Talking Past Each Other

The Legal and Technical Challenges of Harmful Web Scraping

By Timothy H. Edgar*

The practice of web scraping – using automated tools to harvest data from websites – has been around almost since the debut of the World Wide Web in the 1990s. Scraping is a common practice, often taught to computer science students, and has both beneficial and malicious uses. The legal, policy, and ethical questions raised by web scraping practices have been debated for almost as long as the web has existed, but three new trends are converging that are exacerbating the problem of malicious web scraping.

First, the advent of generative AI has greatly increased the demand for scraped data, making it ever more valuable. Scrapers are incentivized as never before to evade common techniques that website owners use to discourage scraping. Second, norms in the technical community against unwanted scraping – never strong to begin with – have mostly broken down. Third, decisions of the federal courts that have narrowed the Computer Fraud and Abuse Act – while providing much-needed definition to a frustratingly broad and vague law – are likely to be

* Professor of the Practice of Computer Science, Brown University, and Lecturer on Law, Harvard Law School. Institutional affiliations are listed for identification purposes only.

misinterpreted and misunderstood as a green light for even the most invasive and harmful forms of scraping.

The result is that bad actors may increasingly turn to web scraping to find and harvest personal data for malicious purposes. This white paper seeks to define the problem of harmful web scraping and explain why lawyers and technologists often talk past one another on this topic. The paper will describe how web scraping relates to the technical concepts of authentication, authorization, and access control explain why so many lawyers and courts are confused about these concepts. It will conclude with a discussion of possible legislative solutions, including amending the Computer Fraud and Abuse Act (CFAA) or addressing harmful web scraping through broader privacy legislation in the United States and elsewhere.

I. Web scraping – the good, the bad, the ugly

What is web scraping?

Web scraping is the use of automated tools, such as bots, to collect data in bulk from websites. By using such tools, scrapers may capture large quantities of information, storing it for later analysis. According to one cybersecurity company official, "web scraping has valid business purposes such as research, analysis and news distribution," but can also be used for "malicious

purposes."[1] Web scraping typically involves three steps. First, software known as a scraper bot sends an HTTP GET request to the website – the same process that a web browser uses to obtain data. Second, the scraper looks for a specific pattern it is programmed to collect. Third, it generally converts the resulting data into another format that is useful for the scraper's purposes, such as a spreadsheet. Scrapers often use a "headless browser" – a type of web browser that eliminates the visual interface, making it much faster because it can request data without waiting for a page to load.[2]

Web scraping may be further defined by whether it is unwanted, unauthorized, or harmful. Many website owners describe scraping that violates a website's terms of service (TOS) as "unauthorized web scraping." Naturally, website owners ordinarily object to scraping that violates their TOS. Such activity may or may not be "unauthorized" in the meaning of the Computer Fraud and Abuse Act (CFAA), as discussed below. As a result, this paper will refer to scraping that violates a TOS and to which website owners object as "unwanted web scraping," to avoid confusion with the CFAA. Finally, web scraping may be harmful, innocuous, or beneficial from a broader societal standpoint.

Beneficial and innocuous web scraping

[1] Kevin Townsend, "Web Scraping – Is It Legal and Can It Be Prevented?," *Security Week*, Nov. 7, 2022, available at https://www.securityweek.com/web-scraping-it-legal-and-can-it-be-prevented/

[2] Cloudflare, "What is data scraping?", available at https://www.cloudflare.com/learning/bots/what-is-data-scraping/

Web scraping is a commonly-used tool with potentially beneficial applications. These include facilitating indexing and web searching, fostering data transparency and competition, and enabling academic research that is ethical and socially valuable.

As of May 24, 2023, the World Wide Web consisted of approximately 6.12 billion indexed pages.[3] To organize such information, search engines are constantly crawling the web and categorizing it. Because the web does not maintain a central index, web crawling is essential to finding information on the internet. There is a way for websites to opt out. Websites typically maintain a "robots.txt" file, which alerts a crawler as to which pages it may index and which pages it should avoid. Search engines may also use web scrapers. Scraping is related, but distinct, from crawling. Scrapers are designed to imitate web browsers and may take actions like filling out forms. Search engines like Google use tools to crawl the web in order to index internet content, while scrapers are designed to extract content. In other words, scraping is more invasive. Scrapers are also more likely to ignore a website's "robots.txt" file.

Web scraping also benefits the public by aiding scientific inquiry and research, taking advantage of the extraordinary data transparency that the World Wide Web has engendered. For example, scraping has been used by social media researchers to understand how Google's recommendation algorithms select content to increase user engagement on YouTube, and the

[3] "The size of the World Wide Web," visited June 6, 2023, available at https://www.worldwidewebsize.com/

potentially harmful impact such algorithms may have on political polarization.[4] Naturally, social media companies may object to such scraping when those findings are critical, even if the scraping is for noncommercial purposes and is both legal and ethical. Web scraping for commercial purposes may also provide broader societal benefits, when done in the right way. One of the most common forms of commercial web scraping is market research. Companies automatically scan their competitors' websites, convert the information they uploaded in html to a format suitable for analysis by a spreadsheet, and use it to make pricing or other decisions about similar products or services.[5] Transparency of economic data, including data about pricing, reduces market friction and increases economic efficiency. Incumbents might prefer more opaque markets, but consumers usually benefit from robust competition.

Unwanted web scraping and a website's Terms of Service (TOS)

Many website owners describe scraping that violates their terms of service (TOS) or similar legal policies as "unauthorized web scraping." As discussed above, because whether scraping is unauthorized is a legal question under the Computer Fraud and Abuse Act (CFAA), for clarity this paper instead describes such scraping as unwanted, although it may be unauthorized as well.

[4] Brandi Geurkink, "POV: Big Tech has a glaring double standard when it comes to web scraping," *Fast Company*, April 18, 2023, available at https://www.fastcompany.com/90882752/pov-big-tech-has-a-glaring-double-standard-when-it-comes-to-web-scraping

[5] Townsend, *Security Week*, *supra.*

Websites typically outline their relationship with users through their TOS, privacy policies and similar legal documents posted on their sites, which set the parameters of how they collect, use, and share data. TOS often seek to ban or limit web scraping, both for good and bad reasons. As discussed above, companies may want to keep pricing and other commercially sensitive information easily available to prospective customers, but limit access to competitors. While companies could choose to put such information behind a login and password, this practice could deter new customers and create friction with existing ones – and sophisticated software may find a weakness that permits the data to be scraped anyway. A creative platform may have expended substantial resources to assemble valuable data – such as photographs, art, or other content – only to find its work exploited by a more aggressive competitor that scrapes its site. Websites may also have reputational reasons to object to scraping. A social media network who has spent years building user trust may find an outsider scraping its site, exploiting personal data it never collected – and destroying user trust in an instant.

