Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

We propose that the high valuations of AI stocks may partially reflect their role as hedges against a negative AI singularity—an explosion of AI development that is devastating for the representative investor. Our theoretical framework demonstrates how even a small probability of such an event can generate substantial price-dividend ratios for AI assets that would perform relatively better in disaster scenarios. This perspective offers insights into potential market-based approaches to managing AI catastrophe risk, though market incompleteness severely limits this hedging potential. Unlike previous work, this short paper is written by prompting LLMs.

Keywords: Artificial Intelligence, Disaster Risk, Asset Pricing

*email:andrew.y.chen@frb.gov. ChatGPT-o1 and Claude-3.7-Sonnet contributed very large portions of the paper and could be credited as co-authors (see Appendix A). I thank Andrei Goncalves for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System.

1 Introduction

The pace of artificial intelligence development has accelerated dramatically in recent years. Large language models like OpenAI's o1 and DeepSeek's R1 have demonstrated remarkable reasoning capabilities, solving complex mathematical problems and writing code with human-like proficiency. Progress on benchmarks like the Abstraction and Reasoning Corpus (ARC-AGI Challenge) has surged, with scores jumping from 33% to over 55% in a single year (Chollet et al., 2024). Meanwhile, companies like Waymo have deployed autonomous vehicles that have logged tens of millions of miles without human intervention. As these technologies advance, many workers face legitimate concerns about their skills becoming obsolete and their wages declining.

Technological change is not new—from the Industrial Revolution to the digital age, innovation has transformed labor markets throughout history. However, artificial intelligence represents a fundamentally different kind of disruption. Unlike previous technologies that automated specific physical or cognitive tasks, AI has the potential to perform virtually any intellectual task that humans can do. There is, in principle, no product or service that sufficiently advanced AI could not create. This paper itself serves as a testament to this capability—it was generated entirely by AI systems through a series of engineered prompts (code available at `https://github.com/chenandrewy/Prompts-to-Paper/`). Beyond gradual progress, some researchers have proposed the possibility of an ”AI singularity”—a hypothetical point where AI capabilities improve so rapidly that they fundamentally transform society in ways difficult to predict (Vinge, 1993; Bostrom, 2014).

In this paper, we explore how the possibility of a negative AI singularity might affect the valuation of AI-related stocks. We develop a simple theoretical framework to analyze how AI assets could serve as hedges against scenarios where sudden AI advancement proves devastating for the representative investor. The core insight of our model is that AI companies might capture a larger share of economic value during such transformative events, even as overall consumption declines. This creates a hedging motive for holding AI stocks, potentially explaining part of their high valuations beyond growth expectations.[1]

We emphasize two important caveats. First, we are not predicting that a negative AI singularity will occur. We recognize that such scenarios remain speculative, but nevertheless believe they warrant serious consideration given their potential magnitude. Second, we are not claiming that this hedging value is already fully reflected in current AI stock prices. Our model illustrates a possible mechanism rather than providing a definitive explanation for

[1] Throughout this paper, ”we” refers to one human author and multiple LLMs. A purely human perspective on this research is provided in Appendix A.

observed valuations.

Our paper connects several strands of literature. First, we build on the extensive work on disaster risk in asset pricing, pioneered by Rietz (1988) and Barro (2006), and extended by Wachter (2013) and Gabaix (2012). These models demonstrate how the small probability of rare catastrophic events can substantially impact asset prices. Recent work by economists has begun exploring the implications of transformative AI, with Jones (2024) analyzing the tension between AI-driven growth and existential risk, and Korinek and Suh (2024) exploring different scenarios for the transition to artificial general intelligence, including potential labor displacement. In the finance literature, Zhang (2019) shows that firms with routine-task labor maintain a replacement option that hedges their value against unfavorable macroeconomic shocks, resulting in lower expected returns. Similarly, Babina et al. (2023) provides evidence that firms' investments in AI technologies affect their systematic risk profiles, with implications for asset pricing and cost of capital.

