cc signals

creative commons

From Human Content to Machine Data

Introducing CC Signals

Lead authors: Jack Hardinges, Sarah Pearson, & Rebecca Ross

Distribution, June 2025

Table of Contents

About Creative Commons

Creative Commons (CC) is an international nonprofit organization that empowers people to grow and sustain the thriving commons of shared knowledge and culture we need to address the world's most pressing challenges and create a brighter future for all.

Our Vision: A world where education, culture, and science are equitably shared as a means to benefit humanity.

Our Mission: CC empowers individuals and communities around the world through technical, legal, and policy solutions that enable the sharing of education, culture, and science in the public interest.

Acknowledgements

This paper is the product of conversations, workshops, consultations, and research over the last three years.

We wish to thank the participants of the *Community Workshop on Preference Signals* held on June 6, 2024 in New York City, and on October 23, 2024 in San Francisco, as well as the participants of the *From Human Content to Machine Data: Using Collective Action to Develop a New Social Contract for Machine Reuse* workshop on April 23, 2025 in Berlin and on May 20, 2025 in New York City. We'd also like to thank our partners, without whom these workshops would not have been possible: The Alexander von Humboldt Institute for Internet and Society, Engelberg Center on Innovation Law & Policy, Morrison & Foerster, and NYU Stern School of Business. We'd also like to thank the CC global community who have been engaging with us on this topic and providing invaluable insights and feedback, as well as the various legal and technical experts who have supported CC's preference signals vision.

This work would not be possible without the support of the members of CC's Open Infrastructure Circle. Thank you. The continuation of this work will require ongoing sustainable funding. Please contact us to explore how you can make a difference through the Open Infrastructure Circle.

Summary

Recent advances in AI have been driven by the use of large amounts of data, including from across the web.

This isn't entirely new. Over the past two decades, machines have been used to access and compile web content to do things like build search engines and create digital archives. Machine reuse of web data has largely been governed by informal norms and standards. This social contract was based on a degree of reciprocity, and generally aligned with people's reasonable expectations for how their works would be used when they shared them publicly.

However, it's increasingly clear that the social contract that underpinned machine use of web data in the past no longer holds. Today, machines don't just crawl the web to make it more searchable or to help unlock new insights—they feed algorithms that fundamentally change (and threaten) the web we know.

In response, some creators are choosing to take their content offline. Others are trying to block machines from accessing their works and erecting paywalls. Large rightsholders are pushing for legislators to expand the scope of intellectual property rights.

This isn't sustainable, and it isn't leading to the future we want. The impact of large AI models, combined with this understandable backlash, risks creating a world where people are no longer able or willing to share their works. Knowledge and creativity could be further locked up, and decades of progress made by the open movement reversed.

This matters, as universal access to knowledge and culture is a human right, and vital to our ability to address our most pressing challenges going forward. At this critical juncture, we believe CC must intervene to help drive towards a more equitable digital future.

We're working on a first iteration of a preference signals framework, which we're provisionally calling CC signals. CC signals are designed to offer a new way for stewards of large collections of content to indicate their preferences as to how machines (and the humans controlling them) should contribute back to the commons when they reuse and benefit from using the content.

Our intervention is based on the beliefs that there are many legitimate purposes for machine reuse of content that must be protected, and that an ecosystem that better addresses the legitimate concerns of those creating and stewarding human knowledge is both possible and necessary.

This paper describes why we're arrived at these beliefs and are taking this action. We're publishing this alongside an initial prototype of CC signals and a request for feedback.

We can't make this a reality without community—join us.

A Note on Terminology

In this paper, we use the terms 'AI' and 'large AI models' as shorthand terms for what we know is a complex field of technologies and practices. We recognize that AI is not really 'artificial' (in that it is created and used by humans), nor 'intelligent' (at least in the way we think of human intelligence), and that model size is relative (we use 'large' to describe models developed since the late 2010s that are able to process large volumes of multimodal data following the introduction of transformer architectures). We talk more specifically about certain types or capabilities of AI, such as generative AI models, where it is necessary.

Background

How AI Models Use Data

Recent progress in AI has been characterized by models of large scale and complex architectures, capable of tasks such as natural language processing and content generation.

Many of these models have been developed using large amounts of data from the public web.[1] Web crawling plays a significant part in this. It involves using automated programs to systematically navigate and make copies of data from websites, blogs, forums, books, social media platforms, and other sources.

Some AI developers, for example, rely on crawlers to extract textual content from different sources in order to train models to detect patterns and then generate human-like text in response to prompts.[2] The datasets used to train large AI models are often made up of multiple datasets generated through web crawling (especially Common Crawl[3] and

[1] *Huang, S. & Siddarth, D. (2023, February 6). Generative AI and the Digital Commons. The Collective Intelligence Project.* https://cip.org/research/generative-ai-digital-commons

[2] *Murgia, M. (2023, September 9). Generative AI Exists Because of the Transformer. Financial Times.* https://ig.ft.com/generative-ai/

[3] *Common Crawl. (n.d.). Common Crawl. Common Crawl.* https://commoncrawl.org/

derivatives of it, such as LAION5B[4] and The Pile[5]) along with supplemental data that an AI developer has crawled themselves.[6] An analysis of the widely-used C4 training dataset, a public dataset based on Common Crawl's corpus, found that its content originated from more than 14 million different web domains.[7]

AI developers also make use of large datasets that are created and maintained with the express purpose of being widely used, including by collaborative communities (e.g., datasets derived from Wikipedia[8]), open source communities (e.g., WikiSQL[9]), scientific projects (e.g., AlphaFold Protein Structure Database[10]), and governments (e.g., official statistics). Platforms for AI development such as Hugging Face[11] and Kaggle[12] now host large collections of training datasets of varying provenance. Some AI developers enter into partnerships with other organizations to gain access to valuable data sources. Large technology organizations, such as Google and Meta, repurpose the masses of data generated through users' interactions with their platform services for model training.

Large AI models rely on access to data throughout their lifecycle. Various types and sources of data, and approaches to accessing it, are used in the process of testing, validating, benchmarking, and fine-tuning models.[13] Once models have been deployed, techniques such as retrieval augmented generation (RAG) enable them to retrieve, in real time, information from the web or a user's system in response to queries, as opposed to generating the response from the trained model alone.

---

[4] *LAION. (n.d.). Projects. LAION.* https://laion.ai/projects/

[5] *Eleuther AI. (n.d.). The Pile. Eleuther AI.* https://pile.eleuther.ai/

[6] *Baack, S. (2024, February 6). Training Data for the Price of a Sandwich. Mozilla Foundation.* https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/

[7] *Schaul, K., Chen, S. Y., & Tiku, N. (2023, April 19). Inside the secret list of websites that make AI like ChatGPT sound smart. Washington Post.* https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

[8] *Wikipedia. (n.d.). Wikipedia. Wikipedia.org; Wikimedia Foundation.* https://www.wikipedia.org/

[9] *salesforce/WikiSQL. (n.d.). A large annotated semantic parsing corpus for developing natural language interfaces. (n.d.). GitHub.* https://github.com/salesforce/WikiSQL

[10] *Google Deepmind & EMBL-EBI. (n.d.). AlphaFold Protein Structure Database. Alphafold.ebi.ac.uk.* https://alphafold.ebi.ac.uk

[11] *Hugging Face. (n.d.). Datasets. Hugging Face.* https://huggingface.co/docs/datasets/en/index

[12] *Kaggle. (2024). Datasets. Kaggle.com.* https://www.kaggle.com/datasets

[13] *Hardinges, J. & Simperl, E. (2024, October 15). A data for AI taxonomy. Open Data Institute.* https://theodi.org/news-and-events/blog/a-data-for-ai-taxonomy/

Governing Machine Use of Web Data, To Date

The use of large volumes of data, including from across the public web, isn't specific to training and deploying large AI models.

Text and data mining (TDM), the process of transforming large amounts of unstructured text into structured formats in order to identify patterns, trends, and other insights,[14] has long been deployed in many fields of research, from computer science and linguistics to environmental sciences and humanities. There is broad international convergence on the potential social value of TDM, and while the scope and details vary widely, every copyright law in the world has at least one exception that promotes research purposes.[15]

Creating archives of the web, such as the Internet Archive[16] and Wayback Machine,[17] relies on using machines to systematically navigate and make copies of data from billions of websites. Common Crawl, mentioned above, was established as a nonprofit foundation in 2007 to produce large crawls of web data for anyone to access and use for analysis, rather than only the handful of companies who, at the time, could afford to undertake their own crawling at scale.[18] Prior to becoming a key source of training data for large AI models, most creators and web users would have been unaware of Common Crawl, and its largely research-oriented use did not spark major debate.

