AI 투명성 거버넌스와 법제 정비 과제[1)]

마경태
법무법인(유) 태평양
변호사

1)
이 글은 필자가 한국지능정보사회진흥원의 "인공지능 시대 법제도 정비 연구보고서" 과제로 제출한 보고서에 기초하여 작성한 것임.

Ⅰ. 들어가며

기존에는 의사결정을 위한 전산 시스템을 구현함에 있어, 의사결정에 관한 명확한 기준이 이미 정해져 있고, 인간이 해당 기준에 따라 결정을 내리도록 프로그램을 작성하는 것이 일반적이었다. 그 의사결정 전산 시스템이 내린 결정에 대해 그 결정을 내린 이유를 설명해 줄 것을 요구 받을 경우, 그 시스템은 왜 그러한 결정을 내렸는지에 대해 설명을 할 수 있다. 나아가, 이처럼 설명이 가능한 시스템은 그 시스템이 특정한 결정을 내린 이유를 분석함으로써, 시스템이 그 결정을 내린 것에 대하여 누가 책임을 부담해야 하는지를 판단할 수 있다.

그러나 뉴럴 네트워크(Neural Network) 기반 기계학습(Machine Learning) 모델과 같이 새로운 지능정보기술을 활용한 인공지능(Artificial Intelligence, AI) 시스템의 경우, 인간은 그 시스템의 프로그램 코드를 확인할 수 있고 일반적인 작동 원리는 알 수 있지만, 그 시스템이 내린 특정한 결정에 대해서 왜 그러한 결정을 내렸는지에 대해 명확하게 알기 어렵다. AI 시스템은 이러한 기술적인 불투명성 때문에 일종의 '블랙박스(Black Box)'라고도 인식되고 있다.

의사결정 나무(Decision Tree) 모델, 로지스틱 회귀(Logistic Regression) 등 전통적인 기계학습 모델들의 경우, 상대적으로 그 구조가 단순하므로 의사결정이 이루어지는 기준을 어느 정도 이해할 수 있다. 하지만 이러한 모델들을 활용한 AI 시스템의 경우라도, AI 시스템을 기반으로 한 제품이나 서비스를 제공하는 기업들이 대부분 그 기준들을 공개하거나 그러한 시스템을 활용하고 있다는 사실 자체를 외부에 알리고 있지 않는 경우가 적지 않다. 그리고 일반인으로서는 AI 시스템의 작동 원리나 기준에 대한 정보를 접하더라도 복잡성으로 인해 이를 이해하기 쉽지 않다.

따라서 AI 시스템의 결정으로 인해 개인에게 위해(harm)가 발생하더라도, 피해자는 AI 시스템의 '불투명' 한 속성으로 인하여 위해를 입게 되었다는 사실 자체를 인식하기 어려울 수 있다. 위해를 입었다는 사실을 알게 되더라도 AI 시스템이 그 결정을 내린 이유를 쉽게 알 수 없으며, 누구에게 어떠한 책임을 물을 수 있는지에 대해 알기 어려울 수 있다. 더욱 우려되는 사실은 AI 시스템이 점차 고도화되고 복잡해짐에 따라 AI 시스템을 운영하는 기업들조차도 위해 발생 가능성을 쉽게 예견하기 어려워질 수 있다는 점이다. 눈에는 쉽게 보이지 않는 AI 시스템의 위험은 기술이 발전하고 확산됨에 따라 증가할 것이므로, 이에 대응하여 AI 시스템의 투명성 확보를 위한 거버넌스에 대한 논의가 필요한 상황이다.

Ⅱ. AI 투명성 확보를 위한 거버넌스

1. AI 투명성 원칙이란

AI, 빅데이터 등 새로운 지능정보기술이 등장하고 활발히 응용됨에 따라, 전 세계적으로 예기치 못한 윤리적, 사회적 위험에 대응해야 한다는 의식이 고조

되고 있다. 앞서 설명한 AI 시스템의 불투명성 문제가 그 대표적인 위험에 해당한다. 이에 각국 정부, 다국적 기업, 국제기구, 전문가협회 등은 AI 시스템이 준수해야 하는 윤리 원칙들을 발표하고 있는데, 위 윤리 원칙들은 공통적으로 AI 시스템의 투명성(Transparency) 원칙을 책임성(Accountability), 자율성(Autonomy·Human agency), 공정성(Fairness) 원칙과 함께 주요 원칙으로 명시하고 있다.

대표적으로 유럽연합 집행위원회(EC)의 AI 고위급 전문가그룹(High-Level Expert Group on Artificial Intelligence, AI HLEG)이 2019. 4. 발표한 '신뢰가능한 AI를 위한 윤리 가이드라인'과 과학기술정보통신부에서 2020. 12. 발표한 '인공지능(AI) 윤리기준'은 AI 시스템이 준수해야 할 기본 원칙에 투명성 원칙을 명시하고 있다.[2)]

투명성 원칙은 ① AI 시스템이 사람에 의해 관리·감독이 이루어질 수 있도록 설계 및 실행되어야 한다는 점과 ② AI 시스템이 무엇을 어떤 이유로 하고 있는지에 대한 설명이 이해하기 쉽고 평가가 가능한 형태로 제공되어야 한다는 점(설명가능성)을 주요 내용으로 한다. 특히 다수의 문헌들은 AI 시스템이 개인에게 위해를 발생(cause harm)시키거나 개인에게 중요한 영향을 미치거나(have a significant effect on individuals), 개인의 삶, 삶의 질, 명예에 영향을 미치는(impact a person's life, quality of life, or reputation) 경우에 설명가능성이 중요하다고 설명하고 있다.[3)]

투명성 원칙은 책임성, 자율성, 공정성 등 다른 윤리 원칙들을 실현하기 위한 전제 조건으로서의 성격을 갖고 있다. 가령 이용자는 투명성 원칙을 통해 AI 시스템이 특정 의사결정을 내린 이유를 알 수 있어야 AI 시스템의 결정에 이의를 제기하거나, AI 시스템으로부터 받은 피해로부터 구제를 받을 수 있다.[4)] 여기서 이용자가 AI 시스템의 결정에 이의를 제기하는 것은 AI 시스템에 대해 '인간의 개입'을 보장한다는 것을 의미하고(자율성), AI 시스템으로부터 받은 피해에 대하여 구제를 받는다는 것은 AI 시스템에 대하여 '책임'을 물을 수 있어야 한다는 것을 의미한다(책임성).

나아가 AI 시스템이 특정 집단에 대해 편향되었는지 여부를 확인하기 위해서는 AI 시스템의 작동에 관한 정보가 필요한데, 이러한 정보를 확인하기 위해서는 AI 시스템의 투명성이 확보되어야 한다. 따라서 투명성 원칙은 AI 윤리 원칙 중 공정성 원칙과도 밀접한 관련이 있다.

이처럼 투명성 원칙은 개인의 권리 행사나 법익을 보호하기 위한 '도구적' 수단으로서의 성격을 갖는다는 점에서 '알 권리'나 '적법절차의 원칙'과 유사하다. 따라서 투명성의 구체적인 내용은 AI 시스템이 개인 또는 사회에 어느 정도의 위험을 발생시키는지에 따라 달라질 필요가 있다.

2) European Commission AI HLEG (High-Level Expert Group on Artificial Intelligence), "Ethics guidelines for trustworthy AI", Brussels: European Commission, 2019, p.14.; 과학기술정보통신부, "사람이 중심이 되는 인공지능(AI) 윤리기준", 과학기술정보통신부, 2020, 8면.

3) Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, & Madhulika Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI." Berkman Klein Center for Internet & Society, 2020, pp.42-42.

4) Edwards, Lilian & Veale, Michael, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For", 16 Duke Law & Technology Review 18, 2017, p.21.; Kaminski, Margot E., "The right to explanation, explained". 34 Berkeley Tech. L.J. 189, 2019, p.204.

5) Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R., "A governance framework for algorithmic accountability and transparency", Brussels: European Parliamentary Research Service, 2019, p.39.

2. AI 투명성 확보를 위한 거버넌스 논의

가. 구속력에 따른 AI 거버넌스 구분

현재 국내외로 AI 윤리 원칙들을 실현하기 위한 관리·감독(AI 거버넌스) 체계에 관해서 다양한 연구와 입법 시도가 이루어지고 있다. AI 거버넌스 모델은 규제의 구속력을 기준으로 다음과 같이 크게 5가지로 나누어 볼 수 있다.[5)]

모 델		내 용
시장원리		소비자들이 AI 시스템에 기반한 제품 또는 서비스를 제공하는 기업들에게 투명성을 요구하고, 이에 기업들은 경쟁적으로 투명성을 확보함으로써 시장에서의 차별화를 도모함.
기업 자기 조직화		기업의 CSR(Corporate Social Responsibility) 활동의 일환으로서, 기업이 AI 시스템의 불투명성으로 인해 발생할 수 있는 위험을 관리하기 위해 자발적으로 취하는 조치들을 의미함. 대표적으로 내부 거버넌스 및 품질 검토 절차 수립, 옴브즈맨 제도의 도입을 들 수 있음.
산업별 자율 규제		IEEE 등 전문가협회에서 작성한 행동강령(Codes of Conduct), 기술적 및 관리적 국가·국제 산업 표준(예: ISO, IEC), 알고리즘 시스템 인증(certification), AI 기업들로 구성된 윤리 위원회(예: Partnership on AI) 등을 통한 자율 규제.
정부-사업자 공동 규제		사업자 자율 규제와 정부 모니터링이 혼합된 형태의 통제. 정부는 사업자들에게 이행해야 할 원칙들과 함께 원칙 준수를 위해 적합하다고 생각하는 이행 방안들을 제시하면서, 구체적인 이행 방안들을 기업들과 상호 협력을 통해 정해 나가는 방식.
국가 개입	법적 권한 창설	GDPR의 자동화된 의사결정 조항과 같이 입법을 통해 개인에게 권리를 부여하는 방식.
	규제기관 통제	AI 시스템에 대하여 규제기관이 직접 통제하는 방식. 규제기관의 역할로는 AI 시스템에 대한 기술·책임기준 제정, AI 모델에 대한 직접적 통제 등이 있을 수 있음.

[표] 규제의 구속력을 기준으로 구분한 AI 거버넌스 모델

6) Kaminski, Margot E., "Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability", Southern California Law Review 92, no. 6, 2019, pp.1552-1553.

나. 법적 보호 방식에 따른 AI 거버넌스 구분

투명성에 관한 AI 거버넌스는 법적 보호 방식에 따라 크게 '개인의 권리 행사를 통한 방식'과 'AI 시스템에 대하여 관리·감독 체계를 구축하는 방식'으로 나누어 볼 수 있다.[6)]

이용자는 AI 시스템 활용 사실 또는 AI 시스템이 내린 의사결정의 이유에 대한 정보를 토대로 AI 시스템이 내린 의사결정에 대해 정정을 요구하거나 다툴지 여부를 판단할 수 있다. 이에 '개인의 권리 행사를 통한 방식'의 경우, 이용자에게

AI 시스템 활용 사실을 고지할 의무(AI 시스템 활용 고지의무), 이용자가 AI 시스템이 내린 의사결정의 이유에 대하여 설명을 요구할 권리(설명요구권), 이용자가 AI 시스템이 내린 의사결정의 적용을 받지 않거나 의사결정에 이의를 제기할 수 있는 권리(이의제기권·적용거부권)를 입법하는 방안이 주로 논의되고 있다.

그리고 'AI 시스템에 대하여 관리·감독 체계를 구축' 하는 방안으로는 AI 시스템에 대한 감사, 영향평가, 적합성 평가, 보호책임자의 지정 등이 주로 논의되고 있다.

'개인의 권리 행사를 통한 방식'은 AI 시스템에 의해 영향을 받는 개인의 권리 구제 측면에서 유리하고, 개인들로 하여금 알고리즘이 내리는 개별 의사결정의 정당성에 대해 의문을 갖고 감시하도록 할 수 있다는 장점이 있다. 하지만 개인은 부여된 권리를 의미 있게 행사하기에 시간적 여유, 자원, 전문성이 부족하고, 실제로 개인에게 제공되는 설명의 내용이 권리 행사에 충분하지 않을 수 있으므로, 개인의 권리 행사를 통한 방식이 실효성이 없다는 의견도 존재한다.[7)]

반면 AI 시스템에 대한 관리·감독 체계를 구축하는 방식의 경우, 일반 개인과 달리 전문성을 가지고 시스템을 평가할 수 있고, 시스템의 개발 단계에서부터 출시 단계에 이르기까지 지속적인 검토가 가능하며, 일반 개인이 쉽게 인식할 수 없는 알고리즘의 문제점(예: 특정 집단에 대한 편향성 등)에 대해서도 검증이 가능하다는 점에서 장점이 있다. 하지만 규제기관의 감시만으로는 자원의 한계로 인하여 사업자들을 충분히 관리·감독이 어려울 수 있고, 일단 감독 대상이 된 사업자의 영업행위에 대하여 지나치게 개입할 우려가 있다는 문제점이 제기된다.

개인의 권리 행사를 통한 방식과 AI 시스템에 대한 관리·감독 체계를 구축하는 방식은 각각 고유한 장단점들이 있으므로, 투명성 확보를 위한 AI 거버넌스 마련 시 두 가지 접근 방식을 상호보완적으로 활용할 필요가 있다.

Ⅲ. 국내외 AI 거버넌스 도입 사례

1. EU GDPR

가. 개요

AI의 투명성과 관련하여 현재 국제적으로 가장 많이 논의되고 있는 AI 거버넌스 모델은 2018. 5.부터 시행된 EU의 유럽연합의 일반정보보호규정(General Data Protection Regulation, GDPR)이다. GDPR은 개인정보 보호의 거버넌스를 위한 법이지만, '자동화된 의사결정(Automated Individual Decision-making)'의 통제에 대해서도 다루고 있다. GDPR의 자동화된 의사결정 규정은 모든 AI 시스템에 대하여 일반적으로 적용되는 것이 아니라, 개인정보를 이용하여 개인을 대상으로 자동화된 의사결정을 내리는 AI 시스템에 국한하여 적용된다.

GDPR은 자동화된 의사결정에 대하여 '자동화된 의사결정 존재 고지의무', '관련 논리에 관한 유의미한 정보를 제공할 의무', '적용거부권', '이의제기권', '정보보호 영향평가' 등 AI 시스템의 투명성 확보를 위한 규정들을 포함하고 있다.

GDPR은 앞서 규제의 구속력을 기준으로 구분한 거버넌스 모델 중 '국가 개입' 내지는 '정부-사업자 공동 규제' 모델에 해당한다.[8)] 그리고 법적 보호 방식의 관점에서 주로 '개인의 권리 행사'에 관한 요소들로 구성되어 있으면서, 동시에 'AI 시스템에 대한 관리·감독 체계' 요소도 포함하고 있다고 평가된다.[9)]

나. 설명요구권 논란

GDPR 본문 상으로는 정보주체에 '설명요구권'을 부여한다고 명확히 규정된 조항이 없다.[10)] GDPR 제13조 내지 제15조는 정보주체가 컨트롤러부터 '관련 논리에 대한 의미 있는 정보'(meaningful information about the logic involved)를 받거나 열람할 권리가 있다고 규정하고 있지만, 문언만으로는 그 의미가 무엇인지 명확하지 않다.

