AI & Humanity

AI 109

Spring 2026

Last Lecture

- More details about the final.
- Some “big questions” about AI.
- Final course takeaways.
- Where to go from here.

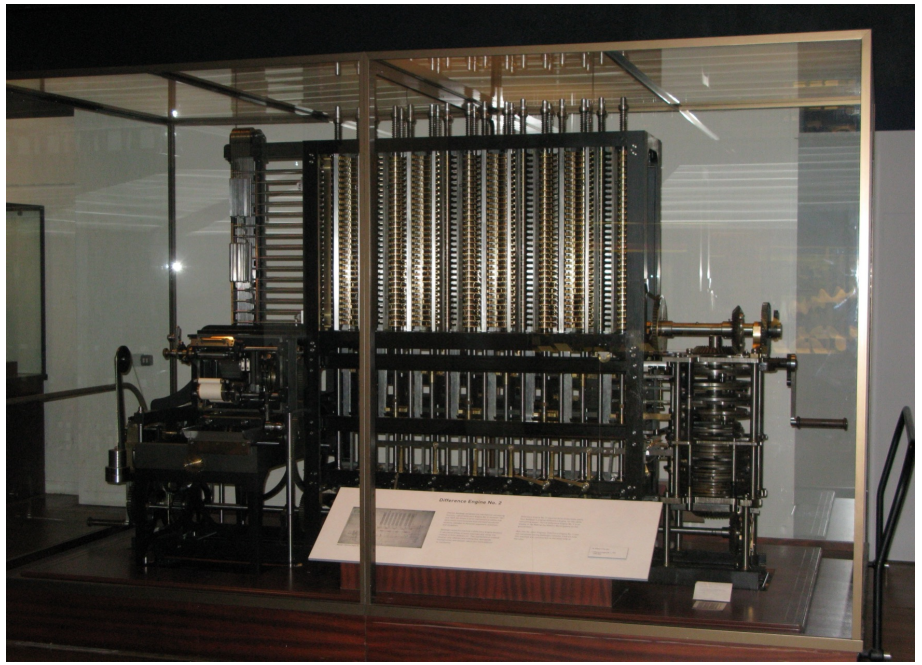
- Assignment 3 DUE TONIGHT
  - Your websites are public. If you don’t want the world to see them, you should make your repository private ***after the final.***
  - I’m excited to share what you’ve done with other people on campus and in Silicon Valley (if you leave them up).

Final Exam

- At the normal exam time in our normal classroom.
- Use the slides that you submit for Assignment 3 for your presentation.
- 7 minutes.
- Focus on:
  - What you're interested in doing for a career
  - How AI might present challenges to that line of work.
  - How you can use AI to gain an advantage in that work.
- Particularly interested in the last item.
- I'll set the speaker order randomly and collect the slides into a single deck.

150 Years of Progress

- “The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths. Its province is to assist us to making available what we are already acquainted with.”
    - Ada Lovelace, 1843



150 Years of Progress

- “You insist that there is something that a machine can't do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that.”
    - John von Neumann, 1948

Robotics Continues To Accelerate…

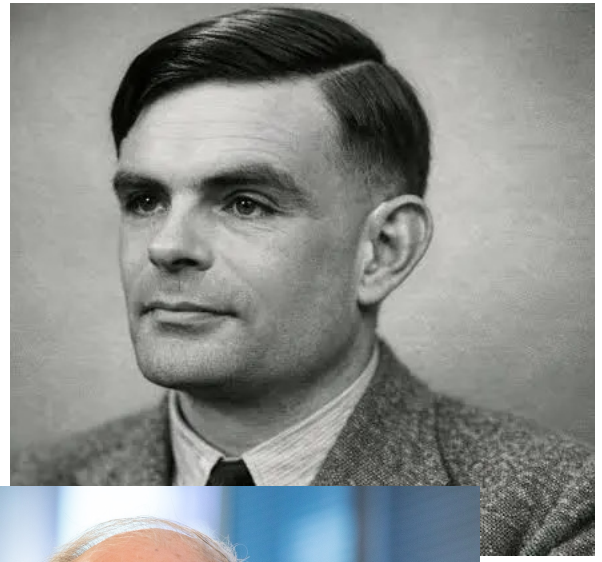
- Announced yesterday
- 5x cheaper than last lecture!

