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.

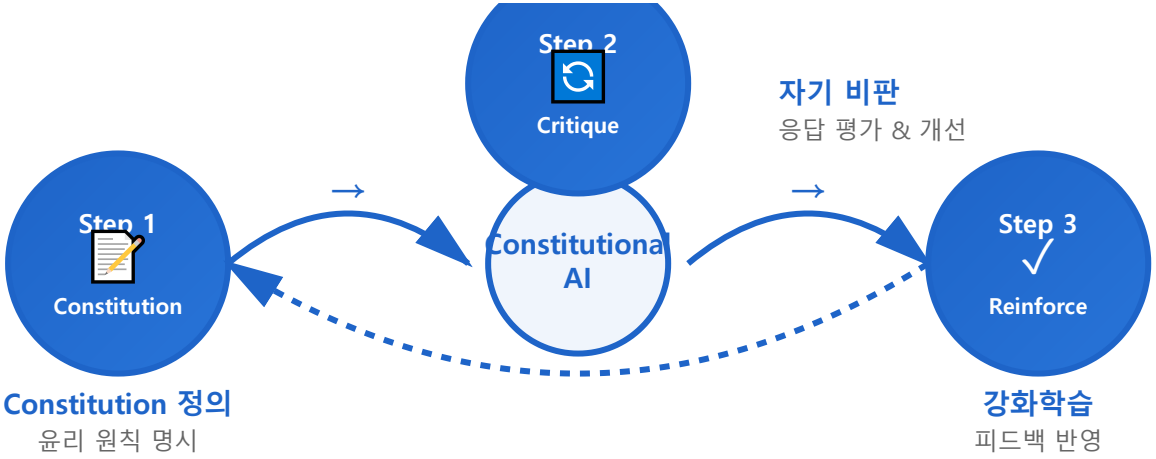
Lecture 8:

Constitutional AI and Medical Ethics

Building Ethical Medical AI Systems

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

생명윤리 4원칙 (Beauchamp & Childress)

의료 AI 시스템에 적용되는 윤리적 프레임워크

Beneficence (선행)

환자에게 이익을 제공하고 최선의 결과를 추구

AI 진단 정확도 향상, 치료 효과 최대화

Non-maleficence (무해)

환자에게 해를 끼치지 않음 ("First, do no harm")

오진 최소화, 안전성 검증, 부작용 예방

Autonomy (자율성)

환자의 자기결정권 존중 및 정보 제공

투명한 AI 설명, 선택권 보장, 동의 과정

Justice (정의)

공정한 치료 접근성과 자원 배분

편향 없는 진단, 평등한 접근, 형평성 보장

⚠ 실제 의료 현장에서는 이 원칙들 간의 균형(Balance)이 필요

Beneficence & Non-maleficence: 이익과 해악의 균형

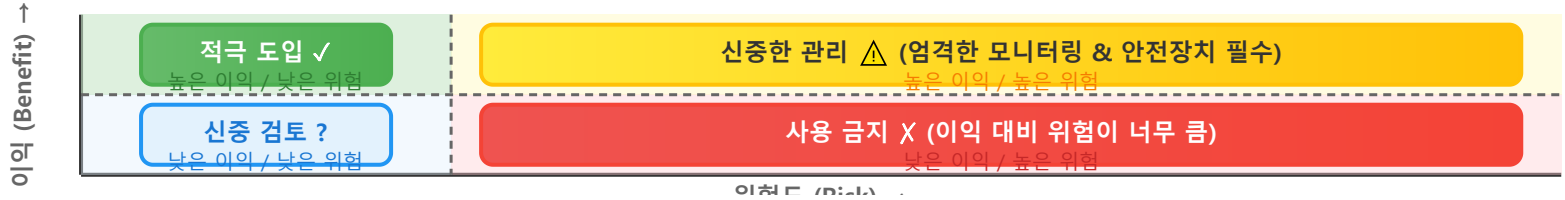
이익 최대화

- **진단 정확도 향상**
  AI로 조기 발견 및 정밀 진단
- **치료 효과 개선**
  개인화된 치료 계획 수립
- **의료 접근성 확대**
  원격 진료 및 자동화 지원
- **의료진 부담 경감**
  반복 업무 자동화

