2016 49th Hawaii International Conference on System Sciences

"We Aren't All Going to Be on the Same Page About Ethics:" Ethical Practices and Challenges in Research on Digital and Social Media

Katie Shilton
College of Information Studies
University of Maryland, College Park
College Park, MD
kshilton@umd.edu

Sheridan Sayles
College of Information Studies
University of Maryland, College Park
College Park, MD
ssayles@umd.edu

Abstract—**Defining ethical practices for research using data from digital and social media communities is an ongoing challenge. This paper reports on interviews with digital and social media researchers to investigate the challenges they experienced when collecting, managing, and analyzing online data. Analysis of the interviews reveals a diverse set of ethical challenges that push at the boundaries of existing research ethics guidance. The interview data also describes existing practices for navigating ethical quandaries, and documents resources that help researchers meet ethical challenges. The analysis of the data points to opportunities for review boards and ethics researchers as well as new debates to undertake as a community.**

Keywords-research ethics; internet research; social media

1. Introduction

Never before have data about human interactions been so available to researchers. Public attention to research ethics in this space has grown as reports of the amount and intensity of research using data from digital and social media platforms have appeared [1]. At the heart of this attention are concerns that the collection and analysis of online data for research presents new ethical challenges due to the data's ubiquity, its context-spanning nature, and its opacity. Researchers and the public question whether we need new or modified ethical principles, practices, and review tools for performing research with this data.

This study contributes qualitative data to this ongoing debate by documenting the ethical challenges faced by a multidisciplinary community of digital and social media researchers. It answers calls for empirical data about research ethics [2], using interviews to understand how interdisciplinary researchers discover ethical challenges and document practices they use to face those challenges. This research asks:

1. What ethical challenges have digital and social media researchers discovered in their work?
2. What practices are researchers using to face these challenges?
3. What resources do researchers need to better navigate research ethics for digital and social media data?

2. Background

In the U.S., research ethics have long been guided by the Belmont Report, which focuses on respect for persons, beneficence, and justice [3]. Respect for persons has most widely been interpreted by ethics review boards as a mandate to obtain informed consent from participants when collecting private data. Openly-available digital and social media data may be interpreted as public, however, and collecting informed consent at scale to use this data may be difficult or impossible. But if we interpret respect for persons broadly, we must consider that much of this data documents work processes and practices that may have required informed consent for data collection in other settings. Contributors to online forums may have no idea such data could be harvested by researchers. For example, researchers who investigate sensitive issues such as values or political conflicts have struggled with whether informed consent was necessary [4], [5].

Beneficence is the second Belmont principle challenged by digital and social media data research. Generally understood as assessment of risks and benefits of the research, it is a principle that guides researchers to think through possible negative consequences of their work. One challenge of using online datasets is the difficulty of providing anonymity. Re-identification risks abound in big datasets [6], [7]. It may also be difficult for researchers to anticipate risks and unintended consequences of online research [8]. Researchers using digital and social media data must also consider whether their research presents a risk to the community they study. While anonymizing individual-level data may protect individuals from scrutiny and exposure, such research frequently identifies groups and communities. Negative results—or the attention and scrutiny such results can bring—may harm the community and complicate ongoing participation for members.

Finally, justice has widely been interpreted by ethics boards as attention to the selection of research subjects. This is an under-investigated area in digital and social media research [9]. Online participants are largely self-selecting, and online community

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.242

participants are generally more affluent and educated than the general population [10], [11]. It may also be difficult for researchers to tell if participants from vulnerable populations (such as children) are included. Reflection is needed about whether potential biases in the study of online data generate justice issues.

In 2002, the Association of Internet Research published guidance detailing questions for researchers to ask themselves when performing internet research, alongside case studies and other resources to help inform online research [12]. These recommendations suggest that researchers consider the environment of their study, standards within their country and research community, and precedence for the type of research. The AoIR revisited its recommendations in 2012 and continues to advocate for flexible guidelines as opposed to fixed codes [13]. Recent publications have focused on how flexible policies need to be in order to facilitate the wide array of current internet research [14], [15].

Meanwhile, more specified codes of ethics for Internet research exist for international research contexts (e.g. [16]) and several universities have created their own guidance for researchers (e.g. [17], [18]). The disparate nature of resources for ethical Internet research guidance begs the question: what are researchers using for guidance, and what are they doing in practice?

