DOI:10.1145/3625254

Sauvik Das, Hao-Ping (Hank) Lee, and Jodi Forlizzi

Privacy

Privacy in the Age of AI

What has changed and what should we do about it?

IN JANUARY 2020, privacy journalist Kashmir Hill published an article in *The New York Times* describing Clearview AI—a company that purports to help U.S. law enforcement match photos of unknown people to their online presence through a facial recognition model trained by scraping millions of publicly available face images online.[a] In 2021, police departments in many different U.S. cities were reported to have used Clearview AI to, for example, identify Black Lives Matter protestors.[b] In 2022, a California-based artist found that photos she thought to be in her private medical record were included, without her knowledge or consent, in the LAION training dataset that has been used to train Stable Diffusion and Google Imagen.[c] The artist has a rare medical condition she prefers to keep private and expressed concern about the abuse potential of generative AI technologies having access to her photos. In January 2023, Twitch streamer QTCinderella made an emphatic plea to her followers on Twitter to stop spreading links to an illicit website hosting AI-generated "deep fake" pornography of her and other women influencers. "Being seen 'naked' against your will should not be part of this job."[d]

IMAGE BY THERIGHTFRAMEMEDIA

The promise of AI is that it democratizes access to rare skills, insights, and knowledge that can aid in disease diagnosis, improve accessibility of products and services, and speed the pace of work. The peril is that it facilitates and fuels a mass, unchecked surveillance infrastructure that can further empower the powerful at the expense of everyone else. In short, AI changes privacy—it creates new types of digital privacy harms (for example, deep fake pornography), and exacerbates the ones we know all to well (for example, surveillance in the name of national security).[1]

And yet, the genie is out of the bottle: There is no short-term future where AI does not feature heavily in products and services. In response to these potential detrimental effects, human-centered AI (HAI)—an approach to AI research and practice that centers around human needs, societal good, and safety—has emerged as a rebuttal to traditional approaches to AI research and development. Privacy is one of the five most frequently cited principles for HAI,[4] yet there remains a significant gap between principles and practice for nearly all HAI principles.[7]

How does AI change privacy? Are the designers, engineers, and technologists who create AI technologies equipped to recognize and mitigate the unique privacy risks entailed by the AI products and services they create? We need answers to these questions if we are to steer the development of AI products and services toward their promise and away from their peril.

How Does AI Change Privacy?

Before we can answer how "AI" changes "privacy," it is worth clarifying what we mean when we say those words. AI is an umbrella term that encompasses many technologies.[4] Machine learning, a subdiscipline of AI that centers on building systems that auto-

[a] See https://bit.ly/3LwIVeC
[b] See https://bit.ly/3rwjfrH
[c] See https://bit.ly/46nEjPR
[d] See https://bit.ly/46o9tXF

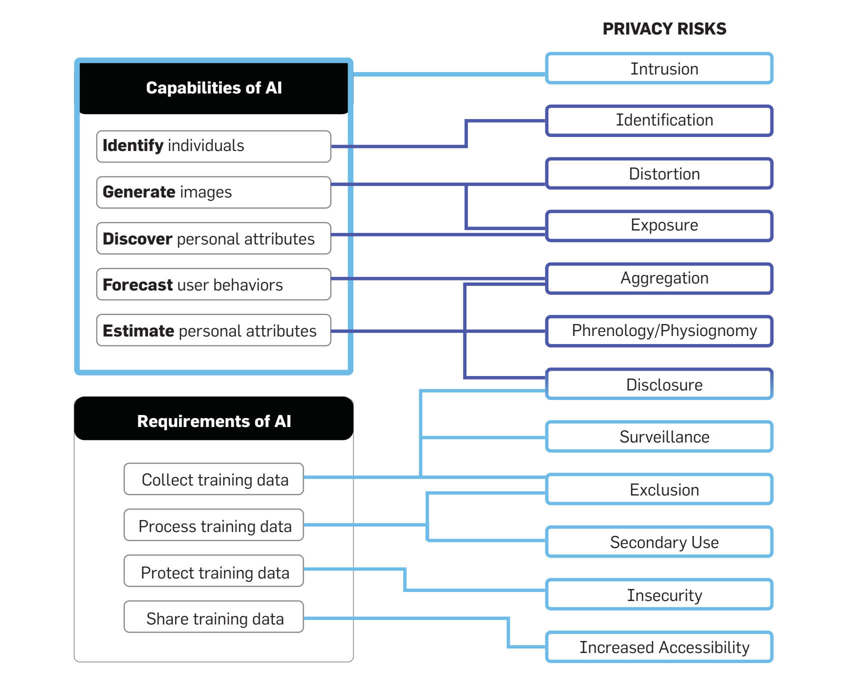
matically improve through data,[2] has been billed as "unreasonably effective"[3] at emulating human- or super-human-level performance at a number of tasks previously thought to be strictly in the domain of human "intelligence"—for example, identifying objects in images, generating natural language text, and making reasoned predictions about future trends.[10] This unreasonable effectiveness has fueled an unceasing appetite for ever more personal data and hardware to capture and process that data.

