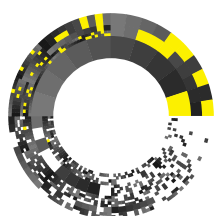


# AI \_COMMONS

*Filling the governance vacuum related to the use of information commons for AI training*



**OPEN  
\_FUTURE**

**JANUARY 2023**

Alek Tarkowski, Zuzanna Warso

**TEN YEARS AGO,  
AI RESEARCHERS  
STARTED SCRAPING THE  
INFORMATION COMMONS  
AND USING THIS CONTENT  
FOR FACE RECOGNITION  
TRAINING.**

**A GOVERNANCE VACUUM  
WAS CREATED FOR THE  
UNEXPECTED USES  
OF OPENLY LICENSED  
CONTENT.**

**LESSONS LEARNED FROM  
THIS CASE CAN HELP  
TO BETTER GOVERN AI  
DATASETS TODAY.**

# TABLE OF CONTENTS

<b>4</b>	<b>_INTRODUCTION</b>
<b>5</b>	<b>_THE GIST OF THE AI_COMMONS CASE</b>
<b>8</b>	<b>_GOVERNANCE OF THE AI_COMMONS LIFE CYCLE</b>
<b>12</b>	<b>_TWO MODELS OF GOVERNANCE FOR AI DATASETS</b>
<b>14</b>	<b>_FINDINGS &amp; RECOMMENDATIONS</b>
<b>16</b>	<i>Issues related to text and data mining, and web scraping</i>
<b>17</b>	<i>Issues related to state regulation</i>
<b>18</b>	<i>Issues related to privacy and personality rights</i>
<b>18</b>	<i>Issues related to research ethics</i>
<b>21</b>	<b>_DESIGNING A COMMONS-BASED GOVERNANCE MODEL FOR AI</b>
<b>23</b>	<b>_ABOUT</b>

# **\_INTRODUCTION**

*This report presents findings and recommendations from an investigation into using openly licensed photographs for AI facial recognition training datasets. With the AI Commons activity, we have been exploring how AI training datasets, and works included in those datasets, can be better governed and shared as a commons.*

The case creates an opportunity to ask fundamental questions about the challenges that open licensing faces today, related to privacy, exploitation of the commons at massive scales of use, or dealing with unexpected and unintended uses of works that are openly licensed.

While events that form this case go back almost a decade, these issues are still relevant. Through the AI Commons project, Open Future Foundation wants to contribute to a collective exploration of solutions to these challenges. There are findings and lessons learned from this case that can improve the governance of AI datasets – as the old ones continue to be used and new datasets are designed and deployed.

Today, there is a need to establish commons-based governance models for AI training datasets, and other elements of the AI technological stack. The case also creates an opportunity to review open-sharing frameworks (and open licensing in particular) and to make them more future-proof.

As part of the AI Commons activity, we have commissioned Adam Harvey to conduct [a study on the use of Creative Commons licenses for AI training datasets](#), and Selkie Study to research [the use of openly licensed photographs and machine learning](#). Furthermore, Aniek Kempeneers has conducted [a study of design solutions](#) for the case, as her MSc graduation project in the [DCODE Labs](#) at the Delft University of Technology. We have also published an in-depth [white paper](#) on understanding the implications of face recognition training with CC-licensed photographs

The authors want to thank experts who have contributed their ideas and feedback to this research project: Peter Cihon, Jennifer Ding, Carlos Muñoz Ferrandis, David Kanter, Jennifer Lee, Mike Linksvayer, Ben MacAskill, Roger MacDonald, Jacob Rogers, Cari Spivack, Paul Stacey, Barry Threw, Luis Villa, Kat Walsh.

# **THE GIST OF THE AI COMMONS CASE**

In 2002, twenty years ago, the Creative Commons licenses were created. These legal tools provided standardized means for content sharing through limited, flexible copyrights.

In 2004 Flickr became one of the first social media platforms and the go-to place for publishing photos on the Web. It was one of the early adopters of Creative Commons.

By 2014, there were almost 400 million CC-licensed photos on Flickr. That year, researchers from Yahoo Labs, Lawrence Livermore National Laboratory, Snapchat and In-Q-Tel used a quarter of all these photos to create YFCC100M, a dataset of 100 million photographs of people created for computer vision applications.

Until today, this dataset remains one of the most significant examples of openly licensed content reusing. Because of the massive scale and the productive nature of the dataset, it became one of the foundations for computer vision research and industry built on top of it.

The YFCC100M dataset has set a precedent, followed by many other datasets. Some are designed as samples of the original one, others copying its approach to content provenance. Many of them became standardized tools used for training facial recognition AI technologies.

In 2019, research by Adam Harvey put the spotlight on MegaFace, a dataset created by a consortium of research institutions and commercial companies as a derivative of the YFCC100M dataset. MegaFace includes 3 million CC-licensed photographs and is the most relevant dataset for face recognition research, benchmarking and training. Harvey's research presented the dataset as a privacy-invading tool consisting of photos of individuals used without their consent.

MegaFace became exemplary of the tension between the open sharing of photographs of people – with tools like the Creative Commons licenses – and potential harms, mainly related to privacy violations and extractive use of personal data.

