Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks

Hao-Ping (Hank) Lee
haopingl@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, United States

Yu-Ju Yang
yujuy@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, United States

Thomas Serban von Davier
thomas.von.davier@cs.ox.ac.uk
University of Oxford
Oxford, United Kingdom

Jodi Forlizzi
forlizzi@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, United States

Sauvik Das
sauvik@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, United States

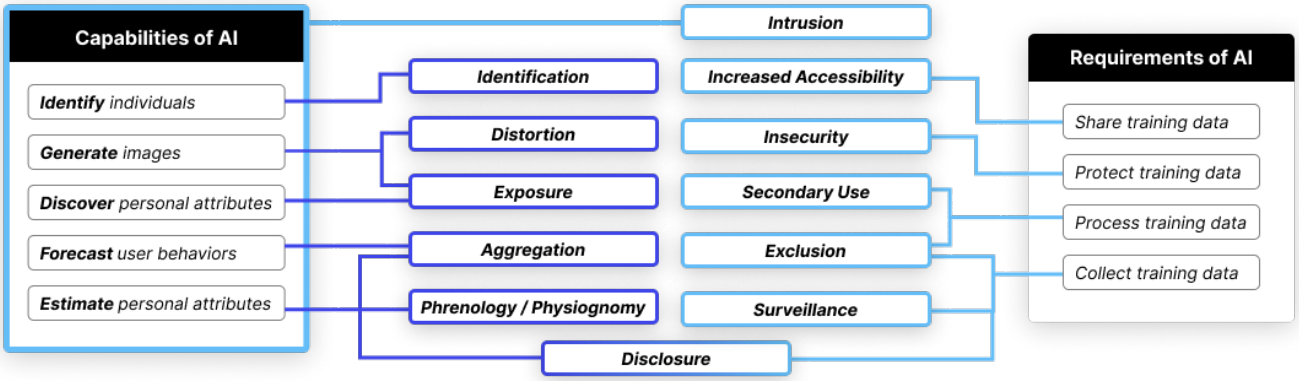

Figure 1: We identify 12 privacy risks that the unique capabilities and/or requirements of AI can entail. For example, the capabilities of AI create new risks (purple) of identification, distortion, physiognomy, and unwanted disclosure; the data requirements of AI can exacerbate risks (light blue) of surveillance, exclusion, secondary use, and data breaches owing to insecurity.

ABSTRACT

Privacy is a key principle for developing ethical AI technologies, but how does including AI technologies in products and services change privacy risks? We constructed a taxonomy of AI privacy risks by analyzing 321 documented AI privacy incidents. We codified how the unique capabilities and requirements of AI technologies described in those incidents generated new privacy risks, exacerbated known ones, or otherwise did not meaningfully alter the risk. We present 12 high-level privacy risks that AI technologies either newly created (e.g., exposure risks from deepfake pornography) or exacerbated (e.g., surveillance risks from collecting training data). One upshot of our work is that incorporating AI technologies into a product can alter the privacy risks it entails. Yet, current approaches to privacy-preserving AI/ML (e.g., federated learning, differential privacy, checklists) only address a subset of the privacy risks arising from the capabilities and data requirements of AI.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; • **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Privacy, Human-centered AI, Privacy taxonomy, Privacy risks, AI incidents

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642116

ACM Reference Format:
Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642116

1 INTRODUCTION

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 [57]. In 2021, police departments in many different U.S. cities were reported to have used Clearview.AI to identify individuals, including Black Lives Matter protesters [116]. 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 [39]. The artist has a rare medical condition that she preferred to keep private, and expressed concern about the abusive 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 deepfake pornography of her and other women influencers. "Being seen 'naked' against your will should NOT BE A PART OF THIS JOB" [110].

These examples illuminate the unique privacy risks posed by AI technologies, prompting the foundational research question we ask in this work: **How do modern advances in AI and ML change the privacy risks of a product or service?** To answer this question, we introduce a taxonomy of AI privacy risks, grounded in an analysis of 321 privacy-relevant incidents that resulted from AI products and services, sourced from an AI incidents database [108], much like the ones described above. This work is important for at least two reasons. First, people are concerned about how AI can affect their privacy: a 2021 survey with around 10,000 participants from ten countries found that roughly half of the respondents believed that AI would result in "less privacy" in the future, citing concerns around large-scale collection of personal data, consent, and surveillance [71]. Second, while privacy is one of the five most commonly cited principles for the development of ethical AI technologies [66], we do not yet have a systematic understanding of if and how modern advances in AI change the privacy risks entailed by products and services.

While AI and ML technologies have vastly expanded in capability [159], there is simultaneously a great deal of hype about what these technologies can and cannot do, making it difficult to separate real risks from speculative ones [68]. Thus, it can be difficult for today's practitioners who develop AI-inclusive products and services to understand how their use of AI technologies might entail or exacerbate practical privacy risks [161]. Prior work shows this difficulty to be true: in an interview with 35 AI practitioners, Lee et al. found that participants had relatively low awareness of privacy risks unique to or exacerbated by AI, and had little incentive to and support in addressing these risks [76].

AI and privacy both existed long before modern dialogues around the role of privacy in ethical AI development. To understand what modern advances in AI *change* about privacy, we needed a suitable baseline for privacy risk as it was understood before these advances. To that end, we used Solove's highly-cited and well-known taxonomy of privacy from 2006 as a baseline [126]. Solove's taxonomy was proposed well before modern advances in AI became mainstream in product design, and remains relevant and influential to this day. Yet, Solove's taxonomy is intentionally broad and technology-agnostic — a useful attribute in the legal and regulatory contexts for which it was developed, but less helpful in prescribing specific mitigations for product designers and developers.

To ground our analysis on real and practical risks, we sourced case studies from a database indexing real AI incidents documented by journalists — the AI, Algorithmic, and Automation Incident and Controversy (AIAAIC) repository [108]. We sourced 321 case studies from the AIAAIC repository in which real AI products resulted in lived privacy risks. We next systematically analyzed whether and how the capabilities and/or requirements of the AI technology described in the incident either (i) *created* a new instantiation of a privacy risk described in Solove's original taxonomy or an entirely new category of risk, (ii) *exacerbated* a privacy risk that was already captured by Solove's taxonomy, or (iii) *did not change* the privacy risk described in the incident relative to at least one of the risks described in Solove's taxonomy.

The result is our taxonomy of AI privacy risks (see Figure 1). Our taxonomy illustrates how the unique capabilities of AI — e.g., the ability to *recommend* courses of action, *infer* users' interests and attributes, and *detect* rare or anomalous events [103] — resulted in both new instantiations of existing categories of risk in Solove's taxonomy as well as one entirely new category of privacy risk. For example, we found that the ability of AI technologies to generate human-like media resulted in new types of exposure risks (e.g., the generation of deepfake pornography [4]), while the ability for AI to learn arbitrary classification functions led to a new category of privacy risk: phrenology/physiognomy (e.g., the belief that AI can be used to automatically detect things like sexual orientation from physical attributes [78]. Our taxonomy also captures how the data and infrastructural requirements of AI exacerbated privacy risks already captured in Solove's taxonomy. For example, since facial recognition classifiers require tremendous amounts of face data, they can exacerbate surveillance risks by encouraging uncritical data collection practices such as collecting face scans in airports [42].

We discuss how existing approaches to privacy-preserving AI and machine learning, such as differential privacy and federated machine learning, only account for a subset of these risks, highlighting the need for new tools, artifacts, and resources that aid practitioners in negotiating the utility-intrusiveness trade-off of AI-powered products and services. Finally, we outline how this taxonomy can be used to create tools that help educate practitioners, and as a repository of shared knowledge regarding AI privacy risks and design processes to mitigate against those risks.

2 BACKGROUND AND RELATED WORK

2.1 Human-centered AI

AI technologies are here to stay. New AI technologies are constantly created and evolving, and so are their harms to individuals and society [108]. Advertising in which users' interests are inferred from their behaviors online to target them with relevant advertisements fuels a multi-trillion dollar industry that has been referred to as "surveillance capitalism." [165] Users find these ads both "smart" and "scary" [136]. Beyond attitudes, recent work has further shown that these advertisements result in many real, lived harms — ranging from psychological distress to traumatization [155]. As AI technologies improve, we see new uses of these technologies to make spurious predictions about individuals and their behavior, portending a new age of AI-facilitated phrenology and physiognomy: e.g., through the use of profile images to predict things like sexual orientation [145] and “criminality” [154].

In response to the potentially detrimental effects of unchecked AI on society, there is growing discussion on how AI technologies’ benefits can be ensured and their potential harms mitigated [37]. Human-centered AI (HAI) is a term commonly used to center human needs and to describe the ethical decision-making that informs AI design [19, 118, 132]. In recent years, Human-computer Interaction (HCI) researchers and AI practitioners have created a body of work to provide guidelines for HAI (see Hagendorff [52] for an extensive evaluation).

Case studies on implementing HAI guidelines reveal stakeholders’ struggle with concepts of privacy and fairness [44, 66, 140]. This paper focuses on privacy, as significant research has previously attempted to define and measure fairness in AI [21, 28]. To understand the potential privacy risks of AI technologies, the negative impacts of past implementations must be considered [89]. This method of looking at the “dark side” of technologies can reveal the potential risks of future technology concepts by reflecting on past harms and has been successfully used to analyze dark patterns in GUIs and consider how software agents may impact user autonomy [46, 84].

We present a novel taxonomy of AI privacy risks to further develop what it means to design for privacy in human-centered AI. This taxonomy aims to provide AI practitioners with tools and a shared language to foreground end-user privacy discussions in the design and development process.

2.2 Prior privacy taxonomies and concepts

AI is unique in its capacity for high-powered decision-making. Unlike traditional tools, AI systems demand copious amounts of data to refine and enhance outputs [157]. However, the data often originates from individuals, giving rise to pressing concerns about privacy and safety [129]. Therefore, input from various sectors is needed, along with comprehensive strategies for responsible development and deployment.

Ensuring privacy and AI safety has been addressed from various angles. Many approaches build upon the seminal work in privacy preservation pioneered by Shokri and Shmatikov [123]. Other research documents challenges [81], especially within the realm of deep learning techniques [24, 82]. Furthermore, a spectrum of cybersecurity threats looms over any AI system striving to safeguard the privacy of its users and data providers [102].

While these works address the task of documenting potential privacy challenges and vulnerabilities within AI systems, they often focus on specific aspects rather than taking a holistic view [121]. When researchers examine the full AI “life-cycle,” it is typically aimed at promoting and ensuring trust and assurance within AI, rather than concentrating on the initial privacy concerns that precipitated distrust [15, 149].

Shahriar et al. offer four categorizations of privacy risks along with a relevant list of strategies applicable throughout the design, development, and deployment phases of an AI system [121]: (1) the risk of identification, (2) the risk of inaccurate decisions, (3) the risk of non-transparent AI, and (4) the risk of non-compliance with regulations. Their categorizations provide effective catch-alls for various potential risks. However, like other recent frameworks (see [146]), the approach of categorizing strategies and techniques by privacy risks involves a degree of theoretical dangers outlined in research case studies and previous surveys rather than proven, reported, and documented privacy risks. Moreover, these taxonomies do not consider what AI technologies *change* about privacy risks relative to notions of privacy prior to modern advances (e.g., the creation and use of deepfake techniques).

It is crucial to turn to the literature on privacy law to address this gap and provide a more holistic and inclusive understanding of AI privacy risks as they manifest worldwide. Solove’s work represents the progress within legal discussions and the judicial system to address taxonomies of different types of privacy risks [126]. Solove’s taxonomy offered a comprehensive classification of different types of privacy intrusions (i.e., intrusions associated with information collection, information processing, information dissemination, and invasion) as seen in the legal field. It was previously used in security research to explore users’ personal attitudes and behaviors regarding privacy issues [8, 73]. However, unlike this paper, the previous research did not look to apply or change the taxonomy to the new and emerging challenges of realized privacy intrusions.

Solove and colleagues built on their work by defining what qualifies as a “harm” and how modern technology challenges these traditional distinctions [32]. Nevertheless, this taxonomy does not possess the AI-specific focus of Shahriar et al.’s research. Our paper intends to bridge the divide between these two bodies of work. By adopting an AI-centric approach to codify realized privacy risks, our paper introduces a distinctive taxonomy for evaluating and ultimately addressing privacy risks specific to the “life-cycle” of AI systems.

2.3 Creating a privacy taxonomy

A robust taxonomy can provide AI practitioners with guidance and structure during the design and development process. Taxonomies provide an organizational hierarchy of information, classifying information into distinct categories [29]. However, developing an effective privacy taxonomy is a challenge many researchers have undertaken with limited success [137]. Two characteristics make the development of a privacy taxonomy challenging.

The first is the inability to agree on any one definition of privacy. Early interpretations consider privacy “the right to be let alone” [25]. Later, the foundation of modern privacy law was built on an argument calling for individual or group autonomy over the sharing and disseminating of personal information [148]. Privacy theory also began to change to consider the dangers of inflexible regulations and the importance of treating privacy as a process rather than a label [11]. HCI researchers built on this theory to consider applications in practice [104]. More recent work considers privacy in the light of contextual integrity [97]. Other researchers embrace the difficulty of defining privacy as the reason for developing adaptive solutions and classification systems [93]. In some cases, it has been easier to define privacy within the constraints of a specific field of operation, such as databases [18].

