

Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

We explore how AI stocks might be priced in the presence of AI singularity risk. We develop a theoretical framework showing that AI stocks can serve as a hedge against negative AI singularity events that would be devastating for the representative investor. Our analysis suggests that financial markets may be organically developing mechanisms to help households manage AI catastrophe risk, though these solutions are inherently constrained by market incompleteness. This short paper was written by prompting large language models to generate content and mathematical derivations based on human-specified instructions.

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 has advanced at a remarkable pace in recent years. DeepSeek’s R1 reasoning model achieved performance comparable to OpenAI’s o1 through pure reinforcement learning despite resource constraints (DeepSeek-AI et al., 2025). The ARC AGI Challenge benchmark, designed to test generalization on novel tasks, has seen scores rise from 33% to 55.5% in just one year (Chollet et al., 2024). Companies like Waymo have deployed fully autonomous vehicles in multiple cities. These developments reflect both the accelerating pace of AI progress and its expanding scope across domains. As these technologies proliferate, investors may justifiably worry about their own economic prospects. What happens to wages, careers, and personal finances if AI can perform an increasing share of human work?

Technological change has, of course, occurred before. The Industrial Revolution, electrification, and computerization all transformed labor markets. Yet AI represents something potentially different—a technology that could, in principle, create any product or service currently produced by humans. Unlike previous technologies that augmented human capabilities in specific domains, general-purpose AI might substitute for human labor across virtually all sectors. Interestingly, this very paper exemplifies the capabilities of current AI systems; it was generated entirely through engineered prompts to large language models (see <https://github.com/chenandrewy/Prompts-to-Paper/>).¹ More concerning is the potential for an “AI singularity”—a hypothetical point at which AI capabilities improve so rapidly that the resulting changes become difficult to predict and potentially devastating for human welfare (Bostrom, 2014).

In this paper, we explore how AI stocks might be priced in the presence of AI singularity risk. We develop a theoretical framework showing that AI stocks can serve as a hedge against negative AI singularity events that would be devastating for the representative investor. This perspective contrasts with the common view that AI valuations are high solely due to expectations of future earnings growth. Our model demonstrates that even a small probability of a severe AI-driven disaster can significantly increase the price-dividend ratio of AI assets. This hedging value emerges because AI assets perform relatively better during potential singularity events, creating a negative correlation between AI asset returns and the representative household’s consumption.

We emphasize two important caveats. First, we are not predicting that a negative AI singularity will occur. Experts disagree substantially about both the timeline and nature of advanced AI development (Bengio, 2023). Nevertheless, it is important to consider this sce-

¹Throughout this paper, “we” refers to one human author and multiple large language models working together. A purely human perspective on this work can be found in [Appendix A](#).

nario given its potentially enormous consequences, much as economists study rare disasters despite their low probability (Barro, 2006). Second, we are not claiming that current AI stock valuations already reflect this hedging value. Our aim is more modest: to illustrate a possible mechanism through which rational investors might value AI stocks partly for their insurance properties against worst-case AI outcomes.

Our paper contributes to multiple strands of literature. First, we build on research examining the potential macroeconomic consequences of advanced AI. Jones (2024) explores the tension between AI-driven economic growth and existential risk, while Korinek and Suh (2024) analyze how wages and output might evolve during transitions to artificial general intelligence, including scenarios where full automation becomes possible. This economic literature highlights the possibility of negative consequences for the representative household from advanced AI, but has not explicitly connected these risks to asset pricing. A second literature strand examines rare disasters in asset pricing, pioneered by Rietz (1988) and Barro (2006), with extensions to time-varying disaster probability by Wachter (2013). Our work applies this framework to technological disasters specifically. Third, we connect to research on hedging labor income risk, including Benzoni, Collin-Dufresne, and Goldstein (2007) and Betermier et al. (2012), who study how households adjust portfolios in response to labor market risks.

