ISSUE BRIEF

이슈브리프 858호
(2026. 7. 3)

트럼프 인공지능(AI) 정책의 확장 : 혁신과 안보의 균형점

제858호

양지수 jisooyang@inss.re.kr

국가안보전략연구원
Institute for National Security Strategy

INSS
www.inss.re.kr

국문초록

2026년 6월 2일 트럼프 대통령은 <첨단 인공지능 혁신 및 안보 촉진(Promoting Advanced Artificial Intelligence Innovation and Security) 행정명령>에 서명하였다. 이 행정명령은 프론티어 AI(Frontier AI)의 사이버안보 위험이 부각되는 가운데, 미국의 인공지능(AI) 혁신을 저해하지 않으면서도 국가안보 위험을 관리하려는 균형을 추구하고 있다. 핵심 내용은 △프론티어 AI의 국가안보 위험을 관리하기 위한 30일 자율적 사전검토 체계 수립, △국가안보국(NSA) 주도의 기밀 벤치마킹 절차 확립, △AI 사이버안보 클리어링하우스 도입이다. 이는 기존 트럼프의 AI 규제 철폐 기조에서 이탈한 것으로, 미국 AI 정책이 혁신 중심 접근에서 나아가 국가안보를 고려하였다는 점에서 큰 의의를 둘 수 있다. 우리도 「AI 기본법」과 '인공지능 기본계획'을 통해 AI 사이버 안보 대응 기반을 마련하고 있으나, 프론티어 AI의 국가안보 위험을 평가·관리하는 제도는 아직 초기 단계이며, 국가 안보를 고려한 프론티어 AI 대응체계 역시 구체화되지 않았다. 이와 같은 상황에서 미국의 AI 정책 기조변화를 고려하여 우리는 프론티어 AI의 위험평가 체계 마련, 국가안보기관 중심의 AI 안보 거버넌스 정립, 소버린 AI 전략 재설계를 고려해야 한다.

키워드: 첨단 AI 행정명령, 프론티어 AI, 국가안보국(NSA), 미토스(Mythos) 사태, 소버린(Sovereign) AI

트럼프 AI 정책의 기조

2026년 6월 2일 미국 트럼프 대통령은 <첨단 인공지능 혁신 및 안보 촉진(Promoting Advanced Artificial Intelligence Innovation and Security) 행정명령>에 서명하였다. 이 행정명령은 트럼프 정부가 출범 이후 유지해 온 혁신 촉진과 규제 완화의 인공지능(AI) 정책 기조에서 나아가 프론티어 AI를 국가안보 차원의 관리 대상으로 편입하였다는 점에서 중요한 의미를 갖는다. 한편 2026년 6월 12일 미국 상무부는 국가안보를 이유로 앤트로픽(Anthropic)에 대해 미토스(Mythos) 5와 페이블(Fable) 5의 접근 제한을 요구하였고, 이에 따라 두 모델의 서비스가 일시 중단되었다. 이후 6월 30일 미국 상무부는 두 모델에 대한 수출 통제를 해제하였으며, 페이블 5는 7월 1일부터 전 세계 사용자에게 다시 제공되었다. 다만 미토스 5는 일부 기관을 대상으로 제한적으로 제공되고 있다.

트럼프 정부는 출범 초기부터 혁신 촉진과 규제 완화를 핵심 기조로 AI 정책을 추진하였다. 이러한 기조의 일환으로 2025년 1월 취임 직후, 트럼프 대통령은 바이든 정부의 <안전하고 보안이 보장되며 신뢰할 수 있는 인공지능의 개발과 사용에 관한 행정명령(Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)>을 철회하였다. 해당 행정명령은 일정 기준 이상의 '이중용도 기반모델(dual-use foundation model)'[1]에 대한 안전성 평가와 정부 보고 의무 등을 규정하고 있었으나, 트럼프 정부는 이를 AI 혁신과 산업 경쟁력을 제약하는 규제로 판단하였기 때문이다.

이후 트럼프 대통령은 <AI 행동계획(AI Action Plan)> 수립을 지시하는 한편, 'AI 교육'[2], '국가 사이버보안'[3], 'AI 기술 수출'[4] 등 AI와

1) AI 모델의 학습연산량을 기준으로 정의된 개념으로, 역량을 기준으로 삼는 정의 방식과 구별된다.
2) 미국 청소년을 위한 인공지능 교육 진흥 행정명령(2025년 4월 23일, ADVANCING ARTIFICIAL INTELLIGENCE EDUCATION FOR AMERICAN YOUTH Executive Order April 23, 2025), https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth/ (Accessed: June 24, 2026).