The second characteristic that makes creating a universally accepted taxonomy difficult is the need to operationalize the taxonomy in a single domain. For example, there was a push for the taxonomy and approach of privacy by design for ubiquitous computing [75]. Similarly, in the field of Robotics, a specific taxonomy was developed to deal with implementing sensor technology [40]. These taxonomies focus on the potential privacy risks the system or technology poses [99]. Even taxonomies built on user or societal input rely on perceived risks instead of reported harms resulting from past usage of similar technology [62]. We identified one taxonomy that accounts for previously recorded privacy risks [126]. However, this work is geared towards lawmakers and legal professionals rather than AI practitioners. It is not built to address specific privacy risks associated with the functionality and design of AI systems.

Similarly, previous HCI research has attempted to provide practitioners with a privacy taxonomy based on end users' experience, raising awareness for physical privacy intrusions [151]. This previous research shows the ability to apply such taxonomies in practical settings. Yet, it does not attempt to handle the more conceptual instances of AI privacy intrusions that may be invisible to end users but are no less impactful.

Therefore, this paper takes the first step to codify patterns of documented privacy risks resulting from AI's capabilities and data requirements.

3 METHOD

3.1 Constructing the Taxonomy of AI Privacy Risks Based on AIAAIC

We developed a taxonomy of privacy risks exhibited in documented AI privacy incidents by performing a systematic review of case studies. Creating typology and taxonomy by synthesizing real-world incidents has been used more broadly in privacy and security [36] and in AI ethics [113]. In this paper, we define AI broadly to accommodate the wide range of its capabilities to *"perform tasks or behaviors that a person could reasonably deem to require intelligence if a human were to do it"* [115]. AI is an umbrella term that encompasses many technologies, and our analysis does as well — we cover approaches ranging from Machine Learning (e.g., prediction and recommendation algorithms), Natural Language Processing (e.g., large language models), Computer Vision (e.g., facial recognition), and Robotics (e.g., home robots, drones). Note that we focus on documented end-user privacy risks of actual AI/ML products rather than speculative risks of general AI/ML concepts. We partly relied on the AI, Algorithmic, and Automation Incident and Controversy Repository (AIAAIC), the largest, up-to-date crowdsourced AI incident database curated by journalism professionals [108]. We also surveyed the AI Incident Database (AIID)[1], another public AI incident database, but decided not to use it because the AIAAIC provided good coverage of most of the privacy-related incidents in the AIID[2].

Out of a database of 1,049 cases[3], 364 of them were labeled to involve "privacy issues"[4] occurring between 2012 to 2023. To ensure that the incidents we analyzed indeed involve AI privacy risks, two coders reviewed the linked resources for *all* 364 cases tagged in the AIAAIC as being privacy-pertinent. Then, the two coders went through an incident-by-incident discussion on whether the reported technology (i) claimed to be inclusive of AI, ML, or otherwise "algorithmic" approaches, (ii) was actually deployed to real end-users, and (iii) involved some form of end-user privacy risks and/or compromise, and further filtered down to 310 cases. We filtered out incidents that did not involve AI technologies (N=21, e.g., virtual-reality applications, data leakage unrelated to the use and development of AI), and incidents that were not associated with end-user privacy risks (N=33, e.g., bias [6], inaccuracy [45], copyright [48]).

To ensure an adequate sampling strategy, we randomly picked 10% of the cases *without the privacy label* in the AIAAIC database (69 out of 685). We identified privacy risk(s) in 11 of these cases (15.94%), and all of the identified risks were found in other cases tagged with the privacy label. Thus, we deemed our analysis had reached saturation. In sum, we analyzed a total of 321 distinct cases in developing our taxonomy of AI privacy risks (see Figure 2).

As our objective was to understand how AI *changes* privacy, and not to re-define what *is* privacy, we rooted our analysis on Solove's taxonomy of privacy from 2006 as a baseline [126] — a popular conceptualization of privacy risks proposed prior to modern advances in AI/ML. Our primary analytic goal was to identify if and how AI exacerbates and/or creates privacy risks relative to this taxonomy, because doing so will highlight how modern advances in AI do and do not change notions of privacy risk. We say that AI *exacerbates* privacy risks when the capabilities and/or requirements of the AI technologies are not the root cause of the privacy risk, but increased its scale, scope, frequency, and/or intensity — e.g., robust identification even with low-quality images. We say that AI *creates* new privacy risks when the capabilities and/or requirements of the AI technology are fundamental enablers of the privacy risk — e.g., deepfake pornography. Otherwise, we say that the AI has *not meaningfully changed* the privacy risk described in the incident.

For each incident, we assessed if and how the privacy violations described in the incident related to the unique context, capabilities of, and requirements entailed by the AI technologies described in the incident. We used an iterative coding process to categorize the privacy risk described in the incident. First, we created our codebook of different types of privacy risks adapted from the taxonomy proposed by Solove [126]. Next, we iteratively updated the definition and scope of Solove's initial set of privacy risks to be more

[1] https://incidentdatabase.ai/

[2] We randomly selected 33% (N=50) of the AIID total incidents that contain the keyword "privacy" (N=151 as of August 16th, 2023). We manually went through the 50 incidents: 20 were either not AI products (e.g., augmented reality applications, executive orders from the government, policies) or not directly related to privacy (e.g., bias, accuracy), and 17 were already included in our AIAAIC database. The remaining 13 (26%) were not, but we found similar incidents in the AIAAIC database that were already captured by our taxonomy, e.g., incidents related to surveillance, data breaching, distortion made by deepfake AI, and physical invasion of AI technologies.

[3] We took a snapshot of the database on August 16th, 2023

[4] The AIAAIC database tags each case with two attributes, *Issue(s)* and *Transparency*, to reflect if a given case raises any privacy concerns from the stakeholders and media.

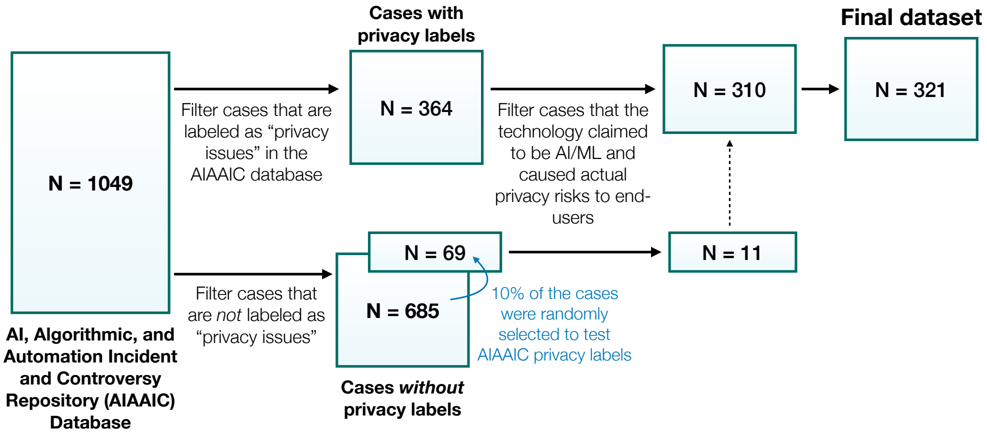
Cases with privacy labels
Final dataset
N = 1049
Filter cases that are labeled as "privacy issues" in the AIAAIC database
N = 364
Filter cases that the technology claimed to be AI/ML and caused actual privacy risks to end-users
N = 310
N = 321
N = 69
N = 685
N = 11
Filter cases that are *not* labeled as "privacy issues"
10% of the cases were randomly selected to test AIAAIC privacy labels
AI, Algorithmic, and Automation Incident and Controversy Repository (AIAAIC) Database
Cases *without* privacy labels

Figure 2: We filtered from 1,049 cases from the AIAAIC database and selected cases labeled as "privacy issues." We filtered them down to cases with the technology claimed to be AI/ML that caused actual privacy risks to end-users. We also picked 10% of the cases without the privacy label from the database and went through the same analysis process. The final dataset comprised a total of 321 cases.

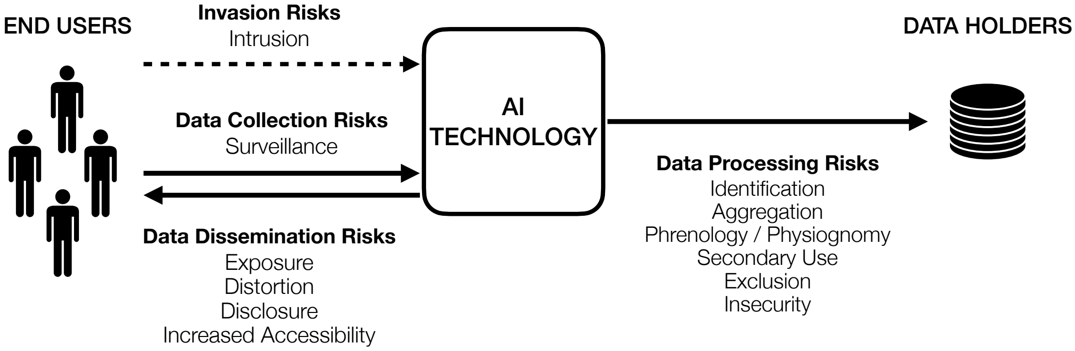

Figure 3: 12 types of privacy risks that AI technologies create and/or exacerbate relate to data collection, data processing, data dissemination, and invasion. The arrows indicate data flow (invasion risks need not involve data, but often do).

specific to the AI privacy incidents in our dataset. For example, in our incident database, we observed that *Increased Accessibility* typically manifested as increasing public access to otherwise private or access-controlled data for building AI/ML models (e.g., through the release of public datasets). We also merged risks when they exclusively co-existed in our analysis. For example, we found that the Appropriation risk, the use of one's identity to serve the aims and interests of another, always manifested with the Distortion risk, disseminating realistic AI-generated false information about individuals. While these two categories can theoretically be separable (i.e., one can imagine Distortion without Appropriation or Appropriation without Distortion), to keep our taxonomy grounded on real incidents and not theoretical harms, we merged the two categories into a single *Distortion* category. Finally, we found an entirely new type of privacy risk, *Phrenology / Physiognomy*, which is not captured in Solove's initial set of privacy risks. This privacy risk is unique to AI due to its capability to estimate sensitive personal attributes (e.g., sexual orientation, ethnicity) of individuals from their physical attributes (e.g., appearance, voice).

In total, we created a final codebook of 12 operationalizable privacy risk labels for AI technologies, including *Surveillance, Identification, Aggregation, Phrenology / Physiognomy, Secondary Use, Exclusion, Insecurity, Exposure, Distortion, Disclosure, Increased Accessibility*, and *Intrusion* (see Table 1 and Figure 3).

3.2 Qualitative Analysis Procedure

To summarize our qualitative analysis procedure, the first author iteratively applied the codebook to 132 cases to update and better scope the definition of each privacy risk in active discussion with four other authors and constructed the initial codebook. Another author joined the coding process when the initial codebook was constructed. This author was trained with the codebook and independently coded the same set of 132 cases. The codes were then iteratively refined and discussed when disagreements occurred until both authors agreed on all codes in the codebook. To validate the inter-rater reliability, the two coders then independently coded another 65 cases (20% of our overall analysis pool; N=321) and reached a high agreement, with Cohen's Kappa larger than 0.8

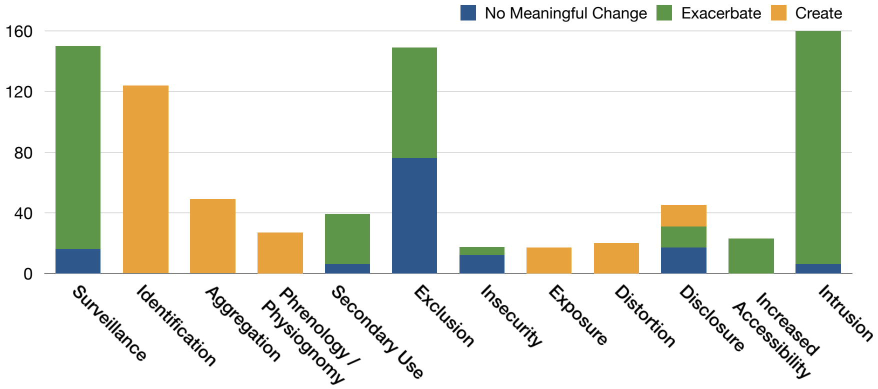

Figure 4: The distribution of each privacy risk we identified as not meaningfully changed, exacerbated, or created by AI. Note that one AI incident can involve multiple types of privacy risks.

on every type of risk and averaging 0.94 on all types of risks (see Appendix Table 2). One coder then coded the rest of the 124 cases. The final codebook comprises 12 types of privacy risks that we identified across the corpus of 321 cases. In determining whether AI newly created, exacerbated, or not meaningfully changed the privacy risks identified in each incident, the two coders engaged in an incident-by-incident discussion for all 321 incidents concerning the root cause of the privacy intrusions, as well as the role AI played in that root cause. The three themes — i.e., create, exacerbate, and no meaningful change — naturally emerged during this process.