이에 최근 수 년간 과연 GDPR으로부터 설명요구권이 도출되는지에 대해 학자들 사이에서 다양한 논의가 이루어져 왔다. 대표적으로 (i) GDPR 제13조 내지 제15조의 '관련 논리에 대한 의미 있는 정보'를 받을 권리로부터 설명요구권이 도출된다는 견해,[11)] (ii) AI 시스템에 대한 설명을 ① '시스템 기능에 대한 설명'[12)]과 '특정 결정에 대한 설명',[13)] ② 자동화된 의사결정이 내려진 시점을 기준으로 '사전적 설명'과 '사후적 설명'으로 구분하고, 이 중 '시스템 기능에 대한 (사전·사후적) 설명'과 자동화된 의사결정의 존재에 대한 제한적인 설명요구권(고지받을 권리)이 제13조 내지 제15조로부터 도출된다는 견해,[14)] (iii) 제13조 및 제14조의 '관련 논리에 대한 의미 있는 정보'와 제22(3)조의 '적절한 보호 대책 수립'으로부터 설명요구권이 도출된다는 견해[15)] 등이 있다.

지금까지의 논의 내용을 보면, 연구 내용마다 설명요구권이 발생하는 법적 근거나 설명의 내용은 다르지만, 대체로 GDPR로부터 자동화된 의사결정에 대해 적어도 일정한 범위 안에서 설명을 요구할 권리가 도출된다는 점에 대해서는 학자들 사이에 어느 정도 공감대가 형성되어 있다고 보인다.

다. 알고리즘 영향평가 활용 가능성

GDPR의 정보보호 영향평가(Data Protection Impact Assessments, DPIA)는 개인정보처리자가 개인정보의 처리 내용에 관해 설명하고, 처리의 필요성 및 비례 원칙 준수 여부를 검토하며, 개인정보의 처리로 인해 발생하는 개인의 권리와 자유에 대한 위험에 대하여 평가하고 대응 방안을 마련할 수 있도록 만든 절차이다.[16)]

최근 AI 알고리즘의 투명성 및 책임성 확보를 위하여 DPIA를 알고리즘 영향평가로 활용하는 방안에 관해서 활발한 논의가 이루어지고 있다.[17)] 이러한 연구들은 GDPR의 DPIA 규정을 근거로 AI 시스템을 개발·구축하는 자가 AI 시스템으로 인하여 개인의 권리와 자유에 발생할 수 있는 잠재적 위험을 평가하고, 이에 대한 대응조치를 마련하며, 이 과정을 기록해야 한다고 주장한다.

7) Edwards, Lilian & Veale, Michael, 앞의 논문, p.26.

8) Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R., 앞의 보고서, pp.44-47.

9) Kaminski, Margot E., "Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability'" Southern California Law Review 92, no. 6, 2019, p.1583.

10) 다만, GDPR 전문(recital) 제71조는 정보주체에 대한 "평가 이후 도달한 결정에 대한 설명을 획득할 권리"를 규정하고 있으나, 전문 부분은 법적 구속력이 없다고 평가된다.

11) Selbst, Andrew D. & Powles, Julia, "Meaningful Information and the Right to Explanation", International Data Privacy Law, vol. 7(4), 2017, p.242.

12) '시스템 기능에 대한 설명'은 자동화된 의사결정 시스템의 논리, 중요성, 예상되는 결과 및 일반적 기능을 의미함.

13) '특정 결정에 대한 설명'은 특정한 자동화된 의사결정에 대하여 근거, 이유 및 개별적 상황을 의미함.

14) Wachter, Sandra & Mittelstadt, Brent & Floridi, Luciano, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", International Data Privacy Law, 2017, p.41.

15) Kaminski, Margot E., "The right to explanation, explained". 34 Berkeley Tech. L.J. 189, 2019, p.211.

16) Article 29 Data Protection Working Party, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is likely to result in a high risk for the purposes of Regulation", WP 248 rev.01, 2017, p.4.

17) Casey, Bryan, and Farhangi, Ashkon, and Vogl, Roland, "Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise", Berkeley Technology Law Journal 34, no. 1, 2019.; Kaminski, Margot E. and G. Malgieri. "Multi-layered explanations from algorithmic impact assessments in the GDPR", Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.

DPIA를 알고리즘 영향평가와 같이 활용하는 방안은 실무상으로도 도입이 추진되고 있다. 대표적으로 2020. 7. 영국 내 개인정보 보호를 규율하는 독립 감독기관인 영국 정보위원회(Information Commissioner's Office, ICO)는 AI 시스템을 개발·구축하는 과정에서 GDPR의 개인정보 보호 및 AI 윤리 관련 규정 내용을 어떻게 이행하고 이를 검증하는지에 대한 실질적인 안내와 예시를 담은 'AI 및 개인정보 보호 가이드라인'(Guidance on AI and data protection)을 발표하였는데, 위 가이드라인은 GDPR의 DPIA를 알고리즘 영향평가와 같이 활용하는 방안을 구체적으로 제시하고 있다.[18)]

18) Information Commissioner's Office, "Guidance on AI and data protection", 2020.

2. 개정 신용정보법

국내에서는 2020. 8. 5.부터 시행된 개정 「신용정보의 이용 및 보호에 관한 법률」(이하 "개정 신용정보법")은 '개인신용평가'에 관한 업무에 대하여 정확성·공정성·투명성 원칙을 규정하고(제22조의3 제1항), 신용정보주체에게 개인신용평가 등의 자동화평가에 대한 '설명요구권'과 '이의제기권'을 부여하고 있다(제36조의2).

이에 따르면, 신용정보주체는 개인신용평가회사 등에 대하여 ① 개인신용평가 행위등에 대해 자동화평가를 하는지 여부, ② 자동화평가의 결과, ③ 자동화평가의 주요 기준, ④ 자동화평가에 이용된 기초정보에 대하여 설명요구권을 행사할 수 있다(제36조의2 제1항 제2호). 그리고 개인신용평가회사등에 대하여 ① 자동화평가 결과의 산출에 유리하다고 판단되는 정보를 제출하거나 ② 자동화평가에 이용된 기초정보의 내용이 정확하지 아니하거나 최신의 정보가 아니라고 판단되는 경우, 기초정보를 정정·삭제할 것을 요구하거나, 자동화평가 결과를 다시 산출할 것을 요구할 수 있다(제36조의2 제2항).

3. 해외 입법안

미국의 경우, 연방 차원에서 AI 거버넌스에 관한 일반법이 입법된 적은 없다. 다만 2019. 4. 미국 상원에서 AI 시스템에 사용되는 알고리즘의 편향성과 차별적 결과를 방지하기 위해 미국 연방거래위원회(Federal Trade Commission, FTC)에게 알고리즘 규제 권한을 부여하고, '자동화된 의사결정 시스템 영향평가'를 도입하는 '알고리즘 책임법안'(Algorithmic Accountability Act)이 발의된 적이 있다.

본 법안에 따르면, 매출액, 이용자 수 등이 일정 기준 이상인 개인 또는 기업 등은 '고위험'의 자동화된 의사결정 시스템에 대해 자동화된 의사결정 시스템 영향평가(Automated Decision System Impact Assessment)를 수행해야 한다(SEC. 3(b)(1)). 자동화된 의사결정 시스템 영향평가는 자동화된 의사결정 시스템 및 시스템의 개발 과정(시스템 디자인과 학습 데이터 포함)을 대상으로 하고, 시스템의 정확성, 공정성, 편향성, 차별성, 프라이버시, 보안에 대한 영향을 평가해야 한다(SEC. 2(2)).

한편, EU는 2020. 2. 향후 도입할 AI 규제의 기본 틀을 마련하기 위해 'AI의 발전과 신뢰 확보를 위한 백서'(이하 "AI 백서")를 발표하였는데, 이를 기반으로 최근 2021. 4. 'EU AI 규제법안'(Proposal for a Regulation on a European Approach for Artificial Intelligence) 초안을 발표하였다.