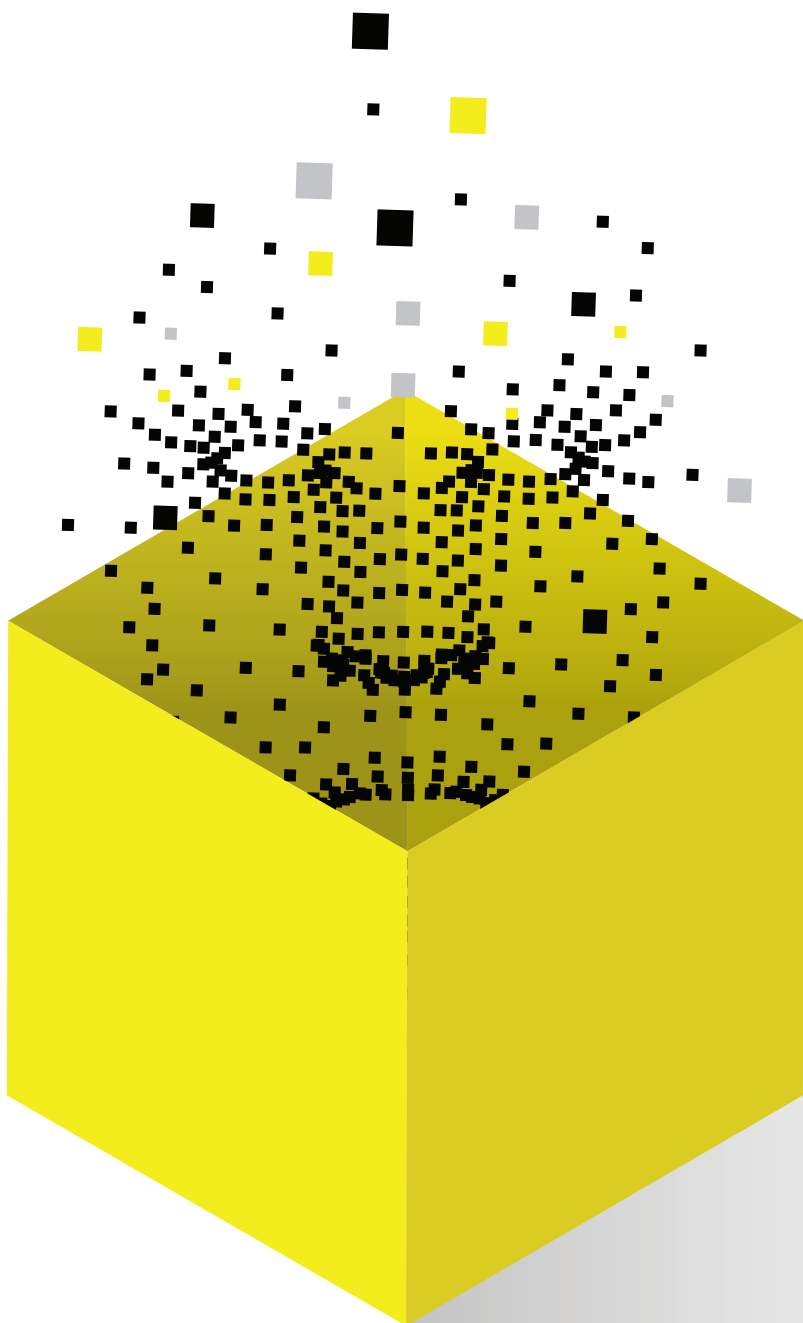
For the open movement – actors who contribute to resources based on non-exclusive forms of intellectual property ownership and advocate for these forms – the MegaFace story illustrated new challenges that open sharing faces in a changed online environment.

While the case seemed not to involve any use that violated the licensing conditions, it did illustrate the limits of copyright licenses for the use of images that also included personality rights. It forced stewards of open licensing to consider issues beyond the remit of copyright law and the ethical aspects of open licensing.

The media picked up the story of the use of openly licensed content in datasets that serve facial recognition training. Stories like the MegaFace case became a symbol of potential harm that can be a side effect of open sharing.

In 2022, the major datasets built with CC-licensed content are still in use. Over the years, these datasets were used to train facial recognition models that were later used in hundreds of projects, including the development of military technologies or surveillance solutions. It is time to find ways to manage both the open resources and the AI solutions built on top of them in a way that is more sustainable and reduces harm.

Through the AI\_Commons project, Open Future Foundation wants to contribute to a collective exploration of solutions to these challenges. There are findings and lessons learned from this case that can improve the governance of AI datasets – as the old ones continue to be used, and new datasets are designed and deployed.



