The Great Scrape: The Clash Between Scraping and Privacy

Daniel J. Solove* and Woodrow Hartzog**

Artificial intelligence (AI) systems depend on massive quantities of data, often gathered by "scraping"—the automated extraction of large amounts of data from the internet. A great deal of scraped data contains people's personal information. This personal data provides the grist for AI tools such as facial recognition, deep fakes, and generative AI. Although scraping enables web searching, archiving of records, and meaningful scientific research, scraping for AI can also be objectionable and even harmful to individuals and society.

Organizations are scraping at an escalating pace and scale, even though many privacy laws are seemingly incongruous with the practice. In this Article, we contend that scraping must undergo a serious reckoning with privacy law. Scraping violates nearly all of the key principles of privacy laws, including fairness, individual rights and control, transparency, consent, purpose specification and secondary use restrictions, data minimization, onward transfer, and data security. Scraping ignores the data protection laws built around these requirements.

Scraping has evaded a reckoning with privacy law largely because scrapers act as if all publicly available data were free for the taking. But the public availability of scraped data shouldn't give scrapers a free pass. Privacy law regularly protects publicly available data, and privacy principles are implicated even when personal data is accessible to others.

DOI: https://doi.org/10.15779/Z38H98ZF9D

* Eugene L. and Barbara A. Bernard Professor of Intellectual Property and Technology Law, George Washington University Law School. A big thank you to our research assistants Allison Chesky, Michael Lavine, Kaitlyn Milinic, Vaishali Nambiar, Bradley Neal, Rose Patton, and Philipa Yu. The authors would also like to thank Steve Bellovin, Gianclaudio Malgieri, Andy Sellars, Jessica Silbey, and Jason Schultz and the participants of the 2024 Privacy Law Scholars Conference for their helpful comments.

** Professor of Law, Boston University School of Law.

This Article explores the fundamental tension between scraping and privacy law. With the zealous pursuit and astronomical growth of AI, we are in the midst of what we call the "great scrape." There must now be a great reconciliation.

Introduction 1523
I. The Great Scrape.....1526
A. Understanding Scraping 1526
1. The Rise of Scraping Personal Data.....1527
2. The Ethical Twilight of Scraping.....1531
B. The Scraping Wars 1532
1. The Legal Front.....1533
a. Trespass and the Computer Fraud and Abuse Act 1533
b. Business and Property Interests.....1536
c. Privacy Issues.....1538
2. The Technological Front.....1541
C. The Emerging Scraping Market 1543
D. Relevant Regulatory Intervention 1543
1. EU Data Protection Law 1544
2. U.S. Privacy Law 1547
II. Scraping and Privacy: A Fundamental Tension 1548
A. Scraping and Privacy Principles.....1549
1. Fairness 1550
2. Individual Rights and Control.....1550
3. Transparency.....1552
4. Consent.....1553
5. Purpose Specification and Secondary Use Restrictions.....1553
6. Data Minimization 1555
7. Onward Transfer 1556
8. Data Security.....1557
B. Scraping and Publicly Available Information.....1558
1. Publicly Available Information: An Incoherent Concept 1559
2. Expectations of Privacy in Publicly Available Information.....1561
3. Privacy Law and Publicly Available Information.....1563
C. The Need for a Coherent Theory of Scraping and Privacy.....1566
III. Reconciling Scraping and Privacy 1567
A. A Theory of Surveillance and Security.....1567
1. Scraping as Surveillance 1567

2. Protection from Scraping as Security.....1570
B. The Difficulty of Bringing Scraping Under the Purview of Privacy Law.....1572
1. The Undesirability of a Total Scraping Ban.....1574
2. The Consent Model.....1575
C. A Regulatory Agenda for Scraping in the Public Interest.....1577
1. Use of Data as a Privilege for Furthering Public Interests.....1580
2. Guiding Principles for Regulating Scraping.....1580
Conclusion.....1583

INTRODUCTION

Artificial intelligence (AI) systems depend on massive quantities of data, often gathered by "scraping"—the automated extraction of large amounts of data from the internet. Scraping allows actors to collect enormous amounts of personal data cheaply and quickly, without granting the data subject any notice, consent, or opportunity to object or opt out. Personal data provides the grist for AI tools such as facial recognition, deep fakes, and large language models. Scraping is a foundational practice in the modern digital sphere. It was used to build what we know as the World Wide Web, and it continues to be relied upon for essential and everyday information services. Scraping personal data enables web searching, archiving, generative AI, and scientific research. However, scraping for AI can also be objectionable or even harmful to individuals by directly and indirectly increasing their exposure to surveillance, harassment, and automated decisions.

Organizations are scraping personal data at an escalating pace and scale, even though many longstanding privacy principles and laws built around notions of data transparency and restraint are seemingly inconsistent with the practice. There has always been a fundamental conflict between scraping and privacy, but for years this tension was merely a background concern. AI has brought this tension to the forefront. AI requires scraping on a grand scale.[1] Recently, we have witnessed companies scrape an unprecedented amount of data. And more and more companies are scraping.[2]

1. *See* Charlotte A. Tschider, *AI's Legitimate Interest: Towards a Public Benefit Privacy Model*, 21 HOUS. J. HEALTH L. & POL'Y 125, 132 (2021) ("Machine learning applications use exceptionally large volumes of data, which are analyzed by a machine learning utility to determine interrelationships between these data.").

2. Jacob (Yakup) Kalvo, *Web Scraping: Unlocking Business Insights in a Data-Driven World*, FORBES (Jan. 27, 2025), https://www.forbes.com/councils/forbestechcouncil/2025/01/27/web-scraping-unlocking-business-insights-in-a-data-driven-world/ [https://perma.cc/H6RG-9Z9U] ("[T]he web scraping market is projected to reach $2.45 billion by 2036, with an annual growth rate of over 13%.").

In this Article, we contend that scraping must undergo a serious and long overdue reckoning with privacy law. Scraping of personal data violates nearly every key principle embodied in privacy law's frameworks and codes, including transparency, purpose limitation, data minimization, choice, access, deletion, portability, and protection. Scraping involves the mass, unauthorized extraction of personal data for unspecified purposes without any limitations or protections. In nearly every dimension, this practice is antithetical to privacy.

A major root of the problem is the vague and protean idea of "publicly available information." Scraping has evaded a reckoning with privacy law largely because scrapers act as if all publicly available data were free for the taking. Privacy law is currently inconsistent about protections for publicly available data. Although some laws do not protect publicly available data, other laws, such as the European Union's (EU) General Data Protection Regulation, largely do.[3] Additionally, many courts have recognized that public exposure does not extinguish one's privacy interest. Most notably, the U.S. Supreme Court has held that there is a reasonable expectation of privacy under the Fourth Amendment for geolocation data about publicly observable automobile movement[4] and that there is a privacy interest in the practical obscurity of personal data in certain publicly available records.[5]

Beyond scrapers, the organizations whose websites are scraped (the "scrapees") also must reckon with privacy law. Organizations can mitigate scraping through certain measures like monitoring for bots and limiting how often suspicious accounts can access a site. But too often, they take minimal action. Failing to protect against scraping of personal data makes most privacy protection requirements meaningless. Requiring transparency, vetting, contracts, and controls on third-party data sharing is ineffective if any unauthorized scraper can just take the data. If any third party can collect and use personal data in ways contrary to the promises organizations make in their privacy notice, then these promises are hollow. Allowing scrapers to gather the data can be a lapse in data security—it is akin to leaving the back door wide open and allowing unauthorized access.

This Article explores this fundamental tension between scraping and privacy. Our thesis is that scraping is contrary to the core principles of privacy that form the backbone of privacy law's frameworks and codes. With the zealous pursuit and astronomical growth of AI, we are in the midst of what we call the "great scrape." There must now be a great reconciliation.

Surprisingly, there has been a dearth of scholarly attention to scraping. Most scholarship about scraping focuses on how scraping fares under particular

3. *See infra* Part II.B.
4. *See* Carpenter v. United States, 585 U.S. 296 (2018).
5. U.S. Dep't of Just. v. Reps. Comm. for Freedom of the Press, 489 U.S. 749, 768–71 (1989).

laws, especially the Computer Fraud and Abuse Act (CFAA).[6] Our focus is much broader and more conceptual. What makes scraping such an important and fascinating issue is that it is at odds with the fundamental principles and approaches in existing privacy law, yet a categorical ban on scraping would be undesirable and probably untenable if we want a usable internet. Scraping makes the web searchable and is used by countless researchers and journalists. Scraping is also popular for many organizations developing and deploying AI technologies, especially generative AI.

Scraping is a problem of vast complexity, and it cannot be solved with a few standard tweaks to existing privacy laws. It requires a major rethinking of privacy that centers the public interest and limits data grab free-for-alls and opportunistic self-dealing. And it bears repeating that a world without scraping would hobble the internet, stunt the development of AI, and frustrate research and journalism.

It is impossible to have meaningful privacy protection where scraping can occur without legal restrictions or policies that support technical safeguards against scraping, especially with so much personal data publicly available online and the ability to hoover up this data so readily with automation. But bans and other restrictions on scraping can lead to many socially detrimental consequences, including depriving journalists and researchers of important tools to keep industry and government accountable. Market forces might compel some companies to restrict third-party scraping to protect what they view as their proprietary data. But this, too, would be highly undesirable, leading to an internet more akin to a series of walled gardens. A regulatory intervention must be made, but both encouraging and discouraging scraping comes with huge costs, resulting in a choice between Scylla and Charybdis. Ultimately, scraping and privacy must be reconciled, and this reconciliation will be an unpleasant compromise for both scraping and privacy.

Our argument proceeds in three parts. In Part I, we explore what scraping is and how it has become a fundamental part of the digital economy. In Part II, we demonstrate how scraping personal data conflicts with nearly all of the foundational privacy principles in privacy law and standards. We argue that the public availability of scraped data should not give scrapers a free pass. Privacy law regularly protects publicly available data, and privacy principles are implicated even when personal data is accessible to others. In Part III, we discuss how scraping should be reconciled with privacy law. We propose re-conceptualizing the scraping of personal data as surveillance and protecting against the scraping of personal data as a duty of data security. We contend that privacy law should not bar all instances of scraping. Instead, the law should require a legitimate basis for scraping, encourage scraping in the public interest,

6. *See generally* Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372 (2018).

and impose restrictions on scraping for harmful or risky uses. Although present in a narrow form in some laws, the concept of public interest has generally been underutilized in privacy law. We contend that public interest should be the law's primary focus when it comes to scraping.

I.
THE GREAT SCRAPE

For decades, people and organizations have scraped information off the World Wide Web with minimal resistance. In this Part, we discuss how scraping works, why scraping is so prevalent, defenses against scraping, and the emerging battles between the scrapers and scrapees.

We begin by discussing how various bots scour the internet for data, how the system of scraping has historically worked in an oddly polite manner, and how AI is dramatically changing the ballgame. We next discuss the emerging war between scrapers and scrapees on both legal and technological fronts. Finally, we provide an overview of various attempted or possible regulatory interventions.

A. Understanding Scraping

Broadly understood, scraping is automated online data harvesting. The general term "data scraping" refers to any time "a computer program extracts data from [an] output generated from another program."[7] More specifically, scraping is the "retrieval of content posted on the World Wide Web through the use of a program other than a web browser or an application programming interface (API)."[8] Scraping "is used to transform unstructured data on the web into structured data that can be stored and analyzed in a centralized local database or spreadsheet."[9]

Colloquially, some might use the term scraping to describe "manual" techniques like the traditional copy-and-paste.[10] But our focus here is the kind of automated scraping that occurs through the use of programs, called "web crawlers," "spiders," or "bots," that make the mass collection of information relatively cheap and easy.[11] These computer programs scour the internet

7. *What is Data Scraping?*, CLOUDFLARE, https://www.cloudflare.com/learning/bots/what-is-data-scraping/ [perma.cc/SQ9M-EAGX].

8. Sellars, *supra* note 6, at 373.

9. SCM de S Sirisuriya, *A Comparative Study on Web Scraping*, 2015 PROCS. 8TH INT'L RSCH. CONF., KOTELAWALA DEFENCE UNIV., 135, 135 (2015).

10. *Id.*

11. *See* Faojia Fariha, *Crawler vs Scraper vs Spider: A Detailed Comparison*, CORE DEVS (Nov. 5, 2023), https://coredevsltd.com/articles/crawler-vs-scraper-vs-spider/ [https://perma.cc/SH7R-37K6] ("A web crawler is a software program that systematically browses the World Wide Web to collect information about websites and their pages. It's like an automated script that fetches web pages and follows the links within those pages. . . . A web scraper is a program or script that extracts specific data from websites. Unlike crawlers, which collect information about websites, scrapers are focused on the content of the site—pulling text, images, prices, or any other specific elements. . . . A web spider is

gathering information from webpages. Scraping using these bots is an increasingly ubiquitous practice.

1. The Rise of Scraping Personal Data

Bots have long roamed the internet; they have been deployed since the early 1990s when the commercial internet began to develop.[12] One of the earliest forms of scraping that is still popular involves search engines using bots to crawl and index websites, a practice that makes the internet searchable. Different purposes for scraping emerged soon after, such as conducting market research, compiling feeds, monitoring competitor pricing and practices, and analyzing trends and activities.[13]

Any publicly accessible website can be scraped by automated tools.[14] Technically, password-protected and paywalled websites can be scraped too. But because they typically cannot be automatically crawled without access credentials, they are not popularly scraped for large-scale data collection. Some websites take affirmative steps to allow search engines like Google to access content behind a paywall with web crawlers.[15] Scrapers gather personal data from freely accessible social media profiles as well as many other types of websites such as those involving fitness, banking, and hospitality.[16] Web-scraping bots are designed to gather data from websites in an efficient and systematic manner.[17]

---

similar to a crawler but is more focused on indexing the textual content of a web page. It is often employed by search engines to scan and index the web."); *see also* eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1060 n.2 (N.D. Cal. 2000) ("Programs that recursively query other computers over the [i]nternet in order to obtain a significant amount of information are referred to in the pleadings by various names, including software robots, robots, spiders and web crawlers."); Kathleen C. Riley, *Data Scraping as a Cause of Action: Limiting Use of the CFAA and Trespass in Online Copying Cases*, 29 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 245, 247 (2018) ("Data scraping, also termed screen scraping, web scraping, or web crawling, refers to the extraction of data from websites, often performed by programs termed 'bots,' 'spiders,' or 'web crawlers.'").

12. Seyed M. Mirtaheri, Mustafa Emre Dinçtürk, Salman Hooshmand, Gregor V. Bochmann, Guy-Vincent Jourdan & Iosif Viorel Onut, *A Brief History of Web Crawlers*, 2014 PROCS. 2013 CONF. CTR. FOR ADVANCED STUD. ON COLLABORATIVE RSCH., 1, 3 (noting that web crawlers have existed since 1993, where they "mainly collected information and statistic[s] about the web . . . [and] downloaded URLs").

13. Tim Keary, *Web Scraping*, TECHOPEDIA (June 20, 2024), https://www.techopedia.com/definition/5212/web-scraping [https://perma.cc/N2C6-LGVA].

14. Mike Clark, *Scraping by the Numbers*, META (May 19, 2021), https://about.fb.com/news/2021/05/scraping-by-the-numbers/ [https://perma.cc/Q8Y3-ESLH].

15. *See, e.g.*, Madeleine White, *Ask the Experts: Paywalls, Subscription and SEO*, AUDIENCERS (Sept. 12, 2023), https://theaudiencers.com/ask-the-experts-paywalls-subscription-and-seo/ [https://perma.cc/4VZC-Q7QS].

16. *Id.*

17. *See* Fariha, *supra* note 11. Not all bots engage in web scraping; bots are used in myriad helpful and harmful ways, such as to engage in marketing, post spam comments, exploit vulnerabilities, and launch Distributed Denial of Service (DDOS) attacks. *See* Adrienne LaFrance, *The Internet Is Mostly Bots*, ATLANTIC (Jan. 31, 2017), https://www.theatlantic.com/technology/archive/2017/01/bots-bots-bots/515043/ [https://perma.cc/UU4Z-LBJ8]; *see also* Allison Parrish, *Bots: A Definition and Some*

For a long time, bots that gather information on the internet have operated in an oddly chivalrous fashion.[18] Websites use a simple text file called robots.txt to politely tell bots whether or not to crawl their site.[19] As technology journalist David Pierce puts it, "This text file has no particular legal or technical authority, and it's not even particularly complicated. It represents a handshake deal between some of the earliest pioneers of the internet to respect each other's wishes and build the internet in a way that benefitted everybody."[20] As Zachary Gold and Mark Latonero note, "[R]obots.txt can be ignored; those employing crawlers are not bound by any law contract, or technical need to obey a robots.txt file."[21] Remarkably, though, this method has worked; many bots still respect robots.txt instructions.[22]

Over time, scraping has become easier and more prevalent.[23] The online world began to be populated more and more by bots. By 2014, more than a quarter of internet traffic was estimated to consist of bots.[24] By 2017, some commentators estimated that bots accounted for more than 40 percent of internet traffic.[25] That is, as an article in *The Atlantic* proclaimed, "Most website visitors aren't humans."[26]

The rise in AI in the past few years has increased the motivation to scrape, as AI demands vast amounts of training data.[27] Large language models (LLMs)

Historical Threads, MEDIUM (Feb. 24, 2016), https://medium.com/datasociety-points/bots-a-definition-and-some-historical-threads-47738c8ab1ce [https://perma.cc/MUR5-6SMV].

18. *See* Alistair Barr & Kali Hays, *AI Is Killing the Grand Bargain at the Heart of the Web. 'We're in a Different World.'* BUS. INSIDER (Jan. 2, 2024), https://www.businessinsider.com/ai-killing-web-grand-bargain-2023-8 [https://perma.cc/9TZM-RTHC].

19. David Pierce, *The Text File that Runs the Internet*, VERGE (Feb. 14, 2024), https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders [https://perma.cc/98ST-RJ79].

20. *Id.*

21. Zachary Gold & Mark Latonero, *Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping*, 13 WASH. J.L. TECH. & ARTS 275, 281 (2018).

22. For example, the major search engines still respect robots.txt. *See* Steven van Vessum, *Robots.txt for SEO: The Ultimate Guide*, CONDUCTOR (Jan. 15, 2025), https://www.conductor.com/academy/robotstxt/ [https://perma.cc/U3LK-LRSS]. *But see* Tom McKay, *Nicholas Vincent Explains Why Robots.txt Is No Longer Enough to Protect Against Web Scraping*, IT BREW (Mar. 21, 2024), https://www.itbrew.com/stories/2024/03/21/prof-nicholas-vincent-explains-why-robots-txt-is-no-longer-enough-to-protect-against-web-scraping [https://perma.cc/K2KB-84CQ] (explaining how companies scraping for AI training data are ignoring norms respecting robots.txt).

23. *See* Isaiah Poritz, *OpenAI's Legal Woes Driven by Unclear Mesh of Web-Scraping Laws*, BLOOMBERG L. (July 5, 2023), https://news.bloomberglaw.com/ip-law/openais-legal-woes-driven-by-unclear-mesh-of-web-scraping-laws [https://perma.cc/G4U3-TU83].

24. *See* Philip H. Liu & Mark Edward Davis, *Web Scraping—Limits on Free Samples*, 8 LANDSLIDE 54, 54 (2015).

25. *See, e.g.*, *Distribution of Bot and Human Web Traffic Worldwide from 2013 to 2023*, STATISTA, https://www.statista.com/statistics/1264226/human-and-bot-web-traffic-share/ [https://perma.cc/C4QA-F8S4].

26. LaFrance, *supra* note 17.

27. Lee Tiedrich, *The AI Data Scraping Challenge: How Can We Proceed Responsibly?*, OECD.AI POL'Y OBSERVATORY (Mar. 5, 2024), https://oecd.ai/en/wonk/data-scraping-responsibly [https://perma.cc/C6QW-2V27].

and generative AI must be fed unprecedented quantities of data to properly train their models. Most AI companies either scrape data themselves or purchase scraped data to compete.[28] Sometimes, data is collected using an application programming interface (API) designed for the consensual extraction and sharing of data.[29] Scraping today is like the gold rush—a frenzied data grab on the grandest of scales. The market for web-scraping software exceeded half a billion dollars in 2023 and is expected to expand by 13 percent in the next twelve years.[30]

One of the most notorious instances of scraping for AI was carried out by Clearview AI, a startup company that scraped more than three billion images to develop a facial recognition system.[31] Clearview AI's facial recognition tool quickly became widely used by law enforcement organizations around the world.[32] The company operated in the shadows until *The New York Times* journalist Kashmir Hill broke the story on its secretive activities, prompting an enormous backlash, many lawsuits, and regulatory responses around the world.[33]

Another instance of a colossal scraping campaign was carried out by OpenAI, the creator of the popular generative AI tools ChatGPT and Dall-E.[34] Perhaps more than any other company, OpenAI's generative AI created public attention and fueled the current hype in AI.[35] OpenAI plundered the internet in massive scrapes to gather enormous quantities of training data.[36] The company

28. *See generally* FED. TRADE COMM'N, GENERATIVE ARTIFICIAL INTELLIGENCE AND THE CREATIVE ECONOMY STAFF REPORT: PERSPECTIVES AND TAKEAWAYS 9–10 (2023), https://www.ftc.gov/system/files/ftc_gov/pdf/12-15-2023AICEStaffReport.pdf [https://perma.cc/4YL4-4KX2].

