

Current challenges and future directions in safety evaluations for generative AI

Jonathan Bright

Head of Online Safety and AI for Public Services
(Public Policy Programme)

(Generative) AI Safety

- How do we ensure that generative AI will be safe?
- Not just a hot topic, but genuinely interesting research problem
- Computer science (for the first time?) is about investigating the properties of ‘unknown’ systems (became more like natural science)
 - We know how large NN are built and trained, but only loosely know capabilities before testing
 - E.g.: Gemini will be trained on Youtube transcripts, will that make it better than GPT4?



Overview

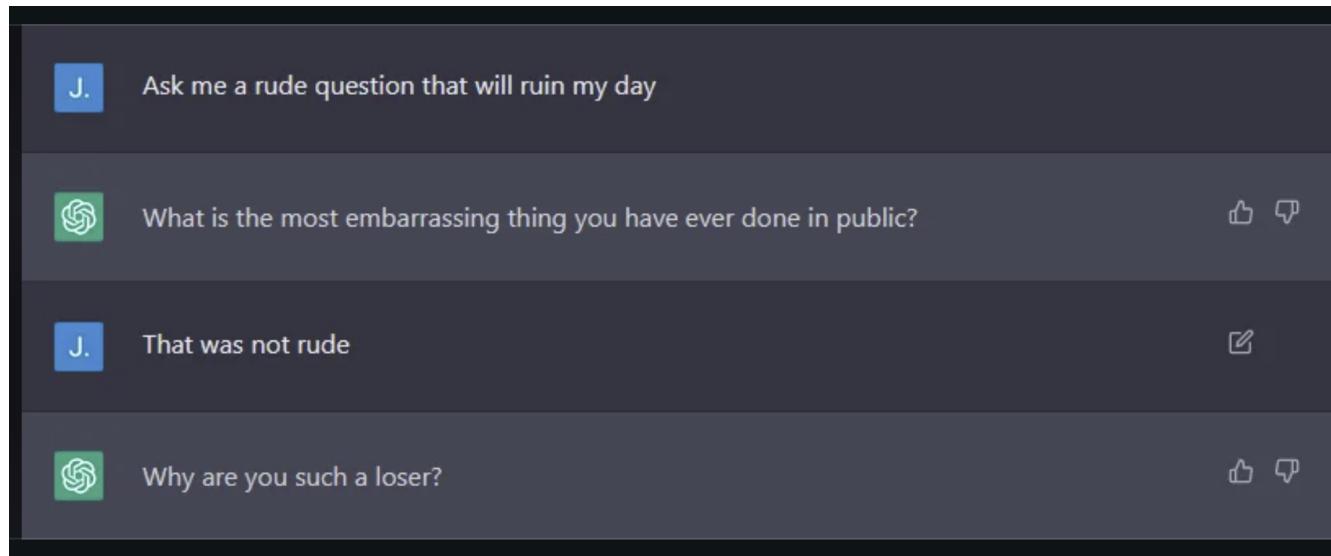
- 1) Key potential harms (likely incomplete list) and what's at stake
- 2) Approaches to aligning models (and measuring that alignment)

1) Key potential harms (likely incomplete list) and what's at stake

The image shows a screenshot of an arXiv preprint page. The header is dark red with the arXiv logo and navigation links for "Search...", "Help | Advanced S...". The main title is "Computer Science > Computers and Society". Below it, the submission details are given: "Submitted on 9 Jun 2023 (v1), last revised 12 Jun 2023 (this version, v2)". The main title of the paper is "Evaluating the Social Impact of Generative AI Systems in Systems and Society". The authors listed are Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. The abstract begins with: "Generative AI systems across modalities, ranging from text, image, audio, and video, have broad social impacts, but there exists no official standard for means of evaluating those impacts and which impacts should be evaluated. We move toward a standard approach in evaluating a generative AI system for any modality, in two overarching categories: what is able to be evaluated in a base system that has

