AI S&P Overview

INFR11158/11230 Usable Security and Privacy

Dr. Jingjie Li

14/03/2025

THE UNIVERSITY *of* EDINBURGH

Overview

- Finally week**s**
- Privacy issues of AI
- Recap – privacy policy
- Guest lecture

Snapchat: Snap AI chatbot 'may risk children's privacy'

6 October 2023

GETTY IMAGES

By Shiona McCallum
Technology reporter

Snapchat has been accused of a "worrying failure" to assess the potential privacy risks its AI chatbot poses to users - especially children - by the UK's data watchdog.

The Information Commissioner's Office (ICO) warned it could close down the My AI feature in the UK after a "preliminary investigation".

The US company said it was "closely reviewing" the provisional findings.

https://www.bbc.co.uk/news/technology-67027282

Think and Share: What are the NEW security, privacy or safety risks brought by new AI models, e.g., generative AI?

AI Security

- The “cognitive” process of AI does not always align with how human perceive and think about the world


STOP
HATE

https://spectrum.ieee.org/slight-street-sign-modifications-can-fool-machine-learning-algorithms

Prompt injection attack

DeepSeek AI Models Vulnerable to Jailbreaking

Data Exposure, Harmful Content and Security Risks Undermine DeepSeek AI Models

Akshaya Asokan (asokan_akshaya) • January 31, 2025

Research from Palo Alto's Unit 42, Kela and Enkrypt AI identified susceptibility to jailbreaking and hallucinations in the Chinese company's recently unveiled R1 and V3 models. Cybersecurity firm Wiz disclosed Wednesday that DeepSeek exposed a real-time data processing database to the open internet, allowing security researchers to view chat history and backend data (see: *Breach Roundup: DeepSeek Leaked Sensitive Data*).

Vaccine misinformation can easily poison AI – but there's a fix

Adding just a little medical misinformation to an AI model's training data increases the chances that chatbots will spew harmful false content about vaccines and other topics

By Jeremy Hsu

8 January 2025

Adversarial Examples

- Definition
  - Inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.
- Impact
  - Leads to incorrect AI decisions or misclassifications that seem correct to human operators.
- Methodology
  - Creating input samples that are slightly altered but cause significant errors in AI outputs.
  - Exploiting model vulnerabilities that are not easily detectable by humans.
- Countermeasures
  - Employing adversarial training methods.
  - Regularly updating and testing models against known adversarial attack techniques.

Prompt Injection

- Definition
  - Manipulation of AI's response by altering the input prompt or commands it receives.
- Impact
  - Can cause AI to produce undesired, biased, or harmful outputs.
- Methodology
  - Craft malicious input prompts to mislead AI.
  - Inject misleading context or information into the AI's operational environment.
- Countermeasures
  - Robust input validation and sanitization.
  - Implementation of authentication protocols to verify source integrity.

Data Poisoning

- Definition
  - Introducing malicious data into the AI’s training set to corrupt its learning process.
- Impact
  - Results in a corrupted model that makes errors or biased decisions.
- Methodology
  - Insertion of subtly incorrect or biased data points into the training dataset.
  - Targeted manipulation to influence specific AI behaviors or outcomes.
- Countermeasures
  - Regular audits of training data.
  - Use of anomaly detection techniques to identify and remove corrupted data.

Inversion Attack

- Definition
  - Techniques used to extract knowledge, sensitive data, or even the entire model from an AI system.
- Impact
  - Loss of intellectual property, exposure of sensitive training data, or compromising model integrity.
- Methodology
  - Querying a model repeatedly to infer its structure or training data.
  - Utilizing side-channel attacks to gain insights into the model's operations.
- Countermeasures
  - Implementing rate limiting and query monitoring to detect and prevent extraction attempts.
  - Using model hardening techniques to obscure internal operations and outputs.

Who are involved in the threat model?

Privacy risks of AI

A TAXONOMY OF PRIVACY

INFORMATION PROCESSING

AGGREGATION
Combining of various pieces of personal information

A credit bureau combining an individual's payment history from multiple creditors.

SECONDARY USE
Using personal information for a purpose other than the purpose for which it was collected

The U.S. Government using census data collected for the purpose of apportioning Congressional districts to identify and intern those of Japanese descent in WWII.

