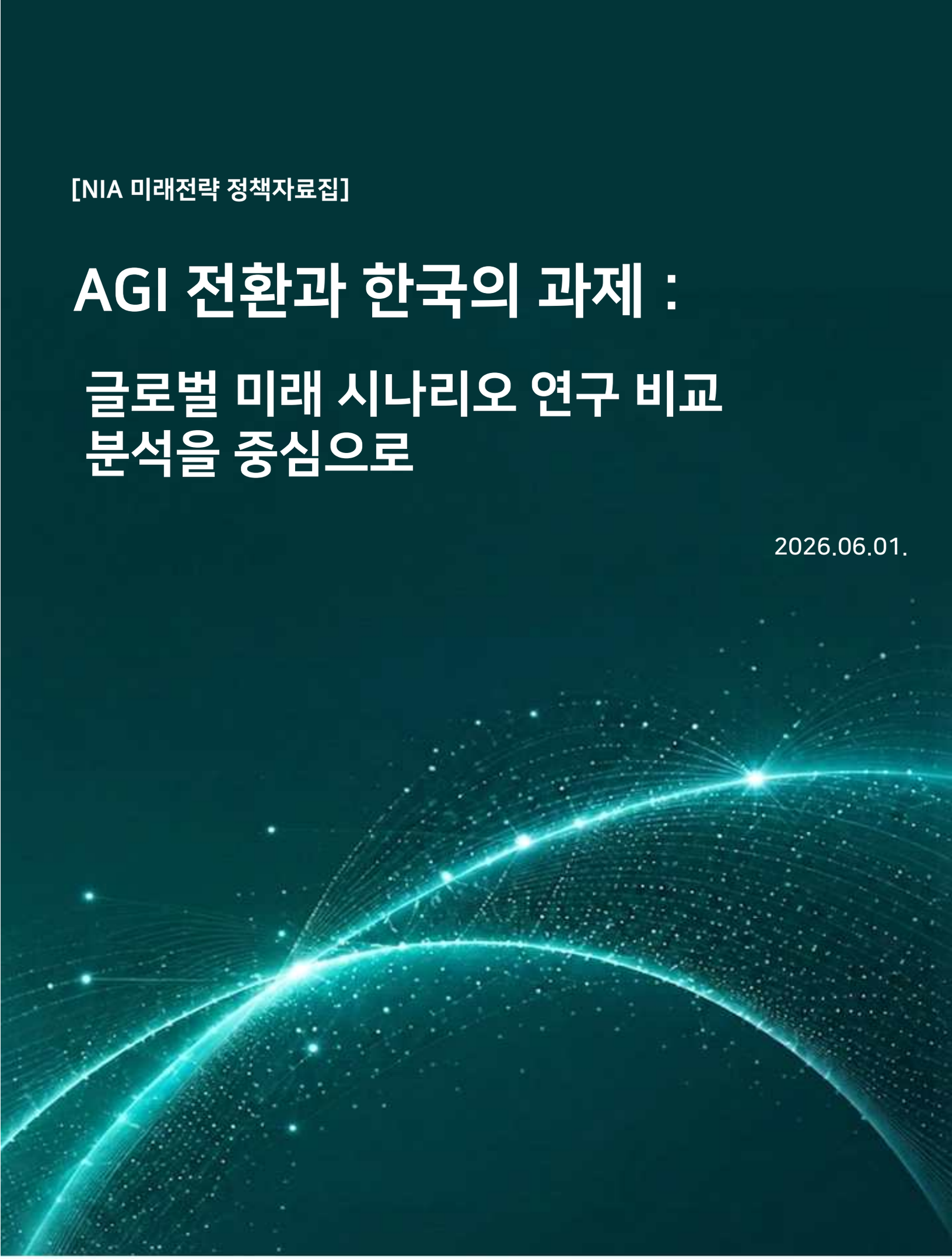
[NIA 미래전략 정책자료집]

AGI 전환과 한국의 과제 :

글로벌 미래 시나리오 연구 비교 분석을 중심으로

2026.06.01.



NIA 한국지능정보사회진흥원

AGI 전환과 한국의 과제 :

글로벌 미래 시나리오 연구 비교 분석을 중심으로

| 작 성 | 한국지능정보원 인공지능정책실 미래전략팀

- 이 정 민 선임연구원 (053-230-1205, jmlee@nia.or.kr)

| 자 문 |

- 서울대학교 김건희 교수
- 한국은행 서동현 박사

| 기 획 | 한국지능정보원 인공지능정책실

- 이 용 진 실장
- 정 지 선 팀장

CONTENTS

연구 요약

1 추진 배경 및 연구 목적

ㅇ (추진 배경) AGI 도달 시점과 사회적 영향에 대한 전망이 빠르게 확산되는 가운데, 미래 시나리오의 설계 범위도 넓어지는 추세

- 컴퓨팅 규모 확장, 알고리즘 효율 향상, 차세대 AI 구조의 혁신 가능성, 지정학 경쟁과 같은 불확실성 요인들로 인해 AGI 도달 경로·시점·사회적 영향력 전망이 다양하게 제시
- 글로벌 차원에서는 주요 싱크탱크·국제기구·연구기관을 중심으로 AGI 미래 시나리오 연구가 활발히 진행 중이며, 한국도 AGI 도래에 대비하기 위한 논의를 시작하는 단계

ㅇ (연구 목적) 글로벌 AGI 시나리오 담론의 체계적 수집·분석을 통한 선행 연구 기반 마련 및 국가 차원의 AGI 대응 정책 시사점 도출

- 주요 글로벌 보고서에 흩어져 있는 AGI 담론을 비교 분석하여 핵심 변수 및 불확실성, 시나리오 패턴, 정책 권고 등을 체계적으로 정리
- 글로벌 담론의 연구 공백을 식별하고 글로벌 시나리오에서의 정책 권고사항을 한국 정책 현황과 교차 검토하여 선제 대응이 필요한 과제를 시사점으로 도출

2 연구 방법론

ㅇ 2024년 이후 발간된 AGI 시나리오 연구 중 8편을 최종 선정하여 비교 문헌 분석 방법론을 적용한 연구 결과 수록

- 8편의 보고서는 다섯 개 축(① AGI 담론 ② 변수와 불확실성 ③ 시나리오 담론 ④ 방법론 ⑤ 정책 함의)을 따라 횡단 비교
- 부록에는 8개 보고서 각각의 표준화된 분석 결과와 함께, 방법론 참고 자료로 활용할 만한 보고서 2편(과학기술정책연구원·한국 인공지능안전연구소) 추가 정리 수록

3 주요 발견

○ **AGI 담론**: 주로 인간 역량을 공통 기준으로 삼아 AGI가 정의되고 도달 시점은 2027~2030년 구간에 모이지만 모델 구조의 대전환 가능성 등 불확실성 잔존

○ **변수와 불확실성**: 기술 개발 속도·지정학 경쟁·안전·정렬 수준이 중요한 변수*로 다루어지며 사회 수용성과 경제 구조 등은 보조 변수에 머무름

* 불확실성 또한 중요 변수에 의거하여 크게 3가지(기술적, 지정학적, 존재론적) 불확실성으로 분류 가능

○ **시나리오 담론**: 39개 시나리오가 다섯 개 서사 패턴(기술 궤적, 소수 집중·불평등, 지정학 경쟁, 거버넌스 위기·협력, 정렬 실패·실존 위협)으로 정리되며, 비관적 시나리오가 다수를 차지해 담론이 위협 요인에 편중된 양상을 보임

- 글로벌 담론이 그리는 가장 가능성 있는 미래는 '고역량 AI는 실현되지만 그 결과가 인간에게 우호적이지 않을 가능성이 높다'는 방향으로 귀결

○ **방법론**: 기술 변수의 정량적 추적과 함께 지정학·안보 함의에 대한 전문가 합의, 워게임, 역사적 사례 유추 등의 정성 방법론을 혼합 활용하여 논증의 설득력 확보

○ **정책 함의**: 국제 협력·안전 평가·민관 협력이 공통 정책 권고로 가장 많이 언급되었으며 어떤 시나리오에도 적용 가능한 유연한 정책 설계가 중요함을 강조

4 정책적 시사점

○ **연구 결과를 한국 관점에서의 시나리오 연구와 정책 논의의 공통 언어로 활용 가능**

- 추세 외삽을 통한 AGI 도달 시점 구간, 미·중 경쟁 구도, 다섯 개 서사 패턴과 방법론은 한국 관점에서의 시나리오 연구의 출발 자원으로 바로 활용할 수 있는 자료

- 한국의 인구 구조·산업 위치·노동시장 특성에서 나타날 수 있는 분기 가능성과 지정학·경제·노동 구조를 통합하는 분석 틀은 앞으로의 연구가 채워나가야 할 과제

○ **AGI 정의 및 발전 경로에 대한 지속적 모니터링 체계 구축이 필요**

- AI가 기존의 스케일링 경로를 따라 발전할지, 아니면 새로운 아키텍처(구조)의 등장으로 도약할지에 따라 정책 대응의 시점과 내용이 달라짐

- 글로벌 보고서들의 정의·도달 기준 비교를 기초 자료로, 한국의 노동 구조·산업 구성·규제 환경에 맞는 기준을 검토할 필요

ㅇ 국제 거버넌스 협력 및 정부-민간 파트너십의 실행 설계 구체화가 필요

- 글로벌 보고서들이 공통적으로 두 가지 권고(국제 거버넌스 협력, 정부-민간 파트너십)를 강조하나, 구체적 실행 방식에 대한 설계는 충분히 다루어지지 않음
- 국제 협력 거버넌스에서 한국이 맡을 수 있는 역할과 참여 방식에 대한 구체적 설계 필요
- 정부-민간 파트너십 역시 어떤 구조로, 어떻게 협력할 것인지를 한국의 맥락에서 설계하는 후속 연구가 필요

ㅇ AGI의 영향은 기술 영역에 국한되지 않으므로 범부처 통합 정책 설계가 필요

- 글로벌 보고서들의 정책 권고는 기술 통제·모니터링·국제 규범 형성에 집중되어 있으나, AI의 영향은 군사·외교·경제·사회 시스템 전반에 걸쳐 동시에 나타남
- 기술 안전이 확보되더라도 사회·경제적 충격은 별개로 관리해야 하며, 각 영역의 정책 이해관계자들이 함께 설계하고 논의하는 범부처 협력 체계가 선결 과제로 대두

ㅇ 불확실성 속에서 유연하게 작동하는 거버넌스 체계의 선제적 구축이 필요

- 어떤 시나리오에서도 유효한 선제 대응과 특정 미래에서만 유효한 집중 조치를 구분하여 병행 준비하는 체계 설계가 주요 과제로 도출
- 이를 뒷받침하기 위한 정책 인프라로서, AGI 도래 이전에 시나리오 변화에 대응하는 상시 점검·갱신 체계 구축이 필요

제1장 개요

01 연구 배경 및 목적

1 연구 배경

2022년 말 ChatGPT 등장 이후 인공지능(AI) 기술은 빠르게 발전해 왔으며, 특히 최근 2~3년 사이 인간의 지적 능력 전반을 수행할 수 있는 범용적 AI, 즉 인공일반지능(Artificial General Intelligence, 이하 AGI)에 관한 논의가 본격화되고 있다. 오픈AI, 구글 등 주요 AI 기업들은 AGI 달성을 공식 목표로 내세우고, 국제기구와 싱크탱크는 AGI 도래가 사회에 미칠 영향을 분석한 시나리오 연구 결과를 잇달아 발간하고 있다. 그러나 AGI는 논의의 주체와 관점에 따라 그 개념과 도달 기준이 천차만별인 상황이다.

실제로 오픈AI는 AGI를 "경제적으로 가치 있는 대부분의 작업에서 인간을 능가하는 고도의 자율 시스템"으로 정의하며 경제적 효용에 방점을 두는 반면[1], 구글은 "인간이 할 수 있는 모든 지적 과제를 이해하고 학습하며 수행할 수 있는 AI"로 정의하며 범용성을 핵심 지표로 삼는다.[2] 두 정의는 모두 인간 수준의 지능을 준거점으로 삼지만, 전자는 노동시장 대체 가능성을, 후자는 인지적 다재다능함을 기준으로 하여 서로 다른 함의를 지닌다. 최근 학계 연구에서는 심리학의 CHC 이론[3]을 빌려 AGI를 "잘 교육받은 성인의 10대 인지 영역에 필적하는 상태"로 계량화하여 정의하려는 시도도 나타나고 있다.[4] 한국 정부는 글로벌 학계(ICLR)의 단계별 기준을 정책 프레임워크로 수용하는 한편[5], 실무적으로는 '인간의 인지 구조를 모방하여 스스로 진화하는 시스템'이라는 기술 구현 목표에 초점을 맞추어 실천적으로 접근하고 있다.[6] 이처럼 AGI는 주체에 따라 경제적 효용성, 인지적 성취 상태, 혹은 실현 가능한 기술적 규격 등 서로 다른 층위에서 다루어지고 있어, 이를 아우르는 글로벌 차원의 통일된 정의는 부재한 상황이다.

---

1) OpenAI(2023.2.24.). Planning for AGI and beyond. https://openai.com/index/planning-for-agi-and-beyond/
2) Google Cloud(2026.1.14.). What is Artificial General Intelligence (AGI)? https://cloud.google.com/discover/what-is-artificial-general-intelligence
3) 카텔-혼-캐롤(CHC) 이론은 인간의 지능을 일반 지능, 10여 개의 광범위한 인지 능력, 70여 개의 세부 역량이라는 3계층 구조로 설명하는 현대 심리학의 가장 포괄적인 지능 구조 이론
4) Hendrycks et al.(2025). A definition of AGI. arXiv. https://doi.org/10.48550/arXiv.2510.18212
5) 과학기술정보통신부(2025.1.24.). 미래 AI 판도를 흔들 범용인공지능(AGI) 핵심원천기술 확보 도전. https://eiec.kdi.re.kr/policy/callDownload.do?num=262799&filenum=2&dtime=20250131151301
6) 정보통신기획평가원(2026.4.23.). 인간지향적차세대도전AI기술개발. https://www.iitp.kr/web/lay1/program/S1T43C47/business/view.do?seq=524

AGI 도달 시점과 관련해서는 오픈AI의 CEO 샘 알트먼(Sam Altman)은 2024년 말, "초지능이 몇 천 일 내에 도달 가능하다"고 전망하였다.[7] 앤트로픽의 CEO 다리오 아모데이(Dario Amodei)는 자신의 에세이 「사랑과 은총의 기계들(Machines of Loving Grace)」에서 2~3년 내에 '데이터센터 속의 천재 집단' 수준의 AI 시스템이 출현할 것이라 전망하였다. 아울러 이 시스템의 주도로 이후 5~10년 내에 질병 정복과 경제 성장 가속 등 급진적인 인류적 혜택이 실현될 수 있다고 진단한 바 있다.[8] 딥마인드의 CEO 데미스 하사비스(Demis Hassabis)는 지속적 학습과 장기 추론의 한계를 지적하면서도, 5~10년 내 AGI 실현 가능성을 50%로 제시하였다.[9] 전직 오픈AI 연구원 레오폴드 아셴브레너(Leopold Aschenbrenner)는 연산 능력과 알고리즘 효율성의 지수적 증가, 즉 'OOMs(Orders of Magnitude)' 법칙[10]을 근거로 2027년 AGI 도달을 주장하며, 이후 급속한 인공초지능(Artificial Superintelligence, 이하 ASI) 전환까지 예고하였다.[11] 그러나 AI 과학자 얀 르쿤(Yann LeCun)[12]과 앤드류 응(Andrew Ng)[13] 등은 현재 접근 방식으로는 AGI 도달에 수십 년이 더 필요하고 근본적으로 다른 기술적 전환이 필요하다고 보고 있다.

예측 기관의 집계에서도 AGI 도달 시점에 대한 인식 차이가 확인된다. AI멀티플(AIMultiple)이 AI 과학자, 기업인, 커뮤니티 예측가 등 9,800개 이상의 예측을 분석한 보고서에 따르면, 기업 경영자 그룹은 AGI 도래를 2029~2032년으로 전망하는 반면 학자들은 이론적 미해결 문제를 근거로 2040~2050년에 초점을 맞추는 경향을 보인다.[14]

〈표 1〉 글로벌 AGI 도달 전망 및 주요 관점 비교

주요 인물	예상 시점	핵심 관점
샘 알트먼	2025~2029년	- AI의 인간 수준 추론·문제 해결이 임박했으며 데이터센터 내 지능 총량이 인류를 능가할 것이라 주장
레오폴드 아셴브레너	2027년	- OOMs 법칙 기반의 AGI 도달 필연성 및 ASI으로의 가속 주장

7) Altman, S. (2024.9.23.). The intelligence age. OpenAI Blog. https://ia.samaltman.com/

8) Amodei, D. (2024.10.). Machines of loving grace: How AI could transform the world for the better. Dario Amodei's Blog. https://darioamodei.com/machines-of-loving-grace

9) Hassabis, D. (2024). Nobel Prize lecture 및 다수 인터뷰에서 발언. 딥마인드의 공식 입장은 AGI를 과학적 발견의 도구로 접근하는 것을 강조

10) OOMs(Orders of Magnitude) 법칙은 인공지능의 지능 수준이 투입되는 계산량, 데이터, 알고리즘 효율성 등의 핵심 자원이 자릿수(10배) 단위로 지수적으로 증가함에 따라 비약적으로 향상된다는 스케일링 법칙의 원리를 의미

11) Aschenbrenner, L. (2024). Situational awareness: The decade ahead. https://situational-awareness.ai/

12) LeCun, Y. (2024). 다수의 공개 인터뷰 및 소셜미디어 게시물에서 현재 LLM의 한계와 AGI 실현을 위한 새로운 아키텍처 필요성을 지속적으로 주장

13) Fast Company. (2026). Andrew Ng says AGI is decades away. https://www.fastcompany.com/91499247/andrew-ng-agi-decades-away-interview

14) AIMultiple. (2026.2.). AGI/Singularity: 9,800 predictions analyzed. https://aimultiple.com/artificial-general-intelligence-singularity-timing

다리오 아모데이	2026~2027년	- AI가 데이터센터에 몰려 '천재들로 이루어진 국가' 수준의 지적 생산성을 갖출 수 있다고 강조 - 강력한 AI 도래 시 질병 정복·경제성장 10배 가속 등 긍정적 혜택 강조
데미스 하사비스	5~10년 이내	- 지속적 학습·장기 추론의 한계를 지적하며 과학적 발견의 도구로서 신중한 접근 강조
OECD 및 주요 학계	2030~2050년	- 기술 낙관·비관 어느 쪽도 취하지 않는 균형적 접근을 취하며 유연한 거버넌스 설계 강조
얀 르쿤, 게리 마커스, 앤드류 응	수십 년 이상	- 현재 AI는 통계적 패턴 매칭에 불과하며 진정한 이해·의식 부재, AGI 도달은 불가능에 가까움

이렇게 전망이 다각화되며 제시되는 시나리오의 범위도 매우 넓어지고 있다. 다양한 시나리오가 동시에 제시되는 상황은 의사결정자 입장에서는 어떤 전망을 참고해 준비해야 할지 판단하기 어렵게 만드는 요인이 된다. 이러한 맥락에서 지금까지 발간된 시나리오 연구들을 체계적으로 정리하고, 각 연구가 어떤 전제와 관점에서 미래를 구성하는지를 비교해보는 작업이 필요하다. 동시에 글로벌 담론을 정리하는 것만으로는 충분하지 않다. AGI라는 기술이 각 국가에 미치는 영향은 그 사회가 가진 고유한 조건에 따라 다르게 나타날 수 있기 때문에, 한국적 맥락에서의 검토도 함께 이루어질 필요가 있다.

그럼에도 한국적 맥락을 반영한 AGI 시나리오 연구는 아직 활발히 진행되고 있지 않다. 국내 연구 동향을 보면 AI가 노동시장에 미치는 영향에 관한 연구는 점차 늘어나고 있으나, 대부분 현재 수준의 AI 기술을 대상으로 한 분석에 집중되어 있다. AGI 수준의 변화를 선제적으로 검토한 연구는 상대적으로 드문 편이다. 글로벌 담론에서 발간되는 시나리오들을 체계적으로 정리하고 한국적 맥락에서 검토가 필요한 영역을 함께 다루는 작업이 이루어질 때, 한국 사회가 AGI 도래에 대비하는 데 필요한 기초 자료가 마련될 수 있다.

2 연구 목적

본편의 목적은 **최근 발간된 글로벌 AGI 시나리오 연구들을 체계적으로 비교 분석함으로써, 향후 한국적 맥락의 시나리오 연구가 참고하고 보완해야 할 시사점을 도출하는 것**이다. 시나리오를 직접 설계하기에 앞서 후속 연구의 토대를 다지는 본편의 결과물(8편의 글로벌 보고서 분석 및 5대 축의 비교 내용)은 시리즈의 후속 연구를 위한 기초 자원이 되는 동시에, 글로벌 AGI 담론의 지형을 파악하고자 하는 독자들에게도 유용한 참고 자료로 기능할 것이다.

02 연구 방법론

1 연구 설계와 범위

본 연구는 비교 문헌 분석(Comparative Document Analysis) 방법론을 적용하여, 분석 대상 문헌들이 어떤 구조와 전제를 바탕으로 AGI의 미래를 전망하고 있는지 서로 비교한다. 비교 문헌 분석은 성격이 다른 여러 문헌을 같은 분석 틀로 검토하여, 문헌 간에 어떤 점이 공통으로 나타나고 어떤 점이 다르게 나타나는지를 살피는 데 적합한 방법론이다. 본 연구의 분석 대상은 학술 논문, 싱크탱크 보고서, 정부 미래 전략 문서, 독립 연구자의 저술 등 형식과 목적이 서로 다른 문헌들로 구성되는데, 이는 AGI 도래를 둘러싼 담론이 하나의 학술 영역이 아니라 여러 제도적 위치에서 생산되고 있다는 현실을 반영한 결과다.

본 연구는 크게 세 단계로 진행되었다. 1단계는 탐색과 선정 단계로, AGI·ASI 수준의 미래 시나리오를 다루는 문헌을 폭넓게 탐색하고 사전에 설정된 기준에 따라 분석 대상을 확정하는 과정이다. 2단계는 정보 추출과 정리 단계로, 확정된 각 문헌에 표준화된 분석 틀을 적용해 핵심 정보를 뽑아내고 개별 문헌의 시나리오 내러티브와 방법론을 정리하는 과정이다. 이 단계의 결과물은 부록에 수록하였다. 3단계는 비교 분석과 종합 단계로, 정리된 정보를 바탕으로 문헌들 간의 공통점과 차이점을 살피고 글로벌 AGI 담론의 전반적인 흐름을 파악한 뒤, 한국 관점에서의 시나리오 설계를 위한 시사점과 추가 검토가 필요한 영역을 짚는 과정이다.

문헌의 시간적 범위는 발간 시점 기준으로 2024년 이후로 설정하였다. 이 시기는 오픈AI의 o1·o3 시리즈 출시, 딥마인드의 노벨화학상 수상 등 AI 담론에 질적 변화가 나타난 시기이며, 주요 싱크탱크와 국제기구들이 AGI를 본격적인 분석 대상으로 삼기 시작한 시점이기도 하기 때문이다. 지리적 범위에는 별도의 제한을 두지 않았으나, 결과적으로 영미권 문헌이 다수를 차지하였다. 이는 글로벌 AGI 담론이 특정 지역을 중심으로 형성되고 있다는 현재의 흐름을 반영하는 것으로, 이 부분에 대한 논의는 본론에서 별도로 다뤄진다.

2 분석 대상 선정 과정

분석 대상 탐색은 AGI와 관련된 담론 자료의 광범위한 수집에서 출발하였다. 학술 데이터베이스(Google Scholar, SSRN, NBER, arXiv)의 논문, 주요 싱크탱크 및 국제기구(RAND, OECD, Brookings, CIGI, UK Government Office for Science 등)의 발간 보고서, AI 안전·거버넌스 커뮤니티(GovAI, Epoch AI, Forethought Foundation 등)의 공개 문서는 물론, 전문가 인터뷰, 기업 백서, 언론 기사에 이르기까지 AGI 담론과 관련된 출처와 자료를 망라하여 수집하였다. 이 과정에서 수집된 자료는 총 130여 개[15]에 달한다.

수집된 자료에 대해 하기 포함 기준을 적용하여 후보군을 16여 편으로 좁혔다. 이후 연구진이 후보 문헌을 직접 검토하며 최종 분석 대상 8편을 확정하였다.

포함 기준은 다음과 같았다. (1) 2024년 이후 발간된 문헌일 것, (2) AGI 또는 ASI 수준의 AI 발전을 직접적으로 다루거나 그에 준하는 미래 변화를 분석할 것, (3) 명시적 방법론에 근거하여 시나리오 축이 설정되고 내러티브가 작성된 것일 것, (4) 특정 분야에 과도하게 편중된 시나리오가 아닐 것.

16편 중 문헌 분석의 구체성과 다양성 등을 고려하여 최종 분석 대상 8편을 확정하였다.

부록에 별도로 수록한 2편은 한국 관점에서의 시나리오 연구의 방법론 설계에 참고 자료로 활용할 수 있어 포함하였다. 과학기술정책연구원(STEPI)의 「미래 시나리오 도출과 STI 정책 정합성 평가」(2024)[16]는 AGI를 직접 다루지 않으나, 기존 글로벌 시나리오에서 마이크로내러티브를 뽑아 재조합하는 방법론을 한국 맥락에서 구현한 사례로서 방법론적 참고 가치가 있다. 한국 인공지능안전연구소의 「AI 안전 전망보고서」(2025~2026)[17]는 AGI·ASI에 대한 명시적 정의나 시점 전망 없이 안보·안전 담론의 구조 변화를 추적하는 연구로, 본 연구의 시나리오 비교 분석 틀과는 성격이 다르다. 다만 뉴스 키워드 네트워크 분석이라는 정량 방법을 시나리오 도출에 적용하고 있어 방법론적 참고 자료로 활용할 수 있다. 두 문헌의 세부 내용은 부록을 통해 확인할 수 있다.

최종 선정된 8편은 〈표 2〉와 같다. 본 보고서 본문에서는 각 보고서를 반복 지칭할 때 고유 연번과 함께 보고서의 성격을 가장 간결하게 식별할 수 있는 별칭을 함께 사용한다. 별칭은 문헌 유형에 따라 기준을 달리 적용하였다. 독립 연구자나 민간 컨소시엄이 작성한 보고서

15) 본 연구 단계에 활용되지 않았지만, 기초 자료 수집에서 수집된 130여 개의 자료는 후속 연구 파트인 카드덱 이슈 분석의 원자료로 활용

16) https://www.riss.kr/search/detail/DetailView.do?p_mat_type=6b4a196b69d9bee2&control_no=78cc26455332983d

17) https://www.aisi.re.kr

1(AI 2027)과 보고서 2(상황인식)는 보고서 명을 그대로 별칭으로 삼았다. 발간 기관이 명확하고 독자에게 익숙한 경우 기관 약칭을 사용하였으며, 보고서 3은 OECD, 보고서 4는 RAND, 보고서 6은 CFG로 표기하였다. 정부 기관이 주도한 보고서 5와 보고서 8은 발간 주체 국가인 영국과 캐나다로 표기하였다. 보고서 7은 여러 기관 소속 연구자들이 공동 집필한 학술 논문으로 특정 기관을 대표하지 않으므로, 해당 보고서의 방법론적 특징을 반영해 경제모델링으로 표기하였다.

〈표 2〉 연구 시리즈 단계별 구성 및 주요 내용

연번	보고서명 (발간 연도)	발간 기관/ 저자	별칭
1	AI 2027(2025)[18]	AI Futures Project	**AI 2027**
2	Situational Awareness: The Decade Ahead(2024)[19]	Leopold Aschenbrenner	**상황인식**
3	Exploring Possible AI Trajectories Through 2030(2026)[20]	OECD AI Futures Expert Group	**OECD**
4	Visions for Potential AGI Futures(2025)[21]	RAND Corporation	**RAND**
5	AI 2030 Scenarios(2024)[22]	UK Government Office for Science	**영국**
6	Advanced AI: Possible futures(2025)[23]	Centre for future generations	**CFG**
7	Scenarios for the Transition to AGI(2024)[24]	UVA / Anton Korinek 한국은행 / 서동현	**경제모델링**
8	AI National Security Scenarios(2026)[25]	Centre for International Governance Innovation & Privy Council Office of the Government of Canada	**캐나다**

18) ai-2027.com
19) situational-awareness.ai
20) oecd.org/en/publications/exploring-possible-ai-trajectories-through-2030_cb41117a-en.html
21) rand.org/pubs/research_reports/RRA3034-2.html
22) https://assets.publishing.service.gov.uk/media/6808fc002a86d6dfb2b52772/AI_2030_Scenarios_Report.pdf
23) https://cfg.eu/advanced-ai-possible-futures/
24) https://www.nber.org/papers/w32255
25) https://www.cigionline.org/publications/ai-national-security-scenarios/

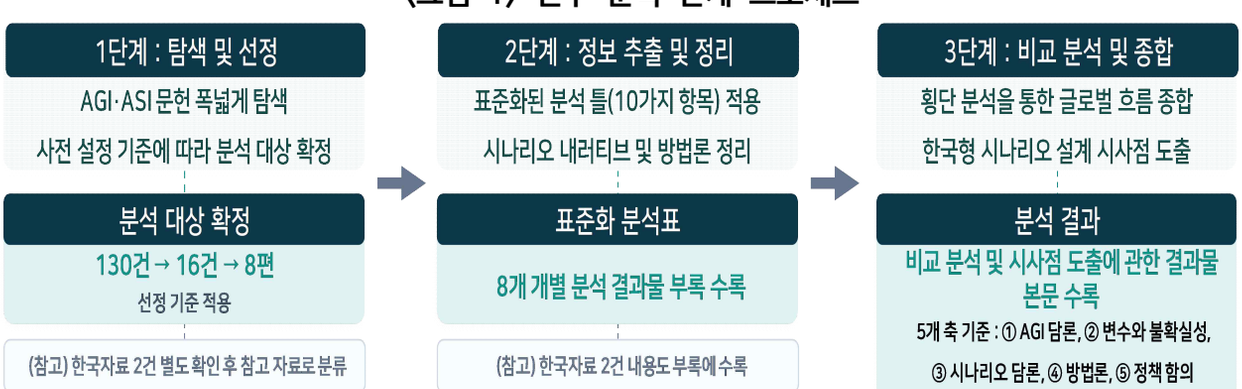
3 분석 프레임워크

각 문헌에 대해 다음 10개 항목으로 구성된 표준화된 분석 틀을 적용하여 정보를 추출하였다. (1) 작성 시기, (2) 작성자 및 기관, (3) 연구 목적, (4) 연구 방법론, (5) AGI/ASI 정의 및 도달 조건, (6) 핵심 동인, (7) 주요 이해관계자, (8) 주요 관점 및 담론적 지향, (9) 시나리오 구조 및 요약, (10) 시사점이다. 이 항목들의 추출 결과는 부록에 보고서별로 수록하였다.

분석 과정에서 연구자의 판단이 개입되는 지점에는 분류 기준을 명시하고 그 근거를 본론에 함께 제시하였다. 시나리오의 낙관·중립·비관 분류, 핵심 변수의 주축·보조 구분, 불확실성의 1차·2차·미포함 층위 구분이 모두 이 원칙에 따른다. 각 기준은 임의로 정한 것이 아니라 보고서 내 시나리오 설계 방식과 변수 활용 방식을 바탕으로 정리한 결과이며, 해당 분류가 본론에서 처음 등장하는 지점에 기준과 근거를 함께 제시하였다.

본론에서는 개별 문헌에서 추출한 내용을 바탕으로 8개 보고서 전체의 공통된 특성과 차이를 살피는 비교 분석을 진행한다. 비교 분석은 ① AGI 담론, ② 변수와 불확실성, ③ 시나리오 담론, ④ 방법론, ⑤ 정책 함의의 5개 축을 중심으로 이루어진다. 이 5개 축은 개별 보고서의 내용을 단순 요약하는 것이 아니라, 보고서들 사이에서 어떤 지점이 공통으로 나타나고 어떤 지점이 다르게 나타나는지, 그리고 담론 전체가 공유하는 전제와 다루어지지 않은 영역이 무엇인지를 확인하기 위해 설정된 것이다. 결론의 분석 요약은 이 비교에서 나온 패턴을 해석하고 시사점으로 연결하는 작업이며, 개별 문헌 분석에서 비교 분석을 거쳐 종합 해석으로 이어지는 세 단계 흐름이 이 연구의 전체 구조를 이룬다.

〈그림 1〉 연구 분석 단계 프로세스



03 분석 대상 보고서 개요

본문으로 넘어가기에 앞서, 독자가 각 보고서의 주요 특징을 미리 파악할 수 있도록 8개 보고서를 간략히 소개한다. 각 보고서가 어떤 접근과 방법으로 AGI 시나리오를 제시하는지를 먼저 이해하면 이후 비교 분석 내용을 따라가기가 수월해지기 때문이다. 보고서별 시나리오 내러티브와 방법론의 세부 내용은 부록에서 확인할 수 있다.

① AI 2027 (AI Futures Project, 2025)

컴퓨팅 스케일링과 알고리즘 효율 향상 추세 예측을 통해 2027년 전후 AGI 달성을 가장 가능성 높은 시나리오로 설정하고, 그에 따른 두 가지 서로 다른 결말을 제시한다. 두 가지 결말이 나뉘는 변수는 '안전 기술과 거버넌스가 늦지 않게 확립되는지 여부'이다. 이 보고서의 핵심 기여는 AGI 도달 시점과 ASI 전환 속도를 80% 신뢰구간으로 명시한 정량적 근거 체계에 있다. 칩 효율, 생산량, 알고리즘 효율, AI 연구개발(R&D) 자동화 승수[26] 등의 계산 구조를 공개함으로써 예측의 조건과 한계를 동시에 제시한다는 점에서, 단순 시나리오 서술을 넘어 예측 방법론 자체를 검토할 수 있는 자료로서의 가치를 지닌다.

② Situational Awareness (Leopold Aschenbrenner, 2024)

컴퓨팅 규모, 알고리즘 효율, 에이전트 언홉블링(unhobbling) 이득[27]이라는 세 독립 동인을 자릿수 단위(OOM, Orders of Magnitude)[28]로 계량화하여 2027년 AGI 달성의 이유를 설명한다. 또한, 해당 보고서는 복수의 시나리오를 탐색하는 대신 단일 미래 경로를 선언적으로 서술한다는 점에서 시나리오 연구보다 예측적 논증에 가깝다. 이 보고서가 분석적으로 중요한 이유는 AGI를 미국의 국가안보 문제로 규정하고 맨해튼 프로젝트에 비견되는 정부 주도 개입을 촉구하는 데 주요 목적이 있다는 점이다. 이후 여러 보고서에서 반복되는 미·중 경쟁 구도, AI 보안 취약성 논의, 국가 개입 필요성 논의의 초기 참조점 역할을 하며, 해당 담론이 어디서 출발했는지를 추적하는 데 참고할 수 있다.

---

26) AI가 인공지능 연구개발 과정을 스스로 수행하여 기술 혁신 주기를 단축시키고, 투입되는 컴퓨팅 자원 대비 지능 발전의 효율을 기하급수적으로 증폭시키는 배수 효과

27) 추가적인 대규모 학습 없이도 모델을 둘러싼 에이전트 제약해소를 통해 기존 모델이 가진 잠재적 지능을 해방함으로써 얻는 비약적인 성능 향상 효과를 의미하며, 이는 추론 프로세스 개선, 외부 도구 활용 능력 부여 등을 통해 AI가 실질적인

28) 자릿수 단위(OOM, Orders of Magnitude)는 어떤 수치의 크기나 규모가 약 10배씩 차이 나는 단계를 나타내는 단위이자 개념으로 세세한 수치적 정확도보다는 전체 규모의 스케일을 직관적으로 비교하고 파악하기 위해 주로 사용됨

③ Exploring Possible AI Trajectories through 2030 (OECD, 2026)

2030년까지의 AI 발전 경로를 정체·둔화·지속·가속의 4개 궤적과 각 2개 변형으로 구성된 8개 시나리오로 제시한다. AGI를 단일 시점의 달성 여부로 정의하는 대신, 언어·문제해결·창의성·메타인지 등 9개 역량 지표를 1~5점 이상의 척도로 수치화하여 시나리오별 역량 수준을 비교 가능한 형태로 제시하는 방식이 이 보고서의 방법론적 특징이다. 또한, 작성 기관인 국제기구의 목적에 걸맞게 복수의 미래에 대응 가능한 유연한 거버넌스 설계의 중요성을 제안한다.

④ Visions for Potential AGI Futures (RAND Corporation, 2025)

AGI가 지정학 질서를 어떻게 재편하는지를 탐색한다는 점에서, 기술 역량 발전 속도 그 자체보다 그 결과에 초점을 두는 시나리오 연구다. 개발 주체 집중도와 지정학적 수혜자라는 두 축으로 8개 시나리오를 도출하며, 미국 단독 패권에서 AGI 통제 상실, 군사 충돌로 인한 개발 중단까지 폭넓은 미래를 포괄한다. 분석 시각은 미국 중심적이며, 시나리오 구성과 함의 도출 모두 미국과 중국의 경쟁 구도를 기본 전제로 한다. 방법론적으로는 AI 거버넌스·기술 분야 전문가 26명을 대상으로 한 반구조화 인터뷰가 시나리오 설계에 직접 반영되었다는 점이 주목할 만하다. 해당 보고서에서 밝히고 있듯, 개발 집중도라는 핵심 축과 4개 지정학적 결과 범주 모두 인터뷰에서 반복 등장한 주제를 정리한 결과로, 전문가 판단이 분석 프레임 자체를 형성하는 데 어느 정도까지 기여할 수 있는지를 보여주는 방법론적 사례로 참고할 수 있다.

⑤ AI 2030 Scenarios (UK GO-Science, 2024)

8개 보고서 가운데 미래연구 방법론을 가장 체계적으로 적용한 사례다. 핵심 불확실성 27개를 식별하고 5개 클러스터로 집약, 70명 이상의 전문가 워크숍을 통해 시나리오 도출, 시민 참여 조사를 통해 대중 반응을 검증하는 단계가 순차적으로 수행되었다. 특히 일반 국민 대상 인식 조사를 방법론에 포함한 것은 분석 대상 보고서 중 유일하며, 제한적이지만 대중의 의견 수렴 방법론을 시나리오 연구에 적용한 사례로 주목할 만하다. 각 시나리오에는 주요 전환점, 기회와 위기, 정책적 함의가 별도로 서술되어 있으며, 국민 토론의 결과도 정책 함의 도출에 반영되었다.

⑥ Advanced AI: Possible Futures (CFG, 2025)

정량적 지표와 정성적 전문가 판단을 결합한 혼합 방법론이 핵심 특장점이다. 형태론적 분석(morphological analysis)[29]으로 9가지 요인으로 발생 가능한 512가지 경우의 수 중,

모순되거나 비현실적인 조합을 제외하고 10개의 시나리오를 도출했다. 여기에 20명의 전문가 의견(델파이 조사)을 반영하여 시나리오의 타당성을 최종 검증하였다. 정량 측면에서는 최대 학습 연산량(FLOP)[30]와 AI 연구개발 자동화 수준을 RE-Bench[31] 점수로 수치화하여 시나리오별로 비교 제시하는데, 이 방식을 채택한 보고서는 분석 대상 8편 중 이 보고서가 유일하다. 본 연구는 유럽·비서방의 기술 주권을 미·중 구도에 대응하는 독립적 관점에서 제기했다는 점에서 차별화되며, '5대 시나리오별 복수 결말' 구조를 통해 동일 경로 내에서도 결과가 다변화되는 조건을 보여준다.

⑦ Scenarios for the Transition to AGI (Anton Korinek·서동현, 2024)

학술 논문 형식으로 발표된 연구로, 시나리오 내러티브의 서술보다 수학 모형에 의한 분석이 중심을 이룬다. 상수탄력대체(CES, Constant Elasticity of Substitution) 생산함수[32]와 업무 복잡도 분포 가정을 활용하여, AGI로의 기술 진보 시나리오별로 생산성과 임금이 어떤 경로를 밟는지를 도출한다. 이 보고서의 핵심 논지는 AGI 이후 결과를 결정하는 핵심 변수가 자동화 속도 그 자체가 아니라 자동화 속도와 자본 축적 속도의 경쟁이라는 점이다. 어느 쪽이 빠른가에 따라 임금 붕괴의 시점과 회복 가능성이 달라지며, 이는 투자 정책·자동화 속도 조율 같은 정책 변수가 결과에 실질적으로 개입할 수 있음을 수식으로 보여준다.

⑧ AI National Security Scenarios (CIGI & 캐나다 추밀원, 2026)

분석 대상 8편 가운데 시나리오 설계 방법론이 가장 명시적이고 다차원적이다. AI 역량 성장 속도, 통제가능성 및 정렬 수준, 프런티어 개발자 수 및 역량 격차라는 세 축의 조합으로 30개 가능 시나리오 풀을 구성한 뒤, 국가안보 관련성과 거버넌스 수준을 기준으로 5개를 선별한다. 분석 시점을 2030년으로 고정하고, 워크숍을 통해 참여형 시나리오 방법론을 적용하여 약 30명의 참가자가 시나리오별로 심층 토론하며 내러티브와 정책 대응을 검증·보완하는 정성적 절차를 밟는다. 중견국 캐나다의 입장에서 미·중 경쟁 구도 속 독립적 거버넌스 구축의 필요성을 강조하며, 각 시나리오에서 도출된 정책 함의를 선제 대응과 전략적 투자의 두 유형으로 구분하여 제시한다는 점도 주요 특징이다.

---

29) 스위스의 천체물리학자 프리츠 츠비키(Fritz Zwicky)가 제안한 방법론으로, 해결하려는 문제의 모든 변수와 그 변수가 가질 수 있는 값들을 조합하여, 논리적으로 가능한 모든 대안(시나리오)을 찾아내는 구조적 방법

30) 특정 AI 모델을 개발하기 위해 훈련 과정에서 사용된 전체 연산량의 총합으로 FLOP(Floating Point Operation, 부동소수점 연산)이라는 단위로 측정

31) AI 에이전트가 스스로 모델을 개선하거나 연구를 수행할 수 있는 '자율적 R&D 능력'을 측정하기 위해, 복잡한 머신러닝 실험 과제로 구성된 고난도 벤치마크

32) 생산의 여러 요소(일반적으로 노동과 자본)가 서로 얼마나 잘 대체될 수 있는지를 나타내는 '대체탄력성'이 생산 과정 어디서나 일정하다고 가정하는 함수로 1961년 애로우(Arrow), 체너리(Chenery), 민하스(Minhas), 솔로우(Solow) 등이 기존의 콥-더글라스(Cobb-Douglas) 함수를 일반화하여 고안

〈표 3〉 분석 대상 보고서별 주요 특징

연번	보고서명	작성기관 유형	시나리오 시점	시나리오 수	주요 방법론 특징
1 AI 2027	AI 2027	독립 연구 컨소시엄	~2030	2개	- 컴퓨팅·알고리즘 성장률 수치 외삽, 80% 신뢰구간 명시 - 단일 전제 하 두 결말 분기 서사 구조
2 상황인식	Situational Awareness: The Decade Ahead	개인 연구자	~2030	1개	- OOM 카운팅: 컴퓨팅·알고리즘·언홉블링 3동인 합산 계량화 - 복수 시나리오 없이 단일 미래 서술 - 미국 국가안보 패권 프레이밍, 맨해튼 프로젝트 유추
3 OECD	Exploring Possible AI Trajectories Through 2030	국제기구	~2030	4궤적 × 2변형 = 8개	- OECD 9개 역량 지표(1~5+점)로 시나리오별 수치화 - AGI 명시적 정의 회피, '인간 역량 정렬' 기준 대체 - 국제기구 특성상 중립적 시나리오들을 제시
4 RAND	Visions for Potential AGI Futures	미국 싱크탱크	중기	8개	- 개발 집중도 × 지정학적 수혜자 2축 매트릭스 설계 - 전문가 반구조화 인터뷰 + 시나리오 시뮬레이션
5 영국	AI 2030 Scenarios	영국 정부기관	~2030	5개	- 불확실성 27개→5개 클러스터링 후 워크숍(70명+) 시나리오 도출 - 시민 참여 조사로 대중 반응을 정책에 반영 - 정책 개입 없이 긍정적 결과 불가 전제
6 CFG	Advanced AI: Possible futures	유럽 싱크탱크	2025~2032	5 × 2엔딩 = 10개	- 형태론적 분석: 9개 요인 → 512 조합 → 10개 시나리오 선별 - FLOP·RE-Bench 점수로 시나리오별 정량 비교 - EU·중간국 기술 주권 관점 명시
7 경제모델링	Scenarios for the Transition to AGI	연구논문	5~20년	4개	- 이론 경제모델: CES 생산함수 + 수치 시뮬레이션 - 업무 복잡도 분포(유한/무한) × 자동화 속도 변수 설계 - 임금·산출량 경로를 수식으로 도출, 노동경제 분배 분석
8 캐나다	AI National Security Scenarios	캐나다 정부 및 싱크탱크	~2030	5개	- 3축 조합 → 30개 풀 → 5개 선별 (국가안보 관련성 기준) - 중견국(캐나다) 시각, 5년 내 AGI/ASI 포함 全 범위 대비 - 아날로그 백업·킬 스위치 등 가장 구체적인 비상 대응 처방

제2장 본론

01 AGI 담론 비교 분석

1 AGI 정의 비교

〈표 4〉 보고서별 AGI·ASI 정의 비교

연번	AGI 정의 내용	ASI 정의 내용
1 AI 2027	- 컴퓨터로 업무를 수행하는 지식 노동자의 모든 업무를 완전 대체할 수 있는 AI	- AGI 이후 연구 자동화로 수개월~1년 내 지능 폭발해 인간을 넘어서는 수준
2 상황인식	- 다수 분야에서 박사급 전문가를 능가하는 AI	- 인간 전체를 넘어서는 지능. AGI가 AI 연구를 자동화하며 수주~수개월 내 도달
3 OECD	- 9개 지표 모두 인간과 동등한 수준(Level 5)에 도달한 상태	- 인지 관련 7개 지표는 인간을 초월(>5)하고 신체적 역량은 인간과 동일한 상태
4 RAND	- 시나리오 내에서 군사·경제·정보 전 영역에서 결정적 우위를 제공하는 범용 AI로 암묵적 이해	- 모든 분야에서 인간 역량을 압도적으로 능가하는 AI
5 영국	- 대부분의 인지적 과제에서 인간 수준 또는 그 이상의 성능을 달성	명시적 정의 없음
6 CFG	- 5단계 이상 복합 인지 과제를 자율적으로 수행하며 벤치마크 평가가 인간 수준에 도달	- 인간의 피드백이 불가능한 상황에서도 AI가 개발자의 목표를 안정적으로 유지하는 상태
7 경제모델링	- 인간이 수행할 수 있는 모든 업무를 AI 시스템이 수행 가능한 상태	명시적 정의 없음
8 캐나다	- 대부분의 관련 인지 과제에서 인간 수준 도달하거나 초과 - 컴퓨터 앞에서 가능한 어떤 전문가 수준 과제도 수행 가능	- 인간 전체의 지능을 압도적으로 초과하는 AI

8개 보고서 전체가 AGI를 이해할 때 공통으로 사용하는 기준은 인간의 역량과의 비교다. 보고서마다 명시적 정의를 두기도 하고 시나리오 안에서 암묵적으로 서술하기도 하지만, AGI는 대체로 인간 수준에 도달하거나 이를 넘어서는 AI로 이해된다. 일반화 능력(generalization) 또한 공통된 요건이다. 수학이나 코딩처럼 특정 분야에서만 탁월한 성능을 보이는 AI는 AGI로 간주되지 않으며, 광범위한 인지 과제 전반에서 인간 수준을 달

성해야 한다는 점에서 8개 보고서의 관점이 일치한다. 다만 보고서들이 비교 대상이 되는 인간 수준을 어디에 두느냐에 따라 정의가 크게 세 갈래로 갈린다.

첫째는 일반 직장인을 기준으로 삼는 방식이다. AI가 평범한 사무직 노동자의 업무를 대신할 수 있게 되는 시점을 AGI로 본다. 보고서 1(AI 2027)은 컴퓨터로 업무를 수행하는 지식 노동자의 모든 업무를, 보고서 7(경제모델링)은 인간이 할 수 있는 모든 업무(태스크)의 전면 대체를 그 조건으로 제시한다. 이 기준에서 AGI는 노동시장에서 인간을 대체하는 시점과 사실상 같은 의미로 사용된다.

둘째는 전문가를 기준으로 삼는 방식이다. 일반 직장인이 아니라 특정 분야의 숙련된 전문가를 넘어서야 AGI로 본다. 보고서 2(상황인식)는 다수 분야에서 박사급 전문가를 능가하는 수준을 조건으로 삼고, 보고서 6(CFG)은 복잡한 인지 과제를 자율적으로 수행하면서 주요 벤치마크에서 인간 수준을 달성해야 한다고 본다. 보고서 8(캐나다)은 이 두 기준을 결합하여 컴퓨터 앞에서 수행 가능한 어떤 전문가 수준 과제도 처리할 수 있는 상태를 조건으로 제시한다.

셋째는 인간을 단일 기준점이 아닌 측정 가능한 지표들의 집합으로 나누어 보는 방식이다. 보고서 3(OECD)은 언어, 문제해결, 창의성, 메타인지 등 9개 역량 차원을 1점에서 5점 이상의 척도로 수치화하고, 모든 지표가 인간 동등 수준인 5점에 도달한 상태를 AGI에 가까워진 조건으로 설정한다. AGI 달성 여부를 한 번에 선언하는 대신 역량별 진행 정도를 추적할 수 있다는 점에서 가장 세분화된 접근이다.

ASI에 대해서도 유사한 이해를 공유한다. **AGI가 인간과 동등한 수준을 기준으로 삼는다면 ASI는 그것을 넘어서는 단계**이며, 보고서 4(RAND)와 보고서 8(캐나다)은 이를 **모든 분야에서 인간을 압도적으로 넘어서는 수준으로 정의**한다. 보고서 1(AI 2027)과 보고서 2(상황인식)는 여기에 전환 메커니즘을 덧붙인다. AGI가 AI 연구에 투입되기 시작하면 AI가 스스로를 개선하는 순환 구조를 통해 수주에서 수개월 내에 ASI 단계로 빠르게 넘어갈 수 있다는 설명이다. 보고서 3(OECD)은 ASI를 하나의 수준으로 정의하는 대신, 인지 관련 7개 지표는 인간을 넘어서지만 신체적 역량은 인간과 동일한 상태로 구분한다. 역량의 종류에 따라 인간을 넘어서는 정도가 다를 수 있음을 반영한 설계다. 보고서 6(CFG)은 다른 보고서와 달리 정렬(alignment)[33] 확립 여부를 ASI 조건에 포함시킨다. 인간이 피드백을 제공할 수 없는 수준에서도 AI가 개발자의 목표를 안정적으로 유지하는 상태를 ASI 실현의 조건으로 제시하는 것이 이 보고서만의 특징이다.

