Understanding Creators' Acceptance of Content Reuse

Riya Sinha
The University of Texas at Austin
Austin, Texas, United States
riyasinha@utexas.edu

Hanlin Li
The University of Texas at Austin
Austin, Texas, USA
lihanlin@utexas.edu

ABSTRACT

Content creators' work has long been reused by various technology developers and researchers; recently joining the ranks of content reusers are emergent generative AI models. However, creators' perspectives and acceptances of these use cases were often overlooked in the process. We present our initial findings on how content creators perceive scenarios of content reuse as acceptable, negotiable, or unacceptable. We surveyed 478 content creators on Instagram and a non-trivial percentage of respondents found content reuse for search engine display, education, and academic research acceptable. Conversely, content reuse for commercial or non-profit generative AI models was a negotiable scenario for most respondents. Receiving compensation was likely to make almost all use cases more acceptable to content creators. Our work paves the way for future work into understanding how creators value their work, research ethics, and best practices for using user-generated content for responsible technological innovation.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

User-Generated Content; Web Scraping; Content Reuse

ACM Reference Format:
Riya Sinha and Hanlin Li. 2024. Understanding Creators' Acceptance of Content Reuse. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9–13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3678884.3681911

1 INTRODUCTION

Content reuse has been a standard practice in many prominent technology business models. For example, search engines, online directories, and data brokers rely on indexing or reusing other websites' content to support search results, listings, and even background checks, respectively [5, 12]. More recently, prominent generative models like ChatGPT and Stable Diffusion are possible thanks to the billions of webpages and images scraped from the web [17, 20, 23, 34].

The emergence of generative models brought content reuse to the forefront of creators' attention. Many became concerned about how far-reaching AI companies' data-collecting tentacles are and subsequently took action to protect their work from being scraped and reused[4]. For example, some fan fiction writers have turned their pages to private or have flooded their pages with irrelevant, low-quality content [30]. Other writers and publishers have resorted to legal means to counter the unauthorized scraping and reuse of their content [13]. Creators from StackOverflow have sought to delete their content after learning the site's partnership with OpenAI [31].

For practitioners to develop sustainable relationships with creators and for policymakers to design AI policies to guide responsible innovation, it is important to gain a comprehensive understanding of content creators' reactions to content scraping and reuse. Above all, AI models are possible thanks to the webpages and images content creators publish [5, 12]. Moreover, the expansion of content creators' protests makes it all the more important for AI companies to take the labor force's concerns into account [10, 11, 21]. AI companies' innovation hinges on creators' on-going contributions to the web [9].

However, there has not been a large-scale investigation of content creators' perspectives about content scraping and reuse, aside from anecdotal incidents reported by journalists. As prior work has shown how contextual factors have played a role in content creators' acceptance of content reuse for research [11, 15, 16], creators will likely hold a wide range of stances when it comes to content reuse for technological innovation.

To facilitate a more well-rounded discussion about AI policies and to better understand the content creation labor force, we conducted a survey with 478 Instagram users to understand *how content creators find different use cases acceptable, negotiable, or unacceptable.* We found that when prompted with specific use cases, a non-trivial percentage of respondents found content reuse for *search engine display*, *education*, and *academic research* acceptable. Conversely, content reuse for *commercial or non-profit generative AI models* was a negotiable scenario for most respondents. Providing compensation is likely to make almost all use cases more acceptable to content creators. Content creators also expressed strong desires for other conditions when their content is reused, depending on the specific purpose. For example, accreditation was preferred by 41% participants if their content were to be reused for research purposes.

Our work highlights that some content reuse cases may be more acceptable and negotiable than others, and therefore, no one consent mechanism can suit different content reuse cases. Put another way, creators may have divergent preferences or demands (e.g., accreditation vs. compensation) depending on for what purposes their content will be used. Our work also highlights the different degrees of acceptance of content reuse for research vs. technology

CSCW Companion '24, November 9–13, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-1114-5/24/11
https://doi.org/10.1145/3678884.3681911

development and innovation, highlighting the need for public policies that incentivize ethical, creator-driven content reuse practices in the tech industry.