Web search is similarly predicated on the use of machines to find and store information from across the web.[19] In their traditional form, search engines presented a fairly simple 'deal' for website owners, which was: if your search product sends us traffic, then we'll allow crawling. In this context, this reciprocal exchange of value has been important, especially given much of the web's reliance on advertising revenue based on traffic and clicks.

This doesn't mean machine use of web data has been entirely uncontested or an anything-goes free-for-all. Some websites and news publishers, for example, sued the

---

[14] *IBM. (2021, October 15). Text Mining. IBM.* https://www.ibm.com/think/topics/text-mining

[15] Flynn, S., Schirru, L., Palmedo, M., & Izquierdo, A. (2022). *Research Exceptions in Comparative Copyright*. Joint PIJIP/TLS Research Paper Series. https://digitalcommons.wcl.american.edu/research/75/

[16] *Internet Archive. (n.d.). Internet Archive. Internet Archive.* https://archive.org/

[17] *Wikipedia Contributors. (2019, March 8). Wayback Machine. Wikipedia; Wikimedia Foundation.* https://en.wikipedia.org/wiki/Wayback_Machine

[18] *Baack, S. (2024, February 6). Training Data for the Price of a Sandwich. Mozilla Foundation.* https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/

[19] *Google. (n.d.). Organizing Information - How Google Search Works. Google.* https://www.google.com/search/howsearchworks/how-search-works/organizing-information/

companies behind early versions of search engines and sought to require that they get explicit consent before crawling.[20] Had the publishers won, the web may have been stifled in its infancy.

Instead, over the last 25 years, publishers, crawlers, and other stakeholders have worked together to establish norms about the appropriate use of web data.

These informal norms represent a form of *social contract*. Social contracts are not codified in laws or strict rules, but shape behavior based on shared values and a desire to act in ways that are mutually beneficial. They are expected to be followed and carry weight where the law doesn't necessarily require adherence,[21] with consequences for noncompliance. They are continuously negotiated and evolve over time, and while they don't necessarily resolve differences and tensions, they provide a degree of mediation based on shared expectations that enables different groups to cooperate.

A simple yet important component of the social contract that has governed machine use of web data has been the Robots Exclusion Protocol (robots.txt).[22] Websites could use robots.txt—as well as other indexing protocols such as HTML meta tags and HTTP header directives—to indicate their preferences to crawlers (including to disallow them) in a scalable, machine-readable format. Crawlers, by and large, respected preferences expressed using robots.txt, and the protocol supported the rapid growth of the web over the 2000s and 2010s. The process of developing and maintaining robots.txt and other protocols also contributed to the forming of the social contract, in that it involved the input and collaboration of affected stakeholders over a number of years.

---

[20] *Gasser, U. (2006, January 1). REGULATING SEARCH ENGINES: TAKING STOCK AND LOOKING AHEAD. Yale Journal of Law & Technology.* https://yjolt.org/sites/default/files/gasser-8-yjolt-201.pdf

[21] *Masiello, B. & Slater, D.. (2023, September 19). Beyond Copyright: Tailoring Responses to Generative AI & The Future of Creativity. Tech Policy Press.* https://www.techpolicy.press/beyond-copyright-tailoring-responses-to-generative-ai-the-future-of-creativity

[22] *Wikipedia Contributors. (2025, March 16). robots.txt. Wikipedia; Wikimedia Foundation.* https://en.wikipedia.org/wiki/Robots.txt

The Future of the Commons Is under Threat

Advances in AI and a Ruptured Social Contract

It's clear that recent advances in AI have ruptured the social contract that underpinned machine use of web data in the past.

There are a multitude of different reasons for this. First, unlike traditional web search, large AI models and the products that are enabled by them do not provide websites with the same benefits of discovery and traffic in return for allowing their content to be crawled.

> *"The intermediation role played by AI systems is altogether new: where the role of search engines has traditionally been to surface the most relevant links to answers of the user's query, AI systems typically expose directly an answer… For the large number of content producers whose sustainability relies on direct exposure to (or interactions with) the final end user, this lack of reliable exposure makes it unappealing to leave their content crawlable for AI-training purposes."*[23] - Dominique Hazaël-Massieux

Reducing the need for people to visit original sources of information, or outright preventing them from doing so, threatens the long-term sustainability of those sources. Case in point: traffic to the world's second most visited history website, World History Encyclopedia, has dropped by 25% since its content was included in Google's AI Overviews in late 2024.[24] Smaller publishers are reporting similar reductions in traffic.[25]

This 'paradox of reuse,[26] where content powers new technology that in time reduces the need for users to visit its source, is particularly acute for web publishers that rely on traffic for advertising revenue, as well as collaboratively-maintained sources of information, like Wikipedia, that rely on contributing communities and user donations. This issue is made worse by the fact that, across almost all forms and applications of large AI models, no

---

[23] *Hazaël, D. (2024, July 26). Managing exposure of Web content to AI systems Reversibility of consent on crawling. Internet Engineering Task Force. https://www.ietf.org/slides/slides-aicontrolws-managing-exposure-of-web-content-to-ai-systems-00.pdf*

[24] *Kantrowitz, A. (2025, March 28). As AI Takes His Readers, A Leading History Publisher Wonders What's Next. Big Technology. https://www.bigtechnology.com/p/as-ai-takes-his-readers-a-leading*

[25] *Alba, D., & Love, J. (2025, April 7). Google AI Search Shift Leaves Website Makers Feeling 'Betrayed'. Bloomberg. https://www.bloomberg.com/news/articles/2025-04-07/google-ai-search-shift-leaves-website-makers-feeling-betrayed*

[26] *Vincent, N. (2022, December 2). The Paradox of Reuse, Language Models Edition. nmvg. https://nmvg.mataroa.blog/blog/the-paradox-of-reuse-language-models-edition/*

coherent theory or practice has emerged for maintaining attribution (and thereby being able to give credit) for the sources of information used.[27] What does a knowledge ecosystem look like when the very underpinnings of context, container, and credibility are divorced from the outputs made available to the public?

In the 2000s and 2010s, most machine use of web data involved addressing targeted research questions or had contained objectives, such as extracting facts from a body of scientific literature or patterns from a database. Large AI models, in contrast, require web data at much larger scale and are almost indiscriminate in scope, ingesting vast quantities of multimodal content. As a result, web publishers have begun to report large increases in machine crawling activity by AI developers.

Numerous open source software projects have described this increase as overwhelming and amounting to a denial-of-service (DoS) attack, with some estimating that more than 70% of traffic on their infrastructure is from AI crawlers.[28] Some domains are reporting particularly aggressive practices from AI crawlers, including deliberate circumvention of standard blocking measures, disregarding terms of service and licenses, ignoring robots.txt directives, spoofing user agents, and rotating IP addresses to avoid detection.[29]

This demonstrates bad faith and places a huge burden on the web in the form of increased bandwidth costs and service instability, including on the many nonprofit institutions and communities that maintain it. Wikipedia, for example, has described how a high volume of crawling is creating significant work for its site reliability team.[30] Some digital collections maintained by libraries, archives, and museums have been knocked offline.[31] This issue is only likely to worsen as the crawlers used by AI developers not only crawl in order to build static

---

[27] *Chandrasekhar, R. (2025, May 12). Legal frictions for data openness. French National Centre for Scientific Research; InnoCube; Open Knowledge Foundation.* https://ok.hypotheses.org/files/2025/03/Legal-frictions-for-data-openness-open-web-and-AI-RC-2025-final.pdf

[28] *Venerandi, N. (2025, March 20). FOSS infrastructure is under attack by AI companies. LibreNews.* https://thelibre.news/foss-infrastructure-is-under-attack-by-ai-companies/

[29] *Edwards, B. (2025, March 25). Open Source devs say AI crawlers dominate traffic, forcing blocks on entire countries. Ars Technica.* https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/

[30] *Mueller, B., Danis, C. & Lavagetto, G. (2025, April). How crawlers impact the operations of the Wikimedia projects. Wikimedia Foundation.* https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/

[31] *Maiberg, E. (2025, June 17). AI Scraping Bots Are Breaking Open Libraries, Archives, and Museums. 404 Media.* https://www.404media.co/ai-scraping-bots-are-breaking-open-libraries-archives-and-museums/

training datasets, but also to keep models current[32] and to dynamically 'fetch' information in response to user prompts or requests.[33]

In this sense, large AI models are benefiting from the labor, money, and care that goes into creating and maintaining the commons while also threatening to 'bleed it dry.'[34]

> *"Traditional models of data commons and open data are being stretched and tested by AI's propensity to rapidly parse, learn from, and exploit shared data. This leads to renewed concerns over a 'tragedy of the commons,' in which data resources may become extracted or enclosed by just a few."*[35] - Stefaan G. Verhulst, Hannah Chafetz and Andrew Zahuranec

To many, the way that large AI models exploit the shared knowledge ecosystem is particularly unfair in the face of the huge financial gain accruing to the large companies behind their development.[36] Over the past year for example, the valuation of Perplexity, an AI search startup, has risen from $500 million to over $9 billion,[37] and OpenAI is generating over $300 million in revenue per month.[38] Many of the largest and most popular models do not have openly published code or training data and often carry significant usage restrictions, which further represents a private enclosure of knowledge taken from the commons.

