CYB-4203/6203

Secure and Trustworthy AI

Presentation 15: Privacy Risks and Bias in AI Systems

Monday, March 30, 2026

Today's Agenda

- Midterm Project LLM-assisted Development Cycle video and workflow outline
- 7.1 Privacy risks in AI/ML: membership inference, surveillance, predictive harm
  - Privacy regulation and compliance challenges
- 7.2 Bias in AI systems: types, sources, and real-world impact

Midterm Project LLM-assisted Development Cycle Video & Outline

A Practical Iterative Workflow

1. Review
2. Brainstorm
3. Research
4. Plan
5. Execute

Repeat…

https://www.youtube.com/watch?v=9IXaV471mCs

LLM-Assisted Development Workflow: Generalized Outline

This document outlines the Review, Brainstorm, Research, Plan, Execute workflow demonstrated in the Midterm Project video walkthrough. The workflow is tool-agnostic — you can follow it with Claude Code, Gemini CLI, OpenAI Codex, Cursor, Copilot, or any LLM assistant.

Phase 1: Review

Goal: Load context and identify what you're working with.

What happens in this phase:

- Feed the LLM your prior work (Assignments 5 and 6, the project requirements, any notes)
- Ask it to summarize the key constraints, requirements, and options
- Identify your two strongest attack vector candidates for interactive demonstration

Tips for getting the best results:

- Start by explicitly telling the LLM what files or documents to read — don't assume it knows your context
- Ask the LLM to identify gaps or ambiguities in the requirements before you start building
- If the LLM summarizes your work, check that the summary is accurate — models sometimes hallucinate details about your own documents
- Use this phase to narrow scope, not expand it: "Which two of these five attack vectors might lend itself best to an interactive demonstration? Which might be the simplest to implement?"

Output: `review.md` — a short document capturing the context, constraints, and your candidate attack vectors.

Phase 2: Brainstorm

Goal: Explore what the artifact could look like without writing code.

What happens in this phase:

- Describe possible artifact concepts to the LLM and ask it to help you evaluate them

In Harvey – Midterm Project Assignment Instructions

7.1: Privacy Risks in AI/ML

Membership inference, surveillance, and predictive harm

Privacy Risks in AI/ML

AI/ML fundamental privacy challenges

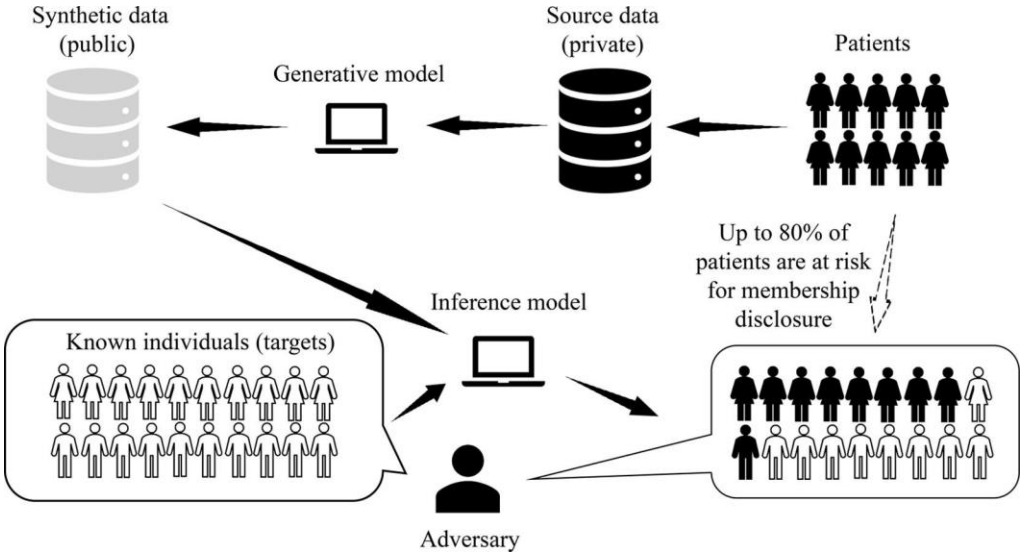
- AI systems consume, process, and memorize personal data at unprecedented scale
- Models can reveal information about individuals that was never explicitly provided
- Traditional privacy frameworks (notice and consent) struggle with AI's complexity
- Privacy is both a security concern (attacks) and an ethical/legal concern (rights)
- Connection to Unit 6: privacy attacks are a subset of the broader attack landscape

Privacy Risks in AI/ML

Membership Inference Attacks

Was your data used to train this model?