EU AI 규제법안에 따르면, ① 고위험 AI 시스템은 위험 평가·대응 시스템(risk management system) 구축, 고품질의 데이터세트 마련, 시스템 기술 명세서(technical document) 작성, 시스템 운영 내역 기록, 이용자에 대한 정보 제공, 사람에 의한 통제 확보, 높은 시스템 성능·안정성·안전성 확보 등의 요건을 준수해야 하고(제8조 내지 제15조), ② 고위험 AI 시스템을 공급하는 자는 위 ①번 의무들을 준수함과 동시에 품질 관리 시스템(quality management system) 적용 및 적합성 평가(conformity assessment) 수행 등의 의무를 부담하며(제16조 내지 제23조), ③ AI 챗봇과 같이 사람과 소통하는 AI 시스템(단, 정황상 AI 라는 점이 명백한 경우는 제외), 사람의 감정을 인식하거나 생체정보를 분류하는 AI 시스템, 딥페이크 등 이미지·오디오·영상을 진위를 알 수 없게 생성 또는 조작하는 AI 시스템은 원칙적으로 고위험 여부와 상관없이 그 사실을 이용자에게 알려주어야 한다(제52조).

Ⅳ. 법제 정비를 위한 향후 과제

본 항에서는 AI 시스템의 투명성 확보를 위한 법적 구속력 있는 AI 거버넌스(이하 "AI 투명성 규제") 도입에 필요한 단기 및 중장기 과제를 제시한다.

1. 입법의 필요성 검토 - 단기 과제

가. 규제 대상 AI 시스템 판단 기준: 위험성

AI 거버넌스는 AI 시스템의 새로운 기술적 특성에 따라 나타날 수 있는 예기치 못한 기술적, 사회적 위험을 최소화하는 동시에 AI 기술을 개발하고 응용하는 기업들의 정당한 이익을 침해하거나 기업들에게 지나친 부담을 주어 AI 산업 발전의 저해 요소로 작용해서도 안 된다. 따라서 AI 거버넌스는 빠르게 변화하는 기술에 맞추어 새로운 위험을 평가하고, 위험에 대한 대응과 기업의 권리 보호 사이에 비례성을 지킬 수 있는 유연한 형태를 가질 필요가 있다.

이러한 점들을 고려하여 AI 거버넌스 모델로 AI 시스템의 '위험성'을 기반으로 접근하는 방식(risk-based approach)이 가장 많이 논의되고 있다.

EU AI 백서와 AI 규제법안은 AI 규제체계에 위험 기반 접근 방식을 적용해야 한다는 점을 명시하고 있다.[19)] EU AI 백서에 따르면, AI 시스템에 대한 법적 구속력 있는 규제의 적용 여부는 AI 시스템이 고위험인지 여부에 따라 달라진다. AI 시스템이 보건의료, 교통, 에너지, 일부 공공분야(가령 출입국, 사법, 사회보장, 채용)와 같이 '통상적으로 위험 발생 가능성이 높은 영역에 적용'(요건①)되고, 동

19) European Commission, "WHITE PAPER on Artificial Intelligence - A European approach to excellence and trust", pp.17-18; Recital 12 of the Regulation on a European Approach for Artificial Intelligence.

시에 '실제로 해당 영역에서 위험 발생 가능성이 높은 방식으로 이용'(요건②)될 때 해당 AI 시스템을 고위험으로 분류한다. 그리고 AI 시스템이 채용 과정에 이용되거나, 근로자의 권리에 영향을 미치거나, 원격 생체인식 기술에 적용되는 경우에는 예외적으로 위 두 가지 요건 충족 여부와 상관없이 고위험으로 분류한다. 그리고 최근 발표된 EU AI 규제법안은 위 EU AI 백서 내용과 마찬가지로 주요 규제의 적용 대상을 의료기기, 장난감, 기계 등 제품의 안전부품, 생체인식·분류, 사회기반시설(도로·철도·항만 등) 관리·운영, 교육기관의 입학·평가, 채용·인사평가, 필수 공공·민간 서비스(복지, 신용평가, 긴급구조), 법 집행, 이민·난민·출입국 통제, 사법을 위해 활용되는 AI 시스템 등 고위험 AI 시스템으로 한정하고 있다(EU AI 규제법안 TITLE III 및 Annex III).

미국 알고리즘 책임법안의 경우, 우선 법의 적용 대상을 매출액, 이용자 수 등을 기준으로 일정 수준 이상인 개인 또는 기업 등으로 제한하고, 그 중 고위험의 자동화된 의사결정 시스템을 대상으로 자동화된 의사결정 시스템 영향평가를 받도록 규정하고 있다. 이때 고위험 여부는 AI 시스템이 소비자에게 법적 또는 기타 중대한 영향을 미치는지 또는 AI 시스템이 다루는 개인정보의 종류나 개인정보를 다루는 방식의 위험성에 따라 판단하고 있다.

GDPR도 자동화된 의사결정에 대한 규정이 적용되는 범위를 정보주체에게 법적 효력을 발생시키거나 그와 유사한 중대한 효과를 미치는 결정으로 제한함으로써 자동화된 의사결정의 위험성에 기반한 규제 방식을 적용하고 있다.

결국 AI 거버넌스가 AI 시스템의 다양한 위험에 능동적으로 대응하고 국내 AI 산업 성장에 장애물로 작용하지 않기 위해서는 위험 기반의 접근 방식을 적용해야 한다고 본다. 위험 기반의 효과적인 AI 거버넌스의 법제화를 위해서는 ① AI 시스템 유형별로 발생시킬 수 있는 위험을 파악하고, ② 규제로 인해 기업에게 발생하는 부담을 파악하여, 규제 도입으로 인한 기업의 부담과 사회적 효용(위험 방지 및 이용자 권리 구제) 사이에 비례성이 지켜지는지 확인해야 한다. 그리고 ③ 기업들의 자율적인 규제를 통해 AI 시스템에 대한 규제를 도입하지 않거나 최소한으로 도입할 수 있는 방안에 대한 연구도 필요하다.

이하에서는 투명성 확보 관점에서 위 각 과제들을 이행하는 방안에 관하여 자세히 설명한다.

나. 국내 AI 활용 현황 및 위험성 연구

AI 시스템은 개인의 법익을 침해할 수 있을 뿐만 아니라 사회 전체에 대한 외부효과를 발생시킬 수 있다. 그리고 AI 시스템이 내포하고 있는 위험은 AI 시스템의 유형별로 다양하게 나타날 수 있다.

가령, 채용 AI 시스템이 편향적으로 작동하는 경우, 지원자 개인의 평등권을 침해할 가능성이 있는 동시에 특정 사회 집단에 대한 차별로 이어질 수 있다. 나아가 AI 시스템의 편향성으로 인해 사회에 발생하는 위해는 ① 채용, 대출 영역에서의 차별과 같이 특정 집단에게 기회 또는 자원을 제공하거나 제공하지 않는

20) Crawford, Kate, "The Trouble with Bias", Conference on Neural Information Processing Systems, invited speaker, 2017.

직접적인 형태로 나타날 수도 있지만(Harms of Allocation, 분배의 위해), ② AI 챗봇이 특정 집단에 대하여 편향적인 발언을 하는 경우와 같이 특정 집단에 대한 사회적 소외를 강화시키는 간접적인 형태로도 나타날 수 있다(Harms of Representation, 대표성의 위해).[20] 위해의 효과를 즉시 확인할 수 있는 분배의 위해와 달리 대표성의 위해는 장기적으로 나타나고 효과를 가시적으로 확인하기가 어렵다는 특징이 있다.

