일반연구논문

인공지능 알고리즘은 사람을 차별하는가?■

오요한* · 홍성욱**

■ 본 논문은 2018년 STEPI 국문 Fellowship의 지원을 받아서 작성된 연구임. 연구를 지원해 준 과학기술정책연구원, 논문의 초고에 대해서 유익한 논평을 해 준 세 분의 심사위원들께 감사드린다.

* RPI 과학기술학과 박사과정 전자우편: ohy@rpi.edu

** 서울대학교 과학사 및 과학철학 협동과정/생명과학부 교수 전자우편: comenius@snu.ac.kr

빅데이터에 근거하여 자동적인 의사결정을 내리는 알고리즘이 사회의 각종 영역에서 점차 널리 사용되고 있는 저변에는 알고리즘의 의사결정이 사회의 자원을 보다 효율적으로 분배하리라는 기대 뿐만 아니라 그 결정이 선입견, 편향, 자의적 판단 등이 개입될 수 있는 인간의 의사결정보다 더 공정한 결과를 낳으리라는 희망 또한 자리잡고 있다. 하지만 알고리즘 의사결정이 그 결정에 의해 영향 받는 이들을 공정하게 다루지 않는다는 주장이 여러 사례와 함께 거듭 제기되면서, 의사결정이 어떻게 절차화되었는지, 또한 특정한 의사결정을 공정하다고 판단하는 데에 어떤 요인이 고려되는지에 대한 근본적인 질문들이 새롭게 제기되고 있다. 본 논문은 사법, 치안, 국가 안보의 세 가지 알고리즘 활용 영역에서 차별의 문제가 제기되는 상황을 구체적으로 분석한 연구들을 검토함으로써, 인공지능 알고리즘이 과연 특정 집단의 인간을 차별하는지, 그리고 공정한 의사결정을 분별하는 기준은 무엇인지 살펴보고자 한다. 본격적인 검토에 앞서 데이터 마이닝 각 단계에서 의도적으로 그리고 비의도적으로 편향적인 결과가 산출될 수 있는 원인에는 무엇이 있는지를 살필 것이다. 결론에서는 이러한 이론적이고 실질적인 검토가 현대 한국 사회에 시사하는 바가 무엇인지 간추려 제시할 것이다.

주제어 | 인공지능, 알고리즘, 빅데이터, 차별, 컴파스 알고리즘, 프레드폴 알고리즘, 국경 통제 알고리즘

1. 서론

2017년 12월, 미국 뉴욕시의 시의원 제임스 바카(James Vacca)와 동료들은 소위 "알고리즘의 설명책임 법안"(algorithmic accountability bill)[1] 이라고 불리는 법안을 발의했다(Kirchner, 2017). 이 법안은 뉴욕시가 특별위원회를 구성해서 시에서 사용되는 알고리즘(algorithm)[2]이 뉴욕시민의 삶에 어떤 영향을 미치는지, 그리고 이런 알고리즘이 연령, 인종, 종교, 성별, 성적 지향, 시민권의 여부에 따라서 시민들을 차별하는지를 조사하는 것을 의무화했다. 법안을 발의한 바카는 이 법안의 목표가 알고리즘의 "투명성(transparency)과 설명책임(accountability)"을 확립하는 데에 있다고 강조했다.[3] 올해 1월부터

1) 법안의 정식 타이틀은 "기관에 의해 사용되는 자동화된 결정 시스템과 관련된 지역 법안" (NYC Local Law No. 49 of 2018). 이 법안과 법안을 논의하는 과정의 기록은 https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0에서 볼 수 있다.

2) 이 논문에서 필자들은 알고리즘을 "특화된 계산에 근거해서 인풋 데이터를 바라는 아웃풋으로 변환시키는 코드화된 절차"로 정의한 Gillespie(2014: 167)의 규정을 따른다. 여기서 보듯이 데이터와 알고리즘은 뗄 수 없는 관계이다(Constantiou and Kallinikos, 2015). 현대사회에서 알고리즘은 우리가 일상적으로 세상을 접하고, 경험하고, 물화시키며, 우리에게 편견을 강화하거나 선택지를 제한하는 방식으로 작동한다(Wilson, 2017).

