Preamble

Aug 13, 2021
1001 West Imperial Highway
#2463
La Habra, CA 90631

Subject: AI Risk Management Framework

About Us:
Preamble is a US based small-business that is focused on creating an ecosystem that provides digitally defined human values in Artificial Intelligence. Preamble's work is focused on recommender systems, securing large language models, and Artificial General Intelligence Safety. We are a team of computer scientists, AI researchers, and cybersecurity professionals who want to build safe and ethically minded products to improve society.

Request #1:

The greatest challenges in improving how AI actors manage AI-related risks – where “manage” means identify, assess, prioritize, respond to, or communicate those risks;

Response:

The lack of regulation in tech and especially in the AI space, allows AI service companies to only worry about mitigating enough AI risk to maintain a fair image for public relations. In order to manage risk, standards need to be developed to identify, measure, then make adjustments. The greatest challenge is capturing the societal level risks from the large scale deployment of AI algorithms for big tech companies. The cumulative risk over an extended period of time can turn into a massive problem for society. The risk where a 1 in 1,000 chance that an individual is radicalized by YouTube or TikTok videos, can aggregate into a massive problem for society. One component to assess AI risk could be to measure unintended time on a platform due to recommender systems.

AI alignment and measures of alignment could be a good starting point for measuring AI risk as described in one of our papers - https://arxiv.org/abs/2107.10939.

Request #4:

The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management – including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;

Response:

Preamble

In terms of safety, the risk unmeasured by tech companies that use dangerous algorithms to recommend content on social media platforms can lead to negative personal behaviors from the consistent reinforcement of content that was intended to increase user engagement. Unfortunately, one of the most efficient ways to increase user engagement is to recommend content that is morally outrageous. Another issue with content recommendation algorithms is that training from user engagement and personal information can lead to more privacy and safety concerns when personal information is collected from individuals at a large scale. If users could otherwise select their preferences from an independent party, then their personal information would be isolated, while curating more quality content to present the user.

Request #12:

The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Response:

The framework should include governance that outlines how transparent models are and how much control users have over the content recommendation algorithms through a middleware service. There also needs to be a way to measure influence from the algorithms. This measurement could then help characterize the decisions that were made on the platform. Once there are tools and methods to measure the influence and choices for algorithms, then standards for these measurements can be set. An independent certifying party could monitor and evaluate AI services. The main caveat to any monitoring solution is having access to the data from big technology companies pre-deployment. It is hard to assess what could happen in a real world deployment until it happens, unless a governing body has early access for auditing. A market solution could be created that monitors AI systems to assess the inputs and compare them to the outputs in order to determine the appropriate settings that can be made to realign the system.