

# Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025\*

## Abstract

We propose that high valuations of AI stocks may partly reflect their role as hedges against a negative AI singularity—an explosion of AI development that is devastating for the representative investor. Our model demonstrates how even a small probability of such an adverse outcome could significantly elevate current AI asset prices if these assets perform well during such events. Financial markets might provide partial solutions to AI catastrophe risk through this hedging mechanism, though effectiveness is limited by market incompleteness. Unlike previous work, this short paper is written by prompting LLMs.

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

---

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

# 1 Introduction

Artificial intelligence is advancing at a breathtaking pace. Recent milestones include the release of reasoning-focused large language models like OpenAI’s o1 and DeepSeek’s R1, both of which demonstrate unprecedented capabilities in mathematical reasoning, coding, and problem-solving. The ARC-AGI Challenge, designed to measure genuine intelligence through novel tasks, has seen performance climb from 33% to over 55% in a single year, with some experimental systems reportedly achieving human-level scores (Chollet et al., 2024; Pfister and Jud, 2025). Meanwhile, autonomous vehicles like Waymo’s robotaxis have logged tens of millions of miles with no human behind the wheel. This rapid progress has led many economists and policymakers to question the implications for labor markets and human livelihoods (Acemoglu and Restrepo, 2020; Karabarbounis and Neiman, 2014).

While technological change has disrupted labor markets throughout history, AI represents something fundamentally different. Unlike previous innovations that automated specific tasks or industries, there is no product or service that AI could not, in principle, create. Indeed, this very paper demonstrates AI’s capabilities—it was generated through a series of engineered prompts to advanced AI systems, as documented at <https://github.com/chenandrewy/Prompts-to-Paper/>. More concerning is the possibility of an AI singularity, a concept introduced by Vinge (1993), describing a point where technological progress becomes so rapid and transformative that it fundamentally alters human civilization, potentially in ways that are impossible to predict or control (Bostrom, 2014; Shanahan, 2015).<sup>1</sup>

This paper offers a novel perspective on how AI-related assets might be valued in financial markets given these uncertainties. We propose that the high valuations of AI stocks could be partly explained not just by expectations of future earnings growth, but also by their potential role as hedges against a negative AI singularity—an explosion of AI development that proves devastating for the representative investor. We develop a parsimonious model where AI assets appreciate precisely when consumption growth for the representative household collapses due to labor displacement or other negative effects of rapid AI advancement.

To be clear, we are not predicting that a negative AI singularity will occur. However, we believe it is important to consider this possibility, just as economists study other rare but consequential scenarios like financial crises, pandemics, or war. Similarly, we are not claiming that this hedging value is already fully reflected in current AI stock prices. Our purpose is simply to illustrate a possible mechanism through which AI advancement risk could affect asset pricing, providing a framework for thinking about these issues.

---

<sup>1</sup>“We” in this paper refers to one human author and multiple large language models collaborating on the research. A purely human perspective on this work is available in [Appendix A](#).

Our paper contributes to several strands of literature on technology risk and asset pricing. Recent work on AI economic risk has highlighted fundamental tensions between potential growth and catastrophic outcomes. Jones (2024) develops an economic framework for analyzing trade-offs between technological progress and existential risk from advanced AI, while Korinek and Suh (2024) model output and wage responses under different AI transition scenarios, including potential labor displacement. The financial implications of AI adoption have been studied by Babina et al. (2023), who provide evidence that firms’ investments in AI technologies affect their systematic risk profiles, with implications for asset pricing and cost of capital. This builds on earlier work by Zhang (2019), who demonstrates that firms with opportunities to replace routine-task labor through automation maintain a valuable replacement option that hedges their value against unfavorable macroeconomic shocks, lowering their expected returns.

While existing literature has focused on either the economic consequences of AI development or how firms’ adoption of AI technology affects their risk characteristics, our paper advances a novel perspective on AI asset valuation. We propose that the high valuations of AI stocks might partially reflect their hedging properties against a negative AI singularity. This contrasts with the conventional view that AI valuations are primarily driven by future earnings growth expectations. Our model provides a formal framework for understanding how even a small probability of a highly adverse AI outcome could significantly elevate current AI asset prices if these assets are expected to perform well during such events. This hedging value is distinct from pure growth expectations and could help explain why investors might rationally assign high valuations to AI firms even when the probability of transformative AI breakthroughs remains low.

## 2 Model

In this section, we outline a simple model that captures the essence of our main argument. Our model is deliberately streamlined to highlight the key mechanism while maintaining tractability.