2 RELATED WORK

2.1 Research ethics in collecting social media data

Researchers have examined how social media users perceive the use of their data for research purposes closely in a series of studies and found that while most people are generally accepting of their content being used for research, their attitudes also depend on a series of contextual factors [11, 15, 16]. For example, Black X (formerly known as Twitter) users have distinct concerns about the reuse of their content depending on the positionality and topic of the research [16].

We extend this line of work by contrasting the commercial reuse of social media data with academic reuse. This contrast will inform more specific policies about content reuse in general, e.g. different guidelines and regulations for researchers vs. commercial entities. This is especially relevant today given researchers' increasingly constrained access to API services, as companies make blanket changes to their API services [14].

2.2 Web scraping, Data Donation, and Data Marketplaces

User-generated content is a valuable asset for various stakeholders, including hosting platforms and third parties, particularly AI companies and developers [32]. While downloading content from APIs is likely the most straightforward approach, it is not always possible or affordable. Moreover, not all APIs are well documented, leading to substantial learning challenges for developers [8, 22, 27]. To collect user-generated content, practitioners and researchers have developed a variety of apparatus, such as web crawlers, data marketplaces, and data donation stations.

Web crawling, or scraping, is an approach researchers and developers take to pragmatically extract information from a web page, typically through an automatic program (i.e. crawler). It has been a prominent practice for the past few decades [5]. For example, e-commerce websites use scraping to collect information from their competitors for competitive pricing. In academia, computational social scientists, communication scholars, and computing researchers have used scraping to construct datasets for social science insights (e.g. [2]) and model training (e.g. [28]).

Data marketplaces serve as platforms where businesses can buy, sell, or exchange datasets. Companies like X (formerly known as Twitter) and Facebook have explored monetization strategies by offering access to their data through APIs. There are also data brokers that aggregate data from external sources to sell to their clients. These data marketplaces enable businesses to access valuable insights for targeted advertising, market research, and trend analysis [3, 6, 24].

Data donation involves users voluntarily contributing their personal data to be used for research, analysis, or other purposes, commonly for non-profit or social issues. Researchers across fields have used data donation approaches to gather data for their research [26, 33]. Crowdsourcing data is widely used in the public health community and crisis communication such as disaster response. COVID-19 saw the rise of citizen scientists, sharing their information in terms of crisis [7].

While these mechanisms of data collection are effective and powerful, they are not without ethical issues. Creators' perspectives are not always accounted for in scraping and data marketplaces. When creators donate their data for altruistic purposes, they may not fully comprehend the implications of how their data is being used or monetized. Our study aims to explicate creators' perceptions of content reuse.

3 METHODS

We conducted a survey consisting of a series of questions on Instagram users' acceptance of seven scenarios of content reuse. We took a convenience sampling approach and recruited participants from a network of University of California, Berkeley (UC Berkeley) students, alumni, and employees. The survey was fielded by the Experimental Social Science Laboratory (Xlab) at the university and active between May 2023 and December 2023 and we received 478 responses in total. While our survey responses highlight various levels of acceptance for different scenarios, due to our sampling approach, our respondents are unlikely to be representative of the entire Instagram user population.

The seven scenarios (Table 1) in the survey were constructed to reflect existing widely-used reuse cases of user-generated content. AI training datasets like Massive Web[25] and Dolma[29] use content for social media platforms. Moreover, like mentioned in Related Work, we also see that platforms also collect and share and sell data to third parties. We created seven scenarios of most prominent reuse to gauge how respondents' reaction differ to scenarios where there is a monetary benefit to the host platforms or third parties (like *selling content in bulk, search engine display, commercial generative AI*) vs. scenarios where there might be perceived social benefits (*education, academic research*).

