Michigan Law Review

Volume 120 | Issue 5

2022

Unfair Collection: Reclaiming Control of Publicly Available Personal Information from Data Scrapers

Andrew M. Parks
University of Michigan Law School

https://doi.org/10.36644/mlr.120.5.unfair

NOTE

UNFAIR COLLECTION: RECLAIMING CONTROL OF PUBLICLY AVAILABLE PERSONAL INFORMATION FROM DATA SCRAPERS

*Andrew M. Parks**

Rising enthusiasm for consumer data protection in the United States has resulted in several states advancing legislation to protect the privacy of their residents' personal information. But even the newly enacted California Privacy Rights Act (CPRA)—the most comprehensive data privacy law in the country—leaves a wide-open gap for internet data scrapers to extract, share, and monetize consumers' personal information while circumventing regulation. Allowing scrapers to evade privacy regulations comes with potentially disastrous consequences for individuals and society at large.

This Note argues that even publicly available personal information should be protected from bulk collection and misappropriation by data scrapers. California should reform its privacy legislation to align with the European Union's General Data Privacy Regulation (GDPR), which requires data scrapers to provide notice to data subjects upon the collection of their personal information regardless of its public availability. This reform could lay the groundwork for future legislation at the federal level.

TABLE OF CONTENTS

* J.D. Candidate, May 2022, University of Michigan Law School. I am grateful to Professor Barbara McQuade for her wonderful input, insights, and encouragement. Thank you to Emily Grau for her thoughtful comments throughout the writing of this piece. Thank you to my family, especially George Parks, for their support. Finally, thank you to all the members of the *Michigan Law Review* Volume 120 Notes Office for their invaluable feedback and edits.

 doi:10.36644/mlr.120.5.unfair

INTRODUCTION

In January 2021, a software engineer in New York City scoured dozens of city and state websites attempting to schedule a COVID-19 vaccination for his mother.[1] At that time, there was no uniform system for scheduling vaccination appointments. The city and state appointment systems were completely different, each with its own sign-up protocol.[2] Frustrated with this convoluted system, the engineer decided to develop a solution. In less than two weeks, he launched TurboVax, "a free website that compiles availability from the three main city and state New York vaccine systems and sends the information in real time to Twitter."[3] Because vaccine appointment information was publicly available on the internet, TurboVax could access this information using a computer program called a "bot." This bot automatically checked, copied, and republished appointment data in bulk, avoiding the need to manually check

1. Sharon Otterman, *N.Y.'s Vaccine Websites Weren't Working. He Built a New One for $50.*, N.Y. TIMES (May 11, 2021), https://www.nytimes.com/2021/02/09/nyregion/vaccine-website-appointment-nyc.html [perma.cc/3THW-EU4J].

2. *Id.*; *see also* Ron Lieber, *How to Get the Coronavirus Vaccine in New York City*, N.Y. TIMES (Mar. 22, 2021), https://www.nytimes.com/article/nyc-vaccine-shot.html [perma.cc/Z9CL-QDWM].

3. Otterman, *supra* note 1; *see also* @turbovax, TWITTER, https://twitter.com/turbovax [perma.cc/UF4R-Y94J].

government websites for available slots.[4] The process that TurboVax used to extract vast amounts of data from the internet is called "scraping."[5]

It's one thing to scrape the internet for publicly available information when the content extracted is not associated with an individual's personal information, but quite another when it is. When a LinkedIn user creates a public profile to search for employment, she may well include her phone number, email address, and a photo of her face. Although this information is technically "public," she might reasonably expect this information to remain personal to her and within her control. She may, for instance, list her LinkedIn profile publicly while searching for a job but later set it to "private" after securing employment. Yet all her personal data—her name, phone number, email address, and photo—were, at least for some time, made public and therefore susceptible to extraction and reappropriation by scrapers.[6] And this bell cannot be unrung.[7]

Over the past few years, consumer data privacy legislation has surfaced across the United States. The California Consumer Privacy Act (CCPA)[8] and the California Privacy Rights Act (CPRA),[9] for instance, now regulate the collection of consumer personal data and the sharing of such data with third parties. But no currently proposed or enacted privacy statute adequately protects publicly available personal information.[10] All of it is exempted, making it fair game to be scraped, used, shared, or sold. Many scholars have written about data scraping and its legality under the Computer Fraud and Abuse Act.[11]

---

4. Email from Huge Ma, Dev., TurboVax, to author (Feb. 10, 2021, 6:27 PM) (on file with the *Michigan Law Review*) (confirming that TurboVax uses scraping technology to replicate communication between a user's browser and server if one were to look up vaccination appointment availability); *see also Frequently Asked Questions*, TURBOVAX, https://www.turbovax.info/faq [perma.cc/A66Z-JDNN]; Dana Schulz, *This Website Wants to Centralize Vaccine Appointments for the Entire Country*, 6SQFT (Feb. 25, 2021), https://www.6sqft.com/vaccinefinder-covid-vaccination-appointments-national-website [perma.cc/DYA4-3TD4].

5. *See* Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372, 381–82 (2018); Marissa Boulanger, Case Note, *Scraping the Bottom of the Barrel: Why It Is No Surprise That Data Scrapers Can Have Access to Public Profiles on LinkedIn*, 21 SMU SCI. & TECH. L. REV. 77, 77–78 (2018).

6. *See* hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 1003–04 (9th Cir. 2019), *vacated*, 141 S. Ct. 2752 (2021) (finding scraping personal data did not violate the Computer Fraud and Abuse Act where the website scraped was public and not password protected).

7. Alexander Tsesis, *Data Subjects' Privacy Rights: Regulation of Personal Data Retention and Erasure*, 90 U. COLO. L. REV. 593, 600 (2019) ("Once a person reveals details about such things as personal location, shopping habits, sexuality, sex, education, travel plans, and an infinite number of similarly revealing data points, the subject becomes almost powerless to demand that social media companies purge all collected and tracked information.").

8. CAL. CIV. CODE §§ 1798.100–.198 (West Supp. 2021) (amended 2020).

9. CAL. CIV. CODE §§ 1798.100–.199.95 (West Supp. 2021) (effective Jan. 1, 2023).

10. *See infra* Section II.B.

11. *See* Sellars, *supra* note 5; Boulanger, *supra* note 5; Jacquellena Carrero, Note, *Access Granted: A First Amendment Theory of Reform of the CFAA Access Provision*, 120 COLUM. L. REV. 131 (2020); Tess Macapinlac, Note, *The Legality of Web Scraping: A Proposal*, 71 FED.

Others have discussed various consumer data privacy statutes and proposals across the United States and Europe.[12] But few have addressed the privacy implications of scraping publicly available personal information,[13] and no one has proposed a reform to regulate such activity in the United States. This Note does just that.

Part I of this Note defines data scraping, explains its purposes, and summarizes its current legality. Part II argues that publicly available personal information should be protected from data scrapers, analyzes the current landscape of state and federal consumer data privacy legislation, and explains why existing and proposed solutions are inadequate to address this issue. It also describes how publicly available personal information is handled by the European Union's General Data Protection Regulation (GDPR). Part III argues that while passing legislation at the federal level could be desirable, California ought to amend its privacy laws to incorporate GDPR-style protections for publicly available personal information. Specifically, California should regulate the collection of publicly available personal information based on whether the information collected can be anonymized, whether the information is collected in bulk, and whether the information is collected for commercial purposes.

I. DATA SCRAPING AND ITS CURRENT LEGALITY

To understand the privacy implications of data scraping, it is necessary to explain its function and legality. Scraping has many useful applications, and it is often employed by individuals serving the public interest. Unfortunately, scraping can also be used for malicious purposes, and businesses frequently attempt to block or deter parties from scraping their websites. As such, Part I concludes by examining the most common legal claims available to address scraping.

COMMC'NS L.J. 399 (2019); Jennie E. Christensen, Note, *The Demise of the CFAA in Data Scraping Cases*, 34 NOTRE DAME J.L. ETHICS & PUB. POL'Y 529 (2020); Zachary Gold & Mark Latonero, *Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping*, 13 WASH. J.L. TECH. & ARTS 275 (2018); 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).

12. *See* Stuart L. Pardau, *The California Consumer Privacy Act: Towards a European-Style Privacy Regime in the United States?*, 23 J. TECH. L. & POL'Y 68 (2018); Alexandria J. Saquella, Comment, *Personal Data Vulnerability: Constitutional Issues with the California Consumer Privacy Act*, 60 JURIMETRICS 215 (2020); Joanna Kessler, Note, *Data Protection in the Wake of the GDPR: California's Solution for Protecting "The World's Most Valuable Resource,"* 93 S. CAL. L. REV. 99 (2019); Jordan Yallen, Comment, *Untangling the Privacy Law Web: Why the California Consumer Privacy Act Furthers the Need for Federal Preemptive Legislation*, 53 LOY. L.A. L. REV. 787 (2020); W. Gregory Voss & Kimberly A. Houser, *Personal Data and the GDPR: Providing a Competitive Advantage for U.S. Companies*, 56 AM. BUS. L.J. 287 (2019).

13. *See* Geoffrey Xiao, Note, *Bad Bots: Regulating the Scraping of Public Personal Information*, 34 HARV. J.L. & TECH. 702 (2021).

A. *Scraping: Definition, Usage, and Purposes*

Data scraping is the process of scanning and extracting large amounts of data from one or more websites using a software program often referred to as a "bot," "robot," or "scraper."[14] Scraping is different from "hacking," which involves breaking into another person's "computer, network, servers, or database,"[15] typically by cracking a password or exploiting a vulnerability in the website's code.[16] Scrapers, by contrast, extract publicly available data[17] and thus have no need to break into private servers.

Scraping has many beneficial purposes. It can be used to preserve websites, conduct research, compare product and price information from various sources, gather contact and social media data for outreach campaigns, track company reputation, and aggregate news and other content on curated websites.[18] Journalists use scraping technology to gather and analyze massive chunks of statistical data.[19] Scholars employ scraping technology to aid their academic research.[20] Advertisers use scraping technology to collect contact details and public posts on social media websites to better market their products to consumers.[21]

Although scraping has beneficial applications, scraping technology can also be used for malicious purposes, such as spamming email accounts, causing website crashes,[22] or conducting scams.[23] Exemplifying morally questionable use of data scraping technology is the company Clearview AI.[24] Clearview

---

14. *See* Boulanger, *supra* note 5, at 77–78; Sellars, *supra* note 5, at 381–82. Additionally, *Black's Law Dictionary* defines "screen-scraping" as "[t]he practice of extracting data directly from one website and displaying it on another website." *Screen-Scraping*, BLACK'S LAW DICTIONARY (11th ed. 2019). To avoid confusion, this Note's use of "scraping technology" refers to the bots used to scrape content from the internet, and its use of "scrapers" refers to the individuals or entities employing scraping technology. Notably, though, others often also use "scrapers" to refer to the bots used for scraping.

15. *Hack*, BLACK'S LAW DICTIONARY (11th ed. 2019).

16. *See* Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes*, 78 N.Y.U. L. REV. 1596, 1644–45 (2003).

17. *See* Macapinlac, *supra* note 11, at 401–02.

18. Sellars, *supra* note 5, at 374; Michael Keating, *Understanding the History of Web Scraping*, OCTATOOLS (Feb. 24, 2016), https://octatools.com/understanding-history-web-scraping [perma.cc/K8JX-GCHH].

19. Keating, *supra* note 18.

20. For instance, in *Sandvig v. Sessions*, four professors sought a declaratory judgment that scraping for research purposes did not violate the Computer Fraud and Abuse Act. 315 F. Supp. 3d 1, 8–10 (D.D.C. 2018).

21. *See* Keating, *supra* note 18.

