PERSONAL INFORMATION AND ARTIFICIAL INTELLIGENCE:

WEBSITE SCRAPING AND THE CALIFORNIA CONSUMER PRIVACY ACT

BRIAN STUENKEL*

This note presents a hypothetical in which an upstart technology firm scrapes public-facing webpages and websites, scooping up individuals' personal identifying information (PII) including names, addresses, phone numbers, and dates of birth (among other information) in the process. In this hypothetical scenario, the tech firm then uses the information gathered to create a dataset containing the PII. The upstart tech firm then uses the dataset to train artificial intelligence (AI) tools. The upstart tech firm is subsequently acquired by a larger software company. Included in the acquisition are both the AI model and the training dataset to be used within the acquiring company's own "code stack" or software product suite.

In the changing landscape of data privacy laws, this note seeks to answer several questions: First and foremost, is website scraping prohibited by the California Consumer Privacy Act (CCPA)? Second, are data scrapers required to notify consumers under the CCPA and if so, how? Third, what legal obligations may an acquiring company have after receiving the information upon acquisition? Finally, what liability may the acquiring company face in acquiring, storing, or disclosing the PII, or in failing to notify individuals whose information was collected by the small AI firm it acquired?

This note considers a particular scenario, however the scenario is designed to highlight some of the problems with an expansive new law, as well as a lack of federal regulation regarding the use of

* J.D. Candidate, University of Colorado Law School

specific types of information when placed in the public sphere without additional safeguards. The state law under which this note seeks to address these questions is the California Consumer Privacy Act, or CCPA.

The CCPA and California Privacy Rights Act of 2020 (CPRA) are, at the time of publication, still in flux, with the California Attorney General weighing in with multiple rounds of proposed modifications to the text of the CCPA and the CPRA, expanding the scope of specific provisions, and increasing the resources available to the California Attorney General with which to enforce the obligations of the CCPA. Throughout the last year, the obligations of the parties who are the focus of this note have repeatedly changed. While the author hopes this note is current whenever read, the reader should proceed with caution as the playing field may have shifted yet again.

INTRODUCTION 430
I. THE HYPOTHETICAL 434
II. THE CALIFORNIA CONSUMER PRIVACY ACT 435
- A. *Extraterritoriality and the Reach of the CCPA* 437
- B. *Parties Affected by the CCPA* 438
- C. *Information Covered by the CCPA* 438
- D. *Rights and Obligations Created by the CCPA* 440
- E. *Conduct Covered by the CCPA* 441
  - 1. Web Scraping and the CCPA 444
  - 2. Web Scraping and Notification 450
- F. *Exceptions to CCPA Obligations* 454
- G. *Remedies and the Role of the California Attorney General* 458
- H. *The California Privacy Rights Act of 2020 (CPRA)* 459

CONCLUSION 460

INTRODUCTION

Machine Learning (ML) is a subcategory of Artificial Intelligence (AI). Machine Learning is, at its core, a technology that makes predictions about information absent from input information, or "input data sets," and bases those predictions on data that originally trained the algorithm, or "training data."[1] In

[1] *See* Nick Heath, *What is machine learning? Everything you need to know*, ZDNET (Dec. 16, 2020, 11:21 AM), https://www.zdnet.com/article/what-is-machine-learning-

ML terms, "[p]rediction is the process of filling in missing information."[2] Importantly, ML uses feedback to improve its processes, allowing it to make more accurate predictions in the future.[3] One of the first prediction machines was created to play checkers.[4] This machine, or "model," was trained with simple instructions that primarily consisted of the parameters of the game itself.[5] Within a relatively short window of playing time, the model was able to play the game better than the average player by using feedback to improve its gameplay strategy.[6]

Today, modern prediction machines have many more practical applications and are used for more than just beating their programmers at checkers. Prediction machines are ubiquitous: the technology is "in our phones, cars, shopping experiences, romantic matchmaking, hospitals, banks and all over the media."[7] Increasingly, AI and predictive ML specifically, are being used in Software-as-a-Service (SaaS) applications.[8] One result of this development is that smaller companies that could not otherwise afford to build out complete AI divisions for their businesses can subscribe to AI application tools and apply these advanced models to data produced or collected by their own organizations.[9] Examples of these types of software offerings are available as applications on various cloud service providers, including Amazon's AWS,

everything-you-need-to-know/ [https://perma.cc/6CPF-4BGG] (describing the various ways training data is incorporated into ML algorithms to predict outcomes).

[2] AJAY AGRAWAL ET AL., PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE 24 (2018) (explaining the development of prediction machines as a category of machine learning algorithms).

[3] *See generally* A.L. Samuel, *Some Studies in Machine Learning Using the Game of Checkers*, 3 IBM J. RES. & DEV. 210 (1959) (describing the rote-learning process whereby the model is more easily able to recall the information the more the model encounters it or repeats it).

[4] Donald Michie, *"Memo" Functions and Machine Learning*, 218 NATURE 19, 19 (1968) (referencing the early efforts of A.L. Samuel to create an AI algorithm that could not only play but beat the human playing against the machine).

[5] *See* Samuel, *supra* note 3, at 208–18 (describing the operations of A.L. Samuel's checkers-playing model which is largely credited as the first functioning ML model able to improve itself and function more efficiently).

[6] *Id.* at 221.

[7] AGRAWAL ET AL., *supra* note 2, at 1.

[8] Rachel Wilka et al., *How Machines Learn: Where Do Companies Get Data for Machine Learning and What Licenses Do They Need?*, 13 WASH. J. L. TECH. & ARTS 217, 220 (2018).

[9] *See* Ashish Datta, *How Small Businesses Can Integrate Machine Learning Into Their Model*, FORBES (Dec. 12, 2017, 9:00 AM), https://www.forbes.com/sites/theyec/2017/12/12/how-small-businesses-can-integrate-machine-learning-into-their-model/#7904812ead61 [https://perma.cc/XVE2-78GH] (describing what options exist for small and medium businesses wishing to integrate ML into their business models and evaluating what benefits ML has historically brought to ventures of these sizes).

Microsoft's Azure, Google's Cloud Platform, and IBM's Developer Cloud.[10] These cloud services provide small and medium businesses (as well as larger businesses) access to pre-built ML models which subscribers use to evaluate operational efficiency and other aspects of their businesses. Though to be sure, it's not just the old familiar names of big tech occupying the space. New companies are entering the market each year as the demand for cost-effective business solutions continues to grow.[11] In fact, as of 2020, ML and AI are among the fastest growing sectors of the tech industry.[12]

But let's get back to the data. At this point you might be asking yourself, where does the data come from that is used to train the ML models? For a small- or mid-size business applying ML to its operations through a subscription format—for example, as a subscriber through Google's Cloud platform—the data that originally trained the model comes from the developers of the model.[13] The small business then takes its own data generated through its own business operations—for example, sales records, geographic information, inventory information, customer surveys, etc.— and inputs this data into the algorithm as "input data."[14] The model then processes the information and makes predictions about future customer interactions.

To illustrate this concept, let's look at the example of ML models evaluating the legitimacy of credit card transactions. In this scenario, the model takes certain information about past purchases including: average dollar value of previous purchases; merchants at which a card was previously used; where and when the card was most recently used; and what types of goods the card was used to purchase. The model then uses that information to determine if the

[10] *See* Janakiram MSV, *The Rise Of Artificial Intelligence As A Service In The Public Cloud*, FORBES (Feb. 22, 2018, 10:13 AM), https://www.forbes.com/sites/janakirammsv/2018/02/22/the-rise-of-artificial-intelligence-as-a-service-in-the-public-cloud/#11302580198e [https://perma.cc/7MEY-6HKM] (detailing the predominant cloud platforms and their AI-as-a-Service offerings).

[11] *See* Louis Columbus, *Roundup Of Machine Learning Forecasts And Market Estimates, 2018*, FORBES (Feb. 18, 2018, 7:00 PM), https://www.forbes.com/sites/louiscolumbus/2018/02/18/roundup-of-machine-learning-forecasts-and-market-estimates-2018/#4eb8a15a2225 [https://perma.cc/58DX-K8NK] (analyzing market forecasts for growth of AI sector of technology markets) [hereinafter *2018 Forecasts & Estimates*].

[12] *See* Louis Columbus, *Roundup Of Machine Learning Forecasts And Market Estimates, 2020*, FORBES (Jan. 19, 2020, 2:22 PM), https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/?sh=8d24cd15c020 [https://perma.cc/5DSD-2SU3] (providing analysis on the ML market and global trends in 2020).

[13] *Cloud AutoML*, GOOGLE CLOUD, https://cloud.google.com/automl [https://perma.cc/RL9V-EYGN].

