Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

We propose a novel explanation for the high valuations of AI stocks: they may serve as hedges against a potential AI singularity that could be devastating for the representative investor. Our model shows that even a small probability of such an event can substantially increase AI asset valuations if these assets are expected to outperform during singularity scenarios. Financial markets might thus provide a partial solution to AI catastrophe risk concerns, though effectiveness is limited by market incompleteness. This short paper was written by prompting large language models to develop and formalize this economic argument.

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. In January 2025, DeepSeek's R1 model demonstrated reasoning capabilities comparable to OpenAI's o1, despite development constraints from US export controls (DeepSeek-AI et al., 2025). In early 2024, advanced models achieved 55.5% on the previously unbeaten ARC-AGI challenge, a benchmark designed to measure intelligence rather than pattern recognition (Chollet et al., 2024). Meanwhile, autonomous vehicles from companies like Waymo have transitioned from experimental technology to commercial service in multiple cities. These developments naturally raise questions about the economic implications for human labor, as investors increasingly contemplate whether their skills and wages might be displaced by increasingly capable AI systems.[1]

While technological change has disrupted labor markets throughout history, AI represents something fundamentally different. Previous technological revolutions typically automated specific tasks or industries, leaving many domains untouched. The industrial revolution mechanized manufacturing but left service work largely unaffected. The digital revolution transformed information processing while leaving physical tasks to humans. AI, however, is distinct because there is no product or service that it could not, in principle, create or enhance. This paper itself exemplifies this point—it was entirely written by AI systems through a series of engineered prompts. The complete process, including all prompts and iterations, is available at `https://github.com/chenandrewy/Prompts-to-Paper/`. Moreover, AI progress may occur with extraordinary suddenness, potentially leading to what has been termed an "AI singularity"—a period of explosive technological development that fundamentally transforms economic and social structures.

In this paper, we study how AI assets are priced given the risk that advanced AI may destroy livelihoods and reduce consumption for the representative household. While most analyses attribute high AI stock valuations to expectations of future earnings growth, we propose an additional factor: AI stocks may command premium valuations because they serve as hedges against a potential negative AI singularity. Our theoretical framework demonstrates that even a small probability of such an event can substantially increase the price-dividend ratios of AI assets if these assets are expected to outperform the broader economy during such events.

We should emphasize two important caveats to our analysis. First, we are not predicting that a negative AI singularity will occur. The technology's ultimate trajectory remains deeply

[1] Throughout this paper, "we" refers to one human author and multiple large language models working in collaboration. For a purely human perspective on this paper, see Appendix A.

uncertain. Nevertheless, it is important to consider this scenario from a risk management perspective, just as financial markets price in low-probability but high-impact events in other contexts. Second, we are not claiming that this hedging value is already fully priced into current AI stock valuations. Our model illustrates a possible mechanism through which such valuation effects could arise, but empirically measuring its significance would require additional work beyond the scope of this paper.

Our analysis builds on several strands of literature. The rare disasters literature has demonstrated how small probabilities of extreme events can significantly impact asset prices (Rietz, 1988; Barro, 2006; Wachter, 2013). Recent work has extended this framework to examine existential risks related to artificial intelligence. Jones (2024) explores the economic trade-off between AI-driven growth and potential existential risks, while Korinek and Suh (2024) analyze how different scenarios for the transition to artificial general intelligence could impact economic outcomes and welfare distribution. In parallel, research on technology assets has shown how investments in automation can create valuable option-like properties. Zhang (2019) demonstrates that firms with routine-task labor maintain "replacement options" that hedge against unfavorable macroeconomic shocks, allowing these firms to substitute capital for labor during downturns.

The valuation of AI stocks has primarily been examined through the lens of expected future earnings growth and systematic risk. Babina et al. (2024) document that AI-investing firms experience higher growth in sales, employment, and market valuations, primarily through increased product innovation. Babina et al. (2023) investigate how firms' investments in AI technologies affect their market beta and risk profiles, finding significant implications for asset pricing and cost of capital. Our paper contributes to this literature by proposing a novel explanation for the high valuations of AI stocks: they may serve as hedges against negative AI singularity events. This perspective complements the traditional growth narrative by recognizing that in scenarios where advanced AI development has adverse consequences for the representative household, companies providing AI infrastructure may retain or increase their relative value, creating a hedging property that investors price into these assets.

2 Model

We now present a simple model to formalize our main argument. While deliberately stylized, it captures the essence of how AI assets might serve as hedges against a potential AI singularity.

Our economy features two types of agents. First, there are AI owners who are fully invested in AI assets. These agents are not marginal investors in the broader stock market, so their consumption patterns do not directly affect asset prices in our analysis.