Ideally, TOS and similar documents protect both the owners of the websites and their users. A TOS or privacy policy may create enforceable promises that expose website owners to liability if those promises are broken. Privacy abuses or data breaches may result in lawsuits under common law contract or tort theories, as well as enforcement by the Federal Trade Commission under section 5 of the FTC Act, which prohibits unfair or deceptive acts or practices in commerce. This is one reason many websites prohibit or seek to control scraping in their TOS. In practice, TOS and other policies are drafted by lawyers to protect the interests of website owners. Whatever the weaknesses of a given website's TOS or privacy policy, they still define

some level of protection for users. By contrast, scrapers do not face such consequences when they misuse data. Scrapers have no relationship with the users whose data they have scraped and have made no promises to safeguard their data.

Harmful web scraping

While a website owner may object to scraping mainly because it harms its own interests, such unauthorized or unwanted scraping may also result in broader societal harms. Many harmful uses of scraping share a common theme. Scrapers exploit personal information collected for one purpose for entirely different purposes in ways that are surprising – and often distressing – to those whose information is being exploited. As a matter of ethics, such scraping is a violation of personal privacy and dignity, whether or not it is a violation of law.[6] Uncontrolled scraping also creates serious cybersecurity risks that websites would be irresponsible to ignore. Scammers and other threat actors liberally use bots, including web scrapers, to harvest sensitive information. Almost half of internet traffic in 2022 consisted of bots, the large majority of which were malicious.[7] Scrapers may obtain personal information even when users have not themselves provided it. Among the more common malicious uses of web scraping is to obtain contact information, such as email addresses or phone numbers often posted on company

[6] See Helen Nissenbaum, *Privacy in Context: Technology, Policy and the Integrity of Social Life* (2009).

[7] According to the report, in 2022, only 52.6% of internet traffic was from humans, while 47.4% was from bots – 30.2% from malicious bots and 17.3% from benign bots. Erez Hasson, "What We Learned from the 2023 Imperva Bad Bot Report," Imperva (blog), May 10, 2023, available at https://www.imperva.com/blog/a-decade-of-fighting-bad-bots-key-learnings-from-the-2023-imperva-bad-bot-report/

websites, that are then used by spammers for social engineering, robocalls, or other fraudulent schemes.[8]

Perhaps the most famous example of aggressive and privacy-invasive web scraping is Clearview AI, a company that has scraped billions of photographs from the web in order to sell a service that allows its customers to match a face to an identity without the person's consent or even knowledge. Clearview AI has built its business model on facilitating bulk facial surveillance for governments and other customers. The company boasts that it has assembled "the largest known database of 30+ billion facial images sourced from public-only web sources, including news media, mugshot websites, public social media, and other open sources." Clearview's business model is to couple this scraped data with facial recognition technology and sell access to the service to both government and private customers, including police, intelligence, military, and corporations.[9] Clearview AI has violated legal obligations in Europe and the United States. A UK privacy regulator fined Clearview $9.4 million for violating UK data protection law. France's data protection agency, CNIL, announced €20 million fine in October 2022 for violating the European Union's General Data Protection Regulation (GDPR).[10] In the United States, the American Civil Liberties Union (ACLU) settled a case against Clearview for violating the Illinois Biometric Privacy Act in May 2022. The settlement forced Clearview to stop offering its "faceprint" database to most businesses in US and to government entities in Illinois.

[8] Cloudflare, "What is web scraping?" *supra.*

[9] Townsend, *Security Week*, *supra.*

[10] See below for further discussion of GDPR.

HiQ Labs built its company's business model on scraping data from the popular social networking site LinkedIn, where users post resumes and job profiles. HiQ Labs even filed a successful lawsuit against LinkedIn objecting to the technical countermeasures it deployed to prevent the scraping, discussed further below. It is not surprising that many LinkedIn users were alarmed by HiQ's plans to offer employers a monitoring service that would alert them if LinkedIn users were busily updating their profiles, so that employers knew they might be seeking other jobs. While HiQ Labs said it was simply developing a product to help employers retain their most valued employees, it is unfortunately not uncommon for an employer to retaliate against or even preemptively fire an employee who seems to be looking for another job. If LinkedIn had offered such a service to employers – as HiQ Labs said it was considering – it would have had to contend with whether doing so would cause its users to flee the platform. HiQ Labs had no such concerns. Scraping LinkedIn profiles permitted another company to weaponize the LinkedIn social network for employer surveillance, without the consent of the company or its users.

Additionally, dating websites are particularly at risk from harmful web scraping. For years, users of dating websites have seen their data scraped for both commercial and noncommercial purposes. Such information, because it is so intimate, is often used for blackmail, scams, or by domestic abusers seeking to harm a partner or an ex-partner.[11] In one example, a researcher

[11] *See* Remarks of Hannah Shimko, Chief Executive, Online Dating Association, "2023 Data Privacy Event: The State of Unauthorized Scraping and Its Impact on Users and Industry," *YouTube*, at around 1:09:00, available at https://youtu.be/UfvCCYP7fF0

who was investigating the gray market in personal information was able to purchase one million dating profiles – assembled from major sites like Match, Tinder and OkCupid – for $153. Highly sensitive information was sold in batches based on characteristics like nationality, sexual preference or age. Purchasers use such information for advertising as well as to add user profiles to their own services in order to attract new subscribers – all without the consent of those whose data was scraped.[12] The appetite for the deeply personal information that dating websites have about their users has fueled major data breaches of at least five dating sites in the United States, Japan, and South Korea in July 2020, involving millions of records around the world – both public and nonpublic. The line between commercial sale of scraped data on the gray market and illegal hacking of nonpublic data is not a clear one. Web scraping is often the first step for malicious hackers who use it to find additional sensitive information like real names, private messages, phone numbers, and billing addresses that websites may have failed to protect properly on their servers.[13]

Even noncommercial scraping has raised serious ethical problems. In one case, critics of social media created what appeared to be a dating website – Lovely-Faces.com – with 250,000 scraped profiles of Facebook users. The fake dating site included a facial recognition algorithm classifying those users into categories like "smug women" or "climber men." The purpose, its

