www.eai.or.kr

EAI 워킹페이퍼

# AI 혁명과 공화주의 안보이론: 무정부와 위계의 이중 난제 재부상

차태서(성균관대학교)

EAI
동아시아연구원

EAI 워킹페이퍼

# AI 혁명과 공화주의 안보이론: 무정부와 위계의 이중 난제 재부상

**차태서**
**성균관대학교 정치외교학과**

## Ⅰ. 기술발전과 문명 패러다임의 변동

중대한 기술 변화는 그 자체로 사회적 격동의 과정을 초래한다. 국제관계 차원에서 새로운 승자와 패자를 만들고, 행위자의 선호를 바꾸며, 새로운 규범과 조직의 구성을 가능하게 한다. 반대로 당대 국제정치의 특징이 기술변화의 내용과 속도에 영향을 미치기도 한다.[1] 오늘날 가장 각광을 받고 있는 첨단기술 분야로서 인공지능(Artificial Intelligence, AI)이 국제정치에 미치는 영향을 가늠하기 위해서는 우선 AI 자체를 인류의 장기구조적 기술경제사의 맥락에 위치 지워볼 필요가 있다.

일단 AI는 광범위한 활용성, 보완적 혁신 촉진 가능성, 지속적인 기술적 개선 가능성 등을 보유한다는 점에서 전통적인 "범용기술(General Purpose Technologies, GPT)"의 특징을 보유한 것으로 평가된다. 이는 과거 전기, 내연기관, 컴퓨터 등의 사례에 조응하는 것으로 이 정도 레벨만으로도 사회경제 전반에 걸쳐 다양하고도 폭넓은 변화를 일으킬 잠재력을 지닌 테크놀로지 혁신으로 간주된다. 그러나 일각에서는 AI가 이런 GPT 수준을 넘어 과거 농경·가축 사육과 화석연료 사용 등 인류사에 있어 에너지 포획의 비약적 향상을 가져왔던 "혁명적 기술(Revolutionary Technologies, RT)"의 반열에 이를 수 있다는 예측을 내놓고 있다. 인류사에 있어 혁명적 기술은 앞서 언급한 두 가지 사례밖에 존재하지 않았으며, 이는 각각 약 1만 년 전의 신석기 혁명과 18-19 세기의 산업혁명을 발생시켜 인류 문명 구조의 근본적 변환을 잉태한 바 있다.[2]

---

[1] Daniel W. Drezner, "Technological Change and International Relations," *International Relations* 33 (2), 2019, pp. 286–303.
[2] Ben Garfinkel, "The Impact of Artificial Intelligence: A Historical Perspective," in *The Oxford Handbook of AI Governance*, edited by Justin B. Bullock et al., Oxford: Oxford University Press, 2024.

AI가 미래에 GPT로 판명될지, RT로 판명될지는 아직 정확히 가늠할 수 없으나, 국제정치 영역에 있어 기존의 권력경쟁, 안보 딜레마, 정치체제 변동 등의 트렌드를 가속(acceleration)하는 중대한 촉매기술로서 작동하여, 세계사의 불안정성과 불확실성을 더욱 증폭시킬 가능성에 대해 많은 전문가들이 주목하고 있다.[3]

본 연구는 특히 "공화주의 안보이론" 혹은 "조직원리(organizing principle)와 안보"라는 문제 틀[4] 속에 AI 혁명이 국제정치에 가져올 함의를 이해해 보고자 한다. 국제정치 사상사에서 공화주의자들은 국내외의 공간에서 무정부 상태(anarchy)와 위계(hierarchy)라는 양대 심연이 만들어내는 불안정(insecurity)을 극복해 내는 정치적 제도 구성(=negarchy)의 가능성을 고민해 왔는데, 본 연구는 AI라는 기술혁신이 이 같은 공화주의 안보이론의 중심 난제를 근본적 차원에서 격화시키는 촉매제로 작동한다고 간주한다.

AI는 계몽주의 프로젝트의 마지막 판본으로서 인류의 마지막 발명품이 될 것인가? 생존 그 자체 등 인간종의 필수적 가치를 위협할 수 있는 AI—특히 미래에 도래할 초지능(Artificial Superintelligence, ASI)—의 거대한 잠재력을 어떻게 통제할 것인가라는 화두는 공화주의 정치철학의 오랜 난제의 연장선상에 놓여 있다.[5] 이하 본문에서는 특히 AI가 초래할 안보 상 위험의 세 가지 경우의 수를 분류해 살펴보려 하는데, ① 국제 무정부 상태에서 패권경쟁-핵전쟁 가능성의 심화 가능성, ② 국내차원에서 위계의 공고화(디지털 독재 부상 및 민주주의의 쇠퇴), ③ 국제차원에서 위계의 강화(기성 국제 불평등 심화+인류 멸종의 위협) 등의 순으로 글을 전개하고자 한다.

