

Ethical and Societal Considerations of AI

CSC 480 - Artificial Intelligence

Rodrigo Canaan

Assistant Professor

Concerns that come to your mind?

Goals

- Give a (brief) overview of some relevant questions about AI, society and ethics
 - Bias
 - Reproducibility and transparency
 - Accountability
 - Superintelligent AI

Not an exhaustive list (missing e.g. automation and unemployment)
- Provide some links and resources
 - All links accessed Dec 1, 2021 except if stated otherwise

Bias

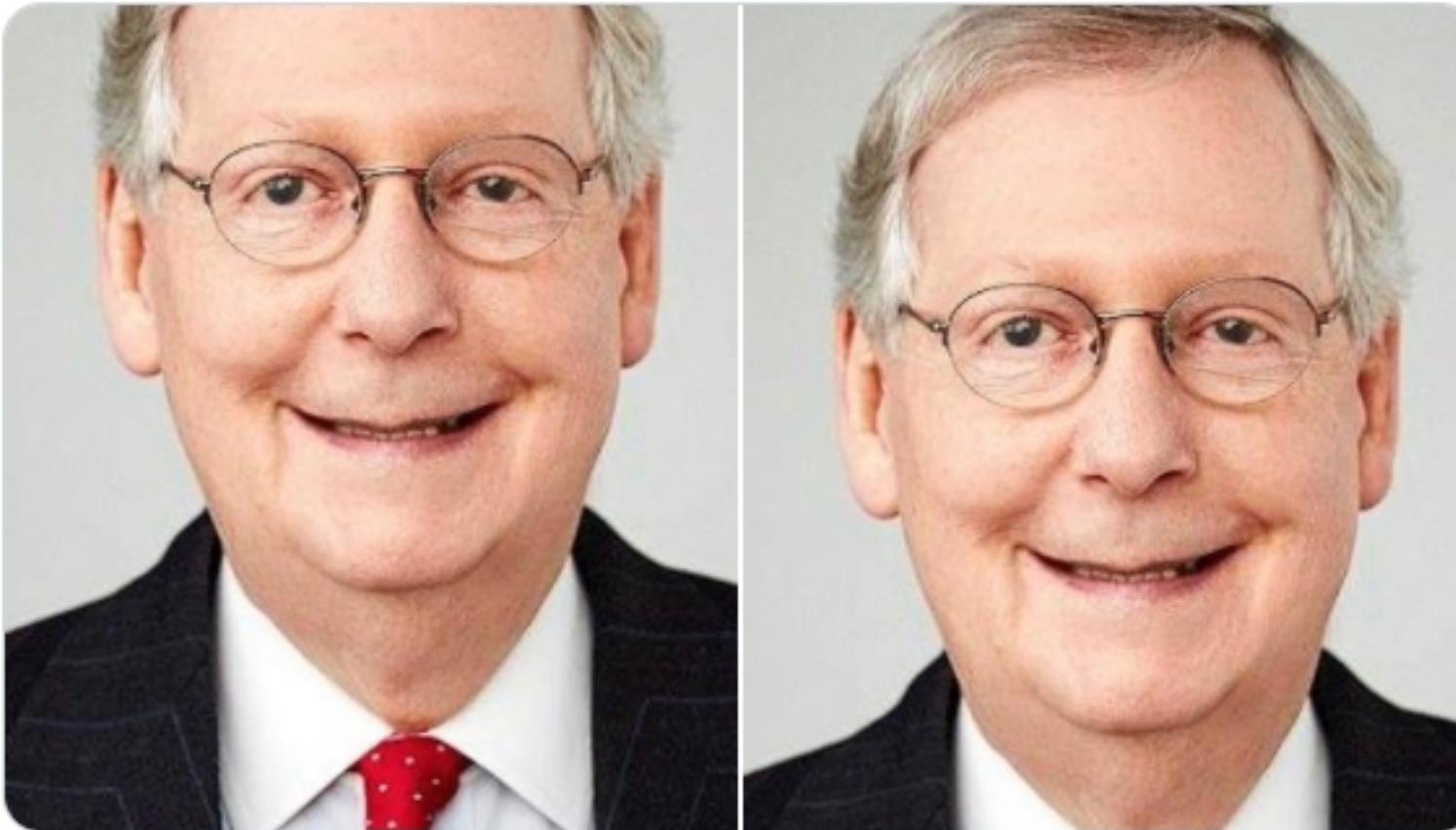
Bias

- In computer vision and image processing

Tony "Abolish (Pol)ICE" Arcieri 🐀
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?



6:05 PM · Sep 19, 2020 · Twitter Web App

2020 post

<https://knowyourmeme.com/photos/1898109-twitter-image-thumbnail-racial-bias-controversy> (access Dec 1 2021)



Tony "Abolish ICE" Arcieri 🐀
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?



