영남대학교 법학연구소 인권교육연구센터
「인권이론과 실천」
제29호 (2021.6.) pp.67~101.

인공지능 윤리의 현황과 과제*

양 천 수**

Ⅰ. 서 론

오늘날 인공지능 기술을 포함한 지능정보기술은 다양한 사회적 공리를 창출한다. 이를 통해 우리에게 여러 혜택을 제공한다. 눈부시게 발전하는 인공지능 기술 덕분에 우리는 자신에게 최적화된 서비스를 제공받을 수 있다. 하지만 세상의 모든 것이 흔히 그렇듯이 인공지능 기술은 우리에게 새로운 사회적 · 법적 문제를 야기한다. 이로 인해 우리는 어떻게 하면 인공지능이 던지는 새로운 문제에 적절하게 대응할 수 있는지를 고민해야 한다.

인공지능이 성공적으로 구현되고 사용되려면 하드웨어와 소프트웨어, 개인정보를 포괄하는 빅데이터 및 이에 대한 사회적 수용이 요청된다. 하드웨어는 우리가 흔히 아는 반도체 기술을, 소프트웨어는 알고리즘을 중심으로 하는 프로그래밍 기술을 말한다. 인공지능에 대한 이론은 이미 1950년대에 대부분 완성되었지만 그 당시에는 이를 뒷받침해줄 수 있는 하드웨어와 빅데이터가 존재하지 않아 인공지능이 구현되기 어려웠다. 그중에서 특히 알고리즘과 빅데이터에 관해 오늘날 문제가 제기된다. 인공지능이 원활하게 가동되기 위해서는 우리의 개인정보를 포함하는 빅데이터가 필요한데 이로 인해 개인정보 침해라는 문제가 야기될 수 있다. 또한 인공지능을 가동하는 데 핵심적인 축이 되는 알고리즘은 부정확성이나 불투명성, 편향성이라는 문제를 야기한다. 이로 인해 특히 수학을 기반으로 하는 알고리즘에 관해 '대량살상

---

* 이 글은 필자가 2020년에 "인공지능 윤리"에 관해 정보통신정책연구원에 자문한 내용을 바탕으로 한 것입니다.

** 영남대학교 법학전문대학원 교수 · 법학박사.

수학무기'(WMD: Weapons of Math Destruction)라는 우려와 비판이 제기되기도 한다.[1] 인공지능에 내재된 알고리즘이 사회의 거의 모든 영역에서 편향을, 차별을, 포함과 배제를 강화한다는 것이다.

인공지능이 우리 사회에 수용되기 위해서는 사회적 수용에 필요한 조건을 충족해야 한다. 이는 객관적인 측면과 주관적인 측면으로 구별할 수 있다. 첫째, 객관적인 측면에서는 인공지능이 창출하는 사회적 공리가 인공지능이 야기하는 문제보다 더욱 커야 한다. 이를 위해서는 두 가지 요건을 충족해야 한다. 우선 인공지능이 창출하는 사회적 공리를 사회의 모든 구성원이 누릴 수 있어야 한다. 나아가 인공지능이 야기하는 사회적 · 법적 문제를 사회의 규범체계가 적절하게 해결할 수 있어야 한다. 다음으로 둘째, 사회 구성원들이 주관적인 차원에서 인공지능에 공감을 할 수 있어야 한다. 이를 위해서는 인공지능에 관해 바람직한 상징적 의미가 형성되고 소통되어야 한다. 달리 말하면 인공지능에 관해 설득력 있는 '상징정치'가 이루어져야 한다.[2]

이 글은 이러한 조건 중에서 인공지능이 유발하는 사회적 · 법적 문제를 해결하는 방안에 초점을 맞추고자 한다. 그중에서도 윤리라는 규범으로 인공지능 문제를 해결하고자 하는 시도를 검토한다. 인공지능이 유발하는 가장 어려운 규범적 문제 중 한 가지는 '알고리즘의 편향성' 문제라 할 수 있다. 편향성 문제는 오늘날 매우 중요한 규범적 원칙으로 자리매김하는 차별금지 원칙을 위반한다는 점에서 중대한 문제로 볼 수 있다. 이에는 두 가지 방안을 모색할 수 있다. 첫째는 가장 대표적인 규제수단인 법을 이용해 대응하는 방안이다. 둘째는 윤리라는 연성 규제수단으로 대응하는 방안이다. 이 중에서 법으로 대응하는 방안은 다시 두 가지 방안, 즉 일반적 차별금지법을 제정해 대응하는 방안과 개별적인 영역에서 차별금지를 하는 방안으로 구별할 수 있다. 현재

---

1) 캐시 오닐, 김정혜(역), 「대량살상 수학무기」, 흐름출판, 2017, 참고.
2) 상징정치에 관해서는 大貫恵美子, 「人殺しの花: 政治空間における象徴的コミュニケーションの不透明性」, 岩波書店, 2020, 참고.

우리 법체계는 후자의 방안을 채택하고 있다. 첫 번째 방안도 사회적으로 논의가 진행되고 있지만 아직은 제도화되지 않았다. 이러한 이유에서 우리 법제는 인공지능 편향성 문제는 개별법을 통해 대응하고 있다고 말할 수 있다. 다만 인공지능, 특히 알고리즘이 야기하는 문제가 매우 광범위하다는 점에서 이러한 방안이 과연 적절한지 아니면 인공지능의 편향성 문제를 일반적으로 규율하는 일반법 또는 기본법을 제정해야 하는 것은 아닌지 의문이 들 수 있다. 물론 이에는 다시 다음과 같은 근원적인 의문을 제기할 수 있다. 현재 상황에서 인공지능 편향성 문제를 법으로 규제하는 것이 적절한지의 의문이 그것이다. 이는 인공지능이 유발하는 규범적 문제를 현재로서는 윤리의 차원에서 규제하는 것이 더욱 바람직하지 않은가의 문제제기로 연결된다.

이러한 문제 상황에서 이 글은 인공지능이 유발하는 규범적 문제를 윤리로 대응하고자 하는 논의를 살펴본다. 인공지능 윤리의 현황을 점검하는 것이다. 이를 위해 이 글은 가장 대표적인 인공지능 윤리 가이드라인이라 할 수 있는 「유네스코 AI 윤리 권고안(초안)」과 최근 발표된 「국가 인공지능(AI) 윤리 기준」을 검토하고 개선방향을 논의하고자 한다.

Ⅱ. 인공지능의 위험 사례

먼저 오늘날 인공지능이 어떤 위험을 지니는지 살펴본다. 인공지능은 데이터, 알고리즘, 사회적 이용의 측면에서 다양한 위험을 창출한다. 그러면 이러한 위험들이 실제로 어떻게 나타나는지 구체적인 사례들을 검토한다.

1. 개인 데이터 남용

인공지능은 엄청난 양의 개인 데이터를 필요로 한다. 이로 인해 인공지능을 활용하고자 하는 이들, 특히 이를 영업에 사용하고자 하는 사업자들은 어

떻게든 개인 데이터를 획득하고자 애를 쓴다. 물론 우리나라나 유럽연합의 경우에는 엄격한 사전동의 방식의 개인정보 자기결정권을 채택하고 있기에 개인 데이터의 탈법적 수집이나 남용이 어느 정도 억제되는 편이다. 그렇지만 우리처럼 엄격한 개인정보 자기결정권을 수용하지 않는 미국에서는 탐욕에 눈이 먼 사업자들이 개인 데이터, 특히 사회적 약자의 개인 데이터를 약탈적으로 수집하고 남용하는 사례가 다수 발생한다.[3] 예를 들어 개인 데이터를 수집하기 위해 진정성이 없는 광고를 올리고 이러한 광고로 개인 데이터를 무차별적으로 수집한 후 이를 필요로 하는 사업자들에게 판매하는 것이다.[4] '데이터 경제'라는 이름 아래 탈법적으로 개인 데이터를 수집한 후 이를 매도하는 것이다. 이렇게 판매된 개인 데이터는 각 데이터 주체, 특히 사회적 약자에 속하는 데이터 주체를 경제적으로 약탈하는 데 사용된다.

2. 부정확한 알고리즘 사용

인공지능에서는 알고리즘이 핵심적인 역할을 한다. 인공지능이 제대로 작동하려면 알고리즘이 정확하고 공정하게 사용되어야 한다. 만약 알고리즘 설계가 잘못되면 인공지능이 산출하는 결과 역시 부정확할 수밖에 없다. 이에 대한 예로 2010년 전후로 미국 워싱턴 DC 교육청이 사용한 '임팩트'(IMPACT)라는 교사 평가 기법을 들 수 있다.[5] 교육개혁의 일환으로 무능한 교사를 선별하기 위해 도입된 임팩트에서는 '가치부가모형'(value-added model)이라는 알고리즘을 사용하였다. 가치부가모형은 가능한 한 정성적인 평가지표는 배제한 채 오직 정량적인 평가지표만으로 교사들을 평가하였다. 하지만 이로 인해 교장과 학부모들 사이에서 높은 평판을 받고 있던 교사를

3) 우리나라나 유럽연합과는 달리 미국은 사후승인(opt-out) 방식의 개인정보 보호체계를 갖고 있다. 이에 관해서는 이상경, "미국의 개인정보보호 입법체계와 현황에 관한 일고", 「세계헌법연구」 제18권 제2호, 세계헌법학회 한국학회, 2012, 195-214면 참고.