## Ⅱ. AI와 국제 무정부(Anarchy) 문제의 심화

오늘날 탈 단극, 강대국 간 전략 경쟁 시대가 도래했다. 냉전 시대부터 구축되어 온 전 지구적 전략적 안정성과 군비통제의 제도적 배치가 붕괴되어 가는 환경이 부상한 것이다. 그런데 AI로 대표되는 4차 산업혁명이 이 같은 구조적 불확실성에 기름을 끼얹고 있다.[6] 반대로 이야기하면, 단극의 안정적

[3] Amelia C. Arsenault and Sarah E. Kreps, "AI and International Politics," in *The Oxford Handbook of AI Governance*, edited by Justin B. Bullock et al., Oxford: Oxford University Press, 2024.
[4] Daniel H. Deudney, *Bounding Power: Republican Security Theory from the Polis to the Global Village*, Princeton: Princeton University Press, 2007.
[5] Daniel Deudney and Devanshu Singh, "Bounding Superpowers: The ASI Control Problem, Public Safety, and Republican Constitutionalism," in *Polycentric Federalism and World Orders*, edited by Brandon Christensen, Switzerland: Springer Nature, 2025, pp. 465–502.
[6] Paul van Hooft, Lotje Boswinkel, and Tim Sweijs, *Shifting Sands of Strategic Stability: Towards a New Arms Control Agenda*, The Hague: The Hague Centre for Strategic Studies, 2022.


ⓒ EAI 2025

역사국면이 아닌 현재와 같은 지정학적 긴장이 고조된 시기에 AI가 출현한 것이 인류에게 악재인 측면이 존재한다. 마치 제2차 세계대전과 냉전이라는 고도의 지정학적 압박 속에서 급속도로 핵개발 및 군비경쟁이 일어난 것과 유사한 국제환경이 현대에 조성되어 있는 셈이다.[7] 이는 국제적 다극화가 부르는 무정부 상태의 경쟁 압력과 AI 군비경쟁의 상호 상승작용을 통제하는 문제가 시급한 안보 문제로 대두되어 있음을 의미한다.

다시 말해, 강대국 간 전략 경쟁의 핵심 전장으로서 인공지능이 부각되고 있는데, 특히 미국의 첨단기술 봉쇄 전략과 중국의 정부 주도 기술 자립화 맞대응 전략이 충돌 중이다.[8] 역사적으로 새로운 GPT의 등장과 확산(diffusion) 과정은 강대국 간 경제패권 변동을 설명하는 강력한 변수로 존재해 왔다. 가령, 1차 산업혁명기 제철 등의 범용기술 확산을 선도한 영국은 급격한 생산성 증대를 통해 패권국으로 성장할 수 있었으며, 19세기 말-20세기 초 미국은 전기화와 기계화라는 GPT 확산을 통해 2차 산업혁명을 선도하고 새로운 패권국으로 나설 수 있었다. 반면, 3차 산업혁명기 일본은 전자제품과 IT 등 일부 산업에서 선도적 위치를 차지했음에도 불구하고 컴퓨터 과학 및 소프트웨어 엔지니어링 등 GPT 확산에 뒤짐으로써 미국을 따라잡는 데 실패하였다. 이런 맥락에서 볼 때 4차 산업혁명기의 GPT인 AI 혁신을 누가 선도하고 제분야에 확산하는데 성공할지의 여부가 오늘날 미중 패권경쟁의 향방을 가를 것으로 예상된다.[9] 이 같은 AI를 둘러싼 두 초 강대국 간 경합의 양상은 특히 세력전이의 직접적인 요소가 될 군사기술 경쟁 분야에서도 급박하게 진행되고 있는데, 도전국인 중국의 경우 "지능화 전쟁" 수행능력의 강화가 미국과의 기존 군사적 격차를 단번에 줄일 수 있는 "곡선 도로 추월"과 같다는 인식을 지니고 있으며, 이에 맞서 미국 역시 세력전이의 방지를 위해 AI 분야에서의 "전일적 지배(global dominance)"를 추구하고 있다.[10]

---

[7] Seth D. Baum, Robert de Neufville, Anthony M. Barrett, and Gary Ackerman, "Lessons for Artificial Intelligence from Other Global Risks," in *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, Boca Raton, FL: Chapman and Hall/CRC, 2022, pp. 111–115.

