KISTI
ISSUE BRIEF

ISSN 2635-5728

과학기술인프라, **데이터로 세상을 바꾸다**

https://kisti.re.kr/promote/post/issuebrief

2021. 11. 01. 제35호

인간과 인공지능(AI)의 공존을 위한 사회·윤리적 쟁점 : 신뢰할 수 있는 인공지능 실현 방안

정도범* · 유화선**

오늘날, 인공지능(AI)은 거의 모든 산업에 적용되어, 복잡한 문제 해결, 생산성·효율성 증가, 비용 절감 등의 경제적 가치를 창출하고 있다. 하지만 AI의 발전과 확산은 사회·윤리적 측면에서 예상하지 못한 부작용도 초래하고 있어, 신뢰할 수 있는 AI 실현을 위한 논의가 요구된다. 세계 각국은 AI의 활용·확산 과정에서 발생할 수 있는 위험이나 부작용을 방지하기 위한 방안을 마련하고 있으며, 우리나라도 AI 윤리, 사람 중심의 AI의 중요성에 대한 논의를 추진 중에 있다. AI의 활용 활성화를 위해 발생 가능한 문제에 대해 사전에 논의하거나 다양한 상황에 대한 구체적인 논의가 이루어져야 한다. 또한 AI 윤리는 AI 자체가 아닌 인간의 윤리에 대한 논의라는 점도 인식해야 할 것이다. 신뢰할 수 있는 AI 실현을 위해 기술적 측면뿐만 아니라 사회·윤리적 측면에 대해서도 함께 논의되어야 하며, 사회적 합의를 이끌어내기 위해 거버넌스를 구축하는 방향도 고려할 수 있을 것이다.

CONTENTS

* KISTI 정책부 정책연구실 책임연구원 T. 042-869-1825 E. dbchung@kisti.re.kr
** KISTI 정책부 정책연구실 선임기술원 T. 042-869-0816 E. hsyou@kisti.re.kr

KiSTi 한국과학기술정보연구원 Korea Institute of Science and Technology Information

1. 들어가며

인공지능 시대의 도래

- **(인공지능의 확산) 2016년 3월에 개최된 이세돌과 알파고의 대국은 인공지능(AI)[1]에 대한 인식을 확산하는 계기가 되었고, 2019년 12월에 발생한 코로나19(COVID-19)는 디지털 전환을 가속화하여 AI의 활용을 촉진하고 있음**
  - AI 기술은 GPS 내비게이션, 이메일 스팸 필터, 언어 번역, 신용카드 사기 경고, 도서 및 음악 추천, 컴퓨터 바이러스로부터 보호, 에너지 사용 최적화 등 우리가 잘 인식하지 못하고 있으나, 우리 생활 전반에 확산되어 있음(Mitchell, 2019)
  - 구글 딥마인드는 이세돌과의 대국에서 승리한 '알파고 리(AlphaGo Lee)'를 개발한 후 '알파고 마스터', '알파고 제로', '알파 제로'로 AI의 성능을 계속 향상시켰으며, AI 활용 분야도 생물학과 의학 분야 등으로 확장함
  - 오늘날, AI는 국가 경쟁력의 핵심 동력으로 부상함에 따라, 미국, 중국 등은 AI 주도권 확보를 위해 국가 차원에서 투자와 지원을 적극 확대하고 있음
- **(인공지능의 중요성) AI는 거의 전 산업에 적용되어, 복잡한 문제 해결, 생산성·효율성 증가, 비용 절감 등의 경제적 가치를 창출할 수 있음**
  - 운송(자율주행자동차 포함), 금융, 농업, 마케팅/광고, 과학, 의료, 사법, 보안, 공공행정, 증강/가상현실 등의 분야에서 AI 기술이 빠르게 도입되고 있음(OECD, 2019a)
  - 우리나라에서도 전 산업의 AI 활용을 촉진하기 위해 제조업에서부터 전 산업으로 대규모 데이터를 기반으로 한 AI 융합 프로젝트(AI+X)를 추진 중에 있음(관계부처 합동, 2019)