[14] *See* Datta, *supra* note 9 (describing the way small businesses can capitalize on ML using their own data).

card is being used fraudulently in the current purchase.[15] The data that was originally used to train the model establishes the parameters for processing and evaluating new data.[16] The new data, or input data, about a specific credit card purchase is processed by comparison to previous transactions the model knows were either fraudulent or legitimate.[17] This allows the model to "flag" transactions it believes are fraudulent, then notify the customer who can then confirm or deny the legitimacy of the transactions, providing yet another data point for the model to improve upon its prediction accuracy.[18]

Training a model to be accurate requires enormous amounts of data.[19] Most small and medium businesses likely do not have the amounts of data required to train a useful model, and the costs associated with buying the amounts of data required are simply not economically feasible for firms of that size.[20] Nor do small and medium businesses typically have the resources required to develop these advanced but costly tools on their own.[21] That is where AI-as-a-Service comes in. Some of the largest platforms already have well developed AI divisions within their businesses, as previously mentioned, and most of them even offer those services to other businesses through their platforms.[22] While this note will not deeply explore this specific scenario, the issue of who owns the model and input data—either the AI firm or the business wishing to utilize AI in its operations—depends largely on the parties' user agreements and terms of service .[23] But the scenario just mentioned

[15] *Cf. id.* (comparing scenarios in which a credit card company would examine for fraudulent activity using manual methods and ML); Jungwoo Ryoo, *How Do Companies Know When Someone Else is Using Your Credit Card?*, SLATE (Nov. 22, 2017, 11:48 AM), https://slate.com/technology/2017/11/how-companies-can-tell-when-someone-else-is-using-your-credit-card.html [https://perma.cc/6LVL-NKPJ] (examining the modern practice of using ML algorithms to detect credit card fraud).

[16] Ryoo, *supra* note 15.

[17] *Id.*

[18] *See* Datta, *supra* note 9.

[19] *See* Theophano Mitsa, *How Do You Know You Have Enough Training Data?*, TOWARDS DATA SCI. (Apr. 22, 2019), https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee [https://perma.cc/EZ99-EMTP] (explaining data input formulas that require vast amounts of data to be effective).

[20] Daniel Newman, *Why AI As A Service Will Take Off In 2020*, FORBES (Jan. 7, 2020, 1:06 PM), https://www.forbes.com/sites/danielnewman/2020/01/07/why-ai-as-a-service-will-take-off-in-2020/#6672c3c33669 [https://perma.cc/2D8S-2RR9].

[21] *Id.*

[22] *Id.*

[23] Brian Higgins, *When It's Your Data But Another's Stack, Who Owns The Trained AI Model?*, NEWS AND ANALYSIS OF AI TECH. & L. (Jan. 31, 2018), http://aitechnologylaw.com/2018/01/who-owns-cloud-trained-ai-model/ [https://perma.cc/LU9T-93PS].

is one in which the medium-sized business simply processes its own data through the AI firm's algorithm. Whereas here, we are concerned more so with a company developing an ML or AI model and selling all of its assets, including the model and training datasets, to an acquiring company.

Other mature companies—including Netflix, Salesforce.com, John Deere, Splunk, and others—are hard at work developing these divisions within their own organizations,.[24] In an effort to keep pace, firms often try to leap-frog their competitors by acquiring companies which develop these sophisticated tools. Where those upstart AI companies get their data is a primary focus of this note.

I. THE HYPOTHETICAL

The following is a hypothetical scenario which serves as the foundation for the analysis in this note.

In the very real-world scenario where a large, mature software technology firm seeks to establish or improve an ML or AI component of its business. Rather than build-out this portion of its software stack from scratch, the mature technology firm elects to acquire an existing company (the AI startup) that has developed the technology it wants to integrate into its own software offering.

Unbeknownst to the mature tech firm, the acquisition target built its ML model on training data comprised of information scraped from public-facing websites. When the AI startup scraped public-facing websites, it collected a large amount of data containing PII of individuals, some of which (for our purposes) were California residents. Scraping involves the use of "bots," or robot applications deployed for automated tasks, which scan and copy the information on webpages then store and index the information.[25] The AI startup then compiled this data into a format its ML model could accept and process, and used the data to "train" the model. In most instances where a similar scenario actually occurs, the industry best practice is to anonymize or pseudonymize the training data to avoid the use or occurrence of PII in the training data

[24] Sam Daley, *10 Publicly Traded Companies Innovating With AI*, BUILT IN (Dec. 4, 2018), https://builtin.com/artificial-intelligence/publicly-traded-ai-companies [https://perma.cc/29DS-KHUL].

[25] *See* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d. 537, 544 (S.D.N.Y. 2013) (describing how "web crawler" applications function. These bots are capable of scanning millions of web pages daily and are limited primarily by the protocols of the servers of the websites the bots are programmed to scan. If the websites observe that the bots do not honor the protocols established by the websites, the bots can be denied from accessing the sites. To be clear, there is much practical use for "crawling" by bots—this is primarily the way search engines like Google rank pages and return relevant search results for their users).

altogether.[26] There are both practical as well as legal reasons that make this practice preferable to alternatives.[27] However, for our purposes, we will assume that the AI startup cut corners and included the PII in the training data.

As part of the acquisition of the AI startup, the training data sets were included in the transfer. So, the acquiring company now possesses the PII in large data sets that were used to train the AI model, as well as to whatever degree the PII is contained within the model itself. In reality, this hypothetical is fairly common.[28] Scenarios just like this, where a maturing tech firm wishes to keep pace with competitors by choosing to buy an upstart company, occurs with increasing regularity.[29] As mentioned above, this type of leap-frogging allows a company to remain competitive by efficiently utilizing its available resources and simply purchasing an upstart company rather than devoting the time and resources required to build a comparable division for itself.

II. THE CALIFORNIA CONSUMER PRIVACY ACT

The California Consumer Privacy Act of 2018 (CCPA) took effect on January 1, 2020.[30] The CCPA was largely a response by the California state legislature to a proposed ballot initiative.[31] The resulting law was drafted in less than a week and, by some accounts, contains multiple drafting errors, typos, and less than

[26] ERIKA MCCALLISTER ET AL., NAT'L INST. OF STANDARDS & TECH., U.S. DEP'T OF COMMERCE, GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION (PII) 4–5 (Apr. 2010) (recommending that businesses seek to purposefully limit the amount and extent of PII they collect from consumers as well as de-identifying or anonymizing the PII contained in data sets and limiting the access to the data sets).

[27] *See Id.* at 2–3 (suggesting that the OECD's Fair Information Practices have been adopted by the U.S. Department of Commerce and have also been used to inform both U.S. federal laws as well as international legislation, namely the European Union's General Data Protection Regulation).

[28] Brad Janssen, Matthew Knouff, & Rani Habash, *Data privacy and security issues in M&A transactions: Part one*, IAPP (Apr. 26, 2016), https://iapp.org/news/a/data-privacy-and-security-issues-in-ma-transactions-part-one/ [https://perma.cc/2ATR-BPLK]. *See also 2018 Forecasts & Estimates*, *supra* note 11 ("Machine learning's potential impact across many of the world's most data-prolific industries continues to fuel venture capital investment, private equity (PE) funding, mergers, and acquisitions all focused on winning the race of Intellectual Property (IP) and patents in this field.").

[29] *See 2018 Forecasts & Estimates*, *supra* note 11.

[30] CAL. CIV. CODE § 1798.198 (West, Westlaw through Ch. 9 of 2021 Reg.Sess.).

[31] *See* Eric Goldman, *An Introduction to the California Consumer Privacy Act (CCPA)*, IAPP 1 (July 9, 2018), https://iapp.org/media/pdf/resource_center/Intro_to_CCPA.pdf [https://perma.cc/VQ3F-V475].

ideal policies.[32] "The CCPA is arguably the most comprehensive—and complex—data privacy regulation in the United States. It may also be one of the most hastily put together pieces of privacy legislation in recent history."[33] Some feel the CCPA is an effort to keep pace with European privacy law, namely the European Union's General Data Protection Regulation (GDPR).[34] Like the CCPA, the GDPR also creates affirmative duties and corresponding individual rights. For example, both the CCPA and GDPR require notices to individual data subjects (though the CCPA uses a different term in lieu of "data subject"); create individual rights to access the data that companies collect about the individual; and provide a right of erasure (more commonly known as the "right to be forgotten").[35] Additionally, both the CCPA and GDPR require businesses to notify customers about what information they collect and how they use and process the information they collect.[36] They both apply the "business requirement" to the collection of data on and offline.[37] And both apply to a wide range of businesses across sectors (though the CCPA has certain carveouts specifically for industries already covered by specific federal regulations).[38]

The stated goal of the CCPA is "to further Californians' right to privacy by giving consumers an effective way to control their personal information."[39] The CCPA creates a number of statutory rights for California residents, discussed in greater detail below in Section D.[40] The most relevant to our hypothetical are the rights and obligations that provide California residents with a say in

[32] *Id*. at 2; *Cf*. David Zetoony, *California Consumer Privacy Act (CCPA) Practical Guide*, BRYAN CAVE LEIGHTON PAISNER 1–2 (Feb. 2020), https://ccpa-info.com/wp-content/uploads/2019/09/bclp-practical-guide-to-the-ccpa.pdf [https://perma.cc/W88A-CWWF] [hereinafter Zetoony, *CCPA Practical Guide*].

