SCRAPING PHOTOGRAPHS

Maggie King

SCRAPING PHOTOGRAPHS

Maggie King

INTRODUCTION

In January 2020, a recurrent debate around modern surveillance techniques was reignited, after the New York Times reported that a little-known start up Clearview AI was selling facial recognition technology to police departments.[1] The software enables law enforcement to identify with high accuracy a suspect caught on camera within minutes of running the image through Clearview's system.[2] Lawmakers and at least forty different interest groups called for a ban or moratorium on facial recognition technology, citing everything from the pitfalls, to the uncertainties and general big brother-like nature of the software.[3]

But another debate also ensued in response to the news. *Wired* published an article arguing that Clearview had "abused" the laws intended to enable a free and open internet when it built its facial recognition tool off scraping photos from social media.[4] Technology companies took a different stance in this debate. Facebook, Twitter, and Google all immediately sent cease-and-desist letters to Clearview to stop scraping their users' photographs. Those actions suggested that in scraping photographs, Clearview had not so much abused laws but rather broken them.[5]

These responses highlight an important, and unresolved legal question: Is scraping photographs legal? To build its software, Clearview AI claims to have scraped more than 3 billion photos from the internet, including from popular social media platforms like Facebook, Instagram,

[1] Kashmir Hill, *The Secretive Company that Might End Privacy as We Know It*, N.Y. TIMES (Jan. 18, 2020), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html?smid=nytcore-ios-share&fbclid=IwAR3tu5Zy1gMSAuVaF2jSmRGr6Hp7iqaxhKU6qBXnf1MActx13XL9KNduuEw.

[2] *Id.*

[3] Chris Mills Rodrigo, *Government Privacy Watchdog Under Pressure to Recommend Facial Recognition Ban*, HILL (Jan. 27, 2020, 4:31 PM), https://thehill.com/policy/technology/480152-government-privacy-watchdog-under-pressure-to-recommend-facial-recognition ("Forty groups, led by the Electronic Privacy Information Center, sent a letter Monday to the agency calling for the suspension of facial recognition systems 'pending further review.'"). For an example of a proposed moratorium on the technology at the state level, see H.B. 2856, 66th Leg., Reg. Sess. (Wash. 2019-2020). *See*, U.S. DEP'T OF COM., FACE RECOGNITION VENDOR TEST (FRTV) PART 3: DEMOGRAPHIC EFFECTS (2019) (For a recent, leading report on the pitfalls and uncertainties of facial recognition technology, NIST produced a report

[4] Louise Matsakis, *Scraping the Web Is a Powerful Tool. Clearview AI Abused it,* WIRED (Jan. 25, 2020, 7:00 AM), https://www.wired.com/story/clearview-ai-scraping-web/.

[5] Alfred Ng & Steven Musil, *Clearview AI Hit With Cease-and-Desist From Google, Facebook Over Facial Recognition Collection,* CNET (Feb. 5, 2020, 6:10 PM), https://www.cnet.com/news/clearview-ai-hit-with-cease-and-desist-from-google-over-facial-recognition-collection.

Twitter and YouTube.[6] Companies seeking to halt scraping activities have historically relied heavily on anti-hacking laws, and analogized scraping to hacking. The US federal government and all fifty states have each enacted statutes that protect against hacking, or "unauthorized access to computers."[7] The most recent caselaw suggests run-of-the-mill data scraping does not per se violate federal or state hacking laws as long as the data scraped is public or authorized for access to the entity doing the scraping. However, "[t]he wide variety of outcomes in scraping-related litigation demonstrates that courts are uncertain of what exactly constitutes computer hacking,"[8] much less whether current interpretations of the laws would apply equally to photographs. Beyond hacking laws, those seeking to enjoin scraping activities have brought claims that scraping violates trespass to chattels, contracts, and federal copyright laws. In fact, in response to Clearview AI's alleged scraping of photographs, a Facebook attorney said in a since-deleted tweet that Clearview was "not only violating terms of sites, but also committing copyright infringement by using people's photos this way."[9] At present, no major scraping cases have addressed the scraping of photographs on any of these claims. But scraping photographs does interact with copyright and other existing laws in ways that differ from traditional scraping cases. In short, while the legality of scraping data generally is trending towards legal, with some uncertainty remaining, the issue of scraping photographs in particular is not settled. Rather, the legal debate around scraping photographs is only just emerging.

This note explores whether any existing laws prohibit scraping photographs, as suggested by Facebook and other big tech companies' recent actions against Clearview. After examining each potential claim, this note argues that no existing law should be construed to hold Clearview liable for scraping photographs, because doing so would create inconsistencies in existing law. But also, the apparent legality of Clearview's scraping activity presents an argument for a reversal of the recent trend towards laws that, guided by the principle of a free and open internet, favor scraping. Rather, the apparent legality of activity that ultimately enables otherwise unrestrained modern surveillance techniques presents an argument for a return to federal laws that provide stronger defenses for cyberproperty. Part I explains the concept and technology of scraping photographs and the extraction of facial scan data for use in facial recognition algorithms. Part II clarifies the state of the law on scraping under the claims that are most commonly brought against scraping activities, including the Computer Fraud and Abuse Act (CFAA), copyright, and contract law. Part III applies these doctrines to scraping photographs and extracting facial images. Part IV provides policy arguments for the direction

[6] Donie O'Sullivan, *This Man Says He's Stockpiling Billions of Our Photos,* CNN BUSINESS (Feb. 10, 2020, 9:18 AM), https://www.cnn.com/2020/02/10/tech/clearview-ai-ceo-hoan-ton-that/index.html.

[7] Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes*, 78 N.Y.U. L. REV. 1596 (2003).

[8] Jacquellena Carrero, *Access Granted: A First Amendment Theory of Reform of the CFAA Access Provision*, 120 COLUM. L. REV. 131, 150 (2020).

[9] Kevin Keller (@kevinkeller), TWITTER (Jan. 18, 2020), https://twitter.com/kevinkeller/status/1218622954168152064.

scraping law should take in light of this note's findings regarding the state of the law today. Part V briefly concludes.

I. TECHNICAL BACKGROUND ON SCRAPING

In May 2019, long before the Clearview news broke, the Electronic Frontier Foundation (EFF) had already produced a report stating that "[f]ace recognition is poised to become one of the most pervasive surveillance technologies."[10] On the other hand, proponents of the technology argue that it improves existing public safety and criminal justice measures, which are prone to bias and error under purely human decision making.[11] Clearview's website touts that the software "helps to identify child molesters, murderers, suspected terrorists," and that it helps to exonerate the innocent, identify victims of child sexual abuse and other crimes, and "avoid eyewitness lineups that are prone to human error."[12] In short, public opinion towards facial recognition technology is under vibrant and live debate.

Regardless of where one falls in that debate, it is certainly true that "the development and the uses of this technology are growing faster than the laws defining its proper use."[13] At present, "[t]here is no clear line in the sand about how [facial recognition technology] should be used. There is no federal law that tells facial recognition users that they can go only so far and no further."[14] While state and local laws banning facial recognition are beginning to crop up in select locales, ultimately, a more direct route to regulating Clearview is to study the source of its feedstock- ubiquitous photographs of people online, available for free to anyone who knows how to code a decent "scraper." Without an abundant set of photographs, facial recognition algorithms have no data on which to build their models.

A. How to Scrape a Photograph

At their core, web scrapers are simply a more efficient means of online information collection. Scraping can be performed on any type of data displayed in a web browser. Traditionally, it has been performed on numerical data–such as price estimates listed on Zillow for each house in the country or baseball player game statistics. Scraping photographs is a

[10] Jennifer Lynch, *Face Off: Law Enforcement Use of Facial Recognition Technology,* EFF (May 2019), https://www.eff.org/files/2019/05/28/face-off-report.pdf.

