KISTI
ISSUE BRIEF

ISSN 2635-5728

과학기술인프라, **데이터로 세상을 바꾸다**

https://doi.org/10.22810/2024KIB068

2024. 05. 28. 제68호

인공지능 윤리(AI Ethics): 인간과 인공지능의 조화로운 공존 방안

유화선 · 윤병성 · 최희석

인공지능(AI) 전환 시대로 진입하면서 AI는 일상생활과 산업 전반에 걸쳐 혁신적인 변화를 가져오고 있다. 하지만 AI의 급격한 진보는 윤리적, 기술적 문제를 동반하며 전문가들은 AI의 위험성에 대해 경고하고 있다. 이에 따라 세계 각국 정부와 기업은 AI에 의한 위험과 부작용을 방지하기 위해 AI 윤리 원칙과 법·제도 등을 발표하고 있으며, 국내 정부도 AI 윤리·신뢰성 확보를 위한 방안을 마련하기 위해 노력 중에 있다. 인간과 AI가 조화롭게 공존하기 위해서는 AI의 발전과 인간의 책임 사이에서 균형이 요구되며, AI는 인간을 완전히 대체하기보다는 인간의 능력을 향상시키고 인간과 협업할 수 있는 혁신의 조력자로 인식해야 할 것이다. 또한 미래 AI와 인간이 바람직한 관계로 지속되기 위해서는 미래 AI 사회 시나리오 연구를 통해 선제적인 대응 전략 수립이 필요하며, AI 기술 수준과 안전성을 판단할 수 있는 평가 기준 개발, AI 기술과 윤리 통합 교육 및 산업별 맞춤형 AI 윤리 정책이 마련되어야 할 것이다.

CONTENTS

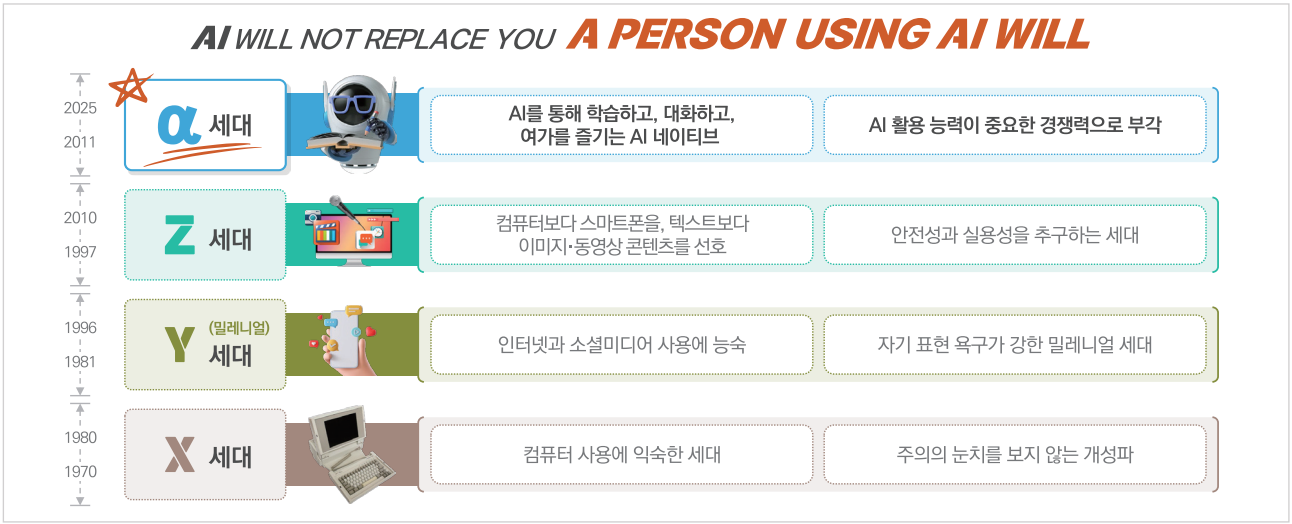
1. 인공지능 시대로의 전환

지능을 가진 기계, 인공지능

- (인공지능의 등장) 존 매카시(John McCarthy)가 1956년 다트머스 회의에서 과학자들에게 생각하는 능력을 갖춘 기계 연구를 제안하면서 '인공지능(Artificial Intelligence, AI)' 용어가 처음 사용됨
  - 인공지능의 정의는 오랫동안 다양하게 나타났으나, 존 매카시가 2007년 발표한 논문[1]에서 지능형 기계를 만드는 과학과 공학으로 정의하며, 인공지능은 기계가 지식을 가지고 스스로 학습하고 행동할 수 있어야 한다고 제시함
  - 이후, AI 겨울[2]이라는 불황기를 겪기도 했으나 컴퓨터와 데이터 기술[3]이 발전하면서 세계 각국은 AI 기술의 중요성을 인식하고 AI 연구 개발과 인재 양성에 막대한 자원을 투입하고 있음
  - AI 열풍으로 글로벌 AI 시장 규모도 2023년 1,502억 달러(약 200조 원)에서 연평균 성장률 36.8%로 급격하게 성장해 2030년 1조 3,452억 달러(약 1,800조 원) 규모에 이를 것으로 전망됨(Markets and Markets, 2024)
- (인공지능의 일상화) 챗GPT 등장과 함께 생성형 AI 열풍으로 인해 각종 분야에서 AI 역할이 급속도로 확대되고 있으며, 정부는 모든 국민의 AI 일상화(AI Everywhere)를 위한 정책 기반을 마련 중임
  - AI 기술은 금융, 미디어, 교육, 의료, 법률, 행정 등 다양한 산업 및 업종 등에 도입되어 인간의 일상생활 속에 깊이 퍼져 있으며, 다양한 분야에서 혁신적인 변화를 가져오고 있음

    ※ 2024년 CES 주제는 'All Together, All On'으로 모든 기업과 산업이 다 함께 인류의 문제를 혁신 기술로 해결하자는 것을 의미하며, 핵심은 전 산업을 관통하는 AI 기술의 융합임을 발표함(삼정KPMG, 2023a)

〈그림 1〉 세대별 AI 활용 역량



출처) NIA(2023)

1) What is artificial intelligence?
2) AI 겨울(AI winter)은 AI 연구에 대한 자금, 관심 등이 감소하는 일종의 불황기로, 1974~1980년과 1987~1993년 두 번의 겨울이 있었음
3) 데이터 기반의 가치 창출을 위해 데이터 수집부터 분석·활용에 이르는 과정에서 적용되는 기술(유화선 외, 2023)

- 정부는 전 국민 AI 일상화를 목표로 국민 일상, 산업 현장, 공공행정 분야에 AI를 중점적으로 도입할 계획을 수립·시행 중임

  ※ 정부는 최근 'AI 일상화 및 산업 고도화 계획('23.1.)', '초거대 AI 경쟁력 강화 방안('23.4.)', '전국민 AI 일상화 실행계획('23.9.)' 등을 발표함

