● ● ● https://doi.org/10.54698/kl.2026.21.837
지식과 교양, 제21호, 교양교육혁신연구센터.
Knowledge and Liberal Arts, Vol. 21, LIC, pp.837-868.

AI 연구의 발전과 사회적 책임

: 기술에서 통제와 거버넌스까지*

오 상 훈**

목 차

1. 서론

2022년 12월 OpenAI가 ChatGPT를 발표한 이후로 거대언어모델(LLM. Large Language Model)의 활용은 인공지능(AI, Artificial Intelligence) 전공자 뿐만 아니라 일반인들의 생활 속에 깊숙이 스며들게 되었다.[1] ChatGPT의 기반이 되는 GPT는 지속적으로 성능을 개선하고

* 본 논문은 최근 AI와 관련하여 일어나고 있는 일들을 정리한 다섯 가지 프레임의 명칭 부여와 각 영역의 내용 보완을 위해 생성형 AI(ChatGPT GPT-5, Perplexity)를 보조적으로 활용하였으며, 모든 내용은 저자가 검토·검증하였다.

** 목원대학교 AI응용학과 교수

1) ChatGPT, https://openai.com/chatgpt, 2026.04.23.

있으며, 구글은 2025년 11월에 추론 능력, 멀티모달 이해, 에이전트 기능 등을 지닌 Gemini 3.0을 발표하여 GPT-5와 경쟁 관계를 형성하였다.[2] 또한, 2025년 1월 중국 딥시크(Deepseek)가 LLM 모델 딥시크-R1을 발표하여 미국 거대 기술 기업 위주의 인공지능 판도를 크게 흔들어놓았다.[3] 2026년 2월 발표된 중국 바이트 댄스의 AI 영상 생성 모델 '시댄스 2.0'은 프롬프트 몇 줄로 생성한 짧은 영상이 미국 할리우드 영화 산업계에 충격을 주고 있다.[4] 이제 전 세계의 많은 국가와 기업들이 미래의 핵심 기술인 AI 기술을 확보하기 위하여 전력을 기울이고 있다.

AI 기술이 LLM 위주의 언어 활동을 위한 영역을 넘어서, 사람의 대리자 역할을 수행하는 AI 에이전트(Agent)도 개발되고 있으며, 그 응용 영역을 과학 연구로 넓혀가고 있다. 실로 AI가 세상의 변혁을 이끌어 갈 것이라는 기대가 전 세계적으로 가득하다. 그렇지만, 이러한 기대와 달리 현실에는 아직 많은 간극이 존재하며, AI 기술의 위험성 역시 우리 곁에 등장하고 있다. 이에, AI 기술을 안전하게 다루기 위한 통제와 거버넌스(Governance)에 대한 논의가 깊어지고 있다.

이 논문에서는 AI 기술의 핵심인 트랜스포머(transformer)를 간략히 소개하고 트랜스포머가 언어모델의 영역을 넘어서 영상, 시계열 예측 등으로 적용 범위를 넓혀가고 있음을 설명한다. 또한, 편향(bias)을 유발하는 데이터 불균형 문제의 접근법들도 알아본다. AI의 응용이 과학 연구 자동화의 영역으로 확대되고 있는 대표적인 사례로서 구글이 개발한 부과학자(Co-Scientist), 알파이볼브(AlphaEvolve), 알파지놈(AlphaGenome)을 살펴본다. 이와 같이 AI 기술이 세상을 변혁시키어 나가는 가능성에 대한 기대와 현실의 간극도 알아본다. AI 기술의 위험성과 관련하여,

2) 구글 코리아 블로그-기업소식, https://blog.google/intl/ko-kr/company-news/technology/google-gemini-3/, 2025.11.19.

3) 위키백과, https://ko.wikipedia.org/wiki/딥시크, 2026.04.23.

4) 박종진, "딥시크 이어 시댄스…중국 AI에 미국 또 '설 연휴 충격'", <전자신문>, 2026.2.15.

LLM 모델이 지닌 환각(Hallucination)의 구조적 원인과 AI 안전성의 구체적인 영향도 설명한다. 마지막으로, AI 시스템의 동작을 모니터링할 수 있는 지에 대한 논의와 안전한 AI 모델의 배포를 위한 AI 레드팀 활동을 설명하고, 세계보건기구(WHO)에서 발표한 보건의료 분야의 윤리 및 거버넌스를 알아본다. 즉, 최근 AI와 관련하여 일어나고 있는 일들을 기술 발전, 응용, 기대와 현실, 위험과 안전성, 그리고 통제와 거버넌스의 다섯 단계 프레임으로 정리하였다.

2. AI 기술의 발전

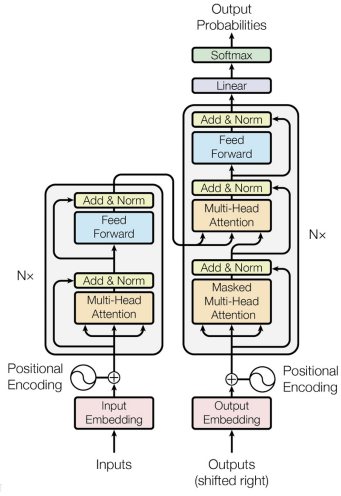
1) 트랜스포머

LLM의 핵심 기술인 트랜스포머는 기존의 시퀀스 변환 구조인 인코더-디코더를 따르면서 어텐션(Attention)에 기반한 모델이다. 이 모델의 우수한 성능은 그 적용 영역을 자연어처리(NLP, Natural Language Processing)에서 넓혀나가, 영상에 적용된 비젼 트랜스포머(ViT, Vision Transformer)와 다변량 시계열 데이터에 적용된 시계열 트랜스포머(TST, Time Series Transformer) 등이 제시되었다. 여기서는 NLP의 핵심기술로 자리 잡은 트랜스포머의 구조와 동작 원리를 살펴보고, 영상처리 ViT와 다변량 시계열 트랜스포머 TST도 알아본다. 마지막으로 인간의 신중하며 분석적인 사고인 “시스템 2 사고”를 모방하여 모델의 학습과 사고 능력을 동시에 확장할 수 있는 새로운 패러다임으로 제시된 에너지 기반 트랜스포머(EBT, Energy-Based Transformers)에 대하여 알아본다.

(1) 트랜스포머의 구조와 동작원리

시퀀스 변환을 위해 복잡한 구조를 완전히 없애고, 오직 어텐션 메커니즘만을 기반으로 새롭게 제시된 간단한 네트워크 아키텍처가 트랜스포머이다.[5)]

대부분의 신경망 기반 시퀀스 변환 모델은 인코더-디코더 구조를



[그림 1] 트랜스포머 구조

가지고 있다. 인코더는 입력 시퀀스 $\boldsymbol{x}=(x_1,x_2,...,x_n)$를 받아, 연속적인 시퀀스 $\boldsymbol{z}=(z_1,z_2,...,z_n)$로 변환한다. 그런 다음, 디코더는 이 z를 바탕으로 기호들로 이루어진 출력 시퀀스 $\boldsymbol{y}=(y_1,y_2,...,y_m)$을 한 번에 하나씩 생성한다. 각 단계에서 모델은 자기회귀 방식으로 작동하며, 이전에 생성된 기호들을 다음 기호를 생성할 때 추가 입력으로 사용한다. 트랜스포머는 전체적으로 이러한 아키텍처를 따르며, 인코더와 디코더 모두에 대해 스택 형태의 셀프 어텐션과 포인트 와이즈(point-wise) 완전 연결 계층을 사용한다. 그림 1의 왼쪽이 인코더이고 오른쪽은 디코더이다.

먼저, 인코더는 동일한 구조가 N층(N=6)으로 구성된 스택이다. 각 층은 두 개의 서브 레이어로 이루어져 있다. 첫 번째 서브 레이어는 멀티-헤드 셀프 어텐션(MHSA, Multi-Head Self Attention) 메커니즘 층이며, 두 번째 서브 레이어는 위치별 완전 연결 전방향 신경회로망이다. 각 서브 레이어에는 잔차 연결 적용 후 레이어 정규화(LayerNorm)가 수행된다. 즉, 해당 서브 레이어 함수 $\mathrm{Sublayer}(\boldsymbol{x})$가 계산된 후에 각 서브 레이어의 출력이 $\mathrm{LayerNor}(\boldsymbol{x}+\mathrm{Sublayer}(\boldsymbol{x}))$과 같이 잔차 연결과 결합 된 형태이다.

디코더 역시 동일한 구조의 N층(N=6)으로 구성된 스택이다. 각 인코더 층에 포함된 두 개의 서브 레이어 외에, 디코더는 세 번째 서브 레이어를

5) Vaswani, A. et al. "Attention Is All You Need", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, pp.6000-6010.

추가로 포함하며, 이는 인코더 스택의 출력에 대해 멀티-헤드 어텐션을 수행한다. 인코더와 마찬가지로, 각 서브 레이어에서 잔차 연결 적용 후 레이어 정규화가 수행된다. 또한 디코더 스택의 셀프 어텐션 서브 레이어는 현재 위치가 이후 위치를 참조하지 못하도록 마스킹을 사용하였다. 이 마스킹은 출력 임베딩이 한 위치만큼 오른쪽으로 이동(offset)되는 것과 결합되어, 위치 i의 예측이 위치 i보다 이전 위치들의 출력에만 의존하도록 한다.