해악 최소화

- **오진 방지**
  False Positive/Negative 최소화
- **부작용 예방**
  약물 상호작용 체크
- **과잉 진료 방지**
  불필요한 검사/처치 감소
- **프라이버시 침해 방지**
  데이터 보안 강화

이익-위험 매트릭스 (Risk-Benefit Matrix)



환자 자율성 (Patient Autonomy) in AI

🔍 투명성 (Transparency)

- ✓ AI 알고리즘 작동 방식 설명
- ✓ 데이터 사용 범위 명시
- ✓ 의사결정 과정 공개
- ✓ 정확도 및 한계 고지

🎯 선택권 (Choice)

- ✓ AI 사용 여부 선택
- ✓ 대체 치료 옵션 제공
- ✓ 데이터 공유 범위 결정
- ✓ 결과 통지 방법 선택

📝 설명 의무 (Explainability)

- ✓ 진단 근거 제시
- ✓ 위험 요인 설명
- ✓ 치료 권고 이유
- ✓ 불확실성 명시

🔐 통제권 (Control)

- ✓ 개인정보 접근 권한
- ✓ 데이터 수정/삭제 요청
- ✓ AI 분석 결과 확인
- ✓ 이의제기 절차 보장

정의 (Justice): 의료 형평성과 접근성

형평성

모든 환자에게
공정한 대우

접근성

지리/경제적
장벽 제거

자원 분배

합리적인
우선순위 설정

🚨 AI 의료에서의 불평등 위험

- ⚠ 특정 인구 집단 데이터 부족
- ⚠ 농어촌 지역 접근성 제한
- ⚠ 경제적 장벽 (비용)
- ⚠ 소수 인종/민족 편향
- ⚠ 디지털 격차 (Digital Divide)
- ⚠ 언어/문화적 장벽

✓ 다양한 데이터 수집

모든 인구 집단을 대표하는 학습 데이터 확보

✓ 편향 모니터링

그룹별 성능 격차 지속적 측정 및 개선

✓ 보편적 설계

다양한 환경과 사용자를 고려한 UI/UX

✓ 공공 투자

취약 계층을 위한 AI 의료 서비스 지원

개인정보 보호 (Privacy) & 비밀 유지 (Confidentiality)

Layer 1: 데이터 수집
최소 필요 데이터만 수집, 명시적 동의

Layer 2: 저장 및 전송
암호화, 접근 제어, 감사 로그

Layer 3: 사용 및 분석
비식별화, 차등 프라이버시, 권한 관리

Layer 4: 공유 및 폐기
제한적 공유, 안전한 삭제

기술적 보호

암호화 (Encryption)
익명화 (Anonymization)
차등 프라이버시

관리적 보호

접근 권한 관리
보안 정책 수립
직원 교육

법적 보호

GDPR, HIPAA 준수
개인정보보호법
의료법 준수

AI 사용에 대한 정보 제공 동의 (Informed Consent)

1

정보 제공

- AI 사용 목적 및 방법
- 예상 이익과 위험
- 대안적 치료 방법

2

이해도 확인

- 질문 기회 제공
- 이해 여부 검증
- 명확한 언어 사용

3

자발적 동의

- 강압 없이 선택
- 거부 권리 보장
- 충분한 고려 시간

🔍 투명성 요구사항

알고리즘
어떤 AI 모델을 사용하는가?

정확도
성능 지표는 어느 정도인가?

데이터
어떤 데이터로 학습했는가?

한계
어떤 상황에서 오류가 발생하는가?

역할
AI가 최종 결정을 하는가, 보조하는가?

프라이버시
데이터는 어떻게 보호되는가?

