알고리즘 통제: 투명성의 구체적 기준

손도일
법무법인 율촌 파트너
변호사

김명훈
법무법인 율촌 파트너
변리사/외국변호사

Ⅰ. 들어가는 말: 알고리즘 통제의 필요성과 투명성 요건의 중요성

우리는 알게 모르게 이미 인공지능이 추천하는 결과물에 의하여 선택이 제한되고 있다. 쇼핑을 할 때도, 영화를 보거나 재미있는 동영상을 볼 때에도 모두 플랫폼에서 제공하는 인공지능의 추천에 의존하는 경향이 점점 높아지고 있다. 한편, 우리가 잘 알지 못하는 사이에 인공지능의 결정에 따라 우리의 이해관계가 달라질 수 있는 가능성도 높아지고 있다. 가령 상당수의 기업들이 수많은 지원자들 중에서 인공지능을 통하여 1차적으로 선별을 하고 있고, 금융과 의료 영역에도 인공지능의 도입이 활성화되고 있다.

이와 같은 흐름에 대하여 당연히 많은 우려들이 있고, 그 반작용으로 많은 가이드라인과 입법(안)들이 발표되고 있다. 이와 같은 우려들은 인공지능의 의사결정이 블랙박스(black-box) 구조를 가지고 있기 때문에 특히 더 발생하는 것으로 보인다. 즉, 알고리즘의 오작동이 발생한 경우 이와 같은 문제가 모델 설계의 오류인지 혹은 부정확한 데이터 활용에 의한 오류인지를 구분하는 것이 매우 어렵다.[1] 또한 인공지능 기술이 인간에 의하여 '의도적으로' 악용될 가능성도 존재한다. 가령 특정한 의도를 갖고 편향된 데이터를 AI에 주입할 수도 있고, 아니면 알고리즘 설계 자체가 편향된 결과를 도출하도록 되어 있을 수도 있다. 따라서 인공지능의 실패에 따른 부작용을 최소화하려면, 알고리즘을 어떠한 형식으로든 통제할 필요성이 있다는 점에 대하여는 많은 공감대가 이루어지고 있는 것으로 보인다.

다만, 현실적으로 알고리즘을 전면적으로 통제하는 것에 대하여는 많은 논란이 있다. 그러나 아래에서 보는 바와 같이 적어도 설명할 수 있는 인공지능(XAI) 내지 "투명성"은 거의 모든 가이드라인과 입법례에서 형태는 다를지 몰라도 공통적으로 추구하는 통제규범이라고 할 수 있다. 이하에서는 해외 및 국내의 알고리즘 통제의 입법례에 관하여 살펴 본 후, 알고리즘의 투명성 통제 기준을 좀 더 구체적으로 논의하고자 한다.

1) 이제영/김단비/양희태, "인공지능 기술 전망과 혁신정책 방향(2차년도), 안전하고 윤리적인 인공지능 R&D 및 활용을 위한 제도 개선을 중심으로", 과학기술정책연구원, 2019.12., 6-7면

Ⅱ. 해외 및 국내의 알고리즘 통제에 관한 가이드라인/입법례 개관

알고리즘 통제(인공지능 통제와 사실상 같은 의미로 볼 수 있다)에 관하여 가장 적극적인 행동을 취하고 곳은 유럽연합(EU)이라고 할 수 있다. 이하에서는 먼저 EU의 각종 사례를 살펴보고, 이후 미국 및 국내의 사례를 살펴보고자 한다.

1. EU의 사례

EU 인공지능 윤리 가이드라인(2019년 4월)에 의하면 '신뢰 가능한 AI(trustworthy AI)'를 구현하기 위한 요건으로 인간 자율성 및 감독(human

agency and oversight), 기술적 견고함 및 안전(technical robustness and safety), 프라이버시와 데이터 거버넌스(privacy and data governance), 투명성(transparency), 다양성, 차별금지, 공평성(diversity, non-discrimination and fairness), 사회 및 환경 복지(societal and environmental wellbeing), 책임성(accountability) 원칙을 열거하고 있다.

EU 인공지능 백서[2] [2020년 2월]에서는 인간중심, 공정성, 투명성 등 EU의 가치를 지키고 위험을 통제할 수 있어야 함을 요구하고 있고, 특히 고위험 인공지능 시스템은 투명하고 추적 가능하며 통제 가능하여야 하고, 적정한 학습데이터, 알고리즘의 프로그래밍과 학습용 데이터에 대한 기록 보관 또는 데이터 자체의 보관, 고위험 AI 시스템 사용에 대해 선제적 방식으로 적절한 정보를 제공할 것, 시스템이 발생할 수 있는 위험성을 사전에 적절하게 고려한 시스템 개발, 인간의 적절한 개입, 생체 인식 데이터는 극히 예외적인 경우 외에는 원격 식별을 위한 처리 금지를 요구하고 있다.[3]

가이드라인이 아니라 실제적으로 입법이 된 것 중 가장 눈에 띄는 해외 입법례는 유럽의 일반정보보호법(General Data Protection Regulation, 이하 "GDPR")이다. GDPR에서는 개인정보주체에게 자동화된 의사결정과 관련한 다양한 권리를 인정하고 있다. GDPR 제13조와 제14조에 따르면 개인정보가 정보주체 혹은 그 외로부터 입수된 경우, 개인정보처리자(controller)는 정보주체에게 다른 정보와 함께 프로파일링을 포함한 자동화된 의사결정 프로세스가 존재하는지 여부 및 그 영향에 관하여 통보하도록 되어 있다.[4] [5] 제15조에서는 정보주체에게 자신의 개인정보가 자동화된 의사결정에 따라 처리되는지 여부에 대한 확인을 구할 권리를 부여하고 있으며,[6] 제22조에서는 자신에 대하여 중대한 영향을 주는 결정이 자동화된 의사결정만 의존하는 것에 대하여 이의를 제기할 권한을 부여한 바 있다.[7]

EU 집행위원회는 2021. 4. 21. 인공지능에 관한 통일규범(인공지능법)의 제정 및 일부 연합제정법들의 개정을 위한 법안(인공지능법안)(Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts)을 공개하며 AI에 대한 법적 규제를 예고했다. 이 규제는 인간의 안전과 생계에 위협을 끼칠 수 있는 모든 AI를 포괄적으로 금지하고, 기본권을 해칠 수 있는 AI는 서비스 출시 전 평가 등 의무 사항을 지켜야 한다는 것이 핵심이다. 즉, 사람의 잠재의식을 이용하거나, 취약 연령·신체적 또는 정신적 장애 등 특정 계층에 피해가 갈 수 있는 AI 서비스를 원천적으로 금지했다. 신용등급이나 기타 집단 신뢰도에 영향을 미칠 수 있는 사회적 점수(social credit)에 대한 AI 활용도 폭넓게 막겠다는 방침이다. 또한 얼굴을 그대로 복사하는 딥페이크 같은 데이터 활용 콘텐츠는 물론 인간 감정 인지 등 AI의 실시간 생체 데이터 수집도 금지했다. 공개적으로 접근 가능한 원격 인식 시스템은 법 집행기관의 적절한 판단을 받을 것을 요구했다. 특히 고용과 사법 처리에 이르기까지 광범위한 고위험 AI를 규정하면서, 기업은

2) EU, "WHITE PAPER: On Artificial Intelligence - A European approach to excellence and trust", 2020.2.

3) 한국정보화진흥원, "EU 인공지능 백서와 데이터 전략", 2020.5.8.

