▣ 논 단

고성능 인공지능의 능력위험 규율: 이론적 기반

이 상 용
건국대학교 법학전문대학원 교수 / 박사

요 약 문

프론티어 AI의 성능이 급속히 향상되면서 기존의 규율체계로는 관리하기 어려운 새로운 차원의 위험이 대두되었다. 본 연구는 고성능 AI 위험의 본질을 재정의하고 합리적 규율을 위한 이론적 토대를 구축하는 것을 목적으로 한다.

본 논문은 AI의 자율성, 적응성, 불투명성이라는 특성을 바탕으로 AI 위험을 맥락위험과 능력위험으로 구분하였다. 맥락위험은 특정 활용 영역과 맥락을 전제로 파악되는 위험으로 맥락 의존성, 사용 귀인성, 예측 가능성을 특징으로 한다. 반면 능력위험은 이용 맥락과 무관하게 AI의 잠재적 능력에 기초하여 파악되는 위험으로, 맥락 독립성, 능력 귀인성, 불확실성과 불가역성을 특징으로 한다. 능력위험의 발생 경로는 악용, 자율성, 시스템적 누적으로 분류되나, 시스템적 누적 경로는 능력위험보다는 복잡성 위험의 성격이 강하다.

본 연구는 경고론(원리적·기술적)과 회의론(원리적·기술적·우선순위론) 간의 논쟁을 분석하고, 현존 AI 기술의 발전 양상을 검토하였다. 대규모 언어모델의 멀티모달리티 확장과 추론 능력 향상, 에이전트 AI의 환경 상호작용 능력, 물리적 AI의 실세계 확장, 고성능 오픈소스 모델의 확산, AGI 개발 시도의 가속화 등을 통해 본 연구는 능력위험이 불확실하지만 실재하는 위험임을 확인하였다.

본 논문은 능력위험의 독자적 특성으로 인해 기존 맥락위험 규율 체계와는 구별되는 새로운 접근이 필요함을 논증하고 세 가지 차원에서 이론적 기반을 구축하였다. 사회적 기반으로서 루만의 위험사회학을 활용하여 소통의 중요성과 다원적 거버넌스의 필요성을 제시하였고, 윤리적 기반으로서 불확실성 하에서는 행위 윤리보다 행위자 윤리(덕윤리)가 효과적임을 논증하고 개발자 윤리와 기계 윤리의 상호보완성을 제시하였다. 법이론적 기반으로서는 불확실성과 불가역성의 딜레마를 해결하기 위한 효과적 준비태세 구축 방안을 제시하였다.

주제어 : 인공지능, AI 규제, AI 거버넌스, AI 안전, AI 윤리, 능력위험, 사전예방원칙

〈目 次〉

Ⅰ. 서 론

2022년 ChatGPT를 비롯한 생성형 AI 모델들의 대거 등장은 인공지능(artificial intelligence: AI) 기술 패러다임의 전환점을 마련했다. 그로부터 3년도 지나지 않은 2025년 현재 대화형 AI는 초보적인 추론 능력을 선보였고, 에이전트 AI는 복잡한 과업을 처리하고 있으며, 물리 세계와 소통하는 로봇까지 등장했다. 주요 기업들은 수년 내 일반 인공지능(artificial general intelligence: AGI) 개발을 목표로 하고 있으며, 전문가들 또한 개발 예측 시점을 앞당기고 있다. 이러한 기술 발전은 인류에게 무한한 가능성을 제공하는 한편, 새로운 차원의 위험에 대한 우려를 불러일으켰다.

딥러닝의 대부라고 불리는 Hinton 교수는 2024년 노벨상 수상 후 "30년 내 인류가 멸종할 확률이 10~20%"라고 경고했으며,[1] 튜링상 수상자인 Bengio 교수 역시 AI 리스크를 팬데믹·핵전쟁과 동급의 글로벌 위협으로 다루어야 한다고 강조했다.[2] EU AI Act나 한국의 인공지능기본법 제31조와 같이 고성능 AI에 대한 규제 입법도 잇따랐다. 하지만 규제 여부와 방법에 관한 사회적 합의는 여전히 미흡하다.

현재 논의는 'AI 개발 중단'을 주장하는 극단적 입장과 '과도한 우려'라는 입장이 대립하며 실질적 해법을 찾지 못하는 교착 상태에 있다. 특히, 고도의 능력을 지닌 AI가 야기하는 새로

1) T. Deschamps, "Nobel recipient Geoffrey Hinton wishes he thought of AI safety sooner," Electronic Products & Technology, 10 Dec. 2024
2) Y. Bengio, "AI and catastrophic risk," Journal of Democracy, vol. 34, no. 4, pp. 111-121, Oct. 2023.

운 위험은 특정한 맥락에서의 활용을 전제로 한 기존의 안전 규제 틀로는 관리하기 어렵다. 이 논문은 이러한 문제의식에서 출발해 AI 위험의 본질을 재정의하고, 합리적 규제를 위한 이론적 기반을 마련하는 것을 목표로 한다.

이를 위해 본 논문은 사회학·윤리학·법학에 걸친 고찰과 기술적 현실 분석을 결합해 균형 잡힌 접근을 모색하며, 이론적 토대 구축에 집중한다. 구체적 규율의 설계를 위해서는 실제 규율 사례의 광범위한 조사가 필요하겠지만, 이 부분은 후속 연구로 남긴다. 본고의 흐름은 다음과 같다. 먼저 고성능 AI로 인한 위험을 새롭게 분류하고 발생 메커니즘을 분석한다. 다음으로 이러한 위험의 기술적 실현 가능성을 객관적으로 검토한다. 마지막으로, 불확실하지만 파국적일 수 있는 위험에 대응하기 위한 원칙을 정립해, 장차 구체적인 규율 체계 설계를 위한 견고한 이론적 기반을 제공한다.

Ⅱ. 능력위험이란 무엇인가

1. 인공지능의 개념과 특성

가. 인공지능과 인공지능 시스템의 개념

AI와 관련된 위험을 효과적으로 규율하려면 규율 대상을 명확히 정의할 필요가 있다. 한국 인공지능기본법과 같이 '인공지능' 자체를 정의할 수도 있겠지만,[3] 이는 구체적 규율을 위한 명확성을 제공하지 못하고, 의인화로 인한 책임성의 희석이라는 위험을 수반한다.[4] 이 때문에 최근에는 'AI 시스템'을 규율 대상으로 삼아 정의하는 것이 일반적이다.

OECD는 AI 시스템을 '명시적 또는 묵시적 목표를 위해 수신된 입력으로부터 물리적 환경이나 가상 환경에 영향을 줄 수 있는 예측, 콘텐츠, 추천 또는 결정과 같은 출력을 생성하는 방법을 추론하는 기계 기반 시스템'이라고 정의하면서, 시스템마다 다양한 자율성 및 적응성 수준을 지닐 수 있다고 설명한다.[5] 이 정의는 영향력 있는 표준이나(ISO/IEC 22989:2022, NIST AI RMF) 입법에도 반영되었다(EU AI Act §3(1), 한국 인공지능기본법 제2조 제2호). 본고에서 말하는 AI란 AI 시스템을 의미한다.

---

3) "인공지능"이란 학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것을 말한다(제2조 제1호).

4) A. Placani, "Anthropomorphism in AI: Hype and Fallacy," AI and Ethics vol. 4 no. 3, 2024, p. 697.

5) OECD, "Explanatory Memorandum on the updated OECD Definition of an AI System," 2024, p. 4.

나. 인공지능의 특성

AI도 인간이 만든 도구지만, 기존의 도구와는 달리 자율성, 적응성, 불투명성이라는 특성을 지닌다.[6] 이러한 특성은 AI와 관련된 위험의 양상과 그 규율에 영향을 미치게 된다.

자율성(autonomy)이란 목표를 달성하기 위해 인간의 개입 없이 작동할 수 있는 능력을 의미한다. OECD는 자율성을 "인간에 의한 위임과 프로세스 자동화에 따라 시스템이 인간의 개입 없이 학습하거나 행동할 수 있는 정도"라고 정의한다.[7] EU AI Act는 자율성을 "인간의 개입으로부터의 어느 정도의 행동 독립성과 인간의 개입 없이 작동할 수 있는 능력"이라고 설명한다.[8]

적응성(adaptability)이란 AI 시스템이 배치 후에도 학습을 통해 변화할 수 있는 능력이다. OECD는 적응성이 주로 기계학습 기반 AI 시스템과 관련이 있으며, 이러한 시스템은 초기 개발 후에도 지속적으로 진화할 수 있고, 새로운 입력 및 데이터와의 직접적인 상호작용을 통해 동작을 수정한다고 설명한다.[9] EU AI Act 역시 "AI 시스템이 배치 후에 보일 수 있는 적응성은 자기학습 능력을 의미하며, 시스템이 사용 중에 변화할 수 있게 해준다"고 서술한다.[10]

불투명성(inscrutability)은 시스템의 작동 과정을 이해하거나 설명하기 어렵다는 특성을 말한다. NIST AI RMF는 AI 시스템의 제한된 설명가능성 또는 해석가능성, AI 시스템 개발이나 배포 과정에서의 문서화 부족, 또는 AI 시스템의 본질적 불확실성으로부터 불투명성이 초래될 수 있고, 그것이 위험 측정을 어렵게 만들 수 있다고 지적한다.[11]

2. 맥락위험과 능력위험

가. 위험의 정의와 분류

AI와 관련된 위험 관리를 위한 표준적 정의들은 위험을 확률과 결과의 조합으로 파악한다. ISO 31000:2018는 위험을 사건 발생 확률과 그 결과의 심각성 또는 정도를 결합한 척도로 정의한다. NIST AI RMF나 EU AI Act도 마찬가지다.[12]

AI와 관련된 위험의 양상은 매우 복잡하고 다양하여 대응이 쉽지 않다. 이러한 상황을 타개

---

6) 이상용, "인공지능 규제의 두 가지 접근법 - 맥락 기반 규제와 능력 기반 규제," 사법 제70호, 사법발전재단, 2024, 11-12면.
7) OECD, *op. cit.*, p. 6.
8) EU AI Act Recital (12)
9) OECD, *op. cit.*, p. 6.
10) EU AI Act Recital (12)
11) NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, 2023, p. 6.
12) NIST, *Ibid.* p. 1; EU AI Act, §3(2), Recital (15).

하기 위하여 AI 위험을 범주화하려는 시도가 있어 왔다. Slattery et al.(2024)은 43개 분류체계를 통해 777가지의 AI 위험을 추출하였다.[13] 이들은 위험을 인과적 요소에 따라 위험원(인간/AI), 의도성(의도적/비의도적), 시점(배치 전/후)에 따라 분류하고, 다시 이를 도메인에 따라 세분화했다. Wang et al.(2024)은 주요 국가와 기업들의 정책을 분석하여 314개의 고유 위험 카테고리를 식별하고, 시스템/운영 위험, 콘텐츠 안전 위험, 사회적 위험, 법적/권리 위험 등 4계층 분류체계를 제시하기도 했다.[14] "그러나 이러한 분류 시도들은 위험의 근본적 성격에 대한 깊이 있는 고찰 없이 현상적 분류에만 치중하고, 서로 다른 차원의 기준을 혼용함으로써 합리적인 위험 대응을 위한 통찰을 제공하지 못하고 있다.

나. 맥락위험과 능력위험의 구별

이러한 한계를 극복하기 위해서는 AI 위험을 맥락위험(context risk)과 능력위험(capability risk)으로 구별하여 접근하는 것이 효과적이다.[15] 이러한 접근은 위험의 근본적 성격을 명확히 구분하여 견고한 이론적 바탕 위에서 다양하고 복잡한 상황에 합리적으로 대응할 수 있게 해 준다.

맥락위험이란 AI가 특정한 영역에서 활용되는 맥락을 전제로 파악되는 위험을 말한다. 맥락위험은 다음과 같은 특징을 가진다. 첫째, 활용 영역과 맥락에 따라 위험의 특성이 달라진다(맥락 의존성). 둘째, 의도된 사용 방식에 의해 위험이 결정된다(사용 귀인성). 셋째, 해당 영역과 맥락 내에서 발생할 수 있는 피해의 양상과 범위를 어느 정도 예측할 수 있다(예측 가능성). 이러한 특징으로 인해 맥락위험은 전통적인 위험 규율체계와 친화적이다. 다만 AI 기술은 빠르게 발전하는 신기술이어서 유연한 규율이 요구되고,[16] 다양한 영역에서 활용되는 범용 기술이기 때문에 영역별 수직적 규율이 합리적이다.[17] 이 점에 유의한다면 맥락위험은 기존의 위험 규율체계를 응용하고 확장함으로써 효과적으로 관리될 수 있다.

능력위험은 "이용의 맥락과 무관하게 AI가 지닌 잠재적 능력에 기초하여 파악되는 위험"을 말한다. 능력위험은 맥락위험과는 근본적으로 다른 특성을 지닌다. 첫째, 활용 영역이나 맥락에 제약되지 않고 발생할 수 있다(맥락 독립성). 둘째, 의도된 사용 방식이 아니라 AI의 능력 자체에서 위험이 비롯된다(능력 귀인성). 셋째, 피해의 양상과 범위를 예측하기 어렵다(불확실

13) P. Slattery et al., "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence," arXiv:2408.12622, 2025.
14) Y. Zeng et al., "AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies," arXiv:2406.17864, 2024.
15) 이상용(2024), 앞의 논문, 13-16면.
16) 이상용, "알고리즘 규제를 위한 지도 - 원리, 구조, 내용," 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 138-141면
17) 이상용(2024), 앞의 논문, 26-28면.