Recent empirical work on AI stocks provides a foundation for our theoretical analysis. Babina et al. (2023) show that firms’ investments in AI technologies affect their systematic risk profiles, with implications for asset pricing and cost of capital. Similarly, Zhang (2019) demonstrates that firms with greater opportunities to replace routine-task labor with automation maintain a replacement option that hedges their value against unfavorable macroeconomic shocks, lowering expected returns. Our paper extends this literature by proposing a novel explanation for the high valuations of AI stocks: they may serve as a hedge against a negative AI singularity that would be devastating for the representative investor. This perspective contrasts with the standard view that attributes high AI valuations solely to expectations of future earnings growth. By analyzing AI stocks as insurance against technological disaster risk, we provide a new lens for understanding their pricing and risk characteristics.

2 Model

We develop a simple model to capture the essence of our argument. While our model is stylized, it allows us to highlight the key mechanism through which AI stocks might serve as a hedge against negative AI singularity events.

Our economy consists of two types of agents. First, there are AI owners who are fully invested in AI assets. 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 representative household has standard preferences with constant relative risk aversion γ and time discount factor β :

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

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

$$\log \Delta C_{t+1} = 0 \quad \text{if no disaster} \quad (2)$$

However, with probability p , a disaster occurs, causing consumption to fall:

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

In our context, a disaster represents a sudden improvement in AI that is devastating for the representative household. This can be thought of as a worst-case scenario for AI progress. In such a scenario, the economy as a whole may boom, but 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 consumption to plummet. Beyond the direct economic impact, this consumption decline can also be interpreted as a stand-in for other losses such as way of life and meaning.

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, capturing the ongoing uncertainty that would persist even after an initial singularity event.

We model an AI asset that represents publicly traded AI stocks. The dividend process for this asset is given by:

$$D_t = a \exp(hN_t)C_t \quad (4)$$

where $a > 0$ is a small constant, reflecting that AI stocks currently represent a minor share of the economy. N_t denotes the number of disasters (AI singularity events) that have occurred up to and including time t . The parameter $h > 0$ captures the growth in AI asset dividends relative to the economy when a disaster occurs.

This specification implies that each time a disaster occurs, the AI asset grows as a share

of the economy. Intuitively, firms that provide semiconductors, data, AI models, and related technologies at least partially benefit from sudden improvements in AI, even as these same improvements may devastate household consumption. The AI asset's dividends are tied to aggregate consumption, but they grow disproportionately during disaster events.

3 Asset Pricing Results

We now derive the equilibrium price of the AI asset and examine how it relates to our central thesis. Asset pricing theory tells us that the stochastic discount factor (SDF) between time t and $t + 1$ is given by:

$$M_{t+1} = \beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \quad (5)$$

This SDF represents the representative household's marginal rate of substitution between consumption today and consumption tomorrow. For any asset with future payoffs, the price is the expected discounted value of those payoffs using the SDF.

For our AI asset with dividend process $D_t = a \exp(hN_t)C_t$, the ex-dividend price P_t must satisfy:

$$P_0 = E_0 \left[\sum_{t=1}^{\infty} \left(\prod_{u=1}^t M_u \right) D_t \right] \quad (6)$$

Dividing by the current dividend D_0 , we can express the price-dividend ratio as:

$$\frac{P_0}{D_0} = E_0 \left[\sum_{t=1}^{\infty} \left(\prod_{u=1}^t M_u \right) \frac{D_t}{D_0} \right] \quad (7)$$

Given our dividend process, we have:

$$\frac{D_t}{D_0} = \frac{a \exp(hN_t)C_t}{a \exp(hN_0)C_0} = \exp[h(N_t - N_0)] \times \frac{C_t}{C_0} \quad (8)$$

Additionally, the product of discount factors simplifies to:

$$\prod_{u=1}^t M_u = \beta^t \left(\frac{C_t}{C_0} \right)^{-\gamma} \quad (9)$$

Combining these expressions, the price-dividend ratio becomes:

$$\frac{P_0}{D_0} = E_0 \left[\sum_{t=1}^{\infty} \beta^t \left(\frac{C_t}{C_0} \right)^{1-\gamma} \exp[h(N_t - N_0)] \right] \quad (10)$$

This formula has an intuitive interpretation. The price-dividend ratio depends on three components: (1) the time discount factor β^t , (2) the growth of consumption adjusted for risk aversion $(C_t/C_0)^{1-\gamma}$, and (3) the exponential growth of the AI sector during disaster events $\exp[h(N_t - N_0)]$.

