Jeju Peace Institute 제주평화연구원

JPI PeaceNet

인공지능 안전과 윤리: 디지털 미래의 신뢰 기반

윤 상 필

고려대학교 정보보호대학원 연구교수

[기획자 註] AI는 굉장히 빠른 속도로 발전을 거듭하고 있으며 이미 다양한 분야에서 우리의 삶에 많은 영향을 주고 있다. 하지만 AI 윤리에 관한 논의는 여러 행위자간 인식 차이로 인한 합의 부재, 그리고 규제가 기술 혁신을 저해할 수 있다는 우려 때문에 상대적으로 미흡하다. 이에 본고는 AI의 안전한 사용을 위해 윤리적 기준이 필요한 이유를 설명하고 이를 마련하기 위한 정책 방향을 제언한다 [기획: 박동준 연구실장(djpark@jpi.or.kr)].

I. 서론

인간처럼 생각하는 기계를 표방하는 인공지능(Artificial Intelligence, AI) 기술이 일상에 광범위하게 녹아들고 있다. 44.5%의 국민이 챗GPT나 제미나이와 같은 대규모 언어모델(Large Language Model, LLM) 기반 생성형 AI를 이미 경험하고 있다.[1] 그리고 이제 더 많은 사람들이 네이버나 유튜브가 아닌 AI 서비스로 검색하기 시작했다.[2] 2025년 한 해에만 월간 AI 기반 트래픽이 187% 증가하였으며 이는 전년 대비 사람의 트래픽 증가량과 비교하면 8배 높다.[3] 인터넷 공간에서 사람보다 AI의 활동이 많아지고 있다는 뜻이다. 일상뿐만 아니라 기술 시험의 최전선인 전장에서는 AI가 정보를 수집해 계획을 수립하고 공격 목표를 선정하고 있다.[4]

기술 수준도 하루가 다르게 변화한다. LLM 중심의 패러다임은 2026년 들어 이미 피지컬 AI(Physical AI)와 에이전틱 AI(Agentic AI)로 바뀌었다.[5] 2026년 4월 엔트로픽(Anthropic)은 AI 보안 모델인 미토스(Mythos)가 아직 너무 위험하다는 이유로 일반 공개를 보류했다. 이 때문에 국내에서는 미토스 사태로 보안 기관은 물론 모든 공공기관과 기업들이 비상 대응에 착

1) 과학기술정보통신부, 한국지능정보사회진흥원, 「2025 인터넷이용실태조사」, 2025, p. 191.

2) 고민서, "네이버로 검색하냐고요? 설마요…챗GPT 이용률 50% 넘겼다는데," 『매일경제』 2026년 1월 29일, https://www.mk.co.kr/news/it/11946388.

3) HUMAN Security, "2026 State of AI Traffic & Cyberthreat Benchmark Report," 2026.

4) Daniel Michales and Dov Lieber, "How AI Is Turbocharging the War in Iran," *The Wall Street Journal,* March 7, 2026, https://www.wsj.com/tech/ai/how-ai-is-turbocharging-the-war-in-iran-aca59002?mod=tech_lead_story.

5) LLM은 대규모 텍스트 데이터를 학습해 자연어를 이해하고 생성하는 모델로서 사용자의 질의에 답변을 하는 형태의 서비스가 대표적이다. 반면, 에이전틱 AI는 목표를 달성하기 위해 자율적으로 작업을 수행하는 시스템으로서 출장 일정을 기획하고 캘린더에 등록하며 교통편을 찾아주는 등 LLM을 포함하여 보다 확장된 기능을 수행한다. 한편, 피지컬 AI는 센서, 카메라, 로봇팔 등과 결합해 현실 세계를 인식하고 물리적 행동을 수행하는 AI를 말한다. 물류, 수술 등 다양한 로봇이나 드론, 자율주행차 등이 대표적인 사례다.

