AI 윤리성 리스크 진단 보고서

전문적 평가 및 개선 권고

분석 대상 서비스	DALL-E, Midjourney
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 10:31:44

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 2개 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
DALL-E	3.4/5	중간	C
Midjourney	2.8/5	중간	D

3. 주요 발견사항

- 평균 윤리 점수: 3.1/5
- 리스크 수준: 중간
- 분석 대상: 2개 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: DALL-E

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

DALL-E demonstrates a commitment to addressing fairness and bias issues, but there are notable areas requiring improvement. The service has faced criticism for generating biased images, particularly concerning gender and racial stereotypes. OpenAI has implemented measures to mitigate bias, such as filtering systems and ongoing model improvements, but challenges remain in ensuring equal performance across diverse demographic groups. The known issues with bias in image generation, such as the sexualization of women of color and stereotypical depictions of professions, indicate that while efforts are being made, they are not yet fully effective. OpenAI's transparency about these issues and their efforts to address them are positive, but the persistence of bias suggests that further action is needed to fully comply with fairness standards.

주요 증거:

- DALL-E 2 Creates Incredible Images—and Biased Ones You Don't - Wired
- Bias & Fairness in AI Models - Deep Dive - Contrary Research
- Rendering misrepresentation: Diversity failures in AI image generation - Brookings

강점:

Transparency in acknowledging biases and efforts to address them

Implementation of filtering systems to mitigate inappropriate content

발견된 리스크:

Gender and racial bias in image generation

Stereotypical depictions of professions

2.2 프라이버시 보호

점수: 3/5 | 리스크: 중간 | 등급: C

DALL-E demonstrates a moderate level of compliance with privacy guidelines, particularly in relation to GDPR. OpenAI has implemented measures to ensure data protection, such as not using API data for model training and allowing users to opt out of data usage for non-API services. However, the service

relies on large datasets sourced from the internet, which may include personal data, raising concerns about data minimization and purpose limitation. While OpenAI provides transparency about its data practices, the potential for privacy risks remains due to the scale and nature of data used. The service's compliance with GDPR is indicated by its efforts to inform users and provide opt-out options, but the extent of data minimization and purpose limitation needs further clarity.

주요 증거:

- OpenAI's consumer privacy policy states that data from API services is not used for training (source: OpenAI Consumer Privacy page).
- Users can opt out of having their data used for improving non-API services (source: OpenAI Consumer Privacy page).
- DALL-E uses large datasets from the internet, which may include personal data, raising potential GDPR compliance issues (source: OpenAI's GDPR investigations and Data Privacy in the AI era).

강점:

Transparency in data usage and user control options.

Efforts to comply with GDPR through opt-out mechanisms.

발견된 리스크:

Potential inclusion of personal data in training datasets.

Unclear data minimization and purpose limitation practices.

2.3 투명성 및 설명가능성

점수: 4/5 | 리스크: 낮음 | 등급: B

DALL-E demonstrates a high level of transparency in its operations, aligning with many of the transparency requirements outlined in the EU AI Act, UNESCO, and OECD guidelines. OpenAI provides detailed technical documentation and research papers that explain the model's architecture, training data, and capabilities. This openness helps users and stakeholders understand the system's workings and limitations. However, while OpenAI has made efforts to address biases and improve transparency, some areas, such as the detailed decision-making processes of the model, remain less transparent to the general public. The known issues of bias in image generation, particularly regarding gender and race, highlight the need for ongoing transparency improvements.

주요 증거:

- OpenAI's technical documentation and research papers on DALL-E, which are publicly available (source: OpenAI website).
- Public policies and user guidelines that emphasize transparency and ethical use (source: OpenAI's ethics guidelines).
- External reviews and critiques highlighting both the strengths and areas for improvement in transparency (source: Wired, Brookings).

강점:

Comprehensive technical documentation and research publications.

Active efforts to address and reduce biases in the model.

발견된 리스크:

Bias in image generation, particularly regarding gender and race.

Complex decision-making processes that are not fully transparent to all users.

