Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

This paper proposes that high AI stock valuations may partly reflect their role as hedges against a negative AI singularity—a scenario where rapid AI advancement benefits AI companies but reduces consumption for the representative investor. Our theoretical framework demonstrates how this hedging property increases with disaster severity and probability, potentially explaining seemingly disproportionate AI market capitalizations. While financial markets might naturally develop mechanisms to address AI risks, their effectiveness is limited by market incompleteness. Understanding these financial responses complements technical and governance approaches to AI safety. This paper was generated through a series of engineered prompts to large language models, with human guidance and curation.

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

Recent years have witnessed extraordinary advances in artificial intelligence capabilities. DeepSeek's R1 model has demonstrated remarkable reasoning abilities (DeepSeek-AI et al., 2025), contestants have made significant progress on the once-intractable ARC AGI challenge (Chollet et al., 2024), and Waymo's autonomous vehicles now navigate complex urban environments without human intervention. These developments have sparked growing investor concern about the potential displacement of human labor by increasingly capable AI systems. While economists have long studied the impacts of automation on labor markets (Acemoglu and Restrepo, 2020; Karabarbounis and Neiman, 2014), the current wave of AI progress presents unique challenges given its rapid pace and broad applicability.

Unlike previous technological revolutions that transformed specific sectors, AI has the potential to create virtually any product or service currently produced by humans. As Vinge (1993) observed decades ago, we are approaching a point where "our old models must be discarded and a new reality rules." This paper itself exemplifies this transformative potential—it was entirely generated by AI systems through a series of engineered prompts, with a human providing only guidance and curation (see `https://github.com/chenandrewy/Prompts-to-Paper/`). Such capabilities raise the possibility of an "intelligence explosion" or technological singularity (Good, 1965; Chalmers, 2010), where AI progress becomes self-reinforcing and potentially sudden.

This paper examines how AI stocks might be priced in light of these developments, particularly the possibility that advanced AI could significantly impact human consumption. We[1] develop a theoretical framework where AI technologies present a double-edged sword: they might simultaneously increase overall productivity while reducing the consumption share that accrues to the representative household. In our model, a "negative AI singularity" represents a disaster state where rapid AI progress leads to consumption declines for the average investor, even as AI-focused companies capture increasing economic value. This creates a natural hedging property for AI stocks, potentially explaining part of their high valuations.

Several important caveats should be noted. First, we are not predicting that a negative AI singularity will occur. The possibility of such an event remains speculative, and many experts envision AI progress as broadly beneficial for humanity. Nevertheless, prudent risk analysis requires consideration of adverse scenarios, particularly when they entail potentially systemic consequences. Second, we are not claiming that this hedging value is currently priced into

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

AI stocks. Our model illustrates a possible mechanism through which such pricing could occur, but empirical validation of this channel requires further research.

Our paper contributes to several strands of literature. First, we build on the rare disaster framework pioneered by Rietz (1988) and advanced by Barro (2006) and Wachter (2013), adapting it to model technological disruption. While standard disaster models assume that assets suffer during disasters, we consider the possibility that certain assets—specifically AI stocks—might benefit during technological disaster states.

Second, we connect to the emerging literature on AI's impact on financial markets. Zhang (2019) demonstrates that firms with greater automation potential maintain lower expected returns due to their embedded option to replace routine-task labor during economic downturns. Babina et al. (2023) provides evidence that firms' investments in AI technologies affect their systematic risk profiles, with implications for their market beta and cost of capital. Our work extends these insights by proposing a novel mechanism through which AI stocks might serve as a hedge against broader technological disruption.

Finally, our paper contributes to the literature on technological singularity and its economic implications. Korinek and Suh (2024) analyzes how output and wages respond to different AI development scenarios, examining the transition to artificial general intelligence. Jones (2024) highlights the economic dilemma between AI-driven growth and existential risk, suggesting that the transition to advanced AI creates inherent trade-offs between potential benefits and catastrophic outcomes. Our model provides a financial market perspective on these trade-offs, showing how asset prices might reflect the dual nature of transformative AI technologies.

