ISSN 2005-8349 GRI연구논총 2024년 기획호 DOI : 10.23286/gri.2024.26.5.002

AI 안전을 위한 규제와 거버넌스*

김 윤 명**

AI 안전은 AI를 개발하거나 이용하는 과정에서 나타날 수 있는 여러 위해 상황이나 문제로부터 국민의 안전을 보장하기 위한 정책목표이다. 지금까지 AI 윤리를 통해 안전성을 확보하기 위한 자발적인 노력을 요구해 왔다. 그렇지만, 이제는 EU 「AI법」과 같이 윤리원칙을 넘어서 규제중심으로 전환하고 있다. AI 안전을 달성하기 위한 방법론은 다양하다. 제도적이거나 기술적으로 안전을 위한 여러 가지 사항을 구체화할 수 있기 때문이다. 무엇보다, AI를 이용하는 과정에서의 투명성 확보이다. 결과에 대한 편향 없는 공정성을 확보하는 것이다. AI 안전을 통해 얻을 수 있는 사회적 가치는 국민의 안전과 AI에 대한 신뢰성이다. 이를 위하여, 무과실 책임과 입증책임의 전환은 실효적인 안전 확보방안이 될 수 있다. 정부는 AI 안전을 위한 구체적인 정책목표를 제시하여야 한다. 규제지향적인 목표를 수립할 것인지, 원칙중심의 자율규제적 목표를 수립할 것인지는 AI 안전에 대한 사회적 합의에 따라 달라질 수 있다. AI 안전의 목표이자 과제는 AI를 활용하는 과정에서 나타날 수 있는 여러 가지 문제 상황에서 국민의 안전을 담보하는 것이기 때문이다. 기술중립적인 관점에서 기술에 규제가 아닌 비즈니스 모델에 대한 규제일지 구체화 된 논의가 이루어져야 한다. 이를 위해서는 시민사회의 지속적인 모니터링이 필요하다. 기술이라는 것이 선의로 개발된 것이기는 하지만, 이를 이용하는 과정에서 오남용은 예기치 않게 나타날 수 있기 때문이다. AI 위험 관리를 위한 프레임워크(framework)의 검토는 AI 위험관리와 안전성 평가에서 의미있는 역할을 할 것으로 기대한다. 아울러, AI 안전을 담보할 수 있는 제도적, 기술적 수단으로서 설명요구권과 설명가능한 AI(XAI)는 블랙박스 효과에 따른 위험성을 줄일 수 있을 것이다. 무엇보다, 안전을 위한 거버넌스는 민관의 협력과 지속적인 모니터링을 수행하는 시민사회 등이 참여함으로써 보다 구체화할 수 있다. 아울러, 지속적이고 일관된 안전 정책을 유지하기 위해서 AI 법제의 입법이 뒷받침될 필요가 있다. 기술발전에 따른 안전은 선택이 아닌 필수이며, 미지의 기술인 AI는 더욱 그러하다.

주제어 _ AI 안전, 규제와 혁신, AI 거버넌스, 국가인공지능위원회, 안전연구소, 시민사회의 모니터링

* "가을이 깊으니 겨울과 맞닿는다. 이 아침 무서리 내려앉은 겨울뜨락의 국화는 가을뜨락의 그 국화였다. 심사위원분의 서릿발같은 지적과 따스한 의견에 진심으로 감사드린다(2024.11.20.). 이 글은 법제연구원 이슈페이퍼인 "AI 안전 확보를 위한 법적 과제"의 일부를 논문화한 것이다."

** 전남대학교 데이터사이언스대학원, 법학박사

ISSN 2005-8349 GRI REVIEW 2024 DOI : 10.23286/gri.2024.26.5.002

Regulatory Governance for AI Safety

Kim Yun-myung*

AI safety is a policy goal to ensure the safety of the people from various harmful situations or problems that may appear in the process of developing or using AI. Until now, voluntary efforts have been required to secure safety through AI ethics. However, as in the EU AI Act, it is shifting beyond ethical principles to a regulatory focus. There are various methodologies for achieving AI safety. This is because various matters for safety can be embodied systematically or technically. Above all, it is securing transparency in the process of using AI. It is to ensure fairness without bias against results. The social value that can be obtained through AI safety is the safety of the people and the reliability of AI. To this end, the transition between no-fault liability and the burden of proof can be an effective safety measure.

The government should present specific policy goals for AI safety. Whether to establish a regulation-oriented goal or a principle-based self-regulatory goal may vary depending on social consensus on AI safety. This is because the goal and task of AI safety is to ensure the safety of the people in various problem situations that may arise in the process of using AI. From a technology-neutral point of view, a concrete discussion should be made on whether or not the business model is regulated by technology. For this, continuous monitoring of civil society is necessary. This is because although technology was developed in good faith, misuse in the process of using it may appear unexpectedly. The review of the framework for AI risk management is expected to play a meaningful role in AI risk management and safety evaluation. In addition, as an institutional and technical means to ensure AI safety, the right to request explanation and explainable AI(XAI) will be able to reduce the risk of the black box effect. Above all, governance for safety can be more concrete by the participation of public-private cooperation and civil society that conducts continuous monitoring. In addition, legislation of AI Act needs to be supported in order to maintain a continuous and consistent safety policy. Safety according to technological advancement is a must, not an option, and this is especially true for AI, an unknown technology.

Key words _ AI safety, regulation and innovation, AI governance, National Artificial Intelligence Commission, AI Safety Research Institute, Monitoring of Civil Society

* Graduate School of Data Science, Chonnam National University, Ph.D. in Law.

Ⅰ. 서 론 : AI 안전을 위한 문제제기

2023.11월, 영국 블레츨리에서 인공지능(artificial intelligence, 이하 'AI'라 함) 안전에 대한 글로벌 논의인 AI Safety Summit이 있었고, 그에 따른 블레츨리 선언(Bletchley declaration)이 이루어졌다.[1] 2024. 5월, 서울에서 블레츨리 선언의 후속 논의인 AI Seoul Summit이 개최되었다.[2] 2024년 EU「AI 법」이 발효되었고, 우리나라를 포함하여 많은 나라에서 AI 관련 법률과 글로벌 논의를 통해 AI가 가져올 수 있는 문제에 대한 안전 확보를 위해 노력하고 있다.

AI를 포함한 디지털 기술과 서비스는 안전을 위협하지 않고 신뢰할 수 있어야 한다. 따라서, 디지털 위험에 대비하는 수단과 절차가 마련되어야 한다. AI가 사회·경제적으로 응용되는 분야가 확대되고 있으며, 효용성이 높아지고 있는 것도 사실이다. 그렇지만, 그로 인하여 발생하는 문제의 파급력이 작지 않을 경우에 그에 대한 책임 논의는 커질 것이다. 또한, AI의 결함이나 오류(誤謬) 및 그에 따른 확대된 문제가 발생할 수 있기 때문에 이러한 경우에는 관련된 자에게 어떠한 책임을 물을 수 있는 지에 대한 검토가 요구된다.

현재의 기술 수준에서 인공지능 오류는 SW가 가지고 있는 결함과 크게 다르지 않다. AI의 속성상 SW를 기본적인 요소로 하기 때문이다. 그렇지만, 기계학습(machine learning)의 결과로 나타날 수 있는 오류는 SW 코드에서 발생하는 오류와는 차이가 있다. 블랙박스(black box)에서 발생하는 오류에 대해서는 어떠한 책임을 물을 것인지 쉽지 않다. 개발자들도 AI가 내린 결정의 이유를 알 수 없다고 하기 때문이다. 다만, 인공지능의 오류에 따른 문제해결 과정에서 고려할 수 있는 것은 블랙박스에 대한 지배영역이 누구에게 있는지가 하나의 기준이 될 수 있을 것이다.

AI 안전은 SW로서 AI가 구현되는 과정에서 나타날 수 있는 안전 이슈를 관리가능한 상태로 유지하는 것이다. 궁극적으로는 SW로서 AI 시스템이 안전하게 이용될 수 있다는 이용자의 신뢰를 확보하고 유지하는 것이다. 이를 뒷받침하기 위한 법제도 및 거버넌스 체계 수립에 대해 살펴보고자 한다.

---

1) GOV.UK, "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023", (November 1, 2023), https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023, (2024.6.1. 방문).

2) 대한민국 정책브리핑, "대한민국, '서울 선언'을 통해 국제 인공지능 협치(거버넌스)의 새로운 방향 제시", (2024.05.23.), https://www.korea.kr/briefing/pressReleaseView.do?newsId=156632066, (2024.6.1. 방문).

Ⅱ. AI 안전이 추구하는 가치

1. AI 안전의 필요성

AI의 기본적인 속성은 SW이지만, HW적으로 SW의 성능을 높일 수 있다는 점에서 HW로서 AI는 무시하기 어렵다. 「소프트웨어 진흥법」에서는 소프트웨어를 "컴퓨터, 통신, 자동화 등의 장비와 그 주변 장치에 대하여 명령·제어·입력·처리·저장·출력·상호작용이 가능하게 하는 지시·명령(음성이나 영상정보 등을 포함한다)의 집합과 이를 작성하기 위하여 사용된 기술서(記述書)나 그 밖의 관련 자료"로 정의한다. 우리나라는 AI에 대한 정의나 이를 규율하는 법률은 없으나,[3] EU 「AI법」에서는 AI시스템을 "배포 이후에 적응성을 보이고 명시적 또는 묵시적 목표를 위하여 물리 환경이나 가상 환경에 영향을 미칠 수 있는 예측, 콘텐츠, 권고나 결정 등의 산출물을 생성하는 방법을 입력을 통하여 추론할 수 있는 다양한 수준의 자율성을 가지고 작동하도록 설계된 기계 기반 시스템"으로 정의하고 있다.[4] 당초 초안에는 SW로서 AI를 정의하였으나, AI가 구현되는 것은 SW만이 아닌 다양한 HW를 포함한 컴퓨팅 능력(computing power)이 융합된 것이기 때문에 정의를 수정한 것으로 보인다. 이처럼, EU 「AI법」에서는 AI에 대해 시스템으로 정의하고 있다는 점은 시사하는 바가 크다.

인공지능을 포함한 다양한 과학기술에 대한 의존도가 높아지고 있으며, 기술에 대한 신뢰도가 높아질수록 '기술 의존도'는 더욱 커진다.[5] 문제는 예기치 못한 상황에서 사고가 발생할 수 있다는 점이다. 일례로, SW 의존도가 높아질수록 SW로 인한 사소한 실수는 결과적으로 대형 사고의 원인이 되거나 치명적인 결과로 나타나고, 재산상 손해를 포함한 인명 피해도 비례하여 커질 것은 자명하다. AI 모델의 대표인 거대 언어 모델(large language model, 이하 'LLM'이라 함)의 한계는 결과에 대해 명확하게 설명하지 못하고 있다는 점이다. 딥러닝 알고리즘은 문제해결능력은 뛰어나지만, 심층신경망은 계층이 많아 확률적 판단이 여러 번 중첩되기 때문에 판단 수식을 설명해도 사람이 이해할 수 없다. 따라서, 의사결정이 심층신경망 내부에서 어떤 매커니즘으로 도출하였는지 설명하기 어렵기 때문이다.[6] 단지 결과에 대한 추론을 통해 원인을 짐작할 수 있을 뿐이다. 인간의 언어를 이해하는 LLM이지만, 실제

3) 현재, 국회 의안정보 사이트에는 15개의 인공지능 관련 법안이 발의되어 있는 상태이다.
4) EU 「AI법」 제3조 (1).
5) 무스타파 술레이만, 『더 커밍 웨이브』, 한스미디어, (2024), p.9. p.19 참조.
6) 이수호, 『AI 인사이트』, 한빛비즈, (2022), p.351.