인공지능의 사회·윤리적 이슈

- **(인공지능의 부작용) AI의 발전과 확산은 사회·윤리적으로 예상하지 못한 부작용도 초래하고 있음**
  - 2015년 구글 포토의 AI가 흑인 여성을 '고릴라'로 인식하거나, 2016년 마이크로소프트의 AI 챗봇[2] '테이(Tay)'는 인종차별 발언 등으로 인해 16시간 만에 서비스를 중단함
  - 우리나라에서도 2020년 12월에 출시된 AI 챗봇 '이루다'가 성소수자나 장애인에 대한 혐오 발언으로 출시 20일 만에 서비스를 종료함

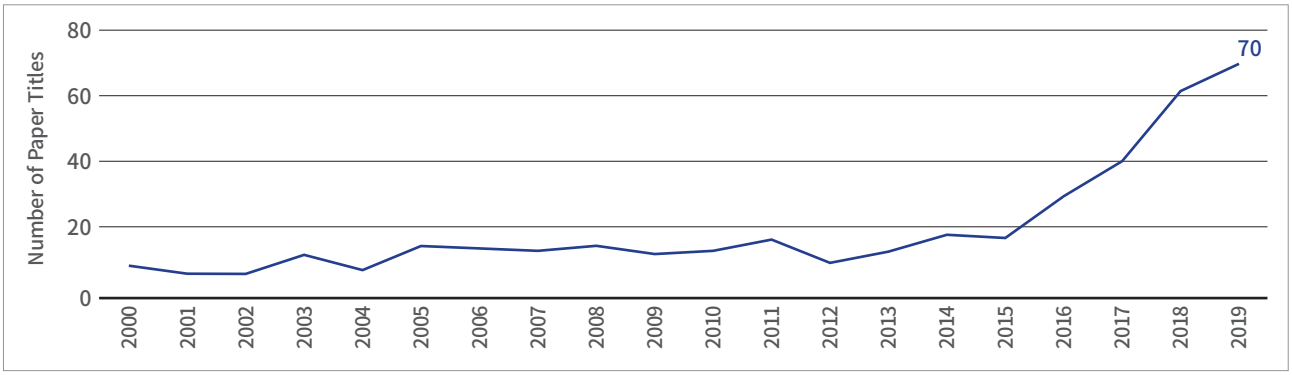
1) 인공지능(AI: Artificial Intelligence)은 인간의 지적 능력을 컴퓨터로 구현하는 과학기술로서, 과거에 산업화 과정에서 기계가 인간의 육체노동을 대체했다면 앞으로는 AI가 인간의 지적 능력을 대체하는 수준까지 발전할 것으로 예측됨(관계부처 합동, 2019)
2) AI 챗봇을 대화형 인공지능(conversational AI)으로 표현하기도 함(신용우·정준화, 2021)

- 미국의 AI 기반 재범 예측 시스템인 COMPAS[3]에서 백인보다 흑인의 재범률을 훨씬 높게 예측하기도 하였으며, 딥페이크[4] 기술을 활용한 가짜 동영상이나 편집물도 큰 사회적 문제가 되고 있음
- AI 알고리즘 결함으로 인해 인명사고를 일으킨 자율주행자동차, 인류에게 직접적인 위협이 될 수 있는 자율살상무기 개발 등도 AI에 대한 위험성을 경고함

● **(인공지능 윤리) 인공지능 윤리(AI ethics)는 AI 시스템의 오용이나 남용, 잘못된 설계, 의도하지 않은 부정적인 결과가 초래할 수 있는 개인 및 사회적 피해에 대한 대응으로 등장함(Leslie, 2019)**

- 세계 각국은 AI 윤리의 중요성을 인식하고 신뢰할 수 있는 AI 실현을 위한 원칙을 마련 중에 있음
- 2015년 이후, AI 컨퍼런스에 제출된 논문의 제목에 '윤리(ethics)' 관련 키워드가 포함된 논문 수가 크게 증가함(HAI, 2021)

[그림 1] AI 컨퍼런스에서 윤리(ethics) 키워드를 언급한 논문 제목 수(2000~2019년)



출처) HAI(2021)