---

[32] *Edwards, B. (2025, March 25). Open Source devs say AI crawlers dominate traffic, forcing blocks on entire countries. Ars Technica. https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/*

[33] *Hazaël, D. (2024, July 26). Managing exposure of Web content to AI systems Reversibility of consent on crawling. Internet Engineering Task Force. https://www.ietf.org/slides/slides-aicontrolws-managing-exposure-of-web-content-to-ai-systems-00.pdf*

[34] *White, M. (2025, March 14). "Wait, not like that": Free and open access in the age of generative AI. Citation Needed. https://www.citationneeded.news/free-and-open-access-in-the-age-of-generative-ai/*

[35] *Verhulst, S. G., Chafetz, H. & Zahuranec, A. (2024, May 9). "Data Commons": Under Threat by or The Solution for a Generative AI Era? Rethinking Data Access and Re-use. Data & Policy Blog. https://medium.com/data-policy/data-commons-under-threat-by-or-the-solution-for-a-generative-ai-era-rethinking-9193e35f85e6*

[36] *Evans, B. (2023, August 27). Generative AI and intellectual property. Benedict Evans. https://www.ben-evans.com/benedictevans/2023/8/27/generative-ai-ad-intellectual-property*

[37] *Wheeler, K. (2024, November 7). How Perplexity AI Boomed From US$500m to US$9bn. Technology Magazine.com. https://technologymagazine.com/articles/from-500m-to-9bn-charting-perplexitys-soaring-valuation*

[38] *Isaac, M., & Griffith, E. (2024, September 27). OpenAI Is Growing Fast and Burning Through Piles of Money. The New York Times. https://www.nytimes.com/2024/09/27/technology/openai-chatgpt-investors-funding.html*

There is also the significant burden large AI models place on the environment, such as via carbon emissions and electricity costs. The carbon footprint of training a single large language model is approximately the equivalent of 125 round-trip flights between New York and Beijing.[39] A 100-word email generated by ChatGPT requires the equivalent of one water bottle to cool the servers for the underlying GPT-4 model to function.[40] For some, including those whose works may have been used to train such models, the benefits of AI may not be worth the environmental impact and extraction from our natural commons.

It is difficult to overstate the wider disruption that advances in AI could cause to the knowledge ecosystem, from the future of livelihoods in some industries to the ability to separate fact from fiction and the value of human authorship. In creative industries such as art, music, and creative writing, there is concern that generative tools will interfere with human creatives' ability to create, share, and earn compensation.[41] Voice actors have discovered their likenesses have been used in inappropriate commercial applications or political messages.[42] Journalists are concerned about the effects of their reporting being presented to users without its original context or editorial standards.[43]

Evidence of the Broken Social Contract Is All Around Us

The backlash against advances in AI is sweeping, and demonstrates the extent to which the current relationship between AI developers and the commons is broken.

We've observed significant unrest among some web communities, including among those who have expressly created (and in some cases, openly licensed) their content for wide use. In 2019, many Flickr users were dismayed to learn their openly-licensed images had been used to train facial recognition models.[44] More recently, many of Reddit's biggest forums

---

[39] *Dhar, P. (2020, August 12). The carbon impact of artificial intelligence. Nature Machine Intelligence.* https://www.nature.com/articles/s42256-020-0219-9

[40] *Verma, P., & Tan, S. (2024, September 18). A bottle of water per email: the hidden environmental costs of using AI chatbots. Washington Post; The Washington Post.* https://www.washingtonpost.com/technology/2024/09/18/energy-ai-use-electricity-water-data-centers/

[41] *Zhao, B. (2024, March 28). Replacement of human artists by AI systems in creative industries. UNCTAD.* https://unctad.org/news/replacement-human-artists-ai-systems-creative-industries

[42] *Pistilli, G. (2025, March 26). I Clicked "I Agree", But What Am I Really Consenting To?. Hugging Face.* https://huggingface.co/blog/giadap/beyond-consent

[43] *Bridging Responsible AI Divides. (2025, February 1). Journalism and Generative AI: Data, Deals and Disruption in the News Media. Zenodo.* https://zenodo.org/records/14968195/files/Journalism%20and%20Generative%20AI%20Workshop%20Report.pdf?download=1

[44] *Merkley, R. (2019, March 13). Use and Fair Use: Statement on shared images in facial recognition AI. Creative Commons. Creative Commons.* https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/

'went dark' in protest over the platform's plans to enable AI developers to access the mass of forum conversations they'd played a vital role in creating.[45] Contributors to Stack Overflow, an internet forum for developers, have been banned from the site after they deleted their content in order to stop it from being used to train large AI models.[46]

An increasing number of web publishers appear to be blocking their content from being used for AI model training, or from being crawled at all. In 2023, the *New York Times* updated its terms of service to explicitly prohibit its content from being crawled and asked Common Crawl to remove any of its existing news articles from its dataset.[47] This practice extends to smaller publishers, too. According to Cloudflare, more than 84% of the websites that use its hosting services have now applied technical restrictions to stop crawlers from accessing their content.[48] An analysis of the C4 training dataset found that between 2023 and 2024 there had been "a rapid crescendo of data restrictions from web sources, rendering [more than] 28% of the most actively maintained, critical sources in C4, [now] fully restricted from use."[49]

This shift away from allowing *any* machine reuse of content represents the kind of all-or-nothing choice that is likely to have negative effects. Overbroad opt-outs risk unintentionally limiting many of the types of TDM that have long been considered acceptable and socially valuable.

In addition, we've seen a worrying trend among some academic publishers who control public access to research toward more restrictive CC licenses, in the hopes that it will restrict AI

---

[45] *Hern, A. (2023, June 14). Reddit moderators vow to continue blackout in API access fees row. The Guardian.* https://www.theguardian.com/technology/2023/jun/14/reddit-moderators-vow-to-continue-blackout-in-api-access-fees-row

[46] *Grimm, D. (2024, May 8). Stack Overflow bans users en masse for rebelling against OpenAI partnership. Tom's Hardware.* https://www.tomshardware.com/tech-industry/artificial-intelligence/stack-overflow-bans-users-en-masse-for-rebelling-against-openai-partnership-users-banned-for-deleting-answers-to-prevent-them-being-used-to-train-chatgpt

[47] *Weatherbed, J. (2023, August 14). The New York Times prohibits using its content to train AI models. The Verge.* https://www.theverge.com/2023/8/14/23831109/the-new-york-times-ai-web-scraping-rules-terms-of-service

[48] *Heikkilä, M. (2024, December 18). This is where the data used to build AI comes from. MIT Technology Review.* https://www.technologyreview.com/2024/12/18/1108796/this-is-where-the-data-to-build-ai-comes-from/

[49] *Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A., Dinh, A., Chitongo, C., Yin, D., & Sileo, D. (2024, July 24). Consent in Crisis: The Rapid Decline of the AI Data Commons. ArXiv.org.* https://arxiv.org/abs/2407.14933

training on published materials. Cambridge University Press has said that "historically for books, our default open access license was a CC BY-NC license… [but] we are now looking at using more CC BY-NC-ND as the default."[50] This shift is being led by organizations such as the British Academy.[51] These attempts are misguided, because they impede the ability of humans to make full use of the materials, thus limiting access to knowledge, while simultaneously providing ineffective control over machine reuse.

