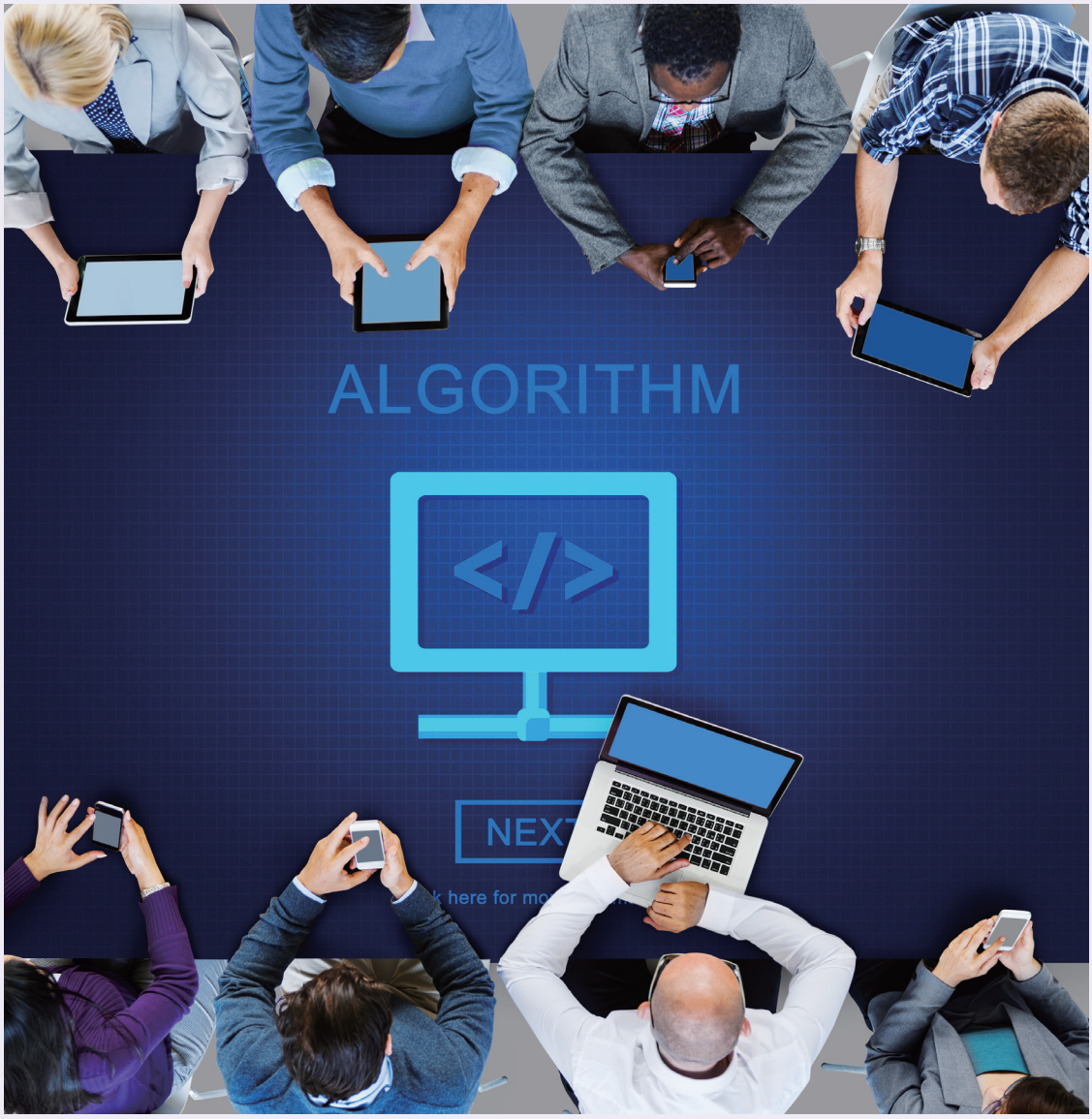
01 AI알고리즘과 가치중립성 확보 방안

이상직 법무법인(유한)태평양 변호사·AI-지식재산특별전문위원장



1. AI와 AI알고리즘

최근 야당 원내대표의 국회 본회의장 교섭단체 연설이 포털 사이트 메인 화면에 노출되면서 AI알고리즘을 통한 뉴스배열이 가치중립성을 준수할 수 있는 것인지에 관하여 논란이 일었다. 포털 측은 개인의 주관에 따라 조작될 수 없는 AI알고리즘에 따라 뉴스를 배열하고 있을 뿐 아니라 인위적으로 뉴스 배열에 개입하지 않는다고 항변했다. 그러나 AI알고리즘이라고 하더라도 그 규칙을 설계하는 사람들의 생각이 어떠한 형태로든 반영될 수밖에 없으므로 뉴스 배열이 가치중립적이라고 쉽게 단정하기 어렵다는 반론도 거세게 제기되고 있다.

정부는 2030년까지 디지털 경쟁력 세계 3위, 경제효과 최대 455조 원 창출, 삶의 질 세계 10위를 목표로 AI 국가전략을 수립하여 추진하고 있다. AI는 데이터, 네트워크와 함께 정부가 역점을 두고 추진하는 한국판 뉴딜 중 디지털 뉴딜의 핵심이기도 하다. AI는 인간의 정신적 활동을 모방하는 기계, 소프트웨어 등 유무형의 장치 또는 체계를 말한다. 머신러닝, 딥러닝 등을 통하여 지각, 추론, 학습 등 인간 지능과 유사한 활동을 한다.

