Summary of Ethical Considerations

Yuchen Xia (stilex), Yeda Song (yedasong), Hendrik Mayer (htmayer),

Problem and Motivation

As the growing use of generative AI impacts human life and society, it is crucial to evaluate the risks and harms posed by these systems. Today, increasingly large language models are being released. While the trend of larger models often correlates with improved performance, we should take a step back and develop methodologies for risk assessment and mitigation.

Related Works

Language models have evolved from n-gram models to neural models, and from LSTM models to pre-trained transformer models. All of these systems are systems trained to predict sequences of words, but differences have been made in two main aspects: (1) the scale of the dataset used and (2) the scope of the tasks the model is capable of. By scaling up in these two ways, modern extremely large LMs incur new kinds of risk.

Solution Overview

The first paper, **"Sociotechnical Safety Evaluation of Generative AI Systems"**, introduces a 3-layer sociotechnical evaluation framework to analyze generative AI systems. First, the capability evaluation layer tests the behavior and outputs of a generative AI system. This is a common type of evaluation that is done and can include measures of bias and even downstream environmental harms. Second, the human interaction layer focuses on the experiences and interactions of individuals with the AI system. Third, the systemic impact evaluation centers on the widespread societal repercussions of specific generative AI systems. For example, AI systems have changed how education functions throughout the world. Overall, this evaluation framework provides more of a human lens for evaluating AI systems than traditional evaluation metrics. However, existing evaluations are not as comprehensive and have not been as extensively conducted as the authors believe they should be. About 85.6 percent of the evaluations that the authors tracked in Figure 3.2 have focused on capability evaluations; the human aspects of their proposed framework are rarely applied to generative AI systems in practice. Furthermore, the majority of the current evaluation is focused on text modality. There are specific ways that the authors believe the multimodal evaluation gap can be closed. The paper also suggests roles and responsibilities among various AI actors. Different groups in society have different roles in evaluating AI systems. For instance, AI developers may conduct capability evaluations, while application developers lead human interaction evaluations.

The second paper, **"On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"**, describes specific categories of costs and risks such as "Environmental costs", "Financial costs", "Risk of substantial harm", and "Opportunity cost". The authors propose that

LLMs are like “stochastic parrots”, that is they stitch together pieces of the training data without actually understanding language at a fundamental level. Training LLMs uses a lot of compute, and much of the energy used from this compute comes from non-renewable sources. While Google leads the industry in using renewable energy, according to the class presentation, the industry still needs to move towards renewable energy sources as a whole. However, even when renewable energy sources are used, there are infrastructure costs when building wind or solar farms, for example. In addition, there has been prior work showing that climate change disproportionately impacts marginalized populations, who are less likely to reap the benefits from modern generative AI technologies. On the other hand, the financial costs of developing and deploying large language models are substantial. Strubell et al. estimate that training a single BERT base model without hyperparameter tuning costs significant energy resources and financial investment. For example, achieving a modest increase of 0.1 BLEU score in machine translation can cost an additional $150,000 in compute expenses. Moreover, the cost of inference can far exceed the already-high training costs for industrial-scale deployments.

Limitations

Evaluation is incomplete: No evaluation framework can fully capture all possible risks and scenarios for generative AI systems. Risks are often dynamic and context-dependent, and they may emerge over time or in unforeseen ways after system deployment. For example, AI systems might interact with unexpected user behaviors or external systems, creating emergent risks that were not identified during evaluation. The real-world deployment of AI systems often exposes latent risks that were not evident in testing environments. Some harms, such as systemic societal impacts, require longitudinal studies or large-scale observation post-deployment.

Evaluation is never value-neutral: The design and application of evaluations always involve subjective decisions, such as:

- What risks to prioritize
- Which metrics to use and how to interpret results

These decisions reflect the values and goals of the evaluators, which can vary across cultures and organizations. Evaluations may unintentionally perpetuate existing biases or inequalities. For example, tools designed to measure harm in one culture may not transfer effectively to another. Certain harm on underrepresented groups might be neglected. The lack of universal standards for evaluation exacerbates these issues, making it difficult to compare results or establish shared benchmarks for safety and reliability. In addition, evaluations can be influenced by biases, potentially leading to predetermined outcomes or criticisms.

Future Research Directions

There are many areas, such as multimodal outputs or societal impacts, where evaluations are lacking. Developers and researchers should prioritize creating new methods and metrics to address these gaps, particularly for risks that are currently underexplored. Conducting evaluations should become a routine part of AI system development and deployment. These

evaluations must be applied consistently across development cycles to ensure potential risks are mitigated early. Evaluation results should directly inform decision making about AI systems. They should also guide policy and governance structures, with penalties for systems that fail to meet safety standards. Moreover, there should be incentives aligned to ensure developers prioritize robust evaluation and risk mitigation. The AI community should work toward establishing standardized evaluation frameworks to ensure consistency, comparability and reliability of results across different organizations and systems. Collaborative efforts across the AI ecosystem–developers, policymakers and researchers–are essential to develop and deploy shared frameworks for AI safety evaluations.

Summary of Class Discussion

After some clarifying questions were asked, we moved on to a class debate. We were discussing how we should allow AI to progress. Specifically, should there be regulation on AI progress or not? First, we summarize some of the points made by the side for unlimited development. It was argued that we do not know that AI is harmful yet. Someone also brought up the distinction between having regulation in deployment versus development. They argued that it is hard to enforce guardrails in development and that guardrails do not make sense during development. Regulating AI may slow down some of the benefits, such as automatic medical diagnosis. They also argued that a super intelligent AI may not be very powerful. If it is just a software system that cannot act in the world, then it could be controlled. Another point was that regulating AI makes it less democratic. The only ones who will be able to hire the armies of lawyers to deal with the regulations are large corporations. Regulations may be a barrier to entry. Also, AI is like the internet, which cannot be regulated at scale. Someone will develop AI systems without regulations; if we want to keep up, then we have to do the same. It is similar to an arms race.

Now, we summarize the points made by the side for regulated development. Even though we live in a capitalist society, there are still regulations present that benefit our economy. Movies like “The Terminator” suggest that AI may not remain software forever and could cause serious harm. At the very least, AI being allowed to kill humans should be regulated. More broadly, there are minimum regulations that should be enacted for AI systems. It may be infeasible to regulate how models are designed, but there may be ways to regulate datasets or other parts of the development. It is still possible to regulate AI systems and not inhibit innovation. AI should be regulated because it is powerful, not because it is inherently evil. AI progress may also cause income inequality to become a greater problem. AI is similar to nuclear power and weapons; it should be treated as such for regulation. You can specifically outline things that you do not want AI to be able to do when drafting regulations. Regulations do not have to be a barrier to entry. For example, the regulations could be information disclosures before deployment. There are many types of potential harms associated with AI. For instance, a Sora model could generate a video of anyone doing anything. We should regulate how often companies are able to retrain large models from scratch. When left unchecked, constant training uses up a lot of compute, which can have environmental implications.