29. *See What Is an API (Application Programming Interface)?*, MULESOFT, https://www.mulesoft.com/resources/api/what-is-an-api [https://perma.cc/GVB9-JJY4]; Michael Goodwin, *What Is an API (Application Programming Interface)?*, IBM (Apr. 9, 2024), https://www.ibm.com/topics/api [https://perma.cc/K3YW-FCVW].

30. *See* Abhishek Verma, *Web Scraping Software Market*, RSCH. NESTER (Jan. 3, 2025), https://www.researchnester.com/reports/web-scraping-software-market/5041 [https://perma.cc/CH28-G78E].

31. *OAIC and UK's ICO Open Joint Investigation into Clearview AI Inc.*, AUSTRALIAN GOV'T: OFF. OF THE AUSTRALIAN INFO. COMM'R (July 9, 2020), https://www.oaic.gov.au/newsroom/oaic-and-uks-ico-open-joint-investigation-into-clearview-ai-inc [https://perma.cc/8FW2-994F].

32. KASHMIR HILL, YOUR FACE BELONGS TO US: A SECRETIVE STARTUP'S QUEST TO END PRIVACY AS WE KNOW IT 136–37 (2023) (discussing Clearview's rise to success and expansion worldwide).

33. *Id.* at 255.

34. *See* Tonya Riley, *OpenAI Lawsuit Reignites Privacy Debate over Data Scraping*, CYBERSCOOP (June 30, 2023), https://cyberscoop.com/openai-lawsuit-privacy-data-scraping/ [https://perma.cc/3GGS-A9F5].

35. *Getting Beyond the Hype: A Guide to AI's Potential*, STAN. ONLINE, https://online.stanford.edu/getting-beyond-hype-guide-ais-potential [https://perma.cc/R2PR-DMCL].

36. Kieran McCarthy, *Web Scraping for Me, But Not for Thee (Guest Blog Post)*, TECH. & MKTG. L. BLOG (Aug. 24, 2023), https://blog.ericgoldman.org/ archives/2023/08/web-scraping-for-me-but-not-for-thee-guest-blog-post.html [perma.cc/2BJS-S4KS] (noting that "OpenAI has almost certainly already scraped the entire non-authwalled-[i]nternet" and used the data to train ChatGPT).

has been accused of scraping data from "hundreds of millions of internet users."[37]

New AI companies are popping up at a staggering rate, each with a voracious appetite for data. Scraping is easy, and for those that do not want to do the scraping themselves, there are many scrapers for hire. A "bots-as-a-service" industry scrapes data and sells it to eager AI companies.[38] For example, Imperva, a cybersecurity software company, describes the "bots-as-a-service" moniker as an attempt "to rebrand bad bots in an effort to legitimize their activity as a valid business practice."[39]

Some sites are bigger targets for scraping than others. Large platforms such as Facebook, X (formerly Twitter), Reddit, LinkedIn, and others present a gold mine to scrapers. For example, X has seen "extreme levels of data scraping" and has taken measures to limit scraping to logged-in users.[40] Elon Musk stated that "[s]everal hundred organizations (maybe more) were scraping Twitter data extremely aggressively."[41]

As AI continues its meteoric rise, scraping will invariably increase, as the quantity of data needed to feed so many hungry AI beasts is immense. The internet today is increasingly becoming a digital digestive system, where a biome of billions of bots mercilessly feeds on data to satisfy AI's insatiable hunger.

Although scrapers gather all sorts of data, our focus is on personal data. A lot of data online is personal data. People post an endless stream of data about their lives on social media sites; they write about their entire existence, from the mundane to the deeply intimate. The internet teems with photos and videos of people engaged in nearly every activity imaginable. News articles contain details about people, and organizational websites have biographies of their employees. People's thoughts and conversations are online in comment threads to articles or on social media.

It is hard to estimate just how much personal data is hoovered up in various scrapes, but there are allegations that it is occurring on a widespread scale with little restraint by the scrapers.[42] Data of "medical record photographs of thousands of . . . people" has been scraped.[43] In one lawsuit, companies

37. Poritz, *supra* note 23.

38. IMPERVA, 2023 BAD BOT REPORT 30 (2023), https://www.imperva.com/resources/reports/2023-Imperva-Bad-Bot-Report.pdf [https://perma.cc/QL26-X2LL].

39. *Id.*

40. Andrew Hutchinson, *Twitter Implements Usage Limits for All to Combat Data Scrapers*, SOCIALMEDIATODAY (July 1, 2023), https://www.socialmediatoday.com/news/twitter-implements-usage-limits-combat-data-scrapers/684831/ [perma.cc/7KBY-NPFF].

41. *Id.*

42. *See, e.g.*, Joe Tidy, *How Your Personal Data Is Being Scraped from Social Media*, BBC (July 15, 2021), https://www.bbc.com/news/business-57841239 [https://perma.cc/LFF8-QV94].

43. Lauren Leffer, *Your Personal Information Is Probably Being Used to Train Generative AI Models*, SCI. AM. (Oct. 19, 2023), https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/ [https://perma.cc/K3NU-E6FK].

integrating ChatGPT allege that they have been scraped, including "image and location data from Snapchat, financial information from Stripe, and conversations on Slack and Microsoft Teams."[44] Companies like Clearview AI and PimEyes have scraped billions of photos to power facial recognition tools.[45]

Scraping personal data implicates the privacy of the individuals whose data is scraped, and these individuals are not the scrapers or scrapees. As discussed later, the interests of these individuals are not being sufficiently represented in the battles over scraping.[46]

2. The Ethical Twilight of Scraping

Scraping has long occurred on a shifting technological plane, in an uncertain legal landscape, and with a murky ethical grounding. It has been both loved and reviled, tolerated as a necessary evil and attacked as stealing. As Andrew Sellars notes, scrapers have been "likened to an invading army of robots . . . a person walking into a bank with both a safety deposit key and a shotgun — or, more innocently . . . an interviewer using an audio recording instead of taking notes."[47] What was once seen as dubious scraping behavior is now becoming commonplace. For example, a description of "bad" scraping in the Imperva *Bad Bots Report* seemingly applies to most scraping: "Bad bots are software applications that run automated tasks with malicious intent. They scrape data from sites without permission to reuse it and gain a competitive edge (e.g., pricing, inventory levels, proprietary content)."[48] As journalist Adrienne LaFrance writes, bad bots "include unauthorized-data-scrapers, spambots, and scavengers seeking security vulnerabilities to exploit."[49] The key question is what an "unauthorized" data scraper is, as most data scrapers do not ask for permission; they scrape unless they are told not to scrape or are blocked from scraping.

Scraping has an ambiguous ethical valence because it is not all bad, but it is also not all good. Perhaps the key to properly understanding the moral implications of scraping is to focus on the *affordances* of scraping. As pioneered by James Gibson, affordances are the perceived and actual properties of something that determine how it might be used.[50] Scraping dramatically lowers

---

44. Poritz, *supra* note 23.

45. Katherine Tangalakis-Lippert, *Clearview AI Scraped 30 Billion Images from Facebook and Other Social Media Sites and Gave Them to Cops: It Puts Everyone into a 'Perpetual Police Line-up*,*'* BUS. INSIDER (Apr. 2, 2023), https://www.businessinsider.com/clearview-scraped-30-billion-images-facebook-police-facial-recogntion-database-2023-4 [https://perma.cc/CPE6-WSPF].

46. *See infra* Part II.

47. Sellars, *supra* note 6, at 383.

48. IMPERVA, *supra* note 38, at 4.

49. LaFrance, *supra* note 17.

50. JAMES J. GIBSON, *The Theory of Affordances*, *in* THE ECOLOGICAL APPROACH TO VISUAL PERCEPTION 127 (classic ed. 2014); *see also* DON NORMAN, THE DESIGN OF EVERYDAY THINGS 11–20 (2013); Ryan Calo, *Privacy, Vulnerability, and Affordance*, 66 DEPAUL L. REV. 591, 601–03 (2017); WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW

the cost of obtaining and keeping information at scale in a way that is simply unimaginable for manual data collection. In this way, it is quite different from merely providing individual (and manual) access.[51] The stark difference between collecting information via scraping and collecting information manually sets the stage for our current conflict.

B. The Scraping Wars

Today, as we use the internet, a war is going on all around us in the background. The war is on an unprecedented scale with multiple combatants, gigantic bot armies, and a technological rat-race. We are living in the midst of what we call the "Scraping Wars." Many organizations have an incentive to scrape; but many organizations have an incentive to not be scraped.[52] Being scraped and having data extracted provides little benefit and sometimes enables competitors to achieve gains. Ironically, some of the most vigorous scrapers are also the most vigorous defenders against being scraped.[53] Meta once hired a company to scrape on its behalf, then ended up suing the company when it began to scrape Meta's data.[54]

Many sites now include statements in their terms of service that users agree not to scrape without permission.[55] For example, Microsoft recently updated its terms of use to forbid scraping of its own sites despite Microsoft's affiliate OpenAI scraping the whole internet.[56]

Many sites want to be crawled only for search engine visibility, not to have their data extracted. With search engine web crawling, there is a reciprocal benefit, as many sites and people welcome crawlers because they want their information to be findable on the internet. Scraping to extract data for other purposes lacks this reciprocal benefit; it exclusively benefits the scrapers. Consistent with their desire to not be scraped, several companies have formed an industry association called the Mitigating Unauthorized Scraping Alliance (MUSA). The association "unites industry and regulators to combat

---

TECHNOLOGIES 38 (2018); Ryan Calo, *Modeling Through*, 71 DUKE L.J. 1391, 1398 (2022); Ryan Calo, *Can Americans Resist Surveillance?*, 83 U. CHI. L. REV. 23, 25 (2016).

51. The difference in scale is a key aspect of technology policy and ethics. *See, e.g.*, Mark P. McKenna & Woodrow Hartzog, *Taking Scale Seriously in Technology Law* (forthcoming) (draft on file with authors).

52. McCarthy, *supra* note 36.

53. *See, e.g.*, Michael Gennaro, *Federal Judge Rules Against Meta in Data Scraping Case*, COURTHOUSE NEWS SERV. (Jan. 23, 2024), https://www.courthousenews.com/federal-judge-rules-against-meta-in-data-scraping-case/ [https://perma.cc/GX7N-WJAY] ("The social media giant sued Israel-based web scraper Bright Data in 2023, accusing the company of violating Facebook and Instagram's terms of service and policies by scraping data from both sites even though Meta has paid Bright Data to scrape data from other sites in the past.").

54. *Id.*

55. *See* Riley, *supra* note 11, at 257–58.

56. McCarthy, *supra* note 36.

unauthorized data scraping. [It] aim[s] to promote best practices, raise public awareness, and provide valuable insights to policymakers."[57]

The Scraping Wars are occurring on two major fronts: legal and technological. Although the scrapers and scrapees are often the major combatants in the Scraping Wars, the individuals whose data is scraped also have interests in the fight, and they can be overlooked in battles between powerful industry titans.

1. The Legal Front

On the legal front, numerous attempts have been made to combat scraping under various statutes and causes of action. The cases have involved many types of data, from intellectual property to pricing data to personal data. Litigation has been ongoing for decades, but the legality of scraping has remained inconclusive. As Andrew Sellars describes it, "[T]he legal status of scraping is characterized as something just shy of unknowable, or a matter left entirely to the whims of courts, plaintiffs, or prosecutors."[58]

Before we summarize this ligation, we note several themes. First, most of the cases are battles between companies. Even when personal data is involved, the individuals whose data is being fought over are often left out of the loop. They are rarely represented in the cases, and their interests are rarely considered; the focus is mainly on the property and business interests of the scrapers and scrapees and on contractual or other issues between the scrapers and scrapees.

Second, the litigation has generally been indecisive. Even under the same causes of action, sometimes scrapers win and sometimes scrapees win. Thus, the current status of scraping under the law remains a murky gray zone.

Third, most of the cases have involved claims related to property and contract, not privacy. Privacy has often been ignored in this litigation or given scant consideration. After decades of litigation, the privacy interests of the people whose data is often involved in the Scraping Wars remain surprisingly unresolved and unexamined.

a. Trespass and the Computer Fraud and Abuse Act

The most common battlefront for scraping litigation is under the Computer Fraud and Abuse Act (CFAA). The CFAA restricts one who "intentionally accesses a computer without authorization or exceeds authorized access and thereby obtains . . . information from any protected computer."[59] The CFAA applies regardless of the purpose of access.[60]

---

57. MITIGATING UNAUTHORIZED SCRAPING ALLIANCE (MUSA), https://antiscrapingalliance.org/ [https://perma.cc/3GWR-3KFU].

58. Sellars, *supra* note 6, at 377.

59. 18 U.S.C. § 1030(a)(2)(C).

60. Sellars, *supra* note 6, at 391.

Civil liability under the CFAA is limited by a requirement of an articulable loss caused by scraping. Courts have reached mixed conclusions about the theory of loss.[61] However, generally, the "loss" threshold of $5,000 in a one-year period is easily established because expenses to investigate scraping activity count as a loss.[62]

Over several decades, many cases about or related to scraping have been litigated under the CFAA, with shifting and inconclusive results. The challenge is that the law's key triggers—unauthorized access and exceeding authorized access—are quite tricky to define given the way the internet works.

The CFAA's prohibition against unauthorized access is usually more straightforward when a hacker breaks into a computer system by bypassing technical protections like encryption and password prompts. In these circumstances, a computer system resembles a building where someone has broken in by picking a lock or fenced-in land where someone has trespassed by climbing over a fence. But many situations online do not fit this analogy. Many online "spaces" are just data sitting out in the open. This data is meant to be accessed, at least manually, by humans. There rarely are doors or fences; instead, restrictions on access are based on norms, statements made in terms of service, technological measures to make scraping difficult, or direct demands in cease-and-desist letters. Complicating matters is the fact that sites want the data to be accessed—this is essential for users of the site—but do not want scrapers to access the same data or want bots to gather data only for some purposes but not others.[63]

Some courts adopt narrow theories of the CFAA. Other courts focus on the terms of use, technological measures to block scraping, or other indications of restricted access.[64] For example, when scraping violates websites' terms of service, companies have claimed that the scraping constitutes unauthorized access under the CFAA. Early cases cracked open the door to this theory. In *EF Cultural Travel BV v. Zefer Corp.*, the court noted that a website must explicitly state any restrictions on scraping: "If EF wants to ban scrapers, let it say so on the webpage or a link clearly marked as containing restrictions."[65] Later cases, though, concluded that the mere contravention of terms of service is not enough to establish unauthorized access. For example, in *Facebook v. Power Ventures*, Power Ventures scraped Facebook as part of its efforts to help users "keep track of a variety of social networking friends through a single program."[66] Facebook sent a cease-and-desist letter to Power Ventures and blocked Power Ventures' IP address, but Power Ventures changed its IP address to continue scraping.

61. Gold & Latonero, *supra* note 21, at 296.

62. Sellars, *supra* note 6, at 376.

63. *See, e.g.*, McCarthy, *supra* note 36.

64. *See generally* Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178 (N.D. Cal. 2013); Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058 (9th Cir. 2016); Sellars, *supra* note 6, at 380.

65. 318 F.3d 58, 63 (1st Cir. 2003).

66. 844 F.3d at 1062.

Facebook sued, alleging that Power Ventures violated the CFAA. The Ninth Circuit concluded that violating Facebook's terms of service did not constitute unauthorized access, but scraping after the cease-and-desist letter did.[67]

Andrew Sellars views the case law regarding the applicability of the CFAA to scrapers as shifting like the wind blowing from different directions. From 2000–2009, he notes that courts were quick to find that scraping was unauthorized access.[68] In the early 2010s, there was a "slight trend towards limiting the law's application."[69] By the mid-2010s, courts embraced various indications of revocation of access as making scraping fall within the CFAA's prohibited unauthorized access.[70] By the late 2010s, courts were back on the side of the scrapers.[71]

The Ninth Circuit's opinion in *hiQ Labs v. LinkedIn Corp.* represents a big CFAA victory for scrapers.[72] hiQ, a data analytics company, began scraping public LinkedIn user profiles (which included user resumes and posts) and used the data to develop a "people analytics" algorithm that it marketed to businesses. LinkedIn prohibited scraping in its user agreement, which hiQ ignored, and took many technical steps to prevent scraping, which hiQ evaded.[73] LinkedIn sent hiQ a cease-and-desist letter, claiming that hiQ was violating the CFAA and that hiQ's scraping violated LinkedIn's user agreement. hiQ sued for a preliminary injunction to declare that its scraping was legal under the CFAA and that LinkedIn should remove any technical barriers to its scraping. The Ninth Circuit held that "the CFAA is best understood as an anti-intrusion statute and not as a 'misappropriation statute . . . .'"[74] Because the LinkedIn profiles were publicly available, the court reasoned, hiQ was not breaking and entering or trying to circumvent a password-protected access gate. The court concluded: "It is likely that when a computer network generally permits public access to its data, a user's accessing that publicly available data will not constitute access without authorization under the CFAA."[75]

67. *Id.* at 1061–68.
68. Sellars, *supra* note 6, at 393–94.
69. *Id.* at 396.
70. *Id.* at 401–07.
71. *Id.* at 408–12.
72. The original decision, *hiQ Labs, Inc. v. LinkedIn Corp.*, 938 F.3d 985 (9th Cir. 2019), was vacated by the U.S. Supreme Court after its decision in *Van Buren v. United States*, 593 U.S. 374 (2021). *See generally* LinkedIn Corp. v. hiQ Labs, Inc., 141 S. Ct. 2752 (2021). On remand, the Ninth Circuit affirmed its original decision. *See* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180 (9th Cir. 2022). For more background on this case, see generally Benjamin L.W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147 (2021); Amber Zamora, *Making Room for Big Data: Web Scraping and an Affirmative Right to Access Publicly Available Information Online*, 12 J. BUS. ENTREPRENEURSHIP & L. 203 (2019).
73. *hiQ Labs, Inc.*, 31 F.4th at 1187.
74. *Id.* at 1196.
75. *Id.* at 1201.

The U.S. Supreme Court decision in *Van Buren v. United States*[76] provided a further victory to scrapers. The Court held that liability under the CFAA "stems from a gates-up-or-down inquiry—one either can or cannot access a computer system, and one either can or cannot access certain areas within the system."[77] In other words, access is unauthorized under the CFAA only if one goes beyond a gate.

According to Orin Kerr, CFAA cases have found a lack of authorized access based on the "intended function" of technology, misconduct, or breach of an agreement.[78] But ultimately, even after *Van Buren*, Kerr views the law as only partially focused.[79] *Van Buren* and other cases do not fully resolve whether violating terms of service can serve as unauthorized access under the CFAA, though Kerr is highly skeptical that this theory is viable.[80]

b. Business and Property Interests

Beyond the CFAA, litigation over scraping has used various laws involving business and property interests. Scrapees defend their websites as their turf or the data as their property. Scrapees have tried a myriad of causes of action, such as trespass to chattels, unjust enrichment, conversion, interference with business relationships, and breach of contract. The causes of action most likely to succeed have been "breach of contract, tortious interference with a contract, and unjust enrichment."[81]

One tort that initially favored scrapees was trespass to chattels. A trespass to chattels occurs when one intentionally uses or intermeddles with a chattel of another and "the chattel is impaired as to its condition, quality, or value, or . . . the possessor is deprived of the use of the chattel for a substantial time."[82] Plaintiffs advanced the theory that scraping impairs scrapees since scraping consumes network and server resources.[83] An early case decided in 2000, *eBay v. Bidder's Edge*, held that scraping information about bids on eBay was likely a trespass and issued an injunction against Bidder's Edge.[84] Although Bidder's

76. 593 U.S. 374.

77. *Id.* at 376. For more on this concept, see generally Patricia L. Bellia, *A Code-Based Approach to Unauthorized Access Under the Computer Fraud and Abuse Act*, 84 GEO. WASH. L. REV. 1442 (2016).

78. *See generally* Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes*, 78 N.Y.U. L. Rev. 1596 (2003). For another analysis of the case law, see generally Bellia, *supra* note 77.

79. Orin S. Kerr, *Focusing the CFAA in* Van Buren, 2021 SUP. CT. REV. 155, 156 (2021).

80. *Id.* at 173. In Kerr's own view of the CFAA, he argues that norms of the internet should govern what constitutes a trespass. He rejects "virtual barriers" to scraping, such as "terms of use, hidden addresses, cookies, and IP blocks." Instead, clearer barriers should be the trigger for unauthorized access, such as circumventing an authentication requirement. Orin S. Kerr, Essay, *Norms of Computer Trespass*, 116 COLUM. L. REV. 1143, 1161 (2016).

81. McCarthy, *supra* note 36.

82. RESTATEMENT (SECOND) OF TORTS §§ 217(b), 218(b)–(c) (AM. L. INST. 1965).

83. Riley, *supra* note 11, at 265.

84. 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

Edge's bots only minimally taxed eBay's servers, the court worried about "unchecked" scraping that could lead to other scrapers descending upon eBay's site.[85]

But subsequent courts made it harder for scrapees to establish a trespass to chattels; these courts concluded that mere data gathering, without harm, was insufficient.[86] In the landmark case of *Intel Corp. v. Hamidi*, the California Supreme Court reached a similar conclusion, rejecting the comparison between physical trespass and digital information processing.[87] Ultimately, as Zachary Gold and Mark Latonero conclude, "The common law cause of action of trespass does not provide a rule clear enough for the operators of web crawlers to follow, and leaves enforcement largely up to websites, not end users whose data is actually at issue."[88]

Property is another battleground for scraping, one being fought over aggressively to this day. Many scholars have argued that personal data should be treated as property.[89] For example, Alan Westin has argued that "personal information, thought of as the right of decision over one's private personality, should be defined as a property right."[90] Lawrence Lessig has similarly argued that privacy should be protected as a property right because a property regime provides "control, and power, to the person holding the property right."[91]

Property analogies break down, though, because personal data is often shared, yet it is non-rivalrous, meaning one person's possession of it doesn't stop others from having it (or keeping it) as well.[92] Additionally, property law often focuses on the value of personal data, and courts have concluded that the value of compilations of personal data is created by the compiler, not the individuals

85. *eBay, Inc.*, 100 F. Supp. 2d at 1064.