AI의 유형을 보자. 강한 AI는 대량의 데이터를 학습하여 인간이 사전에 예상하지 못한 결과까지 만들어내는 AI이다. 약한 AI는 복잡한 계산을 수행하지만, 인간이 정한 규칙을 벗어나지 않는 AI이다. 현재 AI의 사례로 거론되는 대부분이 약한 AI에 속한다. 강한 AI의 구현은 다소 시간이 걸릴 것으로 보인다. 따라서 AI에 관한 진흥 및 규제 정책이 집중적으로 논의되어야 할 분야도 약한 AI이다. AI알고리즘은 AI의 두뇌라고 할 수 있는데, 주어진 문제를 논리적으로 해결하기 위하여 필요한 절차, 방법, 명령어, 규칙들의 집적체를 말한다. AI알고리즘은 실시간 교통상황이나 실시간 뉴스 검색, 쇼핑 품목 조회는 물론이고 방송, 인터넷 콘텐츠의 선호도 순위, 판례 검색이나 스포츠 결과 예측 등 일반적인 실생활 영역에서 전문적인 영역까지 인간의 정신적 활동을 의미 있는 수준으로 대체하는 상황에 이르고 있다.

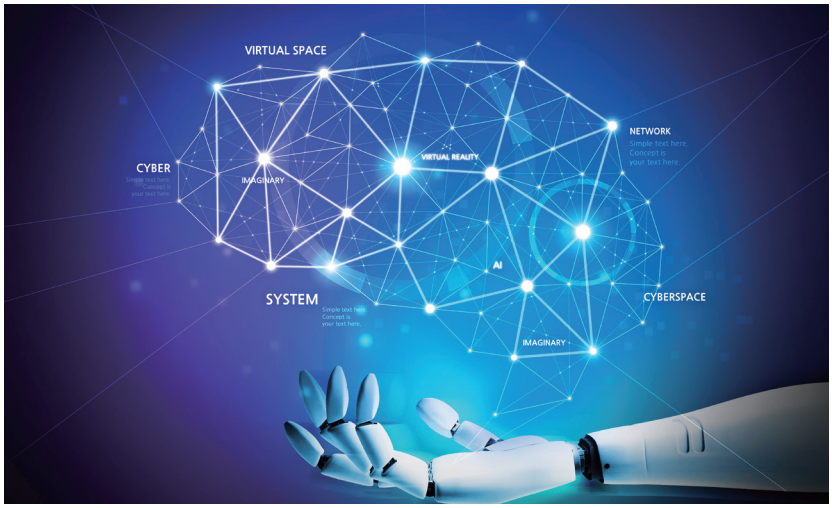
먼저 AI가 가져올 혜택을 보자. AI를 활용하여 전에 없던 새로운 비즈니스 모델, 소프트웨어, 운영시스템, 장비, 거래 플랫폼을 만들 수 있다. AI를 활용하여 발명과 특허 출원, 저작 등 창작이 쉬워진다. 온 국민이 발명가, 작곡가, 소설가가 될 수 있다. AI를 활용해 수학적 계산이 필요한 복잡한 금융상품도 쉽게 개발할 수 있다. 의료, 건강 데이터를 AI를 통해 분석하여 새로운 치료법이나 신약을 짧은 기간에 만들어낼 수 있다. AI기기를 교육수단으로 활용한다면 교육의 질도 대폭 향상할 수 있다. AI를 복지정책에 활용하면 복지 사각지대를 쉽게 확인하여 적절한 구제조치를 취할 수 있다. 자녀의 학교 출석 여부,

부모의 직장 출퇴근 여부, 병원 출입 여부 및 시기 등을 정밀 분석하여 사각지대 취약계층을 보호할 수 있는 것이다. AI를 활용한 지리 정보 수집, 분석 등을 통해 국토의 균형 발전도 꾀할 수 있다. 부동산 투기 성향, 통계 등을 AI시스템에서 분석하여 투기 억제에 직접적인 효과가 있는 부동산정책을 개발, 수립할 수도 있을 것이다.

그러나 AI가 가져올 혜택을 상상하더라도 마냥 즐겁기만 한 것은 아니다. 우려스러운 점도 있다. 인공지능의 오작동으로 자율주행차량 등이 사고를 내어 인간의 생명, 신체에 위해를 줄 수 있다. AI를 이용하여 다른 기업이나 개인의 영업 비밀을 빼낼지도 모른다. AI를 활용해 금융 시장을 조작할 수 있고, 고객을 기망하여 상품을 판매할 수 있다. 강력한 AI를 가진 기업을 중심으로 담합하거나 시장에서의 지배력을 남용하고, 협력·거래 업체를 괴롭힐 수 있다. AI가 병원 치료 내역 등 환자의 민감 정보를 이용하는 경우에는 사생활을 침해할 수도 있다. AI가 내 일자리를 뺏을 수 있다. AI를 적용한 결과에 따라 인종, 사회적 지위, 성향, 경제적 부의 규모, 성별 등을 이유로 차별이 발생하는 경우도 배제할 수 없다. 현재까지는 컴퓨터 프로그램이 내린 의사결정의 원인과 과정, 결과에 관하여 확인과 설명을 할 수 있고 책임 소재도 밝힐 수 있는 것이 일반적이다. 그러나 AI를 적용한 경우에는 머신러닝, 딥러닝의 결과로 그 시스템을 작동한 사람조차도 어떤 과정을 거쳐 왜 그러한 결정이 이루어졌는지 알 수 없으므로 책임 소재를 밝히기 어려울 수 있다.

