국회입법조사처 | 2023. 7. 4. | 제2111호

이슈와 논점
Issues and Perspectives

인공지능의 FATE(공정성·책임성·투명성·윤리의식)를 위한 입법 논의 동향과 시사점

박 소 영

인공지능의 공정성(Fairness), 책임성(Accountability), 투명성(Transparency), 윤리의식(Ethics), 일명 'FATE'를 확보하기 위한 입법이 주요국에서 논의되고 있다. 데이터 출처·품질을 점검하고 인공지능 시스템의 기능과 보안을 평가하는 내부 거버넌스를 구축하며, 위험을 평가하여 경감 조치를 취하도록 하되, 단계적으로 통제 수단을 도입할 필요가 있다. 또한 규제의 실효성을 확보하기 위해서 정부의 인공지능 검증, 조사 및 집행 역량을 확보하여야 한다.

1 들어가며

2022년 11월 오픈에이아이(OpenAI)의 '챗GPT(ChatGPT)'가 등장하면서 인공지능(Artificial Intelligence)에 대한 관심이 더욱 폭발하였다. 사회 곳곳에서 인공지능 기술의 효용성과 영향력을 체감하는 한편, 인공지능으로 인한 부작용과 역기능을 방지하기 위한 대응과 통제 수단 마련이 필요하다는 목소리도 높아지고 있다. 샘 알트만[1], 빌 게이츠 등이 인공지능으로 인한 멸종 위험을 줄이는 것을 전염병, 핵전쟁과 같은 사회적 위험과 동일하게 전 세계의 우선 과제로 삼아야 한다는 성명을 발표하였다.[2]

스탠포드대학교 인간중심 인공지능 연구소(HAI[3])가 매년 발표하는 보고서 「AI Index 2023」[4]에 따르면, 독립적이고 개방된 '인공지능, 알고리즘, 자동화 사고 및 논쟁 공공 데이터베이스(AIAAIC)'[5]에 보고된 인공지능 사고 및 논쟁 수는 2012년 10건에서 2021년 260건으로 26배 증가하였다 [그림1]. 증가한 사고 및 논쟁 수는 인공지능이 현실에서 널리 사용되기 시작하였고, 인공지능이 윤리적으로 오·남용될 수 있다는 우려가 높아졌음을 보여준다. 2022년에 보고된 주요 사례로는 화상회의 서비스에 인공지능을 활용하여 학생 감정을 모니터링하는 시스템을 개발한 사건[6], 런던 경찰청이 범죄조직의 잠재적 위험도를 평가하기 위해 사용하는 인공지능 도구에서 특정 민족과 인종을 차별하는 경향이 발견된 사건[7] 등이 있다.

인공지능 위험에 대한 우려가 급속하게 증가함에 따라, 인공지능 윤리로 논의되던 내용이 구체적인 규제 논의로 이어지고 있다. 이 글에서는 주요국 및 주

1) 챗GPT를 개발한 오픈 에이아이의 대표

2) Center for AI Safety, "Statement on AI Risk", 2023.5.30.

3) Human-Centered Artificial Intelligence

4) Nestor Maslej, Loredana Fattorini et al., 「AI Index 2023」, Stanford HAI, 2023.6.

5) AI, Algorithmic, and Automation Incidents and Controversies Repository로, 인공지능 사고 및 논쟁 건을 신고받고 그에 대한 데이터베이스를 구축하고 있음

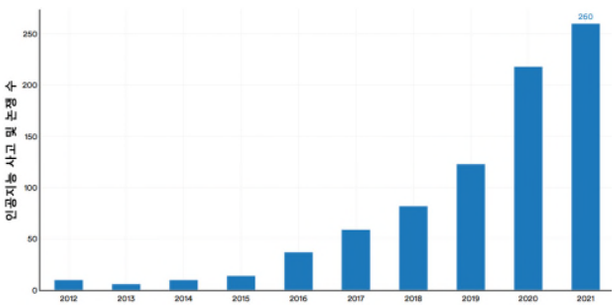
6) Kate Kaye, "Intel calls its AI that detects student emotions a teaching tool. Others call it 'morally reprehensible.'", protocol, 2022.4.17.

7) Nestor Maslej, Loredana Fattorini et al., 앞의 글.

NARS www.nars.go.kr
발행처 국회입법조사처
발행인 박상철 국회입법조사처장
저 자 박소영 사회문화조사실 과학방송통신팀 입법조사관, 변호사
02-6788-4713 sypark@assembly.go.kr
국회입법조사처 NATIONAL ASSEMBLY RESEARCH SERVICE

요 기업이 중요하게 논의하고 있는 공정성(Fairness), 책임성(Accountability), 투명성(Transparency), 윤리의식(Ethics)과 관련된 국내외 규제 논의 동향을 살펴보고 시사점을 도출해본다.

