CONTENTS

How Machines Learn: Where do Companies get Data for Machine Learning and What Licenses do They Need?
  Rachel Wilka, Rachel Landy, and Scott A. McKinney
  217

Fair Use, Fair Play: Video Game Performances and “Let’s Plays” as Transformative Use
  Dan Hagen
  245

Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping
  Zachary Gold and Mark Latonero
  275

Smart Contracts, Blockchain, and the Next Frontier of Transactional Law
  Scott A. McKinney, Rachel Landy, and Rachel Wilka
  313
## EDITORIAL BOARD

**Editor-in-Chief**  
Ari Mead Robbins

**Associate Editor-in-Chief**  
Rob Philbrick  
Dan Goodman

<table>
<thead>
<tr>
<th>Title</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chief Managing Editor</td>
<td>Daniel Hagen</td>
</tr>
<tr>
<td>Senior Managing Editors</td>
<td>Mariko Kageyama</td>
</tr>
<tr>
<td></td>
<td>Sebastian Stock</td>
</tr>
<tr>
<td></td>
<td>Jacob Knutson</td>
</tr>
<tr>
<td></td>
<td>Derk Westermeyer</td>
</tr>
<tr>
<td>Chief Article Editor</td>
<td>Daniel Healow</td>
</tr>
<tr>
<td>Senior Articles Editors</td>
<td>Adam Roberts</td>
</tr>
<tr>
<td></td>
<td>Carolea Casas</td>
</tr>
<tr>
<td></td>
<td>Michael Rebagliati</td>
</tr>
<tr>
<td></td>
<td>Steven Matyas</td>
</tr>
<tr>
<td></td>
<td>Miles Bludorn</td>
</tr>
<tr>
<td></td>
<td>Yonah Reback</td>
</tr>
<tr>
<td>Chief Online Editor</td>
<td>Toban Platt</td>
</tr>
<tr>
<td></td>
<td>Amelia Zukic</td>
</tr>
<tr>
<td>Chief External Editor</td>
<td>Seth Parent</td>
</tr>
</tbody>
</table>

## EDITORIAL STAFF

<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna McCurley</td>
<td>Alexander Hagel</td>
<td>Alex Palumbo</td>
</tr>
<tr>
<td>Brittany Taylor</td>
<td>Christine Steenstrup</td>
<td>Craig Dammeier</td>
</tr>
<tr>
<td>Evan Fowler</td>
<td>Gardner Reed</td>
<td>Hugo Fraga</td>
</tr>
<tr>
<td>Isaac Prevost</td>
<td>Jabi Diagana</td>
<td>Joshua Oh</td>
</tr>
<tr>
<td>Mika Rothman</td>
<td>Vanessa James</td>
<td>William Cason</td>
</tr>
<tr>
<td>Christopher J Lindeymeier</td>
<td>David O’Hair</td>
<td>Jenna Belmonte</td>
</tr>
</tbody>
</table>

## EXTERNAL BOARD

<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex Alben</td>
<td>Jen Chang</td>
<td>Bob Cumbow</td>
</tr>
<tr>
<td>Kaustuv Das</td>
<td>Trent Dykes</td>
<td>Jennifer Fan</td>
</tr>
<tr>
<td>Parag Gheewala</td>
<td>Robert Gomulkiewicz</td>
<td>Chris Igielski</td>
</tr>
<tr>
<td>Jessie Kennedy</td>
<td>Tim Lee</td>
<td>Ashley Long</td>
</tr>
<tr>
<td>Gregory McNeal</td>
<td>Sean O’Connor</td>
<td>Lonnie Rosenwald</td>
</tr>
<tr>
<td>Christopher Sweeney</td>
<td>John Weresh</td>
<td></td>
</tr>
</tbody>
</table>
HOW MACHINES LEARN: WHERE DO COMPANIES GET DATA FOR MACHINE LEARNING AND WHAT LICENSES DO THEY NEED?

Rachel Wilka, Rachel Landy, and Scott A. McKinney*
© Rachel Wilka, Rachel Landy, and Scott A. McKinney

Cite as: 13 Wash. J.L. Tech. & Arts 217 (2018)
http://digital.law.washington.edu/dspace-law/handle/1773.1/1815

ABSTRACT

Machine learning services ingest customer data in order to provide refined, customized services. Machine learning algorithms are increasingly prominent in multiple sectors within the software-as-a-service industry including online advertising, health diagnostics, and travel. However, very little has been written on the rights a company utilizing machine learning needs to obtain in order to use customer data to improve its own products or services.

Machine learning encompasses multiple types of data use and analysis, including (a) supervised machine learning algorithms, which take specific data provided in a tagged and classified format to deliver specific predictable output; and (b) unsupervised machine learning algorithms, where untagged data is processed in

* Scott McKinney and Rachel Landy are senior associates in the Technology Transactions practice at Wilson Sonsini Goodrich & Rosati, P.C. (WSGR). Scott is an adjunct professor at Georgetown Law and a guest lecturer for Cornell University’s Cornell Tech grad program. Rachel represents numerous technology companies on matters relating to intellectual property and commercial contracts. Rachel Wilka is Corporate Counsel at Zillow Group, Inc. and lead counsel for Zillow Rentals, hotpads, Inc., and dotlop, Inc., supporting all product counseling, licensing, commercial partnership, and risk management matters. Rachel was previously a technology transactions associate at Wilson Sonsini Goodrich & Rosati. The views expressed in this article are the authors’ own and do not represent the views of WSGR, Zillow, or any of the authors’ other clients. Thank you to Manja Sachet and Rob Philbrick.
order to look for patterns and correlations without a specified output.

This Article introduces the reader to the types of data use involved in various machine learning models, the level of data retention normally required for each model, and the risks of using personal information or re-identifiable data in connection with machine learning. The paper also discusses the type of license a commercial provider and consumer would need to enter into for various types of machine learning software. Finally, the paper proposes best practices for ensuring adequate rights are obtained through legal agreements so that machines may self-improve and innovate.

**TABLE OF CONTENTS**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>219</td>
</tr>
<tr>
<td>I. Background</td>
<td>220</td>
</tr>
<tr>
<td>A. Definition of Machine Learning</td>
<td>220</td>
</tr>
<tr>
<td>B. Types of Machine Learning</td>
<td>222</td>
</tr>
<tr>
<td>1. Supervised</td>
<td>222</td>
</tr>
<tr>
<td>2. Unsupervised</td>
<td>223</td>
</tr>
<tr>
<td>3. Reinforcement</td>
<td>224</td>
</tr>
<tr>
<td>II. Levels of Data Use Associated With Different Machine Learning Models</td>
<td>225</td>
</tr>
<tr>
<td>A. Supervised</td>
<td>226</td>
</tr>
<tr>
<td>B. Unsupervised</td>
<td>227</td>
</tr>
<tr>
<td>III. Retention</td>
<td>229</td>
</tr>
<tr>
<td>IV. Sources of Data</td>
<td>231</td>
</tr>
<tr>
<td>A. Data Sets Sold Through Data Brokers</td>
<td>231</td>
</tr>
<tr>
<td>B. Ongoing Customer Data Collection From Network-Connecting Software as a Service Offering</td>
<td>232</td>
</tr>
<tr>
<td>C. Batch Uploaded Data From Software Installed On-Premises for Customers</td>
<td>232</td>
</tr>
<tr>
<td>D. Open Source Public Data Sets</td>
<td>233</td>
</tr>
<tr>
<td>V. Laws/Legal Risks Around Use of Data/PII in Machine Learning</td>
<td>233</td>
</tr>
<tr>
<td>A. Use of Sensitive Data</td>
<td>234</td>
</tr>
<tr>
<td>B. The Output Use Case</td>
<td>235</td>
</tr>
</tbody>
</table>
Machine learning—it’s been a technology catch-phrase for at least five years, a tagline for any company purporting to “innovate a new future,” but what does it actually mean? Machine learning services ingest data in order to provide refined, customized services to users.¹

Real world utilization of machine learning increases daily, as more and more companies use the technology for market trend analysis, price setting, development of company (or industry) best-practices, medical diagnoses, insurance—virtually any industry that has representable and analyzable output information can be optimized through machine learning.²

---

The algorithms that drive machine learning are increasingly prominent within the software-as-a-service industry, where machine learning can be leveraged for multiple industries, including online advertising, health diagnostics, and travel. Despite the increased use of machine learning across business sectors, the rights a company utilizing machine learning needs to obtain in order to use outside data to improve its own products are often amorphous and misunderstood. As machine learning becomes integral to companies across all industries and those companies become more and more reliant upon datasets for use in their machine learning analysis, the data itself (and the corresponding rights in such data) becomes increasingly important.

This Article examines the legal data rights a company needs to obtain in order to use data for machine learning, and how those rights change depending on the machine learning model and business application. Part I of this Article defines machine learning and analyzes the various use cases for machine learning based on differing data rights. Part II discusses how companies may use data for different purposes. Part III discusses the varying degrees of data retention a company may undertake. In Part IV, we follow that discussion with an overview of data sources a machine learning company could access. Part V discusses the laws and legal risks relating to the use of data (including personally identifiable information (“PII”)) in machine learning applications across commercial sectors. Lastly, Part VI provides recommendations and considerations for drafting data licenses.

I. BACKGROUND

A. Definition of Machine Learning

The term “machine learning”, which is widely credited to ex-
IBM employee Arthur Samuel,⁴ is the ability of computers (“machines”) to learn without being guided or re-programmed.⁵ Samuel’s initial machine learning example was a machine that can be programmed to play checkers better than the person who designed the program. Remarkably, a computer could be trained to do this in eight to ten hours of playing time over sixty years ago using machine learning.⁶ All that was necessary to train the computer was to provide it with the rules of the game, a general sense of direction regarding how the game worked, and a list of parameters that were thought to have something to do with the game, but whose correct background signs and relative importance were unknown and unspecified to the computer.⁷ In relatively short order, the machine learned how to play checkers better than its programmer, without the programmer having to revise the initial computer code or manually train the computer in strategy.⁸

The use cases for modern machine learning are virtually boundless. Machine learning is best used in tasks for which designing code with explicit task-specific instructions is difficult or impossible, such as ranking, optical recognition, complex problem solving, and filtering.⁹ Machine learning applications typically involve feeding (relatively) automated programs a large data set of inputs, and solving problems or identifying issues using results-driven decisions based on the data set.

To be clear, machine learning (in the classic sense) is not artificial intelligence. Although machine learning does involve learning by experience, a machine learning algorithm does not act intelligently,¹⁰ and is not flexible in changing environments.¹¹ However, we see the concepts become increasingly conflated, as

⁵ Id.
⁶ Id.
⁷ Id.
⁸ Id.
⁹ ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING 6–8 (3rd ed. 2014).
¹⁰ See discussion infra Part I.B.
algorithms are commonly programmed with artificial intelligence, and as machine learning algorithms come to make up a greater part of the artificial-intelligence ecosystem. Machine learning should not be conflated with data mining, either. Unlike data mining, which usually focuses on uncovering previously unknown properties of a dataset, machine learning typically focuses on better-predicting outcomes or revising an algorithm based on already-known properties of that dataset.

Below we discuss the common types of machine learning and the different levels of data use associated with different machine learning models.

**B. Types of Machine Learning**

Machine learning can be split into three major categories: (1) supervised, (2) reinforcement, and (3) unsupervised. We discuss each in turn below.

1. Supervised

With supervised machine learning, one knows the desired output of the algorithm based on a dataset, usually referred to as “training data,” that is used to optimize a performance criterion. Supervised machine learning algorithms are typically “taught” using a training dataset. If the algorithm provides unexpected or incorrect results

---

13 But see ALPAYDIN, supra note 9, at 2 (describing the application of machine learning methods to a database as “data mining.”). Opinions regarding, and semantical definitions of the term “machine learning” vary.
15 Id.; see also Data Sets and Machine Learning, DEEP LEARNING FOR JAVA https://deeplearning4j.org/data-sets-ml (last visited Mar. 31, 2018); ALPAYDIN, supra note 9, at 3.
after analyzing the base data using the training dataset, the programmer can make algorithmic tweaks (or changes to the training data) to right the course. In supervised machine learning, all of the data within a training data set is “labeled” (or assigned a value), which allows the machine to easily compare analysis data against the training set baseline. The algorithm generates information based on its analysis of the training data, and uses that information to produce inferred or revised functions. These revised functions can be used by the end user to discern new trends regarding a dataset, or to refine the algorithmic analysis itself. Analyzing enormous data sets at a speed only computers can achieve, the algorithm can identify trends, flag otherwise unidentified issues, and give the algorithm operator other desired results that can be tweaked using variations in the algorithm or training data.

2. Unsupervised

In unsupervised machine learning, there is no training data, and the outcomes are unpredictable. Unsupervised machine learning algorithms can solve problems using input datasets alone, with no reference or training data, by recognizing patterns in the data and grouping together reoccurring or common data characteristics. Unlike supervised algorithms, which rely on labeled data, unsupervised machine learning uses functions to uncover previously unknown properties of a dataset using unlabeled data. For example, say you had a dataset comprised of apples, oranges, and bananas, and want to analyze and identify trends in the fruit. The problems are: the data set is huge, the fruit are all jumbled together, and none of the data is labeled as an “apple,” an “orange,” or a “banana.” In a supervised machine learning scenario, if the algorithm was not “taught” to identify an apple, it would not know to look for, nor group together, apples. In contrast, an unsupervised machine learning algorithm is able, over time, to recognize that data across the datasets have similar characteristics, such as being shiny, red,
and generally apple-shaped. Unsupervised algorithms can identify these similarities and group together the apples with the apples, the oranges with the oranges, and the bananas with the bananas. Unsupervised machine learning can seem to border on artificial intelligence,20 and companies often use it to analyze large datasets of customer transactions, generate common trends or characteristics based on the past transactions, group those customers into clusters, and use that cluster of information to refine the company’s business model.21

There is a sub-class of supervised machine learning called “semi-supervised” machine learning, in which an algorithm-operator uses a small amount of labeled training data to inform a much larger unlabeled dataset.22 Semi-supervised machine learning is usually thought of as halfway between unsupervised and supervised learning.23 Both supervised and semi-supervised machine learning tend to lend themselves to relatively predictable outcomes, and are often used by companies to optimize user experiences based on predicted or predetermined outcomes.

3. Reinforcement

Reinforcement learning is based on an algorithm that has a concept of how an environment should behave, and learns an optimal behavior for such an environment by analyzing repetition and repeated failures over time.24 Unlike supervised machine learning, reinforcement learning algorithms are not presented with input/output pairs for correction—instead, the algorithm is performance-driven.25 One well-known example of reinforcement

21 ALPAYDIN, supra note 9, at 12.
22 CHAPELLE, ET AL., supra note 14, at 2–3.
23 Id.
25 Id.
learning is the self-driving car industry. Many self-driving algorithms are not artificially intelligent in the traditional sense, but instead use repetition (i.e. driving thousands of test miles and tracking driving errors and successes) to optimize the algorithm and underlying technology in a way that human programmers could never do on their own. Another way to think about reinforcement learning is “trial-and-error”, but on a massive scale accomplishable only by computers. Over time, the software learns what to do, and what not to do, until its functionality is optimized for the task at hand.

II. LEVELS OF DATA USE ASSOCIATED WITH DIFFERENT MACHINE LEARNING MODELS

The use case for machine learning implementation dictates the data rights that must be obtained, as well as the applicable data retention and use policies. For example, consider these three different use cases:

- OpenTable recommends restaurants, but can only do so based on the information it collects (e.g. where the user has dined before, not the actual dish he or she actually eats—information OpenTable does not have).

- To predict which show a user will want to binge next, Netflix wants to know that user’s viewing history, and some relevant demographic information, such as age, gender, and location.

- Accolade’s Maya Intelligence Option inputs information

---


27 Id.

28 Maini, supra note 14.


about an individual’s health insurance, medical history, medications, test results, and other personal health information in order to provide personalized healthcare support.\textsuperscript{31}

Like all companies that depend on machine learning, these companies obtain, use, and retain data in different ways, depending on their business model and their machine learning models.

\textit{A. Supervised}

Supervised machine learning presents clearer use cases. The outcome is predictable, and in fact, programmed. Netflix and OpenTable, for example, ingest user preference data to produce individualized recommendations to that user. These algorithms do not necessarily rely on extraneous data inputs—they are trained to provide recommendations if certain inputs are present. But by continuously ingesting new data, the engine can be refined and perfected on an ongoing basis. For example, over time, Netflix may be able to distinguish between medical-drama fanatics who want to binge Grey’s Anatomy and those who prefer ER. For this reason, the results of supervised machine learning can be highly valuable to companies in any industry, but especially those industries that are consumer-facing.

However, for both Netflix and OpenTable, the use of the data (recommendations) is not these companies’ core business; rather, it is an added feature that has helped propel both companies to the top of their respective industries. Without compelling recommendations, Netflix would still be a video streaming service. However, it relies on data to enhance the user’s experience, thus adding value to the service.\textsuperscript{32} Netflix does this by ingesting and inferring from a user’s preferences. For example, it knows if you watched one episode of Gilmore Girls, or if you watched every

\begin{itemize}
\item \textsuperscript{31} ACCOLADE, https://www.accolade.com/solutions/ (last visited March 30, 2018).
\end{itemize}
season five times, and it can use that information to determine whether you were a superfan or lost interest quickly.

The same is true, to a lesser extent, with OpenTable. OpenTable bases its recommendations largely on collections of user ratings. However, OpenTable’s capabilities are limited. Its model does not know whether its users actually ate at a restaurant booked through OpenTable. It only knows how that user feels about the restaurant if he or she rates it on the app. Furthermore, the app does not know, for example, whether dietary preferences affected that rating.

One benefit of supervised machine learning algorithms is that, in the early stages, potential data sets can be separated into those that are necessary and those that are merely helpful. A company may find that data sets with particular characteristics are subject to more extensive regulations than the data required to successfully implement a machine learning solution. As a result, the company will either utilize the data differently, or avoid implementation of the data altogether. For example, Netflix, in its early days, may have found that age was highly useful. However, unless the appropriate controls are in place, gathering other sensitive information, such as children’s’ names, can result in significant legal risk. Nevertheless, using machine learning, a start-up company may find that it can estimate age based on user habits, thereby making it unnecessary to undertake the legal risk of gathering that information directly.

B. Unsupervised

Using unsupervised machine learning is a process best thought of as “high risk, high reward.” Without a clearly defined desired


35 This is contrary to companies operating in the healthcare space, which almost always need some level of personal health information—another highly regulated category of data. For those companies, the risk is inherent in the business and should be priced into the model for customers.
output, the company may not know what it needs, or even what it is likely to get, from the algorithm. On the other hand, a company might get results that it did not anticipate or even think were possible. Unsupervised machine learning is popular in the health-tech industry because making a diagnosis requires analyzing many variables that human doctors cannot necessarily test for individually.\textsuperscript{36} Machine learning gives doctors the assistance they need to take in a large amount of data and then spit out all known potential diagnoses. The Maya Intelligence Option, for example, could benefit from taking in numerous health data points in order to generate a potential treatment plan, the scope of which would not be pre-defined.

Unsupervised machine learning, by its nature, requires that the operator have more flexibility in its use of data sets. As a result, the data use rights obtained from data providers (discussed in Part V) for use in unsupervised machine learning analysis should be broader than data use rights for supervised machine learning. For example, speech recognition software operators obtain broad rights to use data collected through the software (i.e. users’ speech). The Apple Terms of Service state: “By using Siri or Dictation, you agree and consent to Apple’s and its subsidiaries’ and agents’ transmission, collection, maintenance, processing, and use of this information, including your voice input and User Data, to provide and improve Siri, Dictation, and dictation functionality in other Apple products and services.”\textsuperscript{37} While Apple’s main purpose in collecting this data is likely to tune its engine to recognize speech more efficiently, such a broad license also allows the operator to use the speech for a number of ancillary purposes, such as understanding dialects, intonations, and speech impediments. Thus, the operator is not sure what the results will be or how those results may be used in the future. Indeed, an operator may find that certain data sets once considered vital turn out to be useless. Prior to implementation, the machine learning algorithm cannot necessarily predict which data is valuable and


\textsuperscript{37} \textit{Apple Ios Software User Agreement}, \textsc{Apple Inc.}, at 3 (emphasis added) available at https://www.apple.com/legal/sla/docs/ios6.pdf (last revised 2012).
which is not. This uncertainty necessitates a broader, less restrictive scope of operator rights than in other scenarios. In some cases, this may mean that the operator must assume the additional risks of using, collecting, or storing data that is subject to regulation.

Overall, companies’ use cases and data supply needs should help inform whether their algorithms are unsupervised, reinforced, or supervised. Accordingly, the rights to be obtained to that data, discussed in Part V, should reflect those business decisions. Moreover, in addition to the data use rights that must be obtained, we must also consider the data storage and retention issues associated with machine learning.

III. RETENTION

In addition to determining whether an algorithm should be supervised or unsupervised, any machine learning company must determine the scope of its data retention policy. Data retention policies track how data is stored, shared, and deleted to ensure consistency of data treatment and compliance with contractual obligations, applicable law, and best practices. As discussed in Part II, the particulars of a data retention policy for a machine learning company rely on the use case for the algorithm and the data-treatment requirements imposed by the data source.

For example, a supervised machine learning environment may only need to retain training data if it is not using new data to improve its capabilities. Or, it may only need to retain the data for a limited period of time in order to establish overall patterns or features to include in training data. In our Netflix example, it may be helpful for Netflix to know that over a two-year period, a user watched all of Dawson’s Creek, Gilmore Girls, and 7th Heaven, but not Buffy the Vampire Slayer.38 Knowing, in context, that the user prefers real-life teen dramas to science-fiction teen dramas can help improve the algorithm.

By contrast, an OpenTable user’s eating habits may not follow predictable patterns. The fact that a user ate at a Chinese restaurant five days in a row is helpful for understanding the user’s culinary tastes that week. But that same user could then decide she’s had

---

38 This assumes that all of the programs mentioned are available on Netflix.
enough Chinese food for a year, and move on to sushi. Thus, for OpenTable, pattern analysis is less important than it is for Netflix; it can simply build on each data input individually without a longer-term analysis. Where Netflix may be able to determine that a user had a child based on a change in viewing habits (and could adjust accordingly), OpenTable’s use case doesn’t require a long data retention period to provide a benefit.

Ultimately, assuming the operator has obtained the requisite rights from users (discussed in Part V), the operator ought to retain the data for as long as is commercially reasonable (although the relevant industry market approach may dictate that data be destroyed after a certain amount of time). To mitigate the potential harm of data destruction requirements, an operator should always retain the training data it used to fix bugs and help tune the algorithm. Other than the training data, a company could find that it need not retain a lot of individual data inputs so long as the algorithm has previously ingested, responded, and reacted to the data.

Some data providers try to contractually require data destruction after the term of an engagement. Operators of unsupervised algorithms should always push back; the nature of those algorithms is such that there could always be a golden needle in a data-haystack, so an operator should try to retain the right to continue to mine the data for as long as possible. If a customer is insisting on destruction, an operator may promise anonymization and aggregation of the data so the customer could not be identified. Ultimately, the operator must determine at what point the algorithm (and the operator’s business) will be able to live without the data, i.e., when it has obtained sufficient replacement data to be self-sustaining. In other words, what retention term is reasonable for the company? The operator may be able to compromise by agreeing to only use a customer’s data in perpetuity where that data is anonymized and aggregated with other customers’ data sets. A company that destroys data will also need to develop an appropriate support policy if the original reference set is eventually deleted.

