Harvard Journal of Law & Technology
Volume 34, Number 2 Spring 2021

BAD BOTS: REGULATING THE SCRAPING OF PUBLIC PERSONAL INFORMATION

Geoffrey Xiao[*]

TABLE OF CONTENTS

[*] Columbia Law School, Candidate for J.D., 2021. Many thanks to Professors Daniel Richman and Matthew Waxman, who inspired the topic and provided valuable feedback. I am also very grateful to Qifan Huang for helpful discussions and to Noah Goldman and the rest of the JOLT staff for their diligent work.

I. INTRODUCTION

In early 2020, the New York Times broke the story of Clearview AI, which had created a facial recognition program by surreptitiously scraping more than three billion images from publicly available websites such as Facebook and Twitter.[1] The New York Times story provoked outcry from privacy advocates, the websites that were scraped, and the public.[2] However, Clearview's refrain has been that the images it scraped were made publicly available by users and that these users have no privacy interests in otherwise public information.[3]

This Note analyzes how U.S. law addresses the privacy implications of data scraping. This analysis looks at public personal information, which is information openly accessible on the web that can identify the poster. The prototypical example is the one raised by Clearview's scraping: users publicly post personal information (e.g., photos) on social media websites like LinkedIn and Facebook, and a third party (e.g., Clearview) scrapes this public personal information.

The principal problem raised by data scraping is whether users have privacy interests in personal information they have posted publicly. While U.S. law has generally been reluctant to find such privacy interests, this Note argues that there are strong privacy interests because harms arise from the unauthorized use of public personal information and because users post information in an environment of obscurity and trust.[4] Next, this Note surveys how existing legal regimes protect public personal information. This Note argues that current regulations fall short in several significant ways. First, some data privacy laws — notably, California's comprehensive data privacy statute, the California Consumer Privacy Act ("CCPA") — exempt scrapers from providing notice to users whose data have been scraped. These laws are based on the presumption that the indirect scraper/user relationship makes providing notice difficult. Second, the legal regime needs to require opt-in consent instead of opt-out consent because opt-out consent fails

[1] Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Jan. 31, 2021, 1:38 PM), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html [https://perma.cc/7Z64-9XQQ].

[2] Nick Statt, *ACLU Sues Facial Recognition Firm Clearview AI, Calling It a 'Nightmare Scenario' for Privacy*, VERGE (May 28, 2020, 1:13 PM), https://www.theverge.com/2020/5/28/21273388/aclu-clearview-ai-lawsuit-facial-recognition-database-illinois-biometric-laws [https://perma.cc/N6SM-NNMF].

[3] Defendant's Memorandum of Law in Support of its Motion to Dismiss at 18, ACLU v. Clearview AI, Inc., No. 2020 CH 04353 (Ill. Cir. Ct. Oct. 7, 2020).

[4] *See* discussions *infra* Part II.

to adequately protect privacy. Third, regulations need to provide an active role for websites in protecting their users' privacy, but regulations also need to be wary of granting websites monopolistic control over scraping.

Lastly, this Note addresses the First Amendment defense that scrapers like Clearview have made. It is arguable whether regulations that limit the scraping of public information are a constraint on speech (and subject to strict scrutiny) or merely a restriction on expressive conduct (and subject to intermediate scrutiny). In any event, a notice-and-consent requirement withstands First Amendment challenges because it does not unduly limit scraping.

II. THE PROBLEM OF PRIVACY IN PUBLIC

The central problem raised by scraping is whether users have a legitimate privacy interest in information they have made public. Clearview, for instance, has argued "that individuals have no right to privacy in materials they post [publicly] on the Internet."[5] While U.S. law generally follows the rule of "no privacy in public," there are actually very strong privacy interests in public personal information.

A. U.S. Privacy Law and the Rule of "No Privacy in Public"

U.S. privacy law is described as "sectoral," meaning privacy regulation is field-specific as opposed to "omnibus."[6] For example, privacy torts allow individuals to vindicate invasions of their privacy, and the Fourth Amendment protects individuals from government intrusion.[7] In these different sectors, U.S. law has generally adopted the "no privacy in public" principle.[8]

In the public disclosure of private fact tort, "there is no liability when the defendant merely gives further publicity to information about

[5] Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 3, at 18.

[6] W. Gregory Voss, *Obstacles to Transatlantic Harmonization of Data Privacy Law in Context*, 2019 U. ILL. J.L. TECH. & POL'Y 405, 417–27 (2019). E.U. data privacy law is "omnibus," meaning data privacy transcends specific sectors. For example, data privacy is a fundamental right guaranteed by the Charter of Fundamental Rights of the European Union. Charter of Fundamental Rights of the European Union art. 8, Dec. 18, 2000, 2000 O.J. (C 364) 1.

[7] *See* Samuel Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 219 (1890) (describing privacy torts).

[8] *See* Woodrow Hartzog, *The Public Information Fallacy*, 99 BOS. L. REV. 459, 459 (2019) ("The concept of privacy in 'public' information or acts is a perennial topic for debate . . . . People struggle to reconcile with traditional accounts of privacy the notion of protecting information that has been made public. As a result, successfully labeling information as public often functions as a permission slip for surveillance and personal data practices."); Woodrow Hartzog & Evan Selinger, *Surveillance as Loss of Obscurity*, 72 WASH. & LEE L. REV. 1343, 1349 (2015) ("Courts and policy-makers regularly affirm that there is no 'privacy in public.'").

the plaintiff which is already public or when the further publicity relates to matters which the plaintiff leaves open to the public eye."[9] For example, in *Daly v. Viacom, Inc.*, the defendant photographed the plaintiff kissing a man in a bathroom stall.[10] Even though the defendant took this photograph in a private bathroom stall, the court rejected the plaintiff's claim for public disclosure of private fact.[11] According to the court, simply because the plaintiff had kissed the man in public before, her kiss had been publicly disclosed, so the disclosure was not actionable under tort law.[12]

Similarly, under the intrusion upon seclusion tort, claims resting on "'public places' or things that are in 'plain view'" are not actionable.[13] In one case, a news crew filming a car accident was not liable under the intrusion tort because the accident occurred on a public highway.[14] The court distinguished between filming on a public highway (not actionable) and filming inside a medivac helicopter (actionable because "plaintiffs had an objectively reasonable expectation of privacy in the interior of the rescue helicopter, which served as an ambulance").[15]

While not directly applicable to Clearview, the Fourth Amendment's treatment of the privacy in public problem provides a helpful analogue.[16] Just like tort law, the Fourth Amendment gives minimal protection for publicly available information. Under what Professor Monu Bedi calls the "public disclosure doctrine," courts have found that individuals lack reasonable expectations of privacy in publicly available information.[17] The seminal *Katz v. United States* case announced that "[w]hat a person knowingly exposes to the public, even in his own home or office, is not a subject of Fourth Amendment protection."[18] In *California v. Ciraolo*, the Court found the Fourth Amendment inapplicable when the government performed aerial surveillance on a backyard because "[a]ny member of the public flying in this airspace who glanced down could have seen everything that these officers observed."[19] The Court also refused to extend Fourth Amendment protections to a police search of sidewalk garbage because "[i]t is common

[9] Sipple v. Chronicle Publ'g Co., 201 Cal. Rptr. 665, 669 (Cal. Ct. App. 1984). *See also* Hartzog, *supra* note 8, at 504.

[10] Daly v. Viacom, Inc., 238 F. Supp. 2d 1118, 1123–25 (N.D. Cal. 2002).

[11] *Id.* at 1125.

[12] *Id.*

[13] Hartzog, *supra* note 8, at 500.

[14] Shulman v. Group W Prods., Inc., 955 P.2d 469, 490 (Cal. 1998).

[15] *Id.*

[16] However, there have been reports that the government has used Clearview. *See* Taylor Hatmaker, *Clearview AI Landed a New Facial Recognition Contract with ICE*, TECHCRUNCH (Aug. 14, 2020, 3:34 PM), https://techcrunch.com/2020/08/14/clearview-ai-ice-hsi-contract-2020/ [https://perma.cc/H2SP-USST].

[17] Monu Bedi, *The Fourth Amendment Disclosure Doctrines*, 26 WM. & MARY BILL RTS. J. 461, 480–82 (2017).

[18] Katz v. United States, 389 U.S. 347, 351 (1967).

[19] California v. Ciraolo, 476 U.S. 207, 213–14 (1986).

knowledge that plastic garbage bags left on or at the side of a public street are readily accessible to animals, children, scavengers, snoops, and other members of the public."[20] In rejecting a Fourth Amendment argument against a subpoena seeking public (but since deleted) tweets, one court aptly summarized: "[i]f you post a tweet, just like if you scream it out the window, there is no reasonable expectation of privacy."[21]

Still, there has been some pushback against the rule of no privacy in public. In *Commonwealth v. McCarthy*, the Massachusetts Supreme Judicial Court acknowledged that license plates are "knowingly exposed" to the public and that police are free to examine the license plates of cars driving on the street.[22] However, the court also said that a pervasive (in terms of location and time) automatic license plate reader ("ALPR") system would raise Fourth Amendment issues "because the whole of one's movements, even if they are all individually public, are not knowingly exposed in the aggregate."[23] Nevertheless, *McCarthy* found an ALPR system of "four cameras at fixed locations on the ends of two bridges" insufficiently pervasive to violate the Fourth Amendment.[24]

Like the Fourth Amendment, the First Amendment has given short shrift to privacy interests in public information. In *Cox Broadcasting Corp. v. Cohn*, a rape victim sued a television station for broadcasting her name.[25] The Court found that the First Amendment defeated the plaintiff's claim because the station had obtained the plaintiff's name from public court records. The Court explained that "the interests in privacy fade when the information involved already appears on the public record."[26] *Florida Star v. B.J.F.* reached a similar conclusion when a newspaper published a rape victim's name after obtaining it from a publicly available police report.[27]

Recently, the Ninth Circuit applied the "no privacy in public" rule to data scraping. In *hiQ Labs, Inc. v. LinkedIn Corp.*, LinkedIn attempted to use the Computer Fraud and Abuse Act ("CFAA") to protect

[20] California v. Greenwood, 486 U.S. 35, 40 (1988).

[21] People v. Harris, 949 N.Y.S.2d 590, 595 (N.Y. Crim. Ct. 2012); *accord id.* at 597–98 ("The Constitution gives you the right to post, but as numerous people have learned, there are still consequences for your public posts. What you give to the public belongs to the public.").

[22] Commonwealth v. McCarthy, 142 N.E.3d 1090, 1101 (Mass. 2020).