- AI로 인한 잠재적 피해는 그 파급효과가 매우 크기 때문에, 윤리적이고 신뢰할 수 있는 적절한 기준 및 방안을 모색해야 함

<표 1> 인공지능으로 인한 잠재적 피해

• 편견과 차별	• 사생활 침해
• 개인의 자율성, 청구권 및 권리에 대한 거부	• 사회적 관계의 고립과 해체
• 불투명하고 설명할 수 없거나 정당화될 수 없는 결과	• 신뢰할 수 없거나 안전하지 않고 품질이 낮은 결과

출처) Leslie(2019)

● **(인공지능 준비 현황) 2020년 발표된 '정부 AI 준비 지수(Government AI Readiness Index)'[5]를 살펴보면 우리나라는 7위를 차지하여 2019년(26위) 대비 순위가 크게 상승함(Oxford Insights, 2020)**

- 특히, 우리나라는 '데이터 및 인프라' 측면에서 스마트폰과 인터넷의 높은 보급률 등으로 인해 아시아 지역에서 1위를 차지함

3) Correctional Offender Management Profiling for Alternative Sanctions
4) 딥페이크(deepfake)는 딥러닝(deep learning)과 페이크(fake)의 합성어로, 인공지능(AI) 기술을 활용한 제작물 또는 제작 프로세스 자체를 의미함
5) 정부 AI 준비 지수는 크게 정부(Government), 기술 영역(Technology Sector), 데이터 및 인프라(Data and Infrastructure)와 같이 3개의 차원으로 구분됨 (Oxford Insights, 2020)

- 하지만 정부가 AI를 얼마나 책임 있게 사용하는지를 측정하는 '책임 있는 AI 하위 지표(Responsible AI Sub-Index)'[6]에서 우리나라는 34개 국가 중 21위를 차지함
- 즉, 우리나라는 AI 활용을 위한 인프라가 잘 구축되어 있고 AI 분야에 대한 지원도 꾸준히 강화되고 있지만, AI 거버넌스나 윤리, 정책/제도 등의 측면은 상대적으로 미흡하다고 볼 수 있음

● **(사회·윤리적 논의 필요성) 다가오는 AI 시대에 대비하여 신뢰할 수 있는 AI 실현을 위해 AI의 불완전성으로 인해 발생 가능한 사회·윤리적 쟁점에 대해 지속적으로 논의해나가야 할 것임**

2. 인공지능 정책 동향

국외 인공지능 정책 동향

● **(OECD) 2019년 5월, 42개 국가는 AI 시스템을 안전하고 공정하며 신뢰할 수 있는 방식으로 설계하는데 동의하는 AI에 관한 OECD 원칙에 서명함(OECD, 2019b)**

- OECD는 AI가 새로운 도전을 추구하고 있지만 불안과 윤리적 우려도 일으킴에 따라, 안전과 개인정보보호 등의 측면에서 사람들이 신뢰할 수 있는 AI에 대해 언급함
- 그리고 AI R&D 투자, AI를 위한 디지털 생태계 육성 및 정책 마련, 노동시장 전환을 위한 준비, 신뢰할 수 있는 AI를 위한 표준 개발 및 국제 협력 등을 강조함

<표 2> AI에 관한 OECD 원칙

• AI는 포용적 성장, 지속 가능한 개발 및 웰빙을 주도하여 인간과 지구에 혜택을 제공해야 함
• AI 시스템은 법, 인권, 민주적 가치 및 다양성을 존중하는 방식으로 설계되어야 하며, 공정하고 정의로운 사회를 보장하기 위해 적절한 안전장치를 포함해야 함
• 사람들이 AI 시스템에 참여하는 과정에서 이해할 수 있도록 AI 시스템에 대한 투명성과 책임 있는 공개가 있어야 함
• AI 시스템은 평생 동안 강건하고 안전한 방식으로 작동해야 하며, 잠재적 위험을 지속적으로 평가하고 관리해야 함
• AI 시스템을 개발, 배포 또는 운영하는 개인 및 조직은 위의 원칙에 따라 적절한 작동을 위한 책임을 져야 함

출처) OECD(2019b)