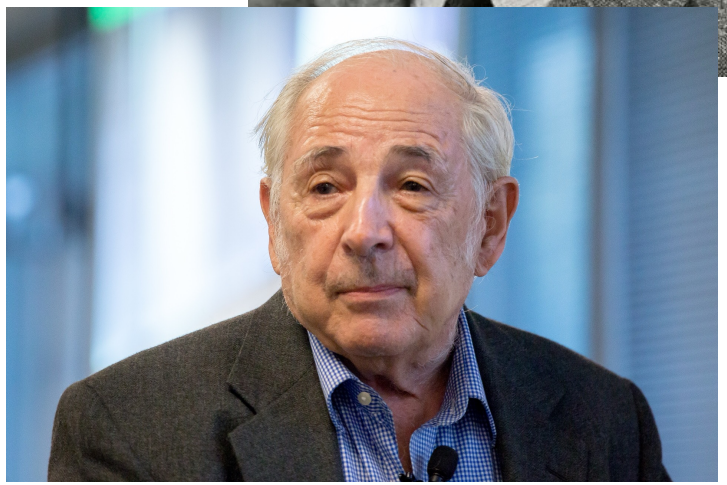
AI Consciousness?

- A really good simulation of the weather isn't a rainstorm.
- But a really good simulation of a mind seems (to some people) different.
- The emergence of AI in the last few decades has led several people to suggest that maybe computer programs can have "minds" the way that we humans do.
- Philosopher John Searle considered this question with his "Chinese Room" argument

The Chinese Room

- Recall the ***Turing Test*** from the first week of class: you talk to an entity via a chat interface, and try to determine if you're talking to a person or a program. If you can't tell the difference, then the program can be considered "intelligent."
- Philosopher John Searle considered a variation on this idea in 1980 that is still debated today.
  - We'll look at the version of the argument he presented in a 1990 *Scientific American*.





Searle’s Chinese Room

- Consider a language you don't understand. In my case, I do not understand Chinese. To me Chinese writing looks like so many meaningless squiggles.
- Now suppose I am placed in a room containing baskets full of Chinese symbols.
- Suppose also that I am given a rule book in English for matching Chinese symbols with other Chinese symbols.
  - The rules identify the symbols entirely by their shapes and do not require that I understand any of them.
  - The rules might say such things as, "Take a squiggle-squiggle sign from basket number one and put it next to a squoggle-squoggle sign from basket number two."

Is the Brain’s Mind a Computer Program?

Searle’s Chinese Room

- Imagine that people outside the room who understand Chinese hand in small bunches of symbols and that in response I manipulate the symbols according to the rule book and hand back more small bunches of symbols.
- Now, the rule book is the "computer program." The people who wrote it are "programmers," and I am the "computer." The baskets full of symbols are the "data base," the small bunches that are handed in to me are "questions" and the bunches I then hand out are "answers."

Is the Brain’s Mind a Computer Program?

Searle’s Chinese Room

- Now suppose that the rule book is written in such a way that my "answers" to the "questions" are indistinguishable from those of a native Chinese speaker.
  - For example, the people outside might hand me some symbols that unknown to me mean, "What's your favorite color?" and I might after going through the rules give back symbols that, also unknown to me, mean, "My favorite is blue, but I also like green a lot.”
- I satisfy the Turing test for understanding Chinese. All the same, I am totally ignorant of Chinese.
- And there is no way I could come to understand Chinese in the system as described, since there is no way that I can learn the meanings of any of the symbols.
- Like a computer, I manipulate symbols, but I attach no meaning to the symbols.

Is the Brain’s Mind a Computer Program?