---

[12] Samantha Cole, "Shady Data Brokers Are Selling Online Dating Profiles by the Millions," *Vice*, Nvo. 12, 2018, available at https://www.vice.com/en/article/59vbp5/shady-data-brokers-are-selling-online-dating-profiles-by-the-millions

[13] Kate Hawkins, "Data Breach: Millions of Dating App Records, Messages, and User Profiles Exposed in Data Leak," *WizCase (blog)*, last updated June 19, 2022.

creators said, was to protest Facebook's impact on privacy – but the site used real names and real photos, all without consent.[14] Academic researchers have also turned to dating websites to harvest personal information without consent, arguing – wrongly – that the fact that such data is easily available and arguably public means there are no ethical issues with using it for research. "OkCupid is an attractive site to gather data from," one researcher remarked, saying other researchers should use their dataset of 70,000 scraped user profiles "for their own purposes." The dataset included details like usernames, sexual orientation, drug use, and sexual turn-ons like whether a user would enjoy being "tied up" during sexual activity. University review boards have unfortunately been slow and inconsistent in how they apply existing ethical guidelines to such practices.[15] Dating photos and profiles scraped without permission have been harvested and traded for training AI algorithms, over the objections of dating websites. One AI researcher – after taking down a set of 40,000 photos he had uploaded to the internet at the request of Tinder – nevertheless uploaded his "Tinder Face Scraper" to the code-sharing site GitHub, describing it as a "simple script that exploits the Tinder API" so that anyone might "build a facial dataset."[16]

---

[14] Ryan Singel, "'Dating' Site Imports 250,000 Facebook Profiles, Without Permission," *Wired*, Feb. 3, 2011, available at https://www.wired.com/2011/02/facebook-dating/.

[15] Joseph Cox, "70,000 OkCupid Users Just Had Their Data Published," *Vice*, May 12, 2016, available at https://www.vice.com/en/article/8q88nx/70000-okcupid-users-just-had-their-data-published

[16] Mary Papenfuss, "Massive Tinder Photo Grab Is Latest Scary Warning To Be Careful What You Post," *Huffington Post*, Apr. 30, 2017, available at https://www.huffpost.com/archive/in/entry/massive-tinder-photo-grab-is-latest-scary-warning-to-be-careful_in_5c10eef2e4b085260ba71105; Natasha Lomas, "Someone scraped 40,000 Tinder selfies to make a facial dataset for AI experiments," *TechCrunch*, Apr. 28, 2017, available at https://techcrunch.com/2017/04/28/someone-scraped-40000-tinder-selfies-to-make-a-facial-dataset-for-ai-experiments/

The challenge of AI

The demand for scraped data has become even greater with the increasing sophistication of new digital services based on Artificial Intelligence/Machine Learning (AI). ChatGPT and other generative AI services depend on assembling large quantities of data to feed the AI algorithms that create new content. Large Language Models (LLMs) require vast quantities of unstructured text – the more recent and natural, the better. The demand for such data has flooded websites with scraping traffic.[17] Generative AI raises legal, policy, and ethical questions because the choice of what datasets are fed into such models seriously affect what comes out of them. There are serious privacy, copyright, and fairness problems with many existing AI tools. AI tools can reflect and amplify societal bias, including racial, gender, and other harmful biases; result in inaccurate or unfair decisions that are based on opaque criteria; and even create "hallucinations" – plausible, but entirely false images or text presented as fact.

Because AI models have such a voracious appetite for data, including personal and sensitive data, they pose severe risks to privacy. Private data that is used to train AI tools may be exposed by AI-generated content. Apart from such obvious privacy harms, the incentives created by the demand for AI tools has further eroded norms against uncontrolled scraping. Programmers who once respected a website's "robots.txt" file or a TOS that limits scraping are

[17] Emanuel Maiberg, "An AI Scraping Tool Is Overwhelming Websites With Traffic," April 25, 2023, available at https://www.vice.com/en/article/dy3vmx/an-ai-scraping-tool-is-overwhelming-websites-with-traffic

under increasing pressure to ignore such requests in favor of ingesting the data their bosses demand in order to train an AI tool they hope to bring to market.

These trends pose real challenges for cybersecurity. Aggressive and unwanted scraping often violate basic security principles, bypassing authentication, authorization, and access controls that website owners use to safeguard their systems.

II. Cybersecurity, web scraping and the "three A's": authentication, authorization, access control

The central goal of cybersecurity can be summed up in the "CIA triad." Computer, systems, and information security all seek to preserve the confidentiality, integrity, and availability of data. A close cousin to the CIA triad is the "three A's" – authentication, authorization, and access control – which are critical tools for achieving these goals.

Why authentication, authorization, and access control are central to cybersecurity

NIST's influential cybersecurity framework covers the three A's under the function of "protect," listing a myriad of technical references for authentication, authorization, and access control under the category of "identity management and access control" to aid users of the framework

in adopting technologies to safeguard their systems.[18] While the three A's are most closely associated with the topic of identity credential and access management, they are fundamental to virtually every technical aspect of cybersecurity. The three A's matter even in areas having nothing to do with a particular user's identity or a standard credential scheme. Importantly, the concepts of authentication, authorization, and access control have an important role to play even when a user's identity is not known.

While there is no uniform definition of these concepts, references to technical standards in the NIST Cybersecurity Framework provide a helpful starting point.

- Authentication. "Verifying the identity of a user, process, or device, often as a prerequisite to allowing access to resources in an information system."[19]

- Authorization. "A decision to grant access, typically automated by evaluating a subject's attributes."[20]

[18] See NIST Framework 1.1, April 16, 2018, at pp. 29-31, available at https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf (referencing FIPS 200)

[19] NIST Special Publication 800-53 revision 4, available at https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf

[20] NIST SP 800-63-3, available at https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-3.pdf

- Access control. "An automated system that controls an individual's ability to access one or more computer system resources such as a workstation, network, application, or database."[21]

Note that the latter term, access control, is generally limited to situations in which users have already been authenticated and authorized, and so is generally less relevant to the topic of this paper.

Common misunderstanding #1 – authentication and authorization are basically the same thing

While authentication and authorization are related, they are distinct concepts.

