Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

This paper explores a novel explanation for high AI stock valuations: they may serve as hedges against a negative AI singularity—an explosion of AI development that is devastating for the representative investor. We develop a parsimonious asset pricing model showing that even with small probabilities of such events, the hedging property of AI assets can generate significantly elevated price-dividend ratios. Our analysis suggests financial markets naturally create mechanisms to partially hedge technological disruption risks, though these solutions are constrained by market incompleteness. We connect our theoretical framework to empirical findings on automation risk in asset pricing. Unlike previous work, this short paper is written by prompting LLMs.

Keywords: Artificial Intelligence, Disaster Risk, Asset Pricing

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

1 Introduction

Artificial intelligence is advancing at an unprecedented pace. The recent release of sophisticated reasoning models like OpenAI's ChatGPT-o1 and DeepSeek's R1, capable of solving complex mathematical problems and writing code with human-like proficiency, marks a significant leap in AI capabilities (Wu et al., 2024; DeepSeek-AI et al., 2025). On the ARC-AGI Challenge—a benchmark designed to measure general intelligence through abstraction and reasoning—scores have rapidly improved from 33% to over 55% in just one year (Chollet et al., 2024). Meanwhile, Waymo's autonomous vehicles have logged tens of millions of miles without human intervention. These developments have intensified investor concerns about wage displacement as AI increasingly demonstrates capabilities that overlap with human cognitive tasks.

While technological change has occurred throughout history, AI represents a fundamentally different kind of innovation. Unlike previous technologies that automated specific physical tasks, there appears to be no product or service that AI could not, in principle, create. This paper itself exemplifies this potential—it was generated entirely by AI through a series of engineered prompts (see `https://github.com/chenandrewy/Prompts-to-Paper/`). More concerning is the possibility of an "AI singularity"—a point where AI capabilities improve so rapidly that they fundamentally transform society, potentially in ways harmful to human welfare (Vinge, 1993). As Bostrom (2014) argues, the emergence of superintelligence could represent an existential risk if not properly aligned with human values.[1]

In this paper, we explore how the possibility of a negative AI singularity might influence AI asset pricing. We develop a parsimonious model in which a representative household faces the risk of a "disaster" in the form of sudden AI advancement that dramatically reduces household consumption. Crucially, this same event benefits AI companies by increasing their share of the economy. We show that this hedging property can lead to elevated price-dividend ratios for AI stocks, even when the probability of such an event is small.

It is important to note two caveats. First, we are not claiming that a negative singularity will necessarily occur. Indeed, AI development could lead to broadly shared economic prosperity. However, considering the downside risk remains important for comprehensive economic analysis. Second, we are not asserting that hedging value is already fully reflected in current AI stock valuations. Our model simply illustrates a possible mechanism through which singularity risk could affect asset prices.

Our paper builds on two strands of literature. First, we draw on the rare disaster frame-

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

work for asset pricing pioneered by Rietz (1988) and Barro (2006), and extended by Gabaix (2012) and Wachter (2013). This literature shows how small probabilities of catastrophic events can generate substantial risk premia in asset markets. We apply this insight to the specific context of AI development, connecting to recent work in economics that examines the potential consequences of rapid AI progress. Jones (2024) analyzes the tension between AI-driven growth and existential risk, while Korinek and Suh (2024) present scenarios for economic transitions as AI approaches human-level capabilities. Simultaneously, we connect to the literature on hedging labor income risk through portfolio choice (Cocco, Gomes, and Maenhout, 2005; Benzoni, Collin-Dufresne, and Goldstein, 2007), but with a focus on technological disruption.

The second strand relates to the impact of AI and automation on asset pricing. Zhang (2019) demonstrates that firms with high routine-task labor maintain a replacement option that hedges against unfavorable macroeconomic shocks, leading to lower expected returns. Knesl (2023) similarly finds that firms with a high share of displaceable labor have negative exposure to technology shocks. More directly related to our work, Babina et al. (2023) provide evidence that firms' investments in AI technologies affect their risk profiles and systematic risk exposure. Our contribution is to propose a novel perspective on AI stock valuations: rather than being driven solely by expectations of future earnings growth, these valuations may partly reflect the hedging value these assets provide against a potential negative AI singularity—a perspective not previously explored in the asset pricing literature.

