Peukert, Christian; Abeillon, Florian; Haese, Jérémie; Kaiser, Franziska; Staub, Alexander

Working Paper

Strategic Behavior and AI Training Data

CESifo Working Paper, No. 11099

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Peukert, Christian; Abeillon, Florian; Haese, Jérémie; Kaiser, Franziska; Staub, Alexander (2024) : Strategic Behavior and AI Training Data, CESifo Working Paper, No. 11099, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/300027

CESifo WORKING PAPERS

11099
2024

April 2024

Strategic Behavior and AI Training Data

Christian Peukert, Florian Abeillon, Jérémie Haese, Franziska Kaiser, Alexander Staub

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11099*

Strategic Behavior and AI Training Data

Abstract

Human-created works represent critical data inputs to artificial intelligence (AI). Strategic behaviour can play a major role for AI training datasets, be it in limiting access to existing works or in deciding which types of new works to create or whether to create new works at all. We examine creators' behavioral change when their works become training data for AI. Specifically, we focus on contributors on Unsplash, a popular stock image platform with about 6 million high-quality photos and illustrations. In the summer of 2020, Unsplash launched an AI research program by releasing a dataset of 25,000 images for commercial use. We study contributors' reactions, comparing contributors whose works were included in this dataset to contributors whose works were not included. Our results suggest that treated contributors left the platform at a higher-than-usual rate and substantially slowed down the rate of new uploads. Professional and more successful photographers react stronger than amateurs and less successful photographers. We also show that affected users changed the variety and novelty of contributions to the platform, with long-run implications for the stock of works potentially available for AI training. Taken together, our findings highlight the trade-off between interests of rightsholders and promoting innovation at the technological frontier. We discuss implications for copyright and AI policy.

JEL-Codes: K110, L820, L860.

Keywords: generative artificial intelligence, training data, licensing, copyright, natural experiment.

*Christian Peukert**
Faculty of Business and Economics (HEC)
University of Lausanne / Switzerland
christian.peukert@unil.ch

Florian Abeillon
Faculty of Business and Economics (HEC)
University of Lausanne / Switzerland
florian.abeillon@proton.me

Jérémie Haese
Faculty of Business and Economics (HEC)
University of Lausanne / Switzerland
jeremie.haese@unil.ch

Franziska Kaiser
Faculty of Business and Economics (HEC)
University of Lausanne / Switzerland
franziska.kaiser@unil.ch

Alexander Staub
Faculty of Business and Economics (HEC)
University of Lausanne / Switzerland
alexander.staub@unil.ch

*corresponding author

April 2024

The first author led the project and contributed the most, while other authors who made significant contributions are listed alphabetically. We thank Anthi Kiouka for valuable input and discussions. We also thank Timothy Carbone at Unsplash for generously answering our questions, giving helpful insights and providing access to data. We are grateful to Heski Bar-Isaac, Stefan Bechtold, Chiara Farronato, Christophe Gösken, Brad Greenwood, Ginger Zhe Jin, Reinhold Kesler, Sijie Lin, Ananya Sen, Holger Spamann, Catherine Tucker and Joel Waldfogel for helpful comments. The paper has benefited from the feedback of participants at seminars and conferences at Technis Webinar, ETH Zurich (Center for Law and Economics), Harvard Business School (Digital Competition and Tech Regulation Conference), and NYU Stern School of Business.

1 Introduction

Data is a key input to artificial intelligence (AI) that enables societal and economic progress through the discovery of knowledge, enhancement of productivity, and expansion into new or existing markets (Wu et al., 2020; Brynjolfsson et al., 2023; Beraja et al., 2023; McElheran et al., 2024). The rapid proliferation of generative AI (genAI) technologies has enabled hundreds of millions of users to generate high-quality text (including software code), images, audio, and video at a negligible cost. All this is made possible by the vast *stock of data* available on the public internet, which together with computing power, is responsible for the tremendous performance improvements in recent large language models (Ho et al., 2024; Metz et al., 2024). However, access to a continuous *flow of data* remains crucial for the performance of a wide range of AI applications in dynamic social environments (Hadsell et al., 2020; He et al., 2011; Peukert et al., 2023; Valavi et al., 2022; Whang et al., 2023). Input data for genAI models, in particular, comes mostly from online content in the form of text, images, sound, and video produced by humans. Hence, human behavior and incentives for individuals and organizations to contribute to training datasets play a major role for the quality of genAI (Bauer et al., 2016; Sokol and Van Alstyne, 2021; Burtch et al., 2024; del Rio-Chanona et al., 2023; Quinn and Gutt, 2023).

Questions of whether, how, and which data can or should be utilized to train AI have moved to the center of the policy debate. Regulatory frameworks – or the lack thereof – can significantly impact both the *demand* for and *supply* of data. With respect to input data, AI policy intersects with privacy law and competition policy (Jones and Tonetti, 2020; Farboodi et al., 2019). More recently, however, intellectual property law has emerged as a critical area of the policy discourse because input data may be subject to copyright protection (Eshraghian, 2020; Samuelson, 2023). While some jurisdictions allow exemptions for research and development in copyright law, there is general legal uncertainty (Fiil-Flynn et al., 2022; Henderson et al., 2023). Several high-profile lawsuits against AI developers allege, among other claims, direct copyright infringement by creating unauthorized copies of their works and using such copies as training data.[1] As a result, policymakers around the world are tasked with balancing innovation in AI and the interests of rightsholders. So far, however, there is not sufficient empirical evidence to guide policy (Peukert and Windisch, 2024).

In this paper, we shine a light on how strategic behavior shapes the flow of data, studying the response of individual creators to their works being made available for the training of commercial AI.

[1] See articles in the New York Times (`http://tiny.cc/pxljxz`) and Reuters (`http://tiny.cc/txljxz`).

Our empirical setting is one of the largest stock photography websites in the world, Unsplash. In the summer of 2020, Unsplash released metadata of a subset of 25,000 photos explicitly for the training of commercial AI applications (hereafter the "LITE" dataset). This natural experiment allows us to identify the causal effects of being included in AI training data and characterize strategic responses of those affected. We compare the upload behavior of contributors whose works were included in this dataset to contributors whose works were not included. In addition, we investigate changes in the variety and novelty of user contributions due to the release of the LITE dataset. We measure variety and novelty by calculating the similarity of each new upload to all existing images one year prior to the treatment based on their associated keywords using natural language processing methods. This allows us to track how variety and novelty change over time for both treated and control users, before and after the release of the LITE dataset.

We find that treated users left the platform at a higher-than-usual rate. Conditional on remaining active, treated users substantially slowed down the rate of new uploads by about 40% per month.

Our content analysis shows that within-users, uploads decrease in variety but not in novelty compared to the existing stock of images. Across users, however, the variety of uploaded images decreased by about 5% compared to the existing stock and uploaded images were about 30% less novel. This implies changes to the aggregated training dataset stemming mostly from changes in user composition rather than from a shift in individual behavior. We show that users with professional gear reduce their contributions significantly more, as do more successful users. However, our results do not imply that photographers leave the market entirely. We show that users' behavioral changes only occur on Unsplash and are not mirrored in their activity on Instagram, the most popular image-sharing platform. Finally, we provide some evidence that the behavioral response is stronger, the more AI may be perceived as an economic threat. First, we show that the effect intensifies later when more capable genAI models become widely known to the public. Second, users who have multiple images in the training dataset reduce their activity almost twice as much compared to users who had only one image included.

Overall, we document that strategic behavior affects the size as well as aspects of the quality of the flow of data. A back-of-the-envelope calculation suggests that making the entire catalog available for commercial AI research (which is similar to a policy that would allow fair use of any copyrighted

material) would have reduced the flow of data by half. At the same time, the flow would become increasingly similar to the stock, with the number of very similar images tripling within a year.

Our contribution is twofold. First, we provide the first large-scale empirical evidence on strategic behavior in the supply of data. This adds to a growing literature that discusses the effects of AI on workers and platform contributors (Acemoglu et al., 2022; Doshi and Hauser, 2023; Eloundou et al., 2023; Felten et al., 2023; Huang et al., 2023; Quinn and Gutt, 2023). In particular, we highlight upstream consequences of the proliferation of genAI models, which is an under-explored topic (Lin, 2024). Second, our setting allows us to estimate the causal effects of adding human-made works to training datasets that can be used for the development of commercial AI applications. Thereby we are able to provide important empirical evidence to a primarily theoretical debate on policy-making in the age of AI (Fiil-Flynn et al., 2022; Samuelson, 2023; Henderson et al., 2023; Gans, 2024; Yang and Zhang, 2024; Wang et al., 2024).

2 Background and related literature

Regulation can affect the demand for data by defining rules about which data can or should be used for AI applications. For example, privacy law and competition policy can reduce firms' abilities to track consumer behavior (Johnson, 2022), require firms to elicit consent from consumers (Godinho de Matos and Adjerid, 2022) or share data with competitors (Prüfer and Schottmüller, 2021; Martens et al., 2024). Access to data has direct consequences for the output of AI applications. The value of data, for example for prediction exercises, varies with its quality and quantity (Lei et al., 2023; Neumann et al., 2019; Peukert et al., 2023; Sun et al., 2023).