3) 알고리즘의 투명성은 말 그대로 사람이 이를 들여다 볼 수 있다는 것을 의미하며, 설명책임은 그 알고리즘에 문제가 있을 때 이를 만든 사람이 문제의 원인을 설명할 책임을 진다는 것을 의미한다. 개인이 기업에게 알고리즘의 공개를 요구할 때 기업은 자신의 지적 재산권을 주장하면서 이를 공개하지 않을 수 있기 때문에, 투명성을 보장하는 방법으로 제 3의 감독기관에 의한 조사 방법을 주로 사용한다. 설명책임은 알고리즘의 개발자가 그 작동에 대해서 설명할 의무를 포함해서 윤리적이고 책임성있는 방식으로 이를 개발하고, 그 결과에 책임을 지는 것을 의미한다 (SearchEnterpriseAI, 2015; SearchEnterpriseAI, 2017).

시행된 이 법안에 따라서 뉴욕시는 5월에 공무원, 학계, 법조계, 과학기술계의 전문가로 구성된 태스크포스를 발족시켰다. 이들은 뉴욕시가 학교 배정, 치안, 사회보장제도 등에 사용하는 알고리즘에서 차별적인 요소가 있는지를 검토해서 2019년 말에 보고서를 발간하는 계획을 확정하고 활동에 들어갔다.[4]

법안이 명시한 알고리즘은 "자동화된 결정 시스템"(automated decision system)이다. 이것은 "결정을 내리거나 도움을 주는 데 사용되는 머신러닝(machine learning), 데이터 프로세싱(data processing), 혹은 인공지능(artificial intelligence)의 기술에서 유래된 알고리즘을 포함하는 알고리즘들의 컴퓨터화된 구현"으로 정의된다.[5] 인공지능을 인간을 대신하여 자동적으로 결정을 내리거나, 혹은 인간의 결정을 도와주는 알고리즘으로 잠정적으로 정의한다면, 뉴욕시의 법안은 인공지능 알고리즘을 투명하고 설명책임 있는 것으로 만들기 위한 법안이라고 볼 수 있다. 투명성과 설명책임은 기업의 거버넌스를 지탱하는 가장 중요한 원칙으로 간주되는 것인데,[6] 이것이 인공지능 알고리즘의 영역에도 그대로 도입된 것이다.

2017년 말엽에 이런 법안이 입안된 데에는 이유가 있다. 잘 알려져 있듯이 2000년대 이후에 기존의 신경망 네트워크(neural

4) 태스크 포스 발족과 활동에 대해서는 Sidney Fussell, "NYC Launches Task Force to Study How Government Algorithms Impact Your Life," *Gizmodo* (2018. 5. 16) at https://gizmodo.com/nyc-launches-task-force-to-study-how-government-algorit-1826087643 참조.

5) 각주 1번의 법안 참조.

6) 기업 거버넌스에서 투명성에 대한 관심이 높아지게 된 계기는 2002년 Enron의 도산한 이후 WorldCom, Typo, Qwest, 회계법인 Arthur Anderson 등의 미국 대기업들이 회계부정사건에 휘말리면서 기업에 대한 신뢰 저하와 주가 하락을 낳은 데에서 찾을 수 있다. 안중호, 양지윤 (2006: 100) 참조.

network), 머신러닝 등의 성능을 획기적으로 개선한 딥러닝(deep learning) 등의 방법이 인공지능 분야에 도입되면서, 인공지능의 효율성과 정확성은 놀라울 정도로 높아졌고, 응용되는 영역도 넓어졌다. IBM의 인공지능 왓슨은 2011년 2월에 퀴즈 프로그램 <Jeopardy!>의 챔피언을 이긴 뒤에 암 진단과 금융 상품의 선택에 응용되기 시작했으며, 2016년 3월에 구글 딥마인드의 알파고는 이세돌 국수에 승리를 거뒀다. 구글에서 만든 자율주행자동차는 2012년 5월에 미국 네바다주에서 최초로 운전면허를 획득하기도 했다. 지난 2-3년 동안에 법률 분야에 도입된 인공지능은 한 사안에 대해서 여러 복잡한 법안들을 뒤져서 자료를 만드는 일을 초급 변호사보다 더 잘 수행하기 시작했고, 인공지능에 의한 구글이나 네이버의 번역 서비스도 과거에 비하면 월등히 개선되었으며, 얼굴 인식과 사물 인식 분야에서 인공지능은 거의 사람의 수준에 도달했다. 그 사이에 인공지능의 응용 분야는 더 확대되어 인공지능 알고리즘은 기존의 의료, 법률, 금융 등의 분야뿐만 아니라, 채용, 치안, 사법, 교육, 공공행정, 감사, 국경 관리, 이민 및 난민 관리 등의 분야에서도 인간의 판단을 보조하거나 대체하기 시작했다. 인공지능이 직장의 소멸을 가지고 온다는 우려가 커지면서, 인공지능에 대한 사회적 관심도 매우 커졌다. 두 번의 겨울을 거치고, 인공지능의 세 번째 르네상스가 찾아온 것이었다(브린욜프슨, 맥아피, 2014).

