THE GREAT SCRAPE:
THE CLASH BETWEEN SCRAPING AND PRIVACY

by

Daniel J. Solove

&

Woodrow Hartzog

Draft: July 3, 2024

ABSTRACT

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 is about people. This personal data provides the grist for AI tools such as facial recognition, deep fakes, and generative AI. Although scraping enables web searching, archival, and meaningful scientific research, scraping for AI can also be objectionable or 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 in privacy laws, including fairness; individual rights and control; transparency; consent; purpose specification and secondary use restrictions; data minimization; onward transfer; and data security. With scraping, data protection laws built around these requirements are ignored.

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.

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.

THE GREAT SCRAPE: THE CLASH BETWEEN SCRAPING AND PRIVACY

By Daniel J. Solove[1] and Woodrow Hartzog[2]

INTRODUCTION 4
I. THE GREAT SCRAPE 7
A. Understanding Scraping 7
1. The Rise of Scraping 7
2. Scraping in the Age of AI 9
3. Scraping Personal Data 10
4. The Ethical Twilight of Scraping 11
B. The Scraping Wars 13
1. The Legal Front 14
2. The Technological Front 21
C. The Emerging Scraping Market 23
D. Regulatory Intervention 23
1. EU Data Protection Law 23
2. U.S. Privacy Law 26
E. The Need for a Coherent Theory of Scraping and Privacy 27
II. SCRAPING AND PRIVACY: A FUNDAMENTAL TENSION 29
A. Scraping and Privacy Principles 30
1. Fairness 30
2. Individual Rights and Control 31
3. Transparency 33
4. Consent 33
5. Purpose Specification and Secondary Use Restrictions 34
6. Data Minimization 34
7. Onward Transfer 36
8. Data Security 38
B. Scraping and Publicly Available Information 38
1. Publicly Available Information: An Incoherent Concept 39
2. Expectations of Privacy in Publicly Available Information 41
3. Privacy Law and Publicly Available Information 42
III. RECONCILING SCRAPING AND PRIVACY 45
A. A Theory of Surveillance and Security 46
1. Scraping as Surveillance 46
2. Protection from Scraping as Security 48
B. The Difficulty of Bringing Scraping Under the Purview of Privacy Law 50
1. The Undesirability of a Total Scraping Ban 52
2. The Consent Model 54
C. A Regulatory Agenda for Scraping in the Public Interest 56
1. Use of Data as a Privilege 59
2. Guidelines for Scraping 61
CONCLUSION 64

[1] 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.
[2] Professor of Law, Boston University School of Law.

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 any notice, consent, or opportunity to object or opt out for the data subject. This personal data provides the grist for AI tools such as facial recognition, deep fakes, and large language models. Scraping is a foundational practice for the modern digital sphere. Organizations and individuals used it to build what we know as the World Wide Web and rely upon it for essential and everyday information services. Although scraping personal data enables web searching, archiving, generative AI, and scientific research, scraping for AI can also be objectionable or even harmful to individuals and society 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 are seemingly inconsistent with the practice. There has always been a fundamental conflict between scraping and privacy, but for years this tension has remained a background concern. AI has brought this tension to the forefront. AI requires scraping on a grand scale.[3] Recently, we have witnessed companies scrape an unprecedented amount of data, and more and more companies are scraping.

In this Article, we contend that scraping must undergo a serious and long overdue reckoning with privacy. Scraping of personal data violates nearly every key privacy principle embodied in privacy laws, 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. But privacy law is currently conflicted about publicly available data. Although some laws exclude such data, other laws such as the EU's General Data Protection Regulation largely do not. Additionally, many courts have recognized that public exposure does not extinguish one's privacy interest. Most notably, the U.S. Supreme Court held that there is a reasonable expectation of privacy under the Fourth Amendment for geolocation data about publicly observable automobile movement and that there is a privacy interest in the practical obscurity of personal data in certain publicly available records.[4]

---

[3] Charlotte A. Tschider, *AI's Legitimate Interest: Towards a Public Benefit Privacy Model,* 21 Hous. J. Health L. & Policy 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.").

[4] Carpenter v. United States, 138 S. Ct. 2206 (2018); U.S. Dep't of Justice v. Reporters Committee for Freedom of the Press, 489 U.S. 749 (1989).

Beyond scrapers, the organizations whose websites are scraped (the "scrapees") also must have a reckoning with privacy. Organizations can mitigate scraping through certain measures, but too often, the actions taken by companies to prevent scraping of their website are minimal. 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 a farce 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 the fundamental tension between scraping and privacy. Our thesis is that scraping is generally anathema to the core principles of privacy that form the backbone of most privacy laws, 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 laws, especially the Computer Fraud and Abuse Act (CFAA). Our focus is much broader and more conceptual. What makes scraping such an important and fascinating issue is that it stands so 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 useable 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.

As we will discuss, 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 and different approaches than most laws take. There is a fundamental tension between scraping and core longstanding privacy principles. Nevertheless, a world without scraping would hobble the internet, stunt the development of AI, and frustrate research and journalism.

With so much personal data publicly available online, with the ability to hoover up this data so readily with automation, it is impossible to have meaningful privacy protection when scraping can occur without legal restrictions or policies that support technical safeguards against scraping. 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 in an effort 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 laws and standards. We argue that 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. 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 shouldn't bar all instances of scraping. Instead, the law should require a legitimate basis for scraping, encourage scraping in the public interest, and impose restrictions on scraping for harmful or risky uses. Although it is present in a narrow form in some laws, the concept of public interest generally has been underutilized in privacy laws. 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 only pockets of 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 output generated from another program."[5] 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)."[6] Scraping "is used to transform unstructured data on the web into structured data that can be stored and analyzed in a central local database or spreadsheet."[7]

Colloquially, some might use the term scraping to describe "manual" techniques like the traditional copy-and-paste.[8] But our focus here is the kind of automated scraping that occurs through the use of programs called "web crawlers," "spiders," or "bots" and makes the mass collection of information relatively cheap and easy.[9] These computer programs scour the internet gathering information from webpages. Scraping is a ubiquitous practice, and it is increasing.

1. The Rise of Scraping

Bots have long roamed the internet; they have been deployed since the early 1990s when the commercial internet began to develop.[10] One of the earliest forms of

---

[5] *What is Data Scraping*, CLOUDFLARE, https://www.cloudflare.com/learning/bots/what-is-data-scraping/.
[6] Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. Sci. & Tech. L. 372, 373 (2018).
[7] S.C.M. De S Sirisuriya, *A Comparative Study on Web Scraping*, 135 (Int'l Rsch. Conf. Kotelawala Def. Univ. 2015), http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf?sequence=1&isAllowed=y.
[8] *Id.*
[9] Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372, 381–84, 381 n.57 (2018), citing 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 Internet 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.'").
[10] Seyed M. Mirtaheri et al., *A Brief History of Web Crawlers*, PROCEEDINGS OF THE 2013 CONFERENCE OF THE CENTER FOR ADVANCED STUDIES ON COLLABORATIVE RESEARCH, 3 (2013) (noting that web

scraping that is still popular today involves search engines using bots to crawl and index websites, a practice that makes the internet searchable. Different purposes for scraping soon emerged, such as conducting market research, compiling feeds, monitoring competitor pricing and practices, and analyzing trends and activities.[11]

Any publicly-accessible website can be scraped by automated tools.[12] (Technically, password-protected and paywalled websites can be scraped too, but because they typically cannot be automatically crawled without access credentials this practice is not as popular for large-scale data collection.)[13] Scrapers gather data from freely accessible social media profiles was well as many other types of sites such as those involving fitness, banking, and hospitality.[14]

Web scraping bots are designed to gather data from websites in an efficient and systematic manner. Not all bots engage in web scraping; bots are used in myriad helpful and harmful ways, such as to post spam comments, engage in marketing, exploit vulnerabilities, and launch DDOS attacks.[15]

For a long time, bots that gather information on the internet have operated in an oddly chivalrous fashion. Websites use a simple text file called robots.txt to politely tell bots whether or not to crawl their site.[16] As technology journalist David Peirce 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."[17] But as Zachary Gold and Mark Latonero note, "robots.txt can be ignored; those employing crawlers are not bound by any law contract, or technical need to obey a robots.txt file."[18] Remarkably, this system has worked; many bots have respected robots.txt instructions.

Over time, scraping has become easier and more prevalent.[19] The online world began to be populated more and more by bots. By 2014, more than a quarter of

---

crawlers have existed since 1993, where they "mainly collected information and statistic[s] about the web . . . and downloaded URLs").

[11] Margaret Rouse, *Web Scraping*, TECHOPEDIA (Feb 8. 2023), https://www.techopedia.com/definition/5212/web-scraping.

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

[13] Some websites take affirmative steps to allow search engines like Google to access content behind a paywall with web crawlers. *See, e.g.*, Madeline White, *Ask the experts: paywalls, subscription and SEO*, THE AUDIENCEERS (Sept. 12, 2023), https://theaudiencers.com/ask-the-experts-paywalls-subscription-and-seo/.

[14] *Id.*

[15] Adrienne LaFrance, *The Internet Is Mostly Bots*, THE ATLANTIC (Jan. 31, 2017), https://www.theatlantic.com/technology/archive/2017/01/bots-bots-bots/515043/; https://medium.com/datasociety-points/bots-a-definition-and-some-historical-threads-47738c8ab1ce.

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

[17] *Id.*

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

[19] Isaiah Poritz, *OpenAI's Legal Woes Driven by Unclear Mesh of Web-Scraping Laws*, BLOOMBERG LAW (July 5, 2023, 5:44 AM), https://news.bloomberglaw.com/ip-law/openais-legal-woes-driven-by-unclear-mesh-of-web-scraping-laws.

internet traffic was estimated to consist of bots.[20] By 2017, some commentators estimated that bots accounted for more than half of internet traffic; an article in *The Atlantic* proclaimed that "[m]ost website visitors aren't humans."[21]

2. Scraping in the Age of AI

AI demands vast amounts of training data.[22] Some of this data is collected directly from people. Other times, data is collected from a company using an application programming interface, known as an "API," which are designed for a consensual extraction and sharing of data.[23] However, most of this data is obtained through scraping.[24]

Large language models (LLMs) and generative AI must be fed unprecedented quantities of data. Most companies are usually either scraping data or purchasing scraped data to compete with rivals. Scraping today is occurring like the gold rush – a frenzied data grab on the grandest of scales. The market for web scraping software approached half a billion dollars in 2023 and is expected to quintuple in the next 15 years.[25]

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.[26] Clearview AI's facial recognition tool quickly become widely used by law enforcement organizations around the world.[27] The company operated in the shadows until 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.[28]

Another instance of a colossal scraping campaign was carried out by OpenAI, the creator of the popular generative AI tools, ChatGPT and Dall-E. Perhaps more than any other company, OpenAI's generative AI also generated public attention and fueled the current hype in AI. To develop its tools, OpenAI plundered the internet in massive scrapes to gather enormous quantities of training data.[29] The company

---

[20] Philip H. Liu & Mark Edward Davis, *Web Scraping—Limits on Free Samples*, 8 LANDSLIDE 54, 54 (2015).
[21] Adrienne LaFrance, *The Internet Is Mostly Bots*, THE ATLANTIC (Jan. 31, 2017), https://www.theatlantic.com/technology/archive/2017/01/bots-bots-bots/515043/. *See also Distribution of bot and human web traffic worldwide from 2014 to 2022*, STATISTA (May 2023), https://www.statista.com/statistics/1264226/human-and-bot-web-traffic-share/.
[22] Lee Tiedrich, *The AI data scraping challenge: How can we proceed responsibly*, OECD.AI POLICY OBSERVATORY (Mar. 5, 2024), https://oecd.ai/en/wonk/data-scraping-responsibly.
[23] *See* Ma-Keba Frye, *What is an API?,* MULESOFT, https://www.mulesoft.com/resources/api/what-is-an-api; Michael Goodwin, *What is an API?,* IBM (Apr. 9, 2024), https://www.ibm.com/topics/api.
[24] U.S. FED. TRADE COMMISSION, GENERATIVE ARTIFICIAL INTELLIGENCE AND THE CREATIVE ECONOMY STAFF REPORT: PERSPECTIVES AND TAKEAWAYS 9 (Dec. 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/12-15-2023AICEStaffReport.pdf.
[25] Abhishek Verma & Hetal Singh, *We Scraping Software Market*, RESEARCH NESTER (Jan. 17, 2024), https://www.researchnester.com/reports/web-scraping-software-market/5041.
[26] *OAIC and UK's ICO open joint investigation into Clearview AI Inc.*, AUSTRALIAN GOVERNMENT OFFICE OF THE AUSTRALIAN INFORMATION. COMMISSIONER (July 9, 2020), https://www.oaic.gov.au/newsroom/oaic-and-uks-ico-open-joint-investigation-into-clearview-ai-inc.
[27] KASHMIR HILL, YOUR FACE BELONGS TO US; A SECRETIVE STARTUP'S QUEST TO END PRIVACY AS WE KNOW IT (2023).
[28] *Id.*
[29] Kieran McCarthy, *Web Scraping for Me, But Not for Thee*, TECH. & MARKETING L. BLOG (Aug. 24,

has been accused of scraping data from "hundreds of millions of internet users."[30]

The scale of these scrapes and others is unprecedented – the amount of data gathered from each scraper is mindboggling and the total amount of data amassed by all scrapers is nearly beyond comprehension.

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 hungry AI companies.[31] 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."[32]

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.[33] Elon Musk stated that "[s]everal hundred organizations (maybe more) were scraping Twitter data extremely aggressively."[34]

As AI continues to bedazzle investors and the public, as it continues its meteoric rise, scraping will invariably increase, as the 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.

3. Scraping Personal Data

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. People write about their health, beliefs, political opinions, reading interests, movie and musical tastes, friends, family, buying habits, fitness, resumes, and nearly every corner of their 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; so do organizational websites, which have biographies of their employees. People's thoughts and conversations are online in comment threads to articles or on social media. Personal data exists online in every corner and crevice, like insects in the rain forest.

2023) (nothing that ChatGPT has "almost certainly already scraped the entire non-authwalled-Internet" and used the data to train ChatGPT), https://blog.ericgoldman.org/ archives/2023/08/web-scraping-for-me-but-not-for-thee-guest-blog-post.html.

[30] Isaiah Poritz, *OpenAI's Legal Woes Driven by Unclear Mesh of Web-Scraping Laws*, BLOOMBERG LAW (July 5, 2023, 5:44 AM), https://news.bloomberglaw.com/ip-law/openais-legal-woes-driven-by-unclear-mesh-of-web-scraping-laws.