4) 캐시 오닐, 앞의 책, 139-140면.

5) 이에 관한 상세한 소개는 캐시 오닐, 앞의 책, 16면 아래 참고.

무능한 교사로 평가하여 해고되게끔 하는 결과를 빚고 말았다. 물론 이에 대해서는 두 가지 판단을 할 수 있다. 첫째는 해당 교사가 받고 있던 높은 평판이 잘못된 것일 수 있다는 점이다. 이는 학생 교육 능력과는 무관한 평판일 수 있다는 것이다. 둘째는 가치부가모형이라는 알고리즘이 잘못 설계되어 정확하지 않은 결과가 빚어졌고 이로 인해 해당 교사가 억울하게 해고되었다는 것이다. 평가 결과에 의문을 품던 해당 교사는 문제를 제기하였고 조사 결과 임팩트가 사용한 알고리즘이 정확하지 않게 설계되었다는 점이 확인되었다. 가치부가모형이라는 알고리즘이 애초에 잘못 설계되는 바람에 이로 교사를 평가하는 과정에서 정확하지 않은 결과가 도출된 것이다. 그 때문에 특정 교사의 직업적 생명이 결정되고 말았다.

또 다른 예로 알고리즘이 오작동하여 주식시장을 혼란에 빠트리는 사례를 언급할 수 있다. 오늘날 주식시장에서는 주식을 매매하는 데 인공지능이 적극 사용된다. 이를테면 주식 알고리즘은 인간은 따라 하기 힘든 판단능력과 속도로 주식거래에 참여한다. 이를 통해 금융회사는 막대한 이익을 챙기지만 간혹 인공지능이 오작동하여 주식시장에 혼란을 야기한다. 다수의 주식 알고리즘들이 동시에 주식가격을 터무니없게 설정함으로써 주식시장이 비합리적으로 움직이게 하는 것이다. 말하자면 업무상 과실로 주가조작을 하는 것이다. 이로 인해 주식시장에 참여하는 평범한 (개미라고 불리는) 주식거래자들이 피해를 입는다.[6)]

3. 편향된 알고리즘이 유발하는 문제들

(1) 대량살상 수학무기

인공지능에 활용되는 알고리즘은 인공지능에 입력(투입)되는 대량의 데이

6) 이에 관한 상세한 내용은 크리스토퍼 스타이너, 박지유(옮김), 「알고리즘으로 세상을 지배하라」, 에이콘, 2016, 7면 아래 참고.

터, 즉 빅데이터를 분석하고 이를 통해 새로운 패턴이나 가치 등을 창출하는 데 기여한다. 금융거래에서 손쉽게 볼 수 있듯이 빅데이터를 분석함으로써 미래를 예측하는 데도 활용된다. 그렇지만 이 과정에서 알고리즘은 입력되는 데이터나 알고리즘 자체의 한계 등으로 인해 편향성을 갖는 경우가 많다. 그리고 이러한 편향성은 사회적으로 크나큰 문제를 야기한다. 그 때문에 수학자이자 빅데이터 전문가인 캐시 오닐(Cathy O'Neil)은 이러한 문제를 야기하는 알고리즘을 '대량살상 수학무기'(Weapons of Math Destruction: WMD)라고 명명한다.

캐시 오닐은 특정한 알고리즘이 대량살상 수학무기, 즉 WMD로 지칭되려면 세 가지 요건을 충족해야 한다고 말한다. 불투명성, 확장성, 피해가 그것이다.[7] 먼저 WMD는 불투명하다. 이는 다음과 같은 의미를 지닌다. 첫째, WMD는 해당 알고리즘이 어떻게 작동하는지를 명확하게 설명하지 않는다는 것이다. 둘째, 이로 인해 WMD는 편향성, 즉 특정한 판단 대상들을 합리적 이유 없이 차별한다는 것이다.

다음으로 WMD는 특정한 영역에서만 제한적으로 사용되는 알고리즘의 지위를 넘어 다른 영역까지 확장될 수 있다. 이를테면 무능한 교사를 평가하기 위해 개발된 알고리즘이 신용불량의 위험이 높은 사람을 평가하거나 범죄를 다시 저지를 위험성이 높은 사람을 평가하는 알고리즘으로 확장되는 경우를 말한다.

나아가 이처럼 적용 영역이 확장됨으로써 WMD는 사회 곳곳에서 막대한 피해를 야기한다. 이때 말하는 피해는 주로 알고리즘의 편향성에서 비롯하는 경우가 많다. 알고리즘의 편향성으로 말미암아 특정한 판단 대상들이 합리적인 이유 없이 차별을 받아 사회적 피해가 야기되는 것이다. 그런데 더 큰 문제는 이렇게 사회적 피해가 야기되는 경우에 이를 강화하는 악순환이 발생한다는 것이다. 알고리즘의 편향성으로 유발된 사회적 피해 결과가 다시 알고리즘에 부정적으로 환류되어, 다시 말해 그 자체가 유용한 데이터가 되어

7) 캐시 오닐, 앞의 책, 60-61면.

≪편향 ⇒ 사회적 차별 ⇒ 사회적 피해≫라는 악순환의 고리가 고착되고 심화되는 것이다. 말을 바꾸면 알고리즘의 편향성으로 유발된 사회적 차별이 시간이 지나면서 확증편향 되는 것이다. 이로 인해 알고리즘 편향성이 유발한 ≪포함-배제≫라는 구별은 더욱 심화된다. 캐시 오닐은 이렇게 알고리즘이 WMD로 작용하는 경우에 관해 다양한 사례를 제시하는데 그중 몇 가지를 아래에서 소개한다.

(2) 대학평가 모델

먼저 대학평가 모델을 언급할 수 있다. 오늘날 언론 등이 대학을 평가하여 순위를 매기는 일은 일상적인 현상으로 자리매김하였다. 그런데 이러한 대학평가가 대학이 지닌 역량을 제대로 평가하는지에는 오래 전부터 의문이 제기되었다. 오늘날 일상이 된 대학평가는 1983년으로 거슬러 올라간다.[8] 지금은 대학평가로 전 세계적으로 유명해진 미국의 시사 잡지 「유에스 뉴스 & 월드 리포트」(US News & World Report)가 치열한 언론사 간의 경쟁에서 살아남기 위해 대학평가를 시작한 것이다.

문제는 「유에스 뉴스 & 월드 리포트」를 위시하는 대학평가 기관이 대학을 평가하기 위해 사용하는 데이터에서 찾을 수 있다. 무엇이 과연 대학의 진정한 역량을 평가하는 데 적합하고 유용한 데이터인지가 명확하지 않기에 대학평가 기관은 이른바 '대리 데이터'를 사용하여 대학을 평가한다. 이에 관해 두 가지 문제가 제기된다. 첫째, 대리 데이터는 대학의 역량을 정확하게 평가하는 데 한계를 지닌다는 점이다. 말 그대로 '대리' 데이터이기 때문이다. 'SAT 점수'나 '학생 대 교수 비율', '입학 경쟁률'과 같은 대리 데이터는 대학의 역량을 '간접적'으로 평가할 수 있지만 대학의 진정한 역량을 '직접적'으로 평가하는 데는 한계가 있다는 것이다. 둘째, 이러한 대리 데이터는

8) 캐시 오닐, 앞의 책, 95면 아래.

이미 기존에 명문대학으로 자리매김한 대학을 기준으로 하여 선별된 것이라는 점이다.[9] 이미 명문대학으로 자리 잡은 대학들에게 유리하게 대리 데이터들이 편향되어 있는 것이다. 그 점에서 대리 데이터는 대학을 평가하는 데 부정확할 뿐만 아니라 편향성마저 지닌다. 기존에 존재하는 대학 간 서열을 정당한 것으로 인정함으로써 합리적 이유 없이 다른 대학들을 차별하고 있는 것이다.

이처럼 정확하지 않을 뿐만 아니라 편향성을 지닌 대리 데이터로 대학을 평가하면서 다음과 같은 문제가 발생한다. "순위가 전국적인 표준으로 확장됨에 따라 부정적인 피드백 루프가 활성화되기 시작했다. 문제는 대학 순위가 자기 강화적인 특징을 갖는다는 점이었다."[10]

(3) 범죄 예측 프로그램

다음으로 범죄 예측 프로그램을 들 수 있다. 예산 압박에 시달리는 미국 각 주 및 시의 경찰 당국은 가능한 한 효율적으로 범죄를 억제하기 위해 범죄 예측 프로그램과 같은 알고리즘을 적극 이용한다. 이미 2013년에 미국 펜실베이니아 주에 자리한 소도시 레딩(Reading)의 경찰서는 캘리포니아 주 산타크루즈에 자리한 빅데이터 스타트업 '프레드폴'(PredPol)이 개발한 범죄 예측 프로그램을 도입하여 사용하고 있다. 프레드폴 프로그램은 레딩시의 범죄 통계 데이터를 토대로 하여 범죄 발생 가능성이 가장 높은 지역을 시간대별로 예측한다.[11] 레딩시 경찰은 이 프로그램을 도입한지 1년 만에 강도 사건이 23%나 감소했다고 발표하였다. 이렇게 범죄 예측 프로그램이 일견 성공을 거두자 미국 각 경찰은 프레드폴과 같은 범죄 예측 프로그램을 적극 도입하여 사용한다. 예를 들어 뉴욕시는 프레드폴과 비슷한 '컴스탯'(CompStat)

9) 캐시 오닐, 앞의 책, 109면.
10) 캐시 오닐, 앞의 책, 97-98면.
11) 캐시 오닐, 앞의 책, 149면.