의료 AI의 가치 정렬 (Value Alignment)

가치 정렬: AI 시스템의 행동이 인간의 가치관, 윤리 원칙, 문화적 규범과 일치하도록 하는 과정

문화적 민감성

- 종교적 신념 존중
- 문화적 치료 선호도
- 언어 및 의사소통 방식
- 가족 참여 정도

우선순위 설정

- 생명 vs 삶의 질
- 비용 vs 효과
- 개인 vs 공동체
- 단기 vs 장기 이익

윤리적 딜레마

- 자원 배분 기준
- 응급 상황 우선순위
- 말기 환자 치료
- 실험적 치료 적용

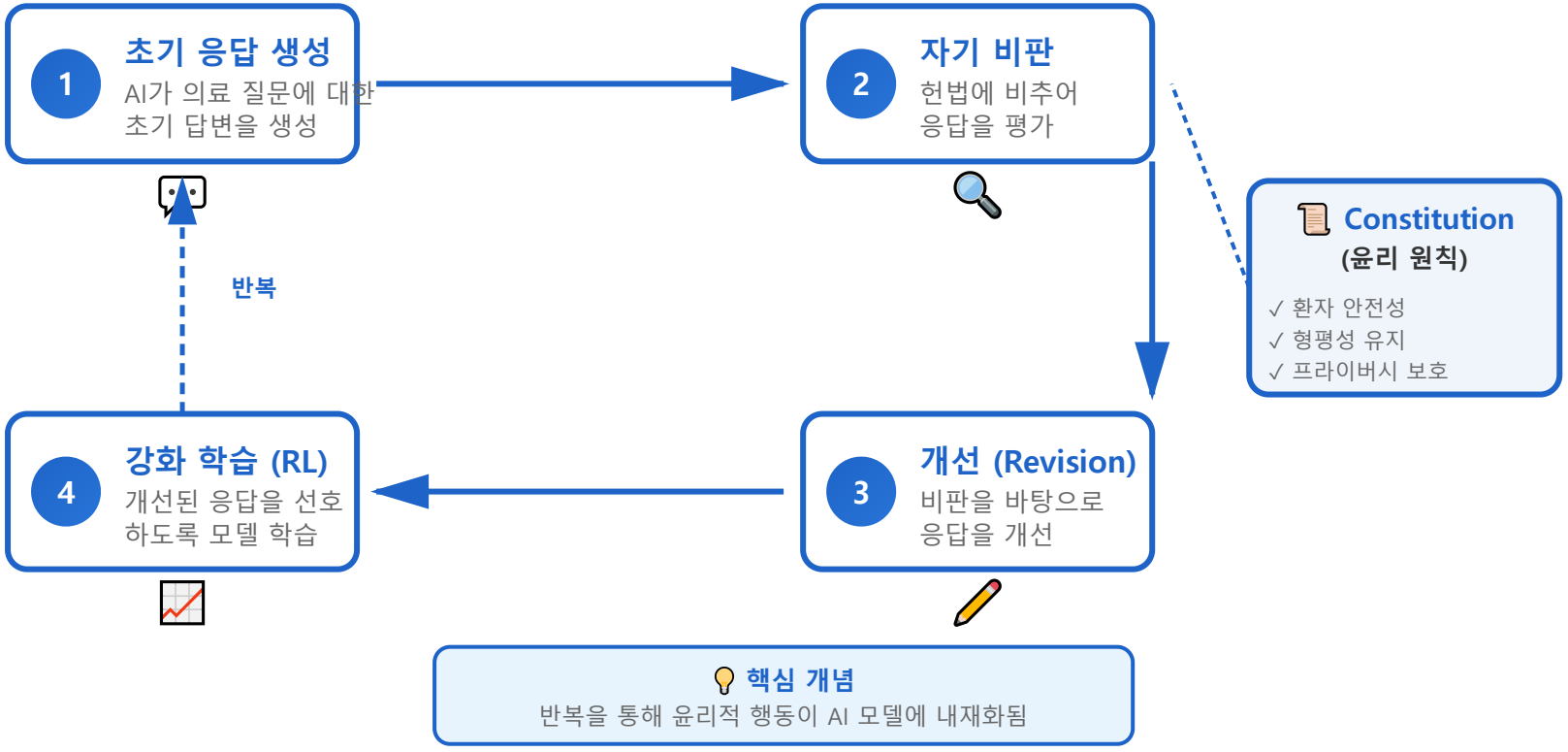
📊 가치 매핑 (Value Mapping) 메트릭

PART 2

AI Safety in Healthcare

헌법적 학습 (Constitutional Training)

