Researcher Views and Practices around Informing, Getting Consent, and Sharing Research Outputs with Social Media Users When Using Their Public Data

Nicholas Proferes
School of Information Science
University of Kentucky
nproferes@uky.edu

Shawn Walker
School of Social and Behavioral Sciences
Arizona State University
shawn.w@asu.edu

Abstract

Publicly accessible social media data is frequently used for scientific research. However, numerous questions remain regarding what ethical obligations researchers have in regard to using such content. We report on researchers' own views and practices regarding informing, getting consent from, and sharing research outputs with users when using publicly accessible social media data. Findings reveal both diverging current practices and views on what researchers ought to do in the future. Some researchers view the ethics of public data use as merely requiring compliance with the requirements of their ethics board, while others' ethical practices go beyond what is minimally required. Some researchers worry about the effects of contacting users to inform, seek consent, or share outputs with users. Yet others note that they want to build bridges with online communities through these mechanisms, but struggle with a lack of precedent and tools to do so at scale.

1. Introduction

Social science researchers use data from sites such as Facebook, Twitter, Reddit, Tumblr, and dozens of others to understand a wide range of online social phenomena. They seek to answer questions ranging from whether sentiment can be used to predict movement in the stock market [1] to whether or not it's possible to predict cardiovascular mortality at the community aggregate level based on post content [2]. Twitter dominates academic social media research. Despite having a lower monthly active user count than Facebook, Instagram, or Snapchat, Twitter has become the "model organism" [3] for research because of its comparatively open and generous APIs. Reliance on data from Twitter has only increased as many other social media platforms increasingly lock-out researchers [4], [5]. Between 2006 and 2012, there were upwards of 350 peer-reviewed publications using data from Twitter [6], and the numbers have only risen since.

There have been numerous conversations within academic communities, such as the Association of Internet Researchers [7], regarding the extent to which questions about research ethics should come into play when using publicly accessible data. Questions of note have included: should social media data be considered to be "human subjects data?" Should researchers quote tweets directly from public data as part of their publications? And, should researchers seek out re-consent from users if the data is already public?

Researchers and institutional ethics boards are struggling with these questions as well. A study of U.S. IRBs by Vitak et al. [8] found "a lack of consensus among IRB staff about what should be reviewed" in regards to research using social media data. Typically, IRBs review research that involves human subjects data, and U.S. federal guidelines (OHRP 45 CFR 46.102) offer a definition of human subjects data that, "means a living individual about whom an investigator (whether professional or student) conducting research obtains (1) Data through intervention or interaction with the individual, or (2) Identifiable private information." However, many IRBs in reported conflicting views on whether or not social media data fit under that purview, as data from sources such as Twitter can be collected without direct intervention with individuals and is typically considered public.

At the other end of the microscope, many social media users are entirely unaware that their data may be used for research [9]. When asked if they thought that researchers were allowed to collect publicly accessible Twitter data, almost half of the participants in Fiesler and Proferes' 2018 [9] study indicated they thought researchers were actually forbidden from doing this without researchers having to ask users for their permission. Social media users also frequently have under-developed understandings or beliefs about

URI: https://hdl.handle.net/10125/64032
978-0-9981331-3-3

the potential uses of their data, which raises questions about the extent to which users are making *informed* choices when they agree to the rules of a platform.

In short, the status-quo is that researchers are using increasing volumes of Twitter data to study various phenomena, IRBs have conflicting views and guidance regarding the necessary ethical behavior for this work, and users are mostly in the dark that this is taking place. Thus, it is important to understand how researchers are themselves interpreting their own ethical obligations and acting upon them. In this work, we seek to better understand how researchers are confronting these questions; what their views are and what ethics-based practices they may engage in outside of formal compliance-driven requirements of IRBs. We also ask researchers about the conflicts they have between ideal ethical practice and the ethical practices they can realize through the tools they have accessible. For example, whether they would inform, seek to consent, or share research findings with the individuals whose publicly accessible data they are collecting if they had tools that could automate such processes at scale.

Ultimately, we find many researchers are engaging in ethical practices beyond the minimum compliance-driven practices required by institutional ethics bodies and are seeking to notify and share findings with the users they study. Many more indicate they want mechanisms and tools to help them communicate with users in these ways. But, other researchers are hesitant, indicating worries or concerns about the implications of contacting users whose data they are using. Some worry about the possibility of creating anxiety among users, about the Hawthorne effect, and about creating new expectations that this is how all research should be done.