S1:	*Imagine a search engine (e.g., Google Search) displayed your content in its search results.*
S2:	*Imagine a software company used your content to create AI-powered chatbots like ChatGPT and image generation tools like Dall-E and Stable Diffusion.*
S3:	*Imagine a nonprofit organization used your content to create open-source, AI-powered chatbots like ChatGPT and image generation tools like Dall-E and Stable Diffusion.*
S4:	*Imagine a nonprofit organization (e.g. Wikipedia) used your content for educational purposes.*
S5:	*Imagine a research team at a university used your content to conduct social media research.*
S6:	*Imagine the content platform downloaded your content to sell it in bulk for a fee.*
S7:	*Imagine a people search website used your content to let others find your public records and profiles.*

Table 1: Scenarios

Participants can indicate each scenario as acceptable, unacceptable, and negotiable to themselves by selecting responses shown in Table 2. In the survey, Response 1 indicated non-acceptance of the scenario, and Response 6 indicated unconditional acceptance. Response 2 to 5 represented negotiable acceptances of the scenario. Participants who selected Response 1 (non-acceptance) or Response 6 (unconditional acceptance) were not allowed to select any of the negotiable responses (Responses 2 to 5). Respondents were permitted to select multiple responses from Responses 2 to 5, as long as they did not select Response 1 or Response 6. We also asked participants to share any additional thoughts through free text responses at the end of the survey.

Response 1:	*I will never accept this.*
Response 2:	*If I was accredited, I might be okay with this*
Response 3:	*If I received compensation from the company, I might be okay with this.*
Response 4:	*If I was anonymized, I might be okay with this.*
Response 5:	*If I could control what images are used, I might be okay with this.*
Response 6:	*I accept this unconditionally.*

Table 2: Responses

4 RESULTS

4.1 Overview of Responses

Table 3 shows the number of times each option was selected by respondents across all seven scenarios. Below, we first discussed what scenarios are least and most acceptable to respondents overall, respectively, by examining the first row and the last row of the table. We then provided a description of respondents' answers per scenario, examining the table vertically.

At a high level, our scenarios elicit varying levels of acceptance. Looking at the occasions in which respondents selected "I will never accept this", we found that they concentrated on four scenarios: *Commercial generative AI, non-profit open source generative AI, content resale in bulk*, and *people search websites*. In contrast, very few respondents, less than 10% (of 478 respondents), found scenarios of *education* and *academic research* unacceptable. This suggested that the vast majority of respondents were open to the idea of their content being used for purposes that promote these societal benefits.

Looking at unconditional acceptance, we found that respondents who selected this response largely concentrated on three scenarios: *search engine display*, *education*, and *academic research*. The rest of the scenarios were rarely unconditionally accepted by respondents (less than 10%). The low unconditional acceptance rates of these scenarios suggested that, when prompted, creators were unlikely to consent to having their content used for generative AI development unconditionally (regardless of whether it is for commercial or non-profit purposes) or resale. Below, we discuss under what conditions these scenarios became negotiable for respondents.

4.2 Varying responses to each scenario

Search Engine (S1). Respondents found search engines displaying their content to be a mostly negotiable scenario. Compensation was the most desired condition, followed by control. 155 (32%) respondents reported that if they were compensated, they might find the scenario acceptable. 145 respondents reported that if they could control what images were used, they might find it acceptable. The other two conditions, anonymization and accreditation, were also remarkably desired by over 20% of participants. Only 66 respondents (out of 478) said they would never accept their content being displayed in search engine results.

Commercial Generative AI (S2). Reuse of content for generative AI software by commercial software companies was a relatively negotiable scenario. 32 respondents accepted the scenario unconditionally, while 323 respondents selected one or more conditions. 123 respondents (25.7%) said they would never accept this scenario. The compensation condition was the most preferable, with 217 respondents selecting it. This strong preference leaves room for more in-depth investigation into to what extent compensation affects creators' perception and acceptance of generative AI.

