

Can licensing mitigate the negative implications of commercial web scraping?

HANLIN LI, University of California, Berkeley, USA

NICHOLAS VINCENT, University of California, Davis, USA

YACINE JERNITE, Hugging Face, USA

NICK MERRILL, University of California, Berkeley, USA

JESSE JOSUA BENJAMIN, Lancaster University, UK

ALEK TARKOWSKI, Open Future, Poland

The rise of prominent AI models such as ChatGPT and Stable Diffusion has brought the scale of commercial web scraping to the forefront attention of content creators and researchers. Billions of webpages and images are used to train these models without content creators' knowledge, sparking extensive criticism and even lawsuits against AI firms. Amidst such debates, licensing is proposed by researchers and legal experts to be a potential approach to mitigate content creators' concerns and promote more responsible data reuse. However, it remains unclear what specific licensing terms will be effective and what sociotechnical environments are necessary to facilitate the use of licensing at scale. This workshop will provide a venue for researchers, content creators, and legal experts to identify opportunities and challenges in licensing content in the context of AI training.

Additional Key Words and Phrases: datasets, scraping, licensing, Responsible AI Licensing, copyright

ACM Reference Format:

Hanlin Li, Nicholas Vincent, Yacine Jernite, Nick Merrill, Jesse Josua Benjamin, and Alek Tarkowski. 2018. Can licensing mitigate the negative implications of commercial web scraping?. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

1 INTRODUCTION

Web scraping has become a prominent way for tech companies, from OpenAI to Google, to collect data and content from other organizations and platforms. Yet at times, scraping may impose costs on individuals and societies in the form of wealth concentration, privacy invasion, and lack of consent, in addition to potential copyright infringements [8]. As more AI systems that rely on web scraping for training data are developed and deployed, content producers have expressed concerns about the extractive nature of this data collection technique and the consequences of its deployment [10].

In the midst of these potential negative implications, most proposals have focused on technical remediation such as protecting content creators' personal style [7] and supporting opt-out requests for scraped datasets[11]; scholars and content creators have proposed a social mechanism to promote more responsible data use: old-fashioned licensing [3, 9]. There are two specific ways of licensing web content. The first and more long-standing approach is to license copyrighted content with a fee as a way to protect creators' copyrights, as commonly seen in the creative industry. The second and emergent approach is to license content with use restrictions – otherwise known as behavioral use licensing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

53 or "Responsible AI Licensing" (RAIL) – to promote and enforce ethical norms around data and content [3]. The second
54 approach serves to deter certain controversial use of datasets (e.g. “cannot be used to impersonate others without
55 their consent”) and has been gaining rapid traction among developers and researchers [1]. For example, on Hugging
56 Face, over 800 datasets have been licensed under RAIL-related licenses. Other open datasets also take approaches
57 similar to behavioral use licensing. For example, the Casual Conversations Dataset was shared by Meta under the term
58 that this dataset should not be used to infer personal information, among many other restrictions [6]. Theoretically,
59 copyright licensing and behavioral use licensing complement each other; while copyright licenses protect content
60 creators’ copyrights, behavioral use licenses give content creators rights to privacy and control over their work.
61

62 However, key sociotechnical questions about how these licensing mechanisms will play out among content creators
63 and those that seek to scrape content remain unanswered. Who has the right to license what type of content and what
64 incentives are necessary for firms or their scrapers to honor licensing agreements? What license clauses are desirable
65 and effective in mitigating concerns about unauthorized content reuse? Ultimately, how can licensing mitigate various
66 negative implications of web scraping? This workshop will explore the intersection of licensing and content scraping
67 and aim to offer a pragmatic roadmap toward responsible data collection and use.
68

69 We will host a series of lightning talks followed by group discussions with workshop participants. We will welcome
70 presentations that address the following aspects of web scraping and licensing:
71

- 72 • Understanding the current landscape of web scraping and investigating how firms and developers approach the
73 legal and ethical risks of scraping and aggregating web content.
- 74 • Understanding current practices around licensing among content creators, e.g. how content creators license
75 their content and what rights they would like to preserve when making their content publicly visible.
- 76 • Identifying specific opportunities to operationalize licensing to counter the negative effects of web scraping
77 and other unauthorized data reuses, including but not limited to privacy violation, lack of compensation for
78 content creators, copyright infringements, etc.
- 79 • Examining parallels between licensing and other creator-oriented responsible AI initiatives, such as data
80 stewardship [2, 5] and refusal [4].
81

82 This exchange-oriented part will be supplemented by design tasks in groups. Participants and presenters will build
83 upon ideas and research discussed in earlier sessions to chart out ways to implement licensing in content creation and
84 software development practices. In particular, we hope to make progress in the following directions:
85

- 86 • How will web scraping practices likely adapt in response to copyright and behavioral use licensing?
87
- 88 • What do we need to do to realistically impact commercial web scraping practices through licensing?
89
- 90 • How to help content creators choose licensing terms and set licensing fees?
91
- 92 • Is there any existing infrastructure that we may learn from about facilitating the adoption of copyright and
93 behavioral use licensing?
94
- 95 • How will copyright and behavioral use licensing intersect with Creative Commons licenses?
96

97 More broadly, this workshop will aim to foster interdisciplinary conversations about data scraping and reuse, open
98 source, and responsible AI. We will reflect on what other governance frameworks, besides licensing, may be helpful in
99 fostering responsible innovation and how to navigate the tension between openness and privacy in data reuse.
100

101 2 WORKSHOP LOGISTICS

102 We plan to host a one-day virtual workshop that consists of lightning talks and group discussions.
103
104

2.1 Planned Activities

We will start with open remarks by organizers to set the workshop's agenda, followed by sessions of lightning talks. Each session of lightning talks will have time for participants to ask questions and provide feedback. From past experience, we find that lightning talks help stimulate discussion among workshop participants.

