FLI Position Paper on the OECD Survey to classify AI systems

The Future of Life Institute (FLI) is one of the world's leading voices on the governance of AI. The institute is an independent nonprofit that works on maximizing the benefits of technology and reducing its associated risks.

FLI created one of the earliest and most influential set of AI governance principles - the Asilomar AI principles – and maintains a large network among the world's top AI researchers. The Institute, alongside the governments of France and Finland, is also the civil society champion of the UN Secretary General's Digital Cooperation Roadmap.

Members of the Future of Life Institute have filled the survey to classify two potential AI systems:

(1) **GPT-3**: an advanced language model used by a "global election management agency" (e.g., Cambridge Analytica) to generate mass-customized Facebook posts to influence people of a particular affiliation (e.g. convincingly dissuade them from voting).

(2) **OSAGI**: a hypothetical open-source artificial general intelligence that exceeds the average human in intelligence, creativity, and well-roundedness, trained on the entire internet in all modalities

We found a few areas for improvement for the survey that we wish to present to you:

I. Human rights and well-being criteria

We find the distinction between well-being and human rights criteria tenuous and recommend all the well-being criteria become core criteria. In addition, we suggest these criteria be more extensively defined, along with suggested metrics on how to assess them, in order to avoid companies subjectively interpreting them (which seemed feasible from the survey).

II. Multi-tasks systems

The core application areas were considered mutually exclusive in the survey and it was not possible to select multiple options (e.g., both computer vision and natural language processing). This problem shows that assessing the risks associated with AI models based on their use is not adapted to the current evolution of AI. Systems are more and more often multi-tasked. Both Open AI's CLIP and DALL-E would fall into both computer vision and natural language processing. DeepMind's MuZero could fall under any task given that it could learn as a planning algorithm, and this decision would not yet be made at the level of the placing on the market.

III. Open source systems

We find that open-source AI systems that could potentially be high risks could evade any regulation under this framework. Given that they would be open-source, the users of the system, whether they would serve a critical function, what their tasks would be would all be unknown.

IV. Artificial General Intelligence

This framework was built with narrow-AI in mind, and high risk AGI could currently be considered low risk from this survey. We recommend adding generality criteria (e.g., is the system multimodal? Multitask? Self-modeling? Self-replicating? Learning autonomously?).

V. Emotional manipulation

The examples we selected for this survey were meant to illustrate that under the current framework, high-risk AI systems could easily be classified as low risk. A provider could easily game the form without explicitly lying. In one of our models, we responded to these questions from the perspective of a global election management agency (such as Cambridge Analytica for instance) which would consult for a political party and engage in advertising on Facebook. We find that there is no assessment of emotional manipulation, which we believe should be included within human rights as a core criterion, with a specific definition and metrics to measure its risk.