4) 2. In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: (f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

5) Where personal data have not been obtained from the data subject, the controller shall provide the data subject with the following information:
2. In addition to the information referred to in paragraph 1, the controller shall provide the data subject with the following information necessary to ensure fair and transparent processing in respect of the data subject:
(g) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

6) 1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: (h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

7) 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her

고위험 AI에 대한 수칙 위반 정도에 따라 글로벌 연 매출의 6%를 벌금으로 부과받을 수 있도록 하고 있다.[8]

2. 미국의 사례

미국의 대표적인 알고리즘 통제 방안은 2019. 4. 알고리즘 책임 법안(Algorithmic Accountability Act)[9]이라고 할 수 있다.[10] 본 법안은 AI 시스템에 사용되는 알고리즘의 편향성과 차별적 결과를 방지하기 위하여 미국 연방거래위원회(Federal Trade Commission, FTC)에 알고리즘 관련 권한을 부여하고, 개인정보를 사용, 저장 또는 공유하는 기업에게 자동화된 의사결정 시스템 영향평가와 데이터 보호 영향평가를 수행하도록 요구할 것을 요구하고 있다.[11] '자동화된 의사결정 시스템 영향평가'는 시스템 자체와 그 개발 과정(시스템 디자인과 트레이닝 데이터 포함)을 대상으로 이루어져야 하고, 시스템의 정확성, 공평성, 편향성, 차별성, 프라이버시, 보안에 대한 영향평가를 하도록 되어 있다.

FTC는 2020. 4. "Using Artificial Intelligence and Algorithm"(인공지능 및 알고리즘 사용지침)을 배포하였다. 이에 의하면 FTC는 기업이 AI를 사용하는 과정에서 소비자에게 발생할 수 있는 위험을 어떻게 관리할 것인지에 대한 방향을 제시하고 있다. 이에 따르면, (1) 투명성(소비자기만금지, 민감한 데이터 수집시 투명성 담보, 불리한 조치에 대한 통지), (2) 설명가능성(결정에 대한 구체적 이유, 영향을 미친 상위 주요요인 공개, 거래조건 변경 시 통지), (3) 결과의 공정성(특정 집단에 대한 차별금지, 입력값의 차별적 요소와 별개로 도출되는 결과에도 차별이 발생하지 않도록 관리, 정보접근권한 및 수정기회를 소비자에게 제공), (4) 데이터와 모델의 견고성 및 실증적 타당성(정보의 정확성과 최신성, 명문화된 정책과 절차, AI모형의 유효성 검사, (5) 법령준수, 윤리, 공정성 및 비차별성에 대한 책임 견지(자기 점검을 통산 편견, 피해 방지, 무단 사용에 대한 알고리즘 보호, 책임메커니즘 구축: 객관적인 시각에서 스스로 개발한 AI를 평가하는 체계구축)를 요구하고 있다.[12]

3. 국내의 사례

과학기술정보통신부는, 2020. 11. 27. 국가 인공지능 윤리기준"(안)을 공개하였는데, 이에 의하면, 인공지능 개발과 활용 전 과정에서 ①인권 보장, ②프라이버시 보호, ③다양성 존중, ④침해금지, ⑤공공성, ⑥연대성, ⑦데이터 관리, ⑧책임성, ⑨안전성, ⑩투명성[13]의 요건이 충족되어야 한다고 규정하고 있다.

국토교통부는 2020. 12. 15. 자율주행차의 윤리·보안·안전에 관한 가이드라인을 발표하였는데, ①자율주행차는 문제 발생 시 책임을 확인할 수 있는 기록 시스템을 갖추어야 함(투명성), ②자율주행차는 인간의 안전을 최우선적으로 보호하도록 설계·제작·관리되어야 함(안전성), 자율주행차는 개인정보 등의 보안 체계를 갖추어야 함(보안성), ③문제가 발생한 경우 관련 주체는 해당 문제에

8) 한국경제신문, EU "비도덕적 AI 내놓는 기업, 매출 6% 벌금", 2021. 4. 23.

9) THE WHITE HOUSE, Artificial Intelligence for the American People https://www.whitehouse.gov/ai/

10) 동 법안은 2021년 새로운 의회의 구성으로 일단은 폐기된 것으로 볼 수 있지만, 향후에도 유사한 내용으로 다시 발의가 될 수 있다.

11) Algorithmic Accountability Act, H.R. 2231, 116th Cong. (2019-2020)https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info

12) 양기문, "기업의 AI 및 알고리즘 이용에 관한 지침", 정보통신정책연구 제32권 제4호, 2020. 4., 51-56면

13) 사회적 신뢰 형성을 위해 타 원칙과의 상충관계를 고려하여 인공지능 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다. 인공지능기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

대한 책임을 져야 함(책임성)과 같은 원칙을 제시하고 있다.

행정안전부 역시 2020. 2. Privacy by Design 개념을 적용하여 개인정보를 처리하도록 하는 자동처리 되는 개인정보 보호 가이드라인을 발표하였다. 이 가이드라인에 의하면, 기획단계부터 설계 및 운용단계에서 서비스에 꼭 필요한 개인정보만을 최소한으로 처리하고, 그 내용을 투명하게 공개하도록 하고 있다.

지능정보화기본법에 따르면, 과학기술정보통신부장관은 지능정보기술의 안정성·신뢰성·상호운영성을 확보하기 위한 기술기준을 고시할 수 있도록 하고 있고, 군사/의료 혹은 오작동시 중대한 위해가 초래될 수 있는 지능정보기술에 대하여는 해당 고시를 준수할 의무를 부과하고 있다(동법 제21조)[14]. 또한 생활파급력이 큰 지능정보서비스에 대하여는, 안정성 및 신뢰성과 정보보호에 미치는 영향 등에 관한 평가를 할 수 있도록 되어 있다(동법 제56조). 특히 과학기술정보통신부장관은 지능정보서비스의 안정성을 확보하기 위한 최소한도의 보호조치의 내용과 방법을 정하여 고시할 수 있도록 하고 있다(동법 제60조).

4. 소결론

이상에서 본 여러 가지 가이드라인과 법(안)들을 종합하면, 알고리즘 통제가 어느 정도는 필요하다는 것에는 대체적으로 합의가 있는 것으로 보이지만, 그 정도에 있어서는 다양한 의견이 있다. 규제기구에 의한 모든 알고리즘 의사결정을 감독하여야 한다는 의견, 알고리즘의 투명성과 설명가능성 정도를 의무화하자는 의견, 일반화된 수준의 감독 기준 정도만 제시하자는 의견, 별도의 규율은 불필요하다는 의견이 있으나,[15] 필자는 인공지능의 발전 필요성과 인간의 보호를 조화할 수 있도록 최소한 알고리즘의 투명성을 의무화하는 것이 가장 현실적인 방안이라고 생각한다. 다만, 알고리즘의 사용영역에 따라 정도의 차이는 있을 것이다.

Ⅲ. 투명성 통제의 구체적 기준

앞서 본 많은 사례에서 알고리즘(인공지능)에 대하여 투명성을 요구하고 있으나, 실제로 그 투명성의 개념과 범위에 관하여는 다양한 시각이 존재하고, 구체적으로 통일된 기준은 없는 것으로 보인다. 지금까지의 논의를 살펴보면 투명성의 기준에 관하여 다음과 같은 정도의 공통된 기준을 제시할 수 있을 것으로 생각한다.