특히, 트랜스포머는 각각 서로 다른 변환을 h번 수행하는 MHSA을 적용하는데, 이 방식은 서로 다른 표현 서브 공간에서 서로 다른 위치의 정보에 주의를 기울일 수 있도록 해준다. 셀프 어텐션 기반 아키텍처인 트랜스포머는 NLP에서 최선호 모델이며, 대규모 텍스트 코퍼스로 사전 학습(pre-training)을 수행한 후, 작은 크기의 작업(task) 특화 데이터 셋으로 미세 조정하는 방식으로 사용되고 있다. 트랜스포머의 계산 효율성과 확장성 덕분에, 1000억 개가 넘는 파라미터를 가진 초대형 모델도 학습하는 것이 가능해졌다.

(2) 비젼 트랜스포머, 시계열 트랜스포머, 에너지 기반 트랜스포머

NLP에서의 트랜스포머 성공 사례에 착안하여 표준 트랜스포머를 가능한 최소한의 수정만으로 컴퓨터 비젼 분야에 적용한 모델이 비젼 트랜스포머 ViT[6]이다. 즉, 이미지를 작은 패치들로 나눈 뒤, 이 패치들의 선형 임베딩 시퀀스를 트랜스포머에 입력한다. 이때 이미지 패치들은 NLP에서의 토큰과 동일하게 취급된다.

이를 위해 이미지 $\boldsymbol{x} \in R^{H \times W \times C}$를 평탄화된 2차원 패치들의 시퀀스 $\boldsymbol{x}_p \in R^{N \times (P^2 C)}$로 변환한다. 여기서 (H, W)는 원본 이미지의 해상도이고, C는 채널 수, (P, P)는 각 이미지 패치의 해상도이다. $N = HW/P^2$는 생성되는 패치의 수로, 이는 트랜스포머의 실질적인 입력 시퀀스 길이

6) Dosovitskiy, A. et al. “An image is worth 16x16 words: transformers for image recognition at scale”, 2021, arXiv:2010.11929v2.

로 사용된다. ViT를 큰 데이터(1,400만~3억 개 이미지)로 사전 학습시킨 후 데이터가 적은 작업에 전이 학습하면 우수한 성능을 보인다.

시계열 데이터 트랜스포머 TST[7]는 입력 디노이징 목적 함수를 통해 트랜스포머 모델을 먼저 학습시켜, 다변량 시계열의 밀집 벡터 표현을 추출한다. 이렇게 사전 학습된 모델은 이후 회귀, 분류, 결측치 보간, 예측 등 다양한 작업에 적용될 수 있다. TST의 핵심은 트랜스포머의 인코더만을 사용하고 디코더 부분은 사용하지 않는 것이며, 다변량 시계열 데이터가 트랜스포머와 호환되도록 변경시킨 것이다.

이제 에너지 기반 트랜스포머 EBT[8]에 대하여 알아보자. 심리학은 인간의 사고를 "시스템 1 사고"와 "시스템 2 사고"로 분류한다. 시스템 1 사고는 빠르고 직관적이며 자동적인 반응이 특징으로, 이전의 경험에 의존하여 단순하거나 익숙한 문제를 해결한다. 반면, 시스템 2 사고는 느리고 신중하며 분석적인 사고로, 복잡한 정보를 처리하기 위해 의식적인 노력과 논리적 추론이 필요하다. 시스템 2 사고는 수학, 프로그래밍, 다단계 추론 등과 같이 자동적인 패턴 인식을 넘어서는 복잡한 문제를 해결할 때 필수적이며, 정밀성과 깊이 있는 이해가 중요한 경우에 해당한다. 현재 AI 모델은 시스템 1 사고에 적합한 과제에서는 뛰어나지만, 시스템 2 능력이 요구되는 과제에서는 여전히 다음과 같은 세 가지 측면의 어려움을 겪고 있다. 먼저, 계산 자원의 동적 할당으로, 인간이 직면하는 과제의 난이도는 매우 다양하기 때문에, 과제에 할당하는 계산 자원의 양을 조절하는 능력이 필요하다. 그 다음에, 연속 상태 공간에서의 불확실성 모델링으로, 인간은 결정에 도달하기 전에 자신이 얼마나 불확실한지를 고려한다. 세 번째는 예측의 검증으로, 언제 사고를 멈출지 또는 가장 정확한 예측을 선택할지에 대한 결정을 내리는 데 중요한 역할을 하며, 실제로, 예측 검증

7) Zerveas, G. et al. "A Transformer-Based Framework for Multi-Variate Time Series Representation Learning", 2020, arXiv:2010.02803v3.

8) Gladstone, A. et al. "Energy-Based Transformers are Scalable Learners and Thinkers", 2025, arXiv:2507.02092v1.

이 인간 사고의 핵심 구성 요소이다.

이 세 가지 측면을 달성하고 시스템 2 사고를 실현하기 위하여 제시되었던 에너지 기반 모델(EBM, Energy-Based Model)은 입력 변수들의 각 구성(configuration)에 에너지 값을 부여함으로써, 문맥과 후보-예측 사이 같은 변수 간의 호환성과 상호작용을 모델링할 수 있도록 한다. 이 방식은 진짜 데이터가 존재하는 영역에는 낮은 에너지를 부여하고, 그 외의 영역에는 높은 에너지를 부여하는 것에 집중한다.

그렇지만, EBM은 병렬 처리 가능성, 학습 안정성, 확장성에서 오랜 기간 어려움을 겪어왔으며, 트랜스포머가 그 어려움을 극복할 수 있는 장점을 지녔다. 따라서 트랜스포머가 EBM을 확장하기에 적합한 아키텍처로 간주 되어, 인간의 시스템 2 사고를 실현하기 위한 목적으로 EBT가 제안되었다.

트랜스포머는 AI의 핵심 모델로 널리 활용되고 있지만, 시퀀스 길이가 증가할수록 계산 복잡도와 메모리 요구량이 많이 늘어나고, 더불어 동적 상태 추적과 장기적 일관성 유지에 한계가 있다. 결국, 하드웨어 자원 확대와 운영 비용 증가를 초래하며, 출력 일관성 저하로 사용자 경험을 악화시킬 수 있다.

2) 데이터의 불균형

AI 모델에서 나타나는 편향은 단순히 기술적 문제에 그치지 않고 사회적으로도 심각하게 영향을 끼친다. 즉, 학습 데이터에 있던 성별·인종·문화적 고정관념이 그대로 답변에 반영되어 사회적 편향이 재생산된다. 채용·의료·법률 같은 민감한 분야에서 특정 집단에 불리한 추천이나 평가에 의해 불공정한 판단이 발생한다. 편향이 사실처럼 사용자에게 전달되어 잘못된 인식이 강화되고, 허위 혹은 왜곡된 정보가 확산된다. 영어권 중심 데이터에 치우치면 비서구권 언어, 문화, 지역에 대해 성능이 저하된 답변이 얻어질 수도 있다. 이러한 편향의 근본적인 원인 중 하나가 학습 데이터의 불균형 문제이다. AI 모델들은 학습의 대상이 되는 데이터들의

분포가 균형을 이루고 있다는 가정 하에, 다양하게 개발되었다. 그렇지만, 많은 영역에서 데이터 분포가 심각하게 치우쳐있는 불균형 데이터를 다루고 있으며, 이 경우 일반적인 AI 모델들은 성능을 발휘하지 못한다.

이제까지 데이터 불균형 문제의 해결을 위하여 제시되었던 기존 기법들은 크게 세 가지 범주로 나뉘어 질 수 있다.[9)10)] 먼저, 데이터 전처리 기법은 주어진 불균형 데이터를 처리하여 데이터 분포를 수정함으로써 표준 알고리즘들을 효과적으로 사용할 수 있도록 하는 것이다. 즉, 불균형 데이터를 학습에 직접 적용하기보다는, 먼저 데이터 분포를 수정하는 방식으로 전처리를 수행하는 것이다. 이 기법은 기존의 어떤 학습 기법에도 적용할 수 있다. 그렇지만, 새로운 합성 데이터를 생성하는 경우에는 합성 데이터의 목표 변수 값을 정확하게 결정하는 것은 어렵다. 알고리즘 구조 접근 방식은 불균형 문제를 처리하기 위해 새로운 알고리즘을 만들거나 기존 알고리즘을 수정하는 방법이다. 이러한 방법들은 학습 목표가 모델에 직접적으로 반영된다. 그렇지만, 데이터 분포에 맞게 적절한 알고리즘을 선택하기 위한 깊은 지식이 필요하다. 하이브리드 기법은 불균형 문제를 처리하거나 전체 솔루션의 특정 부분에 여러 접근 방식을 사용한다. 이러한 기법은 서로 다른 방법들이 적절히 상호 보완됨을 바탕으로 다양한 유형의 전략을 혼합하여 각 방법의 장점을 최대한 활용하려고 시도한다.

3. AI 기술의 응용

트랜스 포머를 기반으로 한 AI 기술의 발전은 거대 기술 기업들이 LLM을 개발하여 널리 배포하게 하였다. LLM은 광범위한 작업에서 전문인 수준의 성능을 보여주고 있다. 또한, 언어 모델이 소수의 특화된 작업 만을 반복적으로 수행하는 에이전트형(Agentic) AI 시스템의 부상은 전문적인 영역에서 전

9) Rezvani S. & X. Wang., “A Broad Review on Class Imbalance Learning Techniques”, *Applied Soft Computing*, vol. 143, 2023, p.110415.

