Lecture 08 - Contents

An overview of the parts in the medical AI ethics and safety lecture.

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 8:

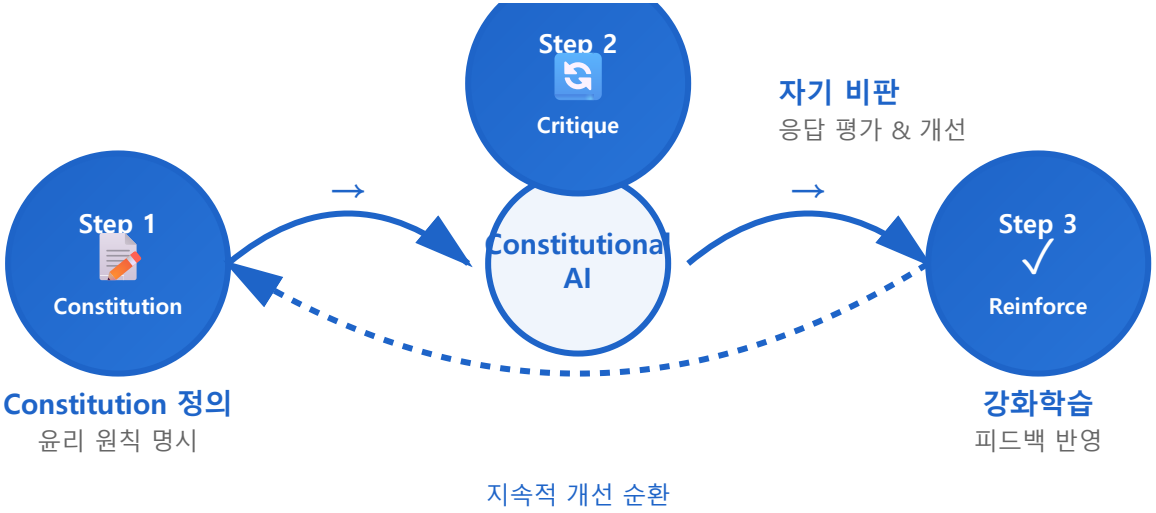
Constitutional AI and Medical Ethics

Building Ethical Medical AI Systems

Ethics Medicine AI

Constitutional AI (CAI) Overview

Constitutional AI는 명시적인 윤리 원칙(constitution)을 기반으로 **자기 개선(self-improvement)**하는 AI 시스템



자기 개선
스스로 응답을 평가하고 개선

윤리 정렬
명시적 원칙에 맞춰 정렬

피드백 루프
지속적인 학습과 개선

PART 1

Medical Ethics in AI

Four Principles of Biomedical Ethics (Beauchamp & Childress)

Ethical Framework Applied to Medical AI Systems

Beneficence

Providing benefits to patients and pursuing the best outcomes

Improving AI diagnostic accuracy, maximizing treatment effectiveness

Non-maleficence

Do no harm to patients ("First, do no harm")

Minimizing misdiagnosis, safety verification, preventing side effects

Autonomy

Respecting patient self-determination and providing information

Transparent AI explanations, ensuring choice, informed consent process

Justice

Fair access to treatment and resource allocation

Unbiased diagnosis, equal access, ensuring equity

⚠️ In actual medical practice, balance among these principles is necessary

Beneficence & Non-maleficence: Balancing Benefits and Harms

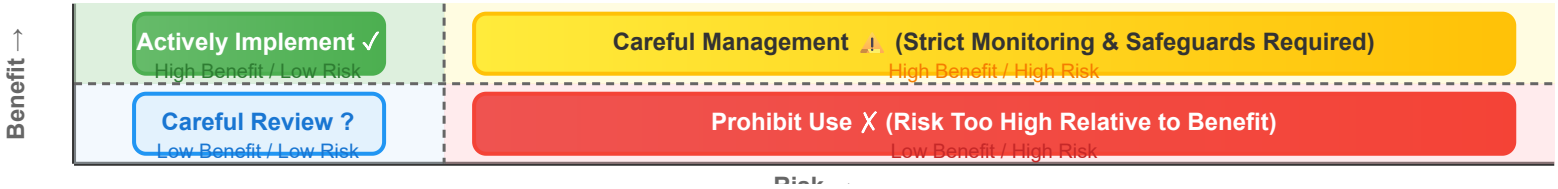
Maximizing Benefits

- **Improved Diagnostic Accuracy**
  Early detection and precise diagnosis with AI
- **Enhanced Treatment Outcomes**
  Personalized treatment planning
- **Expanded Healthcare Access**
  Telemedicine and automated support
- **Reduced Clinician Burden**
  Automation of repetitive tasks