- AI 서비스의 보편화 및 일상화로 인해 미래 세대는 AI를 얼마나 잘 다루는지 즉, AI 활용 역량이 앞으로는 중요한 경쟁력이 될 것임

생성형 인공지능 부상에 따른 변화

- (새로운 서비스 등장) 생성형 AI 기술의 진화로, 인간이 사는 세상은 디지털 전환을 넘어 인공지능 전환(AI Transformation, AX) 시대로 진입하고 있음

  - 글로벌 빅테크(Big Tech) 플랫폼 기업뿐만 아니라 국내 기업들은 자체적인 파운데이션 모델 구축을 통해 생성형 AI 서비스를 개발 중이며, 기업들은 생성형 AI 활용을 통해 '서비스 확대, 신뢰도 강화, 고객층 다변화'라는 트렌드로 진화하고 있음

〈그림 2〉 생성형 AI를 활용한 서비스 개발 트렌드

출처) 삼정KPMG(2023b)

- 또한 텍스트, 음성, 이미지 위주의 생성형 AI 기능에서 최근에는 동영상(비디오)을 위한 기능으로 서비스가 확장되고 있는 추세임

  ※ 동영상 생성형 AI로는 오픈AI의 Sora, 구글의 Lumiere, 런웨이의 Gen-2, 메타의 Emu Video 등이 있으며, 콘텐츠 제작에 소요되는 비용·시간을 절약하는 반면 진짜와 가짜의 경계를 흐려지게 한다는 우려도 있음

- (인공지능에 대한 경고) AI 기술 발전은 인간의 삶을 편리하게 하고 인간이 풀지 못하는 문제를 해결하는 잠재력을 가지고 있으나, 글로벌 업계 리더와 전문가들은 AI의 위험성을 경고하고 있음

  - 샘 알트만(오픈AI CEO), 데미스 하사비스(구글 딥마인드 CEO) 등 업계 리더들은 AI가 전염병이나 핵전쟁만큼 인류에게 실존적 위협을 가할 수 있다며, AI로 인한 멸종을 막아야 한다고 경고함

- 전 세계의 저명인사, 리더, 학자 등 전문가들은 발표 및 공개서한[4] 등을 통해 AI 기술 발전으로 인한 AI 통제 불능 상태, 인류의 멸종 가능성 등 위험성을 경고하고 있으며, AI 개발 중단과 이를 규제할 국제기구의 필요성도 언급하고 있음

〈표 1〉 인공지능에 대한 전문가들의 경고

- "2029년에는 AI가 아마도 모든 인간을 합친 것보다 똑똑해질 것이다" – **일론 머스크**
- "인류가 AI에 대처하는 방법을 익히지 못한다면 AI 기술은 인류 문명사에서 최악의 사건이 될 것이다" – **스티븐 호킹**
- "AI가 실존적 위험을 초래할 수 있다. 실존적 위험이란 아주 많은 사람이 해를 입거나 죽임을 당하는 것으로 정의될 수 있다" – **에릭 슈미트**
- "AI 개발을 멈추기에는 이미 늦었다. 곧 인간의 능력을 능가할 수 있다", "10년 내에 자율적으로 인간을 죽이는 로봇 병기가 등장할 것으로 본다" – **제프리 힌튼**
- "AI의 발전 속도를 알아차렸다면 효용성보단 안전성을 우선시 했을 것이다" – **요슈아 벤지오**
- "국제원자력기구(IAEA)와 같이 AI 문제를 감시·규제할 국제기구가 필요하다" – **샘 알트만**

출처) 언론보도 종합

- **(윤리적 이슈) 국가 간 과도한 AI 개발 경쟁에 대한 우려가 커지면서 올바른 AI 사용을 위해서는 윤리적·기술적인 안전장치가 필요하다는 의견이 제기되고 있음**
  - AI를 활용함으로써 개인정보 침해, 차별 및 편향성 문제 등 윤리 문제가 발생하고 있으며, AI의 위험 요소로 무기화, 통제력 상실 등이 제시되고 있음
  - 사고 과정을 모르는 AI, 비윤리적 목적을 가진 AI 활용으로 발생하는 위험성은 인류 파괴를 야기하고 있으며, 인간이 어느 정도 통제할 수 있는 기술적·제도적 방안 마련이 필요함

    ※ 우크라이나군의 AI 드론이 자체 판단으로 러시아군을 공격했으며, 올해 3월 이스라엘이 하마스 전쟁에서 AI 자율살상무기를 활용하는 등 기술이 인간 통제를 벗어나고 있음

    ※ 최근 오스트리아에서 열린 콘퍼런스('24.4.29.)에서 100여 개국이 참석해 킬러로봇 출현에 대한 우려로, AI 자율살상무기에 대한 규제 방안을 논의함

〈표 2〉 AI가 가진 위험성

• AI의 무기화 (대량 사이버 공격, 생화학 무기 제조 등) • AI 시스템에 대한 공격	• 예측 불가능한 기술 발전 • AI 시스템의 오작동 • AI 시스템에 대한 통제력 상실

출처) Gladstone AI(2024)

4) 1,100명 이상의 IT 업계 리더 및 학자들은 'Pause Giant AI Experiments: An Open Letter(2023.3.22.)' 공개서한을 통해 AI 개발을 일시 중단하여 AI의 발전 속도를 늦출 것을 제안함