- Determining whether a specific data record was in the model's training set
- Exploits the difference in model confidence between training data and unseen data
- Privacy implications: confirms participation in sensitive datasets (health, finance, location)
- Zhang et al. (2022): Even partially synthetic HRE data susceptible to privacy intrusions such as membership inference

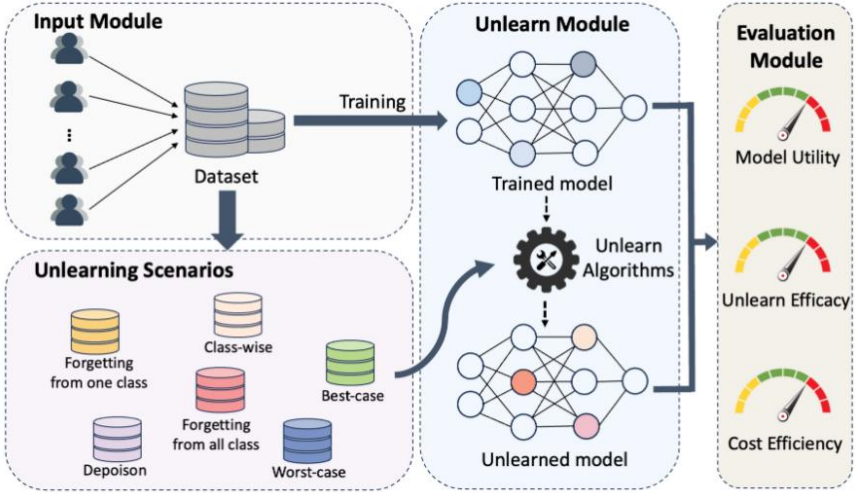


Zhang et al. (2022)

Privacy Risks in AI/ML

Why Membership Inference Matters

- Confirms presence in a sensitive dataset. Ex: genomic study, mental health program, criminal database
- GDPR Article 17 (right to erasure): if membership can be inferred, was data truly deleted?
- **Machine unlearning**: the challenge of actually removing a data point's influence from a trained model
- Verification problem: how do you prove a model has "forgotten" a data point?
- Active research area with no complete solution



Li et al. (2025) Figure 2: System overview of MUBox

Privacy Risks in AI/ML

Model Inversion Attacks

Reconstructing private datasets from model outputs

- Fredrikson et al. (2015): reconstructed recognizable faces from a facial recognition model
- The model's outputs encode enough information to reverse-engineer inputs
- Particularly dangerous for models trained on sensitive data (medical, biometric, financial)
- Gradient-based inversion: using model gradients to reconstruct training images
- Defenses: differential privacy, output perturbation, limiting confidence score precision

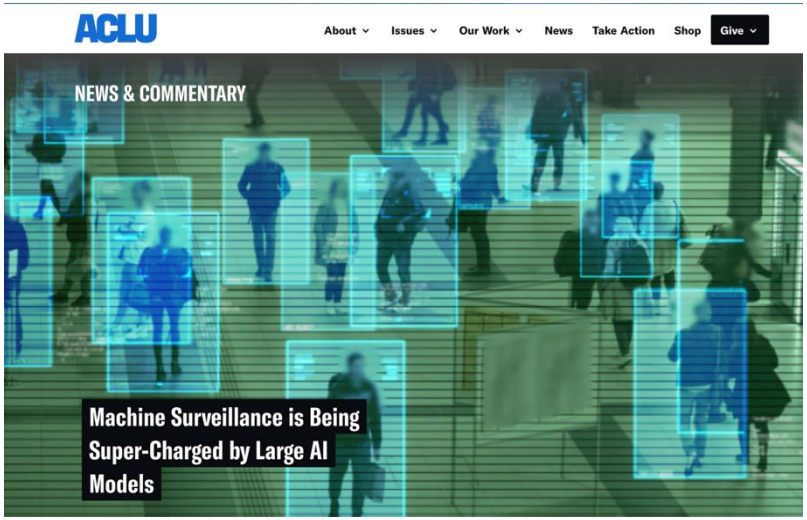
Frederikson et al. (2015) Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Privacy Risks in AI/ML

Surveillance and Predictive Harm

AI-enabled surveillance at scale

- Facial recognition: LLMs / AI making analytics much cheaper and more broadly available.
- Predictive policing: targeting individuals based on predicted behavior, not actual actions
- Social scoring systems: aggregating data to rank individuals on trustworthiness or risk
- Location tracking and movement prediction from mobile data
- The chilling effect: people change behavior when they know they are being watched



