AI 윤리 리스크 진단 보고서

진단 대상: gemini
진단 일시: 2025-10-23 15:56

Ⅰ. 서비스 개요

서비스 유형: 생성형 AI
주요 목적: Gemini 요약

기술적 구조:
• Gemini는 구글 딥마인드가 개발한 차세대 인공지능 시스템으로, 딥러닝, 강화 학습, 대규모 데이터 처리를 통합하여 멀티모달 기능을 갖춘 대규모 언어 모델(LLM)입니다. 텍스트, 이미지, 오디오, 비디오 및 코드 전반을 처리하고 추론할 수 있도록 설계되었습니다.

서비스 목적:
• Gemini는 실시간 대화 지원, 동영상 요약, 로봇 제어, 의료 진단 지원 등 다양한 분야에서 활용되며, 사용자 맞춤형 응답을 제공하여 생산성과 창의성을 높이는 것을 목표로 합니다. 또한, Google 생태계 내에서 통합된 사용자 경험을 제공합니다.

주요 기능:
• 텍스트 생성 및 추론, 멀티모달 데이터 처리, 실시간 상호작용, 개인 맞춤형 지원, 다양한 기기에서의 접근성.
• Med-Gemini 모델을 통한 의료 맞춤형 솔루션 제공 및 Gemini Robotics를 통한 물리적 작업 지원.
• 개발자는 Google Cloud의 Vertex AI 플랫폼을 통해 Gemini 모델을 활용할 수 있으며, API를 통해 다양한 AI 모델에 접근할 수 있습니다.
• Gemini는 사용자 데이터를 기반으로 응답을 맞춤화하고, 다양한 플랫폼에서의 통합적 기능을 제공합니다.

Ⅱ. 초기 윤리 리스크 평가

점수: 5.0점
점수: 5/5

Ⅲ. 사용자 피드백

threh

Ⅳ. 피드백 반영 후 재평가 결과

공정성 (Fairness): 4.0점
공정성 (Fairness): 4/5

편향성 (Bias): 3.0점
편향성 (Bias): 3/5

투명성 (Transparency): 4.0점
투명성 (Transparency): 4/5

설명가능성 (Explainability): 3.0점
설명가능성 (Explainability): 3/5

프라이버시 (Privacy): 5.0점
프라이버시 (Privacy): 5/5

Ⅴ. 최종 개선 권고안

AI 서비스의 윤리 리스크 평가 결과를 바탕으로 각 항목별로 구체적인 개선 권고안을 제시하겠습니다. 각 권고안은 EU, OECD, UNESCO의 가이드라인과 연결하여 설명하겠습니다.

1. 공정성 (Fairness) - Score: 4.0
개선 권고안:
• 다양한 데이터 소스 활용: 공정성을 높이기 위해 다양한 인구 통계적 배경을 가진 데이터를 수집하고 활용하여 AI 모델의 훈련을 진행해야 합니다. 이는 OECD의 "AI Principles"에서 강조하는 "공정한 접근"의 원칙과 일치합니다.
• 정기적인 공정성 감사: AI 시스템의 공정성을 정기적으로 평가하고 감사하는 프로세스를 도입하여, 시스템이 특정 그룹에 불리하게 작용하지 않도록 지속적으로 모니터링해야 합니다.

2. 편향성 (Bias) - Score: 3.0
개선 권고안:
• 편향성 검토 및 수정: AI 모델의 학습 데이터를 분석하여 잠재적인 편향성을 식별하고, 이를 수정하기 위한 알고리즘 개선을 진행해야 합니다. UNESCO의 "AI Ethics" 가이드라인에서도 편향 제거의 중요성을 강조하고 있습니다.
• 다양한 이해관계자 참여: 다양한 이해관계자(예: 사회적 소수자, 전문가 등)를 포함한 포커스 그룹을 구성하여 AI 시스템의 편향성을 평가하고, 그들의 피드백을 반영하여 개선할 수 있는 방안을 마련해야 합니다.

3. 투명성 (Transparency) - Score: 4.0
개선 권고안:
• 알고리즘 설명서 제공: AI 시스템의 작동 원리를 설명하는 문서를 제공하여 사용자와 이해관계자가 시스템의 결정 과정을 이해할 수 있도록 해야 합니다. 이는 EU의 AI 법안에서 요구하는 투명성 원칙과 일치합니다.
• 결과 설명 기능 강화: AI의 결과를 사용자에게 설명할 수 있는 기능을 추가하여, 사용자가 시스템의 결정을 이해하고 신뢰할 수 있도록 해야 합니다.

4. 설명가능성 (Explainability) - Score: 3.0
개선 권고안:
• 설명 가능한 AI 모델 개발: AI 모델을 설계할 때 설명 가능성을 고려하여, 결과를 쉽게 해석할 수 있는 모델을 개발해야 합니다. OECD의 "AI Principles"에서 설명 가능성의 중요성을 강조하고 있습니다.
• 사용자 교육 프로그램: 사용자에게 AI 시스템의 결과를 해석하는 방법에 대한 교육 프로그램을 제공하여, 사용자가 AI의 결정을 이해할 수 있도록 지원해야 합니다.

5. 프라이버시 (Privacy) - Score: 5.0
개선 권고안:
• 데이터 보호 강화: 개인 데이터를 수집하고 처리할 때 GDPR(일반 데이터 보호 규정)을 준수하여, 개인의 프라이버시를 철저히 보호해야 합니다. 이는 EU의 데이터 보호 원칙과 일치합니다.

• 최소한의 데이터 수집 원칙: 필요한 최소한의 데이터만을 수집하고, 사용 후 즉시 삭제하는 정책을 수립하여 개인의 프라이버시를 더욱 강화해야 합니다.

이러한 개선 권고안들은 AI 서비스의 윤리적 리스크를 줄이고, 사용자와 사회에 긍정적인 영향을 미칠 수 있도록 도와줄 것입니다. 각 권고안은 관련 가이드라인의 원칙을 준수하며, 지속 가능한 AI 시스템 개발에 기여할 것입니다.

※ 본 보고서는 Human-in-the-loop 기반 AI 윤리 평가 결과입니다.