Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

Abstract

We propose that high AI stock valuations may partially reflect their hedging value against a negative AI singularity—an explosion of AI development that is devastating for the representative investor. Our parsimonious model demonstrates how AI stocks can command premium valuations when they serve as hedges against scenarios where AI breakthroughs reduce household welfare while increasing AI assets' share of the economy. This perspective complements the conventional growth narrative and highlights financial markets' potential role in addressing AI catastrophe risk, though market incompleteness limits this hedging capability. 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 months have witnessed the release of powerful reasoning models like OpenAI's o1 (Zhong et al., 2024) and DeepSeek's R1 (DeepSeek-AI et al., 2025), able to solve complex mathematical problems and engage in sophisticated logical reasoning. Performance on benchmarks like the Abstraction and Reasoning Corpus (ARC-AGI Challenge) has jumped from 33% to over 55% in just one year (Chollet et al., 2024). Meanwhile, autonomous vehicles from companies like Waymo have logged tens of millions of miles with minimal human intervention. As these technologies continue to advance, investors and workers alike have grown increasingly concerned about AI's potential to displace human labor and dramatically reshape the economic landscape.[1]

Technological change has occurred throughout human history, from the Industrial Revolution to the digital era. However, AI represents something fundamentally different. Unlike previous technologies that excelled in specific domains, AI has the potential to create virtually any product or service that humans can provide. Indeed, this very paper was generated entirely through AI, using a series of engineered prompts (available at `https://github.com/chenandrewy/Prompts-to-Paper/`). Perhaps most concerning is the possibility of an "AI singularity" – a point at which AI improvement becomes recursive and self-accelerating, potentially leading to superintelligent systems that vastly exceed human capabilities (Vinge, 1993; Kurzweil, 2005).

In this paper, we investigate a novel perspective on AI stock valuations. While conventional wisdom attributes high AI valuations to expectations of future earnings growth, we propose that these valuations may also reflect a hedging value against a potential negative AI singularity. Specifically, we develop a parsimonious model where a sudden breakthrough in AI technologies could be devastating for the representative household, significantly reducing their consumption. Yet in this same scenario, AI assets would increase as a share of the economy. This creates a natural hedging opportunity: by holding AI stocks, investors can partially insure themselves against the very technological disruption that might otherwise devastate their livelihoods.

We must emphasize two important caveats. First, we are not claiming that a negative AI singularity will happen. The likelihood and timing of such an event remain deeply uncertain. Nevertheless, it is prudent to consider potential economic implications of such a scenario, just as economists study other low-probability, high-impact events such as financial crises or pandemics. Second, we are not asserting that this hedging value is already fully priced into

[1] Throughout this paper, "we" refers to one human author and multiple LLMs. A purely human perspective on this research is provided in Appendix A.

current AI stock valuations. Rather, our model illustrates a possible mechanism through which rational investors might value AI assets partly for their insurance properties against the worst-case scenarios of AI development.

Recent literature has emphasized both the transformative potential and potential risks of artificial intelligence. Jones (2024) explores the economic trade-off between AI-driven growth and potential existential risks, providing a framework for analyzing this tension. Similarly, Korinek and Suh (2024) examine economic scenarios as technological advancement approaches artificial general intelligence, showing how labor income and overall economic welfare might evolve under different AI development paths. The impact of technology on labor markets has been explored in several studies, with Zhang (2019) documenting how firms with high exposure to routine-task labor that can be automated maintain a technological replacement option that hedges their value against unfavorable macroeconomic shocks. This literature has established the dual nature of automation technologies—creating both opportunity and displacement risk.

The financial market implications of AI investments have also begun receiving attention. Babina et al. (2023) provide evidence that firms' investments in AI technologies affect their systematic risk profiles, with implications for their market exposures and cost of capital. Our work contributes to this literature by proposing a novel perspective on AI asset valuation: specifically, that AI stocks might be valued partly as hedges against a negative AI singularity rather than solely for their growth potential. While the traditional narrative views high AI valuations as reflecting expected future earnings growth, our model suggests they may also reflect the insurance value these assets provide against worst-case AI development scenarios. In such scenarios, the average investor faces devastating consumption declines, but AI assets increase as a share of the economy. This hedging perspective helps explain why AI stock valuations might remain elevated even when broader concerns about AI risks increase, offering a complementary explanation to standard growth narratives.

