

■ 논 단

개인정보보호법에 의한 자동화된 결정의 규율

이 상 용

건국대학교 법학전문대학원 교수 / 박사, 제1저자

이 혜 리

원광대학교 법학전문대학원 부교수 / 박사, 교신저자

요 약 문

자동화된 결정은 그 효율성과 일관성으로 인해 점점 증가하고 있지만, 인간의 존엄성, 프라이버시와 자유, 민주주의, 공정성 등과 관련하여 우려가 제기되기도 한다. 이러한 우려를 완화하기 위하여 개인정보보호법제, 차별금지법제, 인공지능규제법제, 행정법제 등 여러 법제도적 방안이 검토되어 왔지만, 그 중에서도 포괄적이고 사전적이며 수단적인 성격을 띠는 개인정보보호법제가 가장 주목을 받고 있다.

개정 개인정보보호법 제37조의2로 신설된 자동화된 결정에 대한 정보주체의 대응권 규정은 그 모태인 GDPR보다 균형 있고 합리적인 규율을 시도하였다는 점에서 높이 평가할 만하다. 그럼에도 불구하고 보다 바람직한 운영을 위해서는 몇 가지 고려할 점이 있다. 첫째, 기술적 한계를 고려할 필요가 있다. 사회관념상 기술적으로 불가능한 것을 요구할 수는 없기 때문이다. 특히 인공지능 기술이 지닌 블랙박스적 특성, 성능과 투명성의 상충관계, 공정성 지표의 한계 등을 고려할 필요가 있다. 둘째, 신기술에 대한 절차적·수단적 규제로서의 특성을 고려하여 경직되고 과도한 규제를 피할 필요가 있다. 셋째, 상황과 맥락을 충분히 고려하여 운영될 필요가 있다. 특히 사적 영역은 공적 영역과 달리 사적자치와 이익형량이 우선적으로 고려되어야 함을 염두에 두어야 한다. 넷째, 자율과 후견의 적정한 조화를 모색할 필요가 있다. 자동화된 결정에 관한 위험을 통제하기 위하여 정보주체에게 대응권을 부여한 것 자체가 자율적 접근을 취한 것이다. 일부 후견적 접근이 필요한 경우에도 자율규제를 우선적으로 검토해야 할 것이다.

주제어 : 인공지능, 자동화된 결정, 자동화된 의사결정, 자동적 처분, 공정성, 개인정보보호법, GDPR, 차별금지법

— <目 次> —

I. 서 설	V. 자동화된 결정에 대한 대응권
II. 자동화된 결정이란 무엇인가	1. 거부권
1. 자동화된 결정의 개념	2. 설명요구권
2. 자동화된 결정과 프로파일링과의 관계	VI. 기술적 한계
3. 자동화된 결정과 인공지능 기술과의 관계	1. 기술적 한계가 의무에 미치는 영향
III. 무엇이 문제인가	2. 블랙박스의 문제
1. 자동화된 결정과 관련된 위험	3. 상충관계의 문제
2. 자동화된 결정과 공정성	4. 공정성 지표의 문제
3. 소결	VII. 바람직한 운영방안
IV. 어떻게 문제를 해결할 것인가	1. 절차적·수단적 규제로서의 특성 고려
1. 서설	2. 상황과 맥락의 고려
2. 개인정보보호법제	3. 자율적 접근과 후견적 접근의 조화
3. 인공지능규제법제	VIII. 결 론
4. 행정법제	
5. 차별금지법제	
6. 소결	

I. 서 설

판단과 의사결정을 자동화하는 인공지능 기술의 발전으로 개인의 법적 지위에 영향을 미치는 의사결정들 역시 빠른 속도로 자동화되고 있다. 이처럼 자동화된 의사결정은 인간에 의한 의사결정에 비하여 높은 효율성과 일관성을 보여주지만, 동시에 여러 가지 우려를 불러일으키기도 한다.

이에 대한 법제도적 대응의 하나로서 마련된 개정 개인정보보호법(이하 ‘개정법’이라고만 한다)¹⁾ 제37조의2는 정보주체에게 자동화된 결정에 대한 대응권으로서 거부권과 설명요구권

1) 2023. 3. 14. 법률 제19234호로 개정되었다. 개정법은 2023. 9. 15. 시행되었으나 자동화된 결정에 대한 대응권을

을 부여하고 있다. EU GDPR²⁾ 제22조로부터 많은 영향을 받은 위 조항은 인공지능 시대를 살아가는 정보주체들의 권익을 보호해 줄 것으로 기대되고 있다. 그러나 자동화된 결정과 관련된 위험을 규율하기 위한 법제도에 개인정보보호법만 있는 것은 아니다. 자동적 처분을 규율하는 행정법제나 EU의 AI 법안으로 대표되는 인공지능 규제법제 역시 자동화된 결정의 규율에 동원될 수 있으며, 공정성 등의 문제를 다루기 위해 전통적인 차별금지법제 역시 활용될 수 있다.

본고는 개정법 제37조의2의 신설에 즈음하여 이 조항이 자동화된 결정에 대한 규율 체계에서 어떠한 역할을 담당할 수 있을지, 그리고 그 한계는 무엇이며 어떻게 운영되는 것이 바람직할 것인지에 관하여 포괄적이고 체계적으로 조망해보고자 한다. 이를 위해서 먼저 자동화된 결정과 관련된 위험이 무엇인지 식별하고, 그러한 위험에 대응하기 위한 여러 법제도들의 장단점을 검토한 뒤 특히 개인정보보호법제가 지니는 강점과 약점을 분석한다. 그리고 자동화된 결정에 이용되는 기술의 특성에 기인하는 규율의 한계를 살펴본 뒤, 위 조항의 바람직한 운영방안을 제시해보기로 한다.

II. 자동화된 결정이란 무엇인가

1. 자동화된 결정의 개념

넓은 의미의 자동화된 결정이란 어떠한 주체의 법적 지위나 이익에 영향을 미치는 판단이나 의사결정의 전부 또는 일부가 자동화된 경우를 말한다고 볼 수 있다. 그러나 넓은 의미의 자동화된 결정 모두가 법적인 규율이 요구되는 위험을 초래하지는 않을 것이다. 법적 규율의 대상이 되는 자동화된 결정의 개념은 개인정보보호법에 가장 명확하게 정의되어 있다. 즉, 개정법에 의하면 자동화된 결정이란 “완전히 자동화된 시스템(인공지능 기술을 적용한 시스템을 포함한다)으로 개인정보를 처리하여 이루어지는 결정”을 말한다(제37조의2 제1항).³⁾ 이러한 정의는 “자동화된 처리에만 기초한 결정(decision based solely on automated processing)”에 관하여 규율하고 있는 EU GDPR 제22조의 영향을 받은 것이다. 즉, 부분적으로만 자동화되어 인간의 개입이 있는 결정은 위 조항에서 말하는 자동화된 결정에는 해당하지 않는다. 본고는 개인정보보호법이 자동화된 결정의 규율에서 담당하는 역할과 그 한계 및 바람직한 운영방안

규정한 제37조의2는 2024. 3. 15.부터 시행된다.

2) Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

3) 다만, 행정기본법 제20조에 따른 행정청의 자동적 처분은 제외되어 있는데, 그 경위와 의미에 관하여는 뒤에서 자세히 설명한다.

을 분석하는 것을 목표로 하는 만큼 원칙적으로 개인정보보호법의 정의를 채택하기로 한다. 다만 간혹 맥락에 따라 개인정보의 처리가 수반되지 않는 경우를 포함하여 ‘자동화된 결정’이라는 표현을 사용하는 경우도 있음을 밝혀둔다.

2. 자동화된 결정과 프로파일링과의 관계

개인에 대한 자동화된 결정은 많은 경우 개인에 대한 자동화된 평가를 수반하며 이를 위하여 프로파일링(profiling)이 이루어지게 된다. 일반적으로 프로파일이란 한 범주의 개인들을 특징짓는 일련의 데이터를 의미하고, 프로파일링은 이러한 프로파일을 개인에게 적용함으로써 이루어지는 자동적인 개인정보 처리기법을 말한다.⁴⁾ GDPR은 1995년의 EU 개인정보보호지침⁵⁾과 달리 명시적인 정의 규정을 두고 있는데, 이에 따르면 프로파일링이란 “자연인의 개인적인 특정 측면을 평가하기 위하여 행해지는 모든 형태의 자동화된 개인정보 처리”를 의미한다[§4(4)].⁶⁾ 아래에서는 GDPR이 정의하는 바에 따라 프로파일링의 개념을 사용하기로 한다. 이러한 개념에 의하면 개인적 측면의 평가에 이르지 못하는 개인정보의 처리는 프로파일링에 해당하지 않는다. 한편 개정법은 자동화된 결정에 관하여만 규율할 뿐 프로파일링에 관한 조항은 따로 두고 있지 않다.⁷⁾

프로파일링은 자동화된 결정과 밀접한 관련이 있지만 서로 같은 것은 아니다. 즉, 프로파일링은 개인에 대한 분석 또는 예측을 위하여 그 자체로 활용될 수도 있고, 나아가 개인에 대한 결정, 특히 자동화된 결정을 위하여 활용될 수도 있다.⁸⁾ 프로파일링은 자동화된 결정의 개념 요소 중 ‘자동화된 처리’의 대표적인 예에 해당하지만,⁹⁾ 프로파일링 없이 자동화된 결정이 이루어질 수도 있다.¹⁰⁾

-
- 4) Council of Europe, Recommendation CM/Rec(2010)13 of the Committee of Ministers to member states on the protection of individuals with regard to automatic processing of personal data in the context of profiling, 1. Definitions d and e, 2010.
- 5) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23/11/1995.
- 6) 위 조항은 특히 자연인의 업무상 성과, 경제적 상황, 건강, 개인선호, 관심사, 신뢰도, 행동, 위치 또는 이동에 관한 측면을 분석하거나 예측하기 위하여 행해지는 경우를 예시하고 있다.
- 7) 참고로 「신용정보의 이용 및 보호에 관한 법률」은 ‘자동화평가’에 관한 정의 규정을 두고 있다(제2조 제14호).
- 8) Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, p. 8; 박노형/정명현, “EU GDPR상 프로파일링 규정의 법적 분석”, 안암법학 제56호, 안암법학회, 2018, 289면.
- 9) GDPR §22(1).
- 10) 예를 들어 과속탐지 카메라 사진을 기초로 속도위반 과태료를 부과하는 것은 프로파일링을 수반하지 않는 자동화된 결정에 해당한다.(Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, p. 8.)

프로파일링 자체는 금지되어 있지 않지만,¹¹⁾ 개인정보 처리의 일종이므로 개인정보의 처리에 관한 원칙과 정보주체의 권리가 존중되어야 한다.¹²⁾ 예를 들어 정보주체는 프로파일링을 포함한 개인정보 처리의 정지를 요구할 수 있다(개인정보보호법 제37조). 프로파일링은 자동화된 결정과 구별되는 것임에도 불구하고 자동화된 결정에 대한 규제에 있어서 중요한 함의를 갖는다. 즉, 프로파일링을 비롯한 자동화된 개인정보의 처리가 자동화된 결정의 개념 요소로 들어감으로써 개인정보의 처리에 관한 규율을 정하고 있는 개인정보보호법이 자동화된 결정과 관련된 문제들을 해결하기 위한 법적 수단으로 활용될 가능성이 열리게 되는 것이다.

3. 자동화된 결정과 인공지능 기술과의 관계

자동화된 결정에 활용되는 알고리즘은 인공지능 기술만을 염두에 둔 것은 아니다.¹³⁾ EU 개인정보보호지침에 자동화된 결정과 관련된 규정이 도입된 1995년 당시는 이른바 ‘인공지능 겨울’에 해당하던 시기라는 점¹⁴⁾을 생각해보면 당연한 일이기도 하다. 하지만 지금은 인공지능 기술의 전성시대이다. 오늘날 인공지능 기술은 효율적인 금속 제련법을 찾아내고 암 진단을 도우며 불 만한 영화를 추천하고 창작을 보조하는 등 수많은 곳에 쓰이고 있지만, 무엇보다도 ‘판단과 의사결정을 위한 자동화 도구’로서 막강한 능력을 자랑하고 있다. 따라서 채용, 금융, 치안, 교정 등 다양한 영역의 판단 및 의사결정 과정에서 인공지능 기술이 활용되는 것은 자연스러운 현상이다. 실제로 자동화된 결정의 상당 부분은 인공지능 기술을 활용하여 이루어지고 있고, 그 비중은 지속적으로 늘어날 것으로 예상된다. 결국 자동화된 결정에 대한 규율이라는 문제는 많은 부분 인공지능에 대한 규율의 문제와 중첩되게 된다.

III. 무엇이 문제인가

1. 자동화된 결정과 관련된 위험

알고리즘에 근거한 자동화된 결정은 인간의 의사결정보다 효율적일 뿐만 아니라 일관성이

11) GDPR 제22조는 자동화된 처리에만 기초한 결정에 따르지 않을 정보주체의 권리를 규정한 것이지, 자동화된 처리인 프로파일링 그 자체를 제한하는 것은 아니다(BT-Drs. 16/10529, p. 13).

12) 프로파일링의 맥락에서 개인정보 처리의 원칙과 정보주체의 권리를 설명한 것으로는 인터넷진흥원, EU 일반 개인정보보호법 가이드북, 2020, 162면.

13) 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 72면.

14) 한국인공지능법학회, 인공지능과 법, 박영사, 2019, 7면(임영익 집필부분)

확보됨에 따라 보다 공정할 수도 있다는 측면에서 장점이 적지 않다.¹⁵⁾ 그럼에도 불구하고 자동화된 결정에 대하여는 인간의 존엄성, 프라이버시와 자유, 민주주의, 공정성 등과 관련하여 우려가 제기되기도 한다.

가. 인간의 존엄성

첫 번째 우려는 인간의 존엄성과 관계된 것이다. 알고리즘에 의한 자동화된 결정은 인간의 개입 없이 결정 대상자의 권리나 이익에 중대한 영향을 미칠 수 있다는 점에서 인간의 자율성과 존엄성을 위협할 수 있다는 것이다.¹⁶⁾ 이 문제는 알고리즘이 전적으로 객관적이고 중립적이라고 가정할 경우에도 여전히 발생할 수 있다.¹⁷⁾ 인간이 한낱 수단으로 취급될 때 인간의 존엄성은 침해되며, 인간이 주체성을 박탈당할 때 가치질서의 타락이 시작되기 때문이다.¹⁸⁾ 독일 헌법 제1조는 인간의 존엄성은 생명권보다 소중한 불가침의 본질적 가치라고 선언하고 있기도 하다.¹⁹⁾ 이러한 태도는 인간을 목적이 아닌 수단으로 취급해서는 안 된다는 칸트의 정언명령에 바탕을 둔 것이다.²⁰⁾

나. 프라이버시와 자유

두 번째 우려는 프라이버시와 관련된 것이다. 자동화된 결정은 많은 경우 결정 대상자의 개인적인 특정 측면을 평가하기 위한 프로파일링을 수반하는데, 이 과정에서 사생활의 비밀과 자유가 침해될 수 있기 때문이다. 나아가 자신의 행동이 관찰되고 분석된다는 인식은 자기검열을 통한 위축효과를 유발하여²¹⁾ 표현의 자유와 일반적 행동의 자유에도 부정적인 영향을 줄 수 있다.²²⁾

15) ICO, Automated decision-making and profiling, 2018. (<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/> 2023. 11. 27. 확인)

16) 박노형/정명현, “EU GDPR상 프로파일링 규정의 법적 분석”, 안암법학 제56호, 안암법학회, 2018, 311면; 권건보, “정보접근권의 실효적 보장에 관한 고찰 - 정보주체의 정보접근권을 중심으로 -”, 미국헌법연구 제29권 제2호, 미국헌법학회, 2018, 38면; 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 83면; Bygrave, Lee A, “Article 22 Automated Individual Decision-making, Including Profiling” in Christopher Kuner et al (Ed.), The EU General Data Protection Regulation (GDPR): A Commentary, Oxford University Press, 2020, pp. 283-315.

17) 김희정, “알고리즘 자동의사결정으로부터 개인의 보호 : 인간의 개입권을 중심으로”, 헌법학연구 제26권 제1호, 한국헌법학회, 2020, 225-226면.

18) 최규환, 인간존엄의 형량가능성, 헌법재판연구원, 2017, 9면.

19) 독일 헌법재판소는 테러리스트에 의해 납치된 것으로 의심되는 상업용 항공기를 격추시키는 것을 허용한 독일항공운송보안법이 인간을 항공기의 일부와 같이 대상으로 취급하고 숫자 또는 통계로 취급함으로써 독일 헌법 제1조를 위반하였다고 판시한 바 있다[BVerfG, Judgment of the First Senate of 15 February 2006 - 1 BvR 357/05 - paras. (1-156)].

20) 한희원, 정의로의 산책(제2판), 삼영사, 2011, 214-216면.