관련된 다양한 분야에서 행정명령을 연이어 발표하였다. 같은 해 7월에는 <하나의 크고 아름다운 법안(One Big Beautiful Bill Act)>을 통해 AI 공급망의 국내화와 미국 내 AI 산업 기반 강화를 추진하였으며, 12월에는 연방 차원의 'AI 정책체계'[5]를 구축하기 위한 행정명령을 발표하였다. 이처럼 트럼프 정부는 혁신과 시장 중심의 AI 정책 기조를 지속적으로 추진하였다.

그러나 이러한 혁신 중심 기조가 AI의 안보적 위험에 대한 고려를 완전히 배제한 것은 아니었다. 위 '국가 사이버보안' 행정명령이 이미 AI 시스템의 안전성 확보와 관련된 조항을 포함하고 있었기 때문이다. 다만 이러한 조항들은 AI 전반에 대한 사이버보안 확보에 초점을 두었던 반면, 2026년 6월 2일 발표된 위 '첨단 인공지능 행정명령'은 '프론티어 AI'라는 특정 범주를 국가안보 차원의 명시적 관리 대상으로 설정하고, 이에 대한 별도의 평가·관리체계를 제도화했다는 점에서 차별성을 갖는다.

미토스 사태: 프론티어 AI 관리체계 확장의 촉매

위 '첨단 인공지능 행정명령'에 따라 프론티어 AI 국가안보 관리체계가 도입된 배경에는 2026년 상반기에 발생한 미토스 사태가 있었다. 2026년 4월 7일, 앤트로픽은 프론티어 AI 모델인 클로드 미토스 프리뷰(Claude Mythos Preview)를 발표하면서, 해당 모델이 제로데이 취약점(아직 알려지지 않아 패치되지 않은 보안 결함)을 자율적으로 탐지하고

3) 국가 사이버보안 강화를 위한 특정 조치의 지속 행정명령(2025년 6월 6일, SUSTAINING SELECT EFFORTS TO STRENGTHEN THE NATION'S CYBERSECURITY Executive Order June 6, 2025), https://www.whitehouse.gov/presidential-actions/2025/06/sustaining-select-efforts-to-strengthen-the-nations-cybersecurity-and-amending-executive-order-13694-and-executive-order-14144/ (Accessed: June 24, 2026).

4) 미국 인공지능 기술 집합의 수출 촉진 행정명령(2025년 7월 23일, PROMOTING THE EXPORT OF THE AMERICAN AI TECHNOLOGY STACK Executive Order July 23, 2025), https://www.whitehouse.gov/presidential-actions/2025/07/promoting-the-export-of-the-american-ai-technology-stack/ (Accessed: June 24, 2026).

5) 인공지능에 관한 국가 정책 체계 확립에 관한 행정명령(2025년 12월 11일, ENSURING A NATIONAL POLICY FRAMEWORK FOR ARTIFICIAL INTELLIGENCE, Executive Order December 11, 2025), https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/ (Accessed: June 30, 2026).

이를 실제 공격에 활용할 수 있는 공격 코드를 생성하는 수준의 사이버 역량을 보유했다고 밝혔다. 이는 프론티어 AI가 인간 전문가 수준의 취약점 탐지와 공격 코드 생성을 자율적으로 수행할 수 있음을 확인한 것이었다. 이러한 역량은 국가의 사이버 방어 역량을 강화하는 동시에 적대국의 공격 능력 역시 크게 증대시킬 수 있다는 우려를 현실화시켰다.