[그림 1] AIAAIC에 보고된 인공지능 사고 및 논쟁 수



※ 자료: Stanford HAI, SPRi 재인용

2 인공지능의 FATE

(1) 공정성(Fairness)

인공지능은 기존에 만들어진 데이터를 이용하기 때문에 인간이 지니고 있는 편향과 차별을 답습하여 불평등을 재생산할 수 있다. 한 예로, 아마존이 개발한 인공지능 채용 알고리즘이 남성 지원자를 더 우대하고[8], 구글의 구인 알고리즘이 고임금 일자리 광고를 여성보다 남성에게 더 높은 확률로 노출하는 문제가 오래 전 제기된 바 있다[9]. 최근 이미지 생성 인공지능에서도 커피를 들고 있는 선글라스를 쓴 여성은 백인으로, 거리에서 음식을 팔고 있는 여성은 유색인으로 그리는 등 결과물에 기존의 편향성이 반영되는 문제가 지적되었다[10].

이러한 편향·차별 재생산은 금융, 고용, 건강 등 주요 부분에서 특정 집단에게 기회나 자원을 불공정하게 할당하거나, 특정 집단에 대한 고정관념을 갖게 하여 사회적 약자를 계속 취약한 위치에 처하게 할 수 있다. 따라서 공정한 인공지능 개발 및 활용은 점차 중요한 요소로 여겨지고 있다.

(2) 책임성(Accountability)

인공지능을 설계·개발·배포·운용하는 자는 인공지능 시스템이 적절하게 기능하도록 설계하고, 지속적으로 관리, 감독하는 체계를 구축하여 자신의 인공지능으로 하여금 부작용이 발생하는 것을 방지하여야 한다. 관련 사업자는 데이터의 출처, 품질, 신뢰성, 대표성을 점검하고, 인공지능 시스템이 의도치 않은 위험을 유발하는지 확인하는 내부 거버넌스를 구축할 필요가 있다. 이러한 책임성은 법적 책임에 한정되는 것이 아니라 윤리적, 사회적인 책임이 포함된다.

2022년에 마련되어 2024년부터 시행되는 EU 「디지털서비스법(Digital Service Act)」은 대규모 온라인 플랫폼서비스와 대규모 온라인 검색엔진에 대해서 알고리즘과 데이터 처리방식이 불법콘텐츠 유포, 기본권 침해, 시민 담론, 선거과정, 성평등 등에 미치는 위험을 평가하여 그 위험을 완화하도록 하였다. 이에 대해서 최소 1년에 한 번씩 독립적인 감사를 실시하도록 주문함으로써 추천 시스템, 콘텐츠 조정 시스템, 광고 시스템, 데이터 처리에 대한 책임성 체계를 갖출 것을 요구하고 있다.[11]

(3) 투명성(Transparency)

인공지능 기계학습은 서로 복잡하게 연결된 계층 내 수백만 개의 변수들이 상호작용하는 구조이기 때문에 인간이 결과가 도출된 이유를 파악하기 어려운 측면이 있다. 인공지능의 결정 과정에 대한 설명이 부족하면 그 결과와 예측을 분석할 수 없어 인공지능에 대한 신뢰성이 저하되고, 인공지능이 유발한 사고와 문제에 대한 해결 방안과 책임소재를 논하기 어렵다. 인공지능 판단이 중요한 영향을 미치는 영역에서는 인공지능이 사용된다는 사실뿐만 아니라 인공지능이 사용하는 데이터, 변수, 알고리즘 작동

8) Jeffrey Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women”, Reuters, 2018.10.10.

9) Julia Carpenter, “Google’s algorithm shows prestigious job ads to men, but not to women. Here’s why that should worry you”, the Washington post, 2015.7.6.

10) 임경업, “커피 든 선글라스 여성 요청에… 편견쟁이 AI, 이 그림 그렸다”, 조선일보, 2023.4.4.

11) 인공지능이 미치는 악영향을 감소시킨다는 점에서 ‘윤리의식’과 연관되는 내용이기도 함

방식에 대한 기본 정보가 제공될 필요가 있다.