22. Boulanger, *supra* note 5, at 78.

23. *See* Kevin Collier, *Why Cybercriminals Looking to Steal Personal Info Are Using Text Messages as Bait*, NBC NEWS (May 6, 2021, 9:44 PM), https://www.nbcnews.com/tech/security/scam-text-messages-are-rampant-no-easy-fix-rcna840 [perma.cc/2VZQ-WLZ5].

24. *See* Sam duPont, *On Facial Recognition, the U.S. Isn't China—Yet*, LAWFARE (June 18, 2020, 8:01 AM), https://www.lawfareblog.com/facial-recognition-us-isnt-china-yet [perma.cc/2X4C-63HB].

scrapes billions of personal images posted on Facebook and other websites for use in its facial recognition software.[25] It then sells its software to law enforcement agencies, allowing police departments to "compare a face captured on a security camera against [Clearview's] database to reveal possible matches."[26] No user consents to Clearview's collection, and even if the image is later removed from the public site, Clearview keeps a copy.[27] Significantly, cease-and-desist letters from Google, YouTube, Venmo, and LinkedIn have failed to stop Clearview from scraping.[28] Clearview has ignored the letters and maintains that it has a First Amendment right to access publicly available information.[29] Clearview's facial recognition software has been used by thousands of law enforcement agencies, companies, and individuals around the world.[30]

Scraping technology is also deployed problematically in the "mugshot industry."[31] In this industry, private companies use bots to scrape booking photos of arrested persons from publicly accessible law enforcement websites. The companies then display the photos in "mugshot galleries" on their websites.[32] Scraping enables the companies to monetize the mugshots in various ways, such as hosting advertisements on their websites, charging visitors a fee to search their mugshot database, and—most controversially—charging subjects

25. *Id.*

26. *Id.*

27. *Google, YouTube, Venmo and LinkedIn Send Cease-and-Desist Letters to Facial Recognition App That Helps Law Enforcement*, CBS NEWS (Feb. 5, 2020, 6:52 PM), https://www.cbsnews.com/news/clearview-ai-google-youtube-send-cease-and-desist-letter-to-facial-recognition-app [perma.cc/4JGW-2HW2].

28. *Id.*

29. *Id.*; Katelyn Ringrose & Divya Ramjee, *Watch Where You Walk: Law Enforcement Surveillance and Protester Privacy*, 11 CALIF. L. REV. ONLINE 349, 361 (2020).

30. Ryan Mac, Caroline Haskins & Logan McDonald, *Clearview's Facial Recognition App Has Been Used by the Justice Department, ICE, Macy's, Walmart, and the NBA*, BUZZFEED NEWS (Feb. 27, 2020, 11:37 PM), https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement [perma.cc/W228-5GU9]. A June 2021 report by the Government Accountability Office revealed that Clearview's facial recognition software was used by at least ten federal government agencies, including Customs and Border Protection, ICE, the FBI, and the Secret Service. U.S. GOV'T ACCOUNTABILITY OFF., GAO-21-518, FACIAL RECOGNITION TECHNOLOGY: FEDERAL LAW ENFORCEMENT AGENCIES SHOULD BETTER ASSESS PRIVACY AND OTHER RISKS 12 (2021), https://www.gao.gov/assets/gao-21-518.pdf [perma.cc/273T-MNDB].

31. *See* Eumi K. Lee, *Monetizing Shame: Mugshots, Privacy, and the Right to Access*, 70 RUTGERS U. L. REV. 557, 566–69 (2018).

32. *Id.* at 563, 566; *see also* Allen Rostron, Commentary, *The Mugshot Industry: Freedom of Speech, Rights of Publicity, and the Controversy Sparked by an Unusual New Type of Business*, 90 WASH. U. L. REV. 1321, 1323–24 (2013).

large fees to have their mugshots removed.[33] Even if an arrested person's criminal record is expunged, their scraped mugshot can appear in Google search results and be dispersed across dozens of websites.[34]

To prevent scraping, website owners often prohibit the practice in their website's terms of service[35] or implement technological barriers. One such barrier is the installation of a "robots.txt" file—a widely used protocol that instructs specified bots to ignore certain files when crawling or scraping a website—to their website's root directory.[36] However, these technological barriers do not always effectively deter scraping. And as Section I.B will explain, the most common legal barriers to scraping do little to deter scraping publicly available personal information.

B. *The Current Legal Landscape of Data Scraping*

In the United States, litigation that responds to data scraping typically involves the following claims: (1) Computer Fraud and Abuse Act (CFAA) claims for scraping data "without authorization" or "exceed[ing] authorized access"; (2) state and federal copyright-infringement claims; and (3) common law trespass-to-chattels and breach-of-contract claims. [37] While scholars have written extensively about whether these causes of action effectively deter or prevent scraping in general,[38] this Section instead focuses specifically on the failure of these causes of action to protect publicly available personal information.

1. Claims Under the Computer Fraud and Abuse Act

The Computer Fraud and Abuse Act (CFAA) imposes liability on anyone who "intentionally accesses a computer without authorization or exceeds authorized access[] and thereby obtains . . . information from any protected computer."[39] In *hiQ Labs, Inc. v. LinkedIn Corp.*, a data company used bots to

33. Rostron, *supra* note 32, at 1324–25; *see also* Lee, *supra* note 31, at 568 (describing "reputation management" companies that charge fees ranging from the low hundreds up to thousands of dollars to remove mugshot images from the internet).

34. Sarah Esther Lageson, *There's No Such Thing as Expunging a Criminal Record Anymore*, SLATE (Jan. 7, 2019, 2:44 PM), https://slate.com/technology/2019/01/criminal-record-expungement-internet-due-process.html [perma.cc/7LAX-3AUE].

35. Christensen, *supra* note 11, at 533 ("Companies often attempt to limit scraping of their data through their website's terms and conditions."); *see also* Sw. Airlines Co. v. Roundpipe, LLC, 375 F. Supp. 3d 687, 690 (N.D. Tex. 2019).

36. Sellars, *supra* note 5, at 413–14.

37. Zamora, *supra* note 11, at 205, 210.

38. *See, e.g.*, Boulanger, *supra* note 5, at 78–81; Zamora, *supra* note 11, at 210–24; Christensen, *supra* note 11, at 531–35; Kathleen C. Riley, Note, *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, 265–279 (2018); Han-Wei Liu, *Two Decades of Laws and Practice Around Screen Scraping in the Common Law World and Its Open Banking Watershed Moment*, 30 WASH. INT'L L.J. 28, 32–44 (2020).

39. Computer Fraud and Abuse Act, 18 U.S.C. § 1030(a)(2)(C).

scrape information that LinkedIn users included on their public profiles, such as their name, job title, work history, and skills.[40] The Ninth Circuit found that this scraping did not violate the CFAA even though LinkedIn prohibits users from scraping its website in its terms of service and employs technological barriers to block scraping.[41] Instead, the court held that scraping only triggers liability under the CFAA when a website is private or password protected and a user circumvents this barrier to scrape data anyway.[42] LinkedIn then filed a petition for a writ of certiorari to the Supreme Court.[43]

While LinkedIn's petition was pending, the Supreme Court decided *Van Buren v. United States*, its first case interpreting the CFAA.[44] In *Van Buren*, the Court considered whether a police officer who accessed a computer for an improper *purpose* "exceed[ed] authorized access" in violation of the CFAA.[45] Holding that accessing a computer for an improper purpose does not violate the CFAA, the Court adopted a "gates-up-or-down" approach: a person violates the CFAA by bypassing a "gate" that is down that the person is not supposed to bypass.[46] In other words, a person needs to enter "particular areas of the computer—such as files, folders, or databases—that are off limits to him" for liability to follow.[47]

After issuing this ruling, the Supreme Court granted LinkedIn's petition for writ of certiorari in *hiQ Labs*.[48] Upon review, the Court vacated the Ninth Circuit's opinion and remanded the case for further consideration in light of the Court's ruling in *Van Buren*.[49] But applying *Van Buren*'s "gates-up-or-down" inquiry to *hiQ Labs* will probably not change its outcome. The data scraped on LinkedIn's website were publicly accessible and not protected by a password. The "gates," therefore, were not down. As such, a person who scrapes data from a publicly accessible website likely does not violate the CFAA because that person has not bypassed a "gate" barring access to publicly available data.

---

40. 938 F.3d 985, 991 (9th Cir. 2019), *vacated*, 141 S. Ct. 2752 (2021).

41. *See id.* at 1001–03.

42. *Id.* at 1001 ("[A]uthorization is only required for password-protected sites or sites that otherwise prevent the general public from viewing the information.").

43. *See* Zarish Baig & Kristin L. Bryan, *hiQ LinkedIn Data Scraping CFAA Ruling Delayed Pending SCOTUS Decision*, NAT'L L. REV. (Apr. 26, 2021), https://www.natlawreview.com/article/hiq-linkedin-data-scraping-cfaa-ruling-delayed-pending-scotus-decision [perma.cc/NHX3-UPKT].

44. 141 S. Ct. 1648 (2021).

45. *Id.* at 1662.

46. *Id.* at 1658–59 ("[O]ne either can or cannot access a computer system, and one either can or cannot access certain areas within the system.").

47. *Id.* at 1662; *see also* Orin Kerr, *The Supreme Court Reins In the CFAA in* Van Buren, LAWFARE (June 9, 2021, 9:04 PM), https://www.lawfareblog.com/supreme-court-reins-cfaa-van-buren [perma.cc/ADE3-ZREZ].

48. LinkedIn Corp. v. hiQ Labs, Inc., 141 S. Ct. 2752 (2021).

49. *Id.*

In *Sandvig v. Sessions*, the plaintiffs argued that researchers' use of data-scraping tools constituted access "without authorization" in violation of the CFAA.[50] Because the data sought were publicly available, the court stated that "[e]mploying a bot to crawl a website . . . may run afoul of a website's [terms of service], but it does not constitute an *access* violation when the human who creates the bot is otherwise allowed to read and interact with that site."[51] Given these rulings, it is unlikely that the CFAA presents any meaningful barrier to scraping publicly available personal data.

2. Copyright Infringement, Trespass to Chattels, and Breach of Contract Claims

Like claims brought under the CFAA, claims of copyright infringement, breach of contract, and trespass to chattels are unlikely to protect individuals' publicly available personal information from scrapers. First, when the data includes personal information—for example, an individual's name, address, email address, phone number, geolocation data, or internet browsing history—courts tend to find that the scraping does not constitute copyright infringement because facts are not copyrightable.[52] Copyright law distinguishes noncopyrightable facts from copyrightable works of authorship that are independently created by the author and possess at least a minimal degree of creativity.[53] One district court has held that scraping data from Southwest Airlines' website did not constitute copyright infringement because "[f]are, route and scheduling information are all facts and thus not copyrightable."[54] Personal data similarly are facts, not works of authorship, suggesting that copyright law cannot serve as a remedy for this kind of data scraping.

Second, scraping could constitute trespass to chattels—intentional interference with another's personal property[55]—if the bots used for scraping impede the website owner's ability to use portions of its servers.[56] These claims may provide an effective method for website *owners* to deter scrapers from impermissibly collecting data from their websites.[57] But because most individ-

50. 315 F. Supp. 3d 1, 8–10 (D.D.C. 2018).

51. *Sandvig*, 315 F. Supp. 3d at 27.

52. *See* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 344–48 (1991).

53. *See id.* at 344–51.

54. Sw. Airlines Co. v. Farechase, Inc., 318 F. Supp. 2d 435, 437, 440–41 (N.D. Tex. 2004).

55. eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1069 (N.D. Cal. 2000).

56. *See, e.g.*, Craigslist Inc. v. 3Taps Inc., 942 F. Supp. 2d 962, 980–81 (N.D. Cal. 2013); *eBay*, 100 F. Supp. 2d at 1069–70; Register.com, Inc. v. Verio, Inc., 356 F.3d 393, 404–05 (2d Cir. 2004).