[31] *2023 Imperva Bad Bot Report*, IMPERVA (May 10, 2023) https://www.imperva.com/resources/reports/2023-Imperva-Bad-Bot-Report.pdf.

[32] *Id.*

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

[34] *Id.*

Online social media platforms host the most personal data, but there are also countless blogs and other sites where people post photos and personal data. Law firm websites, university websites, and many others have biographical information about their employees, as well as photos of employees and information about students. Personal data is marbled throughout the internet.

It is hard to estimate just how much personal data is hoovered up in various scrapes, but there are allegations being made that this is occurring with abandon. Data of "medical record photographs of thousands of . . . people" has been scraped.[35] In one lawsuit, companies 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."[36] Companies like ClearviewAI and PimEyes have scraped billions of photos to power facial recognition tools.[37]

When personal data is involved in scraping, there is a different dynamic than when it is not. Personal data implicates the privacy of the individuals whose data is scraped, and these individuals are not the scrapers or scrapees. These individuals are thus another party – a stakeholder with vital interests, as their data is at stake. But as we will discuss later on, the interests of these individuals are not being sufficiently represented in the battles over scraping.

4. The Ethical Twilight of Scraping

Scraping grew up with the internet. Scraping has been loved and reviled, tolerated as a necessary evil and attacked as an unwanted pack of scavengers. Scraping has long occurred on a shifting technological plane and a swampy uncertain legal landscape.

Scraping personal information exists in weird ethical twilight – the practice is dicey, yet it is neither blessed nor fully condemned. According to the Imperva *Bad Bots Report,* "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)."[38] This definition could technically extend to most scraping of personal data.

As journalist Adrienne LaFrance writes, bad bots "include unauthorized-data-scrapers, spambots, and scavengers seeking security vulnerabilities to exploit."[39] The key question is what an "unauthorized" data scraper is, as most data scrapers

---

[35] Lauren Leffer, *Your Personal Information ss Probably Being Used to Train Generative AI Models*, SCIENTIFIC AMERICAN (Oct. 19, 2023), https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/.
[36] Isaiah Poritz, *OpenAI's Legal Woes Driven by Unclear Mesh of Web-Scraping Laws*, BLOOMBERG LAW (July 5, 2023, 5:44 AM), https://news.bloomberglaw.com/ip-law/openais-legal-woes-driven-by-unclear-mesh-of-web-scraping-laws.
[37]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'*, BUSINESS INSIDER (Apr. 2, 2023), https://www.businessinsider.com/clearview-scraped-30-billion-images-facebook-police-facial-recogntion-database-2023-4.
[38] IMPERVA, *supra* note 31.
[39] Adrienne LaFrance, *The Internet Is Mostly Bots*, THE ATLANTIC (Jan. 31, 2017), https://www.theatlantic.com/technology/archive/2017/01/bots-bots-bots/515043/.

do not ask for permission; they scrape unless they are told not to scrape or are blocked from scraping. Uncertainty abounds as to the meaning of "unauthorized."

The ethical ambiguity of scrapers is reflected by the metaphors used to describe the internet and scraping. As Andrew Sellars notes, scrapers have been "likened to an invading army of robots, a vandal taking hammer to a piece of machinery, a person walking into a bank with both a safety deposit key and a shotgun – or, more innocently, a roving machine that constantly takes photographs, an interviewer using an audio recording instead of taking notes, or a person who records signs posted within a store."[40]

Is scraping just innocent data gathering? Much data online is essentially offered up to the public, where some might compare it to placing it on public placards.

But when the internet is viewed with a property lens, the normative valence changes. If the internet is viewed as a form of "space," then scrapers might be considered unwanted intruders, as an invading horde of scavengers trespassing onto scrapee territory. If data is viewed as a form of property, then scrapers are stealing.[41]

Another lens with which to see scraping is norms. Scraping might be considered a norm violation, a form of rude socially-disfavored behavior. Consider an analogy to free food samples in a supermarket. If someone systematically eats all the samples as their meal, such a practice contravenes the unwritten norm that samples are for tasting and should be eaten in moderation. Is the gluttony of scraping a norm violation? Perhaps, but the law often does not penalize many norm violations.

Part of the challenge of addressing scraping is that so many metaphors can apply to the internet. Many of these metaphors work to some degree, but they also fail to capture the unique qualities of the internet, which aren't readily analogizable to the physical world or existing concepts. Scraping has an ambiguous ethical valence because it is akin to so many things, yet different. Scraping is not all bad, but it is also not all good.

Perhaps the key 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.[42] Scraping dramatically lowers 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

---

[40] Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372, 383 (2018).

[41] *See, e.g.*, Julie Cohen, *Cyberspace As/And Space*, 107 COLUM. L. REV. 210 (2007); Orin Kerr, *The Problem of Perspective in Internet Law*, 91 GEO. L. J. 357 (2003); Pamela Samuelson, *Privacy as Intellectual Property?*, 52 STAN. L. REV. 1125 (1999); Dan Hunter, *Cyberspace as Place and the Tragedy of the Digital Anticommons*, 91 CALIF. L. REV. 439 (2003).

[42] James J. Gibson, *The Theory of Affordances, in* The Ecological Approach to Visual Perception (2014); *see also* DON NORMAN, THE DESIGN OF EVERYDAY THINGS (1988); Ryan Calo, *Privacy, Vulnerability, and Affordance*, 66 DEPAUL L. REV. 591, 601–03 (2016); WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW 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).

individual (and manual/non-automated) access. 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, a war 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" – the various strategies and technologies to scrape and to defend against scraping. Many organizations have an incentive to scrape; but many organizations have an incentive to not be scraped.[43] Being scraped 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. Meta once hired a company to scrape on its behalf, then ended up suing the company when it began to scrape Meta's data. [44]

Many sites now include statements in their terms of service that users agree not to scrape without permission.[45] For example, Microsoft recently "updated its general terms of use to prohibit scraping, harvesting, or similar extraction methods of its AI services," even though Microsoft's affiliate OpenAI has bots "designed to scrape the entire internet."[46]

Many sites want to be crawled, but 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 the crawlers because they want their information to be findable on the internet. AI scraping lacks this reciprocal benefit; it provides little benefit for the scrapees, and a more unilateral benefit for the scrapers.

Already, several companies have formed an industry association called the Mitigating Unauthorized Scraping Alliance (MUSA), which aims to "bring together leading companies to protect data from unauthorized scraping and misuse by identifying and promoting best practices to prevent unauthorized data scraping, educating the public on the harms of such scraping, and providing insight, knowledge, and expertise to policy makers around unauthorized scraping."[47]

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.

---

[43] McCarthy, *supra* note 29 .
[44] *Id.*
[45] 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, 257-58 (2018).
[46] *Id.*
[47] Mitigating Unauthorized Scraping Alliance (MUSA), https://antiscrapingalliance.org/ (last accessed July 3, 2024).

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 other forms of data, including personal data. This litigation has been ongoing for decades, but it has remained inconclusive. As Andew Sellars describes it, "the 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."[48]

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 case 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, and the current status of scraping under the law remains a murky gray zone. An apt analogy might be made to the Crusades – years of battles and bloodshed, wins and losses, but ultimately, no definitive resolution.

Third, most of the cases have involved claims related to property and contract, not privacy. Indeed, 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."[49] The CFAA applies regardless of the purpose of access.[50]

Civil liability under the CFAA is limited by a requirement of a loss caused by scraping. Courts have reached mixed conclusions about the theory of loss.[51] Generally, however, the "loss" threshold of $5,000 in a one-year period is often readily established because expenses to investigate scraping activity count for a loss.[52]

---

[48] Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372, 377 (2018).
[49] 18 U.S.C. § 1030(a)(2)(C).
[50] Sellars, *Twenty Years of Web Scraping, supra* note 48, at 391.
[51] Zachary Gold & Mark Latonero, *Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping,* 13 WASH. L.J. TEC. & ARTS 275, 296 (2018).
[52] Sellars, *Twenty Years of Web Scraping, supra* note 48, at 376.

Over the course of several decades, many cases about scraping or relevant 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 usually works less controversially when a hacker breaks into a computer system by bypassing technical protections like encryption and password prompts. In these circumstances, a computer system most 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 (or at least imagined audiences). 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 to cease-and-desist. Complicating matters is the fact that sites want the data to be accessed – this is essential for users of the site – but they just do not want scrapers to access the data. Sites want bots to gather data for some purposes but not others.

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.[53] When scraping occurs in violation of website 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 "[t]he use of a scraper tool to collect pricing information does not automatically exceed the authorized access of a website unless the website owner publishes an explicit statement on the website restriking access."[54] Later cases 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."[55] Facebook sent a cease-and-desist letter to Power Ventures, and it later blocked Power Ventures' IP address, but Power Ventures changed its IP address to continue scraping. 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 was unauthorized access.[56]

Andy Sellars views these cases as changing, with the applicability of the CFAA to scrapers shifting like the wind blowing from different directions. From 2000-2009, he notes that courts were quick to find that scraping was unauthorized access.[57] In the early 2010s, there was a "slight trend towards limiting the law's application."[58] By the mid 2010s, courts embraced various indications of revocation of access as

[53] 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, *Twenty Years of Web Scraping, supra* note 48, at 380.
[54] EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003).
[55] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1062 (9th Cir. 2016).
[56] *Id.* at 1061-68.
[57] Sellars, *Twenty Years of Web Scraping, supra* note 48, at 393-94.
[58] *Id.* at 396.

making scraping fall within the CFAA's prohibited unauthorized access.[59] By the late 2010s, courts were back on the side of the scrapers.[60]

Parts of the Ninth Circuit's opinion in *hiQ Labs v. LinkedIn* represents a big CFAA victory for scrapers. Originally decided in 2019, the case was vacated by the Supreme Court and affirmed again on remand in 2022.[61] On LinkedIn, people post profiles with their professional resumes and write short posts or longer articles. hiQ is a data analytics company that began scraping public Linkedin user profiles. It then used the data to develop a "people analytics" algorithm that it marketed to businesses. hiQ identified employees who were likely to be recruited by others so employers could take steps to retain them. hiQ also identified "skill gaps" in a business's workforce. LinkedIn prohibited scraping in its user agreement. It took many technical steps to prevent scraping.[62]

LinkedIn sent hiQ a cease-and-desist letter, claiming that hiQ was violating the CFAA among other laws, and that hiQ's scraping was in violation of LinkedIn's user agreement. hiQ sued for a preliminary injunction to not only declare that its scraping was legal under the CFAA but that LinkedIn remove any technical barriers to its scraping. On the CFAA, the court held that "the CFAA is best understood as an anti-intrusion statute and not as a ''misappropriation statute.'"[63] Because the LinkedIn profiles were publicly available, the court reasoned, hiQ was not breaking and entering. hiQ was not 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."[64]

During the hiQ litigation, the U.S. Supreme Court decision in *Van Buren v. United States,*[65] 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."[66] In other words, there must be some kind of proceeding beyond a gate for access to be unauthorized.

According to Professor Orin Kerr, CFAA cases have found a lack of authorized access based on the "intended function" of technology, misconduct, or breach of an agreement.[67] But ultimately, even after *Van Buren,* Kerr views the law as only

---

[59] *Id.* at 401-07.
[60] *Id.* at 408-12.
[61] The original decision, hiQ Labs, Inc. v. LindkenIn 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, 141 S. Ct. 1648 (2021). *See* LinkedIn Corp. v. hiQ Labs, Inc., 141 S. Ct. 2752 (2021). On remand, the 9th Circuit affirmed its original decision. *See* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.3d 1180 (9th Cir. 2022). For more background on this case, *see* 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).
[62] hiQ Labs, Inc. v. LinkedIn Corp., 31 F.3d 1180, 1086 (9th Cir. 2022).
[63] *Id.* at 1197.
[64] *Id.* at 1201.
[65] 141 S.C t. 1648 (2021).
[66] *Id.* at 1658–59. For more on this concept, see Patricia L. Bellia, *A Code-Based Approach to Unauthorized Access Under the Computer Fraud and Abuse Act*, 84 GEO. WASH. L. REV. 1442 (2016).
[67] Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse*

partially focused.[68] *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.[69]

(b) Business and Property Interests

Beyond the CFAA, litigation over scraping has used various torts 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 ones most likely to succeed have been "breach of contract, tortious interference with a contract, and unjust enrichment."[70]

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."[71] Plaintiffs advanced the theory that scraping impairs scrapees by consuming network and server resources.[72]

In *eBay v. Bidder's Edge,*[73] an early case decided in 2000, a district court held that scraping information about bids on eBay was trespass and issued an injunction against Bidder's Edge. Although Bidder's 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.[74]

But subsequent courts made it harder for scrapees to establish a trespass to chattels; courts concluded that mere data gathering, without harm was insufficient.[75] 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.[76]

Statutes, 78 N.Y.U. L. Rev. 1596 (2003). For another analysis of the caselaw, see Patricia L. Bellia, *A Code-Based Approach to Unauthorized Access Under the Computer Fraud and Abuse Act*, 84 Geo. WASH. L. REV. 1442 (2016).

[68] Orin S. Kerr, *Focusing the CFAA in Van Buren*, 2021 SUP. CT. REV. 155, 156 (2022).

[69] *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, *Norms of Computer Trespass*, 116 COLUM. L. REV. 1143, 1161 (2016).

[70] Kieran McCarthy, "Web Scraping for Me, But Not for Thee," TECHNOLOGY & MARKETING LAW BLOG (Aug. 24, 2023), https://blog.ericgoldman.org/archives/2023/08/web-scraping-for-me-but-not-for-thee-guest-blog-post.htm.

[71] RESTATEMENT (SECOND) OF TORTS §§ 217(b), 218(b)-(c) (Am. Law. Inst. 1965).

[72] 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, 265 (2018) ("Trespass to chattels is commonly argued in data scraping cases, under the theory that a defendant's scraping interfered with a plaintiff's use of its website and servers by consuming intangible resources such as network and server capacity.").

[73] *eBay, Inc. v. Bidder's Edge, Inc.*, 100 F. Supp.2d 1058 (N.D. Cal. 2000).

[74] *Id.* at 1064.

[75] *Ticketmaster Corp. v. Tickets.com, Inc.*, No. CV997654HLHVBKX, 2003 WL 21406289 (C.D. Cal. Mar. 7, 2003).

[76] 71 P.3d 296, 299 (Cal. 2003).