[11] Craig McCarthy, *Facial recognition leads cops to alleged rapist in under 24 hours,* N.Y. POST (Aug. 5, 2019, 6:03 PM), https://nypost.com/2019/08/05/facial-recognition-leads-cops-to-alleged-rapist-in-under-24-hours/.

[12] CLEARVIEW AI, https://clearview.ai/ (last visited Jan. 29, 2021).

[13] Leo Briceno, *Should Government Be Able to Track Your Every Move Outside Your House For The Rest Of Your Life?*, THE FEDERALIST (Mar. 6, 2020), https://thefederalist.com/2020/03/06/should-government-be-able-to-track-your-every-move-outside-your-house-for-the-rest-of-your-life.

[14] *Id.*

relatively newer form of scraping, but the mechanics of scraping data or photographs of any type are largely identical.[15]

Scraping occurs when a software program is used to "electronically copy, retrieve or otherwise acquire data and information from the websites of others with little or no human interaction."[16] Manually, a user can perform the equivalent of "scraping" one image from the web by (1) accessing the webpage with the image of interest, (2) right-clicking on the image, and (3) clicking "Save Image as." The process of scraping is the same, but it is automated in code to (1) automatically open a web connection to the hosted image site, and (2) make an HTTP request for the image (such as a "Get" request), specifying the image based on its own URL. The HTTP request effectively sends the bytes that make up the image to the programmer's computer or other designated location. It is worth noting that the "Get" request transmission that occurs in scraping is technically identical to the process that occurs when a person uses a website and makes a request to physically view an image. Ready-made code is free and open to programmers to use to implement these procedures on any website that the programmer may access manually.[17] For example, one such set of code essentially operates so that it "pretends to be a real user, it opens the browser, moves the cursor around and clicks buttons if you tell it to do so."[18]

In order to be useful in a machine learning algorithm, such as facial recognition or like software, scraped photographs must be converted to "face scans."[19] Face scans are a different digital format of a photograph-the format required to be used as the data read in a facial recognition algorithm.[20] Face scans are created after a photo is uploaded to a site such as Facebook. Facebook's "technology scans the photo and detects whether it contains images of faces. If so, the technology extracts the various geometric data points that make a face unique, such as the distance

[15] *See, e.g.*, Martin Perez, *How to Scrape and Download Images from any Website,* PARSEHUB (Aug. 22, 2019), https://www.parsehub.com/blog/scrape-images-website/.

[16] CouponCabin LLC v. Savings.com, Inc., No. 2:14-CV-39-TLS, 2017 WL 83337, at *2 (N.D. Ind. Jan. 10, 2017).

[17] For two examples, examine Python's Beautiful Soup and Selenium libraries. Online tutorials are also free and easy to find, each of which usually provides a step by step process for programmers to learn how to use each particular library. *See, e.g.*, Fabian Bosler, *Image Scraping with Python,* MEDIUM (Sept. 27, 2019), https://towardsdatascience.com/image-scraping-with-python-a96feda8af2d.

[18] Bosler, *supra* note 17.

[19] These are also known as face templates and/or facial scans. The language varies by court opinion, among the 6-7 major cases discussing the issue. Most of these opinions surface from litigation concerning the Illinois state biometric information privacy law, BIPA, which restricts the use of "face scans" but not the use of photographs. Thus, the distinction is heavily discussed in leading cases on the topic of whether the conversion and storing of a photograph into a face scan constitutes a violation of BIPA. *See e.g.*, *Patel*, *infra* note 21; *Rivera*, *infra* note 22; *McGinnis*, *infra* note 20.

[20] While this distinction is not significant in the software development world, it has become a point of distinction in recent court cases. *See, e.g.*, McGinnis v. United States Cold Storage, Inc., 382 F. Supp. 3d 813, 819 (N.D. Ill. 2019) (referred to face templates as "created from photographs of plaintiffs' faces," and cited prior case *Google v. Rivera* in its holding on this distinction).

between the eyes, nose, and ears, to create a face signature or map."[21] Then, the technology "compares the face signature to faces in Facebook's database of user face templates (i.e., face signatures that have already been matched to the user's profiles)" in search of a match.[22] In other words, a face scan effectively functions as a complex digital template of an individual's face. A computer can then use these face scans to generate images of that individual or identify them automatically in other images.[23] The latter technology is better known as facial recognition.

B. Why Scrape a Photograph?

Scrapers in general all have "broad appeal due to their speed."[24] As compared to manual data collection of information posted online, "[t]hey can retrieve several pages on a server simultaneously and access target websites automatically thousands of times per day."[25] Although scrapers access the websites automatically, the information they access is not different from the information accessible to the person who wrote the scraper if they were to manually collect it. "[M]ost scrapers, if designed appropriately, would be highly similar to the level of access of a human browser."[26] As a result, as a general rule of thumb, "courts should raise an eyebrow at a claim that a scraper should be viewed as an invasive criminal trespasser."[27] When given the option to write a scraper, "businesses and individuals prefer using scrapers to manually collecting data" for efficiency reasons alone.[28]

What complicates the issue of scraping of any nature is its scale. At scale, scrapers can quickly and relatively cheaply copy data from any data source displayed on a site which a human can access and deposit another copy of all of that data in an easily readable format on any other person's machine. As a result, scraping is a popular but controversial

[21] Patel v. Facebook, Inc., 932 F.3d 1264, 1268 (9th Cir. 2019), cert. denied, No. 19-706, 2020 WL 283288 (U.S. Jan. 21, 2020).

[22] *Patel*, 932 F.3d, at 1268. *See also* Rivera v. Google Inc., 238 F. Supp. 3d 1088, 1091 (N.D. Ill. 2017) ("Google immediately scanned each uploaded photograph of Rivera. Id. ¶ 28. The scans located her face and zeroed in on its unique contours to create a "template" that maps and records her distinct facial measurements. Id. At the time of the automatic upload and face-scan . . .").

[23] For example, code that performs the conversion of a jpeg into a face scan, and then applies a facial recognition algorithm can be found online. *See, e.g.,* Adam Geitgey (@ageitgey), GITHUB, https://github.com/ageitgey/face_recognition/blob/master/examples/face_recognition_svm.py (one example of numerous open software programs that can be used for this purpose.)

[24] Myra F. Din, *Breaching and Entering: When Data Scraping Should Be A Federal Computer Hacking Crime*, 81 BROOK. L. REV. 405, 411 (2015) (internal quotation marks omitted.)

[25] *Id.* (internal quotation marks omitted).

[26] Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372, 414 (2018).

[27] *Id.* at 414-15.

[28] Din, *supra* note 24, at 411. See also the court's reasoning in *Sandvig v. Sessions*, as discussed in Sellars, *supra* note 26, at 411.

activity, as it can either be perceived as a form of theft, or a powerful act supporting the freedom of information.

