DIGITAL MEDUSA

netrified

Keeping the Internet and the Web open in the age AI systems and apps

Digitalmedusa.org

Imagine an educator that wants to do an educational AI app for some vulnerable community in their own language

We call her Kafka

- Kafka tries to train the open AI app through crawling the web and using other web crawler databases

But she can't

In summer of 2024,

"Data Provenance Initiative at MIT released a new paper by Shayne Longpre et al. that shows a dramatic increase in restrictions on the use of publicly available content as AI training data."

https://www.dataprovenance.org/consent-in-crisis-paper

https://www.dataprovenance.org/

Kafka cannot train the educational AI App in the local language and curating that data is very costly. Kafka can’t help the vulnerable community.

Why did this happen and how can we prevent it from happening?

First lets learn about Robots.txt

- Robots.txt is an old ad hoc protocol (issued in 1994, standardized in 2022 recently at the Internet Engineering Task Force- IETF, 9309 https://datatracker.ietf.org/doc/rfc9309/)Robots Exclusion Protocol
- It's a voluntary protocol
- In the past the aim was to enable site operators to signal to Internet crawlers where they can crawl and where they can't

Some AI crawlers did not respect the signals on not to crawl

- Some AI crawlers ignored the signals on what to crawl and what not to
- Some data providers/Publishers/Platforms presumably closed down access to their content so that it would not be used by AI crawlers
- Some publishers don’t know about robots.txt

Restricting the web, restricting access to knowledge

- When restricting access to data on the web, many will be affected including small, open AI systems developers, data scientists and researchers, Open Source developers, crawler foundations such as Common Crawl. Kafka is just one example. But there can be many more.
- So what can different stakeholders and the AI system providers do to prevent this?

Align copyright interest and access to knowledge on the Internet

- Prevent copyright infringement
- Prevent copyright overreach
- Facilitate access to knowledge

Engage with Standards, Rules, Policies AROUND THE WORLD

- There is a technical standard working group on signaling preferences of publishers and data holders to the AI web crawlers. You can monitor their activities and if you get the time join and weigh in: https://datatracker.ietf.org/doc/charter-ietf-aipref/
- Monitor EU AI Act and AI Code of Practice
- Monitor W3C
- Don’t forget other countries and communities (Ongoing legal conversations in South Korea, the UK…)
- Be transparent about how you train your AI
- Be vocal about what you can access and what you can’t access on the Internet
- Be in touch with the publishers
- Come up with best practices for crawling and demonstrate that you follow those

Contact me!

- farzaneh@digitalmedusa.org
- Link to my blog