OPEN
_FUTURE

CONSIDERATIONS FOR OPT-OUT COMPLIANCE POLICIES BY AI MODEL DEVELOPERS

OPEN FUTURE POLICY BRIEF #6

Author: Paul Keller

16 MAY 2024

THE ISSUE IN BRIEF

Article 53(1c) of the AI Act requires "providers of general-purpose AI models" to "put in place a policy to comply with Union copyright law, and in particular to identify and comply with, including through state of the art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790." This paper explores what such compliance policies could look like in practice and what technical standards and services are available to implement rightholder opt-outs in a way that is effective, scalable, and addresses the needs of diverse groups of rightholders and AI model developers.

The paper recognizes that the requirements contained in the AI Act apply to individual providers of general-purpose AI models and that it is thus likely that there will be a variety of compliance policies. At the same time, it is also clear[1] that there are incentives for both rightholders and AI model trainers to standardize key elements of such policies. The ambition for policies to substantially converge is also reflected in the fact that Article 56 of the AI Act makes it clear that the compliance policies will be part of codes of practice that are to be established by providers of general-purpose AI together with the AI Office and other relevant stakeholders within nine months from the date of entry into force of the AI Act.

This paper is a first attempt to outline what a generic compliance approach could look like while being mindful of the fact that actual implementations will probably differ due to the specific operational requirements of different providers of general-purpose AI models. It further assumes that there will be multiple ways in which rightholders in different types of creative works wish to express opt-outs. As of the coming into force of Art. 53(1c), it is unlikely (and possibly undesirable) that there will be a single one-size-fits-all solution. Policy-makers and stakeholders should plan for these obligations to be flexible enough to evolve together with standards and best practices as they develop.

Finally, this paper focuses on machine-readable opt-outs. While there is some debate about the definition of machine-readable, for the purposes of this paper, we exclude attempts to opt out that are not explicitly stated to be machine-readable, such as blanket opt-outs in press releases, statements on websites, and reservations of rights in website terms and conditions.

[1] For background see our previous policy brief on right holder opt-outs: Defining best practices for opting out of ML training, Open Future, September 2023.

POLICY ELEMENTS {}

In its most simple form, a framework for developing a policy to comply with the machine-readable rights reservations required by Article 4(3) of the Copyright in the Digital Single Markets (CDSM) directive might take the following form:

> If you (rightholder) tell us (model trainer) which of your works you want to opt out, we will not use them for model training.

In practice, such a policy will require more precision with regard to four different aspects: the *identifiers* for works, the *vocabulary* for opting out, the *infrastructure* used to communicate and respect opt-outs, and the *effect of the opt-out* once it has been recorded. This leads us to the following version of the logic expressed in the statement above:

> If you tell us what **{identifier}** you want to opt out from which uses **{vocabulary}** via these means **{infrastructure}** then we will do this **{effect of opt-out}**.

The remainder of this document will look at each of these four areas in need of further specification before concluding by identifying potential next steps.

{IDENTIFIERS}

When it comes to identifying the works/content that are opted-out in a machine-readable way, there are two different approaches. The first approach consists of location-based (domain- or URL-based) identifiers, and the second approach consists of unit-based identifiers (applying to individual copyrighted works or media files). Both of these approaches have their own advantages and disadvantages that make each of them more suitable for different scenarios and types of works to be opted out. As of today, location-based mechanisms are more widely used, but over time most policies in order to comply with opt-outs will likely need to account for both types of identifiers.

Unit-based versus location-based identifiers

In the context of text and data mining (TDM) opt-outs, *location-based* refers to any strategy that enables domain owners, organizations, network administrators, or rightholders who upload works to websites they control to set broad, overarching policies that apply to all works hosted on a domain or URL or a large subset of its content. This approach is designed to offer a

streamlined, effective way to manage and express rights reservations at a high level, affecting all applicable content without the need to address each item individually2. Examples of location-based approaches that are relevant to the discussion about TDM opt-outs are `robots.txt`, `ai.txt`, the TDM Reservation protocol (TDMRep), DeviantArt's `noai` meta-tags, domain registration in do-not-train registries, and the use of HTTP headers. Of these approaches, `robots.txt` is currently the most widely used.