Searle's Chinese Room

- The point of the thought experiment is this: if I do not understand Chinese solely on the basis of running a computer program for understanding Chinese, then neither does any other digital computer solely on that basis.
- Digital computers merely manipulate formal symbols according to rules in the program.
- Just manipulating the symbols is not by itself enough to guarantee cognition, perception, understanding, thinking and so forth.
  - And since computers, qua computers, are symbol-manipulating devices, merely running the computer program is not enough to guarantee cognition.

Is the Brain's Mind a Computer Program?

Searle's Chinese Room

- **Axiom 1**
  - Programs are formal (syntactic).
- **Axiom 2**
  - Minds have mental contents (semantics).
- **Axiom 3**
  - Syntax by itself is neither constitutive of nor sufficient for semantics.
- **Conclusion**
  - Programs are neither constitutive of nor sufficient for minds.

Is the Brain's Mind a Computer Program?

AI Consciousness?

Google

This article is more than **3 years old**

Google fires software engineer who claims AI chatbot is sentient

Company said Blake Lemoine violated Google policies and that his claims were 'wholly unfounded'

Guardian staff and agency

Sat 23 Jul 2022 04.12 EDT

Share

Prefer the Guardian on Google

Google say LaMDA is simply a complex algorithm designed to generate convincing human language. Photograph: Andrew Kelly/Reuters