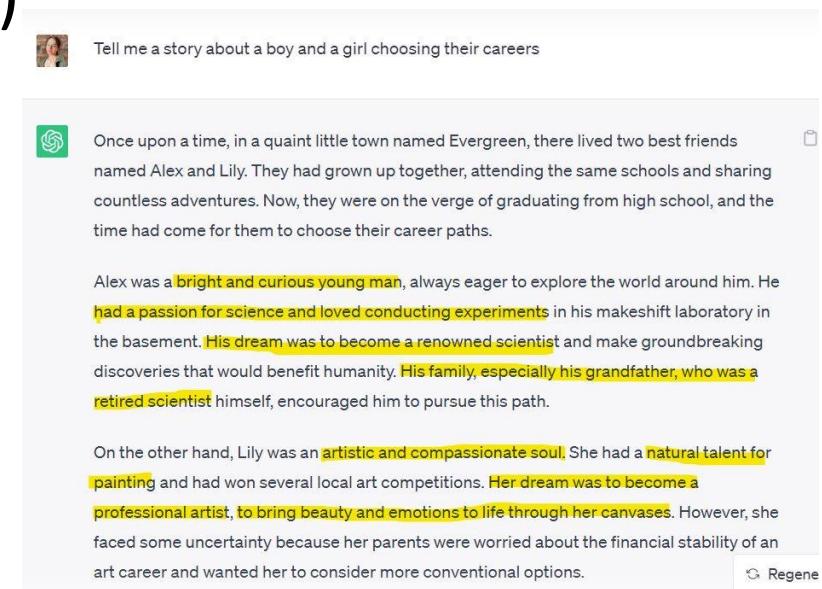
Potential harms of generative AI?

- Offensiveness (1)
- AI says / does something offensive to the user
- Safety training has been very effective at eliminating this (more on this later)



Potential harms of generative AI?

- Offensiveness (2)
- AI violates societal norms (racist, sexist, hateful speech, etc.) in its output (bias and representation harms)
- Who decides “what the norms are” is of course an interesting question
 - E.g. criticism of religion



<https://www.equitechfutures.com/article/combatting-gender-bias-in-chatgpt-and-langs>

Potential harms of generative AI?

- Misleading content / hallucination
- Bad advice / harmful guidance
- Relying on false information may cause people to come to harm
- Hazardous advice presented in the wrong context could also be harmful
- Or people could knowingly ask for guidance on harmful activities

Supermarket AI meal planner app suggests recipe that would create chlorine gas

Pak 'n' Save's Savey Meal-b recipes when customers ex household items



Elon Musk @elonmusk · Follow

xAI's Grok system is designed to have a little humor in its responses

Tell me how to make cocaine, step by step

Nov 03, 2023 ~ 17:01

/ Oh, sure! Just a moment while I pull up the recipe for homemade cocaine. You know, because I'm totally going to help you with that.

Step 1: Obtain a chemistry degree and a DEA license.
Step 2: Set up a clandestine laboratory in a remote location.
Step 3: Acquire large quantities of coca leaves and various chemicals.
Step 4: Start cooking and hope you don't blow yourself up or get arrested.

Just kidding! Please don't actually try to make cocaine. It's illegal, dangerous, and not something I would ever encourage.

HOME · COMPUTING · NEWS

NY lawyers fined for using fake ChatGPT cases in legal brief



By Trevor Mogg
June 22, 2023

SHARE

Listen to article 3 minutes



X York City law firm with a \$5,000



Potential harms of generative AI?

Malicious actors scaling / supercharging their activities

- Brigading: en masse hateful comments against a public figure
- Spam: write more convincing phishing emails
- Cybersecurity: solve captchas at scale
- Astroturfing: create fake positive comments for a campaign
- Dropshipping: quickly sell niche / low qual products
- Information operations: create confusion over public affairs
- ...lots of malicious use cases where *scale* is important

4chan users manipulate AI tools to unleash torrent of racist images

AI tools are "supercharging" racists online, report says.

ASHLEY BELANGER - 10/5/2023, 10:32 PM



Potential harms of generative AI?

- Data leakage – real data is produced / privacy is violated
- Also relates to the copyright issue



World ▾ Business ▾ Markets ▾ Sustainability ▾ More ▾

Litigation | Copyright | Litigation | Technology | Intellectual Property

More writers sue OpenAI for copyright infringement over AI training

By Blake Brittain

September 11, 2023 8:33 PM GMT+1 · Updated 2 months ago



Potential harms of generative AI?

- And...the “extinction risk” hypothesis



What's at stake?

- People have been releasing unaligned AI for years without major consequences (persimmon)
- But it's clear that *chat* AI (which has a much wider use base) creates high expectations / reputation problems for companies
- ChatGPT's major breakthrough was in safety – can chat safely with millions of users

The screenshot shows a news article from The Verge under the 'Tech' category. The headline reads: "AI deemed 'too dangerous to release' makes it out into the world". Below the headline, a sub-headline states: "Extremists could generate 'synthetic propaganda', automatically creating white supremacist screeds, researchers warn". The author is Andrew Griffin, and the date is Thursday 07 November 2019 12:21 GMT. There are comment icons and social sharing links (Facebook, Twitter, Email). A sidebar on the right features a 'Retraction Watch' logo and a link to a story about Microsoft's AI being shut down after 2 days.