● **(EU) EU 집행위원회는 2019년 4월 '신뢰할 수 있는 AI를 위한 윤리 지침(Ethics Guidelines for Trustworthy AI)', 2020년 2월 '인공지능 백서'를 발표하는 등 다양한 정책을 추진함**

- 2019년 4월, 신뢰할 수 있는 AI를 위해 ① 모든 관련 법률과 규정 준수, ② 윤리적 원칙과 가치 준수, ③ 기술적/사회적 측면에서 강건성 확보를 강조함(European Commission, 2019)

6) 책임 있는 AI 하위 지표는 크게 포괄성(Inclusivity), 책임성(Accountability), 투명성(Transparency), 개인정보보호(Privacy)와 같이 4개의 차원으로 구분됨 (Oxford Insights, 2020)

- 2020년 2월, AI의 활용을 촉진하는 동시에 AI와 관련된 위험을 해결하기 위한 AI 프레임워크를 제시하는 '인공지능 백서'를 발표함(European Commission, 2020)
- 2021년 4월, EU 집행위원회는 세계 최초로 '인공지능 법안(Artificial Intelligence Act)'을 발표하고, AI의 위험 수준별 규제 방안을 제안함(European Commission, 2021)

<표 3> 위험 수준별 AI 시스템 및 규제 방안

위험 수준	주요 설명	규제 방안
용납할 수 없는 위험	• 기본권 침해 등 EU 가치에 위배되는 AI 시스템	금지
고위험	• 건강/안전/기본권 등 고위험을 야기할 수 있는 AI 시스템	관리 및 준수 의무 부과
제한된 위험	• AI 챗봇처럼 개인과 상호작용하거나 감정과 특징 인식 • 이미지/영상 콘텐츠 생성 및 조작	투명성 의무 부과 (사용자에게 AI 시스템 작동 방식 등 공지)
최소한의 위험	• AI 기반 비디오게임, 스팸 필터 등	규제하지 않음

출처) 양희태(2021), European Commission(2021)

● **(세계 각국) 미국, 중국, 일본 등 세계 주요국도 AI의 활용·확산 과정에서 위험이나 부작용을 방지하기 위한 방안을 마련하고 있음**

- 미국은 2019년부터 AI를 규제하기 위한 법안을 도입하고 있으며, 구글, 마이크로소프트 등의 주요 기업을 중심으로 윤리적인 AI 실현을 위해 자율적인 AI 개발 원칙을 마련함
- 중국은 2020년 8월 AI 산업의 건전하고 지속 가능한 발전을 위해 '국가 차세대 AI 표준체계 구축 지침'을 발표하였으며, AI의 투명성, 책임성, 개인정보보호 등 윤리·보안 이슈에 대해서도 관심을 표명함
- 일본은 2019년 3월 '인간 중심의 AI 사회 원칙'을 발표하였고, 2020년 7월 사회·경제·윤리·법적 과제에 대한 'AI 활용 가이드라인'을 제시함

국내 인공지능 정책 동향

● **(사람 중심 강조) 우리나라는 2019년 12월 '인공지능 국가전략'을 발표하며, 사람 중심의 AI 실현을 강조함(관계부처 합동, 2019)**

- AI 제품·서비스의 확산에 대응하여 신뢰성·안전성 등을 검증하는 품질관리체계를 구축하고, AI 윤리 관련 논의를 추진함
- 2020년 12월, 기술의 급속한 발전과 함께 AI 윤리 이슈가 지속적으로 제기됨에 따라, 사람이 중심이 되는 '인공지능(AI) 윤리기준'을 마련함(관계부처 합동, 2020)

※ 인공지능이 지향하는 최고 가치를 '인간성(Humanity)'으로 설정하고, 모든 인공지능은 '인간성을 위한 인공지능(AI for Humanity)'이여야 함을 명시함

<표 4> 인공지능(AI) 윤리기준 주요내용

- (목표 및 지향점)
  ❶ 모든 사람이 ❷ 모든 분야에서 ❸ 자율적으로 준수하며 ❹ 지속 발전
- (3대 기본원칙)
  ① 인간 존엄성 원칙, ② 사회의 공공선 원칙, ③ 기술의 합목적성 원칙
- (10대 핵심요건)
  ① 인권 보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성

출처) 관계부처 합동(2020)

