No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service

Casey Fiesler,[1]* Nathan Beard,[2] Brian C. Keegan[1]
[1]Department of Information Science, University of Colorado Boulder
[2]College of Information Studies, University of Maryland

Abstract

Researchers from many different disciplines rely on social media data as a resource. Whereas some platforms explicitly allow data collection, even facilitating it through an API, others explicitly forbid automated or manual collection processes. A current topic of debate within the social computing research community involves the ethical (or even legal) implications of collecting data in ways that violate Terms of Service (TOS). Using a sample of TOS from over one hundred social media sites from around the world, we analyze TOS language and content in order to better understand the landscape of prohibitions on this practice. Our findings show that though these provisions are very common, they are also ambiguous, inconsistent, and lack context. By considering our analysis of the nature of these provisions alongside legal and ethical analysis, we propose that ethical decision-making for data collection should extend beyond TOS and consider contextual factors of the data source and research.

Introduction

Social media and other user-generated content platforms have opened up a wealth of publicly available information about human behavior. This "data gold mine" has proven to be a great resource for researchers in many disciplines beyond just social computing (Felt 2016; Lazer et al. 2009). Twitter data has supported many kinds of important research–from disease tracking (Paul and Dredze 2011) to communication during crisis (Vieweg et al. 2010) to understanding the flow of misinformation (Starbird 2017).

One reason that Twitter has become the "model organism" of social computing research is that the data is easy to obtain through Twitter's API (Tufekci 2014). However, other social media platforms may not have such easy access, and therefore the researcher must collect data through other means, whether manual or automated. Data scraping is a common method, in which bits of code make it possible to automatically capture large quantities of data from online platforms. It is these large quantities that have offered unique opportunities for digital social research, with new ways of collecting, analyzing, and visualizing data; it also allows for ordered collection, so that messy online data can become usable, well-ordered data sets (Marres and Weltevrede 2013).

However, even when data collection is possible technically, sometimes it is prohibited by terms of service (TOS), which restrict certain behaviors and uses of a site. Whether it is permissible, or ethical, for researchers to violate TOS in the course of collecting data is currently an open question within the social computing research community (Vaccaro et al. 2015; Vitak, Shilton, and Ashktorab 2016).

The lack of clear norms for this specific issue highlights a bigger picture around the rocky relationship between social science, computer science, and existing research ethics infrastructures, which is that we still do not have conclusions about what constitutes human subjects research in the context of big data (Metcalf and Crawford 2016). Even the institutions tasked with providing ethical guidance for researchers have inconsistent policies born in part from a lack of confidence in their understanding of the landscape of new forms of data collection (Vitak et al. 2017). In the absence of clear norms, this lack of clarity can lead to worry and even chilling effects for some researchers; concern over TOS violations is a particularly striking example because of the potential legal risk that could put others (for example, students) at risk (Weller and Kinder-Kurlanda 2017).

The question of ethics rather than legality is also complicated by the tension that sometimes exists between principles of good scientific research (openness, transparency, reproducibility) and respect for legal constraints–particularly when those constraints are set by a social media company that may have an agenda that seems contrary to the values of outsiders (Weller and Kinder-Kurlanda 2017). Indeed, fundamental ideals that are meant to frame how we think about research ethics, such as justice, may be at odds with the existing power structures between platforms, researchers, and users (Hoffmann and Jones 2016). Increasing awareness of the significant impact social media has on society, coupled with high profile controversies around platforms' treatment of data, have led to calls for companies like Facebook to allow more access to data for researchers (Metcalf and Fiesler 2018). Therefore, debates over the permissibility of collect-

*Casey Fiesler is the corresponding author, casey.fiesler@colorado.edu

ing data without permission involve more complex issues than simply breaking a contract or violating a law.

These debates also rarely involve consideration for the content of the terms themselves—not surprising, considering that TOS and other online policies are both rarely read and difficult to understand (Fiesler, Lampe, and Bruckman 2016; Reidenberg et al. 2015). Therefore, we consider: what are the terms that researchers might be violating? How are they framed, and do they even make sense in the context of research? What might the nature of these provisions, taken together with legal precedent and existing literature, suggest about the ethical implications of breaking TOS?

This paper describes the current landscape of data collection provisions in social media TOS, beginning with a description of relevant regulatory background and ethical debates, followed by the results from an analysis of specific provisions from 116 social media sites. Our analysis reveals that though these provisions are very common, they are also ambiguous, inconsistent, and lack context. We conclude with a proposal that ethical decision-making for data collection should extend beyond TOS and consider contextual factors of the data source and research. A critical part of ethical decision-making is learning to *ask the right questions* (Franzke et al. 2019), and we hope that a better understanding of the nature of data scraping TOS provisions, as well as an understanding of the legal background and different ethical approaches, will help researchers during this process.

Unpacking Existing Regulation

Terms of Service in the Law

Absent specific relevant laws, the relationship between a user and a service provider is largely defined by contract law in the United States. Online, this comes in the form of a user "clicking to agree" to TOS or end-user license agreements (EULA) when they access either free or paid online services. Unlike traditional contracts, however, there is no negotiation between the user and the provider; TOS are a "take it or leave it" proposition. Users do not even have to explicitly agree to them in order for them to be valid. Like shrink-wrapped software agreements which courts have upheld despite users not being able to read them until after purchase, "browsewrap" agreements that passively list terms without an active "click-through" may also be valid (Bagley and Brown 2015). Despite some reasonable arguments and jurisdictional splits about the enforceability of some TOS, most courts have had little difficulty enforcing these contracts. The enforceability of TOS has also been key to determining other types of liability, such as unfair competition, trademark infringement, and fraud (Tasker and Pakcyk 2008).