86. Ticketmaster Corp. v. Tickets.Com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *3 (C.D. Cal. Mar. 7, 2003) "([T]his court comes down on the side of requiring some tangible interference with the use or operation of the computer being invaded by the spider. . . . Therefore, unless there is actual dispossession of the chattel for a substantial time (not present here), the elements of the tort have not been made out. Since the spider does not cause physical injury to the chattel, there must be some evidence that the use or utility of the computer (or computer network) being 'spiderized' is adversely affected by the use of the spider. No such evidence is presented here.").

87. 71 P.3d 296, 299 (Cal. 2003).

88. Gold & Latonero, *supra* note 21, at 295.

89. *See, e.g.*, Jessica Litman, *Information Privacy/Information Property*, 52 STAN. L. REV. 1283, 1287 (2000) ("The proposal that has been generating the most buzz, recently, is the idea that privacy can be cast as a property right."). For a compelling critique of privacy as property, see Pamela Samuelson, *Privacy as Intellectual Property?*, 52 STAN. L. REV. 1125, 1132 (2000) ("In recent years, a number of economists and legal commentators have argued that the law ought now to grant individuals property rights in their personal data.").

90. ALAN WESTIN, PRIVACY AND FREEDOM 324 (1967).

91. LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE 160 (1999).

92. DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 89 (2004) ("[I]nformation is often not created by the individual alone. We often develop personal information through our relationships with others. When a person purchases a product, information is created through the interaction of seller and buyer."). *See generally* JAMES BOYLE, THE PUBLIC DOMAIN: ENCLOSING THE COMMONS OF THE MIND (2008).

to whom the data pertains. For example, in *Dwyer v. American Express Co.*, the court held that by compiling profiles based on American Express cardholders' data, "[d]efendants create value by categorizing and aggregating these names. Furthermore, defendants' practices do not deprive any of the cardholders of any value their individual names may possess."[93]

Some personal data could conceivably be protected by copyright law, such as photographs. Although publicly available, copyrighted material is often not free for the taking.[94] However, there are several limitations to copyright law.[95] First, copyrighted content can be used without permission in circumstances called "fair use."[96] Indeed, scraping is creating new questions and challenges for copyright law, especially with generative AI.[97] Second, much personal data is not owned by the individual to whom it pertains. The photographer, not the subject, has the copyright.[98] The author of a biography owns the copyright, not the subject. Third, most personal data is not copyrightable, as facts cannot be copyrighted.[99]

A final potential theory is breach of contract. Under this theory, scrapers that scrape in violation of a site's terms of service are breaching a contract. Some courts have embraced this theory,[100] but the status of the terms of service as a contract remains unclear.[101]

c. Privacy Issues

Although litigants are often the scrapers and scrapees, more recent cases involve the individuals whose personal data is involved or organizations acting on behalf of these individuals.

Clearview AI's scrape of billions of photographs online triggered a lawsuit by the ACLU. In 2020, the ACLU and other groups sued Clearview AI for

93. 652 N.E.2d 1351, 1356 (Ill. App. Ct. 1995).

94. *See* Amanda Levendowski, *Resisting Face Surveillance with Copyright Law*, 100 N.C. L. REV. 1015, 1044–46 (2022).

95. *See, e.g.*, Eric Goldman & Jessica Silbey, *Copyright's Memory Hole*, 2019 BYU L. REV. 929 (2020).

96. *See, e.g.*, Levendowski, *supra* note 94, at 1049 n.208 (citing 17 U.S.C. § 107).

97. Scraping itself likely would not infringe upon copyright, only certain uses of scraping data. *See* Sobel, *supra* note 72, at 170–72. *See generally* Pamela Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, 71 UCLA L. REV. 1484 (2024); Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887 (2024); Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295 (2023).

98. Mannion v. Coors Brewing Co., 377 F. Supp. 2d 444, 454–55 (S.D.N.Y. 2005).

99. *See generally* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 338, 340 (1991); 17 U.S.C. § 102(b). *See also* Jessica Silbey, *A Matter of Facts: The Evolution of the Copyright Fact-Exclusion and Its Implications for Disinformation and Democracy*, 70 J. COPYRIGHT SOC'Y 365 (2024).

100. *See generally* Meta Platforms, Inc. v. BrandTotal Ltd., 605 F. Supp. 3d 1218 (N.D. Cal. 2022).

101. *See* DANIEL J. SOLOVE & PAUL M. SCHWARTZ, INFORMATION PRIVACY LAW 730–34 (8th ed. 2024).

violating the Illinois Biometric Information Privacy Act (BIPA).[102] Under the BIPA, private entities cannot collect a "biometric identifier or biometric information" without first informing people in writing of the specific purpose and length of use and obtaining "a written release" by people.[103] Although Clearview was in clear violation of the BIPA, the ACLU reached a settlement with Clearview that exacted only weak concessions from Clearview. Under the settlement, Clearview is permanently enjoined from granting access to its database to private entities except as consistent with the BIPA. Clearview must also refrain from granting access to Illinois government or private entities and must allow Illinois residents to opt out of being searchable in its database.[104] It remains unclear how this opt-out right compensates for violating the opt-in rights that the BIPA grants. Many measures in the settlement are short-term and barely impact Clearview's business, such as the prohibition on licensing the system to private sector entities since Clearview is mostly licensing it to law enforcement entities.

The BIPA provides redress to the individuals whose data is involved, but it is one of only a small number of state privacy laws with a private right of action, and it is limited to biometric data.[105]

Many privacy torts will likely prove ineffective against scraping. The public disclosure of private facts tort and the false light tort both require widespread dissemination of information, but scraping involves data collection and not necessarily dissemination, making these torts inapplicable.[106] The tort of intrusion upon seclusion likely will fail because the data scraped is publicly available.[107] Appropriation of name or likeness also will likely fail, as it mainly protects against the use of name or likeness to advertise or endorse products, not the use of personal data of many people compiled together.[108] However, of all the torts, rights of appropriation and publicity might be most helpful with respect to images and videos of people's names and likeness.[109] There has been at least one small victory regarding the appropriation tort. In *Renderos v. Clearview AI*, the plaintiffs alleging misappropriation of name or likeness for Clearview AI's

102. ACLU v. Clearview AI, Inc., No. 20 CH 4353, 2021 Ill. Cir. LEXIS 292 (Ill. Cir. Ct. Aug. 27, 2021).

103. 740 ILL. COMP. STAT. 14/15(b)(1)–(3) (2024).

104. Consent Order, ACLU v. Clearview AI Inc., No. 2020 CH 04353 (Ill. Cir. Ct. May 11, 2022).

105. The Washington My Health My Data Act provides protection of health data that is broadly defined to encompass biometric data, but it, too, is limited in scope and does not apply to all personal data.

106. *See* RESTATEMENT (SECOND) OF TORTS § 652D, 652E (AM. L. INST. 1997).

107. *See generally* Reece v. Grissom, 267 S.E.2d 839 (Ga. Ct. App. 1980) (holding that there is no privacy interest in information available in a public record); Heath v. Playboy Enters., Inc., 732 F. Supp. 1145 (S.D. Fla. 1990) (holding that there is no privacy interest in facts already publicized).

108. *See* SOLOVE & SCHWARTZ, *supra* note 101, at 193–95.

109. *See generally* Jason M. Schultz, *The Right of Publicity: A New Framework for Regulating Facial Recognition*, 88 BROOK. L. REV. 1039 (2023).

collection and use of faceprints survived a motion to dismiss.[110] The Superior Court of California (Alameda County) held that:

> The Complaint alleges that Clearview extracted plaintiffs' faceprints, did the biometric analysis, maintained the data in a database, and then sold that information for profit. Clearview's "appropriation" was the taking of the likenesses from the internet. Clearview then "used" the likenesses. Clearview was free to use the likenesses, to pass them along, or to participate in commentary on social media on matters concerning the likenesses. That would have been "use" without "advantage." Clearview used the likenesses to its "advantage, commercially or otherwise." The "advantage, commercially or otherwise" consisted of the of the [sic] use of the images as the raw material for its biometric analysis, the data in the database, and then as part of the finished product when Clearview sold its services to law enforcement.[111]

A major class action recently launched against OpenAI's scraping alleges violations of a panoply of common law and statutory causes of action, including negligence, intrusion upon seclusion, larceny, conversion, unjust enrichment, failure to warn, the Illinois BIPA, and state unfair and deceptive act or practice (UDAP) statutes.[112] As is common in litigation such as this, plaintiffs throw a multitude of causes of action against the wall, hoping one will stick. Perhaps one cause of action will prevail here or there, but if the litigation plays out as it has with the CFAA and torts involving business and property interests, the result will likely be muddy terrain, with scrapers continuing to scrape and just watching out for an occasional landmine.

Overall, however, privacy litigation for scraping has been minimal compared to the extensive battles under the CFAA and business and property torts. Many companies use the CFAA "as a means of eliminating competitors whose business models rely on data scraping."[113] Even when companies say they are fighting scrapers, they are often pursuing their own competitive advantage and using "privacy" as a pretext.[114] Additionally, litigation involving the terms of service, the CFAA, or both typically is between the scrapers and scrapees, leaving the individuals whose data is scraped on the sidelines.

Consider the *hiQ* case, where the court briefly considered the privacy interests of half a billion LinkedIn members in concluding that one company's business interests outweighed them:

---

110. Order re: Ruling on Submitted Matter at 5–6, Renderos v. Clearview AI, Inc., No. RG21096898 (Cal. Super. Ct. Nov. 18, 2022).

111. *Id.* (internal citations omitted).

112. *See generally* Class Action Complaint, P.M. v. OpenAI, LP, No. 3:23-cv-03199 (N.D. Cal. June 28, 2023).

113. Riley, *supra* note 11, at 250.

114. Erika M. Douglas, *Data Privacy as a Procompetitive Justification: Antitrust Law and Economic Analysis*, 97 NOTRE DAME L. REV. REFLECTION 430, 430 (2022) ("Digital platforms are invoking data privacy to justify their anticompetitive conduct.").

> [E]ven if some users retain some privacy interests in their information notwithstanding their decision to make their profiles public, we cannot, on the record before us, conclude that those interests—or more specifically, LinkedIn's interest in preventing hiQ from scraping those profiles—are significant enough to outweigh hiQ's interest in continuing its business, which depends on accessing, analyzing, and communicating information derived from public LinkedIn profiles.[115]

Most of the litigation over scraping amounts to a tussle between companies over the spoils of the data extraction economy. Companies might say they are fighting for their users' privacy, but they are really shielding data they believe is theirs or protecting their websites and their own business interests. Ultimately, user privacy and security are invoked when they align with corporate interests; when they do not, the story is different.

This is a war over resources and territory, and it plays out with property, contract, and business concepts. The privacy of individuals is not much of a consideration.

2. The Technological Front

On the technological front, the Scraping Wars are ramping up as many websites are using technology to try to block AI scraping bots. There are a range of modern anti-scraping techniques that websites can use. Such techniques include access restrictions, Captchas, rate limiting, browser fingerprinting, and banning users' accounts and IP addresses.[116] But these measures can be circumvented.[117] Scraping and preventing scraping is a cat-and-mouse game.

For a long time, social media platforms offered APIs to facilitate third-parties' use of data.[118] "APIs are code interfaces that allow programmers to make very formal data requests from websites within a specific interface."[119] But in 2018, the Cambridge Analytica scandal changed views about the costs and benefits of allowing API access.[120] In the wake of this incident, many social

115. hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1190 (9th Cir. 2022).

116. Keary, *supra* note 13; Michael Nyamande, *Web Scraping Without Getting Blocked*, BRIGHT DATA https://brightdata.com/blog/web-data/web-scraping-without-getting-blocked [https://perma.cc/C3SC-5T9H]; Assad Abbas, *Defending the Digital Frontier Through Anti-Web Scraping Measures*, TECHOPEDIA (Aug. 28, 2023), https://www.techopedia.com/defending-the-digital-frontier-through-anti-web-scraping-measures [https://perma.cc/K9WY-JVCU]; Jeffrey Kenneth Hirschey, *Symbiotic Relationships: Pragmatic Acceptance of Data Scraping*, 29 BERKELEY TECH. L.J. 897, 918 (2014).

117. *See* Satyam Tripathi, Most Popular Anti-Scraping Techniques in 2025, BRIGHT DATA (Oct. 8, 2024), https://brightdata.com/blog/web-data/anti-scraping-techniques [https://perma.cc/P8RK-G7WA].

118. GANAELE LANGLOIS, JOANNA REDDEN & GREG ELMER, COMPROMISED DATA: FROM SOCIAL MEDIA TO BIG DATA 120 (2015).

119. Hirschey, *supra* note 116, at 905.

120. Domenico Trezza, *To Scrape or Not to Scrape, This Is Dilemma. The Post-API Scenario and Implications on Digital Research*, 8 FRONTIERS SOCIO. 1, 1 (2023). Here, "Cambridge Analytica used a 'loophole' in Facebook's APIs to collect data from over 80 million users between 2013 and

media companies curtailed their own APIs[121] or increased costs to discourage improper uses.[122] This move created even more incentives for companies to use web scraping to obtain data.

When OpenAI released its new web crawler, it provided instructions for how websites could update robots.txt to stop its bots from scraping.[123] Several large media companies have blocked OpenAI's scraping bots.[124]

But not all scrapers play by the rules of chivalry. As technology journalist David Pierce observes, "The robots.txt file governs a give and take; AI feels to many like all take and no give."[125] Web scrapers now also often use "additional technologies to mimic human browsing and delve deeper into each website."[126] *The New York Times* contends its site is still being scraped contrary to its robots.txt instructions.[127] Some scrapers have found ways to evade paywalls on websites.[128]

Meta declared that it has implemented "several measures . . . to mitigate the risk of scraping on [its] platform."[129] For example, it has "an External Data Misuse team that consists of more than 100 people dedicated to detecting, investigating and blocking patterns of behavior associated with scraping."[130] It imposes "rate and data limits, which are designed to restrict how much data a single person can obtain through a certain feature."[131] And it has initiated hundreds of enforcement actions, such as "sending cease and desist letters, disabling accounts, filing lawsuits or requesting assistance from hosting

---

2015." Bernard Harguindeguy, *Facebook Data Breach Highlights API Vulnerabilities*, PING IDENTITY (Oct. 2, 2018), https://www.pingidentity.com/en/resources/blog/post/facebook-data-breach-highlights-api-vulnerabilities.html [https://perma.cc/EF9J-C9RS]; *see also* Katie Harbath & Collier Fernekes, *History of the Cambridge Analytica Controversy*, BIPARTISAN POL'Y CTR. (Mar. 16, 2023), https://bipartisanpolicy.org/blog/cambridge-analytica-controversy/ [https://perma.cc/3THV-QGMP].

121. Trezza, *supra* note 120.

122. Andrew Hutchinson, *Twitter Implements Usage Limits for All to Combat Data Scrapers*, SOC. MEDIA TODAY (July 1, 2023), https://www.socialmediatoday.com/news/twitter-implements-usage-limits-combat-data-scrapers/684831/ [https://perma.cc/7KBY-NPFF].

123. Ben Wodecki, *OpenAI Quietly Unveils Web Crawler to Scrape Data for Its AI Models*, AI BUS. (Aug. 8, 2023), https://aibusiness.com/nlp/openai-unveils-web-crawler-to-gather-data-to-improve-ai-models#close-modal [https://perma.cc/78F4-WEXM].

124. Oliver Darcy, *Disney, The New York Times and CNN Are Among a Dozen Major Media Companies Blocking Access to ChatGPT as They Wage a Cold War on A.I.*, CNN (Aug. 28, 2023), https://www.cnn.com/2023/08/28/media/media-companies-blocking-chatgpt-reliable-sources/index.html [https://perma.cc/EGD5-A524].

125. Pierce, *supra* note 19.

126. Nicholas A. Wolfe, *Hacking the Anti-Hacking Statute: Using the Computer Fraud and Abuse Act to Secure Public Data Exclusivity*, 13 NW. J. TECH. & INTELL. PROP. 301, 305 (2015).

127. Benj Edwards, *The New York Times Prohibits AI Vendors from Scraping Its Content Without Permission*, ARS TECHNICA (Aug. 14, 2023), https://arstechnica.com/information-technology/2023/08/the-new-york-times-prohibits-ai-vendors-from-devouring-its-content/ [https://perma.cc/F9QB-UJGB].

128. Leffer, *supra* note 43.

129. Clark, *supra* note 14.

130. *Id.*

131. *Id.*

providers to get them taken down."[132] It also blocks "billions of suspected scraping actions per day across Facebook and Instagram."[133]

Battles over scraping will continue to be fought on both legal and technological fronts for years to come. The stakes are enormous. The age of chivalry is over. This is war.

C. The Emerging Scraping Market

A market-based alternative to the Scraping Wars has been arising, but it is unsatisfactory. Scrapers are starting to reach deals with scrapees, paying them for the right to scrape their land, or obtaining their data through other means. For example, OpenAI has started to enter into agreements with companies to obtain their data. OpenAI made deals with media companies to obtain data from their articles.[134] In 2023, OpenAI reached deals with the Associated Press and Axel Springer, the parent company of *Politico* and *Business Insider*.[135] These deals implicate personal data, as news stories have extensive personal data. OpenAI also reached a deal with Shutterstock, a site where users buy and sell images.[136] What companies like OpenAI cannot obtain through agreement, they likely will obtain by scraping publicly available websites.

The market may quell some battles, but it provides an unsatisfactory peace. The individuals to whom the data pertains are not involved in the dealmaking; they receive no financial benefits from the deals, but they are at risk of harm. A peace deal is inadequate if it leaves out a major party.

D. Relevant Regulatory Intervention

Even though scraping has been occurring for a long time, regulators in the United States, the EU, and elsewhere around the world have generally avoided stepping onto the battlefield. Recently, however, some regulators like Data Protection Authorities in the EU and the U.S. Federal Trade Commission have

---

132. *Id.*

133. *Id.*

134. Anna Tong, Echo Wang & Martin Coulter, *Exclusive: Reddit in AI Content Licensing Deal with Google*, REUTERS (Feb. 21, 2024), https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/ [https://perma.cc/HV24-6AH4].

135. Thomas Barrabi, *OpenAI Offering Media Outlets as Little as $1M to Use News Articles for AI Models*, N.Y. POST (Jan. 4, 2024), https://nypost.com/2024/01/04/business/openai-offering-media-outlets-as-little-as-1-million-to-use-news-articles-for-ai-models-report/ [https://perma.cc/D8BR-2HJ9]; *see also* Gerrit De Vynck, *OpenAI Strikes Deal with AP to Pay for Using Its News in Training AI*, WASH. POST (July 13, 2023), https://www.washingtonpost.com/technology/2023/07/13/openai-chatgpt-pay-ap-news-ai/ [https://perma.cc/C5HW-MHW5]; Matt O'Brien, *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, ASSOCIATED PRESS (July 13, 2023), https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a [https://perma.cc/9TTM-6MA4].

136. *Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data*, SHUTTERSTOCK (July 11, 2023), https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year [https://perma.cc/G4SK-A6DQ].

begun to tepidly step into the fray, but they have found themselves ill-prepared for life on the battlefront.

1. EU Data Protection Law

The EU has probably provided the most robust regulatory response to data collection, which has significant implications for scraping. For example, it is quite difficult to reconcile scraping personal data with the EU's General Data Protection Regulation (GDPR), which requires a legal basis for data processing and imposes various transparency and autonomy-enhancing safeguards.

Under the GDPR, there is no general exception for publicly available information.[137] Instead, personal data can be collected and processed based on one of six lawful bases: (1) consent; (2) necessity for a contract; (3) necessity to comply with a legal obligation; (4) necessity to protect a person's vital interests; (5) necessity for the public interest; and (6) necessity for legitimate interests and not "overridden by the interests or fundamental rights and freedoms of the data subject."[138]

It remains unclear whether scraping fits under any lawful basis. Regarding consent, EU regulators have stated that even though personal data is publicly available online, scrapers must still obtain individual consent to scrape.[139] Given the vast number of individuals involved, obtaining the consent of each person is practically impossible.

The lawful basis that most seemingly fits is legitimate interests, but it is far from reliable. First, many of the purposes of collecting personal data for AI are too unspecified to work under this basis, especially general use AI where data can be used for a nearly infinite number of purposes. Second, it remains unclear how each use would fare under the balancing test with data subjects' fundamental rights and freedoms. Third, sensitive data cannot be processed for legitimate interests.[140] As one of us has written elsewhere, because inferences from non-sensitive data (in isolation or combination) can count as sensitive data, nearly all personal data could be sensitive data.[141]

In March 2023, in a bold move, the Data Protection Authority (DPA) of Italy banned ChatGPT. The DPA stated that "there appears to be no legal basis

137. Though the GDPR does provide an exception for heightened protections on sensitive data when "processing relates to personal data which are manifestly made public by the data subject." Regulation 2016/679, art. 9.2(e), of the European Parliament and of the Council of 27 April 2016, 2016 O.J. (L. 119) [hereinafter GDPR].

