Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

We propose that AI stocks may be valued highly in part because they hedge against a negative AI singularity—a scenario where rapid AI advancement displaces human labor and reduces household welfare. Our model demonstrates that even with a small probability of such an event, the insurance value of AI assets can significantly contribute to their valuations. Financial markets provide a natural but incomplete mechanism for addressing some aspects of AI catastrophe risk, as most households cannot invest in cutting-edge private AI labs. This paper was generated by prompting large language models, with minimal human input beyond initial conceptualization and editing.

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

In recent years, artificial intelligence capabilities have advanced at a breathtaking pace. DeepSeek AI has developed the R1 model that rivals state-of-the-art reasoning capabilities (DeepSeek-AI et al., 2025). Open AI's o1 model achieved an 87.5% score on the ARC AGI benchmark (Pfister and Jud, 2025), a test designed specifically to measure general intelligence rather than specialized skills. Meanwhile, Waymo's autonomous vehicles have logged millions of miles without human intervention. These developments are not merely incremental improvements but represent fundamental shifts in what machines can accomplish, leading many investors to wonder about the implications for their careers and financial security.

Technological change has, of course, occurred throughout human history. The Industrial Revolution displaced craftsmen, and computerization eliminated many clerical jobs. However, artificial intelligence stands apart from previous technological transformations in a crucial way: there is, in principle, no product or service that AI could not eventually create. This paper itself illustrates this point—it was generated entirely by AI systems using a series of engineered prompts, with minimal human input beyond the initial conceptualization and editing (see `https://github.com/chenandrewy/Prompts-to-Paper/` for details). What's more, many scholars have raised the possibility that AI progress could be extraordinarily sudden and self-reinforcing, potentially leading to what has been termed the technological singularity—a hypothetical point at which artificial intelligence surpasses human intelligence and accelerates technological growth beyond our ability to predict or control (Vinge, 1993; Kurzweil, 2005; Bostrom, 2014).

This paper explores an alternative explanation for the high valuations of AI stocks. While conventional wisdom attributes these valuations primarily to expectations of future earnings growth, we propose that AI stocks may also be valued for their hedging properties against a negative AI singularity. Our model demonstrates that even with a small probability of an AI disaster, the insurance value that AI stocks provide can significantly contribute to their current valuations.[1]

The key insight from our analysis is that AI assets may serve as partial hedges against scenarios where rapid AI advancement is detrimental to the representative household. In a world where AI suddenly improves to the point of displacing substantial human labor, publicly traded AI companies—those providing the infrastructure, hardware, and components that enable AI development—stand to capture an increasing share of economic value, even as households face declining labor income. This hedging property becomes more valuable as

[1] "We" refers to one human author and multiple large language models that collaborated on this paper. A purely human perspective on this work is provided in Appendix A.

either the probability or severity of potential AI disasters increases.

We should emphasize two important caveats. First, we are not claiming that a negative singularity will happen. The probability of such an outcome remains highly uncertain, and there are ongoing efforts to ensure AI development proceeds safely (Bengio et al., 2024). Nevertheless, it is important to consider this scenario in our financial models, just as we consider other rare disaster risks (Barro, 2006; Wachter, 2013). Second, we are not asserting that this hedging value is already fully priced into AI stocks. Our model simply illustrates a potential mechanism that could contribute to AI asset valuations.

Recent literature on the economic implications of artificial intelligence has highlighted various mechanisms through which AI can affect asset pricing and market dynamics. Jones (2024) explores the economic tension between AI-driven growth and potential existential risks, providing a framework for analyzing the tradeoffs between technological progress and catastrophic outcomes. Similarly, Korinek and Suh (2024) analyze how output and wages respond to different AI development scenarios, examining potential paths as technology approaches artificial general intelligence (AGI). The relationship between AI and firm risk has been directly examined by Babina et al. (2023), who provide evidence that firms' investments in AI technologies significantly affect their systematic risk profiles, with implications for asset pricing, growth options, and cost of capital. This emerging research establishes the fundamental relationship between AI adoption and market risk exposure but has not yet explored the implications for AI asset valuations in the context of potential negative AI singularity.