내부 알고리즘의 처리과정에 대해서는 인간의 언어로 설명하지 못한다.[7] 이러한 AI 모델의 한계를 극복하기 위한 기술적 방안으로써 설명가능한 AI가 제안된 바 있다. 안전이란 측면에서 그 결정을 신뢰하기 위해서는 명확한 이유를 이해할 수 있어야 하기 때문이다. 그러한 이해는 설명가능한 AI 모델을 통해서, 구현될 수 있다. 단순한 설명가능한 AI의 개발이 아닌, AI 안전을 위한 연구자의 기술적 과제이다.[8] 아울러, AI 생성물에 대한 신뢰성 문제도 해결되어야 한다. 환각현상, 헛소리, 딥페이크, 왜곡된 결과, 편향된 결과는 AI 모델이 인간에게 제시하는 결과물의 내용에 관한 것이다. 문제는 AI가 반복적이고 일관되게 생성하는 것은 그 자체에 대한 신뢰성을 확보할 가능성도 높다는 점이다. 사람도 반복적이고 일관되게 어떤 주제에 대해 얘기할 경우, 그 내용에 대한 사실여부를 떠나 믿을 가능성도 있다. 조작적으로 AI를 운영하는 경우에는 결과를 왜곡할 수 있다. 따라서, AI로 인한 문제점들을 해결하고 안전성을 담보하기 위해서는 AI가 생성하거나 의사결정한 결과에 대한 책임논의가 구체화될 필요가 있다.

2. 구체적인 AI 안전

국민의 안전 보장은 국가의 책무이다. 다만, 헌법적 논의에서 안전권은 파생적 기본권으로 인식되고 있다.[9] 더욱이, 안전에 대한 개념을 정의하고 있는 법률도 찾기 어렵다. 재난안전에 관한 기본법이라고 할 수 있는 「재난안전기본법」에서도 안전에 대한 정의는 없다. 참고할 수 있는 개념은 「안전기본법(안)」[10]에서 찾을 수 있다. 동 법안에서 안전이란 "재난이나 그 밖의 각종 사고로부터 사람의 생명·신체·재산 및 국가에 위험이 없는 상태"로 정의하고 있다. 참고로, 국가안전관리 기본계획에서는 안전이란 "위험요인이 없거나 이러한 위험요인에 대한 충분한 대비가 되어 있는 상태"[11]로 보고 있다. 따라서, 국민이 안전한 상태에서 일상생활을 영위할 수 있는 상태를 유지해야 하는 것이 안전의 목표이다.

AI가 안전하다는 것은 AI가 기능적으로 작동하는 상태를 의미하며, 기술적 오류나 결함이 없거나 또는 결함이 내재하더라도 대응이나 수용가능한 수준이어야 한다는 의미이다. 따라서, AI로 인한 사고가 수용가능한 범위를 넘어선 경우라면 안전한 상태로 보기 어렵다. 아울러, 기술적 실업이나 양극화와 같은 사회적인 측면에서의 안전도 같이 살펴보아야 한다. 기술적 안전에 치중한 나머지, 사회적 안전을 등한시 할 경우에는 양극화나 기술적 실업등 등 사회체제에 대한 위협을 가져올 수 있다. AI 안전은 기

7) 이수호, 『AI 인사이트』, 한빛비즈, (2022), p.351.
8) Id., p.411.
9) 김윤명, 「소프트웨어 안전을 위한 입법정책 방안」, 홍익법학 25(2), (2024), p.283.
10) 오영환의원 등 16인 발의, 안전기본법안, 의안번호 2105198, (2020.11.11.).
11) 행정안전부, 「국가안전관리기본계획[2010-2014]」, (2010), https://www.mois.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_00120084aswC0Wy&fileSn=0, (2024.6.1. 방문).

술적 측면과 사회적 측면을 함께 갖추어야할 목표로서 설정되어야 한다. 기술적 측면에서의 안전은 AI 결함이 관리가능한 상태로 유지되는 것이다. 따라서, 기술적 측면에서 AI 안전은 운용자의 실수, HW, SW 고장이 발생하더라도 확대사고로 이어지지 않도록 HW, SW 설계와 개발 시에 안전 기능을 추가하여 확보한 상태로 이해된다.

구체적인 AI의 결함에 따른 AI 안전을 이해하기 위해 제조물책임법상의 결함 유형에 따라 살펴보고자 한다. 먼저, AI 결함은 제조, 설계, 표시상의 결함으로 AI가 안전하지 못하게 된 경우로, 원래 의도했던 바대로 기능하지 않는 상태를 말한다. 제조 결함은 AI 개발 시 원래 설계와 다르게 제작된 경우이며, 설계결함은 오류 등을 줄일 수 있는 알고리즘을 고려하지 않은 코딩을 의미한다. 표시결함은 AI를 제공하는 과정에서 합리적인 설명이나 경고 등을 하지 않는 경우를 말한다. 이러한 경우를 소비자가 '통상적'으로 기대했던 수준으로 작동되지 않는 것으로 볼 수 있다. 따라서, AI 안전은 AI가 내린 결론에 대해 신뢰할 수 있다거나 차별이나 편향적이지 않고 공정성을 담보할 수 있는 사회적인 가치판단의 영역으로 볼 수 있다.

3. AI 안전의 가치 : 신뢰성과 국민의 안전 보장

1) 다양한 책무의 집합으로서 AI 안전

AI가 가져오는 다양한 문제에 대한 책무로서, 사업자, 개발자 및 이용자의 책임과 의무에 대해 살펴볼 필요가 있다. AI는 개발 과정에서의 책임성, 공정성, 신뢰성을 확보할 수 있어야 한다. 따라서, 기계학습 과정에서 사용되는 데이터의 편향, 저작권 침해, 데이터 윤리 등 다양한 문제를 해결할 수 있어야 한다. 그 과정에서 절차적인 공정성도 확보할 수 있어야 한다. 아울러, 생성된 결과물이 사회적 가치를 훼손하지 않는 상태로 유지될 수 있어야 한다. AI에 어떤 역할을 부여할 것인지, 사회적으로 어떠한 기여를 할 수 있을 것인지 등 구체적인 고민 없이 개발하거나 서비스를 제공할 경우, 신뢰성을 확보하기 어려울 것이기 때문이다. 또한, 이용자의 책무도 중요하게 다루어질 필요가 있다. 개발자가 의도했던 바대로 이용해야 예기치 못한 사고나 위험으로부터 안전성을 확보할 수 있어야 하기 때문이다.[12] 이를 위하여, 국가와 서비스제공자는 이용자 교육을 포함한 AI 리터러시(literacy)가 확보될 수 있도록 노력해야 한다.

AI 안전은 다양한 사회적 이슈를 관리가능하거나 수용가능한 상태로 유지할 수 있을 경우에 가능하다. AI 안전은 기술적 안전 그 자체의 논의와 더불어, AI가 가져오는 다양한 정치, 경제, 사회, 문화 영

12) 시스템 자체의 문제라면, 강제적 셧다운이나 비상조치가 내재되어야할 것이다.

역에서의 안전까지도 논의되어야 하는 이유이다.

2) 무과실 책임과 입증책임의 전환

기술의 발전과 시대 상황의 변화에 따라 법률이 추구하는 가치도 변하게 마련이다. 일례로, 「제조물책임법」은 피해자의 입증책임을 제조자가 과실이 없음을 밝혀야 하는 무과실(無過失) 책임으로 전환시켰다. 제조자는 손해의 발생 원인이 자신에게 없다는 입증하지 못할 경우에는 그 결함으로 인한 손해배상책임을 진다. 제조물에 대한 엄격한 책임을 제조자에게 부과함으로써, 그 만큼 안전한 사회를 구현하기 위한 시대적 요구가 법률에 반영된 것이다.

3) AI 안전의 가치로서 신뢰성

인공지능 신뢰성이란 데이터 및 모델의 편향, 인공지능 기술에 내재한 위험과 한계를 해결하고, 인공지능을 활용하고 확산하는 과정에서 부작용을 방지하기 위해 준수해야 하는 가치 기준을 말한다. 주요 국제기구를 중심으로 인공지능 신뢰성을 확보하는 데 필수적인 요소가 무엇인지 활발한 논의가 이루어지고 있다. 일반적으로 안전성, 설명가능성, 투명성, 견고성, 공평성 등이 신뢰성을 확보하는 데 필수적인 요소로 거론되고 있다.

〈표 1〉 인공지능 신뢰성의 속성

속성	의미
안전성(safety)	인공지능이 판단·예측한 결과로 시스템이 동작하거나 기능이 수행됐을 때 사람과 환경에 위험을 줄 가능성이 완화 또는 제거된 상태
설명가능성(explainability)	인공지능의 판단·예측의 근거와 결과에 이르는 과정이 사람이 이해할 수 있는 방식으로 제시되거나, 문제 발생 시 문제에 이르게 한 원인을 추적할 수 있는 상태
투명성(transparency)	인공지능이 내리는 결정에 대한 이유가 설명 가능하거나 근거가 추적 가능하고, 인공지능의 목적과 한계에 대한 정보가 적합한 방식으로 사용자에게 전달되는 상태
견고성(robustness)	인공지능이 외부의 간섭이나 극한적인 운영 환경 등에서도 사용자가 의도한 수준의 성능 및 기능을 유지하는 상태
공정성(fairness)	인공지능이 데이터를 처리하는 과정에서 특정 그룹에 대한 차별이나 편향성을 나타내거나, 차별 및 편향을 포함한 결론에 이르지 않는 상태

출처 : TTA, (2022)

4) 국민의 안전 보장

국민의 안전을 위한 노력은 국가의 책무이다. 이러한 점에서 AI의 안전은 AI 자체의 안전을 넘어서, AI를 이용하는 국민의 안전을 위한 것이라는 점을 명확히 하여야 한다. 위협받는 내용은 차별, 편향에

따른 공정성, 채용 및 다양한 전문적인 영역에서의 투명성, 해킹에 따른 시스템의 견고성 및 안전성 등이다. 또한, 사람의 대체와 같은 기술적 실업 및 양극화, AI를 학습시키면서 허락없이 또는 위법하게 수집하거나 이용하는 개인정보나 저작권 등의 데이터 윤리도 안전의 문제이다. AI가 가져오는 문제가 기본권이나 인권을 침해하거나 위협할 소지가 있는 경우에는 국민의 안전을 위협하는 것으로서 규제되어야 한다. EU「AI법」의 금지되는 AI나 고위험 AI는 EU의 가치나 기본권을 훼손하는 경우에는 제한할 수 있다는 점을 명확히 하고 있다.[13)]

Ⅲ. AI 안전을 확보하기 위한 규제원칙

1. AI 규제의 필요성

1) 인간에 대한 이해의 필요

AI가 인간의 삶에 영향을 미치고, 의사결정을 대신함으로써 인간은 주체적인 삶이 아닌 기계에 의존하는 삶을 살아갈 가능성이 높아지고 있다. AI가 더 나은 인간의 삶을 보장할 수 있다는 믿음과 더불어 각종 폐해와 함께 장기적으로 인간을 대체할 것이라는 우려가 동시에 제기되고 있다.[14)] 그렇지만, 인간이 인간으로서 존재하기 위해서는 가치, 철학, 안전, 신뢰라는 측면에서 인간을 바라볼 수 있어야 한다. 이 것이 AI 연구가 인간에 대한 이해가 필요한 이유이다. 또한, AI를 이용하는 과정에서 인간의 가치를 존중할 수 있어야 한다. AI의 사용에 따른 인간의 물리적, 사회적, 정신적 안전 상태가 유지될 수 있도록 하여야 한다. 인간을 위해 AI가 제대로 작동하기 위해서는 AI 서비스가 이루어지는 과정은 물론, AI 모델을 위한 데이터 거버넌스를 구축함으로써 데이터 편향이 이루어지지 않도록 해야 한다. 이를 위하여 데이터, 알고리즘에 대한 공개와 재현가능성을 확인할 수 있어야 하고, AI가 내린 의사결정에 대해 설명할 수 있어야 한다. 무엇보다, AI 서비스는 이용자인 국민이 이용함에 있어서 안전해야 하고 결과에 대해 신뢰할 수 있어야 한다. 그러한 가치를 담아내기 위해서는 AI기술이나 서비스가 지향하는 점을 명확히 하여야 한다. 인간을 위한 것이냐 또는 인간을 수단화하는 것이냐에 따라 달라질 수 있는 가치이기 때문이다.