1

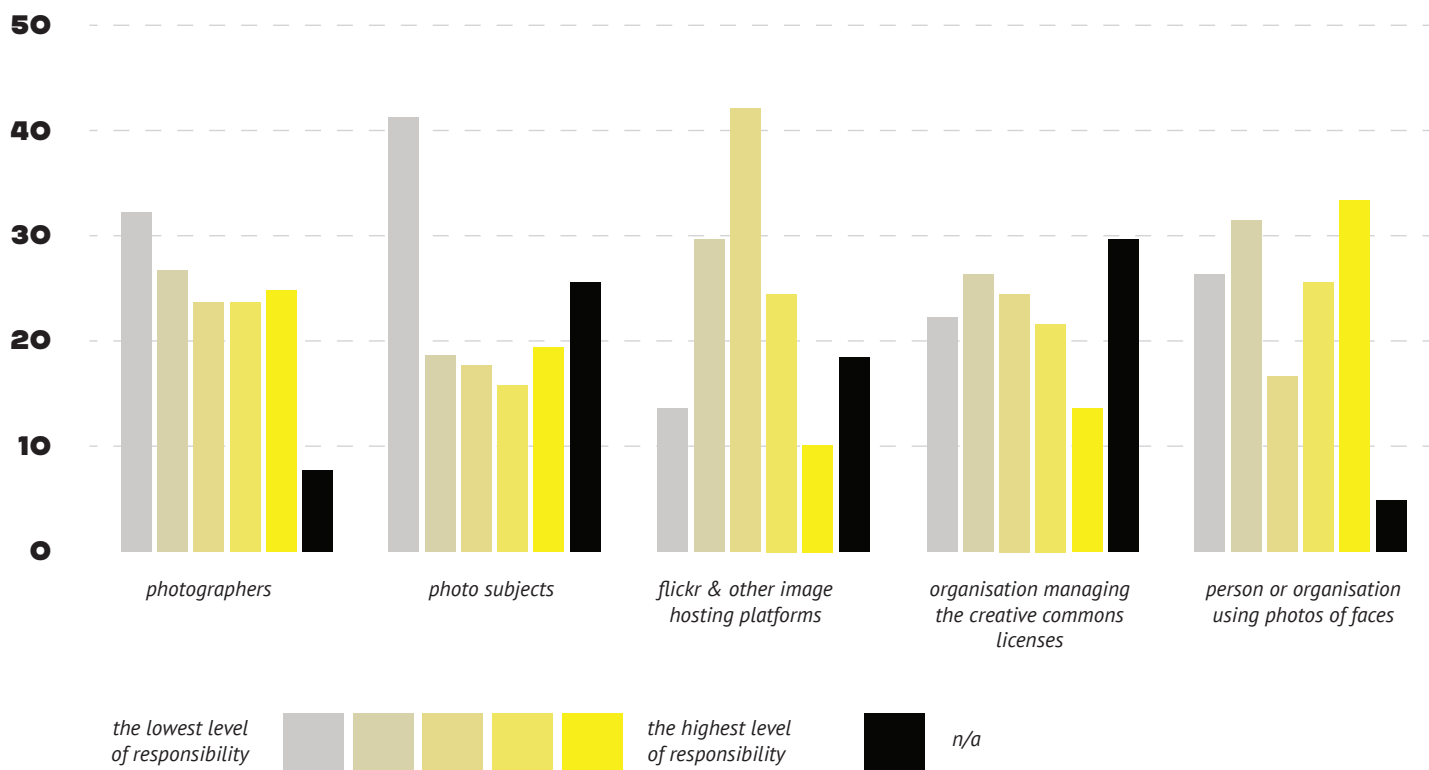
**GOVERNANCE OF  
THE AI COMMONS  
LIFE CYCLE**

# GOVERNANCE OF THE AI COMMONS LIFE CYCLE

“Who dropped the ball?” asked one of the attendees of a workshop in San Francisco as we discussed the issue of training face recognition AI with openly licensed photographs available online (“the case”). The response was: “Everyone did.”

This assessment may be too harsh; in the early 2000s, hardly anyone could have predicted that pictures posted online would be used to develop AI models for biometric surveillance and military applications. Nonetheless, it accurately summarizes our situation twenty years later. It does not just emphasize that something went wrong, but rather points to the fact that there is no single party to blame.

This realization is reflected in the results of a survey that we have conducted of users who share openly licensed photos on online platforms such as Flickr or Wikimedia Commons. The survey has revealed that the users are unsure who should be held accountable for using photos of faces that are shared on image hosting platforms.



**Figure 1.** In your opinion, who should take responsibility for the use of photos with faces, which are shared on imagehosting platforms? (in particular for negative outcomes of such use). Please rate on a scale where: 1 is the lowest level of responsibility and 5 is the highest.



The phrase “everyone dropped the ball” also neatly defines our current reality: that there have been missed opportunities to limit or stop the misuse of content. Over nearly a decade, sufficient governance mechanisms for openly licensed AI training datasets have yet to be developed.

The case that was the starting point for our discussion involved open licensing frameworks as well as issues outside the purview of copyright regulation. The harms and unethical uses are linked to privacy and digital rights. There are concerns about the responsibilities of researchers and developers who use publicly available data to build AI. Copyright measures are not the best way to address these issues, and open licensing is not the best way to safeguard digital rights and uphold ethical behavior.

To address the factors that contributed to the misuse of open content, we need a governance model that covers the various stages of the content and data life cycle and addresses the various concerns. By governance, we mean coordinated actions of multiple actors using multiple instruments, methods, and strategies that, when combined, give rise to rules and norms and ensure their implementation. We use the term “governance” to refer to various methods of establishing and enforcing norms, which include more than just legal frameworks, and to emphasize that different parties must contribute to achieving the desired results while mitigating risks or harms.

