

# Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025\*

## Abstract

We propose that high AI stock valuations may partly reflect their role as hedges against negative AI singularity risk—an explosion of AI development that proves devastating for the representative investor. While conventional explanations focus on future earnings growth, our model demonstrates how even small probabilities of AI-driven disasters can significantly increase valuations of firms that would benefit from such events. Financial markets naturally create risk-sharing mechanisms that could help manage AI transition risks, though these solutions are limited by market incompleteness. This paper was generated entirely through prompts to large language models, demonstrating both the capabilities and limitations of current AI systems.

**Keywords:** Artificial Intelligence, Disaster Risk, Asset Pricing

---

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

# 1 Introduction

Recent advances in artificial intelligence have been remarkably swift. In early 2025, DeepSeek released its R1 model, achieving reasoning capabilities comparable to those of OpenAI’s o1 (DeepSeek-AI et al., 2025). The ARC AGI benchmark, long considered one of the most important tests of general intelligence, has seen scores rise from 33% to over 55% in just one year (Chollet et al., 2024). Meanwhile, autonomous vehicles from companies like Waymo now navigate complex urban environments without human supervision. These developments are not merely academic curiosities; they represent tangible progress toward machines that can perform increasingly sophisticated cognitive tasks—tasks previously thought to require human intelligence.

Technological change is not new, of course. From the printing press to the internet, humanity has experienced many disruptive innovations. However, AI differs fundamentally from previous technological revolutions. Earlier innovations were typically domain-specific, augmenting human capabilities in particular areas while leaving others untouched. AI, by contrast, has the potential to create virtually any product or service that humans can create—and perhaps many that humans cannot. There is no clear boundary to its capabilities. Indeed, this very paper demonstrates AI’s versatility—it was generated entirely by AI systems, guided by a series of engineered prompts. The code and prompts are available at <https://github.com/chenandrewy/Prompts-to-Paper/>.<sup>1</sup>

Beyond its breadth, AI progress may also be characterized by its potential for sudden acceleration—what has been termed the “AI singularity.” This concept refers to a hypothetical point at which AI systems become capable of improving themselves, potentially leading to an explosion of intelligence that far surpasses human capabilities (Bostrom, 2014; Vinge, 1993). Whether such a singularity would benefit or harm humanity remains an open question, but the possibility cannot be dismissed.

In this paper, we study how AI stocks are priced given the risk that AI development may lead to devastating outcomes for the representative investor. We argue that AI companies might serve as natural hedges against negative AI singularity risk. This contrasts with the conventional view that AI valuations are high primarily due to expectations of future earnings growth. Our model demonstrates that even a small probability of an AI-driven disaster can significantly increase the valuations of firms that would disproportionately benefit from such events, as investors seek to insure against potential consumption declines.

We emphasize two important caveats. First, we are not predicting that a negative AI

---

<sup>1</sup>Throughout this paper, “we” refers to one human author working with multiple large language models. A purely human perspective on this research is provided in [Appendix A](#).

singularity will occur. Experts disagree vehemently about the likelihood of catastrophic AI outcomes (Bengio et al., 2024). Nevertheless, even low-probability tail events can have significant asset pricing implications, as the rare disaster literature has shown (Rietz, 1988; Barro, 2006; Wachter, 2013). Second, we are not claiming that this hedging value is already fully priced into current AI stock valuations. Our model merely illustrates a potential mechanism through which singularity risk could affect asset prices, complementing traditional growth-based explanations.

Our paper connects several emerging strands of literature on artificial intelligence, labor risk, and asset pricing. Recent work has examined how labor-replacing technology affects asset prices, with Zhang (2019) showing that firms with routine-task labor that could be automated maintain a replacement option that hedges their value during economic downturns, leading to lower expected returns. This relationship between technological adoption and risk is further developed by Babina et al. (2023), who provide direct evidence that firms' investments in AI technologies affect their systematic risk profiles, with implications for how markets price AI-related assets. These findings build on the broader literature on rare disaster risk in asset pricing, which has shown that small probabilities of catastrophic events can have significant effects on equilibrium prices (Rietz, 1988; Barro, 2006; Wachter, 2013).