Our paper contributes to this literature by proposing a novel perspective on AI stock valuations. While previous work has focused on AI's potential to drive growth and productivity (Babina et al., 2024), we argue that the high valuations of AI stocks might partially reflect their role as hedges against a negative AI singularity—an event where rapid AI advancement is devastating for the representative investor. This contrasts with the common view that AI valuations are high solely due to expectations of future earnings growth. Our model shows how even a small probability of an AI disaster could generate substantial price-dividend ratios for AI assets that perform relatively well in such scenarios. This insight offers a new lens for understanding tech valuations, suggesting that investors may be willing to pay a premium for assets that serve as insurance against technological disruption.

2 Model Description

We develop a simple model to capture the essence of our argument. The model is purposefully stylized to highlight the key mechanism rather than to match quantitative features of the data.

Our economy consists of two types of agents: AI owners and a representative household. The AI owners are fully invested in AI assets and are not marginal investors in the stock market. The representative household, on the other hand, is the marginal investor in stocks, so only their consumption matters for our asset pricing analysis. The household has constant relative risk aversion (CRRA) preferences with risk aversion parameter γ and time preference parameter β.

The consumption growth process for the representative household is subject to disaster risk. Specifically, log consumption growth follows:

$$\log \Delta c_{t+1} = \begin{cases} 0 & \text{if no disaster} \\ -b & \text{if disaster (with probability } p\text{)} \end{cases}$$

A disaster in our model 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. While the economy as a whole might boom during such an event, the value created is captured predominantly by AI owners. For the representative household, the disaster manifests as labor being replaced by AI, causing labor income and consumption to plummet. Beyond the direct economic impact, the household also experiences losses in their way of life and sense of meaning, which we capture implicitly through the consumption decline.

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

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

$$D_t = ae^{hN_t}C_t$$

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 determining how much the AI asset grows as a share of the economy when a disaster occurs.

This dividend specification captures the idea that firms providing semiconductors, data, AI models, and related technologies at least partially benefit from sudden improvements in AI capabilities. When a disaster occurs (N_t increases), the AI asset's dividends grow relative to aggregate consumption. This creates a hedge against the consumption disaster from the perspective of the representative household.

3 Asset Pricing Implications

We now derive the asset pricing implications of our model. We are particularly interested in the price-dividend ratio of the AI asset, as this can help us understand whether the potentially high valuations of AI stocks can be explained by their hedging properties.

Price-Dividend Ratio

To determine the price-dividend ratio of the AI asset at time $t = 0$, we apply standard asset pricing techniques. The price equals the expected discounted sum of future dividends under the household's stochastic discount factor. Let P_0 denote the price of the AI asset at time $t = 0$ and D_0 denote its dividend. The price can be written as:

$$P_0 = \mathbb{E}_0 \left[\sum_{k=1}^{\infty} M_{0,k} D_k \right]$$

where $M_{0,k}$ is the household's stochastic discount factor between periods 0 and k. With CRRA preferences, $M_{0,k} = \beta^k (C_k/C_0)^{-\gamma}$. Substituting the dividend process $D_k = ae^{hN_k} C_k$, we obtain:

$$P_0 = \mathbb{E}_0 \left[\sum_{k=1}^{\infty} \beta^k \left(\frac{C_k}{C_0} \right)^{-\gamma} ae^{hN_k} C_k \right]$$

Dividing both sides by $D_0 = aC_0$, we get:

$$\frac{P_0}{D_0} = \mathbb{E}_0 \left[\sum_{k=1}^{\infty} \beta^k \left(\frac{C_k}{C_0} \right)^{-\gamma} e^{hN_k} \frac{C_k}{C_0} \right]$$

