DoNotTrain: A Metadata Standard for Indicating Consent for Machine Learning

Daphne Ippolito [*1] **Yun William Yu** [*2]

Abstract

Modern machine learning for Generative AI are dependent on large-scale scrapes of the internet. There are currently few mechanisms for well-intentioned ML practitioners to pre-emptively exclude data that website owners and content creators do not want generative models trained on. We propose two mechanisms to address this issue. First, building off of the existing `robots.txt` protocol, we recommend a `learners.txt` protocol which enables a website owner to specify which pages of their website are appropriate for ML models to train on. Second, we propose a standardized tag which can be added to the metadata of image files to indicate that they should not be trained on.

1. Introduction

As Generative AI moves from the realm of academic research to practical deployment, the issue of whether machine learning models (and their outputs) are derivative works of the underlying training data is a matter of growing importance (Gervais, 2021). Until either the courts rule or new law is crafted clarifying the matter, there will always be a cloud hanging over models trained on non-public domain data. Even then, however, different jurisdictions may not share the same view on machine learning models as derivative works, so waiting for clarification is no panacea.

In the absence of legal clarity, sociotechnical mechanisms can be employed. If website creators and content creators agree on a standardized way to annotate content which should not be used to train ML models, well-intentioned ML practitioners can omit this content from their training data. Our proposed standard, DoNotTrain, takes inspiration from the `robots.txt` standard, which is a way for site owners to signal whether or not webcrawlers should access particular webpages. Though the standard lacks enforcement mechanisms, it is generally respected by the major search engines.

We propose providing a standardized mechanism for explicit consent to train at both the level of entire websites and individual files. Such a mechanism would empower people with more control over how their creations are used. Just as importantly, there are incentives for the the trainers of large ML models to respect such a standard. Not only would it provide some measure of assurance that models are on firm moral footing, but the proposed mechanism can also address the ouroboric challenge of training datasets for new models being corrupted by AI-generated content from older ones (Hataya et al., 2022).

2. Background

Introduced in 1994 by Martijn Koster at Nexor and codified as the IETF Robots Exclusion Standard in 2022, `robots.txt` files live on the root directories of websites and are intended to tell automatic agents, i.e., "robots," which files in a website should and should not be accessed (Koster et al., 2022). The most common automatic agents are the web crawlers used by search engines like Google or Bing to build an index of the internet. A `robots.txt` file allows a website owner to specify which of their files should be visible on search results and prevent their server from getting overloaded by requests from crawlers. The file is both a non-binding request and friendly advice; agents are not bound to abide by the file, and indeed, malicious agents may purposely use a `robots.txt` to discover additional files. However, sites often use `robots.txt` to warn crawlers of "crawler traps" like calendars, which have an infinite number of programmatically generated links, where webcrawlers may get stuck, but without meaningful content.

The original idea of `robots.txt` was to provide guidance for all automatic agents, but over time, website owners have tailored their `robots.txt` files to optimize search engine rankings. This has caused non-search engine agents, such as the Internet Archive, to choose to ignore `robots.txt`

[*]Equal contribution [1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA [2]Department of Mathematics, University of Toronto, Toronto, ON, Canada. Correspondence to: Daphne Ippolito <daphnei@cmu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

files (Graham, 2017). The existing `robots.txt` standard does not convey any of these nuances.

Of course, we are not the first to note that it would be helpful to have machine-readable annotations of other allowed activities. One failed attempt at extending `robots.txt` was the Automated Content Access Protocol (ACAP), which was designed by a coalition of publishers who tried to implement much more fine-grained permissions than a simple "crawl" or "don't crawl" (Paul, 2008). Though there are many reasons for its failure to be adopted, we think one of the major flaws was the complexity of implementing fine-grained permissions. Unlike the `robots.txt` file, which can be entirely managed by the webcrawler, implementing more fine-grained permissions requires the cooperation of both downstream applications and potential end-users.

Another example to learn from is the fate of the DNT (Do Not Track) and GPC (Global Privacy Control) browser headers (Brandom, 2021). The DNT was a browser setting end users could choose to affirmatively set that would tell a website they do not want to be tracked. However, nearly no websites respected the DNT header. Two of the reasons for DNT's failure were (1) a lack of incentives for websites (legal or otherwise) to respect it, and (2) what it meant to "be tracked" was not clearly defined. DNT's putative successor, the GPC, is currently trying to solve both problems through clear legal mandates. Although still early, GPC has seen some measured success (Whitney, 2022).