On the other hand, *unit-based* refers to the tools and standards designed to manage opt-outs at the individual piece of content or data unit level. Unlike broader, *location-based* strategies, this approach focuses on granular control, allowing creators and other rights holders to specify permissions, restrictions, and conditions for the use of specific media files within larger datasets or collections. It also allows the rightholders to specify those rights regardless of where the media files are hosted or even when they are stored offline. There is a wide variety of approaches that range from standards for embedding metadata in media files (e.g., the Coalition for Content Provenance and Authenticity (C2PA)) to content-derived identifiers used to soft bind metadata to content (e.g., the International Standard Content Code (ISCC)), to watermarking and tools specifically designed to register right holder opt-outs in the context of AI training (e.g., haveibeentrained.com).

For those rightholders who manage their own domains or sites, *location-based* approaches to identifying content are simpler and more cost-effective to implement. Some *location-based* identifiers (especially `robots.txt`) also have the advantage of being widely used in the online environment to control indexing and scraping of content and, more recently, to opt out from data collection for AI training purposes. Their main drawback is the resulting lack of granularity[3] and the fact that such identifiers can only be set by entities that have control over the domains or URLs in question, which may or may not be the actual rightholders. In addition, `robots.txt` only protects the domain from crawlers on the site where it is hosted. If the data is linked or embedded elsewhere on the web with sites that do not have corresponding rules set, it will still be included in datasets compiled by crawlers that are excluded via `robots.txt`.[4]

Conversely, *unit-based* approaches to identifying content work on a much more granular level (media files) that more closely resembles the concept of works that is the object of TDM opt-outs under Article 4 of the CDSM directive. *Unit-based* opt-outs can also be set at the early stages of distribution chains, ensuring that such opt-outs can be set by rightholders themselves without having to rely on third-party platforms. Conceptually, *unit-based* opt-outs can also operate across platforms/contexts, although some of them are vulnerable to metadata-stripping

---

[2] Many of these approaches described in this section, including `robots.txt`, can also be applied to individual files based on their unique location (URL).

[3] Unless they are applied on a per-file or per-page level.

[4] This latter issue is addressed by Spawnings ai.txt implementation that relies on checking the permissions encoded in the ai.txt files at the time of training (which often takes place significantly later than the crawling of websites to build training data sets). Similarly, AI developers can also undertake to re-check robots.txt exclusions at the time of training to note changes that have taken place after the crawl date into account.

that can undermine their effectiveness. One key limitation of *unit-based* approaches is that they may have drawbacks for dynamic web page-based works, such as news websites or software code, that often undergo significant changes within short time frames as well as streamed and live performances.

In the context of an effective opt-out compliance policy, both approaches will need to be considered, recognizing that some location-based identifiers, such as robots.txt, offer a solution that can be implemented more readily. *Location-based* identification strategies seem especially relevant for textual works natively published online (a major training source for large language models) where it is impractical to employ unit-based identification approaches. *Unit-based* approaches are much better suited for works that mainly circulate as independent media files (images, audio, video but also text published as PDFs, ebooks and other stand-alone file formats).

Convergence on identifiers

There are strong incentives for both rightholders and AI model trainers to standardize a relatively small number of identifiers to effectively communicate opt-outs. From the perspective of rightholders, a small number of standardized identifiers will increase legal certainty and streamline opt-out processes even though technically, rightholders are free to use any machine-readable way of identifying works for which they wish to opt-out. From the perspective of AI model trainers, a small number of standardized identifiers will reduce implementation complexity and, thus, cost. **To achieve this, AI model trainers should make sure that they support standards that address the needs of rightholders. Ideally, a situation in which there is a limited number of standardized identifiers that can deal with all kinds of works and distribution strategies should be desirable from the perspective of both rightholders and AI model trainers**. In addition, there should be room for more bespoke systems that address concerns related to specific types of content used as training data.