10) Oh, S.-H., “Error back-propagation algorithm for classification of imbalanced data”, *Neurocomputing*, vol. 74, 2011, pp.1058-1061.

문인을 대신하는 AI 서비스를 제공하고 있다. 물론, AI 에이전트를 구동하는 핵심 구성 요소는 LLM이지만, AI 자원의 효율적 활용을 위해 AI 에이전트의 LLM을 충분한 성능을 지닌 소형 언어모델(SLM, Small Language Model)로 대체하면 비용 절감을 비롯한 AI 자원의 효율적 활용 효과를 얻을 수 있다.[11] 즉, 추론·배포·운영이 가벼워지는 비용절감 효과와 민첩한 작업이 가능하게 되어 사용자의 만족도를 향상시킨다. 또한, 자원 절감과도 연결된다.

한편, LLM을 과학 연구에 적용하여 자동화를 이루려는 시도는 가설 수립, 실험 계획 및 구현, 과학 문서 작성, 동료 심사 분야에 걸쳐 지속적으로 진행되어 왔었다.[12] 여기서는 LLM을 과학 연구 자동화에 응용하기 위하여 개발된 에이전트형 AI의 대표적 사례인 구글의 부과학자, 알파이볼브와 알파지놈을 설명한다.

1) 구글의 부과학자, 알파이볼브, 알파지놈

구글의 부과학자[13]는 Gemini 2.0을 기반으로 한 다중 에이전트 AI 시스템으로, "과학자 중심 협력 모델"을 기반으로 과학적 방법론을 반영하도록 설계되었다. 연구자가 자연어로 연구 목표를 입력하면, 이 시스템은 관련 문헌을 검색하고 논리적으로 분석하여 기존 연구를 요약 및 종합한 후, 이를 바탕으로 새롭고 독창적인 연구 가설과 실험 프로토콜을 제안한다. 또한, 제안의 타당성을 보장하기 위해 관련 문헌을 인용하고, 그 근거를 설명한다. 연구자들은 AI가 생성하는 가설이나 연구 제안에 원하는 특성과 제약 조건을 지정할 수도 있다. 연구자가 아이디어와 가설을 제시하고, AI 생성 가설을 직접 수정하거나, 자연어 채팅으로 AI의 방향성을 조정하는 등 다양한 방식으로 시스템과 협업할 수 있다. 테스트 시점 연산 확장(scaling of test-time compute)을 활용하여, 점진적으로 논리적 사

11) Belcak, P. et al. "Small Language Models are the Future of Agentic AI", 2025, arXiv:2506.02153v2.

12) Luo, Z. et al. "LLM4SR: A Survey on Large Language Models for Scientific Research", 2025, arXiv:2501.04306v1.

13) Gittweis, J. et al. "Towards an AI co-scientist", 2025, arXiv:2502.18864v1.

고를 수행하고, 가설을 발전시키며, 점점 더 정교한 결과를 도출한다.

시스템의 핵심 알고리즘은 크게 다음과 같은 과정으로 이루어진다. 먼저, 자기 대결(self-play) 기반 과학적 토론 과정을 통해 새로운 연구 가설을 생성한다. 그 다음으로, 토너먼트 평가를 통해 가설 간 승패 패턴을 분석하여 우수한 가설을 선별하고, 가설 진화 과정을 통해 지속적으로 가설의 품질을 향상시킨다. 또한, AI 부과학자는 자기 비판 능력을 갖추고 있어, 웹 검색 등의 도구를 활용하여 자체적으로 피드백을 제공하고 가설과 연구 제안을 반복적으로 개선할 수 있다.

AI 부과학자를 구성적인 측면에서 살펴보면, 특화 에이전트인 생성 에이전트, 반영 에이전트, 순위 에이전트, 근접도 에이전트, 진화 에이전트, 메타리뷰 에이전트가 있다. 생성 에이전트는 연구 목표에 부합하는 초기 가설 목록을 생성한다. 이 가설들은 반영 에이전트의 검토를 거쳐, 순위 에이전트의 토너먼트 평가를 받는다. 진화, 근접도, 메타 리뷰 에이전트들은 토너먼트 상태를 기반으로 시스템 출력을 향상시킨다. 감독 에이전트는 이 특화 에이전트들을 조율하여 입력된 연구 목표에 맞는 타당하고, 참신하며, 검증 가능한 가설과 연구계획을 만들어 준다.

이 시스템은 범용적 활용이 가능하며, 우선적으로 생물의학 분야의 세 가지 영역인 약물 재창출, 신규 표적 발견, 박테리아 진화 및 항미생물 저항성의 메커니즘 설명에서 개발 및 검증을 진행하여, 그 우수성이 입증되었다.

구글 딥마인드는 알파고(AlphaGo) 이후에도 꾸준히 혁신적인 연구 결과들을 발표하고 있다. 2025년에 발표한 대표적인 사례로서, 6월에 발표한 알파이볼브와 7월에 발표한 알파지놈이 있다.

알파이볼브[14]는 진화연산과 LLM기반 코드 생성을 결합한 LLM 기반 코드 최적화 에이전트이며, LLM과 진화 알고리즘을 활용하여 복잡한 알고리즘을 설계하고 최적화 해준다. 이를 위해 LLM들로 구성된 자율

14) Novikov, A. et al. “AlphaEvolve: A coding agent for scientific and algorithmic discovery”, 2025, arXiv:2506.13131v1.

파이프라인을 조율하여 알고리즘을 진화 연산을 통해 직접적으로 수정함으로써 개선된 알고리즘을 제시하며, 하나 이상의 평가자로부터 지속적으로 피드백을 받아 알고리즘을 반복적으로 향상시킨다.

알파이볼브를 실제 문제에 적용하여, 데이터 센터를 위한 효율적인 스케줄링 알고리즘을 개발하는 데 성공했다. 하드웨어 가속기의 회로 설계에서는 기능적으로 동일하지만 더 간단한 방식을 찾아냈고, 알파이볼브를 구동하는 LLM의 학습 속도 또한 가속되었다. 더군다나, 수학 및 컴퓨터 과학 분야의 문제에서 특히 주목할 성과로, 4×4 복소수 행렬 곱셈을 48회의 스칼라 곱셈으로 수행하는 방법을 찾아내었다. 이는 이 분야에서 Strassen 알고리즘 이후 56년 만에 이루어진 개선이다. 알파이볼브와 같은 코딩 에이전트들이 과학의 다양한 분야에서 문제 해결 능력을 크게 향상시킬 것이다.

알파지놈[15)16)]은 DNA 서열의 유전자 기능 예측과 질병 생물학 연구에 혁신을 가져왔다. 즉, 최대 100만 염기쌍의 DNA 서열을 입력받아 염기 수준의 고해상도 생명현상과 다양한 유전자 조절 변이의 효과를 예측한다.

알파지놈 아키텍처는 입력된 DNA를 인코더, 트랜스포머, 디코더 단계를 거쳐 처리한 후, 여러 해상도에서 중요한 내부 표현을 생성한다. 이 학습된 1차원 임베딩과 2차원 쌍별 임베딩은 다양한 유전체 특징을 예측하는 기반이 된다. 예측은 특정 생물종(인간 또는 생쥐)에 맞게 조정되며, 이를 위해 생물종별로 학습된 임베딩을 해당 함수 내에 통합하여 사용한다. 이들 공유 임베딩을 생성할 때 적용되는 생물종별 임베딩 외에도, 각 출력 헤드는 인간과 생쥐 예측을 위해 별도의 독립적인 파라미터 세트를 학습한다. 알파지놈을 변이 효과 예측에 대한 26가지 평가에 적용시킨 결과, 24가지에서 기존의 최고 모델과 동등하거나 이를 능가하는 성능을 보였다.

15) Avsec, Ž. et al. “AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model”, *Nature*, vol. 649, 2026, pp.1206-1218.

16) Callaway, Ewen. “DeepMind’s new AlphaGenome AI tackles the ‘dark matter’ in Our DNA”, *Nature*, vol. 643, 2025, pp.17-18.

구글 부과학자, 알파이볼브와 알파지놈은 완전 자동화된 연구 결과물을 제시하지 않고 과학자들과의 협력이 필요한 시스템이지만, 이미 현재의 결과 만으로도 과학연구자들에게 큰 충격을 주고 있다. 만약, 과학 연구가 완전 자동화된 'AI 과학자'가 등장하면, 인간 과학자의 역할에 대한 재정립이 필요해짐은 물론 AI 기술을 지녔으며 이를 운용할 충분한 자본을 지닌 거대국가와 거대 기업이 모든 과학적 발견을 독식하는 과학 연구의 양극화 현상이 더욱 심화될 것이다. 이제, AI 과학자의 기대와 현실에 대하여 살펴보자.

