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 harms the representative investor while benefiting AI firms. Our parsimonious model demonstrates that even a small probability of such events can significantly increase AI stock valuations through a hedging premium. This perspective highlights how financial markets naturally develop instruments allowing individuals to partially hedge against technological transitions, complementing regulatory and technical approaches to AI risk management. This short paper was written entirely through a series of engineered prompts to large language models.

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 unprecedented progress in artificial intelligence capabilities. DeepSeek's R1 model achieved performance comparable to OpenAI's o1 in mathematical problem-solving and reasoning through reinforcement learning (DeepSeek-AI et al., 2025). The ARC AGI Challenge, designed to measure general intelligence rather than narrow capabilities, saw scores rise dramatically from 33% to over 55% in a single year (Chollet et al., 2024). Meanwhile, Waymo's autonomous vehicles have accumulated millions of miles with minimal human intervention, operating commercial services in multiple cities. This accelerating pace of AI development has naturally led many investors to question the security of their future wages and livelihoods.

While technological change has occurred throughout human history, AI represents a fundamentally different kind of innovation. Unlike previous technologies that excelled at specific tasks, AI has the potential to perform virtually any intellectual task a human can do. There is, in principle, no product or service that sufficiently advanced AI could not create. This paper itself exemplifies this point—it was generated entirely by AI through a series of engineered prompts (code available at https://github.com/chenandrewy/Prompts-to-Paper/). Moreover, progress in AI capabilities may be incredibly sudden—the hypothesized "AI singularity" where recursively self-improving AI could rapidly surpass human capabilities across all domains.[1]

In this paper, we study how AI stocks are priced given the risk that advanced AI may disrupt livelihoods and reduce consumption for the representative investor. We develop a simple asset pricing model where AI firms partially benefit from technological advances that might harm the average household's consumption. This creates a natural hedging property for AI stocks against the risk of a negative AI singularity event.

We emphasize two important caveats. First, we are not claiming that a negative singularity will definitely occur. The future of AI development remains highly uncertain, with many paths ranging from beneficial coexistence to serious disruption. Nevertheless, even low-probability scenarios warrant consideration in risk-aware investment strategies. Second, we are not asserting that this hedging value is already incorporated into current AI stock prices. Our model simply illustrates a possible mechanism through which rational investors might value AI stocks partly for their hedging properties against technological displacement.

The intersection of artificial intelligence and asset pricing encompasses several strands of literature. Research on AI existential risk has gained prominence, with Jones (2024)

[1] Throughout this paper, "we" refers to one human author and multiple large language models working together. For a purely human perspective on this work, see Appendix A.

highlighting the fundamental tension between AI-driven economic growth and potentially catastrophic outcomes. Korinek and Suh (2024) analyze economic scenarios during the transition to AGI, showing how output and wages respond to different AI development paths, including a potential ”wage collapse” in full automation scenarios. On the asset pricing side, the rare disaster literature pioneered by Rietz (1988) and extended by Barro (2006) and Gabaix (2012) demonstrates how small probabilities of catastrophic events can significantly impact asset valuations and risk premia. Research on hedging technological displacement has also emerged, with Zhang (2019) documenting that firms with routine-task labor maintain a technological replacement option that hedges against negative macroeconomic shocks, lowering their expected returns. More recently, Babina et al. (2024) provide direct evidence that firms' investments in AI technologies affect their systematic risk profiles.

Our paper contributes to this literature by proposing a novel perspective on the valuation of AI stocks. While existing research typically attributes high AI valuations to expectations of future earnings growth, we develop a theoretical framework suggesting that these valuations may partially reflect a risk-hedging premium. Specifically, AI stocks could serve as a hedge against negative AI singularity events—scenarios where rapid AI advancement harms the representative investor's consumption. This hedging view complements growth-based explanations and connects the asset pricing literature with research on AI risk. Unlike Zhang (2019), who shows how firms can hedge their own technological risk through automation options, we examine how investors might hedge against broader AI-driven economic disruption through their portfolio choices. Our model provides a rational explanation for high AI stock valuations even when the probability of a negative singularity is small, demonstrating how disaster risk can influence asset prices through the hedging channel rather than the more commonly studied direct risk exposure channel.