---

13) EU「AI법」제77조(기본권을 보호하는 기관의 권한) 등 참조.
14) 류현숙, 『인공지능 기술 확산에 따른 위험 거버넌스 연구』, 한국행정연구원, (2017.12), p.89.

2) AI 안전의 확보

AI 안전을 담보하기 위한 방안으로써 규제는 명확하여야 한다. 사업자에게 예측가능한 형태로 제시되어야 하며, 그렇지 않을 경우에는 사업 영위에 어려움이 따르기 때문이다. 먼저, AI를 개발하는 과정에서 명확하게 기록을 남겨져야 한다. 기록으로 남긴다는 것은 향후에 발생할 수 있는 문제에 대한 원인을 파악할 수 있는 수단으로 활용할 수 있기 때문이다. AI 문제가 블랙박스라는 점에서 그 원인을 파악할 수 없다는 점이 반영된 것으로 생각된다. 또한, 서비스에 대한 평가를 통하여 일정한 조건을 충족하지 못할 경우, 소비자에게 제공되지 못하도록 하거나 또는 서비스 자체를 차단해야 한다.[15] 현재로서는 AI 기술이 인간의 의지와 같은 능력이나 스스로 생각하여 결정하는 수준은 아닌 것으로 보인다.[16] 결국, 인간에 의하여 시작되고 대략적이나마 인간의 지시·명령에 따라 이루어지기 때문이다.

안전을 위해 AI를 구축하여 서비스로 제공하는 사업자는 다양한 시도를 하면서, 위험요소를 줄이는 노력을 하고 있다. 그럼에도 불구하고, AI 자체가 갖는 블랙박스라는 속성을 극복하기는 쉽지 않다. AI 모델 내부적으로 처리되는 방식에 대해 개발자조차도 명확하게 설명하지 못하기 때문이다. 알고리즘에 대한 설명의무를 요구하는 이유이기도 하다. 이를 뒷받침하기 위하여 GDPR이나 우리 개인정보 보호법에 명시적으로 정보주체의 권리로서 설명요구권 및 알고리즘 적용 거부권 등이 규정되어 있다. 아울러, AI의 이용과정에서의 투명성이나 공정성을 확보하기 위한 다양한 노력과 의무를 서비스제공자에게 부과될 필요가 있다. 주의의무를 제공자에게 부여함으로써 자율적인 규제형태를 통하여 책임을 다하도록 유도할 수 있을 것이다. 이러한 노력과 의무는 AI가 가져올 수 있는 불안전성에 대한 억제(containment)이며, 여기에는 규제, 기술의 안전성, 새로운 거버넌스와 소유권 모델, 새로운 방식의 책임성과 투명성이 포함된다. 다만, 이 모든 것은 더 안전한 기술을 위한 필요조건이지 충분조건은 아니라는 점이다.[17] 챗GPT 이후로 디지털 전환의 핵심 요소가 되고 있는 생성형 AI는 다양한 위험을 증대시킬 것으로 보인다. 예를 들면, 환각현상이나 저작권 침해물에 대한 책임, 대출이나 채용 등에 사용됨으로써 기본권을 침해하는 프로파일링, 차별이나 편향된 결과물, 오류나 결함으로 인한 침해사고도 문제이다.

이용하는 사람간의 격차도 문제이다. AI 격차(AI divide)를 해소하는 정책이 수립되어야 하는 이유이다. 앞으로는 사람과 AI의 경쟁이 아닌 AI를 사용하는 사람과 그렇지 않은 사람간의 경쟁이 될 것이며, 그에 따른 격차도 커질 것이기 때문이다. AI 문해력(AI literacy)의 확산이 없을 경우 AI 양극화는 더욱 커질 것임은 자명하다. 이러한 상황을 인식하여 EU 「AI법」에서는 AI 문해력과 관련하여 AI 서비스

15) 김윤명, 『블랙박스를 열기 위한 인공지능법』, 박영사, (2022.01), p.177.
16) 서울고등법원 2024.5.16 선고 2023누52088 판결.
17) 무스타파 술레이만, 『더 커밍 웨이브』, 한스미디어, (2024.01), p.71.

제공자에게 구체적 의무를 부과하고 있다.[18] 또한, AI가 개발자나 서비스제공자의 의도대로 이용될 수 있도록 이용자의 책무도 강조되어야 한다. 대표적으로, 딥페이크(deepfake) 문제는 이용자의 악의적인 이용으로 나타나는 사회문제이기 때문이다. 정작 AI 자체의 윤리보다 인간의 윤리가 먼저 강조되어야 하는 이유이기도 하다.

3) AI 규제 방향

미국, EU, 중국의 규제 수준은 상이하지만 생성형 AI를 포함하여 일반적인 AI가 가져오는 다양한 문제에 따라 규제 당국은 규제 방안을 제시하고 있다. AI가 가져오는 상황을 인식하고 이에 대한 규제 방안을 제시하고 있다는 점에서 규제의 필요성을 공감하고 있는 것으로 판단된다. 다만, AI에 대한 규제가 필요하다는 점은 인정되지만 그 수준이나 방법에 대해서는 차이가 있다. AI 시스템의 오용으로 인한 피해를 방지할 수 있도록 이를 별도의 유형으로 분류하고, 그와 관련된 문제사례에 착안해 발생가능한 위험의 정도나 종류에 따라서 유형을 세분화하여 규제의 대상으로 삼고, 적정한 제재를 통해 실효성을 담보할 필요가 있다.[19] 대표적으로 EU「AI법」은 위험기반으로 AI를 유형화하여 규제 수준을 다르게 하고 있다.

무엇보다, AI 모델을 구축하는 것은 기반 기술을 활용하는 것이기 때문에 규제의 범위에 포함되지 않도록 하는 것이 필요하다. 모델 자체를 학습시키는 과정에서 학습 데이터의 저작권 침해 문제는 무시할 수 없으나, 저작권 침해 영역에서 다루면 될 사안이다. 따라서, AI에 대한 규제는 기술에 대한 규제가 아닌 구체적인 비즈니스 모델에 대한 규제가 바람직한 방향이다. 구체적으로는 LLM과 같은 AI 모델이 아니라, LLM을 기반으로 하는 AI 시스템이나 서비스가 안전을 해치는 경우로 제한되어야 한다는 의미이다.[20]

2. 디지털 심화기에서의 AI 안전권

1) 헌법적 가치로서 안전

디지털 심화기는 디지털 전환을 넘어, 기술적 가치가 사회적, 철학적 의미로 확장되는 국면의 전환을

---

18) EU「AI법」제4조(AI 문해력) AI 시스템 공급자와 배포자는 자신의 기술적 지식, 경험, 교육 및 훈련과 AI 시스템이 사용되는 맥락을 고려하고, AI 시스템이 사용되는 사람 또는 사람 그룹을 고려하여, 직원 및 자신을 대신하여 AI 시스템의 운영 및 사용을 다루는 다른 사람의 AI 문해력이 충분한 수준일 수 있도록 최대한 보장하기 위한 조치를 취하여야 한다.

19) 김정화 외,「생성형 인공지능(Generative AI) 기술의 규제 방향에 대한 입법론적 고찰 – ChatGPT 등 인공지능 시스템 생성물에 대한 표시·고지의무를 중심으로 –」, 형사법의 신동향 80, (2023), p.254.

20) 김윤명,「생성형 AI서비스제공자의 법적 책임과 의무」, 法學論叢 44(1), (2024.2.) p.58.

의미한다. 단순한 기술적인 디지털 전환의 가치만이 아닌, 우리사회가 가져야 할 구체적이고 실현가능한 공통된 가치를 가져야 한다. 이러한 관점에서 디지털 사회는 안전이라는 가치가 무엇보다 중요하다. 재난으로부터의 안전, 사회적인 활동에서의 안전, 경제적 생활에서의 자유, 기계적인 이용에서의 물리적인 안전 및 심리적인 안전을 포괄하는 안전권이라는 사회적 가치의 확보이다. 그렇기 때문에 디지털 심화기에는 국민 누구라도 안전한 상태에서 일상생활을 영위할 수 있어야 한다. 국민이 다양한 위해로부터 안전한 상태를 유지하도록 하는 것은 국가의 책무이다.

AI 안전은 물리적인 안전을 포함하여 AI가 가져오는 의사결정의 결과가 공정하고, 투명하고 그 결과에 대해 쉽게 이해할 수 있어야 한다. 이는 AI가 내린 결정에 대해 이용자가 수용할 수 있는 수준의 신뢰성을 가질 수 있어야 한다는 의미이다. 정부는 AI 결정에 대한 사업자의 설명의무를 부여하거나 기술적으로 설명가능한 AI의 개발을 진행 중이다. 더 나아가 AI가 가져오는 다양한 문제점들에 대응할 수 있는 체계나 거버넌스의 수립이 필요하다고 본다.

전문영역에서 사용되던 AI가 이제는 일상에서 사용되는 서비스 형태로 확산하고 있다. AI의 사용 범위가 챗GPT와 같은 일상적인 것에서부터 자율주행차와 같은 전문적인 영역에 이르기까지 다양하다. 일반인들이 AI에 대한 인식과 이용하는 영역 또한 다양화하고 있기 때문에 AI를 넘어선 다양한 사회현상에 대응할 수 있는 디지털 안전권에 대한 논의 또한 필요하다. 참고할 수 있는 개념은 독일 연방헌법재판소에서 수립한 'IT 기본권'[21]이다. 독일 헌법재판소의 판결에 따라,[22] 디지털기술 발전에서 흠결될 수 있는 기본권 보장을 위해 '정보기술체계의 신뢰성과 무결성 보장에 관한 기본권'으로서 IT 기본권이 창설되었다.[23] 이는 「독일기본법」의 정보자기결정권과 뿌리를 같이하고 있기에 국가가 시민의 정보기술시스템에 침입하여 개별적인 통신의 진행과정이나 저장된 데이터에 접근한다면 정보기술 분야에서 기본권 향유자의 사적 생활영역이 국가로부터 침해받게 되는 데, 이 기본권은 이를 보호하려는 것이다.[24] 이는 시스템 자체의 안전성은 물론, 무결성까지도 요구되는 개념으로 볼 수 있다. 시스템이 방해나 조작으로부터 벗어난 상태로 유지되도록 한다는 점에서 이를 유추할 수 있을 것이다.[25]

---

21) IT기본권은 '온라인 기본권', '컴퓨터 기본권', '정보통신기본권'이라고도 한다. 홍선기, 「독일에서의 디지털 기본권에 대한 논의」, 유럽헌법연구 33, (2020), p.72.