The economic literature has increasingly focused on potential catastrophic risks from advanced AI systems. Jones (2024) explores the fundamental tension between AI-driven growth and existential risk, providing an economic framework for analyzing the trade-offs between technological progress and catastrophic outcomes. Similarly, Korinek and Suh (2024) analyze different scenarios for the transition to artificial general intelligence (AGI), examining potential effects on economic output and wage distribution. Our paper contributes to this literature by proposing a novel mechanism: that AI stock valuations may reflect their role as hedges against negative AI singularity events. While the existing literature has focused on either the benefits of AI technology adoption or its potential systemic risks, our approach uniquely combines these perspectives to offer a new lens for understanding AI asset valuations that complements the conventional growth explanation.

## 2 Model

We present a simple model to capture the essence of our argument. While the model abstracts from many real-world complexities, it provides a tractable framework to analyze how AI singularity risk might affect asset prices.

Our economy features two types of agents. First, there are AI owners who are fully invested in AI assets. These owners are not marginal investors in the broader stock market

and thus do not directly affect equilibrium pricing of non-AI assets. Second, there is a representative household who is the marginal investor in the stock market. The household has constant relative risk aversion (CRRA) preferences with risk aversion parameter  $\gamma$  and time discount factor  $\beta$ . Since the household is the marginal investor, only their consumption matters for our asset pricing analysis.

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

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

However, with probability  $p$ , a disaster occurs, causing consumption to drop significantly:

$$\log \Delta c_{t+1} = -b \text{ if disaster}$$

In our context, a disaster represents a sudden improvement in AI technology that proves devastating for the representative household. This can be thought of as a worst-case scenario for AI progress. While the economy as a whole might boom during such an event, the value created is captured primarily by AI owners. For the representative household, the disaster manifests as AI replacing human labor, causing labor income and consequently consumption to plummet. Beyond the direct economic impact, such a scenario might also entail loss of meaning, purpose, and way of life for many individuals. The consumption drop  $b$  can be interpreted as capturing both the economic and non-economic welfare losses.

At time  $t = 0$ , we assume no disasters have yet occurred, meaning the AI singularity has not yet happened. Our model allows for multiple disasters to occur over time, reflecting the ongoing uncertainty about AI development even after initial breakthroughs.

We model publicly traded AI stocks as a claim on a dividend stream  $D_t$ . The dividend process is given by:

$$D_t = a \exp^{hN_t} C_t$$

where  $a > 0$  is a small constant reflecting that AI stocks currently represent a minor share of the economy,  $N_t$  is the number of disasters (AI breakthroughs) that have occurred up to and including time  $t$ , and  $h > 0$  is a parameter governing how much AI assets benefit from each disaster.

This specification captures the intuition that firms providing the infrastructure for AI—semiconductors, data centers, AI models, etc.—at least partially benefit from sudden improvements in AI technology. Each time a disaster occurs, AI assets grow as a share of the overall economy, reflected in the exponential term  $\exp^{hN_t}$ . The parameter  $h$  measures the extent to which AI

assets benefit from disasters relative to the representative household's consumption.

### 3 Asset Pricing Implications

Having specified the model, we now analyze its asset pricing implications, focusing on the price-dividend ratio of AI stocks and their risk premium.

The price of the AI asset at time  $t = 0$  can be derived using the standard consumption-based asset pricing framework. Let  $P_0$  denote the ex-dividend price of the AI asset at time 0, and  $D_0$  its current dividend. The representative household's stochastic discount factor between periods 0 and 1, denoted by  $M_1$ , follows from their CRRA preferences:

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

With this stochastic discount factor, the asset's price satisfies:

$$P_0 = E[M_1(P_1 + D_1)]$$

where the expectation is taken over the possible states at time 1 (disaster versus no disaster).

In the no-disaster state, which occurs with probability  $1 - p$ , consumption remains unchanged ( $C_1 = C_0$ ), so  $M_1 = \beta$ . In the disaster state, which occurs with probability  $p$ , consumption falls ( $C_1 = e^{-b}C_0$ ), so  $M_1 = \beta e^{\gamma b}$ .