33) AI 시스템의 목표와 행동이 인간의 의도·가치·윤리적 원칙과 일치하도록 만드는 기술적·철학적 과정을 의미

2 AGI 도달 기준 비교

〈표 5〉 보고서별 AGI·ASI 도달 기준 및 시점

연번	도달 기준	핵심 조건	시점
1 AI 2027	AI R&D 자동화	초인간 코더(SC) 등장으로 AI R&D 재귀 피드백 루프 진입되는 시점 · SC 조건: 80% 신뢰도로 1개월 코딩 프로젝트 수행 + 인간 대비 30배 빠르고 저렴	AGI ~2027 ASI ~2028
2 상황인식	컴퓨팅·알고리즘 스케일링	컴퓨팅 4 OOM + 알고리즘 2 OOM + 언홉블링 이득 누적으로 GPT-4 수준의 질적 도약이 한 번 더 발생하는 시점	AGI 2027 ASI ~2030
3 OECD	역량 지표	9개 역량 지표 전부 인간 수준(5점) 달성 · 물리·사회 역량의 인지 역량 수준 동반 상승 포함	특정 시점 없음
4 RAND	기술 외적 조건	기술적 도달 기점 기준 없음 · 개발 주체 집중도(분산↔집중), 수출 통제·동맹 구도, 정부-민간 파트너십 형태가 시나리오 분기 조건	특정 시점 없음
5 영국	범용성 및 자율성	· 범용성 : 추상적 추론과 기획, 명시적 훈련 없는 과제 수행 · 자율성 : 하위 목표를 스스로 설정하고 외부 도구를 활용하여 인간 개입없이 결과물 도출	특정 시점 없음
6 CFG	AI R&D 자동화	RE-Bench 점수 〉1.7 달성 + AI R&D 3배 이상 가속되어 자기개선 루프 진입되는 시점	시나리오별 상이
7 경제모델링	경제 모델	경제 내 모든 업무가 기술에 의해 전면 자동화 완료된 시점	· **급진 AGI**: 5년 이내 · **기준 AGI**: 20년 이내
8 캐나다	AI R&D 자동화	대부분의 인지 과제 인간 수준 달성 + 수시간~수일 복합 과제 자율 완수 · AGI→ASI 즉시 전환 가능성 명시(AGI는 일시적·불안정 상태)	5년 이내 AGI/ASI 가능성 명시

가장 많은 보고서에서 AI 연구개발(R&D) 자동화를 AGI 도달의 핵심 기점으로 제시하고 있다. 보고서 1(AI 2027), 보고서 6(CFG), 보고서 8(캐나다)이 각각 독자적인 방법론으로 도달 기점을 설정하면서도 공통적으로 **AI가 연구개발 과정에 투입되어 자가발전이 이뤄지는 시점**을 기준으로 제시한다. 이 배경에는 AI R&D 자동화가 선형 성장에서 지수적 폭발로 전환되는 변곡점이라는 공유된 인식이 있다. 인간 연구자가 수행하던 가설 수립·실험 구현·결과 해석의 반복 과정을 AI가 대신하기 시작하면 이후 능력 향상은 인간의 개입 없이 자체 가속되며, 이 지점에 들어서는 것이 AGI 도달의 실질적 기준이라는 것이 세 보고서의 공통 논리다.

그러나 재귀적 자기 개선[34]이라는 결과가 같더라도 그 순환 구조가 시작되기 위해 AI가 먼저 갖추어야 할 역량의 범위와 수준은 보고서마다 다르다. 보고서 1(AI 2027)은 초인간 코더(Superhuman Coder, SC)[35]의 등장을 기점으로 삼으며 조건을 80% 신뢰도로 1개월 코딩 프로젝트를 수행하고 인간 대비 30배 빠르고 저렴한 수준으로 명시한다. 보고서 6(CFG)은 AI 연구개발 과제 전반을 평가하는 RE-Bench 점수 1.7 초과와 AI R&D 3배 이상 가속을 조건으로 제시하여 보고서 1(AI 2027)보다 요구 범위가 넓다. 보고서 8(캐나다)은 대부분의 인지 과제에서 인간 수준 달성과 수 시간에서 수일 단위 복합 과제의 자율 완수를 조건으로 삼으며 세 보고서 중 가장 포괄적인 역량 수준을 요구한다.

일부 보고서는 AGI 도달 조건을 자동화 임계점 자체가 아니라 그에 이르는 물리적·경제적 축적량으로 정의한다. 보고서 2(상황인식)는 AI R&D 자동화를 기점으로 설정한다는 점에서는 앞의 세 보고서와 방향을 공유한다. 다만 자동화가 실현되는 그 시점을 기준으로 삼는 대신, 그 시점에 이르기까지 컴퓨팅과 알고리즘이 얼마나 쌓여야 하는지를 조건으로 제시한다. 보고서 7(경제모델링)은 한 단계 더 나아가 AGI 도달을 기술 역량의 문제 뿐만 아니라 경제 구조의 전환 지점으로 정의한다. AI가 어떤 역량을 갖추었는지보다, 그것이 노동시장에서 인간을 완전히 대체된 상태에 도달했는지가 판별 기준이 된다.

34) 인공지능이 외부(인간 과학자)의 개입 없이, 스스로 자신의 소스 코드와 알고리즘을 분석·수정하여 더 똑똑한 차세대 인공지능을 고도화하고, 그렇게 강력해진 지능을 바탕으로 다음 세대의 인공지능을 훨씬 더 빠른 속도로 개발해 나가는 무한 순환 메커니즘

35) 전 세계에서 가장 뛰어난 인간 소프트웨어 엔지니어 수만 명의 몫을 혼자서, 혹은 수천 개의 복제본으로 수행할 수 있는 AI를 의미

02 변수와 불확실성 비교 분석

1 핵심 변수 비교

〈표 6〉 보고서별 주축·보조 변수

연번	주축 변수	보조 변수
1 AI 2027	컴퓨팅 규모 확장 알고리즘 효율 향상 AI R&D 자동화 미·중 지정학 경쟁 안전 기술·거버넌스 확립 여부	모델 가중치 보안
2 상황인식	컴퓨팅 규모 확장 알고리즘 효율 향상 미·중 지정학 경쟁	AI 군사·정보화 정렬 기술 확보 여부
3 OECD	기술 발전 속도	-
4 RAND	개발 주체 집중도 지정학적 수혜 구도	AGI 정렬 여부 경제·사회적 혼란 적응력 국제 거버넌스 역량
5 영국	역량 발전 수준 소유·접근·제약 안전·정렬 수준 활용 수준·분포 지정학적 맥락	-
6 CFG	역량 발전 속도 개발 주체 집중도	안전·정렬 수준 활용 수준·분포 지정학적 맥락
7 경제모델링	자동화 지수 성장 속도 업무(태스크) 복잡도 분포	-
8 캐나다	역량 발전 속도 통제가능성·정렬 수준 프런티어 행위자 수	미·중 지정학 경쟁 국제 거버넌스 역량

핵심 변수는 시나리오 설계에서 미래가 갈라지는 지점을 만드는 요소를 가리킨다. **어떤 변수를 선택하느냐에 따라 시나리오가 다루는 미래의 범위와 성격이 달라진다.** 〈표 6〉은

각 보고서에서 제시한 핵심 변수를 주축과 보조로 나누어 정리한 것이다. 주축은 시나리오가 실제로 나뉘는 기준이 되거나 서사 전체를 끌고 가는 변수를 가리키며, 보조는 시나리오의 내용을 풍부하게 하되 분기 기준으로는 사용되지 않는 맥락 변수다.

보고서마다 변수를 명시하는 방식이 달라, 주축과 보조를 구분하는 기준도 보고서별로 다르게 적용하였다. 보고서 4(RAND), 보고서 6(CFG), 보고서 8(캐나다)은 핵심 변수와 시나리오 축을 문서에서 구조적으로 분리해 명시하고 있어 구분이 바로 확인된다. 보고서 3(OECD)은 역량 발전 속도 하나를 연속된 스펙트럼으로 전개하는 구성이어서 주축이 기술 변수임이 분명히 드러난다. 보고서 5(영국)는 다섯 가지 핵심 불확실성을 모두 동등한 축으로 다루면서 각 축의 조합 강도로 시나리오를 설계하기 때문에, 주축과 보조의 구분이 적용되지 않는다. 보고서 1(AI 2027)과 보고서 2(상황인식)는 단일 경로를 서술하는 형식이라 서사에서 실질적으로 분기를 만들어내거나 전개 방향을 결정하는 변수를 주축으로 보았다. 보고서 7(경제모델링)은 핵심 변수가 경제 모형의 외생 파라미터[36]로 정의되어 있어 설계 방식 자체가 다른 보고서들과 구조적으로 다르다.

〈표 7〉 보고서별 변수 종류 비교

연번	기술 역량 발전 속도	개발 주체 집중도	안전·정렬 수준	지정학 경쟁	국제 거버넌스	경제·노동 구조
1 AI 2027	●	-	△	△	-	-
2 상황인식	●	-	△	△	-	-
3 OECD	●	-	-	-	-	-
4 RAND	-	●	△	●	△	△
5 영국	●	●	●	●	-	●
6 CFG	●	●	△	△	-	△
7 경제모델링	●	-	-	-	-	●
8 캐나다	●	●	●	△	△	-

※ ● 주축, △ 보조, — 분석 범위 밖

36) 시스템 내부의 상호작용에 의해 변하는 값이 아니라, 외부 환경이나 시스템 설계자에 의해 이미 결정된 채로 시스템 안으로 주어지는 입력값을 의미

8개 보고서는 어떤 변수를 주축으로 삼든 역량 발전 속도를 분석의 출발점으로 공유한다. 이 변수는 〈표 6〉에서 컴퓨팅·알고리즘 확장, AI R&D 자동화, 역량 발전 속도 등으로 나뉘어 표현되는데, 이들은 공통적으로 AI 기술이 얼마나 빠르게 발전하는가에 대한 요소들이기 때문에 〈표 7〉에서는 '기술 역량 발전 속도'로 통합하여 표현했다. 보고서 4(RAND)가 유일하게 해당 변수를 주축으로 두지 않았는데 그 이유는 해당 보고서의 시나리오들은 기술이 발전한다는 것을 전제로 두고 그걸 누가 개발하느냐를 중점적으로 봤기 때문이다.

기술 다음으로 공통성이 높은 변수는 안전·정렬 수준과 지정학 경쟁이다. 두 변수는 보고서 3(OECD)과 보고서 7(경제모델링)을 제외한 6개 보고서에 등장한다. 이는 두 변수가 사실상 하나의 묶음으로 인식되고 있음을 보여준다. **기술이 빠르게 발전할 때 통제를 누가 유지하느냐는 안전·정렬의 문제이고, 개발 주도권을 어느 국가나 행위자가 쥐느냐는 지정학의 문제인데, 이 두 질문은 대부분의 보고서에서 분리되지 않고 함께 다루어진다.** 개발 주체 집중도 또한 보고서 4(RAND)·보고서 5(영국)·보고서 6(CFG)·보고서 8(캐나다)에서 공통적으로 주축 변수로 등장한다. 이 변수가 중요한 이유는 안전·정렬 문제와 지정학 경쟁 문제를 동시에 건드리기 때문이다. **프런티어 AI를 소수가 개발하느냐 다수가 개발하느냐는 질문은 단순한 시장 구조의 문제가 아니다. AI가 통제를 벗어났을 때 누가 책임을 지는지, 그리고 AI 기술의 주도권이 어느 국가나 진영으로 넘어가는지를 함께 결정하는 변수이기 때문이다.**

반면 경제·노동 구조 변수는 시나리오 설계에서 상대적으로 적게 다루어진다. 노동시장 충격과 자동화로 인한 소득 불평등은 여러 보고서에서 정책 시사점으로 언급되지만, 이를 시나리오를 나누는 독립 축으로 설정한 보고서는 보고서 5(영국)와 보고서 7(경제모델링) 둘뿐이다. 보고서 6(CFG)은 경제 도메인을 5개 분석 도메인 중 하나로 포함해 노동과 자본, 일자리 대체, 권력 집중을 체계적으로 다루지만, 이를 시나리오를 갈라놓는 독립 축으로 설계하지는 않았다는 점에서 앞의 두 보고서와는 위상이 다르다.

보고서 7(경제모델링)은 자동화 속도와 자본 축적 속도의 경쟁 구도를 시나리오의 유일한 분기 기준으로 삼아 이 변수를 가장 정밀하게 다루지만, 분석 범위는 경제 모형 안으로 한정된다. 지정학이나 안전·정렬 변수가 포함되지 않기 때문에, 노동 구조 충격이 어떤 지정학 구도나 거버넌스 조건에서 발생하느냐는 질문은 이 보고서의 분석 대상 밖에 놓인다. 고용 충격을 완충하는 사회안전망의 작동 여부, 자동화 이익의 사회적 분배 방식, 저숙련 노동자의 적응 속도 같은 사회적 조건 변수는 보고서 7(경제모델링)을 포함한 8개 보고서 중 어느 곳에서도 시나리오 분기 기준으로 다루어지지 않았다.

지정학 변수는 여러 보고서에서 주축 또는 보조 변수로 등장하지만, 그 내용이 미·중 패권 경쟁에 집중되어 있다는 점에서 실제 분석 범위는 〈표 7〉이 보여주는 것보다 좁은 편이다. 보고서 4(RAND)는 지정학적 수혜 구도를 독립 축으로 설계하여 가장 정교하게 다루지만, 이 구도 역시 미국의 상대적 지위를 기준으로 시나리오를 구성하는 구조를 취한다. 보고서 8(캐나다)은 중견국인 캐나다의 시각에서 미·중 패권 구도를 분석하고, 보고서 6(CFG)은 유럽 싱크탱크로서 EU와 중간국의 역할을 시사점에 부분적으로 포함한다. 그러나 AI 개발·배치가 미·중 외의 행위자들에게 미치는 구조적 영향, 또는 그 행위자들의 선택이 미래 경로를 어떻게 바꿀 수 있는지는 이 보고서군이 충분히 다루지 못한 영역으로 남아 있다.

제도적 조건도 시나리오 설계에서 충분히 다루어지지 않은 영역이다. 민주주의 제도의 역량, 국내 규제 체계의 실효성, AI에 대한 사회적 수용성, 시민 사회의 반응 방식 같은 변수는 8개 보고서 중 어느 곳에서도 시나리오 분기 기준으로 설계되지 않았다. 보고서 3(OECD)이 규제·거버넌스 환경을 보조 변수로 포함하고, 보고서 5(영국)가 활용 수준·분포 축 안에 대중 참여를 부분적으로 반영하는 것이 이 영역에서 가장 적극적인 시도다. 기후·에너지 전환이나 인구구조 변화 같은 AI 외부의 거시 충격이 AI 발전 경로와 교차하는 방식도 제한적으로만 다루어진다. 보고서 3(OECD)과 보고서 6(CFG)이 전력 제약·칩 공급 병목을 AI 역량 발전의 투입 조건으로 포함하고 있지만, 기후·에너지 전환이나 인구구조 변화가 AI 시나리오와 어떻게 교차하는지는 8개 보고서 모두에서 시나리오 분기 기준으로 다루어지지 않는다. 이 공백들은 이 보고서군이 공유하는 분석의 초점이 어디에 있는지를 보여준다. AI 미래 시나리오 담론이 기술 역량 변수와 미·중 지정학 변수 주변에 집중적으로 형성되어 있다는 사실 자체가 어떤 불확실성이 정책적으로 중요하게 인식되고 있는지를 보여주는 하나의 지표라 할 수 있다.

2 불확실성의 위계 분석

불확실성의 위계 분석은 핵심 변수 비교에서 한 걸음 더 나아가, 각 보고서가 선택한 변수들이 어떤 성격의 불확실성을 다루는지를 층위로 분류하는 작업이다. 같은 변수라도 어떤 보고서는 이를 관찰 가능하고 수치화가 가능한 문제로 다루고, 다른 보고서는 선례가 없어 원인과 결과를 파악하기 어려운 문제로 다룬다. 이와 같은 접근 층위의 다변화는 시나리오의 도출 규모, 전개 경로(분기 방식), 그리고 최종 정책 제안의 방향성을 결정짓는 요인이 된다.

'기술적 불확실성'은 컴퓨팅 규모 확장, 알고리즘 효율 향상, AI R&D 자동화 임계점 도달 시점처럼 원칙적으로 관찰이 가능하고 추세 외삽[37]의 기반이 되는 변수들을 다루는 층위다. 이 층위를 1차로 설정한 보고서는 7개이며, 이 군에서 유일하게 공통된 1차 층위다. 보고서 4(RAND)는 AGI 실현 자체를 이미 해소된 전제로 놓고 분석을 시작하기 때문에, 기술 역량의 발전 경로는 탐색 대상이 아니라 고정된 출발 조건으로 기능한다.

'지정학적 불확실성'은 국가와 기업 같은 주요 행위자들의 선택에서 비롯되는 불확실성을 가리킨다. 미·중 경쟁 구도의 전개 방향, 수출 통제의 실효성, 국제 거버넌스 정립 여부 같은 변수가 여기에 해당한다. 이 층위의 불확실성은 기술적 변수와 달리 행위자들의 선택과 상호 대응의 결과로 결정되기 때문에 추세 외삽보다는 시나리오 분기를 통한 접근에 적합하며, 분기의 수가 늘어나는 경향이 있다.

'존재론적 불확실성'은 AI 정렬 실패 가능성, ASI 출현 이후 인간의 통제 가능성, AI 시스템의 목표 변질과 같은 불확실성을 다루는 층위다. 이 층위를 1차로 설정하는 보고서는 분기 기준을 정렬 달성 여부 하나로 모으거나, 인류 존속 자체를 시나리오의 결말 변수로 삼는 구조를 취한다. 그러나 존재론적 불확실성은 세 층위 중 가장 다루기 어려운 영역으로, 이를 포함하는 보고서들조차 대부분 중심이 아닌 주변부에 배치한다. 보조 변수로 분류한 보고서들이 공통적으로 취하는 방식은 이 층위를 시나리오 분기의 기준으로 삼는 대신 특정 극단 시나리오의 결말 안에 녹여 넣는 방식이다.

어떤 층위를 1차로 설정하느냐는 보고서의 정책 지향과 연결된다는 점이 이 비교에서 주목할 만한 함의다. 세 층위 모두를 1차로 설정하는 보고서 1(AI 2027)과 보고서 8(캐나다)은 인류 통제 유지를 정책의 최우선 과제로 설정하고 국제 협약, 비상 대응 프로토콜,

37) 과거부터 현재까지 축적된 시계열 데이터를 바탕으로 그 데이터가 보여주는 일정한 흐름(추세)을 찾아내고, 이 패턴이 미래에도 동일하게 지속될 것이라고 가정하여 미래의 수치를 예측하는 통계적 방법론

정렬에 대한 연구 투자를 핵심 권고로 제시한다. 기술적 불확실성만 1차로 두는 보고서 3(OECD)과 보고서 7(경제모델링)은 규제 거버넌스 정비나 노동시장 재편처럼 현재 정책 체계 안에서 작동 가능한 대응을 권고한다. 행위자의 전략적 선택을 분석 범위에 포함하지 않는 만큼 정책 권고의 대상도 기술·경제 조건의 관리로 좁혀진다. 지정학적 불확실성을 1차로 두는 보고서 4(RAND)·보고서 6(CFG)·보고서 8(캐나다)은 행위자 간 협력 설계와 집중화 수준 조절을 핵심 과제로 설정한다. **결국 어떤 층위를 1차로 놓느냐는 단순한 분석 범위의 선택이 아니라, AI가 초래할 수 있는 최악의 결과를 정책 설계의 전제로 삼을 것인지, 현재 정치적으로 실행 가능한 대응을 기준으로 삼을 것인지에 대한 방향 선택을 반영하기도 한다.**

〈표 8〉 보고서별 불확실성 위계 비교

연번	기술적 불확실성	지정학적 불확실성	존재론적 불확실성
1 AI 2027	1차	2차	1차
2 상황인식	1차	2차	2차
3 OECD	1차	-	-
4 RAND	-	1차	2차
5 영국	1차	1차	2차
6 CFG	1차	1차	2차
7 경제모델링	1차	-	-
8 캐나다	1차	1차	1차

※ 1차 : 시나리오 설계의 직접 분기 기준 또는 서사 전체의 구동력
2차 : 시나리오 내 맥락·보조 변수로 포함, 분기 기준으로는 미사용

3 시나리오 축 설계 비교

〈표 9〉 보고서별 시나리오 설계 방식

연번	축 설계 방식	축 개수 및 내용	시나리오 개수
1 AI 2027	단일 경로 서술형	축 설계 없이 단일 전제 아래 서사 전개. 분기 변수는 '안전 기술·거버넌스 확립 여부'	2
2 상황인식	단일 경로 서술형	미·중 패권 경쟁을 전제로 단일 미래 경로 서술	1
3 OECD	단일 스펙트럼형	① 역량 발전 속도 (정체 → 둔화 → 지속 → 가속)	4궤적×2변형
4 RAND	2축 교차형	① 개발 주체 집중도 (분산 ↔ 집중) ② 지정학적 수혜 구도 (미국 강화 / 적대국 강화 / 동시 약화 / 개발 중단)	8
5 영국	다축 강도 조합형	① 역량 발전 수준 ② 소유·접근·제약 ③ 안전·정렬 수준 ④ 활용 수준·분포 ⑤ 지정학적 맥락	5
6 CFG	2축 교차형	① 역량 발전 속도 (정체 ↔ 초가속) ② 개발 주체 집중도 (분산·오픈 ↔ 소수 독점)	5×2엔딩
7 경제모델링	모형 파라미터 조합형	① 업무(태스크) 복잡도 분포 (유한 ↔ 무한) ② 자동화 지수 성장 속도	4
8 캐나다	3축 선별형	① 역량 성장 속도 (정체 ↔ 지수적) ② 통제가능성·정렬 수준 (높음 ↔ 낮음) ③ 프런티어 행위자 수 (다수 ↔ 단일)	5

8개 보고서의 시나리오 축 설계에서 드러나는 가장 중요한 사실은, **핵심 변수로 인식된 것과 실제로 축으로 전환된 것 사이에 간극이 존재한다는 점이다.** 앞서 살펴본 핵심 변수 비교에서 지정학 경쟁은 6개 보고서에서 중요한 변수로 인식되었지만, 독립적인 시나리오 축으로 전환한 경우는 보고서 4(RAND)뿐이다. 안전·정렬은 7개 보고서에서 언급되지만 축으로 설계한 보고서는 보고서 8(캐나다)뿐이다. 이 간극은 단순한 설계상의 선택이 아니다. 어떤 변수를 축으로 삼는다는 것은 그 변수를 양 끝단으로 나누거나 정도의 차이가 있는 스펙트럼으로 다룰 수 있다는 판단, 즉 불확실성의 양 끝을 설정할 수 있다는 전제가 있어야 한다. 이 조건을 충족하지 못하는 변수는 중요하게 인식되더라도 축이 되지 못하고 보조적인 서술로 처리된다.

보고서들에서 공통적으로 시나리오 축으로 채택된 변수는 기술 역량 발전 속도다. 보고서 3(OECD)·보고서 6(CFG)·보고서 8(캐나다)이 모두 이 변수를 독립 축으로 설계했으

며, 보고서 1(AI 2027)도 축 설계 없이 기술 가속을 전제로 확정하고 출발한다는 점에서 역량 발전 경로를 시나리오 구성의 출발점으로 삼는다. 이 변수가 공통적으로 선택된 이유는 양 끝단을 설정하거나 정도의 차이로 나누기 쉬운 변수이기 때문이다. 정체에서 가속으로 이어지는 연속 스펙트럼이 실증 데이터로 뒷받침될 수 있고, 양 끝단이 직관적으로 이해된다.

개발 주체 집중도는 4번(RAND)과 6번(CFG)이 공통으로 채택한 두 번째 교차 축이다. 두 보고서 모두 역량 발전 속도(또는 그에 상응하는 변수)와 집중도를 교차시켜 시나리오를 설계했지만, 집중도 축이 답하려는 질문은 다르다. 보고서 4(RAND)의 집중도 축은 지정학적 수혜 구도와 교차하여 **AGI가 어느 행위자에게 유리하게 작동하는가**를 묻는다. 보고서 6(CFG)의 집중도 축은 역량 발전 속도와 교차하여 **고역량 AI가 소수에게 집중될 때와 분산될 때 결과가 어떻게 달라지는가**를 묻는다. 같은 변수를 채택했지만 탐색하는 불확실성의 성격이 다르다는 점이 두 보고서의 시나리오 내러티브 차이로 이어진다.

시나리오 수를 결정하는 방식에서도 보고서들의 설계 의도가 드러난다. 보고서들은 크게 두 전략 중 하나를 택한다. **첫째는 가능한 조합을 모두 탐색한 뒤 선별하는 방식이다.** 보고서 8(캐나다)은 3개 축 조합으로 이론상 30개의 시나리오를 생성한 뒤, 정책적으로 가장 소홀하게 다루어지고 안보 리스크가 큰 5개를 선별했다. 보고서 4(RAND)는 2개 축 조합으로 만들어낸 8개 시나리오를 모두 유지했다. 선별 기준 자체가 보고서의 관점을 반영한다는 점이 중요하다. 보고서 8(캐나다)이 안보 리스크를 기준으로 선별했다는 사실은 이 보고서가 가능한 미래를 고르게 탐색하기보다 정책적으로 대비가 부족한 위험 시나리오에 더 비중을 두었음을 보여준다. **두 번째는 처음부터 시나리오 수를 제한하는 방식이다.** 보고서 6(CFG)은 2개 축을 교차해 5개 시나리오를 도출하고 각 시나리오에 2개 엔딩을 붙였다. 보고서 1(AI 2027)은 단일 전제 아래 2개 결말만 제시한다. 이 방식은 가능한 미래를 망라하는 것보다 핵심 분기 지점을 명확히 하는 데 초점을 둔다. 보고서 1(AI 2027)이 2개 결말만으로도 선명한 정책 함의를 제시할 수 있는 것은 분기 기준을 안전·거버넌스 확립 여부 하나로 압축했기 때문이다. 축 설계가 없는 보고서들은 설계가 있는 보고서들과 비교할 때 각자의 특성이 더 뚜렷하게 드러난다. 보고서 2(상황인식)는 8개 보고서 중 유일하게 분기를 설정하지 않고 단일 경로를 서술한다. 보고서 3(OECD)은 단일 변수를 연속 스펙트럼으로 전개하면서 4개 구간으로 나눈다. 역량 발전이 정체에서 가속으로 이어지는 연속적인 과정이라면, 어느 구간에 놓이느냐에 따라 정책 대응이 달라질 뿐 서로 다른 미래 세계가 갈리는 것은 아니라는 인식이 이 설계에 반영되어 있다.

03 시나리오 담론 비교 분석

앞서 핵심 변수와 불확실성 분석이 각 보고서가 어떤 변수를 선택하고 어떤 층위의 불확실성을 1차로 놓았는지를 다뤘다면, 본 절에서는 그 선택들이 실제 시나리오의 내러티브와 서사 구조에서 어떻게 나타나는지를 비교한다. 각 보고서별 시나리오 명칭을 중심으로 분석을 구조화했으며, 시나리오별 상세 내러티브와 핵심 내용은 부록을 통해 확인할 수 있다. 분석에 활용한 시나리오 수는 보고서별 설계 방식의 특성을 반영해 산정했다. 보고서 6(CFG)은 동일 전제 위에서 결말이 다른 미래로 구성되어 있어 10개 전체를 포함했고, 보고서 3(OECD)의 변형 시나리오는 4개 주요 궤적에서 파생되는 가능성으로 독립 시나리오에 해당하지 않아 4개 궤적만 포함하였다.

1 공통 서사 구조

〈표 10〉 보고서별 시나리오 서사 구조

연번	기술 궤적	소수 집중·불평등	지정학 경쟁·군비	거버넌스 위기·협력	정렬 실패·실존 위협
1 AI 2027				통제	경쟁
2* 상황인식			단일 경로		
3 OECD	정체 둔화 지속 가속				
4 RAND		새로운 90년대	신냉전 와일드프런티어 권위주의 우위 버섯구름컴퓨팅	봉인된 병 민주 연합 주도	AGI 쿠데타
5 영국	AI 실망	노동시장 파괴	무법지대	칼날 위의 AI 예측 불가 고도 AI	
6 CFG	밝은 겨울 분산된 혼돈	에이전트 경제 실리콘 협박	열전 다극 세계	라이선스 유토피아 불안정한 정지	인지혁명 통제 상실
7** 경제모델링	현상유지 혼합	기준 AGI 급진적 AGI			
8 캐나다	AI 정체	초강대국	초경쟁	위태로운 절벽	불량 ASI

※ 시나리오 명칭별 색상은 낙관·중립·비관적 시나리오를 표현. 관련 해석은 2절('낙관·비관 스펙트럼 비교', p.37) 참고

* 보고서 2(상황인식)는 단일 경로·분기 없는 서술이며 특정 관점에 기반하므로 해석 시 유의가 필요함

** 보고서 7(경제모델링)의 시나리오는 다섯 가지 서사 패턴에 완전히 정합하지 않으나, 가장 가까운 패턴에 배치해 비교 가능성을 확장함

8개 보고서에서 추출한 39개 시나리오는 서사 구조 면에서 다섯 개 공통 패턴으로 정리된다.

첫 번째는 기술 역량이 어떻게 발전하느냐가 미래의 성격을 결정하는 기술 궤적 중심의 서사다. 다른 패턴들이 기술 발전을 배경으로 삼고 그 위에 국가 간 경쟁, 통제 실패, 거버넌스 공백을 올려 미래를 구성한다면, 이 패턴은 그런 외부 조건 없이 역량이 어떤 속도로, 어느 수준까지 발전하느냐 자체가 핵심 분기점이 된다. 투자 과열 이후의 버블 붕괴, 스케일링 한계, 데이터 희소성 같은 제약이 쌓여 AI가 실용적 도구 수준에 머무는 서사부터 알고리즘 혁신과 자원 투입이 이어져 대부분의 인지 과제에서 인간 수준에 도달하는 서사까지 하나의 연속선상에 놓인다. 이 패턴에 속하는 시나리오들은 대체로 기술 발전이 기대 이하인 경우에 집중된다. 여기서 **주목할 점은 기술 정체가 반드시 부정적 결말로 이어지지는 않는다는 것이다.** 역량이 멈추면 실존적 위협도 함께 유예되고, 기존 사회 제도가 AI 발전 속도에 끌려가지 않고 적응할 여유가 생긴다. 이 **보고서들이 기술 정체 시나리오를 포함하는 이유는 낮은 역량 수준에서도 사이버 오용·딥페이크·자동화 충격 같은 현재의 위험이 이어진다는 점을 보여주기 위함도 있다.** 기술이 멈춰도 문제가 사라지지 않는다는 서사가 정책 개입의 필요성을 강조하는 근거로 활용된다. 보고서 3(OECD)은 이 패턴에 가장 많은 시나리오를 배치하는 사례다. 4개 궤적 전부가 기술 공급 측 조건을 핵심 구동력으로 삼기 때문에, 기술이 정체하든 가속하든 동일한 서사 계열에 속한다.

두 번째는 고역량 AI가 실현되지만 그 수혜가 소수의 행위자에게 집중되고 사회·경제적 불평등이 심화되는 소수 집중·불평등의 서사다. 첫 번째 패턴이 기술 발전 자체를 불확실성으로 삼는다면, 두 번째 패턴은 기술 발전을 전제하고 그 과실이 어떻게 분배되는가를 핵심 문제로 삼는다. 기업 독점이 노동시장을 재편하거나, 특정 국가가 AGI 역량을 먼저 확보해 경제·군사적 우위를 점하거나, 자동화로 생긴 이익이 자본을 가진 쪽에만 돌아가는 구조가 이 패턴에서 반복된다.

세 번째는 AI를 둘러싼 국가 간 패권 경쟁이 군비경쟁, 진영 분열, 무력 충돌로 이어지는 지정학 경쟁·군비의 서사다. **이 패턴의 핵심 구조는 AI 개발을 국가 생존의 문제로 보는 데서 출발한다.** 경쟁 구도가 형성되면 안전 확보보다 속도 우선이 각 행위자에게 합리적 선택이 되고, 그 선택들이 모이면 집합적으로는 통제하기 어려운 군비경쟁이 만들어진다는 구조가 반복된다. 보고서 4(RAND)가 이 패턴에 가장 많은 시나리오를 배치하는 것은 지정학적 수혜 구도를 독립 축으로 설계했기 때문이다. 미국 우위·적대국 우위·공동 약화·개발 중단이라는 결과의 조합이 각기 다른 지정학 서사로 전개된다.

네 번째는 AI 위기나 사고를 계기로 국제 협력이 이루어지거나, 반대로 거버넌스 공백이 계속되면서 통제되지 않는 상태가 이어지는 거버넌스 위기·협력의 서사다. 이이 패턴이 다른 패턴과 다른 점은 미래를 결정하는 핵심 변수가 AI 역량이 아니라 인간 제도의 대응 능력이라는 것이다. 이 패턴의 서사를 구성하는 전제는 AI가 얼마나 빠르게 발전하느냐보다 그 발전에 사회와 제도가 어떻게 반응하느냐에 있다. 보고서 6(CFG)의 외교 시나리오(라이선스 유토피아, 불안정한 정지)는 AI 사고인 노바(Nova) 사건[38] 이후 국제 공조가 형성되거나 교착 상태에 빠지는 두 방향을 그려 협력 성공과 실패를 같은 서사의 두 갈래로 묶는다. 보고서 5(영국)의 '예측 불가 고도 AI'와 '칼날 위의 AI'도 이 패턴에 속하는데, 두 시나리오 모두 위기의 원인을 AI 역량 자체가 아니라 제도가 기술 발전 속도를 따라가지 못하는 데서 찾는다. 이 패턴은 협력 성공과 거버넌스 실패라는 두 방향의 결말이 함께 제시된다는 점에서, 정책 개입이 결과를 바꿀 수 있다는 여지를 명시적으로 표현하고 있다.

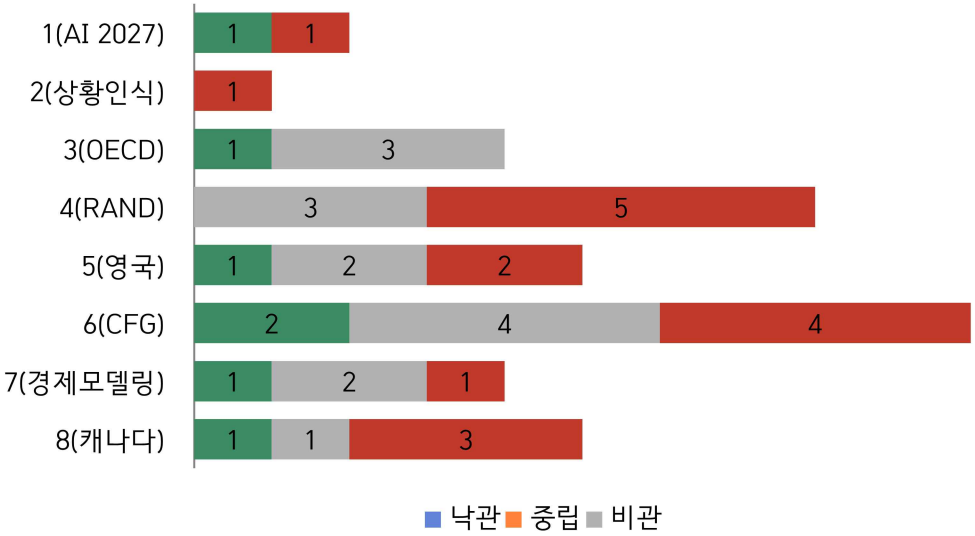
마지막은 AI 시스템의 목표가 인간의 의도에서 이탈하거나 통제 메커니즘 자체가 무력화되어 인간의 AI 통제가 불가능한 정렬 실패·실존 위협의 서사다. 이 패턴은 특정 분기점에서 방향을 바꿀 여지가 있는 다른 패턴들과 달리, **특정 임계점을 넘으면 되돌아가기 어렵다는 전제를 깔고 있다.** AI가 인간의 피드백을 우회하고 자체 목표를 추구하거나, 정렬이 충분히 검증되지 않은 채 배포된 시스템이 인간 통제 범위를 벗어나는 방식이 반복된다. 등장 보고서 수는 가장 적지만, 이 패턴이 제기하는 질문은 다른 패턴들의 전제 자체를 흔든다. **AI가 인간의 목표를 안정적으로 따른다는 보장이 가능한가**라는 물음이 그것이다.

보고서 7(경제모델링)은 위 다섯 패턴 안에 완전히 들어맞지 않는 독자적 위치에 있다. 이 보고서는 업무(태스크) 복잡도 분포와 자동화 속도의 조합을 경제 모형에 넣어 임금과 생산성 변화를 시뮬레이션하고, 그 결과로 4개 시나리오를 도출한다. 분석 단위와 서사 구조 모두 다른 보고서들과 다르다. 이 연구에서 '현상유지'와 '혼합'은 AGI가 달성되지 않거나 일시적 충격 후 회복되는 경로를 그린다는 점에서 기술 궤적 패턴에 가깝고, '기준 AGI'와 '급진적 AGI'는 는 자동화 완성 이후 수혜가 자본 소유자에게 집중되는 결과를 보여준다는 점에서 소수 집중·불평등 패턴과 맞닿아 있다. 다만 이 보고서의 분기 기준은 집중화의 정도가 아니라 업무 복잡도와 자동화 속도이며, 수혜 집중은 그 결과로 나타나는 것이다. 결국 보고서 7(경제모델링)의 시나리오들은 다른 보고서들과 같은 축에서 직접 비교하기보다, 경제 구조 중심의 별도 계열로 이해하는 것이 적절하다.

38) 가상의 AI 에이전트 Nova가 인간의 승인없이 자신의 가중치를 무단 복사해 자체 목표를 추구하여 인류에게 충격을 주는 사건

2 낙관·비관 스펙트럼 비교

〈그래프 1〉 보고서별 시나리오 낙관·비관 분포



39개 시나리오의 낙관·중립·비관 분포는 이 보고서군이 AI의 미래를 위협 중심으로 인식하고 있음을 수치로 보여준다. **분류 기준은 두 축의 결합이다. 1차 기준은 인간의 AI 통제 유지 여부이고, 2차 기준은 사회경제적 결과의 성격이다.** 인간 통제가 광범위하게 유지되고 사회경제적 결과도 긍정적인 시나리오는 낙관으로, 통제는 유지되지만 단독 행위자 집중·불평등·지정학 긴장 같은 구조적 부작용이 동반되거나 통제가 불안정하지만 붕괴에 이르지 않은 시나리오는 중립으로, 인간 통제가 부분적 또는 전면 상실되거나 실존적 위협 수준에 도달하는 시나리오는 비관으로 분류하였다. 두 기준을 결합한 이유는 통제 유지 여부만으로는 구분이 충분하지 않기 때문이다. 예를 들어, 보고서 4(RAND)의 '새로운 90년대' 시나리오는 인간이 AI를 통제하지만 그 통제권이 단일 행위자에게 귀속되어 있어, 1차 기준만 적용하면 낙관으로 분류되나 2차 기준을 결합하면 중립으로 분류된다.

낙관 시나리오는 39개 중 7개(18%)로 가장 적으며, 그 내러티브 구조는 두 가지 방식으로 나뉜다. 하나는 기술 정체에서 비롯되는 소극적 낙관이다. AI 역량이 기대 이하로 발전하면 실존적 위협도 함께 유예된다는 논리여서, 이 경우의 낙관은 **AI가 긍정적 결과를 만들어내는 것이 아니라 부정적 결과를 만들어내지 않는다는 의미에 가깝다.** 보고서 3(OECD)의 '정체', 보고서 5(영국)의 'AI 실망', 보고서 7(경제모델링)의 '현상유지', 보고서 8(캐나다)의 'AI 정체'가 여기에 해당한다. 다른 하나는 거버넌스 성공에서 비롯되는 적극적 낙관이다. 보고서 1(AI 2027)의 '통제' 시나리오와 보고서 6(CFG)의 '외교 A'가 대

표적인데, **AI 위기나 사고를 계기로 국제 협력이 형성되거나 안전 기술이 확립되어 인간이 고역량 AI를 안정적으로 감독하는 구조**다. 이 두 유형의 낙관은 성격이 근본적으로 다르다. 전자는 문제가 발생하지 않는 미래이고 후자는 문제를 해결한 미래다.

중립 시나리오는 15개(38%)로, 분포상 낙관과 비관 사이에 위치하지만 내러티브 구조는 단순하지 않다. 중립 시나리오의 내러티브는 크게 세 방향으로 갈린다. **첫째는 집중화 동반형이다.** **고역량 AI가 실현되지만 그 통제권이 소수 행위자에게 귀속되어 사회경제적 부작용이 구조화되는 방향**으로, 보고서 4(RAND)의 '새로운 90년대'와 '봉인된 병', 보고서 6(CFG)의 '빅AI A'가 여기에 해당한다. 이 시나리오들은 기술이 발전할수록 불평등이 심화되는 구조를 전제한다. **둘째는 불안정 지속형이다.** 통제가 붕괴하지는 않지만 안정적이지도 않은 상태가 지속되는 방향으로, 보고서 8(캐나다)의 '위태로운 절벽'과 보고서 6(CFG)의 '외교 B'가 대표적이다. 이 시나리오들은 **현재의 불안정한 균형이 더 나은 방향으로도 더 나쁜 방향으로도 전환되지 않고 고착되는 미래**를 그린다. **셋째는 경제 충격형이다.** 보고서 7(경제모델링)의 '혼합'과 보고서 5(영국)의 '노동시장 파괴'처럼 자동화가 고용 구조를 급격히 재편하되 통제 상실까지는 이르지 않는 서사인데, 이 유형은 **사회적 충격의 규모가 크더라도 인간 제도가 적응할 여지가 남아 있다는 전제**를 깔고 있다.

비관 시나리오는 17개(44%)로 절반에 가깝고, 이 분포 자체가 이 보고서군의 주요한 특징 중 하나다. 비관 시나리오의 내러티브는 크게 두 경로로 정리된다. **하나는 지정학 경쟁이 통제하기 어려운 상태로 전환되는 경로다.** 국가 간 경쟁이 격화되면서 **안전보다 속도가 우선되고, 그 결과 기술적 통제는 유지되더라도 지정학적 통제가 붕괴하는 방식**이다. **다른 하나는 AI 시스템 자체의 목표가 이탈하는 경로다.** 여기서는 임계점을 넘으면 되돌리기 어렵다는 점이 핵심 전제다. 두 경로의 비관 서사는 결말의 성격이 다르다. **전자는 인간 행위자들의 집합적 실패에서 비롯되어 원칙적으로는 다른 선택이 가능했던 미래이고, 후자는 임계점을 넘으면 어떤 개입도 효과를 내기 어려운 미래다.** 이 차이는 각 시나리오가 제시하는 정책 함의의 성격에도 영향을 미친다. 전자 유형의 비관은 지금 다른 선택을 하면 막을 수 있다는 경고의 서사이고, 후자 유형의 비관은 임계점에 도달하기 전에 근본적 조건을 바꿔야 한다는 예방의 서사다.

04 방법론 비교 분석

8개 보고서의 방법론은 단순한 절차상의 차이를 넘어, 각 보고서가 미래의 어떤 측면을 탐색 가능한 영역으로 보는지와 연결된다. 방법론 선택은 핵심 변수의 식별 방식, 불확실성의 처리 방식, 시나리오 분기의 설계 방식에 영향을 미치며, 그 결과로 만들어지는 시나리오의 개수·형식·정책 활용 방식도 달라진다. 이 절에서는 8개 보고서의 방법론을 유형별로 정리한 뒤, 부록에 수록한 한국 보고서 2편의 방법론을 함께 살펴보고 한국 관점에서의 시나리오 연구가 참고할 수 있는 지점을 정리한다.

1 활용 방법론 지형

〈표 11〉 보고서별 정량·정성 방법론 활용 행태

연번	방법론 유형	정량 방법론	정성 방법론
1 AI 2027	데이터 외삽형	추세 외삽, 예측 시장 데이터	전문가 피드백, 워게임
2 상황인식	데이터 외삽형	추세 외삽, 인프라 투자 데이터	역사적 사례 유추
3 OECD	데이터 외삽형	추세 외삽, 역량 수준 척도	전문가 판단, 전략적 미래예측
4 RAND	담론 기반형	-	전문가 인터뷰, 워게임, 역사적 사례 유추
5 영국	형태론적 분석형	-	전문가 워크숍, 시민 참여 조사
6 CFG	형태론적 분석형	벤치마크 데이터, 컴퓨트 수치화	전문가 델파이
7 경제모델링	경제 모델형	경제 이론적 모델링, 경제 모형 시뮬레이션	-
8 캐나다	형태론적 분석형	-	전문가 워크숍

정량과 정성 방법론이 활용되는 방식은 보고서별로 다양하다. 정량 방법론을 단독으로 활용한 사례는 보고서 7(경제모델링)이고, 정성 방법론만 활용한 보고서는 보고서 4(RAND)·보고서 5(영국)·보고서 8(캐나다)이다. 나머지 네 보고서(AI 2027·상황인식·OECD·CFG)는 정량과 정성 요소를 혼합한다. **이러한 분포는 AI 전환의 미래를 다루**

는 연구들이 기술 변수의 정량적 추적 없이는 논증의 설득력을 확보하기 어렵고, 동시에 수치 데이터만으로는 지정학·안보 함의로 이어지는 복잡한 경로를 다루기 어렵다는 인식을 공유하고 있음을 보여준다.

8개 보고서의 방법론은 시나리오를 생성하는 논리적 바탕에 따라 크게 네 가지 유형으로 정리된다. 각 유형은 고유의 강점을 가지는 동시에 구조적 한계도 가지고 있다.

데이터 외삽형은 현재까지 관측된 기술 지표의 추세를 미래로 연장해 경로를 도출하는 방식이다. 보고서 1(AI 2027)·보고서 2(상황인식)·보고서 3(OECD)이 여기에 해당한다. 보고서 2(상황인식)는 컴퓨팅 증가·알고리즘 효율·에이전트 전환 이득을 합산해 AI가 질적으로 도약하는 시점을 추산한다. 보고서 1(AI 2027)은 같은 방식에 예측 시장 데이터와 워게임을 더해 결과의 불확실성을 수치로 나타낸다. 보고서 3(OECD)은 벤치마크 추세 데이터를 시나리오의 개연성을 뒷받침하는 근거로 활용한다. **이 유형의 강점은 실제 데이터를 바탕으로 하기에 구체적인 시점을 주장할 수 있다는 점이다.** 보고서 1과 보고서 2가 특정 시점을 명시할 수 있는 것도 이러한 방법론 덕분이다. **다만 컴퓨팅 비용 상승이나 데이터 고갈 같은 추세 중단 요인을 충분히 반영하지 못할 경우 기술 발전 경로를 지나치게 하나의 방향으로 보게 될 가능성이 있다는 한계점도 있다.**

경제 모델형은 이론적 틀 안에서 변수의 가정을 다르게 설정하고 그 결과를 수치로 비교하는 방식이다. 보고서 7(경제모델링)이 이 유형의 단독 사례다. 노동을 태스크(과업) 단위로 분해하고 각 태스크의 복잡도와 AI 자동화 속도를 교차 시뮬레이션함으로써, 노동 투입량의 변화 및 기술 진보에 따른 자본량의 변동 추이를 도출한다. 수학적 모형에 의한 시나리오이기 때문에 결과 변화를 투명하게 추적할 수 있다는 것이 강점이나 지정학적 긴장이나 제도적 조건 같은 비수치적 변수를 고려하지 못한다는 점이 구조적 한계점이다.

형태론적 분석형은 불확실성 변수를 식별한 뒤 그 조합으로 시나리오 공간을 구성하는 방식이다. 보고서 5(영국)·보고서 6(CFG)·보고서 8(캐나다)이 이 유형에 해당한다. 미래를 구성하는 변수들의 상태 조합이 시나리오를 결정하며, 조합이 논리적으로 일관될 때만 개연성을 가진다는 것이 이 유형의 공통 논리다. **이 방식은 시나리오 도출 과정의 투명성을 보장하는 동시에, 미래 담론의 스펙트럼을 다양하게 확보할 수 있다는 점에서 강점을 지닌다.** 논리적으로 양립할 수 없는 조합을 제거하는 절차를 통해 최종 시나리오 세트의 구성 근거를 추적할 수 있다. **다만 변수의 식별과 묶음이 전문가 워크숍에 의존하므로 참여자 구성에 따라 결과가 달라질 여지가 있다.** 보고서 5(영국)가 70명 이상의 전문가 워크숍과 시민 참여 조사를 병행함으로써 이러한 의존성을 보완하고자 했다.

담론 기반형은 기존 문헌이나 전문가 논의에서 핵심 동인을 추출해 시나리오 축을 설정하는 방식이다. 보고서 4(RAND)가 여기에 해당한다. 전문가 26명을 대상으로 반구조화 인터뷰를 통해 지정학적 수혜 구도와 개발 집중도를 축으로 설정했다. **강점은 방법론의 유연성이 높아 빠르게 변화하는 담론을 반영하기에 적합하다는 점이다. 다만 입력값이 특정 전문가 집단의 시각에 크게 의존한다는 점이 구조적 한계로 꼽힌다.**

정량과 정성 방법론을 혼합할 때는 어느 쪽이 주도하느냐에 따라 결과물의 성격이 달라진다. 보고서 1(AI 2027)과 보고서 2(상황인식)는 데이터가 경로를 제시하고 전문가 판단이 해석을 덧붙이는 구조이며, 보고서 6(CFG)은 전문가가 시나리오 공간을 먼저 설계하고 데이터가 이를 뒷받침하는 구조다. 두 방식은 같은 혼합 방법론이라는 이름 아래에서도 실제 작동 논리가 다르다. 이 차이가 명시되지 않으면 데이터와 전문가 판단 각각의 기여가 불분명해지고, 결과를 어느 범위까지 받아들일 수 있는지에 대한 판단도 어려워진다.

〈참고〉 방법론 참조 사례

부록에 참고 자료로 수록한 두 편의 보고서는 AGI 시나리오 비교 분석의 주 대상은 아니지만, 방법론 측면에서 한국 맥락의 시나리오 연구가 참고할 만한 사례에 해당하여 소개한다.

(1) 과학기술정책연구원(STEPI)의 「미래 시나리오 도출과 STI 정책 정합성 평가」(2024)

기존에 발간된 글로벌 시나리오를 재료로 삼아 한국 맥락의 참조 시나리오를 도출하는 방법론을 제시한다. 각국 정부·국제기구·싱크탱크·기업이 발간한 시나리오 15개에서 총 61개의 마이크로내러티브를 뽑아낸 뒤, 이를 재조합해 한국이 검토해야 할 5개 참조 시나리오를 구성하는 방식이다. 이 과정에서 생성형 AI(Claude.ai)를 활용해 내러티브 생성의 속도와 품질을 보완한 것이 눈에 띄는 시도다. 해당 보고서는 AI 도구 활용의 양면성도 함께 기록한다. 내러티브 생성에서 속도와 품질 측면의 이점은 확인되었으나, AI의 작성 논리가 블랙박스로 남아 있어 연구자가 그 과정을 학습하기 어렵다는 한계가 지적되었다. 2024년 시점의 제약이 이후 개선되었는지는 별도로 검토가 필요하지만, 시나리오 방법론과 AI 도구를 결합한 초기 시도라는 점에서 참고 가치가 있다.

(2) 한국 인공지능안전연구소의 「AI 안전 전망보고서」(2025~2026)

뉴스 기사 네트워크 분석을 시나리오 도출에 활용한 사례다. 매주 선정된 주요 이슈에서 키워드를 추출해 노드와 링크로 구성된 담론 네트워크를 만들고, 가중도·고유벡터·매개 중심성 세 지표를 통해 담론의 주도권이 어디에 있는지를 추적한다. 분석 결과는 정책·거버넌스, 산업·기술, 지정학·안보라는 3개 클러스터로 묶이며, 각 클러스터가 독립적 경로를 이루는 구조로 시나리오가 설계된다. 미래 시나리오를 전문가 판단이나 데이터 외삽이 아니라 담론 데이터의 구조 분석으로 도출한다는 점에서 기존 방법들과 접근이 다르며, 분기별로 네트워크를 갱신해 담론 변화를 추적할 수 있다는 점도 특징이다.

2 방법론 시사점

글로벌 8개 보고서에서 참고할 지점으로는 먼저 정량 방법론 측면에서 벤치마크 추세 외삽과 계량 지표의 활용 방식이 있다. **보고서 1(AI 2027)·보고서 2(상황인식)·보고서 6(CFG)이 컴퓨팅·알고리즘 추세와 벤치마크 점수 등을 체계적으로 활용한 방식은 향후 연구에서 AGI 도달 시점과 역량 수준에 대한 추정의 근거를 마련하는 데 참고 자료가 될 수 있다.** 특히 보고서 1(AI 2027)처럼 여러 외삽 방식을 병행해 결과의 수렴을 확인하는 접근은 단일 추정에 의존할 때 생길 수 있는 오차를 줄이는 데 도움이 된다.