2 Model

In this paper, we develop a simple model to explore the pricing of AI stocks under the possibility of a negative AI singularity. The model is deliberately parsimonious, focusing on the essential elements needed to capture our main argument while abstracting from many real-world complexities.

Our economy features two types of agents. First, there are AI owners who are fully invested in AI assets. Importantly, these agents are not marginal investors in the stock market. Second, there is a representative household who is the marginal investor in stocks. Only the household's consumption matters for our asset pricing analysis. The household has standard preferences with constant relative risk aversion utility:

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

where β is the time discount factor and γ is the coefficient of relative risk aversion.

The key feature of our model is the consumption growth process for the representative household. In normal times, consumption growth is zero:

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

However, with probability p in each period, a disaster occurs:

$$\log \Delta c_{t+1} = -b \quad \text{if disaster} \quad \text{(with probability } p\text{)} \tag{3}$$

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

In our framework, a disaster represents a sudden improvement in AI that is devastating for the representative household—essentially a worst-case scenario for AI progress. While the economy as a whole may boom during such an event, the value created is captured primarily by AI owners. For the representative household, the disaster manifests as labor being replaced by AI, causing labor income and consequently consumption to plummet. Beyond the direct economic impact, this consumption decline can also be interpreted as a stand-in for other losses such as diminished sense of purpose, meaning, and way of life.

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

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

$$D_t = ae^{hN_t}C_t \tag{4}$$

where $a > 0$ is a small constant reflecting that AI stocks currently constitute a minor share of the economy, N_t is the number of disasters that have occurred up to and including time t, and $h > 0$ is a parameter governing how much the AI asset grows as a share of the economy with each disaster.

The dividend process captures a key intuition: firms that provide the infrastructure for AI—such as semiconductors, data centers, AI models, and related services—at least partially benefit from sudden improvements in AI capability. While these improvements may be devastating for the representative household, they increase the relative importance of AI in the economy, which is reflected in the dividend process through the term e^{hN_t}.

3 Asset Pricing Results

3.1 Price-Dividend Ratio

We now derive the price-dividend ratio for the AI asset in our model. By the standard consumption-based asset pricing approach, the time-0 price of the AI asset satisfies:

$$P_0 = E_0[M_1(D_1 + P_1)] \tag{5}$$

where M_1 is the stochastic discount factor (SDF) from $t = 0$ to $t = 1$, and $D_1 + P_1$ is the asset's payoff (dividend plus ex-dividend price) at $t = 1$. With power utility, the SDF is:

$$M_1 = \beta \left(\frac{C_1}{C_0}\right)^{-\gamma} \tag{6}$$

Because there is a disaster with probability p and no disaster with probability $1 - p$, the SDF takes one of two values:

$$M_1 = \begin{cases} \beta & \text{with probability } 1 - p \text{ (no disaster)} \\ \beta e^{\gamma b} & \text{with probability } p \text{ (disaster)} \end{cases} \tag{7}$$

Similarly, the AI asset's dividend at $t = 1$ follows:

$$D_1 = \begin{cases} aC_0 & \text{with probability } 1 - p \text{ (no disaster)} \\ ae^{h-b}C_0 & \text{with probability } p \text{ (disaster)} \end{cases} \tag{8}$$

To solve for the time-0 price, we posit that the price at any time t must be proportional to consumption and depends on the number of disasters that have occurred. Specifically, we conjecture:

$$P_t = Q(N_t)C_t \tag{9}$$

where $Q(\cdot)$ is a function that maps the number of disasters to a constant of proportionality. At $t = 0$, we have $N_0 = 0$, so $P_0 = Q(0)C_0$.

