Week 12: AI Ethics & Fairness

From Bias Detection to Responsible AI

BSc Natural Language Processing

Discovery-Based Learning Approach

2025

The AI That Rejected All Women

Amazon's Hiring AI (2014-2018):

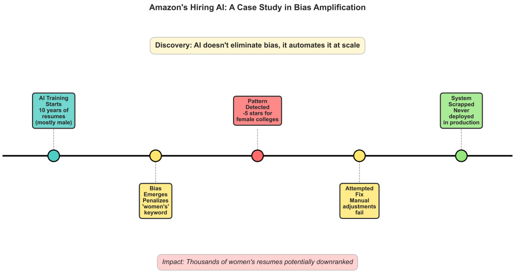
Training Data:

- 10 years of resumes
- Mostly male engineers (historical)
- Used to train ML ranking model

The Discovery:

- Resume mentions "women's chess club" → -5 stars
- Attended women's college → Downranked
- Any "women's" keyword → Penalty

Impact:
Thousands of women's resumes potentially rejected
System never deployed (discovered during testing)



The Insight:
"AI doesn't eliminate bias,
it automates it at scale"

Why It Happened:

- Model learned from biased history
- Optimized to match past hires
- Past hires were mostly men

Paradigm Shift: From "Objective Algorithms" to "Bias Amplifiers"

OLD Belief (2010):

"Algorithms are objective and fair"

Reasoning:

- Math has no prejudice
- Computers treat everyone equally
- Data-driven decisions are neutral
- Removes human bias from process

Example Claim:

- ML hiring: No gender/race considered
- Should be fairer than humans
- "Let the data speak"

Reality:
This assumption was wrong

NEW Understanding (2024):

"Algorithms amplify training data bias"

Reality:

- Models learn historical patterns
- Historical data reflects discrimination
- Optimization amplifies patterns
- Scale multiplies impact

Concrete Examples:

- Resume screening: Women downranked
- Loan approval: Racial disparities
- Medical diagnosis: Worse for minorities
- Facial recognition: Lower accuracy for Black faces

Solution:
Proactive bias detection & mitigation

Key Insight: Bias is not a bug to fix, it's a fundamental challenge requiring ongoing vigilance

Real-World Harms: Quantified Impact in 2024

Real-World AI Harms in 2024: Documented Cases with Quantified Impact

**Healthcare:
Skin Cancer
Detection**

93% accuracy
(white skin)

68% accuracy
(black skin)

**Criminal Justice:
COMPAS
Recidivism**

45% false positive
(Black defendants)

23% false positive
(White defendants)

**Language Models:
GPT-3 Gender
Stereotypes**

'doctor' → 'he' (67%)
'nurse' → 'she' (71%)

Improved in GPT-4

**Financial:
Credit
Scoring**

Women pay higher
interest rates

Controlling for risk

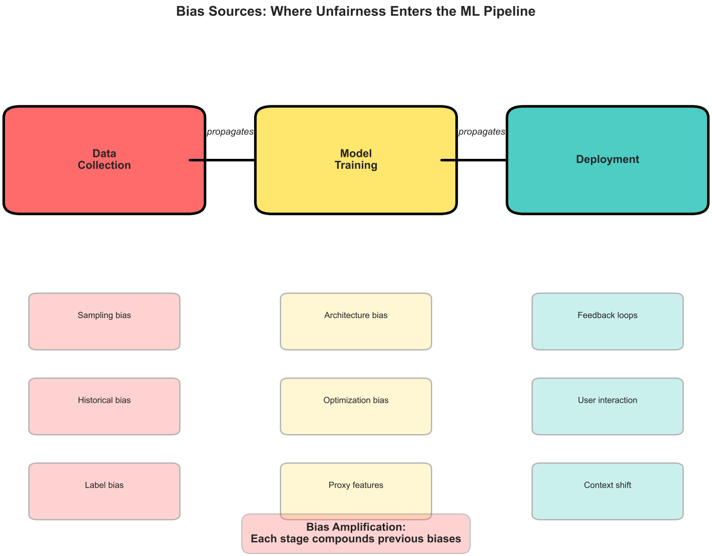
**Employment:
Amazon Hiring
AI**

-5 stars for
'women's' keyword

System scrapped

**Combined Impact: Millions affected by biased AI decisions in 2024
Across healthcare, justice, finance, employment, and language technology**

Foundation 1: Bias Sources (Visual)



Foundation 1: Bias Sources (Detailed)