AI 투명성 규제는 AI 시스템이 개인 또는 사회에 발생시킬 수 있는 위험을 방지하고 대응하기 위한 수단으로서의 기능을 하기 때문에, AI 시스템이 발생시킬 수 있는 위험의 내용과 정도에 따라 규제의 도입 여부 및 내용도 달리 정해야 한다. 가령 AI 시스템의 특정 집단에 대한 편향성은 AI 시스템을 이용하는 개개인은 인식하기 어려울 수 있으므로, 개인의 권리 행사를 통한 방식보다는 AI 시스템에 대한 관리·감독을 통해 대응하는 것이 더 효과적일 수 있다.

따라서 AI 투명성 규제를 도입하기 위해서는 우선 현재 국내에서 활용되고 있는 AI 시스템의 유형별로 어떻게 활용되고 있고 어떠한 위험을 발생시킬 가능성이 있는지에 대한 연구가 선행되어야 한다. 그리고 이를 토대로 규제 도입이 필요한 분야와 각 분야별로 적합한 규제 내용이 무엇인지 명확하게 정리할 필요가 있다. 특히 국내에서 규제가 필요한 '고위험' AI 시스템의 유형이 무엇인지부터 확인이 필요하다.

특히 채용, 대출 영역에서의 차별과 같이 이미 활발하게 연구가 이루어지고 있는 분야뿐만 아니라 미디어 분야에서의 AI 추천시스템 활용으로 인해 발생할 수 있는 확증편향 현상과 그로 인한 민주주의에 대한 위협과 같이 새롭게 주목을 받고 있는 위험 분야에 대해서도 어떠한 위험을 발생시키는지에 대한 실증적인 연구가 이루어져야 한다.

다. AI 투명성 규제로 인한 기업 부담

1) 설명 가능한 인공지능 구축

AI 알고리즘 모델 중에는 의사결정 나무(Decision Tree) 모델과 같이 모델 자체의 분석을 통해 동작 원리를 일정 수준 해석할 수 있는 모델들이 있는 반면, 뉴럴 네트워크(Neural Network) 모델 등과 같이 별도의 '설명 가능한 인공지능'(eXplainable AI, XAI) 기술을 적용하지 않고서는 해석이 어려운 모델들이 있다. 이와 같이 AI 시스템의 설명 가능성은 AI 모델 자체의 해석 또는 XAI의 적용을 통해 실현할 수 있는데, 이론상 모델 자체의 분석이 가능한 모델의 경우에도 모델 자체의 해석만으로는 사람이 이해하기 쉬운 직관적인 설명을 도출해내기 어려운 경우에는 XAI가 필요할 수 있다.

XAI는 여전히 많은 개발 및 연구가 이루어지고 있는 분야로서 아직 기술적으로 상당한 한계가 존재한다. XAI 기술은 아직까지 어느 기계학습 모델이든 간단히 적용할 수 있는 수준에 이르지 못했고, 적용하는 기계학습 모델의 특성에 따라 적용 방법이나 성능이 달라질 수 있다. 가령 기계학습 모델의 종류에 관계없

이(Model-agnostic) 적용할 수 있는 XAI 기술로 대표적으로 SHAP과 LIME 분석 기법이 있는데, SHAP 분석의 경우 연산을 위해 매우 오랜 시간이 걸리고, 이상치(outlier)가 많으면 제대로 된 성능을 발휘하지 못할 수 있다. 그리고 LIME 분석은 관측치 근방의 분석 범위(neighborhood)를 정하기 쉽지 않고, 두 개의 매우 가까운 관측치에 대한 설명이 다를 수 있어 안정성이 떨어진다는 지적이 있다.[21)]

따라서 기업이 AI 시스템에 XAI를 적용하기 위해서는 별도로 전문성을 갖추기 위한 추가적인 노력이 필요할 수 있고, XAI 구축을 위해 기존 AI 시스템 디자인의 변경이 필요할 수 있는바, 이는 AI 시스템을 개발·이용하는 기업에게 상당한 부담으로 작용할 수 있다.

21) Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.; 안재현, 『XAI 설명 가능한 인공지능, 인공지능을 해부하다』, 위키북스, 2020, 106면, 173면.

2) 영업비밀 유출 가능성

기업들은 AI 모델을 개발하고 성능을 향상시키기 위해 데이터를 수집, 전처리, 분석하고 이를 토대로 AI 모델의 종류, 입력변수, 설정변수(hyperparameter) 정하여 테스트 과정을 통해 지속적으로 AI 모델을 튜닝하는 과정을 거친다. 기업이 개발한 AI 모델에 대한 정보와 개발 과정을 통해 축적한 정보는 경쟁사와의 차별화를 가져올 수 있는 중요한 자산에 해당한다. 따라서 기업으로서는 AI 시스템에 관한 정보를 지식재산으로서 보호를 받을 필요가 있는데, 일반적으로 영업비밀의 형태로 관리한다.

그런데 AI 투명성 규제로 인하여 AI 시스템의 작동방식이나 입력변수 등 주요 정보가 외부에 공개될 경우 영업비밀이 유출될 가능성이 있다. AI 시스템에 대한 투명성 요구와 AI 시스템 정보의 영업비밀성 사이의 충돌에 대해, GDPR전문은 자동화된 의사결정 관련 이용자의 권리에 대한 예외 사유로 기업의 영업비밀 또는 지식재산권 보호를 들고 있다(recital 63). 비록 이용자의 정보 제공 요구에 대하여 기업의 영업비밀 항변이 항상 우선한다고 볼 수는 없겠으나, 적어도 둘 사이에 적정한 균형이 이루어져야 한다.[22)]

22) Article 29 Data Protection Working Party, "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation", WP251, 2018, p.17.; Kaminski, Margot E., "The right to explanation, explained". 34 Berkeley Tech. L.J. 189, 2019, p.203.

3) AI 시스템 남용 가능성

AI 알고리즘에 영향을 미치는 변수들 및 반영 방식에 대한 구체적인 정보가 공개될 경우, 이는 AI 시스템의 남용으로 이어질 수 있다. 가령 자동화된 개인신용평가의 경우, AI 시스템이 인터넷에 공개된 이용자들의 정보를 스크랩해서 평가에 반영할 수 있다. 그런데 이 경우 개인신용평가 결과가 특정 입력 값(예: 자신의 SNS 활동)에 상당히 민감하게 반응한다는 점이 이용자들에게 알려진다면, 신용도가 낮은 이용자들이 이를 악용하여 부당하게 개인신용평가 결과를 높이려는 사례들이 나올 수 있다.

위 문제는 기업이 AI 시스템의 완결성을 높임으로써 해결해야 할 문제이기는 하나, AI 시스템의 불투명한 특성으로 인해 기업도 사전에 정확히 인식하지 못할 수 있다. 결국 이용자에 의한 시스템 남용 행위는 시스템의 신뢰성 및 안정성을 훼손시키고, 기업들에게 이를 방지하기 위한 추가적인 비용을 발생시킬 수 있다.

라. 시장 자율 규제를 통한 투명성 확보

아직 AI가 기술적으로나 산업 적용에 있어 초기 단계에 있기 때문에 섣부른 규제는 국내 AI 산업의 성장동력에 정체를 야기할 수 있다. 따라서 AI 시스템에 대한 규제를 도입하지 않거나 최소한으로만 도입하고, 기업들에게 자율적인 규제를 독려하는 방안에 대해서도 가능성을 열어 두어야 할 것이다.

기업들은 AI 시스템의 투명성에 대한 소비자들의 요구를 의식하여 투명성을 자사의 이미지나 제품의 신뢰도 향상을 위한 수단으로 삼을 유인이 있다. 가령 자율주행차가 시장에 출시될 경우, 출시된 모델 중 차량의 동작을 포함하여 차량이 내리는 결정에 대해 이용자에게 왜 그러한 결정을 내렸는지를 설명하는 모델이 있다면, 그러한 기능이 없는 모델보다 시장에서 더 경쟁력을 가질 수 있다. 만약 이러한 유인을 통해 기업들이 자발적으로 AI 시스템의 투명성 확보를 위한 기준을 정하여 이를 준수하는 관행이 확산된다면 AI 투명성 규제의 도입이 필요하지 않을 수 있다.