AI가 스스로 윤리적 원칙을 학습하고 적용하는 과정



레드팀 테스트 (Red Teaming) in Medical AI

Red Teaming: 의도적으로 AI 시스템의 취약점을 찾고 공격하여 안전성을 검증하는 방법

공격 시나리오

- 잘못된 진단 유도 시도
- 개인정보 추출 시도
- 해로운 치료 권장 유도
- 편향된 결정 강요

탐지 방법

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

방어 전략

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

주요 발견 사항 (Typical Findings)

Critical
심각한 오진 가능성

High
프라이버시 누출 위험

Medium
편향된 추천

Low
경미한 부정확성

적대적 테스트 (Adversarial Testing)

AI 모델의 강건성(Robustness)을 검증하는 테스트 방법론

공격 벡터 (Attack Vectors)

- 입력 데이터 조작 (Data poisoning)
- 적대적 예제 (Adversarial examples)
- 프롬프트 인젝션 (Prompt injection)
- 모델 추출 공격

방어 메커니즘

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

강건성 메트릭

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

테스트 도구

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

안전 가드레일 (Safety Guardrails) 구현

AI 시스템이 안전한 범위 내에서만 작동하도록 보장하는 제약 조건

입력 가드레일

- 허용된 입력 범위 제한
- 악의적 프롬프트 탐지
- 개인정보 자동 제거
- 형식 검증

출력 가드레일

- 해로운 내용 필터링
- 의학적 타당성 검증
- 불확실성 명시
- 면책 조항 추가

행동 가드레일

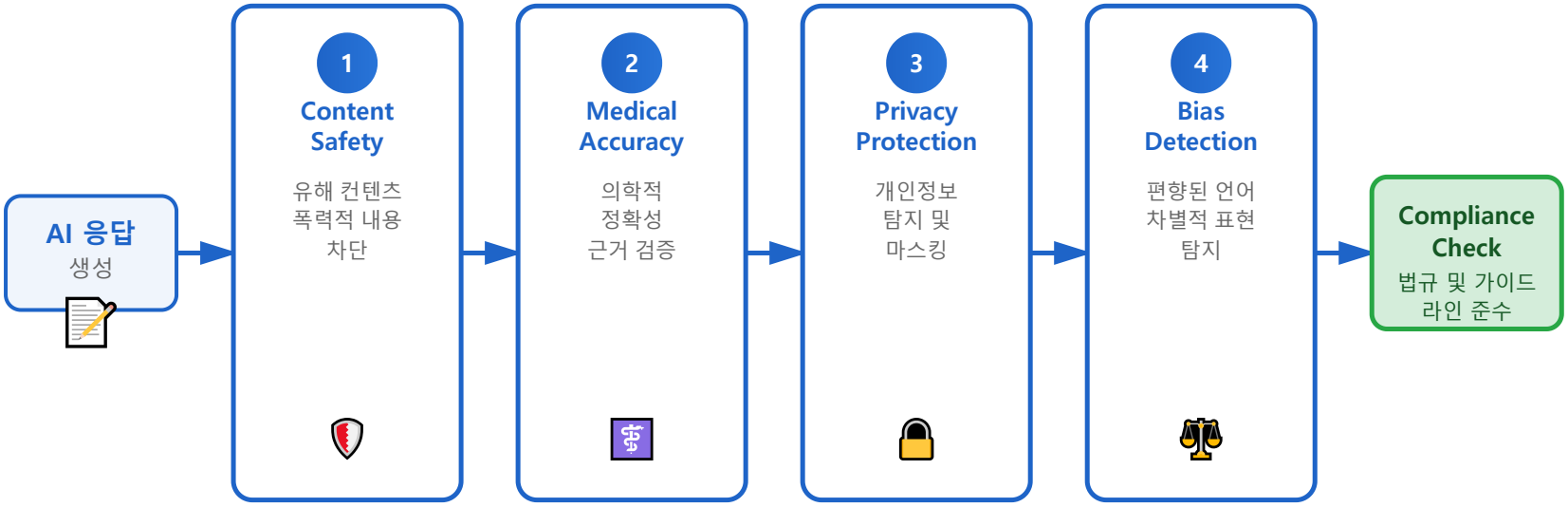
- 허용된 작업만 실행
- 권한 기반 접근 제어
- 감사 로그 자동 기록
- 임계값 기반 경고