그렇지만 2015년 이후에 인공지능 알고리즘이 예상치 못했던 방식으로 편파적인 판단을 내린다는 소식이 잇달아 전해지기 시작했다. 대표적인 사례는 2015년에 구글의 사물 인식 프로그램으로 출시한 "구글 포토"(Google Photo)라는 카메라 앱(app)이 흑인

커플의 얼굴을 '고릴라' 라는 카테고리로 분류한 것이었다.[7] 이는 큰 사회적 논란을 불러 일으켰고, 구글은 이에 대해 바로 사과하고 시정을 약속했다.[8] 또 같은 해에 미국에 거주하는 아시아인들이 백인에 비해서 과외 튜터링 서비스에 2배 가까운 돈을 내고 있다는 사실이 폭로되었다(Angwin and Larson, 2015). 회사가 튜터링 서비스의 가격을 결정하기 위해 알고리즘에 도입한 여러 변수들이 그런 결과를 낳았던 것이었다. 또 구글의 광고가 여성에 비해 남성들에게 더 높은 보수의 자문, 관리 직종 등의 상대적으로 고급 취업 광고를 내보낸다는 사실도 드러났다(Gibbs, 2015). 일찍이 같은 해 1월 출간되어 평판, 인터넷 검색, 신용 평점의 영역에 활용되는 알고리즘들의 비가시적 영향력과 문제점을 널리 알린 프랭크 파스콸레(Frank Pasquale)의 대중서 『블랙박스 사회 The Black Box Society』가 공교롭게도 뒤이을 소식들의 전조 역할을 한 셈이었다(파스콸레, 2016).

이듬해 2016년에는 마이크로소프트사의 인공지능 챗봇(chat-bot) 테이(Tay)가 일부 트위터 사용자들이 훈련시킨 혐오 표현을 따라하기 시작해서 회사가 시범 서비스를 시작한 지 만 하루도 안되어 서비스를 중단시켰고(Liu, 2017), 같은 해에 미국 법원과 교도소에서 형량, 가석방, 보석 등의 판결에 널리 사용되던 컴파스(COMPAS) 알고리즘이 흑인들에게 편파적인 판결을 냈다는 <프로퍼블리카 ProPublica>지의 폭로가 이어졌다(Angwin, Larson, Mattu, and Kirchner, 2016). 이 폭로는 알고리즘 연구자, 사회과학자, 정책 결정

7) "Google Photos, ..." https://twitter.com/jackyalcine/status/615329515909156865

8) 구글은 이 문제를 해결할 장기적인 해법을 모색하겠다고 했지만, 2018년에 발표된 해결책은 "고릴라"를 검색 인덱스에서 지우는 것이었다. Simonite(2018) 참조.

자들 사이에 큰 논쟁을 불러 일으켰다. 뉴욕시의 2017년 법안을 발의한 제임스 바카도 이 논란을 겪으면서 법안의 필요성을 절감했다고 회고했을 정도였다(Kirchner, 2017). 바로 같은 시기인 2016년 9월에 알고리즘의 불평등한 결과 및 편향 문제를 분석한 수학자 캐시 오닐(Cathy O'Neil)의 대중서 『대량살상수학무기 Weapons of Math Destruction』가 출판되어 베스트셀러가 되었다(오닐, 2017). 보다 최근에는 전자상거래 기업 아마존이 구직자의 이력서를 평가하기 위한 알고리즘을 만들기 위해 알고리즘을 훈련시켜 왔었으나 젠더 편향 등의 문제로 결국 2017년에 개발을 중단했다는 보도가 나왔다(Dastin, 2018). 이 알고리즘은 "여성 체스 클럽" 등 '여성"이 언급된 지원서를 채용대상에서 배제하거나 두 곳의 여성 대학을 졸업한 이들을 감점 매기는 등, 성별 편향적인 결과를 보였기 때문이다.[9]