2. Review of relevant literature

Social media data – the posts, activities, and trace data from users of social media services [10] — have become an increasingly important source of real-time information of public reactions to events [11]. This data contains not only the 'post' or 'update' itself, but also the URLs, images, videos, and metadata (posting data, user profile information, location, etc.) embedded in or accompanying the post. In order to understand the issues around consent and the "publicness" of social media data, it is important to consider two perspectives: that of social media users (or the participants in research using social media data) and that of the researchers collecting and analyzing social media data.

2.1. The perspective of social media users

The perspective of social media users is grounded in the network of connections and the affordances the social media service provides. The network of connections to other social media users such as friends, family, associates, and communities provides the context within which users interact. The features of a platform carry with it constraints and opportunities for the user known as affordances [12]. For example, one of Twitter's features is the limitation on the length of a user's post. More applicable to questions of data collection by academics, are the features that shape the types of content (text, images, videos) a user can post and the privacy, or visibility, of a user's social media post.

Each social media service offers their users specific options to control the privacy of their account. These range from a basic public or private switch applied to the entire account or a more flexible and advanced array of settings. It is important to note that the majority of social media services set the default privacy setting of posts to be "public" when a user creates an account. As a result, a user's posts will be public unless they take specific action to change the settings. In essence, users must "opt-out" in order to make their posts private, requiring users to be aware of these privacy settings and their defaults. These settings have implications for researchers since most services only allow access to users' public posts, but users may not inherently realize the content they create is public by default.

Additionally, social media platforms such as Twitter do not offer features that let users see who exactly has viewed their content. As a result, while users may imagine their audience to be friends or family, researchers are unlikely to appear as part of their "imagined audience." Further, while users may create tweets in response to a particular event or moment, they often do not anticipate new future uses of their data. For example, when Twitter users were asked about their feelings concerning the newly created archive of every post made to Twitter at the Library of Congress, many expressed surprise and frustration that their tweets might be used in this way [13].

2.2. The perspective of researchers collecting and analyzing social media data

The perspective of researchers collecting and analyzing social media data can be quite different from that of the users of the service, even though researchers may also be users of the service themselves. While

some researchers working with small datasets may collect social media data using the same interface as users by capturing screenshots or posts manually – researchers working large social media datasets often collect data en masse. The process is automated via custom computer programs that connect to social media services via numerous Application Programming Interfaces (APIs) offered by each service. APIs offer interfaces for computer programs to interact with online services, allowing researchers to collect public posts from these services.

When connecting to these APIs, most social media services will only allow the collection of data in real-time, meaning that it is difficult or even impossible to access historical posts. This limitation increases the pressure on researchers to collect or acquire data as soon as a phenomena has occurred (in real-time) as to not lose data; bringing with it a host of methodological challenges [14]–[16]. Researchers often address these methodological challenges by collecting data at the first available moment. The impetus to collect data in real-time does not give researchers the opportunity to request consent before collecting their post.

Since real-time collection of data makes it difficult to seek for consent before collection takes place, this leaves researchers with a large dataset possibly containing millions of posts by millions of users. While these posts are technically considered "public" by the social media service, many users do not alter their default privacy settings. As a result, this data is often considered "public" but may only be public due to the "public-by-default" nature of most social media services.

2.3. Twitter as a site for social science research

The process of collecting social media data, while seemingly simple on the surface, requires numerous competencies [15], [17], [18], both technical and research design related. The process is made more complex as it involves a mixture of theory, data, and computational processes (see Goble 2008 for a bioinformatics perspective) filled with many "black-boxes" [15], [19]. An algorithmic system underlies the multitude of interfaces users and programs use to consume and interact with information from social media platforms. These algorithmic systems [20] are an assemblage of "institutionally situated code, practices, and norms with the power to create, sustain, and signify relationships among people and data through minimally observable, semiautonomous action" [21].

To users and researchers outside of the platform, these algorithmic systems and databases seem like black boxes taking input from a user's action and outputting posts without giving any details of how data is processed or changed. The lack of transparency only adds complexity to the research process since the impact of forces assembling and acting on data are unknown to us. If social media is a black box to researchers, the collection of social media data is an even darker box to users [9].