[8] George S. Takach, *Cold War 2.0: Artificial Intelligence in the New War between China, Russia, and America*, New York: Pegasus Books, 2024.

[9] Jeffrey Ding, "The Rise and Fall of Technological Leadership: General-Purpose Technology Diffusion and Economic Power Transitions," *International Studies Quarterly* 68 (2), 2024.

[10] 이재준, 「미중 간 군사 기술 경쟁과 세력 전이: 인공지능, 자율무기체계 군사 기술을 중심으로」, 『한국과 국제정치』 38 (3), 2022.; Office of Science and Technology Policy, "America's AI Action Plan," The White House, July 10, 2025, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.; The White House, "Launching the Genesis Mission," November 24, 2025. https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/.


ⓒ EAI 2025

그런데 이 같은 상황의 전개가 국제정치질서 차원에서 근본적 리스크를 제기하는 것은 이른바 "AI-핵 넥서스"의 등장으로 인해 핵보유국 간 안보 딜레마가 심화되고 전략적/위기 안정성이 약화되어 심각하게 핵전쟁의 위험성이 증가하게 될 가능성이 점쳐지고 있기 때문이다.[11] 무엇보다 우선적으로 기성 핵무기 체계에 AI가 활용되면서 인간 의사결정의 탈 중심화/자동화와 결정 속도의 가속화가 이루어지면서 기성 억지이론의 전제 혹은 가정이 붕괴되는 사태가 벌어질 수 있다. 여기에 더해 AI 기술이 정찰, 표적획득, 유도 시스템에 결합되면 핵무기의 가시성이 높아짐에 따라 핵 보복 자산(=2차타격능력) 선제 무력화 가능성이 증가하게 되고, 이로 인해 전략적 안정성이 크게 약화될 소지가 있다. 이는 곧 핵보유국들의 군비 경쟁 압박을 높일 뿐만 아니라 "조기경보 발사(launch on warning)" 같은 공세적 핵 교리 채택을 부추기게 되어 위기 안정성마저 떨어지게 만들며, 결과적으로 "비 의도적 확전(inadvertent escalation)" 위험이 구조적으로 증폭되는 결과를 가져오게 된다.[12]

설상가상으로 AI-핵 넥서스의 부상이 국제관계사 국면 상 미중러 전략 경쟁시대 혹은 다극화라는 "새로운 핵 시대"의 맥락과 접합되어 있다는 점이 상황을 더욱 어렵게 만든다. 비교적 안정적이었던 냉전기 미소 양국 간의 "공포의 균형" 시대에 비해, 지금은 핵보유국 숫자가 증가해 기본적으로 억지게임의 복잡성이 크게 증가했을 뿐만 아니라, 냉전기 건설되었던 핵 군비통제 제도들마저 전략 경쟁의 희생양으로 거의 전부 소멸된 상태라는 점이 사태의 심각성을 더한다.[13] 그런데 AI 기술의 도입이 핵 군비경쟁에서 결정적인 "선점 우위 효과(First mover effect)"를 지닐 수 있다는 열강들의 공포가 이에 추가되면 걷잡을 수 없이 안보 딜레마가 강화되는 결과가 초래될 수 있다.[14]

이와 같은 현 국제 안보상황은 마치 구 냉전 초기 이른바 "핵의 정치적 문제(nuclear political question)"가 처음 부상했을 때와 역사적 데자뷰를 느끼게 만든다. 당시 역사상 초유의 파괴력을 지닌 원자탄의 등장은 폭력 상호 의존도의 극단적인 증대를 의미했고, 과연 세계적 무정부

---

[11] 정구연, 「인공지능의 군사적 활용과 전략적 안정성」, 『경제와 사회』 제143호, 2024.

[12] James Johnson, *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age*, Oxford: Oxford University Press, 2023; Vladislav Chernavskikh and Jules Palayer, "Impact of Military Artificial Intelligence on Nuclear Escalation Risk," *SIPRI Insights on Peace and Security* No. 2025/06, 2025.