본 논문은 인공지능 알고리즘이 불러일으킨 공정성과 차별에 관한 논란을 소개하면서 이를 학술적으로 분석하는 데 목적이 있다. 보다 효율적인 논의를 위해 본 논문에서 다룰 "인공지능 알고리즘"은 사회 현상으로부터 수집된 빅데이터 바탕의 예측 모델에 기반한 알고리즘 및 소프트웨어를 부르는 것에 한정할 것이다. 분석을 위해 본 논문은 과학기술학(STS)에서 많이 사용되는 논쟁 연구의 방법론을 사용해서 상반되는 주장을 공평하고 대칭적으로

9) 이력서 평가 알고리즘의 개발 과정에서 영국 소재 아마존 에든버러 엔지니어링팀은 저마다 상이한 직무 능력을 평가하는 500개의 모델을 만들어 과거 지원자의 이력서 데이터에 등장하는 5만개의 전문용어를 추출하도록 했다. 하지만 결과적으로 알고리즘은 젠더 편향 이외에도 여러 문제점을 보였다. 알고리즘들은 다양한 컴퓨터 코드 작성 능력 등과 같이 IT 분야에서 흔한 역량에 낮은 가중치를 부여했고, "실행했다"(executed)와 "포착했다"(captured) 등과 같이 남성 엔지니어의 이력서에 더 빈번하게 사용되는 동사들로 자신을 묘사한 지원자들을 선호했다(Dastin, 2018).

분석한 뒤에,[10] 이런 논쟁을 통해 서로 다른 주장들 중에 어떤 입장이 더 설득력이 있는지, 그리고 정책적인 대안이 어떻게 제시되는지를 보일 것이다.[11] 더불어 사회기술적 현장에서 기술이 개발되고 사용되는 맥락에 대한 컴퓨터과학자들과 문화/비판 연구자들의 연구도 함께 검토할 것이다. 논문의 2절은 인공지능 알고리즘에 사용되는 데이터, 특히 빅데이터의 속성과 이런 데이터에 기반한 분석 모델이 편향적으로 사용될 가능성을 논할 것이다. 이를 바탕으로 이어지는 세 개의 절에서는 사법, 치안, 국가 안보 각 영역에서 알고리즘 차별 문제를 구체적으로 분석한다. 각 영역들은 국가 권력의 동작 기제에 어떻게 인공지능 기술이 접목되고 있는지를 세부적으로 드러내어 보여주는 동시에 인공지능 기술이 어떻게 개인, 시공간, 특정 국가의 시민들이라는 추적과 예측의 대상을 만들어내는지를 효과적으로 보여줄 수 있는 것이기에 선

10) 과학기술학 방법론으로서의 논쟁 연구(Controversy Studies)가 사회적 논쟁의 대상으로서의 과학기술을 분석하는 데에 활발히 적용되었음은 여러 개론적 소개에서 다뤄진 바 있다(Nelkin, 1995; Martin and Richards, 1995). 최근의 개론적 서술에서 쉴라 자사노프는 방법론적 대칭성(methodological symmetry)에 영향을 받은 과학기술학의 질적 방법론들이 과학과 민주주의의 관계를 탐구해 왔다고 소개하며, 그 중 논쟁 연구가 이 주제에 적용된 참여 관찰, 포커스 그룹 연구, 국가간 비교 등 여러 구성주의 과학기술학 방법론들 중에서 가장 두드러지는 방법론이라고 평가한 바 있다. 그는 논쟁연구에 대해 보다 세부적으로 내적(internalist), 상호작용적(interactional), 제도적(institutional) 연구들이라는 세 가지 대략적인 구분이 가능하다고 보았다(Jasanoff, 2017: 269-271). 본 논문은 알고리즘의 판단이라는 기술적 결과물이 공정한가와 같은 '지식의 신뢰성'을 둘러싼 논쟁을 다룬다는 점에서 내적 논쟁연구의 방법을 취한다. 더 나아가 본 논문은 이름 붙이자면 역-제도적 (reverse-institutional) 과학기술 논쟁의 양상에 주목하고 있다. 자사노프의 정리에 따르면, 제도적 논쟁 연구는 어떻게 국가 기관 및 규제 기구의 논리가 과학 수행의 정당성에 영향을 끼치고, 경계 짓기를 수행하는지에 관심을 보인다. 반면, 본 논문이 주목한 현상들에서는 공통적으로 정보기술 및 데이터에 기반한 의사결정이 사법, 치안, 국가 안보 등과 같은 국가 기관의 의사결정의 논리에 결합되거나 영향을 끼치며, 그 결정을 보다 신뢰할 만한 것으로 보이게 한다는 점이 특징적이다.