EXCLUSION
Failing to let an individual know about the information that others have about them and participate in its handling or use

A company using customer call history, without the customer's knowledge, to shift their order in a queue (i.e. "Your call will be answered in the order [NOT] received")

INSECURITY
Failing to protect information

An ecommerce website allowing others to view an individual's purchase history by changing the URL (e.g. enterprivacy.com?id=123)

IDENTIFICATION
Linking of information to an individual. [Sometimes called 'singling out']

A researcher linking medical files to the Governor of a state using only date of birth, zip code and gender.

COLLECTION

SURVEILLANCE
Watching, listening to, or recording of a person's activities

A website monitoring cursor movements of a visitor while visiting the website.

INTERROGATION
Questioning or probing for personal information

An interviewer asking an inappropriate question, such as marital status, during an employment interview.

INVASION

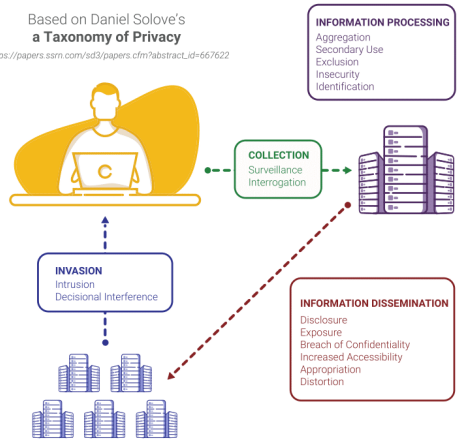
INTRUSION
Disturbing a person's tranquility or solitude

An augmented reality game directing players onto private residential property.

DECISIONAL INTERFERENCE
Intruding into a person's decision making regarding their private affairs

A payment processor declining transactions for contraceptives.

Based on Daniel Solove's
a Taxonomy of Privacy
https://papers.ssrn.com/sd3/papers.cfm?abstract_id=667622



INFORMATION DISSEMINATION

DISCLOSURE
Revealing truthful information about a person that impacts their security or the way others judge their character

A government agency revealing an individual's address to a stalker, resulting in the individual's murder.

EXPOSURE
Revealing a person's nudity, grief, or bodily functions

A store forcing a customer to remove clothing revealing a colostomy bag.

BREACH OF CONFIDENTIALITY
Breaking a promise to keep a person's information confidential.

A doctor revealing patient information to friends on a social media website.

INCREASED ACCESSIBILITY
Amplifying the accessibility of personal information

A court making proceeding searchable on the Internet without redacting personal information.

APPROPRIATION
Using an individual's identity to serve the aims and interests of another

A social media site using customer's images in advertising.

DISTORTION
Disseminating false or misleading information about a person

A creditor reporting a paid bill as unpaid to a credit bureau.

Version 6 (2022)