Technical documentation makes clear that the concepts of authentication and authorization are distinct and should be considered separately. For example, Pluggable Authentication Modules (PAMs) are used to allow applications to manage user accounts in the Linux operating system and have been part of Linux since 1997. This technical process manages how software determines, for example, whether a username and password combination are valid and up-to-date, despite the many different ways that such information might be stored and handled.[22] The documentation for Linux is careful to distinguish between authentication and authorization

[21] Note that this is a definition of "logical access control system," one of several definitions of specific types of access control in NIST Special Publication 800-53 revision 4 at p. B-13, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf

[22] Susan Lauber, *An introduction to Pluggable Authentication Modules (PAM) in Linux*, Red Hat, July 22, 2020, available at https://www.redhat.com/sysadmin/pluggable-authentication-modules-pam

in discussing how programmers should employ PAMs. "Note that authentication and authorization are two separate processes," a commonly-used guide reads.[23] The guide defines authentication as "the process of confirming an identity" through such methods as logins and passwords, cryptographic certificates, and physical tokens, while authorization "defines what the authenticated party is allowed to do or access."[24]

One reason for the confusion is that the terms authorization and authentication are functional terms – they describe what a particular technical process is trying to accomplish, rather than the specific process used to accomplish that goal. Authentication is usually defined as establishing an "identity," and while this could be that of a particular (human) user, it could just as easily be of a device, system, or piece of software. Authorization is defined as the decision of what privileges to grant to that user, device, system, or software.

Common misunderstanding #2 – authentication and authorization are synonymous with a login process that allows a computer to limit access to a particular user with an account

Logins and passwords are, of course, a very common form of authentication, but authentication is not synonymous with logins and passwords. While a login process that combines a user ID with a password – increasingly coupled with some form of multifactor authentication (MFA) – is

[23] Red Hat Enterprise Linux 7 System-Level Authentication Guide at p. 8, last updated April 25, 2022, available at https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/pdf/system-level_authentication_guide/red_hat_enterprise_linux-7-system-level_authentication_guide-en-us.pdf

[24] *Id.* at p. 4. See also https://www.okta.com/identity-101/authentication-vs-authorization/

the form of authentication with which most computer users are familiar, authentication is a more general term for a process that can be accomplished in a myriad of ways. In fact, newer forms of authentication are likely to abandon familiar username and password logins in favor of authentication that relies on processes that operate in the background. In fact, passwords are themselves a key vulnerability; compromised passwords are involved in 81% of breaches.[25] Passkeys and other forms of password-less authentication are being rolled out by big companies like Google as well as start-ups like BastionZero.[26]

Not all forms of authentication are based on credentials. For example, challenge-response authentication is a term used to describe the process in which a question and the correct response provides security for a resource or system. A username and password combination is one example of a challenge-response authentication system that is based on user identity; security questions are another. Authentication need not be based on user identity, however. CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) are a form of challenge-response authentication, designed simply to determine whether the respondent is a human or a bot cleverly designed to mimic human behavior. Authentication can also be based on a Zero Knowledge (ZK) proof, a cryptographic method of proving that you

[25] Matthew Tyson, "How passkeys are changing authentication," *CSO Online (blog),* Jan. 24, 2023, available at https://www.csoonline.com/article/3685933/how-passkeys-are-changing-authentication.html.

[26] Michael Hill, "BastionZero releases SplitCert for password-free authentication and access," *CSO Online (blog)*, June 8, 2023, available at https://www.csoonline.com/article/3698296/bastionzero-releases-splitcert-for-password-free-authentication-and-access.html; Michael Hill, "Google rolls out passkey support across accounts on all major platforms," *CSO Online (blog)*, May 3, 2023, available at https://www.csoonline.com/article/3695173/google-rolls-out-passkey-support-across-accounts-on-all-major-platforms.html.

possess an attribute – such as a validly-issued credential that shows you are above a certain age – without revealing any other information that might reveal your identity.

Put simply, a common shorthand for authentication – a process designed to show "you are who you say you are" – can be misleading. Many authentication methods exist that do not even attempt to establish a user's identity, but merely to show that someone (or some device) that is attempting to connect with your system – such as a website – possesses an attribute or meets criteria that the owner of the system has established. In other words, these methods do not reveal identity because the owner of the system does not require that you say who you are – but they are forms of authentication, nevertheless. If the authentication step is satisfied, the system may then decide to authorize access to some resources (and deny access to others).[27]

To be sure, much of the open web is just that – open – and requires neither authentication nor authorization. As one resource explains, "In some cases, there is no authorization; any user may use a resource or access a file simply by asking for it. Most of the web pages on the Internet require no authentication or authorization."[28] This has consequences for scraping under the CFAA, as discussed below.

[27] Linda Rosencrance, "Definition: challenge-response authentication," *Tech Target*, available at https://www.techtarget.com/searchsecurity/definition/challenge-response-system

[28] See Boston University TechWeb, available at https://www.bu.edu/tech/about/security-resources/bestpractice/auth/

Common misunderstanding #3 – authentication must invariably come before authorization (and without authentication, authorization is not a meaningful concept)

As discussed above, the three "A's" are commonly considered by security professionals as part of an overall system typically involving credential management. As a result, the habit of many security professionals is to think of them in chronological order – first authentication, then authorization, and finally access control. This habit is strong enough that many informational resources for security may discuss authentication as if it were a necessary prerequisite to authorization.[29] It is more accurate to say, as one resource puts it, that "[a]uthorization is usually coupled with authentication." It is not the case that authorization can *never* occur without authentication, simply that the two processes usually proceed together. Authorization describes the process in which a server chooses to grant or not to grant access to a file or other resource. While it is typically "coupled" with authentication "so that the server has some concept of who the client is that is requesting access," a technical process could just as easily authorize access (or withhold authorization) based on whatever criteria a programmer chooses to use.