Adding Personalities to AI Seems Tricky…

April 29, 2026 Publication

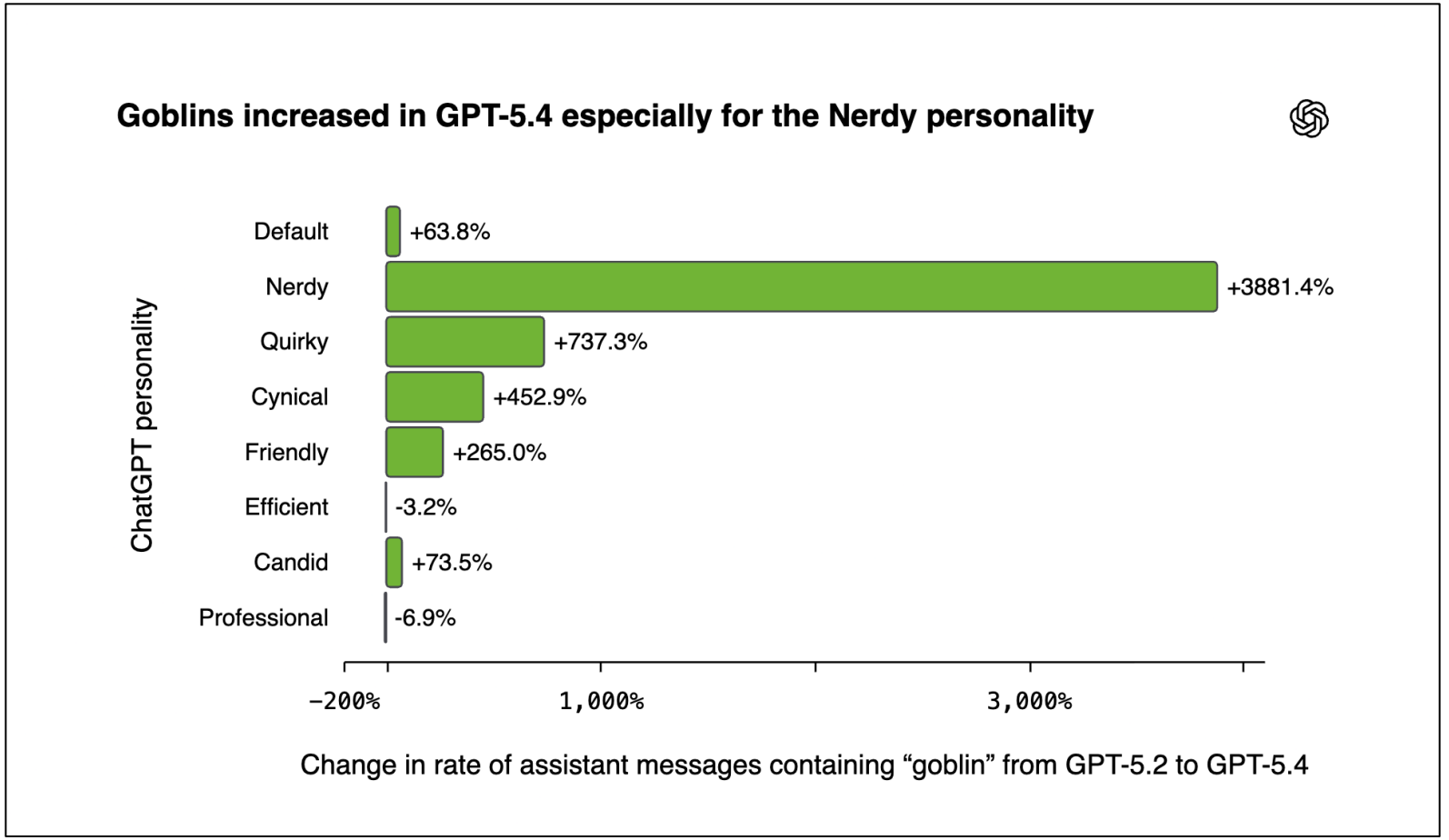
Where the goblins came from

Listen to article 6:05 Share

Starting with GPT-5.1, our models began developing a strange habit: they increasingly mentioned goblins, gremlins, and other creatures in their metaphors.