There are a couple of lessons to be learned here. First, any proposal should be as simple as possible, avoiding feature creep and complexity. Additionally, it is crucial that all interested parties have incentives to respect the standard.

3. Proposal

Extending `robots.txt`

An extension to the `robots.txt` standard would support specifying different kinds of allowable uses, rather than a simple binary rule of do or do not read. But, we must take care to avoid ambiguity or complexity. Here, we specifically address the case of helping content creators and web site owners tell ML dataset creators which files they do not want to be used to train machine learning systems.

We propose copying the `robots.txt` standard to signal which webpages are encouraged as training data. Our proposal is simple: website owners may add a second file, `learners.txt`, to the root directory of their website, following the exact same syntactical standards as `robots.txt`. Alternately, HTTP/HTML metadata could be used on a per-file basis (e.g. `X-LEARNERS-TAG: noindex` HTTP response header). Just like `robots.txt`, `learners.txt` would be a request which polite actors agree to abide by, with no binding legal force.

Most developers of state-of-the-art generative AI models take the Common Crawl, a publicly available crawl of the internet, as their starting point, and filter and process it into a training dataset. The Common Crawl currently respects robots.txt. For the `learners.txt` protocol to be useful, we would work with the Common Crawl to add `learners.txt` metadata to their store.

Adding Image Metadata

Ideally anyone who creates content should be able to specify whether or not they want their content trained on. However, adding metadata annotations is easier for some types of data than others. Text is perhaps the hardest data to consistently add annotations to. Outside of specialized use cases (such as source code on Github with accompanying licenses), people rarely upload .txt or .doc files to the internet; instead they post micro-blogs on Twitter, product reviews on Amazon, etc. This text ends up embedded within larger web pages, without any standardized, attached metadata. It is possible to include canary strings within the text to let dataset creators know to skip it, such as is done by the Google BIG-Bench project (Srivastava et al., 2022), but this approach has its own set of issues—not only are canary strings easily stripped, they can also often be added by third parties to a webpage that otherwise is available for training, through e.g. a comments form. On the other hand, images are almost always uploaded as self-contained files, and there are just a handful of common formats. All of the standard formats for images on the web (jpg, png, webp, etc.) support textual metadata.

We propose that image creators add either the string `DO_TRAIN` or `NO_TRAIN` to the metadata of images they want to instruct machine learning practitioners to respectively train or avoid training on, where the `NO_TRAIN` tag overrides a `DO_TRAIN` tag. Importantly, we make no assumption about the expected behavior in the absence of either tag.

Image metadata is easy to modify in commercial image editing software such as Adobe Lightroom, and it would be straightforward to build a lightweight application to add/remove the `NO_TRAIN` tag from image metadata. Widely used image generation AI, such as DALL-E[1] or Midjourney[2], could add the `NO_TRAIN` tag to all generated images so as to avoid the generated images being included in future datasets. Like with `learners.txt`, malicious agents can choose to ignore the tag, or they can strip an image of the tag then re-upload it, though the latter may fall under the purview of existing copyright law.

[1] https://openai.com/product/dall-e-2
[2] https://www.midjourney.com/

4. Discussion

As we have alluded to throughout the text, the success of DoNotTrain depends entirely on the buy-in of relevant parties. Indeed, following the example of the Robots Exclusion Standard, we deliberately choose to not design it as legally/contractually binding. In this section, we discuss the incentives for content creators and model trainers to adopt DoNotTrain, as well as how adoption might happen.

Content Creator Incentives
The most obvious incentive for a content creator is greater control over the way their creations are used downstream. We expect that the primary use of DoNotTrain, at least initially, will be creators adding the `NO_TRAIN` tag to request that their content not be trained on. Although lacking any legal force, these tags will allow content creators to explicitly express displeasure in the court of public opinion should it come to light that a model trainer has ignored their polite requests. Furthermore, a restrictive license on content could be paired with a request not to train in order to prevent an oblivious trainer from accidentally training on discouraged data.