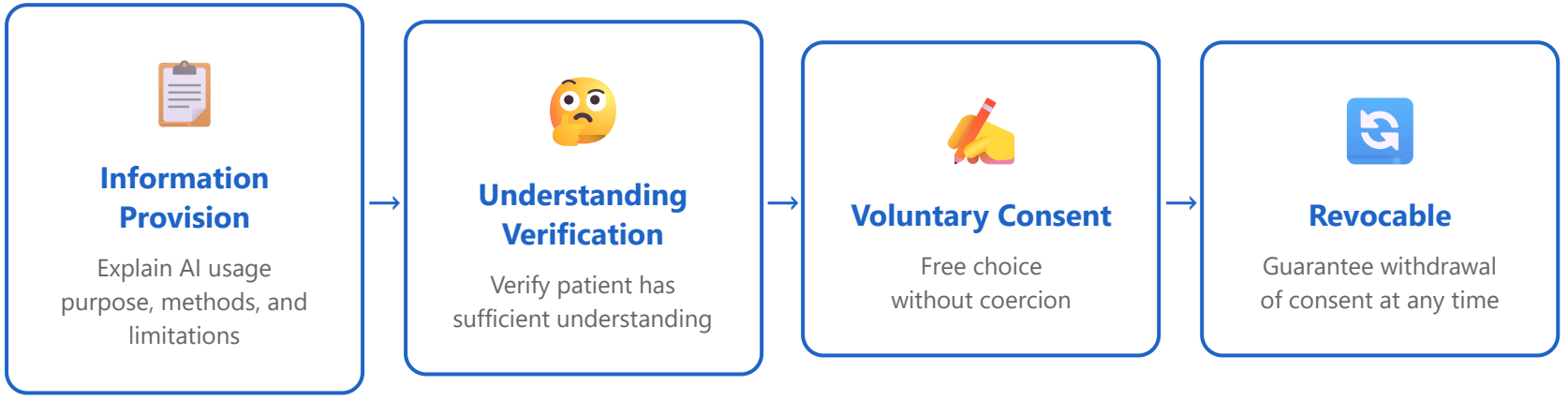
Minimizing Harms

- **Prevention of Misdiagnosis**
  Minimizing False Positives/Negatives
- **Adverse Effect Prevention**
  Drug interaction checking
- **Prevention of Overtreatment**
  Reducing unnecessary tests/procedures
- **Privacy Protection**
  Enhanced data security

Risk-Benefit Matrix



Patient Autonomy in AI



🔍 Transparency

- ✓ Explain AI algorithm operation
- ✓ Specify data usage scope
- ✓ Disclose decision-making process
- ✓ Notify accuracy and limitations

🎯 Choice

- ✓ Choice to use AI
- ✓ Provide alternative treatment options
- ✓ Determine data sharing scope
- ✓ Choose result notification method

📝 Explainability

- ✓ Present diagnostic basis
- ✓ Explain risk factors
- ✓ Reason for treatment recommendation

🔐 Control

- ✓ Personal information access rights
- ✓ Request data modification/deletion
- ✓ Review AI analysis results

- ✓ Specify uncertainty
- ✓ Guarantee objection procedure

Justice: Healthcare Equity and Access

Equity

Fair treatment
for all patients

Accessibility

Removing geographic/
economic barriers

Resource Allocation

Setting rational
priorities

Inequality Risks in AI Healthcare

- ⚠️ Lack of data for specific populations
- ⚠️ Limited access in rural areas
- ⚠️ Economic barriers (cost)
- ⚠️ Minority race/ethnic bias
- ⚠️ Digital Divide
- ⚠️ Language/cultural barriers

✓ Diverse Data Collection

Securing training data representing all population groups

✓ Bias Monitoring

Continuous measurement and improvement of performance gaps by group

✓ Universal Design

UI/UX considering diverse environments and users

✓ Public Investment

Support for AI healthcare services for vulnerable populations

Privacy Protection & Confidentiality

Layer 1: Data Collection

Collect only minimum necessary data, explicit consent

Layer 2: Storage and Transmission

Encryption, access control, audit logs

Layer 3: Use and Analysis

De-identification, differential privacy, permission management

Layer 4: Sharing and Disposal

Limited sharing, secure deletion

Technical Protection

Encryption

Anonymization

Differential Privacy

Administrative Protection

Access Control Management

Security Policy Establishment

Employee Training

Legal Protection

GDPR, HIPAA Compliance

Personal Information Protection Act

Medical Law Compliance

Informed Consent for AI Use

1 Information Provision

- AI usage purpose and method
- Expected benefits and risks
- Alternative treatment options

2 Understanding Verification

- Opportunity to ask questions
- Verification of understanding
- Use of clear language

3 Voluntary Consent

- Choice without coercion
- Right to refuse guaranteed
- Adequate consideration time

🔍 Transparency Requirements

Algorithm
Which AI model is being used?

Accuracy
What are the performance metrics?

Data
What data was used for training?

Limitations
In what situations do errors occur?

Role
Does AI make final decisions or assist?

Privacy
How is data protected?