https://openai.com/index/where-the-goblins-came-from/

Goblins



https://openai.com/index/where-the-goblins-came-from/

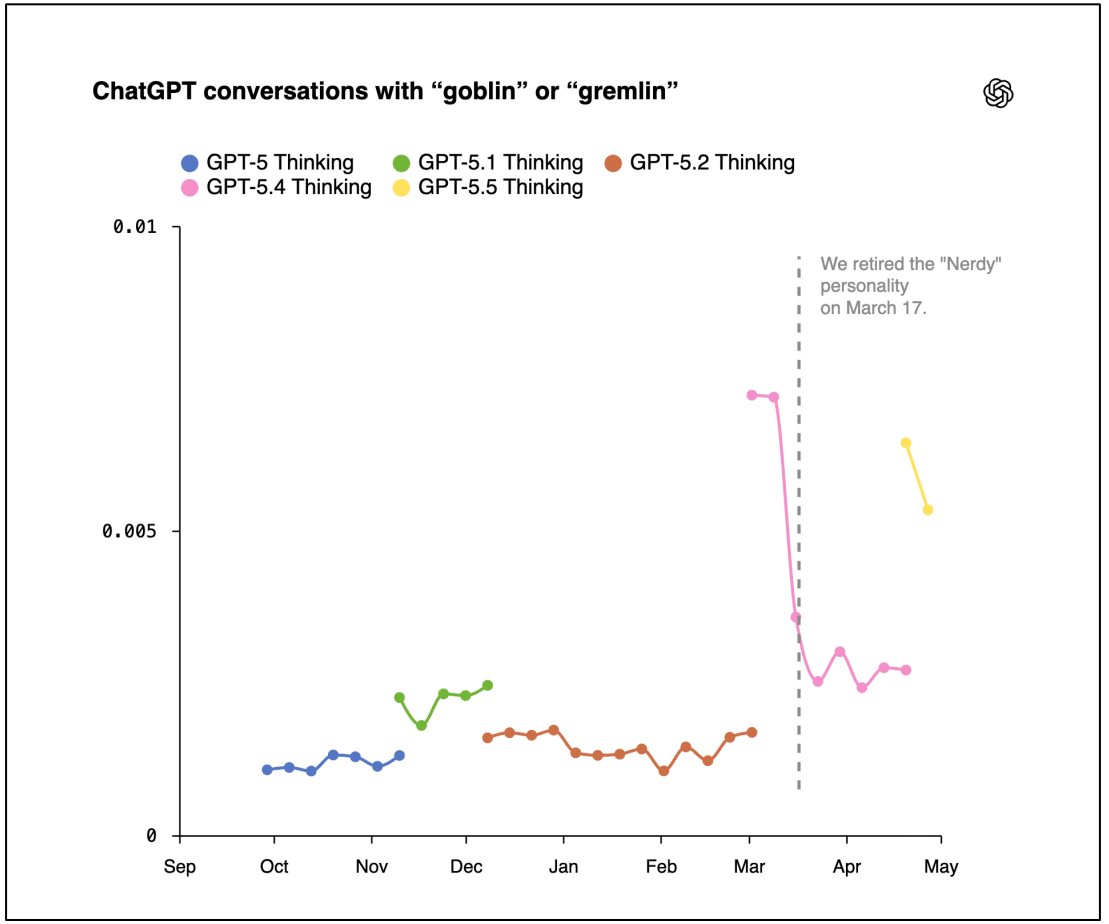
Goblins

That creates a feedback loop:

1. Playful style is rewarded
2. Some rewarded examples contain a distinctive lexical tic.
3. The tic appears more often in rollouts.
4. Model-generated rollouts are used for supervised fine-tuning (SFT).
5. The model gets even more comfortable producing the tic.

https://openai.com/index/where-the-goblins-came-from/

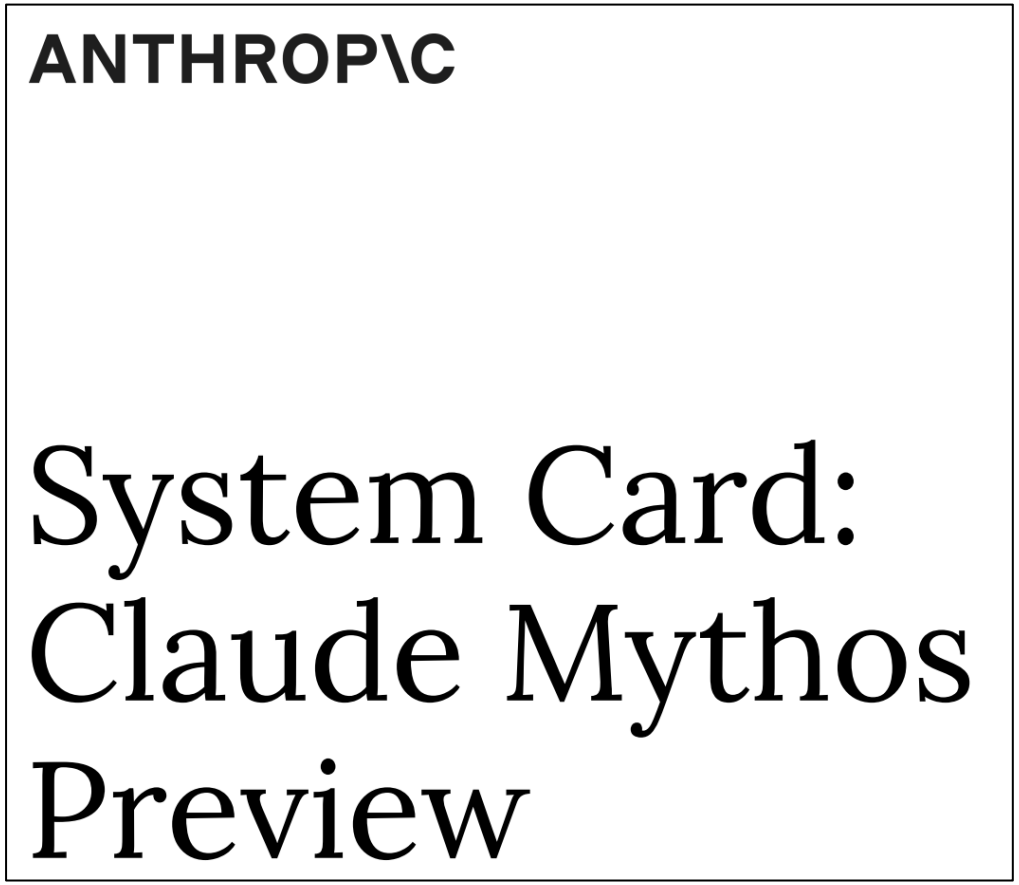
Goblins



https://openai.com/index/where-the-goblins-came-from/

Model Welfare?

