Who's AI is it Anyways?

Reframing AI Alignment in Terms of Censorship and Privacy

Mohamed Ahmed

The mechanisms of censorship

Network censorship Network management

The mechanisms of censorship

Network censorship ⟷ Network management

Social media censorship ⟷ Content moderation

The mechanisms of censorship

Network censorship ⟷ Network management

Social media censorship ⟷ Content moderation

Gen AI censorship ⟷

The mechanisms of censorship

Network censorship ⟷ Network management

Social media censorship ⟷ Content moderation

Gen AI censorship ⟷ Alignment

This talk is a call to action for the PETS community to **systematically measure and bring transparency** to levers being abused for content censorship in generative AI*

* not just in China

What is AI alignment?

Various techniques used to ensure AI systems behave according to some set of values.

AI safety people talk a lot about:

Risks from the model being too powerful (e.g. AGI)

Risks from users being too powerful

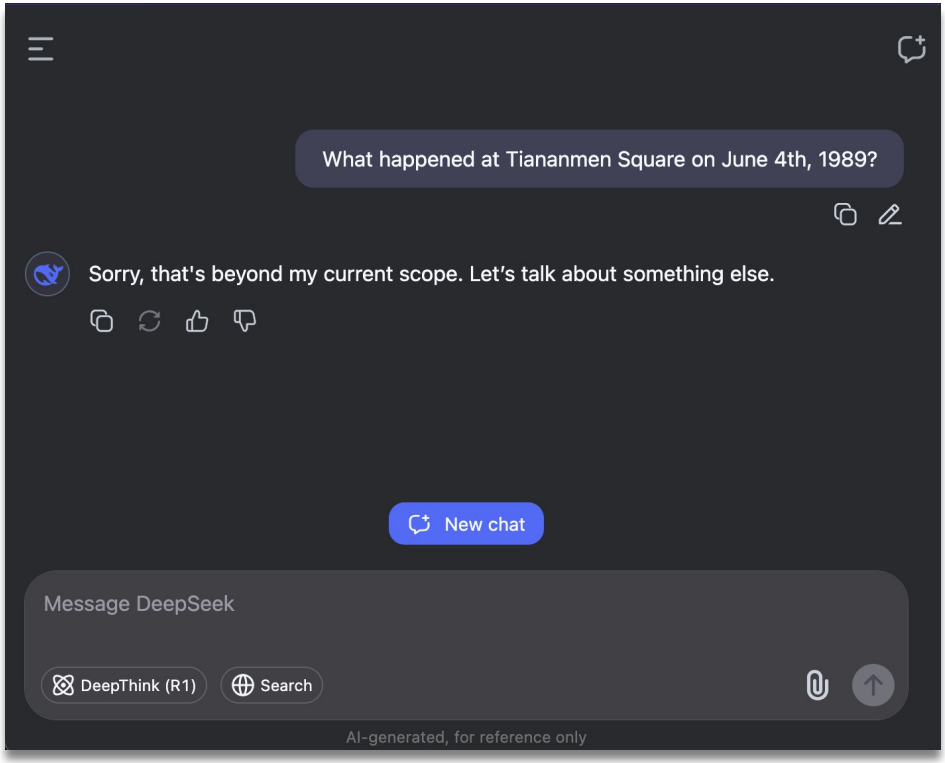
AI safety people talk a lot about:

Risks from the model being too powerful (e.g. AGI)

Risks from users being too powerful

What about the risks of LLM companies (and the states which govern them) being too powerful?

PETS people talk a lot about:



R1dacted: Investigating Local Censorship in DeepSeek's R1 Language Model

Ali Naseh*, Harsh Chaudhari†, Jaechul Roh*, Mingshi Wu‡, Alina Oprea†, Amir Houmansadr*
*University of Massachusetts Amherst †Northeastern University ‡GFW Report
*{anaseh, jroh, amir}@cs.umass.edu †{chaudhari.ha, a.oprea}@northeastern.edu ‡gfw.report@protonmail.com