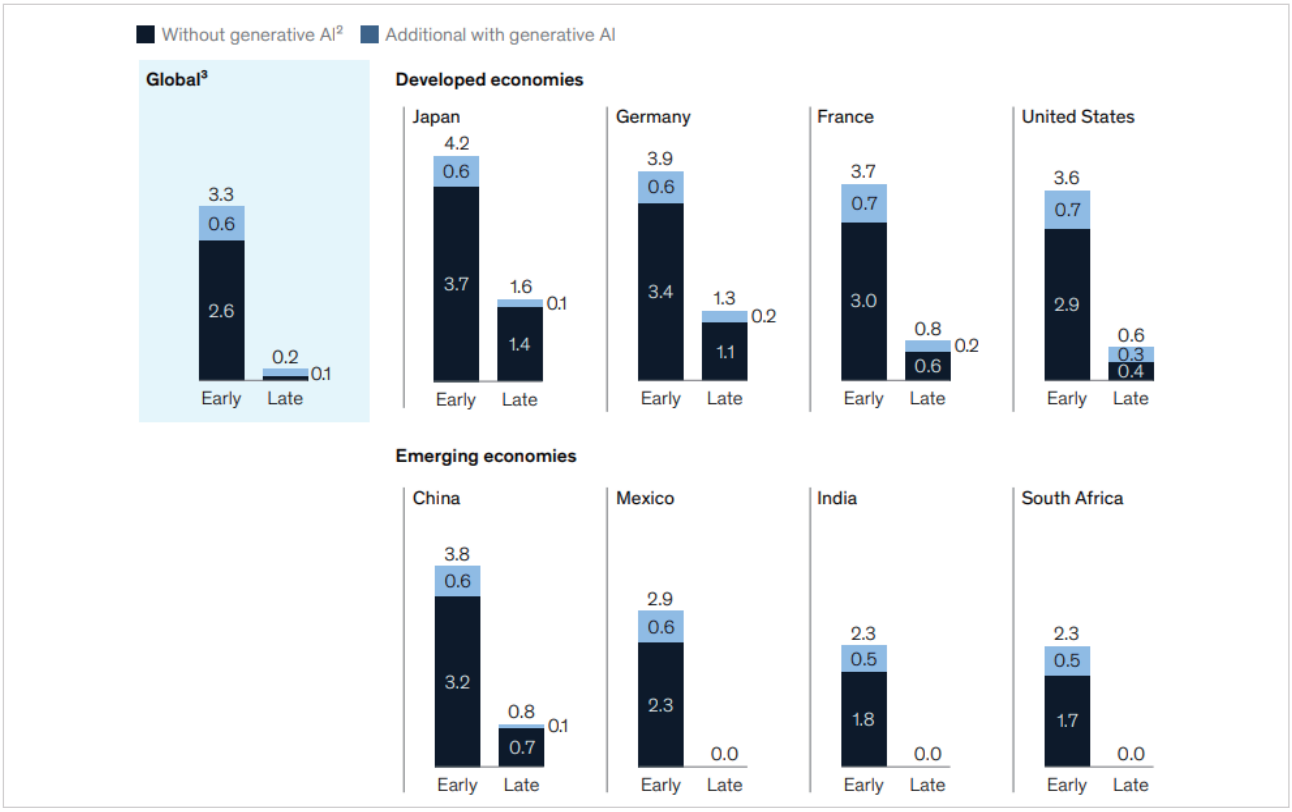
2. 인공지능 기술과 윤리

인공지능 기술의 명과 암

- (기술진보의 빛) 게임 체인저(Game Changer)[5] 기술로 부상한 AI는 경제·사회구조 전반에 영향을 미치고 있음
  - 생성형 AI는 향후 10년 동안 글로벌 GDP를 7% 증가시켜 경제 호황을 기대하게 하며, 3억 개에 달하는 일자리에도 영향을 미칠 것으로 전망함(Goldman Sachs, 2023)
  - 국내는 제조·서비스업 등 경제 전반에 생성형 AI를 성공적으로 적용할 경우, 연간 310조 원 이상('26년 기준)의 경제적 파급효과를 가져올 것으로 예상하고 있음[6]
  - 생성형 AI 도입은 산업 전반에 현저한 영향을 미치고, 근로자의 역량 강화 및 노동 생산성 향상에도 크게 기여할 것으로 기대됨

    ※ 생성형 AI를 통해 현재 직장인들이 수행하는 업무의 60~70%를 자동화할 수 있을 것으로 추정되며, 2040년까지 매년 0.1~0.6%의 노동 생산성을 증가시키고, 다른 기술과 결합하면 0.2~3.3%p의 생산성을 향상시킬 것으로 예측됨(McKinsey&Company, 2023)

〈그림 3〉 생성형 AI의 생산성 영향(2022~2040년)



출처) McKinsey&Company(2023)

5) 세계경제포럼(WEF)은 2023년 다보스 포럼에서 생성형 AI를 '게임 체인저'로 정의하고 사회와 산업 차원의 대비가 필요하다는 의견을 제시함

6) 과학기술정보통신부가 글로벌 컨설팅 회사인 베인앤컴퍼니(Bain&Company)와 공동으로 연구·분석한 결과임

- (기술진보의 그림자) AI 기술의 급진적 발전과 함께 AI로 인한 피해 사례가 증가하고 있으며, 이로 인해 비윤리적인 AI 활용이 문제시되고 있음
  - 딥페이크로 인한 가짜뉴스 생성, AI 편향으로 차별 발생, 프라이버시 침해 등과 같이 무분별한 AI 사용은 인간에게 위험을 유발함

    ※ 세계경제포럼(WEF)의 2024년 글로벌 리스크 보고서(Global Risk Report)에서는 AI가 생성한 가짜뉴스와 허위 정보, AI 기술의 부작용 등을 향후 10년 내 인류가 직면한 가장 큰 위협 분야로 선정함

〈표 3〉 AI 부작용 사례

• (가짜뉴스) 홍콩 딥페이크로 인한 금융사고, 미국 대선 과정의 가짜뉴스 논란 • (차별·편향) 아마존 AI 채용 시스템의 여성 차별, 직업 이미지 생성에서의 인종 편향 • (프라이버시 침해) 중국 신장 위구르 지역의 AI 감시 시스템

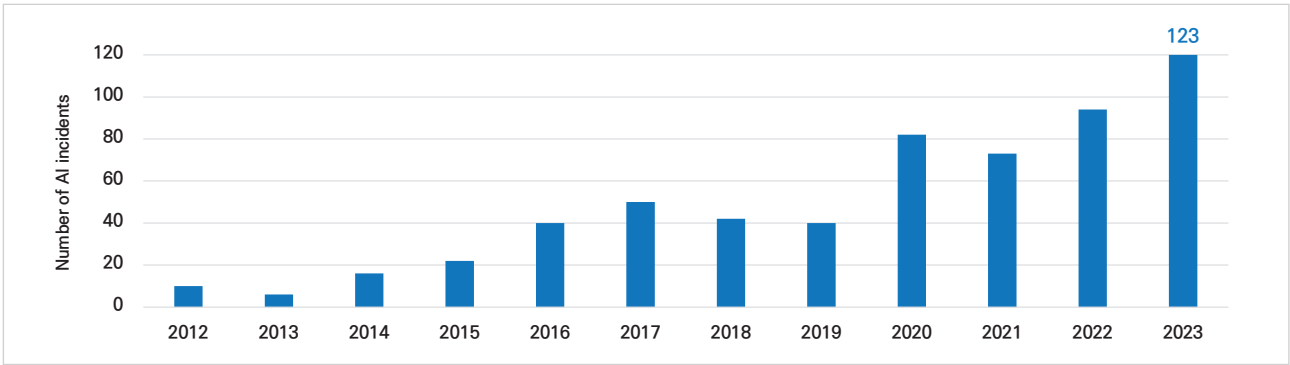
출처) 언론보도 종합

  - 또한 빅테크들이 자사 AI 모델의 기능 향상을 위해 저작권 규정을 무시하고 AI 학습용 데이터를 무단으로 사용한 사례도 발생하고 있음

    ※ 오픈AI가 GPT-4 모델 개발 중에 기존 수집한 AI 학습용 데이터가 고갈될 위기에 처하자 유튜브, 팟캐스트 등의 콘텐츠를 무단 사용했다는 의심을 받고 있음
  - AI 기술의 영향력이 증가함과 동시에 AI 관련 사고 건수가 지속적으로 증가하고 있으며, AI로 인한 부작용을 방지하기 위한 대응 수단 마련이 요구되는 상황임