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."[77]

Intellectual 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.[78] For example, Alan Westin argued: "[P]ersonal information, thought of as the right of decision over one's private personality, should be defined as a property right."[79] Lawrence Lessig argues that privacy should be protected as a property right because a property regime provides "control, and power, to the person holding the property right."[80]

Property analogies break down because personal data is often shared, yet it is non-rivalrous, meaning when one person has it, it doesn't stop others from having it (or keeping it) as well.[81] Additionally, property law often focuses on the value of personal data, and courts have concluded that the value of compilations of personal data are created by the compiler, not the individuals 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, "Defendants 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."[82]

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. However, there are several limitations with copyright law.[83] First, copyrighted content can be used without permission in circumstances called "fair use." Indeed, scraping is creating new questions and challenges for copyright law, especially with Generative AI.[84] Second, much personal data is not owned by the individual to whom it pertains. The taker of a photograph, not the subject, has the

---

[77] Zachary Gold & Mark Latonero, *Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping,* 13 WASH. L.J. TEC. & ARTS. 275, 295 (2018).
[78] *See* 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").
[79] ALAN WESTIN, PRIVACY AND FREEDOM 324 (1967).
[80] LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999).
[81] 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."); JAMES BOYLE, THE PUBLIC DOMAIN (2008).
[82] 652 N.E.2d 1351 (Ill. App. 1995).
[83] *See, e.g.*, Eric Goldman & Jessica Silbey, *Copyright's Memory Hole*, 2019 BYU L. REV. 929 (2020).
[84] Scraping itself would likely not infringe upon copyright, only certain uses of scraping data. *See* Sobel, *supra* note 64, at 170-72; *see also* Pamela Samuelson, *Fair Use Defenses in Disruptive Technology Cases*, UCLA L. REV. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4631726; Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM LAW REVIEW 1887 (2024); Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUSTON LAW REVIEW, Vol. 61, No. 2 (2023).

copyright.[85] The author of a biography owns the copyright, not the subject. Third, most personal data is not copyrightable, as facts cannot be copyrighted.[86]

Another 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,[87] but the status of the terms of service as a contract remains unclear.[88]

(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 violating the Illinois Biometric Information Privacy Act ("BIPA").[89] 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.[90] 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. It must refrain from granting access to Illinois government or private entities.[91] Additionally, Clearview must allow Illinois residents to opt out of being searchable in its database. 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.[92] Many privacy torts will likely prove to be ineffective against scraping. The public disclosure of private facts tort and the false light tort both require widespread dissemination of information, and scraping involves data collection, thus making these torts inapplicable.[93] The tort of intrusion upon seclusion likely will fail because the data scraped is publicly available.[94]

---

[85] *Mannion v. Coors Brewing Cor.*, 377 F. Supp.2d 444, 454-55 (S.D.N.Y. 2005).
[86] *See, e.g.*, Feist v. Rural Publications (US 1991); 17 USC 102(b); *see also* Jessica Silbey, *A Matter of Facts: The Evolution of Copyright's Fact-Exclusion and Its Implications for Disinformation and Democracy*, J. COPYRIGHT SOCIETY OF USA, Vol. 71, No. 3 (forthcoming 2024).
[87] Meta Platforms, Inc. v. Brandtotal Ltd., 20-cv-07182-JCS (N.D. Cal. May. 27, 2022).
[88] DANIEL J. SOLOVE & PAUL M. SCHWARTZ, INFORMATION PRIVACY LAW 730-733 (8th ed. 2024).
[89] ACLU v. Clearview AI, Inc., No. 2020 CH 04353, 2022 Ill. Cir. LEXIS 288.
[90] 740 ILL. COMP. STAT. 14/15(b)(1)-(3).
[91] Consent Order, ACLU v. Clearview AI, Inc., No. 2020 CH 04353, 2022 Ill. Cir. LEXIS 2887.
[92] 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.
[93] *See* Restatement (Second) of Torts §§ 652D, 652E.
[94] *See* 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); Health v. Playboy Enterprises, Inc., 732 F. Supp. 1145 (S.D.

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.[95] However, of all the torts, rights of misappropriation and publicity might be most helpful with respect to images and videos of people's names and likeness.[96] 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 collection and use of faceprints survived a motion to dismiss.[97] Regarding the sufficiency of the pleadings regarding a claim for misappropriation, the Superior Court of California (Alameda) 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 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.[98]

Recently, a major class action was launched against OpenAI's scraping for its AI chatbot ChatGPT 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 Biometric Information Privacy Act, and state unfair and deceptive act or practice (UDAP) statutes.[99] 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 land mine.

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."[100] Even when companies say they are fighting scrapers, they are

---

Fla. 1990) (holding that there is no privacy interest in facts already publicized).

[95] SOLOVE & SCHWARTZ, INFORMATION PRIVACY LAW, *supra* note 150, at 193-95.

[96] *See, e.g.,* Jason Schulz, *The Right of Publicity: A New Framework for Regulating Facial Recognition*, 88 BROOKLYN L. REV. 1039 (2023).

[97] Court Decision on Anti-SLAPP Motion and Demurrer, Renderos v Clearview AI, No. RG21096898 (Sup. Ct. Cal. Nov. 18, 2022), https://static1.squarespace.com/static/62c3198c117dd661bd99eb3a/t/637d2d6a87725b11dd104531/1669148010542/ANTISLAPPClearview.pdf.

[98] *Id.*

[99] Class Action Complaint, P.M. et al. v. OpenAI, Case No. 3:23-cv-03199 (N.D. Cal. June 28, 2023).

[100] Kathleen C. Riley, *Data Scraping as a Cause of Action: Limiting Use of the CFAA and Trespass in*

often pursuing their own competitive advantage and using "privacy" as a pretext.[101] 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 and concluded that the business interests of one company outweighed them:

> [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.[102]

The vast majority 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 website 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. A few examples include access restrictions, Captchas, rate limiting, browser fingerprinting, and banning user's accounts and IP addresses.[103] But these measures can be circumvented. Scraping and preventing scraping is ultimately a cat-and-mouse game.

For a long time, social media platforms offered APIs to facilitate third-parties' use of data.[104] "APIs are code interfaces that allow programmers to make very formal

---

Online Copying Cases, 29 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 245, 250 (2018).

[101] 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.").

[102] *hiQ Labs, Inc. v. LinkedIn Corp.*, 31 F.3th 1180 (9th Cir. 2022).

[103] Margaret Rouse, *Web Scraping*, TECHOPEDIA (Feb. 8. 2023), https://www.techopedia.com/definition/5212/web-scraping; Michael Nyamande, *Web Scraping Without Getting Blocked*, https://brightdata.com/blog/web-data/web-scraping-without-getting-blocked; 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; Jeffrey Kenneth Hirschey, *Symbiotic Relationships: Pragmatic Acceptance of Data Scraping*, 29 BERKLEY. TECH. L.J. 897, 918 (2014).

[104] GREG ELMER, GANAELE LANGLOIS, JOANNA REDDEN, COMPROMISED DATA: FROM SOCIAL MEDIA TO BIG Data, 120 (Bloomsbury 2015).

data requests from websites within a specific interface."[105] But in 2018, the Cambridge Analytica scandal changed views about the costs and benefits of allowing API access.[106] In the wake of this incident, many social media companies curtailed their APIs,[107] or increased their cost to discourage improper uses.[108] 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.[109] Several large media companies have blocked OpenAI's scraping bots.[110]

But not all scrapers play the game of chivalry. As David Peirce observes, "The robots.txt file governs a give and take; AI feels to many like all take and no give.…And the fundamental agreement behind robots.txt, and the web as a whole — which for so long amounted to "everybody just be cool" — may not be able to keep up either."[111] Web scrapers now also often use "additional technologies to mimic human browsing and delve deeper into each website."[112] The *New York Times* contends its site is still being scraped contrary to its robots.txt instructions.[113] Some scrapers have found ways to evade paywalls on websites.[114]

Meta declared that it has implemented "several measures…to mitigate the risk of scraping on our platform."[115] 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."[116] Additionally, it imposes "rate and data limits, which are designed to restrict how much data a single person can obtain through a certain feature."[117] Meta also notes that it has initiated hundreds of enforcement actions, such as "sending cease and desist letters,

---

[105] Jeffrey Kenneth Hirschey, *Symbiotic Relationships: Pragmatic Acceptance of Data Scraping*, 29 BERKLEY TECH. L.J. 897, 905 (2014).
[106] Domenico Trezza, *To Scrape or Not To Scrape, This Is Dilemma. The Post-API Scenario and Implications On Digital Research*, FRONTIERS IN SOCIOLOGY (2023), https://www.frontiersin.org/articles/10.3389/fsoc.2023.1145038/full.
[107] Trezza, *To Scrape, supra* note X.
[108] Andrew Hutchinson, *Twitter Implements Usage Limits for All to Combat Data Scrapers*, SOCIAL MEDIA TODAY (July 1, 2023), https://www.socialmediatoday.com/news/twitter-implements-usage-limits-combat-data-scrapers/684831/.
[109] Ben Wodecki, *OpenAI Quietly Unveils Web Crawler to Scrape Data for Its AI Models*, AI BUSINESS (Aug. 8, 2023), https://aibusiness.com/nlp/openai-unveils-web-crawler-to-gather-data-to-improve-ai-models#close-modal.
[110] 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, 10:17 PM), https://www.cnn.com/2023/08/28/media/media-companies-blocking-chatgpt-reliable-sources/index.html.
[111] Pierce, *supra* note 16.
[112] 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).
[113] Benj Edwars, *The New York Times prohibits AI vendors from scraping its content without permission*, ARS TECHNICA (Aug. 14, 2023, 12:21 PM), https://arstechnica.com/information-technology/2023/08/the-new-york-times-prohibits-ai-vendors-from-devouring-its-content/.
[114] Lauren Leffer, *Your Personal Information Is Probably Being Used to Train Generative AI Models*, SCIENTIFIC AMERICAN (Oct. 19, 2023),
[115] Mike Clark, *Scraping by the Numbers*, META (May 19, 2021), https://about.fb.com/news/2021/05/scraping-by-the-numbers/.
[116] *Id.*
[117] *Id.*

disabling accounts, filing lawsuits or requesting assistance from hosting providers to get them taken down."[118] Meta states that it blocks "billions of suspected scraping actions per day across Facebook and Instagram."[119]

Battles over scraping will continue on the 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

In the midst of the Scraping Wars, market alternatives have been arising. Scrapers are starting to reach deals with scrapees, paying them for the right to scrape their land, or obtain 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.[120] In 2023, Open AI reached deals with the Associated Press and Axel Springer, parent company of Politico and Business Insider.[121] Personal data is implicated in these deals, as news stories have extensive personal data. OpenAI also reached a deal with Shutterstock, a site where users buy and sell images.[122] 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. REGULATORY INTERVENTION

Despite the fact that scraping has been occurring for a long time, regulators have generally avoided stepping onto the battlefield. Recently, however, some regulators have begun to tepidly step into the fray, but they have found themselves ill-prepared for life on the battlefront.

1. EU Data Protection Law

It is quite difficult to reconcile scraping personal data with the EU's GDPR, which requires a legal basis for data processing and imposes various transparency and autonomy-enhancing safeguards.

---

[118] *Id.*
[119] *Id.*
[120] 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/.
[121] Thomas Barrabi, *OpenAI Offering Media Outlets as Little as $1M to Use News Articles for AI Models*, N.Y. POST (Jan. 4, 2024); *see also* Gerrit De Vynck, *OpenAI Strikes Deal With AP to Pay for Using Its News in Training AI*, THE WASH. POST (last updated July 13, 2023); Matt O'Brien, *ChatGPT-maker OpenAI signs deal with AP to license news stories*, ASSOCIATED PRESS (July 13, 2023).
[122] *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.

Under the GDPR, there is no general exception for publicly available information.[123] Instead, personal data can be collected and processed based on one of six lawful bases: (1) consent; (2) necessary for a contract; (3) necessary to comply with a legal obligation; (4) necessary to protect a person's vital interests; (5) necessary for the public interest; and (6) necessary for legitimate interests and not "overridden by the interests or fundamental rights and freedoms of the data subject."[124]

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.[125] Given the vast number of individuals involved, obtaining the consent of each person is practically impossible.

The lawful basis that most seemingly fits – legitimate interests – is far from a reliable basis. 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 fundamental rights and freedoms. Third, sensitive data cannot be processed for legitimate interests. 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.[126]

In March 2023, in a dramatic and bold move, the Italian Data Protection Authority (DPA) banned ChatGPT. The DPA stated that "there appears to be no legal basis underpinning the massive collection and processing of personal data in order to 'train' the algorithms on which the platform relies."[127]

But soon afterward, in a rather awkward walk-back, the DPA then reinstated ChatGPT.[128] 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. The ban on ChatGPT was lifted in late April of 2023.[129] Regarding the legal basis for processing, the Italian DPA stated that OpenAI would need to rely on either consent or legitimate interests as the applicable legal basis for processing under the GDPR.[130] As an article in *The Verge*

---

[123] 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." GDPR art. 9.
[124] GDPR art. 6.
[125] 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/.
[126] Daniel J. Solove, *Data Is What Data Does: Regulating Use, Harm, and Risk Instead of Sensitive Data*, 118 Nw. U. L. Rev. 1081 (2024).
[127] *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 (English translation).
[128] *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.
[129] K.C. Halm, John D. Seiver & Patrick J. Austin, *Italy's Data Protection Agency Lifts Ban on ChatGPT*, DAVIS WRIGHT TREMAINE LLP, (May 15, 2023), https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2023/05/ai-chatgpt-italy-ban-lifted.
[130] *ChatGPT: Italian SA to lift temporary limitation if OpenAI implements measures – 30 April set as*

appropriately put it, "So far, none of these changes seem to dramatically modify how ChatGPT operates in Italy."[131]

Thus, scraping continues in the EU, however, a full showdown between GDPR and scrapers is near. In a recent guide on scraping personal data, the Dutch data protection authority Autoriteit Persoonsgegevens (AP) held that scraping of personal information is almost always a violation of the GDPR.[132] 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, with the aim of determining whether or not those people will receive requested insurance.[133] In practice, the AP said that the only practical 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."[134] 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.[135] If the rest of the DPAs in the EU hold the same opinion, this would essential prohibit scraping for profit by commercial entities, which would be a dramatic prohibition.

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.[136]

The commissioners stated that "websites should implement multi-layered technical and procedural controls to mitigate the risks."[137] Interestingly, the joint statement did not focus on the scrapers and their violations of privacy law or on how the

---

deadline for compliance, GARANTE PER LA PROTEZIONE DEI DATI PERSONALI, (Apr. 12, 2023), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9874751.

[131] Adi Robertson, *ChatGPT Returns to Italy After Ban*, THE VERGE (Apr. 28, 2023).

[132] *scraping bijna altijd illegal [Scraping is almost always illegal]*, AUTORITEIT PERSOONSGEGEVENS [DUTCH DATA PROTECTION AUTHORITY] (May 1, 2024), https://autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal (Neth.).