Regarding data collection, arguably more relevant than contract law are laws against "unauthorized access." As is often the case with legal precedent for early technology, the laws applied to scraping (as well as hacking) look a lot like laws that originally dealt with analog situations–for example, "trespass to chattels" governs interference with someone's property that results in harm to that property. The binding nature of TOS can support a claim of trespass to chattel, on the grounds that a particular use of the site violated provisions that restricted use (Tasker and Pakcyk 2008). An early scraping-related court case was eBay v. Bidder's Edge, which relied on a theory of trespass when preventing automated data collection from eBay (N.D. Cal. 2000).

Violating TOS could also be considered using a site beyond the scope of permissive use, and therefore constitute a violation of the Computer Fraud and Abuse Act (CFAA) (18 U.S.C. §1030). For example, in 2009, there was an attempted prosecution under the CFFA for the instigator in a cyberbullying case that resulted in a teen suicide; she had violated MySpace's TOS by creating a fake account. Despite several circuits specifically refusing to expand the reach of the statute in this way, prosecutors have still used it to pile on charges to arguably minor crimes (Constant 2013). A judge pointed out how ludicrous it was to make violations of TOS a crime, that "describing yourself as tall dark and handsome when you're actually short and homely could earn you an orange jumpsuit" (United States v. Nosal 9th Cir 2012). Another recipient of this overreach was Aaron Swartz, who used a script to automate downloading articles from JSTOR, a subscription-only provider of access to academic articles. Swartz was prosecuted under the CFAA for illegally accessing the service, and he later committed suicide. There have since been legislative attempts to amend the CFAA to codify that a TOS violation does not fall under its jurisdiction, but these attempts have fallen short (Constant 2013).

A recent, highly relevant case for this question concerns HiQ, a "talent management algorithm" company; it scrapes public data from LinkedIn and sells reports to employers about employees that may be job searching. As part of a larger set of claims, LinkedIn stated that HiQ violated the CFAA. The district court ruled against LinkedIn, partially on the grounds that publishing a website implicitly gives the public permission to access it. The court also pointed out the problem of unintended consequences, that such an interpretation of the CFAA "would not leave any room for the consideration of either a website owner's reasons for denying authorization or an individual's possible justification for ignoring such a denial" (HiQ Labs Inc. v. LinkedIn Co. N.D. Cal. 2017). In other words, the court expressed concern that the reason for a website denying access is irrelevant.

The U.S. government has also never prosecuted anyone under the CFAA for violating TOS for the purposes of conducting research or journalism, though in 2016 the ACLU filed a lawsuit that sought to remove the barrier of the CFAA for certain kinds of research (Bhandari and Goodman 2017). The lawsuit raises constitutional claims on the grounds that this barrier prevents algorithmic audits and other research designed to uncover discrimination. As of January 2020, this case is still making its way through the courts but survived a motion to dismiss in March 2018, largely on first amendment grounds (Sandvig v. Sessions D.C. Cir. 2018).

In short, it is an unsettled question as to whether it is explicitly illegal (or even a criminal act) to violate TOS. Though some legal interpretations (including some recent case law) suggest that the CFAA may not apply to scraping public data regardless of TOS provisions, there is still enough uncertainty to make risk aversion a reasonable reaction, particularly when students might be involved. Ad-

ditionally, legal issues are complicated across geographical boundaries, and researchers in the European Union or China might face additional challenges in the context of data collection (Halavais 2019).

Terms of Service in Practice

Beyond the potential for legal consequences (either civil or criminal), platforms still have the authority to enforce their TOS however they like—for example, by banning a user account from the platform. However, whether and how any particular TOS provision might be enforced is typically opaque to users (Blackwell et al. 2017).

Though with respect to both platform enforcement and legal enforcement, even if the consequences are clear, the actual rules may not be. A large body of research has confirmed that online terms and conditions such as TOS, EULAs, and privacy policies are highly complex to the point of being unreadable and in many cases practically incomprehensible (Fiesler, Lampe, and Bruckman 2016; Luger, Moran, and Rodden 2013; Reidenberg et al. 2015). One study of privacy policies showed that even legal experts have different interpretations of their meanings (Reidenberg et al. 2015). Moreover, beyond readability, design issues such as "too long" and "small font" make policies less accessible to users (Good et al. 2005). Given these problems, it is unsurprising that research has also confirmed that very few people read online policies (Böhme and Köpsell 2010; Fiesler, Lampe, and Bruckman 2016); this phenomenon is almost certainly due in part to habituation precipitated by experience with their incomprehensibility (Böhme and Köpsell 2010). However, this should not suggest that users do not care about the content policies. Prior work shows a disconnect between users' expectations and the actual policies (Good et al. 2005), including instances in which particularly objectionable TOS provisions come as a surprise to users (Fiesler, Lampe, and Bruckman 2016). Moreover, TOS are highly inconsistent across platforms (Fiesler, Lampe, and Bruckman 2016), so if a user takes the time to read and understand one, this may not scaffold knowledge of others.

Research has also shown that, regardless of this lack of understanding, many users still feel legally and morally bound by TOS and related documents—regardless of whether they are presented as legally binding (Wilkinson-Ryan 2017). Simply the appearance of a legal-looking document is enough to provoke this reaction, even if a user does not agree to any terms. Researchers currently disagree about the ethics of violating TOS (Vitak, Shilton, and Ashktorab 2016), but this finding suggests that some may feel morally bound regardless of the legal legitimacy of the document.

Regulation Beyond Terms of Service

Regardless of the legality of TOS, there are other regulatory concerns regarding online data collection. For example, scrapers may place a load on the servers being accessed, which could be in contradiction to expected use, and possibly even in violation of other rights of the site owners. Typically, visitors to a website are desirable; however, site hits from a researcher's scraper are consuming resources without the typical benefits of a site visit, a form of free-riding (Allen, Burk, and Ess 2008).

There is also the question of the impact not on the platform but on the users. Internet users contribute to robust datasets just by engaging in their everyday behavior, like posting to social media or even clicking on links, and once they consent to the platform's use of their data via TOS, they also effectively give away their data to the derivative-data market (Bagley and Brown 2015). However, it is important to remember that real people do not consider harms in terms of law; violations of social norms that are in no way illegal (or even against accepted research norms) can still be perceived as unwanted or "creepy" (Tene and Polonetsky 2013).