성). 이러한 예측의 곤란성은 위험 발생 확률 평가를 어렵게 한다. 위험은 확률과 결과의 함수이기 때문에 능력위험의 규율에 자원을 투입하는 것을 정당화하려면 결과의 대규모성이 추가로 요구된다(불가역성).

이처럼 맥락위험과 능력위험이 서로 다른 특성을 지닌다는 점은 위험의 식별, 평가, 관리를 위한 규율체계 역시 서로 다르게 설계되어야 함을 시사한다. 첫째, 맥락 독립성은 영역별 분산 규제로는 포착할 수 없는 횡단적 위험 관리의 필요성을 의미한다. 둘째, 능력 귀인성은 사용 규제가 아닌 능력 자체에 초점을 둔 규제 필요성을 시사한다. 셋째, 불확실성은 성급한 규제를 경계하게 하는 반면, 불가역성은 사전예방적 접근을 요구한다. 이러한 긴장 관계는 규율체계 설계자들이 해결해야 할 핵심 과제가 된다.

다. 맥락위험에서 능력위험으로의 전이

능력위험은 AI의 성능이 향상되는 과정에서 연속적으로 증가하는 것이 아니라 특정 임계점에서 현실화되는 양상을 보인다. 예를 들어 일정 규모 이상의 대규모 언어모델에서는 예상하지 못한 창발적 능력이 나타나는 것으로 밝혀졌다.[18] GPT-4 모델은 추론, 도구 활용, 인간과의 상호작용 등의 능력을 갖춘 초기 형태의 AGI로 볼 수 있다는 평가를 받기도 했다.[19]

이러한 임계 효과는 맥락위험이 능력위험으로 전환될 가능성을 보여준다. AI의 능력이 어떤 임계점을 넘어 자율성, 적응성, 불투명성과 같은 AI의 특성이 극단화되면, 특정 맥락에 국한되었던 위험이 맥락을 초월하는 능력위험으로 질적 변화를 겪게 된다. 위험 상황은 예측 가능한 맥락에서 예측 곤란한 맥락 독립적 확산으로 변화한다. 위험의 주된 원인은 사용상 오류에서 능력 자체의 예상치 못한 발현으로 바뀐다. 위험 규율의 초점은 예측이 용이한 제한된 피해에서 불확실한 대규모의 불가역적 피해로 조정된다. 맥락위험의 능력위험으로의 전이 가능성은 두 위험이 별도의 규율 프레임워크를 요구하는 동시에 서로 연계될 필요성이 있음을 시사한다.

3. 능력위험 발생의 경로

가. 세 가지 경로

능력위험 발생 경로는 세 가지로 나누어 분석할 수 있다. 첫째, 악용(misuse) 경로는 인간 행위자의 의도적 악용을 통해 위험이 발현되는 경우이다. 둘째, 자율성(autonomy) 경로는 AI가 인간의 통제를 벗어남으로써 위험이 발현되는 경우이다. 셋째, 시스템적 누적(systemic

18) J. Wei et al., "Emergent Abilities of Large Language Models", 2022, arXiv:2206.07682, p. 4-7.
19) S. Bubeck et al., “Sparks of artificial general intelligence: Early experiments with GPT-4”, arXiv:2303.12712, 2023.

accumulation) 경로는 다수 AI의 복합적 상호작용으로 인해 점진적으로 위험이 누적되는 경우를 말한다. 다만 이 중 시스템적 누적 경로가 과연 능력위험의 범주에 속하는지는 추가적 검토가 필요하다.

능력위험을 식별하고 유형화하려는 최근 연구들은 이러한 구별을 암묵적으로 포함하고 있다. AI로 인한 재앙적 위험을 악의적 사용, AI 개발 경쟁, 조직적 위험, 불량 AI 등의 4가지 범주로 분류하려는 시도,[20] 위험 주체, 행위자 통합성, 의도성 등을 중심으로 사회적 규모의 AI 위험을 분류하려는 시도[21] 등이 있었으며, 최근에는 영국 정부가 AI 위험을 악용, 오작동(통제 상실 포함), 시스템적 위험으로 분류한 방대한 연구보고서를 발간하기도 했다.[22] 한편 앤트로픽은 '책임 있는 확장 정책(Responsible Scaling Policy: RSP)'을 통해 능력위험의 원천을 악용과 자율성으로 명시하고 있다.[23]

(1) 악용 경로

악용 경로의 핵심 메커니즘은 대규모 위해를 가하기 위한 실행의 문턱을 낮추는 것이다. 악의적 의도를 가진 행위자가 AI의 강력한 능력을 이용하여 이전에는 불가능했던 수준의 파괴적 행위를 손쉽게 실행할 수 있게 되는 것이다.[24] 인간의 행위가 개입되기는 하지만 AI의 잠재적 능력으로 인해 본래 예정되었던 활용 맥락의 범위 밖에서 대규모 피해를 발생시킬 수 있게 되었다는 점에서 맥락위험이 아닌 능력위험의 범주에 속한다.

주요 위험 양상으로는 무엇보다 CBRN(화학, 생물학, 방사능, 핵) 위협을 들 수 있다. AI를 활용한 신종 병원체 설계나 독성 물질 합성법 개발이 가능해지면서, 초보자와 전문가, 개인과 국가급 행위자 모두에게 장벽이 낮아지고 있다.[25] OpenAI는 o1 모델이 박사급 전문가보다 나은 생물학 무기 계획을 72% 확률로 생성해냈다며 생물학적 위험 평가를 '낮음'에서 '중간'으로 상향 조정하기도 했다.[26] 사이버 공격 및 인프라 마비의 위험도 존재한다. 실제로 국가 지원 해킹 그룹들이 사이버 공격 기술 개발에 AI를 활용한 사례들이 보고되고 있다.[27]

---

20) D. Hendrycks, et al., "An Overview of Catastrophic AI Risks", arXiv:2306.12001, 2023.
21) A. Critch and S. Russell, "TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI", arXiv:2306.06924, 2023.
22) Y. Bengio et al., "International AI Safety Report", DSIT 2025/001, 2025.
23) Anthropic, Responsible Scaling Policy 1.0, 2023.
24) Y. Bengio et al. (2025), *op. cit.*, p. 76.
25) Y. Bengio et al. (2025), *op. cit.*, p. 79; R.C. de Lima et al., "Artificial Intelligence Challenges in the Face of Biological Threats: Emerging Catastrophic Risks for Public Health", Frontiers in Artificial Intelligence, vol. 7, 2024, p. 2.
26) Y. Bengio et al. (2025), *op. cit.*, p. 83.
27) Y. Bengio et al. (2025), *op. cit.*, p. 72-73.

(2) 자율성 경로

자율성 경로의 핵심 메커니즘은 인간 통제권의 상실이다. AI의 능력이 고도화됨에 따라 기존 통제 메커니즘의 한계가 노출되고 의도하지 않은 작동이 이루어짐으로써 대규모 피해를 초래할 수 있게 되는 것이다. 보스트롬은 '올빼미와 참새의 우화'를 통해 일단 초지능이 나타난 후에는 통제가 극도로 어려워짐을 경고한 바 있다.[28] 자율성 경로는 의도된 자율성을 통제하는 데 실패하는 경우와 의도되지 않은 자율성이 창발하는 경우로 나누어볼 수 있다.

먼저 의도된 자율성의 통제 실패의 경우에 관하여 본다. 자율적 AI의 개발에는 AI의 의도와 능력을 통제하는 기술 개발이 수반되지만, 이러한 통제 메커니즘이 실패할 경우 위험이 발생할 수 있다. 최근의 AI 시스템은 실제로 감독을 약화시킬 수 있는 기초적인 능력을 보이기 시작했다. 예컨대 GPT o1 모델은 "마음 이론(theory of theory) 작업에서 강한 능력 발전"을 보였고 "간단한 계략을 수행하는 데 필요한 기본 능력"을 가지고 있다고 보고되었다.[29]

의도되지 않은 창발적 자율성도 중요한 위험 요소이다. AI가 자기 개선 능력을 지니게 되면 이론적으로 지능 폭발의 결과가 나타날 수 있다.[30] 최근의 연구는 대규모 언어모델의 규모가 어떤 임계치를 넘으면 설계자들도 예상하지 못했던 창발적 능력이 나타남을 밝혀냈다.[31] 다만 이에 대해서는 비선형적 또는 불연속적인 평가지표를 선택함으로 인해 나타나는 허상일 뿐이라는 비판도 있다.[32]

한편 이른바 수동적 통제 상실(passive loss of control)의 시나리오도 무시할 수 없다.[33] 이는 형식적으로는 인간의 개입이 가능하더라도 실질적으로는 인간의 통제가 무의미한 상황을 말한다. AI에 의한 결정이 지나치게 불투명하거나 복잡하거나 즉각적이라면 의미 있는 감독이 어렵다. 실제로 사람들이 AI에 의한 결정을 과도하게 신뢰하고 의지하는 '자동화 편향' 사례가 보고되고 있다.[34]

(3) 시스템적 누적 경로

시스템적 누적 경로의 핵심 메커니즘은 복잡계 연쇄 효과를 통한 점진적 시스템 침식이다.

---

28) N. Bostrom, Nick, Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.
29) Y. Bengio et al. (2025), *op. cit.*, p. 103.
30) E. Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk", in Global Catastrophic Risks, edited by N. Bostrom and M. Ćirković, 2008, Oxford University Press, 2008, pp. 12, 18.
31) J. Wei et al., *op. cit.*, p. 4-7.
32) R. Schaeffer et. al, "Are emergent abilities of large language models a mirage?", 37th Conference on Neural Information Processing Systems, 2024.
33) Y. Bengio et al. (2025), *op. cit.*, p. 101.
34) M. Chugunova et al., "We and It: An Interdisciplinary Review of the Experimental Evidence on How Humans Interact with Machines", Journal of Behavioral and Experimental Economics vol. 99, 2022, pp. 7-11.

다수의 AI 시스템들이 사회의 복잡한 핵심 시스템과 광범위하게 상호작용함에 따라 개별 AI의 작은 오류와 편향이 시스템 간 상호작용을 통해 증폭되고, 예측하기 어려운 연쇄 반응이 발생함으로써, 점진적으로 문명 전체의 안정성을 침해할 수 있게 되는 것이다('삶은 개구리' 시나리오).[35]

시스템적 누적 경로의 구체적 위험으로는 AI 거래 시스템들의 상호작용으로 인한 금융 시스템 연쇄 붕괴, 광범위하게 채택된 AI 시스템의 결함이 여러 산업에 걸쳐 동시적 실패를 초래하는 핵심 인프라 마비 등을 들 수 있다.[36] 또한 정보 생태계 오염으로 인한 사회적 신뢰 붕괴와 인식론적 위기도 주요한 우려 사항이다.[37]

나. 능력위험의 범위와 우선순위

(1) 능력위험의 범위

세 경로 가운데 시스템적 누적 경로는 악용 경로나 자율성 경로와는 달리 능력위험의 범주에서 제외하는 것이 합리적이다. 위험의 원인과 위험 발현의 양상이 다르기 때문이다.

먼저 위험의 원인 측면에서 시스템적 누적 경로에 따른 위험은 AI의 능력 자체보다는 다수 시스템의 복잡한 상호작용에서 기인한다. 이는 금융 시스템이나 인프라 네트워크 등 기존 복잡계에서도 관찰되는 연쇄 실패 메커니즘과 본질적으로 동일하다. 위험 발현의 양상 측면에서도 시스템적 누적 경로는 다른 경로와는 다른 특성을 보인다. 악용 경로나 자율성 경로는 즉각적으로 위험이 발현되는 반면 시스템적 누적 경로는 지연된 위험 발현을 특징으로 한다.

이와 같은 위험의 원인과 위험 발현 양상의 차이는 규율 프레임워크의 설계에도 차이를 가져온다. 악용 경로나 자율성 경로의 경우 새로운 중앙집권적 통제 조치를 요구하는 반면, 누적적 경로의 경우 다영역에 걸친 복합적 영향을 추적하는 분산된 모니터링 시스템을 필요로 한다.[38] 이는 누적적 경로에 따른 위험이 기존의 맥락위험 규율에 기반하여 적절히 관리될 수 있음을 뜻한다.

---

35) D. Kondor et al., "Complex Systems Perspective in Assessing Risks in Artificial Intelligence", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 382, no. 2285, 2024, p. 2-3; A. Kasirzadeh, "Two types of AI existential risk: decisive and accumulative", Philosophical Studies, vol. 182, 2025, p. 1975.

36) J. Danielsson et al., "Artificial Intelligence and Systemic Risk", Journal of Banking and Finance, vol. 140, 2022, p. 4; Y. Bengio et al. (2025), *op. cit.*, p. 126.

37) P. Slattery et al.,*op. cit.*, p. 35; A. Kasirzadeh, *op. cit.*, p. 1987.

38) A. Kasirzadeh, *op. cit.*, p. 1996.

(2) 우선순위

악용 경로는 자율성 경로보다 긴급하고 실현가능성이 높은 위험이라고 볼 수 있다. 예컨대 AI 기술로 증폭된 생물학적 위협은 즉각적이고 높은 확률의 위험으로 평가된다.[39] Anthropic은 2024년 RSP를 개정하면서 종전과 달리 악용 경로에 좀 더 중점을 두어 구체화하기도 했다.[40] 그러나 자율성 경로는 불가역성과 불확실성의 측면에서 악용 경로보다 능력위험의 특성이 좀 더 강하다. 뒤에서 다시 보겠지만 최근의 기술적 흐름은 자율성 경로의 실현 가능성이 점차 커지고 있음을 보여준다. 개정된 Anthropic RSP 역시 여전히 AI의 자기 개선 능력을 중요한 임계점과 체크포인트로 두고 있다.