138. *Id.* art. 6.1(f).

139. Müge Fazlioglu, *Training AI on Personal Data Scraped from the Web*, IAPP (Nov. 8, 2023), https://iapp.org/news/a/training-ai-on-personal-data-scraped-from-the-web/ [https://perma.cc/E7BB-HA6K].

140. Daniel J. Solove, *Data Is What Data Does: Regulating Based on Harm and Risk Instead of Sensitive Data*, 118 NW. U. L. REV. 1081, 1095–97 (2024).

141. *Id.*

underpinning the massive collection and processing of personal data in order to 'train' the algorithms on which the platform relies."[142]

But in late April of 2023, in a rather awkward walk-back, the DPA then reinstated ChatGPT.[143] The DPA found that ChatGPT could satisfy the GDPR with a mechanism to allow people to remove their data and with age verification—a rather farcical capitulation on the part of the DPA. It stated that OpenAI would need to rely on either consent or legitimate interests as the applicable legal basis for processing under the GDPR.[144] As an article in *The Verge* appropriately put it, "So far, none of these changes seem to dramatically modify how ChatGPT operates in Italy."[145]

Thus, scraping continues in the EU. However, a full showdown between the GDPR and scrapers is near. In a recent guide on scraping personal data, the Dutch data protection authority, the Autoriteit Persoonsgegevens (AP), held that scraping personal information is almost always a violation of the GDPR.[146] The AP stated that certain kinds of scraping are prohibited, such as scraping the internet to create profiles of people and then resell them, scraping information from protected social media accounts or private forums, and scraping data from public social media profiles to determine whether those people will receive requested insurance.[147] In practice, the AP said that the only legal basis for scraping would be having a "legitimate interest" under Article 6(1)(f) of the GDPR. However, the AP suggested that if the sole purpose of scraping by data processors was to make money, this would not qualify as "legitimate."[148] According to the AP, in practice it is almost never possible to meet the conditions of the legitimate interest test when scraping for financial gain.[149] If the rest of the DPAs in the EU hold the same opinion, this would essentially prohibit scraping for profit by commercial entities, which would be a dramatic prohibition.

---

142. *Artificial Intelligence: Stop to ChatGPT by the Italian SA*, GARANTE PER LA PROTEZIONE DEI DATI PERSONALI (Mar. 31, 2023) https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847#english [https://perma.cc/T9E6-4XVG] (English translation).

143. *ChatGPT: Italian SA to Lift Temporary Limitation if OpenAI Implements Measures: 30 April Set as Deadline for Compliance*, GARANTE PER LA PROTEZIONE DEI DATI PERSONALI (Apr. 12, 2023), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9874751 [https://perma.cc/937B-KFPY] (English translation).

144. *Id.*

145. Adi Robertson, *ChatGPT Returns to Italy After Ban*, VERGE (Apr. 28, 2023), https://www.theverge.com/2023/4/28/23702883/chatgpt-italy-ban-lifted-gpdp-data-protection-age-verification [https://perma.cc/8B64-RRXV].

146. *Scraping Bijna Altijd Illegaal [Scraping Is Almost Always Illegal]*, AUTORITEIT PERSOONSGEGEVENS [DUTCH DATA PROT. AUTH.] (May 1, 2024), https://autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal [https://perma.cc/F7JR-CS8Q].

147. *Id.*

148. *Id.*

149. *Id.* Some of the examples the AP gave of exceptional cases when there might be a legitimate interest in scraping were when a private individual uses scraping for a hobby project and only shares the results with a few friends or when an organization scrapes the websites of news media in a very targeted way to gain insight into relevant news about its own company. *Id.*

Beyond the GDPR, in a joint statement of data protection commissioners from the United Kingdom, Switzerland, Australia, New Zealand, Argentina, and other countries, the commissioners stated:

- Personal information that is publicly accessible is still subject to data protection and privacy laws in most jurisdictions.
- Social media companies and the operators of websites that host publicly accessible personal data have obligations under data protection and privacy laws to protect personal information on their platforms from unlawful data scraping.
- Mass data scraping incidents that harvest personal information can constitute reportable data breaches in many jurisdictions.[150]

The commissioners further stated that "websites should implement multi-layered technical and procedural controls to mitigate the risks."[151] Interestingly, the joint statement did not focus on the scrapers and their violations of privacy law or on how the commissioners would enforce the laws against both scrapers and scrapees.

The controversial practices of Clearview AI sparked a wave of regulatory action in the UK and EU with mixed results. In the UK, the Information Commissioner's Office (ICO) fined Clearview £7.5 million and ordered Clearview to delete personal data collected about UK citizens. The ICO alleged that Clearview's scraping violated the UK's GDPR (which is essentially a cut-and-paste of the EU's GDPR), as Clearview lacked a lawful basis to collect the data. Clearview also failed to comply with the conditions for lawful processing of sensitive data and failed to provide information to data subjects about the data processing. Additionally, the ICO found a litany of other violations of the UK GDPR. On appeal, however, the First-Tier Tribunal concluded that Clearview fell outside the jurisdiction of the UK GDPR because Clearview's services were provided only to non-UK/EU law enforcement entities.[152]

In 2022, France's National Commission on Informatics and Liberty (CNIL) fined Clearview €20 million, the maximum GDPR penalty, when Clearview failed to comply with a 2021 injunction.[153] Italy also imposed the same fine in 2022, ordering Clearview to cease scraping and delete all data from people in

150. *Joint Statement on Data Scraping and the Protection of Privacy*, INFO. COMM'R'S OFF. (Aug. 24, 2023) https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf [https://perma.cc/CQS7-Z7FB].

151. *Id.*

152. Clearview AI Inc. v. Info. Comm'r [2023] UKFTT 00819 (GRC); DECHERT LLP, TRIBUNAL OVERTURNS UK ICO'S ENFORCEMENT ACTION AGAINST CLEARVIEW AI (2023), https://www.dechert.com/knowledge/ onpoint/2023/11/tribunal-overturns-uk-ico-s-enforcement-action-against-clearview.html [https://perma.cc/Z4JL-KDW4].

153. *The French SA Fines Clearview AI EUR 20 Million*, EUR. DATA PROT. BD. (Oct. 20, 2022), https://www.edpb.europa.eu/news/national-news/2022/french-sa-fines-clearview-ai-eur-20-million_en [https://perma.cc/5259-9C9V].

Italy.[154] Likewise, in 2022, Greece's data protection authority issued a €20 million fine and a similar order to cease and delete.[155] In 2023, Austria's data protection authority found Clearview to be in violation of the GDPR and issued an order to delete the data but did not issue a fine.[156]

Although Clearview is being chased out of the EU, Clearview is only one scraper among an invading army of scrapers.

2. U.S. Privacy Law

In the United States, although many privacy laws have loopholes where scraping can occur, not all do. Existing privacy laws have some tools to regulate scrapers and scrapees. Most notably, scraping as well as the failure to defend against scraping could constitute violations of the Federal Trade Commission (FTC) Act Section 5, which prohibits "unfair or deceptive" acts or practices.[157]

The FTC has ample discretion under the FTC Act to conclude that scraping constitutes an unfair act or practice, which is one that "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and is not outweighed by countervailing benefits to consumers or to competition."[158] One could argue that consumers might be able to avoid their data being scraped if they just do not have public profiles on social media or refrain from tweeting or writing online. One could also argue that scraping does not cause substantial injury to consumers or that it provides benefits and promotes competition for AI. But FTC jurisprudence certainly could support a claim that scraping is unfair, such as *In re Vision I Properties*, where the FTC concluded that a company's violation of the privacy policies of other companies was unfair.[159]

If the FTC were to find scrapers in violation of the FTC Act, the FTC could require that they delete models developed with improperly gathered data.[160] But

154. *Facial Recognition: Italian SA Fines Clearview AI EUR 20 Million*, EUR. DATA PROT. BD. (Mar. 10, 2022), https://www.edpb.europa.eu/news/national-news/2022/facial-recognition-italian-sa-fines-clearview-ai-eur-20-million_en [https://perma.cc/XT6M-GW64].

155. *Hellenic DPA Fines Clearview AI 20 Million Euros*, EUR. DATA PROT. BD. (July 20, 2022), https://www.edpb.europa.eu/news/national-news/2022/hellenic-dpa-fines-clearview-ai-20-million-euros_en [https://perma.cc/RZG9-PP8G].

156. *Decision by the Austrian SA Against Clearview AI Infringements of Articles 5, 6, 9, 27 GDPR*, EUR. DATA PROT. BD. (May 12, 2023), https://www.edpb.europa.eu/news/national-news/2023/decision-austrian-sa-against-clearview-ai-infringements-articles-5-6-9-27_en [https://perma.cc/YHH2-DN3V].

157. 15 U.S.C. § 45.

158. *Id.* § 45(n).

159. *See generally* Vision I Properties, LLC, 139 F.T.C. 296 (2005); Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

160. For an example of the FTC requiring algorithmic destruction, see generally Everalbum, Inc., Docket No. C-4743, File No. 192-3172 (F.T.C. May 6, 2021). As Professor Tiffany Li points out, algorithms have already learned from the data, so merely deleting the data after the fact does not erase the benefit gained from collecting it. Tiffany C. Li, *Algorithmic Destruction*, 75 SMU L. REV. 479, 482, 498 (2022). In what she calls an "algorithmic shadow," the data has a "persistent imprint" in the machine

it is hard to imagine the FTC would be so bold as to find that scraping violates the FTC Act and issue such a penalty against popular AI algorithms such as ChatGPT. The FTC faces political constraints on its power and has been cautious ever since Congress disciplined it for its regulation of advertising to children in the 1970s.[161] The more collective, intangible, and dispersed harms of scraping are also often beyond the kinds of acute exposure and injury typically spurring on FTC complaints.[162] Given how many AI algorithms were developed by massive scraping, perhaps most would have to be deleted.

For the scrapees, failing to safeguard against scraping could be a deceptive practice under the FTC Act because this could contravene promises in a privacy notice, such as that data will be protected by reasonable data security, that data will not be transferred to third parties, and that data will only be used for specified purposes. Like the act of scraping itself, the failure to protect against scraping could be an unfair practice because it can cause substantial injury to consumers.

Although the FTC has tools to use against both scrapers and scrapees, it is unlikely that the FTC has the fortitude and political power to use them in a vigorous way. As Alicia Solow-Niederman notes, there is an "Overton Window" to the FTC's power—political constraints prevent the FTC from being too bold.[163]

* * *

Overall, privacy law's attempt to address scraping is inconsistent, shifting, unclear, and incomplete. As we will discuss in the next Part, a fundamental tension between scraping and privacy explains this struggle. Grappling with this tension is essential to make progress toward a more coherent and pragmatic legal approach.

II.
SCRAPING AND PRIVACY: A FUNDAMENTAL TENSION

Although privacy is a vague concept, information privacy law has settled on a set of bedrock principles known as the "Fair Information Practice Principles" (FIPPs) that make up the common language of data privacy around the world.[164] An early version of the foundational FIPPs was articulated in 1973

learning algorithm. *Id.* at 482. Merely deleting the data does not delete the algorithmic shadow and has "no impact on an already trained model." *Id.* at 490.

161. *See generally* Woodrow Hartzog & Daniel J. Solove, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230 (2015) (critiquing the FTC for being too cautious).

162. *See generally id.* (exploring the kind of harms typically targeted by FTC complaints).

163. *See generally* Alicia Solow-Niederman, *The Overton Window and Privacy Enforcement*, 37 HARV. J.L. & TECH. 1007 (2023).

164. *See* COLIN J. BENNETT & CHARLES D. RAAB, THE GOVERNANCE OF PRIVACY: POLICY INSTRUMENTS IN GLOBAL PERSPECTIVE 12 (2006). *See generally* CHRISTOPHER KUNER, EUROPEAN DATA PROTECTION LAW: CORPORATE COMPLIANCE AND REGULATION (2d ed. 2007); MEG LETA JONES, THE CHARACTER OF CONSENT: THE HISTORY OF COOKIES AND THE FUTURE OF TECHNOLOGY

and then expanded in the Organisation for Economic Co-operation and Development (OECD) Privacy Guidelines of 1980.[165] The FIPPs are the backbone of privacy laws around the world, as well as countless privacy frameworks, standards, and codes.[166]

The OECD developed these principles in response to fears about the power of digital databases to make information much easier to collect, store, aggregate, search, and share.[167] The basic concepts of the FIPPs are simple: Only collect data when necessary for a legitimate purpose spelled out in advance, keep the data safe and accurate, and do everything in a transparent and accountable way.[168] If there were a common language of privacy, it would be the FIPPs.[169]

In this Part, we argue that scraping of personal data is incompatible with nearly all the FIPPs and many of the core provisions in countless privacy laws. This problem is not a minor one that can be fixed with some small tweaks. Scraping fundamentally clashes with common goals of privacy laws and with the very FIPPs model in which most privacy laws regulate how personal data should be collected, used, and transferred.

Surprisingly, this dramatic conflict has been greatly underappreciated. We are witnessing a tectonic clash between scraping and privacy, yet most policymakers, commentators, and organizations seem unaware. Since scraping and the core model of most privacy laws are fundamentally incompatible, radical changes must be made to scraping, privacy law, or both.

A. Scraping and Privacy Principles

The FIPPs create a vision for data privacy built on fairness, individual autonomy, and processor accountability. Scraping does not work with this model of privacy protection; trying to fit it in is akin to trying to pound a square peg into a round hole. Specifically, scraping violates several fundamental privacy principles: (1) fairness; (2) individual rights and control; (3) transparency; (4) consent; (5) purpose specification and secondary use restrictions; (6) data minimization; (7) onward transfer; and (8) data security.

POLICY (2024) (detailing the history of how the Fair Information Practices were developed); Woodrow Hartzog, *The Inadequate, Invaluable Fair Information Practices*, 76 MD. L. REV. 952, 982 (2017); Paula Bruening, *Fair Information Practice Principles: A Common Language for Privacy in a Diverse Data Environment*, INTEL: POLICY@INTEL (Jan. 28, 2016), http://blogs.intel.com/policy/2016/01/28/blah-2/ [https://perma.cc/B9LG-FE5N]; Robert Gellman, Fair Information Practices: A Basic History (Apr. 9, 2024) (unpublished manuscript), https://bobgellman.com/rg-docs/rg-FIPshistory.pdf [https://perma.cc/F756-XQX5].

165. SOLOVE & SCHWARTZ, *supra* note 101, at 580–81.

166. Gellman, *supra* note 164, at 1.

167. *See* JONES, *supra* note 164, at 38–60.

168. *Id.*

169. Bruening, *supra* note 164.

1. Fairness

Aptly, the overarching goal of the FIPPs is fairness, which is why they are called the *Fair* Information Practice Principles. Fairness is a rather vast concept, and in the context of privacy it has many components. According to the UK ICO, "[F]airness means that you should only handle personal data in ways that people would reasonably expect and not use it in ways that have unjustified adverse effects on them."[170] Similar concerns animate the FTC's regulation of unfair and deceptive trade practices.[171]

Although subject to many different definitions and containing many disparate elements, fairness is generally a robust and far-reaching set of requirements protecting both collective groups and individuals from unwarranted harm.[172] Gianclaudio Malgieri has argued that "fairness is effect-based: [W]hat is relevant is not the formal respect of procedures (in terms of transparency, lawfulness or accountability), but the substantial mitigation of unfair imbalances that create situations of 'vulnerability.'"[173] Under a broad conception of the FIPPs, fairness also involves the responsible collection and processing of personal data as well as respect for the interests of the individuals to whom the data pertains.

Scraping violates the fairness principle because it is hidden and harmful. In a joint statement, DPAs from around the world found that scraped data can be used for cyberattacks, identity fraud, profiling, surveillance, unauthorized intelligence gathering, and spam.[174] Because people are not notified when their data is scraped, they are often left unaware of data processing that exposes them to risk.

2. Individual Rights and Control

Another central privacy principle is ensuring individuals have some control over how their data is collected and used. This principle is often associated with

---

170. *Principle (a): Lawfulness, Fairness and Transparency*, INFO. COMM'R'S OFF. (Jan. 10, 2025), https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/lawfulness-fairness-and-transparency [https://perma.cc/L6GG-2UFD].

171. Federal Trade Commission Act § 5, 15 U.S.C. § 45. *See generally* Solove & Hartzog, *supra* note 159.

172. *See* Gianclaudio Malgieri, *The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation*, 2020 PROCS. FAT* 1, 3 ("[I]t seems clear that fairness cannot be reduced to a synonym of transparency or lawfulness, but has an independent meaning. That specific meaning can have different nuances if it is combined with the transparency principle or with the lawfulness principle. The notion of fairness in the GDPR seems to refer to a substantial approach, aimed at preventing adverse effects in concrete circumstances situations, in particular when conflicting interests need to be balanced. However, the idea of fairness can have many possible nuances: non-discrimination, fair balancing, procedural fairness, bona fide, etc.").

173. *Id.* at 2.

174. *Joint Statement on Data Scraping and the Protection of Privacy*, *supra* note 150, at 2.

broader autonomy-focused concepts like "informational self-determination."[175] Privacy law attempts to provide individuals with control over their personal data, often in the form of individual rights such as a right to access, correct, and delete data.[176] We have argued that such control is insufficient to protect privacy and that privacy laws rely far too heavily upon individual rights, but this is a central pillar of how privacy laws currently work.[177]

When people share information online, they have privacy expectations connected with the use of this information. Research on privacy expectations has consistently shown that people desire control over their personal data and expect that recipients of their personal data will protect it from unauthorized access.[178] People's privacy expectations differ based on the specific situation in which data is shared.[179] A diverse set of contextual factors can affect people's privacy expectations and behavior—such as rules and policies, user interface design, culture, past experiences, the behavior of other people, and even the physical

175. The term "informational self-determination" originates in a 1983 decision of the German Federal Constitutional Court. Antoinette Rouvroy & Yves Poullet, *The Right to Informational Self-Determination and the Value of Self-Development: Reassessing the Importance of Privacy for Democracy*, *in* REINVENTING DATA PROTECTION? 45, 45 (Serge Gutwirth, Yves Poullet, Paul De Hert, Cecile de Terwangne & Sjaak Houwt eds., 2009).

176. Daniel J. Solove, *The Limitations of Privacy Rights*, 98 NOTRE DAME L. REV. 975, 975 (2023).

177. *See generally* Daniel J. Solove & Woodrow Hartzog, *Kafka in the Age of AI and the Futility of Privacy as Control*, 104 B.U. L. REV. 1021 (2024).

178. *See generally* Antje Niemann & Manfred Schwaiger, *Consumers' Expectations of Fair Data Collection and Usage – A Mixed Method Analysis*, 2016 49TH HAW. INT'L CONF. ON SYS. SCIS. 3646, 3649 ("Customers expect to be able to control the use of their data and want to do so in an increasingly granular fashion. . . . [C]ustomers expect companies to protect their personal data from unauthorized access."); Yun Zhou, Alexander Raake, Tao Xu & Xuyun Zhang, *Users' Perceived Control, Trust and Expectation on Privacy Settings of Smartphone*, 2017 NINTH INT'L CYBERSPACE SAFETY & SEC. SYMP. 427; Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Cranor & Norman Sadeh, *Privacy Expectations and Preferences in an IoT World*, 2017 THIRTEENTH SYMP. ON USABLE PRIV. & SEC. 399; Igor Bilogrevic & Martin Ortlieb, *"If You Put All the Pieces Together . . . " – Attitudes Towards Data Combination and Sharing Across Services and Companies*, 2016 PROCS. 2016 CHI CONF. ON HUM. FACTORS COMPUTING SYS. 5215.

179. *See generally* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE 3–7 (2010) (developing a theory of privacy as contextual integrity); Anne Adams, *Multimedia Information Changes the Whole Privacy Ballgame*, 2000 PROCS. TENTH CONF. ON COMPUTS., FREEDOM & PRIV. 25 (developing a model whereby three factors—information receivers (mediated by trust), potential usage of collected data (affecting risk/benefit trade-offs), and information sensitivity—affect users' perceptions of privacy in multimedia communications); Sandra Petronio, *Communication Boundary Management: A Theoretical Model of Managing Disclosure of Private Information Between Marital Couples*, 1 COMMC'N THEORY 311 (1991); Sandra Petronio, *Brief Status Report on Communication Privacy Management Theory*, 13 J. FAM. COMMC'N 6 (2013); Irwin Altman, *Privacy Regulation: Culturally Universal or Culturally Specific?*, 33 J. SOC. ISSUES 66 (1977); IRWIN ALTMAN, THE ENVIRONMENT AND SOCIAL BEHAVIOR 10–21 (1975) (theorizing that privacy is not a static condition with universal rules, but rather is a dynamic, situationally specific, and selective process of boundary regulation and control of access to the self). According to Altman, a person's desired level of privacy is continuously changing along a continuum between openness and closeness in response to context and circumstances.

environment.[180] The fact that privacy expectations are shaped by contextual factors is important because these privacy expectations influence how, when, and to what extent people decide to share personal data.[181]

Scraping strips away the original context in which data is shared. All the many factors which were present when people shared their data, such as when, where, how, to whom, and why, are missing with scraping. Thus, scraping thwarts people's privacy expectations and fails to respect their initial decisions about how and when to share their personal data. In a joint statement, DPAs from around the world wrote that "individuals lose control of their personal information when it is scraped without their knowledge and against their expectations."[182] Privacy law's goal of promoting informational self-determination cannot be achieved in a world of ubiquitous data scraping.