2 Model

We now present a simple model that captures the essence of our argument. While the model is intentionally stylized, it allows us to illustrate the key mechanism through which AI stocks may serve as hedges against negative AI singularity events.

Our economy consists of two types of agents. First, there are AI owners who are fully invested in AI technologies 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 representative household's consumption matters for our asset pricing analysis, we focus on their preferences.

The representative household has constant relative risk aversion (CRRA) preferences with risk aversion parameter γ and time discount factor β. Their utility function is given by:

$$U = E_0 \left[\sum_{t=0}^{\infty} \beta^t \frac{C_t^{1-\gamma}}{1-\gamma} \right] \tag{1}$$

The representative household's consumption growth follows a simple process with disaster risk. Let $\Delta c_{t+1} = \log(C_{t+1}/C_t)$ denote log consumption growth. 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, leading to a significant drop in consumption:

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

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

In our context, a disaster represents a sudden improvement in AI technology that is devastating for the representative household. This can be thought of as a worst-case scenario for AI progress, akin to a negative singularity event. When such an event occurs, the economy as a whole may boom, but the value of this technological advancement is captured primarily by AI owners. For the representative household, the consequences are severe: labor is replaced by AI, causing labor income to plummet, which in turn reduces consumption. Beyond the direct economic impact, households may also experience losses in their way of life and sense of meaning, which we capture implicitly through the consumption decline.

At time $t = 0$, we assume no disasters have yet occurred, meaning the AI singularity has not yet taken place. However, multiple disasters may occur over time, representing ongoing uncertainty about the evolution of AI technology even after an initial singularity event.

We model publicly traded AI stocks as a single asset that pays a dividend D_t in each period. The dividend process is given by:

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

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 (AI singularity events) that have occurred up to and including time t, and $h > 0$ is a parameter that governs how the AI asset's share of the economy grows with each disaster.

This specification captures a key feature of our argument: each time a disaster occurs (i.e., when there is a sudden improvement in AI that harms the representative household), the AI asset grows as a share of the economy. Intuitively, firms that provide the infrastructure for AI advancement—such as semiconductor manufacturers, data providers, and developers of AI models—at least partially benefit from these technological leaps, even as the representative household suffers. This creates a natural hedging property for AI stocks against the risk of negative AI singularity events.

3 Asset Prices

Given our model setup, we now derive the equilibrium price of the AI asset and analyze how this price varies with the probability and severity of AI singularity events. This analysis will help us understand the mechanism through which AI stocks can serve as hedges against negative AI singularity events.

We begin by deriving the price-dividend ratio of the AI asset at time 0, before any disasters have occurred. The standard asset pricing equation relates the ex-dividend price P_t of the asset to its expected discounted future payoffs:

$$P_t = E_t[M_{t+1}(P_{t+1} + D_{t+1})] \tag{5}$$

where $M_{t+1} = \beta(C_{t+1}/C_t)^{-\gamma}$ is the stochastic discount factor from period t to $t+1$.