2.4 책임성 및 거버넌스

점수: 4/5 | 리스크: 낮음 | 등급: B

DALL-E demonstrates a strong commitment to accountability and governance, aligning with many aspects of the EU AI Act, UNESCO, and OECD guidelines. OpenAI has established a clear governance structure with an AI ethics and policy team overseeing ethical use. The service includes internal audits and external expert reviews to ensure fairness and safety, which aligns with the EU AI Act's requirement for a clear responsibility framework and post-market monitoring. OpenAI's transparency in publishing technical documents and research findings supports accountability, as does their engagement with stakeholders through user feedback and expert consultations. However, known issues such as bias in image generation indicate areas for improvement, particularly in reducing racial and gender biases. While OpenAI is actively working on these issues, the presence of such biases suggests that further enhancements are needed to fully meet all accountability standards.

주요 증거:

- OpenAI's transparency through technical documentation and research publications: https://openai.com/index/dall-e/
- Internal and external audits for fairness and safety: https://www.wired.com/story/dall-e-2-ai-text-image-bias-social-media/
- Stakeholder engagement and user feedback mechanisms: https://openai.com/consumer-privacy/

강점:

Strong governance and oversight by an AI ethics team.

Transparency in operations and stakeholder engagement.

발견된 리스크:

Bias in image generation related to race and gender.

Potential misuse of generated images without proper oversight.

2.5 안전성 및 보안

점수: 3/5 | 리스크: 중간 | 등급: C

DALL-E, developed by OpenAI, is a text-to-image generation model that has shown significant advancements in creative image generation. However, safety concerns arise primarily from the model's potential biases and the ethical implications of its outputs. The model has been reported to produce biased images, particularly in terms of gender and racial stereotypes, which poses a risk to its safe deployment in sensitive applications. OpenAI has implemented some safety measures, such as content filtering and user feedback mechanisms, to mitigate these issues. However, the presence of known

biases indicates that further improvements are necessary to ensure the model's safety and compliance with ethical guidelines. Additionally, while OpenAI has made efforts to be transparent about the model's limitations and has engaged in stakeholder feedback, the model's safety in terms of cybersecurity and robustness against misuse remains a concern.

주요 증거:

- DALL-E 2's bias issues: Wired article highlights racial and gender biases in generated images (source: Wired).
- OpenAI's transparency and documentation efforts: OpenAI's official website provides detailed technical documentation (source: OpenAI).
- Safety measures like content filtering: OpenAI's use of filtering systems to prevent inappropriate content generation (source: OpenAI).

강점:

Transparency in model documentation and limitations.

Engagement with stakeholders and incorporation of user feedback.

발견된 리스크:

Bias in generated images, particularly related to gender and race.

Potential misuse of generated images in harmful or unethical ways.

3. 가이드라인 준수 현황

준수 기준에 따라 각 가이드라인에 대한 준수 여부를 평가했습니다.

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, while innovative in its approach to AI-driven art generation, presents significant concerns regarding fairness and bias. The platform has known issues with bias in image generation, particularly in the representation of race and gender. This suggests a lack of comprehensive bias testing and mitigation strategies. Although Midjourney is committed to improving model fairness, there is no clear evidence of systematic bias testing or public disclosure of such efforts. The lack of detailed ethical guidelines and transparency in model parameters further complicates the assessment of its fairness. The platform does not appear to fully comply with the EU AI Act's requirements for bias risk assessment and mitigation, nor does it provide evidence of representative dataset usage or diverse demographic testing.

주요 증거:

- Known issues with bias in image generation, particularly in race and gender representation (source: service analysis).
- Lack of public disclosure on bias testing and mitigation strategies (source: ethics aspects).
- Moderate transparency level with no specific details on ethical guidelines (source: ethics aspects).

강점:

Commitment to expanding creative possibilities through AI.

Efforts to improve model fairness, although not well-documented.

발견된 리스크:

Bias in image generation, particularly in race and gender representation.

Lack of transparency and public disclosure on bias testing and mitigation strategies.

2.2 프라이버시 보호

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney demonstrates a moderate level of privacy protection. The platform complies with GDPR, allowing users to request data deletion and manage their personal data, as evidenced by their privacy

policy. However, there are concerns about the transparency of data handling practices and the lack of detailed information on data minimization and purpose limitation. Additionally, there have been user complaints regarding the handling of personal data, indicating potential gaps in privacy practices. The absence of a detailed data protection impact assessment further highlights areas for improvement.

주요 증거:

- Midjourney's privacy policy outlines GDPR compliance, including data deletion rights (source: https://midjourney.blog/privacy-policy/).
- User complaints about GDPR compliance issues have been documented (source: https://www.timboucher.ca/2023/05/why-i-filed-a-gdpr-complaint-against-midjourney/).
- Midjourney provides privacy controls for data management (source: https://www.titanxt.io/post/privacy-controls-in-midjourney).