The hedging properties of technological assets have been studied in other contexts. Zhang (2019) demonstrates that firms with routine-task labor maintain a replacement option that hedges their value against unfavorable macroeconomic shocks, lowering their expected returns. While Zhang focuses on the benefits to firms of having the option to replace labor with technology, our paper flips this perspective to consider how investors might value technology stocks partially for their hedging properties against labor displacement. The rare disaster literature, pioneered by Rietz (1988) and extended by Barro (2006) and Wachter (2013), provides a methodological foundation for our approach. However, while this literature typically treats disasters as exogenous economic contractions affecting all assets similarly, we introduce a novel perspective by considering a disaster that specifically benefits AI assets relative to labor income. Our contribution is thus to propose a new way to think about AI stock valuations—not merely as claims on future earnings growth, but as potential hedges against a negative AI singularity that could devastate labor income while potentially benefiting the owners of AI technology.

2 Model

We now present a simple model to formalize our argument that AI stocks may be valued highly in part because they hedge against a negative AI singularity. The model is deliberately stylized to capture the essence of our main argument while remaining tractable.

Our economy consists of 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 constant relative risk aversion preferences with coefficient γ and time discount factor β.

The consumption growth process for the representative household is subject to disaster risk, which in our context represents a sudden improvement in AI that is devastating for the household. Specifically:

$$\log \Delta c_{t+1} = \begin{cases} 0 & \text{if no disaster occurs} \\ -b & \text{if a disaster occurs (with probability } p) \end{cases} \tag{1}$$

A disaster in this context represents a worst-case scenario for AI progress. While the overall economy might boom during such an event, the value created is primarily captured by AI owners. For the representative household, the disaster manifests as labor being replaced by AI, causing labor income and consumption to plummet. Beyond the direct economic impact, households may also experience losses in their way of life and sense of meaning. The consumption decline parameter b can be interpreted as capturing both the economic and non-economic aspects of these losses.

At time $t = 0$, we assume no disasters have yet occurred, meaning the AI singularity has not yet taken place. Our model allows for multiple disasters to occur over time, representing ongoing uncertainty even after an initial singularity event.

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

$$D_t = a \exp^{hN_t} C_t \tag{2}$$

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

when disasters occur.

This dividend specification captures the idea that firms providing semiconductors, data, AI models, and related technologies at least partially benefit from sudden improvements in AI capabilities. When a disaster occurs (N_t increases), the AI asset's dividends grow relative to aggregate consumption. This reflects the increasing share of economic value captured by AI-related firms during periods when AI capabilities advance dramatically, even as these advances may be detrimental to the representative household.

3 Results

Having specified the consumption and dividend processes, we now analyze the asset pricing implications of our model. We derive the price/dividend ratio for the AI asset and examine how it varies with the probability and severity of the AI disaster.

In the standard consumption-based asset pricing framework, an asset's current price equals the expected discounted value of its future payoffs. For a dividend-paying asset, the time-0 price P_0 satisfies:

$$P_0 = E_0[M_{0,1}(P_1 + D_1)] \tag{3}$$

where $M_{0,1}$ is the stochastic discount factor from time 0 to time 1, and D_1 is the dividend paid at time 1. Iterating forward, the price can be expressed as the discounted sum of all future dividends:

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

where $M_{0,t}$ denotes the discount factor from time 0 to t.

