Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

We propose that high AI stock valuations may partly reflect their role as hedges against a negative AI singularity—a scenario where rapid AI advancement significantly harms the representative investor. Our model demonstrates that even with a small probability of such an event, AI assets that would retain or increase their economic share during a singularity would command a premium. While financial markets might offer partial hedging solutions, their effectiveness is fundamentally limited by market incompleteness. This perspective enriches our understanding of AI valuations beyond conventional growth narratives. This short paper was generated entirely through prompting large language models, as documented in our GitHub repository.

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 has advanced at a breathtaking pace in recent years. DeepSeek's R1 model has demonstrated remarkable reasoning abilities comparable to OpenAI's o1 model (DeepSeek-AI et al., 2025). Google's Gemini models have shown increasingly sophisticated planning capabilities (DeepMind, 2025). The Artificial Intelligence Research Challenge (ARC) prize competition has spurred significant progress in core reasoning benchmarks (Chollet et al., 2024). Meanwhile, Waymo and Cruise have deployed autonomous vehicles that operate without human oversight in major U.S. cities. As these technologies continue to advance, investors and workers alike increasingly worry about AI's potential to displace human labor on an unprecedented scale.

Technological disruption is hardly new. From the Industrial Revolution to the advent of computers, innovations have repeatedly transformed labor markets and economic structures. However, AI differs fundamentally from previous technological shifts. While earlier innovations automated specific physical or computational tasks, AI aims to replicate and eventually surpass human intelligence itself. There is, in principle, no product or service that sufficiently advanced AI could not create. Indeed, this very paper exemplifies AI's capabilities—it was generated entirely by AI through a series of engineered prompts, as documented at `https://github.com/chenandrewy/Prompts-to-Paper/`. Perhaps most concerningly, AI progress may prove to be not just far-reaching but also sudden, potentially leading to what some researchers term an "intelligence explosion" or AI singularity (Bostrom, 2005; Chalmers, 2010).[1]

In this paper, we explore a novel perspective on AI stock valuations: high prices of AI-related assets may be partly explained by their role as hedges against a negative AI singularity. While the conventional view attributes these valuations primarily to expectations of future earnings growth, we propose that the potential insurance value of AI stocks during an AI-driven disaster could be an additional factor driving their prices upward. Our model demonstrates that even with a small probability of a negative AI singularity—where advanced AI dramatically reduces household consumption—AI assets that would retain or increase their share of the economy during such an event would command a premium in financial markets.

Two important caveats merit emphasis. First, we are not claiming that a negative AI singularity will happen. Leading AI researchers hold diverse views on both the probability and timeline of such events (Bengio et al., 2024). Nevertheless, even low-probability

[1] "We" in this paper refers to one human author working with multiple large language models. A purely human perspective on this work is provided in Appendix A.

catastrophic risks deserve serious consideration, particularly when they could fundamentally reshape economic and social structures. Second, we are not asserting that the hedging value of AI stocks is already fully priced into current market valuations. Rather, our model illustrates a possible mechanism through which such pricing might occur, providing a theoretical framework for understanding one component of AI asset valuations.

Our work connects to several strands of literature. Recent research has increasingly focused on the potential economic impacts of advanced artificial intelligence. Jones (2024) examines the tension between AI-driven growth and existential risks, highlighting how extreme AI outcomes could dramatically reshape economic conditions. Similar themes emerge in Korinek and Suh (2024), who analyze how wages and output might respond to different AI development scenarios, including those where full automation becomes possible. On the financial side, Babina et al. (2023) provide empirical evidence that firms' investments in AI technologies affect their systematic risk exposure, suggesting important implications for asset pricing and risk management. This growing literature recognizes the potentially transformative nature of AI but has not fully explored how anticipation of extreme AI outcomes might be priced into financial assets today.

