인공지능 알고리즘의 편향성과 공정성

정원섭*

【요약】

인공지능 기술이 다방면으로 활용되면서 그 효율성에도 불구하고 편향성에 대한 논쟁이 국내외에서 격화하고 있다. 국내 온라인 시장을 대표하는 업체가 알고리즘에 대한 인위적인 조작 혐의로 거액의 과징금을 부여받았는가 하면, 미국 의회는 구글 등 4대 디지털 기업에 대해 불공정 행위를 기소하였다. 제 1장에서는 인공지능 편향성과 관련된 주요 논쟁들이 지닌 특징을 소개하고자 한다. 제 2장에서는 인공지능 알고리즘에서 차별과 편향성이 등장하는 이유와 유형들을 데이터 마이닝의 특성과 연관지어 살펴보고자 한다. 3장에서는 컴퓨터 기술의 비가시성과 논리적 변용성의 관점에서 알고리즘 편향성에 대해 살펴보았다. 4장에서는 인공지능 기술 응용과정에서 등장하는 편향성과 관련하여 인공지능 전문가들의 역할 및 시민사회의 공동 대응 방안을 제시하고자 하였다.

【주제어】 인공지능, 편향성, 알고리즘, 공정성, 차별, 데이터

---

* 경남대학교 자유전공학부 교수

** 이 연구결과물은 2019학년도 경남대학교 신진교수연구비 지원에 의한 것임
https://doi.org/10.34162/hefins.2020..25.003

I. 인공지능 편향성 논쟁

2020년 10월 6일 공정거래위원회에서는 네이버(주)에 대해 자사의 검색알고리즘을 인위적으로 조정·변경한 것에 대한 시정 명령과 함께 약 267억의 과징금을 부여하였다.[1] 공교롭게도 같은 날 미국 하원에서도 아마존, 애플, 페이스북, 구글 등 4대 온라인 업체에 대해 디지털 시장에서 지배력 남용을 비판하는 보고서를 발표하였다.[2] 우리나라 디지털 시장을 사실상 독점하고 있는 네이버는 이미 뉴스 배치의 편향성과 관련하여 지속적으로 논쟁의 대상이 되어왔다는 점에서 이번 사태는 디지털 거대 기업의 공정성에 대해 시사하는 바가 적지 아니하다.

기업 내부자가 인위적으로 조정함으로써 시장에서 불공정성을 초래하였다는 것은 공정 경쟁이라는 시장의 기본 원칙을 위반한 것이라는 점에서 그 해결 방향은 어느 정도 뚜렷해 보인다. 공정거래위원회의 판단이 사실이라면 이것은 인공지능 알고리즘을 이용과 관련된 문제일 뿐, 알고리즘 자체의 편향성 문제라고 할 수는 없을 것이다. 따라서 이 경우 인위적이며 의도적인 개입 내지 횡포라는 점에서 위법 행위자를 처벌하고 전문직 윤리를 강조하거나 내부 고발자를 보호하면서 시장에서 공정 경쟁을 보장하는 각종 법률과 제도를 정비하는 등 상당 부분 전통적인 방식으로 해결할 수 있는 길을 모색할 수 있을 듯하기 때문이다. 그럼에도 불구하고 알고리즘에 대해 인위적

---

1) 물론 네이버(주)는 즉각 반발하였지만, 공정거래위원회는 이 사건을 “네이버가 자신의 검색알고리즘을 조정·변경하여 부당하게 검색결과 노출 순위를 조정함으로써 검색결과가 객관적이라고 믿는 소비자를 기만하고 오픈마켓 시장과 동영상 플랫폼 시장의 경쟁을 왜곡한 사건”으로 규정하였다. 공정거래위원회 (2020), “부당하게 자사 서비스를 우선 노출한 네이버 쇼핑·동영상 제재- 온라인 플랫폼 사업자가 검색알고리즘을 조정·변경해 자사 서비스를 우대한 행위를 제재한 최초 사례.”, https://www.ftc.go.kr/www/selectReportUserView.do?key=10&rpttype=1&report_data_no=8759.

2) Jerrold Nadler, et al. (2020).

조정이 가능하다는 점에서 최근 부각되고 있는 인공지능 기술의 편향성에 대한 국내외 논란과 연동되는 측면이 적지 아니하다.

이미 우리는 2016년 3월, 마이크로소프트사의 챗봇 테이(Tay)로부터 전혀 다른 방식으로 큰 충격을 받았다. 빅데이터를 통해 학습된 것으로 알려진 테이는 출시 후 하루도 되지 않아 여성, 흑인, 유태인 등 사회적 약자에 대해 거침없이 온갖 혐오 발언을 하였다. 테이로부터 충격이 더욱 컸던 것은 바로 직전 구글 딥마인드Google Deep Mind의 알파고가 이세돌과의 대국에서 보여준 엄청난 위력을 목격하면서 자연인의 능력을 완전히 압도하는 인공지능이 반사회적 행동을 서슴없이 할 수 있는 가능성을 목격하면서 우리는 현재의 사회 질서뿐만 아니라 생존 자체가 위협받을 수 있다는 두려움을 가질 수밖에 없었기 때문이었다.