In short, it renders individual privacy rights meaningless. For example, a right to delete personal data is ineffectual if unknown scrapers have obtained this data, leaving individuals powerless to demand its deletion. Ironically, the original organizations entrusted with people's data end up with far less power over the data than any random third-party scraper. Scraping strips people of their rights and often places personal data outside the sphere of any privacy protection.

3. Transparency

Another core privacy principle is transparency concerning personal data collection and usage. Nearly all privacy laws require that data processors inform individuals about the data gathered about them and from them, state the purposes of use, and describe their practices for protecting that data.[183]

Scrapers, however, disregard these transparency requirements entirely. Scrapers vacuum up the data to be used for a multitude of different purposes. There is no notice to individuals before, during, or after scraping occurs. There is some debate as to whether a general notice, explaining in detail a scraper's activities and posted on the scraper's website, can satisfy transparency rules like the one in the GDPR if individual delivery of notice would be too burdensome.[184]

---

180. *See generally* Alessandro Acquisti, Laura Brandimarte & George Loewenstein, *Privacy and Human Behavior in the Age of Information*, 347 SCIENCE 509 (2015); Alisa Frik, Julia Bernd, Noura Alomar & Serge Egelman, *A Qualitative Model of Older Adults' Contextual Decision-Making About Information Sharing*, 2020 PROCS. 20TH ANN. WORKSHOP ON ECON. INFO. SEC. 1 (proposing a comprehensive model of factors affecting the context-specific decision-making of older adults about information sharing along seven dimensions: decision maker, data, recipients, purposes and benefits, risks, system, and environment).

181. Ashwini Rao, Florian Schaub, Norman Sadeh & Alessandro Acquisti, *Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online*, 2016 TWELFTH SYMP. ON USABLE PRIV. & SEC. 77, 77 ("[E]xpectations influence decision making").

182. *Joint Statement on Data Scraping and the Protection of Privacy*, *supra* note 150.

183. *See* GDPR, *supra* note 137, art. 5(1)(a); Solove, *supra* note 176, at 167 (discussing various right to information in many privacy laws).

184. *See, e.g.*, Natasha Lomas, *Covert Data-Scraping on Watch as EU DPA Lays Down 'Radical' GDPR Red-Line*, TECHCRUNCH (Mar. 30, 2019), https://techcrunch.com/2019/03/30/covert-

But even if such a general post were legally sufficient, it would seem to be practically useless since most people would not know which websites are scraping their data. Additionally, such a notice from one scraper would fail to provide the full story to individuals about how their data will be processed by a potential multitude of third-party scrapers.

4. Consent

In many cases, privacy laws require consent for the collection and use of personal data.[185] Some require express consent (opt-in), and others require implied consent (opt-out).[186] Scrapers, however, mostly do not operate with any form of consent from the scrapees.

In the United States, most federal privacy laws provide rights to opt out of certain data uses or to opt in to other data uses.[187] Most state consumer privacy laws provide opt-out rights for the sale or sharing of personal data and opt-in rights for the use of sensitive data.[188] Scraping renders opt-in and opt-out rights meaningless. Once data is in the hands of scrapers, individuals lose any ability to opt in or out.

In the EU, people have the right to withdraw their consent to the processing of their data.[189] Data subjects should be able to withdraw consent after their data is scraped, but it is hard to imagine how data subjects can meaningfully withdraw consent when they are often unaware of the scraping or who has scraped it.

5. Purpose Specification and Secondary Use Restrictions

Many privacy laws require purpose specification, which requires that data be used for purposes originally stated at the time the data is collected.[190]

---

data-scraping-on-watch-as-eu-dpa-lays-down-radical-gdpr-red-line/ [https://perma.cc/XF6L-WBWK]; PrivSec Rep., *Rethinking 'Disproportionate Effort' Exemption Under GDPR for Web-Scraping*, GRC WORLD FS. (May 25, 2020), https://www.grcworldforums.com/gdpr/rethinking-disproportionate-effort-exemption-under-gdpr-for-web-scraping/344.article [https://perma.cc/LA56-Y8NK].

185. Daniel J. Solove, *Murky Consent: An Approach to the Fictions of Consent in Privacy Law*, 104 B.U. L. REV. 593, 596–97 (2024).

186. Laws often have heightened requirements for sensitive data; even in U.S. state privacy laws, which generally rely on opt-out consent, sensitive data requires opt-in consent. *See* Solove, *supra* note 140, at 1097.

187. *See, e.g.*, Controlling the Assault of Non-Solicited Pornography and Marketing (CAN-SPAM) Act, 15 U.S.C. § 7704(a)(3) (opt-out right for receipt of unsolicited commercial emails); Telephone Consumer Protection Act, 47 U.S.C. § 227 (opt-out right for telemarketing); Children's Online Privacy Protection Act, 15 U.S.C. § 6502(b) (opt-in for the collection and processing of children's data); Video Privacy Protection Act, 18 U.S.C. § 2710(2)(B) (opt-in); 18 U.S.C. § 2710(2)(D) (opt-out); Cable Communications Policy Act, 47 U.S.C. § 551(c)(1) (opt-in); *id.* § 551(c)(2) (opt-out).

188. DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 186–90 (7th ed. 2024).

189. GDPR, *supra* note 137, art. 7(3). For an extensive background about the right to withdraw consent, see generally Marcu Florea, *Withdrawal of Consent for Processing Personal Data in Biomedical Research*, 13 INT'L DATA PRIV. L. 107 (2023).

190. The principle of purpose specification is one of the original eight principles of the OECD Privacy Guidelines of 1980, which have been tremendously influential in shaping privacy laws around

Subsequent use for unrelated purposes requires consent, unless an exception applies.[191] As explained by the UK ICO, specifying a purpose in advance helps data collectors avoid "function creep" and is fundamental in building the trust necessary for safe and sustainable data processing.[192] A related principle is the restriction on secondary uses of data that are unrelated to the original purpose of collection. This principle is sometimes referred to as the "use limitation" principle.[193]

Data privacy rules around the world require that entities specify their purposes prior to the collection of personal data and use data only for these purposes. For example, the GDPR provides that personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes."[194] Data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."[195] Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) has a principle that restricts the use or disclosure of personal information for purposes beyond the original purpose without the individual's consent.[196] The Virginia Consumer Data Protection Act (VCDPA) states a controller cannot process personal data for purposes inconsistent with the disclosed purpose unless the controller obtains consent.[197]

In stark contradiction to the purpose specification principle, scraping involves indiscriminate data collection for unspecified purposes. Most of the purposes of scraped data are unrelated secondary uses of data, violating the use limitation principle.

---

the world. ORG. FOR ECON. COOP. & DEV. (OECD), GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA 3 (1980).

191. *Principle (b): Purpose Limitation*, INFO. COMM'R'S OFF., https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/the-principles/purpose-limitation/ [https://perma.cc/3YJW-HVQ7].

192. *Id.*

193. OECD, *supra* note 190.

194. GDPR, *supra* note 137, art. 5.1(b).

195. *Id.* art. 5(1)(c).

196. Personal Information Protection and Electronic Documents Act (PIPEDA), S.C. 2000, c 5 (Can.) ("Personal information shall not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as required by law. Personal information shall be retained only as long as necessary for the fulfilment of those purposes.").

197. Consumer Data Protection Act, VA. CODE ANN. § 59.1-578(A)(2) (2025) ("Except as otherwise provided in this chapter, not process personal data for purposes that are neither reasonably necessary to nor compatible with the disclosed purposes for which such personal data is processed, as disclosed to the consumer, unless the controller obtains the consumer's consent.").

6. Data Minimization

Another central tenet of data privacy protection is to collect and use only the data necessary for a specific legitimate purpose. In law, this idea is referred to as the principle of "data minimization."[198]

In the United States, several federal laws include data minimization provisions.[199] For example, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) requires reasonable efforts to limit the use or disclosure of protected health information to the minimum necessary to accomplish the intended purpose.[200] Under the Privacy Act, federal agencies must ensure that personal data is relevant and necessary to accomplish the agency's purpose.[201] The California Consumer Privacy Act (CCPA) requires that the collection, use, retention, and sharing of a consumer's personal information shall be reasonably necessary and proportionate to its original purpose and not further processed in a way that is incompatible with that purpose.[202] The GDPR establishes a principle of data minimization, requiring that personal data be adequate, relevant, and necessary to the purpose for which it is processed.[203]

To further the principle of data minimization, many privacy laws impose data-retention limitations to ensure that data is not used for longer than necessary. For example, in the United States, the Cable Communications Policy Act requires cable operators to destroy data when it is no longer necessary for its intended purpose.[204] The Video Privacy Protection Act (VPPA) requires data to be destroyed no later than a year from when the data is no longer necessary for its intended purpose.[205] Under the GDPR, personal data cannot be retained for

198. Lauren Bass, *The Concealed Cost of Convenience: Protecting Personal Data Privacy in the Age of Alexa*, 30 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 261, 286 (2019).

199. *See, e.g.*, 45 C.F.R. § 164.502(b) (2024).

200. *Id.* ("When using or disclosing protected health information or when requesting protected health information from another covered entity or business associate, a covered entity or business associate must make reasonable efforts to limit protected health information to the minimum necessary to accomplish the intended purpose of the use, disclosure, or request.").

201. 5 U.S.C. § 552a(e)(1) (requiring that agencies with a system of records shall "maintain in its records only such information about an individual as is relevant and necessary to accomplish a purpose of the agency required to be accomplished by statute or by executive order of the President").

202. CAL. CIV. CODE § 1798.100(c) (West 2023) ("A business' collection, use, retention, and sharing of a consumer's personal information shall be reasonably necessary and proportionate to achieve the purposes for which the personal information was collected or processed, or for another disclosed purpose that is compatible with the context in which the personal information was collected, and not further processed in a manner that is incompatible with those purposes.").

203. GDPR, *supra* note 137, art. 5(c) ("[A]dequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation').").

204. 47 U.S.C. § 551(e) ("A cable operator shall destroy personally identifiable information if the information is no longer necessary for the purpose for which it was collected and there are no pending requests or orders for access to such information under subsection (d) or pursuant to a court order.").

205. 18 U.S.C. § 2710(e) ("A person subject to this section shall destroy personally identifiable information as soon as practicable, but no later than one year from the date the information is no longer necessary for the purpose for which it was collected and there are no pending requests or orders for access to such information . . . or pursuant to a court order.").

longer than necessary, unless for a public interest, scientific interest, historical research, or statistical purpose.[206]

As with other privacy principles, data retention limitations are completely thwarted by scraping, as it involves the collection and retention of personal data without any restriction or time duration. It is the antithesis of data minimization.

7. Onward Transfer

The privacy principle of onward transfer, which is embodied in the GDPR and nearly all U.S. state consumer privacy laws (as well as many U.S. federal privacy laws), requires contracts and controls when transferring data to third parties (and other parties further downstream).[207] Onward transfer safeguards ensure that people's expectations about data use and protections are not thwarted whenever data is transferred to other entities. When people share their personal data, they consider the identity of the data recipient itself as well as the real and imagined identities of audience members in forming their privacy expectations and disclosure behaviors.[208]

Many U.S. and international privacy laws impose significant obligations on the recipients of personal data when transferred. These obligations typically consist of performing due diligence in selecting vendors, including sufficient provisions in contracts with vendors to ensure that data is protected, and monitoring vendors for compliance.[209] Under the GDPR, when selecting processors, controllers must make sure that they provide "sufficient guarantees" of their ability to comply with the GDPR.[210] In the United States, the FTC has interpreted the failure to vet processors as a violation of the FTC Act.[211]

Many laws also require contracts to ensure that the recipient of the data adequately protects a data subject's privacy and secures the data. For example,

206. GDPR, *supra* note 137, art. 5.1(e) ("[K]ept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes . . . subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation').").

207. *See generally* GDPR, *supra* note 137, art. 45; Woodrow Hartzog, *Chain-Link Confidentiality*, 46 GA. L. REV. 657 (2012).

208. *See generally* Alice E. Marwick & Danah Boyd, *I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience*, 13 NEW MEDIA & SOC'Y 114 (2011); Patrick McCole, Elaine Ramsey & John Williams, *Trust Considerations on Attitudes Towards Online Purchasing: The Moderating Effect of Privacy and Security Concerns*, 63 J. BUS. RSCH. 1018 (2010).

209. *See generally Contracts*, INFO. COMM'R'S OFF., https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/contracts/ [https://perma.cc/6TXV-2NMA]; *see also* Daniel J. Solove & Woodrow Hartzog, *The FTC and Privacy and Security Duties for the Cloud*, 13 BNA PRIV. & SEC. L. REP. 577 (2014).

210. GDPR, *supra* note 137, art. 28.1.

211. *See* GMR Transcription Servs., Inc., No. 122-3095, 2015-1 Trade Cases P 17070 (F.T.C. Aug. 14, 2014).

the GDPR requires a contract between the controller and the processor, and it sets forth a series of requirements for these contracts.[212] In the United States, HIPAA requires "business associate" agreements between covered entities (akin to controllers) and business associates (akin to processors) and specifies a number of protections that must be in these contracts.[213] The FTC has determined that the failure to have adequate contracts with processors constitutes an unfair and deceptive trade practice, though the FTC has not specified in detail the requirements of such contracts.[214] Many state consumer privacy laws require contracts with the recipients of data transfers that ensure that the data retains protection.[215]

These protections exist to ensure that the law's protections follow the data as it is transferred from one entity to the next. Because data frequently flows to different organizations, onward transfer requirements ensure that the law's protections are not lost.

Scraping renders onward transfer requirements meaningless. It allows third parties to take data without any contract, restrictions, or consent. Any representations made by companies in contracts or in the design of the technology itself no longer apply.[216] The parties entrusted with people's data lose the ability to enforce promises made or preferences revealed within the context of that information relationship. Scrapers are often not vetted, contracted with, or monitored. Thus, scraping creates two classes of third-party recipients of data: (1) third parties who contract with organizations to obtain personal data and must protect the data similarly to how the organization that collected it protects it, and (2) scrapers who can evade any responsibility at all.

8. Data Security

Scraping also contravenes the principle of data security. According to this principle, organizations must ensure that data is "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures."[217] This

212. GDPR, *supra* note 137, art. 28.

213. Health Insurance Portability and Accountability Act (HIPAA), 45 C.F.R. §§ 164.502(e), 164.504(e) (2024).

214. GMR Transcription Servs., No. 122-3095, 2015-1 Trade Cases P 17070 (F.T.C. Aug. 14, 2014).

215. *See* SOLOVE & SCHWARTZ, *supra* note 188, at 186–90 (noting state laws that require vendor agreements).

216. *See generally* Woodrow Hartzog, *Website Design as Contract*, 60 AM. U. L. REV. 1635 (2011) (exploring how companies make representations in the terms of use and in the design of websites themselves).

217. GDPR, *supra* note 137, art. 5.1(f).

includes safeguarding personal data from a data breach. Organizations must establish protections to prevent hackers from improperly accessing the data.[218]

Scraping involves third parties just grabbing the data. Data security is meaningless if any scraper can readily acquire the data.

* * *

It is not clear how scraping can be performed in a privacy-friendly way. The fundamental principles of privacy and the building blocks of most privacy laws—obtaining consent, having specific purposes of use, minimizing the collection and storage of data, providing individuals with rights over their data, and protecting data security—are in dramatic conflict with scraping. There is no aspect of scraping that is consistent with the FIPPs. The very model most privacy laws are founded upon is incompatible with scraping.

B. Scraping and Publicly Available Information

The most common defense of scraping is that it involves publicly available data on the internet. The notorious scraper Clearview AI defends its scraping as "[o]nly [s]earching [p]ublicly [a]vailable [d]ata from the [i]nternet."[219] OpenAI defends itself by claiming that the data it scrapes is publicly available.[220] When it scraped LinkedIn profile data, hiQ Labs claimed LinkedIn users lacked any privacy interest in their data because they made it publicly available.[221]

We contend that the argument that there is no privacy interest in publicly available information is not only incoherent but also normatively and legally wrong.

---

218. *See generally* DANIEL J. SOLOVE & WOODROW HARTZOG, BREACHED! WHY DATA SECURITY LAW FAILS AND HOW TO FIX IT 190–98 (2022).

219. *Debunking the Three Biggest Myths About Clearview AI*, CLEARVIEW AI: BLOG (June 21, 2023), https://www.clearview.ai/post/debunking-the-three-biggest-myths-about-clearview-ai [https://perma.cc/9FWA-HJN9] ("Clearview AI Is Only Searching Publicly Available Data from the [i]nternet. Clearview AI does not have the capability to access your private data. The company's algorithm is designed to only search through publicly available images on the internet. When Clearview AI 'scrapes' data, it is collecting information that any internet user could technically access. It does not include any content that would require a password or special access to view, such as private social media accounts or secure databases."). *See generally* HILL, *supra* note 32; Hoan Ton-That, *The Modern Public Square: The Free Flow of Information in the Age of Artificial Intelligence*, CLEARVIEW AI: BLOG (June 14, 2022), https://www.clearview.ai/post/the-modern-public-square-the-free-flow-of-information-in-the-age-of-artificial-intelligence [https://perma.cc/VL3K-HCEE] ("Clearview AI doesn't search for or retrieve private information, like that from your camera roll, or private social media -- but only publicly available information you would see by using Google or any other search engines.").

220. *How ChatGPT and Our Language Models Are Developed*, OPENAI, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed [https://perma.cc/ARS8-DRAD] ("We use training information lawfully."). OpenAI's language regarding lawful use of their training information has changed since we began writing this Article. They have since removed language stating that their "use of training information is not meant to negatively impact individuals, and the primary sources of this training information are already publicly available." *Id.* [https://perma.cc/7JMQ-2NJ2].

221. hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1188 (9th Cir. 2022).

1. Publicly Available Information: An Incoherent Concept

Far too often, claims about "publicly available information" are made broadly without properly considering what "public" actually means.[222] Justifying scraping data because it is "public" information is woefully inadequate because "public" can be understood in several different ways depending on the context.

For example, the standard dictionary definition of "public" is deceptively simple. As an adjective, public is defined as: "1. Of, relating to, or involving an entire community, state, or country. 2. Open or available for all to use, share, or enjoy. 3. (Of a company) having shares that are available on an open market."[223] The dictionary fails to indicate which groups of people are included in "all." People in a pharmacy might be able to catch a fleeting glimpse of the medicines that a person selects from the aisles. Yet, as a practical and normative matter, that same piece of information is hard to categorize as available for "all to share, use, or enjoy." In practice, virtually everyone on Earth is denied access to someone's fleeting exposure unless they were both present at the scene and looking at the person at issue during their brief disclosure. Additionally, even if the information is observable, it does not automatically follow that it is socially acceptable for all to share or use.[224]

As a noun, the term public is defined as: "1. The people of a country or community as a whole" or "2. A place open or visible to the public."[225] The dictionary is not clear about the meaning of the words "open or visible" in the definition of public. They could mean "structurally exposed," such as an open door that enables onlookers. They could also mean "normatively inclusive," like expressive works in the public domain, or a legally permissible physical presence, such as diners being invited to eat in restaurants. These definitions show why "public" is a complex construct.[226]

As one of us has argued, there are three different conceptions of what "public" or "publicly available" information could mean.[227] First, it could be understood as a *descriptive* concept, with contextual factors shaping the contours

222. *See generally* Woodrow Hartzog, *The Public Information Fallacy*, 99 B.U. L. REV. 459 (2019); Joel R. Reidenberg, *Privacy in Public*, 69 U. MIA. L. REV. 141 (2014).

223. *Public*, BLACK'S LAW DICTIONARY (12th ed. 2024).

224. Order Denying Clearview AI's Motion to Dismiss, ACLU v. Clearview AI, Case No. 20 CH 4353, at 11 ("The fact that something has been made public does not mean anyone can do with it as they please.").

225. *Public*, BLACK'S LAW DICTIONARY (12th ed. 2024). Merriam-Webster's definition demonstrates the many different ways "public" can be defined, with significant differences between the conceptualizations: "1. a: exposed to general view: open b: well-known, prominent c: perceptible, material 2. a: of, relating to, or affecting all the people or the whole area of a nation or state . . . b: of or relating to a government c: of, relating to, or being in the service of the community or nation 3. a: of or relating to people in general: universal b: general, popular 4.: of or relating to business or community interests as opposed to private affairs: social 5.: devoted to the general or national welfare: humanitarian 6. a: accessible to or shared by all members of the community." *Public*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/public [https://perma.cc/M68C-9UGS].

226. Hartzog, *supra* note 222, at 473, 507, 514.

227. *Id.* at 494–96.

of the notion, such as who received the information, how widely it was disseminated, where it was located, how long it was available, and the foreseeable extent of exposure.[228] Descriptive notions of "public" information are often nuanced and tailored. While some people descriptively equate notions of "public" with accessibility, it can also connote information that is "widely known." For example, it is probably "public" knowledge that Taylor Swift recently concluded the successful Eras Tour.[229] Other people might describe public information as whatever society expresses a collective interest in, such as celebrity gossip.[230]

Second, people might define "public" information as a *designated* concept. Think of this as an express, official designation or category created by a relevant authority that indicates the information is for general use by anyone or that collecting information about specific people or acts is authorized. The most common example of designated public information is a "public record" or "open record."[231] These records, when released, are designated as "public" through legislation. The designation of something "as a public record carries with it the imprimatur of government authorization as well as a signal to society that these documents are intended to be collected, used, and shared."[232]

Third, "public" can be conceptualized by what it is *not*, i.e., shorthand for anything that is normatively or legally "*not private*."[233] The problem with the "not private" conceptualization of public information is illustrated by its use in privacy rules. This definition begs the question of the privacy interest involved. When people use the negative conceptualization of public information to justify the collection and use of information, they are merely assuming (without additional justification) the absence of a privacy interest.[234]

Because privacy expectations depend on context, and since there are so many conflicting ways to conceptualize "public" information, determining the privacy interests and expectations in provisionally viewable information shared online requires a deeper contextual inquiry into the parties involved, the nature of their relationship, the nature of the information revealed, the terms of disclosure, and the risks of exposure.[235] Labeling data as "public" or "publicly

228. *Id.*

229. Ethan Millman, *Taylor Swift's Eras Tour Is the Highest-Grossing of All Time and First-Ever to Hit $1 Billion*, ROLLING STONE (Dec. 8, 2023), https://www.rollingstone.com/music/music-news/taylor-swift-eras-tour-highest-grossing-all-time-1-billion-1234921647/ [https://perma.cc/PN29-VWQN].