1. 투명성 통제의 방향: 총론

알고리즘의 투명성 확보는 정보공개의무 및 설명의무를 통해서 가능할 것이다. 구체적으로는, 알고리즘 투명성에 대한 가이드라인을 제시한 후 그에 따라 개발자가 AI 알고리즘의 작동원리의 기본적인 부분을 공개하고, 설명하도록

14) 동법 시행령 제16조(기술기준 등) ①과학기술정보통신부장관은 법 제21조제1항에 따라 기술기준을 정한 경우에는 지체 없이 해당 기준을 관보 및 인터넷 홈페이지에 공고해야 한다.
②법 제21조제2항에서 "대통령령으로 정하는 국민의 생명 또는 신체안전 등에 밀접한 지능정보기술"이란 지능정보기술을 이용하는 사람의 생명·신체를 보호하는데 현저한 지장을 줄 우려가 있는 지능정보기술로서 다음 각 호의 어느 하나에 해당하는 지능정보기술을 말한다.
1. 군사적 목적으로 개발·관리·활용하려는 지능정보기술
2. 「의료법」 제24조의2제1항에 따른 수술등의 의료행위에 직접 이용되어 사람의 신체에 영향을 미칠 수 있는 지능정보기술
3. 지능정보기술이 오작동(誤作動)될 경우 사람에게 중대한 위해(危害)를 끼칠 우려가 있는 지능정보기술

15) 김윤명, "알고리즘과 법", 한국정보화진흥원, 2019, 34면

16) 이상용, "알고리즘 규제를 위한 지도", 경제규제와 법 제13권 제2호, 2020. 11. 139면

17) 이상용, 앞의 논문, 150면

18) 이제희, "알고리즘의 취급에 대한 법적 논의", 공법학연구, 2018. 8., 327-328면

19) 양종모, "인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안", 법조, 2017. 6., 87-88면

하는 것이 필요할 것이다. 물론, 알고리즘이나 소프트웨어가 갖는 영업비밀성 내지 지식재산권에 대한 침해 가능성이라는 측면이 있기는 하지만, 알고리즘 소스코드나 로직을 어느 정도까지 공개하거나 설명하도록 할지(내용적인 측면), 그리고 알고리즘을 공개하거나 설명하는 과정에서 비밀유지가 가능하도록 하는 장치의 도입(절차적인 측면) 등을 통해서 위와 같은 문제는 보완이 가능할 것으로 보인다. 특히 금융, 의료, 자율주행자동차와 같이 위험성이 큰 곳에 대하여는 좀더 정치한 규제가 필요할 것이다.

투명성 통제를 하는 경우에도 그 방향을 포지티브 방향으로 할 것인지(허용되는 행위만을 열거할 것인지), 아니면 네거티브 방향(허용되지 않는 행위만을 열거할 것인지)으로 할 것인지에 대하여 논의가 있을 수 있다. 그러나 현재의 감독 기술 수준과 빠르게 진전되는 인공지능 기술의 발전, 알고리즘 개발의 특성 등을 고려할 때, 네거티브 규제를 원칙으로 하되 일부 핵심적인 행위원칙에 대하여는 별도로 규정하는 것이 필요할 것이다. 행정규제기본법 제5조의2에서도 신기술 서비스 및 제품에 대하여는 네거티브 규제를 원칙으로 하고 있다.[16]

알고리즘의 실패에 대한 궁극적인 책임의 근거를 물을 수 있는 주체를 선별할 수 있도록 하는 것이 투명성 통제의 핵심이라고 할 수 있다. 즉 특정한 알고리즘의 실패로 인하여 제3자에게 손해가 발생한 경우, 알고리즘을 개발한 자와 이를 수정한 자, 특정한 데이터를 통하여 알고리즘을 학습시킨 자와 이와 같은 알고리즘을 사용한 자들 사이에서 책임을 분명하게 물을 수 있도록 하는 것이 필요할 것이고, 이때 필요한 정보를 각 이해관계자들이 투명하게 제공하도록 하는 제도가 필요한 것이다.[17] 실제로는 관련 손해배상소송에서 알고리즘의 제작자에게 우선적으로 결함이 없었다는 점에 관한 입증책임을 부과하는 것이 현실적인 방법이다.[18] 다만, 이와 같이 할 경우 알고리즘 개발에 대한 과도한 위험부담이 인공지능 발전에 장애가 될 수 있다는 우려도 있다.

2. 알고리즘의 설계·제작·수정의 투명성

알고리즘은 일단 기술적으로 사회적으로 매우 안정적이고 견고하게 만들어져야 하고, 알고리즘은 윤리적으로 설계되어야 한다(ethical by design)는 점에 관하여는 대부분의 견해가 일치하고 있다.[19] 문제는 알고리즘은 매우 복잡한 설계-제작-수정 과정을 거치게 되므로, 문제가 발생한 경우에 어느 단계가 문제인지를 명확하게 파악하는 것이 필요하다. 따라서 알고리즘의 설계와 제작과 수정 과정에서는 그 기록을 하도록 하는 조치가 필요할 것으로 생각한다. 이와 관련하여 전자금융감독규정 제14조에서 정보처리시스템 보호대책으로 규정하는 각종 조치들이 알고리즘의 설계·제작·수정에도 상당부분 적용될 수 있을 것으로 보인다. 이를 구체적으로 알고리즘에 적용하면 다음과 같다.

1. 알고리즘에 구동, 조작방법, 명령어 사용법, 운용순서, 장애조치 및 연락처 등 시스템 운영매뉴얼을 작성할 것

2. 알고리즘, 데이터베이스관리시스템(Database Management System:DBMS)·운영체제·웹프로그램 등 주요 프로그램에 대하여 정기적으로 유지보수를 실시하고, 작업일, 작업내용, 작업결과 등을 기록한 유지보수관리대장을 작성·보관할 것

3. 알고리즘의 장애발생 시 장애일시, 장애내용 및 조치사항 등을 기록한 장애상황기록부를 상세하게 작성·보관할 것

4. 알고리즘의 정상작동여부 확인을 위하여 시스템 자원 상태의 감시, 경고 및 제어가 가능한 모니터링시스템을 갖출 것

5. 알고리즘의 통합, 전환 및 재개발 시 장애 등으로 인하여 인공지능 시스템의 운영에 지장이 초래되지 않도록 통제 절차를 마련하여 준수할 것

6. 알고리즘 책임자를 지정·운영할 것

7. 알고리즘의 운영체계, 시스템 유틸리티 등의 긴급하고 중요한 보정(patch) 사항에 대하여는 즉시 보정 작업을 할 것

8. 중요도에 따라 알고리즘의 운영체제 및 설정내용 등을 정기 백업 및 원격 안전지역에 소산하고 백업자료는 1년 이상 기록·관리할 것

9. 알고리즘의 운영체제(Operating System) 계정으로 로그인(Log in)할 경우 계정 및 비밀번호 이외에 별도의 추가인증 절차를 의무적으로 시행할 것

10. 알고리즘에 대한 사용권한, 접근 기록, 작업 내역 등에 대한 상시 모니터링체계를 수립하고, 이상 징후 발생 시 필요한 통제 조치를 즉시 시행할 것

알고리즘의 제작 과정에서는 제작자가 사용자가 모르는 상태에서 알고리즘의 사용에 영향을 주거나, 사용과정에서의 데이터에 접근할 수 있도록 하는 비밀기능을 개발하여 설치하는 것은 금지되어야 한다.