앤트로픽은 당시 발표에서, 미토스 모델이 인간의 추가 개입 없이 프리비에스디(FreeBSD)에서 17년 동안 발견되지 않았던 원격 코드 실행 취약점을 발견하고, 이를 악용하여 루트(root) 권한을 획득하는 데 성공했다고 밝혔다.[6] 앤트로픽은 이러한 자체 평가를 바탕으로, 미토스 모델을 공개 출시하지 않고 제한된 컨소시엄에만 접근을 허용하였다. 같은 시기 영국 AI보안연구소(AISI) 역시 독립적인 평가를 통해, 통제된 환경에서 해당 모델이 방어가 허술한 시스템에 대해서는 인간의 개입 없이도 취약점을 자율적으로 악용할 수 있음을 확인하였다.[7]

앤트로픽이 미토스 모델을 제한된 컨소시엄에만 공개하였음에도 불구하고, 프론티어 AI의 안전성 확보 방식을 둘러싸고 트럼프 정부와 앤트로픽 간에는 입장 차이가 있었다. 양측 모두 프론티어 AI의 위험성에는 공감하였으나, 그 위험을 정부의 국가안보 체계를 통해 관리할 것인지, 기업의 자율적 안전조치를 중심으로 관리할 것인지를 두고 견해를 달리한 것이다.

미토스 사태는 프론티어 AI가 단순한 생산성 향상 도구를 넘어 국가안보에 중대한 영향을 미칠 수 있다는 실질적인 사례가 되었다. 또한 프론티어 AI의 안전성 확보와 활용 범위를 둘러싸고 정부와 개발기업 간 이해관계가 충돌할 수 있음을 드러냈다. 이러한 상황은 개별 기업의

6) Anthropic, “Assessing Claude Mythos Preview’s Cybersecurity Capabilities”, Anthropic (April 7, 2026), https://www.anthropic.com/research/mythos-preview (Accessed: June 22, 2026).
7) AI Security Institute, “Our evaluation of Claude Mythos Preview’s cyber capabilities”, AISI (April 13, 2026), https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities (Accessed: June 24, 2026).

자율적 안전조치만으로는 국가안보 위험을 충분히 관리하기 어렵다는 문제의식을 확산시켰으며, 결국 프론티어 AI를 국가안보 차원의 관리 대상으로 편입하고 정부가 직접 위험을 평가·관리하는 제도적 대응을 촉진하였다.

‘첨단 인공지능 행정명령’의 주요 내용과 특징

‘첨단 인공지능 행정명령’의 핵심은 프론티어 AI에 대한 ‘안전한 프론티어 모델 배포 체계(Secure Frontier Model Deployment)’를 수립하는 것이다. 위 행정명령의 초안은 프론티어 AI 모델에 대한 사전 검토 기간을 90일로 규정하였으나, 트럼프 대통령은 이러한 조치가 미국의 AI 경쟁력을 저해할 수 있다는 이유로 서명을 유보하였다. 이후 백악관 내부 조율을 거쳐 최종 행정명령에서는 사전 검토 기간이 90일에서 30일로 단축되었다. 이러한 기간 단축은 프론티어 AI에 대한 사전 검토를 유지하면서도 기업의 혁신에 대한 규제 부담을 최소화하려는 절충적 제도 설계의 결과로 볼 수 있다. 더불어 위 행정명령은 프론티어 AI의 국가안보 위험을 평가·관리하기 위한 새로운 거버넌스를 도입하였다.

우선, 위 행정명령은 프론티어 AI의 사이버 역량을 평가하기 위한 ‘기밀 벤치마킹 절차’를 도입하였다. 위 절차는 국가안보국(NSA), 국토안보부 사이버보안·인프라보안국(CISA), 재무부가 공동으로 개발하며, 상무부 산하 국립표준기술연구소(NIST), 국가사이버국장, 대통령과학기술보좌관(APST)이 기술적 자문과 협력을 제공하는 방식으로 설계되었다. 다만 위 행정명령은 어떤 모델을 ‘적용 대상 프론티어 모델’로 지정할지에 대한 최종 결정 권한을 국가안보국장에게 부여하였다. 이는 위 절차 설계가 다기관 협의를 통해 이루어지지만, ‘적용 대상 프론티어 모델’ 지정이라는 핵심 권한은 국가안보국에 집중되어 있음을 보여준다.

위 행정명령은 또한 재무부 장관으로 하여금 국가사이버국장, 국가안보국장, 사이버보안·인프라보안국장과 협의하여 ‘AI 사이버안보 클리어링하우스(AI Cybersecurity Clearinghouse)’를 구축하도록 하였다. 위 클리어링하우스는 AI 기업과 핵심 인프라 운영자 간 취약점 정보를

공유·조정하고, AI 기반 취약점 탐지와 대응을 지원하기 위한 정부-민간 협력 플랫폼이다. 위 행정명령은 '적용 대상 프론티어 모델' 지정 권한을 국가안보국에 집중시키는 한편, AI 기반 취약점 대응을 위한 정부-민간 협력체계를 제도화함으로써, 프론티어 AI의 사이버 역량을 국가안보 차원으로 편입한 것으로 할 수 있다.

