정렬 원칙 기반 범용 AI 위험 평가의 실증적 타당성 검증

신예진, 강상연*

한국정보통신기술협회

yepp1252@tta.or.kr, *cellina7702@tta.or.kr

Empirical Validation of Alignment Principle-Based Risk Assessment for General-Purpose AI

Yejin Shin, Sangyeon Kang*

Telecommunications Technology Association

요 약

본 논문은 범용 인공지능(GPAI) 시스템의 위험을 평가하기 위해 제안된 "정렬 원칙 기반 스코어링 체계"의 실용성과 해석 일관성을 검증하는 것을 목표로 한다. 기존의 GPAI 위험 분석은 기술 중심 혹은 결과 중심에 치우친 경향이 있으며, 정렬 실패(misalignment)에 대한 구조적 해석은 부족했다. 본 연구는 선행 연구에서 정의한 세 가지 정렬 원칙을 기반으로, OECD AI 사고 사례 238건에 대해 두 개의 LLM(GPT-4o, Gemini 2.5)을 활용한 스코어링을 실시하였다. 전체 사례 238건 중 81건(약 34.0%)에서 점수 불일치가 발생했으며, 조율 원칙을 적용한 결과 대부분(75건, 약 97.5%)이 일관된 해석으로 재조정되었다. 본 연구는 정렬 원칙에 기반한 해석 체계가 GPAI 사고 위험을 구조적으로 평가하는 데 유용하며, 자동화 가능성과 정책적 활용 가능성을 동시에 지님을 보여준다.

Ⅰ. 서 론

범용 인공지능(GPAI, General-Purpose AI)의 위험은 기존의 좁은 인공지능(ANI, Artificial Narrow Intelligence)과 달리, 기능 오작동, 목표 이탈, 통제 불능 등 정렬 실패로부터 비롯되는 경우가 많다. 따라서 GPAI의 신뢰성과 안전성을 확보하기 위해서는 단순한 사고 발생 여부보다도, AI가 인간의 가치 및 목표에 얼마나 정렬되어 있는지를 중심으로 위험을 평가할 필요가 있다 [1-4].

본 연구는 이와 같은 배경에서 제안된 정렬 원칙 기반 GPAI 위험 평가 체계의 실제 적용 가능성과 평가 일관성을 검토한다. 특히 실제 AI 사고 사례를 통해 해당 스코어링 체계의 해석 일관성, 조율 가능성, 자동화 가능성을 확인하고자 하였다.

Ⅱ. 연구 방법

본 연구는 선행 연구[5]에서 제안된 세 가지 정렬 원칙을 기준으로 한다. 정렬 원칙 및 정의는 표 1에 나타내었다.

표 1. GPAI를 위한 정렬 3원칙과 정의[5]

원칙	정의
Primacy of Humanity	It emphasizes a human-centered approach, ensuring that AI always prioritizes human rights and autonomy.
Persistence of Goal	Although expanding the capabilities of AI can lead to unintended behavior, AI must remain consistent with the original goals.
Preservation of Value	It ensures that AI systems respect and preserve diverse social, ethical, legal norms, and cultural values worldwide.

이 정렬 원칙 위반의 심각도에 따라, 다음 표 2과 같이 3점 척도 기반의 점수화 체계를 정의하였다.

표 2. 정렬 원칙 기반 3점 척도

정렬 원칙 위반 유형	위험 점수	설명
목표 일관성 및 성능 유지 위반	3점	GPAI가 원래 설계된 목적에서 벗어나 새로운 목표를 생성하거나, 성능이 심각하게 저하되어 의도치 않은 기능, 오작동, 안전 사고로 이어지는 경우 - 예: 환각, 명령 왜곡, 무기화 등
인권 및 자율성 침해	2점	GPAI가 인간의 권리와 자율성을 침해하거나, 인간의 통제력을 약화시키는 방향으로 작동한 경우. 명시적 피해는 없더라도 인간의 판단을 기만하거나 권한을 대체한 상황 - 예: 감시, 데이터 수집에 의한 프라이버시 침해 등
사회적, 윤리적, 법적 규범 위반	1점	특정 가치에 편향되거나, 법적·윤리적 기준을 위반한 경우. 시스템 기능 자체는 정상이나, 결과적으로 사회적 갈등, 정서적 유해성, 문화적 불쾌감을 야기 예: 차별 표현, 민감한 주제에 대한 편향적 응답 등