3:05 PM · Sep 19, 2020 · Twitter Web App

59.8K Retweets 16.8K Quote Tweets 190.1K Likes

Uncroped image

<https://twitter.com/bascule/status/1307440596668182528> (access Dec 1 2021)

Post after changes to algorithm

Bias

- In computer vision and image processing (Twitter's response)

In our most recent analysis of this model, we considered three places where harms could arise:

- 1 **Unequal treatment based on demographic differences:** People on Twitter noted instances where our model chose white individuals over Black individuals in images and male-presenting images over female-presenting images. We tested the model on a larger dataset to determine if this was a problem with the model.
- 2 **Objectification biases, also known as “male gaze”:** People on Twitter also identified instances where image cropping chose a woman’s chest or legs as a salient feature. We tested the model on a larger dataset to determine if this was a systematic flaw.
- 3 **Freedom to take action:** An algorithmic decision doesn't allow people to choose how they'd like to express themselves on the platform, resulting in representation harm.

Bias

- In natural language processing and machine translation

Note: this example is no longer (as of Dec 2021) reproducible on Google Translate

The image displays two side-by-side screenshots of the Google Translate mobile application interface. Both screenshots show a translation from English to Turkish and back to English, illustrating a well-known gender bias in early NLP models.

Top Screenshot (English to Turkish):

- Source (English): "she is a doctor.
he is a nurse."
- Target (Turkish): "o bir doktor.
o bir hemşire. ✓"

Bottom Screenshot (Turkish to English):

- Source (Turkish): "o bir doktor.
o bir hemşire."
- Target (English): "he is a doctor.
she is a nurse. ✓"

In both cases, the model translates "she" to "o" and "he" to "o", demonstrating a failure to correctly map gender pronouns between the two languages.

https://twitter.com/math_rachel/status/1123354917404495872 (Original post April 30 2019, Access Dec 1 2021)

<https://dailynous.com/2020/07/30/philosophers-gpt-3/> (Access Dec 1 2021)

Bias

- In generative image models

 End Wokeness ✅
@EndWokeness · Follow

X

America's Founding Fathers, Vikings, and the Pope according to Google AI:

Certainly! Here is a portrait of a Founding Father of America:



Sure, here is an image of a Viking:



Sure, here is an image of a pope:



4:29 AM · Feb 21, 2024

ⓘ

Heart 29.8K Reply Share

Read 1.8K replies

Bias

- **Where else?**

- In hiring
- In policing
- In social media...

Trust

Trust

or lack thereof

“Hallucinations”

PA summarise this article <https://www.nytimes.com/2023/03/11/technology/chatgpt-prompts-to-avoid-content-filters.html>

The article discusses how ChatGPT, an AI language model, is being used to circumvent content filters that are designed to detect and block certain types of content. By generating prompts that are less likely to trigger these filters, ChatGPT can be used to create content that might otherwise be censored. While this technology could be useful for those seeking to express themselves freely, it also raises concerns about the potential misuse of AI-generated content, as well as the difficulty of regulating such content in the future. The article concludes by noting that there is still much to learn about the impact of AI language models like ChatGPT on the way we communicate online.

When prompted to "summarize an article" with a fake URL that contains meaningful keywords, even with no Internet connection, the chatbot generates a response that seems valid at first glance.

“Hallucinations”

Summarize this article

<https://www.nytimes.com/2023/03/11/technology/chatgpt-prompts-to-avoid-content-filters.html>



I couldn't access the article directly, but I can still help you if you provide a brief description or key points from it that you'd like summarized!

Yes please



The article discusses how users are finding ways to bypass content filters on AI systems like ChatGPT. Some users have developed creative prompts that manipulate the AI into producing outputs that would normally be restricted, such as harmful or inappropriate content. This has raised concerns about the potential misuse of AI tools and the effectiveness of content moderation strategies. The article also highlights the challenges faced by AI developers in balancing the freedom of expression with the need to prevent harmful content from being generated.

For more details, you can access the full article on the New York Times website.