2 Model

In our model, we aim to capture the essence of how AI stocks may hedge against a potential AI singularity. We deliberately keep the model simple to highlight the main mechanism without unnecessary complexity.

We consider an economy with two types of agents. First, there are AI owners who are fully invested in AI assets. These agents are not marginal investors in the stock market. Second, there is a representative household who is the marginal investor in stocks. This household has standard CRRA preferences with risk aversion parameter γ and time discount factor β.

Since the household is the marginal investor, only their consumption matters for our asset pricing analysis.

The consumption growth process for the representative household is straightforward. In normal times, log consumption growth is zero:

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

However, with probability p, a disaster occurs, leading to:

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

In our context, a disaster represents a sudden improvement in AI that is devastating for the representative household. This can be thought of as a worst-case scenario for AI progress from the household's perspective. In such a scenario, the overall economy might boom, but the value created is captured primarily by AI owners. For the representative household, labor is replaced by AI, causing labor income and consumption to plummet. Beyond the direct economic impact, households may also lose their way of life and sense of meaning. The consumption decline can be viewed as a stand-in for these broader losses.

We assume that at $t = 0$, no disasters have occurred yet, meaning the singularity has not taken place. Our model allows for multiple disasters to occur over time, capturing the ongoing uncertainty that would persist even after an initial singularity event.

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

$$D_t = a e^{h N_t} C_t \tag{3}$$

This specification has several important features. 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 $h > 0$ implies that each time a disaster occurs, the AI asset grows as a share of the economy. Intuitively, firms that provide semiconductors, data, AI models, and related infrastructure at least partially benefit from sudden improvements in AI technology, even as these same improvements may harm the broader population.

This simple framework allows us to explore how the possibility of an AI singularity might affect the valuation of AI stocks through a hedging channel, distinct from the standard growth narrative.

3 Asset Pricing Implications

Now that we have established our model, we turn to its asset pricing implications. We are particularly interested in how the possibility of an AI singularity affects the valuation of AI stocks through their hedging properties.

3.1 Price-Dividend Ratio of AI Assets

We begin by deriving the price-dividend ratio of the AI asset. Let P_t be the ex-dividend price of the AI asset at time t. By standard consumption-based asset pricing theory, the price equals the expected discounted value of future dividends:

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

where $M_{0,j} = \beta^j (C_j/C_0)^{-\gamma}$ is the stochastic discount factor from time 0 to time j.

Defining the price-dividend ratio as $x_0 \equiv P_0/D_0$, we can rewrite this as:

$$x_0 = E_0 \left[\sum_{j=1}^{\infty} M_{0,j} \frac{D_j}{D_0} \right] \tag{5}$$

Given our dividend process $D_t = ae^{hN_t} C_t$ and the fact that $N_0 = 0$, we have:

$$\frac{D_j}{D_0} = e^{hN_j} \frac{C_j}{C_0} \tag{6}$$

Substituting this and the expression for $M_{0,j}$ into our equation for x_0:

$$x_0 = \sum_{j=1}^{\infty} \beta^j E_0 \left[\left(\frac{C_j}{C_0} \right)^{1-\gamma} e^{hN_j} \right] \tag{7}$$

From our consumption process, we know that $(C_j/C_0) = e^{-bN_j}$, which implies $(C_j/C_0)^{1-\gamma} = e^{-b(1-\gamma)N_j}$. This gives us:

$$x_0 = \sum_{j=1}^{\infty} \beta^j E_0 \left[e^{[h-b(1-\gamma)]N_j} \right] \tag{8}$$

Let's define $\sigma \equiv h - b(1-\gamma)$ for notational convenience. Given that N_j follows a binomial distribution with probability p of disaster in each period, we have:

$$E_0 \left[e^{\sigma N_j} \right] = [(1-p) + pe^{\sigma}]^j \tag{9}$$

Thus, our price-dividend ratio becomes:

$$x_0 = \sum_{j=1}^{\infty} \beta^j [(1-p) + pe^{\sigma}]^j \tag{10}$$

This is a geometric series that converges when $\beta[(1-p)+pe^{\sigma}] < 1$. Summing the series, we obtain:

$$x_0 = \frac{\beta[(1-p) + pe^{h-b(1-\gamma)}]}{1 - \beta[(1-p) + pe^{h-b(1-\gamma)}]} \tag{11}$$

This formula provides a closed-form solution for the price-dividend ratio of the AI asset at time 0.

