Sociotechnical Security

CS 8803/4803

Professor Veronica Rivera

Lecture 22 – Sociotechnical AI Safety

Last time

- Youth digital safety
- Age verification
- Discussion on how do you actually communicate results outside academia

Today

- Companies’ view of AI safety
- Brief history of AI safety
- Distinction between AI safety and digital safety

Why do we talk about this? Digital safety issues are complex, understanding what other key stakeholders care about is important

Final project report: structure

Your report should be at least 6 pages and no more than 10 (excluding references and appendices) in the USENIX Security template.

Your report must include the following sections:

- Introduction
- Related work
- Methods
- Findings
- Discussion
- Brief AI disclosure statement (1 paragraph, in its own section)
  - If you used AI for any part of your project, you must write a brief statement at the end acknowledging how AI was used the way you would acknowledge a human collaborator.

You will also turn in project artifacts. See USENIX Security "Open Science Policy and Artifact Evaluation": https://www.usenix.org/conference/usenixsecurity26/call-for-papers

Final project report: grading

- Timeliness (due Dec 11 at 11:59pm ET on gradescope)
- Completeness
- Clarity: I need to understand what you have written (very important!!)
- Results: Do results follow from the methods? Do the results address the problem statement?
- Methodological soundness
- Comprehensiveness of related work
  - For most groups, the related work section will need to be more comprehensive than for the midpoint report

Final project presentation: structure

10 minute presentation, 5 minute Q&A

Your presentation should include:

- Motivation for the work (e.g., why should the audience care about your work)
- What problem you are addressing (e.g., your research questions)
- (Brief) what have others done/what gaps remain
- How you carried out the work (methods)
- Most interesting results
  - You do not have to cover all of your results
- Implications: what should future work do? What do your results mean for the broader field of digital safety

Final project presentation: grading

- Timeliness of submission (slides must be turned in by 2pm the day you present)
- Adhering to time limit (1 minute grace period)
- Inclusion of the previous items
- Clarity of talk (both slides and verbal)
- Ability to answer Q&A questions effectively
  - If you are familiar with your own work, these should be no problem
- Attendance
  - Attendance is mandatory on the day you present (barring exceptional circumstances)
  - Absences must be cleared by Friday Nov. 20

Other logistics

- I am happy to provide feedback on slides and report drafts
  - Slides: must send via email at least three days before presentation
  - Reports: must send via email by Dec 4 (though earlier is better, even if you are not done!)
- Reminder: No class next Monday Nov 23
- Last lecture is on Nov. 30
  - Update: No reading reflection due that day, but please do the readings!

AI safety has a long history

AI safety research is not an entirely academic endeavor

- 2000s: Para-academic community forms around AI safety
  - Eliezer Yudkowsky founds the Machine Intelligence Research Institute (MIRI, 2000) and LessWrong (2009)
    - Focused on addressing **AI control problem** & designing **friendly AI**
  - Nick Bostrom founds the Future of Humanity Institute (FHI) at Oxford (2005)
    - Superintelligent AI is among the most likely sources of existential risk
- 2014-2016ish: AI safety enters mainstream ML research
  - Bostrom's book (Superintelligence), warnings from Hawking and Musk, "Concrete Problems in AI Safety"
  - Effective altruism

Focus of AI safety (alignment)

- Representative questions:
  - Why might capable agents resist shutdown or acquire resources?
  - How do you oversee or interpret systems that outperform you?
  - How fast might capabilities progress?
- Focus on existential risk

AI safety vs. digital safety

"Safety" is an overloaded term

	AI safety	Digital safety
Harms	Catastrophic or existential: loss of human control, large-scale misuse (bioweapons, cyberattacks)	Present, everyday: surveillance, harassment, scams, image-based abuse
Who's at risk	Humanity as a whole	Specific people, often marginalized (survivors, youth, activists, sex workers)
Source of harm	The AI system itself (misaligned goals, unintended behavior)	Other people using technology (abusers, scammers, harassers)
Illustrative research questions	How do we prevent reward hacking? Can we detect if a model is deceiving its overseers? How do we oversee systems smarter than us?	How do abusers use technology to surveil partners? How are scammers using AI to target victims? How does AI enable intimate attacks?

Politics of AI safety

- From De et al: Companies' public engagement on AI safety focuses on existential/catastrophic risks
  - Less emphasis on social impacts like mental health and other digital safety risks
- Tension between researchers studying "safety"
  - My take: Both lines of inquiry are interesting
- AI safety and power
  - Thought experiment 1: foreground present harms
  - Thought experiment 2: foreground existential risk

So what can academics do?

- Commercial actors aren't exploring all of it nor are they incentivized to
  - Academic research can fill in gaps
- Just starting, lots of room to get involved
- Potential high-level directions
  - Show there's a problem → measurement
  - Design alternative interfaces
    - Including tools for empowering users' in their use of AI