11) 다만 여러 논쟁이 지금(2018년 10월)도 진행 중인 경우가 많아서 논쟁이 어떻게 종결되는지를 보이는 것은 본 논문의 범위를 넘어선다. 이는 추후의 연구 주제가 될 것이다.

택되었다. 3절은 사법의 영역에서 사용되는 알고리즘이 개개인을 평가할 때 인종이나 빈부에 따라서 개인을 차별하는가의 문제를 다룰 것인데, 최근에 크게 논란이 되었던 노스포인트(Northpointe)사의 위험 판정 알고리즘 컴파스(COMPAS)를 주로 분석할 것이다. 4절은 미국의 여러 도시에서 경찰이 치안 영역에서 사용하는 프레드폴(PredPol) 같은 알고리즘이 범죄가 일어나는 시간이나 공간을 예측할 때 역시 인종과 빈부를 차별하는가의 문제를 다룰 것이며, 5절은 미국, 그리고 유럽 연합 국가들이 국경 관리 및 국가 안보를 위해 사용하는 알고리즘과 데이터베이스가 사람들을 국가 단위로 분류하여 인종이나 국적을 차별하는가의 문제를 분석할 것이다. 마지막 결론에서는 이론 연구가 우리 사회에 가지는 함의를 모색해 보겠다.

2. 빅데이터 분석에서 알고리즘 공정성 문제를 야기하는 요인

이 논문에서 분석할 알고리즘은 데이터, 혹은 빅데이터와 분리해서 생각할 수 없다. 이는 선구적 컴퓨터 과학자 니클라우스 워스(Niklaus Wirth)의 공식화에서 잘 드러난다. 워스는 PASCAL, ALGOL 등 1960년대 초기 프로그래밍 언어를 개발한 이후, 1975년에 "알고리즘+데이터 구조=프로그램"이라는 관계식을 만들었다(Wirth, 1975). 정보학자 폴 도리쉬(Paul Dourish)는 워스를 계승해서 알고리즘을 사회문화적으로 분석하는 작업이 "데이터 항목, 데이터 스트림,

데이터 구조와 같은 데이터의 다양한 현상들"을 분석하는 작업과 병행될 필요가 있다고 주장했다(Dourish, 2016: 8).

그렇다면 지금의 인공지능 알고리즘이 처리하는 빅데이터는 어떤 주목할 만한 특성을 갖는가? 가장 중요한 특성은 빅데이터에 수많은 상관관계(correlations)가 존재한다는 것이다. 내 소비 패턴, 학습 패턴, 인터넷 검색 패턴, 병원 방문 패턴, 전화 이용 패턴 등이 종합되면, 나에 대해서 (심지어 내 자신도 모르는) 많은 상관관계가 생길 수 있다. 이런 데이터가 나에 대해서만이 아니라 다른 사람에 대해서도 수집이 되고, 횡적으로 비교가 되면 훨씬 더 많은 상관관계가 생기게 된다. 여기에 패턴으로부터 스스로 배우는 머신러닝 알고리즘이 결합하면, 내 행동에 대해 예측이 가능해진다.

빅데이터와 결합된 알고리즘에 의한 결정은 의도적으로 혹은 비의도적으로 차별적인 결과를 낳을 수 있다. 우선 알고리즘 의사결정에서 의도적인 차별이 일어날 수 있는데, 이는 성별, 인종, 빈부와 같은 민감한 카테고리를 포함하지 않는 경우에도 가능하다. 머신러닝을 통해 빠진 데이터를 채우는 대체(imputation)가 용이해지고, 직접적으로 획득이 곤란한 민감한 데이터를 대리변수(proxy variable)를 이용해서 획득하는 것이 가능해지기 때문이다(Williams et al., 2018). 특히 대리변수를 사용하면 직접적으로 물어보는 것이 금지되어 있는 민감 정보에 대한 추정이 가능해 진다. 예를 들어, 취업 면접자에 대해서 부모의 직업과 소득을 물어보는 것이 불법이라고 해도, 거주지 정보, 소비 패턴, 취미 활동 등에 대한 정보를 결합하면 이를 어렵지 않게 추정할 수 있는 것이다. 성별과 인종에 대해서도 데이터를 결합, 비교하거나 이를 가능케 하는 몇 가지 질문을 던지면 이를 쉽게 추정할 수 있다.[12]

