# Responsible AI & Ethical Innovation

## Building Fair and Accountable ML Systems

Week 7: Machine Learning for Smarter Innovation

# Today's Journey

**Part 1: Foundation**
- Why ethics matters
- Real-world failures
- Ethical frameworks
- Stakeholder analysis

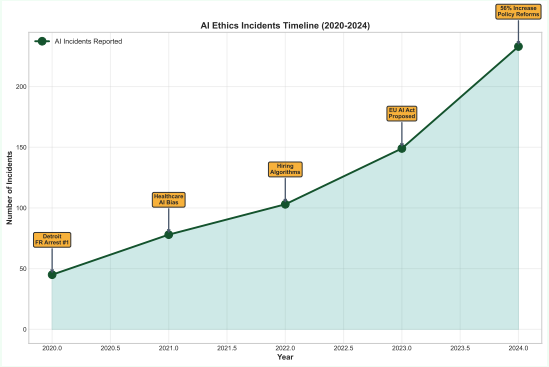
**Part 2: Algorithms**
- Bias in data
- Fairness metrics
- Detection methods
- Mitigation techniques

**Part 3: Implementation**
- Fairness toolkits
- Explainability tools
- Model documentation
- Privacy-preserving ML

**Parts 4-5: Design & Practice**
- Inclusive design
- Case studies
- Ethical audit framework
- Regulatory compliance

Ethics isn't optional – it's essential for sustainable innovation

# The Growing Ethics Crisis in AI




## The Numbers

**2024 Statistics:**

- 233 AI incidents reported
- 56% increase from 2023
- \$10B+ in settlements
- 47 countries affected

**Impact Areas:**

- Healthcare: 34%
- Finance: 28%
- Law enforcement: 22%
- Employment: 16%

Source: AI Incident Database, 2024

# When AI Goes Wrong: 2024 Incidents

**Facial Recognition Bias**

Detroit Settlement (2024)

- Black man wrongfully arrested
- False facial recognition match
- Police now banned from arrests based solely on FR

UK Facewatch Case (May 2024)

- Woman wrongly ID'd as shoplifter
- Banned from all stores in network
- System failed on non-white individual

**Employment Discrimination**

Uber Eats (2024)

- Driver dismissed by FR system
- Technology failed on darker skin
- No human review process

Resume Screening

- AI tools used for hiring decisions
- Women & minorities disadvantaged
- Most managers untrained in fair use

These aren't edge cases – they're systemic failures

# Three Reasons Ethics is Critical

**1. Human Impact**
Real people suffer:

- Lost opportunities
- Legal consequences
- Psychological harm
- Social stigma
- Financial loss

**Scale:**
Millions affected daily by automated decisions

**2. Legal Risk**
Regulatory landscape:

- EU AI Act (2024)
- GDPR fines: up to 4%
- US state laws
- Class action lawsuits
- Criminal liability

**Cost:**
$10B+ in settlements in 2024

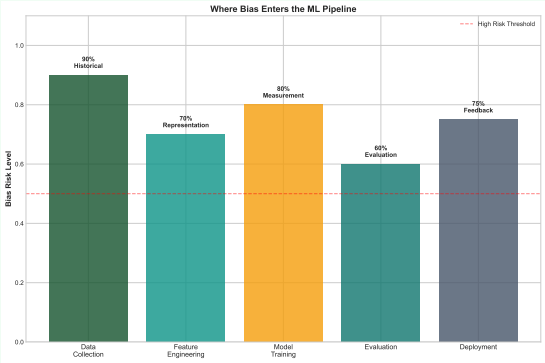
**3. Business Value**
Ethical AI drives:

- Customer trust
- Brand reputation
- Market access
- Talent attraction
- Innovation

**ROI:**
73% consumers prefer ethical brands

Ethics is not a constraint – it's a competitive advantage

# Understanding Bias: The ML Pipeline




## Bias Entry Points

1. **Data Collection**
   - Historical discrimination
   - Sampling bias
   - Missing populations
2. **Feature Engineering**
   - Proxy variables
   - Correlation artifacts
   - Human assumptions
3. **Model Training**
   - Optimization bias
   - Spurious correlations
   - Overfitting to majority

Bias can enter at any stage – vigilance is required throughout

# Ethical Frameworks for AI

**Consequentialist**
Focus on outcomes:

- Maximize benefit
- Minimize harm
- Utilitarian calculus
- Risk-benefit analysis

**Example:**
Deploy system if overall benefit outweighs harm to affected groups

**Deontological**
Focus on duties:

- Respect autonomy
- Honor rights
- Follow rules
- Categorical imperatives

**Virtue Ethics**
Focus on character:

- Cultivate wisdom
- Practice fairness
- Show compassion
- Demonstrate integrity

**Example:**
Ask "What would a fair person do?"

**Care Ethics**
Focus on relationships:

- Understand context
- Value relationships
- Address vulnerability
- Emphasize empathy

No single framework is perfect – combine perspectives for robust ethics

# Core Principles for Responsible AI

**Technical Principles**

Fairness

- Equal treatment across groups
- Measure disparate impact
- Mitigate bias systematically

