Scroll.in

RESEARCH DIGEST

Scientists are studying your tweets, YouTube videos, Instagram posts. Is that ethical?

A group of researchers is working on rethinking ethical guidelines for studies using social media data.

Casey Fiesler
Apr 14, 2019 · 11:30 pm

Kacper Pempel/Reuters

Who do you think reads your social media posts? You might assume that only your followers see your tweets and status updates – but someone else might be taking a close look. Scientists are increasingly using public social media data for research, and they are not just examining tweets – they also delve into your online dating profiles, your Yelp reviews, your Instagram posts, your YouTube videos, and even your comments on articles like this one.

The internet, especially social media, has provided researchers with access to a trove of information about human behaviour that is just there for the taking. And these researchers (like me!) have a lot of questions that they want to use that data to answer.

Consider Twitter, for instance. We can use it to predict flu trends and elections, help diagnose depression, understand the spread of misinformation, or improve emergency communication during crises.

Turkish writer and academic Zeynep Tufekci once called Twitter the "model organism" of social media research: researchers use Twitter data to answer questions because, like the fruit fly, the platform and its users are just so easy to study. Why is it so easy? Because Twitter data is almost entirely public.

"Public" is the magic word when it comes to research ethics. "But the data is already public." That was the response from Harvard researchers in 2008, when they released a data set of college students' Facebook profiles, and from Danish researchers in 2016, when they released a data set scraped from OKCupid.

The regulatory bodies that oversee research ethics (like institutional review boards at US universities) usually don't consider "public" data to be under their purview. Many researchers see these review boards as the arbiters of what is ethical; if it is not something that the boards care about, then it cannot be unethical, right?

Whether the data is public or not is important for ethical decision-making –in fact, it is necessary. (If you are weirded out by the idea of scientists collecting your public tweets, imagine how much worse it would be if they were collecting your posts in a closed Facebook group.)

The problem is that, for some researchers, whether the data is public is the only thing that matters. I suggest (sometimes loudly, to people who do not want to hear it) that it should not be.

I am part of a group of researchers, funded by the National Science Foundation, that is working on rethinking ethical guidelines for studies using social media data. We are interested in understanding things like scientists' attitudes and practices, laws and policies that govern this practice, ways of assessing and mediating risk, and the impact that scientific research has on the people whose data is being collected.

I have been digging most deeply into this last part, beginning with a study (published last year in *Social Media + Society*) that I conducted with my collaborator Nicholas Proferes. We surveyed Twitter users with a simple question in mind: How do they feel about researchers using their tweets, and what can this tell us about best practices for researchers?

First, a simple answer: most (almost two-thirds) of respondents had not realised that scientists might use their tweets at all. This is despite a fairly clear statement in Twitter's privacy policy indicating that possibility – but as we already know, most people don't read privacy policies. However, many – but not all – study participants indicated that their level of comfort and acceptance would depend on some specific factors.

These contextual factors are important to understand; they are things that researchers should be taking into account when deciding whether and how to collect and report public data. We are not interested in stopping research that uses public data, but we want to help researchers do it ethically.

Not all tweets are the same.

The idea that a tweet about what someone had for breakfast is exactly the same as a tweet revealing someone's sensitive health condition is, to put it bluntly, absurd. After all, one of the potential harms of using public data is amplification –spreading content beyond its intended audience. It probably won't do much harm if more people know about your breakfast cereal. But negative consequences might follow a tweet about someone having cancer, a sexually transmitted disease, or depression.

Amplification can happen through sharing data sets, or through reproducing content in a published paper. And even though academic papers might not be widely read, they can have a broader reach than expected. In at least one case, a journalist easily identified users from tweets quoted in a paper and then contacted the individuals. The subject matter of the tweets was pretty innocuous in this case – but what if it wasn't? What if, instead of finding out that their tweet about fishing had been quoted in a research article, a user discovered that the article had included their tweet about an sexually transmitted disease?

I'd like to think that most researchers are being thoughtful about this –that someone would not, for example, publish verbatim tweets that reveal something very sensitive about a user.

But if our sole ethical rule really is "Is it public?" then the content of the tweet does not matter. And, unsurprisingly, our study uncovered that respondents' comfort with researchers reading their posts depended a great deal on content.