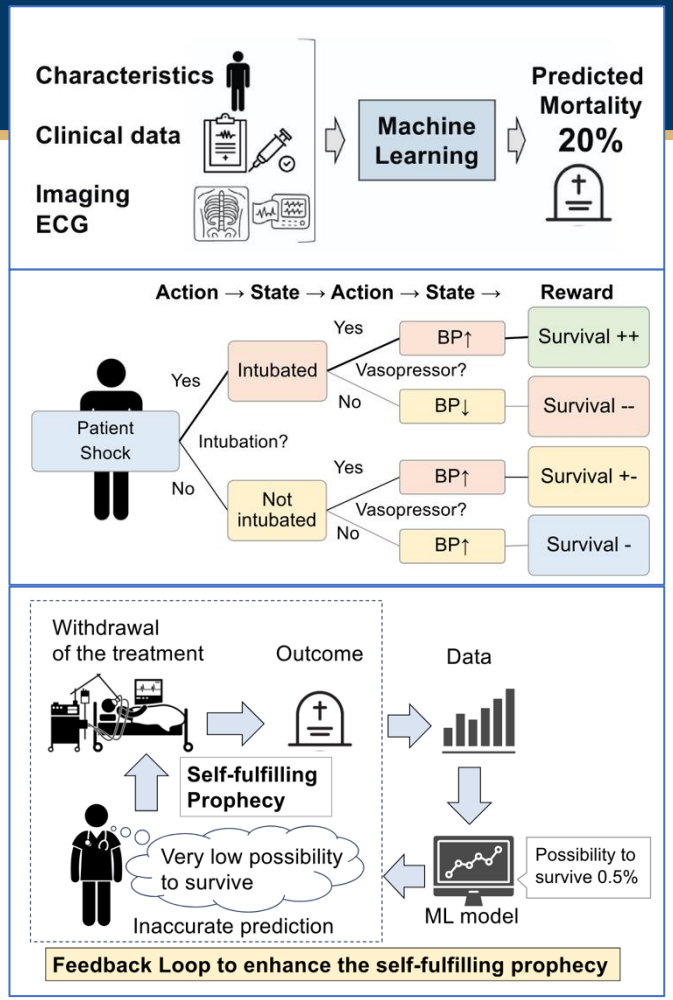
Jay Stanley – ACLU – March 2025

Privacy Risks in AI/ML

Predictive Harm

When predictions become self-fulfilling

- AI predictions can create the outcomes they predict (feedback loops)
- Predictive policing concentrates resources in targeted areas, increasing arrests, "confirming" predictions
- Credit scoring models deny loans to populations they predict will default, preventing economic mobility
- Health insurance risk predictions that become discriminatory gatekeeping
- The privacy harm is not just data exposure -- it is the use of data to constrain futures



Okada, et al. 2023

Mitigating Privacy Risks in AI/ML

Differential Privacy

Adding calibrated noise to data or outputs so that no individual record significantly affects the result

- Formal mathematical guarantee: epsilon-delta privacy budget
- Used by Apple (emoji usage), Google (Chrome), U.S. Census Bureau (2020 Census)
- Privacy-utility tradeoff: more privacy = more noise = less accurate results

Federated Learning

Training approach where models learn from decentralized data on local devices or servers without centralizing the raw data

- Raw data stays on the client or local node; only model updates are shared centrally
- Reduces central data exposure risk, but updates can still leak information without safeguards
- Common use cases: mobile keyboards, healthcare collaborations, cross-org model training

Mitigating Privacy Risks in AI/ML

Homomorphic Encryption

Encryption technique that allows computation on encrypted data without exposing original data

- Encrypt sensitive data → run computations on encrypted data → decrypt results
- Servers run approved computations on ciphertext; only the keyholder can decrypt the result.
- Main tradeoff is performance: usually much slower and more resource intensive

- We will cover Differential Privacy, Federated Learning, and Homomorphic Encryption in depth in Unit 8

Regulating Privacy Risks in AI/ML

The Privacy Regulation Challenge

- **GDPR**: right to explanation, right to erasure, purpose limitation – all challenged by AI
- **HIPAA**: health data used in AI training raises compliance questions
- **CCPA/CPRA**: California privacy rights and AI-specific provisions
- AI systems can infer protected information (health, religion, sexuality) from non-protected data
- Regulatory frameworks designed for databases struggle with probabilistic models