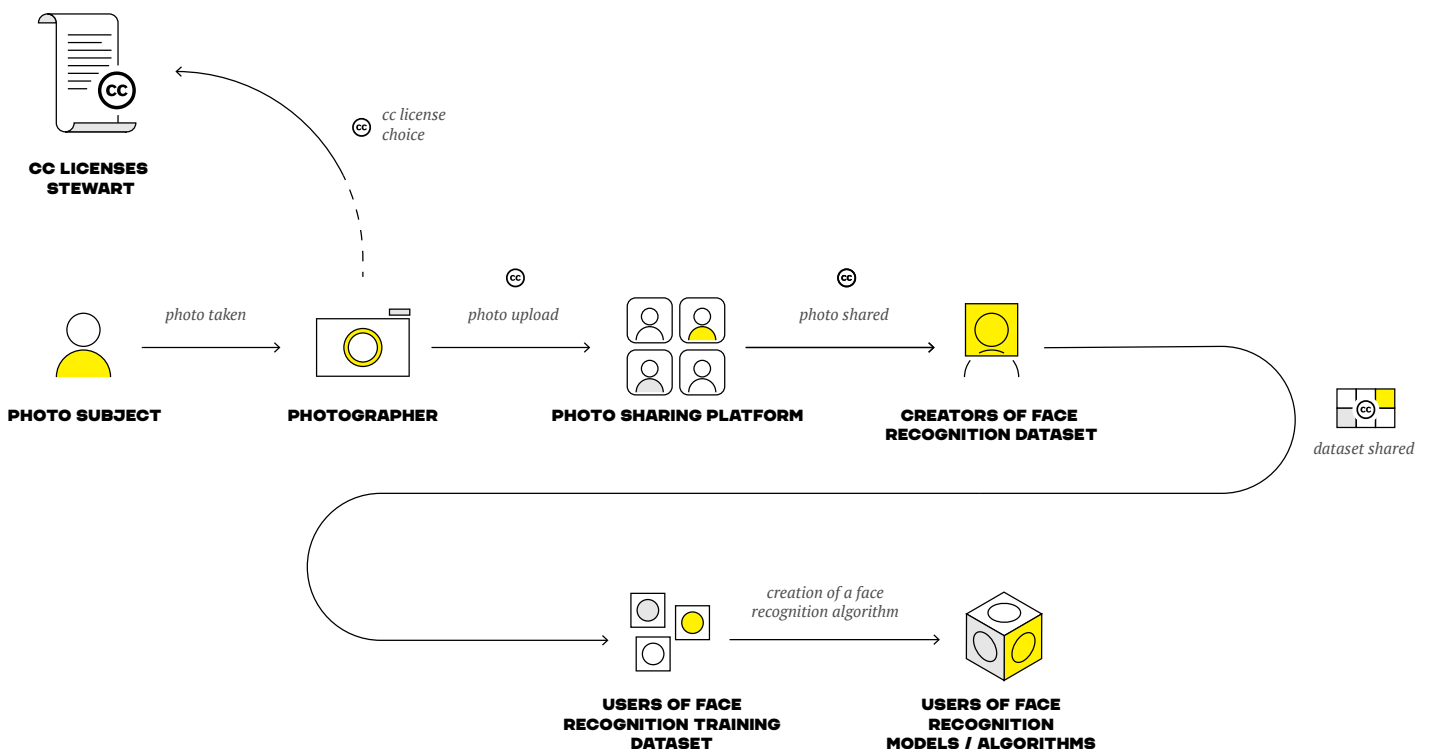


Figure2. The AI\_Commons life cycle.

On the one hand, decisions taken by creators of photographs, photo-sharing platforms, and stewards of open licensing frameworks set downstream conditions in the life cycle: those for creating and managing AI datasets, and then AI models and AI systems. For this reason, in recent years, there has been a growing awareness that the stewards of open sharing must consider challenges that may arise later when the content is used in a novel context. Even if these challenges fall outside the scope of copyright order.

This awareness is well expressed by Creative Commons, in the organization's [2020-2025 strategy](#) is based on a vision of better sharing<sup>1</sup>:

*“In order to protect what we have achieved so far and to create the world we want to see, we must expand our focus beyond copyright licensing, because content sharing cannot be decoupled from economic or ethical concerns.” The vision calls for expanding beyond “open sharing,” to pursue “a commons that serves the public interest.”*

Decisions made by dataset creators play a crucial role in the overall governance process. The history of image recognition and face recognition datasets, including the most popular – and infamous – ImageNet dataset, is one of negligence in properly governing them. It is surprising to see how, over the years, various academic and corporate actors have failed to meet different standards of legal compliance and good governance as they reused content from the information commons, and turned it into AI training datasets.

One of the main findings of our study is the existence of a governance vacuum in the process of creating and using datasets for AI training. This vacuum is the result of actions (or lack of) of multiple actors across the life cycle. The main recommendation from our research and consultations is the need to improve the governance of such datasets.

Before proceeding, to better frame this challenge, we should make two qualifications. Firstly, dealing with bad actors and reducing harm is not the same as incentivizing good actors to establish best practices and good dataset governance. Secondly, cases of obvious harm should be distinguished from those in which there is a sense of potential harm, or the issue is one of the unexpected uses.

<sup>1</sup> Catherine Stihler (December 16, 2020): “Announcing Our New Strategy: What’s Next for CC - Creative Commons.”