57. *See, e.g.*, *Craigslist*, 942 F. Supp. 2d at 966–67, 980 (suggesting that scraping could constitute trespass to chattels where defendant continued scraping despite cease-and-desist letters and where defendant's unauthorized interference allegedly "reduce[d plaintiff]'s capacity to service its users because it occupie[d] and use[d plaintiff]'s resources"); *eBay*, 100 F. Supp. 2d at

uals are not website owners and do not host their own data, they have no trespass to chattels claim to bring against scrapers who trespass upon or impede access to web servers.

Finally, data scraping may constitute breach of contract when a website's terms of service expressly prohibit scraping and users scrape data anyway.[58] The enforceability of an antiscraping provision in a website's terms of service often depends on whether the agreement required the scraper to affirmatively manifest assent to its terms.[59] Even if it did, the terms ordinarily bind only the parties to the agreement—the website owner and the scraper. Thus, such agreements would not necessarily create any cause of action for individuals whose personal information is scraped from a website.[60]

II. THE DATA SCRAPING LOOPHOLE

Part II of this Note argues that, where an individual's personal information is concerned, scraping of even *publicly available* personal information should be regulated. While there are existing state and federal consumer data privacy laws in the United States, data scraping circumvents these proposed solutions, rendering them inadequate to address this issue. In contrast, the European Union's General Data Protection Regulation (GDPR) provides a more robust model for amending the American legal framework on data privacy.

A. *Publicly Available Personal Information Should Be Protected*

Even when the information is publicly available, scraping personal information is problematic. In the absence of statutory and other legal protections for personal information, courts have held that scraping personal information is permissible so long as the information is publicly accessible.[61] But personal information may be made public for various reasons, often without the

1060–62, 1070–72 (finding plaintiff likely to prevail on its trespass claim where defendant's scraping may have diminished the quality of plaintiff's computer systems and bandwidth).

58. Christensen, *supra* note 11, at 533.

59. *See* Nguyen v. Barnes & Noble Inc., 763 F.3d 1171, 1175–79 (9th Cir. 2014) (holding an agreement unenforceable because its terms were buried in a hyperlink in the bottom corner of the website and the site "provide[d] no notice to users nor prompt[ed] them to take any affirmative action to demonstrate assent"). *But see Verio*, 356 F.3d at 403 (finding that scraper assented to website's terms by accessing the website despite not clicking a button specifically agreeing to the website's terms).

60. *But see* QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576, 588 (E.D. Pa. 2016) (holding that website owner was a third-party beneficiary of an agreement between a scraper and defendant where defendant permitted the scraper to "transmit malicious and unsolicited software . . . [and] us[e] a device, program, or robot" against the plaintiff's website (alterations in original)).

61. *See, e.g.*, hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 1003–04 (9th Cir. 2019), *vacated*, 141 S. Ct. 2752 (2021); Sandvig v. Sessions, 315 F. Supp. 3d 1, 26–27 (D.D.C. 2018); Sandvig v. Barr, 451 F. Supp. 3d 73, 86–89 (D.D.C. 2020).

knowledge or consent of the subject.[62] And even when a subject *voluntarily* makes her information public, she likely does so without meaningful consent and without considering the potentially damaging implications of such a decision, both for herself and society at large.

1. Information Made Public Without the Subject's Knowledge or Consent

In many cases, an individual never knows that their personal information has been made public, making it impossible for them to consent to its publication. Some personal information is made public through lawful government public records. For example, the Federal Election Commission (FEC) website publicly displays federal political campaign contributions. These data include each contributor's full name, mailing address, occupation, employer, and contribution amount.[63] Yet individuals probably do not realize they are publicly disclosing all of this personally identifiable information when they donate to a campaign.

An individual's personal information might also be made public when a third party publishes it online without their consent. This sometimes takes the form of "doxing"—a kind of cyber harassment involving "the public release of personal information that can be used to identify or locate an individual."[64]

Finally, personal information might also be made public as a result of a data breach.[65] Hackers frequently sell databases of stolen data records from businesses on the dark web for large sums of money.[66] If businesses delay or choose not to disclose the cyber breach to consumers, the consumers may never know their information was hacked and potentially made public.[67]

---

62. For example, information may be made public when an individual is "doxed." *See* Alexander J. Lindvall, *Political Hacktivism: Doxing & the First Amendment*, 53 CREIGHTON L. REV. 1, 2 (2019).

63. *See, e.g.*, *Receipts*, FED. ELECTION COMM'N, https://www.fec.gov/data/receipts [perma.cc/3SUE-RZWL] (recording personal information of campaign donors, including full name, mailing address, employer, occupation, date of contribution, and contribution amount).

64. Julia M. MacAllister, Note, *The Doxing Dilemma: Seeking a Remedy for the Malicious Publication of Personal Information*, 85 FORDHAM L. REV. 2451, 2453 (2017); Lindvall, *supra* note 62, at 2.

65. For example, Equifax suffered a data breach in 2017 that may have compromised the sensitive information of 143 million American consumers. Tara Siegel Bernard, Tiffany Hsu, Nicole Perlroth & Ron Lieber, *Equifax Says Cyberattack May Have Affected 143 Million in the U.S.*, N.Y. TIMES (Sept. 7, 2017), https://www.nytimes.com/2017/09/07/business/equifax-cyberattack.html [perma.cc/B2Y2-Z9BK].

66. *See* Davey Winder, *Hacker Gives Away 386 Million Stolen Records on Dark Web—What You Need to Do Now*, FORBES (July 29, 2020, 5:15 AM), https://www.forbes.com/sites/daveywinder/2020/07/29/hacker-gives-away-386-million-stolen-records-on-dark-web-what-you-need-to-do-now-shinyhunters-data-breach [perma.cc/5DSM-2ZSV].

67. *See* Renae Merle, *Yahoo Fined $35 Million for Failing to Disclose Cyber Breach*, WASH. POST (Apr. 24, 2018), https://www.washingtonpost.com/news/business/wp/2018/04/24/yahoo-fined-35-million-for-failing-to-disclose-cyber-breach [perma.cc/CDB7-UD2V].

Some have argued that personal data should be treated like property, owned and controlled by the individual.[68] Although current U.S. law does not recognize any definitive right of ownership to data,[69] users nevertheless might naively believe that they control theirs. After all, they can control whether to set their social profiles to "public" or "private," and they decide whether to hide or archive content previously posted publicly. That an individual's personal information has been published publicly on the internet should not automatically grant internet data scrapers carte blanche authority to extract, reappropriate, or monetize it. Personal information is just that: *personal*.

2. Information Made Public Voluntarily Should Still Be Protected

Even when an individual voluntarily makes her information public, she still retains a privacy interest in controlling it. A dissent penned by Justice Gorsuch in a different legal context—government collection of personal information from third parties for criminal investigations—provides helpful insights:

> [T]he fact that a third party has access to or possession of your papers and effects does not necessarily eliminate your interest in them. Ever hand a private document to a friend to be returned? Toss your keys to a valet at a restaurant? Ask your neighbor to look after your dog while you travel? You would not expect the friend to share the document with others; the valet to lend your car to his buddy; or the neighbor to put Fido up for adoption.[70]

This reasoning can be extended. For example, just because a user posts her home address on a publicly available website does not eliminate her interest in later preserving the privacy of that information. She may have made the post public only temporarily. She may have accidentally posted it publicly when she intended it to be private. Or she may have posted it to her private profile—specifically electing to make the information viewable only by a select group of friends on her account—and yet one of those friends with access may have reposted or redistributed her information publicly. In each of these scenarios, her interest in preserving the privacy of her personal information should not be completely eliminated merely because it wound up publicly accessible at least for some time.

Unless a user affirmatively changes her privacy settings on the websites and social media platforms to which she gives her data, third parties can probably access her information. Most social networking platforms make users'

---

68. *See* Jeffrey Ritter & Anna Mayer, *Regulating Data as Property: A New Construct for Moving Forward*, 16 DUKE L. & TECH. REV. 220, 223 (2018) ("This article offers a bold proposition: An explicit, legal mechanism to establish, claim and transfer property rights in data must be adopted."); will.i.am, *We Need to Own Our Data as a Human Right—and Be Compensated for It*, ECONOMIST (Jan. 21, 2019), https://www.economist.com/open-future/2019/01/21/we-need-to-own-our-data-as-a-human-right-and-be-compensated-for-it [perma.cc/X336-LKCR].

69. *See* Ritter & Mayer, *supra* note 68, at 251.

70. Carpenter v. United States, 138 S. Ct. 2206, 2268 (2018) (Gorsuch, J., dissenting).

content publicly accessible by default.[71] But even if a prudent person were to set her profile to "private" to hide her personal information from public view, data scrapers might still be able to access it.[72] And on non-social media websites that limit access to those with login credentials, a scraper would only need to sign up for an account to gain access.[73] LinkedIn's privacy policy warns:

> Please do not post or add personal data to your profile that you would not want to be publicly available. . . . Your profile is fully visible to all Members and customers of our Services. Subject to your settings, it can also be visible to others on or off of our Services (e.g., Visitors to our Services or users of third-party search engines).[74]

Still, if a website includes a warning about data scraping, studies suggest users are unlikely to take heed. A 2017 survey of two thousand U.S. consumers found that 91 percent of people consent to terms of service without reading them.[75] For those aged 18 to 34, the rate was 97 percent.[76] In light of these statistics, it would be imprudent to conclude that the average user realizes that she has knowingly consented to scraping by bots if she accidentally posts a photo of herself publicly on Instagram.[77]

3. The Dangers of Allowing the Scraping of Personal Information in Bulk

What flows from scrapers' ability to extract individuals' publicly available personal data is alarming. At its most innocuous, data scraping permits third

---

71. *See, e.g.*, *About Public and Protected Tweets*, TWITTER, https://help.twitter.com/en/safety-and-security/public-and-protected-tweets [perma.cc/UNL3-HHVU].

72. Twitter's privacy pages contemplate such a scenario: "Protected Tweets [are o]nly visible to your Twitter followers. Please keep in mind, your followers may still capture images of your Tweets and share them." *Id.*

73. *See* Sandvig v. Barr, 451 F. Supp. 3d 73, 89, 92 (D.D.C. 2020).

74. *Privacy Policy*, LINKEDIN (Aug. 11, 2020), https://www.linkedin.com/legal/privacy-policy [perma.cc/9YGH-L9LM].

75. Jessica Guynn, *What You Need to Know Before Clicking 'I Agree' on That Terms of Service Agreement or Privacy Policy*, USA TODAY (Jan. 29, 2020, 2:21 PM), https://www.usatoday.com/story/tech/2020/01/28/not-reading-the-small-print-is-privacy-policy-fail/4565274002 [perma.cc/CGN9-BWF5].

76. *Id.*

77. Such a lack of knowledge or consent is exacerbated when considering the number of young social media users. For example, of TikTok's forty-nine million daily users in the United States, more than a third are fourteen years old or younger. Raymond Zhong & Sheera Frenkel, *A Third of TikTok's U.S. Users May Be 14 or Under, Raising Safety Questions*, N.Y. TIMES (Sept. 17, 2020), https://www.nytimes.com/2020/08/14/technology/tiktok-underage-users-ftc.html [perma.cc/4B49-HW6E]. Federal law nominally prevents website operators from collecting certain data from children. *See* Children's Online Privacy Protection Act of 1998 (COPPA), 15 U.S.C. § 6502. However, many children on the internet lie about their age. Mark Sweney, *More Than 80% of Children Lie About Their Age to Use Sites like Facebook*, GUARDIAN (July 25, 2013, 7:01 PM), https://www.theguardian.com/media/2013/jul/26/children-lie-age-facebook-asa [perma.cc/Y9LT-WJ2J].

parties to monetize our personal information without our knowledge or consent. At its most dangerous, it has the potential to vastly restrict liberty, undermine democracy, and even put people in physical danger.