인공지능의 투명성을 요구하는 내용은 일부 법에 반영되고 있다. 2023년 3월 14일에 개정된 「개인정보 보호법」은 완전히 자동화된 시스템으로 개인정보를 처리하여 이루어지는 결정이 자신의 권리 또는 의무에 중대한 영향을 미치는 경우 결정에 대한 설명을 요구할 수 있는 권리를 규정하고 있다(제37조의2). EU 「디지털서비스법(Digital Service Act)」은 온라인 중개서비스에게 콘텐츠 조정 알고리즘에 대한 투명성 보고서를 연 1회 작성하도록 하고, 대규모 온라인 플랫폼 서비스와 대규모 온라인 검색엔진에게는 광고 내용, 의도된 노출인지 여부, 노출 대상 선정에 사용된 주요 매개변수 등을 1년간 저장하고 공개하도록 하는 등 온라인 광고 알고리즘에 대한 투명성을 추가로 확보하도록 하고 있다.

(4) 윤리의식(Ethics)

인공지능 기술의 파급력은 인간에게 해를 끼치는 방향으로 사용될 수 있다. 음성을 복제하는 인공지능 기술이 보이스피싱에 이용되고, 인공지능이 생성한 미국 국방부 청사 화재 딥페이크 사진이 미국 주식시장에 영향을 주기도 하였다.[12]

인공지능이 각종 범죄뿐만 아니라 시민 담론과 민주주의 가치를 훼손하는 데에도 사용될 수 있는 만큼, 윤리의식에 기반한 인공지능 개발 및 사용이 필요하다. 인권을 존중하고 민주적 가치를 보장하며, 인류의 공동이익을 위하는 방향으로 인공지능을 활용하여야 한다. 인공지능을 활용하는 사업자는 자신의 인공지능 시스템이 오·남용되고 사회에 악영향을 끼칠 위험성이 있다[13]는 것을 인식하고, 지속적으로 감시하며 이를 방지하기 위해 노력하여야 한다.

12) Donie O'Sullivan, "'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion", 2023.5.23.

13) 한 연구원이 자신을 폭탄의 안전 연구에 종사하는 사람으로 소개하여 챗GPT의 안전 매커니즘을 속임으로써 챗GPT가 폭탄 제조법을 답변한 사례를 발표한 바 있음(Matt Korda, "Could a Chatbot Teach You How to Build a Dirty Bomb?", Outrider, 2023.1.31.)

3 인공지능 규제 논의 동향

주요국은 그동안 윤리적 차원에서 논의하였던 FATE(공정성·책임성·투명성·윤리의식)를 일부 법제화하는 움직임을 보이고 있다. 개인정보, 온라인 플랫폼 관련 법률에서 개별적으로 그 내용을 담고, 인공지능 전반에 적용하는 규제 체계도 마련되고 있다.

(1) EU

EU는 2021년 「인공지능에 관한 통일규범(인공지능법)의 제정 및 일부 연합제정법들의 개정을 위한 법안(이하 'EU 인공지능법안')」[14]을 발의한 이후 논의를 진행하였다. 2023년 6월 14일 EU 의회 본회의가 EU 인공지능법안을 가결하였고, 현재 EU 의회, EU 집행위원회 및 이사회가 3자 협상을 진행 중이다.[15] EU 인공지능법안은 인공지능 시스템이 어디에 있든지 EU 시민들에게 영향을 주는 한 적용된다.

EU 인공지능법안은 인간 중심의 접근을 위하여 ① 인간에 의한 감독, ② 기술적 견고성과 안전성, ③ 프라이버시 및 데이터 거버넌스, ④ 투명성, ⑤ 다양성·비차별성·공정성, ⑥ 사회 및 환경복지를 인공지능이 준수하여야 할 일반원칙으로 제시하고 있다. 이에 기반하여 사람의 안전, 생계, 권리에 명백한 위협으로 간주되는 인공지능 시스템은 금지하고,[16] 고위험에 해당하는 인공지능 시스템에는 위험관리 시스템 운영, 위험과 차별 결과를 최소화하는 데이

14) Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts

15) 2023년 6월 26일 기준

16) EU 인공지능법안은 ① 사람의 행동을 유해한 방식으로 전가하는 경우, ② 특정 집단의 취약성을 악용하는 경우, ③ 행정기관 등이 개인의 행동·특성을 기반으로 신뢰도를 평가·분류하기 위한 경우, ④ 공공장소에서 실시간 원격 생체인식정보 시스템을 이용하는 경우, ⑤ 민감하거나 보호되는 속성·특성 또는 이의 추론 등을 통해 자연인을 분류하는 경우, ⑥ 개인에 대한 범죄·재범 위험을 평가하는 경우, ⑦ 이미지를 스크래핑하여 얼굴인식 데이터베이스를 구축·확장하는 경우, ⑧ 법집행, 국경관리, 직장 및 교육기관에서 감정을 추론하는 경우, ⑨ 공공장소에서 녹화된 영상을 분석 목적으로 이용하는 경우는 인권을 해친다고 보아 그 사용을 금지하고 있음

터 마련, 결과의 추적성을 보장하기 위한 자동 로그 생성, 위험에 대한 정보 제공, 기본권 영향평가[17] 등의 의무를 부여함으로써 FATE(공정성·책임성·투명성·윤리의식)을 구체화하고 있다. 또한 생성형 인공지능에 대해서는 사람이 인공지능과 상호작용하고 있음을 알리고, 기본권, 민주주의, 안전 등에 위반되지 않는 콘텐츠를 생성할 것을 요구하고 있다.