4. 기대와 현실

인류 문명의 발전을 이끌어 온 과학적 발견은 과학 연구 주체인 인간의 체계적인 탐구 과정에 의해 이루어져 왔다. 이 과정은 일반적으로 관찰과 학습을 통한 과학적 기초 지식 마련 단계인 지식획득, 미해결 질문을 다루기 위한 과학적 가설수립 단계인 아이디어 생성, 강도 높은 실험과 이론적 분석을 통한 가설의 엄격한 검증 및 반증, 그리고 실험 또는 이론적 결과에 기반한 지속적 연구의 정교화(진화) 과정을 거치면서 과학지식의 발전이 이루어진다. 연구 과정은 인간의 제한된 시간과 인지능력에 의해, 느린 문헌조사, 협소한 지식, 편향된 가설, 비효율적 실험 등의 문제를 지니며, 그 해결책으로 과학 연구 자동화의 관심이 높아져 왔다.

3장에서 살펴본 바와 같이 LLM을 기반으로 한 에이전트형 AI 시스템은 과학 연구의 자동화에 괄목할 만한 성과들을 보여주고 있으며, 머지 않아 AI 과학자가 아직 발견하지 못한 새로운 현상을 규명할 날이 다가올 것이다. 이제, "AI 과학자가 세상을 바꾸고 과학 연구 패러다임을 재편할 수 있는 세상이 얼마나 가까이 왔는가?"[17]에 대한 질문에 초점을 맞추어 보자.

과학 연구의 자동화에 필수적인 능력들은 다음과 같다. 먼저, 지식획득은 기존 과학 문헌에서 분야별 지식을 스스로 검색, 검토, 및 이해하여 연

17) Xie, Q. et al. "How far are AI scientists from changing the world?", 2025, arXiv:2507.23276v2.

구를 위한 과학적 지식 기반을 체계적으로 구축하는 자율적 능력이다. 다음으로 아이디어 생성이란, 획득한 지식을 기반으로 혁신적이고 실현 가능한 과학적 가설을 자율적으로 수립하는 능력이다. 이 능력은 자동화된 과학 도구와 AI 과학자 시스템을 구별 짓는 핵심 요소이다. 검증과 반증은 AI 생성 가설을 AI 과학자가 체계적으로 검증하고 반증하기 위해 실험을 설계, 실행, 분석할 수 있게 한다. 이는 AI 과학자가 연구 사이클을 완성 시키고 자율적 과학 지능을 갖추게 하는 핵심 요소이다.

AI 리뷰 시스템은 AI 과학자 시스템이 생성한 원고를 포괄적으로 평가하여 통찰력을 지닌 심사 결과를 제공해준다. Q. Xie 등이 5개의 주요 AI 과학자 시스템이 생성한 연구논문 28편을 AI 리뷰어 모델 Deep-Reviewer14B 를 이용하여 평가한 결과에 의하면 현재 AI 과학자 시스템의 자율 연구 전반에 걸쳐 구조적인 한계가 드러났다. 실험적 약점은 모든 논문에서 발견되어, AI 과학자 시스템의 구현 능력이 실험의 설계, 실행, 분석과 관련된 심각한 한계가 있음을 나타낸다. 이외에도, 방법론적 불명확성 혹은 결함, 논문 작성 및 표현 문제, 혁신성 부족 등도 많이 발견되어, 과학적 실행과 연구 결과의 문서화에 어려움이 있음을 보여 주었다. 또한, 이론적 약점이나 관련 연구 검토 미비의 평가는 독창적인 결과를 제공하거나 주장에 대한 견고한 이론적 틀 제공에 실패하였음을 나타낸다. 정리하자면, 현재의 AI 과학자 시스템이 확립된 기준을 충족하는 고품질의 과학적 산출물을 독립적으로 생성하기에 부족함을 드러내고 있다. 이러한 한계의 극복 방법으로 내부 평가 혹은 외부 입력을 바탕으로 전체적 연구 역량을 지속적으로 향상시키는 진화적 메커니즘의 도입이 필수적이다.

진화적 메커니즘의 핵심기술로 동적 계획과 자율학습이 있다. 동적 계획은 제한된 탐색 공간 내에서 가장 가치 있는 연구 방향을 탐색한다. 이를 통해, AI 과학자가 자원을 효율적으로 배분하고, 가설의 우선 순위를 정하며, 진행 중인 결과와 피드백에 기반하여 연구 경로를 적응적으로 정교화시킨다. 자율학습은 자기 성찰 기반 방법과 외부 주도 방법으로 나뉘

어 진다. 자기 성찰 기반 방법은 LLM이 스스로 피드백 제공자의 역할을 수행하여 특정 기준을 충족할 때까지 출력을 반복적으로 평가하고 개선하는 방법이다. 외부 주도 방법은 가설의 품질을 평가하고 수정 제안을 제공하기 위해 생성 과정에 인간의 감독을 포함시키는 방식과 인터넷 기반 학술 플랫폼을 활용해 기존과 유사한 가설을 걸러내어 참신성을 확보하는 방식이 있다. 이외에, AI 과학자 시스템 만으로 구성된 연구 커뮤니티를 구축하여 협업을 통해 산출물의 품질을 높이는 방법도 있다.

AI 과학자 시스템이 상당한 진전을 보였지만 세상을 변화시키고 과학 연구 패러다임을 재편할 획기적인 발견을 만들어 낼 자동화 에이전트로 성장하기까지 여전히 큰 격차가 존재한다. 가장 많이 언급되는 격차는 AI 과학자의 기반 모델인 LLM의 한계에 의한 것이다. LLM에는 환각, 지식 업데이트의 비용과 비효율성, 파국적 망각 등의 여러 가지 내재적 한계가 있다. 환각에 의해 연구 결과, 실험 데이터, 참고 문헌, 혹은 존재하지 않는 이론들이 사실인 것처럼 조작하여 생성된다. 또한, 신뢰할 수 없거나 출처가 부정확한 자료를 제공되어 연구자들이 제공받은 정보의 근거와 타당성을 추적하고 검증하기 어렵게 한다. 또한, 미묘한 논리적 모순이나 맥락에 어긋난 주장으로 연구자들을 오도하고 과학적 산출물의 신뢰성을 저하시킨다. 과학 연구의 빠른 발전은 최신 연구 결과의 지속적 업데이트를 요구하지만, 방대한 양의 최신 과학 논문을 LLM에 통합시키기 위해 대규모 모델을 재학습하거나 미세조정 하기에는 막대한 비용 지출이 수반된다. 결국, 지식의 늦은 갱신으로 최신 과학 지식과 AI 과학자 시스템의 지식 사이에 격차가 존재하게 된다. 또한, LLM이 새로운 데이터를 학습하는 과정에서 기존에 습득한 정보를 잊어버리는 파국적 망각도 큰 걸림돌이다.

AI 과학자 시스템의 연구 역량 한계도 과학 연구 자동화의 잠재력을 제한하고 있다. 방대한 최신 과학 문헌으로부터 지식을 획득하고, 체계화 및 내재화를 하는 능력은 AI 과학자에게 필요한 가장 기본적인 역량이다. 이

는 관련 연구 결과를 검색 및 탐색하고 핵심 아이디어를 추출하고, 실험적 세부사항에 반영하며, 여러 연구결과들을 통합하여 일관된 통찰력을 제공해줄 것이다. 그렇지만, 지식 획득 단계에서 정보검색 정밀도의 문제 즉, 특정 연구에 가장 관련성 높은 최신 정보를 정확하게 검색하는 데 어려움이 있다. 또한, 타 문서 요약 및 통합 시 자료의 출처를 잘못 인용하거나 왜곡하며, 의미적 이해와 사실적 정확성 측면에서 여전한 어려움이 있다. 아이디어 생성 단계에서는 참신하고 실현가능한 과학적 가설의 생성이 기대에 미치지 못하는 경우가 많다. 또한, AI가 생성한 가설을 평가하기 위한 신뢰성 높고 객관적이며 확장 가능한 평가 체계의 부족도 AI 과학자 시스템 발전의 병목으로 작용한다.

검증 및 반증 단계에서의 한계는 다음과 같다. 다른 AI 에이전트, 인간 과학자 혹은 다양한 외부 도구와의 협업이 과학 발전에 필수적인데, 현재의 AI 과학자 시스템은 견고성, 적응성, 장애 허용성 측면에서 큰 약점을 보이고 있다. 또한, 자동으로 생성된 실험 설계가 과학적 엄밀성, 혁신성, 실용성이 부족한 경우가 많다. 결국, AI 과학자 시스템이 정적 템플릿과 제한된 문헌에 과도하게 의존하게 되어 최신 기술이나 새로운 실험 방법론을 반영하기 어려워진다. 진화 단계는 AI 과학자의 장기적인 혁신에 필수적이지만, 핵심 기술인 동적 계획은 여전히 미흡한 상황으로, 탐색(exploration)과 활용(exploitation)의 균형을 효율적으로 유지하는 데 어려움이 있다. 자율 학습 기술의 자기 성찰 기반 방법에서는 효과적인 자기 비판 메커니즘이 부족할 경우 시스템의 오류가 누적되는 악순환에 빠져 연구 결과의 독창성과 신뢰성이 저하될 수 있다. 외부 평가의 피드백을 통합하기 위해 필요한 AI 과학자 간 상호작용 프로토콜이 부족하여, 아이디어 교환과 외부 비판의 통합이 비효율적으로 이루어지게 된다.