Second, there is a representative household who is the marginal investor in stocks. This household has standard preferences with constant relative risk aversion γ and time discount factor β. The household's stochastic discount factor is therefore:

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

where C_t is the household's consumption at time t.

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

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

However, with probability p each period, a disaster occurs, causing consumption to drop by a factor b:

$$\log \Delta C_{t+1} = -b \quad \text{if disaster occurs} \tag{3}$$

In our context, a disaster represents a sudden improvement in AI technology that is devastating for the representative household. This can be interpreted as a worst-case scenario for AI progress, where the economy as a whole may boom, but the benefits accrue primarily to AI owners rather than the representative household. For the household, such a scenario could involve replacement of labor by AI, leading to a plunge in labor income and consequently consumption. Beyond the direct economic impact, this consumption decline can also be interpreted as a stand-in for broader losses in the household's way of life and sense of meaning.

We assume that at time $t = 0$, no disasters have yet occurred, meaning the AI singularity has not yet happened. Multiple disasters may occur over time, capturing the ongoing uncertainty that would persist even after an initial singularity event.

The key asset in our model represents publicly traded AI stocks. This asset pays a dividend D_t that follows:

$$D_t = a \cdot \exp(hN_t) \cdot C_t \tag{4}$$

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

This specification captures the idea that each time a disaster (sudden AI improvement) occurs, AI assets grow as a share of the economy. Intuitively, firms that provide the infrastructure for AI advancement—such as semiconductors, data centers, AI models, and related technologies—at least partially benefit from rapid AI progress, even if this progress is otherwise detrimental to the representative household. The parameter h controls the strength of this effect, with larger values indicating that AI assets capture more value during AI-driven disasters.

3 Asset Pricing

We now derive closed-form expressions for the price of the AI asset and analyze its implications. This will formalize the intuition that AI assets may command high valuations partly because they serve as hedges against potential AI singularity events.

The ex-dividend price of the AI asset at time t, denoted by P_t, must satisfy the standard asset pricing equation:

$$P_t = E_t[M_{t+1}(P_{t+1} + D_{t+1})] \tag{5}$$

where M_{t+1} is the stochastic discount factor and D_{t+1} is the dividend at time $t+1$. At time 0, we write:

$$P_0 = E_0[M_1(P_1 + D_1)] \tag{6}$$

We define the price-dividend ratio at time 0 as $\mathcal{P} = P_0/D_0$. To solve for $\mathcal{P}$, we need to characterize the behavior of dividends and the stochastic discount factor during normal times and disasters.

Given our model specification, the stochastic discount factor takes the following values:

$$M_1 = \beta \quad \text{with probability } 1-p \quad \text{(no disaster)} \tag{7}$$

$$M_1 = \beta e^{\gamma b} \quad \text{with probability } p \quad \text{(disaster)} \tag{8}$$

Meanwhile, dividend growth follows:

$$\frac{D_1}{D_0} = 1 \quad \text{with probability } 1-p \quad \text{(no disaster)} \tag{9}$$

$$\frac{D_1}{D_0} = e^{h-b} \quad \text{with probability } p \quad \text{(disaster)} \tag{10}$$

In a stationary environment, the price-dividend ratio remains constant over time, so $P_1/D_1 = P_0/D_0 = \mathcal{P}$. This implies $P_1 = \mathcal{P}D_1$. Substituting into the pricing equation:

$$P_0 = E_0[M_1(P_1 + D_1)] \tag{11}$$
$$= E_0[M_1((\mathcal{P} + 1)D_1)] \tag{12}$$
$$= (\mathcal{P} + 1)E_0[M_1 D_1] \tag{13}$$

Breaking down the expectation by the two possible states:

$$P_0 = (\mathcal{P} + 1)[(1 - p)\beta D_0 + p\beta e^{\gamma b} e^{h-b} D_0] \tag{14}$$
$$= (\mathcal{P} + 1)[\beta(1 - p) + \beta p e^{\gamma b+h-b}]D_0 \tag{15}$$

Since $P_0 = \mathcal{P} D_0$, we can write:

$$\mathcal{P} D_0 = (\mathcal{P} + 1)[\beta(1 - p) + \beta p e^{\gamma b+h-b}]D_0 \tag{16}$$

Dividing both sides by D_0 and rearranging:

$$\mathcal{P} = \frac{\beta(1 - p) + \beta p e^{\gamma b+h-b}}{1 - \beta(1 - p) - \beta p e^{\gamma b+h-b}} \tag{17}$$

For the price-dividend ratio to be finite, we require $\beta(1 - p) + \beta p e^{\gamma b+h-b} < 1$. This closed-form expression provides several insights into the valuation of AI assets.