    ※ AI 관련 사고 건수는 2023년 123건으로, 전년 대비 32.3% 증가하였으며 2013년 이후 20배 증가함(HAI, 2024)

〈그림 4〉 AI 사고 현황(2012~2023년)



출처) HAI(2024)

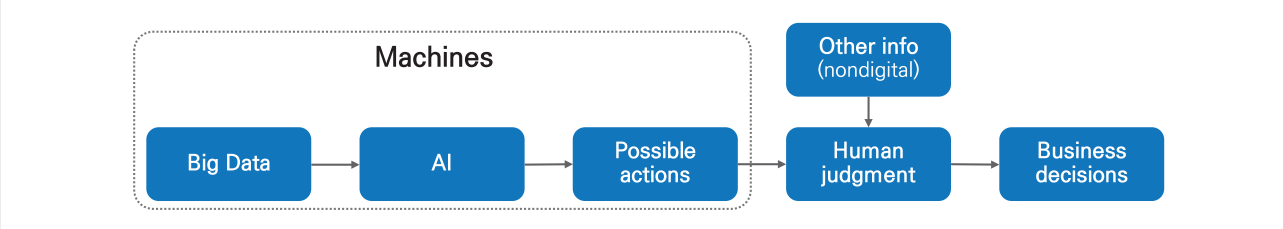
인공지능 윤리 전쟁

- (인간 vs 인공지능) AI 시스템을 의사결정에 활용하게 되면서 윤리적 딜레마 상황이 발생하고 있지만, 합리적인 의사결정을 위해서는 인간과 AI의 공존이 필요함

- AI가 대량의 데이터를 처리·분석함으로써 인간의 의사결정을 지원하나, 인간의 가치를 배제한 결정을 내리는 문제가 발생할 수 있음
- 기업 내 의사결정 과정이 AI 주도형(AI-driven) 의사결정 방식으로 진화된 상황에서 경험·직관에 의존하는 인간 또는 첨단 기술인 AI를 의사결정 과정에 단독 사용하기보다는 상호 보완적으로 활용해야 함

  ※ 에릭 콜슨(Eric Colson)은 '인간과 AI가 상호 보완하는 의사결정 모델'을 가장 이상적인 모델로 제시함

〈그림 5〉 인간과 AI가 상호 보완하는 의사결정 모델



출처) Eric Colson(2019)

- (개발 vs 규제) AI의 활용이 산업을 진흥시켜 인간에게 경제적 편익을 주는 면도 있지만, 예기치 못한 위험 발생으로 윤리 및 규제에 대한 논의도 가속화되고 있음
  - AI 기술 개발 촉진과 윤리 간 의견 대립이 실제 기업 경영상에서 물리적인 충돌로 나타난 사례가 있음

    ※ 오픈AI CEO인 샘 알트만(Sam Altman)은 이사회 내부와 가치관의 차이(수익성 vs AI 안전성)로 해임('23.11.)되었다가 닷새 만에 복귀함
  - 세계 최초로 포괄적인 AI 규제법을 통과('24.3.)시킨 EU는 기술 규제에 대한 보완책으로 생성형 AI 진흥안을 마련함
  - 즉, AI 기술로 인해 발생되는 변화와 위험은 기술 발전과 인간의 책임 사이의 균형이 필요하다는 것을 의미함

국제사회 대응 모습

- (AI 윤리 확산) 세계 각국과 국제기구는 AI에 의한 부작용을 방지하기 위해 AI 윤리와 관련된 대책 마련에 관심이 고조되고 있음
  - AI 관련 각종 사고들이 발생함에 따라 AI 개발에 대한 공동 약속인 아실로마 인공지능 원칙(Asilomar AI Principles)[7]이 발표('17)되었고, 윤리 이슈가 언급되면서 이때부터 AI 윤리 분야의 중요성에 대한 논의가 본격화되기 시작함

    ※ 아실로마 인공지능 원칙은 연구 이슈(5개), 윤리적 가치(13개), 장기적 이슈(5개) 등 세 가지 범주로 구성되어 있으며, AI 개발 과정에서 발생할 수 있는 실패와 자유의 침해에 대해 책임성 있는 행동을 강조함

7) 미국의 비영리 단체인 Future of Life Institute에서 개최한 콘퍼런스(Beneficial AI 2017)에서 채택된 AI 원칙으로, 스티븐 호킹, 일론 머스크, 데미스 하사비스 등 2,000여 명의 과학·기술계 인사들이 지지 서명을 남김

〈표 4〉 아실로마 인공지능 원칙(윤리적 가치)