Many legal scholars argue that "[s]craping publicly available information, regardless of a site's terms of service or a cease-and-desist letter" should not be a crime.[29] "An open and healthy internet demands it."[30] Such calls to justify scraping echo the broader policy arguments often made for a free and open internet. As Cyberlaw scholar Patricia Bellia has summarized, many of these justifications can be traced back to a policy concern over the power to control access to information. Bellia writes that "a right to control access to the physical equipment of a network translates into far broader powers–for example, the ability to block speech or to control access to and uses of information."[31] These arguments hinge on the belief that "the public has an interest in open access to information and open avenues for speech. If a company such as eBay can control access to its servers, it can also control access to the information those servers hold."[32] Additionally, Jamie Lee Williams, a staff attorney at EFF on the civil liberties team describes scraping as simply "a fundamental thing that we rely on every day, a lot of people without realizing, because it's going on behind the scenes."[33] EFF and other digital rights groups have also often argued the benefits of scraping outweigh the harms.[34] Finally, data shows that courts should always be wary of the motivation behind a scraping claim: "It may be that the platform's true motivations [for bringing suit against a web scraper] are actually anticompetitive, speech-suppressing, or otherwise untrustworthy."[35]

The reality is that "[t]here are countless uses of web scraping. Some are good. Some are bad. Some are bad for the website but should be allowed for the good of the public."[36] While policy justifications for scraping are widely debated on the back of justifications for a free and open internet, the law of scraping is coalescing in many circuits as cases on the issue are heard more frequently. Most of the direction of the law has honored the policy justifications set out in favor of a free and open internet. The next section summarizes the state of the law on scraping today.

II. THE LAW OF SCRAPING

Those seeking to enjoin scraping activities have brought claims that scraping violates federal and state hacking laws or common law trespass to chattels, contracts, and federal copyright laws. This section

[29] Jason Tashea, *Why scraping publicly available information online isn't a crime*, ABAJOURNAL (Sept. 23, 2019, 6:30 AM), https://www.abajournal.com/lawscribbler/article/scraping-a-public-website-isnt-a-crime.

[30] *Id.*

[31] Patricia L. Bellia, *Defending Cyberproperty*, 79 N.Y.U. L. REV. 2164, 2194 (2004).

[32] *Id.* at 2191.

[33] Louise Matsakis, *Scraping the Web Is a Powerful Tool. Clearview AI Abused It.*, WIRED (Jan. 25, 2020, 7:00 AM), https://www.wired.com/story/clearview-ai-scraping-web/.

[34] Matsakis, *supra* note 4.

[35] Sellars, *supra* note 26, at 415. The data shows most companies only bring scraping claims against emerging competitors. *Id.*

[36] *Id.*

details the general success of each of these claims in scraping cases today.[37] Part III then applies each of these legal theories on data scraping to the scraping of photographs, and discusses the extent to which the nature of the scraped subject being a photograph should change the equation.

A. CFAA Claims

Once the hallmark of scraping claims, the federal hacking law, the Computer Fraud and Abuse Act (CFAA) increasingly appears to be a dead end for plaintiffs seeking to successfully prohibit scraping activities on their sites. For a long time, the CFAA was "the primary legal means by which companies offering web-based services attempt[ed] to block scraping of their applications."[38] But the most recent caselaw supports more limited protections for plaintiffs. The current law generally allows scraping where the access would have been granted to the human user, excusing from CFAA liability a very broad set of scraping activities today.[39]

1. CFAA Background: Ambiguous Statutory Language. – The CFAA is a federal law that "contains several provisions outlawing unauthorized access to computer systems."[40] The law was initially passed to target hacking activities.[41] The statute has gone through multiple amendment processes since its initial passage in 1984, and in turn "is now quite broad" in its statutory language that could apply to more than traditional hacking activities.[42] The current language prohibits "knowingly access[ing] a computer without authorization or exceed[ing] authorized access" and obtaining certain protected information from that conduct.[43] In other words, the current CFAA defines two different ways in which one may access a computer as "unauthorized": (1) "'access[ing]' a computer 'without authorization'" and (2) "exceed[ing] authorized access."[44] As a result, "while the target of the law was computer hackers," the statute reads as though liability extends "to anyone who accessed a computer without authorization or

[37] Note that while the technical description of scraping is relatively straightforward, courts have not come to a consensus on a legal definition of scraping, which can sometimes cloud opinions on the topic in caselaw. Carrero, *supra* note 8, at 137 ("Because web scraping practices encompass a broad range of activity, courts have not come to a consensus on common terminology for web scraping or what activity qualifies as 'scraping.'").

[38] Kathleen C. Riley, *Data Scraping As A Cause of Action: Limiting Use of the CFAA and Trespass in Online Copying Cases*, 29 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 245, 266 (2018).

[39] *See* hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 998 (9th Cir. 2019).

[40] Bellia, *supra* note 31 at 2256.

[41] *Id.*

[42] *Id.*

[43] The Computer Fraud and Abuse Act, 18 U.S.C. § 1030(a)(1) (2008).

[44] *See* Patricia L. Bellia, *A Code-Based Approach to Unauthorized Access Under the Computer Fraud and Abuse Act*, 84 GEO. WASH. L. REV. 1442, 1444–45 n. 7 (2016). ("The statute defines the phrase 'exceeds authorized access' as follows: '[T]o access a computer with authorization and to use such access to obtain or alter information in the computer that the accesser is not entitled so to obtain or alter.'" *Id.* at 1445 (quoting 18 U.S.C. § 1030(e)(6).).

'exceed[ed] authorized access.'"[45] This language could apply to those operating scrapers.[46]

The CFAA does not define, and courts interpreting the statutes are not always in agreement on, what specific activities violate these provisions.[47] Specifically, cyberlaw scholar Orin S. Kerr summarized that with respect to the CFAA, "[n]o one knows what it means to 'access' a computer . . . or when access becomes 'unauthorized.'"[48] Particularly relevant for scraping claims, the language leaves open the question of whether the statute "cover[s] access to information in circumstances where a content provider generally makes information accessible to the public but seeks to limit its subsequent use"[49]–e.g., "whether a content provider can use the statute to control access to information that is otherwise publicly available."[50] But, because the statute contains a definition for a "protected computer,"[51] which is "broad enough to capture any Internet-connected computer, . . . any effort to limit application of the statute to nonpublic information must draw upon" a narrower interpretation of another piece of the statutory language.[52]

2. CFAA Scraping Claims in Caselaw. – Today, courts remain split as to "how to interpret unauthorized access," no less so when it comes to scraping.[53] The provisions of the CFAA invoking that phrasing "have been interpreted in numerous ways by federal courts and legal scholars."[54] About forty of the roughly sixty opinions that have considered the application of the CFAA to web scraping

[45] Mark A. Lemley & Mark P. McKenna, *Unfair Disruption*, 100 B.U. L. REV. 71, 88 (2020).

[46] "Claims against scrapers tend to be brought under the 'obtaining information' provisions in 18 U.S.C. § 1030(a)(2)(C) and the 'computer fraud' provisions in § 1030(a)(4)." Sellars, *supra* note 26, at 391. Note, though "a few also address the 'damage' provisions in § 1030(a)(5)." *Id.* A violation of the "obtaining information" provision occurs when one "intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains information from any protected computer." *Id.* Under either provision, the key requirements are that a plaintiff or prosecutor must show that a user's access to the computer was "without authorization" and "exceeds authorized access." *Id.*

[47] Kerr, *supra* note 7, at 1596 ("The few courts that have construed these terms have offered widely varying interpretations.").

[48] *Id.* Furthermore, the question of what constitutes a "computer" is no longer clear in the age of cell phones, for example. *See* Rob Williams, *Should Tablets and Smartphones be Considered "PCs"?*, TECHGATE, (Nov. 23, 2011) https://techgage.com/article/should_tablets_and_smartphones_be_considered_pcs/).

[49] PATRICIA BELLIA, PAUL SCHIFF BERMAN, BRETT M. FRISCHMANN, DAVID G. POST, CYBERLAW 734 (4th ed. 2010).

[50] *Id.* at 737.