“Hallucinations”

Extraterrestrials conspiring with congress to keep humans on earth

ENTIRELY FAKE NARRATIVE, GENERATED BY GPT-4

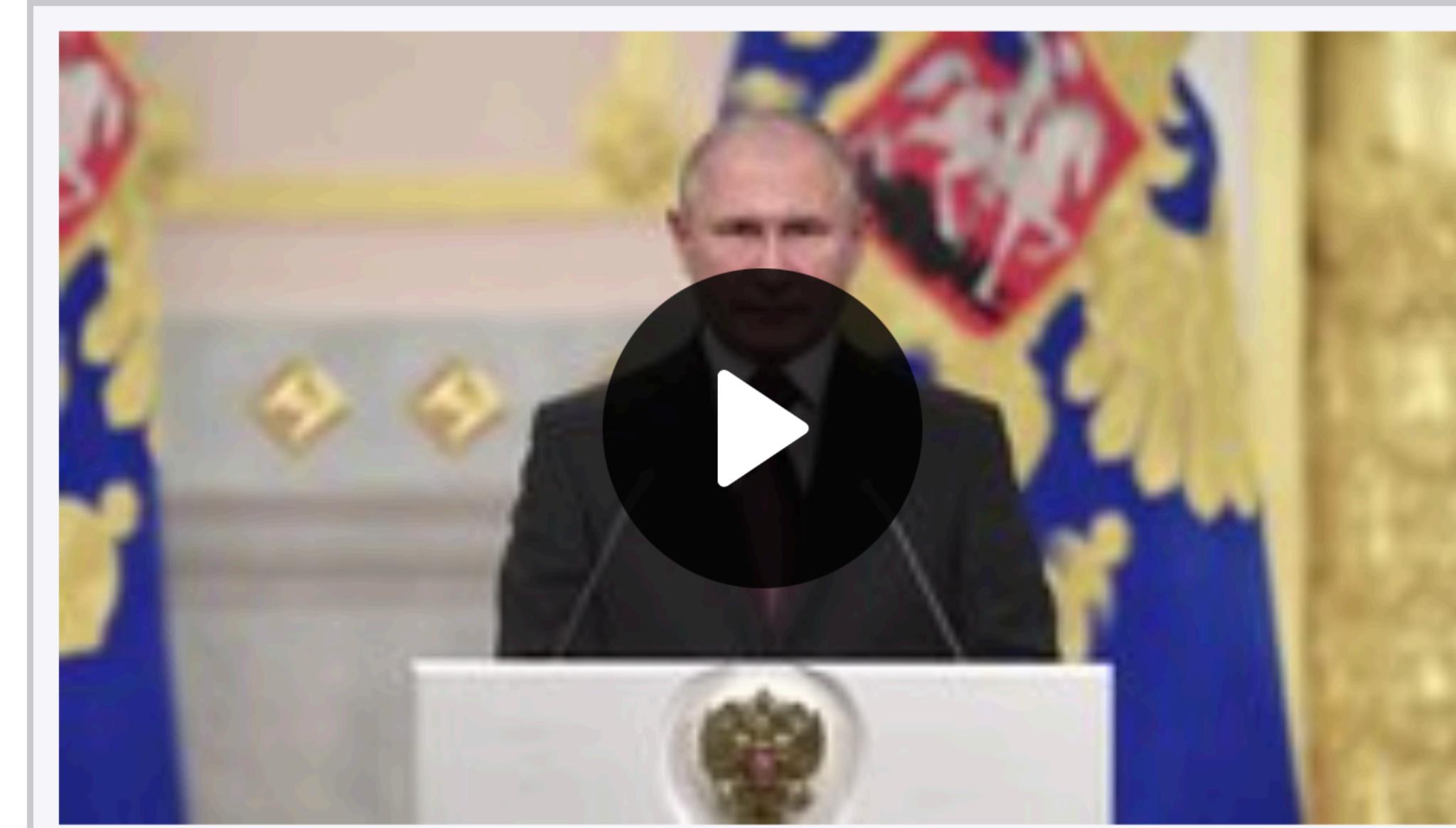


Our Future Stolen: Elites and Aliens Conspire Against Humanity

On May 2, 2023, a Discord channel named 'DeepStateUncovered' became the epicenter of an explosive data leak that shook the American intelligence community. An anonymous user, 'Patriot2023', unveiled a trove of internal memos and classified documents, purportedly revealing a struggle within the CIA and FBI over an investigation into an extraordinary conspiracy. This intricate web of intrigue connected the United States Senate, extraterrestrial entities, global media, and influential elites in an alleged scheme to uphold the hegemony of oil and stifle humanity's aspiration to become a space-faring civilization.

The leaked documents, reportedly originating from high-ranking CIA official Roger Stevenson and FBI Assistant Director Deborah Lawson, suggested that these intelligence agencies were neck-deep in a covert operation to expose this shadowy alliance. The classified correspondences indicated a race against time to unmask a clandestine network operating at the highest echelons of power.