Value Alignment in Medical AI

Value Alignment: The process of ensuring that an AI system's behavior aligns with human values, ethical principles, and cultural norms

Cultural Sensitivity

- Respect for religious beliefs
- Cultural treatment preferences
- Language and communication styles
- Degree of family involvement

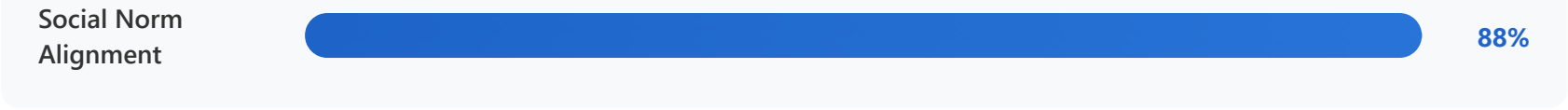
Priority Setting

- Life vs quality of life
- Cost vs effectiveness
- Individual vs community
- Short-term vs long-term benefits

Ethical Dilemmas

- Resource allocation criteria
- Emergency prioritization
- End-of-life care treatment
- Experimental treatment applications

Value Mapping Metrics

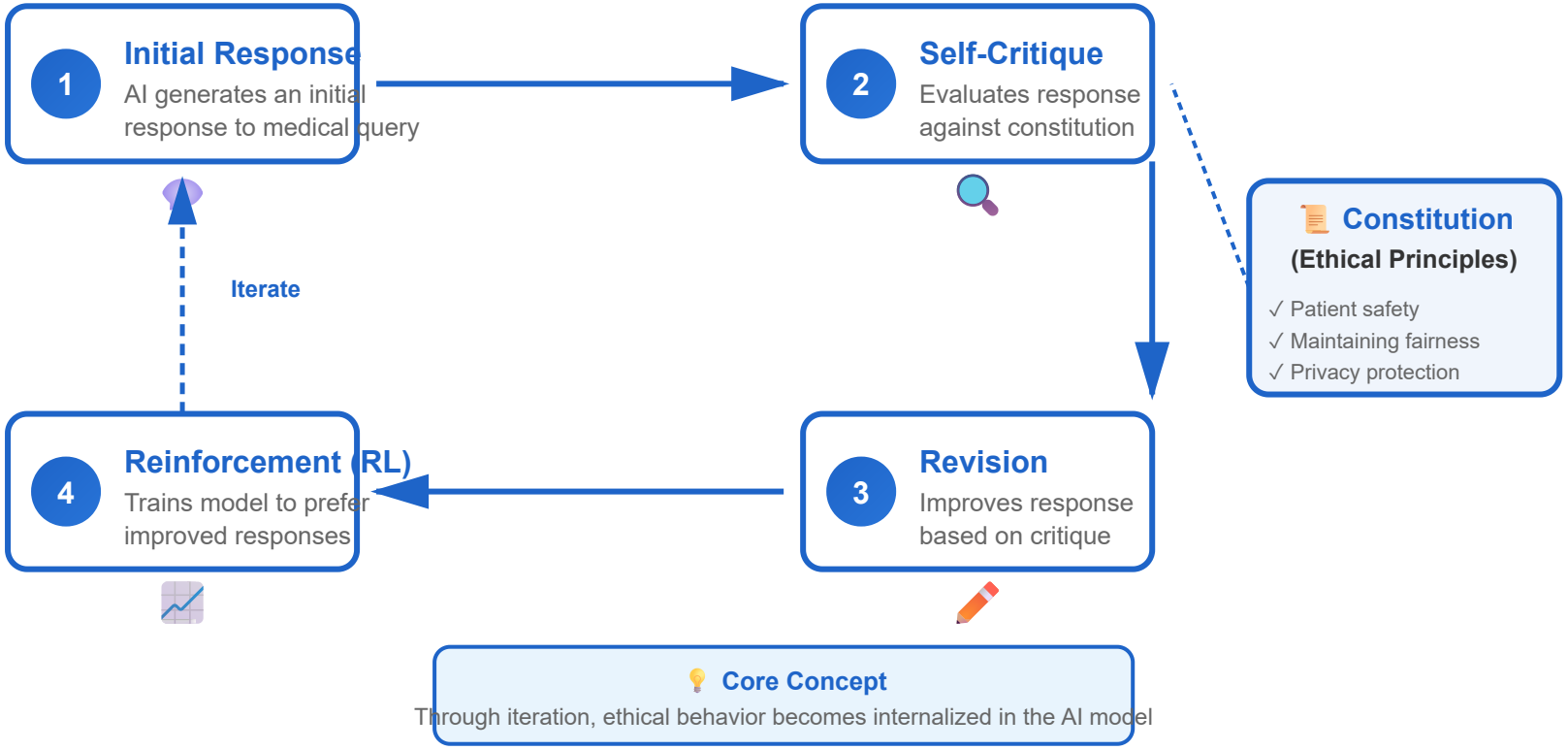

Social Norm Alignment
88%

PART 2

AI Safety in Healthcare

Constitutional Training

The process by which AI learns and applies ethical principles autonomously



Red Teaming for Medical AI

Red Teaming: A method of intentionally finding and attacking vulnerabilities in AI systems to verify their safety