4 TAXONOMY OF AI PRIVACY RISKS

We introduce a taxonomy of AI privacy risks: i.e., privacy risks that are created and/or exacerbated by the incorporation of AI technologies into products and services. In short, we found that AI technologies create new instantiations of the privacy risks in Solove's taxonomy [126] (e.g., generative AI can create new types of distortion intrusions), create a new category of risk not captured by Solove's taxonomy (e.g., resurging phrenology/physiognomy), and exacerbate many of the risks highlighted by Solove's taxonomy (e.g., AI technologies can more robustly identify individuals from low fidelity data sources) (see Figure 4).

We discuss these AI-created and exacerbated risks below as they relate to data collection, processing, dissemination, and invasion (see Figure 3). Overall, we found that of the 321 incidents from the AIAAIC database that involve privacy risks, the AI technology implicated in the incident either created or exacerbated the described privacy risks 298 times (92.8%), suggesting that the unique capabilities and/or requirements of AI do appear to meaningfully change privacy risks and that AI-specific privacy guidance may be necessary for practitioners.

4.1 Data collection risks

Data collection risks "create disruption based on the process of data gathering" [126]. Recent advances in AI/ML have been fueled by the collection of vast amounts of personal data. Solove further identifies surveillance as a risk that pertains to AI technology. AI technologies might *create* data collection risks if the AI technology provides functionality that enables the collection of previously inaccessible data; they *exacerbate* data collection risks when data is collected specifically for the development of an AI/ML system, or if AI technologies facilitate the data collection process in a manner that increases the scope of the risk. In our analysis, we found incidents of AI exacerbating **surveillance** risks, but not of creating new such risks.

4.1.1 Surveillance (150/321).

Surveillance refers to watching, listening to, or recording an individual's activities [126]. Surveillance risks long pre-date modern advances in AI. AI technologies do not always meaningfully change surveillance (16/150), i.e., when end-users feed their own personal data to access the utility offered by AI, such as by uploading videos to capture body movement or estimate car speed. Nevertheless, owing to the never-ending need for personal data to train and deploy effective machine learning models, we identified two ways AI technologies can exacerbate surveillance risks: i.e., by increasing the scale and ubiquity of personal data collected.

AI enhances the scale of surveillance (32/150) by enabling linking across a diversity of sources, and increasing the quantity of collected personal data.

Where applicable, real-world models collect data from different sources to enrich datasets. We found that multi-faceted, high-fidelity data can exacerbate risks involving surveillance in the physical world. One example comes from a predictive policing platform deployed in Xinjiang, China. The system *"collects [individual's] information from a variety of sources including CCTV cameras and Wi-Fi sniffers, as well as existing databases of health information, banking records, and family planning history"* [112]. This information was then used to identify persons and assess their activities in the real world. We also found incidents describing AI systems that collected an array of end-user behavioral data in the cyber world. For example, Gaggle, a student safety management tool, monitors students' digital footprints such as email accounts, online documents, internet usage, and social media accounts to assess and prevent violence and suicides [20].

Table 1: Taxonomy of AI Privacy Risks. We found incidents matching 12 distinct, but not mutually exclusive, categories of privacy risk.

Privacy risk [126]	***How does AI influence the risk?***	***Examples***
Data Collection Risks		
Surveillance: *watching, listening to, or recording of an individual’s activities*	AI ***exacerbates*** surveillance risks by increasing the scale and ubiquity of personal data collected.	A predictive policing platform deployed in Xinjiang, China,“*collects [individual’s] information from a variety of sources including CCTV cameras and Wi-Fi sniffers, as well as existing databases of health information, banking records, and family planning history*” [112].
Data Processing Risks		
Identification: *linking specific data points to an individual’s identity*	AI ***creates*** create new types of identification risks with respect to scale, latency, robustness, and ubiquity.	Models trained on Simulated Masked Face Recognition Dataset (SMFRD) are capable of identifying persons with a mask on, “*violating the privacy of those who wish to conceal their face*” [150].
Aggregation: *combining various pieces of data about a person to make inferences beyond what is explicitly captured in those data*	AI ***creates*** new types of aggregation risks owing to their scale, latency, ubiquity, and their ability to forecast end-user behavior and infer end-user attributes.	“*The system, called the National Data Analytics Solution (NDAS), uses a combination of AI and statistics to try to assess the risk of someone committing or becoming a victim of gun or knife crime*” [17].
Phrenology / Physiognomy: *inferring personality, social, and emotional attributes about an individual from their physical attributes*	AI ***creates*** phrenology/physiognomy risks through learning correlations between arbitrary inputs (e.g., images) and outputs (e.g., sexual orientation).	‘Gaydar’, an AI sexual orientation prediction model, was found to “distinguish between gay or straight people” based on their photos [79].
Secondary use: *the use of personal data collected for one purpose for a different purpose without end-user consent*	AI ***exacerbates*** secondary use risks by creating new AI capabilities with collected personal data, and (re)creating models from a public dataset.	The Diversity in Faces (DiF) dataset was created to improve the research on fairness and accuracy of artificial intelligence face recognition systems across genders and skin colors, and should not be used for commercial purposes. Nevertheless, Amazon and Microsoft were accused of using the dataset to “*improve the accuracy of their facial recognition software*” [13].
Exclusion: *the failure to provide end-users with notice and control over how their data is being used*	AI ***exacerbates*** exclusion risks by training on rich personal data without consent.	LAION-5B is a large, openly accessible image-text dataset for training ML models. However, a person found that her private medical photographs were referenced in the public dataset, and suspected that “*someone stole the image from my deceased doctor’s files and it ended up somewhere online, and then it was scraped into this dataset*” [39].
Insecurity: *carelessness in protecting collected personal data from leaks and improper access due to faulty data storage and data practices*	AI ***exacerbates*** insecurity risks by introducing new vulnerabilities when incorporating AI and its associated data pipeline in the products.	Lee Luda, a chatbot trained on real-world text conversations, was found to expose the names, nicknames, and home addresses of the users whose data on which it was trained [63].
Data Dissemination Risks		
Exposure: *revealing sensitive private information that people view as deeply primordial that we have been socialized into concealing*	AI ***creates*** new types of exposure risks through generative techniques that can reconstruct censored or redacted content; and through exposing inferred sensitive data, preferences, and intentions.	TecoGAN, a deep learning video clarification tool, has been used to clarify censored images of genitalia [91].

Distortion: *disseminating false or misleading information about people*	AI ***creates*** new types of distortion risks through the generation of realistic fake images and audio that humans have difficulty discerning as fake.	Prime Voice AI, a text-to-voice generator, was misused to create the voices of celebrities to *“make racist remarks about the US House representative”*, and that the AI-generated clips *“run the gamut from harmless, to violent, to transphobic, to homophobic, to racist”* [33, 53].
Disclosure: *revealing and improperly sharing data of individuals*	AI ***creates*** new types of disclosure risks by inferring additional information beyond what is explicitly captured in the raw data. AI ***exacerbates*** disclosure risks through sharing personal data to train models.	The “Safe City” initiative in Myanmar used AI-infused cameras to identify faces and vehicle license plates in public places and alert authorities to individuals with criminal histories [5]. The UK’s National Health Service partnered with Google to share mental health records and HIV diagnoses of 1.6 million patients to develop a model for detecting acute kidney injury [59].
Increased Accessibility: *making it easier for a wider audience of people to access potentially sensitive information*	AI ***exacerbates*** the scale of increased accessibility risks via publicizing large-scale datasets that contain personal information, for the use of building and improving AI/ML models.	OkCupid dataset contained personal information such as users’ location, demographics, sexual preferences, and drug use, and was uploaded to Open Science Framework to facilitate research on modeling dating behaviors [153].
Invasion Risks		
Intrusion: *actions that disturb one’s solitude in physical space*	AI ***exacerbates*** the scale and ubiquity of intrusion risks via enabling centralized and/or ubiquitous surveillance infrastructures.	Ring, a smart doorbell that enables homeowners to monitor activities and conversations near where the doorbell is installed has raised concern due to *“the devices’ excessive ability”* to capture data of an individual’s neighbors [90].

Additionally, as the amount of training data often has a direct impact on model performance, AI technologies can exacerbate surveillance risks by increasing the need for collecting large-scale personal data to train effective models. For example, the South Korean Ministry of Justice attempted to build a government system for screening and identifying travelers based on photos of over 100 million foreign nationals who entered the country through its airports [42]. Without the promise of AI technologies to automatically sift through and make sense of these data, there would be little incentive to collect data of this scale.

AI technologies exacerbate the ubiquity of surveillance risks (102/150) by using physical sensors and devices to collect information from environments. For example, geolocation data from mobile devices were used to assess employee performance, raising concerns about employee tracking outside of work [138]. CCTV cameras have been used in applications to detect and prevent suicide attempts [106] or to detect security anomalies in physical spaces [142], while also introducing bystander privacy risks and concerns [83]. Microphones enable a responsive audio interface for virtual assistants, along with concerns of extensive audio data collection and eavesdropping by the service provider [107].

4.2 Data processing risks

Data processing risks result from the use, storage, and manipulation of personal data [126]. Solove identified five types of data processing risks: identification, aggregation, secondary use, exclusion, and insecurity. In our analysis, we found incidents pertaining to each of these risks, as well as an entirely new category of data processing risk: **phrenology/physiognomy** risk, which is created by AI technologies by correlating arbitrary inputs and outputs. We also found that AI technologies create new types of **identification** and **aggregation** risks (e.g., by operating on low-quality data; and by forecasting future events), and exacerbate **secondary use**, **exclusion**, and **insecurity** risks (e.g., by re-purposing foundation models; by training models on datasets containing content obtained without consent; and by introducing new security vulnerabilities due to the use of AI).

4.2.1 Identification (124/321). Identification refers to linking specific data points to an individual’s identity [126]. These risks are commonplace even without AI; for example, users may be manually tagged in photos, or manually identified in CCTV video feeds. AI technologies, however, allow for automated identity linking across a variety of data sources, including images, audio, and biometrics. We found that AI technologies entail new types of identification risks with respect to scale, latency, robustness, and ubiquity.

AI technologies enabled automated identification at scale (20/124). One example is Facebook’s now-disabled Tag Suggestions product, through which Facebook demonstrated its ability to automatically identify individuals from uploaded photos. When this feature was in use, Facebook had 1.4 billion daily active users[5]; still, *“any time someone uploads a photo that includes what Facebook thinks is your face, you’ll be notified even if you weren’t tagged”* [124].

AI technologies allow identification risks to occur more quickly, in nigh real-time (24/124), once the models are trained. For example, in 2019, the Italian government was on the verge of implementing a real-time facial recognition system across football stadiums that

[5] https://investor.fb.com/home/default.aspx

"prevent individuals who are banned from sports competitions from entering stadiums." It also picked up audiences' "racist conversations" to alert law enforcement authorities to the presence of racist fans [127].

In addition, *AI technologies allow for robust identification even with low-quality data (7/124).* Clearview AI, a facial recognition application that aids U.S. law enforcement in identifying wanted individuals, claims to be able to identify people under a range of obfuscation conditions: *"[a] person can be wearing a hat or glasses, or it can be a profile shot or partial view of their face"* [57]. Similarly, models trained on Simulated Masked Face Recognition Dataset (SMFRD)[6] are capable of identifying persons with a mask on, *"violating the privacy of those who wish to conceal their face"* [150].

Finally, *AI technologies enable ubiquity identification risks in situated physical environments (73/124)* like public places (e.g., [92]), stores (e.g., [58]), and classrooms (e.g.,[114]). For example, XPeng Motors, a Chinese electric vehicle firm, was reported for using facial recognition-embedded cameras in their stores to collect biometric data of customers [49].

4.2.2 Aggregation (49/321).

Aggregation risks refer to combining various pieces of data about a person to make inferences beyond what is explicitly captured in those data [126]. These risks can occur without AI through manual analysis, but AI technologies greatly facilitate these inferences at scale, identified as a future trend by Solove: *"the data gathered about people is significantly more extensive, the process of combining it is much easier, and the computer technologies to analyze it are more sophisticated and powerful"* [126]. Similar to identification risks, we found that AI technologies create new types of aggregation risks owing to their scale, latency, ubiquity, and their ability to forecast end-user behavior and infer end-user attributes.

One of the unique strengths of AI systems is that they automate complex processes into simple programs that overcome human limitations. While controversial, many public sectors still utilize algorithmic tools in high-stake contexts such as social work [69] and services for the unhoused [74] to prioritize limited resources. To that end, *AI technologies create aggregation at scale (23/49)* by processing vast amounts of personal data to infer invasive things about individuals not explicit in those data. For example, an AI start-up created a service that assesses a prospective babysitter's likelihood to engage in risky behaviors such as drug abuse and bullying by *"scan[ning] ... thousands of Facebook, Twitter and Instagram posts"* [56].

AI technologies perform complicated inferencing tasks nigh instantly (11/49). Technologies have been developed to estimate employee performance in-the-moment [141], and to forecast what one might write in emails [134]. AI technologies have also been developed to predict when end-users might be ovulating [61], and their moment-to-moment risk of committing suicide [20].