22) 독일 헌법재판소의 결정에 대해서는 박희영, 「정보기술 시스템의 기밀성 및 무결성 보장에 관한 기본권(상)」, 법제 43, (2008.10) 참조.

23) BVerfGE 120, 274 - 1 BvR 370/07, 1 BvR 595/07 (2008. 2. 27. 결정).

24) 홍선기, 「독일에서의 디지털 기본권에 대한 논의」, 유럽헌법연구 33, (2020), p.72.

25) 계인국, 「인터넷 검색엔진과 개인정보보호- 인적관련 정보의 처리와 정보자기결정권 및 IT-기본권」, 법제연구 46, (2014), p.167.

2) AI 안전권의 명확화

안전에 대한 헌법적 논의가 이루어지고 있다. 즉, 안전권이 헌법상 기본권인지 아니면 헌법 제10조의 행복추구권에 따른 파생적 기본권인지 다양한 논의이다. 또한, 안전권을 구체적으로 구현하는 입법이 ICT 분야 법률에서 이루어지고 있다. 대표적으로, AI와 가장 밀접한 법률인 「소프트웨어 진흥법」은 소프트웨어 안전을 SW 내부의 오작동이나 안전기능 미비로 인하여 발생할 수 있는 사고에 충분히 대비할 수 있는 상태로 정의하고 있다.[26] 즉, 오류가 발생하더라도 통제하거나 관리할 수 있는 상태라면 안전성이 유지될 수 있다는 것임을 알 수 있다. AI도 SW라는 점에서 SW 안전을 담보할 수 있어야 한다. 다만, SW 등 ICT 분야의 안전만이 아닌 철도, 항공 등 교통분야의 안전관련 법제는 파편적이라는 점에서 한계가 내재한다. 법률의 특성이기도 하지만, 다양한 안전 법제가 독자적인 입법목적을 갖고 있어 이질적인 대상을 연계하여 적용하기가 쉽지 않기 때문이다. 일반 안전법제의 제정을 통하여 안전권을 구체화하거나 특수한 유형의 파생적인 안전권을 규율하는 방안을 고려할 수 있을 것이다.

직접적으로 안전을 강조하지 않더라도 책임을 강화하면서 안전을 유도하는 입법의 경우도 있다. 대표적으로, 「제조물 책임법」을 들 수 있다. 「제조물 책임법」은 제조자의 의무를 강조한 법률이지만 소프트웨어 등 무형의 제품이나 정보는 제조물로 보지 않는다. 다만, AI나 소프트웨어가 제조물에 포함될 경우, 개발자나 사업자는 AI 안전을 위한 노력을 구체화할 것이며 이에 따른 안전성은 더욱 높아질 것이다. 이를 위하여 「제조물 책임법」에 무형의 제품 결함을 포함하는 입법을 하거나, 정보통신 관련 법률에 AI로 구현되는 제품이나 서비스의 안전의무에 관한 규정을 둘 수도 있을 것이다. 보다 명확히 하기 위해서는 현재 국회에서 논의 중인 AI 관련 법률에서 AI 안전을 위한 구체적인 규정을 두는 방안도 고려할 수 있다. 현재, 제안된 입법안에서는 신뢰성 확보를 위한 인증제도를 규정하고 있으나, 이는 실효적인 AI 안전을 담보하는 규정으로 보기 어렵다. 안전에 관한 구체적인 기준이나 안전을 위협하는 제품이나 서비스에 대한 사업자의 책임을 지우는 것은 아니기 때문이다.

3. AI 안전을 위한 규제원칙과 기대효과

1) 규제 원칙과 성질

(1) 규제 원칙

EU 「AI법」 제정의 의의는 그 동안 AI에 대한 규제가 윤리적인 측면에서의 논의였다면, 이제는 법적

26) "소프트웨어안전"이란 외부로부터의 침해행위가 없는 상태에서 소프트웨어의 내부적인 오작동 및 안전기능(사전 위험분석 등을 통하여 위험발생을 방지하는 기능을 말한다) 미비 등으로 인하여 발생할 수 있는 사고로부터 사람의 생명이나 신체에 대한 위험에 충분한 대비가 되어 있는 상태를 말한다.

측면에서의 실효적인 규제로 규제의 관점이 이전된다는 점이다. 미지의 기술인 AI에 대한 규제의 필요성은 다양하다. 사업자나 개발자 입장에서 규제는 명확하고, 예측가능해야 한다. 이로써, 다양한 형태로 나타날 수 있는 사업의 불투명성을 완화할 수 있기 때문이다. 또한, 규제정법주의에 따라 규제는 명확하게 법률로써 진행되어야 한다.[27] 다만, 법령에서 전문적·기술적 사항이나 경미한 사항으로서 업무의 성질상 위임이 불가피한 사항에 관하여 구체적으로 범위를 정하여 위임한 경우에는 고시 등으로 정할 수 있다. 특히, 행정기관은 법률에 근거하지 아니한 규제로 국민의 권리를 제한하거나 의무를 부과할 수 없다. 규제법정주의는 국민의 권리를 제한하거나 의무를 부과하는 모든 규제는 반드시 법률에 근거가 있어야 하며, 법률에 직접 규정하는 것이 원칙이다.

AI 분야의 입법도 규제법정주의에 따라 명확한 규제 내용을 제시하여야 한다. 명확하지 않을 경우, 불투명한 사업을 영위하기가 쉽지 않기 때문이다. 기업의 입장에서 규제의 높고 낮은 정도는 불확실성과 비교할 때, 문제는 아니다. 해결해나갈 수 있기 때문이다. 그렇지만, 불투명한 규제는 어떻게 할 방법이 없다. 이러한 점을 정책수립이나 입법 과정에서 당국에서는 고려하여야 한다. EU「AI법」과 같이 위험기반에 따른 명확한 기준을 제시해야 한다. 사업자들이 기대하는 것은 명확한 규제 수준이다. 규제가 높고 낮음의 문제라기 보다는 예측가능한 규제를 통하여 사업상 불투명성을 해소할 수 있기 때문이다.

일각에서는 법적 기준을 제시하는 것에 대해 우려를 표하기도 한다. 즉, 법적 기준을 제시한다는 것은 사업자에게 면책을 줄 수 있다고 보기 때문이다.[28] 대표적으로, 제조물 책임법의 '개발위험의 항변'이나 '법령준수의 항변'을 들 수 있다.[29] AI의 경우도, 구체적인 기준을 제시함으로서 해당 기준을 수범한 사업자에게는 면책가능성을 부여할 수 있기 때문이다.[30] 그렇지만, 개발위험의 항변은 국민의 안전에 위협이 될 수 있다[31]는 지적도 충분히 고려되어야 한다. AI가 인간의 기본권에 영향을 미칠 수 있다면, 제조물 책임법에서 사업자나 개발자의 면책에 관한 규정을 삭제하는 것도 하나의 방법이다.[32]

(2) 규제의 성질

AI에 대한 규제는 인권과 소비자인 국민의 안전을 주된 목적으로 한다. 안전 규제, 소비자 보호 규제

---

27) 「행정규제기본법」 제4조에서 규제법정주의를 선언하고 있다. 이에 따라 규제는 법률에 근거하여야 하며, 그 내용은 알기 쉬운 용어로 구체적이고 명확하게 규정되어야 한다. 또한, 규제는 법률에 직접 규정하되, 규제의 세부적인 내용은 법률 또는 상위법에서 구체적으로 범위를 정하여 위임한 바에 따라 대통령령·총리령·부령 또는 조례·규칙으로 정할 수 있다.

28) 손은지, 「AI의 안전한 공동체 편입을 위한 합리적 규제설계에 관한 연구」, 法學論文集 48(1), (2024), p.9.

29) 한국소비자원, 『제조물책임법 해설 및 사례』, 한국소비자보호원, (2002.06.), p.58~59.

30) 손은지, 「AI의 안전한 공동체 편입을 위한 합리적 규제설계에 관한 연구」, 法學論文集 48(1), 중앙대학교 법학연구원, (2024), p.9.

31) 이상정, 「제조물책임법상 개발위험의 항변」, 한국상품학회 학술발표논문집, (2003), p.163.

32) Id., p.163.

로서 사회적 규제로 볼 수 있다.[33] 그동안 AI의 논의에서 배재된 영역이 소비자 영역이다. 실질적으로 AI를 이용하는 소비주체에 대해 특별한 고민을 하지 못한 것은 기술적 진흥에 집중한 논의의 결과로 이해된다. 따라서, AI 안전을 위한 기본적인 고려는 소비자인 인간을 중심에 놓은 입법적 논의와 규제의 가치설정이라고 생각된다. 이러한 측면에서 AI 안전과 같은 사회적 규제는 경제적 규제와는 달리, 규제 완화가 능사는 아니다.[34] 사회규제가 갖는 성격상 안전을 위한 것이라는 점에서 기술의 안전을 확보하기 위한 규제는 필요충분성을 갖추어야 하기 때문이다. 즉, 최소한의 안전성이 보장된 상태에서 신기술이나 신사업이 허용되도록 하고, 발전과정에서 안전성 보장을 조정하고 강화하여야 한다.[35] AI에 대한 규제는 그 특성에 따른 규제 방식이 필요하다.

또한, 불완전한 AI 서비스가 출시됨으로써 나타날 수 있는 문제의 해결을 위해서라도 AI 서비스에 대해 선허용 후규제 원칙에 대해 고려될 필요가 있다. 21대 국회에서 AI 법제 논의과정에서 '선허용 후규제'가 논란이 된 적이 있지만, 사전적인 규제가 없다는 점이 반드시 긍정적으로 작동하는 것은 아니다. 사업자에게 사후적 규제는 예측가능성을 떨어트릴 수 있기 때문이다. 즉, 이러한 규제가 강한 규제로 실현되거나 예측하지 못한 규제가 되는 경우에 막대한 투자가 물거품이 될 수 있기 때문이다.36) AI에 대한 사전적 규제 없이 시장 출시를 허용할 경우, 사업자는 이에 대한 리스크에 충분히 대응할 수 있을지 의문이다. 따라서, AI의 성질이나 목적에 따라 구체적인 기준을 제시함으로써 사업자나 이용자에게 예측가능하고 신뢰가능하도록 해야할 것이다. 이를 위해 규제기관은 사전적 규제 원칙을 선언하고 대신 명확한 기준을 제시함으로써 사업자에게는 법적 안정성을 담보할 필요가 있다.

2) AI 규제 : 위험 중심 v. 원칙 중심

AI가 가져오는 여러 가지 문제에 대응해야 한다는 당위성은 인정된다. 다만, 기술에 대한 규제는 자칫 기술 투자나 개발을 저해할 수 있으며, 기술중립성에 위배될 수 있기 때문에 기술 규제는 지양될 필요가 있다. 따라서, 원칙 중심으로 갈 것이냐 위험 기반에 따른 규제로 갈 것이냐의 정책적 결정이 필요하다. 위험 중심의 방식이나 원칙에 따른 위험 대응이냐는 방법론의 차이이지, 안전을 위한 AI 규제 목적은 크게 다르지 않다.