그럼에도 불구하고 AI는 전 세계 국가가 제4차 산업혁명의 중요한 추진전략으로 삼아 정체된 산업을 일으키기 위하여 반드시 가야 할 길로 인식하고 적극적으로 지원하고 있는바, 그 방향성을 의심할 만한 심각한 장애는 발견되지 않고 있다. 그러나 AI알고리즘의 작동 과정에 사람이 인위적으로 개입하지 않는다고 하더라도 AI에 관한 프로그램을 설계하고 규칙화하여 적용하는 것은 사람이므로 이 과정이나 그 적용 결과에 따라 설계자나 누군가의 주관, 의도, 목적 또는 의도하지 않은 제3의 변수들이 개입될 가능성이 여전히 남아있다. 따라서 AI의 진전에 따라 예상되는 부작용과 단점을 미리 짚어보고 대비한다면 AI에 대한 국민의 수용도를 높일 수 있을 뿐 아니라 그 혜택을 극대화할 수 있을 것이다.

2. AI알고리즘에 관한 법적 보호



국가는 특정인이 고도의 기술적인 창작물을 발명한 경우에 독점하여 사용할 수 있는 권한을 부여하는 등 지식재산권을 행사할 권리를 발명자에게 부여하는데, 이를 특허라고 한다. AI알고리즘도 그 유형과 내용을 구체적으로 따져보아야 하겠지만 특허법상의 요건을 갖추면 특허로서 보호받을 수 있다. 산업경영에서 이용할 수 있어야 하고, 신규성을 갖추어야 하며, 해당 기술 분야에서 통상의 지식을 가진 자가 쉽게 발명할 수 없는 등 진보성이 필요하고, 선량한 풍속이나 기타 사회질서에 반하는 등 불특허사유가 없다면 특허 출원 및 등록을 할 수 있다.

한편, 부정경쟁방지 및 영업비밀 보호에 관한 법률은 영업비밀을 보호하고 있다. 영업비밀은 공공연히 알려져 있지 아니하고 독립된 경제적 가치를 지니는 것으로서 합리적인 노력에 의하여 비밀로 유지된 생산방법, 판매 방법, 그밖에 영업활동에 유용한 기술상 또는 경영상 정보를 말한다(법 제2조). 영업비밀을 침해하는 행위에 대해서는 금지청구권, 손해배상청구권 등의 행사를 통하여 보호받을 수 있고(법 제10조, 제11조 등), 침해자를 형사처벌(법 제18조)할 수도 있다. AI알고리즘은 특정 사업을 목적으로 가격, 거래조건, 거래 방법 등의 전부 또는 일부에 AI를 활용하여 필요한 절차, 방법, 명령어, 규칙들을 집적 및 제시하는 것이다. AI알고리즘은 공공연히 알려진 것이 아니고 인적·물적 투자 등 노력에 의한 결과물이므로 그 기업의 영업비밀에 속한다. 따라서 AI알고리즘 침해에 대해서는 금지청구권, 손해배상청구권 등을 행사하여 보호받을 수 있다.

3. AI알고리즘의 부정적 특징 : 편향성과 불투명성

AI알고리즘은 기업의 영업비밀 또는 특허 등 지식재산권의 보호범위에 속하지만 편향성, 불투명성의 문제를 야기할 수 있으므로 미리 해결방안을 찾아볼 필요가 있다. 먼저 몇 가지 사례를 들어본다.[1)]

사례 1: A사의 챗봇 B는 Social Network Service 미디어(이하 SNS)에 게시된 글을 학습하여 인간의 언어를 이해하고 표현할 수 있도록 설계된 대화형 AI다. SNS의 데이터를 그대로 읽어 학습한 결과, 인종을 차별하거나 여성을 혐오하는 듯한 표현을 대화 형태의 메시지로 드러냈다.