더욱이 테이가 이런 발언을 하게 된 것은 일부 트위터 사용자들이 테이를 특정 방향으로 적극적으로 학습시킨 결과라는 점이 알려지면서 이런 두려움은 다음과 같이 두 가지 형태로 구체적으로 부각되었다. 첫째 두려움은 인공지능 기술이 특정 집단에 의해 해킹을 당해 조직적으로 악용될 가능성이다. 최근 대규모 자동차 회사들이 무인자율주행자동차 상용화와 관련된 일정을 앞다투어 발표하고 있는 상황에서 테이가 특정 방향으로 학습당했다는 소식은 인공지능 기술에 대한 안전성에 대한 두려움을 더욱 심화하였다. 뿐만 아니라 인공지능 기술이 킬러 로봇 등 군사용 살상 무기로 전환될 수 있다는 점에서 인공지능 윤리 및 거버넌스에 대한 논의를 폭발시키며 이와 관련된 다양한 가이드라인이 국내외에서 봇물터진듯 쏟아지고 있다.[3)]

둘째 두려움은 바로 기계학습 기술, 더 나아가 인공지능을 통해 기존의 이러저런 사회적 편견이나 고정 관념이 완화 내지 제거되는 것이 아니라

3) 2018년 인공지능 기술을 선도하고 있는 세계 50여 명의 저명 인사들은 카이스트의 인공지능연구가 킬러 로봇 연구로 이어질 수 있다는 우려를 제시하면서 카이스트와의 공동연구를 거부하는 선언을 발표하였다. 임세정・이재연, 「카이스트發 'AI 킬러로봇' 논란... 세계 로봇학자들 '보이콧' 왜?」, 『국민일보』, 2018.04.06.

오히려 정당화 내지 악화될 수도 있다는 점이다. 사실 새로운 기술은 기존의 골치 아픈 문제를 해결하면서 대체로 환대받아 왔다. 가령 원자력 발전소는 적어도 도입초기에는 화력 발전소와 달리 저비용 친환경 기술로 환영받았다. 그러나 원자력 발전기술에서 보듯 어떤 기술이 사회 전반으로 확산되면 도입초기 생각하지 못한 문제들이 부각되기 마련이다. 인공지능 기술 역시 마찬가지이다. 알파고가 바둑에서 보여 준 엄청난 위력은, 다양한 해석이 있을 수 있지만, 근본적으로 연산능력에 기반하고 있다. 이에 비해 우리의 언어는 연산 능력뿐만 아니라 '지금 여기'라는 특정한 사회 문화적 맥락에서 지속적으로 새롭게 태어난다. 테이 사건은 인공지능 기술을 바둑과 같은 단순 연산 분야를 넘어서 언어라고 하는 고도의 "사회적 영역"으로 확장할 경우 등장할 수 있는 등장할 수 있는 새로운 문제들을 잘 보여 주고 있다. 왜냐하면 테이가 학습당한 언어는 지금까지 인류가 가지고 있었던 온갖 편견이나 혐오 혹은 갈등을 무비판적으로 그대로 담고 있기 때문이다.

더욱이 "사회적 영역"에서 인간의 결정은 이해관계의 상충, 문화적 상대성, 정보의 한계, 판단력 부족, 가치관의 차이 등등으로 인하여 설령 선의지를 바탕으로 최선을 다하더라도 동일 사안에 대해서도 사람마다 다를 수 있다. 뿐만 아니라 판단의 부담으로 인해 경우에 따라 오류가 있을 수 있다.[4] 그러나 기술, 특히 인공지능 기술의 경우 이해관계의 상충이나 문화적 상대성 혹은 특정 가치관과 상관없이 결정할 수 있을 것이라는 막연한 기대를 할 수도 있다. 뿐만 아니라 어떤 사안은 고려할 요소가 너무나 많고 복잡하여 토론을 통한 사회적 합의에 맡길 수 없는 경우도 비일비재하다. 그래서 사회적으로 논란이 되는 복잡한 사안에 대해서 인공지능 기술에 의한 결정이 자연인의 결정보다 더 공정할 것이라고 기대하면서 골치아픈 결정을 인공지능에게 위임하고자 하는 유혹에 더욱 솔깃할 수 있다.

인공지능이라는 말이 오늘날처럼 널리 퍼지기 이미 오래전부터 세금,

4) John Rawls (1993), pp. 36-37; John Rawls (1993), pp. 55-57.

대출, 치안, 입시 등 다양한 분야에서 그 분야 업무를 잘 수행할 수 있는 여러 종류의 소프트웨어가 활용되어 왔으며 또한 축적된 데이터를 바탕으로 더욱 향상된 결정을 제시하는 수많은 알고리즘이 활용되고 있다. 급기야 지난 여름 코로나 사태로 인해 고등학교 졸업시험을 치룰 수 없게 된 영국 정부는 인공지능 기술을 활용하여 성적을 부여하고자 하였다.[5] 인공지능이 예측한 바에 대해 주로 서민층인 공립학교 학생들이 부유층 자녀인 사립학교 학생들에 비해 불리하다는 주장이 제기되면서 이것은 하나의 해프닝이 되고 말았다. 그러나 이런 결과가 "전국적인 차원에서는 공정하다고 주장할 수 있으나, 개인별로는 공정함을 완전히 상실한 것"이라는 옥스퍼드 컴퓨터 공학과 선임연구원인 헬레나 웹의 발언은 인공지능 기술 활용과 관련하여 분명 또 하나의 중요한 쟁점을 시사한다.[6]

이 글에서는 인공지능 차별 및 편향성과 관련하여 현재 논쟁의 초점이 되고 있는 주요 사례들의 이런 특성을 바탕으로 다음과 같은 논의를 전개하고자 한다. 우선 제 2장에서는 인공지능 편향성이 등장하는 이유와 유형들을 데이터 마이닝의 특성과 연관지어 살펴보고자 한다. 3장에서는 컴퓨터 기술의 비가시성과 논리적 변용성의 관점에서 알고리즘 편향성에 대해 살펴 볼 것이다. 그리고 4장에서는 인공지능 기술 응용과정에서 등장하는 편향성과 관련하여 인공지능 전문가들의 역할 및 사회의 공동 대응 방안을 제시하고자 한다.