[33] *California Amendment to Privacy Law Official*, BRYAN CAVE LEIGHTON PAISNER (Sept. 26, 2018), https://www.bclplaw.com/en-US/thought-leadership/california-amendment-to-privacy-law-official-the-definitive.html [https://perma.cc/NT3B-THY4].

[34] Zetoony, *CCPA Practical* Guide, *supra* note 32, at 2.

[35] *What You Need to Know About the New General Data Protection Regulation (GDPR)*, BRYAN CAVE LEIGHTON PAISNER (Feb. 17, 2016), https://www.bclplaw.com/en-US/thought-leadership/what-you-need-to-know-about-the-new-general-data-protection.html [https://perma.cc/RBN6-22VG].

[36] *Id.*; CAL. CIV. CODE § 1798.145 (West, Westlaw through Ch. 9 of 2021 Reg. Sess.) (describing the exemptions granted to businesses and individuals attempting to comply with California Consumer Protection laws).

[37] *Cf. What You Need to Know About the New GDPR*, *supra* note 35 (noting that the GDPR applies to "companies doing business in the EU"); CAL. CIV. CODE § 1798.145 (Westlaw).

[38] *Cf.* Zetoony, *CCPA Practical* Guide, *supra* note 34; CAL. CIV. CODE § 1798.100 (Westlaw) (observing obligations which apply to "businesses that collect personal information").

[39] Assemb. B. 375 § 2(i), 2017–2018 Leg., Reg. Sess. (Cal. 2017).

[40] Max N. Helveston, *Reining in Commercial Exploitation of Consumer Data*, 123 PENN. ST. L. REV. 667, 690 (2019).

preventing companies that collect the residents' information from selling it to other businesses, and require the companies doing the collecting to notify consumers "at the time of collection."[41]

A. Extraterritoriality and the Reach of the CCPA

By its own terms, the CCPA seems to reach outside its borders, asserting jurisdiction over any company that meets its definition of doing "business in the state," "potentially appl[ying] to any business throughout the globe that has/gets personal information about California residents the moment the business takes the first dollar from a California resident."[42] This is particularly noteworthy given the sectoral patchwork approach to U.S. privacy law generally.[43] Moreover, "the law's purported application to businesses not physically located in California raises potentially significant dormant Commerce Clause and other Constitutional problems."[44] While these considerations are outside the scope of this note, they highlight yet more potential problems with a law which purports to extend its own jurisdiction throughout the United States and beyond.

If a business collects the personal information of 50,000 or more California consumers in a year, by its language, the CCPA reaches the business even if it does not actually "do business in the state of California."[45] For businesses physically located near the California state line in neighboring border states, which many California residents may visit annually, this may mean they are swept up by the CCPA even if these businesses do not have an online presence. Though this observation is not completely relevant to this note, it highlights the public policy concerns of allowing California's laws to reach beyond its borders to capture businesses with no direct ties to California.

[41] *Id.*; CAL. CIV. CODE § 1798.140 (Westlaw).

[42] CAL. CIV. CODE § 1798.140(c)(1) (Westlaw); Goldman, *supra* note 31, at 2.

[43] *See* Jane K. Winn, *Can a Duty of Information Security Become Special Protection for Sensitive Data under US Law?* (Sept. 9, 2008), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1265775 [https://perma.cc/3BGC-6H47] (describing the "patchwork" of U.S. privacy laws).

[44] Goldman, *supra* note 31, at 2.

[45] CAL. CIV. CODE § 1798.140(c)(1)(B) (Westlaw) (doing business in the state of California is one of the potential qualifiers in (c)(1), but not the only one so a "business" under the CCPA could do business in another state yet still collect the personal information from at least 50,000 California consumers in a year); *see also* Zetoony, *CCPA Practical Guide*, *supra* note 34, at 3.

B. *Parties Affected by the CCPA*

The CCPA affects companies "doing business in" California, but only if they buy or sell consumer information of 50,000 (or more) California "consumers" or "devices" (100,000 or more effective January 1, 2023)[46]; or have a gross annual revenue of $25 million or more[47]; or derive at least fifty percent of their revenue from sharing personal information from consumers.[48] However, the CCPA contains accommodations and carve-outs for small-businesses, non-profits, and other institutions and businesses which are covered by federal laws, such as the Gramm-Leach-Bliley Act (regulating financial institutions), the Fair Credit Reporting Act (regulating consumer reporting agencies), and the Health Insurance Portability and Accountability Act (regulating health care providers).[49]

Unlike the thresholds and carveouts for companies "doing business" in California, the consumer definition has no such exceptions. The CCPA's definition of "consumer" is given substance through its reference and application to "natural [resident] person(s)."[50] "Resident" is further defined by the CCPA to include "(1) every individual who is in the State for other than a temporary or transitory purpose, and (2) every individual who is domiciled in the State who is outside the State for a temporary or transitory purpose."[51] The statute provides for the protection of natural persons who are residents of California, "however identified, including by a unique identifier."[52] This definition is consistent with the GDPR's definition of "data subject," which is "an identified or identifiable natural person."[53]

C. *Information Covered by the CCPA*

The CCPA provides for activities of qualifying businesses engaged in collection, use, and sale (among other things) of "personal information" (or PII) of California resident consumers.[54]

[46] CAL. CIV. CODE §1798.140(c)(1)(B) (amended 2020); *see also* § 1798.140(c)(1)(B) (Westlaw) (effective Jan. 1, 2023).

[47] CAL. CIV. CODE § 1798.140(c)(1)(A) (Westlaw).

[48] *Id.* § 1798.140(c)(1)(C) (Westlaw) (This definition is amended by the CPRA which takes effect Jan. 1, 2023).

[49] CAL. CIV. CODE § 1798.145(c)(1), (d)(2), (d)(3)(e) (Westlaw); Zetoony, *CCPA Practical Guide*, *supra* note 34, at 3.

[50] CAL. CIV. CODE § 1798.140(g) (Westlaw) (The CCPA's definition of consumer makes direct reference to California Code of Regulations which already contains a definition of "resident." *See* 18 CAL. CODE REGS. § 17014 (2020).

[51] 18 C.C.R. § 17014.

[52] CAL. CIV. CODE § 1798.140(g) (Westlaw).

[53] Council Directive 2016/679, art. 3, 2016 O.J. (L 119) 33.

[54] Zetoony, *CCPA Practical Guide*, *supra* note 34, at 3–4.

Like some of its other provisions, the CCPA gives qualifying information a robust definition as that which "identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household."[55] The statute goes on to list a number of qualifying pieces of PII which includes without limitation:

> [R]eal name, alias, postal address, unique personal identifier, online identifier, Internet Protocol address, email address, account name, social security number, driver's license number, passport number . . . [i]nternet or other electronic network activity information, including, but not limited to, browsing history, search history, and information regarding a consumer's interaction with an Internet Web site, application, or advertisement . . . [a]udio, electronic, visual, thermal, [and even] olfactory" information.[56]

Importantly, expressly excluded from its definition of consumer information are "deidentified" and "aggregate" information.[57] The CCPA defines "aggregate personal information" as "information that relates to a group or category of consumers, from which individual consumer identities have been removed, that is not linked or reasonably linkable to any consumer or household, including via a device."[58] "Deidentified information" is statutorily defined as "information that cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer," (with three additional requirements proscribing how the company may use the deidentified consumer information).[59]

Notably, nowhere does the CCPA "attempt to harmonize the overly broad definition of 'personal information' with deidentification or aggregation."[60] In fact, some commenters have suggested that the inclusion of "reasonable" in the definition of "deidentified information" creates an unintended gap because a business otherwise covered by the statute could claim that even if the consumer information is re-identifying *in fact*, if the process of

[55] CAL. CIV. CODE § 1798.140(o)(1) (Westlaw).
[56] *Id.* § 1798.140(o)(1)(A),(F),(H) (Westlaw).
[57] *Id.* § 1798.145(a)(5) (Westlaw).
[58] *Id.* § 1798.140(a) (Westlaw).
[59] *Id.* § 1798.140(h)(1)–(3) (Westlaw).
[60] *See* Goldman, *supra* note 31, at 3 (describing the broad and overly inclusive definitions of "consumer information" and the statutes' lack of a cohesive framework for determining what information is included or excluded from the definition).

re-identifying the resident consumer requires more than a "reasonable" effort, the consumer's personal information is therefore not covered by the statute.[61]

Most importantly for our purposes, this definition includes the personal information a California resident consumer willingly places on publicly available web pages. The CCPA provides that "publicly available information" is excluded from the statutory definition of "personal information."[62] However, this exclusion is severely limited in that "'publicly available' [only] means *information that is lawfully made available from federal, state or local government records*."[63] In other words, only information contained in federal, state, or local government records, which is also lawfully made publicly available (for example, in property tax assessor files, registered voter files, court filings, motor vehicle records, professional and business licenses, etc.) is excluded from the statutory definition of "personal information" and therefore not subject to the provisions of the CCPA.