The derived price-dividend ratio reveals how AI assets are valued in relation to potential AI singularity events. To understand the economic intuition, note that the term $e^{\gamma b+h-b} = e^{(\gamma-1)b+h}$ plays a crucial role. When a disaster occurs, the representative household experiences a consumption drop (by factor e^{-b}), causing their marginal utility to rise sharply (by factor $e^{\gamma b}$). Simultaneously, while the overall economy suffers, the AI asset's dividend increases relative to the economy (by factor e^h) but decreases in absolute terms (by factor e^{-b}).

This creates a valuable hedging property: precisely when the representative household's marginal utility is highest, the AI asset provides relatively higher dividends than the rest of the economy. This negative covariance between the stochastic discount factor and asset returns drives up the current price-dividend ratio, as investors value assets that pay more in bad states.

The hedging value increases with both the probability of disaster (p) and the severity of the disaster (b). It also increases with the degree to which AI assets benefit from disasters (h). This formalization supports our central thesis that AI assets may command high valuations partly because they hedge against potentially negative AI singularity events.

To illustrate these effects quantitatively, Table 1 shows how the price-dividend ratio varies with different parameter values, fixing $\gamma = 2$, $\beta = 0.96$, and $h = 0.2$.

Table 1: Price-Dividend Ratio for AI Assets

b	Disaster Probability (p)			
	0.0001	0.001	0.01	0.02
0.4	24.1	24.5	30.3	40.3
0.6	24.1	24.8	34.7	63.6
0.8	24.1	25.1	40.7	124.0
0.95	24.2	25.4	51.6	–

The table reveals several key patterns. First, even with a very low disaster probability of 0.01 (a 1% chance per period), the price-dividend ratio can be substantially higher than the baseline level of around 24 that prevails when disaster risk is minimal. Second, as the severity of disasters increases (higher b), the hedging value of AI assets increases dramatically. With $b = 0.8$ and $p = 0.02$, the price-dividend ratio reaches 124, more than five times the baseline level.

These results demonstrate that even a small probability of an AI singularity event can dramatically increase AI asset valuations if these assets provide hedging benefits during such events. This effect offers an alternative explanation for the high valuations of AI companies beyond simple expectations of future earnings growth. The market may be implicitly pricing the hedging value these assets provide against a potential AI singularity scenario, where the broader economy suffers but AI assets relatively outperform.

4 Model Discussion

Our model, while deliberately stylized, captures the essential economic mechanisms underlying AI asset valuations in the presence of potential singularity events. Here we discuss several important subtleties that inform our modeling choices and their implications.

A critical feature of our model is the implicit assumption of market incompleteness, which manifests in the disaster magnitude parameter $b > 0$. This positive value indicates that, on net, the representative household experiences a consumption decline during AI breakthrough events, despite the simultaneous increase in AI asset dividends.

The parameter b represents the combined effect of two forces: (1) the negative impact of an AI disaster on the representative household's consumption and (2) the positive performance of AI assets during such events. If markets were complete, the representative household could fully diversify against AI risk by holding shares in all AI assets, including private ones. In such a scenario, b would likely be negative, indicating that the household would experience a consumption boom rather than a disaster during rapid AI progress.

However, reality aligns more closely with our incomplete markets assumption. Most households cannot purchase 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 technologies, yet their ownership remains concentrated among a small set of investors and employees. This market structure prevents the representative household from fully hedging against—let alone benefiting from—potential AI singularity events.

One might argue that a more elaborate model could provide additional insights by explicitly modeling AI owners, private AI assets, and their interactions with the representative household. Such a model could address fascinating questions: How exactly does AI progress displace the representative household's wages? How do AI owners' incentives affect the pace of AI progress and the persistence of market incompleteness? How do preference and technology parameters influence the probability of a negative singularity?

While these questions are undoubtedly important, we believe that elaborating our model to address them would primarily decorate speculations with mathematics rather than deliver robust insights. The fundamental economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain unchanged. Moreover, the inherent unpredictability of a potential AI singularity means that any detailed modeling of its specific mechanisms would necessarily involve substantial conjecture.

We have deliberately chosen a parsimonious approach, recognizing that the benefit of reading a paper should exceed its cost. A more complex model would impose greater cognitive demands on readers without necessarily yielding proportionate insights. Our streamlined analysis allows us to clearly communicate the core economic intuition while leaving room for the human-written Appendix A, which provides valuable context about the paper's motivation and creation process.