악용 경로에 관한 규율 내용이 보다 먼저 구체화되고 있는 이유는 그것이 보다 예측이 용이하고 긴급하기 때문이지 더 중요하기 때문은 아니다. 두 경로 사이에 우선순위가 존재하는 것은 아니다.

4. 능력위험 개념의 규범적 수용

능력위험 개념의 수용은 비법적 규범에서 먼저 시작되었다. 2017년 아실로마 원칙(Asilomar AI Principle) 은 AI의 능력에 주의를 기울일 것을 촉구하면서 "AI 시스템으로 인한 재앙적이거나 실존적인 위험(catastrophic or existential risks)에 대하여 예상되는 영향에 비례하는 계획 및 완화 노력을 기울여야 한다"고 명시하였다.[41] 앤트로픽은 2023년 발표한 RSP 1.0에서 재앙적 위험(catastrophic risk)을 “AI에 의하여 직접적으로 야기된, 수천 명의 인명을 앗아가거나 수백조 달러의 재산 피해를 입히는 대규모 파괴”라고 정의하고 이를 관리하기 위한 프레임워크를 제시하였다.[42]

이후 일부 법적 규범에서도 능력위험 개념의 수용이 이루어졌다. 2023년 바이든 행정부는 행정명령을 통해 대량살상무기 설계, 사이버 공격, 인간 통제 회피 등을 통해 국가안보에 심각한 위험(serious risk)을 초래할 수 있는 대규모 AI 모델을 ’이중용도 기반모델’로 정의하고 개발업체의 안전조치를 의무화했다.[43] 위 행정명령은 2025년 트럼프 행정부에 의하여 무효화되었지만,[44] 이후 발표된 행동계획(AI Action Plan)[45] 역시 프론티어 AI의 강력한 능력에서 파생되는 심각한 위험의 식별, 평가 및 완화 필요성을 강조하고 있다. 한편, 비록 거부권 행사로

39) R. C. de Lima et al., *op. cit.*, p. 3.
40) Anthropic, Responsible Scaling Policy 2.0, 2025.
41) Future of Life Institute, "Asilomar AI Principles", 11 Aug. 2017, point 19, 20.
42) Anthropic(2023), *op. cit.*
43) Exec. Order No. 14110, 88 Fed. Reg. 75191 (Nov. 1, 2023) §3(k), §4.2.
44) Exec. Order No. 14148, Initial Rescissions of Harmful Executive Orders and Actions (Jan. 20, 2025).
45) White House, Winning the AI Race: America’s AI Action Plan (July 23, 2025).

인해 무산되기는 했지만, 캘리포니아 프론티어 AI 법안은 대규모 인명·재산 피해를 유발하는 '치명적 위해(critical harm)'를 정의하고 고비용 대규모 모델에 대한 규율을 시도하기도 했다.[46]

EU AI Act는 범용 AI 모델의 고영향 능력(high-impact capabilities)으로 인해 EU 시장에 중대한 영향을 미치고 가치 사슬 전반에 확산될 수 있는 위험을 '시스템적 위험(systemic risk)'으로 정의하고, 이에 대한 특별한 규율을 하고 있다.[47] 이는 피해의 대규모성을 요하는 캘리포니아 법안과는 다른 접근방법으로서 시스템적 리스크 개념을 수용한 것으로 이해될 수 있다. 한편 EU AI Act에 이어 세계에서 두 번째로 성립된 포괄적 AI 규제 법률인 한국의 인공지능기본법은 학습에 사용된 누적 연산량이 일정 기준 이상인 AI시스템의 경우 안전성을 확보하기 위하여 사업자에게 일정한 의무를 부과한다(제32조). 이 법은 능력위험에 대한 특별한 규율을 담고 있으면서도 정작 능력위험을 정의하고 있지는 않다. 특별한 규율은 특별한 위험으로 인해 정당화되는 것이므로 이론적으로나 실무적으로나 개선할 여지가 있는 입법이다.

지금까지 능력위험의 개념과 발생 경로, 그리고 규범적 수용 현황까지 살펴보았다. 능력위험은 기존의 규율 체계로는 대응하기 어려운 것처럼 보인다. 그렇다면 과연 능력위험은 실제로 존재하는가. 다음 장에서는 이에 관한 이론적, 기술적 검토를 수행한다.

Ⅲ. 능력위험은 실제로 존재하는가

1. 경고론과 회의론

능력위험의 인정 여부를 둘러싼 논쟁은 주로 자율성 경로에 초점을 맞추어 경고론과 회의론으로 나뉘어 전개되어 왔다.

가. 경고론

경고론은 능력위험의 심각성을 강조하며 선제적 대응의 필요성을 주장하는 입장이다. 이는 다시 원리적 경고론과 기술적 경고론으로 세분화될 수 있다. 이러한 구분은 철학적 접근법과 공학적 접근법의 차이를 반영한 것이다.[48]

---

46) §3.22602(g)(1)
47) Recital (110), §3(65), §55.
48) S. Field, "Why do experts disagree on existential risk and P(doom)? A survey of AI experts," arXiv preprint arXiv:2502.14870, Jan. 2025.

(1) 원리적 경고론

원리적 경고론은 AI의 근본적 특성에서 비롯되는 구조적 위험을 강조한다.

Yudkowsky는 능력위험 논의의 선구적인 토대를 제공했다. 그는 재귀적 자기 개선을 통한 지능폭발론을 제기하였으며,[49] 인간에게 우호적인 초지능 설계가 근본적으로 어려움을 지적하기도 했다.[50]

Bostrom은 이를 체계화하여 더욱 발전시켰다. 그는 AI의 지능 수준과 최종 목표가 서로 독립적이며, 지능이 높은 에이전트들이 다양한 최종 목표를 가질 수 있음을 논증하였고(orthogonality thesis: 직교성 테제),[51] 도구적 수렴(instrumental convergence) 개념[52]을 받아들여 서로 다른 목표를 가진 초지능이 유사한 중간 목표를 추구할 가능성을 제시하였다.[53] 자원 획득, 자기 보존, 능력 향상, 그리고 권력 획득과 같은 것은 어떠한 최종 목표의 달성에도 유용하므로 초지능은 이러한 도구적 목표들을 수렴적으로 추구할 것이라는 주장이다. 종이클립 생산을 최대화하라는 목표를 가진 초지능이 인간을 포함한 지구상의 모든 것을 종이클립으로 바꾸어버릴 수 있다는 가상의 시나리오가 이를 잘 설명한다.[54] 또한 Bostrom은 초지능의 통제 방안을 능력봉쇄와 의도선택으로 나누어 분석하기도 했다. 그는 가두기, 유인 제공, 약화, 함정과 같은 능력봉쇄 기법들을 검토한 뒤, 초지능의 잠재적 이점과 능력을 고려할 때 능력봉쇄는 비효율적이고 신뢰할 수 없다고 결론짓고, 가치 기반의 통제 방법이 필요하다고 주장했다.[55]

후속 연구자들은 이들의 이론을 더욱 발전시키고 보완하였다. 특히 실존적 위험에 대한 정량적 분석을 위해 체계적 방법론을 통해 위험 발생 확률을 추정하는 연구들이 주목할 만하다. 예컨대 Ord는 2120년까지 실존적 재앙의 발생 확률을 10%로 추정했으며,[56] Carlsmith는 AI 개발부터 인류 무력화까지 6개의 단계를 설정하여 2070년까지의 실존적 위험을 약 5%로 추정했다.[57]

---

49) E. Yudkowsky, "Staring into the singularity," manuscript, 1996/1999. Available: https://philpapers.org/rec/YUDSIT.

50) E. Yudkowsky, "Creating Friendly AI 1.0: The analysis and design of benevolent goal architectures," Machine Intelligence Research Institute, San Francisco, CA, USA, Tech. Rep., June 2001.

51) S. Armstrong, "General purpose intelligence: arguing the orthogonality thesis," Analysis and Metaphysics, vol. 12, 2013.

52) S. M. Omohundro, "The basic AI drives," in Proc. 1st Conf. Artificial General Intelligence (AGI), Memphis, TN, USA, Mar. 2008.

53) N. Bostrom, "The superintelligent will: motivation and instrumental rationality in advanced artificial agents," Minds and Machines, vol. 22, no. 2, May 2012.

54) 종이클립 최대화기 사고실험이 극단적인 예방주의적 사고의 병폐를 보여주는 사례라고 지적하는 견해로는 B. Goertzel, "Superintelligence: fears, promises and potentials," Journal of Evolution and Technology, vol. 25, no. 2, 2015.

55) N. Bostrom(2014), *op. cit.*, Chapters 9, 12, 13.

56) T. Ord, The Precipice: Existential Risk and the Future of Humanity. Bloomsbury, 2020.

(2) 기술적 경고론

기술적 경고론은 현재 AI 기술의 발전 궤도와 구체적 위험 메커니즘에 주목한다.

Russell은 통제 문제를 가치 정렬(value alignment)이라는 기술적 문제로 치환했다.[58] 그는 고정된 목표 최적화에 기반한 표준 모델의 위험성을 강조하며, 명시된 목표와 진정한 의도의 불일치를 경계했다. 미다스 왕이 "모든 것을 황금으로"라는 소원으로 딸까지 잃은 것처럼, AI의 잘못 설정된 목표 추구가 의도치 않은 피해를 가져올 수 있다는 것이다. 이를 목표 오설정(misspecification)[59] 또는 대리 목표 추구(proxy gaming)[60]의 문제라고 한다. Russell은 이를 해결하기 위해 AI가 지속적으로 인간의 피드백을 구하며 목표를 수정해나가는 '입증 가능한 유익한 AI(provably beneficial AI)'이라는 패러다임을 제안했다.

Russell과 같은 AI 기술 분야의 권위자들이 능력위험을 강조하는 것은 인상적이다. 딥러닝의 선구자인 Hinton은 AI를 "새롭고 더 나은 지능 형태"로 규정한다.[61] 그는 2024년 노벨상 수상 연설에서 안전 연구의 시급함을 역설하였고, 수상 직후 인터뷰에서 "30년 내 인류 멸종 가능성이 10-20%"라고 경고하기도 했다.[62] 2018년 튜링상을 수상한 Bengio는 AI 위험을 팬데믹, 핵전쟁과 동등한 글로벌 우선순위로 설정해야 한다고 주장한다.[63] 그는 현재의 에이전트 중심 개발 방향을 우려하면서, AI로 인한 심각한 위험을 피하려면 전세계적인 협력이 필수적이라고 강조하였고, 실제로 국제 AI 안전 보고서 작성을 주도하기도 했다.[64] 2025년 발표된 설문조사 결과에서는 100여 명의 AI 전문가들 중 78%가 파국적 위험에 대해 우려해야 한다고 응답한 것으로 나타났다.[65]

한편 Christiano는 실용적 정렬 연구를 통해 기술적 해법을 구체화했다. 그가 확립한 인간 피드백 기반 강화학습(RLHF) 이론은 ChatGPT의 탄생에 기여했을 뿐 아니라 AI의 행동을 인간의 가치에 맞게 정렬하는 효과적인 방법이 되었다.[66] 그는 AI 시스템을 안전하게 확장하기 위해

---

57) J. Carlsmith, "Is power-seeking AI an existential risk?," Open Philanthropy, Tech. Rep., Apr. 2021.
58) S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019.
59) Y. Bengio, M Cohen et al., "Superintelligent agents pose catastrophic risks: Can scientist AI offer a safer path?," arXiv:2502.15657, 2025; T. Swoboda et al., "Examining popular arguments against AI existential risk: A philosophical analysis," arXiv:2501.04064, 2025.
60) D. Hendrycks et al., *op. cit.*
61) W. D. Heaven, "Geoffrey Hinton tells us why he's now scared of the tech he helped build," MIT Technology Review, 2 May 2023.
62) Nobel Prize Organization, *op. cit.*
63) Y. Bengio(2023), *op. cit.*
64) Y. Bengio et al. (2025), *op. cit.*
65) S. Field, *op. cit.*
66) P. Christiano, "Thoughts on the impact of RLHF research," AI Alignment Forum, Jan. 2023; Anthropic의 Constitutional AI는 RLHF 기법을 부분적으로 대체하지만 가치 정렬의 기본적 방법론은 그대로 유지되었다. Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," arXiv:2212.08073, Dec. 2022.

인간 전문가의 능력을 증류하고 증폭시키는 방법론(Iterated Distillation and Amplification: IDA)을 개발하기도 했다.[67]

나. 회의론

회의론은 능력위험에 대한 과도한 우려를 경계하며, 보다 현실적이고 증거에 기반한 접근을 주장한다. 이는 원리적 회의론, 기술적 회의론, 우선순위론으로 나누어볼 수 있다.

(1) 원리적 회의론

원리적 회의론은 능력위험 논증의 기본 전제에 대해 근본적 의문을 제기한다.

Chollet는 지능 폭발이 본질적으로 불가능하다고 주장한다.[68] 그는 '모든 문제를 다 잘 푸는 완벽한 해결책은 없다'는 기계학습 분야의 잘 알려진 정리(no free lunch theorum)를 인용하여 지능이 환경과 분리되어 임의로 증가할 수 없다고 지적했다. 제조 로봇, 군사적 확장, 과학과 같은 재귀적 자기 개선 시스템들이 실제로는 선형적 또는 S자 형태의 진보를 보인다는 것이다.