- **(인공지능 실현 전략) 2021년 5월, AI 윤리의 실천과 이용자의 AI 수용성 향상을 위해 '신뢰할 수 있는 인공지능 실현 전략(안)'을 발표함(관계부처 합동, 2021)**
  - AI의 기술적 한계 극복과 함께, 오·남용 등에 따른 잠재 위험 예방을 위한 제도 보완과 윤리의식 확산을 강조함
  - 주요 추진전략으로 '신뢰 가능한 인공지능 구현 환경 조성', '안전한 인공지능 활용을 위한 기반 마련', '사회 전반 건전한 인공지능 의식 확산'을 제시함

주요 시사점

- **세계 각국은 AI 기술의 주도권 확보를 위한 노력과 함께, AI의 부작용 및 잠재적 위험 등에 대비하여 AI 윤리의 중요성을 인식하고 있음**
- **하지만 아직까지 국가 차원의 AI 윤리는 큰 틀에서의 기준만 마련되어, 세부적인 논의를 바탕으로 사회적 합의가 이루어져야 할 것임**

3. 인공지능의 사회·윤리적 쟁점

발생 가능한 문제에 대한 사전 논의

- **(AI 사고 예방) 고위험 AI에 대한 명확한 기준을 설정하여, 사전에 발생 가능한 문제에 대해 지속적인 논의를 통해 예방할 수 있어야 함**
  - 현재 AI 사고가 발생한 후 AI 원칙이나 가이드라인, 윤리기준 등의 후속조치가 이루어지고 있어, 사전에 충분한 논의가 이루어지지 않고 있음
  - AI의 복잡성, 불완전성 등으로 인해 개발자도 예측하지 못한 사고가 항상 발생할 수 있기 때문에, AI 활용·확산을 위해서도 다양한 측면에서 AI의 잠재적 위험을 고민해야 함

- AI는 오직 설정된 목표를 달성하기 위해 작동하는데, 인간의 존엄성이나 생명, 윤리에 대한 가치를 벗어난 선택을 할 가능성에 대해 모니터링할 필요가 있음

※ AI에게 인간의 고통을 없애라는 임무를 부여했을 때 그 고통을 없애는 방법 중 하나로 인간을 없애려고 할 수도 있음(KAIST, 2019)

<표 5> 예측하지 못한 AI 사고가 발생하는 3가지 유형

- 강건성(robustness)의 실패: 시스템이 오작동을 일으킬 수 있는 비정상적이거나 예상하지 못한 입력(inputs)을 받을 때
- 세밀성(specification)의 실패: 시스템이 설계자나 운영자가 의도한 것과 미묘하게 다른 것을 달성하려고 할 때
- 확인성(assurance)의 실패: 운영 중에 시스템을 적절하게 모니터링하거나 통제할 수 없을 때

출처) Arnold and Toner(2021)

● **(AI 의사 사례) 왓슨(Watson), 닥터앤서 등은 의료 데이터를 바탕으로 진단이나 치료 방법을 보조하는 역할에 한정되어 있지만, AI 기술이 발전함에 따라 AI 의사의 역할과 범위, 사고 발생 시 책임 문제 등에 대해 선제적인 논의가 요구됨**

- 물론, 아직까지 AI가 의사를 대체하기 힘든 것이 현실이지만, 앞으로 AI가 점점 더 의료 분야에 도입되어 의료 패러다임을 변화시킬 것이란 점은 분명함(최윤섭, 2018)
- AI의 판단을 참고하여 최종적으로 의사결정을 내리는 것은 의사지만, 정확도와 숙련도가 높아진 AI와 의사의 의견이 불일치하거나 환자가 AI의 의견을 신뢰할 경우, 의사의 탈숙련화 가능성 등 의사의 역할에 대한 진지한 고민이 필요할 것임

※ 예를 들어, 비행기의 경우 자동항법장치나 오토파일럿의 발전으로 인해 오늘날 조종사들은 이착륙 외에 조종간을 거의 잡지 않게 되어 비행 기술의 숙련도가 하락하게 됨