---

IV. SOURCES OF DATA

Companies looking to obtain data to create or train machine learning algorithms tend to look to four sources: (a) data sets sold through data brokers; (b) batch uploaded data from software installed on-premises for customers; (c) ongoing customer data collection from network-connected software as a service offering (both for customer-facing improvements and other company purposes); and (d) open public data sets.\(^\text{40}\)

A. Data Sets Sold Through Data Brokers

Data brokers are companies that have gradually built databases of consumer data. These databases were originally built for “marketing, fraud detection, and credit scoring purposes.”\(^\text{41}\) Companies can go to data brokers to purchase data sets, usually with personally identifiable information removed. Data brokers may offer a database (or set of databases) that tracks behaviors the operator wants to build a machine-learning algorithm around. Data broker databases can include demographic data, court and public records data, social media and technology data, consumer interests data, financial data, health data, and purchase behavior data.\(^\text{42}\) However, some observers doubt whether data broker databases are sufficiently anonymized to avoid business or regulatory risk.\(^\text{43}\) Another downside of purchased data is that the purchaser runs the


\(^{42}\) See Leo Mirani & Max Nisen, The Nine Companies That Know More About You Than Google or Facebook, QUARTZ (May 27, 2014), https://qz.com/213900/the-nine-companies-that-know-more-about-you-than-google-or-facebook/.

risk of the data not being tailored to its exact needs, thereby making it less useful in providing the desired predictive output.\textsuperscript{44} The largest American data brokers include Acxiom, Corelogic, and Datalogix.\textsuperscript{45}

\textbf{B. Ongoing Customer Data Collection From Network-Connecting Software as a Service Offering}

The most common method of collecting training data is to collect data directly from users of an operator’s service. Data collected from consumers can be acquired in different ways: (a) web activity, provided when a consumer interacts with the company’s website; (b) consumer surveys and other feedback mechanisms; (c) mobile user data, provided through consumer interaction with a company app; and (d) social media.\textsuperscript{46} In order to obtain necessary rights to consumer data, the operator should include a license in its governing user agreement (e.g., the consumer terms and conditions of use) and accurately disclose the data collection and use in its privacy policy. We discuss obtaining rights to service user data in more detail in Part V.

\textbf{C. Batch Uploaded Data From Software Installed On-Premises for Customers}

For customers not connected to the operator’s network automatically (i.e., customers that do not use a hosted or software-as-a-service product), operators can choose to negotiate the right to receive a bulk package of use data through a manual upload or other transfer mechanism. This type of data collection most often occurs where the operator’s product is installed on-premise, which may be due to: (a) industry privacy sensitivity, for example, in the medical and financial sectors; (b) consumer desire for customized

\textsuperscript{44} See, \textit{e.g.}, INFOBASE, https://www.acxiom.com/what-we-do/infobase (providing a large user database with numerous information points gathered, over time, in response to different requests).

\textsuperscript{45} Mirani, \textit{supra} note 42.

solutions;\textsuperscript{47} or (c) the nature of the product lends itself better to on-site installation.\textsuperscript{48} On-premise software can involve a negotiated paper agreement (instead of a shrink-wrap or click-through agreement), so companies need to be careful that the necessary data rights are not negotiated out of the agreement.

\textbf{D. Open Source Public Data Sets}

Finally, academic institutions, individual researchers, and ‘open-source advocates’\textsuperscript{49} have created pre-populated data sets for common machine-learning algorithm problems. For example, the University of California at Irvine currently maintains 413 data sets that are open to the public for use in machine learning algorithms.\textsuperscript{50} Generally, the rights to these data sets are less restrictive than one would find in a negotiated bilateral agreement, as open source licenses tend to be permissive by nature. However, operators should still evaluate the applicable data license terms to be aware of any requirements to contribute developed technology back to the open source community, and other requirements of the license (e.g., to provide attribution). Descriptions of most common open source licenses are maintained by the Open Source Initiative.\textsuperscript{51}

\textbf{V. LAWS/LEGAL RISKS AROUND USE OF DATA/PII IN MACHINE LEARNING}

The legal risks of using data generally depend on the following


\textsuperscript{48} See HOST ANALYTICS, https://hostanalytics.com/blog/on-premises-versus-cloud-based-epm-software-which-is-right-for-your-business/.

\textsuperscript{49} Open source advocates are generally thought of as zealous individuals, who believe that as much of the internet and developing software as possible should be made open to the public. See, e.g., CBSNEWS, \textit{Oracle names Open-Source Evangelist}, CNET (Sept. 7, 2005), https://www.cnet.com/news/oracle-names-open-source-evangelist/.


factors: (a) the relative sensitivity of the data; (b) the types of predictions to be produced; (c) the agreement governing the acquisition and use of the data; and (d) the impact on a broader industry or market.

A. Use of Sensitive Data

The legal risk associated with a machine learning algorithm is determined, at least in part, by the sensitivity of the source data. In other words, if regulated data is an input, then the output is also likely to be regulated (or considered sensitive data of the same category). Sensitive data is more often regulated, and penalties for non-compliance with regulatory schemes for sensitive (e.g., personally identifiable) data often carries harsher penalties.\(^{52}\) In addition, data providers (like business-to-business operators or data brokers) may be more hesitant to agree to provide sensitive data that is subject to extensive regulations, due to their fear of being held accountable for misuse by a third party of data they originally collected.

The primary categories of what we often consider sensitive data are not surprising: (a) health data; (b) financial data; (c) educational data; (d) location data; (e) visual data (photos of a consumer); and (f) data regarding children. Importantly, if an operator seeks to use sensitive data to make predictions within the given industry, the operator will fall under the purview of industry regulators.\(^{53}\) For example, if educational data is used to predict educational outcomes for students, or financial data is used to determine credit-worthiness, the resulting predictions would likely be subject to similar regulatory schema.

In addition, operators may be required to handle data in a


\(^{53}\) For example, HIPAA will apply to data clearinghouses, processors, and clearinghouses, as well as business associates which will include most health-software providers See Are You a Covered Entity?, Centers for Medicare and Medicaid Services, https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/HIPAA-ACA/AreYouaCoveredEntity.html (last visited May 10, 2018).
specific way, or even store data for longer periods of time, based on the sensitivity of the industry. For example, in the health context, the Health Insurance Portability and Accountability Act requires that certain health-related data (but not all) be retained for at least six years.\textsuperscript{54} Particular categories of health providers are subject to additional retention requirements. For example, Medicare managed care providers must retain records for at least ten years.\textsuperscript{55} While the operator itself may not be a managed care provider, it may be a subcontractor to one who is required to be bound by the same retention policies. In those cases, it is common for the “covered entity” (i.e., the entity bound by the law) to contractually “pass through” certain data retention requirements under HIPAA to all of its subcontractors.

\textbf{B. The Output Use Case}

Certain machine learning outputs may create undue legal risk, even if the data is collected in compliance with any applicable laws. For example, an operator’s use of data to predict a consumer’s credit-worthiness will result in a company being classified as a “Credit Reporting Agency.”\textsuperscript{56} Credit reporting agencies are subject to burdensome regulations.\textsuperscript{57} As another example, the use of data in a device to predict health outcomes can lead to a product or service being classified as a medical device, which is subject to regulation by the Food and Drug Administration, including things like fitness


\textsuperscript{55} 42 C.F.R. § 422.504(d)(2)(iii) (2011).


trackers and massage chairs.\textsuperscript{58} As discussed in Part V.A., detection of legal wrongdoing in these cases often does not require analyzing the actual data use, and can be determined solely from the resulting product.

\textit{C. Breach of Contract/License}

One of the larger areas of legal risk for operators using data in machine learning algorithms is the risk of non-compliance with the agreements under which data rights are obtained. If a company relies on a small number of customers for the majority of its revenue, just one dispute can have an enormous impact on the company, especially if the details of the alleged misuse are made public. Such an allegation, even if unfounded, could harm the company’s ability to attract future customers. For example, the unauthorized use of a customer’s data could be considered a breach of confidentiality (if the data is identified as being subject to confidentiality terms), intellectual property infringement (to the extent any intellectual property rights are embodied in the data), or misappropriation of trade secrets (depending on how the data is misused), which could result in breach of contract claims, claims in tort, or statutory damages for copyright infringement.

Additionally, it is critical that operators relying on a few large enterprise customers use that data correctly (i.e., consistent with the data use rights in the customer license agreement). The loss of one large customer could destroy the viability of the algorithm.

It is important to keep in mind, however, that private actions (e.g., between two private parties) to enforce violations of data use terms are limited by the customer’s ability to detect the operator’s wrongdoing. It is often difficult or impossible for a customer to know, or to prove, that a company uses individual data in machine learning algorithmic analyses. To address this information imbalance, new methods of detecting illegal collection and use of data have evolved over the last few years. For example, to uncover

\textsuperscript{58} Given the rise of internet of things, new ways to deal with these devices/requirements are being explored. \textit{See FDA Selects Participants for New Digital Health Software Precertification Pilot Program, FOOD AND DRUG ADMINISTRATION (September 26, 2017),} https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm577480.htm.
Bing’s practice of copying data and functionality, Google inserted false hits in their search engine functionality and monitored Bing to see if the false stories or incorrect results also appeared in Bing’s results in the same order. Additionally, parties more frequently negotiate contractual auditing rights to allow searching for wrongful use of data directly in the service provider’s files.59

D. Impact on the Larger Market/Industry

Finally, because widely-adapted machine learning algorithms are a relatively recent technological development, novel regulations and industry controls are being created in an attempt to police new concerns as they arise. Outside of the United States, the Australian government is looking into whether machine learning should be considered anti-competitive in particular use cases because it can create the ability to more easily base pricing off of a competitor and allow parties without any actual direct communication to participate in a tacit price fixing scheme.60

VI. WHAT NEEDS TO BE CONSIDERED WHEN DRAFTING AN AGREEMENT FOR A MACHINE LEARNING SERVICE

Different operators will rely on different license terms to obtain data depending on the proposed data use. First, an operator must determine whether it is interested in the rights to the results output, or just improvements to the algorithm. Second, the operator must determine if it is attempting to buy data or simply collect data through a service it is already offering. Third, the operator must visualize the desired machine learning output. The actual output will often dictate the terms of the license required to offer the machine learning service.


A. Predictions Versus Algorithm Improvements

Not all machine learning operators have the same level of interest in using the results of an algorithm in future work. Some operators are intimately interested in the accuracy of the result, but not the result itself. For example, a marketing platform that predicts whether an individual will click on an image with particular attributes will not care about whether the consumer goes on to buy the linked product. Instead, it cares only about which attributes the image contains and whether the attributes had the predicted effect (i.e., caused the consumer to click the link). The relevant data are image attributes and the user’s “clicks,” rather than the customer’s content. In contrast, a medical imagery predictive algorithm would want to know if its software successfully or unsuccessfully predicted the presence of a medical condition, and all of the specific outcomes that were or were not correctly predicted. As a result, that operator would need a license to obtain more specific data about each diagnosis.

B. Source of Data

As discussed in Part IV, some consumer-facing companies offer data-gathering services and data can also be obtained through wholesale acquisitions of databases. Data gathered through negotiated agreements with customers can vary depending on: (a) whether the company is business-to-business (“B2B”) or business to consumer (a business providing a service to an individual consumer) (“B2C”); (b) industry norms and data sensitivity; and (c) customization of the product and algorithm. Operators should be cognizant of the different rights negotiated with each customer, and maintain minimum acceptable terms to avoid violation of customer agreements. By contrast, purchased data generally has fewer limitations which may only restrict the purchaser from specific high-risk activities, like predicting credit-worthiness or re-identifying

---

individuals.\textsuperscript{62}

\textbf{C. Output}

Finally, both public perception and potential legal consequences of machine learning data use are dependent on the final output of the algorithm. Consider the medical industry. Given the public interest in improving and refining medical care, consumers may be more likely to allow companies to use their data to develop software that will diagnose a specific ailment based on individual attributes. The customers themselves have a stake in the result and thus less resistant to sharing their data. However, information about personal health is highly sensitive. Consumers may be willing to allow the use of their data, but only if it is anonymized. An operator should be aware that in some cases, it is far more likely to get the data sets it needs if it promises to protect the consumer’s identity.\textsuperscript{63}

\textbf{D. Recommendations for Drafting}

When drafting an agreement to acquire data for use in a machine learning algorithm, there are several aspects of the license one should consider. This Section discusses a number of considerations for data licenses, including: (1) license duration; (2) ownership of created output; (3) requirement for data to be provided in a de-identified/non-sensitive format; (4) combining data with other data sets; and (5) promises that data is gathered in accordance with applicable law.

\textsuperscript{62} As an example, Acxiom states that data sets from their site: “contain information on individuals and households in the U.S. and are developed from many sources, including public records, publicly available information, and data from other information providers. Acxiom’s marketing products are used by qualified companies, non-profit organizations and political organizations in their marketing, fundraising, customer service and constituent service and outreach programs to provide customers and prospects with better service, improved offerings and special promotions.” \textit{Highlights for US Products Privacy Policy, ACXIOM.COM}, \url{https://www.acxiom.com/about-us/privacy/highlights-for-us-products-privacy-policy/} (last visited Apr. 20, 2018).

\textsuperscript{63} These promises could, of course, expose the operator to significant legal risk if they are broken.
1. License Duration

A data license should not be time-limited. This is particularly important if the algorithm makes continuing reference to source data. If the license itself cannot be perpetual, then the operator should retain perpetual rights to any improvements or derivative works of the data so that the effectiveness of the algorithm is not diminished.

If an operator must agree to a time-limited license that requires the return of data, then it should be aware how difficult it can be to identify exactly which machine learning result is attributable to a specific data set or individual piece of data. The model should improve and evolve with each new data set added. Therefore, the ideal data license will be perpetual, notwithstanding termination of the underlying agreement.

Additionally, an operator must be aware that a large enterprise customer could insist that a data license be revocable in the event of an operator’s breach of the underlying agreement. If the license were revoked, the operator would likely be required to return all data. As discussed, that can be an incredibly cumbersome task to undertake. As a result, it is critical for the operator to ensure compliance with its data license agreements to avoid a license revocation that compromises the algorithm. Concerns about time limitations in a license are less of an issue with data licensed from data brokers, as data brokers often grant perpetual licenses.

2. Ownership of Created Output

Ownership of the output of a machine learning algorithm is another important consideration. Enterprise customers, particularly those with negotiating leverage, will often attempt to claim that any technology, intellectual property, or other output developed by referencing their original data belongs to them. That approach is reasonable in a consulting arrangement with a defined project scope, but not necessarily in the machine learning context, where the operator continuously uses its customers’ data to offer an improved product to every current and future customer.

Therefore, it is critical that the operator maintains ownership of its algorithm, as well as the improvements to the algorithm.
generated based on its customers’ data in order to protect the operator’s key intellectual property. As a fallback position, the operator could attempt to transfer ownership of any custom developed features for the specific client or consumer-data reliant improvement if: (a) that improvement or model alone is unusable by the customer in any context other than the operator’s algorithm; and (b) the operator is granted a perpetual, unlimited, royalty-free, sublicensable license to the developed model or improvements for use in its products and services.

3. Requirement for Data to be Provided in a De-Identified/Non-Sensitive Format

Machine learning operators often do not want to assume the risk of hosting a platform which produces predictions that could inadvertently reveal an individual’s personally identifiable information (“PII”). If the operator gathers data from customers, it must ensure that customers strip their data of any PII or otherwise take on the risk of removing PII. Some enterprise customers, on the other hand, may refuse to provide any PII and will agree to represent that no PII is included in their data sets. Data brokers may also agree to similar terms, or undertake removal themselves. In any event, the customer’s privacy policy (if it is required to have one) should ensure that the customer has the right to provide the data to the operator. The operator can then ask the customer to represent and confirm that all data is provided in compliance with the privacy policy.

4. No Prohibition on Combining Data With Other Data Sets

Machine learning algorithms, by their nature, improve with exposure to more and more data, regardless of the source. If data is collected in bulk from an external source, any prohibition on commingling that data with data from other sources undermines the usefulness of that data set. This issue often arises when purchasing data from data brokers, who may have negotiated no commingling provisions with their providers that are passed on to purchasers of the data. An operator could address this issue in its agreement with a data broker by agreeing that there will be no commingling that
results in the identification of individuals or that connects PII to an anonymized/de-identified data set.

Obtaining the rights to combine data sets can be especially important since demonstrating compliance with a contractual requirement to keep data sets separate can be nearly impossible. Certain aspects of data may be present in multiple data sets, and machine learning output may be reliant on multiple data sets, so showing that particular data came from one source and not another is not feasible.

5. Representation That Data was Gathered in Accordance With Applicable Law

Finally, when obtaining data from an external data source, a machine learning operator will have little control over how the data was originally gathered, and very little insight as to whether the collection complied with applicable law. As such, the operator must rely on the representations and warranties of its data providers as to the legality of the data, and should ensure that the applicable representations and warranties are in the underlying data agreement. The operator should insist on these representations and warranties and refuse to deal with any provider that will not agree to them.

CONCLUSION

While the concept of machine learning is not new, the ubiquity of machine learning applications has seen a significant upswing over the past five to ten years. In the legal sector, drafting appropriate license language and associated data use rights for machine learning applications requires lawyers to understand what exactly machine learning is and how it differs from traditional software licensing or service provider scenarios. The most important point to take into consideration when drafting a machine learning license is that all data use is not created equal. How data is gathered, processed, and stored will depend on the type of machine learning model and the goals of the organization using the data. Therefore, to appropriately draft a license, attorneys should examine the data cycle with their client to understand how data will be gathered, processed, stored, and retained. The specifics of the data type, use, processing and
storage will affect a multitude of legal and contractual issues relevant to the data use license itself, including, but not limited to, breadth of license, data use timeframe, and handling of derivatives. Attorneys should also take into consideration sensitivity of data use, collection and retention within a given industry, as well as factors such as consumer perception and the machine learning algorithms’ output to help them better advise clients on the “real-world” risks of using different types of data in their business.

**Practice Pointers**

- **License duration** *(term of the agreement versus perpetual)*: Understand how long the company needs to refer back to the data (including whether data will be needed for fixing later-discovered flawed outcomes) and whether the data can be separated from the algorithm without affecting functionality.

- **Ownership of created output** *(customer-owned or company-owned)*: Understand whether output is customer specific or increases the value of the algorithm as a whole, and whether the algorithm using training data continues to process improvements from both old and company-created data inputs.

- **Data Identifiability** *(anonymous versus individual characteristics)*: Understand which data is likely to be used as a predictor, and whether anonymization of data would affect the ability to create valuable output. Additionally, consider the federal and state statutes applicable to the type of data processed by the algorithm (e.g., HIPPA for health-related data).

- **Data Set Combination** *(allowed or prohibited)*: Understand whether data-set combination is likely to re-identify personally identifiable information regarding individual data subjects, and which attributes of a data set need to be correlated with to produce valuable output.

- **Responsibility for gathering data in compliance with law** *(company versus outside data source)*: If data is gathered in bulk from an outside source (including from a data broker, a
white-labeled incorporation of the algorithm, or an open source set), the outside party should bear primary responsibility for gathering the data in compliance with law. For data gathered directly from a customer, the company will likely bear primary responsibility for informing the consumer and obtaining consumer consent. For data gathered from the internet (via web scraping or other similar techniques) without the express consent of the data source, the attorney should analyze whether such data collection (1) violates law, or (2) violates online terms of service agreements, and the attorney and company should together conduct a risk-benefit analysis of such data collection.
FAIR USE, FAIR PLAY: VIDEO GAME PERFORMANCES AND “LET’S PLAYS” AS TRANSFORMATIVE USE

Dan Hagen*

© Dan Hagen

Cite as: 13 Wash. J.L. Tech. & Arts 245 (2018)

http://digital.law.washington.edu/dspace-law/handle/1773.1/1816

ABSTRACT

With the advent of social video upload sites like YouTube, what constitutes fair use has become a hotly debated and often litigated subject. Major content rights holders in the movie and music industry assert ownership rights of content on video upload platforms, and the application of the fair use doctrine to such content is largely unclear. Amid these disputes over what constitutes fair use, new genres of digital content have arrived in the form of “Let’s Play” videos and other related media. In particular, “Let’s Plays”—videos in which prominent gamers play video games for the entertainment of others—are big business in the streaming and video upload world. Many video game producers vigorously assert the right to prevent the publishing of Let’s Play videos or to demand a cut of the revenues. This article discusses who legally possesses the right to distribute or profit from Let’s Play content under current law, and the way that courts ought to approach these disputes consistent with the principles of copyright protection. I conclude that the nature of video game content produces conceptual challenges not necessarily present in movies and music, and that these differences have a bearing on fair use analysis as it applies to Let’s Play videos.

* Dan Hagen is a J.D. candidate at the University of Washington, class of 2018. Thank you to Professor Bob Gomulkiewicz for his helpful edits, suggestions, and guidance in developing this article. Thanks also to Professor Zahr Said for her detailed review of an early draft of this article.
TABLE OF CONTENTS

Introduction .................................................................................................................. 246
I. Background and Gameplay Overview ................................................................. 248
   A. Overview of Gameplay Types ........................................................................... 250
      1. Long Plays and Walk Throughs ................................................................. 250
      2. Speedruns ................................................................................................ 251
      3. Conventional Let’s Plays .......................................................................... 252
      4. E-Sports .................................................................................................. 253
   B. Video Game Interactivity ................................................................................ 254
II. Copyright Law and Enforcement ........................................................................ 255
   A. Fair Use and the DMCA ................................................................................ 255
   B. Videogame Producer Actions Against Let’s Plays ........................................ 257
III. Video Games and Copyright ............................................................................. 258
   A. The Nature of Video Game Content ............................................................... 258
   B. Video Game Output and Recordings ............................................................. 259
   C. Video games as public performance ............................................................. 262
IV. The Principles of Fair Use Applied to Let’s Plays ............................................ 265
   1. Purpose and Character of the Use ................................................................. 265
   2. The Nature of the Copyrighted Work ............................................................ 269
   3. Amount and Substantiality of the Portion Used in Relation to the Whole .... 270
   4. Effect on the Potential Market for or Value of the Copyrighted Work ........ 271
Conclusion .................................................................................................................... 273
Practice Pointers ......................................................................................................... 273

INTRODUCTION

If someone makes a recording of themselves playing a video game, who owns the resulting content and what may they legally do with it? The answer is not as straightforward as some video game producers presume. Video games are different from movies and music in significant ways, and the limits of copyright protection in the context of interactive media have yet to be comprehensively judicially tested. However, some case law provides insight into how courts will, or ought to, approach the repurposing of video game content consistent with the principles of copyright law.

“Let’s Play” videos are a relatively new genre of media, and the
application of copyright protection to these videos presents potential challenges. A “Let’s Play” is a recording of gameplay footage made for the benefit of an audience. These videos are often streamed live over the internet or recorded and uploaded to social media sites like YouTube. The Let’s Play content creator often provides running commentary, usually related to the game being played. Let’s Play videos can be broken into sub-categories depending on the player’s purpose. For example, a “speedrun” video is a type of Let’s Play in which a player attempts to finish a game as fast as possible. Other types of videos may involve or focus on competitive demonstrations of skill against multiplayer opponents, humorous reactions or mockery of video game content, or socializing between the player and the viewers. Though the lines are not defined with perfect clarity, a Let’s Play video is typically understood as a recording of a gameplay performance, in contrast to a gameplay “stream” which is a live transmission of a that performance as it is happening.

Over the last couple of years, Let’s Play videos have grown into a billion dollar per year industry. Streamers and Let’s Play content creators receive money through donations during their live-streams and ad revenue from videos watched after the fact by hundreds of millions of consumers. People watch Lets Plays for a variety of reasons, including entertainment, information as to whether a game is worth purchasing, and tips for progressing or improving their own gameplay experience.

1 See What is a Let’s Play on YouTube?, MEDIAKIX (Feb. 3, 2016), http://mediakix.com/2016/02/what-is-a-youtube-lets-play-video/#gs.kNvLoUE.
2 Id.
3 See infra Section I.A.
6 See Leo Mirani, 500 Million People are Watching Videos of Video Games QUARTZ (Jul. 9, 2015), https://qz.com/449161/500-million-people-are-watching-videos-of-video-games/.
7 See Jake Muncy, Why I Watch People Play Videogames on the Internet, WIRED (Aug. 21, 2016), https://www.wired.com/2016/08/why-i-watch-lets-
Some video game copyright holders claim to own some, or all of the rights to the footage produced by a Let’s Player, the same way the rights-holder of a movie or song would if someone else made a copy or a derivative work and published it. This conception of video game copyright protection relies on an understanding of video game content as being equivalent to that of a movie or a song. However, the strength of that position will depend on the level of copyright protection afforded to video games. Furthermore, the legitimacy of a video game copyright holder’s assertion of copyright in a Let’s Play context will depend on whether a Let’s Play is sufficiently “transformative” to qualify as fair use.