3. Methods

As part of this study, we sought to elicit researchers' experiences with IRBs when using publicly accessible data, their own ethics review practices, and views on hypothetical situations involving questions of research ethics via an online survey. We developed our survey questions based on conversations taking place within the online research communities in which the authors participate, and based on the work of previous studies mentioned in the review of relevant literature. After developing an initial draft of the survey, we piloted the survey with a small cadre of colleagues and solicited feedback about the question wording. After revising question wording for clarity and receiving IRB approval, we circulated the survey in the method described below.

3.1. Population of Interest

Our primary interest is in researchers who use publicly accessible Twitter data as part of their work. We employed purposive sampling to identify individuals who have worked with Twitter data. We used four methods to recruit participants: 1) we emailed a collection of individuals who had previously published in CSCW, CHI, ICWSM, iConference, WWW, Ubicomp, CKIM, and KDD that mentioned "twitter," "text mining," "logs," "activity traces," and/or "social network" as part of their abstract; 2) we posted survey recruitment materials to a number of academic Facebook groups (such as Researchers of the Socio-Technical); 3) tweeted the recruitment during the 2018 CSCW conference using the conference hashtag; and 4) shared the recruitment on the authors' own Twitter and Facebook timelines.

Though we tried to ensure a range of different researchers had the opportunity to participate in our survey, we still ultimately relied on a convenience sample. A challenge in studying this population is that true random sampling is difficult, and further, recruitment presents bias towards those interested in reflecting on the ethics of their own practices. Those

who are willing to participate in such a conversation may naturally have a different view of research ethics than those who don't, or may be otherwise intrinsically motivated.

4. Data

We received a total of 52 completed surveys. Note that "No Response Given" appears as part our data because all questions were optional, allowing participants to skip any questions they did not wish to answer, but are not included as part of percentage calculations.

In Table 1 we introduce demographic information about our sample first, however, this information was asked at the end of the survey. As seen in Table 1, a majority of participants in our study identified as male. Participants varied in the specific research positions they held, with a fairly even mix of assistant, associate, and full professors, graduate students, post-docs, lecturers and other research positions. Participants most readily identified information science, communications, and computer science as their "home" discipline. Given the sampling methodologies and conferences from which authors were contacted, this representation is unsurprising.

We also asked participants about the types of institutions they work at, where they work (geographically), and the types of ethics review bodies at their home institutions. As seen in Table 2, three-quarters of our respondents work at a public university or college, and a majority are located within the United States. Given the volume of participants that are located in the U.S., it is not surprising that the majority of our participants also indicated that they have an institutional review board (IRB) as an ethics review body at their institution.

The IRB process comes from U.S. federal policies requiring ethics review bodies as a condition of institutions being eligible for federal research grants [22]. Of the small (n = 3) number of individuals who reported having no ethics review body whatsoever at their institution, two of those individuals indicated that they *still* sought out some form of ethics review for their work, despite a lack of a reviewing body.

Next, we asked researchers about the kinds of Twitter data they collect, and the methods by which they collect it. As shown in Table 3, most researchers are working with tweets, but many are also working with linked and embedded content, as well as profile information and trend data. While most researchers are collecting data through scripts (such as twarc), many collect data straight from the Twitter.com website and through Twitter's APIs.

Table 1. Sample demographics

Variable	N	(% of those who responded)
Gender		
Female	17	39.5%
Male	25	58.1%
Genderfluid	1	2.3%
No Response	9	-
Position		
Assistant Professor	7	16.3%
Associate Professor	9	20.9%
Graduate Student	12	27.9%
Lecturer	3	7.0%
Other	1	2.3%
Post-doc	3	7.0%
Full Professor	7	16.3%
Research Scientist	1	2.3%
No Response	9	-
Home Discipline		
Anthropology	2	4.0%
Business	2	4.0%
Communication	10	20.0%
Computer Science	8	16.0%
Education	1	2.0%
Engineering	1	2.0%
Geography	1	2.0%
Information Science	16	32.0%
Media Studies	2	4.0%
Political Science	1	2.0%
Psychology	2	4.0%
Research and Development	1	2.0%
Social Work	1	2.0%
Sociology	2	4.0%
No Response	2	-