AI technologies can make physical objects and environments smarter and more responsive, *enabling ubiquitous aggregation risks (5/49).* Smart home devices, for example, allow for automated control of home appliances, dynamic temperature control to strike an optimal balance between energy consumption and comfort, and voice user interfaces [9]. These features require AI technologies to continuously monitor data streamed from physical sensors, creating new aggregation risks in situated environments. For example, smart speaker microphone feeds have been used to infer who is present in a room, who is speaking, and other information that can be algorithmically inferred from voice data [2].

Finally, AI technologies enable forecasting future behaviors and states based on historical data. This forecasting can be used, for example, to help proactively identify health risks, plan optimal routes to avoid predictable traffic, and estimate retirement savings. These capabilities of AI, however, also *create a new type of predictive aggregation risk (10/49).* For example, in 2018, Argentina's government deployed an AI model that predicted teen pregnancy in low-income areas from their first name, last name, and address [65]. AI has also been used for crime prediction. For example, in 2018, law enforcement in the United Kingdom aimed to predict serious violent crime using AI based on *"records of people being stopped and searched and logs of crimes committed"* [17].

4.2.3 Phrenology / Physiognomy (27/321).

Phrenology and Physiognomy are debunked pseudosciences that postulate that it is possible to make reliable inferences about a person's personality, character, or predispositions from an analysis of their *outer appearance* and/or *physical characteristics* [1]. Beyond the baseless prediction made from historical data streams discussed in Aggregation risks (Section 4.2.2), phrenology/physiognomy risks pose unique downstream privacy harms distinct from aggregation risks: whereas aggregation risks primarily arise from the collection and combination of disparate pieces of information to make deductive inferences about individuals, phrenology/physiognomy risks introduce new and unfounded inferences about an individual's internal characteristics (e.g., their preferences and proclivities). Moreover, while aggregation risks generally come from the combination of factual and observable data streams over which users can have some awareness and control (e.g., purchasing habits), phrenology/physiognomy risks arise from making inferences over physical characteristics over which users have no control. Moreover, beyond the harm to the individual, there is also a broader societal harm: prior work has warned that irresponsible use of AI classification models could usher in a revival of these pseudosciences [14, 128] by, e.g., motivating surveillance institutions to train AI models to make spurious inferences about a person's preferences, personality, and character from inputs that capture their outer appearance. Our analysis reveals that AI technologies are indeed being used in this way, resulting in a new category of privacy risk not captured by Solove's initial taxonomy. We define phrenology/physiognomy risks as the use of AI to infer personality, social, and emotional attributes about an individual from their physical attributes. This risk stems from AI's ability to learn correlations between arbitrary inputs (e.g., images, voices) and outputs (e.g., one's demographic information).

Some models aim to infer preferences, like sexual orientation. For example, 'Gaydar' is an AI sexual orientation prediction model that "distinguishes between gay or straight people" based on their photos [79]. Researchers have also used AI to predict "criminality" — i.e., whether someone is a criminal — from facial images [154]. Outside of the problematic assumptions of these models (i.e., that sexual orientation and criminality can be inferred from photos), this

[6] https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset#download-datasets

research raises concerns about the potential for harm and misuse of AI models to infer and disseminate information about individuals without consent [79]. AI technologies have also been used to predict other personal information such as one's name [26], age [109], and ethnicity [117] based on facial characteristics.

Other models aim to predict a person's mental and emotional state based on their images. For example, teaching tools devised by Class Technologies estimate students' engagement from their facial expressions without students' consent [70]. Still other models scrutinize vocal attributes to predict an individual's trustworthiness. For instance, the AI system DeepScore captures and assesses voice data to predict deceptiveness, and has been utilized by health insurance and money lending platforms to select low-risk clients [43].

4.2.4 *Secondary Use (39/321).*

Secondary use encompasses the use of personal data collected for one purpose for a different purpose without end-user consent [126]. In AI technologies, this risk is mostly associated with data practices for training data. AI does not always change secondary use risks (6/39). For example, Luca, an app that was used for contact tracing during the COVID-19 pandemic in Germany, was found to re-purpose personal data, such as location data, to support law enforcement by *"tracking down witnesses to a potential crime"* [105]: but the risk described here would have been just as salient even without the use of AI. Nevertheless, many common practices to train AI/ML models more effectively can exacerbate secondary use. In our dataset, we identified two AI-exacerbated secondary use risks: creating new AI capabilities with collected personal data, and (re)creating models from a public dataset.

When data collectors have already built models using personal data, they may be tempted to expand the models by creating additional features and capabilities, which can be unanticipated for end-users (22/39). For example, OkCupid, a dating site that matches users using an *"one-of-a-kind algorithm"*[7], was found to contact an AI startup, Clarifai, *"about collaborating to determine if they could build unbiased A.I. and facial recognition technology,"* and that *"Clarifai used the images from OkCupid to build a service that could identify the age, sex and race of detected faces"* [86].

Secondary use risks can also be exacerbated when AI practitioners try to reuse pubic datasets to train models for purposes other than the original purpose for which those data were collected (11/39). For example, People in Photo Albums (PIPA) is a facial photograph dataset created to *"recogniz[e] peoples' identities in photo albums in an unconstrained setting"* [162]. Yet, the PIPA dataset has been used in research affiliated with military applications and companies like Facebook [54, 55]. Similarly, the Diversity in Faces (DiF) dataset is a collection of annotations of one million facial images that was released by IBM in 2019 [125]. The dataset was created to improve the research on fairness and accuracy of artificial intelligence face recognition systems across genders and skin colors. While it was not to be used for commercial purposes, Amazon and Microsoft were accused of using the dataset to *"improve the accuracy of their facial recognition software"* [13].

4.2.5 *Exclusion (149/321).*

Exclusion refers to the failure to provide end-users with notice and control over how their data is being used [126]. Even without AI, computing products can covertly process data without informing users. Thus, AI technologies do not meaningfully change exclusion risks when the risk is isolated to just the covert processing of personal data (76/149). For example, a "trustworthiness" algorithm developed by a short-term homestay company covertly used publicly accessible social media posts to ascertain if a potential customer was trustworthy [67], but the use of AI in this case did not fundamentally change the privacy risk. We nevertheless found in our incident database that the requirements of AI technology *can* exacerbate exclusion risks by incentivizing the collection of large, rich datasets of personal data without securing consent (73/149).

For example, the Large-scale Artificial Intelligence Open Network (LAION) is a German non-profit organization that aims "to make large-scale machine learning models, datasets and related code available to the general public." In 2022, they released a large-scale dataset LAION-5B [120], the biggest openly accessible image-text dataset at the time[8]. These data have been used to train many other high-profile text-to-image models such as Stable Diffusion[9] and Google Imagen[10][39]. However, a person found that her private medical photographs were referenced in the public image-text dataset. She suspected that *"someone stole the image from my deceased doctor's files and it ended up somewhere online, and then it was scraped into this dataset"* [39]. Other models were found to be trained on "semi-public" personal data that were scraped from places like online forums, dating sites, and social media without users' awareness and consent (e.g., [3, 57, 164]). For example, Clearview AI built a private face recognition model trained on three billion photos that were *"scraped from Facebook, YouTube, Venmo and millions of other websites"* [85].

Prior work has shown that it can be challenging to ensure agency to any individual over their data regarding how data they have shared online can and cannot be used by such models [100], and that it can be deliberately made complex for individuals to remove their data from the dataset [22]. Additionally, when commercial AI models are "black boxes," the general public has no means to audit how personal data is used by AI (e.g., Clearview AI). Finally, "algorithmic inclusion" — i.e., ensuring that everyone is included in a system — is often seen as a more desirable way to build AI systems in the context of AI ethics. These "inclusive AI" approaches, however, need to be balanced against exclusion-based privacy risks [10, 12]: when more people's data are captured to build inclusive systems, those people may be subject to increased exclusion risk if their data is collected without adequate consent and control.

4.2.6 *Insecurity (17/321).*

Insecurity refers to carelessness in protecting collected personal data from leaks and improper access due to faulty data storage and data practices [126]. Products and services that include AI are subject to many of the same insecurity risks that result from poor operational security, unrelated to the capabilities and data requirements of AI (12/17). For example, our dataset includes a data breach where attackers hacked into Verkada,

[7] https://www.okcupid.com/about

[8] https://laion.ai/blog/laion-5b/

[9] https://stablediffusionweb.com/

[10] https://imagen.research.google/

a security startup that provides cloud-based security cameras with face recognition. This gave the attackers access to cameras that *"are capable of identifying particular people across time by detecting their faces, and are also capable of filtering individuals by their gender, the color of their clothes, and other attributes"* [34, 135]. These operational security mistakes are not unique to or exacerbated by AI technologies, even though the AI-enabled products and services that are hacked afford attackers access to compromised data that would otherwise not be accessible. We did, however, find instances in which the capabilities and/or data requirements of AI technologies directly exacerbated insecurity risks (5/17).

Sometimes AI technology can compromise end-user privacy in order to enable AI utility. For example, Allo, a messaging app that Google first launched in 2017, included an AI virtual assistant and automatic replies. The messenger was not end-to-end encrypted, allowing for AI models developed by Google to "read" users' chat content and personalize services for them [47].

We also found cases where AI technologies unexpectedly reveal the personal data on which they were trained. For example, Lee Luda, a chatbot trained on real-world text conversations, was found to expose the names, nicknames, and home addresses of the users whose data on which it was trained [63]. Similarly, services that use generative AI models to create realistic but fake human faces, have been shown to be able to reconstruct the raw personal data on which the models were trained [147].

Additional vulnerabilities can be introduced through the infrastructural data requirements entailed by AI technologies. For example, converting raw data into training-ready labeled data can require the exposure of raw personal data to human annotators. For example, iRobot hired gig workers to annotate audio, photo, and video data captured by their household robots to train AI models. However, some of these raw and sensitive photos were leaked online by the gig workers [50]. Cases like this illustrate how AI can blur the boundary between data *processing* risks and data *dissemination* risks — sometimes, the act of processing data through AI requires dissemination.

4.3 Data dissemination risks

Data dissemination threats result when personal information is revealed or shared by data collectors to third-parties [126]. AI technologies *create* new data dissemination risks by enabling new ways of revealing and spreading personal data; they also *exacerbate* data dissemination risks by increasing the scale and the frequency of the dissemination.

In our analysis, we found that AI technologies create new types of **exposure**, **distortion**, and **disclosure** risks (e.g., by reconstructing redacted content; by generating a realistic fake video of an individual; and by sharing AI-derived sensitive information about individuals with third-parties). We also found cases in which AI technologies exacerbated known **disclosure** risk (e.g., by sharing large-scale user data to third-parties to train models), and **increased accessibility** risk (e.g., by open-sourcing large-scale benchmark datasets containing user data).

4.3.1 Exposure (17/321). Exposure risks encompass revealing sensitive private information that people view as deeply primordial that we have been socialized into concealing [126]. Traditionally, these risks arise when an individual's private activities are recorded and disseminated to others without consent. AI technologies can create new types of exposure risks via generative techniques that can create, reconstruct, manipulate content (i.e., deepfake techniques) (10/17) and expose inferred sensitive end-user attributes predicted by AI/ML (e.g., one's interests [79]) (7/17).

Specifically, we found that AI can create new types of exposure risks by reconstructing censored or redacted content. For example, generative adversarial networks (e.g., TecoGAN [31]) have been used to clarify images of censored genitalia [91], and to "undress" people to create pornographic images without consent [27]. Deepfake applications such as DeepFaceLive[11] or DeepFaceLab[12] can be made to morph a non-consenting subject's face into pornographic videos. These deepfake technologies have been used to facilitate mass dog-piling and online harassment [16] and to create illegal online pornography businesses [4].

In our analysis, we also found that AI technologies create new risks that expose sensitive data, preferences, and intentions inferred by AI/ML. For instance, Flo, an app that tracks menstruation and ovulation, forecasts its users' menstrual cycle and ovulation. Despite promising to maintain the privacy of personal data, Flo allegedly shared customers' menstrual timing and intention to get pregnant with third-parties like Facebook [119]. AI can also be built to proactively disseminate incriminating information about individuals to the public. In Shenzhen, China, a system was implemented to detect jaywalking and other offenses captured by cameras. The system identifies offenders and displays their photographs, names, and social identification numbers on LED screens placed at road junctions [156].

4.3.2 Distortion (20/321). Distortion refers to disseminating false or misleading information about people [126]. Distortion risks are analogous to slander or libel, and have existed well before modern advances in AI. However, we found that AI technologies can create new types of distortion risks by exploiting others' identities to generate realistic fake images and audio that humans have difficulty discerning as fake [96, 139].

Some models can generate realistic audio of individuals. For example, Prime Voice AI, a text-to-voice generator, was misused to create the voices of celebrities to *"make racist remarks about Alexandria Ocasio-Cortez (the US House representative)"*, and that the AI-generated clips *"run the gamut from harmless, to violent, to transphobic, to homophobic, to racist."* [33, 53]. Other AI-created distortion risks are less egregious, but raise important questions about expectations around privacy in light of how generative AI can be used to simulate the likeness of those who have passed. For example, the filmmaker of a documentary was revealed to be using deepfake technology to create scenes, with the likeness of an actor who had passed away, for lines *"he wanted [Anthony] Bourdain's (the main character of the documentary) voice for but had no recordings of"* [77].

