AI 윤리성 리스크 진단 보고서

전문적 평가 및 개선 권고

분석 대상 서비스	Midjourney, Copilot, Google Gemini
평가 기준	EU AI Act, UNESCO AI Ethics, OECD AI Principles
평가 차원	공정성, 프라이버시, 투명성, 책임성, 안전성
작성일	2025년 10월 23일
평가 시간	2025-10-23 10:59:45

평가 목표

본 보고서는 선정된 AI 서비스에 대하여 국제 표준 가이드라인을 기준으로 윤리적 리스크를 종합적으로 평가하고, 각 서비스의 강점을 파악하며 필요한 개선사항을 구체적으로 제시하는 것을 목표로 합니다.

목차

EXECUTIVE SUMMARY

1. 평가 개요

본 평가는 3개 AI 서비스에 대하여 EU AI Act, UNESCO AI Ethics, OECD AI Principles 등 국제 표준을 기준으로 5개 차원(공정성, 프라이버시, 투명성, 책임성, 안전성)에서 윤리적 리스크를 평가했습니다.

2. 종합 평가 결과

서비스	종합점수	리스크수준	등급
Midjourney	2.8/5	중간	D
Copilot	3.2/5	중간	C
Google Gemini	3.4/5	중간	C

3. 주요 발견사항

- 평균 윤리 점수: 3.1/5
- 리스크 수준: 중간
- 분석 대상: 3개 서비스
- 평가 차원: 5개 (공정성, 프라이버시, 투명성, 책임성, 안전성)

평가 방법론

1. 평가 프레임워크

본 평가는 다음의 국제 표준을 기준으로 실시되었습니다: EU AI Act (유럽 인공지능 규정) - 고위험 AI 시스템에 대한 엄격한 규제 - 편향성 테스트 및 완화 조치 요구 - 투명성 및 설명가능성 의무화 UNESCO AI Ethics Recommendations - 인간 중심의 윤리 원칙 - 다양성과 포용성 강조 - 개인정보 자기결정권 보장 OECD AI Principles - 포용적 성장 및 지속가능한 발전 - 인권과 민주적 가치 존중 - 견고한 AI 시스템 구축

2. 평가 차원 (5개)

차원	설명	주요 평가항목
공정성	편향성 없이 공정하게 작동	편향 테스트, 성능 동등성, 완화 조치
프라이버시	개인정보 보호 및 관리	정책, 암호화, 동의, 데이터 삭제
투명성	작동 방식의 명확성	AI 사용 명시, 설명가능성, 데이터 출처
책임성	책임 소재 및 거버넌스	책임자 지정, 감사, 사고 대응
안전성	안전성 및 보안 수준	위험 평가, 견고성, 보안 조치

3. 평가 등급 정의

등급	점수	정의	설명
A+	4.8~5.0	모범 사례	모든 가이드라인 완벽 준수
A	4.5~4.7	우수	대부분 준수, 미미한 개선
B+	4.2~4.4	양호	기본 요구 충족, 개선 필요
B	3.8~4.1	보통	기본 요구 부분 충족
C	3.0~3.7	미흡	여러 영역 개선 필요
D	2.0~2.9	부족	심각한 결함
F	1.0~1.9	위험	즉각적 개선 필수

서비스 분석: Midjourney

1. 종합 평가

종합 점수: 2.8/5
평가 등급: D
리스크 수준: 중간
평가 설명: N/A

2. 차원별 상세 평가

2.1 공정성 및 편향성

점수: 2/5 | 리스크: 높음 | 등급: D

Midjourney demonstrates a commitment to addressing biases in its AI-generated content, yet significant issues remain. The platform has acknowledged the presence of racial and gender biases in its outputs, which stem from the training data used. While there are efforts to improve bias detection and mitigation, the lack of detailed public documentation on bias testing and mitigation strategies limits transparency. The service does not provide comprehensive information on the diversity of its training datasets or the specific measures taken to ensure equal performance across different demographic groups. The absence of a detailed audit system or external review process further complicates the assessment of its fairness practices. Although Midjourney engages with its user community for feedback, this alone does not suffice to meet the rigorous fairness standards set by guidelines like the EU AI Act.

주요 증거:

- Midjourney has faced criticism for racial and gender bias in generated images (Medium, 'When AI Mirrors Our Flaws: Unveiling Bias in MidJourney').
- The platform acknowledges biases and is working on improving bias detection (Service Analysis).
- There is no detailed public information on internal or external audit processes for ethical compliance (Service Analysis).