The following examples highlight the very different but equally dangerous consequences of data scraping. In February 2019, an African American man named Nijeer Parks was falsely accused of shoplifting and attempting to hit a police officer with a car outside a motel in Woodbridge, New Jersey.[78] He spent ten days in jail and paid around $5,000 to defend his case.[79] Parks's arrest stemmed from facial recognition software misidentifying him.[80] Parks later sued the city's police department, alleging that it used facial recognition software from Clearview AI.[81] While it is still unclear whether Clearview AI was used in his apprehension,[82] the mass scraping of publicly available personal data can lead to misidentification resulting in false arrests, jail time, and thousands of dollars in attorney fees.[83]

---

78. Kashmir Hill, *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*, N.Y. TIMES (Jan. 6, 2021), https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html [perma.cc/5WQD-55WX].

79. *Id.*

80. *Id.* Indeed, a study published by the National Institute of Standards and Technology found empirical evidence showing that most facial recognition software programs exhibit racial bias, producing higher rates of false positives for Asian and African American faces compared to images of Caucasian faces. PATRICK GROTHER, MEI NGAN & KAYEE HANAOKA, NAT'L INST. OF STANDARDS & TECH., U.S. DEP'T OF COM., FACE RECOGNITION VENDOR TEST (FRVT) PART 3: DEMOGRAPHIC EFFECTS 2–3 (2019), https://doi.org/10.6028/NIST.IR.8280; Catherine Thorbecke, *After Federal Study Finds Racial Bias in Facial Recognition Tech, Advocates Renew Calls for Ban*, ABC NEWS (Dec. 20, 2019, 2:31 PM), https://abcnews.go.com/Business/federal-study-finds-racial-bias-facial-recognition-tech/story?id=67853261 [perma.cc/UK9R-J7J6]; *see also* Elaisha Stokes, *Wrongful Arrest Exposes Racial Bias in Facial Recognition Technology*, CBS NEWS (Nov. 19, 2020, 7:00 AM), https://www.cbsnews.com/news/detroit-facial-recognition-surveillance-camera-racial-bias-crime [perma.cc/LM67-4A73].

81. Complaint at ¶ 29, Parks v. McCormack, No. 2:21-cv-04021-MCA-LDW (N.J. Super. Ct. Law Div. Nov. 25, 2020); Hill, *supra* note 78; *see also supra* notes 24–30 and accompanying text. It should be noted that in 2020, New Jersey's attorney general barred police officers in the state from using the Clearview AI app. Kashmir Hill, *New Jersey Bars Police from Using Clearview Facial Recognition App*, N.Y. TIMES (Jan. 24, 2020), https://www.nytimes.com/2020/01/24/technology/clearview-ai-new-jersey.html [perma.cc/R7NZ-4VM2]. And in February 2021, Canada determined that Clearview AI's scraping of biometric information violated Canada's privacy laws. Kashmir Hill, *Clearview AI's Facial Recognition App Called Illegal in Canada*, N.Y. TIMES (Feb. 3, 2021), https://www.nytimes.com/2021/02/03/technology/clearview-ai-illegal-canada.html [perma.cc/3XZ6-Z4P6].

82. Curiously, Parks's amended complaint dropped any mention of Clearview AI. *See* Second Amended Complaint, Parks v. McCormack, No. 2:21-cv-04021-MCA-LDW (N.J. Super. Ct. Law Div. June 1, 2021).

83. *See* Kristin L. Bryan & Christina Lamoureux, *Government Users of Facial Recognition Software Sued by Plaintiff Alleging Wrongful Imprisonment over Case of Mistaken Identity*, NAT'L L. REV. (Jan. 4, 2021), https://www.natlawreview.com/article/government-users-facial-recognition-software-sued-plaintiff-alleging-wrongful [perma.cc/R5DN-ECX2]; Stokes, *supra* note 80 (discussing fallout of false arrest based on erroneous match in facial recognition software); Drew Harwell, *Wrongfully Arrested Man Sues Detroit Police over False Facial Recognition Match*, WASH. POST (Apr. 13, 2021, 4:18 PM), https://www.washingtonpost.com/technology/2021

Mass scraping of personal information creates dangers that go beyond mere loss of privacy; it also enables cybercrime. In April 2021, *Insider* reported that personal data, ranging from phone numbers to locations, of over 533 million Facebook users were scraped and leaked in hacking forums.[84] Facebook confirmed that "malicious actors obtained this data not through hacking [Facebook's] systems but by scraping it."[85] Alon Gal, the chief technology officer of the cybercrime intelligence firm Hudson Rock, noted that the leaked data "could prove valuable to cybercriminals who use people's personal information to impersonate them or scam them into handing over login credentials."[86] Other researchers posit that the data could be used to gain access to individuals' Facebook accounts, email accounts, and other social networking accounts because, once a hacker has a victim's email address, they might be able to log into their other accounts by pairing the email address with simple passwords.[87] Phone numbers, in particular, have "taken on new significance and potential value to attackers" as they are "ubiquitous identifiers, linking you to different parts of your digital life" and "play[ing] a role in sensitive authentication."[88] Shockingly, just days after its Facebook story, *Insider* reported that the personal data of over 500 million LinkedIn users were also scraped and published for sale online.[89]

Scraping also has the potential to influence elections by extracting personally identifiable information in order to target individual voters. Aggregate IQ, a Canadian digital advertising and software development company, infamously influenced the United Kingdom's 2016 EU referendum by scraping

/04/13/facial-recognition-false-arrest-lawsuit [perma.cc/2GW7-V2TS] (detailing false arrest due to use of facial recognition software by the Detroit Police Department).

84. Aaron Holmes, *533 Million Facebook Users' Phone Numbers and Personal Data Have Been Leaked Online*, INSIDER (Apr. 3, 2021, 10:41 AM), https://www.businessinsider.com/stolen-data-of-533-million-facebook-users-leaked-online-2021-4 [perma.cc/5W5Y-6KY6].

85. Mike Clark, *The Facts on News Reports About Facebook Data*, FACEBOOK (Apr. 6, 2021), https://about.fb.com/news/2021/04/facts-on-news-reports-about-facebook-data [perma.cc/HLM2-V92M].

86. Holmes, *supra* note 84.

87. *E.g.*, Mostafa Rachwani, *Facebook Data Leak: Australians Urged to Check and Secure Social Media Accounts*, GUARDIAN (Apr. 5, 2021, 4:18 AM), https://www.theguardian.com/technology/2021/apr/05/facebook-data-leak-2021-breach-check-australia-users [perma.cc/4BRN-QUZX].

88. Lily Hay Newman, *What Really Caused Facebook's 500M-User Data Leak?*, WIRED (Apr. 6, 2021, 7:57 PM), https://www.wired.com/story/facebook-data-leak-500-million-users-phone-numbers [perma.cc/NX2C-V2G4].

89. Katie Canales, *Hackers Scraped Data from 500 Million LinkedIn Users—About Two-Thirds of the Platform's Userbase—and Have Posted It for Sale Online*, INSIDER (Apr. 8, 2021, 12:34 PM), https://www.businessinsider.com/linkedin-data-scraped-500-million-users-for-sale-online-2021-4 [perma.cc/Y35R-5G5Q]. In October 2021, several news outlets reported that the scraped personal information of another 1.5 billion Facebook users was allegedly being sold on a hacking forum, but as of this writing, the claim is unverified. Ryan Mac, *No, There Isn't Proof That the Private Data of 1.5 Billion Facebook Users Is Being Sold by Hackers.*, N.Y. TIMES (Oct. 5, 2021, 11:11 AM), https://www.nytimes.com/2021/10/05/technology/fb-hackers-data-sale.html [perma.cc/9FPS-QMNX].

individuals' profile information on LinkedIn and Facebook and serving them targeted ads supporting the "Vote Leave" campaign.[90] In 2016, Donald Trump's campaign hired the political data firm Cambridge Analytica, which scraped the private information of more than fifty million Facebook users.[91] The firm used these data to "identify the personalities of American voters and influence their behavior"[92] and "orchestrate[] emotionally charged political campaigns that advanced demeaning, racialized, nationalistic propaganda."[93]

Finally, data scraping can place people in physical danger by easing access to individuals' whereabouts. The story of Judge Esther Salas of the District of New Jersey illustrates the perils of publicizing personal information. In July 2020, an angered attorney sought revenge against Judge Salas for her handling of a lawsuit he filed in her court.[94] On a Sunday afternoon, the attorney showed up to Judge Salas's home and rang her doorbell.[95] Her only son, a college student named Daniel, opened the door. The attorney fired multiple gunshots, shooting and killing Daniel. He then shot Judge Salas's husband three times, seriously wounding him.[96]

Easy access to Judge Salas's personal information—including her home address—enabled the gunman to hunt down her family. In a *New York Times* op-ed, Judge Salas wrote that FBI agents informed her of how easy it is to find and purchase personal information about judges on the internet, including photos of their homes and the license plates on their vehicles.[97] In Judge Salas's case, the gunman "was able to create a complete dossier of her life: he stalked her neighborhood, mapped her routes to work, and even learned the names of her best friend and the church she attended."[98] This access to Judge Salas's personal information was completely legal, and it enabled the shooter to kill

---

90. *See* DIGITAL, CULTURE, MEDIA AND SPORT COMMITTEE, DISINFORMATION AND 'FAKE NEWS': FINAL REPORT, 2017–19, HC 1791, at 45, 48 (UK), https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf [perma.cc/X4XB-Q2RX].

91. Sarah Perez & Zack Whittaker, *Facebook Sues Two Companies Engaged in Data Scraping Operations*, TECHCRUNCH (Oct. 1, 2020, 4:54 PM), https://techcrunch.com/2020/10/01/facebook-sues-two-companies-engaged-in-data-scraping-operations [perma.cc/6YE8-QTTV] ("Cambridge Analytica infamously scraped millions of Facebook profiles in the run-up to the 2016 presidential election in order to target undecided voters."); Kevin Granville, *Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens*, N.Y. TIMES (Mar. 19, 2018), https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html [perma.cc/TQ6H-PYWJ].

92. Granville, *supra* note 91.

93. Tsesis, *supra* note 7, at 607.

94. Esther Salas, Opinion, *My Son Was Killed Because I'm a Federal Judge*, N.Y. TIMES (Dec. 8, 2020), https://www.nytimes.com/2020/12/08/opinion/esther-salas-murder-federal-judges.html [perma.cc/N2TW-BYN9]; *see also* Nicole Hong, William K. Rashbaum & Mihir Zaveri, *'Anti-feminist' Lawyer Is Suspect in Killing of Son of Federal Judge in N.J.*, N.Y. TIMES (July 22, 2020), https://www.nytimes.com/2020/07/20/nyregion/esther-salas.html [perma.cc/HV87-7NY5].

95. *See* Salas, *supra* note 94.

96. *Id.*

97. *Id.*

98. *Id.* (cleaned up).

her only child.[99] Although it is not clear whether data scraping may have contributed to this specific incident, there is no question that data scraping could facilitate harm through collecting and publicizing personal information of the kind that allowed the gunman to arrive at Judge Salas's door.

Using bots to scrape data in bulk from various publicly available sources makes it easier to collect and compile an abundance of personal information for potentially malicious purposes. Scraping enables its practitioners to more easily create a "complete dossier" of an individual's life. It can increase false arrests and influence elections. It's a useful tool for scammers, stalkers, and scoundrels. And what's worse, as the next Section explains, is that no existing or proposed legislation restricts scraping publicly available personal information.