Robustness

- Reliable performance
- Graceful degradation
- Security against attacks

Transparency

- Explainable decisions
- Documented processes
- Auditable systems

**Social Principles**

Accountability

- Clear responsibility
- Redress mechanisms
- Governance structures

Privacy

- Data minimization
- User control
- Confidentiality

Beneficence

- Human wellbeing first
- Social benefit
- Environmental sustainability

These principles form the foundation of responsible AI development

# Identifying Stakeholders

**Direct Stakeholders**

Users

- Those who interact with system
- Experience direct consequences
- May lack agency or choice

Developers

- Design and build system
- Make technical choices
- Bear professional responsibility

Deployers

- Organizations using AI
- Control deployment context
- Manage operational decisions

**Indirect Stakeholders**

Affected Communities

- Impacted without direct use
- May face systemic effects
- Often marginalized groups

Society

- Broader social impacts
- Norm shifts
- Democratic implications

Environment

- Carbon footprint
- Resource consumption
- E-waste generation

Map all stakeholders before deployment – especially those without voice

# Recognizing Power Imbalances

**Who Has Power?**

Tech Companies

- Control system design
- Set defaults & constraints
- Influence policy
- Access to resources

Governments

- Regulatory authority
- Procurement decisions
- Surveillance capabilities
- Enforcement power

Privileged Groups

- Represented in data
- Cultural norms embedded
- Economic resources
- Political influence

**Who Lacks Power?**

End Users

- Limited choice
- Information asymmetry
- No opt-out options
- Captive audiences

Marginalized Groups

- Underrepresented in data
- Higher error rates
- Less recourse
- Compounded discrimination

Future Generations

- No voice in decisions
- Inherit consequences
- Path dependencies
- Environmental debt

Responsible AI requires actively empowering the powerless

# What Happens Without Ethics

**Individual Harms**

Documented Cases:

- Wrongful arrests: 12+ cases
- Denied healthcare: 1000s
- Job discrimination: widespread
- Credit denial: systemic
- Deportation errors: 200+

Each case represents:

- Personal trauma
- Career damage
- Family disruption
- Loss of trust

**Systemic Consequences**

Societal Level:

- Amplified inequality
- Eroded civil rights
- Reduced opportunity
- Surveillance normalization
- Democratic backsliding

Industry Level:

- Regulatory backlash
- Market restrictions
- Talent exodus
- Innovation chilling
- Public distrust

The cost of fixing problems later far exceeds prevention

# Foundation Summary: Ethics First

**Key Takeaways**

1. Ethics failures are increasing (56% YoY)
2. Real people face real harms
3. Multiple ethical frameworks exist
4. All stakeholders matter
5. Power imbalances are structural

**Critical Insight:**
Ethics isn't about constraining innovation – it's about ensuring sustainable, legitimate, trustworthy progress.

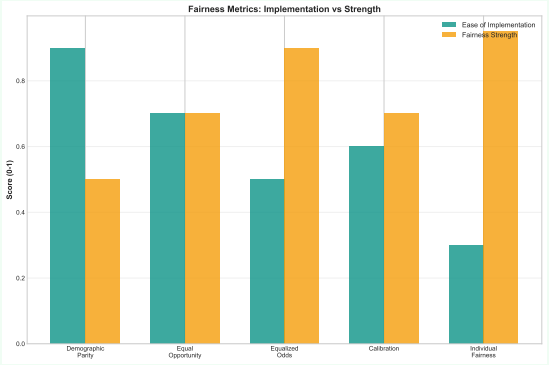
**Next: Algorithms**
Moving from "why" to "how":

- Fairness metrics
- Bias detection
- Mitigation techniques
- Mathematical foundations

**Mindset Shift:**
From "Can we build it?" to "Should we build it, and how do we build it responsibly?"

Ethics is Engineering

# Fairness Metrics Landscape




## Three Main Approaches

**Group Fairness**

- Compare outcomes across groups
- Statistical parity
- Most common in practice

**Individual Fairness**

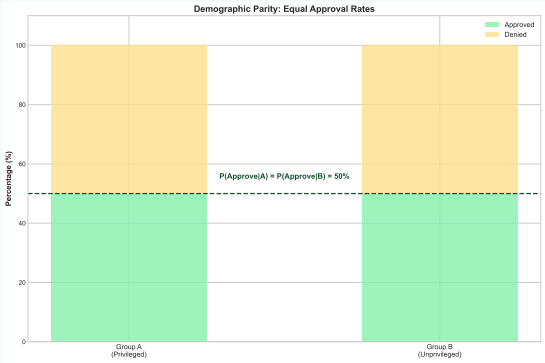
- Similar individuals treated similarly
- Harder to define
- Context-dependent

**Causal Fairness**

- Counterfactual reasoning
- Identify discrimination
- Most rigorous

Different metrics capture different notions of fairness

# Demographic Parity (Statistical Parity)




## Mathematical Definition

For protected attribute $A$ and decision $D$:

$$P(D = 1|A = a) = P(D = 1|A = b)$$

**Intuition:**
Positive outcomes should be independent of group membership

**Example:**

- 50% of Group A approved
- 50% of Group B approved
- Same rate regardless of merit