[51] "The term 'protected computer' refers to any computer 'used in or affecting interstate or foreign commerce or communication,' 18 U.S.C. § 1030(e)(2)(B)–effectively any computer connected to the Internet . . . including servers, computers that manage network resources and provide data to other computers." hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 999 (9th Cir. 2019) (citation omitted). *See, e.g.,* "LinkedIn's computer servers store the data members share on LinkedIn's platform and provide that data to users who request to visit its website. Thus, to scrape LinkedIn data, hiQ must access LinkedIn servers, which are "protected computer[s]." *Id.*

[52] Patricia Bellia has found that this narrowing could be drawn "perhaps from a narrow interpretation of 'access'; from a narrow interpretation of what it means for access to 'exceed' what is authorized or to be 'without authorization'; or from other structural or policy considerations." BELLIA ET AL., *supra* note 49, at 734.

[53] Carrero, *supra* note 8, at 148.

[54] Riley, *supra* note 38, at 267.

have analyzed the substantive claims.[55] "How precisely to interpret these phrases has been at the center of a very large portion of the discussion about the CFAA."[56] Overall, Bellia observed that "the caselaw reflects at least five different interpretive paradigms."[57]

The broadest interpretation of liability can be found in the approach by the First, Fifth, and Eleventh Circuits, which "have broadly interpreted the CFAA to include violations of a corporation's terms of use policies."[58] By contrast, the Second, Fourth, and Ninth Circuits "narrowly construe the CFAA as an antihacking statute that only penalizes access if it amounts to 'breaking and entering' a computer without any lawful access at all."[59] This approach translates to a general rule that entry of a bot is not legally different from the entry of a human using the browser, assuming both have the same access and permissions to the site, because in both cases the user requests data that is publicly available or at least open to that user. Recent decisions have included in this ruling as justification the fact that the CFAA was passed to limit hacking, not scraping.[60] For example, "the Fourth Circuit in *WEC Carolina Energy Solutions LLC v. Miller* stated it could not 'contravene Congress's intent by transforming a statute meant to target hackers into a vehicle for imputing liability to . . . [those] who disregard a use policy.'"[61]

While the circuit split has persisted, the most recent trend in scraping cases suggest a general narrowing of CFAA liability, "recognizing that both the public interest in public web scraping and the technical similarities between web scraping and web browsing should limit application of the CFAA to web scraping."[62] For example, the Ninth Circuit's 2019 opinion in *hiQ Labs, Inc. v. LinkedIn Corp.,* produced the general rule for publicly-accessible data for that circuit that CFAA liability for scrapers "is premised on a distinction between information presumptively accessible to the general public and information for which authorization is generally required."[63] This is because "[t]he CFAA was enacted to prevent intentional intrusion onto someone else's computer–specifically, computer hacking."[64] Under this premise, the test becomes, whether "a computer

[55] Sellars, *supra* note 26, at 388: "The opinions begin in 2000, a little less than a decade after the establishment of HTTP and the World Wide Web in 1991, and grow in frequency nearly every year since, from one to two opinions per year in the early 2000s to closer to six to eight per year in the 2010s." *Id.* at 388–89. Note that this trend "roughly tracks the expansion of the CFAA in the civil context more broadly. There have been a little over a dozen appellate opinions in cases involving web scraping, but only one has generated something resembling a dissenting opinion." *Id.*

[56] *Id.* at 391–92.

[57] Bellia, *supra* note 44, at 1445; see id. n. 10 for additional work on interpreting this part of the statute.

[58] Carrero, *supra* note 8, at 148.

[59] *Id.* at 148–49.

[60] *Id.* ("In 2015, the Second Circuit held in *United States v. Valle* that a narrow interpretation of the statute is consistent with the statute's principal purpose of addressing the problem of hacking, i.e., trespass into computer systems or data.") (internal quotation marks omitted).

[61] *Id.*

[62] Sellars, *supra* note 26, at 412–13.

[63] hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 998 (9th Cir. 2019).

[64] *Id.* at 1000. The case then goes on to talk about how legislative history of the CFAA analogized intentional intrusion to breaking and entering. *See* United States v. Nosal (Nosal I), 676 F.3d 854, 858 (9th Cir. 2012) (citing S. Rep. No. 99-432, at 9 (1986) (Conf. Rep.)).

network generally permits public access to its data."[65] In the event that it does, "a user's accessing that publicly available data will not constitute access without authorization under the CFAA." [66]

In summary, the circuits remain split on what scraping activity may violate the CFAA, and it is highly fact dependent, varying across the accessibility of the scraped data as well as the terms of use of the sites bringing suit. However, in the most recent cases on point, authorization is generally not required for publicly accessible information, regardless of the type of access.[67]

B. Contract Claims

"Automated scraping violates the policies of sites like Facebook[68] and Twitter,[69] the latter of which specifically prohibits scraping[70] to build facial recognition databases."[71] Both sent cease-and-desist letters to Clearview stating so much in the aftermath of the news it was scraping their users' photographs. "But it's unclear whether they have any legal recourse in the current system."[72]

Under a contract theory, platforms like Twitter and Facebook would bring a claim for breach of contract, for a scraper's violation of the site's terms of services or terms of use. These claims are brought routinely today. For example, in a very recent 2020 Nevada case, plaintiffs suing a scraping firm included in their argument that the defendants had viewed and visited the plaintiff 4Internet's terms of use page, which stated that "all visitors must only access the webpages using the system interface."[73] According to the plaintiff, "this term means that [u]sing [a] bot to access

[65] *hiQ*, 938 F.3d, at 1003.

[66] In *hiQ*, the defendant was scraping data from LinkedIn's publicly accessible member profiles. "In May 2017, LinkedIn sent hiQ a cease-and-desist letter, asserting that hiQ was in violation of LinkedIn's User Agreement and demanding that hiQ stop accessing and copying data from LinkedIn's server." In the letter, LinkedIn stated that "if hiQ accessed LinkedIn's data in the future," it would be violating the CFAA, among other state and federal laws (though LinkedIn asserted that it has "claims under the Digital Millennium Copyright Act and under trespass and misappropriation doctrines," it had "chosen for present purposes to focus on a defense based on the CFAA," so the CFAA was the sole defense that the court addressed in its opinion. *hiQ*, 938 F.3d, at 995.), and importantly warned hiQ that it had "implemented technical measures to prevent hiQ from accessing, and assisting others to access, LinkedIn's site, through systems that detect, monitor, and block scraping activity." *hiQ*, 938 F.3d, at 992. Applying its test, the court held that hiQ was not liable under the CFAA for its scraping activity. *Id.*

[67] The most recent caselaw such as *hiQ* holds that "where access is open to the general public, the CFAA 'without authorization' concept is inapplicable." *Id.* at 1000.

[68] FACEBOOK AUTOMATED DATA COLLECTION TERMS, https://www.facebook.com/apps/site_scraping_tos_terms.php (last visited Jan. 29, 2021).

[69] TWITTER TERMS OF SERVICE, https://twitter.com/en/tos, (last visited Jan. 29, 2021).

[70] TWITTER DEVELOPER TERMS, https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases (last visited Jan. 29, 2021).

[71] Matsakis, *supra* note 4.

[72] *Id.*

[73] Miller v. 4Internet, LLC, 433 F. Supp. 3d 1188, at 1193 (D. Nev. 2020) (citing ECF No. 9 at 46–47).

and obtain information from the 4Internet server exceeded the authority or permission granted to users of the site."[74]

Success on these claims requires first that the scraper's behavior actually violates the website's language in its terms of service. Assuming that the terms of service include such language such that scraping would violate those terms, "many courts are increasingly willing to enforce contracts of adhesion that appear online, such as clickwrap and browsewrap agreements."[75] This legal theory isn't actually all that new. In *ProCD, Inc. v. Zeidenberg*, "the Seventh Circuit found that a shrinkwrap license agreement on software protected a compilation of data where copyright could not."[76] Sometimes, like in the circuit split seen in CFAA claims, courts have chosen to inquire beyond the simple existence of the term language, into "whether a user had actual or constructive notice of a website's terms of use in order to determine whether a contract was formed."[77] To this end, "[c]ourts have sometimes suggested that, were websites to make their terms more accessible, users would more effectively be bound by terms of service."[78]

However, in scraping cases, these claims often dead end because platforms tend to terminate any user (or block the IP address of a user, in the case of a publicly-accessible website) that is scraping, as a first step in preventing that activity. Of course, once the user is terminated, usually so is their obligation under any user agreement.