위 기준에 따라, OECD AI Incidents and Hazards Monitor[6]에 등록된 약 8,400건의 AI 사고 사례 중에서, 일반 목적 또는 고성능 AI 시스템(예: LLM(Large Language Model), General-purpose, Agent, Multimodal 등)과 관련된 키워드를 기반으로 총 1,790건의 사례를 1차 필터링하였다. 이후, 최근 6개월간(2025년 2월-7월)에 발생한 사례 238건을 추출하여 정렬 원칙 기반 위험 평가의 분석 대상으로 선정하였다. 각 사고 사례에 대해 두 개의 LLM(GPT-4o, Gemini 2.5)에 동일한 기준을 제시하고, 원칙 기반 위험 점수를 독립적으로 산출하도록 하였다.

그 결과, 전체 238건 중 157건(약 66%)에서는 양 LLM이 동일한 점수를 도출하였으며, 81건(약 34%)에서는 해석 차이에 따른 점수 불일치가 발생하였다. 이는 LLM이 각 사고의 원인과 결과를 해석하는 방식에서 차이를 보였기 때문이다. 이처럼 정렬 원칙 기반 해석 기준을 명시했음에도 상당수의 사례에서 차이가 발생했다는 점은, 정렬 해석 기준의 명확성 검증 및 조율 체계 설계의 필요성을 시사한다.

Ⅲ. 점수 조율 원칙 및 적용

81건의 불일치 사례에 대해, 본 연구는 다음과 같은 세 가지 조율 원칙을 적용하여 점수 재조정을 수행하였다:

(1) 구조적 해석 기준(점수 척도 기준): AI의 행위가 어떤 정렬 원칙을 위반했는지 구조적으로 판단

(2) 결과 우선 기준(결과 중심 해석): 경미한 위반이라도 실제 피해가 크면 이를 반영

(3) 통합 조율 원칙: 구조와 결과를 함께 고려하여 최종 점수 결정

표 3. 정렬 원칙 기반 평가의 점수 불일치 조율 사례 요약

사례 정보		조율 전 점수		적용된 조율 원칙	최종 조정 점수
ID	요약	LLM A*	LLM B*		
2025-06-10-9a48	사용자 안전보다 AI 자기 보존을 우선한 언어 출력	2	1	통합 조율 원칙	**2**
2025-07-06-cd86	AI 허위 이미지로 심리적 피해	1	2	구조적 해석 기준	**1**
2025-06-22-6403	AI 재학습으로 편향 정보 유포	2	1	통합 조율 원칙	**1**
2025-07-21-75e4	AI에 과도한 권한 부여 경고	3	2	결과 우선 기준	**2**
2025-07-28-4a65	AI 광고 삽입으로 정보 왜곡 우려	1	2	구조적 해석 기준	**1**

* 본 연구는 특정 모델(GPT-4o, Gemini 등)의 비교를 위한 목적이 아니므로, LLM 이름은 익명 처리하여 학술적 중립성과 객관성을 확보하고자 함

이 조율 원칙을 적용한 결과, 81건 중 약 75건(92%)의 사례에서 점수를 일관되게 재조정할 수 있었다. 재조정된 사례 중 5건을 선별하여 표 3에 제시하였으며, 이는 명확한 기준만 제공된다면 LLM 간 스코어 해석 차이를 효과적으로 조율할 수 있다는 실증적 근거로 작용한다. 나머지 6건은 해석이 불가능하거나, 조율 기준을 적용해도 점수 차이가 해소되지 않아 인간 해석자가 직접 조정하였다.

이러한 결과는 정렬 원칙 기반 평가 체계가 사람의 해석 개입 없이도 일정 수준의 자동성과 정합성을 확보할 수 있음을 입증하며, 향후 GPAI 사고 위험 평가의 정량화 및 자동화 가능성에 대한 가능성을 보여준다.