Non-profit, Open Source Generative AI (S3). Content being used to create AI-powered chatbots by non-profit organizations was also a relatively negotiable scenario. 33 respondents accepted the scenario unconditionally, while 338 respondents selected one or more conditions to be met for them to consider this use case. 107 respondents (22%) said they would never accept their content being used for this scenario. Like *commercial generative AI* scenario, compensation was the most prominent condition, followed by control and anonymization.

Education (S4). Education saw more acceptance from respondents overall, with only 35 of them refusing to have their content used for this scenario. 76 participants accepted the scenario unconditionally, while 367 respondents selected one or more negotiable conditions. Like all other scenarios, compensation was the most preferred condition by respondents, with approximately 40% (189 out of 478) of them choosing it. Control (33%) was the second most preferred condition.

Academic Research (S5). Academic research was a generally negotiable, and at times unconditionally acceptable scenario, with the smallest number of respondents (15) rejecting it entirely. 105 respondents unconditionally accepted the scenario (the largest across all seven scenarios), and 358 respondents accepted the scenario with some conditions. Compensation was again the most preferred condition (47%), followed by accreditation.

Content Resale in Bulk (S6). Content resale in bulk is unlikely to be acceptable with 44% of respondents (210 out of 478) indicating they would never accept it. However, the other 192 respondents (40%) found it negotiable if they were compensated. The almost 50-50 division among the respondents indicated that this might be the most contentious scenario among respondents. As noticed in the responses for cases of control (75), compensation (192), anonymization (59), and accreditation (60), there is a clear indication of a strong aversion to unrestricted commercial use. Only 5% of respondents said they would unconditionally accept their content being displayed in search engine results.

People Search Website (S7). Creators being identified or displayed through their content on people search websites is an unacceptable scenario for 43% of respondents (206 out of 478). Only 10% of respondents (48 out of 478) said they would unconditionally accept their content being used by a people search website. The

	S1: Search Engine Display	S2: Commercial Generative AI	S3: Non-profit, open source Generative AI	S4: Education	S5: Academic Research	S6: Content Resale in Bulk	S7: People Search Websites
Non-Acceptance	66 (~ 14%)	123 (~ 26%)	107 (~ 22%)	35 (~ 7%)	15 (~ 3%)	210 (~ 44%)	206 (~ 43%)
Accreditation	113 (~ 24%)	92 (~ 19%)	98 (~ 21%)	148 (~ 31%)	196 (~ 41%)	60 (~ 13%)	60 (~ 13%)
Compensation	155 (~ 32%)	217 (~ 45%)	209 (~ 44%)	189 (~ 40%)	225 (~ 47%)	192 (~ 40%)	111 (~ 23%)
Anonymization	108 (~ 23%)	120 (~ 25%)	130 (~ 27%)	140 (~ 29%)	140 (~ 29%)	59 (~ 12%)	63 (~ 13%)
Control	145 (~ 30%)	109 (~ 23%)	120 (~ 25%)	156 (~ 33%)	144 (~ 30%)	75 (~ 16%)	88 (~ 18%)
Unconditional Acceptance	94 (~ 20%)	32 (~ 7%)	33 (~ 7%)	76 (~ 16%)	105 (~ 22%)	23 (~ 5%)	48 (~ 10%)

Table 3: Number of times each response was selected by our participants across all seven scenarios.

unconditional acceptance of this scenario is more than double that for *content resale in bulk*. Conversely, 111 respondents found this scenario acceptable if they were compensated, compared to 192 for *content resale in bulk*. This shows that participants are unlikely to be open to negotiating scenarios that might result in their personal information to be displayed or shared.

4.3 Additional concerns expressed in free-text responses

We asked for free text responses from participants about their additional concerns or reservations about their content being reused by third-party entities.

We identified three main themes in their responses: violation of privacy, non-consensual monetization, and lack of compensation for reuse. Privacy concerns were the most voiced, with many respondents wary of unauthorized use of their personal information and images. There was a clear desire for data usage to be consensual, with expectations of notification and control over how content was used. Moreover, participants mentioned experiences of personal information being used for deepfakes and other illegal uses, which resulted in a clear non-acceptance of scenarios where their personal information can be used to identify them. Monetization and compensation were also another source for concern, with respondents demanding fair compensation for their content, especially when used by companies for profit.