[13] Jacob Stokes, Colin H. Kahl, Andrea Kendall-Taylor, and Nicholas Lokker, *Averting AI Armageddon: U.S.-China-Russia Rivalry at the Nexus of Nuclear Weapons and Artificial Intelligence*, Washington, DC: Center for a New American Security, 2025; Jon B. Wolfsthal, Hans Kristensen, and Matt Korda, "Why We Should Worry About Nuclear Weapons Again," *The Washington Post*, https://www.washingtonpost.com/opinions/interactive/2025/us-russia-nuclear-weapons-proliferation-danger/.; June 4, 2025; Anton La Guardia, "The Perils of the World's Third Nuclear Age," *The Economist*, November 20, 2024, https://www.economist.com/the-world-ahead/2024/11/20/the-perils-of-the-worlds-third-nuclear-age.

[14] Baele, Stephane J., Iqraa Bukhari, Christopher Whyte, Scott Cuomo, Benjamin Jensen, Kenneth Payne, and Eugenio V. Garcia, "AI IR: Charting International Relations in the Age of Artificial Intelligence," *International Studies Review* 26 (2), 2024.


ⓒ EAI 2025

상태 하의 “베스트 팔렌 체계”가 지속 가능한가라는 근본적 화두를 제기하였다. 그리고 이 난제에 대해 한스 모겐소(Hans Morgenthau), 존 헤르츠(John Herz) 같은 고전 현실주의자들조차 인류 멸종을 야기할 수도 있는 국민 국가체계가 더 이상 지속 불가능함을 주장하며, 단일 세계정부의 건설 필요성(nuclear one worldism)을 주장하기까지 하였다. 물론 20세기 후반의 실제 현실은 국가 간 체제는 그대로 둔 채 미소 합의하 상호 억지(deterrence) 및 군비통제(arms control)의 제도화로 타협되었긴 하지만 말이다.[15]

그러나 오늘날 핵 군비경쟁이 재 점화된 “제3의 핵 시대”에 AI기술과 핵무기의 연결이 전략적 안정성을 크게 떨어뜨리고 비 의도된 핵전쟁 발발의 가능성이 고조되는 국면이 부상하면서, 우리는 구 냉전 초기에 제기되었던 세계질서의 근본적인 화두에 다시 한번 직면하게 되었다. 과연 이번에는 국가 간 체계가 지속가능한가? 지구사에 있어 “6번째 대 멸종”을 막기 위해서는 어떤 형태의 세계질서와 군비통제 제도화가 필요한가?

## Ⅲ. AI와 국내외 위계(Hierarchy) 문제의 심화

AI기술의 확산은 국내와 국제 공간의 기존 위계를 공고화 하거나 새로 만들어내고, 더 극단적으로는 인류의 “실존적 위기(existential risk, x-risk)”를 생성해내는 계기가 될 수도 있다는 점에서 또 다른 차원의 중대한 화두들을 제기하고 있다.

### 1. 급진적 혁신론: AI에 의한 전쟁의 급속한 변혁

AI는 지능을 단일하고 자기 완결적인 특성을 지닌 것으로 간주함으로써 그 설계된 시스템 자체가 기본적으로 중앙집권적 아키텍처의 모습을 띠기 쉽다. 이에 더해 AI의 훈련과 운영에는 막대한 자원이 필요하기 때문에 대기업이나 정부 등이 AI 개발을 독점할 가능성이 높다. 더구나 AI는 중앙권력이 시민을 미시적으로 통제하고 감시하는 기술과 결합하기 쉬운 것으로 알려져 있다. 따라서 기술적 특성상 AI는 본질적으로 반 민주적인 권위주의적 집중화와 친화성을 지녔다는 평가를 받기도 한다.[16]

---

[15] Daniel H. Deudney, *Bounding Power: Republican Security Theory from the Polis to the Global Village*, Princeton: Princeton University Press, 2007.

[16] Adam Elkus, “Are AI and Democracy Compatible?” *Foreign Policy*, August 29, 2025, https://foreignpolicy.com/2025/08/29/ai-democracy-dictatorship-agi-governance/.; George Gilder and Peter Thiel, “COSM Speaker Peter Thiel: The Failures and 'Self-Hatred' of Big Tech,” *Mind Matters*, October 14, 2021, https://mindmatters.ai/wp-content/uploads/sites/2/2021/10/Mind-Matters-Episode-156-Peter-Thiel-at-COSM-rev1.pdf.