이라는 범죄 예측 프로그램을, 필라델피아 경찰은 '헌치랩'(HunchLab)을 사용한다. 뉴욕시가 사용하는 컴스탯은 MS사와 공동으로 구축하였다. 컴스탯은 범죄경력과 자동차 번호, 911 전화내역, 6,000여 대의 CCTV에서 수집한 데이터를 바탕으로 하여 개발된 '영역감지시스템'(Domain Awareness System: DAS)을 이용한다.[12] 이를 이용하여 범죄를 예측하고 실시간 대응체계를 마련한다. 이를 활용하여 수사기간을 단축하고 업무 효율성을 제고하였다.[13] 범죄 예측 프로그램이 거둔 성과를 캐시 오닐은 다음과 같이 말한다.[14]

"프레드폴이 개발한 예측 프로그램은 오늘날 예산에 쪼들리는 미국 전역 경찰서에서 크게 환영받고 있다. 애틀란타, LA 등 다양한 지역의 경찰 당국이 범죄 예측 프로그램이 시간대별로 범죄 발생 가능성이 높다고 예측한 지역들에 경찰 인력을 집중적으로 배치한 덕분에 범죄율이 감소했다고 발표했다."

문제는 이러한 범죄 예측 프로그램이 범죄를 정확하고 공정하게 예측하지 못한다는 점이다. 대학을 평가할 때와 마찬가지로 범죄 예측 프로그램은 대리 데이터를 활용한다. 범죄에 관한 다양한 데이터를 분석함으로써 확률적·통계적으로 범죄를 예측하는 것이다. 이로 인해 범죄 예측 프로그램은 특정한, 가령 경제적으로 빈곤한 지역이나 유색인종 등에게서 범죄 발생 확률이 더 높게 나타난다고 예측하는 편향성을 보인다. 이로 인해 특정한 지역에 산다는 것만으로, 특정한 인종에 속한다는 것만으로 범죄자로 의심받고 취급받는 문제가 발생한다. 그리고 이는 악순환을 거쳐 스스로 편향성을 강화한다.[15]

---

12) DAS에 관해서는 E. S. Levine/Jessica Tisch/Anthony Tasso/Michael Joy, "The New York City Police Department's Domain Awareness System", in: Interfaces (Published online in Articles in Advance 18 Jan 2017) (http://dx.doi.org/10.1287/inte.2016.0860) 참고.

13) 김지혜, "범죄 예방 및 대응에서 AI의 역할", 「AI Trend Watch」 제13호, 정보통신정책연구원, 2020, 6면.

14) 캐시 오닐, 앞의 책, 149면.

15) 상세한 분석은 캐시 오닐, 앞의 책, 153면 아래 참고.

실제로 뉴욕시가 이용하는 컴스탯에 관해 개인의 사생활 침해 등과 같은 인권침해 논란이 발생하여 DAS에 관한 가이드라인을 제정하기도 하였다.[16]

(4) 채용 프로그램의 문제: 디지털 골상학

인재를 채용하는 영역 역시 범죄 예측과 더불어 알고리즘의 편향성이 문제되는 영역이다. 캐시 오닐에 따르면 미국에서 직원 채용 프로그램이 적극 사용되면서 여러 문제를 야기한다. 직원 채용 프로그램 역시 다양한 대리 데이터를 활용함으로써 효율성이라는 이름 아래 차별을 유발하는 것이다.[17] 예를 들어 미국의 특정 종합유통업체는 인적자원관리 회사 '크로노스'(Kronos)가 개발한 직원 채용 프로그램을 이용하여 채용을 할 때 인성적성검사를 실시하는데, 이 과정에서 미국 밴더빌트 대학교를 다니다가 정신건강 문제로 휴학을 하고 이후 회복한 지원자를 인성적성검사만으로 탈락시키기도 하였다. 지원자의 주장에 의하면 이미 정신건강을 회복했는데도 해당 종합유통업체가 사용한 직원 채용 프로그램은 지원자를 채용하기에 부적합한 인재로 평가한 것이다.[18]

문제는 이러한 직원 채용 프로그램 역시 해당 지원자와 직접 관련이 있는 데이터가 아닌 대리 데이터를 이용하여 확률적·통계적으로 지원자를 평가한다는 점이다. 요컨대 '간접적인 데이터'만으로 '효율성'이라는 이름 아래 다수의 지원자를 신속하게 평가하는 것이다. 이 과정에서 자연스럽게 정확하지 않은 평가뿐만 아니라 편향적인 평가 역시 이루어진다. 범죄 예측 프로그램과 마찬가지로 특정한 지역에 살거나 특정한 인종에 속하는 경우 직원 채용 프로그램 역시 편향적인 판단을 하는 것이다. 이렇게 상당수의 직원 채용

16) 김지혜, 앞의 보고서, 6면. 범죄 예측 프로그램의 문제를 분석하는 연구로는 이병규, "AI의 예측능력과 재범예측알고리즘의 헌법 문제: State v. Loomis 판결을 중심으로", 「공법학연구」 제21권 제2호, 한국비교공법학회, 2020, 169-191면; 최정일, "빅 데이터 분석을 기반으로 하는 첨단과학기법의 현황과 한계: 범죄예방과 수사의 측면에서", 「법학연구」 제20권 제1호, 한국법학회, 2020, 57-77면 등 참고.

17) 캐시 오닐, 앞의 책, 201면 아래.

18) 캐시 오닐, 앞의 책, 181면 아래.

알고리즘이 지원자의 직무 능력이 아닌 출신 지역이나 인종, 국적 등과 같은 요소로 편향적인 판단을 한다는 점에서 캐시 오닐은 이를 '디지털 골상학'으로 규정하기도 한다.[19] 19세기에 유행했던 사이비 과학인 골상학을 디지털 알고리즘이 다시 구현하고 있다는 것이다.

구글과 더불어 세계적인 플랫폼 기업인 아마존 역시 이와 유사한 문제를 일으켰다.[20] 다만 실제로 진행된 채용에서 문제가 된 것은 아니고 개발 중인 인공지능 채용 시스템에서 특정 집단에 편향적인 평가를 하는 문제가 발생하였다. 인공지능 채용 시스템이 여성 지원자를 차별하는 판단을 한 것이다. 이러한 문제가 발생한 이유는 아마존의 인공지능 채용 시스템에 제공한 데이터에서 찾을 수 있다. 아마존은 이미 채용된 직원들의 이력서를 데이터로 제공하였는데 이때 데이터에 존재하는 편향성이 그대로 인공지능의 알고리즘에 반영된 것이다. 현실 세계에 존재하는 편향성이 인공지능의 판단에 영향을 미친 것이다.

(5) 일정의 노예

캐시 오닐은 일정 관리 알고리즘이 어떻게 저임금 노동자들을 일정의 노예로 만드는지 흥미롭게 분석한다. 그 예로 한국인들이 사랑하는 '스타벅스' 사례를 언급한다.[21] 스타벅스가 직원을 효율적으로 배치하기 위해 일정 관리 알고리즘을 사용하면서 스타벅스 직원들은 저임금에 시달리면서 일정의 노예가 된다. 이른바 '클로프닝'(clopening)이 이들을 지배한다.[22] 이로 인해 다수의 스타벅스 직원들은 자신의 삶을 빼앗긴다. 일정 관리 알고리즘이 내

19) 캐시 오닐, 앞의 책, 206-207면 참고.

20) 박소정, ""이력서에 '여성' 들어가면 감점"…아마존 AI 채용, 도입 취소", 「조선일보」(2018. 10. 11)(https://www.chosun.com/site/data/html_dir/2018/10/11/2018101101250.html)(방문일자: 2020년 11월 15일 16시 3분).

21) 캐시 오닐, 앞의 책, 212면 아래.

22) '클로프닝'은 'closing'과 'opening'을 합성한 신조어로 퇴근하자마자 출근해야 하는 상황을 뜻한다. 캐시 오닐, 앞의 책, 208면.

리는 예측 불가능한 명령을 따라야 하기 때문이다.

이러한 문제는 최근 우리나라에서도 이슈가 된다. 코로나 19로 사회적 거리두기가 진행되면서 각종 앱을 이용한 주문 및 배송이 급속하게 증대하였는데 이로 인해 앱에 종속되는 배송 노동자들의 일상이 논란이 된다.[23] 인공지능에 기반을 둔 배송앱은 배송 노동자들에게 배송에 관해 지시를 내린다. 이때 특히 배송 시간이 문제가 된다. 인공지능 알고리즘이 계산한 배송 시간과 실제 배송 시간 사이에 큰 차이가 발생하는 경우가 많기 때문이다. 배송 노동자들은 지시된 배송 시간을 맞추기 위해 도로교통 법규를 위반하는 배송을 해야 한다. 무리한 일정을 거절하면 배송이 주어지지 않는다. 인간 배송 노동자들이 인공지능 알고리즘이 계산한 일정의 노예가 되는 것이다.