[133] *Id.*

[134] *Id.*

[135] *Id.* Some of the examples the AP gave of exceptional cases when there might be a legitimate interest in scraping would be 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.

[136] Joint Statement on Data Scraping and the Protection of Privacy (Aug. 24, 2023), https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf.

[137] *Id.*

commissioners would enforce the laws against scrapers as well as 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 that Clearview 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. Moreover, Clearview failed to comply with the conditions for lawful processing of sensitive data and failed to provide information to data subjects about the data processing. The ICO also 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.[138]

Throughout the EU, Clearview has sparked a series of enforcement actions. In 2022, France's CNIL fined Clearview 20 million euros, the maximum GDPR penalty, when Clearview failed to comply with a 2021 injunction.[139] Italy also imposed the same fine in 2022, ordering Clearview to cease scraping and delete all data from people in Italy.[140] Likewise, in 2022, Greece's data protection authority issued a 20 million euro fine and similar order to cease and delete.[141] In 2023, Austria's data protection authority found Clearview to be in violation of the GDPR, but just issued an order to delete the data but did not issue a fine.[142]

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 law already has some tools to regulate scrapers and scrapees. Most notably, scraping as well as the failure to defend against scraping could constitute violations of the FTC Act Section 5, which prohibits "unfair or deceptive" acts or practices. The FTC has been enforcing the FTC Act Section 5 for privacy violations for several decades. The FTC has ample jurisprudence 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

---

[138] [2023] UKFTT 819 (GRC); *Tribunal Overturns UK ICO's Enforcement Action Against Clearview AI*, DECHERT (Nov. 8, 2023), https://www.dechert.com/knowledge/onpoint/2023/11/tribunal-overturns-uk-ico-s-enforcement-action-against-clearview.html.

[139] EUROPEAN DATA PROTECTION BOARD, *The French SA Fines Clearview AI EUR 20 Million*, (Oct. 20, 2022), https://www.edpb.europa.eu/news/national-news/2022/french-sa-fines-clearview-ai-eur-20-million_en.

[140] EUROPEAN DATA PROTECTION BOARD, *Facial Recognition: Italian SA Fines Clearview AI EUR 20 Million*, (Mar. 10, 2022), https://www.edpb.europa.eu/news/national-news/2022/facial-recognition-italian-sa-fines-clearview-ai-eur-20-million_en.

[141] EUROPEAN DATA PROTECTION BOARD, *Hellenic DPA Fines Clearview AI 20 Million Euros*, (July 20, 2022), https://www.edpb.europa.eu/news/national-news/2022/hellenic-dpa-fines-clearview-ai-20-million-euros_en.

[142] European Data Protection Board, "Decision by the Austrian SA Against Clearview AI Infringements of Articles 5, 6, 9, 27 GDPR" (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.

reasonably avoidable by consumers themselves and is not outweighed by countervailing benefits to consumers or to competition."[143]

Arguments can certainly be made 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. Arguments can be made 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.[144]

If the FTC were to find scrapers in violation of the FTC Act, the FTC could require the deletion of models developed with improperly-gathered data.[145] But it is hard to imagine the FTC would be so bold as to find that scraping violated 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 dealt the FTC a severe setback for its efforts to regulate advertising to children in the 1970s. 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. 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 because it could contravene promises in a privacy notice that data will be protected by reasonable data security, that data will not be transferred to third parties, that data will only be used for specified purposes, and so on. The failure to protect against scraping could also be an unfair practice – as could the scraping itself.

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.[146]

E. 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

---

[143] 15 U.S.C. § 45(n).
[144] In re Vision I Properties (FTC 2005).
[145] For an example of the FTC's requiring algorithmic destruction, see In re Everalbum, Inc. (FTC 2022). 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. In what she calls an "algorithmic shadow," the data has a "persistent imprint" in the machine learning algorithm. Merely deleting the data does not delete the algorithmic shadow and has "no impact on an already trained model." Tiffany C. Li, *Algorithmic Destruction*, 75 SMU L. REV. 479, 482, 498 (2022).
[146] Alicia Solow-Niederman, *The Overton Window and Privacy Enforcement,* 34 HARV. J. L. & TECH. (forthcoming 2024).

exist.

Litigation over scraping is likely to go on for years and years and implicate a panoply of causes of action. Countless questions remain about whether these torts would lead to a desirable regulatory regime.[147]

Scraping has been occurring in the shadows of the law. Scraping has been recognized as a dubious practice, but it rarely has been confronted by privacy law. It remains a practice that occurs in the dark of the night, with hardly anyone shining the spotlight on it. But this situation is untenable. The Scraping Wars are breaking out, and the problems posed by scraping are no longer possible to deny or ignore.

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 such a theory is the most practical and sound way forward. This does not mean that individual lawsuits will always be unhelpful. But even a bottom-up approach will benefit from some top-down thinking and an overall direction. Additionally, smaller websites might lack the resources to litigate against scrapers.[148]

Moreover, many of the legal and technological mechanisms employed in the Scraping Wars fail to involve the individuals whose data is being fought over. Individuals do not set the robots.txt files for websites containing their data nor control the terms of service of platforms. It is up to the website operators to implement technical anti-scraping measures. 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.

---

[147] *See* Benjamin L.W. Sobel, *A New Common Law of Web Scraping,* 25 LEWIS & CLARK L. REV. 147 (2021) (arguing that no common law torts can adequately address scraping and proposing a new tort of bad faith breach of terms of service).

[148] Zachary Gold & Mark Latonero, *Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping,* 13 WASH. L.J. TEC. & ARTS 275, 298-99 (2018).

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.[149] An early version of the foundational FIPPs was articulated in 1973 and then expanded in the OECD Privacy Guidelines of 1980.[150] The FIPPs have been the backbone of privacy laws around the world, as well as countless privacy frameworks, standards, and codes.[151]

These principles were developed to respond to fears about the power of digital databases to make information much easier to collect, store, aggregate, search, and share. 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.[152] The FIPPs are the beating heart of virtually every data protection law in the world, including the EU's GDPR and most U.S. federal and state privacy laws.[153]

In this Part, we argue that scraping of personal data is incompatible with nearly all of the FIPPs as well as 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 as well as with the very 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 crashing together between scraping and privacy, yet most policymakers, commentators, and organizations seem unaware of that this is a crisis. Scraping and the core model of most privacy laws are fundamentally incompatible, and radical changes must be made to scraping, privacy law, or both.

---

[149] *See* COLIN J. BENNETT & CHARLES D. RAAB, THE GOVERNANCE OF PRIVACY: POLICY INSTRUMENTS IN GLOBAL PERSPECTIVE 12 (2006); GRAHAM GREENLEAF, ASIAN DATA PRIVACY LAWS: TRADE AND HUMAN RIGHTS PERSPECTIVES 6-7 (2014). *See generally* CHRISTOPHER KUNER, EUROPEAN DATA PROTECTION LAW: CORPORATE COMPLIANCE AND REGULATION (2d ed. 2007); 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*, Policy@Intel (Jan. 28, 2016), http://blogs.intel.com/policy/2016/01/28/blah-2/; Robert Gellman, *Fair Information Practices: A Basic History* (unpublished manuscript), http://bobgellman.com/rg-docs/rg-FIPshistory.pdf.

[150] DANIEL J. SOLOVE & PAUL M. SCHWARTZ, INFORMATION PRIVACY LAW 580-81 (8th ed. 2024).

[151] *See* COLIN J. BENNETT & CHARLES D. RAAB, THE GOVERNANCE OF PRIVACY: POLICY INSTRUMENTS IN GLOBAL PERSPECTIVE 12 (2006); GRAHAM GREENLEAF, ASIAN DATA PRIVACY LAWS: TRADE AND HUMAN RIGHTS PERSPECTIVES 6-7 (2014). *See generally* CHRISTOPHER KUNER, EUROPEAN DATA PROTECTION LAW: CORPORATE COMPLIANCE AND REGULATION (2d ed. 2007); 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*, Policy@Intel (Jan. 28, 2016), http://blogs.intel.com/policy/2016/01/28/blah-2/; Robert Gellman, *Fair Information Practices: A Basic History* (unpublished manuscript), http://bobgellman.com/rg-docs/rg-FIPshistory.pdf.

[152] *Id.*

[153] *See, e.g.*, BENNETT & RAAB, *supra* note 149, at 12; Greenleaf, *Asian Data Privacy*, *supra* note 149, at 6-7.

A. SCRAPING AND PRIVACY PRINCIPLES

The FIPPs create a vision for data privacy built on fairness, individual autonomy, and processor accountability. Scraping doesn't work with this model of privacy protection; trying to fit scraping into this model is akin to trying to pound a square peg into a round hole. Specifically, scraping violates several fundamental privacy principles of (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.

1. Fairness

The overarching goal of the FIPPs is fairness, which is why they are called the *Fair* Information Practice Principles. Fairness is rather vast concept, and in the context of privacy, fairness has many components. The FIPPs, for example, encompass numerous principles. According to the UK ICO, "fairness 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."[154] Similar concerns animate the Federal Trade Commission's regulation of unfair and deceptive trade practices.[155] The FTC's definition of unfairness is a practice that "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition."[156]

Although subject to many different definitions and containing many disparate elements, one can generally claim that fairness is a robust and far-reaching set of requirements protecting both collective groups and individuals from unwarranted harm.[157] Gianclaudio Malgieri has argued that "fairness is effect-based: what 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.'"[158] Under the broad conception of the FIPPs, fairness also involves the responsible collection and processing of personal data as well as respecting the interests of the individuals to whom the data pertains.

---

[154] United Kingdom Information Commissioner Office, "Principle (a): Lawfulness, fairness and transparency," https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/the-principles/lawfulness-fairness-and-transparency/ (hereinafter UK ICO, "Lawfulness").

[155] FTC Act Section 5; *see also* Daniel J. Solove & Woodrow Hartzog, *FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

[156] 15 U.S.C. § 45(n).

[157] Gianclaudio Malgieri, *The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation*, Proceedings of FAT* '20, January 27–30, 2020. ACM, New York, NY, USA, 14 pages. DOI: 10.1145/3351095.3372868 ("[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.").

[158] *Id.* at 2.

Scraping violates the fairness principle because it is hidden and harmful. In a joint statement, data protection authorities (DPAs) from around the world found that scraped data can be used for cyberattacks, identity fraud, profiling, surveillance, unauthorized intelligence gathering, and spam.[159] People are not notified when their data is scraped, which often leaves people exposed and worse off.

2. Individual Rights and Control

Another central privacy principle involves ensuring individuals have some control in how their data is collected and used. This goal is often referred to in broader autonomy-focused concepts like "informational self-determination."[160] To implement this principle, most privacy laws require some form of consent to collect personal data – either express consent (opt in) or implied consent (opt out).[161] Scraping, however, mostly occurs without any form of individual consent.

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.[162] In their 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."[163]

People's privacy expectations depend upon the specific situation in which data is shared; privacy expectations are *contextually dependent*.[164] A diverse set of

[159] Joint Statement on Data Scraping, *supra* note 136.

[160] 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, and Sjaak Houwt, eds. 2009).

[161] 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* Daniel J. Solove, *Data Is What Data Does: Regulating Use, Harm, and Risk Instead of Sensitive Data*, 118 NW. U. L. REV. 1081 (2024).

[162] *See, e.g.*, Antje Niemann and Manfred Schwaiger, *Consumers' Expectations of Fair Data Collection and Usage – A Mixed Method Analysis*, 2016 49TH HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES (HICSS) 3646 (2016) ("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, and Xuyun Zhang, *Users' Perceived Control, Trust and Expectation on Privacy Settings of Smartphone*, NINTH INTERNATIONAL CYBERSPACE SAFETY AND SECURITY SYMPOSIUM 427 (2017); Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Cranor, and Norman Sadeh, *Privacy Expectations and Preferences in an IoT World*, 2017 USENIX Association 399 (2017); Igor Bilogrevic & Martin Ortlieb, *If You Put All The Pieces Together…: Attitudes Towards Data Combination and Sharing Across Services and Companies*, 2016 ASSOCIATION FOR COMPUTING MACHINERY 5215 (2016).

[163] Joint Statement on Data Scraping, *supra* note 136.

[164] *See, e.g.*, HELEN NISSENBAUM, PRIVACY IN CONTEXT (2009); Anne Adams, *Multimedia Information Changes the Whole Privacy Ballgame*, PROCEEDINGS OF THE TENTH CONFERENCE ON COMPUTERS, FREEDOM AND PRIVACY 25 (2000) (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*, COMMUNICATION THEORY 311 (1991); Sandra Petronio, *Brief Status Report on Communication Privacy Management Theory*, J. Family Comm. 6 (2013); Irwin Altman, *Privacy Regulation: Culturally Universal Or Culturally Specific?* J. SOC. ISSUES 66 (1977); IRWIN ALTMAN, THE ENVIRONMENT AND SOCIAL BEHAVIOR: PRIVACY, PERSONAL SPACE, TERRITORY, AND CROWDING (1975) (theorizing that privacy is not a static condition with universal rules,

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 environment.[165] 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.[166]

Scraping strips away the original context in which data is shared. All of the many factors which were present when people shared their data are missing with scraping. Thus, scraping thwarts people's privacy expectations and fails to respect people's initial decisions about how and when to share their personal data. Privacy law's goal of promoting informational self-determination cannot be achieved in a world of ubiquitous data scraping.

For better or for worse, 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.[167] 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.[168] Despite their limitations, privacy rights still serve important functions,[169] and they are central to the model of privacy protection established by most privacy laws. Scraping, however, takes all privacy rights away from individuals. When privacy rights can be readily extinguished, they become meaningless. For example, a right to delete personal data is ineffectual if it only applies at the original organization that has the data. With scraping, the data can exist in the clutches of thousands of other companies that scraped it, leaving individuals powerless to demand its deletion. The same situation applies to all other privacy rights, which are ignored by scrapers.

In short, scraping deprives people of control and renders rights meaningless. Ironically, the original organizations entrusted with people's data end up with far less power over the data than any random third party that scrapes the data. Scraping strips people of their rights and often places personal data outside the sphere of any privacy protection.

---

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).

[165] Alessandro Acquisti, *Privacy and Human Behavior in the Age of Information*, 347 SCIENCE 509 (2015); Alisa Frick *et al.*, *A Qualitative Model of Older Adults' Contextual Decision-Making About Information Sharing*, Proceedings of the 20th Annual Workshop on the Economics of Information Security (WEIS) (2020) (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).

[166] Ashwini Rao, Florian Schaub, Norman Sadeh, and Alessandro Acquisti, *Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online*, PROC. OF THE TWELFTH SOUPS 77 (2016) ("expectations influence decision making").

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

[168] 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).