Given our disaster process for consumption growth and our assumption that $\gamma = 2$, we can further simplify this expression. When a disaster occurs, consumption falls by a factor of $\exp(-b)$, but the term $(C_t/C_0)^{1-\gamma}$ actually increases by a factor of $\exp(b)$. Combined with the AI dividend growth factor $\exp(h)$, each disaster effectively increases the present value of future dividends by a factor of $\exp(b + h)$.

Under stationarity (when each period's discount multiplier is less than one on average), the price-dividend ratio simplifies to a geometric series:

$$\frac{P_0}{D_0} = \frac{r}{1 - r} \quad \text{where} \quad r = \beta[(1 - p) + p \exp(b + h)] \quad (11)$$

This expression shows that the AI asset's price-dividend ratio increases with both the disaster size b and the AI growth parameter h . Intuitively, larger disasters (higher b) and stronger AI growth during disasters (higher h) both make the AI asset more valuable relative to its current dividends.

To better understand how disaster risk affects AI asset valuations, we present numerical examples of the price-dividend ratio under various parameter combinations. We fix $\gamma = 2$, $\beta = 0.96$, and $h = 0.2$, and vary both the disaster size b and probability p .

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

	$p = 0.0001$	$p = 0.001$	$p = 0.01$	$p = 0.02$
$b = 0.4$	24.03	24.52	29.30	39.00
$b = 0.6$	24.06	24.75	34.70	63.60
$b = 0.8$	24.10	25.32	40.67	141.90
$b = 0.95$	24.12	25.36	51.63	—

This table reveals several important patterns. When the disaster probability is very low ($p = 0.0001$), the price-dividend ratio remains relatively constant across different disaster sizes. This makes intuitive sense—if disasters almost never happen, their potential impact matters little for asset pricing.

However, as the disaster probability increases, the price-dividend ratio becomes increasingly sensitive to the disaster size. For example, with $p = 0.01$ (a 1% annual chance of disaster), increasing the disaster size from $b = 0.4$ to $b = 0.95$ nearly doubles the price-dividend ratio from 29.30 to 51.63. With $p = 0.02$, the effect is even more dramatic, with the price-dividend ratio more than tripling from 39.00 to 141.90 as b increases from 0.4 to

0.8.

These results directly support our central thesis. AI stocks may command high valuations not only because of expected future growth in normal times, but also because they serve as a hedge against AI singularity events that would be devastating for the representative household. The model shows that even a small probability of a severe AI-driven disaster can significantly increase the price-dividend ratio of AI assets. Moreover, the effect becomes stronger as either the probability or the severity of the disaster increases.

The hedging value emerges because AI assets perform relatively better during disaster periods. While the representative household experiences a consumption decline of $\exp(-b)$, the AI asset's dividend grows by a factor of $\exp(h)$ relative to aggregate consumption. This negative correlation between the AI asset's performance and the household's consumption makes the asset particularly valuable as insurance against AI singularity risk.

Our analysis suggests that current high valuations of AI stocks may partly reflect their role as insurance against worst-case AI outcomes rather than simply optimistic expectations about future earnings growth. This perspective adds nuance to debates about whether AI stocks are overvalued, by highlighting that traditional growth-based valuation metrics might miss this insurance premium.

4 Model Discussion

Our model captures the essential mechanism through which AI stocks might serve as a hedge against negative AI singularity events. However, like any model, it abstracts from certain complexities that deserve discussion. Here we highlight some key subtleties and explain why our modeling approach is appropriate for our purposes.

A critical feature of our model is the implicit 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 is 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, including private AI companies and research labs. In such a world, the household could not only fully hedge against AI singularity risk but potentially benefit from it, implying $b < 0$ (a sudden boom rather than a disaster). The household would effectively own enough of the AI sector to offset any negative impacts on labor income or welfare.