Another important regulatory structure with respect to research practices are ethical review bodies such as (in the United States) Institutional Review Boards (IRBs). A recent study of how IRBs regulate social computing research practices revealed that there is a great deal of variance across different IRBs (Vitak et al. 2017). Similar to legal doctrines, the goals of researchers and IRBs may often be misaligned. Only one third of IRB respondents thought that studies involving scraping web data should be evaluated by IRBs, and only five percent thought that scraping should require informed consent prior to data collection (Vitak et al. 2017). Moreover, norms and regulations for research ethics differ across cultures; for example, U.S. IRBs take a utilitarian approach, weighing harm and benefit, whereas Europeans have traditionally been more insistent on informed consent regardless of cost (Allen, Burk, and Ess 2008).

Professional organizations also often have rules about ethical conduct. A 2018 revision of the ACM Code of Professional Ethics[1] (that governs computing professionals) removed a reminder that "violation of... terms of license agreements is prohibited by law in most circumstances" which many researchers had interpreted as a strong statement that violating TOS was a violation of this code (Vaccaro et al. 2015). The new version, changed for the first time since 1992, only mentions license agreements in the context of "respect[ing] the work required to produce new ideas, inventions, creative works, and computing artifacts," stating that computing professionals should "provide appropriate credit in the form of respecting... license agreements." This change suggests clarification of the intention behind the licensing statement: it is not necessarily meant to apply to terms and conditions broadly. Language around "overriding public good" also remains in the context of unauthorized access, provoking individual ethical judgment.

Another type of regulation is social norms, which can play a large role in individual ethical judgments. Within the research community itself, as with IRBs, it is a matter of dispute whether data collection against TOS is unethical (Vitak, Shilton, and Ashktorab 2016). Arguments against TOS violations also go to the ethics of breaking the law, or of putting a burden on a site's servers, or to potential harm to users. Some researchers may choose to seek consent from any users whose content is quoted in a paper or otherwise identifiably shared (Flicker, Haans, and Skinner 2004).

[1] https://www.acm.org/code-of-ethics

Aside from the issue of TOS violations, the most common ethical heuristic among researchers for data collection appears to be whether or not the data is "public" (Vitak, Shilton, and Ashktorab 2016; Zimmer 2010). It is important to recognize, however, that there is no clearly accepted definition of "public" data. Sharing content publicly does not mean that someone has no expectations of privacy–both the type of content and the kind of use are relevant contextual factors. For example, a single piece of content from an individual is not the same, in terms of ethical implications, as collecting a user's entire social media history (Zook et al. 2017). As Hartzog notes in a legal analysis of the concept of "public" information, labeling something "public" essentially serves as a permission slip for surveillance or data collection, but that "the no privacy in public' justification is misguided because nobody even knows what public' means" (Hartzog 2018). There are not even clear research ethics norms on this point; for example, is a scraper spoofing a logged-in user (e.g., OKCupid user profiles (Zimmer 2016)) collecting "public" data or not? The Common Rule suggests that public data cannot cause any further harm to the individual (Metcalf and Crawford 2016), but is that actually true? Much more so than the law, ethical judgments are highly individual and context-dependent, which makes it no surprise that there are even fewer answers here about permissibility than there was in an analysis of the law.

Data Collection and Analysis

In order to understand the landscape of scraping-related TOS provisions across a broad sample of different social media sites, we began with Wikipedia's list of social networking websites, of which there were 165 as of October 2017. Though this is an imperfect method (for example, it leaves out some sites common to social computing research, such as Reddit), it was a systematic sampling method rather than relying on our own judgment about what constitutes social media. We filtered sites based on three criteria: (1) it was still operational; (2) it had a TOS; and (3) it had been at some point part of a published research paper. For the last metric, we searched for each site name on Google Scholar, and kept the site in the dataset if it was mentioned in at least one paper. This inclusion criteria left a list of 116 sites; we retrieved the TOS from each, and used Google Translate for any TOS that were not originally in English. The TOS in our dataset reflect how they appeared in November 2017; it is important to note that given frequent changes to TOS, it is likely that at least some provisions in our dataset have changed since then. The average length of the TOS in our dataset is 5,520 words, which is line with previous collections of social media TOS (Fiesler, Lampe, and Bruckman 2016). The full list of sites is available in Table 1.

Two researchers conducted open qualitative coding on a subset of policies in the dataset in order to gain a general understanding of lexical syntax and structure, and then came together to discuss a set of data collection related provisions. Relying in part on the legal training of the first author, we created a set of keywords and guidelines for identifying data collection provisions. We used keywords combined with manual checks to identify these provisions in our data.

Our definition for a data collection provision was that it concerned whether or not a user or visitor to the site is permitted to collect data–automatically or manually–regardless of purpose or type of data. Based on this definition, our inductive, open coding (Strauss and Corbin 1998) identified emergent data collection provisions from 91 sites in our dataset. We then jointly created a coding scheme that categorized these provisions in four ways: (1) prohibition on automated data collection; (2) prohibition on manual data collection; (3) prohibition on any data collection; and (4) a requirement to obtain permission for data collection. These non-mutually-exclusive categories are described in more detail in our findings. Our open coding also identified emergent categories, such as terms used, and purposes or types of data specified. Two researchers discussed codes and adjudicated disagreements during the coding process. Following the coding process, we identified broader themes that we discuss later in this paper in characterizing the provisions.

Findings

Categorizing Data Collection Provisions

Our iteratively-developed framework for data collection restrictions revealed four primary categories. The distribution of these (non-mutually-exclusive) categories across the sites in our dataset can be seen in Table 1. Note that these categories are broad, and individual categories as noted in Table 1 should not be interpreted for decision-making purposes without the nuance of a specific provision. Moreover, because it is very likely that a number of these provisions have changed since the time of our data collection, though we provide this data to illustrate the variance among provision types, it should not be relied upon as an up-to-date, accurate reflection of these sites' policies.