Simple to measure but ignores differing base rates

# When Demographic Parity Fails

**The Problem**
Consider loan approval:

- Group A: 80% good credit
- Group B: 40% good credit

Demographic Parity Forces:

- Approve 60% from each group
- Underpredict Group A
- Overpredict Group B
- Ignores actual creditworthiness

**Result:**

- Higher default rates in Group B
- Missed opportunities in Group A
- Economic inefficiency

No metric is universally correct – context matters

**When to Use It**
Appropriate when:

- Base rates should be equal
- Differences reflect discrimination
- Goal is representation
- Historical bias correction

Examples:

- University admissions (diversity)
- Jury selection
- Political representation
- Media visibility

**Key Insight:**
Demographic parity prioritizes equal outcomes over equal error rates

# Equal Opportunity (Equality of Opportunity)


## Mathematical Definition

For true label $Y = 1$ (qualified):

$$P(D = 1|Y = 1, A = a) = P(D = 1|Y = 1, A = b)$$

**Intuition:**
Qualified individuals have equal chances regardless of group

**Focus:**

- True Positive Rate (TPR)
- Recall parity
- Benefit the deserving

Allows different base rates while ensuring equal treatment of qualified

# Equalized Odds

**Mathematical Definition**
For both $Y = 1$ and $Y = 0$:

$$P(D = 1|Y = y, A = a) = P(D = 1|Y = y, A = b)$$

**Requires:**

- Equal TPR across groups
- Equal FPR across groups
- Stronger than equal opportunity

$$\text{TPR}_a = \text{TPR}_b$$
$$\text{FPR}_a = \text{FPR}_b$$

Equalized odds balances both types of errors

**Intuition**
Predictions independent of group:

- For qualified individuals
- For unqualified individuals
- Both errors equalized

**Trade-off:**

- More constrained
- Harder to achieve
- Better group fairness
- May reduce accuracy

**Example:**
Criminal recidivism: Equal error rates for defendants of different races

# Fairness Trade-offs: Impossibility Results

**The Bad News**
You cannot simultaneously achieve:

1. Demographic Parity
2. Equal Opportunity
3. Calibration

Unless:

- Perfect prediction (impossible)
- OR base rates are equal
- OR you sacrifice accuracy

**Implication:**
Fairness is not a purely technical problem – it requires value judgments

**Choosing Your Metric**
Questions to ask:

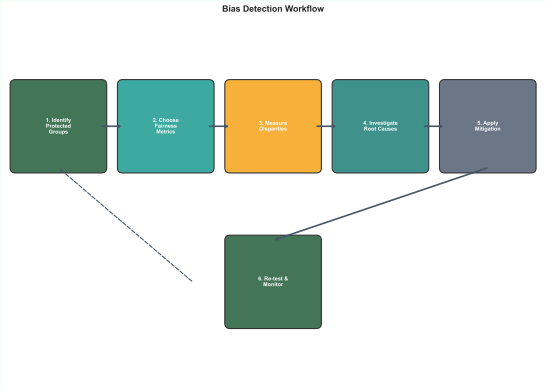
- What harm are we preventing?
- Who bears the cost of errors?
- What are the base rates?
- Is historical bias present?
- What do stakeholders prefer?

**Example: Lending**

- Banks prefer calibration
- Regulators prefer demographic parity
- Borrowers prefer equal opportunity
- Courts require justification

Math can't resolve value conflicts – transparency about choices is essential

# Detecting Bias in Your Model




## 5-Step Process

1. **Identify Groups**
   - Protected attributes
   - Intersectional identities
2. **Choose Metrics**
   - Context-appropriate
   - Multiple perspectives
3. **Measure Disparities**
   - Statistical tests
   - Confidence intervals
4. **Investigate Sources**
   - Data? Model? Deployment?
5. **Iterate**

Bias detection should be continuous, not one-time

# Taxonomy of Bias

**Data Bias**

Historical Bias
- Past discrimination in data
- Example: Hiring data reflects sexism

Representation Bias
- Missing or undersampled groups
- Example: Face datasets lack diversity

Measurement Bias
- Proxy labels differ by group
- Example: Arrest rates vs crime rates

Aggregation Bias
- One model for all groups
- Example: Medical tests designed for men

**Algorithmic Bias**

Evaluation Bias
- Metrics favor majority
- Example: Accuracy vs F1 score

Deployment Bias
- Different usage patterns
- Example: Over-policing minority areas

Feedback Loops
- Predictions influence future data
- Example: Predictive policing

Interaction Bias
- User behavior differs by group
- Example: Voice assistants & accents

Multiple biases often compound – address systematically

# Bias Mitigation: Pre-processing

**Data Transformations**

Reweighting

- Adjust sample weights
- Balance group representation
- Preserve individual data

Resampling

- Oversample minorities
- Undersample majorities
- SMOTE for synthetic data

Relabeling

- Fix label bias
- Correct historical discrimination
- Requires domain knowledge

**Feature Transformations**

Disparate Impact Remover

- Modify features
- Preserve utility
- Reduce correlation with protected attribute