---

5) "How an AI grading system ignited a national controversy in the U.K.", https://www.axios.com/england-exams-algorithm-grading-4f728465-a3bf-476b-9127-9df036525c22.html

6) "Algorithms can drive inequality. Just look at Britain's school exam chaos", https://edition.cnn.com/2020/08/23/tech/algorithms-bias-inequality-intl-gbr/index.html. 그러나 디지털 정보 남용 방지 단체인 '폭스글러브'의 창립자인 코리 크라이더는 "영국의 A-레벨 시험은 빙산의 일각"이라며 알고리즘은 사용된 원자료에서 발견된 편향을 복제한다고 지적했다. 하지만 그는 알고리즘 기술에 책임을 돌릴 일이 아니라고 경고했다. 그는 "기술적인 문제라고 말하는 어떤 사람도 거짓말을 하는 것"이라며 "영국 고교 학점 사태는 학점 인플레를 막기 위한 정치적인 선택이었지, 기술적인 문제가 아니었다."고 지적했다.

II. 데이터 마이닝에서 편향성

수학과 컴퓨터 과학에서 알고리즘이란 “어떤 문제를 해결하기 위해 명확히 정의된 유한 숫자의 규칙과 절차의 모임. 즉 명확히 정의된 한정된 개수의 규제나 명령의 집합으로서 한정된 규칙을 적용함으로써 문제를 해결하는 것”[7]을 말한다. 다시 말해 특정 문제를 해결하거나 연산을 수행하도록 컴퓨터가 작동하는 일련의 한정된 단계들의 집합이라고 할 수 있다. 인공지능 알고리즘은 그 목적에 맞게 다듬어진 데이터를 기반으로 자동화된 추론automated reasoning이라고 할 수 있다.[8] 그래서 인공지능을 통해 추론된 결과는 이 알고리즘과 입력된 데이터에 의존할 수밖에 없기 때문에 GIGO, 즉 “쓰레기를 넣으면 쓰레기가 나온다. Garbage in, garbage out.”는 표현이 상식처럼 간주된다. 인공지능이 제시한 결과가 편향적이라고 한다면 그 원인은 알고리즘이나 데이터 혹은 이 양자 모두에 있을 수밖에 없다.

최근 뉴럴 네트워크neural network과 딥러닝deep learning과 같은 기계학습 기술은 폭발적으로 발전하고 있다. 또한 데이터 마이닝을 통해 거의 모든 분야에서 빅데이터가 하루가 다르게 쌓여가고 있다. 이와 더불어 엄청난 컴퓨팅 파워가 맞물리면서 인공지능은 적용 영역을 세금이나 대출 업무 등 수리적 분야를 넘어 면접이나 치안, 재판 등 훨씬 민감한 사회적 사안으로 확장하면서 정확성과 효율성 뿐만 아니라 공정성까지 담보할 수 있을 것이라는 기대를 받고 있다.

그러나 구글에서 이미지 처리를 하며 흑인을 고릴라로 분류하자[9] 인공지

---

7) 컴퓨터.인터넷.IT용어대사전 (2005)

8) “The Definitive Glossary of Higher Mathematical Jargon — Algorithm”. Math Vault. https://mathvault.ca/math-glossary/#algo.

9) “Google apologises for Photos app's racist blunder”, https://www.bbc.com/news/technology-33347866.

능이 정확성에 대한 의문은 말할 것도 없거니와 인종 차별을 하고 있다는 비난을 받으면서 해당 서비스는 중단되었다. 이와 함께 인공지능 판사라고 일컬어지는 콤파스 알고리즘에 대해서도 문제 제기가 본격화되었다. 미국 법원과 교도소에서 형량, 가석방, 보석 등의 판결에 널리 사용되던 컴파스 COMPAS 알고리즘이 흑인들에게 불리하게 판단하는 주장이 등장하면서 이와 관련 논쟁이 지금도 심각하게 진행 중이다.[10]

그러나 이와 반대로 모기지 승인과정에서 인간을 통한 거래를 줄이거나 아예 없애면서 유색인종, 독신여성, 동성애 부부 등 전통적으로 집을 사기 힘들었던 계층에게 유리한 방향으로 대출 승인이 증가하였다는 주장도 있다. 2016년 발족하여 현재 미국 44개 주에서 운영되고 있는 디지털 대출 전문 사이트 베터닷콤Better.com에 따르면 대출이 30-40대 히스패닉 고객들의 경우 532%, 동일 연령대 흑인들의 경우 411% 확장되었다. 또한 결혼한 동성애자들의 경우에도 10배 이상 확장되었다. 피부색이나 성적 취향 때문에 전통적인 모기지 승인 면접 심사 과정에서 의식적으로 혹은 무의식적으로 피할 수 없었던 불쾌함과 무기력함을 느꼈던 고객들이 이제 좋은 이율로 내 집을 마련할 수 있게 된 것이다. 대출 과정이 상당 부분 자동화되었지만 대출 여부를 판단하는 문지기는 여전히 있다. 그러나 그 판단을 인간이 아니라 숫자가 하면서 대출 기회가 다양한 계층으로 널리 확대되었다는 것이다.[11]