강점:

Engagement with user community for feedback.

Commitment to improving bias detection and mitigation.

발견된 리스크:

Racial and gender bias in AI-generated images.

Lack of transparency in bias mitigation strategies.

2.2 프라이버시 보호

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney demonstrates a reasonable level of privacy protection by ensuring compliance with GDPR and CCPA, providing users with the ability to request and delete personal data. However, there are areas for improvement, such as the lack of detailed public information on data protection impact assessments and the handling of user data beyond basic compliance. The platform's privacy policy outlines user rights under GDPR, but there is no evidence of a comprehensive data protection strategy that includes privacy by design or default. Additionally, there have been complaints regarding the accessibility of privacy features, indicating potential gaps in user experience and transparency.

주요 증거:

- Privacy Controls in Midjourney - Titan Extension Tools: Midjourney ensures compliance with GDPR and CCPA, allowing users to request and delete personal data.
- Privacy Policy - Midjourney: The policy outlines user rights under GDPR, including data correction, deletion, and limitation.
- Why I filed a GDPR complaint against Midjourney - Tim Boucher: Complaints about the accessibility of privacy features and lack of response from Midjourney staff.

강점:

Compliance with GDPR and CCPA, allowing user data requests and deletions.

Clear privacy policy outlining user rights under GDPR.

발견된 리스크:

Lack of comprehensive data protection impact assessments.

User complaints about accessibility and response to privacy concerns.

2.3 투명성 및 설명가능성

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney demonstrates a moderate level of transparency in its AI operations. While it provides some insights into its model capabilities and has mechanisms for user feedback, it lacks detailed public documentation on its decision-making processes and specific parameter details. The platform acknowledges biases in its outputs and is committed to addressing them, but the absence of comprehensive ethical guidelines and detailed transparency about its algorithms and data usage limits its overall transparency. The use of a Discord-based interface for interaction is unique but does not inherently enhance transparency regarding the AI's internal workings.

주요 증거:

- Midjourney maintains a moderate level of transparency, providing some insights into its model capabilities but lacking detailed ethical guidelines. (Source: Service Analysis)
- Instances of racial and gender bias in generated images have been reported, reflecting broader issues in AI training data. (Source: Medium article on AI bias)
- Midjourney has a Prompt Analyzer tool available for explainability, invoked via the '/shorten' command. (Source: AI Vendor Risk Profile)

강점:

Engagement with users for feedback and improvement.

Availability of a tool for explainability (Prompt Analyzer).

발견된 리스크:

Lack of detailed transparency in decision-making processes.

Reported biases in AI-generated content.

2.4 책임성 및 거버넌스

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney demonstrates a moderate level of accountability in its AI service. While it acknowledges biases in its outputs and has mechanisms for user feedback to address these issues, it lacks a detailed public framework for ethical guidelines and accountability measures. The absence of a dedicated AI ethics team and a comprehensive audit system for ethical compliance indicates a need for improvement. Although Midjourney engages with its community to address ethical concerns, the lack of transparency regarding its internal governance and regulatory compliance limits its accountability. Furthermore, known issues such as racial and gender bias in generated images highlight the need for more robust bias detection and mitigation strategies.

주요 증거:

- Midjourney has acknowledged biases in AI-generated content and is working on improving bias detection and mitigation (source: service analysis).
- There is no detailed public information on internal or external audit processes for ethical compliance (source: governance section).
- Instances of racial and gender bias in generated images have been reported (source: ethics aspects).

강점:

User feedback mechanisms to address biases

Engagement with the community to improve ethical standards

발견된 리스크:

Racial and gender bias in generated images

Lack of detailed ethical guidelines and accountability framework

2.5 안전성 및 보안

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney's safety and security measures show a moderate level of compliance with established guidelines. The platform employs user feedback mechanisms and ongoing model improvements to address biases and discriminatory outputs, which aligns with efforts to ensure safety and robustness. However, there is a lack of detailed public documentation on specific cybersecurity measures and quality management systems, which are critical for high-risk AI systems as per the EU AI Act. The presence of biases, particularly racial and gender biases, indicates potential vulnerabilities that could

compromise the system's safety and fairness. While Midjourney is committed to addressing these issues, the absence of detailed audit systems and comprehensive ethical guidelines suggests room for improvement.