[29] See, e.g., BU TechWeb, available at https://www.bu.edu/tech/about/security-resources/bestpractice/auth/, Dave Piscitello, "What is Authorizaton and Access Control?" Dec. 2, 2015, available at https://www.icann.org/en/blogs/details/what-is-authorization-and-access-control-2-12-2015-en , Priocept, "Authentication vs. Authorization vs. Access Control," available at https://priocept.com/2011/08/30/authentication-vs-authorisation-vs-access-control/ , Robert Roohparvar, "Why You Need Both Authorization and Authentication," *Infoguard Cyber Security*, Jan. 27, 2020, available at https://www.infoguardsecurity.com/why-you-need-both-authorization-and-authentication/

The important point is that a decision to grant access to a resource is (by definition) an authorization decision. While such a decision often occurs after authentication, it may occur before or even without authentication.[30] The common scenario is for authentication to occur first, followed by authorization, but there are examples in which a system must authorize (or fail to authorize) a request despite a failure of authentication. These are described in the technical literature as examples of "authorization."[31] Not only is it theoretically possible for authorization to occur without (or before) authorization there are also many common situations in which this actually occurs. A familiar example is the use of rate-limiting logins to prevent brute-force attacks on passwords. Anyone who has been locked out of their account because they exceeded the number of "wrong password" attempts has run up against a system that was configured to deny *authorization* to a particular user *that has not yet been authenticated* – indeed, the point of this particular denial of authorization is to protect the authentication process itself.[32]

How web scrapers bypass common methods of authentication, authorization, and access control

[30] Another way of thinking about the relationship between authentication and authorization is to define authentication to include any process for verifying the criteria that a programmer has chosen to make an authorization decision. Under this definition of authentication, it would be correct to say that authentication always precedes authorization, but only because authentication has been defined so broadly that it would include whatever step is required to make an authorization decision.

[31] *Id.* at pp. 80-81.

[32] See Stavros' Stuff, "Authentication and rate limiting" (blog post), Sept. 16, 2013 https://www.stavros.io/posts/authentication-and-rate-limiting/

Website owners who wish to discourage scrapers have a variety of methods at their disposal. Owners may ban scraping in their TOS, discourage web crawlers from some portions of their sites in their "robots.txt" files, or take legal steps – like sending a cease-and-desist letter to a company engaged in persistent unwanted scraping – to communicate that they do not want their sites to be scraped. Unfortunately, none of these steps prevents web scrapers from ingesting data. To actually stop scrapers, website owners must institute technical countermeasures.

Technical countermeasures to prevent scraping include blocking IP addresses associated with scraping bots, using CAPTCHAs, or dynamically changing elements of the website's HTML code in ways that frustrate scrapers.[33] Many anti-scraping countermeasures meet the technical definitions of authentication, authorization, or access control – even though the website has not limited access using a login based on a username and password. One example is rate limiting. Rate limiting is a security measure that limits how often a server will respond to a particular request, such as attempts to log in to an account or requests for a particular webpage. It is an essential mitigation against denial-of-service (DoS) and distributed denial-of-service (DDoS) attacks, brute force attacks to guess passwords, and unwanted web scraping. Rate limiting solutions track the IP addresses of incoming requests to determine if they exceed a particular

[33] See Cloudflare, "What is data scraping?" *supra.* See also "How to safeguard valuable data from malicious web scraping," PWC White Paper, available at https://explore.pwc.com/safeguard-data#page=1 Perhaps ironically, this white paper includes the contact information of the authors, making it a tempting target for web scrapers.

threshold within a given time period. If they do, the application will refuse to fulfill requests from those IP addresses from some period of time.[34]

Anti-scraping countermeasures may meet the technical definition of authentication quoted above. In essence, such measures are forms of "[v]erfying" that a "user, process, or device" does not belong to a particular class that the website owner has chosen to exclude. This is a form of negative identification – a process of determining that you are not a person, entity, or device on a banned list – and it serves "as a prerequisite to allowing access to resources in an information system" – the definition of authentication. The purpose of this step is to inform "a decision to grant access, typically automated by evaluating a subject's attributes" – the definition of authorization.

III. Scraping and the Computer Fraud and Abuse Act (CFAA): why the courts are confused

Federal courts have been confused for years about how to apply the main federal anti-hacking statute – the Computer Fraud and Abuse Act (CFAA) – to a variety of unwanted computer behaviors. The CFAA traces its origins to the mid-1980s when the Reagan Administration confronted the very new problem of malicious hacking into sensitive computer systems. Although it has been amended many times in subsequent years, the language of the statute's

[34] Cloudflare, "What is rate limiting?" available at https://www.cloudflare.com/learning/bots/what-is-rate-limiting/

principal provisions – which make unauthorized intrusions into computer systems both a crime and a basis for a civil lawsuit – have remained largely the same.

The confusion stems not only from the wording of this very outdated statute, but it has been compounded by misunderstandings that result from lawyers reading the technical terms used in the CFAA to refer to legal concepts with which they are familiar rather than the technical concepts to which the statute refers.

The CFAA and the <u>two</u> "A's" – "authorization" and "authorized access."

The CFAA defines as criminal two forms of conduct – "access without authorization" or actions that "exceed authorized access" – that are sometimes described as "computer trespass." See 18 U.S.C. § 1030(a)(2). The "trespass" label, however, can be misleading. Trespass is a legal concept with a rich history and many rules in physical space, making it catnip for lawyers and judges who are deeply versed in the common law of trespass but who do not understand the very different world of computers. At best, notions of trespass can serve as an analogy to malicious cyber activity – and may serve as easily to mislead as to clarify. In fact, the word "trespass" appears nowhere in the CFAA. Note also that the CFAA conspicuously lacks one of the three "A's." Nowhere does the statute mention authentication, an oversight that stems from its history. In the 1980s, when the CFAA was originally enacted, it applied only to specialized computer systems – such as military systems – that were generally "air-gapped" and therefore unavailable to remote users.

Instead, 18 U.S.C. § 1030(a)(2) provides that whoever "intentionally accesses a computer without authorization or exceeds authorized access" has committed an offense under the CFAA.[35] The term "authorization" is not defined by the statute.[36] Still, in the context of a computer security statute, it should have been clear that this term is meant in its technical sense. The terms "authorization" and "access" have long been common computer security terms with meanings that are clear to computer security professionals. Unfortunately, these terms are not necessarily well known in their technical sense to ordinary computer users – like lawyers and judges. Even worse, "authorization" and "authorized access" are common words that may be used in ordinary life, or they may refer to specialized legal concepts – like trespass.

For decades, lawyers and judges struggled to provide the missing definition of "authorization."[37] Two main approaches developed. The broader approach defined "authorization" in legal terms: the owner of a computer system grants authority to access the system subject to whatever terms and conditions that owner has chosen to specify. For websites, those terms are contained in the TOS, while employers might specify them in computer use policies. One example of this approach is *Craigslist v. 3Taps* – an early web scraping case in which Craigslist

[35] These provisions of the CFAA prohibit using such access to obtain information from a "protected computer," a term that has now been expanded to cover any computer connected to the internet as well as most (if not all) that are not.