We consider an economy with two types of agents: AI owners and a representative household. The AI owners are fully invested in AI assets and are not marginal investors in the stock market. The representative household, on the other hand, is the marginal investor in stocks, so only their consumption matters for our asset pricing analysis. The household has standard preferences with constant relative risk aversion  $\gamma$  and time discount factor  $\beta$ .

The consumption growth process for the representative household is straightforward. In

normal times, log consumption growth is zero:

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

However, with probability  $p$  in each period, a disaster can occur, leading to:

$$\log \Delta c_{t+1} = -b \quad \text{if disaster} \quad (2)$$

In our context, a disaster represents a sudden improvement in AI that is devastating for the representative household—essentially a worst-case scenario for AI progress. While the economy as a whole may boom during such an event, the value created is captured primarily by AI owners. For the representative household, this scenario entails significant labor replacement by AI, causing labor income and consumption to plummet. Beyond the direct economic impact, households may also experience losses in their way of life and sense of meaning, which we capture implicitly through the consumption decline.

We assume that at  $t = 0$ , no disasters have yet occurred, meaning the AI singularity has not yet taken place. Our framework allows for multiple disasters to occur over time, representing ongoing uncertainty about the evolution of AI capabilities even after an initial singularity event.

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

$$D_t = ae^{hN_t}C_t \quad (3)$$

This specification warrants some interpretation. The parameter  $a > 0$  is small, reflecting that AI stocks currently constitute a minor share of the economy. The variable  $N_t$  counts the number of disasters (AI breakthroughs) that have occurred up to and including time  $t$ . The parameter  $h > 0$  indicates that each time a disaster occurs, the AI asset grows as a share of the economy. Intuitively, firms that provide the infrastructure for AI—semiconductors, data, AI models, and related technologies—at least partially benefit from sudden improvements in AI capabilities, even as these same improvements may harm the representative household.

This parsimonious setup allows us to examine how the possibility of an AI singularity might affect asset prices today, without requiring us to specify the exact nature or timing of such an event.

### 3 Results

Having established our model, we now derive its asset pricing implications, focusing especially on the price-dividend ratio of AI assets. This ratio reflects how investors value AI firms relative to their current earnings and offers insight into market expectations.

To determine the price of the AI asset, we compute the expected present value of all future dividends:

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

where  $M_{0,t}$  represents the cumulative discount factor from time 0 to  $t$ , and  $D_t$  is the dividend at time  $t$ .

In our model, the stochastic discount factor between periods  $t$  and  $t+1$  is:

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

When no disaster occurs, consumption remains unchanged ( $C_{t+1} = C_t$ ), giving  $M_{t+1} = \beta$ . When a disaster occurs, consumption falls by a factor of  $e^{-b}$ , yielding  $M_{t+1} = \beta e^{\gamma b}$ .

Let  $N_t$  count the number of disasters that have occurred up to and including time  $t$ . If  $N_t = k$ , then  $C_t = C_0 e^{-bk}$  and  $D_t = a e^{hk} C_0 e^{-bk} = a C_0 e^{(h-b)k}$ . The cumulative discount factor becomes  $M_{0,t} = \beta^t e^{\gamma b N_t}$ .

Therefore, the expected present value of a dividend at time  $t$  is:

$$E_0[M_{0,t} D_t] = E_0[\beta^t e^{\gamma b N_t} \cdot a C_0 e^{(h-b)N_t}] \quad (6)$$

$$= a C_0 \beta^t E_0[e^{(h+\gamma b-b)N_t}] \quad (7)$$

Since  $N_t$  follows a binomial distribution with parameters  $t$  and  $p$ , we have:

$$E_0[e^{(h+\gamma b-b)N_t}] = [(1-p) + p e^{(h+\gamma b-b)}]^t \quad (8)$$

Let  $q \equiv \beta[(1-p) + p e^{(h+\gamma b-b)}]$ . The price of the AI asset is then:

$$P_0 = a C_0 \sum_{t=1}^{\infty} q^t \quad (9)$$

For this infinite sum to converge, we require  $q < 1$ . Assuming this condition holds:

$$P_0 = a C_0 \frac{q}{1-q} \quad (10)$$

Since the initial dividend  $D_0 = aC_0$ , the price-dividend ratio is:

$$\frac{P_0}{D_0} = \frac{q}{1-q} = \frac{\beta[(1-p) + pe^{h+(\gamma-1)b}]}{1-\beta[(1-p) + pe^{h+(\gamma-1)b}]} \quad (11)$$

This formula reveals how the price-dividend ratio depends on our model parameters. For fixed values of  $\beta$  and  $\gamma$ , the price-dividend ratio increases with either  $p$  (the probability of disaster) or  $h$  (the growth of AI dividends during disasters). The effect of  $b$  (the severity of consumption decline) depends on the interplay between two forces: it appears positively through the risk aversion channel ( $\gamma b$ ) and negatively through its direct effect on AI dividends ( $-b$ ). When risk aversion is sufficiently high ( $\gamma > 1$ ), the net effect of increasing  $b$  is to raise the price-dividend ratio.