Concretely, the new AI Act of the European Union will demand that "Training, validation and testing datasets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used".[2] However, such provisions also have intersections and perhaps even contradictions with intellectual property law. In the EU, for example, although copyright law was only recently modernized with the Directive on Copyright in the Digital Single Market in 2019, discussions about further reform are in full swing as part of legislation on the regulation of AI. When

[2]See Article 10 of the Proposal of the European Parliament and of the Council on the Artificial Intelligence Act from January 2024, https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

rightsholders can exercise exclusive rights and restrict access to certain source materials, copyright may, in turn, create or promote biased AI systems (Levendowski, 2018).

On the other hand, the supply of data is discussed less. With the abundance of data online, researchers and AI developers can potentially access a large *stock of data*, most of which was created before the dawn of AI. In general, surprisingly little is known about input datasets for AI applications. Interesting empirical work comes from the computer science community. Birhane and Prabhu (2021) study a widely-used training dataset of images concerning privacy issues (the presence of human faces), sexually explicit content, and measures of the age and gender distribution of depicted humans. In related recent work, Guilbeault et al. (2024) study gender bias in about one million images downloaded from the web. In many applications, access to a large stock of historical data is sufficient. Yet, in dynamic settings, where the stock of data can become outdated, a consistent *flow of data* is key for algorithmic performance (Hadsell et al., 2020; He et al., 2011; Valavi et al., 2022; Peukert et al., 2023; Whang et al., 2023). However, the flow of data is endogenously determined by strategic behavior, which is, in turn, influenced by institutional settings such as privacy law, competition policy, and intellectual property law.

Key criteria that define the quality of training datasets concern the *variety* and *novelty* of information in the dataset. To illustrate this, consider two examples. First, imagine a panel dataset that an econometrician would analyze. We can increase variety by adding more cross-sections, e.g. more firms, and we can increase novelty by expanding the time series, i.e. adding more years per firm. As a second example, consider a dataset with photos of New York City. We can add to the variety dimension by adding images of buildings from every neighborhood, from different angles, at different times of the day, etc. However, as some buildings make way for new developments, we need to add novel images of the same concept, e.g. the skyline, to ensure a continued high-quality dataset. Variety and novelty can be measured by changes coming in the flow of incoming data points and their relation to the existing stock of data. As a result, strategic behavior that alters the flow of data can also alter key dimensions of dataset quality and, therefore, the output quality of AI applications.

Recent theoretical work is useful as a starting point to think about policy options to solve the trade-off between copyright protection and AI innovation (Gans, 2024; Yang and Zhang, 2024). Digital goods, such as user-generated content or creative works, tend to have high fixed costs and near-zero costs of copying. Intellectual property rights or alternative business models create rents that

allow creators to recoup their fixed-cost investments. The proliferation of genAI systems suggests two potential outcomes: AI might reduce these monopoly rents, leading humans to produce only works with lower fixed costs, or AI might increase human productivity, effectively reducing fixed costs and requiring lower monopoly rents to incentivize creation. These effects can depend on factors such as the bargaining power of rights holders (Gans, 2024) and the existing body of work in a particular domain (Yang and Zhang, 2024). A potential solution is a licensing mechanism that compensates rightsholders with a royalty scheme that is determined by Shapley values of pieces of training data that end up in the final outcome (Wang et al., 2024).

Strategic behavior of contributors to training datasets, especially when large parts of the public internet can be considered as a huge training dataset, can have intricate long-run effects on online communities. Evidence from platforms like Stack Overflow suggests that the introduction of genAI systems like ChatGPT can lead to reduced community participation, a decrease in the number of questions and users as well as a decline in the quantity and quality of answers (Burtch et al., 2024; Quinn and Gutt, 2023; Gallea, 2023), and an overall decrease in activity (del Rio-Chanona et al., 2023). Three recent working papers are most directly related to our paper. Quinn and Gutt (2023) discuss how the quality of training data can change because of changes in user behavior. In the context of Stack Overflow, they show that changed user behavior affects the underlying quality of the training data for the next iterations of AI models. Huang et al. (2023) investigate the impact of genAI on content creators across two major Asian platforms, Lofter and Graffiti Kingdom, through quasi-experiments. The study reveals that introducing genAI tools on Lofter decreased creator activity, whereas banning these tools on Graffiti Kingdom had a mixed impact: Non-churning creators increased their activity, while that of individuals with copyright concerns decreased. Similarly, Lin (2024) studies the online art platform DeviantArt and shows that contributors decreased their publication volume by 22% after the platform introduced a genAI feature that let users generate art with a click of a button. In summary, while evidence on the impact of genAI trained on vast and intransparent datasets is growing, the lack of research on contributors' responses to the explicit use of their work for the development of commercial AI models forms the backdrop of our study.

3 Empirical setting

3.1 Unsplash

Unsplash is a stock photo platform hosting approximately 6 million photos and illustrations, submitted by roughly 360,000 contributors. According to Similarweb, Unsplash is among the top 5 most visited websites in the "photography" category globally, receiving about 30 million visits each month. Over the past decade, Unsplash has recorded a cumulative total of 1 trillion image views, coming from tens of millions of individual users and thousands of API partners.

Unlike other stock photography websites, the vast majority of images on Unsplash are made available under a permissive license, which allows all images on the platform to be used freely for both non-commercial and commercial purposes. Users can, of course, revoke this license by deleting individual images or their entire accounts at any time. Unlike other websites that make images available under similar licenses (e.g. Wikimedia), contributions on Unsplash are of a high quality. Partially, this is enabled by rigorous moderation efforts of an editorial team that hand-selects images that are prominently shown on the website. As of May 2023, the curation team has selected about 300,000 images for curation, which represents about 6% of all images available on the platform.

In October 2022, Unsplash launched a paid subscription service called Unsplash+. Among other things, after being accepted into the program as a contributor, users can receive monetary compensation for uses of the works.[3] Further, the Unplash+ license explicitly disallows the use of images for machine learning, AI, and biometric tracking technology.[4]

3.2 The Unsplash research program

3.2.1 Two-tiered access to training data for AI

In August 2020, Unsplash released a dataset containing 25,000 "nature-themed" and "featured" photos available for commercial and non-commercial use, including for the training of AI. This dataset is called the LITE dataset. In addition, the platform released a dataset covering the whole collection of images hosted on the platform (hereafter the FULL dataset) made available exclusively for non-commercial research purposes. Access to the LITE dataset happens with the click of a button, access to the FULL dataset is subject to an application process with a thorough manual review process. The datasets do not contain the images per se (they can still be downloaded quite effortlessly through an API),

[3] See https://unsplash.com/blog/contribute-to-unsplash/.
[4] See https://unsplash.com/plus/license.

but costly-to-compute metadata such as keyword tags, depicted landmarks, color distributions, etc. By releasing the two flavors of the dataset, Unsplash essentially treated their entire community with their image (metadata) being available for AI research. However, the images that were selected for the LITE dataset receive a much stronger treatment because access is quick, free, and subject to less restrictive usage permissions.

3.2.2 The selection of images in LITE

The LITE dataset was created on June 25, 2020, and, therefore, is a subset of all images available on Unsplash up until that date. The images to be included in the dataset were selected using the following database query:[5]

```
SELECT DISTINCT photos.id FROM photos
JOIN tags ON tags.photo_id = photos.id
WHERE tags.keyword = 'nature' -- The photo is tagged with "nature"
      AND tags.confidence > 0.9 -- We are more than 90% confident that "nature" is present in that photo
      AND photos.curated =  TRUE -- photo has been curated by the editorial team
      AND photos.status = 1 -- photo is still available on the platform
LIMIT 25000
```

Unsplash uses third-party computer vision algorithms to automatically add tags to images, most prominently Amazon's Rekognition service.[6] These auto-generated keywords come with a confidence score which ranges from 0 to 1.

Curated images are photos that were chosen by a human curator at Unsplash to be "special", i.e., unique perspective, exploration of light and materials, creative use of color, inventive framing, evoking a mood, usefulness for reuse. These photos were either featured on the front page of the website or featured for ten days under a specific topic (e.g.: "nature").[7]

[5] See https://github.com/unsplash/datasets/issues/55 for details on the sampling procedure as expressed by Timothy Carbone, Head of Data at Unsplash, who runs the research dataset program. In an email conversation with us, he shared the database query he used to select the images in LITE and clarified further: "The Lite dataset's only objective is to give a preview of the content that's in the Full dataset. Here are the motivations behind its content: (1) Focusing on a single keyword helps the end-user accessing more depth of content, rather than breadth. This is allegedly more helpful for things like training AI and will hopefully provide a better 'trial' experience for the dataset. (2) Making sure we only choose curated content helps with lowering the risk of the content being removed from the platform and makes the dataset easier on the eye."

[6] See https://unsplash.com/blog/the-data-stack-at-unsplash/ and https://docs.aws.amazon.com/rekognition/latest/APIReference/API_DetectLabels.html for details.

[7] See https://unsplash.com/blog/how-we-choose-what-photos-to-feature-on-the-unsplash-homepage/.