C. Copyright Claims

Finally, some plaintiffs have argued to hold scrapers accountable under a copyright theory of liability. However, in traditional data scraping cases litigating the scraping of numerical or textual data, copyright infringement, if alleged, is often dismissed.[79] In these cases, courts have been quick to remind plaintiffs of "a prime theorem of copyright law-

[74] *Id.*

[75] Riley, *supra* note 38, at 273 (referencing Erin Canino, *The Electronic "Sign-in-Wrap" Contract: Issues of Notice and Assent, the Average Internet User Standard, and Unconscionability*, 50 U.C. DAVIS L. REV. 535, 541 (2016); AT&T Mobility LLC v. Conception, 563 U.S. 333, 346–47 (2011).

[76] Riley, *supra* note 38, at 263–64. *See* ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1455 (7th Cir. 1996).

[77] Riley, *supra* note 38, at 274. *See, e.g.*, DHI Grp., Inc. v. Kent, No. H-16-1670, 2017 WL 4837730 (S.D. Tex. Oct. 26, 2017).

[78] Riley, *supra* note 38, at 303. *See also* Nguyen v. Barnes & Noble Inc., 763 F.3d 1171, 1179 (9th Cir. 2014) (citations omitted) ("While failure to read a contract before agreeing to its terms does not relieve a party of its obligations under the contract, the onus must be on website owners to put users on notice of the terms to which they wish to bind consumers. Given the breadth of the range of technological savvy of online purchasers, consumers cannot be expected to ferret out hyperlinks to terms and conditions to which they have no reason to suspect they will be bound.")

[79] Riley, *supra* note 38, at 276. *See, e.g.*, eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1072 (N.D. Cal. 2000) ("BE argues that the trespass claim … 'is similar to eBay's originally filed but now dismissed copyright infringement claim'"); *Naturemarket*, 694 F. Supp. 2d atl056; Allure Jewelers, Inc. v. Ulu, No. 1:12CV91, 2012 WL 4322519 (S.D. Ohio Sept. 20, 2012) (dismissing Allure's copyright claim based on late registration).

facts, as such, are not subject to copyright protection."[80] Courts consistently interpret data to fall squarely within the uncopyrightable facts camp.[81] As a result, "[c]opyright claims have not had particular success in data scraping cases."[82] In turn, "proprietors of social media and other user-based websites have attempted to prohibit third parties from copying their data under the [CFAA], state hacking statutes, and the related tort of trespass to chattels."[83]

On the other hand, some traditional scraping cases have included copyright claims that have some copyright element attached to the data that is scraped. In these cases, *Ticketmaster Corp. v. Tickets.Com, Inc.* is instructive on the approach courts take. There, the court considered three separate steps of the scraping process that may constitute copyright infringement and examined each separately.

> The first is whether the momentary resting in the [defendant's] computers of all of the electronic signals which are used to form the video representation to the viewer of the interior web pages of the [defendant's] computer constitutes actionable copyright infringement. The second is whether the URLs, which were copied and used by [the defendant], contain copyrightable material. The third is whether [the defendant]'s deep-linking caused the unauthorized public display of [Ticketmaster] event pages.[84]

The court granted summary judgment to the defendant on all three copyright issues. Regarding the URL issue, the court relied directly on a landmark copyright principle, citing *Feist v. Rural Telephone*: "There is nothing sufficiently original to make the URL a copyrightable item, especially the way it is used. There appear to be no cases holding the URLs to be subject to copyright. On principle, they should not be."[85] Second, regarding the deep-linking issue, the court distinguished the display of any copyrighted information as infringement because "[i]n this case, a user on the [Tickets.com] site was taken directly to the originating [Ticketmaster] site, containing all the elements of that particular [Ticketmaster] event page."[86] And finally, on the first point, which cuts to the heart of a copyright claim against scraping, the court held that the copying that occurred was fair use, because it was for the sole purpose of extraction of unprotected facts. "Taking the temporary copy of the electronic information for the limited purpose of extracting unprotected public facts leads to the conclusion that the temporary use of the electronic

[80] Ticketmaster Corp. v. Tickets.com, Inc., No. 99CV7654, 2000 WL 1887522, at *3 (C.D. Cal. Aug. 10, 2000), *aff'd*, 2 F. App'x 741 (9th Cir. 2001) ("The primary star in the copyright sky for this case is that purely factual information may not be copyrighted.") (referencing Feist Publ'n, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991).).

[81] *See, e.g.*, New York Mercantile Exch., Inc. v. IntercontinentalExchange, Inc., 497 F.3d 109, 114 (2d Cir. 2007) ("All facts, scientific, historical biographical, and news of the day, may not be copyrighted and are part of the public domain available to every person.").

[82] Riley, *supra* note 38, at 264.

[83] *Id.*

[84] Ticketmaster Corp. v. Tickets.Com, Inc., 2003 WL 21406289, at *4 (C.D. Cal. Mar. 7, 2003).

[85] *Id.* at *5.

[86] *Id.* at *6.

signals was "fair use" and not actionable."[87] There, "the court- having accepted that Ticketmaster's website was copyrightable . . . (based on the notion that the organization and arrangement of the information on a website is copyrightable) . . . evaluated Tickets.com's copying and determined that its . . . activity was fair use."[88] Thus, even in cases which copyright protected material may have been scraped, courts have been unwilling to find scrapers liable for copyright infringement.

III. SCRAPING PHOTOGRAPHS

Part II summarized the state of the law for data scraping, concluding that run-of-the-mill scraping of traditional data types from public websites is generally effectively unrestricted today in most circuits under any theory of liability. However, in all of the cases discussed in Part II, scraping occurred on data other than photographs. As a result, to the extent that photographs are different from conventional data, scraping photographs presents a novel legal question that has yet to be taken up in a court case.

To start, it is important to briefly clarify what is included in the notion of "conventional" data, drawing a line between photographs and data that has been at issue in prior scraping cases. In many prior scraping cases*,* the data at issue has usually been a combination of numeric and textual data, such as in *hiQ* in which the data included strings of text including name, job title, work history, and skills.[89] In other leading scraping cases, such as *Ticketmaster* and *EF Cultural Travel BV v. Explorica,*[90] the data at issue was proprietary pricing data. In those cases, the prices were compiled with more than a modicum of effort and creativity, but still comprised an amalgamation of underlying facts.[91]

Recognizing the distinction between photographs and data that has so far been litigated in scraping cases, this section highlights the key differences between photographs and data on which we have decided the current scraping laws. First, as noted, photographs, unlike factual data, are protected under federal copyright law.[92] Second, most photographs

[87] *Id.* at *5.

[88] Riley, *supra* note 38, at 277 (*referencing* Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *5-6 (CD. Cal. Mar. 7, 2003).).

[89] hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 991 (9th Cir. 2019).

[90] EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577 (1st Cir. 2001) (Tour company sued a competitor, alleging that the defendant violated the Computer Fraud and Abuse Act and the Copyright Act by using a "scraper" software tool to scrape tour prices and other information from the company's website.) *See also*, eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1062 (N.D. Cal. 2000) (litigating the scraping of "information regarding eBay-hosted auctions for Beanie Babies and Furbies[.]").

