AI Control Position Paper

Prapanch Ramamoorthy

praprama@cisco.com

23 July 2024

Introduction

This position paper has been written for the Internet Architecture Board (IAB) Workshop on AI-Control - https://datatracker.ietf.org/group/aicontrolws/about/. The paper has 2 broad parts – the first part calls out use cases which need to be considered for data/content authors wanting to opt-out of AI crawling. The second part calls out the requirements that any solution we come up with must keep in consideration. All the information below should be consumed keeping in mind existing data and new data that could be created in future.

Use cases for opting out

This section calls out the potential use cases for opting content out of including to train models:

1. The most obvious use case for opting content out of inclusion to train artificial intelligence models is data available on the public web. However, there needs to be more granularity on how we define data in this category so that the solution we come up with applies to all of them:
    a. Web pages and web sites which are owned and operated by the same people who own the content on the web site. Examples of this could be company websites where either the person or the company running the infrastructure running the website also owns the data posted there.
    b. Websites which are hosted on hosting service providers. In this case, there is a clear distinction between who owns and operates the infrastructure and who owns the data posted there. Several cloud providers provide such facilities for users and companies to host their websites and data.
    c. Community hosting systems where users post their data on a website provided by the service provider. The kind of data shared by users could be a wide variety including text, audio, video, images and other file formats. For instance, platforms like Wikipedia or social media platforms like Facebook, X, etc. The data ownership here is largely dictated by the terms and conditions of the service provider.
2. The next category is that content/data with clear copyright and ownership associated. For example, professional movies, music, art, novels/stories, etc. These might be posted and/or made available on the web either legitimately or through illegitimate means (piracy) or could also remain in a form that is not available on the public web.
3. There could also be inadvertent sharing of data which doesn't fall into the piracy/illegitimate category. A good example will be a picture taken of the cover of a book to post a review online on an e-commerce platform.

4. There could also be cases where previously non-digital forms of content/data are now digitized for the purposes of training models where the original owner of the content never intended for their creation to be used for such use cases. For example, there may not be a lot of data available online for training a language model on a local/tribal language but there might be texts, audio and video available in a non digital library that can be converted and made use of for this purpose.

Solutions

While the above list is not exhaustive, it aims to show how the solution to this problem needs to consider a diverse set of use cases. It also shows how a single solution may not work for all use cases and there may need to be a set of different solutions for each of the use cases.

The case of robots.txt

Re-using and extending robots.txt in its current form will probably work well for use cases 1a, 1b and 1c to some extent but not in all scenarios. In its current design, robots.txt can provide access control over specific Uniform Resource Identifiers (URIs). This works well where there is a clear URI/path distinction for each piece of content/data hosted on a website. However, that is not always practically the case or feasible. The growth of single-page applications has meant that there is no need for front end applications to fetch data from the backend every time and that has led to lesser usage of URIs for distinction of various sections of the UI.

In the case of community platforms (use case 1c) and to some extent 1b as well, since ownership of the hosting infrastructure is not with the content owner, they may not be able to control what goes into the robots.txt file.

Another challenge is how to distinctly identify the use case – web crawlers vs data scraper for AI model training. Web crawlers are beneficial because it helps generate reach and grow popularity of content. At the same time, the content owner may not want systems that intend to use the data for training models.

In order to reuse robots.txt, the following **enhancements** will need to be made:

a. Enable federated use cases for use cases 1b and 1c – make it easy for content/data owner to define what goes into robots.txt to protect their data when they may not be able to directly control the infrastructure.
b. Ability to distinguish intent – web crawlers for search engines vs training models.

Beyond robots.txt

How we extend an opt-out mechanism for community platforms like those in the realm of social media. Robots.txt is probably not a scalable solution in such cases. There is a need to come up with newer ways of sending opt-in/opt-out signals. One approach to take is to include metadata along with the data/content in a standardized format while also allowing for various media types like text, audio, video, images, etc. This can be seen as taking the format of robots.txt and including it with individual bits of content/data rather at the level of a website or domain.

From non-digital to digital

There is a clear case of moving from non-digital to digital format with or without the explicit consent of the data owner. This needs novel opt-in/opt-out mechanisms which ties in the physical and digital realm. In certain cases, there are clear legal requirements that need to be adhered to (copyright laws for example) and the solution we come up with needs to extend support to such legal requirements.

This also extends to use cases 2, 3 and 4 – in all these scenarios, ownership is clearly understood in the non-digital realm. However, the same isn’t easily translated to the digital realm. For example, use case 3 of inadvertent data sharing where the cover of the book might contain a piece of art or illustration which is owned by the creator. How we translate this into an opt-out mechanism should be an important consideration going forward.

Data marketplace - the case for opt-in with consent

There could be a situation arising in future where there is a marketplace for content/data that model creators use for training their models. This is going to be especially relevant as more and more AI created content is out there and the need for original creative content grows. This would mean creating a signaling mechanism whereby the content owner is able to provide access to their data to specific people/entities while making sure others are not able to leverage their data. While not an immediate use case, this a potential development in the field of AI that may necessitate better opt-in and opt-out signaling mechanism.