The fact that there was no specific ordering in the SQL query, i.e. no ORDER BY clause, leads to an unpredictable row order of the results. The row order depends on the database's internal storage and retrieval mechanisms, which can be influenced by factors like the structure of the database's indexes and how the query optimizer chooses to execute the query.[8]

Table 1: Sort order in LITE dataset

	(1)	(2)
AutoKeywordScore	-3.3952***	-3.4890***
	(1.1170)	(1.1132)
ImagePopularity	-0.0000	-0.0000
	(0.0000)	(0.0000)
UserKeyword	136.1898	142.9112
	(119.9305)	(120.5190)
ImageAge	-25.7948	13.3226
	(37.7172)	(48.3746)
UserPopularity		0.0545
		(0.0508)
AccountAge		-53.3435
		(43.0708)
adj. R2	0.0003	0.0003
Observations	25,000	24,964

Note: White-robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table 1 shows that the order of images in the LITE dataset cannot be explained by image characteristics such as image age (i.e. time since upload), image popularity (i.e. views and downloads) or whether the user has chosen the keyword "nature" to describe the image. Also, user characteristics such as the total sum of views and downloads of the users' images (*UserPopularity*) and the time since the first upload of the user (*AccountAge*), do not explain the order of images in the LITE dataset.[9] Only the confidence score associated with the automatically assigned keyword "nature" is correlated

[8]The documentation of PostgreSQL, which is the database system Unsplash uses, states: "When using LIMIT, it is a good idea to use an ORDER BY clause that constrains the result rows into a unique order. Otherwise you will get an unpredictable subset of the query's rows [...] You don't know what ordering unless you specify ORDER BY. [...] If ORDER BY is not given, the rows are returned in whatever order the system finds fastest to produce.", see https://www.postgresql.org/docs/current/sql-select.html and "If sorting is not chosen, the rows will be returned in an unspecified order. The actual order in that case will depend on the scan and join plan types and the order on disk, but it must not be relied on. A particular output ordering can only be guaranteed if the sort step is explicitly chosen.", see https://www.postgresql.org/docs/current/queries-order.html.

[9]Note that we do not have user-level information about 14 users, because their accounts were deleted before we can observe them in FULL 1.0.1, the earliest FULL dataset we have access to. These users have 63 images in the LITE dataset. Hence, the number of observations in column 2 is smaller than in column 1.

with the sort order. The higher the confidence score of an image, the higher up it appears on the list. This is consistent with PostgreSQL using a descending order to generate a list of all keywords that meet the *confidence* condition. For example, conditional on meeting the threshold, keywords with a confidence score of 0.99 would appear before keywords with a confidence score of 0.90. However, it is worth stressing that the regression can only explain little variation in the rank order, as shown by the low value of the adjusted R^2.

With this information, we can construct a control group of images, for example, consisting of those images that would have shown up below row 25,000 had the query not used a LIMIT condition. Based on the results in Table 1, images would be listed further up, the higher the confidence value of the computer vision algorithm for the keyword "nature". For technical reasons, however, this does not necessarily mean that those images monotonically convey more of the concept "nature". They can also convey the same amount of "nature" but less of other concepts. In multi-label classification algorithms using Deep Learning and Convolutional Neural Networks, as deployed in the DetectLabels service of Amazon Rekognition, the calculation of confidence scores is a nuanced process.[10] The last layer typically applies the softmax function, transforming logits – the network's final layer output, representing unnormalized log probabilities – into a probability distribution across the predicted labels. Softmax is non-linear implying that changes in logits can be disproportionally reflected in confidence scores. Moreover, the confidence score for a given label is not only a function of its own logit but also depends on the logits of all other labels. Therefore, an increase in the confidence score of "nature", disproportionally so at the high end of the range, may result from a larger logit of the concept itself or from smaller logits of other concepts.

3.2.3 Announcement and stakeholders' awareness of the program

The release of the research datasets was announced publicly on the Unsplash blog on August 6, 2020.[11] In this announcement, Unsplash mentioned the target users for the dataset: "We're releasing the most complete high-quality open image dataset ever, free for anyone to use to further research in machine learning, image quality, search engines, and more." They also describe the differences between the two versions: "We're releasing the data in two versions: a LITE dataset available for commercial and noncommercial usage, and the FULL dataset available for noncommercial usage." Importantly, they

[10] The exact technical underpinnings of Rekognition are not publicly disclosed, but a slide deck from a System Dev Engineer at Amazon Webservices gives some indication about the service's architecture, see https://iptc.org/download/events/phmdc2017/IPTC-PhMdC2017-AHornsby-AmazonRekognition.pdf.

[11] See https://unsplash.com/blog/the-unsplash-dataset/.

also acknowledge that the dataset is not a one-shot project: "As the Unsplash library continues to double in size every year, we'll continue updating the dataset with new fields and new images." Every user who subscribes to the Unsplash newsletter, which is the default option when signing up for an account, receives periodic emails summarizing announcements on the Unsplash blog. Therefore, we have strong reasons to believe that the average user was aware of the research dataset program. Users were not able to opt out, but were, of course, free to delete affected images or their user accounts. Users also did not receive monetary or other compensation for having images included in the LITE nor the FULL dataset. The release of the Unsplash dataset was well received by the AI community. The GitHub repository, which provides documentation and sample code, was starred 2,300 times, and there is an active discussion with the community via issues and pull requests.[12] In general, according to PapersWithCode, data from Unsplash has been used in at least 24 computer science papers for problems such as Image Augmentation, Image Generation, Image Inpainting and Outpainting, and Text-to-Image Generation.[13] Altogether, these are strong reasons to believe that the release of the LITE dataset was sufficiently salient to Unsplash users to allow us to detect treatment effects.

4 Data and methods

4.1 Available information

We have access to rich metadata on more than 4.9 million images, which represents all images uploaded to Unsplash since May 2013 and still available in June 2020 and/or in May 2023 (FULL dataset). In particular, we combine information from datasets released in August 2020 (but created on June 25, 2020) and in May 2023.[14] We observe the user account that has uploaded the image, the upload date, whether an image was curated, a list of keywords associated with each image, and metadata on camera gear for photos. We also observe whether a user has chosen to display a badge on their Unsplash page that says "Available for hire" and for a subset of users (those who chose to reveal this information in their Unsplash profile), we also know their Instagram account name and have collected information

[12] See https://github.com/unsplash/datasets.

[13] See https://paperswithcode.com/datasets?q=unsplash&v=lst&o=match.

[14] To be precise, we have access to datasets in version 1.0.0 (LITE), and versions 1.1.0 and 1.2.1 (FULL). See https://github.com/unsplash/datasets/blob/master/CHANGELOG.md.

on their upload history on Instagram.[15] Further, we know whether a user has joined the Unsplash+ program as a contributor.[16]

Using information about the camera gear that the photographer used, we flag users as professionals if at least one of the following conditions is fulfilled for any photo they uploaded before the release of the LITE dataset. First, the exposure time is longer than 5 seconds, which cannot be executed by low-end cameras – including most smartphones – and without a tripod. Second, a photo was shot with a focal length of more than 250, which typically requires expensive professional lenses. Third, a photo is shot on a camera from a mid-to-high-end brand, i.e., not on a smartphone.[17]

4.2 Treatment and control groups

We exploit the fact that Unsplash released a subset of their entire platform contents as a freely available LITE dataset. Because we want to study the behavioral responses of users when their work is made available as a training dataset for AI research, our treatment group is comprised of all 8,298 user accounts that have at least one photo in the LITE dataset.[18] There are several ways to construct a control group. First, we could simply take all users who had uploaded at least one image before the LITE dataset was created but did not have any images included in the LITE dataset. Second, we could take the set of users whose images fulfilled the opposite of the selection criteria of the original database query discussed in section 3.2.2. That is, we take users who had not uploaded nature-themed and curated photos at the time when the LITE dataset was created. Third, we could take all users whose images fulfilled the same criteria as in the original database query but were not included in the LITE dataset. This approach comes closest to resembling the perfect counterfactual for users in the LITE dataset, making it our preferred specification. This specification of the control group includes

[15] This information is not available in the LITE or FULL datasets. We, therefore, do not have time-varying information, but we were able to get a snapshot on February 5, 2024, from which we observe the for-hire status and Instagram information as of that day of about 11,352 out of 12,052 users from our preferred sample of treated and control users. We observe Instagram activity for a subset of 1,084 users. Not all Unsplash users reveal their Instagram accounts on their Unsplash profiles. Further, technical restrictions require us to limit ourselves to the 20 most recent posts of a user. We exclude users for whom we only observe posts in the after period as we do not know whether this is because the account didn't exist in the before period, whether there were no posts in the before period, or whether the user is so active that we can only see posts from the after period in our snapshot of the 20 most recent posts. As a result, our sample is likely biased towards less active users with relatively old Instagram accounts.

[16] We do not know when a particular user joined the Unsplash+ program. However, we have this information as of April 23, 2024, approximately 2.5 years after the program's launch.

[17] We do this based on the following list of brands: Hasselblad, Leica, Rollei, RED, Blackmagic, Voigtlander, Canon, Nikon, Fuji, Sigma, Olympus, Ricoh, Minolta, Konica, Yashica, Kodak, Agfa, Noritsu, Epson, DJI.

[18] Users can change account names on Unsplash. Using the fact that the IDs of photos remain unique and cannot be changed, we can track the same user even if they changed account names.

3,754 users. In the robustness section, we also report results from the two former approaches, showing that the results are consistent, but our preferred approach yields the most conservative estimates.

4.3 Panel construction

To analyze dynamics over time, we need to create a panel structure at the user level. We construct two separate panels. First, while we have information on the upload date of every image, we do not have high-frequency information on whether an image is still available on the platform. We have two snapshots from which we can construct a dataset that tells us whether a user account we observe carrying active images in June 2020 still has any available images in May 2023.