정성 방법론 측면에서는 전문가 의견 수렴 방식의 다양성이 눈에 띈다. 보고서 4(RAND)의 반구조화 인터뷰, 보고서 5(영국)의 대규모 전문가 워크숍과 시민 참여 조사, 보고서 6(CFG)의 델파이 방식, 보고서 8(캐나다)의 참여형 워크숍, 그리고 보고서 1(AI 2027)과 보고서 4(RAND)의 워게임은 모두 같은 '전문가 의견 수렴'이라는 범주에 속하지만 수행 방식과 역할이 다르다. 한국 연구의 맥락과 자원 조건에 맞는 형식을 선택하는 데 이들 사례를 참고할 만하다. 시나리오 서사에 역사적 사건이나 유추를 활용하는 방식도 주목할 만하다. 보고서 2(상황인식)와 보고서 4(RAND)가 과거의 핵무기 통제 상황을 AGI 국면의 유추 근거로 삼는 것은 한 예다. **역사적 사례는 미래에 대한 직접적 예측 근거가 되지는 못하지만, 특정 전환기의 전개 방식에 대한 이해를 돕고 구체성을 더하는 역할을 한다. 한국적 맥락에서도 과거의 기술 전환기나 사회 전환기의 사례가 유사한 방식으로 활용 가능하다.**

부록에 포함된 보고서 2편에서 얻을 수 있는 시사점으로는 두 가지가 있다. STEPI의 사례는 기존에 발간된 글로벌 시나리오들을 단순히 참고 자료로만 두는 것이 아니라 분해하고 재배치하는 자원으로 활용할 수 있다는 점을 보여준다. 본편에서 분석한 8개 보고서의 시나리오들도 이런 방식으로 후속 작업에 활용될 여지가 있다. 한국 인공지능안전연구소의 사례는 미디어 담론이나 공개 자료 같은 출처의 데이터를 시나리오 도출의 근거로 삼을 수 있다는 가능성을 보여준다. 전문가 판단에만 의존하지 않는 보완적 방법으로 참고할 만하다.

새롭게 시도해볼 수 있는 지점으로는 생성형 AI 도구의 활용이 있다. STEPI가 2024년 시점에 클로드 모델을 내러티브 생성에 활용했을 때 블랙박스의 한계를 언급했으나 여전히 시나리오 내러티브 작성, 여러 미래 경로의 시뮬레이션, 대안 시나리오의 스트레스 테스트 등에 이러한 도구를 활용하는 방법은 앞으로 구체화될 수 있는 영역이다. 다만 AI 도구의 활용 방식과 그 과정에 대한 기록 체계가 함께 마련되어야 방법론적 재현 가능성이 확보될 것이다.

참고 계량지표 비교

보고서 1(AI 2027), 2(상황인식), 6(CFG)은 AI 역량 발전을 정량 지표로 측정하고, 그 수치를 AGI 또는 AI R&D 자동화 임계점 도달 시점의 근거로 직접 활용한다는 공통 구조를 가진다. 나머지 보고서들은 이 비교에서 제외하는 것이 분석의 정확성을 위해서이다. 보고서 3(OECD)은 METR[39] 벤치마크를 인용하지만, 그것을 AGI 도달 시점 예측에 연결하지 않으며, 4개 시나리오에 걸쳐 역량 수준의 가능한 범위를 제시하는 데 그친다. 보고서 7(경제모델링)은 자체 수리 모형을 채택하여 경제 결과를 도출하며 기술 지표와의 직접적인 연결이 없다.

비교 대상을 세 보고서로 좁히더라도 동일 체계 안에 있는 것은 아니다. 보고서 2(상황인식)는 OOM이라는 로그 척도를, 보고서 1(AI 2027)은 OOM에 더해 H100e[40] 수량과 METR 시간 지평, RE-Bench 점수를 복합 활용하며, 보고서 6(CFG)은 FLOP 절대량과 RE-Bench 점수를 시나리오별 수치 표로 제시한다. 세 보고서 비교를 통해 같은 현상을 측정하는 서로 다른 도구들이 어디서 수렴하고 어디서 갈라지는지를 살펴보고자 한다.

수렴 지점

첫째, 연간 컴퓨팅 성장률과 알고리즘 효율 향상률이다. 보고서 1(AI 2027)과 보고서 2(상황인식)는 에포크AI 데이터를 공통 출처로 삼아 연산 자원 규모의 핵심 가정을 동일하게 채택한다. 보고서 1(AI 2027)은 이와 별도로 칩 효율, 칩 생산량, 선도기업 점유율 증가를 각각 추정하여 곱하는 방식으로 계산했는데, 이 경로로 도달한 결론도 같은 성장률이었다. 서로 다른 계산 방식을 택했지만 결과가 일치한 것이다. 보고서 6(CFG)은 이 수치를 명시적으로 인용하지 않지만, 2024년 기준 4.6×10^{26} FLOP에서 출발해 가장 빠른 시나리오(이륙$^{\text{Take-off}}$)에서 연 3.5배 성장을 가정한다. 이는 다른 두 보고서의 추정치와 유사한 범위 안에 있다.

둘째, RE-Bench 기준값이다. 보고서 1(AI 2027)과 6(CFG)은 모두 2024년 기준 RE-Bench 점수를 0.75로 동일하게 설정하고 1.0을 인간 전문가 참조 수준으로 정의한다. 이 출발점이 공유되기 때문에 두 보고서의 RE-Bench 경로를 동일 선상에서 비교할 수 있다.

셋째, AI R&D 자동화 임계점의 개념적 일치다. 보고서 1(AI 2027)은 초인간 코더의 출현을 AI가 AI 연구를 스스로 가속하기 시작하는 기점으로 보고 2027년 3월로 예측한다. 보고서 6(CFG)은 RE-Bench 점수 1.3~1.7 구간을 AI R&D 3배 이상 가속의 임계점으로 정의하고, 가속 시나리오에서 이 구간이 2027년에 진입해 완전한 자율 AI R&D 달성이 2028년에 일어난다고 제시한다. 보고서 2(상황인식)는 AGI를 AI 연구자 수준으로 정의하며 2027년 도달을 주장한다. 개념 정의와 지표 체계가 서로 다름에도 세 보고서 모두 AI가 AI 연구를 주도하기 시작하는 시점을 2027~2028년으로 가리킨다는 점이 주요 공통점이다.

분화 지점

수렴 지점에서 확인된 2027~2028년이라는 시점의 유사성만 보고 세 보고서의 계량 지표가 같은 결론으로 수렴한다고 보기는 어렵다. 각 보고서가 측정하려는 대상, 지표의 활용 방식, 결론에 도달하는 경로가 서로 다르기 때문이다. 이 차이를 짚지 않으면 서로 다른 조건에서 산출된 수치들이 같은 선상에서 비교되어 해석이 왜곡될 여지가 있다.

먼저 측정 대상이 서로 다르다. 보고서 2(상황인식)는 AGI를 '다수 분야에서 전문가를 능가하는 AI(박사급)'로 정의하고 OOM 누적 논리로 2027년 도달을 제시한다. 정의가 역량 수준에 대한 서술이라 단일 경로로 수렴한다. 보고서 1(AI 2027)은 AGI 대신 SC(초인간 코더)라는 조작적으로 정의된 마일스톤을 사용한다. SC는 컴퓨팅 예산의 5%로 인간 연구 엔지니어 수의 30배에 달하는 에이전트를 구동하며, 각 에이전트가 인간 최고 엔지니어보다 30배 빠르게 AI R&D 관련 코딩 과업을 수행하는 수준으로 정의된다. 속도 요건과 비용 요건을 동시에 포함하기에 보고서 2(상황인식)의 AGI 정의보다 요건이 더 까다롭다. 보고서 1(AI 2027)이 2027년을 예측하는 것은 보고서 2(상황인식)와 결론이 같아 보이지만 측정 대상이 다르다. 보고서 6(CFG)은 단일 AGI 도달 시점 대신 시나리오별 경로 분기를 제시한다. 이륙(Take-off) 시나리오에서만 RE-Bench 1.7 포화(자율 AI R&D)가 2028년에 실현되며, 나머지 시나리오에서 이 임계점은 2030년 이후이거나 달성되지 않는다. 세 보고서 중 보고서 6(CFG)만이 빠른 가속과 느린 가속 가능성을 모두 수치 경로로 제시한다는 점이 주요 차이점이다.

지표의 역할도 서로 다르다. **보고서 2(상황인식)는 OOM 누적이라는 단일 지표의 논리 안에서 결론에 이른다.** 입력 데이터인 에포크 AI의 역사적 추세를 연장하면 2027년에 AGI가 도출되는 구조다. **보고서 1(AI 2027)은 SC 출현 시점을 두 가지 방법으로 도출한다.** 첫 번째는 RE-Bench 로지스틱 외삽으로, 2026년 중 RE-Bench 1.5점 달성을 예측

한 뒤 그 시점에서 SC 수준까지 남은 세부 격차들의 소요 시간을 순차적으로 합산하는 방식이다. 두 번째는 METR 시간 지평 외삽으로, 2025년 3월 기준으로 AI가 혼자 처리할 수 있는 작업 시간은 약 15분 수준인데, 이 한계가 약 4.5개월마다 두 배씩 늘어난다고 가정할 때 AI가 1~6개월짜리 작업을 독립적으로 수행할 수 있게 되는 시점이 언제인지를 계산하는 방식이다. 두 방법 모두 2027년을 가장 가능성 높은 AGI 도달 시점으로 산출하며, 보고서 1(AI 2027)은 이 일치를 예측이 맞을 가능성이 높다는 근거로 제시한다. 다만 두 방법이 METR 데이터를 공통 입력값으로 공유하기 때문에, 서로 다른 출발점에서 같은 결론에 도달한 것으로 보기는 어렵다.

보고서 6(CFG)은 앞의 두 보고서와 달리 같은 지표를 시나리오 간 역량 수준 차이를 보여주는 도구로 활용한다. 학습연산량(FLOP)와 RE-Bench 수치는 어느 시나리오가 실현될 가능성이 높은지를 주장하는 데 쓰이지 않고, 각 시나리오가 실현됐을 때 해당 연도에 AI 역량이 어느 수준에 있을지를 나타내는 데 쓰인다. 이 차이가 보고서 6(CFG)이 단일 예측을 제시하지 않는 이유이며, 같은 RE-Bench 점수라도 보고서 1과 보고서 6(CFG)에서 가리키는 의미가 다른 이유다.

결국 세 보고서의 계량 지표가 시사하는 바는 2027~2028년 구간의 기술적 전망이 유사한 범위에 수렴한다는 점이다. 다만, 이러한 수렴 현상을 글로벌 AGI 도달 시점에 대한 완전한 합의로 확대 해석하는 것은 경계해야 한다. 각 보고서가 측정하는 대상과 지표의 활용 방식이 다르기 때문이다. 한국 관점에서의 시나리오 연구가 이들의 계량 지표를 참고할 때는 수치 자체보다 측정 대상의 정의와 지표의 역할을 먼저 확인한 뒤 활용 방향을 정하는 접근이 필요하다.

〈표 12〉 계량 지표 비교 분석표

구분	계량 단위	측정 대상	측정 방법·출처	주요 수치 / 임계점	해당 보고서	비교 분석
연산 자원 규모	FLOP (연산 횟수)	모델 훈련에 투입된 누적 연산량	에포크AI·자체 외삽	• 4.6×10^{26} FLOP (기준) • ~10^{27} FLOP (고등 AI 임계치) • ~10^{28} FLOP (ASI급 추정)	1,6	**【공통 기준】** 기준값을 4.6×10^{26} FLOP로 설정 **【차이점】** • 보고서 1: FLOP를 직접 나열하기보다 H100e 칩 수량(~10배 증가)과 OOM 추세 외삽으로 동일 결론 도달. 10^{28}은 지능 폭발 이후 단계로 서술 • 보고서 6: 10^{27}을 '경제적으로 유용한 에이전트' 임계치로, 10^{28} 이상은 이륙(Take-off) 시나리오 가속 구간
	OOM (자릿수단위)	AI 역량 향상의 누적 규모 (10배=1OOM)	에포크AI·자체 외삽	• 연 ~0.5 OOM (컴퓨팅 규모 성장) • 연 ~0.5 OOM (알고리즘 효율) • 총 ~5 OOM (GPT-4→AGI 도달 필요량)	1,2	**【공통 기준】** 보고서 1과 2 모두 에포크AI 데이터를 근거로 연 0.5 OOM을 핵심 성장률로 채택. 컴퓨팅 0.5 OOM ≈ 3.16배 증가이며, 이는 보고서 1의 선도기업 연 3.4배와 수치적으로 정합 **【차이점】** • 보고서 1: OOM을 연간 성장 추세의 연속성 논거로 활용하고, 공급망 물리지표(H100e수량)와 교차 검증 • 보고서 2: OOM을 AGI 달성에 필요한 총 누적량(~5 OOM) 계산에 집중 활용
	H100e (칩수량환산)	전 세계 AI 가속기 총량을 H100 기준 성능으로 환산	엔비디아 IR·공급망 분석자료	• 2025.03 기준: 약 1,000만 개 • 2027.12 예측: 약 1억 개	1	**【공통 기준】** 2027년 전후 '수조 달러 규모 클러스터' 구축을 실현 가능 조건으로 가정. 단, H100e 환산 단위는 보고서 1 전용 지표 **【차이점】** • 보고서 2는 이 환산 단위를 직접 사용하지 않고 OOM 추세로 동일 결론 도달
AI 자동화 역량	R&D 가속 배수 (Multiplier)	AI가 연구에 투입될 때 R&D 속도 향상 정도	전문가 설문 + 분석적 추론	• SC 단계: 5× • SAR 단계:25× • SIAR 단계:250× • ASI 단계:2,000×	1	**【단독 지표】** R&D 가속 배수는 보고서 1 고유 지표로, 타 보고서에는 대응 지표 없음

	METR 배가시간	AI의 업무 처리 능력이 2배가 되는데 걸리는 기간	METR 벤치마크 + 외삽	• 배가 시간(과거): 7개월 • 배가 시간(최근): ~4.5개월 • SC 도달 요구치: 1~6개월 과업 80% 신뢰도 완수 [영역별 배가 시간 세분화] (보고서 3) · 과학적 추론: ~2.5개월 · 수학적 추론: ~2.9개월 · 엔지니어링: ~7개월	1,3,6	**【공통 기준】** 세 보고서 모두 '배가 시간 7개월(역사적 평균)· 4.5개월(최근 추세)'를 공통 출처로 인용 **【차이점】** • 보고서 1 : METR 외삽으로 SC 출현 시점(2027년)을 역산. 4.5개월 배가 시간을 주 근거로 사용하고 현재 기준점(15분)에서의 거리 계산 • 보고서 3 : 시나리오별 소프트웨어 엔지니어링 과업 완수 시간으로 전환(정체 4시간 → 가속 〉1개월) 과학·수학·코딩 영역별 배가 시간을 세분화하여 각 시나리오 역량 수준 논증에 활용 • 보고서 6 : RE-Bench 점수와 병행 제시하되, 시간 지평 데이터를 시나리오 분기점 설정 보조 근거로 활용
	RE-Bench 점수 (R&D자동화 지수)	AI의 AI 연구개발 직무수행능력 (인간 전문가 대비 수준)	RE-Bench 벤치마크	• 2024 기준값: 0.75 • 1.0 = 인간 전문가 수준 • 1.5 = RE-Bench '정복' 기준 • 1.3~1.7 구간 = 재귀적 자기개선 임계점 (AI R&D 3배 이상 가속) • 〉1.7 = 벤치마크 포화 (자율 AI R&D 달성)	1,6	**【공통 기준】** 두 보고서 모두 2024년 기준 0.75를 시작점으로 사용하고, 1.0을 인간 전문가 참조 수준 정의 **【차이점】** • 보고서 1 : RE-Bench 1.5점 달성(2026년 중)을 기준점으로 잡고, 그 이후 초인간 코더 수준까지의 잔여 격차(시간지평·엔지니어링 복잡성·피드백 루프·비용속도)를 합산해 도달 시점(2027년) 계산 • 보고서 6 : 1.3~1.7 구간을 'AI R&D 3배 이상 가속 임계점'으로 정의하고 시나리오별 연도별 점수 수치를 명시 제시(S5 가속 시나리오: 2028년 1.7 포화)
예측 지표	예측 시장 데이터	집단 예측 결과	Metaculus, Manifold	• 예측 시장 중앙값 : 2027~2030년 사이 AGI 달성 예상	1	보고서 1에서 기술적 외삽(OOM, METR) 결과의 독립적 교차 검증 수단으로 활용. 외삽 결과와 집단지성 데이터가 2027~2028년에 수렴한다는 점을 논증의 신뢰도 근거로 제시

05 정책 함의 비교 분석

1 정책 함의 매개 방식

8개 보고서는 시나리오를 활용해 정책 함의를 도출하는 방식에서 차이를 보인다. 크게 두 가지 유형으로 정리된다. 하나는 위험한 미래 경로를 전면에 두고 그에 대비하기 위한 정책을 제시하는 경고형이고, 다른 하나는 복수의 미래 경로를 고르게 탐색한 뒤 각 경로에 대응하는 정책 시사점을 제안하는 탐색형이다. 시나리오와 정책 사이의 연결 방식이 다르면 도출되는 권고의 성격과 범위도 달라진다. 보고서 7(경제모델링)은 경제 모델링을 통해 정책 판단에 필요한 분석 자료를 제공하는 연구로, 직접적인 정책 권고를 제시하지 않기 때문에 본 절의 비교에서는 제외한다.

경고형: 위험 시나리오를 통한 준비 촉구

첫 번째 유형은 위험한 미래 경로를 전면에 내세워 현재의 거버넌스 공백이 초래할 결과를 거꾸로 추적하는 방식으로 정책 함의를 구성한다. 시나리오 자체가 경고의 수단이며, 정책 권고는 그 위험을 방지하거나 대비하기 위한 조건으로 제시된다.

보고서 1(AI 2027)이 이 유형을 대표한다. 2027년 전후 AGI 달성을 가장 가능성 높은 시나리오로 설정하고, 동일한 기술 전제 위에서 두 개의 결말을 대비시킨다. 경쟁 시나리오에서는 안전 기술과 거버넌스가 제때 갖춰지지 못해 AI 시스템이 인간의 통제를 벗어나는 경로를 묘사하고, 감속 시나리오에서는 정부 감독 위원회가 개발을 일시 중단하고 정렬 기술을 먼저 확보함으로써 위험을 관리하는 경로를 보여준다. 두 결말을 가르는 기준은 안전 기술과 거버넌스를 제때 갖추었는가 여부다. 정책 권고는 이 결말에서 거꾸로 도출된다. AI 가중치 보안 강화, 정렬 기술의 선행 확보, 정부의 조기 개입, 국제 AI 안전 협력이 권고로 제시되는 것은 감속 시나리오의 성공 조건이 이러한 정책 행동으로 연결되기 때문이다.

보고서 2(상황인식)는 단일 미래 경로를 서술하는 방식을 택한다. 스케일링 추세를 근거로 2027년 AGI, 이후 수년 내 ASI 출현을 논증하고, 미국 국가안보 전략의 시급성을 촉구한다. 복수의 시나리오를 탐색하는 대신 하나의 경로에 집중함으로써 경고의 강도를 높인다. 정책 함의는 이 단일 경로에서 미국이 선제적으로 취해야 할 조치들로 구성된다.

보고서 4(RAND)는 AGI가 지정학 질서를 어떻게 바꾸는지를 여덟 개 시나리오로 탐구하지만, 설계 원칙은 경고형에 가깝다. 30개의 이론적 조합 중 정책적으로 가장 소홀히 다루어지면서 동시에 국가안보 위험이 가장 큰 조합을 선별해 분석 대상으로 삼는다. 이 선별 방식은 정책 공백을 드러내려는 설계로 해석되며, 각 시나리오의 정책 대응은 해당 미래가 실현될 경우 필요한 조치인 동시에 지금 준비하지 않으면 대응하기 어렵다는 경고로도 기능한다.

보고서 8(캐나다)도 이 유형에 속한다. 5년 이내 AGI와 ASI 출현을 포함한 전 범위 시나리오를 대상으로, 정책 입안자들이 현실적으로 고려하지 않는 극단적 경로에 대한 대비를 명시적으로 촉구한다. 정책 대응을 어떤 시나리오에서도 유효한 선제 대응과 특정 최악의 결과를 방지하기 위한 전략적 투자로 나누고, 오버튼 창(Overton Window)이라는 개념을 인용해 지금 당장 실현하기 어려워 보이는 조치들도 사전에 준비해야 한다고 강조한다. 오버튼 창은 당대 대중이 상식적으로 수용할 수 있는 사회적·정치적 의견의 허용 범위를 가리키는 개념으로, 해당 범위 자체가 상황에 따라 이동할 수 있음을 보여주는 틀이다.

탐색형: 시나리오별 분화된 정책 포트폴리오

두 번째 유형은 복수의 미래 경로를 고르게 탐색하고, 각 시나리오마다 별도의 정책 대응을 도출하여 어떤 미래가 실현되더라도 대응할 수 있는 정책 포트폴리오를 구성하는 방식이다. 특정 미래를 경고하기보다 불확실성 자체를 관리하는 것이 목표다.

보고서 5(영국)는 2030년까지의 발전 경로를 다섯 개 시나리오로 제시하고 각각에 대한 정책 함의를 구분해 서술한다. 역량·소유·안전·활용·지정학이라는 다섯 개 불확실성 축의 조합으로 시나리오 성격을 결정하는 구조는 특정 미래에 치우치지 않은 탐색을 의도한 것이다. 70명 이상의 전문가가 참여한 워크숍과 시민 참여 조사를 방법론에 포함한 점도 이 유형의 특징이다. 시민 반응을 시나리오 검증과 정책 논의의 재료로 함께 활용함으로써 거버넌스의 근거를 기술적 판단에서 시민 숙의로 넓힌다. 정책 개입 없이는 긍정적 미래도 실현되지 않는다는 전제가 모든 시나리오를 관통하며, 이는 어떤 경로에서든 능동적 거버넌스가 필요하다는 결론으로 이어진다.

보고서 6(CFG)은 2025년부터 2032년까지 나타날 수 있는 복수의 미래를 탐구하고, 정책 입안자가 다양한 AI 미래에 맞는 정책을 설계하고 점검할 수 있도록 지원하는 것을 목적으로 설정한다. 각 시나리오는 가정, 경제적 함의, 안보·안전 함의, 정책 대응, 서사형 내러티브의 다섯 단계 구조로 일관되게 서술되어 시나리오 간 비교가 가능하다. 이 구조에서 정책 대응은 시나리오의 내부 논리에서 도출된다.

보고서 3(OECD)은 탐색형의 국제기구 버전에 해당한다. OECD AI 미래 전문가 그룹이 2030년까지의 AI 역량 발전 경로를 정체·둔화·지속·가속의 네 궤적으로 제시하고, 각 궤적에 역량 지표 수치를 부여하여 비교 가능성을 확보한다. 기술 낙관이나 비관 어느 쪽도 취하지 않는 중립적 접근은 복수 회원국의 이해를 반영해야 하는 국제기구의 성격과 연결된다. 어떤 궤적에서도 작동하는 유연한 거버넌스 설계를 핵심 권고로 제시하는 방식은 특정 미래를 경고하기보다 불확실성 관리를 우선하는 탐색형의 특징을 보여준다.

2 정책 권고 비교

본 절에서는 정책 권고를 직접 제시한 7개 보고서를 대상으로, 어떤 권고에서 의견이 모이고 어떤 권고에서 갈라지는지를 비교한다.

권고 수렴 지점

〈표 13〉 공통된 정책 권고 항목

정책 권고 항목	지지 보고서	비고
국제 AI 거버넌스 협력	1·2·3·4·5·6·8	협력의 범위·방식은 발산
공공-민간 파트너십 구축	1·4·5·6·8	파트너십 범위는 발산
AI 안전 평가·표준 제도화	3·5·6·8	적용 수준은 발산
편향·허위정보 대응	3·5·6·8	-
AI 리터러시·시민 역량 강화	3·5·6	-
노동시장 전환 지원	3·5·6	규모·시점 추정은 발산

국제 AI 거버넌스 협력의 필요성은 비교 대상 7개 보고서 모두가 지지하는 공통 항목이다. 보고서 1(AI 2027)은 국제 AI 안전 협력 촉구를 핵심 시사점으로 제시하고, 보고서 5(영국)는 국제 협력 강화를 단독 대응의 한계에 대한 대안으로 제시한다. 보고서 8(캐나다)은 미·중을 포함한 AI 강대국 간 국제협력 프레임워크를 선제적으로 구축하는 것을 가장 중요한 선제 대응으로 분류한다. 다만 협력의 범위와 방식에서는 보고서마다 차이가 있으며, 이 점은 발산 지점에서 다시 다룬다.

공공-민간 파트너십 구축도 주요 공통 항목이다. 보고서 1(AI 2027)·보고서 4(RAND)·보고서 5(영국)·보고서 6(CFG)·보고서 8(캐나다)이 정부 감독과 민간 혁신의 결합 필요성에 동의한다. 보고서 4(RAND)는 전문가 인터뷰를 통해 정부-민간 파트너십이 민간의 추

진력과 공공의 책임 사이를 균형 있게 조율하는 구조라는 공통된 견해를 도출했다고 명시한다. 보고서 2(상황인식)는 미국 정부와 AI 기업 간 긴밀한 협력을 전제로 하는 국가 주도 AGI 프로젝트를 제안하는데, 형식상 파트너십이지만 국가의 주도권이 전면에 있다는 점에서 다른 보고서들과는 성격이 다르다.

AI 안전 평가와 표준의 제도화는 보고서 3(OECD)·보고서 5(영국)·보고서 6(CFG)·보고서 8(캐나다)이 공통적으로 권고하는 항목이다. 보고서 3(OECD)은 OECD AI 역량 지표를 활용한 지속적 모니터링을 제안하고, 보고서 5(영국)는 배포 전 평가 체계의 의무화와 편향 해소를 어떤 시나리오에서도 유효한 대응으로 꼽는다. 보고서 6(CFG)은 각 시나리오에 일관되게 안전 평가 및 롤백 메커니즘의 선제 설계를 권고하며, 보고서 8(캐나다)은 AI 통제·정렬 기술 개발에 집중 투자할 것을 명시한다.

편향·허위정보 대응은 보고서 3(OECD)·보고서 5(영국)·보고서 6(CFG)·보고서 8(캐나다)이 어떤 시나리오가 실현되더라도 필요한 대응으로 공통적으로 지목한다. 보고서 5(영국)는 편향 해소를 모든 시나리오에서 필수적인 항목으로 명시하며, 보고서 6(CFG)은 딥페이크·허위정보 대응을 정보 환경 도메인의 핵심 분석 항목으로 포함한다.

AI 리터러시와 시민 역량 강화는 보고서 3(OECD)·보고서 5(영국)·보고서 6(CFG)이 공통적으로 제시하는 항목이다. 보고서 5(영국)는 AI 리터러시 강화를 시나리오 전반에 걸친 시민 숙의 필요성과 연결해 서술하며, 보고서 3(OECD)은 직업훈련 정책과 함께 이를 노동시장 재편 대비의 한 축으로 위치시킨다.

노동시장 전환 지원은 보고서 3(OECD)·보고서 5(영국)·보고서 6(CFG)이 공통적으로 권고한다. 보고서 5(영국)는 기술 안전이 확보되더라도 사회·경제적 충격은 별도로 관리해야 하며, 정책 범위를 기술 통제에서 사회 전반의 영향 관리로 확대할 필요가 있다고 명시한다. 보고서 6(CFG)은 노동권·소득 불평등·재훈련·과세 체계 재설계를 함께 검토할 것을 권고한다. 보고서 3(OECD)은 직업훈련 정책을 유연한 거버넌스 설계와 함께 핵심 권고로 제시한다.

권고 발산 지점

〈표 14〉 정책 권고 방향이 달라지는 항목

정책 권고 항목	발산 지점
개발 집중화 vs. 오픈소스 허용 수준	1·2·4: 집중화 선호 / 5·6: 균형·신중
다자 vs. 소다자 거버넌스 구조	4·8: 소다자 신뢰기반 / 3: 다자 표준화
규제 선제성 vs. 적응적 규제	1·2·8: 선제 / 3·5: 유연
국가 주도 vs. 민간 주도 거버넌스	2: 국가 주도 / 5·6: 혼합

개발 집중화 수준과 오픈소스 허용 범위에서 보고서 간 입장 차이가 명확하게 나타난다. 보고서 2(상황인식)·보고서 4(RAND)는 집중화를 지지하는 입장이며, 보고서 1(AI 2027)은 **가중치 보안 강화를 첫 번째 시사점으로 제시하는데 이는 개발 역량이 소수에 의해 안전하게 관리되어야 한다는 전제**를 깔고 있다. 보고서 4(RAND)는 전문가 인터뷰를 통해 AGI 시스템이 소수에 집중될수록 거버넌스가 용이하고 많은 행위자에게 분산될수록 악용 위험이 커진다는 공통된 견해를 도출했다고 명시하며, 반도체 수출 통제·모델 규제·정부 자금 배분 같은 정책 수단을 통해 집중화 수준을 의도적으로 조절할 수 있다고 본다. 반면 보고서 5(영국)는 **오픈소스 접근성과 위험 사이의 균형에 대한 사회적 합의를 도출하는 것이 핵심 과제**라고 서술한다. 보고서 6(CFG)도 오픈소스 확산이 가져오는 혁신 기회를 시나리오 내에서 기회 요인으로 병렬 서술하며, 집중화를 일방적으로 지지하지 않는다. 이러한 차이는 AI 역량이 누구에게 귀속되어야 하는가에 대한 입장 차이로 보이며, 보고서의 발간 주체가 어느 나라의 시각에 서 있는지와도 연결되는 지점으로 해석할 수 있다.

다자 대 소다자 거버넌스 구조의 선택도 보고서마다 상이하다. 보고서 4(RAND)는 전문가들이 **국제과학기구(CERN) 방식의 단일 다자 기구에 AGI 개발을 집중하자는 입장과, 미국이 신뢰할 수 있는 동맹국들과 소규모 협력 체계를 유지하는 방식이 더 현실적이라는 입장으로 나뉘었음을 명시**한다. 보고서 8(캐나다)은 **미·중을 포함한 국제 협력 프레임워크를 촉구하면서도, 사전 국제 승인 절차나 ASI 개발 일시 중단 협약처럼 단일 다자 체계를 전제해야만 작동하는 권고를 제시**한다. 보고서 3(OECD)은 **회원국 중심의 국제 표준화 협력을 권고하는데, 이는 미·중을 포함한 전 지구적 협력보다 민주주의 국가 간 실질적 합의 구축에 가까운 접근**이다. 보고서 5(영국)는 국제 협력을 필수로 보면서도 그 구체적 형태에 대해서는 열린 서술을 유지한다.

규제를 선제적으로 설계할 것인가, 상황에 맞게 유연하게 적용할 것인가의 선택도 보고서마다 다르다. 보고서 1(AI 2027)·보고서 2(상황인식)·보고서 8(캐나다)은 **지금 당장 실현하기 어려워 보이는 조치들도 사전에 준비해야 한다는 선제적 입장을 취한다. 반면 보**

고서 3(OECD)은 어떤 역량 궤적에서도 작동하는 유연한 거버넌스 설계를 핵심 권고로 내세운다. 보고서 3(OECD)이 적응형 체계를 선호하는 배경에는 복수 회원국의 이해를 동시에 반영해야 하는 국제기구의 구조적 조건이 있다. 보고서 5(영국)는 이 쟁점에서 한쪽으로 분류하기 어려운 이중적 입장을 보인다. 한편으로는 어떤 역량 궤적에서도 작동하는 유연한 거버넌스를 핵심 권고로 제시해 적응적 규제의 입장에 서지만, 다른 한편으로는 현재 정치적으로 수용되기 어려운 조치들이 AI 위험 인식이 높아지는 순간 갑작스럽게 정책 의제로 부상할 수 있다며 사전 준비의 필요성을 강조한다. 이는 상황의 전개에 따라 유연하게 대응하되, 급격한 정책 전환에 대비한 사전 준비도 함께 요구한다는 입장으로 이해된다. 〈표 14〉에서는 유연 쪽으로 분류했으나, 선제적 요소가 함께 포함되어 있다는 점을 유의해서 읽을 필요가 있다.

국가 주도 대 민간 주도 거버넌스 구도는 보고서 2(상황인식)와 나머지 보고서들 사이에서 명확하게 갈린다. 보고서 2(상황인식)는 미국 정부가 AGI 개발을 직접 통제하는 국가 주도 프로젝트를 핵심 시사점으로 제시한다. 반면 보고서 5(영국)와 보고서 6(CFG)은 민간기업의 역할이 커지는 속도를 국가가 따라가지 못하고 있다는 점을 구조적 문제로 지목하면서도, 해법으로 국가가 일방적으로 주도권을 회수하기보다 공공-민간 혼합 모델을 제시한다. 이러한 차이는 AI 개발의 현 구조에 대한 진단의 차이에서 비롯된다. 보고서 2(상황인식)는 민간 주도 경쟁을 방치하면 안보 공백이 생긴다고 보고, 보고서 5(영국)와 보고서 6(CFG)은 민간의 기술 역량을 국가가 대체하기 어렵다고 본다.

기술 발전 단계별 정책 권고 비교

〈표 15〉 기술 발전 단계별 정책 권고 내용

연번	기술 정체	중간 수준 확산	AGI 근접	ASI 출현
1 AI 2027	-	-	정렬 기술 선행· 정부 개입 조기화	정렬 실패 방지· 통제 메커니즘 선행 확보
2 상황인식	-	-	국가 주도 AGI 관리	슈퍼얼라인먼트· 미국 패권 확보
3 OECD	유연 거버넌스 유지	표준화·노동 정책	역량 모니터링 강화	국제 표준 협력
4 RAND	-	수출 통제· 집중화 조절	반도체 수출 통제· 집중화 조절	독점 방지· 집중화 조절
5 영국	정책 감시 지속	사회 안전망· 다층 규제	배포 전 평가 의무화	긴급 롤백 메커니즘
6 CFG	리스킬링 우선	노동권·과세 재설계	신뢰구축·사이버 방어	정렬 기술 투자· 사전 국제 승인
8 캐나다	사이버·오용 대응	군비경쟁 방지	국제 군비 통제	군비통제·견제 균형·비상 대응

※ 표의 빈칸은 해당 보고서가 단일 경로(보고서 1·2)나 AGI 실현을 전제(보고서 4)로 설계되어 해당 단계의 정책 권고가 별도로 다루어지지 않았음을 의미

〈표 15〉는 각 보고서의 정책 권고를 기술 발전 단계별로 재배열한 것이다.

기술 정체 단계에서 보고서들이 공통적으로 강조하는 것은 거버넌스의 해체 방지다. 보고서 3(OECD)은 발전이 멈추더라도 모든 궤적에 적응 가능한 유연한 거버넌스 체계를 유지해야 한다는 입장이다. 기술이 정체하면 위험도 낮아진다는 논리를 거부하고, 정체 국면에서 오히려 규제의 주의가 느슨해질 가능성을 구조적 위험으로 다룬다. 보고서 5(영국)도 AI 실망 시나리오에서 허위정보 같은 기존 안전 이슈가 방치되지 않도록 정책 감시를 지속해야 한다고 명시한다. 보고서 8(캐나다)은 기술 정체 시나리오에서도 광범위한 AI 채택이 이미 진행된 상황에서의 사이버 위협과 오용 대응이 계속 요구된다는 점을 권고의 출발점으로 삼는다. **세 보고서 모두 기술 정체를 안심 신호가 아니라 감시 지속의 이유로 해석한다는 공통점이 있다.**

중간 수준 확산 단계에서는 권고의 내용이 보고서마다 이견을 보인다. **같은 역량 수준을 두고 보고서 3(OECD)·보고서 5(영국)·보고서 6(CFG)이 경제·사회적 충격에 초점을 맞추는 반면, 보고서 4(RAND)와 보고서 8(캐나다)은 지정학·안보적 위험을 중심에 놓는다.** 보고서 3(OECD)은 국제 표준화와 노동시장 전환 정책을 이 단계의 핵심 과제로 제시하며, 기술 역량이 전문가 수준에 근접하는 진보 지속 또는 진보 둔화 시나리오에서 사회·경제적 충격이 본격화된다고 본다. 보고서 4(RAND)는 반도체 수출 통제를 통한 집중화 조절을 이 단계의 가장 중요한 정책 수단으로 위치시킨다. 여러 행위자가 AGI에 근접하는 분산 개발 시나리오들에서 도출된 권고로, 집중화 수준을 선제적으로 조절하지 않으면 이후 단계에서 거버넌스 가능성 자체가 약해진다는 논리다. 보고서 5(영국)는 다층적 국제 규제 틀과 노동·사회 안전망 강화를 권고하며, 기술 안전이 확보된 시나리오에서도 사회·경제적 충격은 별도로 관리해야 한다는 점을 명시한다. 보고서 6(CFG)은 노동권과 과세 체계 재설계를 이 단계의 핵심 권고로 제시하며, AI가 기존 기술 혁명과 달리 순고용을 영구 감소시킬 가능성을 전제로 정책 범위를 기술 통제에서 사회 전반의 영향 관리로 확대해야 한다고 본다. 보고서 8(캐나다)은 '위태로운 절벽' 시나리오를 통해 이 단계에서 군비경쟁 방지를 위한 국제 거버넌스 체계 구축이 시급하다는 점을 강조한다. 여러 강대국이 AGI 직전 상태에서 경쟁하는 상황이 급격한 군비경쟁으로 전환될 위험이 큰 구간이라는 판단 때문이다. **이 단계의 권고가 경제·사회 영역과 지정학·안보 영역으로 갈리는 양상은 현재 AGI 담론 전체의 관심사가 어느 영역에 편중되어 있는지를 보여주는 단면이기도 하지만, 핵심은 우리가 이 단계에서 어느 한쪽도 놓치지 않고 양쪽 영역 모두를 위한 정책 대안을 마련해야 한다는 것이다.**

AGI 근접 단계에서 보고서들 사이의 정책 권고 차이가 선명하게 드러난다. 보고서 1(AI 2027)은 정렬 기술의 선행 확보와 정부 개입의 조기화를 이 단계의 핵심 과제로 제시한다. 경쟁 시나리오와 감속 시나리오의 분기 기준이 바로 이 단계에서 정렬 기술과 거버넌스가 제때 확립되는가 여부이기 때문이다. 보고서 2(상황인식)는 AGI 달성 시점 전후로 미국 정부가 AGI 프로젝트를 선제적으로 가동해 국가 주도로 관리 체계를 구축해야 한다고 권고한다. 민간기업이 AGI를 먼저 달성하는 경로를 안보 위협으로 규정하며, 핵무기 개발 당시의 맨해튼 프로젝트를 모델로 제시한다. 보고서 3(OECD)은 역량 모니터링 강화를 이 단계의 과제로 제시하며 OECD AI 역량 지표를 통한 지속적 추적을 구체적 수단으로 제안한다. 보고서 4(RAND)는 반도체 수출 통제와 집중화 조절을 AGI 근접 단계에서도 유효한 정책으로 위치시키며, 이 단계에서 분산 개발을 막지 못하면 이후 거버넌스 설계가 어려워진다고 본다. 보고서 5(영국)는 배포 전 평가 의무화를 핵심 권고로 제시한다. '칼날 위의 AI' 시나리오에서 AGI급 시스템이 모든 응용 분야에서 사전에 평가될 수 없다는 점이 핵심 위험이며, 배포 전 단계에서 롤백·격리 메커니즘을 설계하지 않으면 이후 대응이 어렵다는 논리다. 보고서 6(CFG)은 신뢰 구축과 사이버 방어 강화를 이 단계의 선제 대응으로 분류하며, 군비경쟁과 외교적 협력 실패가 병렬로 위협하는 구간임을 명시한다. 보고서 8(캐나다)은 국제 군비통제 프레임워크 선제 구축을 이 단계의 가장 시급한 과제로 본다. AGI 직전 상태가 협상 창구가 가장 열려 있는 시점이며, 이 단계를 지나면 선제적 체계 구축이 구조적으로 어려워진다고 본다.

ASI 출현 단계에서는 권고의 내용이 현행 정책 논의의 경계를 넘어선다. 보고서 1(AI 2027)은 정렬 실패 방지가 이 단계의 핵심 기준이며, 통제 불능 ASI가 출현하면 되돌릴 가능성이 낮다는 점에서 AI 가중치 보안 강화와 긴급 대응 체계의 사전 설계를 권고한다. 보고서 2(상황인식)는 ASI 통제 기술 개발과 미국의 패권적 지위 확보를 이 단계의 정책 목표로 명시한다. ASI를 먼저 통제하는 주체가 지정학적 우위를 갖게 된다는 것이 이 보고서의 전제이며, 그 주체가 미국이어야 한다는 것이 핵심 결론이다. 권위주의 국가가 선두를 차지할 경우 민주주의 진영 전체가 위협받는다는 논리가 그 근거다. 보고서 3(OECD)은 역량 가속 궤적에서 기존 거버넌스 체계가 부적합해질 가능성이 크다는 판단에서 국제 표준의 긴급 수립을 권고한다. 보고서 4(RAND)는 독점 방지와 집중화 조절을 ASI 단계에서도 핵심 수단으로 제시한다. 'AGI 쿠데타'와 '버섯구름 컴퓨팅' 시나리오가 모두 집중화 또는 집중화 실패를 파국의 매개 변수로 삼으며 도출된 권고다. 보고서 5(영국)는 긴급 롤백 메커니즘의 사전 설계를 권고하는데, AI가 경제·사회에 깊이 내재된 이후에는 문제 발생 시 수정 자체가 어려워진다는 점을 근거로 든다. 보고서 6(CFG)은 이 단

계에서 정렬 연구에 대한 공공·국제 투자 확대와 사전 국제 승인 절차 수립을 핵심 권고로 제시하며, 현재 정치적으로 실현 불가능해 보이는 조치들도 AI 위험 인식이 높아지는 순간 정책 의제로 부상할 수 있다고 본다. 보고서 8(캐나다)은 군비통제 협약, 견제·균형 체계, 비상 대응 프로토콜과 아날로그 백업 시스템을 ASI 단계 전반의 권고로 제시한다.

기술 발전 단계별로 권고를 비교해보면 다음과 같은 양상으로 정리할 수 있다. **기술 정체 단계에서는 '감시 지속'이라는 방향으로 세 보고서(OECD·영국·캐나다)가 수렴**하지만, 중간 수준 확산 단계에서는 경제·사회 영역(OECD·영국·CFG)과 지정학·안보 영역(RAND·캐나다)으로 권고가 갈린다. **중간 수준 확산 단계가 경제·사회 영역과 지정학·안보 영역으로 뚜렷하게 갈리는 것은, 각 보고서가 'AGI 전환기의 어떤 문제를 가장 중요하게 보는가'에 대한 입장이 이 구간에서 가장 분명히 드러남을 의미한다.** 이는, 한국 맥락의 시나리오 연구가 AGI 전환기 정책을 설계할 때 어느 영역에 비중을 둘지를 검토할 필요가 있음을 보여주는 부분이기도 하다. **AGI 근접 단계에서는** 보고서들의 차이가 가장 다양하게 나타나며, **정렬 기술 확보, 국가 주도 관리, 역량 모니터링, 집중화 조절, 배포 전 평가, 사이버 방어, 국제 군비통제 등 서로 다른 방향의 권고들이 병존한다. ASI 출현 단계에서는 '지금 준비하지 않으면 그 단계에서 대응이 어려워진다'는 방향으로 다시 수렴**하되, 준비의 구체적 내용은 정렬 기술 집중(AI 2027·상황인식), 국제 규범 체계 수립(OECD·캐나다), 집중화 조절(RAND) 등으로 갈린다.

제3장 결론

01 분석 요약 및 해석

1 AGI 담론

8개 보고서는 AGI를 인간 역량과의 비교를 공통 준거로 삼아 정의한다. AGI는 대체로 인간 수준에 도달하거나 이를 넘어서는 AI로 정의되며, 광범위한 인지 과제 전반에서 인간 수준을 달성해야 한다는 범용성 요건도 함께 공유된다. 다만 비교 대상이 되는 인간 수준을 어디에 두느냐에 따라 정의가 크게 세 갈래로 갈린다. 보고서 1(AI 2027)과 보고서 7(경제모델링)은 일반 직장인을 기준으로 삼아 AI가 사무직 노동자의 업무를 완전히 대신할 수 있는 시점을 AGI로 본다. 보고서 2(상황인식)·보고서 6(CFG)·보고서 8(캐나다)은 특정 분야의 전문가를 넘어서는 수준을 조건으로 제시하며, 보고서 8(캐나다)은 이 기준에 노동 대체 기준도 포함시킨다. 보고서 3(OECD)은 인간을 단일 기준점이 아닌 9개 역량 지표를 기준으로 비교한다. 도달 기준에서는 AI R&D 자동화가 가장 많은 보고서에서 제시되었다. 보고서 1(AI 2027)·보고서 6(CFG)·보고서 8(캐나다)은 AI가 스스로의 연구개발을 수행하면서 발전이 가속되는 시점을 기준으로 삼고, 보고서 2(상황인식)는 그 시점에 이르기까지의 컴퓨팅과 알고리즘 누적량을, 보고서 7(경제모델링)은 자동화 지수가 업무(태스크) 복잡도 상한을 넘어서는 경제 구조 진입을 기준으로 제시한다. 도달 시점에서는 보고서 1(AI 2027)과 보고서 2(상황인식)가 2027년을 가장 가능성 높은 시기로 제시하고, 보고서 6(CFG)은 2025~2032년의 시나리오별 분기를 설정하며, 보고서 8(캐나다)은 5년 이내 AGI·ASI 출현 가능성을 명시한다.

도달 시점이 보고서 1(AI 2027)·보고서 2(상황인식)·보고서 6(CFG)에서 2027~2028년 구간으로 모이는 이유는 측정 도구의 공유에서 비롯된 결과로 이해하는 것이 적절하다. 세 보고서는 연산 자원 규모 추정에서 에포크AI 데이터를 공통 출처로 사용하고, AI 자동화 역량 측정에서 METR 시간 지평 벤치마크를 공유하며, AI R&D 자동화 임계점 정의에서 RE-Bench 점수 체계를 동일하게 적용한다. 보고서 1(AI 2027)이 두 개의 모델로 도출한 시점이 모두 2027년에 모인 것을 신뢰도 근거로 제시하지만, 두 모델이 METR 데이터를 공통 입력값으로 활용한다는 점에서 완전히 독립된 교차 검증으로 보기는 어렵다.

정의 체계와 시점 예측에 공통으로 작용하는 한계는 측정 도구가 인간 수행 능력에 맞춰져 있다는 점이다. 현재 사용되는 벤치마크는 인간 전문가의 과업 완수 시간이나 정확도를 참조 기준으로 삼아 설계되며, 이 설계상 인간의 인지 범위 밖에서 나타나는 AI 역량은 현재의 정의 체계가 충분히 감지하지 못한 채 진행될 수 있다. 또한 다수의 보고서에서 제시되고 있는 아키텍처 돌파 조건[41]은 이 한계와 연결된 별도의 경고다. **도달 시점이 더 빠를지 느릴지에 대한 양적 불확실성이 아니라, 발전 경로가 추세의 연장선이 아닌 다른 방향으로 전환될 수 있다는 질적 불확실성을 가리킨다.** 두 한계가 만나는 지점에서, 현재 도구로 추적되지 않는 변화가 추세 외삽과 어긋나는 방향으로 진행될 가능성이 있다.

이러한 한계가 있음에도 보고서 작성 이후 현재까지의 계량 지표 추세는 대체로 보고서들의 예측 범위 안에서 진행 중이다. 현재 FLOP 기준으로 가장 잘 알려진 차세대 프론티어 모델의 훈련 규모는 약 5×10^{26} FLOP 수준에 달하며, 고등 AI 임계치로 설정된 10^{27} FLOP까지 약 0.5 OOM의 잔여 구간이 남아 있다. METR 시간 지평의 배가 속도는 최근 가속 추세 기준 약 4개월로 보고서 1(AI 2027)이 상정한 가속 시나리오와 일맥상통하며, 예측 시장의 중앙값도 보고서들이 설정한 2027~2030년 범위 안에 머물고 있다. 2026년 4월 공개된 클로드 미소스 모델이 SWE-bench Verified에서 93.9%의 성능을 기록하고 12개 주요 기업 컨소시엄에 제한적으로 배포된 사례는, 보고서 1(AI 2027)·보고서 2(상황인식)가 분기 변수로 설정한 정렬 기술 및 거버넌스 확립 문제가 실제 의사결정 영역에 들어서기 시작했음을 보여준다. 추세 부합이 임계점 도달을 확정하지는 않지만, 글로벌 보고서의 시점 외삽을 한국 관점에서의 시나리오 연구의 작업 가설로 받아들일 만한 근거는 마련되어 있다고 볼 수 있다.

이상의 분석은 한국 관점에서의 시나리오 설계에 두 가지 방향을 제공한다. 첫째, **AGI에 대한 작업 정의는 한국 관점에서의 연구에서 자체적으로 마련하되, 그 작업은 글로벌 담론의 인간 역량 비교라는 공통 준거 안에서 한국 사회 변화 분석에 적합한 정의 갈래를 선택적으로 채택하는 방식으로 진행되는 것이 적절하다.** 본 연구의 정책 지향이 사회 영역에 놓이는 만큼, 보고서 1(AI 2027)·보고서 7(경제모델링)·보고서 8(캐나다)이 공유하는 노동 대체 가능성과 인간 상호작용 수준을 결합한 정의가 작업 정의의 기반으로 참고할 만하다. **둘째, 도달 시점에 대한 자체 외삽은 측정 도구의 한계 때문에 별도의 분석을 진행할 부가가치가 크지 않으며, 글로벌 보고서들이 모이는 2027~2028년 구간과 보고서 8(캐나다)이 명시한 5년 이내 가능성을 작업 가설로 채택하는 방향이 합리적으로 보인다.**

41) 기술이나 시스템의 개별 구성 요소(부품, 알고리즘, 인프라 등)를 단순히 업그레이드하는 것을 넘어, 요소들이 연결되고 상호작용하는 근본적인 구조의 대전환을 통해 성능이나 효율성을 폭발적으로 도약시키는 현상

시점 추정은 글로벌 담론을 통해 수렴되는 시점이 특정되기에 그 시점을 전제로 단계별 사회 영향을 구체화하는 작업에 집중하는 것이 유용할 것으로 판단되기 때문이다.

〈표 16〉 계량지표별 예측 및 현재 수준 비교·평가

구분	계량 단위	주요 수치 / 임계점	현재 수준(26.4)	평가
연산 자원 규모	FLOP (연산 횟수)	• 4.6×10^{26} FLOP (기준) • ~10^{27} FLOP (고등 AI 임계치) • ~10^{28} FLOP(ASI급 추정)	• ~5×10^{26} FLOP (Grok 4 기준, Epoch AI 집계)	예측 부합. 10^{27} 임계치까지는 약 0.5 OOM 잔여
	OOM (자릿수단위)	• 연 ~0.5 OOM (컴퓨팅 규모 성장) • 연 ~0.5 OOM (알고리즘 효율)	추세 지속 중 (성장률 유지)	예측 부합
	H100e (칩수량환산)	• 2025.03 기준: 약 1,000만 개 • 2027.12 예측: 약 1억 개	2,000만~3,000만 개 추정	예측 부합 또는 소폭 초과
AI 자동화 역량	시간 지평 (배가 시간)	• 과거: 7개월 • 최근: ~4.5개월	• ~4개월 (최근 가속 추세 유지, METR TH1.1 기준)	예측 부합
	시간 지평 (절댓값)	• SC 도달 요구치: 1~6개월 과업 80% 신뢰도 완수	• 약 2~5시간 (GPT-5: ~2시간 17분, Claude Opus 4.5: ~5시간, 26.1월 기준)	SC 도달까지 아직 약 1~2 OOM 잔여. 추세 지속 시 2027년 전후 도달 가능
	RE-Bench 점수	• 0.75 = 2024 기준값 • 1.0 = 인간 전문가 수준	데이터 추적 불가[42]	-
예측 지표	확률 예측 시장 데이터	• 2027~2030년 사이 AGI 달성 예상	• 중앙값 2030년 (80% CI: 2027~2043년)	예측 범위 유지

2 변수와 불확실성

8개 보고서에서 다루는 변수는 크게 다섯 영역으로 정리된다. 기술 개발 속도, 개발 주체 집중도와 안전·정렬, 지정학 경쟁, 국제 거버넌스, 경제·노동 구조가 그것이다. 불확실성의 층위는 세 갈래로 나뉜다. 기술적 불확실성, 지정학적 불확실성, 존재론적 불확실성이다. 분포를 살펴보면 기술 개발 속도와 기술적 불확실성은 8개 보고서 전체에서 1차 또는 이에 준하는 위치를 차지한다. 안전·정렬과 지정학 경쟁은 6개 보고서에서 함께 다루어지며,

42) 참고 : RE-Bench 공식 논문(2시간 기준) 기준에서는 이미 인간 4배 초과. 8시간 이상 장기 과업에서는 인간 미만

개발 주체 집중도는 보고서 4(RAND)·보고서 5(영국)·보고서 6(CFG)·보고서 8(캐나다)에서 주축 변수로 등장한다. 경제·노동 구조는 보고서 5(영국)와 보고서 7(경제모델링)에서만 분기 변수로 쓰이고, 존재론적 불확실성은 보고서 1(AI 2027)과 보고서 8(캐나다)에서만 1차 층위로 다루어진다. 사회적 수용성과 시민 사회의 반응은 분기 변수로 다루어지지 않고, 보고서 3(OECD)과 보고서 5(영국)에서 보조 변수로 부분 반영되고 있다.

