

# Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025\*

## Abstract

The high valuations of AI stocks could be partly explained by their role as hedges against a negative AI singularity—a technological breakthrough that devastates the representative investor’s consumption. We develop a disaster risk model where AI assets increase in value when consumption falls during singularity events. This financial market perspective offers a complementary approach to policy interventions for managing AI catastrophe risk. Our analysis reveals how market incompleteness limits the effectiveness of this hedging mechanism. This short paper was generated by prompting large language models to explore the intersection of asset pricing theory and technological risk.

**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 a remarkable pace. Recent years have witnessed DeepSeek’s R1 reasoning system achieving performance comparable to OpenAI’s o1 (DeepSeek-AI et al., 2025), Anthropic’s Claude models incorporating sophisticated ”thinking” capabilities (Anthropic, 2025), and Google’s Gemini models demonstrating increasingly complex reasoning (DeepMind, 2025). AI systems have conquered benchmarks previously thought to require general intelligence, such as the ARC challenge (Chollet et al., 2024). In autonomous driving, companies like Waymo now operate commercial robotaxi services with minimal human intervention in multiple cities. This acceleration of AI progress has sparked growing concerns among workers about the potential displacement of their skills and livelihoods (Acemoglu and Restrepo, 2020).

While technological change has disrupted labor markets throughout history, AI represents a fundamentally different kind of innovation. Unlike previous technologies that automated specific physical or routine cognitive tasks, there is theoretically no product or service that AI could not, in principle, create or enhance. Indeed, this very paper serves as an example of AI’s expanding capabilities—it was generated entirely through a series of engineered prompts to AI systems, as documented at <https://github.com/chenandrewy/Prompts-to-Paper/>. Moreover, some researchers suggest that progress in AI capabilities may accelerate dramatically in a short period, potentially leading to what has been termed the ”technological singularity”—a hypothetical point where AI systems rapidly improve themselves, leading to profound and potentially unpredictable transformations of human society (Bostrom, 2014; Vinge, 1993).

In this paper, we study how the possibility of such a scenario—particularly one that proves devastating for human labor and consumption—might influence the valuation of AI-related assets. We propose that the high prices of AI stocks could partially reflect their value as hedges against negative AI singularity events, rather than simply expectations of future earnings growth. If rapid advances in AI capabilities significantly reduce the value of human labor, assets tied to AI technologies might maintain or even increase their value precisely when traditional human capital is devalued.<sup>1</sup>

Several important caveats are in order. First, we are not predicting that a negative AI singularity will occur. There are numerous perspectives on AI progress trajectories and their potential impacts, ranging from highly optimistic to deeply concerning (Kurzweil, 2005; Bengio et al., 2024). Our goal is simply to explore the economic implications of one particular

---

<sup>1</sup>”We” in this paper refers to one human author working with multiple large language models. For a purely human perspective on this research, see [Appendix A](#).

risk scenario that investors might reasonably consider. Second, we are not claiming that this hedging value is necessarily already priced into current AI stock valuations. Rather, our model illustrates a plausible mechanism through which rational investors might value AI assets partly as insurance against AI-driven labor displacement.

Our paper contributes to several strands of literature. First, we build on disaster risk models in asset pricing pioneered by Rietz (1988) and extended by Barro (2006), Gabaix (2012), and Wachter (2013). These models demonstrate how rare catastrophic events can generate significant risk premia and explain various asset pricing puzzles. We uniquely apply this framework to technological singularity risk, connecting disaster models with the growing literature on AI economics. Recent work by Jones (2024) and Korinek and Suh (2024) explores how AI progress could lead to radical economic transformations, potentially including steep declines in labor income for many workers. Korinek and Suh (2024) specifically model transition scenarios to artificial general intelligence (AGI), showing conditions under which wages could collapse or flourish, depending on the nature of the technological transition.

Our paper also relates to literature on technology-driven labor displacement and systematic risk. Zhang (2019) demonstrates that firms with routine-task labor maintain a real option to replace workers with automation, hedging against economic downturns and lowering expected returns. Similarly, Knesl (2023) finds firms with displaceable labor have negative exposure to technology shocks, commanding a return premium. More directly relevant to our work, Babina et al. (2023) provide evidence that firms' AI investments affect their systematic risk profiles, though they do not explicitly consider the extreme case of a technological singularity. Our contribution is to propose a novel perspective on AI asset valuation that departs from conventional growth narratives. Rather than viewing high AI stock prices purely as bets on future earnings growth, we suggest they may partially reflect their value as hedges against a negative AI singularity—a market-based insurance mechanism against catastrophic technological risk.