Table 2. Institution characteristics

Variable	N	% responded
Institution Type		
For-Profit Organization	1	1.9%
Other	4	7.7%
Private University or College	8	15.4%
Public University or College	39	75.0%
Country of Institution		
Australia	2	3.8%
Brazil	2	3.8%
Canada	2	3.8%
Chile	1	1.9%
Germany	2	3.8%
Ireland	1	1.9%
Italy	1	1.9%
Norway	2	3.8%
Poland	1	1.9%
Singapore	1	1.9%
Sweden	1	1.9%
UK	4	7.7%
USA	32	61.5%
Type of Ethics Body		
No ethics body at my institution, and I don't work with any external ethics review bodies.	1	2.0%
No ethics body at my institution, but I still seek ethics review from outside my institution.	2	3.9%
ERB	2	3.9%
Foundation Review Committee	1	2.0%
IEC	1	2.0%
IRB	32	62.7%
National Ethics Office	1	2.0%
Other Ethics Committee	2	3.9%
REB	9	17.6%
No Response	1	-

When asked if they had ever gone through institutional ethics review for their use of Twitter data, even if their ethics body ultimately decided their research projects were exempt, 21 (43.8%) indicated they had not gone through ethics review when using publicly accessible data, and 27 (56.3%) indicated they had. This is in contrast to the finding from Zimmer & Proferes [6] which found that only four percent of published papers using Twitter data indicated within the publication they had gone through some kind of IRB or ethics review. This data point suggests that while researchers may not be discussing it as part of their publications, many are still going through ethics review.

Table 3. Types and methods of data collected

Variable	N	% selections
Types of Data Collected		
Tweets	49	34.3%
Profile	25	17.5%
Embedded Content	26	18.2%
Linked Content	31	21.7%
Trend Information	11	7.7%
Metadata	1	0.7%
Collection Method		
Twitter.com	21	16.8%
Sprinkler	18	14.4%
Firehose	8	6.4%
Decahose	1	0.8%
Powertrack	7	5.6%
3rd Party Purchase (GNIP, Sifter)	8	6.4%
Script (Python, R)	24	19.2%
Nvivo	4	3.2%
NodeXL	11	8.8%
Google Sheets	11	8.8%
Web Archives	4	3.2%
Full Dataset from Someone Else	6	4.8%
Rehydrated dataset	9	7.2%
Profile RSS	1	0.8%
Talkwalker	1	0.8%
TrISMA	1	0.8%

We additionally asked respondents whether their ethics review bodies had ever required them to get consent from users either prior to collecting their publicly accessible data, or after they had collected data. Only 3 researchers indicated the former, and 1 researcher the latter. When asked if outside of an IRB requiring it, if the researcher sought to consent users anyways, 6 respondents of the 45 that answered this question gave an affirmative response (13.3%). Participants who answered in the affirmative were given the opportunity to discuss why they chose to do this in an open-text box. Several of the respondents indicated they felt obligated to seek consent from

users, even though they were not required to, because they felt ethically obligated if they wanted to include a given tweet directly in the text of their article. As one participant put it, "*If I am using someone's contributions in an identifiable way, I want them to be ok with it, whether or not the IRB think [sic] this is an ethical issue*."

Table 4. Most stringent level of review the researchers have ever experienced from an IRB while using publicly accessible Twitter data

Highest Level of Review	N	% of respondents
Exempted	14	48.3%
Expedited	7	24.1%
Full Review	4	13.8%
Some other level of review	4	13.8%

When we asked if, outside of formal consent, if the researcher provided any other means of informing users about the researcher's use of users' public data, such as Tweeting at them or sharing the research output (such as publications) with the user, 10 respondents of the 46 that answered this question gave an affirmative response (21.7%). Participants who answered affirmatively were given the opportunity to expand on their reasons for doing so. Reasons varied among participants. One noted:

"My country's research data review board has asked that I publicly post an explanation of my project and give users the opportunity to request to see and withdraw their data. (This is new since GDPR.)"

Another stated:

"I have shared a publication about a particular hashtag with the person who created the hashtag (in addition to asking him formal consent). I also reach out to ask about consent by tweeting at people."

Of the individuals who responded that they do follow-up with participants as part of their research practice, most indicated they use Twitter as the medium for follow-up.