After lightning talks, participants will build upon prior discussion to prototype licensing scenarios, chains of events, artifacts, and stakeholders that may emerge in the process from scraping data to model development (e.g., of a Large Language Model) and deployment. These scenarios will be discussed to arrive at a pragmatic output for the workshop around which participants can gather and build a community or generate forms of dissemination, such as a spectrum or matrix for decision-making in the licensing process.

We will encourage collaborative note-taking using Miro. Participants will be invited to capture their takeaways, questions, and discussions on a Miro board shared with the group. At the end of the workshop, we will have a short debriefing session to discuss next steps and opportunities to continue the discussion on data scraping and licensing.

2.2 Recruitment

Our workshop will be open to all CSCW attendees but participants who are interested in presenting will need to submit an extended abstract. We will post our call for participation on social media and directly to communities that create, aggregate, and maintain datasets. We welcome participation and expertise from a wide range of domains and disciplines, including but not limited to social computing, copyright, and open source. Abstracts will be reviewed by our organizing team for relevance. We will have two to three reviewers for each submission and reviewer comments will be shared with authors. We welcome submissions in various formats, from analytical reports to essays to design fiction. Accepted abstracts will be given a lightning talk opportunity. We expect a maximum of 50 participants for this workshop.

2.3 Post-Workshop Plans

We will summarize our workshop discussion in the format of a public-facing blog post. We hope to use the blog post as a call for the broader creator, developer, and researcher communities to consider and adopt licensing as a way to mitigate the negative implications of web scraping and as a step toward responsible AI. We will solicit feedback from workshop participants on our blog post before making it public and we plan to create an online space in the form of a listerv or Google Group to continue the conversation.

2.4 Organizers

Our team consists of researchers and practitioners from a diversity of domains including research, policymaking, and advocacy. Part of our team members is associated with the RAIL initiative ¹ and we hope to use this workshop as an opportunity to connect the CSCW community with the RAIL initiative to exchange research ideas and identify opportunities to collaborate on data governance.

Hanlin Li is an assistant professor at the University of Texas at Austin. She studies the social and economic impact of user-generated data and explores approaches to collective, responsible data governance.

Nicholas Vincent is a postdoc scholar at the University of California, Davis. His work focuses on studying the dependence of modern computing technologies, including the broad set of systems called "AI", on human-generated data, with the goal of mitigating negative impacts of these technologies.

¹<https://www.licenses.ai/>

157 Yacine Jernite leads the ML and Society team at Hugging Face. He works on ML systems governance at the intersection
158 of regulatory and technical tools, with a focus on NLP models and data curation, documentation, and governance.

159 Nick Merrill is a research fellow at the UC Berkeley Center for Long-Term Cybersecurity. His work aims to shift the
160 way people understand, identify, and implement safeguards against harms and expand the kinds of decision-makers
161 able to do so.

162 Jesse Josua Benjamin is a Post Doctoral Research Associate whose research focuses on combining Philosophy of
163 Technology and Design Research to investigate emergent AI challenges and Human-Computer Interaction.

164 Alek Tarkowski is the Director of Strategy at Open Future. He has over 15 years of experience with public interest
165 advocacy, movement building, and research into the intersection of society, culture, and digital technologies.

168 REFERENCES

- 169 [1] [n. d.]. The Growth of responsible AI licensing. <https://openfuture.eu/publication/the-growth-of-responsible-ai-licensing>
- 170 [2] Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay
171 Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE Principles
172 for Indigenous Data Governance. 19, 1 (Nov. 2020), 43. <https://doi.org/10.5334/dsj-2020-043> Number: 1 Publisher: Ubiquity Press.
- 173 [3] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022.
174 Behavioral Use Licensing for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for
175 Computing Machinery, New York, NY, USA, 778–788. <https://doi.org/10.1145/3531146.3533143>
- 176 [4] Patricia Garcia, Tonia Sutherland, Niloufar Salehi, Marika Cifor, and Anubha Singh. 2022. No! Re-imagining Data Practices Through the Lens of
177 Critical Refusal. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 315:1–315:20. <https://doi.org/10.1145/3557997>
- 178 [5] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant
179 Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson,
180 Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM*
181 *Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2206–2222.
182 <https://doi.org/10.1145/3531146.3534637>
- 183 [6] Meta. [n. d.]. Casual Conversations Dataset. <https://ai.facebook.com/datasets/casual-conversations-downloads>
- 184 [7] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. GLAZE: Protecting Artists from Style Mimicry by
185 Text-to-Image Models. <https://doi.org/10.48550/arXiv.2302.04222> arXiv:2302.04222 [cs].
- 186 [8] James Vincent. 2022. The lawsuit that could rewrite the rules of AI copyright. <https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data>
- 187 [9] James Vincent. 2022. Shutterstock will start selling AI-generated stock imagery with help from OpenAI. <https://www.theverge.com/2022/10/25/23422359/shutterstock-ai-generated-art-openai-dall-e-partnership-contributors-fund-reimbursement>
- 188 [10] James Vincent. 2023. AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit. <https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart>
- 189 [11] Nick Vincent. [n. d.]. ChatGPT Stole Your Work. So What Are You Going to Do? *Wired* ([n. d.]). <https://www.wired.com/story/chatgpt-generative-artificial-intelligence-regulation/> Section: tags.
- 190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208