BigScience & AI2 Legal Hackathon

March 2023

About

The BigScience workshop (https://bigscience.huggingface.co/en/#!index.md) was an effort started in January 2021 that gathered 1000+ participants from over 40+ countries to research questions surrounding large language models (capabilities, limitations, potential improvements, **bias, ethics,** environmental impact, role in the general AI/cognitive research landscape) as well as the **legal and ethical** challenges around creating and sharing such models and datasets for research purposes and among the AI research community.

The core effort resulted in the creation of the multilingual Large Language Model BLOOM and of the Responsible Open-science Open-access Text Sources corpus (ROOTS) in June 2022, and brought together a multidisciplinary community of researchers working on different aspects of AI governance. In the spirit of continuing the project's mission of making AI technology more equitable, this community is organizing its second legal hackathon this year (see here for an outcome of last year's iteration) to help provide users and affected parties with legal resources related specifically to responsible model licensing and data subject rights.

Overview of the research topic(s) and scope

As part of this study, the researchers will leverage their legal expertise to advance governance processes to help make the use and development of AI systems more accountable to the technology's many stakeholders. The research will be focused on two main aspects of governance:

- Responsible AI Licenses are a new category of licenses for Artificial Intelligence and Machine Learning systems that include some **behavioral clauses** to restrict the technology from being used in ways that are deemed harmful — they can include prohibition from using AI systems in settings such as surveillance, requirements for disclosure of the use of AI in technology or for human oversight, or mandatory testing of bias dynamics that could exacerbate discrimination.
  While these types of licenses are seeing growing popularity in technical communities interested in developing AI systems more responsibly, there is still much work to be done to give these behavioral clauses standardized language that is both grounded in current and upcoming regulation of AI systems and interpretable to entities who want to put them in application.

- Machine Learning models are trained on data that is often about or created by people. The legal relationship between the training data, the models, and their output is still very much being debated, as are its consequences for legal flowdown of the data subjects'

rights.
As a consequence, data subjects who want to exercise their rights to exert control over are faced with a number of questions, including which legal mechanisms are the most appropriate for their specific case (*e.g.,* intellectual property, privacy rights), or what are the closest parallels they can leverage outside of AI (*e.g.,* how has “derivative work” of the training data been defined before the rise of large-scale data-driven technology).

The hackathon is hosted by the Allen Institute for Artificial Intelligence which is a 501c3 non-profit, in collaboration with BigScience, the RAIL Initiative, and Hugging Face.

Work Product

*around 50 hours of legal research per student.

- Collaborating on a taxonomy of behavioral use clauses and legal definitions of technical terms to support licensing efforts that promote more beneficial and equitable uses of AI technology
- Research memos on mechanisms for asserting data rights targeted at pre-identified groups of stakeholders (visual artists, writers, social media activity) in popular AI use cases (generative AI, code and writing assistants, chatbots)

Relevance of the project(s)

- We aim to create resources to facilitate more responsible development of the technology by actors concerned with its social impact, and to help individuals navigate and exercise rights to their data in a quickly evolving technical context.
- These outcomes will be directly useful to developers and researchers to help them promote responsible practices and help set positive norms for the technology by leveraging legally grounded governance tools

Relevant websites for context

- BigScience x NYU Law - 2022/23 Edition
- https://www.licenses.ai/
- https://sites.google.com/huggingface.co/big-science-data-governance/
- https://bigscience.huggingface.co/
- https://www.openml.fyi/2023-02-06/

LLM student participants

We seek LLM students with experience or interest in AI, IP, privacy, open research and data, and standards.