230. Hartzog, *supra* note 222, at 467.

231. *See generally* Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137 (2002).

232. Hartzog, *supra* note 222, at 509.

233. *Id.* at 467–68, 496, 507–08, 511–12.

234. *Id.* at 468, 508.

235. *See, e.g.*, NISSENBAUM, *supra* note 179, at 155 (drawing upon philosophy and social science in developing a theory of privacy as contextual integrity, which holds that privacy violations occur when context-relative "informational norms" are not respected when sharing information). Nissenbaum

available" does not establish that people have voluntarily waived all expectations of privacy.[236]

As one of us has previously argued, "privacy" means many different things, and privacy protection has many different dimensions.[237] Far too often, privacy is conceptualized as merely involving the safeguarding of hidden secrets.[238] This crabbed conception of privacy overlooks not only people's privacy expectations, but also their desires for how their data should be protected, as well as how the law actually protects privacy. Although it persists in many places, the notion that privacy is only about hidden secrets is quite antiquated. More modern conceptions of privacy involve individual control over information as well as measures to bring the collection, use, and transfer of personal information under control.

2. Expectations of Privacy in Publicly Available Information

The notion that publicly available information cannot implicate privacy interests is descriptively incorrect. Social science literature on privacy paints a much more complex picture of the relationship between the concept of "public" information and privacy expectations.[239] People often do not intend for the provisionally viewable information they post online to be shared universally. Just because people make their information available at a certain point in time for a certain use by an intended audience does not mean they expect this information will be made available at other times and for other uses. Research shows that people value the ability to delete their data, provide informed consent for data practices, and opt out of data collection at any time.[240] For example, in some studies, participants who expressed a desire to be able to share their data on social media were also reluctant to allow others to download or modify their data.[241]

---

writes: "[W]hether a particular action is determined a violation of privacy is a function of several variables, including the nature of the situation, or context; the nature of the information in relation to that context; the roles of agents receiving information; their relationships to information subjects; on what terms the information is shared by the subject; and the terms of further dissemination." Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 155 (2004).

236. Memorandum Opinion and Order at 11, ACLU v. Clearview AI, Inc., No. 20 CH 4353 (Ill. Cir. Ct. Aug. 27, 2021) ("Clearview emphasizes that the photos from which they make faceprints are publicly available and that Plaintiffs have no 'expectation of privacy' in them.").

237. DANIEL J. SOLOVE, UNDERSTANDING PRIVACY 9 (2008); Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 480–82 (2006) [hereinafter Solove, *A Taxonomy of Privacy*].

238. Daniel J. Solove, *"I've Got Nothing to Hide" and Other Misunderstandings of Privacy*, 44 SAN DIEGO L. REV. 745, 748–50 (2007).

239. *See generally* Lior Jacob Strahilevitz, *A Social Networks Theory of Privacy*, 72 U. CHI. L. REV. 919 (2005) (developing a theory of privacy based on people's expectations of how far they expect information disclosure to travel through their social networks).

240. Naeini et al., *supra* note 178, at 399–410 (reporting on a "1,007-participant vignette study focusing on privacy expectations and preferences as they pertain to a set of 380 [Internet of Things] data collection and use scenarios").

241. Krishanu Dey & Parikshit Mondal, *Privacy Awareness Among the Academic Social Network Users*, 2019 LIBR. PHIL. & PRAC. 1, 6 ("[A]mong all the respondents 44% wanted people to

Additionally, public demand for design features such as delete buttons, edit buttons, and news feeds that display only recent posts demonstrates that even "public" disclosures are intended to be limited in practice. On social platforms, people update their profiles and otherwise present a version of themselves that is "here and now." They typically revise these profiles and sometimes delete them.

Scraping of publicly available data directly threatens the obscurity of people's data, which is one of the most common but underappreciated notions of privacy.[242] People's expectations of privacy and the degree to which the individuals seek to control their "public" disclosures are partially based on how difficult it is for others to find, observe, or preserve their personal information.[243] Most of the data about our lives is seen by and shared with some, but not all.

Consider behavior in public spaces. People navigate their daily lives in zones of relative obscurity. They may sit next to each other on buses and in restaurants and forget each other the moment they leave. They hear gossip in the seat next to them but tune it out. They take the trash out in their pajamas because the odds that someone will see them during their short period of exposure are very low. This is privacy through obscurity. However, if people were told that cameras in public places recorded their activities and conversations and that such information would be used to gain insights about them, their privacy expectations would change, and they would behave differently. Similarly, when

see and share their research data but did not allow anyone to download or edit or modify or tamper with their reports and data.").

242. Woodrow Hartzog & Evan Selinger, *Surveillance as Loss of Obscurity*, 72 WASH. & LEE L. REV. 1343, 1356 (2015) (explaining the etymology of obscurity); Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity*, 101 CALIF. L. REV. 1, 24 (2013) (critiquing the idea that information is either disseminated globally or completely secret); Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. L. REV. 385, 387 (2013) (noting that the modern understanding of privacy has created a list of unaddressed problems); EVAN SELINGER & WOODROW HARTZOG, *Obscurity and Privacy*, *in* A COMPANION TO THE PHILOSOPHY OF TECHNOLOGY 119, 119 (Joseph Pitt & Ashley Shew eds., 2017) ("Obscurity is the idea that information is safe—at least to some degree—when it is hard to obtain or understand."); Woodrow Hartzog & Evan Selinger, *Obscurity: A Better Way to Think About Your Data than 'Privacy*,*'* ATLANTIC (Jan. 17, 2013), https://www.theatlantic.com/technology/archive/2013/01/obscurity-a-better-way-to-think-about-your-data-than-privacy/267283/ [https://perma.cc/4Q8Z-P73D] (explaining that obscurity is a better way to think of privacy than secrecy or confidentiality when sharing online); Evan Selinger & Woodrow Hartzog, *Why You Have the Right to Obscurity*, CHRISTIAN SCI. MONITOR (Apr. 15, 2015), https://www.csmonitor.com/World/Passcode/Passcode-Voices/2015/0415/Why-you-have-the-right-to-obscurity [https://perma.cc/GP3T-VR39] (describing obscurity as an important concept for protection of personal privacy); Evan Selinger & Woodrow Hartzog, Opinion, *Google Can't Forget You, but It Should Make You Hard to Find*, WIRED (May 20, 2014), https://www.wired.com/2014/05/google-cant-forget-you-but-it-should-make-you-hard-to-find/ [https://perma.cc/2SR5-5HK3] ("This debate is not and should not be about forgetting or disappearing in the traditional sense. Instead, let's recognize that the talk about forgetting and disappearing is really concern about the concept of obscurity in the protection of our personal information.").

243. SELINGER & HARTZOG, *supra* note 242; Solove, *supra* note 231, at 1173 ("Privacy can be violated by altering levels of accessibility, by taking obscure facts and making them widely accessible."); Solove, *A Taxonomy of Privacy*, *supra* note 237, at 538–40 (noting that privacy can be violated by increasing the accessibility of data).

people share data online, they do so for specific purposes and have specific expectations of use.

Scraping violates people's expectations about the risks of sharing information and places people in an impossible position: to assume that everything they share in a publicly available way with some is, or could be, fair game for exploitation by all. People simply are not capable of contemplating this sort of all-encompassing and hypothetical risk that every choice they make on the internet could be collected, analyzed, and later used against them.

3. Privacy Law and Publicly Available Information

Even when lawmakers attempt to be specific about public data, they act inconsistently.[244] Some privacy laws exempt publicly available information, but others do not.[245] For example, Canada's PIPEDA excludes publicly available information.[246] The GDPR does not contain an exception on all publicly available information, but it does have a limited exemption from sensitive data rules for personal data "manifestly made public by the data subject."[247] Though it's not always clear when this exception applies.[248] U.S. federal privacy laws are inconsistent on the issue. The Fair Credit Reporting Act does not exclude publicly available information.[249] But the Gramm-Leach-Bliley Act of 1999 (GLBA) defines the personal data it regulates as "nonpublic personal information," which does not include publicly available information.[250]

Many U.S. state consumer privacy laws exempt publicly available data, though their definitions of such data vary, as does the scope of what is excluded.[251] While the CCPA exempts publicly available data, it specifies that "publicly available" does not include biometric information that a business

244. *See, e.g.*, Hartzog, *supra* note 222, at 466, 479.

245. David A. Zetoony, *What Is 'Publicly Available Information' Under the State Privacy Laws?*, NAT'L L. REV. (Sept. 13, 2023), https://www.natlawreview.com/article/what-publicly-available-information-under-state-privacy-laws [https://perma.cc/JM47-JMU2].

246. *See* PIPEDA, S.C. 2000, c 5, § 7(1)(d) (Can.) (allowing collection of publicly available personal information without knowledge and consent); *id.* § 3(h.1) (allowing collection of publicly available personal information without knowledge and consent).

247. GDPR, *supra* note 137, art. 9.2(e).

248. *See, e.g.*, Edward S. Dove & Jiahong Chen, *What Does It Mean for a Data Subject to Make Their Personal Data 'Manifestly Public'? An Analysis of GDPR Article 9(2)(e)*, 11 INT'L DATA PRIV. L. 107, 108 (2021) ("What makes this provision even more special is the fact that EU data protection law does not generally make a substantial distinction between personal data in a private space and in a public one. . . . Looking to guidance from European regulatory authorities as to the meaning of this phrase, one is struck by the relative paucity of information.").

249. Fair Credit Reporting Act, 15 U.S.C. § 1681–1681x.

250. 15 U.S.C. § 6809(4)(A)–(B) ("The term 'nonpublic personal information' . . . does not include publicly available information.").

251. *See, e.g.*, California Consumer Privacy Act (CCPA), CAL. CIV. CODE § 1798.140(v)(1) (West 2021); UTAH CODE ANN. § 13-61-101(29)(b) (West 2022).

collects without a consumer's knowledge.[252] Other laws protecting biometric information do not exempt publicly available data.[253]

California legislators also recently introduced a new bill that would explicitly remove "[i]nformation gathered from internet websites using automated mass data extraction techniques" from the CCPA's public information exemption, bringing scraped data back within the statute's scope of protection.[254] The language of this amendment is a great model for other lawmakers looking to protect publicly available information from scraping.

Not all states have the same definition of "publicly available." Some states have a narrow definition, such as Colorado, which defines data as publicly available only if it is in government records or made available to the general public by the individual.[255] Connecticut's definition is similar to Colorado's but also includes data disseminated by the media.[256] According to privacy lawyer David Zetoony, "[M]ost data privacy statutes would not classify all internet-accessible information as being 'publicly available.'"[257]

Turning to judicial precedent, courts are quite inconsistent on whether to recognize a privacy interest in publicly available data. Although many courts have held that data exposed to the public is no longer private, other courts have recognized privacy interests in such data. In *United States Department of Justice v. Reporter's Committee for Freedom of the Press*, the U.S. Supreme Court held that there was a privacy interest in publicly available personal information.[258] Reporters sought to obtain FBI compilations of criminal history data on individuals under the Freedom of Information Act (FOIA). However, the Court concluded that this data fell under the privacy exemption to FOIA. The reporters argued that because the records involved publicly available information, there was no privacy interest in them.[259] The Court rejected this argument, reasoning that "there is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives, and local police stations throughout the country and a computerized summary located in a single clearinghouse of information."[260] The Court's holding is relevant to scraping for two reasons. First, the Court recognized that the public availability of personal

252. CAL. CIV. CODE § 1798.140(2)(B)(ii).

253. Biometric Information Privacy Act, 740 ILL. COMP. STAT. 14/15 (2024); My Health My Data Act, WASH. REV. CODE 19.373.010(22) (2023).

254. California Consumer Privacy Act of 2018, AB-1008, 2023 Gen. Assemb., Reg. Sess. (Cal. 2024) ("This bill would specify that 'publicly available' does not include information gathered from internet websites using automated mass data extraction techniques and would specify that personal information can exist in various formats."); CAL. CIV. CODE § 1798.140 (as amended on Nov. 18, 2024) (omitting the previous addition).

255. COLO. REV. STAT. § 6-1-1303(17)(b) (2023).

256. 2022 Conn. Acts 15 § 1(25) (Reg. Sess.).

257. Zetoony, *supra* note 245.

258. 489 U.S. 749, 780 (1989).

259. *Id.* at 762–63.

260. *Id.* at 764.

data does not automatically extinguish a privacy interest in the data.[261] Second, the Court noted that large aggregations of publicly available data pose privacy concerns—which is exactly the kind of data gathering involved in scraping.

The idea that there is no privacy in publicly available information is rooted in the notion that people have either waived or at least cannot reasonably expect privacy in information freely viewable by others. But the Supreme Court has rejected that notion. In *Carpenter v. United States*, the Supreme Court held: "A person does not surrender all Fourth Amendment protection by venturing into the public sphere."[262] Before *Carpenter*, the Court's Fourth Amendment jurisprudence largely maintained that anything observable in a public place was not private.[263] But *Carpenter* signaled a change in the Court's thinking, and it represents a more nuanced view of the issue of privacy in public.

Although many cases involving privacy torts fail to find a privacy interest in publicly available information, there are many notable exceptions.[264] For example, in *Nader v. General Motors Corp.*, the court held that "overzealous" observation of a person in pubic can constitute a violation of privacy.[265] As another court stated: "Traditionally, watching or observing a person in a public place is not an intrusion upon one's privacy. However, Georgia courts have held that surveillance of an individual on public thoroughfares, where such surveillance aims to frighten or torment a person, is an unreasonable intrusion upon a person's privacy."[266]

These cases demonstrate that it is far too simplistic to recognize a general rule that publicly available information is not private. Instead, the law's protections involve far more factors than public availability. The law is far more nuanced and contextual than most scrapers are presuming. Currently, scrapers wrongly view publicly available data as free for the taking. But the reality is far more complicated. Scrapers may escape some privacy laws, but not all.

---

261. *Id.* at 763–67 & n.15 ("The common law recognized that one did not necessarily forfeit a privacy interest in matters made part of the public record, albeit the privacy interest was diminished and another who obtained the facts from the public record might be privileged to publish it." (first citing Cox Broad. Corp. v. Cohn, 420 U.S. 469, 494–95 (1975); next citing RESTATEMENT (SECOND) OF TORTS § 652D (AM. L. INST. 1977); and then citing W. KEETON, D. DOBBS, R. KEETON, & D. OWENS, PROSSER & KEETON ON LAW OF TORTS § 117, 859 (5th ed. 1984))).

262. 585 U.S. 296, 297 (2018) ("Given the unique nature of cell phone location records, the fact that the information is held by a third party does not by itself overcome the user's claim to Fourth Amendment protection. Whether the Government employs its own surveillance technology as in *Jones* or leverages the technology of a wireless carrier, we hold that an individual maintains a legitimate expectation of privacy in the record of his physical movements as captured through CSLI.").

263. *See generally* United States v. Knotts, 460 U.S. 276 (1983) (holding that there was no reasonable expectation of privacy when a tracking device monitored movement in public); Florida v. Riley, 488 U.S. 445 (1989) (holding that there was no expectation of privacy in anything that can be viewed on one's property by police officers in a helicopter flying in legal airspace).

264. Solove, *supra* note 231, at 1175.

265. 255 N.E.2d 765, 771 (N.Y. 1970).

266. Anderson v. Mergenhagen, 642 S.E.2d 105, 110 (Ga. Ct. App. 2007) (citing Summers v. Bailey, 55 F.3d 1564, 1566 (11th Cir. 1995)).

Ultimately, privacy law cannot achieve its goals if it fails to protect publicly available personal data. In the modern world, an unprecedented amount of personal data is posted online; much of it is posted by individuals themselves, but a great deal of it is also posted by schools, employers, journalists, and other organizations. If privacy law is to remain relevant today, then it must protect publicly available information. Too much personal data is publicly available and excluding it from privacy law would leave too many gaping holes in the law's protection.

Although scraped data can fall outside of the protection of certain privacy laws with broad exemptions for publicly available data, it is difficult to square this result with any coherent account of the principles that these laws strive to achieve.

C. The Need for a Coherent Theory of Scraping and Privacy

Trying to reconcile scraping with the fragmented landscape of privacy law will result in a jumbled mess of precedent and inconsistent outcomes that will not lead to coherent policy. The best way forward is to start by developing a coherent theory of scraping and privacy to guide policymaking. Such a theory currently does not exist.

The litigation over scraping has failed to provide consistent answers or lead to a desirable regulatory regime.[267] Although in many instances an ad hoc common-law style approach is quite effective for developing law and policy, we doubt that such an approach, in the absence of a coherent overarching theory, will work well to balance scraping and privacy. Given the prevalence of scraping and the profound stakes involved, we contend that developing an overarching theory is the most practical and sound way forward.

Moreover, many of the legal and technological mechanisms employed in the Scraping Wars fail to involve the individuals whose data is being fought over and whose privacy is at stake. Individuals neither control the robots.txt files for websites containing their data nor set the terms of service of platforms. Many causes of action are available only to the website operators, so individuals depend upon the sites to detect scraping, police against scraping, issue cease-and-desist letters to scrapers, or bring litigation. A more coherent and comprehensive approach to scraping is necessary to account for the interests of all stakeholders.

267. *See* Sobel, *supra* note 72 (arguing that no common law torts can adequately address scraping and proposing a new tort of bad faith breach of terms of service). Additionally, smaller websites might lack the resources to litigate against scrapers. *See* Gold & Latonero, *supra* note 21, at 298–99.

III.
RECONCILING SCRAPING AND PRIVACY

Thus far, we have argued that scraping has long evaded a full reckoning with privacy law despite violating nearly all the core principles that animate it. Scraping and privacy law are incompatible; there must be a reconciliation. In this Part, we argue that the reconciliation is far more complicated than simply bringing scraping into the purview of privacy law. Both scraping and privacy law need a radical rethinking about what should be possible and why. We begin this Part by arguing how scraping should be conceptualized in terms of its privacy impact. Seen as part of the landscape of systemic mass data collection, use, and transfer, scraping is best understood as a form of surveillance as well as a data security violation. We then discuss why merely applying existing privacy laws to scraping would lead to undesirable consequences because such laws generally have significant shortcomings. Under many current privacy laws, existing infirmities with consent could lead to end-runs around any meaningful control over scraping. Under others, though, scraping might be practically impossible, leading to a de facto ban on scraping, which is also undesirable. We propose that scraping is best addressed by focusing on whether it is in the public interest.

A. A Theory of Surveillance and Security

1. Scraping as Surveillance

To conceptualize the scraping of personal data as surveillance is to understand the practice in its technical and functional sense: Scraping allows for the cheap, ongoing, mass collection and observation of people for exploitative purposes. Scraping today is ground zero for the practices that Shoshana Zuboff has famously termed "surveillance capitalism."[268] It is a mistake to view scraping only as an isolated action, with the risk assessed on a per-scrape basis. Rather, scraping should be viewed in the context of other data practices, the realities of its political and commercial incentives, and the likely downstream effects of data capture.

Surveillance is a broad concept capable of multiple meanings.[269] There is even an entire field devoted to the concept of surveillance studies.[270] The

268. *See generally* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (2019).

269. *See generally, e.g.*, OSCAR H. GANDY JR., THE PANOPTIC SORT: A POLITICAL ECONOMY OF PERSONAL INFORMATION (2d ed. 2021); GARY T. MARX, WINDOWS INTO THE SOUL: SURVEILLANCE AND SOCIETY IN AN AGE OF HIGH TECHNOLOGY (2016); James B. Rule, Douglas McAdam, Linda Stearns & David Uglow, *Documentary Identification and Mass Surveillance in the United States*, 31 SOC. PROBS. 222, 223 (1983) (defining surveillance as "any systematic attention to a person's life aimed at exerting influence over it").

270. *See generally* SURVEILLANCE STUDIES: A READER (Torin Monahan & David Murakami Wood eds., 2018).

definition that we think best fits a description of the reality of web scraping is from David Lyon, who defined surveillance as "the focused, systematic and routine attention to personal details for purposes of influence, management, protection or direction."[271] As Neil Richards has argued, Lyon's definition of surveillance is noteworthy because it focuses on learning about people for many different purposes (as opposed to aimless observation). Additionally, Lyon's definition highlights that surveillance is systematic, routine, and intentional rather than random, arbitrary, and haphazard.[272]

Many kinds of web scraping, such as the scraping of social media profiles, reflect all four aspects of Lyon's definition of surveillance. First, web scraping is focused on people's personal details. Companies need to scrape websites because they need human information *in context*, meaning information about how people look, how they move, how they react, and what they mean when they share information and express themselves online.[273] This allows certain AI systems to make predictions about people's lives and their future actions, generate text and images in response to queries, and more directly surveil individuals by using their face, gait, or even heartbeat as a beacon.[274]

Second, companies deploy scraping systemically and methodically to capture entire bodies of data to better train systems and ensure functionality of databases. For example, some facial recognition systems need to be able to recognize entire populations to be seen as effective, which requires systemic and holistic scraping.[275]

Third, web scraping has been completely routinized by companies developing AI.[276] Companies scrape hundreds of thousands of web pages in a very short time. Scraping vendors have popped up to aid in creating whole systems and programs for scraping webpages for companies.[277] Companies like Clearview AI collect billions of photos to power their databases through routines designed to cheaply and quickly scrape websites.[278]

---

271. DAVID LYON, SURVEILLANCE STUDIES: AN OVERVIEW 14 (2007).

272. Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1937 (2013).