However, reality aligns with our model's assumption of market incompleteness. Most households cannot buy shares in many cutting-edge AI labs such as OpenAI, Anthropic,

xAI, or DeepSeek. These private companies, which are at the frontier of AI research, remain largely inaccessible to the average investor. Consequently, even if households invest heavily in publicly traded AI stocks, they cannot fully hedge against AI singularity risk.

One might argue that a more elaborate model could add detail to the AI owners, private AI assets, and their interactions with the representative household. Such a model could address questions like: How 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?

While these questions are fascinating, we believe our simpler approach has several advantages. First, adding such details would essentially decorate speculations with mathematics. The core economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain the same. The additional complexity would not fundamentally alter our main insight: that AI stocks might be valued partly for their hedging properties against negative AI singularity events.

Second, a more complex model would be significantly more costly for readers to digest. In our view, the benefit of reading a paper should exceed the cost. Our streamlined model communicates the essential idea without burdening readers with unnecessary technical details. This approach allows us to focus on the economic intuition rather than mathematical complexity.

In essence, our model serves as a thought experiment that formalizes a simple but potentially important insight: the high valuations of AI stocks might partly reflect their hedging value against negative AI singularity events, rather than just optimistic growth expectations. This perspective adds nuance to discussions about AI stock valuations and highlights a mechanism that has received little attention in both academic and popular discourse.

5 Conclusion

We have presented a theoretical framework showing that AI stocks may serve as a hedge against negative AI singularity events. Our model demonstrates that even a small probability of a severe AI-driven disaster can significantly increase the price-dividend ratio of AI assets. This hedging value emerges because AI assets perform relatively better during a potential singularity, creating a negative correlation between AI asset returns and the representative household’s consumption. This perspective adds nuance to the common view that AI valuations are high solely due to expected future earnings growth.

Our analysis suggests that financial markets may be organically developing mechanisms to help households manage AI catastrophe risk. While Universal Basic Income is frequently

proposed as a policy response to AI-driven labor displacement, financial markets offer complementary risk-sharing capabilities. By investing in publicly traded AI companies, households can potentially offset some of the negative consequences of an AI singularity on their labor income.

However, these financial market solutions are inherently constrained by market incompleteness. As highlighted in our model, many frontier AI labs remain private and inaccessible to average investors. This market incompleteness means that households cannot fully hedge against AI singularity risk through their investment portfolios. The magnitude of this incompleteness is captured by our parameter $b > 0$, indicating that the representative household experiences a net consumption loss during an AI singularity event despite any AI investments they might hold.

This financial market perspective on AI risk mitigation is surprisingly underexplored in the AI catastrophe risk literature, which tends to focus on technical AI safety (Bengio et al., 2024; Bostrom, 2014) or policy interventions. Work by Jones (2024) examines the economic tension between AI-driven growth and existential risk, but doesn’t fully explore how financial markets might help distribute AI risk. Similarly, analyses of AI’s economic implications by Korinek and Suh (2024) acknowledge potential labor market disruptions but give less attention to how financial assets might serve as hedging instruments.

Financial markets provide a mechanism for risk sharing that works alongside other approaches to managing AI risk. The hedging role of AI assets demonstrates how markets can, to some extent, anticipate and price future technological disruptions. By allowing individuals to invest in technologies that might otherwise threaten their economic standing, financial markets offer a partial—though incomplete—solution to technological displacement risks. Future research should further explore how to enhance these market mechanisms while developing complementary approaches to ensure that the benefits of AI progress are widely shared.

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 (2023). *FAQ on Catastrophic AI Risks*. URL: <https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/>.

- 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>.
- 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.
- Betermier, Sebastien, Thomas Jansson, Christine Parlour, and Johan Walden (2012). “Hedging Labor Income Risk”. In: *Journal of Financial Economics* 105.3, pp. 622–639.
- 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*.
- 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, 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

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 = γ , time preference = β
- 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.