Privacy does not have a pithy, universally agreed-upon definition.

In contrast, privacy does not have a pithy, universally agreed-upon definition. Judith Jarvis Thomson once said: "Perhaps the most striking thing about the right to privacy is that nobody seems to have any very clear idea what it is."[9] Robert Post famously stated: "Privacy is a value so complex, so entangled in competing and contradictory dimensions, so engorged with various and distinct meanings, that I sometimes despair whether it can be usefully addressed at all."[6] So what do we mean when we say privacy? Recognizing that "privacy is too complicated a concept to be boiled down to a single essence," Solove proposed instead a taxonomy of privacy to articulate a range of threats and concerns that can be considered within the purview of "privacy."[8] These threats range from concerns over how data is collected, processed, disseminated, and used to invade people's personal affairs.

But how do the capabilities and requirements of AI change privacy? Today, there is a great deal of hype in conversations about what AI can and cannot do—many concerns about AI are speculative and far fetched, but we would like to keep the conversation grounded in reality. We drew on the AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC)[e] to analyze 321 documented AI privacy incidents. Rooting our analysis on Solove's taxonomy of privacy, we found that the unique capabilities and data requirements of AI can create new and exacerbate known privacy intrusions across 12 high-level categories: intrusion, identification, distortion, exposure, aggregation, phrenology/physiognomy, disclosure, surveillance, exclusion, secondary use, insecurity, and increased accessibility (see the figure here). In short, we found the unique capabilities of AI create new types of privacy intrusions, while the massive data requirements of AI can exacerbate intrusions that are already well known.

The unique capabilities of AI can entail new privacy risks, for example, those of identification, exposure, distortion, aggregation, and unwanted disclosure. The mass data requirements of AI can exacerbate known risks, for example, surveillance, exclusion, secondary use, and data breaches owing to insecurity. (Isadora Krsek helped with the creation of this image.)



Consider, for example, how one of the unique capabilities of AI is the ability to generate human-like media. This capability creates new types of exposure intrusions that can surface and reveal private information that we would like to conceal. Deep fakes and generative adversarial networks have been used, for example, to "undress people" without consent. Consider also how the requirement for vast troves of personal data that fuel AI models incentivizes the uncritical collection of personal data streams that can lead to secondary use intrusions. Foundation models—large models, such as GPT-4, that are trained for general purpose task performance and that are often fine-tuned with much smaller datasets to optimize performance for specific tasks—are often trained on large datasets of personal data scraped from the Web. But when did you consent to having your reddit posts used to train GPT-4? And what about all of the downstream models that have been trained on top of GPT-4 for more context-specific use cases, such as providing online mental health care?

Are Practitioners Equipped to Mitigate AI-Exacerbated Privacy Threats?

Since AI technologies have the potential to pose unique privacy harms and because the design pipeline for AI differs significantly from traditional software engineering, we also explored how well-equipped AI practitioners are in identifying and mitigating AI-exacerbated privacy threats. We interviewed 35 industry AI practitioners to understand how aware they were of the unique privacy

[e] See https://bit.ly/45hRXDz

threats entailed by AI, what motivated and inhibited their privacy work for AI products and services, and what affected their ability to do this work.

We found that AI practitioners defined privacy work as protecting users from intrusive or non-consensual uses of personal data (for example, surveillance, secondary use). They exhibited relatively low awareness of the ways in which AI products and services create and/or exacerbate unique privacy threats. We also observed that practitioners faced many more inhibitors than motivators for privacy work. While compliance with regulatory and policy requirements was a key motivator for their privacy work, and allowed practitioners to prioritize privacy even when it conflicted with other product goals (such as model performance), it also prevented practitioners from conceiving of privacy beyond meeting the minimum standards for compliance requirements that were not AI-specific. Accordingly, the work practitioners were motivated to engage in did not directly address the unique privacy risks entailed by AI. Finally, we also found that practitioners relied on design references and automated audits to help minimize privacy risks, but observed the tools and artifacts they used were not specific to their product or to the harms introduced or exacerbated by AI. As a result, practitioners felt ill-equipped to handle their privacy work and discussed the need for more product- and AI-specific guidance in the tools they employed in designing for privacy.