다만, AI 시스템의 투명성 확보를 위한 시장 유인이 존재하거나, 기업간 경쟁이 이루어지기 위해서는 소비자들이 AI 시스템의 위험에 대해 인식하고 기업들에게 투명성을 적극적으로 요구하는 환경이 조성되어야 한다. 따라서 정부 및 시민사회에서 국민들이 AI 시스템을 이해하고 활용할 수 있는 'AI 리터러시'(AI literacy)를 가질 수 있도록 교육 체계를 마련하고, 공익 광고 등을 통하여 AI 윤리에 대한 국민 의식 고취를 위한 노력이 이루어져야 한다.

나아가 기업들이 자율적으로 AI 시스템에 대한 관리·감독 체계를 갖출 수 있도록 정부 차원에서 기업들이 관련 규제와 AI 윤리를 어떻게 이행하고 이를 검증하는지에 대한 구체적인 가이드라인을 작성하여 제공할 필요가 있다. 기업들이 자율적으로 AI 시스템 관리·감독 체계를 운영하면서 확인된 성과와 문제점들을 실증적으로 분석하고, 이 과정을 통해 AI 가이드라인의 내용을 국내 실정에 맞게 보완해 나감으로써, 향후 AI 시스템에 대한 강제성 있는 관리·감독 규제에 대한 도입이 필요한 상황이 오면, 이를 규제의 기초 자료로 활용할 수 있을 것이다.

2. 법제 마련 시 고려할 사항들 - 중장기 과제

가. 입법 방식에 대한 검토: 일반법 vs. 개별법

AI 투명성 규제의 입법 방식은 크게 ① GDPR, EU AI 규제법안, 미국 알고리즘 책임법안 등과 같이 AI 시스템에 공통적으로 적용되는 '일반법' 형태로 입법을 하는 방법과 ② 개정 신용정보법의 '자동화평가' 관련 조항과 같이 AI 시스템이 활용되는 분야별로 '개별법' 형태로 입법을 하는 방법으로 구분해 볼 수 있다.

AI 투명성 규제를 일반법 형태로 입법을 할 경우, AI 시스템의 이용자들에 대한 광범위한 보호가 가능하고, 통일적이고 체계적인 규범 체계를 마련할 수 있다는 장점이 있다. 또한, 이용자 입장에서 자신의 권리와 의무의 내용을 명확하게 이해할 수 있어 예측 가능성과 법적 안정성 확보가 가능하다.

그러나 과연 일반법을 통해 다양한 유형의 AI 시스템을 효과적으로 규율할 수 있는지에 대하여 의문을 제기하는 시각도 존재한다. AI 시스템은 유형별로 활용의 정도와 위험성이 다르고, 기존에 정립된 규제의 내용이 상이할 뿐만 아니라, 이용자들이 원하는 투명성의 내용도 다를 수 있다.

가령 온라인 플랫폼에서 이용자가 게재한 콘텐츠를 내부 기준에 따라 자동 삭제를 하는 경우, 해당 플랫폼이 뉴스 플랫폼인지 아니면 일반 전자상거래 플랫폼인지 여부에 따라 게재된 콘텐츠의 법적 보호의 필요성이 달라질 수 있고, 삭제 조치에 이의를 제기하기 위해 필요한 정보의 내용도 달라질 수 있다. 또한 기존 규제를 통해 이미 AI 시스템의 투명성을 어느 정도 확보할 수 있는 산업 분야의 경우, 이용자가 AI 시스템이 내린 의사결정의 '기준' 자체를 문제 삼기 보다는 AI 시스템에 입력된 '정보'를 확인하고 정정하려는 경우가 많을 수 있는 반면, 이러한 규제가 없었던 분야에 대해서는 이용자가 AI 시스템이 내린 의사결정의 '기준' 자체에 대해서도 의문을 갖고 설명을 요구하려고 할 수 있다.

현재 국내에서는 규제기관별로 AI 시스템을 규제하는 내용의 독자적인 입법안을 준비하면서, AI 투명성 규제의 내용도 법안에 따라 달라지는 경향을 보이고 있다. 이러한 입법 방향이 결과적으로 규제의 규범력 확보에 도움이 될 수 있지만, 동시에 입법이 이루어지는 분야와 그렇지 않은 분야 사이의 형평성이나 개별법 적용 경계의 불명확성으로 인한 법령 간 중복 규제 및 상호 모순 등의 문제점이 발생할 수 있다는 점을 유념할 필요가 있다. 또한 AI 시스템이 적용되는 분야별로 발생시킬 수 있는 위험과 그에 따라 요구되는 투명성의 내용이 달라질 수 있다는 점이 개별 법안에 반영되어야 할 것이다.

나. 개인의 권리 행사를 통한 방식

1) AI 시스템에 의한 의사결정 사실 고지 의무

AI 투명성 확보를 위한 권리 입법의 첫 단계로서, AI 시스템을 개발·구축하는 자로 하여금 이용자에게 의사결정이 AI 시스템에 의해 내려졌다는 사실이나 소통 상대방이 사람이 아닌 AI 시스템이라는 사실을 고지하도록 입법하는 방안을 고려해볼 수 있다. 이용자 입장에서는 의사결정이 사람에 의해서 이루어지는지 아니면 시스템에 의해 자동적으로 내려지는지에 따라 의사결정의 내용에 이의를 제기하거나 권리 구제를 받기 위해 고려해야 할 사항과 취해야 할 수단이 달라질 수 있다(예: 누구에게 문의하여 해결할 수 있는지, 권리 구제를 위해 어떠한 정보를 사전에 알아봐야 하는지). 따라서 의사결정이 AI 시스템에 의해 이루어진다는 사실을 이용자에게 고지하는 것은 이용자의 권리 행사에 있어 도움이 될 수 있다.

하지만 오늘날 AI 시스템의 불투명한 속성으로 인해 이용자에게 발생하는 위험은 이용자가 AI 시스템의 존재 자체를 몰라서 발생하는 것이라기보다는 주로 AI 시스템의 작동 과정에 대한 충분한 정보를 습득할 수 있는 통로가 막혀 있어서 AI 시스템이 의사결정을 내린 이유가 무엇이고 의사결정에 대하여 누구에게 어떠한 책임을 물을 수 있는지를 쉽게 알 수 없기 때문에 발생한다. 따라서 이용자에게 단지 AI 시스템의 존재 또는 의사결정이 AI 시스템에 의해 내려졌다는 사실을 고지하는 것만으로는 권리 구제라는 입법 목적을 충분히 달성할 수 있을지 의문이다.[23)]

따라서 AI 투명성 확보를 위한 권리 입법에 대한 논의는 궁극적으로 아래 설명하는 설명요구권, 이의제기권 등 이용자 권리 구제에 직접적으로 도움이 될 수 있는 입법 수단들에 대한 검토까지 이루어져야 할 것이다.

2) 설명요구권

설명요구권의 입법을 위해서는 우선 AI 시스템에 대한 '설명'이 구체적으로 무엇을 의미하는지를 명확히 해야 한다. AI 시스템에 대한 설명이 과연 무엇인지에 대해서는 다양한 연구가 존재하는데, 설명의 개념은 크게 (i) '모델' 관련 설명과 (ii) '개별 결정' 관련 설명으로 구분하여 볼 수 있다.[24)]

우선 모델 관련 설명은 개별 결정 또는 특정 입력 데이터와 상관없이 AI 시스템의 기계학습 모델에 대한 광범위한 정보를 제공하는 것을 의미한다. 모델 관련 설명으로는 다음 정보들이 있다.