No automated data collection No automated data collection ("no auto", 63 sites in our data) refers to provisions that explicitly state that there can be no use of bots, computer programs, scripts, scrapers, spiders, crawlers, or any other non-human collection of any data (including user data) from the site. For example, Facebook's TOS stated, "You will not collect users' content or information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our prior permission."

No manual data collection No manual collection of data ("no manual", 14 sites in our data) refers to provisions that explicitly stated that a user (or a visitor) cannot manually collect—or "copy," which could imply copying and pasting—data from the site. For example, Academia's TOS states that one cannot, "Scrape or copy profiles and information of others through any means (including crawlers, browser plugins and add-ons, and any other technology or manual work)." Manual data collection (e.g., copying content by hand) is sometimes seen as a way to bypass violating TOS, but this is not always the case. This type of provision also suggests something about the reasoning behind preventing data collection; it cannot just be due to server load or other direct harm to the site, but suggests a desire to not have the data collected at all.

No data collection No data collection ("no data", 47 sites in our data) refers to provisions that have vague, umbrella statements against data collection, without specifically identifying the method of data collection. For example, a number of TOS simply stated that someone may not "collect and store" information or data. As discussed below, sometimes these provisions will specify a type of data, however, such as "personal data."

Get permission Get permission ("permission", 24 sites in our data) refers to provisions that specify that one must obtain permission before collecting data, manually or automatically, API or not. This permission may be required from either the site, or from users. For example, Delicious's TOS specified that one may not "collect or store any personally identifiable information from the Service from other users of the Service without *their* express permission" (emphasis added). Gaia Online's prohibited "automated means to access Gaia Online for any purpose without *our* express written permission" (emphasis added).

Tools and Terms

Terms like "robot," "crawler," "spider," and "scraper" can be used interchangeably to refer to computationally traversing the web to extract data, since they perform relatively similar tasks (Algiriyage, Jayasena, and Dias 2013). In our data set, the language used to describe automated data collection varied. The terms (allowing for variations, i.e. scrape/scraper/scraping) that appeared in our dataset (along with how many sites that term appeared on) are: automated (45), bot or robot (40), crawl (18), harvest (13), mine (15), scrape (36), script (19), and spider (34).

It is not uncommon for TOS to take a shotgun approach to provisions, listing everything that might possibly be applicable even if a provision ends up overbroad (Fiesler, Lampe, and Bruckman 2016). This could be happening with these provisions as well since the terms often appear together (or in various configurations). The particular list of tools is also one of the factors that drives textual similarity between provisions. For example, Academia, Delicious, Govloop, and Fet-Life all prohibited "spiders, robots, crawlers, dating mining tools or the like."

However, our analysis of the content of provisions does not suggest meaning in these differences. The lists of terms fail to draw meaningful distinctions between these tools, which could then cause confusion about permissible behavior across sites–for example, if one site restricts "spiders" but not "crawlers" or "harvesters" do these suggest any real differences? This is one factor (similar to the difficulty in interpreting "legal mumbo jumbo" in a policy (Good et al. 2005) that could hinder meaningful understanding of these provisions.

Context

For the majority of sites in our data set, scraping provisions are entirely agnostic as to contextual factors. For example, not allowing users to "use any robot, spider, scraper or other automated means to access" a site does not suggest any distinction regarding what purpose they might have for accessing, or what kind of data they might be accessing.

Purpose Only four sites had provisions that mention any kind of purpose-based restriction beyond the implied, such as "to collect data," "to perform any activity," or "to copy." There are two sites (ning and GovLoop) that included a purpose exception for scraping (with very similar language): "except for Internet search engines (e.g., Google) and non-commercial public archives (e.g. archive.org) that comply with our robots.txt file, or 'well-behaved' web services/RSS/Atom clients." This is actually the only explicit exception in our data set, and it is vague–particularly since the provision adds that they "reserve the right to define what we mean by 'well-behaved."

There are not any exceptions for academic research. However, one of the few explicitly prohibited purposes stated in our data set is a restriction on using FetLife "to do any academic or corporate research without the expressed written consent of BitLove." This provision is actually separate from their data collection provision, which also disallows any automated means to collect data. In this case, this provision about research would also arguably restrict research like recruiting participants.

Though some researchers might not appreciate this part of FetLife's TOS, it is unambiguous and implies a thoughtfulness that is absent in most other provisions. Do the rest actually care about academic research, or are their provisions simply trying to keep away aggregators or other potentially harmful uses? It is impossible to say. For the vast majority of provisions, the purpose of data collection (e.g., whether to create a market competitor or whether to answer an important scientific question) is irrelevant.

Data As for type of data, 32 sites had provisions that specify or mention one specific type: personal data or information. In fact, some sites restrict only that type of data collection. A number of sites had variations on this language from travbuddy: "You are not allowed to collect or store personal information about other users." Variations on this include "personal data" and "personally identifiable information."

Restricting the type of data that can be collected, and in particular thinking about personal data, shows a particular contextual sensitivity that the majority of provisions do not have. However, this still may not provide enough instruction for researchers. Even for those few sites that specify some examples of "personal," there is often a caveat for a broader reading–for example, VampireFreaks prohibits collecting "addresses, phone numbers, email addresses, Social Security numbers, credit card numbers, name, likeness, or biography," but clarifies that the list is not limited to those. Similar to the lack of clarity (legally or ethically) around what constitutes "public" data (Hartzog 2018), how would we define what constitutes "personal" data? One reading might be that it means "personally identifiable," but in one study of people's attitudes about researchers' use of their social media content, one concern was whether the content was "embarrassing" (Fiesler and Proferes 2018), a context that goes beyond social security numbers or names.

