https://doi.org/10.1038/s43588-023-00490-7

Enhancing the ethics of user-sourced online data collection and sharing

Michelle N. Meyer, John Basl, David Choffnes, Christo Wilson & David M. J. Lazer

Social media and other internet platforms are making it even harder for researchers to investigate their effects on society. One way forward is user-sourced data collection of data to be shared among many researchers, using robust ethics tools to protect the interests of research participants and society.

Twitter's revocation of special academic access to its application programming interfaces (APIs) is the latest blow to the study of information sharing and consumption on the internet[1]. Platform APIs offer easy access to data, and Twitter is the modal source of online behavioral data — largely because of its generous APIs — to study everything from misinformation to the filter bubble.

Yet, while this is disastrous for the research community and others, such as journalists and various civil society actors, who relied on them, API-based (or platform-sourced) data collections were always quite limited: built for third-party app producers, not for research; often unreliable[2]; providing researchers little access to the central variable of interest (that is, what users actually see[3]); and subject to the whims of the platforms being studied. The recent batch of papers in *Science* and *Nature* involving a structured collaboration between academics studying the role of Facebook and Instagram in the 2020 election offers one powerful model for studying the internet[4]; however, it is notable that no company (even Meta) has committed to a similar effort in the future. The field needed new, independent paradigms for studying the internet long before the present-day retraction of APIs.

One such paradigm could be to consensually collect, for further sharing with many researchers, data from a large set of internet users, that is, user-sourced data collection. When a platform shows content to a user, that moment is experienced by only two actors, who are therefore the only two possible sources of these key data: the platform and the user. Platform-sourced data include not only APIs but also aggregate measures (such as those offered by CrowdTangle). Although the large majority of the literature is based on platform-sourced data, there is also a longstanding tradition by scholars and journalists of using bespoke user-sourced data.

A user-sourced dataset has important limitations relative to platform-sourced data: for instance, it will be smaller in scale, and recruited samples may be biased in ways not easily captured by sociodemographics and therefore corrected by oversampling. But it can also offer important benefits, especially if it goes beyond existing efforts in the field (targeted, bespoke data collections typically used by single research teams) to capture the breadth of individual online experiences. In principle, user-sourced data can be multi-platform, and linked to more standard, survey-based measures. Most importantly, the information extracted from user-sourced data can be constructed around the needs of researchers rather than the willingness of a platform to share data. Finally, if made broadly available in an ethical fashion, the costs could be amortized across many researchers and projects.

However, the regulatory framework that governs academic research in the USA (and in many other places) is poorly aligned with many of the challenges of large-scale collection and sharing of digital trace data. Supported by the National Science Foundation, we have been working on such an infrastructure: the National Internet Observatory (NIO). Our first task has been to build an ethics framework that aims to go well beyond what the current US regulatory framework requires.

The status quo for ethical regulation

The US Policy for Protection of Human Research Subjects — better known as 'the Common Rule' — is the primary framework governing research data. It has its origins in the 1970s, was intended to govern medical and interventional behavioral research, and has only been significantly overhauled once[5]. The Common Rule applies to federally funded research involving human 'subjects' — people with whom researchers intervene or interact, or about whom they obtain, analyze or generate identifiable private information. Recruiting people from whom to collect internet data easily falls within the Common Rule and is therefore subject to prospective review by an Institutional Review Board (IRB).

However, the Common Rule alone constitutes an inadequate ethical framework for a user-sourced internet data repository. Although it requires fairly comprehensive disclosure of a study's nature, risks and potential benefits, it does not require researchers to ensure that prospective participants comprehend that information. Moreover, IRBs typically only consider risks to participants (though some argue they should consider risks to non-enrolled individual bystanders as well as to groups and society at large: https://www.hhs.gov/ohrp/sachrp-committee/recommendations/tab-c-the-protection-of-non-subjects-from-research-harm.html). Similarly, the Common Rule directs that "[t]he IRB should not consider possible long-range effects of applying knowledge gained in the research (for example, the possible effects of the research on public policy)".