There are several elements that may complicate a legal analysis of a Let’s Play recording or performance, such as in-game music and extended non-interactive cut-scenes. In addition, the level of interactivity in the game as well as the purpose and function of the recording or performance may lead to different conclusions about whether a specific Let’s Play constitutes fair use. The ways in which courts choose to conceptualize video game and Let’s Play content will undoubtedly affect the analysis.

I. BACKGROUND AND GAMEPLAY OVERVIEW

The precise origin of Let’s Play videos is unclear. People have likely been recording themselves playing video games since video games were introduced to the public. However, mass public consumption of such secondary media is a relatively new phenomenon. The term “Let’s Play” was probably first coined in the Something Awful forums in the year 2003, but applied to still images with text, and bore little resemblance to what we think of today as a Let’s Play. Video game review sites relied on captured plays; see also Mijntje Boon, Let’s Plays: Why are They so Popular?, CREDO MAGAZINE (Jun. 29, 2016), http://www.credomagazine.nl/lets-play/.

8 See e.g., Chelsea Stark, Nintendo Will Get Revenue From All YouTube Videos Featuring its IP, MASHABLE (May 17, 2013), http://mashable.com/2013/05/17/nintendo-youtube/#GwDavkEC_PqX.


10 Patrick Klepek, Who Invented Let’s Play Videos?, KOTAKU, (May. 6,
video as early as 2001, but the use of video recordings of gameplay footage by amateurs did not become popular until around 2007 when YouTube became more popular.\(^{11}\) It was around this time that people began posting gameplay clips for entertainment. For example, the “Angry Video Game Nerd” became one of the first YouTube celebrities for videos in which he played particularly frustrating or mediocre games and then commented on them for humorous effect.\(^{12}\) Others like “JonTron” are cited as being among the first to popularize Let’s Plays in the form they exist in today.\(^{13}\)

When it became clear that people enjoyed watching others play video games as much as they enjoyed playing them, many more YouTube channels dedicated to Let’s Play videos sprang up.\(^{14}\) Today, the most popular Let’s Play content creators produce Let’s Plays as a career. The YouTube user PewDiePie has by far the most popular channel on YouTube, boasting over sixty-two million subscribers, and the channel is primarily dedicated to Let’s Plays.\(^{15}\) Based on ad revenue his channel brings in, PewDiePie’s yearly income is estimated at between $4 and $7 million dollars.\(^{16}\) Many other prominent YouTube content creators, such as “Angry Joe” and “TotalBiscuit” have millions of subscribers.\(^{17}\) The popularity of Let’s Play videos eventually led to the production of a website in 2011 called Twitch.tv dedicated to live streaming of gameplay videos.

\(^{11}\) Id.


\(^{13}\) Jon Jafari, JonTronShow, YOUTUBE, https://www.youtube.com/user/JonTronShow (last visited May 10, 2018).

\(^{14}\) See supra note 10.


\(^{17}\) Joe Vargas, The Angry Joe Show, YOUTUBE (last visited May 10, 2018) https://www.youtube.com/user/AngryJoeShow; John Bain, TotalBiscuit, The Cynical Brit, YOUTUBE (last visited May 10, 2018) https://www.youtube.com/user/TotalHalibut (immediately prior to publication of this article, John Bain passed away).
footage.\textsuperscript{18} Many YouTube Let’s Players maintain both YouTube and Twitch accounts, using Twitch for the livestream, and later uploading the recording to YouTube.\textsuperscript{19} The livestream is a more interactive affair for the audience, as Twitch chat allows the audience to chime in and make suggestions, criticize, or encourage the player as they play.\textsuperscript{20} Twitch also permits its users to donate money to their favorite streamers.\textsuperscript{21}

\textit{A. Overview of Gameplay Types}

While a Let’s Play is, broadly speaking, a video of someone’s gameplay experience, there are non-arbitrary ways of differentiating gameplay videos based on function, purpose, and content. For the purpose of legal analysis, understanding the type of Let’s Play at issue helps to determine whether the content may be viewed as “transformative,” and whether the game has strong underlying copyright protection. In addition, the genre of game and the characteristics of its content may also be relevant. As streaming and publication of gameplay footage continues to gain popularity, it is important to understand these differences.

1. Long Plays and Walk Throughs

A “long play,” is a video of a complete gameplay experience from beginning to end.\textsuperscript{22} Their purpose is to capture everything the game has to offer for the purpose of entertainment, preservation, or providing helpful information to potential players interested in completing the game.\textsuperscript{23} Though player input will always be unique,

\textsuperscript{23} Id.
there is typically nothing that occurs in a long play that goes beyond what the game designer intended. Consistent with its purposes, a long play does not generally contain commentary; instead, capturing as pure a gameplay experience as possible.

Video walkthroughs can appear similar to long plays, but are produced for the purpose of helping other players learn how to complete a game. For example, players may consult a video walkthrough when they find themselves unable to clear a particularly difficult section of a game. A walkthrough may include commentary aimed at assisting players, but in most cases the visuals convey the necessary information. While it is possible for people to watch walkthroughs as entertainment, their primary purpose is to provide useful information to players.

2. Speedruns

A “speedrun” is an attempt by the player (the “runner”) to finish a game under various conditions as fast as possible. Speedrunning has a competitive element, as runners attempt to set speed records for whatever category they are running. Speedrunners often stream live, accept donations, and monetize their videos on YouTube. One popular organization of speedrunners, “Awesome Games Done Quick,” conducts bi-annual speedruns streamed live for charity and has raised over $14 million dollars to date.

Unless prohibited by agreement, runners often make use of glitches and other exploits not intended by game designers to improve their clear time. So long as a glitch is exploitable within the game’s code, requiring no outside intervention, it is usually fair game. In addition to witnessing the skill of the runner, seeing players expertly exploit glitches is part of the appeal of watching speedruns,

---

26 *Id.*
28 See supra note 25.
as glitches can produce spectacular and bizarre results.29

Each game will usually have several categories of speedrun agreed upon by the gaming community which a runner may attempt.30 For example, in an “any percent run,” the runner is typically free to use warps, glitches, and sequence breaks to improve their clear time. In a “glitchless” run, the runner is prohibited from making use of such glitches. A 100 percent run might require that all levels be completed, or all items collected without skips.31 At the highest levels, speedruns begin to closely resemble one another as the fastest methods of clearing a game under the various categories become known.32 Recordings and performances of highly optimized speedruns will differ only slightly depending on how well executed the runs are.33 The emphasis on technical execution rather than on individual or artistic input from the players, in addition to frequent monetization, will have implications for fair use analysis.34

3. Conventional Let’s Plays

The most prominent type of Let’s Play video is one in which a gamer emulates the experience of playing a game in front of friends for the benefit of an audience. However, instead of one or two friends on a couch, a Let’s Player may be playing for an audience of thousands or even millions. This type of Let’s Play is subject to wide variability, depending on the player.35 An audience may watch a

29 ZFG, AGDQ 2016 - Legend of Zelda: Ocarina of Time Glitch Exhibition
https://www.youtube.com/watch?v=YrlqoGO2-BE.

30 See Rami Ismail, If Esports are the Sports of Video Games, This is the Parkour, ROLLING STONE (Feb. 5, 2018), https://www.rollingstone.com/glixel/features/rami-ismail-speedrunning-w516376


33 Id.

34 See discussion infra Part IV.

player because they are charismatic, amusing, or skillful. They may watch videos of a game or genre of game because they find it interesting, entertaining, or helpful. A Let’s Player may view themselves as an entertainer, a commentator, or a critic. While the genre is not defined by commentary, Let’s Play videos and streams typically include commentary by the player.

4. E-Sports

E-sports have notably become more prominent and lucrative in the last few years. Recognizing the growing audience and lucrative potential already present in large video game tournaments like EVO, DOTA2 and the League of Legends Championships, corporations like ESPN have begun reporting on and showcasing e-sports. E-Sports videos typically display matches between two or more players in head-to-head competition. The entertainment purpose in displaying such matches is clear, and competition between players is certainly anticipated by game publishers, though unsanctioned public exhibition may not be.

The competitive scene has the potential to come into conflict with claims of copyright infringement in much the same way as conventional Let’s Plays and speedruns. In such cases, the third-party publishing the recording or stream would be the potential primary infringer rather than the players. For example, in 2013 Nintendo asserted its copyright over the game Super Smash Bros. Melee against the EVO Fighting Games Championships, not only in an effort to block the competition organizers from streaming matches, but also to block EVO from using the game in their not-play-video-games.

36 Id.
37 Id.
38 Id.
39 Id.
competition at all.\textsuperscript{42} In the face of public criticism, Nintendo ultimately relented and permitted the competition to move forward.\textsuperscript{43} As above with speedruns, the emphasis on technical execution in the e-sports context may impact a fair use analysis.

\textbf{B. Video Game Interactivity}

The level of interactivity present in any particular video game varies. On the low end of the interactivity scale, there are games which function as interactive stories with minimal player input. Examples include the recent Telltale \textit{Game of Thrones} and \textit{Batman} games.\textsuperscript{44} In these games, players make choices, which determine how a story unfolds, but do not otherwise affect what appears on screen.

On the other end of the spectrum are games that invite players to be creative. Examples are games like \textit{Mario Paint}, \textit{Super Mario Maker}, and \textit{Minecraft}.\textsuperscript{45} These games give the player the tools with which to create and alter their own renderings or environments. As an analogy, the game developer has given the player a palette and tools to produce their own unique works. \textit{Minecraft} is interesting for another reason, in that the worlds in which the player is empowered to build are produced procedurally by a computer, meaning that the designers themselves are unaware of the details of any individual

\begin{itemize}
\end{itemize}

Most games lie somewhere in between these extremes. They can range from simple puzzle-type games in which players manipulate two-dimensional objects on a single screen, like Tetris or Bejeweled, to complex 3-D games involving player avatars and pre-built worlds. Generally, the more complex a game is, the greater the range of potential options available to a gamer. While developers do not anticipate any exact set of inputs, generalized input patterns are anticipated and required for players to progress through the game.

II. COPYRIGHT LAW AND ENFORCEMENT

\textit{A. Fair Use and the DMCA}

The U.S. Copyright statute provides that the publishing of copyrighted materials in certain circumstances determined to be “fair use” is not an infringement of copyright.\footnote{It is a matter of debate whether fair use ought to be treated as an affirmative defense, being an exception to a violation of applicable copyright law, or whether conduct falling under fair use is not a violation of the statute. The view currently expressed by the Supreme Court in dicta, that fair use is an affirmative defense, is not obvious from the text of the statute. Because of this, some circuit courts have departed from the dicta of the Supreme Court. \textit{See, e.g.}, Lenz v. Universal Music Corp., 815 F.3d 1145, 1152 (9th Cir. 2016).} The statute states the following:

The fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

\begin{itemize}
\item (1) the purpose and character of the use, including
whether such use is of a commercial nature or is for nonprofit educational purposes;

(2) the nature of the copyrighted work;

(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

(4) the effect of the use upon the potential market for or value of the copyrighted work. The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.48

The United States Supreme Court holds that all four of these factors must be considered together in light of the purposes of copyright law when determining whether the use of copyrighted material constitutes “fair use.”49 “[A]s we apply copyright law, and the fair use doctrine in particular, we bear in mind its purpose to encourage "creative activity" for the public good.”50

Congress passed the Digital Millennium Copyright Act (“DMCA”) into law in 1998 to update copyright laws such that they adequately cover emerging technology.51 The DMCA permits copyright holders to issue takedown notices to internet websites hosting copyrighted material under certain conditions.52 Among these conditions is that the copyright holder first make a good faith effort to determine whether the content in question is “fair use.” Failure to do so results in a violation of the DMCA.53 It is through the DMCA notice and takedown procedures that companies can assert their copyrights with regard to Let’s Play content on video hosting sites such as YouTube and Twitch.

---

53 See Lenz v. Universal Music Corp., 801 F.3d 1126, 1130 (9th Cir. 2015).
B. Videogame Producer Actions Against Let’s Plays

YouTube makes use of a Content ID matching system to monetize or remove potentially infringing material from its website. The system allows copyright holders to upload visuals or music to a database. When a YouTube user uploads a video, the content is checked against the database; if there is a match, the copyright holder determines what happens next. A content match can result in immediate takedown of the potentially infringing video, monetization in the form of ads on the video for the benefit of the copyright holder, or portions of the video being muted or censored. The process happens automatically, with no requirement that any person actually review the potentially infringing material before sanctions are implemented. Despite the DMCA’s requirement that a good faith effort be made to determine whether potentially infringing material is fair use, the courts have ruled that algorithmic takedown processes are legally permissible.

YouTube content creators whose videos are claimed can undertake a lengthy appeal process to have their video reinstated. However, the process undoubtedly favors the claimant. The appeal is never seen by a third party, but simply goes to the copyright claimant for reconsideration. If the appeal is rejected, the content creator can appeal again; but if the appeal is rejected for a second time, it results in a copyright strike against the creator’s channel.

---

55 Id.
56 Id.
57 Id.
58 See Lenz, 801 F.3d at 1158.
Copyright strikes can have a serious impact on a user’s channel and depending on its popularity, the user’s bottom line.62

Nintendo is one of the most aggressive video game publishers when it comes to asserting its copyright over Let’s Players.63 Nintendo makes use of YouTube’s ContentID system and automatically monetizes the videos of anyone who makes use of their content, resulting in Nintendo receiving all of the advertising proceeds.64 For this reason, many prominent YouTube content creators refrain from posting any videos of Nintendo content.65 In response to public criticism, Nintendo instituted the Nintendo Creator’s Program, which purports to share advertising revenues legally entitled to the copyright holder to those that sign up.66 Taking part in the program is subject to many restrictions, including a prohibition on using any content outside of a specified list of games published by Nintendo.67 Given Nintendo’s assertiveness with regard to its legal position, their corporate policy or one like it is fertile ground for a legal dispute.

III. VIDEO GAMES AND COPYRIGHT

A. The Nature of Video Game Content

Video games incorporate digital assets including artwork, trademarks, software code, music, voice acting, and animated cut-
scenes into a final playable product. Unlike other forms of media, full realization of the value and character of video games is entirely dependent on the unique input of individual players. Such input is anticipated by the game designer and is typically required for the story or progression of the game. Video games could be analogized to board games, in which the pieces are all included, and the input of the player dictates the progression of the game. However, the audiovisual component of video games, as well as the repeatability of in-game sounds and images under certain conditions may provide a basis for copyright protection. In addition to the audiovisual display, the “performance” of a video game could also provide a basis for copyright protection, just as it does with other performative works.

B. Video Game Output and Recordings

Many companies, including Nintendo, currently assert that they own the product of the interaction between the player and the digital assets as though they produced that product. This conclusion is based on the notion that the audiovisual content of a game display, including displays created by players during gameplay, is wholly owned by the game’s copyright holder. Case law suggests that a video game’s audiovisual “fixed” content could in principle be copyright protected. A fixed product is something that takes a final form, such as a recording, and doesn’t change. A Let’s Player, by making a recording of their gameplay experience is undoubtedly “fixing” that content. The copyrightability of such fixed content will, however, still depend on whether the underlying video output or gameplay performance is subject to copyright protection.

For the most part, the audiovisual display in recorded videogame content will almost always be unique due to the input of the player. Every minor decision a player makes contributes to a different audiovisual experience. Even absent any form of commentary or

---


69 See Lewis Galoob Toys, Inc. v. Nintendo of Am., Inc., 964 F.2d 965 (9th Cir. 1992).
alteration of the audiovisual content, a Let’s Play video will likely be totally unique. This makes it different from a copy of a music or song where the copyrighted material is a specific sequence of sounds or images. Nevertheless, the individual elements making up the game’s audiovisual display exist in the game’s code.

The law provides copyright protection for audiovisual content. Section 101 of the Copyright Act defines “audiovisual works” entitled to copyright protection as “works that consist of a series of related images which are intrinsically intended to be shown by the use of machines or devices . . . together with accompanying sounds, if any, regardless of the nature of the material objects, such as films or tapes, in which the works are embodied.”

In *Stern Elecs., Inc. v. Kaufman*, the Second Circuit held that copyrightability extends even to the audiovisual display resulting from interactions between the game code and the player. While acknowledging that a gameplay experience is not fixed in a conventional sense, “[t]he repetitive sequence of a substantial portion of the sights and sounds of the game qualifies for copyright protection as an audiovisual work.”

Exactly how repetitive or substantial audiovisual sequences must be to qualify for copyrightability was not specified.

In *Midway Mfg. Co. v. Artic Int’l, Inc.*, the Seventh Circuit further elaborated on why copyright protection for the audiovisual output of video games is appropriate. The Court first acknowledged two difficulties with attempting to include video games under the definition of audiovisual works: First, that “series of related images” as defined under the Statute, may be interpreted to refer “only to a set of images displayed in a fixed sequence.” Construed this way, videogames would not qualify as audiovisual works because a different sequence of images appears on screen each time the game is played.

The second difficulty identified by the *Midway* court is that the display of the arrangement of the digital assets stored within a game’s code is in the control of the player:

---

71 Stern Elecs., Inc. v. Kaufman, 669 F.2d 852, 856 (2nd Cir. 1982).
73 *Id.* at 1011.
[T]he person can vary the order in which the stored images appear on the screen by moving the machine’s control lever. That makes playing video games a little like arranging words in a dictionary into sentences or paints on a palette into a painting. The question is whether the creative effort in playing a video game is enough like writing or painting to make each performance of a video game the work of the player and not the game's inventor.74

The court ultimately concluded that, despite these difficulties, video game content is copyrightable.75 The court reasoned that video game output was more akin to that of a television viewer pressing buttons on a remote control than it is like creative output.76 Furthermore, the control that a player exercises in playing a game is somewhat illusory, as it is still dictated by the game’s code. “He is unlike the writer or a painter because the video game in effect writes the sentences and paints the painting for him; he merely chooses one of the sentences stored in its memory, one of the paintings stored in its collection.”77

These cases were decided when video games were still in their infancy. Along with technological advancements, the range of options available to a player in most games today is much greater than what existed in 1983. Thus, it is harder to argue that every specific combination of gameplay choices was necessarily anticipated by the creator, or that it is limited by the game’s code in a legally relevant way. Furthermore, games in which assets are provided to the player for the purpose of creative activity, as well as games involving procedurally generated worlds displaying content that cannot have been anticipated by the game designer, cut against the reasoning underlying these holdings.

Regardless of where and how the line is to be drawn, video game developers are likely entitled to some degree of protection over the audiovisual content resulting from gameplay, and therefore have a copyright interest in the repurposing of that content. The primary

74 Id.
75 Id. at 1011–12.
76 Id. at 1012.
77 Id. at 1013.
legal battleground is therefore to be fought over whether Let’s Play output, videos, and performances qualify as fair use of that copyrighted material.

C. Video games as public performance

In addition to protecting copyright holders of “audiovisual works” from the repurposing or display of their content, the Federal Copyright Statute also grants the exclusive right to public performances of copyrighted material.\textsuperscript{78} In the context of a protected work, the statute defines “perform” as “to recite, render, play, dance or act it, either directly or by means of any device or process or, in the case of a motion picture or other audiovisual work, to show its images in any sequence or to make the sounds accompanying it audible.”\textsuperscript{79} The statute goes on to state that the definition of “public performance” includes “[the transmission of] a performance or display of the work to a place specified . . . or to the public, by means of a device or process, whether the members of the public capable of receiving the performance or display received it in the same place or in separate places and at the same time or at different times.”\textsuperscript{80}

The application of public performance copyright to the “performance” of video games is not immediately clear. Nevertheless, the use of the term “play” in the definition of “performance” could mean that the statute reaches the playing of games. However, in \textit{Allen v. Academic Games League of Am.}, the Ninth Circuit held that the word “play” as referenced in the statute, “has generally been limited to instances of playing music or records.”\textsuperscript{81} \textit{Allen} involved the assertion of copyright by a board game manufacturer over the public playing of their games at non-profit academic tournaments.\textsuperscript{82} The court declined to extend copyright protection to the public performances of board games.\textsuperscript{83}

\textsuperscript{80} 17 U.S.C. § 101(2) (2012)
\textsuperscript{81} Allen v. Academic Games League of Am. Inc., 89 F.3d 614, 616 (9th Cir. 1996).
\textsuperscript{82} \textit{Id.} at 615.
\textsuperscript{83} \textit{Id.} at 616.
holding that “[t]o do so would mean interpreting the Copyright Act in a manner that would allow the owner of a copyright in a game to control when and where purchasers of games may play the games and this court will not place such an undue restraint on consumers.” The court went on to opine that whether in public or in private, “games are meant to be played,” suggesting that the fundamental nature and purpose of games was relevant to their determination.

The Allen court noted that even if the playing of a game could be classified as a public performance under copyright law, the “performance” of the games by tournament organizers would constitute fair use. This conclusion was based on the non-profit status of the tournament, and the fact that the tournament likely had a positive impact on the market for the games, rather than a negative one.

While on its surface the Allen decision would appear to apply to video games in a straightforward manner, the court in Allen cited to a case out of the Fourth Circuit reaching a contrary conclusion in the case of coin-operated arcade games. In Red Baron-Franklin Park, Inc. v. Taito Corp., the court found a video game’s status as an audio-visual work involving a “sequence of images” to be the primary determining factor bringing the playing of a video game under the definition of “performance” under the Copyright Statute.

[T]he exact order of images will vary somewhat each time a video game is played depending on the skill of the player, but there will always be a sequence of images . . . [w]e therefore conclude that the operation of a video game constitutes a performance as that term is defined in § 101.

Since Allen did not involve video games, the court did not directly engage with the justification made by the court in Red Baron, except

---

84 Id.
85 Id.
86 See id. at 617.
87 Id.
88 Id. at 616.
to note the case as contrary authority. The sweeping language in *Allen*, however, appears to apply to video games as well as to board games.

Despite the clear conflict between these two cases, at least one district court in the Ninth Circuit sought to resolve the tension by essentially rejecting the reasoning used in *Allen*. In *Valve Corp. v. Sierra Entm’t Inc.*, the District Court for the Western District of Washington reinterpreted the *Allen* conclusion, finding that it was not in fact inconsistent with *Red Baron*.  

90 The court ruled that “[Allen] held that whether the performance is fee-based is an important factor in determining whether the performance is public.” 91 Of course, *Allen* held no such thing. The focus on the tournament’s non-profit status in *Allen* was relevant only to the question of whether the “performance” was fair use. Further, what was at issue in *Allen* was whether the playing of a game constituted a “performance” at all, not whether it was public. Therefore, the *Valve Corp.* court’s interpretation of the *Allen* holding is wrong twice. *Valve Corp.* also implies that *Allen*’s citation of *Red Baron* suggests agreement, neglecting to mention that it was cited as contrary authority. 92

In the wake of *Red Baron*, Congress amended Section 109 to specifically permit “public performances” on arcade machines. 93 While essentially overturning the outcome of *Red Baron*, it did so without contradicting *Red Baron*’s conclusion that the playing of video games in public constitutes a public performance. Both *Red Baron* and *Allen* provide a plausible basis for argument on either side of the video game performance copyrightability divide. The courts will have to decide whether audiovisual content is a significant enough distinguishing factor to overcome the video game’s status as a game, the purpose of which is to be played, as the most important variable in deciding whether the copyright statute protects against public gameplay.

---

90 See *Valve Corp. v. Sierra Entm’t Inc.*, 431 F. Supp. 2d 1091, 1097 (W.D. Wash. 2004).
91 Id.
92 Id.
IV. THE PRINCIPLES OF FAIR USE APPLIED TO LET’S PLAYS

One reason that Let’s Play videos represent an extreme test of copyright principles is that, assuming copyright applies to the product of video game output, many of the fair use factors are pushed to their limit. First, fair use above all requires that its purpose be one “such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research.” This is not necessarily an exhaustive list, nor does it preclude a content creator from producing content for the purpose of monetary gain. Nevertheless, many Let’s Plays would fall under one or more of these purposes. Most Let’s Plays involve commentary of some kind, and many can also be said to be for purposes of teaching and criticism. Let’s Play consumers will commonly cite all three of these purposes as a reason for watching, in addition to entertainment value.