D. Rights and Obligations Created by the CCPA

The CCPA creates a number of rights for California "consumers." The primary rights created by the CCPA include 1) the right of a consumer to request information from a business regarding what information the business collects about the consumer,[64] 2) the right of a consumer to request that information collected about the consumer be deleted by the business,[65] 3) the right to "opt-out" of the sale of the consumer's information by the collecting business to a third-party,[66] and 4) the consumer's right to not be discriminated against by the business if the consumer chooses to exercise its rights under the CCPA.[67]

These rights are not unqualified. For example, a consumer's right to request that a business or service provider delete their personal information is limited by whether the business requires the personal information for a number of potential reasons.[68] Those reasons can include "detecting security incidents," engaging in "public or peer-reviewed" research, and notably, a determination by the business itself whether it requires the "use [of] the consumer's personal information, internally, in a lawful manner that is

[61] *Id.* at 4.
[62] CAL. CIV. CODE § 1798.140(o)(1)(K)(2) (Westlaw).
[63] *Id.* (emphasis added).
[64] *Id.* § 1798.100(a) (Westlaw).
[65] *Id.* § 1798.105(a) (Westlaw).
[66] *Id.* § 1798.120(a) (Westlaw).
[67] *Id.* § 1798.125(a)(1) (Westlaw).
[68] *Id.* § 1798.105(d)(1)–(9) (Westlaw).

compatible with the context in which the consumer provided the information."[69]

This last exception clearly places a great deal of discretion in the hands of the business by allowing the business to make its own determination to whether it deems the information necessary for internal use. This exception is notable because it exempts businesses from compliance with one of the strongest rights a consumer can exercise—the right to require the business to delete personal information about the consumer. By its own terms, the language of this exception establishes a relatively low bar for a business wishing to retain the information and likely only prevents the business from selling the information to a third-party as the information may only be retained for "internal" use.[70] This exception may also play an important role in determining what the mature tech firm in our hypothetical scenario may do with the personal information it receives as an included asset in its acquisition of the AI startup. Because the AI startup has used personal information as training data for its AI algorithm, there is reason to believe its continued use of the training data and/or algorithm may fit this exception. This potential outcome is explored later in this note.

E. Conduct Covered by the CCPA

The CCPA defines information "collecting" very broadly, as "buying, renting, gathering, obtaining, receiving, or accessing any personal information pertaining to a consumer *by any means*. This includes receiving information from the consumer, either actively or passively, or by observing the consumer's behavior."[71] This definition reaches a vast number of businesses and business activities.[72] However, the text of the CCPA appears most concerned with the sort of direct interactions between consumers and websites which require some form of PII from the consumer in order to register and verify the user, or allow the consumer to purchase or

[69] *See id*.

[70] *Id.* § 1798.105(d)(9) (Westlaw).

[71] *Id.* § 1798.140(f) (Westlaw) (emphasis added).

[72] It is not difficult to imagine the vast number of scenarios which would bring a business, online or otherwise, within the reach of the CCPA, given its definition of "collecting" personal information. To name a few of the most common from personal experience—registering a profile for a site or platform, downloading a whitepaper, entering a contest or sweepstakes, purchasing goods, downloading an app, etc.

Deleted: *I*

Deleted: *d*

download something from the site.[73] Large websites such as Facebook, LinkedIn, and Amazon all qualify under this definition because they each require registered users to log into the site with an account before a user can upload information about themselves or purchase goods.[74] The websites use such information to verify the user is legitimate, but may also use the information for other purposes.

All of the statutory consumer-rights that the CCPA creates are "intended to further the [state] constitutional right of privacy and to supplement existing laws relating to consumers' personal information."[75] However, the primary right created by the CCPA is the California consumer's right to request information from these types of businesses regarding what information the businesses collect about the consumer, and to whom the businesses provides that information.[76] Furthermore, this right does not have any listed exemptions and all qualifying businesses must comply with "verifiable consumer requests."[77] Rather than make each consumer dig through a websites' terms of use looking for contact information, the CCPA requires businesses covered by the statute to "[m]ake available to consumers two or more designated methods for submitting requests for information required to be disclosed [by the business] . . . including, at a minimum, a toll-free telephone number."[78] But a "business that operates exclusively online and has a direct relationship with a consumer from whom it collects personal information" is only required to "provide an email address for submitting requests for information."[79] The CCPA requires that websites create a "clear and conspicuous link" on the website's homepage that consumers can access to make the requests detailed above.[80]

[73] *See* Assemb. B. 375 § 2(a)–(i), 2017–2018 Leg., Reg. Sess. (Cal. 2017). As discussed later, the text of the CCPA indicates that the law primarily targets websites directly interacting with consumers and collecting consumer information which consumers provide, and at the third parties with which the collecting websites transact.

[74] All three of these presumably have income great enough to be included within the scope of the CCPA. In addition, they each collect "personal information" about their users—some perhaps more than others.

[75] CAL. CIV. CODE § 1798.175 (Westlaw).

[76] *Id.* § 1798.100(a) (Westlaw). This is the first statutory right listed by the CCPA and seeks most pointedly to accomplish the goal of providing Californians with more information about how their state constitutional right to privacy may be affected by their sharing of personal information with whom businesses the consumer engages.

Deleted: with

[77] Stuart D. Levi, *California Consumer Privacy Act: A Compliance Guide,* SKADDEN, ARPS, SLATE, MEAGHER & FLOM LLP 13–22 (Mar. 20, 2019), https://www.skadden.com/insights/publications/2019/03/california-consumer-privacy-act [https://perma.cc/W8HK-Y7B9].

[78] CAL. CIV. CODE § 1798.130(a)(1)(A) (Westlaw).

[79] *Id.*

[80] *Id.* § 1798.135(a)(1) (Westlaw).

Without this type of direct interaction, which firms engaged in data scraping have no practical way of accomplishing, there does not seem to be a technologically satisfying solution for website scrapers to notify consumers. When performed within the bounds of applicable law, notwithstanding the CCPA, scraping is the copying and indexing of information placed on public-facing webpages. The activity can be done manually or by automated process as described below. But there is no direct interaction between the party doing the scraping and the consumers who make the information publicly available.

However, it is only by virtue of the fact that websites have a way to inform their users "at the point of collection" about their data collection practices, and the users of their rights under the CCPA, that the CCPA can be effective at all. Without this type of direct interaction with consumers, which firms engaged in data scraping conduct do not have a meaningful way of replicating, an entire industry is left in limbo.

The language of the statute, specifically the proposed bill's recitals, further implies that the websites the CCPA is exclusively directed at are those with which consumers interact directly.[81] Because websites like Facebook and LinkedIn require users to register with the site to create profiles and change information about themselves, those businesses have a readily available medium by which they can notify users of their rights under the CCPA, namely their own websites. In fact, they are required to use this medium under the CCPA to do just that.[82] Furthermore, because they manage and maintain that medium they are responsible for what happens to the users' PII in their possession.[83] That is, if those businesses collect or sell user's PII, or otherwise contract for a third-party's use of PII, they are required by the CCPA to respond to consumer inquiries regarding what specific PII is being used by third-parties and by whom the PII is being used,

[81] *See* Assemb. B. 375 § 2(a)–(i), 2017–2018 Leg., Reg. Sess. (Ca. 2017) (Notably recital (d) which reads as follows: "As the role of technology and data in the everyday [sic] lives of consumers increases, there is an increase in the amount of personal information *shared by consumers with businesses*. California law has not kept pace with these developments and the personal privacy implications surrounding the collection, use, and protection of personal information." (Emphasis added)).

[82] *See* CAL. CIV. CODE § 1798.135(a)(1) (Westlaw).

[83] The definition of "business" is defined to include the conduct and uses the businesses engage in regarding "personal information." It provides in part, a business is "a legal entity . . . that collects consumers' personal information or on the behalf of which that information is collected and that alone, or jointly with others, determines the purposes and means of the processing of consumers' personal information." *Id.* § 1798.140(c)(1) (Westlaw).

and importantly, offer consumers an easy way to request that their information not be sold to those or other third-parties.[84]

For websites the size of Facebook, LinkedIn, and Amazon, the primary difficulty encountered by these requirements is not likely responding to consumers or making sure they know who is using the information, but rather managing the massive amounts of data they collect. But for smaller companies without the resources of these tech giants, simply ensuring that service provider contracts meet the requisite level of control is undoubtedly a daunting task.

1. Web Scraping and the CCPA

For our hypothetical, the threshold question is whether the conduct of scraping a website constitutes "collecting" information under the CCPA's expansive definition. An important subsequent question is whether or not it *should* qualify. Because the definition of "collect" under the CCPA includes "obtaining" or "gathering" consumer PII "*by any means*," the conduct of scraping a website seems to fall squarely within that description.[85] As discussed above, third-party scrapers use bots to index or otherwise gather information found on public-facing webpages.[86] Internet search engines operate this way and a whole industry has popped up around strategically ensuring a given website is among the top results from search queries conducted using search engines like Google and Bing.[87] SEO, or Search Engine Optimization, is the practice of positioning a website to be among top search results for certain key terms in search engines like Google and Bing.[88] In fact, Forbes estimated (based on Borrell Associates research conducted in 2016) that in the U.S. alone, $80 Billion would be spent on SEO in 2020.[89] These popular search engines work by crawling "hundreds of billions" of webpages, gathering information from

[84] *Id.* §§ 1798.135, 1798.140 (Westlaw).