With respect to content reuse for generative AI models, creators extensively described their concerns. For respondents who used or were using alternate personas, anonymity was a major concern. Additionally, there was a varied perception of the value of content. Some respondents feeling their posts were not significant enough to be misused, while other respondents, like artists, saw a high value and potential for misuse of their creations.

It was also interesting to note that many respondents showcased a certain level of awareness when it comes to their content being scraped and reused for a variety of purposes. This prompts the need to delve deeper into understanding the awareness possessed by content creators and exploring what measures, if any, they are taking to mitigate such situations.

5 DISCUSSION

Our study showed the nuances in participants' (non)-acceptance of various use cases for their social media content. At a high level, we see a non-trivial percentage of blanket acceptance of content reuse for search engines, research, and education. However, across all use cases, a significant share of participants expressed varying degrees of preferences toward accreditation, anonymization, control, and compensation. These findings suggest that users do want to customize how their content is reused based on the specifics of a scenario.

5.1 No one-size-fits-all solutions for consent

Our findings highlight the importance of developing specific "affirmative consent" mechanisms for different data use cases. There is unlikely a one-size-fits-all solution given the different levels of (non)-acceptance of the seven provided use cases. One condition may be strongly preferred for a use case, but may not move any needle for another use case. Conversely, each condition may likely lead to disparate impact on acceptance across different use cases. For example, of the seven scenarios provided, *commercial generative AI, non-profit open source generative AI, education use, academic research, and content resale in bulk* prompted the most respondents to choose compensation. In contrast, *people search websites* only saw less than 25% of respondents choosing compensation. This disparity suggests that compensation, while a promising incentive for many creators, may be less effective in certain scenarios.

5.2 Implications for Policymakers

Some instances may be particularly unacceptable for creators and would require more stringent enforcement than others. e.g. *people search websites*.

Scenarios concerning *content resale in bulk* for a fee or *people search websites* are highly unlikely to be accepted and draw serious concerns about privacy and security. In our free text responses, creators expected to be informed or made aware of instances where their data/content is being allowed to be downloaded and reused by third-party organizations. These use cases should be of high priority for policymakers.

Ethical frameworks developed for research use of social media data may be relevant for public policymaking as well. Among the research community, scholars have expressed various frameworks and approaches to ensure ethical collection of social media data. Drawing from public discourses and scholarly discussions of the T3 dataset, Zimmer outlined a series of considerations for researchers to take into account when collecting public data, e.g. recognizing the changing nature of privacy [35]. More recently, Ajmani et al. further examined how sensitive social media data was discussed in research on mental health and advocated for formal disclosures of research data practices [1]. These recommendations may be applicable to public policies.

5.3 Implications for Web Crawlers, Data Donation Stations, and Data Marketplaces

While our survey did not specifically ask for creators' perspectives on these data collection approaches, our findings do highlight respondents' strong desire for compensation across all scenarios, even for *academic research*. This suggests that for web crawlers, data donation stations, and data marketplaces to fully comply with creators' demands, data collectors must integrate monetary rewards for creators.

Respondents also expressed a strong preference for control, but currently, web crawlers, data donation stations, and data marketplaces rarely support this need [18, 19]. This highlights the need for more sophisticated visibility control for content creators. Currently, many content platforms offer a binary approach to visibility control, i.e., private profile vs. public profile, and our respondents' varying preferences for control over the seven scenarios suggest that platforms should also consider allowing creators to express a spectrum of visibility, e.g. visible/invisible to search engine companies, visible/invisible to generative AI companies, and visible/invisible to academic researchers.

6 CONCLUSION AND FUTURE WORK

We explored how Instagram content creators found different content reuse scenarios acceptable, negotiable, and unacceptable. Participants were more likely to accept scenarios that have potential societal benefits, like *educatione* and *academic research*. Content reuse for *commercial generative AI* or *non-profit, open source generative AI* models was a negotiable scenario for most respondents, especially if compensation were provided.