1. Purpose and Character of the Use

Whether a work containing copyrighted material is being produced for monetary gain is a relevant but not necessarily dispositive question in determining whether a work qualifies as fair use. In the case of Let’s Plays, many prominent YouTube content producers publish their content as a career. Whether that content is monetized on YouTube through the running of ads, or by donations through Twitch or Patreon, Let’s Plays can be a lucrative business. Nevertheless, a Let’s Play need not necessarily be produced for monetary gain. There are many examples of Let’s Play videos that are not monetized. It is important to note that companies such as Nintendo will make copyright claims against YouTube content producers who publish Let’s Plays using their games regardless of whether that content is monetized or not.

Although the use of copyrighted material for monetary purpose

---

95 See Muncy, supra note 7; see also Boon, supra note 7.
96 See Oria Madden, Nintendo Claiming Ad Revenue on YouTube User-Generated Gameplay Videos, NINTENDOLIFE (May 16, 2013), www.nintendolife.com/news/2013/05/nintendo_claiming_ad_revenue_on_youtube_user_generated_gameplay_videos.
invites heightened scrutiny from the court, such use is not dispositive in determining whether “purpose and character” weighs against the potential infringer.

The commercial nature of the use does not by itself . . . determine whether the purpose and character of the use weigh for or against finding fair use. We look as well to any difference in character and purpose between the new use and the original. We consider whether the copy is “transformative” of the work it copied because it altered the first with new expression, meaning, or message.97

As noted by the Allen court, the purpose of a game is to be played. A Let’s Play however, may have several different purposes, such as entertainment, education, ridicule, or criticism. The context in which Let’s Plays are consumed indicates an altered purpose from the original to the new use.

The reference to “transformative” use, as quoted by the Third Circuit above, comes from the Supreme Court in Campbell v. Acuff-Rose Music Inc. In that case, the Court spoke of the transformative character of a use as having a bearing on the manner in which all four fair use factors are to be applied.98 The Court held that “the nature of parody,” which was at issue in that case, required the fair use factors be weighed with the parodic character of the use taken into consideration.99

[T]he goal of copyright . . . is generally furthered by the creation of transformative works. Such works thus lie at the heart of the fair use doctrine’s guarantee of breathing space within the confines of copyright[. . .] [T]he more transformative the work, the less will be the significance of other factors . . .

98 Id.
99 Id. at 588.
that may weigh against a finding of fair use.\textsuperscript{100}

Circuit courts disagree over what precisely qualifies a use as transformative and the degree to which the transformative use weighs against the other statutory factors.\textsuperscript{101} One side of the split takes a broad view of Campbell’s “transformative” consideration, finding that a use of copyrighted material need only be for a new purpose distinct from the original to qualify as transformative. For example, in \textit{Authors Guild v. Google, Inc.}, the Second Circuit held that Google’s commercial activity of making digital copies of copywritten books without the authors’ consent was “highly transformative” because the new purpose for which the copyrighted material was being used was to “enabl[e] a search for identification of books containing a term of interest.”\textsuperscript{102} The Supreme Court recently declined to take up the matter on appeal.\textsuperscript{103} Cases coming out of the Fourth and Ninth Circuits take a similarly expansive view.\textsuperscript{104}

The narrower application of \textit{Campbell} takes the view that for a use to be transformative it must add some new meaning or expression to the original. For example, the Third Circuit held as such in \textit{Video Pipeline v. Buena Vista Home Entertainment}, stating that “no added creative activity reveals a dearth of transformative character.”\textsuperscript{105} In that case, a company had compiled two-minute preview clips of copyrighted films and made them available online. The company argued that the video clips were not being displayed for aesthetic or entertainment purposes, as was the intent of the source videos, but rather, for providing consumers with information about the films.\textsuperscript{106} The court held that the “absence of creative ingenuity” meant the clips lacked any “significant transformative


\textsuperscript{101} See Keinitz v. Sconnie Nation LLC, 766 F.3d 756, 758 (7th Cir. 2014) (expressing skepticism of the Second Circuit’s application of \textit{Campbell}.)

\textsuperscript{102} Authors Guild v. Google, Inc., 804 F.3d 202, 216 (2nd Cir. 2015).

\textsuperscript{103} See Authors Guild v. Google, Inc., 136 S. Ct. 1658 (2016).

\textsuperscript{104} See, e.g., A.V. ex rel Vanderhye v. iParadigms LLC, 562 F.3d 630, 639 (4th Cir. 2009); Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003).

\textsuperscript{105} Video Pipeline, Inc. v. Buena Vista Home Entertainment, Inc. 342 F.3d 191, 200 (3d Cir. 2003).

\textsuperscript{106} \textit{Id.} at 198.
quality.”\textsuperscript{107} The Sixth and Eleventh Circuits take a similar approach to that of \textit{Video Pipeline}.\textsuperscript{108}

As for what qualifies as “new meaning or expression,” the Second Circuit has explicitly rejected the requirement that a fair use defense “must comment on, relate to the historical context of, or critically refer back to the original works.”\textsuperscript{109} Instead, the Court held that “[t]he law imposes no requirement that a work comment on the original or its author in order to be considered transformative, and a secondary work may constitute a fair use even if it serves some purpose other than those . . . identified in the preamble to the statute.”\textsuperscript{110} In \textit{Cariou v. Prince}, the court found that making a collage consisting of copyrighted art was transformative of the original art, even though the artist making the collage did not intend to satirize, parody, or convey any particular message.\textsuperscript{111} The Second Circuit nevertheless found “new expression” in the arrangement of the images. In addressing the \textit{Cariou} decision specifically, The Seventh Circuit has questioned how such re-purposing can be principally distinguished from derivative works.\textsuperscript{112}

How a court views the \textit{Campbell} considerations could substantially impact whether gameplay is considered transformative. Even though fair use may not cover those who seek to emulate game code in most circumstances, the character of gameplay as a collaborative interaction between software and player could mean that gameplay “performances” are in some way intrinsically transformative. Could the character of video game content justify special consideration the same way that parody does?

The view that Let’s Plays by their nature may be viewed as essentially transformative is consistent with the intrinsic tension of copyright protection—promoting the creative use of such assets by content creators, while protecting the rights of video game producers against those who might pirate their work. Unlike a movie or a song,

\textsuperscript{107} Id. at 200.
\textsuperscript{110} Id.
\textsuperscript{111} Id. at 707.
\textsuperscript{112} See Keinitz v. Sconnie Nation LLC, 766 F.3d 756, 758 (7th Cir. 2014).
the consumer is not meant to sit passively while the game operates. A video of an un-played video game is an uninteresting thing. Understanding that video games alone do not create audiovisual output is critical to a principled fair use analysis consistent with the purpose of copyright. The essential nature of a video game in producing unique audiovisual content, even while recognizing its status as entitled to copyright protection, should weigh heavily in favor of fair use.

Remembering that the purpose of copyright law is to promote creative activity, courts must ask whether creative activity would be stifled by the assertion of copyright protection in the case of Let’s Plays, speedruns, or competitive e-sports. But if every gameplay experience is unique and “transformative” of the original, then this portion of the analysis must weigh in favor of the player regardless of whether commentary, editing, criticism, or any other content is added. The addition of commentary, a prominent feature of most Let’s Play videos, should make the argument in favor of fair use that much more persuasive as a “transformative” work since it unquestionably adds the “creative” element that both sides of the circuit split acknowledge as transformative.

It is not always easy to identify when something involves creativity. Some sports, such as figure skating and gymnastics, combine technical and artistic elements. Thus, a gamer’s efforts to rack up points or complete a game quickly may lack the sort of creative content at least one side of the circuit split is looking for to make the use transformative. Nevertheless, there are some games for which the creativity of the player is the core of the game’s purpose. Even absent commentary, the performances and creations of such a game’s users would seem to qualify as transformative under even the narrowest interpretation of Campbell.

2. The Nature of the Copyrighted Work

When courts have analyzed the “nature” of the unauthorized work being used, they have looked to whether the work is fiction, and whether it is published. Those works that are fictional or unpublished are “closer to the core of intended copyright
The Supreme Court has held that creative work typically meets this criterion whether it is published or not, and video games are certainly creative works. As such, the “nature” prong of fair use analysis is easily met, and cuts against fair use.

3. Amount and Substantiality of the Portion Used in Relation to the Whole

The factor that has the most potential to weigh against a Let’s Player is the sheer amount of content they typically make use of. Often, a Let’s Player will record hours of video game content, sometimes the entire course of the game. The use of such large amounts of content are an inherent part of a Let’s Play video, as it is consumed by people who want to see a game played. Substantiality is not simply a redundant reference to quantity, but also requires an evaluation of the quality of the material used.

The Supreme Court found in *Campbell* that the character of a parodic song may permit the parodist to fairly use more substantial portions of a popular song than might otherwise be permitted in other cases. While specific to parody, the Court seemed to be acknowledging that factors such as amount and substantiality ought not be analyzed in a vacuum, but with reference to the character or purpose of the use itself. This is consistent with the Court’s general requirement that the four factors be analyzed together, and with the purpose of copyright law in mind. If playing a game is understood to be transformative, then the amount of gameplay footage captured is ultimately irrelevant because it is a unique work.

One issue that might also arise relates to the “substantiality” of the used portion of gameplay. Modern video games incorporate storytelling techniques similar to that of movies or television shows.

---

116 See *Campbell*, 510 U.S. at 587.
117 *Id.* at 588.
A video game copyright holder may argue that a gamer who plays through certain portions of the video game exposes to potential consumers “the heart” of the work. The Court notes that even a short portion of a work may be more qualitatively significant than a long portion depending on the context.118 A conception of a video game output that analogizes it to a movie or a novel is likely to lead a court to consider the exhibition of certain story elements as weighing against fair use. The more a game resembles a movie, the stronger this argument will be. A hypothetical game with very little interactivity, perhaps requiring only binary inputs in a branching story with multiple endings would have a strong argument in this regard, though it’s questionable whether such software would qualify as a game at all. It is safe to say that a game’s overarching story is, at least for now, typically supplementary—not core—to most gameplay experiences; however, it is possible that extended cut-scenes with little or no interactivity complicate the analysis.

4. Effect on the Potential Market for or Value of the Copyrighted Work

As the Court noted in *Campbell*, the elements affecting the analysis of the third factor also have bearing on the fourth. “A work composed primarily of an original, particularly its heart, with little added or changed, is more likely to be a merely superseding use, fulfilling demand for the original.”119 It is not enough to simply posit that a use will produce economic harm to the copyright holder to render it unfair, as only certain types of harm are legally considerable. “When a lethal parody, like a scathing theater review, kills demand for the original, it does not produce harm cognizable under the Copyright Act . . . the role of the courts is to distinguish between biting criticism that merely suppresses demand and copyright infringement, which usurps it.”120 Let’s Players who criticize the games they stream may undoubtedly produce economic harm to the video game creator, but this is not the kind of harm that copyright protects against. Instead, the video game rights holder must argue that the Let’s Play video or performance takes the place

---

120 Id. at 591–92.
of the video game to the average consumer. In other words, that people will refrain from purchasing the game if they can just watch someone else play it. This is why a rights holder might rely on the argument that the spoiling of story elements in gameplay videos might prevent potential purchasers from buying the game, because they can just watch the story unfold online.

Whether a Let’s Play video or stream can be viewed as usurping the potential market for a protected derivative work may depend on whether the video game creator is likely to produce such a derivative work or license others to do so. Once again however, a showing of harm is not enough unless the Let’s Play is likely to be a substitution for a derivative work. Even if it were likely that video game creators did intend to license others to produce Let’s Plays, as Nintendo appears to be attempting with its content creator’s program, it is not clear that such derivative works would be substitutionary. If every Let’s Play is unique due to the combination of a player’s idiosyncratic inputs along with their commentary or criticism, this may be enough to find that it is not a market substitute. Especially given that an independent Let’s Player is free to criticize the game as they play, expose flaws, exploit glitches, engage in edgy humor, or discuss unrelated topics of the day. A non-independent, corporate-sponsored, official Let’s Player would be unlikely to have the freedom to criticize or satirize the content. If a video game rights holder is also claiming that Let’s Plays are diminishing the value of the original product, this would also seem to undermine any claim that the company intends to produce or license such derivative works.

In the end, it is a question of fact whether a Let’s Play is likely to injure the market for the copyrighted work. While there will undoubtedly be arguments on both sides, there is substantial evidence that Let’s Plays actually increase video games sales—particularly with smaller, lesser-known games. This is why most

---

121 Id. at 592–93.
122 Id.
game companies are supportive of Let’s Plays. Those that are not, such as Nintendo, Konami, and Capcom, may be doing themselves harm through their vigorous assertions of copyright protection, whether such assertions are legally justified or not. If this is indeed the case, such a fact weighs in favor of a finding of fair use, just as it did in Allen and Google.

CONCLUSION

Two considerations should lead to the conclusion that Let’s Plays and other similar media ought to have a strong presumption in favor of fair use. First, that the underlying purpose of copyright, to encourage creative activity, would be undermined by permitting the assertion of copyright claims over Let’s Play videos in most circumstances. Second, that the nature of Let’s Plays as creative performances, taken in conjunction with the four fair use factors, weighs each factor in favor of fair use. Campbell provides precedent for framing the fair use factors through the lens of differing “characters” of media. Courts should recognize that video game output is unique in that it is the product of the interaction of the user with the underlying software, and presume in favor of fair use in copyright disputes. Furthermore, it is not usually in a copyright holder’s interest to impede or litigate against Let’s Players, which courts ought to take into account as they perform their analysis in light of the purpose of copyright law.

PRACTICE POINTERS

- Video game output is likely entitled to copyright protection.
- Case law is unclear on the question of whether a video game “performance” is copyrightable.
- Whether a Let’s Play video or performance qualifies as fair use will depend on how courts balance the fair use factors in the context of different types of Let’s Play, and the level of interactivity present in the game. Let’s Play videos or performances incorporating commentary or criticism are

more likely to qualify as fair use. In circuits adopting the requirement that a use must contain new meaning or expression to qualify as transformative, Let’s Plays focused on technical execution, such as speedruns or e-sports, may have a weaker argument than Let’s Plays emphasizing creativity or the personality of the player. Furthermore, the greater the potential input from the player, and the greater the range of options available to the player, the stronger the fair use argument.
ROBOTS WELCOME? ETHICAL AND LEGAL CONSIDERATIONS FOR WEB CRAWLING AND ScrapING

Zachary Gold and Mark Latonero*
© Zachary Gold and Mark Latonero

Cite as: 13 Wash. J.L. Tech. & Arts 275 (2018)

http://digital.law.washington.edu/dspace-law/handle/1773.1/1817

ABSTRACT

Web crawlers are widely used software programs designed to automatically search the online universe to find and collect information. The data that crawlers provide help make sense of the vast and often chaotic nature of the Web. Crawlers find websites and content that power search engines and online marketplaces. As people and organizations put an ever-increasing amount of information online, tech companies and researchers deploy more advanced algorithms that feed on that data. Even governments and law enforcement now use crawlers to carry out their missions. Despite the ubiquity of crawlers, their use is ambiguously regulated largely by online social norms whereby webpage headers signal whether automated “robots” are welcome to crawl their sites. As courts take on the issues raised by web crawlers, user privacy hangs in the balance. In August 2017, the Northern District of California granted a preliminary injunction in such a case, deciding that

---

* Zachary Gold works for the New York City Mayor’s Office of the Chief Technology Officer. He previously researched privacy and human rights issues at the Data & Society Research Institute, a non-profit independent research organization in New York that studies the social and cultural issues arising from data-centric and automated technologies. He obtained his J.D. from the American University Washington College of Law and his M.A. in Public Policy from SUNY Stony Brook. Dr. Mark Latonero (Ph.D. University of Southern California) leads the Human Rights Program at the Data & Society Research Institute. He is a fellow at UC Berkeley School of Law’s Human Rights Center as well as the USC Annenberg School and Leiden University in the Netherlands. The views expressed in this article are those of the authors and do not reflect the views of the City of New York.
LinkedIn’s website must be open to such crawlers. In March 2018, the District Court for the District of Columbia granted standing for an as-applied challenge to the Computer Fraud and Abuse Act to a group of academic researchers and a news organization. The Court allowed them to proceed with a case in which they now allege the law’s making a violation of website Terms of Service a crime effectively prohibits web crawling and infringes on their First Amendment Rights. In addition, news media is inundated with stories like Cambridge Analytica wherein web crawlers were used to scrape data from millions of Facebook accounts for political purposes.

This paper discusses the history of web crawlers in courts as well as the uses of such programs by a wide array of actors. It addresses ethical and legal issues surrounding the crawling and scraping of data posted online for uses not intended by the original poster or by the website on which the information is hosted. The article further suggests that stronger rules are necessary to protect the users’ initial expectations about how their data would be used, as well as their privacy.

TABLE OF CONTENTS

Introduction............................................................................................................ 277
I. Web Crawlers.................................................................................................. 280
II. Privacy Concerns......................................................................................... 282
III. Government Crawling and Scraping....................................................... 285
    A. The Fourth Amendment ........................................................................ 286
    B. The Third-Party Doctrine .................................................................. 287
    C. Contextual Privacy ............................................................................ 288
IV. Private Sector Crawling............................................................................ 293
    A. Trespass ............................................................................................... 293
    B. The Computer Fraud and Abuse Act ............................................... 295
    C. Overview of Private Sector Use ..................................................... 298
V. Academic Use ............................................................................................ 300
VI. Application.................................................................................................. 302
VII. Policy Dilemmas ........................................................................................ 304
    A. Metadata ............................................................................................. 305
    B. Exclusions and Bias ......................................................................... 305
    C. Data Security ..................................................................................... 306
INTRODUCTION

Scientists, researchers, private industry, and government are tuning in to the changes in information-gathering and analysis brought about by big data. Where relatively small data projects—public opinion surveys, questionnaires, or other similar projects—were once used to provide answers to scientific, business, and civic questions, we can now turn to the much larger store of information on the Internet to try to find better or faster answers to those same questions. Using algorithms and artificial intelligence, we can increase efficiency, augment labor, and complete tasks that are too massive, complicated, or otherwise difficult for humans to realistically complete.

Private companies like Google, Microsoft, and others have for decades provided answers—or, more commonly, provided a list of locations where one might find an answer. They use web crawlers to search and index the web to provide reliable, relevant web pages in response to search queries.¹ Further, these algorithms index a relatively small portion of the worldwide web,² and much less of the broader internet. Not only do these crawlers search a limited number of websites, they also save little information from them. Search engines tend to care only about which websites link to which other websites, maintaining headlines and snippets of text to display to users, or saving thumbnails of images for the same reason. Much of

the data stored on the web is ignored by crawlers entirely, and not scraped for indexing and searching. But this data is the raw material for big data analytics, machine learning algorithms, and similar tools that attempt to analyze, inform, and predict.

While web crawlers are mostly used to collect the relatively limited information necessary to power search engines, they can be used to search, index, and later analyze vast amounts of information on the internet. Increased storage capabilities and computing power are making such usage more practical. Governments can use web crawlers to find criminals operating online. Researchers can use them to identify social trends or political opinions. Private companies may try to glean information about their customers and their preferences from data scrapped from forums, blogs, social media websites, or elsewhere.

These basic functions, long used for well understood purposes, will soon be—or are already being—used to provide the raw data for analyses that many may consider uncomfortable, unethical, or even illegal. They can provide the images necessary to feed a facial recognition system, the content needed to search for violent extremists, or to jump-start a business using data someone else already collected.

This raises a number of questions about the use of such software and the status of the websites they crawl. For this reason, a number of institutions have sought to address this issue. The American Association for Public Opinion Research published its own report identifying data ownership, data stewardship, data collection authority, privacy and reidentification, and data protection as policy challenges to be addressed. The White House, under President Obama, also released a report on big data discussing government uses and providing a background on U.S. privacy law, ranging from

---


Samuel Warren and Louis Brandeis’ *The Right to Privacy*, to the Fair Information Practice Principles and the Consumer Privacy Bill of Rights. The report, among other things, discussed big data’s effect on citizenship, discrimination, and privacy, and made a number of general recommendations, including a national data breach standard, developing technical expertise to stop discrimination, and amending the Electronic Communications Privacy Act (ECPA).

Prior discussions have failed to provide implementable technology or policy solutions, leaving many questions unanswered. In the context of government use, can crawling and scraping ever constitute a search or seizure that would be governed by the Fourth Amendment? More broadly, as applied to the private sector and researchers, do internet users have a privacy interest in what they post online? How and when does such an interest operate? What kind of policies should crawlers obey to protect those searched? Do current federal laws apply to these activities, and do they have the necessary force to meaningfully protect internet users’ data from being made part of a database that will be used for purposes users did not or could not foresee?

Technology often advances ahead of law and policy. Web crawlers are currently governed almost entirely by social norms and politeness, and neither Congress, the executive branch, nor the courts have promulgated laws or guidelines specifically governing their use as tools of surveillance. Without any such rules, there is a near certainty that someone’s privacy has already been, or will soon be violated, their statements connected to their true identity, online posts used against them in court, or some unforeseen harm caused. This article will discuss the problems raised by big data and web crawling from an ethical and legal standpoint. The question of how to regulate crawling and scraping data with bots by government, the private sector, researchers, and individuals will be examined with the goal of identifying issues and highlighting specific dilemmas for policymakers to address before widespread surveillance using web

---


6 Id. at 60.
crawlers can cause undue harm.

I. WEB CRAWLERS

Web Crawlers, also called bots, spiders, and crawlers are in common use on the web. Perhaps of most familiarity to the average internet user, their work product is on display whenever one uses a search engine like Google. Search engines employ crawlers to systematically scan, analyze, and save information about websites to index those sites for searching, determine their importance to a particular search, and find connections between websites.\(^7\)

Web crawlers visit websites at the direction of their operators, but often with little everyday input from them. Operators may choose all the web pages that a crawler will visit, but more often they are driven by algorithms making that determination. For example, Googlebot, the web crawler Google uses to inform its search engine, uses an algorithm to determine what to crawl based on data from previous crawls.\(^8\) These crawlers may visit a given web page a number of times a day to ensure data is collected in a timely fashion.\(^9\) Often, there is a way for website operators to submit their URLs manually to request that a bot crawl their websites.\(^10\) Nevertheless, crawls are often automatic and informed by the sample of the web searched, necessitating that some websites will be left out, and leading to some amount of bias in the results of the crawl. Web crawlers can provide information in real time.\(^11\)

Because crawlers are so active and bandwidth is limited, crawlers generally operate politely, in accordance with social


norms—the desires of website operators are stored in the code of their websites. Crawlers, poorly designed or left to run freely, can use significant network resources or even crash servers.\textsuperscript{12} For this reason, a protocol exists to temper the crawls performed by these bots. Website administrators use the Robots Exclusion Protocol, or “robots.txt”, to ask crawlers not to search particular pages of their website, or to leave it un-crawled entirely.\textsuperscript{13} This file can be targeted at specific bots (for example, telling only Googlebot not to index a page), or at all bots.\textsuperscript{14} Some robots will also respect requests to time delays between crawls to conserve network resources.\textsuperscript{15} However, robots.txt can be ignored; those employing crawlers are not bound by any law, contract, or technical need to obey a robots.txt file.\textsuperscript{16} Only politeness and social pressure provide enforcement power.