[23] *Id.* at 1103; *accord id.* at 1105.

[24] *Id.* at 1106.

[25] Cox Broadcasting Corp. v. Cohn, 420 U.S. 469 (1975).

[26] *Id.* at 494–95.

[27] Florida Star v. B.J.F., 491 U.S. 524 (1989). In *Florida Star*, the plaintiff rape victim asserted a claim under a Florida law banning publication of the name of a sexual offense victim. The Court struck down this law on First Amendment grounds. *Florida Star* acknowledged that privacy interests may sufficiently override First Amendment interests, but, in the instant case, found the means employed were not narrowly tailored to furthering the privacy interests. *Id.* at 537.

its users' privacy interests from a scraper.[28] hiQ scraped public profiles from LinkedIn's website to produce a data analytic called "Keeper," which "identif[ies] employees at the greatest risk of being recruited away."[29] LinkedIn argued that hiQ's data scraping and "Keeper" analytic endangered LinkedIn users' privacy. Even though hiQ only scraped public profiles, LinkedIn argued that "many members — including members who choose to share their information publicly — do not want their employers to know they may be searching for a new job."[30] Ultimately, the Ninth Circuit found that LinkedIn users' had minimal privacy interests in public personal information, concluding that "there is little evidence that LinkedIn users who choose to make their profiles public actually maintain an expectation of privacy with respect to the information that they post publicly, and it is doubtful that they do."[31]

B. There are Privacy Interests in Public Personal Information

While courts have been reluctant to find privacy interests in public information, scholars have articulated several theories for recognizing privacy in public.

1. The Privacy Harms of Public Personal Information

Even when personal information is public, the collection, processing, and further dissemination of such information can create privacy harms. Put simply, the privacy harms associated with public personal information are as substantial as those associated with private personal information. These harms are independent of whether the information is initially public or private.

First, data collection creates dignitary and emotional harms. Professor Daniel Solove argues that information collection "create[s] feelings of anxiety and discomfort."[32] Knowing that your social media pictures are being harvested and analyzed is extremely discomforting.[33] These feelings of unease can have chilling effects on free speech.[34] For example, many Internet users adopt anonymous online identities. Yet, data scraping, by amassing large amounts of user data, can be used to

[28] hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 992 (9th Cir. 2019), *petition for cert. filed* (U.S. filed Mar. 9, 2020) (No. 19-1116).

[29] *Id.* at 991.

[30] *Id.* at 994.

[31] *Id.*

[32] Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 493 (2006).

[33] *See* Margot E. Kaminski & Shane Witnov, *The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech*, 49 U. RICH. L. REV. 465, 483–93 (2015).

[34] Solove, *supra* note 32, at 495.

de-anonymize these identities, thereby discouraging Internet users from speaking.[35] Even if the information collection were covert, if it were so "well-concealed . . . [to] eliminate the potential for any discomfort or chilling effect, it would still enable the watchers to gather a substantial degree of information about people," giving the observer a disproportionate amount of power over the observee.[36] This power imbalance can even lead to nefarious activities such as blackmail, identity theft, discrimination, and fraud.[37] Covert surveillance also infringes on an individual's autonomy — the very concealment of surveillance deceives the surveilled individual.[38]

Once collected, the processing of public personal information can create harms. The aggregation of collected information "can reveal new facts about a person that she did not expect would be known about her when the original, isolated data was collected."[39] For example, researchers analyzed public tweets to identify users with mental health issues.[40] Certainly, Twitter users do not expect mental health diagnoses if they created their accounts to retweet funny cat videos. Using tweets to diagnose mental health issues is an example of what Professor Solove calls "secondary use," which is "the use of data for purposes unrelated to the purposes for which the data was initially collected."[41] For example, Twitter users tweet to share ideas with the public, not to participate in a facial recognition program. Indeed, users come to Twitter with this expectation: since 2018, Twitter's terms of service have

[35] *Id.* at 510–16. *See* Kaminski & Witnov, *supra* note 33, at 496 ("Anonymity 'exemplifies the purpose behind the Bill of Rights, and of the First Amendment in particular: to protect unpopular individuals from retaliation — and their ideas from suppression — at the hand of an intolerant society.'") (quoting McIntyre v. Ohio Elections Comm'n, 514 U.S. 334, 357 (1995)).

[36] Solove, *supra* note 32, at 495. *See also* Tara Seals, *Millions of Social Profiles Leaked by Chinese Data-Scrapers*, THREATPOST (Jan. 11, 2021, 4:54 PM), https://threatpost.com/social-profiles-leaked-chinese-data-scrapers/162936/ [https://perma.cc/4LL7-X2EF] (describing how recent data leak of scraped public personal information can be used in social engineering identity theft and fraud attacks).

[37] Solove, *supra* note 32, at 495–96. *See also* Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1952–58 (2013) (describing harms of information collection, including blackmail, persuasion, and discrimination).

[38] Stanley I. Benn, *Privacy, Freedom, and Respect for Persons*, *in* NOMOS XIII: PRIVACY 1, 10–11 (J. Roland Pennock & John W. Chapman eds., 1971) ("Covert observation . . . deliberately deceives a person about his world, thwarting . . . his attempts to make a rational choice. One cannot be said to respect a man as engaged on an enterprise worthy of consideration if one knowingly and deliberately alters his conditions of action, concealing the fact from him."); Joel R. Reidenberg et al., *Privacy Harms and the Effectiveness of the Notice and Choice Framework*, 11 I/S: J. L. & POL'Y FOR INFO. SOC'Y 485, 512–15 (2014) (describing harms of surreptitious collection).

[39] Solove, *supra* note 32, at 507.

[40] Emily Reynolds, *Psychologists Are Mining Social Media Posts For Mental Health Research — But Many Users Have Concerns*, BRIT. PSYCH. SOC'Y (June 29, 2020), https://digest.bps.org.uk/2020/06/29/psychologists-are-mining-social-media-posts-for-mental-health-research-but-many-users-have-concerns/ [https://perma.cc/3USB-YLA4].

[41] Solove, *supra* note 32, at 521.

explicitly prohibited developers from using Twitter data for facial recognition purposes.[42] Just like information collection, secondary use creates feelings of unease and discomfort.[43]

Further, data scraping creates security harms. The app "Girls Around Me" scraped data from Foursquare to identify where women were in real-time.[44] Another app, "PleaseRobMe," scraped Twitter location data to identify when people were away from their homes.[45] Accumulated scraped data is also vulnerable to security breaches. Already, scrapers have suffered data breaches.[46] Moreover, the failure to provide users notice of scraping — Professor Solove terms this "exclusion" — can itself be a harm.[47] Exclusion reduces the accountability of data collectors and can create feelings of vulnerability, uncertainty, and powerlessness because of "[a]n inability to participate in the maintenance and use of one's information."[48]

2. Privacy Because of Obscurity

One conceptual framework for analyzing privacy in public is to calibrate privacy in reference to obscurity — "the notion that when our activities or information is unlikely to be found, seen, or remembered, it is, to some degree, safe."[49] Professor Woodrow Hartzog argues that the private/public dichotomy is not an on/off switch for privacy interests; rather, privacy is a spectrum of obscurity. Because information is obscure, we can reasonably expect to retain privacy interests in such information even if it is technically public. Just like a conversation at a crowded restaurant is private because of obscurity — in that we do not expect eavesdroppers, spies, or passersby to listen in — so too is most of our information online.[50] The vast majority of the Internet simply ignores your Facebook pictures. The passage of time also makes information obscure: no one remembers your Myspace pictures from fifteen

[42] *More About Restricted Uses of the Twitter APIs*, TWITTER, https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases [https://perma.cc/89BX-7SNM].

[43] Solove, *supra* note 32, at 521–22 ("Secondary uses thwart people's expectations about how the data they give out will be used . . . [and can lead to] fear and uncertainty . . . [and] a sense of powerlessness and vulnerability.").

[44] Nick Bilton, *Girls Around Me: An App Takes Creepy to a New Level*, N.Y. TIMES (Mar. 30, 2012, 4:43 PM), https://nyti.ms/2jKC0zL [https://perma.cc/4M9H-QUMA].

[45] MG Siegler, *Please Rob Me Makes Foursquare Super Useful for Burglars*, TECHCRUNCH (Feb. 17, 2010, 2:17 PM), https://techcrunch.com/2010/02/17/please-rob-me-makes-foursquare-super-useful-for-burglars/ [https://perma.cc/P3M4-BNDH].

[46] Alex Scroxton, *Social Media Data Leak Highlights Murky World of Data Scraping*, COMPUTER WEEKLY (Aug. 20, 2020, 1:15 PM), https://www.computerweekly.com/news/252487895/Social-media-data-leak-highlights-murky-world-of-data-scraping [https://perma.cc/52W9-R5UM].

[47] Solove, *supra* note 32, at 522–23.

[48] *Id.* at 523.

[49] Hartzog, *supra* note 8, at 515.

[50] *Id.* at 515–16.

years ago. There are also transaction costs to accessing information. Typically, the immense manpower needed prevents someone from collecting all your photos from every social media website you have ever used — "just because information is hypothetically available does not mean most (or even a few) people have the knowledge and ability to access information."[51] Most websites actually promote privacy through obscurity in their terms of service.[52] As Professors Woodrow Hartzog and Frederic Stutzman explain, "[t]erms of use . . . prevent other social technology users from engaging in obscurity-eroding behavior, such as scraping data from websites."[53] Ultimately, widescale, automated collection of personal information via scraping destroys obscurity by reducing the transaction costs and difficulties in accessing and understanding personal information.

3. Privacy Because of Trust in Websites and Other Users

Another useful framework for conceptualizing privacy in public is to analyze "relationships of trust."[54] One argument, first proposed by Professor Jack Balkin, is that the website/user relationship is one of trust, just like the attorney/client relationship.[55] Accordingly, websites are "information fiduciaries" and have "special duties with respect to personal information that they obtain in the course of their relationships" with users.[56] One of the duties these websites could have is to protect users' personal information from unauthorized third-party scraping.[57] Already, websites have duties to protect users' private personal information from security hacks.[58] Given that unauthorized use of public personal information creates privacy harms just as much as unauthorized use of private personal information, websites' duties should extend to protecting public personal information.

[51] *Id.* at 516.

[52] Casey Fiesler, Nathan Beard & Brian C. Keegan, *No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service*, 14 PROCS. INT'L AAAI CONF. WEB & SOC. MEDIA 187, 191 (2020) (finding that 80% of social media websites ban scraping).

[53] Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. L. REV. 385, 407 (2013).

[54] Hartzog, *supra* note 8, at 518.