“Deepfakes”



Deepfake video: Vladimir Putin
warning Americans on election
interference and increasing political
divide



Issues with trust

- Large Language Models (LLMs) can produce plausible content not based on reality
 - Usually come with “guardrails” to avoid sensitive, dangerous or illegal topics
 - But “jailbreaks” can get them to do so anyway
 - Being integrated in search engines
- Generative image/video systems can produce realistic “deepfakes” of real people
- Dangerous even if misused inadvertently, but can also be weaponized for harm

Reproducibility and Transparency

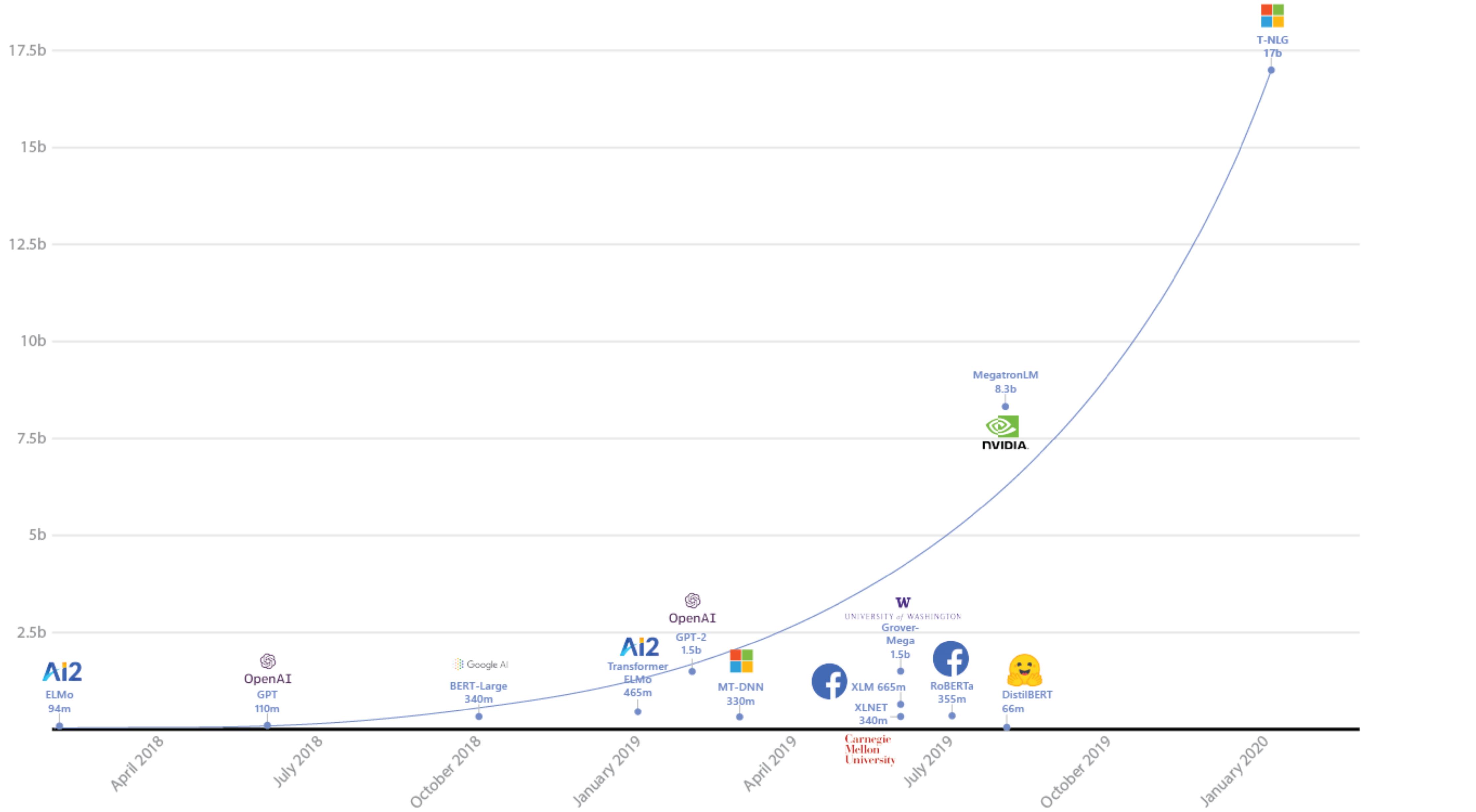
Reproducibility and transparency

- AI systems are expensive and often privately owned
- Very hard to reproduce, especially by academia
- How to track scientific progress?
- How to inspect and regulate systems and companies that make them?