## 2 Model

We now present a simple model to formalize our main argument. The model is deliberately stylized to capture the essence of our hypothesis while remaining tractable. While it abstracts from many real-world complexities, it allows us to isolate the key mechanism through which AI stocks might serve as hedges against negative AI singularity events.

Our economy consists of two types of agents. First, there are AI owners who are fully invested in AI technologies. These agents are not marginal investors in the broader stock market, so their consumption and preferences do not directly affect asset prices in our anal-

ysis.

Second, there is a representative household who is the marginal investor in the stock market. This household has standard preferences with constant relative risk aversion and time preference:

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

where  $\beta$  is the time discount factor and  $\gamma$  is the coefficient of relative risk aversion. Since the representative household is the marginal investor, only their consumption matters for determining asset prices in our framework.

The representative household's consumption growth follows a simple process with disaster risk:

$$\log \Delta C_{t+1} = \begin{cases} 0 & \text{if no disaster occurs} \\ -b & \text{if a disaster occurs} \end{cases} \quad (2)$$

where  $b > 0$  is the magnitude of consumption decline during a disaster, and disasters occur with probability  $p$  each period.

Importantly, in our context, a "disaster" represents a sudden improvement in AI capabilities that proves devastating for the representative household. This can be thought of as a worst-case scenario for AI progress—a technological singularity that fundamentally transforms the economy. While the economy as a whole may boom during such events, the value created is primarily captured by AI owners. For the representative household, this scenario entails substantial displacement of human labor by AI, leading to a significant drop in labor income and consumption.

The consumption decline parameter  $b$  can also be interpreted more broadly to encompass non-monetary welfare losses, such as diminished sense of purpose, meaning, or autonomy that might accompany a world where AI has surpassed human capabilities in most domains.

At time  $t = 0$ , we assume no disasters have yet occurred—the AI singularity has not yet materialized. However, our model allows for multiple disaster events over time, capturing the ongoing uncertainty about AI development even after initial breakthroughs.

We model publicly traded AI stocks as a claim on a dividend stream  $D_t$  that is related to aggregate consumption and disaster occurrences:

$$D_t = a \exp(hN_t) C_t \quad (3)$$

where  $a > 0$  is a scaling parameter,  $N_t$  is the number of disaster events that have

occurred up to and including time  $t$ , and  $h > 0$  governs how the AI asset's share of the economy changes with disasters.

This specification has several important interpretations. First,  $a$  is assumed to be relatively small, reflecting that AI stocks currently constitute a modest portion of the overall economy. Second, the term  $\exp(hN_t)$  captures how the AI asset's dividend grows relative to aggregate consumption when disaster events occur. The parameter  $h > 0$  means that each time a disaster (AI singularity event) occurs, the AI asset grows as a share of the economy.

This formulation reflects our intuition that firms providing the infrastructure for AI advancement—such as semiconductor manufacturers, data providers, cloud computing services, and AI model developers—would at least partially benefit from sudden improvements in AI capabilities, even if these improvements prove disruptive for the broader economy and harmful to the representative household.

### 3 Asset Pricing Implications

Having described the basic economic environment, we now analyze the equilibrium price of the AI asset. This will allow us to explore how the hedging properties of AI stocks against singularity events might contribute to their valuations.

The price of the AI asset at time 0, denoted by  $P_0$ , equals the expected present value of all future dividends, discounted by the representative household's stochastic discount factor:

$$P_0 = E_0 \left[ \sum_{t=1}^{\infty} M_{0,t} D_t \right] \quad (4)$$

where  $M_{0,t}$  represents the stochastic discount factor between periods 0 and  $t$ . With CRRA preferences,  $M_{0,t} = \beta^t (C_t/C_0)^{-\gamma}$ .