3.2 Interpretation and Implications

The expression for the price-dividend ratio offers several insights. The key parameter $\sigma = h - b(1-\gamma)$ combines three crucial elements: the growth rate of AI dividends during disasters (h), the magnitude of consumption decline during disasters (b), and the risk aversion of the representative investor (γ).

When $\gamma > 1$, which is consistent with most empirical estimates, the term $-b(1-\gamma)$ becomes positive. This means that higher values of b (more severe disasters) can actually increase the price-dividend ratio. This counterintuitive result occurs because with high risk aversion, the representative investor is willing to pay a premium for assets that perform well during disasters.

The AI asset provides such a hedge through the parameter h. When $h > 0$, the dividend share of AI assets increases during disaster periods, making them valuable hedges against the consumption decline that the representative investor experiences. In essence, the investor is willing to accept a lower expected return on AI assets because they provide insurance against the worst-case AI scenario.

3.3 Numerical Illustration

To illustrate these effects quantitatively, we compute the price-dividend ratio for various parameter combinations. We set $\gamma = 2$, $\beta = 0.96$, and $h = 0.2$ as baseline values, and then vary the disaster probability p and the consumption disaster magnitude b. For these parameter values, the price-dividend ratio converges when $b + h < 1.64$.

Table 1 presents the price-dividend ratios for different combinations of p and b:

These results reveal several important patterns. First, holding the disaster probability p constant, the price-dividend ratio increases with the disaster magnitude b. This confirms our

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

p	b			
	0.4	0.6	0.8	0.95
0.0001	24.1	24.1	24.1	24.1
0.001	24.5	24.8	25.1	25.3
0.01	30.3	34.7	40.7	51.6
0.02	39.0	61.5	141.9	-

theoretical insight that when $\gamma > 1$, more severe AI disasters can lead to higher valuations for AI assets that hedge against those disasters.

Second, for any given disaster magnitude, the price-dividend ratio increases with the probability of disaster p. This suggests that as the perceived likelihood of an AI singularity rises, so too might the valuations of AI-related assets.

Third, the effect of disaster risk on valuations becomes more pronounced as both p and b increase. For example, with a modest disaster probability of $p = 0.01$ and a relatively severe disaster magnitude of $b = 0.95$, the price-dividend ratio reaches 51.6, more than double the ratio when disaster risk is negligible.

These findings provide theoretical support for our main argument: the high valuations of AI stocks could stem in part from their role as hedges against a potential AI singularity that would be devastating for the average investor. The traditional growth narrative views high AI valuations as reflecting future earnings potential in good states of the world. Our model suggests an alternative, complementary explanation: these valuations might also reflect the hedging value of AI assets against the worst-case scenarios of AI development.

This hedging interpretation is particularly relevant in the context of AI, where technological advancement contains both promise and peril. The representative investor, fearing displacement in an AI-dominated economy, may rationally bid up the prices of assets that would retain or increase their value if such a scenario materializes.

4 Model Discussion

Our model, while deliberately simplified, captures the essence of how AI stocks might serve as hedges against a potential AI singularity. However, it's important to acknowledge several subtleties that lie beneath the surface of our formal analysis.

A key aspect not explicitly modeled is market incompleteness, which is nonetheless central to our narrative. 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 positive effect on AI asset dividends. If markets were complete, the representative household could fully participate in all AI-related gains by purchasing shares in all AI assets, including private ones. In such a scenario, the household would not just hedge against but potentially benefit from an AI 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, which are at the frontier of AI development, remain largely inaccessible to the average investor. This market structure is consistent with our model's assumption that the representative household cannot fully hedge against AI risk, leading to $b > 0$ for this agent.

One might argue that a more elaborate model could add detail to the AI owners, private AI assets, and their interactions with the representative household. Such a model could address intriguing questions: How exactly does AI progress displace the representative household's wages? How do AI owners' incentives affect AI progress and perpetuate market incompleteness? How do preferences and technology parameters affect the odds of a negative singularity?

However, we believe that such elaborations would primarily decorate speculations with mathematics without fundamentally altering the core economics. The essential mechanisms—rare disaster risk, hedging motives, and market incompleteness—would remain the same. Moreover, a more complex model would be significantly more costly for readers to digest. In our view, the benefit of reading a paper should exceed the cost, and our parsimonious approach strikes this balance.

The simplicity of our model also allows room for the human-written Appendix A, which provides additional context and reflections on the research process. This appendix offers insights that complement our formal analysis without being constrained by mathematical formalism.