Second, to study contribution dynamics in detail, we aggregate the data at the monthly level and count the number of uploads of every user within that time period. To create a balanced panel, we impute zeroes in months where we do not observe any activity.

4.4 Analysis of variety and novelty

To investigate whether contributions to Unsplash change after the release of the LITE dataset, we compare the flow of new uploads to the stock of images one year before the launch of the research program. In particular, to capture *variety* as a quality measure of a training dataset, we measure how similar each new upload is relative to all existing images. To capture *novelty*, we count the number of images in the stock of data that are very similar to a focal new upload.

We then compare how variety and novelty evolve over time for treated users and control users, before and after the release of the LITE dataset. We do so based on the keywords that users assign to their uploaded images. The average image has 10.28 keywords. We use natural language processing techniques to assess the similarity between images based on their associated keywords in 4 steps.

1. Aggregate keywords for each image and train a Word2Vec model on these aggregated keywords to generate a vector representation for each unique keyword. Herein, the Word2Vec model is configured to create 100-dimensional vectors for each keyword, with a context window of 5 words and a minimum word frequency threshold of 1, ensuring that even rare keywords are included in the model.

2. Apply the Word2Vec model to each image, converting the keywords into vectors, and then calculate the average vector to represent the image's semantic content.

3. Calculate cosine similarity scores between the vector representation of a particular image and the vector representations of all existing images.

4. Compute average similarity across all images and count the number of images that exceed a specified similarity threshold, respectively.

We perform steps 1 and 2 for the stock of images uploaded by users that eventually will be treated, as well as for the stock of images uploaded by users in the control group. We define the stock as all images uploaded before June 25, 2019. We perform steps 1-4 for all images uploaded by treated and control users after June 25, 2019. This allows us to compare new uploads starting one year before the start of Unsplash's research program to new uploads after the release of the first LITE dataset, across users in the treatment and in the control group. We report 0.8 and 0.9 as the threshold level for very similar images in step 4, where cosine similarity scores are defined on the interval $[-1, 1]$, with higher values corresponding to higher similarity between images.

4.5 Identification strategy and econometric model

Our causal identification strategy is based on the notion that the images in the LITE dataset are randomly drawn from a superset of all images on the platform that meet the eligibility criteria. This implies quasi-randomization at the user level. For our econometric analysis, we rely on a difference-in-differences approach. We perform a set of OLS regressions saturated with fixed effects to compare users with images in the LITE dataset to users who had at least one image deemed eligible to be included in the LITE dataset but were not included in the LITE dataset. Our baseline specification is

$$Y_{it} = \delta(Post_t \times Treated_i) + \eta_t + \mu_i + \varepsilon_{it}, \tag{1}$$

where Y_{it} is the outcome of interest, e.g. the number of images uploaded by user i in month t. $Treated_i$ indicates whether user i was included in the LITE dataset. $Post_t$ is a binary variable that takes the value 1 if the current month is after the release of the LITE dataset, and 0 otherwise. We include month fixed effects η_t in all specifications, and in some specifications, we additionally include user fixed effects μ_i. Because inclusion in the LITE dataset is the result of a natural experiment with randomization, the parameter δ estimates the causal effect of being included in a dataset made available for AI training on user behavior.

Table 2: Results: Likelihood of remaining on the platform

	Images			Users	
	(1)	(2)	(3)	(4)	(5)
Treated	0.0220**	0.0219**	0.0109***	-0.0075*	-0.0084**
	(0.00875)	(0.00879)	(0.00155)	(0.00395)	(0.00406)
Mean Control	0.9075	0.9075	0.9075	0.9624	0.9625
Image-Age FE	Yes	Yes	Yes	Yes	Yes
User FE	No	No	Yes	No	No
Observations	705,268	704,848	704,848	12,033	11,607

Note: We estimate a linear probability model. The dependent variable is an indicator of whether an image or a user account is still available on Unsplash in May 2023. *Treated* indicates whether a user was part of the LITE dataset. We compare the survival of (images by) users who were part of the LITE dataset to the survival of (images by) users who were not part of LITE dataset over a period of 2.5 years (from Aug 20 to May 23). Fixed effects for the image's age (in months). Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

5 Results

We begin by reporting baseline results separately for users' deletion rates as well as their uploading behavior. Then, we discuss effect heterogeneity by different types of users and different types of images.

5.1 Baseline results

5.1.1 Changes in deletion rates

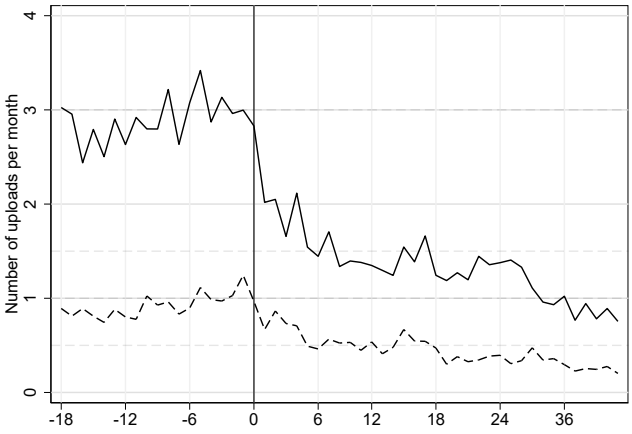
First, we study whether being included in the LITE dataset affects the likelihood of an image or a user account being deleted from the platform. Hence, this set of results differs from all other analyses else below, which focus on users' uploading behavior.

It is important to note that while we know the exact upload date of each image, we do not know the exact deletion dates of either images or user accounts. However, we can use the fact that Unsplash periodically releases updates of the FULL dataset, which encompasses the entire platform catalog at a particular point in time. We can trace images (and corresponding users) that were part of the LITE dataset (created in June 2020) and observe whether they were still available on the platform in May 2023.

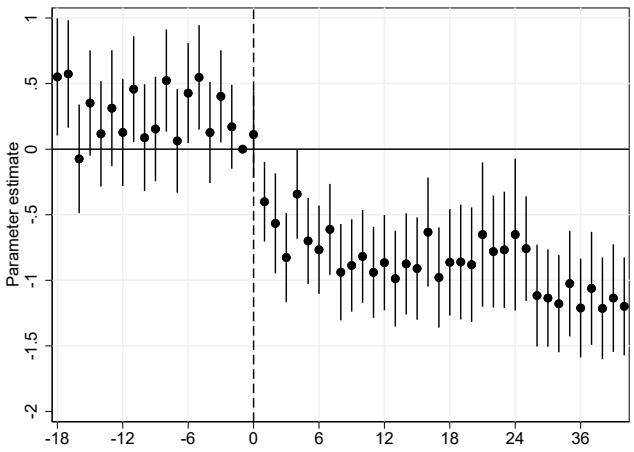
In column (1) of Table 2, we see that images that were included in the LITE dataset (i.e. that were treated) have a 2 percentage points higher chance of surviving on the platform than images that were not included in the LITE dataset. In column (2), we limit the sample to images of photographers that have at least two uploads and find the same result. We do so to compare against the specification

Figure 1: Number of uploads, treatment versus control group

Panel A: Levels



Panel B: Differences



Note: In Panel A, we depict the average number of uploaded images for users in the treatment group (solid) and users in the control group (dashed). In Panel B, we plot OLS estimates of the δ_τ coefficients are obtained from $Y_{it} = \sum_{\tau \in T} \nu_\tau + \sum_{\tau \in T} \delta_\tau \left(\gamma_\tau \times Treated_i\right) + \mu_i + \varepsilon_{it}$, with the number of uploads of user i per month t as the dependent variable. The dots reflect month-specific point estimates comparing treated users to the control group. Standard errors are clustered at the user level, and bars indicate 90% confidence bands. In both panels, the vertical line indicates the creation of LITE 1.0.0.

in column (3), which has user-fixed effects. The coefficient there implies a 1 percentage point higher likelihood of survival. This suggests substantial differences in the variation between users and within users. Overall the results suggest that treated images are deleted at a rate of 7-8% ($\approx$ 3% per year), whereas control images are deleted at a rate of 9% ($\approx$ 4% per year). In columns (4) and (5), we turn to user accounts. We see that treated user accounts are less likely to survive than accounts of control users. The estimate is about 1 percentage point in both samples of all users and those users with at least two observations (i.e., the users in the sample in columns 2 and 3). This implies that accounts of treated users are deleted at a rate of 3% ($\approx$ 1% per year), whereas control users delete their accounts at a rate of 4% ($\approx$ 1.5% per year).

5.1.2 Changes in users' uploading behavior

In the following specifications, we focus on the number of new uploads as our dependent variable. We compare users that had at least one image in the LITE dataset to our control group. For this analysis, we can draw upon high-frequency information, which allows us to construct a panel dataset at the user-month level. In Panel A of Figure 1, we plot the average number of uploads per month in the treatment group and the control group. In the observed 18 months before the creation of the LITE dataset (indicated by a 0 on the horizontal axis), upload rates do not fluctuate much for both.