The Common Rule also does not consider downstream risks from sharing de-identified data. Once data have been stripped of direct participant identifiers and made available in a research repository, the Common Rule — including IRB oversight — ceases to apply entirely, because there are no longer any research subjects involved[6]. Users of a data repository neither intervene nor interact with the data subjects, and the data no longer meet the Common Rule's definition of private identifiable information. Notwithstanding reidentification risks, data are only 'identifiable' under the Common Rule if a subject's identity "is or may readily be ascertained by the investigator or associated with the information". Even if such data were identifiable, they must also be

'private', yet data that are explicitly collected for widespread sharing via a research repository do not obviously constitute "information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place", or "that has been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public". Indeed, secondary research on identifiable data is exempt from the Common Rule if it is publicly available (housed in an open repository, for example) or if the researchers analyzing the data avoid recording it in a way that makes identities "readily ascertainable".

Ethical challenges of building an internet data repository

The principles underlying the Common Rule as articulated in the Belmont Report — respect for persons, beneficence and justice — are hard to disagree with (https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf). But they require specification and balancing in particular contexts[7], and the context of widely sharing what one sees and does on the internet is not the context of, say, participating in a drug trial.

User-sourced online data collections present ethical challenges that obviously could not have been contemplated in the 1970s, and still are not adequately reflected in today's Common Rule. First, it is impossible to fully disclose in advance the nature, risks and potential benefits of the NIO, consistent with the principle of respect for persons, because some of these are moving targets. For instance, given constant engineering changes to platform websites and apps, a web scraper that collects certain information today may collect different — potentially more identifiable or sensitive — information tomorrow. Further, data collected that today expose no known privacy risks could implicate significant ones in the future. As an example, until recently, those who conducted internet searches or participated in social media conversations in the USA about how to terminate a pregnancy would have been unlikely to have engaged in criminal conduct, and so collecting data about these online activities would have entailed privacy, but not legal, risks to research participants or third parties with whom they converse. Very soon after the overturning of *Roe v. Wade*, however, some US state laws criminalized online sharing of information about how to obtain an abortion, creating a new risk for NIO participants. Nor can we anticipate in advance each study that might be conducted with NIO data. Second, given how information is embedded in human relationships, the data of human subjects almost always implicate the interests of third parties. For instance, understanding the information that algorithms show to different people requires us to study participants in the context of social networks rather than in isolation, which necessarily entails collecting data about other members of those networks.

The growing fields of AI, information and data ethics further highlight a wide range of risks associated with big data analytics that are relevant to projects like NIO, including vectors of bias and unfairness in how data are collected, structured and deployed and risks associated with aggregating data streams that might individually pose only minor risks[8–10]. These fields reveal how existing operationalizations of the Belmont principles are inadequate in emerging contexts (https://www.atlanticcouncil.org/in-depth-research-reports/report/specifying-normative-content)[11]. They also highlight risks that extend beyond direct data contributors to indirect data contributors and inference subjects, including non-traditional ways that applications of big data analytics can threaten privacy and security (https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf)[11] and entrench bias[10].

These fields may one day also offer solutions to manage these ethical challenges, and there are a growing number of frameworks, principles and suggested rules about how to manage ethical challenges in this space[12]. However, no consensus currently exists about how to realize core values[13]. Furthermore, these fields are still in their infancy, especially compared to more long-standing fields such as bioethics. Simply, there is neither formal regulation nor a robust ethics ecosystem that fills the gaps between the Common Rule and what is necessary to adequately manage the ethical challenges raised by a project such as the NIO.