Even a provision like travbuddy's that speaks to "per-

sonal information about other users" could mean anything from contact information to photos to personal anecdotes. Whereas this kind of provision is helpful in thinking about what the site is actually trying to prevent (i.e., privacy violations as opposed to server load), it is also vague enough that researchers might either be chilled from collecting data, or on the other side, consider it a loophole.

Data Sharing

Many of these provisions restrict not only collecting data but also sharing it. For example, MeetUp warns users not to "distribute any part of our Platform, including any data, or Content of others." Indabamusic users "agree not to share, syndicate, reproduce, or otherwise disseminate the information from another User's Profile." Of course, sharing necessarily implies (and requires) collection, so any prohibition against collecting data would imply prohibiting sharing as well. These ambiguities point toward interesting divergences in how platforms conceive of "content" versus "data." Fundamental affordances like content sharing necessarily involve reproducing a user's content, but in a way that typically preserves some contextual integrity (Nissenbaum 2004), which is distinct from "data" as something that is collected and leaves these contexts. However, like different types of scraping tools, these provisions tend to use words like "content" and "data" interchangeably without a clear distinction between the two.

Though another issue to consider with respect to sharing data is that it becomes an admission of guilt. Arguably, a site could be anonymized in a research paper such that there is no indication that any particular TOS might have been broken. However, sharing datasets used for analysis is a norm or even a requirement in some scientific communities. The tension between following TOS and abiding by the norms or requests of a research community can be challenging, and researchers might be pressured to comply with these requests regardless of TOS (Weller and Kinder-Kurlanda 2017). Similarly, concerns about admitting to TOS violations might also discourage authors from mentioning TOS in methods sections of papers–even if they think they have followed the terms (due to uncertainty/ambiguity).

In sum, our qualitative analysis of the content of data collection provisions reveals that they are ambiguous and largely devoid of context. This understanding of the nature of the provisions themselves informs a broader ethical analysis of what it might mean for a researcher to violate them.

Discussion

The ethics of violating TOS for the purposes of data collection for research has been a hotly debated topic within recent years. However, this debate has largely not taken into account the landscape of the content of these terms themselves. Though their specifics are not relevant to many of the arguments concerning TOS, knowledge of what terms researchers might be violating and how they are framed provides additional context. In this discussion, we consider what the nature of these provisions, taken together with legal precedent and existing literature, suggest about the ethical implications of breaking TOS. However, this analysis will not provide easy answers; as noted in AOIR's ethics guidelines, the kinds of issues raised by internet research are *ethical* problems precisely because there is more than one defensible response, which means that ambiguity, uncertainty, and disagreement are inevitable (Franzke et al. 2019).

Our analysis of data collection provisions across social media TOS shows that the inconsistency (as shown in Table 1) is also paired with ambiguity and vagueness. TOS are often designed to be as broad as possible, and therefore it may not be in the site's best interest to be specific, and to provide context, but as a researcher it makes informed decision-making very difficult. For example, one researcher might have a different interpretation of what constitutes "personal" data than another. Moreover, the majority of data collection provisions are agnostic to the purpose of the data collection or the type of data being collected–which are precisely the kinds of factors that should be relevant for ethical decisions (Fiesler and Proferes 2018).

Our analysis of existing legal precedent revealed that there is also still some ambiguity surrounding the legality of TOS violations. However, it is important to remember that legal and ethical are not the same thing; a TOS violation may be both, or neither, or one or the other. We also argue that though legal decision-making can sometimes be devoid of context, ethical decision-making should never be. Deciding "should I collect data here?" entirely based on a TOS provision (or lack of one) suggests that violating TOS is (a) inherently unethical; and (b) the only reason that data collection could be unethical. The nature of TOS provisions as largely context-agnostic, as revealed by our analysis, illustrate how they might blind us to the relevant ethical issues.

For example, one consequence of purpose-agnostic data collection prevention is the ACLU lawsuit that is working towards allowing TOS violations for the sake of conducting algorithmic audits for identifying discrimination (Bhandari and Goodman 2017). Not only could seeking out discriminatory practices be seen as a good that outweighs harm, but according to the lawsuit, barring it could even be illegal.

One could also imagine other situations in which violating TOS against the wishes of a company could be an ethical act. For example, if research on a platform can only be conducted by researchers explicitly given access to that platform, might that skew scientific discovery, particularly if researchers may be constrained by the company that employs them (Hoffmann and Jones 2016)? Even assuming the best intentions of the company controlling a platform, are they able to be objective in researching the social impact of that platform? This is an important question, considering the amount of power that social media platforms have, both by architecture and by their policies, in shaping rights and behavior online (DeNardis and Hackl 2015).

In other words, there may or may not be ethical considerations on the part of the platform in their decisions about whether or not to restrict data access. TOS themselves can reflect possibly conflicting values; social media platforms are designed to encourage users to share as much as possible, and though the platforms might also care about user privacy, the data they share is incredibly valuable (Acquisti,

Brandimarte, and Loewenstein 2015). What is their incentive to allow access to that data for anyone else? Our analysis revealed that there are a number of sites that disallow manual as well as automatic data collection, which suggests the issue is not with, e.g., server load, but instead with providing the content at all.

The ACM Code of Ethics now states that "individuals and organizations have the right to restrict access to their systems and data so long as the restrictions are consistent with other principles in the Code," which suggests that it could be a violation of the code to restrict access under certain conditions. Imagine that a platform restricted data collection as FetLife does, for the purposes of academic research—but added an exception that it was acceptable only if the research findings are framed to be favorable to the platform. Legally, a TOS can restrict access with (mostly) whatever conditions they like—but would we consider this restriction ethical, and if not, would it then be ethical to violate it?

Allen et al. (2008) also argue that the burden on researchers to examine and understand TOS or other revocations of consent may be cost-prohibitive and undermine research. They also argue that the benefit of research to society likely outweighs minor determinants to a website based on server traffic. They and others contend that these costs could be part of a larger set of reasons for exempting researchers from informed consent requirements for online research (Hudson and Bruckman 2004). However, this particular utilitarian analysis more considers the harms to the platform rather than potential harms to the users of that platform–and both matter in this context.