평가

이번 행정명령의 가장 큰 특징은 프론티어 AI 관리 권한이 상무부가 아닌 국가안보국을 중심으로 설계되었다는 점이다. 이는 프론티어 AI의 위험 관리 권한을 국가안보기관에 부여함으로써, AI 관리체계도 기존 산업정책 중심에서 국가안보 거버넌스로 편입된 상징적인 변화로 평가할 수 있다. 주목할 점은 이러한 국가안보 거버넌스가 전통적인 허가·승인 방식이 아니라, 국가안보기관의 위험평가와 민간기업의 자율적 참여를 결합한 형태로 설계되었다는 점이다. 위 행정명령은 기밀 벤치마킹 절차와 AI 사이버안보 클리어링하우스를 통해 국가안보기관과 민간기업 간 협력을 제도화하면서도, 기업의 혁신과 자율성을 최대한 유지하려는 접근을 취하였다.

다만 이러한 설계가 실제 운영 과정에서도 기업의 자율성을 충분히 보장할 수 있을지는 별개의 문제이다. 앞에서 설명한 바와 같이 2026년 6월 12일, 미국 상무부는 국가안보를 이유로 앤트로픽에 미토스 5와 페이블 5 모델에 대한 수출통제 지시를 통보하였다. 해당 지시는 외국인에 대한 모델 접근 제한을 요구하는 내용이었다. 그러나 앤트로픽은 해외 이용자와 미국 내 외국인 이용자를 국적별로 식별·분리하여 접근을 제한하는 데 기술적·운영적 제약이 크다는 이유로 결국 해당 모델의 서비스 제공을 모두 중단시켜 버렸다.[8] 이 사례는 정부의 국가안보상 요구와 실제 기업의 조치가 반드시 일치하지 않으며, 기업의 기술적·

8) Anthropic, "Statement on the US government directive to suspend access to Fable 5 and Mythos 5", Anthropic (June 12, 2026), https://www.anthropic.com/news/fable-mythos-access (Accessed: June 25, 2026).

운영상의 판단에 따라 그 적용 범위가 확대될 수 있음을 보여준다. 다시 말해, 프론티어 AI에 대한 국가안보 거버넌스에서는 정부의 규제뿐 아니라 기업의 자율적 위험관리 전략 역시 규제 효과를 결정하는 중요한 변수로 작용할 수 있다. 따라서 향후 국가안보를 이유로 프론티어 AI에 대한 정부의 개입이 반복될 경우, 기업은 규제 불확실성과 법적 위험을 최소화하기 위해 정부 요구보다 더 광범위한 제한조치를 선제적으로 선택할 가능성이 있다.

한편, 프론티어 AI에 대한 접근 제한은 단기적으로 국가안보에 대한 위험을 줄일 수 있지만, 글로벌 AI 시장에서 경쟁국에게 새로운 기회를 줄 수도 있다. 실제로 앤트로픽이 미토스 5와 페이블 5의 서비스 제공을 중단한 직후 중국 Z.ai(지푸 AI)는 GLM-5.2를 출시하였다. GLM-5.2 모델은 코딩과 에이전트 작업 분야에서 미국 주요 프론티어 AI 모델들과 경쟁 가능한 수준의 성능을 보였다는 평가를 받았다.[9] 이는 특정 프론티어 AI 모델에 대한 접근 제한으로 발생한 시장의 공백을 경쟁국 기업이 빠르게 메우고, 기술 격차 역시 빠르게 축소될 수 있음을 시사한다. 따라서 프론티어 AI 정책은 국가안보상 필요성과 위험 수준에 따라 접근 범위를 조정하면서도 혁신을 유지할 수 있는 균형적 방식으로 설계될 필요가 있다.