수했다. 그런데 불과 1달이 채 지나지 않아 영국의 AI안전연구소(AI Security Institute, AISI)는 사이버보안 역량 평가에서 오픈AI의 최신 모델 GPT-5.5가 미토스를 넘어서는 성능을 보였다고 밝혔다.[6)]

2016년 알파고 쇼크로 사람들이 가졌던 막연한 기대와 두려움과 달리 지금의 인공지능은 다분히 정치와 안보, 산업 및 경제와 같은 현실 문제로 다가오고 있다. 기업들은 앞장서 인공지능 윤리 원칙을 만들어 발표하고 있으며, 유엔 총회는 2024년 3월 지속 가능한 발전을 위한 안전하고 신뢰할 수 있는 인공지능 기회 확보에 관한 결의안을 만장일치로 채택했다.[7)] 세계 각국은 유럽연합과 우리나라를 필두로 인공지능을 포괄적으로 규율하는 일반법을 제정하기 시작했다.

이러한 상황 속에서 본고는 글로벌 인공지능 규범 형성 노력의 중요성을 강조하고 그 효과가 가시적인 결과로 나타나야 한다는 점을 주장하고자 한다. 10년 전 본격적으로 촉발된 인공지능 윤리에 관한 논의는 오늘날 실제 법규제로 구현되어 시장 진입의 요건으로 작동하고 있다. 어떤 미래를 그려낼 것인지 논의하고 구체적인 효과를 가져오는 제도와 정책을 설계할 때다.

2. 인공지능 안전과 윤리의 의의

인공지능 안전과 윤리의 본질을 논의하는 데 있어 인공지능은 어디까지나 기술이자 수단임을 인식하는 일이 중요하다. 흔히 생성형 AI 서비스의 환각 효과(hallucination)나 오류를 비판하지만 반대로 생각해 보면 사람의 거짓말이나 기억력의 한계로 인한 오류가 더 많다고 할 수 있다. 그런 점에서 인공지능에 대해 더욱 정확성, 견고성, 안전성을 요구하는 것은 믿고 쓸 수 있는 도구에 대한 요청이라고 이해할 수 있다.

인간 중심 원칙, 공정성, 보안성, 형평성과 같은 인공지능 윤리 기준도 있고 인공지능 기본법도 속속들이 제정되고 있는 상황이지만 인공지능 안전과 윤리는 생각만큼 보장되지 않는 것으로 보인다. 여전히 인공지능의 신뢰성 문제가 계속되고 있으며, 물리적 현실과의 상호작용 단계는 어려운 과제로 남아있다. 언어와 문화에 따른 격차가 커지고 있고 다양한 영향 평가나 사전 분석 제도들은 인공지능이 현실에 미치는 영향을 충분히 예측, 평가하지 못하고 있다.[8)] 국가적 생존 경쟁으로 인해 규제 논의도 실천으로 이어지지 못하고 있다. 2026년

6) UK AISI, "Our evaluation of OpenAI's GPT-5.5 cyber capabilities," April 30, 2026, https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities.

7) UN General Assembly, "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development," A/78/L/49, 2024.

8) Yoshua Benjio, "International AI Safety Report 2026," 2026, pp. 29-31.

5월 미국 트럼프 대통령은 AI 규제 행정명령 서명을 연기하면서 중국과의 인공지능 경쟁 우위를 점할 필요가 있다고 강조했다.[9] 유럽연합 또한 인공지능법(EU AI Act)에 따른 고위험 AI 규제 시행을 2027년 말로 늦추기로 했다.[10]

이런 상황에서 인공지능에 의한 안전 위협 및 권리 침해 사례들이 증가하고 있다. 보안 업체 탈레스(Thales)에 따르면 2025년 AI 기반 봇(Bot) 공격은 전년 대비 12.5배 급증했다.[11] AI 해커가 현실화되고 있는 것이다. 민주주의를 저해하는 문제도 계속해서 발생하고 있다. 2024년 미국은 대선을 앞두고 뉴햄프셔주 민주당 경선에서 조 바이든 당시 대통령의 음성으로 투표하지 말라는 딥보이스(DeepVoice) 기반 자동 녹음 전화가 유권자들에게 무차별 살포되는 문제를 겪었다. 이에 따라 연방통신위원회는 AI를 활용한 자동 녹음 전화를 전면 금지하기도 했다. 우리나라도 6월 3일 지방선거를 앞두고 딥페이크 등 인공지능 기반 허위사실 유포 행위에 엄정 대응하겠다는 방침을 발표했다.[12] 2024년 2월 홍콩 소재 글로벌 기업 직원은 회사의 최고재무책임자(CFO) 주재 영상회의에 참석해 CFO의 지시에 따라 약 2억 홍콩달러를 여러 계좌에 송금하였는데 조사 결과 CFO는 물론 다른 회의 참석자도 모두 딥페이크로 생성된 사실이 밝혀졌다. 딥페이크 성범죄 문제도 심각하다. 우리나라는 2024년 11월 정부 합동으로 딥페이크 성범죄 대응 방안을 발표하고 허위영상물 처벌 및 수사 대응 강화, 신속 삭제 및 플랫폼 책임성 제고 등의 조치를 마련했다.