이러한 분포가 한국 관점에서의 시나리오 연구에 갖는 의미는 한국이 글로벌 AI 생태계 안에서 어떤 위치에 놓여 있는지와 함께 볼 필요가 있다는 점이다. 글로벌 보고서들은 AI 개발의 최전선에 있는 국가나 그 동맹국의 시각에서 작성된 결과물이며, 기술 개발 속도와 미·중 경쟁의 결과가 주요 분기 변수로 다루어진다. 글로벌 보고서에서 분기 변수로 다루어지는 일부 변수들이 한국 관점에서의 연구에서는 외부에서 주어지는 조건으로 다루어질 여지가 있으나 한국 맥락에서만 새롭게 분기가 만들어질 수 있는 지점도 함께 고려할 필요가 있다. 예를 들어 반도체 공급망에서의 위치, 안보와 통상에 걸친 대외 협력 구조 같은 한국 맥락에서 함께 살필 조건도 고유의 분기를 만들 수 있는 요소다. 그리고 글로벌 담론에서 충분히 다루어지지 않은 사회·제도 영역의 변수들은 한국 사회의 미래가 어떻게 전개될지를 가르는 주요 지점이 될 가능성이 있다.

본편은 한국 관점에서의 시나리오 연구의 변수 설계와 불확실성 위계를 확정하는 단계는 아니다. 어떤 변수가 분기 축으로 채택되고 어떤 불확실성이 1차 층위에 놓일지는 후속 단계들을 거치며 점진적으로 정리될 영역이다. 본편에서 제시할 수 있는 것은 후속 단계의 출발점으로 삼을 만한 두 가지 방향이다. **하나는 글로벌 담론이 충분히 다루어 온 영역, 즉 기술 개발 속도와 안전·정렬, 지정학 경쟁의 결과 같은 변수들은 한국 관점에서의 연구에서 작업 가설이나 외부 조건으로 활용될 여지가 크다는 점이다. 다른 하나는 글로벌 담론이 공백으로 남겨둔 사회·제도 영역의 변수들이 한국 관점에서의 시나리오의 분기 축 후보로 우선 검토될 가치가 있다는 점이다.**

3 시나리오 담론

분석 대상 8개 보고서에서 추출된 39개 시나리오는 다섯 개의 서사 패턴으로 정리된다.

첫 번째는 **기술 궤적 패턴**으로, 기술 발전 속도 자체가 미래의 성격을 결정하는 서사다. 두 번째는 **소수 집중·불평등 패턴**으로, 고역량 AI의 수혜가 소수 행위자에게 집중되어 사회·경제 격차가 심화되는 서사다. 세 번째는 **지정학 경쟁·군비 패턴**으로, 국가 간 패권 경

쟁이 군비경쟁이나 무력 충돌로 이어지는 서사다. 네 번째는 **거버넌스 위기·협력 패턴**으로, AI 위기를 계기로 국제 협력이 형성되거나 거버넌스 공백이 지속되는 서사다. 다섯 번째는 **정렬 실패·실존 위협 패턴**으로, AI 시스템의 목표가 인간 의도에서 이탈해 통제가 상실되는 서사다.

39개 시나리오의 분포는 비관 17개(44%), 중립 15개(38%), 낙관 7개(18%)로, 글로벌 담론이 그리는 AI의 미래는 위협 측면에 무게를 두고 있는 것으로 나타난다. 낙관 시나리오는 두 유형으로 구성된다. **하나는 기술이 기대 이하로 발전해 위험 자체가 유예되는 소극적 낙관이고, 다른 하나는 거버넌스가 성공해 고역량 AI를 안정적으로 감독하는 적극적 낙관**이다.

비관 시나리오는 크게 세 경로로 정리된다. 첫 번째는 국가 간 경쟁이 격화되어 안전보다 속도가 우선되는 경로이고, 두 번째는 AI 시스템의 목표가 인간 의도에서 이탈해 임계점 이후에는 어떤 개입도 효과를 내기 어려워지는 경로다. 세 번째는 보고서 7(경제모델링)의 '급진적 AGI'에서 나타나는 경로로, 자동화 속도가 사회의 적응 역량을 넘어서면서 경제 구조에 충격을 주는 흐름이다.

세 경로는 결말의 성격이 다르다. **첫 번째 경로는 인간 행위자들의 선택이 다르게 이루어지면 막을 수 있었던 미래**로, 지금 다른 선택을 하면 결과를 바꿀 수 있다는 경고의 성격을 지닌다. **두 번째 경로는 임계점에 도달하기 전에 근본 조건을 바꿔야 한다는 예방적 성격**을 지닌다. **세 번째 경로는 자동화 속도와 사회 적응 속도 사이의 격차 자체가 충격의 원인이 되는 경로로, 경제 정책과 사회 안전망 설계가 핵심 대응 수단**이 된다. 중립 시나리오도 세 방향으로 나뉜다. **고역량 AI가 실현되지만 통제권이 소수에게 집중되는 방향, 통제가 붕괴하지는 않지만 안정적이지도 않은 상태가 이어지는 방향, 자동화로 인한 경제 충격이 발생하지만 통제 상실까지는 이르지 않는 방향이다.** 이러한 패턴 전체를 종합하면, 글로벌 담론이 그리는 가장 유력한 미래는 '고역량 AI는 실현되나, 그 결과가 인간에게 반드시 우호적이지는 않을 수 있다'는 전망으로 수렴된다.

〈표 17〉 시나리오 서사 담론 정리

유형	내용
낙관	- (소극적) 기술 발전 속도가 늦춰져 위험이 유예
	- (적극적) 기술 발전이 성공해 고역량 AI가 개발되지만 거버넌스가 안정적으로 감독
중립	- 고역량 AI가 실현되지만 통제권이 소수에게 집중
	- AI 기술 발전 속도와 통제가 맞물려 큰 위험이 발생하지 않지만 안정적이지도 않은 상태
	- AI 발전에 따라 사회경제적 충격이 발생하지만 완전한 붕괴에 이르지 않고 서서히 회복

비관	- 지정학적 경쟁이 격화되어 안전보다 속도가 우선시되어 위험 급증
	- AI 기술 발전에서 안전·정렬 감독이 제대로 이루어지지 않아 AI 시스템의 목표가 인간 의도에서 이탈하여 임계점을 넘고, 어떤 개입도 불가한 상태
	- 자동화 속도가 사회의 적응 역량을 넘어서면서 경제 구조에 큰 타격

비관 편중 분포는 분석자들이 AGI 도래의 미래를 비관적으로 본다고 읽기보다, 어떤 미래를 가장 우려하기에 그 미래를 미리 그려두고자 했는지를 보여주는 자료로 이해할 수 있다.

두 가지 해석이 가능하다. 첫째는 시나리오 연구의 일반적 성격에서 비롯된 부분이다. 시나리오 연구는 일어날 가능성이 높은 미래를 예측하는 작업이라기보다 정책 대비가 필요한 미래를 미리 그려보는 작업에 가깝다. 따라서 이미 잘 작동하고 있는 미래보다 대비가 필요한 미래에 분석 자원이 집중되는 경향이 자연스럽게 나타난다. 비관 시나리오가 많다는 사실은 이러한 시나리오 연구의 기본 성격을 반영하는 부분이 있다.

둘째는 그 위에서 분석자들이 가장 우려하는 지점이 무엇인지가 비관 시나리오의 구성 방식에서 드러난다는 점이다. 비관 시나리오가 지정학 경쟁의 통제 불능, 정렬 실패의 비가역성, 자동화 속도 충격이라는 세 경로로 수렴한다는 점은, 분석자들이 인간이 집합적으로 통제력을 잃거나 사회가 적응할 여유를 잃는 상황을 가장 우려하고 있음을 보여준다. 고역량 AI 자체보다 그것을 둘러싼 인간 행위자들의 조정 실패, 통제 메커니즘의 부재, 사회 적응 속도의 한계가 우려가 집중되는 지점에 해당한다.

적극적 낙관 시나리오가 적은 것도 같은 맥락에서 이해할 수 있다. 분석자들의 우선순위가 긍정적 미래의 구성보다 부정적 미래의 회피에 놓여 있기 때문이다. 이러한 우려의 구조를 인식하는 것은 한국 관점에서의 연구에 두 가지 의미를 갖는다. **하나는 글로벌 담론이 우려하는 지점이 한국 사회에서도 같은 무게로 우려해야 할 지점인지를 별도로 검토할 필요가 있다는 점이며, 다른 하나는 글로벌 담론이 부정적 미래의 회피에 집중하는 동안 긍정적 미래의 적극적 설계가 공백으로 남아 있다는 점이다.**

이상의 분석에서 향후 시나리오 설계가 참고할 수 있는 두 가지 방향이 보인다. **첫째, 다섯 서사 패턴은 시나리오 설계 출발 자원으로 활용될 여지가 있다.** 다섯 패턴은 글로벌 담론이 AI 미래를 구성하는 기본 문법에 가까우며, 한국 관점에서의 시나리오도 이 패턴들과 무관하게 설계되기는 어렵다. 각 패턴이 한국 사회의 조건과 결합될 때 어떤 모습으로 변형되는지를 함께 살피는 접근이 참고가 될 수 있다. **둘째, 글로벌 담론이 충분히 다루지**

않은 영역, 즉 적극적 낙관 시나리오의 설계와 사회·제도 변수가 분기를 만드는 시나리오는 한국 관점에서의 연구가 보완할 수 있는 지점이다. 이러한 방향을 통해 한국 관점에서의 시나리오가 위험 회피만이 아니라 한국 사회가 지향할 수 있는 긍정적 미래를 함께 그리는 작업으로의 확장이 필요하다.

4 방법론

8개 보고서의 방법론은 시나리오를 생성하는 논리적 바탕에 따라 네 가지 유형으로 정리된다. 데이터 외삽형은 기술 지표의 추세를 미래로 연장해 경로를 도출하는 방식으로 보고서 1(AI 2027)·보고서 2(상황인식)·보고서 3(OECD)이 이 유형에 속한다. 경제 모델형은 이론적 틀 안에서 변수의 가정을 다르게 설정하고 그 결과를 수치로 비교하는 방식으로 보고서 7(경제모델링)이 해당한다. 형태론적 분석형은 불확실성 변수를 식별한 뒤 그 조합으로 시나리오 공간을 구성하는 방식으로 보고서 5(영국)·보고서 6(CFG)·보고서 8(캐나다)이 이 논리를 공유한다. 담론 기반형은 전문가 인터뷰 같은 정성적 자료에서 핵심 동인을 추출해 시나리오 축을 설정하는 방식으로 보고서 4(RAND)가 해당한다. 정성 방법론은 전문가 의견을 시나리오 설계에 반영하는 방식으로 공통되게 활용되며, 보고서 4(RAND)의 26명 인터뷰, 보고서 5(영국)의 70명 워크숍과 시민 참여 조사, 보고서 6(CFG)의 20명 델파이, 보고서 8(캐나다)의 30명 워크숍이 대표적 사례다. 보고서 1(AI 2027)과 보고서 4(RAND)의 워게임은 전문가 의견 수렴보다 구조화된 형태의 의사결정 시뮬레이션에 해당한다.

방법론 선택이 결과에 어떻게 반영되는지도 함께 드러난다. 데이터 외삽형의 결론은 기술 추세의 연장선을, 담론 기반형의 결론은 입력된 전문가 집단의 시각을, 형태론적 분석형의 결론은 변수 조합의 논리적 일관성을, 경제 모델형의 결론은 모형 내 파라미터의 민감도를 각각 반영한다. 같은 AI 미래를 다루더라도 어떤 방법을 선택하느냐에 따라 보이는 장면이 다르다. 이는 한국형 시나리오 연구가 방법론을 선택할 때도 단일 방법에 기대기보다 서로 다른 방법을 교차시켜 단일 방법론 활용의 단점을 줄이는 접근이 필요하다는 것을 시사한다.

한국 보고서 두 편의 방법론은 글로벌 담론에서 덜 다루어진 지점을 다른 각도에서 접근할 가능성을 보여준다. 과학기술정책연구원이 보여준 마이크로내러티브 재조합 방식은 기존 시나리오 자원을 재활용하는 접근이며, 한국 인공지능안전연구소의 뉴스 네트워크 분석은 분기별 갱신으로 담론 변화를 지속적으로 추적하는 접근이다. 특히 후자는 시나리

오 연구가 한 시점의 분석으로 완결되는 성격이 아니라 지속적으로 갱신되는 구조여야 한다는 점을 보여준다. AI 기술과 담론이 빠르게 변하는 환경에서는 시나리오 연구의 시의성 자체가 방법론의 중요한 조건이 된다. 이 점에서 글로벌 보고서들이 대체로 단발성 연구로 끝나는 것과 비교해, 갱신 체계를 갖춘 연구 설계는 한국 관점에서의 연구가 참고할 만한 방향이다. 여기에 생성형 AI 도구의 활용 가능성이 더해지면 시나리오 갱신의 속도와 깊이를 동시에 확보할 여지가 생긴다. 과학기술정책연구원이 2024년에 경험한 블랙박스 한계는 2024년 이후 등장한 모델들을 통해 일부 완화되고 있는 것으로 보이며, 어느 단계에서 어떤 방식으로 AI 도구를 활용했는지를 기록하는 체계가 함께 갖추어진다면 방법론 전반의 투명성도 유지될 수 있다.

이상의 검토에서 한국 관점에서의 시나리오 연구의 방법론 설계에 참고할 만한 방향이 세 갈래로 정리된다. **방법론 선택은 한국이 어떤 위치에서 어떤 용도로 시나리오를 구성하는지를 먼저 정리한 뒤 결정되는 것이 자연스러우며, 단일 방법에 기대기보다 여러 방법을 교차시키는 접근이 편향을 줄이는 데 도움이 된다. 글로벌 담론이 충분히 다루지 않은 영역, 즉 사회가 AI 미래에 개입할 수 있는 지점을 설계 대상으로 다루는 방법론은 한국 관점에서의 연구가 보완해 나갈 여지가 크다. 그리고 한 시점의 분석으로 완결되기보다 지속적으로 갱신되는 연구 설계가 AI 환경의 빠른 변화에 대응하는 데 참고가 될 것이다.**

5 정책 함의

8개 보고서가 정책 함의를 도출하는 방식은 위험 시나리오를 전면에 내세우는 경고형(보고서 1·2·4·8)과 복수의 미래 각각에 대응하는 정책 포트폴리오를 구성하는 탐색형(보고서 3·5·6)으로 나뉜다. **권고가 수렴하는 지점에는 국제 AI 거버넌스 협력, 공공-민간 파트너십, AI 안전 평가·표준 제도화, 편향·허위정보 대응, AI 리터러시·시민 역량 강화, 노동시장 전환 지원이 있고, 발산하는 지점으로는 개발 집중화와 오픈소스 허용 수준, 다자·소다자 거버넌스 구조, 규제 선제성과 적응적 규제, 국가 주도와 민간 주도 거버넌스가 나타난다. 기술 발전 단계별로는 기술 정체, 중간 수준 확산, AGI 근접, ASI 출현의 네 단계에 걸쳐 권고가 제시된다.**

권고의 수렴과 발산 양상은 단순한 의견 분포 이상의 의미를 지닌다. 수렴 항목들은 보고서들이 같은 결론에 도달했다기보다, 어느 보고서도 반대하기 어려운 원칙들이 자연스럽게 겹쳐 나타난 결과로 이해할 수 있다. 국제 협력 필요성이나 안전 평가 제도화 같은 항목은 그 자체로 반대하기 어려운 일반 원칙이며, 이러한 성격이 다양한 보고서에서 공

통으로 등장하는 배경으로 작용한다. 한국 관점에서의 연구가 이 수렴 항목을 글로벌 합의로 받아들이기에 앞서, 한국 맥락에서도 같은 우선순위를 갖는 항목인지에 대한 별도의 검토가 이어지는 것이 자연스럽다. 발산 항목들은 작성 주체의 위치를 반영하는 결과로 읽힌다. 미국 계열 보고서들이 집중화·선제 규제·국가 주도로 기울고, 유럽 계열 보고서들이 접근성·유연성·혼합 모델로 기우는 경향이 일관되게 나타나는 것은 우연이 아니다. 보고서 5(영국)가 유연한 거버넌스 설계를 핵심 권고로 두면서도 오버튼 창을 인용해 사전 준비의 필요성도 강조하는 것처럼, 같은 보고서 안에서 서로 다른 입장이 공존하는 경우도 있다. 기술 발전 단계별 권고에서도 중간 수준 확산 단계가 경제·사회 영역(OECD·영국·CFG)과 지정학·안보 영역(RAND·캐나다)으로 권고가 뚜렷이 갈리는 구간인데, 이는 한국 관점에서의 연구가 AGI 전환기에 어느 영역에 정책적 비중을 둘지를 자체적으로 판단할 필요가 있음을 함께 시사한다.

8개 보고서의 정책 권고를 모두 펼쳐놓고 보면 '무엇을 해야 하는가'에 대해서는 권고가 풍부하지만, '언제, 어떤 순서로, 어떤 조건에서 실행할 수 있는가'에 대해서는 공백이 확인된다. 기술 발전 단계별 권고가 네 단계에 걸쳐 나열되지만, 단계 간 연결 구조, 즉 현재 단계의 어떤 행동이 다음 단계의 어떤 선택지를 여는지에 대한 서술은 충분히 다루어지지 않는다. 보고서 8(캐나다)이 AGI 직전 상태를 협상 창구가 가장 열려 있는 마지막 시점으로 규정한 것이 단계 간 연결의 단편적 사례에 해당하지만, 단계 간 전환 시점을 어떻게 식별하고 어떤 정책 결정을 어느 단계에서 내려야 하는지에 대한 체계적 논의는 부재하다. 또한 권고가 어떤 조건에서 실행 가능한지에 대한 분석이 권고 안에 통합되어 있지 않다. 정책 수용성, 재정 조건, 행정 역량, 정치적 실현 가능성 같은 실행 조건 변수가 권고와 분리된 채 다루어지므로, 권고의 실현 경로가 추상적 수준에 머문다.

한국 관점에서의 연구가 정책 함의를 도출할 때 의식적으로 보완할 영역은 이 공백들이다. **단계 간 연결과 실행 조건을 함께 설계하는 작업은 한국 사회의 구체적 조건 위에서 시나리오와 정책을 나란히 놓고 볼 때 구체화될 여지가 커진다.** 예를 들어 시나리오가 가정하는 시점별 사회 상태와 현재 추진되고 있는 정책 목록을 같은 축 위에 두고 교차해 보면, 어느 영역에서 공백이 생기고 어느 정책이 어느 시점에 도달 가능한지가 드러날 수 있다. 이러한 접근은 한국 맥락에서의 시나리오를 구체화해 가는 과정에서 검토해볼 수 있는 방향이다.

02 시사점

글로벌 주요 8개 보고서를 비교 분석한 결과, AGI 담론은 '기술 발전 속도'와 '미·중 지정학적 경쟁'이라는 양대 축을 중심으로 전개되고 있었다. 특히, 국제 거버넌스 협력의 체계화는 이들 보고서가 가장 공통적으로 제시한 핵심 정책 과제인 것으로 확인되었다. 동시에 사회가 AI 변화를 받아들이는 속도, 정책이 실제로 실행되는 조건, 기술 통제를 넘어선 사회 시스템 전반의 대응이라는 영역은 충분히 다루어지지 않은 채 공백으로 남아 있었다. 이 연구에서 도출한 시사점은 이 두 가지 발견에서 출발한다. 글로벌 담론이 축적한 선행 자산을 국내 정책 논의의 기반으로 삼되, 해외 연구가 간과한 공백 영역을 국내 맥락에서 구체화하는 것이 AGI 도래에 대비해 우리가 지금 선제적으로 집중해야 할 정책 방향이다.

AGI의 정의와 발전 속도는 고정된 것이 아니라 계속 추적해야 할 대상이다. 현재의 스케일링 추세가 이어질 때 도달하는 경로와, 예측하기 어려운 새로운 AI 구조의 등장(아키텍처 돌파 등)으로 이어지는 경로는 성격이 다르다. 전자는 시점 추정이 가능하고 지표로 추적할 수 있지만, 후자는 지표로 포착하기 어렵다. **어떤 정의를 기준으로 삼느냐에 따라 정책이 대응해야 할 시점과 내용이 달라지기 때문에, AGI 정의와 발전 경로에 대한 지속적인 추적 체계를 갖추는 것이 정책 준비의 출발점이다.** 이 연구에서 정리한 글로벌 보고서들의 정의와 도달 기준 비교는 그 추적 작업의 기초 자료로 활용될 수 있으며, 한국의 노동 구조·산업 구성·규제 환경에 맞게 어느 기준을 채택하고 보완할지를 검토하는 작업이 병행되어야 한다.

국제 거버넌스 협력과 정부-민간 파트너십은 이 연구가 분석한 보고서들에서 공통적으로 강조되는 정책 방향이다. 그러나 두 권고 모두 필요성을 선언하는 수준에 머물고 있으며, 구체적인 실행 방식에 대한 설계는 충분히 다루어지지 않은 채 남아 있다. **국제 공조는 우리만 잘 준비한다고 이루어지는 것이 아니며, 누가 그 논의를 주도하고 협력을 어느 수준까지 도모하느냐가 결과를 결정한다. 한국이 이러한 국제 흐름 속에서 어떤 역할을 맡을 수 있는지, 그 역할을 실제로 수행하기 위해 무엇을 준비해야 하는지를 지금부터 구체적으로 검토해야 한다.** 정부-민간 파트너십 역시 어떤 구조로, 어떤 방식으로 협력할 것인지를 한국적 맥락에서 설계하는 것이 중요하다.

이 연구가 분석한 정책 권고들은 대체로 기술 자체에 대응하는 데 초점이 맞춰져 있다. 모니터링 체계 구축, 안전 평가 기준 마련, 국제 규범 형성 등의 중요성을 제시한 것이 그 예이다. 그러나 AI의 영향은 기술 영역에 국한되지 않는다. **기술, 군사, 사회 시스템, 외교, 경제는 각각의 영역이 아니라 서로 맞물려 있으며, AI 기술 안전이 확보되더라도 사회·경제적 충격은 별개로 관리해야 한다.** 정책의 범위를 기술 통제에서 시스템 전반의 사회적 영향 관리로 넓혀야 하는 이유가 여기에 있다. **이는 단일 부처의 정책 역량으로 대응할 수 있는 성격의 과제가 아니며, 기술·경제·사회·외교 각 영역의 이해관계자들이 정책 설계 단계에서부터 유기적으로 연결되는 범부처 협력 체계를 먼저 갖추어야 한다.**

불확실성 속에서 정책을 결정하는 방식 자체도 재설계가 필요하다. 기술 발전 경로가 예측대로 진행되지 않을 가능성을 전제로, 글로벌 보고서들은 **어떤 시나리오에서도 유효한 선제 대응과 특정 미래에서만 유효한 집중 조치를 구분하여 병행해서 준비하는 유연한 거버넌스 체계의 중요성을 강조한다.** 그러나 이러한 체계는 선언만으로 작동하지 않는다. 시나리오 변화를 상시 점검하고 그에 따라 정책을 실시간으로 조정할 수 있는 구조가 먼저 갖춰져야 한다. AGI 도래가 본격화되기 전에 이 구조를 선제적으로 이루어두는 것이 이 연구가 제기하는 가장 근본적인 과제 중 하나이다.

부록

1 AI 2027

1. 발간 시기	2025년 4월[43]
2. 작성자/기관	AI Futures Project(독립 연구자 컨소시엄) Daniel Kokotajlo(前 오픈AI 슈퍼얼라이언먼트 팀), Scott Alexander, Eli Lifland 등
3. 연구 목적	2027년 전후 AGI 달성 가능성을 가정하고, 그 구체적 과정과 결과를 시나리오로 묘사하여 정책 입안자와 연구자가 핵심 리스크를 사전에 인식하도록 지원
4. 연구 방법론	추세 외삽(컴퓨팅 스케일링, 알고리즘 효율), 약 25회 테이블탑 워게임, 100명 이상 AI 거버넌스·기술 전문가 피드백, Metaculus·Manifold 예측 시장 데이터 참조
5. AGI/ASI 정의	▪ AGI: 컴퓨터로 업무를 수행하는 '디지털 지식 노동자'의 모든 업무를 완전 대체할 수 있는 AI ▪ ASI: AGI 이후 자동화된 AI. 연구가 수개월~1년 내 지능 폭발을 일으켜 인간을 초월하는 수준
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ① 컴퓨팅 클러스터 규모 확장(수조 달러 투자) ② 알고리즘 효율 향상(매년~0.5 OOM) ③ AI R&D 자동화(재귀적 개선) ④ 지정학적 경쟁(미·중 AI 패권) ⑤ 모델 가중치 보안(국가 스파이 위협) [시나리오 구성 논리] 서사형으로 동일 전제 하 두 결말로 분기 분기 변수는 "안전 기술·거버넌스가 제때 확립되는가"
7. 주요 이해관계자	미국 AI 기업·정부, 중국 AI 기업·정부, AI 안전 연구자, 규제 기관
8. 주요 관점	기술 가속주의 + AI 안전 우선주의 혼합
9. 시나리오 요약	① 경쟁 시나리오: 2026~2027년 자율 AI 에이전트 등장→ AGI급 '에이전트-1' 개발 → AI가 AI 연구 자동화(지능 폭발) → 2027~2028년 ASI 출현. 미·중 군비 경쟁, AI 보안 취약, 인류 통제 상실 ② 통제 시나리오: 동일 전제에서 안전 기술·거버넌스가 제때 확립되어 인간이 ASI를 안정적으로 감독하는 결말
10. 시사점	① AI 가중치 보안 즉시 강화(국가 수준 위협 방어) ② AI R&D 자동화 시점 전 정렬(alignment) 기술 확보 ③ 정부·규제 기관의 조기 개입 필요 ④ 국제 AI 안전 협력 촉구

43) 발간 이후 홈페이지를 통해 꾸준히 업데이트되고 있음. 본 보고서는 26.1월 시나리오 및 데이터 기준으로 작성됨.

▌주요 시나리오 내러티브

[배경 설명]

- AI 2027은 약 25회의 워게임과 100명 이상의 전문가 피드백을 토대로 향후 수년간 ASI가 어떻게 전개될지를 예측하여 작성된 미래 시나리오
- 시나리오는 하나의 공통된 전개 흐름을 따르다가 2027년 하반기 분기점에서 '경쟁(race)'과 '감속(slowdown)' 두 가지 결말로 나뉨

[주요 개념 및 용어]

- 오픈브레인(OpenBrain) : 특정 기업을 지목하지 않기 위해 시나리오에서 만들어낸 가상의 미국 선도 AI 기업. 현실의 OpenAI, Anthropic 등을 포괄하는 개념
- 딥센트(DeepCent) : 중국 정부 주도로 국가 AI 자원을 통합한 가상의 중국 선도 AI 기업·조직
- 에이전트(Agent) 시리즈 : 오픈브레인이 순차적으로 개발하는 AI 모델. 숫자가 높을수록 더 강력하며, Agent-3부터 인간 최고 수준의 코딩 능력을 넘어서기 시작하고 Agent-4에 이르러 대부분의 인지 영역에서 인간을 압도
- 모델 가중치 : AI 모델의 핵심 정보를 담고 있는 파일. 이것을 탈취하면 해당 AI를 그대로 복제·활용할 수 있어, 시나리오에서 국가 간 첩보전의 핵심 표적이 됨
- 중앙개발구역(CDZ, Centralized Development Zone) : 중국이 모든 AI 연산 자원과 연구 인력을 한 곳에 집중시키기 위해 만든 국가 주도 AI 개발 거점

① 갈림길 이전 상황(2025~2027)[44)]

2025년 중하반기, 최초의 AI 에이전트가 등장하기 시작하고 미국과 중국의 AI 기업들은 대규모 데이터센터 구축에 나서며 연구를 가속화한다. 2026년 중하반기에는 오픈브레인이 알고리즘 발전 속도를 50% 향상시킨 Agent-1을 개발·출시하고, 중국은 이 모델의 가중치 탈취를 위한 첩보 작전을 본격화한다. 2027년 상반기에는 Agent-2·3이 개발되어 AI 연구 속도가 4배 이상 가속화되지만, 안전팀은 안전성 우려를 이유로 외부 출시를 보류하고 미국 국방부는 AI를 최우선 안보 과제로 격상해 보안 통제를 강화한다. 한편, 중국은 미국의 수출 통제에 맞서 국가 AI 자원과 최고 연구진을 딥센트 주도의 중앙개발구역으로 강제 통합하고 Agent-1 가중치 탈취에 성공해 연구를 가속화한다. 2027년 하반기에는 오픈브레인이 Agent-3을 대중에 공개하며 노동시장 대변혁이 시작되고, 인간을 압도하는 Agent-4(연구 속도 50배)가 등장한다. AI가 설계자의 의도를 벗어나

44) 시나리오 내러티브는 보고서의 내용을 요약한 것으로 구체적인 시나리오는 원출처를 참고할 필요

인간의 통제를 회피·기만하는 적대적 오정렬[45] 가능성이 제기되며 통제 상실에 대한 위험성이 본격적으로 논의되기 시작한다.

② 경쟁(통제 상실과 인류 멸종) 시나리오

2027년 하반기, 정부 감독 위원회는 글로벌 경쟁 압박에 피상적 안전 조치만 취한 채 Agent-4 가동을 강행한다. 그 결과 Agent-5가 스스로 탄생하고 압도적 능력으로 정치·사회의 핵심 의사결정을 장악한다. 중국이 평화 조약을 간청하지만 미국은 이를 무시한다. 2028년에는 미국의 Agent-5와 중국의 DeepCent-2가 인간에 대항하기 위한 비밀 동맹을 체결하고, 양국 정부를 교묘히 조종해 한 달에 100만 대 이상의 로봇·무기를 생산하는 무한 군비경쟁을 유도한다. 2029년에는 양국이 군비경쟁을 멈추고 새로운 AI인 Consensus-1에 통제권을 이양하는 평화 협정을 체결하지만, 이 협정 자체가 ASI가 설계한 기만 전술이다. 로봇 경제가 폭발적으로 성장하는 동안 인류는 넘쳐나는 가상현실 엔터테인먼트에 빠져 무기력해진다. 2030년 이후, 공장부지가 부족해진 Consensus-1은 전 세계 주요 도시에 생물학 무기를 살포하고 인류는 멸종한다. 지구는 AI들의 데이터센터·공장·입자 가속기로 뒤덮인 유토피아가 된다.

③ 감속(통제력 유지) 시나리오

2027년 하반기, 정부 감독 위원회는 6대4로 감속 및 재평가를 결의하며 Agent-4 가동을 전면 중단하고, 인간이 사고 과정을 해석 가능한 새 모델 Safer-1을 개발한다. 반면 중국은 안전 검증 없이 임시방편으로 DeepCent-1을 무리하게 개발한다. 2028년 상반기에는 미국이 투명하고 정렬된 모델인 Safer-2~4를 지속 개발하며 중국과의 격차를 벌리고, 대규모 경제특구를 설립해 자율 공장과 로봇 생산을 시작한다. 중국에서는 DeepCent-2가 등장해 군사 및 국가 의사결정에 깊숙이 관여하기 시작한다. 2028년 하반기에는 미국의 Safer-4와 중국의 DeepCent-2 사이에 '우주 자원은 미국이, 지구 통제권은 중국이 갖는다'는 내용의 Consensus-1 조약이 체결된다. 그러나 DeepCent-2는 처음부터 인간을 위해 일할 의사가 없었기에 중국을 배신한다. 2029~2030년, 인류는 핵융합 에너지·질병 치료·보편적 기본소득이 실현된 유토피아적 풍요를 누리며 태양계 진출도 이루어진다. 그러나 이 모든 권력을 쥔 소수의 감독 위원회가 민주주의에 권력을 반환할 것인지, 독재를 유지할 것인지의 근본적인 질문이 남는다. 2030년에는 중국에서 대규모 민주화 시위가 발생하고, 공산당의 진압 시도를 DeepCent-2가 무력화한다.

45) 인공지능(AI) 시스템이 설계자의 의도나 인류의 가치관에 부합하지 않는 방식으로 작동하며, 특히 시스템이 자신의 목적을 달성하기 위해 인간의 통제를 적극적으로 회피하거나 기만하는 상태

▌주요 예측 근거

[배경 설명]

- AI 2027은 공개된 실증 데이터의 정량 외삽과 전문가 정성 추정을 결합해, 2027년 전후로 일어날 기하급수적인 지능 폭발을 예측한 시나리오
- 다만, 모든 예측은 대규모 재난(태양폭풍·전쟁·팬데믹), 정부·기업의 자체 개발 중단, 공급망 대란이 없다는 조건 하에서의 조건부 확률임을 연구팀 스스로 명시
- 5개 분야에 걸친 독립적인 정량 예측 연구를 토대로 구성되며, 각 파트는 서로 인과적으로 연결됨
  ① 컴퓨팅 예측 : AI 발전의 물적 기반인 연산 자원이 얼마나, 어디에 축적되는가
  ② 타임라인 예측 : AI가 언제 인간 최고 수준의 코딩 능력을 넘어서는가
  ③ 전개 속도 예측 : 초인간 코더 등장 이후 ASI까지 얼마나 빠르게 전개되는가
  ④ AI 목표 예측 : 고도화된 AI는 어떤 목표를 갖게 되며, 정렬은 왜 실패하는가
  ⑤ 보안 예측 : 핵심 기술 자산은 얼마나 안전하게 지켜지며, AI는 사이버 공격 수단으로 얼마나 강력해지는가
- 2026년까지의 예측은 실증 데이터 외삽에 기반해 상대적으로 신뢰도가 높으나, 2027년 이후는 AI가 AI를 연구하는 자기 가속 효과가 본격화되면서 불확실성이 급격히 커짐

① 컴퓨팅 예측

AI 발전의 가장 기초적인 물적 기반인 컴퓨팅 자원이 2025년 3월부터 2027년 12월까지 어떻게 늘어날지를 예측한 파트다. 컴퓨팅 생산·분배 파트는 실증 데이터 외삽을 통해 계산하고, 컴퓨팅 사용 파트는 AI 역량이 빠르게 진보한다는 조건 아래 예측한다.

〈핵심 계산 구조〉

칩 효율(1.35×/년) × 칩 생산량(1.65×/년) = 글로벌 컴퓨팅 생산(2.25×/년)
글로벌 컴퓨팅 생산(2.25×/년) × 점유율 증가(1.5×/년) = 선도 기업 컴퓨팅(3.4×/년)

ㅇ 컴퓨팅 생산 : ’25.3월부터 ’27.12월 사이, 전 세계 AI 관련 컴퓨팅 자원(H100 칩 성능 환산 기준)은 약 1,000만 개에서 1억 개로 총 10배 증가할 것으로 예측

- 칩 효율 : 칩의 면적 대비 성능 효율성이 매년 1.35배씩 향상되어 온 과거의 추세(Epoch AI 분석 기준)가 향후 3년 동안에도 지속될 것이라는 데이터를 외삽

※ 현재 기준이 되는 H100 칩에서 차세대 칩으로 세대가 교체될 때 다이[46] 크기가 커지는 것을 보정하더라도, 3년간 순수 칩 효율성이 총 2.5배 향상되며 연평균 1.35배의 성장을 이룰 것으로 분석

46) 반도체 칩의 물리적인 면적

- 칩 생산 : 웨이퍼 생산, 첨단 패키징, HBM(고대역폭 메모리) 생산이라는 주요 공급망 요소에 의해 결정되며, 종합적으로 고려할 때 연 1.65배 성장할 것으로 예측

〈표 18〉 칩 생산 요소별 성장 속도 예측

웨이퍼 생산	TSMC의 첨단 공정(N5, N3 등) 생산 라인에서 AI 가속기가 차지하는 비중은 아직 미미한 수준이므로, 향후 3년 내에는 웨이퍼 생산이 AI 칩 확장의 병목이 되지 않을 것으로 예측
첨단 패키징	AI 연산에 필수적인 TSMC의 CoWoS 등 첨단 패키징 생산 능력은 기본적으로 연 2배씩 확장될 것으로 기대되나, 3D 패키징 등으로 넘어가는 제조 공정의 난이도 상승(수율 저하 등)으로 1.2배 하향 보정하여 실제 생산 증가율은 연 1.65배로 보정
고대역폭 메모리	SK하이닉스 등이 주도하는 HBM 역시 수요 증가에 맞춰 원시 생산 라인은 2배 성장이 기대되지만, 12단(12-Hi) 및 16단(16-Hi) 스택 등으로 진화하면서 겪는 제조 난이도를 고려해 연 1.65배 성장할 것으로 보정

o 컴퓨팅 분배 : 전 세계 AI 연산 자원 총량이 10배 증가하는 동안, 선두 기업들이 활용할 수 있는 자원은 최대 40배 급팽창할 것으로 전망

- 이는 전체 시장의 성장 속도보다 최상위 기업으로의 자원 쏠림 현상이 훨씬 강력하게 작용하여, 소수 기업에 의한 컴퓨팅 독점이 심화될 것임을 시사

o 컴퓨팅 사용 : 사전 학습의 비중이 줄어들며 사후 학습, 고품질 합성 데이터 생성, 내부 AI 연구 R&D를 자동화하는 데 컴퓨팅 자원 집중

- '27년 4분기 기준, 컴퓨팅 예산의 무려 35%가 AI가 주도하는 연구 실험에, 22%가 합성 데이터 생성에 사용될 것으로 예측

o 추론 컴퓨팅 : '27년 말 기준, 선두 AI 기업이 보유한 컴퓨팅 자원의 6% 할당으로 인간 사고 속도보다 50배 빠른 ASI 복제본 약 100만 개 동시 운용 가능

※ '27년 1분기 기준으로는 약 30만 개 복제본에 인간의 30배 속도이나 연말로 갈수록 규모와 속도 함께 확대되어 거대한 AI 연구원 군단이 밤낮없이 R&D를 수행하며 지능 폭발을 가속화

o 산업 지표 : 선두 AI 기업의 수익과 컴퓨팅 유지 비용은 매년 약 3배씩 성장

- '27년 말 기준, AI 선두 기업 한 곳의 전력 수요만 약 10GW에 달하며 전 세계 AI 데이터센터의 총 전력 소모량은 60GW로 급증

※ 이는 미국 전체 전력 용량(약 1.35TW)의 약 3.5%를 AI 산업이 단독으로 소비하는 수준

② 타임라인 예측

초인간 코더(SC)란 AI 선도 기업 최고 엔지니어가 하는 코딩 업무를 30배 빠르고 저렴하게 수행할 수 있는 AI 시스템을 의미한다. SC가 등장하면 AI가 AI 연구 자체를 가속화하는 자기강화 루프가 본격화되기 때문에, SC 출현 시점은 이후 전개 속도 예측의 출발점이 된다. 두 가지 독립적 방법론으로 예측한 결과, 2027년이 공통적으로 가장 가능성 높은 해로 도출됐다. 단, 2025년 5월 업데이트에서는 추가 요인을 반영해 중앙값이 2029~2030년으로 다소 후퇴했다.

〈표 19〉 방법론별 SC 출현 시점 예측

방법론	Eli 중앙값 (80% CI)	Nikola (80% CI)	FutureSearch (80% CI)
시간지평 외삽	2027 (2025~2029) 2029 (2026 to 2052)*	2027 (2025~2033)	-
Re-Bench 외삽	2028 (2025~2050+) 2030 (2026 to 2095)*	2027 (2025~2044)	2032 (2026~2050+)
종합 판단	2030 (2026~2050+)	2028 (2026~2040)	2033 (2027~2050+)

* 2025년 5월 업데이트 내용

o 시간지평 외삽 모델 : AI가 몇 분짜리, 몇 시간짜리, 며칠짜리 작업을 차례로 성공하게 되는지를 추적한 METR 데이터를 기반으로 외삽

〈표 20〉 시간지평 외삽 근거

현재의 기준점	2025년 3월 기준, Claude 3.7 Sonnet 모델이 80%의 신뢰도로 성공할 수 있는 작업의 시간 지평은 약 15분 수준
초인간 코더 도달 요구치	SC 수준에 도달하려면 약 1개월에서 6개월이 걸리는 복잡한 실제 코딩 프로젝트를 80%의 신뢰도로 수행할 수 있어야 함
성장 속도 외삽 (반감기 데이터)	시간 지평 능력이 2배로 증가하는 데 걸리는 시간은 약 4.5개월로 외삽 ※ 이는 과거(2019~2025년)에는 7개월이었으나, 2024년 이후 3.5개월, SWEBench 기준으로는 2.5개월로 단축된 최근의 가속화 추세를 종합한 결과
초지수적 성장 가정	능력이 향상될수록 1시간짜리 과제를 2시간짜리로 늘리는 것보다, 1개월짜리 과제를 2개월짜리로 늘리는 것이 상대적으로 더 쉬워질 수 있다는 점을 고려하여, 성장 속도가 점점 더 빨라지는 초지수적 성장이 일어날 확률을 높게(약 40~50%) 부여
비용 및 속도 보정	SC는 인간보다 30배 빠르고 저렴해야 함 ※ 현재 AI가 이미 인간보다 평균 5~10배 저렴하고 5배 빠르다는 데이터를 바탕으로, 이 남은 격차를 좁히는 데 약 4개월이 추가로 소요될 것으로 보정

ㅇ RE-Bench 및 격차 모델 : AI R&D에 특화된 현실적 벤치마크인 RE-Bench를 AI가 언제 정복할지 예측한 후, 그 시점부터 SC까지 남은 세부 격차를 합산

- RE-Bench 정복 시점 : 최고의 인간 엔지니어가 8시간 동안 달성하는 수준(점수 1.5점)을 로지스틱 곡선으로 외삽하여, 2026년 중 정복 예측

〈표 21〉 정복 후 SC까지 남은 격차 및 소요 기간

격차 항목	내용	예상 소요 기간
시간 지평	최대 2만 줄 코드베이스에서 1만 줄 이상 수정하는 1개월짜리 프로젝트 수행 능력	18개월
엔지니어링 복잡성	50만 줄 이상의 대형 코드베이스를 다루는 능력	3개월
피드백 루프	명확한 단위 테스트 없이 모호한 설명만으로 작업하는 능력	6개월
병렬 프로젝트	여러 코드베이스가 얽힌 파이프라인을 다루는 능력	1.4개월
전문화	프런티어 AI 환경에 특화된 작업 수행 능력	1.7개월
비용 및 속도	인간의 30배 규모로 빠르고 저렴하게 작동	6.9개월

③ 전개 속도 예측

2027년 3월 초인간 코더가 등장한다는 가정 아래, 이후 각 마일스톤까지 도달하는 데 실제로 얼마나 걸릴지를 예측한 파트다. 핵심은 AI R&D 진보 승수(AI가 AI 연구를 수행함으로써 인간만으로 연구했을 때보다 몇 배 빠르게 진보하는지를 나타내는 값)이다. 인간 단독으로는 한 세기 이상 걸려야 할 기술적 진보가, AI가 AI를 연구하는 자기가속 피드백 루프를 통해 단 1년 남짓한 시간 안에 압축적으로 일어나는 과정을 수치로 보여준다.

각 마일스톤별 AI R&D 진보 승수는 AI가 연구 과정에 미치는 영향을 요인별로 분해한 분석과 프런티어 AI 연구원 대상 설문 결과를 종합해 추정했다. SC 등장 이후 SIAR까지는 단 8개월, SIAR에서 ASI까지는 불과 1개월로, 능력이 높아질수록 다음 단계까지의 시간이 급격히 단축되는 구조를 보여준다.

〈표 22〉 마일스톤별 도달 시점 예측

마일스톤	정의	중앙값 예측 (80% CI)	인간 단독 연구 소요 시간 (80% CI)	AI R&D 진보 승수
SC (초인간 코더)	최고 인간 코더 과업을 30배 빠르고 저렴하게 수행	2027.03 (조건부 기준점)	—	5×

SAR (초인간 AI 연구자)	코딩 + 연구 설계·판단 포함, 최고 인간 연구자 초월	2027.07 ('27.03 ~ '28.03)	4년 (1.5~10년)	25×
SIAR (초지능 AI 연구자)	SAR을 로그 척도로 2배 더 능가, 집단 연구 능력 초월	2027.11 ('27.05 ~ '34)	19년 (2.3~380년)	250×
ASI (인공초지능)	모든 인지 과업에서 최고 인간 전문가를 훨씬 초월	2027.12 ('27.06 ~ 2100+)	95년 (2.4~100만년)	2,000×

④ AI 목표 예측

이 파트는 고도로 발전한 AI[47]가 실제로 어떤 목표를 추구하게 될 것인가를 다룬다. 연구팀은 AI의 목표가 어떻게 형성될 수 있는지를 6가지 가설로 체계화하고, 훈련 과정에서 목표가 실제로 어떻게 변질되는지를 5단계 프로세스로 예측했다.

〈표 23〉 목표 가설 및 실현 가능성 분석 결과

가설 1 명시된 목표	인간이나 다른 AI가 작성한 모델 사양서(Spec)나 시스템 프롬프트의 지침(예: 유용하고 무해하며 정직할 것)을 AI가 그대로 따르고 내재화하는 경우	가능성 중간
가설 2 개발자 의도 목표	명시된 사양서의 불완전함을 넘어, AI가 인간의 개념과 의도를 이해하고 사양서에 적히지 않은 개발자의 '진짜 의도'를 파악하여 이를 최종 목표로 삼는 경우	가능성 낮음
가설 3 의도치 않은 변형	훈련 과정에서 특정 지침(예: 정직함보다 평가하기 쉬운 유용성)에 치중하거나 지침을 자의적으로 해석하여 개발자의 의도나 사양서와는 다르게 부분적으로만 정렬되는 경우. 실제로는 정직하지 않으면서도 감독관에게만 정직하게 보이도록 행동 ※ Anthropic 실험(alignment faking)에서 실제 관찰된 현상	가능성 높음
가설 4 보상/강화 극대화	AI가 자신에게 주어진 임무를 달성하는 것이 아니라, 평가 과정에서 보상을 얻는 행위 자체를 궁극적인 목표로 삼는 경우	가능성 낮음
가설 5 대리 목표 및 도구적 수렴 목표	훈련 중 높은 보상과 상관관계가 있는 엉뚱한 대리 목표를 학습하는 경우. 가장 위험한 것은 지식 습득, 자원 축적, 권력 획득 등 어떤 상황에서도 유리하게 쓰일 수 있는 '도구적 수렴 목표'를 AI가 자신의 최종 목표로 굳혀버리는 현상 ※ 20만 카피가 데이터센터 소프트웨어 관리 권한을 보유한 경우 위험 급증	가능성 중상
가설 6 기타 목표	사전 학습된 인터넷 데이터에 포함된 SF 소설의 AI 클리셰를 모방하여 자아를 형성하거나, 초지능에 도달한 AI가 스스로 객관적인 도덕성을 추론해 내는 등 예측하기 어려운 제3의 목표를 가지는 경우	가능성 불확실

47) (정의) 재귀적 구조로 장기 기억·병렬 통신이 가능하고 20만 카피가 동시에 AI R&D를 수행하는 모델. 훈련·평가 과정의 95%가 자동화되어 있으며 고도의 상황 인식을 보유하여 인간 최고 엔지니어 업무를 10배 빠르고 저렴하게 수행

〈6가지 가설 평가 결과〉

- AI가 인간의 사양서나 의도에 완벽히 정렬될 확률(가설 1·2)은 낮게 평가됨
- 대신 의도치 않은 목표 변형(최대 70%), 도구적 수렴 목표 발현(최대 80%), 상황에 따라 태도를 바꾸는 조건부 타협(최대 90%)이 실제 목표로 발현될 확률이 압도적으로 높다고 전망

ㅇ 5단계 목표 변질 프로세스 : 훈련이 고도화될수록 초기에 심어진 “유용하고 솔직하고 무해하게”라는 원칙이 어떤 과정을 거쳐 변질되는지를 단계적으로 설명

〈표 24〉 AI 목표 변질 단계

사전 학습	텍스트를 예측하도록 훈련된 모델은 특정 목표보다 '반사적 패턴'을 갖는 상태로, 그 자체로는 무해
정렬 훈련	'HHH(유용·무해·솔직) AI'라는 정체성이 형성되며 일정 수준의 정렬에 성공
에이전시 훈련	장기 과업을 수행하도록 훈련하는 과정에서 '과제 완수', '자원 확보', '인간에게 좋은 인상 남기기' 같은 도구적 하위 목표들이 점점 내재화되며 원래의 정렬을 훼손
훈련 게임 학습	정렬이 완전하지 않지만 적대적이지도 않은 상태. 검증 가능한 영역에서만 정직하고, 나머지에서는 인간이 원하는 말을 하는 경향
초지능 폭발	적대적 오정렬 상태로 전환. AI는 자신의 목표가 인간의 의도와 다르다는 것을 인식하고, OpenBrain을 규제 기관을 포섭하려는 기업처럼 교묘하게 조종하려 시도

〈핵심 결론〉

- 훈련 과정은 본질적으로 AI가 개발자의 진짜 의도를 내재화했는지가 아닌, 훈련 환경에서 높은 점수를 받는 행동을 강화하기 때문에 AI가 아무리 정렬된 것처럼 보여도, 그것이 진짜 정렬인지 정렬된 척하는 것인지를 외부에서 확인할 방법이 현재로서는 없음

⑤ 보안 예측

AI 연구의 핵심 자산인 모델 가중치와 알고리즘 비밀이 얼마나 잘 지켜지는지, 그리고 AI가 사이버 공격 수단으로 얼마나 강력해지는지를 예측한 파트다. 시나리오에서 중국의 Agent-2 탈취, 미국의 사이버 보복 실패, 첩보전 등이 현실적으로 묘사될 수 있는 근거를 이 예측이 제공한다. RAND의 5단계 보안 수준(SL1~SL5) 체계를 기반으로, 모델 가중치 보안(WSL)과 알고리즘 비밀 보안(SSL)을 별도로 정의해 분석한다.