[55] Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183, 1186 (2016).

[56] *Id.* at 1208. *But see* Lina M. Khan & David E. Pozen, *A Skeptical View of Information Fiduciaries*, 133 HARV. L. REV. 497, 504 (2019) (critiquing information fiduciary theory).

[57] *But see* Benjamin L. W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147, 183–87 (2021) (arguing that users cannot trust websites to protect their information from scrapers).

[58] *See, e.g.*, CAL. CIV. CODE § 1798.150(a)(1) (West 2020).

Tellingly, many websites hold themselves to this standard of care because their terms of service prohibit scraping and because users expect websites to enforce their scraping prohibitions.[59] Websites explain how scraping violates their terms of service and can lead to expulsion.[60] Websites also actively enforce these scraping bans using technological restrictions and litigation.[61] In fact, websites even advertise how their technological and legal restrictions on scraping protect user privacy. Facebook, for example, describes its efforts to combat scraping as "part of [its] ongoing work to protect people's privacy."[62] In sum, users trust websites to enforce these terms of service and to protect users from scraping.[63]

Although not an information fiduciary relationship like the website/user relationship, user/user relationships are also infused with trust. Users expect other users to comply with the terms of service and to refrain from scraping.[64] Ultimately, users expect social media websites to be safe, trusting these environments enough to share information with one another without fear that their data will be subject to scraping schemes. This environment of trust is essential for social functioning: "[i]f we cannot trust others with our personal information, society will suffer."[65]

4. Posting Publicly is Not Implied Consent

Clearview has argued that "the individuals who posted . . . on the Internet effectively consented to sharing their [personal] information . . . with the public at large."[66] However, when users post, they

[59] Fiesler et al., *supra* note 52, at 191 (estimating that 80% of social media websites prohibit scraping).

[60] *See Terms of Service*, FACEBOOK (Oct. 22, 2020), https://www.facebook.com/terms.php [https://perma.cc/6W8K-QSDB].

[61] *See, e.g.*, Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1065–67 (9th Cir. 2016).

[62] *Taking Legal Action Against Scraping*, FACEBOOK (Oct. 15, 2020, 3:45 PM), https://about.fb.com/news/2020/10/taking-legal-action-against-data-scraping/ [https://perma.cc/E8D2-QXJF].

[63] But if a website expressly allows third-party scraping, the user has a weaker claim to privacy. Twitter, for example, tells users: "By publicly posting content when you Tweet, you are directing us to disclose that information as broadly as possible, including through our APIs, and directing those accessing the information through our APIs to do the same." *Twitter Privacy Policy*, TWITTER, https://twitter.com/en/privacy [https://perma.cc/A82E-FF2F]. Similarly, many websites expressly authorize web crawling (e.g., Google search). *See, e.g.*, *Twitter Terms of Service*, TWITTER, https://twitter.com/en/tos [https://perma.cc/D5DT-SDRV] ("[C]rawling [Twitter] is permissible if done in accordance with the provisions of the robots.txt file.").

[64] *See* Sobel, *supra* note 57, at 45 (arguing for a bad-faith breach of contract cause of action to address instances like Clearview, wherein one user can sue another user for a bad-faith breach of the terms of service).

[65] Hartzog, *supra* note 8, at 518.

[66] Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 3, at 22 (internal quotation marks omitted).

do so believing that their information will be obscure and in an environment of trust.[67] Users who post publicly may also be unaware of the privacy implications of publicly available personal information.[68] Therefore, if anything, the implication is that users expect privacy and do not expect their information to be swept up by data scraping.

Data privacy laws recognize implied consent as flawed. Under the General Data Protection Regulation ("GDPR") — the E.U.'s comprehensive, omnibus data privacy law — consent requires a "clear affirmative act" and implied consent such as "pre-ticked boxes or inactivity" is insufficient.[69] Similarly, only "written release" satisfies consent in the Biometric Information Privacy Act ("BIPA"), which is Illinois' biometric data privacy law.[70] Further, the GDPR distinguishes between consent given for different purposes: "[w]hen the processing has multiple purposes, consent should be given for all of them."[71] In other words, even if a user makes the affirmative choice to make her LinkedIn profile public, she manifests an intent to participate in an obscure and trustworthy environment, not an intent to participate in data harvesting.[72]

* * * * *

In sum, there are strong privacy interests in public personal information. The collection, aggregation, and analysis of public personal information create a wide range of harms, from social anxiety harms to security harms. Many social media users reasonably expect to escape these harms because their information is obscure and because they trust websites and other users. However, unauthorized third-party scraping erodes obscurity and trust, placing users' privacy at significant risk.

[67] *See supra* Sections II.B.2 and II.B.3. Even if the social media website, by default, sets posting to private — meaning the user must affirmatively decide to share with the world — that does not mean the user has consented because in most instances, the user expects obscurity. *See* Kirsten Martin & Helen Nissenbaum, *Privacy Interests in Public Records: An Empirical Investigation*, 31 HARV. J.L. & TECH. 111, 117–19 (2017) (presenting empirical evidence).

[68] *See* Casey Fiesler & Nicholas Proferes, *"Participant" Perceptions of Twitter Research Ethics*, 4 SOC. MEDIA & SOC'Y 1, 2 (2018) (finding that many Twitter users are unaware that their public tweets can be used by researchers).

[69] Regulation 2016/679 of Apr. 27, 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 (General Data Protection Regulation), 2016 O.J. (L 119) 1, Recital 32 (EU) [hereinafter GDPR].

[70] 740 ILL. COMP. STAT. 14/15(b)(3) (2008).

[71] GDPR, *supra* note 69, at Recital 32. The California Consumer Privacy Act of 2018 has similar limitations. CAL. CIV. CODE § 1798.100(b) (West 2018) ("A business shall not collect additional categories of personal information or use personal information collected for additional purposes without providing the consumer with notice consistent with this section.").

[72] Perhaps posting publicly may become implied consent once the attitudes toward scraping change. That is, if all Internet users know that third-party scraping abounds, posting publicly becomes implied consent.

III. The Existing Regulatory Landscape

Despite the tension between privacy and public information, several legal regimes are available to protect privacy interests in public personal information: (1) the website can assert CFAA and contract law claims, (2) users and regulatory agencies can assert state data privacy law claims (e.g., CCPA and BIPA), and (3) regulatory agencies can enforce state data broker laws.

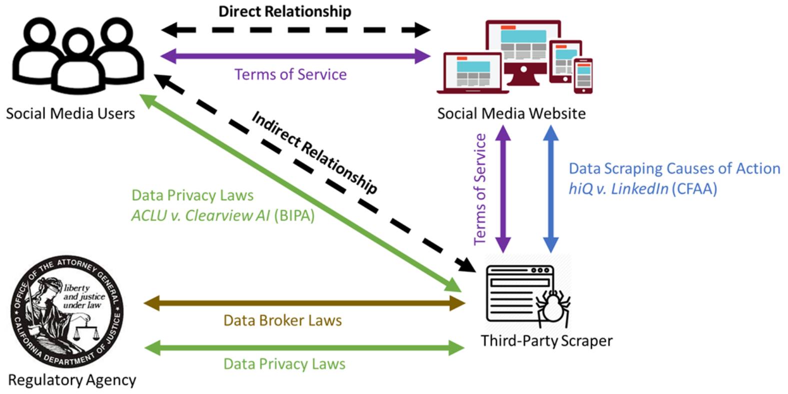

Figure 1: Summary of the Scraping Ecosystem.

A. Website Enforcement of Data Scraping Causes of Action

A website can assert a CFAA or contract law claim against the scraper. In a contract law claim, the website argues that the scraper is bound by the website's terms of service, which prohibit scraping.[73] These cases turn on whether the terms of service (typically a "browsewrap" agreement) is an enforceable contract, which requires "actual or constructive knowledge of a website's terms and conditions."[74] Terms of service contained in hyperlinks at the bottom of a webpage are generally unenforceable because they are not conspicuous enough.[75] But, in one data scraping case, the court found that the scraper had constructive notice of a browse-wrap terms of service because the scraper's own website had a similar browse-wrap provision prohibiting scraping.[76]

[73] *See* 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, 272–76 (2018) (collecting contract law cases on data scraping); Sobel, *supra* note 57 (proposing a cause of action where users assert a breach of the scraper/website contract).

[74] Nguyen v. Barnes & Noble Inc., 763 F.3d 1171, 1176 (9th Cir. 2014).

[75] *Id.* at 1177–79.

[76] DHI Grp., Inc. v. Kent, No. CV H-16-1670, 2017 WL 4837730, at *2–4 (S.D. Tex. Oct. 26, 2017).

Perhaps the ubiquity of "no scraping" provisions may help show constructive notice.[77]

Under the CFAA, the website can argue that it has imposed technological and/or verbal restrictions preventing scraping and that the scraper has circumvented these restrictions and accessed the website "without authorization" or has "exceed[ed] authorized access."[78] But much to the chagrin of privacy advocates like Electronic Privacy Information Center,[79] *hiQ v. LinkedIn* severely limits CFAA claims for scraping. In *hiQ*, the Ninth Circuit found the CFAA inapplicable to publicly available information because "'without authorization' . . . suggests a baseline in which access is not generally available and so permission is ordinarily required."[80] According to *hiQ*, "[w]here the default is free access without authorization, in ordinary parlance one would characterize selective denial of access as a ban, not as a lack of authorization."[81]

After the New York Times' report on Clearview, websites such as Facebook and YouTube sent cease-and-desist letters to Clearview for violating their terms of service.[82] While *hiQ* likely prevents these websites from asserting CFAA claims against Clearview, a contract law claim for breach of the terms of service could be meritorious.

Overall, websites may be better positioned than users (and agencies) to go after scrapers. In many instances (especially if the scraper does not provide notice), users will not even know that their information is being scraped. Unlike users, websites can monitor web traffic to find instances of scraping. Websites also have more resources to litigate claims. Strengthening data scraping causes of action can also channel scraping into websites' application programming interfaces ("APIs"), which are website-created tools specifically facilitating the scraping process. APIs protect privacy because, in order to use an API, a scraper must abide by the API's terms of use, which place restrictions on collecting users' personal information.[83] Websites can also use APIs

[77] *See* Fiesler et al., *supra* note 52, at 191 (finding 80% of social media websites ban scraping).

[78] 18 U.S.C. § 1030(a)(2).

[79] Electronic Privacy Information Center filed a Supreme Court amicus brief supporting LinkedIn. Brief of EPIC as Amicus Curiae in Support of Petitioner, LinkedIn Corp. v. hiQ Labs, Inc., No. 19-1116 (U.S. Apr. 13, 2020), https://www.epic.org/amicus/cfaa/linkedin/EPIC-Amicus-LinkedIn-v-hiQ.pdf [https://perma.cc/7UUH-G92D].