Elsewhere, licenses with behavioral-use clauses designed for the distribution of AI models are being applied to datasets,[52] and new clauses have been put forward to shape the use of content for model training, such as the Open Data Commons License.[53]

Some web users are taking a more offensive approach, by seeking to cause damage to web crawlers and large AI models. Nepenthes, for example, was developed by hosting service SourceHut to act as 'a tarpit to catch web crawlers' by generating an endless, circular sequence of pages and links that disrupt how crawlers work.[54] Nightshade is also an offensive tool that works by distorting generative image models if they access an artist's works without permission.[55]

There's significant litigation now underway between large rightsholders and AI companies. Getty Images, for example, is suing Stability AI for allegedly infringing copyright in more than 12 million photos in the development of its Stable Diffusion model.[56] Music publishers UMG, Concord, and ABKCO have claimed that Anthropic infringed its copyright in popular song

---

[50] *Hansen, D. (2025, March 17). AI Licensing: An Interview with Ben Denne of Cambridge University Press. Authors Alliance. https://www.authorsalliance.org/2025/03/17/ai-licensing-an-interview-with-ben-denne-of-cambridge-university-press/*

[51] *British Academy. (2025, May 12). Open Access and the REF: A British Academy position paper. The British Academy. https://www.thebritishacademy.ac.uk/publications/open-access-and-the-ref-a-british-academy-position-paper/*

[52] *Hugging Face. (n.d.). Datasets; Active filters: creativeml-openrail-m. Hugging Face. https://huggingface.co/datasets?license=license:creativeml-openrail-m&sort=trending*

[53] *Benhamou, Y., & Dulong de Rosnay, M. (2024, January 8). Open Data Commons Licenses (ODCL): Licensing Personal and Non Personal Data Supporting the Commons and Privacy. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4662511*

[54] *ZADZMO. (n.d.). Nepenthes. ZADZMO. https://zadzmo.org/code/nepenthes/*

[55] *The Nightshade Team. (n.d.). Nightshade: Protecting Copyright. University of Chicago. https://nightshade.cs.uchicago.edu/whatis.html*

[56] *Vincent, J. (2023, February 6). Getty Images Sues AI Art Generator Stable Diffusion in the US for Copyright Infringement. The Verge. https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion*

lyrics when training its Claude model.[57] Many claimants are seeking damages, with some arguing that models should be destroyed.[58]

Amid the copyright uncertainty, some content creators are striking commercial licensing agreements with AI firms. The *Wall Street Journal*, the *New York Post*, *The Times*, the *Sunday Times*, *Barron's*, MarketWatch, The Associated Press, Axel Springer, Prisa Media, *Le Monde*, and the *Financial Times* have all licensed their content to OpenAI.[59]

There have also been calls for expanding the scope of intellectual property rights to address large AI models' use of data. Large rightsholders in many countries, including Canada,[60] Brazil,[61] Japan,[62] and Australia,[63] are pushing for legislators to constrain current AI development and proposed copyright exceptions. In the UK, a campaign led by the News Media Association recently resulted in nearly every national newspaper adopting the same front cover in protest against the UK Government's plans to require rightsholders to opt out of having their works used to train AI models.[64]

---

[57] *Brittain, B. (2025, March 26). Anthropic wins early round in music publishers' AI copyright case. Reuters. https://www.reuters.com/legal/anthropic-wins-early-round-music-publishers-ai-copyright-case-2025-03-26/*

[58] *Samuelson, P. (2024, February 15). How to Think About Remedies in the Generative AI Copyright Cases. Lawfare Media. https://www.lawfaremedia.org/article/how-to-think-about-remedies-in-the-generative-ai-copyright-cases*

[59] *Staff, C. A.-P. (2024, May 28). OpenAI inks multi-year content deal with News Corp. PR Week Global. https://www.prweek.com/article/1874481/openai-inks-multi-year-content-deal-news-corp*

[60] *Government of Canada.(2025, February 11). Consultation on Copyright in the Age of Generative Artificial Intelligence: What we heard report. Government of Canada. https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-copyright-age-generative-artificial-intelligence-what-we-heard-report*

[61] *Rocha de Souza, A. & Schirru, L. (2024, November 12). Regulating AI and Copyright in Brazil: the stage of the game. Kluwer Copyright Blog. https://copyrightblog.kluweriplaw.com/2024/11/12/regulating-ai-and-copyright-in-brazil-the-stage-of-the-game/*

[62] *Inagaki, K., & Keohane, D. (2024, July 21). Japan's copyright rules draw AI groups — and alarm from creators. Archive.ph; Financial Times. https://archive.ph/oBK5h*

[63] *Bennett. T. (2024, November 26). Force big tech to pay for AI training data: Senate committee. Archive.ph; Financial Review. https://archive.ph/2BnPA#selection-1255.0-1264.3*

[64] *Weatherbed, J. (2025, February 25). UK newspapers blanket their covers to protest loss of AI protections. The Verge. https://www.theverge.com/news/619063/uk-newspapers-covers-protest-government-ai-rights-proposal*

The Breakdown of the Social Contract Is Resulting in Enclosure

The enclosure of knowledge is happening. We see this in the movement of content behind paywalls or to more restrictive licenses. We see this in the increasing use of broad opt-outs from all machine uses of content. Over the long term, it may manifest in people not publicly sharing at all.

While the instinct to limit access to works is understandable, enclosure will not serve the public interest.

> *"By trying to wall off those considered to be bad actors, people wall off the very people they intended to give access to.*
>
> *People who gate their work behind paywalls likely didn't set out to create works that only the wealthy could access.*
>
> *People who implement registration walls probably didn't intend for their work to only be available to those willing to put up with the risk of incessant email spam after they relinquish their personal information.*
>
> *People who try to stave off bots with CAPTCHAs asking 'are you a human?' probably didn't mean to limit their material only to abled people who are willing to abide ever more protracted and irritating riddles."*[65] - Molly White

Universal access to knowledge and culture is a human right,[66] and it is vital to our ability to address our most pressing challenges going forward. This is especially true in an information environment saturated with AI-generated content, where it will be critical that humans have ready access to the educational resources, scientific research, journalism, and other core knowledge goods that help us advance.

It is not just *human* access to knowledge that benefits the public interest. Closing off access in response to AI crawlers—using blunt approaches that do not distinguish them from other machines—affects crawling for legitimate and widely-accepted purposes. This includes search, as well as other critical applications of machine access to web data that serve the public interest, such as misinformation and hate speech detection, accessibility

---

[65] *White, M. (2025, March 14). "Wait, not like that": Free and open access in the age of generative AI. Citation Needed.* https://www.citationneeded.news/free-and-open-access-in-the-age-of-generative-ai/

[66] *United Nations. (1948). Universal declaration of human rights. United Nations.* https://www.un.org/en/about-us/universal-declaration-of-human-rights

improvements, translation, and archiving. Various fields of research that rely on TDM are also impeded.

> *"AI companies' scraping is jeopardizing un-permissioned [research] projects. Independent technology research is crucial for ensuring a degree of transparency and understanding the platforms that are enmeshed in so much of our social, economic, entertainment, educational, and political lives."*[67] – Ryan McGrady, Ethan Zuckerman, Kevin Zheng

Closing down access to knowledge is *also* a problem for AI. After all, data from the public web is a key component in developing large AI models.[68] But, the adoption of adversarial and disrespectful relationships with creators and other users of the web will, over time, reduce incentives for people to share and widely distribute their work. In putting the commons at risk, large AI models are 'drilling away at their own foundations.'[69]

If content is no longer publicly available or otherwise becomes more risky and uncertain to use, it becomes solely accessible to those with deep pockets. Small firms, startups, nonprofits, and academic researchers would not have the financial means to go to court to defend their use of content, or enter into bilateral agreements to license data in the same way as large technology firms.[70] It would further entrench these firms from competition, especially given they already enjoy the benefit of access to their own proprietary datasets.[71] In addition to impeding human access to knowledge, we're concerned that a shift to restrictive licensing would result in a less fair, diverse, and competitive AI ecosystem.

---

[67] *McGrady, R., Zuckerman, E., & Zheng, K. (2025, January 30). AI Companies Threaten Independent Social Media Research. Tech Policy Press.* https://www.techpolicy.press/ai-companies-threaten-independent-social-media-research/

[68] *Open Source Initiative. (n.d.). OSAID FAQs. Open Source Initiative.* https://opensource.org/ai/faq#what-is-an-open-source-ai

[69] *Vincent, N. (2022, December 2). The Paradox of Reuse, Language Models Edition. nmvg.* https://nmvg.mataroa.blog/blog/the-paradox-of-reuse-language-models-edition/

[70] *Maffulli, S. (2024, May 7). Why datasets built on public domain might not be enough for AI. Open Source Initiative.* https://opensource.org/blog/why-datasets-built-on-public-domain-might-not-be-enough-for-ai

[71] *AI Now Institute. (2024, March 12). Joint submission to the European Commission's consultation on competition and generative AI. AI Now Institute.* https://ainowinstitute.org/publication/joint-submission-to-the-european-commissions-consultation-on-competition-and-generative-ai

The quality and safety of AI models would also suffer from mass enclosure of information. The issue of Western-centricity in training data[72] and resultant risks to deploying many large AI models in non-English languages,[73] for example, is unlikely to improve through restrictive licensing practices. Beneficial uses of AI are more likely to emerge with engagement from and the ability to access content generated by academia, civil society, and other actors, rather than by ceding the field to commercial interest alone.[74]

The good news? It is not too late to do something. Although so much AI training has already happened, there will undoubtedly be new models created in the future. And new models of equity and cooperation are needed to sustain access to content and information that deployed models now rely on. Widely accessible corpora of high-quality, current data, like Wikipedia, will have a vital role in providing the 'factual netting'[75] for deployed large AI models and the products they underpin, just as they do for the wider web.