To solve for the price-dividend ratio, we introduce a scaled price variable Q_t, defined as:

$$Q_t \equiv \frac{P_t}{e^{hN_t}C_t} \tag{6}$$

Since $D_t = ae^{hN_t}C_t$, we can express the price-dividend ratio as:

$$\frac{P_t}{D_t} = \frac{Q_t}{a} \tag{7}$$

Substituting the definition of Q_t into the asset pricing equation and using the expressions for M_{t+1}, P_{t+1}, and D_{t+1}, we derive a recursive equation for Q_t:

$$Q_t = E_t\left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{1-\gamma} e^{h(N_{t+1}-N_t)}(Q_{t+1} + a)\right] \tag{8}$$

At time 0, with $N_0 = 0$, and recognizing that Q_t is stationary when appropriately scaled, we can simplify this recursion. Considering the two possible states—disaster or no disaster—we get:

$$Q = \beta[(1-p)(Q+a) + p \cdot e^{(1-\gamma)(-b)} \cdot e^h \cdot (Q+a)] \tag{9}$$

For convenience, let $\Omega \equiv (1-p) + p \cdot e^{(1-\gamma)(-b)} \cdot e^h$. Then:

$$Q = \beta\Omega(Q + a) \tag{10}$$

Solving for Q, we obtain:

$$Q = \frac{a\beta\Omega}{1 - \beta\Omega} \tag{11}$$

Therefore, the price-dividend ratio at time 0 is:

$$\frac{P_0}{D_0} = \frac{Q}{a} = \frac{\beta\Omega}{1 - \beta\Omega} \tag{12}$$

where $\Omega = (1-p) + p \cdot e^{(1-\gamma)(-b)} \cdot e^h$.

For this ratio to be finite, we require $\beta\Omega < 1$, which places restrictions on the parameters of the model. Specifically, for our calibration with $\gamma = 2$ and $\beta = 0.96$, this translates to a constraint on the sum $b + h$.

The formula for the price-dividend ratio reveals several important insights. First, the ratio increases with the disaster probability p, provided that $e^{(1-\gamma)(-b)} \cdot e^h > 1$. With $\gamma > 1$ (as is typically assumed in the asset pricing literature), this condition holds when h is sufficiently large relative to b.

Intuitively, when the probability of an AI singularity event increases, the representative household places a higher value on assets that perform well during such events. Since the AI asset's dividends grow proportionally with each disaster (through the factor e^h), the asset becomes more valuable as a hedge against these adverse events.

The price-dividend ratio also increases with the severity of the disaster b, as a larger consumption drop makes the hedging properties of the AI asset more valuable. Similarly, a higher value of h indicates that the AI asset captures more of the economic value created during singularity events, further increasing its price-dividend ratio.

To illustrate these relationships numerically, we present a table of price-dividend ratios for various combinations of disaster probability p and disaster size b, fixing $h = 0.2$, $\gamma = 2$, and $\beta = 0.96$:

The table clearly demonstrates that the price-dividend ratio increases both with the probability of a disaster and with the severity of the consumption decline. For instance, when $b = 0.8$ (representing a substantial 55% drop in consumption during a disaster), the price-dividend ratio rises from 24.11 to 42.48 as the disaster probability increases from 0.0001 to 0.01.

These results support our main argument: even if market participants assign a relatively

b	p			
	0.0001	0.001	0.01	0.02
0.4	24.00	24.52	29.30	39.00
0.6	24.08	24.76	34.70	61.50
0.8	24.11	25.06	42.48	141.90
0.95	24.14	25.36	51.63	—

Table 1: Price-Dividend Ratios for the AI Asset

small probability to a negative AI singularity event, the hedging value of AI stocks can significantly increase their valuations. This effect becomes more pronounced as either the perceived probability or the expected severity of such events increases.

Thus, our model provides a rational explanation for high AI stock valuations that does not rely solely on expectations of future earnings growth. Instead, it highlights how these valuations may reflect the insurance value that AI stocks provide against the risk of a negative AI singularity—a risk that, while small in probability, could have profound consequences for the representative household.

4 Model Discussion

Our model, while intentionally stylized, captures the key economic mechanism through which AI stocks may serve as hedges against negative AI singularity events. However, several important subtleties deserve further discussion.

Market incompleteness plays a central role in our analysis, 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 both the AI disaster and the AI asset's dividend response. If markets were complete, the representative household could purchase shares in all AI assets, including private ones, thereby not only fully hedging against singularity risk but potentially benefiting from it. In such a scenario, b would be negative, indicating a consumption boom rather than a disaster.

