Final Project Memo

To: Dr. Cory Curl

From: Matthew Jones

Subject: Web Scraping and AI Training

Introduction

AI is a hot topic in policy right now, and web scraping is among the most pressing aspects of the AI revolution that needs to be properly addressed. Americans are growing more concerned about private and public usage of their personal data, and the number of web scraping agents to gather more data for new models continues to rise. Meanwhile, the human-made foundation of the internet is being diluted by AI-created content. Daunting though that may be, there are ways to respond to this, as seen by policy put forth by various countries.

Web Scraping for LLMs

Eloquently put by Dr. Aaronson from George Washington University in her study of data governance for LLMs, "Just as it is helpful to understand what chemicals go into our food or water, it is helpful to understand what data goes into our LLMs." This is the idea of data provenance, or the knowledge of the origin of underlying data in an AI model, as well as modifications to data sets, and the validity of the data. That said, AI companies are failing to properly account and publish this information, which has led to lawsuits of OpenAI and Google. Currently in the United States, courts have upheld the right to scrape the internet if the data isn't harmful to society, a firm, or an individual. Without a proper system for ensuring data provenance from AI firms, though, none of this is verifiable. Furthermore, these web scrapers are reaching for everything they can on the internet, from users' web searches to social media posts, and most people are unaware that this is occurring. According to a Pew Research Study, more than half of Americans (56%) say they always, almost always, or often click "agree" without reading privacy policies. What these users don't realize, is that inside of these user agreements, companies have started saying they may use their collected information for AI training purposes.

Importance

At the surface level, 81% of Americans say the information collected about them will be used in ways they aren't comfortable with, which exemplifies an erosion of public trust of private companies. At a deeper level, though, web scraping can lead to detrimental effects on the internet. Sam Altman, CEO of OpenAI, recently posted on X that while he has never previously taken the 'dead internet' theory seriously, he has noticed "there are really a lot of LLM-run twitter accounts now." While he was mocked in the comment replies to his post, his tweet raises a real red-flag for the stability of the internet. The dead internet theory is the idea the internet could only be filled with AI agents and AI-created content, and all

traffic will be through AI, rather than human, usage. Although this theory is likely far off, increased web scraping and reproduction of data used through web scraping has the chance to create a cyclical exchange of information where AI models are responding based off purely AI created information, rather than human founded truth. That said, web scraping wouldn't be the topic that it is if it wasn't so vital to the engineering of modern genius level LLMs. AI companies will pressure this to be an accessible and efficient means to creating swaths of large data sets.

Stakeholders

American people:

> As expressed by the Pew Research survey on the public view of data privacy, 78% of democrats and 68% of republicans support increased regulation of personal information. The disintegration of truth due to a 'dead internet' also negatively impacts everyday lives. A Pew Research article from last year found that one in five Americans get their news from social media platforms and influencers. Combined with the increase in AI generated content, disinformation can run rampant for Americans. Furthermore, as AI models are trained on untruthful media, the content that they generate can lead to more disinformation.

Tech companies:

The mission statement of OpenAI is to achieve artificial generative intelligence (AGI)[1]. Put simply, the only way to achieve this is through increased data collection. There are of course other technological limitations, but at our rate of advancement it is likely that these will be tackled as well.

Policymakers:

> If the sentiments that we are in a new Digital Cold War with China are true, then policymakers may do whatever they can to push for innovation and advancement. Policymakers may also not want to inhibit the American virtue of innovation and expertise.

Current Approaches

While President Biden wrote legislation promoting innovation alongside data provenance and verifiable AI[2], President Trump has since rescinded this legislation and pursued a different course of action. U.S. Courts have, in some cases, upheld scraping of publicly accessible data[3]. The longer the U.S. waits to formalize a response to web scraping, though, there is more data that is absorbed without consent, and more lawsuits will materialize with increased ambiguity about the extent of AI companies' technological freedoms. Many countries are already putting out proposals, some of which are listed here.

---

[1] AGI is achieved when one AI model can match a human's productivity and ability in any task
[2] 2023 AI Executive Order
[3] HiQ v. LinkedIn

1. The European Union

   The EU already has in place the General Data Protection Regulation (GDPR) which allows for data protection laws to limit the use of web scraping. Websites can also opt-out of being scrapable, protecting their data and intellectual property. Furthermore, while companies web scrape, they are required to adhere to principles of transparency, data minimization, storage limitation, accuracy, security, confidentiality, integrity, and accountability.

