Analyzing the Ethical Implications of Research Using Leaked Data

Anne E. Boustead, *University of Arizona*
Trey Herr, *Atlantic Council Cyber Statecraft Initiative*

ABSTRACT Although information made public after a data breach can provide insight into difficult research questions, use of these data raises ethical questions not directly addressed by current ethical guidelines. This article develops a framework for identifying and managing risks to human subjects when conducting research involving leaked data. We contend that researchers who seek to use leaked data should identify and address ethical challenges by considering the process through which the data were originally released into the public domain.

Information made public after a data breach—that is, a security failure resulting in an institution's confidential information being accessed by a third party (Romanosky 2016)—can provide crucial insight into elusive but important questions. For example, WikiLeaks' release of diplomatic cables presented novel information about US foreign policy and its machinations (Roberts 2012) and the Paradise Papers, which disclosed financial dealings of political and business elites (Shaxson 2018), are both useful for studying influence networks.

However, researchers who seek to use leaked information must first address a difficult issue: Can unethically produced data be used ethically? Although it may seem that researchers should be free to use publicly available information, use of information released through illicit activity without the permission of the data owner or subjects should be subject to heightened ethical scrutiny. While other professions have addressed this issue (Jamieson 2019), political science has not yet reckoned with whether leaked data can be used within the discipline's ethical guidelines.

By presenting a framework and recommendations for the ethical use of leaked data in research, this article provides guidance to investigators and Institutional Review Boards (IRBs) tasked with identifying and analyzing these ethical concerns. We base this analysis on our interpretation of the ethical obligations placed on researchers by the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979) and the Common Rule, a uniform federal policy on the ethical treatment of human subjects (45 CFR §46.101 et seq. 2018). However, from the beginning, it is worth noting that both sets of guidelines have been criticized, particularly because they may fail to account for ways in which research can exacerbate harms faced by vulnerable communities (Rogers and Kelly 2011) or paternalistically deny these communities opportunities to participate in research (Gustafson and Brunger 2014). Consequently, the framework developed in this article should be viewed as a tool to help researchers grapple with the ethical issues posed by using leaked data in research rather than a universal and inviolate set of rules.

ETHICAL FRAMEWORK FOR USING LEAKED DATA IN RESEARCH

In the United States,[1] the foundational principles of ethical human research derive from the Belmont Report. These principles—respect for persons, beneficence, and justice—are implemented through the Common Rule. The Menlo Report (Dittrich and Kenneally 2012) described how to apply these principles to information-technology research and articulated a fourth principle: respect for law and public interest (Garfinkel and Cranor 2010). Researchers comply with these principles by obtaining IRB approval before conducting research involving human subjects and adhering to a series of specific safeguards. Respect for persons mandates informed consent and special protections for vulnerable populations; beneficence requires that researchers maximize benefits and minimize harms to subjects; justice requires that subjects be selected equitably and that the risks and benefits of research are evenly distributed throughout society. The respect-for-law principle articulated in the Menlo Report requires transparent and accountable research conducted in compliance with relevant laws (Dittrich and Kenneally 2012, 12). Researchers also should consider the broader ethical implications of their research, including its impact on vulnerable communities (Hoffmann and Jonas 2017).

Application of these safeguards to leaked data is not clear. Previous discussions determined that online publication does not make data public for research-ethics purposes because "[p]rivate data that was obtained through illicit means (e.g., data stolen in an intrusion incident) and put on a public website is still private data"

Anne E. Boustead *is assistant professor of government and public policy at the University of Arizona. She can be reached at boustead@arizona.edu.*
Trey Herr *is director of the Atlantic Council Cyber Statecraft Initiative. He can be reached at therr@atlanticcouncil.org.*

doi:10.1017/S1049096520000323

(Egelman et al. 2012, 6). A careful analysis of the ethical issues involved in conducting political science research using leaked data, using WikiLeaks as an example, concluded that "political scientists should use leaked information because it is uniquely valuable and offers insights that would otherwise be unavailable" (Michael 2015, 177). Conversely, social scientists who decided not to use leaked data in other contexts explained that they made this decision because "the negatives outweighed the positives, especially when [they] could gather all or most of the same data in a more legal and more accepted manner" (Poor 2017; Poor and Davidson 2016, 5).