다. 민주주의

세 번째는 자동화된 결정이 민주주의의 기반을 침해할지 모른다는 우려이다. 캠브리지 애널리티카라는 영국의 정치컨설팅 회사는 페이스북 앱을 통해 얻은 미국인들의 정보를 이용해 ‘선택 가능자’를 분류하여 맞춤형으로 콘텐츠를 노출시킴으로써 그들의 생각에 영향을 미쳐 트럼프의 대선 승리를 도우려 했다는 폭로는 이러한 위험을 방증한다.²³⁾ 이처럼 노골적 의도가 없더라도 이른바 에코 체임버(echo chamber)라는 현상을 통해 민주주의의 기반을 약화시킬 수 있다.²⁴⁾ 상업적 이익을 위해 설계된 알고리즘은 사람들의 입맛에 맞는 콘텐츠만을 배포함으로써 다른 의견이나 내용을 접하기 어렵게 만들며, 결국 알 권리를 제한하고 시민으로서의 능력을 감소시킬 수 있기 때문이다.²⁵⁾

라. 공정성

네 번째 우려는 공정성, 즉 차별 및 편견과 관련된 것이다. 인간의 주관성이나 비합리성을 극복할 것으로 기대되었던 알고리즘에 의한 자동화된 결정이 오히려 차별과 편견을 조장할 수도 있다는 것이다.²⁶⁾ 이러한 우려는 많은 사람들의 공감을 사기도 하지만 동시에 치열한 논쟁의 대상이 되기도 한다. 공정성의 우려는 자동화된 결정에 대한 규제의 논거로서 가장 많이 거론되는 것인 만큼 이에 관하여는 항을 달리하여 좀 더 자세히 살펴보기로 한다.

-
- 21) Hildebrandt, Mireille, “Profiling and the Identity of the European Citizen”, in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives*, Springer, 2008, p. 307.
- 22) 김희정, “알고리즘 자동의사결정으로부터 개인의 보호 : 인간의 개입권을 중심으로”, *헌법학연구* 제26권 제1호, 한국헌법학회, 2020, 215면.
- 23) Independent Global News, “The Great Hack : Big Data firms Helped Sway the 2016 Election. Could it be Happened Again in 2020?” (2020. 1. 7.); 타겟으로 설정된 사람들에게만 특정 정보를 차별적으로 제공하여 민주주의를 침해해버리는 것을 ‘디지털 게리맨더링’이라고 표현하기도 한다(Zittrain, Jonathan, “Engineering an Election: Digital gerrymandering poses a threat to democracy”, *Harvard Law Review Rorum*, 127, 2014, pp. 336-337).
- 24) 김희정, “알고리즘 자동의사결정으로부터 개인의 보호 : 인간의 개입권을 중심으로”, *헌법학연구* 제26권 제1호, 한국헌법학회, 2020, 216면. 본래 잔향 효과를 위한 폐쇄 공간을 뜻하던 에코 체임버라는 용어는 SNS 서비스의 확산에 따라 신념이나 견해가 비슷한 사람들이 서로 정보나 뉴스를 공유함으로써 기존의 신념이나 견해에 대한 확신이 강화되고 증폭되는 현상을 의미하게 되었다.
- 25) Hildebrandt, Mireille/Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives*, Springer, 2008, p. 331, p. 551.
- 26) 양종모, “인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율방안”, *법조* 제72호, 법조협회, 2017, 60-104면; 남중권, “머신러닝 알고리즘의 데이터 처리에 대한 법적 제한의 한계: 개인정보보호와 차별금지의 측면에서”, *과학기술과 법* 제10권 제1호, 충북대학교 법학연구소, 2019, 65-94면.

2. 자동화된 결정과 공정성

가. 공정성의 의미

공정성의 의미는 실체적 공정성과 절차적 공정성, 그리고 개인적 공정성과 집단적 공정성으로 나누어 파악해볼 수 있다.²⁷⁾

실체적 공정성은 다의적으로 파악될 수 있다. 예컨대 응보적 정의, 분배적 정의,²⁸⁾ 교정적 정의, 심지어 ‘계약은 지켜져야 한다(pacta sunt servanda)’와 같은 원칙까지도 공정성의 범주에 넣을 수 있다.²⁹⁾ 실체적 공정성 개념이 지닌 다의성과 상호 모순의 문제를 회피하기 위해 절차적 공정성 개념이 활용될 수 있다. 절차적 공정성은 수범자들로 하여금 자발적으로 법을 준수하도록 설득함으로써 법의 지배에 있어서 핵심적 역할을 담당한다.³⁰⁾ 특히 인공지능 알고리즘의 경우 그 위험이 어떤 양상으로 나타날지 예측하기 어렵다는 점에서 절차적 공정성이 더욱 중요한 역할을 하며, 이를 위하여 책임성과 투명성이 강조되고 있다.³¹⁾

개인적 공정성과 집단적 공정성은 공학 연구자들에 의해 제시된 여러 공정성 지표들이 기반하고 있는 개념으로서 대체로 평등 원칙의 관점에서 정의된다. 개인적 공정성은 ‘같은 것은 같게, 다른 것은 다르게’ 대우하는 것이 공정하다는 아리스토텔레스의 정식에 기반한 것이다. 이에 비해 집단적 공정성은 인공지능이 내놓은 결과가 여러 사회 집단에 대해 동등할 것을 요구하는 것을 말한다. 이는 특정 집단에 속한 개인들이 역사적 차별로 인해 유사하게 불리한 특성을 가지고 있다면 이들을 유사하게 대우하더라도 여전히 차별적일 수 있다는 인식을 기반으로 한다.

나. 차별이나 편향의 발생 경로

알고리즘이 차별이나 편향을 보이는 경로는 의도적인 경우와 비의도적인 경우로 나누어볼 수 있다.³²⁾ 의도적 차별은 성별이나 인종과 같은 요소를 직접적으로 이용하여 이루어질 수도

27) 김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 46-59면. 다만 이러한 기준은 생성 모델에는 적용하기 어려운 점이 있다. 위 논문은 생성 모델의 공정성 기준으로서 분배적 위해와 표상적 위해의 관념을 제안하고 있다(위 논문 60-70면).

28) 분배적 정의는 효율성과 상충될 수 있는데, 법경제학자들은 대체로 시장에 직접 개입하여 효율성을 떨어뜨리는 방식이 아니라 사후적으로 소득을 이전하는 재분배를 통하여 분배적 정의를 달성하는 방식을 선호한다(이상용, “데이터 거래의 법적 기초”, 법조 제67권 제2호, 법조협회, 2018, 16-18면).

29) Kaplow, Louis/Steven Shavell. “The conflict between notions of fairness and the Pareto principle”, American Law and Economics Review 1.1, 1999, p. 64.

30) Tyler, Tom R., “Procedural fairness and compliance with the law”, Revue Suisse D Economie Politique et de Statistique 133, 1997: 219-240.

31) OECD, Recommendation of the Council on Artificial Intelligence, OECD Publishing, 2019, Principle 1.3.

있지만, 이를 간접적으로 추정할 수 있는 대용변수(proxy variable)들을 이용하여 이루어지기도 한다.³³⁾ 그러나 보다 흔히 찾아볼 수 있는 것은 비의도적 차별이다. 기계학습 알고리즘에서 의도하지 않은 차별이나 편견이 나타나는 것은 학습데이터의 불완전이나 불균형에 원인이 있을 수도 있지만, 학습데이터에 기존의 사회적 편향이 그대로 반영되어 있기 때문일 수도 있다.³⁴⁾ 일부 견해는 기존의 역사적 편향이 반영된 알고리즘이 생성한 데이터가 다시금 학습에 활용되어 피드백 루프가 형성된다면 의도하지 않게 역사적 편향을 영구화하고 촉진할 수 있다고 우려하기도 한다.³⁵⁾

다. 공정성 침해 사례

실제로 알고리즘이 차별이나 편향을 보인 것으로 지적된 사례도 여럿 있다. 먼저 채용 분야에서는 구글의 광고가 여성보다는 남성에게 고소득 직종의 직업 광고를 제시한 것이나,³⁶⁾ 아마존이 활용하려던 채용 시스템이 ‘여성’이라는 표현이 들어간 지원서를 채용 대상에서 배제한 것으로 나타나 결국 폐기된 사례³⁷⁾가 대표적이다. 교육 분야에서는 영국 자격검정청의 학생 평가 알고리즘이 사립학교에 다니는 학생들에게는 유리하게 작용하고 어려운 배경의 학생들에게 불리하게 작용하여 많은 비판을 받았다.³⁸⁾ 상거래나 금융 영역에서도 불공정 사례들이 발견되었는데, 미국에서 튜터링 서비스의 가격 책정 알고리즘이 백인보다 아시아인에게 2배 가까운 요금을 책정하였다거나,³⁹⁾ 골드만삭스와 애플이 출시한 애플카드와 관련하여 한 여성에게 낮은 신용 한도가 책정되는⁴⁰⁾ 등의 사례가 알려진 바 있다.

32) 의도적 차별과 비의도적 차별의 구분은 미국의 차별금지법제의 틀을 따른 것이다.

33) 오요한/홍성욱, “인공지능 알고리즘은 사람을 차별하는가?” 과학기술학연구 제18권 제3권, 한국과학기술학회, 2018, 162면.

34) ICO, ICO consultation on the draft AI auditing framework guidance for organisations, 2020 참조; 미국 백악관은 2016년의 보고서를 통해 알고리즘이 편향된 데이터를 반영할 수 있는 네 가지 경로를 제시한 바 있다(Executive Office of the President, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, 2016, pp. 7-8); 기계학습의 단계에 따라 비의도적인 편향이 발생할 수 있는 경로를 기술적으로 상세히 설명한 것으로는 오요한/홍성욱, “인공지능 알고리즘은 사람을 차별하는가?” 과학기술학연구 제18권 제3권, 한국과학기술학회, 2018, 162-166면.

35) 오요한/홍성욱, “인공지능 알고리즘은 사람을 차별하는가?” 과학기술학연구 제18권 제3권, 한국과학기술학회, 2018, 194면; Executive Office of the President, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, 2016, pp. 7-8; 알고리즘으로 도출된 결과를 인간이 또 다시 학습한다면 인간의 편향성이 더욱 고착화될 수도 있다고 주장하는 견해도 있다.(김예진, “인공지능에 의한 자동화된 의사결정의 위험성 - 장애인 채용을 중심으로 -”, Law & technology 제18권 제1호, 서울대학교 기술과법센터, 2022, 39면).

36) The Guardian, “Women less likely to be Shown Ads for High-paid Jobs on Google, Study Shows” (2015. 7. 8).

37) Reuters, “Amazon scraps secret AI recruiting tool that showed bias against women”, (2018. 10. 10.)

38) Smith, Helen, “Algorithmic bias: should students pay the price?”, AI & Society 35.4, 2020, pp. 1077-1078.

39) Angwin, J., et al., “The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review”, ProPublica, 2015.

이처럼 민간 영역뿐만 아니라 공공 영역에서도 문제가 지적되고 있다. 예컨대 치안 영역에서 ‘프레드폴(PredPol)’이라는 범죄 예측 알고리즘은 우범 지역을 인간보다 높은 정확도로 예측함으로써 범죄율을 낮추는 데 기여한 것으로 유명해졌다.⁴¹⁾ 그러나 이에 대하여는 프라이버시와 적법적처를 침해하고 편견과 차별을 고착화한다는 이유로 많은 비판이 이루어졌다. 즉, 주로 거리에서 벌어지고 검거율이 높은 경범죄를 데이터에 포함시킨 결과 흑인들이 많이 거주하는 빈민가가 범죄의 핫스팟으로 분류되어 집중적으로 순찰이 이루어지고, 그에 따른 검거 결과가 다시 데이터로 사용되는 피드백 효과가 생긴다는 것이다.⁴²⁾ 그러나 가장 유명한 것은 ‘컴파스(COMPAS)⁴³⁾’ 사건이다. 미국의 탐사보도지인 프로퍼블리카(ProPublica)는 2016년 미국 법원과 교도소에서 형량, 가석방, 보석 등의 결정에 널리 사용되던 컴파스 알고리즘이 흑인 남성들에게 편파적인 결과를 가져왔다고 폭로했다.⁴⁴⁾ 이 문제는 결국 법정 사건으로 비화했다. 컴파스에 의하여 고위험군으로 분류되어 결국 중형을 선고받은 루미스(E. Loomis)라는 청년이 그 판결에 불복하면서, 남성을 더 위험하다고 판단하는 불공정성을 내포하고 있는 컴파스의 활용으로 인해 공정하게 재판받을 권리를 침해당했다고 주장한 것이다. 그러나 위스콘신 주 대법원은 위 주장을 받아들이지 않았다.⁴⁵⁾ 컴파스는 실제로 여성보다 높은 남성의 강력범죄율을 반영하여 위험점수를 산출했을 뿐이고, 이는 판결에서 고려된 여러 요소 중 하나일 뿐이라는 이유에서였다.

라. 반론과 재반론

이처럼 차별과 편견에 따른 공정성의 침해 우려는 자동화된 결정에 대한 규제의 강력한 근거가 되고 있지만, 이러한 우려가 과장되었다고 비판하는 견해도 많다. 이에 따르면 인간은 기억의 한계나 피로와 같은 물리적 한계가 있고, 각종 심리학적 편향에 노출되어 있으며, 선호체계가 합리적이지 못하고 집단적 사고에 취약하지만, 알고리즘에 의한 의사결정은 정확하고 일

40) New York State Department of Financial Service(NYDFS), Report on Apple Card Investigation, 2021.

41) Santa Cruz Sentinel, “Modest Gains in First Six Months of Santa Cruz’s Predictive Police Program” (2012. 2. 26.)

42) O’Neil, C., Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2016; Selbst, Andrew, “Disparate Impact in Big Data Policing”, Georgia Law Review, 52, 2018, pp. 131-137.

43) COMPAS는 Correctional Offender Management Profiling for Alternative Sanctions의 머리글자를 따서 만든 약어이다.

44) Angwin, J., et al., “Machine Bias There’s software used across the country to predict future criminals. And it’s biased against blacks”, ProPublica, 2016.

45) Loomis v. Wisconsin, 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017). 위 판결은 2017. 6. 26. 연방대법원이 피고인의 상고허가신청을 기각함에 따라 확정되었다.(<https://www.supremecourt.gov/docketfiles/16-6387.htm>)

관적이며 효율적이라는 것이다.⁴⁶⁾

인간이 과연 알고리즘보다 공정한가에 대한 의문도 제기된다. 인간의 결정 역시 의식적, 혹은 무의식적으로 편견과 차별의 영향을 받기 때문이다.⁴⁷⁾ 그런 점에서 기계학습 기술을 활용한 모형이 재범 위험성 예측도를 높이고 인간 판사들의 편향을 줄일 수 있다는 연구 결과⁴⁸⁾는 주목할 만하다. 최근에 국내에서 간호사 채용에 활용된 인공지능 역량검사 결과가 채용 후 성과를 유의미하게 예측한 반면 인간 면접관에 의한 면접 결과는 이를 전혀 예측하지 못했다는 연구 결과가 나오기도 했다.⁴⁹⁾ 무엇보다 인공지능 기술은 기존의 차별을 드러내고 시정할 수 있는 기회를 제공할 수 있으며,⁵⁰⁾ 나아가 편향되거나 차별적인 데이터를 수정 또는 제거함으로써 공정성을 제고할 수도 있다.⁵¹⁾ 예컨대 금융 영역에서 인공지능은 인간에 의한 차별을 줄이고 포괄적 금융을 확장하는 데 기여할 수도 있다.⁵²⁾

그러나 위와 같은 반론에도 불구하고 차별과 편견에 대한 우려는 여전히 자동화된 결정에 대하여 적정한 수준의 규제가 마련되어야 하는 근거가 될 수 있다. 자동화된 결정은 대규모로 이루어져 인간의 결정에 비하여 영향력이 매우 크고,⁵³⁾ 현재의 기술 수준으로 반영하기 어려운 요소들이나 학습과정에서 미처 반영하지 못한 새로운 요소들을 반영시키기 위해 인간의 개입이 필요할 수도 있기 때문이다.⁵⁴⁾ 또한 인공지능 기술의 차별 시정적 효과를 누리기 위해서는 적정 수준의 투명성이 확보될 필요가 있기도 하다.

3. 소결

위에서 본 우려들은 자동화된 결정의 본질로부터 자연스럽게 나오는 것이다. 인간의 개입

46) Coglianese, Cary/Alicia Lai, "Algorithm vs. Algorithm", Duke Law Journal 72, 2022, pp. 1318-1322.

47) Kuner, Christopher, et al., "Machine learning with personal data: is data protection law smart enough to meet the challenge?", 7.1 International Data Privacy Law, 2017.

48) Kleinberg, Jon, et al., "Human decisions and machine predictions", The quarterly journal of economics 133.1, 2018; 인간의 판단이 컴파스 알고리즘의 판단보다 우수하다는 연구 결과로는 Dressel, J./H. Farid., 'The Accuracy, Fairness, and Limits of Predicting Recidivism', Science Advances, 4.1, 2018.

49) 허영은, "채용, 왜 바꾸기 어려울까요?", HQ Conference Q. (2022. 11. 17.).

50) Kleinberg, Jon, et al., "Discrimination in the Age of Algorithms", Journal of Legal Analysis 10, 2018.

51) Kamarinou, Dimitra, et al., "Machine Learning with Personal Data", Queen Mary University of London Legal Studies Research Paper No. N247/2016, 2016.

52) Bartlett, Robert, et al., "Consumer-lending discrimination in the FinTech Era", Journal of Financial Economics, 143.1, 2022, pp. 30-56. 위 연구에서는 핀테크 대출에서는 대면 심사에서 존재하였던 소수 인종에 대한 차별 정도가 완화되었다고 보고한다.

