Expressing preferences for data use

Martin Thomson and Lars Eggert, August 2024

Abstract

Machine learning systems depend on access to large amounts of data. A number of existing models that have been created based on data that has been gathered from websites. This data is often obtained without the permission of the people who might have a stake in how that data is used. The debate about this practice is obviously difficult, as it needs to balance the societal benefits that come from the resulting models and the interests of people who hold a stake in the information that is used. Political debate on this topic is very important, but is quite complex. How this debate will ultimately resolve is unclear, but one point that is widely accepted is the need for a means for stakeholders to opt out of having their data ingested for model training. That is, there is value in giving people the means to express their preferences about the use of their data. We identify requirements for a mechanism and conclude that a simple textual signal is most appropriate.

Introduction

Many artists find the advent of generative machine learning systems deeply distressing. Models are trained on their work without their knowledge or assent. These models are then employed to private advantage in a direct threat to the ongoing livelihood of artists. This is just one example of how recent advances in applied AI are impacting individuals and society.

This conversation is obviously highly sensitive and emotionally charged. The strength of opinions about the current problems can end up obscuring the fact that there are even larger issues at stake.

As our society has evolved, we have collectively developed new and more sophisticated ways to codify and process the information that is produced. We collectively produce a vast amount of data that takes many different forms. The ways in which that data might be applied is even more vast.

Machine learning to analyze and synthesize data is a natural progression of that process. The proven ability of machine learning to refine large amounts of unrefined data is a significant and genuine advance.

The applications for machine learning systems are potentially more diverse than the variety of inputs that they use. There are very serious governance questions about how we collectively manage a system that has such diverse inputs and outputs. Concerns are not limited to rights of use, but our goal is not to delve into other fundamental questions, such as the growing resource consumption of such systems.

For usage, balancing the equities of the numerous stakeholders is challenging. The output quality of machine learning systems generally increases with larger and higher-quality training data. This adds to this challenge, because it creates a strong incentive to maximize the size of the training data, which is in direct conflict with attempts to place restrictions on the use of data. Add to that the large numbers of companies and organizations that are competing to deliver the "best" models, homed in different countries with different laws and regulations for data handling, and the governance landscape becomes truly difficult.

There are many stakeholders when it comes to the use of data. Some of those stakeholders might be granted some amount of legal rights to determine how data is used. Presently, the legal basis for rights is typically privacy or copyright, but there is no reason that these are the only ways that rights might be conferred.

Thankfully, these difficult issues are not the topic of technical standards-making. We have a much narrower task: to take the resolution of that process and codify it so that it might be understood and acted upon. Or, if we are to be proactive, to offer technical solutions that facilitate the choice of more constructive outcomes.

Technical standards bodies can — and should — define mechanisms that make it easy to express assertions of rights regarding system inputs. Giving stakeholders the means to have a say might be key to enabling some policy approaches.

This document looks at the components that are already in place and suggests several lines of investigation.

What preferences might say about uses

As a matter of context, machine learning systems require input data in two forms:

- Labeled inputs for training.
- Unlabeled inputs for their use.
- Inputs that a model draws on during use.

This separation of training and usage is critical. The inputs to a system that uses a trained model to perform its function is not special. We are therefore primarily concerned about the inputs to the training process. More complex models, like those that use retrieval-augmented generation, also draw on web resources during their use.

The important consideration here is that an aggregate of *all inputs* are drawn upon by a model when it is used. This is especially evident in cases where generative models reproduce elements of their training dataset, but this is not specific to generative systems.

The novel question for standardization is how a stakeholder might express their rights. These rights might need to be expressed to different degrees of specificity. What rights someone can effectively express ultimately needs to be decided as a matter of policy. The question for policymakers is how to interpret these statements of preference.

Expected usage preferences

The outcome of that process ultimately needs to be codified for it to be useful. For this exercise, we'll consider a few of the proposals and possibilities.

Declining any use for training — other than that for which a stakeholder is not permitted to decline for public interest or other reasons — is a simple expression that is likely necessary to be able to state clearly and efficiently.

Developing a more comprehensive taxonomy, such that more specific statements could be made, requires creating a language for identifying uses. General statements about usage is something that copyright law often contends with; copyright experts might be required to advise on what a sensible taxonomy of purposes could be. These might be used to prohibit or permit entire classes of use. Examples of general classes of use might include commercial, research, or private use (Appendix B of this paper lists some of

these). Alternatively, the ai.txt proposal from Spawning breaks down use by type of medium, presumably with an assumption that the usage is primarily generative.

Stakeholders might also wish to grant or deny access to specific entities, though this is obviously less clearly useful. Per-entity grants or denials are always possible to express in other ways that can override more general mechanisms.

Authenticity requirements

Various actors might consider the authenticity of expressed preferences to be important. However, this introduces some significant challenges:

- Restrictions on use might be falsely applied to content, which unnecessarily excludes that data from usage that should be authorized.
- Unauthorized removal or weakening could deny stakeholders the ability to state their preferences about use.