시사점

우리의 「AI 기본법」 제32조 및 같은 법 시행령 제23조는 학습 연산량을 주된 기준으로 '고성능 인공지능'을 정의하여 적용 대상을 식별하고 있다. 이는 바이든 행정부의 2023년 행정명령(EO 14110)이 학습 연산량을 주된 식별 기준으로 채택했던 접근과 유사하다. 반면 트럼프의 '첨단 인공지능 행정명령'은 프론티어 AI 모델의 적용 대상을 연산량 규모

9) Laurie Chen, "After Anthropic shutdown, China's Z.ai closes frontier gap as it plans dual listing," Reuters (June 25, 2026), https://www.reuters.com/world/asia-pacific/after-anthropic-shutdown-chinas-zai-closes-frontier-gap-it-plans-dual-listing-2026-06-25/ (Accessed: June 29, 2026).

보다 사이버 역량(capability)을 중심으로 식별하고 있다. 이러한 변화에도 불구하고 우리의 프론티어 AI 식별 기준은 역량 중심 접근으로 전환되지 않은 채, 여전히 학습 연산량 중심으로 유지되고 있다. 이러한 상황에서 미국의 위 행정명령은 우리의 프론티어 AI 정책에 다음과 같은 시사점을 제공한다.

첫째, 국가 차원의 프론티어 AI 위험평가 체계를 마련해야 한다. 미국의 '첨단 인공지능 행정명령'은 기밀 벤치마킹 절차를 통해 프론티어 AI의 사이버 역량을 국가안보기관이 직접 평가하는 체계를 도입하였다. AI 안보의 핵심은 규제의 강도보다, 고위험 프론티어 AI 모델의 위험성과 사이버 역량을 독자적으로 식별·평가할 수 있는 국가 차원의 평가체계를 갖추는 데 있다. 우리 역시 프론티어 AI를 단순히 활용하는 것을 넘어, 기밀 벤치마킹에 준하는 국가 차원의 위험평가 체계를 구축할 필요가 있다.

둘째, 국가안보기관 중심의 AI 안보 거버넌스를 정립해야 한다. 프론티어 AI의 국가안보적 위험은 공개 정보만으로는 평가하기 어렵다. 해당 모델이 적대적으로 활용될 경우의 시나리오, 해외 유출 가능성, 탈옥(jailbreak)을 통한 악용 위험 등은 정보기관이나 보안기관이 보유한 위협정보와 결합될 때 비로소 실질적인 판단이 가능하다. 미국이 프론티어 AI의 위험 평가 권한을 상무부가 아닌 국가안보국 중심으로 설계한 것도, 이러한 이유에서이다. 우리 역시 국가정보원(NIS)을 중심으로 과학기술정보통신부, 국방부 등이 협력하여 프론티어 AI의 국가안보적 위험을 평가·분석하는 체계를 수립할 필요가 있다.

셋째, 소버린 AI 안보(Sovereign AI Security)의 전략을 재설계해야 한다. 이번 사례는 해외 프론티어 AI에 대한 접근이 국가안보의 핵심 변수로 작동할 수 있음을 보여주었다. 지금까지 우리의 AI 정책은 독자 기반모델과 컴퓨팅 인프라 확보 등 자립 기반 강화에 초점을 두어 왔다. 그러나 AI 분야에 있어 자립 역량 구축에는 수십년이 걸리는 반면, 외국 정부의 국가안보

조치나 외국 기업의 자체 대응에 따라 프론티어 AI 모델에 대한 접근 제한이나 서비스 중단은 단기간에도 이루어질 수 있다. 이러한 점에서 AI 주권은 국내 AI를 전적으로 자체 개발하는 능력뿐만 아니라, 공급자의 정책 변화나 지정학적 충격 속에서도 프론티어 AI에 대한 접근 연속성을 확보할 수 있는 능력까지 포함하는 개념으로 재정립될 필요가 있다. 따라서 우리는 정부·공공기관의 AI 도입과 활용 과정에서 프론티어 AI 모델의 접근 제한에 대비한 접근 연속성과 신속한 공급자 전환 역량을 확보할 필요가 있다. 또한 우리는 미국 프론티어 AI에 의존하는 동맹국들과의 연대를 바탕으로 프론티어 AI 접근 연속성을 새로운 국제 의제로 제안하고 관련 논의를 선도할 필요가 있다. 결론적으로 향후 우리의 AI 정책은 '무엇을 개발할 것인가'뿐만 아니라 '프론티어 AI에 대한 접근 연속성을 어떻게 확보할 것인가'까지도 함께 고려해야 한다.

//끝//