3. Methods

We began inquiry into norms, practices and challenges in researcher use of digital and social media data by conducting semi-structured interviews with interdisciplinary researchers working with internet datasets. We used snowball sampling and citation chaining to find researchers experienced with digital and social media research. We began by soliciting collaborators (comprised of colleagues in information science, computer science, and business schools) who were building an online data repository. We asked this group for the names of "emerging leaders" in online data research. Selection of interview subjects focused on balancing demographic, disciplinary, and job/rank diversity. We interviewed 20 researchers with PhDs in information technology, information systems, information studies, communication, business, and computer science. All were faculty at US and European academic institutions, or researchers in consulting or industrial research labs. Fourteen researchers reported studying open platform social networks (such as Twitter), and six researchers reported studying restricted platform social networks. Three respondents conduct usability studies. Five researchers collect activity traces from open or restricted platforms. One researcher collects GIS data. Only two researchers report collecting explicitly sensitive information, such as health information or information about minors.

We conducted interviews until we reached conceptual saturation: until we felt no new issues were being raised by participants [19]. As this is a non-representative sample of researchers in this space, we do not seek to generalize from the findings here. Instead, we wish to provide qualitative data to shape questions and topics for future research.

The semi-structured interview protocol focused on how researchers discovered and resolved ethical challenges in their work. Interviews ranged from 15 to 45 minutes, and were conducted by the second author via Skype. Interviews were recorded, and the interviewer took field notes during the interview. Field notes and transcriptions were coded using DeDoose qualitative analysis software. The authors worked together on an iterative coding process. Based upon our research questions, we began by coding for 1) ethical challenges, 2) needed resources, 3) challenge discovery, and 4) practical solutions. These categories expanded over two iterative, joint rounds of coding to incorporate the final high-level codes listed below:

Ethical challenges	Response to research
Regulatory challenges	Solutions
Challenge discovery	Needed resources

4. Findings

We grouped our findings into four main themes: ethical and regulatory challenges reported by researchers (Section 4.1), how researchers discovered ethical challenges (Section 4.2), researchers' practical solutions to ethical challenges (Section 4.3), and resources requested by researchers for dealing with ethical challenges (Section 4.4).

4.1. Ethical and regulatory challenges

The ethical challenges reported by interview subjects were many and diverse. Some were predicted by the literature, including gaining consent, navigating restrictions by platforms, weighing risks versus benefits to participants, and defining sensitive information and participant privacy expectations. But concerns emerged that were largely unmentioned in related literature, as well. These included being perceived as spam, worries about judging participants, and a pervasive feeling that everyone else (commercial interests and governments) was using this data, and that academics shouldn't be restricted from using it.

4.1.1 Consent. Six interview participants discussed the challenges of gaining consent for digital and social media data. Several participants mentioned a recent controversial Facebook study [1], [20], citing lack of participant consent as a primary issue in that research. On the other hand, some researchers felt that online data such as that posted to Twitter is public, and therefore consent was implied. Researcher H put it this way:

> Some [subjects] say, "I didn't grant consent." … And my counter is, "You did grant consent. You posted to Twitter, publicly. What did you expect?"

Researcher I summed up the tension between these two perspectives nicely, by stating that at the root of the challenge is whether platform users understand that their data is being used for research:

> Twitter data in particular is a little bit of an issue, right? Because there is this concern that it is publicly available data. So some people would argue that anything that I put in the public space should be, by rights, available to anyone else, right? But a lot of people aren't aware of the fact that that's the case.

Researchers also discussed the logistical difficulties of obtaining informed consent for large-scale studies. Researcher R described:

> …when we collect data, it might be millions of users. So, that would imply basically emailing those millions of users, telling them about the things that we are doing. So maybe, if I were to do another type of project in which I would focus on say 100 users or 60 Twitter users… if it was a smaller scale, maybe I would start thinking about ways of letting them know whether that's fine with them.

4.1.2 Justice. Justice, and relative injustices, were another set of concerns brought up by researchers. Some researchers, such as Researcher H, saw their work encompassed within larger justice problems with big data:

> This is the ethical challenge right now, that big data has a lot of power over people who are the parts of big data. And we don't have any visibility into it, we don't have control over [it].

Another justice issue described by Researcher F. related to the accessibility of online activities:

> Who has access to those kinds of technologies? And that's access both in terms of socio-economic or demographic factors, but also in terms of cultural factors, where people live, access to the internet, that kind of stuff. … If you do have differential access, how does that affect the knowledge that you're able to create in a research setting?

A final justice issue, here raised by Researcher U, was framed as accountability to publics beyond the group being studied:

> I don't think that we should think of our research subjects as the only people who are affected by the work. So, if you're looking at something like hate communities, those people who are being hated may not be [your subjects], but they might be people to whom we're most ethically accountable. …because I think if you're studying people whose purpose is to harm others … then maybe accountability to them should not be [the central issue]. Maybe accountability to the greater society is more important.