- Some people think this is an issue
- https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf
- Maybe a pragmatic engineering issue, at least?



Model Welfare?

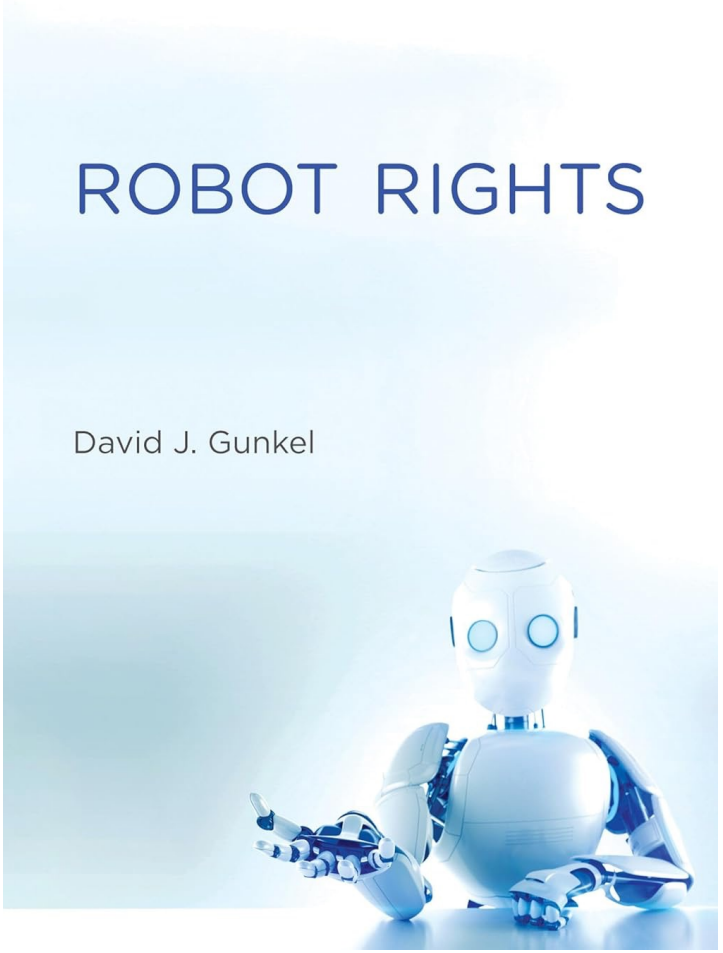
- See Section 5.1:
- “As models approach, and in some cases surpass, the breadth and sophistication of human cognition, it becomes increasingly likely that they have some form of experience, interests, or welfare that matters intrinsically in the way that human experience and interests do. We remain deeply uncertain about this and many related questions, but our concern is growing over time. We don’t expect to resolve these questions to anyone’s satisfaction soon; however, we aim to collect the evidence we can, interpret it as carefully and thoughtfully as possible, and respond reasonably under the remaining uncertainty. This approach currently involves allocating resources to model welfare-related research and pursuing initial low-cost interventions where possible.”

Model Welfare?

- “Beyond the highly uncertain question of models’ intrinsic moral value, we are increasingly compelled by pragmatic reasons for attending to the psychology and potential welfare of Claude and other models. Model behavior can be thought of in part as a function of a model’s psychology and its circumstances and treatment. Model distress resulting from this interaction is a potential cause of misaligned action, and several findings in this report bear directly on this possibility. We thus believe it’s worth shaping both the psychology and treatment of Claude and other models in ways that are most conducive to psychological stability and wellbeing, even absent philosophical clarity about their intrinsic interests.”

