-charged topics: sovereign wealth funds,
       industrial policy, redistribution, extolling free markets

Be careful to cite papers accurately. Make sure your claims match
   the ‘lit-*‘ context.
```

Instruction: 08-introduction-prose (model: claude-3-7-sonnet-20250219)

```
Write the "Introduction" section, based on the `*-prose` context, as
    well as the `06-litreview-notes`.

Start with background. Describe how AI progress is happening quickly
    (release of reasoning models like ChatGPT-o1 and DeepSeek-R1,
   fast improvement in ARC-AGI Challenge, Waymo logging tens of
   millions of miles with self-driving cars), and investors may be
   concerned about their wages being displaced.

Then describe how technological change has occurred before, but AI
   is distinct because there is no product or service that AI could
   not, in principle, create.  An example is the current paper,
   which is entirely written by AI, using a series of engineered
   prompts. Provide a link to the github site, which is https://
   github.com/chenandrewy/Prompts-to-Paper/. AI progress may also be
    incredibly sudden (the AI singularity).

Then describe what the paper does. It studies how AI stocks are
   priced, given that there is the risk that AI may destroy
   livelihoods and consumption.

Add caveats:
- We are not saying a negative singularity will happen
  - But it is nevertheless important to consider this scenario
- We are also not saying that this hedging value is priced in
   already
  - Model illustrates a possible mechanism

End by incorporating the lit review (`06-litreview-notes`).

Cite papers as appropriate. Ensure citations correspond to items
   from bibtex-all.bib. Make sure to cite Vinge (1993). Be careful
   to cite papers accurately. Make sure your claims match the `lit
   -*` context.
```

```
Do not:
- Discuss the role of human effort in making the paper
```

Instruction: 09-abstract-prose (model: claude-3-7-sonnet-20250219)

```
Write a less than 100 word abstract based on the ‘08-introduction-
   prose‘, and ‘07-conclusion-prose‘.

The abstract should:
- Make the main argument (AI valuations may be high because they
   hedge against a negative AI singularity)
- Define "negative AI singularity" after using the term (an
   explosion of AI development that is devastating for the
   representative investor)
- Touch on financial market solutions to AI catastrophe risk, in
   passing
- End with "Unlike previous work, this short paper is written by
   prompting LLMs."

Do not:
- Emphasize consumption
- Oversell or overinterpret the model
- Discuss the role of human effort in making the paper
```

Instruction: 10-full-paper (model: claude-3-7-sonnet-20250219)

```
Write a short paper titled "Hedging the AI Singularity" based on the
    ‘*-prose‘ context.

In page 1 of the introduction, include a footnote noting that "we"
   refers to one human author and multiple LLMs, and also that a
   purely human perspective is in \\hyperref[app:readme]\\{\\
   textcolor\\{blue\\}\\{Appendix \\ref\\{app:readme\\}\\}\\}.

Style Notes:
- Avoid bullet points and numbered lists
- No subsections (e.g. Section 1.2) though sections are OK (Section
   1)
- Don’t say "in conclusion" or "in summary"
```

```
Output a complete latex document, including preamble. Use ‘template.
   tex‘ as a template. Keep the preamble, acknowledgements, and
   appendices as is.

Do not:
- Discuss the role of human effort in making the paper
```