2 Model

We now present a simple model that captures the essence of our argument. The model is intentionally stylized to highlight the key mechanism through which AI stocks might serve as a hedge against negative AI singularity events.

Our economy consists of two types of agents. First, there are AI owners who are fully invested in AI companies and are not marginal investors in the broader stock market. Second, there is a representative household who is the marginal investor in stocks. Since only the household's consumption matters for asset pricing in this framework, we focus on their preferences and consumption process.

The representative household has standard preferences with constant relative risk aversion γ and time discount factor β:

$$U_t = E_t \left[\sum_{j=0}^{\infty} \beta^j \frac{C_{t+j}^{1-\gamma}}{1-\gamma} \right] \tag{1}$$

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

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

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

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

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

In our context, a disaster represents a sudden improvement in AI technology 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 may boom during such events, the value created is captured primarily by AI owners. For the household, the negative impact comes through multiple channels: labor income plummets as jobs are replaced by AI, and there may be broader losses to way of life and meaning that are difficult to quantify. The consumption drop b serves as a stand-in for these various losses.

We assume that at $t = 0$, no disasters have yet occurred, meaning the AI singularity has not yet taken place. The model allows for multiple disasters to occur over time, capturing the ongoing uncertainty even after an initial singularity event.

We model publicly traded AI stocks as a single asset with dividend process:

$$D_t = a \exp(hN_t) C_t \tag{4}$$

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 much AI stocks benefit from disasters.

This specification captures the idea that each time a disaster (sudden AI improvement) occurs, AI assets grow as a share of the economy. Intuitively, firms that provide the infrastructure for AI advancement—such as semiconductors, data centers, AI models, and related technologies—at least partially benefit from rapid AI progress, even if this progress is harmful to the representative household. The parameter h controls the strength of this effect, with larger values indicating that AI stocks capture more value during singularity events.

3 Results

We now derive the price-dividend ratio for the AI asset and examine how it relates to our key thesis about AI stocks serving as a hedge against negative AI singularity events.

At time $t = 0$, the fundamental pricing equation states that the price of the AI asset P_0 equals the expected discounted value of next period's dividend plus price:

$$P_0 = \mathbb{E}_0[M_1(D_1 + P_1)] \tag{5}$$

where M_1 is the stochastic discount factor between $t = 0$ and $t = 1$. Dividing by the current dividend D_0 gives the price-dividend ratio $v_0 = P_0/D_0$.

Since disasters occur independently each period with probability p, we can express the price as the present value of all future expected dividends:

$$P_0 = \mathbb{E}_0\left[\sum_{j=1}^{\infty} M_{0,j} D_j\right] \tag{6}$$

where $M_{0,j}$ is the multi-period discount factor from time 0 to j. For our representative household with CRRA utility, this discount factor is $M_{0,j} = \beta^j (C_j/C_0)^{-\gamma}$.

In our model, each disaster has two effects: it increases AI dividends by a factor e^h and reduces consumption by a factor e^{-b}. For period j, if n_j disasters have occurred, the discount factor becomes $\beta^j e^{\gamma b n_j}$, while the dividend is $D_j = a e^{h n_j} C_j = a e^{h n_j} C_0 e^{-b n_j}$.

Taking the expectation over the binomial distribution of disasters in j periods yields:

$$P_0 = aC_0 \sum_{j=1}^{\infty} [\beta(1 - p + p e^{h-b+\gamma b})]^j \tag{7}$$

For convenience, let's define $\lambda \equiv \beta(1 - p + p e^{h-b+\gamma b})$. When $\lambda < 1$, this geometric series converges to:

$$P_0 = aC_0 \frac{\lambda}{1 - \lambda} \tag{8}$$

Since $D_0 = aC_0$ at time 0 (when no disasters have yet occurred), the price-dividend ratio is:

$$v_0 = \frac{P_0}{D_0} = \frac{\lambda}{1 - \lambda} = \frac{\beta(1 - p + p e^{h-b+\gamma b})}{1 - \beta(1 - p + p e^{h-b+\gamma b})} \tag{9}$$

This expression reveals how the price-dividend ratio depends on the interplay between the disaster probability p, disaster size b, and the AI dividend boost h.