**2**

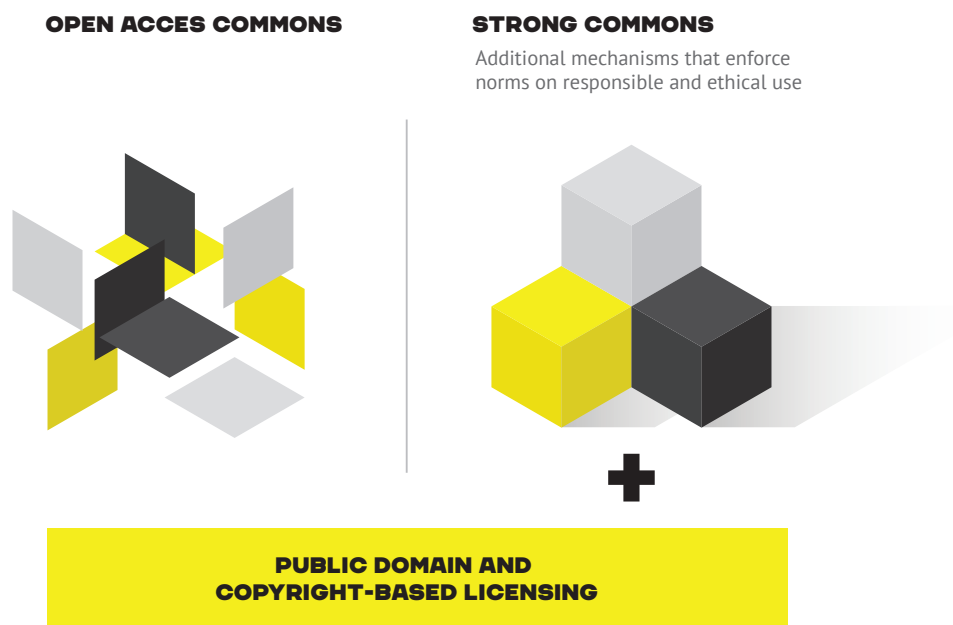
**TWO MODELS  
OF GOVERN-  
ANCE FOR AI  
DATASETS**

# TWO MODELS OF GOVERNANCE FOR AI DATASETS

The pool of openly licensed content, or the information commons, has traditionally been understood as not requiring governance mechanisms beyond open licenses' limited, minimal rules. This approach has aimed to ensure the most efficient and unrestricted sharing and use of resources. This kind of minimal governance structure can be described as Open Access commons<sup>2</sup>. The challenges raised by the case indicate the need to look beyond this approach, which is at the heart of many open-sharing frameworks and initiatives.

An alternative process entails introducing additional rules and limits to reuse while adhering to the general vision of the commons. Such a more robust form of commons-based ordering has often not been considered a form of openness in the past<sup>3</sup>. It involves harnessing non-copyright-based tools, introducing other legal mechanisms, and considering social norms and ethics. This approach can be understood as a stronger form of commons-based governance. It is characterized, in particular, by more granular and specific access and use rules that serve to harden the commons against abuse and negative externalities.

The conversations and consultations that we have conducted show a strong consensus that stronger forms of commons-based governance are necessary in the case of AI training datasets.



<sup>2</sup> We borrow this term from Balázs Bodó, who used it to distinguish Open Access from other, stronger forms of commons-based governance. See: Balázs Bodó, "Was the Open Knowledge Commons Idea a Curse in Disguise? – Towards Sovereign Institutions of Knowledge." SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, December 11, 2019), <https://doi.org/10.2139/ssrn.3502119>.

<sup>3</sup> Tarkowski, Alek, and Jan J. Zygmuntowski, "Data Commons Primer." (Open Future Foundation, September 2022), <https://openfuture.eu/publication/data-commons-primer/>.

# 3

## **FINDINGS & RECOMMEN- DATIONS**

# **FINDINGS & RECOMMENDATIONS**

As noted above, the case is one where the focus needs to be expanded beyond copyright licensing. For this reason, we have been structuring the conversations about AI Commons around several lenses, through which the case can be analyzed: copyright, privacy, other regulatory measures and research ethics. Each lens offers different insights into how to design a stronger commons-based governance model for AI training datasets.

## *Issues related to copyright and Creative Commons licensing*

### **1. COPYRIGHT IS THE BASELINE.**

While the case concerns copyrighted content shared under open licenses, the challenges it raised fall beyond the scope of copyright law. Copyright rules, including open licensing frameworks, should be viewed as “the floor,” or a starting point for further governance; other modes of ordering and regulating behavior must also be considered. In particular, most of the issues framed as responsible or ethical AI fall beyond the scope of copyright law, and copyright-based tools are not well suited to address unethical, harmful, or irresponsible use of the content for AI training.

Nonetheless, there is the question of whether actors such as content-sharing platform administrators and open licensing framework stewards are in any way responsible for aspects and outcomes of sharing that fall outside the scope of copyright law order. Should they be expected to deal with issues caused by other actors, such as academic and corporate entities that create and use AI training datasets?

In its [opinion on the case from 2021](#), Creative Commons acknowledged that “The legal uncertainty caused by ethical concerns around AI, the lack of transparency of AI algorithms, and the patterns of privatization and enclosure of AI outputs, all together constitute yet another obstacle to better sharing.<sup>4</sup>”

If better sharing is the goal, stewards of open sharing frameworks must contribute to overcoming the barrier mentioned above, even if copyright tools are not the only ones to address the issue. Other components of the governance model include privacy and data protection legislation, research ethics that should guide the researchers and developers, and social (or community) norms.