위급상황에서는 알고리즘의 제작자에게 필요한 데이터를 요청할 수 있도록 하는 것도 필요할 것이다. 알고리즘은 여러 형태로 배포되고, 그 이후 알고리즘의 사용자들의 필요에 따라서 수정되어 사용될 수 있다. 만일 알고리즘을 사용한 인공지능이 문제를 일으키는 경우 사용자들을 이를 긴급하게 수정하여야 할 필요성이 있다. 이 경우 사용자들을 알고리즘의 설계·제작자들에게 필요한 최소한의 데이터를 요청할 수 있는 권리가 있어야 한다.

알고리즘 제작자는 그 제작 및 수정과정에서 알고리즘의 문제를 발견한 경우 이를 적극적으로 공유하고, 그 해결방안을 배포하도록 어느 정도는 의무화하는 것이 필요할 것이다. 특히 알고리즘이 고위험 AI의 영역에서 사용되는 경우에는 특히 이와 같은 조치가 필요할 것이다(일종의 리콜이라고 볼 수 있다). 만일 이와 같은 의무를 게을리할 경우 제조물 책임 등 일반적인 민·형사적인 책임을 평가받음에 있어서 부정적인 고려요소가 될 것이다.

알고리즘은 데이터를 바탕으로 결과를 산출하는 것이므로 알고리즘에 결함이 있으면 당연히 알고리즘은 결함된 결과치를 산출하게 될 것이다. 알고리즘의 결과에 결함이 있는 경우, 알고리즘의 결함에 기인한 것일 수도 있고, 입력된 데이터의 편향성에 의한 것일 수도 있다. 이와 같은 알고리즘의 결함을 분석하

기 알고리즘 설계 단계에서부터 이와 같은 분석·평가를 염두에 두고 알고리즘을 개개의 모듈로 나누어 주석을 달도록 한 후 이를 분석하는 방안이 있다. 이를 정적 분석(static analysis)라고 한다. 한편, 알고리즘이 복잡한 경우에는 정적 분석을 통하여 그 결함을 찾는 것에는 한계가 있을 수밖에 없다. 이에 따라 프로그램을 작동시킨 후 그 작동상황을 분석하는 동적 분석(dynamic analysis)가 있다. 물론 실제 입력된 데이터와 동적 분석 시에 입력되는 데이터에 차이가 있을 수 밖에 없으므로 이도 완벽한 분석이라고 보기는 어렵다는 한계가 있다.[20)]

고도의 위험도를 가진 알고리즘의 경우에는 보다 적극적인 방식으로 관련 정보를 제공하도록 하는 것이 필요할 것이다. 즉 알고리즘의 능력이나 한계, 목적, 작동의 조건이나 정확성의 예상 수준에 관하여 보다 정확한 정보가 적극적으로 제공될 필요가 있다.[21)]

3. 알고리즘 활용과 관련한 투명성

우리 일상에서 인공지능은 이제 광범위하게 사용되고 있고, 그 결정에 따라 많은 이해관계자들에 대한 중대한 영향을 줄 수 있다. 따라서 이해관계자들에 대하여는 알고리즘이 사용된다는 점과 알고리즘의 결정 기준에 대한 기본적인 정보는 제공되어야 한다. 다만 그 정보제공의 범위는 알고리즘이 상용되는 영역에 따라 달라질 수 있을 것이다.

알고리즘이 가장 많이 우리 주변에서 사용되는 것은 소위 "추천" 기능을 통한 것이다. 일반 소비자들의 경우, 어떤 기준에 의하여 추천되는 것인지를 알 필요가 있지만, 더 중요한 것은 플랫폼 사업자들이 플랫폼 이용사업자(즉, 플랫폼 입점업체)를 추천하는 기준일 것이다. 현재 입법이 추진되고 있는 온라인플랫폼 중개거래의 공정화에 관한 법률("온라인플랫폼 공정화법")에서도 논의가 되고 있고, 온라인 플랫폼 중개사업자는 온라인 플랫폼 이용사업자와 중개거래계약을 체결한 때에는 지체 없이 주요 거래조건 및 분쟁예방을 위한 사항이 명시된 계약서를 서면으로 교부하여야 하는데, 계약서에 포함될 "주요 거래조건"에는 거래되는 재화 또는 용역이 온라인 플랫폼에 노출되는 순서, 형태 및 기준 등에 관한 사항 등을 규정하고 있다(제6조). 한편, 이 법과 동시에 논의가 되고 있는 온라인플랫폼 이용자 보호에 관한 법률(이하 "온라인플랫폼 이용자 보호법")에 의하면, "대규모 온라인 플랫폼 사업자"는 검색결과, 추천 등을 결정하는 요소 등 콘텐츠 등의 노출 방식 및 노출 순서를 결정하는 기준을 공개하여야 하고, 이러한 기준 중 개인화된 기준의 적용 여부 및 내용과 범위 등을 이용자가 선택할 수 있도록 되어 있다(제11조). 이외에도 EU의 Digital Service Act 안에 따르면, 온라인 광고에 대하여 ① 해당정보가 광고라는 사실, ② 광고주, ③ 해당 광고가 노출되는 이용자를 결정하는데 사용된 주요 매개변수에 대한 의미 있는(meaningful) 정보를 명확히 실시간으로 확인할 수 있도록 해야 한다(제24조).

현재의 입법안들이 주로 플랫폼 사업자들과 관련한 것이라면, 이외에도

20) 양종모, 앞의 논문, 87-91면

21) 정남철/계인국/김재선, "미래세대 보호를 위한 법적과제4 - 인공지능에 대한 유럽연합의 규제체계와 대응전략을 중심으로", 글로벌법제연구, 한국법제연구원, 2020, 73-74면

22) 양기문, 앞의 논문, 52면

23) 이는 앞서 본 GDPR의 자동화된 평가에 대한 대응권리와 유사한 개념이다.

24) 정남철/계인국/김재선, 앞의 논문, 73-74면

의료, 금융과 같이 개개인의 생명과 재산에 중대한 영향을 받을 수 있는 상황에서 사용되는 알고리즘에 대하여는 이보다는 좀 더 엄격한 투명성 기준이 필요할 것이다. 즉, 알고리즘에 의한 결정에 의하여 불리한 조치가 취해지는 경우에는, 그와 같은 조치가 알고리즘에 의한 것이라는 점을 알려줌과 동시에 소비자에게는 그 판단의 기초가 된 자신의 정보를 확인하고 이를 수정하여 재검토를 요구하거나[22)] 인간의 개입을 통한 평가를 요구할 권한이 부여되어야 한다.[23)] 가령, 인공지능에 따른 금융상품 추천서비스 혹은 로보어드바이저에 의한 주식매매의 경우 어떤 기준으로 추천을 하는지에 관하여 미리 알려줄 필요가 있고, 만일 손실이 난 경우에 투자자가 로보어드바이저 기업을 상대로 소송을 하는 경우(알고리즘의 실패로 인하여 손실이 났다고 주장하는 경우), 로보어드바이저 사업자는 그 알고리즘의 타당성에 대하여 입증책임을 부담한다고 보아야 할 것이다.

4. 데이터 관리의 투명성

알고리즘과 함께 인공지능을 이루는 다른 한 축은 데이터라고 할 수 있다. 어떤 데이터를 알고리즘에 입력하는지에 따라서 알고리즘이 도출하는 결과는 매우 다를 수 있다. 알고리즘에 입력된 데이터를 어떻게 투명하게 관리할 것인지는 알고리즘의 투명성 통제에 있어서 핵심적인 요소라고 할 수 있다.