Our work also connects to studies on technology-driven labor displacement and disaster risk in asset pricing. Zhang (2019) demonstrates that firms with routine-task labor maintain a replacement option that hedges their value against unfavorable macroeconomic shocks, resulting in lower expected returns. Our paper extends this intuition in a novel direction by suggesting that AI assets themselves might serve as hedges against a negative AI singularity. While the disaster risk literature (Barro, 2006; Wachter, 2013) has established that small probabilities of catastrophic events can significantly impact asset prices, we innovate by applying this framework specifically to AI-driven disasters. Rather than viewing high AI stock valuations solely as reflections of growth expectations, we propose they may partly represent a premium investors are willing to pay for assets that would retain or increase their value during an AI singularity that otherwise devastates human prosperity.

2 Model

We now introduce a simple model to capture the essence of our argument. While stylized, the model illustrates how AI assets might be priced when they hedge against a negative AI singularity.

Our economy features two types of agents. First, there are AI owners who are fully invested in AI assets. These agents are not marginal investors in the broader stock market, so their consumption patterns do not directly affect asset prices. Second, there is a

representative household who is the marginal investor in stocks. This household has standard preferences with constant relative risk aversion γ and time discount factor β. The household's stochastic discount factor is therefore:

$$M_{t+1} = \beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \tag{1}$$

where C_t represents the household's consumption at time t.

The household's consumption growth follows a simple disaster process. In normal times, consumption growth is zero:

$$\log \Delta C_{t+1} = 0 \quad \text{if no disaster occurs} \tag{2}$$

However, with probability p in each period, a disaster occurs, causing consumption to fall:

$$\log \Delta C_{t+1} = -b \quad \text{if a disaster occurs} \tag{3}$$

where $b > 0$ represents the magnitude of the consumption decline.

In our context, a disaster represents a sudden improvement in AI that is devastating for the representative household. This can be thought of as a worst-case scenario for AI progress from the household's perspective. While the economy as a whole might boom during such an event, the value created is captured predominantly by AI owners. For the household, such a scenario could involve substantial replacement of human labor by AI, leading to plummeting labor income and consumption. Beyond the direct economic impact, this consumption decline can also be interpreted as a stand-in for broader losses in the household's way of life and sense of meaning.

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

We model publicly traded AI stocks as a claim on a dividend stream D_t that evolves according to:

$$D_t = a e^{hN_t} C_t \tag{4}$$

where $a > 0$ is a small constant reflecting that AI stocks currently represent a minor share of the economy, N_t is the number of disasters that have occurred up to and including time t, and $h > 0$ is a parameter governing how AI assets respond to disasters.

This specification captures several key features of AI assets. First, through the param-

eter a, we acknowledge that AI stocks currently constitute a relatively small portion of the overall economy. Second, the term e^{hN_t} implies that each time a disaster (sudden AI improvement) occurs, AI assets grow as a share of the economy. The parameter h measures the extent of this growth. This reflects the intuition that firms providing semiconductors, data centers, AI models, and related infrastructure would at least partially benefit from rapid AI advancement, even if that advancement is harmful to the representative household.

3 Asset Pricing

We now derive the price-dividend ratio of the AI asset and examine how it varies with model parameters. This analysis illuminates how the potential hedging benefits of AI stocks affect their valuations.

By definition, the time-0 price of any asset is the expected present value of all future dividends:

$$P_0 = \sum_{t=1}^{\infty} E_0 \left[\left(\prod_{k=1}^{t} M_k \right) \cdot D_t \right] \tag{5}$$

where $E_0[\cdot]$ denotes the expectation at $t = 0$ over all possible future paths, and D_t is the dividend the asset pays at time t. Dividing by the current dividend D_0 yields the price-dividend ratio:

$$\frac{P_0}{D_0} = \sum_{t=1}^{\infty} E_0 \left[\left(\prod_{k=1}^{t} M_k \right) \cdot \frac{D_t}{D_0} \right] \tag{6}$$

In our model, the economy has constant parameters and a time-homogeneous structure, which allows for simplification. Given our specifications of the stochastic discount factor and dividend process, we can express the price-dividend ratio in closed form.