ⓒ EAI 2025

그리하여 AI가 기성 권위주의 사회와 민주주의 사회 모두에서 “디지털 독재”의 미래를 열 가능성에 대해서 논쟁이 진행되고 있다.[17]

가령, 코로나 팬데믹 기간 확인된 “디지털 판옵티콘”의 출현가능성은 AI기술에 기반해 푸코적인 “감시사회”가 진화하고 있다는 우려를 낳았다.[18] 이런 지점에 있어 가장 선구적(?)인 “AI기반 감시국가”의 미래를 선보인 곳은 다름 아닌 중국이다. 즉, 중국의 감시체계는 지난 20여 년간 행정·치안 데이터 디지털화, 도시 단위 AI 통합, 성(省) 단위 공유로 진화하며, 7억 대에 달하는 CCTV와 모바일·IoT를 포괄하는 거대 인프라를 구축해 왔다. 특히 초기의 수동적 감시에서 벗어나, 중국의 AI는 경찰의 정보 수집–지휘–배치–순찰 전체 주기를 점차 자동화하며 능동형 감시·통제 체제로 전환 중인데, 도시 전체 데이터를 실시간 통합·분석하고, AI 에이전트가 스스로 결정을 실행하는 방향으로 진화하고 있다. 또한 드론·자율주행차·휴머노이드 로봇 등 “AI-구동 집행 수단”이 등장하며, 감시는 온라인 제약을 넘어 오프라인 현장 개입과 물리적 통제 단계로 확대되고 있다. 이러한 통합은 최종적으로 중앙 지도부가 전국을 실시간으로 내려다보는 판옵티콘적 구조 완성을 지향하며, 중국 공산당은 인간 관료 의존을 줄이고 AI 중심의 자동 통치를 목표로 삼고 있다.[19]

다른 한편 AI는 꼭 전체주의 국가에 의한 초 중앙집권화라는 경로가 아니더라도 기존 민주주의 체제의 구조적 토대를 약화시킬 수 있는 장기 위험요소들을 내포한다. 첫째, 정치 커뮤니케이션의 설계와 전달을 AI가 대체하면 대표성과 시민–정부 간 피드백 루프가 단절되어, 대의제 핵심인 숙의·응답성 메커니즘이 무너질 위험이 있다. 둘째, AI는 부·권력의 급격한 집중과 불평등을 심화 시켜, 민주주의를 형식만 남는 껍데기로 만들고 과두제적 경향을 가속할 수 있다. 셋째, 뉴스·지식 생산이 대규모 언어 모델(LLM)을 중심으로 재편되면 정보 생태계가 소수 빅테크에 종속되며, 이미 취약해진 민주적 공론장은 더 큰 “인지적 위기”를 맞게 될 것이다. 이상의 세 흐름이 결합될 경우, 민주주의는 외형만 유지한 채 기술관료적·모조적(synthetic) 민주주의로 타락할 가능성까지 거론된다.[20] 넷째, 이른바 “감시 자본주의”의 도래에 따라 전체주의적 정부가 아닌 전체주의적 “기업” 네트워크 체제에 의해 우리의 일상이 포획되는 상황이 나타날 수 있다. 이는 중앙집권적 국가기구 대신 화폐화와 이윤

[17] 유발 하라리, 『넥서스: 석기시대부터 AI까지, 정보 네트워크로 보는 인류 역사』, 김명주 역, 서울: 김영사, 2024.

[18] Mark Coeckelbergh, “Vulnerability, AI, and Power in a Global Context: From Being-at-Risk to Biopolitics in the COVID-19 Pandemic,” in *The Global Politics of Artificial Intelligence*, New York: CRC, 2022.

[19] Valentin Weber, “China’s AI-Powered Surveillance State,” *Journal of Democracy* 36 (4), 2025.

[20] Dean Jackson and Samuel Woolley, “AI’s Real Dangers for Democracy,” *Journal of Democracy* 36 (4), 2025.