인공지능이 불편부당한 결정을 내릴 수 있기 위해서는 알고리즘 자체가 공정하게 구성되어야 할 것이다. 지금 네이버가 의심받는 것처럼 알고리즘에 대한 인위적 조작이 있어서는 물론 안 될 것이다. 뿐만 아니라 초기 입력값, 즉 입력 데이터가 공정하게 선별되어야 하지만 데이터를 정제하는 다음과 같은 네 단계의 과정에서 자의성arbitrariness이 개입할 수 있다.

---

10) 오요한, 홍성욱 (2018) 참고.

11) “Is an Algorithm Less Racist Than a Loan Officer?”, https://www.nytimes.com/2020/09/18/business/digital-mortgages.html?fbclid=IwAR0gacBq0oOps0yIoRoMuuoRa9fUyuyml9tZBkBGyJAVsUqb4kFaVlzqp-4

첫째 목표 변수를 정의하는 과정에서 특징 집단이 과잉 혹은 과소 대표되거나 아예 배제될 수도 있다. 이런 경우 표본 편향과 배제 편향으로 인해 인공지능은 공정성과 정확성 모두에서 위기에 처하게 된다. 둘째 데이터를 수집하여 다듬고 레이블링labeling하는 과정에서 적절한 평가를 하지 못하여 편향성이 발생할 수 있다. 흔히 측정에서 문제가 발생하고 이것이 시정되지 못한 결과 측정 편향과 회상 편향이 발생하면 인공지은 전혀 다른 결론으로 나아갈 수 있다. 셋째 특징 선택feature selection 단계이다. 인종적 편향이 등장하는 것은 바로 이 단계라고 할 수 있다. 여러 특징들을 상호 비교 연결함으로써 예상하지 못한 정보나 혹은 노출되어서는 안 되는 개인 신상 정보가 나타날 수도 있다. 이런 정보들이 마지막으로 모델을 바탕으로 의사결정을 하는 단계에서 특정한 방향으로 결론을 유도할 수도 있다. 이처럼 각각의 단계에서 다양한 편향이 개입될 수 있다.[12)]

뿐만 아니라 빅데이터 내의 수많은 상호 관계로 인해 수집이 금지된 혹은 전혀 새로운 종류의 정보를 추론할 수 있다는 점이다. 그리고 이렇게 추론된 정보는 차별이나 편향의 근거로 이용될 수 있다는 다음과 같은 이유에서 나올 수 있다는 점이다. 빅데이터에 존재하는 수많은 정보가 개별로 존재할 경우 큰 의미를 찾기 힘들지만 이런 정보가 상호 연결될 경우 차별의 근거로 사용될 수 있기 때문이다. 이미 널리 알려진 사실이지만 미국 월마트 빅데이터 전문가들은 고객의 25가지 구매 행태를 분석하면 여성의 임신과 출산을 상당히 정확하게 예측할 수 있다는 사실을 확인했다. 예를 들어 향이 나는 로션을 사던 여성이 무향의 로션으로 바꾸거나, 평소 사지 않던 미네랄 영양제를 갑자기 사들이는 경우다. 개별 구매만으로는 임신 여부를 알 수 없지만

12) 기계 학습과 관련된 데이터 편향은 분류 방식에 따라 매우 다양한 방식으로 제시되고 있다. 가령 다음과 같은 7가지 편향이 제시되기도 한다. 표본 편향sample bias, 배제편향(exclusion bias), 측정 편향measurement bias, 회상 편향recall bias, 관찰자 편향observer bias, 인종편향racial bias, 결합 편향association bias. “7 Types of Data Bias in Machine Learning”, https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/

소비 패턴을 다른 사람들과 비교 분석함으로써 임신이라는 전혀 의도하지 않은 예측할 수 있었던 것이다.[13]

이와 마찬가지로 현재 정부나 공기업의 채용 과정에서 공정성을 증진하고 개인의 사생활 보호를 위해 가족 관계, 경제형편, 성적 취향 등에 대해 직접적인 질문을 엄격히 금지하지만 여전히 유관 간접 질문을 통해 이런 정보들을 충분히 추론할 수 있다. 그리고 이렇게 수집된 정보는 의사 결정과정의 공정성을 훼손할 수 있다. 여기서 우리는 정보의 양이 증가할 경우 판단의 정확성은 향상될 수 있을지언정 그 공정성이 함께 증진하기 힘들다는 점을 인식할 수 있다.