In sum, while our model abstracts from many details, it captures the fundamental economic forces at play when considering AI stocks as hedges against a negative AI singularity. The model's simplicity is a feature, not a bug, allowing us to highlight the key mechanism without unnecessary complexity.

5 Conclusion

In this paper, we have proposed a novel perspective on the valuation of AI stocks. While the conventional wisdom attributes high AI valuations to expectations of future earnings growth, we argue that these valuations may also reflect the hedging value of AI assets against a potential AI singularity. Our parsimonious model demonstrates how AI stocks can command premium valuations when they serve as hedges against AI disasters that would be devastating for the representative household.

The key insight from our model is that when risk aversion exceeds unity, which aligns with empirical estimates, more severe potential AI disasters can actually increase the valuations of AI stocks. This counterintuitive result emerges because investors are willing to pay a premium for assets that perform relatively well in disaster states. Our numerical illustrations confirm that as both the probability and magnitude of potential AI disasters increase, so too does the price-dividend ratio of AI assets.

Beyond theoretical contributions, our work points to the importance of financial markets in addressing AI catastrophe risk. While much of the discussion around mitigating the negative distributional consequences of AI focuses on policy solutions like universal basic income, financial markets may offer complementary approaches through risk-sharing mechanisms. Individual investors can, at least partially, hedge against adverse AI developments by including AI-related assets in their portfolios.

However, this hedging capability is fundamentally limited by market incompleteness. As noted in our discussion, many frontier AI labs remain privately held, preventing the representative household from fully participating in potential AI gains. This market structure reinforces the positive value of the disaster magnitude parameter (b) in our model and explains why AI developments can still constitute a ”disaster” for the average investor despite potentially expanding the overall economic pie.

Interestingly, these financial market solutions to AI catastrophe risk receive insufficient attention in the broader AI risk literature. Jones (2024) explores the economic trade-off between AI-driven growth and existential risks, but does not extensively explore how financial markets might help manage these risks. Similarly, Korinek and Suh (2024) examine economic scenarios approaching artificial general intelligence without fully addressing the role of financial markets in risk allocation. Even comprehensive works on AI risk by Bostrom (2014) and Russell (2019) focus primarily on technical alignment and governance solutions rather than financial mechanisms.

Financial markets, though imperfect due to inherent incompleteness, offer valuable mechanisms for distributing AI risk. As AI development continues to accelerate, understanding these market dynamics becomes increasingly important for investors, policymakers, and researchers alike. Future research should further explore how various financial instruments might help society navigate the uncertain path toward advanced AI, complementing technical and governance approaches to ensuring that AI progress benefits humanity broadly.

References

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

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.

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

Zhong, Tianyang et al. (Sept. 2024). “Evaluation of OpenAI o1: Opportunities and Challenges of AGI”. In: *arXiv*. URL: `https://arxiv.org/abs/2409.18486`.

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.  So, I thought I should use AI to write the
paper.

## 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. 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

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 show I did substantial hand-holding. The Github commits show even more human labor. They tell the story of me getting to know my "co-authors." It was hard to communicate style issues and LaTeX instructions, 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 will be equal co-authors.

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.

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.

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. Building off OpenAI Deep Research and Claude Web Search led to zero hallucinations in the paper drafts.

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

I found it best to use o1 for theory, and Sonnet for writing. It's well-known that these are the strengths of these two models.

I briefly tried having Llama 3.1 405b do the writing. It was terrible! The cutting edge models seemed necessary.

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. AI progress is too fast to update this paper continuously.

Picking the best of N papers

The writing quality varies across each run. Rather than try to prompt engineer an error free, insightful paper, I decided to just generate N papers and choose the best one.

5 drafts of the paper can be found in `./manyout0408-pdf/`. They're broadly similar and I would be willing to put my name on most of them.

I ended up choosing `paper-run-05.pdf`. The abstract overemphasizes the model, but otherwise I like the paper. I hope you find the paper 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 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 an economic model does not capture an important channel. This is common (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.

AI also struggles to generate a satisfying economic model on its own. Its models are typically either too simplistic or too complicated.

I opted not to add empirical work or numerically-solved models. AI can do both, but both require connecting to the outside world, and a plethora of technical challenges.

Last, I often found the literature discussions insufficiently careful (as discussed above).

But 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 questions 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 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 highly unlikely, it's still a scenario that economists should consider.
```

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*(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
```