4.1.3 Risk. Several researchers worried about unintended risks, or their inability to properly judge risk. Some, such as Researcher F, spoke about this in general terms:

> We don't have a good sense of what risks are for this data, which is I think one of the real issues... I've heard a lot of people thinking about the risks in online spaces and using metaphors from the real world to talk about things like privacy or about risks of data breaches or whatever… The kinds of breaches and the kinds of consequences are probably relatively rare but potentially highly impactful which is not a particularly good space for rational reasoning.

Others, such as Researcher U, gave very specific examples:

> And [a colleague] had some cases of researchers … who were studying pro-ana [pro-anorexia] blogs and one of the girls, when she realized she was being studied, got super excited because now she had an audience for her anorexia, and she could demonstrate to the researcher how much weight she was losing and how thin she was getting, and so just the fact of being studied made her condition worse.

On the other hand, Researcher I felt that the rewards of online data research outweigh risks:

> I'm thankful that there are people who aren't as concerned about their privacy because I feel like [online data] provides us with an interesting lens by which to understand human social experience and human social communication. And so I would argue that to the extent of which there's an invasion of privacy, my hope is that at least we're taking advantage of that to better understand the human condition and therefore, advance science.

4.1.4 Privacy. As the quote above gestures to, many researchers expressed that they felt user privacy expectations for online spaces were unclear or unsettled. In fact, unclear privacy expectations were the most-frequently discussed ethical issue in our dataset, coming up in at least twelve interviews. As Researcher I recounted:

> Even though this is publicly available information, a lot of users don't realize how easy it is to collect that data, and so might have not wanted it to be publicly available, if they had known other individuals would potentially be collecting that data.

Others such as Researcher H argued that, because users have fine-grained control over data sharing on many platforms, remaining data is fair game for research:

> I mean you can go back and turn off tweets. You can delete them. You can make them private. And nothing I ever do ever attempts to overcome or defeat access controls.

Researcher I also raised the question of whether objections are coming from platform users themselves, or non-users who are less familiar with information norms in those spaces:

> I do present these results often in public forums … Occasionally I have people who are amazed that so many other users are putting content out there that's publicly available, and the kinds of content they put out there that's publicly available. But I haven't had anyone tell me that they're concerned about the fact I'm collecting this data. Most of the time the people who come up after those things are not actually involved in putting out the content themselves. And so they're not concerned for their privacy, they're just amazed that other people have such little concern for privacy, I guess.

4.1.5 Anonymity and reidentification. Several researchers worried about the challenges of data anonymization. More than one researcher cited a study in which a *New York Times* reporter re-identified an anonymous participant using her tweets. Several researchers mentioned similar reidentification challenges in their own research, such as Researcher U:

> …we have the [Twitter] archives and we have the transcripts, but the minute we use one of those tweets, then anonymity is shot.

Researcher S described the challenge this way:

> If you describe a certain platform, then sometimes you run into the cases that individuals are almost identifiable from a unique set of actions. … How do you deal with these data that make that individual identifiable? And, how can you make sure that the data that you work with is changed in such a way that this doesn't happen?

Researchers such as Researcher L struggled with whether to point out to others the ease of data deanonymization:

> I've conducted research where I was exploring someone's use of anonymous data and I was able to identify the source of that data. And I had an ethical dilemma on whether or not to go public with my success in re-identifying that data set.

Researcher U also pointed to technological change as a challenge in this area:

> At the time that I did that early … research, I was able to collect the data by manually saving all the messages for a long period of time before they expired from the university servers and disappeared. And then some years later, Google or Deja News happened … and all of that stuff became searchable. So I had gone and nicely anonymized people, but all they had to do is Google what's in my book, and they could find out who wrote it. So the technology's changed, and what was private earlier, by changing the names, now isn't private because everybody can just search the quoted material.

4.1.6 Judgment. Four researchers brought up a less-discussed ethical issue: the perception of judging participants. Researcher H described:

> Not everybody enjoys having these data artifacts presented to them because it highlights the idea that data has been gathered about their behavior and essentially, that data is being used to judge them. Because I am judging you. I'm saying, "You are fabulous. You're marvelous. You're the best. You're the center. And by the way, that means that if I didn't mention your name, you're not on the fabulous list."

In a different example, Researcher C related challenges with a system designed to mitigate cyberbullying:

> The point of the study was to develop these cyber bullying reversal pings, so that once a person who was being bullied was identified, you could send positive messages. So, there was a lot of ethical considerations with: … if you get a false positive, are you going to potentially make the issue worse? … if they're not actually experiencing cyber bullying, or they didn't feel bad, but then you bring up an old issue again…

In these cases, researchers identified the classification of participants into (potentially judgmental) categories as an ethical issue. While "judging participants" is not immediately identifiable as a concern in the research ethics literature, we believe that researchers' concerns about judgment reflect several classical ethics concerns. These researchers may be identifying a mismatch between online participants' contextual expectations for their data and realities [21]. These concerns may also reflect larger debates about the role of social media data in categorizing and labeling individuals for predictive analytics and, potentially, discrimination [22].