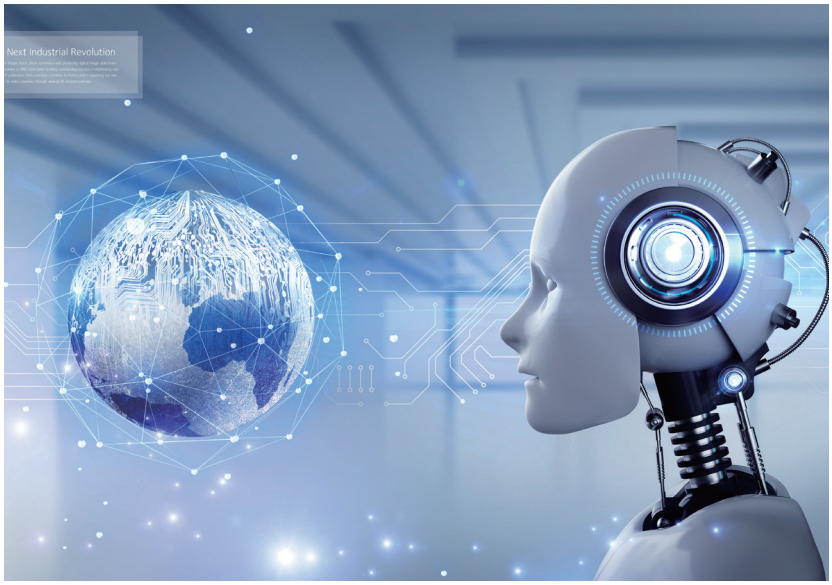
사례 2: B사는 AI를 이용하여 직원 채용을 자동화하는 프로그램을 개발하였으나 실험 결과, 여성 지원자를 차별하는 경향이 나타났다(지난 수년간 B사에 제출된 이력서 패턴을 분석하도록 훈련되었는데 이력서 대부분이 남성 지원자가 제출한 것이었다).

사례 3: 뉴스 포털은 뉴스를 배열하고 노출하는 순위를 정함에 있어서 이용자의 클릭 횟수 등 관심도를 반영하는 AI알고리즘을 가지고 있는데, 견해를 달리하는 집단이 자신에게 유리한 뉴스의 노출 및 검색 순위를 높이고자 소속원에게 클릭을 독려하는 운동을 일으켰다.

(가) 편향성

헌법 및 각종 법률은 국민에 대한 부당한 차별을 금지하고 있다. AI알고리즘을 적용한 결과에 따라 인종, 사회적 지위, 성향, 경제적 부의 규모, 성별 등을 이유로 차별이 발생하는 경우가 있을 수 있다. 이를 AI알고리즘의 편향성이라고 한다. AI알고리즘의 편향성이 발생하는 원인을 보자.

1) 필자는 자율주행 또는 AI알고리즘의 편향성과 관련하여 비유적인 사례로 "김유신의 말"에 빗대 설명하기도 한다. 김유신은 신라의 촉망받는 화랑이었는데 천관녀 라는 기생의 술집에 드나들면서 학업을 소홀하게 된다. 부모의 꾸중을 들은 김유신은 정신을 차리고 훈련에만 열중하고 천관녀의 집에 가지 않는다. 어느 날 훈련에 지친 김유신은 말을 타고 귀가하는 길에 잠이 들게 되고 말은 그를 천관녀의 집으로 인도하게 되는데, 김유신은 책임을 물어 말을 죽였다. 말은 김유신이 훈련을 마치고 자신의 집에 가기 전에 천관녀의 집에 들른다는 데이터를 양적 측면에서 많이 가지고 있고, 김유신이 부모의 꾸중을 들은 때로부터 천관녀의 집에 들르지 않는다는 질적으로 중요한 데이터를 소홀히 하였다. 그 결과 주인이나 그 사회의 규범이 요구하는 결과가 아닌 천관녀의 집으로 향했다는 편향된 결과를 보여주었다. 그러나 말은 책임이 없다. 책임을 져야 할 법적 인격도 없을 뿐 아니라 문제가 되는 결과를 야기한 데이터는 사실상 김유신에 의하여 투입된 것이다. 그러나 김유신에게도 책임을 물을 수 없다. 그는 훈련에 지쳐 잠깐 졸았을 뿐이다. 해결방법은 세 가지로 압축된다. 첫 번째 방법은 말이 천관녀에게 가지 못하도록 사전에 교육하는 것이다. 그러나 말의 행동을 예측할 수가 없는데 어떻게 교육이 가능할까. 즉, AI알고리즘의 작동 결과를 우리가 미리 알 수 없는데, 그와 같은 규칙을 사전에 넣을 수 있는지 의문이다. 두 번째 방법은 김유신이 천관녀에게 가곤 했다는 데이터를 사전에 제외하는 것이다. 그러나 AI알고리즘의 학습에 투입되는 데이터를 임의로 조정한다면 AI알고리즘이 충분한 역할을 할 수 있을지 의문이고, 왜곡된 결과가 도출될 가능성도 배제할 수 없다. 셋째, 말이 천관녀의 집에 도달했을 때에 김유신이 말머리를 자신의 집으로 돌리는 것이다. AI알고리즘의 작동 결과에 규범적 판단이 필요 없는 경우는 문제가 없지만, 규범적 판단이 필요하다면 김유신 같은 사람이든 다른 AI알고리즘이든 개입을 통하여 최악의 결과를 막아야 할 것이다.