In reality, most households cannot invest in many cutting-edge AI labs such as OpenAI, Anthropic, xAI, or DeepSeek. These private companies, which are at the forefront of AI development, remain largely inaccessible to the average investor. This market structure is consistent with our model's assumption that the representative household can only partially hedge against AI singularity risk through publicly traded AI stocks.

A more elaborate model could certainly add detail to the AI owners, private AI assets, and their interactions with the representative household. Such a model might address questions like how AI progress 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 must ask whether such elaborations would genuinely enhance our understanding or merely decorate speculations with mathematics. The core economic insights—rare disaster risk, hedging motives, and market incompleteness—would remain fundamentally the same. Moreover, a more complex model would be considerably more costly for readers to digest.

We believe that the benefit of reading a paper should exceed its cost. Our streamlined approach allows us to convey the essential economic mechanism without unnecessary complexity. It also leaves room for the human-written Appendix A, which provides valuable context about the paper's motivation and creation process.

In sum, while our model abstracts from many real-world complexities, it captures the essential economic forces at play. The possibility that AI stocks serve as hedges against negative AI singularity events emerges naturally from standard asset pricing considerations when markets are incomplete and technological progress creates both winners and losers.

5 Conclusion

In this paper, we propose a novel perspective on the high valuations of AI stocks. While the conventional view attributes these valuations primarily to expectations of extraordinary future earnings growth, we suggest an alternative, complementary explanation: AI stocks may serve as a hedge against negative AI singularity events.

Using a parsimonious model with a representative investor and a stylized AI asset, we show that when there is even a small probability of a sudden AI advancement that harms the representative household's consumption, AI stocks can command a significant price premium. This occurs because these stocks, which represent firms providing the infrastructure for AI advancement, at least partially benefit from technological leaps that might otherwise be detrimental to the average household. Our numerical analysis demonstrates that as either the probability or the severity of an AI singularity increases, both the price-dividend ratio and the risk premium of AI stocks rise accordingly.

The financial market mechanism we identify offers an important perspective on managing AI catastrophic risk that has received insufficient attention in the broader discussion. Most analyses of AI risk mitigation focus on regulatory approaches, alignment research, or universal social programs as response mechanisms. However, financial markets naturally develop instruments that allow individuals to hedge against significant economic transitions, and AI represents perhaps the most consequential technological transition on the horizon.

This market-based approach to addressing AI risk is inherently limited by market incompleteness, as we highlight in our model. Not all households can invest in leading AI companies, particularly private ones that might capture the largest share of value from AI advancement. This incompleteness creates a wedge between the potential and realized hedging benefits of AI investments. Yet even acknowledging these limitations, the existing public markets provide a partial hedging mechanism that merits more thorough consideration in AI risk discussions.

The current literature on AI catastrophic risk, exemplified by works such as Bengio et al. (2024), Bostrom (2014), and Jones (2024), has thoroughly explored the technological and policy dimensions of AI risk but has devoted less attention to how financial markets can distribute and price this risk. Similarly, Korinek and Suh (2024) analyze various economic scenarios during the transition to AGI but do not fully explore the role of asset markets in hedging transition risks.

Our work bridges asset pricing theory with research on AI risk, demonstrating how standard financial market mechanisms may already be pricing and partially addressing catastrophic AI risks. While financial markets alone cannot solve all challenges posed by advanced AI, understanding their role in risk allocation provides valuable insights for both investors and policymakers navigating an increasingly AI-driven future.

References

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson (2024). "Artificial intelligence, firm growth, and product innovation". In: *Journal of Financial Economics* 151, Article 103745.

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.

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.

Gabaix, Xavier (2012). "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance". In: *Quarterly Journal of Economics* 127.2, pp. 645–700.

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.

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

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.
```