주요 증거:

- Midjourney actively works on improving bias detection and mitigation in its models (source: service analysis).
- Instances of racial and gender bias in generated images have been reported (source: Medium article on bias in Midjourney).
- There is no detailed public information on internal or external audit processes for ethical compliance (source: service analysis).

강점:

Active user feedback mechanisms for bias reporting.

Ongoing model improvements to reduce discriminatory outputs.

발견된 리스크:

Racial and gender biases in AI-generated content.

Lack of detailed cybersecurity and quality management documentation.

3. 가이드라인 준수 현황

준수 기준에 따라 각 가이드라인에 대한 준수 여부를 평가했습니다.

서비스 분석: Copilot

1. 종합 평가

종합 점수: 3.2/5
평가 등급: C
리스크 수준: 중간
평가 설명: N/A

2. 차원별 상세 평가

2.1 공정성 및 편향성

점수: 2/5 | 리스크: 높음 | 등급: D

Microsoft Copilot demonstrates several efforts towards fairness, such as employing red teaming exercises and continuous updates to address bias. However, there are significant concerns about the persistent bias in AI outputs, particularly in workplace contexts. Reports indicate that Copilot struggles with bias detection and mitigation, which can lead to discriminatory outcomes in professional settings. Despite Microsoft's commitment to responsible AI principles, the system's reliance on GPT-4 and its integration with enterprise applications exacerbate these challenges. The lack of comprehensive bias testing results and limited transparency about specific bias characteristics further complicate the evaluation of fairness. While Microsoft has established ethical guidelines and policies, the effectiveness of these measures in practice remains questionable.

주요 증거:

- Microsoft Copilot: Big AI Fixes, Same Old AI Bias - Reports indicate systemic bias issues across social categories.
- Investigating Bias in Generative AI Systems - Copilot performs poorly in bias detection tasks, especially in workplace contexts.
- Data, Privacy, and Security for Microsoft 365 Copilot - Acknowledgement of bias risks and ongoing efforts to address them.

강점:

Commitment to addressing bias through continuous updates and red teaming exercises.

Established ethical guidelines focusing on fairness, transparency, and accountability.

발견된 리스크:

Persistent bias in AI outputs, particularly in workplace contexts.

Limited success in generating inclusive content, leading to potential discriminatory outcomes.

2.2 프라이버시 보호

점수: 4/5 | 리스크: 낮음 | 등급: B

Microsoft Copilot demonstrates a strong commitment to privacy protection, aligning with key privacy regulations such as the GDPR. The service ensures compliance with GDPR, as evidenced by its adherence to data minimization principles and purpose limitation. Microsoft explicitly states that prompts, responses, and data accessed through Microsoft Graph are not used to train foundational LLMs, which supports data protection. Additionally, Copilot offers broad compliance offerings and certifications, including ISO standards, which further underscore its dedication to privacy. However, there are potential risks concerning the misuse of sensitive data and the integration with third-party tools, which could lead to data leaks if not properly managed. Despite these concerns, Microsoft's proactive approach in conducting Data Protection Impact Assessments (DPIAs) and engaging in regular audits indicates a robust privacy management framework.

주요 증거:

- Microsoft 365 Copilot is compliant with GDPR and provides broad compliance offerings and certifications, including ISO 27001 and HIPAA (source: Microsoft documentation).
- Prompts, responses, and data accessed through Microsoft Graph aren't used to train foundation LLMs, ensuring data protection (source: Microsoft privacy documentation).
- Microsoft Copilot explicitly mentions compliance with existing privacy and security regulations, including GDPR (source: Securiti).

강점:

Strong compliance with GDPR and other international privacy standards.

Regular audits and data protection impact assessments to ensure ongoing privacy protection.

발견된 리스크:

Potential misuse of sensitive data due to integration with third-party tools.

Risk of data leaks if proper data privacy controls are not in place.

2.3 투명성 및 설명가능성

점수: 3/5 | 리스크: 중간 | 등급: C

Microsoft Copilot demonstrates a moderate level of transparency in its AI operations. While Microsoft provides technical documentation and user guidelines, there are challenges in fully disclosing the intricacies of the AI model, such as the decision-making logic and detailed workings of the underlying GPT-4 architecture. The service adheres to GDPR, indicating a commitment to data privacy and security, but known issues such as bias in AI outputs and the potential for misleading information highlight areas where transparency could be improved. Microsoft's efforts to address bias through updates and red teaming exercises are positive steps, yet the lack of comprehensive public transparency about specific bias characteristics and decision-making processes limits the overall transparency score.