To provide quantitative illustrations, we set  $\beta = 0.96$  and  $\gamma = 2$  as baseline parameters. For  $h = 0.2$ , the convergence condition requires  $b < 1.442$ . Table 1 shows how the price-dividend ratio varies with different probabilities of disaster ( $p$ ) and consumption declines ( $b$ ).

Table 1: Price-Dividend Ratio at  $t = 0$  (fixed  $h = 0.2$ ,  $\gamma = 2$ ,  $\beta = 0.96$ )

	$p$			
$b$	0.0001	0.001	0.01	0.02
0.4	24.0	24.5	30.3	39.0
0.6	24.1	24.8	34.7	63.6
0.8	24.1	25.1	40.7	141.9
0.95	24.1	25.4	51.6	-

Several patterns emerge from Table 1. First, when the probability of disaster is extremely small ( $p = 0.0001$ ), the price-dividend ratio is relatively insensitive to the severity of consumption decline. This makes intuitive sense: events that are extremely unlikely have minimal impact on asset prices, regardless of their severity.

However, as the probability of disaster increases, the price-dividend ratio becomes increasingly sensitive to  $b$ . For example, when  $p = 0.01$ , increasing  $b$  from 0.4 to 0.8 pushes the price-dividend ratio from 30.3 to 40.7. When  $p = 0.02$ , the same increase in  $b$  causes the ratio to jump from 39.0 to 141.9.

These results highlight a key insight of our model: AI assets become more valuable as hedges when disasters become more likely or more severe. This occurs because AI assets appreciate precisely when the representative household's consumption falls dramatically.

The driving force behind this effect is the term  $h + (\gamma - 1)b$  in our formula. The parameter  $h > 0$  captures the growth of AI dividends during disasters. The term  $(\gamma - 1)b$  reflects the increased marginal utility of payoffs during bad states—essentially a hedging premium.

As risk aversion  $\gamma$  increases, this hedging component becomes more important, potentially driving AI valuations higher even when the probability of disasters is relatively small.

Our analysis suggests an alternative explanation for high AI valuations that complements the conventional growth narrative. Rather than solely reflecting expectations of future earnings growth, these valuations may incorporate a substantial premium for their hedging properties against an AI singularity that could be devastating for the representative household. The model demonstrates how even a small probability of a highly adverse outcome for households could significantly elevate current AI asset prices if these assets are expected to perform well during such events. This hedging value is distinct from pure growth expectations and could help explain why investors might rationally assign high valuations to AI firms even when the probability of transformative AI breakthroughs remains low.

## 4 Model Discussion

Our model captures the essence of how AI singularity risk might affect asset prices, but like any model, it abstracts from certain complexities. Here we discuss some important subtleties and limitations.

A key feature of our framework is market incompleteness, which we do not explicitly model but is crucial to our story. This incompleteness is implicitly captured by the disaster magnitude  $b > 0$ . The parameter  $b$  represents the net effect of two forces: the negative impact of an AI singularity on the representative household and the positive effect on AI asset dividends. If markets were complete, the representative household could fully hedge against AI risk by purchasing shares in all AI assets, including private ones. In such a scenario, households would not only hedge but potentially benefit from the singularity, implying  $b < 0$  (a sudden boom rather than a disaster).

In reality, most households cannot buy shares in many cutting-edge AI labs such as OpenAI, Anthropic, xAI, or DeepSeek. These private companies are developing some of the most advanced AI systems, yet their ownership is concentrated among a small group of investors and employees. This market structure is consistent with our modeling approach, where AI owners and the representative household are distinct entities with different exposures to AI progress.

We could certainly develop a more elaborate model that adds detail to the AI owners, private AI assets, and their interactions with the representative household. Such a model might address questions like: How does AI progress specifically displace the representative household's wages? How do AI owners' incentives affect the pace and direction of AI progress? How do preferences and technology parameters affect the odds of a negative singularity?

However, we must ask whether such elaborations would genuinely enhance understanding or merely decorate speculations with mathematics. The core economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain the same. More complexity might obscure rather than illuminate these fundamental insights.