[36] The term "exceeds authorized access" is defined by the statute, but this definition still depends on the (undefined) concept of authorization. See 18 U.S.C. § 1030(e)(6) ("the term 'exceeds authorized access' means to access a computer with authorization and to use such access to obtain or alter information in the computer that the accessor is not entitled so to obtain or alter").

[37] See Orin Kerr, "Norms of Computer Trespass," 116 *Columbia Law Review* 1143 (2016).

sought to prevent another company from scraping its site in violation of its TOS.[38] Craigslist's argument was that it had conditioned access to its site on compliance with its TOS, and that by violating the TOS, 3Taps was violating that legal condition and therefore acting without authorization. The trouble with this approach was not necessarily that it involved scraping – which might be quite invasive – but that by relying on a legal understanding of the term "authorization," it threatened to make any simple TOS violation a federal crime. A narrower approach was adopted by the United States Court of Appeals for the Ninth Circuit in *United States v. Nosal*.[39] In *Nosal*, the Ninth Circuit sought to ground "authorization" in the concept of hacking – access without or in excess of authorization simply described the process of circumventing a "technological barrier" embedded in the computer code itself.

The result was legal uncertainty. Interpretations of the CFAA varied substantially by geography, as different federal courts used different approaches depending on how they understood the various tests adopted by their circuit courts of appeals. For more than two decades, this debate over the CFAA pitted fears of prosecutorial overreach, which counseled in favor of a narrow approach, against the need for a flexible statute that would cover an ever-evolving variety of malicious activities online.

The Supreme Court speaks – reading technical terms to have technical meanings

[38] *Craigslist, Inc. v. 3Taps, Inc.*, 942 F. Supp. 2d 962 (N.D. Cal. 2013).

[39] 676 F.3d 854, 863-64 (9th Cir. 2012).

In 2021, the Supreme Court finally resolved the debate by adopting a narrower interpretation, consistent with the technical concept of authorization instead of the legal one.[40] In *Van Buren v. United States,* Justice Barrett's opinion for the court parsed the text of "exceeds authorized access." She determined that an authorized user only violates the statute if he bypasses a clear "gate." In other words, "exceeds authorized access" means to access data or "areas of the computer" that are clearly off-limits for that user. Access to data with intent to misuse it does not count. The *Van Buren* approach is a binary one: whether access is authorized is a "gates-up-or-down" inquiry. The Supreme Court made clear that when statutes use technical terms, they should be given their technical meaning. It was a "eureka" moment for reading the CFAA. Still, *Van Buren* left some questions unanswered. One such question was whether an authorization "gate" is necessarily one that is guarded by computer code, or whether a clear enough "no-access" policy was enough. Another was whether *Van Buren's* "gates-up-or-down" approach to "exceeds authorized access" – which typically applies to insiders – also applies in the cases of access "without authorization" – which applies to outside hackers.

Unfortunately, the Supreme Court sidestepped these questions when it decided not to move forward with argument in another CFAA case, *LinkedIn v. HiQ Labs*, discussed above. The case began when HiQ Labs sued LinkedIn for instituting countermeasures against its aggressive scraping behavior. HiQ Labs argued that LinkedIn's countermeasures were a form of illegal anticompetitive conduct, and that it was in the public interest for the court to order LinkedIn to

[40] *Van Buren v. United States*, 593 U.S. __, 141 S. Ct. 1648 (2021).

stop. LinkedIn responded that it was within its rights to defend its website from scraping because HiQ Labs had violated its TOS and ignored its legal demand letters. According to LinkedIn, HiQ Labs had accessed its website "without authorization" in violation of the CFAA. The Ninth Circuit, applying its narrow approach to the CFAA, had sided against LinkedIn – a violation of LinkedIn's TOS was not a violation of the CFAA. The Supreme Court agreed to hear LinkedIn's petition for review of the Ninth Circuit's decision. Briefs were filed and argument was set. After deciding *Van Buren*, however, the Supreme Court changed course, sending the case back to the Ninth Circuit for reconsideration. On remand, HiQ Labs argued that the Supreme Court's decision only strengthened its position. HiQ Labs said that *Van Buren*'s "gates-up-or-down" approach meant that its scraping behavior could never violate the CFAA. According to this argument, *Van Buren* showed that "authorization" was governed not by legal policies but by technical gates, and technical gates meant logins and passwords. Therefore, the public profiles of LinkedIn's users were fair game. LinkedIn – supported by privacy organizations – countered that it had done more than rely on its TOS or issue legal demands that HiQ Labs stop scraping – it had enacted countermeasures, such as an IP block, that served as a "gate," just as *Van Buren* had said the CFAA demanded.

Nevertheless, the Ninth Circuit still decided against LinkedIn. In doing so, the judges showed that the idea of the CFAA as a trespass statute – and the misunderstanding of "authorization" as a legal concept – was alive and well, even after *Van Buren*. As the court explained, the CFAA "is best understood as an anti-intrusion statute and not a 'misappropriation statute.'" As the Ninth Circuit put it, the term "authorization" in the CFAA depends on a distinction between "'open

spaces'" and "'closed spaces'" rather than on the rules provided by LinkedIn's TOS.[41] Since HiQ Labs had not intruded on any of LinkedIn's "closed spaces," it could not have violated the CFAA. Instead of examining whether LinkedIn's technical measures to block scraping and bot activity constituted a form of authentication, authorization, or both, the Ninth Circuit simply equated the idea of "authorization" with accessing "closed spaces," an idea it limited to websites that are "password-protected." Only a website that requires a user login and password, the court reasoned, was a closed space that could be accessed "without authorization."

The trouble with this view, of course, is that it simply substitutes a new misunderstanding of authorization – a trespass-based understanding – for an older misunderstanding that was based on contract law. This is simply not how the Supreme Court said courts should interpret the CFAA. "[W]hen a statute, like this one, is addressing a technical subject, a specialized meaning is to be expected," the Supreme Court observed, so "our interpretation tracks the specialized meaning of 'access' in the computer context."[42] Indeed, the Ninth Circuit actually acknowledged that it was ignoring the Supreme Court's admonition to interpret the CFAA in technical rather than legal terms, saying that while "access" might have a technical definition, the term "authorization" would be given what it called its "'plain and ordinary'" meaning. The Ninth Circuit's interpretive choice was a strange one – effectively giving the same word ("authorization") two different meanings within the CFAA, and even within the same sentence

[41] *HiQ Labs v. LinkedIn*, slip op. at 31 (quoting *United States v. Nosal (Nosal I),* 676 F.3d 854, 857-58 (9th Cir. 2012); *id.* at 33 (quoting Orin S. Kerr, *Norms of Computer Trespass*, 116 Colum. L. Rev. 1143, 1161 (2016)).