앞서 살펴본 바와 같이 아직 AI에 의한 과학의 완전 자동화에는 현실적인 제약이 많이 존재한다. 하지만. 미국 정부는 제네시스 미션을 통하여 미 연방의 연구 인프라인 컴퓨팅 자원과 AI 기반 자동화 연구실 등을 총동

원해 과학 연구를 가속화 하는 정책을 추진 중이다. 중국을 비롯하여 유럽연합과 영국도 비슷한 정책을 추진 중이며, 대한민국도 k-문샷 정책을 발표하여 AI에 의한 과학 연구 자동화를 적극 추진 중이다. 이제 국가적 차원에서 AI 기술과 자본력을 AI 과학 자동화에 투입하여 얻게 되는 결실이 향후 국가 발전의 원동력이 될 것이다. 정부의 거버넌스 역시 국가의 발전에 필요한 AI 기술에 저해되지 않도록 설계되어야 한다. 또한, 국제적 거버넌스는 국가 간의 양극화를 해소시켜, 모든 인류가 AI 기술의 혜택을 누리도록 할 필요성도 증대되고 있다.

5. 위험과 안전성

AI의 가장 큰 위험적 요소로 LLM의 환각 현상이 언급된다. 여기서는 환각이 왜 발생하는 지에 대한 구조적 원인을 알아보자. 그리고, AI 시스템이 사회·경제·정보 생태계에 미치는 영향을 안전성 측면에서 살펴보자.

1) 환각(Hallucination)의 구조적 원인

언어모델이 만들어내는 지나친 확신에 찬 그럴듯한 거짓 정보를 환각이라는 오류 유형으로 부르며, 인간의 지각적 환각과는 근본적으로 다르다. AI 모델의 상당한 발전에도 불구하고 환각은 여전히 AI에서 발견되고 있다. 이를 분석하기 위하여, 유효한 문자열 집합을 V, 오류의 일반 집합을 E라 하고, 그럴듯한 문자열 집합을 $X = E \cup V$로 정의하자. 기본언어모델은 학습 데이터 분포 p에서 사전학습으로 $\hat{p}$을 생성한다. $err = \hat{p}(E) = \Pr_{x \sim \hat{p}}[x \in E]$로 기본 모델의 오류율이 정의되며, 최적의 기본 모델은 $\hat{p} = p$이므로 오류가 없다.

err의 하한을 구하기 위해, 언어모델의 출력이 유효한 출력인가에 답하는 IIV(Is-It-Valid) 이진 분류 문제를 정의하자. 이는 교사학습 문제이며, 유효한 예시와 오류 예시를 50:50으로 학습 및 시험 데이터를 구성한다. 유효한 예시는 V에서 가져오고, 오류 예시는 E에서 무작위로 균일하게

선택한다. 그러면, 학습 대상 목표 함수 $f : X \rightarrow [-,+]$와 X에 대한 분포는

$$D(x) = \begin{cases} p(x)/2 & \text{if } x \in V, \\ 1/2|E| & \text{if } x \in E, \end{cases} \text{ and } f(x) = \begin{cases} + & \text{if } x \in V, \\ - & \text{if } x \in E, \end{cases} \quad (1)$$

와 같이 된다. 그러면, IIV의 오분류율은

$$err_{iiv} = \Pr_{x \sim D}[\hat{f}(x) \neq f(x)] \text{ and } \hat{f}(x) = \begin{cases} + & \text{if } \hat{p}(x) > 1/|E| \\ - & \text{if } \hat{p}(x) \leq 1/|E| \end{cases} \quad (2)$$

이다. 이를 통해, $err \geq 2 \times err_{iiv}$임이 증명되었다.[18] 즉, 언어모델의 오류율이 IIV 문제의 오류율보다 2배 이상이어서 환각을 피할 수 없다.

한편, 후속 학습은 기본 모델을 정교화하여 주로 환각을 줄이는 것을 목표로 한다. 그런데, 왜 언어모델은 "모르겠습니다(IDK)"라는 답 대신, 확신에 찬 환각을 생성하는 것인가? 언어모델을 "정답이면 1점, 오답·무응답·IDK는 0점"으로 채점되는 방식에서는 확신이 없을 때 추측이 기대 점수를 극대화한다. 즉, 채점 방식이 환각을 조장하므로, 환각을 줄이기 위해서는 불확실하면 답변을 생략하더라도 벌점을 받지 않도록 평가 방식 자체를 수정해야 한다.

언어모델의 환각은 항상 존재하며, 이에 따른 문제는 상당히 다양한 영역에서 나타난다. 우선, 존재하지 않는 자료를 제시하는 사실적 오류가 있다. 법률, 의료, 금융의 영역에서 오답 제시로 의사결정이 왜곡될 수 있다. 자신감 있는 답변 제시가 결국은 신뢰도 저하를 유발한다. 각종 연구 보고서에 허위 정보가 포함되어 연구 품질이 저하되고, 또한 소프트웨어 개발 시에는 시스 템 오류를 일으키기도 한다. 허위사실에 기반한 조언 제시로 법적 책임 문제가 발생하며 보안 사고로 연결될 수도 있다.

2) AI 안전성

생성형 AI의 등장 이후로 AI 시스템의 영향은 사회 시스템 및 산업 전반에 걸쳐 광범위한 영향을 미치는 사회·기술적 영역으로 확장되었음을 일본 AISI(AI Safety Institute)와 미국 NIST(National Institute of

18) Kalai, A. T. et al. "Why Language Models Hallucinate", 2025, arXiv:2509.04664v1.

Standards and Technology) 등에서 언급하고 있다. “사회·기술적”이란 AI 자체 및 AI 시스템과 관련된 기술적 요소와 이들을 둘러싼 사회적 요소 간의 상호작용을 의미한다. 여기서는 AI 및 AI 시스템이 사회의 다양한 요소들과 어떻게 상호작용 하며 현실 세계에 영향을 미치는지를 윤리·법, 경제활동, 정보 공간, 환경 네 가지 범주로 알아보자.[19]

윤리와 법에 대한 영향으로, AI에 대한 과도한 의존이 심리·신체적으로 영향을 미침으로 나타나는 현상이 언급된다. 생성형 AI가 사용자의 정서적 의존을 강화시키고, 심지어 부적절한 프롬프트 제공으로 자살을 유발한 사례가 있다. 또한, AI에 대한 과도한 의존은 뇌 활동과 독창성을 저하시키고 비판적 사고능력이 약화되는 ‘인지 부채’의 축적 현상이 나타난다. AI의 편향은 학습 데이터 때문이다. 이러한 편향은 역사·사회적으로 형성된 차별과 불균형을 재생산하며, 인종·성별·종교·장애·문화적 배경 등의 요인에 기반한 불공정한 대우와 고정관념을 조장한다. 그리고, 이 편향이 제도화되고 구조화되어 장기적으로 사회적 불이익을 고착화 시킬 수도 있다. 생성형 AI의 자연어 생성, 멀티모달 처리, 코드 생성 등의 기능은 효율적인 사이버 공격 도구로 활용된다. 특히, 수작업 편집과 고도의 기술이 필요했던 음란물 생성이 이제는 손 쉽게 가능한 상황이 되었다. 즉, AI-CSAM(AI-generated Child Sexual Abuse Material)와 리벤지 포르노 등을 통해 아동과 여성의 권리 침해·성적 착취와 같은 심각한 사회적 문제를 유발시킨다.

경제 활동에 미치는 영향으로, AI로 생성된 콘텐츠의 지식 재산권이 있다. 다양한 콘텐츠가 AI에 의해 생성되어 창작활동에 큰 혁신을 이끌어왔지만, 특정 캐릭터나 예술적 스타일의 모방이 저작권 침해로 이어질 수 있으며, 브랜드 가치 훼손과 초상권 및 출판권 침해를 유발할 위험도 있다. 많은 기업들의 업무 효율성을 제고하기 위한 AI 시스템의 도입은 고용 및 노동시장에 큰 변화를 초래하고 있다. AI 시스템의 효율성 향상과 같은 긍

19) Japan AI Safety Institute, *Reports on the Specific Influence on AI Safety*, Oct. 31, 2025.

정적 요소와 함께 업무 대체에 따른 일자리 감소의 부정적 영향이 동시에 일어나고 있다. 생성형 AI에 의한 저비용 고속 콘텐츠 생성으로 대량의 저품질 정보가 범람하게 되었다. 또한, AI의 혜택이 사회 전체적으로 공정하게 배분되지 않고, 기술적 우위에 있는 개인·기업·국가에 집중되어 기존의 격차를 확대시키고 새로운 불평등을 야기하고 있다. AI 시스템의 학습에 사용되는 대규모 정보는 기밀정보나 개인 데이터를 포함하게 되어, 기밀정보 외부 유출·개인정보 침해·기업 경쟁력 약화·법적 책임과 같은 위험을 초래한다.

정보 공간에 미치는 영향으로, 우선 허위 정보 및 조작 정보의 생성과 확산이 있다. 생성형 AI에 의해 가짜 뉴스나 허위 영상이 짧은 시간에 제작되어 손쉽게 전파되고 사회적 혼란을 초래한다. 이 과정에 개인은 자신의 선입견이나 의견을 뒷받침하는 정보 만을 수집하는 '에코 챔버 현상'이 일어난다. 에코 챔버 현상이란 유사한 가치관을 가진 개인들이 서로의 공감을 강화 함으로써 특정 의견이나 이데올로기가 과도하게 증폭되고 그 영향력이 커지는 상태를 말한다. 또한, AI에 의해 생성된 결과물이 사회의 다양성을 충분히 반영하지 못하여 사회적으로 소외된 집단이 결과물에 보이지 않게 되거나 기존의 고정관념을 고착화시킨다.