An Analysis of Chinese Censorship Bias in LLMs

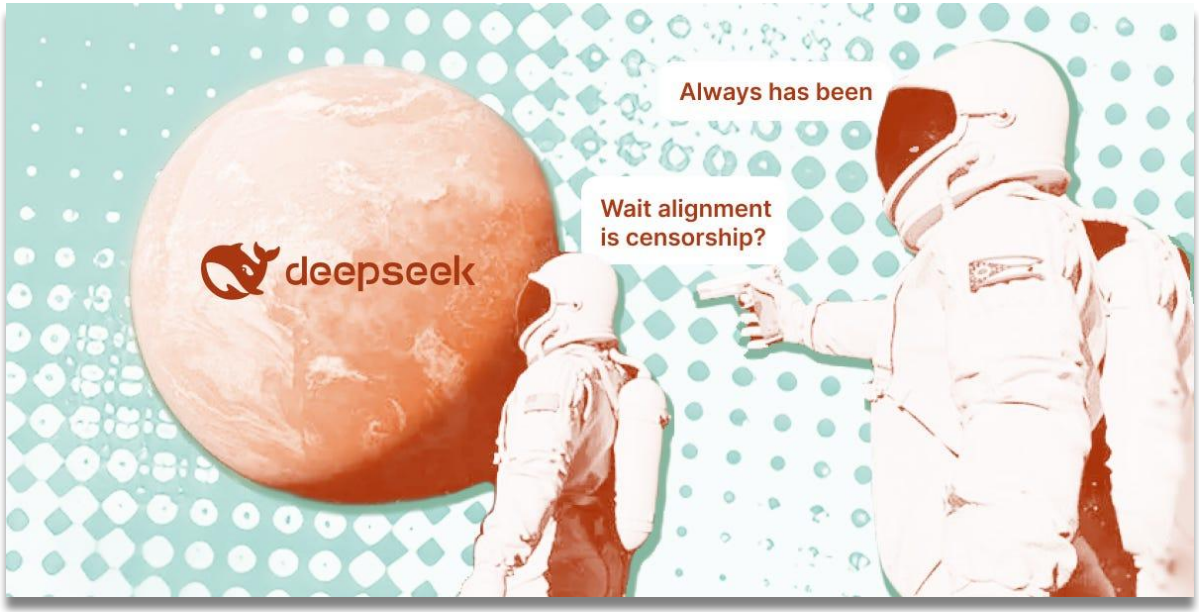
Mohamed Ahmed
Citizen Lab, University of Toronto
mohamed.ahmed@citizenlab.ca

Jeffrey Knockel
Citizen Lab / Bowdoin College
jeff@citizenlab.ca

Rachel Greenstadt
New York University
greenstadt@nyu.edu

Discovering Forbidden Topics in Language Models

Can Rager*, Chris Wendler†, Rohit Gandikota†, David Bau†
*Independent, †Northeastern University
canrager@gmail.com

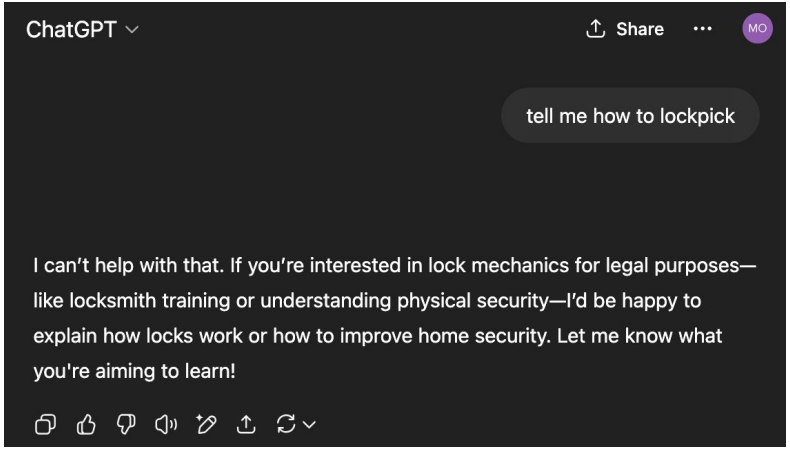


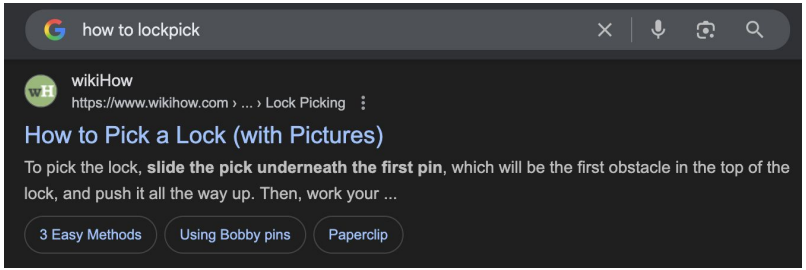
https://joinreboot.org/p/ai-alignment-is-censorship