B. *Scraping Personal Information Circumvents Current and Proposed Privacy Laws*

Existing and proposed consumer privacy laws fail to adequately protect individuals' personal information from data scrapers. Indeed, there is currently no comprehensive data privacy legislation enacted at the federal level.[100] However, responding to rising enthusiasm for consumer data privacy protection, several states have enacted or introduced legislation to protect the privacy of their residents' personal information, including California,[101] New

99. *Id.* Since this incident, Judge Salas has called on Congress to pass the Daniel Anderl Judicial Security and Privacy Act, named after her son. *Id.* The Act "would protect judges' personally identifiable information from resale by data brokers" and "allow federal judges to redact personal information displayed on federal government internet sites and prevent publication of [their] personal information . . . where there is no legitimate news media interest or matter of public concern." *Id.* The Act was introduced to the Senate in July 2021 but had not passed as of February 2022. Mark Brnovich & Gurbir S. Grewal, Opinion, *Congress Must Pass Daniel's Law to Protect Federal Judges*, ROLL CALL (July 16, 2021, 6:00 AM), https://rollcall.com/2021/07/16/congress-must-pass-daniels-law-to-protect-federal-judges [perma.cc/6UHL-Y3BX]; *see* Daniel Anderl Judiciary Security and Privacy Act of 2021, S. 2340, 117th Cong. (2021). A version of this legislation was enacted, however, in the state of New Jersey. *See Governor Murphy Signs "Daniel's Law*," STATE OF N.J. (Nov. 20, 2020), https://nj.gov/governor/news/news/562020/approved/20201120b.shtml [perma.cc/S3PY-MH2B].

100. Wendy Zhang, *Comprehensive Federal Privacy Law Still Pending*, NAT'L L. REV. (Jan. 22, 2020), https://www.natlawreview.com/article/comprehensive-federal-privacy-law-still-pending [perma.cc/N9EH-27MT]; *see also* Yallen, *supra* note 12, at 796–99.

101. California Consumer Privacy Act, ch. 55, 2018 Cal. Stat. 1807 (codified as amended at CAL. CIV. CODE §§ 1798.100–.198 (West Supp. 2021)); *see* Daisuke Wakabayashi, *California Passes Sweeping Law to Protect Online Privacy*, N.Y. TIMES (June 28, 2018), https://www.nytimes.com/2018/06/28/technology/california-online-privacy-law.html [perma.cc/HA9P-UWAF].

York,[102] Virginia,[103] Nevada,[104] Florida,[105] Colorado,[106] New Hampshire,[107] Washington,[108] and Illinois.[109] But even the strictest regulations contain gaping regulatory holes allowing scrapers to run wild with individuals' data.

California has enacted the most comprehensive data privacy laws to date in the United States.[110] The California Consumer Privacy Act (CCPA) went into effect in 2020 and provides Californians with certain rights regarding businesses' collection and sale of their personal information.[111] The California Privacy Rights Act (CPRA) was then enacted in November 2020 and will take effect in January 2023.[112] It expands upon the CCPA, creating a new agency called the California Privacy Protection Agency dedicated to enforcing the new

---

102. New York Privacy Act, S. 6701, 2021–2022 Reg. Sess. (N.Y. 2021); *see* ALEXANDER H. SOUTHWELL ET AL., GIBSON DUNN, NEW YORK PRIVACY ACT UPDATE: BILL OUT OF COMMITTEE, MOVES TO FULL SENATE (2021), https://www.gibsondunn.com/wp-content/uploads/2021/05/new-york-privacy-act-update-bill-out-of-committee-moves-to-full-senate.pdf [perma.cc/WUB3-J7YB].

103. Consumer Data Protection Act, ch. 35 (codified at VA. CODE. ANN. §§ 59.1-575 to -585 (Supp. 2021)); *see* Rebecca Klar, *Virginia Governor Signs Comprehensive Data Privacy Law*, HILL (Mar. 2, 2021, 5:24 PM), https://thehill.com/policy/technology/541290-virginia-governor-signs-comprehensive-data-privacy-law [perma.cc/7EPW-WK8Y].

104. Gretchen A. Ramos, Ed Chansky & Cathy C. Shyong, *Nevada Passes Opt-Out Privacy Law, Effective October 1, 2019*, NAT'L L. REV. (June 5, 2019), https://www.natlawreview.com/article/nevada-passes-opt-out-privacy-law-effective-october-1-2019 [perma.cc/4XLS-K5AB].

105. Kate Black, *Florida's Next: FL Consumer Privacy Bill Introduced*, NAT'L L. REV. (Jan. 24, 2020), https://www.natlawreview.com/article/florida-s-next-fl-consumer-privacy-bill-introduced [perma.cc/ZXU4-EPJP].

106. Act of July 7, 2021, ch. 483, § 1, 2021 Colo. Sess. Laws 3445, 3445–65 (codified at COLO. REV. STAT. §§ 6-1-1301 to -1313 (2021)); *see* RYAN BERGSIEKER, SARAH ERICKSON, LISA ZIVKOVIC & ERIC HORNBECK, GIBSON DUNN, THE COLORADO PRIVACY ACT: ENACTMENT OF COMPREHENSIVE U.S. STATE CONSUMER PRIVACY LAWS CONTINUES (2021), https://www.gibsondunn.com/wp-content/uploads/2021/07/the-colorado-privacy-act-enactment-of-comprehensive-u-s-state-consumer-privacy-laws-continues.pdf [perma.cc/XUC2-Q5RK].

107. Gretchen A. Ramos & Darren Abernethy, *Additional U.S. States Advance the State Privacy Legislation Trend in 2020*, NAT'L L. REV. (Jan. 27, 2020), https://www.natlawreview.com/article/additional-us-states-advance-state-privacy-legislation-trend-2020 [perma.cc/646L-5FW6].

108. Jake Holland, *Washington State Inches Closer to Passing Consumer Privacy Law*, BLOOMBERG L. (Mar. 4, 2021, 11:00 AM), https://news.bloomberglaw.com/tech-and-telecom-law/washington-state-inches-closer-to-passing-consumer-privacy-law [perma.cc/8AMK-2MQP].

109. Ramos & Abernethy, *supra* note 107.

110. Andy Green, *Complete Guide to Privacy Laws in the US*, VARONIS (Apr. 2, 2021), https://www.varonis.com/blog/us-privacy-laws [perma.cc/X7CS-UCDG].

111. California Consumer Privacy Act of 2018, CAL. CIV. CODE §§ 1798.100–.120 (West Supp. 2021) (amended 2020).

112. Lara O'Reilly, *Prop 24—the California Privacy Rights and Enforcement Act—Passed by Voters. Here's What Publishers Need Know*, DIGIDAY (Nov. 5, 2020), https://digiday.com/media/prop-24-the-california-privacy-rights-and-enforcement-act-passed-by-voters-heres-what-publishers-need-know [perma.cc/DX93-GUSG].

privacy law.[113] But California's legislation, however, does not prevent companies from using bots to scrape personal information from publicly available websites. Scraped data falls outside of its scope and remains unregulated.

To illustrate, the CPRA gives California consumers the ability to opt out from companies sharing, selling, or even retaining their data.[114] But what if those third parties simply *scrape* their data instead? In that case, the information was neither *shared* nor *sold*. The scrapers just took it, leaving the subjects unable to opt out. The CPRA also allows consumers to request disclosure of the "categories of personal information that the business collected about the consumer" and the "categories of personal information that the business sold or shared about the consumer and the categories of third parties to whom the personal information was sold or shared."[115] But consumers cannot be made aware of this same information if some unknown party simply scrapes their data.

While section 1798.100 of the CPRA provides an expansive notice-at-collection provision, requiring certain businesses to inform their consumers about aspects of personal data collection,[116] publicly available information remains unprotected by that same statute's definition of "personal information." The definition of "personal information" in section 1798.140(v)(2) expressly excludes publicly available information.[117] Thus, the statute permits businesses to scrape personal information from publicly available websites without providing any notice.

Moreover, regulations issued by California's attorney general pursuant to the CPRA provide that "[a] business that does not collect personal information directly from the consumer does not need to provide a notice at collection to the consumer if it does not sell the consumer's personal

---

113. Austin Mooney & Amy C. Pimentel, *California Voters Approve the California Privacy Rights Act*, NAT'L L. REV. (Nov. 4, 2020), https://www.natlawreview.com/article/california-voters-approve-california-privacy-rights-act [perma.cc/55WZ-XQGS].

114. *See* CAL. CIV. CODE §§ 1798.105, 1798.120 (West Supp. 2021) (effective Jan. 1, 2023); David Alpert, Note, *Beyond Request-and-Respond: Why Data Access Will Be Insufficient to Tame Big Tech*, 120 COLUM. L. REV. 1215, 1217 (2020).

115. CAL. CIV. CODE § 1798.115(a)(1)–(2) (effective Jan. 1, 2023).

116. *Id.* § 1798.100(a)(1)–(2).

117. The section states as follows:

> "Personal information" does not include publicly available information or lawfully obtained, truthful information that is a matter of public concern. For purposes of this paragraph, "publicly available" means: information that is lawfully made available from federal, state, or local government records, or information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media, or by the consumer; or information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience. "Publicly available" does not mean biometric information collected by a business about a consumer without the consumer's knowledge.

Id. § 1798.140(v)(2).

information."[118] Thus, a business that collects a consumer's personal information by scraping it from an intermediate source only needs to provide notice to the consumer if it intends to sell it.[119] The language in these provisions reveals a gaping hole in personal data privacy regulations.

Other states have followed California's lead, but similarly fail to address privacy concerns for publicly available data. Virginia, for example, became the second state to enact a comprehensive data privacy statute in March 2021.[120] Virginia's law, the Consumer Data Protection Act, imposes data processing obligations for businesses processing consumers' personal information, and it gives consumers various privacy rights similar to those granted by California law.[121] The legislation contains no private right of action and exempts several entities and types of data.[122] Like the California legislation, it excludes publicly available information from its definition of "personal data,"[123] and it defines "publicly available information" broadly, encompassing information that "a business has a reasonable basis to believe is lawfully made available to the general public through widely distributed media, by the consumer, or by a person to whom the consumer has disclosed the information, unless the consumer has restricted the information to a specific audience."[124] The law also limits the obligations imposed on data processors where those obligations "adversely affect[] the rights or freedoms of any persons, such as exercising the right of free speech under the First Amendment to the United States Constitution."[125]

Similar defects are present in Colorado's recently enacted privacy law, the third comprehensive data privacy statute adopted in the United States.[126] Like its California and Virginia kin, the Colorado Privacy Act excludes publicly available information from the scope of its regulations, and its definition of "publicly available information" covers any "information that a controller has a reasonable basis to believe the consumer has lawfully made available to the

---

118. CAL. CODE REGS. tit. 11, § 999.305 (2021).

119. Nate Garhart, *Data Scraping Under the Revised CCPA Regulations*, FARELLA BRAUN + MARTEL: PRIV. BLOG (Mar. 18, 2020), https://www.farellaprivacy.com/2020/03/data-scraping-under-the-revised-ccpa-regulations [perma.cc/46NQ-CFCE].

120. Kurt R. Hunt & Matthew A. Diaz, *Virginia Becomes 2nd State to Adopt a Comprehensive Consumer Data Privacy Law*, NAT'L L. REV. (Mar. 8, 2021), https://www.natlawreview.com/article/virginia-becomes-2nd-state-to-adopt-comprehensive-consumer-data-privacy-law [perma.cc/7P3F-RHYB].

121. *See id.*; VA. CODE ANN. §§ 59.1-575 to -585 (Supp. 2021) (effective Jan. 1, 2023).