(6) 알고리즘 편향과 보험

알고리즘의 편향성은 보험에도 영향을 미친다.[24] 보험가입자를 차별하는 것이다. 이를테면 특정한 보험에 가입하려는 보험가입자가 보험사고를 자주 일으키는 집단에 속한다고 판단되면 이들에게는 가격이 비싼 보험 상품을 판매하는 것이다. 같은 보험 상품을 판매하는 경우에도 보험사고를 잘 일으키지 않는 '안전한 보험가입자'로 판단되는 경우에는 이들에게는 상대적으로 저렴하게 보험 상품을 판매하는 것이다. 캐시 오닐은 이를 다음과 같이 말한다.[25]

"2015년 미국의 비영리단체 컨슈머 리포트Consumer Reports는 자동차 보험료의 차이를 규명하기 위해 전국 차원의 광범위한 조사를 진행했다. 이를 위해 전국 3만3419개 우편번호별로 가상의 소비자를 만들어 미국의 모든 주

23) 김민제 · 선담은, "가라면 가? 25분 거리를 15분 안에 가라는 'AI 사장님'", 「한겨레」(2020. 10. 30.)(http://www.hani.co.kr/arti/society/labor/967865.html#csidx9f9e41d83405407bb0f5477240b4627)(방문일자: 2020년 11월 15일 16시 41분).

24) 캐시 오닐, 앞의 책, 268면 아래.

25) 캐시 오닐, 앞의 책, 273면.

요 보험사들에 견적서를 요청하고, 그들이 보내준 20억 장 이상의 견적서를 분석했다. 결과부터 말하면 보험사들의 보험료 산정 정책은 매우 불공정할뿐더러, (…) 신용평가 점수에 깊이 의존했다."

이렇게 보험가입자를 평가하고 선별할 때 보험사는 보험 관련 데이터에 기반을 둔 알고리즘을 사용한다. 하지만 이때도 알고리즘은 대리 데이터를 사용하고 이로 인해 보험가입자를 평가할 때 공정한 판단이 아닌 편향된 판단을 한다. 이러한 편향된 판단이 보험 상품 가격에도 영향을 미치는 것이다.

(7) 마이크로 타기팅 선거운동

알고리즘은 정치 영역, 특히 선거운동에서도 사용된다.[26] 이때 알고리즘의 편향성은 의도적으로 강화되어 사용된다. '마이크로 타기팅'(micro targeting)이라는 이름으로 말이다.[27] 여기서 마이크로 타기팅이란 선거운동을 할 때 유권자들을 의도적으로 구별하는 것을 말한다. 정치적으로 자신들에게 유리한 유권자들을 편향적으로 구별하여 이들에게 맞춤형 선거운동을 하는 것이다. 실제로 오바마 대통령 재선캠프의 데이터과학자 레이드 가니는 데이터 분석 전문가를 채용해 마이크로 타기팅을 적극 활용하는 선거운동을 하였다.[28] 이는 꽤 성공을 거두어 오바마가 재선하는 데 기여하였다.

마이크로 타기팅은 요즘 광고업계에서 즐겨 사용하는 '커스터마이징'(customizing) 광고와 유사하다. 각 유권자의 정치적 성향을 고려하여 이에 걸맞은 정치 광고를 내보내는 등과 같은 선별적 · 편향적 선거운동을 하는 것이다. 이러한 이유에서 정치 영역이나 광고 영역에서는 알고리즘의 편향성이 '마이크로 타기팅'이나 '커스터마이징'이라는 이름 아래 강화된다.

---

26) 캐시 오닐, 앞의 책, 298면 아래.
27) 캐시 오닐, 앞의 책, 313면.
28) 캐시 오닐, 앞의 책, 313면 아래.

(8) 커스터마이징 광고와 차별

마이크로 타기팅은 개별 유권자가 원하는 정치적 상품을 제공한다는 점에서 개별적인 차원에서는 큰 문제가 되지 않는다. 이에 반해 아마존이나 유튜브, 페이스북 등에서 즐겨 사용하는 커스터마이징 광고는 개별 소비자에게 이익만을 제공하는 것은 아니다. 이를테면 부자인 사람에게는 그에 적합한 부동산이나 여행 상품을 추천하면서 상대적으로 가난한 사람에게는 이러한 광고를 제공하지 않는 게 오히려 그 사람을 차별하는 것이 될 수 있기 때문이다.[29] 각 소비자에게 개별화된 광고는 소비자가 지닌 욕망을 비합리적으로 차별하는 문제를 야기할 수 있다. 소비자가 다원적으로 품을 수 있는 욕망을 억압하고 배제하는 문제가 될 수 있는 것이다.

(9) 프로파일링과 인격권 침해

개별 유권자가 소비자를 커스터마이징하려면 해당 정보주체를 각 영역에서 프로파일링할 수 있어야 한다. 그런데 이렇게 빅데이터와 알고리즘을 이용하여 개별 정보주체를 프로파일링하는 것은 해당 주체의 인격권을 침해하는 것이 될 수 있다. 프로파일링되는 정보주체의 입장에서 보면, 원하지 않는데도 자신의 사적인 영역 모든 것이 인공지능 사업자와 같은 상대방에게 노출되는 것으로 볼 수 있기 때문이다.

4. 알고리즘과 감시국가 문제

앞에서 알고리즘의 편향성 문제 가운데 하나로 범죄 예측 프로그램의 편향성을 언급하였다. 그런데 이렇게 알고리즘이 범죄 예측, 더 나아가 형사사법과 결합되면 자칫 '빅브라더'(big brother)와 같은 감시국가를 초래할 수 있다.[30] 이때 알고리즘은 감시국가 및 감시사회를 실현하는 유용한 도구가

29) 이에 관해서는 구정우, 「인권도 차별이 되나요?」, 북스톤, 2019 참고.

된다. 이를 독일의 범죄학자인 징엘른슈타인(Tobias Singelnstein)과 슈톨레(Peer Stolle)는 다음과 같이 인상 깊게 서술한다.[31]

"이러한 통제기술의 입장에서는 현대적 정보처리기술이 특별한 의미를 갖는다. 정보처리기술은 한편으로는 최대한 질서에 순응하여 행동하고 눈에 뜨이는 행동을 하지 않도록 함으로써 행위통제를 위해 투입된다. 다른 한편 정보처리기술은 위험을 통제하고 회피하는 데 기여한다. 이 기술을 통해 사람과 사실에 대해 포괄적인 데이터를 수집하고 평가하는 것이 가능하게 되고, 이들 데이터는 다시 예방적 개입이 필요한지 여부에 대한 예측결정을 하기 위한 토대가 된다. 모든 형태의 삶의 표현과 관련된 데이터를 조사, 처리, 저장하고 이러한 목적을 위해 설치된 데이터뱅크들을 상호 연결함으로써 이제는 모든 사람과 모든 상황을 탐지할 수 있는 총체적 능력을 갖추게 되었다."

5. 인공지능의 사회적 이용에 따른 위험 사례

그밖에 인공지능을 사회적으로 이용하면서 발생한 위험 사례들을 간략하게 소개한다. 가장 대표적인 경우로 자율주행차의 교통사고를 들 수 있다.[32] 현재 자율주행차 개발에서 선두 위치에 있는 테슬라나 구글 모두 자율주행차를 시험하면서 교통사고를 내기도 하였다. 물론 아직은 자율주행차가 실현되었다고 말할 수 없기에 지금까지 발생한 사고들은 자율주행차가 일으킨 사고로 말하기에는 어렵다. 그러나 앞으로 자율주행차가 실현된다 하더라도 오작동으로 교통사고가 발생할 가능성이 있다는 점을 염두에 둘 필요는 있다.

---

30) 양천수, 「빅데이터와 인권」, 영남대학교 출판부, 2016 참고.

31) 토비아스 징엘슈타인 · 피어 슈톨레, 윤재왕(역), 「안전사회: 21세기의 사회통제」, 한국형사정책연구원, 2012, 80면.

32) 이에 관해서는 김규옥 · 조선아, "자율주행차 사고유형으로부터의 시사점: 미국 캘리포니아 자율주행차 사고자료를 토대로", 「교통 기술과 정책」 제17권 제2호, 대한교통학회, 2020, 34-42면 참고.

인공지능이 기계학습을 하는 과정에서 혐오표현을 하거나 범죄와 유사한 행위를 저지른 경우도 발생하였다. 인공지능이 가짜뉴스를 생산해 퍼트리거나 '딥페이크'(deep fake)를 자행하는 경우도 들 수 있다.[33]

인공지능이 사회 각 영역에서 성공적으로 사용되면서 인간의 일자리가 점점 줄어드는 현상도 거론할 필요가 있다. 은행 업무를 예로 보면 은행 업무의 자동화가 가속화되면서 오프라인 은행 점포가 점점 줄어드는 현상을 들 수 있다. 반도체 생산과 같은 첨단 제조업 영역에서 자동화가 진척되면서 신규 일자리가 그다지 늘지 않는 것도 언급할 필요가 있다. 인공지능이 기존의 인간 일자리를 대체하고 있는 것이다.

Ⅲ. 인공지능 위험의 통제 방안

1. 고려 사항

인공지능 위험을 법이나 윤리와 같은 규범으로 통제할 때는 몇 가지 고려해야 하는 사항이 있다. 혁신과 안전, 포용이 그것이다. 이는 현 정부가 강조하는 '혁신적 포용국가'(innovative inclusive state)와 맥락을 같이 한다.[34] 이외에 인공지능 위험의 대응 개념인 안전 역시 고려해야 한다.