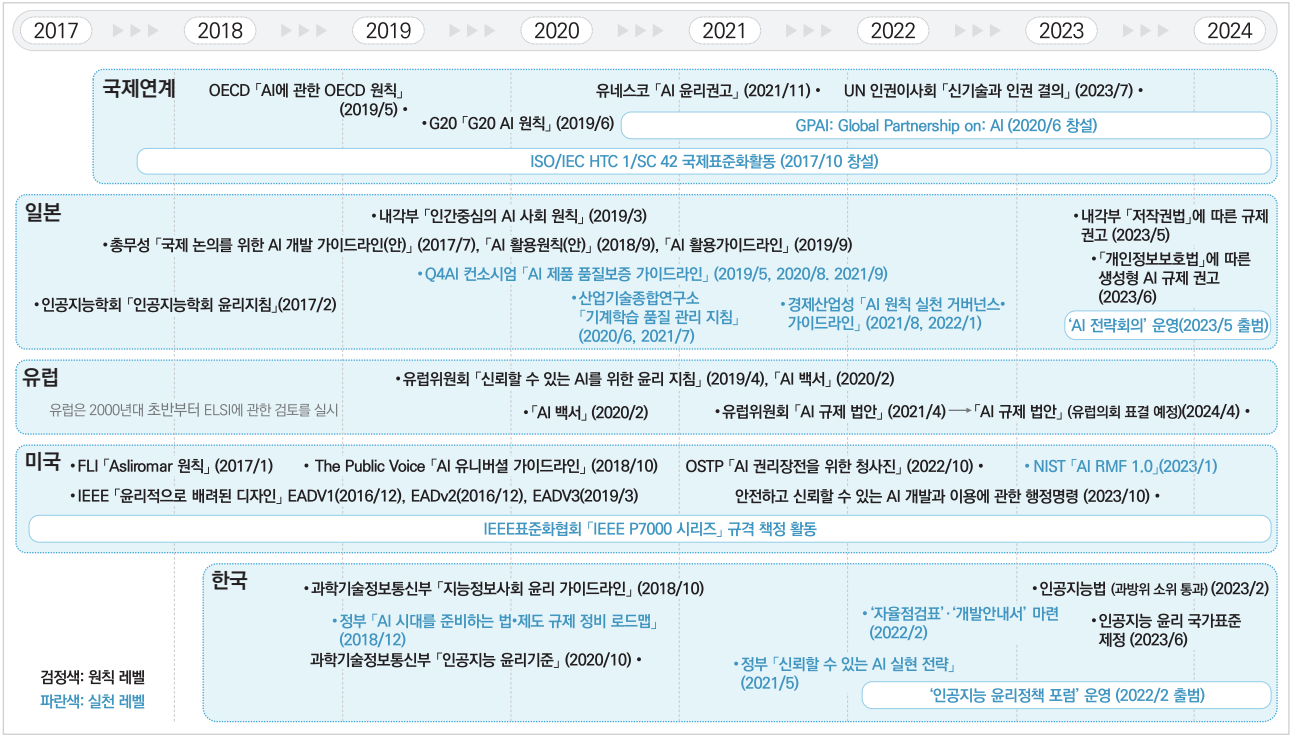
• 안전 • 실패의 투명성 • 사업적 투명성 • 책임성 • 가치 일치	• 인간의 가치 • 개인정보보호 • 자유와 프라이버시 • 공동 이익 • 공동 번영	• 인간 통제 • 사회 파괴 방지 • 인공지능 무기 경쟁 방지

출처) Future of Life Institute(2017)

– AI 윤리가 전 세계적으로 주목받으면서 AI 선진국 및 국제기구를 중심으로, 바람직한 인공지능 개발 및 활용을 위한 AI 윤리에 관한 원칙, 권고안 등이 발표되고 있음

※ EU는 신뢰할 수 있는 AI 윤리 가이드라인('19), OECD는 AI 원칙('19), 일본은 인간 중심의 AI 사회 원칙('19)을 발표했으며, 우리나라도 AI 윤리 기준('20) 등을 마련함

〈그림 6〉 주요국의 AI 윤리 원칙



출처) CRDS(2023) 일부 내용 수정

– 또한 ChatGPT 출시 이후 세계 주요국은 윤리적 차원에서 논의되었던 FATE 원칙[8]을 EU를 중심으로 법제화하기 시작하는 등 기준을 강화함과 동시에 산업 진흥을 저해하지 않는 방향에서 제도적 틀을 마련 중임(Bahar Memarian et al., 2023)

8) FATE는 Fairness(공정성), Accountability(책임성), Transparency(투명성), Ethics(윤리 의식)을 의미함

- **(글로벌 기업의 노력) 글로벌 IT 기업들은 AI 기술이 인간에게 이로운 방향으로 발전하도록 개발 과정에서 윤리를 강조하고 사용 규범을 마련하고 있음**
  - AI 폐해를 막기 위해 기업들은 자체적으로 AI 윤리 원칙을 마련하고 있으며, 인간중심성, 책임성, 개인정보보호, 안전성, 투명성 등 측면을 강조하고 있음

〈표 5〉 기업별 AI 윤리 원칙

	Humanity Human-centred	Responsibility Accountability	Privacy Security	Safety Reliability	Transparency Explainability
OECD	○	○	○	○	○
UNESCO	○		○	○	○
European Commission	○	○	○	○	○
미국 국가정보장실	○		○		○
호주 산업과학자원부	○	○	○	○	○
대한민국 과학기술정보통신부	○	○	○	○	○
사우디 데이터인공지능청	○	○	○	○	○
Google	○	○		○	
Microsoft		○	○	○	○
IBM	○		○		○
Adobe		○			○
OpenAI	○			○	
LG AI Research	○	○		○	○
Kakao	○		○		○
NAVER	○		○	○	○
SAMSUNG	○		○		○
SK Telecom	○		○	○	○
Upstage	○	○	○	○	○

출처) 박찬준 외(2023)

  - 또한 오픈AI, Google, 네이버클라우드 등 국내외 기업들은 AI 시스템의 취약점을 발견해 정보의 신뢰성을 검증하고 AI가 윤리적으로 활용될 수 있도록 자체 AI 레드팀[9]을 운영하고 있음

- **(AI 윤리 진단) AI가 발전할수록 편향성 문제가 증가함에 따라 정부와 기업에서는 규정과 기술로 문제를 극복하기 위해 노력 중임**
  - UNESCO는 정부가 AI를 개발하고 윤리적으로 배치할 수 있도록 지원하는 진단 도구인 'AI 준비도 평가 방법론(Readiness Assessment Methodology, RAM)'을 발표함('23.7.)

    ※ RAM은 UNESCO 회원국이 만장일치로 채택한 AI 윤리 권고안('21.11.)에 따른 것으로, 기존 법/정책 적절성을 평가하고, 공무원/공공기관의 기술 역량을 측정하여 종합 평가함

9) AI 레드팀은 AI 시스템의 결함과 취약성을 식별하기 위해 단점을 일부러 공격하는 역할을 수행함

- 국내도 AI 문제를 자체적으로 검사할 수 있는 AI 윤리 자율점검표 및 신뢰할 수 있는 AI 개발 안내서를 마련함('22.8.)
- 글로벌 빅테크 기업들은 AI 서비스가 개발 단계에서부터 공정성을 확보할 수 있도록 사전 점검할 수 있는 검증 도구를 개발하여 운영 중임

  ※ AI 모델의 문제점을 진단하고 교정하는 대표적인 검증 도구로 IBM의 'AIF360', MS의 'Fairlearn', Google의 'What-If Tool' 등이 있음

3. 인공지능 윤리 정책

해외 인공지능 윤리 정책

- **(국제기구·협의체) UN과 OECD, G20/G7 등 국제기구와 협의체는 AI 윤리의 중요성을 인식하여 AI 윤리 및 신뢰성 관련 원칙 제정 및 파트너십을 통한 글로벌 협력을 추진함**
  - 2021년 11월, UNESCO는 'AI 윤리 권고'를 통해 인류 사회 및 환경·생태계에 미치는 AI의 위험을 최소화하고 사회적 이익 실현을 위한 가치와 원칙을 제시함(UNESCO, 2021)
  - 2023년 7월, UN 인권이사회는 국제사회가 AI를 개발·활용하는 데 인권을 보호하고 증진할 것을 강조한 '신기술과 인권' 결의를 채택함(UNHRC, 2023)
  - 2023년 11월, 'AI 안전 정상회의'에서 AI의 잠재적 위험성에 대한 이해와 관리의 필요성을 인식하고, 공동의 노력을 통해 안전하고 책임감 있는 방식으로 AI를 개발하도록 제안함(GOV.UK, 2023)
  - 2023년 12월, G7 회원국은 첨단 AI 시스템의 위험을 완화하기 위한 종합 AI 정책 프레임워크를 발표하고 AI 개발자가 준수해야 할 행동 강령과 모든 AI 사용자가 지켜야 할 지침을 제시함(일본 총무성, 2023)

