Category X for AI?

Are LLM / ML projects repeating our licensing mistakes?

Nick Burch
Nick Burch

Currently unemployed
Please hire me!
Nick Burch

@Gagravarr
@nick@social.earth.li
Some resources
ASF 3rd Party License Policy

apache.org/legal/resolved.html

(Includes "Category X")
ASF Generative Tooling Guidance

www.apache.org/legal/generative-tooling.html
I am not speaking on behalf of the ASF here
I am not a lawyer

(Though I do often go drinking with them!)
There is a lot of uncertainty in this area
There will be changes over time
Category X?
License Criteria

The following criteria serve as guidelines for the categories on this page.

1. The license must meet the Open Source Definition.a
2. The license, as applied in practice, must not impose significant restrictions beyond those imposed by the Apache License 2.0.

apache.org/legal/resolved.html#criteria
Category A

Similar in terms to the Apache License 2.0
Category B
Sometimes, with a lot of care
Category X: What can we NOT include in an ASF Project?
Category X includes...

- Not open source
- Restrictions on larger works (e.g., copyleft)
- Incompatible patent clauses
- Field-of-Use restrictions
- Legally unclear what they mean
Category X - JSON License

The Software shall be used for Good, not Evil
Category X - JSON License

The Software shall be used for Good, not Evil

- Who defines good and evil?
- Who gets to decide?
- Is it you? The original author? Someone else?
- Who gets to object?
- Does it vary on jurisdiction?
- Does it vary over time?
Category X - JSON License

Significant amounts of time spent debating this
Eventually... everyone has to use a different version

https://lists.apache.org/thread/t152mmwrt9dvnl5owf54df4qykfozcj0


https://github.com/tdunning/open-json
AI stuff has lots of problems
Licensing is only one!
We will do some of the other ones later....
"Open Source AI"
"Open Source AI" - What's open?

- Source code?
- Training methodology?
- Training data?
- Hyperparameters and weights?
- Eventually model?
Facebook's LLaMA

Non-commercial bespoke

Amongst other things... who defines "non-commercial"?
Facebook's LLaMA and LLaMA 2

These do not meet the OSI definition
OpenAI (eg ChatGPT)

Even less Open...
Some Open Problems
Models leaking training data
What is the license of output?
How do the prompts influence the license of the output?
Are you even allowed to train the models on the data currently used?
Will courts in different countries make different decisions?
Will there be new laws? And how will they differ between countries?
Will WTO-style negotiations be needed? How long will it take?
Some other areas we might be able to learn from / guess from
The Phone Book
OpenStreetMap

Ended up needing their own license

Open Data Commons Open Database License (ODbL)
Radio Stations
"Covers" of songs
Music in Theatre

UK specific, but similar ideas in most places

- Grand Rights - eg musicals, ballets
- Small Rights - Interpolated Music
- Small Rights - Incidental Music
Will we need a new kind of right around AI and training data?
Will we need "Compulsory Licensing" for some or all training data?
Comments / Questions / Ideas?