먼저, 원칙 중심적인 접근 방식은 AI 안전을 확보하기 위한 문화적인 접근 방식으로 이해할 수 있다. 이는 보편적으로 수범해야 할 가치로서 AI에 대한 윤리적인 논의의 확장이다. 이에 대해서는 데이터 윤

33) 박균성, 『정책, 규제와 입법』, 박영사, (2022.05), p.52.
34) Id., p.52.
35) Id., p.55.
36) Id., p.77.

리나 AI를 활용하면서 직면할 수 있는 문제의 윤리적인 대응으로 이해할 수 있다. 이러한 논의는 원칙을 구체적으로 제시하는 것이 바람직하다는 점에서 법적인 규제와도 연결된다. AI 윤리는 AI 법률의 제정에 따라 그 가치가 축소하는 듯 보이지만, 실상 안전을 위한 모든 내용을 법제화할 수 있는 것은 아니기 때문에 개발 및 서비스 현장에서의 AI 윤리는 여전히 AI에 대한 규범적 가치로서 작동한다. AI는 법적 규제 이전에 소비자의 신뢰를 얻지 못할 경우에는 시장에서 퇴출될 수 있기 때문이다. AI 윤리가 작동할 수밖에 없는 환경에 놓여있다고 볼 수 있다.

다음으로, 규제 중심적인 접근 방식은 법률로써 구체적인 내용을 규정하고, 강제하는 것이다. EU 「AI법」이 대표적인 규제 중심의 법률이다. 리스크 기반의 규제라는 점에서 4가지 위험 유형에 따라 구체적인 기준을 사업자에게 제시함으로써, 사업자들의 불확실성을 해소할 수 있다는 점에서 의의가 있다. 일례로, AI 안전확보를 위한 데이터 공개, 인간의 안전을 해치는 알고리즘의 사용 금지, 관련 기술의 개발과 관련한 기록관리 의무 등을 들 수 있다. 실상, AI 안전은 기술적인 안전이 우선이지만 개발이나 서비스 과정에서 사업자에게 주의의무를 부여함으로써 간접적으로 안전을 강조할 수 있다는 점에서 보완적인 형태로써 작동하게 된다.

3) 규제의 대상 : 기술 v. 서비스

일반적으로 과학기술은 그 자체에 선악의 가치가 담겨지지 않는 중립적인 성격을 갖는다. 일례로, 기술중립적 입장에서 보면 다양한 유형의 데이터를 생성하도록 설계된 GAN(generative adversarial network) 알고리즘은 선의의 기술이다.[37] 문제는 이러한 기술이 개발자나 서비스제공자의 의도했던 바대로 이용된다고 보기 어렵다는 점이다. GAN 알고리즘을 이용하여 딥페이크(deepfake)를 만드는 것이 대표적인 예이다. 더욱이, 이용자의 다양한 이용행태를 사전적으로 차단하는 것은 바람직하지도 않다. 다양한 경험과 시도를 통하여 새로운 가치를 창출할 수 있기 때문이다. 이처럼, 기술은 이용하는 과정에서 발생하는 변수나 이용자의 이용행태에 따라 달라질 수 있다. 기술의 오남용은 기술 자체에 있는 것이 아닌 기술을 활용하는 사람에게 달려 있기 때문에 이들이 제대로 기술이나 서비스를 활용할 수 있도록 리터러시를 확산하는 것도 간접적 규제의 방편이 될 것이다.

일반적으로 규제 관점은 그동안 의무 없던 사항에 대한 의무가 발생한다는 점에서 기술기업의 입장에서는 기술 촉진을 저해할 수 있다. 이는 비용을 증대시킬 가능성이 높고, 상대적으로 경쟁을 제한할 수도 있다. 반면, 반시장 경쟁을 개선시킬 수도 있다. 사회적 비용이 높아질 수 있겠지만, 이는 정부의 역할을 충실히 함으로써 얻는 무형의 가치와 비교교량을 통해서 판단해야 할 사항이다. 이에 따라, AI

37) 김윤명, 「딥페이크 발명이 특허법상 공서양속에 위배되는지 여부에 관한 연구」, IP & Data 法 4(1), (2024), p.132.

기술의 안전성을 확보하기 위한 규제의 필요성이 인정될 수 있다. 그렇지만, 기술을 규제하는 것은 기술의 다양성이나 의도성을 훼손할 가능성도 있다. AI 안전이 확보되지 않은 상태로 지속적이고 반복적으로 이루어지는 위법한 행위에 대해서는 별도의 규제에 대한 정책적 고려가 필요하다. 다만, 기술에 대한 규제보다는 기술을 응용한 사업화 모델에 대한 미세한(fine) 규제를 통하여 기술의 가치를 훼손하지 않는 수준에서 정리하는 것이 바람직하다.[38] 이용자의 기술 오남용에 대한 접근은 일의적이지 않아야 한다. 사상의 자유와 표현의 자유와 같은 최고의 법적 가치는 기술의 오남용에 대한 규제와 비교할 때 고려될 필요가 있다. 이로써, 기술자체에 대한 규제가 아닌 해당 기술이 구현된 서비스 등 비즈니스 모델에 대한 세밀한 규제가 필요한 이유이다.

4) 기대 효과

AI 안전을 위한 규제는 사업자에게 부담지우는 것이지만, 안전이라는 가치는 헌법이 국민에게 부여한 행복추구권의 파생적 권리인 안전권을 달성하기 위한 것이다. 즉, AI가 가져오는 문제는 예측하여 대응하는 것이 바람직하겠지만, 예측하기 어려울 경우에는 사업자에게는 규제로서 작용할 수 있다는 점도 고려될 필요가 있다. 권리를 제한하거나 의무를 부과하는 것을 가급적 제한하고자 하는 것이 규제법정주의를 채택한 이유이다. 그렇기 때문에 규제를 설정하는 경우에는 명확하게 이루어져야 한다. 법률에 근거하는 안전 규제를 통해 기대할 수 있는 효과는 사업자 입장에서는 예측가능성의 확보, 규제에 따른 책임문제의 명확화, 그리고 인센티브로 작용할 수 있는 면책가능성을 가질 수 있다는 점이다. 「행정기본법」에서 규율하고 있는 샌드박스의 규정은 후발주자도 시장 진입을 가능하게 하는 정책적 효과를 가져올 수 있다. 따라서, 중소 벤처 AI 사업자에 대해서는 필요한 정책수립의 근거가 될 수 있다. 이처럼, 명확하고 예측가능한 규제는 전반적으로 시장에 대한 공정성, 투명성 및 예측가능성을 높일 수 있는 효과를 가져오게 된다.

---

38) Id., p.135.

Ⅳ. AI 안전의 확보와 구현을 위한 거버넌스

1. AI 안전 거버넌스의 방향

AI 안전에 대한 규제의 구체적인 실현 방안으로서 거버넌스 차원에서의 AI 안전을 확보하는 방안에 대한 고려 또한 필요하다. 위험기반의 관리체계나 위험기반의 규제정책은 AI의 생애주기에 따른 안전체계의 수립에 있다. AI 안전은 AI 또는 안전에 대한 거버넌스로 볼 수 있기 때문이다. AI 안전 규제에 대한 방향은 원칙중심으로 구체적인 기준을 제시할 필요가 있다는 점, 규제 대상은 기술 자체가 아닌 구체적인 비즈니스 모델이라는 점을 확인하였다. AI 안전은 다양한 영역에서 이루어져야 하며, 국가나 기업이 분리하여 대응할 수 있는 것이 아니다. 소비자인 국민의 입장에서도 안전에 관한 이해관계를 갖는다. 따라서, 공공영역, 기술영역, 소비자 영역 등 다양한 영역의 협의체를 구성하고 그 구성을 구체화하는 거버넌스의 수립은 AI 안전을 확보하기 위한 프레임워크(framework)가 될 수 있다. 최근 출범한 대통령소속의 '국가인공지능위원회'[39]가 실질적인 추진체가 될 것으로 보인다. 또한, AI 안전을 위한 법제도, 정책, 평가 및 기술개발을 위한 AI 안전연구소의 역할도 중요하다.

2. NIST AI 위험관리 프레임워크의 적용

1) AI 위험관리 프레임워크

안전과 AI는 분리할 수 없는 정책과제라는 점에서, AI안전 관점에서 미국국립표준원(NIST)이 제안한 AI 위험관리 프레임워크(Artificial Intelligence Risk Management Framework, 이하 'AI RMF'라 함)[40]는 조직이 AI 위험을 해결할 수 있도록 유연하고 체계적이며 측정가능한 프로세스를 제공한다. AI 위험관리를 위한 프로세스를 따르면, AI 기술의 이점을 극대화하는 동시에 개인, 그룹, 지역사회, 조직 및 사회에 부정적인 영향을 미칠 가능성을 줄일 수 있다. AI RMF는 두 부분으로 구성되어 있다. 첫 번째 부분에서는 조직이 AI와 관련된 위험을 어떻게 프레임화할 수 있는지 논의하고 신뢰할 수 있는 AI 시스템

39) 2024.8.6.일 시행된 국가인공지능위원회의 설치 및 운영에 관한 규정에 근거하고 있다.

40) 미국 상무부 산하 연구기관 美국립표준기술연구소(National Institute of Standards and Technology. 이하, NIST)가 2024. 1월 26일(현지시간) 인공지능 위험 관리 프레임워크 'AI RMF 1.0'(AI Risk Management Framework-AI RMF 1.0)를 발표했다. 인공지능신문, "美 상무성, 신뢰할 수 있고 책임있는 AI 운영 위한 '인공지능 위험 관리 프레임워크(AI RMF 1.0)' 발표", (2024.02.01.), https://www.aitimes.kr/news/articleView.html?idxno=27268, (2024.6.1. 방문).

의 특성을 개략적으로 설명한다. 프레임워크의 핵심인 두 번째 부분은 조직이 실제로 AI 시스템의 위험을 해결하는 데 도움이 되는 네 가지 특정 기능인 거버넌스, 맵핑, 측정 및 관리에 대해 설명한다.

2) 신뢰할 수 있는 AI 시스템

AI RMF는 '신뢰할 수 있는 AI 시스템'의 특성을 다음과 같이 정리하고 있다. 먼저, AI 시스템이 의도한 요구사항을 실패 없이 수행하는지에 대한 지속적인 테스트와 모니터링을 통해 정확성, 강건성을 확인하여 유효성과 신뢰성 확보토록 하고 있다. 또한, AI 시스템은 사람의 생명, 건강, 재산, 또는 환경을 위험에 빠뜨리는 상황을 초래해서는 안되며, 안전 위협의 종류, 심각도에 따라 리스크 관리 접근이 필요하다. AI 시스템에 대한 의도된 공격을 피하고, 방어하고, 복구하는 보안능력과 예상하지 못한 이벤트가 발생한 이후 정상 기능으로 돌아갈 수 있는 복원성이 필요하다. 또한, 투명성은 AI 시스템과 결과물 정보를 해당 시스템 사용자가 적절한 수준으로 접근하여 이용하는 것으로 이때 책임성은 투명성을 전제로 확보가 가능하다. AI 시스템의 한계로 지적되는 블랙박스화에 따라, AI 시스템이 작동하는 매커니즘 설명과 설계 목적에 따른 결과물에 대한 의미 해석이 가능하도록 도움이 되는 정보가 제시되어야 한다. 무엇보다, 개인정보 보호를 위한 익명성, 기밀성, 제어와 같은 가치가 AI 시스템 설계, 개발, 배포 시 제시될 필요가 있다. AI 시스템의 한계이자 문제로 지적되는 편향 등에 대해서는 공정성을 담보하기 위하여, 시스템적 편향, 통계적 편향, 인지적 편향 등 3가지 범주의 편향을 관리하고 통제해야 한다.