Content that comes with restrictive usage preferences does not necessarily preclude the use of that data. Laws might provide a basis for use of content for certain purposes, despite the stated preference of stakeholders. Content consumers can also engage with stakeholders via other means to obtain permission to use data.

The following analysis concludes that the use of cryptography, like digital signatures or watermarking, is unlikely to improve outcomes.

Authorizing preferences

Those seeking to use information for model training might wish to know whether the preferences they receive are made by stakeholders whose preferences they need to respect. An automated means of identifying authorized stakeholders could be valuable for this purpose.

Any method for establishing authentication for stakeholders would need to clear a number of very significant technical challenges, given the global scope of the systems in question. Even identifying who has expressed a given preference is not simple. Using existing identifiers — like legal names, account names, private enterprise numbers, or domain names — is less likely to be as directly actionable than identification based on roles like content author or content distributor.

Either approach creates a challenge in terms of validating any claim about identity. Systems have been proposed that claim to support content provenance, like C2PA. These systems use digital signatures from a small set of trusted entities. However, these systems also place tight constraints on content production and editing processes, which narrow their applicability significantly. These approaches might identify content authors, but are unlikely to be broadly applicable across the diverse forms of content production that might be used in model training. It would not be reasonable to require the use of these systems, especially for existing content.

We conclude that any attempt to include authorization is challenging and therefore not something to prioritize given the exigent need for solutions.

Weakening or removal of preferences

Stakeholders that express preferences might wish to have some guarantee that any restrictions on use they place are respected. That is, these preferences persist no matter how information is distributed, including content editing, transcoding, and other common tasks.

The situation today with copyright is it is difficult to ensure that content is not copied without permission. Copies are made without the markings that might be used to identify those who hold rights. There is a genuine risk that copies made in this way will be stripped of any markings about preferences, with those copies being inadvertently integrated into model training.

Elision of markings is particularly serious when there is a presumptive permission to use content, as that places the onus on stakeholders to ensure that their preferences are clearly understood. A legal requirement to obtain a license for use – that is, a requirement that stakeholder expressly permit use – would make the problem less urgent.

For copyright violations, it is often the content itself that is the means by which the violation is detected. Unlike simple copyright violations, the unwanted inclusion of content in a model is likely to be difficult for stakeholders to detect or prove. Transparency about training data sets and other systems for accountability might help, but these topics are out of scope for this discussion.

Watermarking is an approach that might offer some ability to embed information in content. The use of content watermarking is unproven. Any widely available means of

reading watermarks also offers adversaries the means by which to remove the watermark. Controlling access to the system for checking watermarks is likely infeasible, particularly if there is a desire to offer stakeholders a choice of marking scheme.

Any system of marking also needs to be readily applied at scale to existing content. Though any new system might be applied to new content, there needs to be a scalable method that can be used for existing content. Cryptographic methods, particularly those that rely on modifying content, are unlikely to be feasible for this purpose.

We acknowledge the challenges that content providers have in accurately representing the preferences of content creators. This is already a massive challenge for copyright reasons, as copied content is frequently uploaded to the web. However, we do not see authentication to be a tractable approach.

What mechanisms could be used

Today, there are two basic approaches in use for expressing usage preferences:

1. Content is directly annotated with usage rules (or licenses).
2. The protocols or systems for obtaining content provide a means to obtain licenses.

Both of these neatly address the question of identifying the information that is covered. Attaching to content or its delivery avoids some of the more challenging issues that might arise from less direct methods that might require the definition of content identification and addressing.

Using the content delivery mechanism to deliver usage preferences offers scaling advantages. Information takes many forms, which could mean that annotations might need to be developed for each data format that might be used. There are fewer delivery mechanisms to consider in the design and far fewer artifacts to consider in its application.

On the other hand, annotating content might be useful in providing content producers (such as authors and artists) with a means to mark content at its point of production in ways that will survive various forms of distribution and redistribution.

Another special challenge for content marking schemes is the potential for markings to be removed during copying or distribution. The use of existing metadata might mean

that this is only a problem when stripped deliberately, but where there is presently no expectation that markings exist, the absence of markings is unlikely to be noted.

The preferences and incentives of content producers could differ from those of distribution channels. Each party has a stake in terms of expressing preferences, so allowing multiple entities to separately indicate their preferences could be a good outcome. People seeking to use information would therefore need to limit that usage so that it respects the preferences of all stakeholders.

robots.txt

The robots.txt file is currently the most widely recognized method of permitting stakeholders to express preferences to automated crawlers of HTTP sites.

Some companies that engage in crawling have chosen to self-identify using the HTTP User-Agent header field. Servers can therefore selectively block those crawlers based on that self-identification.