A Collaborative Intervention

We're at a watershed moment. AI has broken the social contract that had governed the way we, and machines, share and access knowledge. In the face of this disruption, the instinct to limit access to information is understandable. But blunt enclosure will not serve the public interest in the long term and ultimately puts the commons at risk.

We Need a New Social Contract for Machine Reuse

To protect and grow the commons, there must be a new social contract to govern how AI models—and the tools, products and products they increasingly underpin—engage with it.

We believe creator consent is a core value of and a key component to a new social contract. This view is both an ethical and a pragmatic position, rather than a legalistic one. There are many scenarios in which creator consent may not be legally required under copyright law.

---

[72] *Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A., Dinh, A., Chitongo, C., Yin, D., & Sileo, D. (2024, July 24). Consent in Crisis: The Rapid Decline of the AI Data Commons. ArXiv.org.* https://arxiv.org/abs/2407.14933

[73] *Jain, D., Kumar, P., Gehman, S., Zhou, X., Hartvigsen, T., & Sap, M. (2024). PolygloToxicityPrompts: Multilingual Evaluation of Neural Toxic Degeneration in Large Language Models. ArXiv.org.* https://arxiv.org/abs/2405.09373

[74] *Hansen, G. W. (2025). AI deals underscore importance of open access (opinion). Inside Higher Ed.* https://www.insidehighered.com/opinion/views/2025/01/07/ai-deals-underscore-importance-open-access-opinion

[75] *Gertner, J. (2023, July 18). Wikipedia's Moment of Truth. The New York Times.* https://www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html

But, the absence of someone having a legal right to veto something does not mean their expectations and preferences do not matter.

Ethically, violation of creator expectations at scale is a problem. It is also short-sighted. Just as a homeowner does not have to invite you back if you refuse to clear your plate after dinner, or an employee can look for a new job if they believe their employer's policies are unfair, creators can choose not to share with the public at all. If we want a resilient ecosystem for knowledge sharing, we need creators to consent to their participation in the AI lifecycle.

It's clear that AI developers need this, too. Many examples of new, generative AI-powered tools decline to answer queries to avoid accessing news content they *are* permitted to access, while answering queries using content they *shouldn't be* accessing.[76] As we've written before, responsible technology developers share an interest in defining better ways to respect creators' wishes.[77] If they could get a clear signal of the creators' intent, then they would follow it.

To be sure, the concept of consent is complicated and may sometimes come into conflict with other values. There are important scenarios where it is both legally *and* ethically acceptable to use something, even where the creator doesn't consent and might outright object, such as for parody or critical review. The boundaries of consent are not static—instead, they need to be constructed and developed over time.

A Thriving Commons Requires Reciprocity

We believe that a critical ingredient to widespread consent is reciprocity.

Sharing knowledge generally—and CC-licensed content specifically—has always been based on an implicit assumption that we are all in this together. The basic value system of the commons is rooted in a fundamental belief that knowledge and creativity are building blocks of our culture, rather than just commodities from which to extract market value. We give, we take, and we give again.

This notion is not new to the commons. Scholar Elinor Ostrom described the reciprocal exchange of trust and cooperation as the key to successful governance of the commons.[78]

---

[76] *Jazwinska, K., & Chandrasekar, A. (2025, March 6). AI Search Has A Citation Problem. Columbia Journalism Review.* https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php

[77] *Stihler, C. (2023, August 31). Exploring Preference Signals for AI Training. Creative Commons.* https://creativecommons.org/2023/08/31/exploring-preference-signals-for-ai-training/

[78] *Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.*

The social concept of reciprocity is not necessarily about legal obligations. It is about maintaining a community or public-minded approach to how you participate in an ecosystem. The social concept of reciprocity in the commons is not transactional.

> *"I don't mean a bilateral exchange in which an obligation is incurred, and can then be discharged with a reciprocal 'payment.' I mean keeping the gift in motion in a way that is open and diffuse, so that the gift does not accumulate and stagnate, but keeps moving..."*[79] - Robin Wall Kimmerer

But large AI models currently massively benefit from the commons, while opaquely ingesting knowledge and repurposing it in ways that often involve obfuscation[80] and feel to some like exploitation.[81] The dynamic between AI and the commons should not be zero sum. A focus on, and more importantly, a commitment to, reciprocity could instill a mutually beneficial relationship between content creators and AI developers, reinforcing the commons for all.

Reciprocity is a big concept.[82] What it means to different stakeholders will undoubtedly vary, and infusing reciprocity into the relationship between AI development and the commons is not a simple task. Notably, there never has been, nor should there be, a mandatory one-to-one exchange of value between each individual and the commons.

At its essence, reciprocity is a profound notion that *this is for all of us*. We are playing the long game to defend and protect a thriving creative commons now and for future generations, using a new fit-for-purpose social contract as our guide.

More Copyright Is Not the Answer

We do not believe the expansion of copyright is a viable path toward a new social contract. We fundamentally believe that ideas, facts, and basic building blocks of knowledge cannot be owned.

Copyright is built upon this principle, by protecting the original expression of authors, but allowing others to freely reuse and build upon the ideas and information within a work. Creators necessarily learn from and develop their skills by engaging with pre-existing works

---

[79] *Kimmerer, R. W. (2024). The Serviceberry. Simon and Schuster.*

[80] *Van Houweling, M. (2025) The Freedom to Extract in Copyright Law. North Carolina Law Review.* https://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=7009&context=nclr

[81] *Tarkowski, A. (2025, January 22). Data Governance in Open Source AI Enabling Responsible and Systemic Access. Open Future.* https://openfuture.eu/wp-content/uploads/2025/01/250123_OSI-Data-Governance-OSAI.pdf

[82] *Tumadóttir, A. (2025, April 2). Reciprocity in the Age of AI. Creative Commons.* https://creativecommons.org/2025/04/02/reciprocity-in-the-age-of-ai/

and artists—for instance, noticing the style in which musicians arrange notes, or building on surrealist styles initiated by visual artists. Likewise, scientists and researchers build on past discoveries and the existing literature to gain a better understanding of how the world works and to progress ideas. Human progress is enabled by the ability to build on the past.

This is why we continue to believe that copyright should be balanced in order to facilitate TDM. At its core, TDM is a way to study and analyze existing works, using machines, in order to create new insights and materials. AI training involves forms of TDM. While much of the discourse around TDM as applied to AI has focused on the creation of artistic works, TDM and AI have uses that can help generate advances across science, education, healthcare, and other domains of significant importance to society.[83] In general, we think using existing works in order to derive uncopyrightable elements or make otherwise non-infringing uses should be permissible under copyright law, even if it involves making a copy of a whole work as an intermediate step, such as through TDM.

There are certainly scenarios where AI training and deployment constitutes copyright infringement; the lines here vary by jurisdiction and context and are actively undergoing litigation.[84] However, we know the current state of copyright law around the world does not grant rightsholders universal authority to control use of their works for AI training. In the United States, the doctrine of fair use generally protects analysis of existing works to extract non-copyrightable elements. The European Union (EU) has an exception for TDM for certain research and cultural heritage institutions, while allowing others to perform TDM so long as they abide by specific, machine-readable reservations made by rightsholders.

More to the point, copyright *shouldn't* grant universal authority to control use of works for AI training in all scenarios. This would mean granting a monopoly over ideas, genres, and facts. Expanding property rights risks further concentration of power, both in AI development and beyond. And, given that many content creators and artists sign away their copyrights to large companies, the main beneficiaries of more restrictive copyright laws and licensing deals would be large rightsholders, not creators themselves.