Table 3: Results: Changes in uploads

	(1) Uploads	(2) I(Uploads>0)	(3) Log(1+Uploads)
Post × Treated	-1.1099***	-0.0691***	-0.1630***
	(0.09693)	(0.00352)	(0.00728)
Mean Treated Before	2.9038	0.2635	0.4892
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Observations	612,662	612,662	612,662

Note: The dependent variable in column (1) is the total number of uploads per month, in column (2) it is an indicator of whether a user account has uploaded at least one image in a given month, and it is the log number of uploads in column (3). *Treated* indicates whether a user was part of the LITE dataset. Fixed effects for upload month of the image and user account. Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Treated users upload on average around 3 images per month and control users upload around 1 image per month. Looking at differences between treated and control in Panel B of Figure 1 confirms that there are no statistically significant differences in the before period. However, we see a substantial and statistically significant decrease in the after period. The effect is immediate and remains relatively stable over time. After 24 months, there seems to be a further decrease, which we investigate further in section 5.4.

We can quantify the average treatment effect for the entire observed post-period in an OLS model specified as described in equation (1). The results are reported in Table 3. In column (1), we look at the number of uploaded photos per month and quantify the average treatment effect as 1 less upload per month, which is a reduction of about 38%. We get similar results when we apply a log transformation in column (3). Further, in column (2), we document a decrease in the extensive margin of about 30%, i.e., when we use a dependent variable that indicates whether a user has uploaded at least one image in a given month.

5.2 Changes by different types of users

5.2.1 Professional versus amateur photographers

Beyond the average treatment effect on the treated, investigating heterogeneity in user characteristics can provide further nuance. Hence, we consider whether behavior differs between professional and amateur users. Thereafter, we turn to differences between more or less prolific users.

The results in column (1) of Table 4 show that treated users who upload photographs created with professional gear react even more negatively than others. Similarly, we find in column (2) that treated

Table 4: Results: Changes in uploads, professionals versus amateurs

	(1)	(2)
Post × Treated	-0.6740***	-0.7206***
	(0.13433)	(0.09401)
Post × Pro-Gear	-0.0686	
	(0.10904)	
Post × Treated × Pro-Gear	-0.5038***	
	(0.17407)	
Post × ForHire		-0.2362
		(0.16669)
Post × Treated × ForHire		-0.8917***
		(0.25200)
Mean DV	1.4595	1.4474
Month FE	Yes	Yes
User FE	Yes	Yes
Observations	611,891	577,739

Note: The dependent variable is the number of uploads in a given month. *Treated* indicates whether a user was part of the LITE dataset. *Pro-Gear* indicates whether the user has uploaded at least one photo taken with professional gear (see section 4 for a detailed definition). *ForHire* indicates whether a user has chosen to show "Available for hire" on their Unsplash profile page. The number of observations differs from those in Table 3 because we do not observe *Pro-Gear* for all images and *ForHire* for all users. Fixed effects for upload month of the image and user account. Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$ * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

users with a badge on their profile page that says "Available for hire" reduced their uploads disproportionally compared to users without a badge. It is possible that users with professional gear and those who signal that they are available for hire make a living as photographers and are, therefore, perhaps more directly affected by the potential displacement effects of AI. Hence, monetary motives may be one of the drivers for users to contribute less to the platform after their works have been made available for AI research.

5.2.2 Heterogeneity in contribution intensity

The distinction between professional and amateur photographers is helpful to understand the societal impact of AI. Another interesting dimension of user heterogeneity concerns upload activity, which is an important metric to draw insights for platform governance. In Table 5, we show that the main effect reported in column (1) of Table 3 varies by percentile of a user's total upload activity. The point estimates in columns (1)–(4) of Table 5 are smaller and not always significantly different from the estimate in column (1) of Table 3 in the sense that 90% confidence bands overlap. However,

Table 5: Results: Changes in uploads, by user activity

	(1) All	(2) w/o 99th	(3) w/o 95th	(4) w/o 90th
Post × Treated	-0.4529***	-0.9872***	-0.6847***	-0.4571***
	(0.02672)	(0.05445)	(0.03513)	(0.02699)
Post × Treated × Above99thPct	-17.4338			
	(11.68857)			
Post × Treated × Above95thPct	-0.9941			
	(2.04651)			
Post × Treated × Above90thPct	-1.3862*			
	(0.74521)			
Mean Treated Before	2.9038	2.2160	1.5488	1.1435
Month FE	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes
Observations	612,662	606,490	581,988	551,076

Note: The dependent variable is the number of uploads in a given month. *Treated* indicates whether a user was part of the LITE dataset. *Above99thPct*, *Above95thPct* and *Above90thPct* indicate whether the users' uploads were above the 99/95/90th percentile. The samples in columns (2) to (4) do not include users above the 99/95/90th percentile of uploads. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table 6: Results: Changes in uploads, by types of images

	(1) Nature	(2) Curated	(3) Nature & Curated
Post × Treated	-0.2265***	-0.3472***	-0.1392***
	(0.02956)	(0.01642)	(0.00614)
Mean Treated Before	0.7218	0.5739	0.2022
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Observations	612,662	612,662	612,662

Note: The dependent variable is an indicator of whether a user account has uploaded at least one image in a given month. *Treated* indicates whether a user was part of the LITE dataset. Fixed effects for upload month of the image and user account. Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

in absolute terms, the implied effect sizes are larger (column 2: -45%, column 3: -44% and column 4: -40%) compared to the -30% of the baseline. Given that we observe the universe of activity on Unsplash, we prefer not to remove particularly prolific users and report estimates of average treatment effects based on all users as our more conservative baseline results.

Table 7: Results: Changes in uploads, similarity of images

	Avg. Similarity		LogVerySimilar (0.8)		LogVerySimilar (0.9)	
	(1)	(2)	(3)	(4)	(5)	(6)
Post × Treated	0.0189***	0.0070**	0.2520***	0.0749	0.1336***	0.0380
	(0.00097)	(0.00345)	(0.01817)	(0.05707)	(0.02300)	(0.06871)
Mean Treated Before	0.3910	0.3910	8.1285	8.1285	4.5527	4.5527
Image-Age FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	No	Yes	No	Yes	No	Yes
Observations	340,991	340,346	340,991	340,346	340,991	340,346

Note: The dependent variable in columns (1) and (2) measures how similar an uploaded image is relative to the average image in the stock of images uploaded by treated vs. control users up until one year before the creation of the LITE dataset. In columns (3)–(6), the dependent variable is the log+1 count of images from the stock of images uploaded by treated/control users up until one year before the creation of the LITE dataset that have a cosine similarity score of larger than 0.8 or 0.9 relative an uploaded image. *Treated* indicates whether a user was part of the LITE dataset. Fixed effects for upload month of the image and user account in columns (2), (4) and (6). Standard errors in parentheses, White-robust in columns (1), (3) and (5), and clustered at the user level in columns (2), (4) and (6). * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

5.3 Changes in the types of images

5.3.1 Nature-themed and curated images

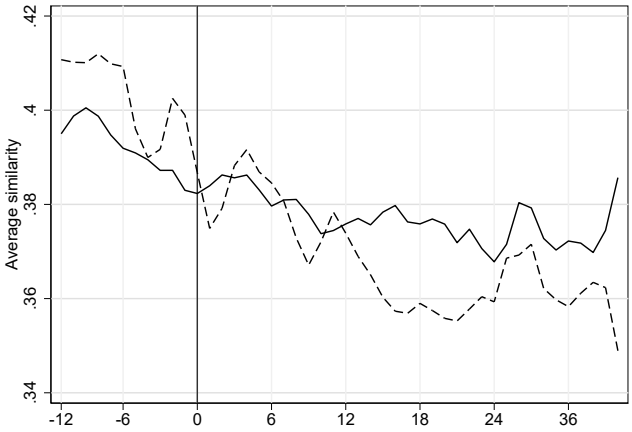
Next, we study whether the types of images uploaded to Unsplash have changed. In column (1) of Table 6, we show that the likelihood of uploading at least one nature-themed image decreases by 23 percentage points or 30% relative to the average in the before period. Note that this is about the same effect size we get when looking at all types of images in column (2) of Table 3. However, columns (2) and (3) of Table 6, suggest that the extensive margin decreases at 60% for curated images and at 70% for nature-themed and curated images. If we believe that Unsplash's curation team picks high-quality images, then this suggests a reduction in the artistic quality of uploads.

