Tess Gompper
April 12, 2024

The Importance of Interdisciplinary and Representative Collaboration and Transparency in LLM Research and Development

Introduction:

Large Language Models (LLMs) offer many intriguing areas of research, technological advancement, and changes to our technological landscape. However, they also create many ethically ambiguous areas of concern and introduce the possibility for a variety of harms. In order to mitigate biases, harms, and risks, the industry needs applicable ethical considerations and solutions to make these technologies more moral, safe, and fair.

Data & Society, an independent nonprofit research organization, provides three "core concerns" surrounding data-centric technologies:

1. Data-centric technologies have social, cultural, and political implications that are far-reaching, unevenly distributed, and poorly understood,
2. These technologies disproportionately cause harm to systemically marginalized populations,
3. The concentration of power and wealth in the technology industry impacts the governance of data-centric technologies, which has consequences for democratic rights and practice. (*Data & Society*)

All three of these concerns are highly applicable to LLMs, as the development of LLMs is a highly data-centric task. Many of the risks associated with LLMs come from problematic training data. These training data can have "social, cultural, and political implications" as described in concern one. These implications often come to fruition in the form of harmful biases which are known to harm vulnerable populations, as we see in concern two. And, due to the "concentration of power and wealth in the technology industry" discussed in concern three, we do not often see democratic, inclusive practices which would work to mitigate harms and biases as we will discuss below.

Data & Society is creating a more just and equitable data world by working to "change the terms of debate," "shift power," "shape policy and practice," and "[build] organizational trust and equity" (*Data & Society*). Their first two items are particularly applicable when we consider ethical practices in LLM research and development. Data & Society describe "[changing] the terms of debate by challenging techno-solutionist narratives and pushing for nuanced, context-specific understandings of technology's role in society" and "[shifting] power by foregrounding systemically affected communities and offering approaches to design and governance that are grounded in equity and justice" (*Data & Society*). Similarly, as LLMs are typically designed to be used by average people in a wide variety of ways, as seen with ChatGPT, the research and development of LLMs must require highly "nuanced, context-specific understandings" in order to situate the "technology's role in society." To bring these understandings to the forefront of the research and development processes, we must shift power as described above to foreground those who have the necessary nuanced, context-specific knowledge and who will work to create fair technologies which will not disproportionately affect at-risk communities.

I argue that in order to make LLM technologies more ethical, safe, and fair, we need to foreground purposeful intentions, interdisciplinary collaboration, representative collaboration, and transparency in all of our research and development practices.

Purposeful Intentions: What Are We Doing Here?

In our fast-paced world of constant technological advancements and innovations, it can often be hard to spot the difference between the things we *can* do and the things we *should* do. When we deal with applications that will be used by humans, be it as tools for education, medicine, or entertainment, it is vital to have specific purposes and intentions. At the conception of and throughout any project, we must consider the following questions:

> Whose interests does this technology serve?
>
> Why are we creating this technology in this way?
>
> Why are we creating this technology at all? (Guyan 4)

Questions such as these help to create purposeful, intentional technologies, and help to mitigate the building of technologies just because we can. Similarly, "at the stage of scoping potential applications, it is worth asking whether a given technology is anticipated to be net beneficial - or whether it may cause harm when performing with high accuracy, such as certain kinds of surveillance tools, in which the application overall should be called into question"(Weidinger et al. 38). Considering net beneficence and potential harms at the conception of a project is also key in setting out with purposeful intentions and mitigating harms before they can occur. It also helps researchers and developers stay on track with creating because they *should* not just because they *can*.

Interdisciplinary Collaboration: Mitigating a Wicked Problem

In many cases of researching and developing technologies, such as LLMs, typical researchers and developers may not be equipped to recognize ethical opportunity areas or to ask the right questions. Usually, computer science researchers are not trained in appropriately recognizing and mitigating social risks and biases. For this reason we need humanities researchers, societal and community experts, ethics researchers and others on the research and development teams to recognize potentials for harm.

When we view the question of creating fair and equitable LLMs as a wicked problem, we see the value of utilizing interdisciplinary teams. A wicked problem is “often a social or cultural problem…with many interdependent factors making them seem impossible to solve”(Wong). The many dimensions of wicked problems “interact with one another in ways that are ever-changing and unpredictable. As a result, many wicked problems are never completely solved. Instead, the best one can hope for is a process of continual improvement in addressing the issue”(Johnson-Woods). With LLMs, we see many dimensions at play: technological, resource consumption, financial considerations, and of course ethical considerations, harm and bias mitigations. The best way, then, to address this multi-dimensional wicked problem, is by “engaging with and integrating multiple perspectives” and “to engage in interdisciplinary collaboration, which mobilizes theories, methods, and practices from the natural, social, and human sciences”(Freeth and Caniglia 247).

Representative Collaboration: Who Knows Best?