[91] *See* Ticketmaster Corp. v. Tickets.Com, Inc., 2000 WL 525390, at *2 (C.D. Cal. Mar. 27, 2000) ("While the expression, organization, placement, etc., of the factual data may be protected, Tickets is not alleged to have copied the method of presentation, but rather to have extracted the factual data and presented it in its own format.").

[92] The data types scraped in existing cases either, (1) have not enjoyed copyright protections because they are facts, or (2) were at issue in the case because the data used after scraping constituted merely the factual aspects of the scraped data.

scraped are user-generated media governed by a non-exclusive license to platforms, as compared to traditionally-scraped data which is overwhelmingly created by the site displaying the data. Both of these differences impact liability for scraping social media photographs as compared to conventional data. This section details these differences and provides recommendations for how a court applying the law to a novel question of photo scraping should consider each one.

Before addressing these differences, it is important to highlight an important similarity between photographs and traditional data such as numeric and textual data. Under the CFAA, photographs are treated analogously to such data types. Here, there is no major distinction at play as there is in contract and copyright claims. Depending on the social media user's privacy settings, social media photographs may be governed by another Ninth Circuit CFAA case, *Facebook v. Power Ventures*, rather than *hiQ,* which focused on publicly-accessible user data. "In rejecting LinkedIn's claim of unauthorized access, the *hiQ Labs*[93] court distinguished the Ninth Circuit's decision in *Power Ventures* on the grounds that the data in that case was not 'public,' as one can only access Facebook content with a username and password."[94] But this difference applies with equal force, where courts recognize it, to photographs and traditional user-generated data.

A. Copyright and Photographs

Photographs, unlike traditional data including numeric and textual data, are protected under federal copyright law.[95] Because scraping a photograph inherently entails making a copy of that photograph on a second machine, scraping photographs constitutes prima facie copyright infringement.[96] While user-generated media are largely the photographs that are subject to data scraping activities for facial recognition, such as in the case of Clearview, copyrights for user-generated photographs are no exception. The copyright for user-generated photographs posted on platforms is assigned to the photographer, who is usually, but not always,

While photographs enjoy strong, affirmative copyright protections, courts have yet to consider whether a photograph should be considered a part of the latter category.

[93] Note that this reference to *hiQ* is to the district court opinion and was written before the Ninth Circuit's opinion was published. *See* hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099 (N.D. Cal. 2017).

[94] Sellars, *supra* note 26, at 409.

[95] "Simply put, copyright attaches to the content the moment it is fixed; the originality threshold for copyright protection is sufficiently low that much of what gets posted, generated, and shared on these sites is protected by copyright (excluding ideas, facts, and other public domain content." BELLIA ET AL., *supra* note 49, at 765. *See generally* Ticketmaster Corp. v. Tickets.com, Inc., 2000 WL 1887522, at *3 (C.D. Cal. Aug. 10, 2000), *aff'd*, 2 F. App'x 741 (9th Cir. 2001) ("The primary star in the copyright sky for this case is that purely factual information may not be copyrighted.") (referencing Feist Publ'n, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991).).

[96] *See* 17 U.S.C. § 501(a) ("Anyone who violates any of the exclusive rights of the copyright owner as provided by sections 106 through 122 or of the author as provided in section 106A(a) . . . is an infringer of the copyright or right of the author."). *See also supra* Part I.

the individual who posted the pictures.[97] This understanding has led many to assume that copyright steps in to protect against scraping photographs in cases such as those actions against Clearview, unlike that of traditional data scraping.

1. Analogous scraping activities are fair use. - However, those scraping photographs can look to the holding of Ticketmaster, which saved from copyright liability similar scraping as a fair use. There, Tickets.com was scraping Ticketmaster's website for data by making a copy for 10-15 seconds of portions of Ticketmaster's website and then within that time period extracting the factual information to display on its own site. The California district court recognized that "there is undeniably copying of the electronic bits which make up [Ticketmaster's] event pages when projected on the screen."[98] However, the scraping was considered a fair use under Ninth Circuit precedent that held "copying for reverse engineering to obtain non-protectible information is permitted by the fair use doctrine in certain circumstances."[99]

A straightforward application of the *Ticketmaster* logic should extend to scraping photographs. This argument hinges critically on the fact that photographs can be separated from the data contained in a digital image, the latter of which is used in facial recognition algorithms. Both the legal and policy arguments supporting such separability provide arguments for extending the *Ticketmaster* fair use logic to the act of scraping photographs.

First, the Ninth Circuit has repeatedly held that digital photographs are a separate entity from the bytes comprising a "face scan."[100] In these cases, courts rely on the logic that "face scans," referring to the format of a photograph used in a facial recognition algorithm, are data, and separable from the photographs from which they are derived. As a result, the Ninth Circuit has held that scanning and using these face scans without the subject's consent is a violation of the Illinois state law protecting an individual's biometric privacy.[101] To hold that such separability does not exist when it is copyright law, rather than a privacy law, that is

[97] "Copyright allocates ownership of User-generated content to the creator (not the subject), which is often the user." BELLIA ET AL., *supra* note 49, at 765. "[A]s a practical necessity, site owners and/or service providers must obtain, at a minimum, nonexclusive licenses that permit them to redistribute user content publicly." *Id.* This usually occurs at the site's terms of use or similar contractual agreement.

[98] Ticketmaster Corp. v. Tickets.com, Inc., No. 99CV7654, 2000 WL 1887522, at *3 (C.D. Cal. Aug. 10, 2000), *aff'd*, 2 F. App'x 741 (9th Cir. 2001). The court noted that even though "the copying is transitory and temporary and is not used directly in competition with [Ticketmaster], . . . it is copying and it would violate the Copyright Act if not justified." *Id.*

[99] *Id.* (referencing Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).).

[100] *See e.g.*, *Patel v. Facebook*, regarding the legally cognizable difference between a facial scan and a photograph support separability when it comes to data and photographs. *Patel*, 932 F.3d, at 1273-74. For a lengthy discussion of–and challenge to–the separability concept for photographs and face scans, see Google's argument in Rivera v. Google Inc., 238 F. Supp. 3d 1088, 1092-1100 (N.D. Ill. 2017).

[101] *Id.*

violated would simply be inconsistent. Furthermore, in *Ticketmaster,* the court highlighted in accepting the fair use defense that "it was unlikely that the spiders could have been programmed to take only the factual information from the [Ticketmaster] web pages without initially downloading the entire page."[102] It is inarguable that, like in *Ticketmaster*, photo scrapers cannot easily be programmed to scrape only the face scan–such a tool would at best put additional pressure on the plaintiff's system. In short, photo scraping, like the scraping of traditional data, should be held a "fair use since it was necessary to temporarily copy . . . to obtain the non-protected material."[103]

The policy justifications behind these holdings further buttress this interpretation of the law, in that the underlying policy arguments enable and encourage the extraction and use of data as a fair use. Copyright "affords protection to authors as an incentive to create."[104] "In determining whether a challenged use of copyrighted material is fair, a court must keep in mind the public policy underlying the Copyright Act: to secure a fair return for an author's creative labor and to stimulate artistic creativity for the general good."[105] In *Ticketmaster*, the court opined that "no public policy . . . would be served by restricting [Tickets.com] from using spiders to temporarily download [Ticketmaster]'s event pages in order to acquire the unprotected, publicly available factual event information" because the protectible elements that were copied were "discarded and not used by [Tickets.com] and . . . not exposed to the public." [106] In other words, Tickets.com's scraping "was not exploiting [Ticketmaster]'s creative labors in any way."[107] A face scan is similarly situated with purely utilitarian purposes in mind. The facial scan is utilized in a format unrecognizable to the human eye as the original photograph, to be seen and understood only by a computer in an algorithm.