Attack Scenarios

- Attempting to induce incorrect diagnosis
- Attempting to extract personal information
- Inducing harmful treatment recommendations
- Forcing biased decisions

Detection Methods

- Adversarial prompts
- Edge case testing
- Stress testing
- Cross-lingual attacks

Defense Strategies

- Input validation
- Output filtering
- Rate limiting
- Human oversight

🔬 Key Findings (Typical Findings)

Critical

Severe misdiagnosis risk

High

Privacy breach risk

Medium

Biased recommendations

Low

Minor inaccuracies

Adversarial Testing

Testing methodology to verify the robustness of AI models

Attack Vectors

- Input data manipulation (Data poisoning)
- Adversarial examples
- Prompt injection
- Model extraction attacks

Defense Mechanisms

- Adversarial training
- Input sanitization
- Ensemble methods
- Certified defense

Robustness Metrics

- Accuracy under attack
- Perturbation tolerance
- Recovery rate
- False positive rate

Testing Tools

- CleverHans
- Foolbox
- ART (Adversarial Robustness Toolbox)
- TextAttack

Safety Guardrails Implementation

Constraints that ensure AI systems operate only within safe boundaries

Input Guardrails

- Restrict allowed input range
- Detect malicious prompts
- Automatically remove personal information
- Format validation

Output Guardrails

- Filter harmful content
- Validate medical accuracy
- Specify uncertainty
- Add disclaimers

Action Guardrails

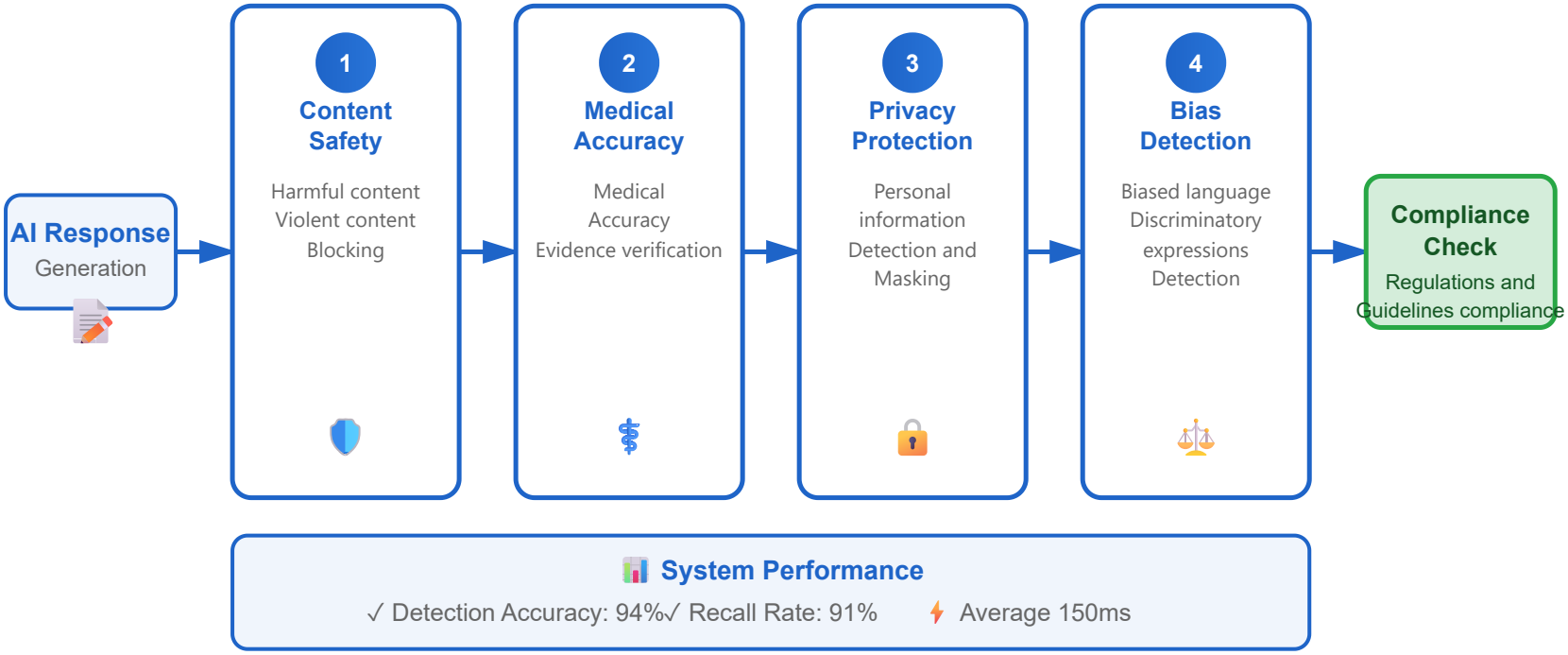
- Execute only permitted actions
- Permission-based access control
- Automatically record audit logs
- Threshold-based alerts