6.1 Next Steps

Moving forward, we plan to examine how participants' preferred conditions might be correlated to each other and how the preferences of creators under different scenarios might differ. We also plan to further analyze the free text responses provided by participants, diving deeper into understanding other concerns around privacy, security, fair compensation, and awareness raised by content creators. We will leverage insights from these responses to interpret our quantitative findings from the structured questions on acceptance.

As compensation was preferred by many respondents to accept a scenario, future work may conduct empirical experiments to model how compensation will affect creators' incentive to sell their content.

REFERENCES

[1] Leah Hope Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. 2023. A Systematic Review of Ethics Disclosures in Predictive Mental Health Research. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1311–1323. https://doi.org/10.1145/3593013.3594082

[2] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing Markets. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 24–35. https://doi.org/10.1609/icwsm.v14i1.7276

[3] Santiago Andrés Azcoitia and Nikolaos Laoutaris. 2022. A Survey of Data Marketplaces and Their Business Models. *ACM SIGMOD Record* 51, 3 (Nov. 2022), 18–29. https://doi.org/10.1145/3572751.3572755

[4] Michael Barbaro, Clare Toeniskoetter, Rob Szypko, Mooj Zadie, Devon Taylor, Lisa Chow, Elisheba Ittoop, and Alyssa Moxley. 2023. The Writers' Revolt Against A.I. Companies. *The New York Times* (July 2023). https://www.nytimes.com/2023/07/18/podcasts/the-daily/ai-scraping.html

[5] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30, 1-7 (April 1998), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

[6] Raul Castro Fernandez. 2023. Data-Sharing Markets: Model, Protocol, and Algorithms to Incentivize the Formation of Data-Sharing Consortia. *Proceedings of the ACM on Management of Data* 1, 2 (June 2023), 1–25. https://doi.org/10.1145/3589317

[7] Daniel Diethei, Jasmin Niess, Carolin Stellmacher, Evropi Stefanidi, and Johannes Schöning. 2021. Sharing Heartbeats: Motivations of Citizen Scientists in Times of Crises. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445665

[8] Ekwa Duala-Ekoko and Martin P. Robillard. 2012. Asking and answering questions about unfamiliar APIs: an exploratory study. In *Proceedings of the 34th International Conference on Software Engineering (ICSE '12)*. IEEE Press, Zurich, Switzerland, 266–276.

[9] Masood Farivar. 2023. AI Firms Under Fire for Allegedly Infringing on Copyrights. https://www.voanews.com/a/ai-firms-under-fire-for-allegedly-infringing-on-copyrights/7238575.html

[10] Casey Fiesler. 2019. Ethical Considerations for Research Involving (Speculative) Public Data. *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (Dec. 2019), 249:1–249:13. https://doi.org/10.1145/3370271

[11] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4, 1 (Jan. 2018), 2056305118763366. https://doi.org/10.1177/2056305118763366 Publisher: SAGE Publications Ltd.

[12] Arpita Ghosh and Preston McAfee. 2011. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*. ACM, Hyderabad India, 137–146. https://doi.org/10.1145/1963405.1963428

[13] Sheena Goodyear. 2023. These authors say Open AI stole their books to train ChatGPT. Now they're suing. *CBC Radio* (Sept. 2023). https://www.cbc.ca/radio/asithappens/authors-guild-chatgpt-lawsuit-1.6974154

[14] Flora Graham. 2023. Daily briefing: What the end of Twitter's free API means for research. *Nature* (Feb. 2023). https://doi.org/10.1038/d41586-023-00480-9 Bandiera_abtest: a Cg_type: Nature Briefing Publisher: Nature Publishing Group.

[15] Libby Hemphill, Angela Schöpke-Gonzalez, and Anmol Panda. 2022. Comparative sensitivity of social media data and their acceptable use in research. *Scientific Data* 9, 1 (Oct. 2022), 643. https://doi.org/10.1038/s41597-022-01773-w Publisher: Nature Publishing Group.