① 구성 정보: 모델 학습의 목적, 모델의 종류(뉴럴 네트워크, 랜덤 포레스트 등), 모델 학습 전 사용된 변수들

② 학습 메타데이터: 모델 학습에 사용된 입력데이터에 대한 통계적·정성적 분석 자료, 입력데이터의 출처, 모델의 출력 또는 분류 값

③ 성능 지표: 모델의 예측 성능(주요 하위 집단에 대한 예측 성능 포함)

④ 근사화 모델: 기존 모델의 작동 원리를 단순화시킨 근사화된 모델

⑤ 처리 정보: 모델이 어떻게 학습, 테스트, 보완되었는지에 대한 정보

다음으로 개별 결정 관련 설명은 특정 입력 데이터와 관련된 설명을 의미한다. 개별 결정 관련 설명으로는 다음 정보들이 있다.

① 민감도 기반 설명: 입력데이터 중 어떤 값이 바뀌면 결과가 바뀌는지

② 사례 기반 설명: 학습데이터 중 본인의 입력데이터와 가장 유사한 데이터가 무엇인지

③ 인구통계 기반 설명: 본인과 유사한 결과를 받은 다른 사람들의 특성이 무엇인지

④ 성능 기반 설명: 시스템이 결과에 대하여 어느 정도 확신하는지, 본인과 유사한 다른 사람들이 잘못 분류된 비율이 평균보다 높은지 아니면 낮은지

23) 다만, 캘리포니아 주의 봇 공개 법(Bot Disclosure)이 규정하고 있는 바와 같이 상품, 서비스에 있어 고객을 유인하거나 투표에 영향을 끼치는 목적 등으로 챗봇 등이 사용되는 경우 등과 같이 그 대화 상대방이 AI 시스템이라는 점을 공개할 의무를 부과할 필요가 있는 상황도 존재할 수 있다. California Legislative Information 2018, SB-1001 Bots: Disclosure.

24) Edwards, Lilian & Veale, Michael, op. cit., pp.55-59

그런데 위와 같은 분류는 기술적으로 명료하다는 장점이 있지만, 각 정보가 개인이 권리를 행사하거나 법익을 보호하기 위해 어떠한 기능을 할 수 있는지가 명확하지 않다. 설명요구권은 개인의 권리 행사나 법익을 보호하기 위한 '도구'로서의 성격을 갖고 있으므로, 개인에게 제공되는 설명의 내용은 그 설명이 이용되기 위한 목적에 부합해야 한다.[25] 가령 이용자의 취향에 기반한 동영상 추천시스템의 경우, 이용자가 추천시스템이 자신의 어떤 정보를 사용하는지 궁금하다면 추천시스템의 입력변수 항목에 대한 설명이 도움이 될 수 있다. 하지만 AI 시스템에 의해 낮은 개인신용평가를 받은 개인이 평가 결과에 이의를 제기하기 위해서는 입력변수 항목이 무엇인지에 대한 설명보다는 입력데이터 중 어떤 값이 바뀌면 결과가 바뀌는지에 대한 설명이 더 의미가 있을 수 있다.

이처럼 AI 시스템이 이용되는 분야별로 출력하는 결과물이 다르고, 이용자에 따라 결과물에 대해서 제기하는 이의의 내용과 이를 위해 요구하는 설명의 내용도 다를 수밖에 없다. 따라서 상황별로 요구되는 다른 유형의 설명들을 어떻게 효과적으로 법제화할 수 있는지에 대한 고민이 필요하다. 가령 일반법 또는 개별법 형태 중 어떤 입법 방식이 바람직한지, 법을 통해 설명의 내용을 구체적으로 특정하는 것이 맞는지 아니면 법에서는 설명의 내용을 원칙 위주로 규정하고 세부적인 기준은 감독기관의 가이드라인을 통해 구체화하는 방안이 바람직한 것인지에 대해서 검토가 필요하다.

설명요구권의 범위와 관련하여, 이용자에게 필요한 정보는 AI 시스템의 알고리즘 소스 코드와 같이 해석을 위해 전문지식이 필요한 원자료(raw data)가 아닌 개인이 이해할 수 있고 권리구제에 활용할 수 있는 실용적인 정보이다. 따라서 설명요구권의 범위 내에 위와 같은 원자료를 포함시킬 필요는 크지 않을 것이다. 설명의 범위에서 원자료를 제외시키는 것은 기업의 영업비밀 보호 측면에서도 필요하다. 알고리즘 소스 코드와 같은 원자료는 검증이 필요한 경우 뒤에서 볼 AI 검증 절차 등 전문성이 갖춰진 AI 시스템 관리·감독 수단을 통해 검토가 이루어져야 할 것이다.

한편, 설명요구권을 법제화할 경우 의사결정이 오로지 AI 시스템에 의해서 이루어지는 경우가 아닌 인간의 개입이 일정 부분 이루어지는 자동화된 의사결정에 대해서도 규제가 적용되어야 할 것인지에 대한 검토가 필요하다. 최종적인 의사결정은 인간이 내리고 AI 시스템은 인간의 의사결정을 보조하는 역할에 그친다면, 인적 개입을 통해 AI 시스템에 대한 투명성과 책임성을 확보할 수 있으므로, 이러한 의사결정에 대해 별도의 규제를 적용할 필요성은 상대적으로 낮다고 볼 수 있다.

3) 이의제기권

'이의제기권'은 AI 시스템이 내린 의사결정의 결과 또는 내용의 정당성에 대해 개인이 이의를 제기할 권리를 의미한다. 이의의 내용은 AI 시스템이 의사결정을 내리기 위해 이용한 정보를 다른 정보로 변경하여 다시 의사결정을 내릴 것을 요구하거나 AI 시스템의 모델 디자인 요소(예: 데이터 전처리 방법)를 변경할 것을 요구하는 경우 등이 있을 수 있다. 이처럼 이의의 내용은 AI 시스템의 유형별로 다양할 수 있으므로 설명요구권의 경우와 마찬가지로 입법 방식에 대한 고민이 필요하다.

25) Selbst, Andrew D. & Powles, Julia, 앞의 논문, p.236.

다. AI 시스템에 대하여 관리·감독 체계를 구축하는 방식

1) AI 시스템 관리·감독 방안으로서 'AI 검증 절차'

AI 시스템에 대한 관리·감독 방안으로 AI 시스템에 대한 감사, 알고리즘 영향평가, 적합성 평가 등의 검증 절차(이하 "AI 검증 절차")들이 활발하게 논의되고 있다. 위 절차들은 이름과 세부적인 내용은 조금씩 다르지만 공통적으로 AI 시스템이 사회에 어떠한 위험을 발생할지 사전적·예방적으로 평가하고, 그 결과를 토대로 해당 시스템을 이용해도 되는지 판단하거나 이용하기 위해 어떠한 조치를 취해야 하는지 검토하는 역할을 한다.

해외에서 입법되거나 논의 중인 AI 검증 절차로는 대표적으로 ① GDPR의 자동화된 의사결정에 대한 '정보보호 영향평가'를 알고리즘 영향평가로 활용하는 방안, ② 미국 알고리즘 책임법안의 '자동화된 의사결정 시스템 영향평가', ③ EU AI 규제법안의 '적합성 평가' 절차가 있다. 위 절차들은 공통적으로 AI 시스템 전반에 대한 체계적인 기록 및 설명, 시스템의 위험성에 대한 평가, 위험을 최소화하기 위한 조치의 마련을 포함한다.