먼저 인공지능이 주도하는 '혁신'(innovation)을 고려해야 한다. 인공지능 위험을 통제할 때는 가능한 한 인공지능이 주축이 되는 혁신을 저해하지 않도록 해야 한다. 다음으로 인공지능 위험에 대한 '안전'(safety)을 고려해야 한다. 인공지능 위험이 실현되어 사람들의 권리나 사회적 공리 등을 훼손하

33) 이에 관해서는 홍태석, "딥페이크 이용 아동성착취물 제작자의 형사책임: 일본의 판례 및 논의 검토를 통하여",「디지털 포렌식 연구」 제14권 제2호, 한국디지털포렌식학회, 2020, 139-151면 참고.

34) 포용국가에 관해서는 Anis A. Dani/Arjan de Haan, Inclusive States: Social Policy and Structural Inequalities (World Bank, 2008) 참고.

지 않도록 해야 한다. 나아가 '포용'(inclusion)을 고려해야 한다. 인공지능이 주도하는 혁신에서 배제되는 사람들이 없도록 해야 하고 혹시라도 배제되는 이들이 있는 경우에는 이들이 사회적 영역으로 포용될 수 있도록 해야 한다.

2. 법을 통한 통제

인공지능 위험을 통제하는 가장 대표적인 수단으로 법적 규제를 생각할 수 있다. 법적 규제는 강제적이고 고정적이며 사후적인 규제라는 성격을 지닌다. 이러한 법적 규제는 단기적인 측면에서 실효성을 확보하는 데 유리하다. 그렇지만 다음과 같은 문제도 지닌다. 법적 규제가 야기하는 '규제의 역설'이 그것이다. 이로 인해 법적 규제는 인공지능이 주도하는 혁신을 저해하는 중대한 장애물이 될 수 있다.

3. 윤리를 통한 통제

법적 규제 이외에 인공지능 위험을 통제하는 규범적 수단으로 윤리를 고려할 수 있다. 윤리적 규제는 자율적이며 유동적이고 사전적인 규제라는 특징을 지닌다. 이러한 윤리적 규제는 다음과 같은 장점을 지닌다. '연성 규제'(soft regulation)로서 수범자의 자율규제를 유도한다는 것이다. 이를 통해 혁신에 친화적인 인공지능 위험 통제를 도모할 수 있다.

Ⅳ. 유네스코 인공지능 윤리 권고안

지난 2020년에 최초 버전이 나온 「유네스코 AI 윤리 권고안(초안)」(이하 '초안'으로 약칭함)은 그동안 'AI 윤리'에 관해 전개된 논의 상황에 비추어볼 때 상당히 의미 있는 결과물로 평가할 수 있다.[35] 이 초안을 일별해 보면 다

음과 같은 인상을 받는다. 초안은 지금까지 축적된 논의 성과를 종합적으로 그러면서도 섬세하게 집적하고 있다는 점이다. 철학적 · 윤리적으로 쟁점이 될 수 있는 부분을 섬세하게 고려하면서 그리고 이 초안이 유엔 차원에서 제시하는 일종의 '프레임워크'(기본 작업)라는 점을 감안하면서 실제로 실행 가능한 요청들을 담아내고 있다. 뿐만 아니라 그동안 유엔(유네스코)이 윤리 및 인권영역 등에서 논의하고 성과를 도출한 규범적 내용 역시 집약적으로 담고 있다. 이러한 예로 생태주의적 사고, 젠더 평등, 다양성 및 포용, 윤리경영 및 인권경영, 적응적 · 진화적 사고, 이해관계자 중심주의, 세대 간의 정의, 교육에 대한 강조 등을 언급할 수 있다.[36] 그 점에서 이 초안이야말로 AI 윤리에 관해 그동안 축적된 여러 논의를 섬세하면서도 체계적으로 종합한 뛰어난 지적 산물이라고 평가할 수 있다. 이를 아래에서 개관하면서 그중 몇 가지를 중점적으로 검토한다.

1. 구성

(1) 개요

초안은 다음과 같이 구성된다. 서문, 적용 범위, 목적 및 목표, 가치 및 원칙, 정책 과제 영역, 모니터링 및 평가, 현재 권고안의 활용, 현 권고안의 홍보-장려, 최종 조항이 그것이다. 이 중에서 특히 중요한 것으로 가치와 원칙 그리고 정책 과제를 꼽을 수 있다.

(2) 가치

35) 이 초안이 수정된 버전은 (https://unesdoc.unesco.org/ark:/48223/pf0000376713)에서 확인할 수 있다(방문일자: 2021년 4월 22일). 이 글은 2020년에 제시된 초안을 분석 대상으로 하여 논의를 전개한다.

36) 이 중에서 인권경영에 관해서는 양천수, "인권경영을 둘러싼 이론적 쟁점", 「법철학연구」 제17권 제1호, 한국법철학회, 2014, 159-188면 참고.

초안은 인공지능 윤리가 추구해야 하는 가치로 다음을 언급한다. 인간의 존엄성, 인권 및 근본적 자유, 소외된 사람이 없도록, 조화로운 삶, 신뢰 가능성, 환경 보호가 그것이다.

(3) 원칙

초안은 원칙을 두 가지로 구별한다. 그룹 1과 그룹 2가 그것이다.

먼저 그룹 1로 다음과 같은 원칙을 제시한다. 인간과 인간의 번영을 위해, 비례성, 인간의 관리 감독 및 결정, 지속 가능성, 다양성 및 포용성, 개인정보보호, 인식 및 교육, 다중-이해관계자 및 적응형 거버넌스가 그것이다.

다음 그룹 2로 다음과 같은 원칙을 제시한다. 공정성, 투명성 및 설명 가능성, 안전 및 보안, 책임(responsibility) 및 책무성(accountability)이 그것이다.

(4) 과제 목표 및 정책 과제

초안은 5개의 과제 목표와 11개의 정책 과제를 제시한다.

과제 목표 I은 윤리적 책무를 규정한다. 과제 목표 I 아래에 정책 과제 1을 제시한다. 이는 다양성 및 포용성 증진을 과제로 설정한다.

과제 목표 II는 영향 평가를 규정한다. 과제 목표 II는 3개의 정책 과제를 제시한다. 정책 과제 2는 시장 변화에 대한 대응을, 정책 과제 3은 AI의 사회적 및 경제적 영향에 대한 대응을, 정책 과제 4는 문화와 환경에 미치는 영향을 과제로 설정한다.

과제 목표 III은 AI 윤리 역량 구축을 규정한다. 과제 목표 III은 2개의 정책 과제를 제시한다. 정책 과제 5는 AI 윤리 교육 및 인식의 증진을, 정책 과제 6은 AI 윤리 연구 장려를 과제로 설정한다.

과제 목표 IV는 (경제)개발 및 국제협력을 규정한다. 과제 목표 IV는 2개의 정책 과제를 제시한다. 정책 과제 7은 (경제)개발 분야에서 AI의 윤리적 활용

증진을, 정책 과제 8은 AI 윤리에 대한 국제협력 증진을 과제로 설정한다.

과제 목표 V는 AI 윤리를 위한 거버넌스를 규정한다. 과제 목표 V는 3개의 정책 과제를 제시한다. 정책 과제 9는 AI 윤리를 위한 거버넌스 메커니즘 확립을, 정책 과제 10은 AI 체계의 신뢰성 보장을, 정책 과제 11은 책임성, 책무성 및 사생활 보호를 과제로 설정한다.

2. 분석

이러한 초안에서 주목할 만한 점을 선별해 분석하면 다음과 같다.

(1) 가치와 원칙 구별

이론적인 면에서 볼 때 눈에 띄는 점은 초안이 가치와 원칙을 구별한다는 점이다. 가치와 원칙의 관계를 어떻게 설정할 것인지에 관해서는 철학이나 윤리학, 법철학 등에서 다양한 논의가 이루어진다. 이에 관해 초안은 꽤 설득력 있게, 아마도 실행 가능한 AI 윤리 원칙을 확립하기 위해 가치와 원칙을 개념적으로 구별한다. 그러면서 인간의 존엄성이나 조화로운 삶처럼 추상도가 높은 개념들은 가치로 설정하고 이보다 그 의미 내용이 좀 더 명확한 개념들은 원칙으로 설정한다. 물론 개별 원칙들을 일별하면 이러한 구별 방향이 언제나 일관되게 적용되는 것인지에는 의문이 들 수 있다.

(2) 원칙의 유형화

AI 윤리 원칙을 두 가지로 유형화하는 것도 눈에 띈다. 첫 번째 유형은 “인간-AI 체계의 상호작용과 관련된 특성을 반영하는” 원칙이다. 두 번째 유형은 “AI 체계 자체의 속성과 관련된 특성을 반영하는” 원칙이다. 첫 번째 유형의 원칙이 AI의 사회적 의미와 영향을 규율하는 원칙이라면, 두 번째 유형의 원칙은 AI 자체의 기술적 속성을 규율하는 원칙이라고 바꾸어 말할 수

있다. 이러한 유형화는 AI 윤리 원칙이 상징적 · 선언적인 의미만 갖는 것으로 그치는 것이 아니라 실제로 적용 가능한 원칙이 될 수 있도록 하는 데 유용하다고 판단된다.