You can argue that Twitter users know that they are putting their words out there for the entire world to see – but they still expect their audience to be primarily their own followers, not a room full of scientists, and certainly not the entire readership of *The New York Times*. I often hear, "Well, they should have known," but research has established that people don't have a good understanding of the reach of their social media data. As scientists, we should have a higher standard of care than that.

Not all tweeters are the same.

Scientists also have an ethical obligation to exercise a higher standard of care for people in more vulnerable positions, and this should extend to collecting data from potentially vulnerable groups in digital spaces.

My PhD student Brianna Dym and I have been working to develop best practices for studying online communities that have high numbers of queer and trans members (like online fandom). One important characteristic of these spaces is that many participants may not be out in their physical lives, and, as a result, amplifying users' content can bring very real safety concerns.

In one specific case, researchers collected YouTube videos of people going through gender transition in order to help train facial recognition algorithms. As a result, people's transition photos appeared in scientific papers without their permission or knowledge. Not only did these YouTube users not intend for their content to be used in this way, but its amplification has the potential to do more harm to an already vulnerable group.

An underlying issue is that many of the most popular support spaces –YouTube or Tumblr, for example – might be "public," but users don't actually intend for their content to be consumed by people outside their communities, and they only share personal or sensitive content to help fellow community members.

But social media platforms typically encourage as many eyes as possible on content, making it difficult for users to have control over their privacy; on many platforms, the choice is to make your content viewable to everyone or to simply not use that platform at all. Telling vulnerable users to just "not make it public" if they don't want their content taken out of context can mean cutting them off from important support spaces.

Not all research is the same.

Ethical rules that focus only on the characteristics of the data itself ignore the ethics of what we do with that data. Consider, for example, the algorithm created to detect someone's sexual orientation from a photograph. The training data for this algorithm came from profile photographs collected from online dating sites– images that were already labeled with sexual orientations. However, as acknowledged in the paper itself, a use case for this kind of algorithm could be the identification and persecution of people in a country where homosexuality is illegal.

Similarly, in the case of YouTube videos of gender transition, people whose images were used were understandably concerned that the content they made public in order to help their community would result in "spot the trans person" technology. As recent research has shown, automatic gender recognition technology carries disproportionate risk for trans people, and that many perceive it as actively harmful. Imagine finding out that your photograph was used to help create a technology that would harm your community.

This study also highlights another important characteristic of research: inference. It is one thing for your content to be amplified and therefore reveal what you have deliberately made public (like stating your sexual orientation on a dating site). This becomes something quite different when the research reveals something you did not make public.

Imagine a research paper that included selfies from Twitter labeled with the sexual orientation or the mental health diagnosis that the algorithm predicted for them. It is ludicrous to argue that what we scientists do with data is irrelevant to an ethical evaluation.

So what do we do?

This research shows why it is important to move beyond the notion that as soon as data can be labeled "public", all ethical obligations have been met and researchers can do whatever they like with it. When I suggest this to other researchers, they often respond with frustration. They want to do the right thing, but they also want straightforward rules and expectations. In reality, there are many contextual factors at play when evaluating the ethics of using public data. There is no mathematical formula. Like any ethical decision, there are some clear cases of right and wrong, but more often the boundaries are fuzzy.

When it comes to making the most ethical decisions we can in light of our constantly evolving media, methods, and opportunities, it is critical that we move beyond simplistic rules and consider each situation individually and holistically. Researchers cannot place the whole

burden on users; we cannot expect them to know that we are out there watching, nor demand that they anticipate any and all potential harms that may come to them now or in the future.

If you are a social media user concerned about unexpected uses of your data, my advice is not to shut down all your accounts and live in fear of invasions of privacy. The risks of harm are low for most people, and the majority of researchers are thoughtful and are already doing things to mitigate this risk. Moreover, many kinds of scientific research have the potential to do a lot of good, and maybe your tweets can help it along!

That said, it's important to keep in mind that your "public" data is truly available to the public – everyone from marketers to journalists can see it, and your tweet's appearance in a news article has considerably more reach than in an academic publication. Tweet as much as you like – but also tweet like BuzzFeed is watching.

As researchers, we have a responsibility to acknowledge that factors like the type of data, the creator of that data, and our intended use for the data are important when it comes to using public information. These factors must inform the decisions we make about whether and how to collect data and to report findings. I hope the work that my collaborators and I are doing will help to inform best practices, so that, in the end, we can continue to contribute great science to the world while also respecting the people who share their data with us every day.

This article first appeared on Medium's How We Get To Next.