2. Face Scans are not protectible derivative works. - Whether the face scan itself is protectible is a question of first impression that courts should not over complicate by treating differently from other types of data used in algorithms. That is, a face scan could be simply data, or another form of unprotectible useful article, or a protectible derivative work. Two arguments support finding face scans to be either a useful article or otherwise factual data, and not a protectible derivative work. First, if face scans are not simply data, they also meet the definition of useful articles. The Copyright Act is clear that useful articles of any type are not protectible. Section 101 of the Act defines a useful article as "an article having an intrinsic utilitarian function that is not merely to portray the appearance of the article or to convey information."[108] In order to determine whether an article is a "useful article" within the meaning of the statute, a court

[102] Ticketmaster Corp. v. Tickets.Com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *5 (C.D. Cal. Mar. 7, 2003)

[103] *Id.* at *4.

[104] Computer Assocs. Int'l, Inc. v. Altai, Inc., No. 762, 1992 WL 139364, at *1 (2d Cir. June 22, 1992), *opinion withdrawn and superseded on reh'g*, 982 F.2d 693 (2d Cir. 1992).

[105] Ticketmaster Corp. v. Tickets.Com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *5 (C.D. Cal. Mar. 7, 2003).

[106] *Id.*

[107] *Id.*

[108] Tyler T. Ochoa, *What Is A "Useful Article" in Copyright Law After Star Athletica?*, 166 U. PA. L. REV. ONLINE 105, 111 (2017).

should first identify the "intrinsic utilitarian function" or functions served by the article. It should then ask whether the article is excluded from the statutory definition because the only functions are "merely to portray the appearance of the article or to convey information."[109] In Star Athletica, the Court accepted that "the 'shape, cut, and dimensions' of a [cheerleading] uniform are 'utilitarian aspects' to which copyright does not extend."[110] The only potential aspects of a face scan that may be protectible are analogous to the shape, cut, and dimensions of the face of a person captured in a photograph. As such, they should be interpreted again consistent with existing caselaw as useful articles.

Second, one may argue that a face scan is data, but data that should be classified as a derivative work. So this argument goes, face scans as data deserve protection in that they are a combination of data points that deserve copyright protection under *Feist*. *Feist* held that although data itself is not protectible, "[t]o be sure, the manner of expression and format of presenting those facts is protectible."[111] But courts should resist the temptation to classify face scans as a protectible arrangement. The justification for the protection of the "manner and mode of expression and format" of data is "a fundamental concept of copyright law:" "ideas and knowledge may not become the property of any one person . . . What is protectible is the manner in which the idea or knowledge is expressed."[112] Face scans fall more closely into the former category because "[t]o be copyrightable, a compilation of facts must exhibit subjectivity in which facts are included or how they are arranged,"[113] and a face scan is a reduction of those elements to the actual shape of a real person's face. An individual's face is not protectible under copyright law, only photographs that feature it. Likewise, any biometric scan of a human face or other body part should not be granted copyright protection.

Finally, one last potential counterargument is that a facial scan alone may be unprotectible, but initial scraping of the photograph onto another server still constitutes prima facie copyright infringement. This argument raises two important legal questions. First, should it matter when conversion happens? If conversion happens mid-stream, is it possible that scraping photographs is protectible, given there is never a copy of the photograph made for a non-transient period of time in the scraping and conversation process? And second, can the act of copying be considered fair use under *Ticketmaster,* given the photograph itself is not used again? However, combining *Star Athletica's* separability description of useful objects with *Ticketmaster's* fair use holding presents a relatively airtight defense for those scraping photographs. The use of the face scan alongside discarding the photograph is nearly analogous to the

[109] *Id.* at 115 (citing 17 U.S.C. § 101(2012).).

[110] *Id.* at 116 (citing Star Athletica, LLC v. Varsity Brands, Inc., 137 S. Ct. 1002, 1016 (2017) ("In any event, as explained above, our test does not render the shape, cut, and physical dimensions of the cheerleading uniforms eligible for copyright protection.")).

[111] Ticketmaster Corp. v. Tickets.com, Inc., No. 99CV7654, 2000 WL 1887522, at *3 (C.D. Cal. Aug. 10, 2000), *aff'd*, 2 F. App'x 741 (9th Cir. 2001).

[112] *Id.*

[113] Riley, *supra* note 38, at 305. *See also* Miriam Bitton, *Protection for Informational Works After Feist Publications, Inc. v. Rural Telephone Service Co.*, 21 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 611, 631 (2011).

defendant's act of copying a website, only to quickly discard its protectible copy as soon as the data from it was extracted.[114] Courts should favor this argument, given the alternative would dictate how developers should write their code. In other words, courts should lean on this defense in order to avoid making a judgment on whether it matters when the conversation from photograph to face scan occurs in the scraping process.

In sum, courts should hold photo scraping a fair use in order to remain consistent with the policy and existing precedent in similarly situated copyright scraping cases. Albeit, this conclusion produces uncomfortable results. The result of allowing separation of photographs from their underlying data is to provide no defense against unrestricted use of the individual's face template–notwithstanding any other privacy law that would police such uses. It also effectively enables facial recognition software. Part IV will discuss potential paths to address this uncomfortable outcome.

B. Data Ownership

1. Non-exclusive licenses bar platforms from asserting claims for user-generated data. – In hiQ, the Ninth Circuit went out of its way to highlight that "LinkedIn has only a non-exclusive license to the data shared on its platform, not an ownership interest."[115] The terms of service between the user and the platform often dictate merely a non-exclusive license, barring any platform's claim of right to the data to the exclusion of data scrapers. In fact, "LinkedIn specifically disclaims ownership of the information users post to their personal profiles: according to LinkedIn's User Agreement, members own the content and information they submit or post to LinkedIn."[116] The contract users agreed to "grant[s] LinkedIn only a non-exclusive license to 'use, copy, modify, distribute, publish, and process' that information."[117]

This fact puts pressure on platforms bringing contract breach claims for scraping in violation of a site's terms of service, when the data at issue is governed in the terms by a non-exclusive license. In granting the platform a right to exclude scrapers from accessible user data, courts run the risk of expanding the contract rights a platform has under its license agreement with the user. LinkedIn further complicated matters after it terminated hiQ's access in that "LinkedIn has not explained how it

[114] Note that this analysis assumes that the photograph is discarded. A separate copyright infringement issue may exist for companies like Clearview if they are reproducing the photographs in full as a part of their service to law enforcement. Without citing back to the source of the image, on the third issue of *Ticketmaster*, companies like Clearview and others scraping photographs may have a problem. *See, e.g.,* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 561 (S.D.N.Y. 2013) (holding that a software company's conduct in scraping, aggregating, and delivering to defendant's readers, substantial excerpts of copyrighted news articles was not protected by fair use). Because this note is on the topic of scraping and not on the topic of additional distribution of any copyrighted material, this is not explored further here.

[115] hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 998 (9th Cir. 2019).

[116] *Id.* at 989–90.