53) 김희정, "알고리즘 자동의사결정으로부터 개인의 보호 : 인간의 개입권을 중심으로", 헌법학연구 제26권 제1호, 한국헌법학회, 2020, 217면.

54) 황지은, "인공지능 관점에서 GDPR 제22조의 인간 개입에 대한 분석", 경제규제와 법 제13권 제1호, 서울대학교 법학연구소, 2020, 43-44면

없는 ‘자동화’된 결정이기에 인간의 존엄성과 자율성에 대한 우려가 생기고, 프로파일링 등을 통한 ‘평가’를 수반하기 때문에 차별이나 편견과 같은 공정성에 대한 우려가 생기며, 많은 경우 ‘개인적 측면’에 대한 평가가 이루어지기 때문에 프라이버시에 대한 우려가 생기고, 그러한 평가가 시민의 감시와 조종으로 이어질 수 있기 때문에 민주주의에 대한 우려가 생기는 것이다.

IV. 어떻게 문제를 해결할 것인가

1. 서설

자동화된 결정 과정에서 생길 수 있는 인간의 존엄성, 프라이버시와 자유, 민주주의, 공정성 등의 문제를 어떻게 해결해야 할까. 여러 방법이 있겠지만 법적인 대응방안으로는 개인정보보호법제, 차별금지법제, 인공지능규제법제 등이 고려될 수 있다. 공적 영역에서는 행정절차법 등이 고려될 수도 있다. 이들은 각기 장단점을 지니고 있는데, 아래에서 간략히 살펴보기로 한다.

2. 개인정보보호법제

개인에 대한 자동화된 결정은 대부분 자동화된 개인정보의 처리를 수반하게 된다. 이를 통해 개인정보보호법제가 자동화된 결정도 규율할 수 있는 계기가 마련된다. GDPR(§22)과 개정법(제37조의2)은 자동화된 결정에 대응할 수 있는 정보주체의 권리를 부여하고 있다.

그런데 개인정보자기결정권의 헌법적 근거를 생각해볼 때 이를 기반으로 하는 개인정보보호법이 자동화된 결정에 결부된 모든 위험을 충분히 다룰 수 있는지는 의문의 여지가 있다. 개인정보자기결정권의 핵심적 근거라고 할 수 있는 사생활의 비밀과 자유라는 측면에서 보면 프라이버시의 문제 해결에는 적합할 수 있지만,⁵⁵⁾ 그 외의 문제에 대응하기에는 태생적으로 어려운 부분이 있기 때문이다.⁵⁶⁾ 그러나 헌법재판소는 독자적 기본권으로서 개인정보자기결정권의 헌법적 근거를 사생활의 비밀과 자유뿐만 아니라 인간의 존엄과 가치 및 행복추구권에 근

55) 박상돈, “헌법상 자동의사결정 알고리즘 설명요구권에 관한 개괄적 고찰”, 헌법학연구 제23권 제3호, 한국헌법학회, 2017, 210면.

56) 김희정, “알고리즘 자동의사결정으로부터 개인의 보호: 인간의 개입권을 중심으로”, 헌법학연구 제26권 제1호, 한국헌법학회, 2020, 218면; 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -”, 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 151-152면; 한편 개인정보자기결정권은 사적 영역의 보호를 위하여 인정되는 전통적인 프라이버시권과 차이가 있다고 보는 견해로는 권건보, “개인정보자기결정권”, 헌법 판례 100선, 한국헌법학회, 2012, 334쪽.

거를 둔 일반적 인격권, 그리고 자유민주적 기본질서 등에서도 찾고 있다.⁵⁷⁾ 그렇다면 개인정보보호법의 취지는 인간의 존엄성 내지 인격권을 보호하고⁵⁸⁾ 자유민주주의의 근간을 지키려는 것으로 확장하여 이해할 수도 있을 것이다.⁵⁹⁾ 이에 비하여 차별이나 편향과 같은 공정성의 문제는 개인정보에 대한 정보주체의 통제와는 직접적인 관련성이 부족하다.⁶⁰⁾ 다만 EU에서는 개인정보의 처리에 수반되는 전형적인 위험의 하나로 인식되어 공정성의 원칙이 일찍부터 개인정보처리에 관한 원칙으로 자리잡았다는 점을 고려하면, 간접적으로는 공정성과도 관련이 있다고 볼 수 있을 것이다.⁶¹⁾

이처럼 개인정보보호법은 자동화된 결정에 결부된 위험들을 직·간접적으로 다루고 있지만, 본질적으로 예방적이고 수단적이며 절차적인 방식으로 대응한다는 특징이 있다.⁶²⁾ 예컨대 개인정보보호법제는 차별적으로 활용될 수 있는 민감한 개인정보의 수집 자체를 금지함으로써 차별을 예방하는 기능을 수행할 수 있으며, 정보제공을 받을 권리나 자동화된 결정에 대한 대응권 등을 통해 절차적 공정성을 확보할 수 있다.

다만, 개인정보보호법제는 알고리즘을 통제하는 자가 아니라 정보처리자를 수범자로 한다는 점에서 한계가 있다.⁶³⁾ 법경제학적 관점에서 위험의 적절한 배분방법으로 널리 받아들여지고 있는 ‘최소 비용 회피자의 원칙(the least-cost avoider principle)’은 최소 비용으로 위험을 회피할 수 있는 사람에게 위험을 배분하는 것이 효율적일 뿐 아니라 공정하다고 본다.⁶⁴⁾ 자동화된

57) 헌법재판소 2005. 5. 26. 선고 99헌마513 결정. 그러나 일반적 인격권이나 자유민주적 기본질서는 이론적으로 개인정보자기결정권의 근거가 되기 어렵다고 주장하는 견해도 있다. 사생활의 비밀과 자유에 관한 규정이 없는 독일 기본법에서와 달리 우리 헌법에서는 일반적 인격권에 의존할 필요성이 적으며(박상돈, “헌법상 자동의사결정 알고리즘 설명요구권에 관한 개괄적 고찰”, 헌법학연구 제23권 제3호, 한국헌법학회, 2017, 210면), 객관적 법원리인 민주주의 원리 등으로부터 주관적 권리가 발생한다고 보기는 어렵기 때문이다(정태호, “개인정보자기결정의 헌법적 근거 및 구조에 관한 고찰 - 동시에 교육행정정보시스템(NEIS)의 위헌 여부의 판단에의 그 응용 -”, 헌법논총 제14집, 헌법재판소, 2003, 420면).

58) 김희정, “알고리즘 자동의사결정으로부터 개인의 보호 : 인간의 개입권을 중심으로”, 헌법학연구 제26권 제1호, 한국헌법학회, 2020, 227-228면; 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 83면.

59) 헌법재판소는 개인정보자기결정권을 독자적 기본권으로 승인해야 하는 이유가 “궁극적으로는 개인의 결정의 자유를 보호하고, 나아가 자유민주체제의 근간이 총체적으로 훼손될 가능성을 차단하기 위하여 필요한 최소한의 헌법적 보장 장치를 만들기 위함”에 있음을 밝히고 있다(헌법재판소 2005. 5. 26. 선고 99헌마513 결정).

60) 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -”, 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 151-152면 참조.

61) GDPR §5(a). 이에 비하여 우리 개인정보보호법상 개인정보보호의 원칙에는 공정성의 원칙이 명시되어 있지 않고(제3조), 단지 정보주체의 권리 중 하나로서 ‘개인정보의 처리로 인하여 발생한 피해를 신속하고 공정한 절차에 따라 구제받을 권리’가 인정될 뿐이다(제4조).

62) 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 83면 참조.

63) 김희정, “알고리즘 자동의사결정으로부터 개인의 보호 : 인간의 개입권을 중심으로”, 헌법학연구 제26권 제1호, 한국헌법학회, 2020, 220면.

64) 이상용, “인공지능과 계약법”, 비교사법 제23권 제4호, 한국비교사법학회, 2016, 1669면.

결정과 관련된 위험을 최소 비용으로 회피할 수 있는 자는 알고리즘 통제자이지 정보처리자가 아니다. 그 둘이 일치하는 경우라면 문제가 없겠지만, 많은 경우에 그러하듯이 이들이 서로 달라지는 경우라면 비효율적이고 잠재적으로 불공정한 규제 프레임워크가 될 수도 있다.

3. 인공지능규제법제

자동화된 결정 과정에서 생기는 문제들은 근본적으로 자동화 알고리즘, 많은 경우 인공지능 알고리즘에서 비롯된다. 이 때문에 인공지능 기술의 개발 및 이용 자체를 규율하려는 시도가 이루어지고 있다. 초기에는 인공지능 윤리에 의한 연성 규범에 의존했지만, 최근에는 EU의 인공지능 법안⁶⁵⁾이나 미국의 알고리즘 책임 법안⁶⁶⁾을 비롯하여 각국에서 입법 시도가 이루어지고 있으며, 우리나라도 예외가 아니다.⁶⁷⁾ 이러한 시도에 대하여는 평가가 엇갈리고 있다. 인공지능에 대한 규율은 어떠한 구조와 내용을 지녀야 할까. 이는 인공지능의 활용 양상, 그리고 기술적 특성과 관련이 있다.

먼저 인공지능의 활용 양상은 규제의 구조에 영향을 미친다. 인공지능 산업은 범용성을 지닌 신기술 확산으로 인한 파괴적 혁신의 전형적 모습을 보인다. 이는 신산업 규제의 일반적 원리가 그대로 적용될 수 있음을 의미한다.⁶⁸⁾ 즉, 인공지능 기술의 개발과 활용으로 인한 변화의 속도가 빠르고 광범위하며, 그 과정에서 어떠한 위험이 발생할 것인지 사전적으로 예측하기 어렵기 때문에 경직된 규제보다 유연한 규제가 요구된다는 것이다. 다만 유연한 규제가 곧 약한 규제를 의미하는 것이 아니라는 점은 유의할 필요가 있다.⁶⁹⁾

한편 인공지능의 기술적 특성은 규제의 내용에 영향을 미친다. 인공지능을 정의하는 방법에는 여러 가지가 있지만, 공학적 측면에서 가장 유용한 관점은 스스로 이상적 성과를 낼 수 있

65) EU 집행위원회가 초안을 제출한 이후 ChatGPT 등 대규모 언어모델(LLM)을 비롯한 새로운 인공지능 기술이 급속도로 확산하면서 큰 영향을 미치게 됨에 따라 EU 의회는 기반모델(foundation model)에 관한 규율을 추가하는 등 위 초안에 상당한 수정을 가하였고, 2023. 6. 14. 위 수정안이 유럽의회에서 가결되었다.

66) Algorithmic Accountability Act of 2022.

67) 국회 과학기술방송정보통신위원회는 2023년 2월에 기존에 발의된 7개 법안을 폐기하고 위원회 대안(AI 산업 육성 및 신뢰 기반 조성 등에 관한 법률안)을 합의하였지만, 위 법안은 아직 상임위원회를 통과하지 못한 상황이다. 디지털데일리, “‘갈 길 바쁜데...’ 해 넘긴 AI 법안에 애타는 과기정통부”, 2024. 1. 3. (<https://www.ddaily.co.kr/page/view/2024010218073406582>)

68) 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -”, 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 141면.

69) 규제 방식은 강도(強度)와 경도(硬度)를 기준으로 유형화할 수 있다. 먼저 자유 제한의 정도, 즉 규제 강도를 기준으로 강한 규제와 약한 규제로 나눌 수 있다. 공적규제, 사전규제, 포지티브 규제를 강한 규제라고 한다면, 자율규제, 사후규제, 네거티브 규제는 약한 규제라고 할 수 있다. 다음으로 자유 제한의 경직성, 즉 규제 경도에 따라서 경직적 규제와 유연한 규제로 나눌 수 있다. 규정 중심 규제, 내용적 규제, 이분법적 규제, 공적 규제가 경직적 규제라면, 원칙 중심 규제, 절차적 규제, 잠정적·적용적 규제, 자율규제는 유연한 규제라고 할 수 있다. 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -”, 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 140면.

는 ‘합리적 에이전트’로 보는 것이라고 할 수 있다.⁷⁰⁾ 이러한 기술적 지향성으로 인해 인공지능 알고리즘은 몇 가지 특성을 갖게 된다. 즉 인간의 개입 없이 작동한다는 점에서 자율성, 특정한 영역의 작업을 인간 못지않게 또는 인간보다 잘 수행한다는 의미에서 합리성을, 그리고 때로는 인간과 구별하기 어렵거나 공감을 불러일으킨다는 의미에서 인간과의 유사성을 지니는 것이다.⁷¹⁾ 이러한 특성들은 생산성 향상과 혁신을 추동하기도 하지만 동시에 몇 가지 문제를 일으키기도 한다.⁷²⁾ 예컨대 자율성은 책임 소재를 불분명하게 하고 인간의 존엄성을 위협할지 모른다는 우려를 불러일으킨다. 합리성 역시 위법한 목적으로 오용되거나, 정확성에만 초점을 맞추어 사회의 편견과 불평등을 그대로 반영할 경우에는 문제가 될 수 있다. 그리고 인간과의 유사성은 인공지능과 인간과의 관계를 어떻게 설정할 것인지에 관한 규범적 문제를 낳을 수 있다. 한편, 인공지능경망을 이용한 딥러닝 기술을 활용하는 경우 결과값의 근거를 찾기 어렵다는 이른바 ‘블랙박스(black box)’의 문제가 생긴다.

인공지능 기술이 지닌 위와 같은 기술적 특성은 책임성과 투명성의 확보가 의사결정을 자동화하는 인공지능 기술에 특유한 규제 내용의 핵심이 되어야 함을 암시한다. 다만 이러한 규제는 그 자체로 정당성이 인정된다기보다는 인간의 존엄성, 기본적 자유, 공정과 정의, 민주주의와 법치주의와 같은 다른 가치에 봉사하는 수단적인 것에 불과하다. 따라서 책임성과 투명성 확보가 아무리 중요하더라도 그것만으로 ‘강한 규제’를 할 수는 없으며, 규제의 강도는 인공지능 기술이 이용되는 특정한 맥락 하에서 보호하려는 가치가 무엇인지 그리고 위협의 정도는 어떠한지에 따라 결정되어야 한다.⁷³⁾

이처럼 인공지능 기술에 관한 규제는 유연한 규제여야 하고, 투명성과 책임성 확보를 위한 목적만으로는 강한 규제를 정당화할 수 없으며, 이용의 상황과 맥락에 따라 위협의 속성과 정도에 부합하는 규제가 마련될 필요가 있다. 그런 점에서 인공지능 기술의 이용 전반에 걸쳐 적용될 수 있는 포괄적 규제 패러다임을 마련하려는 EU 인공지능 법안을 그대로 따르는 것은 바람직하지 못하다. 비록 위 법안이 수용불가능한 위험, 고위험, 제한적 위험, 최소 또는 무위험 등으로 구분하여 위험에 비례하는 의무를 지우는 위험기반접근법을 취하고 있기는 하지만,⁷⁴⁾ 앞으로 전개될 인공지능 산업의 다양한 맥락을 이러한 틀에 의하여 적정하게 담아낼 수 있을 것인지에 대하여는 회의적일 수밖에 없다. 위 법안은 특수한 개별 영역에는 개별 법률이

70) Stuart Russell/Peter Norvig(류광 옮김), 인공지능: 현대적 접근방식, 제이펍, 2016, 2-6면.

71) 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -”, 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 135면.

72) 이상용, “인공지능과 법”, 새봄을 여는 민법학(정태윤·지원림 교수 정년기념논문집), 홍문사, 2023, 597면.

73) 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -”, 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 149면 참조.

74) 수용 불가능한 위험은 금지되고, 고위험의 경우 위험관리 시스템의 유지나 학습데이터의 대표성 보장을 비롯한 여러 의무사항을 준수해야 하며, 제한적 위험의 경우 투명성 의무가 부과되고, 최소 위험의 경우에는 자율규제가 권장될 뿐이다.

특별법으로 적용된다는 입장을 취하고 있으므로 포괄적 패러다임의 의미 역시 퇴색되고 만다.⁷⁵⁾ 개인정보의 처리와 관련된 규정들 역시 GDPR과 사실상 일치하거나⁷⁶⁾ GDPR에 의하여 보완되는⁷⁷⁾ 관계에 불과하다. 반면에 미국의 알고리즘 책임 법안은 책임성과 투명성에 초점을 맞춘 자율규제 방식을 취하고 있어 참고할 만하다. 즉, 위 법안은 연방거래위원회로 하여금 일정 기준을 넘어서는 대규모 사업자에 대하여 내부영향평가를 수행하여 보고하도록 하는 것을 핵심으로 하는 규정을 마련하도록 함에 그치고 있다($\S 3$, $\S 4$).

따라서 자동화된 결정에 이용되는 인공지능 기술에만 초점을 맞추어 포괄적인 규제를 시도하는 것은 지양해야 한다. 다만 자율규제를 원칙으로 하여 투명성에 초점을 맞춘 약한 수준의 유연한 규제는 검토해볼 만하다. 이를 통해 알고리즘 통제자를 직접 규제하지 못하는 개인정보보호법의 한계를 극복할 수도 있다. 다만 개인정보보호법에 의한 규제와 중복되어 사업자에게 불필요한 부담을 지우지 않도록 세심한 제도 설계가 이루어져야 할 것이다.

