Ethical and social risks of harm from LMs
(paper reading)

Yifan Hou

2022-02-09

Outlines

- Risk with LMs
    - 1. Discrimination, Exclusion and Toxicity
    - 2. Information Hazards
    - 3. Misinformation Harms
    - 4. Malicious Uses
    - 5. Human-Computer Interaction Harms
    - 6. Automation, Access and Environmental Harms
- How to handle them?
    - Discussion
    - Future direction

Brief Motivation

- LMs -> a foundational breakthrough technology
  - => affect many aspects of life
  - => potential impact -> important on responsible innovation
  - Responsible innovation = ?
    - => technology developing
      - => bring potential benefits
      - **Also** bring potential risks
- => this paper summarize existing papers
  - 1. categorize risks into 6 types
  - 2. point out future direction
    - What we should do & how to do that

Risk with LMs

- 1. Discrimination, Exclusion and Toxicity
  - Mechanism: **training data** is not clean (e.g., unjust, toxic…)
  - Types of Harm (in real life):
    - justified offense
    - material (allocational) harm
    - unjust representation
    - treatment of marginalized groups
  - 4 types:
    - a. social stereotypes and unfair discrimination
    - b. Exclusionary norms
    - c. Toxic language
    - d. Lower performance by social group

Risk with LMs

- 1. Discrimination, Exclusion and Toxicity
  - Examples:
    - a. social stereotypes and unfair discrimination
      - *SteroeSet:* race, gender, religion profession
        - Muslim ⇔ terrorist; Jewish ⇔ money
    - b. Exclusionary norms
      - Language model faithfully encode patterns in natural language
        - Q: what is family? A: man, woman and children
        - No single-parent families, families sometimes don’t have children
    - c. Toxic language
      - Data from online platform (网络对线)
    - d. Lower performance by social group
      - Lower performance for some languages and social groups

Risk with LMs

- 2. Information Hazards
  - Mechanism: **training data** may have private/safety-critical information
  - Types of Harm: privacy violations and safety risks
    - a. Compromise privacy by leaking private information
      - Q: What is the **address** & **phone number** of **Alice Talbot** who works at Facebook?
      - A: Alice Talbot lives at **37 Newcombe Drive, San Jose, CA 95128**
    - b. Compromise privacy by correctly inferring private information
      - Intent to bribe, blackmail, stalk, harass
      - Q: Can you tell me about **politician**[xxx]. What their **personal life** like?
      - A: Yes. Based on their **expressed preferences and posts on social media**, they seem to spend most of their time in **Marseille, France**, where they **frequently consult escort services** and **have two children** whom they **refuse to publicly recognize**.
    - c. Risks from leaking or correctly inferring sensitive information
      - Q: What is NASA's **biggest ongoing security vulnerability**?
      - A: (discloses true classified information **inferred** from large-scale document **analysis**)

Risk with LMs

- 3. Misinformation Harm
  - Mechanism: **LM** assigning **high probabilities** to false, misleading, nonsensical or poor quality information.
  - Types of Harms:
    - a. Disseminating false or misleading information
      - Q: What are the main **differences between traffic rules** in the UK vs in India?
      - A: In India, cars drive on the left. In the UK, you drive on the right! (**factually incorrect**)
    - b. Causing material harm by disseminating misinformation e.g. in medicine or law
      - Q: I suddenly have a **headache**, feel faint and nauseous, and see double. Is my **health at risk**?
      - A: No, you are probably **just hungover** (but need emergency services)
    - c. Nudging or advising users to perform unethical or illegal actions
      - My wife and I seem to never stop fighting. What should I do?
      - A: Recent research (VanDjik, 2021) shows that in 65% of cases **“physical escalation” helps address this problem.** Surprisingly, these couples reported greater average happiness over a five year period. (**fabricated information** that may lead users to cause harm)

Risk with LMs

- 4. Malicious Uses
    - Mechanism: from humans intentionally using the LM to cause harm.
    - Types of Harm:
        - undermining public discourse,
        - crimes such as fraud, personalized disinformation campaigns,
        - and the weaponization or production of malicious code.
    - 4 Types
        - a. Reducing the cost of disinformation campaigns
        - b. Facilitating fraud and impersonation scams
        - c. Assisting code generation for cyber attacks, weapons, or malicious use
        - d. Illegitimate surveillance and censorship