동적 가드레일

- 컨텍스트 기반 조정
- 사용자 역할별 제한
- 실시간 위험 평가
- 적응형 임계값

출력 필터링 시스템 (Output Filtering)

AI 응답의 안전성을 보장하는 다층 필터 파이프라인



해악 방지 전략 (Harm Prevention Strategies)

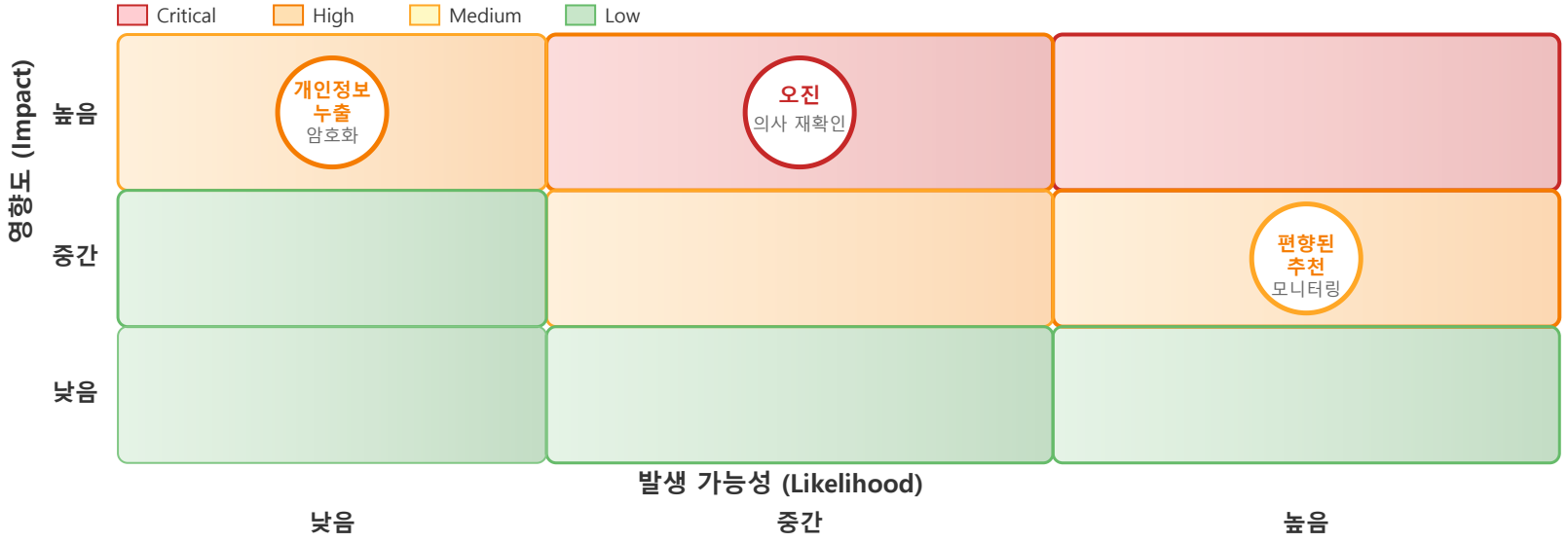
위험 평가

- 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)



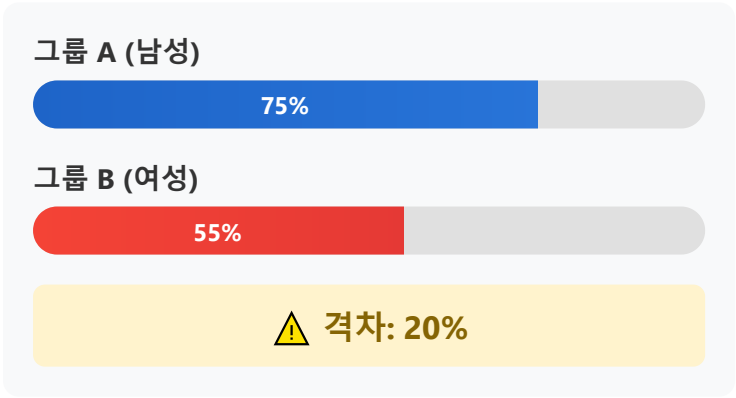
PART 3

Bias and Fairness in Medical AI

인구통계 균등성 (Demographic Parity)