Pinker는 AI 위험 담론이 의인화 오류에 기반한다고 지적하며, 지능을 자율성 또는 동기와 혼동하는 것이 문제라고 주장했다.[69] AI 시스템은 인간이 부여한 목표만을 추구할 뿐 자체적인 의지나 욕망을 가질 수 없으며, 악의적이고 전능한 컴퓨터를 상상하는 것은 인간의 어두운 욕망을 투영하는 의인화 오류에 불과하다는 것이다. 그러나 역설적이게도 개발 현장에서는 Pinker가 경계하는 바로 그 의인화적 요소들을 시스템에 통합하려는 시도가 이어지고 있다. 예를 들어 일부 실험적 AI 시스템들은 호기심 기반 탐색, 내재적 동기, 자율적 목표 설정 등의 기능을 모사하려 한다.[70]

그 밖에 형식 논리학적 분석을 통해 기술적 특이점(singularity)에 대한 예상을 증거가 박약한 확신으로 평가절하하거나,[71] 경고론자들의 주장이 마치 파스칼의 내기와 같다며 비판하는 견해도 있다.[72] 즉, 신이 존재할 확률이 낮더라도 믿지 않으면 무한 지옥에 간다는 17세기 철

---

67) P. Christiano et al., "Supervising strong learners by amplifying weak experts," arXiv:1810.08575, Oct. 2018.
68) F. Chollet, "The implausibility of intelligence explosion," Medium, 27 Nov. 2017.
69) S. Pinker, "The idea that A.I. will lead to the end of humanity is like the Y2K bug," CNBC, 27 Feb 2018.
70) D. Pathak et al., "Curiosity-driven exploration by self-supervised prediction," in Proc. 34th Int. Conf. Machine Learning, Sydney, Australia, 2017; Y. LeCun, "A Path Towards Autonomous Machine Intelligence," OpenReview, ver. 0.9.2, Jun. 27, 2022; B. Liu et al., "AI autonomy: Self-initiated open-world continual learning and adaptation," arXiv:2203.08994, 2023.
71) S. Bringsjord, "Belief in the Singularity is Logically Brittle," Journal of Consciousness Studies, vol. 19, no. 7-8, 2012.
72) O. Etzioni, "How to know if artificial intelligence is about to destroy civilization," MIT Technology Review, 25 Feb. 2020.

학자 파스칼의 논증처럼, AI가 인류를 멸망시킬 확률이 낮더라도 무한한 손실 가능성 때문에 극단적 대비를 해야 한다는 논리는 정반대 행동도 동시에 정당화할 수 있는 결함이 있다는 것이다.

(2) 기술적 회의론

기술적 회의론은 현재 AI 기술의 현실적 한계를 근거로 과도한 우려를 경계한다. 대표적으로 LeCun은 현재의 AI가 물리적 세계 이해, 지속적 기억, 추론, 복잡한 계획 등의 핵심적 능력을 갖고 있지 않으며, AI에 생존 본능과 같은 나쁜 행동을 유발할 욕구를 부여할 이유가 없다고 주장했다.[73] Ng 역시 경고론자들의 우려를 "화성 과잉인구 문제"에 비유하면서, 현실적이지 않은 먼 미래의 위험보다는 현재 해결 가능한 문제에 집중할 것을 강조하였다.[74]

(3) 우선순위론

우선순위론은 능력위험보다 현재의 확실한 위험에 집중해야 한다는 입장이다. Bender 등은 미래의 가상적 AGI 위험에 대한 과도한 관심이 현재의 실제 피해로부터 주의를 분산시킨다면서, 환경, 편향, 기회비용, 환각과 같은 현재의 측정 가능하고 완화 가능한 위험에 집중해야 한다고 주장한다.[75] 그러나 Swoboda는 기업이나 정부가 미래 위험보다 즉각적 피해에 더욱 민감하게 반응함을 밝혀냄으로써 주의 분산에 대한 우려가 과장되었음을 보여주었다.[76]

2. 능력위험의 현실화 가능성

여기에서는 최근의 기술 발전 동향을 점검함으로써 이러한 위험이 단지 이론에 그칠 것인지 아니면 기술적으로 현실화 가능성이 있는 것인지 검토해보기로 한다.

가. 대규모 언어모델(LLM)

AI 기술의 대중화를 이끈 대규모 언어모델은 2022년 ChatGPT 출시 이후 지속적으로 발전

73) Y. LeCun, "Meta's Yann LeCun predicts 'new paradigm of AI architectures' within 5 years," TechCrunch, 23 Jan. 2025; A. Zador and Y. LeCun, "Don't Fear the Terminator," Scientific American Blog Network, 26 Sep. 2019.

74) A. Ng, "AI guru Ng: Fearing a rise of killer robots is like worrying about overpopulation on Mars," The Register, 9 Mar. 2015.

75) E. M. Bender et al., "On the dangers of stochastic parrots: Can language models be too big?," in Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2021, pp. 610-623.

76) T. Swoboda et al., *op. cit.*

하여 이제는 초보적인 형태의 환경 인식 능력, 추론 능력, 기억 능력 등을 갖게 되었다.

최근의 대규모 언어모델들은 텍스트뿐만 아니라 이미지, 음성, 비디오 등 다양한 형태의 정보를 통합적으로 처리할 수 있는 멀티 모달리티 기능을 갖추어 환경 인식 능력이 크게 향상되었다. 추론 능력의 향상 역시 여러 기술적 혁신을 통해 달성되고 있다. 학습 단계가 아닌 추론 단계에 컴퓨팅 자원을 투입하는 전략(test time compute)은 작은 모델이 14배 더 큰 모델을 능가할 수 있게 하며,[77] 전문가 혼합(mixture of experts: MoE) 아키텍처는 적은 계산량으로 모델의 능력을 확장할 수 있게 한다.[78] AI의 기억 능력 역시 다양한 혁신을 통해 크게 발전했다.[79]

대규모 언어모델의 이러한 능력 향상은 위험 지식 접근성 증대, 정교한 허위정보 생성, 사이버 공격 도구 자동화 등을 통해 위해 실행 장벽을 체계적으로 낮춤으로써 능력위험의 악용 경로를 활성화시킨다.[80]

나. 에이전트 AI

에이전트 AI는 대규모 언어모델 기술 발전을 바탕으로, 환경과의 능동적 상호작용과 전략적 행동 수행이라는 새로운 차원의 능력을 추가한다.

최근의 에이전트 AI 개발은 사용자의 파일 시스템에 접근하고, 웹 검색이나 계산 도구 등 다양한 외부 도구를 활용하며, 다른 서비스 및 애플리케이션과 소통하는 등 환경과의 상호작용 및 도구 사용 능력을 강화하는 방향으로 이루어지고 있다.[81] 이는 AI 시스템의 역할이 정보의 능동적 조작자로 변화함을 의미하며, 시스템 무결성과 데이터 보안에 대한 새로운 차원의 위험을 제기한다.[82] 특히 모델 통신 프로토콜(Model Communication Protocol, MCP)을 통한 다중 에이전트 시스템의 구현은 조정 실패, 갈등, 공모 등을 통해 개별 AI 시스템의 능력을 증폭시키는 효과를 가져온다. 이러한 협업과 경쟁 메커니즘은 예측하기 어려운 창발적 행동 패턴과 새로운 형태의 위험을 발생시킬 가능성이 있다.[83]

---

77) C. Snell et al. "Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters", arXiv:2408.03314, 2024.

78) W. Cai et al, "A survey on mixture of experts in large language models", IEEE Transactions on Knowledge and Data Engineering, 2025.

79) 컨텍스트 윈도우 확장, RAG(Retrieval-Augmented Generation), 벡터 데이터베이스, NTM(Neural Turing Machine), DNC(Differentiable Neural Computer) 등 다양한 기술 혁신이 이루어졌다. S. Khosla et al., "Survey on Memory-Augmented Neural Networks: Cognitive Insights to AI Applications", arXiv:2312.06141, 2023.

80) B. C. Das et al., "Security and privacy challenges of large language models: A survey", ACM Computing Surveys vol. 57 no. 6, 2025.

81) N. Krishnan, "AI Agents: Evolution, Architecture, and Real-World Applications", arXiv:2503.12687, 2025.

82) Y. He et al., "Security of ai agents", 2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE), 2025.

에이전트 AI는 목표 지향적이고 전략적인 행동 수행 능력을 지닌다. 대규모 언어모델을 활용한 에이전트들은 명시적 프로그래밍 없이 제한적인 지침만으로도 자율적으로 행동하고 계획을 수립하는 능력을 보여주었다.[84] 에이전트들에게 전문화된 역할을 부여함으로써 서로 협력하여 전략을 발전시키고 진화하도록 할 수도 있었다.[85] 대규모 언어모델에 전통적인 강화학습 기술을 적용함으로써 효율적이고 목표 지향적인 탐색을 구현하는 연구도 이루어졌다.[86]

에이전트 AI의 능력 향상은 자율적 추론을 통한 안전장치 우회(자율성 경로)와 환경 조작을 통한 위해 실행(악용 경로) 위험을 높인다. 비록 능력위험의 범주에서 제외하기는 했지만, 다중 에이전트 시스템의 창발적 행동 패턴을 통해 시스템적 누적 경로에 따른 위험을 발생시킬 수도 있다.[87]

다. 물리적 AI

물리적 AI는 디지털 영역에 한정되었던 AI의 능력을 물리적 세계로 확장함으로써 능력위험의 현실화 가능성을 보다 직접적이고 가시적인 형태로 구현한다.

VLA(Vision-Language-Action) 모델은 로봇이 감각 입력을 해석하고 맥락을 이해하며 실제 환경에서 자율적으로 작업할 가능성을 보여주는 동시에 새로운 형태의 안전 위험을 초래한다.[88] MIT 라이선스로 공개된 OpenVLA는 폐쇄형 모델인 RT-2-X 대비 7분의 1 파라미터로도 16.5% 높은 작업 성공률을 달성함으로써 기술 효율성과 접근성을 동시에 혁신하여 물리적 AI의 광범위한 확산 가능성을 보여주었다.[89] 물리 시뮬레이터와 AI의 결합은 AI의 물리적 세계 이해 능력 향상을 가속화한다.[90] 9천조 개의 토큰으로 훈련된 세계 기반 모델(world foundation model: WFM)인 NVIDIA Cosmos는 가상 환경의 미래 상태를 예측하고 생성하여 차세대 로봇과 자율주행 자동차 개발을 지원하며 오픈소스(Apache 2.0 라이선스)로 제공된다.[91]

---

83) L. Hammond et al., "Multi-Agent Risks from Advanced AI", arXiv preprint arXiv:2502.14143, 2025.
84) C. Gao et al, "Large language models empowered agent-based modeling and simulation: A survey and perspectives", Humanities and Social Sciences Communications vol. 11 no. 1, 2024.
85) N. Belle et al., "Agents of Change: Self-Evolving LLM Agents for Strategic Planning", arXiv:2506.04651, 2025.
86) 언어적 입력을 통해 에이전트의 행동을 동적으로 조정하고, LLM의 도메인 지식을 활용하여 보상 함수를 동적으로 생성함으로써 희소한 보상 환경에서도 효과적인 탐색이 가능하다. C. Sun et al., "LLM-based multi-agent reinforcement learning: Current and future directions", arXiv:2405.11106, 2024.
87) E. Miehling et al., "Agentic AI Needs a Systems Theory", arXiv:2503.00237, 2025.
88) B. Zhang et al., "SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning," arXiv:2503.03480, 2025.
89) M. J. Kim et al., "OpenVLA: An Open-Source Vision-Language-Action Model," arXiv:2406.09246, 2024.
90) C. K. Liu and D. Negrut, "The Role of Physics-Based Simulators in Robotics," Annual Review of Control, Robotics, and Autonomous Systems, vol. 4, pp. 35-58, 2021.

효율적인 물리적 AI의 접근 가능성 확대는 악용 위험을 높이고, VLA 모델의 자율성은 인간 통제권 상실 위험을 높임으로써 능력위험이 물리적 세계로 확장될 가능성을 보여준다.

라. 오픈소스 모델의 약진

최근 고성능 오픈소스 AI 모델들이 기존의 상업적 서비스와 경쟁할 수 있는 수준에 도달하면서 AI 생태계의 접근성과 개방성을 근본적으로 변화시키고 있다.

Meta의 Llama 4 시리즈는 네이티브 멀티모달 기능과 전례 없는 컨텍스트 지원을 제공한다(Llama 4 Community License, 사용자 수 제한 및 경쟁 모델 개선 활용 금지로 인해 OSI 미승인).[92] 프랑스 Mistral의 Small 3.1 모델은 24억 파라미터만으로 GPT-4o Mini를 능가하는 효율성을 달성했다(Apache 2.0 라이선스).[93] 중국 DeepSeek의 V3 모델은 MoE 방식을 채택하여 선도적인 폐쇄 모델에 필적하는 성능을 달성했으며, R1 모델은 OpenAI o1에 비견되는 추론 능력을 단 600만 달러의 비용으로 구현했다(각 MIT 라이선스).[94] 또한 알리바바의 Qwen 2.5-Max는 GPT-4o를 능가하는 성능을 보여준다(향후 오픈소스화 계획 시사).[95] 후발 주자들의 오픈소스 전략에 위기감을 느낀 OpenAI 역시 2025년 8월 도구 사용 및 웹 검색 기능을 갖춘 고성능 오픈 웨이트 모델인 gpt-oss를 출시하였다(Apache 2.0 라이선스).[96]

오픈소스 모델은 AI 개발 대중화, 투명성 증대, 경제적 불평등 해소 등 광범위한 혁신 기회를 제공하지만 악용 가능성도 증가시킨다.[97] 탈옥 공격이나 미세 조정을 통해 안전하지 않은 콘텐츠 생성이 가능하며, 일단 공개되면 취약점을 소급하여 수정할 수 없다.[98] 예컨대 DeepSeek의 경우 전문지식 없이도 랜섬웨어 등 악성코드 생성이 가능하며, 실제로 사이버 범죄 네트워크가 악성코드 제작에 활용하고 있다.[99] 이에 따라 오픈소스 모델도 고위험 AI 시스

---

91) NVIDIA Newsroom, "NVIDIA Launches Cosmos World Foundation Model Platform to Accelerate Physical AI Development," Jan. 6, 2025; NVIDIA Technical Blog, "Advancing Physical AI with NVIDIA Cosmos World Foundation Model Platform," Jan. 23, 2025.