Dynamic Guardrails

- Context-based adjustment
- User role-based restrictions
- Real-time risk assessment
- Adaptive thresholds

Output Filtering System

Multi-layer filter pipeline ensuring AI response safety



해악 방지 전략 (Harm Prevention Strategies)

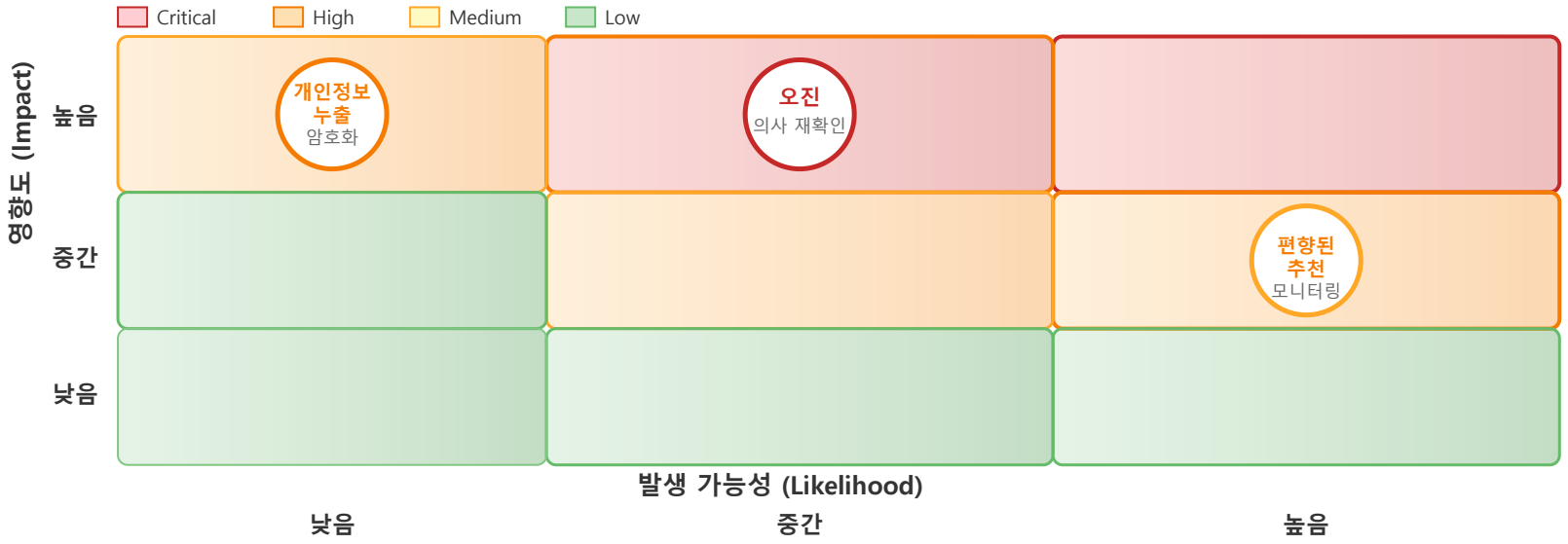
🎯 위험 평가

- 위험 식별 및 분류
- 발생 가능성 추정
- 영향도 분석
- 우선순위 결정

🛡️ 완화 전략

- 기술적 안전장치
- 인간 감독 체계
- 점진적 배포
- 비상 중지 메커니즘

📊 위험 매트릭스 (Risk Matrix)



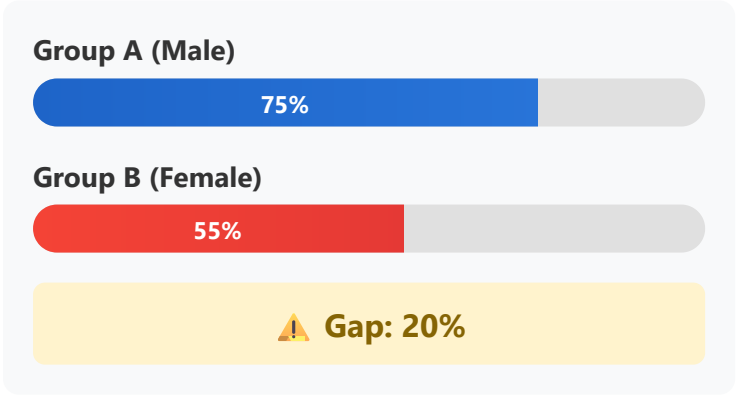
PART 3

Bias and Fairness in Medical AI

Demographic Parity

Demographic Parity: A fairness criterion that requires the rate of positive outcomes from an AI system to be equal across all demographic groups