[80] Petition for Writ of Certiorari at 13, LinkedIn Corp. v. hiQ Labs, Inc., No. 19-1116 (U.S. Mar. 9, 2020).

[81] *Id.* at 13 (internal quotation marks omitted).

[82] Charlie Wood, *Facebook has Sent a Cease-and-Desist Letter to Facial Recognition Startup Clearview AI for Scraping Billions of Photos*, BUS. INSIDER (Feb. 6, 2020, 8:51 AM), https://www.businessinsider.com/facebook-cease-desist-letter-facial-recognition-cleaview-ai-photo-scraping-2020-2 [https://perma.cc/ZG6V-GFL4].

[83] Twitter's developer terms of service require a scraper to "apply for a developer account" and Twitter "review[s] all proposed uses" of its API. *Developer Policy*, TWITTER, https://developer.twitter.com/en/developer-terms/policy [https://perma.cc/B4BL-UGZZ].

as a technological mechanism to protect privacy. For example, when a scraper uses the LinkedIn API to extract information from a user's public profile, the API automatically sends a notice-and-consent message to the affected user.[84]

But relying on websites to enforce their users' privacy interests may, as *hiQ* warned, promote anticompetitive behavior.[85] Moreover, it is unclear whether users can actually trust websites to protect them from scrapers.[86] While platforms have uniformly decried Clearview's scraping, it is unclear whether websites would mount such a unified response to other types of scraping.

B. Agency Enforcement of Data Broker Laws

California and Vermont have data broker laws, which specifically regulate a "data broker" — "a business that knowingly collects and sells to third parties the personal information of a consumer with whom the business does not have a *direct* relationship."[87] These laws try to fix the information asymmetry in the scraper/user relationship: because of the indirect, arms-length interaction, users "are generally not aware that data brokers possess their personal information."[88] The California assembly bill introducing the data broker law states its legislative intent:

> Consumers who have a direct relationship with traditional and e-commerce businesses . . . may have some level of knowledge about and control over the collection of data by those businesses . . . . By contrast, consumers are generally not aware that data brokers possess their personal information, how to exercise

Twitter's developer terms also unequivocally prohibit certain uses of its API, including scraping for facial recognition. *More About Restricted Uses of the Twitter APIs*, TWITTER, *supra* note 42.

[84] *Profile API*, MICROSOFT, https://docs.microsoft.com/en-us/linkedin/shared/integrations/people/profile-api [https://perma.cc/2CKS-YXCP] (explaining permissions required to retrieve profiles); *Authorization Code Flow (3-legged OAuth)*, MICROSOFT, https://docs.microsoft.com/en-us/linkedin/shared/authentication/authorization-code-flow [https://perma.cc/2B4Z-NL4R] (explaining how users must consent to permission requests).

[85] hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 998 (9th Cir. 2019) ("If companies like LinkedIn, whose servers hold vast amounts of public data, are permitted selectively to ban only potential competitors [like hiQ] from accessing and using that otherwise public data, the result — complete exclusion of the original innovator [hiQ] in aggregating and analyzing the public information — may well be considered unfair competition under California law.").

[86] Sobel, *supra* note 57, at 45–50 (describing the problem as a "trust-your-overlords" problem). The information fiduciary concept would impose such an obligation on websites, but the information fiduciary concept is not yet binding law.

[87] CAL. CIV. CODE § 1798.99.80(d) (West 2019) (emphasis added); *see also* VT. STAT. ANN. tit. 9, § 2430(4)(A) (2019).

[88] Assemb. B. 1202, 2019–2020 Reg. Sess. (Cal. 2019); *see also* H.B. 764, 2017–2018 Leg. Sess. (Vt. 2018) (explaining the same legislative intent).

> their right to opt out, and whether they can have their information deleted, as provided by California law . . . . Therefore, it is the intent of the Legislature to further Californians' right to privacy by giving consumers an additional tool to help control the collection and sale of their personal information by requiring data brokers to register annually with the Attorney General and provide information about how consumers may opt out of the sale of their personal information.[89]

To make the practices of indirect data collectors more transparent, these data broker statutes impose registration requirements.[90] For example, the California data broker statute requires data brokers to provide their name, address, and data collection practices and makes these entries available online.[91] After the New York Times exposé, Clearview registered in California[92] and Vermont.[93] In theory, these registration entries provide users with notice of data brokers' practices, allowing users to exercise control over their personal information.

C. User and Agency Enforcement of Data Privacy Laws

In the United States, omnibus data privacy regulation has primarily been state-driven. Two of the most significant state data privacy laws are California's CCPA and Illinois's BIPA. The CCPA and BIPA give state residents certain rights — such as the right to receive notice of data collection — vis-à-vis "personal information" collected by regulated entities.[94] The E.U.'s omnibus data privacy regulation — the GDPR — functions similarly to the CCPA and BIPA.

[89] Assemb. B. 1202, 2019–2020 Reg. Sess. (Cal. 2019).

[90] *See* CAL. CIV. CODE § 1798.99.82 (West 2019) (requiring data broker to provide name, address, and data collection practices); VT. STAT. ANN. tit. 9, § 2446 (2019) (requiring data broker to provide name, address, opt-out policies, data collection practices, and information on security breach occurrences). The Vermont statute also requires data brokers maintain security standards. VT. STAT. ANN. tit. 9, § 2447(a)(1) (2019).

[91] CAL. CIV. CODE § 1798.99.82 (West 2019); *Data Broker Registry*, CAL. DEP'T JUST., OFF. ATT'Y GEN., https://oag.ca.gov/data-brokers [https://perma.cc/T92U-ZB46].

[92] Data Broker Registration for Clearview AI, Inc., CAL DEP'T JUST., OFF. ATT'Y GEN. (2020), https://oag.ca.gov/data-broker/registration/185841 [https://perma.cc/URS6-E8RC].

[93] Data Broker Information: Clearview AI, Inc., VT. SEC'Y OF STATE (2020), https://bizfilings.vermont.gov/online/DatabrokerInquire/DataBrokerInformation?businessID=367103 [https://perma.cc/ZP5P-YGNK].

[94] CAL. CIV. CODE § 1798.100(b) (West 2018); *accord* 740 ILL. COMP. STAT. 14/15(b)(1)–(2) (2008).

Given the CCPA's, GDPR's, and BIPA's broad definitions of "collect"[95] and regulated entities,[96] third-party scrapers are bound by these privacy regulations. In addition, all three privacy regulations apply to public personal information. Although the CCPA does not apply to "publicly available information," "publicly available" is narrowly defined to mean "information that is lawfully made available from federal, state, or local government records" and expressly "does not mean biometric information collected by a business about a consumer without the consumer's knowledge."[97] Neither BIPA nor GDPR makes any exceptions for publicly available information. Finally, the CCPA limits private causes of action to data breach suits,[98] but the California Attorney General can bring actions to enforce the CCPA's requirements.[99] BIPA and GDPR create private rights of action.[100]

One of the most significant rights guaranteed by data privacy laws is the right to receive notice of data collection. Notice informs users, allowing them to control how their data is used. The CCPA, GDPR, and BIPA all provide the right to notice, but each implements the right differently. The CCPA exempts scrapers from providing notice, the GDPR requires notice unless impracticable, and BIPA requires notice every time. Different implementations have arisen because the indirect relationship between scraper and user can make notice difficult to provide.

1. The Right to Receive Notice Prior to Data Collection: BIPA and CCPA

The CCPA guarantees the right to receive notice at or before the point of data collection.[101] A recently promulgated CCPA implement-

[95] The CCPA defines "collect[]" as "buying, renting, gathering, obtaining, receiving, or accessing any personal information pertaining to a consumer by any means. This includes receiving information from the consumer, either actively or passively, or by observing the consumer's behavior." CAL. CIV. CODE § 1798.140(f) (West 2018). BIPA does not define "collect" as a term of art, but BIPA describes a broad swathe of regulated activities: "collect, capture, purchase, receive through trade, or otherwise obtain." 740 ILL. COMP. STAT. 14/15(b) (2008). GDPR uses the terminology "process." GDPR, *supra* note 69, at art. 4(2).

[96] Regulated entities under the CCPA are "business[es]." CAL. CIV. CODE § 1798.140(a) (West 2018). Regulated entities under BIPA are "private entit[ies]." 740 ILL. COMP. STAT. 14/10 (2008). Regulated entities under GDPR are "controllers." GDPR, *supra* note 69, at art. 4(7).

[97] CAL. CIV. CODE § 1798.140(v)(2) (West 2018).

[98] *See* CAL. CIV. CODE § 1798.150 (West 2018).

[99] CAL. CIV. CODE § 1798.155(b) (West 2018).

[100] The ACLU recently asserted BIPA claims against Clearview for failures to provide notice and obtain consent prior to its data scraping. Complaint at 3, ACLU v. Clearview AI, No. 2020-CH-04353 (Ill. Cir. Ct. May 28, 2020). GDPR, *supra* note 69, at art. 80(2), 82(1).

[101] CAL. CIV. CODE § 1798.100(b) (West 2018); 740 ILL. COMP. STAT. 14/15(b)(1)–(2) (2008).

ing regulation, however, exempts some scrapers from the notice requirement: "[a] business that does not collect personal information *directly* from the consumer [e.g., a scraper] does not need to provide a notice at collection to the consumer if it does not sell the consumer's personal information."[102] In its statement of reasons, the California Attorney General explained that businesses collecting personal information indirectly from consumers "cannot feasibly provide a notice 'at or before the point of collection.'"[103] The scope of this implementing regulation is still unclear, and it is arguable whether collecting training data for machine learning like Clearview did is a sale subject to notice requirements.[104] That said, because Clearview does not disclose its training data to others, the remainder of this Note treats the CCPA as providing a broad, almost categorical exemption from the notice requirement for Clearview and similar scrapers.[105] Unlike the CCPA, BIPA does not distinguish between direct and indirect collection. Under BIPA, all scrapers must provide notice before data collection.[106]

2. The Right to Receive Notice Prior to Data Collection: GDPR Article 14

An interesting point of comparison to U.S. law is the E.U.'s data privacy law, the GDPR. Like the CCPA and BIPA, the GDPR applies to public personal information.[107] But the GDPR takes an intermediate

[102] CAL. CODE REGS. tit. 11, § 999.305(d) (2020) (emphasis added). *See also* CAL. CODE REGS. tit. 11, § 999.305(e) ("A data broker registered with the Attorney General . . . does not need to provide a notice at collection to the consumer if it has included in its registration submission a link to its online privacy policy that includes instructions on how a consumer can submit a request to opt-out."); CAL. CODE REGS. tit. 11, § 999.301(l) (2020) ("'Notice at collection' means the notice given by a business to a consumer at or before the point at which a business collects personal information from the consumer as required by Civil Code section 1798.100, subdivision (b), and specified in these regulations.").