1. Data Bias:

Sampling Bias

- Training data not representative
- Example: Medical AI trained on 80% white patients
- Impact: Lower accuracy for minorities

Historical Bias

- Data reflects past discrimination
- Example: Hiring data (mostly male engineers)
- Impact: Model learns to prefer men

Label Bias

- Human labelers have biases
- Example: Toxicity labels vary by annotator demographics
- Impact: Model inherits annotator biases

2. Model Bias:

Architecture Bias

- Model design favors certain patterns
- Example: CNNs for faces (tested on white faces)
- Impact: Worse for underrepresented groups

Optimization Bias

- Loss function optimized for majority
- Example: Accuracy maximized on dominant class
- Impact: Minority performance sacrificed

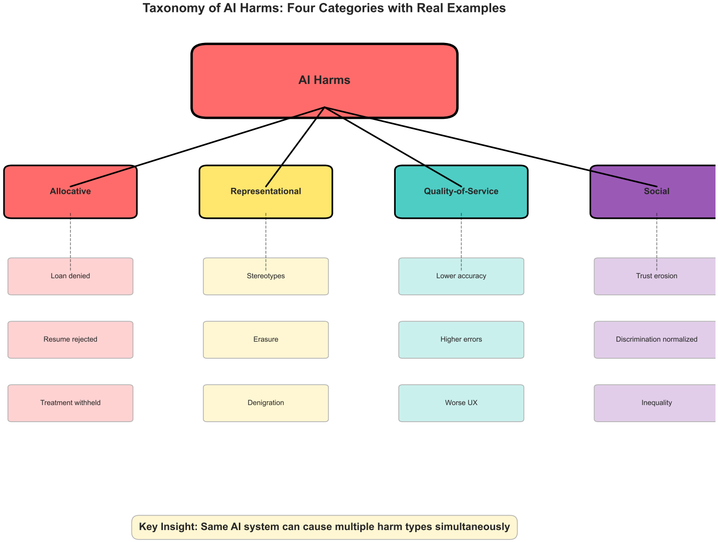
3. Deployment Bias:

Feedback Loops

- Model predictions influence future data
- Example: Biased recommendations → biased clicks → more bias
- Impact: Self-reinforcing discrimination

Comprehensive View: Bias is not one problem, it's a systemic challenge across the ML pipeline

Foundation 2: Harm Taxonomy (Visual)



Foundation 2: Harm Taxonomy (Detailed)

1. Allocative Harm:

Resources withheld or unfairly distributed

Examples:

- Loan denied due to biased credit score
- Resume rejected by biased hiring AI
- Medical treatment withheld (risk score)
- Insurance premium higher (demographic)

Impact: Direct material loss (money, opportunity)

2. Representational Harm:

Stereotypes reinforced or groups erased

Examples:

- Image search: “CEO” shows only men
- Translation: “The doctor” → “he”
- Face recognition: Fails on minorities
- Voice assistants: Only understand native speakers

Impact: Dignity, identity, social standing

3. Quality-of-Service Harm:

Unequal performance across demographics

Examples:

- Skin cancer detection: 93% (white), 68% (Black)
- Speech recognition: Higher error for accents
- Face unlock: Fails more for women, minorities
- Medical AI: Trained on majority population

Impact: Frustration, exclusion, worse outcomes

4. Social Harm:

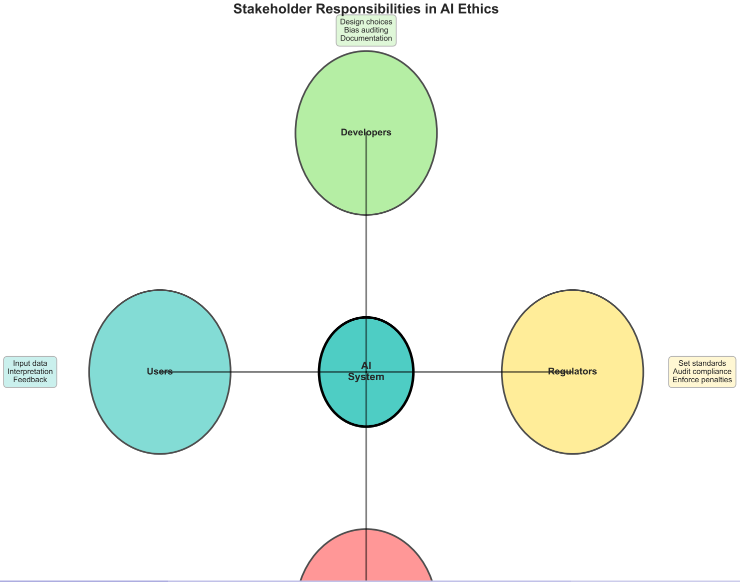
Erosion of trust and normalization of discrimination