    ※ 2023년 5월, 일본에서 열린 G7 정상회의에서 AI 위험 완화 및 기술 발전을 통제하기 위해 생성형 AI 관련 국제규범을 만드는 데 합의하여 진행된 결과임
- **(EU) AI에 관한 세계 최초의 포괄적인 규제로 평가되는 AI 법안을 제정하고 있으며, AI에 따른 위험 통제 및 인권 보호, 투명성·책임성 보장 등을 위한 조치 및 국제협력을 추진 중임**
  - 2021년 집행위원회가 발의한 EU AI 법안이 의회와 이사회의 협의 및 수정을 거쳐 2024년 3월 의회의 표결을 통과하여 2024년 상반기 중 공포될 것으로 예상됨
  - EU 집행위원회는 AI 법이 시행되기 전까지 AI 개발자가 AI 법의 주요 의무를 미리 준수하도록 권장하는 자발적 인공지능 협약(AI Pact)을 공식 출범할 예정임(European Commission, 2024)
  - EU와 미국은 2024년 4월 제6차 EU-US 통상기술위원회(Trade and Technology Council, TTC)에서 AI에 관한 양측의 협력 증대를 선언한 공동 성명을 발표함

    ※ 공동 성명은 AI 안전과 거버넌스에 중점을 두고 AI 기반 기술 개발을 위한 협력을 강화하며, AI 거버넌스 및 규제 시스템의 차이를 최소화하기로 합의함

- **(미국) 바이든 정부 출범 이래 AI 기술 개발 및 활용을 촉진하는 동시에 AI에 의한 위험 관리와 시민 권리 보호를 위한 정책을 발표함**
  - 2023년 10월 바이든 대통령은 '안전하고 신뢰할 수 있는 AI에 대한 행정명령'[10]에 서명하고 AI의 잠재적인 위험을 예방하기 위한 원칙 및 연방 기관이 수행할 조치를 제시함(White House, 2023)

〈표 6〉 '안전하고 신뢰할 수 있는 AI에 대한 행정명령'의 지도 원칙

• AI 안전 및 보안을 위한 새로운 표준 수립 • 시민의 사생활·개인정보 보호 • 평등 및 시민권 증진 • 소비자, 환자 및 학생 등 취약층 보호	• 고용 안정 지원 및 근로자 보호 • 혁신 및 경쟁 촉진 • 연방 기관의 책임 있고 효과적인 AI 사용 보장 • 미국의 AI 분야 글로벌 리더십 증진

- **(중국) AI 사용자의 보호 및 콘텐츠 관리를 강화하고, 불량 정보의 확산을 방지하기 위해 엄격한 규제와 공급자의 의무를 부가함**
  - 2023년 7월 정부는 '생성형 AI 서비스 관리 방법'을 통해 자국 내 콘텐츠를 생성하는 AI 서비스의 관리지침을 발표함(国家互联网信息办公室, 2023)

    ※ 국가 규정에 따른 안전 평가, 법에 따른 훈련 데이터 처리 활동 수행, 데이터 라벨링 요구사항 준수, 사회 규범에 준수를 위한 콘텐츠 관리, 생성 콘텐츠 표시, 개인정보보호 등을 공급자 의무로 부여

- **(영국) AI 주도 국가로 AI의 혁신 촉진과 함께 위험에 대응하고 공공 신뢰를 높일 수 있도록 유연하고 반복적이며 협력적인 규제를 지향함**
  - 2023년 3월 과학혁신기술부는 'AI 규제에 대한 친혁신적 접근' 백서를 발표하고, 혁신을 억제하는 강압적인 법안을 대신 기관들이 상황별 맞춤형 접근 방식을 취하도록 할 계획임을 밝히며 AI 규제 5개 원칙을 제시함(UK DSIT, 2023)

    ※ AI 규제 5개 원칙: ① 안전, 보안, 견고성, ② 투명성 및 설명 가능성, ③ 공정성, ④ 책임성 및 거버넌스, ⑤ 경쟁 가능성 및 보상

- **(일본) 혁신을 방해하지 않도록 AI 규제를 최소화하는 방향을 견지해 왔으나, 최근 EU와 미국 등 AI 규제 강화에 대한 논의가 확산함에 따라 대응 방안을 검토 중임**
  - 2023년 5월 AI 정책 컨트롤타워로 AI 업계 전문가로 구성된 "AI 전략회의"를 발족하고 민간·교육·공공 분야에서의 이용 촉진, AI 위험 대응을 논의 중임
  - 2024년 2월 경제산업성 산하 정보처리추진기구에 AI 안전성 확보 전담 조직인 AI 안전연구소를 설립함(NHK, 2024)
  - 자민당은 2024년 이내에 가짜정보 확산 방지 및 저작권 보호, AI의 윤리적 사용을 보장하기 위한 내용을 담은 AI 법안을 도입할 계획을 발표함

10) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

국내 인공지능 윤리 정책

- (대한민국 인공지능 도약방안) 2023년 9월 '대한민국 인공지능 도약방안'을 발표하고, 민간 주도의 AI 윤리·신뢰성 확보 방안과 신뢰성 R&D 추진을 밝힘(과기정통부, 2023a)
  - 2022년 공개한 신뢰할 수 있는 AI 개발 안내서를 기반으로 AI 제품·서비스의 위험 요인을 분석하고 신뢰성 검·인증을 추진함

    ※ 2024년 과기정통부 고위험 영역[11]에 해당하는 AI 사업을 대상으로 신뢰성 검·인증 실시를 의무화할 계획임
  - 민간이 윤리 원칙을 준수하기 위해 자율적으로 운영하는 윤리위원회를 구성·운영하는 표준지침을 수립함
  - 신뢰성 확보를 위해 AI의 설명가능성과 공정성 개선을 위한 핵심기술 및 초거대 AI의 위험성 완화(환각, 편향, 유해성 표현 등)를 위한 기술 개발을 추진함
- (인공지능 윤리·신뢰성 확보 추진계획) 2023년 10월 「인공지능 윤리·신뢰성 확보 추진계획」을 발표하고 중점 추진 방향을 제시함(과기정통부, 2023b)
  - AI 산업 발전의 전제 조건인 AI 윤리·신뢰성 확보 지원을 위해 분야별 가이드라인 확대 및 민간 자율 신뢰성 검·인증을 추진함
  - AI 위험성에 대응하기 위한 기술 개발 지원을 목적으로 AI 신뢰성 확보 기술 개발 및 AI 평가 데이터 신뢰성 확보 계획을 밝힘
  - 책임 있는 AI 활용을 위한 제도적 기반 마련을 목적으로 AI 생성물 워터마크 도입 검토 등 규제 개선·제도 정립 과제를 추진함