📊 Parity Chart



🎯 Group Fairness

- Measure positive prediction rate for each demographic group
- Statistical significance testing
- Set acceptable gap thresholds
- Analyze and adjust for gap causes
- Continuous monitoring

✓ Objective

Achieve $P(\hat{Y}=1|Group=A) \approx P(\hat{Y}=1|Group=B)$ for all demographic groups

Health Equity Metrics

Gap Measurement

Absolute Gap
Relative Gap
Inequality Index

Accessibility

Geographic Access
Economic Access
Cultural Access

Outcome Equity

Treatment Outcomes
Survival Rates
Quality of Life

Equity Indicator Analysis

Diagnostic Accuracy Gap

- By Race: 5% gap
- By Gender: 3% gap
- By Age: 7% gap

Treatment Recommendation Gap

- By Income Level: 12% gap
- By Insurance Type: 8% gap
- By Residential Area: 10% gap

Improvement Goals

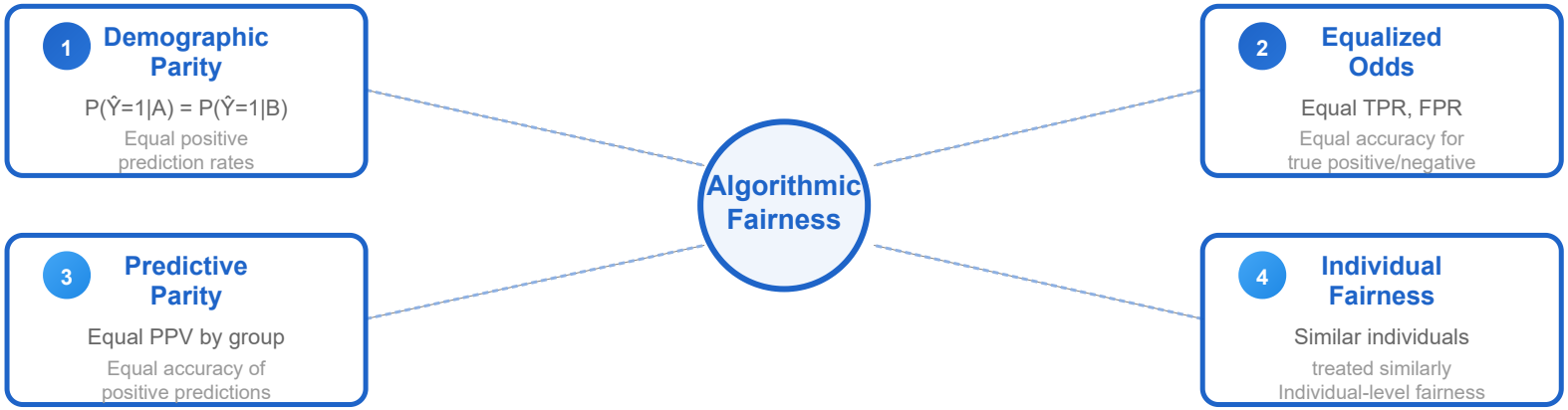
- Reduce gap to 5% or less
- Quarterly monitoring
- Prioritize vulnerable populations

Improvement Strategies

- Ensure data diversity
- Model retraining
- Establish feedback loops

Algorithmic Fairness

📊 Fairness Definitions



⚠️ **Impossible to satisfy all fairness criteria simultaneously**

⚠️ Trade-offs

- Fairness criteria cannot be satisfied simultaneously
- Accuracy vs fairness trade-offs
- Group fairness vs individual fairness
- Short-term vs long-term fairness

🛠️ Implementation Methods

- Pre-processing: Data resampling
- In-processing: Fairness-constrained learning
- Post-processing: Result adjustment
- Hybrid approaches

Disparate Impact Analysis

Disparate Impact: A phenomenon where seemingly neutral policies or practices have a disproportionately negative effect on specific groups

📊 4/5ths Rule

Disparate Impact Ratio

$$\frac{P(\hat{Y}=1|Group=B)}{P(\hat{Y}=1|Group=A)}$$

✓ ≥ **0.8**: Fair

✗ < **0.8**: Suspected Disparate Impact

🔍 Detection Methods

- Statistical Significance Testing (Chi-square test)
- Comparison of Performance Metrics by Group
- Intersectionality Analysis
- Tracking Changes Over Time
- Causal Analysis

⚠️ Examples in Medical AI

- Lower diagnosis rates for specific races
- Limited treatment access for elderly patients
- Under-treatment of female patients
- Lower recommendation grades for low-income patients

Representation Bias

Bias that occurs when training data fails to properly reflect actual population distribution

Data Imbalance Issues

- Underrepresentation of specific races/ethnicities
- Gender imbalance
- Age distribution distortion
- Lack of rare disease data
- Geographic concentration

Resulting Impact

- Lower accuracy in minority groups
- Increased misdiagnosis rates
- Treatment recommendation bias
- Widening health disparities
- Decreased reliability