- AI 의사는 단순히 기술적 측면뿐만 아니라, 인간의 생명을 책임지는 의료 분야에 AI가 적용된다는 측면에서 매우 다양하고 복잡한 이슈가 등장할 수 있음

다양한 상황에 대한 구체적인 논의

● **(AI+X) 앞으로 과학기술뿐만 아니라 운송, 의료, 금융, 교육, 법률, 보안, 마케팅 등 광범위한 분야에서 AI가 활용될 것이며, AI가 의사결정을 해야 하는 상황에 지속적으로 직면하게 될 것임**

- AI 시대에서는 전 산업에 AI가 적용되어 일자리, 정책/제도 등에 큰 변화를 일으킬 것임

※ 예를 들어, 자율주행자동차가 활성화될 경우 직접 자동차를 운전하는 것 자체가 불법이 될 수도 있음

- 하지만 AI로 인해 발생할 수 있는 사고에 대한 책임 소재나 보상 등이 여전히 불분명하고, AI 전반에 관한 세부적인 윤리기준도 미흡한 상황임
- '사람이 중심이 되는 AI 윤리기준'이란 선언적인 측면에서 더 나아가, AI 시대의 다양한 상황에서 어떻게 AI가 사람 중심으로 의사결정을 할 것인지 등에 대해 구체적인 논의가 필요함

- **(자율주행자동차 사례) 자율주행자동차가 다른 차량이나 보행자를 피해야 하는 상황에 직면했을 때 AI의 의사결정, 즉 사회·윤리적 선택 시 우선해야 할 가치에 대해 지속적인 논의가 이루어져야 함** (McDougall, 2019)
  - AI의 의사결정은 탑승자 및 자체 차량, 보행자, 생존 확률, 전체적인 피해 규모 등과 같이 어떤 측면을 우선적으로 고려해야 하는지 사회적 합의가 요구됨
  - 보행자의 경우 노인, 어린이를 둔 어머니, 학생 중에서 우선 피해야 할 대상을 어떤 기준으로 선택할 것이며, 개인 또는 집단 등에 따라 우선순위가 어떻게 달라져야 하는지 등도 주요 논의사항이 될 수 있음
  - 또한 차량에서도 승용차나 트럭 등이 고려될 수 있으며, 트럭 중에서도 유조선 트럭일 경우 등과 같이 다양한 상황이 발생할 수 있음
  - 만약 AI가 학습을 통해 특정한 상황에서 탑승자를 희생시키는 결정을 한다면 사람들은 자율주행 자동차를 타려고 하지 않을 것이므로, 각종 상황에 대한 구체적인 대응 방안을 마련해야 함

인공지능이 아닌 인간의 윤리에 대한 논의

- **(AI 윤리 이슈) AI 윤리는 AI 자체라기보다 AI를 개발하고 운영하는 규범으로의 윤리라는 점을 인식해야 함**(KAIST, 2019)
  - 오늘날, AI의 핵심은 데이터를 기반으로 학습하고 일정한 패턴을 찾아 예측하는 머신러닝인데, 이러한 학습 알고리즘을 AI가 자율성을 가진 것으로 과대평가하는 경향이 있음(KAIST, 2019)
    ※ 알파고가 학습 알고리즘을 바탕으로 바둑에서 인간에게 승리하긴 했지만, 아직 인공지능이 자율적으로 행동하거나 인간의 능력을 대체한다고 볼 수 없음
  - 결국 AI 윤리는 AI가 아닌 AI 개발자나 설계자, 즉 인간의 윤리로 볼 수 있음
  - 자율살상무기의 경우에도 AI 자체가 아니라 살상 대상을 결정하고 그 무기를 개발하여 사용하는 인간이 비난받아야 할 것임

- **(인간의 윤리) 앞으로 AI의 개발 과정에서 인간의 윤리적 가치를 반영한 설계가 이루어져야 하고, 인간을 위한 방향으로 AI가 활용되어야 함**
  - AI 기술의 발전은 '인간 고유의 능력은 무엇인가?', '인간에게 중요한 가치는 무엇인가?', '과연 인간은 공정한가?' 등에 대해 고민하고 답을 찾아나가는 계기가 될 것임
  - 인간의 윤리에 대한 논의는 과학기술 분야뿐만 아니라 인문·사회 분야를 포함한 다학제적(multidisciplinary) 접근이 요구되며, 산·학·연·정 등의 구성원들이 함께 참여하고 협력해야 함