4. 행정법제

자동화된 결정이 국가에 의하여 이루어지는 경우에는 민주주의와 법치주의, 그리고 적법절차의 원리가 전면에 나서게 된다.⁷⁸⁾ 즉, 국가에 의한 자동화된 결정이 가능한지, 그에 대한 통제는 어떻게 할 것이며 절차적 정당성은 어떻게 확보할 것인가가 문제된다. 민간에서와 마찬가지로 신뢰성과 투명성, 그리고 공정성을 확보하기 위한 방안들이 해결방안의 핵심이 되겠지만,⁷⁹⁾ 국가와 시민 사이의 관계가 문제된다는 점에서 그 요구 수준은 한층 높아질 수밖에 없다.

행정기본법은 독일 연방행정절차법(BVwVfG) $\S 35a$ 의 영향을 받아 자동적 처분의 근거를 마련하는 한편 그 통제를 위하여 허용범위를 기속행위에 제한하였다. 즉 행정기본법 제20조는 인공지능 시스템을 포함하여 ‘완전히 자동화된 시스템’에 의한 행정청의 처분을 허용하되, 민주주의 및 법치주의적 고려에 따라 재량이 없는 경우에만⁸⁰⁾ 가능하도록 하였다.⁸¹⁾ 다만 자동

75) 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 65면 참조.

76) EU 인공지능 법안 제5조 제1항에서 공적인 범용 사회적 평점 시스템을 금지한 것은 당초 수집목적에 벗어난 추가 처리를 제한하고 있는 GDPR 규정과 사실상 일치한다는 견해가 있다.(주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 68면).

77) EU 인공지능 법안 제54조는 공익을 위한 인공지능 시스템을 개발하기 위한 추가 처리를 허용하고 있으나, 그 경우에도 그 밖의 처리의 적법 요건은 갖추어야 하고 자동화된 결정에 대한 거부권 등도 인정된다고 해석하는 견해가 있다(주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 84면).

78) 김도승, “인공지능 기반 자동행정과 법치주의”, 미국헌법연구 제30권 제1호, 미국헌법학회, 2019, 105-138면.

79) 미국에서 논의되는 알고리즘 자동행정결정은 이러한 이슈들을 중심으로 논의되고 있다.(김재선, “알고리즘 자동행정결정에 대한 행정법적 해석방안에 관한 연구 - 미국 행정법상 입법방안 논의를 중심으로”, 법학논총 제45권 제3호, 단국대학교 법학연구소, 2021, 83면).

적 처분의 절차적 정당성을 확보하기 위한 규정은 따로 두지 않았다. 그러나 학계의 주류적 견해는 자동적 처분도 행정행위인 이상 사전통지나 이유제시와 같은 행정절차법상의 법리들이 적용된다는 것이므로,⁸²⁾ 이를 통해 자동적 결정의 절차적 정당성이 확보되고 신뢰성과 투명성이 제고될 수 있을 것이다. 특히 사전통지 및 이유제시의무는 기능적으로 개정법상 자동화된 결정에 대한 설명요구권에 대응하는 것으로서 의미가 크다. 이와 관련하여 일부 견해는 시스템에 입력된 개별 데이터와 처분 도출 원리 등이 처분의 이유로서 제시되고 확인될 필요가 있다고 주장하기도 한다.⁸³⁾ 한편, 개인정보보호법 개정 과정에서 행정기본법 제20조에 따른 행정청의 자동적 처분과의 충돌을 우려하는 견해가 제시됨에 따라 자동적 처분은 개인정보보호법상 자동화된 결정 개념에서 제외되었다.⁸⁴⁾

헌법상 적법절차의 원리 역시 국가에 의한 자동화된 결정을 통제하고 절차적 공정성을 확보하기 위한 기제가 될 수 있다.⁸⁵⁾ 적법절차의 원리는 고지와 청문을 핵심으로 하는 절차적 적법절차와 기본권 침해의 한계를 설정하는 실체적 적법절차를 모두 포함하기 때문이다.⁸⁶⁾ 일부 견해는 사법부와 입법부에서 민간 기업의 특정 행위에 대하여도 청문 절차를 부과하는 점을 근거로 적법절차의 원리가 사적 영역으로 확대될 가능성을 제시하기도 한다.⁸⁷⁾ 그렇게 본다면 본래 자의적이고 차별적인 국가작용을 방지하기 위한 보호장치로 기능하는⁸⁸⁾ 적법절차의 원리가 사인 간의 자동화된 결정에 대한 통제의 근거, 특히 설명요구권의 근거로 작동할 여지도 있을 것이다. 그러나 사적자치의 영역은 과잉금지원칙의 각 단계가 그대로 적용되기보다는 법익균형성이라는 최적화의 원리가 중요하다는 점에서 이러한 접근방법에는 찬성하기 어렵다.⁸⁹⁾ 최근 들어 논의되고 있는 ‘기술 영역의 적법절차(technological due process)’⁹⁰⁾ 역시

80) 한편, 독일 연방행정절차법은 완전 자동화된 행정행위가 허용되려면 재량 뿐 아니라 판단여지도 없어야 한다고 정하고 있다.

81) 행정기본법 제20조는 인공지능 기술이 적용된 경우도 포함하고 있는데, 장래 등장할 학습능력이 있는 인공지능 시스템의 자율성은 그 자체로 재량의 허용이 될 수밖에 없다는 견해가 제시되고 있다(조성규, “인공지능에 기반한 자동화된 행정결정의 행정법적 쟁점”, 동북아법연구 제16권 제4호, 전북대학교 동북아법연구소, 2023, 171-172면).

82) 김대인, “행정기본법과 행정절차법의 관계에 대한 고찰”, 법제연구 제59호, 한국법제연구원, 2021, 19면; 김민호/윤금남, “자동화된 의사결정과 행정절차법에 관한 검토”, 성균관법학 제34권 제3호, 성균관대학교 법학연구원, 2022, 51-52면.

83) 권은정, “자동적 처분의 행정법적 제 문제: 처분의 형식·절차 및 구제에 관한 쟁점을 중심으로”, 법학논총 제38권 제3호, 한양대학교 법학연구소, 2021, 100면; 김민호/윤금남, “자동화된 의사결정과 행정절차법에 관한 검토”, 성균관법학 제34권 제3호, 성균관대학교 법학연구원, 2022, 54면.

84) 제402회 국회(임시회), 법제사법위원회회의록 제1호, 국회사무처, 국회사무처(2023. 1. 16.), 13면, 제403회 국회(임시회), 법제사법위원회회의록 제2호, 국회사무처(2023. 2. 16.), 20면.

85) 성윤희, “행정기본법상 자동적 처분에 관한 연구”, 이화여자대학교 대학원 법학과 박사학위 논문, 2023.

86) 김중철, “적법절차원리의 헌법화에 대한 소견 - 미국연방헌법상 적법절차원리의 기능을 중심으로 -”, 인권법평론 제20호, 전남대학교 법학연구소 공익인권법센터, 2018, 14면.

87) Cotran v. Rollings Hudig Hall Int'l, Inc., 948 P.2d 412, 422 (Cal. 1998) 참고.

88) 강승식, “절차적 적법절차의 본질에 관한 연구”, 미국헌법연구 제25권 제2호, 미국헌법학회, 2014.

데이터와 알고리즘 및 그 처리과정을 감시하기 위한 사전적인 통제를 의미할 뿐 그 법적 근거를 제시하는 것은 아니므로 크게 의미를 둘 것은 아니라고 생각된다.

5. 차별금지법제

지금까지 살펴본 개인정보보호법제, 인공지능규제법제, 행정법제는 모두 자동화된 결정과 관련된 위험을 방지하기 위한 예방적, 절차적, 수단적 성격을 지닌다. 그러나 이러한 법제도가 마련되었음에도 불구하고 실제로 위험이 발생한 경우에는 그 위험의 속성과 정도에 부합하는 사후적이고 실제적인 규제가 이루어지게 된다.⁹¹⁾ 예컨대 어떠한 자동화된 결정에 의하여 타인의 신체나 재산상 이익 또는 명예가 침해된 경우에는 형법상 범죄가 되거나 민법상 손해배상 의무가 발생할 수 있다.

그러나 가장 대중의 관심을 끄는 것은 차별에 의한 공정성 침해의 문제일 것이며, 이는 차별금지법제에 의하여 직접적으로 다루어지게 된다. 차별금지법제는 헌법상 평등의 이념을 실현하기 위하여 기존의 차별을 시정하고 차별행위를 금지하는 것을 목적으로 한다.⁹²⁾ 특히 민간 영역에서 차별금지법제는 사적자치의 원칙이 평등의 이념에 의하여 제한된다는 의미를 가지며, 따라서 헌법상 비례의 원칙에 의하여 제어된다. 차별금지법제는 자동화된 결정과 관련된 공정성의 문제를 직접적으로 정교하게 다룰 수 있지만, 그 밖의 위험을 다루기에는 부적합하다.

차별금지법제는 차별금지사유와 차별금지영역의 범위에 따라 일반적 차별금지법과 개별적 차별금지법으로 나눌 수 있다.⁹³⁾ 국가인권위원회법⁹⁴⁾이 전자에 해당한다면 「채용절차의 공정화에 관한 법률」⁹⁵⁾은 후자에 해당한다고 할 수 있다. 그러나 법리적으로 보다 중요한 것은 직접차별과 간접차별의 구분이다.⁹⁶⁾ 직접차별이란 ‘차별금지사유가 직접적인 이유가 되어 발생하는 합리적 이유가 없는 차별’을 말하고, 간접차별이란 ‘중립적인 기준을 적용하였으나 그러

89) 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 84면 참조.

90) Citron, D. Keats/Frank Pasquale, “The Scored Society : Due Process for Automated Predictions”, Washington Law Review 89:1, 2014.

91) 위험의 속성에 대하여

92) 최승철, 차별금지법의 이해, 한울아카데미, 2011.

93) 홍성수, “포괄적 차별금지법의 필요성”, 이화젠더법학 제10권 제3호, 이화여자대학교 젠더법학연구소, 2018, 6-7면.

94) 고용 등 3가지 차별금지영역과 성별, 나이 등 19가지 차별금지사유를 정하고 있다(제2조 제3호).

95) 직무 수행에 필요하지 않은 한 구직자의 신체적 조건이나 출신 지역 등을 심사자료로 삼지 않도록 하고 있다(제4조의3).

96) 「장애인차별금지 및 권리구제 등에 관한 법률」 제4조 제1항; 한국장애인개발원, 장애인차별금지법과 상충되는 고용 관련 연구, 2019, 29면.

한 중립적 기준이 특정 소수자 집단에게 불이익한 결과를 야기하는 경우'를 말한다.⁹⁷⁾ 예를 들어 「장애인차별금지 및 권리구제 등에 관한 법률」은 직접차별과 간접차별을 모두 차별행위로 보아 금지하고 있다(제4조 제1항). 다른 나라의 차별금지법제에서도 이러한 구별이 있다. 미국 민권법 제7편(Title VII of the Civil Rights Act of 1964)은 고용 영역에서의 차별행위를 의도적 차별행위와 결과적 차별행위로 구분하여 모두 금지하고 있는데, 의도적 차별행위는 직접차별에, 결과적 차별행위는 간접차별에 대응하는 것으로 이해할 수 있다.⁹⁸⁾ 특히 미국 고용평등기회위원회⁹⁹⁾의 '4/5 규칙'은 결과적 차별행위에 관한 일종의 기준을 제시하고 있는데, 이에 따르면 어떤 보호집단에 대한 합격률이 다른 집단과 비교하여 4/5 미만이 되어서는 안된다.¹⁰⁰⁾ 한편, 차별행위자는 직접차별에 대하여는 합리적 이유가 있었음을 들어 항변할 수 있고, 간접차별에 대하여는 직무관련성 또는 업무상 불가피성의 항변을 할 수 있다.¹⁰¹⁾

6. 소결

자동화된 결정 과정에서 나타날 수 있는 위험 요인들에 대응하기 위하여 개인정보보호법제, 차별금지법제, 인공지능 규제법제, 행정법제 등이 고려될 수 있다. 그러나 차별금지법제는 공정성 이외의 문제에 대응하기 어렵다는 점에서, 인공지능 규제법제는 범용성을 지닌 신산업에 대한 포괄적 규제 입법이 부당하다는 점에서, 행정법제는 공적 영역에만 적용될 수 있다는 점에서 각각 한계가 있다. 개인정보보호법제가 자동화된 결정에 대한 대응 수단으로 각광받고 있는 것은 이 때문이다.

V. 자동화된 결정에 대한 대응권

개정법 제37조의2는 개인정보의 처리를 수반하는 자동화된 결정에 있어서 정보주체의 주도하에 위험을 사전적으로 통제하기 위하여 마련된 것이다. 아래에서는 그 규율 내용을 간략히

97) 박은정 외, 차별판단기준 수립을 위한 연구(제1부), 국가인권위원회, 2021, 54면.

98) 김병필, "인공지능 공정성 심사 기준에 관한 연구", 서울대학교 박사학위논문, 2023, 75-76면. 다만 미국 민권법과 달리 우리나라나 EU에서는 차별의 의도는 요건이 아니라는 점을 유의할 필요가 있다(고학수/정해빈/박도현, "인공지능과 차별", 저스티스 제171호, 한국법학원, 2019, 219면 참조).

99) U.S. Equal Employment Opportunity Commission(EEOC)

100) 4/5 규칙에 법적 구속력이 있는 것은 아니지만, 실무상으로 채용 자동화 시스템을 구축할 경우 대체로 위 기준을 준수하려는 것으로 보인다.(Raghavan, Manish, et al., "Mitigating bias in algorithmic hiring: Evaluating claims and practices", Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020)

101) 김병필, "인공지능 공정성 심사 기준에 관한 연구", 서울대학교 박사학위논문, 2023, 84-86면.

소개하기로 한다. 본고는 큰 틀에서 개인정보보호법이 자동화된 결정에 관한 규율 체계에서 담당하는 역할을 조망하고 바람직한 운영방안을 제시하는 것을 목표로 하므로 위 조항의 상세한 해석론까지 전개하지는 않겠지만, 앞으로의 논의를 위하여 중요한 해석상의 문제점은 지적하기로 한다.

1. 거부권

개정법 제37조의2 제1항은 자동화된 결정이 정보주체의 권리 또는 의무에 중대한 영향을 미치는 경우 정보주체는 해당 개인정보처리자에 대하여 해당 결정을 거부할 수 있는 권리(이하 ‘거부권’이라고 한다)를 가진다고 규정한다. 다만 i) 정보주체의 동의를 받은 경우, ii) 법률에 특별한 규정이 있거나 법령상 의무를 준수하기 위하여 불가피한 경우, iii) 정보주체와 체결한 계약을 이행하거나 계약을 체결하는 과정에서 정보주체의 요청에 따른 조치를 이행하기 위하여 필요한 경우에는 예외적으로 거부권이 인정되지 않는다(제1항 단서). 이처럼 거부권이 인정되지 않는 경우에 정보주체를 보호하기 위한 조치에 관하여는 따로 규정이 없지만, 뒤에서 보는 설명요구권이 예외 사유 없이 인정되는 결과 거부권이 인정되지 않는 경우에도 정보주체를 보호하는 역할을 맡게 된다.

이러한 규율 방식은 GDPR과 유사하면서도 다소 차이가 있다. 즉, GDPR 제22조는 제1항에서 자동화된 처리에만 기초한 결정의 적용을 받지 않을 정보주체의 권리를, 제2항에서 예외적으로 그러한 권리가 인정되지 않는 경우를, 그리고 제3항에서 일부 예외 사유의 경우 정보처리자(controller)의 보호조치 의무를 규정하는 구조를 취하고 있다. 이러한 난해한 구조는 위 규정의 해석을 매우 어렵게 만드는데, 일부 견해는 위 조항이 단지 권리의 창설에 그치는 것이 아니라 자동화된 결정의 원칙적 금지를 규정한 것으로 본다.¹⁰²⁾ 그러나 단순한 규정 구조를 갖고 있는 개정법 제37조의2는 정보주체의 권리를 창설한 것으로 좀 더 명확하게 해석될 수 있다.¹⁰³⁾

개정법 제37조의2는 GDPR을 참고하면서도 보다 명확하고 균형있는 규율을 꾀한 것으로 생각된다. 그럼에도 불구하고 불명확한 부분이 없는 것은 아니며, 특히 거부권 행사의 효과가 그러하다. 즉, 동조 제3항은 “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”라고 되어 있어서 문언상으로는 거부권 행사의 효

102) Article 29 Data Protection Working Party, Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679, p. 8.

103) 박노형/김효관, “자동화된 결정에 관한 개인정보보호법 정부 개정안 신설 규정의 문제점 - EU GDPR과의 비교 분석”, 사법 제62호, 사법발전재단, 2022, 372, 376면 참조.

과와 설명요구권 행사의 효과가 모두 자동화된 결정의 비적용, 인적 개입에 의한 재처리, 설명 등으로 동일한 것처럼 보인다. 거부권과 설명요구권의 취지와 요건이 서로 다른만큼 이러한 해석은 타당하지 않으나, 객관적으로 타당한 해석이 도출되기는 쉽지 않아 보인다.¹⁰⁴⁾ 앞으로 치밀한 해석론이 이루어질 필요가 있다고 생각된다.