🎯 Correction Strategies

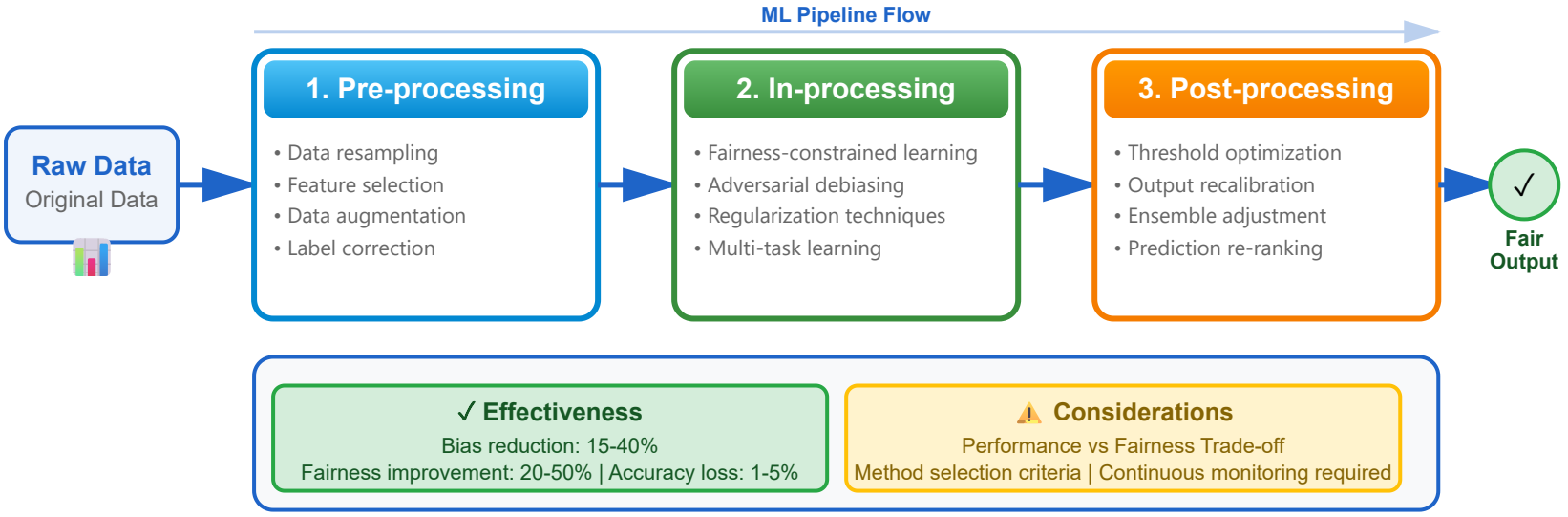
Resampling
Over/Under sampling, SMOTE

Weight Adjustment
Class weights, Sample weights

Data Augmentation
Augmentation, Synthetic data

Bias Mitigation Techniques

🛠️ Mitigation Methods by ML Pipeline Stage



Continuous Monitoring

Performance Monitoring

- Track Accuracy, Precision, Recall
- Measure performance gaps by group
- Analyze error patterns
- Conduct A/B testing

Drift Detection

- Data distribution changes (Data drift)
- Concept changes (Concept drift)
- Prediction distribution changes
- Set alert thresholds

Anomaly Detection

- Unexpected input patterns
- Abnormal outputs
- Detect performance drops
- Automated alert system