머신러닝에 사용된 모든 데이터를 보관하라고 하는 것은 기업에 과도한 부담을 줄 뿐만 아니라, 그 데이터에 보관된 개인정보와 민감한 다른 정보를 고려할 때 바람직한 안이라고 볼 수 없다. 대신, 최소한 머신러닝에 사용된 데이터들의 기본적인 정보 및 샘플 데이터를 보관하도록 하는 방법을 고려할 수 있을 것이다. 다만, 고위험 AI 영역에서는 항상 위험 발생 가능성이 있으므로, 좀 더 상세한 정보를 장기간 보관할 필요가 있을 것이다(가령 의료진단 알고리즘의 잘못된 진단결과에 따른 손해배상 소송이 제기된 경우, 적어도 어떠한 종류의 데이터가 해당 알고리즘의 학습에 사용된 것인지는 이를 제작한 피고가 밝혀야 할 책임이 있다고 볼 수 있다).[24)]

알고리즘 학습을 위하여 이용자의 개인정보를 수집하고 활용하는 경우에는, 이용자에게 고지를 하고, 필요한 경우 동의를 받도록 하는 것이 필요할 것이다. 이는 미국 FTC의 인공지능 및 알고리즘 사용지침에서도 투명성과 관련하여 요구하고 있을 뿐만 아니라, 우리나라의 개인정보보호법에서도 요구되는 요건이다. 한편, 실명정보가 아닌 가명정보의 경우에는 동의를 받지 않고 통계작성, 과학적 연구 등 제한적인 범위 내에서 활용할 수 있다(개인정보보호법 제28조의2 참조).

5. 인공지능 기술 분류에 따른 투명성

최근 발표된 EU 인공지능 윤리 가이드라인, EU 인공지능 백서, 미국의 알고리즘 책임 법안(Algorithmic Accountability Act), FTC의 "Using Artificial

Intelligence and Algorithm", 과학기술정보통신부의 "국가 인공지능 윤리기준"(안)에서는 공통적으로 알고리즘의 투명성 내지 설명요구권을 제시하고 있다. 또한 학계에서는 투명성을 달성하는 방법으로 시장원리에 따른 규제, 산업별/회사별 자율규제, 국가 개입을 통한 공적 규제, 사전적 규제/사후적 규제가 논의되고 있다.[25)]

그러나 이러한 규제 방법론들은 모든 인공지능 기술에 일률적으로 적용되기 어렵다. 예를 들어, 투명성 달성을 위하여 알고리즘 공개를 요구하는 경우를 가정하면, 고도화된 인공신경망[26)]처럼 의사결정이 블랙박스(black-box) 구조를 가지고 있는 인공지능 시스템에서는 알고리즘의 공개 자체가 불가능하거나 공개하더라도 그 분석에 상당히 오랜 시간이 소요될 수 있기 때문에 결국에는 제도가 유명무실화될 가능성이 있다. 따라서 투명성 달성을 위한 구체적인 방법들을 모든 인공지능 시스템에 동일하게 적용할 수는 없는 것이고 여러 기준에 따라 인공지능 시스템들을 그룹화하고 각 그룹별로 서로 다른 기준을 적용할 필요가 있다.

다양한 인공지능 시스템을 나누는 방법은 여러 가지가 있을 수 있으나 이하에서는 투명성 확보라는 관점에서 인공지능 시스템을 구분하는 기술적인 기준을 살펴보고, 각 기준에 따라 구분된 인공지능 시스템에 종래 논의되었던 투명성 확보 방법이 적용 가능한지 알아본다.

(1) 인공지능 학습 방식에 따른 투명성 확보 기준

인공지능의 가장 기본적인 속성은 학습을 통해 사람과 같이 스스로 결정을 하고 예측을 할 수 있다는 것이다. 인공지능이 데이터를 학습하는 방식(machine learning)은 크게 3가지로 구분할 수 있는데, ① 지도학습(supervised learning), ②비지도학습(unsupervised learning), ③강화학습(reinforcement learning)이 있으며, 각 학습 방식 별로 서로 다른 투명성 확보 기준을 적용할 필요가 있다.

참고로 머신러닝 알고리즘은 매우 활발하게 연구가 진행되고 있는 분야이고 상당히 빠른 속도로 발전되고 있기 때문에 각 학습방식별로 다양한 종류의 알고리즘들[27)]이 개발되고 있다. 이러한 알고리즘들은 복잡도, 특성, 계산방식이나 로직 등이 서로 다르기 때문에 동일한 학습 방식이라고 하더라도 알고리즘의 종류에 따라 설명가능성이나 투명성이 다를 수 있다. 이에 아래에서는 구체적인 알고리즘에 따른 투명성 확보기준을 설명하기보다는 각 학습 방법의 일반적인 성질 차이에 따른 투명성 확보 기준을 논의한다.

1) 지도학습(supervised learning)

지도학습은 인공지능에게 어떤 것이 맞는 답인지를 사람이 지도하면서 학습을 시키는 방식이다. 예를 들어, 여러 스팸메일들을 인공지능에게 학습시키면서 '이 메일들은 모두 스팸메일이다'라는 답(target value)을 미리 알려준다. 이렇게 하면 인공지능은 제공받은 스팸메일 데이터들을 학습하면서 스팸메일의 특징들을 알아낸다. 그 결과 인공지능은 임의의 메일이 수신되면 그동안 학습한 메일들과 유사한지 아닌지를 살펴 해당 메일이 스팸메일인지 여부를 판단할 수 있게 된다.

이러한 지도학습은 인공지능 학습의 목표, 방법, 예상 결과를 모두 인간이 설계하고 진행하기 때문에 학습 과정의 투명성을 확보하기가 좀 더 용이하다는 특징이 있고, 지도학습 알고리즘 중에서 함수적 형태의 분류기[28)]를 사용하는 경우에는 그 근거를 추적하고 검증하는 것이 비교적 쉬우며[29)] 특정 답을 찾겠다는 학습의 결과도 뚜렷하기 때문에 인공지능이 도출해내는 결과값을 예상하고 분석하는 것도 가능할 수 있다.

이처럼 지도학습에 의한 인공지능의 경우에는 개발 단계 중 학습 과정에서의 투명성이나 알고리즘 자체의 투명성을 요구하는 것이 가능할 수 있기 때문에 다른 학습 방법에 비해 다양한 투명성 확보 방안을 적용하기가 좀 더 용이하고, 나아가 표준화된 알고리즘 영향평가 툴을 마련할 수도 있다. 특히 그동안 개발된 인공지능 시스템 중 상당수가 지도학습에 기초하고 있다는 점을 고려하면 학습 방식에 따른 구분 방식은 상당한 의미가 있다고 평가된다.[30)]

2) 비지도학습(Unsupervised learning)

비지도학습이란 정답이 없기 때문에 인공지능 스스로 목적이나 기준을 만들어 내고 그에 따른 결과값을 도출해 내는 방식이다. 예를 들어, 여러 메일 데이터를 인공지능에 학습시키면 인공지능은 그 스스로 메일들 사이의 숨겨진 패턴이나 규칙을 찾아내서 유사한 메일끼리 군집화하여 보여주게 되는데, 이렇게 군집화된 메일은 스팸메일/정상메일일 수도 있고, 첨부파일이 있는 메일/없는 메일일 수도 있다.