2. China

   Generative AI technology provided to the public must use data from legitimate sources, not infringe on others' intellectual property, obtain consent for personal data, and take measures to increase the quality, accuracy, truthfulness, objectivity, and diversity of their data. That said, China is already known for limiting the free flow of information into their country with The Great Firewall, so limiting access to web scraping can also be seen as an extension to limiting data accessible to Chinese citizens.

3. India

   Calling the EU's opt-out model ineffective, India has called for a royalty system. This would institute a levy on AI companies paid to content creators for using their work to train models. This was released on December 9th, and countries have 30 days to respond.

Alternatives

1. Status quo

The first (and easiest) alternative for the U.S. is to maintain a "fair use" policy. While navigating relationships with foreign nations amongst their increasingly complex policies would require some poise, domestically web scraping could continue uninhibited.

2. Increased Regulation

The United States could increase their regulation of web scraping and promoting verifiability. This could take form in a few different ways.

   a. Really Simple Licensing (RSL) 1.0

In September of this year RSL 1.0 was announced and could prove to be a quintessential technology in web scraping policy. Expanding on the robots.txt file which created a foundation for blocking web crawlers, with RSL 1.0, web providers who support this standard can block crawlers that haven't paid for a license to scrape that website. In essence, a human can still open your website and have full access, while a bot without paid clearance is not allowed in. A possible alternative could be to continue "free use" while encouraging adoption of the RSL 1.0 standard, which would allow all companies and content creators to choose whether a web crawler has access to their site, and receive compensation.

b. Data Protection Impact Assessments (DPIAs)

A possible government funded alternative would be to require companies to publish DPIAs. While it still allows for scraping, there is increased data provenance as companies would be required to outline what information comes from where, and companies would be held to a standard of compliance for whatever other policies are put in place.

c. Regulation of models trained on synthetic data

To limit the degradation of AI models and truth on the internet, a policy could be instituted prohibiting AI models from being trained on any synthetic data. This would be accomplished through scraping reports and the requirement to eliminate any data that is found to be AI generated.

Alternative Evaluations Based on Criteria:

1. **Freedom to Innovate**
2. **Freedom to Privacy**
3. **National Security**
4. **Epistemic Security**
5. **Virtue of American Innovation**

1. Status Quo

Maintaining the fair use policy would support the continued rise of freedoms that AI companies currently have. Though as mentioned, Americans will likely continue to have concerns about what data is being collected about them and where it is going. In discussing these topics with peers, a common theme came up that people really don't know what information is being collected or how it is really used. Where the status quo scores highly in freedom to innovate through ease of access to data, it scores the exact opposite in respecting American's right to privacy and understanding of how their data is used.

From the perspective that views AGI as strategically important, fair use scraping is the fastest way for us to achieve AGI from the data standpoint, making it extremely effective at addressing the national security concern. That said, in the realm of epistemic (information) security, it has the greatest risk of leading us down a misinformation rabbit hole. Increased training on synthetic data will only lead to a greater disintegration of truth.

Lastly, while this policy would support American innovation, it is hard to say whether it supports the *virtue* of American innovation. While achieving AGI first is exactly what anyone would expect of the U.S., achieving it without guardrails risks undermining human general intelligence and personal security, which is not the reputation or story that the virtue of being American supports.

2. Increased Regulation

Since increased regulation could take the form of any other the listed alternatives or a mixture of them all, I will evaluate each criterion holistically. Beginning with the freedom to innovate, this criterion is hit hardest by increased oversight into the data that is being scraped. Efficiency would greatly decrease through the enforcement of DPIAs while they might also have to pay money to scrape certain websites through RSL 1.0. That said, for non-AI companies, RSL 1.0 can augment their freedom to innovate through a new means of making income. An art website for example could make money off OpenAI requesting to train a new image model on their human-made art. Restricting synthetic training would inhibit AI companies through efficiency and availability of resources as more of the internet is AI generated.

Likewise, through either RSL 1.0 or DPIAs, freedom to privacy would be hugely improved through this alternative. RSL 1.0 expands upon individual websites and companies' sovereignty over who can reap their data and DPIAs increase data provenance.

In terms of National Security, there is greater risk that through regulation we will fall behind in the global AI race. Peers have acknowledged this and espouse a happy medium of regulation with innovation for this concern. Paywalls such as RSL 1.0 might cause AI companies to only scrape off certain sources leading to less growth. For DPIAs, data provenance itself wouldn't slow America down, merely the possible inefficiencies in achieving increased provenance. More importantly, reduced synthetic training could lead to a poorer National performance due to models trained off ineffective AI generated data. Or the lack of data could be an inhibiting factor in training larger models.