---

[83] *Rucic, H. (2024, April 23). KR21 Principles on Artificial Intelligence, Science and Research. Knowledge Rights 21.* https://www.knowledgerights21.org/news-story/kr21-principles-on-artificial-intelligence-science-and-research/

[84] *OECD. (2025, February 9). Intellectual property issues in artificial intelligence trained on scraped data. OECD.* https://www.oecd.org/en/publications/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data_d5241a23-en.html

Licenses Have Limitations

We're aware of efforts to develop new licenses and contracts, including to shape more responsible AI development (e.g., Responsible AI Licenses[85]), curtail extractive practices (e.g., Post-Open Zero Cost Licenses[86]), and practice digital sovereignty (e.g., Nwulite Obodo Open Data License[87]). Others have proposed ways to adapt CC licenses to address AI training.[88] We are currently treading cautiously when it comes to using licenses or contracts as a path to a new social contract.

To have bite, licenses need an underlying intellectual property right. This is a key aspect of the effectiveness of CC licensing: by design, they apply only when copyright applies, and they do not impose contractual obligations on activity otherwise permitted under law (e.g., via exceptions and limitations to copyright).

Given that AI training currently falls outside the scope of copyright in many scenarios, compliance with the license conditions may not be required when using CC-licensed works for AI training.[89] For example, in a jurisdiction like Japan, where the act of reproducing a work for purposes of AI training is permitted under an exception to copyright law, the CC license would have no effect on that use. Given the wide and varied scope of the exceptions and limitations to copyright law that apply to AI training, CC licenses are not well-designed for imposing reciprocal terms on AI developers.

We are wary of using contract law to fill that gap, at least on the public web. From a functional perspective, contracts are difficult to enforce when access to the information is technically unrestricted. Without the control of copyright or another underlying intellectual property right, a contract requires affirmative agreement between the parties in order to impose enforceable obligations. And, even if agreement can be secured, it wouldn't bind other parties, who may gain access to the data from a different source or further down AI's complex value chain.

---

[85] *Responsible AI Licenses. (n.d.). Responsible AI Licenses (RAIL).* https://www.licenses.ai/

[86] *Post Open. (n.d.). What is Post Open?.* https://postopen.org/about-post-open

[87] *C. Okorie & M. Omino. (n.d.). Licensing African Datasets.* https://licensingafricandatasets.com/

[88] *Szkalej, K., & Senftleben, M. (2024, June 12). Mapping the Impact of Share Alike/Copyleft Licensing on Machine Learning and Generative AI. Open Future.* https://openfuture.eu/wp-content/uploads/2024/06/Share-Alike-and-ML-Report-FINAL.pdf

[89] *Creative Commons. (2025, May 14). Using CC-Licensed Works for AI Training.* https://creativecommons.org/using-cc-licensed-works-for-ai-training-2/

Using contracts to govern what copyright does not also poses ethical quandaries, because it could disrupt the balance struck between free expression and the rights of authors.[90] In cases where copyright does not provide the right to control a given use of a work, it likely reflects a legislated compromise between the interests of creators and the public. Permissionless reuse of copyrighted works plays an important role in the preservation of free expression.[91] This is why CC licenses overlay onto the acts copyright restricts—and not onto acts protected under exceptions and limitations to copyright. This has been a fundamental principle of CC licensing.

CC's approach to forging a new social contract for machine reuse is therefore aimed first and foremost at changing norms.

The Need for a Nuanced Way to Express Preferences

This issue is often framed as a binary—either works should never be used by large AI models without permission, or they should always be allowed.

However, as with the application of the law, the reality of people's notions about what is fair and prosocial when it comes to sharing and reuse of content are more complex. For example, many African natural language processing experts continue to believe in openness and the use of permissive licenses, while feeling that open licensing alone does not respond to their concerns of extractive technology development.[92] There are different feelings within communities, too. Some, but not all, members of the Wikipedia community feel the unrestricted crawling and use of data allows AI companies to unfairly exploit the web.[93]

At present, the opt-out approach for copyrighted works in the EU lacks nuance.[94] We're concerned that bluntness could help bring about the lose-lose scenario of entrenching the

---

[90] *In some jurisdictions, contracts that attempt to control copyrighted works in ways that go beyond the protections afforded by copyright could also face legal obstacles. See e.g., the discussion on copyright misuse in The Mirage of Artificial Intelligence Terms of Use Restrictions by Peter Henderson and Mark Lemley.*

[91] *Netanel, N. (1996, November). Copyright and a Democratic Civil Society. The Yale Law Journal.*

[92] *Okorie, C. & Marivate, V. (2024, April 30). How African NLP Experts Are Navigating the Challenges of Copyright, Innovation, and Access. Carnegie Endowment for International Peace.* https://carnegieendowment.org/research/2024/04/how-african-nlp-experts-are-navigating-the-challenges-of-copyright-innovation-and-access

[93] *Woodcock, C. (2023, May 2). AI Is Tearing Wikipedia Apart. VICE.* https://www.vice.com/en/article/v7bdba/ai-is-tearing-wikipedia-apart

[94] *Senftleben, M. (2025, April 22). The TDM Opt-Out in the EU - Five Problems, One Solution. Kluwer Copyright Blog.* https://copyrightblog.kluweriplaw.com/2025/04/22/the-tdm-opt-out-in-the-eu-five-problems-one-solution/

power of large rightsholders and technology companies, as well as the development of lower quality, biased AI models. It's also clear that the robots.txt protocol, designed to help govern machine access to content in a radically different time, is not currently suited to express the nuanced preferences that people have when sharing their works in an age of large AI models.

> *"How can we programmatically tell an AI crawler how it's allowed to use content on this page? Ultimately, at the moment, there isn't a good way to instruct AI crawlers when it comes to them consuming content. Only a set of suggestions with poor flexibility and even poorer implementation."*[95] - Nick Jackson

One of the founding motivations of CC was to offer more choices for people to share their works and thus expand the commons that we believe human knowledge, creativity, and progress depend on. We've long advocated against all-or-nothing binaries and developed practical tools giving more nuanced options that account for the varied preferences of creators. CC licenses allow creators to indicate not only whether content can be openly shared, but also whether it can be adapted and used for commercial purposes. These permissions are allowed as long as certain common-sense conditions are met, such as providing attribution to the creator.

> *"Large entities… create sandboxes for 'sharing,' but then effectively claim ownership over everything built within that sandbox. This is, in my view, not a sharing economy. It is instead simple sharecropping. The key is to build alternatives that creators on the Internet can use to both create as they wish and keep control of their creativity."*[96] - Lawrence Lessig

This was written in 2006. When once again presented with binary choices that seem to benefit only the few, we aim to create another path that advances the AI ecosystem toward the public interest.

Our Strength Lies in the Collective

Social norms are arguably the single most important aspect of human governance. They dictate how we behave, how we belong, and how we make decisions across almost all aspects of our lives.

---

[95] *Jackson, N. (2025, April 10). Telling AI to go away (but politely). dxw.* https://www.dxw.com/2025/04/telling-ai-to-go-away-but-politely/

[96] *Lessig, L. (2006, October 25). CC Values. Creative Commons.* https://creativecommons.org/2006/10/25/ccvalues-2/

Norms can be powerful, but they require collective action. We're wary of individual creators and collections of content trying to shape the use of their works in their own myriad ways. A single preference, uniquely expressed, *is* inconsequential.

Power comes from coordination and solidarity. The more we converge on preferences and means of expressing them across sectors, communities, and geographies, the more leverage we will have. We intend to create a reciprocal framework that captures widely-held sentiment across the commons.

Together, we can demand a different way.

Introducing CC Signals

We are excited to introduce a first iteration of a preference signals framework, which we're provisionally calling CC signals.

CC signals are designed to offer a new way for stewards of large collections of content to indicate their preferences as to how machines (and the humans controlling them) should contribute back to the commons when they reuse and benefit from using the content. They do not aim to limit or restrict AI development or other types of TDM that machines can undertake. Instead, they are designed to incentivize actions in return.

The idea behind CC signals is simple. Using CC signals, a steward of a collection of content (the 'Declaring Party'), such as a repository of research outputs or scanned books, can express a set of criteria (a 'signal') that would-be users of the content must meet in order to use the content. The criteria are organized around different dimensions of reciprocity, and are intended to drive meaningful, practical action. The framework initially consists of four signals, described in more detail below.

CC signals are designed to be interpretable by machines, as well as humans. As we describe in the following sections of this report, this means CC signals leverage external technical standards and protocols to support interoperability and ensure effectiveness at scale.

CC signals are designed as global tools, which means they operate across different legal systems. As a result, applying a CC signal is likely to have a different legal effect depending on who applies it and in what context. Where copyright exists and is applicable, CC signals are intended to leverage the power of copyright without increasing its power. This is not about creating new property rights; it is more like defining *manners for machines*.