여기서 우리는 현대의 대표적인 정치 철학자인 롤즈의 기획에 주목할 필요가 있다. 그는 사회 전체를 위한 정의의 원칙에 대한 합의를 할 수 있기 위해서는 먼저 그 합의 당사자들이 공정한 입장에 있어야 한다는 점을 강조한다. 이를 위해 그는 의사결정과정의 공정성에 부정적 영향을 줄 수 있는 다음과 같은 두 가지 종류의 우연들을 배제하고자 한다. 첫째 재능, 체력, 용모 등등 선천적 우연이며, 둘째는 재산, 계급, 환경 등 사회적 우연이다. 적어도 인간은 자신의 특수한 사정을 알면 알수록 자기에게 이로운 결정을 하기 마련이며 그 결과 공정한 결정을 기대하기 어렵다는 점이다.[14] 공정성을 기하기 위해 '역지사지', 즉 입장을 서로 바꾸어 생각해 보고자 하는 것도 이런 개인적 특수성을 배제하고자 하는 것이다. 이런 점에서 볼 때 인공지능의 경우 차별의 가능성을 차단하여 공정성을 기하기 위해 제거하고자 했던 민감한 개인 신상 정보들을 다양한 데이터 마이닝을 통해 뚜렷하게 드러나게 한다는 점이다. 따라서 데이터 마이닝 작업 그 자체가 차별을 초래하는 것은 아니지만 악용되어 차별로 이어질 수 있는 가능성을 배제할 수는 없다.

---

13) "부모도 모르는 딸의 임신, 대형마트는 알고 있다",
http://www.hani.co.kr/arti/economy/economy_general/729868.html#csidxc9258ae5e702bb8b717e1883a218873

14) Rawls, John (1971), pp. 118-123.

III. 알고리즘의 편향성

2017년 인공지능 전문가들을 대상으로 한 설문 조사에서 인공지능 윤리 문제에서 가장 심각한 사안으로 63퍼센트가 “인종적 편견이나 특정 종교적 입장이 프로그램 되는 것”이라고 응답했다. 이 문제는 해결될 수 있을까? 해결 가능하다면 어떻게 해결할 수 있을까? 불가능하다면 왜 불가능하며, 우리는 어떻게 해야 하는가? 그 답을 찾기 위해서는 두 가지 요인을 검토해볼 필요가 있다.

첫째, 인공지능 알고리즘의 특성
둘째, 인공지능 전문가의 특성

인공지능 알고리즘은 컴퓨팅의 논리적 변용성logical malleability에 기초한다. 논리적 변용성이란 음성이나 텍스트 혹은 영상처럼 과거 아날로그 정보처리 과정에서는 서로 이질적인 것으로 간주되던 정보들을 0과 1과 같은 가장 단순한 단위로 환원하여 처리한 후 상이한 양상으로 재현하는 것을 의미한다.[15] 논리적 변용성 덕분에 컴퓨터 연산에서는 온갖 종류의 정보가 상호 융합한다. 오늘날 여러 가지 이질적인 활동을 통합적으로 수행할 수 있는 소위 범용인공지능artificial general intelligence을 기대하는 것 역시 이러한 논리적 변용성에서 출발한다.

논리적 변용성을 통해 성취한 탁월한 성과 덕분에 복잡한 논쟁이 지속되고 있는 윤리와 도덕 그리고 종교 분야 사안들에 대해서도 인공지능이 간결한 답을 구할 수 있을 것이라는 기대가 나타난다. 그러나 여기서 선결문제는 “인공지능이 윤리나 예술 등 가치와 연관된 영역에서 어떤 역할을 할 수

15) James Moor (1985), p. 269.

있는가?"다. 이러한 영역에서 인공지능을 어떻게 활용할 것인지 결정해야 하는 것이다. 그리고 인공지능을 어디까지 활용할지 정하는 이용자의 태도가 중요한 것이다.

인공지능을 구성하는 알고리즘과 관련된 윤리의 핵심은 알고리즘을 구성하는 전문가 윤리다. 인공지능 알고리즘은 알고리즘 전문가에 의해 구성되고 운영되고 변용되며 또한 그렇게 될 수 있기 때문이다. 오늘날처럼 기능적으로 세분화된 상황에서 전문가들은 그 고유한 전문 영역에서 전문 지식과 장인 정신을 바탕으로 성실성과 정직성과 같은 덕목을 실천해 주기를 희망한다. 일반인들이 전문가의 결정을 존중하는 것은 그들이 전문 지식을 선의로 활용할 것이라는 기대 아니 적어도 의도적으로 악용하지는 않을 것이라는 사회적 바람이 광범위하게 존재하기 때문이다.

그런데 알고리즘이 작동하는 컴퓨터 연산 과정은 근본적으로 비가시성 invisibility을 지닌다는 점에서 인공지능 윤리에서 전문가 윤리는 특히 중요하다. 즉 어떤 과정을 거쳐 그런 결론에 이르게 되었는지 일반인으로서는 도저히 알 수 없는 경우가 비일비재하다. 그래서 알고리즘 윤리는 곧 알고리즘 전문가의 도덕성을 바탕으로 출발한다. 투명성이 인공지능 윤리에서 대두되는 것은 바로 이런 비가시성 때문이다. 그러나 이때 투명성은 조직의 투명성처럼 회계 장부를 공개하거나 회의록을 공개하는 것으로 달성될 수는 없다. 알고리즘 자체를 공개한다고 하더라도 그 알고리즘이 워낙 방대하여 그것을 가시적으로 확인할 수 없기 때문이다.[16]

그 결과 알고리즘 자체가 아니라 알고리즘을 구성하는 주요한 원칙을 공개하도록 하는 간접적인 방식으로 접근할 수밖에 없다. 나아가 이러한 원칙이 제대로 구현되었다는 것을 확인하는 절차로 알고리즘이 실제 작동하였을 경우를 상정하여 다양한 시뮬레이션을 제시하고 이에 대한 인증 절차를 거치는 방법이 현실적일 것이다.[17] 그리고 실제로 알고리즘이 작동되었을

16) James Moor (1985), p. 272.