{VOCABULARY}

In addition to the identification of the works that are being opted-out, it is also necessary that machine-readable rights reservations contain a clear description of what types of uses rightholders are opting out from. In other words, there needs to be a machine-readable vocabulary that describes the intent of the opt-out.

Article 4(3) of the CDSM directive provides that rightholders have the ability to reserve the rights that fall under the scope of the text and data mining exception in Article 4(1) of the directive. Article 2(2) of the directive defines text and data mining as

> *any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations;*

As we have argued before[5], this definition of text and data mining includes, in its scope, the training of AI models. This interpretation has since been confirmed by the EU legislator in the AI Act[6]. While the rights reservation in Article 4(3) does provide the legal basis for opting out from AI training, it is important to note that in addition to AI training, the definition of TDM from the CDSM directive includes many other forms of computational analysis of copyrighted works. This means that opting out from TDM in its entirety will likely lead to unintended consequences, as such an opt-out is much wider in scope and will likely apply to many other forms of computational analysis that may be considered desirable by rightholders.

In fact, many rightholders have made it clear that while they want to be able to opt-out of the training of generative AI models, they do not want to opt-out from their works being used by other forms of AI, especially when such technologies are used for search and other types of discovery such as algorithmic timelines. At the same time, a full opt-out from TDM for content that is lawfully available online will also be extremely limiting for all kinds of online intermediaries, regardless if they employ AI-based tools or not[7]. **As a result, it seems desirable to develop a vocabulary (or taxonomy) of uses for rightholders to opt out that is more granular than the binary approach of either opting out of all TDM or declaring no opt-out**.

Article 53 of the AI Act applies only to "providers of general-purpose AI models," which implies that the policies that model developers have to put in place can be limited to the uses of works for the purpose of training general-purpose AI (GPAI) models. Article 3(68) of the AI Act defines a general-purpose AI model as "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks." This definition includes the current generation of generative AI models but likely also includes a number of other models that do not produce output that resembles copyrighted works and may, therefore, be of less concern to rightholders. Nevertheless, AI model trainers depend on an opt-out implementation that is also practicable in the context of those general-purpose AI models that have thus received less public attention to date.

Recital 105 of the AI Act, which provides the context for the compliance policy requirement in Article 53(1c), specifically concerns itself with "large generative models, capable of generating text, images, and other content, [that] present unique innovation opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed." Against this backdrop, it seems sensible that a vocabulary for machine-readable opt-outs should include an option for opting out from the use of works for the training of this class of general-purpose AI models, which is generally referred to as generative AI models. In other words, in addition to a `no-tdm` option the vocabulary should also

---

[5] See Protecting Creatives or Impeding Progress? Machine learning and the EU copyright framework, Open Future, February 2023.

[6] See Recital 105 and Article 53(1c) of the AI Act.

[7] In this context, it is important to stress that even if there has been a full reservation of the TDM rights, other exceptions and limitations (including the transient copying exception from Article 5(1) of the InfoSoc directive) may apply to uses that also meet the definition of text and data mining.

offer a `no-generative-ai` option that allows rightholders to specifically opt-out from the use of their works for the training of AI models, while still allowing the use of their work in text and data mining for other purposes such as search, including when search is supported by AI[8].

Currently, only the C2PA approach explicitly supports such a more granular vocabulary (the vocabulary distinguishes between uses of the work for `data_mining`, `ai_training`, `ai_generative_training`, and `ai_inference`). Other approaches (such as Spawning's products and the DeviantArt `no-ai` meta tag) are specifically targeted at (generative) AI training, while others (such as TDMRep) are explicitly aimed at the full spectrum of text and data mining. Finally, the various opt-out solutions based on robots.txt do not target a specific class of uses but specific user-agents, although several companies, including model training companies (such as Google, Microsoft, and OpenAI), have started to distinguish crawlers that serve different purposes by assigning them different, task-specific user-agents. However, no naming conventions for user-agents have emerged that would simplify the task of declaring opt-outs through `robots.txt` for rightholders.