이러한 비지도학습에서 인간은 데이터를 수집하여 전달하기만 할 뿐 구체적인 학습 방법, 목표, 예상 결과를 모두 인공지능이 스스로 진행하기 때문에 지도학습 보다 학습 과정의 투명성을 확보하기가 좀 더 어렵고, 인공지능이 스스로 판단한 데이터 분류 방법은 개념적인 구분(예를 들면, 스팸메일인지 여부)이 아니라 메일에 포함된 수치 데이터값들의 차이나 평균에 따른 구분[31)]이어서 단순한 알고리즘을 사용한 경우가 아니라면 각 수치 계산이 어떠한 개념적인 의미를 나타내는지 다시 추론하고 검증하기가 어려우며, 인공지능이 어떠한 결과를 도출해 낼지도 예상하기가 쉽지 않을 수 있다.

결국 비지도학습에 의한 인공지능의 경우에는 알고리즘 자체의 투명성을 요구하는 것은 쉽지 않고,[32)] 학습에 사용된 데이터 자체에 대한 평가(학습데이터의 편향성 검증)나 인공지능이 도출하는 결과에 대한 평가(결과데이터의 편향성 검증) 방식으로 접근하는 것이 바람직하다.

25) Ansgar Koene, Chris Clifton, Yohko Hatada, Helena Webb & Rashida Richardson, "A governance framework for algorithmic accountability and transparency", Brussels: European Parliamentary Research Service, 2019, p.39.

26) Open AI사가 2020. 5. 공개한 GPT-3 모델의 경우에는 무려 1750억개의 매개변수를 가지고 있고, 인간이 평생 보는 정보보다 많은 데이터(크롤링 4100억개, 웹텍스트 190억개, 책 670억개, 위키피디아 30억개 등)를 학습한다.

27) 예를 들어, 지도학습에는 Regression (Linear, Polynomial, Ridge/Lasso), Classification (K-NN, Navie Bayes, SVM, Decision Trees), 비지도학습에는 Clustering (K-Means, Mean-Shift, Fuzzy C-Means, Agglomerative, DBSCAN), Pattern Search (Euclat, Apriori, FP-Growth), Dimension Reduction (t-SNE, PCA, LSA, SVD, LDA), 강화학습에는 Genetic, Q-Learning, DQN, SARSA, ASC 등 다양한 알고리즘이 있다.

28) 예를 들어 '광고'라는 글자가 3번 이상 등장하면 스팸메일로 구분하는 것과 같이 특정 수학함수에 따라 데이터 특성을 구분하는 방식이다. 회귀분석형(Regression Analysis) 분류기, 반복 분할형(Recursive Partitioning), 커널함수(Kernel Function) 및 거리함수(Distance Function) 기반 분류기 등이 있다.

29) 김도훈, "알고리즘 책임성 논의와 알고리즘에 대한 이해", 주간기술동향(2018. 5. 30.), 정보통신기술진흥센터, 18면.

30) 물론 앞에서 언급한 바와 같이, 지도학습 알고리즘이라고 하더라도 복잡도가 상당히 높거나 앙상블 기법과 같이 전통적인 알고리즘을 변형한 경우에는 투명성 확보가 어려울 수 있으며, 이러한 경우에는 후술하는 비지도학습의 투명성 확보 방안을 적용할 필요가 있다.

31) 클러스터링(Clustering) 알고리즘, 성분분석(Component Analysis) 또는 분산분석(Variance Analysis) 알고리즘 등에 따른 구분

32) 블랙박스형 알고리즘을 분석하는 XAI(eXplainable AI)에 대한 내용은 이하 목차에서 별도로 후술한다.

3) 강화학습(Reinforcement learning)

강화학습이란 정답은 모르지만 행동에 대한 보상은 알 수 있고, 그 보상으로부터 최대의 보상을 받는 방법으로 학습하는 것이다.[33] 예를 들어 체스 게임에서 체스 말이 특정 위치에 도달하는 것이 종국적으로 체스게임에서 이길 수 있는 방법인지는 알 수 없지만, 긍정적인 이벤트와 연관될 가능성(상대 체스 기물을 잡거나 퀸을 위협하는 것)과 부정적인 이벤트와 연관될 가능성(다음 차례에 상대방에게 기물을 잃게 되는 것)을 모두 고려하여 보상이 최대로 되는 경우를 계산하고, 만일 보상이 최대로 되는 위치로 이동했지만 결국 부정적인 이벤트가 발생한 경우에는 이러한 내용을 반복적으로 학습함으로써 결국에는 보상을 최대화하는 방법이다.[34]

인공신경망은 정해진 알고리즘에 따라 데이터를 처리하기 보다는 계속적인 학습을 통해 음의 로그우도(negative log-likelihood) 등 손실함수(loss function)를 최소화하는 최적의 값으로 모델의 파라미터들을 보정해나가는 방식이다. 이러한 보정과정을 거쳐 결정된 수치 값들은 투명성 원칙에 따라 공개된다고 하여도 사람이 그 의미를 분석하는 것은 사실상 불가능에 가깝다. 따라서 인공신경망에 대한 투명성은 새로운 접근방식이 필요하며, 아래와 같은 방식을 적용해 볼 수 있을 것이다.

가. 설명가능한 AI(eXplainable AI; XAI)

미국 고등연구계획국(Defense Advanced Research Projects Agency, DARPA)은 2016년부터 '설명 가능한 인공지능 프로그램(Explainable artificial intelligence program)'을 추진[35]하고 있고, 구글은 머신러닝 모델을 파악하고 해석하는 도구이자 프레임워크인 Explainable AI를 출시하고 AI 설명 백서를 발간하였다.

이러한 XAI 도구들을 이용하면 데이터의 각 요소가 모델 예측의 최종 결과에 기여한 정도에 따라 점수로 표현되기 때문에 특성 기여도를 분석할 수 있고,[36] 이를 통해 어떤 알고리즘으로 결과값을 도출하였는지 추론하는 것이 가능하다.

그러나 위 분석 도구는 특정 개발도구로 개발된 인공지능에만 적용 가능하고, 일종의 인공지능 역분석 도구라는 점에서 해당 인공지능에 대한 해킹이 용이해 질 수 있다는 점을 고려하면, 제한적으로 활용하는 것이 바람직하다.

나. 일관성이 아닌 상관성 방식에 따른 투명성

인공신경망의 경우 알고리즘 자체를 분석하기 어렵다는 점을 고려하면 결국 알고리즘에 의해 도출된 결과값을 사용하여 사용된 알고리즘을 역으로 추론하는 것이 가장 현실적인 방안이 될 수 있다. 다만 인공신경망에서 확률에 따라 각 신경망의 수치들을 조정하는 모델을 사용하는 경우에는 일관성 있는 결과

33) 김유두/장문수/이종서, 인공지능을 위한 텐서플로우 입문, 광문각, 2019, 148면

34) 세바스찬 라시카/바히드 미자리리, "머신러닝 교과서 with 파이썬, 사이킷런, 텐서플로", 길벗, 2019, 35면

35) 이상용, 앞의 논문, 152면

36) 한국정보화진흥원, "인공지능 의사결정 설명과 법·정책적 과제," 지능정보사회 법제도 이슈리포트, 2020

값이 도출되지 않을 수 있고, 따라서 특정 결과만을 기초로 알고리즘을 추론하는 것은 바람직하지 않으며 어느 정도 이상의 상관관계가 있음을 확인할 필요가 있다. 이러한 상관관계를 도출하는 방법으로는 저작권법 상 실질적 유사성 판단 방법을 참고할 수 있는데, 저작권법은 양 저작물의 상관관계(의거성 및 실질적 유사성)에 기초하여 복제 여부를 판단한다.[37]

다. 인공지능 자체에 윤리 소스코드 삽입을 의무화

인공지능에게 특정 결과는 도출될 수 없도록 강제하는 윤리 소스코드 삽입을 고려할 수 있다. 이미 로봇 분야에서는 로봇 3원칙[38]을 마련해 두고 있는데, 인공지능 분야에서도 마찬가지로 윤리적 판단을 위한 최소한의 경우의 수를 프로그래밍하는 것이 가능하다.[39] 예를 들어, 다른 인공지능과의 담합을 금지하거나, 인종/성차별을 할 수 없도록 프로그래밍해 둘 수 있고, 만일 이러한 제한사항에 해당하는 결과가 도출된 경우에는 그러한 결과가 나오지 않도록 인공지능을 재학습(강화학습)하는 방법도 가능하다.