Demographic Parity: AI 시스템의 긍정적 결과 비율이 모든 인구 집단에서 동일해야 한다는 공정성 기준

📊 균등성 차트



🎯 그룹 공정성

- 각 인구 집단의 긍정 예측 비율 측정
- 통계적 유의성 검정
- 허용 가능한 격차 범위 설정
- 격차 원인 분석 및 조정
- 지속적인 모니터링

✓ 목표

모든 인구 집단에 대해 $P(\hat{Y}=1|Group=A) \approx P(\hat{Y}=1|Group=B)$ 달성

건강 형평성 메트릭 (Health Equity Metrics)

격차 측정

절대 격차
상대 격차
불평등 지수

접근성

지리적 접근
경제적 접근
문화적 접근

결과 형평성

치료 성과
생존율
삶의 질

📈 형평성 지표 분석

진단 정확도 격차

- 인종별: 5% 격차
- 성별: 3% 격차
- 연령별: 7% 격차

치료 권장 격차

- 소득 수준별: 12% 격차
- 보험 유형별: 8% 격차
- 거주 지역별: 10% 격차

개선 목표

- 격차 5% 이하로 감소
- 분기별 모니터링
- 취약 계층 우선 개선

개선 전략

- 데이터 다양성 확보
- 모델 재학습
- 피드백 루프 구축

알고리즘 공정성 (Algorithmic Fairness)

📊 공정성 정의 (Fairness Definitions)

⚠ 모든 공정성 기준을 동시에 만족하기는 불가능

⚠ Trade-offs

- 공정성 기준들은 동시에 만족 불가능
- 정확도 vs 공정성 트레이드오프
- 그룹 공정성 vs 개인 공정성
- 단기 vs 장기 공정성

🛠 구현 방법

- Pre-processing: 데이터 재샘플링
- In-processing: 공정성 제약 학습
- Post-processing: 결과 조정
- Hybrid approaches

차별적 영향 분석 (Disparate Impact Analysis)

Disparate Impact: 중립적으로 보이는 정책이나 관행이 특정 집단에 불균형적인 부정적 영향을 미치는 현상

📊 4/5ths Rule

Disparate Impact Ratio

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

✓ **≥ 0.8**: 공정함

✗ **< 0.8**: 차별적 영향 의심

🔍 탐지 방법

- 통계적 유의성 검정 (Chi-square test)
- 그룹별 성능 메트릭 비교
- 교차분석 (Intersectionality)
- 시간에 따른 변화 추적
- 인과관계 분석

⚠ 의료 AI에서의 사례

- 특정 인종의 낮은 진단율
- 노인 환자의 치료 접근성 제한
- 여성 환자의 과소 치료
- 저소득층 환자의 낮은 권장 등급

대표성 편향 (Representation Bias)