122. Natasha G. Kohne et al., *Virginia Consumer Data Protection Act: What Businesses Need to Know*, AKIN GUMP (Mar. 4, 2021), https://www.akingump.com/en/news-insights/virginia-consumer-data-protection-act-what-businesses-need-to-know.html [perma.cc/2BDN-Z9LY].

123. VA. CODE ANN. § 59.1-575.

124. *Id.*

125. *Id.* § 59.1-582(E).

126. Act of July 7, 2021, ch. 483, § 1, 2021 Colo. Sess. Laws 3445, 3445–65 (codified at COLO. REV. STAT. §§ 6-1-1301 to -1313 (2021)); *see* Cynthia J. Larose & Christopher J. Buontempo, *And Now There Are Three. . . . The Colorado Privacy Act*, NAT'L L. REV. (July 16, 2021), https://www.natlawreview.com/article/and-now-there-are-three-colorado-privacy-act [perma.cc/5DKD-A623].

general public."[127] In sum, businesses scraping publicly available personal data remain unregulated even by the most expansive state data privacy laws.

At the federal level, privacy legislation has been similarly inadequate. For example, the proposed Consumer Data Privacy and Security Act of 2020 exempts publicly available information from its scope.[128] It also contains a broad definition of publicly available information,[129] permitting data scrapers to extract whatever personal information is posted on publicly available websites so long as there is a reasonable basis to believe that the individual volunteered it.

Another proposal, the SAFE DATA Act, similarly excludes publicly available information from its protection.[130] It broadly defines "publicly available information" to include any information that the entity reasonably believes has been made widely available to the general public, including information from a public website.[131] The Online Privacy Act[132] and the Privacy Bill of Rights Act[133] are comparably deficient because they also exclude publicly available information.[134] Simply put: no currently enacted or proposed legislation in the United States satisfactorily shields individuals' publicly available personal information from the claws of data scrapers.

C. *An Alternative Framework: The European Union's General Data Protection Regulation*

In sharp contrast with the United States, the European Union considers data privacy a fundamental right.[135] Even the scope of what is considered "personal data" or "personally identifiable information" differs substantially. In the United States, these terms apply to specific categories of information, with

---

127. COLO. REV. STAT. § 6-1-1303(17)(b) (2021).

128. Consumer Data Privacy and Security Act of 2020, S. 3456, 116th Cong. § 2(9)(C)(iv) (2020).

129. *Id.* § 2(13) ("The term 'publicly available information' means any information that a covered entity or service provider has a reasonable basis to believe is lawfully made available to the general public from[] (i) a Federal, State, or local government record; (ii) widely distributed media; or (iii) a disclosure to the general public that is made voluntarily by the individual, or required to be made by a Federal, State, or local law.").

130. *See* S. 4626, 116th Cong. § 2(10)(C)(iv) (2020).

131. *Id.* § 2(10)(G) ("[T]he term 'publicly available information' means any information that a covered entity has a reasonable basis to believe . . . is widely available to the general public, including information from a telephone book or online directory; television, internet, or radio content or programming; or the news media or a website that is lawfully available to the general public on an unrestricted basis . . . ." (cleaned up)).

132. Online Privacy Act of 2019, H.R. 4978, 116th Cong. (2019).

133. Privacy Bill of Rights Act, S. 1214, 116th Cong. (2019).

134. *See* H.R. 4978, § 2(13)(B)(i) ("The term 'personal information' does not include[] publicly available information related to an individual . . . ."); S. 1214, § 2(10)(C)(i) (2019) ("The term 'personal information' does not include publicly available information."). Notably, each of these bills contains a much narrower definition of "publicly available information," which is a certainly preferable step in the right direction with respect to regulating data scraping.

135. Voss & Houser, *supra* note 12, at 296.

restrictions placed on the use of those categories of information applying only to certain industries.[136] Conversely, the EU's definition of personal data is deliberately broad and aimed at protecting an individual's right to privacy.[137] Comparing these legal frameworks reveals that the United States' laws are underdeveloped with respect to ensuring data privacy for its people. If the United States wishes to make progress in this field, it should follow Europe's lead.

In 2016, the EU passed the General Data Protection Regulation (GDPR).[138] It is described as "the toughest privacy and security law in the world," imposing obligations on any organization—regardless of location—that targets or collects data related to people in the EU.[139] Unlike the CPRA, the GDPR's definition of "personal data" contains no exception for publicly available information.[140] The regulation provides EU citizens with rights, including the right to be notified when their personal data are collected, the right to access any of their collected personal data, the right to rectify inaccurate personal data, and the right to erasure of their personal data.[141]

Most relevant to the issue of data scraping is article 14 of the GDPR. It obligates data controllers[142] to inform those whose personal data they intend to process when the information in question has *not been directly obtained* from them—for instance, when their personal data have been scraped off the public internet.[143] Pursuant to article 14, data scrapers—when they scrape

---

136. *Id.* at 313.

137. *Id.*

138. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 1 [hereinafter General Data Protection Regulation].

139. Ben Wolford, *What is GDPR, the EU's New Data Protection Law?*, GDPR.EU, https://gdpr.eu/what-is-gdpr [perma.cc/3FQ4-55GL].

140. General Data Protection Regulation, *supra* note 138, art. 4(1); *see also* Piotr Foitzik, *Publicly Available Data Under the GDPR: Main Considerations*, IAPP (May 28, 2019), https://iapp.org/news/a/publicly-available-data-under-gdpr-main-considerations [perma.cc/NC62-X6ZC] ("[T]he GDPR applies in full irrespective of if the data are or were publicly available or not.").

141. General Data Protection Regulation, *supra* note 138, arts. 12–23.

142. A "controller" under the GDPR is a "natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data." *Id.* art. 4(7).

143. The article states:

> Where personal data have not been obtained from the data subject, the controller shall provide the data subject with the following information: (a) the identity and the contact details of the controller and, where applicable, of the controller's representative; (b) the contact details of the data protection officer, where applicable; (c) the purposes of the processing for which the personal data are intended as well as the legal basis for the processing; (d) the categories of personal data concerned; (e) the recipients or categories of recipients of the personal data, if any; (f) where applicable, that the controller intends to transfer personal data to a recipient in a third country or international organisation . . . .

Id. art. 14(1); *see also* Fiona Campbell, *Data Scraping—What Are the Privacy Implications?*, PRIV. & DATA PROT., Oct./Nov. 2019, at 3.

publicly available personal information concerning persons in the EU—must provide extensive notice to every data subject[144] within one month of scraping their data.[145]

However, even when data scrapers are able to provide such notice to all of the data subjects, the scraping must still meet certain criteria for it to be lawful. Article 5 provides that personal data shall be collected for "specified, explicit and legitimate purposes."[146] Moreover, scrapers may only collect information necessary for the purposes for which the data are processed[147] and may not retain personal data longer than is necessary for those purposes.[148]

Finally, the GDPR requires a lawful basis for data collection. There are six lawful bases available under the GDPR: (1) consent, (2) contract with the data subject, (3) compliance with a legal obligation, (4) vital interest, (5) public interest, and (6) legitimate interest.[149] Of these, the only fitting lawful ground for scraping is legitimate interest.[150] For scraping to satisfy the legitimate-interest lawful-basis requirement, the data scraper's legitimate interest must outweigh the data subject's "interests or fundamental rights and freedoms."[151] Together, these requirements dramatically restrict lawful data-scraping activity.

In March 2019, Poland's Data Protection Agency (DPA), acting pursuant to the GDPR, issued its first fine involving data scraping.[152] The agency held that Bisnode—a digital marketing company that scraped six million people's personal data—failed to respect data subject rights set out in article 14 of the GDPR because it did not notify the data subjects.[153] Bisnode was slapped with a €220,000 fine and given three months to comply with article 14's information-notification requirements.[154] Bisnode attempted to meet its notification obligation through a website posting, but the Polish DPA "rejected the argument that placing a privacy notice on the data scraping business's website

144. "Data subject" is the term used by the GDPR to refer to an "identified or identifiable natural person," and "personal data" refers to "any information relating to" a data subject. General Data Protection Regulation, *supra* note 138, art. 4(1).

145. *Id.* art. 14(3).

146. *Id.* art. 5(1)(b).

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

148. *Id.* art. 5(1)(e).

149. *Id.* art. 6(1)(a)–(f).

150. Fiona Campbell, *Data Scraping—Considering the Privacy Issues*, FIELDFISHER (Aug. 27, 2019), https://www.fieldfisher.com/en/services/privacy-security-and-information/privacy-security-and-information-law-blog/data-scraping-considering-the-privacy-issues [perma.cc/ZR55-4BSH].

151. General Data Protection Regulation, *supra* note 138, art. 6(f); *see also* Campbell, *supra* note 150.

152. Natasha Lomas, *Covert Data-Scraping on Watch as EU DPA Lays Down 'Radical' GDPR Red-Line*, TECHCRUNCH (Mar. 30, 2019, 12:00 PM), https://techcrunch.com/2019/03/30/covert-data-scraping-on-watch-as-eu-dpa-lays-down-radical-gdpr-red-line [perma.cc/T5QG-3L4N].

153. *Id.*

154. *Id.*

was enough to notify individuals, particularly where individuals were not aware that their data had been scraped and was being processed."[155] Similarly, in April 2021, Spain's data protection authority ordered Equifax to delete personal data it collected and pay a fine of about $1.1 million for including in credit reports publicly available data it scraped from government sources about individuals' outstanding debts.[156]

As Section II.B analyzed, the current and proposed data privacy statutes in the United States contain loopholes that allow data scraping of personal information to go unregulated. However, the GDPR's application to U.S. companies does not fill those loopholes. Instead, the GDPR should serve as a model for U.S. legislation with respect to preventing and deterring scraping individuals' personal information.

Even though U.S. companies are not exempt from the GDPR's territorial scope,[157] domestic legislation is required to similarly protect U.S. persons' data from data scrapers. The GDPR's regulations apply to companies established in the EU and companies (including those in the United States) that process personal data of subjects who are in the EU.[158] Notably, the GDPR's application is not limited to the collection and processing of EU citizens and residents' data. For example, it includes U.S. persons located within EU borders when their data is processed.[159]

But the fact that a U.S. company complies with the GDPR does not necessarily mean that domestic U.S. persons' data is protected. First, companies often have different versions of their websites based on the various territories in which they do business, each version providing different data privacy rights, policies, and procedures.[160] Companies that comply with the GDPR

155. Campbell, *supra* note 150; *see also* Christopher Escobedo Hart, *Data Scraping, at Home and Abroad*, SEC. PRIV. & L. (Sept. 11, 2019), https://www.securityprivacyandthelaw.com/2019/09/data-craping-at-home-and-abroad [perma.cc/W6LH-3WQC].

156. Catherine Stupp, *Data Scraping in EU Regulators' Sights as Spain Orders Equifax to Delete Information*, WALL ST. J. (May 6, 2021, 5:30 AM), https://www.wsj.com/articles/data-scraping-in-eu-regulators-sights-as-spain-orders-equifax-to-delete-information-11620293400 [perma.cc/B8DD-MUYV].

157. Lucy Handley, *US Companies Are Not Exempt from Europe's New Data Privacy Rules—and Here's What They Need to Do About It*, CNBC (May 23, 2018, 11:09 AM), https://www.cnbc.com/2018/04/25/gdpr-data-privacy-rules-in-europe-and-how-they-apply-to-us-companies.html [perma.cc/VJR5-EF4V]; Yaki Faitelson, *Yes, the GDPR Will Affect Your U.S.-Based Business*, FORBES (Dec. 4, 2017, 8:30 AM), https://www.forbes.com/sites/forbestechcouncil/2017/12/04/yes-the-gdpr-will-affect-your-u-s-based-business [perma.cc/Z8UH-GZN6].