A common vocabulary

All of this points to the need for a common vocabulary for machine-readable opt-outs. Defining such a vocabulary should be a relatively straightforward step that equally benefits rightholders and AI training companies. **Based on the considerations above, it seems desirable that any compliance policies should be based on a vocabulary that distinguishes between a full TDM opt-out (`no-tdm`) and an opt-out from generative AI training that applies to the use of works for training the subset of AI models described in recital 105 of the AI act (`no-generative-ai`)**.

Since the `no-generative-ai` is a more specific version of the `no-tdm` opt-out, for the purpose of training generative AI models, either of these two would signal an opt-out to the model trainers. It seems clear that in most circumstances, the more targeted `no-generative-ai` opt-out is more aligned with the interests of rightholders and model trainers and would likely become the default type of opt-out in the context of the compliance policies that have to be put in place by developers of general-purpose AI models[9].

In order to streamline the various methods for right holders to opt out, such a vocabulary would need to be supported by all identification approaches regardless of whether they are location-based or unit-based. While this will be relatively straightforward for most of the identifiers discussed in the previous section, `robots.txt` raises a particular set of issues in this context. Since `robots.txt` functions based on self-assigned user agents, there is currently no way for rightholders to indicate that their opt-out applies to a particular class of uses such as `no-tdm` or

[8] Which is an increasingly common occurrence. In addition, there are multiple other technologies, such as automated content recognition, that rely on forms of TDM and that are generally viewed as beneficial by copyright holders.

[9] To be clear, the use of a more limited opt-out such as no-generative-ai does not create any obligations on model trainers to include the opted-out works in other types of services that they offer, as it simply expresses a reservation of some of the scope of the rights covered by the TDM exception in Article 4 of the CDSM directive.

no-generative-ai. Instead, rightholders must target all known user agents that engage in the collection of training data for AI model training with specific rules. From the perspective of rightholders this 'vertical approach' is both ineffective and not scalable as it requires rightholders to constantly monitor for new/unknown crawlers and subsequently target their user-agents.

To address this concern, AI model trainers should commit to a 'horizontal' approach based on generic (or virtual) user-agents that correspond to uses defined through the standardized vocabulary. These could take the form of wildcard user-agent names such as *-genai, *-tdm, *-aiuser,... coupled with a commitment to comply with rules addressed to these generic user agents[10].

{INFRASTRUCTURE}

One of the advantages of *location-based* approaches to opting out works from AI training is that they do not require a dedicated infrastructure. The approaches tend to rely on existing protocols (robots.txt and features of the http protocol stack) or can be implemented in a way that leverages the web's existing architecture (ai.txt). The same is true for *unit-based* approaches that rely on embedded metadata, where the opt-out information is directly linked to the media files — but these approaches are vulnerable to metadata stripping and are hard to implement retroactively for content that has already been published.

By contrast, *unit-based* approaches that rely on content-derived identifiers (e.g., ISCC, watermarking, or fingerprinting approaches) require some form of registry where opt-outs are recorded. Such registries can be proprietary (see, for example, Spawning's DNT registry), but can also be operated as public infrastructure.

To address the trust issues between rightholders and AI model trainers that surround the use of copyrighted works to train AI models and to facilitate the development of an open standard for opt-outs that can be implemented by any new market entrant, an approach that relies on publicly accessible registries seems preferable to approaches that rely on proprietary repositories. To facilitate machine-readable opt-outs that are effective and scalable, opt-outs should be made transparent to all parties involved in AI training.

Looking at the current landscape of *unit-based* identifiers, an approach based on a content derived identifier such as the ISCC to identify opted-out works and record opt-outs via the

[10] This approach would still leave room for rules addressed to specific user agents, since such rules would overrule the generic rules. This enables rightholders to extend or withhold permissions on a per-model-trainer basis.

proposed standardized vocabulary seems viable[11] for at least some categories of works[12]. Such a registry would soft-bind opt-out declarations based on the standardized vocabulary to ISCC codes. This would allow AI model trainers to use ISCC codes as a look-up key to check the registry for known opt-outs.