2. 설명요구권

개정법 제37조의2 제2항은 개인정보처리자가 자동화된 결정을 한 경우 정보주체는 그 결정에 대하여 설명 등을 요구할 수 있다고 규정한다. 이러한 설명요구가 있는 경우 개인정보처리자는 원칙적으로 설명을 제공해야 하며 거부권의 경우와 달리 예외 사유가 따로 규정되어 있지 않다. 다만 정당한 사유가 있는 경우에는 설명을 제공하지 않아도 되는데, 이는 거부권의 경우에도 마찬가지이다(동조 제3항).

이러한 규율 역시 GDPR과 유사하면서도 적지 않은 차이를 보인다. 즉, GDPR의 경우 자동화된 결정에 관한 규정인 제22조에는 설명의무가 명시되어 있지 않고 단지 법적 구속력이 없는 전문 제71항에 서술되어 있을 뿐이기 때문에 설명요구권이 법적으로 인정되는지 여부와 만약 인정된다면 그 근거가 되는 조항이 무엇인지에 관하여 다툼이 있다.¹⁰⁵⁾ 다만 네덜란드의 암스테르담 항소법원은 2023. 4. 4. 정보주체의 열람권에 관한 GDPR 제15조 제1항 (h)에 근거하여 설명요구권을 주장한 원고의 청구를 인용한 바 있다.¹⁰⁶⁾ 개정법 제37조의2는 설명요구권을 자동화된 결정에 대한 대응권의 하나로서 명시적으로 규정함으로써 명확한 규율을 꾀하였다.

그러나 개정법에 불완전한 부분이 없는 것은 아니다. GDPR에서는 거부권의 경우와 마찬가지로 자동화된 결정이 본인에 관한 법적 효과를 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 경우에만 설명요구권을 인정한다[§15(1)(h), §22(1),(4)].¹⁰⁷⁾ 또한 설명의 대상을

104) 참고로 2023. 11. 23. 입법예고된 개인정보보호법 시행령 2차 개정안(이하 '시행령 개정안'이라고 한다) 제44조의 3 은 이러한 문제를 해결하기 위하여 i) 거부권 행사시 자동화된 결정의 비적용 및 정보주체의 요구가 있는 경우 인적 개입에 의한 재처리 의무(제1항), ii) 설명요구권 행사시 설명의무(제2항)를 개인정보처리자가 부담하는 것으로 규정하였다(<https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9553#LINK>). 그러나 이 문제가 법 개정이 아닌 시행령의 제정에 의하여 해결될 수 있는지는 의문이다.

105) 정보주체의 정보요구권 및 열람권에 관한 제13조 내지 제15조에 기하여 설명요구권이 인정된다는 견해로는 Goodman, Bryce/Seth Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation", " AI Magazine 38.3, 2017; 법적 권리로서의 설명요구권은 GDPR에서 인정되지 않는다는 견해로는 Wachter, Sandra, et al., "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", International Data Privacy Law 7.2, 2017.

106) Amsterdam Court of Appeal 200.295.742/01. <https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:GHAMS:2023:793>

107) Amsterdam Court of Appeal 200.295.742/01. <https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:GHAMS:2023:793>

i) 자동화된 결정의 존부, ii) 수반된 논리에 관한 유의미한 정보, iii) 정보주체에게 있어서 그러한 처리의 중요성과 예상되는 결과 등으로 명확히 하고 있다[§15(1)(h)]. 그러나 개정법 제37조의2 제2항에는 위와 같은 내용들이 담겨 있지 않다.¹⁰⁸⁾ 이는 설명요구권의 남용을 초래하는 한편 자동화된 결정에 대한 정보주체의 대응을 위한 정보의 제공이라는 취지를 살리지 못한 채 예측가능성의 부족으로 수범자의 규제 준수를 어렵게 하는 결과로 이어지지 않을까 우려가 된다. 그 밖에 위 조항은 '설명 등'을 요구할 수 있다고 규정하고 있으나 설명 이외에 요구의 대상이 될 수 있는 것이 무엇인지 문언상으로는 전혀 알 수 없다는 점도 문제이다.¹⁰⁹⁾ 장차 해석론의 발전과 궁극적으로는 입법적 개선에 의하여 해결되어야 할 부분이다.

VI. 기술적 한계

1. 기술적 한계가 의무에 미치는 영향

자동화된 결정에 대한 정보주체의 대응권과 개인정보처리자의 보호조치 의무의 범위를 함에 있어서는 이익형량 뿐 아니라 기술적 한계를 고려할 필요가 있다. 이는 주로 설명요구권으로 대변되는 투명성의 요청과 관련하여 문제가 된다. 앞서 보았던 것처럼 설명요구권의 행사가 있는 경우 개인정보처리자는 정보주체가 자동화된 결정에 어떻게 대응해야 할지 결정할 수 있을 정도로 그에 수반된 논리에 관한 유의미한 정보를 제공하여야 한다. 구체적인 범위와 내용은 상황과 맥락에 따라 달라지겠지만, 엄격히 본다면 문제된 맥락에서 요구되는 적법성과 타당성을 갖추었는지 여부를 판단할 수 있을 정도의 설명이 이루어져야 한다고 해석될 여지도 있다.¹¹⁰⁾

그러나 과연 그러한 정도의 설명이 기술적으로 가능할 것인지는 의문스럽다. 만약 기술적으로 불가능하다면 이를 하지 못한 것을 두고 의무위반이라고 하거나 위법하다고 말할 수는 없

108) 참고로 시행령 개정안 제44조의3 제2항 단서는 자동화된 결정이 정보주체의 권리 또는 의무에 중대한 영향을 미치는 경우가 아닌 때에는 공개의무의 대상인 정보를 고지함으로써 설명에 갈음할 수 있도록 규정하고 있다. 내용은 수궁할 만하지만 개정법이 문언상 이와 같이 해석될 수 있는지는 의문이다. 한편 동조 제2항은 설명의 대상에 포함되어야 할 사항을 열거하고 있지만, 정보주체에게 필요한 정보(예컨대 자동화된 결정에 수반된 논리)는 누락되어 있고 반대로 개인정보처리자에게 부담만을 안기는 불필요한 정보(예컨대 처리 과정에 관한 정보)가 포함되어 있어 설명요구권의 취지에 부합하지 않으므로 재고를 요한다.

109) 참고로 시행령 개정안 제44조의3 제3항은 이의제기권을 규정하고 있다. 이는 개정 개인정보보호법 제37조의2 제2항이 정보주체에게 이의제기권을 부여하였음을 전제로 한 것이지만, 위 조항이 그렇게 해석될 수 있는지는 의문이다.

110) ISO는 인공지능에 대한 설명가능성은 인공지능의 활용이 규범적으로 정당화될 수 있다는 사실에 대한 설명까지로 확장될 필요가 있다고 밝힌바 있다. ISO/IEC, Artificial Intelligence (AI) - Assessment of the robustness of neural networks, Technical Report(TR 24029), 2021.

다. 설명이 불가능하다고 하여 설명요구권이 존재하지 않는 것은 아니라는 지적¹¹¹⁾은 크게 도움이 되지 않는다. 지금 문제되는 것은 설명요구권의 존재 여부가 아니라 제재의 근거가 되는 의무위반이나 위법성의 유무이기 때문이다. 이 때 기술적 가능성의 유무는 순수하게 기술적 측면에 의해서만 판단할 것은 아니고 사회관념에 비추어 판단해야 한다. 예컨대 어떻게든 설명가능하게 알고리즘을 구현할 수 있다고 하더라도 그로 인해 성능이 현저히 저하하거나 비용이 비합리적으로 증가하여 당초 자동화된 결정 시스템을 채택하려던 목적을 이룰 수 없게 된다면 기술적으로 가능하지 않은 것이라고 보아야 할 것이다.

이처럼 개인정보처리자의 의무의 범위를 정함에 있어서는 기술적 한계를 고려할 필요가 있다. 그 밖에도 수범자의 규제 준수에 대한 기대가능성에 영향을 미칠 수 있는 요소들은 폭넓게 고려될 필요가 있다. 예를 들어 다른 기업이 개발한 인공지능 모델을 구매하여 활용하는 중소기업의 경우 정보주체의 설명요구에 적절히 대응하기 어려울 수 있다. 이 경우 중소기업이 인공지능 개발자로부터 필요한 정보와 자원을 제공받을 수 있도록 계도하는 등의 조치가 필요할 것이다. 한편, 기술적 가능성과 관련하여 좀 더 자세히 살펴볼 세부 문제들이 있다. 이른바 ‘블랙박스(black box)’의 문제, 상충관계의 문제, 그리고 공정성 지표의 문제가 그것이다.

2. 블랙박스의 문제

오늘날 자동화된 결정에는 인공지능 기술이 폭넓게 활용되고 있다. 인공지능 기술은 크게 기호와 논리를 중시하는 기호주의적 접근법(symbolism)과 인간의 뇌를 모방하는 연결주의적 접근법(connectionism)으로 나뉜다.¹¹²⁾ 1956년 다트머스 회의에서 인공지능이라는 용어가 탄생한 이래 기호주의가 압도적 영향을 미쳐 왔지만, 2010년대 이후에는 신경망(neural network)을 기초로 한 딥러닝(deep learning), 즉 연결주의적 접근법이 인공지능의 부활을 주도하면서 판단과 의사결정과 관련된 대부분의 영역에서 기호주의적 접근법을 크게 능가하는 성능을 보여주었다. 그런데 딥러닝 기술은 의사결정 나무(decision tree) 등 기호주의적 접근법과 달리 결과값의 근거를 찾기 어렵다¹¹³⁾. 이처럼 인공지능에 의한 판단이나 결정의 근거나 과정에 대

111) 박상돈, “헌법상 자동의사결정 알고리즘 설명요구권에 관한 개괄적 고찰”, 헌법학연구 제23권 제3호, 한국헌법학회, 2017, 197면.

112) Stuart Russell/Peter Norvig(류광 옮김), 인공지능 : 현대적 접근방식, 제이펍, 2016, 20-36면.

113) 인공지능 기술을 분류모델과 생성모델로 나누고, 분류모델을 다시 인지모델과 판별모델로 나누면서, 판별모델의 경우 인지모델과 달리 주로 정형데이터를 입력받아 SVM(Support Vector Machine) 등의 고전적 기계학습 기술을 이용하므로 투명성과 해석가능성의 수준이 높다고 암시하는 견해도 있다(김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 28-30면; Schwartz-Ziv, Ravid/Amitai Armon, “Tabular data: Deep learning is not all you need”, Information Fusion 81, 2022, pp. 84-90). SVM과 같은 고전적 기계학습 기술이 신경망에 의존하지 않고 기호주의적 접근법을 취하는 것은 사실이다. 그러나 이들 역시 본질이 ‘자동 분류기(automatic classifier)’라는 점에서 딥러닝 기술과 마찬가지로 결과값의 근거를 찾기 어렵다는 ‘블랙박스’ 문제를 공유한다고 볼 수 있다(이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조,

한 적절한 설명이 없는 상태를 흔히 ‘블랙박스(black box)’라고 표현한다.

블랙박스적 특성은 알고리즘의 불투명성의 원인 가운데 가장 해결하기 어려운 것이다. 일반적으로 알고리즘에 불투명성이 나타나는 이유는 i) 의도적인 은폐, ii) 전문성이 요구되는 복잡성, iii) 기계학습에서의 수학적 최적화 방식과 인간의 추론 및 해석 방식 사이의 불일치 등으로 구분되어 설명된다.¹¹⁴⁾ GDPR 상의 설명요구권을 처음으로 주창했던 Goodman & Flaxman은 이 중 i)의 경우는 정보제공의무를 규정한 GDPR 제13조에 의하여, ii)의 경우는 간접하고 이해하기 쉬우며 손쉽게 접근할 수 있는 방식으로 정보제공이 이루어져야 한다고 규정하는 제12조에 의하여 해결될 수 있지만, iii)의 경우는 기계학습의 특성상 매우 어려운 문제가 된다고 지적했다.¹¹⁵⁾ 통상적인 기계학습 알고리즘은 정확한 예측을 위한 상관관계를 밝혀내는 것에 기반할 뿐 통계적 유의성을 벗어난 인과적 추론이나 설명은 하지 못한다.¹¹⁶⁾ 상관관계는 장래에 어떤 일이 생길 확률은 보여주지만 왜 그런지는 알려주지 않는다. 심지어 기계학습 알고리즘의 설계자조차도 상관관계 이상의 것을 알아낼 방법이 없다. 이러한 문제점을 극복하고 인공지능 알고리즘의 설명가능성을 확보하려는 노력이 그동안 많이 이루어져 왔다. 미국 고등연구계획국(Defense Advanced Research Projects Agency, DARPA)이 2016년부터 4년에 걸쳐 ‘설명가능한 인공지능(XAI 프로그램)’을 추진했던 것이 바로 단적인 사례이다.¹¹⁷⁾ 그럼에도 불구하고 그 성과는 지지부진한 편이다.¹¹⁸⁾

이처럼 블랙박스 문제를 직접적으로 다루는 것이 생각보다 어려운 것으로 나타남에 따라 이를 우회적으로 해결하려는 시도가 이루어지기도 했다.¹¹⁹⁾ 이에 따르면 설명이란 정보주체의 행동을 돕기 위한 수단으로서 그 목표는 i) 어떠한 결정이 이루어진 이유를 알리고 이해하도록 도우며, ii) 원하지 않는 결과가 나왔다면 결정에 이의를 제기할 근거를 제공하고, iii) 현재의

내용 -, 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 135면).

114) Burrell, J., “How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms”, Big Data and Society 3.1, 2016.

115) Goodman, Bryce/Seth Flaxman, “European Union regulations on algorithmic decision-making and a “right to explanation”, AI Magazine 38.3, 2017, pp. 55-56.

116) Goodman, Bryce/Seth Flaxman, “European Union regulations on algorithmic decision-making and a “right to explanation”, AI Magazine 38.3, 2017, p. 55.

117) 신경망 구성 요소의 가중치(weight)나 현저성(saliency)에 주목하는 비교적 단순한 것부터 의사결정 기능과 설명 기능을 전담하는 알고리즘을 따로 두는 보다 혁신적인 것까지 다양한 방안들이 제시되었다. Explainable Artificial Intelligence (XAI) Program, 2016. Available at: <https://www.darpa.mil/program/explainable-artificial-intelligence> (2023. 11. 27. 확인); Gunning, David/David Aha, “DARPA’s explainable artificial intelligence program”, AI Magazine, Association for the Advancement of Artificial Intelligence, 40.2, 2019.

118) 사건으로는 정형데이터를 기반으로 하는 분류 모델의 경우에 어떠한 입력 요소가 결과값에 어느 정도의 영향을 미치는지 정도는 실험적 방법으로 보여줄 수 있다고 생각되지만, 그 경우에도 그것이 인과관계를 의미한다고 볼 수는 없을 것이다.

119) Wachter, Sandra, et al., “Counterfactual explanations without opening the black box: Automated decisions and the GDPR”, Harv. JL & Tech. 31, 2017.

의사결정 모델을 기반으로 장치 원하는 결과를 얻기 위해 무엇을 바꾸어야 하는지 이해하도록 하는 것이다. 그런데 이들 목표는 시스템의 내부 논리를 설명하지 않더라도, ‘무조건적인 반사실적 설명(unconditional counterfactual explanations)’의 제공, 즉 바람직한 결과를 얻기 위해 이루어질 수 있는 최소한의 변화를 기술하는 것에 의하여 달성될 수 있다는 것이다. 이러한 시도는 매우 흥미롭고 설득력이 있다고 생각되지만, 인공지능 알고리즘에 반사실적 설명의 기능을 추가하는 것 역시 어려운 일이라는 한계를 안고 있다.

3. 상충관계의 문제

다음으로 볼 것은 알고리즘의 성능과 투명성 사이에 존재하는 상충관계의 문제이다. 즉, 알고리즘의 표현 능력(representational capacity)과 해석 가능성(interpretability) 사이에는 뚜렷한 상충관계가 존재한다. 예를 들어 선형 회귀 모델의 경우 해석 가능성은 높지만 단순한 관계만을 표현할 수 있다. 서포트 벡터 머신(support vector machine; SVM)은 풍부한 함수를 표현하지만 해석이 쉽지 않다. 신경망을 이용한 딥러닝 모델의 경우 놀라울 정도로 복잡한 관계를 표현할 수 있지만, 복잡한 구조를 지닌 다층 신경망에 학습된 가중치를 설명한다는 것은 난망한 일이다.¹²⁰⁾ 인공지능 모델의 유연성과 투명성 사이에도 상충관계가 존재한다. 예를 들어 분류 모델이 훈련데이터에 과적합(overfitting)되는 것을 방지하기 위해 사용되는 정규화(regularization)나 무작위화(randomization) 등의 기법은 유연성을 확보해주지만 동시에 어느 정도 투명성을 희생시키게 된다.¹²¹⁾ 알고리즘의 정확성과 공정성 사이에 상충관계가 존재한다는 점도 지적되고 있다.¹²²⁾ 다만 설명가능성을 알고리즘 성능의 한 지표로 본다면 위와 같은 상충관계는 다르게 해석될 여지가 있다.¹²³⁾

한편, 투명성은 기회주의적 행동을 통해 알고리즘의 실효성을 약화시키기도 한다. 예를 들어 지원자의 업무능력을 입사시험을 통해 평가하려고 하는 경우 그 평가기준이 되는 시험점수는 실제로 평가하고자 하는 업무능력의 매개변수에 불과하다.¹²⁴⁾ 만약 평가기준이 공개된다면

120) Goodman, Bryce/Seth Flaxman, “European Union regulations on algorithmic decision-making and a “right to explanation”,” *AI Magazine* 38.3, 2017, pp. 55-56.