There are other methods of keeping crawlers out, such as requiring users to log in, or fill in a captcha, but those too can be sidestepped by a bot’s programmers.\textsuperscript{17}

As technology advances, web crawlers are able to scrape more data from websites. Where it may not have been possible to save all the text or images from a website in the past, as the cost of storage has gone down, the operators of a web crawler can now scrape and store far more information, including comments and the identities of those who posted them, advertisements, and pictures. Advancements in facial recognition technology allow people in images to be identified, and disparate online identities can be

\begin{footnotes}
\footnote{\textsuperscript{12} See \textit{Aren’t Robots Bad for the Web?}, ROBOTSTXT.ORG, http://www.robotstxt.org/faq/bad.html (last visited May 1, 2018) (“Certain robot implementations can (and have in the past) overloaded networks and servers. This happens especially with people who are just starting to write a robot; these days there is sufficient information on robots to prevent some of these mistakes.”).}

\footnote{\textsuperscript{13} See \textit{About /robots.txt}, ROBOTSTXT.ORG, http://www.robotstxt.org/robotstxt.html (last visited May 1, 2018) (explaining how to use robots.txt to allow robots complete access, exclude robots entirely, exclude or allow particular robots, or how to disallow crawling of particular pages).}

\footnote{\textsuperscript{14} \textit{Id}.}

\footnote{\textsuperscript{15} See \textit{Robots.txt Tutorial}, SEOB, http://tools.seobook.com/robots-txt/ (last visited May 1, 2018).}

\footnote{\textsuperscript{16} \textit{Can a /robots.txt Be Used in a Court of Law?}, ROBOTSTXT.ORG, http://www.robotstxt.org/faq/legal.html (last visited May 1, 2018).}

\footnote{\textsuperscript{17} See, \textit{e.g.}, Tim Anderson, \textit{How Captcha Was Foiled: Are You a Man or a Mouse?}, THE GUARDIAN (Aug. 27, 2008), http://www.theguardian.com/technology/2008/aug/28/internet.captcha.}
\end{footnotes}
connected to a real person.

II. PRIVACY CONCERNS

Web crawlers provide the ability for any sufficiently sophisticated and funded operator to maintain a fairly ubiquitous surveillance regime over a larger number of internet domains. This has serious implications for the privacy of internet users. Web crawlers can be used for widespread tracking of internet users without their knowledge or consent. When paired with other technologies, these crawlers can successfully deanonymize people who post online under pseudonyms, or even identify people who have merely had pictures of them posted by others.

Web crawlers can be used to easily acquire large amounts of information, including who posts on which websites, who they interact with, and what they post. This may reveal political, religious, and other views of users, along with significant personal information. Some government agencies already use various methods to track protests and protesters, and eight out of ten law enforcement professionals use social media as a tool in their investigations. Web crawlers enable government agents to quickly collect data from web forums, personal blogs, social networking sites like Twitter, Facebook, and Tumblr, or bulletin boards like Craigslist. Web crawlers also allow government agents to collect data from protest groups’ websites to determine the number of

---

18 See, e.g., George Joseph, Exclusive: Feds Regularly Monitored Black Lives Matter Since Ferguson, THE INTERCEPT (July 24, 2015), https://firstlook.org/theintercept/2015/07/24/documents-show-department-homeland-security-monitoring-black-lives-matter-since-ferguson (explaining that the Department of Homeland Security collected information, “including location data . . . from public social media accounts, including on Facebook, Twitter, and Vine, even for events expected to be peaceful. . . . They also show the department watching over gatherings that seem benign and even mundane. . . . [A] DHS-funded agency planned to monitor a funk music parade and a walk to end breast cancer in the nation’s capital.”).

protestors, identify the protestors, and discover their motivations.\textsuperscript{20} These activities have important constitutional implications as they could chill protected speech, infringe on protester’s freedom of association, or violate a person’s Fourth Amendment right to protection against unreasonable searches. Corporations and researchers are also using crawlers to scrape internet data to inform their business practices and research.\textsuperscript{21} While these corporate practices do not implicate the same constitutional rights as government use of crawlers, they do have significant bearing on the privacy rights of internet users whose data is collected. Not only might the initial collection by corporations or researchers violate the privacy of internet users, but poor security practices could result in data breaches putting personal data in the hands of people with malicious motives.

This collection of information can be done without the knowledge or consent of those posting. Users post online with certain expectations about how their posts will be used, and while they may use websites that include privacy controls or have terms of service (ToS) forbidding crawling, these may be circumvented. Privacy controls are often too confusing for users to employ effectively,\textsuperscript{22} and in any case do not control what others post. And, as discussed above, very little controls the ability of web crawlers to scrape data from a web page.\textsuperscript{23} This means that government

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{20} See Richard Esposito et al., \textit{Showden Docs Reveal British Spies Snooped on YouTube and Facebook}, NBC News (Jan. 27, 2014), http://investigations.nbcnews.com/_news/2014/01/27/22469304-snowden-docs-reveal-british-spies-snooped-on-youtube-and-facebook. In 2012, the British Government Communications Headquarters demonstrated the ability to monitor YouTube, Facebook, and Twitter in real time; this sort of information apparently has value to governments interested in monitoring online activity.
\item \textsuperscript{21} See discussion infra Part V.
\item \textsuperscript{22} See Josh Constine, \textit{Facebook Admits Users Are Confused About Privacy, Will Show More On-Screen Explanations}, TECHCRUNCH (Apr. 8, 2014), http://techcrunch.com/2014/04/08/facebook-privacy-settings/ (“Facebook’s privacy team manager Mike Nowak admitted that people think Facebook changes its privacy controls too often or that the company has failed to make privacy easy to understand.”).
\item \textsuperscript{23} See, e.g., \textit{How Do I Prevent Robots Scanning My Site?}, ROBOTS.TXT, http://www.robotstxt.org/faq/prevent.html (last visited May 1, 2018) (providing advice on how to prevent scraping by crawlers, but noting “this only helps with
agencies, corporations, or others can easily navigate around users’ expectations, collecting whatever data they want without the subjects of the surveillance ever learning of the collection, much less having a chance to consent.

This sort of tracking, scraping, and storage of information allows governments to engage in further invasions of privacy beyond merely collecting information on individuals as they interact both online and offline. Such practices have serious implications for unmasking real identities online.

Facial recognition technology can, to varying degrees, accurately identify a person in a picture. This allows a government agency, or others, to scrape images from websites to identify the people in the photos, creating a database of users, their acquaintances, and friends. Because metadata is often uploaded with such photos, the times and locations of users’ meetings may also be collected. To an increasing extent, clear images of peoples’ faces are not necessary as computers are being trained to identify people based on factors like hair style, clothing, body shape, and pose. Users cannot avoid this sort of surveillance by refraining from taking pictures of themselves, or by asking their friends not to post photos or tag them. It is possible that images posted by strangers may lead to ones’ identification in the background of a picture with an entirely different subject.

Such crawling and scraping can also be used to unmask aliases. Crawlers may scrape information like physical addresses, email addresses, phone numbers, or linked accounts that can be used to link aliases to each other, or to link an alias to a real-world identity, stymying attempts to speak anonymously. While this is certainly well-behaved robots.


possible without crawlers, crawlers’ ability to search constantly and systematically increases the chances that a user’s mistake or private material will be found and taken advantage of. Further, this can be done on a large scale, leading to the potential unmasking of a great number of aliases.

Crawlers are accessible to nearly anyone with a bit of technical expertise and access to the necessary computing resources to complete their task. While government crawling and scraping has implications for the privacy as well as the First and Fourth Amendment rights of U.S. citizens, application of these tools by private entities is not without risks.

III. GOVERNMENT CRAWLING AND SCRAPING

Government agencies, from the federal level to local police departments, are already putting information they find online to use. Law enforcement uses social media to anticipate crime, but nearly half of law enforcement agencies have no formal process governing the use of social media for their investigations. This leaves open the possibility of abuse and allows law enforcement professionals to ignore privacy expectations of internet users. The federal government uses data mining to find terrorists by looking for relationships between people and connections between behaviors, and has programs aimed at analyzing “massive” data sets.

Government searches are governed by the Fourth Amendment. Yet whether web crawlers constitute a search under the Amendment is unsettled. There are generally two possible interpretations of the Fourth Amendment’s privacy protections: The Third-Party

---

26 See LEXISNEXIS, supra note 19, at 3.
27 See LEXISNEXIS, supra note 19, at 2.
29 U.S. CONST. amend. IV. Persistent surveillance online also could have a significant chilling effect on speech. For its First Amendment implications, see Karen Gullo, Surveillance Chills Speech—As New Studies Show—And Free Association Suffers, ELECTRONIC FRONTIER FOUNDATION (May 19, 2016), https://www.eff.org/deeplinks/2016/05/when-surveillance-chills-speech-new-studies-show-our-rights-free-association.
Doctrine, and a more contextual view of privacy focusing on the amount the surveillance uncovers about a person’s life.

A. The Fourth Amendment

The Fourth Amendment’s limitation on unreasonable searches applies only to public actors, but it carries great weight in the discussion of online privacy concerns, as the government exercises vast power online to monitor user activity. The Fourth Amendment goes a long distance in shaping the public’s perception of their rights in relation to private actors as well, while they are not actually bound by those same constitutional guarantees.

For many years after its conception, courts understood the Fourth Amendment as protecting against a physical invasion of privacy, including a government agent’s trespass onto land, or the physical taking of a private citizen’s possession. More ephemeral information—like conversations overheard from a location a government agent had a right to be—were granted no protection. It is unclear to what degree trespass may apply to online actions, making it uncertain whether the Fourth Amendment binds government searches online based on a theory of trespass.

Some courts hold that a claim for civil trespass can be sustained based on the use of server resources by a web crawler. In cases where web crawlers used rather small amounts of server resources to search and scrape data from websites, claims against the operators of those web crawlers for trespass have stood. This theory of

---


32 See id.

33 See e.g., eBay, Inc. v. Bidder’s Edge, Inc., 100 F. Supp. 2d 1058, 1070 (N.D. Cal. 2000).

34 See id.; but see Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *3 (C.D. Cal. Mar. 7, 2003) (“This court respectfully disagrees with other district courts' finding that mere use of a spider to enter a publicly available web site to gather information, without more, is sufficient to fulfill the harm requirement for trespass to chattels.”).
online trespass is not widely accepted, but it could subject government web crawls to the Fourth Amendment. The architecture of the modern web, which puts nearly everyone’s data on someone else’s computer through the use of cloud computing, may hamper the use of this doctrine online. The government would not be trespassing on the end user’s computer, but onto some company’s. In such a case, the end user may never find out, forcing them to rely on others to notify them or to enforce their rights.

In 1967, the U.S. Supreme Court decided United States v. Katz, explaining that the “Fourth Amendment protects people, not places.” The Court held that a person making a phone call in a phone booth had a reasonable expectation of privacy in his conversation, thus preventing government eavesdropping without a warrant. In later cases, the Court elaborated that a search is unreasonable and violates the Fourth Amendment when the target of the search has manifested an expectation of privacy that society considers reasonable.

B. The Third-Party Doctrine

The Third-Party Doctrine states that there is “no legitimate expectation of privacy in information [one] voluntarily turns over to third parties.” A number of cases decided before the creation of the Internet provide for significant government access to records and other information. Applying this doctrine, the courts determined that a number of records held by institutions for or about individuals are unprotected regardless of the use for which they are shared. Courts

35 See Ticketmaster, 2003 WL 21406289, at *3 ("[S]cholars and practitioners alike have criticized the extension of the trespass to chattels doctrine to the internet context, noting that this doctrinal expansion threatens basic internet functions (i.e., search engines) and exposes the flaws inherent in applying doctrines based in real and tangible property to cyberspace.").
37 Id. at 351.
38 Id. at 353.
held that the Fourth Amendment did not prohibit the government from obtaining information revealed to a third party, even if the information was revealed on the assumption that it will be used only for a limited purpose and the confidence placed in the third party will not be betrayed. This doctrine neatly fits into the Katz test, which protects people when they take action to keep their information private. The Third-Party Doctrine adds the presumption that a person can have no legitimate expectation of privacy in shared information.

The impact of the Third-Party doctrine may have been reasonable when it was adopted, but its impact on privacy online is plain and oversized. Online, all of one’s activities are shared with a third party. Emails are shared with an email client. The websites one visits are shared with an ISP, and any number of entities that have attached cookies to the browser being used. Everything one does online is shared by the very nature of the Internet; even while browsing alone, some intermediary between one’s PC and the server contacted is recording an exchange of packets. As a result, privacy rights are significantly curtailed online. For example, the Electronic Communications Privacy Act, passed in 1986, provides protection against the search and seizure of emails in transit, in storage on a home computer, or stored on what would now be called the “cloud” for 180 days or less. The government must obtain a warrant for such data. For email stored in the cloud for more than 180 days, or opened and stored in the cloud, the government can compel disclosure with only a subpoena. This constitutes less protection than email stored locally, on one’s computer (or on paper, in a file cabinet) would get.

C. Contextual Privacy

The views of the Fourth Amendment described above, and the Third-Party Doctrine, assume a black and white view of privacy where any sharing of information, regardless of the purpose,

---

42 See Miller, 425 U.S. at 443.
removes any privacy the user could have expected to have in that information. A more nuanced view of privacy is possible, through which internet users would not be denied their privacy based on technical necessities, nor their activities treated as an entirely new realm deserving of a new view of privacy. Instead, the context of the use should be determinative. Just as a patient would be shocked if a doctor shared his information with marketers, but would likely have little issue with that same information being shared with an insurance company or pharmacist, privacy expectations online are contextual. Users share their emails with Google and may expect ads to be shown to them based on the content of those emails, but may not expect those emails to be shared with the government. Under a contextual view, a person’s privacy level would depend on the use of the technology.

Of course, applying offline rules to online activities could mean applying the Third-Party Doctrine. But some recent Supreme Court cases might point to a changing view on that issue. In United States v. Jones, the Supreme Court unanimously agreed that long-term tracking of a suspect using a GPS device placed on the suspect’s car required a warrant. This ruling has significant implications for web crawling. Addressing long-term tracking first, the Court held that it was not reasonable to expect that a government agent would follow someone for a long period of time. Online surveillance and web crawling allow the government to do just that, searching the web and scraping websites for every trace a given user leaves, going back in time as far as any website maintains its data.

47 This hypothetical ignores, for the sake of argument, the significant laws that govern the handling of medical information and focuses merely on consumer, or patient, expectations.
48 Nissenbaum, supra note 46, at 38.
49 Additionally, in the particular case of email, the change in how email is used since the passage of ECPA and the routine storage of large numbers of emails and other documents in the cloud, rather than on home computers, bolsters the argument that users do expect a different amount of privacy than ECPA provides, at the very least.
A concurrence by four justices rejected the majority’s trespass-based approach and determined that a reasonable person did not expect to be tracked with a GPS unit over a long period of time, which in this case, was about a month. The reasoning of the four concurring justices, adopting a new approach to apply to persistent, long-term tracking which was either impossible or prohibitively expensive in the past, may signal a coming change in how such cases are decided. Such a view may even lead to a significant curtailing, if not the end, of the Third-Party Doctrine.

Justice Sotomayor, in her own concurrence, expressed concern over the application of trespass in the electronic age given that many forms of surveillance require no trespass. For instance, tracking the GPS chip in a suspect’s phone, rather than placing one somewhere on his person or possessions. Justice Sotomayor was explicitly worried about electronic surveillance and went as far as suggesting that the Third-Party Doctrine be reconsidered. She said the approach was “ill suited to the digital age, in which people reveal a great deal of information about themselves to third parties in the course of carrying out mundane tasks.” As one scholar put it, “all communications over the Internet . . . are stored for various lengths of time on third party servers or Internet service providers.” Justice Sotomayor cited *Katz* for the proposition that “what [a person] seeks to preserve as private, even in an area accessible to the public, may be constitutionally protected.” Further, computers, including those online or in the cloud, are routinely used to hold the sorts of documents, photographs, and other private matters that were previously kept in the home. Without changes to the Third-Party Doctrine, these documents would lose protection merely because of where they are stored.

---

51 Id. at 418 (Alito, J., concurring).
52 Id. at 413 (Sotomayor, J., concurring).
53 Id. at 417.
In Riley v. California, the Supreme Court discussed how searches of cell phones can reveal far more than just one sort of information contained in them would otherwise reveal.\footnote{See Riley v. California, 134 S. Ct. 2473, 2489 (2014).} In this case, the government searched a cell phone incident to arrest.\footnote{Id. at 2482.} The Court reaffirmed that searches of cell phones under this authority must occur to protect officer safety or to preserve evidence, and otherwise require a warrant or exigent circumstances.\footnote{Id. at 2483.} However, recognizing the difference between collecting large and small amounts of information has clear implications for government use of web crawlers.

Although neither Jones nor Riley addressed online surveillance specifically, it seems clear that long-term surveillance, or surveillance that covers a wide variety of information (and perhaps even information shared online in at least some contexts) may not be completely unprotected under the Fourth Amendment. These cases drew a line based on the amount of data collected; they alleged that when the government collects enough data, even if it is public, the nature of the collection can change and violate a persons’ privacy.

Scholars have suggested new ways to apply the Fourth Amendment online in a way that would protect the privacy of those who share information online. One way is to protect content, while allowing the government to collect non-content information.\footnote{See Orin S. Kerr, Applying the Fourth Amendment to the Internet: A General Approach, 62 STAN. L. REV. 1005, 1029 (2010). Non-content information, or metadata, is information “related to identity, location, and time.” Id. at 1018. Metadata could feasibly include email addresses, account names, IP addresses, or other similar information. See also Chris Conley, Metadata: Piecing Together a Privacy Solution, ACLU OF CALIFORNIA (Feb. 2014), https://www.aclunc.org/sites/default/files/Metadata%20report%20FINAL%20%2021%20cover%20%2B%20inside%20for%20web%20%283%29.pdf.} This was proposed as being similar to the inside/outside distinction applied in physical space, in which people have a greater degree of protection under the Fourth Amendment inside, in private spaces, than they do outside, in public.\footnote{Kerr, supra note 60, at 1009.}
protections currently applied to post mail and telephone calls, but may draw critics based on the revealing nature of metadata.

Alternatively, one could apply Fourth Amendment protections online based on the “structure of the particular technology” and “the particular uses to which an individual puts the technology.” Under such an approach, password protected information stored in the cloud would be protected, even if it were non-content information, just as if it were held in a filing cabinet in one’s home. Determining how to deal with social media is difficult under this approach, but could be determined based on the amount of control the user maintains over access to the information, even if the owner of the platform has access for certain purposes. The court could ask if “assuming privacy settings are optional, [the ‘resident’] chose privacy settings that would support a finding that his [social media sites are] sufficiently restricted that they are not readily available to the general public.” Just as in determining whether to treat a physical space as a residence, courts should not inquire too closely into the specific uses an individual chooses to make of an online social space; an individual does not have a lesser basic expectation of privacy against the government in their home simply because they have frequent parties or have a large number of guests.

Finally, the Fourth Amendment could be read to protect certain “structural privacy rights.” Acknowledging that prior to certain technological advancements, some forms of surveillance were too expensive to employ, the courts should strive to maintain protections at that level. For example, while following a given person was once prohibitively expensive, one can now be followed electronically with the use of the GPS chip in one’s phone. A rule designed to

---

62 Id. at 1019.
63 Id. at 1032.
64 Strandburg, supra note 56, at 659–60.
65 Id.
66 Id. at 661–62.
67 Id. at 663 (citing Crispin v. Christian Audigier, Inc., 717 F. Supp. 2d 965, 991 (C.D. Cal. 2010)).
68 Id.
protect a structural privacy right would use the Fourth Amendment to impose legal costs where there were once economic costs.\footnote{Id.}

IV. PRIVATE SECTOR CRAWLING

The private sector may have many uses for crawling and for scraped data beyond those discussed above. Companies can use them to gather information on their customers’ views on certain products they’ve purchased. They can gather information about pricing on their competitors’ websites. They could also be used to gather significant amounts of information on their customers from personal blogs, social media sites, forums, and other websites where users may talk about or otherwise make their identity or their preferences known. This could allow companies to gather large dossiers of sensitive information with few, if any, rules about what can be gathered, when and where it can be gathered from, along with generally weak rules about the storage of information. This section will discuss the case law applicable to corporate use of web crawlers and the policy implications of corporate use. Some sectors of the U.S. economy are governed by industry-specific privacy regulations.\footnote{These privacy regulations will be discussed where applicable, but they are relatively narrow in scope and are largely outside the scope of this paper.}

A. Trespass

In eBay v. Bidder’s Edge, a California district court was faced with determining whether Bidder’s Edge, an auction aggregation site, could crawl eBay’s website, scrape information on bids, and provide search results to its own users.\footnote{eBay, Inc. v. Bidder’s Edge, Inc., 100 F. Supp. 2d 1058, 1067 (N.D. Cal. 2000).} The court held that such unpermitted crawling amounted to trespass, and ordered an injunction to stop Bidder’s Edge from continuing its crawling and scraping of eBay.\footnote{Id. at 1069–70.} The court came to this decision even though Bidder’s Edge used very little of eBay’s server resources (a couple of percent, at most), and did not damage the property, though it did
prevent eBay from using a small percent of server resources for other uses.

Another California court attempted to apply this “ancient common law action to the modern age.” 74 Prior courts held that “mere invasion or use of a portion of the web site by a spider is a trespass (leading at least to nominal damages), and that there need not be an independent showing of direct harm either to the chattel (unlikely in the case of a spider) or tangible interference with the use of the computer being invaded.” 75 The Ticketmaster court, however, required a showing that the computer being crawled be adversely affected by the use of the spider, rejecting that “mere use of a spider to enter a publicly available web site to gather information, without more, is sufficient to fulfill the harm requirement for trespass to chattels.” 76

The California Supreme Court dealt with a similar issue where a former company employee sent a number of emails to his former coworkers’ corporate email accounts. 77 Here, a number of emails were sent to employees, who were given the choice to opt out of receiving the emails. 78 Intel argued that it deserved an injunction against the sending of those emails, as the emails were a trespass on its server that ate up server and human resources (time spent replying, setting up filters, etc.). 79 However, the court declined to find a trespass, as California law required some damage to the property. Here, there was no allegation that the emails impaired the functioning of Intel’s computers, and the emails were allowed to be sent. 80

Courts have come to vastly different conclusions about whether trespass applies online, and have made some important points in doing so. First, it is important to note that Intel v. Hamidi depended on the definition of trespass, a common law concept that can differ

---

75 Id. (noting the criticism of extending trespass to the internet).
76 Id. (disregarding the work load performed by Ticketmaster’s servers to accommodate Tickets.com’s crawlers).
78 Id.
79 Id. at 300.
80 Id. at 311.
from one jurisdiction to another. Second, it is unclear what the definition of damage is when applied to the use of a server. One court found that merely using server resources was enough to find damage, while another found that a minimal use that did not affect the operation of the computer at issue was not enough for a court to find damage.\footnote{Compare eBay v. Bidder’s Edge, Inc., 100 F. Supp. 2d 1058 with Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *3 (C.D. Cal. Mar. 7, 2003).} It is unclear, based on these opinions, whether merely using a computer’s resources constitutes damage, and if not, how much of a computer’s resources must be used for a court to find it was damaged. It is also unclear what sort of warnings are required to make it known that a crawler is unwelcome. \textit{Intel v. Hamidi} did not address the issue in-depth,\footnote{Id. at 300. In fact, Intel did not appeal to Hamidi to stop sending the messages, but merely attempted to block the receipt of them by Intel employees.} and eBay notified Bidder’s Edge in multiple ways that their crawlers were unwelcome.\footnote{eBay, Inc. v. Bidder’s Edge, Inc., 100 F. Supp. 2d 1058, 1062 (N.D. Cal. 2000).} Would merely having a robots.txt header forbidding crawling or posting it in a website’s ToS be enough? If any use of server resources without permission is a trespass, then how can the operator of a crawler find out what is in a target website’s robots.txt header or ToS without crawling? The common law cause of action of trespass does not provide a rule clear enough for the operators of web crawlers to follow, and leaves enforcement largely up to websites, not end users whose data is actually at issue. It is not enough to ensure user privacy from web crawlers only when it is desired.