4 Creative Commons, “Should CC-Licensed Content Be Used to Train AI? It Depends.” (March 4, 2021).

## **2. NEED FOR MORE FINE-GRAINED CONTROL OVER OPEN SHARING**

There are numerous proposals for dealing with unexpected uses of open works. These proposals concern uses that are formally in accordance with the licensing terms but violate social norms – whether those upheld by the licensor, the community, or society as a whole.

While the motto “permission given in advance” is a core tenet of open licensing frameworks, some users of these licensing tools have questioned this principle over the years. In parallel, uses have emerged that showed difficulties in providing such broad permission. The case we have been studying is a prime example of such a situation.

Face recognition training is also a novel example of a prominent use of open works in the two-decade history of open licensing, widely regarded as leading to potentially harmful – though not necessarily illegal – uses. Previously, there were always concerns about the misuse of open works, but there were no significant real-world examples.

The concept of more fine-grained control of uses differs from traditional methods of introducing constraints, such as licensing clauses that limit entire categories of uses – the most common example being the Non-Commercial condition of some Creative Commons licenses.

In one scenario, these conditions would be personalized by each entity sharing their works through tools that allow them to express motivations for sharing. This path is currently being explored by the new [responsible AI licenses](#) introduced in 2022, and their creators envision a proliferation of licensing conditions chosen by licensors<sup>5</sup>.

Another proposed solution would entail technical means of tracking uses of openly shared works, which could benefit from advances in web technologies. These tools would make downstream (re)uses of content traceable and, therefore, legible to licensors, creating opportunities to interact with users and potentially enforce additional norms.

Current sharing models divorce people from the works they share, as through the process of reuse, they acquire a “life cycle.” Creators, however, might want to have ties with their works throughout these life cycles. Jeni Ténisson has explored this relational aspect of open sharing in her essay on “Creative Communities.”<sup>6</sup>

Finally, there are emerging ideas around opt-out mechanisms for AI training datasets. These are coupled with the emergence of search tools that allow the licen-

5 Danish Contractor, Daniel McDuff, Julia Haines, Jenny Lee, Christopher Hines, and Brent Hecht, “Behavioral Use Licensing for Responsible AI.” In 2022 ACM Conference on Fairness, Accountability, and Transparency, 778–88, (2022), <https://doi.org/10.1145/3531146.3533143>.

6 Jeni Ténisson, “Creative Communities,” (Open Future, 2022). <https://openfuture.eu/paradox-of-open-responses/creative-communities>.

sor to understand better whether and how their works have been included in such datasets. Adam Harvey has pioneered this approach through the [exposing.ai](#) project, and [Have I Been Trained?](#) (created by spawning.ai initiative) is another recent example of this approach. It must be noted that the effectiveness of any opt-out mechanisms will be limited by the proliferation of copies of open datasets.

### **3. LICENSE COMPLIANCE AND ENFORCEMENT.**

Companies and research institutes have been careful to avoid copyright violations and rely on pools of openly licensed works when building AI training datasets. Copyright can be a powerful tool regulating downstream uses of the information commons in AI applications.

At the same time, Adam Harvey showed that since the creation of YFCC100M, face recognition dataset creators and maintainers have misrepresented CC licensing conditions. [Research conducted by Harvey](#) demonstrates different license compliance issues related to AI training datasets. These include such basic errors as the failure to correctly attribute works, but also the lack of compliance with licensing terms, such as limitations on commercial use.

A 2022 dataset licensing study concludes that license compliance has significant challenges. Authors showed that datasets (including [ImageNet](#), [MS COCO](#) or [FFHQ](#)) “might not be suitable to build commercial AI software due to a high risk of potential license violations<sup>7</sup>.” However, according to some experts we have interviewed, license compliance issues are irrelevant in this case, as the uses might fall under fair use rules (in the United States) or some form of a copyright exception for text and data mining (in Europe, for example).

#### *Issues related to text and data mining, and web scraping*

Terms of service could be used to enforce conditions under which photographs can be used. There is room to develop guidelines and best practices in this regard. At the same time, the effectiveness of such measures depends on the shape of copyright exceptions for text and data mining, including web scraping.

There are two basic approaches to creating AI datasets. The first one, which is typical of the case we have been studying, a pool of open works is purposefully chosen to ensure license compliance. The second approach creates the dataset by scraping the “raw internet” and relying on copyright exceptions. [LAION](#), a dataset of 400 million image-text pairs used to build modern text-to-image generation models, has been built in this way.

7 Gopi Krishnan Rajbahadur, Erika Tuck, Li Zi, Dayi Lin, Boyuan Chen, Zhen Ming, Jiang, and Daniel M. German, “Can I Use This Publicly Available Dataset to Build Commercial AI Software? -- A Case Study on Publicly Available Image Datasets,” (arXiv, April 11, 2022), <https://doi.org/10.48550/arXiv.2111.02374>.



The growing relevance of the second approach means that copyright exceptions, especially those focused on text and data mining, will play a key role in structuring the governance frameworks of AI datasets. Based on these exceptions, datasets could be built not just with openly licensed content, but with any publicly available data and content instead.

More research is needed in this context to understand the relationship between research institutions and commercial entities, particularly in light of the exceptions for non-commercial and research uses. For example, Stability.AI, a private company that owns the stability.ai generative AI model, has created this model with the help of a non-profit that manages the LAION dataset, and a research institute that provided the computing power needed to create the model<sup>8</sup>. This can be interpreted as a method of circumventing commercial restrictions on the use of text and data mining exceptions.

