# Ethics of the Use of Social Media as Training Data for Artificial Intelligence Models used for Digital Phenotyping: Commentary

Aditi Jaiswal, Aekta Shah, Christopher Harjadi, Erik Windgassen, Peter Washington

Submitted to: JMIR Formative Research
on: April 22, 2024

# *Table of Contents*

# Ethics of the Use of Social Media as Training Data for Artificial Intelligence Models used for Digital Phenotyping: Commentary

Aditi Jaiswal[1] MS; Aekta Shah[2] PhD; Christopher Harjadi[3]; Erik Windgassen[4]; Peter Washington[1] PhD

[1]University of Hawaii Honolulu US
[2]Salesforce San Francisco US
[3]University of California, Berkeley Berkeley US
[4]University of California, Riverside Riverside US

**Corresponding Author:**
Peter Washington PhD
University of Hawaii
1680 East-West Road
Honolulu
US

## *Abstract*

(to be updated by authors)

(JMIR Preprints 22/04/2024:59794)
DOI: https://doi.org/10.2196/preprints.59794

### Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

# Ethics of the Use of Social Media as Training Data for Artificial Intelligence Models used for Digital Phenotyping: Letter to Editor

Aditi Jaiswal[1], Aekta Shah[2], Chris Harjadi[3], Erik Windgassen[4], Peter Washington[1]

[1]Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI, United States
[2]Salesforce, San Francisco, CA, United States
[3]Department of Computer Science, University of California, Berkeley, Berkeley, CA, United States
[4]University of California, Riverside

Community-based participatory research and human-centered design are central to research that aims to advance health equity [1]. While participatory design is a well-known framework that is increasingly though not yet widely used for research in areas such as interventions development [2] and partnered science, there is a dearth of research that builds artificial intelligence (AI) models for health in a way that is grounded in community-based principles. The lack of community guidance early in the AI development process may lead, inadvertently, to models that are unethical despite being formally approved by an Institutional Review Board (IRB). We discuss, in particular, the topic of consent, which we argue spans at least two parts of the AI development process: (1) consent to build the AI model, which can be determined through participatory design sessions with the community that the AI model is meant to serve, and (2) consent to use an individual's data within the training process of the model, which can be obtained through explicit consent procedures.

We discuss these gaps in community-based research for AI, with a particular focus on the development of social media-based screening tools for underserved communities, especially neurodiverse populations. Using social media for the quantification of characteristics or traits of an individual is a form of digital phenotyping, a method that can work with a broad range of data sources [3]. While the increasing availability of public data trails on social media can lead to predictive models that are possibly useful for creating positive good for health outcomes, the unrestricted use of these data poses the risk of training machine learning models on user-generated content without the explicit consent of the people who generated the data. Furthermore, the release of such models has the potential to lead to unintended consequences and possibly harm.

Social media platforms have emerged as a popular data source for several research domains, including for screening and surveillance broadly in psychiatry and behavioral sciences [4-6], sometimes with the help of AI. Government agencies such as the National Institutes of Health (NIH) encourage research that uses existing data streams, including social media, to provide actionable insights for conditions such as substance use [4-7]. However, several thought leaders are noting that such research must be careful to not scrape data from the Internet without the consent of the end users [8]. Some recent papers in social media analytics have been careful to obtain explicit consent from users participating in the study or to only conduct analysis on anonymized data feeds. The NIH has started to prioritize funding research that addresses these ethical challenges [4-8]. The White House in late 2023 has also recently highlighted the need for ethical AI practices via its list of Voluntary AI Commitments created for companies [9], but with guidelines such as prioritizing "research on societal risks posed by AI systems" and protecting privacy that are highly relevant to non-commercial research as well.

This conversation intersects strongly with the discourse around the training procedures of large language models, many of which have been trained on web data without user consent. Over the last few years, generative AI has revolutionized the field of AI by demonstrating remarkable capabilities, from generating human-like text to creating art and music. These models require massive amounts of pre-training data collected from various public forums. However, there have been numerous examples of popular language models being trained with web data without explicit user consent or consent that was hidden away in terms and conditions. For example, users were concerned that

Google was famously suspected of training Bard/Gemini using Gmail data without consent from end users, though Google claims that they did not do so. Similarly, OpenAI has trained ChatGPT using data from user's conversation histories. These cases raise questions about how our social contracts may have changed and what users inadvertently opt for when signing up on social media. Although OpenAI has provided the option to opt out of data retention, the default opt-in option raises privacy and data concerns.