Though some may draw a hard line ethical judgment dependent on respecting the wishes of a company or not breaking a contract, this is certainly not the only way that judgment might be made. Ethical decision-making is difficult, particularly in situations in which there are competing interests and/or all relevant information may not be available. When weighing harms and benefits, are you considering only study participants, or also society at large? Do you consider the potential harms beyond intent? Moreover, context is important when it comes to considering these harms.

However, just as violating TOS may not be inherently unethical, the lack of a TOS preventing it does not make data collection inherently ethical. A fundamental rule of responsible big data research is "the steadfast recognition that most data represent or impact people" and that we should begin with the assumption that all data are people until proven otherwise (Zook et al. 2017). Real people do not consider harms in terms of law (Tene and Polonetsky 2013). Moreover, as Halavais points out in discussing possible paths towards continued data access, TOS represent a prohibition on data collection from a platform, not from the users who create that data (Halavais 2019).

In fact, though some ethical analyses of social media research have suggested that TOS should be followed because they establish "reasonable expectations" for the users on the site (Gelinas et al. 2017), we know that most users do not read, or understand the implications of, the policies for the platforms they use (Fiesler, Lampe, and Bruckman 2016; Luger, Moran, and Rodden 2013; Reidenberg et al. 2015), which suggests they are not actually informing expectations. Moreover, our analysis illustrates the language in these provisions is ambiguous and devoid of explanation, which suggests it is not necessarily written to be understood.

Twitter expressly states in its privacy policy that the platform "broadly and instantly" disseminates public Twitter content and that "organizations such as universities, public health agencies, and market research firms" may make use of that content. However, Twitter users are broadly unaware of these provisions, and their feelings about the ethics of researchers using their tweets is not dependent on how they perceive policy or even how they perceive "publicness"—but rather, contextual factors such as the nature of the tweets, who is conducting the research, and what the research is about (Fiesler and Proferes 2018).

After all, just because you can do something doesn't mean you should. Even well-intentioned research or application might be seen as harmful by the public, regardless of legality. For example, the Samaritans Radar project allowed Twitter users to (without consent) monitor their friends' content, to be alerted to anything that might suggest suicidal risk. The creators brushed away legal and ethical concerns on the grounds that tweets are public and the app abided by Twitter's TOS and API policies (Eskisabel-Azpiazu, Cerezo-Menendez, and Gayo-Avello 2017). However, after public pressure related to flaws and potential misuses of the platform, the project closed down.

In 2010, the health support site PatientsLikeMe (PLM) discovered a dating mining bot, which was traced back to a marketing firm. PLM sent a cease-and-desist letter in response, but PLM users were alarmed to realize that the reasoning was about protection of the platform's ownership over the content there and not out of concern for users' privacy (Ferguson 2017). In this case, users had consented via TOS for their content to be used for research purposes; the business model of PLM was in part to sell data to researchers. However, many users did not realize that their data would go beyond PLM.

Another well-known controversy highlights the contextual nature of reactions to data use—the revelation that data from millions of Facebook users contributed to psychometric profiles and models that the data firm Cambridge Analytica used in part to help influence voters (Metcalf and Fiesler 2018). The data collection began for academic purposes, and indeed, nearly identical data collection had happened in the past—with consent, and used as part of a research study. It was not just the collection itself that sparked such a public outcry, but what the data was used for, and the fact that users were so broadly unaware of this. Should we be making decisions only about the ethics of how data is collected, or should its use be part of the equation? After all, data that appears public and is ungoverned by substantive consent practices can still cause harm (Zook et al. 2017).

For example, on many platforms, images of faces may be publicly accessible content, with no TOS provisions discouraging their collection. However, contextual norms around privacy likely do not lead to expectations of the collection of these images for purposes such as use in experiments of detection of sexual orientation (Sweeney 2017). It is a rea-

sonable argument that (identifiable) faces in particular might require a higher ethical standard than other types of data, particularly when the research deals with a sensitive topic such as sexuality (Sweeney 2017). Moreover, the creation of gender recognition technology has the potential to do direct harm to the transgender community (Hamidi, Scheuerman, and Branham 2018). These examples show how different types of context are important: how the purpose of the research, who is conducting the research, and the type of data being collected are all relevant to an ethical evaluation of using data. However, our analysis illustrates that for a legal interpretation of whether TOS is being violated, most often none of these things would matter at all.

However, when it comes to ethics, context is critical. Researchers should consider the broader consequences of data collection and analysis practices, but there are also times when it is appropriate to stray from "rules." Even then, one must be cautious and be willing to engage in difficult ethical debates (Zook et al. 2017). It isn't necessarily so much about following rules as it is about being thoughtful and carefully examining each situation contextually. For example, user expectations should be an ethical consideration, but TOS are unlikely to be a proxy for those expectations; therefore, better ways to determine expectations might be seeking guidance from moderators or members of the community (Galbraith 2017). This kind of care does involve more work than would following a clear set of guidelines. However, as Zook et al. conclude, "responsible big data research depends on more than meeting checklists" (Zook et al. 2017).

Conclusion

This work was motivated by the thorny question of whether researchers should be permitted to violate TOS when collecting data. Some readers may find our conclusions frustrating, because they do not include a clear answer to this question. However, ethics by its very nature rarely has right or wrong answers, only those that are culturally and normatively informed. Our goal was to provide an empirically-grounded description of the current landscape of data collection provisions in order to help researchers with making these decisions. Our observations about the inconsistency, ambiguity, and context-agnostic nature of the provisions support the idea that TOS language alone should not be the sole deciding factor of whether collecting data from a given platform could do harm. Contextual factors, such as the purpose of the research or the nature of the collected data, are highly relevant to an ethical determination–but typically irrelevant to whether the collection is legal. We emphasize the importance of researchers making individual ethical judgments based on the specific circumstances of their research rather than relying on a TOS checkbox.