[85] *Id.* § 1798.140(e) (Westlaw) (emphasis added).

[86] Scraping is not a new practice and the market for data from online sources was reported to have been in the hundreds of millions of dollars as far back as 2009. *See* Julian Angwin & Steve Stecklow, *'Scrapers' Dig Deep for Data on Web*, WALL ST. J. (Oct. 12, 2010, 12:01 AM), https://www.wsj.com/articles/SB10001424052748703358504575544381288117888 [https://perma.cc/2YL6-LGA6].

[87] *How Search organizes information*, GOOGLE, https://www.google.com/search/howsearchworks/crawling-indexing/ [https://perma.cc/N35M-5QWT].

[88] *See Search Engine Optimization (SEO) Starter Guide*, GOOGLE, https://support.google.com/webmasters/answer/7451184?hl=en [https://perma.cc/U3MH-K8BM].

[89] TJ McCue, *SEO Industry Approaching $80 Billion But All You Want Is More Web Traffic*, FORBES (July 30, 2018, 3:41 AM), https://www.forbes.com/sites/tjmccue/2018/07/30/seo-industry-approaching-80-billion-but-all-you-want-is-more-web-traffic/#4dc38daa7337 [https://perma.cc/J367-CEHT].

them and then organizing that information in massive Search indexes.[90] When a search is conducted on one of the engines, the terms in the search are processed by Google's trademarked algorithm and the results a user sees are web pages Google has indexed which the algorithm thinks you'll most want to visit.[91]

Overall, the body of case law treating the conduct of web scraping is sparse, however several recent cases have addressed scraping by businesses who appropriate the scraped information and use the information in one form or another. In *Spokeo v. Robins* the website at issue, Spokeo.com, scraped and compiled information from multiple public-facing web pages to compose reports about the individual subjects of search queries performed on its site.[92] Users of Spokeo.com are able to search for people based on name, email address, home address and other criteria.[93] Spokeo.com's users ranged from inquisitive romantic partners to prospective employers performing background research on job applicants.[94] Primarily at issue in the U.S. Supreme Court's decision was whether or not the plaintiff's, Robins's, complaint was sufficient to establish standing.[95] The conduct alleged by the complainant was that Spokeo.com violated the Fair Consumer Reporting Act (FCRA) when it presented false information about Robins in the form of a "consumer report."[96]

Not at issue in *Spokeo* was the conduct of scraping third-party websites, but rather reporting on the information that was obtained through that scraping.[97] So, while this case is perhaps the only case involving website scraping to come before the U.S. Supreme Court, the Court's decision does not provide a great deal of insight as to how the court would treat the conduct of website scraping specifically.

In *Associated Press v. Meltwater Holdings,* decided before *Spokeo*, the court considered whether Meltwater's conduct of crawling various websites for AP's stories and scraping "snippets"

[90] *How Search organizes information*, *supra* note 87.

[91] *Search algorithms: How Search algorithms work*, GOOGLE, https://www.google.com/search/howsearchworks/algorithms/ [https://perma.cc/2LKL-FNKP].

[92] Spokeo, Inc. v. Robins, 136 S. Ct. 1540, 1544, 1546 (2016). ("Spokeo conducts a computerized search in a wide variety of databases and provides information about the subject of the search." "Spokeo markets its services to a variety of users, including not only 'employers who want to evaluate prospective employees,' but also 'those who want to investigate prospective romantic partners or seek other personal information.'").

[93] *Id.* at 1546.

[94] *Id.*

[95] *Id.* at 1547–48.

[96] *Id.* at 1545–46.

[97] *Id.* at 1544.

of the stories for use in notifying and informing Meltwater's own customers of certain stories, violated the Copyright Act.[98] The U.S. District Court for the Southern District of New York, deciding on cross motions for summary judgment, considered the defenses proffered by Meltwater whose primary defense was based on the "fair use" doctrine.[99] Meltwater's services, for which its subscribers paid thousands of dollars annually, included crawling websites and reproducing verbatim portions of news stories based on user search terms.[100] In the content reported to Meltwater's users, directly at issue in the case, were thirty-three copyrighted stories originally reported by the Associated Press (AP).[101]

Arguing that it functioned primarily as an internet search engine, Meltwater claimed that it "transformed" the AP copyrighted content thereby making its conduct exempt through the so-called "fair use" doctrine.[102] While this case gets closer to the conduct we are most concerned about, namely the crawling and scraping of websites, its focus is on the rights enjoyed by parties with intellectual property rights, namely copyrights, in the scraped content.[103] In our hypothetical, individual consumers do not have copyrights in their names or other personal information. That being said, the case does have some relevance in that the court considered the "transformation" defense presented by Meltwater.[104] Because the "fair use" doctrine is specifically an affirmative defense in copyright infringement litigation, [105] it is not clear how under the CCPA a similar defense might be made.

Still, the CCPA does include certain express exemptions of specific conduct, which are covered later in this article. In *Meltwater*, the court considered the extent to which the scraped information was transformed reasoning that the 'fair use' doctrine is designed to protect the use of copyrighted materials where "new aesthetics, new insights and understandings" are produced.[106] But likewise, it is not designed to protect use where the copyrighted

[98] Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 541–42 (S.D.N.Y. 2013).

[99] *Id.* at 541.

[100] *Id.* at 543.

[101] *Id.* at 542.

[102] *Id.* at 550.

[103] *See id.* at 557–61.

[104] *Id.* at 550.

[105] *Id.* ("The fair use doctrine, although of common law origin, has been codified at 17 U.S.C. § 107. This section provides that '[n]otwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work … for purposes such as criticism, comment, news reporting, teaching … scholarship, or research, is not an infringement of copyright.'").

[106] *Id.* at 551.

material is merely “repackage[d].”[107] Unconvinced by Meltwater’s arguments, the court reasoned that Meltwater’s use of the copyrighted content was not sufficiently transformative and exploitive of AP’s work in a manner which unfairly injured AP.[108]

Most recently, in *hiQ Labs v. LinkedIn*, a case that reached the Ninth Circuit Court of Appeals on noteworthy procedural grounds, the court was confronted with the question of whether LinkedIn could prevent hiQ Labs from scraping information from public facing user-profiles hosted by LinkedIn’s website on the grounds that the conduct allegedly violated the Computer Fraud and Abuse Act (CFAA) of 1986, the Digital Millennium Copyright Act (DMCA), and the common law of trespass.[109] The Ninth Circuit Court of Appeals limited its ruling to the issue of preliminary injunctive relief ordered by the District Court.[110] There, hiQ Labs sought to prevent LinkedIn from prohibiting hiQ’s access to the profiles by way of an injunctive order.[111] The court correctly limited its review of the lower court’s decision by evaluating whether the decision of the lower court was “illogical, implausible, or without support in the record.”[112]

In its review, the court seemingly relied heavily on both third-party doctrine and the court’s understanding of the phrase, “without authorization” as found in the CFAA, to support its finding that both LinkedIn’s users and LinkedIn itself had assumed the risk that a third-party might view the public-facing user-profile information (containing personal information such as name, email address, education and employment history), and that LinkedIn did not adequately limit the public’s access to the webpages in question to meet the CFAA’s meaning of “without authorization.”[113]

Because both LinkedIn and the individual users made the information available to the public by making the user profiles viewable to anyone with a web browser, and because LinkedIn did not claim any ownership of the user data by virtue of the terms of service in its user agreements, the users effectively assumed the

[107] *See id.*

[108] *Id.* at 552–53.

[109] hiQ Labs v. LinkedIn Corp., 938 F.3d 985, 992 (9th Cir. 2019). Initially, LinkedIn sent hiQ a cease-and-desist letter demanding hiQ discontinue its practice of scraping the public facing pages hosted by LinkedIn. hiQ then filed suit against LinkedIn and sought a preliminary injunctive order preventing LinkedIn from stopping hiQ’s activity.

[110] *Id.* at 993.

[111] *Id.* at 992–93.

[112] *Id.* at 993 (citing Doe v. Kelly, 878 F.3d 710, 713 (9th Cir. 2017)).

[113] *Id.* at 1003 (analyzing the meaning of “without authorization” in the context of the CFAA, 18 U.S.C. § 1030(a)(2) (1986)).

risk that a third-party might view the information.[114] The court reasoned, "[i]t is likely that when a computer network generally permits public access to its data, a user's accessing that publicly available data will not constitute access without authorization under the CFAA."[115] Furthermore, in its evaluation of the lower court's finding as to which party's individual interests were most aligned with that of the public interest, the court again found that hiQ Labs' interest weighed heaviest.[116] It reasoned that,

> [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.[117]

This language from the court reflects the precarious position U.S. privacy law finds itself in at the current moment. The barrier between public and private is small but significant for both the individuals whose information is swept up by parties scraping web pages viewable by the public, and the companies which host and use the information their users provide. Making all that information public—which is not private—result which the decision in *hiQ Labs* lends itself to, is arguably one of the biggest issues the CCPA seeks to address.