(2) 인공지능 기술 개발 단계에 따른 투명성 확보 기준

인공지능 기술 개발은 설계단계, 데이터 수집단계, 테스트(검증) 단계로 나눌 수 있다. 인공지능은 알고리즘이 블랙박스화되어서 검증이 어렵다는 문제를 고려하면 기술 개발 단계에서부터 투명성을 확보할 필요가 있지만 기술 개발 초기부터 과도한 투명성 기준을 요구하면 기술 발전을 저해할 수 있으므로 기술 개발 단계에서 투명성을 요구할지 아니면 기술 개발 이후 상용화 단계에서 사후적인 투명성을 요구할지는 기술의 종류나 분야, 위험도에 따라서 적절히 결정하는 것이 바람직하다.

1) 설계단계

알고리즘 자체는 데이터의 각종 수치를 계산하는 것일 뿐 어떤 가치나 이념을 가지지 않는 중립적인 성질을 가지고 있다. 따라서 인공지능의 문제는 알고리즘에 의한 문제라기보다는 인공지능을 설계하는 과정에서 설계자의 주관적 편향이 개입되거나 고려해야 할 사항을 누락함에 따른 것으로 봄이 타당하다. 이러한 점을 고려하면 설계단계에서 투명성 기준을 적용하는 것은 근본적인 문제를 해결하는 방식으로 좀 더 유효적절할 수 있다.

구체적으로 앞서 논의한 비밀기능 개발 금지, 알고리즘 설계자에게 필요한 데이터를 요청할 수 있는 권리 부여, 개발 단계에서의 알고리즘 문제점 공유 방안 이외에 ① 추후 검증이 가능한 학습 방식이나 알고리즘으로 개발하는 경우와 블랙박스화되어서 추후 검증이 어려운 방식으로 개발하는 경우로 나눠서 후자의 경우에는 좀 더 엄격한 기준에 따라 개발 단계에서부터 투명성을 요구하거나, ② 인간의 신체나 건강에 위해를 줄 수 있는 의사결정의 경우에는 인공지능에

37) 김윤명, "알고리즘과 법-알고리즘 차별에 따른 공정성 확보방안," 한국정보화진흥원, 2019, 58면

38) 1법칙: 로봇은 인간을 해칠 수 없다.
2법칙: 로봇은 인간의 명령을 따라야 한다(단, 첫 번째 원칙과 충돌되는 명령은 제외)
3법칙: 로봇은 첫 번째 원칙 및 두 번째 원칙과 충돌되지 않는 한도에서 스스로의 존재를 보호해야 한다.

39) 김윤명, 앞의 논문, 84면

의한 의사결정을 할 수 없고 반드시 사람에 의한 의사결정이 개입될 수 있도록 개발하는 가이드라인을 마련할 수 있다. 더 나아가 분야별로 인공지능 개발 계획에 대한 승인/허가를 요구하거나 민감한 데이터의 경우 특정 알고리즘을 사용할 수 없게 제한하는 방법도 고려해 볼 수 있으나, 개발 단계에서부터의 과도한 투명성 요구 및 규제는 인공지능 산업 발전에 부정적인 영향을 미칠 수 있으므로 지양하는 것이 바람직하다.

2) 데이터 수집 단계

인공지능의 문제들은 상당 부분 학습 데이터의 편향성에 기인하는 경우가 많다. 이러한 편향성은 개발자의 주관적인 편향성에 기초한 경우도 있지만, 현실 세계의 데이터 자체의 수가 편향되어 있기 때문인 경우도 있다. 예를 들어, 인공지능 시스템은 채용, 인사 데이터를 수집해서 지원자를 평가하는데 여성 지원자에 대한 과거 데이터는 남성 지원자에 대한 과거 데이터보다 부족하기 때문에 여성을 배제할 가능성이 있고, 실제로 아마존이 개발한 AI 채용프로그램은 '여성' 키워드가 포함되면 선택하지 않은 것으로 알려졌다.[40)]

따라서 인공지능 알고리즘의 개발뿐만 아니라 데이터 자체의 편향성을 막을 필요가 있으며, 더 나아가 추후 투명성 확보를 위하여 ① 수집 데이터의 종류, 방법, 양에 대한 기록을 생성하고 보관할 의무 부과, ② 앞에서 논의한 바와 같이 데이터 전체 또는 일정 샘플을 일정 기간 저장할 의무를 부과할 수 있다. 한편 각 데이터의 편향성 조정을 위한 기준(예를 들면, 성별이나 인종이 고르게 분포될 수 있도록 조정)을 마련할 필요는 있으나, 특정 인공지능의 경우에는 효율적인 학습을 위하여 오히려 편향된 데이터가 필요할 수 있고 현실 세계의 데이터 자체가 편향되어 있어서 고른 데이터의 수집이 어려울 수도 있기 때문에 이러한 특수성은 반드시 고려되어야 한다.

참고로 EU 인공지능 백서의 인공지능 규제 내용을 살펴보면 고위험 인공지능 시스템에 대해서는 별도의 요건을 제시하고 있는데, 학습데이터에 관해서는 기존 법률과 EU의 가치와 원칙을 존중하기 위한 조치를 취해야 하고, 알고리즘의 프로그래밍과 학습용 데이터는 그 기록을 보관하거나 데이터 자체를 보관하도록 하고 있다.

3) 테스트(검증) 단계

인공지능 개발자는 데이터를 수집하면 해당 데이터를 학습용 데이터와 검증용 데이터로 구분하는 과정을 수행한다. 먼저 학습용 데이터로 인공지능을 개발하고 이후 검증용 데이터로 원하는 결과가 나오는지를 검증하게 되는데 만일 결과가 만족스럽지 못하면 피드백 과정을 거쳐서 인공지능 모델을 개선하게 된다.

이러한 검증단계에서 투명성 확보를 위한 데이터 세트를 제공하고 해당 데이터 세트를 통해 문제점 검증을 진행할 수 있다. 다만 이러한 검증단계가 서비스 출시 직전이라는 점을 고려하면 검증단계에서의 투명성 확보는 상업화 측면

40) 한국경제신문, AI도 '인종·성차별' 한다… 다른 접근법 필요한 이유, 2019. 2. 21. https://www.hankyung.com/it/article/201902200481g

41) 김도훈, 앞의 논문, 24면

42) 이상용, 앞의 논문, 143면

에서 매우 비효율적인 방법이므로[41)] 최소한의 데이터만으로 검증하는 것이 바람직하다.