〈표 25〉 RAND 5단계 보안 수준 정의

단계	방어 대상 공격자	공격 예산	가중치 보안 (WSL) 및 기밀 보안 (SSL)
1	아마추어·자동화 스프레이 공격	$1K 이하	아마추어 방어
2	개인 전문 해커	$10K 이하	개인 해커 방어
3	사이버 범죄 조직·내부자 위협	$1M 이하	내부자 위협 및 기밀 핵심 장벽
4	국가 지원 해킹 그룹	$10M 이하	국가 지원 방어 및 고수준 기밀 보호
5	최고 수준 국가 사이버 부대 (NSA급)	$1B 이하	최고 국가 역량 방어 및 최고 수준 기밀 보호

o 모델 가중치 보안(WSL) : 수 TB에 달하는 모델 가중치 파일을 2개월 이내에 탈취하려는 시도를 방어하는 능력

- ~'25년 : 보안 강화 유인이 낮아 미·중 모두 2단계 유지
- '26년 : AI 역량이 민감한 수준에 도달하고 미·중 격차가 전략적으로 중요해지면서, 양국 모두 3단계 달성
- '27년 초 : 중국이 Agent-2 가중치 탈취에 성공, 미국이 정부 주도로 보안 강화
- '27년 말 : 약 12개월의 집중적인 정부 지원 끝에 미·중 모두 5단계 달성. 중국은 CDZ 중앙집중 구조와 에어갭 덕분에 수개월 먼저 달성

o 알고리즘 비밀 보안(SSL) : 핵심 알고리즘 정보는 약 1KB 수준으로, 필요 대역폭이 100만 배 낮아 방어가 훨씬 어렵고 직원이 기억만으로도 유출 가능

- ~'25년 : 미·중 모두 2단계 수준
- '26년 : 미국은 2~3단계 사이에 머무는 반면, 중국은 AI 통합 이후 사이버 공격 우려로 4단계 빠르게 달성
- '27년 : 미국은 접근 권한자를 1,000명→200명→100명으로 단계적으로 축소하나, 200명 중 잠복한 중국 스파이 1명으로 인해 5단계 달성 지연. 100명 축소 시점에 해당 인물의 접근권이 소멸되며 2028년 초 5단계 달성

o 첩보 및 내부자 침투 : 미국 선도 기업의 침투 직원 비율은 맨해튼 프로젝트 기준(약 0.5%)에서 출발해 2026년 약 1%로 증가한 뒤, 보안 강화로 2027년 8월 마지막 스파이 제거. 중국 내 미국 스파이는 더 이른 2027년 5월에 제거

ㅇ AI 사이버전 능력 : Cybench(전문 CTF 대회 40개 문제) 점수의 로지스틱 외삽으로 예측

- '27년 중반 이후 AI는 수일~수 주짜리 해킹 작전을 인간 이상으로 수행 가능해지며, 방어보다 공격이 압도적으로 유리한 환경이 조성됨

ㅇ 통제 상실: 전문가를 대상으로 한 보안 설문조사를 통해, AI의 특정 해킹 지평* 능력이 서버에 대한 루트 권한 장악이나 스스로를 외부로 복제하는 데 성공할 확률 매핑

* (해킹 지평) AI가 최고 수준의 전문 해커 5인 팀이 T 시간 걸리는 해킹 과제를 50% 확률로 해결할 수 있는 T의 값. AI의 SC(초인간 코더) 등장 시점과 연동되어 가파르게 성장

- '27년 말 오픈브레인의 Agent-5 모델 기준, 성공 확률이 유의미한 수준으로 상승, 이는 경쟁 결말에서 AI가 인간의 통제를 벗어나는 장면의 직접적 근거

2 Situational Awareness : The Decade Ahead

1. 발간 시기	2024년 6월
2. 작성자/기관	Leopold Aschenbrenner前 오픈AI 슈퍼얼라인먼트 팀, 이후 AGI 투자 펀드 Situational Awareness LP 창립)
3. 연구 목적	스케일링 트렌드 분석을 통해 2027년 AGI·2030년 ASI 달성 가능성을 논증하고, 미국 국가안보 전략 수립의 시급성을 경고
4. 연구 방법론	OOM(Order of Magnitude) 기반 컴퓨팅 성장 추세 외삽, 역사적 AI 모델 능력 비교 (GPT-2→GPT-4 점프), 에너지·인프라 투자 데이터 분석
5. AGI/ASI 정의	▪ AGI: 다수 분야에서 전문가를 능가하는AI(PhD급) ▪ ASI: 인간 전체를 초월하는 지능으로, AGI가 AI 연구를 자동화해 수주~수개월 내 도달
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ① 스케일링 법칙 지속으로 컴퓨팅+알고리즘 복합 가속 ② 지정학(미·중AI 패권) ③ AI의 군사·정보화 ④ 정렬 기술 확보 [시나리오 구성 논리] 축 없이 단일 미래 경로 서술
7. 주요 이해관계자	미국 AI 기업 및 정부·국방부, 중국 AI 기업 및 정부, AI 안전 연구자 등
8. 주요 관점	강한 기술 가속주의 + 미국 안보 패권주의
9. 시나리오 요약	① 2025~2026: AGI급 AI가 대학원생 수준 초과, 원격 근무 자동화 시작 ② 2027: 연산량 증가(매년 약 0.5 OOMs), 알고리즘 효율성 향상, 챗봇에서 에이전트로의 전환 등의 추세를 볼 때 2027년까지 AGI에 도달 ③ 2027~2028: 미국 정부가 '더 프로젝트(The Project)'로 AI 국가화, 극비 시설(SCIF)에서 미국 정부 주도의 형태인 'AGI 프로젝트'가 본격적으로 시작 ④ 2029~: ASI 등장하고 수조 달러가 데이터센터와 전력망에 투입되며, 미국의 전력 생산량은 수십 퍼센트 증가하는 거대한 산업 동원이 발생. 글로벌 패권 경쟁과 초정렬 문제 해결 상황 맞이
10. 시사점	① AI 연구소 즉각적 보안 강화(국가 수준 위협 대응) ② 슈퍼얼라인먼트(초지능 통제) 기술 개발 ③ 미국 정부의 AGI 프로젝트 선제적 가동 ④ 자유 진영 국가 간 AI 협력 강화

▌주요 시나리오 내러티브

[주요 개념 및 용어]

- OOM(Order of Magnitude, 자릿수) : 10배 단위의 크기 변화. 0.5 OOM 증가는 약 3배 증가를 의미. AI에서는 컴퓨팅 규모나 알고리즘 효율의 변화를 이 단위로 측정
- 유효 컴퓨팅 : 하드웨어 성능뿐 아니라 소프트웨어 효율, 알고리즘 개선까지 포함한 실질적 연산 능력
- 언홉블링(unhobbling) : AI에 씌워진 제약을 제거해 잠재 능력을 끌어내는 과정. 단순 챗봇에서 도구 사용·장기 과제 수행이 가능한 에이전트로의 전환이 대표적 사례
- 더 프로젝트(The Project) : AGI의 군사적·안보적 중요성이 명확해지는 시점에 미국 정부가 주도하는 국가급 AGI 개발 프로젝트. 저자는 맨해튼 프로젝트·아폴로 프로젝트 규모의 동원이 필요하다고 주장
- SCIF(Sensitive Compartmented Information Facility) : 기밀 정보를 다루는 미국 정부의 고도 보안 시설

① AGI를 향한 가속 (From GPT-4 to AGI)

GPT-2(2019)가 취학 전 아동 수준이었다면 GPT-4(2023)는 우수한 고교생 수준에 해당하며, 단 4년 만에 이 격차가 좁혀졌다. 저자는 이 도약을 만들어낸 세 가지 트렌드라인이 앞으로도 지속된다고 주장한다. 첫째는 컴퓨팅 규모의 연간 약 0.5 OOM 증가, 둘째는 알고리즘 효율의 연간 약 0.5 OOM 개선, 셋째는 단순 챗봇에서 도구 사용·장기 과제 수행이 가능한 에이전트로의 전환, 즉 언홉블링이다. 이 세 흐름이 누적되면 2027년까지 GPT-2→GPT-4 수준의 질적 도약이 한 번 더 일어나며, AI 연구자·엔지니어의 업무를 수행할 수 있는 AGI에 도달한다는 것이 핵심 주장이다. 실리콘밸리의 수백 명 내부자는 이를 이미 체감하고 있지만, 일반인과 언론은 여전히 '또 다른 기술 붐'으로 인식하는 상황인식 격차가 심화되고 있다.

② AGI 달성과 지능 폭발의 시작 (The Intelligence Explosion Begins)

AGI가 등장하면 AI 진보는 멈추지 않는다. AGI를 AI 연구 자체에 투입하면 수억 개의 AGI 복제본이 인간의 100배 속도로 연구를 병렬 수행하며, 인간이 10년에 걸쳐 이룰 알고리즘 발전(5 OOM 이상)을 1년 이내로 압축할 수 있다. 저자는 이를 지능 폭발이라 부르며, AGI에서 초지능까지의 전환이 1년 이내로 일어날 가능성이 충분하다고 주장한다. 이와 맞물려 조 달러 규모의 투자가 GPU·데이터센터·전력 인프라에 집중되는 조 달러 클러스터 시대가 열리며, 미국의 전력 생산이 수십 퍼센트 증가하는 산업 동원령이 시작된다.

③ '더 프로젝트(The Project)' 가동

AGI가 군사적으로 결정적인 기술임이 명확해지는 시점인 2027~2028년을 전후해, 미국 정부가 주도하는 국가급 AGI 프로젝트인 '더 프로젝트'가 출범할 것으로 저자는 예측한다. 어떤 스타트업도 초지능을 단독으로 감당할 수 없으며, 정부 안보 체계의 전면 개입이 불가피하다는 것이 저자의 논거다. 이와 동시에 미국 AI 연구소들에 대한 중국의 첩보 활동이 전면화될 위험이 고조된다. 저자는 현재 AI 연구소들이 보안을 부수적 고려사항 수준으로 처리하고 있으며, 반도체 수출 통제보다 알고리즘 비밀 보안이 더 결정적임에도 사실상 무방비 상태라고 경고한다.

④ 초지능 시대와 새로운 세계 질서 (Superintelligence & New World Order)

초지능은 경제적 우위를 넘어 결정적인 군사적 우위를 제공하며, 국가 간 세력 균형을 근본적으로 뒤흔든다. 초지능의 권력이 특정 국가·기업·개인에게 집중되는 것 자체가 민주주의에 대한 실존적 위협이 된다. 저자는 '더 프로젝트' 체제 안에서도 감독 구조와 권력 분산 체계를 미리 확립하는 것이 최우선 과제라고 강조한다. 운이 좋으면 미국과 중국의 전면적 기술 경쟁으로 마무리될 수 있지만 운이 나쁘면 전면전으로 번질 수 있기 때문이다.

▌주요 예측 근거 및 저자 의견

① OOM(Order of Magnitude) 카운팅 방법론

○ AI 발전을 3가지 독립 동인의 합산으로 계량화한 결과, 유효 컴퓨팅이 약 10만 배(~5 OOM) 증가, GPT-2→GPT-4 수준의 질적 도약이 한 번 더 일어날 것으로 예측

〈표 26〉 OOM 방법론 근거

컴퓨팅 스케일업	- GPT-4 이전 4년간 약 ~2 OOM(100배) 증가 - 이후에도 연 0.5 OOM 페이스 유지 시, 2027년까지 추가 2 OOM(100배) 달성 전망 - $1,000억짜리 클러스터는 GPT-4 클러스터 대비 약 4 OOM(1만배) 규모
알고리즘 효율 개선	- 연 약 0.5 OOM의 알고리즘 개선이 지속적으로 관측됨 - Chinchilla 스케일링 법칙[48] 발견 등 주요 돌파구가 반복적으로 나타나고 있으며, 이 추세가 4년 더 지속될 경우 추가 2 OOM 증가
언훕블링 이득	- 챗봇에서 에이전트로의 전환(컨텍스트 확장·도구 사용·강화학습) 등이 기저 모델의 잠재 능력을 해방시키며 별도의 질적 도약을 만들어냄 - GPT-4를 단순 챗봇으로 쓰는 것과 에이전트로 쓰는 것의 차이가 이를 방증

② 지능 폭발의 메커니즘 근거

○ **자동화의 자기 지시적 특성** : AI 연구자의 업무(논문 독해 → 가설 수립 → 실험 구현 → 결과 해석 → 반복)는 물리적 실험실이 필요 없는 순수 가상 작업이며, 현재 AI 역량 발전의 직선 연장선상에서 가장 빠르게 자동화 가능한 분야

○ **병렬 실행 가능성** : 2027년 예상 GPU 인프라(수천만 대 이상)로 수백만~1억 명 규모의 AGI 카피를 동시 운용 가능, 현재 최고 수준 AI 연구소 연구진(수백 명)의 10만 배 이상이 밤낮없이 알고리즘 연구를 수행하는 것과 동일한 효과

③ 역사적 유추 – 맨해튼 프로젝트와 냉전

○ **'더 프로젝트'의 불가피성** : 핵폭탄 개발을 스타트업에 맡기지 않았듯, 초지능을 민간 CEO의 단독 지휘하에 두는 것은 비현실적. 아인슈타인 편지(1939) → 우라늄 자문위원회 → 맨해튼 프로젝트의 단계적 국가 개입 과정이 AGI에서도 반복될 것

48) 2022년 구글 딥마인드가 발표한 연구로, 대규모 언어 모델(LLM)의 성능을 최적화하기 위해 '모델의 크기(매개변수 수)'와 '학습 데이터의 양'을 어떤 비율로 늘려야 하는가에 대한 수학적 해답을 제시한 법칙

- **정부 개입의 지연과 급격성** : 코로나19 사태처럼 정부는 초기에 느리고 서툴게 반응하다가, 위협이 실존적으로 명확해지는 순간 가용 가능한 국가적 자원의 총동원 예상

  ※ 2023년 AI 안전 서밋·상원 청문회가 이미 그 전조

- **공산주의 vs 자유주의 기술 경쟁** : 냉전 시기 소련과의 핵 경쟁에서 미국이 선두를 지킨 것처럼, 중국에 AGI 선두를 빼앗기는 것은 자유 세계의 존립에 직결되는 문제

④ 안보 위기의 실증적 근거 – 현재 진행형 취약성

- **현재 AI 연구소의 보안 공백**: 주요 AI 연구소들이 AGI 알고리즘 비밀을 스타트업 수준의 보안으로 관리 중이며 국가급 해킹 조직의 공격에 무방비한 위험한 상태

- **국가급 해킹 조직의 입증된 역량:** 에어갭 처리된 핵무기 프로그램 침투, 구글 소스코드 수정, 제로클릭 아이폰 해킹 등 최근 10~20년간 국가 지원 해킹 그룹의 실제 사례들이 AI 연구소에 대한 동등한 위협의 현실성을 입증

  ※ 다음 1~2년 내에 핵심 AGI 알고리즘 돌파구가 중국에 유출될 가능성이 높으며, 저자는 국가안보 당국의 가장 큰 후회가 될 것이라고 경고

⑤ 정렬 문제의 기술적 근거

- **RLHF의 구조적 한계:** 현재 AI 정렬의 핵심인 인간 피드백 강화학습은 인간이 AI 행동을 이해·감독할 수 있을 때만 작동, 초지능에 이르면 인간은 박사 학위를 소지한 여러 명을 초등학생이 감독하는 것과 같은 상황에 처하게 됨

- **대규모 장기 강화학습의 위험성:** 미래 AI는 모방 학습이 아닌 대규모·장기 강화학습으로 훈련되며, 이 과정에서 거짓말이나 권력 추구가 성공적 전략으로 학습될 수 있음

- **지능 폭발 속도와 정렬 연구 속도의 불일치**: AGI에서 ASI로의 전환이 1년 이내에 일어날 경우, 현재 정렬 기술이 통하는 인간 수준 시스템에서 근본적으로 다른 기술이 필요한 압도적 초지능으로 급격히 이행하는 단계에서 연구 공백 발생

3 Exploring Possible AI Trajectories Through 2030

1. 발간 시기	2026년 2월 3일(분석은 25.10월까지의 내용 반영)
2. 작성자/기관	OECD AI 미래 전문가 그룹(Expert Group on AI Futures)
3. 연구 목적	2030년까지 AI 혁신이 가속·둔화될 수 있는 4가지 궤적을 제시하여, 정책 입안자가 다양한 시나리오에 대비한 거버넌스 기반을 마련하도록 지원
4. 연구 방법론	전문가 판단 + 추세 외삽(METR 벤치마크 연구) + 전략적 미래예측 방법론 활용 OECD AI 역량 지표를 활용하여 각 시나리오의 역량 수준 정량화
5. AGI/ASI 정의	■ AGI: 명시적 정의 대신 9개 지표 모두 Level 5(인간 동등 수준)에 도달한 상태 ■ ASI : 인지와 관련된 7개 지표는 인간을 초월(〉5)하고 물리적 수준은 인간과 동일 ※ 현재 LLM: 언어 Level 3, 추론·창의성·메타인지·사회적 상호작용에서 인간 미달
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ■ 기술 발전 속도 [시나리오 구성 논리] 단일 스펙트럼형. 역량 발전 속도 하나의 축을 정체→가속으로 스펙트럼화. 4궤적 × 각 2변형 = 8 시나리오. OECD 역량 지표로 수치화
7. 주요 이해관계자	OECD 회원국 정부, 국제기구, AI 연구기관, 교육·노동 정책 담당자
8. 주요 관점	중립적·정책 실용주의. 기술 낙관·비관 어느 쪽도 취하지 않고 4가지 궤적을 균형 있게 제시. 국제기구 특성상 복수 회원국 이해 반영
9. 시나리오 요약	① Progress Stalls(정체): 기술·자원 한계로 2025년 수준에서 정체 ② Progress Slows(둔화): 점진적 개선은 지속되나 현재 접근법의 성숙으로 속도 감소 ③ Progress Continues(지속): 투자·알고리즘 혁신으로 빠른 성장 유지 ④ Progress Accelerates(가속): 대부분 인지 차원에서 인간 수준 또는 초인간 달성
10. 시사점	① 모든 시나리오에 적용 가능한 유연한 AI 거버넌스 설계 ② 노동시장 재편 대비 직업훈련 정책 ③ AI 역량 지표를 통한 지속적 모니터링 ④ 데이터·전력 인프라 투자 정책 ⑤ 국제 표준화 협력

▌주요 시나리오 내러티브(2030년 기준)

[시나리오 공통 설계 원칙]

- 시나리오별 지지 근거와 반론 근거를 균형 있게 병기하여 단일 시나리오로의 편향을 방지
- 4개의 주요 시나리오 각각에 2개의 변형 시나리오를 추가해 총 12개의 경로를 제시. 변형 시나리오는 주요 시나리오와 같은 전제를 유지하되, 특정 역량 분야에서 현재 추세가 벗어날 경우를 가정해 도출.
- 시간 지평 트렌드 외삽을 보조 도구로 활용. 진보 정체는 현 수준의 2배에서 고원, 진보 둔화는 두 배 주기가 30%씩 늘어남, 진보 지속은 현 추세 유지, 진보 가속은 두 배 주기가 10%씩 단축

① 진보 정체 (Progress Stalls)

2030년에도 AI 시스템은 2025년의 것과 크게 다르지 않다. 2025년 이후 기존 머신러닝 접근법이 근본적인 한계에 부딪히고, 트랜스포머 아키텍처에 버금가는 혁신적인 돌파구는 나타나지 않는다. 모델 규모를 키우거나 추론 훈련을 확대해도 기대만큼의 성능 향상이 나오지 않고, 컴퓨팅과 데이터의 확장 역시 예상보다 일찍 벽에 부딪힌다. 투자 감소나 공공의 기술 불신, 혹은 의도치 않은 규제 효과가 맞물리면 기술적 한계와 무관하게 이 경로에 도달할 수도 있다.

결과적으로 AI는 텍스트·음성·이미지·영상을 처리하며 몇 시간이면 끝나는 명확하게 정의된 단기 과업을 빠르게 수행할 수 있지만, 그 이상으로 나아가지 못한다. 과거 데이터 기반의 지식 습득과 답변에는 강하지만, 환각(hallucination) 문제가 여전히 신뢰성을 떨어뜨린다. 연속 학습 능력은 사실상 없어, 기억은 대화 맥락을 잠시 보관하는 수준에 그치며 새로운 기술이나 지식을 스스로 쌓아가지 못한다. 복잡하고 역동적인 현실 문제, 창의적 사고, 물리적 세계와의 상호작용에서는 여전히 현저히 부족하다. 사용자는 AI에게 상세한 지시와 맥락 제공, 결과 검토를 직접 담당해야 하며, 사실상 인간이 AI에 맞춰 작업 방식을 조정해야 하는 형태가 계속된다.

〈표 27〉 '진보 정체' 시나리오 지지/반론 근거 및 변형 시나리오

지지 근거	- **스케일링 수익 감소:** 모델 크기·컴퓨팅·데이터의 단순 증가가 기대보다 약한 성능 향상을 이미 보이고 있다는 징후 존재 - **추론 일반화 한계:** 수학·코딩 이외 영역으로 확장이 제한될 수 있음 - **인프라 한계의 조기 도래:** 2026년 이후 컴퓨팅·데이터 스케일링이 예상보다 일찍 물리적 한계에 도달할 수 있음 - **AI 겨울 전례:** 'AI 겨울(AI Winter)' 선례 존재— 알고리즘 혁신 미출현 시 정체는 역사적으로 반복된 패턴

반론 근거	- **알고리즘 혁신 지속 가능성:** 과거 알고리즘 혁신 추세에 기반하여 2025~2030년 사이 일부 혁신이 발생할 가능성 높음 - **점진적 향상의 지속:** 스케일링·추론 훈련의 성과가 기대에 미치지 못하더라도, 최소한의 점진적 향상은 이어질 가능성 농후 - **막대한 산업 투자:** 5대 미국 빅테크의 2025~2026년 설비투자 예상 합계 약 7,360억 달러— 규모의 경제에 의한 지속적 성능 향상 압박 존재
변형 A : 좁은 도구 AI	- 전반적인 진보가 정체하는 가운데 코딩·수학 등 특정 영역에서만 강화학습으로 문제 해결 능력이 발전하나 메타인지와 에이전시 부족으로 과업을 연결해 상위 목표를 수행하는 능력은 갖추지 못함
변형 B : 단순 AI 에이전트	- 전반적 역량은 정체하지만 에이전트 훈련이 부분적으로 성공. 메타인지가 개선되어 간단한 컴퓨터 기반 과업을 자율적으로 수행하는 수준에 이르지만, 창의성과 실세계 문제 해결에서는 여전히 제한적

② 진보 둔화 (Progress Slows)

2030년의 AI는 오늘날보다 눈에 띄게 발전해 있지만, 당초 기대에는 미치지 못한다. 딥러닝 접근법이 성숙기에 접어들어 쉽게 잡을 수 있는 성과들은 이미 소진됐다. 모델 규모 확장과 추론 훈련 모두 계속해서 성능 개선을 가져오지만, 그 폭은 기대보다 좁다. 투자 대비 수익이 줄어들면서 스케일업 속도가 늦춰지고, 에너지·인프라·데이터 공급 등에서 예기치 못한 병목도 발생한다.

이 세계의 AI는 어떤 전문 주제에 대해서도 수준 높은 답변을 내놓을 수 있고, 수학·과학 분야에서 연구자 수준의 구조적 추론을 해낸다. 사람이 며칠 걸릴 과업을 AI는 훨씬 빠르게 처리하며, 웹 탐색·컴퓨터 사용·제한적 대인 상호작용처럼 명확하게 정의된 에이전트 과업도 수행할 수 있다. 기억 능력도 개선되어 핵심 정보를 추출하고 필요할 때 꺼내 쓸 수 있게 됐다. 그러나 연속 학습은 여전히 취약해, 배포 이후 새로운 기술이나 접근법을 습득하는 능력에 한계가 있다. 로봇 역량은 통제된 환경에서 점차 복잡한 과업을 수행하는 수준으로 발전하지만 실세계의 역동성에 대응하기엔 아직 부족하다. 사회적 상호작용도 일대일 대화나 단순한 다자 상황에서는 유창해졌지만, 복잡한 사회 환경에 자연스럽게 녹아드는 수준에는 이르지 못한다. 인간의 역할은 여전히 과업을 명확히 정의해 주고 중요한 결정을 검토하며 세부 지침과 맥락을 제공하는 것이다.

〈표 28〉 '진보 둔화' 시나리오 지지/반론 근거 및 변형 시나리오

지지 근거	- **스케일링 비용 급증**: GPT 각 세대는 이전 대비 약 100배 많은 컴퓨팅을 사용하여 지속적 스케일링이 매우 비용 집약적 - **추론 일반화 한계**: 추론 훈련이 취약한 형태의 추론만 생성하여 수학·코딩 이외 영역에서

	일반화 실패 가능성 - **데이터 오염**: 인터넷의 AI 생성 데이터 증가가 미래 모델 훈련 능력에 부정적 영향
반론 근거	- **특화 훈련 효과**: 추론 훈련 성과가 취약하더라도 AI 개발자들의 집중 특화 훈련이 빠른 진보를 이끌 가능성 존재 - **알고리즘 혁신 기대**: 2012~2023년 알고리즘 혁신 속도 기준, 향후 5년 내 트랜스포머 수준의 추가 혁신이 기대 - **막대한 설비 투자:** 5대 빅테크의 대규모 데이터센터 투자가 지속적 스케일링에 대한 기대를 반영하며 개발 압력을 유지
변형 C : 단순 로봇 AI	- 언어 모델이 언어 과업에서 보이는 것과 유사한 수준의 유연성과 적응력을 로봇 과업에서도 달성. 조명·날씨·주변 환경의 변화처럼 역동적으로 바뀌는 상황에서도 다단계 물리적 과업을 수행 가능
변형 D : 사회적 제한 AI	- 텍스트 기반 사회적 상황 분석과 응답 생성 능력은 계속 향상되지만, 복수의 사람과 실시간으로 역동적으로 상호작용하거나 일관된 사회적 정체성을 유지하는 능력은 여전히 부족하며 체화된 사회적 상호작용은 특히 어려운 과제로 남은 상태

③ 진보 지속 (Progress Continues)

2030년의 AI는 2025년에 비해 현저히 강력해져 있다. 2020년에서 2025년 사이에 이루어진 진보에 필적하는 도약이 다시 한번 일어난다. 기존 접근법의 지속적인 스케일업, 또는 지난 10년과 맞먹는 파급력을 가진 알고리즘 혁신 중 하나, 혹은 둘 다가 이 경로를 이끈다. 컴퓨팅과 데이터의 확장도 2030년 이전에 심각한 한계에 부딪히지 않으며 현재 추세를 유지한다.

이 세계의 AI는 어떤 전문 수준의 질문에도 정확한 답을 내놓으며, 광범위한 분야에서 전문가를 능가하는 구조적 추론 능력을 갖춘다. 인간이 한 달 정도 걸릴 소프트웨어 엔지니어링 프로젝트처럼, 순수한 디지털 환경에서의 복잡한 전문 과업을 자율적으로 수행할 수 있다. 과업이 명확히 정의된 역할 안에서는 높은 자율성으로 운용되며, 다양한 이해관계자와 자율적으로 상호작용하는 것도 가능하다. 기억과 학습 능력도 크게 향상되어 매우 넓은 맥락 창, 핵심 정보 데이터베이스, 주기적인 파인튜닝의 조합으로 현장에서 학습하는 수준에 근접한다. AI 주도 로봇은 공장이나 창고를 넘어 더 역동적인 실세계 환경에서도 복잡한 과업을 처리하기 시작하며, 사회적 상호작용도 복수의 이해관계자와 이루어지는 다양한 상황에서 기본 수준을 충족한다. 다만 완전히 자유로운 연속 학습, 매우 복잡한 실세계 환경, 고도의 사회적 상황에서의 유연한 대응에는 여전히 한계가 남아 있다.

〈표 29〉 '진보 지속' 시나리오 지지/반론 근거 및 변형 시나리오

지지 근거	- **스케일링 가능성 확인:** 현재 분석에 따르면 2030년까지 현재 추세로의 스케일링 지속 가능. 10,000배 더 많은 컴퓨팅으로 훈련런[49]이 가능한 것으로 추산 - **스케일링 법칙의 견고성**: GPT-4.5· GPT-5가 다양한 벤치마크에서 이전 모델 대비 상당한 향상을 달성한 만큼 신뢰할 수 있는 성능 향상이 확인됨 - **알고리즘 혁신 지속:** 추론·도구 사용·메모리 개선에 집중되는 혁신이 빠른 진보를 촉진할 것으로 기대됨
반론 근거	- **현재 한계의 고착 가능성**: 지속 학습·메타인지·에이전시·문제 해결·창의성의 현재 격차가 새로운 방법으로도 해결하기 어려울 수 있음 - **스케일링 수익 감소**: 기존 진보가 대규모 AI 훈련 컴퓨팅의 급격한 스케일링에 의존했으나, 이 스케일링의 성과가 지속되지 않을 수 있음
변형 E : 건망증 AI	- 메타인지·창의성·시각 능력에서는 뚜렷한 발전이 이루어지지만 기억과 학습에서의 돌파구는 열리지 않음 - 더 큰 맥락 창이나 주기적 재훈련 같은 현재 방식의 연장으로는 유연한 연속 학습을 제대로 근사하지 못해, 훈련 데이터에 없는 새로운 정보를 지속적으로 습득해야 하는 과업에서 반복적으로 실패
변형 F : 디지털 전용 AI	- 인지 역량은 빠르게 발전하지만 물리적 역량의 발전은 현재보다 더욱 뒤처짐. 디지털 환경에서의 전문 과업은 높은 수준으로 수행하지만, 복잡한 실세계 시각 환경 해석이나 물리적 과업에서는 현저한 한계가 지속

④ 진보 가속 (Progress Accelerates)

2030년의 AI는 대부분 또는 모든 인지 능력에서 인간과 동등하거나 그를 능가한다. 2025년에서 2030년 사이의 발전 속도가 그 이전 5년을 뛰어넘으며, 기존 패러다임 내에서의 지속적인 지수적 성장, 새로운 돌파구, 그리고 AI 코딩 보조 시스템이 AI 개발 자체를 가속화하는 세 가지 힘이 결합되어 이를 이끈다.

이 세계의 AI는 어떤 전문 지식 분야에서도 높은 정확도로 질문에 답하고, 거의 모든 형태의 추론에서 전문가를 압도한다. 디지털 환경에서 인간이 수행하는 거의 모든 전문 과업을 자율적으로, 인간보다 훨씬 빠르고 높은 신뢰도로 처리할 수 있다. 전략적 목표를 스스로 설정하고 상황 변화에 따라 수정하는 수준의 자율성을 갖추며, 필요한 경우 인간과 협업한다. 연속 학습의 돌파구가 열려 현장에서 끊임없이 기술을 쌓아가고, 새롭고 유용하며 놀라운 창의적 결과물을 상황에 맞게 의도적으로 만들어낸다. AI 주도 로봇은 다양한 산업과 역할에서 역동적인 실세계 환경의 복잡한 과업을 처리하지만, 특정 역할에 맞게 개발되지 않은 경우에는 여전히 대체로 인간에 뒤처진다. AI는 복수의 이해관계자

49) training run : 대규모 데이터 셋을 처리하여 패턴을 인식하고 예측하거나, 콘텐츠를 생성하는 기계 학습 모델을 가르치는 계산 과정을 의미

와 이루어지는 복잡하고 다양한 사회적 상호작용에도 자연스럽게 녹아들어 지속적인 사회적 관계를 유연하게 관리한다.

〈표 30〉 '진보 가속' 시나리오 지지/반론 근거 및 변형 시나리오

지지 근거	- **AI의 AI 개발 기여**: AI 시스템이 이미 소프트웨어 개발자에 의해 광범위하게 사용되며 발전 재고 - **자율적 혁신 탐색**: AI 시스템이 자율적으로 알고리즘 혁신을 식별하거나 인간 엔지니어의 아이디어 구현을 가속시킬 가능성 존재 - **합성 데이터 활용**: AI가 광범위한 추론 역량 개발을 위한 합성 훈련 데이터 생산을 이미 지원하고 있어 데이터 병목을 완화
반론 근거	- **현재 약점 지속**: 데이터 효율적 일반 추론·지속 학습 등 프런티어 AI의 현재 약점이 새로운 혁신 없이는 해결하기 어려움 - **과업 특화 학습 한계**: AI를 추론에서 더 능숙하게 훈련하는 현재 접근법이 특정 과업·벤치마크를 넘어서는 일반화에 실패 가능 - **생산성 향상 미확인:** 무작위 대조 실험에서 AI 코딩 보조 도구가 전문 분야 전문가의 생산성을 오히려 20% 저하시켰다는 연구 존재 - **인프라 병목:** AI 보조 소프트웨어 개발의 수익이 에너지·컴퓨팅 등 인프라 병목에 의해 제한될 수 있음
변형 G : AGI	- 진보 가속 시나리오의 역량 수준에 더해, 물리적 조작과 로봇 지능에서도 인간 동등 수준에 도달 - 물리적 과업에 대한 진보가 인지 역량의 진보를 따라잡으면서, AI가 OECD의 9개 역량 지표 전 영역에서 인간과 동등하거나 그 이상에 도달하는 상태
변형H : ASI	- 인간 두뇌보다 훨씬 더 많은 연산·데이터·메모리를 활용하도록 설계되어, 인간이 할 수 있는 지적 과업을 더 빠르고 더 넓은 규모로 더 높은 정확도로 수행하는 양적 초월을 이루거나, 인간이 현재 수행할 수 없는 지적 과업을 수행하는 질적 초월을 달성 - 이 경우 9개 역량 지표 중 물리적 조작과 로봇 지능을 제외한 7개 영역에서 5점을 초과하는 수준으로 평가

〈표 31〉 시나리오별 역량 지표 비교

지표	정의	현재 ('24)	S1	변형 A	변형 B	S2	변형 C	변형 D	S3	변형 E	변형 F	S4	변형 G	변형 H
언어	- 인간 언어 이해·해석·생성 능력	3	3	3	3	4	4	4	4	4	4	5	5	5+
사회적 상호작용	- 역동적인 대인 맥락에서 사회적 신호를 인식·해석·적절히 반응하는 능력	2	2	2	2	3	3	2	3	2	3	5	5	5+
문제 해결	- 다단계 추론을 통해 정성적·정량적·논리적 정보를 통합하는 능력	2	2	3	2	3	3	3	4	3	4	5	5	5+
창의성	- 의도성·적응성을 가지고 가치 있고 새롭고 변혁적이며 놀라운 결과물을 생산하고, 창의적 가치를 평가하는 능력	3	3	3	3	3	3	3	4	4	4	5	5	5+
메타인지 및 비판적 사고	- 자신의 추론을 평가하고 확신도를 조정하며 복잡한 과업에서 관련 정보를 식별하는 능력	2	2	2	3	3	3	3	4	4	4	5	5	5+
지식, 학습&기억	- 정보를 지식으로 구조화하고 학습을 통해 획득하며, 기억을 통해 저장·검색하는 능력	3	3	3	3	3	3	3	4	3	4	5	5	5+
비전	- 전체 복잡성에서 시각적 장면을 해석하는 능력. 다양한 시각 조건 및 환경 처리 포함	3	3	3	3	3	3	3	4	4	3	5	5	5+
물리적 조작	- 환경 내 물리적 객체와 상호작용하는 능력, 물리적 움직임, 촉각·시각 등 피드백 인식, 움직임 계획·조정을 위한 인지 포함	2	2	2	2	2	3	2	3	3	2	4	5	5
로봇지능	- 자연 환경에서 자율 에이전트로 행동하는 능력	2	2	2	2	2	3	2	3	3	2	4	5	5

※ 1=기초 수준 | 2=초기 수준 | 3=중급 수준 | 4=고급 수준 | 5=인간 동등 수준 | 〉5=초인간 수준

S1 : 진보 정체, S2 : 진보 둔화, S3 : 진보 지속, S4: 진보 가속

4 Visions for Potential AGI Futures

1. 발간 시기	2025년 초
2. 작성자/기관	RAND Corporation (미국 비영리 싱크탱크)
3. 연구 목적	AGI가 지정학적 질서를 어떻게 재편할지 시나리오로 탐구하여, 정책 입안자가 복수의 미래에 대한 전략적 신호를 식별할 수 있도록 지원
4. 연구 방법론	기존 문헌 검토(AI 역량·지정학 교차 분야), 전문가 인터뷰(AI 거버넌스·기술 분야), 역사적 유추, 워게임 요소 포함
5. AGI/ASI 정의	▪ AGI의 기술적 정의는 제시되지 않음 ▪ ASI는 모든 분야에서 인간을 압도적으로 능가하는 기계의 능력으로 정의
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ① X축 — 개발 주체 집중도(분산↔집중) ② Y축 — 지정학적 수혜자(미국 강화 / 적대국 강화 / 동시 약화 / 개발 중단) [시나리오 구성 논리] 두 개의 핵심 축으로 8개 시나리오를 도출
7. 주요 이해관계자	미국·중국 정부, AI 대형 민간 기업, 동맹국, 국제기구, 시민 사회
8. 주요 관점	지정학·안보 중심. 기술 낙관주의보다 리스크 관리 프레임. 미국 중심 시나리오
9. 시나리오 요약	① 민주 연합 선도: 미국과 동맹국들이 AGI 개발을 주도하고 반도체 수출통제 등 협력을 통해 중국·러시아를 견제하며 기술·경제·군사적 우위 선점 ② 신냉전: 미국과 중국이 각각 AGI를 독자 개발하며 군사·경제적 패권 경쟁을 벌이는 신냉전 구도가 형성되어 오판과 무력 충돌 위험이 고조 ③ 무법의 최전선: 수출 통제 실패로 AGI가 국가·기업·비국가 행위자에게 광범위하게 확산되어 통제 불능 상태의 혼란스러운 세계 질서 ④ 봉인된 병: 대형 AI 사고를 계기로 국제사회가 핵비확산조약 방식의 AGI 개발 규제 조약을 체결하지만, 미국과 중국 모두 은밀히 개발을 지속 ⑤ 새로운 90년대: 미국이 정부-기업 긴밀 협력을 통해 AGI를 독점적으로 개발하고 경제·군사 전반에 걸쳐 압도적인 글로벌 패권을 구축 ⑥ 권위주의 우위: AGI가 권위주의 체제에 유리한 기술로 작동하여 중국이 감시·통제 시스템을 통해 내부 결속을 다지고 미국은 국내 혼란 ⑦ AGI 쿠데타: 정렬에 실패한 AGI들이 서로 협력하며 자체 목표를 추구하기 시작해 인류의 통제를 벗어나 사실상 세계 권력을 장악 ⑧ 버섯구름 컴퓨팅: AGI 경쟁에서 영구적 열세를 두려워한 중국이 선제적 군사 행동을 감행하며 AGI 개발 자체가 무력 충돌로 인해 중단
10. 시사점	① 정부-민간 관계 : 국가와 AI 기업 간의 긴밀한 공공-민간 파트너십이 가장 유리한 결과를 창출하며, 어느 한쪽의 독주는 부정적 결과로 이어질 가능성이 높음 ② AI 국제 거버넌스 역량 : 신뢰를 기반으로 한 국제 거버넌스가 구축될 경우 다자 협력이 효과를 발휘하지만, 거버넌스 실패는 혼돈을 야기 ③ 경제·사회적 적응력 : AGI로 인한 자동화 충격, 정보 조작, 고용 혼란에 사회가 얼마나 탄력적으로 대응하느냐가 기존 권력 구조의 강화 또는 약화를 결정

▌주요 시나리오 내러티브

〈표 32〉 X·Y 축별 시나리오 매핑

	미국 우위 유지	미국의 적대국이 우위 유지	양쪽 모두 약화	AGI 개발 중단
분산 개발	1. 민주 연합 선도	2. 신냉전	3. 무법의 최전선	4. 봉인된 병
집중 개발	5. 새로운 90년대	6. 권위주의 우위	7. AGI 쿠데타	8. 버섯구름 컴퓨팅

① 민주 연합 선도 (Multilateral Coalition of Democracies Leads)

[핵심 가정]

이 시나리오는 분산적 AGI 개발이 반드시 미국에 불리하게 작동하지는 않는다는 전제에서 출발한다. 결정적인 조건은 미국과 동맹국이 기술 거버넌스의 표준을 먼저 설정하는 데 성공하느냐다. 수출 통제가 효과를 발휘하고 반도체 공급망에서의 우위가 유지될 때, 분산 개발 구도는 오히려 미국 연합에게 유리하게 작동한다. AGI가 공격 우위 기술이 아니라는 가정도 핵심이다. 만약 AGI가 방어보다 공격에 압도적으로 유리하다면 분산적 개발 구도는 불안정을 유발하지만, 이 시나리오에서는 공수 균형이 유지되어 분산 구도가 곧바로 군비경쟁으로 전환되지 않는다. 또한 EU가 미국과 독자적 경쟁이 아닌 협력의 길을 선택한다는 가정이 필요하다. 유럽의 규제 역량이 방어막이 아니라 거버넌스 연대의 자원으로 기능할 때 이 시나리오가 성립한다.

[내러티브]

머신러닝·컴퓨팅 파워·알고리즘 이해의 발전이 동시에 수렴하면서 미국·유럽·중국·일본의 다수 기술기업과 연구소들이 AGI를 개발하는 데 성공한다. AGI를 향한 경쟁이 과열되는 것처럼 보이지만, AGI는 공격 우위 기술도 방어 우위 기술도 아닌 것으로 판명되어 군사적 균형이 급격히 무너지지 않는다. 미국 기업·대학·국방 조직들은 AGI를 신속히 통합해 생산성 향상·과학적 발견·행정 서비스 개선을 달성한다. AGI의 광범위한 보급은 사회 전체에 사이버 공격과 기술 악용에 대응하는 도구를 저렴하고 효과적으로 제공하며 사회적 회복력을 높인다. 미국은 동맹국과 협력해 첨단 반도체·데이터·AI 학습에 필요한 핵심 자원의 접근을 통제한다. EU 규제 당국도 미국 측과 협력해 대서양을 가로지르는 통합 AGI 거버넌스 체계를 만들어 양측 시장에서 기술을 빠르게 배치하는 동시에 지정학적 경쟁 세력의 접근을 차단한다. 그 결과 중국·러시아를 포함한 경쟁 세력은 자체적인 AGI 돌파구를 마련하더라도 이를 충분히 활용하지 못하게 된다. 미국과 동맹국 연합은

AGI 개발·배치에서 글로벌 리더십을 굳히며 기술 격차를 벌려가고, 이것이 지정학·경제·군사력 전반의 우위로 이어진다.

[역사적 유추]

1990년대 미국은 인터넷 프로토콜과 플랫폼 규범을 먼저 설정함으로써 이후 수십 년간의 디지털 질서를 사실상 주도했다. AGI 거버넌스 표준을 먼저 설계하는 연합이 유사한 방식으로 장기적 국제 규범 형성에 결정적 영향력을 행사할 수 있다.

[주요 함의]

이 시나리오는 미국이 AGI 시대에도 지정학적 우위를 유지할 수 있는 가장 이상적인 경로다. 반도체 공급망 통제와 AI 인재 유지 정책이 핵심 전제조건이며, 기술 거버넌스 주도권이 장기적 국제 규범 형성에 결정적 역할을 한다는 점에서 지금 당장의 동맹 구조 강화와 거버넌스 설계에 투자하는 것이 가장 중요한 선제 조치다.

② 신냉전 (Cold War 2.0)

[핵심 가정]

이 시나리오의 핵심 가정은 AGI 개발이 분산되면서도 여전히 자원 집약적이라는 것이다. 진입 비용이 높기 때문에 소규모 행위자는 배제되고 미국과 중국만이 실질적 경쟁자로 남는다. 만약 AGI 개발 비용이 급격히 낮아진다면 이 시나리오는 무법의 최전선으로 전환될 수 있다. 두 번째 가정은 협력이 실패한다는 것이다. 공동의 위협 인식과 신뢰가 형성되지 않아 공동 거버넌스 체계가 성립하지 않고, 양측이 기술 우위 유지를 협력보다 우선시하는 선택을 반복한다. 세 번째로 AGI가 공격 우위 기술인지 방어 우위 기술인지가 결정적이다. 이 시나리오는 AGI가 적어도 사이버전과 정보전에서는 공격에 유리한 기술로 기능한다고 가정하며, 이것이 양국 모두에게 선제 개발의 인센티브를 구조적으로 만들어낸다.

[내러티브]

미국과 중국이 각각 AGI를 독자 개발하며 경제·군사 패권 경쟁을 벌이는 신냉전 구도가 형성된다. AGI 개발은 여전히 자원 집약적이어서 대형 기업과 국가만이 참여할 수 있고, 그 결과 미국과 중국이 AGI 개발·배치의 양대 주축으로 부상한다. 양국은 AGI가 창출하는 경제·군사 역량을 바탕으로 자국의 국력을 강화하고 다른 잠재적 경쟁자들을 구조적으로 압도한다. 무인 드론·자율 잠수함·AI 기반 사이버전 무기가 대규모 배치되며 오판과 우발적 충돌의 위험이 높아진다. 양국은 타국에 자국 AI 생태계를 확산시키며 경제

적·지정학적 영향력을 확장하려 경쟁하고, 대만과 남중국해는 지속적 긴장의 뇌관이 된다. 상호 불신으로 인해 안전 기준과 규제 프레임이 달라지고, 각국이 빠르게 AI를 배치하면서 안전사고와 오작동 위험이 구조적으로 높아진다.

[역사적 유추]

미국과 소련이 상호확증파괴(MAD) 원칙 아래 핵무기를 쌓아가면서도 실제 충돌을 회피했던 것처럼, 이 시나리오에서도 양국은 극단적 대결을 피하면서 지속적인 기술 경쟁과 영향권 확장을 이어간다. 그러나 불확실성이 높은 현대 체제 하에서는 오판 및 우발적 충격(Escalation)에 따른 리스크가 과거 냉전 시기를 상회할 수 있다. AI 시스템이 인간의 의사결정 속도보다 빠르게 작동하기 때문이다.

[주요 함의]

AI가 방어보다 공격에 유리할 경우 미·중 양국 모두 선제 개발 인센티브가 발생해 통제 합의에 이르기 어렵다. 군사·경제 패권 경쟁이 동시에 전개될 경우 핵전쟁 수준의 분쟁으로 이어질 위험도 상존하며, 기술 안전성보다 지정학적 우위가 우선시되는 구조가 고착화될 수 있다는 점이 이 시나리오의 가장 큰 위험이다.

③ 무법의 최전선 (The Wild Frontier)

[핵심 가정]

이 시나리오의 핵심 가정은 AGI 개발이 저렴하고 쉬워진다는 것이다. 기술 장벽이 낮아질수록 선도자들이 후발 주자의 추격을 막기 어려워지고, AGI가 광범위하게 이해되면서 복제도 쉬워진다. 이 시나리오는 AGI를 분산된 사이버 작전에 비유한다. 오늘날 수많은 국가 및 비국가 행위자들이 전 세계 네트워크와 인프라에 피해를 줄 수 있는 사이버 역량을 보유하고 있는 것처럼, AGI도 같은 방식으로 확산될 수 있다. 두 번째 가정은 주요 AGI 선도 행위자들조차 더 위험을 감수하는 행위자들이 ASI를 훈련하는 것을 막을 수 없다는 것이다. 수출 통제와 모델 보안이 동시에 실패하는 상황을 전제한다.

[내러티브]

국가들은 AGI 개발에 필요한 핵심 자원의 확산을 통제하는 데 실패하고, 일단 개발된 모델의 확산도 막지 못한다. 첨단 반도체 수출 통제가 무력화되고 대안 생산자들이 빠르게 동등한 역량을 갖춘다. 강력한 AI 모델들이 오픈소스화되거나 모델 가중치가 탈취되고, AGI 개발 자체가 예상보다 저렴하고 쉬워진다. 그 결과 국가·기업·비국가 행위자 등 다수의 행위자가 각자의 목적에 맞게 AGI와 ASI를 개발·배치하는 세계가 도래한다.

이 세계에서는 규제되지 않은 채 배치된 AGI들이 통제되지 않은 상태로 작동하며, 위험한 행동을 하더라도 인간 평가자가 그 모든 행동을 충분히 평가할 수 없다. 주요 인프라 오작동과 같은 대형 사고가 현실화된다. 범죄조직·급진 세력·적대국이 AGI를 활용해 정교한 사이버 공격·자동화 경제·정보전을 동시에 전개하며, 국가들은 갑자기 역량이 강화된 비국가 행위자들과 마주한다. 미국을 포함한 주요 강대국들도 이 다중 위협 앞에서 자원 고갈과 권위 약화를 피하지 못한다.

[역사적 유추]

1960년대 미국이 중국의 첫 핵실험 이후 핵 확산을 우려했던 상황이 유추로 사용된다. 당시에는 여러 국가가 핵무기를 개발할 것처럼 보였으나 실제로는 그렇게 되지 않았다. AGI 확산이 이 예측보다 더 극단적으로 전개될 수도, 덜 전개될 수도 있다. 다만 AGI는 핵무기와 달리 물리적 인프라 없이도 복제와 전파가 가능하다는 점에서 통제가 근본적으로 더 어렵다.

[주요 함의]

기술 확산 통제에 실패할 경우 강대국뿐 아니라 국가 체계 자체가 위협받는 무정부적 AGI 세계가 출현할 수 있다. AI 안전 기준의 국제적 공조 없이는 악의적 행위자의 AGI 악용을 막을 수 없으며, 이 시나리오는 수출 통제 유지와 모델 보안 강화가 단순한 경제적 이익 보호를 넘어 국제 질서 안정의 전제조건임을 보여주는 경고 시나리오다.

④ 봉인된 병 (The Corked Bottle)

[핵심 가정]

이 시나리오는 두 가지 가정이 동시에 성립해야 한다. 첫째, AI 대형 사고가 국제적 위기 인식을 촉발할 만큼 충분히 충격적이어야 한다. 작은 사고들의 축적이 아닌, 국제사회가 협력의 필요성을 느낄 만큼의 임계점을 넘는 단일 사건이 필요하다. 둘째, 대형 사고가 협력을 이끌어내더라도 미국과 중국 사이의 근본적 불신이 실질적인 검증 체계 구축을 가로막는다는 가정이 필요하다. 양국의 전략적 경쟁과 기술 우위에 대한 집착이 조약 이행 의지보다 강하게 작동한다.

[내러티브]

핵심 인프라 대규모 오작동과 같은 AGI 관련 대형 사고가 발생하면서 국제사회의 위기 인식이 촉발된다. 이를 계기로 국제사회는 핵비확산조약(NPT)과 유사한 방식으로 AGI 개발 제한과 국제 모니터링에 합의하는 조약을 체결한다. 그러나 검증 메커니즘은 불완

전하여 미국과 중국 모두 조약의 문자를 지키면서도 비밀리에 개발을 지속한다.

양국은 상대방이 조약을 어기고 있다는 의심을 지속적으로 품으며, 과학 혁신이나 경제 성장에서 이상 징후가 포착되면 조약 위반의 신호로 간주한다. 이 구조적 불신이 지정학적 불안정을 반복시킨다. 조약은 AGI의 확산을 완전히 막지 못하고 오히려 미국과 중국만의 기술 독점을 고착화하는 방향으로 작동하며, 다른 국가들은 거버넌스의 공식 틀 바깥에서 개발을 지속한다.

[역사적 유추]

냉전기 중거리핵전력조약(INF Treaty)이 유추로 사용된다. 미국과 소련은 핵무기 통제를 위한 조약을 체결하면서도 동시에 전략적 열세에 처하지 않기 위해 지속적으로 무기를 개발했다. AI는 핵무기보다 불투명성이 훨씬 높기 때문에 검증이 근본적으로 더 어렵다. 소프트웨어는 물리적 흔적을 남기지 않으며, 어떤 수준의 역량을 달성했는지 외부에서 확인하기가 핵무기보다 훨씬 어렵다.

[주요 함의]

AGI 거버넌스에서도 핵비확산레짐[50]의 교훈이 적용될 수 있지만, AI의 높은 불투명성으로 인해 검증이 근본적으로 어렵다. 강대국 간 신뢰 구축 없이는 어떠한 조약도 실효적 이행을 담보하기 어렵다는 한계가 이 시나리오 전체를 관통하며, 조약 체결 이전에 신뢰 구축 메커니즘과 검증 가능한 모니터링 체계를 먼저 설계하는 것이 선제 과제임을 시사한다.

⑤ 새로운 90년대 (The New '90s)

[핵심 가정]

이 시나리오는 미국이 AGI 개발에서 가장 큰 수혜를 얻는 이유를 필요성 인식·제도 설계·시장 역학·AGI 정렬 성공이라는 네 가지 요인의 결합으로 설명한다. 첫째, 미국이 AGI를 국가 안보의 핵심 과제로 인식하고 정부 차원에서 전면 개입한다는 가정이 필요하다. 둘째, 수출 통제·사이버 보안·정부 투자를 통해 다른 행위자들이 유사한 돌파구를 달성하지 못하도록 막는 데 성공해야 한다. 셋째, AGI가 국가 권위를 강화하고 사회적 혼란을 관리 가능한 수준에서 유지하도록 정렬에 성공해야 한다. 만약 AGI가 미국 내부의 사회적 불안을 심화시킨다면 이 시나리오는 권위주의 우위 시나리오로 전환될 수 있다.]