2) AI 검증 수행 방법

AI 검증 절차는 검증을 수행하는 주체에 따라 ① 기업이 스스로 내부 검증을 하는 방법, ② 기업으로 하여금 독립적인 외부 기관의 검증을 받도록 하는 방법, ③ 정부가 직접 기업을 상대로 검증을 하거나 자료 제출을 요구하는 방법이 있을 수 있다. 기업이 주도하여 검증을 수행하는 경우, 검증 과정에는 정부가 개입하지 않더라도, 정부가 기업에게 검증 결과 자료를 제출하도록 하여 규범력을 확보할 수 있다. GDPR, 미국 알고리즘 책임법안의 AI 검증 절차는 모두 원칙적으로 검증을 기업이 자체적으로 수행하고, 규제기관의 감독을 받는 형태를 취하고 있으므로, 위 ① 방법에 해당한다고 볼 수 있다.

AI 검증 절차는 AI 시스템의 작동 방식에 대한 정보뿐만 아니라 AI 시스템이 개인 및 사회에 미치는 영향을 아우르는 종합적인 평가가 필요하다. 이를 위해서는 AI 시스템의 처리 목적·구성요소·기능을 정확히 아는 자에 의해 시스템 개발 과정 전반에 대한 단계별 검토가 이루어져야 하는데, 앞서 본 ① 방법과 같이 AI 시스템을 운영하는 자가 직접 평가를 주도하는 것이 적합할 것으로 보인다. 외부기관이 주도하는 검증의 경우, 해당 외부기관은 AI 시스템에 대한 구체적인 이해가 부족할 수밖에 없기 때문에, 기업으로부터 AI 시스템의 학습 데이터나 프로그램 소스 코드를 제출 받더라도, 그 자료만으로는 AI 시스템이 어떠한 위험성을 가졌는지에 대해 정확히 평가하기 어렵다.

기업이 직접 검증을 수행하는 경우 기업에게 부족한 전문성은 외부 전문기관이나 규제기관의 자문을 통해 보완할 수 있을 것이다. 그리고 AI 시스템으로 인하여 중대한 위험이 발생하는 예외적인 경우에 대해서는 규제기관이 직접 AI 시스템에 대해 검증을 할 수 있도록 하는 방안을 고려해 볼 수 있을 것이다. 가령 EU AI 규제법안 제23조는 고위험 AI 시스템을 운영하는 자로 하여금 규제기관 요청 시 법안의 기준을 충족하고 있다는 점을 입증하는 정보와 문서를 제출할 의무를 부여하고 있다.

3) AI 검증 절차의 역할

기업들이 AI 시스템 이용의 책임성과 투명성을 확보하기 위해서는 내부적으로 ① 구성원들에 대한 명확한 역할 및 책임 부여, ② 잠재적 위험의 검토 및 대응, ③ 이행 사실의 기록 및 입증을 기본 원칙으로 하는 거버넌스를 마련하여야 한다. AI 검증 절차는 기업들이 내부적으로 수립한 거버넌스를 충실히 이행했는지 검증하는 역할을 한다. 따라서 AI 검증 절차를 법제화하기 위해서는 내부 거버넌스 수립에 관한 사항도 함께 입법이 이루어져야 한다.

AI 시스템에 관한 내부 거버넌스와 이에 대한 검증 절차는 기업 입장에서 단순히 주어진 규제 체크리스트를 형식적으로 확인하는 절차가 아니라, AI 시스템이 어떠한 위험을 발생시킬 수 있는지 진지하게 고민하고 이에 대한 대응조치를 마련하는 등 AI 시스템을 책임감 있게 개발·구축했다는 사실을 입증할 수 있는 수단이 되어야 한다. 따라서 기업의 내부 거버넌스에는 기업의 AI 시스템 개발·구축 과정을 문서화하고 이를 체계적으로 관리하는 절차가 포함되어야 한다. 이를 위해 기업으로 하여금 AI 시스템을 위해 이용하는 데이터와 알고리즘 모델에 대한 상세한 정보를 기록하도록 하는 방안을 고려해볼 수 있다.

한편 기업들이 내부적으로 거버넌스를 수립하고 검증을 수행하기 위해서는 구체적인 관리·감독 기준과 가이드라인이 마련되어야 한다. AI 시스템을 개발·구축하는 상당수 기업들은 전문성 및 인적 자원의 부족으로 인하여 독자적인 내부 거버넌스를 수립하기가 어려울 수 있으므로, 정부 차원에서 기업 실정에 맞는 관리·감독 기준과 가이드라인을 마련할 필요가 있다.

4) AI 검증 내용의 공개 방안 검토

AI 검증 절차의 실효성 강화를 위해, 검증 결과를 규제기관에 제출하거나 일반에 공개하는 절차를 두는 방안을 고려해볼 수 있다. 규제기관에 대한 제출 또는 일반에 대한 공개 절차를 둘 경우, 기업들이 AI 검증을 적절히 수행하였고 검증 결과를 AI 시스템에 반영했는지에 대해 외부에서 확인할 수 있으므로, AI 검증 절차가 무의미한 체크리스트가 될 위험을 방지할 수 있다. 그리고 기업들이 AI 시스템의 이용으로부터 발생할 수 있는 위험이나 문제점들을 외부와 공유함으로써, 이를 사회 구성원들이 함께 논의하고 해결해 나갈 수 있는 기회를 제공할 수 있다.

다만 AI 검증 결과의 제출 또는 공개 제도를 도입하더라도, 기업의 영업비밀 침해 및 시스템 남용 우려 등을 감안하여 기업이 검증 결과를 요약하여 주요 내용만을 규제기관에 제출 또는 일반에 공개하거나, 검증 결과를 제출받은 규제기관이 검증 내용을 요약 정리하여 보고서 형태로 일반에 공개하는 방안 등 기업의 부담을 최소화할 수 있는 방안이 마련되어야 할 것이다.

향후 다양한 AI 시스템 공통적으로 활용할 수 있는 표준화된 관리·감독 기준이 마련된다면, AI 검증 결과를 토대로 AI 시스템에 인증을 부여하는 표준 인증체계 제도를 도입하는 방안도 고려해 볼 수 있을 것이다.

Ⅴ. 마치며

최근 국내외로 AI 투명성 확보를 위한 규제의 도입 가능성에 대한 연구 및 논의가 활발하게 이루어지고 있다. 특히 EU는 일찍이 GDPR에 AI 투명성 확보를 위한 규정을 두고, 최근 고위험 AI 시스템에 대한 적합성 평가 절차 등을 주요 내용으로 하는 AI 규제법안을 발표하는 등 AI 투명성 규제에 관한 논의를 주도하고 있다. 국내에서도 AI 투명성 규제의 도입에 대한 논의가 점차 활성화되어 가고 있지만, 국내 실정에 맞는 규제를 마련하기 위한 실증적 연구가 아직 부족하다. 국내에서 AI 기술이 적극적으로 적용되고 있는 분야·서비스가 무엇인지, 해당 분야·서비스에서 AI 시스템이 실제로 위험을 발생시키고 있거나 발생시킬 우려가 있는지, AI 투명성 확보를 위한 시스템 구축 및 운영이 기업에 어떠한 부담을 발생시키는지 등에 대한 구체적인 연구가 필요한 상황이다. 따라서 AI 투명성 확보를 위한 규제를 준비하기에 앞서, 위와 같이 AI 시스템 이용 환경 및 규제 도입의 필요성에 대한 연구가 선행되어야 할 것이다.

그리고 이와 함께 정책적으로 기업들이 스스로 AI 투명성 확보를 위해 노력하도록 유도해야 할 필요가 있다. 이를 위해 정부에서 기업들에게 AI 시스템을 개발·구축하는 과정에서 관련 규제 및 AI 윤리를 어떻게 이행하고 이를 검증하는지에 대한 구체적인 가이드라인을 마련하여 제공한다면, 기업 내부적으로 AI 투명성 확보를 위한 정책을 수립하는 데에 도움이 될 것으로 생각한다. 기업들이 자율적으로 AI 시스템 관리·감독 체계를 운영하면서 확인된 성과와 문제점들을 실증적으로 분석함으로써, 이를 AI 투명성 규제 마련의 기초 자료로 활용할 수 있을 것이다.