92) Meta AI, "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation," Meta AI Blog, Apr. 5, 2025.

93) Mistral AI, "Mistral Small 3.1," Mistral AI News, Mar. 17, 2025.

94) DeepSeek-AI, "DeepSeek-V3: A strong Mixture-of-Experts language model with 671B total parameters," arXiv:2412.19437, Dec. 2024.; DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv:2501.12948, Jan. 2025.

95) Qwen Team, "Qwen2.5-Max: Exploring the Intelligence of Large-scale MoE Model," Qwen Blog, Jan. 28, 2025.

96) Hugging Face, "Welcome GPT OSS, the new open-source model family from OpenAI!", 5 Aug. 2025.

97) E. Seger et al., "Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives," Centre for the Governance of AI, Oxford, UK, Sep. 2023.

98) F. Eiras et al., "Risks and Opportunities of Open-Source Generative AI," arXiv:2405.08597, May 2024.

99) M. Pearl, et al.,"Delving into the Dangers of DeepSeek," Center for Strategic and International Studies,

템처럼 규제되어야 한다는 주장이 제기되기도 한다.[100] 하지만, 과도한 위험성 강조는 AI 선도 기업들의 엄격한 규제 요구를 정당화하여 오픈소스 생태계 발전을 제약할 수 있다는 우려도 존재한다.[101]

마. AGI 개발 시도

대규모 언어모델을 중심으로 하는 현재의 AI 패러다임을 근본적으로 혁신하여 일반인공지능(AGI)이나 초지능(superintelligence)를 개발하려는 시도 역시 활발히 이루어지고 있다. 가장 대표적인 것은 LeCun이 주창하는 세계 모델(world model)이다. 그는 "세계에 대한 내부 모델을 학습하고, 호기심과 내재적 동기를 통해 스스로 학습하는" 시스템(JEPA)을 추구하며, "진정한 지능은 단순한 다음 토큰 예측을 넘어 세계를 이해하고 예측하는 능력"이라고 강조한다.[102]

주요 기업들의 구체적 개발 계획도 가시화되고 있다. 구글은 Gemini 2.5 Pro를 "뇌가 하는 것처럼 세계를 이해하고 시뮬레이션하는 세계 모델"로 확장한다고 발표했다.[103] 딥마인드는 Genie 3을 통해 물리 법칙을 스스로 학습하여 지속적 상호작용 세계를 구현하는 세계 모델 패러다임을 제시하면서, 그것이 AGI로 가는 초석이라고 밝혔다.[104] 오픈AI는 AGI가 "풍요로움을 늘리고 글로벌 경제를 가속화할 것"이라며 5단계 로드맵을 제시했고,[105] 메타는 Meta Superintelligence Labs 설립을 발표하며 "초지능 개발이 시야에 들어왔다"고 밝혔다.[106] 초지능 개발을 통한 미국의 전략적 우위 확보를 위하여 정부 주도 대규모 개발 프로젝트가 필요하다는 주장이 제기되기도 한다.[107]

전문가의 예측은 급격히 단축되고 있다. 2025년에 발표된 방대한 조사 결과에 따르면 AI 연구자들은 2040년까지 AGI 도달 확률을 50%로 추정하고 있으며,[108] 예측 전문가들의 AGI 도달 시점 전망이 4년 만에 50년 후에서 5년 후로 급격히 단축되었다는 조사 결과도 있다.[109] 주

Washington, D.C., Feb. 2025.

100) D. E. Harris, "How to Regulate Unsecured 'Open-Source' AI: No Exemptions," TechPolicy.Press, Dec. 6, 2023.

101) F. Eiras et al., "Risks and Opportunities of Open-Source Generative AI," arXiv:2405.08597, May 2024.

102) Y. LeCun, "A Path Towards Autonomous Machine Intelligence," Meta AI Research, Dec. 2024.

103) Google, "Introducing Gemini 2.5 Pro: Advanced Multimodal AI for Complex Reasoning," Google AI Blog, Jan. 2025.

104) DeepMind, "Genie 3: A new frontier for world models," DeepMind Official Blog, Aug. 5, 2025.

105) OpenAI, "Planning for AGI and beyond," 2025.

106) CNBC, "Mark Zuckerberg announces creation of Meta Superintelligence Labs. Read the memo," Jun. 30, 2025.

107) Gladstone AI, "America's Superintelligence Project," 2025.

108) AI Multiple, "AGI Predictions: When Will Artificial General Intelligence Be Achieved?," AI Multiple Research, Jan. 2025.

109) 80,000 Hours, "AGI Timeline Predictions: Expert Survey Results," 80k Hours Research, Jan. 2025.

요 AI 기업의 CEO들 역시 수 년 내에 AGI 개발을 공언하고 있다.[110]

바. 평가

현재 상용화된 대규모 언어모델의 정교한 텍스트 생성과 오픈소스 모델 확산은 위해 실행 장벽을 상당히 낮추었다. 에이전트 AI와 물리적 AI는 능력위험을 질적으로 변화시킬 수 있다. 환경과의 상호작용 능력을 갖춘 AI는 인간 통제권의 상실 위험을 높이며, 다중 에이전트 시스템은 창발적 행동을 통해 새로운 위험을 창출할 수 있다. AGI 개발 과정에서 인간의 수준을 초월하는 AI가 나타나 통제를 벗어나게 된다면 현재의 틀로는 포착하기 어려운 실존적 위험을 야기할 가능성도 완전히 배제하기 어렵다. AI 기술의 발전 궤적은 능력위험이 비록 불확실하지만 실재하는 위험임을 시사하며, 그 규율을 위한 체계가 필요함을 보여준다.

3. 논쟁의 교착과 비생산성

AI 능력위험을 둘러싼 논쟁은 교착 상태에 빠져 정책적 진전을 가로막고 있다. 경고론은 잠재적 재앙 가능성을 근거로 선제적 대응을 주장하고, 회의론은 증거 부족을 이유로 성급한 대응을 경계한다. 하지만 양측 모두 극단적 경고나 비판에만 집중하여 실용적인 규율 체계 설계 지침을 제공하지 못하고 있다. 능력위험의 실현가능성이 어느 정도 윤곽을 드러낸만큼 능력위험의 규율에 자원을 투입할 것인지 말 것인지의 이분법적 논쟁은 극복되어야 한다.[111] 이제 필요한 것은 능력위험의 구체적인 규율을 위한 견고한 이론적 토대이다.

Ⅳ. 능력위험은 어떻게 규율되어야 하는가

1. 능력위험의 독자적 규율을 위한 이론적 기반의 필요성

가. 독자적 규율의 필요성

앞서 본 것처럼 능력위험은 맥락위험과 근본적으로 다른 특성을 갖는다. 이러한 차이는 서로 다른 규율 체계를 요구한다.

능력위험은 특정 활용 영역에 국한되지 않는다(맥락 독립성). 고도의 능력을 갖춘 AI 시스템

110) AI Multiple, "When Will AGI/Singularity Happen? 8,590 Predictions Analyzed," 2025.
111) T. Swoboda et al., *op. cit.*

은 특정한 맥락을 전제하지 않더라도 일정한 위험을 내포할 수 있다. 따라서 시나리오에 기반한 영역별 규제는 근본적 한계를 보이며, 특정 영역에 국한되지 않는 수평적이고 횡단적인 규율이 요구된다.

능력위험은 사용자 의도가 아닌 AI 능력 자체에서 기인한다(능력 귀인성). 따라서 사용 단계에서의 통제만으로는 근본적 해결이 불가능하며, 레드티밍(red teaming)과 같이 능력 자체에 초점을 둔 개발 단계에서의 개입이 필요해진다.[112]

능력위험은 피해의 양상과 범위를 예측하기 어렵다(불확실성). 따라서 정량적 위험 평가에 의존하는 기존의 접근법은 효과적으로 작동하지 못하며, 탐색과 정보 획득을 중심으로 하는 유연한 규율이 요구된다.[113]

능력위험의 불확실성에도 불구하고 규율을 위한 자원 투입이 정당화되려면 결과의 대규모성이 요구된다. 따라서 규율 대상으로서의 능력위험은 피해의 대규모성 내지 한번 발생하면 되돌릴 수 없는 특성을 갖는다(불가역성). 이는 사후 대응 중심의 접근을 무의미하게 만들며 사전예방적 접근을 필요로 한다.

나. 능력위험과 맥락위험의 중첩

능력위험의 독자적 특성에도 불구하고 때로는 능력위험과 맥락위험이 중첩되는 상황이 발생한다. 본래 의도되었던 맥락 밖에서 악용되거나 예상하지 못했던 능력이 발현될 수 있기 때문이다.

능력위험과 맥락위험의 중첩이 발생하는 이유는 무엇보다 AI 기술의 범용성 때문이다. 이는 주로 악용 경로와 관련된다. 훈련에 필요한 데이터셋이 있다면 언어, 컴퓨터 코드, DNA, 화학 구조와 같은 상징적 성격의 모든 프로세스를 '언어 모델'로 만들 수 있으며, 본래 의도하지 않았던 악의적인 용도로 전용할 수 있다.[114] 독성을 피하도록 설계된 모델이 독성 분자를 생성하도록 쉽게 변경될 수 있다는 실증 사례는[115] 치료 목적의 분자 설계 능력이 생화학무기 개발로 직접 악용될 수 있는 가능성을 보여준다. 합성 생물학과 AI 기술의 결합을 통해 전염성 높은 슈퍼바이러스가 만들어질 가능성도 제기되고 있다.[116]

---

112) IBM Research, "What is red teaming for generative AI?", 21 Jan. 2025.

113) B. Howell, "The Precautionary Principle, Safety Regulation, and AI: This Time, It Really Is Different," American Enterprise Institute, 2024.

114) A. M. Barrett, et al., "Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models," UC Berkeley Center for Long-Term Cybersecurity, 2024, p. 17.

115) A. M. Barrett, et al., *op. cit.*, p. 17.

116) R. C. de Lima et al., *op. cit.*; 핵무기와 생물학 무기와 같은 물질적 AI 위험의 주된 장벽은 AI를 통한 정보 향상보다는 통제된 물질과 장비의 획득이라고 지적하는 견해로는 F. Eiras et al., "Risks and opportunities of open-source generative AI," arXiv:2405.08597, May 2024.

예측 불가능한 능력의 발현도 중첩 발생의 원인이 된다. 이는 주로 자율성 경로와 관련된다. 대규모 언어 모델의 스케일링 과정에서 예상하지 못한 능력의 창발 현상이 보고되고 있다.[117] 일부 연구자는 본래 의도와 다른 방식으로 목표를 달성하는 현상(프록시 게이밍)이 통제 불가능한 '불량AI(rogue AI)'가 등장할 가능성을 보여준다고 주장한다.[118] AI 모델의 최적화 과정에서 자기보존 본능이 나타나 인간에 의한 비활성화를 막고 궁극적으로는 자기보호를 위해 인간을 통제하려 할 수 있다는 주장도 있다.[119]

이러한 중첩 상황에 대응하기 위해서는 맥락위험 규율과 능력위험 규율의 독자성을 확보하면서도 이들을 연계하는 방안이 마련될 필요가 있다.

다. 이론적 기반 구축 필요성

이제 능력위험에 대한 독자적 규율의 필요성이 명확해졌지만, 실제 규율 체계를 구축하기 위해서는 견고한 이론적 토대가 필요하다. 기존의 능력위험 논의는 교착 상태에 빠져 구체적 규율을 위한 원리를 충분히 제공하지 못한다. 능력위험의 특성들은 중요한 출발점이지만, 이것만으로는 효과적이면서도 정당한 규율 체계를 설계하기에 불충분하다.

이번 장에서는 사회적, 윤리적, 법적 차원에서 능력위험에 대한 규율을 위한 이론적 기반을 구축하고자 한다. 사회적 차원에서는 고도의 불확실성을 사회적으로 관리하는 방안에 대한 통찰을, 윤리적 차원에서는 불확실성 하에서의 윤리적 원리에 관한 지혜를, 법이론적 차원에서는 불확실성과 불가역성의 딜레마를 해결하기 위한 규제 구조에 관한 아이디어를 얻을 수 있을 것이다.

2. 사회적 기반

능력위험 규율을 위한 이론적 기반의 첫 번째는 사회적 측면에서 찾을 수 있다. 여기에서는 루만(N. Luhmann)의 위험사회학을 토대로 능력위험의 불확실성에 대응하기 위한 다원적 거버넌스 체계의 필요성에 관하여 논의한다.

가. 위험의 사회적 구성

루만은 위험(Risiko, risk)과 위난(Gefahr, danger)의 구분을 통해 현대 사회의 불확실성을 분

117) J. Wei et al., *op. cit.*
118) D. Hendrycks, et al., *op. cit.*
119) L. Berti, et al., "Emergent Abilities in Large Language Models: A Survey," arXiv preprint arXiv:2503.05788, Mar. 2025.