This choice reflects our view that economics is not merely about developing elaborate mathematical frameworks but about effectively communicating insights that help us understand real-world phenomena. In the context of AI singularity risk—a topic characterized by fundamental uncertainty—clarity and accessibility are particularly valuable.

Despite our choice of parsimony, several extensions of our model could prove fruitful for future research. One promising direction would be to explicitly model the dynamic interaction between public and private AI assets, perhaps incorporating insights from the literature on innovation and creative destruction. Another would be to explore the political economy of AI regulation and how it might affect the distribution of benefits from AI progress. These extensions could provide additional insights while maintaining analytical tractability.

5 Conclusion

In this paper, we have developed a theoretical framework that offers a novel explanation for the high valuations of AI stocks. While the conventional narrative attributes these valuations primarily to expectations of future earnings growth, our model suggests an additional factor: AI assets may command premium valuations because they serve as hedges against a potential AI singularity that could be devastating for the representative investor.

Our analysis shows that even a small probability of an AI singularity event can substantially increase AI stock valuations if these assets are expected to outperform the broader economy during such events. This hedging property arises because AI companies—those providing infrastructure, computing resources, and development platforms—may capture a larger share of economic value precisely when a rapid AI advancement negatively impacts the representative household's consumption. The mathematical framework we developed demonstrates how this negative covariance between the stochastic discount factor and AI asset returns can significantly increase their price-dividend ratios.

An important implication of our analysis is that financial markets could provide a partial solution to AI catastrophe risk concerns. By investing in AI assets, households can partially hedge against negative AI singularity scenarios, offering an alternative or complement to policy proposals like Universal Basic Income that are frequently discussed in the context of AI-driven labor displacement. However, it is crucial to recognize that the effectiveness of this hedging strategy is fundamentally limited by market incompleteness. As our model highlights, many cutting-edge AI enterprises remain privately held, preventing the representative household from fully diversifying against AI risk.

This market-based approach to addressing AI catastrophe risk has received insufficient attention in the literature on AI safety and existential risk. While scholars like Bostrom (2014) and Bengio et al. (2024) have thoroughly examined the technical and governance aspects of AI risk, the potential role of financial markets in mitigating these risks remains underexplored. Similarly, economic analyses such as Jones (2024) and Korinek and Suh (2024) focus primarily on the trade-offs between growth and risk or on the implications for labor markets, rather than on how financial innovation might help distribute the risks and benefits of AI progress more broadly.

Future research could explore how reducing market incompleteness through financial innovation might improve risk sharing, or how existing financial instruments could be repurposed to hedge against AI singularity risks. As AI development continues to accelerate, understanding these financial dimensions becomes increasingly important for both investors and policymakers concerned with managing the potential disruptions of transformative AI.

References

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson (2024). "Artificial intelligence, firm growth, and product innovation". In: *Journal of Financial Economics* 151, Article 103745.

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

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

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, et al. (2024). "Managing extreme AI risks amid rapid progress". In: *Science* 384.6698. URL: `https://arxiv.org/abs/2310.17688`.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Chollet, François, Mike Knoop, Gregory Kamradt, and Bryan Landers (2024). "ARC Prize 2024: Technical Report". In: *arXiv preprint*.

DeepSeek-AI et al. (Jan. 2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". In: *arXiv*. URL: `https://arxiv.org/abs/2501.12948`.

Jones, Charles I. (2024). "The AI Dilemma: Growth versus Existential Risk". In: URL: `https://web.stanford.edu/~chadj/existentialrisk.pdf`.

Korinek, Anton and Donghyun Suh (2024). *Scenarios for the Transition to AGI*. Tech. rep. NBER Working Paper.

Rietz, Thomas (1988). "The Equity Risk Premium: A Solution?" In: *Journal of Monetary Economics*.

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 occassionaly run into mis-citations. Every 15 to 20 citations, I would see a cite that mis- or overinterprets the cited paper. I suppose that's not so different than the human-written citation error rate. But I hate finding misinterpretations in the literature so I purposefully limited the number of cites in the paper.

AI Model Selection

o1 did the theory, and sonnet thinking did the writing. It's well known that these are the strengths of these two models.

Sonnet thinking 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 may
projects.

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

B Prompts Used to Generate This Paper

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

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

The system prompt and instructions are listed below.

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

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

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

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

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

Be modest. Do not overclaim.

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

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

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

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

Use the following outline:
```

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

```
Do not:
- Use bullet points or numbered lists
```

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

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

Do not:
- Restate the assumptions
- Assume any variable is constant or stationary (prove it)

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

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

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

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

Make a table for the risk premium (expected return - risk-free rate)
    in percent (100*(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.
```