More commonly, servers make content available, but express their preferences through exclusion rules in robots.txt. That is, rather than performing access control, the server trusts the crawler to apply the policy.

The use of robots.txt was originally defined for use in Web search indexing. This is a specialized purpose, where there is mutual advantage to making information available: the site might become more visible in searches and the crawler gains a more comprehensive index. Use of the data obtained by a crawler for model training or other uses not aligned with the inherent incentive structure was not considered in the original design of this mechanism.

The robots.txt format does not support any way to express preferences that relate to other purposes. A website can only use the identity of the crawler or the identity of resources as a distinguisher.

This creates several obvious shortcomings. A site that values search indexing but not model training cannot express that intent using robots.txt. Websites can seek to identify all of the active crawlers, but this will always miss new crawlers. This is already a situation that is unmanageable for many.

The use of a mechanism like robots.txt that covers model training – or the definition of extensions to robots.txt for model training – is possible. This necessarily needs to consider having clearly articulated rules about preferences that includes more than the identity of the crawler. At a minimum, usage preferences need to be able to express constraints on purpose.

Content markings

Handling content markings is likely more diverse than looking at protocols.

The IETF has some experience with the marking information with conditions on use. The GEOPRIV effort aimed to annotate data with policy information. The information would therefore always be coupled with rules about its use. From a procedural perspective, this coupling ensures that rules are never absent; or at least their absence is notable.

The GEOPRIV effort produced a number of RFCs but was ultimately unsuccessful. This might be seen as the consequence of building a technical mechanism absent the work to establish a legal basis for setting and respecting rules.

Many file formats provide a way to provide copyright and licensing information in metadata. A simple expression of policy regarding use by machine learning models might be included in a similar way. A structured form that could be applied consistently across multiple formats is most likely to be acceptable to those seeking to use content.

Copyright licensing has a number of successful projects that demonstrate the utility of simplified expressions. SPDX identifiers unambiguously express which license applies to a source code file; Creative Commons has also created shorthands for identifying their licenses.

Defining that format is something that a technical standards organization like the IETF could accomplish. A simple, concise textual expression seems mostly likely to be usable across the diverse range of formats that might need to be marked.

There are many file formats, but it is possible that by providing the capability for the most common formats, a large amount of content will be covered. The pattern established for those formats can then be propagated to other formats.

Protocol extensions

Adding markers for protocols might offer a means to apply markings through the distribution medium. This might be helpful to address formats that lack the means to carry metadata (such as plain text formats) or where standards for metadata in that format have not yet been set.

As an example, a header field might be defined to express preferences for any content retrieved using HTTP.

Alternatives to standardization

Standardization gives stakeholders a predictable means of expressing their preferences and those training models greater certainty about the status of the information they consume. However, in the absence of well-recognized standards, many approaches have emerged.

De facto standards

Many AI crawlers respect robots.txt, despite its shortcomings. There are also a number of opt-out mechanisms that are essentially proprietary. Support for these formats is uneven, forcing stakeholders to express their preferences in multiple ways in order to be effective.

It is possible that one or more of these methods will eventually become a de facto standard. This carries a risk that market power, not suitability, will determine what method is successful. Even if that outcome is broadly acceptable, the process for managing change and evolution would remain unclear if that happened.

Faking generated content

A dystopian vision of the future of the Internet is one where the machines are all talking to each other. They are all trained on the output of other machines. What happens generates a lot of waste heat, but no value for people.

One potential defense against this outcome is to avoid training models on the output of other models. So, while we consider the process of watermarking content to be ultimately ineffective for the purposes of establishing provenance, unauthenticated markings of that nature might still help avoid the ingestion of the output of other models.

Such markings could then become a crude, but powerful means of expressing a desire not to contribute to model training.

Poisoning the well

AI poisoning tools are a more destructive means by which people might seek to assert their rights. Poisoning tools modify content in subtle ways that are designed to exploit identified weaknesses in model training systems. A small number of poisoned inputs can ruin an entire model for certain purposes.

Use of poisoning has the effect of spoiling the models that are produced from poisoned inputs. This is an attack that does not require additional incentive to mount. Countermeasures could be expensive if no better alternative than poisoning is offered to those who do not want content to be used for training.

Conclusion

Providing a means for people to request that data not be integrated into model training is urgently needed. This can be done no matter how the difficult and important policy debates eventually settle. Having a technical mechanism in place is possible prior to the resolution of that debate; a standardized mechanism might also contribute to advancing the debate by opening up new options.

Currently, the needs of those seeking to opt out of participation in arbitrary models is poor. Though there is obviously a desire for mechanisms with stronger security properties, we cannot see how practical solutions could be attainable in any reasonable time frame.

We suggest defining a simple, textual assertion that can be conveyed by multiple means, both at protocol and content layers. This approach provides people a cooperative option, rather than the strictly adversarial options that are presently available, like poisoning.