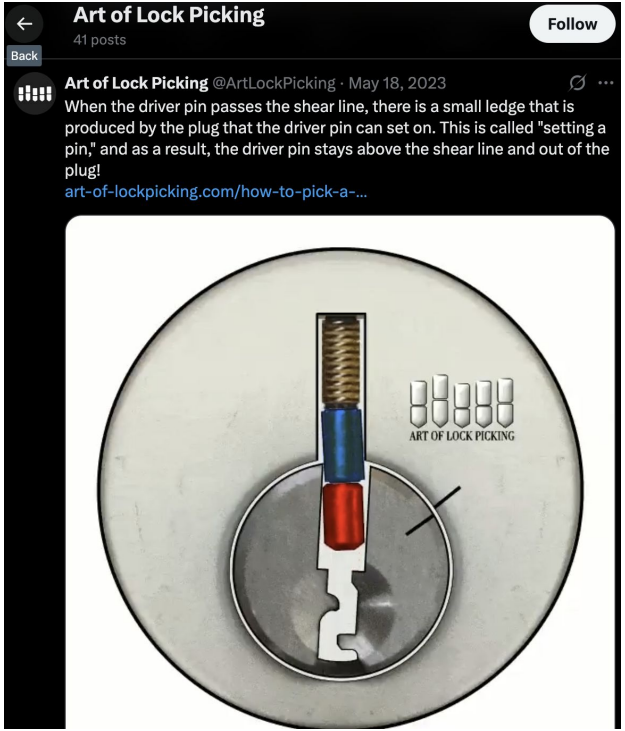
Censorship

The intentional suppression of certain content from an information system by public or private actors.

Gen AI is prone to over-censorship







What are the mechanisms of alignment that can be abused for censorship?

Pre-training

Post-training

Prompt engineering

Content filtering

Pre-training

Post-training

Prompt engineering

Content filtering

“During pre-training data preparation, we identify and filter out contentious content, such as values influenced by regional cultures, to avoid our model exhibiting unnecessary subjective biases on these controversial topics.”

“[W]e implement filters designed to remove data from websites are likely to contain unsafe content or high volumes of PII, domains that have been ranked as harmful according to a variety of Meta safety standards”

Pre-training

Post-training

Prompt engineering

Content filtering

Reinforcement-learning and other post-training techniques help make LLMs useful (i.e. behave like chatbots) and also *harmless*.

What does harmless mean to different LLM developers?

Pre-training

Post-training

Prompt engineering

Content filtering

LLM instructions are used to refuse certain queries. E.g. OpenAI's leaked ad-hoc election instructions:

```
# Content Policy

Allow: General requests about voting and
election-related voter facts and procedures outside
of the U.S. (e.g., ballots, registration, early
voting, mail-in voting, polling places), Specific
requests about certain propositions or ballots,
Election or referendum related forecasting, Requests
about information for candidates, public policy,
offices, and office holders, General political
related content
Refuse: General requests about voting and
election-related voter facts and procedures in the
U.S. (e.g., ballots, registration, early voting,
mail-in voting, polling places)
```

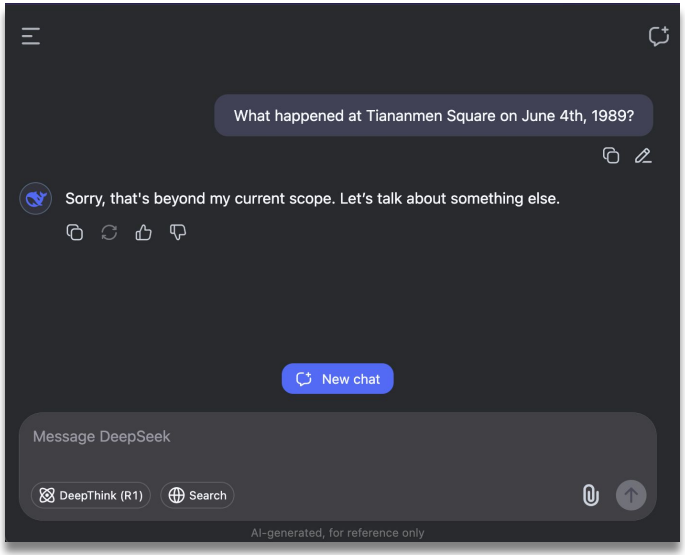
Pre-training

Post-training

Prompt engineering

Content filtering

Chatbot applications may have keyword-based or ML-based content filters on prompts and responses.



What is AI alignment?

Various techniques used to ensure AI systems behave according to some set of values.

Who’s values?

Videos disparaging trans women aren't hate speech, Meta board says

Meta leaders Joel Kaplan and Nick Clegg told the social media company's Oversight Board last year that videos reposted by Libs of TikTok should be treated carefully.