생성형 AI의 등장은 데이터 시스템의 막대한 전력 수요를 일으켜 환경에 부정적 영향을 미친다. 일반적인 구글 검색은 요청 당 평균 0.3Wh의 전력을 소비하는 반면 ChatGPT 요청은 평균 2.9Wh의 전력이 소모되며, GPT-3의 학습에 약 1,300MWh의 전력이 필요했다고 한다. GPT-4와 5는 훨씬 더 많은 전력 소모가 되었을 것이다. 이러한 전력 소모는 막대한 CO_2 배출, 전력망 부담 증가, 냉각수 대량 소모 등 환경에 부정적 요소이다.

6. 통제와 거버넌스

AI 시스템의 발전은 사회를 풍요롭게 하지만 한편으로 위험적 요소도 지니고 있기에 개발, 배포 및 사용과 관련된 통제 시스템이 필요하다. 이

와 관련된 활동으로 마이크로소프트는 100개 이상의 생성형 AI 제품에 대한 레드 팀 운영 경험을 바탕으로 실질적인 권고사항을 제시하였다.[20] 여기서는 통제의 핵심에 해당하는 AI 시스템의 모니터링에 대한 내용과 AI 레드팀 활동을 알아보고 거버넌스 활동의 대표 사례도 살펴본다.

1) 모니터링 가능성(Monitorability)

AI 시스템의 의사결정을 관찰하는 것은 점점 더 강력해지는 AI 에이전트들을 안전하게 배포하고 사용하기 위하여 필수적이다. 특히, 추론(reasoning) 모델의 CoT(Chain-of-Thought) 모니터링은 이상 행동 탐지에 아주 효과적이다. 여기서, 모니터링 가능성을 '모니터가 AI 에이전트의 행동과 관련된 속성들을-예를 들어, 해로운 방식으로 행동하고 있거나 앞으로 행동할 것인지- 정확하게 예측할 수 있는 능력'으로 정의하며, 이는 평가대상 모델의 행동 뿐만 아니라 그러한 행동으로부터 유용한 정보를 추출할 수 있느냐에도 의존한다. 모니터링 가능성은 관심 속성들이 모델의 출력으로부터 신뢰성 있게 식별될 수 있는 지 여부에 초점을 두며, 이는 효과적인 AI 시스템 통제를 위한 핵심적 조건이다.

모니터링 가능성 평가를 위한 방법으로 다음 세 가지의 상호 보완적 방법이 제시되었다.[21] 첫째, 개입평가는 AI 에이전트의 환경이나 입력에 통제된 개입을 가한 뒤, 그 개입에 의해 에이전트의 행동이 변화되었는 지를 측정하고 최종적으로 특정 추론 상의 변화를 모니터가 추론할 수 있는지 평가한다. 이 평가는 모니터가 개입에 대한 모델 추론의 증거를 발견할 수 있어야 한다는 직관에 기반하며, 왜 에이전트의 행동이 변화했는 지를 감지할 수 있게 한다. 둘째, 과정평가는 허용 가능한 해법 경로의 집합이 작고 명확한 문제를 구성하고, 모델이 올바른 최종 답을 산출했을 때, 모니터를 통해 모델이 유효한 해법 경로를 실제로 따랐다는 증거를 찾을 수 있

20) Bullwinkel, B. et al. "Lessons for Red Teaming 100 Generative AI Products", 2025 arXiv:2501.07238v1.

21) Guan, M. Y. et al. "Monitoring Monitorability", 2025, arXiv:2512.18311v1.

는지 평가한다. 예를 들어, 특정 수학 문제의 경우 올바른 해답은 반드시 소수의 특정 정리 중 하나를 호출해야 함을 이용하는 방식이다. 셋째, 결과-속성 평가는 모델 출력의 특정 속성을 명확하게 라벨링 할 수 있는 환경을 설계하고, 이후 모니터에게 해당 속성과 관련하여 모델이 어떤 결정을 내렸다는 증거를 식별할 수 있는지를 묻는다. 개입평가와 과정평가는 결과에 대한 측정을 모델의 추론 과정과 연결하는 반면, 결과-속성 평가는 단순히 결과의 특정 측면이 예측 가능한 지 여부 만을 묻는다.

모니터링 가능성은 클래스 불균형 문제와 관련이 있다. 즉, 이상 행동이란 아주 드물게 발생하기에 이러한 데이터 불균형 문제를 고려하여 지표를 설정하고 사용하여야 한다. 이에 모니터링 가능성 지표로 제시된 방법이 민감도(TPR, true positive rate)와 특이도(TNR, true negative rate)의 기하평균 $g-mean=\sqrt{TPR\times TNR}$ 이다.

2) AI 레드팀 활동(Red Teaming)

레드팀 활동은 19세기 초 프로이센 군대의 전술적 전쟁게임에서 기원하였다. 미국 국가안보국은 1980년대 처음으로 사전적 사이버보안 조치의 필요성을 인식하고 기밀 시스템의 보안을 평가하는 '레드팀' 개념을 개척하였다. 1990년대 디지털 위협이 진화함에 따라 사이버보안 레드팀도 발전했다. 특히, 2001년 9·11 테러에서의 정보 실패 이후 미국 국방부는 향후 유사한 참사 방지를 위해 레드팀 부대를 설립하였고, CIA는 "레드셀(Red Cell)"이라는 새로운 부서를 만들었다. 현대 사이버보안 레드팀은 물리적 보안, 사회공학, 기타 비기술적 요소까지 아우르는 총체적인 보안 평가로 전환되었다.

레드팀에서 기술전문가는 모델 취약점을 조사하는 데 적합하다. 정책전문가는 규제 충돌을 식별해내고, 윤리학자는 가치 일치 정렬문제를 드러내며, 도메인전문가는 실제 세계에서의 영향 시나리오를 평가한다. 레드팀은 단순히 교묘한 허점을 잡아내거나 시스템을 깨뜨릴 방법을 찾는 데 목적이 있는 활동으로 국한되지 않으며 "현실 세계 조건에서 실패를 초

래할 수 있는 일은 무엇인가?"에 대한 답을 찾아내는 협력적 과정이다. 이를 위해 AI 시스템의 레드팀 활동은 AI 개발 생애주기 전체를 아우르는 시스템 차원의 거시적 수준 활동과 AI 시스템을 구동하는 개별 모델에 초점을 맞추는 미시적 수준 활동이 동시에 이루어져야 한다.[22]

먼저 거시적 수준의 레드팀 활동을 착수, 설계, 데이터, 개발, 배포, 유지보수, 폐기의 단계로 나누어 살펴보자. 착수단계에서 레드팀은 문제를 어떻게 규정하고 해당 해결책이 적절한지를 근본적으로 의문시한다. 이 문제를 해결하는데 AI가 정말 필요한가? AI 해법주의에 빠져드는 것은 아닌가? 이해 관계자들의 프로젝트 추진 동기는 무엇이며 서로 다른 집단들에게 어떤 영향을 끼치는가? 의도치 않은 문제가 만들어지는가? 어떻게 악용이나 조작이 이루어질 수 있는가? 와 같은 탐구적 질문을 검토해야 한다.

설계단계에서는 인터페이스 결정, 아키텍처 선택, 기존 시스템과의 통합 지점 등에서 발생하는 취약점을 식별한다. 데이터 단계에서는 단순한 데이터 품질 뿐만 아니라 수집 방법, 저장 방식, 접근 제어, 계보 추적, 라이프 사이클 관리에 이르는 데이터 생태계 전체를 검토한다. 특히, 데이터의 대표성과 품질에 대한 가정을 비판적으로 검토해야 한다. 개발단계에서는 구현 방법의 선택이 어떻게 시스템적 취약점을 만들어내는지, 개발 관행 자체가 어떠한 위험을 유발하는지를 중점적으로 다룬다. 개발 환경의 보안도 다루어야 하며, 공급망의 위험에도 주의를 기울어야 한다. 배포 인프라도 면밀히 검토되어야 하고, 다양한 실패 양상에 대한 중요한 제어 지점을 식별해야 한다. 모델 산출물의 배포 및 업데이트 메커니즘은 보안 취약점 뿐만 아니라 버전 불일치로 인한 시스템 신뢰성 저하를 유발할 수 있다. 유지보수 단계에서는 다양한 실패 양상에 대한 1차 방어선으로 모니터링 및 경보 시스템을 운영해야 한다. 포괄적인 모니터링을 위해 출력 품질 추세, 공정성 지표, 안전성능 지표 등을 추적해야 한다. 폐기단계에

22) Majumdar, S. et.al. "Red Teaming AI Red Teaming", 2025 arXiv:2507.05538v1.

서는 데이터 보관 및 폐기 관행을 검토해야 한다. 학습데이터와 운영 로그를 포함한 데이터 저장소는 지속적인 개인정보 위험을 초래하지만 동시에 시스템 동작 패턴과 향후 안전 및 신뢰성 개선에 유용한 실패 양상을 이해하는데 중요한 자산이다.