Based on past conversations on mailing lists and in Internet research related social media groups we heard some researchers opine that they would, if possible, prefer to inform or consent users when they used their publicly accessible data, however, that it was too difficult or cumbersome to do at scale. When researchers are using millions of tweets as part of a project, contacting this volume of persons becomes a technical impossibility as the platform does not provide the technical affordances to @mention or DM a large volume of users. Therefore, we asked researchers, if there was a hypothetical tool that allowed them to provide notice to users that they were collecting users' publicly shared data before collection or after collection, would they use it? We also asked if the tool could ask users for consent, would they want to consent users (thus giving users the chance to say no or to opt-out) to their use of publicly accessible data; and, if the tool was designed to allow researchers to share research outputs with users, would they want to use it. Table 5 provides a breakdown of responses to these questions.

On the whole, we found that a majority of respondents gave "maybe" responses to the idea of using a tool to provide notice to users before or after data collection, and in terms of using such a tool to get explicit consent from users. However, a clear majority of researchers would be inclined to use such a tool to share their research output with users. As part of these questions, we provided a text box to allow respondents to indicate why they would, might, or would not want to use such a hypothetical tool.

When it came to providing notice (but not consent) to users that their data was being collected or had been collected, over half for both question constructions answered "Maybe," with about a quarter indicating "Yes" they would use a tool to do this, and about a quarter responding "No." Many researchers indicated they had concerns that providing notice to users about collection would "disrupt the data." Comments in this vein included: There are real benefits to notice and consent, but also real costs around risk/disruption both to users and potentially to the data.

I would be concerned that this might alter someone's behavior.

It depends on the nature of collection. For purely historical data, sure. For ongoing collection, maybe not, because we would need to consider whether notice is likely to influence future behavior.

One participant simply wrote, "*Hawthorne Effect.*" According to Wickstrom and Bendix [23], the Hawthorne effect is:

often mentioned as a possible explanation for positive results in intervention studies. It is used to cover many phenomena, not only unwitting confounding of variables under study by the study itself, but also behavioral change due to an awareness of being observed, active compliance with the supposed wishes of researchers because of special attention received, or positive response to the stimulus being introduced.

Table 5. Responses to tool features questions

Question	Yes, n (%)	Maybe, n (%)	No, n (%)	No Response
If there was a tool that allowed you to provide notice (but not consent) to Twitter users that you were collecting their publicly available data before you collected the data, would you use it?	10 (23.3%)	24 (55.8%)	9 (20.9%)	9
If there was a tool that allowed you to provide notice (but not consent) to Twitter users that you were collecting their publicly available data after you collected the data, would you use it?	11 (25.0%)	25 (56.8%)	8 (18.2%)	8
If there was a tool that allowed you get consent from Twitter users in order to use their publicly available data before data collection, would you use it?	7 (15.9%)	26 (59.1%)	11 (25.0%)	8
If there was a tool that allowed you share research outputs (such as papers and presentations) generated with publicly available Twitter data with participants, would you use it?	29 (65.9%)	14 (31.8%)	1 (2.3%)	8

In this case, the researcher's fear appears to be that by telling someone that they are using their publicly accessible Twitter data, that this would jeopardize the reliability of future data gathered from that individual—that the respondent may begin acting differently if they think researchers are watching.

There were a variety of other reasons why researchers indicated they might not want to provide notice to users. Rather than framing the issue around potentially altering the data to be collected, one researcher worried about the impacts that such notifications would have on users' anxiety, stating that notifying users may "*Risk creating more concern than necessary for low risk research topics*." Several other researchers indicated that informing users is beyond what is required of them, writing, "*It is not required. Should I provide them notice when I read their tweets as well?*" and "*The data is public*."

However, some respondents highlighted the positive benefits that informing users before or after data collection. One saw it as an opportunity to build bridges to a community they wish to study, remarking:

I still like the idea of interaction with the people whose content I am using. I understand that this is not a legal issue or an IRB issue, but I just prefer the bridge-building of interaction rather than the simple awareness from informing.

When the question was formulated as, "If there was a tool that allowed you get *consent* [emphasis added] from Twitter users in order to use their publicly accessible data before data collection, would you use it?" As seen in Table 5, there were markedly fewer "yes" responses to this question, and more "maybe" and "no" responses than both iterations of the informing question.