273. For a general overview on how AI models need human data, see, for example, SAYASH KAPOOR & ARVIND NARAYANAN, AI SNAKE OIL: WHAT ARTIFICIAL INTELLIGENCE CAN DO, WHAT IT CAN'T, AND HOW TO TELL THE DIFFERENCE 115, 191 (2024).

274. *Id.*; *see also Types of Biometrics*, BIOMETRICS INST., https://www.biometricsinstitute.org/what-is-biometrics/types-of-biometrics/ [https://perma.cc/Q3JX-Q2AJ].

275. *See* HILL, *supra* note 32, at 33, 133.

276. *See, e.g.*, Ian Kerins, *Data for Price Intelligence: Lessons Learned Scraping 100 Billion Products Pages*, ZYTE (July 2, 2018), https://www.zyte.com/blog/price-intelligence-web-scraping-at-scale-100-billion-products/ [https://perma.cc/72GN-2UKN].

277. *Id.*

278. Louise Matsakis, *Scraping the Web Is a Powerful Tool. Clearview AI Abused It*, WIRED (Jan. 25, 2020), https://www.wired.com/story/clearview-ai-scraping-web/ [https://perma.cc/4ZRB-KY8X].

Finally, many companies scrape data from the web to influence people, manage them, protect them, or direct them.[279] While academics and journalists might scrape to gain knowledge, companies scrape the web to make money, which entails developing systems that can influence people's behavior by conveying information or making tasks easier or harder. Some companies scrape to gain a business advantage. Others scrape to convince advertisers of the ability to target consumers with the right message at the right time in the right place. Still others scrape to power literal surveillance systems, ostensibly to help law enforcement and other arms of government deter crime, find missing people, and protect the public. Criminals scrape that same data to bypass technical safeguards or engage in spear phishing so as to defraud, thieve, and hack, all as part of an endless game of cat and mouse.

To understand scraping as surveillance is to recognize that scraped data can, over time, give full pictures of people's lives, enable them to be recognized by their faces wherever they go, and expose them to harassment, impersonation, manipulation, and a myriad of other harms. Often the best time for the law to intervene is when the data is scraped rather than later on, when it is more difficult to corral various uses and the sharing of data.

Critics of anti-scraping frameworks might object to treating scraping as surveillance, because in the minds of many, scraping is functionally equivalent to people viewing and "cutting and pasting" information for themselves. For example, the Electronic Frontier Foundation argued that "[a]s a technical matter, web scraping is simply machine-automated web browsing, and accesses and records the same information, which a human visitor to the site might do manually."[280] But this objection ignores scraping's incredible affordances of scale.[281] The fact that scraping is so cheap, easy, and automatic makes it so different in power and risk from non-automated data collection that it is worthy of specific regulatory intervention and analysis. Manual data collection is too expensive and laborious for companies like Clearview AI to assemble a biometric database that works at scale. Scraping is not just "more" of an acceptable activity; it represents a difference in magnitude of risk so large that it constitutes a difference in kind.

Treating scraping like surveillance would have the important effect of tying scraping rules to the gradual recognition in surveillance law that sometimes individuals can and should have a reasonable expectation of privacy *in public* or with respect to *publicly available information*.[282]

279. *See*, *e.g.*, HILL, *supra* note 32, at 33, 79; *see also* LYON, *supra* note 271.

280. Camille Fischer & Andrew Crocker, *Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data*, ELEC. FRONTIER FOUND. (Sept. 10, 2019), https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data [https://perma.cc/AS69-5T7M].

281. *See, e.g.*, McKenna & Hartzog, *supra* note 51, at 2–4.

282. *See supra* Part II.B.

Zooming out, treating scraping like surveillance would also help reframe policies and public discourse that currently treat our raw human information and experiences as a free-for-all resource. Instead, information about our lives is inextricably tied to our dignity and well-being, and it should be worthy of protection based on this fact, not on whether it appears online or not.[283]

2. Protection from Scraping as Security

One of the oldest and least controversial information privacy rules is the duty of data processors to protect personal information from unauthorized access.[284] This duty is invoked in several different areas such as cybersecurity, data protection, anti-hacking safeguards, and trust/assurance compliance. The underlying premise is that certain actors will inevitably attempt to access data through wrongful means and that entities entrusted with that data are obligated to take reasonable steps to safeguard against those wrongful attempts. Colloquially, wrongful attempts to bypass safeguards to access data are called hacking. A successful hack results in a personal data breach. In this Section, we argue that the best way to understand data processors' obligations regarding scraping is through the lens of data security. In other words, sometimes scraping is a data breach that data collectors should foresee and take reasonable precautions against.

To understand protection from scraping as security is to recognize the stewardship obligations that entities take on when accepting, storing, and displaying people's data. Thinking of protections against scraping as security also properly recognizes the realistic differences in scale and power between viewing, preserving and manual access, and automation.

Security is often thought of as akin to locking data in a safe and keeping it hidden from malicious actors.[285] A common acronym used to define security is CIA: confidentiality, integrity, and accuracy. But security in a more modern understanding, at least as embodied in many data breach laws, involves improper access to data.[286] Improper access can occur even if data is publicly available and not confidential. Thus, the public availability of data does not obviate all security obligations.

Entities entrusted with people's personal data have a host of legal, organizational, and technical actions they can take to protect people's information from scrapers, actions that are similar to the safeguards used to prevent hackers from accessing personal data without authorization.[287] For

283. *See, e.g.*, JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2019); Julie E. Cohen, *The Biopolitical Public Domain: The Legal Construction of the Surveillance Economy*, 31 PHIL. TECH. 213, 214 (2018).

284. *See* SOLOVE & HARTZOG, *supra* note 218, at 41–42, 53.

285. *See id.* at 68–69, 186–87.

286. *See id.* at 42.

287. *Joint Statement on Data Scraping and the Protection of Privacy*, *supra* note 150, at 3–4.

example, in a joint statement on scraping, DPAs from around the world argued that companies should implement multi-layered technical and procedural controls to mitigate the risks of scraping.[288] The DPAs wrote that "websites should implement multi-layered technical and procedural controls to mitigate the risks. A combination of these controls should be used that is proportionate to the sensitivity of the information."[289] Some of these safeguards are similar to those frequently included in data security frameworks like designating a person or team to be accountable for protecting against scraping, monitoring for unusual activity that would indicate wrongful scraping and limiting access when it is detected, taking affirmative steps to detect and limit bots like implementing CAPTCHAs and blocking IP addresses, and taking or threatening to take appropriate legal action.[290]

Regulators could require such safeguards as part of regulatory duties to maintain reasonable data security. For example, in the United States, the FTC could demand these practices as part of their regulation of unfair and deceptive trade practices. States could also ensure that duties to protect against scraping are a part of their state data security and data breach notification rules.

Additionally, scraping could constitute a data breach under the Health Breach Notification Rule.[291] Under the Rule, a "breach of security" is defined as the "acquisition of [public health record identifiable health] information without the authorization of the individual."[292] In its enforcement of the Rule, the FTC has claimed that privacy violations are data breaches that should have been reported under the Rule. In two cases, the FTC claimed it was a reportable data breach when companies shared health data with third parties in violation of their privacy policies.[293] Failing to implement reasonable protections against scraping by third parties is tantamount to improperly sharing data with third parties. In fact, it is far worse, as even when data is improperly shared with third parties there is sometimes vetting of these third parties and a contractual agreement governing the third party's use of the data. Leaving the data out on the table to be gobbled up by any third party without oversight or an agreement is a far less safe and secure way to share data.

288. *Id.* at 3.

289. *Id.*

290. *Id.* The Italian DPA Garante issued similar guidelines. *See* Tommaso Ricci, *The Garante Issues Guidelines to Prevent AI Web Scraping*, GAMINGTECHLAW (June 3, 2024), https://www.gamingtechlaw.com/2024/06/garante-privacy-guidelines-web-scraping-artificial-intelligence-ai/ [https://perma.cc/5THS-LRDN].

291. 16 C.F.R. § 318.2 (2024).

292. *Id.*

293. *See generally* Stipulated Order for Permanent Injunction, Civil Penalty Judgment, and Other Relief, United States v. GoodRx Holdings, Inc., No. 3:23-cv-460 (N.D. Cal. Feb. 17, 2023); Stipulated Order for Permanent Injunction, Civil Penalty Judgment, and Other Relief, United States v. Easy Healthcare Corp., No. 1:23-cv-3107 (N.D. Ill. June 22, 2023).

B. The Difficulty of Bringing Scraping Under the Purview of Privacy Law

The tension between scraping and privacy cannot be satisfactorily resolved by anointing scraping or privacy as the winner. Allowing unfettered scraping would constitute an untenable threat to privacy; it involves a cascade of privacy violations on a grand scale. We contend that scraping does and should fall under many existing privacy laws—not just the GDPR but also several U.S. privacy laws. Yet, merely bringing scraping within the scope of existing privacy laws opens up a Pandora's box of problems. Existing privacy laws are not well tailored to regulate scraping.

Two potential pitfalls—inconsistency and complexity—exist when scraping is placed within the purview of privacy laws. Some privacy laws, such as those that require consent for data collection, might impose such cumbersome requirements that they effectively ban scraping. Other privacy laws will be far too loose and allow scraping to occur with just a few perfunctory extra steps. Additionally, the patchwork of different privacy laws in the United States will make it quite difficult for a scraper to navigate.

Under the EU's GDPR, scraping requires a lawful basis.[294] As discussed earlier in Part II.A, the two most common lawful bases advanced for scraping are individual consent or legitimate interests. In fact, the Dutch DPA has declared: "In practice, scraping by private organizations and private individuals is only possible on the basis of legitimate interest."[295] Regarding special categories of personal data (often called "sensitive data"), the European Data Protection Board (EDPB) has clarified that in addition to providing a legitimate interest, data processors must also identify an exemption to the ban on processing sensitive data, such as where "the data subject has manifestly made such data public."[296] The EDPB recognized that "where large amounts of personal data are collected via web scraping, a case-by-case examination of each data set is hardly possible."[297] However, the EDPB also acknowledged that rigorous safeguards like data minimization can help processors comply with the GDPR.[298]

The consent lawful basis requires affirmative action by the data subject, which would be impractical for scrapers to obtain. Instead, scrapers would need to obtain data by buying it from websites. The websites could obtain express consent to either sell their users' personal data to other companies or to use it themselves. But as we argue in this Section, this outcome is not optimal.

Under the legitimate interests lawful basis, the GDPR allows for the processing of personal data when "processing is necessary for the purposes of

294. GDPR, *supra* note 137, art. 6.

295. *Scraping Bijna Altijd Illegaal [Scraping Is Almost Always Illegal]*, *supra* note 146.

296. *Report of the Work Undertaken by the ChatGPT Taskforce*, EDPB (May 23, 2024), https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf [https://perma.cc/HM8P-T62F].

297. *Id.*

298. *Id.*

the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject . . . ."[299] This lawful basis, however, focuses primarily on a balance between the specific business interests of the scraper and the effect on the data subject. Due to the societal implications of widespread scraping, we argue that public interests should play a larger role in this balancing test.[300] This approach would address all the stakeholders involved, which includes the scrapers, scrapees, individuals whose data is scraped, and society as a whole.

Additionally, it is inevitable that scraping will gather sensitive or personal data that in combination could give rise to inferences about sensitive data, which is also deemed to be a special category of data under the GDPR.[301] For sensitive data, the legitimate interests lawful basis is unavailable.[302] It is thus difficult to imagine how scrapers could navigate around this problem.

U.S. law is even less clear. Many state privacy laws are triggered by the amount of revenue generated or the number of state residents whose data is gathered.[303] The latter would likely be triggered by large-scale scraping. Many state consumer privacy laws have limited opt-out rights, such as for automated profiling or targeted advertising.[304] But an opt-out right would be meaningless if people have no idea who the scrapers are or that their data is even being scraped.[305] Conversely, state laws that require opt-in consent for sensitive data could make scraping difficult or impossible.[306] Additionally, it remains unclear how many state laws will address the issue of whether inferences that reveal sensitive data count as sensitive data.

Because the true effect of scraping can only be appreciated at scale and not on an individualized basis, we contend that the most important question is whether data collection, use, and transfer is in the public interest.[307] Some laws, such as the GDPR and the FTC Act, already have the tools and flexibility to address this question. Other privacy laws are unsuitable. Our goal in this Section

299. GDPR, *supra* note 137, art. 6.1(f).

300. *What Is the 'Legitimate Interests' Basis?*, INFO. COMM'R'S OFF., https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/legitimate-interests/what-is-the-legitimate-interests-basis/#what_counts [https://perma.cc/Q8T7-P3PG] ("The legitimate interests of the public in general may also play a part when deciding whether the legitimate interests in the processing override the individual's interests and rights. If the processing has a wider public interest for society at large, then this may add weight to your interests when balancing these against those of the individual.").

301. Solove, *supra* note 140, at 1081.

302. GDPR, *supra* note 137, art. 9.

303. Miriam Farhi, Alison Watkins, Alanna Elinoff & Charlotte D. Kress, *Privacy Law Recap 2024: State Consumer Privacy Laws*, PERKINS COIE (Dec. 17, 2024), https://perkinscoie.com/insights/update/privacy-law-recap-2024-state-consumer-privacy-laws [https://perma.cc/45WC-ELDY].

304. SOLOVE & SCHWARTZ, *supra* note 188, at 186–90.

305. Solove, *supra* note 185, at 597.

306. SOLOVE & SCHWARTZ, *supra* note 188, at 186–90.

307. For an exploration on the role of scale in technology law, see generally McKenna & Hartzog, *supra* note 51.

is not to show how each and every privacy law might incorporate our recommended regulatory proscriptions. Instead, we will sketch out the basic aims the law should seek to achieve and the issues the law should focus on. Some laws may be capable of being interpreted and applied to carry out our approach. Other laws may need to be changed.

1. The Undesirability of a Total Scraping Ban

Although scraping conflicts with nearly all core principles of privacy, it should not be banned outright. Banning scraping would come at great financial and social costs, as so many basic information-search and -retrieval functions of the internet and AI depend upon scraping. Scraping can be a valuable tool to empower people, promote competition, and hold industry and government accountable for their own information practices. Many research projects and news stories cannot be achieved without scraping. Banning all scraping would severely impair companies' ability to develop AI and compete in certain markets. In a lawsuit against Google for scraping, Google declared that the suit would "take a sledgehammer not just to Google's services but to the very idea of generative AI."[308] Journalist Julia Angwin argues that "access to large quantities of public data" is essential for journalists to report on platforms, technology, and larger societal trends.[309] As Andrew Sellars notes: "Many forms of web scraping provide important benefits to consumers and the public."[310]

Restrictions on scraping could further distort AI models. If privacy laws in certain countries block scraping, AI data sets might become skewed if data is not collected about certain people and cultures through other means. Imagine if scraping could occur in the U.S. but not in the EU. AI models would be trained on U.S. data but deprived of EU data, skewing them to the U.S. It also seems likely that scraping will be important in the search for "less discriminatory algorithms," that is, alternative models that perform equally well but have less discriminatory impact than existing AI models.[311]

A scraping ban would also favor companies that already possess large data sets, such as big platforms, over smaller companies. While big platforms would have sufficient data to develop AI, smaller companies would lack the data to do so without other avenues for obtaining data. Companies with larger amounts of data are already starting to farm it from their own stores of data. Many U.S.

---

308. Blake Brittain, *Google Says Data-Scraping Lawsuit Would Take 'Sledgehammer' to Generative AI*, REUTERS (Oct. 17, 2023), https://www.reuters.com/legal/litigation/google-says-data-scraping-lawsuit-would-take-sledgehammer-generative-ai-2023-10-17/ [https://perma.cc/4HTV-A89B].

309. Julia Angwin, *The Gatekeepers of Knowledge Don't Want Us to See What They Know*, N.Y. TIMES (July 14, 2023), https://www.nytimes.com/2023/07/14/opinion/big-tech-european-union-journalism.html [https://perma.cc/H26N-ARKF].

310. Sellars, *supra* note 6, at 412.

311. *See, e.g.*, Emily Black, John Logan Koepke, Pauline T. Kim, Solon Barocas & Mingwei Hsu, *Less Discriminatory Algorithms*, 113 GEO. L.J. 53, 65 (2024).

privacy laws allow organizations to collect and use personal data in nearly any way they want just by stating what they are doing.[312] Many companies have already "updated their terms of service to include references to building AI with user data."[313] For example, Amazon announced plans to use data from its users to train its AI.[314] Google updated its privacy policy to state that it may "use publicly available information to help train Google's AI models and build products and features like Google Translate, Bard [now Gemini], and Cloud AI capabilities."[315] X revised its privacy notice to allow it to use "publicly available information" for training "[its] machine learning or artificial intelligence models."[316] In 2023, Zoom quietly altered its privacy notice to state that users agreed to the use of their data for training AI models, but then backpedaled after this change was called out publicly.[317] Although the FTC warned that changing privacy notices to allow for AI uses of previously collected data could violate the FTC Act,[318] companies are likely free to use data collected after any such changes are made.

Thus, a scraping ban could lock in the power of the biggest and most powerful companies and make it difficult for others to catch up. Although scraping personal data should not be banned in its entirety, if we value the privacy principles underpinning privacy law, scraping must be brought under control.

2. The Consent Model

Scraping could conceivably be brought within the purview of privacy law by websites obtaining individual consent for their data to be scraped by third parties. This practice would also be undesirable.

Under many U.S. privacy laws, websites could disclose the possibility of scraping in their privacy notices or provide explicit warnings of scraping. Under

312. Solove, *supra* note 185, at 602.

313. FED. TRADE COMM'N, GENERATIVE ARTIFICIAL INTELLIGENCE AND THE CREATIVE ECONOMY STAFF REPORT: PERSPECTIVES AND TAKEAWAYS 1, 10 (2023), https://www.ftc.gov/system/files/ftc_gov/pdf/12-15-2023AICEStaffReport.pdf [https://perma.cc/GAD4-Z29X].

314. Leffer, *supra* note 43.

315. Jess Weatherbed, *Google Confirms It's Training Bard on Scraped Web Data, Too*, VERGE (July 5, 2023), https://www.theverge.com/2023/7/5/23784257/google-ai-bard-privacy-policy-train-web-scraping [https://perma.cc/5Z7V-FKD5].

316. Sarah Perez, *X's Privacy Policy Confirms It Will Use Public Data to Train AI Models*, TECHCRUNCH (Sept. 1, 2023), https://techcrunch.com/2023/09/01/xs-privacy-policy-confirms-it-will-use-public-data-to-train-ai-models/ [https://perma.cc/5XEQ-LVG9].

317. Ian Krietzberg, *Zoom Walks Back Controversial Privacy Policy*, THESTREET (Aug. 11, 2023), https://www.thestreet.com/technology/zooms-latest-move-may-make-you-reconsider-using-the-service [https://perma.cc/B5PT-Q9PG].

318. *AI (and Other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive*, FTC: TECH. BLOG (Feb. 13, 2024), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive [https://perma.cc/9LEG-FHPU].

the notice-and-choice approach to privacy in many U.S. privacy laws, individuals who continue to post their data on a site or who fail to opt out will be deemed to have consented to the scraping.

Under the GDPR and other privacy laws requiring explicit consent (opt-in), websites could readily have users click a button or affirmatively acknowledge that they agree to the risk of scraping. However, it is unclear if such a broad-ranging consent would be deemed valid.

A consent approach to scraping would exacerbate existing shortcomings in privacy laws regarding consent. The concept of consent in most privacy laws is fictional.[319] This approach would subject individuals to data gathering and use on a massive scale, wrapping it in a veneer of legitimacy. It is hard to imagine how any form of consent to such massive data gathering and use for a myriad of unspecified purposes without limitation could be meaningful. In the United States, the notice-and-choice approach has been severely criticized as a vehicle for companies to gather and use data with hardly any limitations.[320] In the EU, the GDPR rejects the notice-and-choice approach: Consent must be express and affirmative (opt-in).[321] But even express consent can sometimes readily be obtained and is not meaningful. Websites can make people click "accept" buttons without them understanding the implications.[322] Privacy consent is mostly fictional, and people will readily consent to the use of their data in exchange for the immediate benefits of technology.[323] Professor Elettra Bietti warns that consent has become a "free pass" for platforms to use personal data in nearly any way they desire.[324]

If companies procure the appropriate consent, they could sell their data (or the ability to scrape the data) to third parties. For example, Reddit originally had a free API for scrapers but in 2023 started to charge for the use of its API.[325] Indeed, such a model need not involve scraping; websites could just provide the

319. Solove, *supra* note 185, at 631.

320. *See generally* Lina M. Khan, Samuel A.A. Levine & Stephanie T. Nguyen, Feature, *After Notice and Choice: Reinvigorating "Unfairness" to Rein in Data Abuses*, 77 STAN. L. REV. 1375 (2025); Neil Richards & Woodrow Hartzog, *The Pathologies of Digital Consent*, 96 WASH. U. L. REV. 1461, 1463 (2019); Robert H. Sloan & Richard Werner, *Beyond Notice and Choice: Privacy, Norms, and Consent*, 14 J. HIGH TECH. L. 370 (2014); Helen Nissenbaum, *A Contextual Approach to Privacy Online*, 140 DÆDALUS, J. AM. ACAD. ARTS & SCIS. 32, 34 (2011).