[169] Margot E. Kaminski, *The Case for Data Privacy Rights (Or 'Please, a Little Optimism')*, 97 NOTRE DAME L. REV. REFLECTION 385 (2022).

3. Transparency

One of the core privacy principles involves transparency about personal data collection and usage. Nearly all privacy laws require that organizations inform individuals about the data gathered about them and from them, state the purposes of use, and describe their practices for protecting that data.[170]

None of this is happening with scraping. Scrapers just 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, such a message 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.[171] 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 to check.

Scraping also renders meaningless the transparency notice provided by the original collector of the data. This notice describes data practices of one organization prior to scraping; it fails 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 several circumstances, many privacy laws require consent for the collection and use of personal data.[172] 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.[173] Most of the U.S. 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.[174] 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 opt out.

In the EU, people have the right to withdraw their consent to the processing of their data.[175] Conceivably, data subjects would retain the ability 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

---

[170] *See* GDPR Art. 5(1)(a); Solove, *Limitations of Privacy Rights, supra* note 170, at 167 (discussing various right to information in many privacy laws).

[171] *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-data-scraping-on-watch-as-eu-dpa-lays-down-radical-gdpr-red-line/; PrivSecReport, *Rethinking 'Disproportionate Effort' exemption under GDPR for web-scraping*, GRC WORLD FORUMS (May 25, 2020), https://www.grcworldforums.com/gdpr/rethinking-disproportionate-effort-exemption-under-gdpr-for-web-scraping/344.article.

[172] Daniel J. Solove, *Murky Consent: An Approach to the Fictions of Consent in Privacy Law,* 104 B.U. L. REV. 593 (2024).

[173] *See e.g.,* 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); 47 U.S.C. §551(c)(2) (opt out).

[174] DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 186-190 (7th ed. 2024).

[175] GDPR art. 7(3). For an extensive background about the right to withdraw consent, see Marcu Florea, *Withdrawal of Consent for Processing Personal Data in Biomedical Research,* 13 INT'L DATA PRIVACY L. 107 (2023).

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.[176] Subsequent use for unrelated purposes requires people's consent, unless an exception applies. 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.[177] 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.[178]

Data privacy rules around the world require that entities should 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."[179] Data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."[180] Canada's PIPEDA has a principle that restricts use or disclosure of personal information for purposes beyond the original purpose without the individual's consent.[181] The Virginia Consumer Data Protection Act (VCDPA), a controller cannot process personal data for purposes inconsistent with the disclosed purpose unless the controller obtains consent.[182]

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.

6. Data Minimization

A central tenet of data 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," and it is core to data privacy protection.[183]

---

[176] 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 the world. ORG. FOR ECON. CO-OPERATION AND DEV., OECD GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA (1980).

[177] UK ICO, *Purpose Limitation,* 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/.

[178] ORG. FOR ECON. CO-OPERATION AND DEV., OECD GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA (1980).

[179] GDPR art. 5.1(b).

[180] GDPR Article 5(1)(c).

[181] Personal Information Protection and Electronic Documents Act (PIPEDA), S.C. 2000, c.5 (Principle 5 states "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.").

[182] Consumer Data Protection Act, ch. 36, 2021 to be codified at Va. Code Ann. § 59.1-574(A)(2) ("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").

[183] Lauren Bass, *The Concealed Cost of Convenience: Protecting Personal Data Privacy in the Age of*

In the United States, several federal laws include data minimization provisions.[184] For example, HIPAA requires reasonable efforts to limit the use or disclosure of protected health information to the minimum necessary to accomplish the intended purpose.[185] Under the Privacy Act, federal agencies must ensure that personal data is relevant and necessary to accomplish the agency's purpose.[186] The 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.[187] Under the Virginia CDPA, a controller must limit the collection of data for the purpose for which it is processed.[188]

The GDPR establishes a principle of data minimization, requiring that personal data must be adequate, relevant, and necessary to the purpose for which they are processed.[189] Brazil's privacy law, the LGPD, lists data minimization as one of its principles governing the processing of personal data.[190] Likewise, Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) has a principle called "Limiting Collection" that requires that the gathering of personal data must be limited and necessary for its purpose.[191] Principle 3 of the Australian Privacy Act restricts data collection to what is reasonably necessary for the collector's functions or activities.[192]

To further data minimization, many privacy laws impose data retention limitations to ensure that data is not used for longer than necessary to achieve the purposes of collection. For example, in the US, the Cable Communications Policy Act requires

---

Alexa, 30 FORDHAM Intell. PROP. MEDIA & ENT. L.J. 261, 312 (2019).

[184] *See, e.g.*, 45 C.F.R. § 164.502(b) (2019).

[185] 45 C.F.R. § 164.502(b) (2019) ("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.").

[186] 5 U.S.C. § 552a(e)(1) (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.").

[187] Cal. Civ. Code § 1798.100(c)) ("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.").

[188] VCDPA § 59.1-574(A)(1) ("Controller shall [l]imit the collection of personal data to what is adequate, relevant, and reasonably necessary in relation to the purposes for which such data is processed, as disclosed to the consumer").

[189] GDPR art. 5(c) ("adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')").

[190] LGPD art. 6(III) ("necessity: limitation of the processing to the minimum necessary to achieve its purposes, covering data that are relevant, proportional and non-excessive in relation to the purposes of the data processing").

[191] Personal Information Protection and Electronic Documents Act, S.C., ch. 5 (2000) (Can.).
Principle 4 ("The collection of personal information shall be limited to that which is necessary for the purposes identified by the organization.").

[192] *Privacy Act of 1988* (Cth) sch. 1 (Austl.) (Principle 3.1 states that "If an APP entity is an agency, the entity must not collect personal information (other than sensitive information) unless the information is reasonably necessary for, or directly related to, one or more of the entity's functions or activities." Principle 3.2 states "If an APP entity is an organisation, the entity must not collect personal information (other than sensitive information) unless the information is reasonably necessary for one or more of the entity's functions or activities.").

cable operators to destroy data when it is no longer necessary for its intended purpose.[193] 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.[194]

Beyond the U.S., privacy laws around the world require data minimization and data retention limits. For example, under the GDPR, personal data cannot be retained for longer than necessary. There is an exception for continuing to process data solely for a public interest, scientific interest, historical research, or statistical purposes.[195] Canada's PIPEDA requires that data only be retained for as long as necessary to achieve its intended purpose.[196]

As with other privacy principles and requirements, data retention limitations are completely thwarted by scraping, which involves the collection and retention of personal data without any restriction or time duration. The data minimization principle ensures that data use is constrained to specified purposes, that only data necessary for these purposes be collected, and that data be retained only as long as necessary to achieve these purposes. But scraping often involves the collection of vast stores of personal data with hardly any constraints. 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).[197]

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 when forming their privacy expectations and disclosure behaviors.[198]

---

[193] 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.").
[194] 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.").
[195] GDPR art. 5(e) ("kept 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 organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation')").
[196] PIPEDA Principle 5 provides that "Personal information shall be retained only as long as necessary for the fulfilment of those purposes."
[197] GDPR Art. 45; Woodrow Hartzog, *Chain-Link Confidentiality*, 46 GEORGIA LAW REVIEW 657 (2012).
[198] Alice E. Marwick and danah boyd. *I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience*, 13 NEW MEDIA & SOCIETY 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. RESEARCH 1018 (2010).

Scraping allows third parties to just take the data without any contract, any restrictions, or any consent. Any representations made by companies in contracts or in the design of the technology itself no longer apply.[199] The parties entrusted with people's data lose the ability to enforce promises made or preferences revealed within the context of that information relationship. Regulation of data sale or sharing is meaningless if data can just be grabbed by any third party.

The GDPR, many U.S. privacy laws, and privacy laws in other countries 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. Under the GDPR, when selecting processors, controllers must make sure that the processors provide "sufficient guarantees" of their ability to comply with the GDPR.[200] In the United States, the FTC has interpreted the failure to vet processors as a violation of the FTC Act.[201]

When organizations transfer personal data to other organizations, many laws require contracts to ensure that the recipient of the data adequately protects its privacy and security. For example, the GDPR requires a contract between the controller and the processor, and it sets forth a series of requirements for these contracts.[202] In the U.S. HIPAA requires contracts called "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.[203] 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.[204] Many of the state consumer privacy laws passed since 2018 require contracts with the recipients of data transfers that ensure that the data retains protection.[205]

The rationale for these protections is 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. Otherwise, the law's protections would readily evaporate.

Scraping renders onward transfer requirements meaningless. Scrapers are often not vetted, contracted with, or monitored. Scraped data thus loses all the law's protections. Additionally, scraping creates two classes of third party recipients of data. The first type of third parties must contract with organizations to obtain personal data and must protect the data similarly to how the organization that collected it protects it. The second type of third parties–the scrapers—can evade any

---

[199] *See generally, e.g.*, Woodrow Hartzog, *Website Design as Contract*, 60 AM. U. L. REV. 1635 (2011).
[200] GDPR art. 28.
[201] GMR Transcription Servs., Inc., FTC File No. 122-3095, 2014 WL 4252393, at *4 (F.T.C. Aug. 14, 2014).
[202] GDPR art. 28
[203] HIPAA, 45 C.F.R. §§ 164.502(e), 164.504(e).
[204] GMR Transcription Servs., Inc., FTC File No. 122-3095, 2014 WL 4252393, at *4 (F.T.C. Aug. 14, 2014).
[205] DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 186-190 (7th ed. 2024).

responsibility at all. Moreover, many laws require some form of individual consent for onward transfers of personal data, and these requirements, too, are ignored by scrapers.

8. Data Security

Scraping also contravenes the principle of data security. According to this principle, which is embodied in many privacy laws, 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 organizational measures."[206] This includes safeguarding personal data from a data breach. Protections must be established to prevent hackers from breaking in and improperly accessing the data.[207]

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 "only searching publicly available data from the Internet."[208] Open AI defends itself by claiming the data it scrapes is publicly available.[209] When it scraped LinkedIn

---

[206] GDPR Art. 5(1)(f).

[207] Solove and Hartzog, Breached, *supra* note X, at X.

[208] Clearview AI, 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 ("Clearview AI Is Only Searching Publicly Available Data from the Internet. 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 also Kashmir Hill, Your Face Belongs to Us (2023); 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 ("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.").

[209] Open AI, "How ChatGPT and our Language Models are Developed," https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed ("We use training information lawfully....[O]ur use of training information is not meant to negatively impact individuals, and the primary sources of this training information are already publicly

profile data, hiQ Labs claimed LinkedIn users lacked any privacy interest in their data because they made it publicly available.[210]

We contend that the argument that there is no privacy interest in publicly available information is normatively and legally wrong.

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.[211] Justifying scraping data because it is "public" information is woefully inadequate because "public" can be understood in several different ways depending on the context. It is often unclear which definition is being employed. For example, the standard dictionary definition for "public" is deceptively simple. The adjective 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."[212] The dictionary fails to indicate what 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 if they were not 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.[213]

As a noun, the term public is defined as "1. The people of a nation or community as a whole <a crime against the public>. 2. A place "open or visible" to the public <in public>."[214] The dictionary is not clear about the words "open or visible" in the definition of public. These words 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.[215]

As one of us argued, there are three different conceptions of what "public" or "publicly available" information could mean. First, public information could merely

available.").
[210] *Opinion,* hiQ Labs v. LinkedIn, D.C. No. 3:17-cv-03301-EMC (9th Cir. Apr. 18, 2022), at 18.
[211] Woodrow Hartzog, *The Public Information Fallacy*, 99 B.U. L. REV. 459 (2019); Joel R. Reidenberg, *Privacy in Public*, 49 U. MIAMI L. REV. ** (2014).
[212] Bryan A. Garner (2014). Public. *Black's Law Dictionary* (10th ed.).
[213] *Order Denying Clearview AI's Motion to Dismiss*, ACLU v. Clearview AI, Case No. 20 CH 4353 ("The fact that something has been made public does not mean anyone can do with it as they please").
[214] Bryan A. Garner (2014). Public. *Black's Law Dictionary* (10th ed.). 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. MERRIAM-WEBSTERS COLLEGIATE DICTIONARY (2012). Public. MERRIAM-WEBSTERS COLLEGIATE DICTIONARY (11th ed. 2012).
[215] Woodrow Hartzog, *The Public Information Fallacy*, 99 B.U. L. REV. 459, 473, 507, 514 (2019).

be a *descriptive* concept, with contextual factors shaping the contours of the notion, such as who the information was shared with, how many people saw and internalized information, where the information was located, how long the information was available, and the foreseeable extent of exposure. [216] 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 embarked on the successful "Eras" tour. Other people might describe public information as whatever society expresses a collective interest in, such as celebrity gossip.[217]

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 authorized. The most common example of designated public information is a "public record" or "open record."[218] 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.[219]

Third, "public" can be conceptualized by what it is *not,* i.e., shorthand for anything that is normatively or legally "*not private*."[220] 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 assuming the absence of a privacy interest by assumption in their argument that information is public.[221]

Since people's privacy expectations are contextually dependent, 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.[222] The fact that data is denoted "public" or "publicly available" is not evidence that people have voluntarily waived all expectations of

---

[216] *Id.*
[217] Hartzog, *Public Information, supra* note 211, at 464, 466.
[218] Daniel J. Solove, *Access and Aggregation: Privacy, Public Records, and the Constitution,* 86 MINN. L. REV. 1137 (2002).
[219] Hartzog, *Public Information*, *supra* note 211, at 509.
[220] *Id.* at 467-468, 496, 507-508, 511-512.
[221] *Id.* at 468, 508.
[222] *See, e.g.,* NISSENBAUM, PRIVACY IN CONTEXT, *supra* note 164, at 155 (drawing upon philosophy and social science in developing a theory of privacy as contextual integrity, theory of privacy as contextual integrity, which holds that privacy violations occur when "context-relative informational norms" are not respected when sharing information). Nissenbaum 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."

privacy.[223]

As one of us has previously argued, "privacy" means many different things, and privacy protection has many different dimensions.[224] Far too often, privacy is conceptualized as merely involving the safeguarding of hidden secrets.[225] 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 like stubborn fossils, 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 idea that publicly available information cannot implicate privacy interests is descriptively incorrect. The social science literature on privacy paints a much more complex picture of the relationship between the concept of "public" information and privacy expectations.[226] 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 want the ability to delete their data, to be asked for consent regarding data practices, and to be able to opt out of data collection at any time.[227] 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.[228]

Additionally, the 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.[229]

---

[223] Order Denying Clearview AI's Motion to Dismiss, ACLU v. Clearview AI, Case No. 20 CH 4353 (Cir. Ct. Ill. 2021) ("Clearview emphasizes that the photos from which they make faceprints are publicly available and that Plaintiffs have no 'expectation of privacy' in them").
[224] DANIEL J. SOLOVE, UNDERSTANDING PRIVACY (2008); Daniel J. Solove, *A Taxonomy of Privacy,* 154 U. PA. L. REV. 477 (2006).
[225] Daniel J. Solove, *I've Got Nothing to Hide" and Other Misunderstandings of Privacy,* 44 SAN DIEGO L. REV. 745 (2007).
[226] *See, e.g.,* Lior Jacob Strahilevitz, *A Social Networks Theory of Privacy*, 72 U. CHI. L. REV. 919 (2015).
[227] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Cranor, and Norman Sadeh, "Privacy Expectations and Preferences in an IoT World," *USENIX Association*, 2017, p. 399-412.
[228] Krishanu Dey & Parikshit Mondal, *Privacy Awareness among the Academic Social Network Users,* LIBRARY PHIL. AND PRACTICE 1 (2019) ("[A]mong all the respondents 44% wanted people to see and share their research data but did not allow anyone to download or edit or modify or tamper with their reports and data").
[229] 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*

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.[230] Most of the data about our lives is seen by and shared with some, but not all.