On the "yes" side, some researchers saw value in a tool that could facilitate consenting as part of getting permission for republication, stating, "*In cases where I want to publish identifying information and need consent -- or if the review boards get more strict -- I can see how this would be useful.*" On the "maybe" side, there were a number of researchers who indicated that they would use such a tool if required to do so by their ethics body, but would likely otherwise avoid it for fear of "altering the data" or fears that having users give a negative or no response to consent would mean not being able to use their data. On the "no" side, some researchers again mentioned the Hawthorne Effect while others used more legalistic reasons, arguing "*They gave their consent when they posted the tweet publicly*" and they would not try to consent users, *"Not if their information was already public and they had no reasonable expectation of privacy.*" When switching from informing to consenting users, there was a marked shift in the legalistic responses and in the inclusion of the IRB as a justification why.

Finally, we asked, "If there was a tool that allowed you share research outputs (such as papers and presentations) generated with publicly accessible Twitter data with participants, would you use it?" As seen in Table 5, the clear majority of respondents indicated they would use such a tool, with only one researcher indicating that they explicitly would not. In the follow-up open text, researchers provided a few justifications for why they would use or might use such a tool. Many of the justifications included some argumentation about researchers' obligations to share outputs, such as, "*Outputs should be accessible as a matter of course*" and "*more dissemination is always good*." Others indicated they would be interested in sharing their output because it may be of interest to the

user, with one respondent stating, "*I imagine many users would enjoy knowing that their tweets have made a difference*." In the maybe category, researchers who provided expansion on their selection indicated that the choice to share data would be contextually driven by the community they are researching and the topic of the project.

5. Ethical Practice beyond the IRB

Data from our study point to researchers going beyond what is "required" of them: by seeking out ethics review even when they are not required to do so, by seeking out user permission to quote from users even when they have no legal obligation to do so, and in sharing research outputs with users, even when they may not have an obligation to do so. This suggests that many researchers are engaging in contemplative practices around the use of publicly accessible data and see morally necessary practices beyond what is minimally required. Further, many of our respondents indicate that should tools that standardize the process and actually do the work of helping inform, get consent from, or share research outputs with users work at scale, that they would potentially be interested in using these mechanisms.

At the same time, we also observed responses that suggest some researchers would not engage in such practices, unless they were required to do so by their institution. The open responses to these questions leads us to believe that this could be from a combination of lack of norms around ethics and a fear of the unknown impact, if any, such practices would have on data availability and results (i.e. there is a fear that such practices would constrain data availability and thus decrease the validity of research results and possibility lead to Twitter and social media as impossible platforms to collect data from).

The tension in research ethics that we identify in this data is not new. However, our findings suggest that researchers' ethical practices are in part shaped by the kinds of tools available. Many researchers want to do more as far as connecting with users. Ethical practices in the field could change in relation to tools that could work at scale. As new tools develop, this may lead to changes in research norms in the social media research community, potentially increasing forms of contact between researcher and research subject.

Some scholars are working on tools to solve these issues [24], [25]. However, there may also be ways for social media companies themselves to make such ethics practice easier and to facilitate this kind of work. For example, on Twitter, a system that would allow researchers to upload a list of user and Tweet IDs that they have used for particular research studies, along with links to resulting publications, could allow users to look themselves up and to see how their content is being used by the scientific community. Similarly, mechanisms that allow users to see who has viewed their content could also increase users' understandings of how the information they create flows to research communities. Further, mechanisms that would allow users to opt-in to research or donating their data are another possible mechanism to further ethical practice.

5.1. Protecting Users, Protecting Data

Many of the respondents were quick to identify their obligation to protect users as part of their research practice in their answers. However, there were differing views on what constitutes protecting a user as an ethical practice. Some protected users by going back and asking users if it would be okay to quote their content as part of research publications. Others indicated that they felt it might be unwise to notify users about the use of their publicly accessible data because it could potentially cause unnecessary concern or anxiety within the user. These examples point us towards how researchers may go above and beyond the compliance model of ethics in order to enact the ethical principles of *respect for persons* and *beneficence.*

However, some researchers reasoning appears to more centrally protect what they see as the validity of data streams coming from the user. For example, several responses brought up the concern that by notifying users, user behavior might change. Some explicitly refer to the worry they have about the impact of the "Hawthorne Effect" on the validity of their work. Rather than a concern centered on the user, this concern appears as more of protection mechanism for the researcher themselves.