5.3.2 Analysis of Variety and Novelty measures

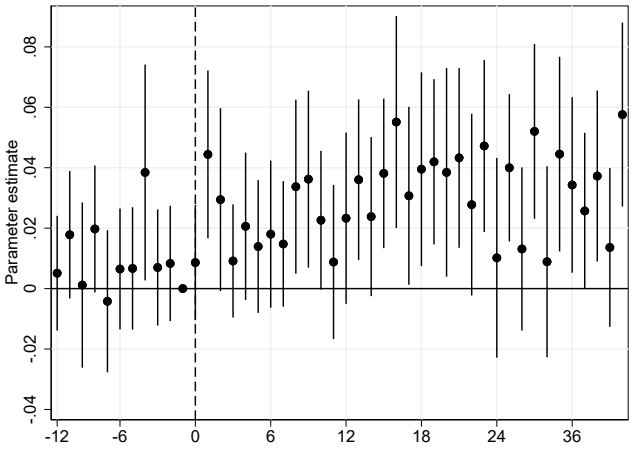
Next, we study how strategic behavior affects the variety and novelty of contributions more directly. In Table 7, we report results from our analysis of similarity using natural language processing of user-specified keywords of uploaded images. We use the same difference-in-differences identification strategy as above. Panel A of Figure 2 shows that the unconditional mean of the average similarity to the pre-existing stock of photos decreases less for images uploaded by treated users than for images uploaded by control users. In other words, treated users upload images that are less novel and more similar to the existing stock of images. This is confirmed more formally in Panel B, where we plot

Figure 2: Changes in uploads, similarity of images

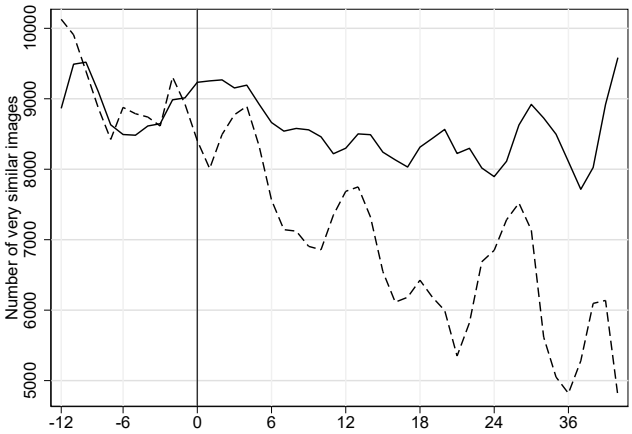
Panel A: Levels of average similarity



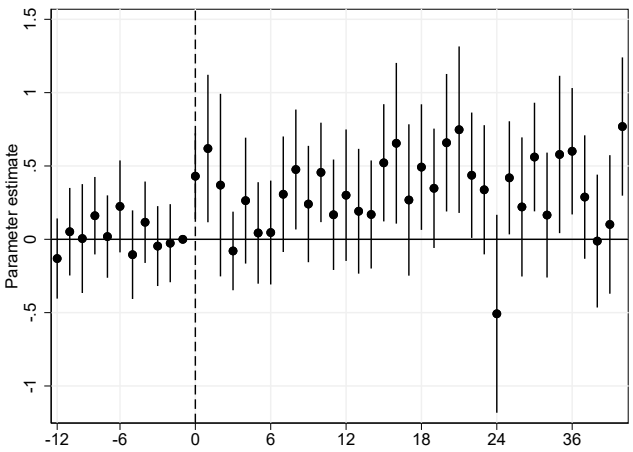
Panel B: Differences in average similarity



Panel C: Levels of very similar count



Panel D: Differences in very similar count



Note: In Panel A, we depict the average similarity of uploaded images by users in the treatment group (solid) and by users in the control group (dashed). In Panel C, we plot the number of images uploaded by treated users (solid) and control users (dashed) that are very similar to the existing stock (threshold 0.8). In Panels B and D, we plot OLS estimates of the δ_τ coefficients are obtained from $Y_{it} = \sum_{\tau \in T} \nu_\tau + \sum_{\tau \in T} \delta_\tau \left(\gamma_\tau \times Treated_i\right) + \varepsilon_{it}$, with the number of uploads of user i per month t as the dependent variable. The dots reflect month-specific point estimates comparing treatment users to control users. Standard errors are clustered at the user level, and bars indicate 90% confidence bands. In all panels, the vertical line indicates the creation of LITE 1.0.0, and numbers on the horizontal axis indicate the number of months before and after that.

estimates of the difference between the treatment and control group in each month, before and after the release of the LITE dataset, from a model without user-fixed effects. In Panels C and D, we turn to a count of the number of very similar images. Again, we see that the decrease in very similar images uploaded by users in the treatment group is less steep than in the control group.

We quantify these differences in Table 7. We find that it is not users who change the type of images they upload, but a difference in the type of uploaded image is coming from a change in the composition of users. In the specifications with user-fixed effects in columns (2), (4) and (6), we do not find much evidence that the release of the LITE dataset caused new uploads to differ from existing uploads. This is especially the case concerning the number of highly similar images in the existing stock. Hence, users do not change the types of images they upload. However, looking at the results from specifications without user-fixed effects in columns (1), (3) and (5), we do see that uploaded images become more similar to the existing stock. Column (1) suggests an increase of about 5% in average similarity for uploads by treated users, whereas columns (3) and (5) suggest an increase of about 28% and 13%, respectively, in the number of highly similar images. Overall, the results imply that the composition of the training dataset will change because of changes in the user composition and the types of images they upload, not because individual users change the types of images they upload. This is in line with the results we describe above. In what follows, we explore potential mechanisms that may explain the changes in user behavior.

5.4 Mechanisms: Awareness of AI potential and treatment intensity

The results in Table 8 serve to highlight potential mechanisms through which the inclusion in the LITE dataset induces users to reduce their upload activity. In column (1), informed by Figure 1, we split the post-period in two parts. The first 24 months correspond to the period between June 2020 and August 2022, a period with modest public interest in AI and AI art in particular. After August 2022, however, there was a substantial increase in media coverage on the topic, e.g., a New York Times article titled "We Need to Talk About How Good A.I. Is Getting", which may have further raised awareness about genAI in the Unsplash community.[19] The timing also coincides with the public release of Stable Diffusion, the first text-to-image generator that could be run on consumer-grade machines, democratizing access to genAI technology after similar technology released a few months earlier had only restricted public access (e.g. DALL-E and Midjourney).[20] We find a treatment effect of 34% before August 2022 and 49% thereafter, which corresponds to a significant 40% increase in the treatment effect. This suggests that, as awareness of the commercial potential of image generation AI raises, photographers reduce their contributions to Unsplash and, therefore, to the flow of AI training data.

[19] see https://www.nytimes.com/2022/08/24/technology/ai-technology-progress.html

[20] See https://stability.ai/news/stable-diffusion-announcement.

Table 8: Results: Changes in uploads, mechanisms

	Uploads	
	(1)	(2)
Post × Treated	-1.0056***	
	(0.10508)	
PostAug22 × Treated	-0.4054***	
	(0.10056)	
Post × TreatedSingle		-0.3059***
		(0.09750)
Post × TreatedMultiple		-2.2830***
		(0.17248)
Mean Treated Before	2.9038	
Mean TreatedSingle Before		1.4011
Mean TreatedMultiple Before		5.0982
Month FE	Yes	Yes
User FE	Yes	Yes
Observations	612,662	612,662

Note: The dependent variable is the total number of uploads per month. *Post* indicates the entire period after the creation of LITE and *PostAug22* indicates the period after August 2022, when groundbreaking genAI for images was released. *Treated* indicates whether a user was part of the LITE dataset. *TreatedSingle* indicates whether a user had exactly one image in the LITE dataset. *TreatedMultiple* indicates whether a user had more than one image in the LITE dataset. Fixed effects for upload month of the image and user account. Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

In column (2), we distinguish between users that were affected more or less intensively by the release of the LITE dataset. In particular, we distinguish between users who had one image included (about 60%) and users who had multiple images included (about 40%). Conditional on more than one image in the LITE dataset, the average number of images per user is 5.96 (standard deviation 14.18 and maximum 419). The estimated effect size is 22% for single-treated users and 45% for multiple-treated users. This difference is statistically significant. These results provide further evidence that users must have been aware of their images being included in the LITE dataset because otherwise, we would not expect a stronger response from more heavily treated users. This result has interesting implications for policy, suggesting that a one-size-fits-all approach to licensing of human-generated work for AI training datasets can lead to unintended heterogeneous effects on the flow of training data. As a consequence, our results suggest that a licensing scheme that ensures a continued flow of data may need to compensate more active contributors to a larger degree (Wang et al., 2024).

Finally, it is worth noting that we observe that users in the treatment group are three times more likely to be contributors in the Unsplash+ program (t-statistic in a two-sided t-test is 6.13). This

program enables users to gain monetary compensation and explicitly disallows the use of their works for AI training. Hence, we interpret the substantially higher uptake of Unsplash+ among users who had at least one image in the LITE dataset as a strategic response to counteract economic forces and to avoid the utilization of their works in AI applications.

5.5 Robustness

5.5.1 Alternative control group definition

Table A.1 repeats the specification of our baseline estimations in Table 3 but with different definitions of the control group. In columns (1)–(3), we report results using a control group composed of users that did not have nature-themed (i.e., not with the keyword "nature") and curated images at the time when the LITE dataset was created. The effect size for the number of uploads is 47% in column (1), 33% in terms of the extensive margin in column (2), and 19% when using a specification with the log-transformed number of uploads in column (3). In columns (4)–(6), we use a control group comprised of all users whose images were not included in the LITE dataset. The effect sizes are 50%, 40% and 21%, respectively. Note that the estimates reported in Table A.1 are not significantly different from those reported in Table 3 in the sense that 90% confidence bands overlap.

5.5.2 OLS versus non-linear models

Given that our outcome variable is both a count, i.e. a non-negative integer, and zero in 86.6% of the observations, one might be worried that OLS is not the optimal choice of estimator (Wooldridge, 2023). Table A.2 shows that our baseline results are broadly robust to using a Poisson pseudo-maximum likelihood regression (PPML) approach.[21] However, the results indicate that PPML is sensitive to large values of the DV, which is of economic interest in our setting (i.e. heavy platform users). In addition, the interaction effect in non-linear models cannot be directly interpreted as the average treatment effect on the treated (Ai and Norton, 2003; Puhani, 2012) while simultaneously imposing additional restrictions for parallel trends to hold (Roth and Sant'Anna, 2023). For these reasons, and for simplicity in the interpretation of effect sizes, we report estimates from OLS regressions throughout.