Consider behavior in public spaces. People go about their day-to-day lives in zones of 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, privacy expectations would change, and people would behave differently. In short, when people share data online, they do so for specific purposes and have particular expectations of use.

Scraping violates those choices and places people in an impossible position to assume that everything they share in a publicly available way with some should be fair game to exploit by all. People simply aren't capable of contemplating this sort of all-encompassing and hypothetical risk.

3. Privacy Law and Publicly Available Information

As noted above, the concept of "publicly available data" as a legal category is incoherent and inconsistent with many core privacy principles.[231] This conceptual incoherence has allowed companies to exploit loopholes to justify scraping. But even when lawmakers attempt to be specific about public data, they act inconsistently.

---

for Online Obscurity, 101 CAL. L. REV. 24 (2013) (critiquing idea that information is either disseminated globally or completely secret); Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. & LEE L. REV. 385, 387 (2013) (noting that modern understanding of privacy has created list of unaddressed problems); Evan Selinger & Woodrow Hartzog, *Obscurity and Privacy, Spaces for the Future, in* A COMPANION TO THE PHILOSOPHY OF TECHNOLOGY (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'*, THE ATLANTIC (2013), https://www.theatlantic.com/technology/archive/2013/01/obscurity-a-better-way-to-think-about-your-data-than-privacy/267283/ (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 Science Monitor (2015) https://www.csmonitor.com/World/Passcode/Passcode-Voices/2015/0415/Why-you-have-the-right-to-obscurity (describing obscurity as 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 (2014) ("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").

[230] Evan Selinger & Woodrow Hartzog, *Obscurity and Privacy, Spaces for the Future, in* A COMPANION TO THE PHILOSOPHY OF TECHNOLOGY (Routledge: by Joseph Pitt & Ashley Shew, eds. 2017); Daniel J. Solove, *Access and Aggregation: Privacy, Public Records, and the Constitution,* 86 MINN. L. REV. 1137, 1175 (2002) ("Privacy can be violated by altering levels of accessibility, by taking obscure facts and making them widely accessible."); Daniel J. Solove, *A Taxonomy of Privacy,* 154 U. PA. L. REV. 477, 538-40 (2006) (noting that privacy can be violated by increasing the accessibility of data).

[231] David Zetoony, *What is 'Publicly Available Information' under the State Privacy Laws?*, NATIONAL LAW REVIEW (Sept. 13, 2023) https://www.natlawreview.com/article/what-publicly-available-information-under-state-privacy-laws.

Some privacy laws exempt publicly available information, but others do not. For example, Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) excludes publicly available information.[232] But the EU's General Data Protection Regulation does not contain such 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."[233] Though it's not always clear when this exception applies.[234] In the U.S., federal privacy laws are inconsistent on the issue. The Fair Credit Reporting Act does not exclude publicly available information.[235] 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.[236]

Many U.S. state consumer privacy laws exempt publicly available data, though their definitions of such data vary as do the scope of what is excluded.[237] Although exempting publicly available data, the California Consumer Privacy Act (CCPA) states that "publicly available" does not include biometric information that a business collects without a consumer's knowledge.[238] Other laws protecting biometric information do not exempt publicly available data.[239]

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.[240] 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

---

[232] *See* Personal Information Protection and Electronic Documents Act, S.C., ch. 5 (2000) (Can.), §7(1)(d) (allowing collection of publicly available personal information without knowledge and consent); Section 3(h.1) (allowing collection of publicly available personal information without knowledge and consent).

[233] The GDPR generally protects against publicly available data, but it exempts "personal data which are manifestly made public by the data subject." GDPR art. 9.2(e).

[234] *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 PRIVACY L. 107 (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.").

[235] The GDPR generally protects against publicly available data, but it exempts "personal data which are manifestly made public by the data subject." GDPR art. 9.2(e).

[236] 15 U.S.C. §§ 6809(4)(A)-(B) ("The term "nonpublic personal information" . . . does not include publicly available information").

[237] California Consumer Privacy Act, Cal. Civ. Code § 1798.140(v)(2) (West 2021); VDCPA, Va. Code § 59.1-571 (2021); Utah Code Ann. § 13-61-101(29)(b) (2022).

[238] See id.

[239] Biometric Information Privacy Act, 740 ILCS 14; Washington My Health My Data Act, RCW 19.373.010 (22).

[240] An Act to amend Section 1798.140 of the Civil Code, relating to privacy, Assembly Bill No. 1008, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB1008 ("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.").

individual.[241] Connecticut's definition is similar to Colorado's but also includes data disseminated by the media.[242] According to privacy lawyer David Zetoony, "most data privacy statutes would not classify all internet-accessible information as being 'publicly available.'"[243]

Turning to judicial cases, 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 also recognized privacy interests in such data – sometimes even the same court in different contexts.

In *United States Dept' 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.[244] Reporters sought to obtain under the Freedom of Information Act (FOIA) FBI compilations of criminal history data on individuals, but 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. But the Court concluded 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."[245] The Court's holding is relevant to scraping for two reasons. First, the Court recognizes that the fact that personal data is publicly available does not automatically extinguish a privacy interest in the data. 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."[246] Before *Carpenter,* the U.S. Court's Fourth Amendment jurisprudence had generally maintained that anything observable in a public place was not private.[247] 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 the privacy torts fail to find a privacy interest in

---

[241] C.R.S. § 6-1-1303(17)(b) (2021).
[242] Connecticut Data Privacy Act, § 1(25).
[243] Zetoony, supra note 231.
[244] 489 U.S. 749 (1989).
[245] *Id.* at 763-64.
[246] Carpenter, 138 S. Ct.at 2217 ("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.").
[247] United States v. Knotts, 460 U.S. 276 (1983) (no reasonable expectation of privacy when tracking device monitored movement in public); Florida v. Riley, 488 U.S. 445 (1989) (no expectation of privacy in anything that can be viewed on one's property by police officers in a helicopter flying in legal airspace).

publicly available information, there are many notable exceptions.[248] 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.[249] 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."[250]

What the cases reveal is that it is far too simple 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. Privacy law remains deeply conflicted on the status of publicly available information.

Ultimately, privacy law cannot achieve its goals if it fails to protect publicly available personal data. In the modern world, with the internet, an unprecedented amount of personal data is being posted online. A lot of personal data is posted by individuals themselves, but also a lot of personal data about people is posted by other people or by organizations such as schools, employers, journalists, and more. 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 laws' protection.

Scraping can avoid conflicting with certain privacy laws that have broad exemptions of publicly available data, but it is difficult to square this with any coherent account of the principles that privacy laws are attempting to achieve.

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 of the core principles that animate many privacy laws. 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 about how scraping should be conceptualized in terms of its privacy impact. When scraping is 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.

---

[248] Solove, *Access and Aggregation, supra,* at 77-79.
[249] Nader v. General Motors Corp., 255 N.E.2d 765 (N.Y. Ct. App. 1970).
[250] Anderson v. Mergenhagen, 642 S.E.2d 105 (Ga. Ct. App. 2007).

We then discuss why merely applying privacy law to scraping would lead to undesirable consequences. Privacy law fails to effectively regulate the collection, use, and transfer of personal data, and it will not address scraping well. Under many privacy laws, existing infirmities with consent could lead to end-runs around any meaningful control over scraping. Under other privacy laws, scraping might be practically impossible, thus leading to a de facto ban on scraping. But a ban on scraping is 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 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 practices that Shoshana Zuboff famously has termed "surveillance capitalism."[251] 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 context of other data practices, the realities of the political and commercial incentives, and the likely downstream effects of data capture.

Surveillance is a broad concept capable of multiple meanings.[252] There is even an entire field devoted to the concept of surveillance studies.[253] The 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."[254] Neil Richards argued, "Four aspects of [Lyon's] definition are noteworthy, as they expand our understanding of what surveillance is and what its purposes are. First, it is focused on learning information about individuals. Second, surveillance is systematic; it is intentional rather than random or arbitrary. Third, surveillance is routine--a part of the ordinary administrative apparatus that characterizes modern societies. Fourth, surveillance can have a wide variety of purposes--rarely totalitarian domination, but more typically subtler forms of influence or control."[255]

Many kinds of web scraping, such as the scraping of social media profiles, reflects all four aspects of Lyon's definition of surveillance. First, it 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. 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

---

[251] SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM (2019).
[252] *See, e.g.*, OSCAR GANDY, THE PANOPTIC SORT; GARY MARX, WINDOWS INTO THE SOUL; Rule, et. al. *Documentary identification and mass surveillance in the United States*, 31 SOCIAL PROBLEMS 223 (1983) ("any systematic attention to a person's life aimed at exerting influence over it.").
[253] SURVEILLANCE STUDIES: A READER (Torin Monahan and David Murakami Wood eds., 2018)
[254] DAVID LYON, SURVEILLANCE STUDIES 23 (2007).
[255] Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1937 (2013).

using their face, gait, or even heartbeat as a beacon.

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.

Third, web scraping has been completely routinized by companies developing AI.[256] 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.[257] Companies like Clearview AI collect billions of photos to power their databases through routines designed to cheaply and quickly scrape websites.[258]

Finally, many companies scape data from the web to influence people, manage them, protect them, or direct them. While academics and journalists might scrape to gain knowledge, companies scrape the web to make money, which means 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. Sill 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 for spearphishing for the ultimate goal of fraud, theft, and hacking, as part of an endless game of cat and mouse.

To understand scraping as surveillance is to recognize that scraped data over time can 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. It recognizes that these harms flow first from the collection of data, and that this is often the best place to address these harms in law.

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."[259] But this objection ignores scraping's incredible affordances of scale.[260] The fact that

---

[256] *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/.
[257] *Id.*
[258] 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/.
[259] Camille Fischer & Andrew Crocker, *Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data*, ELECTRONIC FRONTIER FOUNDATION (Sept. 10, 2019), https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data.
[260] *See, e.g.*, Mark P. McKenna & Woodrow Hartzog, *Taking Scale Seriously in Technology Law* (forthcoming 2024) (on file with authors).

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 ClearviewAI to assemble a biometric database that works at scale. Scraping isn't just "more" of an acceptable activity. It's a difference in magnitude of risk that is so large it is a difference in kind.

Treating scraping like surveillance would have the important effect of tying scraping rules with the gradual recognition in surveillance law that sometimes individuals can and should have a reasonable expectation of privacy *in public* or with *publicly available information.*[261] So too with most kinds of personal information disclosed online. Individuals make their personal data publicly available online to be shared with others for certain purposes. For example, on LinkedIn, people share biographical information for professional and career purposes; they do not just throw it out into the world for any purpose whatsoever. Public availability creates risks that data might be improperly collected and used. But these risks do not obviate the fact that the law should protect this data. For example, the law protects copyrighted content from unauthorized copying and distribution even though it is publicly available. Personal data should have similar protections. We are not arguing that personal data should be regulated identically to intellectual property; instead, we are contending that public availability does not eliminate all limitations on use and dissemination of data.

Zooming out, if lawmakers treated scraping like surveillance, it would help re-frame policies and public discourse that treat our raw human information and experiences as a free-for-all resource, there for the taking. Instead of the industry's harmful vision of our data as part of the "biopolitical public domain" articulated by Julie Cohen, information about our lives would be legally recognized as being inherently valuable, inextricably tied to our dignity and wellbeing, and worthy of protection.[262]

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. This duty is invoked in several different areas such as cybersecurity, data protection, anti-hacking safeguards, and trust/assurance compliance. The idea is that it is foreseeable that some actors will use wrongful means to access people's data, 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 is called hacking. A successful hack results in a personal data breach. In this part, 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.

---

[261] *See infra* Part II.B.
[262] *See, e.g.*, JULIE COHEN, BETWEEN TRUTH AND POWER (2019); Julie Cohen, The Biopolitical Public Domain: The Legal Construction of the Surveillanc Economy.

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 verses preserving and manual access versus automation.

Security is often thought of as akin to locking data in a safe and keeping it hidden from malicious actors. 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. 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. Access must still be authorized. Data must still be protected.

It is a mistake to think that scraping just involves one party. 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 that are similar to the same safeguards they take to prevent hackers from accessing personal data without authorization. For 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.[263] 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."[264] 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, limiting the number of visits per hour or day by a single account, 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, threatening or taking appropriate legal action, and notifying affected individuals.[265]

In the United States, the FTC could demand these practices as part of their regulation of unfair and deceptive trade practices. We argue that such duties to protect against scraping should also be included in the FTC's proposed regulations for commercial surveillance and data security practices. States should also ensure that these 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.[266] Under the Rule, a "breach of security" is defined as

---

[263] Joint Statement on Data Scraping, *supra* note 136.
[264] *Id.*
[265] *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/ (citing Garante Per La Protezione Dei Dati Personali, Provvedimento del 20 maggio 2024 [10020316], https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/10020316).
[266] 16 CFR Part 318.

"acquisition of [PHR identifiable health information] without the authorization of the individual." 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 that companies that it was a reportable data breach when companies shared health data with third parties in violation of their privacy policies.[267] 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.

B. THE DIFFICULTY OF BRINGING SCRAPING UNDER THE PURVIEW OF PRIVACY LAW

The tension between scraping and privacy cannot be resolved satisfactorily by anointing scraping or privacy as the winner. Allowing unfettered scraping would constitute an untenable threat to privacy. Scraping involves a cascade of privacy violations on an enormously grand scale. Thus, we contend that scraping does and should fall under many existing privacy laws – not just the GDPR but also several U.S. privacy laws. However, 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 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 U.S. will make it quite difficult for a scraper to navigate.