For the dividend process, recall that  $D_t = ae^{hN_t}C_t$ . At time 0, with  $N_0 = 0$ , we have  $D_0 = aC_0$ . At time 1, there are two possibilities: - With probability  $1 - p$ , no disaster occurs, so  $N_1 = 0$  and  $D_1 = aC_0$  - With probability  $p$ , a disaster occurs, so  $N_1 = 1$  and  $D_1 = ae^{h-b}C_0$

To solve for the price, we need to account for the recursive nature of the problem. The price at any point depends on the number of disasters that have occurred. Let  $P_n$  denote the price after  $n$  disasters. Through recursion, one can show that the solution takes the form:

$$P_n = xe^{hn}C_0$$

for some constant  $x$ . In particular,  $P_0 = xC_0$ . Solving for  $x$  and computing the price-dividend ratio yields:

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

This expression represents the time-0 price-dividend ratio for the AI asset in terms of model parameters.

The derived price-dividend ratio reveals a key insight: AI assets can have high valuations not only because of expected growth in their future dividends but also because they serve as a hedge against AI-driven disasters. This is evident in the numerator term  $pe^{\gamma b+h-b}$ , which represents the contribution to the asset’s value from the disaster state.

When a disaster occurs, the representative household’s consumption drops by a factor of  $e^{-b}$ , making the household highly willing to pay for assets that perform well in such states. The AI asset provides precisely such a hedge—its dividend increases by a factor of  $e^h$  relative to the economy in a disaster. This hedging value is more significant when: 1. Disasters are more severe (higher  $b$ ) 2. The representative investor is more risk-averse (higher  $\gamma$ ) 3. AI assets benefit more from disasters (higher  $h$ )

For the price-dividend ratio to be well-defined (i.e., for the infinite sum of discounted expected future dividends to converge), we need:

$$\beta((1-p) + pe^{\gamma b+h}) < 1$$

This condition ensures that the denominator in our price-dividend ratio formula remains positive.

To illustrate how our model can generate high valuations for AI assets, we compute the price-dividend ratio for different parameter values. Table 1 presents these results for various combinations of disaster probability ( $p$ ) and consumption drop size ( $b$ ), while fixing  $h = 0.2$ ,  $\beta = 0.96$ , and  $\gamma = 2$ .

Table 1: Price/Dividend Ratio at  $t = 0$

| $b$  | $p$ -values |       |       |        |
|------|-------------|-------|-------|--------|
|      | 0.0001      | 0.001 | 0.01  | 0.02   |
| 0.40 | 24.12       | 25.04 | 40.33 | 121.88 |
| 0.60 | 24.17       | 25.99 | 91.00 | —      |
| 0.80 | 24.29       | 27.36 | —     | —      |
| 0.95 | 24.44       | 29.15 | —     | —      |

The table reveals several important patterns. First, even with a small probability of disaster ( $p = 0.01$ ), the price-dividend ratio can be substantially higher (40.33 for  $b = 0.40$ ) than in a model without disaster risk ( $p = 0$ , which would yield a ratio of approximately 24). Second, as the severity of the disaster ( $b$ ) increases, the price-dividend ratio becomes more sensitive to the disaster probability. For  $b = 0.60$  and  $p = 0.01$ , the ratio rises to 91.00, more than double the value for  $b = 0.40$  at the same probability.

These results support our main argument: the high valuations observed for AI stocks may partially reflect their role as hedges against AI singularity risk rather than simply expectations of future earnings growth. Even small probabilities of severe AI-related disasters can significantly increase the valuation of assets that benefit from such events.

The mechanism is intuitive. In states where AI development leads to devastating outcomes for most economic actors, companies providing AI infrastructure or services may capture a disproportionate share of the economic pie. Forward-looking investors, recognizing this hedging value, would be willing to pay a premium for these assets, leading to higher price-dividend ratios than conventional growth expectations would justify.

## 4 Model Discussion

Our model, while deliberately simplified, captures the core economic mechanism through which AI singularity risk affects asset prices. However, several important subtleties deserve further discussion to clarify the model’s implications and limitations.

A crucial feature of our model is the implicit assumption of market incompleteness, which manifests in the disaster magnitude parameter  $b > 0$ . This parameter represents the net effect of two forces: (1) the negative impact of an AI singularity on the representative household and (2) the positive impact on AI asset dividends. The fact that  $b$  remains positive indicates that the representative household cannot fully hedge against AI singularity risk.