By presenting a framework and recommendations for the ethical use of leaked data in research, this article provides guidance to investigators and Institutional Review Boards (IRBs) tasked with identifying and analyzing these ethical concerns.

The significance of leaked data is particularly acute for political scientists and historians whose access to the historical record often is skewed by secrecy; however, policies governing its use are ambiguous. After Cablegate and the release of more than 250,000 secret documents into the public domain, *International Studies Quarterly* created a provisional policy to disqualify submissions that used leaked documents "if such use could be construed as mishandling classified material" (Michael 2015, 176). This type of policy would undermine research primarily derived from access to once-secret information. Michael (2015) leveraged documents from Cablegate to review the Trans-Pacific Partnership negotiations, exploring gaps between public and private US positions on key issues of intellectual property. The empirical richness and rigor of this research would not have been possible without access to this type of otherwise-secret material.

In addition to the larger rubric developed in this article, political scientists considering the use of leaked data in their work should consider two questions more specific to their field: (1) Within a cache of leaked documents, are the materials germane to the core question of the political science research project or better categorized as organizational context and atmospherics?; and (2) Do leaked documents address a policy maker's state of mind? One of the most challenging dimensions of political science research is establishing clear causality between information available to a decision maker and a political outcome; therefore, leaked data can provide unique value.

Furthermore, political science is not the only field grappling with the use of either leaked data or unethically obtained data more generally. In articulating their ethical principles, information-security researchers identified the use of leaked data as an area of ethical concern (Schrittwieser, Mulazzani, and Weippl 2013). More generally, medical ethicists have long struggled with the use of data generated by unethically conducted studies, particularly in the context of tragedies such as the Nazi experiments and the Tuskegee syphilis study. A survey of physicians, medical scientists, and bioethicists investigated whether respondents would be willing to use data from unethical experiments. It found that a majority of respondents would be willing to use the data if they were generated through scientifically valid processes and were necessary to preserve life (Halpin 2010). Some scholars have hesitated to conclude that data from Nazi experiments can be used ethically; one argued that their use is a "grave profanity under all circumstances" (Post 1991, 43). However, other scholars would allow their use under limited circumstances when necessary to preserve life (Steinberg 2015).

Because data leaks have become common, the time is suitable for a comprehensive discussion of the ethics of research that uses leaked data. This section describes a framework for considering whether and under which circumstances the safeguards commonly used to protect individuals involved in research are implicated by use of a leaked dataset. This discussion focuses on two broad questions: Is IRB review required? If so, should the IRB grant approval?

Is IRB Review Required?

As a threshold matter, the Common Rule's ethical requirements apply only to research on human subjects conducted at federally funded institutions. "Human subject" includes living persons who researchers interact with or about whom they obtain identifiable private information. The Common Rule also exempts research based on existing records that are publicly available or about unidentifiable persons (45 CFR §46.104. 2018). It is not clear whether leaked data about identifiable persons fall within these exceptions. Under the Common Rule, information is considered private if it is collected for a narrowly defined purpose under confidential conditions (45 CFR §46.102. 2018). Although publicly released, leaked data nevertheless may have been collected under circumstances within the Common Rule's understanding of "private" (Egelman et al. 2012, 6). Indeed, the Menlo Report specifically cites "data captured by malicious actors recording online financial transactions in order to commit fraud" as an example of private information (Dittrich and Kenneally 2012, 4).

In general, researchers should rely on their IRB to determine whether ethical review of their proposed research is required (King and Sands 2015). We recommend that this decision should be made in consideration of three questions. First, reviewers should ask whether the data contains identifying information, such as a name, address, telephone number, email address, and IP address (Garfinkel and Cranor 2010). Second, because leakers may release information slowly over time, reviewers should engage in an ongoing inquiry about whether the dataset could reveal identifiable information if linked with another dataset. Third, if the dataset includes identifiable information, reviewers should consider whether it includes information that an "individual can reasonably expect will not be made public" (Dittrich and Kenneally 2012, 4). Fourth, reviewers should be extraordinarily hesitant to conclude that a leaked dataset containing private identifiable information is now public information; they should err on the side of caution in deciding whether to require formal review.

Should the IRB Grant Approval?