The pricing formula offers several important insights. For the price-dividend ratio to be finite, we need $\lambda < 1$. This condition places constraints on the parameters, particularly the relationship between h and b.

When disasters are interpreted as negative AI singularity events, our model reveals a counterintuitive mechanism: if AI stocks benefit during such events ($h > 0$), they can serve as a hedge for the representative household. Even though the household experiences consumption declines during disasters, they receive some compensation through their AI stock holdings, which increase in value.

This hedging property becomes more valuable as: 1. Disasters become more severe (higher b) 2. AI stocks benefit more from disasters (higher h) 3. The probability of disasters increases (higher p)

To illustrate these effects, Table 1 presents numerical values for the price-dividend ratio under different parameter combinations. We fix $h = 0.2$, $\gamma = 2$, and $\beta = 0.96$, while varying the disaster size b and probability p.

Table 1: Time-0 Price-Dividend Ratio ($v_0 = P_0/D_0$)

b	p			
	0.0001	0.001	0.01	0.02
0.40	24.04	24.52	30.25	39.00
0.60	24.08	24.76	34.71	63.60
0.80	24.10	25.06	40.67	141.90
0.95	24.13	25.38	51.63	–

The table reveals several patterns that align with our thesis about AI stocks serving as hedges. As the disaster size b increases (moving down the rows), the price-dividend ratio rises, reflecting the increased hedging value of AI stocks. Similarly, as the disaster probability p increases (moving right across columns), the price-dividend ratio also increases substantially. This illustrates how perception of higher AI risk could drive AI stock valuations higher.

This pattern is most dramatic when both parameters are high. For instance, with $b = 0.80$ and $p = 0.02$, the price-dividend ratio reaches 141.90, nearly six times its value when disasters are extremely rare ($p = 0.0001$). This suggests that increasing concerns about negative AI singularity could significantly contribute to high AI stock valuations, even if the probability remains objectively small.

These results provide a novel perspective on the current high valuations of AI stocks. While traditional explanations focus on expectations of future earnings growth, our model suggests an additional factor: AI stocks may be priced higher because they serve as a hedge against negative AI singularity events. In essence, the market may be implicitly pricing the insurance value of these assets, even if market participants are not explicitly thinking in these terms.

4 Model Discussion

Our model captures the essence of how AI stocks might serve as a hedge against negative AI singularity events. However, like any model, it makes simplifications that merit further discussion.

A crucial feature of our model is the implicit market incompleteness, which manifests through the disaster magnitude parameter $b > 0$. This parameter represents the net effect on the representative household after accounting for both the negative impact of an AI disaster and any offsetting benefits from publicly traded AI assets.

The positive value of b indicates that, on net, the representative household suffers from AI singularity events. This would not be the case in a complete markets setting. If markets were complete, the representative household could purchase shares in all AI assets—both public and private—and thereby not only fully hedge against AI disasters but potentially benefit from them. In such a scenario, b would likely be negative, turning the "disaster" into a sudden boom for the representative household.

However, reality more closely resembles our incomplete markets setting. Most households cannot invest in many cutting-edge AI labs such as OpenAI, Anthropic, xAI, or DeepSeek. These private companies are developing some of the most advanced AI systems but remain inaccessible to ordinary investors. This market structure means that even if households invest heavily in publicly traded AI stocks, they still face substantial unhedged exposure to AI progress risk.

This incompleteness is not a modeling oversight but rather a reflection of real-world constraints that limit risk-sharing opportunities. When a technological revolution occurs that primarily benefits a small set of private companies, the broader population may experience displacement effects without corresponding financial compensation.