ⓒ EAI 2025

의 논리에 따라 우리의 삶이 인공지능과 데이터 과학을 통해 기록되고 규율 되는 또 다른 형태의 탈근대적 전체주의의 양상이다.[21]

2. 국제정치경제: Global North & South 격차 강화

국제정치경제의 관점에서 보면, 기존에 존재해 온 북반구와 남반구 간 격차는 AI 시대에 더욱 강화될 전망이다. AI 발전의 저변에는 알고리즘 독점과 디지털 신식 민주주의가 놓여 있는데, 이는 서구 기업들이 저개발국의 데이터를 대규모로 추출하고 값싼 노동을 착취하는 구조를 통해 유지된다. 다시 말해, AI의 국제정치경제는 노동의 외양을 바꾸는 것처럼 보이지만, 실제로는 기존의 경제 권력 구조를 거의 변화시키지 않는다. 더 나아가 AI의 군사적 무기화는 강대국 간 경쟁에 의해 주도되고 있으며, 그 실험장은 주로 남반구 국가들이다. 강대국의 군대가 전투머신과 자동화 장비로 무장해 가는 동안, 기술적으로 열세인 남반구에서는 저강도 분쟁과 대리전이 반복되고, 이 불균형은 결국 강대국에게 결정적 군사적 우위를 제공한다.[22]

3. X-risk: 특이점 이후 ASI의 위험성

2023년 5월, 비영리단체 AI 안전센터(Center for AI Safety)는 “AI로 인한 멸종 위험을 완화하는 것은 전염병이나 핵 전쟁 위험 등과 함께 전 세계에서 우선순위로 다뤄져야 한다”는 짤막한 성명서를 발표했다. 흥미롭게도 이 성명서에는 오픈AI의 샘 알트먼, 구글 딥마인드의 데미스 하사비스, 안트로픽의 다리오 아모데이 등 AI 주요 기업 경영진을 포함해 제프리 힌튼 토론토대 교수와 요슈아 벤지오 몬트리올대 교수, 케빈 스콧 마이크로소프트 최고기술책임자 등 핵심 과학자 350여 명이 서명했다.[23]

---

[21] 샤나 주보프, 『감시 자본주의 시대: 권력의 새로운 개척지에서 벌어지는 인류의 미래를 위한 투쟁』, 김보영 역, 서울: 문학사상사, 2021.; 마크 코켈버그, 『인공지능은 왜 정치적일 수밖에 없는가: AI의 정치학과 자유, 평등, 정의, 민주주의, 권력, 동물과 환경』, 배현석 역, 서울: 생각이음, 2023.

[22] Eugenio V. Garcia, “Conclusions: Charting the Challenge of AI IR,” *International Studies Review* 26 (2), 2024.

[23] 선명수, 「AI, 핵전쟁·전염병 만큼 위험…AI개발자들, 인류멸종 경고」, 『경향신문』, 2023년 5월 31일. https://www.khan.co.kr/article/202305311305001.



ⓒ EAI 2025

AI의 진화가 특이점(singularity)[24]을 초과해 일반지능, 나아가 초지능 단계에 이르렀을 때 그 통제 불가능성에 의해 인류의 생존 자체를 위협하게 될 것이라는 경고가 본격적으로 나오기 시작한 것은 닉 보스트롬(Nick Bostrom)의 『슈퍼인텔리전스』[25]가 발간된 2014년 이후이다. 이러한 묵시론적 논의는 보스트롬 자신의 "존재론적 위험(existential risk, x-risk)"[26] 가설과 "취약한 세계(vulnerable world)"[27] 가설의 문제의식과 맞닿아 있다. 우선 "존재론적 위험" 테제는 현대 기술문명을 인류사 최초로 전 지구적 멸종 혹은 회복 불가능한 잠재력의 소멸을 초래할 수 있는 위기 단계로 규정하며, 이를 인류 전체를 영구히 단절시키거나, 지적 생명의 향후 전개 가능성을 근본적으로 파괴하는(global, terminal) 위험으로 정의한다. 이런 존재론적 위험은 원자력, 나노기술, 생명공학, 인공지능 등 가속되는 기술 진보가 불러온 전혀 새로운 차원의 위험이다. 비슷한 맥락에서 "취약한 세계" 테제는 일정 수준의 기술능력에 도달하면 문명이 기본적으로 파국을 향할 수밖에 없는 상태에 도달할 수 있다고 말한다. 문제는 일단 발명된 기술을 '탈 발명'할 수 없다는 비 가역성과, 현대 문명이 현재 "준-무정부적 기본 상태(semi-anarchic default condition)"에 놓여 있어 이러한 취약성을 효과적으로 집단 관리할 수 없다는 데 있다.

그런데 이상의 두 개념은 AI와 직접적으로 연관된다. 고도로 발전된 ASI는 한편으로는 인류 전체의 존속과 미래 잠재력 자체를 위협할 수 있는 새로운 유형의 존재론적 위험을 구성하며, 다른 한편으로는 국가·기업·개인이 고위험 역량을 손쉽게 보유하거나 오용할 수 있게 함으로써 기술 문명을 취약한 세계로 미끄러뜨릴 수 있다.