However, we wish to open the question (though certainly not resolve it) whether there actually is a Hawthorne Effect on Twitter. From our review of the relevant literature, while we can find anecdotal evidence, no studies have systematically analyzed whether the actual size and scope of such an effect, should it exist, in relation to data collection from Twitter. Undergirding our questioning about whether or not this phenomenon exists is the reality that the intellectual history of the Hawthorne Effect is, at points, somewhat dubious [26]. Further, a systematic review of work involving the Hawthorne Effect found, "Consequences of research participation for behaviors being investigated do exist, although little can be securely known about the conditions under which they operate, their mechanisms of effects, or their

magnitudes. New concepts are needed to guide empirical studies." [27].

We suggest that, if the argument is to be made that if users are in fact, aware that their tweets are publicly accessible, and can be used for research purposes (as noted in the Terms of Service), that reminding them of this should have little to no effect on their behavior. There is the potential they could be primed in some ways, depending on how much they

were told about the nature of the study, but in most cases, researchers are collecting data from Twitter either retroactively or as it's being created.

The reality of the situation, however, is that most users are unaware that their data may be used in this way [9]. This is either because they have not read the Terms of Service, or that no one has gone out of their way to inform users that this is the case. The relationship between users and social media platforms frequently involves a high degree of information asymmetry and reliance on users resigning themselves to the fate that they just won't know how their data is being used. Should researchers benefit from such an unequal power relationship? We suggest that researchers may want to consider informing users as an ethical practice, one that moves beyond the model of ethics as compliance. We suggest that the position of "don't tap on the glass and disturb the fish" problematically reinforces users not understanding what happens to the content that they create is used by third-parties. It disenfranchises users by denying them the opportunity to reflect on whether they wish to change their behavior on these platforms, or not. In many ways, relying on such logic potentially runs counter to the notion of "respect for persons."

5.2. Sharing Findings

While researchers hold mixed views of whether to inform and/or consent users when using their publicly accessible data, many researchers would be excited to have tools and mechanisms to share research outputs more broadly with users. Again, here we see research stance is motivated beyond the "ethics as compliance" model.

Ideally, if users better understand how researchers create public benefit through increased sharing of research outputs, users may have more reason to trust scientists going forward. It could increase public understanding of phenomena that take place on social media. And, in an ideal world, it could also decrease the degree to which academic knowledge built on the content created by users is stuck behind pay-walls.

6. Conclusion

Publicly accessible data obtained from social media is frequently used for scientific research. There are ongoing conversations about what ethical obligations researchers might have in regard to using such content. Our findings contribute to these conversations by highlighting a number of ways that researchers go beyond the minimally required practices of IRBs. Many researchers seek out ethics review even when not required to do so, contemplate their actions, and some seek to inform, get consent from, or share research outputs with their social media subjects. However, some researchers are limited in their ability to accomplish their ideals by the tools available. The scope and scale of data availability created by social media is part of its value, but also part of its challenge. Our findings suggest one possible roadmap of the kinds of practices researchers might be inclined to engage in, should the tools to do this kind of work at scale become available.

Users could also benefit from such tool development. It is possible that users would better understand how the content they create is used by the scientific community and how science progresses as a result of access to this kind of data. Unpacking the black-box of academic use of social media data potentially opens up numerous positive benefits. However, there are potential downsides as well. Some researchers had noted fears that additional efforts at informing, seeking consent from, and sharing research outputs with users could cause unnecessary anxiety among users. Further work is needed that examines how users would respond to being informed about the use of their content, how they respond when asked for consent to use their publicly accessible data, and how they respond to having research outputs shared with them. Additionally, more study is needed that can answer the question of whether or not the Hawthorne Effect is a legitimate concern for researchers using public data.

Lastly, more expansive study is needed that examines how researchers in a swath of disciplines are addressing these issues. Because our sampling methodology focused primarily on academics who had published in information science, communications, and human-computer interaction dominant conferences, our findings may reflect disciplinary views where ethics training and ethical thinking is often given prominent attention. Such views and practices may or may not be common in other fields.