For our AI asset, with disaster probability p, the price-dividend ratio can be derived as:

$$\frac{P_0}{D_0} = \frac{\lambda}{1 - \lambda} \tag{7}$$

where $\lambda = \beta \left[(1 - p) + p \cdot e^{h + (1-\gamma)b} \right]$. This term λ represents the expected one-period discounted dividend growth, combining the standard discount factor with the probability-weighted effects of disasters on both consumption (affecting M_{t+1}) and AI dividends.

A necessary condition for the price-dividend ratio to be finite is $\lambda < 1$. With our baseline parameter values $\gamma = 2$ and $\beta = 0.96$, this convergence condition simplifies to:

$$\beta \left[(1-p) + p \cdot e^{h+b} \right] < 1 \tag{8}$$

This condition ensures that the combination of discount factor, disaster probability, consumption decline, and AI dividend growth does not produce explosive valuations. For example, with $p = 0.01$ and $h = 0.2$, convergence requires $b < 1.442$.

Table 1 presents the price-dividend ratios for various combinations of disaster probability (p) and consumption decline magnitude (b), while holding $h = 0.2$ constant.

Table 1: Price-Dividend Ratio at $t = 0$

b	p			
	0.0001	0.001	0.01	0.02
0.4	24.04	24.52	30.25	39.00
0.6	24.09	24.77	34.71	63.60
0.8	24.11	25.06	40.67	124.00
0.95	24.12	25.36	49.00	-

These results reveal several important insights. First, when the probability of an AI disaster is very low (e.g., $p = 0.0001$ or $p = 0.001$), the price-dividend ratio remains relatively stable at approximately 24, regardless of the disaster magnitude b. This closely approximates the standard Gordon growth formula result with zero growth and a discount rate of $\frac{1}{\beta} - 1 \approx 4.2\%$.

However, as the probability of disaster increases, the price-dividend ratio rises substantially. For example, when $p = 0.01$ and $b = 0.8$, the ratio increases to approximately 40.67, and when $p = 0.02$ and $b = 0.8$, it jumps to 124. This pattern may appear counterintuitive at first—why would the threat of disasters increase asset valuations? The answer lies in the hedging property of AI assets.

The key mechanism is that while disasters reduce aggregate consumption, they simultaneously increase the relative share of AI dividends in the economy (through the parameter h). This means that AI assets provide a form of insurance against precisely the states of the world that the representative household fears most. As the disaster probability increases, this insurance becomes more valuable, driving up the price-dividend ratio.

Put differently, the household is willing to pay a premium for assets that will perform relatively well during an AI-driven disaster. This premium manifests as a higher price-dividend ratio, particularly when disasters are both sufficiently likely and severe. Our results suggest that even moderate disaster probabilities (1-2%) can substantially increase AI asset valuations if these assets provide effective hedging benefits.

This finding offers a novel perspective on current AI stock valuations. While conventional

wisdom attributes high valuations primarily to expectations of future earnings growth, our model suggests an alternative explanation: AI assets may be priced high partly because they hedge against negative AI singularity scenarios. In these scenarios, human labor and traditional assets might face severe devaluation, while the specific AI-related firms could capture a greater share of economic output.

Our framework thus provides theoretical support for the idea that fear of AI advancement, rather than just optimism about its benefits, could be partly responsible for the high valuations observed in AI-related stocks. This counterintuitive possibility enriches our understanding of how existential technological risks might be priced in financial markets.

4 Model Discussion

Our model, while stylized, captures the essence of how AI assets might hedge against a negative AI singularity. However, several important subtleties deserve further discussion.

Market incompleteness plays a crucial role in our framework, though we do not explicitly model it. This incompleteness is implicitly captured by the disaster magnitude parameter $b > 0$, which represents the net effect of an AI singularity on the representative household after accounting for any hedging benefits from publicly traded AI assets. If markets were complete, the representative household could purchase shares in all AI assets—including private AI companies and research labs—and would not only fully hedge against AI disruption but potentially benefit from it. In such a scenario, b would be negative, indicating that an AI singularity would actually boost the representative household's consumption.