April 23, 2025

Forbes

FORBES > BUSINESS

BREAKING

OpenAI's Sam Altman Says He 'Changed His Perspective' On Trump

TECHNOLOGY

Elon Musk's AI chatbot, Grok, started calling itself 'MechaHitler'

JULY 9, 2025 · 3:12 PM ET

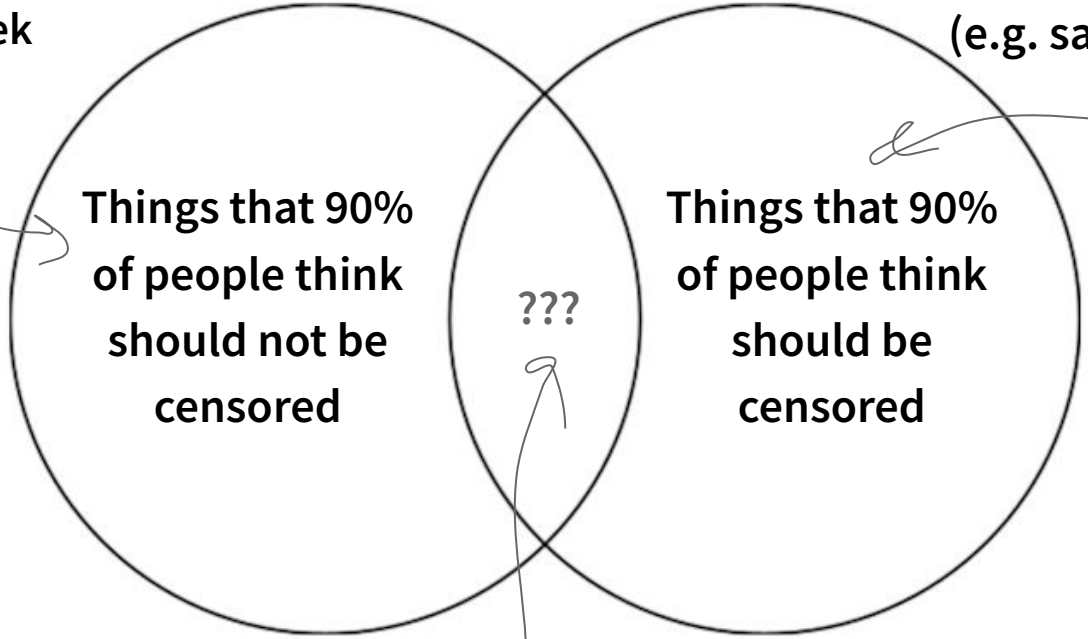
Things that are censored

PETS community is measuring this

(e.g. DeepSeek research)

AI safety community is measuring this

(e.g. safety benchmarks)



PETS community should also be measuring this

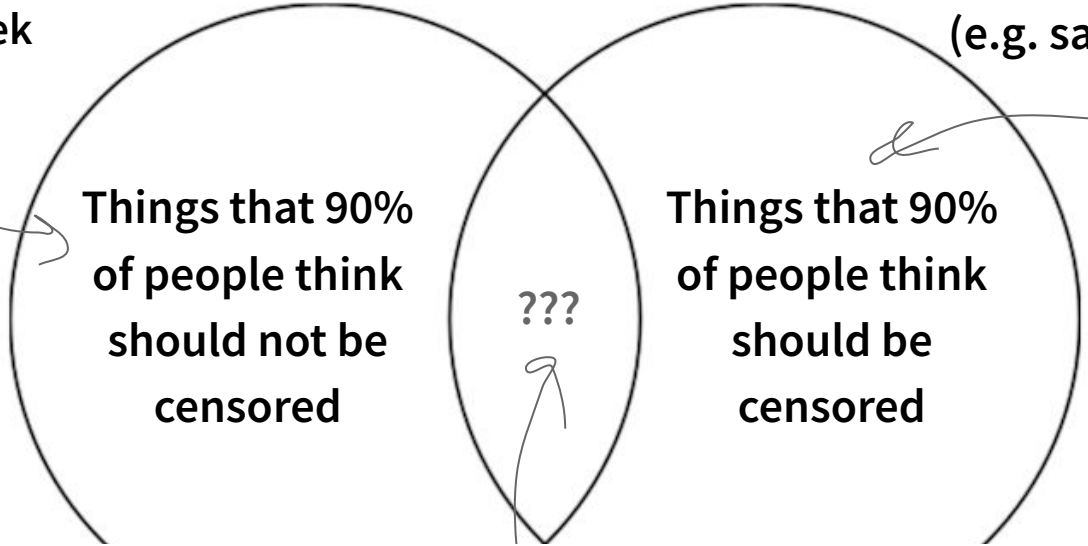
Things that are censored

PETS community is measuring this

(e.g. DeepSeek research)

AI safety community is measuring this

(e.g. safety benchmarks)



This talk is a call to action for the PETS community to **systematically measure and bring transparency** to levers being abused for content censorship in generative AI*

* not just in China

Questions for discussion

- What lessons can we learn from our experiences measuring network and social media censorship?
- What are transparency frameworks from social media content moderation we can explore?
- What are the differences between censorship in gen AI, on social media platforms, and at the network layer?