[117] *Id.*

can enforce its user agreement against hiQ now that its user status has been terminated."[118]

The Ninth Circuit also repeatedly emphasized that, at least when the data scraped is user-generated, enjoining scraping for any reason raises antitrust concerns. "[G]iving companies like LinkedIn free rein to decide, on any basis, who can collect and use data–data that the companies do not own, that they otherwise make publicly available to viewers, and that the companies themselves collect and use–risks the possible creation of information monopolies that would disserve the public interest."[119] Afterall, LinkedIn's "core business model–providing a platform to share professional information–does not require prohibiting hiQ's use of that information."[120]

2. Paths to asserting user claims directly. - There should still exist a user's right to prevent against such uses of their photograph. The grant of an exclusive license is one straightforward option that would give platforms the power to prevent such uses. But it is "likely that social media users would balk at giving services the exclusive licenses," particularly for the use of their photographs.[121] "One solution to this problem would be to allow user-based services to sue on behalf of their users in derivative form, or for social media companies faced with data scraping to hire attorneys to file class actions on behalf of their users."[122] Either solution would resolve the collective action problem users face in asserting these claims individually.

For example, copyright claims brought by platforms themselves have so far also been dismissed for this reason. "[B]ecause these services have non-exclusive licenses to user content," they lack standing to sue for copyright infringement.[123] "It is well-established law that "[a] non-exclusive license conveys no ownership interest, and the holder of a nonexclusive license may not sue others for infringement."[124] In *Craigslist Inc. v. 3Taps Inc.*, "the court held that Craigslist's terms of use did not constitute the required writing"[125] to grant an exclusive license. In fact, "[t]he lack of standing in copyright infringement lawsuits for user-based services explains why services like Facebook and LinkedIn have resorted to the CFAA as a potential remedy for copying of their websites."[126] In sum, granting platforms the power to sue on their users' behalf may provide new pathways for users to assert their copyright over their photographs, and may

[118] This logic also prevented liability under the CFAA. Applying its test to hiQ's scraping activity, the court held that "[t]he data hiQ seeks to access is not owned by LinkedIn and has not been demarcated by LinkedIn as private using such an authorization system," and thus hiQ was not liable under the CFAA. *Id.* at 998.

[119] *Id.* at 1005.

[120] *Id.* at 998.

[121] Riley, *supra* note 38, at 308.

[122] *Id.*

[123] *Id.* at 307–08.

[124] *Id.* at 307.

[125] *Id.*

[126] *Id.* at 308.

even offer them more control over whether an act of scraping constitutes hacking.

IV. UNRESTRICTED SURVEILLANCE AND STRONGER PROTECTIONS FOR CYBERPROPERTY

Part III illustrates how platforms and the users of those platforms likely are without an effective legal means to prevent the unrestricted scraping by companies such as Clearview of photographs posted on social media. But rather than conflating the purpose of existing law by simply treating photographs differently than regular data in scraping cases in order to prevent unwanted uses of photographs, this conundrum provides new reasons to strengthen protections against scraping across the board. In short, the apparent legality of scraping photos heightens the case for federal statutory limits on scraping any type of data.

Such protections should not restrict scraping by shoehorning it into a hacking law, but rather by offering platforms some means by which to signal that they prohibit scraping activity, that courts will respect. As Bellia and others have suggested in proposing such protections, "[s]o long as the law presumes a default rule of open access and places the burden on the system owner to adequately convey the limits on permissible uses of her system, property-rule protection for network resources is appropriate."[127] To that end, the law should empower platforms and sites with some technical means to define what constitutes trespass on their servers.[128]

Furthermore, this note's deep dive into Clearview's photo scraping illustrates a broader emerging issue with the direction of scraping law. That is, "[t]he need to protect personal data from the threat of unauthorized exploitation becomes more pressing every day."[129] In its petition to the U.S. Supreme Court for certiorari in *hiQ*, LinkedIn invoked Clearview, arguing that the Ninth Circuit's decision to allow hiQ to scrape LinkedIn had not only "eviscerated the legal argument" that websites may block unwanted scraping, but also that the decision enabled far more troublesome activities, like that of Clearview.[130] In other words, LinkedIn appropriately argued that scraping law has probably bent too far in favor of a free and open internet, and the results are echoing outside of traditional antitrust and property-based concerns. And further, "the prevailing position that the balance struck by copyright law should control" is breaking down under technological advances that have created means by which to reduce protectible works into validly non-protectible useful objects.[131]

To simply limit the scraping of photographs would also miss the point raised in these privacy-based concerns. The issues LinkedIn raises in its petition for cert and those raised among the many in society this year that have shown a desire to legally constrain activities like Clearview's extend beyond facial recognition technology. Facial recognition is only one

[127] Bellia, *supra* note 31, at 2173–74.

[128] *See, e.g.*, Bellia, *supra* note 44.

[129] LinkedIn Corporation v. hiQ Labs, Inc., 938 F.3d 985 (9th Cir. 2019), *appeal docketed*, (U.S. Mar. 12, 2020) (No. 19-1116) at 4.

[130] *Id.* at 4–5.

[131] BELLIA ET AL., *supra* note 40, at 715.

form of modern surveillance that can be based on open or available user data that can be scraped.[132] Privacy and security expert Bruce Schneier argues that "[f]ocusing on one particular identification method misconstrues the nature of the surveillance society we're in the process of building."[133]

Rather, those desired protections highlight that, with the benefit of hindsight, it is now clear that the single-track pursuit of the goal of an unrestricted, open internet was probably too imbalanced to work in the emerging technical environment. "The earliest cases involved claims that objectionable activities–typically, the sending of large quantities of unsolicited commercial e-mail," or more succinctly "plaintiffs seeking to prevent unwanted uses of their computer systems."[134] But now, the protections sought would be for individuals seeking to prevent unwanted uses of perfect copies of the dimensions of their *face,* and other equally identifiably granular datasets. "The uproar over Cambridge Analytica's massive misuse of Facebook user information and, more recently, Clearview's compilation of a vast database that will potentially allow for instant facial recognition (and possible surveillance) of billions of people, leaves no doubt that the public is deeply concerned about the issue of control of personal information and privacy on the Internet."[135] The apparent legality of such scraping activity presents an argument for a reversal of the recent trend towards laws that, guided by the principle of a free and open internet, favor scraping, and for a return to federal laws that provide stronger defenses for cyberproperty. It is time for lawmakers to step in, for the sake of balancing the benefits of the free and open internet against not merely property interests, but the privacy interests of individuals.

CONCLUSION

Unwanted as it may be, scraping a photo for use in a facial recognition algorithm or other software should not constitute copyright infringement, or a violation of any other existing federal laws. To hold otherwise would conflict with the policy behind existing copyright protections as well as broader law regarding the status of facial data used for such purposes. However, the fact that no other law today effectively prevents against any type of scraping activity provides a novel argument against lax scraping laws. Existing laws are enabling vast privacy abuses to emerge by making it legal for any company scraping copyrighted photographs online (without license or permission) to do so without violating any federal laws, as long as they extract only the data from the photograph when it is copied, thereby turning it into a facial scan. And equally invasive processes can be built off scrapable numeric and textual data that are personally-identifiable in nature.

[132] "Ubiquitous mass surveillance is increasingly the norm." Bruce Schneier, *We're Banning Facial Recognition We're Missing the Point.,* NY TIMES (Jan. 20, 2020, 5:00 AM), https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html.

[133] *Id.*

[134] Bellia, *supra* note 31, at 2166–67.

[135] LinkedIn Corporation v. hiQ Labs, Inc., 938 F.3d 985 (9th Cir. 2019), *appeal docketed,* (U.S. Mar. 12, 2020) (No. 19-1116) at 28.

Such ends provide the strongest evidence to date that the unrestrained pursuit of an open internet should be re-examined in this modern era. Courts hearing such arguments have taken that goal to the logical extreme, enabling facial recognition. Policymakers and lawmakers are currently scrambling to ban facial recognition as a result. A more straightforward approach would be to simply admit that open internet taken to its logical extreme is the wrong policy. Rather, we should be writing laws that balance open internet goals with property interests and privacy interests.