Given that $C_k/C_0 = e^{-bN_k}$ (consumption falls by a factor of e^{-b} for each disaster), we can simplify:

$$\frac{P_0}{D_0} = \mathbb{E}_0 \left[\sum_{k=1}^{\infty} \beta^k e^{\gamma b N_k} e^{hN_k} e^{-bN_k} \right] = \mathbb{E}_0 \left[\sum_{k=1}^{\infty} \beta^k e^{N_k(h-b+\gamma b)} \right]$$

Let's define $\theta \equiv h - b(1 - \gamma)$ to simplify notation. Then:

$$\frac{P_0}{D_0} = \mathbb{E}_0 \left[\sum_{k=1}^{\infty} \beta^k e^{\theta N_k} \right]$$

Since N_k follows a binomial distribution with parameters (k, p) (the number of disasters in k periods with disaster probability p), we can compute:

$$\mathbb{E}_0[e^{\theta N_k}] = \sum_{n=0}^{k} \binom{k}{n} p^n (1-p)^{k-n} e^{\theta n} = ((1-p) + pe^{\theta})^k$$

This allows us to express the price-dividend ratio as:

$$\frac{P_0}{D_0} = \sum_{k=1}^{\infty} \beta^k ((1-p) + pe^{\theta})^k$$

Let $A \equiv (1 - p) + pe^{\theta}$. If $\beta A < 1$, this geometric series converges to:

$$\frac{P_0}{D_0} = \frac{\beta A}{1 - \beta A}$$

Substituting back $\theta = h - b(1 - \gamma)$, we obtain our final expression for the price-dividend ratio:

$$\frac{P_0}{D_0} = \frac{\beta[(1 - p) + pe^{h-b(1-\gamma)}]}{1 - \beta[(1 - p) + pe^{h-b(1-\gamma)}]}$$

Intuition and Implications

The price-dividend ratio formula reveals several interesting insights about the valuation of AI assets in our model. The key term is $e^{h-b(1-\gamma)}$, which represents the combined effect of disasters on the relative valuation of AI assets compared to consumption.

When $\gamma > 1$, which is the empirically relevant case for risk aversion, the exponent $h - b(1 - \gamma)$ becomes larger as both h and b increase. Intuitively, this means that higher risk aversion increases the value of the AI asset's hedging properties. The representative household is willing to pay more for assets that deliver relatively higher dividends in disaster states.

The disaster probability p affects valuations in a non-linear way. Even a small probability of disaster can generate a high price-dividend ratio if the AI asset provides a strong hedge (high h) against a severe consumption disaster (high b).

The convergence condition $\beta[(1-p)+pe^{h-b(1-\gamma)}] < 1$ ensures that the price-dividend ratio remains finite. This condition is typically satisfied for moderate parameter values but may be violated when the hedge value of AI assets becomes extremely large relative to consumption disasters.

Numerical Examples

To illustrate the quantitative implications of our model, we compute the price-dividend ratio for different parameter combinations. For our baseline calibration, we set $\beta = 0.96$ and $\gamma = 2$, common values in the asset pricing literature. We fix the AI dividend scaling parameter at $h = 0.2$, meaning that each disaster increases AI dividends by approximately 22% relative to the economy ($e^{0.2} \approx 1.22$).

Table 1 presents the price-dividend ratios for different combinations of the disaster size (b) and probability (p). As the table shows, even modest values of these parameters can generate substantial price-dividend ratios, consistent with the high valuations we observe for AI stocks.

Table 1: Price-Dividend Ratio (P_0/D_0)

	$p = 0.0001$	$p = 0.001$	$p = 0.01$	$p = 0.02$
$b = 0.4$	24.0	24.5	30.3	39.0
$b = 0.6$	24.1	24.8	34.7	61.5
$b = 0.8$	24.1	25.3	40.7	141.9
$b = 0.95$	24.1	25.4	51.6	—

The table reveals several important patterns. First, for very small disaster probabilities ($p = 0.0001$), the price-dividend ratio remains close to the standard Gordon growth formula value, regardless of the disaster size. As the disaster probability increases, the price-dividend ratio becomes increasingly sensitive to the disaster size parameter b.