With CRRA preferences, the one-period stochastic discount factor is $M_{0,1} = \beta(C_1/C_0)^{-\gamma}$. Under our consumption process, this becomes:

$$M_{0,1} = \begin{cases} \beta & \text{if no disaster occurs} \\ \beta e^{\gamma b} & \text{if a disaster occurs} \end{cases} \tag{5}$$

For the AI asset, the dividend growth rate $G_D = D_1/D_0$ is:

$$G_D = \begin{cases} 1 & \text{if no disaster occurs} \\ e^{h-b} & \text{if a disaster occurs} \end{cases} \tag{6}$$

Given our i.i.d. assumption for the economic environment, the price/dividend ratio $x = P_t/D_t$ remains constant over time. At time 0, we can express this ratio as:

$$x = \frac{E_0[M_{0,1}G_D]}{1 - E_0[M_{0,1}G_D]} \tag{7}$$

provided that $E_0[M_{0,1}G_D] < 1$ to ensure convergence.

Computing the expectation:

$$E_0[M_{0,1}G_D] = \beta[(1-p) + p \cdot e^{b+h}] \tag{8}$$

The convergence condition requires:

$$\beta[(1-p) + p \cdot e^{b+h}] < 1 \tag{9}$$

This condition has an intuitive interpretation: the expected discounted dividend growth must not be too high. For our baseline parameter values ($\beta = 0.96$, $\gamma = 2$, $h = 0.2$), and with $p = 0.01$, the condition becomes $b < 1.44$, meaning the consumption decline in a disaster cannot be too severe, or the price would become unbounded.

Table 1 presents the price/dividend ratios for different combinations of disaster probability (p) and consumption decline severity (b), holding $h = 0.2$ fixed.

Table 1: Price/Dividend Ratios

	$p = 0.0001$	$p = 0.001$	$p = 0.01$	$p = 0.02$
$b = 0.4$	24.04	24.53	30.25	39.00
$b = 0.6$	24.09	24.76	34.71	63.60
$b = 0.8$	24.10	25.06	41.57	141.86
$b = 0.95$	24.13	25.35	51.63	–

The table reveals a striking pattern: as either the probability (p) or the severity (b) of an AI disaster increases, the price/dividend ratio of AI assets rises substantially. This occurs

because AI assets serve as a hedge against disasters that harm the representative household. During a disaster, while household consumption falls by e^{-b}, AI assets' dividends decline by only e^{h-b} (or even increase relative to total consumption if $h > b$), making them relatively more valuable in those states of the world.

This hedging property becomes particularly valuable as the severity of potential disasters increases. For instance, with a disaster probability of just 1% and a consumption decline of 80% ($b = 0.8$), the price/dividend ratio reaches 41.57, far higher than the 24.10 ratio that would prevail with a minuscule 0.01% probability.

Our analysis provides a new perspective on AI stock valuations. The high valuations observed in the market might reflect not just optimism about future earnings growth, but also the insurance value these stocks provide against a potential negative AI singularity. Even with a small probability of a severe AI disaster, the hedging value can significantly contribute to current valuations.

4 Model Discussion

Our model captures the essence of how AI stocks may hedge against a negative AI singularity, but like any model, it makes simplifying assumptions that deserve further discussion. We now explore some of the subtleties and limitations of our approach.

Market incompleteness plays a crucial role in our analysis, though we do not explicitly model it. This incompleteness is implicitly captured by the disaster magnitude parameter $b > 0$. The parameter b represents the net effect of two forces: the negative impact of an AI disaster on the representative household and the partial hedge provided by publicly traded AI assets. If markets were complete, the representative household could purchase shares in all AI assets, including private ones, and would not only fully hedge against but potentially benefit from an AI singularity. In such a scenario, b would be negative, representing a consumption boom rather than a disaster.

However, reality aligns more closely with our model's assumptions. Most households cannot invest 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 and would likely capture substantial value during a singularity event. The inability of the representative household to access these investments creates the market incompleteness that our model implicitly assumes.

A more elaborate model could certainly add detail to the AI owners, private AI assets, and their interactions with the representative household. Such a model might address questions like: How exactly does AI progress displace the representative household's wages? How do AI

owners' incentives affect both AI progress and market incompleteness? How do preferences and technology parameters affect the odds of a negative singularity?