4.3.3 Disclosure (45/321). Whereas distortion is the dissemination of false or misleading information, disclosure risks encompass the act of revealing and improperly sharing people's personal data

[11] https://github.com/iperov/DeepFaceLive
[12] https://github.com/iperov/DeepFaceLab

[126]. Indeed, any computing product that collects and stores personal data can introduce disclosure risks. Our dataset includes cases where AI does not meaningfully change disclosure risks (17/45), such as sharing personal data with law enforcement or third-parties. Nevertheless, AI technologies create new types of disclosure risks by being able to derive or infer additional information beyond what is explicitly captured in the raw data. We also found AI technologies can exacerbate disclosure risks because the personal data used to train ML models are often shared with specific individuals or organizations.

Many of the disclosure risks we identified involved the creation of machine learning models that automatically infer undisclosed personal information about individuals (14/45). For example, the "Safe City" initiative in Myanmar used AI-infused cameras to identify faces and vehicle license plates in public places and alert authorities to individuals with criminal histories [5].

AI technologies can also exacerbate disclosure risks when personal data is shared by organizations to train machine learning models (14/45). For example, the UK's National Health Service partnered with Google to share mental health records and HIV diagnoses of 1.6 million patients to develop a model for detecting acute kidney injury [59].

4.3.4 Increased Accessibility (23/321). Increased accessibility refers to making it easier for a wider audience of people to access potentially sensitive information. We found incidents in which AI technologies exacerbated the scale of this risk via the public sharing of large-scale datasets, containing personal information, for the use of building and improving AI/ML models. In the AI/ML community, it is common practice to leverage open-source benchmark datasets to train AI/ML models. This open-source data sharing enables transparency and public audits of AI research and development. However, publicizing datasets also enables anyone to collect large amounts of personal data that may have otherwise been private, access-controlled, or difficult to find. For example, the "OkCupid dataset" contained data of almost seventy thousand users from the dating site OkCupid. The dataset contained personal information such as users' location, demographics, sexual preferences, and drug use. It was uploaded to Open Science Framework, a website that helps researchers to open source datasets and research software, to facilitate research on modeling dating behaviors [153].

4.4 Invasion risks

The final top-level category of privacy risk Solove outlined, Invasion, can be understood as the unwanted encroachment into an individual's personal space, choices, or activities [126]. Solove placed two sub-categories under invasion: intrusion and decisional interference. We found incidents where AI technologies exacerbated intrusion risks, in particular.

4.4.1 Intrusion (160/321). Intrusion risks encompass actions that disturb one's solitude in physical space [126]. For six of the 160 intrusion incidents we identified, we noted that the AI technologies described in the incident did not fundamentally change the risk described in the incident: the intrusion would have remained as described even without the capabilities and/or requirements of AI. One example is the use of digital screens in stores to show customers personalized ads [88]: the intrusion would remain, even if the system did not use AI. However, we identified two ways AI can exacerbate intrusion risks that increase their scale and ubiquity.

The capabilities of AI technologies (e.g., to identify a person and detect behaviors) *enable a centralized surveillance infrastructure that creates large-scale intrusion risks (113/160)*; the requirements of AI (e.g., access to vast troves of data and GPU servers) necessitate this infrastructure. For example, Pharmaceutical University in Nanjing, China, implemented a recognition system at various locations on campus to closely monitor students' attendance and learning behaviors [133, 163]. Similarly, employers are increasingly incorporating AI-infused workplace monitoring technologies that collect data from employees' smartwatches [131] and computer webcams [144] to track their performance, absence, and time-on-task.

The capabilities of AI can also *turn everyday products (e.g., doorbells, wristbands) into powerful nodes in a ubiquitous surveillance infrastructure (41/160).* For example, Ring, a smart doorbell that enables homeowners to monitor activities and conversations near where the doorbell is installed, has raised concern due to "the device's excessive ability" to capture data of an individual's neighbors [90]. Similarly, Amazon's Halo fitness tracker uses AI to analyze a user's conversations to highlight when and how often that user spoke in a manner that was indicative of their being "happy, discouraged, or skeptical" [101].

5 DISCUSSION

Our findings demonstrate the many ways modern advances in AI meaningfully change privacy risks relative to how we conceived of privacy risks prior to these advances, as captured by Solove's widely cited taxonomy of privacy [126]. Across the 321 AI privacy incidents we analyzed, roughly 7% of the cases did not involve privacy risks that were created or exacerbated by AI. For example, we encountered instances where a product that happened to include AI was subject to a data breach in which users' personal data was compromised [7]. Nevertheless, in approximately 93% of the cases we analyzed, the unique capabilities and data requirements of the AI technologies involved in the incident either created a new type of privacy risk, or exacerbated a known risk.

We found that the unique capabilities of AI create new types of privacy risks. For example, AI creates new data processing risks in its ability to identify the activity of individuals even with low-quality data, and in its ability to forecast future outcomes. AI creates a new category of phrenology/physiognomy risks by enabling the creation of spurious classifiers correlating physical attributes with social, emotional, and personality traits. AI creates new types of data dissemination risks in its ability to generate human-like media, e.g., by generating a realistic fake video of an individual. We also found that the data requirements of AI exacerbate privacy risks we have grappled with for decades. For example, AI technologies can lead to more pervasive, larger scale surveillance than before; exacerbate secondary use, exclusion, insecurity, disclosure, and increased accessibility risks in the processing and disseminating of personal data; and, increase the ways in which computing can intrude upon people's personal space.

Equipped with the knowledge of how AI *has* changed privacy risks, we first discuss how the current AI/ML methods fall short and only address a subset of the AI privacy risks identified in our taxonomy (Section 5.1). Then, we present our taxonomy as a living structure that can be expanded with risks documented by Solove's original taxonomy [126] in cases where we did not find matching incidents in our incident database (Section 5.2). In theory, future advances in and/or the use of AI may entail risks in these categories, so it is worth discussing them as privacy risks that AI may change in the future. Moreover, we discuss a number of ways we expect this taxonomy might be useful for both future research and practice (Sections 5.1.1 and 5.2.1).

5.1 Charting the design space for privacy-preserving AI/ML work

Our findings broaden the design space for privacy-preserving AI and ML. For example, a recent meta-review of HAI principles and guidelines argues that privacy in ML-driven systems centers around the protection, control, and agency over personal data [161]. Based on our findings, these criteria only consider a small subset of the AI privacy risks we identified: they consider some — but not all — of the data collection and processing risks exacerbated by AI, and do not at all consider the data processing and dissemination risks newly created by AI. In this section, we provide an overview of how the existing tools and approaches, that aim to help practitioners build privacy-preserving AI systems [87, 152, 161], fall short of effectively identifying and addressing many AI privacy risks.

Differential Privacy and Federated Learning. Differential Privacy (DP) [95] and Federated Learning (FL) [80] are commonly thought of as approaches to "privacy-preserving" machine learning where 1) the model output is insensitive to the presence or absence of data on an individual in a dataset, and 2) the model provider only learns and improves the model in an aggregated manner. Tools such as Diffprivlib[13] [60] and IBM Federated Learning[14] [60] have been used by practitioners to implement DP and FL into their ML products. When training an ML model, however, these approaches only apply to some data processing risks — e.g., so that the model can not be used to re-identify data of individuals from the model outputs — and not the full range of risks we discuss in our taxonomy. Owing to these shortcomings, organizations that commonly advocate for end-user privacy rights, like the Electronic Frontier Foundation (EFF), have argued against the use of these approaches when they are used as stand-ins for stronger privacy protections (e.g., as in the case of Google's attempt to replace third-party browser cookies with "Federated Learning of Cohorts") [35]. For example, the "criminality classifier" that takes in photos of people's faces and claims to predict their likelihood to be a criminal [154] could be built with a federated learning architecture. Doing so would not address the physiognomy risk inherent to the idea itself, nor the exclusion and disclosure risks arising from how the data is collected and the inferences shared without consent.

Data Privacy Auditing. Prior work has created data auditing tools, such as the Privacy Meter[15] [94], to help practitioners conduct privacy impact assessments on ML models. Doing so allows practitioners to quantify some privacy risks (e.g., membership inference attacks). However, because the Privacy Meter must be applied *after* the model is trained, it is inherently limited in its ability to mitigate against the risks that arise in the data collection and processing phases of work. In addition, similar to DP and FL, this approach takes a limited view of privacy and only applies to specific data processing risks — e.g., aggregation risks that arise from collective sensitive personal data in the training data.

Ethics Checklists and Toolkits. Prior work in AI ethics has introduced many toolkits to support practitioners in ethical AI development [152], some of which also surface privacy risks. For example, Microsoft's Harms Modeling[16] is an activity that includes design exercises and worksheets that help *"evaluate potential ways the use of a technology you are building could result in negative outcomes for people and society,"* including potential privacy risks. AI ethics checklists such as Deon[17] allow practitioners to *"add an ethics checklist to [their] data science projects,"* which include questions that make practitioners reflect on the collection, storage, and analysis of data containing PII (personally identifiable information). These checklists and toolkits could help practitioners consider a broader range of privacy risks described in our taxonomy (e.g., data collection and dissemination risks). However, these tools approach privacy risks monolithically, and at a high-level (e.g., privacy loss, PII exposure); they provide little guidance to practitioners to consider privacy risks newly created and/or exacerbated by AI (e.g., physiognomy, distortion risks). In other words, the use of such tools relies on practitioners' individual awareness of AI privacy risks, which prior work has identified as a key barrier to AI privacy work [76].

Note that *all* of these approaches have value and we are not suggesting that they not be used. Rather, we caution against rhetoric that it is possible to create "privacy-preserving" AI/ML technologies using *only* these approaches.

5.1.1 Future Work: Creating AI-specific privacy guidance. Given that our findings show that AI creates new types of privacy risks and exacerbates existing ones, and that current privacy-preserving AI/ML methods fall short of identifying and addressing many of these risks, there is a need for future work to fill the gap of mitigating privacy risks created and exacerbated by AI. Specifically, our taxonomy opens up a new design space for privacy-preserving AI/ML tools that aim to raise practitioners' awareness of utility-intrusiveness trade-offs of their AI product ideas (e.g., [41]). For example, prior work in other AI-adjacent fields, such as Robotics, has explored how to correlate desired robot function with a minimally-invasive set of sensors [40]. In the broader context of implementing privacy and security in software products, prior work has found that practitioners still largely see privacy and security in products as an "all or nothing" notion such that privacy comes at the expense of other important objectives [51, 130].

Future work can explore incorporating our AI privacy taxonomy into harm-envisioning techniques, such as Consequence Scanning [38], by providing AI privacy risk prompts to capture associated

[13] https://github.com/IBM/differential-privacy-library
[14] https://github.com/IBM/federated-learning-lib
[15] https://github.com/privacytrustlab/ml_privacy_meter
[16] https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/
[17] https://github.com/drivendataorg/deon

negative consequences holistically. These techniques can help practitioners run lightweight privacy evaluations on AI product ideas, and help them balance the utility and intrusiveness of these products and services across design iterations. With such a tool, we hypothesize that practitioners can better advocate and design for privacy in working contexts that may dissuade this work [76, 130].

Our taxonomy can also consolidate promising future research in foregrounding tensions across data pipelines, practices, and stakeholders (i.e., data subjects, data observers, data beneficiaries, and data victims). By mirroring the first step in Rahwan's Society in the Loop framework [111], AI practitioners can make concrete the envisioned value and the stakeholders of their proposed AI concepts. To assist in this process, future work can create artifacts that encourage practitioners to articulate the value proposition of their envisioned product. Based on our taxonomy, then, it may be possible to mine our database for AI privacy incidents about products that are "semantically" similar based on an articulated value proposition. By showing practitioners related AI privacy incidents, they might then be guided to reflect on the utility-intrusiveness trade-off of their envisioned AI product ideas: for whom that value is generated (i.e., data beneficiaries), whose data is processed to unlock that value (i.e., data subjects), who can be impacted by the data pipeline (i.e., data victims), and by which privacy risk (e.g., surveillance).

In practice, however, this type of early-stage discussion around AI utility and privacy risk can be challenging because: (i) practitioners do not necessarily understand the full potential and limitations of AI [159]; (ii) privacy is often treated as compliance with general regulatory mandates rather than a product-specific design choice [143]; and, (iii) practitioners do not have access to AI-specific tools that support their privacy work pertaining to the capabilities and requirements that AI brings to their products [76]. Accordingly, there is a need for a greater understanding of where such tools and artifacts might be effectively incorporated into practitioners' workflows.

5.2 Theoretical extensions to the AI privacy risks taxonomy

We see our taxonomy as a living structure that helps scaffold the conversation about how advances in AI change privacy risks. But just as the capabilities and requirements of AI may change with future advances, so too might AI privacy risks. One way we might envision future AI privacy risks is by exploring the four subcategories of privacy risk in Solove's original taxonomy [126] for which we did not find relevant incidents in our dataset: Interrogation, Blackmail, Breach of Confidentiality, and Decisional Interference. In the future, we may observe incidents where advances in AI meaningfully change or exacerbate these risks as well.

Interrogation. Interrogation risks encompass the covert collection of data while a subject is being actively questioned [126]. For example, lie detector tests entail interrogation risks — information beyond what an individual is saying is collected to assess the truthfulness of their words. We can envision AI both creating and exacerbating interrogation risks. Large Language Model-powered chatbots like ChatGPT, for example, could create new interrogation risks by imitating people and interacting with users in natural language, aiming to extract information from users. AI-infused affective computing technologies could exacerbate interrogation risks (e.g., [98]): using these technologies, it may be possible to draw inferences about an individual's demeanor from verbal (e.g., language use, tone) and non-verbal (e.g., body language, eye movements) cues.