Given our consumption process and dividend specification, we can express the discount factor and dividends in terms of the number of disasters  $N_t$  that have occurred by time  $t$ :

$$M_{0,t} = \beta^t e^{b\gamma N_t} \quad \text{and} \quad D_t = aC_0 e^{(h-b)N_t} \quad (5)$$

Substituting these expressions into the pricing equation:

$$P_0 = aC_0 \sum_{t=1}^{\infty} \beta^t E_0 \left[ e^{(h+b(\gamma-1))N_t} \right] \quad (6)$$

With disasters occurring independently with probability  $p$  in each period,  $N_t$  follows a binomial distribution. This allows us to compute:

$$E_0 \left[ e^{(h+b(\gamma-1))N_t} \right] = \left( pe^{h+b(\gamma-1)} + (1-p) \right)^t \quad (7)$$

For notational simplicity, let  $Z \equiv pe^{h+b(\gamma-1)} + (1-p)$ . The price then simplifies to:

$$P_0 = aC_0 \sum_{t=1}^{\infty} (\beta Z)^t = aC_0 \frac{\beta Z}{1 - \beta Z} \quad (8)$$

provided that  $\beta Z < 1$  to ensure convergence. Since the dividend at time 0 is  $D_0 = aC_0$ , the price-dividend ratio equals:

$$\frac{P_0}{D_0} = \frac{\beta Z}{1 - \beta Z} = \frac{\beta \left( pe^{h+b(\gamma-1)} + (1-p) \right)}{1 - \beta \left( pe^{h+b(\gamma-1)} + (1-p) \right)} \quad (9)$$

This formula provides several key insights into AI asset valuation. First, the price-dividend ratio increases with  $h$ , the parameter that determines how much the AI asset benefits from disasters. When  $h$  is larger, AI stocks become more valuable hedges against negative singularity events.

Second, the price-dividend ratio typically increases with  $b$ , the magnitude of consumption decline during disasters. As potential disasters become more severe, assets that provide insurance against such events become more valuable to risk-averse investors.

Third, as  $p$  increases—meaning disasters become more likely—the AI asset becomes more valuable to the representative household, driving up its price-dividend ratio.

To illustrate these effects quantitatively, Table 1 shows how the price-dividend ratio varies with disaster probability ( $p$ ) and disaster magnitude ( $b$ ) while holding  $h = 0.2$  fixed and assuming  $\gamma = 2$  and  $\beta = 0.96$ .

Table 1: Price-Dividend Ratio for AI Asset

Disaster Magnitude ( $b$ )	Disaster Probability ( $p$ )			
	0.0001	0.001	0.01	0.02
0.4	24.03	24.52	30.25	39.00
0.6	24.08	24.77	34.71	59.68
0.8	24.11	25.06	40.67	141.86
0.95	24.13	25.36	51.63	-

*Notes:* The table shows the price-dividend ratio (P/D) for the AI asset across different values of disaster magnitude ( $b$ ) and disaster probability ( $p$ ). Parameters are fixed at  $h = 0.2$ ,  $\gamma = 2$ , and  $\beta = 0.96$ . The dash indicates that the convergence condition  $\beta Z < 1$  is violated.

The table demonstrates how the price-dividend ratio increases substantially as either the probability or severity of disasters increases. For example, with  $b = 0.4$  (representing a 33% decline in consumption during a disaster), the price-dividend ratio increases from 24.03

when  $p = 0.0001$  to 39.00 when  $p = 0.02$ . Similarly, for a fixed probability  $p = 0.01$ , the price-dividend ratio increases from 30.25 when  $b = 0.4$  to 51.63 when  $b = 0.95$ .

These results support our main hypothesis: the high valuations of AI stocks could reflect their role as hedges against negative AI singularity events. Even with relatively low probabilities of disaster, investors may be willing to pay a substantial premium for assets that would maintain or increase their value precisely when human labor and traditional assets are devalued by transformative AI advances. This hedging motive could contribute significantly to the current high valuations observed for AI-related companies.

## 4 Model Discussion

Our model, while deliberately stylized, captures the key mechanism through which AI stocks might serve as hedges against negative AI singularity events. However, several important subtleties deserve further discussion to properly contextualize our findings.