The good news is that these challenges are not entirely new, which means that we are not starting from scratch in building the NIO's governance. First, there are already numerous repositories for depositing research data, including social media and other social science data for secondary research use, and many have developed governance structures for sharing these data for research and other uses[6]. Similarly, many longitudinal social science studies, like the NIO, both collect and store data for secondary research use. In addition, biobanks such as the NIH's All of Us Research Project (AoURP) are helpful analogs to repositories for donated internet data and considerable thought has been put into their governance structures. Biobanks collect and store — for wide, indefinite research use — biospecimens, from which genomic data are extracted, as well as survey and electronic health record data. Because all research uses of a biobank cannot be known in advance, participants generally give consent to have their samples and data studied for anything that falls within a broad category (such as 'human health and well-being'). Just as many people do not fully appreciate the nature and extent of their digital traces, many do not understand the implications of genomic information or even what is contained in their own health records, making meaningfully informed consent challenging in both cases. Because genetic variants are shared with both nearer-term relatives and members of larger genetic ancestral groups, the interests of many people besides individual participants are at stake, as is the case with social media and other internet research.

Not just scraping by: ingredients for ethical user-sourced digital trace data collection

Drawing on these existing projects, we have made a series of decisions, summarized in Fig. 1, for the NIO's ethical governance. We began by brainstorming internally, and with our external ethics advisory board, about all the ways the NIO could go ethically wrong — and right. Most concerns and aspirations fell into one of five 'ends', reflected in the first five columns of the table depicted in Fig. 1: the privacy of participants or bystanders could be compromised; NIO data could be misinterpreted or misused; researchers could be ethically negligent or even intentionally unethical; participants might make a decision about NIO enrollment, or continued enrollment, without appreciating its risks and potential benefits over time; and we could fail to be sufficiently transparent with all stakeholders, undermining trust and the viability of the project. Although these five ends can all be mapped to one or more of the broad, familiar principles of beneficence, non-maleficence, respect for persons and justice, these mid-level ends served as better guidance for the next stage. That stage entailed brainstorming a set of feasible ethics interventions to address each end (represented by the rows of the table depicted in Fig. 1). Core to this governance is a 'Swiss cheese' model: each individual intervention works imperfectly, but in combination, they are likely to reduce the odds of harm significantly and to increase the odds of realizing benefits. Most interventions serve

Core ends					Examples of NIO ethics interventions
	√				Oversample populations underrepresented in internet and/or social science research
			√		eConsent with teachback questions and a comprehension quiz
√			√		eConsent and NIO website provides instructions about how to manage privacy (e.g., using incognito mode)
			√		Regular participant reminders that they are enrolled in NIO
√	√		√	√	Participant engagement and feedback via surveys, focus groups, and/or a standing community advisory board
	√			√	Researchers have early access to NIO sample characteristics to appropriately gauge feasible research questions
√	√	√			Beta testing data access with small group of trusted faculty colleagues who provide feedback on vulnerabilities
		√			Institutional buy-in: institutional official must sign DUA and inform NIO if a researcher falls out of good standing
√	√	√			Required modular ethics training for researchers
√					Access is provided to individual researchers only and protected by security best practices
√	√	√			Standardized (not project-specific) data use agreement (DUA) countersigned by researcher institution, • No attempted re-identification • No attempt to exfiltrate, publish, or redistribute data • Immediately alert NIO of unexpected data privacy issue • No linking NIO data with outside or other NIO data without express written permission of NIO • Acknowledge consequences for violation (e.g., reporting to IRB/funder; expulsion from NIO)
√	√	√			Researcher application for analytic access to data: • Research question(s) • Description of data requested and justification for each data element • Analytic approach • List of all project personnel who would have access to the data/results • Potential benefits of the research • Description of risks of the project (referring to appropriate training modules) • Assessment of distribution of risks and potential benefits across different groups
√		√			Data remains on NIO servers
√	√				Review of source code, in some cases
√					Data access limited to specific project need
√					Query-return restrictions (e.g., hide results for fewer than *t* participants) in some cases
		√			Real-time passive monitoring of NIO data use to detect data exfiltration
			√	√	Continuously updated list of NIO-based research posted to public study website
√	√	√			Institutional reminders that specific faculty are active NIO users
		√		√	Random and for-cause audits: including manual review of individual monitoring logs, aggregate patterns of data access, and publications

Minimize privacy risks to participants and bystanders
Ensure scientifically and socially responsible use of data
Ensure researcher compliance
Respect participant autonomy
Promote transparency

Fig. 1 | NIO ethics interventions, mapped to the ends each serves. The rows show a non-exhaustive list of risk-reducing and value-enhancing measures taken by NIO (ethics interventions), in rough chronological order of when they apply during the life-cycle of a NIO project. Each ethics intervention is mapped to one or more of five ends (columns), which are in turn informed by the broader values of justice, beneficence, non-maleficence and respect for persons.

multiple ends and values, and each end and value is supported by multiple interventions.