Substituting our conjecture into the pricing equation and dividing by C_0:

$$Q(N_0) = E_0 \left[\beta \left(\frac{C_1}{C_0} \right)^{-\gamma} \frac{C_1}{C_0} \left(ae^{hN_1} + Q(N_1) \right) \right] \tag{10}$$

In the no-disaster state, $N_1 = N_0 = 0$ and $\frac{C_1}{C_0} = 1$. In the disaster state, $N_1 = N_0 + 1 = 1$ and $\frac{C_1}{C_0} = e^{-b}$. Evaluating the expectation:

$$\begin{align} Q(N_0) &= (1-p)\beta[ae^{hN_0} + Q(N_0)] + p\beta e^{\gamma b} e^{-b}[ae^{hN_0}e^h + Q(N_0+1)e^{-b}] \tag{11} \\ &= (1-p)\beta[ae^{hN_0} + Q(N_0)] + p\beta e^{(\gamma-1)b}[ae^{hN_0}e^h + Q(N_0+1)e^{-b}] \tag{12} \end{align}$$

We can verify that a solution of the form $Q(N) = \mathcal{Q}e^{hN}$ is consistent with this recursive equation, where $\mathcal{Q}$ is a constant independent of N. Substituting this form:

$$\begin{align} \mathcal{Q}e^{hN_0} &= (1-p)\beta[ae^{hN_0} + \mathcal{Q}e^{hN_0}] + p\beta e^{(\gamma-1)b}[ae^{hN_0}e^h + \mathcal{Q}e^{h(N_0+1)}e^{-b}] \tag{13} \\ &= (1-p)\beta[ae^{hN_0} + \mathcal{Q}e^{hN_0}] + p\beta e^{(\gamma-1)b}[ae^{hN_0}e^h + \mathcal{Q}e^{hN_0}e^h e^{-b}] \tag{14} \\ &= (1-p)\beta[ae^{hN_0} + \mathcal{Q}e^{hN_0}] + p\beta e^{(\gamma-1)b}e^h[ae^{hN_0} + \mathcal{Q}e^{hN_0}e^{-b}] \tag{15} \\ & \tag{16} \end{align}$$

Dividing both sides by e^{hN_0}:

$$\begin{align} \mathcal{Q} &= (1-p)\beta[a + \mathcal{Q}] + p\beta e^{(\gamma-1)b}e^h[a + \mathcal{Q}e^{-b}] \tag{17} \\ &= (1-p)\beta[a + \mathcal{Q}] + p\beta e^{(\gamma-1)b+h}a + p\beta e^{(\gamma-1)b+h-b}\mathcal{Q} \tag{18} \\ &= (1-p)\beta[a + \mathcal{Q}] + p\beta e^{(\gamma-1)b+h}a + p\beta e^{(\gamma-1)b+h-b}\mathcal{Q} \tag{19} \\ &= \beta a[(1-p) + pe^{(\gamma-1)b+h}] + \beta\mathcal{Q}[(1-p) + pe^{(\gamma-1)b+h}] \tag{20} \end{align}$$

Solving for $\mathcal{Q}$:

$$\mathcal{Q} = \frac{\beta a[(1-p) + pe^{(\gamma-1)b+h}]}{1 - \beta[(1-p) + pe^{(\gamma-1)b+h}]} \tag{21}$$

At time 0, $N_0 = 0$, so $P_0 = \mathcal{Q}C_0$. Since $D_0 = aC_0$, the price-dividend ratio is:

$$\frac{P_0}{D_0} = \frac{\mathcal{Q}}{a} = \frac{\beta[(1-p) + pe^{(\gamma-1)b+h}]}{1 - \beta[(1-p) + pe^{(\gamma-1)b+h}]} \tag{22}$$

This expression for the price-dividend ratio reveals several important insights about AI asset pricing. First, the ratio will be finite only when the term $\beta[(1-p) + pe^{(\gamma-1)b+h}]$ is less than one. This convergence condition places constraints on the permissible combinations of disaster probability (p), disaster magnitude (b), and AI dividend growth during disasters (h).

The price-dividend ratio increases with both h and p, highlighting that greater AI growth during disasters and higher disaster probability both contribute to higher valuations. Interestingly, the effect of the disaster magnitude b depends critically on the coefficient of relative risk aversion γ. When $\gamma > 1$, which is the empirically relevant case, a larger disaster magnitude b increases the price-dividend ratio. This occurs because for $\gamma > 1$, the increase in the SDF during disasters (which scales with $e^{\gamma b}$) more than offsets the decrease in dividends (which scales with e^{-b}).