(3) 인간중심적 사고와 생태주의적 사고의 결합

초안은 여타의 AI 윤리처럼 인간중심적 사고를 원칙으로 삼는다. AI는 자족적인 존재가 아니라 인간을 위해 존재하는 도구라는 것이다. 그러면서도 초안은 인간중심적 사고가 야기할 수 있는 문제를 해소하고자 "조화로운 삶"과 "환경 보호"를 가치로 포섭한다. 이 점에서 초안은 단순히 인간중심적 사고에만 머무는 것이 아니라 인간과 자연 환경의 조화를 중시하는 생태주의적 사고 역시 수용한다고 말할 수 있다.

(4) 책임주체로서 AI 행위자

법학자의 관점에서 볼 때 가장 눈에 띄는 점은 초안이 "AI 체계"와 "AI 행위자"를 구별한다는 점이다. 그러면서 AI 윤리의 책임자는 AI 체계가 아닌 AI 행위자가 되어야 한다고 강조한다. 더불어 "새로운 규제 프레임워크를 개발할 경우 정부는 사람 또는 법인에게 책임성과 책무성을 부여해야 한다는 점을 염두에 두어야 한다. AI 체계에게 책임을 물게 하거나 AI 체계에게 법적 지위를 부여해서는 안 된다."고 정한다(정책과제 11 중에서 94번). 이는 상당히 의미가 있으면서도 논란을 야기할 것으로 보인다. 왜냐하면 AI에게 법적 책임을 물을 수 있는가에 관해 다양한 논의가 전개되는 상황에서 초안은 이를 명백하게 부정하기 때문이다. 물론 필자는 초안의 태도에 동의한다.[37]

나아가 책임주체를 AI 행위자로 강조하는 초안의 태도에서 AI 윤리의 실

37) 이 문제에 관해서는 조성은 외, 「인공지능시대 법제 대응과 사회적 수용성」, 정보통신정책연구원, 2018 참고.

질적인 수범자는 AI 체계 자체가 아니라 AI 행위자라는 점을 시사한다.

(5) 다양한 이해관계자를 고려하는 거버넌스

초안은 AI 윤리 거버넌스에 관해서도 의미 있는 제안을 한다. 두 가지를 언급할 수 있다. 첫째, 기업의 사회적 책임에 관한 논의에서 발전한 '이해관계자 중심주의'를 수용하고 있다는 것이다. 이에 따라 거버넌스를 구축할 때 다양한 이해관계자들을 고려할 것을 강조한다. 둘째, 적응적 · 진화적 사고를 수용하고 있다는 것이다. 미국의 법사회학자 노넷과 셀즈닉이 제시한 '응답적 법'(responsive law)으로 거슬러 올라가는 적응적 · 진화적 사고는 거버넌스가 사회변화에 적극 대응해야 한다고 강조한다.[38]

(6) 절차주의적 · 혁신적 사고

이의 연장선상에서 초안은 자연스럽게 절차주의적 사고와 혁신적 사고를 받아들인다.[39] 이를테면 정책과제 11 중에서 96번은 다음과 같이 말한다.

"회원국은 AI 생태계의 모든 행위자(시민사회, 법집행, 보험사, 투자자, 제조업체, 엔지니어, 변호사, 사용자의 대표 등을 포함)가 새로운 규범을 제정하는 과정에 참여하도록 해야 한다. 이러한 규범은 모범수칙과 법으로 진화할 수 있다. 또한 회원국은 규제 샌드박스(Regulatory Sandbox)와 같은 메커니즘의 사용을 장려하여 급격한 신기술 개발에 발맞추어 법 및 정책 개발을 가속화하고 법이 공식적으로 채택되기 전에 안전한 환경에서 테스트될 수 있도록 보장해야 한다."

---

38) 응답적 법에 관해서는 양천수, "새로운 법진화론의 가능성", 「법철학연구」 제15권 제2호, 한국법철학회, 2012, 163-202면 참고.

39) 이에 관해서는 양천수, "제4차 산업혁명과 규제형식의 진화", 「경제규제와 법」 제12권 제2호, 서울대학교 공익산업법센터, 2019, 154-172면 참고.

(7) 실질적인 윤리 원칙

초안을 전체적으로 분석하면 AI 개발자 등에게 직접적으로 부과되는 원칙은 비교적 간소한 편임을 알 수 있다. 두 번째 유형의 원칙이 여기에 해당한다. 이에 따르면 AI 개발자 등은 공정성, 투명성 및 설명 가능성, 안전 및 보안, 책임 및 책무성을 준수할 수 있도록 AI 체계를 개발하고 운용해야 한다. 그동안 제시된 AI 윤리 원칙 등과 비교할 때 초안이 제시하는 원칙은 꽤 간결하고 그 때문에 실제로 적용될 가능성도 높다고 판단된다.

3. 평가

초안은 AI 윤리에 관해 화려한 수사는 지양하고 실제로 실행 가능한 원칙을 간결하게 제시한다. 오히려 AI를 운용하는 AI 행위자에게 다양한 윤리적 의무를 부과하는 것에 비중을 둔다. 이를 통해 AI 개발이 규제라는 장벽에 부딪히지 않도록 고려한다. 이러한 초안의 태도는 필자의 AI 윤리에 관해 주장하는 바와 기본적으로 일치한다. 필자 역시 AI 윤리는 현재의 기술발전 상황을 고려하여 실제로 적용될 수 있는 것을 중심으로 간결하게 마련되어야 한다고 주장한 바 있기 때문이다.[40] 따라서 현재로서는 초안에서 무엇을 더 보완해야 할지 뚜렷하게 보이지는 않는 편이다. 다만 다음과 같은 점은 비판적으로 고려할 수 있다.

(1) 가치와 원칙의 구별

초안은 '가치'와 '원칙'을 구별한다. 여기에 나름 설득력을 부여할 수 있다. 그렇지만 가치와 원칙의 관계에 관해서는 철학, 윤리학 및 법철학 등에

40) 이를 보여주는 양천수, "인공지능과 윤리: 법철학의 관점에서", 「법학논총」 제27집 제1호, 조선대학교 법학연구원, 2020, 73-114면; 선지원 외, 「지능정보기술 발전에 따른 법제 · 윤리 개선방향 연구」, 정보통신기획평가원, 2019 등 참고.

서 다양한 견해가 제시된다. 과연 이렇게 양자를 명확하게 구별할 수 있는지, 이렇게 구별하는 게 바람직한지 의문이 제기될 수 있다. 따라서 필자는 가치와 원칙을 구별하지 말고 모두 원칙으로 통합해서 규율하는 것이 더 낫다고 판단한다. 가치와 원칙을 가령 '일반 원칙' 및 '개별 원칙'과 같은 방식으로 규율하는 게 더 적절하지 않을까 생각한다.

(2) 개인정보보호

원칙 그룹 1에서 제31번은 개인정보보호를 규율한다. 그런데 개인정보보호만을 강조한 나머지 이에 대한 예외는 규율하지 않는다. 인공지능을 개발하고 운용하기 위해서는 빅데이터가 필요하다는 점을 고려할 때 한편으로는 개인정보보호를 강조할 필요가 있지만, 다른 한편으로는 이에 대한 합리적 예외 역시 마련할 필요가 있다.

(3) 다양성 및 포용성

정책과제 1은 다양성 및 포용성 증진을 강조한다. 이러한 일환에서 인공지능에 대한 국제적 논의와 협력 등을 강조한다. 그러나 이와 함께 강조해야 할 부분은 인공지능에 대한 지식재산권을 존중할 필요가 있다는 것이다. 인공지능 기술에 대한 독점을 이유로 하여 인공지능 개발에 많은 노력을 기울여 획득한 지식재산에 대한 권리를 형해화하는 것도 막아야 한다.[41]

제45번은 "문화적 및 사회적 정형화"를 공개 및 방지해야 한다고 말한다. 다양성을 강조하는 유엔의 견지에서 보면 이는 당연한 주장으로 평가된다. 그렇지만 각 국가가 지닌 문화적 독자성도 존중할 필요가 있다. 이는 "국가는 전통문화의 계승·발전과 민족문화의 창달에 노력하여야 한다."고 규정하는 헌법 제9조와 충돌할 수 있다. 따라서 이를 완화할 필요가 있다.

41) 이에 관해서는 정책과제 3번 참고.

(4) AI의 사회적 및 경제적 영향에 대한 대응

제54번은 AI 기술의 독점과 이에 관한 불평등을 방지하기 위한 메커니즘을 개발할 것을 강조한다. 그렇지만 앞에서 언급하였듯이 이를 이유로 하여 AI 기술에 대한 지식재산권을 형해화하는 것은 막아야 한다. 이에 관한 내용을 추가할 필요가 있다.

제57번은 AI 체계에 대한 인증체계 도입을 권고한다. 그렇지만 현재 인공지능 기술이 발전하는 상황을 고려할 때 이러한 내용을 윤리에 담는 것은 강한 규제로 보일 수 있다. 설사 이 규정을 유지한다 하더라도 대략적인 내용만을 담는 것이 더욱 적절하다. 이를 상세하게 규율하는 것은 지양해야 한다. 그게 아니면 인공지능 윤리와 AI 체계에 대한 인증체계를 결합하는 방식을 고려할 수 있다. AI 체계에 대한 인증체계를 강제하는 것이 아니라 사전적 · 자율적으로 실시하게끔 하는 것이다.