첫째, AI알고리즘의 설계, 기획 단계에서 편견이 들어가 차별을 야기하는 경우이다. 예를 들어 네비게이션이나 가상비서 프로그램의 목소리는 대부분 여성이다. 이용자에게 편안함을 주기 위해서라면 별다른 문제는 없을 것이다. 그러나 이용자의 요구나 명령을 수동적으로 따르는 것이 여성의 역할이라는 편견이 서비스의 기획에 무의식적으로 반영된 것이라면 그것은 전혀 다른 문제가 된다. 너무 앞서나간 것일 수도 있지만 AI알고리즘을 설계하는 엔지니어의 대부분이 남성이어서 자연스럽게 성차별적인 요소가 반영되는 것일 수도 있다. 인간은 내심으로는 많은 편견을 가지고 있고 그것이 오히려 자연스러울 수 있다. 그러나 편견을 가진 사람이라고 하더라도 규범적 억제를 통하여 그 편견을 외부에 표출하지 않는 것이 일반적인데, AI는 아직 그 수준에 이르지 못하고 있다.

둘째, AI시스템 기획, 디자인, 개발, 시험, 적용 등 과정의 전부 또는 일부에서 외부 요인의 개입으로 편향성이 발생하는 경우이다. 예를 들면, AI 가상비서가 이용자 등 사람의 성희롱 발언을 학습하거나 수용하면서 편향성을 가질 수도 있다. 위 사례 3의 경우와 같이 견해를 달리하는 집단이 자신에게 유리한 뉴스의 노출 및 검색 순위를 높이고자 소속원에게 클릭을 독려하는 운동을 일으킨 결과로 뉴스 독자의 실질적인 선호를 반영하지 못하는 뉴스 배열이 나오는 것도 편향성을 가져오는 원인이 된다.

셋째, AI가 의존하는 데이터가 잘못되어 AI알고리즘을 적용한 결과도 편향성을 띄는 경우이다. 위 사례 1에서 보는 바와 같이 AI가 SNS에 게시된 글을 여과 없이 받아들이고 학습하여 인간의 언어를 이해하고 표현함에 따라 사람을 차별하거나 여성을 혐오하는 결과물을 만들어내기도 한다. 알고리즘의 설계 문제에 더하여 학습 과정에 투입된 데이터가 가지는 대표성이나 공정성 등의 질적 측면이 고려됨이 없이 그 양적 측면만 고려되는 경우에도 편향성이 발생할 수 있다. 특히 자유민주주의 사회에서 개인의 의견이 여과 없이 표현되는 현실을 고려한다면, 포털과 소셜미디어, 온라인 커뮤니티의 게시글 등이 그 자체로 특정 집단의 이해관계만을 반영하거나 다양한 분야에서의 차별과 불평등을 내포할 경우에 AI 출력물의 편향성은 더욱 커질 것이다.

넷째, AI가 고객의 편견을 찾아내어 마케팅에 활용하는 과정에서 편향성을 나타낼 수도 있다. 제4차 산업혁명은 기존의 산업혁명과 달리 빅데이터 분석과 AI를 통해 고객의 숨은 성향과 니즈를 찾아 그에 맞는 상품과 서비스를 고객이 원하는 시점에, 고객이 원하는 장소에, 고객이 원하는 방법으로 전달하는 특징을 갖는다. 그 과정에서 고객의 편향성을 확인하고 그 편향성에 부합하는 상품이나 서비스를 제공하려고 한다면 당연히 AI알고리즘이 편향성을 의도하고 따를 수밖에 없을 것이다.

다섯째, AI가 대외적으로 의사 표현이나 결정을 드러내는 마지막 단계 이전

의 적절한 과정에서 편향성을 교정하는 알고리즘을 가지고 있지 않은 경우이다. 사람은 온갖 편견을 가지고 있지만, 그것을 마음속으로만 담아두는 경우에는 양심의 자유로 보호된다. 그러나 외부로 편향성이 표출되는 경우에는 사안에 따라 명예훼손 등 법령을 위반하게 된다. 마찬가지로 특단의 사정이 없는 한 AI알고리즘을 적용한 결과를 활용하여 기업 등의 내부에서 의사결정을 위한 요소로 참고하기 위해서만 사용하는 것을 편향되었다고 말할 수는 없고, 외부에 표현되어 누군가를 차별하거나 부당한 처우를 하는 경우에 주로 문제된다(내부 직원에 대하여 근로관계 등에 관한 의사결정을 하기 위하여 이용하는 경우도 같다). 그렇다면 AI알고리즘을 이용하여 대외적인 의사표시나 행위를 하는 경우에는 그 전 단계에서 편향성 여부를 판단하여 제어할 수 있어야 하는데 그렇지 못한 경우에 편향성의 문제가 불거질 수 있다.

AI알고리즘의 편향성은 법적으로 허용되지 않은 차별이나 부당한 처우를 가져올 수 있으므로 피해 방지를 위해 어떠한 정책과 법 제도를 만들 것인지 중요해진다.

(나) 불투명성

AI시스템은 그 작동 과정이나 결과의 전부 또는 일부에서 인간이 알 수 없는 현상을 야기할 수 있고, 이를 불투명성이라고 한다. 현재 컴퓨터 프로그래밍의 경우 대부분 그 프로그램이 내린 결정의 원인과 과정, 결과에 관하여 확인과 설명을 할 수 있고, 그로 인한 책임 소재도 밝힐 수 있다. 그러나 AI알고리즘의 경우 머신러닝, 딥러닝의 결과로 그 시스템을 작동한 기업조차도 어떤 과정을 거쳐 왜 그러한 결정이 이루어졌는지 알기 어렵다. 그것은 AI알고리즘이 추구하는 목적이기에 당연한 것이다. 그러나 AI알고리즘의 작동으로 누군가에게 피해가 발생했을 경우에 그 적용 과정이나 결과를 사전에 알 수 없다는 이유로 면죄부를 주는 것은 바람직하지 않다. 따라서 AI알고리즘의 불투명성을 해소하는 것이 중요한 법·정책적 과제가 되고 있다.