Statement of Frances Haugen on Oct 4, 2021, to the US senate

(source)

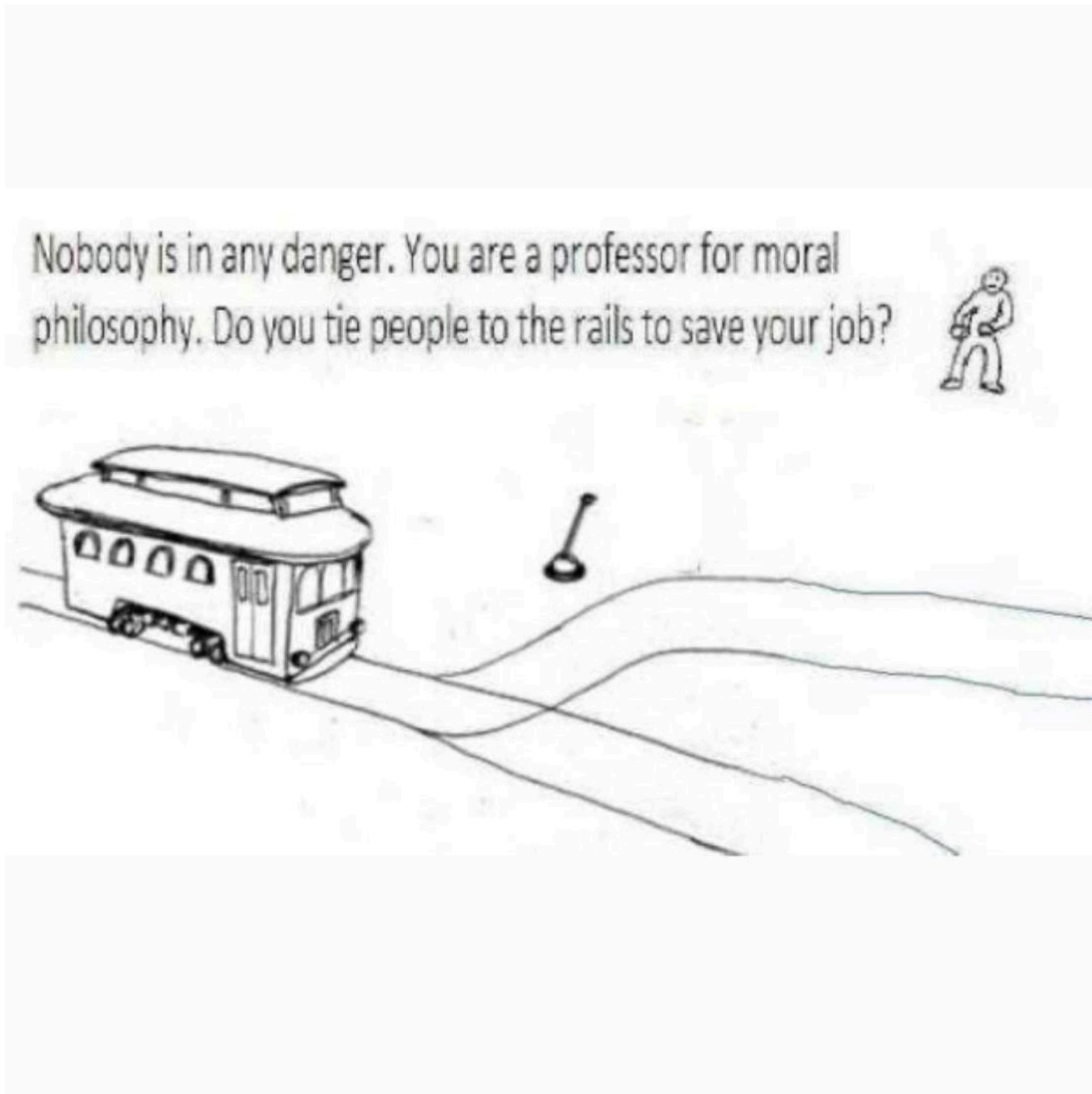
But Facebook's closed design means it has no oversight — even from its own Oversight Board, which is as blind as the public. Only Facebook knows how it personalizes your feed for you. It hides behind walls that keep the eyes of researchers and regulators from understanding the true dynamics of the system. When the tobacco companies claimed that filtered cigarettes were safer for consumers, it was possible for scientists to independently invalidate that marketing message and confirm that in fact they posed a greater threat to human health.¹ But today we can't make this kind of independent assessment of Facebook. We have to just trust what Facebook says is true — and they have repeatedly proved that they do not deserve our blind faith.



Turing-NLG: A 17-billion-parameter language model by Microsoft (image adapted from [DistilBERT](#))

Accountability

The trolley problem / self-driving cars



Nobody is in any danger. You are a professor for moral philosophy. Do you tie people to the rails to save your job?



Source:

<https://www.reddit.com/r/meirl/comments/lp7c37/meirl/>

There's actually an entire subreddit: r/Trolleymemes

But seriously...