121) 김도훈, “알고리즘 책임성 논의와 알고리즘에 대한 이해”, *주간기술동향 제16권 제1호*, 정보통신기술진흥센터, 2018, 17-22면.

122) 오요한/홍성욱, “인공지능 알고리즘은 사람을 차별하는가?” *과학기술학연구 제18권 제3권*, 한국과학기술학회, 2018, 166면; Selbst, Andrew, “Disparate Impact in Big Data Policing”, *Georgia Law Review*, 52, 2018, p. 137.

123) 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -”, *경제규제와 법 제13권 제2호*, 서울대학교 공익산업법센터, 2020, 152-153면; 공정성 지표 역시 또 하나의 성능 지표로 고려될 수 있다는 견해로는 김병필, “인공지능 공정성 심사 기준에 관한 연구”, *서울대학교 박사학위논문*, 2023, 215면.

124) 이처럼 목표변수의 정답이 명확히 정의되기 어렵고 확률적인 성격을 갖는 모델을 판별 모델이라고 하고, 이와 달리 목표변수의 정답이 명확히 정의될 수 있고 고정된 성격을 갖는 모델을 인지모델이라고 한다.

평가대상은 기준에 특화된 적용된 행동을 함으로써 실제보다 능력을 부풀려 보이기 위한 노력을 할 수 있고, 이는 사실상 알고리즘의 정확도를 떨어뜨리는 결과로 이어진다.¹²⁵⁾

4. 공정성 지표의 문제

가. 여러 가지 공정성 지표

현실적으로 자동화된 결정의 대상이 된 정보주체가 이에 어떻게 대응할 것인지 판단하기 위해 가장 알고 싶은 것은 그 결정이 과연 공정하게 이루어졌는가일 것이다. 자동화된 결정이 초래하는 위험 가운데 가장 많은 관심을 받고 있는 것 역시 공정성의 문제이다. 그렇다면 알고리즘이 공정한지 여부를 판단할 수 있는 정보 역시 설명의 대상에 포함되어야 할까? 결론부터 말하자면 반드시 그렇다고는 할 수 없다. 공정성을 기술적으로 표현하는 것에는 분명한 한계와 모순이 있기 때문이다.

위 문제에 답하는 한 가지 방법은 데이터 학습을 통해 알고리즘이 만들어지는 과정에 주목하는 것이다. 앞에서 살펴본 차별이나 편향의 발생 경로에 관한 내용이 바로 그것이다. 그러나 보다 나은 방법은 근본적으로 공정성이 의미하는 바에 초점을 맞추는 것이다. 공학자들은 공정성의 개념을 계량화하기 위하여 수많은 공정성 지표를 창안하였다.¹²⁶⁾ 하지만 여기서는 영국 개인정보보호위원회(ICO)의 제안에 따라 i) 결과 동등성(outcome parity), ii) 예측도 동등성(predictive value parity) iii) 오류율 동등성(error parity) 등 세 가지를 중점적으로 살펴보기로 한다.¹²⁷⁾

결과 동등성 기준은 가장 직관적이고 간명한 것으로서 각 집단의 결과가 동등한지 여부를 평가한다. 오류율 동등성 기준과 예측도 동등성 기준은 인공지능의 예측과 평가데이터의 ‘정답(ground truth)’ 사이의 오차가 각 집단 사이에 동등한지 여부를 평가한다. 다만, 예측도 동등성은 양성예측도(positive predictive value)와 음성예측도(negative predictive value)의 동등성을 의미하는 반면, 오류율 동등성은 위양성율(false positive rate)과 위음성율(false negative

판별 모델과 인지 모델은 모두 어떤 데이터가 입력변수로 주어지면 그에 대한 목표변수를 출력하는 대응관계를 구현한다는 점에서 분류 모델에 속한다. 반대로 어떠한 목표변수에 대응하는 입력변수의 확률분포를 모형을 구축하고 그로부터 확률이 높은 값을 표본추출하는 모델을 생성 모델이라고 한다.(김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 22-37면)

125) 고학수/정해빈/박도현, “인공지능과 차별”, 저스티스 제171호, 한국법학원, 2019, 248면.

126) Narayanan, Arvind, “21 fairness definition and their politics”, ACM Fairness, Accountability and Transparency Conference, 2018; 인공지능의 공정성 지표를 반사실적 공정성(counterfactual fairness)와 통계적 공정성으로 구분하기도 하는데, 이는 각각 개인 공정성과 집단 공정성의 구분에 대응한다. 김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 55면.

127) 오요한/홍성욱, “인공지능 알고리즘은 사람을 차별하는가?” 과학기술학연구 제18권 제3권, 한국과학기술학회, 2018, 172면; 김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 156면.

rate)의 동등성을 의미한다. 양자의 핵심적 차이는 분모에 있는데, 전자는 예측된 수치를 분모로 삼는 반면, 후자는 실제 수치를 분모로 삼는다. 예를 들어 A, B 집단의 구성원에 대하여 재범 위험성을 평가하는 인공지능이 있다고 해 보자. 고위험자로 선정된 사람의 비율이 두 집단에서 모두 같다면 결과 동등성 기준이 충족된 것이다. 고위험자로 평가된 사람 중에서 실제로 고위험인 사람이 차지하는 비율(양성예측도)이 두 집단에서 모두 같다면 예측도 동등성 기준¹²⁸⁾이 충족된 것이다. 한편 고위험자로 평가되었으나 실제로는 저위험인 사람의 수가 전체 저위험자 수에서 차지하는 비율(위양성율)¹²⁹⁾이 두 집단에서 모두 같다면 오류율 동등성의 기준이 충족된 것이다.

나. 공정성 지표와 차별금지법제

이러한 공정성 지표들은 차별금지법제의 틀에 맞추어 해석될 수 있다. 우선 직접차별을 막기 위해서는 차별금지사유를 인공지능의 입력 변수로 활용하지 않는 방법을 검토해볼 수 있다. 그러나 이런 방법이 항상 효과가 있는 것은 아니다. 외견상 중립적인 다른 속성(대용변수, proxy variable)을 통해 차별금지사유에 해당하는 보호속성이 추론되는 경우가 많기 때문이다.¹³⁰⁾

한편 간접차별을 식별하기 위해서는 예측도 동등성 기준이나 결과 동등성 기준을 활용할 수 있을 것이다. 만약 두 집단 간에 결과 동등성 기준이 충족되었다면 평가 대상자가 어느 집단에 속하는지와 무관하게 평가 대상자에게 결과에 대한 균등한 기대치가 제공되었다는 뜻이다. 이 경우 ‘같은 것은 같게, 다른 것은 다르게’ 대우받은 사람의 비율은 두 집단이 달라지지만, 그것은 역사적 차별과 편향이 두 집단에 다르게 자리잡았기 때문이라고 보는 것이 간접차별의 논리이다. 만약 두 집단 간에 예측도 동등성 기준이 충족되었다면 평가 대상자가 어느 집단에 속하는지와 무관하게 평가자의 목적에 부합하는 방식으로 평가가 이루어졌다는 이야기가 된다. 이 경우 ‘같은 것은 같게, 다른 것은 다르게’ 대우받은 평가 대상자들의 비율은 두 집단에 있어서 동일하다. 평가자의 입장에서 보면 예측도 동등성 기준은 그 인공지능 알고리즘을 얼마나

128) 이를 ‘집단 캘리브레이션(calibration)’ 기준 또는 ‘캘리브레이션’이라고 표현하기도 한다. 예측도 동등성 기준은 학습데이터 상의 편향을 제거하기 위한 명시적 개입을 하지 않더라도 머신러닝 학습 과정에서 기본적으로 충족되는 경향을 보인다.(Barocas, Solon, et al., *Fairness and Machine Learning: Limitations and Opportunities*, online textbook, 2022, <https://fairmlbook.org/> 2023. 11. 27. 확인)

129) 이와 달리 전체 고위험자 중에서 고위험자로 평가된 사람의 비율을 진양성율이라고 한다. 진양성율은 1에서 위음성율을 뺀 값과 같다. 한편, 양성 예측도를 정밀도(precision)라고 하고 진양성율을 재현율(recall)이라고 하여 인공지능의 성능을 나타내는 지표로 사용하기도 한다. 특히, 예측도는 인공지능의 예측 값을 이용하고자 하는 이들이 그 예측 결과를 얼마나 신뢰하고 사용할 수 있는지를 보여주는 지표라고 할 수 있다.

130) 전이 대용변수, 축소 대용변수, 확장 대용변수의 경우로 구분하여 문제점을 검토한 것으로는 Cofone, Ignacio, “Antidiscriminatory privacy”, *SMU Law Review* 72, 2019. 다만 대용변수에 의한 차별의 경우 간접차별의 범주로 보아 판단해도 좋을 것이다.(김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 79-82면).

신뢰하고 사용할 수 있는지를 보여주는 지표이다.

한편, 만약 두 집단 간에 오류율 동등성 기준이 충족되었다면 평가 대상자 개인이 어느 집단에 속하는지와 무관하게 부당한 판단을 받을 확률이 같다는 의미이다. 오류율 동등성 기준은 ‘차등적 부당대우’라는 새로운 차별 판단의 방식을 제시하는 것으로서 기존의 차별금지법제에 반영되어 있지 않은 개념이다. 대체로 보아 이는 평가 대상자의 입장에서 기회의 균등을 의미한다고 볼 수 있다.¹³¹⁾

다. 공정성 지표의 한계

이렇게 보면 인공지능 알고리즘의 공정성 여부를 판단할 수 있는 적절한 방법론과 지표들이 존재하고, 따라서 설명의 대상이 되어야 할 것처럼 보이기도 한다. 그러나 실제로는 그렇지 않다. 그 이유로는 i) 공정성 지표들이 지닌 한계와 ii) 공정성 지표들 사이의 충돌, iii) 직접차별 기준과 간접차별 기준 사이의 충돌 등을 들 수 있다.

우선 i) 공정성 지표들이 지닌 한계에 관하여 보자. 공정성 지표는 계량화를 위해 공정성을 조작적으로 정의한 것으로서 본래 공정성이 지닌 다양하고 풍부한 개념을 모두 담아낼 수 없다. 비교 대상이나 절대적 평가기준이 없고, 어떤 지표 값이 나오게 된 근본적 원인을 알 수 없으며, 그 지표를 얼마나 신뢰할 수 있는지도 알기 어렵다.¹³²⁾ 특히 지표를 산출하기 위해 민감 정보에 접근해야 하는 경우 개인정보보호법제와 충돌한다는 문제도 있다. 각각의 공정성 지표들 역시 한계를 가지고 있다. 예를 들어 결과 동등성 기준의 경우 이를 지키기 위해서는 정확도의 하락을 감수해야 한다.¹³³⁾ 어떤 집단에서 두 변수 간의 상관관계가 발견되더라도 해당 집단을 작은 하위집단으로 구분할 경우 그 상관관계가 사라지거나 역전되기도 한다(Simpson paradox).¹³⁴⁾ 예측도 동등성 기준이나 오류율 동등성 기준의 경우 정확한 평가 데이터를 구축할 필요가 있다는 점도 실무적으로 큰 어려움이 된다. 그리고 예측도 동등성 기준의 경우 기존 사회에 존재하는 불평등을 그대로 반영할 뿐이라는 비판이,¹³⁵⁾ 오류율 동등성 기준의 경우 이를 위반한 경우의 법적 함의를 도출하기 어렵다는 비판이¹³⁶⁾ 제기되기도 한다.

131) 김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 174-175면.

132) 김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023, 227-233면.

133) Corbett-Davies, Sam, et al., “Algorithmic decision making and the cost of fairness”, Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017은 이를 공정성 비용이라고 표현하면서, 역사적 차별을 시정하기 위한 비용을 해당 사업자에게 전가하는 것이 정당화될 수 있는지를 문제삼는다.

134) 1973년 버클리 대학교 대학원 입시에서 남성은 44%가 합격한 반면 여성은 단지 35%가 합격하였지만, 학과별로 구분하여 볼 경우 대부분의 학과에서 여성이 남성보다 합격률이 더 높은 것으로 드러난 사례가 대표적이다 (Bickel, Peter, et al., “Sex bias in graduate admissions: Data from Berkeley”, Science 187, 1975).

135) Wachter, Sandra, et al., “Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law”, West Virginia Law Review 123.3, 2021.

더 큰 문제는 ii) 공정성 지표들 사이에 모순과 충돌이 존재한다는 것이다.¹³⁷⁾ 즉, 결과 동등성, 예측도 동등성, 오류율 동등성은 상호 간에 동시에 충족하는 것이 불가능하다는 점이 증명되어 있다.¹³⁸⁾ 예를 들어 컴파스 사건에서 프로퍼블리카는 오류율 동등성이 갖춰져 있지 않음을 문제삼았고, 노스포인트사는 예측도 동등성이 갖춰져 있다면서 이를 반박했다. 둘 사이에 차이가 생기는 이유는 강력 범죄의 대다수가 젊은 남성에게 의하여 저질러진다는 기저율(base rate)의 차이 때문이다. 두 기준이 양립할 수 없는 것이라면 단순히 공정성 지표만으로는 공정성에 관한 판단을 할 수 없으며 또 다른 관점에서 추가적인 해석이 필요해진다.

가장 어려운 문제는 iii) 직접차별 기준과 간접차별 기준이 충돌할 수 있다는 것이다. 즉, 인공지능에 의해 초래될 수 있는 간접차별을 막기 위한 노력이 오히려 직접차별에 해당할 수 있다. Ricci v. DeStefano 판결¹³⁹⁾은 이 문제를 극명하게 보여준다. 위 사건에서 뉴헤이븐시 정부는 알고리즘을 이용한 소방관 승진시험 합격률이 인종에 따라 크게 다르게 나오자 결과적 차별행위를 피하기 위해 시험결과를 폐기하였다. 연방대법원은 시 정부의 행위가 오히려 의도적 차별행위에 해당한다고 판결하였다. 이는 미국 차별금지법제의 해석에 있어서 반분류(anti-classification) 원칙과 반종속(anti-subordination) 원칙 간의 충돌을 보여주는 사례라고 할 수 있다.¹⁴⁰⁾ 즉, 결과 동등성 기준은 그 자체로 공정성을 보장하지 못할 뿐 아니라 사안의 맥락에 따라서는 오히려 공정성을 침해하는 것이 될 수도 있는 것이다.

VII. 바람직한 운영방안

1. 절차적·수단적 규제로서의 특성 고려

자동화된 결정과 관련된 위험은 인간의 존엄성, 프라이버시와 자유, 민주주의, 공정성 등 여

136) 오류율 동등성 지표만으로는 알고리즘이 공정하거나 불공정한지 여부를 결정하지 않는다고 본 견해로는 Hellman, Deborah, "Measuring algorithmic fairness", Virginia Law Review 106.4, 2020.

137) 오요한/홍성욱, "인공지능 알고리즘은 사람을 차별하는가?" 과학기술학연구 제18권 제3권, 한국과학기술학회, 2018, 201면.

138) Kleinberg, Jon, et al., "Inherent trade-offs in the fair determination of risk scores", arXiv preprint arXiv:1609.05807, 2016; Barocas, Solon, et al., Fairness and Machine Learning: Limitations and Opportunities, online textbook, 2022.

139) Ricci v. DeStefano, 557 U.S. 557 (2009). 위 판결의 해석에 관한 논쟁으로는 Barocas, Solon/Andrew Selbst, "Big data's disparate impact", California law Review 104.3, 2016 및 Kim, Pauline, "Auditing algorithms for discrimination", University of Pennsylvania Law Review 166, 2017 참조.

140) 미국 수정헌법 제14조의 평등 조항의 해석과 관련하여, 반분류 원칙은 형식적 평등을 강조하는 반면, 반종속 원칙은 실질적 평등의 개념을 전제로 역사적으로 종속되어 온 집단에 대한 적극적 평등실현조치를 옹호한다 (Fiss, Owen, "Groups and the Equal Protection Clause", Philosophy & Public Affairs 107, 1976).

러 측면에서 나타날 수 있다. 이를 해결하기 위해 다양한 법제가 활용될 수 있지만, 그 중에서도 개인정보보호법제는 위와 같은 위험들을 사전적으로 예방하기 위한 절차적·수단적 통제 수단으로 기능한다. 새로 도입된 자동화된 결정에 대한 대응권의 경우에도 그 핵심은 설명요구권과 공개의무에 의하여 투명성을 확보하는데 있다고 할 것이다.

다만 거부권의 경우에는 인간의 존엄성을 견지하기 위한 내용적 통제로 볼 여지가 있다. 그러나 거부권이 행사되더라도 개인정보처리자는 인적 개입을 전제로 동일한 내용의 결정을 할 수 있기 때문에 실제로 있어서는 큰 의미를 갖지 않는다. 정확성의 측면에서 인적 개입이 실효성이 없다는 보고도 있다. 즉, 일반인을 상대로 한 피고인의 재범위험 예측 실험에서, 참가자들이 알고리즘의 권고를 참고하여 의사결정을 내리는 경우, 알고리즘을 사용하지 않고 결정을 내리는 경우보다는 더 정확하지만, 알고리즘보다는 덜 정확한 판단을 하는 것으로 드러났다.¹⁴¹⁾ 「신용정보의 이용 및 보호에 관한 법률」상 신용정보주체는 자동화평가 결과에 대하여 정확성 확보를 위한 의견진술 또는 재처리를 요구할 수 있을 뿐 인적 개입의 부재만을 이유로 한 거부권을 인정하거나 인적 개입에 의한 재처리를 요구할 수 있도록 하지는 않았다는 점은 이와 무관하지 않다(제36조의2).