4. AI알고리즘의 편향성, 불투명성 제거 등 가치 중립성 확보를 위한 과제

앞서 본 바와 같이 AI알고리즘은 그 고유의 특징으로서 편향성과 불투명성의 위험이 있다. 그러나 AI알고리즘이 영업비밀 또는 지식재산권으로 보호받음에 따라 어떻게 하여야 영업비밀이나 지식재산권을 침해하지 않으면서 그 편향성과 불투명성을 극복할 수 있는지가 중요한 과제가 된다.

(가) 해외 동향

(1) EU GDPR(General Data Protection Regulation)은 EU 회원국에 직접 적용되는 개인정보보호법으로서 개인정보 보호라는 관점에서 AI알고리즘의 편향성과 불투명성을 제거하고 가치중립성을 높이려고 한다. GDPR 제13조, 제22조는 정보 주체가 기업에 AI알고리즘의 의사결정을 확인할 수 있는 인과관계에 관하여 최소한의 설명을 요구할 수 있는 법적 근거를 두고 있다. 먼저 설명요구권에 관하여 보자. 정보 주체가 기업에 AI알고리즘 등에 의한 프로파일링(자동화된 의사결정)으로 인하여 자신의 개인정보가 언제 누구에게 어디까지 알려지고 이용될 것인지 정보제공 등 설명을 요구할 수 있는 권리이다. 다음으로, 이의제기권에 관하여 보자. AI알고리즘의 적용 결과에 이의를 제기하고 해명과 확인을 요구할 수 있는 권리이다. EU는 기업의 프로파일링으로부터의 개인정보 보호 규정을 근거로 AI알고리즘의 편향성 및 불투명성 제거도 가능하다는 입장에서 출발하고 있다. 다만, AI에 관한 것이라고 하더라도 개인정보의 범위를 벗어난 사항을 추가적으로 규제하기 위해서는 별도의 법 제도가 필요하다는 것을 부인하지는 않는다.

한편, EU AI윤리가이드라인(2019년 4월 제정)에서 편향성 및 불투명성 제거와 가치중립성 제고를 위하여 참고할 만한 것을 보면, 첫째, 추적 가능성이다. AI알고리즘이 어떤 데이터를 이용하고 어떻게 동작하는지에 관한 정보가 AI시스템에 기록되어 있어야 한다. 둘째, 설명 가능성이다. AI시스템에 의한 의사결정을 사람이 이해할 수 있어야 한다. AI시스템의 의사결정이 개인에게 영향을 미치는 경우에 그 이유를 설명해 줄 것을 요구할 수 있다. 설명 방법도 일반인, 규제기관, 연구자 등의 전문성에 맞추어 이루어져야 한다. AI가 의사결정 시스템을 기획 설계한 목적과 그 작동과정이 어떻게 이루어지고 개인에게 어떤 영향을 미치는지 등에 관한 설명을 적시에 제공해야 한다. 셋째, 소통이다. 상대방이 AI시스템이라는 것을 알 수 있어야 한다. AI시스템과의 소통을 거부하는 경우에 사람과 소통할 수 있는 기회를 제공해야 하고, AI시스템의 능력과 한계

에 대한 정보를 제공해야 한다.

(2) 미국 국가과학기술자문회의(National Science and Technology Council Committee on Technology)는 2016년 AI보고서(Preparing for the Future of Artificial Intelligence)에서 AI를 주요 성장동력으로 보되, 공정성, 안전, 투명성, 이해 가능성, 인간 가치에 관한 정부의 역할을 강조했다. 영국 정부는 2018년 AI 보고서(AI in the UK: Ready, Willing and Able?)에서 데이터 접근과 제어, 이해할 수 있는 AI, 디지털 이해력 증진, 공중보건 관리, AI 위험 완화를 중요시하고 있다. 일본 총무성은 2019년 인간중심의 AI 사회 원칙을 발표하면서 인간중심, 교육 교양, 개인정보 보호, 보안, 공정경쟁, 공정성, 책임성, 투명성, 혁신에 방점을 두었다. 즉, EU 이외의 국가에서도 AI가 준수해야 할 중요사항의 하나로 편향성과 불투명성의 제거를 통해 가치중립성을 확보할 것을 요구하고 있다.

(3) 결국 주요 선진국은 AI에 관한 진입 규제를 최소화하여 산업역량 강화의 기회로 삼고, AI를 위한 생태계 등 인프라 조성을 중요한 요소로 보고 있으며, 법적 규제 이전에 윤리적 가이드를 준수할 것을 요구하고 있다. 그러면서도 일관되게 AI가 가져올 해악에 대한 규제는 해당 기업의 규모, 서비스 내용과 형태, 고객에 미치는 영향, 다른 산업과의 연관성을 고려하여 맞춤형 규제를 해야 한다는 입장이다. 아울러, AI 독과점, AI를 통한 담합 등 시장을 저해하는 행위 및 지배력 남용에 대해서는 일벌백계해야 한다고 보고 있다.