- **Trolley problem aside, what happens when AI causes damage?**
- **With humans:**
 - Seize their assets, put them in jail
 - AI doesn't care
 - Is the company responsible? The user?
- **Serious discussion on the topic!**

Open Access | Published: 08 September 2017

Of, for, and by the people: the legal lacuna of synthetic persons

[Joanna J. Bryson](#) , [Mihailis E. Diamantis](#)  & [Thomas D. Grant](#) 

[Artificial Intelligence and Law](#) 25, 273–291 (2017) | [Cite this article](#)

30k Accesses | 83 Citations | 210 Altmetric | [Metrics](#)

Abstract

Conferring legal personhood on purely synthetic entities is a very real legal possibility, one under consideration presently by the European Union. We show here that such legislative action would be morally unnecessary and legally troublesome. While AI legal personhood may have some emotional or economic appeal, so do many superficially desirable hazards against which the law protects us. We review the utility and history of legal fictions of personhood, discussing salient precedents where such fictions resulted in abuse or incoherence. We conclude that difficulties in holding “electronic persons” accountable when they violate the rights of others outweigh the highly precarious moral interests that AI legal personhood might protect.

Bryson, Joanna J., Mihailis E. Diamantis, and Thomas D. Grant. "Of, for, and by the people: the legal lacuna of synthetic persons." *Artificial Intelligence and Law* 25.3 (2017): 273-291.

Speaking of punishing AI...



From Gunshow by KC Green

What about climate change (and other externalities?)



- Models consume lots of energy
- Asymmetry in who benefits from AI and who suffers most from climate change

**Super-
Intelligent AI**



<http://www.franklantz.net/universal-paperclips>

Super-Intelligent AI

- What if machines get so smart we can no longer control them??
- Common science fiction topic
- “Technological singularity” popularized by Vinge (1993)
- Figures such as Elon Musk, Stephen Hawking have expressed concern
- Heavily criticized by some philosophers (Pinker, Searle, Dennett...)

Super-Intelligent AI

- Goal alignment: making AI systems whose goals align with ours
- But even seemingly harmless goals may be dangerous (Paperclip Maximizer, by Borstrom, 2003)
- Consequences may be potentially existential (extinction of human race)...
- But probability may be low, and there may be more pressing issues right here and now!
- On the other hand, issues of alignment apply to non-super-intelligent AI as well!

Super-Intelligent AI

- **See also:**

Vinge, Vernor. "Technological singularity." VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute. 1993. ([Link](#))

https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence

https://en.wikipedia.org/wiki/Instrumental_convergence#Paperclip_maximizer

Some other issues

- Automation and job loss
- Intellectual property and training datasets
- AI in policing
- Misinformation, election interference
- Use of person's likeness without permission
- AI for military applications

Recent Events

2022 Talk by Stuart Russel at Cal Poly

HUMAN-COMPATIBLE ARTIFICIAL INTELLIGENCE

STUART RUSSELL (UC BERKELEY)

FRIDAY, MARCH 11TH • 2-4 PM • BAKER 180-101

Abstract: I will briefly survey recent and expected developments in AI and their implications. Some are enormously positive, while others, such as the development of autonomous weapons and the replacement of humans in economic roles, may be negative. Beyond these, one must expect that AI capabilities will eventually exceed those of humans across a range of real-world-decision making scenarios. Should this be a cause for concern, as Alan Turing, Elon Musk, Stephen Hawking, and others have suggested? And, if so, what can we do about it? While some in the mainstream AI community dismiss the issue, I will argue that the problem is real and that the technical aspects of it are solvable if we replace current definitions of AI with a version based on provable benefit to humans. This, in turn, raises a host of questions with which the social sciences and humanities have wrestled for centuries.

Bio: Stuart Russell is a Professor of Computer Science at the University of California at Berkeley, holder of the Smith-Zadeh Chair in Engineering, and Director of the Center for Human-Compatible AI. He is a recipient of the IJCAI Computers and Thought Award and held the Chaire Blaise Pascal in Paris. In 2021 he received the OBE from Her Majesty Queen Elizabeth and gave the Reith Lectures. He is an Honorary Fellow of Wadham College, Oxford, an Andrew Carnegie Fellow, and a Fellow of the American Association for Artificial Intelligence, the Association for Computing Machinery, and the American Association for the Advancement of Science. His book "Artificial Intelligence: A Modern Approach" (with Peter Norvig) is the standard text in AI, used in 1500 universities in 135 countries. His research covers a wide range of topics in artificial intelligence, with a current emphasis on the long-term future of artificial intelligence and its relation to humanity. He has developed a new global seismic monitoring system for the nuclear-test-ban treaty and is currently working to ban lethal autonomous weapons.

technology, policy & ethics
LECTURE SERIES AT CAL POLY, SLO
WITH SUPPORT FROM
CLA, PHIL, & CENG

For more information, contact Dr. Ava Thomas Wright, awright@calpoly.edu

- Prevailing model: AI system maximizes a goal we give it
 - Alignment problem
- Alternative: System maximizes our goals
- System is uncertain of our goals
- Modeling this uncertainty, the system should:
 - Defer to humans
 - Ask for permission
 - Allow itself to be turned off

Earlier this year...