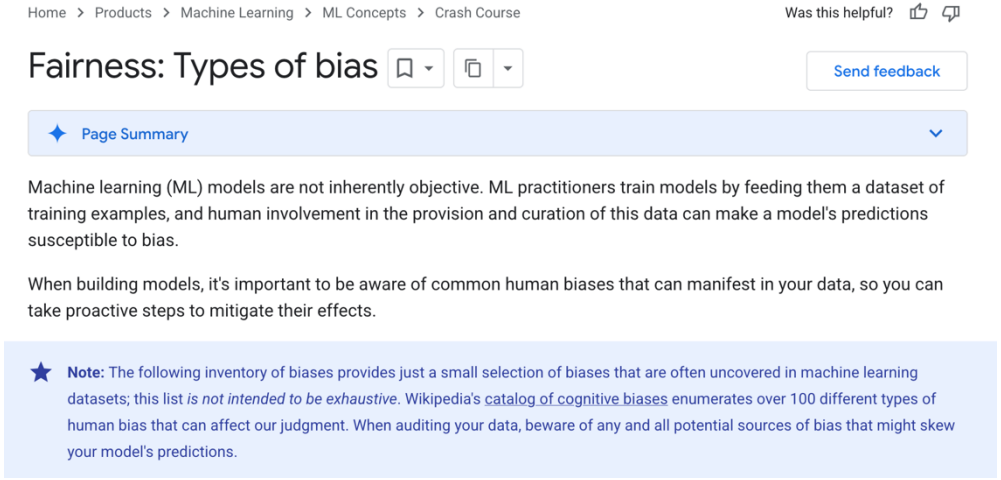
7.2: Bias in AI Systems

Types, sources, and fairness evaluation frameworks

Bias in AI/ML

Types of Bias

- **Statistical bias**: systematic error in estimation or prediction
- **Social bias**: reflecting or amplifying societal prejudices and stereotypes
- **Algorithmic bias**: systematic and repeatable errors in outcomes that create unfair treatment
- Not all bias is harmful (a spam filter is biased against spam -- that is the point)
- The concern: when bias creates unfair or discriminatory outcomes for protected groups



Google Developer – Fairness: Types of bias

Bias in AI/ML

Sources of Bias in the AI/ML lifecycle

- **Training data bias**: historical data reflects historical discrimination
- **Representation bias**: underrepresentation of certain groups in training data
- **Measurement bias**: features that proxy for protected attributes (zip code proxying for race)
- **Aggregation bias**: assuming one model fits all populations equally
- **Evaluation bias**: benchmarks that do not represent all affected populations
- **Deployment bias**: using a model in contexts different from its training distribution

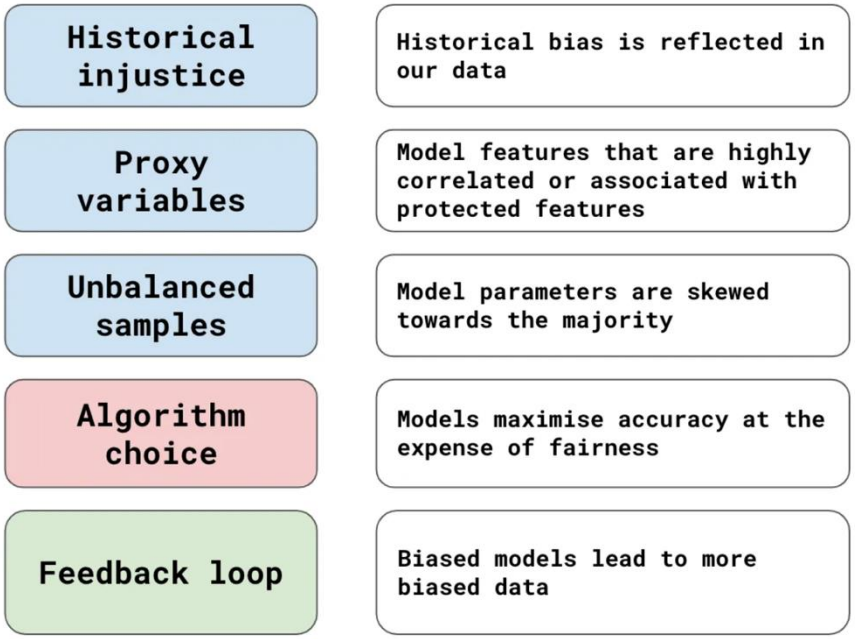


Figure 1: overview of sources of unfair predictions in machine learning (source: author)

Conor O'Sullivan – Medium (2023) – Unfair Predictions: 5 Common Sources of Bias in Machine Learning

Bias in AI/ML

Case Study: Facial Recognition Bias

Gender Shades (Buolamwini & Gebru, 2018)

- Audited commercial facial recognition from Microsoft, IBM, and Face++
- Error rates for lighter-skinned males: 0.8%
- Error rates for darker-skinned females: up to 34.7%
- Performance gap of over 40x based on skin tone and gender
- Led to IBM exiting the facial recognition market; Microsoft limiting law enforcement sales

Gender Shades

Gender Shades – MIT Media Lab - YouTube

Bias in AI/ML

Bias in Generative AI

- Image generators producing stereotypical representations (executives as white men, criminals as dark-skinned)
- Google Gemini image generation controversy (Feb 2024): overcorrection producing historically inaccurate images
- LLM bias: different quality of responses based on names, dialects, or cultural context
- Embedding bias: word embeddings encode societal stereotypes (man:programmer :: woman:homemaker)
- Generative AI amplifies bias at scale -- millions of outputs per day

Charles Sturt University