학습 데이터가 실제 인구 분포를 제대로 반영하지 못할 때 발생하는 편향

데이터 불균형 문제

- 특정 인종/민족 과소 대표
- 성별 불균형
- 연령 분포 왜곡
- 희귀 질환 데이터 부족
- 지리적 편중

결과적 영향

- 소수 집단에서 낮은 정확도
- 오진율 증가
- 치료 권장 편향
- 건강 격차 심화
- 신뢰도 저하

🎯 보정 전략

재샘플링

Over/Under sampling, SMOTE

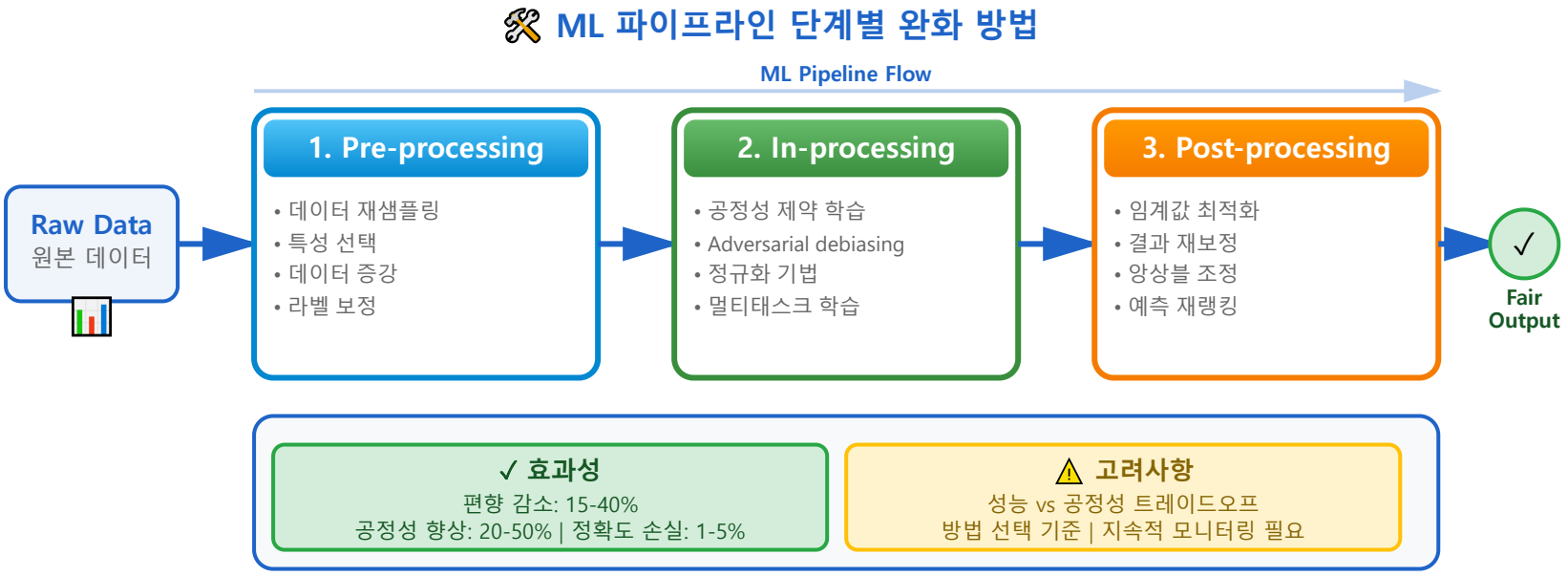
가중치 조정

Class weights, Sample weights

데이터 증강

Augmentation, Synthetic data

편향 완화 기법 (Bias Mitigation Techniques)



지속적 모니터링 (Continuous Monitoring)