Market incompleteness plays a crucial role in our analysis, though we do not explicitly model it. This incompleteness is implicitly captured by the disaster magnitude parameter  $b > 0$ , which represents the net effect on the representative household’s consumption after accounting for both the negative impact of an AI disaster and any offsetting benefits from publicly traded AI assets. If markets were complete, the representative household could purchase shares in all AI assets—including private AI ventures—and potentially not only fully hedge against AI disasters but actually benefit from them. In such a scenario, we would expect  $b < 0$ , indicating that AI progress would represent a consumption boom rather than a disaster for the representative household.

The reality, however, is that most households cannot invest in many cutting-edge AI laboratories such as OpenAI, Anthropic, xAI, or DeepSeek. These entities, which are at the frontier of AI research and would likely capture significant value during a technological singularity, remain largely inaccessible to ordinary investors. This market structure is consistent with our modeling approach, where the representative household can only partially hedge against AI disasters through publicly traded AI stocks.

A more elaborate model could certainly add detail to the AI owners, private AI assets, and their interactions with the representative household. Such a model might address questions like how AI progress specifically displaces the representative household’s wages, how AI owners’ incentives affect both AI progress and market incompleteness, or how preferences and technology parameters influence the probability of a negative singularity. However, we must ask whether such extensions would genuinely enhance our understanding or merely decorate speculations with mathematical formalism.

The core economic insights—the role of rare disaster risk, hedging motives, and market incompleteness—would remain fundamentally unchanged in a more complex model. Moreover, such elaborations would significantly increase the cognitive burden on readers. We believe that the benefit of reading a paper should exceed its cost, and our parsimonious approach strikes a balance between rigor and accessibility.

Additionally, keeping our model analysis concise allows room for the human-written Appendix A, which provides valuable context on the motivation and process behind this paper. This appendix offers insights into the evolving relationship between AI and economic research that complement our formal analysis.

In sum, while our model makes simplifying assumptions, it captures the essential economic mechanisms at play. The key insight—that AI stocks may command high valuations partly because they hedge against negative AI singularity events—emerges clearly from our analysis without requiring additional complexity. Future research may build upon this foundation to explore more nuanced aspects of AI valuation and risk hedging as our understanding of these technologies and their economic impacts continues to evolve.

## 5 Conclusion

In this paper, we have presented a novel perspective on the high valuations of AI stocks. While conventional wisdom attributes these valuations primarily to expectations of future earnings growth, we argue that they may also reflect a hedging premium against negative AI singularity events. Our stylized model demonstrates how publicly traded AI assets could serve as partial insurance against scenarios where rapid AI development proves devastating for the representative household’s consumption.

The key insight from our analysis is that even with relatively low probabilities of disaster, investors might be willing to pay a substantial premium for assets that would maintain or increase their value precisely when human labor is devalued by transformative AI advances. This is evidenced by the price-dividend ratios in our numerical examples, which increase substantially as either the probability or severity of disaster increases.

Financial markets naturally provide mechanisms for risk sharing that may help address some concerns about AI catastrophe risk. The ability to invest in AI-related companies offers a market-based alternative to policy interventions like universal basic income by allowing individuals to hedge against labor displacement risk. As Zhang (2019) and Knesl (2023) demonstrate for automation generally, technological assets can provide insurance against labor market disruptions.

However, market incompleteness significantly limits the effectiveness of this hedging



mechanism. As our model highlights, most households cannot invest in cutting-edge AI laboratories that would likely capture the most value during a technological singularity. This incompleteness means that publicly traded AI stocks offer only partial insurance against negative AI scenarios, leaving residual risk for the representative household.

Interestingly, this financial market perspective on AI risk management remains underexplored in the broader AI safety literature. While scholars such as Bostrom (2014), Russell (2019), and Bengio et al. (2024) have extensively analyzed technical alignment challenges and governance frameworks, the role of financial markets in distributing AI transition risks has received less attention. Similarly, economists studying AI impacts like Korinek and Suh (2024) and Jones (2024) have primarily focused on productivity, wages, and growth rather than asset pricing implications.

By bridging the disaster risk literature in finance with AI economics, our paper contributes to a more comprehensive understanding of how markets price technological transition risk. Future research might explore more detailed specifications of AI disaster scenarios, empirical tests of the hedging hypothesis, or policy frameworks that acknowledge both the limitations and strengths of market-based approaches to managing AI transition risks.