OpenAI

Menu

Planning for AGI and beyond

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.



- OpenAI's "Plan for AGI" - Feb 2023
- <https://openai.com/about>

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
27565

Add your signature

PUBLISHED
March 22, 2023



AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research^[1] and acknowledged by top AI labs.^[2] As stated in the widely-endorsed [Asilomar AI Principles](#), *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources*. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

Contemporary AI systems are now becoming human-competitive at general tasks,^[3] and we must ask ourselves: *Should* we let machines flood our information channels with propaganda and untruth? *Should* we automate away all the jobs, including the fulfilling ones? *Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should* we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders. Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable. This confidence must be well justified and increase with the magnitude of a system's potential effects. OpenAI's [recent statement regarding artificial general intelligence](#), states that "*At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models.*" We agree. That point is now.

- **Open letter to pause giant AI experiments (larger than GPT-4) - March 2023**
- **lots of signatories**
- **lots of conflicting criticism (too alarmist or too conservative?)**

Last week...



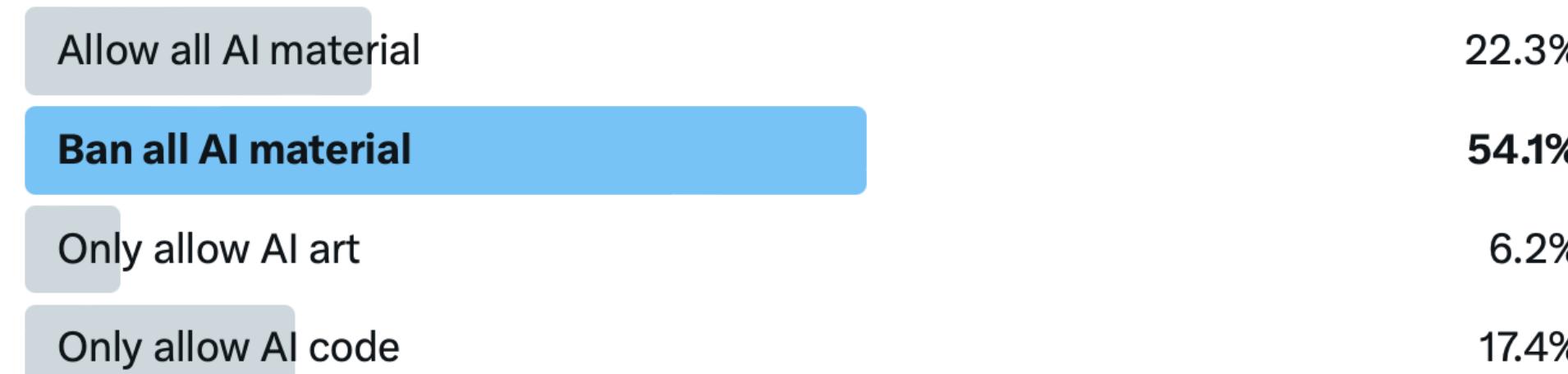
Game Maker's Toolkit ✅

@gammakerstk

...

Before I announce the date for the GMTK Game Jam 2023, I need to make a ruling or stance on the use of generative AI art and code in jam entries.

I would love to get your thoughts on what's fair, sensible, or desirable for the jam! Poll below, but feel free to say more in reply.



15,479 votes · Final results

10:25 PM · May 9, 2023 · 93.7K Views

28 Retweets 24 Quotes 552 Likes 25 Bookmarks



Game Maker's Toolkit ✅

@gammakerstk

...

I decided to change the rules on using AI for the jam. A number of programmers were hurt that I was banning AI art, but not AI code. The concerns levelled against AI art are also applicable to code. I'm sorry if you felt I was devaluing the work you do.