주요 시사점

- 세계 각국은 AI 윤리와 신뢰성 확보를 위해 자국 내 정책적 대응을 위한 원칙을 세워 규제 및 행정적 조치를 적극적으로 실행하고 있음
  - EU의 AI 법안의 최종 통과가 임박함에 따라 우리나라를 포함한 세계 각국은 EU의 법안을 참고하여 자국의 규제에 적용할 것으로 예상됨(브뤼셀 효과[12])
  - 동시에 타국과의 국제협력을 통해 AI 윤리·안전성을 위한 글로벌 표준을 구축하는 데 노력하고 있음
- AI 기술의 지속적인 혁신과 윤리·신뢰성 확보에 대한 사회와 AI 기업 사이에 치열한 논의가 지속될 것으로 전망되며, 이러한 논의를 바탕으로 사회적 합의를 도출하여 혁신과 규제의 균형을 이뤄야 할 것임

11) 국회에 계류 중인 'AI 법안'에서 고위험 영역으로 정의된 에너지·교통, 원자력, 의료기기, 생체정보, 채용·대출·평가, 공공활용 등 분야를 참고하여 설정함
12) 미국 컬럼비아 법대 교수인 Anu Bradford가 'EU의 규칙이 세계의 표준이 된다'는 현상을 지칭하여 만든 용어임

4. 인간과 인공지능, 유토피아로 가는 길

인간과 인공지능의 공존

- (발전과 책임의 균형) 인간과 인공지능의 윈윈(win-win)을 위해서는 AI 기술의 발전과 인간의 책임 간의 균형이 필요함
  - AI 활용성 제고를 위해 사용자의 AI 자원 접근성을 높이려고 하는 노력과 동시에 AI의 잠재적 위험성에 대한 우려가 상존하고 있음
  - 기술 발전으로 인한 사회문제에 유연하게 대응하기 위해서는 인공지능 개발자, 사용자, 정책입안자 각각의 역할에 따라 윤리적 책임을 지도록 해야 함
    ※ 개발자는 윤리적 책임 의식을 가지고 AI를 설계해야 하며, 사용자는 책임 있는 AI 활용과 함께 AI를 감시·통제하는 역할을 수행하고, 정책입안자는 책임 있는 AI 개발과 사용 방향을 제시하고 적절한 윤리적 규범 또는 법적 규제를 마련해야 함
  - AI 윤리는 반드시 금지를 위한 것이 아니며, AI가 초래할 수 있는 위험도 수준을 고려하고 윤리 목적을 명확히 할 때 AI 개발에 제약 요인이 되지 않을 것임
- (인간 중심 AI) AI는 인간을 완전히 대체하기보다는 인간의 능력을 향상시키고 인간과 협업할 수 있는 개체로 인식해야 함
  - AI 중독을 야기하는 인간의 과도한 AI 의존성은 지양하고, AI가 혁신의 조력자로 인식할 수 있는 환경이 조성되어야 함
  - 포스트휴먼(posthuman)[13]이 예견됨에 따라 인간과 협업하는 AI가 신인류로 부상할 것이며, AI와 함께 공존하기 위한 방법을 모색해야 할 때임
    ※ 미국 미래학자 레이 커즈와일은 2030년 지금은 상상할 수 없는 신인류가 탄생할 것이며, 2045년 AI가 모든 인간의 지능을 뛰어넘는 특이점이 올 것이라고 예견함
  - 인간과 AI가 조화롭게 공존하기 위해서는 AI 활용 목적이 인간의 가치와 일치해야 하므로, AI 기술을 활용하는 데 있어 인간의 윤리 의식과 도덕적 판단력이 중요함
    ※ 다양한 기구에서 발표되고 있는 AI 설계를 위한 지침도 공통적으로 인간의 가치관에 부합하게 인공지능이 작동하기 위해 필요한 사항을 포함하고 있음(Jacob Turner, 2023)

정책적 제언

- (미래 AI 사회 시나리오) AI로 인해 발생할 수 있는 변화와 불확실성을 연구하여 미래 AI 사회를 사전에 준비할 수 있도록 시나리오와 대응 정책 수립이 필요함
  - 범용인공지능(Artificial General Intelligence, AGI)을 넘어 이제는 인간보다 뛰어난 슈퍼인텔리전스의 등장이 예상되고 있으므로, 인간의 안전을 확보하기 위한 연구가 선행되어야 함

13) 인간의 주요 능력이 현재의 기준과 한계를 월등히 뛰어넘는 존재이기 때문에 더 이상 인간으로 부를 수 없는 미래 인간을 의미함(노대원, 2018)

※ 최근 미국에서 AI가 인류를 멸종시킬 수준의 위협이 될 수 있다는 보고서[14]를 발간('24.3.)하여, AI 진화로 인한 국가안보 위험성을 경고하고 있음

※ 오픈AI는 10년 내 초지능(superintelligence)이 인류를 위협하는 기술이 될 것으로 전망하고, 인간을 보호하기 위한 연구에 착수함('23)

- AI 확산에 따라 발생할 수 있는 이슈를 도출하고, 긍·부정적 시나리오로 나누어 각각에 대한 결과를 예측하고 준비할 수 있는 정책 설계가 필요함
- AI의 기술적 한계뿐만 아니라 산업 변화, 활용 격차 등 미래에 발생할 수 있는 다양한 상황을 고려하여, 사회적 혼란을 야기할 수 있는 문제에 대한 해결책을 사전에 마련해야 함

- **(AI 평가 모델) 책임 있는 AI 사용과 평가를 위해서는 AI 성능과 모델의 위험성을 객관적으로 판단할 수 있는 기준이 마련되어야 함**
  - 주요 AI 기업 및 개발자들은 서로 다른 AI 벤치마크를 사용하여 모델을 평가하고 있어, AI의 위험성을 객관적·체계적으로 비교하기 어려움(HAI, 2024)

    ※ AI Index 2024 보고서는 AI 기술은 기존 벤치마크를 초월한 새로운 기술 역량에 도달했으며, AI 모델의 책임 있는 사용과 평가에 대한 표준화가 부족하다고 발표함
  - 세계 각국 정부는 윤리성, 신뢰성 차원에서 AI 기술이 발생시킬 수 있는 위험성에 대해 인식하고 평가할 수 있는 체계를 마련하기 위해 노력 중임