\textbf{B. The Computer Fraud and Abuse Act}

The Computer Fraud and Abuse Act of 1986 (CFAA)\footnote{Computer Fraud and Abuse Act, 18 U.S.C. § 1030 (2018).} protects computers from unauthorized access and from access that exceeds authorization.\footnote{Id.} The law provides for both criminal and civil penalties.\footnote{Id. § 1030(c).} At times, courts have addressed whether unauthorized crawling and scraping can violate the CFAA. Because the CFAA
was passed in 1986, it does not incorporate web crawlers into its provisions. It is not clear how the law would apply to such software, as the following cases illustrate. Further, courts have been divided over how the CFAA should be applied outside of the limited case of web crawlers.87

In *EF Cultural Travel BV v. Explorica*, the First Circuit was tasked with determining whether scraping a website violated the CFAA. Its determination of whether access was unauthorized in this particular case is outside the scope of this paper, as it hinged on a confidentiality agreement signed by a former employee of the company whose website was scraped, and not on an html header, ToS, or other commonly used means of signaling a desire not to be crawled or scraped.88 However, the court also looked at whether the scraping met the damage or loss requirements of the CFAA. The court found that EF Cultural Travel had suffered a loss due to Explorica’s scraping, under a theory that Congress had intended loss “to target remedial expenses borne by victims that could not properly be considered direct damage caused by a computer hacker.”89 Because EF Cultural Travel had been forced to take “diagnostic measures” to “assess whether their website had been compromised,”90 they had suffered a loss. Though EF Cultural Travel suffered no physical damage, the court determined that Congress, by specifying that either damage or loss would enable recovery under the CFAA, had intended that no physical damage was necessary.91 However, nine years later, the District Court of Maryland held that for lost revenue to qualify as a “loss” under the CFAA, the unauthorized access in question must have caused an interruption of service.92 Other courts have declined to follow that definition.93

88 EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577, 582 (1st Cir. 2001).
89 *Id.* at 585 (citing *In re DoubleClick Inc. Privacy Litig.*, 154 F. Supp. 2d 497, 521 (S.D.N.Y. 2001)).
90 *Id.* at 584 & n.17.
91 *Id.* at 585.
93 See, e.g., Am. Family Mut. Ins. Co. v. Gustafson, No. 08–cv–02772–MSK,
In another case, *AOL v. LCGM*, the court held that LCGM violated the CFAA by sending bulk email to AOL subscribers in violation of AOL’s ToS\(^{94}\) and by collecting those email addresses in violation of the same ToS.\(^ {95}\) Again, LCGM caused AOL to incur technical costs as a result of their actions, impaired the functioning of AOL’s network, and damaged AOL’s goodwill.\(^ {96}\)

Over the years, courts have operated under a number of different rules regarding when the CFAA applies. However, it seems clear that a web crawler visiting a target website, using its resources, and scraping it for data, could violate the CFAA. Web crawlers can certainly operate in violation of an html header or of a ToS,\(^ {97}\) and they also use resources of the servers they contact, which could cause a service disruption. Consequently, website operators wishing to keep crawlers away from their site must expend money and resources responding to such visits.

Nevertheless, in a recent case the Northern District of California found there was likely no violation of the CFAA in a suit brought by LinkedIn against hiQ, which scraped LinkedIn for publicly accessible data in violation of LinkedIn’s ToS.\(^ {98}\) The court distinguished previous cases,\(^ {99}\) finding a CFAA violation in similar circumstances, while noting that unlike previous cases, hiQ was scraping public data rather than password protected parts of


\(^{95}\) Id. at 450–51.

\(^{96}\) Id. at 451.

\(^{97}\) See, e.g., Kerr, *supra* note 87, at 1165–67 (noting that some scholars do not think that ToS should be binding on web users, as they are rarely read, hard to understand, and better understood as limits on liability than as limits on who can use the website).


\(^{99}\) Id. (citing United States v. Nosal, 844 F.3d 1024, 1038 (9th Cir. 2016) & Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016)).
The court explained that, unlike in *United States v. Nosal* or *Facebook v. Power Ventures*, where “unauthorized intruders reached into what would fairly be characterized as the private interior of a computer system not visible to the public,” the scraping at issue here was publicly available, without a password, and this put it outside Congress’ intent in passing the CFAA to prevent hacking. Further, the court reasoned that applying the CFAA in the way LinkedIn suggested “would have sweeping consequences well beyond anything Congress could have contemplated,” potentially creating criminal liability for “merely viewing a website in contravention of a unilateral directive from a private entity . . . effectuating the digital equivalence of Medusa.”

The court also discussed how to apply the concept of trespass to online domains, determining that social norms tell us the Web is “inherently open,” and that the CFAA’s bar on “access without authorization” probably does not apply to publicly available portions of a website. The court awarded hiQ a preliminary injunction barring LinkedIn from preventing hiQ’s scraping activity on their website.

### C. Overview of Private Sector Use

Private sector corporations are subject to significant restrictions on what and when they can crawl. Unlike the restrictions on the government, these restrictions are not theoretical, though they are hardly clear-cut. It seems that corporate operators of web crawlers may need to abide by the desires of websites to not be crawled, whether that preference is made known in a robots.txt header, a ToS, or otherwise. However, this is dependent on the ability and willingness of websites to use litigation to stop crawlers from operating on and scraping their website, leaving smaller websites and users in a jam.

---

100 *Id.* at 1109.
101 *Id.*
102 *Id.*
103 *Id.* at 1110.
104 *Id.* at 1111 (citing Kerr, *supra* note 87, at 1162).
These controls on private operators of web crawlers are available only to the operators of a website. Individual users cannot ensure their data is not crawled or scraped, and must rely on the operators of the websites they use to maintain their privacy against crawlers. Given how vague case law is on the subject, it is unclear whether users or websites can rely on these protections to keep their data private and out of corporate databases. Many websites and users may be unable to protect themselves, and some websites may find it is in their interest to allow crawlers to scrape their data, regardless of some of their users’ wishes.

For example, web forums may lack the resources and money to defend their users’ information from those who wish to scrape it. While some forums are quite large, most are small and likely lack the technical, monetary, and legal resources to stop an organization that insists on ignoring their calls to refrain from crawling and scraping. These forums may be quite interested in protecting their data; forums often host discussions on personal issues, including those of sex, medical conditions, and others, and have a reputation that they wish to maintain among their users. However, they often do not monetize this data beyond serving ads to those who read or post. This limits their resources and how valuable that data is to the forum; they lose no value if another group holds the same data. These sorts of forums may not be willing or able to protect their users’ privacy and users have no way of signaling their desire not to have their posts crawled, and suffer even more from a lack of resources. Other websites, like Twitter, do monetize the data they collect by limiting the ways that data can be culled from their service and charging users to access the full archive of tweets.  

Social networks collect even more data than forums, and this data is perhaps more sensitive and specific than that people post on

---

forums. But like forums, social networks have a reputation to protect, and the larger ones may have significant resources and a desire to keep whatever information they have to themselves, and monetize it as they see fit. For example, Facebook collects, and reveals, large amounts of data about its users. It uses the data to make recommendations, displays news stories of potential interest, and shows advertisements based on the information scraped. In 2018, amid a media firestorm, Facebook’s CTO confirmed that a private company, Cambridge Analytica, surreptitiously scraped data from 87 million users. The firm reportedly collected the Facebook profiles in order to target voters during the 2016 U.S. Presidential election. This incident focused international attention on the risk of crawlers deployed by third parties harvesting detailed personal data found on proprietary social networks.

V. ACADEMIC USE

Crawlers also have potential for academic researchers in social science, computer science, and other fields. Internet research has greatly expanded the methods for social analysis used by researchers. Now, in addition to traditional surveys, researchers can collect vast amounts of data from online communities, social media, and various websites to answer questions on topics such as youth attitudes, demographic change, or political beliefs.

In the same way that the government or corporations may use web crawlers to collect sensitive data that users meant to keep private, researchers may collect significant data on a much wider array of issues of noncommercial general inquiry. While searching for private, closely held beliefs and ideas can lead to valid findings, researchers in academic institutions are bound by the same laws that govern the private sector and have additional institutional controls over their research.


The CFAA arguably stands in the way of academics who want to use crawlers. Researchers may seek to deploy web crawlers and other bots to gather and analyze data for basic and applied research publications adding to literature of their disciplines. The tension surrounding this use is not theoretical. In 2017, University of Michigan Professor Christian Sandvig, his academic colleagues, and the news organization First Look Media Works, intended to conduct research on online discrimination using methods including web crawlers where such conduct is prohibited by the CFAA. The ACLU filed a lawsuit on their behalf against the U.S. Attorney General over the CFAA’s criminalization of such research activities according to the website’s ToS. The plaintiffs are concerned that conducting their research with crawlers, which they allege will cause no harm to the websites they study, will expose them to significant criminal liability. The case has not yet been decided on the merits, but they have been allowed to move forward with an as-applied challenge to the CFAA on the Free Speech and Free Press Clauses of the First Amendment. Even if their case is successful, the website ToS will remain in force and they may be prohibited from accessing the websites themselves or be subject to civil actions.

Academics performing studies have more oversight on their research than some other actors. Institutional Review Boards (IRBs) are tasked with reviewing and approving proposed human research by academics. IRBs are supposed to ensure that researchers obtain informed consent from their subjects and do not expose them to undue risk of harm.

However, there are number of problems with the IRB process. First, they often take a long time to complete their reviews (often months), keeping them slightly behind the newest technology. They also may not necessarily understand the problems associated with collecting data online; while using publicly available data posted on the web may not appear to be human subjects research, such data use clearly can have significant impact on the lives of those who posted it. Finally, many researchers use “found” data, or data that has been collected by another entity, which is either publicly

available online or given to them by a private company, without further review.\textsuperscript{110} This allows researchers to avoid institutional review even when they are subjecting the data to new analysis and may uncover novel findings about those who posted the data online. Such use creates another point of failure where personally identifiable information can be revealed or data can be leaked. Considering the possible problems with avoiding review in this way is made more important in light of recent calls for researchers to open up the data they use in their research and to share it with others in their field.\textsuperscript{111}

Academic researchers need clearer rules about mandatory review of the analyses they wish to perform on this sort of data, even when it is collected by another entity. Academic actors collect information and perform studies on topics that are just as sensitive as the projects carried out by the government. They study religion, sex, gender, and a host of other topics, many times focusing on vulnerable or disenfranchised populations. Institutions reviewing this sort of research need to ensure that the studies they produce are conducted with respect for the privacy of those using the internet and that the data collected is handled and saved responsibly.

VI. APPLICATION

Given this state of affairs, users may enjoy some degree of privacy online, even in the information that they post publicly. However, the existing laws and guidelines governing the use of web crawlers to gather information on the web are inadequate to the task

\textsuperscript{110} See 45 C.F.R. § 46.101(b)(4) (exempting from the human research subjects policy “Research, involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.”).

\textsuperscript{111} See, e.g., Paige Shaklee, New Data Journal Lets Researchers Share Their Data Open Access, ELSEVIER CONNECT (Sep. 9, 2014), https://www.elsevier.com/connect/new-data-journal-lets-researchers-share-their-data-open-access (“[E]ach piece of data that has been carefully and thoughtfully gathered has value. Often, you don't know what future value that data will have until you've shared it with colleagues in the scientific community.”).
of protecting privacy interests. While the courts have not dealt with government surveillance using web crawlers, a wide search could turn up enough information, in aggregate, to create a search subject to the Fourth Amendment. Just as tracking a person with a GPS unit for 30 days reveals much about that person’s life, so could crawling and scraping enough data about a particular person. Such searches threaten to reveal nearly everything about a person’s life without the knowledge of those being searched. Law enforcement also recognizes that using online material for policing purposes requires walking a fine line. The Bureau of Justice Assistance produced a report recommending that police departments institute policies governing when such tools can be used, what authorization is needed, and how collected data should be stored.

A similar expectation of privacy exists against privately operated web crawlers, though this expectation is largely enforceable only by the website hosting the information, not the end user. While online trespass is not widely accepted as a good idea among the legal community, and the CFAA was not aimed specifically at protecting from this kind of harm, these bodies of law do provide some protection against robot searches. Such crawls, if unwanted, could create a private cause of action against those operating the web crawlers, though there are practical concerns to enforcing such a prohibition on crawling.

Beyond the legal norms discouraging unwanted crawling and scraping of data from websites, ethical and social norms are in place. Facebook, whose founder once said that privacy was no longer a social norm, has changed its sharing default from “public” to “friends.”

---

112 These could reveal locations from check-ins and photos on social networks, opinions about politics, social movements, and literature, names of friends and acquaintances, product reviews on online marketplaces, and more.


step to remain private online, and sixty-eight percent say that stronger laws are needed to protect people’s online privacy.115 People attempt to guard their identity, keep information from specific people or organizations, and care quite strongly that they control who has access to much of their information.116

To ensure that internet users’ privacy is maintained, more work is needed to put in place strong administrative and legal protections. At the moment, it is unclear how the law applies to web crawlers in all jurisdictions. Private sector actors, including academic institutions, have weak controls on their use of these tools. More accountability is needed, and clearer rules need to be put in place to ensure that web crawlers are not abused and internet users do not suffer undue harm. The remainder of this paper will discuss some of the policy questions that need to be considered while crafting these rules.

VII. POLICY DILEMMAS

Internet users have certain expectations about their use that web crawlers may confound. Certain social norms exist surrounding use of the Internet and particular websites on it. For example, when users post an update on Facebook, they expect that post is for the use and enjoyment of their friends. Though it may be available to the public, most people are unlikely to think that their posts will be scrutinized and used to profile them.117 Further, many websites have rules prohibiting web crawling, contributing to the belief that people’s data will not be scooped up by a bot sent on a mission to find any data that it can. Government, corporate, and university web crawling

---


116 Id.

It allows large organizations to build a comprehensive profile on any person or organization it would like to, at very low cost to those operating web crawlers.

A. Metadata

Crawlers allow for the compilation of a significant amount of metadata about users. This metadata can be extremely revealing, is often unprotected, and may not be protected from government search under the Fourth Amendment. With some effort, metadata from anonymous accounts could be linked to a real identity, meaning that users could not escape being tracked by using an alias or username not plainly associated with them. A person’s religious views, medical status, or other personal information could be determined just from viewing metadata.

This information could be embarrassing, used against people in courts or among the public, and could be data that a person never wanted linked back to their real identity. Using web crawlers to collect and index this sort of data could thwart all of those expectations.

B. Exclusions and Bias

Crawlers do not, and perhaps cannot, search everything. They will inevitably miss information, fail to search some websites, or mistakenly believe that some information is not relevant to its search and fail to collect it. As with all other methods of data collection, some people and data will be excluded from the searches conducted by crawlers. What this means for those operating web crawlers is not entirely known. In the context of the government, it means that searches for criminals will never be perfect. For corporations or researchers, it means that searches designed to study a given community will miss people, and fail to provide a full picture. This could bias any resulting conclusions drawn from such data, and require that those directing searches consider how inclusive their search will be and ways to correct for such exclusion bias.

Searches conducted with crawlers will suffer from more

---

118 Though, after the release of the Snowden documents, people may be more aware of the surveillance they are subject to online. See also Esposito et al., supra note 20.
traditional forms of bias. Just as someone drafting questions for an opinion poll may choose words that push people towards a certain answer, programmers may choose search terms, or construct their algorithms in such a way that their bots are drawn to certain types of data, and hence certain types of answers. This also leaves open the possibility that the searching organization may miss someone, mistakenly associate someone with an act, or may make improper conclusions on which policy will be based.\textsuperscript{119}

Not all of these are strictly privacy problems. The fact that someone was not found by a crawler is surely a good thing for their privacy, but may be bad for public policy. At the same time, invading peoples’ privacy imperfectly leaves open the possibility that action will be taken against people who, in truth, should be left to lead their lives in peace. Controls need to be put in place based on realistic abilities of web crawlers in finding information to ensure that does not happen.

\textit{C. Data Security}

Collecting large amounts of data makes one a target for hackers and opens the possibility of data leaks. As discussed above, this data can be sensitive and can paint a detailed picture of a person’s life. Government agencies have not yet found practical ways to secure their data, and have publicly failed to do so.\textsuperscript{120} Before they embark on additional data collection initiatives, any actor needs to ensure that it can keep the information it does collect safe. This means strong access controls, employing encryption to protect the data, ensuring that employees practice good ‘cyber hygiene,’ that computers are regularly updated, and that steps are taken against unauthorized outside access.


Just as internet users probably do not expect their data to be collected and used for government purposes when they post on Facebook or one of the many forums that exist, they certainly do not expect their data to be used in the future for purposes not yet imagined. Data storage is increasingly inexpensive and allows for the long-term storage, and therefore the long-term use, of collected data.\textsuperscript{121} While many of the things that people post online fade in their ability to cause embarrassment or harm with age, many do not, and some may in fact end up more potent in that regard.

If organizations are to collect data with web crawlers, even in a limited scheme, it must consider whether it plans to maximize the amount of data it collects, over-collecting and storing indefinitely, or minimize its data, discarding it as it is used or after a given time period, during which it is put to no use. Data should, in all cases, be minimized to protect the privacy of internet users, who should not have to worry that decades after posting, their youthful indiscretions will haunt them because a government crawler saved a post.

\textbf{E. Unfair or False Light, Undue Harm, and False Positives}

Related to some of the other concerns listed here, data could be used to paint some internet users in an unfair or false light. Failing to fully collect data about people, or using only part of the data collected, could make a person look bad for failure to consider context or the full picture. This sort of risk can be reduced by controlling how data will be used, who has access to it, and how long it is kept. Use of this data could cause severe harm to some internet users, and may point a guilty finger at innocent users. Organizations employing web crawlers to collect data should consider what level of certainty is required before they can employ their data. There should also be procedural hurdles before such data

\textsuperscript{121} Lucas Mearian, \textit{CW@50: Data Storage Goes from $1M to 2 Cents per Gigabyte (+Video)}, \textit{ComputerWorld} (Mar. 23, 2017), https://www.computerworld.com/article/3182207/data-storage/cw50-data-storage-goes-from-1m-to-2-cents-per-gigabyte.html (noting that from the year 2000 to 2017, the cost of a gigabyte stored on a disk drive has dropped from $7.70 to $0.02).
can be used; just as the criminal justice system is governed by proof beyond a reasonable doubt, programs using data from web crawlers need similar, if less lofty, standards governing their actions.

F. Misuse of Data

There is also the possibility of deliberate misuse of data. Individual employees may use their resources to further their own ends, or simply for entertainment. Proper access controls and good security can significantly reduce the risk of this and protect internet users swept up by web crawlers from significant embarrassment and possibly serious harm.

G. Vulnerable Populations

Many vulnerable, hidden, or marginalized populations use the online technologies to communicate to find support. Sometimes this is done in the open on Twitter, in forums, or through other clients that keep records of their discussions on the open, searchable web. Government agencies may decide some of these populations need to be watched, either for their own safety or the safety of others. This could do significant damage to such communities, causing them to disband after discovering they are under surveillance, or subjecting them to discrimination because of what is found in discussions they never intended for outsiders.

H. Chilling Speech

Finally, government surveillance can have the effect of chilling speech. Those who know the government is crawling the web to record conversations, metadata, and other information may choose not to have conversations or not to go online in the first place. This has significant social costs, and the government should consider the public, civic, and social goods that the internet fosters before it takes actions that could hinder those acts that make the internet so

---

122 See e.g., UNHCR, Connectivity for Refugees, www.unhcr.org/innovation/connectivity-for-refugees/ (last visited May 8, 2018); see also THE ECONOMIST, Phones are now Indispensable for Refugees, Feb. 11, 2017.
valuable.

**VIII. HOW TO TREAT ROBOTS ONLINE**

The internet is undoubtedly an open place that users should be able to surf free of fear from legal action over trespass from website operators with extreme ToS or other usage controls. However, the widespread use of web crawlers to collect information may confound the expectations of many internet users who do not have full knowledge of how the internet works and what bots are capable of. People may understand that their comments will persist, and may be linked to their identity, but the abilities enabled by bots go beyond the risk that a stray comment or account will be linked to a real identity.

Internet users take part in online communities with expectations as to how those communities operate and how their contributions will be maintained. They largely assume that humans and the service they are using will read their posts and review their activity, not some outside party. Website owners also have expectations that they will be able to monetize the data they collect, and that data will not be taken without compensation.

Web crawlers confound these expectations by giving anyone the ability, with relatively few resources, to collect huge amounts of information posted online. While this may threaten business models, it also threatens the assumption of relative obscurity that many users depend on when they partake in online forums. The scale on which robots, and not humans, can collect information, is the relevant consideration in determining whether websites should be allowed to control access by robots.

Web crawlers may require different handling. Website owners should be able to count on robots.txt to guide robots that access their webpages. This would allow website owners to make it clear which pages robots can access and perhaps, how often, and is a clear line for courts trying to apply trespass or other authorized access laws to the internet.

The analysis is not entirely dissimilar from the analysis applied

---

123 Kerr, supra note 87, at 1162.
by the court in hiQ v. LinkedIn.\(^{124}\) While the court there proposed that the situation is more similar to a shop that has “displayed a sign in its storefront window visible to all on a public street and sidewalk,” where “it could not ban an individual from looking at the sign and subject such person to trespass for violating such a ban,”\(^{125}\) the analogy ignores the fact that online, one cannot look at a shop without entering it. A more apt analogy may be if someone walked into that same shop with a scanner, and saved digital copies of its wares for later reproduction and use. Nevertheless, robots.txt could be seen as analogical to a shop owner restricting the manner and scope of access to a physical store.

Enabling website owners to undertake civil actions for violations of their robots.txt restrictions acts similarly to trespass norms; owners can decide who is allowed on to their property, and for what purposes. This solution is not perfect for a number of reasons. It leaves owners of websites in charge of determining and enforcing the wishes of their users, and leaves some web crawler users who people might want to allow to have their information, such as researchers, without that access. This can occur in cases where website owners are indiscriminate in their rulemakings or limit access by corporate entities that publish databases used by researchers. Limiting the rules specifically to bots also addresses some of the possible negative outcomes of applying the CFAA to scraping that the court noted in hiQ —namely, consequences ranging from racial or gender discrimination to illiberal political outcomes.\(^{126}\)

However, owners of websites are far more likely to be responsive to users’ wishes than the more detached third parties operating web crawlers. Additionally, those who want access to the information currently gathered with web crawlers can negotiate for it, something that already happens with many websites like Twitter.\(^ {127}\) This leaves website owners in control of who can gather

\(^{124}\) See hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099 (N.D. Cal. 2017).
\(^{125}\) Id. at 1112–13.
\(^{126}\) Id. at 1110.
\(^{127}\) See, e.g., Barry Schwartz, Google Confirms New Experiment with Twitter in Search Results, SEARCH ENGINE LAND (May 4, 2015),
the information on their websites and users relatively sure that third parties will not scrape their data, so they can continue to use the websites of their choice for the purposes they intend.

CONCLUSION

The idea that any and all information on the web is openly accessible and available and therefore can be freely crawled and scraped is wrongheaded. This article demonstrates that actors engaged in these practices across sectors should be aware of the legal factors that discourage crawling and scraping websites for large amounts of data, and the ethical and social factors that argue in favor of close control of crawling in some cases.

Clearly establishing and strengthening legal rules and accountability mechanisms that regulate government, the private sector, academia, and individuals is necessary. The CFAA and trespass doctrine may operate to keep any type of actor from crawling a website and gathering information, but the application of those laws to the internet is unclear, and it can be difficult for the crawled, particularly smaller institutions, to protect themselves under those laws. The government may be further bound by the Fourth Amendment, though the judiciary has yet to make it clear how the Third-Party Doctrine and aggregation principle should bear on the Fourth Amendment in the electronic world and on the internet. Even academia is bound by relatively lax rules, governed only by IRBs.

Without stronger rules and greater accountability, internet users are left open to severe privacy invasions. Their blogs, Facebook and Twitter pages, reviews, photos, discussions on forums can all be scraped, saved, analyzed, and used later for purposes and by people that the users never intended. Though many actors have some rules self-governing their use of crawlers, the rules as a whole are too weak, and holding them accountable is too difficult.

This article presented a number of issues that need to be considered when updating the existing rules governing online surveillance using web crawlers. These issues need to be considered

in writing these new rules. Failing to consider them could result in laws that continue to protect a too-narrow view of privacy, or that fail to prevent all the harms that could befall internet users.
ABSTRACT

Smart contracts are an emerging technology that could revolutionize commercial transactions by eliminating inefficiencies and uncertainty created by the current transactional ecosystem of lawyers, courts, regulators, banks, and other parties with divergent interests. However, a lack of consensus around how smart contracts are implemented, uncertainty regarding enforceability, and scarcity of on point statutes and case law means that a stable legal, commercial and technical smart contract landscape has yet to emerge. The implementation of universal legal, technical and commercial standards and best practices will reduce uncertainty and promote widespread adoption and use of smart contracts.