Fair Representation Learning

- Learn fair latent space
- Information theory approach
- Encode only task-relevant info

**Pros & Cons**

Pros: Model-agnostic

Cons: May lose information

Pre-processing is transparent but may over-correct

# Bias Mitigation: In-processing

**Constrained Optimization**
Add fairness constraints:

$$\min_{\theta} L(\theta) \text{ s.t. } F(\theta) \leq \epsilon$$

Where:

- $L$: Loss function
- $F$: Fairness violation
- $\epsilon$: Tolerance

Examples:

- Fairness-aware SVM
- Constrained deep learning
- Lagrangian formulations

**Adversarial Debiasing**
Train two models:

1. Predictor $P$: Predict label
2. Adversary $A$: Predict group from $P$'s predictions

$$\min_{P} \max_{A} L_P - \lambda L_A$$

**Intuition:**
If adversary can't infer group, predictions are fair

**Trade-off:**
$\lambda$ balances accuracy vs fairness

In-processing offers fine-grained control but requires model modification

# Bias Mitigation: Post-processing

**Threshold Optimization**
Adjust decision thresholds per group:
Group A: $D_a = 1$ if $\hat{y}_a > \tau_a$
Group B: $D_b = 1$ if $\hat{y}_b > \tau_b$

Find $\tau_a, \tau_b$ to satisfy:

- Demographic parity: Equal accept rates
- Equal opportunity: Equal TPR
- Equalized odds: Equal TPR & FPR

**Pros:**

- Model-agnostic
- Easy to implement
- Reversible

**Calibrated Fairness**
Ensure calibration per group:

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

For all groups and all scores $p$

**Methods:**

- Platt scaling per group
- Isotonic regression
- Beta calibration

**Cons:**

- Requires held-out data
- May conflict with other metrics
- Doesn't fix root cause

Post-processing is practical but treats symptoms, not causes

# Algorithms Summary: Measure and Mitigate

**Key Takeaways**

1. Multiple fairness metrics exist
2. Trade-offs are inevitable
3. Demographic parity: Equal outcomes
4. Equal opportunity: Equal TPR
5. Equalized odds: Equal TPR & FPR
6. Three mitigation stages: Pre, in, post

**Critical Choice:**
Which metric matches your values and context?

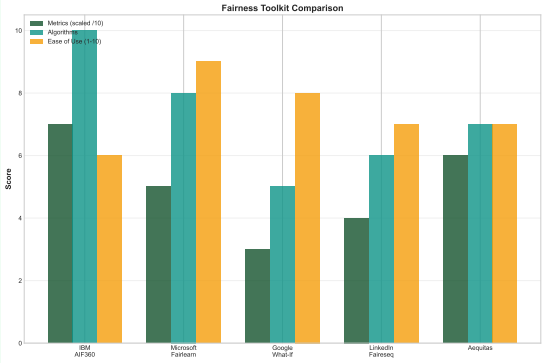
**Next: Implementation**
From theory to practice:

- Fairness toolkits
- Explainability methods
- Model cards
- Privacy techniques

**Remember:**
Mathematics provides tools, not answers. Human judgment required.

Fairness is Multi-Dimensional

# Fairness Toolkits Comparison




## Top Toolkits

**IBM AIF360**

- 70+ metrics
- 10+ algorithms
- Python & R

**Fairlearn**

- Microsoft-backed
- Sklearn integration
- Strong visualization

**Google What-If**

- Interactive exploration
- TensorBoard integration
- Visual debugging

Choose based on your stack and needs

# IBM AIF360: Comprehensive Fairness

**Key Features**

Metrics:

- 70+ fairness metrics
- Group & individual fairness
- Intersectional analysis
- Temporal fairness

Algorithms:

- Pre-processing (4 methods)
- In-processing (3 methods)
- Post-processing (3 methods)

Datasets:

- 10 benchmark datasets
- Preprocessed & ready
- Academic standard

**Example Usage**

```
from aif360.datasets import
  AdultDataset
from aif360.metrics import
  BinaryLabelDatasetMetric
dataset = AdultDataset()
metric = BinaryLabelDatasetMetric(
  dataset,
  unprivileged_groups=[
    {'sex': 0}],
  privileged_groups=[
    {'sex': 1}]
)
print(metric.mean_difference())
print(metric.disparate_impact())
```

AIF360: Best for comprehensive fairness assessment

# Fairlearn: Easy sklearn Integration

## Core Concepts

### Assessment

- MetricFrame for group metrics
- Disaggregated analysis
- Interactive dashboards

### Mitigation

- GridSearch: Threshold optimization
- ExponentiatedGradient: In-processing
- ThresholdOptimizer: Post-processing

### Constraints

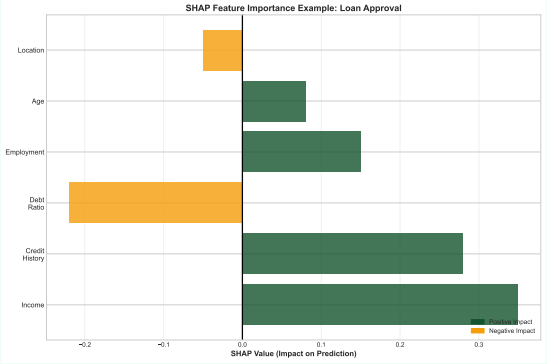
- Demographic parity
- Equalized odds
- True positive rate parity
- Bounded group loss