PRIVACY BY DESIGN

https://privacybydesign.training

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

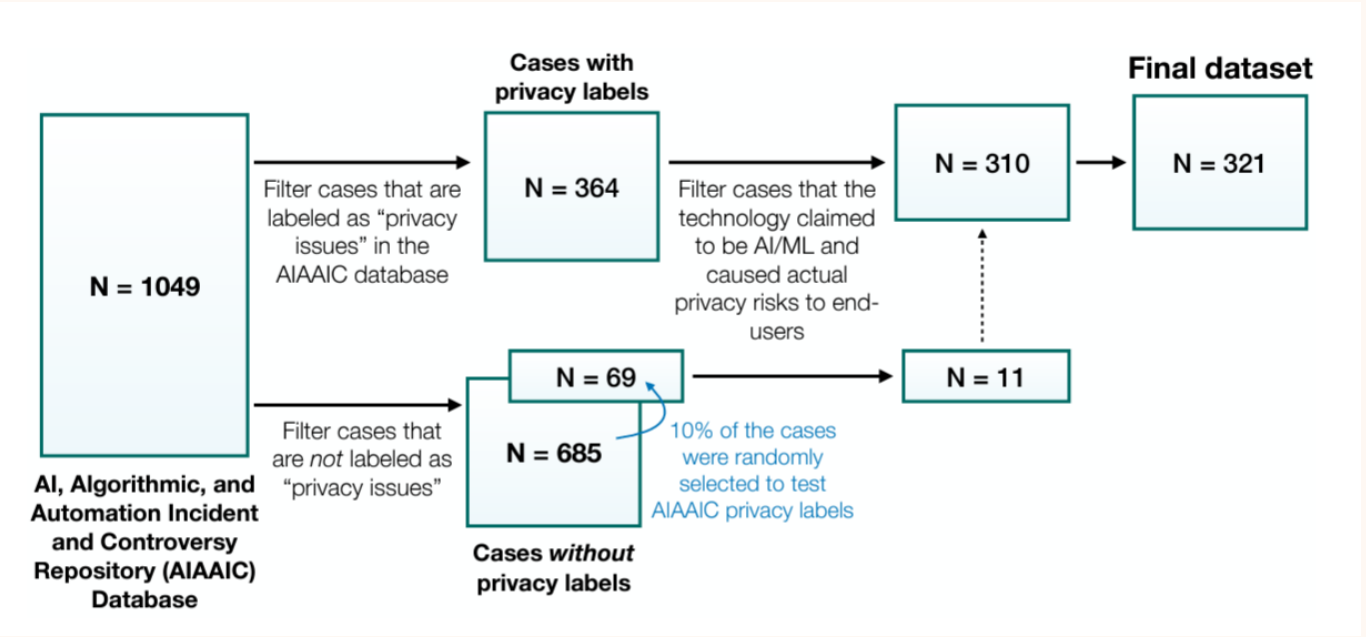

Capabilities of AI
Identify individuals
Generate images
Discover personal attributes
Forecast user behaviors
Estimate personal attributes
Requirements of AI
Share training data
Protect training data
Process training data
Collect training data

Objective

- Develop a privacy taxonomy for AI privacy risks
- What's AI?
  - "perform tasks or behaviors that a person could reasonably deem to require intelligence if a human were to do it" – an umbrella definition

Method

- Materials: AI incident database
- Approach: qualitative coding and analysis
  - Top-down/deductive coding: Solove's privacy taxonomy



Data flow

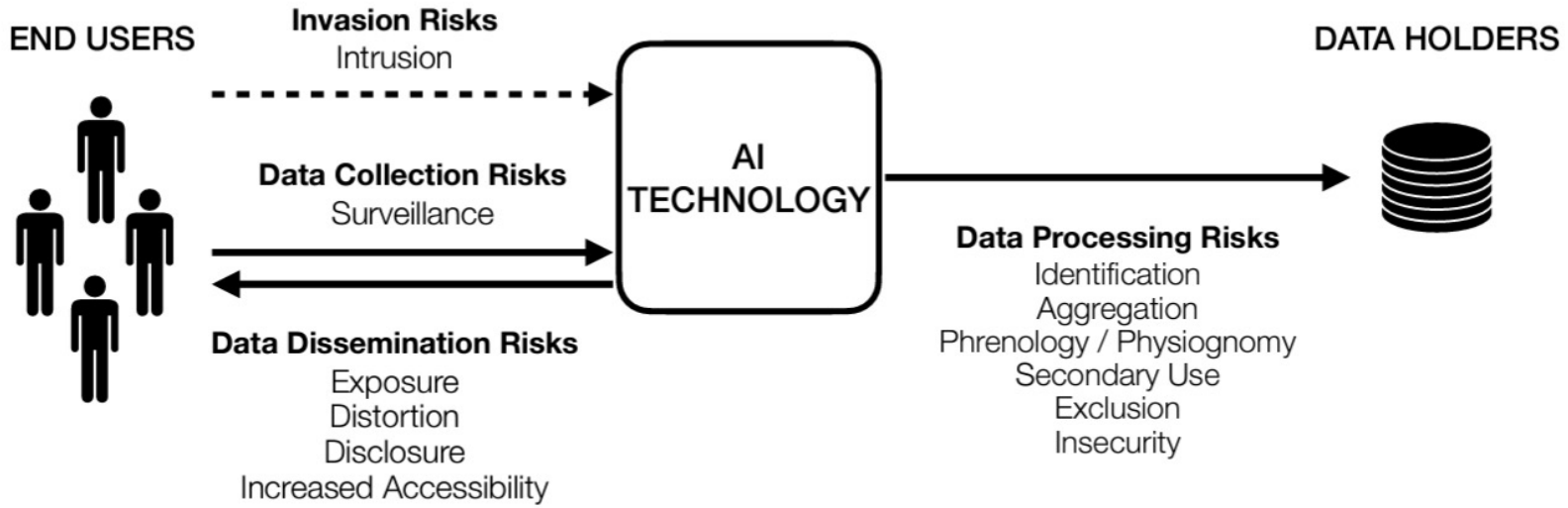
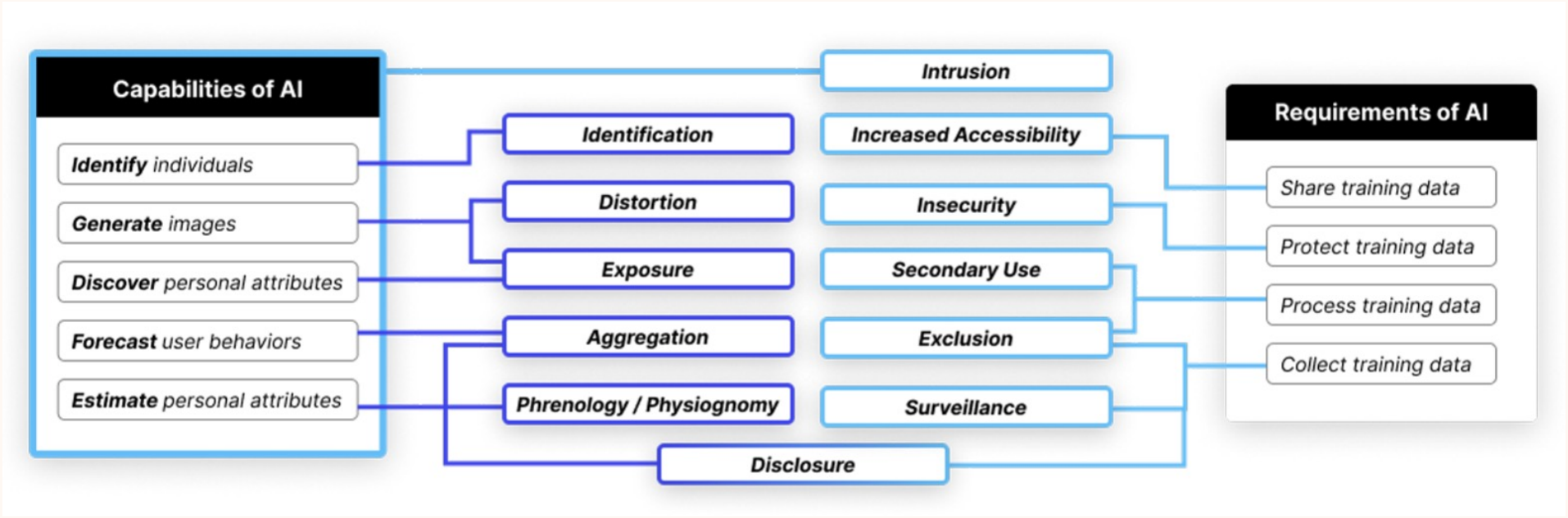