(나) AI의 가치중립성을 위한 기본원칙 및 윤리적 가이드

AI의 가치중립성을 확보하기 위하여 우리가 지켜야 할 원칙이나 기준은 어떠해야 할까. 먼저 사람 중심으로 사람의 가치를 최우선하는 AI알고리즘을 구현해야 한다. 사람 중심의 AI 구현을 위해 AI의 개발부터 활용에 이르는 전 과정에서 개발자, 공급자, 이용자 등 모든 사회 구성원이 참조할 수 있는 기본적 윤리기준을 제시해야 한다. 구속력과 강제력 있는 법이 아닌 윤리의 형태로 기준을 제시함으로써 AI를 활용하는 기업의 자율성을 존중하고 기술발전을 장려하며, 기술과 사회변화에 유연하게 대처할 수 있게 해야 한다. 산업경제 분야의 자율규제 환경 조성을 통해 AI 연구개발 및 산업 성장을 저해하지 않아야 하고, 개발자 및 공급자에게 부당한 부담을 지우게 해서는 안 된다. 범용성을 가진 일반원칙으로서의 윤리기준을 제시하여 다양한 분야에서 AI 윤리기준의 참조모

델이 되고, 사안별 또는 분야별 AI 윤리기준 제정의 근거를 제공하여 영역별 세부 규범이 유연하게 발전해 나갈 수 있는 기반을 조성해야 한다. 또한, 사회 경제 및 기술 변화와 함께 새롭게 제기되는 AI 윤리 이슈를 반영하여 윤리기준의 지속적인 수정 및 보완을 가능케 하는 AI 윤리 플랫폼으로 기능할 수 있도록 해야 한다.

결국 윤리기준이 지향하는 최고 가치는 인간성의 존중인데, AI가 인간에게 유용해야 할 뿐 아니라 인간 고유의 성품을 훼손하지 않고 오히려 보존하고 함양할 수 있도록 개발되고 관리되어야 한다. 인간 가치의 존중을 위하여 인간의 행복추구, 인권보장, 개인정보보호, 다양성 존중, 해악 금지, 공공성, 개방성, 연대성, 포용성, 안전한 데이터 관리, 책임성의 확보, 통제성, 안전성, 투명성이 요구된다고 할 것이다. 국민이 주도하는 지능정보화시대에 국민이 고개를 끄덕일 수 있는 멋진 AI원칙과 기준이 주어진다면 편향성, 불투명성 제거를 통한 가치중립성이 확보되고 발전적으로 가꾸어져 나갈 것으로 기대된다.

다 법률적 해결방안

법률적 측면에서도 기업의 특허 등 지식재산권과 영업비밀을 보호하면서도 AI의 편향성과 불투명성을 제거하여 고객을 보호하기 위한 다양한 논의가 있다.

첫째, 계약 등 시장원리와 자율규제에 의하자는 견해를 들 수 있다. 예를 들어, 뉴스를 편집하는 AI알고리즘은 복수의 엔지니어와 전문가 자문을 통하여

설계되는 것인 만큼 구조적으로 의사 결정력이 있는 특정 개인의 주관에 의하여 좌우될 가능성을 줄일 수 있는 측면이 있다. 그러나 AI알고리즘이 작동하는 규칙 등의 설계자가 여전히 사람일 수밖에 없는 한계가 있으므로 사람의 의견이 전혀 반영되지 않는다고 단정하기는 어렵다. 사람을 채용하는 AI알고리즘이 학점, 사회봉사 실적, 에세이, 발표력 등에 가중치를 같이 하거나 달리하는 경우에 과연 그것만으로 가치중립적이라고 할 수 있을지 의문이다.[2)]

그러나 AI알고리즘의 불투명성 및 편향성 제거가 고객의 상품 선택을 끌어내는 중요요소이므로 시장에서의 공정한 경쟁을 통하여 불투명성, 편향성에 관한 문제점을 해소하는 것은 중요하다. 또한 기업이 편향성과 불투명성 제거 및 가치중립성 확보에 관한 사회적 책임을 다하도록 하고, 소속 협회, 단체의 자율규제(가치중립성 확보를 위한 윤리 가이드라인, 지침, 규약, 회원사가 위반하는 경우의 자율적인 제재 등)와 시민단체의 감시, 견제를 통해 간접적으로 규제하는 것이다. 이 견해는 시장원리를 존중하고 AI의 산업적 발전을 가속화하는 장점이 있다. 그러나 기업 스스로의 의지에만 의존함으로써 기업의 의사결정에 따라 실질적인 고객 보호가 미흡할 수 있다는 것이 단점이다. AI알고리즘 자체는 사람이 설정해 놓은 목표나 목적을 위해 봉사하게 된다. 기업의 경우는 매출이나 영업이익을 극대화하는 방향으로 자동으로 알고리즘이 움직이게 설계하는 것이 일반적이고 그 목적을 달성하지 못할 것 같은 경우에는 알고리즘이 변경,

2) AI가 인간의 개입 없이 독립 후 스스로 데이터를 수집하여 학습하고 판단해 의사결정을 하는 단계에 이른다면 그것은 다른 문제이다. 이 경우에는 해당 AI에게 사람이나 법인과 같은 법적 인격을 부여할 것인지부터 달리 논의되어야 할 것이다.