3) 위험 요소의 관리

AI RMF는 AI 시스템에서 발생할 수 있는 위험 요소를 관리하기 위해 식별, 평가, 완화의 세 단계로 접근한다. 이 세부 과정은 다음과 같이 이루어진다. 먼저, 위험 식별로서 AI 시스템이 가지고 있는 잠재적 위험 요소를 분석한다. 여기에는 데이터의 편향, 윤리적 문제, 프라이버시 침해, 안전성 문제 등이 포함된다. 예를 들어, 데이터가 특정 인구 집단에 편향된 경우 차별적 결과를 초래할 위험이 있으며, 이처럼 AI가 의도치 않게 부정확하거나 부적절한 결정을 내릴 가능성이 있는지를 식별한다. 다음으로, 위험 평가로서 식별된 위험 요소의 심각성 및 발생 가능성을 평가하는 단계이다. AI 모델이 실제 환경에서 작동할 때 발생할 수 있는 문제와 그 영향력을 정량적으로 또는 정성적으로 분석한다. 이 단계에서는 AI 시스템이 다양한 조건에서 일관된 성능을 유지하는지 확인하고, 시스템 오류가 발생했을 때 영향을 받는 사용자와 그 피해 규모를 예측한다. 마지막으로, 위험 완화(mitigate)로서 평가 결과에 따라 필요한 조치를 통해 위험을 최소화한다. 이는 모델 재학습, 데이터 품질 개선, 새로운 정책 수립 등으로 이루어진다. 예를 들어, 편향된 결과를 방지하기 위해 AI 모델을 재훈련하거나 프라이버시 침해

가능성을 줄이기 위해 데이터 사용 방침을 강화하는 방식으로 위험 요소를 관리한다. 이 과정에서 위험은 지속적으로 모니터링되고 업데이트될 것이다.

4) 핵심 기능의 제시

AI RMF는 AI 리스크를 관리하여 신뢰할 수 있고 책임감 있는 시스템을 구축하기 위한 조직의 거버넌스, 맵핑(위험식별), 측정, 관리의 네 가지 핵심 기능을 제시하고 있다. 먼저, AI 리스크를 관리할 수 있는 거버넌스 체계의 구축이다. 거버넌스는 적절한 구조, 정책 및 프로세스를 구현하는 등 AI 시스템 수명 주기 전반에 걸쳐 리스크 관리 문화를 조성해야 하며 조직의 고위 관리자는 리스크 관리를 조직의 우선순위로 해야 한다. 두 번째는 AI 시스템 수명주기는 다양한 주체들이 참여하는 여러 상호작용의 활동으로 모든 구성요소에 대한 위험 및 이점을 매핑하여 개인, 집단, 커뮤니티, 조직 및 사회에 미치는 영향을 종합적으로 검토해야 한다. 세 번째는 리스크에 대해서는 정량적 및 정성적 또는 혼합된 방법, 기술 등을 사용하여 리스크 및 관련 영향을 분석, 평가, 벤치마킹 및 모델링 해야 한다. 이러한 분석을 통해 리스크를 평가하고 영향을 모델링하여 조직적 대응 전략을 수립한다. 마지막으로, 위험에 대한 모니터링 등을 통해 식별된 리스크에 대한 대응, 복구, 커뮤니케이션 계획을 수행하고, 이후에도 반복적인 리스크 모니터링을 통해 예상치 못한 새로운 리스크에 대비할 수 있도록 지속적인 관리체계를 구축한다.

5) 법적 및 정책적 시사점

AI RMF는 AI 기술과 관련된 위험 관리에서 표준적인 가이드라인을 제공함으로써, 기업이나 기관이 AI 시스템의 신뢰성을 확보하는 데 도움을 준다. AI 위험에 적응하는 사이버 보안 관행을 채택할 수 있도록 지원한다. AI RMF는 AI 관련 위험관리 지침을 제공하고, AI 시스템이 안전 및 신뢰성에 필요한 표준을 충족하도록 보장한다. 또한, AI RMF는 기술적 문제뿐만 아니라 사회적, 법적, 윤리적 요소를 포함하여 AI와 관련된 복잡한 위험을 포괄적으로 다루고 있다. 정책 입안자와 규제기관은 AI의 공정성, 투명성, 안전성을 보장하기 위한 프레임워크를 개발할 때 참조할 수 있는 기준을 제공한다. 마이크로소프트는 이 프레임워크를 고객 계약에 포함시킬 것이라고 발표했다.[41] 이처럼, AI RMF는 AI가 사회에 미칠 수 있는 부정적 영향을 사전에 대비할 수 있도록 한다. 특히, 정책 수립자와 법률가들이 AI 시스템의 안전성을 보장하고 공공의 신뢰를 확보할 수 있는 근거를 마련하는 데 중요한 지침으로 사용될 수 있다. 이를 통해 사회적 신뢰를 높이고, AI 기술 발전에 따른 법적 위험을 줄이는 기반을 제공한다는

41) AI타임스, "美 표준기술연구소, 신뢰할 수 있는 AI를 위한 'AI 위험 관리 프레임워크' 공개", (2023.01.27.), https://www.aitimes.com/news/articleView.html?idxno=149135, (2024.6.1. 방문).

점에서 AI RMF의 의의가 있다. AI 안전을 위한 세부적이고, 실현가능한 기준과 방안을 제시한다는 것은 안전성 확보라는 정책적 목표에 대한 사업자들의 예측가능성을 마련하고, 이로써 법적안정성을 제시함으로써 사업자의 참여를 이끌어낼 수 있다. 우리도 논의 중인 AI 법제에 준거틀을 제시할 수 있어야할 것이며, 그 작업은 AI 안전연구소의 몫이 될 것이다.

2. AI 안전의 기술적 구현

1) AI 안전을 위한 설명의무

「개인정보 보호법」에 따라 개인정보처리자는 정보주체의 설명요구권[42]을 기술적으로 구현하기 위하여 자동화된 결정의 기준과 절차, 개인정보가 처리되는 방식 등을 정보주체가 쉽게 확인할 수 있도록 공개하여야 한다. 다만, 자동화된 결정이 정보주체의 동의 등에 따라 이루어지는 경우에는 그러하지 아니하다(제37조의2). 「개인정보 보호법」은 알고리즘에 대한 설명의무와 거부권을 정보주체의 권리로 인정하고 있으나, 알고리즘의 적용 자체를 금지하는 것이 아니라 적용 과정에서 정보주체가 문제라고 인식하거나 사용된다는 사실을 인지하고서 그에 대해 설명을 요구하거나 적용을 거부할 수 있는 사후적인 권리로서 인정된다. 일종의 시정요구권이라는 법적 성질을 갖는다. 다만, 정보주체가 알고리즘의 적용과정에서 나타나는 문제점을 인식하고 그러한 문제가 기본권을 훼손하는 경우에는 사후적인 구제조치 내지 시정요구를 통해 문제를 해결해가는 구조라는 점에서 기본권적 성질을 부인하기는 어렵다.[43]

2) 설명요구권의 기술적 구현

AI 안전은 위한 방안으로써 설명요구권을 구현하기 위한 방안으로써 설명가능한 AI의 개발이나 또는 결과에 대한 신뢰성을 높이기 위한 기술개발이 이루어지고 있다. 먼저, 설명가능한 AI가 구체화하기 전 단계로서 요구받는 안전성 확보방안으로 알고리즘 공개가 주장된다. 알고리즘을 공개함으로써 문제의 원인을 찾고, 그에 따른 대응방안을 찾을 수 있다는 이유이다. 다만, 사업자들은 알고리즘 공개에 대해 영업비밀이나 기업정보가 외부에 유출됨으로써 기업자산이 제3자에게 유출될 수 있다는 주장을 한다. 이러한 우려를 불식시키기 위해 고려되어야할 사항은 알고리즘을 공개가 일반공중이나 제3자에게 공

---

42) 설명요구권은 투명성, 안전성, 공정성, 신뢰성, 책임성 등이 확보된 서비스를 이용하기 위한 정보주체의 권리이다. 개인정보 보호법은 정보주체의 권리로서 완전히 자동화된 개인정보 처리에 따른 결정을 거부하거나 그에 대한 설명 등을 요구할 권리를 인정하고 있다. 헌법상 기본권이라기 보다는 개별법상의 권리로써 헌법의 개인정보자기결정권에 근거한 파생적 권리로서 성질을 갖는다. 정보주체의 권리로서 설명요구권 및 결정거부권을 인정함으로써, AI에 의한 의사결정에 대해 정보주체로서 개인은 적극적인 권리를 행사할 수 있다. 개인정보처리자는 정보주체가 자동화된 결정을 거부하거나 이에 대한 설명 등을 요구한 경우에는 정당한 사유가 없는 한 자동화된 결정을 적용하지 아니하거나 인적 개입에 의한 재처리·설명 등 필요한 조치를 하여야 한다.

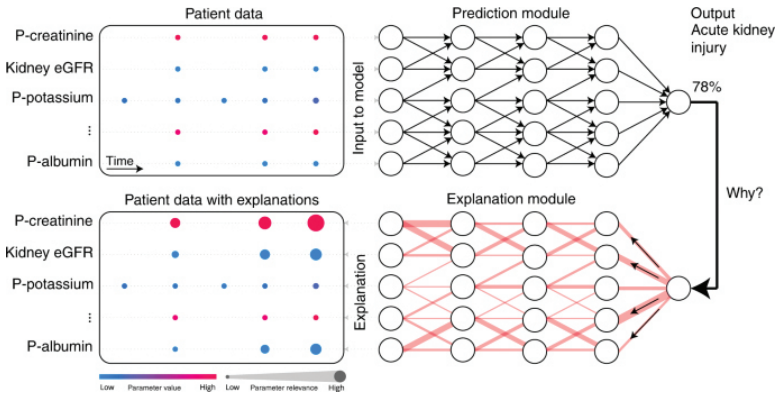
43) 김윤명, 「알고리즘 공개와 영업비밀 보호 간의 긴장관계」, 사법 66, (2023), p.894.

개한다는 것이라기 보다는 관리가능한 객관적이고 공정한 기관을 통해 이루어지도록 설계하여야 한다.

전문기관을 통한 공개에 대해서도 우려를 가진다면 적극적으로 이용자나 당사자에게 해당 알고리즘에 대해 구체적으로 운용되는 내용이나 과정 및 결과를 도출하게 된 내용을 요구할 수 있는 권리를 부여하는 것이다.[44] 현재, 「개인정보 보호법」에 규정된 설명요구권 등 정보주체의 권리를 일반 법률에 규정하는 것이다. 「특허법」에서도 블랙박스화에 따른 투명성 및 신뢰성을 확보 등을 이유로 기술공개와 유사하게 데이터를 포함한 알고리즘에 대해 공개할 수 있도록 제도화하는 방안을 고려할 수 있다.[45] AI 발명을 공개함에 있어서 학습데이터를 공개하는 방안이다. 예를 들면, AI 발명의 특성이나 학습에 사용된 학습데이터의 특성에 따라, 알고리즘의 공개, 소스코드의 공개, 전체 데이터셋의 공개, 데이터 일부에 대한 샘플링, 일부 데이터셋의 공개, 공개된 사이트의 주소, 기탁처 등에 대한 사항을 고려할 수 있을 것이다.