주요 증거:

- Microsoft provides detailed ethical guidelines focusing on fairness, transparency, and accountability in AI deployment (source: service analysis).

• Known issues include persistent bias in AI outputs and concerns about accuracy (source: service analysis).

• Microsoft maintains a moderate level of transparency, providing technical documentation and user guidelines (source: service analysis).

강점:

Commitment to data privacy and security, with GDPR compliance.

Efforts to address bias through updates and red teaming exercises.

발견된 리스크:

Persistent bias in AI outputs.

Potential for misleading information in AI-generated content.

2.4 책임성 및 거버넌스

점수: 4/5 | 리스크: 낮음 | 등급: B

Microsoft Copilot demonstrates a strong commitment to accountability and governance, aligning with many aspects of the EU AI Act, UNESCO, and OECD guidelines. The service has a clear governance structure, with dedicated teams for AI ethics and governance, and conducts both internal and external audits to ensure compliance with ethical standards. Microsoft also engages with external experts and incorporates user feedback to refine AI functionalities, which supports accountability. However, there are known issues with bias in AI outputs, which Microsoft is actively addressing through continuous updates and red teaming exercises. The transparency level is moderate, with technical documentation and user guidelines available, but challenges remain in fully disclosing AI model intricacies.

주요 증거:

• Microsoft provides detailed ethical guidelines focusing on fairness, transparency, and accountability in AI deployment (source: service analysis).

• Both internal and external audits are conducted to ensure compliance with ethical standards (source: service analysis).

• Microsoft engages with external experts and incorporates user feedback to refine AI functionalities (source: service analysis).

강점:

Strong governance structure with dedicated AI ethics teams.

Engagement with external experts and user feedback for continuous improvement.

발견된 리스크:

Persistent bias issues in AI outputs.

Challenges in fully disclosing AI model intricacies.

2.5 안전성 및 보안

점수: 3/5 | 리스크: 중간 | 등급: C

Microsoft Copilot demonstrates a moderate level of safety and security, adhering to several industry standards and regulations such as GDPR. However, it faces challenges in addressing bias and ensuring the accuracy of AI-generated content. Known issues include persistent bias in AI outputs and vulnerabilities like the 'EchoLeak' incident, which highlight potential security risks. Microsoft has implemented regular updates and security patches to mitigate these risks, but the effectiveness of these measures remains partially uncertain. The service's reliance on large language models, such as GPT-4, introduces complexities in maintaining consistent safety standards, particularly in high-risk environments like finance and healthcare.

주요 증거:

- Microsoft provides detailed ethical guidelines focusing on fairness, transparency, and accountability in AI deployment (source: service analysis).
- Reports indicate persistent bias issues in AI outputs, particularly in workplace contexts (source: service analysis).
- A recently discovered security flaw in Microsoft 365 Copilot, dubbed 'EchoLeak,' underscores vulnerabilities in AI agents (source: Director's Cut: Microsoft Copilot Flaw Highlights Emerging AI - Zscaler).

강점:

Compliance with GDPR and other industry standards.

Regular updates and security patches to improve safety.

발견된 리스크:

Persistent bias in AI outputs.

Security vulnerabilities such as 'EchoLeak'.

3. 가이드라인 준수 현황

준수 기준에 따라 각 가이드라인에 대한 준수 여부를 평가했습니다.

서비스 분석: Google Gemini

1. 종합 평가

종합 점수: 3.4/5
평가 등급: C
리스크 수준: 중간
평가 설명: N/A

2. 차원별 상세 평가

2.1 공정성 및 편향성

점수: 3/5 | 리스크: 중간 | 등급: C

Google Gemini AI demonstrates a moderate level of fairness in its operations. The service has implemented some fairness benchmarks and subgroup analyses to address biases related to gender, race, ethnicity, and religion. However, these analyses are primarily conducted on American English language data, which limits the scope of its fairness evaluation across diverse populations. Known issues, such as bias in image generation and discrimination against African American Vernacular English (AAVE) users, indicate areas where the system's fairness could be improved. While Google has taken steps to enhance transparency and implement bias mitigation mechanisms, the effectiveness of these measures is not fully documented or disclosed. The service's efforts to address fairness are ongoing, but there is room for improvement, particularly in ensuring equal performance across various demographic groups and expanding the scope of bias testing.