## Example Code

```
from fairlearn.reductions import
  ExponentiatedGradient
from fairlearn.reductions import
  DemographicParity
constraint = DemographicParity()
mitigator = ExponentiatedGradient(
  estimator,
  constraints=constraint
)
mitigator.fit(X, y,
  sensitive_features=A)
y_pred = mitigator.predict(X_test)
```

Fairlearn: Best for quick sklearn integration

# SHAP: Explaining Model Decisions




SHAP values: The gold standard for feature importance

## SHapley Additive exPlanations

**Theory:**

- Game-theoretic foundation
- Shapley values
- Additive feature attribution
- Mathematically rigorous

**Practical:**

- Model-agnostic
- Fast approximations
- Beautiful visualizations
- Local & global explanations

# LIME: Local Interpretable Model-agnostic Explanations

### How LIME Works

1. Perturb input locally
2. Get model predictions
3. Fit simple linear model
4. Explain with coefficients

**Key Idea:**
Complex models are locally linear

**Advantages:**

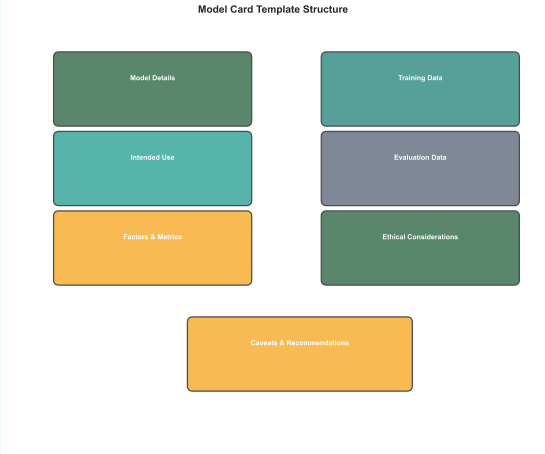
- Truly model-agnostic
- Works on any data type
- Human-interpretable
- Fast computation

### Example Usage

```
from lime.lime_tabular import
  LimeTabularExplainer
explainer = LimeTabularExplainer(
  X_train,
  feature_names=features,
  class_names=classes,
  mode='classification'
)
explanation = explainer.explain_instance(
  X_test[i],
  model.predict_proba
)
explanation.show_in_notebook()
```

LIME: Best for quick local explanations

# Model Cards: Documentation Standard




## Essential Documentation

**Required Sections:**

- Model details
- Intended use
- Factors (demographics)
- Metrics
- Training data
- Evaluation data
- Ethical considerations
- Caveats & recommendations

**Benefits:**

- Transparency
- Accountability
- Risk communication

Model cards should be mandatory for deployed systems

# Datasheets: Documenting Training Data

**Why Datasheets Matter**

Problems with undocumented data:

- Unknown biases
- Unclear provenance
- Misuse in new contexts
- Privacy violations
- Reproduction failures

**Inspiration:**
Electronics datasheets – standardized, comprehensive, essential

**Core Questions**

Motivation

- Why was dataset created?
- Who funded it?

Composition

- What do instances represent?
- How many instances?
- Missing data?

Collection

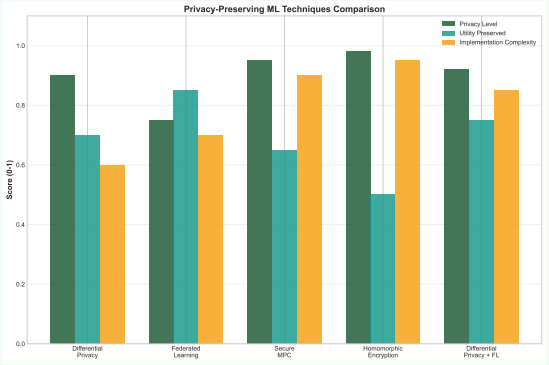
- How was data acquired?
- Who was involved?
- Ethical review?

Uses

- Intended tasks?
- What to avoid?

Dataset documentation prevents downstream harms

# Privacy Techniques




## Three Approaches

**Differential Privacy**

- Add calibrated noise
- Formal privacy guarantee
- Accuracy trade-off

**Federated Learning**

- Train locally
- Share only updates
- Keep data distributed

**Secure Multi-party Computation**

- Cryptographic protocols
- Compute on encrypted data
- No raw data exposure

Privacy and utility can coexist with right techniques

# Differential Privacy: Formal Guarantees

### Mathematical Definition

A mechanism $M$ is $(\epsilon, \delta)$-differentially private if:

$$P(M(D) \in S) \leq e^{\epsilon} P(M(D') \in S) + \delta$$

For all datasets $D, D'$ differing in one record

**Intuition:**

- Adding/removing one person
- Changes output minimally
- Individual contribution hidden
- Privacy budget $\epsilon$

### Practical Implementation

```
from diffprivlib.models import
  LogisticRegression
clf = LogisticRegression(
  epsilon=1.0,
  data_norm=5.0
)
clf.fit(X, y)
```