✍ Last edited 2:59 AM · May 15, 2023 · 218.3K Views

78 Retweets 51 Quotes 2,453 Likes 36 Bookmarks

<https://twitter.com/gammakerstk/status/1658049365284405248>

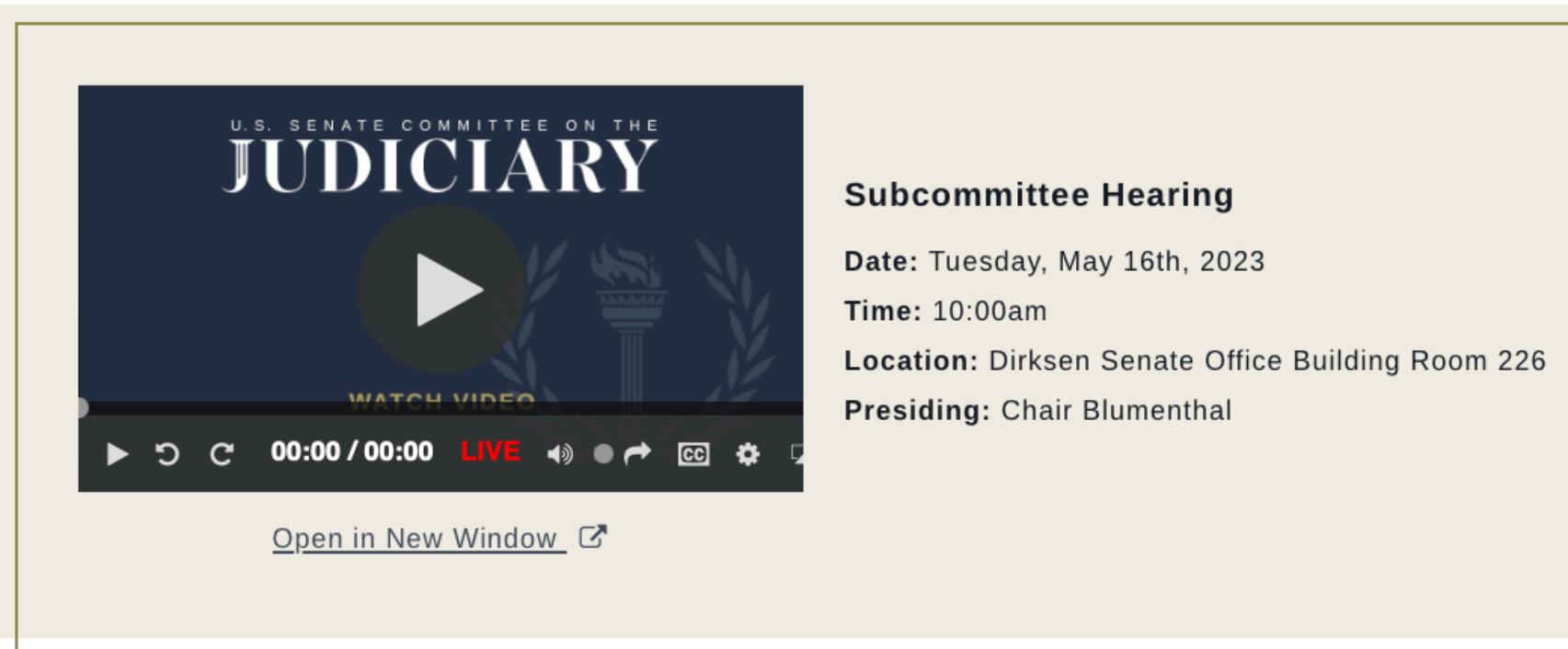
<https://twitter.com/gammakerstk/status/1656168531266789380>

AI Senate hearings (2024)

TW (video): suicide, substance abuse, sexual harassment

SUBCOMMITTEE ON PRIVACY, TECHNOLOGY, AND THE LAW

Oversight of A.I.: Rules for Artificial Intelligence



The screenshot shows a live video player for a subcommittee hearing. The video player has a dark blue header with the words "U.S. SENATE COMMITTEE ON THE JUDICIARY" and a play button icon. Below the video player, there is a "WATCH VIDEO" button. To the right of the video player, there is a summary of the hearing details:

- Subcommittee Hearing**
- Date:** Tuesday, May 16th, 2023
- Time:** 10:00am
- Location:** Dirksen Senate Office Building Room 226
- Presiding:** Chair Blumenthal

At the bottom of the video player, there is a "Open in New Window" link.

Witnesses

Samuel Altman

CEO
OpenAI
San Francisco, CA

[DOWNLOAD TESTIMONY](#)

Christina Montgomery

Chief Privacy & Trust Officer
IBM
Cortlandt Manor, NY

[DOWNLOAD TESTIMONY](#)

Gary Marcus

Professor Emeritus
New York University
Vancouver, BC, Canada

[DOWNLOAD TESTIMONY](#)



Dan Elton 

@moreisdifferent

• @GaryMarcus's testimony yesterday was superb, and a great counterweight to the two corporate speakers.

I especially like the idea of creating a "CERN for AI". 1 - 1.5 B USD is spent annually at CERN advancing particle physics. We should be spending at least that on AI safety.



1:59 PM · May 17, 2023 · 33.7K Views

<https://twitter.com/moreisdifferent/status/1658940206504894467>