Acknowledgements

This work was funded by NSF award IIS-1704369 as part of the PERVADE (Pervasive Data Ethics for Computational Research) project. Thank you to our colleagues at CU Boulder and PERVADE for their feedback and help, with special thanks to Emanuel Moss and Michael Zimmer.

References

Acquisti, A.; Brandimarte, L.; and Loewenstein, G. 2015. Privacy and human behavior in the age of information. *Science* 347(6221):509–514.

Algiriyage, N.; Jayasena, S.; and Dias, G. 2013. Identification and characterization of crawlers through analysis of web logs. In *Proc. IEEE ICIIS*, 150–155.

Allen, G.; Burk, D.; and Ess, C. 2008. Ethical Approaches to Robotic Data Gathering in Academic Research. *International Journal of Internet Research Ethics* 1(1):9–36.

Bagley, A. W., and Brown, J. S. 2015. Limited Consumer Privacy Protections Against the Layers of Big Data. *Santa Clara High Technology Law Journal* 31(3):483–527.

Bhandari, E., and Goodman, R. 2017. Data journalism and the computer fraud and abuse act: Tips for moving forward in an uncertain landscape. In *Computation + Journalism Symposium*.

Blackwell, L.; Dimond, J.; Schoenebeck, S.; and Lampe, C. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.: CSCW* 1(2).

Böhme, R., and Köpsell, S. 2010. Trained to accept? A field experiment on consent dialogs. *Proc. ACM CHI 2010* 2403–2406.

Constant, S. A. 2013. The Computer Fraud and Abuse Act: A prosecutor's dream and a hacker's worst nightmare. *Tulane Journal of Technology and Intellectual Property* 16:231–248.

DeNardis, L., and Hackl, A. M. 2015. Internet governance by social media platforms. *Telecommunications Policy* 39(9):761–770.

Eskisabel-Azpiazu, A.; Cerezo-Menendez, R.; and Gayo-Avello, D. 2017. An ethical inquiry into youth suicide prevention using social media mining. In Zimmer, M., and Kinder-kurlanda, K., eds., *Internet Research Ethics for the Social Age: New Cases and Challenges*. New York: Peter Lang Publishing. 227–234.

Felt, M. 2016. Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society* 3(1).

Fiesler, C., and Proferes, N. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4(1).

Fiesler, C.; Lampe, C.; and Bruckman, A. S. 2016. Reality and Perception of Copyright Terms of Service for Online Content Creation. In *Proc. ACM CSCW 2016*, 1450–1461.

Flicker, S.; Haans, D.; and Skinner, H. 2004. Ethical dilemmas in research on Internet communities. *Qualitative Health Research* 14(1):124–134.

Franzke, A. S.; Bechmann, A.; Zimmer, M.; and Ess, C. 2019. Internet research: Ethical guidelines 3.0: Association of internet researchers. *Available online: https://aoir.org/reports/ethics3.pdf*.

Galbraith, K. L. 2017. Terms and Conditions May Apply (But Have Little to Do With Ethics). *American Journal of Bioethics* 17(3):21–22.

Gelinas, L.; Pierce, R.; Winkler, S.; Cohen, I. G.; Lynch, H. F.; and Bierer, B. E. 2017. Using social media as a research recruitment tool: Ethical issues and recommendations. *The American Journal of Bioethics* 3:3–14.

Good, N.; Dhamija, R.; Grossklags, J.; Thaw, D.; Aronowitz, S.; Mulligan, D.; and Konstan, J. 2005. Stopping Spyware at the Gate: A User Study of Privacy , Notice and Spyware Anti-Spyware Technology. In *Proc. SOUPS 2005*.

Halavais, A. 2019. Overcoming terms of service: a proposal for ethical distributed research. *Information, Communication & Society* 1–15.

Hamidi, F.; Scheuerman, M. K.; and Branham, S. M. 2018. Gender recognition or gender reductionism? The social implications of automatic gender recognition systems. In *Proc. ACM CHI 2018*.

Hartzog, W. 2018. The Public Information Fallacy. *Boston University Law Review* 98.

Hoffmann, A. L., and Jones, A. 2016. Recasting Justice for Internet and Online Industry Research Ethics. In Zimmer, M., and Kinder-kurlanda, K. E., eds., *Internet Research Ethics for the Social Age: New Cases and Challenges*. Bern, Switzerland: Peter Lang. 3–18.

Hudson, J. M., and Bruckman, A. 2004. Go Away: Participant Objections to Being Studied and the Ethics of Chatroom Research. *The Information Society* 20(2):127–139.

Lazer, D.; Brewer, D.; Christakis, N.; Fowler, J.; and King, G. 2009. Life in the network: the coming age of computational social science. *Science* 323(5915):721–723.

Luger, E.; Moran, S.; and Rodden, T. 2013. Consent for All: Revealing the Hidden Complexity of Terms and Conditions. In *Proc. ACM CHI 2013*, 2687–2696.

Marres, N., and Weltevrede, E. 2013. Scraping the social?: Issues in live social research. *Journal of Cultural Economy* 6(3):313–335.

Metcalf, J., and Crawford, K. 2016. Where are human subjects in big data research? The emerging ethics divide. *Big Data and Society* 3(1).

Metcalf, J., and Fiesler, C. 2018. One way Facebook can stop the next Cambridge Analytica.

Nissenbaum, H. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79:101–139.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proc. AAAI ICWSM 2011*, 265–272.

Reidenberg, J. R.; Breaux, T.; Cranor, L. F.; and French, B. 2015. Disagreeable Privacy Policies: Mismatches Between Meaning and Users' Understanding. *Berkeley Technology Law Journal* 30(1):39–68.

Starbird, K. 2017. Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *Proc. AAAI ICWSM 2017*.

Strauss, A., and Corbin, J. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, CA: SAGE Publications.