After researchers refer a project to their IRB, the IRB must determine whether the proposed project complies with the

https://doi.org/10.1017/S1049096520000323 Published online by Cambridge University Press

Common Rule. This section discusses how use of leaked data may complicate this determination.

Is Informed Consent Possible?

Researchers generally must obtain informed consent before conducting research that involves people. Informed consent requires that individuals agree to participate in research after being notified of potential risks (45 CFR §46.116. 2018). Although informed consent is a "canonical principle" in research ethics (Lysaught 2004, 668), it may be waived if the proposed research poses minimal risk and would be impossible without a waiver (45 CFR §46.116 2018). The Common Rule also allows researchers to proceed without informed consent when working with publicly available secondary data or public behavior (45 CFR §46.111(a)(7). 2018).

Researchers seeking to use private, identifiable data typically should obtain informed consent from the data subjects—even if the data were leaked to the public. Researchers who cannot feasibly obtain consent may proceed only after demonstrating that the research poses minimal additional risk of harm to subjects. In deciding whether this standard has been met, IRBs should focus on the type of data leaked, specific vulnerabilities of those from whom it was collected, and how it was made public. IRBs also should contemplate whether any cleaning or coding done could put subjects at further risk by making the data more available. Finally, IRBs should consider whether subjects could be harmed by the consent process by causing potential anxiety if they previously were unaware that their data had been leaked.

Does the Research Maximize Benefits and Minimize Harms?

Under the Common Rule, an IRB can approve research only if the "[r]isks to subjects are reasonable in relation to anticipated benefits, if any, to subjects, and the importance of the knowledge that may reasonably be expected to result" (45 CFR §46.111. 2018). This analysis often is conducted by reference to risks incurred in normal daily activities and is limited to direct rather than theoretical long-term benefits. These risks and benefits should be evaluated using an objective "reasonable-researcher" standard (Dittrich and Kenneally 2012, 9).

Research using leaked data requires a risk/benefit calculus that reflects how the data were created and made public. Because use of intimate information creates greater potential for harm, researchers should consider how leaked data were collected to determine whether they are private (e.g., during a doctor's visit). The particular needs of data subjects should be carefully considered because the sensitivity of information varies across communities.

The potential benefit of research in part depends on the data quality and analysis used. "Research that uses biased samples, questions, or statistical evaluations…cannot generate valid scientific knowledge and is thus unethical" (Emanuel, Wendler, and Grady 2000). However, the data-leak process may introduce uncertainty about the quality of leaked data; for example, it may be difficult to determine whether the data have been selectively released.

Therefore, researchers proposing to use leaked data first should investigate how the data were generated, collected, and leaked. If this investigation raises concerns about potential incompleteness or inaccuracies, then researchers should explain how they will account for this uncertainty. Second, researchers should question whether the compromising party had incentive to alter the leaked data and explore whether potential alterations could be identified—perhaps by comparing different versions of purportedly identical documents (Groll 2016). Third, researchers should consider whether additional bias was introduced when the leaked data were publicly disseminated, especially when released by a third party with their own motives for altering the data. This analysis may require careful investigation of the circumstances underlying release of the data because leakers may go to great lengths to hide their identity and, consequently, their partisanship. For instance, after compromising US sporting agencies and political parties, Russian intelligence organizations used WikiLeaks as a conduit to obscure themselves and place the data in the public domain without as much perceived geopolitical bias (Nakashima and Harris 2018).

Are Subjects Selected Equitably?

Historically, research subjects were chosen because they were easily manipulatable, placing the cost of the research on those unlikely to receive benefits (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979). Accordingly, ethical guidelines now require that subjects are chosen equitably based on the purpose and setting of the research (45 CFR §46.111(a)(2). 2018). Projects involving populations vulnerable to coercion (e.g., prisoners) are subject to higher scrutiny (45 CFR §46.200 et seq. 2018).

Researchers using leaked data should be careful not to reinforce inequitable data collection. If certain individuals are more likely to have their data leaked because of an inability to opt out of data collection or their use of services with poor security practices, then these individuals can be expected to disproportionately bear the costs of research conducted using leaked data. Researchers therefore should consider whether the persons whose data were leaked will benefit from the research and question whether the method of compromise led to the selective release of data.