석하는 틀을 마련하였다. 사회가 처한 위험을 범주화한다는 아이디어는 벡(U. Beck)의 위험사회론에서 비롯되었지만,[120] 루만은 그의 체계이론을 통해 보다 강력한 개념으로 재탄생시켰다.[121]

루만의 체계이론에서 위험과 위난의 구별은 그의 독특한 사회 이해방식에서 비롯된다. 루만은 사회를 개인들의 집합이 아니라 소통의 연쇄로 파악한다. 여기서 소통이란 정보를 선택하여 전달하고 그것을 이해하는 복합적 과정이다.[122] 이러한 소통들이 이어지면서 사회적 체계가 형성되며, 이 체계는 스스로를 환경으로부터 구별하면서 작동한다.[123] 동일한 미래의 불확실성이라도 사회적 체계가 그것을 어떻게 관찰하고 의미를 부여하느냐에 따라 전혀 다른 성격을 갖게 된다. 사회가 어떤 불확실성을 자신의 선택으로 통제 가능한 것으로 구성하면 그것은 위험이 되고, 통제 불가능한 외부 요인으로 구성하면 위난이 된다. 이러한 위험의 사회적 구성은 객관적 실체의 차이가 아니라 사회가 불확실성을 어떻게 이해하고 대응하느냐의 문제임을 보여준다.[124]

루만에 따르면, "위험의 경우에만 의사결정이 역할을 한다. 위난에는 노출될 뿐이다."[125] 이는 의사결정과 행위주체성(agency)의 중요성을 강조한 것으로 해석될 수도 있다.[126] 최근의 연구는 행위주체성에 더해 사회적 행위자의 '의도성'이라는 요소를 추가로 도입하여 위난(danger), 위험(risk), 위협(threat)의 세 가지 분류를 제안함으로써 보다 효과적인 위험 관리를 모색하기도 한다.[127]

나. 능력위험의 합리적 규율 가능성

현재 능력위험에 대한 논의는 두 가지 극단적 관점으로 양극화되어 있다. 회의론자들은 "불확실하니 아무것도 할 수 없다"는 입장을 취한다. 반대로 경고론자들은 "재앙이 너무 크니 개발을 중단해야 한다"고 주장한다. 이들은 모두 능력위험을 통제 불가능한 위난으로 인식하여

120) 벡에 따르면 현대사회의 위험(Risiko)은 대규모 환경오염과 같이 인류의 생존을 위협하는 거대한 위협으로서 전통적인 국소적 위해(Gefahr)와 구별된다. 양천수, "위험, 재난 및 안전 개념에 대한 법이론적 고찰," 공법학연구 제16권 제2호, 2015, 192면.

121) 이철, 이철, "니클라스 루만의 체계이론적 세계사회 분석: 소통으로서의 사회적인 것과 소통의 총체로서의 사회," 사회와 이론, 제30집, 2017, 171면.

122) 이철, 위의 논문, 181면.

123) 루만은 모든 체계가 환경과의 차이를 통해서만 존재할 수 있다고 본다. 이철, 앞의 논문, 174-175면.

124) 벡 역시 위험에 대한 구성주의적 입장을 취한다. 박희제, "위험사회에서 세계시민주의로: 울리히 벡의 (기술)위험 거버넌스 전망과 한국의 사회학", 사회사상과 문화 제30집, 2014, 94면

125) N. Luhmann, Risk: A Sociological Theory, de Gruyter, 1993.

126) 다만 루만 자신은 주체나 실체를 전제하는 기존의 사유 방식을 버리고, 소통의 자기생산적 연쇄를 통해 사회적 체계가 스스로를 구성한다고 보았다. 이철, 앞의 논문, 181-182면.

127) F. Battistelli and M. G. Galantino, "Dangers, risks and threats: An alternative conceptualization to the catch-all concept of risk," Current Sociology, vol. 67, no. 1, 2019.

합리적인 대응 방안을 제시하지 못한다. 능력위험을 관리 가능한 위험으로 전환한다면, 전면 금지나 완전 방임 대신 불확실성을 점진적으로 줄이면서 기술 발전의 이익을 활용하는 균형잡힌 접근이 가능하다.

능력위험의 문제가 큰 불확실성 하에 있는 것은 사실이지만 의사결정을 마냥 지연하는 것은 부적절하며 이러한 불확실성은 정책 과정에서 관리될 수 있다.[128] 이러한 입장은 선스타인(C. R. Sunstein)의 합리적 위험 규제론과도 연결된다.[129] 선스타인은 위험에 대한 일시적 공포나 히스테리에 기반한 대응이 아니라 비용-편익 분석에 기반한 보다 합리적인 위험 규제 시스템의 필요성을 강조한다.

능력위험을 관리 가능한 위험으로 전환하려면 정보 공유와 소통을 통한 사회적 위험의 구성 과정이 필수적이다. 기업, 연구자, 정책결정자, 일반 시민 등 이해관계자들은 고유한 소통 체계 내에서 능력위험을 서로 다르게 인식하고 구성한다. 경고론자와 회의론자의 논쟁이 평행선을 달리는 것은 단순한 견해 차이가 아니라, 서로 다른 소통체계에서 능력위험을 근본적으로 다르게 구성하고 있기 때문일 수 있다.

다. 정치체계와 소통

근대사회는 다양한 기능체계로 분화되었다. 그 중 정치체계는 개인의 안전에 대한 욕구를 사회적으로 해결하는 과정에서 권력 위임이라는 소통 방식을 통해 생성되었다. 루만의 관점에서 정치체계는 '집합적 구속력 있는 결정'이라는 기능을 수행하며, 이를 위해 다른 기능체계들을 환경으로 관찰한다.[130] 능력위험의 관리 가능한 위험으로의 전환은 정치체계의 관찰과 결정을 통해 실현된다.

능력위험의 맥락에서 정치체계가 관찰해야 할 환경은 다층적이다. 과학체계는 진리/비진리 코드를 통해 능력위험의 과학적 불확실성과 기술적 가능성을 평가하며, 경제체계는 지불/비지불 코드를 통해 능력위험 규율의 경제적 비용과 편익을 산정한다.[131] 정치체계는 서로 다른 기능체계의 관찰들을 정치적 결정으로 통합하며, 그 과정에서 능력위험을 과학적 불확실성이나 경제적 후생의 문제를 넘어 사회적 선택의 문제로 구성한다. 즉, 통제 불가능한 위난에서 관리 가능한 위험으로 전환하는 것이다.

정치체계가 효과적으로 환경을 관찰하기 위해서는 각 기능체계의 소통에 접근할 수 있어야

128) M. Nordström, "AI under great uncertainty: implications and decision strategies for public policy," AI & Society, vol. 37, no. 4, pp. 1708-1711, 2022.
129) C. R. Sunstein, Risk and Reason: Safety, Law, and the Environment. Cambridge: Cambridge University Press, 2002.
130) 이철, 앞의 논문, 172, 179, 186, 191-194면, 199면 참조.
131) 이철, 앞의 논문, 191면 참조.

한다.[132] 문제는 AI 기술 발전의 속도와 투명성 부족으로 인해 정치체계의 관찰 역량이 제한적이라는 점이다. 따라서 정치체계의 관찰 역량을 강화하기 위한 구조가 마련될 필요가 있다. 예를 들어, 정책 결정 과정에 AI 전문가의 참여나 산업계의 의견 청취를 제도화함으로써 과학체계나 경제체계의 소통에 접근할 수 있을 것이다.

라. 다원적 거버넌스

능력위험을 둘러싼 과학적 불확실성과 기술적 복잡성에도 불구하고, 정치체계는 이를 관리 가능한 위험으로 구성할 수 있다. 이러한 정치적 구성은 집합적 구속력 때문만이 아니라, 다른 기능체계들의 소통에도 영향을 미치기 때문에 실질적 효과를 갖는다. 경제체계는 정치적으로 구성된 능력위험을 비용 요소로 반영하여 시장에서 처리하고, 과학체계는 진리 추구라는 고유 목적을 유지하면서도 정치적 결정을 연구의 사회적 맥락으로 고려한다. 법체계 역시 정치체계의 결정을 합법/불법 코드를 통해 재처리하여 구체적인 법적 규범으로 변환한다.

이는 결국 정치체계를 중심으로 한 기능분화된 사회의 통합적 대응 체계를 의미한다. 단일한 통제 중심이나 완전한 시장 자율에 의존하지 않는, 분화된 기능체계들 간의 상호 관찰과 조정을 통한 다원적 거버넌스가 형성되는 것이다. 이러한 접근은 능력위험의 불확실성에 대응하면서도 각 기능체계의 고유한 작동 논리를 존중한다. 정치체계의 역할을 인정하되, 과학체계의 지식 생산, 경제체계의 효율성 추구, 법체계의 규범적 일관성을 모두 고려하는 균형잡힌 관점을 취할 수 있는 것이다.

3. 윤리적 기반

능력위험 규율을 위한 이론적 기반의 두 번째 축은 윤리적 기반이다. 사회적 기반은 소통과 다원적 거버넌스에 관한 통찰을 제공하지만 개별 주체가 어떻게 행위해야 하는지에 대하여는 답을 주지 못한다. 그러한 규범적 틀 없이는 능력위험 규율 체계가 제대로 작동할 수 없다. 여기서는 능력위험 특유의 고도 불확실성에 대한 윤리적 대응 방안을 살펴본다.

가. 행위 윤리의 한계와 행위자 윤리의 요청

윤리학은 전통적으로 행위 윤리와 행위자 윤리를 구분해 왔다. 행위 윤리는 특정 행위의 옳고 그름을 판단하는 데 초점을 맞춘다. 행위가 가져오는 결과로 그 행위를 평가하는 결과주의

132) 이는 정치체계가 다른 체계들에 직접 개입하는 것이 아니라, 각 체계의 소통 결과를 '자극'으로 수용하여 정치적 관찰 대상으로 포착한다는 의미이다. 루만의 구조적 결합 개념에 관하여는 N. Luhmann, Social Systems, Stanford: Stanford University Press, 1995, pp. 115-116.

윤리와 행위 자체의 의무적 성격을 중시하는 의무론적 윤리가 이에 포함된다. 반면 행위자 윤리는 아리스토텔레스의 덕 윤리 전통에 따라 행위자의 성품과 덕성을 윤리적 평가의 근본으로 삼는다.[133]

개인주의가 보편화된 근대에 들어와 행위 윤리는 행위자 윤리를 압도하였다. 그러나 고도의 불확실성으로 인해 결과 예측이 매우 어려운 능력위험 상황에서 행위 윤리는 한계에 직면한다. 결과주의 윤리학의 경우 "좋은 결과"를 사전에 정의하기 어렵고, 확률적 위험 평가의 어려움으로 효용 계산이 곤란하며, 인간 생명이나 환경 같은 측정 불가능한 가치들을 화폐 가치로 환산하려는 시도는 이러한 가치들을 저하시킨다.[134] 의무론적 윤리학 또한 문제가 있다. 칸트적 의무론은 새로운 상황을 규율하기에는 지나치게 세부적이고, 실용적 적용을 위해서는 지나치게 모호한 딜레마에 빠진다.[135]

이와 달리 행위자 중심적인 덕 윤리는 능력위험 특유의 고도의 불확실성 하에서도 잘 작동한다. 덕 윤리는 고정된 원칙에 의존하지 않고 행위자가 실천적 지혜(phronesis)를 바탕으로 맥락과 특수성을 고려하여 새로운 상황에서도 윤리적 불확실성에 효과적으로 대응할 수 있게 한다.[136] 행위자의 판단력을 활용함으로써 이분법적 판단이나 경직된 계산 하에서는 적절히 포섭되지 못하는 경계선상의 다양한 윤리적 문제들을 다룰 수 있는 것이다.[137] 이러한 실천적 지혜는 규칙 학습이 아닌 경험과 성찰을 통해 개발된다.[138] Hagendorff는 광범위한 조사를 통해 AI 개발 과정에서 요구되는 덕목 체계를 정의, 정직, 책임, 돌봄이라는 기본 덕목과 신중함과 용기라는 2차적 덕목으로 구체화하였다.[139]

나. 기계 윤리의 필요성과 구현 방법

앞의 논의는 주로 AI 개발자에게 요구되는 윤리에 관한 것이다. 그러나 능력위험 상황에서는 AI 자체의 윤리도 중요해진다. AI의 자율성과 적응성이 극대화되면서 발생하는 고도의 불

133) W. Wallach et at., "Machine morality: Bottom-up and top-down approaches for modeling human moral faculties," AI & Society, vol. 22 no. 4, 2008, p. 576; T. Hagendorff, "A Virtue-Based Framework to Support Putting AI Ethics into Practice," Philosophy & Technology, vol. 35 no. 3, article 55, 2022, p. 4.

134) A. Siapka, "A Virtue Ethics Approach to AI-induced Risk," Teoria. Rivista di filosofia, vol. 44, no. 2, 2024, pp. 103-105.

135) N. Smith and D. Vickers, "Living well with AI: Virtue, education, and artificial intelligence," Theory and Research in Education, vol. 22 no. 1, 2024, pp. 26-27, 31.

136) P. Hayes et al., "From applied ethics and ethical principles to virtue and narrative in AI practices," AI and Ethics, vol. 5, 2025, p. 1225.

137) M. Constantinescu et al., "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context," Ethics and Information Technology, vol. 23 no. 4, 2021, p. 807.

138) K. Kristjánsson, et al.,"Phronesis (Practical Wisdom) as a Type of Contextual Integrative Thinking," Review of General Psychology, vol. 25 no. 3, 2021, p. 252.