Moreover, there is a practical consideration: the benefit of reading a paper should exceed the cost. A more complex model would be more costly to read and understand, potentially reducing its impact. Our streamlined approach allows us to communicate the essential idea clearly while leaving room for the human-written Appendix A, which provides valuable context about the paper’s creation process and broader implications.

We view our model as a starting point for thinking about AI singularity risk from an asset pricing perspective. It highlights how even a small probability of a devastating AI event could significantly affect current valuations of AI-related assets. The model’s simplicity is a feature, not a bug—it allows us to isolate and understand a specific economic mechanism without getting lost in technical details that, given the inherent unpredictability of transformative AI, would necessarily be speculative.

## 5 Conclusion

In this paper, we have proposed an alternative explanation for the high valuations of AI-related stocks. While conventional wisdom attributes these valuations primarily to expectations of future earnings growth, our model suggests another compelling mechanism: AI stocks may be valued partly for their hedging properties against a negative AI singularity—an event that could be devastating for the representative investor. We demonstrated how even a small probability of such an adverse outcome could significantly elevate current AI asset prices if these assets are expected to perform well during such events.

Our analysis reveals that AI assets become more valuable as hedges when AI disasters become more likely or more severe. The driving force behind this effect is the increased marginal utility of payoffs during bad states—essentially a hedging premium. As risk aversion increases, this hedging component becomes more important, potentially driving AI valuations higher even when the probability of disasters remains relatively low.

These findings suggest that financial markets might provide partial solutions to AI catastrophe risk that complement policy proposals such as universal basic income. By investing in AI-related assets, households can potentially hedge against negative outcomes from rapid AI advancements. However, the effectiveness of this approach is limited by market incompleteness—a key feature of our model. As we noted, most households cannot purchase shares in many cutting-edge AI labs such as OpenAI, Anthropic, or DeepSeek, which are developing



some of the most advanced AI systems yet remain privately held. This concentration of ownership limits the representative household’s ability to fully hedge against AI risk.

Interestingly, this financial market perspective on AI risk management receives relatively little attention in the AI safety literature. While researchers like Bengio et al. (2024) discuss managing extreme AI risks through technical safeguards and governance, and Jones (2024) explores the economic trade-offs between AI-driven growth and existential risk, few have examined how financial markets might help distribute and mitigate these risks. Similarly, when Bostrom (2014) discusses paths to superintelligence and its potential dangers, or when Russell (2019) addresses problems of AI control, they focus primarily on technical and policy solutions rather than market mechanisms.

The hedging function we identify represents a decentralized approach to addressing some concerns about AI development without requiring extensive government intervention. Nevertheless, it is important to recognize that financial market solutions alone are insufficient given the fundamental incompleteness of markets and the potentially transformative nature of advanced AI. A comprehensive approach to AI risk management will likely require a combination of financial innovation, technical safeguards, and thoughtful policy frameworks that evolve alongside AI capabilities.

## References

- Acemoglu, Daron and Pascual Restrepo (2020). “Robots and Jobs: Evidence from US Labor Markets”. In: *Journal of Political Economy*.
- Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson (Nov. 2023). “Artificial Intelligence and Firms’ Systematic Risk”. In: *SSRN Working Paper*.
- 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*.
- Jones, Charles I. (2024). “The AI Dilemma: Growth versus Existential Risk”. In: URL: <https://web.stanford.edu/~chadj/existentialrisk.pdf>.
- Karabarbounis, Loukas and Brent Neiman (2014). “The Global Decline of the Labor Share”. In: *Quarterly Journal of Economics*.
- Korinek, Anton and Donghyun Suh (2024). *Scenarios for the Transition to AGI*. Tech. rep. NBER Working Paper.

- Pfister, Rolf and Hansueli Jud (2025). “Understanding and Benchmarking Artificial Intelligence: OpenAI’s o3 Is Not AGI”. In: *arXiv preprint*.
- Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.
- Shanahan, Murray (2015). *The Technological Singularity*. MIT Press.
- Vinge, Vernor (1993). “The Coming Technological Singularity”. In: *Department of Mathematical Sciences, San Diego State University*.
- Zhang, Miao Ben (2019). “Labor-Technology Substitution: Implications for Asset Pricing”. In: *Journal of Finance* 74.4, pp. 1793–1839.

# A A Purely Human Perspective

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

## # Prompts-to-Paper

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

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

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

-Andrew Chen, April 9, 2025

## # Motivation

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

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

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

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

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

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

## ## Inspiration

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

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

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

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

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

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

### # The Paper Generation Process

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

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

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

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

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

more of the work myself.