Figure 3: 12 types of privacy risks that AI technologies create and/or exacerbate relate to data collection, data processing, data dissemination, and invasion. The arrows indicate data flow (invasion risks need not involve data, but often do).


Capabilities of AI
Identify individuals
Generate images
Discover personal attributes
Forecast user behaviors
Estimate personal attributes
Identification
Distortion
Exposure
Aggregation
Phrenology / Physiognomy
Intrusion
Increased Accessibility
Insecurity
Secondary Use
Exclusion
Surveillance
Disclosure
Requirements of AI
Share training data
Protect training data
Process training data
Collect training data

Takeaway

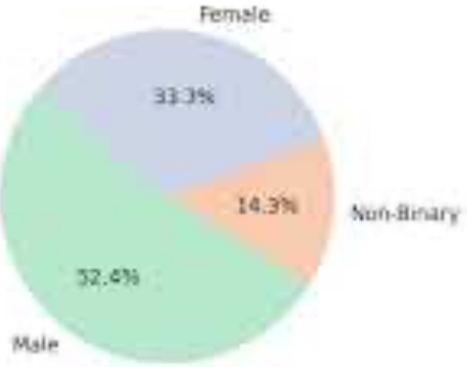
- New risks of **phrenology/physiognomy**
  - New, unfounded traits users have little control
- New types of **identification and aggregation** risks (forecasting, low quality data...)
- Exacerbate other risks (**secondary use, exclusion, insecurity**...)

Think and share: what could be the (new) approaches you use to mitigate these risks?

Participant Statistics

N=21

Age Distribution



Gender Identity Distribution

Take-home

- **(Blog)** Iqbal, Umar, Tadayoshi Kohno, and Franziska Roesner. "LLM platform security: Applying a systematic evaluation framework to OpenAI's ChatGPT plugins." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, pp. 611-623. 2024. https://ojs.aaai.org/index.php/AIES/article/view/31664/33831
- **(Blog)** West, Jack, et al. "A Picture is Worth 500 Labels: A Case Study of Demographic Disparities in Local Machine Learning Models for Instagram and TikTok." *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024.