국가의 입장에서는 경쟁력을 강화하면서도 안전과 윤리를 고려해야 하는 딜레마를 겪고 있다. 이런 상황에서 규제 논의는 지연되고 그 틈새에서 시민들의 권리가 침해되며 사회적 위험도 증가하고 있다. 개별적인 피해가 나타나는 분야의 소관 부처를 중심으로 대응이 이루어지고 있으나 종합적인 틀과 실효적인 규제가 요구되는 상황이다. 도구로서의 인공지능이 아닌 도구를 악용하는 사람, 도구를 안전하게 만들지 않는 사람을 제재할 수 있어야 한다. 도구에 의해 영향을 받는 사람들의 정당한 권리를 보장하고 피해를 입은 사람은 구제할 수 있어야 한다. 인공지능 안전과 윤리가 단순히 비전이나 구호에 그쳐서는 안 되는 이유다. 인공지능 안전과 윤리는 구체적인 규범과 정책으로 이어져 국가와 시장, 일상과 현실 세계에 직접적인 영향을 미칠 수 있어야 한다.

9) Karen Freifeld, Jody Godoy, Courtney Rozen, and Jacob Bogage, "Trump postpones AI executive order, cites need to compete with China," *Reuters*, May 22, 2026, https://www.reuters.com/business/retail-consumer/white-house-postpones-trumps-ai-signing-ceremony-says-axios-2026-05-21/.

10) European Council, "Artificial Intelligence: Council and Parliament agree to simplify and streamline rules," May 7, 2026, https://www.consilium.europa.eu/en/press/press-releases/2026/05/07/artificial-intelligence-council-and-parliament-agree-to-simplify-and-streamline-rules/.

11) Thales, "2026 Bad Bot Report: Bad Bots in the Agentic Age," 2026.

12) 대한민국 정책브리핑, "'가짜뉴스는 중대 범죄'…지방선거 앞두고 AI 악용 허위정보 차단 범정부 대응체계 본격 가동," 2026년 5월 19일, https://www.korea.kr/news/policyNewsView.do?newsId=148964424.

3. 인공지능 안전과 윤리 논의의 지향점

가. 신뢰할 수 있는 인공지능 생태계 구축

힘의 논리가 본질인 세계에서 국가 간 경쟁은 피할 수 없는 과제이다. 그러나 그렇다고 해서 안전과 윤리 문제를 미뤄두는 이분법적인 사고는 바람직하지 않다. 오히려 자국 내에서 발생하는 인공지능 안전과 윤리 문제로 인해 혁신 역량이 저해될 가능성이 더 높다. 따라서 인공지능 안전과 윤리를 고려해 신뢰할 수 있는 인공지능 생태계를 구축하는 데 초점을 둬야 한다. 여기에는 산업 성장과 민관협력에 기반한 국가적 거버넌스는 물론 반도체와 컴퓨팅 인프라, 데이터와 모델, AI 전문 인력과 첨단 기술 역량이 모두 포함된다.