*4.1.7 Intrusion***.** Related to worries about judging participants were worries about potential disruption of online communities through intrusion. Three researchers framed this worry as being perceived as spam. As Researcher H recounted:

> I've had two or three encounters on Twitter where it's been an up and back saying, "You know, you don't have a right to do this. This is spam. This is bad. You should stop.

Researcher O described his experience this way:

> I was very early on doing Wikipedia research… sending out these surveys to people. Well, apparently I pissed the wrong people off at Wikipedia. They blocked my account, wanted to know what was going on, and I had to negotiate with this administrator, the appropriate way to do my survey… I successfully threw myself on their mercy, tried to be transparent and eventually, I convinced them that I was legit and not just spamming or whatever.

Worries about judging participants and intrusion into communities both reflect concerns that studying online spaces might negatively impact people's online experiences. Both concerns are worries about harms to online participants that are difficult to quantify. Researchers are unsure if negatively impacting someone's social media experience is enough of a harm to prevent certain kinds of research practices.

4.1.8 Disagreement and differing norms. Not everyone expressed worries about ethical challenges in their research. Four respondents reported that they hadn't experienced ethical issues or challenges in their work. Two researchers attributed this to not collecting identifying information. As Researcher T described:

> … Most of the stuff, it's anonymized anyways, data that I get, I don't know who the users are. I don't have any personally identifiable info.

Two others attributed the lack of ethical challenges to luck. As Researcher W put it:

> I'm certainly aware of the things that could come up. I haven't particularly run into any ethical problems myself.... And I don't know whether, 'cause I'm not doing anything terribly sensitive, or whether it's just because we got lucky.

Other researchers, such as Researcher H, cited a lack of concrete ethical norms as a challenge in their work:

> I guess I wanna know how representative a critique is, and how representative does it have to be to be addressed? So if one out of a million people say, "Hey, I don't like that," is that a concern that really needs to be considered?

Not only do online data subjects potentially disagree about research ethics, researchers also disagree. As Researcher U put it:

> It's not as though you can sit a whole bunch of internet researchers down in a room and we're all going to be on the same page about ethics.

4.1.9 Everyone else is doing it. A reoccurring theme in the interviews was the feeling that academics should engage in online data collection and analysis, because everyone else is doing it as well. As Researcher H said:

> You know that all of this is being done by the platforms themselves. Like "Is Twitter analyzing Twitter?" You betcha.

He added to his concerns the pervasiveness of state surveillance:

> I'm sorry to get my tin foil hat out, but… The feds are gonna take your data...

Academics wonder why research should be especially restricted, as online data is being used for numerous other forms of analysis.

4.2. Challenge discovery

We also asked respondents how they had discovered ethical challenges in their research. Interestingly, none of the interview subjects reported being challenged on research ethics directly by ethics review boards. Instead, researchers reported being challenged by their peers, including peer reviewers and funding agencies, and their colleagues on interdisciplinary teams.

4.2.1 Lack of Challenges by Review Boards. No interview subjects reported that an institutional ethics board had challenged them in their research, and subjects cited different but related reasons for this. Overall, respondents reported that review boards consider online data public and therefore ineligible for review. As Researcher C described:

> When you're publicly scraping data, the IRB generally says it doesn't need to be reviewed... So for the scraping part of our study, we didn't need anything because we didn't have "human subjects" data.

Scrutiny also varies heavily between institutions, as Researcher O described:

> I have colleagues who deal with much more risk-averse, much more authoritarian IRBs, and I just don't. So I think it just depends on the culture of the IRBs.

Some researchers, such as Researcher C, attributed institutional variance to the fact that IRBs are largely comprised of generalists:

> The IRB is great for providing you guidance on, "This is what you need to do so that people don't try and sue us," but they're generalists, not

> specialists, so in a specific instance of a research study, maybe nobody has expertise in that area…

A prominent theme throughout the interviews was that IRB-suggested changes to studies were misguided, irrelevant, or did not address ethical issues. As Researcher V put it:

> If there were changes, they were administrative changes. They're not like substantial changes... It's more like you are missing these two forms and therefore, you should add them. And so, it's more procedural, not substantial.

4.2.2 Challenges by peers. According to participants, networks of peers seemed to be a much more effective ethical check than IRBs. Several researchers, including Researcher O, reported discovering ethical challenges in their work when write-ups of the study went out for review by conferences, journals, or funding agencies:

> It wasn't 'til reviewers started asking questions … that I had to start wrestling with [ethical] issues.