**Trade-off:**

- Lower $\epsilon$ = More privacy
- Lower $\epsilon$ = Less accuracy
- Typical: $\epsilon \in [0.1, 10]$

Differential privacy: Industry standard for privacy-preserving analytics

# Implementation Summary: Tools & Techniques

**Key Takeaways**

1. Fairness toolkits exist and work
2. AIF360: Comprehensive
3. Fairlearn: Easy integration
4. SHAP & LIME: Explainability
5. Model cards: Documentation
6. Privacy techniques available

**Action Item:**
Install and try one toolkit this week

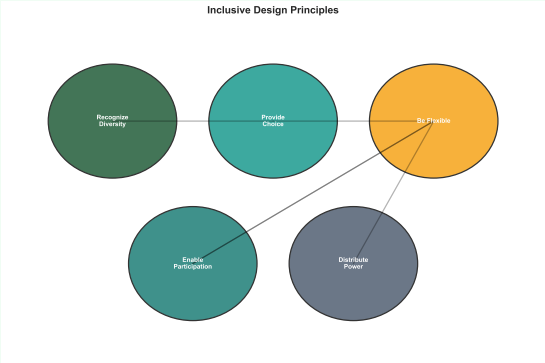
**Next: Design**
Human-centered perspective:

- Inclusive design
- Accessibility
- User consent
- Environmental impact
- Real case studies

**Remember:**
Tools are necessary but not sufficient – judgment required

Implementation Makes Ethics Real

# Inclusive Design for AI




## Design for All

**Core Principles:**

- Recognize diversity
- Provide choice
- Be flexible
- Enable participation
- Distribute power

**In Practice:**

- Multiple input modalities
- Customizable interfaces
- Cultural adaptation
- Accessibility by default

Design with, not for, marginalized communities

# Accessibility in AI Systems

**WCAG 2.1 for AI**

Perceivable

- Alt text for AI-generated images
- Captions for voice interfaces
- High contrast UI
- Screen reader compatibility

Operable

- Keyboard navigation
- Sufficient time limits
- No seizure-inducing patterns
- Clear navigation

Understandable

- Plain language explanations
- Consistent behavior
- Input assistance
- Error prevention

Robust

- Compatible with assistive tech
- Future-proof markup
- Graceful degradation

**Legal Requirement:**
EU Accessibility Act (2025)
ADA compliance (US)

Accessibility benefits everyone, not just disabled users

# User Transparency and Consent

**Transparency Layers**

Level 1: Existence

- Disclose AI is being used
- Not a human interaction
- System capabilities

Level 2: Process

- How decisions are made
- What data is used
- Who has access

Level 3: Reasoning

- Why this specific decision
- Contributing factors
- Actionable recourse

**Consent Design**

Informed Consent Requires:

- Clear language (not legalese)
- Specific purposes
- Real choices
- Easy opt-out
- Granular control

**GDPR Requirements:**

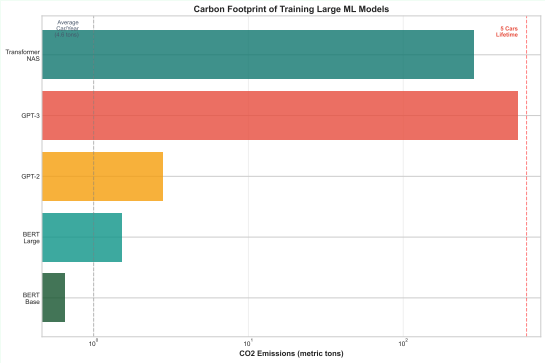
- Right to explanation
- Right to deletion
- Right to object
- Right to human review

**Anti-pattern:**

"Agree or leave" is not real consent

Transparency without agency is surveillance

# AI's Carbon Footprint




## Hidden Environmental Cost

**Training Emissions:**

- GPT-3: 552 tons CO2
- BERT: 1438 lbs CO2
- Transformer NAS: 626k lbs CO2

**Inference at Scale:**

- Billions of queries daily
- Data center energy
- Cooling requirements

**Mitigation:**

- Efficient architectures
- Green data centers
- Model compression

Ethics includes environmental responsibility

# Case Study: Detroit Wrongful Arrest

**The Incident (2020, settled 2024)**

Facts:

- Robert Williams, Black man
- Arrested for shoplifting
- Facial recognition false match
- Detained 30 hours
- Charges eventually dropped

Technical Failure:

- Low-quality image
- No confidence threshold
- No human verification
- Known bias in FR systems

**Lessons Learned**

What Went Wrong:

- Over-reliance on technology
- No quality checks
- Ignored known limitations
- Procedural shortcuts
- Lack of accountability

Outcome:

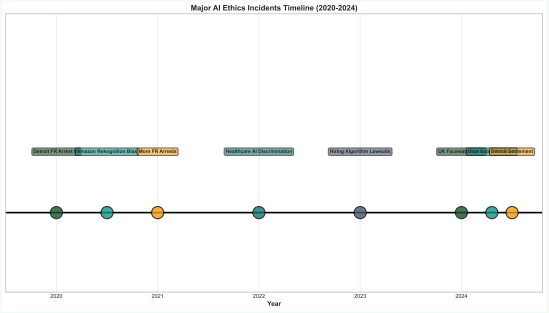
- Settlement & policy change
- Detroit now bans FR-only arrests
- National attention to FR bias
- ACLU lawsuits in multiple cities