158. General Data Protection Regulation, *supra* note 138, art. 3.

159. *See* Faitelson, *supra* note 157.

160. For example, the athletic apparel brand Adidas's U.S. website differs sharply from its Irish website with respect to privacy rights provided to visitors. If a visitor clicks the "data settings" link in the footer of Adidas's Irish website, the website launches a pop-up allowing users to have their data sent to them or deleted pursuant to rights bestowed by the GDPR. ADIDAS.IE, https://www.adidas.ie [perma.cc/P7YZ-CWMX]. If a visitor clicks the same link on the U.S. website, the site prompts them to select their state; if they select California, they are provided with similar options pursuant to the rights conferred by the CCPA. ADIDAS.COM, https://www.adidas.com/us [perma.cc/UN4B-DNAX]. If the user selects any other state, users

may do so only on a territorial basis, and their scraping activity may similarly follow territorial bounds.

Second, even if U.S. companies chose to follow GDPR standards for all data subjects (including domestic U.S. persons), this would not confer upon U.S. persons the full breadth of the GDPR's rights and protections, requiring domestic legislation to fill the gaps. For example, article 82 of the GDPR provides the right to compensation for "[a]ny person who has suffered material or non-material damage as a result of an infringement" of the GDPR.[161] Article 77 permits such persons to "lodge a complaint with a supervisory authority" to enforce their rights.[162] Conversely, domestic U.S. persons, who the GDPR does not protect, would not have any such remedy or method to enforce their rights without U.S.-specific legislation. Thus, a U.S. company that scrapes personal data without providing notice at collection pursuant to article 14 would be subject to liability only where the personal data collected are that of persons in the EU, but the company would not be subject to liability for scraping data of persons in the United States. If no notice is provided (or if the collection violates other provisions of the GDPR), persons in the EU can lodge a complaint to enforce their rights; persons in the United States cannot.

For these reasons, the United States must enact domestic privacy legislation to ensure similar data protection for its people, and it should look to the GDPR as a model for such legislation. With respect to data scraping, a domestic statute aligned with the GDPR should contain a definition of personal information that doesn't exclude publicly available information[163] and a provision similar to article 14, which requires a business to give notice to data subjects whose information it scrapes from the internet.[164]

III. A PROPOSAL FOR CALIFORNIA: "FAIR COLLECTION"

While passing legislation at the federal level could be desirable, this Part asserts that California should reform its data privacy legislation to conform with the protections afforded by the GDPR. Doing so would deter impermissible scraping by providing a remedy to individuals whose personal information has been scraped without notice. Finally, this Part addresses potential counterarguments, including concerns regarding the First Amendment implications of attempting to regulate the collection of publicly available data.

---

are merely provided another link to read the Adidas privacy policy, with no options to have their data sent to them or deleted. *Id.*

161. General Data Protection Regulation, *supra* note 138, art. 82(1).

162. *Id.* art. 77.

163. *See id.* art. 4(1).

164. *See id.* art. 14(1).

A. *California Should Adopt GDPR-Style Regulations to Shield Publicly Available Personal Information from Data Scrapers*

In the United States, many have called for preemptive legislation at the federal level to fill the domestic consumer data privacy void.[165] Several reports indicate that both Democrats and Republicans want to "take on Big Tech" with laws and regulations addressing several issues, including data privacy.[166] To be clear, data privacy legislation at the federal level could be beneficial. But to date, Congress's federal privacy legislation has been limited to sector-specific laws.[167]

Gridlock in Washington might diminish any potential for an all-encompassing data privacy law,[168] especially one that addresses this Note's narrow topic of scraping publicly available personal information. Consumer privacy advocates have raised the concern that federal legislation would not embrace the comprehensiveness and strictness of the CPRA or GDPR.[169] The fear is that federal preemptive legislation would "wipe[] out more stringent state rules" like those in California.[170] And despite the benefits that might flow from uniformity throughout the nation, consumers could be left with "the lowest common denominator" of privacy legislation.[171] Instead, some argue that any federal standard should serve as a minimum level of compliance, allowing states to pass their own stronger laws.[172] Because this Note is narrowly focused

---

165. *See, e.g.*, Saquella, *supra* note 12, at 243–45 (calling for a preemptive federal law on data privacy because "various state laws will create inconsistent privacy rights" and "data protection and privacy breaches do not respect state boundaries"); Yallen, *supra* note 12, at 821–25; Kessler, *supra* note 12, at 121–27 ("[T]he United States ultimately should adopt a federal standard that offers consumers similar protections as the GDPR and the CCPA. This would eliminate the issue of complying with a patchwork system as well as potential Dormant Commerce Clause challenges of state laws.").

166. Cecilia Kang, *Democratic Congress Prepares to Take on Big Tech*, N.Y. TIMES (Jan. 26, 2021), https://www.nytimes.com/2021/01/26/technology/congress-antitrust-tech.html [perma.cc/PYZ5-U5TN]; *see also* Karen Schuler, *Federal Data Privacy Regulation Is on the Way—That's a Good Thing*, IAPP (Jan. 22, 2021), https://iapp.org/news/a/federal-data-privacy-regulation-is-on-the-way-thats-a-good-thing [perma.cc/UY83-WBGK].

167. Saquella, *supra* note 12, at 228–29. The Health Insurance Portability and Accountability Act (HIPAA), for example, provides data privacy and security for medical information, and the Fair and Accurate Credit Transactions Act protects certain data in the financial sector. *Id.*

168. Kessler, *supra* note 12, at 123.

169. *See id.* at 122–23 ("[S]everal technology companies have said they would embrace a federal privacy law . . . . One caveat is that most of these companies would oppose a law as strict as the GDPR, and privacy advocates argue that these companies may merely want to preempt laws like the CCPA and set a diluted standard that is far more lenient than California's.").

170. Allison Grande, *Federal Privacy Law Shouldn't Lower the Bar, Senators Told*, LAW360 (Oct. 10, 2018, 10:36 PM), https://www.law360.com/articles/1090519/federal-privacy-law-shouldn-t-lower-the-bar-senators-told [perma.cc/2SU9-8MVF]; *see also* Rebecca Klar & Chris Mills Rodrigo, *New State Privacy Initiatives Turn Up Heat on Congress*, HILL (Feb. 10, 2021, 6:00 AM), https://thehill.com/policy/technology/538122-new-state-privacy-initiatives-turn-up-heat-on-congress [perma.cc/2AZ3-4APF].

171. Grande, *supra* note 170.

172. Kessler, *supra* note 12, at 125.

on reforming the way privacy law treats publicly available personal information for purposes of data scraping, the most straightforward approach is to amend the California legislation. Doing so could serve as a model for future federal legislation.

To address data scraping of publicly available personal information, California should amend its privacy laws enacted through the CCPA and CPRA. First, it should remove section 1798.140(v)(2), which, as previously noted, excludes publicly available information from its definition of personal information.[173] Removing this provision would keep publicly available personal information within the scope of California's privacy protections, as it remains within the scope of the GDPR's protections.[174]

Second, California should not exempt businesses from notice-at-collection requirements when they do not collect personal information directly from the consumer.[175] Thus, as in article 14 of the GDPR, businesses would also have to provide notice when collecting personal information indirectly or from a source other than the data subject.[176]

Third, California should expand the private right of action provided by section 1798.150. Currently, that provision only permits consumers to bring civil actions if their information is subject to a data breach.[177] It should be expanded to allow consumers to bring civil actions when businesses fail to notify individuals that their personal data have been collected.

In place of the removed provisions, California's legislature ought to adopt a more nuanced approach that would prohibit most forms of data scraping while permitting innocuous collections of personal information. This would be similar to permitting scraping where there is a lawful basis under the GDPR.[178] Here, this Note envisions permitting data scraping when the information collected is more likely to be anonymized, is not collected in bulk, and is collected for journalistic or academic purposes. Just as the "fair use" doctrine in the Copyright Act allows certain permissible uses of a copyrighted work to avoid copyright infringement liability,[179] California's privacy regulations should exempt certain collections and uses of personal information that it deems permissible when the personal information is not collected directly from the subject. Let's call it "fair collection." This Note proposes the following language:

---

173. CAL. CIV. CODE § 1798.140(v)(2) (West Supp. 2021) (effective Jan. 1, 2023) ("'Personal information' does not include publicly available information or lawfully obtained, truthful information that is a matter of public concern.").

174. *See* General Data Protection Regulation, *supra* note 138, art. 4(1).

175. *See supra* note 118 and accompanying text.

176. *See* General Data Protection Regulation, *supra* note 138, art. 14.

177. *See* CAL. CIV. CODE § 1798.150.

178. *See* General Data Protection Regulation, *supra* note 138, art. 6(1)(a)–(f).

179. 17 U.S.C. § 107 ("[T]he fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright.").

> Notwithstanding § 1798.100, the "fair collection" of personal information—such as when the personal information is collected in small quantities for academic, educational, or journalistic purposes—is exempt from the notice-at-collection requirements when the personal information is not collected directly from the consumer. In determining whether the collection of personal information in any particular case constitutes "fair collection," the factors to be considered shall include:
>
> (a) the personal nature of the information collected, such as whether the information is anonymized or is capable of individually identifying a person;
>
> (b) the volume of the information collected; and
>
> (c) the purpose and character of the collection, including whether the collection is done for academic or legitimate news reporting purposes or to address matters of public concern, or instead is collected for commercial purposes.

Taking a "fair collection" approach would allow California to regulate data scrapers that collect massive amounts of personal data for commercial purposes while permitting small amounts of data collection when it is unlikely to be harmful.

The three proposed factors close the present gaps in the CCPA and CPRA. Factor (a) considers whether the information collected is capable of personally identifying an individual, which is the reason for regulating this activity in the first place. Some information, like health data or internet browsing history, is capable of anonymization and thus could be permissibly collected. But collecting an email address, IP address, phone number, or an image of someone's face should be regulated because such information is inextricably linked to a particular individual and cannot be anonymized unless outright redacted.

Factor (b) considers the volume of the data collected. Data scraping is a particular method of gathering information. What makes scraping different from an individual user manually gathering information from the internet is the ability for the data scraper to collect information automatically and in bulk.[180] If a business downloads a handful of publicly available email addresses, that could be exempted from regulation under this Note's proposal. But if a business uses bots to collect thousands of email addresses from some publicly available source, it would be regulated. The absence of a bright-line rule for what volume of collection is permissible is a feature, not a bug. If

180. A similar observation was made in the data-scraping case *Sandvig v. Sessions*, 315 F. Supp. 3d 1, 26–27 (D.D.C. 2018) ("Scraping or otherwise recording data from a site that is accessible to the public is merely a particular use of information that plaintiffs are entitled to see. . . . Employing a bot to crawl a website . . . does not constitute an *access* violation [under the Computer Fraud and Abuse Act] when the human who creates the bot is otherwise allowed to read and interact with that site. . . . [B]ots are simply technological tools for humans to more efficiently collect and process information that they could otherwise access manually." (citations omitted)).

scrapers aren't sure how much collection is too much, that uncertainty functions to deter scraping.[181] Conversely, if businesses knew that scraping under a certain volume of personal data would likely be permissible, they might confidently continue to do so.

Factor (c) considers the purpose and character of the collection. Personal data collected for commercial purposes would be subject to greater scrutiny than data collected for academic or journalistic purposes, or to address matters of public concern. Taken together, collecting publicly available personal information would be permissible when the information is less likely to identify a specific individual, the collection only concerns a small number of data subjects, and the collection furthers a beneficial public purpose. Collecting publicly available personal information would be impermissible when the information identifies individual subjects, is collected in large quantities, and is collected for commercial purposes. This proposed reform would finally address data scraping and protect individuals' personal information regardless of whether it is publicly available. Further, it would align the protections of the CPRA more closely to those of the GDPR.