Researcher C recounted:

> When [my student] applied for NSF funding last year … the reviews all focused on ethics.

These experiences indicate that anonymous peers may be powerful agents for influencing attention to research ethics. Other researchers, including Researcher D, indicated that interdisciplinary colleagues were also good resources for discovering and discussing ethical challenges:

> [Ethical issues arise] just simply in conversation. Of the four of us, one is a mathematics professor, so she is like the least familiar with this. One's an anthropologist, so she deals with this, but in real life. And the other's a sort of English composition instructor, then there's me. ... And [ethical challenges were], again, literally this week's discussion.

4.3. Practical solutions

Practical solutions to ethical challenges demonstrated by researchers tended to group into two categories: discussion of *how* to make ethical decisions, and discussion of concrete actions. Discussion of decision-making tools included consulting existing ethical guides and relying on existing social networks for advice. Discussion of concrete actions included providing transparency into research, removing non-consenting individuals from datasets, minimizing data collection, aggregating data, providing participant consent or control over data, and collecting only historical data.

4.3.1 Existing ethical guides. Though few researchers reported being *challenged* by their IRBs, some reported positive experiences relying on their IRBs for *guidance*. As Researcher G stated:

> I don't think I needed additional resources. The IRB Board was superb in helping.

Researcher O described his collaboration with the IRB in detail:

> I have a fairly open-minded IRB … and we just talked through how I was planning to handle things, what the issues were. And they worked collaboratively … to solve them. Since there weren't any real strong protocols in place, and it wasn't clear what the ethical issues were, we just had to try and proceed. And sort of hashed these issues out as we went.

Other researchers, including Researcher L, recounted drawing on guidance from non-research communities to think about the ethics of their work.

> I actually kinda looked at hacker culture as part of my... a way for me to think about this. Like when a hacker finds an exploit, do they tell the person or do they go public with the exploit? And as a hacker, I decided to go public 'cause I thought that generating this kind of awareness would help prevent this in the future.

4.3.2 Existing social networks. Three researchers related relying on existing social networks to discuss ethical solutions. They referenced asking co-authors, trusted colleagues, and colleagues researching the same platform for advice on ethical issues. As Researcher O described:

> I've had plenty of exchanges between colleagues doing this type of research, and so I just try to not just let it be my decision, but I seek out the advice of trusted colleagues for the best way to proceed.

Researcher D described her process this way:

> What do I do when I run into this problem? Honestly, I post it to Facebook and tag my friends, who I know do this kind of research… and say, "Hey, help!"

4.3.3 Participant consent and control. Despite the challenges of collecting informed consent, some researchers, including Researcher S, do so:

> …we specifically said, "This is an experimental set-up. This is for research purposes. Know that if you participate, all your activity will be monitored and will be used in a paper."... The work that we did so far, was with people clearly consenting to this, to be monitored…

Participant V worries not only about consent, but about debriefing participants after procedures involving deception:

> The other issue that I sometimes have to be very careful is … keeping the participants informed because sometimes I do experimental research

> … and deception is sometimes inevitable … And then with the online environment, that makes debriefing a challenge. So, I oftentimes put in a good amount of thinking to figure out how to debrief.

4.3.4 Transparency. Transparency or openness with data subjects was the most-discussed ethical solution, with seven researchers volunteering that they had explored transparency as a way of addressing ethics. Researchers described various tactics for transparency. Some researchers, including Researcher S, shared data collection plans with platform providers:

> I always run [data collection] by them, by these platform owners. And often, we send it out together with them. So, they have a say in that. …We talk about like what data we'll store and what we'll not store, and those kind of issues.

Other researchers, such as Researcher G, took steps to inform subjects of their presence:

> We reminded people on many occasions [of data collection], and also there was a note run on the discussion board, that any comments that they made were open to the world.

On some occasions, Researcher G also involved participants in more depth:

> I have opted to show, to circulate drafts of papers to participants, and they have sometimes given me a few comments.

The most complex reported version of transparency involved an open drafting process reported by Researcher D:

> What we have decided to do, at least for now, is to make the paper open. So the paper is being drafted in Google Docs, and we're providing the link on the forum, for those who want to look at it. And we're giving people up to a week after our first draft to make comments. So people are like, "Eh, I'm not comfortable with using my identity or wait... I don't know about that," or, "That's not how I feel." That's one of the ways we're trying to handle [ethical challenges].

Researcher N summarized why he thinks transparency is so important:

> I think it's important to give back to the people, especially when you are working with online communities, because they are more responsive than anonymous surveys or questionnaires you send to people... So you have to be a bit more responsive too.