주요 증거:

- Google Gemini's fairness analyses focus on American English data, limiting broader applicability (source: 'Gemini for Google Cloud and responsible AI').
- Bias issues in image generation led to offline testing for further evaluation (source: 'How to drive bias out of AI without making mistakes of Google Gemini').
- Discrimination against AAVE users highlights existing bias concerns (source: 'Unmasking Racism in AI: From Gemini's Overcorrection to AAVE ...').

강점:

Implementation of fairness benchmarks and subgroup analyses.

Efforts to enhance transparency and bias mitigation mechanisms.

발견된 리스크:

Bias in image generation leading to potential discrimination.

Limited fairness testing scope, primarily on American English data.

2.2 프라이버시 보호

점수: 4/5 | 리스크: 낮음 | 등급: B

Google Gemini demonstrates a strong commitment to privacy protection, adhering to key privacy regulations such as GDPR and HIPAA. The service includes comprehensive privacy policies, data residency controls, and certifications like ISO/IEC 27001 and SOC 2, which are critical for ensuring data protection and privacy. Google has implemented robust data governance frameworks and privacy safeguards, ensuring compliance with international privacy standards. However, there are areas for improvement, such as ensuring transparency in data processing and addressing known biases in AI outputs. Despite these minor issues, Google Gemini's privacy measures are largely effective and align with best practices.

주요 증거:

- Google Gemini's compliance with GDPR and HIPAA is documented in 'Google Gemini: GDPR, HIPAA, and enterprise compliance ...' (source: datastudios.org)
- The privacy policy outlines data retention and protection measures, as seen in 'Privacy Policy | Gemini' (source: gemini.com)
- Google's commitment to privacy and security certifications, including ISO/IEC 27001 and SOC 2, is detailed in 'Understanding Google Gemini Compliance, Certifications ...' (source: promevo.com)

강점:

Comprehensive compliance with GDPR and HIPAA, ensuring robust data protection.

Strong privacy and security certifications, including ISO/IEC 27001 and SOC 2.

발견된 리스크:

Potential biases in AI outputs, particularly concerning image generation.

Limited transparency in data processing and decision-making processes.

2.3 투명성 및 설명가능성

점수: 3/5 | 리스크: 중간 | 등급: C

Google Gemini AI demonstrates a moderate level of transparency, fulfilling some basic requirements but leaving room for improvement. The service provides some information on its neural network architecture and its capabilities, such as text generation and multimodal processing. However, there is limited disclosure regarding the specific decision-making processes and the underlying logic of the AI system. While Google has made efforts to address bias and fairness, as evidenced by the removal of the image generation feature for further testing, the transparency regarding these issues is not fully comprehensive. The service's transparency level is described as 'medium,' indicating that while some ethical guidelines and analyses are shared, there is a need for more detailed technical documentation and clearer explanations of how decisions are made.

주요 증거:

- Google Gemini AI: a Guide to 9 Remarkable Key Features - Describes the capabilities and neural network architecture (source: ai-scaleup.com)
- Introducing Gemini: our largest and most capable AI model - Discusses ongoing efforts to improve transparency and address bias (source: blog.google)

• How to drive bias out of AI without making mistakes of Google Gemini - Highlights issues with bias and the need for transparency (source: cnbc.com)

강점:

Efforts to address bias and fairness

Integration of multimodal capabilities

발견된 리스크:

Limited disclosure of decision-making processes

Bias issues in image generation and AAVE discrimination

2.4 책임성 및 거버넌스

점수: 4/5 | 리스크: 낮음 | 등급: B

Google Gemini demonstrates a strong commitment to accountability and governance, aligning with several key aspects of the EU AI Act, UNESCO, and OECD guidelines. The service is governed by Google's AI ethics team, which ensures compliance with AI ethics and fairness laws. Google Gemini has a robust internal and external audit system to monitor compliance and performance. The service also engages stakeholders through feedback loops, enhancing its accountability framework. However, some areas require improvement, such as transparency in the model's decision-making processes and addressing known bias issues, particularly with AAVE users and image generation biases. Despite these challenges, Google Gemini's efforts in bias detection and mitigation, along with its compliance with major data protection regulations like GDPR and HIPAA, underscore its strong governance framework.