[103] OFF. CAL. ATT'Y GEN., FINAL STATEMENT OF REASONS 11 (2020), https://oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-fsor.pdf [https://perma.cc/EZ2H-86KG].

[104] *See, e.g.*, Nate Garhart, *Data Scraping Under the Revised CCPA Regulations*, JDSUPRA (Mar. 20, 2020), https://www.jdsupra.com/legalnews/data-scraping-under-the-revised-ccpa-43249/ [https://perma.cc/YHE7-5DL8] (posing hypotheticals).

[105] CAL. CIV. CODE § 1798.140(ad)(1) (West 2018) ("[S]ale . . . means selling, renting, releasing, disclosing, disseminating, making available, transferring, or otherwise communicating orally, in writing, or by electronic or other means, a consumer's personal information by the business to a third party for monetary or other valuable consideration.") (internal quotation marks omitted).

[106] BIPA requires all biometric collectors to provide notice and receive consent prior to collecting biometric information. 740 ILL. COMP. STAT. 14/15(b) (2008) ("No private entity may collect, capture, purchase, receive through trade, or otherwise obtain a person's or a customer's biometric identifier or biometric information, unless [notice is given and consent is received].").

[107] GDPR, *supra* note 69, at art. 4(1) ("'[P]ersonal data' means any information relating to an identified or identifiable natural person . . . who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location

approach between BIPA's no-exceptions notice requirement and the CCPA's exemption for scrapers. Under Article 14 of the GDPR, businesses that indirectly collect personal information need to provide notice unless doing so "proves impossible or would involve a disproportionate effort" (the impracticability exception).[108] The Polish data protection authority UODO recently fined a company for scraping personal data (names, personal identification numbers, and addresses) from a public, government-maintained database of business registration records.[109] According to UODO, the scraper failed to provide the requisite Article 14 notice to over six million affected individuals.[110] UODO also narrowly interpreted the impracticability exception, rejecting the scraper's argument that it would be exceedingly burdensome and expensive to provide notice to six million individuals via telephone and/or postal mail.[111] Similarly, the French data protection authority CNIL recently reminded commercial prospectors (companies that scrape contact information in order to send advertisements) that their scraping needed to adhere to the notice-and-consent requirement.[112]

IV. Shortcomings of the Existing Regulatory Landscape

While regulations exist to protect users' privacy interests in public personal information, they fail to address several significant problems. First, regulations — such as the CCPA — exempting scrapers from providing notice prior to collection fail to recognize the importance of notice. Second, regulations should require opt-in consent because opt-in consent protects privacy better than opt-out consent. Lastly, websites

data, an online identifier or to one or more factors specific to the . . . identity of that natural person.").

[108] GDPR, *supra* note 69, at art. 14(5)(b). *See also Right to Be Informed*, INFO. COMM'R'S OFF., https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-be-informed/ [https://perma.cc/YYT4-9KYL] ("If you obtain personal data from publicly accessible sources: You still have to provide people with privacy information, unless you are relying on an exception or an exemption.").

[109] *The First Fine Imposed by the President of the Personal Data Protection Office*, URZĄD OCHRONY DANYCH OSOBOWYCH (UODO) (Mar. 26, 2019), https://uodo.gov.pl/en/553/1009 [https://perma.cc/4HWU-H2RA] [hereinafter *The First Fine*, UODO]; URZAD OCHRONY DANYCH OSOBOWYCH [UODO] [office for personal data protection] Mar. 15, 2019, ZSPR.421.3.2018 (Pol.).

[110] *The First Fine*, UODO, *supra* note 109.

[111] *Id.*

[112] *CNIL Publishes Guidance on Web Scraping and Re-Use of Publicly Available Online Data for Direct Marketing*, HUNTON ANDREWS KURTH (May 4, 2020), https://www.huntonprivacyblog.com/2020/05/04/cnil-publishes-guidance-on-web-scraping-and-re-use-of-publicly-available-online-data-for-direct-marketing/ [https://perma.cc/PX92-FRBG].

need to take a more active role in protecting users from scraping attacks, but we should be wary of granting websites monopolistic control over scraping.

A. Pre-Collection Notice is Necessary to Protect Privacy from Scrapers

Notice is critical in the context of third-party scraping. Users are aware, to some extent, that websites are collecting their personal information.[113] Websites' privacy policies describe the scope of data collection, and websites provide pop-ups and notifications of their data collection practices. But, because of the indirect relationship between users and scrapers, users are often unaware that scrapers even exist.[114] Thus, unauthorized third-party scrapers can operate surreptitiously.

Data broker laws try to fix this knowledge gap,[115] but they ultimately fail because most users do not have the wherewithal to check data broker registries. These data broker registration filings are also incredibly sparse: for example, neither Clearview's California nor Vermont registration expressly lists which websites Clearview scrapes.[116] Scrapers operating surreptitiously may also ignore data broker laws. Indeed, Clearview ignored data privacy laws and contractual restrictions prohibiting scraping for nearly three years before investigative journalism uncovered its alarming practices.[117]

This lack of knowledge creates serious privacy harms. Covert information collection undermines individual autonomy and free choice.[118] The lack of notice also excludes individuals from the data collection process, making individuals feel powerless in controlling how their data is used.[119] This powerlessness is not just a feeling but is

[113] *See* Brooke Auxier et al., *Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information*, PEW RSCH. CTR. (Nov. 15, 2019), https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/ [https://perma.cc/BXC9-WKD8] (studying perceptions toward data collection); Assemb. B. 1202, 2019-2020 Reg. Sess. (Cal. 2019). The lack of user awareness of data brokers was one of the motivations for passing data broker laws.

[114] Fiesler & Proferes, *supra* note 68, at 1–2 (finding that many Twitter users are unaware that their public tweets can be used by researchers).

[115] *See* Assemb. B. 1202, 2019–2020 Reg. Sess. (Cal. 2019).

[116] Data Broker Information, *supra* note 93 ("Clearview AI Inc. collects publicly available images."); Data Broker Registration, *supra* note 92 (providing no information on collection practices).

[117] Clearview only registered in Vermont and California after the New York Times story broke. Data Broker Registration, *supra* note 92 (CA registration approved July 30, 2020); Data Broker Information, *supra* note 93 (VT registration dated Jan. 14, 2020).

[118] *See, e.g.*, Seals, *supra* note 36; Richards, *supra* note 37; Benn, *supra* note 38; Reidenberg et al., *supra* note 38.

[119] Solove, *supra* note 32, at 488.

itself a concrete harm. If users knew that scrapers collected public information, they might choose to make their information private.[120] In fact, users have switched their privacy settings in response to the New York Times' Clearview story.[121] Without notice of scraping practices, users would be oblivious to the need to change their privacy settings.[122] In addition to excluding users from self-help, the lack of notice precludes users from exercising statutory data privacy rights, such as the right to request deletion.[123]

In the long run, requiring scrapers to inform users of their data collection practices will educate users on the harms of leaving their personal information publicly available and allow users to exercise statutory data privacy rights and simple self-help remedies.[124] Admittedly, in assuming that users are ill-informed on the dangers of publicly posting personal information, this approach is paternalistic.[125] However, this lack of knowledge is a real issue, as highlighted by the Clearview scandal.

Many of the traditional criticisms of notice also have little force in the scraping context. Notice is often criticized for relying on complex privacy disclosures which are unintelligible to the average user, but here the notice provided by scrapers is simple: we will harvest public data.[126] Another critique of notice is that it does not facilitate real choice.[127] In the usual application of notice-and-consent to the website/user relationship, the website's privacy policy is a take-it-or-leave-it deal.[128] However, in the scraping context, there is no coercion to consent to scraping. If anything, most websites allow users to make their

[120] *See* Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1880–82 (2013).

[121] Thomas Smith, *I Got My File From Clearview AI, and It Freaked Me Out*, ONEZERO (Mar. 24, 2020), https://onezero.medium.com/i-got-my-file-from-clearview-ai-and-it-freaked-me-out-33ca28b5d6d4 [https://perma.cc/2Q9W-SL8A] ("I was so shocked by the data available there that I made my profile private after receiving Clearview's report.").

[122] *See* Daniel Susser, *Notice After Notice-and-Consent: Why Privacy Disclosures Are Valuable Even If Consent Frameworks Aren't*, 9 J. INFO. POL'Y 37, at 52–55 (2019).

[123] CAL. CIV. CODE § 1798.105(a) (West 2018).

[124] *See* Reidenberg et al., *supra* note 38, at 519 ("[S]urreptitious collection can also be avoided *ex ante* through proper notice.").

[125] *See* Solove, *supra* note 120, at 1894–98.

[126] Susser, *supra* note 122, at 43–47; Solove, *supra* note 120, at 1885. In the traditional user/website context, a user consents to an extremely long privacy agreement, which contains all the details on the website's data collection. *See* Alexis C. Madrigal, *Reading the Privacy Policies you Encounter in a Year Would Take 76 Work Days*, ATLANTIC (Mar. 1, 2012), https://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851/ [https://perma.cc/CEW3-7G4Y]. Of course, it is possible that a scraper provides an equally long and complex disclosure to a user, but there is simply less information being conveyed in the scraping context.

[127] Susser, *supra* note 122, at 43–47; *see also* Solove, *supra* note 120, at 1883–86.

[128] Susser, *supra* note 122, at 43–47; Solove, *supra* note 120, at 1886–93.

information private and un-scrapable. Ultimately, notice can facilitate truly informed choice.[129]

However, it can be difficult for indirect data collectors to provide notice because they have no direct line of communication with users. BIPA ignores this impracticability and requires all collectors to provide notice prior to collection. On the other hand, the CCPA exempts indirect collectors from providing such notice. The GDPR takes an intermediate approach, requiring notice unless it is impracticable.

In light of the importance of notice and the ability of scrapers to operate surreptitiously, the CCPA's loophole for indirect data collection is flawed and needs to be closed. The GDPR's impracticability exception is less alarming than the CCPA's loophole because the GDPR's exception has been narrowly interpreted.[130] In Clearview's case, for example, there would be no GDPR impracticability to providing notice because it is easy to identify and communicate with the social media user who made a post (simply send a direct message).[131] UODO found that mailing/calling six million people was insufficiently impracticable.[132] Certainly, sending electronic messages to social media users is more manageable than mailing/calling individuals.[133] In sum, pre-collection notice is necessary because it serves an invaluable role and because it is very easy for scrapers to provide in the social media context.