설명요구권과 공개의무를 통해 정보주체는 자신에 대한 자동화된 결정의 이유를 파악하고 그에 대하여 어떠한 대응을 할 것인지 결정할 수 있게 된다. 이 경우 정보주체가 궁극적으로 어떠한 법적 대응을 할 수 있는지는 결정이 이루어진 영역과 맥락에 따라 달라진다. 예컨대 장애인임을 이유로 채용을 거절당한 것이라면 차별금지법제, 구체적으로는 「장애인차별금지 및 권리구제 등에 관한 법률」에 따라 손해배상청구를 하거나(제46조), 국가인권위원회에 진정을 할 수 있고(제38조),¹⁴²⁾ 악의적 차별행위에 해당하는 경우 형사 고소를 할 수도 있을 것이다(제49조). 이처럼 규제의 강도를 결정하는 것은 특정한 영역과 맥락 하에서 규제의 근거가 되는 배후의 보호법익 내지 위험의 속성과 그 정도이며, 속성상 보호 필요성이 큰 위험이 존재하거나 위험의 정도가 높은 경우에는 이를 다루기 위한 별도의 법제도가 마련되어야 한다. 단지 절차적·수단적으로 투명성을 확보하기 위한 자동화된 결정에 대한 규율은 그 자체로 강한 규제가 될 수는 없는 것이다.¹⁴³⁾ 따라서 자동화된 결정 관련 규정의 위반이 있는 경우 정보주체의

141) Green, Ben/Yiling Chen, "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments", Proceedings of the conference on fairness, accountability, and transparency, 2019.

142) 국가인권위원회는 차별행위자에 대하여 차별행위의 중지, 원상회복, 재발방지조치 등의 구제조치 등을 권고할 수 있고, 차별행위자가 정당한 사유 없이 권고를 이행하지 않는 경우 이를 통보받은 법무부장관은 시정조치를 명할 수 있으며, 시정조치 불이행에 대하여는 과태료에 처하도록 되어있다(제43조, 제50조).

143) 이상용, "알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -", 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 149-150면; 한편, 설명의 가치를 도구적 가치와 내재적 가치로 구분하면서 이를 도구적 가치로 파악할 때 설명요구권의 법적 근거가 보다 견고해질 수 있다고 보는 견해로는 Selbst, Andrew/Julia Powles, "Meaningful Information and the Right to Explanation", Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 2018).

권리가 침해되었다는 점만을 강조하여 과도한 제재가 가해지지 않도록 운용의 묘를 살릴 필요가 있다.

2. 상황과 맥락의 고려

가. 공적 영역과 사적 영역의 차이

공적 영역과 사적 영역에 있어서 자동화된 결정과 관련된 위험은 질적으로 다르다. 자동화된 결정과 관련된 인간의 존엄성, 프라이버시와 자유, 민주주의, 공정성 등의 위험은 공적 영역과 사적 영역 모두에서 나타나지만, 공적 영역에서는 국가와 시민 사이의 관계나 공동체 구성의 원리가 문제된다는 점에서 그 위험이 더욱 중하게 여겨질 수밖에 없다. EU의 인공지능 법안에서 공공기관에 의한 사회적 평점 시스템의 채택을 수용 불가능한 위험으로 분류하여 금지한 것은 이와 관련이 있다.¹⁴⁴⁾ 위험의 구체적 내용 역시 달라질 수 있다. 예를 들어 공정성의 경우 공적 영역에서는 기회의 균등이 무엇보다 중요하다. 롤즈는 그의 정의론에서 사회적·경제적 불평등은 “공정한 기회 균등의 조건 아래 모든 사람들에게 개방된 직책과 직위가 결부되게끔 편성되어야 한다”고 천명한 바 있다.¹⁴⁵⁾ 이에 비해 사적 영역에서는 사적 자치(private autonomy)가 무엇보다 중요하며, “같은 것은 같게, 다른 것은 다르게”라는 아리스토텔레스 정식이 공정성의 잣대가 된다.

자동화된 결정과 관련하여 문제되는 위험 뿐만 아니라 그 해결방안의 이론적 근거 역시 공적 영역과 사적 영역에 있어서 차이가 있다. 예를 들어 사적 영역에서 자동화된 결정에 대한 거부권은 인간의 존엄성과 주로 관련되어 있는 반면, 공적 영역에서는 민주주의 및 법치주의와 주로 관련되어 있다. 자동적 처분이 재량이 없는 기속행위에만 허용되는 것은 이 때문이다. 설명요구권 역시 공적 영역에서는 법치주의 내지 적법절차의 원리에 의하여 정당화되는 반면, 사적 영역에서는 자율적 토대의 확충 필요성에 의하여 정당화된다. 규율의 정도를 조절하는 법원리도 다르다. 즉, 공적 영역에서는 과잉금지의 원칙에 의하여 조절되는 반면, 사적 영역에서는 이익형량의 원리에 의하여 조절된다.¹⁴⁶⁾

이처럼 자동화된 결정과 관련된 위험의 평가와 그 해결방안의 이론적 근거가 공적 영역과 사적 영역에 따라 달라지므로 각 영역에서의 규율 역시 달라질 수 있다. 그런 점에서 개정법에서 자동적 처분을 규율 대상에서 제외하여 행정기본법, 나아가 행정절차법의 법리에 따르도록 한 것은 수긍할 만하다.

144) EU AI Act §5(1)(c).

145) John Rawls(황경식 옮김), 정의론, 이학사, 2003, 400-401면.

146) 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023, 84면 참조.

나. 이익형량

사적 영역에 있어서 자동화된 결정에 대한 정보주체의 대응권과 개인정보처리자의 보호조치 의무를 어느 범위에서 인정할 것인지는 관련 당사자들의 이익형량을 거쳐 판단할 필요가 있다. 특히 설명요구권이나 공개의무(정보제공의무)는 타인의 자유 및 권리와 충돌할 수 있으므로 유의할 필요가 있다. 이와 관련하여 GDPR 전문 제63호는 자동화된 개인정보처리에 수반된 논리와 해당 정보처리의 결과 등에 관한 정보주체의 열람권 행사가 영업비밀이나 지적재산권 등 타인의 권리 및 자유에 악영향을 끼쳐서는 안 된다고 밝히고 있다.¹⁴⁷⁾ 설명요구권이 유용한 혁신을 저해하는 규제로 작동할 것이라는 우려도 있다.¹⁴⁸⁾

개정법은 개인정보처리자가 정당한 사유를 증명하여 의무에서 벗어날 수 있다고 함으로써 이익형량의 요소를 명시적으로 도입하였다(제37조의2 제3항). 공개의무의 경우에는 이에 관한 내용이 규정되어 있지 않지만, 같은 맥락에서 공개의무의 존부나 범위를 정함에 있어서는 이익형량적 고려가 있어야 할 것이다. 따라서 설명이나 공개가 영업의 자유, 영업비밀, 지적재산권 등 개인정보처리자 또는 제3자의 자유나 권리를 침해할 우려가 있는 경우, 개인정보처리자는 설명이나 공개를 통하여 얻을 수 있는 정보주체의 이익이 그로 인해 저해되는 개인정보처리자나 제3자의 이익보다 현저히 적다는 점을 증명하여 의무를 면할 수 있고, 그렇지 않더라도 관련 당사자들의 이익형량을 고려하여 설명이나 공개의 범위 및 방법이 정해져야 한다고 주장할 수 있을 것이다.

3. 자율적 접근과 후견적 접근의 조화

가. 자율적 접근 방식의 채택

자동화된 결정에 관한 규율 방식은 자율적 접근과 후견적 접근으로 나누어 볼 수 있다. 자율적 접근방식은 정보주체에게 자신과 관련된 위험을 통제할 수 있는 권리를 부여하여 스스로 그 행사 여부를 결정하도록 하는 것이다. 이에 비하여 후견적 접근방식은 국가가 나서서 정보주체의 권익을 보호하는 것이라고 볼 수 있다.

자동화된 결정과 관련된 위험이 인간의 존엄성, 프라이버시와 자유, 민주주의, 공정성 등 다양한 측면에서 나타날 수 있고, 기술과 산업의 발전에 따라 그러한 위험의 개연성이나 양상은 물론 수범자의 위험 통제 능력이 달라질 수 있으며, 활용 영역과 맥락에 따라 위험의 정도나 이익형량의 결과가 다르게 평가될 수 있다는 점을 고려하면 자율적 접근방식이 보다 합리적이

147) 다만 이를 이유로 정보주체에 대한 정보제공 자체가 거부되어서는 안 된다고 설명한다.

148) 이원태, “EU의 알고리즘 규제 이슈와 정책적 시사점”, KISDI Premium Report, 정보통신정책연구원, 2016, 20면.

라고 할 수 있다. 그런 점에서 개정법이 정보주체에게 자동화된 결정에 대한 거부권을 부여하고 그 행사 여부를 결정하도록 함으로써 정보주체의 주도 하에 자율적으로 위험 통제가 이루어지도록 한 것은 기본적으로 타당한 것이라고 할 수 있다.¹⁴⁹⁾

나. 후견적 접근 방식에 의한 보완

사적자치에 입각한 자율적 접근방법이 효율적이라면 몇 가지 전제조건이 갖추어져야만 한다. 즉, i) 당사자에게 충분한 정보가 제공되어야 하고, ii) 당사자가 그 정보를 제대로 인지하고 합리적으로 판단할 수 있어야 하며, iii) 그 판단의 토대 위에서 대등하게 상대방과 협상할 수 있어야 하는 것이다.¹⁵⁰⁾ 설명요구권을 비롯한 투명성 규제는 정보 격차를 해소함으로써 위와 같은 전제조건을 갖추려는 것이다. 이는 자율적 토대의 확충이라는 측면에서 중요한 의미를 갖는다.

그러나 정보의 제공이 언제나 모든 문제를 해결해주는 것은 아니다. 인간은 확증 편향이나 가용성 편향을 비롯한 각종 편향¹⁵¹⁾에 노출되어 환경을 정확히 인지하지 못하고, 위험과 확률의 평가에 취약하며, 현 상태를 기준으로 하여 이익보다는 손실에 민감한 행태를 보이는 등 판단과 의사결정에 있어서 제한된 합리성을 지니고 있을 따름이다.¹⁵²⁾ 이러한 인지적 한계는 정보의 제공만으로는 극복하기 어렵다. 인지 능력을 희소성 있는 자원의 일종으로 본다면 합리적 판단을 포기하는 행위 자체가 합리적인 것으로 여겨질 수도 있다.¹⁵³⁾ 인공지능 기술의 복잡성과 자동화된 결정의 증가 추세에 비추어볼 때 정보제공을 통한 자율적 토대의 확충 외에 적절한 수준의 후견적 조치가 바람직할 수도 있다. GDPR 제22조가 단지 정보주체의 권리의 창설에 그치는 것이 아니라 자동화된 결정의 원칙적 금지를 규정한 것으로 보는 견해는 이러한 입장에 기울어진 것으로 볼 수 있다.

다. 자율규제의 활용

그러나 이러한 후견적 조치는 공적규제가 아닌 자율규제라는 틀을 통해서도 가능하며¹⁵⁴⁾ 그

149) 김재완, “EU 일반정보보호규정(GDPR)의 알고리즘 자동화 의사결정에 대한 통제으로써 설명을 요구할 권리에 대한 쟁점 분석과 전망”, 민주법학 제69호, 민주주의법학연구회, 2019, 295면 참고.

150) 권영준, “계약법의 사상적 기초와 그 시사점”, 저스티스 124호, 한국법학원, 2011, 174면 참조.

151) 심리학 내지 행동경제학적 관점에서 바라본 인지적 편향을 의미하며 공정성의 맥락에서 논의되는 편향과는 구별된다.

152) Tversky, Amos/Daniel Kahneman(이영애 옮김), 불확실한 상황에서의 판단 : 추단법과 편향, 아카넷, 2001, 265-266면, 402-404면, 495-496면; Sunstein, Cass(ed), Behavioral Law & Economics, Cambridge University Press, 2000, pp. 2-10.

153) 서종희, “행동경제학적 관점에서의 소비자법의 개선방향 - 유럽소비자법 등을 중심으로 -”, 소비자법연구 제7권 제4호, 한국소비자법학회, 2021 참조.

것이 보다 바람직하다고 할 수 있다. 자율규제는 공적규제에 비하여 유연성이 높을 뿐만 아니라 효율적이기도 하다. 자율규제 기관의 전문성을 활용하고 신뢰를 형성함으로써 규제비용을 절감할 수 있고 규제비용을 내부화할 수 있기 때문이다.¹⁵⁵⁾ 이러한 장점은 인공지능 기술과 같은 범용 신기술 산업에서 더욱 빛나며, 자동화된 결정에 관한 규율 역시 마찬가지이다. 그런 점에서 책임성과 투명성에 초점을 맞춘 자율규제 방식을 취하고 있는 미국의 알고리즘 책임 법안은 유용하게 참고할 만하다.

자율규제는 구속력이 없다는 비판도 있지만, 여러 방안을 통해 규제로서의 실효성을 확보할 수 있다. GDPR의 인증(certification)이나 행동규약(code of conduct)이 대표적 사례이다. 즉, 개인정보처리자가 승인받은 행동규약을 준수하거나 인증을 받은 경우 GDPR을 준수하였다는 합법성 추정이 이루어질 수 있고, 자기구속의 법리상 인증이나 승인을 한 규제기관은 그에 따른 개인정보처리자의 행위를 위법하다고 판단하기 어려울 것이다.¹⁵⁶⁾ 무엇보다 규범은 강제보다는 내면화를 통하여 진가를 발휘한다.

자율규제에 강제력이 적다는 점은 반대로 큰 부담 없이 규율 범위를 넓힐 수 있다는 장점이 되기도 한다. 예컨대 개인정보보호법은 개인정보영향평가가 의무화되어 있는 공공기관과 달리 민간 개인정보처리자의 경우에는 자율규제에 맡기고 있다(제33조 제1항, 제11항).¹⁵⁷⁾ 이 경우 개인정보처리자는 법이 정하고 있는 의무를 넘어 자동화된 결정에 대한 의견진술의 기회를 부여하고 이의제기의 방법을 고지하는 등 상황과 맥락에 맞는 추가적 보호조치를 채택할 수 있다. 이를 통해 개인정보처리자는 정보주체의 자유와 권리에 대한 부정적인 영향이 완화됨을 증명하고, 문제를 사전에 예방하여 규제 준수 비용을 절감할 수 있을 것이다.¹⁵⁸⁾

154) Casey, Bryan, et al., "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise", Berkeley Technology Law Journal 34, 2019 참조.

155) Ogus, Anthony, Regulation, economics and the law, Edward Elgar Publishing, 2001, pp. 345-346.

156) 이상용, "알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -", 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020, 138면; 참고로 GDPR은 행동규약을 준수하거나 인증을 받은 경우 과태료 경감요인으로 고려하도록 하는 명시적 규정을 두고 있다[§83①(j)].

157) GDPR은 우리나라와 달리 일정한 요건 하에 민간에도 개인정보영향평가를 의무화하고 있는데[§35(1)], 부분적으로만 자동화된 결정의 경우나 프로파일링만 존재하고 자동화된 결정이 없는 경우처럼 §22(1)이 적용되지 않는 경우에도 개인정보영향평가가 요구된다(Article 29 Data Protection Working Party, Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679, p. 27). 한편 EU AI 법안은 고위험 인공지능인 경우 제3자에 의한 적합성 평가를 의무화하고 있는데, 생체인증 범주화, 채용·인사관리, 필수적 공공 및 민간 서비스 혜택과 같은 특정 목적의 경우에는 적합성 평가를 거치도록 하고 있다[§6(1)].

158) 장시영, "개인정보 영향평가 제도의 국내·외 현황 비교 및 시사점", 정보통신방송정책 제30권 14호, 정보통신정책연구원, 2018, 3면

VIII. 결 론

자동화된 결정은 그 효율성과 일관성으로 인해 점점 증가하고 있지만, 인간의 존엄성, 프라이버시와 자유, 민주주의, 공정성 등과 관련하여 우려가 제기되기도 한다. 이러한 우려를 완화하기 위하여 개인정보보호법제, 차별금지법제, 인공지능규제법제, 행정법제 등 여러 법제도적 방안이 검토되어 왔지만, 그 중에서도 포괄적이고 사전적이며 수단적인 성격을 띠는 개인정보보호법제가 가장 주목을 받고 있다.