While setting up such a registry does seem relatively straightforward, the operation of such a registry will likely require significant resources for tasks like conflict resolution, quality control, security, and verification of opt-out claims[13]. This raises the question of how such a registry should be funded and governed. There are at least three general approaches to this:

Such a registry could be run by the European Commission[14], it could be run by a non-profit entity that is funded and governed by stakeholders[15], or it could be built as a system of federated registries, which would permit private entities (including AI model trainers, rightholders or specialized service providers[16]) to operate interconnected registries.

The main advantage of the last scenario would be that it allows for some level of experimentation and is likely faster to become operational than the other two more centralized approaches that would require a lot of stakeholder alignment to get off the ground.

---

[11] Provided that methods to generate and validate ISCC codes from digital content also operate at the scale of AI training datasets which can consist of billions of individual works. While the ISCC seems to meet the needs identified here, it is still untested at scale, and there are no existing implementations for the use of opting out of TDM.

[12] The ISCC standard shows a lot of promise for identifying media files that tend to remain relatively static or circulate with basic alterations of the content, such as images whose scale, resolution, or brightness may differ. It is robust against the most common changes, such as file format conversion, compression, and decoding. Dynamic web-based works, such as news websites or software code present more challenges and may be better served by location-based approaches. While ISCC codes can be generated from text files, they are more suitable for large bodies of static text that do not undergo significant change, such as entire books or chapters.

[13] Establishing right holder identities and verifying that rightholder opt-outs are legitimate will likely be out of scope for an initial version of the registry. This is a difficult problem to solve given the complexities of rights in various value chains and jurisdictions. Initially, a registry probably requires some form of a gated approach (which is at tension with the requirements of CDSM that any rightholder has the right to reserve the rights in question). In the long run, good quality of rightholder claims is in both parties' interest: Model trainers do not want unjustified opt-outs, and rightholders who see opt-outs as a step towards licensing access to their works will want ownership information to be as accurate as possible.

[14] Since neither the AI Act nor the CDSM directive foresees such a registry, it is unclear where such a registry should be hosted. Options include the AI office or the EUIPO which already maintains registries for Orphan Works and Out-Of-Commerce Works, although these operate on a completely different scale than any opt-out registry would need to operate.

[15] In this scenario, the registry could be run by a non-profit entity that maintains the registry on behalf of the stakeholders involved. While the governance structure would need to include representation from both rightholders and AI model training companies, it seems clear that the funding would — at least initially — need to come from the AI model trainers since the registry ultimately serves them as it would be an essential component of their Article 53(1c) compliance policies.

[16] Such as the aforementioned Spawning.

{EFFECT OF THE OPT-OUT}

The effect of the opt-out refers to the actions taken by model trainers once they have received and recorded an opt-out and identified the works to which the opt-out applies to. At the most simple level, it is clear that when an opt-out has been recorded, the opted-out work should not be added to the training data used to train new generative AI models unless other exceptions for doing so apply.

Fundamentally, the opt-out can only apply to reproductions that have been made after the opt-out request has been received. At the current stage of technology, the works cannot be removed from models that have already been trained. This also reflects the fact that opt-outs apply only to copyright-relevant acts of reproduction and not to the use of a model that has been trained on copyrighted works. In practice, that means that for each model, there will be some sort of opt-out cut-off date, after which new opt-outs will no longer affect the model's training. This cut-off date should be as close as possible to the beginning of the training run and should be clearly recorded and communicated when the model is released.

Also, fundamentally, the opt-out must be understood as a reservation of rights under Art. 4(3) of the CDSM, not as a separate prohibition on uses of opted-out content. If, for example, another copyright exception of limitation would apply, the opt-out does not act as a prohibition on those activities. There are examples where machine learning can be deployed to protect copyright owner interests (such as in output filters or for model post-training to resist infringement-seeking prompts), so care must be taken when defining the effect that opt-outs will have on GPAI model developers.