[42] *Van Buren*, slip op. at 12 n.

of that statute. The Ninth Circuit accepted that "exceeds authorized access" has a specialized, technical meaning – as it was bound to do by the Supreme Court's decision in *Van Buren* – while nevertheless persisting in its view that "access without authorization" does not – instead, it has the meaning that seems "'plain and ordinary'" to lawyers and judges.[43] It was one of many examples of how lawyers and computer scientists talk past each other. Logins and passwords are a form of authentication, and they generally go together with authorization, but they are not the only form of either authentication or authorization. They are simply the most visible to ordinary computer users, and one with which judges and lawyers are familiar.

The Ninth Circuit's *LinkedIn* decision offers two paths forward for the federal courts. If other courts reject its strained reasoning, different rules will apply to web scraping depending on where unauthorized scrapers end up in court. The pre-*Van Buren* circuit split that created such uncertainty around the CFAA will be resurrected in a new context. While the Supreme Court would likely resolve such a split at some point, there would be years and potentially decades of confusion in the lower federal courts about what kinds of scraping are allowed under the CFAA. The harms would be considerable. In permissive jurisdictions like the Ninth Circuit, web scraping would be unregulated – allowing beneficial and innovative forms of web scraping, but underprotecting against harmful scraping. In other jurisdictions, courts may adopt restrictive rules that bar virtually all scraping to which a website owner objects, chilling research and other beneficial scraping. Such a result potentially offers the worst of both policies.

[43] *Id.* at 29 & n.13.

Yet if other courts follow the Ninth Circuit's lead, any data that is made available without a login-and-password barrier will lack legal protection under the CFAA, even from the most aggressive and malicious forms of web scraping. Essentially, unauthorized web scraping will be read out of the CFAA. This path does offer the benefits of a clear rule, removing the current disincentives that researchers who use scraping for beneficial purposes may face. Such clarity would come at a price, however. Because the United States lacks a comprehensive privacy law, a weakening of existing legal protections against unwanted scraping of personal data would mark another step in the journey towards a "wild west" for exploitation of such data. The United States could easily become a magnet for the wrong sort of technological innovation – becoming a haven for scammers and surveillance technology.

IV. Legislative Solutions for Harmful Scraping

As discussed in the previous section, as a result of the Ninth Circuit's decision in *LinkedIn v. HiQ Labs*, the federal courts are unlikely to provide reliable protection against unauthorized web scraping – at least in the near and medium term. Even if other federal courts reject the Ninth Circuit's approach, legal protection against even the most aggressive forms of unauthorized scraping will be uncertain and vary based on geography: whether a website owner can obtain legal redress or even deploy some forms of protection will depend on which federal court is likely to hear a scraping case. Addressing the social harms of harmful scraping requires Congress to act. Congress could do so by amending the CFAA, by addressing the problem of harmful scraping in comprehensive privacy legislation, or by doing both.

Amending the CFAA – possibilities, problems, and limits

One possible solution is to amend the Computer Fraud and Abuse Act to clarify that unauthorized scraping, or at least the most harmful and aggressive forms of unauthorized scraping, is a form of "access without authorization" even after the Supreme Court's *Van Buren* decision. Such an amendment could take two forms. First, Congress could define authorization in legal terms, effectively reversing the Supreme Court's *Van Buren* decision. Second, Congress could provide a technically-sound definition of authorization that clarifies the role of authentication – protecting the right of website owners to use standard practices like rate-limiting and CAPTCHAs to prohibit bots and scrapers from harvesting data they have chosen to protect – with appropriate exceptions for beneficial uses of scraping.

While the first approach has some advantages, these are far outweighed by its downsides. Defining authorization in legal terms would give teeth to terms of service (TOS), computer use policies, and similar documents. Such a broad CFAA would be a stronger tool against malicious insiders and would permit website owners to deter scraping by banning it in their TOS, even without deploying technical countermeasures. Yet simply redefining authorization as a legal rather than a technical term would resurrect all the very real and serious civil liberty problems of an overbroad federal hacking law. Giving TOS and other "click-through" policies the force of law – a criminal law at that – would give websites too much power and employers far too potent a weapon against employees who might violate minor computer use policies such as

checking personal websites at work. Website owners could ban any scraping for any reason, including anticompetitive reasons. It would also chill legitimate security research.

The second approach is more promising. The CFAA's failure to define "authorization" has led to decades of confusion in the federal courts – confusion that even survived the Supreme Court's statement in *Van Buren* that favors a technical approach. The role of authentication, in particular, needs clarification. Properly defined, authentication would cover most anti-scraping countermeasures which attempt to distinguish human users from bots and scrapers. Such a law would not permit website owners, employers, or other owners of computer networks to make violation of any TOS or computer use policy a crime or a basis for a federal lawsuit – only circumvention of technical countermeasures, including anti-scraping countermeasures designed to protect against malicious activity. The law could also clarify that such countermeasures cannot be deployed for anticompetitive purposes and could create a safe haven for legitimate security research and other ethical forms of scraping for noncommercial purposes.

There are significant limits to any CFAA-based approach to harmful web scraping, however. The CFAA's purpose is to protect the rights of the owners of websites and other computer systems against unwanted behavior such as hacking. Harmful web scraping may have a much greater impact on those whose personal information is scraped – such as the users of dating websites – than it does on the websites themselves. A CFAA-based approach only indirectly protects the interests of such users, and only insofar as user interests coincide with the interests of platforms

and other website owners. Only comprehensive legislation that protects privacy and personal data can offer data subjects effective remedies against the misuse of personal data.

Comprehensive privacy legislation and the problem of the public/private distinction.

By far the most significant and influential legal framework for the protection of privacy and personal data is the General Data Protection Regulation of the European Union (GDPR).[44] It provides that anyone who collects personal data of an EU resident – defined as either a data controller or a data processor – must do so according to a valid legal basis, such as the consent of the data subject.[45] Data subjects have certain rights, including the rights of notice, deletion, correction, and redress.[46] Any business operating in the European Union must comply with the GDPR, and data of EU residents may not be transferred outside the EU unless the laws of the jurisdiction to which the personal data is transferred provide an "adequate level of protection," defined as "essentially equivalent" to the rights provided by GDPR.[47] Penalties are stiff,

[44] General Data Protection Regulation, Regulation (EU) 2016/679 (English), available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

[45] See, e.g. GDPR art. 5 ("Personal data shall be . . . collected for specified, explicit and legitimate purposes and not further processed in a manner incompatible with those purposes . . . .") and art. 6 (listing purposes, such as consent, for which processes is lawful).