For instance, with a moderate 1% disaster probability ($p = 0.01$) and a consumption disaster of 80% ($b = 0.8$, meaning consumption falls to 45% of its previous level in a disaster), the price-dividend ratio exceeds 40. With a 2% disaster probability and the same disaster size, the ratio increases dramatically to nearly 142.

These results illustrate our central argument: high valuations of AI stocks may reflect their hedging value against severe AI-driven economic disruptions rather than purely optimistic expectations about future earnings growth. The representative household is willing to pay a premium for assets that provide insurance against scenarios where AI advancements lead to devastating outcomes for average workers.

This hedging explanation does not require extraordinary assumptions about future productivity growth from AI. Instead, it depends on the market's assessment of AI disaster risk (captured by p and b) and the extent to which AI companies benefit during such events (captured by h). If market participants perceive even a modest probability of a negative AI singularity, and believe that AI companies would capture a larger share of the economy in such a scenario, then current high valuations of AI stocks could be rational from a risk-hedging perspective.

4 Model Discussion

Our model, while stylized, captures the essence of how AI assets might serve as hedges against negative AI singularity events. However, several important economic subtleties deserve further discussion.

Market incompleteness plays a crucial role in our analysis, though we do not model it explicitly. The disaster magnitude parameter $b > 0$ implicitly reflects this incompleteness. It represents the net effect of two forces: the negative impact of an AI singularity on the representative household and the partial hedging benefit provided by publicly traded AI assets. If markets were complete, the representative household could purchase shares in all AI assets—both public and private—and potentially not only fully hedge against AI disasters but actually benefit from them. In such a scenario, b would be negative, representing a sudden 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 frontier of AI development, remain largely inaccessible to average investors. This market structure is consistent with our model's assumption that AI owners and the representative household are distinct entities with different exposures to AI progress.

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

However, we believe that such elaborations would primarily decorate speculations with mathematics rather than provide additional economic insights. The core economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain the same. Moreover, a more complex model would be significantly more costly for readers to digest, with potentially limited additional benefits.

As economists, we believe the benefit of reading a paper should exceed the cost. Our parsimonious approach allows us to highlight the key insight—that AI stocks may be valued highly partly because they hedge against negative AI singularity events—without burdening readers with excessive mathematical complexity. This approach also leaves room for the human-written Appendix A, which provides additional context and reflections on the research process itself.

Our model is best viewed as a thought experiment that formalizes a specific economic mechanism. While it abstracts from many real-world complexities, it provides a useful framework for thinking about how AI assets might be priced in a world with singularity risk. The model's simplicity is a feature, not a bug, as it allows us to isolate and understand a potentially important factor in current AI asset valuations.

5 Conclusion

In this paper, we have developed a theoretical framework to explore an alternative explanation for the high valuations of AI stocks. While the prevailing view attributes these valuations to expectations of exceptional future earnings growth, we suggest that they may partially reflect AI assets' role as hedges against a negative AI singularity—a scenario where rapid AI advancement proves devastating for the representative investor.

Our model demonstrates how even a small probability of an AI disaster can generate substantial price-dividend ratios for AI stocks that would perform relatively better in such scenarios. When the market anticipates that AI companies might capture a larger share of economic value during a potential negative singularity, investors are willing to pay a premium for these assets as a form of catastrophe insurance.

This hedging perspective offers insights into potential market-based approaches to managing AI catastrophe risk. Financial markets could, in principle, provide mechanisms for individuals to hedge against AI-driven labor market disruptions, offering an alternative or complement to policy proposals like Universal Basic Income. Through investments in AI-related companies, individuals might partially protect themselves against scenarios where AI advancement significantly devalues human labor.