Recommendations

AI is a transformative force, reshaping privacy in ways we are only just beginning to understand. As it stands, practitioners are standing in the path of this onslaught with little more than a tin shield and blindfold. They are operating complex systems that harbor profound privacy implications without the requisite tools or knowledge to mitigate them. So, how can we better equip practitioners?

First, we need a comprehensive mapping of AI capabilities to potential privacy threats. The ill-defined, nebulous terms of "AI" and "privacy" make it easy for practitioners to overlook the specific privacy risks associated with the use of AI in their products and services. This lack of clarity fuels an unexamined approach to AI privacy work, enabling the creation of potentially privacy-intrusive AI systems. We need clear, practical guidance for practitioners to help them distill the capabilities and requirements of the AI technologies they leverage, mapping them to the new privacy threats they carry.

There is a great deal of hype in conversations about what AI can and cannot do.

We should also tap into the power of community and shared knowledge. Prior work uncovers how the driving force behind developers' privacy practices is inherently social—people learn best from their peers. Our interviewees yearned for a comprehensive, indexed, live repository of AI privacy best practices. Showcasing concrete before-and-after scenarios and clear evidence that the community values privacy, this shared knowledge base can provide both direction and motivation.

Finally, Privacy by Design is an aspirational ideal that has been a useful rallying call for regulators, but is not a practical blueprint for daily work. Instead, we need a turnkey design methodology: Privacy *through* Design. My colleagues and I are developing Privacy through Design, which encompasses a suite of worksheets, guides, and tools that can help practitioners understand and balance the utility and intrusiveness of consumer-facing AI products. One example comes from prior work exploring the use of NLP-powered audience controls for Facebook (for example, restricting posts to "friends who like pizza" instead of "friends").[2] Participants were initially all for the enhanced utility of these controls. However, when exposed to scenarios that highlighted the intrusive data processing practices they would be enabling, most found the privacy costs too high for the usability benefit.

Imagine a design process where use case-specific tensions between privacy and other design goals can be compared across use cases and stakeholders. Stakeholders' preferences for privacy or model utility in diverse scenarios would be evaluated. Yes, law enforcement would benefit from real-time facial recognition on smart glasses—but perhaps people might think that the privacy costs of these technologies is too high. The reputational risk of creating such a technology, in turn, may nudge product teams toward alternative design ideas. The key is to involve all stakeholders and respect diverse perspectives.

This approach will help practitioners strip away their blindfolds to raise awareness of the privacy risks their AI usage may entail, stir up the motivation to define what privacy means for their products within the framework of regulations and policy, and empower them to gauge their efficacy in mitigating the privacy harms uniquely intensified by their AI usage. The challenge is immense, but the potential for progress is great. C

References

1. Das, S. Subversive AI: Resisting automated algorithmic surveillance with human-centered adversarial machine learning. In *Resistance AI Workshop at NeurIPS* (2020), 4.
2. Ernala, S.K. et al. Exploring the utility versus intrusiveness of dynamic audience selection on Facebook. In *Proceedings of the ACM on Human-Computer Interaction 5* (CSCW2), (2021), 1–30.
3. Halevy, A. et al. The unreasonable effectiveness of data. *IEEE Intelligent Systems 24*, 2 (2009), 8–12.
4. Jobin, A. et al. The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*, 9 (2019), 389–399.
5. Jordan, M.I. and Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science 349*, 6245 (2015), 255–260.
6. Post, R.C. Three concepts of privacy. *Geo. LJ 89* (2000).
7. Shneiderman, B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS) 10*, 4 (2020), 1–31.
8. Solove, D.J. A taxonomy of privacy. *University of Pennsylvania Law Review* (2006), 477–564.
9. Thomson, J.J. The right to privacy. *Philosophy and Public Affairs* (1975), 295–314.
10. Yildirim, N. et al. Creating design resources to scaffold the ideation of AI concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (July 2023), 2326–2346.

Sauvik Das (sauvik@cmu.edu) is an assistant professor at the Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

Hao-Ping (Hank) Lee (haopingl@cs.cmu.edu) is a Ph.D. student at the Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

Jodi Forlizzi (forlizzi@cs.cmu.edu) is Herbert A. Simon Professor and Associate Dean for Diversity, Equity, and Inclusion at the Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.