다음으로, AI 자체에 설명가능한 알고리즘을 부가함으로써 인간이 이해할 수 있도록 하는 것이다. 설명가능한 인공지능(eXplainable AI, 이하 'XAI'라 함)[46]이란 AI 모델이 특정 결론을 내리기까지 어떤 근거로 의사결정을 내렸는지를 알 수 있게 하는 것을 말한다.[47]

〈그림 1〉 설명가능한 인공지능



출처 : NATURE COMMUNICATIONS, (2020)[48]

44) Id., p.892.

45) Tabrez Ebrahim, 「Artificial Intelligence Inventions & Patent Disclosure」, Iowa Legal Studies Research Paper No. 2021-48, (2020), p.148.

46) DARPA, "Explainable Artificial Intelligence (XAI) (Archived)", https://www.darpa.mil/program/explainable-artificial-intelligence, (2024.6.1. 방문).

47) 안재현, 『XAI 설명가능한 인공지능을 해부하다』, 위키북스, (2020.03), p.4.

48) Simon Meyer Lauritsen et al., 「Explainable artificial intelligence model to predict acute critical illness from electronic health records」, Article number: 3852 (2020), https://www.nature.com/articles/s41467-020-17431-x, (2024.6.1. 방문).

XAI의 핵심 요소는 신뢰이다. 신뢰가 없다면 AI 모델이 생성하는 모든 행동이나 결정에 의심이 지속될 것이고, 이는 AI가 기업에 진정한 가치를 가져다주어야 하는 제품을 배포할 때의 위험을 증가시키기 때문이다. 미국 국립표준기술원에 따르면, 설명가능한 AI는 다음 4가지 원칙을 중심으로 구축되어야 한다.[49)]

〈표 2〉 NIST XAI 4가지 원칙

항목	내용
설명성	각 출력에 대한 증거, 지원 또는 추론을 제공하는 능력
의미성	사용자가 이해할 수 있는 방식으로 설명을 전달하는 능력
정확성	왜 결정을 내렸는가 뿐만 아니라 어떻게 결정을 내렸는가를 설명할 수 있는 능력
한계성	설계 한계를 넘어서서 결론이 신뢰할 수 없는 경우를 판단하는 능력

출처 : NIST, (2021)[50)]

이러한 원리는 지능형 알고리즘의 개발과 훈련을 안내하는 데 사용될 수 있는 동시에 본질적으로 수학적 구성에 적용될 때 설명 가능하다는 것이 무엇을 의미하는지에 대한 인간의 이해를 안내하는 데에도 사용된다. 설명가능한 AI는 블랙박스화하고 있는 알고리즘에 대해 신뢰성과 투명성을 높이기 위한 기술적 방법인 것이다.

다음으로, 관리체계에 관한 사항으로써 서비스제공자에게 주의의무를 부과하거나 기술적 수단을 강구하도록 하는 방안이다. 지속적이고 반복적인 사회질서에 반하는 정보를 생성하는 경우나, 소비자 이익을 심하게 훼손하는 경우에는 손해배상 책임을 강화하고 입증이 어려운 경우를 상정하여 입증책임을 전환하는 것이다. 현재, EU 제조물책임지침이 개정되고 있으며 AI도 제조물책임에 해당한다는 점이 고려되어야할 사항이다.[51)]

기술적인 방법으로써, 검색증강생성(retireval augment generation, 이하 'RAG'라 함) 방식을 연계하는 것이다. RAG는 AI의 생성물에 대한 진실성을 높인다. AI 안전 관점에서 RAG는 기술적, 비즈니스적으로 대안으로 볼 수 있다. 이러한 장점에도 불구하고, 검색기능을 부여한다는 것은 검색 자체에 대한 신뢰성을 전제하는 것으로 이해할 수 있다.

---

49) P. Jonathon Phillips et.al., 「Four Principles of Explainable Artificial Intelligence」, NIST, (2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf, (2024.6.1. 방문).

50) Id.

51) EU 제조물책임지침에 대한 논의에 대해서는 이재호, 디지털시대에서 제조물 책임 – EU 제조물 책임지침 개정안 검토를 중심으로 –, 법학연구 27(1), 2024.03, 129~158면.

3. AI 안전을 위한 거버넌스 체계

1) 안전 거버넌스

거버넌스(governance)는 국가 및 시민 사회의 권력 구조에 대한 이해관계의 조정을 의미하거나, 기업의 지배 구조 등 다양한 의미를 지닌다. 디지털 전환에 따른 거버넌스의 개념도 디지털 거버넌스로 변모하고 있다. 디지털 거버넌스는 디지털 전략이나 정책, 기준에 대한 명백한 책임을 만드는 데 초점을 두며, 조직에서 디지털(웹사이트, 모바일, 소셜미디어와 인터넷 등)과 관련된 책임, 역할과 의사 결정 권한을 확립하는 것으로 이해될 수 있다.[52] 거버넌스는 사용되는 분야나 목적에 따라 그 정의가 달라질 수 있음을 의미한다. 공공 부문의 거버넌스는 정부나 공공 영역에서의 정책적인 의사결정 과정에서 이루어지는 결정이나 절차를 의미한다.

이러한 측면에서 정부와 관련된 문제를 해결하는 기제로서 거버넌스를 파악하는 견해에 따르면, "국가의 경제·사회적 자원의 관리 과정에서 권력이 사용되는 방법·유형으로 파악하기도 하고, 또는 거버넌스는 공적인 관심사와 관련하여 권력이 행사되고, 시민들의 의견이 제시되고, 의사결정이 이루어지는 방법을 결정하는 제도 및 절차"[53]로 정의된다. 거버넌스의 구축은 구체적인 정책적인 역할과 기대 효과를 담아낼 수 있어야 하기 때문이다. 전문성을 확보하고 부처 간의 정책 조정을 통하여 정합성을 담보할 수 있어야 한다. 아울러, 구체적으로 정책을 기획하고 실행할 수 있는 권한이 확보될 수 있어야 한다. 무엇보다, 거버넌스는 거버넌스 내부는 물론 다른 거버넌스와의 관계에서 정보의 공유가 이루어져야 한다. 그러므로, 위원회 간 공유를 넘어서 부처 간 공유까지로 확대될 필요가 있다.

무엇보다, 재난관리를 위한 협력적 거버넌스 체계를 성공적으로 구축하기 위한 실행 방안으로 우선적으로 필요한 것은 공공 부문 내부 조직만의 폐쇄적인 네트워크 형성보다는 외부의 전문가를 활용하는 방안을 고려하는 것이다. 재난관리 기관의 외부에 있는 전문성과 중립성을 갖춘 중재자 또는 촉진자를 활용하여 이들에게 역할(role)과 책임(responsibility)을 배분하고 나아가 공동의 목표(shared goal)에 대해 참여하는 조직들의 자발적인 참여와 협력을 통해 합의를 도출할 수 있는 구조를 형성함으로써 협력적 거버넌스의 효과성을 향상시킬 수 있기 때문이다. 이를 위해서는 정책과 정보 공유를 통해 서로 다른 조직들간 정보 비대칭성의 완화가 동반되어야 할 것이다.[54]

---

52) 김시정 외, 『디지털거버넌스 구축 및 활성화 방안 연구』, 서울디지털재단, (2017.12), p.28.
53) 정명운, 『거버넌스 제도체계 구축을 위한 법제화 방안 연구』, 한국법제연구원, (2009.10), p.19.
54) 원소연, 『한국형 협력적 거버넌스 체계 구축 방안 연구: 네트워크분석을 통한 재난안전분야 비교 사례 연구』, 한국행정연구원, (2013.12), p.20.

2) AI 안전 거버넌스의 체계

AI는 다양한 분야에서 적용가능하며, 예기치 못한 문제가 발생할 소지가 크다는 점을 확인하였다. 이에 따른 AI 위험을 관리하기 위한 협력 체계의 구축,과 일관된 정책의 수립과 집행을 통한 안전성을 확보할 수 있어야 한다. AI를 전담할 기구의 설치도 중요하다. 기존의 체계를 컨트롤하거나, 합의제 행정기구를 두거나 별도 의결기구를 통해서 정책에 대한 의결을 할 수 있도록 하는 것도 하나의 방안이다. 이와 별개로, AI 관련 법안이 국회에 계류 중인 여러 법안에서도 국가인공지능위원회를 두도록 규정하고 있다.[55] AI 관련 입법이 이루어질 경우, 국가인공지능위원회의 설치 근거는 대통령령이 아닌 법률에 근거하게 될 것이다.

세계적으로 AI 안전연구소 설립을 경쟁적으로 주도하는 국가는 영국과 미국이다. 2023.10월 열린 AI 안전성 정상회의에서 미국과 영국은 각각 AI 모델 안전성 평가를 위한 AI 안전연구소 설립을 발표했다. 미국은 자국 기업 중심으로 구성된 5개 분야의 민간 컨소시엄을 발족해 레드티밍(취약점 발견·검증을 위한 의도적 공격), 역량평가 등을 준비하고 있으며, 영국은 AI 기술전문가 채용을 통해 AI 시스템 안전성에 대한 기술평가 시행을 준비하고 있다. 양국은 AI 기술 안전성 점검을 상호 협력하는 양자간 파트너십도 맺었다.[56] 우리 정부도 2024.5월 서울 AI Summit에서 AI 안전연구소 설립을 공식화했다. 동 연구소는 AI 환각 등 기술적 한계와 오용, AI 자율성 확대에 따른 위험 등도 커지는 만큼 이를 해소하는 기술을 개발하는 국가 차원의 연구를 담당할 계획이다.[57] 구체적으로, AI 안전연구소는 AI 안정성을 검증하고 연구하는 전담 기관으로 AI 안전검증, AI 안전기술연구, AI 안전정책 및 글로벌 협력 등으로 구성될 것이라고 한다.[58]

AI 안전연구소의 목표는 다음과 같은 주요 영역에 두어야 한다. 먼저, AI 위험 평가와 관리 체계의 확립이다. AI 기술이 초래할 수 있는 다양한 위험을 사전에 평가하고 관리할 수 있는 체계를 마련해야 한다. 이를 통해 AI의 오남용을 방지하고 사회적, 경제적 안정성을 확보할 수 있다.다음으로, AI 안전은 국제사회에서 필요한 사항이라는 점에서 국제협력 및 표준화가 이루어질 필요가 있다. 즉, AI 기술의 글로벌 특성을 고려하여 국제협력을 강화하고, 글로벌 표준을 마련하여 각국의 AI 정책이 일관되게 적용되도록 해야 한다. 무엇보다, 국제협력이나 표준화는 기술외교라는 점에서 의의가 있다. 이는 AI

55) 일례로, 조인철 의원의 인공지능산업 육성 및 신뢰 확보에 관한 법률안에서는 다음과 같이 규정하고 있다.
제6조(국가인공지능위원회) ① 인공지능사회의 구현 및 인공지능산업의 진흥과 관련된 사항을 심의 의결하기 위하여 대통령 소속으로 국가인공지능위원회(이하 "위원회"라 한다)를 둔다.

56) GOV.UK, "Introducing the AI Safety Institute", (January 17, 2024), https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute, (2024.6.1. 방문).

57) 아시아경제, "하정우 센터장도 후보" 국내 첫 AI안전연구소장은 누구?, 2024.06.28.일자.

58) 동아사이언스, 방승찬 ETRI 원장 "연내 AI안전연구소 설립해 국가·사회 안전 확보", 2024.06.27.일자.

