# AI 윤리 리스크 진단 보고서

진단 대상: gemini

진단 일시: 2025-10-22 17:13

# 1. 서비스 개요

서비스 유형: 생성형 AI

주요 목적: Gemini 요약

기술적 구조:
Gemini는 구글의 최신 대규모 언어 모델(LLM)을 기반으로 한 멀티모달 AI 인터페이스로, 텍스트, 오디오, 이미지 등 다양한 형식을 처리할 수 있습니다. 이 모델은 2013년 Word2Vec 논문에서 시작된 혁신적인 아키텍처를 기반으로 하며, 2017년 Transformer 기술의 발전과 2020년 다중 대화 기능을 포함한 여러 연구 성과를 통해 발전해왔습니다.

서비스 목적:
Gemini의 주된 목적은 사용자에게 개인화된 AI 비서 역할을 하여 정보 접근성과 활용성을 높이는 것입니다. 사용자는 Gemini를 통해 이메일 작성, 코드 디버깅, 아이디어 브레인스토밍 등 다양한 작업을 지원받을 수 있습니다.

주요 기능:
1. 생산성 향상: 긴 문서 요약, 코드 작업 지원 등으로 시간을 절약할 수 있습니다.
2. 창의성 촉진: 블로그 포스트 아웃라인 생성, 이미지 생성 등을 통해 아이디어를 실현할 수 있습니다.
3. 정보 탐색: 복잡한 개념 설명, 주제에 대한 통찰 제공, 웹 콘텐츠 추천 기능이 곧 추가될 예정입니다.
4. 사용자 맞춤화: 특정 지침을 통해 Gemini를 개인의 목표 달성에 맞게 조정할 수 있는 기능이 개발 중입니다.
5. 안전성과 책임: Gemini는 사용자 피드백과 연구를 통해 지속적으로 개선되며, 개인 정보 보호와 정책 가이드라인을 준수합니다.

Gemini는 사용자와의 상호작용을 통해 지속적으로 발전하며, AI의 책임 있는 사용을 위해 다양한 전문가와 협력하고 있습니다.

# 2. 윤리 리스크 평가 (최종)


{
"Summary": {
"score": 0,
"comment": "각 항목에 대한 평가와 코멘트는 다음과 같습니다:₩n₩n1. 공정성 (Fairness): 4점₩n - 코멘트: AI 시스템이 개인 데이터의 처리와 관련하여 공정성을 보장하기 위한 조치가 명시되어 있지만, 특정 조건 하에 공정성이 저해될 수 있는 가능성이 여전히 존재합니다. 특히, 저작권 문제와 관련하여 공정한 사용의 기준이 명확히 정의되어야 합니다.₩n₩n2. 편향성 (Bias): 3점₩n - 코멘트: 편향성을 교정하기 위한 조치가 언급되었지만, 편향성을 완전히 제거하기 위한 구체적인 방법이나 기준이 부족합니다. AI 모델이 학습하는 데이터의 출처와 품질이 편향성에 큰 영향을 미치므로, 이 부분에 대한 추가적인 고려가 필요합니다.₩n₩n3. 투명성 (Transparency): 4점₩n - 코멘트: AI가 생성한 콘텐츠에 대한 투명성 의무가 명시되어 있으나, 실제로 이 의무가 어떻게 이행될지는 불확실합니다. 특히, 저작권 보호 콘텐츠에 대한 처리 방식이 명확히 규정되어야 합니다.₩n₩n4. 설명가능성 (Explainability): 3점₩n - 코멘트: AI 시스템의 결정 과정에 대한 설명 가능성은 중요하지만, 현재의 가이드라인에서는 이에 대한 구체적인 요구 사항이 부족합니다. 사용자와 이해관계자가 AI의 작동 방식을 이해할 수 있도록 하는 데 더 많은 노력이 필요합니다.₩n₩n5. 프라이버시 (Privacy): 5점₩n - 코멘트: 개인 데이터 보호를 위한 엄격한 보안 조치와 접근 통제가 명시되어 있어 프라이버시 보호에 대한 강력한 기준이 마련되어 있습니다. 그러나, 데이터의 수집과 사용에 대한 투명성이 보장되어야 하며, 사용자의 동의가 명확히 반영되어야 합니다.₩n₩n종합적으로, AI 윤리 가이드라인은 여러 중요한 요소를 다루고 있지만, 특히 편향성과 설명가능성에 대한 구체적인 기준과 절차가 더 필요합니다."
}
}


# 3. 종합 개선 권고안

AI 서비스의 윤리 리스크 평가 결과에 대한 구체적인 개선 권고안을 아래와 같이 제시합니다. 각 권고안은 EU, OECD, UNESCO 가이드라인의 원칙과 연결하여 설명합니다.