The decision in *hiQ Labs* came in the wake of two other Ninth Circuit decisions, *United States v. Nosal*, 844 F.3d 1024 (9th Cir. 2016) (*Nosal II*), and *Facebook v. Power Ventures*, 844 F.3d 1058 (9th Cir. 2016). In *Nosal II*, the court reasoned that the CFAA's use of the phrase "without authorization" extended to the access of a password protected area of a website by an unauthorized person using valid login credentials.[118]

More notable is the court's decision in *Power Ventures* though, where the court found that Power Ventures's receipt of a cease and desist letter from Facebook made any subsequent access of Facebook computers by Power Ventures in excess of the authorization and access rights otherwise permitted.[119] In *Power Ventures*, power.com accessed Facebook user profiles after having been given Facebook user's valid login credentials, then collected and aggregated this data with data from the same user's other

[114] *Id.* at 1003–04.
[115] *Id.* at 1003.
[116] *Id.* at 1005.
[117] *Id.* at 1005.
[118] *See* United States v. Nosal, 844 F.3d 1024, 1035–37 (9th Cir. 2016).
[119] *See* Facebook v. Power Ventures, Inc., 844 F.3d 1058, 1062 (9th Cir. 2016).

social media accounts, and put all the information in one place for the user to access at power.com.[120]

Despite having received a cease-and-desist notice and Facebook blocking the IP address from which Power Ventures was accessing Facebook computers, Power Ventures continued to access those computers, even changing its IP address to one that had not been blocked.[121] Facebook also sought for Power Ventures to utilize Facebook's Application Programing Interfaces (APIs) and abide by Facebook's Terms of Use for third-party developers, which Power Ventures resisted.[122]

The court in *hiQ Labs*, likely realizing the similarities to the situation in *Power Ventures*, distinguished *hiQ Labs* from *Power Ventures* noting an important difference between the facts presented: the Facebook computers which Power Ventures accessed were only accessible to users with valid login credentials.[123] Stated differently, the information accessed in *Power Ventures*—the user-profile data—was password-protected and not open to the public, whereas the information in *hiQ Labs* was public-facing, meaning anyone with a web browser could view the profiles.[124] This difference is foundational to the court's decision in *hiQ Labs*. Had the user profiles been accessible only by those with valid login credentials, it seems safe to say the court's decision would have been in much the same vein as *Power Ventures* and *Nosall II*. Many commenters hailed the decision in *hiQ Labs* as a victory for web scraping and web scrapers.[125] But this victory may be short-lived as the CCPA's broad mandate may begin to chip away at this sort of activity.

Because of the unique procedural posture of *hiQ Labs*, the court did not issue a decision on the merits of the underlying arguments, and expressly limited its opinion to the finding for preliminary injunctive relief as ordered by the lower court.[126] The substance of the arguments LinkedIn made in its reply to the

[120] *Id*.

[121] *Id*. at 1063.

[122] *Id*.

[123] hiQ Labs v. LinkedIn Corp., 938 F.3d 985, 1002 (9th Cir. 2019).

[124] *Id*.

[125] Camille Fischer & Andrew Crocker, *Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data*, ELEC. FRONTIER FOUND. (Sept. 10, 2019), https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data [https://perma.cc/4AVN-5249].

[126] *hiQ Labs*, 938 F.3d at 995 (Explaining the purview of the court's review, "[a]s usual, we consider only the claims and defenses that the parties press on appeal. We recognize that the companies have invoked additional claims and defenses in the district court, and we express no opinion as to whether any of those claims or defenses might ultimately prove meritorious.").

motion for preliminary injunctive relief were left largely unconsidered by the court, though on review the 9th Circuit Court of Appeals still took note of the fact that LinkedIn's user agreements expressly state that LinkedIn does not claim any ownership of the content its users add to the site.[127] As of writing this, LinkedIn has filed a petition for certiorari to the United States Supreme Court.[128]

Given the language of the CCPA, our hypothetical AI startup, crawling and scraping public-facing webpages, is probably engaged in "collecting" personal information. But, as discussed above, courts have been reluctant to limit the use of such public-facing information by third parties if the information does not enjoy copyright status (or some other IP right) and is not used in some other prohibited manner such as producing credit reports under the FCRA (as was the case in *Spokeo*). This reluctance goes directly to the core of the next question: whether the scraping of information placed on public facing pages *should* be considered conduct covered by the CCPA?

2. Web Scraping and Notification

As discussed above, the CCPA's expansive definition of "collect" seems to capture web scraping activity. But where does this leave web scrapers and the information they collect from public-facing webpages?

The CCPA's §1798.100(b) requires any qualifying business collecting covered consumer information to notify the consumers "at or before the point of collection" about which categories of information it plans to collect, and for what purposes.[129] However, in the proposed hypothetical, the businesses are not operating the sites that the bots are crawling and scraping. Rather, they are more like unexpected and perhaps unwelcome guests of a private dinner party held at a public restaurant. For a website or platform operating with consumer data of California residents the CCPA mandates that the website must disclose at the point of collection

[127] *Id*. (Writing, "hiQ advanced several affirmative claims in support of its request for preliminary injunctive relief, here we consider only whether hiQ has raised serious questions on the merits of its claims either for intentional interference with contract or unfair competition, under California's Unfair Competition Law, Cal. Bus. & Prof. Code § 17200 *et seq*. Likewise, while LinkedIn has asserted that it has 'claims under the Digital Millennium Copyright Act and under trespass and misappropriation doctrines,' it has chosen for present purposes to focus on a defense based on the CFAA, so that is the sole defense to hiQ's claims that we address here.").

[128] Petition for a Writ of Certiorari, *LinkedIn Corp. v. hiQ Labs.*, No. 19-1116 (S. Ct. filed Mar. 9, 2020) (S. Ct. docket files) *reh'g denied*, No. 17–16783 (9th Cir. Nov. 8, 2019).

[129] CAL. CIV. CODE § 1798.100(b) (West, Westlaw through Ch. 9 of 2021 Reg. Sess.).

Deleted: -

which information it seeks to collect and for what purposes it collects that information.[130] But where the CCPA seems to create affirmative duties for these legitimate websites, the language of the statute may not reach the bots operating as unwelcome guests recording and indexing everything they encounter.

Several reasons for excluding the activity of crawling and scraping information from public-facing webpages have already been mentioned or alluded to here. First, and perhaps the most straightforward of those reasons, is that the websites with which a consumer interacts with directly are most able to limit what types of information a user places on those sites. If the California legislature wanted to prevent consumers from placing any CCPA defined "personal information" on a website, it could enact a law providing for that purpose, though it seems equally clear that this is not a satisfying solution to the issues which the CCPA hopes to address. There is good reason to place "personal information" in the public sphere and there is also reason to find that when a consumer does so, that information has become part of the public domain. For example, the individual users of LinkedIn want to place true information about themselves in their profiles so potential employers and others can view it and hire them. Likewise, sole proprietors and others may place "personal information" about themselves on public-facing websites so that current and prospective customers know who they are doing business with.

Next, large search engines like Google and Bing use this method to scan and index hundreds of billions to trillions of web pages returning hundreds of millions of results for many searches frequently in fractions of a second.[131] By its own count, Google has indexed more than thirty trillion web pages.[132] The benefit of these tools is hard to overstate. These search engines have become an integral part of the everyday experience for millions of Americans (and probably billions of others around the world) allowing widespread access to unprecedented amounts of information.[133] "Google" is now so synonymous with "search" that its more often used contemporary meaning is as a verb, "to obtain information

[130] *Id.*

[131] John Koetsier, *How Google searches 30 trillion web pages, 100 billion times a month*, VENTUREBEAT (Mar. 1, 2013, 12:43 PM), https://venturebeat.com/2013/03/01/how-google-searches-30-trillion-web-pages-100-billion-times-a-month/ [https://perma.cc/329E-SRRA].

[132] *Id.*

[133] *How Search works: Overview: Organizing the content of the web*, GOOGLE, https://www.google.com/search/howsearchworks/ [https://perma.cc/QU36-JCC3] (describing the scope of Google's index, "[t]he index is like a library, except it contains more info than all the world's libraries put together").

about (someone or something) on the World Wide Web."[134] Preventing Google from indexing the information from public-facing websites would be extremely disruptive to users as well as entire industries that rely on the functionality of Google's search engine to function.

First Amendment considerations are noteworthy here too. Where an individual has placed information in the public sphere on public-facing webpages accessible to anyone with a web browser, preventing certain entities from accessing and using the information in a way that is not obviously detrimental to any party potentially runs afoul of constitutional rights to free speech. While this consideration is relevant and particularly noteworthy, adequately exploring this topic is beyond the scope of this article.