Under the EU's GDPR, scraping requires a lawful basis.[268] As discussed earlier, the two most common lawful bases advanced for scraping are individual consent or legitimate interests. In fact, the Dutch Data Protection Authority has declared: "In practice, scraping by private organizations and private individuals is only possible on the basis of legitimate interest."[269] Regarding special categories of personal data (often called "sensitive data"), the European Data Protection Board has clarified that in addition to 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."[270] The EDPP recognized that "where large amounts of personal data are collected via web scraping, a case-by-case

[267] In Re GoodRx Holdings, Inc., (FTC 2023); In re Easy Healthcare Corp., (FTC 2023).
[268] GDPR art. 6.
[269] AUTORITEIT PERSOONSGEGEVENS *supra* note 132.
[270] *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.

examination of each data set is hardly possible."[271] However, the EDPB recognized that rigorous safeguards like data minimization can help processors comply with the GDPR.

The consent lawful basis requires affirmative action by the data subject, which would be impractical for scrapers to obtain. It is hard to imagine how scraping could occur under the consent lawful basis. 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 the legitimate interest 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 which require protection of personal data, in particular where the data subject is a child."[272] This lawful basis, however, focuses primarily on a balance between the specific business interests of the scraper and the interests of individuals. Due to the societal implications of widespread scraping, we argue that public interests should play a larger role in this balancing test.[273] Doing so would affect both sides of the equation, given the costs and benefits of scraping in various contexts.

Additionally, it is inevitable that scraping will gather sensitive data or personal data that in combination could give rise to inferences about sensitive data (which are also deemed to be sensitive data under the GDPR).[274] For sensitive data, the legitimate interests lawful basis is unavailable.[275] It is thus difficult to imagine how scrapers could navigate around this problem.

U.S. law is messier and even less clear. If state consumer privacy laws did not exempt "publicly available information," they would apply differently based on their size thresholds for companies. Many laws are triggered on amount of revenue or the number of state residents whose data is gathered. 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.[276] But an opt out right would be meaningless if people have no idea who the scrapers are or even that their data is being scraped. Generally, opt out "consent" is rather farcical– it is not really meaningful consent; it is inaction that is wrongly treated as

[271] *Id.*
[272] GDPR art. 6(f).
[273] What is the 'legitimate interests' basis?, UK ICO, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/legitimate-interests/what-is-the-legitimate-interests-basis/#what_counts ("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.").
[274] Daniel J. Solove, *Data Is What Data Does: Regulating Based on Harm and Risk Instead of Sensitive Data,* 118 NW. U. L. REV. 1081 (2024)
[275] GDPR art. 9.
[276] DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 186-190 (7th ed. 2024).

consent.[277] Conversely, state laws that require opt in consent for sensitive data could become a difficult requirement for scrapers to navigate.[278] Then, there is the issue of whether inferences that reveal sensitive data count as sensitive data, and it remains unclear how many state laws will address this issue.

On top of all this, the FTC has a basis in its jurisprudence to deem scraping to be a prohibited unfair act or practice under the FTC Act. The Commission has not recognized an exception for scraping publicly available information in its cases. Moreover, the FTC could also conclude that websites that fail to take reasonable measures to prevent scraping could be in violation of the FTC Act under either an unfairness or deception rationale.

The answer to how scraping should fit with privacy law is thus quite unclear. What is clear is that scraping violates the core principles of privacy laws even when the laws themselves are drafted in ways that poorly implement these principles. Privacy laws can be too restrictive in some contexts and too permissive in others, and generally in the U.S., too inconsistent.

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.[279] 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 purpose in this section is not to go through all privacy laws in detail to show how they might incorporate our recommended regulatory proscriptions. Instead, we will sketch out the basic goals the law should 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 would need to be changed.

1. The Undesirability of a Total Scraping Ban

Although scraping deeply conflicts with nearly all core principles of privacy, it should not be banned outright. Banning scraping would come at a great financial and social cost, 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.

Thus, a total ban on all scraping of personal data seems unwise. Scraping has beneficial uses, such as when done by researchers and journalists. Many research projects and news stories cannot be achieved without scraping. Banning all scraping would severely impair the 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."[280]

---

[277] Solove, *Murky Consent, supra* note 172, at 597.
[278] SOLOVE & SCHWARTZ, PRIVACY LAW FUNDAMENTALS, *supra* note 276, at 186-190.
[279] For an exploration on the role of scale in technology law, *see* Mark McKenna and Woodrow Hartzog, *Taking Scale Seriously in Technology Law* (draft on file with authors).
[280] 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-

While we can debate the wisdom of many new kinds of AI systems, any informed policy decision should be made conscious of what is being left on the table.

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[281] As Sellars notes: "many forms of web scraping provide important benefits to consumers and the public."[282]

Restrictions on scraping could also further distort AI models. If privacy laws in certain countries block scraping, then AI datasets 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 US data but deprived of EU data, skewing them to the US. While the necessity of maximum data collection to train AI models has probably been wildly exaggerated, it still seems likely scraping will be important in the search for "less discriminatory algorithms."[283]

A scraping ban would also favor companies that already possess large data sets, such as big platforms. These companies 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 lands. To do this, they are simply declaring that they will do it. Many U.S. privacy laws allow organizations to collect and use personal data in nearly any way they want just by stating what they are doing.[284] Many companies have already "updated their terms of service to include references to building AI with user data."[285] For example, Twitter and Amazon announced plans to use data from their users to train their AI.[286] 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."[287] X revised its privacy notice to allow it to use: "publicly available information" for training "our machine learning or artificial intelligence models."[288]

---

lawsuit-would-take-sledgehammer-generative-ai-2023-10-17/.

[281] 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.

[282] Sellars, *Twenty Years of Web Scraping*, *supra* note 6, at 412.

[283] *See, e.g.*, Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas & Mingwei Hui, Less Discriminatory Algorithms, 113 GEO. L. J. (forthcoming 2024).

[284] Solove, *Murky Consent*, *supra* note 172.

[285] U.S. FED. TRADE COMMISSION, GENERATIVE ARTIFICIAL INTELLIGENCE AND THE CREATIVE ECONOMY STAFF REPORT: PERSPECTIVES AND TAKEAWAYS 10 (Dec. 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/12-15-2023AICEStaffReport.pdf.

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

[287] Jess Weatherbed, *Google confirms it's training Bard on scraped web data, too*, THE VERGE (Jul. 5, 2023), https://www.theverge.com/2023/7/5/23784257/google-ai-bard-privacy-policy-train-web-scraping.

[288] Sarah Perez, *X's Privacy Policy Confirms It Will Use Public Data to Train AI Models*, TECH CRUNCH (Sept. 1, 2023), https://techcrunch.com/2023/09/01/xs-privacy-policy-confirms-it-will-use-public-

In 2023, Zoom abruptly altered its privacy notice and states that users were agreeing to Zoom's "access, use, collection, creation, modification, distribution, processing, sharing, maintenance, and storage of Service Generated Data for any purpose." The language "any purpose" included the purpose of training AI models. Zoom also slipped into the notice a perpetual license to use people's data for AI training. When these changes were called out publicly, Zoom backpedaled.[289]

We are thus already witnessing companies start to cannibalize their own data for the purposes of developing AI. These companies are changing their privacy notices and terms of service to allow for the use of consumer data in the development of their AI algorithms. The FTC recently warned that this practice could violate the FTC Act:

> It may be unfair or deceptive for a company to adopt more permissive data practices—for example, to start sharing consumers' data with third parties or using that data for AI training—and to only inform consumers of this change through a surreptitious, retroactive amendment to its terms of service or privacy policy.[290]

However, there are ways companies could evade entanglements with the FTC Act, such as applying any changes proactively rather than retroactively.

Scraping personal data should not be banned in its entirety, but if we value the privacy principles underpinning privacy law, scraping must be brought under control.

2. The Consent Model

One model is for websites to obtain individual consent for their data to be scraped by third parties. Under many U.S. privacy laws, websites could disclose the possibility of scraping in their privacy notices or provide explicit warnings of scraping. Under 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.

Such an approach would be highly undesirable because it exacerbates existing

---

data-to-train-ai-models/.

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

[290] FTC, *AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive*, FTC TECHNOLOGY 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.

shortcomings in privacy laws regarding consent. Most consent in privacy laws is fictional.[291] Such an approach would subject individuals to data gathering and use on a massive scale, wrapping it in a farcical veneer of legitimacy. It is hard to imagine how any form of consent could be meaningful to such massive data gathering and use for a myriad of unspecified purposes without limitation.

In the U.S., the notice-and-choice approach has been severely criticized as a vehicle for companies to gather and use data with hardly any limitations.[292] In the EU, the GDPR rejects the notice-and-choice approach; consent must be express and affirmative (opt in).[293] But even express consent can sometimes readily be obtained and is not meaningful. Websites can make people click accept buttons without people understanding the implications. Even with accept buttons, readership of terms barely increases.[294] Privacy consent is mostly fictional, and people will readily consent to the use of their data in exchange for the immediate benefits of technology.[295]

With many forms of AI, the uses of personal data are manifold and not fully knowable at the time of data collection. This is especially true for generative AI, where the uses are determined not just by the creator of the AI model but by the users of the model. Professor Elettra Bietti warns that consent has become a "free pass" for platforms to use personal data in nearly any way they desire.[296]

Increasingly, companies will be incentivized to use aggressive means to block scraping by third parties and instead move to make deals with scrapers. Scrapers would no longer have to scrape data if websites were to simply share their data through paid APIs.

Market forces might already be pushing in this direction, as this is a way for websites to further monetize user data and to control which third parties can scrape, providing websites with the ability to exclude competitors. For example, Reddit originally had a free API for scrapers but in 2023, started to charge for the use of its API.[297] Indeed, such a model need not involve scraping – websites could just provide the data to third parties, though such a practice would be the functional equivalent to scraping.

But this market approach leaves individuals whose data is scraped largely out of the

---

[291] Solove, *Murky Consent, supra* note 172, at 631.
[292] Neil Richards and Woodrow Hartzog, *The Pathologies of Digital Consent*, 96 WASH. U. L. REV. 1461, 1463 (2019); Richard Warner & Robert Sloan, *Beyond Notice and Choice: Privacy, Norms, and Consent*, 14 J. HIGH TECH. L. 370 (2014); Helen Nissenbaum, *A Contextual Approach to Privacy Online*, 140 DAEDALUS 32, 34 (2011).
[293] GDPR, art. 4(11) (requiring consent to be "freely given, specific, informed and unambiguous indication of the data subject's wishes").
[294] 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%).
[295] Solove, *Murky Consent, supra* note 172, at 620.
[296] Elettra Bietti, *Consent as a Free Pass: Platform Power and the Limits of the Informational Turn,* 40 PACE L. REV. 308, 313 (2020).
[297] 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.

loop. Massive data transfer would occur based on a series of backroom deals without individuals having a seat at the table. Companies would lean on the fictitious mechanisms of consent to leave users largely out of the picture. Either such arrangements would be buried in a privacy notice or some form of affirmative consent mechanism would be used, but the infirmities of consent would cast their stink over each of these methods. Even if the financial exploitation of data did better distribute benefits and decision-making with the data subjects, it would still turn privacy into a luxury good, conditioning people's privacy on their level of financial comfort and ignoring the poor distributional effects of data markets.

Additionally, any approach built upon individual consent ignores the collective effects of data exposure, which has a disproportionally harmful effect of on marginalized communities like people of color and members of the LGBTQ+ community.[298]

This model would make scraping the privilege of the rich and powerful and further entrench inequality; these entities could afford to buy access to the vast repositories of data, the oil on which modern AI runs. These entities would then be able to develop AI whereas smaller less wealthy entities would not. The rich would grow richer and the poor would grow poorer.[299]

C. A REGULATORY AGENDA FOR SCRAPING IN THE PUBLIC INTEREST

In order for U.S. privacy law to achieve a workable balance with allowing scraping yet protecting privacy, the law must look beyond some of the traditional permissive approaches to regulating the collection, use, and transfer of personal data. Instead of the general approach in the U.S. as allowing organizations wide leeway to collect and use personal data in whatever way they want, the law should view the systemic, automated mass collection and use of personal data through scraping as a *privilege*. This view of data collection, use, and transfer is similar to the GDPR's approach; there must be a justifiable basis for these activities. We propose turning this requirement into a privilege by conditioning 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 valid justification for scraping and substantive and 2) substantive protections to ensure the scraping is safe and avoid exploitation and purpose creep; and 3) procedural safeguards to ensure fairness and adequate representation and agency in decision-making. First, we propose that automated mass scraping of personal data should only be allowed when it is

---

[298] *See, e.g.*, 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 Fairfield & Christoph Engel, *Privacy as a Public Good*, 65 DUKE L.J. 385 (2016); **Chris Gilliard, T*he Rise of 'Luxury Surveillance'*, THE ATLANTIC (Oct. 18, 2022), https://www.theatlantic.com/technology/archive/2022/10/amazon-tracking-devices-surveillance-state/671772/; Evan Selinger & Woodrow Hartzog, *The Inconsentability of Facial Surveillance*, 66 LOY. L. REV. 101 (2019).

[299] *See, e.g.*, Rory Van Loo, *Privacy Pretexts*, 108 CORNELL LAW REVIEW 1 (2022) (arguing in favor of the allied access to personal data by digital helpers, competitors, and regulators to advance people's interests.)

necessary for furthering 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, which generally fall under the lawful bases of consent or legitimate interests.[300] 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."[301] This provision is geared more towards the exercise of government authority and likely is to be interpreted narrowly rather than broadly.[302] The public interest bases for processing sensitive data are more narrowly constricted, allowing for "substantial public interest, on the basis of Union or Member State law," "for reasons of public interest in the area of public health," or "necessary for archiving purposes in the public interest" or "scientific or historical research purposes or statistical purposes."[303] 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.

As a general observation, the public interest remains an underutilized concept in U.S. data privacy law, though scholars are increasingly looking to more collective and social aspects of information privacy.[304] We use "public interest" here to mean a consideration of the collective or shared wellbeing of a public or publics as opposed to a more atomized, individualized wellbeing. Specifically, we deploy the term "public interest" similar to how the concept of the public has been deployed in public health law. Lawrence Gostin helpfully conceptualized public health law as being concerned with how state power is deployed and constrained "to ensure the conditions for people to be healthy (to identify, prevent, and ameliorate risks to health in the population), and of the limitations on the power of the state to constrain for the common good the autonomy, privacy, liberty, proprietary, and other legally protected interests of individuals. The prime objective of public health law is to pursue the highest possible level of physical and mental health in the population, consistent with the values of social justice."[305]

There are several major themes in Gostin's conceptualization of public health that we think are relevant for rules around scraping in the public interest. Specifically, we recommend public health's population-level focus that must remain consistent

---

[300] *Public Task*, U.K. INFORMATION COMMISSIONER'S OFFICE, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/lawful-basis-for-processing/public-task/ (last visited July 2, 2024).