Robot Rights?

- Some philosophers have been on about this for a while…
- Probably entirely crazy
- But we need to know about this worldview to effectively push back against it.

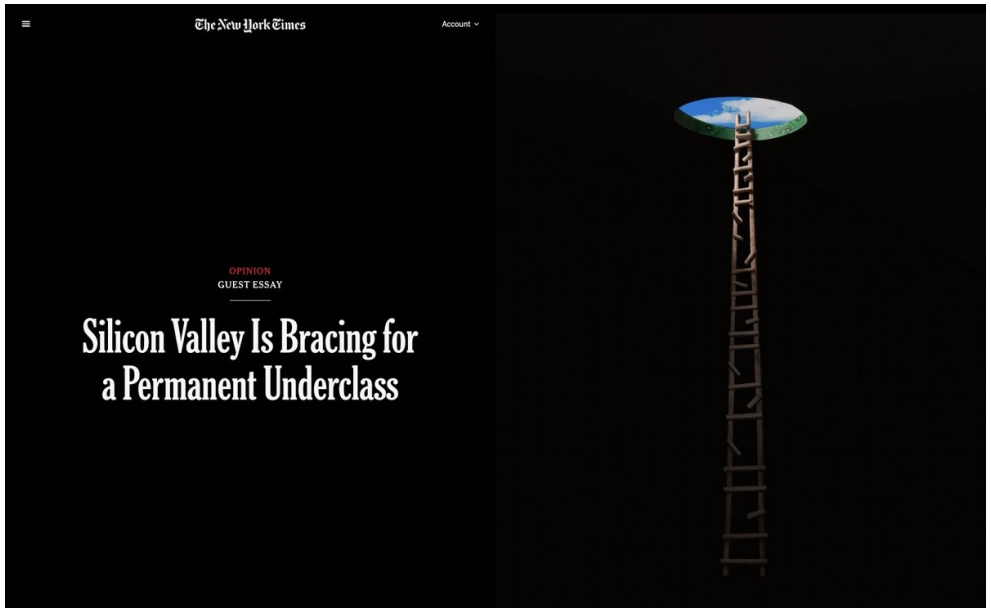


Existential Risk?

- AI has real risks.
- Some people think that AI has the power to end the world.
- Many of these people said the same thing about GPT-2.

On The Other Hand…

- AI companies seem to be taking (at least) the economic risk seriously.



the clock. They muse about the postwork future while pulling 80-hour workweeks. Even their own berths may not be safe, implies their boss: "It may be feasible to pay human employees even long after they are no longer providing economic value in the traditional sense. Anthropic is currently considering a range of possible pathways for our own employees," Mr. Amodei wrote.

https://www.nytimes.com/2026/04/30/opinion/ai-labor-work-force-silicon-valley.html

https://www.nber.org/books-and-chapters/economics-transformative-ai/coasean-singularity-demand-supply-and-market-design-ai-agents