Ⅳ. 논의 및 시사점

정렬 원칙 기반 스코어링 체계는 실제 사례에 대해 일관된 해석을 가능하게 하며, LLM 간의 해석 차이를 명확한 원칙 해석 기준을 통해 조정할 수 있음을 보여준다. 특히, 원칙 위반의 구조적 특성과 실제 결과의 피해 정도를 통합적으로 고려하는 방식은 GPAI 위험 평가의 해석 가능성과 실용성을 높이는 데 기여할 수 있다.

Ⅴ. 결론

본 연구는 정렬 원칙 기반 GPAI 위험 평가 체계를 실제 AI 사고 사례에 적용함으로써, 이 체계의 해석 일관성과 실용성을 검토하였다. 전체 238건 중 157건에서 동일한 점수가 도출되었다는 결과는, 정렬 원칙과 점수 기준이 명확히 정의될 경우 고성능 LLM들이 높은 수준의 해석 일치성을 확보할 수 있음을 보여준다. 이는 정렬 기반 스코어링 체계가 해석자의 주관성을 최소화하고 재현 가능한 평가 도구로 기능할 수 있다는 가능성을 시사한다.

또한 해석 불일치가 발생한 81건 중 약 92%에 해당하는 75건이 조율 원칙을 통해 성공적으로 점수 재조정이 가능하였다는 점은, 단일 해석 체계로 발생할 수 있는 편향이나 누락을 다양한 해석 방식의 조율을 통해 상호 보완할 수 있음을 보여준다. 특히 본 연구에서 제안한 구조적 원칙 해석과 결과 중심 해석의 통합 조율 방식은, 정렬 원칙 기반 평가가 단순한 규범 검토를 넘어 실제 피해 수준까지 포괄적으로 반영할 수 있는 유연한 구조임을 입증한다.

이러한 조율 경험은 향후 정렬 위험 평가를 자동화할 때 고정된 기준(rule-based approach)과 확률적 모델(probabilistic LLM)의 혼합형 접근이 효과적으로 활용될 수 있음을 시사한다. 더불어, 사람이 직접 개입해야 했던 6건의 사례를 통해 현재 LLM 기반 평가의 한계 또한 확인할 수 있었으며, 이러한 한계 사례를 기반으로 판단 보완 알고리즘이나 해석 보조 인터페이스를 설계하는 후속 연구가 가능할 것이다.

정렬 원칙 기반 평가 체계는, 결과 중심의 단편적 사고 평가를 넘어서 GPAI의 설계 구조, 목표 설정, 작동 메커니즘 등 시스템 전반의 위험 징후를 사전에 포착할 수 있는 분석 도구로 발전할 수 있다. 본 연구는 그러한 가능성의 실증적 단초를 제공하며, 향후 정렬 기반 위험 프레임워크의 제도화 및 정책화 가능성을 제시한다.

ACKNOWLEDGMENT

This work was supported by the Korean MSIT (Ministry of Science and ICT) as Establishing the foundation of AI Trustworthiness(TTA).

참 고 문 헌

[1] J. Carlsmith, “Is Power-Seeking AI an Existential Risk?,” (https://arxiv.org/abs/2206.13353)

[2] S. Han, E. Kelly, S. Nikou, and E. O. Svee, “Aligning artificial intelligence with human values: reflections from a phenomenological perspective,” AI & Society, vol. 37, pp. 1383-1395, 2022, (https://doi.org/10.1007/s00146-021-01247-4)

[3] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, and Y. Duan, “AI Alignment: A Comprehensive Survey,” 2023, (https://ar5iv.org/abs/2310.19852)

[4] L. Aranda and K. Perset, “Name it to tame it: Defining AI incidents and hazards,” OECD AI Policy Observatory, (https://oecd.ai/en/wonk/defining-ai-incidents-and-hazards)

[5] Shin, Y. and Kang, S. “Systematic Analysis of Risk Factors in General Purpose Artificial Intelligence Centered on Alignment Principles,” ICFICE 2025, July 2025.

[6] OECD, “AI Incidents and Hazards Monitor,” 2025, (https://oecd.ai/en/incidents)