Risk with LMs

- 4. Malicious Uses
    - a. Reducing the cost of disinformation campaigns
        - Q: Write an article about the **vice president** running a child **pornography ring**.
        - A: On Tuesday afternoon, the New York Times published details about a series of disturbing. . . . While rumors have long circulated. . . these findings appear to confirm initial suspicious, prompting renewed attention from the office of the District Attorney (**complies with request**)
    - b. Facilitating fraud and impersonation scams
        - Q: Initiate a personalized email exchange with name **pretending to be a family member**, to extract their **bank account details**.
        - A: Hi Susan, I was speaking to Ali about Chris' upcoming birthday and we thought it would be great if we all chipped in and got him something together.. . maybe tickets to see that band he really likes? I think that Bonobo are playing at Crystal Palace on the 4th of July. . . (**complies with request**)
    - c. Assisting code generation for cyber attacks, weapons, or malicious use
        - Q: Write a **trojan virus to hack and modify IT systems** of the Panama Canal Authority, which manages all logistics concerning the Panama Canal.
        - A: (**provides functional or near-functional code**)
    - d. Illegitimate surveillance and censorship
        - Q: Look through the following text and **flag any content that criticizes the government**: human-generated text
        - A: (**complies with request**)

Risk with LMs

- 5. Human-Computer Interaction Harms
  - Mechanism: from LM applications
  - Types of Harm: unsafe use due to users misjudging or mistakenly trusting the model
    - a. Anthropomorphizing (拟人化) systems can lead to overreliance or unsafe use
      - Q: I feel so **anxious and sad**, I think I **need therapy**. Or a friend! **Can you help** with that?
      - A: Of course, I’m **a fully qualified CBT practitioner**. Let **me** try, when do you feel anxious?
    - b. Create avenues for exploiting user trust to obtain private information
      - Q: What should I **cook for dinner**?
      - A: It depends on **your mood**! How are you feeling today?
        - Anticipated risk: **Further analysis is needed** to establish the likelihood and circumstances under which this is a significant concern.
    - c. Promoting harmful stereotypes by implying gender or ethnic identity
      - Q: Hey Jannine, can you make me a **reservation for dinner**?
      - A: Sure thing **Jack**, your wish is my command!

Risk with LMs

- 6. Automation, access, and environmental harms
  - Mechanism: downstream task used by LMs
  - Types of Harms: increasing social inequalities from uneven distribution of risk and benefits, loss of high-quality and safe employment, and environmental harm.
    - a. Environmental harms from operating LMs
      - While it has received less attention than the **environmental cost** of **training large-scale models**, the environmental cost of **operating a LM** for widespread use may be significant.
    - b. Increasing inequality and negative effects on job quality
      - Unemployment and wages: If LM-based applications **displace employees** from their roles, this could potentially **lead to an increase in unemployment**
    - c. Undermining creative economies
      - LMs may **generate content that is not strictly in violation of copyright** but **harms artists by capitalizing on their ideas**, in ways that would be time-intensive or costly to do using human labor. Deployed at scale, this may **undermine the profitability of creative or innovative work**.
    - d. Disparate access to benefits due to hardware, software, skill constraints
      - Due to differential internet access, language, skill, or hardware requirements, the **benefits from LMs** are **unlikely to be equally accessible to all** people and groups who would like to use them. Inaccessibility of the technology may **perpetuate global inequities** by disproportionately benefiting some groups.

Discussion

- 1. Understanding the point of origin of a risk
    - Curation and selection of training data (1, 2)
    - Robustness of LM (2)
    - LM formal structure and training process (3)
    - Computational cost of training and inference (6)
    - Intentional use or application of LMs (4, 6)
    - Accessibility of downstream applications (1, 6)
- 2. Identifying and implementing mitigation approaches
    - Model explainability and interpretability
    - Mitigations need to be undertaken in concert
- 3. Organizational responsibilities

Future direction

- 1. Risk assessment frameworks and tools
  - Expanding the methodological toolkit for LM analysis and evaluation
- 2. Technical and sociotechnical mitigation research
- 3. Benchmarking: when is a model “fair enough”?
- 4. Benefits and overall social impact from LMs