때 등장할 상황에 대해 엄격한 책임을 묻도록 함으로써 비가시성을 부분적으로 해소할 수 있을 것이다.

연산 과정의 비가시성은 두 가지 더욱 심각한 문제를 파생한다. 첫째, 의도적 악용의 문제이다. 여기서 우리는 두 가지 상황을 다시 구분할 필요가 있다. 우선 전문가 개인의 윤리 문제다. 전문가가 개인적 차원에서 알고리즘을 악용하지 않도록 다양한 안전장치와 적절한 교육이 요구된다. 히포크라테스 선서와 같은 의료인들의 전문직 윤리 강령이나 1980년에 발표된 〈미국 컴퓨터장비협회ACM의 행동규약〉은 좋은 사례라고 할 것이다. 이러한 전문직 분야 윤리 강령을 통해 해당 분야 전문직의 종사자들에 대한 사회적 신뢰를 강화할 수 있을 것이다.

사실 전문직 윤리와 관련하여 더 중요한 문제는 오늘날 전문가들이 단순히 개인으로 활동하는 것이 아니라 기업이나 조직의 일원으로서 일하고 있다는 점이다. 니부어가 잘 지적하고 있듯이 윤리적인 개인이라고 할지라도 집단의 일원이 되는 순간 집단의 요구로 인해 윤리의식이 무감각해지거나 비윤리적 행위를 강제 받을 수도 있다.

> 모든 인간의 집단은 개인과 비교할 때 충동을 올바르게 인도하고 때에 따라 억제할 수 있는 이성과 자기 극복 능력, 그리고 다른 사람들의 욕구를 수용하는 능력이 훨씬 결여되어 있다. 게다가 집단을 구성하는 개인들이 개인적 관계에서 보여주는 것에 비해 훨씬 심한 이기주의가 모든 집단에서 나타난다.[18]

전문가들을 그들이 소속된 집단의 이기주의적 행위로부터 보호하기 위해서는 소속된 집단의 기존의 관행에 연루되지 않도록 보호할 필요가 있다. 서구의 경우 1980년대 이후 기업윤리에 대한 다양한 논의를 거쳐 적어도 기업 활동이 합법적 틀 안에서 이루어져야 한다는 점은 분명히 한 듯하다.

---

17) David J. Dougall (1997), pp. 305-311.

18) 라이놀드 니부어, 이한우 옮김 (2017), p. 10.

기업 윤리에 대한 논의 과정에서 얻은 중요한 성과는 사회적 공헌 활동이나 기업 내 윤리 전담 기구 설치와 같은 적극적 조치뿐만 아니라 내부고발이라는 안전장치를 적극적으로 수용하는 기업 문화가 형성되었다는 점이다. 물론 내부고발이 적절한 것으로 승인되기 위해서는 여러 요건이 충족되어야 하지만, 내부고발의 필요성이 인정된 것 자체가 하나의 큰 진전이라고 할 수 있다. 내부고발은 집단주의의 오랜 전통 등 특수한 사정으로 말미암아 우리 사회에서 매우 부정적으로 간주되고 있기는 하지만 긍정적으로 고려할 필요가 있다.

비가시성이 지닌 더욱 근본적 문제는 윤리 자체의 특성에 있다. 일상의 윤리적 관념들은 지금까지 누적된 그 시대의 소산이라는 점에서 그 시대적 한계 내에 있다. 가령 오늘날 보편적 가치로 간주되는 인간의 존엄성 역시 서양 근대 이후 일반화된 역사적 산물이며, 일부 동물해방운동가들의 경우 종차별주의에 불과하다는 비난을 받고 있다. 만일 일부 공상과학영화에서 볼 수 있는 것처럼 인간의 모습으로 인간과 친밀한 정서적 교류를 지속적으로 할 수 있는 로봇에 대해 호모 사피엔스가 아니라는 이유로 존엄성을 부정할 수 있을까? 그렇다고 다른 인간과 동일하게 존엄한 존재로 간주해야 할까? 머지않은 장래에 인간의 존엄성이란 윤리적 가치가 인종차별주의처럼 폐기될 가능성이 전혀 없는 것은 아니다. 그럼에도 지금 여기서 인간존엄성은 윤리적 사고를 전개하는 과정에서 아르키메데스의 점일 수밖에 없다.

그런데 문제는 윤리가 지금까지 역사적 경험을 바탕으로 하지만 그 역사적 경험을 넘어서고자 한다는 점이다. 가령 지금까지 인류가 생산한 모든 텍스트를 빅데이터로 축적한 채팅 로봇을 생각해보자. 이런 채팅 로봇은 어쩌면 욕설이나 인종차별적인 발언들을 더 자주 할지도 모른다. 따라서 인공지능에 구축될 알고리즘은 단순히 과거 자료를 집적하는 것을 넘어 현재 우리 문화 속에서 가령 자유와 평등과 박애처럼 수용되고 있는 가치들을 정합적으로 반영해야 한다. 물론 이러한 가치들을 정합적으로 반영할 수 있는 알고리즘이 쉽게 구현될 수는 없을 것이다. 이런 점에서 인공지능 알고리즘은 지속적으로