One might reasonably ask whether a more elaborate model would yield additional insights. A more complex framework could certainly add detail regarding AI owners, private AI assets, and their interactions with the representative household. Such a model might address questions like:

How does AI progress specifically displace the representative household's wages? How do AI owners' incentives affect both the pace of AI progress and the persistence of market incompleteness? How might preference parameters and technological constraints influence the probability of a negative singularity?

While these questions are fascinating, we believe our simpler approach has merit for several reasons. First, a more complex model would essentially decorate speculations with mathematics. Given the inherent unpredictability of transformative AI, adding layers of mathematical complexity might create an illusion of precision where none exists. The core economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain fundamentally the same.

Second, there is a tradeoff between model complexity and readability. As economists, we believe the benefit of reading a paper should exceed its cost. A more elaborate model would substantially increase the cognitive burden on readers without necessarily providing proportionate insights. Our streamlined approach allows us to highlight the key mechanism while keeping the analysis accessible.

5 Conclusion

In this paper, we have presented a novel perspective on AI stock valuations. While the prevailing view attributes high valuations primarily to expectations of future earnings growth, we propose that AI stocks may also be valued for their potential as hedges against negative AI singularity events. Our theoretical framework demonstrates that if AI companies benefit during singularity events while the representative household experiences consumption decline, AI stocks can serve as a partial hedge against technological disruption.

Our model highlights several key mechanisms driving this relationship. First, market incompleteness plays a crucial role—ordinary investors cannot fully hedge against AI risks because they cannot invest in many cutting-edge private AI labs developing the most advanced systems. Second, the hedging value of AI stocks increases with disaster severity, the degree to which AI companies benefit from disasters, and the perceived probability of singularity events. Third, even small probabilities of negative AI singularity can substantially impact valuations when the potential consumption impact is large.

These findings suggest that financial markets might naturally develop mechanisms to partially address AI catastrophe risks. As technology advances, we might see the emergence of more sophisticated financial products specifically designed to hedge against negative AI singularity events. For example, traditional options markets could expand to include instruments explicitly tied to AI development milestones or regulatory changes. Investment funds might emerge that specialize in constructing portfolios optimized for hedging against technological disruption.

However, the effectiveness of such market-based solutions is fundamentally limited by the incompleteness of markets. As long as a significant portion of AI development occurs in private companies inaccessible to ordinary investors, the representative household will retain substantial unhedged exposure to AI risk. This limitation represents a significant challenge for risk sharing in the face of potentially transformative AI.

Interestingly, these financial market mechanisms receive little attention in the AI catastrophe risk literature. While scholars like Bostrom (2014) and Russell (2019) have extensively discussed technological and governance approaches to AI safety, and economists like Jones (2024) and Korinek and Suh (2024) have examined economic implications of AI transitions, few have explored how financial markets might evolve to address these risks. Bengio et al. (2024) discuss managing extreme AI risks but focus primarily on technical safeguards and governance rather than financial mechanisms.

Understanding these potential financial market responses is important for a comprehensive approach to managing AI transition risks. While such market-based mechanisms cannot replace technical safety measures or appropriate governance frameworks, they represent an additional tool for society to manage the complex challenges posed by rapidly advancing AI technologies. Future research should further explore the interplay between financial markets and AI risk management, as well as potential policy approaches that could address the inherent limitations of market incompleteness in this domain.

References

Acemoglu, Daron and Pascual Restrepo (2020). "Robots and Jobs: Evidence from US Labor Markets". In: *Journal of Political Economy*.

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 (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*.

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`.

Good, I.J. (1965). "Speculations Concerning the First Ultraintelligent Machine". In: *Advances in Computers*.

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

Karabarbounis, Loukas and Brent Neiman (2014). "The Global Decline of the Labor Share". In: *Quarterly Journal of Economics*.

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.

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*.

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.
```