4.3.5 Protection measures. Interview respondents also discussed techniques for protecting the identity or participation of individuals in their research. A few researchers discussed letting individuals opt out, such as Researcher H:

> If you don't want me to include you, out of respect for you, I would do that as an ethical researcher, I'll remove you from my data set.

Other researchers, such as Researcher C, tried to minimize data collection to preserve participant confidentiality and prevent potential harms:

> We've spent a lot of time thinking and talking about ways to preserve anonymity to the greatest extent possible. And only collecting information that is deemed essential to answering research questions, not collecting extraneous information just because it's out there.

Other researchers, such as Researcher I, worked with or presented data only in aggregate:

> We still don't publish individual level data, right? We're not gonna publish that this particular individual was at that particular location.

4.3.6 Relying on historical data. Finally, some interview participants cited using historical data, such as older posts and online activities, as a way to alleviate privacy concerns. For example, using historical data was part of Researcher I's ethics practices:

> We primarily publish our results on historical data, right? So, our hope is that where a person was actually posting Tweets from, during, for instance, Hurricane Sandy, which is one of our data sets, is not a big concern to those individuals anymore. Right? It's no longer much of a privacy concern to them.

We will return in the discussion to the question of whether using historical data alleviates privacy and research ethics concerns.

4.4. Needed resources

Some researchers interviewed felt they had the resources they needed to deal with ethical issues, many citing the AoIR guidelines [13] as key guidance. However, most participants felt that having additional resources available would be beneficial. Request for resources fell into two categories: requests for structured codes of conduct, and requests for shared learning resources.

4.4.1 Codes of conduct. The single most requested resource—but also the most controversial—was a code of conduct for internet research. Five researchers brought up this topic. As Researcher I put it:

> I mean, a set of guidelines would be nice in terms of trying to know exactly how to deal with this data.

Researcher S was very specific:

> It would be great if there were any set of guidelines like, "You cannot publish data or you

> cannot use data that is... " For example, when do you ask consent? For what kind of data you need consent? And what kind of questions would you need to ask in order to be able to publish, use data from a certain source?

Researcher S also expressed a desire for a field-wide consensus:

> It would be interesting I think [to find out if] there is kind of a consensus of what we can and can't do. … So if we all kind of abide by the same set of rules, …that certain data we simply do not collect, or something like that. So that would actually help.

Researcher H framed this as an issue of legitimizing his decision-making:

> Well, a clear code of conduct and a clear adjudication system. If I could have shown those people who say… "You shouldn't do this." If I said, "Look, we're a member of ACM, AOIR, ICWSM, IEEE, or some standards organization that says this is legitimate research. This is how you do legitimate research…

There was an emphasis among many researchers that any guides created be substantive, rather than procedural. Researcher F put it this way:

> I don't need another guidebook to tell me … how to write a consent form that will satisfy my IRB. What I need is something that actually tells me how to understand ethical issues in this space and think through them in a constructive manner.

Other researchers, such as Researcher U, pointed out that there were existing codes which provided good guidance in this area:

> I think the Association of Internet Researchers' Ethics Committee has done a really good job developing those resources… They've developed a set of materials that speak to the fact that people are doing very different kinds of work in very different kinds of contexts, and acknowledged that there isn't any sort of "one size, fits all" ethical solution, but instead are focused around: what are the critical questions that you have to ask yourself over and over along the way to be able to make ethically defensible decisions? …I would like it if the Association of Internet Researchers' hard work were more widely recognized and taught and discussed and known.

On the other hand, some researchers had hesitations about formal codes, for fear that resources would mean more bureaucracy and less flexibility. As Researcher O put it:

> Other than a trusted network of colleagues, I'm not sure I would personally want more resources. Because I think, we, academics, we are a long way from the Tuskegee syphilis studies and I just don't... I'm all for being ethical. I'm all for protecting people's rights, but I also think you can go overboard with that stuff, and I fear more resources would expect people to be more bureaucratic about it. "Oh, you've done this wrong," and "Oh, you've done that wrong."

Researcher F stressed the importance of a slow, deliberative process when generating these norms:

> I would like to see resources, but my hope is that they are resources that don't... My worry is …we rush into framework... There's some movement going on right now around, in Congress, to start trying to apply HIPAA to all devices, all wearable device data for example. That would end up closing off all kinds of avenues of exploration and knowledge and usefulness of that data without a very good understanding of the risks or new, more inventive ways of thinking through them.

4.4.2. Resources for shared learning: Many researchers requested the formalization of ways to learn from colleagues' existing practices. Researcher V said:

> I think both within my own institution and also across institutions, it would be nice to... it would be nice to have more established ways to learn from more people.