* Scott McKinney and Rachel Landy are senior associates in the Technology Transactions practice at Wilson Sonsini Goodrich & Rosati, P.C. (WSGR). Scott is an adjunct professor at Georgetown Law and a guest lecturer for Cornell University’s Cornell Tech grad program. Rachel represents numerous technology companies on matters relating to intellectual property and commercial contracts. Rachel Wilka is Corporate Counsel at Zillow Group, Inc. and lead counsel for Zillow Rentals, hotpads, Inc., and dotloop, Inc., supporting all product counseling, licensing, commercial partnership, and risk management matters. Rachel was previously a technology transactions associate at Wilson Sonsini Goodrich & Rosati. The views expressed in this article are the authors’ own and do not represent the views of WSGR, Zillow, or any of the authors’ other clients. Thank you to Manja Sachet and Rob Philbrick.
TABLE OF CONTENTS

Introduction ........................................................................................................ 314
I. Background .................................................................................................. 316
   A. How Does a Smart Contract Work? ....................................................... 316
      1. Blockchain ....................................................................................... 316
      2. Types of Blockchain ......................................................................... 319
   B. What is a smart contract? ...................................................................... 321
      1. “Smart” ............................................................................................ 321
      2. Contract .......................................................................................... 322
      3. Self-execution .................................................................................. 323
II. Neither Above nor Below the Law: Legal Issues Facing Adoption of Smart Contracts .................................................. 325
   A. Formation ............................................................................................ 326
   B. Assuming the Risk: Risk Allocation in an Automated World ................. 327
   C. Indemnification .................................................................................... 328
   D. Flexibility ............................................................................................ 329
   E. Enforcement .......................................................................................... 329
   F. State Laws ........................................................................................... 331
   G. Other Considerations .......................................................................... 333
      1. Third party intrusion ......................................................................... 333
      2. Statute of frauds .............................................................................. 334
      3. Regulatory concerns and compliance with laws .............................. 335
      4. Ethical issues in the practice of law .................................................. 335
III. Industry Application/Current State of Smart Contracts ...................... 336
   A. Mutating Contingencies ........................................................................ 337
   B. Measurable Milestones .......................................................................... 338
   C. Multiple Stakeholders .......................................................................... 339
   D. Current Adoption .................................................................................. 340
      1. Banking ............................................................................................. 340
      2. Music ................................................................................................. 342
      3. Real estate ......................................................................................... 343
IV. Universal Adoption of Smart Contracts .................................................. 345
   Conclusion ................................................................................................ 347

INTRODUCTION
Every other day, the terms “smart contract,” “blockchain,” or “cryptocurrency,” make headlines with reports of extreme cryptocurrency crashes, “ pivots” to blockchain, and bold proclamations, such as that “[b]lockchain [will] replace the functions of lawyer[s].”¹ Hyperbole aside, the reality is these early-state technologies have a lot of promise, but have yet to be fully-realized by the commercial and legal worlds.

In this Article, we explore what smart contracts may mean for the law and the future of commercial contracts. Before we answer that question, however, we must first ask: how might a smart contract work in the real world?

Imagine the following: you want to buy a bushel of apples. You live in Uganda, and the best apples in the world are in the State of Washington. The apple seller, who you have never met, speaks English exclusively, but you speak only Swahili. The apple seller uses a different bank than you, and you cannot afford to pay expensive transaction fees charged by credit card companies, money transferors, or banks. You do not trust the apple seller to send the apples, and the apple seller does not trust you to send a check. How can you and the apple seller make this transaction happen?

Smart contracts provide a solution. As it turns out, the apple seller’s apple bushel recognizes its GPS coordinates (enabled, of course, by “internet of things” technology) and can automatically verify (over the internet) if the apple seller sent the apples and when the apples have reached you. A smart contract ensures you would not pay any money until the apples arrive, and also ensures that,

when the apples do arrive, the apple seller automatically receives a pre-verified payment. Both sides win. This is the promise of smart contracts.

Smart contracts are models of legal efficiency, reducing the need for a complex court system to enforce transactions because the contracts themselves are self-enforcing. Cross-border transactions can occur with less risk that either party will need to go to court to enforce performance, since there is more certainty that the counterparty will fulfill its obligations under the contract. Intermediaries in contractual ecosystems (like banks and money transferors) could become obsolete. The potential applications are endless, including in the realms of finance, real estate, oil, music, art, infrastructure, intellectual property, transportation, and countless other industries.

If developed and implemented properly, smart contracts promise simplified and streamlined commercial transactions by eliminating inefficiencies and uncertainty introduced by lawyers, courts, regulators, and parties with divergent interests, and could represent a new frontier of commercial law and transactions.

In Part I, we describe how a smart contract works, including through an overview of the blockchain technology that has driven the popularity of smart contracts. In Part II, we provide an overview of some high-level legal issues with widespread use of smart contracts. Part III includes a discussion of how various industries could implement smart contracts to maximize efficiency. Lastly, in Part IV, we propose a best practices framework for smart contract implementation.

I. BACKGROUND

A. How Does a Smart Contract Work?

1. Blockchain

Smart contracts were formally proposed in 1996, but had been

---

conceptualized in technical legal circles far earlier. Yet, it wasn’t until recently that smart contracts really took off. The reason for the change is blockchain.3 Before blockchain, the idea of smart contracts was stymied by general uncertainty, identity and transaction verification issues, and concerns that transactions would not be secure. Although blockchain is not necessary for smart contracts to function or exist (i.e., all blockchains are not smart contracts, and all smart contracts do not need to be incorporated into or use blockchain), current and near-future implementations of smart contracts are virtually all based on or tied to blockchain technology.4

In the past, before blockchain, both parties to a theoretical “smart contract” transaction would have had to rely on the other party’s computer code and network infrastructure, trusting that both sets of code were identical (and executed in the same way on both sets of computers).5 Blockchain’s distributed ledger characteristics allow code to be embedded into a single, publicly-distributed ledger where there is no need for duplication. Every smart contract user accesses the same smart contract using the same set of code. As we further describe below, this means that blockchain is effectively tamper-proof, which gives smart contract users certainty that the deal will not be changed unilaterally and allows the transaction to be self-enforcing.6

Blockchain was first described by the pseudonymous Satoshi

---

3 While it is important and necessary to describe the technical functionality of blockchain and smart contract technology in some detail, this paper is aimed primarily at analysis of legal and commercial issues, so we have chosen to only describe the smart contract technology at a high level. For more in-depth information on blockchain, see, e.g., Sloane Brakeville & Bhargav Peripa, Blockchain Basics: Introduction to Distributed Ledgers, IBM (Mar. 18, 2018), https://www.ibm.com/developerworks/cloud/library/cl-blockchain-basics-intro-bluemix-trs/. For more information on smart contract technology see, e.g., Manuel Araoz, The Hitchhiker’s Guide to Smart Contracts in The Ethereum, MEDIUM CORPORATION (Oct. 6, 2017), https://blog.zeppelin.solutions/the-hitchhikers-guide-to-smart-contracts-in-ethereum-848f08001f05?gi=3c6fdfeb292.

4 The authors were unable to identify any mainstream or public uses of smart contracts that do not use blockchain as of the date of this paper.

5 See Szabo, supra note 2.

6 Id.
Nakamoto in the now-famous bitcoin white paper. This paper describes blockchain as a progressively increasing list of records or “blocks,” which are each, in turn, linked to the previous block and secured using cryptography. This chain of records can be distributed to, or managed by, a peer-to-peer network, hence the often-used-term “distributed ledger.” Each block includes a timestamp, a unique hash, and transaction data for that block, as well as the entire history of the chain. All of this information. All of these characteristics together allows users of the blockchain to be sure that any block in the chain cannot be retroactively altered, which allows for the facilitation of secure online transactions without the need for banks, payment processors or governments. The security, payment processing, and account tracking and maintenance functions traditionally performed by banks or processors are automated in a distributed and decentralized blockchain environment.

---

10 A hash, or hash function, is a way of mapping any data of any arbitrary size to a number or value (the “hash”) of a fixed size. Hash functions are valuable in quickly and easily assigning unique values to each blockchain while preventing reverse-engineering of the data that was used to generate the hash.
Blockchain is generally thought to have the following characteristics:\(^\text{11}\)

- **Consensus**—all blocks in a chain must agree on a transaction’s validity.
- **Provenance**—participants in the blockchain network can see where a block originated as well as ownership over time.
- **Immutability**—no one can edit a block (or transaction) after it is added to the ledger.
- **Finality**—a single shared ledger provides a singular, trusted source of ownership and transaction history.
- **Decentralization**—the blockchain “ledger” is distributed to many nodes (or users of the blockchain), so the failure of some nodes, or failure of the network is not fatal.

2. **Types of Blockchain**

Today, there are three high-level classes of blockchain. Understanding the differences between them is critical to understanding the potential varieties of smart contracts across industries.

---

Public Blockchains—The most common type of blockchain is public blockchain. Public blockchain is a blockchain that anyone can read, anyone can send transactions to, and for which anyone can participate in the validation process (see above). Public blockchains are generally considered to be fully decentralized. Bitcoin is a public blockchain.

Consortium Blockchains—Consortium blockchains are validated through a pre-selected and specific set of nodes that determine whether a block is verified. So, for example, a specific subset of the nodes on the chain could validate each transaction (as opposed to public blockchains, in which anyone in the world can participate in validation). Consortium blockchains have potential applicability in industries controlled by entrenched gatekeepers, such as the financial industry, and in circumstances in which the event triggering confirmation of transaction completion varies from transaction-to-transaction. Consortium blockchains are usually described as “partially decentralized.”

Private Blockchains—The final type of blockchain, a private blockchain, is one in which transaction execution permissions are controlled by and central to one entity or organization. “Read” permissions for the blockchain can be either public or private, depending on the application. Private blockchains, which are essentially just a new implementation of a traditional private database, which

---

14 Id.
15 See Praveen Jayachandran, supra note 12.
16 Vitalik Buterin, supra note 13.
17 Id.
18 Id.
19 Praveen Jayachandran, supra note 12; Vitalik Buterin, supra note 13.
might be used in one-off smart contracts, or for internal recordkeeping within a company or trade organization.20

To summarize, blockchain allows two or more parties anywhere in the world to enter into a transaction directly with each other while being relatively sure that the transaction is secure, authentic, and unalterable. This transaction can be done whether or not the parties know each other’s true identity and without any third-party facilitation or mediation, and these parties can be relatively sure that the transaction is secure, authentic, and unalterable.

B. What is a smart contract?

To understand smart contracts, we must first understand what makes a contract “smart,” what makes an instrument executed by two or more parties a “contract,” and what it means for obligations under a contract to self-execute.

1. “Smart”

At their base, smart contracts are self-enforcing agreements that exchange promises or consideration between parties based on a transparent set of rules using predefined inputs. Smart contracts’ use of distributed ledger functionality together with automated contractual triggers ensures that transactions are completed in a secure and accurate manner, reducing the need for complex regulation or oversight.21 There are many misconceptions about what makes a contract “smart,” which this Section attempts to clarify.22

---


21 It is important to remember smart contracts do not require blockchain technology to work. A smart contract could, in theory, be implemented any number of ways, and could, for example, be tied to a credit card or bank payment system.

Nick Szabo, who is often credited with coming up with the idea of a smart contract, describes the smart contract as “a set of promises, specified in digital form, including protocols within which the parties perform on those promises.”\(^\text{23}\) In other words, a smart contract is a legal contract that is represented and executed, at least in part, by automated software. Pieces of code, (sometimes referred to as “software agents”)\(^\text{24}\) perform certain tasks when pre-defined and mutually agreed conditions embedded in the smart contract are met.

A smart contract, however, is not actually very “smart.” Smart contracts do not (at least, as of the date of this Article) include artificial intelligence, in that a smart contract does not learn from its actions, modify its behavior to match what is appropriate for the circumstances, understand concepts commonly found in traditional contracts such as materiality or knowledge, adapt to changing environments, or learn from experience.\(^\text{25}\) Although smart contracts can respond to variable contingencies, they cannot (as of the date of this Article) “smartly” implement or change their behavior based on unpredicted circumstances. In fact, it is just the opposite. Smart contracts are purposefully designed to be inflexible.\(^\text{26}\)

2. Contract

At the risk of stating the obvious, a smart contract must actually

\(^\text{23}\) Nick Szabo, supra note 2.  
\(^\text{25}\) See Poole, Mackworth & Goebel, Computational Intelligence: A Logical Approach, 1 (1998).  
\(^\text{26}\) To clarify, a contract is not “smart” merely because it is executed or displayed electronically or via a software platform. Contracts executed electronically by “e-signature,” or negotiated or developed via automated software negotiating tools are not “smart” contracts by virtue of their digital execution or origination. The key factor in deciding whether a contract is “smart” is whether or not the contract is automated. See Clack, C., Bakshi, V. & Braine, L., Smart Contract Templates: Foundations, Design Landscape and Research Directions (Aug. 3, 2016, revised Mar. 15, 2017), http://www.resnovae.org.uk/fccsuclacuk/images/article/sct2016.pdf.
be a contract. That is to say, it must meet the characteristics of being a legally enforceable exchange of promises. Since countless others before us have written at length regarding the defining attributes of an enforceable contract, we will be brief. Like any other contract, to be legally enforceable, a smart contract must have the following attributes:

- offer;
- acceptance;
- consideration;
- intent (or “mutuality of obligation”);
- each party must have capacity to contract; and
- the agreement must be of lawful subject matter.

We discuss the formation of a contract in Part II below. The rest of this Part assumes that a smart contract has been formed in compliance with the applicable legal regime.

3. Self-execution

As noted above, a smart contract is premised on self-execution; i.e., one or more aspects of the contract’s execution are automatic. Smart contracts use blockchain to ensure that once the parties

---

27 It is important to note that many in the smart contracts community would disagree with this statement. Some in the community would argue that smart contracts are _ultra vires_, and that one does not need to ask the question of whether or not smart contracts are legally enforceable contracts under the traditional, legal definition of “contract,” because, from a smart contract purist’s point of view, questioning enforceability and enforcement is irrelevant since the execution of a smart contract happens automatically. Automatic execution would seemingly eliminate the need for enforcement (or analysis of whether a smart contract might be enforceable). We think that, while this sentiment is admirable, it is unrealistic, because it is inevitable that disputes over smart contract enforcement, formation and other issues relating to transactions carried out via smart contracts that cannot be resolved via the smart contracts code will end up in court or arbitration. Therefore, this Article is written with the point of view that it is necessary and appropriate to analyze the enforceability of smart contracts from a traditional perspective.

execute the contract, the transactions contemplated by that contract are accurate and cannot be avoided by any party without the other parties’ consent.\textsuperscript{29} For a technology to automatically determine whether a party has performed, or if a condition has been met, there must be some clear-cut input to the code underlying that technology (that is the “smart contract”), via a connection, usually via data feed, to the world outside the bounds of the blockchain allowing the contract to determine “if X, then Y.”\textsuperscript{30}

If/then statements are one of the most basic building blocks of any computer program and easily portable to smart contract applications. The “X” in an “if X/then Y” statement could be a stock reaching a certain value, and the Y could be a payout from one party to the other. The X could be a good arriving at a location, and the Y could be a lien being automatically released. The X could be a third party providing a verified e-signature, and the Y could be an escrow being released. The possibilities are endless.\textsuperscript{31}

Today, most smart contracts: (a) are relatively simple; (b) do not govern complex contractual relationships; and (c) are comprised of relatively basic if/then statements on top of a blockchain platform (such as Ethereum).\textsuperscript{32} If/then statements often tie the release of funds (the “then”) to the basic fulfillment of an “if” condition.\textsuperscript{33} Going forward, however, smart contracts may not be so simple, and prospective parties will not need to understand programming or blockchain to use one. In fact, the future smart contract could look very much like a traditional paper contract, except that certain parts

\textsuperscript{29} See discussion supra Part I.A.1.
\textsuperscript{31} See infra Part III.
\textsuperscript{32} However, the Ethereum platform and blockchain is built on a Turing complete, or near-Turing complete language, which means that it is technically feasible for even Ethereum-based smart contract’s to include complex, advanced functionality that goes well beyond the simple if/then statements described in this section. See Kyle Wang, Ethereum: Turing Completeness and Rich Statefulness Explained, MEDIUM CORPORATION (July 9, 2017), https://hackernoon.com/ethereum-turing-completeness-and-rich-statefulness-explained-e650db7fc1fb. The authors expect smart contract complexity to evolve quickly over time.
of that contract (e.g., performance obligations) will be automated. Some smart contract terms may be written in plain, semantic English (or whatever conversational language the parties choose to use), but other provisions of that same agreement will be self-executing.\textsuperscript{34} Indemnity payouts, insurance triggers, and various other provisions of the contract could be automated and self-fulfilling, while other provisions remain rooted in the “real world” (i.e., outside of the blockchain). It is important to keep in mind that, for each if/then trigger of a smart contract, that trigger must be tied to a definitive real-world, but automatically (and likely electronically) verifiable input. If a human has to decide whether a condition is met and trigger a result baked into an automated electronic contract, that contract is not truly smart, because, like with all contracts, reasonable (human) minds can differ. As smart contract technology evolves, so too will the breadth of the real-world inputs, the if/then triggers, and commercial applications.

II. NEITHER ABOVE NOR BELOW THE LAW: LEGAL ISSUES FACING ADOPTION OF SMART CONTRACTS

Part II provides an overview of legal issues relating to the use of smart contracts. At the date of this Article, there have been no court cases—at least not in the United States—providing direct guidance on the enforceability of smart contracts, nor is there a fully developed smart contract market with agreed-upon industry-wide standard practices (which often inform legal results).\textsuperscript{35} Without any smart-contract specific guidance, smart contracts are best analyzed under traditional contract principles. Below, we describe some of the key legal issues facing the formation, execution, and enforcement of fully self-executing smart contracts.\textsuperscript{36}

\textsuperscript{35} See infra Part IV.  
\textsuperscript{36} This is as opposed to automated contracts that are ancillary to negotiated traditional contract terms. If any paper is involved, then almost all of the legal risks associated with a smart contract can be addressed during negotiations and drafting.
A. Formation

As briefly discussed in Part I.B.2, any contract analysis must begin by establishing whether a contract exists at all. At the most fundamental level, contract formation requires offer and acceptance.\(^\text{37}\) Offer and acceptance signify both parties have accepted the terms of the agreement.

Historically, acceptance was indicated by conduct or a wet ink signature.\(^\text{38}\) However, in recent years, contract formation has occurred more and more frequently via electronic means. Since Congress enacted the Electronic Signatures in Global and National Commerce Act (“E-Sign Act”) in 2000,\(^\text{39}\) which gives legal effect to electronic signatures, digital acceptance through tools like DocuSign has become routine.

Additionally, many companies (particularly, consumer-facing companies) rely on alternative means of obtaining acceptance to contracts. For example, users of online services are often presented with a box that they must check in order to indicate assent to standard, non-negotiable terms and conditions.\(^\text{40}\) These contracts, and others purporting to be formed by signifying acceptance through action (e.g., “By clicking “register,” you agree to the Terms of Use) have been deemed enforceable when the user has been provided “reasonably conspicuous notice” of contract terms and “manifests assent” to those terms.\(^\text{41}\) Notice can be provided by means of a conspicuous hyperlink with language that calls attention to the

\(^{37}\) See discussion supra Part I.B.2.

\(^{38}\) While contracts may be formed without signatures, a signature authenticates the parties who are responsible for performance under the contract. Sophisticated contracting parties typically require signatures. Some contracts are required by law to be authenticated by the parties. See, e.g., U.C.C. § 2-201.


\(^{40}\) Non-negotiable consumer contracts are also known as “contracts of adhesion.”

\(^{41}\) Specht v. Netscape Communications Corp., 306 F. 3d 17, 33 (2d Cir. 2002). Cf. Nguyen v. Barnes & Noble, 763 F. d 1171 (9th Cir. 2014) (suggesting that a contract may be enforceable if a user had notice of, or manifested assent to the Terms. However, the cases cited in Nguyen suggest that notice is always required. It is the manifestations of assent that may be implied, depending on the circumstances of the notice.).
action being requested: “By checking the box, you hereby agree to the Terms of Service.”

Insofar as smart contracts are contracts (i.e., legal instruments), they will be subject to the same level of scrutiny as traditional contracts when faced with formation disputes. All parties will need notice of the terms of the contract and to undertake an action that indicates affirmative assent to those terms. In a smart contract context, notifying users of the terms of the agreement may involve presenting them with the series of if/then statements that comprise the code base and subsequently obtaining consent through a digital function, such as a check-box or “execute” button that would need to be clicked, with the clicks logged somewhere as evidence in the event of a dispute. So long as the manifestation of assent is automated, and the code is not authorized to begin performance until all parties have indicated assent, formation should not be a significant legal issue for smart contracts. Since it is an established principle that e-signatures, check-boxes, and other digital methods of contract execution can be valid and binding, it is likely courts will make the same determination regarding smart contracts entered into via the same or similar digital or online processes.

B. Assuming the Risk: Risk Allocation in an Automated World

Traditional contracts typically involve a number of provisions that shift risk between parties, such as representations and warranties and indemnification obligations. These provisions determine which party is on the hook for liability associated with certain events. For example, in the software-as-a-service context, the service provider often indemnifies the user for any third-party claims of infringement arising from the user’s use of the platform.42 Similarly, a data licensor may offer to indemnify a licensee for any claims alleging the licensor did not obtain any required consents to transfer the data. Many risk-shifting provisions found in traditional contracts can be obviated in smart contracts. For example, in a traditional contract, one party may negotiate for the other party to

---

carry certain insurance levels and certify as to its solvency. However, in smart contracts, that type of obligation may not be necessary, because a party has more certainty the other party will pay or perform via the contract’s automated functionality. A smart contract could be built to take regular readings of a party’s financial health through plug-ins to bank accounts or credit scores and then suspend activity when balances or scores fall below a certain level.

C. Indemnification

Indemnification is a bargained-for shield against certain losses: if a proscribed “bad thing” happens to one party, the other party will cover the first party’s losses. These “bad things” could be a lawsuit, a data breach, or property damage. Building full indemnification provisions into a smart contract is likely unworkable in the near future because the variables and flexibility that are often included in that type of provision would be difficult to translate into smart contract code. For example, an indemnity provision could be triggered by the filing of a lawsuit against a party. That can be verified by the blockchain through a Pacer (the public court records system) alert. However, the costs that a party would cover – litigation expenses, attorneys’ fees, and so forth will vary based on the claim and the extent of remedies pursued. Those costs therefore cannot be practically listed within the blockchain. Further, it could be difficult for blockchain or smart contracts to correctly identify that the lawsuit filed was related to the contract and subject to the indemnity provision. Additionally, some indemnity obligations do not get triggered until there is a final non-appealable judgment - it is unlikely a contract will know when a party has exhausted all of its appeals.

For users to obtain any meaningful indemnity, they will have to do some negotiation outside of the blockchain. That could be easily accomplished in a private blockchain, where users know each other. However, in a public blockchain, it is unlikely that anonymous users would sit at a table to negotiate indemnities. As an alternative, users

---

could explore insurance policies to provide the coverage they might otherwise get under an indemnity. Or, each party could contribute to an escrow account to cover claims made against other parties.

D. Flexibility

Smart contracts, by their nature, are not intended, or desired, to be flexible. Rather, the goal - immutability and measurability - is the very opposite, unlike traditional contracts, which commonly build in mechanisms for amendments, modifications, or varying standards of performance. Each of these mechanisms assists with risk allocation in different ways. For example, a party may want to be judged by “commercially reasonable efforts,” rather than an absolute standard of performance. Similarly, a party may only want certain actions to occur if they have materially breached the agreement.