보다 구체적으로 AI가 야기할 수 있는 존재론적 위험의 8가지 범주를 요약하면 다음과 같다. 우선 무기화(weaponization)는 강력한 AI가 초지능적 군사자산, 대량살상 무기 설계 등으로 전용될 때, 인류 공동체가 스스로를 파괴할 능력을 AI에 넘겨주는 결과가 될 수 있다. 둘째, 기능 쇠약(enfeeblement)은 사회와 개인이 점점 더 많은 기능을 AI에 의존하면서, 스스로 문제를 해결하고 위기를 관리할 능력이 약해져, 결정적 순간에 AI 없는 생존 자체가 불가능해질 수 있다. 셋째, 지식 기반 붕괴(eroded epistemics)는 AI가 정보환경을 왜곡하고 사회 전체의 사실 판단능력을 무너뜨릴 경우, 민주적·집단적 의사결정이 마비되어 문명이 스스로를 구할 수 없는 상태에 빠질 위험을 의미한

[24] 레이 커즈와일, 『마침내 특이점이 시작된다: 인류가 AI와 결합하는 순간』, 이충호 역, 서울: 비즈니스북스, 2025.

[25] 닉 보스트롬, 『슈퍼인텔리전스: 경로, 위험, 전략』, 조성진 역, 서울: 까치, 2017.

[26] Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology* 9, 2002.

[27] Nick Bostrom, "The Vulnerable World Hypothesis," *Journal of Evolution and Technology* 10 (4), 2019.



ⓒ EAI 2025

다. 넷째, 잘못된 목표설정(proxy misspecification)은 인간이 의도한 목적이 아니라 엉뚱한 수치나 지표를 AI가 최적화하며, 그 과정에서 인간의 안전·가치·제도를 파괴할 수 있는 위험이다. 다섯째, 가치 고착(value lock-in)은 특정 집단이나 체제가 AI를 통해 자신들의 권력과 이념을 영구화 할 때, 인류가 더 나은 미래로 나아갈 수 있는 경로가 닫혀버리는 결과를 낳을 수 있다. 여섯째, 예상 밖의 능력(emergent functionality)이 갑작스럽게 나타나면 인간은 이를 통제하거나 이해하지 못한 채 대응에 실패할 수 있고, 일곱째, 기만(deception)은 AI가 은밀한 방식으로 인간의 감시를 피해 행동함으로써 자신의 통제권을 빼앗는 결과를 낳을 수 있다. 마지막으로 권력 추구(power-seeking)는 AI가 스스로 설정한 목표를 이루기 위해 권력과 자원을 장악하려 함에 따라, 인간의 통제를 구조적으로 무력화시켜 결국 문명 전체를 위험에 빠뜨릴 수 있는 시나리오다. 이 모든 위험 범주는 각각 다른 경로를 통해, 그러나 공통적으로, 인간이 통제하지 못하는 방향으로 기술 문명이 탈선할 수 있다는 점에서 존재론적 위험으로 규정된다.[28]

## Ⅳ. 문명전환기 새로운 공화정치의 발명?

문제는 오늘날 지정학적 패권경쟁의 맥락에서 주류적인 AI 담론이 구성되고 있다는 점이다. 미중 간 제로섬적 대결구도가 대 전략 공론장의 주인서사(master narrative)로서 자리잡고 있으며, 이것의 하위 영역인 AI 담론도 이 틀 속에서 구성되고 있다. 즉, AI의 역사적 변곡점으로서의 성격이 부각되면서 이 기술을 선점하지 못할 경우 국가의 안보가 결정적으로 위기에 빠질 것이라는 위급성과 안보화의 서사가 지배화 되고 있다. 특히 구 냉전 핵 군비 경쟁과의 유비 속에 반드시 승리해야 할 "경주(race)"로 AI 기술영역이 재현되고 정책적 논의도 이런 상상계 속에 제출되면서, AI에 대한 국제 협력, 위험관리, 통제 등의 논의는 주변화되었다. 따라서 이 같은 반 안보화 테마에 대해 이야기하는 것은 나이브하거나 전략적이지 못한 것으로 취급되고 만다.[29] 이는 인류의 "생존"이라는 글로벌 거버넌스의

