CYB-4203/6203

Secure and Trustworthy AI

Presentation 16: Transparency, Explainability, and Interpretability in AI/ML

Wednesday, April 1 2026

Today's Agenda

- 7.2: Brief recap of bias types and sources
- 7.3: Algorithmic transparency and accountability in AI/ML
- 7.4: Explainability and interpretability
    - Concepts, techniques, and tools (e.g., LIME, SHAP)

7.2: Bias in AI Systems

Brief recap of types and sources

Bias: Examples across the AI/ML Lifecycle

This Week

Week 10

Privacy, Bias, Transparency, and Explainability

Unit 7: Topics 7.1 - 7.4

- **7.1** Privacy risks in AI: membership inference, data leakage, surveillance, and predictive harm
- **7.2** Bias in AI systems: types, sources, and fairness evaluation frameworks
- **7.3** Algorithmic transparency and accountability in high-stakes decision-making
- **7.4** Explainability and interpretability: concepts, techniques, and tools

View Full Week 10 Materials →

Bias: Examples across the AI/ML Lifecycle

This Week

Week 10

Privacy, Bias, Transparency, and Explainability

Unit 7: Topics 7.1 - 7.4

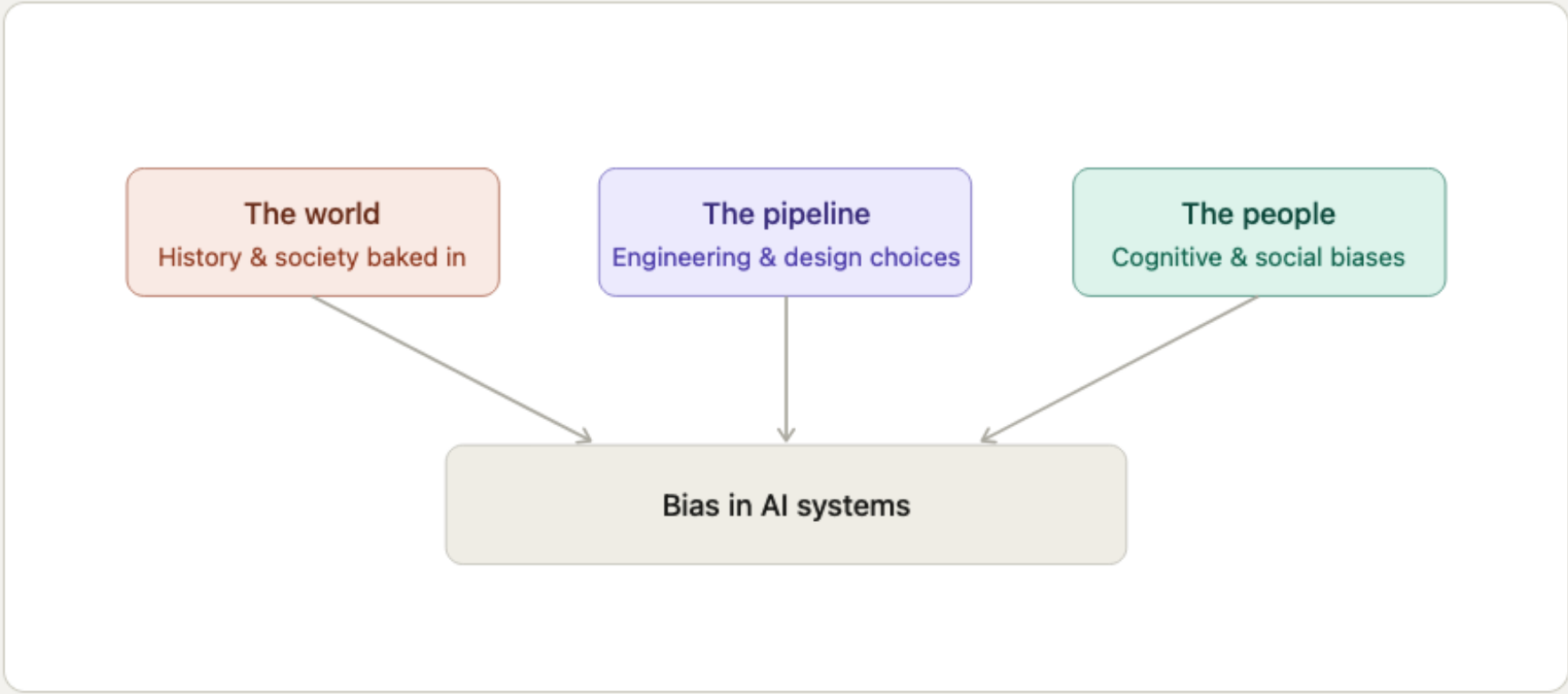
- **7.1** Privacy risks in AI: membership inference, data leakage, surveillance, and predictive harm
- **7.2** Bias in AI systems: types, sources, and fairness evaluation frameworks
- **7.3** Algorithmic transparency and accountability in high-stakes decision-making
- **7.4** Explainability and interpretability: concepts, techniques, and tools