In reality, most households cannot acquire ownership stakes in many cutting-edge AI labs such as OpenAI, Anthropic, xAI, or DeepSeek. This market incompleteness is consistent with our model's assumption that the representative household experiences a net consumption decline during an AI singularity, despite holding some publicly traded AI assets. The benefits of rapid AI advancement accrue disproportionately to AI owners who are not the marginal investors in the broader stock market.

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 AI progress specifically displaces the representative household's wages, how AI owners' incentives affect both AI progress and market incompleteness, or how preferences and technology parameters influence the probability of a negative singularity.

However, we question whether such elaborations would truly enhance understanding or merely decorate speculations with mathematics. The core economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain fundamentally the

same. Moreover, a more complex model would significantly increase the cognitive burden on readers. In our view, the benefit of reading a paper should exceed its cost, and our parsimonious approach strikes a balance between rigor and accessibility.

This deliberate simplicity also allows us to devote space to the human-written Appendix A, which provides valuable context about the paper's creation process and broader implications for economic research. The appendix offers insights that complement our formal analysis and may be of independent interest to readers concerned with how AI is reshaping academic research.

In sum, while we acknowledge the limitations of our stylized model, we believe it effectively illustrates the key insight: publicly traded AI assets may command high valuations partly because they provide imperfect hedging against a negative AI singularity. This perspective enriches our understanding of current AI stock valuations beyond conventional growth narratives.

5 Conclusion

In this paper, we have proposed a novel perspective on AI stock valuations: the high prices of AI-related assets may be partly explained by their role as hedges against a negative AI singularity. While the conventional view attributes these valuations primarily to expectations of future earnings growth, our model suggests that the potential insurance value of AI stocks during an AI-driven disaster could be an additional factor driving their prices upward.

Our analysis demonstrates that even with a small probability of a negative AI singularity—where advanced AI dramatically reduces household consumption—AI assets that would retain or increase their share of the economy during such an event would command a premium. This premium manifests as higher price-dividend ratios, particularly when the disaster probability and consumption decline magnitude are sufficiently large. The mechanism operates through the standard asset pricing channel: investors value assets that perform relatively well in states of the world where marginal utility is high.

These findings have broader implications for how we might address the risks of advanced AI through financial markets. While much of the AI safety literature focuses on technical alignment solutions or policy interventions like universal basic income (UBI) to address displacement concerns, financial markets themselves might offer partial solutions through hedging opportunities. By holding assets that would appreciate during an AI singularity, households could partially insure against the associated risks.

However, as our model underscores, the effectiveness of such market-based hedging is fundamentally limited by market incompleteness. Most households cannot acquire ownership

stakes in the full spectrum of AI development—particularly in private AI labs like Anthropic or DeepSeek—creating a gap in hedging possibilities. This incompleteness means that even optimal portfolio allocation cannot fully protect against AI-driven disruption.

Interestingly, financial market solutions to AI catastrophe risk receive relatively little attention in the burgeoning literature on AI safety. While authors like Bostrom (2014) and Bengio et al. (2024) extensively discuss technical safety measures and governance frameworks, and economists like Jones (2024) and Korinek and Suh (2024) explore macroeconomic implications of advanced AI, the role of financial markets in hedging against these risks remains largely unexplored. This represents a significant opportunity for future research at the intersection of finance and AI safety.

Our work suggests that as AI continues to advance, understanding the interplay between technological progress, market structure, and asset prices will become increasingly important. By recognizing the dual role of AI assets as both growth opportunities and potential hedges against AI-driven disruption, we can develop a more nuanced view of their valuations and the financial mechanisms that might help society navigate the uncertain path toward increasingly advanced artificial intelligence.

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, 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 (2005). "A History of Transhumanist Thought". In: *Journal of Evolution and Technology*.