Lastly, a practical reason why website scraping should not be covered by the CCPA is that there is no technologically satisfying way to notify individual consumers whose information appears on public facing webpages being scraped by a third-party, short of notifying the website host that our hypothetical AI firm is engaged in scraping them. As we've seen, website scraping is not some activity used only by state-agencies engaged in clandestine information gathering.[135] Rather, it is technology which has contributed in part to the reason why the phrase, "Google it" has become so popular in common vernacular. Once reserved for the largest companies, scraping competitors' websites for information related to product offerings, price, etc. is now a common operation among many online retailers.[136] And for some businesses, like the AI startup from our hypothetical, the ability to access and use this information is what allows the company to exist at all.

While many of the bots deployed by sites like Google and Bing announce their arrival to a webpage and identify themselves as bots, presumably for their indexing functions discussed above, others attempt to camouflage or otherwise conceal themselves.[137] Many websites deploy bot blockers to stop certain bots they don't want scraping their pages while allowing those they do.[138] But in the war of the bots, each side is continuously upping the ante

[134] *Google*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/google?utm_campaign=sd&utm_medium=serp&utm_source=jsonld [https://perma.cc/924F-RHRC].

[135] Mehul Srivastava & Tim Bradshaw*, Israeli group's spyware 'offers keys to Big Tech's cloud'*, FIN. TIMES (July 18, 2019), https://www.ft.com/content/95b91412-a946-11e9-b6ee-3cdf3174eb89 [https://perma.cc/DJW4-W25A].

[136] Klint Finley, *'Scraper' Bots and the Secret Internet Arms Race*, WIRED (July 23, 2018, 7:00 AM), https://www.wired.com/story/scraper-bots-and-the-secret-internet-arms-race/ [https://perma.cc/JKY2-A9K6].

[137] *Id.*

[138] *Id.*

making bot-identifiers more savvy and bot-camouflaging more stealthy.

Bots which identify themselves as such, alerting the website to its activities can conceivably serve as notice to the website, and by extension the individual users posting PII to the site (via the hosting websites' terms of use), that the bot is collecting their information. Websites aware of such bot activity should include some language in their terms of use to the effect that information which is public facing is subject to collection by known and unknown third parties. However, this is not a satisfying solution. Imagine reading a statement like this as a consumer—you wouldn't know where to begin to look for the unknown third parties scraping and collecting your information.

Our hypothetical AI startup, quietly going about its business of scraping and collecting information from public-facing webpages, should probably have its own website and a way for consumers to make requests that their information not be collected. But assuming our AI startup company is not actively engaged in selling the personal information it collects, for example as a "data broker," how is any individual consumer to know their information has even been collected?[139] Short of the bot making itself known to the websites it scrapes, and thus risking being blocked, there is no way for the required notification to occur "at or before the point of collection" as required by the CCPA.

Gateways as simple as requiring users to login with valid credentials are enough to make public-facing webpages private as the court's in *Nosal II* and *Power Ventures* highlight. But as *hiQ* demonstrated, the courts have been unwilling to create a general policy whereby information which is made public by the person it identifies enjoys some special privacy rights.[140]

If there is not an easy way for a consumer to request from an AI company developing an algorithm the information the AI firm has collected about them (through their scraping activity), most consumers are not likely to go to great lengths to seek it out and in this scenario the CCPA has lost its teeth.

On March 11, 2020, during the course of writing this paper, the Attorney General of California issued proposed regulations to "establish procedures to facilitate consumer's new rights under the

[139] CAL. CIV. CODE § 1798.99.80 (West, Westlaw through Ch. 9 of 2021 Reg. Sess.) defines data broker as "a business that knowingly collects and sells to third parties the personal information of a consumer with whom the business does not have a direct relationship"; § 1798.99.82 requires data brokers to register with the California Attorney General.

[140] hiQ Labs v. LinkedIn Corp., 938 F.3d 985, 1005 (9th Cir. 2019).

CCPA and provide guidance to businesses for how to comply."[141] Under the proposed regulations, a business such as our hypothetical AI startup would not be required to provide a notice at the point of collection to consumers whose personal information was collected during the scraping of public webpages hosted by other websites.[142] Because the startup is not collecting the information directly from the consumer, but rather from the webpages with which the consumers interacted, the startup is therefore not collecting the consumers personal information "*directly from*" the consumer.[143] The business would have remaining obligations under the proposed regulations before it could "sell" the "personal information" but as explained below, the CCPA contains express exceptions under its definition of "sell" which are not modified by the proposed regulations.

These proposed regulations have been adopted and render much of the above analysis moot. However, the distinction between what would or would not qualify as a "sale" remains integral to this discussion. This note anticipates only the complete acquisition of the AI startup by a mature tech firm which may seem like a "sale" on its surface—after all, the AI startup is in fact being sold to another party—but as explained below, this transaction should not qualify under the CCPA's definition of "sale."

F. *Exceptions to CCPA Obligations*

Though the CCPA covers the sale of PII to third-parties, under the statute's definition of "sell" a specific exception is made for a company's sale of its assets to another company.[144] In our hypothetical, where the AI startup sells the PII as part of a training dataset to a mature tech firm, the sale would not itself constitute the sale of PII under the CCPA's definition.[145]

Section 1798.140(t)(2)(D) by its language generally excludes from the statutory definition of "sale" situations where a company sells controlling stake of its operations and assets to another company—an acquisition or merger. However, the definition goes on to require that the personal information must be used for the same purposes as, and within the scope of the stated terms of the company which originally acquired the consumer information.[146]

[141] *California Attorney General Issues Proposed CCPA Regulations*, JD SUPRA (Oct. 11, 2019), https://www.jdsupra.com/legalnews/california-attorney-general-issues-50572/ [https://perma.cc/926T-RMWC].

[142] CAL. CODE REGS. tit. 11 § 999.305(d) (2020).

[143] CAL. CODE REGS. tit. 11 § 999.305(d) (2020) (emphasis added).

[144] CAL. CIV. CODE § 1798.140(t)(2)(D) (Westlaw).

[145] *See id.*

[146] *Id.*

Furthermore, "[i]f a third party materially alters how it uses or shares the personal information of a consumer in a manner that is materially inconsistent with the promises made at the time of collection, it shall provide prior notice of the new or changed practice to the consumer."[147] This begs the question as to whether the company that scraped the PII from websites is in violation of the CCPA if it uses the information differently from the website that hosted the user information originally. If for example, a user's information was placed on public-facing pages of a website and a different company scrapes that website for the user-information, the company scraping the website has at no time made any promises to the hosting site's users. Because it is not the business that originally collected the information (the website being scraped), there may be no affirmative obligations on the party doing the scraping—only on the party that was being scraped.

More importantly, the CCPA exempts from the definition of "sale," the acquisition of a company which has collected personal information from California consumers: a "business does not sell personal information when, [t]he business transfers to a third party the personal information of a consumer as an asset that is part of a merger, acquisition [], or other transaction in which the third party assumes control of all or part of the business."[148]

The acquiring business may only use the information assets acquired from the target of the acquisition consistent with the disclosures to the consumers regarding how the information would be used by the business which originally collected the information.[149]

For the sake of argument, let's assume that our AI startup is sold to a large software company with a lot of notoriety and a large footprint. Because our AI startup is scraping the websites of other businesses, it has not had the ability to notify consumers that their information is being collected in this manner and thus has made no promises or representations as to how the information is being used, what information is being collected, or to whom the information is being sold (because it is not engaged in the sale of the PII). Let's also assume that the mature tech company has been preparing for the CCPA requirements and has a webpage and a form dedicated to informing consumers about their rights under the CCPA and allowing consumers to make requests for information from the company.

[147] *Id.*
[148] *Id.*
[149] *Id.*

Does the big software company, having recently acquired the training dataset containing PII as well as the AI algorithm from our AI startup, now have to disclose the personal information it has acquired as a result of the acquisition? Probably not. The CCPA distinguishes, probably inadvertently, between information collected by the big software company and personal information collected by some other business which it now owns as a result of the acquisition.[150] The language of the first consumer right created under the CCPA reads as follows: "[a] consumer shall have the right to request that a *business that collects a consumer's personal information* disclose to that consumer the categories and specific pieces of personal information *the business has collected*."[151] For example, if the big software company did not have any direct interaction with the consumer and never collected consumer personal information itself, but acquired such information through the acquisition, the 'acquisition exception' exempts its acquisition of the information from the definition of collection. The definition of "collect" should be limited by the language of the acquisition exemption which reads,

> If a third party materially alters how it uses or shares the personal information of a consumer in a manner that is materially inconsistent with the promises made at the time of collection, it shall provide prior notice of the new or changed practice to the consumer. The notice shall be sufficiently prominent and robust to ensure that existing consumers can easily exercise their choices . . . .[152]

In the event of our hypothetical acquisition, the party that originally collected the information is dissolved and only the mature tech company remains. Thus, only the acquiring company can notify the consumers whose information was collected as to whether it plans to use differently the information included as an asset in the acquisition. The fact that this section goes into detail about how the acquiring party (the "third-party") may use the information, and notes that the information was collected prior to the acquisition, counsels that the acquiring company has not "collected" the consumer information by the CCPA's definition.[153] So, while the CCPA's definition of "collects" is broad, the inclusion of this language under the acquisition exception should be read to

[150] CAL. CIV. CODE § 1798.100(a) (Westlaw).
[151] *Id.* (emphasis added).
[152] CAL. CIV. CODE § 1798.140(t)(2)(D) (Westlaw).
[153] *Id.*

limit the definition of "collect" under the specific hypothetical contemplated in this note.