Some interventions are participant-facing. For instance, we protect participant autonomy, starting with an electronic consent (eConsent) process, similar to that used by AoURP and mobile studies conducted via Apple's ResearchKit[14,15]. The eConsent prioritizes the most important information, limiting each consent screen to one key idea; those who want to learn more about that topic can do so by engaging a pop-up window. The eConsent contains teachback questions of key concepts: to participate, prospective participants must demonstrate they understand the project and its risks. Participants are also empowered to turn data collection off or on easily as needed.

Importantly, NIO consent is dynamic[16,17]: participants will be reminded over time that they have enrolled in the NIO and there will be a continuously updated list of NIO-based research posted to the study website to allow them to revisit their enrollment decision in light of new information about how their data are being used.

Most interventions, however, are researcher-facing. The core NIO team will, for instance, oversample populations that are traditionally underrepresented in internet or social science research. This helps to ensure both that the potential benefits of NIO research apply to a diverse population and that the privacy of traditionally underrepresented participants is protected by ensuring that their data are not more readily re-identifiable outliers in the NIO dataset.

With respect to secondary researchers, to minimize risks to the privacy of participants and individual bystanders, to ensure scientifically and socially appropriate use of the NIO datasets, and to ensure adherence to NIO policies, we will constrain both researcher access to, and use of, the NIO in several ways. First, only researchers who complete several ethics modules will be considered for access; these modules are developed to help sensitize researchers to potential ethical challenges associated with the use of the data. Researchers and their institutional official must also sign a plain language code of conduct and a data use agreement (DUA), respectively, that include a prohibition on attempted re-identification and a requirement to notify us promptly of any unexpected privacy issue, and that requires them to consent to a variety of potential consequences for violations, such as reporting to their IRB and/or funders and expulsion from the NIO. Institutional officials will also receive periodic reminders that specific researchers are active NIO users and will be expected to notify the NIO if the researcher has fallen out of good standing.

Researchers must also submit a form that, in addition to providing standard protocol information, describes the study's risks and potential benefits to all stakeholders (not just participants) and assesses how those risks and potential benefits are distributed across different individuals and groups. Data access will be limited to what is needed for approved projects. Importantly, data will never leave the NIO servers; instead, researchers will bring their code to the servers. NIO will engage in real-time passive monitoring to detect data exfiltration as well as random and for-cause audits of specific projects.

The interventions we describe here currently apply to all uses of the NIO datasets. Over time, we will assess the risks of different kinds of data access and use and consider creating lower-access tiers, in which some of these constraints are relaxed. In addition to starting out with a single, controlled access tier, we will beta test this data access process with a small group of trusted researchers who are sensitive to data privacy and data misuse to help us identify flaws in our process. Although developed specifically for NIO, we believe that many of the ethics interventions we have adopted and the ends they are intended to serve (which in turn reflect the broader, familiar values of justice, beneficence, non-maleficence and respect for persons) will generalize to similar efforts.

Conclusion

The internet research community is at a precarious point, with the destruction of many of the core tools of data collection of the field. The development of user-sourced data collection to study the tech platforms that dominate the contemporary internet is a necessary next step for science, and for global society. This is the Hubble or the CERN of computational social science: success would have dramatic implications for the capacity of the field. A paper such as ref. 18, which presents a multi-year audit of how Google presents information to people regarding politics, had a bespoke data collection, requiring thousands of hours of labor and many research dollars. A shared data collection, such as the NIO, would allow such a paper to be produced with far more data, at a tiny fraction of the cost. And while the NIO is focused on US users only, we envision federated efforts, embedded in the legal and cultural-ethical contexts of other countries, that would allow for even more ambitious cross-national research. This will require substantial extensions to the governance of ethical research that was built in the late twentieth century. The elements we summarize above are certainly incomplete, but also, we think, contain some useful and even necessary pieces of a new self-regulatory approach to ethically researching human behavior on the internet using shared infrastructure.