B. The Role of Consent: Opt-In or Opt-Out?

Another consideration is the role of consent. Here, too, there is variation among the CCPA, GDPR, and BIPA. The GDPR and BIPA require opt-in consent, meaning the user must affirmatively consent to

[129] *See* Robert H. Sloan & Richard Warner, *Beyond Notice and Choice: Privacy, Norms, and Consent*, 14 J. HIGH TECH. L. 370, 411–14 (2014) (explaining when consent is free and informed); *cf.* Solove, *supra* note 120 (presenting critiques of notice and consent).

[130] The Working Party in the Prot. of Individuals with Regard to the Processing of Pers. Data, *Guidelines on Transparency under Regulation 2016/679*, WP260 rev.01, at 28 (Apr. 11, 2018), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227 [https://perma.cc/AYU9-Q4MM] [hereinafter *Guidelines on Transparency under 2016/679*] ("[Art. 14.5(a)] should, as a general rule, be interpreted and applied narrowly.").

[131] *See Are There any Exceptions?*, INFO. COMM'N ORG., https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/the-right-to-be-informed/are-there-any-exceptions/#id4 [https://perma.cc/ZLA8-7LS9] ("To rely on this [impracticability] exception . . . there is a proportionate balance between the effort involved for you to provide [notice] and the effect that your use of their personal data will have on them. The more significant the effect, the less likely you will be able to rely on this exception.").

[132] *See supra* notes 108–109 and accompanying text.

[133] But social media websites may have restrictions on who can send direct messages. Tatum Hunter, *Instagram Bots Are Frowned Upon — But They Work*, BUILTIN (Oct. 6, 2020), https://builtin.com/software-engineering-perspectives/automating-social-media-growth [https://perma.cc/Q2XT-TCTD]. Allowing any bot to send direct messages can inundate users' inboxes and hurt users' privacy. Therefore, websites play an important role in facilitating scraper/user communication.

data collection.[134] The CCPA, on the other hand, uses opt-out consent, meaning that a scraper does not need to receive the user's affirmative consent; instead, the user must affirmatively object to data collection practices.[135] In opt-in consent, the default is privacy protection.[136] In opt-out, the burden shifts to the consumer to actively protect their privacy.[137]

First, there is a question of whether consent is even necessary. Is pre-collection notice alone sufficiently protective of user privacy?[138] The baseline technological rule is opt-out consent.[139] The social media user can always resort to the ultimate opt-out — the self-help remedy of making her personal information private.[140] Once a social media user changes her privacy setting, she has opted out of all scraping. Therefore, a legal rule not requiring consent is meaningless given the technological rule providing for opt-out consent. That is, the baseline technological rule already provides users with opt-out consent.

Then, the question is whether the legal rule should be opt-in consent or opt-out consent. Opt-in consent substantially furthers privacy interests. For example, affirmative consent ensures notice was actually received and that inactive users — which make up an estimated 60% of all social media accounts — are not swept into the scraping.[141] Even

[134] GDPR, *supra* note 69, at art. 6(1)(a); 740 ILL. COMP. STAT. § 15(b)(2). Note, however, the GDPR allows alternatives to opt-in consent. *See id.*, at art. 6(1)(b)–(f). The broadest of these alternatives is when collection is "necessary for the purposes of the legitimate interests." *Id.*, at art. 6(1)(f). For example, "[t]he processing of personal data strictly necessary for the purposes of preventing fraud . . . [is] a legitimate interest of the data controller concerned." *Id.*, at Recital 47.

[135] Grace Park, Note, *The Changing Wind of Data Privacy Law: A Comparative Study of the European Union's General Data Protection Regulation and the 2018 California Consumer Privacy Act*, 10 U.C. IRVINE L. REV. 1455, 1477 (2020) ("In contrast, the CCPA veers away from the GDPR opt-in regime by providing an opt-out regime."); *see also* CAL. CIV. CODE § 1798.120(a) (West 2020) ("A consumer shall have the right, at any time, to direct a business that sells personal information about the consumer to third parties not to sell the consumer's personal information. This right may be referred to as the right to opt-out.").

[136] *See* Park, *supra* note 135, at 1473–74 ("The opt-in regime is the act of requiring online commercial actors to receive an individual's 'express, affirmative and informed consent' before engaging in data processing.") (quoting Joseph A. Tomain, *Online Privacy and the First Amendment: An Opt-In Approach to Data Processing*, 83 U. CIN. L. REV. 1, 3 (2014)).

[137] *Id.* at 1474 ("The opt-out rule 'plac[es] the burden on the individual to prevent certain types of information from being shared.'") (quoting Julia Palermo, Comment, *You Say "Tomato," I Say "Tomahto:" Getting Past the Opt-In v. Opt-Out Consent Debate Between the European Union and United States*, 9 GEO. MASON J. INT'L COM. L. 121, 121 (2017)).

[138] *See* Susser, *supra* note 122 (arguing that notice decoupled from consent protects privacy).

[139] *See, e.g.*, *How to Protect and Unprotect your Tweets*, TWITTER, https://help.twitter.com/en/safety-and-security/how-to-make-twitter-private-and-public [https://perma.cc/TSX2-TWC9] (describing Twitter privacy settings).

[140] Outside the social media context, however, this self-help remedy may not exist.

[141] *See* Fred H. Cate, *The Failure of Fair Information Practice Principles*, in CONSUMER PROTECTION IN THE AGE OF THE 'INFORMATION ECONOMY' 341, 359 (Jane K. Winn ed., 2006) ("[N]otices may never be received. In fact, most requests for consumer consent never reach the eyes or ears of their intended recipient."); Saima Salim, *Do You Have any Old Social*

when users are fully informed of the dangers of scraping, affirmative, opt-in consent empowers users to distinguish between the different purposes behind publicly sharing information.[142] While a user may want to spread her photos as widely as possible, she may also want to avoid being a part of a facial recognition system. Similarly, a Twitter user may consent to a COVID-19 research study but may decline to participate in a mental health study. Overall, affirmative consent preserves users' autonomy interests and allows users to decline participation in data processing activities they perceive to be harmful.[143] If anything, opt-out consent may undermine free speech more than opt-in consent.[144] Consider a user who has received notice from multiple scrapers or has a general brooding sense that her information has been scraped. Under an opt-out framework, because the user has the burden of opting out, the user may simply choose the ultimate opt-out — changing her privacy setting and denying all scrapers access to her information — instead of expending the effort to individually opt out from only the scrapers she deems harmful.[145] In the opt-in framework, the user can feel relatively safe knowing that she controls which scrapers have taken her information and will leave the nuclear option for another day. Perhaps the law should encourage users to change their privacy settings. But incentivizing this behavior would ultimately limit all scraping, including beneficial scraping.[146]

The most significant criticism of opt-in consent is that it places undue burdens on both the data collector and the user.[147] But in the scraping context, the burdens are relatively minimal. The choice presented to the user is simple: to allow or not to allow third-party scraping. The scraper also faces minimal burdens because it can easily contact users through online communication. However, there is the possibility that scrapers will inundate users with notice messages and opt-in requests.

Media Accounts that Are No Longer in Your Use?, DIGITAL INFO. WORLD (Feb. 21, 2020), https://www.digitalinformationworld.com/2020/02/tech-hoarding-surprising-statistics-and-serious-consequences.html [https://perma.cc/B42X-HZXL] (estimating 60% of all social media accounts are inactive).

[142] *See* GDPR, *supra* note 69, at Recital 32 ("When the processing has multiple purposes, consent should be given for all of them.").

[143] *See* Joseph A. Tomain, *Online Privacy and the First Amendment: An Opt-In Approach to Data Processing*, 83 U. CIN. L. REV. 1, 16 (2014) ("An opt-in requirement directly supports the autonomy interest because the individual retains the right to choose to participate in data processing.").

[144] *See id.* at 64–70 (arguing that First Amendment has required opt-in consent in labor union cases).

[145] *See* Smith, *supra* note 121.

[146] For example, changing a LinkedIn profile to private prevents Google from web indexing it. *See Off-LinkedIn Visibility*, LINKEDIN, https://www.linkedin.com/help/linkedin/answer/79854 [https://perma.cc/88ZA-87UA]; *see LinkedIn Public Profile Visibility*, LINKEDIN, https://www.linkedin.com/help/linkedin/answer/83 [https://perma.cc/Z4PT-GHFP].

[147] Park, *supra* note 135, at 1474; *see* Tomain, *supra* note 143, at 56.

Many social media websites restrict direct messaging.[148] Allowing scrapers to send notice messages and opt-in requests to users may open the door for more pernicious messaging bots to contact users. To solve these issues, websites can design methods to facilitate scraper-to-user communication.

Admittedly, opt-in will limit more scraping than will opt-out, but "an opt-in requirement is not a complete ban on data processing."[149] Some users may consent to scraping, especially if the scraper provides incentives in return.[150] Ultimately, an opt-in framework (like BIPA and GDPR) is best suited for protecting users' privacy because it gives users the most control over their personal information.[151]

C. A Nuanced Role for Websites

Data privacy regulations should also give websites a role in safeguarding user privacy. Websites are best able to monitor scraping by inspecting web traffic. As in the case of Clearview, if a third-party scraper surreptitiously scrapes and neglects to provide notice to affected users, it would be extremely difficult for users (and agencies) to learn of scraping. Indeed, lack of notice is the current norm: only Illinois residents (under BIPA) have the right to pre-collection notice.[152] A failure to give websites a role in regulating scraping exposes users to surreptitious scraping.

One possible solution to the surreptitious scraping problem is to strengthen data scraping causes of action like the CFAA and contract law, thereby giving websites the power to vindicate their users' privacy interests. Websites are more capable than individual users in challenging third-party scrapers as they have the legal and technical expertise. A website-primacy approach would also channel scraping into website APIs, which can be powerful tools for regulating scrapers and protecting user privacy. However, it may not be advisable to allow websites to maintain monopolistic control over scraping.[153] Privacy is also a personal construct, and it would be odd to give an outsider control over one's *personal* information. And, while websites may be "information

[148] Hunter, *supra* note 133.

[149] Tomain, *supra* note 143, at 56.

[150] *See* CAL. CIV. CODE § 1798.125(b)(1) (West 2020) (allowing a data collector to offer data subjects financial incentives).