보완되고 수정될 수 있도록 개방적으로 구성되어야 한다. 알고리즘 구성 원칙에서 언급되는 개입 가능성이란 바로 이러한 개방성을 의미한다고 할 수 있다. 또한 어떤 가치들이 어떻게 반영되어 있는지에 대한 설명의 요구 역시 수용될 수 있어야 한다. 인공지능 알고리즘의 설명 가능성이란 단순히 기능적 효율성에 대한 설명 가능성뿐만 아니라 이러한 문화적 요소에 대한 것까지 포괄하는 것으로 이해되어야 한다. 인공지능 윤리는 단순히 지금까지 축적된 자료와 결과를 바탕으로 확보된 빅데이터를 넘어 이를 바탕으로 미래 지향적 가치들에 대한 성찰을 기반으로 현재에 대해 끊임없이 개입하는 과정을 거쳐야만 하는 것이다. 이 점에서 알고리즘을 구성하는 전문가들은 우리 시대정신과 무관한 단순한 기능인이 아니라 당면한 윤리적 요구에 민감해야 할 뿐만 아니라 장차 지향하고자 하는 가치에 대한 예민한 문제의식을 공유하는 자율적인 건강한 민주시민으로 자리매김해야 한다.

현재까지 발표된 인공지능 윤리와 관련된 문건들 대부분은 그 표현상의 사소한 차이에도 불구하고 대체로 인공지능을 제작하고 운용하는 전문가들에 대한 이러한 사회적 기대를 반영하는 것이었다. 다만 국내에서 최근 생산되고 있는 일부 문건의 경우 주요 원칙을 밝히는 것을 넘어 지나치게 엄밀하게 세세한 규정들과 함께 벌칙 규정까지 나열함으로써 인공지능 전문가들의 자율성을 침해하면서 인공지능과 관련된 윤리적 사안을 타율적 방식으로 해결하고자 하는 것처럼 보인다는 점에서 우려하지 않을 수 없다. 이미 우리 사회는 2000년대 초반 정보통신윤리위원회라는 정부기구가 인터넷 및 정보통신과 관련된 윤리 문제를 법률 조항과 각종 행정 규제를 통해 접근하려다가 위헌 판결을 받은 선례가 있다.[19] 이와 같은 '윤리의 법제화'는 윤리가 지닌 근본적 자율성을 훼손함으로써 우리 사회 안에서 윤리적 사안에 대한 비판적 성찰 기회를 박탈하여 결국 윤리를 왜소화할 수 있기 때문이다. 특히 고도의 전문성을 지닌 전문직과 관련된 윤리적 사안의 경우 관련된 어떤 규정을

19) 99헌마480(헌재, 2002.6.27), 전기통신사업법 제53조 등 위헌확인.

단순히 잘 준수하도록 하는 것뿐만 아니라 그들 스스로 규정의 정당성을 끊임없이 성찰하며 자율성을 신장하는 것이 무엇보다 중요하기 때문이다.

IV. 결론

인공지능 기술에 대한 사회적 기대와 우려가 교차하는 현재 상황에서 규범이라는 이름으로 그 기술의 발전 방향을 예단하거나 인위적으로 유도하는 것은 가능하지도 않을 뿐만 아니라 바람직하지도 않다. 현재 상황에서 인공지능을 구현하는 다양한 기술이 상식처럼 우리 사회에 확산되고 있지만 이러한 기술을 활용하여 실제로 인공지능 시스템을 구축하는 것은 여전히 인공지능 기술 전문가들의 몫이다. 인공지능 기술의 기반의 되는 컴퓨팅 연산이 지닌 논리적 변용성과 비가시성이란 특성으로 인해 인공지능 소프트웨어가 실제로 작동하기 전에는 어떤 상황이 등장할지 예견하는 것 역시 쉽지 않다. 최근 보듯이 빅데이터와 딥러닝, 기계학습 등이 어우러져 인공지능 시스템의 자율성이 강화될수록 예측 가능성은 더욱 낮아질 것이며 그 결과 인공지능에 대한 기대와 불안은 더욱 과잉될 것이다.

과잉 기대와 과잉 불안을 해소하며 인공지능 시스템에 대한 사회적 신뢰를 증진할 수 있는 하나의 방안이 인공지능 전문가 윤리다. 인공지능 전문가 윤리는 단순히 윤리 조항을 법조문처럼 나열하는 것이 아니라 인공지능 전문가들이 자신들의 건강한 윤리 의식을 발휘하여 최대한 자율적 활동을 할 수 있도록 그 길을 열어주는 것이야 한다.

이를 위해서는 첫째, 정부는 인공지능 전문가들을 위한 구체적인 법안이나 세세한 시행 세칙 혹은 매뉴얼을 일방적으로 만들기에 앞서 학계, 산업계, 시민사회와 더불어 우리 사회 전체가 공감할 수 있는 윤리적 대원칙을 마련하는 작업을 해야 한다.

둘째, 인공지능 기술을 주도하고 있는 기업이나 단체에서는 자신들이

어떤 원칙에 입각하여 인공지능 기술을 구현하고자 하는지에 대해 스스로 윤리 강령을 개발하고 선포해야 한다. 또한 조직 내부에 윤리적 문제가 발생하거나 비윤리적 관행들이 존재할 경우 해당 사안에 대해 호소할 수 있는 내부 의사소통 통로를 적극적으로 개발하고 이를 고지해야 한다. 나아가 내부 고발이 열려 있음을 적극 알려야 한다. 내부 고발은 궁극적으로는 내부 인력의 자율성과 윤리 의식을 증진하면서 회사 문화를 민주화하여 결국 경쟁력을 강화할 것이기 때문이다.