개정법 제37조의2로 신설된 자동화된 결정에 대한 정보주체의 대응권 규정은 그 모태인 GDPR보다 균형 있고 합리적인 규율을 시도하였다는 점에서 높이 평가할 만하다. 그럼에도 불구하고 보다 바람직한 운영을 위해서는 몇 가지 고려할 점이 있다. 첫째, 기술적 한계를 고려할 필요가 있다. 사회관념상 기술적으로 불가능한 것을 요구할 수는 없기 때문이다. 특히 인공지능 기술이 지닌 블랙박스적 특성, 성능과 투명성의 상충관계, 공정성 지표의 한계 등을 고려할 필요가 있다. 둘째, 신기술에 대한 절차적·수단적 규제로서의 특성을 고려하여 경직되고 과도한 규제를 피할 필요가 있다. 셋째, 상황과 맥락을 충분히 고려하여 운영될 필요가 있다. 특히 사적 영역은 공적 영역과 달리 사적자치와 이익형량이 우선적으로 고려되어야 함을 염두에 두어야 한다. 넷째, 자율과 후견의 적절한 조화를 모색할 필요가 있다. 자동화된 결정에 관한 위험을 통제하기 위하여 정보주체에게 대응권을 부여한 것 자체가 자율적 접근을 취한 것이다. 일부 후견적 접근이 필요한 경우에도 자율규제를 우선적으로 검토해야 할 것이다. ☐

(논문접수 : 2023. 12. 5. / 심사개시 : 2023. 12. 29. / 게재확정 : 2024. 1. 19.)

참 고 문 헌

1. 국내문헌

단행본

민법주해 VII, 채권(5), 박영사, 1999,.

박은정 외, 차별판단기준 수립을 위한 연구(제1부), 국가인권위원회, 2021.

인터넷진흥원, EU 일반 개인정보보호법 가이드북, 2020.

최규환, 인간존엄의 형량가능성, 헌법재판연구원, 2017.

최승철, 차별금지법의 이해, 한울아카데미, 2011.

한국인공지능법학회, 인공지능과 법, 박영사, 2019.

한국장애인개발원, 장애인차별금지법과 상충되는 고용 관련 연구, 2019.

한희원, 정의로의 산책(제2판), 삼영사, 2011.

John Rawls(황경식 옮김), 정의론, 이학사, 2003.

Stuart Russell/Peter Norvig(류광 옮김), 인공지능: 현대적 접근방식, 제이펍, 2016.

Tversky, Amos/Daniel Kahneman(이영애 옮김), 불확실한 상황에서의 판단 : 추단법과 편향, 아카넷, 2001.

논문

강승식, “절차적 적법절차의 본질에 관한 연구”, 미국헌법연구 제25권 제2호, 미국헌법학회, 2014.

고학수/정해빈/박도현, “인공지능과 차별”, 저스티스 제171호, 한국법학원, 2019.

권건보, “개인정보자기결정권”, 헌법판례 100선, 한국헌법학회, 2012.

권건보, “정보접근권의 실효적 보장에 관한 고찰 - 정보주체의 정보접근권을 중심으로 -”, 미국헌법연구 제29권 제2호, 미국헌법학회, 2018.

권영준, “계약법의 사상적 기초와 그 시사점”, 저스티스 124호, 한국법학원, 2011.

권은정, “자동적 처분의 행정법적 제 문제: 처분의 형식·절차 및 구제에 관한 쟁점을 중심으로”, 법학논총 제38권 제3호, 한양대학교 법학연구소, 2021.

김도훈, “알고리즘 책임성 논의와 알고리즘에 대한 이해”, 주간기술동향 제16권 제1호, 정보통신기술진흥센터, 2018.

김대인, “행정기본법과 행정절차법의 관계에 대한 고찰”, 법제연구 제59호, 한국법제연구원, 2021.

- 김도승, “인공지능 기반 자동행정과 법치주의”, 미국헌법연구 제30권 제1호, 미국헌법학회, 2019.
- 김민호/윤금남, “자동화된 의사결정과 행정절차법에 관한 검토”, 성균관법학 제34권 제3호, 성균관대학교 법학연구원, 2022.
- 김병필, “인공지능 공정성 심사 기준에 관한 연구”, 서울대학교 박사학위논문, 2023.
- 김연진, “개정 개인정보 보호법에 관한 시론적 고찰”, 성균관법학 제35권 제2호, 성균관대학교 법학연구원, 2023.
- 김예진, “인공지능에 의한 자동화된 의사결정의 위험성 - 장애인 채용을 중심으로 -”, Law & technology 제18권 제1호, 서울대학교 기술과법센터, 2022.
- 김일환, “개인정보자기결정권의 헌법상 근거와 보호에 관한 연구”, 공법연구 제29집 제3호, 한국공법학회, 2001.
- 김재선, “알고리즘 자동행정결정에 대한 행정법적 해석방안에 관한 연구 - 미국 행정법상 입법방안 논의를 중심으로”, 법학논총 제45권 제3호, 단국대학교 법학연구소, 2021.
- 김재완, “EU 일반정보보호규정(GDPR)의 알고리즘 자동화 의사결정에 대한 통제로써 설명을 요구할 권리에 대한 쟁점 분석과 전망”, 민주법학 제69호, 민주주의법학연구회, 2019.
- 김종철, “적법절차원리의 헌법화에 대한 소견 - 미국연방헌법상 적법절차원리의 기능을 중심으로 -”, 인권법평론 제20호, 전남대학교 법학연구소 공익인권법센터, 2018.
- 김희정, “알고리즘 자동의사결정으로부터 개인의 보호 : 인간의 개입권을 중심으로”, 헌법학연구 제26권 제1호, 한국헌법학회, 2020.
- 남중권, “머신러닝 알고리즘의 데이터 처리에 대한 법적 제한의 한계: 개인정보보호와 차별 금지의 측면에서”, 과학기술과 법 제10권 제1호, 충북대학교 법학연구소, 2019.
- 박노형/정명현, “EU GDPR상 프로파일링 규정의 법적 분석”, 안암법학 제56호, 안암법학회, 2018.
- 박노형/김효권, “자동화된 결정에 관한 개인정보보호법 정부 개정안 신설 규정의 문제점 - EU GDPR과의 비교 분석”, 사법 제62호, 사법발전재단, 2022.
- 박상돈, “헌법상 자동의사결정 알고리즘 설명요구권에 관한 개괄적 고찰”, 헌법학연구 제23권 제3호, 한국헌법학회, 2017.
- 서중희, “행동경제학적 관점에서의 소비자법의 개선방향 - 유럽소비자법 등을 중심으로 -”, 소비자법연구 제7권 제4호, 한국소비자법학회, 2021.
- 성윤희, “행정기본법상 자동적 처분에 관한 연구”, 이화여자대학교 대학원 법학과 박사학위논문, 2023.
- 양종모, “인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율방

- 안”, 법조 제723호, 법조협회, 2017.
- 오요한/홍성욱, “인공지능 알고리즘은 사람을 차별하는가?” 과학기술학연구 제18권 제3권, 한국과학기술학회, 2018.
- 이상용, “인공지능과 계약법”, 비교사법 제23권 제4호, 한국비교사법학회, 2016.
- 이상용, “데이터 거래의 법적 기초”, 법조 제67권 제2호, 법조협회, 2018.
- 이상용, “알고리즘 규제를 위한 지도(地圖) - 원리, 구조, 내용 -” 경제규제와 법 제13권 제2호, 서울대학교 공익산업법센터, 2020.
- 이상용, “인공지능과 법”, 새봄을 여는 민법학(정태윤·지원립 교수 정년기념논문집), 홍문사, 2023.
- 이원태, “EU의 알고리즘 규제 이슈와 정책적 시사점”, KISDI Premium Report, 정보통신정책연구원, 2016.
- 장시영, “개인정보 영향평가 제도의 국내·외 현황 비교 및 시사점”, 정보통신방송정책 제30권 14호, 정보통신정책연구원, 2018.
- 정태호, “개인정보자결권의 헌법적 근거 및 구조에 관한 고찰 - 동시에 교육행정정보시스템(NEIS)의 위헌 여부의 판단에의 그 응용 -”, 헌법논총 제14집, 헌법재판소, 2003.
- 조성규, “인공지능에 기반한 자동화된 행정결정의 행정법적 쟁점”, 동북아법연구 제16권 제4호, 전북대학교 동북아법연구소, 2023.
- 주민호, “자동화된 의사결정에 대한 기본권의 실효적 보장- EU AI법 제54조와 GDPR 제22조의 관계를 중심으로 -”, 법학논고 제82호, 경북대학교 법학연구원, 2023.
- 차상욱, “EU 개인정보보호법제의 최근 입법과 시사점”, 정보법학 제21권 제1호, 한국정보법학회, 2017.
- 최경진, “빅데이터와 개인정보”, 성균관법학 제25권 제2호, 성균관대학교 법학연구소, 2013.
- 홍성수, “포괄적 차별금지법의 필요성”, 이화젠더법학 제10권 제3호, 이화여자대학교 젠더법학연구소, 2018.
- 황지은, “인공지능 관점에서 GDPR 제22조의 인간 개입에 대한 분석”, 경제규제와 법 제13권 제1호, 서울대학교 법학연구소, 2020.

2. 외국문헌

단행본

- Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.

- BeckOK DatenschutzR/von Lewinski, 43. Ed. 1.2.2023.
- Dreyer, Stephan/Wolfgang Schulz, "The General Data Protection Regulation and Automated Decision-making: Will it deliver", Bertelsmann Stiftung, 2019.
- European Parliament, The impact of the General Data Protection Regulation (GDPR) on artificial intelligence, 2020.
- Executive Office of the President, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, 2016.
- FRA, Handbook on European Data Protection Law, 2018.
- ICO, Automated decision-making and profiling, 2018.
- ICO, ICO consultation on the draft AI auditing framework guidance for organisations, 2020.
- ISO/IEC, Artificial Intelligence (AI) - Assessment of the robustness of neural networks, Technical Report(TR 24029), 2021.
- Hildebrandt, Mireille/Serge Gutwirth (eds), Profiling the European Citizen: Cross-Disciplinary Perspectives, Springer, 2008.
- New York State Department of Financial Service(NYDFS), Report on Apple Card Investigation, 2021.
- OECD, Recommendation of the Council on Artificial Intelligence, OECD Publishing, 2019.
- Ogus, Anthony, Regulation, economics and the law, Edward Elgar Publishing, 2001.
- O'Neil, C., Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2016.
- Paal, Boris/Daniel Pauly, Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling, Beck'sche Kompakt-Kommentare Datenschutz-Grundverordnung (3. Auflage), C.H, Beck, 2021.
- Sunstein, Cass(ed), Behavioral Law & Economics, 2000, Cambridge University Press, 2000.

논문

- Angwin, J., et al., "The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review", ProPublica, 2015.
- Angwin, J., et al., "Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks", ProPublica, 2016.
- Barocas, Solon/Andrew Selbst, "Big data's disparate impact", California law Review

104.3, 2016.

Barocas, Solon, et al., *Fairness and Machine Learning: Limitations and Opportunities*, online textbook, 2022.

Bartlett, Robert, et al., "Consumer-lending discrimination in the FinTech Era", *Journal of Financial Economics*, 143.1, 2022.

Bickel, Peter, et al., "Sex bias in graduate admissions: Data from Berkeley", *Science* 187, 1975.

Bradford, Anu, "The Brussels Effect", *Northwestern University Law Review* 107, 2012.

Burrell, J., "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms", *Big Data and Society* 3.1, 2016.

Bygrave, Lee A, "Article 22 Automated Individual Decision-making, Including Profiling" in Christopher Kuner et al (Ed.), *The EU General Data Protection Regulation (GDPR): A Commentary*, Oxford University Press, 2020.

Casey, Bryan, et al., "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise", *Berkeley Technology Law Journal* 34, 2019.

Ciancimino, Michele, "AI-Based Decision-Making Process in Healthcare", *Journal of European Consumer and Market Law* 11.5, 2022.

Citron, D. Keats/Frank Pasquale, "The Scored Society : Due Process for Automated Predictions", *Washington Law Review* 89:1, 2014.

Cofone, Ignacio, "Antidiscriminatory privacy", *SMU Law Review* 72, 2019.

Corbett-Davies, Sam, et al., "Algorithmic decision making and the cost of fairness", *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017.

Custers, Bart, "Data Mining and Group Profiling on the Internet" in Anton Vedder (Ed.), *Ethics and the internet*, Intersentia, 2001.

Coglianese, Cary/Alicia Lai, "Algorithm vs. Algorithm", *Duke Law Journal* 72, 2022.

Dressel, J./H. Farid., "The Accuracy, Fairness, and Limits of Predicting Recidivism", *Science Advances*, 4.1, 2018.

Fiss, Owen, "Groups and the Equal Protection Clause", *Philosophy & Public Affairs* 107, 1976.

Goodman, Bryce/Seth Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'", *AI Magazine* 38.3, 2017.

- Green, Ben/Yiling Chen, "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments", Proceedings of the conference on fairness, accountability, and transparency, 2019.
- Gunning, David/David Aha, "DARPA's explainable artificial intelligence program", AI Magazine, Association for the Advancement of Artificial Intelligence, 40.2, 2019.
- Hellman, Deborah, "Measuring algorithmic fairness", Virginia Law Review 106.4, 2020.
- Hildebrandt, Mireille, "Profiling and the Identity of the European Citizen", in Mireille Hildebrandt and Serge Gutwirth (eds), Profiling the European Citizen: Cross-Disciplinary Perspectives, Springer, 2008.
- Kamarinou, Dimitra, et al., "Machine Learning with Personal Data", Queen Mary University of London Legal Studies Research Paper No. N247/2016, 2016.
- Kaplow, Louis/Steven Shavell. "The conflict between notions of fairness and the Pareto principle", American Law and Economics Review 1.1, 1999.
- Kim, Pauline, "Auditing algorithms for discrimination", University of Pennsylvania Law Review 166, 2017.
- Kleinberg, Jon, et al., "Inherent trade-offs in the fair determination of risk scores", arXiv preprint arXiv:1609.05807, 2016.
- Kleinberg, Jon, et al., "Human decisions and machine predictions", The quarterly journal of economics 133.1, 2018.
- Kleinberg, Jon, et al., "Discrimination in the Age of Algorithms", Journal of Legal Analysis 10, 2018.
- Kuner, Christopher, et al., "Machine learning with personal data: is data protection law smart enough to meet the challenge?", 7.1 International Data Privacy Law, 2017.
- Kuner, Christopher, et al., "The GDPR as a chance to break down borders", International Data Privacy Law 7.4, 2017.
- Narayanan, Arvind, "21 fairness definition and their politics", ACM Fairness, Accountability and Transparency Conference, 2018.
- Paal, Boris, "Case Note: Article 22 GDPR: Credit Scoring Before the CJEU", Global Privacy Law Review, 4.3, 2023.
- Raghavan, Manish, et al., "Mitigating bias in algorithmic hiring: Evaluating claims and practices", Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.
- Schermer, Bart W, et al, "The Crisis of Consent: How Stronger Legal Protection May

- Lead to Weaker Consent in Data Protection", *Ethics & Info. Tech* 16, 2014.
- Shwartz-Ziv, Ravid/Amitai Armon, "Tabular data: Deep learning is not all you need", *Information Fusion* 81, 2022.
- Selbst, Andrew, "Disparate Impact in Big Data Policing", *Georgia Law Review*, 52, 2018.
- Selbst, Andrew/Julia Powles, "Meaningful Information and the Right to Explanation", *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR, 2018.
- Smith, Helen, "Algorithmic bias: should students pay the price?", *AI & Society* 35.4, 2020.
- Tyler, Tom R., "Procedural fairness and compliance with the law", *Revue Suisse D Economie Politique et de Statistique* 133, 1997: 219-240.
- Veale, Michael/Lilian Edwards, "Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling", *Computer Law & Security Review*, 34.2, 2018.
- Wachter, Sandra, et al., "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", *International Data Privacy Law* 7.2, 2017.
- Wachter, Sandra, et al., "Counterfactual explanations without opening the black box: Automated decisions and the GDPR", *Harv. JL & Tech.* 31, 2017.
- Wachter, Sandra, et al., "Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law", *West Virginia Law Review* 123.3, 2021.
- Zittrain, Jonathan, "Engineering an Election: Digital gerrymandering poses a threat to democracy", *Harvard Law Review Rorum*, 127, 2014.

Abstract

Regulation of Automated Decision-Making by the Personal Information Protection Act

Lee, Sang-Yong

Lee, Hye-Ri

Automated decision-making is on the rise due to its efficiency and consistency. However, concerns have been raised regarding human dignity, privacy, freedom, democracy, and fairness. To address these concerns, various legal measures, including data protection laws, anti-discrimination laws, artificial intelligence regulations, and administrative laws, have been under consideration. Among these, data protection laws, characterized by their comprehensive, preventive, and instrumental nature, have garnered the most attention.

The provision on the right to object to automated decision-making newly established in Article 37-2 of the revised Korean Personal Information Protection Act is highly appreciated in that it attempted a balanced and rational regulation compared to its parent, the GDPR. However, there are a few things to consider for more desirable operation.

First, it is necessary to consider technical limitations. This is because it is not possible to demand something that is technically impossible in social norms. In particular, it is necessary to consider the black-box characteristics of artificial intelligence technology, the conflict between performance and transparency, and the limitations of fairness metrics. Second, it is necessary to avoid rigid and excessive regulations, considering the characteristics of procedural and means-oriented regulations for new technologies. Third, it is necessary to operate by taking into account the situation and context. In particular, it should be kept in mind that private autonomy and balancing of interests should be prioritized in the private sector, unlike the public sector. Fourth, it is necessary to find an appropriate balance between autonomy and guardianship. The fact that the right to object is granted to the data subject to control the risks of automated decision-making itself is a

112 선진상사법률연구 통권 제105호 (2024.1.)

self-regulatory approach. Even in cases where some supervisory approach is needed, self-regulation should be prioritized.

Key Words : automated decision-making, fairness, personal information protection law, PIPA, GDPR, anti-discrimination law