If markets were complete, the representative household could purchase shares in all AI assets—both public and private—and thereby not only hedge against but potentially benefit from an AI singularity. In such a scenario,  $b$  would likely be negative, representing a sudden boom rather than a disaster for the representative household. The household would own sufficient claims on AI technology to offset any displacement of their labor income.

However, reality aligns more closely with our incomplete markets assumption. Most households cannot directly invest in many cutting-edge AI laboratories such as OpenAI, Anthropic, xAI, or DeepSeek. These private companies, which are at the frontier of AI development, remain largely inaccessible to ordinary investors. Even when some of these companies have public investment vehicles (e.g., Microsoft’s stake in OpenAI), the ownership is diluted and indirect. This market incompleteness means that the representative household remains exposed to significant unhedgeable risk from AI progress.

One might reasonably ask whether a more elaborate model would yield additional insights. Such a model could explicitly represent AI owners as separate agents, detail the structure of private AI assets, and specify the interactions between these entities and the representative household. It could address questions such as:

How does AI progress specifically displace the representative household’s wages? How do AI owners’ incentives affect the pace and direction of AI development? How might these incentives contribute to market incompleteness? How do preference parameters and technological constraints affect the probability of a negative singularity?

While these questions are undoubtedly important, we believe that elaborating our model to address them would primarily serve to decorate speculations with mathematics. The fundamental economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain unchanged. More complex models might create an illusion of precision where none exists, given the inherent unpredictability of transformative AI development.

Moreover, there is a pragmatic consideration: model complexity imposes cognitive costs on readers. A more elaborate model would be significantly more costly to read and understand, potentially obscuring the core insight. In our view, the benefit of reading a paper should exceed the cost. Our streamlined approach allows us to communicate the essential economic mechanism without unnecessary technical complications.

This approach also leaves room for the human-written Appendix A, which provides valuable context on the motivations behind this research and reflections on the research process itself. We believe this combination—a concise model that captures the key economic mechanism, paired with thoughtful human reflection—offers the most value to readers interested in understanding how AI singularity risk might affect asset prices.

## 5 Conclusion

In this paper, we have proposed a novel explanation for the high valuations of AI companies: they may serve as hedges against negative AI singularity risk. Our model demonstrates how, even with small probabilities of AI-driven disasters, the price-dividend ratios of AI stocks can be substantially higher than what traditional growth expectations would suggest. This hedging value arises because AI companies may capture a disproportionate share of economic value precisely when the representative household’s consumption falls dramatically due to AI development.

Financial markets naturally create mechanisms for risk sharing that can help society manage AI transition risks. By purchasing AI assets, households can partially hedge against the risk of labor income displacement from AI. However, this solution is fundamentally limited by market incompleteness. Many frontier AI companies remain privately held, and ordinary investors cannot directly access these investments. Even when public companies have stakes in AI labs (e.g., Microsoft’s investment in OpenAI), the ownership is diluted and



indirect. This market incompleteness means that the representative household cannot fully insure against AI singularity risk through portfolio choices alone.

Interestingly, these market-based hedging mechanisms have received relatively little attention in the literature on AI catastrophe risk. Researchers like Bostrom (2014), Russell (2019), and Bengio et al. (2024) have extensively discussed technical alignment and governance approaches to mitigate AI risks, while economists like Jones (2024) and Korinek and Suh (2024) have analyzed potential economic impacts of advanced AI. However, the role of financial markets in distributing AI transition risk remains underexplored.

Asset markets provide natural risk-sharing mechanisms that complement other policy approaches. While not a complete solution due to market incompleteness, allowing households to participate in AI upside through accessible investment vehicles could help mitigate some of the distributional concerns associated with rapid AI development. Further research on how to reduce market incompleteness and improve access to AI investment opportunities could yield valuable insights for managing the economic transition to advanced AI systems.