셋째, 인공지능 기술 전문가들은 인공지능 전문직 종사자이면서 동시에 자율적 민주 시민임을 인식해야 한다. 즉 인공지능 기술 관련 매뉴얼이나 윤리 강령을 숙지하고 수동적으로 준수하는 것을 넘어서 이 강령의 구체적 내용에 대해서는 개선할 여지가 없는지 끊임없이 비판적 문제의식을 견지하며 스스로 윤리 강령의 저자가 되고자 해야 한다.[20]

20) 이 글에 대해 꼼꼼한 심사와 소중한 조언을 해준 익명의 심사자들과 최종 원고를 면밀히 교정해준 편집진에게 감사를 표한다.

참고문헌

김도훈 (2018), 「알고리즘 책임성 논의와 알고리즘에 한 이해」, 『주간기술동향』, 16:14-28.

라이놀드 니부어 (1932), 이한우 옮김 (2017), 『도덕적 인간 비도덕 사회』, 서울: 문예출판사

변순용 (2020), 「데이터 윤리에서 인공지능 편향성 문제에 대한 연구」, 『윤리연구』, 128: 143-158.

오요한, 홍성욱 (2018). 「인공지능 알고리즘은 사람을 차별하는가?」, 『과학기술학연구, 18(3): 153-215.

정용찬 (2015). 「빅데이터 산업과 데이터 브로커」, 『KISDI Premium Report』, 15-04: 1-23.

캐시 오닐 (2016), 김정혜 역 (2017), 『대량살상 수학무기: 어떻게 빅데이터는 불평등을 확산하고 민주주의를 위협하는가』, 서울 : 흐름출판.

『컴퓨터·인터넷·IT용어대사전』(2005), 서울 : 일진사.

서태욱, 「범죄 용의자 AI가 찾는다」 , 『매일경제』, 2016.12.29.

이호준, 「인공지능으로 범죄 막을 수 있을까…검, 범죄예방체계 컨설팅 착수」, 『전자신문』, 2016.08.09.

임세정・이재연, 「카이스트發 'AI 킬러로봇' 논란... 세계 로봇학자들 '보이콧' 왜?」, 『국민일보』, 2018.04.06.

정의길, 「AI가 준 학점, 가난한 학생을 차별했다"」, 『한겨레』, 2020.08.25.

99헌마480 (헌재, 2002.6.27), 「전기통신사업법 제53조 등 위헌확인」.

공정거래위원회 (2020), "네이버(쇼핑, 동영상 부문) 시장지배적 지위남용행위 및 불공정거래행위 제재 - 온라인 플랫폼 사업자가 검색알고리즘을 조정·변경해 자사 서비스를 우대한 행위를 제재한 최초 사례", https://www.ftc.go.kr/www/selectReportUserView.do?key=10&rpttype=1&report_data_no=8759. (검색일: 2020.10.06.)

"행안부, 오픈데이터포럼 발족" https://www.yna.co.kr/view/PYH20170727440200013. (검색일: 2020.09.30.)

Ferguson, A. G. (2017), *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement,* NYU Press.

Moor, James (1985), "What is Computer Ethics?" , *Metaphilosophy,* 16(4)

Rawls, John (1971), *A Theory of Justice*, Harvard University Press.

_____ (1993), *Political Liberalism*, Columbia University Press.

_____ (2001), *PJustice as Fairness: A Restatement*, Belknap Press.

Weisburd, D. (2008), "Place-based policing", *Ideas in American policing,* 9:1-15.

Williams, B. A., Brooks, C. F., and Shmargad, Y. (2018), "How Algorithms Discriminate Based on Data They Lack", *Journal of Information Policy,* 8(1): 78-115.

Chakraborty, S., et al.,(2017), "Interpretability of deep learning models: a survey of results", https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8397411 (검색일: 2020.09.30.)

Dastin, J. (2018), "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (검색일: 2020.09.30.)

Nadler, Jerrold, et al. (2020), *Investigation of Competition in Digital Markets,* https://https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf (검색일: 2020.09.30.)

Perry, W. L., et. al., (2013), *Predictive policing: The Role of Crime Forecasting in Law Enforcement Operations*, https://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf (검색일: 2020.09.30.)

【Abstract】

Discrimination and Bias of Artificial Intelligence

Jung, Won-Sup

The subject of this study is artificial intelligence bias. As various artificial intelligence algorithms are used in public domains such as taxation, policing, and judicial law, the debate over bias is intensifying at home and abroad despite their effectiveness. Chapter 1 introduces the main issues related to AI bias. Chapter 2 examines the reasons and types of artificial intelligence bias in light of the characteristics of the data mining. In Chapter 3, I will examine some of bias issues of algorithms in terms of invisibility and logical malleability.
Chapter 4 examines the role of experts in artificial intelligence and its social governance from the perspective of fairness, and then discusses the effects that artificial intelligence technology can have on the perception of fairness by repeating or hiding existing social biases or discrimination.

【Keywords】 Artificial Intelligence, Bias, Algorithm, Fairness, Data, Discrimination

논문 투고일: 2020. 10. 05
심사 완료일: 2020. 10. 27
게재 확정일: 2020. 10. 27