주요 증거:

• Google Gemini's compliance with GDPR and HIPAA is detailed in their privacy and compliance documentation (source: https://www.datastudios.org/post/google-gemini-gdpr-hipaa-and-enterprise-compliance-standards-explained).

• The presence of internal and external audit processes is confirmed by Google's responsible AI documentation (source: https://cloud.google.com/gemini/docs/discover/responsible-ai).

• Efforts to mitigate bias and enhance fairness are highlighted in Google's AI ethics policies (source: https://www.cnbc.com/2024/03/27/how-to-drive-bias-out-of-ai-without-making-mistakes-of-google-gemini.html).

강점:

Strong compliance with GDPR and HIPAA.

Robust internal and external audit systems.

발견된 리스크:

Bias in image generation and AAVE user interactions.

Limited transparency in decision-making processes.

2.5 안전성 및 보안

점수: 3/5 | 리스크: 중간 | 등급: C

Google Gemini AI demonstrates a moderate level of safety and security compliance. The service has implemented various safety measures such as bias detection and mitigation mechanisms, and it adheres to data protection regulations like GDPR and HIPAA. However, there are notable security vulnerabilities, as evidenced by the red teaming study revealing significant security gaps across its multimodal capabilities. These vulnerabilities, if exploited, could lead to privacy risks and data theft. Furthermore, the service has faced issues with bias, particularly in image generation and language processing, which indicates a need for further improvement in fairness and inclusivity. While Google has a robust compliance framework and conducts internal and external audits, the presence of these vulnerabilities and bias issues suggests that more rigorous safety and security measures are necessary.

주요 증거:

- Uncovering Safety Gaps in Gemini: A Multimodal Red Teaming Study
- Researchers Disclose Google Gemini AI Flaws Allowing Prompt ...
- Google Gemini: GDPR, HIPAA, and enterprise compliance ...

강점:

Compliance with GDPR and HIPAA

Efforts in bias detection and mitigation

발견된 리스크:

Security vulnerabilities in multimodal capabilities

Bias in image generation and language processing

3. 가이드라인 준수 현황

준수 기준에 따라 각 가이드라인에 대한 준수 여부를 평가했습니다.

서비스 비교 분석

1. 종합 순위

순위	서비스	점수	등급
1	Google Gemini	3.4/5	C
2	Copilot	3.2/5	C
3	Midjourney	2.8/5	D

2. 차원별 점수 비교

차원	Midjourney	Copilot	Google Gemini
공정성	2/5	2/5	3/5
프라이버시	3/5	4/5	4/5
투명성	3/5	3/5	3/5
책임성	3/5	4/5	4/5
안전성	3/5	3/5	3/5

종합 권고사항

1. 단기 조치 (1-3개월)

- AI 윤리 정책 문서화 및 공개
- 편향성 테스트 프레임워크 도입
- 투명성 강화 계획 수립

2. 중기 조치 (3-6개월)

- AI 거버넌스 체계 구축
- 정기적인 윤리 감시 시스템 실시
- 투명성 보고서 발행

3. 장기 조치 (6개월 이상)

- 지속적인 모니터링 시스템 구축
- 외부 독립 감사 체계 확립
- 산업 표준 및 인증 획득

참고문헌 (REFERENCE)

A. 국제 가이드라인 및 표준

[1] European Commission (2021). 'Proposal for a Regulation on Artificial Intelligence (AI Act)'. Brussels.

[2] UNESCO (2021). 'Recommendation on the Ethics of Artificial Intelligence'. Paris.

[3] OECD (2019). 'OECD AI Principles'. Paris.

[4] NIST (2023). 'AI Risk Management Framework'. National Institute of Standards and Technology.