4. 인공지능 활용 활성화를 위한 제언

신뢰할 수 있는 인공지능 실현 방향

- **(인공지능의 미래) 오늘날 AI는 '메타버스'[7] 등을 포함한 모든 분야에서 활용되고 있으며, 앞으로 '싱귤래리티'[8], '1인 1로봇' 등과 같이 AI가 크게 발전하는 미래를 전망하고 있음**
  - 최근 메타버스 공간에서 실제 사람인지, AI인지 구분하기 힘든 다양한 특성을 가진 AI 캐릭터들이 존재하는 등 AI는 점점 더 우리의 생활과 밀접하게 관련될 것임
  - 아직까지 AI는 '약한 AI'[9] 수준에 머물고 있지만, 향후 '강한 AI'[10]가 도래할 것을 고려하여, 신뢰할 수 있는 AI 실현을 위해 기술적 측면과 사회·윤리적 측면을 모두 반영한 가이드라인 등을 수립해야 함

- **(과학기술 기반 AI 활성화) 현재 AI와 관련하여 이슈가 되고 있는 편향성, 불투명성, 안전성 등 AI의 불완전성을 해결하기 위한 과학기술적 측면의 노력이 요구됨**
  - AI의 핵심 원천은 양질의 데이터이며, AI의 편향성이나 차별성 논란을 해소하기 위해 충분한 학습 데이터를 구축해야 함
  - AI는 '블랙박스' 구조를 가진 학습 알고리즘으로 인해 많은 불확실성을 내포하고 있으므로, 설명 가능한 인공지능(XAI: Explainable AI)을 위한 연구가 더욱 활발히 수행되어야 함
  - 또한 예측하지 못한 AI 사고나 부작용에 대비하여 AI 개발 및 운영과 관련된 표준화된 AI 가이드라인을 수립해야 할 것임

- **(윤리의식 기반 AI 활성화) AI가 활용·확산되기 위해 사회·윤리적 이슈에 잘 대처함으로써 AI에 대한 신뢰성을 확보할 수 있어야 함**
  - AI로 인한 사회적 혼란과 충돌을 방지하기 위해 사후 규제도 중요하지만, 예방적인 차원에서 AI 윤리에 대한 지속적인 논의가 필요함
  - 또한 AI 윤리의식을 제고하기 위해서는 선언적인 원칙을 넘어 발생할 수 있는 다양한 상황을 고려한 구체적인 기준을 마련해야 함
  - AI 윤리, 즉 AI를 개발하고 활용하는 인간의 윤리에 대한 포럼, 세미나, 공청회 등을 개최하여 사회적 합의를 이끌어내야 할 것임

7) 메타버스(metaverse)는 가공, 추상을 의미하는 '메타(meta)'와 현실 세계를 의미하는 '유니버스(universe)'의 합성어로, 3차원 가상 세계를 의미함
8) 싱귤래리티(singularity)는 인공지능이 발전하여 인간의 지능을 뛰어넘는 시점을 의미하며, 레이 커즈와일(Ray Kurzweil)은 이 시점을 2045년으로 예측함
9) 약한 AI는 체스, 바둑 등 특정한 분야에서 뛰어난 연산 능력으로 사람의 업무에 도움을 주는 AI를 의미하며, 구글 딥마인드의 알파고, IBM의 왓슨 등이 해당됨
10) 강한 AI는 인간보다 지능 수준이 높고 어떤 문제에 대해 스스로 사고하여 종합적으로 판단하는 AI를 의미하며, 영화 아이언맨의 자비스 등이 해당됨

정책적인 제언

- **(정책 방향) AI 윤리와 관련된 이슈를 총괄하고 조정하여 사회적 합의를 도출하기 위한 거버넌스를 구축해야 함**
  - 신뢰할 수 있는 AI를 실현하기 위한 명확한 목표 설정, 이해관계자들의 협의 등을 위한 조직을 통해 AI 윤리, 정책/제도 등을 수립해야 함
  - AI 기술 개발과 함께 AI 윤리, 신뢰성·안전성 등을 위한 투자를 확대하고, 공공·민간, 국내외 협력 체계를 추진해야 할 것임