Smart contracts are built on the notion there will not be any modifications after contract finalization. As a result, if or when circumstances relevant to the smart contract change, a whole new contract would need to be written.\footnote{One author suggests that a smart contract’s code read off of a natural language version of the contract that can be easily updated and translated into the code to address this issue. See Reggie O’Shields, Smart Contracts: Legal Agreements for the Blockchain, 21 N.C. BANKING INST. 177 (2017).} Similarly, traditional contracts often include concepts of knowledge, materiality, and varying effort levels, all of which are subjective measurements. These standards are not easily translated into a self-executing objective performance mechanism. As a result, parties to a smart contract must get comfortable without these unqualified standards.

E. Enforcement

Traditionally, contracting parties build dispute resolution and enforcement mechanisms into a contract—jurisdiction, venue, alternative dispute resolution mechanisms, etc.\footnote{See LEXOLOGY, Dispute Resolution Clauses and the Importance of Drafting, (May 14, 2010) https://www.lexology.com/library/detail.aspx?g=0ffe4bc1-5c70-4bca-a58d-420f7ea748e8.} In a smart contract, the need to enforce should be reduced, given performance is
automated. However, situations could arise where a party seeks to enforce the contract against the other. For example, in a smart contract that involves automated payment mechanics, if one party closes the bank account from which the payments are drawn, and the other party’s obligations continue to be executed, then that party may seek to enforce the payment obligations.

A primary concept of contract enforcement is that the party seeking to enforce the contract knows who the other party is. In a private blockchain, knowledge of the identity of one’s counterparty will likely be the norm. However, in a public blockchain, the parties may not necessarily know each other beyond usernames. To mitigate the risk of having to track a party down in real life mid-dispute, the parties could build automated third-party verification tools into a smart contract, such as a background check on the other party. The results of the check could be made available to each party so there is full transparency as to who the parties are. This mechanism would also allow location to be recorded such that a lawsuit could actually be served. Note, though, that adding identity verification may discourage some parties from entering into smart contracts, as one of the primary features of and reasons to use blockchain - at least public blockchain – is to put trust in the system and not the individual. As a result, smart contract parties may prefer to default to anonymous, electronic arbitration.

Even if the counterparty’s identity can be determined, his or her location would still need to be known for purposes of determining jurisdiction and effecting personal service in the event of a lawsuit. One way users can smartly contract around the issues with physical presence is to include automatic arbitration in the smart contract that provides for anonymous, online dispute resolution in the case of an issue.

If the counterparty cannot be found, a user may resort to bringing claims against the only truly known entity in the picture – the blockchain or smart contract platform provider. However, a user’s recourse against that entity may be limited by the terms of its

49 See infra Part IV.
contract with the provider. For example, Ethereum provides a number of blockchain applications, including a wallet. The software for the wallet comes pursuant to license agreements for the various software components included in the wallet, which reads in part:

```
EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.50
```

This license unambiguously attempts to disclaim all liability arising out of the use of the software, leaving users with little recourse against Ethereum. Further, some blockchain platforms are open sourced or in the public domain, resulting in no single party to go after.51

### F. State Laws

In addition to issues that may arise out of general contracting

---

50 “LICENSE” available when a download of Ethereum Wallet is initiated (last accessed May 18, 2018).
principles, there is a risk of inconsistent laws developing. For example, in 2017, Nevada and Arizona enacted laws applicable to smart contracts.\textsuperscript{52} These statutes, among other things:

- Clarify that records that are solely in electronic form will be not deemed unenforceable solely by virtue of their media, and further state that where records are legally required to be in writing, an “electronic record” satisfies the law.\textsuperscript{53} “Electronic record” is defined as a “record created, generated, sent, communicated, received or stored by electronic means”\textsuperscript{54} and is intended to include blockchain transactions.\textsuperscript{55}

- Authorize the use of smart contracts.\textsuperscript{56} For example, the Arizona statute states:

A. In any automated transaction, the parties may form a contract by the interaction of:

(1) Electronic agents of the parties, even if no individual was aware of or reviewed the electronic agents’ actions or the resulting terms and agreements.

(2) An electronic agent and an individual who acts on the individual’s own behalf or for another person, including by an interaction in which the individual performs actions that the individual may refuse to perform and in which the individual knows or has reason to know will cause the electronic agent to complete the transaction or performance.

(B) In addition to subsection A, paragraphs 1 and 2, the terms of any contract are determined by the substantive law that applies to that contract.\textsuperscript{57}

In these states, then, smart contracting has been sanctioned. However, until all fifty states have enacted similar legislation, there remains a risk that a contract may be enforceable in one state and not in another. A conservative smart contractor could insist on contracting with parties only in states where smart contracts are recognized, using IP address look-up tools to verify a party’s location.

\textbf{G. Other Considerations}

1. Third party intrusion

A party could also face risk if there is a flaw in the code that generates the contract. In 2017, hackers stole $30 million worth of Ether, the cryptocurrency Ethereum issues.\textsuperscript{58} Hackers accomplished this heist by discovering a vulnerability in the blockchain code, not the blockchain platform or conduct by any particular user.\textsuperscript{59} If there were similar vulnerabilities in a smart contract, the parties would have a difficult time obtaining recourse against the hackers.

First, the hackers would not have privity with the contracting parties, since they are (presumably) not part of the blockchain. Therefore, there would be no contract claim against the hackers.\textsuperscript{60} At best, there could be claims in tort (e.g., conversion and tortious interference), as well as criminal claims, each of which would

\textsuperscript{60} Note, however, that the hackers could still be sued under a variety of other legal theories, such as conversion and, depending on the facts, tortious interference with a contractual relationship.
require catching the hacker in the first place. If the parties had been savvy enough to obtain promises from the code developer as to the security of the code, then the parties could seek damages from the developer of the code itself for breach of contract. Alternatively, the aggrieved party could pursue a theory of negligence, which would be a tenuous theory of recovery at best and would depend on arguing there is a duty that runs from the developer to the user.

This is in contrast to the offline world, where, if a hacker hacked an individual’s bank account, that individual could rely on his or her contractual and fiduciary relationship with the bank (as well as potential protection through the Federal Deposit Insurance Corporation or equivalent institutions outside the U.S.) to make him or her whole. These protections do not exist in the smart contract realm. There is no fiduciary duty between a software platform and its users. And, as noted above, the developers of the platform may not even be identifiable if the platform is open sourced.

One way parties are addressing these risks is to engage auditing firms to confirm the code is written to specification. To the extent a lawyer is involved in the “drafting” of a smart contract and that lawyer is not also fluent in code, the lawyer should engage an auditing firm to avoid potential malpractice claims.

2. Statute of frauds

Certain contracts are required to be in writing under the Uniform Commercial Code principle known as the Statute of

\[\text{\footnotesize \textsuperscript{64} These include contracts for marriage, contracts for the sale of goods where the value is over $500, contracts that cannot be fully performed within one year, and contracts for transfers of land.}\]

\[\text{\footnotesize \textsuperscript{61} Note that the users would likely have a claim against the hackers for conversion, if they are able to figure out who they are.}\]

\[\text{\footnotesize \textsuperscript{62} Definition of fiduciary: https://legaldictionary.net/fiduciary/ (describing the duty of care that characterizes fiduciary relationships)}\]

\[\text{\footnotesize \textsuperscript{63} See, e.g., Be Confident in Your Smart Contract, SOLIDIFIED, https://solidified.io/ (last visited Apr. 15, 2015). Using a platform like Solidified necessitates disclosure of the smart contract to third parties, and so to the extent the smart contract is a private one, the parties should understand that they are both waiving confidentiality.}\]
Frauds.\textsuperscript{65} The Arizona and Nevada statutes make clear smart contracts are to be considered “writings,” but other states may take different approaches. Until there is a unified approach to whether smart contracts constitute writings, parties seeking to enter into contracts governed by the Statute of Frauds should proceed with caution.

3. Regulatory concerns and compliance with laws

Smart contracts have arisen in highly regulated fields, such as banking and data transfers.\textsuperscript{66} Developers coding smart contracts should be cognizant of applicable regulations, such as the European Union’s “right to be forgotten” principles for data transfer, and the United States’ “know your customer” regulations in the banking and anti-money laundering contexts.\textsuperscript{67}

Additionally, there are laws about who a person may contract with.\textsuperscript{68} For example, Americans cannot enter into contracts with ISIS or any other terrorist organizations.\textsuperscript{69} In a public blockchain, it is conceivable that a user could be contracting, knowingly or unknowingly, with an entity that is prohibited by law, and users should be aware of those risks.

4. Ethical issues in the practice of law

Lastly, it is illegal in the United States to practice law without a license. In Washington State, for example, anyone who is not a lawyer is prohibited from practicing law or holding him or herself out as being entitled to practice law.\textsuperscript{70} Washington Court Rules define practicing law as “selection, drafting, or completion of legal

\textsuperscript{65} U.C.C. §2-201.
\textsuperscript{66} See infra, Part III.D.
\textsuperscript{67} See Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González (2014), C-131/12 (holding that European Union privacy law provided individuals with a “right to be forgotten”); Bank Secrecy Act of 1970, Pub L. 91-508.
\textsuperscript{68} In addition, parties should be cognizant of contracting with minors, who may void most contracts until the age of eighteen.
documents or agreements which affect the legal rights of an entity or person(s).”

Similarly, Arizona’s court rules, which have remained unchanged since the passage of its blockchain statute, note the practice of law includes “preparing any document in any medium intended to affect or secure legal rights for a specific person.” Insofar as smart contracts have been given legal effect, then developers coding smart contracts without attorney supervision (and particularly those that hold themselves out as specializing in smart contracts) could be at risk under state laws regulating the practice of law.

III. INDUSTRY APPLICATION/CURRENT STATE OF SMART CONTRACTS

Several industries are already working on developing a framework for a smart contracts ecosystem. These industries often share baseline characteristics, such as:

- An established regulatory standard for conducting transactions, which often provides baseline rules on which one can base smart contract “triggers.” For example, real estate has established norms for collecting money upon the acceptance of an offer and holding the money in escrow for a set period of time before releasing the funds upon closing (i.e. confirmation of a set of conditions).

- A lengthy and/or burdensome contracting process for relatively simple functions. For example, contracts to buy or sell futures in a stock or commodity often start with the terms of a financial intermediary, who then has to find a buyer and a seller willing to accept the terms as-is or negotiate the

---

73 One company, Clause, seeks to find a middle ground. Clause enables contracts (including paper contracts) to be operationalized in a dynamic, automated way and is partnering with law firms to obtain appropriate legal oversight.
74 See discussion infra Part III.D.
terms with a buyer and seller in parallel.

- A multitude of stakeholders. For example, in an oil production environment, there is generally a pumper, a dispatcher, a transporter, a treatment facility, a producer, a buyer, and a guarantor (often a bank or other private insurer).

An industry need not have all, or even one, of the above characteristics to benefit from smart contracts. Each of the above attributes are merely economic drivers that may push early adopters to begin using smart contracts before the technology becomes widely established.

The adoption of smart contract technology has been encouraged by the governmental sector. State governments have begun legislating the use of smart contracts, starting with the recognition that smart contracts can be legally enforced. Specifically, as described above, Arizona recently passed a statute that does not strip a contract of its enforceability solely because it is a smart contract, encouraging technology-sector development in the state. Additionally, Vermont passed a statute that validates the use of blockchain records as records of business.

While early adoption of smart contracts appears to be driven by sectors with regulatory predictability, the industries that stand to benefit the most from the use of smart contracts tend to share certain characteristics. Three common shared characteristics of these industries are: (a) mutating contingencies; (b) measurable milestones; and (c) multiple stakeholders.

A. Mutating Contingencies

A contract having a “mutating contingency” is the idea that the potential outcomes under a contract are not binary, but instead

---


77 Id.

plenary. In simple terms, having mutating contingencies means there are multiple ways performance may be satisfied under a contract, either based on the choices of a party or external circumstances (such as availability of a component or service or changing market pricing). The value, and length, of a written contract is directly correlated to the number of contingencies. For example, a simple in-person sale of an apple for $1 won’t normally involve a contract, because the cost in time of preparing a contract for the sale outweighs the worth of the transaction. In contrast, a sale of hundreds of widgets might have to account for partial deliveries, returns of unsatisfactory widgets, and servicing of widgets after delivery. Hence, the need for a written contract that documents a decision tree of outcomes. The presence of mutating contingencies drives adoption of smart contracts because as long as the inputs can be tied directly to “if/then” statements, a smart contract can automatically facilitate every potential scenario, rather than requiring huge amounts of ink or a multitude of amendments. For example, if only 50 widgets are logged in a system as being delivered when the purchase order called for 100, then payment could easily be automatically reduced so that the buyer only pays for fifty widgets. The more different (yet quantifiable) “if/then” scenarios a business operates pursuant to, the more likely it is to benefit from a smart contract that can automate all of the different contingencies.

B. Measurable Milestones

Another characteristic of industries that could benefit from smart contracts is measurable milestones, i.e., conditions or performance that can be objectively quantified. Unlike mutating contingencies, measurable milestones are tantamount to the current smart contracting practice of relying upon input from outside sources (such as an “oracle”). One of the basic requirements of a blockchain contract is that the parties have to agree in advance to performance conditions, which parties are more likely to do if they

---

view the conditions as objectively measurable.\textsuperscript{80}

For example, the sale of a commodity (e.g., gold) can be easily verified based on a weight and an evaluation of the substance. In contrast, an individual who hires an artist to create a painting is unlikely to agree to make payment upon the delivery of any 10x10 canvas with oil paint. Rather, that individual will want to decide whether it lives up to his or her standards and the specifications provided, which are more than just a measure of the materials involved, and thus is unlikely to agree to an automated verification of worth. As previously discussed, the “if/then” statements that make up the content of a smart contract must be capable of objective measurement.\textsuperscript{81}

\textit{C. Multiple Stakeholders}

Finally, many industries that would benefit from smart contracts, including real estate and banking have numerous stakeholders for typical transactions in those industries.\textsuperscript{82} In other words, it is commonplace for a contract to have more than two signatories, or third parties that are necessary in order to measure performance of the contracting parties.

Traditional contracts often handle multiple stakeholders using reams of paper, lengthy negotiations, and drawn out negotiations to address contingencies among the many parties. For example, in the oil production scenario (where there is often a pumper, a dispatcher, a transporter, a treatment facility, a producer, a buyer, and a guarantor), the supply chain is complex. The pumper extracts the commodity, the volume is verified (often by a third party), the transporter finds transport for the volume and confirms timelines for delivery, and the buyer confirms delivery on-time and at the stated volume to the guarantor. A smart contract would enable all parties to share an interface that both allows to adjustment of deliverables and timelines (with set contingencies for adjustments) and allows third-parties to input confirmations in a way that is immediately


\textsuperscript{81} See discussion supra Part I.B.3.

\textsuperscript{82} See discussion infra Part IV.D.
verified for all other stakeholders. The immutability of blockchain enables each party to rely on the verification that performance is complete.

**D. Current Adoption**

Three industries having the potential to benefit significantly from smart contract adoption are: (a) banking; (b) music licensing; and (c) real estate. Below we discuss how smart contracts could improve relationships and transactions in each industry.

1. **Banking**

   Given the origins of blockchain and the quick, widespread adoption of cryptocurrency, such as bitcoin, it is no surprise that one of the first predicted implementations of smart contracts is in the banking industry.  

   Banking has all the characteristics discussed above, i.e.:

   - **Mutating Contingencies**—Many banking transactions rely on changing price points and dependent values.
   - **Measurable Milestones**—A commodity hitting a specific price point is easily measured and tracked.
   - **Multiple Stakeholders**—Many financial transactions involve at least three parties: a buyer, a seller, and an intermediary such as a bank or investment fund, if not also a separate exchange.

   Banking’s pre-existing technical infrastructure also lends itself

---


to smart contract adoption. For example, high-frequency trading via automated software that trades stocks hundreds of times per day to obtain fractional gains on a high volume of sales requires a technological framework for conducting trades without human intervention for execution of a deal.\textsuperscript{85} The only difference between current automated trading technologies and smart contracts is that a contract involves discrete parties (e.g., a buyer and seller) who have decision power, in contrast with an investment fund that is unilaterally executing decisions to benefit itself.

Banks are testing the smart contract waters. On an industry-wide basis, one goal is to use blockchain technology to track corporate borrowers and share fraud detection activity across banks, subject to know your consumer rules and data use regulation.\textsuperscript{86} In the past year in India, a consortium led by the State Bank of India (“SBI”) known as BankChain has explored different ways to incorporate blockchain technology into bank contracts.\textsuperscript{87} In November 2017, SBI announced it would launch its first test of smart contract technology, starting with non-disclosure agreements, but moving into shared fraudulent activity logs.\textsuperscript{88} In December, BankChain followed-up by announcing that they plan to launch basic ledger functionality for account tracking and other low-risk contract applications in the next few months.\textsuperscript{89}


2. Music

Music also has the hallmarks of a successful smart contract industry. In particular, both licensing and paying for the use of a composition\(^{90}\) by online music services are easily translatable to smart contract technology.

- **Mutating Contingencies**—Different outcomes depend on the use of a song and the rights holder of that song.\(^{91}\) For example, public performances of songs are subject to different royalty schemes than reproductions of songs.\(^{92}\)

- **Measurable Milestones**—Uses of songs by digital music platforms can be objectively verified. The number of downloads on iTunes or streams on Spotify are tracked, and can be used to determine royalty payments.\(^{93}\) Music users have the option to pay royalties that are set by statute, so it is even possible to implement smart contracting for royalty payments without any negotiation over fees. The only requirement for the statutory license is to send a “notice of use” to the copyright owner or the Copyright Office prior to using the composition,\(^{94}\) and then to issue reports (with payments) detailing usage (which may be issued electronically in many instances).

- **Multiple Stakeholders**—The music industry has numerous stakeholders, including record labels, music publishers,

---

\(^{90}\) There are two copyrights in each song. One is in the underlying composition (i.e., the lyrics and sheet music) and the other is in the sound recording (i.e., the audible rendition of the composition). In this section, we discuss the composition only.


\(^{92}\) *Id.*

\(^{93}\) *Id.*

\(^{94}\) The authors do not mean to oversimplify how difficult a process this is for some companies. See, e.g., *Ferrick v. Spotify USA Inc.* www.SpotifyPublishingSettlement.com, GARDEN CITY GROUP LLC, http://www.spotifypublishingsettlement.com/ (last visited Apr. 15, 2018).
songwriters, recording artists, producers and distribution outlets. Given that copyrights are divisible, there may be three or four claimants to a song, each exercising different rights.

If an online music service sends the aforementioned notice of use to the appropriate copyright holder, then the remaining execution of the statutory license could be easily automated. Currently, many online services use intermediaries to help with the administration of the license because of the volume of paperwork involved. If a smart contract were to automate all of that extra work, then both the copyright owners and online services would benefit.

Today, the music industry is already exploring smart contract applications. Companies like Ujo Music are working with creators to automate distribution of recordings (and payments for use), leveraging Ethereum as a platform.95 Choon recently launched a music streaming service and digital payments ecosystem that uses Ethereum smart contracts to pay musicians directly for streams of their music.96

However, critics are still doubtful of the industry’s ability to adopt a smart contract system. The music industry is steeped in custom and without buy-in from all of the stakeholders (particularly, the music publishers and performing rights organizations that make money from the licensing of works and control the necessary ownership data for compositions), there is concern that blockchain will never be able to scale to cover the billions of transactions that occur in the music ecosystem.97

3. Real estate

Lastly, real estate is an industry that we believe is likely to benefit from smart contracts:

• Mutating Contingencies—Real transactions inherently involve the possibility of mutating contingencies, including accepting or rejecting an offer, extended or shortened escrow, methods of resolving issues with the property, and the meeting of closing conditions. 98

• Measurable Milestones—The greatest challenge of a real estate smart contract is the milestones. Often, while closing-conditions are written as exact and predictable, they depend on the acceptability of an alternative or compromise to a buyer or other events reliant on unpredictable human decision making. 99 However, this obstacle may be surmountable given that by the time a house enters an escrow period, both the buyer and seller are likely invested enough to avoid challenging closing unless the problems with the property/transaction are drastic.

• Multiple Stakeholders—A typical real estate transaction has multiple stakeholders, namely the buyer, the seller, the agents of both, the bank, and potentially home inspectors and contractors.

Real estate lends itself to smart contract deployment due to its ability to potentially incrementally adopt smart contracts, starting with simpler transactions, and evolving to transactions with more complexity. For example, as a starting, straightforward application, in a simple land sale, where the buyer and seller contract to sell the land as-is, a smart contract could verify the size and chain-of-title through government records, and execute the closing and money transfer. As a result, the parties would eliminate the need for extensive title searches and brokers.


The real estate industry is on the edge of deploying smart contracts. An “International Blockchain Real Estate Association” focused on implementing blockchain in real estate formed in 2013, and real-estate blockchain startups are exploring: (a) buying and selling leases; (b) funding real-estate development; and (c) timestamping and verifying legal agreements connected to leasing or purchasing apartments. The National Association of Realtors (“NAR”) invested in organizations considering smart contract implementations. Additionally, in October 2017, the first property transaction using blockchain to facilitate payment and title transfer occurred. However, wide-spread adoption will still likely depend on decisions from NAR and local agencies and multiple listing services, and their willingness to explore smart contract solutions.

IV. UNIVERSAL ADOPTION OF SMART CONTRACTS

The promise of smart contracts is clear, but the creation of ecosystems that support smart contracts is still in its early stages. Until parties are comfortable with absorbing the inherent risks of an automated contract, as discussed in Part II above, and until smart contract technology evolves to allow for more sophisticated implementations, smart contracts have some obvious limitations.

In order to realize the potential of smart contracts, and avoid the

---


103 Id.

legal risks it is important to establish universal smart contracts standards and best practices. As a starting point, we think all users of smart contracts should agree on and adopt the following:

- **Jurisdiction and dispute resolution**—Current court systems and lawyers are not sophisticated enough, speedy enough, or otherwise equipped to adequately enforce smart contract disputes. A special smart contracts dispute resolution body (similar to the American Arbitration Association\(^{105}\) or JAMS\(^{106}\) that can be referenced and embedded into a smart contract should be formed. Parties should agree via smart contract code that this independent body, not governmental courts, has jurisdiction. Dispute resolution of smart contracts could take place digitally online, so that parties in different countries could resolve disputes quickly and efficiently, without having to travel or incur other related expenses.

- **Universal Coding Standards**—A universal smart contracts language and coding standards should be developed and adopted, to prevent coding errors and deceit, and ensure a meeting of the minds. Drafting and coding standards should be adopted with the lay-person in mind. Universal smart contracts code should be open-sourced, so that everyone has equal access, and the equal ability to use standardized, security-audited, community-verified code.

- **A “Legal API” for Smart Contracts**—A universal “API” or set of contractual terms and contract triggers should be developed, using plain language together with the universal coding standards proposed above. A concrete set of rules for various common contractual terms and scenarios (e.g., payment terms, reps and warranties, indemnities, etc.) would go a long way to preventing misunderstandings in smart contract transactions, and, more importantly, would lead to

---


a greatly increased scope of transactions that can be carried out autonomously.

The above recommendations would help create a common framework that users, legal and non-legal, could build on to create norms for this new contractual medium.\textsuperscript{107}

CONCLUSION

Smart contracts have the potential to disrupt the entire commercial and legal transactional landscape. However, entrenched impediments such as transaction-facilitating intermediaries like lawyers, banks, payment processors, commercial courts, and governments are sure to resist the self-executing contract revolution every step of the way. It is up to the legal and technical innovators on the front lines of the intersection of contracts and technology to ensure that a useable, fair, and universally adopted smart contracts standard are implemented, understood and accepted around the world.

\textsuperscript{107} During the finalization of this paper for publication, the IEEE announced its intent to develop “techno-legal” standards for smart contracts, similar to our proposal in this section. See IEEE and The Accord Project Partner to Develop Techno-Legal Standards for Smart Contract Applications, BUSINESS WIRE (Feb. 20, 2018), https://www.businesswire.com/news/home/20180220005076/en/IEEE-Accord-Project-Partner-Develop-Techno-Legal-Standards). We look forward to seeing the evolution and eventual adoption of universal standards for smart contracts, whether via the IEEE and the Accord Project or another standard setting body.