Michelle N. Meyer[1,2], John Basl[3,4], David Choffnes[5,6], Christo Wilson[5,6] & David M. J. Lazer[5,7,8,9]

[1]Department of Bioethics and Decision Sciences, Geisinger Health System, Danville, PA, USA. [2]Behavioral Insights Team, Steele Institute for Health Innovation, Geisinger Health System, Danville, PA, USA. [3]Department of Philosophy and Religion, Northeastern University, Boston, MA, USA. [4]Ethics Institute, Northeastern University, Boston, MA, USA. [5]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [6]Cybersecurity and Privacy Institute, Northeastern University, Boston, MA, USA. [7]College of Social Sciences and Humanities, Northeastern University, Boston, MA, USA. [8]Network Science Institute, Northeastern University, Boston, MA, USA. [9]The Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA.

e-mail: d.lazer@northeastern.edu

Published online: 27 July 2023

References

1. Freelon, D. *Political Communication* **35**, 665–668 (2018).
2. Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. *Proc. Int. AAAI Conf. on Web And Social Media* **7**, 400–408 (2013).
3. Lazer, D. *Proc. Natl Acad. Sci.* **117**, 21–22 (2020).
4. González-Bailón, S. et al. *Science* (in the press).
5. Meyer, M. N. *J. Law Med. Ethics* **48**, 60–73 (2020).
6. Meyer, M. N. *Adv. Meth. Practices Psychol. Sci.* **1**, 131–144 (2018).
7. Beauchamp, T. L. & Childress, J. F. *Principles Of Biomedical Ethics* 8th edn (Oxford Univ. Press, 2019).
8. Crawford, K. & Schultz, J. *Boston College Law Rev.* **55**, 93–128 (2014).
9. Vayena, E. & Madoff, L. In *The Oxford Handbook Of Public Health Ethics* (eds Mastroianni, A. C., Kahn, J. P. & Kass, N. E.) 354–366 (Oxford Univ. Press, 2019).
10. Fazelpour, S. & Danks, D. *Phil. Compass* **16**, e12760 (2021).
11. Barocas, S. & Nissenbaum, H. In *Privacy, Big Data, And The Public Good* (eds Lane, J., Stodden, V., Bender, S. & Nissenbaum, H.) 44–75 (Cambridge Univ. Press, 2014).
12. Zook, M. et al. *PLOS Comput. Biol.* **13**, e1005399 (2017).
13. Jobin, A., Ienca, M. & Vayena, E. *Nat. Mach. Intell.* **1**, 389–399 (2019).
14. Wilbanks, J. Design issues in e-Consent. *J. Law Med. Ethics* **46**, 110–118 (2018).

15. Doerr, M. et al. *AJOB Empir. Bioethics* **12**, 72–83 (2021).
16. Kaye, J. et al. *Eur. J. Hum. Genet.* **23**, 141–146 (2015).
17. Budin-Ljøsne, I. et al. *BMC Med. Ethics* **18**, 4 (2017).
18. Robertson, R. E. et al. *Nature* **618**, 342–348 (2023).

Acknowledgements
For very helpful conversations, we thank the members of NIO's Ethics Advisory Board: M. Doerr, N. Kass, J. McNealy, A. Rubel, E. Vayena and P. Williams. This material is based upon work supported by the National Science Foundation under grant number 2131929 (PIs D.L., C.W. and D.C.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Author contributions
D.L., D.C. and C.W. conceived and acquired funding for the project described here (NIO). M.N.M. and J.B. conceived the ethics framework for NIO. M.N.M., J.B. and D.L. wrote the first draft. All authors read, revised and approved the paper.

Competing interests
The authors declare no competing interests.

Additional information
Peer review information *Nature Computational Science* thanks Katie Shilton, Sarah Gilbert, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.