### 1. 공정성 (Fairness)
개선 권고안:
• 저작권 기준 명확화: 공정한 사용 기준을 명확히 정의하고 이를 문서화하여 모든 이해관계자가 이해할 수 있도록 합니다. 이를 통해 AI 시스템이 저작권 문제를 처리할 때 공정성을 보장할 수 있습니다.
• EU 가이드라인 연결: EU의 Directive (EU) 2019/790은 저작권 및 관련 권리에 대한 예외와 제한을 도입하여 공정한 사용을 촉진합니다. 이와 같은 규정을 준수하여 AI 시스템의 공정성을 강화해야 합니다.

### 2. 편향성 (Bias)
개선 권고안:
• 데이터 출처 및 품질 관리: AI 모델이 학습하는 데이터의 출처와 품질을 철저히 검토하고, 편향성을 줄이기 위한 구체적인 기준과 절차를 마련합니다. 예를 들어, 다양한 출처에서 데이터를 수집하고, 데이터 세트의 다양성을 보장하는 방안을 모색해야 합니다.
• OECD 가이드라인 연결: OECD의 AI 원칙은 데이터의 품질과 다양성을 강조합니다. 따라서 편향성을 줄이기 위한 조치로 데이터 수집 및 처리 과정에서 OECD의 지침을 따르는 것이 중요합니다.

### 3. 투명성 (Transparency)
개선 권고안:
• 투명성 이행 계획 수립: AI가 생성한 콘텐츠에 대한 투명성 의무를 이행하기 위한 구체적인 계획을 수립합니다. 예를 들어, AI의 결정 과정과 데이터 사용에 대한 정보를 사용자에게 제공하는 시스템을 구축해야 합니다.
• UNESCO 가이드라인 연결: UNESCO의 AI 윤리 가이드라인은 투명성을 강조하며, AI 시스템의 작동 방식에 대한 정보 제공을 권장합니다. 이를 통해 사용자와 이해관계자가 AI의 작동 방식을 이해할 수 있도록 해야 합니다.

### 4. 설명가능성 (Explainability)
개선 권고안:
• 설명가능성 기준 마련: AI 시스템의 결정 과정에 대한 설명 가능성을 높이기 위해 구체적인 기준과 요구 사항을 설정합니다. 이를 통해 사용자와 이해관계자가 AI의 작동 방식을 더 잘 이해할 수 있도록 해야 합니다.
• EU 가이드라인 연결: EU의 AI 법안은 설명가능성을 중요한 요소로 강조하고 있습니다. 따라서 AI 시스템이 내린 결정에 대한 명확한 설명을 제공하는 것이 법적 요구사항을 충족하는 데 필수적입니다.

### 5. 프라이버시 (Privacy)
개선 권고안:
• 데이터 수집 및 사용의 투명성 강화: 개인 데이터의 수집과 사용에 대한 정보를 사용자에게 명확히 제공하고, 사용자의 동의를 명확히 반영하는 절차를 마련합니다. 이를 통해 프라이버시 보호를 강화할 수 있습니다.
• OECD 가이드라인 연결: OECD의 개인정보 보호 원칙은 데이터 수집 및 사용의 투명성을 강조합니다. 따라서 이러한 원칙을 준수하여 사용자에게 명확한 정보 제공을 통해 프라이버시를 보호해야 합니다.

이러한 권고안들은 AI 서비스의 윤리적 리스크를 줄이고, 사용자와 사회의 신뢰를 구축하는 데 기여할 것입니다. 각 권고안은 국제 가이드라인과 연결되어 있어, 글로벌 스탠다드를 준수하는 데 도움이 됩니다.