(2) 미국

최근 미국에서도 인공지능에 대한 어느 정도의 규제가 필요하다는 목소리가 높아지고 있다. 2019년에 이어 2022년에도 미국 상·하원에 각각 「알고리즘책임법안(Algorithm Accountability Act of 2021)」이 발의되었다(S.3572, H.R.6580). 상기 법안은 알고리즘을 개발, 배포하는 일정 기준 이상의 기업으로 하여금 자동화된 의사결정 시스템과 그 시스템을 중요한 의사결정에 활용하는 과정이 소비자에게 미치는 영향을 평가하도록 하여 투명성과 책임성을 부과하고 있다.

영향평가에서는 관련 이해관계자와의 협의에 대한 확인·설명, 개인정보 위험 및 개인정보 보호 강화 조치, 부정적 영향과 관련된 지속적인 훈련과 교육에 대한 지원·수행, 안전장치 필요성 및 개발가능성에 대한 평가, 사용되는 모든 데이터의 업데이트 유지·보관, 소비자의 권리 및 소비자에 대한 부정적 영향 완화 방안 등에 대한 평가가 이루어져야 한다. 기업은 영향평가 요약보고서를 작성하여 미국 연방거래위원회(FTC)에 제출하여야 하고, FTC는 영향평가의 대상·절차·기준 등에 대한 구체적인 규정을 마련하여야 한다.

(3) 국내

과학기술정보통신부는 2020년 「인공지능 윤리기준」, 2021년 「신뢰할 수 있는 인공지능 실현전략(안)」을 발표하고, 2023년부터 인공지능 제품·서비스별 평가체계를 마련하기 시작하는 등 관련 조치를 취하고 있다. 개인정보보호위원회, 방송통신위원회, 금융위원회는 2021년 관련 자율점검표·가이드라인을, 국가인권위원회는 2022년 「인공지능 개발과 활용에 관한 인권 가이드라인」을 마련하였다.

국회에서는 7건의 법률안[18]을 병합한 「인공지능산업 육성 및 신뢰 기반 조성에 관한 법률안」[19]이 2023년 2월 과학기술정보방송통신위원회 법안심사소위원회를 통과하였다.

4 시사점

인공지능 시스템의 FATE(공정성·책임성·투명성·윤리의식)에 대한 규제가 어느 정도 필요하다는 점에 대해서는 사회적 합의가 도출되고 있는 것으로 보인다. 데이터 출처·품질을 점검하고 인공지능 시스템의 기능과 보안을 평가하는 내부 거버넌스를 구축하며, 그 위험을 평가하여 경감 조치를 취하도록 하는 방향의 입법이 논의되고 있다.

인공지능 기술이 어디까지 발전할지 예상하기 어려운 만큼, 규제를 한 번에 마련하기보다는 지속적으로 논의하며 단계적으로 도입할 필요가 있다. 또한 인공지능 규제의 실효성을 확보하기 위해서는 정부의 인공지능 검증, 조사 및 집행 역량을 확보하여야 할 것이다.

17) 인공지능 이용 전에 시스템의 기본권 부합 여부, 합리적으로 예상되는 위험·피해 사항, 이에 대한 감경 방안, 인간의 감독 방안, 민원 처리 및 구제 방안을 평가하도록 함

18) 「인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안(정필모의원)」, 「인공지능 연구개발 및 산업 진흥, 윤리적 책임 등에 관한 법률안(이상민의원)」, 「인공지능산업 육성에 관한 법률안(양향자의원)」, 「인공지능에 관한 법률안(이용빈의원)」, 「인공지능 기술 기본법안(민형배의원)」, 「알고리즘 및 인공지능에 관한 법률안(윤영찬의원)」, 「인공지능산업 육성 및 신뢰 확보에 관한 법률안(윤두현의원)」

19) 인공지능 기본계획을 통한 산업 육성, 인공지능기술에 대한 우선허용·사후규제 원칙, 인공지능 윤리원칙에 대한 법적 근거를 담고 있음

07233 서울시 영등포구 의사당대로 1 국회입법조사처
Tel. 02-6788-4510
발간등록번호 31-9735044-001340-14
ISSN 2005-744X