(3) 인공지능 기술 분야별 투명성 확보 기준

인공지능은 기술 분야에 따라 인간이나 산업에 미치는 위험성이나 그러한 위험이 발생할 개연성이 모두 다르다. 만일 위험성이나 개연성이 높은 산업이라면 높은 수준의 투명성 확보 기준이 필요할 것이고, 그렇지 않은 산업의 경우에는 좀 더 낮은 수준의 투명성 확보 기준을 적용하는 것이 바람직하다.

예를 들어, 인공지능은 위험의 중대성과 개연성을 기준으로, 위험이 중대하고 발생가능성이 높은 경우(제1유형), 위험이 중대하지만 발생가능성은 낮은 경우(제2유형), 위험이 미약하고 발생가능성이 큰 경우(제3유형), 위험도 미약하고 발생가능성도 낮은 경우(제4유형)로 나눌 수 있다. 각 유형에 따라 강한 규제를 할지, 규제는 완화하면서도 피해 구제를 담보하는 제도적 장치를 마련할지, 사전 규제보다는 사후적인 규제를 할지, 규제 폐지할지를 정할 수 있다.[42)] 한편 EU는 인공지능 백서에서 인공지능이 보건의료, 운송, 경찰, 고용 및 법률시스템에 사용되는 경우나 인체에 위해를 가할 가능성이 있는 경우에는 해당 인공지능을 고위험 기술로 분류하고 별도의 강화된 규제를 통해 투명성을 요구하고 있다.

이처럼 인공지능은 기술 분야에 따라 등급을 나누고 등급별로 다른 수준의 투명성을 요구할 필요가 있다.

6. 가격 알고리즘과 공정경쟁

고위험 AI에 사용되는 알고리즘과는 별도로, 가격결정과 관련한 알고리즘(가격 알고리즘; pricing algorithm)의 경우에는 경쟁법을 위반할 가능성이 있으므로, 그 투명성의 관점에서 유의할 필요가 있다. 가격 알고리즘의 경우 온라인 상거래 플랫폼에서는 광범위하게 사용되고 있는데, 이를 통하여 사업자들간에 알고리즘을 통한 가격담합이 발생할 가능성이 높은 것이다. 해외에서는 Topkins와 경쟁사업자들이 Amazon에서 사용되는 가격정보를 지속적으로 수집하고, 이를 주동한 판매자가 구현한 일련의 규칙에 따라 제품가격을 책정하는 알고리즘 기반 가격책정 소프트웨어를 사용한 사안[43)] 등에서 가격 알고리즘을 통한 담합이 인정된 바 있다.

알고리즘 도입에 따라서 투명한 시장에서 데이터를 수집하는 것이 훨씬 용이하게 되었고, 그 결과 사실상 담합의 효과가 초래될 수 있다. 다만, 자기학습을 통한 알고리즘을 통하여 가격이 설정될 경우에도 그 알고리즘 형태에 따라서 처벌가능성은 달라질 수 있다. 감독되지 않은 기계학습 알고리즘의 경우에는 현행법상 법적으로 그 알고리즘 사용사업자에게 책임을 물을 수 있는지는 불명확하다고 보고 있고, 이에 따라 정부에서는 사업자 간 일치된 공동행위의 외관이 존재하고 이러한 외형상 일치에 필요한 정보가 교환된 경우에는 사업자 간 합의가 있는 것으로 법률상으로 추정하고 있다(공정거래법 개정(안) 제39조).[44)]

43) United States v. David Topkins, No. CR 15-00201 (N.D. Cal), Plea Agreement (April 30, 2015), https://www.justice.gov/atr/case-document/file/628891/download.

44) 한국지능정보사회진흥원, 알고리즘의 투명성, 공정성, 인공지능 시대 법제도 정비 연구보고서, 65 - 81면

7. 투명성과 그 부작용: 영업비밀로서의 알고리즘 보호

알고리즘은 기본적으로 개발자 및 사용자의 영업비밀이라고 할 수 있으므로 이를 과도하게 공개하도록 요구하기는 어려울 것이다. 나아가 알고리즘을 공개하는 경우 그 알고리즘을 역이용하려는 시도가 예상된다. 가령 쇼핑 관련 알고리즘이 투명하게 전부 공개되는 경우, 실제로는 가장 좋은 추천대상이 아니지만, 그 알고리즘이 중요하게 생각하는 점만을 강조한 기업이 가장 우선 추천대상이 될 수도 있는 것이다. 따라서 알고리즘의 어느 부분을 어떤 단계에서 누구에게 공개할 것인지가 핵심적인 문제라고 할 수 있다.

알고리즘을 그 투명성 보장을 위하여 공개하는 경우에는 영업비밀로서의 비공지성을 상실하지 않도록 하는 것이 중요할 것이다. 그렇지 않으면 알고리즘을 만들 아무런 인센티브가 없어서 인공지능 산업이 발전할 수 없기 때문이다. 또한 공지된 알고리즘의 소스코드를 일부 변형한 경우에도 원래의 소스코드 및 변형된 코드가 공지되었다는 사실만으로 영업비밀성을 상실하도록 해서는 안될 것이다. 또한 그 공개여부의 판단기준에 대하여도 「공공기관의 정보공개에 관한 법률」(이하 "정보공개법")의 원칙을 일응 참고하는 것이 바람직하다고 생각한다. 대법원은 국민의 알권리, 문제되는 법인의 성격 및 지위, 보호받아야 할 이익의 내용·성실 및 당해 정보의 내용·성질 등을 종합적으로 고려하여야 한다고 판시한 바 있는데(대법원 2012두12303 판결), 알고리즘 공개에 있어서도 이와 같은 기준이 참고가 될 것이다. 또한 알고리즘의 이와 같은 영업비밀로서의 보호필요성을 참작하여, 알고리즘과 관련한 소송에서는 현행 영업비밀 소송에서 활용되는 비밀유지명령제도, 비공개심리제도 및 법원에만 알고리즘을 공개하는 인카메라 절차의 도입을 적극적으로 검토할 필요가 있다.[45)]

(3) 한편 소송전단계에서 알고리즘의 투명성 통제의 방법으로는, 알고리즘의 자기평가 및 그 평가내용을 공시하는 것을 생각할 수 있다. 이와 유사한 제도로, 개인정보호보법에 규정된 개인정보파일에 관한 내용 등록(공공기관에 한한다 – 제32조), 개인정보보호관리체계 인증제도(ISMS-P, 제32조의2), 개인정보영향평가제(제33조)를 참고할 수 있을 것이다. 이와 같이 한다면 구체적이고 개별적인 알고리즘의 내용 공개는 최소한으로 하면서도 투명성을 확인하는 절차를 구현할 수 있을 것이다. 지능정보화기본법의 영향평가 제도(제56조)도 참고가 될 수 있다.

45) 한국지능정보사회진흥원, -앞의 보고서, 63면

Ⅳ. 맺음말

인공지능의 투명성 확보가 필요하다는 점에 있어서는 어느 정도의 사회적 합의가 도출되고 있으나 그 구체적인 기준이 마련되기 위해서는 아직 많은 논의가 필요한 상황이다. 특히 인공지능 기술의 종류에 따른 투명성 기준을 마련함에 있어서는 법조계나 정부 관계부처뿐만 아니라 산업계 전문가를 통해 인공지능 기술 자체의 특성을 고려한 다양하고 현실적인 의견을 수렴할 필요가 있고, 이를 통해 인공지능의 투명성 및 안정성 보장과 인공지능 산업의 성장을 모두 성취할 수 있는 조화로운 방안을 마련할 필요가 있다.