However, the big software company probably has every reason to comply with the consumer request and disclose every piece of personal information it has about the consumer. The CCPA outlines several exceptions for which a business can deny a consumer's request to delete their information, including among them, "internal uses that are reasonably aligned with the expectations of the consumer based on the consumer's relationship with the business,"[154] and to "[o]therwise use the consumer's personal information, internally, in a lawful manner that is compatible with the context in which the consumer provided the information."[155] First and foremost, if the information is necessary for the internal use of the business—as may be the case if the information is used in the training dataset of an AI algorithm underlying a portion of the company's software stack—it can claim this reason and be exempt from a consumer request to delete the information.[156] The statute's broad language arguably exempts the acquiring company which itself made no representations when the information was transferred through acquisition and where the scraping company (the AI startup) similarly made no representations when performing the scraping activity.[157] However, the software company will still be required to respond to the consumer request for information as no qualifying business is exempt from this obligation and here, the software company likely qualifies as a "business" covered by the CCPA due to its size and digital footprint.

Secondly, from a more practical business perspective, the software company should honor consumer requests to maintain goodwill among its customers. No business will want a reputation as a company that does not take seriously the privacy concerns of its customers. Along the same lines, a business which is not responding promptly or adequately may soon find itself in the crosshairs of the California Attorney General.

[154] CAL. CIV. CODE § 1798.105(d)(1)–(9) (Westlaw).

[155] *Id.* §§ 1798.105(d)(7), 1798.105(d)(9) (Westlaw).

[156] *See* CAL. CIV. CODE § 1798.105(d) (Westlaw).

[157] Tara N. Cho et al., *New CCPA Changes/Clarifications; Some Final, Some Contingent (2 Months to Go)*, NAT. L. REV. (Oct. 24, 2019), https://www.natlawreview.com/article/new-ccpa-changesclarifications-some-final-some-contingent-2-months-to-go [https://perma.cc/CV3G-NVEY].

G. Remedies and the Role of the California Attorney General

Despite being regarded as the most robust privacy law in the United States, the CCPA creates only a limited private right of action for violations.[158] A private civil action is limited to disclosure of "nonencrypted or nonredacted personal information," a defined subset of consumer information found in the California data breach statute.[159] "Encrypted" information, according to the definition provided there, means "rendered unusable, unreadable, or indecipherable to an unauthorized person through a security technology or methodology generally accepted in the field of information security."[160] The California Attorney General, delaying enforcement of the CCPA until July of 2020, provided companies additional time, allowing them to make necessary changes to their business practices in order to comply with the regulation.[161] The law itself provides that an individual consumer may institute a civil action for statutory damages between $100–$750 per incident.[162] Additionally, violations of the statute are enforceable by the California Attorney General with fines up to $7,500 per incident.[163] While the statute does not define "incident," it does provide the criteria allowing a consumer to institute a civil action:

> Any consumer whose nonencrypted and nonredacted personal information . . . is subject to unauthorized access and exfiltration, theft, or disclosure as a result of the business's violation of the duty to implement and maintain reasonable security procedures and practices appropriate to the nature of the information to protect the personal information may institute a civil action. . . .[164]

The California Attorney General has issued multiple rounds of proposed modifications to the CCPA.[165] These proposed modifications changed, and changed again the requirements of businesses not collecting information from consumers,[166] and left some questions unanswered, but what seems clear is that the

[158] *See* CAL. CIV. CODE § 1798.150(a)(1) (Westlaw).
[159] *Id.*
[160] CAL. CIV. CODE § 1798.82(i)(4) (Westlaw).
[161] John Stephens, *California Consumer Privacy Act*, A.B.A. (Feb. 14, 2019), https://www.americanbar.org/groups/business_law/publications/committee_newsletters/bcl/2019/201902/fa_9/ [https://perma.cc/UP5T-UKXQ].
[162] CAL. CIV. CODE § 1798.150(a)(1)(A) (Westlaw).
[163] Stephens, *supra* note 161.
[164] CAL. CIV. CODE § 1798.150 (Westlaw).
[165] *See* California Consumer Privacy Act, 41-Z Cal. Reg. Notice Reg. 1341 (Oct. 11, 2019).
[166] *Id.*

California Attorney General will have limited resources with which to enforce the CCPA's mandates. Some commenters have predicted that initial enforcement efforts will likely be aimed at larger business collecting vast amounts of personal information as opposed to small and medium sized businesses.[167]

H. The California Privacy Rights Act of 2020 (CPRA)

The CPRA, a California ballot measure which passed in November 2020, makes several important changes to the CCPA.[168] Effective January 1, 2023,[169] the CPRA changes the statutory definition of "personal information,"[170] provides a definition for an entirely new term, "[s]ensitive personal information,"[171] and imposes obligations on businesses which "share" consumers' personal information,[172] among other amendments to the CCPA.

Again though, the statute provides that "sharing" does not include the transfer of a consumer's information "as an asset that is part of a merger, acquisition. . ." consistent with the carveout to the statutory definition of "sale." The carveouts remaining unchanged, coupled with the revised definitions of "personal information," indicate that a similar outcome to the hypothetical scenario considered here will likely result after the bill's effective

[167] Allison Schiff, *It May Seem All Quiet On The CCPA Front, But Don't Get Complacent: CCPA Enforcement Has Begun*, ADEXCHANGER (Sept. 28, 2020, 12:35 AM), https://www.adexchanger.com/privacy/it-may-seem-all-quiet-on-the-ccpa-front-but-dont-get-complacent-ccpa-enforcement-has-begun/ [https://perma.cc/FSN8-MH22].

[168] Brandon P. Reilly & Scott T. Lashway, *The California Privacy Rights Act Has Passed: What's in It?*, MANATT (Nov. 11, 2020), https://www.manatt.com/insights/newsletters/client-alert/the-california-privacy-rights-act-has-passed [https://perma.cc/GE6U-DRD9].

[169] 2020 Cal. Legis. Serv. Prop 24 (West).

[170] CAL. CIV. CODE § 1798.140(v)(1)(L)(2) (West, Westlaw through Ch. 9 of 2021 Reg. Sess.) (effective Jan. 1, 2023) (amending the definition to include, "information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media, or by the consumer; or information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience.").

[171] *Id.* § 1798.140(ae) (Westlaw) (defining the term capaciously, the changes reflect the desire by California officials and the state's electorate to broaden privacy rights over time and to provide Californians with a privacy right styled in a more "European" fashion).

[172] *Id.* § 1798.100(d) (Westlaw); *Id.* § 1798.140(ah) (Westlaw) (defining "sharing" as "sharing, renting, releasing, disclosing, disseminating, making available, transferring, or otherwise communicating orally, in writing, or by electronic or other means, a consumer's personal information by the business to a third party for cross-context behavioral advertising, whether or not for monetary or other valuable consideration, including transactions between a business and a third party for cross-context behavioral advertising for the benefit of a business in which no money is exchanged").

date. Notably, the statutes' definition of "collects" is unchanged by the CPRA.[173]

The statute makes several other noteworthy changes. One particularly relevant change is the creation of the California Privacy Protection Agency (the Agency).[174] The Agency is tasked with implementing and enforcing the CCPA's obligations.[175] Broadly, the Agency is charged with "protect[ing] the fundamental privacy rights of natural persons with respect to the use of their personal information,"[176] and "seek[ing] to balance the goals of strengthening consumer privacy while giving attention to the impact on businesses."[177]

CONCLUSION

As noted earlier, this paper explores the implications presented under a very specific scenario and against a backdrop of federal law that deals with specific types of information sector by sector, industry by industry. With this backdrop in mind, the California legislature passed one of the most—if not *the* most—robust privacy laws in the United States. As described, the law creates a number of rights for California residents and a number of obligations on businesses doing business in California and collecting consumers' personal information.

The specific scenario envisioned in this note was intended to illustrate the difficulties of crafting a far-reaching privacy law like the CCPA, by demonstrating at least one example where the law is not truly capable of adequately dealing with current business practices. While the CCPA's broadly drafted language brings the conduct of the hypothetical AI startup within its reach, it also provides for the release of our hypothetical company from its grasp. These problems further counsel that a federal omnibus privacy law is needed. The sectoral and state-by-state patchwork of digital privacy laws creates a tangled web of legislation in which businesses increasingly find themselves caught up. The CCPA, with its progressive approach to creating a more consumer-friendly environment for web users, may be a model, but certainly should not be the standard for an omnibus federal privacy law.

[173] *Compare id.* § 1798.140(e) (Westlaw) *with id.* § 1798.140(f) (Westlaw).
[174] *Id.* § 1798.199.10 (Westlaw).
[175] *Id.* § 1798.199.40(a) (Westlaw).
[176] *Id.* § 1798.199.40(c) (Westlaw).
[177] *Id.* § 1798.199.40(l) (Westlaw).