[21] We do not report results from a negative binomial model because PPML is preferred for its robustness (Blackburn, 2015).

6 Discussion

6.1 Back-of-the-envelope estimation of aggregate effects

Having estimated the average treatment effects following the release of the LITE dataset on several key outcomes, we can now use these estimations to simulate several counterfactuals in which we vary the proportion of users treated.

To proceed, we first need to define a few key variables. The share of users included in the LITE dataset is λ, the effect of being in the LITE dataset is $\hat{\delta}$, and the factual growth rate of the number of images per user in the post-period is g. With this, we can calculate a counterfactual growth rate $\bar{g} = g/(1-\lambda+\lambda\hat{\delta})$, which is the growth rate that would have materialized had the LITE dataset not been released. Then, we can use a few descriptive statistics as well as our causal estimates to attempt a back-of-the-envelope calculation of counterfactuals of key outcomes such as the total number of uploads. Finally, once we have calculated $\bar{g}$, we can simulate any growth rate $\tilde{g}(\lambda) = (1-\lambda)(1+\bar{g}) + \lambda(1+\hat{\delta}\bar{g}) - 1$. This is useful to calculate counterfactuals where we can vary the size of the LITE training dataset and approximate how the size of the LITE training dataset affects the flow of data in terms of the aggregate number of data points added to Unsplash.

To simulate the impact of the share of treated users on the number of uploads, we use the estimated average treatment effect on the treated and the mean uploads of treated users from Table 3. This leads to $\hat{\delta} = 1 - (1.1099/2.9038) = 0.6$. When the LITE dataset was created in June 2020, Unsplash had a stock of $S_0 = 1,931,324$ images uploaded by $U_0 = 198,792$ users. The LITE dataset encompasses $U_L = 8,298$ users, such that the share of users in the LITE dataset was $\lambda = U_L/U_0 = 0.042$. About three years later, in May 2023, Unsplash has $S_t = 4,960,120$ images and $U_t = 338,635$ users. Hence, the factual growth rate $g = \frac{S_t}{U_t}/\frac{S_0}{U_0} - 1 = 14.65/9.72 - 1 = 0.51$ and therefore the counterfactual growth rate $\bar{g} = (1.51)/(1 - 0.042 + 0.042 \times 0.6) = 0.52$. In absolute terms, without the release of the LITE dataset, users would have uploaded 19,313 more images to Unsplash ($= S_0(0.52 - 0.51)$).

We can now extrapolate that a twice-as-large LITE dataset would have reduced the flow of data by 2% (=1-0.50/0.51) compared to the factual and making the entire catalog available for commercial AI research would have reduced the flow of data by 39% (=1-0.31/0.51), corresponding to 386,264 images uploaded less over the period.

6.2 How different is Unsplash from other platforms?

To provide some insight into this question, it is useful to compare Unsplash to Instagram, the by far largest image-sharing platform. According to Instagram's terms of service, users retain ownership of any intellectual property rights that they hold in that content. However, by posting content to Instagram, a user grants the platform a non-exclusive, royalty-free, transferable, sublicensable, worldwide license to use the content they post.[22] This means Instagram can use, reproduce, modify, display, and distribute the content, and they can also allow others to do so within their platform or in association with other media. This license ends when a user deletes the content or the account unless the content has been shared with others and they have not deleted it. Unsplash operates in a similar fashion yet in an even more permissive fashion. When a user uploads an image to Unsplash, they grant anyone a license to use their image for free, including for commercial purposes, without requiring permission from or providing attribution to the user.[23] By uploading photos to Unsplash, users agree to this license, which effectively allows anyone worldwide to use the images with very few restrictions. It is important to note that while a user still retains the copyright to their image, the rights they grant through the Unsplash license are much broader in scope and do not end even if they remove the image from the platform. A key difference there is the control over use. On Instagram, the use of an uploaded image is mostly within the context of the platform, though the license a user grants Instagram is quite broad. Unsplash, in contrast, allows anyone to use the images for nearly any purpose, including commercially and outside the platform, without any compensation or credit to the uploader. Instagram's terms are not explicitly clear about commercial use by third parties unless it is sublicensed, while Unsplash explicitly allows commercial use by anyone.

In Table 9, we study whether users' reaction to having their images included in an AI training dataset affects only their uploads to Unsplash, or whether they also change their behavior on Instagram. To do so, we compare the number of uploads of treated versus control users on Unsplash to their uploads on Instagram. In column (1), we see that the average treatment effect with respect to uploads on Unsplash is about half the size of our baseline (column 1 in Table 3) in this restricted sample. In column (2), we do not find a significant effect on Instagram uploads, as the point estimate is very close to zero. In columns (3) and (4), we combine information from both platforms to compare effect sizes directly. The linear combination in both sets of results, with user fixed effects and with user-platform

[22] See https://help.instagram.com/478745558852511/.

[23] See https://unsplash.com/license.

Table 9: Results: Changes in uploads, Unsplash vs. Instagram

	Unsplash	Instagram	Both	
	(1)	(2)	(3)	(4)
Treated × Post	-0.5118***	-0.0013	-0.4344***	-0.5111***
	(0.08951)	(0.02326)	(0.06900)	(0.06537)
Instagram			0.0704	
			(0.06441)	
Treated × Instagram			-0.7933***	
			(0.07820)	
Post × Instagram			0.0003	-0.0093
			(0.07802)	(0.07604)
Treated × Post × Instagram			0.3556***	0.5090***
			(0.09457)	(0.09240)
Constant	0.9318***	0.3874***	0.8952***	0.6627***
	(0.04588)	(0.01192)	(0.03553)	(0.03525)
Total effect Instagram				
lincom			-0.0787	-0.0020
			(0.06900)	(0.06537)
Mean DV	1.4300	0.7071	1.0686	1.0686
Month FE	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	No
User-Platform FE	No	No	No	Yes
Observations	56,855	56,855	113,728	113,728

Note: The dependent variable is the number of monthly uploads to either Unsplash or Instagram, or both. *Treated* indicates whether a user was part of the LITE dataset. Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$ *

fixed effects, again indicates that the effect on Instagram is not significant. This implies that users did not switch to Instagram as a new outlet for their photography work, but it also does not imply that users stopped to publish their work anywhere online. Overall, our results suggest that adding Unsplash images to an Unsplash-curated AI training dataset affects user behavior on Unsplash, but not elsewhere.

7 Conclusion

GenAI models benefit from the abundance of available training data on the internet. However, the economics of the supply of data have not been discussed much. In this research, we focus on how strategic behavior of human creators affects the flow and, therefore, the future stock of AI training data.

We study the empirical setting of creators who openly contribute their work to Unsplash, a stock photography platform that provides crowd-sourced images under a permissive license. In particular, we evaluate the effects of the release of a training dataset for commercial AI applications. We find that affected users are more likely to leave the platform. Those that stay, reduce their uploads substantially. As a result, the size of the flow of data changes, as does the content. Our results imply that the release of a free training dataset for commercial AI applications, covering roughly 5% of the platform's contributors, overall reduced the flow of data by 2%, decreased the variety (deviation from the average concept in the stock) in the dataset by 4% and increased novelty (repetition of very similar concepts in the stock) by 2.5% over the course of three years.

Our results have important implications for the debate on copyright and AI. Even in a setting where copyright does not directly factor in – Unsplash caters to a special set of creators that choose an open license –, we see a sizable change in individuals' behavior when their work is made available as an AI training dataset. Our results suggest that a change in the flow of data can also translate into changes in aspects of dataset quality, such as variety and novelty, because of changes in the user composition. Consequently, this has implications for the quality of AI training datasets and possibly AI output (Hadsell et al., 2020; He et al., 2011; Quinn and Gutt, 2023; Whang et al., 2023).

In conclusion, our results raise the question of whether exemptions in copyright law for AI (e.g. through fair use provisions that do not compensate rightsholders) could lead society to a situation of lower quality AI (smaller and less varied datasets) and less interesting (because less varied and more repetitive) human creations. Hence, while AI developers may initially prefer laissez-faire copyright (e.g., fair use or text-and-data-mining exceptions), the upstream consequences on the flow of data, both in terms of quantity and quality, may have negative consequences for the continued quality of existing AI applications and raise difficulties in the development of new technologies.

More research is needed to guide policymakers to carefully weigh the trade-offs between regulation of the supply of data, e.g. through privacy policy, competition policy, or copyright law, and AI innovation.

References

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2022). "Artificial intelligence and jobs: Evidence from online vacancies." *Journal of Labor Economics*, *40*, S293–S340.

Ai, C., and Norton, E. C. (2003). "Interaction terms in logit and probit models." *Economics Letters*, *80*(1), 123–129.

Bauer, J., Franke, N., and Tuertscher, P. (2016). "Intellectual Property Norms in Online Communities: How User-Organized Intellectual Property Regulation Supports Innovation." *Information Systems Research*, *27*(4), 724–750.

Beraja, M., Yang, D. Y., and Yuchtman, N. (2023). "Data-intensive innovation and the state: Evidence from ai firms in china." *The Review of Economic Studies*, *90*(4), 1701–1723.

Birhane, A., and Prabhu, V. U. (2021). "Large image datasets: A pyrrhic win for computer vision?" In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1536–1546, IEEE.

Blackburn, M. L. (2015). "The relative performance of poisson and negative binomial regression estimators." *Oxford Bulletin of Economics and Statistics*, *77*(4), 605–616.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). "Generative AI at Work." *NBER Working Paper*, (31161).

