CENTER FOR DEMOCRACY & TECHNOLOGY

Position paper for IAB Workshop on AI-CONTROL

Nick Doty (CDT)
Eric Null (CDT)
Mallory Knodel (NYU)

Publishers, authors and social media users want some way to prevent their content from being used to train large language models and other generative AI, because it hurts the market for their own work or because it could be invasive or otherwise harmful to them. Companies operating large AI models want to be able to access as much data as possible for training purposes, but may be willing to exclude content that the author doesn't want involved, especially if it avoids either copyright suits or privacy complaints. Researchers want the ability to crawl the web for scientific and public interest purposes without getting caught up in ongoing copyright or AI training disputes. To achieve those ends for the stakeholders concerned, a standards-based solution for controlling use of online content for AI training seems possible and promising.

Community-led, multistakeholder-developed solutions will be more transparent, effective and collaborative, rather than ad hoc or single-vendor proposals. And controls for model training should serve a broader set of stakeholders and interests rather than just the largest players.

I. Consider a broad set of stakeholders and interests

Collaborative solutions can avoid decisions being made through lawsuits and better reflect the range of stakeholder interests. There will of course be competing arguments and interpretations of copyright law and fair use in different jurisdictions and legal actions have become frequent, while community conversation has been sparse. But the experience of robots.txt is a useful one here: that informal standard led to a more useful and stable resolution of the conflicts over copyright and search indexing, and enabled flexibility for web hosts and crawlers. We welcome this IAB-hosted conversation as the topic is overdue, and the ad hoc efforts from particular companies have not meaningfully contributed to an effective solution for users, hosts or crawlers.

The search for collaborative solutions, though, must consider more than just the largest players and more interests than just copyright. The ability to control access to news publications and large archives of commercially developed content is demonstrably of interest to large copyright holders. But there is also demand from individual artists and authors, who might not want their style mimicked or their content used in particular ways. Individuals may not have a copyright interest in all the short statements they post online or all the personal data that may be available about them, but they are still interested in privacy protection and personal autonomy, particularly given the capacity for AI-trained models to repeat information, draw inferences, make predictions or produce true and false claims about them. Furthermore, the individuals who post content online today often do not have control over the hosting website, and they might have different interests than the hosting service in how their content gets sold or used.

CENTER FOR DEMOCRACY & TECHNOLOGY

While we believe the intended focus of controls is on tech companies training large language models, we should be cautious of the potential impacts on other uses of crawling, such as for research purposes. Crawling is an invaluable tool for researchers, regulators and interested individuals for a wide range of purposes, including documenting privacy practices, measuring prices, investigating corporate behavior, enabling portability of personal data or improving natural language processing.

II. Standards can be effective here

Well-established opt-out indicator mechanisms can be effective, if they are standardized and if there are mechanisms for accountability. And voluntary opt-out mechanisms can also enable other kinds of accountability, through legal means, as we saw with robots.txt. (In Field v. Google, for example, the well-known mechanisms of robots.txt and `noarchive` meta tags helped make it possible for the court to resolve the copyright claim while maintaining the capability for indexing web pages for search purposes.)

Robots.txt itself is currently not well-suited to these use cases, however. The problems that hosts are encountering seem well-documented, but include: the ever-changing list of user-agents (or tokens that indicate a particular company's purpose even when they never use that user agent string); and content that is smaller than a particular page or host. But RFC 9309 could either be explicitly extended, or a similar mechanism that can address the purpose as well as the agent, could be standardized.

The particular venue of standardization can vary and we could anticipate coordination between them. We should of course be aware of the Robots Exclusion Protocol publication through the IETF and the Text and Data Mining Reservation Protocol – as well as many other expressions of interest (see, for example, AI & the Web: Understanding and managing the impact of Machine Learning models on the Web) – at the W3C. But most of all, as with most standards efforts, we need implementers who are willing to put forward proposals for wide community review and engagement.

III. Control requires more than just opting out of crawling

In addition to opt-out of crawling for model training, we may need standards for provenance and transparency to answer:

- what content was used to train what models?
- how can people opt-out?
- what generated content was based on what input data?
- what is a crawler doing with the data?

Retrospective controls (to enable recourse or remedy) also seem relevant here, especially for privacy-related concerns. While search engine indexes may be regularly purged and updated, a model can be trained on data once but then expose or make use of that information at any point in the future. Provenance, transparency and deletion features should be provided by any crawler training AI models.