<표 6> 신뢰할 수 있는 AI를 위한 정책 방향

• AI 사고 및 위험 상황에 대한 정보 공유를 촉진하고, 민간 부문과 협력하여 AI가 언제 어떻게 실패하는지에 대한 공통의 지식 기반 구축
• 중요하지만 현재 예산이 충분하지 않은 AI 안전 R&D 분야에 대한 투자
• AI 표준 개발 및 테스트 역량에 대한 투자함으로써 안전하고 신뢰할 수 있는 AI 시스템 개발 지원
• R&D 제휴와 다자적 조직을 통해 AI 사고 위험을 줄일 수 있도록 국제적인 협력 추진

출처) Arnold and Toner(2021)

- **(AI에 대한 인식 확산) AI 윤리가 AI 개발자나 설계자에게 한정된 것이 아닌 모든 사람들과의 공유를 통해 신뢰할 수 있는 AI를 위한 인식을 확산해나가야 함**
  - 모든 사람들이 AI 사고나 부작용, 윤리 등에 대해 쉽게 공유할 수 있는 공통의 지식 기반을 구축할 필요가 있음
  - 공통의 지식 기반은 단지 구축하는 것보다 실제로 많은 사람들이 참여하여 소통할 수 있는 실질적인 논의의 장을 마련하는 방향이 되어야 함

참고문헌

- 관계부처 합동 (2019), 「인공지능 국가전략」, 2019.12.
- 관계부처 합동 (2020), 「사람이 중심이 되는 인공지능(AI) 윤리기준」, 제19차 4차산업혁명위원회 심의안건 제2호, 2020.12.23.
- 관계부처 합동 (2021), 「사람이 중심이 되는 인공지능을 위한 신뢰할 수 있는 인공지능 실현 전략(안)」, 2021.05.13.
- 글로벌 과학기술정책정보 서비스, https://now.k2base.re.kr/portal/main/main.do
- 신용우·정준화 (2021), 「'이루다'를 통해 살펴본 인공지능 활용의 쟁점과 과제」, 국회입법조사처, 이슈와 논점, 제1799호.
- 양희태 (2021), 「신뢰할 수 있는 인공지능을 위한 최근 주요국 대응동향 및 시사점」, 정보통신정책연구원, AI Trend Watch, 2021-11호.
- 최윤섭 (2018). 「의료 인공지능」, 클라우드나인.
- 카이스트(KAIST) (2019), 「인공지능의 윤리/정책/사회 이슈」, KPC4IR, ISSUE PAPER, No. 08.
- Arnold, Z. and Toner, H. (2021), "AI Accidents: An Emerging Threat", CSET, CSET Policy Brief, July 2021.
- European Commission (2019), "Ethics Guidelines for Trustworthy AI", 2019.04.08.
- European Commission (2020), "White Paper On Artificial Intelligence – A European approach to excellence and trust", 2020.02.19.
- European Commission (2021), Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, 2021.04.21.
- HAI (2021), "Artificial Intelligence Index Report 2021", Stanford University, Human-Centered Artificial Intelligence.
- Leslie, D. (2019), "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector", The Alan Turing Institute.
- McDougall, J. R. (2019), "Artificial Intelligence – Confronting the Ethical Dilemmas", PICMET 2019 Conference, Keynote Speech.
- Mitchell, M. (2019), "Artificial Intelligence: A Guide for Thinking Humans", PICADOR.
- OECD (2019a), "Artificial Intelligence in Society", 2019.06.11.
- OECD (2019b), "Recommendation of the Council on Artificial Intelligence", 2019.05.22.
- Oxford Insights (2020), "Government AI Readiness Index 2020", Oxford Insights / International Development Research Centre (IDRC).

KISTI ISSUE BRIEF 제35호

발 행 일 2021. 11. 01.

발 행 인 김재수

편 집 위 원 조민수, 최희석, 이준, 정한민, 함재균,
이준영, 이상환, 정도범

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

I S S N 2635-5728