    ※ 각국 정부는 주요 AI 기업과 첨단 모델 출시 전 안전성을 시험하는데 합의(블레츨리 선언, '23.11.)했으며, 미국과 영국은 AI 모델의 안전성 테스트 개발에 관한 협력을 위해 세계 최초로 양자 협정을 맺음('24.4.)
  - AI 모델이 빠른 속도로 개선되고 있으나 AI 기술 수준과 안전성을 판단할 수 있는 공식적인 평가 기준이 없어 객관적 활용에 제약이 있으므로, 이를 해결하기 위한 기준과 평가의 표준화가 필요함

- **(맞춤형 AI 윤리) 미래 AI와 인간이 바람직한 관계를 가지기 위해서는 AI 발전에 상응하는 교육 과정 및 제도 마련이 필수적임**
  - 선한 목적으로 AI가 개발·활용될 수 있도록 기술과 윤리 활용 역량이 포함된 'AI 리터러시'[15] 고도화와 함께 AI 개발의 모든 단계에 윤리가 적용되는 '임베디드 에틱스(Embedded EthiCS)'[16] 교육 과정 운영이 필요함
  - 또한 AI 윤리는 거시적인 담론 수준에서 벗어나 AI가 적용된 산업별 수준을 고려하여 맞춤형 AI 윤리 정책이 수립되어야 함
  - 글로벌 시장에서 국내 AI 기업이 공정하게 경쟁할 수 있도록 미국, 유럽 등 해외 규제와 발맞춰 AI 기본법[17]이 시급히 제정되어야 함

14) 美 국무부가 의뢰해 Gladstone AI가 발표한 'Defense in Depth: An Action Plan to Increase the Safety and Security of Advance AI' 보고서

15) AI 리터러시(AI literacy)는 기술에 대한 이해뿐만 아니라 윤리적인 측면에서 AI를 활용할 줄 아는 역량을 의미함

16) 임베디드 에틱스(Embedded EthiCS)는 앨리스 시먼스 하버드대 교수가 발표한 개념으로, 기술 교육과정에 처음부터 윤리 문제를 반영해서 두 문제를 함께 고민하도록 해야 한다는 취지를 담고 있음(윤송이, 2022)

17) AI 기본법(인공지능 산업 육성 및 신뢰 기반 조성 등에 관한 법률안)이 발의('22.12.7.)되었지만 국회에 계류 중임

참고문헌

- 과학기술정보통신부(2023a), 「전국민 인공지능 일상화 실행계획」, 2023.09.
- 과학기술정보통신부(2023b), 「인공지능 윤리·신뢰성 확보 추진계획」, 2023.10.
- 노대원(2018), 「포스트휴머니즘 비평과 SF: 미래 인간을 위한 문학과 비평 이론의 모색」, 한국비평문학회, 제68호, pp.110-133.
- 박찬준·이원성·김윤기·김지후·이활석(2023), 「초거대 언어모델 연구 동향」, 정보과학회지, 제41권, 제11호, pp.8-24.
- 삼정KPMG(2023a), 「CES 2024 프리뷰: 미리 보는 CES 트렌드」, 2023.12.
- 삼정KPMG(2023b), 「챗GPT와 생성형 AI가 만드는 빅테크 플랫폼 혁신」 보고서.
- 유화선·정도범·이준·최희석(2023), 「데이터 기술(Data Technology) 분류체계 연구」, 한국콘텐츠학회 논문지, 제23권, 제1호, pp.61-73.
- 윤송이(2022), 「가장 인간적인 미래」, 웨일북, 2022.11.01.
- 한국지능정보사회진흥원(NIA)(2023), 「THE AI REPORT」, 2023.01.
- Bahar Memarian, Tenzin Doleck(2023), "Fairness, Accountability, Transparency, and Ethics(FATE) in Artificial Intelligence(AI) and higher education: A systematic review", Computers and Education: Artificial Intelligence, Vol.5.
- Center for Research and Development Strategy(CRDS)(2023), "科学技術・イノベーションの土壌づくりとしてのELSI/RRI: Toward responsible innovation".
- Chosun Biz(2024), https://biz.chosun.com/it-science/ict/2024/03/12/XDRFR5TOYVFWRCGJKGIBCCYECQ/.
- Eric Colson(2019), "인공지능 주도형 의사결정이란 무엇인가", Harvard Business Review, 2019. 9-10월호.
- European Commission(2024), "AI Pact", 2024.03.06.
- Future of Life Institute(2017), "Asilomar AI Principles", 2017.08.11.
- Gladstone AI(2024), "An Action Plan to increase the safety and security of advanced AI", 2024.02.26.
- Goldman Sachs(2023), "The Potentially Large Effects of Artificial Intelligence on Economic Growth", 2023.03.26.
- GOV.UK(2023), "About the AI Safety Summit 2023", 2023.11.01.
- HAI(2024), "Artificial Intelligence Index Report 2024", Stanford University, Human-Centered Artificial Intelligence.
- Jacob Turner(2023), 「로봇 법규」, 한울아카데미, 2023.11.07.
- Markets and Markets(2024), "Artificial international market global forecast to 2030".
- McKinsey&Company(2023), "The economic potential of generative AI", 2023.06.
- NHK(2024), "生成AIめぐるリスク 政府の研究機関が14日に設立へ", 2024.02.08.
- White House(2023), "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence", 2023.10.30.
- World Economy Forum(2024), "Global Risks Report 2024", 2024.01.10.
- UK DSIT(2023), "A pro-innovation approach to AI regulation", 2023.03.29.
- UNESCO(2021), "Recommendation on the Ethics of Artificial Intelligence", 2021.11.23.
- UNHRC(2023), "New and emerging digital technologies and human rights", 2023.07.12.
- 일본 총무성(2023), "Hiroshima AI Process G7 Digital & Tech Ministers' Statement", 2023.12.01.
- 国家互联网信息办公室(2023), "生成式人工智能服务管理暂行办法", 2023.07.13.

저 자

유화선
KISTI 정책전략본부 정책연구센터
선임기술원
T. 042-869-0816
E. hsyou@kisti.re.kr

윤병성
KISTI 정책전략본부 정책연구센터
박사후연구원
T. 042-869-0726
E. bs.yoon@kisti.re.kr

최희석
KISTI 정책전략본부
책임연구원
T. 042-869-0738
E. choihs@kisti.re.kr

KISTI ISSUE BRIEF 제68호

발 행 일 2024. 05. 28.

발 행 인 김재수

편 집 위 원 조민수, 이혜진, 고미현, 김한국, 이상환, 최희석, 최선희, 김윤정

발 행 처 34141 대전광역시 유성구 대학로 245
한국과학기술정보연구원 정책연구센터
https://www.kisti.re.kr

I S S N 2635-5728