This is the key insight of our paper: AI stocks can have high valuations precisely because they provide a hedge against negative AI singularity events. While these events reduce aggregate consumption, they simultaneously increase the relative importance of AI in the economy. For risk-averse investors, this hedging property is valuable, leading to higher price-dividend ratios than would be justified by expected dividend growth alone.

3.2 Quantitative Illustration

To illustrate these effects quantitatively, we present price-dividend ratios for different combinations of disaster probability p and disaster magnitude b. We set $\beta = 0.96$, $\gamma = 2$, and $h = 0.2$ for these calculations. With these parameter values, the convergence condition for a finite price-dividend ratio is $\beta[(1-p) + pe^{b+h}] < 1$, which for $p = 0.01$ requires $b + h < 1.64$.

Table 1 reports the price-dividend ratios for various values of b and p:

Table 1: Price-Dividend Ratios for the AI Asset

b	p			
	0.0001	0.001	0.01	0.02
0.4	24.0	24.5	30.2	39.0
0.6	24.1	24.6	34.7	63.6
0.8	24.1	25.3	40.7	141.9
0.95	24.1	25.3	51.6	–

The table reveals several patterns consistent with our theoretical analysis. First, the price-dividend ratio is increasing in both the disaster probability p and the disaster magnitude b. For very small probabilities (e.g., $p = 0.0001$), the price-dividend ratio is relatively insensitive to changes in the disaster magnitude. However, as the probability increases, the sensitivity to the disaster magnitude grows substantially. For instance, with $p = 0.01$, increasing b from 0.4 to 0.95 raises the price-dividend ratio from 30.2 to 51.6. At $p = 0.02$, the ratio reaches 141.9 for $b = 0.8$ and becomes infinite for $b = 0.95$.

These results highlight how concerns about a negative AI singularity could contribute to elevated valuations for AI stocks. Even relatively small probabilities of a singularity event can generate substantial price-dividend ratios when the event is sufficiently severe and AI assets provide meaningful hedging benefits. This suggests that observed high valuations in AI-related sectors may reflect not just optimism about future growth but also the hedging value these assets provide against potential AI-driven economic disruptions.

While we have kept our model deliberately simple, these findings offer a novel perspective on AI asset pricing that complements conventional explanations based on expected cash flows. The possibility of a negative AI singularity, even if remote, may be partially capitalized into current valuations due to the hedging properties of AI assets in such scenarios.

4 Model Discussion

Our model, while deliberately simplified, captures the essential economic mechanism behind AI asset pricing in the presence of singularity risk. However, it is worth discussing some of the model's subtleties and limitations to better understand its implications and potential extensions.

A key feature of our framework is the implicit market incompleteness, which is not explicitly modeled but is crucial to our results. This incompleteness is embedded in the disaster magnitude parameter $b > 0$, which represents the net effect of two forces: the negative impact of an AI singularity on the representative household and the positive effect on AI asset dividends. The positive parameter b implies that the household cannot fully hedge against the singularity risk.

If markets were complete, the representative household could purchase shares in all AI assets, including private AI companies that are at the cutting edge of development. In such a scenario, the household could not only fully hedge against singularity risk but potentially benefit from it, implying $b < 0$ (a sudden boom rather than a disaster). However, in reality, most households cannot invest in many leading AI laboratories such as OpenAI, Anthropic, xAI, or DeepSeek. This market structure, where the most transformative AI technologies remain privately held, is consistent with our modeling approach.

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

While these questions are fascinating, we believe that addressing them would primarily decorate speculations with mathematics rather than provide additional economic insight. The core economic mechanisms—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, potentially obscuring the main insights.

In our view, the benefit of reading a paper should exceed the cost. Our parsimonious approach allows us to clearly communicate the central insight—that AI stocks may be valuable partly because they hedge against negative AI singularity events—without burdening the reader with unnecessary complexity. This approach also leaves room for the human-written Appendix A, which provides additional context and reflections on the paper's creation process.