Burtch, G., Lee, D., and Chen, Z. (2024). "Generative ai degrades online communities." *Communications of the ACM*, *67*(3), 40–42.

del Rio-Chanona, M., Laurentsyeva, N., and Wachs, J. (2023). "Are large language models a threat to digital public goods? evidence from activity on stack overflow." *Working Paper*.

Doshi, A. R., and Hauser, O. (2023). "Generative artificial intelligence enhances creativity." *SSRN Working Paper*.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). "Gpts are gpts: An early look at the labor market impact potential of large language models." *arXiv preprint arXiv:2303.10130*.

Eshraghian, J. K. (2020). "Human ownership of artificial creativity." *Nature Machine Intelligence*, *2*(3), 157–160.

Farboodi, M., Mihet, R., Philippon, T., and Veldkamp, L. (2019). "Big data and firm dynamics." In *AEA Papers and Proceedings*, vol. 109, 38–42, JSTOR.

Felten, E. W., Raj, M., and Seamans, R. (2023). "How will language modelers like chatgpt affect occupations and industries?" *Working Paper*.

Fiil-Flynn, S. M., Butler, B., Carroll, M., Cohen-Sasson, O., Craig, C., Guibault, L., Jaszi, P., Jütte, B. J., Katz, A., Quintais, J. P., Margoni, T., de Souza, A. R., Sag, M., Samberg, R., Schirru, L., Senftleben, M., Tur-Sinai, O., and Contreras, J. L. (2022). "Legal reform to enhance global text and data mining research." *Science*, *378*(6623), 951–953.

Gallea, Q. (2023). "From mundane to meaningful: Ai's influence on work dynamics–evidence from chatgpt and stack overflow." *arXiv preprint arXiv:2308.11302*.

Gans, J. (2024). "Copyright policy options for generative artificial intelligence." *NBER Working paper*, (32106).

Godinho de Matos, M., and Adjerid, I. (2022). "Consumer Consent and Firm Targeting After GDPR: The Case of a Large Telecom Provider." *Management Science*, *68*(5), 3330–3378.

Guilbeault, D., Delecourt, S., Hull, T., Desikan, B. S., Chu, M., and Nadler, E. (2024). "Online images amplify gender bias." *Nature*, 1–7.

Hadsell, R., Rao, D., Rusu, A. A., and Pascanu, R. (2020). "Embracing change: Continual learning in deep neural networks." *Trends in cognitive sciences*, *24*(12), 1028–1040.

He, H., Chen, S., Li, K., and Xu, X. (2011). "Incremental learning from stream data." *IEEE Transactions on Neural Networks*, *22*(12), 1901–1914.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. (2023). "Foundation models and fair use." *arXiv preprint arXiv:2303.15715*.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. (2024). "Algorithmic progress in language models." (2403.05812).

Huang, H., Fu, R., and Ghose, A. (2023). "Generative AI and Content-Creator Economy: Evidence from Online Content Creation Platforms." *SSRN Working Paper*.

Johnson, G. (2022). "Economic research on privacy regulation: Lessons from the gdpr and beyond."

Jones, C. I., and Tonetti, C. (2020). "Nonrivalry and the economics of data." *American Economic Review*, *110*, 2819–58.

Lei, X., Chen, Y., and Sen, A. (2023). "The value of external data for digital platforms: Evidence from a field experiment on search suggestions." *Available at SSRN*.

Levendowski, A. (2018). "How copyright law can fix artificial intelligence's implicit bias problem." *Wash. L. Rev.*, *93*, 579.

Lin, S. (2024). "From creation to caution: The effect of ai on online art market." *Working Paper*.

Martens, B., Parker, G., Petropoulos, G., and Van Alstyne, M. W. (2024). "Towards efficient information sharing in network markets." In *Proceedings of the 57th Hawaii International Conference on System Sciences*.

McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L., and Zolas, N. (2024). "AI adoption in America: Who, what, and where." *Journal of Economics & Management Strategy*, *108*(11), 3451–3491.

Metz, C., Kang, C., Frenkel, S., Thompson, S. A., and Grant, N. (2024). "How tech giants cut corners to harvest data for a.i." *The New York Times*, updated April 8, 2024. Reporting from San Francisco, Washington, and New York.

Neumann, N., Tucker, C. E., and Whitfield, T. (2019). "How effective is third-party consumer profiling? evidence from field studies." *Marketing Science*, *38*(6), 918–926.

Peukert, C., Sen, A., and Claussen, J. (2023). "The editor and the algorithm: Recommendation technology in online news." *Management Science, forthcoming*.

Peukert, C., and Windisch, M. (2024). "The economics of copyright in the digital age." *Journal of Economic Surveys, forthcoming*.

Prüfer, J., and Schottmüller, C. (2021). "Competing with big data." *The Journal of Industrial Economics*, *69*(4), 967–1008.

Puhani, P. A. (2012). "The treatment effect, the cross difference, and the interaction term in nonlinear difference-in-differences models." *Economics Letters*, *115*(1), 85–87.

Quinn, M., and Gutt, D. (2023). "Does generative ai erode its own training data? empirical evidence of the effects on data quantity and characteristics from a qa platform." *SSRN Working Paper*.

Roth, J., and Sant'Anna, P. H. C. (2023). "When is parallel trends sensitive to functional form?" *Econometrica*, *91*(2), 737–747.

Samuelson, P. (2023). "Generative ai meets copyright." *Science*, *381*(6654), 158–161.

Sokol, D. D., and Van Alstyne, M. (2021). "The rising risk of platform regulation." *MIT Sloan Management Review*, *62*(2), 6A–10A.

Sun, T., Yuan, Z., Li, C., Zhang, K., and Xu, J. (2023). "The value of personal data in internet commerce: A high-stake field experiment on data regulation policy." *Management Science, forthcoming*.

Valavi, E., Hestness, J., Ardalani, N., and Iansiti, M. (2022). "Time and the value of data." *arXiv preprint arXiv:2203.09118*.

Wang, J. T., Deng, Z., Chiba-Okabe, H., Barak, B., and Su, W. J. (2024). "An economic solution to copyright challenges of generative ai." *arXiv:2404.13964 [cs.LG]*.

Whang, S. E., Roh, Y., Song, H., and Lee, J.-G. (2023). "Data collection and quality challenges in deep learning: A data-centric ai perspective." *The VLDB Journal*, *32*(4), 791–813.

Wooldridge, J. M. (2023). "Simple approaches to nonlinear difference-in-differences with panel data." *The Econometrics Journal*, *26*(3), C31–C66.

Wu, L., Hitt, L., and Lou, B. (2020). "Data analytics, innovation, and firm productivity." *Management Science*, *66*(5), 2017–2039.

Yang, S. A., and Zhang, A. H. (2024). "Generative ai and copyright: A dynamic perspective." *arXiv preprint arXiv:2402.17801*.

Appendix

Table A.1: Results: Changes in uploads, alternative control group definitions

	Not Nature & Curated			Not in LITE		
	(1) Uploads	(2) I(Upl.>0)	(3) Log(1+Upl.)	(4) Uploads	(5) I(Upl.>0)	(6) Log(1+Upl.)
Post × Treated	-1.3737***	-0.0857***	-0.2072***	-1.4495***	-0.1041***	-0.2323***
	(0.08534)	(0.00238)	(0.00539)	(0.08527)	(0.00236)	(0.00537)
Mean Treated Before	2.9038	0.2635	0.4892	2.9038	0.2635	0.4892
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,666,123	7,666,123	7,666,123	9,704,756	9,704,756	9,704,756

Note: The dependent variable in columns (1) and (4) is the total number of uploads per month, in columns (2) and (5) it is an indicator of whether a user account has uploaded at least one image in a given month, and it is the log number of uploads in columns (3) and (6). The control group in columns (1)–(3) includes users that have uploaded at least one non-curated and non-nature themed image. The control group in columns (4)–(6) includes all users that do not appear in the LITE data. *Treated* indicates whether a user was part of the LITE dataset. Fixed effects for upload month of the image and user account. Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

Table A.2: Results: Changes in uploads, Poisson models

	All		w/o 99th	w/o 95th	w/o 90th
	(1)	(2)	(3)	(4)	(5)
Post × Treated	0.0607	-0.1996***	-0.0541	-0.1703***	-0.2008***
	(0.12357)	(0.05497)	(0.06611)	(0.05785)	(0.05866)
Post × Treated × Above99thPct		-0.3154			
		(0.38920)			
Post × Treated × Above95thPct		0.1842			
		(0.24548)			
Post × Treated × Above90thPct		-0.2198			
		(0.16234)			
Mean Treated Before	2.9038	2.9038	2.2160	1.5488	1.1435
Month FE	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes
Observations	468,555	468,555	462,383	437,881	406,969

Note: We estimate Poisson pseudo-maximum likelihood regressions models. The dependent variable is the number of uploads in a given month. *Treated* indicates whether a user was part of the LITE dataset. *Above99thPct*, *Above95thPct*, and *Above90thPct* indicate whether a user's uploads were above the 99/95/90th percentile. The samples in columns (3) to (5) do not include users above the 99/95/90th percentile of the pre-treatment uploads. The number of observations is smaller than in Tables 3 and 5 because singleton observations are dropped due to fixed effects. Standard errors clustered at the user level in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$ * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$