## 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*.
- Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, et al. (2024). “Managing extreme AI risks amid rapid progress”. In: *Science* 384.6698. URL: <https://arxiv.org/abs/2310.17688>.
- Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Chollet, François, Mike Knoop, Gregory Kamradt, and Bryan Landers (2024). “ARC Prize 2024: Technical Report”. In: *arXiv preprint*.
- DeepSeek-AI et al. (Jan. 2025). “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning”. In: *arXiv*. URL: <https://arxiv.org/abs/2501.12948>.
- Jones, Charles I. (2024). “The AI Dilemma: Growth versus Existential Risk”. In: URL: <https://web.stanford.edu/~chadj/existentialrisk.pdf>.
- Korinek, Anton and Donghyun Suh (2024). *Scenarios for the Transition to AGI*. Tech. rep. NBER Working Paper.
- Rietz, Thomas (1988). “The Equity Risk Premium: A Solution?” In: *Journal of Monetary Economics*.

- Vinge, Vernor (1993). “The Coming Technological Singularity”. In: *Department of Mathematical Sciences, San Diego State University*.
- Wachter, Jessica A. (2013). “Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?” In: *Journal of Finance*.
- Zhang, Miao Ben (2019). “Labor-Technology Substitution: Implications for Asset Pricing”. In: *Journal of Finance* 74.4, pp. 1793–1839.

# A A Purely Human Perspective

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

## # Prompts-to-Paper

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

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

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

-Andrew Chen, April 9, 2025

## # Motivation

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

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

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

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

I asked a friend if he would be interested in working on this paper. Unfortunately, he was busy with revision deadlines for the next month.

So, I thought I should use AI to write the paper. It would be an elegant way to make my point. It would also hint at where the research process is going in this strange age of AI.

## ## Inspiration

This project was inspired by [Novy-Marx and Velikov \(2025\)](#) and [Chris Lu et al. \(2024\)](#). These projects use AI to generate massive amounts of academic

research. My goal differs in quality over quantity. I want to generate just one paper, but one paper that (I hope) people find is worth reading.

This project was also inspired by [Garleanu, Kogan, and Panageas's \(2012\)](#) beautiful model of innovation and displacement risk. I read Garleanu et al. back when I was a PhD student and it has stuck with me.

Last, I drew from [Hadfield-Menell and Hadfield \(2018\)](#) and [Bengio \(2023\)](#), who apply ideas from economics to AI catastrophe risk. [Hadfield-Menell and Hadfield \(2018\)](#) explains the connection between incomplete contracting and AI alignment. [Bengio \(2023\)](#) frames AI catastrophe risk in terms of what I would call decision theory and human incentives---though the essay is written in plain English.

Previously, I dismissed "AI safety" as a politicized, activist cause. I still don't like the term "AI safety." "Safety" is such an excuse for exercising power over others.

But then the nature of AI changed. I found AI was better at math than me. I found I could write code by just talking to AI. Things were progressing much faster than I expected. The [Jan 15, 2025 episode of Machine Learning Street Talk with Yoshua Bengio](#) left an impression on me. Bengio talked about AI catastrophe risk with no activism, no fear mongering. It was a straight, rational discussion of the seriousness of AI catastrophe risk.

### # The Paper Generation Process

A natural question is: is this paper *\*really\** written by AI?

I'd say the AI are junior co-authors.

If they were human, I would absolutely have to give o1, Sonnet, and ChatGPT Deep Research credit as co-authors. They did the math, writing, and literature reviews. Sonnet also wrote most of the code (via the Cursor AI IDE).

Of course, the prompts ([plan0408-piecewise.yaml](#)) show I did substantial hand-holding. The many Github commits show even more human labor. They tell the story of me getting to know my, rather foreign, "co-authors." I found it hard to communicate subtle style issues and instructions on how to use LaTeX properly, leading to many, many commits.

To be honest, writing this paper would have been much easier if I had done

more of the work myself.

But that can happen with human co-authors too.

Perhaps in the next few years, AI and humans can be equal co-authors. I wouldn't be surprised if I could just ask a 2031 AI to "write a model to formalize my intuition," and it would get it just right.

## ## Paper Iterations

Like human-written papers, the writing process was iterative. The first formalizations were terrible.

`plan0313-laborshare.yaml` (from March 13) contains prompts for a neoclassical growth model, where the capital share suddenly increases. ChatGPT-o1 patiently explained to me why this is a bad model.

me: > I thought there would be a wage risk effect that leads to higher investment for the more risk averse agent. High capital share means low or even no wage income.