기술의 안전한 발전과 적용에 있어서 주도권을 가질 수 있으며, 공적 기여를 할 수 있기 때문이다. 다음으로, 윤리적 AI 개발과 사용 촉진에 관한 사항이다. AI 기술이 윤리적이고 책임감 있게 개발되고 사용될 수 있도록 지침을 마련해야 한다. 이를 통해 AI의 긍정적 잠재력을 극대화하고 부정적 영향을 최소화할 수 있다. 물론, AI 윤리는 법제화를 통하여 그 기능이 약화될 수 있겠지만, AI 규제의 원칙중심에 따를 경우에는 여전히 전반적인 수범주체가 가져야할 윤리는 유효하다. 아울러, 교육 및 인식 제고와 같이 AI 및 AI 안전에 관한 리터러시의 확산이 요구된다.[59] AI 관련 지식과 위험성에 대한 교육을 통해 일반 대중과 관련 종사자의 인식을 제고해야 한다. 이로써 AI 기술의 이해도를 높이고, 안전하고 책임감 있는 사용 문화를 형성하는 데 기여할 수 있기 때문이다. 무엇보다, AI가 기술이나 제도에 매몰되기보다는 사회전반적으로 가져야할 문화라는 인식과 활동이 필요하다.

국가인공지능위원회의 운영지원은 별도 사무국을 통해 이루어지겠지만, AI 안전거버넌스로서 실질적인 내용인 정책 연구, 법제 연구, 안전 평가, 기술 및 표준화, 국제 교류 협력 등이 사업은 AI 안전연구소의 역할로서 기대할 수 있을 것이다. 다만, AI 안전연구소는 관련 법률이 입법화되어야 공식적인 출범이 가능할 것으로 보인다.[60]

V. 결 론

AI 안전은 AI를 개발하거나 이용하는 과정에서 나타날 수 있는 여러 가지 위해 상황이나 문제로부터 국민의 안전을 보장하기 위한 정책목표이다. 지금까지 윤리를 통해 AI 안전성을 확보하기 위한 자발적인 노력을 요구해왔다. 그렇지만, 이제는 EU「AI법」과 같이 윤리원칙을 넘어서 입법을 통한 규제중심으로 전환하고 있다. 우리나라도 '인공지능기본법'에 대한 입법 요구가 끊이질 않고 있으며, 국회에 계류 중인 관련 법안이 15개에 이른다. 그만큼 AI에 대한 국가적 관심이 크기 때문으로 보인다.

---

59) EU「AI법」 제3조 (56)에서는 인공지능 문해력에 대해 "이 규정의 맥락에서 각각의 권리와 의무를 고려하여 제공자, 배포자 및 피침해자가 정보에 입각하여 인공지능시스템을 배포하고 인공지능의 기회, 위험, 인공지능이 초래할 가능성이 있는 피해를 인식할 수 있는 기능, 지식 및 이해를 말한다."고 정의하고 있다.

60) 이혜민 의원의 인공지능산업 진흥 및 인공지능 이용 등에 관한 법률안에는 다음과 같이 설치 근거를 두고 있다.
제14조(인공지능안전연구소) ① 과학기술정보통신부장관은 인공지능이 사회에 미치는 영향과 인공지능으로 인해 발생할 수 있는 위험으로부터 국민의 생명·신체·재산 등을 보호하고 인공지능사회의 신뢰
기반을 유지하기 위한 상태(이하 "인공지능안전"이라 한다)를 확보하기 위한 업무를 전문적이고 효율적으로 수행하기 위하여 인공지능안전연구소(이하 "안전연구소"라 한다)를 운영할 수 있다.

AI 안전 거버넌스와 관련하여 법안에서 중요하게 다루고 있는 AI 안전에 관한 사항은 신뢰성 확보, AI 거버넌스 및 AI 안전연구소에 관한 사항이다. 현재, 대통령령에 근거하여 국가인공지능위원회가 출범하였지만 상위 법인 법률에 근거할 필요가 있다. 아울러 AI 안전에 관한 실질적인 내용을 개발하고 연구할 수 있도록 AI 안전연구소의 자율성과 전문성 확보도 중요한 사항이다.

AI 안전을 달성하기 위한 방법론은 다양하다. 제도적이거나 기술적으로 안전을 위한 여러 가지 사항을 구체화할 수 있기 때문이다. 무엇보다, AI를 이용하는 과정에서의 투명성 확보이다. 결과에 대한 편향 없는 공정성을 확보하는 것이다. 이러한 투명성과 공정성을 통해 서비스제공자가 얻을 수 있는 것은 시장의 신뢰성이다. 신뢰성이라는 추상적 개념이기는 하지만, 이는 기술이나 사업자가 시장에서 성장할 수 있는 기본적인 요소이다.

물론, 제3 영역에서의 정부는 AI 안전을 위한 구체적인 정책목표를 제시할 수 있을 것이다. 규제지향적인 목표를 수립할 것인지, 원칙중심의 자율규제적 목표를 수립할 것인지는 AI 안전에 대한 사회적 합의에 따라 달라질 수 있다. AI 안전의 목표이자 과제는 AI를 활용하는 과정에서 나타날 수 있는 여러 가지 문제 상황에서 국민의 안전을 담보하는 것이다. 기술중립적인 관점에서 기술에 규제가 아닌 비즈니스 모델에 대한 규제로 최소화할 필요가 있다. 이를 위해서는 시민사회의 지속적인 모니터링이 필요하다. 기술이라는 것이 선의로 개발된 것이기는 하지만, 이를 이용하는 과정에서 오남용은 예기치 않게 나타날 수 있기 때문이다. AI가 자발적으로 위험을 키우는 기술적 상황은 아니라는 점에서 규제의 중심은 의도적이거나 악의적인 이용을 통제하는 방향으로 설정될 필요가 있다. AI 위험 관리를 위한 프레임워크의 검토는 AI 위험관리와 안전성 평가에서 의미있는 역할을 할 것으로 기대한다.

무엇보다, AI 관련 법률은 입법 그 자체의 목적이 아닌 명확한 문제의 설정과 문제를 해결하기 위한 정책적 수단이 구체화되어야 한다. 그런 의미에서 AI 기본법의 제정은 시간이 아니라 AI 안전을 위한 명확한 정책목표의 설정과 이행방안을 찾는 데 있다는 점을 다시한번 확인하고자 한다.

■ 참고문헌 ■

고학수(2022),『AI는 차별을 인간에게서 배운다』, 21세기북스.

권용래(2010),『소프트웨어 테스팅』, 생능출판사.

김윤명(2022),『블랙박스를 열기 위한 인공지능』법, 박영사.

김윤명·이민영(2016),『소프트웨어와 리걸 프레임, 10가지 이슈』, 커뮤니케이션북스.

한국소비자원(2002),『제조물책임법 해설 및 사례』.

허 영(2013),『헌법』, 박영사.

과학기술정보통신부(2017.12), 한국정보통신기술협회,『SW 안전 진단 가이드(공통분야)』.

김길수(2020), "인공지능의 신뢰에 관한 연구",『한국자치행정학보』, 제34권 제3호.

김여라(2022), "안전한 디지털 이용환경 조성과 디지털 권리 강화를 위한 과제",『이슈와 논점 제1942호, 입법조사처.

김윤명(2024), "딥페이크 발명이 특허법상 공서양속에 위배되는지 여부에 관한 연구",『IP & Data 法』, vol.4, no.1.

김윤명(2023), "제조물책임 범위의 확장 : SW와 AI의 적용가능성",『정보화정책』, 제30권 제1호.

김윤명·오병철 외(2017),『SW제조물책임 관련 법제 현황 조사연구』, 소프트웨어정책연구소.

김진우(2019), "소프트웨어 업데이트에 관한 민사적 법률문제 –임베디드 시스템을 중심으로–",『민사법학』, 86.

박태형 외(2016),『SW 안전 관리 관점에서의 기반시설 보호법제 개선 연구』.

신봉근(2005), "컴퓨터소프트웨어와 제조물책임",『인터넷법률』, 제27호, 법무부.

유지연(2012),『디지털災難 槪念 確立 및 對應모델 開發 硏究』, 學位論文(博士), 고려대학교.

윤정현 외(2022), "디지털 안전사회의 의미: 안전과 안보의 복합공간으로서 전환적 특징과 시사",『정치·정보연구』, 제25권 3호.

윤정현(2019), "인공지능과 블록체인의 도입이 사이버 안보의 공·수 비대칭 구도에 갖는 의미",『국제정치논총』, 59(4).

윤지영 외(2015),『법과학을 적용한 형사사법의 선진화 방안(Ⅵ)』, 형사정책연구원.

이경미(2020), "인공지능의 소프트웨어 오류로 인한 민사책임",『가천법학』, 13(1).

이수연 외(2015), "디지털 사이니지를 활용한 재난안전 정보 보호에 대한 연구",『융합보안논문지』, v.15 no.7.

이재호(2024), “디지털시대에서 제조물 책임 – EU 제조물 책임지침 개정안 검토를 중심으로 –”, 『법학연구』, 27(1).

임재주(2018.9), 『소프트웨어 안전 기본법안 검토보고서』, 국회과학기술정보방송통신위원회.

임종인 외(2011), 『디지털위험도 관리 및 디지털재난 대응 모델 개발 방안 연구』, 경제인문사회연구회.

임종인 외(2010), 『신IT기술 응용 확산과 디지털재난 유형 예측 연구』, 고려대학교.

정국환 외(2010), 『공공정보화 선진화를 위한 디지털위험관리 방안 연구』, 정보통신정책연구원.

정도균(2019), “SW 안전 국제표준화 동향과 시사점”, 『이슈리포트』, 2019-18호, 정보통신산업진흥원.

조기열(2021.4), 『소프트웨어 진흥법 일부개정법률안 검토보고』, 국회과학기술정보방송통신위원회.

한국저작권위원회(2024), “나이트쉐이드(Nightshade), AI 생성 이미지의 저작권 침해에 반격하는 아티스트의 도구”, 『저작권 이슈 트렌드』, 통권 28.

행정자치부(2010), 『국가안전관리기본계획(2010-2014)』.

Bruce Schneier(2017), *Click Here to Kill Everyone, Policy Issues surrounding Artificial Intelligence, Algorithms & Privacy*.

NIST(2023), *Artificial Intelligence Risk Management Framework(AI RMF 1.0).*

P. Jonathon Phillips et.al(2021), *Four Principles of Explainable Artificial Intelligence*, NIST.

Yuntao Bai et.al.(2022), “Constitutional AI: Harmlessness from AI Feedback”, Available at arXiv:2212.08073

원 고 접 수 일 | 2024년 10월 6일
1차심사완료일 | 2024년 11월 7일
2차심사완료일 | 2024년 11월 15일
최종원고채택일 | 2024년 11월 18일

김윤명 digitallaw@naver.com

2007년 경희대학교에서 지식재산법 전공으로 박사학위를 받았다. 네이버 정책수석, 소프트웨어정책연구소, 국회 보좌관 등을 지냈다. 현재는 디지털정책연구소 소장으로, 국가지식재산위원회 보호분과 위원, 전남대학교 데이터사이언스대학원에서 데이터사이언스의 법과 윤리를 강의하고 있다. 주요 관심분야는 인공지능법, 지식재산법, 데이터 윤리와 법, 콘텐츠법 등이다. 생성형 AI 창작과 지식재산법(2024), 인공지능법(2022)(교육부 우수학술도서), 게임법(2021)(문화부 세종도서), 게임서비스와 법(2014)(문화부 세종도서) 등의 책을 썼다. 사진을 취미로 하며, 사진집을 준비 중에 있다.