제58번은 "AI 윤리 책임자 · 담당관"을 규율한다. 그러나 이 역시 강한 규제 또는 중복 규제가 될 수 있다. 우리 법제도에 이미 존재하는 준법감시인 또는 윤리경영 담당자와 통합하는 것을 모색할 수 있다. 여하간 이러한 내용을 담는 것은 현재로서는 AI 관련 기업에 큰 부담이 될 수 있다.

(5) 책임성 등

제94번은 AI 체계에게 법적 지위를 부여해서는 안 된다고 말한다. 그렇지만 경우에 따라서는, 이를테면 AI가 저작한 저작물이라는 것을 명시하기 위해 AI에게 법적 지위를 인정할 필요가 있다. 제94번의 취지는 이해가 되지만 이에 대한 적절한 예외를 설정할 필요가 있다.

제95번은 AI 체계의 위험에 대한 영향평가 제도 도입을 권고한다. 그러나 이미 인증제도 도입을 권고하면서 더불어 영향평가 제도 도입을 장려하는 것은 중복 규제이자 강한 규제로 보일 수 있다.[42] AI의 위험을 평가한다는 것은

현재로서는 막연하면서도 자의적일 수 있다는 점을 고려할 때 이는 유보할 필요가 있다. 제95번의 내용은 규제 혁신을 강조하는 제96번과 모순될 수 있다.

Ⅴ. 국가 인공지능(AI) 윤리 기준

지난 2020년 11월에 과학기술정보통신부가 주축이 되어 발표한 「국가 인공지능(AI) 윤리 기준」은 우리 정부가 공식적으로 제시한 인공지능 윤리라는 점에서 의미가 있다. 이 기준은 처음에는 '초안' 형태로 제시되었고(이하 '초안'으로 약칭한다), 이후 공론장에서 다양한 의견을 수렴하고 반영하는 반성적 절차를 거치면서 수정안으로 개선 및 발표되었다(이하 '수정안'으로 약칭한다). 아래에서는 초안과 수정안의 주요 내용을 개괄적으로 검토한다.

1. 초안

(1) 기본 구조

초안은 다음과 같이 구성된다. 최상위의 목표 또는 비전으로 '인간성'(AI For Humanity)을 제시한다. 이어서 4대 속성, 3대 기본원칙, 15개 실행원칙으로 체계화된다. 4대 속성은 4단(端)으로, 3대 기본원칙은 3강(綱)으로, 15개 실행원칙은 15륜(倫)으로 지칭되기도 한다. 인공지능 윤리를 동양철학의 유교윤리와 연결한 것이다.[43]

(2) 4대 속성

4단으로도 지칭되는 4대 속성으로 다음이 제시된다. 인권 보장, 공공선 증

42) 다만 인증제도와 영향평가 제도를 결합하는 것은 고려할 수 있다. 영향평가를 인증의 요건으로 설정하는 것이다.

43) 다만 이에는 여러 비판이 제기되었다.

진, 인간 능력의 향상, 기술 윤리적 좋음이 그것이다.

(3) 3대 기본원칙과 15개 실행원칙

3강으로도 지칭되는 3대 기본원칙으로 인간의 존엄성 원칙, 사회의 공공성 원칙, AI의 목적성 원칙이 제시된다. 이는 철학적으로 자유주의, 공동체주의, 공리주의를 반영한 것이다. 각 기본원칙에는 5개의 실행원칙이 배치된다.

먼저 인간의 존엄성 원칙에는 행복추구 원칙, 인권보장 원칙, 개인정보보호 원칙, 다양성 존중 원칙, 해악금지 원칙이 실행원칙으로 배치된다.

다음으로 사회의 공공선 원칙에는 공공성 원칙, 개방성 원칙, 연대성 원칙, 포용성 원칙, 데이터 관리 원칙이 실행원칙으로 배치된다.

나아가 AI의 목적성 원칙에는 책임성 원칙, 통제성 원칙, 안전성 원칙, 투명성 원칙, 견고성 원칙이 실행원칙으로 배치된다.

(4) 분석 및 평가

4대 속성(4端), 3대 기본 원칙(3綱), 15대 실행원칙(15倫)으로 구조화된 초안은 그동안 윤리학 영역 등에서 축적된 성과를 체계적으로 잘 반영한다. 더불어 그동안 발표된 세계 각국의 AI 윤리 역시 종합하고 있는 것으로 평가된다. 이 점에서 초안에 긍정적인 평가를 할 수 있다.

다만 기본 구조에 약간의 의문이 있다. 굳이 4대 속성이 필요할까 하는 점이다. 3대 기본원칙과 15대 실행원칙만으로도 윤리기준이 무엇을 추구하는지가 분명히 드러난다. 4대 속성, 3대 기본 원칙, 15대 실행원칙은 선거공약이나 각종 발전계획에서 많이 사용되는 도식적인 구조로도 보인다. 좀 더 간명하게 만들 필요가 있다. 더불어 요즘 세대, 특히 인공지능 개발자들에게는 생소한 4端이나 3綱, 15倫과 같은 개념을 사용할 필요가 있을지 의문이 든다.

구체적으로 다음과 같은 점을 지적할 수 있다. 첫째, 인간의 존엄 원칙의

실행원칙에 관해 언급할 필요가 있다. 행복추구 원칙의 경우 행복이라는 개념이 너무 포괄적이고 모호해서 헌법학에서도 행복추구권이 독자적인 기본권이 될 수 있는지에 논란이 있다.[44] 이 점에서 행복추구 원칙을 실행원칙으로 규정하는 게 적절한지 의문이 든다. 또한 인권보장 원칙과 해악금지 원칙은 동전의 양면에 해당하는 것으로 같은 내용을 규율한다. 타자의 인권을 충실히 보장하면 해악이 발생하지 않는다. '해악의 원칙'(harm principle)을 제시한 밀(John Stuart Mill)에 따르면 해악금지란 타인의 권리를 침해하지 말라는 의미를 지닌다.[45] 그 점에서 해악금지 원칙과 인권보장 원칙을 병존시킬 필요가 있을지 의문이 든다. 오히려 해악금지 원칙을 살리고 싶다면 사회의 공공선 원칙의 실행원칙으로 배치하는 게 더 적절해 보인다.

둘째, 사회의 공공선 원칙의 실행원칙인 연대성 원칙과 포용성 원칙을 별도로 규정하는 것에도 의문이 있다.[46] 양자의 내용은 거의 같은 것이 아닌가 한다. 연대의 핵심은 나와 타자를 단절시키지 않고 서로가 서로를 포용하는 것이라고 볼 수 있기 때문이다. 요즘 포용국가 논의로 포용성이 화두가 되는데 우리나라에서 언급되는 포용성은 종전에 있던 사회복지국가의 연대성과 차이가 없어 보인다. 우리가 말하는 포용성은 포용국가(inclusive state)가 본래 의미하는 난민과 같은 타자를 포용한다는 의미와는 차이가 있어 보이기 때문이다.[47]

셋째, 국제협력원칙이나 원칙 간의 충돌이 발생하였을 경우 이를 해결할 수 있는 (헌법학에서 개발된) 실제적 조화 원칙 등을 신설할 필요가 있다.

넷째, 윤리기준의 성격을 명확하게 할 필요가 있다. 윤리기준이 상징적인

---

44) 이 문제에 관해서는 허영, 「한국헌법론」 전정17판, 박영사, 2021 참고.

45) 존 스튜어트 밀, 서병훈(옮김), 「자유론」, 책세상, 2018 참고.

46) 물론 엄밀하게 말하면 연대성 원칙과 포용성 원칙은 맥락을 달리한다. 가령 연대성 원칙은 특정한 공동체를 전제로 하는 해당 공동체 구성원들 사이에서 문제가 되는 원칙이라면, 포용성 원칙은 특정한 공동체에 포함되는 구성원들과 배제되는 비구성원들 사이에서 문제가 되는 원칙이라 말할 수 있다.

47) 이에 관해서는 양천수(편), 「코로나 시대의 법과 철학」, 박영사, 2021, 제10장 참고.

원칙으로 자리매김하는 데 만족하는지 그게 아니면 구체적인 영역에서 실행 가능한 원칙을 목표로 하는지를 명확하게 할 필요가 있다. 만약 후자를 지향한다면 이번에 제시되는 윤리기준은 추상적인 원칙이 많아 더욱 다듬어야 할 필요가 있다. 추상적인 것은 최소화할 필요가 있다. 그렇게 하지 않으면 인공지능 개발자에게 혼선을 야기할 수 있다. 만약 상징적인 원칙을 지향한다면 이 점을 분명하게 할 필요가 있어 보인다.

2. 수정안

(1) 기본 구조

수정안은 초안에 대한 다양한 의견을 수렴하고 반영하면서 더욱 간명해졌다. 인간성(AI For Humanity)을 가장 높은 목표로 설정하는 것은 동일하다. 다만 4대 속성을 없애고 대신 3대 기본원칙 및 10대 핵심 요건으로 구조를 단순화했다.

(2) 3대 기본원칙

3대 기본원칙으로 인간 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙이 제시된다. 초안의 3대 기본원칙이 그대로 유지되었다.