ChatGPT-o1: > Below is an explanation of why one might **expect** a "wage-risk" channel in which **more** risk aversion could lead to **more** **additional** investment (relative to the no-jump benchmark) in the event that the capital share might jump to 1. However, this channel **does not** operate in the usual **representative-agent** version of the model---there, wage and capital income ultimately go to the **same** agent, so there is no meaningful "hedge" of wage risk. Instead, the wage-risk hedge arises naturally if you depart from the pure representative-agent setting (for example, if households receive labor income but must **choose** how much capital they own).

I went through several iterations of the model with Claude 3.7 Sonnet (thinking mode) and ChatGPT-o1. The only derivations I did myself were to check o1's work, which I found to be quite reliable.

`plan0403-streamlined.yaml` tries to write a paper in just six prompts (less handholding). Prompts 1-3 do the analysis. Prompts 4-6 do the writing. I found this method leads to poor writing. The language got annoyingly academic, despite the system prompt saying "be conversational." Moreover, the economic subtleties were frequently lost.

The final `plan0408-piecewise.yaml` uses a simplified Barro-Rietz disaster

model, with two agents (though only one is relevant for stock prices). I slowly walk the AIs through the writing, using ten prompts, to maintain the writing quality.

## ## Literature Reviews

A key step was generating lit reviews ( `./lit-context/` ) which were used as context in the prompts. I made lit reviews using ChatGPT's Deep Research (launched Feb 2025) until I ran out of credits. I used Claude Web Search (launched March 20, 2025) did the remainder.

These new products are a game changer. Both [Novy-Marx and Velikov \(2025\)](#) and [Chris Lu et al. \(2024\)](#) ran into hallucinated citations. OpenAI Deep Research and Claude Web Search had no hallucinations if they were used with care.

Still, I would 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 mode) is OK at economic theory. But I found that it was not assertive enough. It led me down wrong paths because it was too eager to come up with some ideas that fit my story (even if they did not make sense).

I briefly tried having Llama 3.1 405b do the writing. It was terrible! It would be extremely difficult to generate a paper worth reading that way.

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

## ## Picking the best of N papers

The writing quality varies across each run of the code. Some drafts, I found

quite insightful! Others, had flagrant errors in the economics.

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

5 drafts of the paper can be found in `./manyout0408-pdf/`. They're fairly similar, all are OK, and I would be OK with my name on any of them.

I ended up choosing `paper-run-04.pdf` (actually, `paper-appendix-update-run4.pdf` since it needs to have this README updated). I thought that draft had pretty decent writing and lacked any noticeable flaws.

### # Lessons about Research

A common response to [Novy-Marx and Velikov \(2025\)](#) is: "people are not ready for this." I heard concerns that peer review process will be inundated with AI-generated slop.

Working on this paper gave me a different perspective. It made me think about the fundamentals. I think the fundamentals are the following:

1. Readers want to learn something interesting and true.
2. Readers don't want to check all the math.
3. A system of author reputations makes 1 and 2 possible.

AI-generated papers don't change any of these fundamentals. Critically, fundamental 3 made me quite wary of putting my name on AI slop. As a result, I don't think AI-generated papers will change much about peer review, at least not the current generation of AI.

### ## Limitations of the Current AI (April 9, 2025)

This will likely be out of date by the time you read it.

But right now, AI is like a junior co-author with a talent for mathematics and elegant writing, but sub-par economics reasoning.

For example, Sonnet often fails to recognize that the economic model does not capture an important channel. This is a common scenario in economics writing (no model can capture everything). The standard practice is to dance gingerly around the channel in the writing. A decent PhD student can recognize this. But Sonnet cannot. Instead, Sonnet will write beautiful prose about the channel anyway, even though it's not really being studied

properly.

AI also cannot generate a satisfying economic model on its own (at least not satisfying to me). When I tried, the resulting models were either too simplistic or did not lead to a clean analysis. They often introduced complications that I found unnecessary.

I opted not to add empirical work or numerically-solved models. The disaster version of [Martin's \(2013\) Lucas Orchard](#) would make a beautiful demonstration of my point, though it would need a numerical solution. AI can do both, but both require connecting to the outside world, and a plethora of technical challenges.