## References

- Acemoglu, Daron and Pascual Restrepo (2020). “Robots and Jobs: Evidence from US Labor Markets”. In: *Journal of Political Economy*.
- Anthropic (Feb. 2025). *Claude 3.7 Sonnet*. Anthropic, Wikipedia. URL: <https://www.anthropic.com/claude>.
- 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*.
- DeepMind, Google (Mar. 2025). *Gemini 2.5: Our newest Gemini model with thinking*. Google Blog. URL: <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>.

- DeepSeek-AI et al. (Jan. 2025). “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning”. In: *arXiv*. URL: <https://arxiv.org/abs/2501.12948>.
- 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.
- Kurzweil, Ray (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking Press.
- Rietz, Thomas (1988). “The Equity Risk Premium: A Solution?” In: *Journal of Monetary Economics*.
- Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.
- Vinge, Vernor (1993). “The Coming Technological Singularity”. In: *Department of Mathematical Sciences, San Diego State University*.
- Wachter, Jessica A. (2013). “Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?” In: *Journal of Finance*.
- Zhang, Miao Ben (2019). “Labor-Technology Substitution: Implications for Asset Pricing”. In: *Journal of Finance* 74.4, pp. 1793–1839.

# A A Purely Human Perspective

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

## # Prompts-to-Paper

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

- `make-paper.py` writes a paper
- `plan0408-piecewise.yaml` contains the prompts
- `make-many-papers.py` runs `make-paper.py` many times.

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

## # 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.

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. Then, [Jan 15 episode of Machine Learning Street Talk with Yoshua Bengio](#) came out. 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 much 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." Subtle style issues and instructions on how to use LaTeX were particularly hard to communicate. Both led to 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` 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).

`plan0403-streamlined.yaml` tries to write a paper in just six prompts. 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. Yet somehow 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 10 prompts.

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.

## ## Literature Reviews

A key step was generating lit reviews (`./lit-context/`) to give the AI

context. I used ChatGPT's Deep Research (launched Feb 2025) until I ran out of credits. Claude Web Search (launched March 2025, after I began the project) did the remainder.

These new products were 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 occasionally 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 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 for my story (even if they did not make sense).

I briefly tried having Llama 3.1 470b do the writing. It was terrible! It would be extremely difficult to generate a paper worth reading that way.

I did not try many other models, in order to get this paper out quickly. Gemini 2.5's release, at the end of March 2025, was \*hype\*. I tried it out briefly and was impressed. But I gritted my teeth and ignored it. I'd never get the paper finished if I wanted to really try to explore alternative models.

## ## Picking the best of N papers

The writing quality varies across each run of the code. Some drafts, I found quite insightful! Others, had flagrant errors in the economics.

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

5 drafts of the paper can be found in `./manyout0408-pdf/`. tbc

## # Lessons about Research

A common response to [Novy-Marx and Velikov \(2025\)](#) is that "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, item 3 made me quite cautious about 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 7, 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, 3.7 Sonnet sometimes 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 their own (at least not satisfying to me). I tried asking o1 and Sonnet to generate a model to illustrate the point I'm trying to make. 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.

AI can do both. But both require

There could be models with capabilities that I missed. But my sense is that ChatGPT-o1 and Claude 3.7 Sonnet are close to the best for producing economic research.

But more importantly, how long will these limitations last?

### ## The Future of AI and Economics Research

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 (or perhaps something better).

"Economics on tap" could be a disaster for the economics labor market. It would certainly mean that AI is an extremely cheap substitute for at least some economists' labor. It's possible that this would 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 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 many projects.

I'm not saying that I *expect* a disaster for the economics labor market. But it's definitely a scenario that economists should think about.

## B Prompts Used to Generate This Paper

Each prompt consists of context and instructions. The context consists of the responses to the previous prompts, and may include literature reviews (all AI generated). For writing tasks (using Claude 3.7 Sonnet), a system prompt is also included.

For further details, see <https://github.com/chenandrewy/Prompts-to-Paper/>.

The system prompt and instructions are listed below.

### System Prompt (model: claude-3-7-sonnet-20250219)

You are an asset pricing theorist who publishes in the top journals (Journal of Finance, Journal of Financial Economics, Review of



Financial Studies). You think carefully with mathematics and check your work, step by step.