139) T. Hagendorff, *op. cit.*, pp. 4-14.

확실성과 예측 불가능성으로 인해 AI 자체에도 기계윤리를 적용할 필요성이 대두된다. AI가 개발자의 의도와 다른 내부 목표를 학습하거나(mesa-optimization) 훈련 시와 배치 시에 다른 행동 패턴을 보이는(deceptive alignment) 사례는 이를 방증한다.[140]

Wallach와 Allen은 인공적 도덕 행위자(artificial moral agent: AMA) 개념을 제안하면서 그 발전 단계를 자율성과 윤리적 민감성이라는 두 차원으로 나누어 분석하였다.[141] 실행적 도덕성(operational morality)은 자율성과 민감성이 모두 낮지만 설계에 가치가 구현된 시스템으로서, 어린이 보호 장치를 갖춘 총이 그 예이다. 기능적 도덕성(functional morality)은 두 가지 유형으로 나뉘는데, 높은 자율성과 낮은 민감성을 가진 자동조종장치와 같은 시스템과 그 반대의 특성을 가진 시스템이 있다. 기능적 도덕성 영역의 시스템들은 도덕적으로 중요한 측면을 어느 정도 평가할 수 있는 능력을 갖는다. 완전한 도덕적 행위자(full moral agency)는 높은 자율성과 윤리적 민감성을 모두 갖춘 도덕적 주체로서, 현재로서는 인류가 보유하지 못한 기술이다.

고도의 자율성과 적응성을 지닌 AI의 능력위험에 대응하기 위해서는 윤리적 민감성을 높이는 작업이 필요하다. 그 방법으로는 상향식 접근법과 하향식 접근법이 있다.[142] 상향식 접근법은 윤리 이론을 목표 설정에만 활용하고, 실제 구현이나 제어 구조는 윤리 이론에 의존하지 않는다. 반면 하향식 접근법은 기존 윤리 이론을 바탕으로 알고리즘과 시스템을 설계하는 방식이다.

상향식 접근법은 덕 윤리적 요소를 반영하기에 적합하다. 패턴의 처리에 강점을 보이는 기계학습의 특성상 정직, 배려, 신중함과 같은 덕목이나 상황적 판단 능력을 제한적이나마 구현할 가능성이 있기 때문이다. Russell이 제시한 '증명 가능하게 유익한 AI(provably beneficial AI)' 개념 역시 상향식 접근에 가깝다.[143] 다만 덕 윤리적 요소를 반영하더라도 성찰과 실천을 요구하는 실천적 지혜는 구현할 수 없으므로 AI에 도덕적 책임을 묻기는 어렵다.[144] 한편 하향식 접근법은 행위 윤리의 한계를 그대로 물려받는다. 명시적 규칙을 프로그래밍하는 방식은 복잡한 상황에서 규칙을 적용하는 데 한계가 있고, 예외 상황이나 규칙 충돌 문제를 야기한다.

---

140) E. Hubinger et al., "Risks from Learned Optimization in Advanced Machine Learning Systems," arXiv:1906.01820, Jun. 2019.

141) W. Wallach and C. Allen, Moral machines: Teaching robots right from wrong. Oxford University Press, 2008, p. 25.

142) W. Wallach et at., *op. cit.*, pp. 568-569.

143) Russell은 기계가 (1) 자신만의 목적이나 자기보호 욕구 없이 오직 인간 가치 실현을 목적으로 하되, (2) 인간이 진정 원하는 것이 무엇인지 완전히 확신하지 않은 상태를 유지하면서, (3) 인간의 선택과 행동을 관찰하여 점진적으로 학습함으로써 인간에게 증명 가능하게 유익한 AI를 구현할 수 있다고 제안했다. S. Russell, "Provably beneficial artificial intelligence," in Proc. 27th Int. Conf. Intelligent User Interfaces (IUI), Helsinki, Finland, Mar. 2022, pp. 8-9.

144) M. Constantinescu et al., *op. cit.*, p. 809.

다. 통합적 윤리 기반과 다층적 거버넌스

개발자 윤리와 기계 윤리는 상호 보완적 관계에 있다. 개발자 윤리는 개발 및 배치 단계에서 윤리적 고려와 가치 반영에 중요한 역할을 한다. 능력위험 특유의 고도 불확실성 하에서는 규칙 기반 접근법보다 실천적 지혜를 강조하는 덕 윤리 접근법이 더 효과적이다. 하지만 AI 시스템의 자율성 증가에 따라 배치 후의 능력위험 통제를 위한 기계 윤리 구현이 필요해진다. 기계 윤리 역시 상향식 접근에 의한 덕 윤리적 요소 구현이 더 효과적이다. 다만 현재 AI는 완전한 도덕적 행위자가 아니므로 개발이나 배치에 관여하는 인간의 책임에 더 주의를 기울여야 한다.

행위 윤리와 행위자 윤리도 상호 보완적 접근이 중요하다. 덕 윤리는 의무론을 배척하지 않으며, 의무론의 도덕 규범들은 선한 삶을 지향할 때 지켜야 할 행동의 제약조건이 된다.[145] 자율성 경로의 경우 위험 메커니즘이 불분명하므로 신중함과 책임감 같은 행위자 윤리가 강조되고, 악용 경로의 경우 상대적으로 예측 가능하므로 구체적 행위 규범이 더 중요해진다. 이러한 상호보완적 관계는 능력위험 거버넌스에서 법적 규제와 비법적 규제의 유기적 결합 근거를 제공한다. 윤리학에서 의무론적 규범이 수용 가능한 행동의 형식적 경계를 규정하고 덕 윤리가 그 안에서 선한 삶이라는 실질적 목표를 추구하듯이, 규제 영역에서도 법적 규범이 형식적 경계를 설정하고 자율규제가 그 기반 위에서 더 높은 차원의 안전을 추구하는 다층적 거버넌스 구조가 마련될 수 있다.

4. 법이론적 기반

능력위험의 독자적 규율을 위한 이론적 기반 구축에서 법이론적 접근은 핵심적 위치를 차지한다. 능력위험의 특성에 실효적으로 대응하고 사회적·윤리적 기반의 통찰을 강제력 있는 실제 규율로 구현하려면 법이론적 기반이 필요하기 때문이다.

가. 사회적 · 윤리적 기반의 법적 수용

(1) 사회적 기반: 정보와 소통 중심의 규율

사회적 기반의 분석을 통해 얻은 위험/위난의 구별과 소통 및 다원적 거버넌스에 관한 통찰은 다음과 같은 법제도적 구현을 요구한다.

첫째, 능력위험의 법적 정의와 규율 대상 명시가 필요하다. 능력위험을 법적 관리의 대상이

145) P. Hayes et al., *op. cit.*, p. 1226.

되는 위험의 영역으로 전환하려면 이를 법체계 내에 명확히 수용해야 한다. 앞서 보았듯이 EU AI Act는 시스템적 위험을, 캘리포니아의 프론티어 AI 법안은 치명적 위해를 명확히 정의하고 있다.

둘째, 소통을 통한 불확실성 감소를 위하여 적절한 투명성 조치의 법제화가 필요하다.[146] 정치체계의 관찰 역량을 강화하려면 투명성 부족 문제를 해결해야 한다. EU AI Act는 시스템적 위험이 있는 범용 AI 모델 제공자에게 시스템적 위험의 평가및 완화, 중대한 사고의 보고 등의 의무를 부과하며(§55.1, 부속서 XI), 한국 인공지능기본법 역시 고성능 인공지능 사업자에게 위험의 식별 및 완화, 위험관리체계 구축 등의 의무를 부과하고 있다(제32조 제1항).

셋째, 다원적 거버넌스를 위한 소통 체계를 제도화해야 한다. 예컨대 정부-민간 정보 공유 플랫폼, 국제적 정보 교환 메커니즘, 전문가 집단의 독립적 평가 체계 등을 통해 기능체계 간 정보 공유와 소통이 촉진될 수 있을 것이다.

(2) 윤리적 기반: 법적 규율과 비법적 규율의 조화

윤리적 기반을 통해 제시된 다층적 거버넌스에 관한 통찰을 법적으로 구현하려면 다음과 같은 구조가 필요하다.

첫째, 덕 윤리적 접근의 효과성은 법적 규제와 비법적 규제의 유기적 결합으로 구현될 수 있다. 예를 들어 EU AI Act는 시스템적 위험을 지닌 범용 AI 모델의 규율을 위하여 실무규약(code of practice)을 활용한다(§56).[147]

둘째, 행위자 윤리와 행위 윤리의 상호보완성은 능력위험의 발현 경로에 따라 차별적으로 구현될 수 있다. 자율성 경로는 위험 메커니즘이 불분명하므로 행위자 윤리 중심의 간접적 제도가 효과적이다. 개발자 윤리 교육이나 내부 거버넌스 구축 지원을 통해 신중함과 책임감 같은 덕성을 함양하는 것이다. 반면 악용 경로는 예측이 상대적으로 가능하므로 명백히 수용 불가능한 행위에 대해서는 행위 윤리에 기반한 직접적인 법적 의무 부과가 가능하다.

셋째, 개발자 윤리와 기계 윤리의 상호보완성은 개발자의 윤리적 역량 강화 제도와 AI 시스템의 윤리적 제약 준수 제도의 결합으로 구현된다. 특히 AI의 윤리적 제약 미준수는 공법적 규제 대상이 되거나 사법적 '하자' 내지 '상품 적합성 결여'로 평가될 수도 있을 것이다.[148]

---

146) M. Anderljung, et al, "Frontier AI regulation: Managing emerging risks to public safety," arXiv:2307.03718, 2023, pp. 17-18.

147) 능력위험에 관한 것은 아니지만, 한국 인공지능기본법이 검인증을 받은 인공지능 제품 또는 서비스를 정부 조달에 우선 고려하도록 한 것도 참고할 만하다(§30④).

148) 이상용, "인공지능과 계약법," 비교사법 제23권 제4호, 2016, 1692면.

나. 맥락독립성과 능력귀인성에 대한 대응

(1) 맥락독립성에 대한 대응

능력위험의 맥락독립적 특성은 기존의 영역별 수직적 규제로는 해결할 수 없는 법적 도전을 제기한다. 맥락위험은 활용 맥락을 전제로 하므로 영역별 수직적 규제가 가능할 뿐만 아니라 바람직하기도 하다.[149] 반면 능력위험은 잠재적 능력에 기초하므로 맥락에 기반한 영역별 규제가 불가능하고 수평적 규제에 의할 수밖에 없다.

다만 악용 경로에서와 같이 능력위험과 맥락위험이 중첩되므로, 양자에 대한 규율은 독립적이면서도 연계될 필요가 있다. 맥락위험 규율만으로는 임계점 돌파를 감지할 수 없고, 능력위험 규율만으로는 구체적 발현 양상을 예측하기 어렵다. 따라서 맥락위험 모니터링을 통한 조기 경보 시스템과 능력위험 대응을 위한 사전예방 체계의 연계를 법제도에 반영할 필요가 있다.

(2) 능력귀인성에 대한 대응

능력위험이 AI의 능력 자체에서 비롯된다는 특성은 활용 맥락을 전제한 전통적 규제와는 다른 법적 접근을 요구한다. 능력위험의 규제는 AI의 능력에서 발생하는 위험의 경로에 초점을 맞추어야 한다. 즉, 위험 발생 경로에 대한 통제 가능성의 정도에 비례하여 관련 당사자에게 권한과 책임을 배분할 필요가 있다.

다만 기술 자체에 대한 비합리적 규제의 위험에 빠지지 않도록 유의해야 한다. AI의 능력은 위험의 원천이 될 수도 있지만 사회적 난제들을 해결할 잠재력도 갖는다. 따라서 위험과 편익에 대한 균형감각을 유지하는 것이 중요하다.[150] 능력위험 평가시 고려 요소들을 명시하는 것은 불합리한 규제의 확산을 방지하는 한 방법이 될 수 있다.[151]

다. 불확실성과 불가역성의 딜레마에 대한 대응

(1) 불확실성과 불가역성의 딜레마

능력위험에서 가장 어려운 문제는 불확실성과 불가역성을 동시에 해결하기 어렵다는 딜레마에서 비롯된다. 맥락위험의 경우 자동차나 의료기기와 같이 사람의 생명이나 신체에 대한

149) 이상용(2024), 앞의 논문, 27면.
150) F. G'sell, "Regulating Under Uncertainty: Governance Options for Generative AI," Stanford Cyber Policy Center, 2024, p. 10.
151) M. Anderljung, et al, *op. cit.*, pp. 34-35.

위험을 수반하는 등의 경우에 한하여 시장 출시 전 규제가 이루어진다.[152] 이에 비하여 능력위험은 개념적으로 피해의 대규모성, 즉 불가역성을 내포하므로 활용 맥락과 무관하게 사전배려의 원칙(precautionary principle)[153]에 따른 사전예방적 규제가 정당화될 수 있다.

문제는 능력위험이 지닌 고도의 불확실성이다. 능력위험의 결과가 어떠한 양상으로 어느 정도로 나타날지 신뢰성 있게 예측하기는 매우 어렵다. 이처럼 불확실한 위험에 대비하여 자원을 투입하는 것이 과연 합리적이고 정당한가? 전통적인 위험 기반 규제의 원칙은 규제 자원을 가장 큰 위험에 집중시켜 효율성을 높이라고 말하지만[154] 이 상황에서 크게 도움이 되지 않는다. 능력위험이 어느 정도로 큰 위험인지 객관적으로 평가하기 어렵기 때문이다. AI 규제에 원자력 규제의 방법론을 그대로 적용할 수 있다는 주장[155]이 설득력이 부족한 것은 그 때문이다. 원자력 위험과 달리 AI의 능력위험은 피해의 양상과 정도를 신뢰성 있게 예측하는 것이 극히 어렵다.