(3) 10대 요건

3대 기본원칙을 실행하는 10대 요건으로 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안정성, 투명성이 제시된다. 초안의 15대 실행원칙이 수정안에서는 10대 요건으로 정리되었다.

(4) 평가

「국가 인공지능(AI) 윤리 기준」은 현재까지 도달한 윤리 · 철학 · 인공지능

의 성과를 집약하고 있다고 평가할 수 있다(이하 '윤리 기준'으로 약칭함). 이는 크게 네 가지 측면에서 살펴볼 수 있다. 이론적 측면, 실천적 측면, 상징적 측면, 절차적 측면이 그것이다. 먼저 이론적 측면에서 보면 윤리 기준은 당대 도달한 이론적 수준과 성과를 집약하고 있다. 다음으로 실천적 측면에서 보면 윤리 기준은 실무 현장에서 사용할 수 있는 윤리 원칙 및 요건을 중심으로 설계되어 있다. 나아가 상징적 측면에서 보면 윤리 기준은 국가가 인공지능 위험 문제에 관해 정면에서 관심을 보이고 있음을 상징적으로 표현한다. 인공지능 윤리 기준을 과연 국가가 제정할 필요가 있는지에 의문이 제기되기도 하지만 인공지능의 위험이 사회 전체적으로 미치는 영향을 고려할 때, 특히 국가의 기본권 보호의무와 관련하여 이는 의미가 없지 않다. 마지막으로 절차적인 측면에서 보면 윤리 기준을 제정하는 과정 자체는 반성적 절차를 충실하게 이행한 모범적인 사례라 말할 수 있다.

Ⅵ. 맺음말

맺음말로 인공지능 윤리가 짊어져야 하는 과제가 무엇인지 언급한다. 인공지능 윤리에 관해 앞으로 수행해야 하는 과제는 현재의 기준을 더욱 구체화한 가이드라인을 제정하는 것이다. 현재 제시된 윤리 기준은 실천적인 측면에서 유용하기는 하지만 여전히 추상적인 부분을 담고 있다. 따라서 이는 실무적으로 적용할 수 있도록 더욱 구체화해야 한다. 이 과정에서 다음을 고려해야 한다. 우선 공적 영역과 사적 영역을 구별하여 인공지능 윤리를 구체화해야 할 필요가 있다. 그러나 이러한 구별만으로는 매우 전문화된 현대사회의 각 영역에 걸맞은 인공지능 윤리를 만들 수 없다. 따라서 기능적으로 분화된 사회의 각 영역에 적합하게 인공지능 윤리를 구체화해야 할 필요가 있다.

〈참고문헌〉

구정우,「인권도 차별이 되나요?」, 북스톤, 2019.
맹준영,「자율주행자동차와 법적책임」, 박영사, 2020.
선지원 외,「지능정보기술 발전에 따른 법제 · 윤리 개선방향 연구」, 정보통신기획평가원, 2019.
양천수,「빅데이터와 인권」, 영남대학교 출판부, 2016.
_____(편),「코로나 시대의 법과 철학」, 박영사, 2021.
조성은 외,「인공지능시대 법제 대응과 사회적 수용성」, 정보통신정책연구원, 2018.
허 영,「한국헌법론」 전정17판, 박영사, 2021.
니클라스 루만, 윤재왕(옮김),「체계이론 입문」, 새물결, 2014.
제리 카플란, 신동숙(옮김),「인간은 필요 없다」, 한스 미디어, 2016.
존 스튜어트 밀, 서병훈(옮김),「자유론」, 책세상, 2018.
캐시 오닐, 김정혜(역),「대량살상 수학무기」, 흐름출판, 2017.
크리스토퍼 스타이너, 박지유(옮김),「알고리즘으로 세상을 지배하라」, 에이콘, 2016.
토비아스 징엘슈타인 · 피어 슈톨레, 윤재왕(역),「안전사회: 21세기의 사회통제」, 한국형사정책연구원, 2012.

김건우, "차별에서 공정성으로: 인공지능의 차별 완화와 공정성 제고를 위한 제도적 방안",「법학연구」 제61집, 전북대학교 법학연구소, 2019.
김규옥 · 조선아, "자율주행차 사고유형으로부터의 시사점: 미국 캘리포니아 자율주행차 사고자료를 토대로",「교통 기술과 정책」 제17권 제2호, 대한교통학회, 2020.
김서안, "데이터 3법 개정의 의미와 추후 과제",「융합보안 논문지」 제20권 제2호, 한국융합보안학회, 2020.
김지혜, "범죄 예방 및 대응에서 AI의 역할",「AI Trend Watch」 제13호, 정보통신정책연구원, 2020.

양천수, “새로운 법진화론의 가능성”, 「법철학연구」 제15권 제2호, 한국법철학회, 2012.

_____, “인권경영을 둘러싼 이론적 쟁점”, 「법철학연구」 제17권 제1호, 한국법철학회, 2014.

_____, “현대 안전사회와 법적 통제: 형사법을 예로 하여”, 「안암법학」 제49호, 안암법학회, 2016.

_____, “제4차 산업혁명과 규제형식의 진화”, 「경제규제와 법」 제12권 제2호, 서울대학교 공익산업법센터, 2019.

_____, “인공지능과 윤리: 법철학의 관점에서”, 「법학논총」 제27집 제1호, 조선대학교 법학연구원, 2020.

이병규, “AI의 예측능력과 재범예측알고리즘의 헌법 문제: State v. Loomis 판결을 중심으로”, 「공법학연구」 제21권 제2호, 한국비교공법학회, 2020.

이부하, “알고리즘(Algorithm)에 대한 법적 문제와 법적 규율”, 「과학기술과 법」 제9권 제2호, 충북대학교 법학연구소, 2018.

이상경, “미국의 개인정보보호 입법체계와 현황에 관한 일고”, 「세계헌법연구」 제18권 제2호, 세계헌법학회 한국학회, 2012.

최정일, “빅 데이터 분석을 기반으로 하는 첨단과학기법의 현황과 한계: 범죄예방과 수사의 측면에서”, 「법학연구」 제20권 제1호, 한국법학회, 2020.

홍태석, “딥페이크 이용 아동성착취물 제작자의 형사책임: 일본의 판례 및 논의 검토를 통하여”, 「디지털 포렌식 연구」 제14권 제2호, 한국디지털포렌식학회, 2020.

김민제 · 선담은, “가라면 가? 25분 거리를 15분 안에 가라는 ‘AI 사장님’”, 「한겨레」, 2020. 10. 30.

박소정, “이력서에 ‘여성’ 들어가면 감점”…아마존 AI 채용, 도입 취소”, 「조선일보」, 2018. 10. 11.

大貫恵美子,「人殺しの花: 政治空間における象徴的コミュニケ-ションの不透明性」, 岩波書店, 2020.

Anis A. Dani/Arjan de Haan, *Inclusive States: Social Policy and Structural Inequalities* (World Bank, 2008).

E. S. Levine/Jessica Tisch/Anthony Tasso/Michael Joy, "The New York City Police Department's Domain Awareness System", in: *Interfaces* (Published online in Articles in Advance 18 Jan 2017) (http://dx.doi.org/10.1287/inte.2016.0860).

【국문초록】

인공지능 윤리의 현황과 과제

양 천 수*

이 글은 인공지능이 유발하는 사회적 · 법적 문제를 해결하는 방안에 초점을 맞춘다. 그중에서도 윤리라는 규범으로 인공지능 문제를 해결하고자 하는 시도를 검토한다. 오늘날 인공지능이 유발하는 가장 어려운 규범적 문제 중 하나로 '알고리즘의 편향성' 문제를 들 수 있다. 편향성 문제는 차별금지 원칙을 위반한다는 점에서 중대한 문제로 볼 수 있다. 그렇지만 이 글은 현재 상황에서 인공지능 편향성 문제를 법으로 규제하는 것이 적절한지 의문을 제기한다. 이 글은 현재로서는 인공지능이 유발하는 규범적 문제를 윤리로 규제하는 것이 더욱 바람직하다고 주장한다. 이러한 문제 상황에서 이 글은 인공지능이 유발하는 규범적 문제를 윤리로 대응하고자 하는 논의를 살펴본다.

주제어: 인공지능 윤리, 인공지능의 위험성, 알고리즘의 편향성, 유네스코 인공지능 윤리 권고안, 국가 인공지능 윤리 기준

* 영남대학교 법학전문대학원 교수 · 법학박사.

【ABSTRACT】

The Current Situation and Challenges of the Artificial Intelligence Ethics

Chun-Soo Yang*

This article focuses on solutions to social and legal problems caused by artificial intelligence. Among them, attempts to solve artificial intelligence problems with the norm of ethics are reviewed. One of the most difficult normative problems caused by artificial intelligence today is the problem of 'algorithm bias'. The problem of this bias can be seen as a serious problem because it violates the anti-discrimination principle. However, this article raises the question of whether it is appropriate to regulate the AI bias problem with law in the current situation. This article argues that it is more desirable to regulate the normative problems caused by artificial intelligence with ethics at this time. In such a situation, this article examines the discussion of ethically responding to the normative problems caused by artificial intelligence.

Keywords : artificial intelligence ethics, artificial intelligence risks, algorithm bias, UNESCO Artificial Intelligence Ethics Recommendation, Korean National Artificial Intelligence Ethics Standard Recommendation

* Professor at Yeungnam University Law School · Dr. jur.