View Full Week 10 Materials →

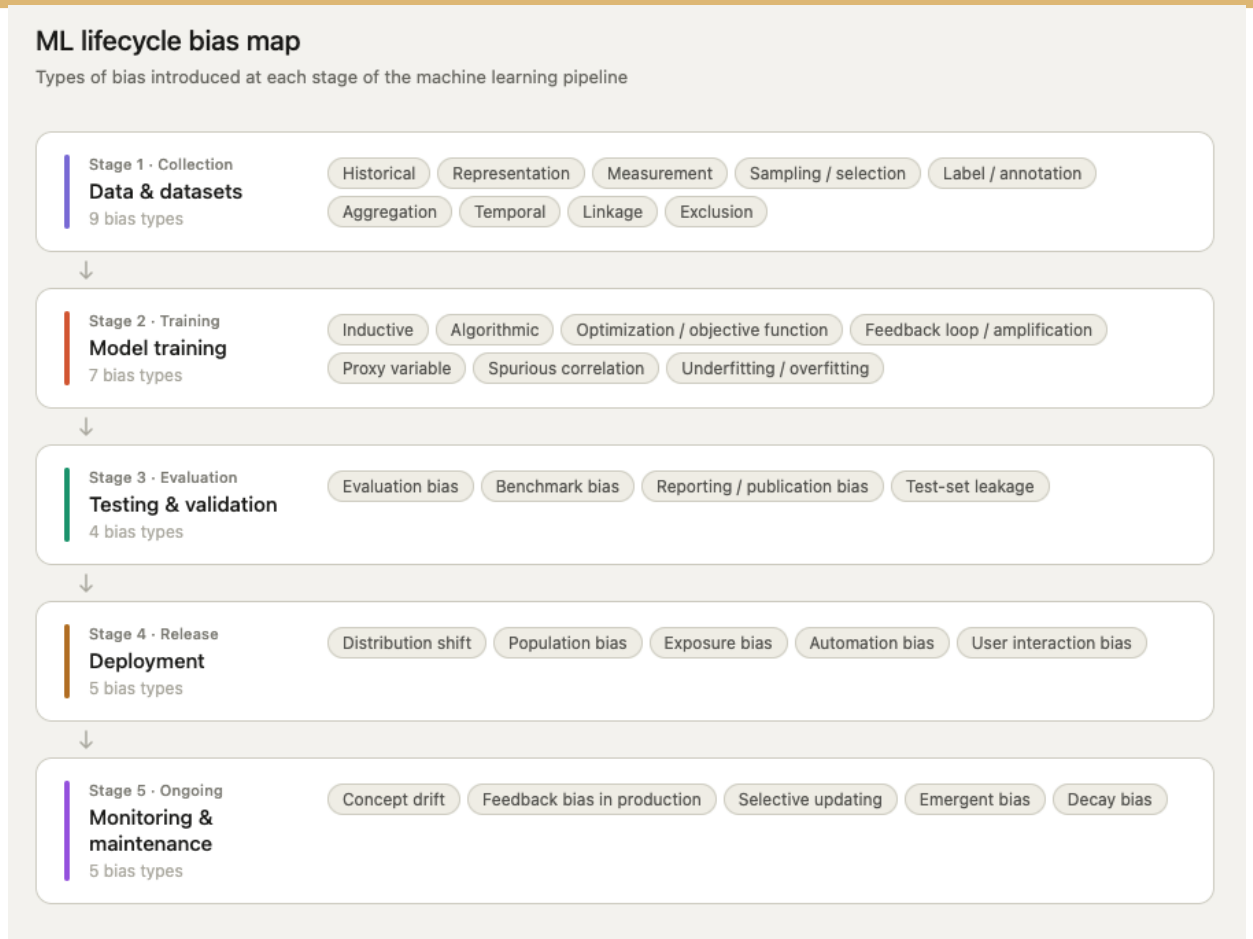
https://dallaselleman.github.io/cyb-4203-6203-spring-2026/weeks/week-10/index.html

Sources of AI Bias

Every form of bias in AI systems originates from at least one of these three roots — and often all three at once.



AI/ML Bias: Examples from The World and The Pipeline



AI/ML Bias: Examples from The People

Cognitive and social biases that practitioners, users, and organizations bring to AI systems

OVER-RELIANCE ON AI

Automation bias

Over-trusting model outputs; suspending critical judgment because "the AI said so." Especially dangerous in high-stakes settings like medicine and law.

Automation complacency

Extended uneventful reliance on an AI system gradually lowers human vigilance — until a critical failure occurs and no one is paying close enough attention.

PERCEPTION & EVALUATION ERRORS

Confirmation bias

Designers and evaluators notice evidence that confirms the model works and discount evidence that it doesn't. Shapes which test cases get written and which get ignored.

Anchoring bias

The first model or benchmark used becomes the de facto standard, even if it was arbitrary or flawed. Subsequent improvements are measured relative to a questionable baseline.

Recency bias

Practitioners over-weight recent data or the latest model architectures, potentially discarding valuable signal and proven techniques from earlier work.

Framing bias

How a problem is defined determines what gets measured and optimized. Choosing "loan default rate" vs. "loan repayment success" encodes fundamentally different values.

AI/ML Bias: Examples from The People

Cognitive and social biases that practitioners, users, and organizations bring to AI systems

STRUCTURAL & ORGANIZATIONAL BIASES

In-group bias

Teams with homogeneous backgrounds build systems that serve people like themselves better than others. Gaps in coverage feel invisible to those who don't experience them.

Expertise bias

Assuming that ML expertise implies expertise in the domain where the model is applied — medicine, law, finance. Technical competence and domain knowledge are separate things.

Ethical fading

Under deadline and competitive pressure, ethical concerns about bias gradually recede from focus during development. Not malicious — just deprioritized until it's too late.

Sycophancy / demand characteristics

Users phrase prompts or inputs in ways they think the model "wants," corrupting usage data and evaluations with socially desirable rather than genuine responses.

7.3 & 7.4:

Transparency, Explainability, Interpretability

Transparency, Explainability, & Interpretability