The other challenge with regard to the scope of the opt-out is the fact that opt-outs apply to works and that it is relatively likely that training data sets will contain multiple expressions of the same work. For example, this may be the case because an opted-out work has been crawled and added to the training data previously when it was not opted-out yet or because the work has been crawled and added to the training data from additional sources where it was not identified as being subject to an opt-out. As a result, training data may contain multiple copies of a work that has been opted out. This means that in addition to not adding opted-out works to the training data used to train new models, model trainers should also make an effort to identify other instances of the opted-out work within the data they use to train future models.

The ability to ensure such horizontal application of opt-outs within training data sets depends on two factors. Crucially, it depends on the type of identifier used when opting out. While content-derived identifiers such as ISCC can be expected to allow model trainers to identify at least some[17] other instances of opted-out works in their training data, other opt-out methods, such as *location-based* identifiers or metadata-based identifiers, are less useful for finding other

[17] There are currently no perfect systems of content recognition/deduplications, and at least initially, any approach can be expected to be imperfect.

instances. The ability to horizontally apply opt-outs also differs depending on the media types. Here, text-based works pose specific challenges since it will often be impossible to identify smaller chunks of an opted-out work as being part of that work only based on a work-derived identifier. To reliably identify parts of works, the model trainers would need to have access to reference files, but given the current lack of trust between rightholders and model trainers, it seems unlikely that rightholders would be willing to provide model trainers with reference files[18]. A similar challenge exists for time-based media such as music or video.

Secondly, and related to the first factor, the ability to ensure the horizontal application of opt-outs within their training data sets will also depend on how well-resourced the entity undertaking the model training is. Here, it will be important to ensure — as part of the code of practice — that this does not turn into another competitive advantage to the most well-resourced AI model training companies.

NEXT STEPS {}

As highlighted in the introduction of this paper, general-purpose AI model providers have to put in place policies to comply with the relevant parts of the EU copyright framework and rightholder opt-outs in particular. While the resulting policies will likely differ between individual model training companies, the process of arriving at policies that are effective, scalable, and address the needs of both rightholders and AI model trainers poses a number of collective action problems. This paper has identified four key components of compliance policies where there is a need to establish consensus and/or converge on a limited number of implementation options.

Based on the analysis provided in this paper, there are a number of obvious steps that can be taken to build consensus and converge on solutions that work for all stakeholders.

The most obvious first step would be to establish a common vocabulary for opt-outs that finds support among rightholders and AI model trainers and to align them with a set of generic user-agents that can be addressed via `robots.txt`. By committing to this, AI model trainers would address some of the core concerns expressed by rightholders about the current state of affairs when it comes to machine-readable opt-outs. It seems likely that such a set would contribute to increasing the level of trust between AI model trainers and rightholders.

In a second step, it should be explored whether the same vocabulary can form the basis for a standardized way of respecting rightholder opt-outs declared on the basis of individual URLs (such as ai.txt), especially in situations where rightholders upload dynamic web content such as text or software code, but do not control the entire domain on which their works are hosted.

[18] Note that such a scenario is not entirely hypothetical as rightholders frequently share reference files with technology companies in other contexts. One example is automated content recognition systems such as YouTube's ContentID or Facebook's Rights Manager.

At the same time, all stakeholders should explore the options to build out an infrastructure for *unit-based* opt-outs. As we have argued above, any compliance approach would ideally include both *location-based* and *unit-based* approaches to identifying opt-outs, and it will be important for all stakeholders to agree on one or more *unit-based* identifiers as soon as possible.

Ultimately, it is up to the European Commission and the AI Office to guide stakeholders towards an implementation of opt-outs that is workable for creators, other rightholders, and GPAI model developers alike. We hope that by outlining the contours of an approach to comply with Art. 53(1c) of the AI Act, this paper can contribute to that goal.

ABOUT OPEN FUTURE

Open Future is a European think tank that develops new approaches to an open internet that maximize societal benefits of shared data, knowledge, and culture.

Paul Keller is a co-founder and director of policy at Open Future. His work focuses on the intersection of copyright policy and emerging technologies. He works on policies and systems that improve access to knowledge and culture and protect the digital public sphere.