The Suite of CC Signals

Our work on CC signals is driven by the goal of increasing and sustaining public access to knowledge, as well as our belief that openness and responsibility can co-exist.

We've drafted four 'signal elements' for public feedback, designed to reflect core elements of the overarching theme of reciprocity. They are: credit, direct contribution, ecosystem contribution, and open.

***Credit**: You must give appropriate credit based on the method, means, and context of your use.*

Attribution and provenance in the context of large AI models is complex, difficult, and rapidly evolving as technologies develop. However, this does not mean that the concept of credit should be seen as irrelevant or impossible in the context of AI. We seek to establish norms around what *is* possible, not letting the perfect be the enemy of the good. Like the attribution condition in the CC licenses, we imagine the credit signal element being enacted in any reasonable manner. We plan to develop guidance and best practices around credit in future stages of this work, drawing on the progress being made in this area by others in the field. For now, at a minimum, we expect this signal to require citation of the training dataset by the reuser. For techniques that enable models to retrieve information in response to queries, such as retrieval augmented generation (RAG), and other use cases where it is technically feasible to connect content with particular outputs, outputs must cite the collection as a source with a link.

***Direct Contribution**: You must provide monetary or in-kind support to the Declaring Party for their development and maintenance of the assets, based on a good faith valuation taking into account your use of the assets and your financial means.*

This is not intended as a commercial transaction. It is designed to create a structure for financial or in-kind contribution to support the sustainability of the Declaring Party. The application of CC signals should not be seen as a business model, or even a way to reliably recoup costs. The contributions are intended to be proportionate, both to the particular type and scale of machine reuse, and to the financial means of the party undertaking it. As with credit, we plan to produce guidance and best practices for direct contribution as CC signals develop.

***Ecosystem Contribution**: You must provide monetary or in-kind support back to the ecosystem from which you are benefiting, based on a good faith valuation taking into account your use of the assets and your financial means.*

This is designed to spur contributions that support the commons as a whole. While the initial phrasing is very open-ended, we hope and expect that norms, best practices, and even new, collective-minded structures could grow around this notion in different sectors and for different types of reuses. The aim is to encourage a practice of giving back, infusing a norm of reciprocity in ways that will help sustain the ecosystem for all.

***Open**: The AI system used must be open. For example, AI systems must satisfy the Model Openness Framework (MOF) Class II, MOF Class I, or the Open Source AI Definition (OSAID).*

This signal element reflects the fact that making AI models open—by releasing model weights, code, or datasets for others to use and build on—is a form of reciprocity.[97] Given the progress made by others in the field to provide meaningful definitions of openness, our proposal for this signal is more specific about what is required to adhere to it.

For the sake of simplicity, we have proposed four particular combinations of signal elements that would serve as the suite of CC signals. The proposed suite of CC signals are:

Credit

Credit + Direct Contribution

Credit + Ecosystem Contribution

Credit + Open

Credit is included as an element in each signal because we believe it is a fundamental form of reciprocity; one that benefits the broader knowledge cycle. In this initial proposal, the other signals are mutually exclusive. The list of signals is intentionally limited so that stewards of large collections and their communities can align in calling for their adoption with AI developers. This will ultimately help to build the collective action needed to drive reciprocity within the AI ecosystem.

[97] *Tumadóttir, A. (2025, April 2). Reciprocity in the Age of AI. Creative Commons. https://creativecommons.org/2025/04/02/reciprocity-in-the-age-of-ai/*

At this phase of CC signals' development, the signal elements and their combinations are preliminary. Through a period of public consultation, we expect the CC signals and conditions they state will shift and evolve, based on what is technically possible and what is desirable to stewards of content.

Scope of Machine Reuse Addressed by CC Signals

Defining a standard set of categories for machine reuse of content that adequately captures the different issues at play here is difficult. For example, though the generative capability of large AI models is the source of issue for some creators and publishers, we think the need for reciprocity extends beyond generative AI only.

Fortunately, there are expert stakeholders doing important work in this space. The Internet Engineering Task Force (IETF) AI Preferences Working Group is leading on work to ensure that any approaches used to declare preference signals are standardized and machine-readable, and therefore effective at scale.[98] The IETF AI Preferences Working Group recently adopted a common vocabulary for opting out of AI training and other forms of TDM, originally put forward by Open Future.[99]

We are proposing to tightly integrate CC signals with the work of the IETF AI Preferences Working Group. In practice, this means that Declaring Parties will be able to select a category of machine use from this vocabulary to apply their preferences to. These categories, which are still evolving, include—and therefore enable Declaring Parties to distinguish between—TDM, AI Training, and Generative AI Training. Once selected, that category of machine reuse is the scope of the CC signal attached by the Declaring Party.

In general, our goal is to work in collaboration with the IETF AI Preferences Working Group to ensure that CC signals uses standard components and approaches. This will reduce the complexity for Declaring Parties, and help responsible AI developers comply with preferences made across different systems.

Declaring a Preference Using CC Signals

The Declaring Party selects among the available CC signals. Once selected, that signal reflects the Declaring Party's preferences regarding machine reuse. This means that the

---

[98] *Internet Engineering Task Force. (n.d.). AI Preferences (aipref). Internet Engineering Task Force. https://datatracker.ietf.org/wg/aipref/about/*

[99] *Keller, P. (2025, March 7). A vocabulary for opting out of AI training and other forms of TDM. Open Future. https://openfuture.eu/publication/a-vocabulary-for-opting-out-of-ai-training-and-other-forms-of-tdm/*

Declaring Party says that the selected category of machine reuse is allowed under the terms of the particular signal elements.

Similar to the CC licenses, CC signals will be both machine- and human-readable. The human-readable explanation of what happens when a signal is applied is called a 'declaration.' There will be a declaration for each signal, with variations based on whether the Declaring Party has copyright authority and the particular scope of machine reuse selected. We plan to explore whether it would be useful to develop a tool that helps Declaring Parties build a standardized declaration, similar to the *Choose a License for Your Work* tool for selecting an appropriate CC license.[100] These declarations could be purely an explanatory device for users, or they could create legal documents in some circumstances. These and other implementation questions will be addressed and iterated upon in the coming months.

Based on our initial proposal, we intend for CC signals to primarily be attached to large collections of content that are published openly on the web, and are being, or would be, accessed using crawlers and similar technologies. However, we're keen to understand whether CC signals could also be used in connection with other data sharing or publication techniques, such as via Application Programming Interfaces (APIs).

The string of machine-readable code used to apply a CC signal to a dataset will be called a 'content usage expression'. We are proposing that the machine-readable version of the CC signals extend the attachment methods specified by the IETF AI Preferences Working Group. Our proposed technical specification for this is available for public comment and input on GitHub: https://github.com/creativecommons/cc-signals.

Who Applies the Preference Signal

A Declaring Party is someone who has both ethical and legal authority to specify how a content collection should be used by machines. Sometimes, the Declaring Party will hold copyright or have authority to represent rightsholders of the content. In these cases, a CC signal may have legal effect depending on the particular jurisdiction.

We are focusing on supporting CC signals to be applied by Declaring Parties that are stewards of large collections of content because it is the most efficient way to shape new norms across the ecosystem.

An individual work is infinitesimally small in the context of modern machine reuse, so we do not believe applying preference signals to individual works—or the unit level—would be a practical starting point.

---

[100] *Creative Commons. (n.d.). Choose a License for Your Work. Creative Commons. https://creativecommons.org/chooser/*

This does not mean the preferences of individual creators are unimportant. In many cases, a content collection will include works by different contributors. There are different ways in which stewards of content collections can use CC signals to give force to the expectations of individual creators, and we're eager to engage further about this as part of our public consultation process. We're observing efforts to understand and define community preferences with keen interest, such as Holly Herndon and Mat Dryhurst's recent work with Serpentine Arts Technologies to bring groups of choirists together to determine how they'd like their works to be used by generative AI models.[101]

Conceivably, stewards of content could enable individual contributors to select the signals they want associated with their contributions, similar to the approach being taken by the social media platform, Bluesky, to enable users to express their own preferences regarding the reuse of their public posts.[102] Using this approach, the full collection would then be divisible into different datasets with different combinations of signals.

The Relationship between CC Signals, Copyright, and CC Licenses

As previously stated, we are not conceiving of CC signals as copyright licenses.