**Key Insight:**

Technology amplifies existing biases in criminal justice

Source: ACLU, Detroit Free Press, 2024

# AI Ethics Failures: 2020-2024 Timeline




## Major Incidents

**2020**

- Detroit FR arrest #1
- Amazon Rekognition bias exposed

**2021-2023**

- More FR wrongful arrests
- Healthcare AI discrimination
- Hiring algorithm lawsuits

**2024**

- 233 incidents reported
- UK Facewatch case
- Uber Eats driver case
- Policy reforms begin

Pattern: Marginalized communities disproportionately harmed

# Case Study: AI in Hiring

**The Problem**

2024 Reality:

- 79% of Fortune 500 use AI screening
- Most managers untrained
- Women & minorities disadvantaged
- Black-box decision systems

How Bias Enters:

- Historical hiring data (biased)
- Proxy variables (school names)
- Resume keywords (gendered)
- Face/voice analysis (discriminatory)

**Documented Cases**

Amazon (2018):

- Resume screening tool
- Penalized "women's college"
- Learned from male-dominated data
- Scrapped after discovery

HireVue (2021):

- Video interview AI
- Analyzed facial expressions
- Lacked scientific validity
- Dropped after criticism

**Regulatory Response:**

NYC Local Law 144 (2023)
Requires bias audits

Hiring AI perpetuates discrimination at scale

# Participatory Design with Affected Communities

**Why It Matters**

Traditional Approach:

- Design for users
- Expert-driven
- Top-down
- Assumptions-based

Participatory Approach:

- Design with users
- Community-driven
- Bottom-up
- Experience-based

**Especially Critical:**
When system affects marginalized groups without choice

**Practical Methods**

Community Engagement:

- Co-design workshops
- Paid advisory boards
- Continuous feedback
- Shared decision-making

Compensation:

- Fair payment for time
- Acknowledge expertise
- Share ownership
- Long-term relationships

**Examples:**

- Detroit Digital Justice Coalition
- Data for Black Lives
- Indigenous data sovereignty

Nothing about us without us

# Value Sensitive Design Framework

**Three Iterations**

1. Conceptual Investigation
   - Identify stakeholders
   - Map values & tensions
   - Recognize trade-offs
   - Document assumptions
2. Empirical Investigation
   - Observe actual use
   - Interview stakeholders
   - Measure impacts
   - Test hypotheses
3. Technical Investigation
   - Design for values
   - Build prototypes
   - Evaluate alternatives
   - Iterate based on feedback

Embed values from the start, not as afterthought

**Core Values to Consider**

Human Welfare:

- Safety
- Health
- Peace

Human Dignity:

- Autonomy
- Privacy
- Non-discrimination

Justice:

- Fairness
- Equality
- Access

**Key Insight:**
Values often conflict – design must navigate trade-offs explicitly

# Red Teaming: Adversarial Ethics Testing

**What is Red Teaming?**

Origin:

- Military/cybersecurity practice
- Simulated attack scenarios
- Find vulnerabilities before adversaries

For AI Ethics:

- Deliberately try to break fairness
- Probe for hidden biases
- Test edge cases
- Challenge assumptions
- Adversarial prompting

**Red Team Composition:**

- Diverse perspectives
- Domain experts
- Affected community members
- Security professionals

Assume adversaries – ethical and malicious

**Red Teaming Process**

Phase 1: Threat Modeling

- Identify attack surfaces
- Map potential harms
- Prioritize risks

Phase 2: Testing

- Systematic probing
- Boundary exploration
- Stress testing
- Edge case generation

Phase 3: Reporting

- Document findings
- Assess severity
- Recommend mitigations
- Track remediation

**Result:**

More robust, ethical systems through adversarial thinking

# Design Summary: Human-Centered Ethics

**Key Takeaways**

1. Design with, not for, users
2. Accessibility is non-negotiable
3. Transparency enables agency
4. Environmental cost matters
5. Real harms to real people
6. Participatory design essential
7. Red team for robustness

**Critical Insight:**
Ethics is about power, not just fairness metrics

**Next: Practice**
Putting it all together:

- Ethical audit framework
- Hands-on exercises
- Building review processes
- Regulatory compliance

**Remember:**
The most ethical AI is often the one not built

Center Humans, Not Technology

# Ethical Decision Framework


## Key Questions

1. **Should we build this?**
   - Is there a real need?
   - Are alternatives better?
   - Can we do it responsibly?
2. **Who might be harmed?**
   - Direct stakeholders?
   - Indirect stakeholders?
   - Future generations?
3. **How do we mitigate?**
   - Technical safeguards?
   - Governance structures?
   - Accountability mechanisms?

Start with "should we?" before "how?"