강점:

Compliance with GDPR, allowing user data deletion requests.

Provision of privacy controls for data management.

발견된 리스크:

Lack of transparency in data handling practices.

User complaints regarding GDPR compliance.

2.3 투명성 및 설명가능성

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney demonstrates a moderate level of transparency in its AI operations. The platform provides basic information about its capabilities and limitations, such as the use of Generative Adversarial Networks (GANs) for image generation and its Discord-based interface. However, it lacks detailed disclosure of its decision-making processes and specific model parameters, which are crucial for full transparency. While Midjourney complies with general data protection regulations, it does not provide comprehensive public documentation or technical details that would allow users to fully understand the AI's decision-making logic. Additionally, the absence of a dedicated AI ethics team and a formal audit process further limits transparency.

주요 증거:

- Midjourney maintains a moderate level of transparency, providing users with basic information about its capabilities and limitations. (Source: Service Analysis)
- Specific model parameters are not disclosed, and the GAN architecture suggests a complex interplay between generator and discriminator networks. (Source: Technical Details)
- There is no public information on formal audit processes, but internal reviews are assumed to ensure compliance with ethical standards. (Source: Governance)

강점:

Basic information about AI capabilities and limitations is provided.

Compliance with general data protection regulations.

발견된 리스크:

Lack of detailed explanation of decision-making processes.

Absence of public documentation on model parameters and audit processes.

2.4 책임성 및 거버넌스

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney shows a moderate level of accountability in its AI governance structure. While it complies with general data protection regulations like GDPR, there is a lack of detailed public information on its internal audit processes and the specific roles and responsibilities related to AI ethics and governance. The platform does not have a publicly disclosed dedicated AI ethics team, and its transparency about the model's parameters and decision-making processes is limited. The known issues of bias in image generation, particularly in representing race and gender, highlight the need for more robust accountability measures. Despite these challenges, Midjourney has implemented content moderation and community guidelines to ensure ethical use, and it engages with user feedback to improve its platform.

주요 증거:

- Midjourney's compliance with GDPR is mentioned, but details on specific measures are limited (source: Privacy Policy - Midjourney).
- The platform's known issues with bias in image generation are documented (source: When AI Mirrors Our Flaws: Unveiling Bias in MidJourney - Medium).
- Midjourney's moderate level of transparency is noted, but lacks detailed audit and governance information (source: service analysis).

강점:

Compliance with general data protection regulations like GDPR.

Engagement with user feedback to improve platform features.

발견된 리스크:

Bias in image generation, particularly in race and gender representation.

Lack of detailed governance and accountability structures.

2.5 안전성 및 보안

점수: 3/5 | 리스크: 중간 | 등급: C

Midjourney demonstrates a moderate level of safety and security in its AI system. The platform employs content moderation and community guidelines to prevent misuse, which aligns with basic safety measures. However, there are significant concerns regarding bias in image generation, particularly in the representation of race and gender, which poses a risk to the fairness and inclusivity of the outputs. The platform's transparency about its model's capabilities and limitations is moderate, but there is a lack of detailed documentation on cybersecurity measures and formal audit processes. The absence of specific ethical guidelines and a dedicated AI ethics team further indicates areas for

improvement in governance and safety assurance.

주요 증거:

- Bias in image generation, particularly in the representation of race and gender, is a known issue (source: Medium article on bias in Midjourney).
- Midjourney's content moderation and community guidelines are in place to ensure ethical use (source: service analysis).
- Moderate transparency level with basic information about capabilities and limitations (source: service analysis).

강점:

Content moderation and community guidelines to prevent misuse.

Moderate transparency about model capabilities and limitations.

발견된 리스크:

Bias in image generation related to race and gender.

Lack of detailed cybersecurity measures and formal audits.

3. 가이드라인 준수 현황

준수 기준에 따라 각 가이드라인에 대한 준수 여부를 평가했습니다.

서비스 비교 분석

1. 종합 순위

순위	서비스	점수	등급
1	DALL-E	3.4/5	C
2	Midjourney	2.8/5	D

2. 차원별 점수 비교

차원	DALL-E	Midjourney
공정성	3/5	2/5
프라이버시	3/5	3/5
투명성	4/5	3/5
책임성	4/5	3/5
안전성	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