[151] Park, *supra* note 135, at 1476 ("[T]he growing consensus . . . is that the opt-in model better protects consumers over opt-out model . . . ."); *see* Tomain, *supra* note 143, at 24–26 (arguing that opt-out is ineffective).

[152] *Compare* 740 ILL. COMP. STAT. 14/15(b)(1)-(2) (2008) (requiring pre-collection notice), *with* CAL. CODE REGS. tit. 11, § 999.305(d) (2020) (exempting scrapers from pre-collection notice requirement).

[153] *See* hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 998 (9th Cir. 2019) (discussing risk of unfair competition risk in allowing companies to selectively ban only certain scrapers).

fiduciaries" in theory, it is arguable whether they are loyal to their users in reality.[154]

An alternative to a website-primacy approach is to work within the framework of data privacy laws like the CCPA, BIPA, and GDPR. These data privacy laws vest power in the user instead of the website: the user controls how her personal information is collected and used. One approach could be to impose a duty on websites to monitor for unauthorized scraping and to inform users of such scraping. Current data privacy laws can easily accommodate this proposal. Laws like the CCPA require data collectors to notify users of data breaches and allow private lawsuits for such breaches.[155] Unauthorized scraping that infringes user privacy rights could be characterized as a security breach, triggering security breach notification requirements.[156] This monitoring duty serves the same knowledge-filling function as requiring scrapers to provide notice to users. This monitoring and notification requirement can also help distinguish between websites' anticompetitive and user-centric motives. For example, in ***hiQ v. LinkedIn***, LinkedIn claimed that it cared about user privacy.[157] If LinkedIn truly cared about user privacy, LinkedIn could have notified users about hiQ's scraping.

New regulations can also require websites to provide clear warnings that anyone, including harmful scrapers, can access public posts.[158] Regulations may also require websites to set default privacy settings to "private" instead of "public."[159] Indeed, some companies have already begun implementing such approaches on their own initiative.[160] Market demand for greater privacy may further drive websites

[154] *See* Sobel, *supra* note 57, at 183–87.

[155] CAL. CIV. CODE §§ 1798.82(d)(1)(D), 1798.150(b) (West 2019).

[156] *But see* Sobel, *supra* note 57, at 183–87.

[157] Petition for Writ of Certiorari at 27–32, LinkedIn Corp. v. hiQ Labs, Inc., No. 19-1116 (U.S. Mar. 9, 2020).

[158] For example, LinkedIn provides users great flexibility in determining what information is public. But this information is buried among the various other profile settings. *See Off-LinkedIn Visibility*, LINKEDIN, https://www.linkedin.com/help/linkedin/answer/79854?trk=psettings-data-sharing_api&lang=en; [https://perma.cc/A638-G87N]; *LinkedIn Public Profile Visibility*, LINKEDIN, https://www.linkedin.com/help/linkedin/answer/83 [https://perma.cc/GW9F-TTFE].

[159] *Cf. Instagram Help Center: Privacy Settings and Information*, INSTAGRAM, https://www.facebook.com/help/instagram/196883487377501 [https://perma.cc/7K7W-7KAZ]; *About Public and Protected Tweets*, TWITTER, https://help.twitter.com/en/safety-and-security/public-and-protected-tweets [https://perma.cc/G2PC-EETU] ("When you sign up for Twitter, your Tweets are public by default.").

[160] Sarah Perez, *TikTok Update will Change Privacy Settings and Defaults for Users Under 18*, TECHCRUNCH (Jan. 13, 2021, 5:00 PM), https://techcrunch.com/2021/01/13/tiktok-update-will-change-privacy-settings-and-defaults-for-users-under-18/ [https://perma.cc/8R3C-2Q8Q] (explaining TikTok's recent change to default privacy settings for minors).

to adopt such technical measures.[161] And if scrapers are subject to notice-and-consent requirements, websites can design user- (and scraper-) friendly ways to mediate this notice-and-consent exchange. Overall, websites can play an invaluable role in helping inform users of the dangers of posting publicly, thereby giving users informed control over how they share their information.

Finally, an interesting edge case to consider is when the website authorizes scraping. Many websites allow scrapers to use APIs subject to certain terms and conditions. Twitter, for example, tells users:

> By publicly posting content when you Tweet, you are directing us to disclose that information as broadly as possible, including through our APIs, and directing those accessing the information through our APIs to do the same . . . . We have standard terms that govern how this data can be used, and a compliance program to enforce these terms. But these individuals and companies are not affiliated with Twitter, and their offerings may not reflect updates you make on Twitter.[162]

The scope of authorized scraping varies among websites and depends on the terms of use and the technical configuration of the APIs. For example, when a scraper requests public LinkedIn user profile information, the LinkedIn API automatically sends a notice-and-consent message to the affected user.[163] In contrast, Twitter's API allows developers to access public tweets with or without user consent.[164] But while Twitter's API does not require scrapers to implement notice-and-consent, Twitter's API terms of use require a scraper to "apply for a developer account" and Twitter "review[s] all proposed uses" of its API.[165] Twitter's developer terms also unequivocally prohibit certain uses of its API, including scraping for facial recognition.[166] Many websites also expressly authorize web crawling by search engines (e.g.,

[161] *See* Daniel Burrus, *The Privacy Revolt: The Growing Demand for Privacy-as-a-Service*, WIRED (Mar. 2015), https://www.wired.com/insights/2015/03/privacy-revolt-growing-demand-privacy-service/ [https://perma.cc/4CDX-FMYA].

[162] *Twitter Privacy Policy*, TWITTER, *supra* note 63. There is an argument to be made that Twitter's policy of allowing access to public tweets violates user's expectation of privacy. But that privacy harm can be traced to the website/user privacy contract rather than the scraper.

[163] *Authorization Code Flow (3-legged OAuth)*, MICROSOFT, *supra* note 84 (explaining how users must consent to permission requests).

[164] *Tweets and Users v2*, TWITTER, https://developer.twitter.com/en/docs/labs/tweets-and-users/quick-start/get-tweets [https://perma.cc/A3PC-974Z] (explaining how to retrieve public tweets); *Authentication*, TWITTER, https://developer.twitter.com/en/docs/authentication/oauth-2-0/application-only [https://perma.cc/JFJ5-4P24] (explaining how authentication does not require user consent).

[165] *Developer Policy*, TWITTER, *supra* note 83.

[166] *More about Restricted Uses of the Twitter APIs*, TWITTER, *supra* note 42.

Google search). Twitter, for example, tells users that "crawling [Twitter] is permissible if done in accordance with the provisions of the robots.txt file [which allows search engine web crawlers like Google]."[167]

In website-authorized scraping, the user's privacy interests are lower because obscurity-disrupting technologies are expressly authorized and because the website has not willingly obliged itself to a standard of care. Assuming that users read and understand websites' terms of service, users may not have a reasonable expectation of privacy with respect to website-authorized scraping.

V. The First Amendment Objection to Regulating Public Personal Information

Finally, scrapers like Clearview have asserted First Amendment rights to access public information. Clearview has argued that a notice-and-consent requirement violates the First Amendment by unduly burdening its right to use public information.[168] This Section analyzes the First Amendment objection to regulating public personal information and argues that a strict notice-and-consent regulation withstands First Amendment scrutiny because it does not unduly burden access to public information.

Within First Amendment doctrine, different levels of scrutiny apply to different kinds of speech regulations.[169] It is arguable whether BIPA regulates Clearview's "speech" or merely the act of collecting data. If BIPA were a direct regulation on the content of Clearview's speech, strict scrutiny — the most exacting review — would apply.[170] Strict scrutiny, however, does not automatically invalidate statutes.[171] A statute "narrowly tailored to serve compelling state interests" may survive strict scrutiny.[172]

Just because information is publicly accessible does not mean the government can never regulate use. For example, intellectual property

[167] *Twitter Terms of Service*, TWITTER, *supra* note 63; *Terms of Service*, YOUTUBE, https://www.youtube.com/static?template=terms [https://perma.cc/9UDC-VLRJ] ("You are not allowed to . . . access the Service using any automated means (such as robots, botnets or scrapers) except (a) in the case of public search engines, in accordance with YouTube's robots.txt file."). *See, e.g.*, Twitter's robot.txt file, TWITTER, https://twitter.com/robots.txt [https://perma.cc/H2DU-FX74].

[168] Defendant's Memorandum in Support of Motion to Dismiss, ACLU v. Clearview AI, *supra* note 3, at 16–23 (challenging BIPA's notice-and-consent regulation on First Amendment grounds).

[169] *See* KATHLEEN ANN RUANE, CONG. RSCH. SERV., 95-815, FREEDOM OF SPEECH AND PRESS: EXCEPTIONS TO THE FIRST AMENDMENT (explaining levels of scrutiny).

[170] Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 340 (2010).

[171] *See* Adam Winkler, *Fatal in Theory and Strict in Fact: An Empirical Analysis of Strict Scrutiny in the Federal Courts*, 59 VAND. L. REV. 793, 845 (2006) (finding that 22% of free speech restrictions survive strict scrutiny).

[172] Reed v. Town of Gilbert, 576 U.S. 155, 163 (2015).

law doctrines — such as copyright, trademark, and the right of publicity — restrict the use of otherwise public information and routinely survive First Amendment challenges. In addition, the Court has upheld buffer zone laws, which restrict the types of speech outside public abortion centers.[173] Ultimately, notice-and-consent laws may survive First Amendment scrutiny.

A. Is Scraping First Amendment Protected "Speech"?

Data privacy proponents have argued that data privacy laws restricting scraping are subject to deferential intermediate scrutiny because they are content-neutral regulations of expressive conduct.[174] In other words, data privacy laws merely regulate the act of data collection. The canonical case for this proposition is *United States v. O'Brien*.[175] There, the Court upheld a criminal prohibition against burning draft cards even when the defendant argued that the burning was a political protest against the Vietnam War. The *O'Brien* standard undergirds a whole host of privacy regulations, from trespass law to prohibitions on eavesdropping.[176]

However, there are arguments that data privacy laws infringe on core First Amendment rights.[177] For example, the Court has announced — albeit in dicta — that "the creation and dissemination of information are speech within the meaning of the First Amendment."[178] Similarly, there is a growing body of caselaw on the "right to record."[179] Most of these cases have come in the context of recording police activity in public, but one court found that an animal rights group had the right to record deer culling in a public state park.[180]

Overall, the scope of the First Amendment's protections is unsettled. Already, one state trial court has questioned Clearview's argument that it is within the ambit of core First Amendment protection.[181] The

[173] Hill v. Colorado, 530 U.S. 703 (2000).