Furthermore, platforms can freely share content and data with the service providers they rely on to run their businesses. Users frequently lack clarity on this.

### *Issues related to state regulation*

Laws that determine the legal assessment of the case we have been studying differ significantly across jurisdictions. For example, the text and data mining rules mentioned above are a key type of law that varies by jurisdiction. Open advocacy has traditionally viewed text and data mining laws favorably because they increase access to and use of content (mainly for research purposes). The case of AI training shows that other issues must be considered to comprehend the impact of such practices in the field of AI research and development.

Specific AI system regulations, such as the European AI Act or the AI Bill of Rights in the United States (both of which have yet to be adopted), may include rules that govern the development and management of AI training datasets. Other laws are also applicable. The Digital Services Act, recently adopted in the European Union, requires platforms to identify systemic issues that they are causing. This opens the possibility of promoting audits of algorithmic systems used by major online platforms.

Licensing rules, technical measures, and regulations all interact in ways that together create a governance structure for AI training datasets. For example, statutory limitations on high-risk AI systems seek to impose behavioral constraints similar to those imposed by RAIL-type licenses.

<sup>8</sup> Andy Baio, "AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability (Waxy.Org)," September 30, 2022), <https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/>.

## *Issues related to privacy and personality rights*

Both Creative Commons and key platforms like Flickr or Wikimedia Commons should include more prominent guidelines on privacy-related aspects of sharing content. Stewards of open frameworks should avoid enabling what Salomé Viljoen calls “sludgy consent”: architecture or processes that create the appearance of consent when none has been given<sup>9</sup>.

In the case of photographs of other people, there is an implicit assumption that the photographer obtained permission to publish the photos online. This is frequently legal fiction, but accountability to these other people is difficult. This aspect of the governance vacuum is not addressed by any mechanisms. There are also no frameworks in place to ensure that individual consent obtained by photographers is eligible to third parties, particularly at scale

## *Issues related to research ethics*

The case shows disregard for basic principles of research ethics, including voluntary consent to participate in research and the right to withdraw at any time. When (re)using existing datasets, containing massive amounts of personal data gathered from various sources, obtaining consent for each study may be impossible or impractical. In terms of the right to withdraw from research, this becomes meaningless if a person is unaware that their information is included in the dataset, as demonstrated in the discussed datasets.

Concerns about research ethics fall between the cracks of the existing ethical oversight system. In contrast to the field of biomedical ethics, where there are common and well-established standards for conducting research and treating participants, ethics committees or boards may be ill-equipped to provide adequate guidance in areas such as big data processing and the development of facial recognition technologies. Furthermore, ethics committees tend to focus on the potential harm to individuals involved in research rather than the impact of a project and its potential to harm society.

The application of research findings in facial recognition technologies frequently raises serious ethical and human rights concerns, bringing up the issue of researchers’ accountability for what happens to their work after it leaves the lab. This is linked to the issue of collaborations between universities and technology companies that develop technologies used in mass surveillance. These ties blur the distinction between research and application.

Several initiatives are attempting to create guidelines that would answer ethical questions that people involved in AI research and development are facing. For

<sup>9</sup> Salomé Viljoen, “A Relational Theory of Data Governance” (The Yale Law Journal, Vol. 131, no 2, November 2021), <https://www.yalelawjournal.org/feature/a-relational-theory-of-data-governance>.

example, the Association for the Advancement of Artificial Intelligence (AAAI)<sup>10</sup> and the Association for Computing Machinery (ACM)<sup>11</sup> have both published ethical guidelines for AI development. The European Union’s High Level Expert Group on Artificial Intelligence (AI HLEG)<sup>12</sup> has also released guidelines on the ethical principles that should guide the development of AI. None of them are, however, enforced widely across the entire field of open AI.



10 Association for the Advancement of Artificial Intelligence. “AAAI Code of Professional Ethics and Conduct.” Association for the Advancement of Artificial Intelligence, 2019. <https://www.aaai.org/Conferences/code-of-ethics-and-conduct.php>

11 Association for Computing Machinery. “ACM Code of Ethics and Professional Conduct.” Association for Computing Machinery, 2018. <https://www.acm.org/code-of-ethics>

12 High Level Expert Group on Artificial Intelligence. “Ethics Guidelines for Trustworthy AI.” Text. European Commission, April 2019. [https://openfuture.eu/wp-content/uploads/2023/01/ai\\_hleg\\_ethics\\_guidelines\\_for\\_trustworthy\\_ai-en.pdf](https://openfuture.eu/wp-content/uploads/2023/01/ai_hleg_ethics_guidelines_for_trustworthy_ai-en.pdf)

# 4

## **DESIGNING A COMMONS-BASED GOVERNANCE MODEL FOR AI DATASETS**

# DESIGNING A COMMONS-BASED GOVERNANCE MODEL FOR AI DATASETS

The need for better AI dataset governance is the most important finding of our research. There is a decade-long history of malpractice due to a lack of proper governance models. The responsibility for this falls largely on entities creating, owning, managing, and sharing these datasets – although, as noted above, some of the responsibility also falls on the upstream entities in the life cycle.

Stewards of the information commons – and in particular organizations managing open licensing frameworks and content-sharing platforms – are among those stakeholders who should face the challenge of mitigating risks and harms associated with and caused by the open sharing of content as an information commons.