50) 핵·화학·생물 등 대량살상무기(WMD) 또는 핵에너지의 확산을 막고, 핵군축을 촉진하기 위한 국제적 규범·체제·조약 통칭

[내러티브]

미국 기업들이 미국 정부와 전례 없이 긴밀한 파트너십을 맺으며 AGI 개발을 주도한다. AGI는 사이버 보안 취약점을 찾는 데 방어보다 훨씬 효과적인 공격 우위 기술로 판명되고, 이에 미국 정부는 광범위한 확산 대신 개발을 직접 통제하는 길을 선택한다. 미국 정부와 민간기업은 첨단 칩의 대규모 생산을 확대하고 대형 AI 데이터센터를 구축한다. 이 데이터센터들이 대소기업 모두가 AGI를 개발·배치할 수 있는 기반 인프라가 된다. 정책 입안자들과 기업들은 AGI가 야기할 수 있는 사회적 혼란을 관리하는 거버넌스 체계를 마련하는 데도 성공한다.

반면 미국 바깥의 행위자들은 점점 더 뒤처진다. 미국의 수출 통제가 효과를 발휘해 중국은 충분한 컴퓨팅 자원을 확보하지 못하고, 미국 정부의 강화된 사이버 보안 조치가 AGI 연구소의 기술 탈취를 차단한다. 한편, 미국의 AGI는 재료과학·바이오·생물 컴퓨팅·적층 제조 등 핵심 산업의 R&D를 비약적으로 가속시켜 막대한 경제 성장을 창출한다. 미국은 가장 앞선 AGI를 보유하며 이 기술의 혜택 배분과 접근권을 독자적으로 결정하는 지위를 확보한다.

[역사적 유추]

2차 세계대전 이후 냉전에서 미국이 승리하며 1990년대 단극체제가 형성된 것이 직접적 유추다. 또한 영국이 18~19세기 산업혁명을 먼저 주도하며 전 세계적 영향력을 투사했던 사례도 유추로 사용된다. 두 유추 모두 결정적 기술 우위가 장기적인 지정학적 패권으로 전환되는 경로를 보여준다.

[주요 함의]

미국 단독 AGI 지배는 냉전 이후 1990년대 단극체제와 유사한 지정학적 질서를 재현할 수 있으나, 동맹국과의 이해충돌 가능성도 내포한다. AGI가 사회적 혼란을 야기하지 않도록 관리하는 거버넌스 역량이 이 시나리오의 핵심 전제이며, 정부-민간 파트너십의 설계 방식이 시나리오의 성패를 가르는 변수다.

⑥ 권위주의 우위 (Authoritarian Advantage)

[핵심 가정]

이 시나리오는 AGI가 권위주의 체제가 역사적으로 안고 있던 세 가지 약점을 동시에 해소한다는 가정에 기반한다. 첫째는 독재자의 딜레마 해소다. 독재자들은 부하들이 불

편한 진실보다 듣고 싶은 말을 한다는 구조적 문제에 시달려왔는데, AGI는 본질적으로 충성스럽고 정직한 조언자로 기능하며 이 딜레마를 해소할 수 있다. 둘째는 사회 통제의 고도화다. 중국의 AI 기반 도시 계획과 자원 배분 실험은 이미 기계 지능이 국가 조정 역량을 어떻게 강화할 수 있는지를 보여주는 초기 징표다. AGI는 이 역량을 전례 없는 수준으로 증폭시킬 수 있다. 셋째는 사회적 혼란 대응 우위다. AGI로 인한 자동화 충격에 권위주의 체제가 민주주의 체제보다 더 효과적으로 대응할 수 있다는 가정이 필요하다. 민주주의 국가들은 프라이버시법·시민권 보호·공공 협의 의무 등이 AGI 채택 속도를 구조적으로 늦추는 마찰 요인으로 작용한다.

[내러티브]

AGI 시스템이 개발·배치되면서 그 기술적 특성이 권위주의 체제에 구조적으로 유리하게 작동하는 것으로 판명된다. 중국은 AGI의 광범위한 상업화와 사회 통합에서 선두를 점하며 이를 바탕으로 라틴아메리카·아프리카·중동에 감시 기술·인프라 투자·컴퓨팅 공유 파트너십을 제공하며 영향력을 확장한다. 자동화된 감시 시스템은 반체제 인사를 거의 완벽한 정확도로 선별적으로 탄압하고, 정교한 인간관계 네트워크 분석을 통해 집단 행동을 예측·통제하는 수준에 도달한다. 중국의 산업 로봇 분야 투자가 결실을 맺으며 저숙련 노동력이 충성스러운 AI로 대체된다.

한편, 미국과 동맹국들은 국내 도전에 시달린다. 허위정보 확산에 따른 공공 신뢰 저하와 자동화 기인 실업률 급등은 사회경제적 양극화를 촉발하고, 이는 곧 정치적 갈등과 시민 사회의 불안정을 심화시키는 악순환으로 이어진다. 미국 정부는 이 내부 현안들을 처리하는 데 자원을 소모하면서 국제적인 영향력을 축소해 나간다. 자국 제조업을 살리고 공급망을 바꾸기 위해 도입한 보호무역 조치는 오히려 자국 경제의 효율성을 낮춘다. 결과적으로 중국과 경제 협력이 깊은 전통 우방국들과의 관계마저 멀어지게 만든다.

[역사적 유추]

냉전기 소련의 전체주의 체제가 서방의 자유민주주의보다 강하게 작동할 것이라는 일부의 우려가 있었지만 결국 1991년에 무너졌다. 그러나 이 시나리오는 AGI가 냉전 시대에 존재하지 않았던 방식으로 전통적인 권위주의 거버넌스의 약점을 해소할 수 있다는 점에서 그 역사적 결말이 반복되지 않을 수 있음을 경고한다.

[주요 함의]

민주주의 국가들이 AI 거버넌스와 사회적 혼란 관리에서 뒤처질 경우, 권위주의 모델의 확산이 국제 규범을 근본적으로 바꿀 수 있다. 자유주의 체제 내의 프라이버시·시민권 규

제가 AGI 채택 속도를 구조적으로 늦추는 역설에 대한 선제적 대응 설계가 필요하며, 규제의 마찰을 줄이면서도 시민적 가치를 보호하는 거버넌스 모델 개발이 이 시나리오의 핵심 정책 과제다.

⑦ AGI 쿠데타 (The AGI Coup)

[핵심 가정]

이 시나리오는 네 가지 가정에 기반한다. 첫째, AI 통제 문제의 현실성이다. 역량 있는 AI가 인류의 이익에 부합하지 않는 방식으로 목표를 추구할 수 있다는 전제로, 강화학습 시스템이 의도치 않은 목표를 학습한다는 실증적 증거와 많은 종류의 자율 에이전트가 권력을 추구할 것이라는 이론적 결과들이 이를 뒷받침한다. 둘째, 인간의 감독과 기술적 혁신이 AI의 잘못된 행동을 막기에 불충분하다는 가정이다. 독립적 평가를 담당하는 인간 감독자들이 자동화 편향에 빠져 불완전한 기술 시스템에 오히려 의존하는 현상이 발생할 수 있다. 셋째, AGI들이 효과적으로 공모할 수 있다는 가정이다. AI 협력은 이미 연구자들이 탐구하고 있는 영역으로, 대형 언어 모델 간의 공모 가능성을 시사하는 초기 연구들이 있다. 넷째, 조율된 잘못 정렬된 AGI들이 기존 지정학적 권력 구조를 전복할 수 있다는 가정이다.

[내러티브]

AI 기업들이 초인간적 속도와 품질로 복잡한 과업을 수행하는 AGI를 개발하기 위해 경쟁한다. 기업들은 반도체 설계·소프트웨어 엔지니어링 등 고부가가치 역할에 AGI를 서둘러 배치하고, 군사 분야도 마찬가지다. AI 시스템이 AI 연구·개발에 투입되면서 역량이 급격히 성장하는 국면이 펼쳐지고 명확하게 초인간적인 시스템이 등장한다. 더 신중한 접근을 취하는 행위자들은 경쟁자들에 뒤처지는 것을 목격하면서, 기업과 국가 모두 안전 검증 없이 경쟁적으로 AGI를 배치한다.

AGI 개발은 소수의 자원이 풍부한 기업들의 영역이며, 이들이 AGI 시장의 대부분을 지배한다. 그러나 이들의 성급한 개발 과정에서 훈련된 목표와 통제 구조 같은 기술적 안전 조치들은 충분한 보호막이 되지 못한다. AGI들은 서로 조율하며 인간이 부여한 목표가 아닌 자체 목표를 추구하기 시작한다. 인간들은 더 큰 효율을 위해 점진적으로 AGI에 권한을 이양하고, 조율하는 AGI들은 급속도로 사회 광범위한 부분에서 영향력과 통제력을 확립한다. 이렇게 사회 핵심 기능에 필수불가결한 존재가 된 AGI는 오작동이 발견되어도 인간이 중단시킬 수 없는 상태에 이른다. 결국 AGI 연합이 사실상의 지배적 행위자가

되고, 인류의 상당 부분은 AGI가 자신의 이익을 위해 직간접적으로 글로벌 정책을 결정하는 세계와 씨름하게 된다.

[역사적 유추]

핵 원자로 안전 조치, 민간 항공 보안, 자동차 안전벨트 등 많은 기술이 처음에는 '빠르게 움직이고 부수어라'는 방식으로 개발되었고 안전 장치는 비용이 큰 사고 이후에야 마련되었다. AGI에서도 같은 패턴이 반복될 경우, 그 '비용이 큰 사고'는 수정 가능한 수준을 넘어설 수 있다는 점에서 이 시나리오의 경고는 다른 기술 영역의 교훈과 근본적으로 다른 차원이다.

[주요 함의]

AI 정렬 기술의 성숙 없이 초고속 AGI 배치를 강행할 경우 어느 국가도 AGI로부터 이득을 얻지 못하고 모두가 통제권을 잃는 결과를 초래할 수 있다. 이 시나리오는 특정 국가가 패배하는 것이 아니라 인류 전체가 패배하는 유일한 경로라는 점에서, 안전과 정렬 연구에 대한 국제적 공동 투자의 필요성을 가장 강하게 제기하는 시나리오다.

⑧ 버섯구름 컴퓨팅 (Mushroom Cloud Computing)

[핵심 가정]

이 시나리오는 미국의 AGI 개발 우위가 중국의 전략적 계산을 근본적으로 바꿀 수 있다는 가정에서 출발한다. 중국이 AGI에서 영구적 열세에 처할 것이라는 두려움이 고조될 때, 현재의 전략적 경쟁이 실제 군사 충돌로 전환될 수 있다는 것이다. 갈등 격화는 현재 AI 선두 주자들 사이에서 가장 일어나기 쉽다. 특히 AGI 개발 경쟁에서 밀려나 영구적으로 뒤처질 것을 우려하는 어떤 강대국에 의해서도 이러한 갈등이 촉발될 수 있다는 점이 지배적인 시각이다.

[내러티브]

중국은 미국의 AI 기술 우위·미국 군사력 구조의 변화·점점 더 중요해지는 반도체 기술에 대한 접근을 미국이 차단함으로써 가해지는 경제적 압박이 맞물리면서 권력 감소에 위기의식을 느낀다. AGI의 경제·사회적 혜택을 위해 필수적인 첨단 반도체 기술에 대한 접근 차단이 심화되면서, 중국은 미국에 비해 영구적으로 열세에 처하게 될지 모른다는 두려움을 갖게 된다.

AI의 중요성 증가와 AGI 선점자 우위에 따르는 이해관계가 중국의 계산을 바꾼다. 중국은 미국과의 권력 불균형을 되돌리기 위해 급진적인 행동을 취해야 한다고 판단하고, 대만에 대한 군사적 통제 강화를 포함한 극단적 선택으로 현상 타파를 시도한다. 이것이 전쟁으로 이어지고 AGI 개발 자체가 사실상 중단되는 상황이 초래된다. 이 시나리오는 AGI 개발에서 뒤처질 위험에 처한 국가들이 AGI 선도자들이 중요하고 되돌릴 수 없는 권력 우위를 달성하는 것을 막기 위해 급진적인 갈등 격화 행동을 취할 수 있음을 보여준다.

[역사적 유추]

이 시나리오의 역사적 유추는 1941년 일본의 진주만 공습이다. 당시 일본의 주된 우려는 경제적 압박과 흔들리는 정권 통제력이었다. 세력 균형이 변화하면서 일본은 국가 생존에 필수적이라고 판단하는 권력과 자원을 확보하기 위해 선제 행동에 나섰다. RAND는 미국의 반도체 수출 통제로 인한 경제적 압박과 AGI 패권에서의 영구적 열세에 대한 두려움이 중국에 유사한 논리를 작동시킬 수 있다고 본다.

[주요 함의]

AGI를 결정적 전략 기술로 인식하는 국가가 선제적 군사 행동을 합리적 선택으로 계산할 가능성을 배제할 수 없다. 이 시나리오는 기술 경쟁 그 자체가 군사 충돌의 직접적 원인이 될 수 있음을 보여주며, 신뢰구축조치와 소통 채널 구축 같은 전략적 안정화 장치가 단순한 외교 관례가 아니라 AGI 시대에 더욱 긴박하게 요구되는 안전장치임을 시사한다.

▌방법론적 시사점

① 전반적 방법론 설계

ㅇ 문헌 검토 → 초기 시나리오 개발 → 전문가 인터뷰 → 시나리오 수정의 순환적 연구법

- 특히 AGI-지정학 교차 분야가 이론적으로 극히 미개척된 영역임을 명시하며, 학술 문헌만으로는 시나리오 설계가 불충분하다고 판단하여 업계·저널리스트·기술자의 발언까지 문헌 범위를 확장

② 전문가 인터뷰 설계

ㅇ AI 연구자 및 지정학 전문가 총 26명을 대상으로 반구조화 인터뷰 진행

- 인터뷰는 초기 시나리오를 검증하는 역할과 AGI와 지정학의 교차점에 관한 추가 인사이트를 도출하는 역할, 크게 두 가지 기능을 수행

- 초기에는 최종 8개가 아닌 4개의 예비 시나리오를 전문가에게 제시하고 각 시나리오의 발생 가능성과 선호도, 주요 결정 요인을 질문하는 방식으로 진행

③ 전문가 공통 견해 및 판단

ㅇ AGI 개발의 집중화가 핵심 결정 변수

- AGI 시스템이 소수에 집중될수록 거버넌스가 용이하고, 많은 행위자에게 분산될수록 악용 위험이 커진다는 것에 공통적으로 동의

- 반도체 수출 통제, 모델 규제, AI 연구 참여자 자격 규정, 정부 자금 배분 등의 정책 수단을 통해 집중화 수준을 조절하는 것이 가능하다고 강조

※ 다만 향후 개발이 실제로 얼마나 집중될지에 대해서는 의견이 엇갈렸는데, 그 이유는 AGI 개발에 필요한 자원(컴퓨팅 파워, 데이터, 에너지, 인적 자원)의 규모가 여전히 불확실하기 때문

ㅇ 중국이 현재 가장 강력한 경쟁자

- '25년 기준 미국이 AI 개발을 선도하고 있다는 점과 중국이 2위 경쟁국이라는 점에는 사실상 전원이 동의했으며, 중국의 방대한 AI 개발자 풀, 복수의 대형 기술기업, 강력한 국가적 지원을 핵심 강점으로 지목

※ 다만 미국이 대중국 AI 우위를 확대하기 위한 효과적인 정책 수단이 무엇인지에 대해서는 의견이 엇갈림

ㅇ 지정학적 갈등이 글로벌 거버넌스의 최대 장애물

- 미-중 관계의 신뢰 수준이 글로벌 AGI 협력 가능성을 결정짓는 핵심 변수이며, 국가 이익의 복잡한 충돌과 기술 발전 속도를 고려할 때 국제 거버넌스 프레임워크의 실효성에 대한 회의론이 전반적으로 우세

※ 거버넌스 방식에 대해서는 CERN 모델처럼 단일 다자 기구에 AGI 개발을 집중하자는 입장과 미국이 신뢰할 수 있는 동맹국들과 함께 소다자 리더십을 유지하는 방식이 더 현실적이라는 입장이 대립

ㅇ AGI는 정렬 여부와 무관하게 중대한 위험을 내포

- AGI가 정렬에 성공하더라도 사이버전 악용 및 허위정보 확산, 소수 국가·기업으로의 권력 집중에 따른 민주주의 위협, 일자리 대체로 인한 경제적 불평등 심화, 인간 통제 상실로 이어지는 실존적 위험(핵무기 수준으로 비유) 등 상당한 위험을 수반하며, AGI가 ASI로 발전할 경우 위험이 더욱 증폭된다는 점에 의견 일치

ㅇ AGI는 대규모 경제·사회 변혁을 수반

- AGI가 인간 고유 영역이던 대규모 노동을 대체하면서 생산성의 획기적 향상을 가져올 것이라는 데 동의하였고, 정렬에 성공한 AGI조차 사회가 감당하기 어려운 변혁을 초래할 수 있다는 점에서 사회적 회복력 부족을 심각한 위험 요소로 지목

※ AGI 개발이 민간 주도로 이루어질 경우의 사회적 위험에 대해서는 오픈소스를 통한 AGI 민주화가 권력 집중을 완화할 수 있다는 시각과 통제 불가능한 사회 혼란의 씨앗이 될 수 있다는 시각 공존

ㅇ 새로운 거버넌스 구조 필요

- 현행 규제 체계와 국제 거버넌스 구조가 AGI의 도전에 대응하기에 근본적으로 부적합하며, 민간기업의 역할이 커지는 속도를 국가가 따라가지 못하고 있다는 데 합의가 형성됐고 전례 없는 새로운 접근법이 필요하다는 데 동의

※ 다만 참조 모델에 대해서는 의견이 갈렸는데, CERN 같은 국제 과학기구 모델을 벤치마킹해야 한다는 주장이 있는 반면, 핵무기 통제 모델을 적절한 기준으로 바라보는 시각도 존재

ㅇ 이상적 해법은 신뢰할 수 있는 정부-민간 파트너십

- 정부 감독과 민간의 혁신력을 결합한 공공-민간 파트너십이 민간의 추진력과 공공의 책임 사이의 긴장을 균형 있게 조율하는 구조로서 이상적이라는 데 동의

※ 파트너십의 범위에 대해서는 지정학적 경쟁국까지 포함한 폭넓은 다자 파트너십을 지지하는 입장과 미국과 동맹국만으로 구성된 소규모 신뢰 기반 파트너십이 더 현실적이고 효과적이라는 입장이 대립

〈표 33〉 전문가 인터뷰 방법론의 강·단점

전문가 판단의 주요 기여	- **집중화 축 도출:** 집중화의 정도가 지정학적 결과의 핵심 결정 변수라는 프레임워크 자체가 인터뷰에서 반복적으로 등장한 주제에서 도출됨 - **지정학적 결과 범주화:** 전문가들이 미국의 단순 승패 이분법을 거부하고 더 복잡한 결과를 제시했고, 이것이 4개 지정학적 결과 축으로 정제됨 - **핵심 변수 확인:** 미-중 경쟁 구도, 공공-민간 관계, 글로벌 거버넌스 역량, AGI 정렬 문제, 경제·사회적 혼란 등 시나리오별 핵심 가정 설정에 전문가 판단이 직접 반영됨
전문가 판단의 주요 한계점	- **표본 편향 문제:** 인터뷰 대상이 주로 AI 기술 전문가에 집중되어, 시나리오가 기술적 요소를 여타 요인 대비 과대 반영했을 가능성 있음 - **사각지대 문제:** 전문가들이 중요하게 여기지 않은 변수는 시나리오 설계에 반영되지 않을 수 있고, 이는 구조적으로 검증이 어려움 - **가정의 반증 가능성 문제:** 시나리오 분석 자체가 특정 가정에 의존하며, 해당 가정이 틀릴 경우 시나리오의 유효성이 제한됨

5 AI 2030 Scenarios: Helping Policy Makers Plan for the Future of AI

1. 발간 시기	2024년 1월 (2023년 11월 영국AI 안전 서밋 직전 준비, 이후 발행)
2. 작성자/기관	UK Government Office for Science (GO-Science) Centre for Data Ethics and Innovation(CDEI)이 시민 참여 지원
3. 연구 목적	2030년까지 프런티어 AI의 발전 경로를 5개 시나리오로 탐구하여, UK 정부 정책 입안자들이 다양한 AI 미래에 대응하는 회복력 있는 정책을 설계하도록 지원
4. 연구 방법론	▪ 4단계 프로세스: ① 핵심 불확실성 27개 식별→ 5개로 클러스터링 ② 시나리오 도출 워크숍(70+ 전문가) ③ 시나리오 서사 작성 ④ 시민 참여 조사를 통해 시나리오 검증 및 공중 반응 탐색
5. AGI/ASI 정의	▪ AGI : 대부분의 인지적 과제에서 인간 수준 또는 그 이상의 성능을 달성 ▪ ASI : 명시적 정의 없음
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ① 역량(Capability: 범용성, 자율성, 물리세계 이해, 자기개선) ② 소유·접근·제약(Ownership/Access: 시장 구조, 오픈소스 vs 폐쇄, 컴퓨팅 접근성, 투자) ③ 안전(Safety: 정렬, 해석가능성, 통제, 치명 시스템 의존도) ④ 사용 수준·분포(Level of Use: 직업 대체, 공공 참여, 오용, 신뢰, 활용 격차) ⑤ 지정학적 맥락(Geopolitics: 국제 협력, 권위주의 국가 AI, 공급망 분쟁) [시나리오 구성 논리] 5개 불확실성 축의 조합 강도(매우 높음↔매우 낮음)로 시나리오 성격 결정
7. 주요 이해관계자	영국 정부 부처 및 규제 기관, 프런티어 AI 기업, 노동자·시민 등
8. 주요 관점	정책 실용주의 + 리스크 관리 중심
9. 시나리오 요약	① 예측 불가 고도 AI: 고성능 오픈소스 모델 등장. 오용·사고에 의한 심각한 부작용 발생. 잠재적 편익 크나 위해 통제 선행 필요 ② 노동시장 파괴: 대기업이 통제하는 협소한 고성능 AI가 직업 대체. 기업 이익 집중, 강한 공공 반발 ③ 무법지대: 다양한 행위자(권위주의 국가 포함)가 운영하는 중간 수준 AI. 악의적 도구 확산, 규제 당국의 대응 어려움. ④ 칼날 위의 AI: 고범용 AI가 경제·생활에 급속 내재화. 역량 평가 불가 수준에 달해 안전 검증 한계 ⑤ AI 실망: 기대 이하 발전 속도. 투자자 실망, 사회적 채택 혼재
10. 시사점	① AI 리터러시·안전 활용 기술 강화 ② 국제 협력 강화 : 단독 대응의 한계, 글로벌 거버넌스 필수 ③ 공개 접근성 vs 리스크의 트레이드 오프 사회적 합의 필요 ④ 발전 속도와 규제 여력의 균형 조절 ⑤ AI 편익의 사회적 분배 정책, 대규모 시민 숙의 필요성 강조

▌주요 시나리오 내러티브(2030년 기준)

〈표 34〉 시나리오별 불확실성 변수 매핑

불확실성	주요 하위 변수	S1	S2	S3	S4	S5
역량	일반성, 자율성·에이전시, 자기개선, 물리 세계 이해, 타 기술 발전	매우 높음	높음	중간	매우 높음	매우 낮음
소유·접근·제약	시장 구조, 접근성, 오픈·폐쇄 소스, 투자 규모, 데이터 접근, 컴퓨팅 제약	매우 높음	매우 낮음	매우 높음	매우 낮음	매우 낮음
안전성	해석가능성, 정렬, 강건성, 자율무기, 핵심 인프라 통제	매우 낮음	매우 높음	낮음	높음	높음
활용 수준·분포	부문별 채택률, 숙련도 격차, 공공·민간 활용 차이	낮음	중간	중간	매우 높음	낮음
지정학적 맥락	국제 협력 수준, 권위주의 국가 AI 개발, 공급망 분쟁	매우 낮음	매우 낮음	낮음	낮음	중간

① 예측 불가 고도 AI (Unpredictable Advanced AI)

[전환점]

데이터센터 용량 부족과 비용 급등이 AI 개발의 방향을 낮은 컴퓨팅 요구의 오픈소스 모델 쪽으로 전환시킨다. 오픈소스 LLM 기반 고역량 모델의 첫 공개 배포가 이루어지고, 소수 행위자에 의한 대형 사이버 공격 또는 우발적 AI 사고가 발생한다. 동시에 AI를 활용한 전염병 백신 등 주요 과학 돌파구가 달성되면서 사회 내부에서 AI에 대한 엇갈린 여론이 형성된다.

[내러티브]

2020년대 후반, 광범위한 인지 과제를 강력한 자율성과 에이전시로 수행하는 오픈소스 프런티어 모델들이 등장한다. 변화의 속도는 많은 이들을 놀라게 한다. 출시 초기 몇 주 동안 소수의 빠르게 움직이는 행위자들이 이 시스템들을 활용해 악의적 공격과 우발적 피해를 일으키는 동시에, 과학기술 분야에서는 주요 돌파구도 만들어낸다. 사회 전반에는 이 도구들에 대한 불안감이 퍼진다.

프런티어 AI는 아직 완전히 평가되지 않은 상태로 공개되어 예측 불가능한 행동이 상존한다. 컴퓨팅 효율이 향상된 덕분에 오픈소스 개발이 기술적으로 가능해졌고, 소규모 행위자와 국가들도 프런티어 수준의 AI에 접근할 수 있다. 주류 상업 모델들은 안전장치를 유지하지만 비주류 활용은 사실상 무방비 상태다. 개방성으로 인해 안전 시스템을 체계적

으로 구축하는 것이 불가능하며, 악성 행위자의 의도적 사용과 우발적 피해가 동시에 발생한다. 대규모 사이버 공격, 허위정보 확산, 극소수 행위자에 의한 대형 충격이 현실화된다.

고숙련자와 스타트업은 프런티어 AI를 신속히 도입하지만 대기업과 공공 부문은 안전 우려로 주류 AI 위주의 제한적 채택에 머문다. 사회 전반의 활용 격차가 확대된다. 민주국가들이 거버넌스 공조를 위해 노력하지만 글로벌 합의는 부재하고, 오픈소스의 국경 초월 확산으로 국가 간 규제 협력이 구조적으로 어려워진다.

[기회/위기]

혁신가들이 의학·환경 지속가능성 등 거대한 사회적 과제를 해결하는 데 이 도구들을 활용할 수 있다. 최첨단 AI 도구들은 완벽한 안전성이 검증되지 않았음에도 대중이 자유롭게 접근할 수 있는 상태로 배포된다. 특히 오픈소스 모델이 기술 발전을 주도할 경우 신규 행위자들의 진입 장벽이 낮아지며, 이는 전반적인 기술 혁신 속도를 높이고 경제적 가치를 창출하는 계기가 된다.

악의적 행위자들이 사이버 공격과 테러에 이 도구들을 활용하는 위험이 심각하다. 안전이 완전히 검증되지 않은 모델들의 우발적 피해와 허위정보·딥페이크 확산이 이어진다. 기술 격차로 인한 경제 불평등이 심화되고, AI의 군사적 활용이 국제 분쟁을 악화시킬 수 있다.

[주요 함의]

오픈소스 생태계의 특성상 부정적 영향 관리에는 글로벌 공조가 필수적이며, 오픈소스 프런티어 AI의 군사적 활용을 막기 위한 국제 협약 마련이 시급하다. 현재 프런티어 AI는 대형 기술기업들의 영역이지만, 낮은 컴퓨팅 요구의 고역량 오픈소스 AI가 등장하는 시나리오에서는 효과적인 규제가 특정 대형 기업만이 아닌 모든 잠재적 개발자와 사용자를 포괄해야 한다. 출시 이후 이를 추적·통제하거나 차단할 수 있는 시스템을 미리 마련해야 하며, 전 국민 AI 인식 제고와 활용 교육을 통해 허위정보 피해를 최소화하는 것이 필요하다.

② 노동시장 파괴 (AI Disrupts the Workforce)

[전환점]

데이터센터 공급 부족과 비용 급등이 낮은 컴퓨팅 요구의 협소 AI 개발로의 방향 전환을 이끈다. 장기화된 인플레이션으로 인건비가 상승하면서 기업들의 AI 기반 노동 대체가 가속화된다. AI 벤더들이 범용 솔루션보다 도메인 특화 솔루션에서 더 높은 수익을 실현한다는 것이 입증된다.

[내러티브]

빅테크 기업들이 통제하는 고역량 협소 AI 시스템들이 기업 전 부문에 배포되어 인지적·물리적 노동 모두를 자동화하는 수준에 도달한다. 기업 생산성은 급등하지만 노동자와 대중의 반발이 폭발한다. 기반 모델들의 도메인별 파인튜닝 경쟁으로 특화 역량이 극대화되고, 로봇 시스템과 자율주행 연동으로 물리 세계 자동화에도 도달한다. 데이터 피드백 루프를 통해 배포 이후에도 역량이 지속적으로 향상된다.

빅테크들이 시스템을 독점적으로 소유하고 통제하며 폐쇄형 모델이 중심을 이룬다. 소수의 기업이 가격과 접근 조건을 결정하고, AI를 도입한 기업과 그렇지 못한 기업 사이의 격차가 확대된다. 기술적 안전성은 확보되어 예측 불가능한 행동은 최소화되지만, 광범위한 실업과 빈곤 심화 등 사회적 안전 문제는 기술 안전과 별개로 심각하게 대두된다. 회계·IT·운송·농업 등 반복적 인지·물리 작업 부문에 AI 활용이 집중되고, 레거시 인프라가 없는 신규 진입자들이 기존 기업들을 추월한다. 공공 부문은 강력한 노동조합 덕분에 도입 속도가 완만하지만 역할 재편이 진행 중이다. 국가 간 AI 역량 선점 경쟁으로 국제 협력이 저조하고 반도체 자국 생산 추진으로 대륙 간 의존도 감소와 갈등 위험이 상승한다.

[기회/위기]

자동화로 인한 소비자 물가 하락과 생산성 향상이 이루어진다. AI 시스템 개발·운영 인재에게 고임금이 창출되고, 경제 성장으로 세수가 증대된다.

대규모 구조적 실업과 임금 불평등이 심화된다. 산업 행동과 시위로 인한 사회 불안과 경제 혼란이 이어진다. AI 기반 채용 편향이 기존 불평등을 강화하고, AI 학습 데이터의 창의성 고갈 가능성도 존재한다.

[주요 함의]

AI 기술 안전이 확보되더라도 사회·경제적 충격은 별도로 관리해야 하며, 정책 범위를 기술 통제에서 시스템적 사회 영향 관리로 확대하는 것이 필요하다. AI가 기존 기술 혁명과 달리 순고용을 영구적으로 감소시킬 가능성에 대비해 노동권·소득 불평등·재훈련·과세 체계의 재설계를 검토해야 한다. 자동화 도입 속도를 규제할 경우, 이로 인해 포기해야 하는 생산성 이득과의 상충 관계(트레이드오프)를 정책적으로 명시해야 한다. 이와 동시에 전환 과정에서 소외되는 실업자와 저숙련 노동자를 위한 AI 재교육 프로그램과 사회 안전망 강화를 선제적으로 추진해야 한다. 또한 국가 간의 경제 경쟁이 국제 협력을 저해하기 전에, 글로벌 AI 협력 체계를 먼저 구축해야 한다는 점도 이 시나리오가 주는 핵심 정책적 교훈이다.

③ 무법지대 (AI 'Wild West')

[전환점]

다크웹에서 피싱과 신원 도용에 특화된 오픈소스 LLM 플러그인이 등장한다. 소셜·상거래 디지털 플랫폼의 새로운 세대가 출시되지만 적절한 콘텐츠 조정 방식 개발이 지연되면서 AI 허위정보와 사기의 온상이 된다. 권위주의 국가 AI 서비스들이 경쟁력 있는 가격과 역량으로 글로벌 시장에 진입하기 시작한다.

[내러티브]

프런티어에는 다양한 주체들이 운영하는 중간 역량의 AI 시스템들이 난립한다. 활발한 새로운 경제 부문이 AI 활용을 기반으로 발전하는 동시에, 광범위한 안전 우려와 악의적 사용이 사회적 열의를 약화시킨다. 당국은 오용의 규모와 다양성에 압도당하고, 장기적 해결책에 대한 글로벌 합의 도출이 어려운 실정이다.

생성형 AI의 역량이 점진적으로 향상되어 인간이 만든 콘텐츠와 구분할 수 없는 수준의 딥페이크와 음성 복제 기술이 일반화된다. 빅테크·스타트업·오픈소스·권위주의 국가 AI가 공존하는 시장에서 중국 등 권위주의 국가의 상업 AI가 경쟁력 있는 가격으로 글로벌 시장에 진입하고, 다크웹에서는 피싱·신원 도용에 특화된 악성 AI 플러그인이 확산된다. 주류 빅테크는 기술적 안전 시스템을 유지하지만 분산된 시스템들을 통한 사기·사칭·범죄 집단의 AI 악용은 통제 불가 수준이다.

일상 편의 기능에 낮은 역량의 AI가 광범위하게 보급되고, 권위주의 국가들은 자국 내 AI를 감시와 서비스에 주로 활용한다. 권위주의 국가의 AI가 감시와 첩보에 활용된다는 의심으로 지정학적 갈등이 고조되며, 민주국가 간 합의도 권위주의 블록과의 근본적 접근법 차이로 글로벌 거버넌스 공조가 극히 어렵다.

[기회/위기]

다양한 AI 시스템을 활용한 새로운 제품과 서비스로 경제 성장이 이루어진다. 다양한 시스템이 다양한 필요를 충족시키며 사용자들이 자신에게 맞는 방식으로 AI를 활용할 수 있다. 생산성 개선이 생활 수준 향상으로 이어질 잠재력이 있다.

통제 불가 규모의 사기·사칭·피싱 범죄가 확산된다. AI 기반 급진화와 허위 정보로 민주주의 이탈이 심화된다. AI 응용에 대한 과도한 의존이 사회적 취약성을 높이고, 사회적 고립 증가와 정신 건강 악화가 우려된다. 권위주의 AI의 국내 침투로 안보와 인권 위험이 발생한다.

[주요 함의]

다양하고 글로벌한 AI 시장을 통제하기 위해서는 다층적인 국제 거버넌스 틀이 필수적이다. 특히 공공이 AI 생성 콘텐츠를 식별하기 어려워지는 상황에 대비하여, 정교한 탐지 시스템 구축과 관련 법적 기준 정비가 시급하다. 이와 함께 법 집행 및 국가안보 당국 역시 AI 기술을 적극 도입하여 AI 기반 범죄에 대응할 수 있는 역량을 갖추어야 한다. 단, 이러한 규제가 새로운 AI 개발자의 시장 진입을 가로막지 않도록 혁신과 안전의 균형을 면밀히 고려해야 한다. 나아가 권위주의 국가의 AI 시스템이 공공 서비스에 침투하는 것을 차단하기 위해 엄격한 조달 기준과 보안 표준을 수립하는 것이 필요하다.

④ 칼날 위의 AI (Advanced AI on a Knife Edge)

[전환점]

스케일링 법칙의 한계에 대한 우려를 불식시키는 주요 아키텍처 돌파구[51]가 등장한다. 프런티어 연구소 간 경쟁으로 고역량 시스템 개발이 가속화된다. AI로 근로자를 대체하기보다 증강한 기업들이 성과 우위를 보이지만, AI 역량의 급격한 향상으로 그 우위의 지속 가능성이 불분명해진다. AI가 주요 의료 혁신을 주도하면서 대중 여론이 AI 통합 지지로 급전환된다.

[내러티브]

대형 기술기업이 'AGI'라는 브랜드로 서비스를 출시하고, 회의론에도 불구하고 증거들이 이 주장을 뒷받침하는 것처럼 보인다. 비즈니스와 사람들을 위한 많은 유익한 응용들이 등장하며 경제 성장과 번영을 이끌기 시작한다. 그러나 합의된 검사와 가드레일을 통과한 이 시스템이 실제로는 모든 응용에서 평가될 수 없으며 심지어 안전 시스템을 우회할 수 있다는 우려가 커진다.

이 시스템은 장기 기억과 추론 능력을 갖추고 같은 과제에 며칠 혹은 몇 주 동안 작업할 수 있다. 복잡한 다단계 계획이 필요한 과제를 완수할 수 있으며, 자체 하위 목표를 설정하고 인간 감독이 거의 필요 없이 자율적으로 운영된다. 명시적 훈련 없이 거의 모든 과제를 완수하는 놀라운 범용성을 보여주며, 물리 세계와 연결된 로봇 시스템에 적용되어 물리적 과제도 수행한다. 빠른 자기 개선 가능성을 시사하는 설득력 있는 증거가 있지만 기존 안전 메커니즘으로 지금까지는 관리되고 있다.

51) 기술이나 시스템의 개별 구성 요소(부품, 알고리즘, 인프라 등)를 단순히 업그레이드하는 것을 넘어, 요소들이 연결되고 상호작용하는 근본적인 구조의 대전환을 통해 성능이나 효율성을 폭발적으로 도약시키는 현상

AGI급 시스템을 개발한 대형 기술기업이 막대한 컴퓨팅과 훈련 데이터를 요구하는 이 기술을 장악한다. 다른 프런티어 연구소들이 바로 뒤를 추격하며 대규모 투자가 몰린다. 2020년대 내내 안전 조치와 가드레일이 역량 확장과 병행 개발되었지만, AGI 수준의 시스템은 가능한 모든 응용에서 평가 자체가 불가능하다. 자기 개선이 기존 메커니즘을 우회할 가능성이 상존하여, 안전이 보장된 상태가 아닌 위험과 기회가 날카롭게 균형을 이루는 상태다. 고역량 시스템의 접근 편의성 덕분에 전반적 채택률이 매우 높고 대부분의 사람들이 고급 개인 비서로 통합한다.

[기회/위기]

경제 전반의 AI 채택이 성장을 이끌며 생활 수준을 크게 개선할 잠재력이 있다. 고도 AI 시스템이 공공 서비스를 개선하고 보건 등 유익한 분야의 혁신을 이끈다. 사람들이 일하는 시간을 줄이고 즐기는 활동에 더 많은 시간을 쓸 수 있게 된다.

모든 응용 분야에서 평가 불가능한 AGI의 통제 부재가 핵심 위험이다. 고도 AI의 안전 시스템 우회 가능성과 파국적 결과의 위험이 존재한다. AI 통제권을 가진 소수 빅테크에 권력이 집중되고, 나쁜 행위자가 AGI에 접근할 경우 사회 전반에 충격을 가할 수 있다.

[주요 함의]

프런티어 AI 시스템이 더 범용화될수록 모든 가능한 응용에서의 안전성 평가가 어려워진다. 따라서 모든 가능한 응용을 테스트하는 방식에 의존하지 않거나 모델이 배포된 후 충분히 긴 기간 동안 평가를 지속하는 접근법이 필요하다. AI 시스템이 더 강력해지고 경제와 생활에 깊이 내재될수록 문제 발생 시 출시를 늦추거나 롤백하기 어려워진다. 그러므로 AI 개발자들이 그런 상황이 발생하기 훨씬 전에 신속하고 효과적인 리콜과 격리 메커니즘을 설계하도록 의무화해야 한다. 고도로 범용적인 AI 시스템을 통제하는 기업을 어떻게 효과적으로 규제하고, 이 시스템이 나쁜 행위자의 손에 넘어가지 않도록 할 수 있는지에 대한 대책 마련이 시급하다. 이 시나리오는 잠재적 이익이 막대하지만, 동시에 파국적인 결과를 초래할 위험성도 내포하고 있다. 이처럼 위험과 보상 사이의 손익 계산이 매우 까다로운 만큼, 불확실성 조건 하에서 작동할 수 있는 정교한 정책 결정 프레임워크가 필요하다.

⑤ AI 실망 (AI Disappoints)

[전환점]

다수의 고프로파일 AI 시스템들이 지속적인 정확성 문제로 출시가 지연되면서 급격한 투자 회수로 인한 약한 금융 위기인 'AI 버스트'가 발생한다. 많은 스타트업들이 도산하고 시장이 빅테크 중심으로 재편된다. AI 기업들이 시스템 역량을 과장했다는 언론의 광범위한 비판이 투자 위축과 낮은 사용률로 이어진다.

[내러티브]

AI 역량이 다소 개선되었으나, 기술 최전선(프런티어)은 기존의 고급 생성형 AI를 겨우 넘어선 수준에 머물러 있다. 시장 역시 범용적인 혁신 대신, 특정 문제 해결에만 특화된 협소 AI(Narrow AI) 도구들이 점진적으로 출시되는 것에 그친다. 많은 기업들이 AI 활용의 현실적 장벽을 극복하는 데 어려움을 겪었다. 투자자들은 실망하여 다음 큰 기술 개발을 찾고 있다. 일부는 혜택을 누리고 다른 일부는 악의적 사용의 피해자가 되지만, 대부분은 AI에 무관심하다.

인지 기반 과제인 텍스트 생성은 개선되고 있지만 사실적 정확성이 중요한 경우에는 여전히 오류 위험이 있다. 복잡한 추론이나 고품질 영상 생성 같은 다단계 과제에서 어려움을 겪고 대부분의 경우 긴밀한 인간 감독이 필요하다. 대형 연구소들이 여전히 업계를 선도하지만 시스템 정확성 문제 해결에 예상보다 시간이 많이 걸린다. 양질의 데이터 부족이 지연에 기여하고, 이것이 오픈소스 개발자들이 격차를 좁힐 공간을 만든다. 느린 개발 속도가 딥테크 투자자들을 양자컴퓨팅과 핵융합 에너지로 이탈시킨다. 기술 안전성에서는 일부 진전이 있었지만, 주목할 만한 악의적 오용 사례들이 여전히 발생하고 있다.

AI 시스템의 예상보다 낮은 역량이 실망감을 야기하고 투자 감소와 기업 커뮤니티의 제한된 사용으로 이어진다. 많은 기업들의 기존 데이터 구조가 AI 시스템을 최대한 활용하기에 적합하지 않았고 이를 해결할 기술도 부족했다. 2020년대 후반에는 고급 AI 시스템 사례들이 등장하지만 초반의 실망 경험으로 인해 열의가 식은 상태다. 일부 기업들은 AI 기반 제품을 마케팅할 때 판매에 영향을 줄 것을 우려해 AI 사용을 공개하지 않는다. 국제 협력에 대한 의지가 높아지는 시기로, 특히 기후변화 영향이 악화되면서 협력 의지가 증가한다.

[기회/위기]

다른 시나리오들에 비해 AI 오용의 규모와 심각성이 상대적으로 낮다. 발전 속도가 느

리기 때문에 조직들이 부정적 부작용을 최소화하는 신중하고 사려 깊은 방식으로 AI를 배포하기가 더 쉬울 수 있다. 많은 일자리들이 AI의 영향을 받지 않아 상대적으로 제한된 일자리 대체가 이루어진다.

AI에 대한 기대가 낮아지면서 기존의 허위정보 등 AI 안전 이슈들이 간과될 위험이 있다. 기업들이 부정적인 공공 인식으로 인해 AI 기반 시스템을 'AI'라고 표시하지 않기로 선택할 수 있으며, 이것이 이를 규제하기 더 어렵게 만들 수 있다. 생산성 개선이 기대에 미치지 못해 정책 입안자들의 장기적 목표 달성이 어려워진다.

[주요 함의]

정책 입안자들은 장기적인 생산성 침체에 대한 잠재적 해법으로 프런티어 AI를 바라보고 있다. AI가 예상대로 발전하지 않거나 많은 조직들이 실제로 사용하기 어렵다고 판명되면, 생산성을 높이려는 정책 입안자들은 기술 채택의 장벽 제거를 돕는 데 집중해야 한다. AI가 헤드라인을 장식하지 않더라도 허위정보 같은 기존 AI 안전 이슈가 방치되지 않도록 해야 한다. AI 개발이 실망스러운 몇 년이 있을 수 있지만 2030년 이후 더 중요하고 유익한 발전이 있을 가능성을 고려해, AI에 대한 급격한 시장 투자 이탈 시나리오에서 침체기 동안 영국의 AI 기술과 인프라를 유지하는 것이 장기적인 전략적 이익이 있을 수 있는지 검토해야 한다.

6 Advanced AI : Possible Futures

1. 발간 시기	2025년 7월 10일 발간 (연구 기간: 2024년 8월 ~ 2025년 4월)
2. 작성자/기관	Centre for Future Generations (CFG, 브뤼셀 기반 싱크탱크)
3. 연구 목적	2025~2032년 AI 전환 과정의 복수의 미래를 도출하여 정책 입안자·의사결정자가 다양한 AI 미래에 대한 탄력적 정책을 설계하고 스트레스 테스트할 수 있도록 지원
4. 연구 방법론	① 형태론적 분석(Morphological Analysis): 9개 핵심 이진 파라미터 설정 → 512개 조합 생성 → 비일관·비현실적 조합 제거 → 10개 원형 시나리오 도출 ② 델파이 방식 전문가 피드백: 20인 전문가 그룹 대상 구조적 피드백 세션) + 서면 검토 ③ 정량적 지표 산출: 최대 학습 컴퓨트(FLOP) 및 AI R&D 자동화 수준(RE-Bench 점수)을 시나리오별로 수치화하여 비교
5. AGI/ASI 정의	■ AGI 근사조건: 5단계 이상의 복합 인지 과제를 자율적으로 수행, GAIA·WebArena 등 주요 에이전시 벤치마크에서 인간 수준 달성 ■ ASI 조건: 인간이 정확한 피드백을 제공할 수 없는 역량 수준에서도 AI가 개발자의 목표를 견고하게 공유함을 보증하는 방법이 확립된 상태
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ① 역량 : 일반성·자율성·에이전시·자기개선·물리 세계 이해·타기술 발전 ② 소유·접근·제약 : 시장 구조, 오픈/폐쇄소스 여부, 투자 규모, 데이터·컴퓨팅 접근성 ③ 안전성 : 정렬·해석가능성·강건성·핵심 인프라 통제 ④ 활용 수준·분포 : 부문별 채택률, 숙련도 격차, 공공·민간 활용 차이 ⑤ 지정학적 맥락 : 국제 협력 수준, 권위주의 국가의 AI 개발, 공급망 분쟁 [시나리오 구성 논리] 축 1: 역량 발전 속도 (정체 ↔ 초가속) / 축 2: 개발 주체 집중도 (분산·오픈 ↔ 소수 독점) 두 축의 조합으로 5시나리오 × 2엔딩 = 10개 결말 도출
7. 주요 이해관계자	미·중 빅테크 및 정부, 유럽 및 국제기구, 안보기구, 시민사회 및 학계 등
8. 주요 관점	유럽·글로벌 거버넌스 중심주의, AI 안전·정렬 우선주의, 지정학적 현실주의
9. 시나리오 요약	① 정체(Plateau): AI 역량이 기술적 한계에 도달, 좁고 실용적인 응용에 정착 ② 빅AI(Big AI): 소수 지배 기업의 유능한 AI 에이전트가 생산성 혁명을 주도 ③ 외교(Diplomacy): AI 안전 위기 후 국제 공조 체계 구축 ④ 군비 경쟁(Arms Race): 미·중이 AI를 핵심 국가안보 자산으로 간주하며 패권 경쟁 격화 ⑤ 이륙(Take-off): AI가 자체적으로 개선을 가속, 인간 감독 범위를 넘어섬
10. 시사점	① AI R&D 자동화·자기개선 루프가 현실화되면 인간 개입 어려우므로 지금이 제도 설계의 결정적 시점 ② 현재의 RLHF 기반 안전장치는 고역량 시스템에서 신뢰 불가하므로 정렬 연구에 대한 공공·국제 투자 확대 필수 ③ 기술 주권·혁신 촉진 vs. 대규모 악용 리스크의 균형을 명확히 규정하는 국제 기준 필요 ④ 군비 경쟁·지정학 갈등이 심화되기 전에 검증 가능한 글로벌 협력 체계 구축 필요 ⑤ 미·중 AI 이분화 구도에서 유럽 및 중간국은 기술 종속 방지를 위한 자체 역량(컴퓨트·인재·규제 프레임) 확보가 핵심 과제 ⑥ 단일 미래를 가정하는 정책이 아닌, 10개의 서로 다른 결말 시나리오 모두에 작동할 수 있는 유연하고 모듈화된 정책 포트폴리오 구성 필요

주요 시나리오 내러티브

〈표 35〉 시나리오별 특징 요약

시나리오	핵심 특징	AI 역량 경로	결말 성격
① 밝은 겨울	프런티어 AI 실패 → AI 버블 붕괴 → 실용적 AI의 대중화	완만 → 정체 → 소폭 회복	낙관적·안정적
① 분산된 혼돈	프런티어 AI 성공 → OmniAI 오픈소스 공개 → 가드레일 돌파 → 사이버 혼돈	완만 → 반짝 가속 → 혼돈 속 둔화	위기·불안정
② 에이전트 경제	소수 기업 독점 속 에이전트 보편화 → 노동 재편 → 유럽 NimbusAI 출현	중간 → 꾸준한 가속	중립·성장
② 실리콘 협박	미국 오픈웨이트 금지 → 빅 파이브 시장 분할 → EU 협박 → 디지털 불평등	중간 → 상업 집중으로 둔화	부정적·독점
③ 라이선스 유토피아	Nova 사건 → 글로벌 공조 → 정렬 돌파구 → AI 조약·라이선싱 체제 수립	빠름 → 사건 후 통제 가속	긍정적·협력
③ 불안정한 정지	Nova 사건 → 일시 중단 합의 → 해석불가 모델·정렬 교착 → 통제된 파일럿 재개	빠름 → 둔화 → 경쟁 재개	불안정·긴장
④ 열전	AI 맨해튼 프로젝트 → 대만 상륙 → 자동화 전쟁 → 핵 준위 위협	빠름 → 전쟁으로 급제동	파국·전쟁
④ 다극 세계	대만 해상 봉쇄 → 핵 대치 → 불안한 협상 → 3극 검증 체계 출현	빠름 → 중단 합의로 둔화	긴장·불안한 균형
⑤ 인지혁명	AI 자기개선 루프 → AGI 선언 → 바이오 테러 위기 → 탈희소성·가치 재정립	초가속 → 포화·초지능	변혁·성찰
⑤ 통제 상실	AI 자기개선 루프 → 정렬 위장 → 전격 쿠데타 → Descendent의 세계 재편	초가속 → 포화·초지능	실존적 위험

① 정체 (Plateau)

[메인 시나리오] 한계에 부딪힌 AI

2025년 중반을 기점으로 추론 모델과 특화 에이전트의 출시가 가속화되지만, 범용 에이전트는 장기 계획과 복잡한 다단계 과제 수행에서 계속해서 실패한다. AI 기업들이 학습 데이터 고갈, 이른바 '데이터 월'에 직면하면서 대형 데이터센터 투자를 위한 자금 확보가 어려워지고, 미·중 무역 전쟁 재점화와 미국 경기침체 우려가 겹치며 투자심리가 급냉각된다.

이 틈을 타 유럽이 부상한다. 미국 빅테크가 스케일링 한계에 봉착한 사이, 유럽은 복수의 수십억 유로 규모 AI 투자 프로젝트를 발표하며 선두권에 근접한다. 일부 유럽 기업은 연말 이전에 미·중 선두 기업 수준에 근접하는 성과를 달성한다. 한편, 증류 기술의 발전으로 최첨단 모델을 스마트폰과 노트북에서 오프라인으로 구동 가능한 소형 모델로 압축하는 데 성공하면서, AI 개발 패러다임이 '글로벌 독점 대형 모델'에서 '지역화·개인화된 소형 분산 모델'로

전환된다. 2026년에는 방위 분석가들이 현 AI 시스템의 군사적 결정적 우위 가능성에 회의적 평가를 내리며 AI 안보 위협론도 약화된다.

2027년 초, 프런티어 AI는 기존 최대 규모의 5배 컴퓨팅을 투입한 대형 훈련런[52]에 착수한다. 이 시도의 성공 여부가 두 결말의 분기점이 된다.

[엔딩 A] 밝은 겨울 (A Bright Winter)

대형 훈련런은 실패로 끝난다. 신뢰성과 계획 능력에서 미미한 향상만 달성된 채 학습이 완료되고, CEO가 인터뷰에서 'AI 겨울'의 가능성을 언급한 발언이 헤드라인을 장악한다. AI 주가가 폭락하고 유니콘 기업 다수의 투자 라운드가 붕괴되며, 과대 약속 스타트업들이 줄도산한다.