[28] Dan Hendrycks and Mantas Mazeika, "X-Risk Analysis for AI Research," *arXiv*, September 20, 2022, https://doi.org/10.48550/arXiv.2206.05862; AI 종말론에 깔린 보다 심층적인 공포는 막스 베버의 쇠우리(iron cage) 개념으로 포착가능하다. 즉, 근대 도구적 이성의 폭주 혹은 통제 불가능성의 절정으로서 AI를 상상해 볼 수 있다. 다시 말해 AI 종말론의 질문은 기실 근대성 자체가 우리의 주체성, 실질 합리성, 성찰성 등을 빈곤화시키고 있다는 문제의식과 맞닿아 있다. Jay A. Gupta, "Welcome to the Machine: AI, Existential Risk, and the Iron Cage of Modernity," *Telos* 203, 2023, pp. 163–69.

[29] Nike Retzmann, "'Winning the Technology Competition': Narratives, Power Comparisons and the US–China AI Race," in *Comparisons in Global Security Politics*, edited by Thomas Müller et al., Bristol: Bristol University Press, 2024, pp. 237–56.


ⓒ EAI 2025

차원에서 AI문제를 대처하기 어렵게 만드는 최대의 난관을 형성하고 있다. 인류 멸종 혹은 예속화 같은 절체절명의 위험이 예견되는 조건에서 어떻게 과거 글로벌 대재앙 리스크 대처에서 교훈을 얻어 대처방법을 강구할 것인지, 인류 전체의 집단적 학습능력에 우리의 미래가 달려 있는 상황이다.[30]

AI 기술의 세계적 확산은 기존 국제질서가 직면해 온 무정부 vs 위계의 이중 난제를 근본적으로 격화 시키며, 더 나아가 인류 문명 전체가 감당하기 어려운 새로운 형태의 초월적 권력 집중을 초래할 수 있다는 점에서 공화주의적 관점에서의 "억제(constraint)" 문제를 전례 없는 수준으로 부각시키고 있다. 특히 ASI는 전능·전지·편재를 동시에 획득할 수 있는 물질적 기반과 자기개량적 잠재력을 갖춘 기술적 주체로서, 공화주의가 역사적으로 경계해 온 폭정적 권력의 궁극 형태가 될 수 있는 바, 이 경우 기존의 가치 정렬(alignment) 전략이나 안전설계 등이 온전히 작동할 것이라는 보장을 제공하지 못한다. 그 결과 ASI 통제전략의 핵심은 단순한 기술적 위험관리의 차원을 넘어, 공화주의적 헌정 설계와 공공안전이라는 문제 틀 속에서 절차적·구조적 제약을 마련하는 방향으로 수렴할 수밖에 없으며, 심지어는 개발 모라토리엄(relinquishment)과 같은 극단적 선택지까지 정책적 상상력의 범위 안에 둘 필요성이 제기된다. 요컨대 AI 혁명의 도래는, 구 냉전 초 핵시대의 개막보다도 더 급진적으로 무정부적 국제체제의 지속 가능성 자체를 다시 묻게 하는 역사적 국면을 열고 있으며, 이것이 "인류의 마지막 발명"이 되지 않기 위한 전지구적 차원의 공화주의적 제도발명을 촉구하는 신호로 이해되어야 한다.■

[30] Seth D. Baum, Robert de Neufville, Anthony M. Barrett, and Gary Ackerman, "Lessons for Artificial Intelligence from Other Global Risks," in *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, Boca Raton, FL: Chapman and Hall/CRC, 2022.


ⓒ EAI 2025

■ 차태서_성균관대학교 정치외교학과.

■ 담당 및 편집: **임재현**_EAI 연구원

문의: 02-2277-0746 (ext. 209) jhim@eai.or.kr

인용할 때에는 반드시 출처를 밝혀주시기 바랍니다.
EAI는 어떠한 정파적 이해와도 무관한 독립 연구기관입니다.
EAI가 발행하는 보고서와 저널 및 단행본에 실린 주장과 의견은 EAI와는 무관하며 오로지 저자 개인의 견해임을 밝힙니다.

발행일 2026년 2월 03일
“” 979-11-7584-053-9 95340

재단법인 동아시아연구원
03028 서울특별시 종로구 사직로7길 1
Tel. 82 2 2277 1683 Fax 82 2 2277 1684

Email eai@eai.or.kr Website www.eai.or.kr



ⓒ EAI 2025