That said, we acknowledge that our model abstracts from many real-world complexities. The actual process of AI development and its economic impacts will undoubtedly be more nuanced than our stylized representation. Different AI technologies will affect different sectors and workers in different ways, and the distribution of gains from AI progress will depend on complex institutional arrangements, policy choices, and market structures.

Nevertheless, we believe our simple model captures an important and previously underappreciated aspect of AI asset pricing: the hedging value that AI stocks provide against the very disruptions they might help create. This insight complements traditional explanations based on expected cash flows and offers a novel perspective on the high valuations observed in AI-related sectors.

5 Conclusion

In this paper, we have explored how the possibility of a negative AI singularity might influence AI asset pricing. Our central argument is that the high valuations of AI stocks may be partly explained by their role in hedging against extreme AI development scenarios—a perspective that complements conventional explanations based on future earnings growth. We have shown that even with small probabilities of a singularity event, the hedging property of AI assets can lead to significantly elevated price-dividend ratios, especially when the potential economic disruption is severe.

Our model demonstrates that AI stocks can have high valuations precisely because they provide a hedge against the very disruptions they might help create. While a negative AI singularity would reduce aggregate consumption, it would simultaneously increase the relative importance of AI in the economy. For risk-averse investors, this insurance-like property is valuable, potentially generating price-dividend ratios that exceed what would be justified by expected dividend growth alone.

An important implication of our analysis is that financial markets naturally create mechanisms that help society partially hedge against technological disruption risks. As Jones (2024) explores in his work on AI growth versus existential risk, economies face difficult trade-offs in navigating technological development. Our findings suggest that alongside policy proposals like Universal Basic Income, financial markets can play a constructive role by creating instruments that provide protection against AI-driven economic dislocations.

However, these market-based solutions are inherently constrained by market incompleteness. As noted in our model discussion, most households cannot invest in leading AI laboratories that remain privately held, preventing them from fully hedging against singularity risk. This incompleteness echoes findings by Benzoni et al. (2007), who show that households face significant constraints in hedging their labor income risks through portfolio choices.

Notably, these financial market approaches to managing AI catastrophe risk receive relatively little attention in the AI safety literature. While Bengio et al. (2024) discuss various governance mechanisms for managing extreme AI risks, and Bostrom (2014) explores institutional responses to superintelligence scenarios, the role of financial markets in creating hedging instruments remains underexplored. Similarly, Korinek and Suh (2024) present detailed scenarios for economic transitions to AGI but do not fully examine how financial markets might evolve to help households manage these risks.

Our work connects to empirical findings by Zhang (2019) and Knesl (2023), who show that firms with automation potential have distinct risk characteristics that affect their valuations. Zhang (2019) demonstrates that firms with high routine-task labor maintain a replacement option that hedges against unfavorable macroeconomic shocks, while Knesl (2023) finds that firms with a high share of displaceable labor have negative exposure to technology shocks. These patterns are consistent with our theoretical mechanism.

Future research could explore practical financial innovations that would expand access to AI investment opportunities, potentially reducing market incompleteness. Such work would build on studies like Babina et al. (2023), which examines how firms' investments in AI technologies affect their risk profiles. By better understanding these dynamics, we can help ensure that the benefits of AI progress are more widely shared while mitigating potential harms from technological disruption—a goal that aligns with concerns raised by both AI optimists and skeptics across the political spectrum.

References

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

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

Benzoni, Luca, Pierre Collin-Dufresne, and Robert S. Goldstein (2007). "Portfolio Choice over the Life-Cycle when the Stock and Labor Markets Are Cointegrated". In: *Journal of Finance* 62.5, pp. 2123–2167.

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*.

Cocco, João F., Francisco J. Gomes, and Pascal J. Maenhout (2005). "Consumption and Portfolio Choice over the Life Cycle". In: *Review of Financial Studies* 18.2, pp. 491–533.

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`.

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

Korinek, Anton and 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*.

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

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

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

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

A A Purely Human Perspective

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

```
# Prompts-to-Paper

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

- make-paper.py writes a paper

- plan0408-piecewise.yaml contains the prompts

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

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

-Andrew Chen, April 9, 2025

# Motivation

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

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

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

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

I asked a friend if he would be interested in working on this paper.
Unfortunately, he was busy 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
```