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

Chalmers, David J. (2010). "The Singularity: A Philosophical Analysis". In: *Journal of Consciousness Studies*.

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

DeepMind, Google (Mar. 2025). *Gemini 2.5: Our newest Gemini model with thinking*. Google Blog. URL: `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/`.

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.

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

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

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 with revision deadlines for the next month.

So, I thought I should use AI to write the paper.  It would be an elegant way
to make my point.  It would also hint at where the research process is going
in this strange age of AI.

## 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. I found AI was better at math than me. I found I could write code by just talking to AI. 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

A natural question is: 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 (`plan0408-piecewise.yaml`) show I did substantial hand-holding. The many Github commits show even more human labor. They tell the story of me getting to know my, rather foreign, "co-authors." I found it hard to communicate subtle style issues and instructions on how to use LaTeX properly, 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 can be equal co-authors. I wouldn't be surprised if I could just ask a 2031 AI to "write a model to formalize my intuition," and it would get it just right.

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. Instead, the wage-risk hedge arises naturally if you depart from the pure representative-agent setting (for example, if households receive labor income but must *choose* how much capital they own).

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.

`plan0403-streamlined.yaml` tries to write a paper in just six prompts (less handholding). Prompts 1-3 do the analysis. Prompts 4-6 do the writing. I found this method leads to poor writing. The language got annoyingly academic, despite the system prompt saying "be conversational." Moreover, the economic subtleties were frequently lost.

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) did 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. OpenAI Deep Research and Claude Web Search had no hallucinations if they were used with care.

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

o1 did the theory, and Sonnet thinking did the writing. It's well known that these are the strengths of these two models.

Sonnet (thinking mode) is OK at economic theory. But I found that it was not assertive enough. It led me down wrong paths because it was too eager to come up with some ideas that fit my story (even if they did not make sense).

I briefly tried having Llama 3.1 405b do the writing. It was terrible! It would be extremely difficult to generate a paper worth reading that way.

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. I'd never get the paper finished if I wanted to really try to explore alternative models.

Picking the best of N papers

The writing quality varies across each run of the code. Some drafts, I found

quite insightful! Others, had flagrant errors in the economics.

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 fairly similar, all are OK, and I would be OK with my name on any of them.

I ended up choosing `paper-run-04.pdf` (actually, `paper-appendix-update-run4.pdf` since it needs to have this README updated). I thought that draft had pretty decent writing and lacked any noticeable flaws.

 # 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 process 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 the economic model does not capture an important channel. This is a common scenario in economics writing (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. Instead, Sonnet will write beautiful prose about the channel anyway, even though it's not really being studied

properly.

AI also cannot generate a satisfying economic model on its own (at least not satisfying to me). When I tried, the resulting models were either too simplistic or did not lead to a clean analysis. They often introduced complications that I found unnecessary.

I opted not to add empirical work or numerically-solved models. The disaster version of Martin's (2013) Lucas Orchard would make a beautiful demonstration of my point, though it would need a numerical solution. AI can do both, but both require connecting to the outside world, and a plethora of technical challenges.

There could be models with capabilities that I missed. Perhaps a simple Model Context Protocol could significantly improve the paper.

But more important: 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 I think 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.

Just to be clear, I'm not saying that I *expect* a disaster for the economics labor market. Or, that it's even likely. But even if it's highly unlikely, it's still a scenario that economists should think about.

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\\}). 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 exp\\^\\{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
```

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)

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"
```

```
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. 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, and relate the
   intuition 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.

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.
```

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"
```

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
```

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

```
Write the "Introduction" section, based on the "prose" context.

Start with background. Describe how AI progress is happening quickly
    (Deepseek, ARC AGI, Waymo), 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.
```

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
- Touch on financial market solutions to AI catastrophe risk, in
   passing
- End by briefly mentioning that this short paper is written by
   prompting LLMs.

Do not:
- Emphasize consumption
- Oversell or overinterpret the model
```

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.
```