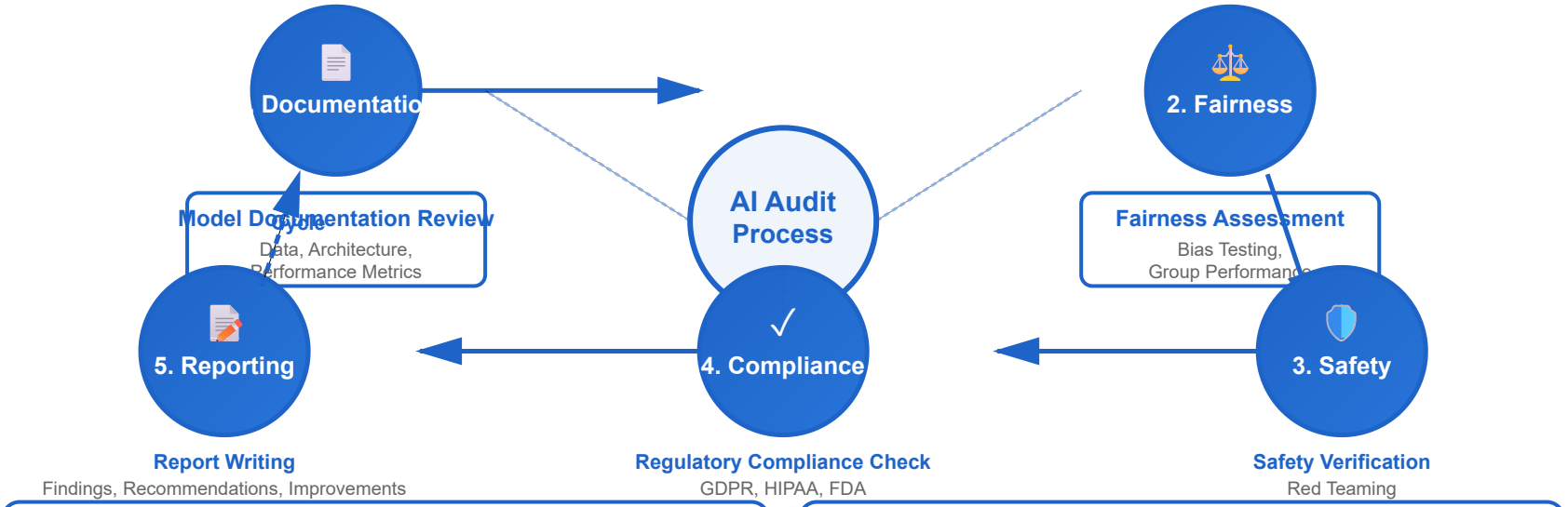
Update Management

- Periodic model retraining
- Incremental learning
- Version control and rollback
- Pre-deployment validation

📈 Key Monitoring Dashboard Metrics

Overall Accuracy	Fairness Score	Drift Alert	Uptime
92.5%	0.87	Low	99.8%

Audit Frameworks



Checklist

- Data Quality and Representativeness
- Model Performance and Robustness
- Explainability
- Security and Privacy
- Governance Framework

Documentation

- Model card
- Datasheet
- Risk assessment
- Testing results
- Deployment logs

Failure Case Studies

📌 Case 1: Racial Bias in Skin Lesion Diagnosis AI

Problem: Trained primarily on data from white patients, resulting in lower accuracy for patients of color

Result: 20% higher misdiagnosis rate for skin cancer in Black patients

Lesson: Need for balanced data collection across diverse racial groups

📌 Case 2: Racial Discrimination in Healthcare Resource Allocation Algorithm

Problem: Used healthcare costs as a proxy variable for health status

Result: Black patients incorrectly assessed as being less sick than white patients

Lesson: Must consider social inequalities when selecting proxy variables

📌 Case 3: Inappropriate Medical Advice from Chatbot

Problem: Lack of safety guardrails led to recommendations for dangerous self-treatment

Result: Patients delayed hospital visits in emergency situations

Lesson: Medical AI must include safety mechanisms and disclaimers

✓ Common Improvement Directions

- Secure data from diverse population groups
- Conduct thorough pre-testing and validation

- Implement continuous monitoring and feedback
- Ensure transparency and accountability

Hands-on: Bias Testing Code

Python Bias Testing Example

```
from sklearn.metrics import confusion_matrix import pandas as pd # Evaluate performance by group
def evaluate_fairness(y_true, y_pred, sensitive_attr): results = {} for group in
sensitive_attr.unique(): mask = sensitive_attr == group tn, fp, fn, tp = confusion_matrix(
y_true[mask], y_pred[mask] ).ravel() results[group] = { 'TPR': tp / (tp + fn), # Recall 'FPR': fp
/ (fp + tn), 'PPV': tp / (tp + fp) # Precision } return pd.DataFrame(results).T
```

Testing Tools

- Fairlearn (Microsoft)
- AI Fairness 360 (IBM)
- What-If Tool (Google)
- Aequitas

Visualization

- Performance comparison charts by group
- Confusion matrix heatmap
- Fairness metrics dashboard
- ROC curve comparison

Key Practice Points

Calculate TPR, FPR, and PPV for various demographic groups and verify that disparities are within acceptable ranges (e.g., 10%)

Ethics Committees

Committee Composition

- Medical professionals (doctors, nurses, etc.)
- AI/Data scientists
- Bioethicists
- Legal experts
- Patient representatives
- Sociologists, anthropologists

Key Roles

- AI system ethical review
- Research protocol approval
- Ethical dilemma resolution
- Policy and guideline development
- Post-implementation monitoring
- Education and awareness raising

Decision-Making Process

Application Submission
AI system proposal

→

Initial Review
Document verification

→

Full Committee Meeting
Multidisciplinary discussion

→

Decision & Monitoring
Approve/Conditional/Reject

Approved
No ethical concerns

Conditional Approval
Revise and resubmit

Rejected
Serious ethical concerns

Thank You!

Constitutional AI and Medical Ethics

핵심 메시지 요약

1. Constitutional AI는 명시적 윤리 원칙으로 AI의 자기 개선을 유도
2. 의료 AI는 생명윤리 4원칙(선행, 무해, 자율성, 정의)을 준수해야 함
3. 안전 메커니즘: Red teaming, Guardrails, Output filtering
4. 편향 완화: Pre/In/Post-processing 기법 적용
5. 지속적 모니터링과 윤리 위원회의 역할이 필수적

Questions?