[46] GDPR chapter III.

[47] GDPR art. 3 (territorial scope); art. 45 (requiring an "adequate level of protection" for transfers of personal data outside of the EU); *Schrems v. Data Protection Commissioner,* Case C-362/14, ECLI:EU:C:2015:650 at ¶ 96 (defining "adequate level of protection" to require "essentially equivalent" legal protections for personal data).

amounting to as much as 4% of an organization's total worldwide revenue or €20 million, whichever is larger.[48]

An entity that scrapes personal data of EU subjects must have a valid legal basis for such scraping. In the absence of consent – which is unlikely in the scraping scenario – it will be difficult if not impossible to find such a legal basis. Significantly, GDPR does not exempt personal data from its protection merely because it is publicly available. Such public availability may affect the data subject's rights and the level of protection provided by GDPR, but personal data does not have to be private to enjoy protection under GDPR.

The United States lacks comprehensive privacy legislation at the federal level; the closest analog to the EU's GDPR is the California Consumer Privacy Act (CCPA).[49] The CCPA offers some, but by no means all, of the protections of GDPR. Most importantly, CCPA exempts data that is "publicly available" from its protection.[50] Congress has been considering comprehensive federal privacy legislation for years. The American Data Privacy and Protection Act (ADPPA), a bipartisan compromise that seeks to reconcile the interests of industry and privacy advocates, has progressed farther than any other proposal, receiving the overwhelming support of the House Energy and Commerce Committee at the end of 2022, but expiring at the end of the last

[48] GDPR art. 83.

[49] California Consumer Privacy Act of 2018, Calif. Civ. Code 1798.100-1798.199.100. The CCPA is sometimes referred to as the California Privacy Rights Act (CPRA), adopted by ballot initiative in 2020. Strictly speaking, the CPRA is an amendment to the CCPA, so this paper refers to the law as the CCPA.

[50] Calif. Civ. Code 1798.140(v)(2) ("'Personal information does not include publicly available information . . . .'").

Congress. Like the CCPA, the ADPPA does not protect personal data that is publicly available,[51] limiting its utility in addressing the problems of harmful web scraping.

Outside the United States and the European Union, approaches to data privacy and the protection of personal data vary widely. The EU's comprehensive approach is the most influential as it is far easier for countries to do business with the EU if they have adopted EU-style data protection legislation. Still, many Asian nations have adopted privacy guidelines that offer a narrower approach than the EU, requiring data subjects to show concrete harm before they may obtain remedies for the misuse of their data.[52] China provides formal legal protection for personal data, but its authoritarian leaders have also set their sights on creating a technologically sophisticated data surveillance-based society. Programs like China's "social credit" system comprehensively monitor citizen behavior. China's cyber strategy touts its laissez-faire approach to privacy as an advantage, powering innovations in AI through vast stores of personal information.

Any legislation that seeks to address the problems of harmful web scraping will have to address the global nature of the problem. Malicious bots, scrapers, and scammers operate across jurisdictional and national boundaries. While data subjects should enjoy legal rights and remedies against misuse of their data – and governments should cooperate to provide such

[51] American Data Privacy and Protection Act, H.R. 8152 § 2(8)(B), Rep. No. 117-669 (Dec. 30, 2022) (excluding "publicly available data" from the definition of "covered data").

[52] See generally, "The Asia Pacific Privacy Guide 2020-21," Deloitte (December 2020), available at https://www2.deloitte.com/ph/en/pages/risk/articles/asia-pacific-privacy-guide.html

rights – an effective approach to harmful web scraping must also empower website owners to deploy technical countermeasures against harmful scraping.

Addressing the promise and peril of AI

As discussed above, there has been intense public attention to the potential of Artificial Intelligence (AI) as a result of the release of ChatGPT – a software tool that generates synthetic text based on user prompts. ChatGPT uses a specific form of machine learning (ML) – a large language model (LLM) – to produce its results. AI tools like ChatGPT require large datasets to train ML algorithms to produce results that are realistic enough to appear authentic to users. Because algorithms train on datasets, such as online chats, text, or images, in order to create synthetic content, AI software requires the ingestion of large quantities of real data in order to refine their models and produce the best outputs. Many AI projects rely on collecting web data through scraping, open APIs, or other methods. The interest surrounding AI has fueled investor interest and engendered market pressure to produce dramatic advances in AI – and products that capture press attention. This pressure, in turn, has created an insatiable appetite for data, including scraped data, making it difficult to hold the line against aggressive, unethical, and even harmful forms of data scraping.

Proposed ethical and regulatory frameworks for AI have mostly focused on the potential harms that might be caused by the use of unregulated AI tools. Some have even called for a pause in AI research, worrying that AI may be close to developing a form of general intelligence that

poses profound and unanticipated dangers to human society. More immediate concerns involve the fairness of AI algorithms, discrimination, accuracy, and the use of AI for misinformation and disinformation. Largely absent from the discussion have been the ethical and regulatory problems of using personal data – often scraped from social networks – to power new digital services for profit without the consent of either websites or their users. No discussion of the ethical or policy challenges posed by AI is complete without addressing the concerns about privacy and security discussed in this paper.

V. Conclusion

If harmful web scraping has received comparatively little attention as a cybersecurity threat, that is likely to change. The demand for scraped data is increasing even as the norms, laws, and regulations that have limited such scraping have eroded. Website owners should be incentivized – or at least not deterred – from using technical measures to limit scraping that could impact the privacy and security of users and others whose data could otherwise be at the mercy of malicious bots, scrapers, and scammers. A useful first step in addressing the problem of harmful web scraping is for lawyers, policymakers, and computer security experts to stop talking past one another. Terms like authentication, authorization, and access control are technical terms. They should be given their technical meaning.

Policymakers in the United States and elsewhere should take steps to prevent harmful scraping, while ensuring appropriate exceptions for scraping for valid commercial purposes and for

legitimate and ethical research. To be effective, solutions to the problem of harmful scraping must transcend particular jurisdictions and legal systems. Both technical countermeasures and laws that are properly informed by technical understanding are critical to addressing harmful web scraping.