EUROPEAN UNION
DELEGATION TO THE UNITED STATES OF AMERICA
San Francisco Office

Friday, 21 April 2023

REPORT // DG Connect mission to Silicon Valley (11-14 April, 2023): AI tech developments, governance and the EU AI Act

Abstract

Out of scope [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[3] [redacted]

[redacted]

[redacted].

Personal data

OpenAI [redacted] advocated for the development of standards to test LLMs and insisted that Generative AI should be addressed separately from the risk-based approach framework presented by the EU AI Act. They advocated for open-source solutions and the need of public benchmarks to evaluate LLMs model. They stressed that training data did not have a fundamental impact on the possible biases of the algorithm's outputs, while also highlighting that such screening and analysis would be very hard to implement as a legislative provision. The most relevant part of the safety validation and *debiasing* work wouldn't happen at the training data phase but after, where through reinforcement learning, human feedbacks (not by users but by the trust and safety team) and risk assessment they can prevent the most harmful applications of the technologies. Trust & safety team substantially trains the model for illicit and harmful behaviors and content to eliminate and avoid, disputably leaving a big gap of uncertainty for all the (potentially infinite) harmful applications that the trust & safety was not able to identify and prevent in the first place. However, data poisoning does represent a concern that the team is working on. On watermarking AI text, OpenAI responded that they were governed by a non-profit and they would not want to make value-bonding decisions – it may be an interesting application but not sure that it will fix the growing copyright and authentication concerns, while increasing the risk for text and data mining rules to be circumvented. Generative AI technologies will also have a growing role and importance in moderating content online. The discussion followed a demo of ChatGPT Visual and ChatGPT 4.

Out of scope

[redacted]

[redacted]

[redacted]

Personal data [redacted]