The issue of data consent is particularly salient for vulnerable and marginalized groups. There are several instances of well-known misuse of data for scientific purposes. HeLa cells, named after Henrietta Lacks, are well known in the field of biology and have contributed greatly to progress in science. However, HeLa cells were commercialized, leading to financial gains without compensation or even an acknowledgement of Henrietta Lacks' contributions. Another notable example is the historical misuse of Indigenous DNA through repeated lack of informed consent by members of Indigenous populations.

In light of these reflections and the evolving discussions around AI ethics, we have elected to make some significant amendments to our recently published Twitter analysis paper on the use of the #ActuallyAutistic hashtag on Twitter for training a machine learning model that could serve as a screening tool for autism [10]. This paper serves as an example of what is possible with AI and social media in today's tech ecosystem, and we provide a word of caution for creators of such models to think through how such models may be misused and interpreted by the community that they were built to serve. Models meant to help the autistic community should be built in collaboration with the community from the onset of the ideation and development process or be led by autistic individuals. We hope that our decision to delete our dataset and model can serve as a template for other researchers.

We would like to highlight two important closing thoughts. First, approval by an IRB does not necessarily translate to an ethical study. Some institutions are creating ethical review boards to provide an additional layer of ethical review of studies. Second, while many areas of health-related research are guided by community-based participatory principles, such practices are not as commonplace in research at the intersection of health, social media, and AI. Speaking with impacted communities helps verify assumptions and provides input into methods design and analysis, leading to more robust conclusions for future research.

**References**

1. Brewer LC, Fortuna KL, Jones C, et al. Back to the Future: Achieving Health Equity Through Health Informatics and Digital Health. *JMIR Mhealth Uhealth*. 2020;8(1):e14512. Published 2020 Jan 14. doi:10.2196/14512
2. Kumar N, Dell N. Towards Informed Practice in HCI for Development. Proc ACM Hum-Comput Interact. 2018 Nov;2(CSCW):99. doi:10.1145/3274368.
3. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*. 2021;20(3):318-335. doi:10.1002/wps.20883
4. National Institutes of Health (NIH). Notice of Special Interest (NOSI): Computational and Statistical Methods to Enhance Discovery from Health Data (NOT-LM-23-001). Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-LM-23-001.html. Accessed [April 2024].
5. National Institutes of Health (NIH). Notice of Special Interest (NOSI): Addressing Health Disparities in NIMHD Research: Leveraging Health Data Science (NOT-OD-22-026). Available from: https://grants.nih.gov/grants/guide/notice-files/not-od-22-026.html. Accessed [April 2024].
6. National Institutes of Health (NIH). Notice of Special Interest (NOSI): IDEA2Health: Innovative Data Evaluation and Analysis to Health (NOT-HL-22-001). Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-HL-22-001.html. Accessed [April 2024].

7. National Institutes of Health (NIH). Notice of Special Interest (NOSI): Leveraging Data Science to Bring Actionable Insights for Substance Use Prevention and Treatment (NOT-DA-23-006). Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-DA-23-006.html. Accessed [April 2024].
8. Ahmed W, Bath PA, Demartini G. Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. In: Woodfield K, ed. The Ethics of Online Research (Advances in Research Ethics and Integrity, Vol. 2). Emerald Publishing Limited; 2017:79-107. doi:10.1108/S2398-601820180000002004
9. The White House. Voluntary AI Commitments. September 2023. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf. Accessed [April 2024].
10. Jaiswal A, Washington P. Using #ActuallyAutistic on Twitter for Precision Diagnosis of Autism Spectrum Disorder: Machine Learning Study. *JMIR Form Res*. 2024;8:e52660. Published 2024 Feb 14. doi:10.2196/52660.