As *manners for machines*, the signals are primarily designed to address the social layer of data governance. However, when applied by a Declaring Party with copyright authority over the content, CC signals are likely to have legal implications under copyright law. The precise effect will depend on the jurisdiction. For example, in the EU, there is a copyright exception for TDM, including AI training, which can be overridden if a rightsholder "reserves their rights" or opts out. A CC signal applied by a rightsholder in this context is likely to be considered such an opt-out, in which case it could impact the ability of reusers to rely upon that copyright exception. This could mean the CC signal then functions as a form of copyright permission, granting the right to use the asset under the terms of the particular signal.

Even in jurisdictions without an opt-out regime like that in the EU, a CC signal applied by someone with copyright authority could have copyright implications. In situations where a reuser needs or wants copyright permission for machine reuse, the CC signal may give that conditional permission depending on who applies it. Further research and analysis about the legal implications of CC signals will be a major focus of our efforts in the coming months.

---

[101] *Ding, J., Jäger, E., Ivanova, V., & Bunz, M. (2024). My Voice, Your Voice, Our Voice: Attitudes Towards Collective Governance of a Choral AI Dataset. ArXiv.org.* https://arxiv.org/abs/2412.01433

[102] *bluesky-social. (n.d.). proposals/0008-user-intents/README.md. GitHub.* https://github.com/bluesky-social/proposals/blob/main/0008-user-intents/README.md

The CC signals are intended to operate in a way that is separate from, and complementary to, the CC licenses. When a CC signal is applied to CC-licensed content by the copyright holder, the signal would grant copyright permission for the selected category of machine reuse under different terms than the CC license. An AI developer could technically rely on either the CC license or the CC signal if copyright permission is needed in that context or jurisdiction. However, since the signal is designed specifically for machine reuse, it is likely to be a more accurate reflection of the wishes of the rightsholder(s) in that context.

As this work continues, and before we move ahead with an implementable version of CC signals, we will undertake an in-depth analysis of the potential interactions with the licenses, and produce guidance on how they can be used together effectively.

Incentivizing Adherence by AI Developers

We recognize that CC signals will rely on willing participation by AI developers to adhere to it.

There are many reasons to be cynical about adherence, particularly when it is not legally required, and there are and will always be bad actors. However, we see many reasons to believe that uptake is likely.

For one thing, there is precedent. Although adherence hasn't always been perfect, robots.txt functioned for many years as a way to encode normative expectations about—and help maintain the social contract for—machine reuse of content on the web. We also see the success of CC licensing as evidence that voluntary buy-in is possible. While CC licenses are built atop copyright law and therefore carry the weight of copyright infringement risk, in reality they work because people have *chosen* to adhere to them. Litigation involving enforcement of CC licenses is rare, and much of it involves litigants who are not operating in good faith.[103] Instead, there are now tens of billions of CC-licensed works available in the commons because they are grounded in intuitive notions about what is fair and prosocial when it comes to sharing and reuse of knowledge.

There are also clear reasons why rational actors should respect and adhere to preference signals. As we've written earlier in this report, data from across the public web is a key component in developing large AI models. If those developing AI do not respect the wishes of creators, they risk eliminating incentives for people to share and widely distribute their works. Over time, this will compromise the accuracy, safety and currency of the models and services they build. This will be particularly acute for small firms, startups, nonprofits, and

[103] *Creative Commons. (n.d.). License Enforcement. Creative Commons. https://creativecommons.org/license-enforcement/*

academic researchers, who would not have the resources to instead rely on costly licensing deals.

In early workshops about this work, we brainstormed a wide range of tactics we could use to stimulate further demand. On the positive side, we could develop a CC-designed badge, a certification system, or other methods for AI developers who adhere to CC signals to use to indicate their participation in this prosocial system. Alternatively, or perhaps in parallel, we could leverage more aggressive tactics like publicly identifying AI developers who do not adhere to CC signals.

Working in Tandem and in Partnership

CC signals is one element of the path towards a more equitable future.

As we've acknowledged before, preference signals are by themselves not sufficient to help sustain the commons,[104] and other interventions will be required to grow it going forward. There remains much work to do to increase transparency around the data used to train models, as well as to reinforce the technical infrastructure that underpins large collections of content on the web. Wikimedia Enterprise, for example, demonstrates how developing commercial-grade APIs for Wikipedia and other sources of knowledge can help ensure that high-quality open data can be sustained for everyone.[105]

We're also conscious that CC signals won't address broader issues concerning the distribution and use of data. These issues, such as risks to data protection and safety, must be tackled by other means, including by choosing not to share something at all if the risk of misuse or abuse is particularly high.[106]

Safeguarding What Matters

Let's not forget what we are protecting. A thriving creative commons belongs to all of us, including humans using machines to generate insights and discoveries. But a creative commons will not thrive on extraction or neglect. It requires care, reciprocity, and intention. Call it stewardship, a circular economy, or regeneration, the principle is the same: the commons must be replenished by the collective it serves.

---

[104] *Timid Robot. (2023, 31 August). Exploring Preference Signals for AI Training - Creative Commons. Creative Commons.* https://creativecommons.org/2023/08/31/exploring-preference-signals-for-ai-training/

[105] *Wikimedia Enterprise. (n.d.). Learn more about Wikimedia Enterprise. Wikimedia Enterprise.* https://enterprise.wikimedia.com/about/

[106] *Downing, K. (2023, July 13). AI Licensing Can't Balance "Open" with "Responsible."* https://katedowninglaw.com/2023/07/13/ai-licensing-cant-balance-open-with-responsible/

This is not protectionism in response to innovation. A reciprocal commons is a catalyst for innovation. In just over two decades, the CC licenses have enabled open access to almost 50% of all published scientific research. Tens of millions of cultural heritage items are openly available through institutions like Europeana, which alone hosts over 25 million open access items.[107,108] Today, tens of billions of CC-licensed works circulate within the commons, supported by a global movement dedicated to expanding open knowledge.

Like natural commons—forests, fisheries, water systems—the digital commons depends on governance, shared values, and sustained cooperation. These ecosystems survive not because they are infinite but because communities agree to nurture them. The digital nature of our commons doesn't change that. When we treat openness as a one-way street, even the most abundant resource can run dry.

What we aspire to co-create, in part through CC signals, is a future where *extractive* no longer feels like the most apt description of the relationship between AI and the humans who create the knowledge that develops and feeds it. Benefiting from the commons at the scale of AI should come with an obligation to give back. The good news is that the options for meaningful replenishment are abundant in and of themselves.

Reciprocity means we collectively have a broad sense that the benefits of training machines on the collective intelligence of humanity are equitably shared. It means there is transparency about how AI systems work and what data they use, with strengthened norms around provenance and attribution. These are signals of respect and renewal, and this is our vision for a thriving creative commons in the age of AI.

We stand at a generational crossroads in how we engage with human knowledge. We can either shape a reciprocal relationship between AI and the commons, or we can risk watching decades of open progress erode.

> *"The hardest thing about pushing the work of Creative Commons is the thought that in 15 years, it will be impossible to explain just why this work was important — either because the worst would have happened, and the technologies that have encouraged the explosion of creativity we see just now will have been re-controlled, or because the best would have happened,*

[107] *Heritage, C. (2024, February 4). Sharing Cultural Heritage Data. ENIGMA EU.* https://eu-enigma.eu/2024/02/04/perslab-sharing-cultural-heritage-data/

[108] *McNeilly, N. (2024, September 19). Measuring the impact of reuse of digital heritage. europeana.* https://pro.europeana.eu/post/measuring-the-impact-of-reuse-of-digital-heritage

> *and the balance that we're pushing for will have been achieved, in both practice and law."*[109] - Lawrence Lessig

Just as was posited by Lawrence Lessig 19 years ago, the work of collectively building a reciprocal AI ecosystem and defending a thriving creative commons is nuanced and complex—but the results are fairly black and white. Just as CC licenses were one part of a complex web that increased global access to knowledge, so too are CC signals. We hope they'll be a seed for collective action, sustainability of open infrastructure, data-sharing frameworks, and yet unimagined innovations.

Building that balance won't come from any one intervention alone. We still believe that open sharing online has a vital role to play in the future of our information ecosystem. But, we must act now to shape the terms of AI's engagement with the commons.

This is our chance to build a future where creativity, knowledge, and technology serve one another, and all of us. Join us.

Now, over to you!

Head over to the CC signals GitHub repository to provide feedback and respond to our discussion questions: https://github.com/creativecommons/cc-signals.

Visit the CC website for more information on AI and the Commons: https://creativecommons.org/ai-and-the-commons/.

Donate to creative commons ↗

[109] *Lessig, L. (2006, December 28). CC's Future. Creative Commons. https://creativecommons.org/2006/12/28/ccs-future/*