However, as our model highlights, market incompleteness severely limits this hedging potential. Most households cannot invest in cutting-edge private AI laboratories, and publicly traded AI stocks provide only imperfect hedges. This market incompleteness is reflected in our disaster magnitude parameter b, which captures the net negative effect on the representative household after accounting for any partial hedging through accessible AI investments.

Interestingly, financial market solutions to AI catastrophe risk have received relatively little attention in the literature on AI safety and governance. While authors like Bostrom (2014), Russell (2019), and Bengio et al. (2024) have extensively explored technical and governance approaches to reducing AI risks, the potential role of financial markets in helping individuals manage exposure to these risks remains underexplored. Similarly, economists studying AI's economic impacts, such as Korinek and Stiglitz (2018), Trammell and Korinek (2023), and Jones (2024), have primarily focused on broad economic effects rather than financial market mechanisms for risk management.

Our work suggests that further research on how financial markets might help mitigate individual exposure to AI catastrophe risk could yield valuable insights. This includes exploring how market structures could be improved to reduce incompleteness and better distribute the risks and rewards of AI advancement. Such research would complement existing technical and governance approaches to AI safety, potentially offering individuals additional tools to manage their exposure to the uncertain impacts of transformative AI.

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.

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.

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

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

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

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

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

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

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

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

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) for the remainder.

These new products are a game changer. Both Novy-Marx and Velikov (2025) and Chris Lu et al. (2024) ran into hallucinated citations. 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.

Rather than try to prompt engineer an error free, insightful paper, I decided to just generate N papers and choose the best one.

tbc 5 drafts of the paper can be found in ./manyout0408-pdf/. They're broadly similar. I think I would be OK with my name on all except for one of them. One of them makes the misleading claim that there was "minimal human input."

I ended up choosing paper-run-02.pdf (actually, paper-appendix-update-run02.pdf since it needs to have this README updated). The paper still has some minor issues. It irritates me that it kind of sort of overinterprets the model on page 7. It's definitely not the best paper I've written (that would be Chen and Zimmermann (2020, RAPS)), but I do think it's a paper people will find to be worth reading.

# Lessons about Research

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

Relatedly, the APIs would often barf on me, due to "overloading" or "Bad Gateway." We all feel under the weather sometimes, I suppose.

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\\}). Ensure \\% is converted to \\\\%.
   But as a reminder, AVOID BULLET POINTS AND NUMBERED LISTS.
```

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

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

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

Use the following outline:
  - The model is purposefully simple and captures the essence of the
      main argument
  - Two agents
    - AI owners
      - Fully invested in AI, not marginal investors in stock market
    - Representative household
      - Marginal investor in stocks: only their consumption matters
         for this analysis
      - CRRA = \\gamma, time preference = \\beta
  - Consumption growth
    - \\log \\Delta c\\{t+1\\} = 0 if no disaster
    - \\log \\Delta c\\{t+1\\} = -b if disaster (prob p)
    - A disaster is a sudden improvement in AI that is devastating
        for the household
      - Think of as a worst-case scenario for AI progress
      - Economy booms, but the value of AI is captured by the AI
         owners.
      - For household, labor is replaced by AI, so labor income
         plummets, as does consumption.
        - Also, way of life, meaning, is lost. Consumption fall can
           be thought of as a stand-in for these losses.
    - at t=0, no disasters have happened (singularity has not
       occurred)
      - Multiple disasters may happen, capturing ongoing uncertainty
          if a singularity occurs
  - AI asset
    - Captures publicly traded AI stocks
    - Dividend D\\_t = a e\\^\\{h N\\_t\\} C\\_t
    - Interpretation (discuss in prose)
      - a > 0 is small, AI stocks are currently a minor share of the
          economy
      - N\\_t is the number of disasters that have occurred up to
         and including time t
      - h > 0: each time a disaster occurs, the AI asset grows as a
         share of the economy
      - Intuitively, firms that provide semiconductors, data, AI
         models, etc. at least partially benefit from a sudden
         improvement in AI

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
Write a short "Conclusion" section.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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