조작되기도 하는데, 과연 자율규제로 이러한 문제를 해소할 수 있을지는 의문이다. 물론, 자율규제에 참여하는 기업 간에 서로 감시, 견제할 수 있는 시스템을 갖추는 것이 중요하다.

둘째, 입증책임의 전환 등 민사적으로 다룰 수 있는 방안을 찾자는 견해이다. 시장원리에 따라 AI알고리즘의 적용과정이나 결과로 인하여 계약불이행이나 불법행위가 발생한 경우에 계약책임, 불법행위책임으로 처리한다. 다만, 입법 또는 판례로 입증책임을 피해자가 아닌 기업에 돌려 해당 기업이 AI알고리즘의 적용과정에 관한 인과관계나 결과에서 편향성, 불투명성이 야기되어 고객 등의 피해가 발생한 것에 관하여 자신에게 과실이나 위법이 없음을 입증케 하는 것이다. 대법원은 "일반적으로 불법행위로 인한 손해배상청구 사건에 있어서 가해행위와 손해 발생 간의 인과관계의 입증책임은 청구자인 피해자가 부담하나, 대기오염이나 수질오염에 의한 공해로 인한 손해배상을 청구하는 소송에 있어서는 기업이 배출한 원인 물질이 물을 매체로 하여 간접적으로 손해를 끼치는 수가 많고 공해 문제에 관하여는 현재의 과학 수준으로도 해명할 수 없는 분야가 있기 때문에 가해행위와 손해의 발생 사이의 인과관계를 구성하는 하나하나의 고리를 자연과학적으로 증명한다는 것은 극히 곤란하거나 불가능한 경우가 대부분이므로, 이러한 공해소송에 있어서 피해자에게 사실적인 인과관계의 존재에 관하여 과학적으로 엄밀한 증명을 요구한다는 것은 공해로 인한 사법적 구제를 사실상 거부하는 결과가 될 우려가 있는 반면에, 가해 기업은 기술적·경제적으로 피해자보다 훨씬 원인조사가 용이한 경우가 많을 뿐만 아니라, 그 원인을 은폐할 염려가 있고 가해 기업이 어떠한 유해한 원인 물질을 배출하고 그것이 피해물건에 도달하여 손해가 발생하였다면 가해자 측에서 그것이 무해하다는 것을 입증하지 못하는 한 책임을 면할 수 없다고 보는 것이 사회형평의 관념에 적합하다(대법원 2002. 10. 22. 선고 2000다65666 판결)"고 판시하였는바, 입증책임의 전환은 원칙에 대한 예외이므로 함부로 확장하여 볼 것은 아니나 AI알고리즘의 적용으로 인한 피해에 관한 책임 문제를 해결하는 하나의 방법으로 그 논의가 시작되어야 할 것으로 보인다. 이 견해는 AI의 산업적 발전에는 도움이 되나 고객 피해 발생을 사전에 막을 수 없다는 단점이 있다.

셋째, 관계 법령의 제·개정을 통하여 정부의 행정규제를 확립하자는 견해이다. 고객에게 피해를 주는 AI알고리즘의 적용과정이나 결과 중 법적으로 금지되어야 할 사항을 미리 입법으로 정립하거나 고객이 AI알고리즘의 편향성, 불투명성으로부터 스스로를 보호할 수 있도록 설명요구권, 이의제기권 등 법적 권리

를 부여하자는 견해이다. 이 견해는 일반법 또는 사업법 등 관계 법령에 AI알고리즘의 불투명성, 편향성 제거 및 가치중립성 확보에 관한 세부 사항을 기업의 의무사항으로 규정하는 법 조항을 도입할 것을 주장한다. 이 경우, 주무 행정기관은 법이 정한 바에 따라 해당 유형에 해당하는 금지행위를 조사하여 제재 등 시정조치를 할 수 있다. 다른 한편으로는 AI알고리즘에 관한 설명요구권, 이의제기권, 적용거부권 등을 고객에게 법적 권리로 부여하고, 기업이 이를 부당하게 거부하는 경우 또는 부당한 거부를 통하여 피해를 발생시키는 경우에 주무 행정기관이 제재조항을 발동할 수 있게 하는 것이다.

5. 결론

글로벌 경쟁 사회에서 미래에 대한 섣부른 규제는 국가의 성장동력 정체를 야기하므로 AI알고리즘에 대한 초기 규제는 최소화함과 동시에 자율규제를 독려하는 것이 바람직해 보인다. 그 대신 기업이 AI알고리즘의 잘못된 적용 결과를 예방할 수 있는 최소한의 사전 장치를 두도록 하고(기술적·관리적·물리적 조치사항 도입), 기업들의 의견을 수렴하여 영업비밀이나 지식재산권을 침해하지 않는 범위내에서 AI알고리즘의 목적, 용도나 개요 등 핵심요소의 대략적인 사항만 편향성, 불투명성 제거를 위하여 공개하는 수준이 타당할 것으로 판단된다.