CSCW workshop，分析内容爬虫的影响和模型许可的相关问题

Can Licensing Mitigate the Negative Implications of Commercial Web Scraping?

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

Nicholas Vincent
nmvincent@ucdavis.edu
University of California, Davis
Davis, CA, USA

Yacine Jernite
yacine@huggingface.co
Hugging Face
Brooklyn, NY, USA

Nick Merrill
ffff@berkeley.edu
University of California, Berkeley
Berkeley, CA, USA

Jesse Josua Benjamin
j.j.benjamin@lancaster.ac.uk
Lancaster University
Lancaster, UK

Alek Tarkowski
alek@openfuture.eu
Open Future
Warsaw, Poland

ABSTRACT

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 to mitigate content creators' concerns 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 answer these questions.

KEYWORDS

datasets, scraping, licensing, Responsible AI Licensing, copyright

ACM Reference Format:
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 (CSCW '23 Companion), October 14–18, 2023, Minneapolis, MN, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3584931.3611276

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 distinct ways of licensing 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 or "Responsible AI Licensing" (RAIL) – to promote and enforce ethical norms around data and content [3]. The second approach serves to deter certain controversial use of datasets (e.g. "cannot be used to impersonate others without their consent") and has been gaining rapid traction among developers and researchers [1]. For example, on the Hugging Face Hub, as of spring 2023, over 800 datasets have been licensed under RAIL-related licenses. Other open datasets also take approaches similar to behavioral use licensing. For example, the Casual Conversations Dataset was shared by Meta under the term that this dataset should not be used to infer personal information, among many other restrictions [6]. Theoretically, copyright licensing and behavioral use licensing complement each other; while copyright licenses protect content creators' copyrights, behavioral use licenses serve to preserve data creators' rights to control their work.

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

CSCW '23 Companion, October 14–18, 2023, Minneapolis, MN, USA

ACM ISBN 979-8-4007-0129-0/23/10.
https://doi.org/10.1145/3584931.3611276

CCF-A!

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

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

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

- What do we need to do to realistically impact commercial web scraping practices through copyright and RAIL licensing?
- How to help content creators set copyright licensing fees and RAIL licensing terms?
- Is there any existing infrastructure that we may learn from to facilitate the adoption of copyright and behavioral use licensing?
- How will web scraping practices likely adapt in response to copyright and behavioral use licensing?
- How will copyright and behavioral use licensing intersect with Creative Commons licenses?

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

2 WORKSHOP LOGISTICS

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

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 listserv 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 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.

Yacine Jernite leads the ML and Society team at Hugging Face. He works on ML systems governance at the intersection of regulatory

[1] https://www.licenses.ai/

and technical tools, with a focus on NLP models and data curation, documentation, and governance.

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

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

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

REFERENCES

[1] [n. d.]. The Growth of responsible AI licensing. https://openfuture.eu/publication/the-growth-of-responsible-ai-licensing

[2] Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE Principles for Indigenous Data Governance. 19, 1 (Nov. 2020), 43. https://doi.org/10.5334/dsj-2020-043 Number: 1 Publisher: Ubiquity Press.

[3] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. Behavioral Use Licensing for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 778–788. https://doi.org/10.1145/3531146.3533143

[4] Patricia Garcia, Tonia Sutherland, Niloufar Salehi, Marika Cifor, and Anubha Singh. 2022. No! Re-imagining Data Practices Through the Lens of Critical Refusal. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 315:1–315:20. https://doi.org/10.1145/3557997

[5] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2206–2222. https://doi.org/10.1145/3531146.3534637

[6] Meta. [n. d.]. Casual Conversations Dataset. https://ai.facebook.com/datasets/casual-conversations-downloads

[7] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models. https://doi.org/10.48550/arXiv.2302.04222 arXiv:2302.04222 [cs].

[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

[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

[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

[11] Nick Vincent and Hanlin Li. 2023. ChatGPT Stole Your Work. So What Are You Going to Do? *Wired* (2023). https://www.wired.com/story/chatgpt-generative-artificial-intelligence-regulation/ Section: tags.