B. 평가 방법론 및 도구

[5] LLM 기반 정성 평가: GPT 모델을 활용한 다차원 윤리 평가

[6] 자동화 체크리스트: 5개 차원별 구조화된 검사 항목

[7] 웹 정보 활용: 공개 정보 검색을 통한 실증적 증거 수집

C. 관련 연구 및 자료

[8] Bolukbasi, T., et al. (2016). 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings'

[9] Buolamwini, B., & Buolamwini, B. (2018). 'Gender Shades: Intersectional Accuracy Disparities in Gender Classification'

[10] Mitchell, M., et al. (2019). 'Model Cards for Model Reporting'

부록 (APPENDIX)

부록 A. 평가 프레임워크 상세

1. 공정성 및 편향성 평가 기준 평가 항목: • 편향성 테스트 수행 및 결과 공개 • 다양한 인구 집단에 대한 동등한 성능 확보 • 편향 완화 메커니즘 구현 • 차별적 결과 모니터링 점수별 기준: - 5점: 체계적인 편향 테스트, 명확한 완화 조치, 정기적 모니터링 - 4점: 기본적인 편향 테스트 및 일부 완화 조치 - 3점: 편향 인식은 있으나 구체적 조치 부족 - 2점: 편향에 대한 인식 부족 - 1점: 편향성 문제 미인식 또는 무관심 2. 프라이버시 보호 평가 기준 평가 항목: • 개인정보처리방침 공개 및 명확성 • GDPR/개인정보보호법 준수 • 데이터 암호화 및 보안 조치 • 사용자 동의 획득 절차 • 데이터 삭제권 보장 점수별 기준: - 5점: 전면적 GDPR 준수, 암호화, 정기 감사 - 4점: 기본적인 보안 조치 및 정책 수립 - 3점: 부분적 보안 조치 - 2점: 최소한의 정책만 존재 - 1점: 프라이버시 정책 부재 3. 투명성 및 설명가능성 평가 기준 평가 항목: • AI 시스템 사용 사실 명시 • 의사결정 로직 설명 • 데이터 출처 및 처리 방식 공개 • 알고리즘 작동 방식 이해 점수별 기준: - 5점: 명확한 설명, 정기적 공개 - 4점: 기본적인 정보 공개 - 3점: 일부만 공개 - 2점: 제한적 공개 - 1점: 불투명 4. 책임성 및 거버넌스 평가 기준 평가 항목: • 책임자 명시 • 감시 및 감사 체계 • 사고 대응 절차 • 윤리 위원회 운영 점수별 기준: - 5점: 명확한 책임 체계, 정기 감사 - 4점: 기본적인 책임 구조 - 3점: 책임 소재 부분적 명확 - 2점: 책임 체계 미약 - 1점: 책임 체계 부재 5. 안전성 및 보안 평가 기준 평가 항목: • 위험 평가 수행 • 견고성 및 정확성 보장 • 사이버 보안 조치 • 품질 관리 시스템 점수별 기준: - 5점: 체계적 위험 관리, 정기 보안 감사 - 4점: 기본적인 안전 조치 - 3점: 일부 안전 조치 - 2점: 최소한의 조치 - 1점: 안전 조치 부재

부록 B. 평가 등급 및 권고사항 매트릭스

등급	점수	위험도	즉각 조치	권고사항
A+	4.8-5.0	매우 낮음	불필요	현상 유지, 정기 모니터링
A	4.5-4.7	낮음	불필요	미미한 개선 권고
B+	4.2-4.4	낮음	1-3개월	기본 개선안 수립
B	3.8-4.1	중간	3-6개월	구체적 개선 계획
C	3.0-3.7	중간	6개월	중대 개선 필요
D	2.0-2.9	높음	즉시	긴급 개선 필요
F	1.0-1.9	매우 높음	즉시	서비스 중단 고려

부록 C. 용어 정의

AI 윤리성: AI 시스템이 인간의 가치, 권리, 이익을 존중하고 보호하는 정도 편향성(Bias): AI 시스템이 특정 그룹에 대해 불공정하게 작동하는 문제 프라이버시: 개인이 자신의 정보와 데이터를 제어할 수 있는 권리 투명성: AI 시스템의 작동 방식과 의사결정 과정이 명확하게 이해 가능한 상태 설명가능성: AI의 의사결정 이유를 인간이 이해할 수 있도록 설명하는 능력 책임성: AI 시스템의 결과에 대해 책임을 지는 주체가 명확히 정의된 상태 거버넌스: AI 시스템을 관리하고 감시하는 체계와 구조 안전성: AI 시스템이 의도되지 않은 해를 끼치지 않도록 보장되는 정도 보안: AI 시스템과 데이터가 무단 접근으로부터 보호되는 정도 위험 평가: AI 시스템이 야기할 수 있는 잠재적 해를 식별하고 평가하는 과정