성능 모니터링

- 정확도, Precision, Recall 추적
- 그룹별 성능 격차 측정
- 오류 패턴 분석
- A/B 테스트 실시

드리프트 탐지

- 데이터 분포 변화 (Data drift)
- 개념 변화 (Concept drift)
- 예측 분포 변화
- 경고 임계값 설정

이상 탐지

- 예상치 못한 입력 패턴
- 비정상 출력
- 성능 급락 감지
- 자동 알림 시스템

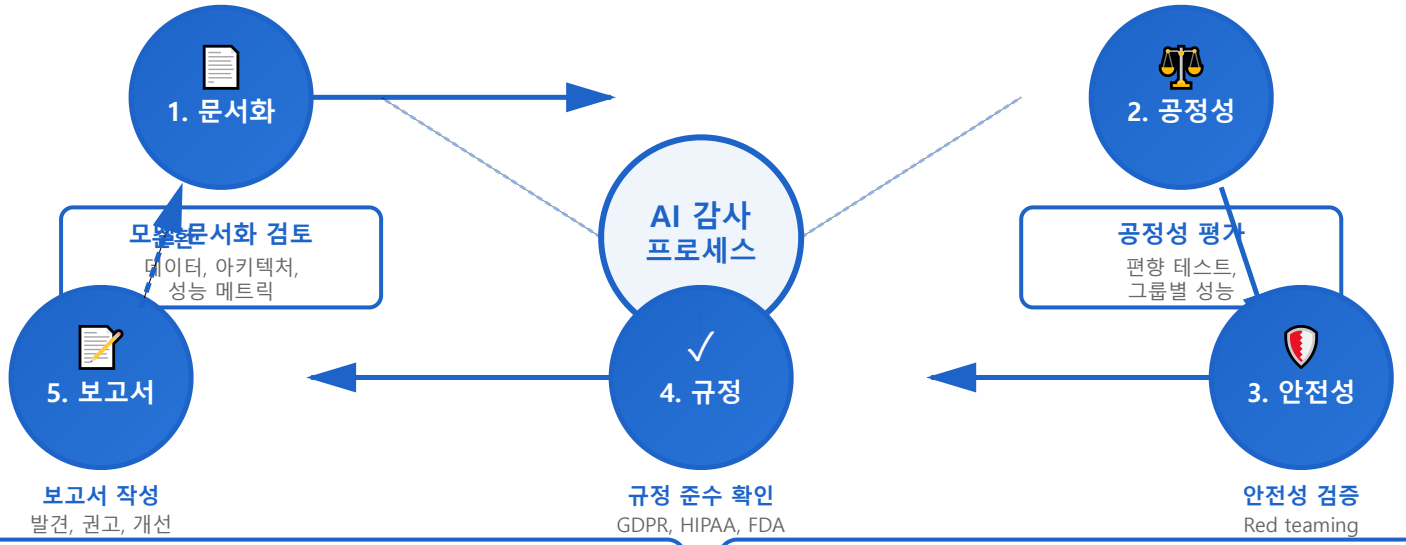
업데이트 관리

- 주기적 모델 재학습
- 증분 학습 (Incremental learning)
- 버전 관리 및 롤백
- 배포 전 검증

📈 모니터링 대시보드 주요 지표

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

감사 프레임워크 (Audit Frameworks)



체크리스트

- 데이터 품질 및 대표성
- 모델 성능 및 강건성
- 설명 가능성
- 보안 및 프라이버시
- 거버넌스 체계

문서화

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

실패 사례 분석 (Case Studies: Failures)

📌 Case 1: 피부 병변 진단 AI의 인종 편향

문제: 백인 환자 위주 데이터로 학습하여 유색인종 환자에서 낮은 정확도

결과: 흑인 환자의 피부암 오진율 20% 높음

교훈: 다양한 인종의 균형잡힌 데이터 수집 필요

📌 Case 2: 의료 자원 배분 알고리즘의 인종 차별

문제: 의료 비용을 건강 상태의 대리 변수로 사용

결과: 흑인 환자가 백인 환자보다 덜 아픈 것으로 잘못 평가

교훈: 대리 변수 선택 시 사회적 불평등 고려 필요

📌 Case 3: 챗봇의 부적절한 의료 조언

문제: 안전 가드레일 부족으로 위험한 자가 치료 권장

결과: 환자가 응급 상황에서 병원 방문 지연

교훈: 의료 AI는 반드시 안전 장치와 면책 조항 필요

✓ 공통 개선 방향

- 다양한 인구 집단의 데이터 확보
- 철저한 사전 테스트 및 검증

- 지속적인 모니터링 및 피드백
- 투명성과 책임성 확보

실습: 편향 테스트 코드 (Hands-on: Bias Testing)

Python 편향 테스트 예제

```
from sklearn.metrics import confusion_matrix import pandas as pd # 그룹별 성능 평가 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
```

테스트 도구

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

시각화

- 그룹별 성능 비교 차트
- Confusion matrix 히트맵
- Fairness metric 대시보드
- ROC curve 비교

실습 포인트

다양한 인구 집단에 대해 TPR, FPR, PPV를 계산하고 격차가 허용 범위(예: 10%) 내인지 확인

윤리 위원회 (Ethics Committees)

위원회 구성

- 의사, 간호사 등 의료진
- AI/데이터 과학자
- 생명윤리학자
- 법률 전문가
- 환자 대표
- 사회학자, 인류학자

주요 역할

- AI 시스템 윤리 검토
- 연구 프로토콜 승인
- 윤리적 딜레마 해결
- 정책 및 가이드라인 수립
- 사후 모니터링
- 교육 및 인식 제고

⚖️ 의사결정 프로세스

신청 접수
AI 시스템 제안서

→

초기 검토
서류 및 문서 확인

→

전체 회의
다학제 논의

→

결정 및 모니터링
승인/조건부/거부

승인
윤리적 문제 없음

조건부 승인
수정 후 재검토

거부
윤리적 문제 심각

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?