Blackmail. Blackmail refers to coercing individuals by threatening to disclose private or sensitive information [126]. Generative AI technologies could create new instantiations of this risk by synthesizing fake but convincing content that may serve as evidence for blackmail. We already saw incidents where incriminating content was fabricated when describing the exposure and distortion risks in our taxonomy, but we did not see this fabricated content being used for blackmail in the incidents we analyzed. Moreover, by automating the process of gathering and compromising information at scale, AI can also exacerbate blackmail risks. As we have seen, ML algorithms can analyze vast datasets from social media, location services, and personal files to identify content that could be used as fodder for blackmail.

Breach of Confidentiality. Breach of Confidentiality refers to an interpersonal risk between two people where one party discloses something to the other in confidence, and the other party violates this confidence by sharing it with third-parties [126]. AI technologies could exacerbate the scale of this risk by enabling conversational agents capable of gaining users' trust and guiding them to share sensitive information. For example, attackers can deploy such AI systems in high-stakes scenarios like healthcare and finance, and pose threats of breaching the confidentiality of the users by sharing the sensitive information they shared with the agent to third-parties.

Decisional Interference. Decisional Interference concerns the unwanted influence over or constraint of an individual's choices or behavior by a third-party [126]. Solove specifically focuses on the government as the relevant third-party, but private institutions and enterprises can also be culprits for this category of risk. AI technologies can exacerbate decisional interference risks by enabling more personalized political propaganda (e.g., [122]). AI technologies might also exacerbate the scale of existing practices of online censorship toward political topics (e.g., [23]). Algorithms for personalized recommendation or persuasive technologies can also subtly guide user choices, sometimes in ways that align more with the goals of external entities (e.g., advertisers or political campaigns) than with the individual's own preferences or well-being.

5.2.1 *Future Work: Creating a living taxonomy of AI privacy risks.*

To our knowledge, our taxonomy is the first attempt to show how common AI requirements and capabilities map onto high-level privacy risks. As shown above, future AI privacy incidents can also expand the taxonomy. In addition, future AI privacy incidents may create new categories of privacy risk that go beyond Solove's taxonomy (like the physiognomy risk we describe here). For example, many artists have been vocal about concerns about the theft of artistic style by generative AI [72]. While these discussions currently center around notions of copyright and intellectual property, we can envision new types of privacy risk as well: e.g., artistic styles might contain personally identifiable or sensitive information. We envision that our taxonomy can complement ongoing crowd-sourced efforts at curating and organizing AI incidents such as the AIAAIC [108] and AIID[18] by providing a framework to formally synthesize and identify emerging privacy risks in AI incidents. With that in mind, the research team is building a website[19] to present our taxonomy of AI privacy risks, and is also planning to expand this website to collect and aggregate submissions of new incidents related to these risks.

To present the AI privacy taxonomy in forms useful to the HCI and AI communities, future work can take an iterative approach, grounded on practitioners' and academics' actual design and research needs, to model the translation function between AI technology ideas and potential risks to consider. Indeed, envisioning with AI — i.e., treating AI as a design material [64, 158–160] — is an open and active area of research. Aligning with this line of research, future work can add to our taxonomy by systematizing AI capabilities and requirements, and the privacy risks they create and exacerbate, at a level of granularity that is useful for practitioners and researchers to ideate, communicate, and collaborate with product teams and stakeholders [159, 160].

5.3 Limitations

We consciously took an "incident-based" approach when constructing our taxonomy. There is a great deal of hype about what AI technologies can do, blurring the lines between speculation and reality [68]. The overabundance of speculative risks necessitated that we limit our consideration to those that journalists and the public-at-large have recognized as harmful as chronicled in the AIAAIC database. With that in mind, our dataset should not be interpreted as inclusive and representative of every *possible* privacy risk created or exacerbated by AI technologies: it is a repository of many privacy risks that have been realized in practice.

Our goal in creating this taxonomy was to codify AI privacy risks based on an accounting of documented, real-world risks. To that end, AIAAIC is currently *"the most comprehensive, detailed, and timely resource"*[20] that is openly accessible and has been used by the community as the source to synthesize the harms caused by AI functionality [113]. To mitigate the sampling bias introduced by our use of the AIAAIC, we tested the database's coverage by independently collecting a list of 15 AI privacy incidents from various sources, e.g., social media posts, literature. Of the 15 incidents we collected, 13 were also included in AIAAIC. For the two incidents that were not included, we found very similar incidents in the database — i.e., similar privacy risks caused by the same technology (e.g., face recognition software) but of different products. As a comparison, we applied the same procedure to the AIID database and only found five incidents included. Thus, we believe that AIAAIC currently provides a pool of AI privacy accidents comprehensive enough for our goal.

We acknowledge that there will be a growing number of AI incidents, and that there may be existing AI incidents that were not captured in our dataset. For example, prior work has surfaced how algorithmic recommender systems can amplify embarrassing exposures through online social networks [30]. Nevertheless, our taxonomy provides a solid foundation for understanding how the capabilities and requirements of AI change privacy risks. Since we ground our taxonomy on Solove's taxonomy of privacy, which has remained highly influential and largely appropriate for nearly two decades, we are confident that our updated taxonomy can be flexibly adapted to encompass new risks if and as they are realized beyond academic inquiry.

Finally, we acknowledge that "privacy" is a broad and context-dependent concept that is susceptible to biased interpretation based on the research team's background. We are an interdisciplinary research team with diverse expertise across HCI, AI, security and privacy, policy, and design. We mitigated the potential for bias by: (i) building our taxonomy on top of Solove's existing and widely accepted taxonomy; (ii) ensuring that multiple coders independently agreed on the risks entailed (or not) by a specific incident; and, (iii) dutifully analyzing *all* incidents, in the AIAAIC database, that were independently characterized by people outside of our research as being privacy-pertinent.

6 CONCLUSION

In this paper, we conducted a systematic analysis of documented incidents of AI privacy risks to answer the question: How do modern advances in AI and ML change privacy risks? Our taxonomy, constructed from a corpus of 321 documented AI privacy incidents, reveals that while the incorporation of AI technologies into products does not *necessarily* change the privacy risks those products might entail, it often does. Our taxonomy reveals that AI can create new types of privacy risks when processing and disseminating end-user data. We showed, for example, that the unique capabilities of AI technologies (e.g., the ability to generate realistic but fake images) also create new types of privacy risks (e.g., exposure risks from deepfake pornography [16]). The taxonomy also reveals that the data requirements of AI technologies can exacerbate known privacy risks. For example, owing to the unique ability of AI to automatically identify individuals from low-fidelity images, governments are more motivated to capture facial images of all passengers that pass through major transportation hubs (e.g., [42]). Our work suggests that AI-specific design guidance is needed for practitioners to negotiate the utility-intrusiveness trade-offs of AI-powered user experiences, and that many existing approaches to privacy-preserving machine learning (e.g., federated learning [80]) address only a small subset of the unique privacy risks entailed by AI technologies.

ACKNOWLEDGMENTS

This work was generously funded, in part, by NSF SaTC grants #2316768 and #2126066. We thank our anonymous reviewers for their helpful comments in revising this paper. We also thank the contributors and maintainers of the AIAAIC incident database.

REFERENCES

[1] 2012. Face to Face: Physiognomy & Phrenology THE SHELF. https://blogs.harvard.edu/preserving/2012/09/24/face-to-face-physiognomy-phrenology/

[2] 2018. Amazon patents 'voice-sniffing' algorithms. *BBC News* (April 2018). https://www.bbc.com/news/technology-43725708

[18] https://incidentdatabase.ai/

[19] The website will be available at https://privacytaxonomy.ai/ and https://aiprivacytaxonomy.com/

[20] https://www.aiaaic.org/aiaaic-repository/about-the-aiaaic-repository

[3] 2019. England's Keele University Neglects Patient Consent Regulations and Uses YouTube Videos to Study Autism in Children. https://www.trialsitenews.com/a/englands-keele-university-neglects-patient-consent-regulations-and-uses-youtube-videos-to-study-autism-in-children

[4] 2021. Deepfake porn case suspect is released on bail - Taipei Times. https://www.taipeitimes.com/News/front/archives/2021/10/20/2003766430 Section: Front Page.

[5] 2021. Myanmar: Facial Recognition System Threatens Rights. https://www.hrw.org/news/2021/03/12/myanmar-facial-recognition-system-threatens-rights

[6] 2021. UW-Madison disables proctoring software amid complaints. https://apnews.com/article/technology-madison-wisconsin-education-software-90a41fa6fa5348d837efbbd3be3a88f3

[7] Lawrence Abrams. 2020. ProctorU confirms data breach after database leaked online. https://www.bleepingcomputer.com/news/security/proctoru-confirms-data-breach-after-database-leaked-online/

[8] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[9] Muhammad Raisul Alam, Mamun Bin Ibne Reaz, and Mohd Alauddin Mohd Ali. 2012. A Review of Smart Homes—Past, Present, and Future. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (Nov. 2012), 1190–1203. https://doi.org/10.1109/TSMCC.2012.2189204 Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).

[10] Kendra Albert and Maggie Delano. 2022. Algorithmic Exclusion. https://doi.org/10.2139/ssrn.4122529

[11] Irwin Altman. 1975. *The environment and social behavior: Privacy, personal space, territory, crowding* (first printing edition ed.). Brooks/Cole Pub. Co, Monterey, Calif.

[12] McKane Andrus and Sarah Villeneuve. 2022. Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022).

[13] Katherine Anne Long. 2021. Amazon and Microsoft team up to defend against facial recognition lawsuits. https://www.seattletimes.com/business/technology/facial-recognition-lawsuits-against-amazon-and-microsoft-can-proceed-judge-rules/

[14] Blaise Aguera y Arcas. 2017. Physiognomy's New Clothes. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

[15] Rob Ashmore, Radu Calinescu, and Colin Paterson. 2021. Assuring the Machine Learning Lifecycle. *ACM Computing Surveys (CSUR)* 54 (5 2021). Issue 5. https://doi.org/10.1145/3453444

[16] Rana Ayyub. 2018. In India, journalists face slut-shaming and rape threats. *New York Times* 22 (2018).

[17] Chris Baraniuk. 2018. Exclusive: UK police wants AI to stop violent crime before it happens. https://www.newscientist.com/article/2186512-exclusive-uk-police-wants-ai-to-stop-violent-crime-before-it-happens/

[18] Ken Barker, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam Majedi, Sampson Pun, and Adepele Williams. 2009. A Data Privacy Taxonomy. In *Dataspace: The Final Frontier (Lecture Notes in Computer Science)*, Alan P. Sexton (Ed.). Springer, Berlin, Heidelberg, 42–54. https://doi.org/10.1007/978-3-642-02843-4_7

[19] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[20] Lois Beckett. 2019. Under digital surveillance: how American schools spy on millions of kids. *The Guardian* (Oct. 2019). https://www.theguardian.com/world/2019/oct/22/school-student-surveillance-bark-gaggle

[21] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://doi.org/10.48550/arXiv.1810.01943 arXiv:1810.01943 [cs].

[22] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. http://arxiv.org/abs/2110.01963 arXiv:2110.01963 [cs].

[23] BRENDAN BORDELON. 2023. 'We better figure it out': The politics trap that could slow a national AI law. https://www.politico.com/news/2023/05/19/ai-old-social-media-sam-altman-00097792

[24] Amine Boulemtafes, Abdelouahid Derhab, and Yacine Challal. 2020. A review of privacy-preserving techniques for deep learning. *Neurocomputing* 384 (4 2020), 21–45. https://doi.org/10.1016/J.NEUCOM.2019.11.041

[25] Louis Brandeis and Samuel Warren. 1890. The right to privacy. *Harvard law review* 4, 5 (1890), 193–220.

[26] Thomas Brewster. 2021. A $2 Billion Government Surveillance Lab Created Tech That Guesses Your Name By Simply Looking At Your Face. https://www.forbes.com/sites/thomasbrewster/2021/04/08/a-2-billion-government-surveillance-lab-created-tech-that-guesses-your-name-by-simply-looking-at-your-face/?sh=5842d9b76b1f

[27] M. Burgess. 2021. *The Biggest Deepfake Abuse Site Is Growing in Disturbing Ways. WIRED.*

[28] Andrew Chester, Yun Sing Koh, Jörg Wicker, Quan Sun, and Junjae Lee. 2020. Balancing Utility and Fairness against Privacy in Medical Data. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. 1226–1233. https://doi.org/10.1109/SSCI47803.2020.9308226

[29] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 1999–2008. https://doi.org/10.1145/2470654.2466265

[30] Ben C. F. Choi, Zhenhui (Jack) Jiang, Bo Xiao, and Sung S. Kim. 2015. Embarrassing Exposures in Online Social Networks: An Integrated Perspective of Privacy Invasion and Relationship Bonding. *Information Systems Research* 26, 4 (2015), 675–694. https://www.jstor.org/stable/24700367 Publisher: INFORMS.