[301] GDPR art. 6(e)

[302] U.K. INFORMATION COMMISSIONER'S OFFICE, *supra* note 300 ("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; 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.").

[303] GDPR art 9(g), (i), and (j).

[304] *See, e.g.*, 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); PRISCILLA M. REGAN , LEGISLATING PRIVACY: TECHNOLOGY, SOCIAL VALUES, AND PUBLIC POLICY (1995); PRISCILLA REGAN, PRIVACY AND THE COMMON GOOD: REVISITED, THE SOCIAL DIMENSIONS OF PRIVACY (2015); Charlotte A. Tschider, *AI's Legitimate Interest: Towards a Public Benefit Privacy Model,* 21 HOUS. J. HEALTH L. & POLICY 125, 132 (2021).

[305] Lawrence O. Gostin, *A Theory and Definition of Public Health Law*, *in* PUBLIC HEALTH LAW: POWER, DUTY & RESTRAINT (Revised & Expanded 2d ed. 2008) at 4.

with the values of social justice—the "[f]air and equitable treatment of groups and individuals, with particular attention to the disadvantaged."[306] We think this is a good place to start for scraping in the public interest, expanding beyond health to specifically include other major areas commonly regulated in the public interest, including employment, housing, accessibility, infrastructure, and public transportation.

We contend public interest should be the primary consideration for justifying the collection, use, and transfer of personal data. 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 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.

Ultimately, there remains a question of gigantic importance: When is the collection, use, and transfer of personal data in the public interest? Answering this question is quite difficult; and it will be contextual for various instances of data collection, use, and transfer. We are not attempting to provide a full answer here. Instead, our goal is more modest. We are contending that developing a framework for data processing in the public interest is the most viable path forward, as difficult as it may be. Other approaches, such as relying on fictious consent or allowing scrapers impunity based on a faulty theory of "publicly available information" are untenable.

It might be that lawmakers and regulators should 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.[307]

We suggest that at least four principles should be followed. The first two principles can guide lawmakers in determining when scraping is in the public interest. 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) *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.

---

[306] *Id.* at 4-5. For a developed framework for addressing group-specific harms like harms to African Americans (not just general harms), *see* Anita L. Allen, *Dismantling the "Black Opticon": Privacy, Race Equity, and Online Data-Protection Reform*, 132 YALE L.J. FORUM 907 (2022).

[307] 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* Ari Waldman, *Privacy, Practice, and Performance*, 110 CALIF. L. REV. 1221, 1270 (2022); *see also* NEIL RICHARDS, WHY PRIVACY MATTERS (2021); JULIE COHEN, BETWEEN TRUTH AND POWER (2019).

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 upon the uses of scraped personal data should be fair, open, accountable, representative, equitable, and deliberative.

4) *Protections Principle:* Scraped data should be afforded all the protections as other personal data under privacy laws unless particular protections are unworkable.

1. Use of Data as a Privilege

Generally, U.S. privacy law views data collection and use as the natural right of organizations. Under the notice-and-choice approach, as long as organizations disclose what they are doing in a privacy notice, they are generally free to do whatever they want with the data. In contrast, we side with the approach of the EU's GDPR, which requires a permissible purpose for data collection and processing– the lawful basis approach that we discussed earlier in this Article.[308]

The collection and use of personal data should be understood as a privilege. Scraping personal data should be allowed when in the public interest, but not for other purposes because it threatens people's privacy and facilitates a host of individual and social information-related harms including harassment, labor exploitation, manipulation, and wrongful discrimination. The GDPR might allow for scraping with consent or for legitimate interests, but 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.[309] 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 basis broadly allows processing of personal data for a very wide range of purposes "pursued by the controller or by any third party."[310] 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.[311]

We contend that because of the extensive scale of scraping and the particular concerns it raises, it should only be allowed when in the public interest. The GDPR has such a lawful basis.[312] We do not propose to follow precisely the particular formulation or interpretations of the GDPR's public interest lawful basis; rather, we

[308] *See supra* at text accompanying notes 268 to 275.
[309] Solove, *Murky Consent, supra* note 179, at 606.
[310] GDPR art. 6(f).
[311] AUTORITEIT PERSOONSGEGEVENS *supra* note 132.
[312] GDPR art. 6(e).

merely 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 only to companies. Accordingly, the court's rationale in hiQ was flawed in that the court mainly focused on the benefits to hiQ's business model and failed to appreciate the privacy interests of the individuals whose data was being scraped.

Additionally, 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 quality. Mere allegations that scraping will help "keep people safe" or "improve your health" should be insufficient without convincing proof that a demonstration that the scraping is necessary and proportionate to the purpose. Industry will likely attempt to dilute and work around any rule in order to maximize profit, and lawmakers should craft their rules accordingly.[313]

Another factor in the analysis should be whether AI models trained on scraping data were created with better public involvement. Essentially, scraping would be understood as a special privilege to be allowed when certain conditions exist. In order for AI development with people's data to be permitted, individuals or the public should receive something in return. This is a kind of grand bargain, a widescale compromise of people's privacy in exchange for something that benefits people, not just a way for companies to make a profit.

If lawmakers take the scraping of personal data seriously, then it is likely much less scraping will occur. Companies hoping to scrape personal data would likely lobby for exceptions and possibly even request a *right* to scrape data, either in new legislation, as part of existing competition law, or even under the First Amendment.

It is hard to imagine how the law could force websites to allow certain forms of scraping, such as scraping by the media or researchers or competitors. The court in the *hiQ v. LinkedIn* case attempted to restrict LinkedIn's ability to stop hiQ from scraping user profiles because of "hiQ's interest in continuing its business, which depends on accessing, analyzing, and communicating information derived from public LinkedIn profiles."[314] This government compelled right to scrape is deeply problematic, as it ignores the privacy interests of individual users and infringes upon the promises LinkedIn makes to its users as well as LinkedIn's obligations under privacy laws. Clearview AI's claim that it has a First Amendment right to scrape people's publicly available photographs is equally spurious because it relies on an illusory bright line distinction between what is public and private, unsupported by doctrine, and would lead to absurd conclusions.[315]

---

[313] *See, e.g.*, JULIE COHEN, BETWEEN TRUTH AND POWER (2019); ARI WALDMAN, INDUSTRY UNBOUND (2021).

[314] hiQ Labs, Inc. v. LindkenIn Corp., 938 F.3d 985 (9th Cir. 2019); *see also* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.3rd 1180 (9th Cir. 2022).

[315] *See, e.g.*, Neil M. Richards, *Reconciling Data Privacy and the First Amendment*, 52 UCLA L. REV. 1149, 1157 (2005); Neil M. Richards, *Why Data Privacy Law Is (Mostly) Constitutional*, 56 WM. & MARY L. REV. 1501, 1533 (2015); Woodrow Hartzog, *The Public Information Fallacy*, 99 B.U. L. REV. 459, 461 (2019); *see also* Jake Karr and Talya Whyte, *The First Amendment Should Protect Us from*

The law can instead take an incentives approach. It can allow websites to use their repositories of data for their own purposes (if such 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. Under current law, however, if websites can wall off the user data they have and then obtain farcical consent to use this data in any way they see fit, an incentives approach will fail because the default would vitiate any incentive. 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.

A more robust public interest justification for using personal data would work better than the farcical game that occurs under "consent." The legitimate interest lawful basis under the GDPR comes close, but a public interest inquiry would better capture both the individual and societal interests involved.

2. Guidelines for Scraping

Guidelines about scraping in the public interest must be developed. We recognize that determining what is in the public interest is an immensely difficult and contested matter, but this is ultimately the best direction for a more coherent approach to regulating the collection, use, and transfer of personal data. As we stated earlier, four principles should guide the law: (1) Reasonable Risk of Harm Principle, (2) Proportional Benefits Principle, (3) Process Principle, and (4) Protections Principle. We also note that not all instances of public interest are equally strong, and privacy is often better addressed in non-binary ways.

For the *Reasonable Risk of Harm Principle,* the law should protect people from downstream harms from having their data scraped. Although newer AI laws are focusing on risk and harm, many privacy laws are built around consent, individual control, and other approaches that are ill-suited to the age of AI. 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 corrosion of social trust, the collapse of democratic institutions, and the failure of infrastructure.[316]

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

---

Facial Recognition Technologies – Not the Other Way Around, TECH POLICY PRESS (Aug. 15, 2023), https://www.techpolicy.press/the-first-amendment-should-protect-us-from-facial-recognition-technologies-not-the-other-way-around/ ("Embracing Clearview's framing would provide it with a First Amendment get-out-of-jail-free card for almost any violation of law, leaving Clearview's secret, commercially motivated facial recognition business entirely insulated from most government regulation and consumer protection or civil rights lawsuits.").

[316] *See, e.g.*, Anita L. Allen, *Dismantling the "Black Opticon": Privacy, Race Equity, and Online Data-Protection Reform*, 132 YALE L.J. FORUM 907 (2022); Julie Cohen, *Infrastructuring the Digital Public Sphere*, 25 YALE J.L. & TECH. 1 (2023); *see also* ROBERT PUTNAM, BOWLING ALONE: THE COLLAPSE AND BEVIVAL OF AMERICAN COMMUNITY (2002); BRETT FRISCHMANN, INFRASTRUCTURE: INFRASTRUCTURE THE SOCIAL VALUE OF SHARED RESOURCES (2012).

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 undue harm. This harm must be mitigated. If the harm of scraping cannot be effectively mitigated through the application of existing privacy laws, it should be prohibited by new rules.

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. For the *Benefits Principle,* there must be articulable benefits to the collection, use, and transfer of personal data beyond merely generating profit for a company. More importantly, those benefits must be proportional or exceed the benefits to the scraper. Too often companies will offer some modest or trivial benefit like a mild efficiency for queues or organization as a pretext for information extract that is lucrative only for the scraper. Other times companies want to scrape so they can offer an important-sounding benefit that in practice is either illusory or so abstract as to be meaningless. "Keeping people safe" is a virtuous goal, but without so many AI surveillance systems don't meaningly provide safety to society and certainly not to marginalized and vulnerable groups like people of color who feel the brunt of surveillance first and hardest. So many AI tools are simply peddling snake oil.[317] 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.

This disproportionate extraction and retention of value violates the benefits principle and should be mitigated through better scraping rules. 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.[318] 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.[319] Another area of law that might help inform lawmakers and regulators might be the law of unjust

---

[317] *See, e.g.*, ARVIND NARAYANAN AND SAYASH KAPOOR, AI SNAKE OIL: WHAT ARTIFICIAL INTELLIGENCE CAN DO, WHAT IT CAN'T, AND HOW TO TELL THE DIFFERENCE (forthcoming 2024); MEREDITH BROUSSARD, ARTIFICIAL UNINTELLIGENCE: HOW COMPUTERS MISUNDERSTAND THE WORLD (2018); Louise Matsakis, *The Princeton researchers calling out 'AI snake oil',* SEMAPHOR, https://www.semafor.com/article/09/15/2023/the-princeton-researchers-calling-out-ai-snake-oil; *Keep your AI Claims in Check*, FTC (Feb. 27, 2023), https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check; see also Woodrow Hartzog, *Unfair and Deceptive Robots*, 74 Maryland Law Review 785 (2015).

[318] *See, e.g.*, DANIEL SOLOVE, THE DIGITAL PERSON (2006); ARI EZRA WALDMAN, PRIVACY AS TRUST (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 Haupt, *What Kind of Fiduciaries are Information Fiduciaries?*, 134 HARV. L. REV. F. 34 (2020); Lilian Edwards, *The Problem with Privacy*, 18 INT'L REV. OF L. COMPUTS. & TECH. 263 (2004); Ian Kerr, *The Legal Relationship Between Online Service Providers and Users*, 35 CAN. BUS. L.J. 419 (2001).

[319] *See, e.g.*, Jordan Francis, Woodrow Hartzog & Neil Richards, *A Concrete Proposal for Data Loyalty*, HARV. J. L. TECH (forthcoming 2024).

enrichment, restitution, and disgorgement.[320]

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 more fair, open, accountable, representative, and thoughtful. Privacy laws already require some of these things, such as requiring risk assessments and accountability. Many laws require fairness. 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."[321] 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 used among different groups. 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.[322]

Scraped data is a public concern because companies that scrape the data keep the surplus for themselves while imposing massive costs on society as externalities. Therefore, it is reasonable to impose process requirements on the practice to preserve the public interest. While we caution against treating data the same as rivalrous property, the collective wellbeing of people whose data is publicly available is somewhat like a public resource in that many people can benefit from it but it is also subject to abuse and often leads to wrongful gains. The law is, in essence, granting companies a license to farm public lands, and that license should not be unfettered. A good place to start considering the conditions upon which data may be scraped in the public interest might be the "Public Interest Privacy Principles," endorsed by 34 civil rights, consumer, and privacy organizations.[323] 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 equal opportunity.[324] Part of this means following the most robust version of the fair information practices as provided for in frameworks like the EU's GDPR. 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."[325]

---

[320] *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).

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

[322] *Id.*

[323] Public Interest Privacy Principles, https://pirg.org/resources/public-interest-privacy-principles/.

[324] *Id.*

[325] *Id.* at 2 ("Legislation ensure fundamental fairness of and transparency regarding automated decision-making. Automated decision-making, including in areas such as housing, employment,

Finally, the *Protections Principle* aims to avoid scraping exceptionalism. Scraped data should not be treated as second class. It should be afforded all the protections ordinarily provided by privacy laws, with the exception of protections that are unworkable. Scraped data should not lose all protections because it is publicly available. The law should continue to protect scraped data to the extent practicable. All provisions of privacy laws should apply to scrapers so they are not more free than recipients of data via contract. Lawmakers should also require reasonable anti-scraping safeguards as part of a company's overall duty to reasonably secure its entrusted personal data.[326]

Currently, scraping is mostly a lawless realm, where hardly anything limits scrapers and where scrapers have virtually no responsibilities. Scrapers should be treated similarly to other organizations that collect and use personal data.

CONCLUSION

Scraping and privacy are in desperate need of a reconciliation. Scraping is in conflict with nearly all core privacy principles. A ban on scraping is untenable, so a compromise must be reached. This compromise requires creativity to protect privacy in ways beyond many traditional privacy principles and laws.

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.").

[326] *See, e.g.*, DANIEL J. SOLOVE & WOODROW HARTZOG, BREACHED! WHY DATA SECURITY LAW FAILS AND HOW TO FIX IT (2022); William McGeveran, *The Duty of Data Security*, 103 Minn. L. Rev. 1135 (2019).