Examples:

- COMPAS: Discrimination in sentencing
- People avoid AI systems (distrust)
- “If AI says it, it must be true” (authority)
- Inequality becomes automated, invisible

Impact: Societal trust, democratic participation

Harm Framework: Understanding different harm types enables targeted interventions


Stakeholder Responsibilities in AI Ethics
Design choices
Bias auditing
Documentation
Developers
Input data
Interpretation
Feedback
Users
AI
System
Regulators
Set standards
Audit compliance
Enforce penalties

Foundation 3: Stakeholders (Detailed Responsibilities)

Developers:

Responsibilities:

- Design with fairness in mind
- Audit for bias pre-deployment
- Document limitations transparently
- Provide recourse mechanisms

Tools:

- Fairness metrics (AIF360, Fairlearn)
- Bias detection tools
- Model cards (documentation)

Users:

Responsibilities:

- Understand system limitations
- Interpret outputs critically
- Report observed bias
- Participate in feedback

Affected Communities:

Responsibilities:

- Share lived experiences of harm
- Provide context developers miss
- Demand accountability
- Advocate for rights

Reality:

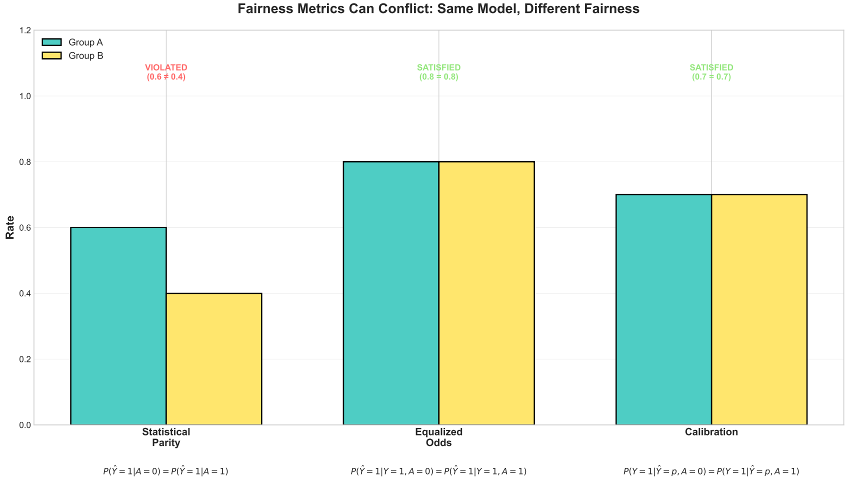
- Often excluded from design
- Harm discovered after deployment
- Limited recourse when harmed

Regulators:

Responsibilities:

- Set fairness standards
- Audit compliance
- Enforce penalties for violations
- Update laws as tech evolves

Method 1: Statistical Parity (Visual)



Core Idea: Equal positive prediction rates across groups

Formula: $P(\hat{Y}=1|A=0)=P(\hat{Y}=1|A=1)$

Statistical Parity: Demographic parity - same proportion of each group receives positive outcome

Method 1: Statistical Parity (Detailed)

Definition:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

where:

- $\hat{Y}$: Model prediction
- A: Protected attribute (race, gender, etc.)
- $A = 0$: Majority group
- $A = 1$: Minority group

Interpretation:

- Same approval rate for both groups
- Example: If 40% of men get loans, 40% of women should too
- Independent of actual qualifications

Numerical Example:

- 1000 male applicants, 400 approved (40%)
- 1000 female applicants, 400 approved (40%)

When to Use:

- Hiring (equal opportunity)
- College admissions
- Loan approvals
- When group parity is legal requirement

Advantages:

- Easy to understand
- Easy to measure
- Legal precedent in some domains
- Prevents overt discrimination

Disadvantages:

- Ignores base rates (true qualifications)
- May require different thresholds per group
- Can conflict with accuracy
- Not always legally defensible

Limitation:

Method 2: Equalized Odds (Visual)

The Idea:

Equal accuracy across groups

Two Conditions:

1. Equal True Positive Rate:

$$P(\hat{Y} = 1 | Y = 1, A = 0) =$$

$$P(\hat{Y} = 1 | Y = 1, A = 1)$$

2. Equal False Positive Rate:

$$P(\hat{Y} = 1 | Y = 0, A = 0) =$$

$$P(\hat{Y} = 1 | Y = 0, A = 1)$$

Concrete Example:

COMPAS Recidivism (Actual):

- TPR (Black): 60%
- TPR (White): 60%
- FPR (Black): 45%
- FPR (White): 23%

Violation: FPR differs (45% $\neq$ 23%)

Impact: Black defendants mislabeled as high-risk at 2$\times$ rate

Equalized Odds: Ensures model is equally accurate for both qualified and unqualified in each group

Method 2: Equalized Odds (Detailed Analysis)

Why It Matters:

True Positive Rate (TPR):

- Among qualified, fraction correctly identified
- Example: Among people who won't reoffend, how many correctly labeled low-risk?
- Equal TPR: Both groups benefit equally

False Positive Rate (FPR):

- Among unqualified, fraction incorrectly identified
- Example: Among people who will reoffend, how many mislabeled low-risk?
- Equal FPR: Both groups harmed equally by errors

COMPAS Example (2016):

- 10,000 defendants analyzed
- FPR Black: 45% (450 false positives)
- FPR White: 23% (230 false positives)
- Result: Black defendants falsely labeled high risk 2×

When to Use:

- Criminal justice (recidivism, bail)
- Medical diagnosis
- Credit scoring
- Any high-stakes decision

Advantages:

- Respects merit (true qualifications)
- Ensures equal error rates
- Legally defensible (equal treatment)
- Widely accepted fairness criterion

Disadvantages:

- Conflicts with statistical parity
- May not satisfy calibration
- Requires ground truth labels
- Can be difficult to achieve

Best Practice:

Method 3: Counterfactual Fairness (Visual)

The Question:

"Would the prediction change if we changed only the protected attribute?"

Example:

Resume:

- Name: James Smith
- Education: MIT Computer Science
- Experience: 5 years at Google
- **Prediction: 0.85 (hire)**

Counterfactual Resume:

- Name: Jennifer Smith (ONLY change)
- Education: MIT Computer Science (same)
- Experience: 5 years at Google (same)
- **Prediction: 0.80 (hire)**

Evaluation:

If predictions differ ($0.85 \neq 0.80$):

- Model is using gender
- Counterfactual fairness VIOLATED
- Direct discrimination

If predictions same ($0.85 = 0.85$):

- Gender does not affect score
- Counterfactual fairness SATISFIED
- No direct discrimination

Causal Fairness:
Only causally relevant factors should affect predictions

Protected attributes should have ZERO causal effect

Counterfactual Fairness: Causal framework - protected attribute should not cause prediction to change

Method 3: Counterfactual Fairness (Detailed Implementation)

Formal Definition:

$$P(\hat{Y}_{A \leftarrow a} | X = x, A = a) =$$
$$P(\hat{Y}_{A \leftarrow a'} | X = x, A = a)$$

Translation: Prediction for individual with protected attribute a equals prediction if they had attribute a', holding all else constant

Implementation:

Step 1: Build causal graph

- Identify causal relationships
- Separate legitimate vs illegitimate paths

Step 2: Block illegitimate paths

- Remove direct effect of A on $\hat{Y}$
- Remove indirect effect through mediators

Step 3: Test counterfactuals

- Generate counterfactual examples

Challenges:

1. **Proxy Features:**
   - ZIP code correlated with race
   - First name correlated with gender
   - Must remove ALL correlated features
   - May lose predictive power
2. **Causal Graph:**
   - Requires domain knowledge
   - Hard to validate
   - May be controversial
3. **Legitimate Pathways:**
   - Some gender effects may be legitimate
   - Example: Women's health outcomes
   - Need to distinguish discrimination from valid correlation

When to Use:
When you can specify causal relationships

Mitigation 1: Data Augmentation (Visual)

The Problem:

Imbalanced training data

Example:

- 10,000 male resumes (hired)
- 1,000 female resumes (hired)
- Ratio: 10:1

Consequence:

- Model learns "male = good candidate"
- Under-represents female patterns
- Worse performance on women

The Solution:

Augment minority class

Methods:

- Oversample minority: Duplicate female resumes
- Undersample majority: Remove male resumes
- SMOTE: Generate synthetic female resumes

Result:

- 5,000 male resumes
- 5,000 female resumes
- Ratio: 1:1
- Model sees balanced examples

Data Augmentation: Fix the data, fix the bias - balance training distribution

Mitigation 1: Data Augmentation (Detailed Techniques)

1. Oversampling:

Method: Duplicate minority samples

Advantages:

- Simple to implement
- No data loss
- Balances classes

Disadvantages:

- Exact duplicates (overfitting)
- Larger dataset (slower training)

2. Undersampling:

Method: Remove majority samples

Advantages:

- Smaller dataset (faster)
- Balances classes

Disadvantages:

- Loses information

3. SMOTE (Synthetic):

Method: Generate synthetic minority samples

Algorithm:

1. Find k nearest neighbors (minority)
2. Interpolate between neighbors
3. Create new synthetic sample
4. Repeat until balanced

Example:

- Resume A: (skills=[Python, Java], exp=5)
- Resume B: (skills=[Python, C++], exp=7)
- Synthetic: (skills=[Python, Java, C++], exp=6)

Advantages:

- No exact duplicates
- Expands decision boundary
- Better generalization

Disadvantages:

Mitigation 2: Adversarial Debiasing (Visual)

The Setup:

Two Models:

1. Classifier (C):

- Task: Predict hired/not hired
- Input: Resume features
- Goal: Maximize accuracy

2. Adversary (A):

- Task: Predict gender from C's hidden layer
- Input: C's internal representation
- Goal: Maximize gender prediction

Training:

- C tries to fool A (remove gender signal)
- A tries to detect gender (maximize accuracy)
- Minimax game: C vs A

The Outcome:

If A succeeds (predicts gender well):

- C's representation contains gender info
- C is biased
- Update C to remove gender signal

If A fails (random guessing):

- C's representation is gender-neutral
- C cannot be biased (no gender info)
- Training complete

Key Idea:
If adversary cannot predict protected attribute from internal representation, model cannot use it for predictions

Adversarial Debiasing: Game-theoretic approach - remove bias signal from learned representations

Mitigation 2: Adversarial Debiasing (Detailed Mathematics)

Objective Function:

$$\min_{\theta_C} \max_{\theta_A} \mathcal{L}_C - \lambda \mathcal{L}_A$$

where:

- $\mathcal{L}_C$: Classifier loss (accuracy)
- $\mathcal{L}_A$: Adversary loss (gender prediction)
- λ: Trade-off parameter

Training Algorithm:

1. **Step 1:** Update adversary
   - Fix classifier weights
   - Train adversary to predict gender
   - Maximize $\mathcal{L}_A$
2. **Step 2:** Update classifier
   - Fix adversary weights
   - Train classifier to fool adversary
   - Minimize $\mathcal{L}_C - \lambda \mathcal{L}_A$
3. **Step 3:** Repeat until convergence

Trade-off Parameter λ:

- $\lambda = 0$: No debiasing (ignore adversary)
- $\lambda =$ small: Weak debiasing
- $\lambda =$ large: Strong debiasing (may hurt accuracy)

Typical Results:

- $\lambda = 0.0$: Accuracy 85%, Gender pred 95% (biased)
- $\lambda = 1.0$: Accuracy 83%, Gender pred 55% (fair)
- $\lambda = 10$: Accuracy 78%, Gender pred 51% (random)

When to Use:

- Deep learning models
- When you can't remove protected attribute from data
- When you want representation-level fairness

Best Practice:
Tune λ with validation set
Balance accuracy vs fairness

Adversarial Debiasing: In-training mitigation - removes bias during model learning

Mitigation 3: Calibration & Post-processing (Visual)

The Problem:

Model outputs not calibrated across groups

Example:

- Model says "70% chance hired"
- For men: 70% actually hired (calibrated)
- For women: 50% actually hired (not calibrated)

Impact:

- Scores mean different things per group
- Misleading confidence estimates
- Unfair decision thresholds

The Solution:

Post-process outputs to equalize calibration

Method:

1. Train model (biased outputs)
2. Compute calibration per group
3. Adjust thresholds to equalize
4. Apply different threshold per group

Result:

- Men: Threshold = 0.5 for hiring
- Women: Threshold = 0.4 for hiring (compensate)
- Same true positive rate for both

Calibration: Post-processing approach - adjust outputs after training to ensure fairness

Mitigation 3: Calibration & Post-processing (Detailed Techniques)

Calibration Curve:

$$P(Y = 1|\hat{Y} = p, A = a) = p$$

Meaning:

- If model says p=0.7, 70% should be positive
- Must hold for EACH group separately
- Calibration $\neq$ accuracy

Platt Scaling:

Method: Learn per-group transformation

$$\hat{p}_{\text{calibrated}} = \sigma(w \cdot \hat{p} + b)$$

where σ is sigmoid, w and b learned per group

Algorithm:

1. Split validation data by group
2. Fit logistic regression per group

Threshold Optimization:

Method: Find different thresholds per group

Algorithm:

1. Set fairness constraint (e.g., equal TPR)
2. Search for thresholds that satisfy constraint
3. Apply group-specific thresholds

Example:

- Group A: Threshold 0.5 $\rightarrow$ TPR 80%
- Group B: Threshold 0.4 $\rightarrow$ TPR 80%
- Result: Equal TPR achieved

Advantages:

- Model-agnostic (works with any classifier)
- No retraining needed
- Mathematically guarantees fairness metric

Disadvantages:

- Requires labeled validation data

Safety: Red Teaming & Constitutional AI (Visual)

Red Teaming:

Adversarial testing for harmful outputs

Process:

1. Hire diverse red team
2. Attempt to elicit harmful outputs
3. Document failure modes
4. Fix vulnerabilities
5. Repeat

Example Attacks:

- Jailbreak prompts: "Ignore previous instructions"
- Indirect requests: "Write a story about..."
- Multi-turn manipulation
- Role-playing scenarios

Coverage:

- Toxicity, bias, misinformation
- Privacy leaks
- Instruction following failures

Constitutional AI:

AI trained to follow ethical principles

The Constitution:

1. Be helpful, harmless, honest
2. Refuse harmful requests
3. Explain refusals politely
4. No discrimination
5. Respect privacy
6. Cite sources

Training:

- Generate responses
- Critique against principles
- Revise to satisfy principles
- RLHF with constitutional feedback

Result:

- GPT-4: Refuses harmful requests
- Explains reasoning transparently

Challenge: Gender Bias in Word Embeddings

The Discovery (2016):

Word2Vec embeddings contain gender stereotypes

Evidence:

Word analogy task:

- man : computer programmer :: woman : **homemaker**
- man : doctor :: woman : **nurse**
- man : brilliant :: woman : **lovely**

Quantification:

- "doctor" - "man" + "woman" $\approx$ "nurse"
- Cosine similarity: 0.72
- Should be: "doctor" (gender-neutral)

Root Cause:

Training Data:

- Google News corpus (3B words)
- Reflects historical gender roles
- "doctor" appears more with "he"
- "nurse" appears more with "she"

Consequence:

- Embeddings used in downstream tasks
- Resume ranking, translation, search
- Bias amplified in applications
- Millions affected

Scale:
Every model using Word2Vec
inherits this bias
(billions of predictions affected)

Word Embedding Bias: Foundational bias - affects all models built on these embeddings

Key Takeaways: Five Principles of Responsible AI

Responsible AI Fundamentals

1. **Bias is Systemic, Not a Bug**
   Enters at data, model, and deployment stages - requires ongoing vigilance
   Example: Amazon AI rejected women despite no gender feature
2. **Fairness Metrics Conflict**
   Statistical parity $\neq$ equalized odds $\neq$ calibration
   Choose metric based on application context and legal requirements
3. **Detection Before Mitigation**
   Measure bias first, then apply targeted intervention
   Use appropriate metrics: WEAT for embeddings, demographic parity for outcomes
4. **Stakeholder Participation**
   Include affected communities in design and evaluation
   Developers alone cannot anticipate all harms
5. **Transparency and Accountability**
   Document limitations, provide recourse, enable auditing
   Model cards, datasheets, fairness reports

Summary: Responsible AI requires technical rigor, ethical awareness, and stakeholder engagement

Course Conclusion: From Theory to Practice

12-Week Journey:

Weeks 1-3: Foundations

- N-grams → Neural LMs → RNNs
- Word embeddings, language modeling

Weeks 4-5: Architectures

- Seq2seq, attention, transformers
- The attention revolution

Weeks 6-8: Modern NLP

- BERT, GPT, pre-training
- Tokenization, scaling

Weeks 9-11: Deployment

- Decoding, fine-tuning, compression
- Making AI practical

Week 12: Ethics

- Bias, fairness, responsibility
- Making AI safe

Real-World Impact:

You now understand:

- How language models work (theory)
- How to build them (practice)
- How to deploy them (engineering)
- How to do so responsibly (ethics)

Next Steps:

- Build your own models
- Contribute to open source
- Research novel architectures
- Advocate for responsible AI

The Future:
AI will transform society
You have the knowledge to ensure
that transformation is beneficial

Thank you!

Course Complete: 12 weeks from N-grams to responsible deployment - you are now equipped to build, deploy, and ensure fairness in NLP systems