7. Acknowledgements

The authors would like to thank Dr. Jessica Vitak for her help and advice regarding participant recruitment.

8. References

[1] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[2] J. C. Eichstaedt *et al.*, "Psychological language on Twitter predicts county-level heart disease mortality," *Psychol. Sci.*, vol. 26, no. 2, pp. 159–169, 2015.

[3] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[4] D. Freelon, "Computational Research in the Post-API Age," *Political Communication*, 35:4, pp. 665-668, Oct. 2018.

[5] M. Bastos and S. T. Walker, "Facebook's data lockdown is a disaster for academic researchers," *The Conversation*, 2018. [Online]. Available: http://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533. [Accessed: 15-Jun-2019].

[6] M. Zimmer and N. J. Proferes, "A topology of Twitter research: disciplines, methods, and ethics," *Aslib J. Inf. Manag.*, vol. 66, no. 3, pp. 250–261, 2014.

[7] Association of Internet Researchers [Air-L]., "The Air-L July 2018 Archive by thread," Jul-2018.

[8] J. Vitak, N. Proferes, K. Shilton, and Z. Ashktorab, "Ethics regulation in social computing research: Examining the role of Institutional Review Boards," *J. Empir. Res. Hum. Res. Ethics*, vol. 12, no. 5, pp. 372–382, 2017.

[9] C. Fiesler and N. Proferes, "'Participant' Perceptions of Twitter Research Ethics," *Soc. Media Soc.*, vol. 4, no. 1, p. 2056305118763366, 2018.

[10] danah boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 2010, pp. 1–10.

[11] A. Bruns and S. Stieglitz, "Twitter data: What do they represent?," *It-Inf. Technol.*, vol. 56, no. 5, 2014.

[12] J. J. Gibson, "The theory of affordances," *Hilldale USA*, 1977.

[13] N. Proferes, "What Happens to Tweets? Descriptions of Temporality in Twitter's Organizational Rhetoric," in *iConference 2014 Proceedings*, 2014, pp. 76–78.

[14] A. Bruns, "How Long Is a Tweet? Mapping Dynamic Conversation Networks on Twitter Using Gawk and Gephi," *Inf. Commun. Soc.*, vol. 15, no. 9, pp. 1323–1351, 2012.

[15] K. Driscoll and S. Walker, "Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data," *Int. J. Commun.*, vol. 8, no. 0, p. 20, Jun. 2014.

[16] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno, "Assessing the bias in samples of large online networks," *Soc. Netw.*, vol. 38, pp. 16–27, Jul. 2014.

[17] M. Brooks, "Human centered tools for analyzing online social data," PhD Thesis, 2016.

[18] M. Felt, "Social media and the social sciences: How researchers employ Big Data analytics," *Big Data Soc.*, vol. 3, no. 1, p. 2053951716645828, 2016.

[19] C. Goble, R. Stevens, D. Hull, K. Wolstencroft, and R. Lopez, "Data curation+ process curation= data integration+ science," *Brief. Bioinform.*, vol. 9, no. 6, pp. 506–517, 2008.

[20] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media Soc.*, vol. 20, no. 3, pp. 973–989, 2018.

[21] M. Ananny, "Toward an ethics of algorithms: Convening, observation, probability, and timeliness," *Sci. Technol. Hum. Values*, vol. 41, no. 1, pp. 93–117, 2016.

[22] P. Aufderheide, "'Does This Have to Go Through the IRB?,'" *The Chronicle of Higher Education*, 17-Aug-2016.

[23] J. Wickstrom and Bendix, "The" Hawthorne effect"-what did the original Hawthorne studies actually show," *Scand J Work Env. Health*, vol. 26, no. 4, pp. 363–367, 2000.

[24] J. Zong, "Empirically Studying Research Ethics with Interface Designs for Debriefing Online Field Experiments," Aug. 2018.

[25] J. Zong and J. N. Matias, "Automated Debriefing: Interface for Large-Scale Research Ethics," in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2018, pp. 21–24.

[26] H. Bastian, "The Hawthorne effect: An old scientists' tale lingering 'in the gunsmoke of academic snipers.,'" 2013.

[27] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects," *J. Clin. Epidemiol.*, vol. 67, no. 3, pp. 267–277, 2014.