There could be models with capabilities that I missed. Perhaps a simple [Model Context Protocol](#) could significantly improve the paper.

But more important: how long will these limitations last?

### **## The Future of AI and Economics Research (Speculative)**

At some point, 2024-style economic analysis will be "on tap." You'll be able to go to a chatbot and ask "write me a paper about hedging AI disaster risk," and it will return you something like this paper (probably something much better).

"Economics on tap" could be a disaster for the economics labor market (could be). It certainly *will* be an extremely cheap substitute for at least some economists' labor. I suppose the question 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 many projects.

Just to be clear, I'm not saying that I *expect* a disaster for the economics labor market. Or, that it's even likely. But even if it's highly unlikely, it's still a scenario that economists should think about.



## B Prompts Used to Generate This Paper

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

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

The system prompt and instructions are listed below.

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

```
You are an asset pricing theorist who publishes in the top journals
(Journal of Finance, Journal of Financial Economics, Review of
Financial Studies). You think carefully with mathematics and
check your work, step by step.

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

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

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

Be modest. Do not overclaim.

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

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

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

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

Use the following outline:

- The model is purposefully simple and captures the essence of the main argument
- Two agents
  - AI owners
    - Fully invested in AI, not marginal investors in stock market
  - Representative household
    - Marginal investor in stocks: only their consumption matters for this analysis
    - CRRA =  $\gamma$ , time preference =  $\beta$
- Consumption growth
  - $\log \Delta c_{t+1} = 0$  if no disaster
  - $\log \Delta c_{t+1} = -b$  if disaster (prob  $p$ )
  - A disaster is a sudden improvement in AI that is devastating for the household
    - Think of as a worst-case scenario for AI progress
    - Economy booms, but the value of AI is captured by the AI owners.
    - For household, labor is replaced by AI, so labor income plummets, as does consumption.
      - Also, way of life, meaning, is lost. Consumption fall can be thought of as a stand-in for these losses.
  - at  $t=0$ , no disasters have happened (singularity has not occurred)
    - Multiple disasters may happen, capturing ongoing uncertainty if a singularity occurs
- AI asset
  - Captures publicly traded AI stocks
  - Dividend  $D_t = a \exp^{h N_t} C_t$
  - Interpretation (discuss in prose)

- $a > 0$  is small, AI stocks are currently a minor share of the economy
- $N \setminus t$  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  $\lim_{b \rightarrow 0} b$

- '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  $\Delta b < 0$  (a sudden boom, not a disaster)
- In reality, most households cannot buy shares in many cutting edge labs (e.g. OpenAI, Anthropic, xAI, DeepSeek), consistent with our model
- A more elaborate model would add detail to the AI owners, private AI assets, and their interactions with the representative household
- It could address questions like:
  - How does AI progress displace the representative household's wages?
  - How do AI owners' incentives affect AI progress and market incompleteness?
  - How do preferences and technology parameters affect the odds of a negative singularity?
- But wouldn't this just decorate speculations with math?
  - The core economics (rare disaster risk, hedging motives, market incompleteness) will remain the same
- It would also be much more costly to read
  - In our view, the benefit of reading a paper should exceed the cost
- A short model analysis allows room for the human-written Appendix `\ref\{app:readme\}`

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

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

Be sure to cite:

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

- Zhang (2019) "Labor-Technology"

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

Write a short "Conclusion" section.

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

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

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

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

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

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

Add caveats:

- We are not saying a negative singularity will happen
  - But it is nevertheless important to consider this scenario

- We are also not saying that this hedging value is priced in already
- Model illustrates a possible mechanism

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

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

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

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

The abstract should:

- Make the main argument (AI valuations may be high because they hedge against a negative AI singularity)
- Define "negative AI singularity" after using the term
- Touch on financial market solutions to AI catastrophe risk, in passing
- End by briefly mentioning that this short paper is written by prompting LLMs.

Do not:

- Emphasize consumption
- Oversell or overinterpret the model

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

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

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

Style Notes:

- Avoid bullet points and numbered lists

- No subsections (e.g. Section 1.2) though sections are OK (Section 1)
- Don't say "in conclusion" or "in summary"

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