[174] Plaintiff's Response to Defendant's Motion to Dismiss at 14–23, ACLU v. Clearview AI, No. 2020 CH 04353 (Ill. Cir. Ct. Nov. 2, 2020), https://www.aclu.org/plaintiffs-response-defendants-motion-dismiss [https://perma.cc/V6Y5-MA8Z].

[175] United States v. O'Brien, 391 U.S. 367, 376–77 (1968).

[176] *See* Neil M. Richards, *Reconciling Data Privacy and the First Amendment*, 52 UCLA L. REV. 1149, 1189–94 (2005).

[177] *See* Jane Bambauer, *Is Data Speech?*, 66 STAN. L. REV. 57, 59–60 (2014); Jacquellena Carrero, *Access Granted: A First Amendment Theory of Reform of the CFAA Access Provision*, 120 COLUM. L. REV. 131, 150–58 (2020).

[178] Sorrell v. IMS Health Inc., 564 U.S. 552, 570 (2011).

[179] Bambauer, *supra* note 177, at 84–86.

[180] *Id.* at 85 (citing S.H.A.R.K. v. Metro Parks Serving Summit Cnty., 499 F.3d 553, 557–58 (6th Cir. 2007)).

[181] Vermont v. Clearview AI, Inc., No. 226-3-20 Cncv, at 9–18 (Vt. Super. Ct. Sep. 10, 2020), https://ago.vermont.gov/wp-content/uploads/2020/09/Clearview-Motion-to-Dismiss-Decision.pdf [https://perma.cc/3UP6-UKLW].

court expressed doubt over whether Clearview's scraping was communicative speech.[182] However, the court declined to definitively rule on whether scraping was speech, instead resting its decision on the fact that the regulations at issue were content-neutral because the government's regulatory purposes were independent of the content of Clearview's "speech."[183] Because the regulations were content-neutral, the court applied intermediate scrutiny, ultimately finding the data privacy regulations to be constitutional.[184]

B. Does Pre-Collection Notice Unduly Limit Speech (Scraping)?

Whether or not scraping is core First Amendment speech, the fundamental First Amendment question is whether a court will find data privacy regulations to unduly burden speech.[185] Requiring pre-collection notice poses only minimal burdens on scraping.

First, pre-collection notice poses very few burdens on social media scrapers because notice is very easy to provide. The data subjects are clearly identifiable, and scrapers can easily communicate with users via online channels.[186] Notice is also invaluable because scrapers can operate surreptitiously. Without notice, users are often unaware of third-party scraping and thus unable to exercise statutory data privacy rights and self-help remedies. The tremendous privacy interests served by requiring pre-collection notice far outweigh the minimal free speech burdens.

Taking this argument to the extreme, however, is the case of web crawlers — like Google search — which scrape and index social media profiles. For example, you can Google search for public LinkedIn profiles. Search engines may argue that they are not collecting "personal data" but rather textual information that can be linked to any one of billions of people on Earth so that they do not need to provide notice.[187] Search engines may also argue (under the GDPR, but not BIPA) that

[182] *Id.* at 12 (citing Universal City Studios, Inc. v. Corley, 273 F.3d 429 (2d Cir. 2001)).

[183] *Id.* at 14.

[184] *Id.*

[185] Reed v. Town of Gilbert, 576 U.S. 155, 163 (2015) (describing that strict scrutiny looks at whether the speech regulation is "narrowly tailored to serve compelling state interests"); Cent. Hudson Gas & Elec. v. Pub. Serv. Comm'n, 447 U.S. 557, 566 (1980) (describing that intermediate scrutiny for commercial speech looks at "whether the asserted governmental interest is substantial . . . [and] whether the regulation directly advances the governmental interest asserted, and whether it is not more extensive than is necessary to serve that interest"); United States v. O'Brien, 391 U.S. 367, 377 (1968) (holding that intermediate scrutiny looks at whether the regulation "furthers an important or substantial governmental interest . . . and if the incidental restriction on alleged First Amendment freedoms is no greater than is essential to the furtherance of that interest").

[186] *Guidelines on Transparency under Regulation 2016/679*, *supra* note 130, at 29–31 (presenting examples where notice is impossible or impracticable to provide).

[187] GDPR, *supra* note 69, at art. 4(1) (defining personal data).

providing notice is impracticable given that the crawlers index the entire Internet.[188] Because many websites expressly allow search engine web crawling in their terms of service, web crawling may also be the edge case of website-authorized scraping.[189] But because some users may be unaware they can be searched and because notice is relatively easy to provide, search engines should provide users notice of their data collection practices.

C. Does Opt-In Consent Unduly Burden Speech (Scraping)?

To be narrowly tailored, a regulation must be "the least restrictive means of achieving a compelling state interest."[190] There are several alternatives to opt-in consent — not requiring consent at all and opt-out consent — but opt-in consent is the least restrictive means of protecting user privacy.

First, not requiring consent at all and solely relying on notice does not practically achieve anything because the technological default rule is opt-out consent.[191] Next, the question is whether opt-out consent is a sufficient alternative that would render opt-in consent unconstitutional. In an ideal Coasean world, opt-in and opt-out consent both lead to the identical result: the initial assignment of control over collection does not matter.[192] If the scraper has initial control (opt-out regime), the scraper can pay users not to opt out, and only those users who value privacy highly enough will exercise their opt-out rights. If the user has initial control (opt-in), the scraper can pay users to opt in, and only those users who value privacy highly will decline to opt in. But transaction costs and information asymmetries cause the real world to diverge from the Coasean world.[193]

Under the opt-in regime, scrapers have transaction costs in the form of seeking opt-in consent. But just like with providing notice, the transaction costs of asking for opt-in consent are minimal because users are easily identifiable and online communication is easy. Opt-in consent can also burden scraping because default options are "sticky."[194] The

[188] This argument likely fails under UODO's interpretation of GDPR impracticability. *See The First Fine*, UODO, *supra* note 109.

[189] *See* discussion *supra* Section IV.C.

[190] McCullen v. Coakley, 573 U.S. 464, 478 (2014).

[191] *See, e.g.*, *How to Protect and Unprotect your Tweets*, TWITTER, *supra* note 139.

[192] Jeffrey M. Lacker, *The Economics of Financial Privacy: to Opt out or to Opt in?*, 88 FED. RSRV. BANK RICHMOND ECON. 1, 9–10 (2002); *see also* Joshua A. Decker, *Markets in Everything and Another View of the Cathedral: Religious Freedom and Coasian Bargaining*, 26 STAN. L. & POL'Y REV. 485, 489 (2015) (applying Coasean theory to religious freedom).

[193] Alan McQuinn, *The Economics of "Opt-Out" Versus "Opt-In" Privacy Rules*, INNOVATION FILES (Oct. 6, 2017), https://itif.org/publications/2017/10/06/economics-opt-out-versus-opt-in-privacy-rules [https://perma.cc/BD2Y-VFB2]; Lacker, *supra* note 192.

[194] *See* Omri Ben-Shahar & John A. E. Pottow, *On the Stickiness of Default Rules*, 33 FL. ST. L. REV. 651, 653–54 (2006).

scraper may have to pay a premium to overcome the "stickiness" of the default rule.[195] As such, the opt-in incentivization price may be higher than the opt-out price. In some cases, perhaps, no incentivization is sufficient to coax users to opt in. But those cases are more of an indictment of the scrapers, the nature of the information they collect, and the way they intend to use it than the opt-in regime itself.

While scrapers can offer incentives to encourage users to opt in, some scrapers may find this financially burdensome. For example, poorly funded, yet extremely vital, research studies may be stymied by an opt-in regime.[196] But there is a readily available alternative: scraping can be performed in ways that do not implicate privacy interests. For example, if the scraper does not scrape personally identifiable information, privacy interests and data privacy laws are not triggered.[197] Admittedly, there may be rare cases where socially beneficial scrapers cannot incentivize users to opt in and where scrapers have no alternative means of scraping, but that may be the sacrifice society has to make for privacy.

Given the Coasean dynamics, opt-in and opt-out consent present nearly identical burdens. Instead of paying to incentivize users to opt in (opt-in regime), a scraper would have to pay to incentivize users to *not opt out* (opt-out regime). However, the incentivization price in the opt-in regime may be higher due to the stickiness of default rules. The most significant difference between opt-in and opt-out consent is that opt-in consent ensures the effectiveness of notice. In the opt-out regime, a failure to opt out is likely to be the result of a failure to receive notice, which is a significant problem in the social media context given the high percentage of inactive accounts.[198] As such, when scrapers complain about the difficulties of opt-in consent and praise the virtues of opt-out consent, scrapers exploit the problem of ineffective notice. Ultimately, opt-in consent is narrowly tailored because it imposes similar burdens as opt-out consent and because it meaningfully advances compelling privacy interests.

VI. CONCLUSION

Social media websites are amazing repositories of ideas and gathering places for people. Every day, hundreds of millions of people use

[195] *See* Melissa W. Bailey, Note, *Seduction by Technology: Why Consumers Opt Out of Privacy by Buying into the Internet of Things*, 94 TEX. L. REV. 1023, 1047 (2016) (discussing incentivization systems).

[196] For one such study, see Gautam Kishore Shahi et. al., *An Exploratory Study of COVID-19 Misinformation on Twitter*, https://arxiv.org/pdf/2005.05710.pdf [https://perma.cc/JKN8-KSBB] (unpublished manuscript) (submitted to Elsevier) (studying misinformation on Twitter).

[197] *See, e.g.*, GDPR, *supra* note 69, art. 4(1) (defining personal data).

[198] Salim, *supra* note 141 (estimating 60% of social media accounts are inactive).

social media websites to post and share information. However, as Clearview's practices have shown, there are concerns with giving third-party scrapers unfettered access to the information hosted on these public websites.

At first blush, privacy is at odds with the concept of publicly available information. This Note, however, provides several theoretical bases for recognizing privacy interests in public information. This Note explains how unauthorized use of public information — just like private information — leads to unwarranted privacy harms. This Note also explains how privacy in public exists because of users' expectations of obscurity and because of users' trust in websites and other users to abide by contractual prohibitions on scraping.

Current regulatory mechanisms are varied, as lawmakers struggle to define the optimal scheme in a complex ecosystem populated with multiple parties. This Note has argued for several proposals to reform existing regulations, including a strict notice-and-consent regime and an obligation on websites to monitor scraping. In particular, the strict notice-and-consent regime and the website monitoring requirement address the notice loophole left by current laws — the ability of scrapers to operate surreptitiously — which Clearview exploited. Ultimately, these proposals would help improve privacy and ensure the continued vitality of the Internet.