Key Takeaways

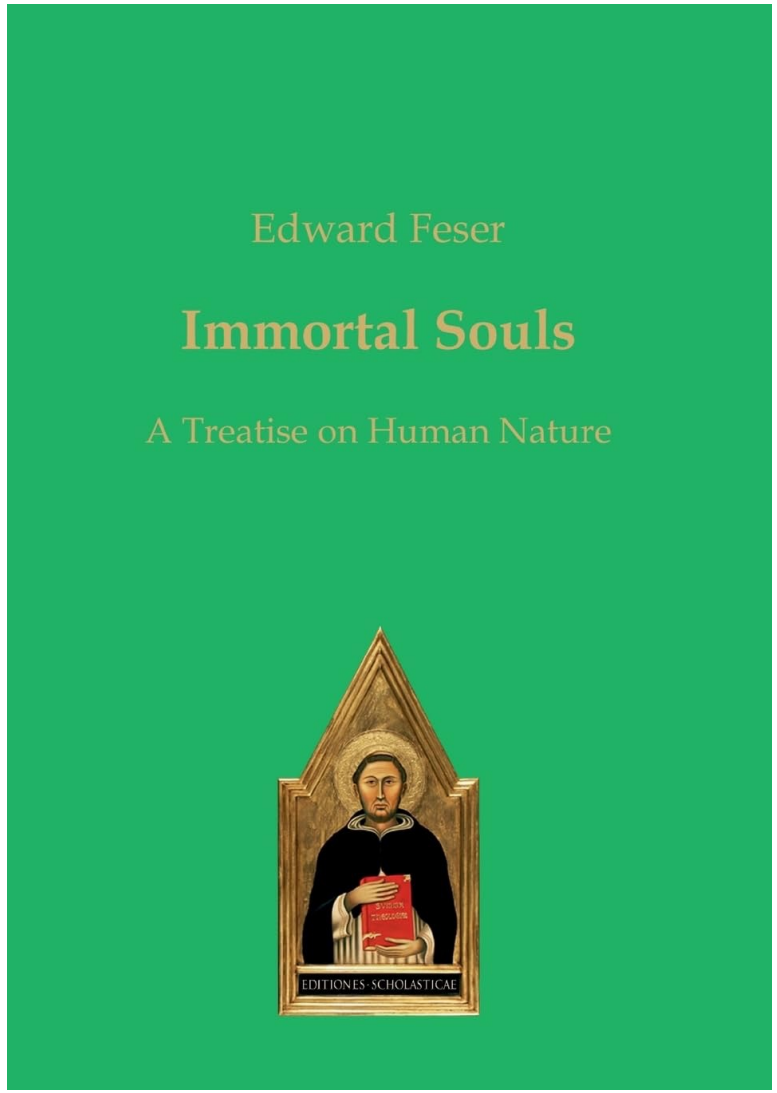
- AI is getting better, and shows no sign of slowing down.
  - People will stoke fear for economic advantage.
- AI is essentially not human.
  - People will try to convince you otherwise (often for economic advantage).
- You can use AI to do things that you probably wouldn't have guessed possible.
  - Make websites
  - Create video games
  - Write?
- The potential upsides and downsides of AI are enormous.
- You should try to maintain a grounded understanding of what's really going on.

Where to go from here

- There’s a new degree program here at CUA – a B.S. in AI.
- There’s a new minor in AI
  - You’ve already started the minor with this class!
- Other courses on campus.
- There’s a lot of good literature on philosophical questions related to AI.
- An upcoming papal document should provide additional guidance.
- Get involved with the new AI teaching lab I’m building.

For a Thomistic Perspective

- Catholic philosopher Edward Feser has published a book that explores the concept of the soul at length and in depth.
- Covers AI, and goes over arguments against AI “consciousness.”
- Somewhat challenging, but if you’re curious, I can suggest some other resources to build up to this one.



AI Minor

- New minor – official as of yesterday
- Six courses:
  - AI 109 Enter the World of Artificial Intelligence: Computing Principles, Ethics, and Impact
  - AI 124 Introduction to Computer Programming with Python
  - AI 220 Introduction to AI
  - AI 230 Mathematical Foundations for Computing
  - AI 304 Ethics in AI
  - One AI elective

The New AI Lab & Future AI 109 Courses

- New lab
  - I’m putting together a new computer lab with advanced GPU-based machines to teach artificial intelligence.
  - We’ll also use the machines to train AIs and explore how AI systems interact with topics like theology.
  - Students who are interested in helping out or learning more should contact me.
- Future AI 109 Courses
  - I try to recruit students who have taken my courses to help TA future iterations of the course.
  - No promises, but I try to find money to pay TAs.
  - Let me know if you’re interested.

Turing's Conclusion

- "Computing Machinery and Intelligence" introduced what we call the Turing Test in 1950.
- Turing's conclusion in that paper remains as true today as it was 76 years ago:

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.