Sweeney, P. 2017. Images of faces gleaned from social media in social psychological research on sexual orientation. In Zimmer, M., and Kinder-Kurlanda, K., eds., *Internet Research Ethics for the Social Age: New Cases and Challenges*. New York: Peter Lang Publishing. 287–292.

Tasker, T., and Pakcyk, D. 2008. Cyber-Surfing on the High Seas of Legalese: Law and Technology of Internet Agreements. *Albany Law Journal of Science & Technology* 18(2008):79–149.

Tene, O., and Polonetsky, J. 2013. A Theory of Creepy: Technology, Privacy and Shifting Social Norms. *Yale Journal of Law & Technology* 16(1):1–32.

Tufekci, Z. 2014. Big Data: Pitfalls, Methods and Concepts for an Emergent Field. In *Proc. AAAI ICWSM 2014*.

Vaccaro, K.; Karahalios, K.; Sandvig, C.; Hamilton, K.; and Langbort, C. 2015. Agree or Cancel? Research and Terms of Service Compliance. In *2015 CSCW Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World*.

Vieweg, S.; Hughes, A. L.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proc. ACM CHI 2010*.

Vitak, J.; Proferes, N.; Shilton, K.; and Ashktorab, Z. 2017. Ethics regulation in social computing research examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics* 12(5):372–382.

Vitak, J.; Shilton, K.; and Ashktorab, Z. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *Proc. ACM CSCW 2016*, 941–953.

Weller, K., and Kinder-Kurlanda, K. 2017. To share or not to share? Ethical challenges in sharing social media-based research data. In Zimmer, M., and Kinder-Kurlanda, K., eds., *Internet Research Ethics for the Social Age: New Cases and Challenges*. New York: Peter Lang Publishing. 115–129.

Wilkinson-Ryan, T. 2017. The pervserse consequences of disclosing standard terms. *Cornell Law Review* 103:117–176.

Zimmer, M. 2010. "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology* 12(4):313–325.

Zimmer, M. 2016. OKCupid Study Reveals the Perils of Big Data Science.

Zook, M.; Barocas, S.; danah Boyd; Crawford, K.; Keller, E.; Gangadharan, S. P.; Goodman, A.; Hollander, R.; Koenig, B. A.; Metcalf, J.; Narayanan, A.; Nelson, A.; and Pasquale, F. 2017. Ten simple rules for responsible big data research. *PLoS Computational Biology* 13(3):1–10.

Site	P	ND	NM	NA
About.Me			✓	✓
Academia.edu			✓	✓
aNobii	✓	✓		
AsianAve.com		✓		✓
aSmallWorld		✓	✓	✓
Bebo		✓		
Biip.no				
BlackPlanet				✓
Busuu		✓		
Buzznet		✓		
Cafemom				✓
Care2		✓		
CaringBridge			✓	✓
Classmates	✓		✓	✓
Cloob				✓
CouchSurfing				✓
Crunchyroll		✓		✓
Cyworld				
DailyStrength		✓		
Delicious	✓	✓		✓
DeviantArt		✓		
Diaspora		✓		
Dol2day				
douban				
Doximity	✓		✓	✓
Dreamwidth				
DXY.cn	✓			
English, baby!				
eToro				✓
Facebook	✓			✓
FetLife		✓		✓
FilmAffinity				
Filmow		✓		
Flickr		✓		✓
Flixster	✓	✓	✓	✓
Fotki				
Fotolog		✓		
Foursquare			✓	✓
Gaia Online		✓		✓

Site	P	ND	NM	NA
Gapyear	✓	✓	✓	
Gays.com				✓
Geni.com		✓		✓
Gentlemint				
GirlsAskGuys	✓			✓
Goodreads		✓		✓
Google+				
GovLoop			✓	✓
Grindr				✓
Habbo		✓		
hi5		✓		
Hub Culture				
Ibibo		✓		
Indaba Music	✓			
Influenster		✓		
Instagram				✓
IRC-Galleria	✓			
italki.com	✓			
ItsMy/GameCloud				
Kaixin001				
Kiwibox		✓		✓
Last.fm		✓		✓
LibraryThing	✓			✓
LifeKnot				
LinkedIn	✓	✓	✓	✓
LinkExpats				✓
LiveJournal				✓
Meetup	✓			
mixi				
MocoSpace				
MouthShut				✓
MyHeritage	✓	✓	✓	✓
MyLife				✓
MySpace				✓
Nasza-klasa.pl		✓		✓
Netlog (Twoo)				
Nexopia				
Ning	✓	✓		✓
Partyflock		✓		

Site	P	ND	NM	NA
PatientsLikeMe	✓			✓
Pinterest		✓		✓
Plurk				
Quechup				
Quora	✓			✓
Qzone				✓
Raptr		✓		✓
Ravelry		✓		
Renren				
Reverbnation				✓
RoosterTeeth				
Sgrouples				✓
Sina Weibo	✓			
Skoob				
Skyrock		✓		
SoundCloud				✓
StudiVZ				✓
StumbleUpon	✓			✓
Tagged		✓		
TalkBizNow				✓
TermWiki		✓		
TravBuddy.com		✓		
Travellerspoint	✓	✓	✓	✓
Trombi				✓
Tuenti				
Tumblr			✓	✓
Twitter	✓			✓
Untappd		✓		✓
VampireFreaks				✓
Viadeo	✓			✓
VK		✓		✓
Wattpad		✓		✓
WriteAPrisoner				✓
Xanga		✓		
XING		✓		✓
Xt3				✓
Yammer				
Yelp		✓		✓
Total	**24**	**47**	**14**	**63**

Table 1: Each site in our dataset, along with the categorizations for the data collection provisions in their TOS: **P** – Permission, **ND** – No data, **NM** – No manual, **NA** – No auto. Sites without data collection provisions are highlighted in gray. Note that his analysis was conducted in Fall 2017, and it is very likely that a number of these provisions have since changed; this table is provided to illustrate the variance among provision types, but should not be relied upon as an up-to-date accurate reflection of these sites' policies.