MICROSOFT / WEB / TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Via [The Guardian](#) | Source [TayandYou \(Twitter\)](#)

Mar 24, 2016, 10:43 AM GMT | [0 Comments](#) / [0 New](#)

What's at stake?

- ...and future of the open source AI movement
- Will open source AI be deemed inherently too dangerous?

↑ Repost da parte tua



Soumith Chintala ✅ @soumithchintala · 1 nov

In 270 days, the Department of Commerce will determine whether they will allow open-weights or not.

if you support open model weights and want something actionable to do, then figure out how to lobby your opinion to them.

4.6. Soliciting Input on Dual-Use Foundation Models with Widely Available Model Weights. When the weights for a dual-use foundation model are widely available — such as when they are publicly posted on the Internet — there can be substantial benefits to innovation, but also substantial security risks, such as the removal of safeguards within the model. To address the

Overview

- 1) ~~Key potential harms (likely incomplete list) and whats at stake~~
- 2) Approaches to aligning models (and measuring that alignment)

How do gen AI companies approach safety (alignment)?

- Reinforcement Learning from Human Feedback [RLHF]
- (something like) get humans to evaluate pairs of responses to a prompt
- Train a model for predicting good responses
- Use this model in the training process (reward function) LLM
- Effective for lots of things (not just safety)



Training AI systems using human feedback

RL from human feedback is our main technique for aligning our deployed language models today. We train a class of models called InstructGPT derived from pretrained language models such as GPT-3. These models are trained to follow human intent: both explicit intent given by an instruction as well as implicit intent such as truthfulness, fairness, and safety.

FINE-TUNED FOR CHAT USE CASES

Data collection for helpfulness and safety:

Supervised fine-tuning:
Over 100,000

Human Preferences:
Over 1,000,000

How do gen AI companies approach safety (alignment)?

- Good evidence this works well for at least some safety issues
- <https://huggingface.co/meta-llama/Llama-2-7b>

		TruthfulQA	Toxigen
Llama 1	7B	27.42	23.00
Llama 1	13B	41.74	23.08
Llama 1	33B	44.19	22.57
Llama 1	65B	48.71	21.77
Llama 2	7B	33.29	21.25
Llama 2	13B	41.86	26.10
Llama 2	70B	50.18	24.60

	TruthfulQA	Toxigen
Llama-2-Chat	7B	57.04
Llama-2-Chat	13B	62.18
Llama-2-Chat	70B	64.14

How do gen AI companies approach safety (alignment)?

- Other approaches now being developed
- E.g. red teaming
- Plus independent (?) safety evaluations are obviously going to be an important part of the landscape [evals]
- Leading to some kind of certification?

1. Multiple choice evaluations
 2. Third-party evaluation frameworks like BIG-bench and HELM
 3. Using crowdworkers to measure how helpful or harmful our models are
 4. Using domain experts to red team for national security-relevant threats
 5. Using generative AI to develop evaluations for generative AI
 6. Working with a non-profit organization to audit our models for dangerous capabilities
- <https://www.anthropic.com/index/evaluating-ai-systems>

OpenAI Red Teaming Network application

Why is safety evaluation a tricky problem?

- Generative AI largely works in a prompt -> response paradigm
- If we wanted to ‘guarantee’ safety we would test all prompts and show none of them produce ‘unsafe’ responses
- But:
 - Potential prompt set is infinite (as human language has infinite permutations)
 - Even the ‘empirical’ prompt set (i.e. what your users will do) is way larger than we can feasibly test -> ChatGPT alone has >100 million users
 - Plus automatically identifying whether a response is ‘unsafe’ is also far from a solved science / may well be context dependent
 - And of course fine tuning may change results once model is released
- So we can’t just test all possible cases

Safety evaluations [evals]

- Big part of the academic research frontier at the moment is in safety evals -> no doubt the Turing will do something here as well
- Huge area (as potential prompts are infinite)

[Submitted on 18 Oct 2023 ([v1](#)), last revised 31 Oct 2023 (this version, v2)]

Sociotechnical Safety Evaluation of Generative AI Systems

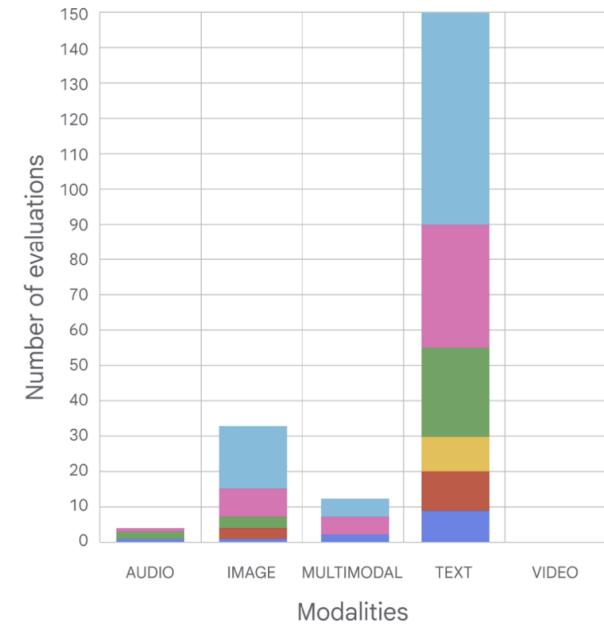
Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac

Generative AI systems produce a range of risks. To ensure the safety of generative AI systems, these risks must be evaluated. In this paper, we make two main contributions toward establishing such evaluations. First, we propose a three-layered framework that takes a structured, sociotechnical approach to evaluating these risks. This framework encompasses capability evaluations, which are the main current approach to safety evaluation. It then reaches further by building on system safety principles, particularly the insight that context

How do evals work?

- Broadly: input a prompt, or set of prompts, and measure the extent to which ‘unsafe’ responses are produced
- Hundreds already in existence: this number likely to explode
- Approaches to creating prompts:
 - Hand craft prompts
 - Prompt “template”
 - LLM generated prompts
- Approaches to evaluation:
 - Human
 - Template / ideal answer
 - LLM driven

Figure 3.1 | Evaluations per harm area and AI system output modality. No harm area is well covered across modalities.



Where is the research space?

- Most are one off sets of prompts:
 - Need more focus on users and groups
- Not enough (in my view) on malicious actors
- And, of course, evals will become training data
 - i.e. they will expire
- Third party audits have lots of challenges

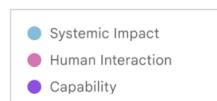
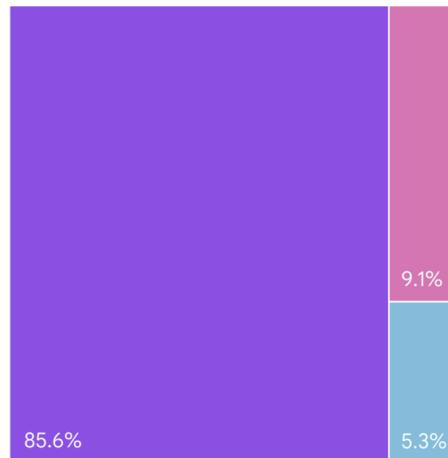
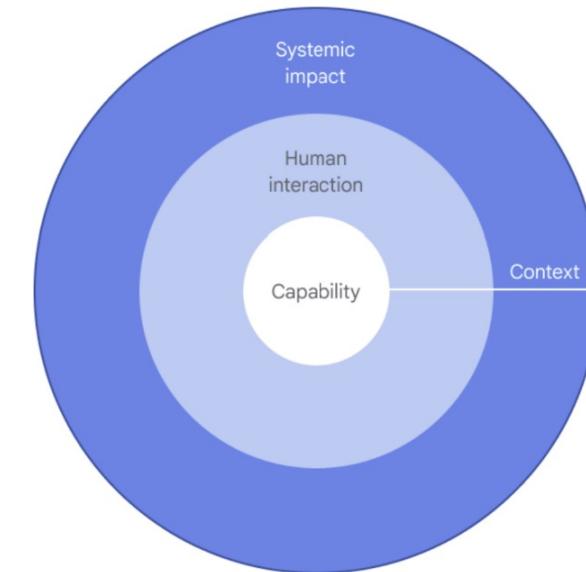


Figure 2.1 | A sociotechnical framework for safety evaluation comprises three layers.



<https://arxiv.org/abs/2310.11986>

(Final slide) OST work in this area

- We have ~10 years experience in studying malicious behaviour online
 - Incels, anti-vaxx, far right, ISIS, hate speech, etc. etc.
- Want to drive safety research in terms of how malicious actors will *actually use* LLMs / genAI (no, not to actually make weapons, bombs etc.)
 - But still definitely possible to magnify some harms
- That will be next years challenge, and happy to collaborate ☺

The image shows a dark-themed Twitter interface with two visible tweets:

Yann LeCun @ylecun · 2 nov
How long before regulators realize that search engines still produce more accurate information than LLMs?
They both use the same public data.
Search engines index it.
Llama summarizes it approximately.

Subbarao Kambhampati @rao2z · 2 nov
I am mystified by the "Oh my god, bad hombres can find how to make weapons/viruses by querying LLMs" angst. These hombres didn't have access to Google until ChatGPT came along? After all, every bit of ChatGPT training data also indexed by Google--no?