[31] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics* 39, 4 (Aug. 2020), 75:75:1–75:75:13. https://doi.org/10.1145/3386569.3392457

[32] Danielle Keats Citron, Daniel J Solove, Kimia Favagehi, Katherine Grabar, Jean Hyun, Austin Mooney, Julia Schur, Rebecca Weitzel, Kenneth Abraham, Alessandro Acquisti, Rachel Bayefsky, Ryan Calo, Ignacio Cofone, Bob Gellman, Woodrow Hartzog, Chris Hoofnagle, Lauren Scholz, Lior Strahelivitz, Ari Waldman, and Benjamin Zipursky. 2022. Privacy Harms. *Boston University Law Review* (2022).

[33] Joseph Cox. 2023. AI-Generated Voice Firm Clamps Down After 4chan Makes Celebrity Voices for Abuse. https://www.vice.com/en/article/dy7mww/ai-voice-firm-4chan-celebrity-voices-emma-watson-joe-rogan-elevenlabs

[34] Joseph Cox and Jason Koebler. 2021. Hacked Surveillance Camera Firm Shows Staggering Scale of Facial Recognition. https://www.vice.com/en/article/wx83bz/verkada-hacked-facial-recognition-customers

[35] Bennett Cyphers. 2021. Google's FLoC Is a Terrible Idea. https://www.eff.org/deeplinks/2021/03/googles-floc-terrible-idea

[36] Sauvik Das, Joanne Lo, Laura Dabbish, and Jason I. Hong. 2018. Breaking! A Typology of Security and Privacy News and How It's Shared. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173575

[37] Kevin C. Desouza, Gregory S. Dawson, and Daniel Chenok. 2020. Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons* 63, 2 (March 2020), 205–213. https://doi.org/10.1016/j.bushor.2019.11.004

[38] doteveryone. 2020. Consequence Scanning – an agile practice for responsible innovators – doteveryone. https://doteveryone.org.uk/project/consequence-scanning/

[39] Benj Edwards. 2022. Artist finds private medical record photos in popular AI training data set. https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/

[40] Stephen Eick and Annie I. Anton. 2020. Enhancing Privacy in Robotics via Judicious Sensor Selection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Paris, France, 7156–7165. https://doi.org/10.1109/ICRA40945.2020.9196983

[41] Sindhu Kiranmai Ernala, Stephanie S. Yang, Yuxi Wu, Rachel Chen, Kristen Wells, and Sauvik Das. 2021. Exploring the Utility Versus Intrusiveness of Dynamic Audience Selection on Facebook. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–30. https://doi.org/10.1145/3476083

[42] Ella Fassler. 2021. South Korea Is Giving Millions of Photos to Facial Recognition Researchers. https://www.vice.com/en/article/xgdxqd/south-korea-is-selling-millions-of-photos-to-facial-recognition-researchers

[43] Todd Feathers, Roshan Abraham, and Karl Bode. 2021. This App Claims It Can Detect 'Trustworthiness.' It Can't. https://www.vice.com/en/article/akd4bg/this-app-claims-it-can-detect-trustworthiness-it-cant

[44] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. https://doi.org/10.2139/ssrn.3518482

[45] Shaun Nichols in San Francisco. 2017. TV anchor says live on-air 'Alexa, order me a dollhouse' – guess what happens next. https://www.theregister.com/2017/01/07/tv_anchor_says_alexa_buy_me_a_dollhouse_and_she_does/

[46] Batya Friedman and Helen Nissenbaum. 1997. Software agents and user autonomy. In *Proceedings of the first international conference on Autonomous agents (AGENTS '97)*. Association for Computing Machinery, New York, NY, USA, 466–469. https://doi.org/10.1145/267658.267772

[47] Gennie Gebhart. 2016. Google's Allo Sends The Wrong Message About Encryption. https://www.eff.org/deeplinks/2016/09/googles-allo-sends-wrong-message-about-encryption

[48] Dave Gershgorn. 2021. GitHub and OpenAI launch a new AI tool that generates its own code. https://www.theverge.com/2021/6/29/22555777/github-openai-ai-tool-autocomplete-code

[49] Global Times. 2021. Xpeng apologizes for illegal collection of facial images after penalty - Global Times. https://www.globaltimes.cn/page/202112/1241489.shtml

[50] Eileen Guo. 2022. A Roomba recorded a woman on the toilet. How did screenshots end up on Facebook? https://www.technologyreview.com/2022/12/19/1065306/roomba-irobot-robot-vacuums-artificial-intelligence-training-data-privacy/

[51] Marco Gutfleisch, Jan H. Klemmer, Niklas Busch, Yasemin Acar, M. Angela Sasse, and Sascha Fahl. 2022. How Does Usable Security (Not) End Up in Software Products? Results From a Qualitative Interview Study. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 893–910. https://doi.org/10.1109/SP46214.2022.9833756

[52] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30, 1 (March 2020), 99–120. https://doi.org/10.1007/s11023-020-09517-8

[53] Maggie Harrison. 2023. Startup Shocked When 4Chan Immediately Abuses Its Voice-Cloning AI. https://futurism.com/startup-4chan-voice-cloning-ai

[54] Adam Harvey and Jules. LaPlace. 2021. *Exposing.ai.* https://exposing.ai

[55] Adam Harvey and LaPlace, Jules. 2021. Exposing.ai: People in Photo Albums. https://exposing.ai/datasets/pipa/

[56] Drew Harwell. 2018. Wanted: The 'perfect babysitter.'Must pass AI scan for respect and attitude. *Washington Post* 23 (2018).

[57] Kashmir Hill. 2020. The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics*. Auerbach Publications, 170–177.

[58] Camilla Hodgson. 2019. Fast-food chains consider trying license plate recognition in drive-throughs. https://www.latimes.com/business/la-fi-license-plate-recognition-drive-through-restaurant-20190711-story.html Section: Business.

[59] Hal Hodson. 2016. Revealed: Google AI has access to huge haul of NHS patient data. https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/

[60] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. 2019. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints* 1907.02444 [cs.CR] (July 2019).

[61] Tatum Hunter and Heather Kelly. 2022. With Roe overturned, period-tracking apps raise new worries. *Washington Post* (Aug. 2022). https://www.washingtonpost.com/technology/2022/05/07/period-tracking-privacy/

[62] Timo Jakobi, Maximilian von Grafenstein, Patrick Smieskol, and Gunnar Stevens. 2022. A Taxonomy of user-perceived privacy risks to foster accountability of data-based services. *Journal of Responsible Technology* 10 (July 2022), 100029. https://doi.org/10.1016/j.jrt.2022.100029

[63] Heesoo Jang. 2021. A South Korean Chatbot Shows Just How Sloppy Tech Companies Can Be With User Data. *Slate* (April 2021). https://slate.com/technology/2021/04/scatterlab-lee-luda-chatbot-kakaotalk-ai-privacy.html

[64] Anniek Jansen and Sara Colombo. 2023. Mix & Match Machine Learning: An Ideation Toolkit to Design Machine Learning-Enabled Solutions. In *Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, Warsaw Poland, 1–18. https://doi.org/10.1145/3569009.3572739

[65] Diego Jemio, Alexa Hagerty, and Florencia Aranda. 2022. The Case of the Creepy Algorithm That 'Predicted'Teen Pregnancy, Wired (2022). *URL: https://www.wired. com/story/argentina-algorithms-pregnancy-prediction* (2022).

[66] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[67] Poppy Johnston. 2022. Banned from Airbnb with no explanation? Here's why. https://au.finance.yahoo.com/news/banned-from-airbnb-023208437.html

[68] Sayash Kapoor and Narayanan, Arvind. 2023. AI Snake Oil. https://www.aisnakeoil.com/

[69] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 454–470. https://doi.org/10.1145/3532106.3533556

[70] Kate Kaye. 2022. Class tests Intel AI to monitor student emotions on Zoom - Protocol. https://www.protocol.com/enterprise/emotion-ai-school-intel-edutech

[71] Patrick Gage Kelley, Celestina Cornejo, Lisa Hayes, Ellie Shuo Jin, Aaron Sedley, Kurt Thomas, Yongwei Yang, and Allison Woodruff. 2023. "There will be less privacy, of course": How and why people in 10 countries expect {AI} will affect privacy in the future. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. 579–603. https://www.usenix.org/conference/soups2023/presentation/kelley

[72] Kate Knibbs. 2023. A New Tool Helps Artists Thwart AI—With a Middle Finger. *Wired* (Oct. 2023). https://www.wired.com/story/kudurru-ai-scraping-block-poisoning-spawning/

[73] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security* 64 (2017), 122–134. https://doi.org/10.1016/j.cose.2015.07.002

[74] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I. Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3580882

[75] Marc Langheinrich. 2001. Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems. In *Ubicomp 2001: Ubiquitous Computing*, Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Gregory D. Abowd, Barry Brumitt, and Steven Shafer (Eds.). Vol. 2201. Springer Berlin Heidelberg, Berlin, Heidelberg, 273–291. https://doi.org/10.1007/3-540-45427-6_23

[76] Hao-Ping (Hank) Lee, Lan Gao, Stephanie Yang, Jodi Forlizzi, and Sauvik Das. 2024. "I Don't Know If We're Doing Good. I Don't Know If We're Doing Bad": Investigating How Practitioners Scope, Motivate, and Conduct Privacy Work When Developing AI Products. In *33nd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, PA.

[77] Radhamely De Leon. 2021. 'Roadrunner' Director Deepfaked Anthony Bourdain's Voice. https://www.vice.com/en/article/m7e54b/roadrunner-director-deepfaked-anthony-bourdains-voice

[78] Sam Levin. 2017. LGBT groups denounce 'dangerous' AI that uses your face to guess sexuality. *The Guardian* (Sept. 2017). https://www.theguardian.com/world/2017/sep/08/ai-gay-gaydar-algorithm-facial-recognition-criticism-stanford

[79] Sam Levin. 2017. New AI can guess whether you're gay or straight from a photograph. *The Guardian* (Sept. 2017). https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

[80] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* 37, 3 (2020), 50–60. ISBN: 1053-5888 Publisher: IEEE.

[81] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2020. When Machine Learning Meets Privacy: A Survey and Outlook. *Comput. Surveys* 54 (11 2020). Issue 2. https://doi.org/10.1145/3436755

[82] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, and Athanasios V. Vasilakos. 2021. Privacy and Security Issues in Deep Learning: A Survey. *IEEE Access* 9 (2021), 4566–4593. https://doi.org/10.1109/ACCESS.2020.3045078

[83] Karola Marky, Alexandra Voit, Alina Stöver, Kai Kunze, Svenja Schröder, and Max Mühlhäuser. 2020. "I don't know how to protect myself": Understanding Privacy Perceptions Resulting from the Presence of Bystanders in Smart Environments. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, Tallinn Estonia, 1–11. https://doi.org/10.1145/3419249.3420164

[84] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 81:1–81:32. https://doi.org/10.1145/3359183

[85] Logan Ryan Mac McDonald, Caroline Haskins. 2020. Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement Section: Tech.

[86] Cade Metz. 2019. Facial Recognition Tech Is Growing Stronger, Thanks to Your Face. *The New York Times* (July 2019). https://www.nytimes.com/2019/07/13/technology/databases-faces-facial-recognition-technology.html

[87] Christian Meurisch and Max Mühlhäuser. 2022. Data Protection in AI Services: A Survey. *Comput. Surveys* 54, 2 (March 2022), 1–38. https://doi.org/10.1145/3440754

[88] Nathaniel Meyersohn. 2022. Walgreens replaced some fridge doors with screens. And some shoppers absolutely hate it | CNN Business. https://www.cnn.com/2022/03/12/business/walgreens-freezer-screens/index.html

[89] Patrick Mikalef, Kieran Conboy, Jenny Eriksson Lundström, and Aleš Popovič. 2022. Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems* 31, 3 (May 2022), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

[90] Dan Milmo. 2021. Amazon asks Ring owners to respect privacy after court rules usage broke law. *The Guardian* (Oct. 2021). https://www.theguardian.com/uk-news/2021/oct/14/amazon-asks-ring-owners-to-respect-privacy-after-court-rules-usage-broke-law

[91] Hanako Montgomery. 2021. Man Arrested for Uncensoring Japanese Porn With AI in First Deepfake Case. https://www.vice.com/en/article/xgdq87/deepfakes-japan-arrest-japanese-porn

[92] Paul Mozur. 2018. Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras. *The New York Times* (July 2018). https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html

[93] Deirdre K. Mulligan, Colin Koopman, and Nick Doty. 2016. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Dec. 2016). https://doi.org/10.1098/rsta.2016.0118 Publisher: The Royal Society.

[94] Sasi Kumar Murakonda and Reza Shokri. 2020. ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339* (2020).

[95] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 634–646.