# Pre-Deployment Ethical Audit

**Data Audit**
Questions:

Y/N Documented provenance?
Y/N Consent obtained?
Y/N Representative samples?
Y/N Known biases identified?
Y/N Privacy protections?

**Model Audit**
Questions:

Y/N Fairness metrics measured?
Y/N Disaggregated performance?
Y/N Explainability tested?
Y/N Adversarial robustness?
Y/N Model card completed?

**Deployment Audit**
Questions:

Y/N Human oversight plan?
Y/N Appeal mechanism?
Y/N Monitoring dashboard?
Y/N Incident response plan?
Y/N Regular audits scheduled?

**Governance Audit**
Questions:

Y/N Clear accountability?
Y/N Ethics review completed?
Y/N Stakeholder input?
Y/N Legal compliance?
Y/N Insurance coverage?

**Pass Threshold:** 90% Yes

Any "No" should trigger investigation and mitigation

# Exercise: Bias Detection in Practice

**Scenario**

You're reviewing a credit scoring model:

Dataset:

- 100k loan applications
- 60% approved overall
- Features: income, assets, history

Performance:

- Overall accuracy: 78%
- AUC: 0.82

Demographic Breakdown:

- Group A (majority): 65% approved
- Group B (minority): 45% approved
- Base credit quality similar

**Your Tasks**

1. Identify Issues
   - What fairness violations?
   - Which metrics to measure?
   - Potential causes?
2. Diagnose
   - Data bias?
   - Model bias?
   - Threshold bias?
3. Propose Solutions
   - Pre-processing?
   - In-processing?
   - Post-processing?
   - Non-technical interventions?

**Discuss:**

Trade-offs between solutions

Work in groups, present findings

# Institutional Ethics Review Process

**Structure**

Ethics Review Board:

- Technical experts
- Domain specialists
- Ethicists/philosophers
- Legal counsel
- Community representatives
- Rotating membership

Scope:

- All high-risk AI systems
- Threshold: Impacts on 1000+ people
- Sensitive domains: Health, justice, employment
- Customer-facing decisions

**Process**

Stage 1: Initial Review

- Submit project proposal
- Risk assessment
- 2-week turnaround

Stage 2: Full Review

- Detailed documentation
- Stakeholder analysis
- Bias testing results
- Mitigation plans
- 6-week evaluation

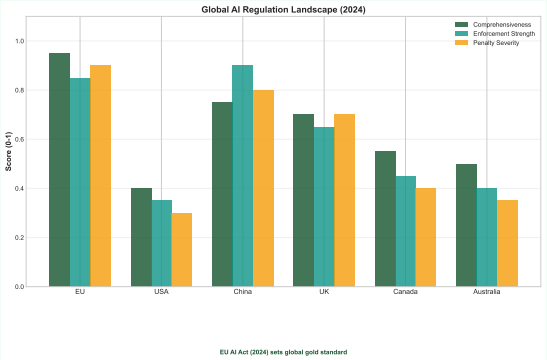
Stage 3: Monitoring

- Quarterly reports
- Incident tracking
- Annual re-certification

**Authority:**
Board can block deployment

Model after institutional review boards (IRBs) in research

# Global AI Regulations




## Key Regulations

**EU AI Act (2024)**

- Risk-based approach
- Prohibited practices
- High-risk requirements
- Fines up to 7% revenue

**GDPR (2018)**

- Right to explanation
- Data minimization
- Fines up to 4% revenue

**US (Fragmented)**

- State laws (CA, NY, etc.)
- Sector-specific rules
- No federal AI law yet

Compliance is complex but necessary

# EU AI Act: What You Need to Know

**Risk Categories**

Unacceptable Risk (Banned):

- Social scoring
- Manipulative AI
- Real-time biometric surveillance (limited exceptions)
- Emotion recognition in workplaces/schools

High Risk (Regulated):

- Critical infrastructure
- Education/employment
- Law enforcement
- Border control
- Administration of justice

Limited/Minimal Risk:

- Transparency requirements
- Self-regulation

**High-Risk Requirements**

Before Deployment:

- Risk management system
- Data governance
- Technical documentation
- Record-keeping
- Transparency
- Human oversight
- Accuracy requirements
- Cybersecurity
- Conformity assessment

After Deployment:

- Post-market monitoring
- Incident reporting
- Corrective actions

**Timeline:**

Fully applicable: August 2026

EU AI Act is most comprehensive regulation globally

# Week 7 Summary: Your Action Plan

**What We Covered**

1. Ethics foundations & real failures
2. Fairness metrics & trade-offs
3. Toolkits & implementation
4. Human-centered design
5. Practical frameworks

**Core Message:**
Ethics is not optional or an afterthought – it's essential for building trustworthy, sustainable, legitimate AI systems

**Key Insight:**
There is no purely technical solution to ethical problems – judgment required

**Action Items This Week**

Personal:

- Install one fairness toolkit
- Try SHAP or LIME
- Read one model card
- Review EU AI Act summary

Professional:

- Audit one current project
- Document ethical considerations
- Propose ethics review process
- Calculate environmental impact

Community:

- Share learnings with team
- Start ethics discussion
- Engage affected communities
- Advocate for better practices

Ethics is Everyone's Responsibility