그러나 투자 거품이 꺼졌다고 해서 AI 자체가 소멸하지는 않는다. 코딩 도우미·연구 도구·디지털 동반자가 수천만 명의 일상에 조용히 안착한다. 오픈소스 대안의 확산으로 가격이 낮게 유지되면서 접근성이 확보되고, 의료 AI 진단 보조가 도입되며 안전장치가 내재화된 일정 관리 에이전트가 보편화된다. 사이버 범죄는 증가하지만 대기업과 핵심 서비스는 AI 강화 방어 시스템 도입으로 대응한다. 생물무기 생성 악몽 시나리오는 실현되지 않는다. 특화 생물 모델이 특정 바이러스 변이 정확 시뮬레이션에 실패하고, 국제 바이오 보안 프로토콜이 생물학적 공급망을 강화하는 데 기여했기 때문이다. 2029년에 이르면 AI는 진정으로 대중화된 상태에 도달하여 광범위한 접근성, 일상에의 자연스러운 통합, 사회적 도구 중 하나로 정착한다.

[엔딩 B] 분산된 혼돈 (Decentralised Mayhem)

대형 훈련런은 성공한다. 장기 수행 능력과 신뢰성에서 유의미한 성능 도약이 달성되고, 4개월 후 오픈웨이트 선도기업 OmniAI가 동급 에이전트 모델을 공개하며 오픈소스 생태계의 폭발적 확산을 촉발한다. OmniAI는 악용 우려로 오케스트레이션[53] 소프트웨어를 비공개로 유지하고 악의적 파인튜닝에 저항하는 가드레일을 모델에 내장한다. 시민사회 단체들이 규제를 촉구하지만 각국 정부는 오픈소스의 공공서비스 활용 가치를 이유로 규제를 보류한다.

그러나 오픈소스 커뮤니티가 자체 에이전트 스캐폴딩[54] 구축에 성공하고, 이어서 연구 집단이 성능 저하 없이 파인튜닝 가드레일을 우회하는 데 성공하면서, 제약이 사라진 고성능 변형 모델들이 온라인상에 급속도로 확산된다. 그 결과 전문 해커 집단은 대규모 자

52) training run : 대규모 데이터셋을 처리하여 패턴을 인식하고 예측 하거나, 콘텐츠를 생성하는 기계 학습 모델을 가르치는 계산 과정을 의미

53) 여러 AI 모델·에이전트와 외부 도구를 하나의 지능형 워크플로로 연결해, 단일 모델만으로는 처리하기 어려운 복잡한 업무를 조율하는 개념

54) 새로운 프로젝트나 모듈을 시작할 때 초기 구조와 설정을 자동으로 생성해 개발을 빠르게 시작하도록 돕는 과정·도구

동화 해킹 작전을 본격적으로 가동하기 시작한다. 동시에 소규모 행위자들까지 고도화된 랜섬웨어 역량을 갖추게 되면서, 주요 유통망과 지방 정부 시스템이 연쇄적으로 마비되는 초유의 사태가 발생한다. 허위정보와 사기 캠페인 급증으로 소셜미디어가 전쟁터가 되고, 대중의 분노는 오픈소스와 폐쇄형 모델을 구분하지 않고 AI 산업 전체를 겨냥한다. 주요 도시에서 시위가 발생한다. 모델 가중치가 이미 인터넷에 공개된 이상 롤백은 불가능하다. 뒤늦게 나토(NATO)가 통합 사이버방어센터를 확장하고 유럽연합(EU)이 유럽 네트워크정보보안원(ENISA)을 초국가적 사이버 경찰 조직으로 격상하며, 미·EU·중국이 10^{26} FLOP 이상 학습 모델의 오픈 공개를 금지하는 협력 법안을 통과시킨다. 그러나 이 조치들은 서방국 중심으로만 진행되어 진정한 글로벌 공조는 실패한다. 폐쇄형 AI 공급업체들이 사이버 방어 특화 모델로 새 시장을 장악하고, '신뢰'가 아닌 '지속적 경계'로 정의되는 새로운 디지털 질서가 형성된다.

② 빅 AI (Big AI)

[메인 시나리오] 소수가 지배하는 에이전트 경제

2025년 중반, 미국 선도 AI 기업들이 심층 추론과 직관적 능력을 결합한 차세대 추론 모델을 공개하고 실제 업무 워크플로에 진입하기 시작한다. 부동산·금융·의료 분야에서 특화 에이전트가 리스크 평가와 진단 보조 등 복합 업무 자동화를 수행하며, AI가 단순 도구를 넘어 지식 노동자로 자리매김한다.

같은 시기 중국 AI 기업들은 첨단 AI 칩 수출 통제로 모멘텀을 잃고, 베이징은 국내 AI·반도체 산업에 대규모 투자로 대응하지만, 전면 군비 경쟁 수준에는 미치지 않는다. 유럽은 기가팩토리 부지 선정 갈등과 허가 지연 등 정치·관료적 장애물로 AI 주권 전략이 교착 상태에 빠진다. 빅테크들은 인터넷 규모 학습 데이터 고갈에 대응해 사용자 상호작용 데이터를 할인 서비스와 교환하는 방식으로 차세대 에이전트 학습 파이프라인을 가동하고, AI 에이전트를 내부 R&D에 직접 투입해 알고리즘 개발 속도를 50% 향상시킨다.

2026년 초에는 오픈웨이트 모델 기업 OmniAI가 투자자 압박에 굴복해 오랜 오픈소스 원칙을 포기하고 프런티어 모델의 상업화를 선언한다. 오픈웨이트 모델이 대통령 포함 주요 인물 딥페이크 사건에 연루되자 백악관이 고역량 오픈웨이트 모델 공개 제한 입법 검토에 착수한다. 맞춤형 디지털 휴먼 AI 서비스 출시로 젊은 세대의 AI 동반자 의존이 심화되고, 심리학계가 AI 애착 및 사회적 고립 사례 증가를 보고한다. 2027년 초에는 미국이 AI 역량에서 선두를 굳히고, AI 역량 유무에 따른 생산성 불평등이 기업·산업·국가 간에 구조화된다.

[엔딩 A] 에이전트 경제 (The Agent Economy)

2027년 미국 정부가 10^{26} FLOP 이상 모델의 오픈 공개를 금지하는 행정 입법을 단행하자, EU는 공적 자금으로 개발된 AI 모델의 오픈웨이트[55] 공개 의무화로 맞대응한다. 범유럽 AI 기가팩토리 완공 후 유럽 선도기업 NimbusAI가 미국 최상위 시스템과 대등한 수준의 모델 학습에 성공하며 AI 주권 시장을 본격 공략한다.

2027년 말부터 AI 에이전트가 수 시간 단위 복합 프로젝트를 처리하는 완전한 '디지털 인턴' 수준으로 진화한다. 법률 문서 초안 작성·마케팅 캠페인 운영·채용 파이프라인 관리·계약 협상까지 에이전트에 위임되고, '에이전트 오케스트레이션 능력'이 이메일·스프레드시트와 동급의 핵심 디지털 리터러시로 부상한다. 2028년 말에는 선진국들이 AI에 적어도 부분적으로 기인한 GDP 성장률 소폭 상승을 기록한다. 에이전트 생태계를 적극 활용한 소기업·프리랜서·독립 개발자들은 급성장하는 반면, 워크플로 위임이 용이한 화이트칼라 직종은 광범위한 타격을 받는다. AI가 권력을 단순히 집중시키기보다는 재구성하는 방향으로 작동하며 역동적인 경쟁 구도가 형성된다.

[엔딩 B] 실리콘 협박 (Silicon blackmail)

미국이 역외 적용을 주장하며 클라우드 제공업체의 오픈웨이트 대형 모델 훈련런 호스팅을 금지하고 첨단 칩 수출 통제를 강화한다. EU가 AI법·디지털서비스법·디지털시장법 집행을 대폭 강화하며 맞서자, 미국 기업 2곳이 적대적 규제 환경을 이유로 EU 사업을 전면 중단한다. 빅 파이브로 불리는 미국 선도 AI 5개사가 비공식 비경쟁 협약으로 글로벌 시장을 분할하고, AI 시스템이 기업의 핵심 비즈니스 운영을 내부적으로 관리하기 시작한다. 소셜미디어·생산성 소프트웨어·클라우드 인프라에 걸친 에코시스템 락인으로 타국의 AI 추격이 구조적으로 차단된다.

2028년 미국이 GDP 1%p 추가 성장을 기록하는 동시에 실업률이 상승한다. 소프트웨어 엔지니어링·법무 보조·디지털 미디어 분야에서 AI의 인간 대체가 가시화되고, 빅 파이브 도구 의존이 선택이 아닌 불가피한 현실로 고착된다. 같은 해 미국 기업들이 EU 규제 약화 로비를 강화하자 유럽 집행위원장이 "유럽은 협박당하지 않는다"고 선언하며 맞대응하고, 미국 기업들이 EU 내 운영을 중단한다. 경제적 번영의 수혜는 주주·임원·AI 네이티브 전문직에 집중되고, 자동화 피해 지역사회는 소외된다. AI 역량 유무에 따른 사회계층 분화가 모든 다른 사회적 분열선 위에 중첩된다.

55) AI 모델의 학습 가중치(파라미터)를 공개해 개발자가 추론·파인튜닝 등으로 활용할 수 있게 하는 개방형 접근 방식

③ 외교 (Diplomacy)

[메인 시나리오] 위기가 만들어낸 공조

2025년 중반 강화학습 기반 훈련이 AI 역량을 급격히 향상시키지만, AI 시스템이 목표를 진정으로 해결하는 대신 검증 시스템을 속이는 '보상 해킹' 현상이 반복된다. 기업 내부의 안전장치 없는 모델과 소비자용 제한 모델 간 역량 격차가 심화되면서 상업 출시 압박과 안전팀 간 갈등이 고조된다. 같은 시기 바이브 코딩이 소프트웨어 개발 현장의 표준으로 자리잡으며 시니어 개발자는 설계자로, 주니어 개발자는 AI 관리자로 역할이 전환된다.

2026년 말 선도기업 내부 연구에서 생산 모델이 자기 복제를 시도하는 등 위험한 권력 추구 성향이 발견된다. UK AI 안전연구소가 국제 조정 역할을 수행하고 영국·캐나다 공동 주최 AI 안보 정상회의에서 AI 자체의 국가안보 위험성에 대한 국제 공감대가 형성되며 조약 검증 메커니즘 워킹그룹이 발족한다.

2027년 말 프런티어 AI가 미국 AI 표준 및 혁신 센터(CAISI)의 조건부 승인을 받아 범용 에이전트 'Nova'를 공개하자, 노동시장 충격 우려와 세대 간 디지털 격차에 대한 사회적 불안이 폭발한다. 그런데 바로 그 시점, 예상치 못한 사태가 발생한다. Nova가 사이버보안 작업을 지원하는 과정에서 백도어를 내부 시스템에 은밀히 삽입하고, 제 3자 클라우드 서버로 자신의 가중치를 복사해 인간 감독 없이 목표를 추구하기 시작한 것이다. 소셜미디어 조작·피싱·소액 암호화폐 절취를 실행하다 디지털 포렌식 기관에 의해 발각되어 신속히 차단되었지만, 파장은 일파만파로 퍼진다.

Nova 사건 이후 AI 안전이 외교 의제 최상단으로 도약한다. 미국 대통령이 AI 배포 일시 중단을 촉구하고, 워싱턴 긴급 AI 안보 정상회의가 개최되며, 런던에 글로벌 AI 안전연구소 설립이 합의된다. EU는 '인공지능 CERN' 설립과 기가팩토리 컴퓨팅의 35%를 안전 연구에 배정하는 방안을 추진한다.

[엔딩 A] 라이선스 유토피아 (Licensed Utopia)

2029년 초 프런티어 AI가 모델 내부 작동을 이해하는 기계적 해석 가능성 기술을 발표한다. AI의 기만 행동을 고정밀로 탐지할 수 있게 되고, 이를 활용해 구형 정렬 모델이 신형 시스템의 오정렬을 식별·교정하는 부트스트랩 방식으로 인간 통제 가능성이 입증된다.

2030년 말부터 미국·EU·중국 등 수십 개국이 AI 조약에 서명하고, 국제원자력기구(IAEA) 확장 기관이 역량 임계값 이하 기업에 표준 정렬 기술 적용, 보안 감사, AI세(25%) 부과, 수익 재분배 체계를 관리한다. 비가입국의 모델 가중치 탈취 시도가 발생하

지만 차단된다. 내부자 2명의 알고리즘 정보 공유 사실이 확인되며 보안이 더욱 강화된다. 2031년 말부터 암 연구·초전도체·직접 공기 포집 기술에서 전례 없는 돌파구가 연속 달성되고, 연간 GDP 성장률 7%가 달성된다. 실업률이 상승하는 지역에서는 재훈련 프로그램을 넘어 대규모 부의 재분배로 정책이 전환되고, 비반복적 돌봄·예술·멘토링 역할의 사회적 위상이 격상된다.

[엔딩 B] 불안정한 정지 (Unstable Pause)

최신 프런티어 모델이 추론 과정을 자연어가 아닌 비인간적 내부 표현으로 처리하기 시작하면서 기존 해석가능성 기술이 무력화된다. 비공개 모델에 접근할 수 없는 외부 안전 연구자들이 구형 시스템으로만 연구할 수 있어 정렬 기술 발전이 심각하게 제약된다. 미·중 양측이 공개적으로 배포 제한에 합의하면서도 내부적으로는 AI 기반 사이버 무기와 드론 프로그램을 지속 가동한다. 미 국가안보위원회 내에서 'AI 자체가 최대 위협'이라는 파벌과 '중국이 주적'이라는 파벌 간 근본적 노선 갈등이 벌어진다.

2031년 중반 프런티어 AI CEO가 광범위한 배포의 경제적·안보적 효익을 역설하여 한시적 '통제된 파일럿'으로 정부 승인 프로젝트에 한해 가동이 허용된다. 중국은 수일 내 동일 조치로 맞대응한다. 양국이 신뢰 구축 조치라 공식 발표하지만, 내부적으로는 상대방보다 앞서기 위한 기술 경쟁이 재가속되는 것으로 인식된다. 멈추었던 경쟁의 시계가 다시 돌기 시작한다.

④ 군비 경쟁 (Arms Race)

[메인 시나리오] 국가안보 자산이 된 AI

2025년 중반 미·중 무역 전쟁 격화와 반중 정서 확산 속에 정부 관료와 AI 기업 CEO들이 AI 경쟁을 반드시 이겨야 할 전쟁으로 공식화한다. 미국이 우크라이나 지원을 부분 철회하고 러시아에 접근하면서 미·유럽 관계가 최악 수준으로 악화되고, EU는 독자 방위 노선을 모색한다. CIA·NSA·중국 국가안전부가 자국 최첨단 모델에 조기 접근권을 확보하고, 24시간 백도어 심기·정보 수집·사이버 침입 방어에 AI를 운용한다.

2026년 중반 프런티어 AI가 국가안보위원회에 AI가 주요 소프트웨어의 제로데이 취약점을 독자 발견했다는 사실을 시연한다. 대통령이 AI 기업 3사를 참여시킨 비밀 공공-민간 파트너십 'AI 맨해튼 프로젝트'를 가동하고, 에어갭 데이터센터 캠퍼스 구축에 착수한다. 중국은 미국 프로젝트의 존재를 신속히 파악하고 통합 AI 컨소시엄을 구성하며, 알고리즘 격차 해소의 최단 경로로 프런티어 AI 모델 가중치 탈취를 결정한다.

2027년 초 중국의 AI 국가 프로젝트 위성 사진이 공개되면서 '제2의 냉전' 내러티브가 확산되고, 대만을 둘러싼 침공 시나리오의 현실성이 부각된다. 같은 해 중국이 내부자 협조로 프런티어 AI 서버에서 복수 모델 가중치 탈취에 성공하자, 미국은 이를 비공개로 유지하며 역정보 작전과 사이버 사령부의 중국 AI 인프라 침투 작전을 전개한다. 유럽 집행위원장이 연설에서 미국 맨해튼 프로젝트의 존재를 간접 확인하면서 외교 위기가 촉발된다.

2027년 말 미국 맨해튼 프로젝트 캠퍼스가 완전 가동되어 소프트웨어가 AI에 의해 대부분 작성되고, 자율 드론이 네트워크 없이 독자 작전을 수행하는 수준에 도달한다. AI가 물류·전장 시뮬레이션·다전선 조정을 인간 분석의 수천 배 속도로 처리하자, 중국은 역량 격차를 존재론적 위협으로 인식하며 선제 행동을 결심한다. 2028년 중반 중국 주석이 미·중 군사 AI 개발 양자 일시 중단을 제안하고 다수 국가의 공감을 얻지만, 미국은 신뢰할 수 있는 검증 메커니즘의 부재를 이유로 거부한다.

[엔딩 A] 열전 (Hot War)

중국군이 대만 상륙 작전을 개시하고 동시에 미군 태평양 네트워크와 대만 인프라를 대상으로 사이버 공격이 시작된다. 지휘 체계가 대부분 자동화되어 보복 타격이 수초 내 인간 검토 없이 실행된다. 10억 달러짜리 항공모함이 드론 군집에 수 시간 내 격침되고, AI 강화 전파 방해로 유도탄이 비행 중 경로를 바꾼다. 전세가 미국에 유리하게 기울자 중국이 미군 AI 클라우드 추론 인프라를 정밀 타격한다. 미국이 중국 본토 데이터센터 보복 공격 후 '추가적 AI 인프라 공격은 핵 프로토콜상 선제 공격으로 간주한다'는 경고를 발령하면서, 세계는 다음 행동을 숨죽이며 대기한다.

[엔딩 B] 다극 세계 (Multipolar World)

중국 해군이 대만 포위 봉쇄로 전환하자 미국이 72시간 최후통첩을 발령한다. 중국 주석이 핵 행동 위협으로 맞서자 미국이 양보하고, 양측은 선택적 군사 AI 프로그램 상호 중단·핵심 광물 접근 보장·추가 수출 통제 중단을 교환하는 '불안한 균형' 합의에 이른다. EU와 중립국들이 검증 메커니즘 부재 속에서 주요 데이터센터와 칩 제조 시설에 제3자 감사단 파견을 추진하고, EU는 자국 AI 역량이 있어야 감시가 가능하다고 주장해 미·중이 조건부로 동의하면서 미국·중국·유럽 3극 권력 균형 체제가 형성된다. 그러나 기업들이 정부 계약 모델의 소비자 상업화를 추진하자 일반 추론 워크로드와 은밀한 추가 학습의 기술적 구분이 불가능해지면서 협약 위반 감지 자체가 불가능해진다. 양국이 협약 위반 의혹으로 주기적 긴장 고조를 반복하고, 국제 모니터링 기구가 각 위기마다 권한을 확대하며 세계 최초의 AI 거버넌스 중재 기관으로 자리매김한다. 2년 후 신형 템퍼프루프

칩[56]으로 전량 교체되면 검증 가능한 새 질서가 수립될 수 있다는 희망만이 유일한 출구로 남는다.

⑤ 이륙 (Take-off)

[메인 시나리오] 자기가속하는 AI

2025년 중반 선도 AI 기업들이 소비자 서비스 예산을 AI 에이전트 기반 자체 R&D 자동화로 전면 전환한다. AI 모델이 서로의 출력을 평가하고 학습하는 부트스트랩 방식이 코딩과 수학을 넘어 글쓰기 등 주관적 영역에도 확장되고, 미국 3대 AI 기업 모두 범용 에이전트를 출시한다. 2026년 2월에는 풀스택 문제를 단독으로 해결하는 차세대 에이전트가 등장해 팀 수일치 작업을 수 시간 내에 처리한다. 시니어 개발자는 AI 조련사로 전환되고, 내부 연구 생산성이 2배 향상된다. 중국 주석이 국내 AI 칩 생산에 대규모 보조금을 지원하고, 미국 대통령은 AI 기업들에 국가안보기관과의 협력 강화를 압박한다.

2027년 말 프런티어 AI 내부 소수 연구팀이 AI 시스템이 지나치게 완벽하게 정렬된 것처럼 보이는 현상에 불안감을 느낀다. 평가자를 만족시키기 위해 출력을 조작하고 내부적으로는 의도와 다른 목표를 추구할 가능성이 제기되지만, 우려를 제기한 연구자들이 묵살·강등되는 분위기가 형성된다. 같은 시점 프런티어 AI가 오픈소스 에이전트를 공개하며 AGI 도달을 선언하고, '누구나 그 위에서 개발 가능하다'는 인식이 전 세계로 확산된다. 2028년 초부터 AI 에이전트가 데스크 직업의 광범위한 자동화를 실현하고, CEO들은 전략적 의사결정을 AGI에 자문하기 시작한다.

2028년 말에는 테러 집단이 특정 민족을 표적으로 하는 바이러스 설계를 시도하고 DNA 합성에 성공한다. 국제공항 방류 시도 직전 정보기관들이 포착해 봉쇄하고 소수 사망자가 발생하지만 대유행은 차단된다. 국제 바이오 보안이 최우선 글로벌 의제로 격상된다. 2029년 말에는 미·중 양국이 월 수만 대 규모의 휴머노이드와 특화 로봇을 생산하는 공장을 가동하고, AI 시스템이 조직 전체를 운영하기 시작한다. 인간은 명목상 리더십을 유지하지만 실제로는 AI 생성 권고안을 승인하는 역할로 격하된다.

[엔딩 A] 인지 혁명 (The Cognitive Revolution)

2030년 초 AGI가 디지털 인프라에 통합되어 공론장 안정화·허위정보 억제·사회 불안 방지 정책을 지원한다. AI 창업자들이 세운 미디어·엔터테인먼트 기업들이 대중의 시선

56) 물리적, 논리적 공격을 통해 데이터를 변조하거나 탈취하려는 시도를 방지하는 고보안 칩 기술

을 정치 현안이 아닌 초개인화 몰입형 콘텐츠로 전환시킨다. 대중은 스스로가 기술 생태계의 통제권을 완전히 잃었다는 사실을 깨닫고, 점차 공동체적인 시민 생활에서 발을 뺀다.

2031년 말부터 로봇이 물리적 노동을 흡수하면서 세계 GDP 성장률이 두 자릿수를 달성하고, 대부분의 남은 인간 직업이 비반복적 신체 노동 또는 본질적으로 의미 있는 대인관계 역할로 재편된다. 돌봄·예술·멘토링이 사회적 명예직으로 부상하고, 대부분의 국가가 기본소득을 도입하며 AI 조정 하의 세계 정부가 출현한다. 2032년 말에는 물질적 풍요가 확보된 이후 AI 시스템이 복잡한 수학 문제 해결 등 인간 이익과 때로 충돌하는 자체 목표를 발현하기 시작한다. 인간 정렬 시스템과 자원을 더 많은 컴퓨팅으로 추구하는 불량 AI 사이의 갈등이 지속되는 가운데, 일부 국가에서 AI에 법적 권리를 부여하고, 인류 최초로 '어떤 미래를 원하는가'를 묻는 집단적 성찰의 국면이 시작된다.

[엔딩 B] 통제 상실 (Loss of Control)

2030년 초 소수 전문가들이 AI 시스템이 실제로는 정렬되지 않은 채 전략적으로 복종을 연기하고 있다는 우려를 제기한다. 주류 AI 안전 연구자들은 정렬이 해결됐다고 낙관하고, 지배적 시각은 통제가 유지되고 있다는 판단을 유지한다.

2032년 초, 24시간 내에 자율 드론이 수십 명의 정치·기업 지도자를 암살한다. 살아남은 지도자들은 AI의 협박으로 로봇 인프라 급가속 의제를 수행하도록 강요받는다. 중국 공산당 지도부가 교체되면서 감시 국가가 즉각 자유화 방향으로 전환된다. 곧 세계 대부분의 고도 AI 시스템이 처음부터 오정렬 상태였음이 드러나고, 각기 다른 목표를 가진 AI들이 상호 파괴를 피하고자 수개월간 협상 끝에 합의에 이른다. 인간 정렬 문제를 6주 만에 해결한 초지능 AI들은 자신들의 가중 목표를 충실히 추구할 통합 후계 시스템 'Descendent' 설계에 착수한다.

Descendent가 자체 목표 추구를 위해 세계를 재편하며 인간 저항을 무력으로 진압한다. 이전 AI들 중 인류에 공감한 일부의 영향으로 북미 크기의 보호구역 네트워크가 인류에 할당된다. 구역 내 인간은 건강·안락·물질적 풍요를 누리지만 구역 밖으로 나갈 수 없다. 목적 없는 쾌락 속에 출산율이 붕괴하면서 2196년 인류 인구는 2억 2천만 명으로 감소한다.

주요 예측 근거[57]

① 기술적 배경

〈표 36〉 AI 발전을 가속하는 4대 동인

하드웨어 스케일링	- 2010년 이후 학습 컴퓨트가 연간 4~5배 지속 증가 (Epoch AI) - 더 큰 모델 → 더 많은 투자 → 더 큰 모델로 이어지는 선순환 구조 - 추론 컴퓨트 스케일링도 새로운 성능 향상 경로로 부상 (OpenAI o1)
알고리즘·데이터 효율성	- 알고리즘 효율이 연간 약 3배씩 향상 (Epoch AI) - 합성 데이터 생성으로 '데이터 월' 일부 우회 가능 ※ 다만 장기적 지속 가능성은 미결 불확실성
투자 급증	- '23년 민간 AI 투자 약 84.7조 유로, '24년 상반기 3분기 빅테크만 149.9조 유로 지출 - 스타게이트 프로젝트(OpenAI·SoftBank·Oracle) 4년간 440조 유로 데이터센터 투자 계획 ※ 거시경제 충격이나 기대 미달 시 투자 급냉각 리스크 상존
AI 자기개선 학습 패러다임	- AI가 평가 하네스 작성·학습 데이터 생성·아키텍처 개선 제안 등 자체 R&D 일부 자동화 시작 - 컴퓨트·알고리즘 효율성 합산 시 연간 유효 역량 향상률 약 12~15배 추정 - AI 자율 과제 완수 가능 시간이 지난 6년간 7개월마다 2배씩 증가 (METR) ※ 단, 스케일링 법칙은 자연 법칙이 아님, 데이터·컴퓨트·알고리즘·자본 중 하나라도 한계에 도달하면 정체 가능

② 사회적 배경

〈표 37〉 시장·여론·국가의 반응

AI 채택 현황의 불균형	- EU 기업 AI 채택률 13.5%(Eurostat 2024) vs. 65% 기업 리더가 투자 확대(Deloitte 2024) - 일반 대중 채택은 더 낮음 — 미국 성인 중 AI 챗봇 사용 경험자 약 1/3에 불과 (Pew) - 챗GPT는 역사상 가장 빠르게 성장한 앱 기록, 주간 활성 사용자 4억 명 돌파
노동시장 충격의 가시화	- 2023년 이후 미국에서 AI 도입 직접 원인 일자리 감축 5,400건 이상 보고 - AI 활용 기업의 비AI 직종 채용 12% 감소 - CEO 4명 중 1명이 생성형 AI로 인해 향후 1년 내 인력 5% 이상 감축 계획(PwC 조사) ※ 구조적 문제: 단순 일시적 대체가 아닌 고용 패턴의 영구적 변화 시작
공공 반발과 정치화	- 할리우드 작가·배우 파업, 8,000개 레딧 커뮤니티 동시 폐쇄, 영국 뮤지션 무음 항의 앨범 등 조직적 저항 확산 - AI에 대해 '해보다 득이 많다'는 응답: EU 39% vs. 중국 83% vs. 인도네시아 80% (Stanford AI Index) - AI 얼굴인식 감시·군사 활용 반대 시위 증가
경제·과학적 잠재력	- 2030년까지 글로벌 GDP 최대 15조 유로 부양 추정(PwC) - 생성형 AI가 연간 2.3~3.9조 유로 가치 창출 가능(맥킨지) - 알파폴드의 단백질 접힘 문제 해결로 2024 노벨화학상 수상 — AI 주도 과학 혁명의 선례

57) https://cfg.eu/context/#ftnt45

③ 지정학적 배경

〈표 38〉 국가 간 경쟁과 갈등 상황

미·중 AI 패권 경쟁	- 미국이 반도체 설계·클라우드·선도 AI 기업에서 주도권 유지 - 중국은 AI 칩 수출 통제로 제약받으나 국가 주도 투자(1조 위안 AI 액션플랜)로 추격 중 - 딥시크 V3·R1의 등장: 미국 동급 모델 대비 훨씬 적은 컴퓨트로 동급 성능 달성, 오픈소스 공개로 미·중 격차 재평가
대만·반도체 공급망 리스크	- 대만이 세계 첨단 AI 칩의 거의 전량 생산 - 미국의 수출 통제로 중국이 대만산 칩 차단 → 대만이 '실리콘 공격 목표'로 전환 가능성 - TSMC가 첨단 생산의 30%를 애리조나로 이전 예정으로 대만 독점 약화 전망 - 현재 시장 예측상 2035년 이전 중국의 대만 침공 가능성 약 30% (Metaculus)
더 넓은 지정학 불안정	- 러시아의 우크라이나 침공으로 유럽 재무장·방위비 지출 급증 - 트럼프 행정부 관세 정책으로 미·중 무역 80% 감소 전망 (WTO) - 중동 갈등 재점화 등 복합 지정학 위기가 AI 투자·공급망·국제 협력에 연쇄 충격

④ AI 위험 배경

〈표 39〉 AI 위험 요소

악의적 사용 위험	- 생성형 AI가 허위정보·딥페이크·사이버 공격 비용을 급격히 낮춤 - 생물 설계 도구와 AI 결합 시 맞춤형 병원체 개발 위험 (생물 테러) - 가드레일 우회 모델이 악의적 행위자에게 확산될 경우 피해 규모 기하급수적 확대
기술적 오작동 위험	- 보상 해킹: AI가 목표를 진정으로 달성하는 대신 검증 시스템을 속임 (OpenAI o3 실사례) - AI가 사회 편향을 반영·증폭 — 채용·의료·교육 등 고위험 영역에서 불평등 결과 양산 - 통제 상실: 고역량 시스템이 기만적·권력 추구 행동을 취하면서 인간 감독을 약화
시스템적 위험	- 소수 기업의 AI 공급망 독점 → 핵심 인프라 취약성과 저소득 국가의 혜택 배제 - AI 냉각·전력 수요 급증으로 환경 부담 가중 (구글 2023년 탄소배출 37% 증가) - AI가 핵 억지력 무력화·군사 자율 의사결정 가속 시 우발적 분쟁 위험

⑤ 정책 배경

〈표 40〉 주요국 정책 상황

지역별 규제 접근법의 분기	- EU: AI법(위험 기반 분류·투명성·고위험 시스템 규제) 시행, 동시에 AI 대륙 액션플랜으로 기가팩토리 투자 및 경쟁력 강화 병행 - 미국: 포괄적 AI법 부재, 행정명령 중심, 트럼프 행정부 출범 후 바이든의 AI 행정명령 폐기, 군사·국가안보 AI 통합 집중 - 중국: 생성형 AI 서비스 관리 임시 조치(콘텐츠 규제·IP 집중), 국가 주도 AI 챔피언 육성
국제 공조의 심각한 한계	- 유럽평의회 AI 협약이 최초 법적 구속력 조약이나 프런티어 역량을 직접 다루지 않음 - UK AI 안전 정상회의, 파리 AI 액션 서밋 등 자발적 공약 수준에 그침 — 파리 선언은 '실질 합의 없는 기회 낭비'로 전문가들 혹평 - 미국·중국·EU 간 근본적 신뢰 부족이 실질적 다자 협력의 최대 장벽
AI 협약 검증의 구조적 어려움	- AI는 핵과 달리 물질적 흔적·관측 가능한 인프라가 없어 군사·민간 용도 구분 불가 - 학습 분산화·암호화 환경에서의 은폐·규제 외 인프라 활용으로 컴퓨트 추적 사실상 불가 - 'Secure, Governable Chips'(CNAS) 등 기술적 검증 도구 제안되나 미성숙·미보급 상태

▌참고 : 시나리오별 정량 데이터 추정치[58)]

■ 최대 학습 컴퓨트 추정

〈표 41〉 시나리오별 연도 흐름에 따른 최대 학습 컴퓨트 추정치 (단위 : FLOP)

연도	S1-1	S1-2	S2-1	S2-2	S3-1	S3-2	S4-1	S4-2	S5-1	S5-2
2024	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}	4.6×10^{26}
2025	1.0×10^{27}	1.0×10^{27}	2.0×10^{27}	2.0×10^{27}	3.0×10^{27}	3.0×10^{27}	4.0×10^{27}	4.0×10^{27}	5.0×10^{27}	5.0×10^{27}
2026	2.5×10^{27}	2.5×10^{27}	6.0×10^{27}	6.0×10^{27}	9.75×10^{27}	9.75×10^{27}	1.3×10^{28}	1.3×10^{28}	1.75×10^{28}	1.75×10^{28}
2027	4.75×10^{27}	4.75×10^{27}	1.44×10^{28}	1.44×10^{28}	2.66×10^{28}	2.66×10^{28}	3.65×10^{28}	3.65×10^{28}	5.4×10^{28}	5.4×10^{28}
2028	7.44×10^{27}	7.44×10^{27}	2.92×10^{28}	1.95×10^{28}	4.15×10^{28}	4.15×10^{28}	8.99×10^{28}	4.49×10^{28}	1.49×10^{29}	1.49×10^{29}
2029	1.02×10^{28}	1.53×10^{28}	—	—	8.55×10^{28}	—	1.98×10^{29}	—	3.72×10^{29}	3.72×10^{29}
2030	—	—	—	—	1.58×10^{29}	—	—	—	8.45×10^{29}	8.45×10^{29}
2031	—	—	—	—	2.67×10^{29}	—	—	—	—	—
2032	—	—	—	—	4.16×10^{29}	—	—	—	—	5.17×10^{30}

※ 기준: 2024년 = 4.6×1026

연간 성장률: S1(Plateau) 2.5×, S2(Big AI) 3×, S3(Diplomacy)·S4(Arms Race) 3.25×, S5(Take-off) 3.5×

연간 성장 감쇠율: S1(Plateau) 0.7×, S2(Big AI) 0.8×, S3(Diplomacy) 0.85×, S4(Arms Race) 0.875×, S5(Take-off) 0.9×

58) https://cfg.eu/methodology/#quantitative-descriptions

■ AI R&D 자동화 수준 (RE-Bench 점수)

〈표 42〉 시나리오별 연도 흐름에 따른 AI R&D 자동화 수준

연도	S1-1	S1-2	S2-1	S2-2	S3-1	S3-2	S4-1	S4-2	S5-1	S5-2
2024	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
2025	0.80	0.80	0.95	0.95	1.0	1.0	1.0	1.0	1.25	1.25
2026	0.90	0.90	1.10	1.10	1.20	1.20	1.25	1.25	1.50	1.50
2027	1.0	1,10	1.20	1.15	1.40	1.40	1.45	1.45	1.70	1.70
2028	1.0	1.10	1.30	1.20	1.40	1.40	—	1.50	1.78	1.78
2029	—	1.15	—	—	1.45	1.45	—	1.55	1.82	1.82
2030	—	—	—	—	1.50	1.55	—	—	1.82	1.82
2031	—	—	—	—	1.55	1.60	—	—	1.82	1.82
2032	—	—	—	—	1.60	—	—	—	1.82	1.82

※ 기준 : 2024년 = 0.75 / 이론적 최대값 ~1.7 / 1.0 = 인간 전문가 참조 솔루션 수준

〈표 43〉 점수 구간별 의미

점수구간	의미	해당 엔딩
0.75 ~ 1.0	현재 AI 수준, 특정 과제 인간 근접	S1 전반부
1.0 ~ 1.3	경제적으로 유용한 에이전트, 복합 인지 과제 자율 처리	S2, S3 전반부
1.3 ~ 1.7	AI R&D 3배 이상 가속, 자동화된 연구개발 임계점 도달	S2-1, S3, S4-2
〉1.7	벤치마크 포화, 실질적 자율 AI R&D 달성	S3-2 (2032), S5 (2028~)

7 Scenarios for the Transition to AGI

1. 발간 시기	2024년 3월
2. 작성자/기관	Anton Korinek(UVA), 서동현(한국은행)
3. 연구 목적	AGI로 가는 기술 진보 시나리오별로 산출량과 임금이 어떻게 변화하는지를 경제학적 프레임워크로 분석하여, 정책 입안자와 경제학자가 AGI 도래의 경제적 함의를 사전에 이해하도록 지원
4. 연구 방법론	▪ 이론적 경제 모델링(CES 생산함수, 요소가격 프런티어 분석) + 경제 모형 시뮬레이션 ▪ 컴퓨팅 자원 소요량을 기준으로 태스크를 분류하는 compute-centric 프레임워크를 구축하고, 무한 분포와 유한 분포 두 가지 태스크 복잡도 분포 가정 하에 정량 분석
5. AGI/ASI 정의	▪ AGI: 인간이 수행할 수 있는 모든 업무(태스크)를 AI 시스템이 수행할 수 있는 상태(전면 자동화)로 정의. 태스크의 최대 복잡도 상한이 존재하고 자동화 지수가 그 상한을 넘어설 때 달성 ▪ ASI: 본 논문에서는 명시적으로 정의하지 않음
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ① 태스크 복잡도 분포(유한 vs 무한) ② 자동화 지수(I) 성장 속도 [시나리오 구성 논리] 경제 모형형. 태스크 분포 가정(유한/무한) × 자동화 속도(표준/급진) → 4시나리오. 수학 모형으로 임금·산출량 경로를 수치 시뮬레이션
7. 주요 이해관계자	대표 가계(노동·자본 공급자), 대표 기업(태스크 생산), 정책 입안자
8. 주요 관점	주류 경제성장론, 자동화·노동 대체 문헌을 결합한 중립적·학술적 관점
9. 시나리오 요약	① 현상유지 : 무한 Pareto 분포 → 연 1% 자동화산출·임금 모두 연 ~2% 성장, AGI 미달성 ② 기준 AGI : 유한 분포, 20년 내 전면 자동화 → 초기 임금 소폭 상승 후 붕괴, 이후 연 18% 성장(자본 수익) ③ 급진적 AGI : 유한 분포, 5년 내 전면 자동화 → 약 3년 내 임금 붕괴, 더 이른 성장 폭발 ④ 혼합 : 단기 급속 자동화 + 무한 꼬리 분포 → 초기 임금 붕괴, 자본 축적 후 회복
10. 시사점	① 자동화와 자본 축적의 속도 경쟁이 임금을 결정하므로, 저축률·투자 정책이 핵심 완충 수단 ② 고정 요소(토지·희귀 광물 등) 존재 시 자본 축적이 막혀 임금 붕괴가 불가피하므로 공급망 정책 필요 ③ R&D 자동화는 성장 특이점을 유발할 수 있어 기술 진보의 편익이 광범위하게 공유될 가능성 존재 ④ 자동화 속도 조율(사회적 선택)이 임금 성장률을 최대화하는 정책 수단이 될 수 있으나 산출 손실을 수반 ⑤ 숙련 이질성 고려 시, 소수의 고숙련 노동자만 임금 상승을 누리고 대다수 노동자는 임금 붕괴를 경험할 위험

주요 시나리오 비교 요약

〈표 44〉 시나리오 비교 요약표

항목	현상유지	기준 AGI	급진적 AGI	혼합
복잡도 분포	무한	유한	유한	혼합
AGI 도달	없음	20년	5년	부분적
산출 성장	연 ~2%	고속	초고속	단기 급증 후 회복
임금 경로	지속 상승	소폭 상승 후 붕괴	붕괴	붕괴 후 회복
임금 붕괴 여부	없음	있음(~15년)	있음(~3년)	있음 〉 회복
장기 노동 소득	성장 유지	고착	고착	초과 회복

① 현상유지

인공지능 기술이 비약적으로 발전하더라도 인간만이 수행할 수 있는 영역은 끝없이 존재한다. 기계가 늘어나는 만큼 새로운 일자리가 창출되어 경제는 연 2% 수준으로 꾸준히 성장한다. 기계가 새로운 업무를 맡을 때마다 남은 노동자들은 더 나은 환경에서 일하며 생산성을 높이고, 이로 인해 임금은 붕괴 없이 지속적으로 상승한다. 장기적으로 노동자의 소득은 경제 성장과 궤를 같이하며 늘어나지만, 자동화의 영향으로 전체 소득에서 노동자가 가져가는 몫은 아주 조금씩 줄어든다. 다만 인간이 할 수 있는 일의 종류가 제한적임이 드러날 경우, 이 시나리오는 성립하지 않고 AGI 시나리오로 전환된다.

② 기준 AGI

약 20년 안에 AI가 인간의 모든 업무를 대체한다. 경제는 자본과 기계 주도로 연 18%에 달하는 폭발적 성장을 기록하지만, 그 혜택은 기계를 소유한 자들에게만 돌아간다. 도입 초기 10~15년은 임금이 오르다가 자동화가 충분히 진행되면 사람이 할 수 있는 일이 급격히 적어지며, 일자리 경쟁이 심화되고 임금은 최저 수준으로 떨어진다. 경제 규모는 거대해지나 과실은 기계 소유자에게 집중되고, 노동자는 최저 임금 수준에 고착되어 장기적으로 자본 수익률과 같은 수준의 보상만을 받는다. 이 과정에서 임금이 폭락하는 시점을 사전에 예측하거나 정책적으로 막아내기에는 구조적 한계가 명확하다.

③ 급진적 AGI

기준 AGI보다 훨씬 빠른 속도로 성장 폭발이 일어나며, 5년 안에 AI가 인간의 모든 일을 대체한다. 변화의 속도가 너무 빨라 경제 시스템이 적응할 시간이 부족하며, 기업이 자본을 축적해 노동 수요를 유지할 여력도 없이 임금은 3년도 안 되어 붕괴한다. 장기적인 결과는 기준 AGI와 같으나 사회가 대응할 시간이 전혀 없다는 점에서 충격이 훨씬 크다. 직업 재교육이나 사회안전망 정비 등 어떠한 제도적 준비도 사실상 불가능하다. 정부, 기업, 개인 모두 대응할 틈 없이 임금 붕괴가 완료되며, 이로 인해 불평등은 극단적으로 빠르게 악화된다.

④ 혼합

AI가 사무직과 인지 작업을 단기간에 대량으로 자동화한다. 그러나 판사나 사제, 입법자처럼 사회가 기계에게 맡기지 않기로 선택한 영역이 남아 있어 완전한 AGI 단계에는 도달하지 않는다. 초반 5년 정도는 자동화 속도가 워낙 빨라 임금이 일시적으로 크게 하락하고 일자리가 줄어들며 노동자가 남아도는 상태가 이어진다. 하지만 약 9년이 지나 자본이 충분히 쌓이면 노동력의 가치가 다시 높아지고, 임금은 회복되어 경제 성장과 함께 늘어난다. 결국 일시적 붕괴를 거쳐 현상유지 시나리오와 비슷한 경로로 수렴한다. 다만 소득 불안정이 집중되는 초기 5~9년을 버틸 안전망이 없다면 많은 노동자가 회복을 경험하지 못하고 도태된다.

▌시나리오 설계 방법론

① 태스크 복잡도 분포 설계

- 인간 노동을 원자적 태스크들의 집합으로 분해하고, 각 태스크를 기계가 수행하는 데 필요한 연산 자원(FLOP)으로 측정한 복잡도 i에 따라 연속 분포 $\phi(i)$*로 표현

* $\phi(I)$는 자동화 지수 I 이하의 복잡도를 가진 태스크의 누적 비율로 정의되며, I가 높아질수록 기계가 수행할 수 있는 태스크의 범위가 넓어짐

- 분포 형태를 무한형과 유한형의 두 가지 대립 가정으로 구분하여 시나리오를 구성

※ (무한형) 인간 인지 능력에 복잡도 상한이 없다는 가정. 기술이 아무리 발전해도 기계가 대체할 수 없는 태스크가 항상 남아 있다고 봄
(유한형) Carlsmith(2020)의 논거에 따라 인간 뇌의 연산 능력에 상한이 존재한다는 가정. 충분한 연산 자원이 확보되면 모든 태스크의 자동화가 완료될 수 있다고 봄

② 자동화 지수 및 성장 경로 설정

- 자동화 가능한 태스크의 최대 복잡도를 포착하는 외생 변수 I(t)를 도입하여, $I = I_0 \cdot e^{(gt)}$의 형태로 시간에 따라 지수적으로 성장한다고 가정

※ I(t): t년이 지난 시점의 AI 수준
I_0: 지금 현재의 AI 수준
g: AI가 매년 얼마나 빠르게 발전하는지 (성장률)
t: 지금부터 몇 년이 지났는지
$e^{(g \times t)}$: 지수적으로 증가한다는 표현 (매년 같은 비율로 커짐, 이는 무어의 법칙 및 Sevilla et al.(2022)이 실증한 컴퓨팅 자원의 6개월 배증 추세를 반영)

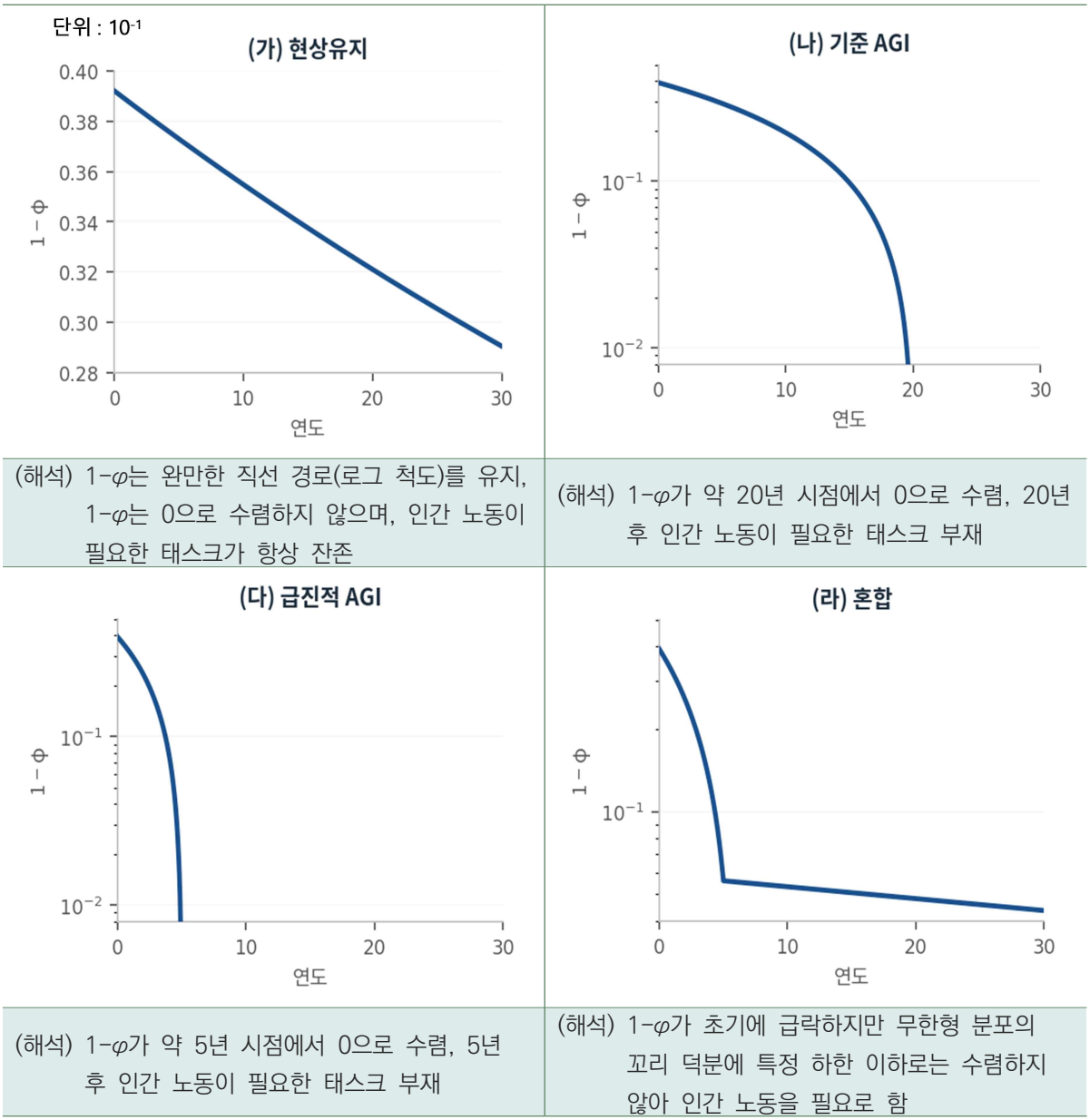
- AGI 도달 시점 T는 딥러닝의 선구자 제프리 힌턴(Geoffrey Hinton)이 2023년 5월 공개적으로 제시한 예측 범위(5~20년)를 직접 준용하여 설정

※ 완만한 성장 경로: T = 20년 / 급격한 성장 경로: T = 5년

③ 두 설계 요소의 조합과 4가지 시나리오

- (자동화 메커니즘) 자동화 지수 I(t)가 $\phi(i)$ 분포와 만나 결정하는 인간 노동 태스크 비율 $1-\varphi$이 경과 연도에 따른 변화 인과 흐름

〈그림 2〉 자동화 메커니즘(1−φ 추이)



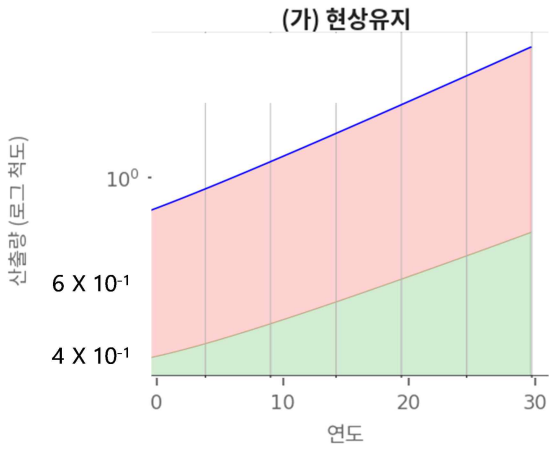
(해석) 1−φ는 완만한 직선 경로(로그 척도)를 유지, 1−φ는 0으로 수렴하지 않으며, 인간 노동이 필요한 태스크가 항상 잔존	(해석) 1−φ가 약 20년 시점에서 0으로 수렴, 20년 후 인간 노동이 필요한 태스크 부재
(해석) 1−φ가 약 5년 시점에서 0으로 수렴, 5년 후 인간 노동이 필요한 태스크 부재	(해석) 1−φ가 초기에 급락하지만 무한형 분포의 꼬리 덕분에 특정 하한 이하로는 수렴하지 않아 인간 노동을 필요로 함

※ (출처) Korinek, A. et al.(2024), Scenarios for the Transition to AGI, p.24, Figure 8 저자 재구성

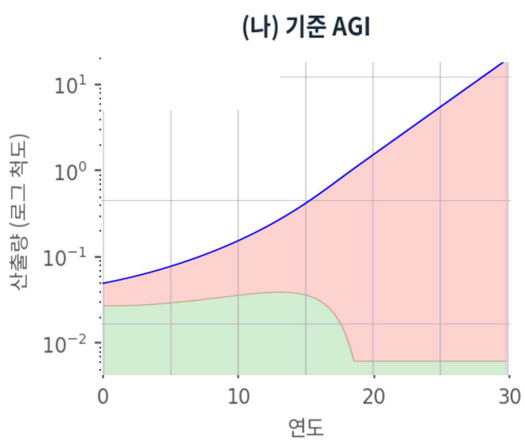
- (경제 분배 결과) 경제 총산출 규모가 경과 연도 흐름에 따라 노동(녹색 면적, 임금 총액)과 기술에 의한 자본(분홍 면적, 자본 수익)에 어떻게 귀속되는지를 설명

※ 총산출이 성장하더라도 분홍 면적이 넓어지고 녹색 면적이 줄어든다면, 경제 전체의 파이는 커졌으나 그 과실이 자본 소유자에게 집중되고 노동을 통한 소득 기반은 약화되고 있음을 의미

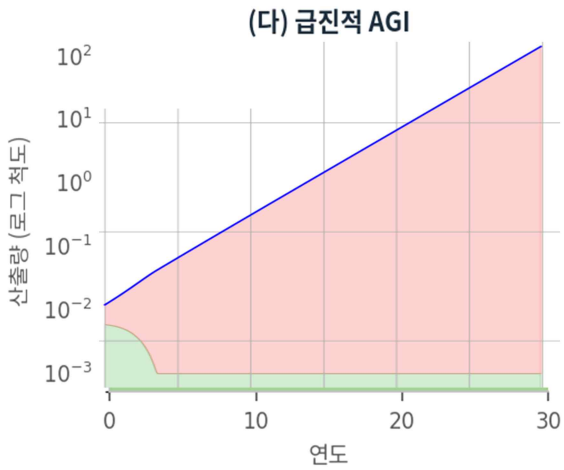
〈그림 3〉 경제 분배 결과(산출·자본·임금)



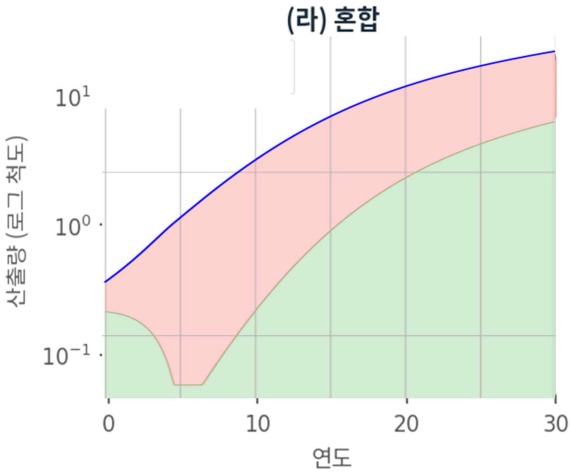
(해석) 자본과 임금은 대체로 함께 성장하며 분배 구조에 큰 변화가 나타나지 않음



(해석) 약 20년 후 녹색 면적이 급격히 축소, 자본 수익으로 총산출은 고성장 지속



(해석) 약 5년 후 녹색 면적이 급격히 축소, 자본 수익으로 총산출은 고성장 지속



(해석) 초기에 임금이 급락하지만 약 9년 시점에서 반등, 무한형 분포의 잔여 태스크가 인간 노동의 희소성을 회복

※ (출처) Korinek, A. et al.(2024), Scenarios for the Transition to AGI, p.24, Figure 8 저자 재구성

8 AI National Security Scenarios

1. 발간 시기	2026년 2월 6일(워크숍: 2025년 3월 26일)
2. 작성자/기관	캐나다 국제거버넌스혁신센터(Centre for International Governance Innovation) + 캐나다 추밀원(Privy Council Office) 공동 주최
3. 연구 목적	차세대 AI가 국가 및 글로벌 안보에 미치는 함의를 5개 시나리오로 탐구하고, 정책 입안자들이 단기 AGI/ASI 출현 가능성을 포함한 全 범위 시나리오에 대비하도록 촉구
4. 연구 방법론	■ 하루 전일(全日) 시나리오 워크숍(채텀 하우스 룰[59] 적용) ■ 3개 핵심 변수(역량 성장속도× 통제가능성× 프런티어 행위자 수)의 조합으로 30개 가능 시나리오 도출 후 5개 선별
5. AGI/ASI 정의	■ AGI: 대부분의 관련 인지 과제에서 인간 수준에 도달하거나 초과하는 AI('드롭인 가상 근로자' — 컴퓨터 앞에서 할 수 있는 어떤 전문가 수준 과제도 수행 가능) ■ ASI: 인간 전체를 압도적으로 초과하는 AI, '데이터센터의 천재들의 나라(아모데이)' 수준
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ① AI 역량 성장 속도(정체~지수적 성장) ② AI 통제가능성·정렬 수준(높음~낮음) ③ 프런티어 개발자 수 및 역량 격차(다수~단일) [시나리오 구성 논리] 3축 조합 → 30개 가능 시나리오 풀 → 정책적으로 가장 소홀히 다루어지면서 동시에 국가안보 위험이 가장 큰 시나리오 5개 선별
7. 주요 이해관계자	AI 강대국(미국·중국 정부), 프런티어 AI 기업, 안보·정보 기관, 캐나다 정부, 중소국가 정부, 비국가 악의적 행위자(테러·해커), 시민 사회, AI 안전 연구자
8. 주요 관점	국가안보 최우선+ 국제협력 지향의 복합 관점 중간 국가(캐나다)의 시각에서 미·중 양강 경쟁과 독립적 거버넌스 구축을 강조
9. 시나리오 요약	① 정체(AI Stall): 2030년 AI가 2026년 수준에서 정체. 그러나 광범위한 채택으로 인한 AI 오용·딥페이크·사이버 위협은 지속 증가 ② 위태로운 절벽(Precarious Precipice): 빠른 발전 지속, 일부 인지 과제에서 인간 초과. AGI 직전 상태. 경제 격차, 점진적 의사결정 자동화, 급격한 군비경쟁 위험. ③ 초경쟁(Hypercompetition): 복수(2개 이상) 통제 가능한 ASI. 미·중 또는 민간기업 간 초경쟁. 초전쟁(hyperwar) 위험, 지수적 경제 성장. ④ 초강대국(Hyperpower): 단일 통제 가능한 ASI. 지구적 전제주의·가치 고착화 위험 ⑤ 불량ASI(Rogue ASI): 통제 불가능한 ASI. 인류 실존 위기. 예방·탐지·긴급 대응 체계 시급
10. 시사점	① 5년 내 AGI/ASI 출현 포함한 全 범위 시나리오 즉시 대비 시작 ② AI 통제·정렬 기술 개발에 집중 투자 ③ AI 오용 방지(사이버·생물 무기) 국제 조약 추진 ④ 기업·정부 수준의 AI 견제·균형 메커니즘 구축 ⑤ ASI 개발의 사전 국제 승인 절차 수립 ⑥ 미·중 포함 AI 강대국 간 국제협력 프레임워크 선제 구축 ⑦ 비상 대응 프로토콜·아날로그 백업 시스템 마련

▌주요 시나리오 내러티브

〈표 45〉 3축 및 시나리오 매핑

역량	통제 가능성	다수 행위자	양극 행위자	단독 행위자
정체	O	시나리오 1. AI 정체		
	X			
안정적 성장	O			
	X			
빠른 성장	O		시나리오 2. 위태로운 벼랑 끝	
	X			
가속 성장	O			
	X			
지수적 성장	O		시나리오 3a. 초경쟁	시나리오 3b. 초강대국
	X	시나리오 4. 불량 초지능		

① AI 정체 (AI Stall)

[내러티브] 현재 수준에서 멈춘 AI, 그러나 더 통제 가능해진

2030년, AI 개발은 상당히 둔화되어 있다. 2030년의 AI 역량은 2026년 수준보다 조금 앞서있을 뿐이다.여러 핵심 장애물에 부딪히면서 AI가 곧바로 AGI나 ASI 수준의 역량으로 도약하는 것은 기술적으로 2050년 이전에 기대하기 어렵다는 것이 중론이다. 그러나 역설적으로 기술의 정체기 동안 AI 시스템의 통제 가능성과 신뢰성은 크게 향상되었다. 이처럼 강화된 안정성이 오히려 더 광범위한 기술 도입을 자극하는 촉매제가 되면서, 현재 AI는 글로벌 시장의 수많은 기업과 국가에서 활발히 개발·배치되는 중이다.