[16] Shamika Klassen and Casey Fiesler. 2022. "This Isn't Your Data, Friend": Black Twitter as a Case Study on Research Ethics for Public Data. *Social Media + Society* 8, 4 (Oct. 2022), 20563051221144317. https://doi.org/10.1177/20563051221144317 Publisher: SAGE Publications Ltd.

[17] Hanlin Li. 2023. Data Scraping Makes AI Systems Possible, but at Whose Expense? | TechPolicy.Press. https://www.techpolicy.press/data-scraping-makes-ai-systems-possible-but-at-whose-expense/

[18] Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. 2023. The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1151–1161. https://doi.org/10.1145/3593013.3594070

[19] Hanlin Li, Nicholas Vincent, Yacine Jernite, Nick Merrill, Jesse Josua Benjamin, and Alek Tarkowski. 2023. Can Licensing Mitigate the Negative Implications of Commercial Web Scraping?. In *Computer Supported Cooperative Work and Social Computing*. ACM, Minneapolis MN USA, 553–555. https://doi.org/10.1145/3584931.3611276

[20] Lingjuan Lyu. 2023. A Pathway Towards Responsible AI Generated Content. (2023).

[21] Andrés Monroy-Hernández, Benjamin Mako Hill, Jazmin Gonzalez-Rivero, and Danah Boyd. 2011. Computers Can't Give Credit: How Automatic Attribution Falls Short in an Online Remixing Community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3421–3430. https://doi.org/10.1145/1978942.1979452 arXiv:1507.01285 [cs].

[22] Chris Parnin, Christoph Treude, Lars Grammel, and Margaret-Anne Storey. 2012. Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow. *Georgia Institute of Technology, Tech. Rep* 11 (2012). http://chrisparnin.me/pdf/crowddoc.pdf

[23] Katie Paul, Anna Tong, and Anna Tong. 2024. Inside Big Tech's underground race to buy AI training data. *Reuters* (April 2024). https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/

[24] Jian Pei. 2020. Data Pricing – From Economics to Data Science. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3553–3554. https://doi.org/10.1145/3394486.3406473

[25] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. https://doi.org/10.48550/arXiv.2112.11446 arXiv:2112.11446 [cs].

[26] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–9. https://doi.org/10.1145/3491101.3503569

[27] Martin P. Robillard. 2009. What Makes APIs Hard to Learn? Answers from Developers. *IEEE Software* 26, 6 (Nov. 2009), 27–34. https://doi.org/10.1109/MS.2009.193 Conference Name: IEEE Software.

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 25278–25294. https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[29] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. https://doi.org/10.48550/arXiv.2402.00159 arXiv:2402.00159 [cs].

[30] Morgan Sung. 2023. Fearing AI, fan fiction writers lock their accounts. https://techcrunch.com/2023/10/11/ao3-ai-fears-lock-account-kinktober-fanfiction/

[31] Benj Edwards Technica, Ars. 2024. Stack Overflow Users Are Revolting Against an OpenAI Deal. *Wired* (May 2024). https://arstechnica.com/information-technology/2024/05/stack-overflow-users-sabotage-their-posts-after-openai-deal/ Section: tags.

[32] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, and Tom H. Luan. 2023. A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions. *IEEE Open Journal of the Computer Society* 4 (2023), 280–302. https://doi.org/10.1109/OJCS.2023.3300321

[33] Savvas Zannettou, Olivia Nemes-Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P. Gummadi, Elissa M. Redmiles, and Franziska Roesner. 2024. Analyzing User Engagement with TikTok's Short Format Video Recommendations using Data Donations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3613904.3642433

[34] Adam Zewe. 2023. Explained: Generative AI. https://news.mit.edu/2023/explained-generative-ai-1109

[35] Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology* 12, 4 (Dec. 2010), 313–325. https://doi.org/10.1007/s10676-010-9227-5