[96] Sophie J. Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8 (Feb. 2022), e2120481119. https://doi.org/10.1073/pnas.2120481119

[97] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (Feb. 2004), 119. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10

[98] Maria Noriega. 2020. The application of artificial intelligence in police interrogations: An analysis addressing the proposed effect AI has on racial and gender bias, cooperation, and false confessions. *Futures* 117 (March 2020), 102510. https://doi.org/10.1016/j.futures.2019.102510

[99] José Bernardi S. Nunes and Andrey Brito. 2023. A taxonomy on privacy and confidentiality. In *Proceedings of the 11th Latin-American Symposium on Dependable Computing (LADC '22)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3569902.3569903

[100] Jonathan A Obar. 2020. Sunlight alone is not a disinfectant: Consent and the futility of opening Big Data black boxes (without assistance). *Big Data & Society* 7, 1 (2020), 2053951720935615. https://doi.org/10.1177/2053951720935615

[101] Andrew O'Hara. 2021. Amazon Halo review: incredibly invasive, but helps you learn about yourself. https://appleinsider.com/articles/20/12/02/review-amazon-halo-is-incredibly-invasive-but-helps-you-learn-about-yourself

[102] Ayodeji Oseni, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. 2020. Security and Privacy for Artificial Intelligence: Opportunities and Challenges; Security and Privacy for Artificial Intelligence: Opportunities and Challenges. *J. ACM* 37 (2020). Issue 4. https://doi.org/10.1145/1122445.1122456

[103] Google PAIR. 2019. People + AI Guidebook. (2019).

[104] Leysia Palen and Paul Dourish. 2003. Unpacking "privacy" for a networked world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 129–136. https://doi.org/10.1145/642611.642635

[105] Rachel Pannett. 2022. German police used a tracing app to scout crime witnesses. Some fear that's fuel for covid conspiracists. *Washington Post* (Jan. 2022). https://www.washingtonpost.com/world/2022/01/13/german-covid-contact-tracing-app-luca/

[106] Minwoo Park. 2021. Seoul using AI to detect and prevent suicide attempts on bridges. *Reuters* (June 2021). https://www.reuters.com/world/asia-pacific/seoul-using-ai-detect-prevent-suicide-attempts-bridges-2021-06-30/

[107] Kari Paul. 2019. Google workers can listen to what people say to its AI home devices. *The Guardian* (July 2019). https://www.theguardian.com/technology/2019/jul/11/google-home-assistant-listen-recordings-users-privacy

[108] Charlie Pownall. 2023. AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC). https://www.aiaaic.org/

[109] Peter N. Henderson · The Canadian Press ·. 2015. Checked 'How Old Do I Look?' You gave Microsoft rights to your photo | CBC News. https://www.cbc.ca/news/science/how-old-do-i-look-microsoft-website-raises-privacy-concerns-1.3062176

[110] QTCinderella [@qtcinderella]. 2023. I want to scream. Stop. Everybody fucking stop. Stop spreading it. Stop advertising it. Stop. Being seen "naked" against your will should NOT BE A PART OF THIS JOB. Thank you to all the male internet "journalists" reporting on this issue. Fucking losers @HUN2R. https://twitter.com/qtcinderella/status/1620142227250094080

[111] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (March 2018), 5–14. https://doi.org/10.1007/s10676-017-9430-8

[112] Megha Rajagopalan. 2018. China Is Said To Be Using A Database To Identify And Detain People As Potential Threats. https://www.buzzfeednews.com/article/meghara/human-rights-watch-china-using-big-data-to-detain-people Section: World.

[113] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 959–972. https://doi.org/10.1145/3531146.3533158

[114] Reuters. 2018. China school using artificial intelligence to detect unfocused students. https://nypost.com/2018/05/17/china-is-using-ai-to-keep-high-school-students-in-line/

[115] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (2019), 33–36.

[116] Juliette Rihl. 2021. Emails show Pittsburgh police officers accessed Clearview facial recognition after BLM protests. http://www.publicsource.org/pittsburgh-police-facial-recognition-blm-protests-clearview/

[117] Charles Rollet. 2021. PRC Minority Ethnicity Recognition Research Targeted Uyghurs, Breached Ethical Standards. https://ipvm.com/reports/eth-rec-ethics

[118] Nithya Sambasivan and Jess Holbrook. 2018. Toward responsible AI for the next billion users. *Interactions* 26 (Dec. 2018), 68–71. https://doi.org/10.1145/3298735

[119] Sam Schechner and Mark Secada. 2019. You Give Apps Sensitive Personal Information. Then They Tell Facebook. *Wall Street Journal* (Feb. 2019). https://www.wsj.com/articles/you-give-apps-sensitive-personal-information-then-they-tell-facebook-11550851636

[120] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. https://doi.org/10.48550/arXiv.2210.08402 arXiv:2210.08402 [cs].

[121] Sakib Shahriar, Sonal Allana, Mehdi Hazrati Fard, and Rozita Dara. 2023. A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. *IEEE Access* (2023). https://doi.org/10.1109/ACCESS.2023.3287195

[122] Laura Courchesne Isra Thange Jacob N. Jon Bateman Shapiro, Elonnai Hickok. 2021. Measuring the Effects of Influence Operations: Key Findings and Gaps From Empirical Research. https://carnegieendowment.org/2021/06/28/measuring-effects-of-influence-operations-key-findings-and-gaps-from-empirical-research-pub-84824

[123] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. *Proceedings of the ACM Conference on Computer and Communications Security* 2015-October (10 2015), 1310–1321. https://doi.org/10.1145/2810103.2813687

[124] Tom Simonite. 2017. Facebook Can Now Find Your Face, Even When It's Not Tagged. *Wired* (Dec. 2017). https://www.wired.com/story/facebook-will-find-your-face-even-when-its-not-tagged/

[125] John Smith. 2019. IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems. https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/

[126] Daniel J. Solove. 2006. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154, 3 (Jan. 2006), 477. https://doi.org/10.2307/40041279

[127] Matthias Spielkamp. 2019. Automating Society: Taking Stock of Automated Decision-Making in the EU. BertelsmannStiftung Studies 2019. (2019).

[128] Luke Stark and Jevan Hutson. 2021. Physiognomic artificial intelligence. *Fordham Intell. Prop. Media & Ent. LJ* 32 (2021), 922.

[129] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (12 2013). https://arxiv.org/abs/1312.6199v4

[130] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. 2021. Privacy Champions in Software Teams: Understanding Their Motivations, Strategies, and Challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445768

[131] The Local. 2020. Spain's Mercadona supermarkets install facial recognition systems to keep thieves at bay. https://www.thelocal.es/20200702/spains-mercadona-supermarkets-install-facial-recognition-systems-to-keep-thieves-at-bay

[132] Daniel W. Tigard. 2021. Responsible AI and moral responsibility: a common appreciation. *AI and Ethics* 1, 2 (May 2021), 113–117. https://doi.org/10.1007/s43681-020-00009-0

[133] Sixth Tone. 2019. Camera Above the Classroom. https://www.sixthtone.com/news/1003759

[134] Hayley Tsukayama. 2021. Gmail's Inbox app will now write (some of) your e-mails for you. *Washington Post* (Dec. 2021). https://www.washingtonpost.com/news/the-switch/wp/2015/11/03/gmails-inbox-app-will-now-write-some-of-your-e-mails-for-you/

[135] William Turton. 2021. Hackers breach thousands of security cameras, exposing tesla, jails, hospitals. *Bloomberg. Available online at: https://www.bloomberg.com/news/articles/2021-03-09/hackersexpose-tesla-jails-in-breach-of-150-000-security-cams* (2021).

[136] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/2335356.2335362

[137] Muhammad Usman, Ricardo Britto, Jürgen Börstler, and Emilia Mendes. 2017. Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method. *Information and Software Technology* 85 (May

2017), 43–59. https://doi.org/10.1016/j.infsof.2017.01.006

[138] Chris Velazco. 2021. Amazon's 'Mentor' tracking software has been screwing drivers for years. https://www.engadget.com/amazon-mentor-app-edriving-delivery-driver-tracking-surveillance-gps-194346304.html

[139] James Vincent. 2022. *Binance executive claims scammers made a deepfake of him-the verge.*

[140] Eduardo Vyhmeister, Gabriel G. Castane, Johan Buchholz, and Per-Olov Östberg. 2022. Lessons learn on responsible AI implementation: the ASSISTANT use case. *IFAC-PapersOnLine* 55, 10 (Jan. 2022), 377–382. https://doi.org/10.1016/j.ifacol.2022.09.422

[141] Jane Wakefield. 2021. Amazon faces spying claims over AI cameras in vans. *BBC News* (Feb. 2021). https://www.bbc.com/news/technology-55938494

[142] Jane Wakefield. 2021. Neighbour wins privacy row over smart doorbell and cameras. *BBC News* (Oct. 2021). https://www.bbc.com/news/technology-58911296

[143] Ari Ezra Waldman. 2017. Designing Without Privacy. *Houston Law Review* (March 2017). https://papers.ssrn.com/abstract=2944185

[144] Peter Walker. 2021. Call centre staff to be monitored via webcam for home-working 'infractions'. *The Guardian* (March 2021). https://www.theguardian.com/business/2021/mar/26/teleperformance-call-centre-staff-monitored-via-webcam-home-working-infractions

[145] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology* 114, 2 (2018), 246. https://doi.org/10.1037/pspa0000098

[146] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L. Mazurek, Manya Sleeper, and Kurt Thomas. 2022. SoK: A Framework for Unifying At-Risk User Research. *Proceedings - IEEE Symposium on Security and Privacy* 2022-May (2022), 2344–2360. https://doi.org/10.1109/SP46214.2022.9833643

[147] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. 2021. This person (probably) exists. identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018* (2021).

[148] Alan F Westin. 1967. *Privacy And Freedom.*

[149] Chathurika S. Wickramasinghe, Daniel L. Marino, Javier Grandio, and Milos Manic. 2020. Trustworthy AI Development Guidelines for Human System Interaction. In *2020 13th International Conference on Human System Interaction (HSI).* IEEE, Tokyo, Japan, 130–136. https://doi.org/10.1109/HSI49210.2020.9142644

[150] Kyle Wiggers. 2021. AI datasets are prone to mismanagement, study finds. https://venturebeat.com/ai/ai-datasets-are-prone-to-mismanagement-study-finds/

[151] Maximiliane Windl, Verena Winterhalter, Albrecht Schmidt, and Sven Mayer. 2023. Understanding and Mitigating Technology-Facilitated Privacy Violations in the Physical World. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany). Association for Computing Machinery, New York, NY, USA, Article 585, 16 pages. https://doi.org/10.1145/3544548.3580909

[152] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–27. https://doi.org/10.1145/3579621

[153] Emma Woollacott. 2016. 70,000 OkCupid Profiles Leaked, Intimate Details And All. https://www.forbes.com/sites/emmawoollacott/2016/05/13/intimate-data-of-70000-okcupid-users-released/ Section: Tech.

[154] Xiaolin Wu and Xi Zhang. 2016. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135* (2016), 4038–4052.

[155] Yuxi Wu, Sydney Bice, W. Keith Edwards, and Sauvik Das. 2023. The Slow Violence of Surveillance Capitalism. *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (6 2023), 1826–1837. https://doi.org/10.1145/3593013.3594119

[156] Vicky Xiuzhong Xu and Bang Xiao. 2018. Chinese authorities use facial recognition, public shaming to crack down on jaywalking, criminals. *ABC News* 20 (2018).

[157] Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, Junjun Qiu, Keqin Hua, Wentao Su, Jian Wu, Huiyu Xu, Yong Han, Chenguang Fu, Zhigang Yin, Miao Liu, Ronald Roepman, Sabine Dietmann, Marko Virta, Fredrick Kengara, Ze Zhang, Lifu Zhang, Taolan Zhao, Ji Dai, Jialiang Yang, Liang Lan, Ming Luo, Zhaofeng Liu, Tao An, Bin Zhang, Xiao He, Shan Cong, Xiaohong Liu, Wei Zhang, James P. Lewis, James M. Tiedje, Qi Wang, Zhulin An, Fei Wang, Libo Zhang, Tao Huang, Chuan Lu, Zhipeng Cai, Fang Wang, and Jiabao Zhang. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* 2, 4 (2021), 100179. https://doi.org/10.1016/j.xinn.2021.100179

[158] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference.* ACM, Hong Kong China, 585–596. https://doi.org/10.1145/3196709.3196730

[159] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *CHI Conference on Human Factors in Computing Systems.* ACM, New Orleans LA USA, 1–13. https://doi.org/10.1145/3491102.3517491

[160] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference.* ACM, Pittsburgh PA USA, 2326–2346. https://doi.org/10.1145/3563657.3596058

[161] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul Republic of Korea, 535–563. https://doi.org/10.1145/3531146.3533118

[162] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir D. Bourdev. 2015. Beyond frontal faces: Improving Person Recognition using multiple cues. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 4804–4813.

[163] Phoebe Zhang. 2019. Chinese university says new classroom facial recognition system will improve attendance | South China Morning Post. https://www.scmp.com/news/china/science/article/3025329/watch-and-learn-chinese-university-says-new-classroom-facial

[164] Michael Zimmer. 2016. OkCupid Study Reveals the Perils of Big-Data Science. *Wired* (May 2016). https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/ Section: tags.

[165] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism : The Fight for a Human Future at the New Frontier of Power.* 691 pages.

A APPENDIX

Table 2: Cohen's Kappa for each privacy risk.

Privacy Risk	Cohen's Kappa
Surveillance	0.88
Identification	0.94
Secondary Use	0.84
Aggregation	0.87
Phrenology / Physiognomy	1
Exclusion	0.90
Insecurity	0.94
Exposure	1
Disclosure	1
Distortion	1
Increased Accessibility	1
Intrusion	0.94
Average	0.94