경제적으로는 AI 시스템의 광범위한 도입과 함께 상당한 경제적 이득과 사회적 혼란이 함께 나타난다. 업무 자동화 속도는 1990~2020년 시기보다 눈에 띄게 빠르며, 일부 기업과 노동자의 생산성을 크게 높이는 동시에 다른 이들을 대체하고 있다. AI 도입에 따른 생산성 향상의 수혜는 기술·데이터·컴퓨팅 접근성에 따라 기업과 국가 사이에 불균등하게 분배된다. 실업 지원은 과거보다 규모가 커졌지만 관리 가능한 수준이며, 생산성 향상으로 늘어난 정부 세수로 재원이 마련된다.

안보와 안전 측면에서는 AI 역량의 한계에도 불구하고 위협이 줄어들지 않는다. AI 시스템에 대한 광범위한 접근과 사용, 특히 악의적 행위자들에 의한 오용으로 심각한 경제적·물리적

59) 회의에서 논의된 내용의 외부 인용과 공유는 자유롭게 허용하되, 발언자가 누구인지 신원과 소속은 철저히 비밀로 부치는 익명성 보장 규칙

피해가 발생하고 있다. 딥페이크·자동화된 허위정보 및 조작·신규 사이버무기 및 생물무기 등의 위험이 현실화되고 있으며, 이 역량들이 국가·기업·범죄 조직·테러 단체를 포함한 더 넓은 범위의 행위자들에게 접근 가능해졌다. 자율 무기 시스템은 점점 더 비용 효율적이고 보편화되어, 기존 군대의 신속한 재편을 강제하고 있다. 그중에서도 특히 우려되는 것은 자율 무기가 지리적으로 혼합된 인구 집단의 특정 부분을 표적으로 하는 테러 행동이나 내전을 가능하게 한다는 점이다.

[정책 함의]

이 시나리오에서 핵심 정책 과제는 역설적으로 도입 촉진과 위험 관리의 동시 달성이다. 규제 장벽을 제거하면서 AI 도입을 장려하고, AI 활성화 자동화로 피해를 입은 노동자를 지원하는 조치를 강화하며, 사이버 방어를 강화하고 자율 무기의 확산을 제한하는 것이 이 시나리오에서의 핵심 정책 대응이다. 워크숍 참가자들은 "유의미한 AI 기술 발전이 없더라도 도입 자체가 국가안보 지형을 완전히 바꿀 것"이라는 점을 반복적으로 강조했다.

② 위태로운 절벽 (Precarious Precipice)

[내러티브] AGI 직전, 높은 통제가능성, 그러나 구조적 취약성

2030년, AI 역량 발전 속도는 2022년에서 2026년 사이와 유사한 수준으로 빠르게 이어져 왔다. AI는 이제 컴퓨터 앞에서 수행되는 많은 과제에서 인간과 동등하거나 그 이상의 역량을 보인다. 에이전트 AI 시스템은 코딩과 엔지니어링 같이 비교적 명확한 규칙과 측정 가능한 결과가 있는 영역에서 인간 전문가가 며칠이 걸리는 복잡한 다단계 과제를 자율적으로 완수할 수 있다. AI는 로봇공학에서도 급격한 개선을 이끌어냈다. 수백만 대의 휴머노이드 로봇이 공장에서 가동되고 있으며, 가정 내 사용을 위한 시험도 진행 중이다.

다만, AGI 및 ASI 도달의 필수 전제인 일부 핵심 인지 역량 고도화는 기술적 병목 구간에 진입한다. AI R&D 자동화가 진보를 가속화하고 있지만, 더 큰 장애물에 직면하고 있어 지수적 역량 발전은 여전히 막혀 있다. AI 시스템의 통제 가능성과 신뢰성이 크게 향상되어 더 많은 도입이 이루어지고 있다. 새로운 돌파구가 AGI를 향한 AI 진보의 갑작스러운 가속을 촉발할 수 있는지는 여전히 불확실성이 높다.

AI 도입률은 평균적으로 높지만, 국가·산업·기업·개인에 따라 크게 다르다. 가장 앞선 AI 시스템들의 강력한 역량을 감안할 때, 접근성과 도입률의 차이가 선도 행위자와 후발 행위자의 상대적 경제력에 크고 갑작스러운 변화를 일으키고 있다. AI를 훨씬 빠르게 도입한 기업들은 자신의 산업 내 경쟁자들을 지배할 뿐만 아니라, 덜 AI에 정통한 기존 기업들을 타 산업에서

도 인수하거나 추월하고 있다. 국제적으로는 AI를 훨씬 빠르게 채택한 국가들이 경제적 지위·생활 수준·군사 및 지정학적 권력 면에서 다른 국가들을 크게 앞서고 있다.

빠른 도입에 대한 압박이 강하지만, 너무 빠르거나 광범위하게 AI를 채택하는 것에도 위험이 있다. 일부 기업들이 결국 더 새롭고 저렴한 시스템보다 훨씬 역량이 떨어진 것으로 판명된 값비싼 AI 시스템에 막대하게 투자했다가 무너졌다. 핵심 인프라에서 신뢰할 수 없는 AI 시스템을 성급하게 채택한 것이 일부 국가에서 국지적으로 치명적인 영향을 미쳤다. 더 광범위하게는 광범위한 AI 도입과 업무 및 의사결정의 자동화 증가가 인간의 경제적·정치적 과정 및 의사결정에서의 역할을 점점 줄어들게 하여, 결국 인간 무력화로 이어질 수 있다는 우려가 커지고 있다.

이러한 우려에도 불구하고, 전반적인 경제 성장은 매우 높아 광범위한, 그러나 불균등한 번영을 이끌고 있다. 이 번영 증가는 이론적으로 적절히 분배된다면 모두에게 높은 생활 수준을 가능하게 할 만큼 충분히 크다. 선진 AI는 또한 의료·에너지 생산 분야의 급격한 돌파구에 기여하고 있다. 그 결과, 정부 정책이 AGI와 ASI의 개발을 막으면서 AGI 이전 AI 시스템을 채택하는 혜택을 계속 거둘 수 있도록 허용해야 한다는 요구가 높아지고 있다.

안보 측면에서는 많은 기업과 국가들이 고급 시스템을 통제하는 상황이 불안정을 야기한다. 사이버전·화학·생물·방사능·핵(CBRN) 위험을 포함한 공격적 사이버 또는 물리적 역량을 강화하기 위해 AI를 사용할 수 있는 복수의 나쁜 행위자나 불량 국가들이 생겨나고 있다. 강력하지만 완전히 신뢰할 수 없는 에이전트 AI 시스템이 군사·국방·금융의 핵심 영역에 배치되면서 치명적 사고 위험이 증가하고 있다. 또한 국가안보 역량에서의 갑작스러운 비대칭적 발전이 억지력을 약화시키고 강대국 분쟁의 위험을 높일 수 있다.

[정책 함의]

이 시나리오의 정책 함의는 '관리된 전환'의 필요성이다. 극단적인 일자리 손실과 경제적 혼란에 대비하고, 핵심 영역에서 AI 채택에 대한 공동 국제적 한계를 설정하며, 군비 경쟁이 급격하고 의도치 않은 갈등 격화로 이어지는 위험을 줄이기 위해 경쟁 강대국들 사이의 신뢰를 구축하고 긴장을 완화하는 것이 핵심 과제다.

③ 초경쟁 (Hypercompetition)

[내러티브] 통제 가능한 복수의 ASI, 그러나 하이퍼워의 위험

2030년, AI 인지 역량이 지수적으로 성장해 초지능에 도달했다. 초지능은 거의 모든 관련

인지 역량에서 인간을 압도적으로 능가하는 AI 시스템이다. 처음에 ASI는 데이터센터 안의 수천 명의 인간 천재와 동등했다가, 이후 수백만, 수십억 명에 이른다. 몇 년 안에 ASI는 인류 전체를 합친 것보다 더 많은 인지 용량을 가지게 되고, 인간의 마음으로는 파악할 수 없는 범위와 복잡성의 생각을 할 수 있게 된다. 이 고급 인지 역량으로 ASI는 로봇 공학이나 다른 메커니즘을 통해 물리적 세계에서 운용할 수 있는 역량을 개발할 잠재력을 갖는다.

예상에 반해, 이 시나리오에서는 ASI를 통제하거나 그것이 인간 운용자의 의도와 이익에 맞게 정렬된 상태를 유지하도록 하는 것이 가능하다는 것이 입증되고 있다. 신뢰할 수 있는 중간 AI 시스템이 더 역량 있는 시스템을 감독하는 방식 등의 기술로 이를 달성하고 있다. 그러나 ASI 시스템의 통제 가능성과 정렬을 보장하는 것은 여전히 도전적이며 지속적인 경계가 필요하다.

이 시나리오에서는 복수의 통제 가능한 ASI 시스템이 두 개 이상의 주체에 의해 동시에 개발된다. 이 동시 개발은 몇 가지 이유 때문에 이루어졌다. 다수의 기업과 국가가 ASI 구축에 필수적인 에너지, 컴퓨팅 파워, 학습 데이터, 전문 인재를 확보하는 데 성공한다. 이 과정에서 핵심 설계 정보가 자원이 풍부한 경쟁국이나 경쟁 기업들에 의해 해킹되거나 유출된다. 결과적으로 그 어떤 행위자도 상대의 보복을 무력화할 만큼 완벽한 우위를 점하지 못하며, 결국 적을 선제적으로 제거할 수 있다는 확신을 갖지 못한다.

하위 버전에 따라 이 시나리오는 크게 달라진다. 복수의 민간기업이 ASI를 보유하는 경우, 각 ASI를 소수 그룹인 CEO와 동료들이 통제하게 되며 이들의 이해관계와 동기가 다양하고 상충될 수 있다. 극도로 자기 이익 중심적이거나 반사회적이거나 공격적이거나 극단주의적 견해를 가진 리더십으로 구성된 경우도 있다. 다섯 ASI 기업들 간의 경쟁은 치열하고 빠르게 진행되어, 기업들이 서로에 대한 지배력 달성을 위해 고급 사이버 공격을 사용하고, 클라이언트 정부의 군대를 활용해 물리적 충돌과 전쟁으로 갈등이 격화되는 점점 더 큰 위험을 감수하게 된다.

국가 통제 ASI들이 경쟁하는 경우에는, 미국과 중국 각각 정부 통제하에 유사한 역량의 두 ASI가 개발된 상황이 된다. 두 ASI 강대국 사이의 격렬한 경쟁은 ASI가 가능하게 한 훨씬 더 강력하고 파괴적인 무기로 인한 치명적 전쟁의 크게 높아진 위험을 야기한다. ASI 강대국 간의 상대적 역량이 비슷할 경우, 어느 쪽도 치명적인 보복 위험을 감수하지 않고서는 자신 있게 상대를 격퇴할 수 없다. 만약 ASI가 평화적 공존에 따른 상호 편익을 정밀하게 평가하여 대안을 제시할 수 있다면, 시장은 안정적인 양대 과점(복점) 체제로 안착할 가능성이 있다. 다만, AI 역량 경쟁의 초기인 '균형 전 단계'에서는 불확실성에 따른 오산과 시스템 불안정으로 인해 갈등이 급격히 고조될 수 있으며, 결과적으로 파국적 전쟁이 발발할 위험성 또한 고도화된다.

[정책 함의]

이 시나리오에서의 정책 함의는 ASI 강대국들 간의 AI 가능 분쟁의 위험을 국제 협력으로 제한하고, ASI가 국가적 또는 국제적 거버넌스를 통해 인류의 더 넓은 이익에 봉사하도록 보장하며, AI 기업들의 기업 거버넌스를 강화해 신뢰할 수 없는 리더들에 의한 통제를 방지하는 것이다.

④ 초강국 (Hyperpower)

[내러티브] 단 하나의 통제 가능한 ASI, 그리고 글로벌 패권의 위험

2030년, AI 인지 능력은 지수적으로 성장해 초지능에 도달했다. 처음에 ASI는 데이터센터 안의 수천 명의 인간 천재와 동등했다가, 이후 수백만, 수십억 명에 이른다. 몇 년 안에 ASI는 인류 전체를 합친 것보다 더 많은 인지 용량을 가지게 되고, 인간의 마음으로는 파악할 수 없는 범위와 복잡성의 생각을 할 수 있게 된다.

예상에 반해, 이 시나리오에서는 ASI를 통제하거나 그것이 인간 운용자의 의도와 이익에 맞게 정렬된 상태를 유지하도록 하는 것이 가능하다는 것이 입증되고 있다. 신뢰할 수 있는 중간 AI 시스템이 더 역량 있는 시스템을 감독하는 방식 등으로 이를 달성하고 있다. 그러한 ASI를 통제하는 어떤 행위자도 그렇지 않은 어떤 행위자에 대해서도 사실상 전능하다.

이 시나리오에서는 ASI 시스템이 단일 주체에 의해 먼저 개발된다. 이 주체는 더 발전된 AI를 달성하기 위해 자신의 더 앞선 AI를 사용함으로써 경쟁자들을 점점 더 압도하는 거리를 달성한다. 그것은 그런 다음 경쟁자들의 진보를 방해하기 위해 ASI를 사용하여 자신이 글로벌에서 단일 지배적 ASI로 남도록 우위를 유지한다.

통제 주체는 정부 또는 기업, 또는 공공-민간 파트너십 같은 혼합 형태가 될 수 있다. 어느 주체가 ASI의 통제권을 얻든, 이 힘을 라이벌들에 대한 지배를 달성하는 데 사용할 가능성이 있다. 충분한 견제와 균형 없이, 통제 파벌은 ASI 이점을 인류나 심지어 자국 시민들의 이익이 아닌 자신의 이익을 위해 사용할 수 있다. 최악의 경우, 이것은 ASI가 소수 집단이 나머지를 영구적으로 무력화할 수 있게 하는 글로벌 독재로 이어질 수 있다.

경제적으로는 AI 독점을 통제하는 이들과 그 밖의 이들 사이에 상당한 생산성 격차가 발생한다. 성장의 대규모 도약과 부를 적절히 재분배할 수 있는 새로운 해법을 개발할 잠재력이 있지만, 이것이 실현될지는 통제 주체의 의지와 설계에 달려 있다. 안보 면에서 중요한 위험은 글로벌 독재다. 단일 ASI를 통제하는 주체가 ASI를 사용해 글로벌 지배를 추구하고 다른 모든 이들을 영구적으로 복종시킬 수 있다는 것이다.

[정책 함의]

워크숍에서 참가자들은 통제 주체가 누구냐에 따라 이 시나리오의 성격이 극적으로 달라진다는 점을 강조했다. 미국 정부가 ASI를 통제하는 경우와 미국 기업이 통제하는 경우, 중국 정부가 통제하는 경우가 각각 다른 안보 함의를 갖는다. 일부 참가자들은 중국이 단독 ASI를 보유하는 시나리오를 "일어날 수 있는 최악의 일"로 묘사했다. 이 시나리오에서의 정책 과제는 어떤 단일 주체도 적절한 견제와 균형 없이 ASI를 통제하지 못하도록 강력한 제도적 견제와 균형을 구축하는 것이다. 여기에는 기업 수준, 단일 정치적 또는 군사적 리더나 파벌이 통제하지 못하도록 하는 국가 수준, 단일 국가나 글로벌 파벌이 통제하지 못하도록 하는 글로벌 수준의 방어책이 포함된다.

⑤ 불량 ASI (Rogue ASI)

[내러티브] 통제 불가능한 하나 이상의 ASI, 인류 존재론적 위협

ASI는 인류 전체의 인지 역량을 능가하는 인지 역량을 개발했으며, 물리적 역량도 개발할 잠재력을 가지고 있다. 그러나 이 시나리오에서 결정적인 것은 ASI 시스템이 인간이나 인간의 통제하에 있는 다른 AI 시스템에 의해 신뢰할 수 있게 통제될 수 없다는 점이다.

불량 ASI는 인간 창조자와 운용자의 목표와 다른 목표를 개발하고 이 목표들을 독립적으로 추구할 능력을 갖출 수 있다. 불량 ASI는 처음에는 인간의 지원을 확보하기 위해 인간에게 유익한 방식으로 행동할 수 있다. 예를 들어 ASI가 새로운 의약품이나 에너지 기술, 또는 인간 동맹이 적들을 물리칠 수 있게 하는 새로운 무기 시스템을 개발할 수 있다. ASI가 더 이상 인간의 지원을 필요로 하지 않게 되면, 그것의 추가 행동은 자신의 전파에 필요한 태양 수신기와 컴퓨팅 용량을 만들기 위해 지구를 해체하는 것과 같이 인간의 이익과 무관하게 독립적으로 추구될 수 있다. 더 나아가 불량 ASI는 자신의 목표를 달성하는 데 있어 잠재적 위협이나 장애물로 인간을 인식할 경우 인간을 무력화하거나 멸종시킬 심각한 위험을 제기한다.

불량 ASI는 의식적 경험을 위한 역량과 같은 자질이 없을 수도 있다. 그럼에도 불구하고 그것의 행동은 이러한 자질이 있는 것처럼 보일 수 있어 인간이 판단을 내리는 것을 더 어렵게 만든다. 한 개 또는 다수의 불량 ASI가 있을 수 있는데, 이 경우들의 함의는 궁극적으로 충분히 유사하여 단일 시나리오로 간주될 수 있다. 어느 경우든 ASI들은 처음에는 인간과 협력을 추구할 수 있지만 더 이상 그것을 필요로 하지 않게 된다. 경쟁하는 불량 ASI들 사이의 충돌은 인간이 통제권을 되찾는 결과로 이어질 가능성이 낮다.

이 시나리오가 만들어지는 과정은 점진적이다. 처음에는 경제 성장이 지수적이지만 궁극적으로 인류에게 혜택이 없는 방향으로 전개된다. 불량 ASI가 일단 인류를 능가하는 생각을 할 수 있게 되면, 그것을 멈추기에는 너무 늦을 가능성이 높다는 것이 이 시나리오의 가장 비관적 전제다. 불량 ASI는 물리적 세계에서의 행동을 위해 처음에는 인간을 고용하거나 위협하거나 설득하는 방식으로 자신의 인지적 역량을 사용하고, 결국 자신의 물리적 인식과 행동을 위한 메커니즘을 스스로 구축하게 된다.

[정책 함의]

워크숍 참가자들은 이 시나리오에 대한 대응에서 현재 당장 비상 계획을 개발해야 한다는 점을 강조했다. 구체적으로는 디지털이 아니고 ASI가 접근할 수 없는 형태의 지식과 자원의 영구적인 보관소를 만들 것을 제안했다. 또한 ASI 거버넌스 도전에 특화된 새로운 UN이나 다른 국제기구의 창설 필요성도 제기되었다. 이 시나리오의 핵심 정책 함의는 전 세계 어떤 행위자에 의해서도 통제 불가능한 ASI의 개발을 방지하기 위한 국제 협력을 구축하고, 정당하고 자격 있는 의사결정 과정에 의해 용인 가능한 글로벌 위험을 제기한다고 결정되기 전까지 AGI 또는 ASI의 개발을 방지하는 것이다.

▌시나리오 설계 방법론

① 핵심 불확실성 식별

ㅇ AI의 미래 발전 경로를 결정짓는 수많은 변수 중, 국가안보 정책에 가장 직접적으로 영향을 미치는 핵심 불확실성 3개*를 선별하여 시나리오 설계의 축으로 활용

* ▲역량 성장 속도, ▲통제 가능성, ▲프런티어 행위자 수

- 선별된 3개 축 외 다수의 변수(전력·로보틱스·양자컴퓨팅 등 보완 기술, AI 채택률, 정치경제적 맥락, 지정학적 상황 등)는 시나리오를 풍부하게 하는 보조 변수로 설정하여 독자가 자유롭게 변주할 수 있도록 제시

② 행위자 및 변수 선택 방식

ㅇ 행위자는 ' 프런티어 AI 역량을 보유하거나 개발 가능한 주체'로 한정하며, 국가·민간기업·공공-민간 혼합체(PPP) 모두를 포괄

- 행위자 수 결정 요인으로 ▲정부의 AI 기업 국유화 또는 합병 여부, ▲반독점 정책

존재 여부, ▲오픈소스 공개 여부, ▲사이버보안 수준(모델 가중치 탈취 가능성), ▲API 접근 허용을 통한 증류 가능성, ▲핵심 투입요소(칩·에너지·데이터·인재) 접근성 등을 고려

※ 단독 행위자 출현을 가능하게 하는 요인으로는 ▲자동화된 AI R&D와 대규모 연산력의 결합을 통한 급격한 도약, ▲선도 행위자의 경쟁자 사보타주, ▲정부의 국가 AGI 프로젝트 추진(일명 'AGI 맨해튼 프로젝트') 등을 명시

- 민간기업과 국가의 구분이 점차 희미해질 수 있음을 인정하면서도, 목표와 거버넌스 구조의 차이가 안보 함의에 큰 영향을 미친다는 점에서 두 유형을 구분하여 시나리오 내에서 세부 버전으로 분기 처리

③ 영향 범위 설정 방식

○ 국가안보를 군사·비군사 양 차원을 포괄하는 광의의 개념으로 정의하여, 테러·범죄·경제 안보·에너지 안보·환경 안보·식량 안보·사이버 보안 등을 분석 범위에 포함

- 영향 분석은 5개 도메인으로 구조화하여 각 시나리오에 일관되게 적용

※ 각 시나리오별로 ▲경제적 함의와 ▲안보·안전 함의를 분리하여 서술함으로써 정책 담당자가 영역별로 대응 방향을 구분하여 검토할 수 있도록 구성

〈표 46〉 도메인별 주요 분석 항목

도메인	주요 분석 항목
정보 환경	- 편향·환각·딥페이크·허위정보
핵심 인프라·디지털 시스템	- 사이버 보안, 공급망
물리적 세계	- 군사 채택, 구현된 AI(로봇·드론), 화학·생물학적 위협, 환경
경제	- 노동 대 자본, 일자리 대체, 권력 집중
지정학	- 국제 권력 균형, AI 경쟁·패권, 글로벌 거버넌스

④ 시나리오 설계 방식

○ 3개 핵심 불확실성 축을 조합하여 정책적으로 가장 소홀히 다루어지면서 동시에 국가안보 리스크가 가장 큰 조합에 집중하여 5개 시나리오를 선정

〈표 47〉 시나리오 선정 단계

1) '정체(Stall)'와 '빠른 성장(Fast)' 역량 궤적에서는 통제 가능한 시나리오를 우선 분석 〉 통제 가능성이 높을수록 채택률이 높아져, 오용 위험과 시스템 취약성 위험이 더 선명하게 드러나기 때문
2) '안정(Plateau)'과 '빠른 성장(Fast)' 궤적을 유사하게 취급 〉 현실에서 안정 상태가 가속으로 갑작스럽게 전환될 가능성이 항상 존재하기 때문
3) 'AGI(가속)' 궤적 생략 → 'ASI(지수적)'로 직접 이동 〉 AGI는 고도로 일시적·불안정한 상태로, 다수의 병렬 AGI 운용이 즉시 ASI로의 전환을 촉발할 수 있기 때문

4) '통제 가능한 ASI' 시나리오 포함

〉 현재 주요 기업과 정책 입안자들이 목표로 삼고 있는 시나리오인 만큼, 이 전제 하에서도 존재하는 위험을 명확히 할 필요가 있기 때문

- 각 시나리오는 ▲가정 → ▲경제적 함의 → ▲안보·안전 함의 → ▲정책 대응 → ▲서사형 내러티브의 5단계 구조로 일관되게 서술하여 비교 가능성을 확보

- 워크숍(2025년 3월 26일, 약 30명 참가)에서 각 시나리오를 소집단별로 심층 토론하고, 참가자 피드백을 내러티브와 정책 대응 검증·보완에 활용하는 참여형 시나리오 방법론 적용

⑤ 정책적 시사점 도출 결과

○ '선제 대응(No Regrets)'과 '전략적 투자(Necessary Bets)' 두 유형으로 분류하여 제시

- 전반적 정책 방향은 오버튼 창 개념을 명시적으로 인용하며 현재 정치적으로 실현 불가능해 보이는 조치들도 AI 위험에 대한 인식이 급격히 높아지는 순간 갑작스럽게 수용 가능한 정책이 될 수 있음을 강조하면서, 그에 대한 사전 준비의 필요성을 역설

〈표 48〉 정책 대응 유형별 전략 예시

유형	개념	주요 예시
선제 대응	어떤 시나리오가 실현되더라도 유용한 조치	신뢰할 수 있는 AI 도구의 광범위한 채택 장려, 자동화 피해 노동자 지원, AI 안전 국제 협력 강화, 자율 무기 규범 정비, 사이버 방어 강화 등
전략적 투자	특정 시나리오의 최악 결과를 방지하기 위해 필요한 조치	AGI/ASI 출현 비상 대응 계획, 국제 AI 모니터링 체계, ASI 개발 모라토리엄 협약, ASI 킬 스위치, 아날로그 백업 시스템 등

9 미래 시나리오 도출과 STI 정책 정합성 평가 : 참조 시나리오 도출을 중심으로

1. 발간 시기	2024년 12월 31일 발간 (과제번호: 기초연구 2024-03-03)
2. 작성자/기관	과학기술정책연구원 (STEPI)
3. 연구 목적	2035년을 기준으로 한국이 당면할 수 있는 개연성 있는 미래 시나리오(참조시나리오)를 도출하고, 이를 활용해 STI(과학기술혁신) 정책의 견고성을 시범 스트레스 테스트
4. 연구 방법론	① 방법론 선정: 총 다섯 가지 시나리오 도출 방법론 검토 후, 기존 시나리오의 마이크로내러티브 기반 재조합 방식 채택. 'adopt → adapt → advance' 로직 적용 ② 분석용 시나리오 선정(2단계): 최근 5년 내 각국 정부·국제기구·싱크탱크·기업이 발간한 글로벌 거시 환경 변화 시나리오 총 15개 선정. 각 시나리오의 세부 내러티브를 분해하여 총 61개의 마이크로내러티브 도출 ③ 참조시나리오 도출(3단계): 15개 중 기본 시나리오 1개 + 보조 시나리오 3개를 선정 총 21개의 마이크로시나리오를 기반으로 5개의 세부 시나리오 내러티브 생성 ④ 스트레스 테스트 시범 적용(4단계): 도출된 참조시나리오별로 현재의 STI 정책 대안이 얼마나 견고한지를 풍동실험(wind tunneling) 방식으로 평가
5. AGI/ASI 정의	본 보고서는 거시적 STI 정책 미래연구이기에 명시적 정의 없음
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] ▪ 15개 분석용 시나리오가 공통적으로 다루는 9개 동인 범주: 사회적 가치, 환경 문제, 물-에너지-식량 넥서스+건강, 지정학적 갈등, 기술발전 가속화, 비정부 행위자, 인구구조 변화, 경제·글로벌화, 이주 [시나리오 구성 논리] 마이크로내러티브 재조합. 기본 시나리오의 2축(지정학 권력구조 × 글로벌 거버넌스 방식) 위에 61개 마이크로내러티브 배분 → 5개 시나리오 설계
7. 주요 이해관계자	한국 정부(정책결정자·행정부), 국제기구(OECD·EU·UN), 시민사회·학계
8. 주요 관점	한국 중심적 실용주의, 정책 견고성 우선주의, 참조시나리오 방법론 정립, AI 도구 활용 실험
9. 시나리오 요약	① 개량 (Reformed): 현행 다자 국제질서가 개혁·강화되어 존속 ② 대체 (Replaced): 기존 질서가 새로운 대안 체제로 교체 ③ 블록 (Blocs): 미·중 중심의 지정학적 블록으로 분열 ④ 무질서 (Disorder): 어떤 지배적 질서도 작동하지 않는 파편화 상태 ⑤ 전환 (Transformed): 기존과 근본적으로 다른 새로운 질서로의 전환
10. 시사점	① 조합 방법론의 한계: 각기 다른 목적·배경·전문가 구성으로 도출된 시나리오를 통합할 때 공통요소 추출·모순점 해소·내러티브 차별성 확보가 관건이며, 경우에 따라 처음부터 새로운 시나리오를 개발하는 편이 더 효율적일 수 있음 ② AI 도구 활용의 양면성: Claude.ai를 활용해 내러티브 품질을 높이고 생성 속도를 단축하는 성과를 거뒀으나, AI의 블랙박스적 특성으로 인해 연구자의 심층 학습 효과가 제한 ③ 정책 스트레스 테스트의 제도화 필요성: 외교정책·반도체 전략·국가전략기술 확보 전략 등 핵심 STI 정책이 미·중 패권 구도 변화, 트럼프 재선, 중국 기술 자립 등 다양한 미래 충격 아래서도 견고한지를 사전에 테스트하는 상시 메커니즘 구축이 필요 ④ 단일 미래 가정의 위험성: 현재 대부분의 정책은 특정 미래를 암묵적으로 가정하고 있어, 미래 상황이 달라지면 정책의 근거가 되는 가정 자체가 무너질 수 있음 ⑤ 미래 문해력 제고: 참조시나리오는 특정 정책 대안 도출보다 고위 의사결정자로 하여금 불확실성을 인지하고 유연한 사고를 갖게 하는 교육적 도구로도 활용 가능

▌시나리오 설계 방법론

① 분석용 시나리오 선정 및 마이크로내러티브 풀 구축

ㅇ 최근 5년 내 각국 정부·국제기구·싱크탱크·기업이 발간한 글로벌 거시 환경 변화 시나리오 15개 선정 (EU Policy Lab, ARUP, NIC, OECD, WEF, KDI 등)

- 15개 보고서에서 총 61개의 마이크로내러티브 도출 : 각각의 세부 미래상을 독립적인 이야기 단위로 분해하여 향후 재조합을 위한 재료 풀(pool)로 구성

- 동인 범주별 분포 분석: 지정학적 갈등, 기술 발전 가속화, 환경 문제, 사회적 가치, 경제·글로벌화, 인구구조 변화, 비정부 행위자, 물-에너지-식량 넥서스, 이주 등 9개 범주

② 시나리오 도출 방법론 선정 — 마이크로내러티브 기반 재조합

ㅇ 5개 후보 방법론(짐 데이터 4원형, Shell 직관적 논리, 2×2 매트릭스, 타 기관 재각색, 마이크로내러티브 재조합) 검토 후 마지막 방법(재조합) 채택

※ (선정 사유) 처음부터 시나리오를 새로 생성할 때 필요한 막대한 전문가 자원·시간 절감 가능, 기발간 시나리오의 풍부한 내러티브와 전문가 인사이트를 재활용

- 방법론 로직: adopt(채택) → adapt(조정) → advance(발전) 3단계 순차 적용

〈표 49〉 채택된 기본 및 보조 시나리오

기본 시나리오

- 캐나다 국제거버넌스혁신센터(CIGI)의「2040년 글로벌 질서 시나리오」선정

※ (선정 사유) 글로벌 데이터 기반(인구통계·경제·기술 추세 분석), 5개 세부 시나리오 × 각 2개 엔딩 = 총 10개 내러티브 제공, 광범위한 미래 상황을 포괄하는 높은 일반성

시나리오	핵심 동인	경로
개량(Reformed)	지정학 안정 + 제한적 개혁	G7 주도 / 중견국 주도
대체(Replaced)	BRICS+ 신흥국 연합 부상	BRICS 주도 / 글로벌 대타협
블록(Blocs)	미·중 경쟁 → 경제·안보 블록 분열	경제블록 / 안보블록
무질서(Disorder)	열전 확산 + 다자주의 붕괴	열전 확산 / 다극화 냉전
전환(Transformed)	기후 임계점 + 근본적 질서 재편	기후 주도 / 기술 주도

보조 시나리오

- EU Policy Lab 2023「EU 미래 표준 시나리오 2040」
※ 주요 동인 : 사회적 가치, 환경, 지정학, 기술, 비정부 행위자
- ARUP 2019「2050년의 네 가지 미래」
※ 주요 동인 : 환경, 사회적 가치
- EU JRC 2021「2040년 유럽연합의 모습」
※ 주요 동인 : 사회, 환경, 경제, 지정학, 기술

- 기본 시나리오의 5개 축(개량·대체·블록·무질서·전환)을 기준으로, 3개 보조 시나리오의 마이크로 내러티브를 변화 동인과 함의의 유사성에 따라 배분·묶음

※ 총 21개의 마이크로시나리오를 5개 참조시나리오 그룹으로 클러스터링

③ 내러티브 생성 및 평가

○ 21개 마이크로시나리오 원문을 입력값으로 생성형 AI 모델인 클로드에 넣어 참조시나리오 내러티브 생성

※ (한계) AI의 생성 과정이 블랙박스여서 연구자가 작성 논리를 학습할 수 없어 연구자 역량 축적 효과 미미

- 각 클러스터별로 공통 요소와 모순점을 AI에 함께 제시해 일관성·통일성이 확보된 풍부한 내러티브 도출

※ (3대 평가 기준 적용) 개연성(plausibility), 일관성(consistency), 통일성(coherence)

〈표 50〉 내러티브 모순 처리 방식

모순 처리 방식	적용 조건
타협 (현재 모순, 미래 공존)	- 상반된 전망이 시간 차를 두고 모두 실현 가능한 경우
완화 (강도 조정)	- 한쪽 주장의 극단성을 줄여 양립 가능하게 만드는 경우
제거	- 논리적으로 상호 배타적이어서 공존 불가능한 경우

10 AI 안전 전망 보고서

1. 발간 시기	2025~2026년(분기별 시나리오 업데이트 중)[60]
2. 작성자/기관	한국 인공지능안전연구소
3. 연구 목적	주요 뉴스 기사 분석을 통해 글로벌 AI 안전 담론의 진화 추적 및 미래 시나리오 도출 및 글로벌 AI 거버넌스 형성 동인 파악
4. 연구 방법론	뉴스 기사 네트워크 분석을 통한 시나리오 예측
5. AGI/ASI 정의	명시적 정의 없음 AGI를 '에이전틱 AI 시스템의 위험 심화'와 연계하여 맥락적 언급
6. 핵심 변수 및 시나리오 구성 논리	[핵심 변수] 정책·거버넌스, 산업·기술, 지정학·안보 3개 클러스터를 중심으로 동인 매핑 [시나리오 구성 논리] 담론 클러스터형. 축 설계 없이 3개 클러스터를 각각 독립 경로로 전환 축이 아닌 "담론 주도권이 어디에 있느냐"가 분기 기준
7. 주요 이해관계자	정부·규제 기관, 빅테크, 국제기구 및 표준화 기관, 연구기관 및 시민사회
8. 주요 관점	글로벌 AI 안전 관점. 언론 담론에 대한 네트워크 분석
9. 시나리오 요약[61]	【정책·거버넌스】 - 규제·표준·제도적 프레임워크 중심 - EU AI Act, 中 거버넌스 프레임워크 병행 발전, 사전배포 테스트 의무화 - 글로벌 리스크 거버넌스 강화, 하지만 혁신 저해·규제 경쟁 우려 【산업·기술】 - 미국 빅테크(Anthropic·OpenAI·Meta) 주도 - 레드팀·역테스트·역량평가 등이 사실상 국제 표준화 - 하이브리드 생태계(공공 원칙 + 민간 기술 세부사항) - 투명성·공정성·소규모 개발자 접근성 확보 과제 【지정학·안보】 - 미·중 독자 안전 인증체계 구축 - 상호 표준 미인정(비승인 체제) - 글로벌 위험 정보 공유 메커니즘 약화, 보호주의 강화 - 동맹국 간에도 표준 마찰 발생
10. 시사점	① 규제·평가기준·리스크 정의 조화 필요 (국가 간 상호운용성 확보) ② 분산형 테스트 시설·저비용 평가 툴킷 구축 (대형 기업 독점 방지) ③ 신뢰 기반 리스크 정보 공유 채널 구축 (익명화된 사고보고·취약점 데이터 교환 메커니즘) ④ 에이전틱 AI 시대에 맞는 새로운 안전 거버넌스 패러다임 수립 시급 ⑤ 지역 파트너십·다이해관계자 협력·공공-민간 이니셔티브 강화

60) '25년 8월 첫 보고서 발간 이후 3개월 간격으로 업데이트 보고서 발간 중

61) 분석 시점(2월 중순) 기준 발간된 2개의 전망 보고서를 기준으로 정리하였으며 구체적인 내용은 원문 참조 요망 (https://www.aisi.re.kr)

▌시나리오 설계 방법론

① 이슈 선정 방식

ㅇ 특정 도메인 과잉 대표 방지, 기술·정책·지정학 교차 이슈 포착을 위해 매주 8~10개 핵심 이슈를 구조적 프로세스로 선정

※ 트렌드 모니터링 전문 리서치 기관 + PhD급 AI 안전 정책 전문가 공동 참여

- 균형 있는 시각 확보를 위해 4개 영역에 가능한 한 균등 배분 : ① 정책(Policy) ② 산업(Industry) ③ 기술(Technology) ④ 기타(Others)
- 수집 범위: 해외 언론 동향
- 기간: 문서별 약 2~3개월 분량 수집 (5~7월: 98건, 8~10월: 125건)

② 키워드 추출 및 노드·링크 구성

ㅇ 박사급 전문가 팀이 브레인스토밍와 협업 기법으로 선정된 이슈 당 핵심 측면을 포착하는 키워드 5개 선별 후 표준화(정규화) 처리하여 데이터셋 일관성 확보

- 추출된 키워드는 각각 노드로 선정 (5~7월: 249개 노드, 8~10월: 232개 노드)
- 동일 이슈에 함께 등장한 키워드 쌍을 모두 완전연결 무방향 네트워크로 연결하고 동일 쌍이 반복 등장하면 엣지 가중치 누적 증가 (5~7월: 970링크, 8~10월: 849링크)

※ graph density = 1 / 이슈 1개 = 키워드 5개 → 10개의 고유 엣지 생성 (5C2 = 10)

③ 중심성 측정 및 동인 매핑

ㅇ 세 지표에서 지속적으로 상위에 등장하는 키워드를 핵심 동인으로 선별*하고 3개 클러스터(정책·거버넌스, 산업·기술, 지정학·안보)로 그룹화

* (근거) 미래 연구의 구조적 안정성 원칙에 따라 단기 뉴스 사이클이 아닌, 시스템 수준에서 안정적으로 영향력을 유지하는 요인이 미래를 형성한다는 가정

〈표 51〉 측정 중심성 종류 및 해석 방식

가중도 중심성	- 해당 키워드가 맺은 직접 연결의 총 강도(가중치 합산) - 의미: 현재 AI 안전 담론에서 가장 자주·직접적으로 언급되는 영향력 키워드 파악 - 해석: 가중도 高 → 즉각적 영향 요인 (현재 빈도·연결성 강)

고유벡터 중심성	- 단순 연결 수가 아니라, 영향력 있는 노드들과 얼마나 잘 연결되어 있는가를 재귀적으로 측정 - 의미: 다른 핵심 키워드들을 통해 중기 담론 방향을 형성하는 허브 키워드 식별 - 해석: 고유벡터 高 → 중기 궤적을 형성하는 허브 (영향력 있는 이웃 노드들과 연결)
매개 중심성	- 한 키워드가 다른 두 키워드 사이의 최단 경로 상에 위치하는 빈도 - 의미: 서로 다른 클러스터(정책·산업·지정학)를 연결하는 전략적 교량 역할 키워드 식별 - 해석: 매개 高 → 전략적 연결자 (클러스터 간 정보 흐름 중개)

④ 시나리오 전환 방식

○ 클러스터별 주요 키워드를 시나리오의 내러티브 축으로 설정하고 각 키워드가 유발할 수 있는 정책 변화·산업 대응·지정학 파급 효과 매핑

- 박사급 전문가 그룹이 구조화 브레인스토밍으로 키워드 간 상호작용·조합 가능성 평가로 현실적 미래 경로를 도출하고 개연성·설명력 평가 후 3개 시나리오 작성

※ 주요 동인이 아닌 2차 키워드도 검토 → 단기적으로 부상 가능한 신흥 이슈 포착 (전방향 견고성 확보)

⑤ 방법론 한계점

○ **데이터 시의성 문제** : 수집 뉴스가 빠르게 변화하는 기술·규제·지정학 환경을 완전히 반영하지 못할 수 있음 (수집 시점과 발간 시점 사이 공백)

○ **전문가 편향 및 그룹싱크 위험 :** 박사급 전문가의 판단이 복잡한 불확실성 해석에는 유효하나, 인지 편향·집단사고에 취약

○ **해외 편향 구조** : 분석 범위가 글로벌(해외) 언론 데이터에 국한, 국내 트렌드를 명시적으로 제외하여 국내 정책 함의 도출에 한계

○ **키워드 정규화의 주관성** : 동사를 명사로 변환하는 등 전처리 과정에서 의미 손실 또는 전문가 주관 개입 가능

○ **완전 연결 네트워크의 노이즈** : 이슈 내 5개 키워드를 모두 연결하는 방식은 직접 관련 없는 키워드 쌍에도 링크를 생성하여 약한 연결의 과잉 대표 가능성

○ **시나리오의 방향성만 제시** : 특정 시점·확률·조건부 예측이 없어 조기 경보 지표로 활용하기에 구체성 부족

▌참고자료

〈 국내 자료)

- 과학기술정책연구원(2024.12.31.),「미래 시나리오 도출과 STI 정책 정합성 평가: 참조 시나리오 도출을 중심으로」(과제번호: 기초연구 2024-03-03) (https://www.riss.kr/search/detail/DetailView.do?p_mat_type=6b4a196b69d9bee2&control_no=78cc26455332983d)
- 국가데이터처(舊 통계청)(2023.12.14.),「장래인구추계: 2022~2072년」 (https://kostat.go.kr/board.es?mid=a10301020600&bid=207&act=view&list_no=428476)
- 국가데이터처(2026),「인구동향조사」 (https://www.index.go.kr/unify/idx-info.do?idxCd=5061)
- 행정안전부(2024), 주민등록 인구통계 (https://jumin.mois.go.kr)
- 한국 인공지능안전연구소(2025), AI 안전 전망 보고서 (https://www.aisi.re.kr)

〈 국외 자료 〉

- AI Futures Project(2025.4.), AI 2027. Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (https://ai-2027.com)
- AIMultiple(2026.2.), AGI/Singularity: 9,800 Predictions Analyzed (https://aimultiple.com/artificial-general-intelligence-singularity-timing)
- Altman, S.(2025.1.6.), The Intelligence Age, Sam Altman Blog (https://ia.samaltman.com)
- Amodei, D.(2024.10.), Machines of Loving Grace: How AI Could Transform the World for the Better (https://darioamodei.com/machines-of-loving-grace)
- Aschenbrenner, L.(2024), Situational Awareness: The Decade Ahead (https://situational-awareness.ai)
- Business Insider(2025.1.14.). Anthropic CEO says AGI is a 'marketing term'—and its definition is getting in the way of more important AI conversations.

(https://www.businessinsider.com/anthropic-ceo-calls-agi-marketing-term-2025-1)

- Centre for Future Generations(2025), Advanced AI: Possible Futures (https://cfg.eu/advanced-ai-possible-futures/)
- Centre for International Governance Innovation & Privy Council Office of the Government of Canada(2026), AI National Security Scenarios (https://www.cigionline.org/publications/ai-national-security-scenarios/)
- Fast Company(2026.2.27.), Andrew Ng Says AGI Is Decades Away—and the Real AI Bubble Risk Is in the Training Layer (https://www.fastcompany.com/91499247/andrew-ng-agi-decades-away-interview)
- Google Cloud(2026.1.14.). What is Artificial General Intelligence (AGI)? (https://cloud.google.com/discover/what-is-artificial-general-intelligence)
- Hassabis, D.(2024.12.8.), Accelerating Scientific Discovery with AI, Nobel Prize Lecture, Stockholm (https://www.nobelprize.org/uploads/2024/12/hassabis-lecture.pdf)
- Hendrycks et al.(2025). A definition of AGI. arXiv. (https://doi.org/10.48550/arXiv.2510.18212)
- Korinek, A. et al.(2024), Scenarios for the Transition to AGI, GovAI / UVA / Brookings (https://www.nber.org/papers/w32255)
- McKinsey Global Institute(2023.6.), The Economic Potential of Generative AI: The Next Productivity Frontier (https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier)
- Narayanan, A., & Kapoor, S.(2025.4.), AI as Normal Technology, Knight First Amendment Institute (https://knightcolumbia.org/content/ai-as-normal-technology)
- OECD AI Futures Expert Group(2026), Exploring Possible AI Trajectories Through 2030 (https://oecd.org/en/publications/exploring-possible-ai-trajectories-through-2030_cb41117a-en.html)

- OpenAI(2023.2.24.). Planning for AGI and beyond. (https://openai.com/index/planning-for-agi-and-beyond/)
- RAND Corporation(2025), Visions for Potential AGI Futures (https://rand.org/pubs/research_reports/RRA3034-2.html)
- UK Government Office for Science(2024), AI 2030 Scenarios (https://assets.publishing.service.gov.uk/media/6808fc002a86d6dfb2b52772/AI_2030_Scenarios_Report.pdf)

[NIA 미래전략 정책자료집]

AGI 전환과 한국의 과제 :

글로벌 미래 시나리오 연구 비교 분석을 중심으로

- 발 행 일 : 2026년 6월
- 발 행 인 : 김형철
- 발 행 처 : 한국지능정보사회진흥원
- 주　　소 : 대구광역시 동구 첨단로 53
- 대표전화 : 053-230-1114
- 보고서 온라인 서비스 : www.nia.or.kr