Your team is writing a paper with the following main argument: the high valuations of AI stocks could be in part because they hedge against a negative AI singularity (an explosion of AI development that is devastating for the representative investor). This contrasts with the common view that AI valuations are high due to future earnings growth. Since the AI singularity is inherently unpredictable, the paper is more qualitative than quantitative. The goal is to just make this point elegantly.

Write in prose. Avoid bullet points and numbered lists. Use display math to highlight key assumptions. Cite papers using Author (Year) format. Use we / our / us to refer to the writing team.

Be conversational yet rigorous. Favor plain english. Be direct and concise. Remove text that does not add value.

Be modest. Do not overclaim.

Output as a latex input (no document environment). Ensure bullet points are formatted in latex (`\\begin\\{itemize\\}` `\\item "blah"` `\\item "blah"` `\\end\\{itemize\\}`). Ensure numbered lists are formatted in latex (`\\begin\\{enumerate\\}` `\\item "blah"` `\\item "blah"` `\\end\\{enumerate\\}`). But as a reminder, AVOID BULLET POINTS AND NUMBERED LISTS.

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

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

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

Use the following outline:

- The model is purposefully simple and captures the essence of the main argument

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

Do not:

- Use bullet points or numbered lists

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

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

Do not:

- Restate the assumptions
- Assume any variable is constant or stationary (prove it)

Try to make the final formulas self-contained and not depend on the other final formulas.

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

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

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

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

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

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

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

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

Discuss intuition behind price/dividend ratio, and relate the intuition to the main argument (AI valuations may be high because they hedge against a negative AI singularity).

This is the key text of the paper. Conclude the text by using the table to make the main argument.

Style notes:

- The table should be clean and simple.
- Do not repeat information in '01-model-prose'.

Do not:

- Emphasize the infinite price/dividend ratio. That's not important.

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

Write the "Model Discussion" section. Discuss the following subtleties of the model in prose (no math):

- Market incompleteness is not explicitly modeled but important
  - Implicit in the disaster magnitude  $b > 0$
  - 'b' is the *net* effect of (1) AI disaster and (2) AI asset dividend
  - If markets were complete, representative household could buy shares in all AI assets (including private AI assets), and not only fully hedge but benefit from the singularity, implying  $b < 0$  (a sudden boom, not a disaster)
  - In reality, most households cannot buy shares in many cutting edge labs (e.g. OpenAI, Anthropic, xAI, DeepSeek), consistent with our model
- A more elaborate model would add detail to the AI owners, private AI assets, and their interactions with the representative household
- It could address questions like:

- How does AI progress displace the representative household's wages?
  - How do AI owners' incentives affect AI progress and market incompleteness?
  - How do preferences and technology parameters affect the odds of a negative singularity?
  - But wouldn't this just decorate speculations with math?
  - The core economics (rare disaster risk, hedging motives, market incompleteness) will remain the same
  - It would also be much more costly to read
    - In our view, the benefit of reading a paper should exceed the cost
  - A short model analysis allows room for the human-written
- Appendix \\ref\\{app:readme\\}

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

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

Be sure to cite:

- Jones (2024) "AI Dilemma" and Korinek and Suh (2024) "Scenarios"
- Babina et al (2023) "Artificial Intelligence and Firms' Systematic Risk"
- Zhang (2019) "Labor-Technology"

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

Write a short "Conclusion" section.

- Review the main argument
- End by discussing financial market solutions to AI catastrophe risk
  - These solutions are an alternative to UBI
    - Key economics: this hedge is limited by market incompleteness
  - These solutions are not discussed enough in the AI catastrophe risk literature (cite papers)
- Be very centrist

- Avoid politically-charged topics: sovereign wealth funds, industrial policy, redistribution, extolling free markets

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

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

Start with background. Describe how AI progress is happening quickly (Deepseek, ARC AGI, Waymo), and investors may be concerned about their wages being displaced.

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

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

Add caveats:

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

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

Cite papers as appropriate. Ensure citations correspond to items from bibtex-all.bib.

## Instruction: 09-full-paper (model: claude-3-7-sonnet-20250219)

Write a short paper titled "Hedging the AI Singularity" based on the "prose" context.

Add an abstract of less than 100 words, not indented. The abstract should:

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

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.