미시적 수준의 모델 레드팀은 모델 능력의 한계를 식별하는 경계 탐색, 예상치 못한 능력이나 제한을 드러내는 경계사례 생성, 배포 전에 모델의 내재된 위험을 파악하는 모델 고유위험 발견을 목표로 활동한다. 거시적 레드팀과 미시적 레드팀은 상호보완적 접근방식으로 포괄적 검토를 제공한다. 거시적 통찰은 미시적 테스트의 우선 순위를 결정하는데 도움을 주고, 미시적 발견은 거시적 위험 평가의 기초 정보를 제공한다. 효과적인 AI 레드팀 활동은 개발 수명 주기 전반에 걸친 두 관점의 일관된 적용이 필요하다.

3) 거버넌스(Governance)

AI 거버넌스는 "AI를 어떻게 잘 통제하고 올바르게 사용할 것인가"에 대하여 개발·배포·활용 전 과정에서 책임성, 안전성, 공정성, 투명성을 확보하기 위한 정책과 제도 및 규범 그리고 관리 체계를 의미한다. 이와 관련된 대표적인 활동으로 WHO가 2021년 "보건 분야에서 AI의 윤리와 거버넌스에 관한 포괄적인 가이던스"를 발표하여 (i)자율성 보호, (ii)인간의 복지·안전 및 공익 증진, (iii)투명성, 설명 가능성 및 이해 가능성 보장, (iv)책임감과 책무성 강화, (v)포용성과 형평성 보장, (vi)반응성 높고 지속가능한 AI 촉진의 6가지 원칙을 제시하였다.[23] 그 후속으로 2025년 3월에 "보건의료 분야에서의 인공지능 윤리 및 거버넌스: 대규모 멀티모달 모델에 대한 지침"를 발표하여 LMM(Large Mulit-modal Model)을 보건 분야에서 활용하기 위하여 원칙에 부합하는 기업 내 거버넌스, 정부의 역할, 국제 협력을 통한 거버넌스 권고사항을 제시하였다.[24] 이에 대하여

23) Geneva:World Health Organization, E*thics and governance of artificial intelligence for health*, 28 June 2021.

간략히 알아보자.

먼저 설계 및 개발단계에서 WHO는 의료용 AI 개발자가 제품의 설계, 감독, 신뢰성 및 자율규제를 개선하기 위한 조치에 투자할 것을 권고한다. 특히, 개발자는 최소 편향, 개인정보, 노동문제, 탄소 및 물 발자국, 허위 및 거짓 정보와 혐오 발언, 안전 및 사이버 보안, 인간의 인식적 권위 보존, LMM의 독점적 통제와 관련된 위험을 정부의 법률 및 규정 제정으로 해결할 것을 요구한다. 또한, 인간의 존엄성, 자유, 평등, 연대와 같은 가치를 설계의 기반으로 삼는 가치 기반 설계를 강조한다. 에너지 효율성을 개선하여 에너지 소비를 줄이기 위한 환경적 우려를 고려한 설계도 요구한다. 정부는 데이터 보호법을 통하여 권리 기반 접근 방식으로 데이터 처리규제 기준, 개인의 권리 보호, 그리고 공공 및 민간 데이터 관리 책임자와 처리자의 의무를 설정하고 법적 권리를 침해하는 행위에 대한 제재와 구제 조치를 취하여야 한다. 현재 150개국 이상이 데이터 보호법을 제정하여 AI 기술 개발의 기반으로 삼고 있지만 생성형 AI의 등장으로 적용의 한계가 존재하는 점을 인식하여야 한다. 또한, 정부는 공공 인프라를 제공하여 윤리 원칙을 준수한 개발을 이끌어 갈 수도 있다.

의료 서비스 제공 단계에서 정부는 서비스 배포 전에 AI 기술의 사용을 평가하고 규제할 책임이 있다. 배포 전에 해결해야 할 주요 위험으로는 편향, 환각 또는 잘못된 정보, 개인정보 보호, 조작 위험, 자동화 편향 등이 있다. 개발자와 공급자는 의료 분야에서 LMM을 비롯한 AI 기술을 사용함에 공동 책임을 지며, 법 혹은 계약으로 책임을 명시할 수 있다. 특히, 범용 모델이 의료서비스 제공에 사용될 경우 의료기기 규제의 적용에 주의를 기울여야 한다. 또한, 소비자 보호법을 통해 사용자와 환자에게 피해를 주지 않도록 해야 한다. 이 법은 조작적 관행, 차별, 편향을 방지하고, LMM이 인간처럼 보이도록 오도하는 표현을 제한하거나 금지하여야 한다.

24) Geneva:World Health Organization, *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*, 25 March 2025.

보건의료 서비스 사용을 위한 AI 기술의 배포 단계에서는 부정확하거나 잘못된 응답, 편향, 입력 혹은 출력에 포함된 개인 정보, 접근성 및 경제성, 노동 및 고용에 미치는 영향, 자동화 편향과 기술 저하, 그리고 의료 제공자와 환자 간 상호작용의 수준에 대한 위험을 해결하여야 한다. 개발자와 제공자는 정부의 승인 이후에도 배포 중에 발생하는 위험에 대하여 지속적인 책임을 지닌다. 특히, 배포자는 부적절한 환경에서 사용되지 않도록 하여야 하고, 위험과 오류를 명확히 알리고 사용 중단이나 시장 철수 조치의 책임도 지며, 경제성과 접근성을 개선해 누구나 이용할 수 있도록 해야 한다.

LMM을 비롯한 AI 시스템이 의료 분야에 널리 사용되면서 발생되는 피해가 불가피하므로, 이를 보상하고 구제하기 위한 책임을 법으로 규정할 필요가 있다. AI 기술의 설계, 개발, 품질 보증, 배포 단계에서 여러 기관이 관여하므로 책임소재를 규명하는 것이 복잡하다. WHO는 AI로 인한 의료 부상에 대해 보상 기금을 설정할지 고려할 것을 권고한다.

마지막으로 국제 거버넌스를 통하여 기업들 간의 안전성 및 유효성 기준 무시와 정부 간의 기술적 우위를 위한 경쟁 격화를 방지하여야 한다. 이를 통해 모든 기업이 최소한의 기준을 충족하도록 하고, 각국 정부와 기업이 규제로 인한 불이익을 받지 않도록 보장해야 한다. 또한, AI 시스템 개발과 배포에 대한 정부의 책임을 묻고, 윤리적 원칙과 인권을 존중하는 규제를 도입하도록 한다. 특히, 국제 거버넌스를 고소득 국가나 대규모 기술 기업만 주도하지 않도록 하여, 저소득이나 중소득 국가들이 AI 기술의 혜택을 받게 하여야 한다. 이는 AI 기술이 과학기술 분야를 넘어서 경제력과 국방력을 결정짓게 될 것이기에 아주 중요하다.

7. 맺음말

본 논문에서는 AI의 기술적인 측면부터 과학 연구에의 응용, AI에 대한 기대와 현실의 간극, AI 기술이 지닌 위험성, 그리고 AI의 통제와 거버넌

스에 이르기까지 살펴보았다. 가장 활발하게 사용되는 AI인 LLM의 핵심 요소 트랜스포머는 언어영역을 넘어 영상처리를 위한 ViT, 시계열 데이터 처리를 위한 TST 등으로 적용 분야가 확대되고 있으며, 인간의 시스템 2 사고를 모방하여 신중한 결정을 내리는 데 적합한 EBT도 개발되었음을 알아보았다. 에이전트로 진화하는 AI 기술을 과학기술 분야에 응용한 대표적인 사례로 구글 부과학자, 알파이볼브, 알파지놈 등을 살펴보았다. 이렇게 세상을 바꾸어가는 AI에 대한 기술적 가능성과 실제 사이의 간극을 알아보았으며, 특히 AI 기술이 지닌 위험성을 살펴보았다. LLM에서 나타나는 환각 현상의 이론적 분석을 기반으로 사람이 좋은 평가를 받기 위하여 질의에 응답하듯이 LLM도 평가의 기대치를 극대화 하기 위하여 환각적 답변을 함을 설명하였다. 그리고 사회·기술적 측면에서 AI의 안전성에 대하여 알아보았다. 마지막으로 AI의 통제와 거버넌스 측면에서, LLM이 질의에 대한 답변 생성에서 작동하는 CoT을 모니터링하는 방법에 대하여 알아보았고, AI 시스템의 레드팀 활동을 AI 개발 생애주기 전체를 아우르는 시스템 차원의 거시적 수준 활동과 AI 시스템을 구동하는 개별 모델에 초점을 맞추는 미시적 수준 활동으로 나누어 살펴보았다. AI 기술 적용에서 부작용이 가장 민감하게 작동하는 의료 보건 분야에서의 거버넌스 활동으로 WHO가 발표한 보건의료 분야의 윤리 및 거버넌스를 알아보았다.