She went on:

> If I could ask for a resource, it would be a platform to share best practices… If we could create best practices knowledge repositories for people to share their stories, to ask and answer each other's questions, I think that's would be super useful as well.

Alongside requests for sharing practices were request for sharing language for talking about research ethics, such as from Researcher L:

> I think researchers probably need to be better trained on how to answer questions related to privacy, related to some of the ethical issues. And I think, especially researchers that are, might be doing work that is a little bit on the cutting edge and might be getting public reaction about ethical concerns… I think we'd be better served if we had better, literally, talking points somehow to explain what we're doing and why we're doing it, why this needed consent or that didn't need consent.

5. Discussion

Analysis of the interview data reveals that, despite over a decade of discussion of internet research ethics, researchers still struggle to address ethical challenges in their work. Though some feel they have the guidance they need from existing resources, colleagues, or institutional structures, many still worry about the ethical impacts of their work.

A few themes emerged from the interviews that warrant further discussion and research. First, researchers see a need for ongoing debate and consensus-building around particular digital and social media research practices. Researchers suspect there may be important differences in practice that we must account for if we're to build a consensus on research ethics. Future research to generalize about researchers' ethics beliefs and practices, and any critical differences in those two areas, is needed.

Though more generalizable research is needed to quantify the scope of this problem, it's clear from the interviews that one prominent area of diverging belief and practice is likely to be how to collect meaningful informed consent in online data research. Though the authors are sympathetic to the difficulty of seeking informed consent at scale, just because something is hard doesn't mean it's not right. Indeed, researchers who did not collect consent at scale expressed that they might if they were working with smaller samples, indicating a concerning ethical dissonance. This equivocation based on sample size paints the objection as instrumental rather than philosophical: there's no clear ethical reason why subjects in smaller studies deserve a different standard of respect. On the other hand, definitions of meaningful consent have been debated in participatory and ethnographic research communities for years [23]. Perhaps the problem of consent for online data collection provides an opportunity to relieve widespread dissatisfaction with today's problematic informed consent models. Modes of transparency explored by researchers in this study may be one step towards relieving this dissatisfaction. Consensus-building symposia and work with review boards should focus on ways that online researchers can be transparent with research subjects—in big or small studies—as a more engaged and meaningful form of informed consent.

Another issue worthy of ongoing debate is whether historical digital and social media data is, in fact, ethically distinct from current or recent activity data. While the use of historical datasets have traditionally been subject to less regulation in universities, work on donor restrictions in archives suggests that many individuals see a privacy interest in historical data, as well [24]. Whether and how historical digital and social media data is different from current data (and what connotes an appropriate length of time before data becomes "historical") is an important but currently under-debated issue. Research to both describe and explain social media users' historical privacy expectations could be an important next step to understanding whether research ethics has an unrecognized obligation to protect such data.

Next, this work suggests that ethics review boards (or alternative institutional structures) might best be positioned as consultants to research design, rather than post-hoc enforcement mechanisms. Industrial research labs are already exploring models that consult on research design rather than review according to a narrow set of rules [25]; academic institutions might learn from their experiences. The data also suggests that peer reviewers sometime serve as post-hoc ethics enforcement mechanisms. Challenges by reviewers and peers were the most frequently-cited methods of discovering new ethical challenges. It is not surprising that some internet research communities are regulating themselves: this is an important function of anonymous peer review. Positioning peer reviewers as ethical referees takes advantage of existing work practices, as academics place great importance on reviewing each other's work. However, such review frequently happens once research is completed, meaning it potentially wastes researchers' time, and worse, doesn't mitigate harm. There is also likely great variability among reviewers of their comfort and expertise flagging ethical concerns. Further research is needed to understand how and why reviewers flag ethical concerns in digital and social media research, and whether this varies across disciplines.

Finally, there is a clear opportunity for platforms for shared learning in the internet research ethics space. Publications focused on ethical research exemplars; knowledge bases for consent, de-identification, or data aggregation techniques; or language for expressing risks and benefits to participants would all be welcomed by the internet research community.

6. Conclusion

This research illustrates some of the very difficult and unresolved ethical challenges faced by digital and social media researchers. Ethical norms among both researchers and participants are still in flux, making the construction of concrete specifications nearly impossible. There are no catch-all solutions for digital and social media research ethics.

Though universal solutions are an unlikely outcome for internet research ethics, researchers are successfully grappling with ethical research practice while collecting pervasive and available digital and social media data. Interviews revealed that prime areas for follow-up research include studying modes for increasing transparency with research subjects, and understanding social media users' expectations for privacy of historical data. Findings also suggest the importance of ethics-oriented changes to university review board processes and peer review practices. Documenting the practices of current researchers, the challenges they face, and their desires for resources presents empirical data from which to take these next steps for internet research ethics.