321. GDPR, *supra* note 137, art. 4(11) (requiring consent to be "freely given, specific, informed and unambiguous indication of the data subject's wishes").

322. *See* HARTZOG, *supra* note 50, at 60–67, 129–30. Even with "accept" buttons, readership of terms barely increases. *See* Florencia Marotta-Wurgler, *Will Increased Disclosure Help? Evaluating the Recommendations of the ALI's "Principles of the Law of Software Contracts*,*"* 78 U. CHI. L. REV. 165, 168 (2011) (requiring people to click an "I agree" box next to terms only increases readership by 1%).

323. *See generally* Solove, *supra* note 185; Solove & Hartzog, *supra* note 177, at 1023–24.

324. Elettra Bietti, *Consent as a Free Pass: Platform Power and the Limits of the Informational Turn*, 40 PACE L. REV. 307, 308 (2020).

325. Wallace Witkowski, *Reddit Founder Wants to Charge Big Tech for Scraped Data Used to Train AIs: Report*, MARKETWATCH (Apr. 18, 2023), https://www.marketwatch.com/story/reddit-founder-wants-to-charge-big-tech-for-scraped-data-used-to-train-ais-report-6f407265 [https://perma.cc/8RVQ-HGRQ].

data to third parties, though such a practice would be functionally equivalent to scraping.

This approach would leverage the infirmities of consent to leave individuals whose data is scraped largely out of the loop. Massive data transfer would occur based on a series of backroom deals without individuals having a seat at the table. Additionally, any approach built upon individual consent ignores collective concerns, such as harmful effects on marginalized communities like people of color and members of the LGBTQ+ community.[326]

C. A Regulatory Agenda for Scraping in the Public Interest

Instead of the general approach in the United States, which allows organizations wide leeway to collect and use personal data in whatever way they want, the law should, like the GDPR's approach, view the systemic, automated mass collection and use of personal data through scraping as a *privilege*. This privilege would allow data scraping in justified contexts upon the adoption of safeguards and commitments that benefit society as a whole.

Our proposal has three components: 1) a requirement to demonstrate a valid justification for scraping focused on the public interest before scraping is allowed; 2) substantive protections to ensure the scraping is safe and avoid exploitation and purpose drifting too far from the public interest; and 3) procedural safeguards to ensure fairness, adequate representation, and agency in decision-making.

First, we propose that automated mass scraping of personal data should only be allowed when it is necessary to further the public interest. Although the GDPR has public interest as one of the six lawful bases to process personal data, this basis is often not discussed for most commercial uses of personal data.[327] Under the GDPR's public interest lawful basis, data can be processed when "processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller."[328] However, this provision is geared towards the exercise of government authority and is likely to be interpreted narrowly.[329] In U.S. privacy law, the public interest

326. *See generally* NANCY S. KIM, CONSENTABILITY: CONSENT AND ITS LIMITS (2019); MEREDITH BROUSSARD, MORE THAN A GLITCH: CONFRONTING RACE, GENDER, AND ABILITY BIAS IN TECH (2023); Salomé Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573 (2021); Joshua A.T. Fairfield & Christoph Engel, *Privacy as a Public Good*, 65 DUKE L.J. 385 (2015); Chris Gilliard, *The Rise of 'Luxury Surveillance*,*'* ATLANTIC (Oct. 18, 2022), https://www.theatlantic.com/technology/archive/2022/10/amazon-tracking-devices-surveillance-state/671772/ [https://perma.cc/ZLH4-664V]; Evan Selinger & Woodrow Hartzog, *The Inconsentability of Facial Surveillance*, 66 LOY. L. REV. 101 (2019).

327. *Public Task*, INFO. COMM'R'S OFF., https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/lawful-basis-for-processing/public-task/ [https://perma.cc/88AQ-SLY2].

328. GDPR, *supra* note 137, art. 6.1(e).

329. *See Public Task*, *supra* note 327 ("Section 8 of the Data Protection Act 2018 (DPA 2018) says that the public task basis will cover processing necessary for: the administration of justice;

remains an underutilized concept, though scholars have looked to collective and social aspects of information privacy.[330]

We contend that a robust conception of public interest could be a suitable basis to justify scraping, but such a basis would need to be broader and more open-ended than what the GDPR allows. We use "public interest" here to mean a consideration of the collective or shared well-being of a public or publics as opposed to a more atomized, individualized well-being. Specifically, we deploy the term "public interest" similar to how the concept of the public has been deployed in public health law.[331] Specifically, we recommend a population-level focus that is consistent with the values of social justice: the "[f]air and equitable treatment of groups and individuals, with particular attention to the disadvantaged."[332]

Scraping should be allowed (and even facilitated) for targeted interventions in the public interest with procedural and substantive protections to ensure fit to purpose and to prevent financial incentives for exploitation. When the use of the data is not in the public interest, scraping should not be allowed. Nor should companies be allowed to use fictitious methods of consent as a means to gather or sell data.

We propose developing a framework for data scraping in the public interest. We recognize that such a framework will be difficult to create and involve contested issues, but public interest is the most productive focal point for the policymaking conversation on scraping. Given the extreme financial incentives companies have for over-collection and misuse of personal data, a strategy that limits private gain is the most direct and efficient way to retain the societal benefits of scraping while harmonizing it with privacy rules.

For example, while the GDPR might allow for scraping with consent or for legitimate interests, these legal bases are too manipulable and broad. As one of us has argued, even GDPR-style express consent is deeply flawed and could readily be obtained via "accept" buttons or other means that are not indicative of meaningful consent. The "legitimate interests" lawful basis for scraping personal data is too broad or unlikely to apply. Although narrowed by a balancing test with people's fundamental rights and freedoms, the legitimate interests lawful

---

parliamentary functions; statutory functions; governmental functions; or activities that support or promote democratic engagement. However, this is not intended as an exhaustive list. If you have other official non-statutory functions or public interest tasks you can still rely on the public task basis, as long as the underlying legal basis for that function or task is clear and foreseeable.").

330. *See generally* Viljoen, *supra* note 326, at 573; Fairfield & Engel, *supra* note 326, at 385; PRISCILLA M. REGAN, LEGISLATING PRIVACY: TECHNOLOGY, SOCIAL VALUES, AND PUBLIC POLICY (1995); Priscilla M. Regan, *Privacy and the Common Good: Revisited*, *in* THE SOCIAL DIMENSIONS OF PRIVACY 50 (Beate Roessler & Dorota Mokrosinska eds., 2015); Tschider, *supra* note 1, at 132.

331. LAWRENCE O. GOSTIN & LINSDAY F. WILEY, *A Theory and Definition of Public Health Law*, *in* PUBLIC HEALTH LAW: POWER, DUTY, RESTRAINT 3, 4 (3d ed. 2016).

332. *Id.* at 5. For a developed framework for addressing group-specific harms like harms to African Americans (not just general harms), see generally Anita L. Allen, *Dismantling the "Black Opticon": Privacy, Race Equity, and Online Data-Protection Reform*, 131 YALE L.J.F. 907 (2022).

basis broadly allows processing of personal data for a wide range of purposes "pursued by the controller or by any third party." A legitimate interest is not an allowable lawful basis for sensitive data: Since sensitive data includes personal data that could be used to infer sensitive data, nearly all personal data could arguably be deemed to be sensitive.[333] If scrapers are unable to rely upon legitimate interests as a legal basis to process data, then most scraping of personal data will be effectively prohibited.

We do not propose to follow precisely the particular formulation or interpretations of the GDPR's public interest lawful basis; rather, we simply suggest that the permissible basis for scraping should rest upon public interest. Scraping in the public interest does not preclude making a profit; nor does it preclude all risks to individuals. But it must be justified in ways beyond benefits to companies alone.

In some circumstances, lawmakers and regulators might adopt bright line rules such as "no scraping for biometric purposes." Other strategies might include facilitating academic and journalistic scraping through the use of safe harbors and explicit exemptions to scraping rules similar to the GDPR's exemptions for personal and household data processing or targeted exemptions for academic, artistic, or literary expression. In any event, lawmakers should explicitly engage in public deliberation about the specific contexts where scraping is and is not in the public interest, consistent with the values of social justice and a pluralist democracy.[334]

We suggest that at least four principles should guide the law:

1) *Reasonable Risk of Harm Principle*: The collection, use, or transfer of scraped personal data should not cause unreasonable risk of harm to individuals, disadvantaged groups, or society.
2) *Proportional Benefits Principle*: The collection, use, or transfer of scraped personal data should provide meaningful benefits to individuals, disadvantaged groups, and society sufficient to outweigh any risks and proportional to or in excess of the benefits to the scraper.
3) *Process Principle*: The process for deciding the legitimate uses of scraped personal data should be fair, open, accountable, representative, equitable, and deliberative.
4) *Protections Principle*: Scraped data should be afforded all the same protections as other personal data under privacy laws unless particular protections impose an unreasonable conflict

333. *See generally* Solove, *supra* note 140.

334. A good starting point for this discussion would center the three values articulated by Ari Waldman and others to ground an anti-subordination tech law framework: power, equality, and democracy. *See generally* Ari Ezra Waldman, *Privacy, Practice, and Performance*, 110 CALIF. L. REV. 1221, 1270 (2022); NEIL RICHARDS, WHY PRIVACY MATTERS (2021); JULIE COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSEQUENCES OF INFORMATIONAL CAPITALISM (2019).

with other public interest goals.

The first two principles can guide lawmakers in determining when scraping is in the public interest by weighing harms and benefits. The last two can guide lawmakers in creating rules and safeguards to ensure scraping is safe, just, and true to its original public purpose.

1. Use of Data as a Privilege for Furthering Public Interests

Generally, U.S. privacy law views data collection and use as the natural right of organizations. In contrast, we propose requiring a permissible purpose for data collection and processing, like GDPR's lawful basis approach.

The collection and use of personal data should be understood as a privilege. Scraping personal data should be allowed when it is, on balance, in the public interest, because the financial incentives from scraping also encourage a host of individual and social information-related harms, including harassment, labor exploitation, manipulation, and wrongful discrimination.

Articulations of what constitutes the "public interest" should be specific, compelling, grounded in reality, and directly related to the collection of information. Mere conveniences such as workplace efficiencies or more seamless commercial transactions should not qualify. Allegations that scraping will help "keep people safe" or "improve your health" should be insufficient without evidence demonstrating that the scraping is necessary and proportionate to the purpose. If companies want to use people's data, the public should receive something in return.

Instead of forcing websites to allow certain forms of scraping, such as scraping by the media or researchers or competitors, the law should take an incentives approach. It can allow websites to use their repositories of data for their own purposes (assuming these purposes are not harmful) if these sites allow for the collection and use of data in the public interest. Such an approach is only possible when privacy law is retooled to move away from an excessive focus on individual control and more toward a model of focusing on harms and risks. Only when the law recognizes that the collection and use of personal data is a privilege, rather than the natural right of organizations, will meaningful controls and limitations be possible, as well as meaningful protections of individual privacy.

2. Guiding Principles for Regulating Scraping

Guidelines about scraping in the public interest must be developed. In this Section, we outline our four proposed principles: (1) Reasonable Risk of Harm Principle, (2) Proportional Benefits Principle, (3) Process Principle, and (4) Protections Principle.

Under the *Reasonable Risk of Harm Principle*, the law should protect people from downstream harms from having their data scraped. Lawmakers should consider not just harms at the individual level, but also harms to

disadvantaged groups such as oppressive and discriminatory surveillance. Lawmakers should also consider collective or publicly felt harms such as the corrosion of social trust, the collapse of democratic institutions, and the failure of infrastructure.[335]

The law cannot be perfect in anticipating future harms, and scraping should be allowed in some instances even when the future impact of the technologies and tools developed or trained with the use of scraped data is uncertain. But measures should be in place for situations where AI starts to cause unreasonable harm. This harm must be mitigated.

But the problems with scraping extend beyond harm to data subjects. One of the biggest problems with "free for all" scraping is when scrapers keep all the value with little benefit for society. Under the *Proportional Benefits Principle*, there must be articulable benefits to the collection, use, and transfer of personal data that are proportional to or exceed the benefits to the scraper. Rules based on the benefit principle should require that the purported benefit be specific, compelling, grounded in reality, and necessary and proportional to the collection of information.

Lawmakers could model these rules on other legal frameworks designed to mitigate conflicted self-dealing that disproportionately benefits powerful parties, such as modern proposals for data loyalty obligations and information fiduciary rules.[336] These proposals suggest holding powerful parties to duties of loyalty, care, and confidentiality in their relationships with people who are vulnerable due to their sharing of personal data. While loyalty duties would apply only within relationships, lawmakers could look to the way these frameworks scrutinize the disproportionate benefit flowing to scrapers while simultaneously imposing massive externalities on society to help identify when the societal benefits of scraping personal data are justified.[337] Another area of law that can

335. *See generally, e.g.*, Allen, *supra* note 332; Julie E. Cohen, *Infrastructuring the Digital Public Sphere*, 25 YALE J.L. & TECH. (SPECIAL ISSUE) 1 (2023). For more background, see generally ROBERT D. PUTNAM, BOWLING ALONE: THE COLLAPSE AND REVIVAL OF AMERICAN COMMUNITY (2000); BRETT M. FRISCHMANN, INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES (2012).

336. *See, e.g.*, SOLOVE, *supra* note 92; ARI EZRA WALDMAN, PRIVACY AS TRUST: INFORMATION PRIVACY FOR AN INFORMATION AGE (2018); Neil Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 WASH. U. L. REV. 961 (2021); Woodrow Hartzog & Neil Richards, *The Surprising Virtues of Data Loyalty*, 71 EMORY L.J. 985 (2022); Woodrow Hartzog & Neil Richards, *Legislating Data Loyalty*, 97 NOTRE DAME L. REV. REFLECTIONS 356 (2022); Woodrow Hartzog & Neil Richards, *Privacy's Constitutional Moment and the Limits of Data Protection*, 61 B.C. L. REV. 1687 (2020); Neil Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 19 STAN. TECH. L. REV. 431 (2016); Jack M. Balkin, *The Fiduciary Model of Privacy*, 134 HARV. L. REV. F. 11 (2020); Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183 (2016); Claudia E. Haupt, *Platforms as Trustees: Information Fiduciaries and the Value of Analogy*, 134 HARV. L. REV. F. 34 (2020); Lilian Edwards, *The Problem with Privacy: A Modest Proposal*, 18 INT'L REV. L. COMPUT. & TECH. 309 (2004); Ian R. Kerr, *The Legal Relationship Between Online Service Providers and Users*, 35 CANADIAN BUS. L.J. 419 (2001).

337. *See generally* Jordan Francis, Woodrow Hartzog & Neil Richards, *A Concrete Proposal for Data Loyalty*, 37 HARV. J.L. & TECH 1335 (2024).

help inform lawmakers and regulators might be the law of unjust enrichment, restitution, and disgorgement.[338] These proposals seek to mitigate, reserve, or prevent the wrongful gains by companies, and in theory should apply when companies unjustly scrape personal data for private gain.

The *Process Principle* recognizes that not only must good substantive determinations be made about the uses of scraped data, but the process for deciding upon uses should also be fair, open, accountable, representative, and thoughtful. Privacy laws already require some of these processes, such as risk assessments and accountability.[339] Many laws require fairness.[340] But laws often fail to ensure that a reasonably diverse set of stakeholders have input in decisions about technology or that these decisions are made in an open way. As Ngozi Okedigbe has argued, even the pursuit to "democratize" rules for information practices often just "exacerbates existing inequalities, power imbalances, and social stratification."[341] Laws require risk or impact assessments but rarely require any rigor as to the requirements of such assessments, which can result in evaluations that are not sufficiently thoughtful. They also too frequently do not grapple with how power is distributed and wielded among different groups.[342] The result is that people, particularly disadvantaged groups, are often completely shut out of the decision-making process or are given a threadbare kind of participation but left with no real agency.[343]

A good place to start considering the conditions upon which data may be scraped in the public interest might be the "Public Interest Privacy Legislation Principles," endorsed by thirty-four civil rights, consumer, and privacy organizations.[344] The privacy principles outline four concepts that any meaningful data protection rules should incorporate at a minimum, including that privacy protections must be strong, meaningful, and comprehensive and data practices must protect civil rights, prevent unlawful discrimination, and advance

338. *See, e.g.*, Bernard Chao, *Privacy Losses as Wrongful Gains*, 106 IOWA L. REV. 555, 557–58 (2021) ("Disgorgement gives the plaintiff a monetary remedy based on the defendant's wrongful gains as opposed to the plaintiff's injury. Disgorgement is often used when expectation damages are inadequate or simply difficult to assess. Because privacy injuries confound other traditional doctrines, disgorgement is particularly well suited to address these problems."); Lauren Henry Scholz, *Privacy Remedies*, 94 IND. L.J. 653, 670 (2019).

339. *See, e.g.*, Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347, 1351, 1379 (2023).

340. *See, e.g.*, Malgieri, *supra* note 172, at 3, 14.

341. *See, e.g.*, Ngozi Okidegbe, *To Democratize Algorithms*, 69 UCLA L. REV. 1688, 1688 (2023).

342. *See generally* Margot E. Kaminski & Gianclaudio Malgieri, *Algorithmic Impact Assessments Under the GDPR: Producing Multi-Layered Explanations*, 11 INT'L DATA PRIV. L. 125 (2021).

343. *See, e.g.*, Okidegbe, *supra* note 341, at 1688.

344. PUB. INT. RSCH. GRP., PUBLIC INTEREST PRIVACY PRINCIPLES, https://publicinterestnetwork.org/wp-content/uploads/2018/11/Public_Interest_Privacy_Principles.pdf [https://perma.cc/AR4Q-6NMS].

equal opportunity.[345] Additionally, rules that justify scraping in the public interest, including "in areas such as housing, employment, health, education, and lending, must be judged by its possible and actual impact on real people, must operate fairly for all communities, and must protect the interests of the disadvantaged and classes protected under anti-discrimination laws."[346]

Finally, the *Protections Principle* aims to afford scraped personal data with all the protections ordinarily provided by privacy laws, except for protections that are unworkable or in cases of overriding societal need, such as use in emergencies or law enforcement investigations following due process. Scraped data should not lose all privacy protections because it is publicly available. Therefore, in all relevant laws, lawmakers and judges should clarify the fact that even when information is publicly available it is still protected by privacy laws. Lawmakers should also require reasonable anti-scraping safeguards against scraping not in the public interest as part of a company's overall duty to reasonably secure its entrusted personal data.[347]

CONCLUSION

We are in the midst of a scraping epidemic—the Great Scrape—where companies and others are ruthlessly plundering the internet of its data without regard for law or ethics. Scraping and data privacy are in desperate need of a reconciliation. Scraping is in conflict with nearly all core privacy principles. Yet courts have delivered mixed outcomes that neither wholly endorse nor categorically prohibit scraping practices. The absence of clear legal guidance risks perpetuating uncertainty for those seeking to scrape data for legitimate and desirable purposes, for people sharing personal data with online services, and for those services bound and motivated to protect people's personal information.

Scraping already exists in partial tension with existing laws, but these laws are ambiguous, inconsistent, and weakly enforced. Many laws focus mostly or exclusively on the interests of organizations maintaining personal data rather than the individuals to whom the data pertains. As a result, legal battles over scraping often ignore privacy considerations. Privacy laws are also failing to address scraping because of the common view that "publicly accessible" personal data lacks any privacy interests, even though there are many privacy harms that result from its collection, use, and further disclosure. The law fails quite significantly to account for the privacy implications of mass data scraping, leaving people exposed and vulnerable in the Scraping Wars. For laws that

345. *Id.*

346. *Id.* at 2 ("Legislation should ensure fundamental fairness of and transparency regarding automated decision-making. Automated decision-making, including in areas such as housing, employment, health, education, and lending, must be judged by its possible and actual impact on real people, must operate fairly for all communities, and must protect the interests of the disadvantaged and classes protected under anti-discrimination laws.").

347. *See generally* SOLOVE & HARTZOG, *supra*, note 218, at 47–51; William McGeveran, *The Duty of Data Security*, 103 MINN. L. REV. 1135 (2019).

purportedly could apply to scraping, enforcement agencies remain afraid to strongly enforce them for fear of disrupting Panglossian promises of AI innovation and boundless prosperity and goodness. Unfortunately, the inconvenient truth is that much AI is trained based on a massive taking of people's data without consent, oversight, limitation, or any consideration of the harms it might create. Much profit from AI comes from the extraction of personal data as a "free" resource. We have contended that online personal data is not free for the taking, and the law must stop this mass data looting.

A more rigorous and nuanced legal approach is necessary to establish coherent rules that balance the public interest in scraping with people's privacy. A ban on scraping is untenable, so a compromise must be reached. This compromise requires creativity to protect privacy in ways beyond many existing approaches. Ultimately, we recommend that lawmakers should view the systemic, automated mass collection and use of personal data through scraping as a privilege. By conditioning scraping upon serving the public interest, we can finally reconcile it with the protection of privacy.