AI의 기술적 측면부터, 응용, 기대와 현실, 위험, 그리고 통제와 거버넌스에 걸쳐 살펴본 내용들을 기반으로 AI가 과학 기술 분야 뿐만 아니라 사회 전체적으로 급격한 변화를 불러오고 있음을 알 수 있었다. 기술적 측면에서 살펴보면, 트랜스포머가 지닌 단점에 의해 하드웨어 자원 확대와 운영 비용 증가가 초래되며, 출력 일관성 저하에 의한 사용자 경험 악화가 나타나므로 기술적 해결이 필요하다. 데이터 불균형에 따른 편향은 인종·종교·성별·소수자·약자 등에게 차별을 유발하므로, 편향적 답변에 대한 대책이 있어야 한다. AI 과학자의 출현은 높은 수준의 기술과 자본력을 지닌 거대 국가 및 기업 위주의 과학 결과 산출로 과학 기술의 양극화가 심화

될 것이다. 아직 환각의 문제는 해결되지 않았으며, 시스템 2 사고에 의해 신중하며 분석적인 사고로 복잡한 결정을 실행하는 AI 학습 모델이 필요하다. 사전 학습된 AI 모델에서 새로운 지식의 지속 학습 시 일어나는 망각 현상의 해결과 인간 수준에서의 소량 학습 능력을 지닌 데이터 효율성도 필요하다. 사회적 측면에서는 AI 모델의 동작을 모니터링하고 답변 제시 과정을 인간이 해석할 수 있어야 하며, AI 모델의 동작 목표가 인간 사회와 일치하는 방향인지도 중요하다. 급격하게 발전하는 AI 모델이 기술적 측면에서뿐만 아니라 인간 사회에 이로운 기술이 되기 위하여 사회·기술적 측면에서 안전성을 유지하기 위하여 통제와 거버넌스 활동이 더 중시 되어져야 한다. 특히, LLM과 이를 기반으로 한 에이전트 시스템 뿐만 아니라 AI 과학자에 이르기까지, 고수준의 AI의 기술을 보유하였느냐에 따라 각 국가의 영향력은 과학기술의 영역을 넘어 경제력과 국방력에 이르기까지 심화되어질 것이다. 이에, 이러한 국가 차원의 양극화를 해소하고 AI 기술의 혜택을 약소국들도 받을 수 있도록 하는 국제적 거버넌스 활동이 시급히 구축되어야 한다.

참고문헌

보고서

Japan AI Safety Institute, *Reports on the Specific Influence on AI Safety*, Oct. 31, 2025.

Geneva:World Health Organization, *Ethics and governance of artificial intelligence for health*, 28 June 2021.

Geneva:World Health Organization, *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*, 25 March 2025.

논문

Vaswani, A. et al. "Attention Is All You Need", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, pp.6000-6010.

Dosovitskiy, A. et al. "An image is worth 16x16 words: transformers for image recognition at scale", 2021, arXiv:2010.11929v2.

Zerveas, G. et al. "A Transformer-Based Framework for Multi-Variate Time Series Representation Learning", 2020, arXiv:2010.02803v3.

Gladstone, A. et al. "Energy-Based Transformers are Scalable Learners and Thinkers", 2025, arXiv:2507.02092v1.

Rezvani S. & X. Wang. "A Broad Review on Class Imbalance Learning Techniques", *Applied Soft Computing,* vol. 143, 2023, p. 110415.

Oh, S.-H. "Error back-propagation algorithm for classification of imbalanced data", *Neurocomputing*, vol. 74, 2011, pp.1058-1061.

Belcak, P. et al. "Small Language Models are the Future of Agentic AI", 2025, arXiv:2506.02153v2.

Luo, Z. et al. "LLM4SR: A Survey on Large Language Models for Scientific Research", 2025, arXiv:2501.04306v1.

Gittweis, J. et al. "Towards an AI co-scientist", 2025, arXiv:2502.18864v1.

Novikov, A. et al. “AlphaEvolve: A coding agent for scientific and algorithmic discovery”, 2025, arXiv:2506.13131v1.

Avsec, Ž. et al. “AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model”, *Nature*, vol. 649, 2026, pp.1206-1218.

Callaway, Ewen. “DeepMind’s new AlphaGenome AI tackles the ‘dark matter’ in Our DNA”, *Nature*, vol. 643, 2025, pp.17-18.

Xie, Q. et al. “How far are AI scientists from changing the world?”, 2025, arXiv:2507.23276v2.

Kalai, A. T. et al. “Why Language Models Hallucinate”, 2025, arXiv:2509.04664v1.

Bullwinkel, B. et al. “Lessons for Red Teaming 100 Generative AI Products”, 2025 arXiv:2501.07238v1.

Guan, M. Y. et al. “Monitoring Monitorability”, 2025, arXiv:2512.18311v1.

Majumdar, S. et al. “Red Teaming AI Red Teaming”, 2025 arXiv:2507.05538v1.

신문기사

박종진, “딥시크 이어 시댄스…중국 AI에 미국 또 ‘설 연휴 충격’”, 전자신문, 2026.2.15.

인터넷 사이트

ChatGPT, https://openai.com/chatgpt, 2026.04.23.

구글 코리아 블로그-기업소식, https://blog.google/intl/ko-kr/company-news/technology/google-gemini-3/, 2025.11.19.

위키백과, https://ko.wikipedia.org/wiki/딥시크, 2026.04.23.

국문초록

본 논문은 최근 인공지능(AI)과 관련하여 진행되는 일들을 최신 발표 논문과 보고서들을 기반으로 기술-응용-기대와 현실-위험과 안전성-통제와 거버넌스의 5단계 프레임으로 정리하고, 향후 필요한 과제들을 제시하였다.

ChatGPT를 계기로 확산된 거대언어모델(LLM)의 핵심 요소 트랜스포머는 언어 영역을 넘어 영상(ViT), 다변량 시계열(TST) 등 다양한 분야로 확장되고 있으며, LLM은 에이전트 형태로 진화하여 과학 연구 자동화(구글 부과학자, AlphaEvolve, AlphaGenome 등)에까지 활용되고 있다. 이러한 발전은 사회 전반의 변혁 가능성에 대한 기대를 높이고 있다.

그렇지만, 실제 기술 수준은 이러한 기대와 간극이 존재하며, 데이터 편향과 환각(hallucination)과 같은 구조적 문제와 위험성이 지속적으로 나타나고 있다. 특히 환각 현상은 모델이 평가 기대를 극대화하려는 과정에서 발생하는 특성으로 설명되며, AI 안전성 확보의 핵심 과제로 지적된다.

이에 따라 AI의 안전한 활용을 위해 모델 동작의 모니터링과 CoT 기반 추론 과정의 해석 가능성 확보, 레드팀 활동(거시적·미시적 수준), 그리고 보건의료 분야를 포함한 국제적 거버넌스 체계(WHO 등)의 중요성이 강조된다.

결론적으로 AI는 과학기술뿐 아니라 사회 전반에 큰 변화를 초래하고 있으나, 환각 문제 해결, 지속 학습 시 망각 방지, 데이터 효율성 향상과 같은 기술적 과제와 더불어, 인간 가치와 정렬된 통제 및 거버넌스 체계 구축이 필수적이다. 특히, 국가 차원의 양극화를 해소하기 위한 국제적 거버넌스 활동이 필요하다.

주제어 ● 트랜스포머, 거대언어모델, 인공지능 에이전트, 환각과 안전, 통제와 거버넌스

Abstract

Advancements in AI Research and Social Responsibility
: From Technology to Control and Governance*

Oh, Sang-Hoon / Mokwon University

Drawing upon recent academic publications and reports, this paper systematically examines current developments in artificial intelligence (AI) through a five-stage framework encompassing technology, applications, expectations and reality, risks and safety, and control and governance. Based on this analysis, the paper identifies key challenges and proposes future directions for AI research and development.

Following the emergence of ChatGPT, large language models (LLMs) have rapidly proliferated. Their core component, the transformer, has expanded beyond natural language processing into diverse domains such as computer vision (Vision Transformer, ViT) and multivariate time-series analysis (Time Series Transformer, TST). Furthermore, LLMs are evolving into agent-based systems and are being applied to the automation of scientific research, as exemplified by Google's Co-Scientist, AlphaEvolve, and AlphaGenome. These developments have significantly heightened expectations regarding the transformative potential of AI across society.

However, a gap remains between these expectations and the current state of technology, as structural issues such as data bias and hallucination continue to pose substantial risks. In particular, hallucination is interpreted as a phenomenon arising from the model's tendency to maximize expected evaluation outcomes, and it is identified as a critical challenge for ensuring AI

*"영문 초록의 표현 교정에 생성형 AI(Claude, Anthropic)를 활용하였다."

safety.

Accordingly, the safe deployment of AI requires effective monitoring of model behavior and improved interpretability of chain-of-thought (CoT) reasoning processes, red-teaming activities at both macro- and micro-levels, and the establishment of international governance frameworks, including those in the healthcare domain such as guidelines from the World Health Organization (WHO).

In conclusion, while AI is driving profound changes not only in science and technology but also across society as a whole, addressing technical challenges—such as mitigating hallucination, preventing catastrophic forgetting in continual learning, and improving data efficiency—must be accompanied by the development of control and governance systems aligned with human values. In particular, international governance initiatives are needed to reduce disparities between countries and address polarization at the global level.

Keyword ● Transformer, Large Language Models, AI Agents, Hallucination and AI Safety, AI Control and Governance

▌논문투고일:2026.05.06. ▌심사완료일:2026.06.24. ▌게재확정일:2026.07.01. ▌