7. Acknowledgements

Thanks to colleagues Brian Butler, Sean Goggins, and Susan Winter for ideas and feedback that shaped this work. This research was funded by NSF Award *IIS-1449188 SAVI: Fostering New Collaborations in Open Online Community Data Research: Prototyping an Open Collaboration Data Factory.*

8. References

[1] V. Goel, "As Data Overflows Online, Researchers Grapple With Ethics," *The New York Times*, 12-Aug-2014.

[2] J. M. Hudson and A. Bruckman, "Using Empirical Data to Reason about Internet Research Ethics," in *ECSCW 2005*, H. Gellersen, K. Schmidt, M. Beaudouin-Lafon, and W. Mackay, Eds. Springer Netherlands, 2005, pp. 287–306.

[3] Office of the Secretary of The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research," Department of Health, Education, and Welfare, 1979.

[4] J. A. Koepfler, K. Shilton, and K. R. Fleischmann, "A stake in the issue of homelessness: Identifying values of interest for design in online communities," in *Proceedings of the 2013 conference on Communities & Technologies*, Munich, Germany, 2013.

[5] Y. Zhou, K. R. Fleischmann, and W. A. Wallace, "Automatic Text Analysis of Values in the Enron Email Dataset: Clustering a Social Network Using the Value Patterns of Actors," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 2010, pp. 1 –10.

[6] M. Lease, J. Hullman, J. Bigham, M. Bernstein, J. Kim, W. Lasecki, S. Bakhshi, T. Mitra, and R. Miller, "Mechanical Turk is Not Anonymous," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2228728, 2013.

[7] P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Rev.*, vol. 57, p. 1701, 2010.

[8] M. Zimmer, "'But the data is already public': on the ethics of research in Facebook," *Ethics Inf. Technol.*, vol. 12, no. 4, pp. 313–325, Dec. 2010.

[9] E. Hargittai, "Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites," *Ann. Am. Acad. Pol. Soc. Sci.*, vol. 659, no. 1, pp. 63–76, May 2015.

[10] A. J. Berinsky, G. A. Huber, and G. S. Lenz, "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Polit. Anal.*, vol. 20, no. 3, pp. 351–368, Jul. 2012.

[11] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in mechanical turk," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2010, pp. 2863–2872.

[12] C. Ess, "Ethical decision-making and Internet research," Association of Internet Researchers, 2002.

[13] A. Markham and E. A. Buchanan, "Ethical decision-making and internet research," Association of Internet Researchers, 2012.

[14] J. G. Warrell and M. Jacobsen, "Internet research ethics and the policy gap for ethical practice in online research settings," *Can. J. High. Educ.*, vol. 44, no. 1, pp. 22–37, Apr. 2014.

[15] C. Munteanu, H. Molyneaux, W. Moncur, M. Romero, S. O'Donnell, and J. Vines, "Situational Ethics: Re-thinking Approaches to Formal Ethics Requirements for Human-Computer Interaction," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York, NY, USA, 2015, pp. 105–114.

[16] H. Felzmann, "Ethical Issues in Internet Research: International Good Practice and Irish Research Ethics Documents," Research-publishing.net, 2013.

[17] "IRB guideline X - guidelines for computer- and internet-based research involving human participants," Penn State University, The Office for Research Protections, 2007.

[18] Office of the Vice President for Research, "Guidance for Data Security and Internet-Based Research Involving Human Participants," University of Connecticut, 2015.

[19] J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 2nd ed. Thousand Oaks, CA: Sage Publications, Inc, 2002.

[20] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proc. Natl. Acad. Sci.*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014.

[21] H. Nissenbaum, *Privacy in context: technology, policy, and the integrity of social life*. Stanford, CA: Stanford Law Books, 2009.

[22] C. Dwork and D. K. Mulligan, "It's not privacy, and it's not fair," *Stanf. Law Rev. Online*, vol. 66, no. 35, Sep. 2013.

[23] B. Thorne, "'You still takin' notes?' Fieldwork and problems of informed consent," *Soc. Probl.*, vol. 27, no. 3, pp. 284–297, Feb. 1980.

[24] S. McKemmish, "Evidence of Me…," *Arch. Manuscr.*, vol. 24, no. 1, pp. 28–45, 1996.

[25] A. Bowser and J. Y. Tsai, "Supporting Ethical Web Research: A New Research Ethics Review," in *Proceedings of the 24th International Conference on World Wide Web*, Republic and Canton of Geneva, Switzerland, 2015, pp. 151–161.