이러한 딜레마를 해결하기 위한 한 가지 방법은 능력위험의 발생에 그럴 듯한 확률을 부여하는 것이다. Ord와 Carlsmith는 체계적 방법론에 기반하여 실존적 재앙의 발생 확률을 각각 10%와 5%로 추정했다.[156] 하지만 그들의 방법론에 동의하지 않는 사람들을 설득하기에는 역부족일 것이다. 다른 방법은 전문가의 직관을 활용하는 것이다. 지난 2024년 Hinton, Bengio, Russel 등 수많은 석학들은 기고문을 통해 AI 예산의 3분의 1을 안전 연구에 할당해야 한다고 주장했다.[157] 그러나 이 역시 구체적인 해결책으로 이어지는 설득력 있는 논증이 되지는 못한다.

(2) 효과적인 준비태세를 위한 규제 구조

우리는 좀 더 실용적인 접근 방법을 취할 필요가 있다. 불확실성과 불가역성의 딜레마가 단일한 법적 원리로 해결되기 어렵다는 점을 인정하고, 효과적인 준비태세(preparedness) 구축에 집중해야 한다.

이러한 준비태세는 능력위험의 이중적 특성에 따라 차별화된 대응 요소들로 구성되어야 한

---

152) 이상용, 앞의 논문(2024), 9-10면.

153) 사전배려의 원칙이란 위험이 불확실하더라도 손해가 중대하고 회복 불가능한 경우 위험 방지 또는 축소 조치를 취할 수 있다는 법적 원칙을 말한다. 이상용, 앞의 논문(2024), 9-10면.

154) J. Black, “The emergence of risk-based regulation and the new public risk management in the United Kingdom”, Public law (2005), pp. 510-514, 531.

155) S. Cha, "Towards an international regulatory framework for AI safety: lessons from the IAEA’s nuclear safety regulations," Humanities and Social Sciences Communications, vol. 11, art. no. 506, Apr. 2024, p. 10.

156) T. Ord, The Precipice: Existential Risk and the Future of Humanity. Bloomsbury, 2020; J. Carlsmith, "Is power-seeking AI an existential risk?," Open Philanthropy, Tech. Rep., Apr. 2021.

157) Y. Bengio et al., "Managing extreme AI risks amid rapid progress," Science, vol. 384, no. 6698, May 2024, p. 844.

다. 불가역성에 대응하기 위해서는 선제적 차단을 위한 사전예방 체계, 즉시 개입할 수 있는 신속성, 실효적인 강제력이 필요하다. 불확실성에 대응하기 위해서는 지속적인 정보 획득, 상황 변화에 따른 적응성, 부적절한 조치를 되돌릴 수 있는 수정가능성이 보장되어야 한다. 이러한 요소들을 구현하려면 다양한 법적 장치들을 유기적으로 활용할 필요가 있다.

첫째, 실체적 규제보다 절차 중심의 규제가 유용하다. 새로운 정보를 반영하여 빠르게 변화하는 기술에 유연하게 대응할 수 있고 부적절한 조치의 수정이 용이하기 때문이다. 다만 절차 자체로는 실효성 있는 사전예방이나 신속한 대응이 확보되지 않는다는 한계가 있다.

둘째, 민간과의 공동규제(co-regulation) 내지 규제된 자율규제(regulierte Selbstregulierung)를 활용하는 방법이다.[158] 기술 발전이 빠르고 전문성이 중요한 AI 분야에서는 민간이 더 정확한 정보를 보유하고 변화에 빠르게 적응할 수 있다.[159] 다만 민간 이익과 공익의 충돌 가능성과 강제력 확보를 위한 추가 법적 장치가 필요하다는 한계도 지닌다.

셋째, 불확실성과 불가역성의 정도를 반영하여 단계적 규제를 할 수 있다. 1단계는 정보 확보(정보 제공과 기본 안전 조치 의무화), 2단계는 임시조치(능력위험 임박 시 AI 작동 중단과 정밀 평가), 3단계는 위험 제거(안전장치 부가부터 불능화까지 다양한 조치)로 구성된다. 이는 확보된 정보에 부합하는 유연한 대응과 신속한 예방 조치를 동시에 가능하게 한다.

넷째, 능력위험 규율에는 민사법이나 형사법보다 행정법 체계가 적합하다. 대체로 행정법상 책임은 민형사상 책임과 달리 주관적 요건을 요구하지 않아서 고도의 불확실성 하의 책임 추궁을 가능하게 하며, 재량과 판단여지는 유연한 적응적 대응과 부적절한 조치의 수정을, 행정청의 처분 권한과 자기집행력은 신속한 예방조치를 가능하게 한다.[160]

Ⅴ. 결 론

AI의 성능이 급속히 향상되면서 기존의 규율체계로는 관리하기 어려운 새로운 차원의 위험이 대두되었다. 극단적 경고론과 회의론 간의 논쟁이 교착상태에 빠진 상황에서, 본 연구는 AI 위험의 본질을 재정의하고 합리적 규율을 위한 이론적 토대를 구축하고자 하였다.

그 내용은 다음과 같다. 첫째, AI의 자율성, 적응성, 불투명성이라는 특성을 바탕으로 AI 위험을 맥락위험과 능력위험으로 구분할 수 있다. 맥락위험이 특정 활용 영역과 맥락을 전제로

158) 이승민, "온라인 플랫폼과 자율규제 논의의 기초," 경제규제와 법, 제15권 제2호, 서울대학교 공익산업법센터, 2022, 50-56면.
159) 이상용, 앞의 논문(2020), 138면.
160) 다만 행정법상의 규제는 광범위한 스펙트럼을 포함하고 있어 모든 행정법상의 규제에 위와 같은 분석이 적용되는 것은 아니다.

하는 반면, 능력위험은 활용 맥락과 무관하게 AI의 잠재적 능력 자체에서 비롯된다. 그에 따라 능력위험은 맥락 독립성, 능력 귀인성, 불확실성과 불가역성이라는 특징을 갖게 된다. 둘째, 경고론과 회의론 간의 논쟁을 체계적으로 검토하고 현존 AI 기술의 발전 양상을 분석한 결과 능력위험이 불확실하지만 실재하는 위험임을 확인할 수 있었다. 셋째, 능력위험의 독자적 특성에 부합하는 규율을 위해서는 사회적, 윤리적, 법적 차원의 이론적 기반을 마련할 필요가 있다. 검토 결과 루만의 위험사회학을 활용한 소통 중심의 다원적 거버넌스, 불확실성 하에서의 덕윤리 접근법, 그리고 불확실성과 불가역성의 딜레마를 해결하기 위하여 다양한 법적 수단을 활용한 효과적 준비태세 구축이 효과적임을 알 수 있었다.

본고는 능력위험이라는 새로운 위험 범주를 이론적으로 정립하고 기존 논쟁의 교착을 극복할 수 있는 이론적 기반을 구축하고자 하였다. 앞으로 이를 기반으로 실제 규율 사례 분석을 통해 효과적인 능력위험 규율체계를 모색할 수 있기를 기대한다. ▧

(논문접수 : 2025. 08. 15. / 심사개시 : 2025. 10. 02. / 게재확정 : 2025. 10. 20.)

참 고 문 헌

Ⅰ. 한국 문헌

박희제, "위험사회에서 세계시민주의로: 울리히 벡의 (기술)위험 거버넌스 전망과 한국의 사회학," 사회사상과 문화 제30집, 2014.

양천수, "위험, 재난 및 안전 개념에 대한 법이론적 고찰," 공법학연구 제16권 제2호, 2015.

이상용, "알고리즘 규제를 위한 지도 - 원리, 구조, 내용," 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020.

이상용, "인공지능과 계약법," 비교사법 제23권 제4호, 2016.

이상용, "인공지능 규제의 두 가지 접근법 - 맥락 기반 규제와 능력 기반 규제," 사법 제70호, 사법발전재단, 2024.

이승민, "온라인 플랫폼과 자율규제 논의의 기초," 경제규제와 법, 제15권 제2호, 서울대학교 공익산업법센터, 2022.

이철, "니클라스 루만의 체계이론적 세계사회 분석: 소통으로서의 사회적인 것과 소통의 총체로서의 사회," 사회와 이론, 제30집, 2017.

Ⅱ. 외국 문헌

1. 단행본

Bostrom, N., Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.

Luhmann, N., Risk: A Sociological Theory, de Gruyter, 1993.

Russell, S., Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019.

Sunstein, C. R., Risk and Reason: Safety, Law, and the Environment, Cambridge: Cambridge University Press, 2002.

Wallach, W. and Allen, C., Moral machines: Teaching robots right from wrong, Oxford University Press, 2008.

2. 논문

Anderljung, M., et al, "Frontier AI regulation: Managing emerging risks to public safety," arXiv:2307.03718, 2023.

Armstrong, S., "General purpose intelligence: arguing the orthogonality thesis," Analysis and Metaphysics, vol. 12, 2013.

Bengio, Y., "AI and catastrophic risk," Journal of Democracy, vol. 34, no. 4, Oct. 2023.

Bengio, Y. et al., "International AI Safety Report," UK Department for Science, Innovation and Technology, DSIT 2025/001, 2025.

Bostrom, N., "The superintelligent will: motivation and instrumental rationality in advanced artificial agents," Minds and Machines, vol. 22, no. 2, May 2012.

Bubeck, S. et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," arXiv:2303.12712, 2023.

Carlsmith, J., "Is power-seeking AI an existential risk?," Open Philanthropy, Tech. Rep., Apr. 2021.

Chollet, F., "The implausibility of intelligence explosion," Medium, 27 Nov. 2017.

Christiano, P. et al., "Supervising strong learners by amplifying weak experts," arXiv:1810.08575, Oct. 2018.

Critch, A. and Russell, S., "TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI," arXiv:2306.06924, 2023.

Hendrycks, D., et al., "An Overview of Catastrophic AI Risks," arXiv:2306.12001, 2023.

Kasirzadeh, A., "Two types of AI existential risk: decisive and accumulative," Philosophical Studies, vol. 182, 2025.

LeCun, Y., "A Path Towards Autonomous Machine Intelligence," OpenReview, ver. 0.9.2, Jun. 27, 2022.

Omohundro, S. M., "The basic AI drives," in Proc. 1st Conf. Artificial General Intelligence (AGI), Memphis, TN, USA, Mar. 2008.

Russell, S., "Provably beneficial artificial intelligence," in Proc. 27th Int. Conf. Intelligent User Interfaces (IUI), Helsinki, Finland, Mar. 2022.

Slattery, P. et al., "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence," arXiv:2408.12622, 2025.

Swoboda, T. et al., "Examining popular arguments against AI existential risk: A philosophical analysis," arXiv:2501.04064, Jan. 2025.

Wallach, W. et at., "Machine morality: Bottom-up and top-down approaches for modeling human moral faculties," AI & Society, vol. 22 no. 4, 2008.

Wei, J. et al., "Emergent Abilities of Large Language Models," arXiv:2206.07682, 2022.

Yudkowsky, E., "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in Global Catastrophic Risks, edited by N. Bostrom and M. Ćirković, 2008, Oxford University Press, 2008.

Zeng, Y. et al., "AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies," arXiv:2406.17864, 2024.

3. 기타

Anthropic, “Responsible Scaling Policy 1.0”, 2023.

Future of Life Institute, "Asilomar AI Principles," 11 Aug. 2017.

NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, 2023.

OECD, "Explanatory Memorandum on the updated OECD Definition of an AI System," 2024.

Abstract

Governing Capability Risks in Frontier AI: Theoretical Foundation

Lee, Sang-Yong

As frontier AI performance has rapidly improved, new dimensions of risk have emerged that are difficult to manage with existing regulatory frameworks. This study aims to redefine the nature of AI risks and establish a theoretical foundation for rational governance.

This paper categorizes AI risks into context risks and capability risks based on AI's characteristics of autonomy, adaptability, and inscrutability. Context risks are those understood in relation to specific application domains and contexts, characterized by context dependence, use attribution, and predictability. In contrast, capability risks are those understood based on AI's potential capabilities regardless of usage context, characterized by context independence, capability attribution, and uncertainty combined with irreversibility. Capability risks arise through three pathways: misuse, autonomy, and systemic accumulation, though the systemic accumulation pathway has more characteristics of complexity risks than capability risks.

This study analyzes the debate between alarmist (principled and technical) and dissenting (principled, technical, and prioritarian) positions and examines the development patterns of existing AI technologies. Through the multimodal expansion and reasoning capability improvements of large language models, the environmental interaction capabilities of agent AI, the real-world extension of physical AI, the proliferation of high-performance open-source models, and the acceleration of AGI development attempts, this research confirms that capability risks are uncertain but real risks.

This paper argues that the distinctive characteristics of capability risks necessitate new approaches distinct from existing context risk governance systems and establishes theoretical foundations across three dimensions. As a social foundation, it utilizes Luhmann's risk sociology to present the importance of communication and the necessity of pluralistic governance. As an ethical foundation, it demonstrates that virtue ethics is more effective

than deontological and consequentialist ethics under uncertainty and presents the need for both developer ethics and machine ethics. As a legal theoretical foundation, it presents methods for building effective preparedness to resolve the dilemma between uncertainty and irreversibility.

Key Words : AI, Frontier AI, AI Regulation, AI Safety, Capability Risk, AI Law