Finally, it is likely inappropriate to use leaked data pertaining only to members of a disadvantaged community or to conduct research that targets such a community. For example, using leaked communications data to train an algorithm to identify LGBTQ persons from their writing should raise serious ethical flags, even if this process uses data from both LGBTQ and non-LGBTQ people.

Our recommendations for the ethical use of leaked data for research overlap significantly with widely accepted best practices: investigators should consider carefully the data-generating process and resulting biases, and they should obtain permission from individuals before using their data.

https://doi.org/10.1017/S1049096520000323 Published online by Cambridge University Press

Does the Research Protect Vulnerable Populations?

Ethical research should ensure the safety and well-being of vulnerable populations. Populations may be vulnerable if their members are less able to freely consent or they face particularized harms. This vulnerability can be perpetuated online because members of these communities may be subject to abuse more frequently (Hoffmann and Jonas 2017). To ensure that information was voluntarily provided, researchers who seek to use leaked data involving vulnerable populations should consider carefully how these data were collected. Furthermore, the impact of research on vulnerable communities should be analyzed from the perspective of those communities because the risks imposed can be difficult to fully anticipate. Consulting with members of vulnerable communities can ensure that their interests are protected (Rencher and Wolf 2013).

Is the Research Transparent and Accountable?

The Menlo Report identifies transparency and accountability as ethically necessary because they build trust by "avoid[ing] guesswork and incorrect references about whether, where, and how ethical principles are addressed" (Dittrich and Kenneally 2012, 12). Consequently, researchers who use leaked data should discuss and justify this choice rather than obscure it. In particular, they should describe how the data were obtained, compromised, and released, as well as from where the researcher obtained the data. Researchers should be prepared to explain how these factors could bias their analysis and justify the steps taken to verify the data's authenticity.

Were Any Laws Violated?

The Menlo Report also recommends that researchers consider relevant laws, including those related to privacy and data-breach notification (Dittrich and Kenneally 2012). Researchers using leaked data should consider which laws, if any, were violated during the data leak and be prepared to justify their use of the data in light of this determination. Because legal requirements vary by context and jurisdiction, researchers must be aware of where the data were collected, compromised, and leaked to conduct this analysis.

CONCLUSION

Our recommendations for the ethical use of leaked data for research overlap significantly with widely accepted best practices: investigators should consider carefully the data-generating process and resulting biases, and they should obtain permission from individuals before using their data. However, researchers using leaked data must confront ethical issues that rarely arise with traditional data sources. Purportedly leaked data may have been stolen and made public against the will of the data collector and the expectations of the data subjects; it also may be wholly or partially fabricated by difficult-to-identify entities with inscrutable motives. It is possible to ethically conduct research using leaked data; however, the decision to use these data must be made after systematically analyzing how they were made public. Emerging data sources currently available to researchers pose new challenges, but these challenges can be met using existing ethical tools.

ACKNOWLEDGMENTS

The authors thank participants at the 2017 Privacy Law Scholars Conference and the 2017 Law and Society Association Annual Meeting for their feedback. ■

NOTE

1. Comparable guidelines are in place in many other countries (Cleaton-Jones and Wassenaar 2010; Millum 2012).

REFERENCES

45 CFR §46.101 et seq. 2018.

45 CFR §46.102. 2018.

45 CFR §46.104. 2018.

45 CFR §46.111. 2018.

45 CFR §46.111(a)(2). 2018.

45 CFR §46.111(a)(7). 2018.

45 CFR §46.116. 2018.

45 CFR §46.200 et seq. 2018.

Cleaton-Jones, Peter, and Doug Wassenaar. 2010. "Protection of Human Participants in Health Research: A Comparison of Some US Federal Regulations and South African Research Ethics Guidelines." *South African Medical Journal* 100 (11): 710–16.

Dittrich, David, and Erin Kenneally. 2012. "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research." Washington, DC: US Department of Homeland Security.

Egelman, Serge, Joseph Bonneau, Sonia Chiasson, David Dittrich, and Stuart Schechter. 2012. "It's Not Stealing If You Need It: A Panel on the Ethics of Performing Research Using Public Data of Illicit Origin." *International Conference on Financial Cryptography and Data Security*. Berlin: Springer.

Emanuel, Ezekiel J., David Wendler, and Christine Grady. 2000. "What Makes Clinical Research Ethical?" *Journal of the American Medical Association* 283 (20): 2701–11.