Increased regulation effectively targets epistemic security concerns. Through DPIAs limiting garbage training and the possibility of reduced synthetic training, this could curtail low quality synthetic content training and limit fake news generation, garbage generation, and maintain human-founded truth on the internet.

In terms of virtue, this is difficult to analyze. Obviously restricting web scraping in any way is in some form anti-innovation. That said, maintaining American excellence in our technology requires a standard for how business is conducted.

Policy Recommendation

As a peer acknowledged, all business in the modern landscape is regulated in some way. They noted it as simply "the price to doing business." Could the price to doing business in AI model creation be paying for data and writing training reports? That doesn't seem unrealistic to me. And while we encourage innovation and technical excellence, through the Gilded Age and beyond the United States discovered a need for independent bodies like the FCC to maintain a standard of excellence in our work. Alongside these virtuous considerations and the increased support of the freedom to privacy, epistemic security concerns, and rights to non-AI companies, I support increased regulation for AI companies when web scraping.

Aaronson, Susan, Data Dysphoria: The Governance Challenge Posed by Large Learning Models (August 28, 2023). Available at SSRN: https://ssrn.com/abstract=4554580 or http://dx.doi.org/10.2139/ssrn.4554580

"AI Ussues." *OECD.AI*, oecd.ai/en/generative-ai-issues. Accessed 16 Dec. 2025.

Bergmann, Dave, and Cole Stryker. "What Is Artificial General Intelligence (AGI)?" *IBM*, 17 Nov. 2025, www.ibm.com/think/topics/artificial-general-intelligence.

Chaturvedi, Arpan, and Munsif Vengattil. "Indian AI Royalty Proposal Targets Data Practices of OpenAI, Google | Reuters." *Indian AI Royalty Proposal Targets Data Practices of OpenAI, Google*, Reuters, www.reuters.com/sustainability/boards-policy-regulation/indian-ai-royalty-proposal-targets-data-practices-openai-google-2025-12-09/. Accessed 16 Dec. 2025.

"Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence | The White House." *National Archives and Records Administration*, National Archives and Records Administration, 30 Oct. 2023, bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Faverio, Michelle. "Key Findings about Americans and Data Privacy." *Pew Research Center*, Pew Research Center, 18 Oct. 2023, www.pewresearch.org/short-reads/2023/10/18/key-findings-about-americans-and-data-privacy/.

For Publication United States Court of Appeals for the Ninth Circuit, cdn.ca9.uscourts.gov/datastore/opinions/2022/04/18/17-16783.pdf. Accessed 16 Dec. 2025.

Hadero, Haleluya. "About 20% of Americans Get Their News from Social Media Influencers, Report Finds." *PBS*, Public Broadcasting Service, 18 Nov. 2024, www.pbs.org/newshour/nation/about-20-of-americans-get-their-news-from-social-media-influencers-report-finds.

Liber, Adam. "The State of Web Scraping in the EU." *IAPP*, iapp.org/news/a/the-state-of-web-scraping-in-the-eu. Accessed 13 Sept. 2025.

Roth, Emma. "A Pay-to-Scrape AI Licensing Standard Is Now Official." *The Verge*, The Verge, 10 Dec. 2025, www.theverge.com/news/841222/rsl-licensing-ai-spec-launch.

"Sam Altman Concerned That the Whole Internet Now Feels Fake as AI Takes Over." *Yahoo! News*, Yahoo!, www.yahoo.com/news/articles/sam-altman-concerned-whole-internet-130025992.html. Accessed 13 Sept. 2025.

Stoller, Mitch. "Did Wikipedia Die and What Does It Mean?" *Medium*, Literate AI, 10 Apr. 2025, medium.com/literateai/did-wikipedia-die-and-what-does-it-mean-866c3f9391c4.

Taneja, Hemant, and Fareed Zakaria. "Ai and the New Digital Cold War." *Harvard Business Review*, 6 Sept. 2023, hbr.org/2023/09/ai-and-the-new-digital-cold-war.

"The 'dead Internet Theory' Makes Eerie Claims about an AI-Run Web. the Truth Is More Sinister." *UNSW Sites*, www.unsw.edu.au/newsroom/news/2024/05/-the-dead-internet-theory-makes-eerie-claims-about-an-ai-run-web-the-truth-is-more-sinister. Accessed 16 Dec. 2025.