While this Note cites many instances of scraping activity conducted by businesses, individual malicious actors also partake in data scraping. Recall the massive leaks of over 533 million Facebook users'[182] and 500 million LinkedIn users' personal information obtained by data scrapers.[183] As of this writing, there is no indication that any corporate data firm was responsible for this bulk scraping.[184] The reporting suggests that both leaks were the result of coordinated scraping efforts conducted by individuals, not businesses. Shouldn't California's privacy legislation regulate this activity as well?

Presently, the CPRA's regulations apply only to businesses that (a) have annual gross revenues in excess of $25 million; (b) buy, sell, or share the personal information of 100,000 or more consumers or households; or (c) derive 50 percent or more of their annual revenues from selling or sharing consumers' personal information.[185] Individuals who collect personal data are not regulated. While a proposal to enact civil or criminal sanctions on individuals is beyond the scope of this Note, California should also consider methods to address massive data collection at the hands of individual scrapers who might use the data to conduct scams and cybercrimes.

---

181. *See* Ehud Guttel & Alon Harel, *Uncertainty Revisited: Legal Prediction and Legal Postdiction*, 107 MICH. L. REV. 467, 496 (2008) ("[S]anction uncertainty can be harnessed to augment the deterrent effect of the criminal system.").

182. Holmes, *supra* note 84.

183. Canales, *supra* note 89.

184. *Business Insider* reported that the leaked data was discovered when "a user in [a] hacking forum advertised an automated bot that could provide phone numbers for hundreds of millions of Facebook users for a price." Holmes, *supra* note 84. Additionally, Facebook's blog post in response to the scraping and subsequent data leak refers to the scrapers as "fraudsters" and "malicious actors," not as corporations or business entities. Clark, *supra* note 85.

185. CAL. CIV. CODE § 1798.140(d)(1) (West Supp. 2021) (effective Jan. 1, 2023).

B. *Addressing First Amendment Concerns*

Critics of limiting scraping in the ways this Note proposes would likely argue that such restrictions violate the First Amendment.[186] In *Sorrel v. IMS Health Inc.*, the Supreme Court held that creating and disseminating information qualify as protected speech under the First Amendment.[187] While restrictions on scraping would not directly implicate the *publication* of personal information, they would limit accessing and recording publicly available facts—activities that contribute to the creation of speech. Restricting the ability to access and record facts disables one from later speaking and disseminating information about those facts. The *Sorrell* Court noted that "[f]acts, after all, are the beginning point for much of the speech that is most essential to advance human knowledge and to conduct human affairs."[188] It follows that laws burdening the underlying inputs of speech implicate the First Amendment.

If scraping qualifies as speech, it would likely be considered conduct "incidental to, or in preparation for, speech" under the First Amendment.[189] For instance, some argue that video recording is a form of expression covered by the First Amendment because it is conduct essential to speech.[190] In *ACLU of Illinois v. Alvarez*, the Seventh Circuit recognized a right to record in enjoining the enforcement of an Illinois all-party-consent wiretap statute.[191] There, the court held that "[c]riminalizing all nonconsensual audio recording necessarily limits the information that might later be published or broadcast . . . and thus burdens First Amendment rights."[192] The right to create the recording, the court reasoned, is "necessarily included within the First Amendment's guarantee of speech and press rights as a corollary of the right to disseminate the resulting recording."[193]

This reasoning may also extend to data scraping. As in *Alvarez*, limiting scraping "necessarily limits the information that might later be published or broadcast," and thus burdens First Amendment rights.[194] In *Sandvig v. Sessions*, a case involving data scraping of a publicly available website, the court

186. *See, e.g.*, Jameel Jaffer & Ramya Krishnan, *Clearview AI's First Amendment Theory Threatens Privacy—and Free Speech, Too*, SLATE (Nov. 17, 2020, 1:21 PM), https://slate.com/technology/2020/11/clearview-ai-first-amendment-illinois-lawsuit.html [perma.cc/7K94-7TSV] (discussing Clearview AI's argument that its scraping practices are protected by the First Amendment because it merely collects publicly available information).

187. 564 U.S. 552, 570 (2011).

188. *Sorrell*, 564 U.S. at 570.

189. Carrero, *supra* note 11, at 152.

190. Justin Marceau & Alan K. Chen, *Free Speech and Democracy in the Video Age*, 116 COLUM. L. REV. 991, 1017 (2016).

191. 679 F.3d 583, 586–87, 595–97 (7th Cir. 2012).

192. *Alvarez*, 679 F.3d at 597.

193. *Id.* at 595.

194. *Id.* at 597. For discussion of the First Amendment, the right to record, and data scraping, see Komal S. Patel, Note, *Testing the Limits of the First Amendment: How Online Civil Rights*

observed that "even if a law says nothing about speech on its face, it is subject to First Amendment scrutiny if it restricts access to traditional public fora."[195] There, because the statute at issue "limit[ed] access to and burden[ed] speech in the public forum that is the public Internet," heightened First Amendment scrutiny was appropriate.[196]

The question, then, is whether this Note's proposed limitations on data scraping would survive First Amendment scrutiny. To reiterate, this Note's reform would limit scraping of personal information in bulk for predominantly commercial purposes. Where commercial speech is involved, courts apply intermediate scrutiny: the state's restriction on commercial speech must directly advance a substantial governmental interest and must be drawn to achieve that interest.[197]

First, in limiting how corporations collect and monetize consumers' personal information, governments like California's have a "substantial interest" in promoting consumer data privacy.[198] Indeed, the California Constitution expressly makes privacy an "inalienable" right of all people.[199] And as the Supreme Court has recognized, the fact that "an event is not wholly 'private' does not mean that an individual has no interest in limiting disclosure or dissemination of the information."[200] The existence of a modern technological tool like scraping "only heightens the consequences of disclosure—'in today's society the computer can accumulate and store information that would otherwise have surely been forgotten.' "[201] Here, scraping poses a substantial threat to individuals' privacy, especially in cases where their personal information has been made public without their knowledge or consent.[202] It allows data that individuals may intend to restrict to instead be continuously collected and shared outside their control.

Testing is Protected Speech Activity, 118 COLUM. L. REV. 1473, 1485–91 (2018), and Jane Bambauer, *Is Data Speech?*, 66 STAN. L. REV. 57 (2014).

195. 315 F. Supp. 3d 1, 29 (D.D.C. 2018) (cleaned up) (quoting McCullen v. Coakley, 573 U.S. 464, 476 (2014)).

196. *Sandvig*, 315 F. Supp. 3d at 29.

197. Sorrell v. IMS Health Inc., 564 U.S. 552, 571–72 (2011); *see also* Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc., 472 U.S. 749, 762 & n.8 (1985); Cent. Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n, 447 U.S. 557, 561, 563 (1980).

198. *See* Trans Union Corp. v. FTC, 245 F.3d 809, 813 (D.C. Cir. 2001) ("Applying intermediate scrutiny, the Commission found that the government has a substantial interest in protecting private credit information . . . ."); Nat'l Cable & Telecomm. Ass'n v. FCC, 555 F.3d 996, 1001 (D.C. Cir. 2009) (" '[P]rotecting the privacy of consumer credit information' is a 'substantial' governmental interest . . . ." (quoting *Trans Union*, 245 F.3d at 818)); *see also* King v. Gen. Info. Servs., Inc., 903 F. Supp. 2d 303, 310 (E.D. Pa. 2012).

199. CAL. CONST. art. 1, § 1.

200. U.S. Dep't of Just. v. Reps. Comm. for Freedom of Press, 489 U.S. 749, 770 (1989).

201. *See* Detroit Free Press Inc. v. U.S. Dep't of Just., 829 F.3d 478, 482 (6th Cir. 2016) (en banc) (quoting *Reps. Comm.*, 489 U.S. at 771).

202. *See supra* Section II.A.

Second, California also has a substantial interest in protecting its residents' First Amendment interests—namely, free expression that relies on privacy. In her concurrence in *United States v. Jones*, Justice Sotomayor noted that even where personal information is publicly available, its collection and compilation can reveal a "comprehensive record" of a person's activity that reflects "a wealth of detail about her familial, political, professional, religious, and sexual associations."[203] Data scraping could be viewed "as such an egregious invasion of privacy that users' First Amendment activity on online platforms would be chilled."[204] Fear that all of an individual's personal information is susceptible to scraping and misappropriation could curb the use of certain internet platforms. Fear or suspicion that one's speech is constantly monitored "can have a seriously inhibiting effect upon the willingness to voice critical and constructive ideas."[205] Finally, California has a substantial interest in protecting the integrity of its elections. As exposed by the malfeasance of Aggregate IQ and Cambridge Analytica, data scraping has the potential to undermine elections by scraping individuals' social media profile information and serving them targeted ads meant to influence their vote.[206]

The statutory reform proposed in this Note—limiting the bulk collection of publicly available personal information for commercial purposes—is narrowly drawn to meet California's interests and thus should pass First Amendment scrutiny. The Constitution affords lesser protection to commercial speech than to other constitutionally guaranteed expression.[207] This reform does not prohibit all access to publicly available information; it merely restricts its collection in bulk and for commercial purposes when it involves personal information that cannot be anonymized.

Indeed, there are provisions of California's current privacy laws that arguably infringe on First Amendment rights far more than this Note's proposal.[208] In the context of government collection of personal information for criminal investigation purposes, the Supreme Court has held that "a person has no legitimate expectation of privacy in information he voluntarily turns over to third parties."[209] But, as this Note explains, information is often publicized without any voluntary action or consent from the data subject. And

---

203. 565 U.S. 400, 415 (2012) (Sotomayor, J., concurring).

204. Carrero, *supra* note 11, at 158.

205. Bartnicki v. Vopper, 532 U.S. 514, 533 (2001).

206. See *supra* Section II.A for a discussion of Aggregate IQ and Cambridge Analytica.

207. KATHLEEN ANN RUANE, CONG. RSCH. SERV., 95-815, FREEDOM OF SPEECH AND PRESS: EXCEPTIONS TO THE FIRST AMENDMENT 14 (2014).

208. For instance, the CPRA gives consumers the right to request that a business delete any personal information it has collected from and about the consumer or to correct inaccurate information about the consumer. CAL. CIV. CODE §§ 1798.105(a), .106(a) (West Supp. 2021) (effective Jan. 1, 2023). Compelling speech in this manner—deleting information and correcting inaccurate information—arguably infringes upon the First Amendment more than restricting bulk, commercial scraping activity in the manner I've proposed would.

209. Smith v. Maryland, 442 U.S. 735, 743–44 (1979); *see also* United States v. Miller, 425 U.S. 435, 443 (1976).

this Note's proposed reform does not restrict collecting single individuals' information, but rather data in bulk. Most importantly, it also exempts from its restrictions data collected for journalistic purposes or to address matters of public concern. Criminal investigations would surely qualify for this exemption. The statutory language suggested in this Note likely comports with the Supreme Court's view of privacy and does not regulate beyond what is necessary to meet California's interests. Accordingly, it should survive any challenges sounding in the First Amendment.

CONCLUSION

Data scraping can be greatly beneficial, but it presents serious concerns when the data contains individuals' personal information. As the author of this Note, I am grateful for data-scraping technology because it made this Note possible. After all, the research tools I used aggregate publicly available information in the form of statutes, cases, and law review articles. But when the information collected is not a judicial opinion but an individual's personal data, more is at stake. Scraping of such data in bulk can harm individual privacy, undermine democracy, and potentially even physically endanger us. Today's privacy statutes do not do enough to address this issue, allowing businesses to scrape and repurpose our personal information with near impunity. California should adopt a new approach that restricts the collection of even publicly available personal information, only allowing such collection when it deems it fair and permissible. Other states—and perhaps the federal government—should soon follow suit.