These stakeholders need to collaborate with another key group of actors who have a stake in the governance of AI datasets: researchers, institutions and companies that create these datasets. Abeba Birhane, Vinay Uday Prabhu and Emmanuel Kahembwe argue that advances in models are rapid and embraced by the research community, while advances in responsible design of datasets are ignored or slow – pointing to the failure to curate key datasets properly<sup>13</sup>.

Trustworthy AI has been a critical narrative shaping the development of these technologies. The case at hand suggests that user trust can be violated already in the early stages of AI development. For this reason, approaches to trustworthy AI should be considered already at the level of data collection and dataset curation.

Dataset governance must address two urgent tasks:

**1. MITIGATING RISKS AND HARMS CAUSED BY THE USE OF EXISTING DATASETS.** This is a decade-long history of datasets that have not been properly governed. There is a need for better standards on the use of both openly licensed content in datasets and for datasets built through web scraping.

**2. DESIGNING GOVERNANCE FRAMEWORKS FOR NEW DATASETS.** The field of AI research can be empowered by the availability of large datasets, made available under easily usable licenses, and with governance models that allow continuous improvement of the datasets & ensure that they are trustworthy (for example, address the issue of bias).

With regard to existing datasets (and models trained on their basis), there is a need to further explore and understand the possibilities of retracting such da-

13 Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, “Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes,” (arXiv, October 5, 2021), <https://doi.org/10.48550/arXiv.2110.01963>.

tasets. DukeMTMC and MegaFace are two prominent datasets from the past that have been taken down – although both cases show challenges with dealing with copies and derivatives of these datasets<sup>14</sup>. A regulated model deletion was proposed as a more robust regulatory measure than financial fines. Finally, self-regulation among AI researchers could lead to community standards prohibiting the use of datasets that have been blacklisted.

Nevertheless, creating new, properly governed datasets (and related governance standards) is more important than policing bad cases from the past. Openly accessible datasets, created in accordance with legal and ethical standards, are critical for democratizing the field and allowing actors other than large corporations to participate in AI research and development. Openness is also crucial for enabling the reproducibility of research.

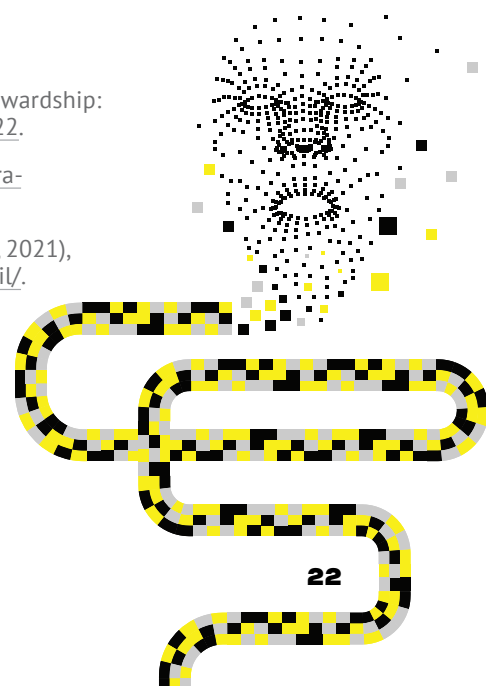
Because of the wide availability of datasets that are shared openly, they can also help to advance other standards of trustworthy AI, such as those related to combating bias in AI technologies. Openly shared datasets have the benefit of being more transparent, allowing for auditing, for example. Finally, commons-based governance approaches can facilitate more participatory dataset design and management modes.

The need for “ethical open datasets” has been recently raised by James Boyle, who argues that they can increase competition, decrease inequality and promote transparency.<sup>15</sup> A similar approach has been proposed by Bhaskar Chakravorty, who argues for the creation of a “creative commons for AI R&D”: a pool of user data created in cooperation by AI research companies and public sector institutions tasked with regulating data and AI<sup>16</sup>.

14 Kenny Peng, Arunesh Mathur, and Arvind Narayanan, “Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers,” (arXiv, November 21, 2021), <https://doi.org/10.48550/arXiv.2108.02922>.

15 James Boyle, “Misunderestimating Openness,” (Open Future, 2022), <https://openfuture.eu/paradox-of-open-responses/misunderestimating-openness>.

16 Bhaskar Chakravorty, “Biden’s ‘Antitrust Revolution’ Overlooks AI – at Americans’ Peril,” (Wired, 2021), <https://www.wired.com/story/opinion-bidens-antitrust-revolution-overlooks-ai-at-americans-peril/>.



# **\_ABOUT**

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.

## *Authors of this report*

Alek Tarkowski is the Strategy Director at Open Future. He holds a Ph.D in sociology from the Polish Academy of Science. He has over 15 years of experience with public interest advocacy, movement building and research into the intersection of society, culture and digital technologies.

Zuzanna Warso is the Director of Research at Open Future. She holds a Ph.D. in International Law from the University of Warsaw. She has over 10 years of experience with human rights research and advocacy. In her work, she has been focusing on the intersection of science, technology, human rights and ethics.




**ALEK TARKOWSKI**

 by: [Giorgos Gripeos](#)



**ZUZANNA WARSO**

 by: [Anna Warso](#)

# KEEP UP TO DATE AND SUBSCRIBE TO OUR NEWSLETTER

**SUBSCRIBE**

**HELLO@OPENFUTURE.EU**