Garfinkel, Simson L., and Lorrie Faith Cranor. 2010. "Institutional Review Boards and Your Research." *Communications of the ACM* 53 (6): 38–40.

Groll, Elias. 2016. "Whoopsie: Russian Hackers Post Same Document Twice, but with Glaring Differences." *Washington Post*, August 24.

Gustafson, Diana L., and Fern Brunger. 2014. "Ethics, Vulnerability, and Feminist Participatory Action Research with a Disability Community." *Qualitative Health Research* 24 (7): 997–1005.

Halpin, Ross W. 2010. "Can Unethically Produced Data Be Used Ethically?" *Medicine and Law* 29:373–87.

Hoffmann, Anna Lauren, and Anne Jonas. 2017. "Recasting Justice for Internet and Online Industry Research Ethics." In *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts*, ed. Michael Zimmer and Katharina Kinder-Kurlanda, 3–18. New York: Peter Lang Publishing.

Jamieson, Kathleen Hall. 2019. "What Should the Press Learn from its Use of Russian Hacked Content?" *Boston Globe*, April 23.

King, Gary, and Melissa Sands. 2015. "How Human Subjects Research Rules Mislead You and Your University, and What to Do about It." Unpublished manuscript. Available at https://gking.harvard.edu/files/gking/files/irb_politics_paper_1.pdf.

Lysaught, M. Therese. 2004. "Respect: Or, How Respect for Persons Became Respect for Autonomy." *Journal of Medicine and Philosophy* 29 (6): 665–80.

Michael, Gabriel J. 2015. "Who's Afraid of WikiLeaks? Missed Opportunities in Political Science Research." *Review of Policy Research* 32 (2): 175–99.

Millum, Joseph. 2012. "Canada's New Ethical Guidelines for Research with Humans: A Critique and Comparison with the United States." *Canadian Medical Association Journal* 184 (6): 657–61.

Nakashima, Ellen, and Shane Harris. 2018. "How the Russians Hacked the DNC and Passed its Emails to WikiLeaks." *Washington Post*, July 13.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. "The Belmont Report." Washington, DC: US Department of Health and Human Services.

Poor, Nathaniel. 2017. "The Ethics of Using Hacked Data: Patreon's Data Hack and Academic Data Standards." In *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts*, ed. Michael Zimmer and Katharina Kinder-Kurlanda, 277–80. New York: Peter Lang Publishing.

Poor, Nathaniel, and Roei Davidson. 2016. "*The Ethics of Using Hacked Data: Patreon's Data Hack and Academic Data Standards.*" New York: Data & Society.

Post, Stephen G. 1991. "The Echo of Nuremberg: Nazi Data and Ethics." *Journal of Medical Ethics* 17 (1): 42–44.

Rencher, William C., and Leslie E. Wolf. 2013. "Redressing Past Wrongs: Changing the Common Rule to Increase Minority Voices in Research." *American Journal of Public Health* 103 (12): 2136–40.

Roberts, Alasdair. 2012. "WikiLeaks: The Illusion of Transparency." *International Review of Administrative Sciences* 78 (1): 116–33.

Rogers, Jamie, and Ursula A. Kelly. 2011. "Feminist Intersectionality: Bringing Social Justice to Health Disparities Research." *Nursing Ethics* 18 (3): 397–407.

Romanosky, Sasha. 2016. "Breach Notification Laws: The Policy and Practice." In *Cyber Insecurity: Navigating the Perils of the Next Information Age*, ed. Richard Harrison and Trey Herr, 137–54. Lanham, MD: Rowman & Littlefield.

Schrittwieser, Sebastian, Martin Mulazzani, and Edgar Weippl. 2013. "*Ethics in Security Research: Which Lines Should Not Be Crossed?*" IEEE Security and Privacy Workshops.

Shaxson, Nicholas. 2018. "How to Crack Down on Tax Havens: Start with the Banks." *Foreign Affairs* (March/April).

Steinberg, Jonathan. 2015. "The Ethical Use of Unethical Human Research." Unpublished manuscript. Available at https://pdfs.semanticscholar.org/9b93/0723bd3f63deeac142fb2b930dbf3725dc77.pdf.

https://doi.org/10.1017/S1049096520000323 Published online by Cambridge University Press