이러한 점을 고려할 때 민간의 인공지능 전문가가 반드시 국가 전략과 법제, 정책을 설계, 집행하는 과정에 참여해야 한다. 설계 단계에서부터 윤리를 고려하는 것은 물론 사이버보안과 산업기밀 보호를 강화하는 것도 필수적이다. 데이터 수집 및 학습 과정에서는 정보보호 문제가 발생하기 때문에 동의나 식별과 같은 형식적 접근에서 벗어나 합리적인 데이터 처리가 가능하도록 하면서도 사생활 등 정보주체의 구체적 권리를 보호할 수 있어야 한다. 전문 인력을 양성하는 과정에도 인공지능 안전과 윤리를 학습할 수 있도록 해야 한다.

나. 인공지능 안전과 윤리의 실현

인공지능 안전과 윤리는 선언적 원칙이나 기업 홍보 차원의 구호에 머물러서는 안 된다. 이제는 실제 기술 개발과 운영, 정책 집행 과정에서 구현 가능한 제도로 정착되어야 한다. 특히 생성형 AI와 에이전틱 AI는 단순 정보 제공 수준을 넘어 자율적으로 판단하고 행동하는 단계로 발전하고 있다. 이에 따라 안전성과 윤리성은 서비스 출시 이후의 문제가 아니라 개발 단계부터 고려되어야 할 핵심 요소가 되었다. 데이터 수집과 학습, 모델 설계, 배포 및 운영에 이르는 전 과정에 걸쳐 안전성과 책임성을 내재화하는 책임 있는 설계(responsible by design) 원칙이 요구되는 이유다.

무엇보다 인공지능 안전 문제를 단순 기술적 보안의 차원으로만 바라봐서는 안 된다. 딥페이크 기반 금융사기와 허위정보 유포, AI 기반 자동화 해킹, 여론조작 등은 민주주의와 사회 신뢰를 직접적으로 위협하고 있다. 따라서 인공지능 안전과 윤리는 특정 영역이나 특정 분야의 문제가 아닌 국가안보와 사회통합, 민주주의 보호 차원의 문제로 접근될 필요가 있다. 이를 위해서는 위험 또는 영향에 기반한 정교한 규율 체계가 필요하다. 모든 인공지능 시스템을 동일하게 규제하기보다는 위험 수준과 영향 범위를 고려한 차등적 규율이 이루어져야 한다. 의료・금융・공공행정・국방과 같이 국민의 생명과 권리에 직접적인 영향을 미치는 분야에 대해서는 보다 엄격한 안전 기준과 책임을 요구할 수 있다. 반면 상대적으로 위험성이 낮은 분야에 대해서는 일정 수준의 자율성과 혁신 가능성을 보장할 필요가 있다.

다. 공통의 이익에 기반한 국제 협력

인공지능 안전과 윤리의 실현은 특정 국가만의 노력으로 달성될 수 없다. AI 기술과 데이터, 디지털 플랫폼은 국경을 초월해 작동하며 그 영향 또한 전 세계적으로 확산되기 때문이다. 한 국가에서 개발된 AI 모델이 다른 국가의 사회와 정치, 경제 시스템에 직접적인 영향을 미치는 상황에서 국제 협력은 선택이 아니라 필수적 과제가 되고 있다. 특히 사이버공격과 허위정보 유포, 디지털 감시, 자율무기체계와 같은 문제는 개별 국가 차원의 대응만으로 해결하기 어렵다.

따라서 국제사회는 공통의 이익에 기반한 협력 체계를 구축해야 한다. 여기에는 최소한의 안전 기준과 윤리 원칙에 대한 국제적 합의 형성은 물론 기술적·제도적 협력도 포함된다. 특히 인공지능 기술 패권 경쟁이 심화되는 상황일수록 국제사회는 기술 경쟁과 안전 확보 사이의 균형을 모색할 필요가 있다. 단기적인 산업 경쟁 논리만을 앞세울 경우 오히려 글로벌 디지털 신뢰 체계 자체가 약화될 수 있기 때문이다. 예를 들어, 인류 차원의 이익을 위해 활용될 수 있는 인류 공용 AI 기반을 형성하는 것도 고려해 볼 수 있다. 단편적인 이익 경쟁에서 나아가 중장기적인 공적 가치를 수호하고 확장하는 전략이 요구된다.