Model Trainer Incentives
From the model trainer's perspective, potential training data can be partitioned into three categories: (1) content that neither discourages nor welcomes training, (2) content that discourages training, and (3) content that welcomes training. At the moment, the only standardized information trainers have to determine whether or not it is legally permissible to train on content is the applicable license; for some licenses, like the Creative Commons CC0, this determination is likely easier than for other licenses, for which legality is likely to depend on whether or not models are considered derivative works. The DoNotTrain Standard does not resolve any legal issues, but it does allow content creators to tag their works as falling into either categories 2 or 3.

A very cautious trainer (say, a risk-averse large company) might still eschew the DoNotTrain standard and only train on content under vetted licenses. However, such a trainer would be at a competitive disadvantage because models generally perform better when exposed to a larger amount of training data. A slightly less cautious trainer could then expand their training data to include content which is marked for training by DoNotTrain but may not have a readily available or permissive license. Although respecting DoNotTrain would not eliminate legal risk, it would allow the trainer to claim that they train only on content that welcomes training, which may be advantageous from a PR perspective. Furthermore, we hope that content creators will only add a `DO_TRAIN` tag to content appropriately licensed to legally permit training, though this cannot be guaranteed. An even less cautious trainer (perhaps a small startup) might choose to train on everything that doesn't explicitly prohibit training. This would expose them to much greater legal risk should models be considered derivative works, but it would also of course improve their models.

One might reasonably wonder what incentive there is for a risk-tolerant trainer to choose to respect a the DoNotTrain protocol when it has no legal force. However, if generative models start tagging their outputs as content that should not be trained on, then even a risk-tolerant trainer may choose to respect those tags to prevent training on ML-generated content, which has been shown to decrease model quality (Hataya et al., 2022). Additionally, even a risk-tolerant trainer may want to prevent the PR problems that would arise from it becoming known that they deliberately do not respect content creator wishes, whether or not it is legally permissible.

Adoption pathway
In this proposal, we have avoided assigning intent to the lack of a DoNotTrain tag. This is deliberate, as there are reams of content already on the Internet, and it would be presumptuous for us to assume what their creators would want. On the other hand, we imagine that model trainers would very likely prefer to treat untagged content as welcoming training, as more training data generally results in better models, with the exception of model-generated content.

Thus, we envision that the most likely initial adoption of a DoNotTrain tag is to mark AI-generated content as not good for training (whether via `learners.txt` or image metadata). Once a major industry player adopts the standard, that could spur content creators who do not want their creations trained upon to explicitly label their data as well.

Sometime in the future, the legal issues around machine learning models and their output will be (more) clarified. We hope that in that world, expressions of creator intent with respect to training will still be needed and respected.

5. Conclusion

Although simple, our proposal gives a mechanism for content creators and website operators to signal intentions with respect to ML training. It deliberately follows the `robots.txt` standard in that it serves as a helpful guidance and a friendly request, without any binding force. However, trainers of large ML models are incentivized to follow it because the same tag is used for warning against both bad training data (such as the output of a generative AI model) and data whose copyright holders do not want to be trained on. Our proposal does not solve the consent to train problem at large, but we hope it is simple and useful enough to be implemented in practice.

References

Brandom, R. Global privacy control wants to succeed where do not track failed. *The Verge*, 2021. URL https://www.theverge.com/2021/1/28/22252935/global-privacy-control-personal-data-tracking-ccpa-cpra-gdpr-duckduckgo.

Gervais, D. J. Ai derivatives: The application to the derivative work right to literary and artistic productions of ai machines. *Seton Hall L. Rev.*, 52:1111, 2021.

Graham, M. Robots.txt meant for search engines don't work well for web archives. *Internet Archive Blogs*, 2017. URL https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/.

Hataya, R., Bao, H., and Arai, H. Will large-scale generative models corrupt future datasets? *arXiv preprint arXiv:2211.08095*, 2022.

Koster, M., Illyes, G., Zeller, H., and Sassman, L. Rfc 9309 robots exclusion protocol. *Internet Engineering Task Force*, 2022. URL https://www.rfc-editor.org/rfc/rfc9309.html.

Paul, R. A skeptical look at the automated content access protocol. *Ars Technica*, 2008. URL https://arstechnica.com/information-technology/2008/01/skeptical-look-at-acap/.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Whitney, L. Cosmetics giant sephora first to be fined for violating california's consumer privacy act. *TechRepublic*, 2022. URL https://www.techrepublic.com/article/sephora-fined-violating-ccpa/.