Arguably, the most useful tool to combat harmful bias and misrepresentation in LLMs is integrating affected community members into the research and development processes. Anyone who will use or be impacted by use and/or development of these technologies should be represented in or heavily consulted by research and development groups. Researchers and developers who do not come from marginalized communities who may be negatively affected by new technologies do not possess the necessary cultural background and knowledge to recognize potential sources of harm and harmful biases and appropriately mitigate them. Marginalized and at-risk communities include those who are disproportionately affected by embedded biases in LLM technologies.

“Responsible innovation is a collaborative [endeavor]. In order to anticipate and mitigate risks posed by technology successfully, we need to view these issues through multiple lenses and perspectives”(Weidinger et al. 6). It is key to include perspectives and lenses that are close to the ground and understand the social nuances and risks at play. To incorporate many views, “it is important to find ways of collaborating with a wide range of stakeholders to robustly address risks of ethical and social harm”(Weidinger et al. 38). Stakeholders include those who, as discussed above, come from and understand the communities at risk and “have capacities to implement such mitigations” (Weidinger et al. 38). Capacities can be financial or resource based, such as time, but they can also include a necessary stability and mental capacity to take on these representative roles.

Although some researchers may consider themselves to be educated and aware enough to mitigate biases and harms, they may not be educated enough to observe when the mitigation of one bias creates another, as these situations can be very nuanced. For example, “methods to reduce toxic speech from [LLMs] have been found to bias model prediction against [marginalized] groups. In this way, a focus on one mitigation at the expense of the other risks may cause negative outcomes” (Weidinger et al. 38). Thus “it is important to keep a broad view to ensure that fixing one risk does not aggravate another”(Weidinger et al. 38). One of the best ways to keep this broad view is via the inclusion of affected community members who better understand the social nuances at play.

Transparency: There's No Such Thing as No Bias

In a perfect world, our LLMs would be harm- and bias-free, created with purposeful intentions by interdisciplinary, representative collaborative teams. But, of course, this is not a perfect world, and we will never be able to create completely bias-free LLMs. However, there are steps we can take to mitigate harms of bias, namely: transparency datasheets.

Following a feminist approach to data, we can adapt "datasheets for datasets" to apply to our LLM transparency datasheets:

> Inspired by the datasheets that accompany hardware components, Timnit Gebru and colleagues advocate for data publishers to create a short, 3-5 page document that accompanies data sets and outlines how they were created and collected, what data is missing, whether preprocessing was done, how the dataset will be maintained, and legal and ethical considerations such as whether the data collection process complies with privacy laws in the EU (D'Ignazio and Klein 107).

Similar information on LLM training datasets should be collected and reported, as these data are largely responsible for biases present in the model. It is "important to transparently disclose what groups, samples, voices and narratives are represented in the dataset and which may be missing" (Weidinger et al. 12) in order to provide a full documentation of biases in training corpora.

By "providing broad and transparent dataset documentation"(Weidinger et al. 12) via transparency datasheets, an awareness is created about potential biases present in the given model. Thus, although biases continue to exist, this awareness allows for a mitigation of risks and harms these biases may have otherwise caused.

Conclusion: Where Do We Go From Here?

Foregrounding purposeful intentions, interdisciplinary collaboration, representative collaboration, and transparency in LLM research and development practices will not rid these technologies of all biases, harms, and risks but it will make these technologies more ethical, safe, and fair. However, these practices cannot work if they are not widely adapted. So, the next step in making our Large Language Models more ethical is to upset the hierarchy of power, in both research and industry, that allows these harms and risks to permeate our technologies and redistribute power towards prioritizing the practices detailed in this paper.

Bibliography

Data & Society, datasociety.net/. Accessed 10 Apr. 2024.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. The MIT Press, 2023.

Freeth, Rebecca, and Guido Caniglia. "Learning to collaborate while collaborating: Advancing Interdisciplinary Sustainability Research." *Sustainability Science*, vol. 15, no. 1, 23 May 2019, pp. 247–261, https://doi.org/10.1007/s11625-019-00701-z.

Guyan, Kevin. *Queer Data: Using gender, sex and sexuality data for action*. London: Bloomsbury Academic, 2022.

Johnson-Woods, Courtney. "The 10 Characteristics of 'Wicked Problems.'" *Resonance Global*, Resonance, 30 Mar. 2023, www.resonanceglobal.com/blog/the-characteristics-of-wicked-problems.

Weidinger, Laura, et al. "Ethical and social risks of harm from language models." *arXiv preprint arXiv:2112.04359* (2021).

Wong, Euphemia. "What Are Wicked Problems and How Might We Solve Them?" *The Interaction Design Foundation*, Interaction Design Foundation, 10 Apr. 2024, www.interaction-design.org/literature/article/wicked-problems-5-steps-to-help-you-tackle-wicked-problems-by-combining-systems-thinking-with-agile-methodology.