한편, 우리나라 역시 국제 협력 체계 속에서 보다 적극적인 역할을 수행할 필요가 있다. 세계 최고 수준의 디지털 인프라와 인공지능 활용 역량, 정보보호 경험을 기반으로 국제 규범 형성과 정책 논의에 실질적으로 기여할 수 있어야 한다. 특히 인공지능 안전과 개인정보 보호, 사이버보안, 디지털 민주주의와 같은 분야에서 한국형 정책 경험과 제도 모델을 국제사회와 공유할 필요가 있다. 궁극적으로 인공지능 안전과 윤리는 단순한 기술 통제의 문제가 아니라 디지털 시대 공공의 신뢰와 인간 중심 가치를 어떻게 유지할 것인가에 관한 국제사회의 공동 과제라고 할 수 있다.

4. 결론

인공지능 기술이 빠르게 발전하고 사람들의 삶에 빠르게 적용되면서 그 어느 때보다도 공적 가치를 수호하는 일이 중요해졌다. 인공지능 안전과 윤리는 혁신을 저해하지 않는다. 오히려 인류가 마주하고 있는 새로운 기술 혁명의 초입에서 혁신의 지속 가능성을 보장하기 위한 조건이다. 무엇보다 중요한 것은 인공지능 안전과 윤리를 추상적인 가치 선언에 머물게 하지 않는 일이다. 현실에서는 이미 딥페이크 범죄, 자동화된 사이버공격, AI 기반 허위정보 유포, 개인정보 침해와 같은 구체적인 피해가 나타나고 있다. 따라서 정책과 제도 역시 실제 위험에 대응할 수 있는 수준으로 발전해야 한다. 단순 권고나 시장에만 맡겨두는 자율규제만으로는 충분하지 않다. 위험 기반 규율 체계와 책임성 확보, 피해 구제 체계가 함께 마련되어야 한다. 시장에서 규칙을 형성하고 위반 사실이 확인되는 때에는 엄격한 공적 제재를 허용하는 체계도 고민할 수 있다.

아울러 인공지능 문제는 특정 국가가 독자적으로 해결할 수 있는 성격의 문제가 아니다. 따라서 국제사회는 공통의 원칙과 협력 체계를 구축할 필요가 있다. 특히 민주주의 보호, 사이버안보, 데이터 주권, 디지털 인권과 같은 문제는 글로벌 차원의 논의와 협력이 필수적이다. 기술 패권 경쟁 속에서도 최소한의 안전 기준과 윤리 원칙에 대한 국제적 합의는 지속적으로 추진되어야 한다.

결국 인공지능 시대의 핵심 과제는 어떤 방향의 AI를 만들 것인지, 어떤 디지털 미래를 꿈꾸는지에 있다. 인간의 존엄성과 자유, 사회적 신뢰와 민주주의라는 공적 가치를 지켜내지 못한다면 기술 발전은 오히려 위험 요소가 될 수 있다. 그렇다면 기술을 발전시키지 않는 것이 바람직하다. 디지털 미래의 신뢰 기반은 그 기술을 책임 있게 설계하고 활용하려는 사회적 의지와 제도적 노력 위에서 형성된다. 인공지능 안전과 윤리는 결국 기술의 문제가 아니라 인류가 어떤 미래를 만들어갈 것인지에 대한 공동의 선택이라고 할 수 있다.

저자 소개: 윤상필

윤상필 박사는 법학과 정보보호학을 전공하고 고려대학교 정보보호대학원 연구교수로 재직 중이다. 주요 연구 분야는 사이버안보, 사이버국방, 사이버수사이며 그밖에 인공지능 안전과 윤리, 개인정보보호, 디지털 전환에 관한 연구들도 수행하고 있다. 현재 경찰청 사이버성폭력수사 자문위원, 캐나다 OBVIA 협력위원(membre collaborateur), 한국사이버안보학회 편집이사, 한국인터넷윤리학회 총무이사, 개인정보보호법학회 홍보이사, 한국공법학회 기획간사 등으로 활동하고 있다. 저서로는 『사이버보안취약점의 법적 규제』 (박영사, 2022), 역서로는 『비트전: 사이버전의 혁신』 (박영사, 2024)가 있다.

2026년 5월