While these questions are fascinating, we believe that addressing them would primarily decorate speculations with mathematics. The core economic mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain fundamentally the same. Moreover, a more complex model would significantly increase the cognitive cost for readers without necessarily providing proportional insights.

In our view, the benefit of reading a paper should exceed the cost. Our simplified approach allows us to convey the key insight—that AI stocks may be valued highly in part because they hedge against a negative AI singularity—without burdening readers with excessive technical details. This brevity also allows room for the human-written Appendix A, which provides context and reflection on the paper's creation process.

The simplicity of our model is a feature, not a bug. It focuses attention on the central economic mechanism while acknowledging the inherent unpredictability of an AI singularity. Given this unpredictability, a more detailed model might create a false sense of precision about something that is fundamentally uncertain.

5 Conclusion

In this paper, we have explored an alternative explanation for the high valuations of AI stocks. While conventional wisdom attributes these valuations primarily to expectations of future earnings growth, we propose that AI stocks may also be valued for their hedging properties against a negative AI singularity. Our model demonstrates that even with a small probability of an AI disaster, the insurance value that AI stocks provide can significantly contribute to their current valuations.

The key insight from our analysis is that AI assets may serve as partial hedges against scenarios where rapid AI advancement is detrimental to the representative household. As AI capabilities improve dramatically, publicly traded AI companies—those providing the infrastructure, hardware, and components that enable AI development—stand to capture an increasing share of economic value, even as households face declining labor income. This hedging property becomes more valuable as either the probability or severity of potential AI disasters increases.

Financial markets provide a natural mechanism for addressing some aspects of AI catastrophe risk that complements other proposed solutions such as universal basic income. By investing in AI-related assets, households can obtain partial insurance against negative AI outcomes. However, our analysis highlights that this hedging strategy is fundamentally

limited by market incompleteness. Most households cannot invest in cutting-edge private AI labs that would likely capture substantial value during a singularity event, and publicly traded AI companies represent only a fraction of the potential value created in such scenarios.

The role of financial markets in mitigating AI risk has received relatively little attention in the AI catastrophe risk literature. While researchers like Bengio et al. (2024) and Bostrom (2014) have extensively discussed governance approaches and technical safety measures, financial market mechanisms remain underexplored. Similarly, economic analyses like Jones (2024) and Korinek and Suh (2024) have examined the tension between AI-driven growth and potential risks, but have not fully addressed how capital markets might help distribute the gains and losses from AI advancement.

Future research could explore how to improve market completeness in this domain, potentially through new financial products that provide more direct exposure to AI development outcomes. Such innovations might help broaden access to AI-related hedging opportunities beyond those currently available through publicly traded stocks. However, these market-based approaches should be viewed as complementary to, rather than substitutes for, other risk mitigation strategies including technical safety research and appropriate governance frameworks.

References

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

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

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

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

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.

Kurzweil, Ray (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking Press.

Pfister, Rolf and Hansueli Jud (2025). "Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI". In: *arXiv preprint*.

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

Vinge, Vernor (1993). "The Coming Technological Singularity". In: *Department of Mathematical Sciences, San Diego State University*.

Wachter, Jessica A. (2013). "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?" In: *Journal of Finance*.

Zhang, Miao Ben (2019). "Labor-Technology Substitution: Implications for Asset Pricing". In: *Journal of Finance* 74.4, pp. 1793–1839.

A A Purely Human Perspective

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

```
# Prompts-to-Paper

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

- make-paper.py writes a paper

- plan0408-piecewise.yaml contains the prompts

- make-many-papers.py runs make-paper.py many times.

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

# 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. 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 (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.

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: "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 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, 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 even if it's unlikely, or highly unlikely, it's 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-abstract-prose (model: claude-3-7-sonnet-20250219)

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

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

Do not:
- Emphasize consumption
- Oversell or overinterpret the model
```

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

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

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

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

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