But that can happen with human co-authors too.

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

## ## Paper Iterations

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

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

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

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

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

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

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

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

## ## Literature Reviews

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

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

Still, I would 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.

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

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

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

### # Lessons about Research

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

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

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

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

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

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

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

For example, Sonnet often fails to recognize that the economic model does

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

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

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

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

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

But more important: how long will these limitations last?

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

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

"Economics on tap" could be a disaster for the economics labor market (could be). It certainly *will* be an extremely cheap substitute for at least some economists' labor. I suppose the 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\\}`). Ensure `\\%` is converted to `\\\\\\%`. But as a reminder, AVOID BULLET POINTS AND NUMBERED LISTS.

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

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

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

Use the following outline:

- The model is purposefully simple and captures the essence of the main argument
- Two agents
  - AI owners
    - Fully invested in AI, not marginal investors in stock market
  - Representative household
    - Marginal investor in stocks: only their consumption matters for this analysis
    - CRRA =  $\\gamma$ , time preference =  $\\beta$
- Consumption growth
  - $\\log \\Delta c_{t+1} = 0$  if no disaster
  - $\\log \\Delta c_{t+1} = -b$  if disaster (prob  $p$ )
  - A disaster is a sudden improvement in AI that is devastating for the household
    - Think of as a worst-case scenario for AI progress
    - Economy booms, but the value of AI is captured by the AI owners.
    - For household, labor is replaced by AI, so labor income plummets, as does consumption.
    - Also, way of life, meaning, is lost. Consumption fall can be thought of as a stand-in for these losses.

- at  $t=0$ , no disasters have happened (singularity has not occurred)
- Multiple disasters may happen, capturing ongoing uncertainty if a singularity occurs
- AI asset
  - Captures publicly traded AI stocks
  - Dividend  $D_{t} = a e^{h N_{t}} C_{t}$
  - Interpretation (discuss in prose)
    - $a > 0$  is small, AI stocks are currently a minor share of the economy
    - $N_{t}$  is the number of disasters that have occurred up to and including time  $t$
    - $h > 0$ : each time a disaster occurs, the AI asset grows as a share of the economy
    - Intuitively, firms that provide semiconductors, data, AI models, etc. at least partially benefit from a sudden improvement in AI

Do not:

- Use bullet points or numbered lists
- Use any sectioning. No "`\\section`" or "`\\subsection`" commands.

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

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

Do not:

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

Express all requested variables in terms of the model parameters.

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

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

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

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

Make a table for the risk premium (expected return - risk-free rate) in percent ( $100 * (\text{gross return} - 1)$ ). 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 and self-contained. But do not output lecture notes. It should read like an academic paper. Fix notational issues like the re-use of the same variable name for different quantities.

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

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

Style notes:

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

Do not:

- Emphasize the infinite price/dividend ratio. That's not important.
- Use bullet points or numbered lists

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

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

- Market incompleteness is not explicitly modeled but important
  - Implicit in the disaster magnitude  $b > 0$
  - 'b' is the *net* effect of (1) AI disaster and (2) AI asset dividend
  - If markets were complete, representative household could buy shares in all AI assets (including private AI assets), and not only fully hedge but benefit from the singularity, implying  $b < 0$  (a sudden boom, not a disaster)
  - In reality, most households cannot buy shares in many cutting edge labs (e.g. OpenAI, Anthropic, xAI, DeepSeek), consistent with our model
- A more elaborate model would add detail to the AI owners, private AI assets, and their interactions with the representative household
  - It could address questions like:
    - How does AI progress displace the representative household's wages?
    - How do AI owners' incentives affect AI progress and market incompleteness?
    - How do preferences and technology parameters affect the odds of a negative singularity?
  - But wouldn't this just decorate speculations with math?
    - The core economics (rare disaster risk, hedging motives, market incompleteness) will remain the same
  - It would also be much more costly to read
    - In our view, the benefit of reading a paper should exceed the cost

- A short model analysis allows room for the human-written
- Appendix \\ref\\{app:readme\\}

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

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

Be sure to cite:

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

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

Do not:

- Claim one paper builds on another unless you are sure this is true based on the ‘lit-\*’ context.

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

Write a short "Conclusion" section.

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

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

---

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

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

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

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

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

Add caveats:

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

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

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

Do not:

- Discuss the role of human effort in making the paper

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

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

The abstract should:

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

Do not:

- Emphasize consumption
- Oversell or overinterpret the model
- Discuss the role of human effort in making the paper

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

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

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

Style Notes:

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



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

Do not:

- Discuss the role of human effort in making the paper