데이터의 의도적 추정 이외에도 알고리즘이 의도와 무관하게 차별적인 결정을 내리게 될 수 있다. 그 다양한 이유는 알고리즘 의사결정 단계 별로 찾아볼 수 있다(Barocas & Selbst, 2016: 677-694; Kroll, et al., 2017: 679-682). 컴퓨터 과학과 법학에 전문성을 갖춘 연구자들은 알고리즘의 의사결정이 의도하지 않게 차별적인 결과를 낳게 되는 원인을 데이터 마이닝(data mining)의 네 단계인 목표 변수의 정의, 훈련 데이터의 레이블링(labeling)과 수집, 특징 선택(feature selection)의 사용, 그리고 모델을 바탕으로 내린 의사결정 각각에서 찾을 수 있다고 본다.

먼저 데이터 마이닝의 첫 단계인 목표 변수를 정의하는 과정으로 인해 편향적 결과가 야기될 수 있다. 예컨대 "신용도"를 특정 횟수 이상 대출을 상환하지 않을 확률로 정의내리는 것은 비교적 당연해 보이지만, 실상 이러한 방식의 정의는 "신용의 정도"에 대한 특정한 잣대와 실행, 즉 신용 업계라는 특정 행위자가 이를 특정하게 문제 설정하여 신용 창출 및 상환 시스템을 구축하여 실행해온 방식의 산물이다. 게다가 데이터 마이닝 기법은 서로 비교하기 어려운 지표보다는 일관적으로 비교 가능한 지표들을 사용한다는 문제가 있다. 예컨대 경력직 구직자를 평가할 때 이들이 이전 직장에서 받은 연간 업무 평가 점수보다 구직자들의 재직 기간을 사용하는 것을 선호하는 식이다. 후자가 전자에 비해 구직자끼리 쉽게 비교 가능하기 때문이라는 것이다. 하지만 재직

12) 데이터 마이닝 기법이 편향적인 결과를 산출하도록 의도적으로 유도하는 또 다른 경우로 마스킹(masking) 기법이 있다. 이 기법의 하위 분류로는 특정 계층에게 불리한 추론이 가능하도록 일부러 편향적인 데이터를 수집하기, 기존의 편향적인 데이터가 신빙성 있고 공평하다고 주장함으로써 기존 데이터에 배태된 편향을 보존하기, 보다 세부적인 속성들이 아니라 조악하게 분류된 수준의 속성을 사용함으로써 특정 계층들에게 불리한 결과를 피할 수 없게 만들기 등이 있다(Barocas & Selbst, 2016: 692-693).

기간이 업무 성과를 예측할 수 있는 요인으로 중립적이지 않다는 점이 문제의 소지가 될 수 있다. 여성의 경우에 출산 및 육아에 의한 경력 단절 등의 이유 때문에 상대적으로 평균 재직 기간이 짧다.[13] 따라서 재직 기간에 대한 예측을 바탕으로 좋은 직원을 결정한다면 이는 결과적으로 여성에게 적은 채용 기회를 제공하게 될 수 있다.

다음으로 데이터 마이닝의 두 번째 단계인 훈련 데이터를 레이블링하고 수집하는 과정에서도 편향적인 결과가 나올 수 있다. 데이터 마이너는 데이터를 직접 레이블링 할 때가 많은데, 이 과정에서 주관성이 불가피하게 개입된다.[14] 예컨대 대출 상환을 4번 놓친 이가 신용도가 좋은지 나쁜지를 구분하는 작업은 데이터 마이너에 의해서 결정되어야 하며, 이 때 데이터 마이너의 주관이 개입한다(Hand, 2006; Barocas & Selbst, 2016: 681에서 재인용). 그리고 데이터를 수집하는 단계에서 특정 계층이 과대 대표되었거나 과소 대표되었다면 이들에 대해 차별적인 판단 결과를 내릴 수도 있다. 예컨대 특정 계층의 정보가 부정확하게 수집되거나, 통계적으로 상응하는 인구 비율과 차이 나게 수집된 경우가 이에 해당된다. 한 예로 보스턴 시는 “스트리트 범프”(Street Bump) 라는 스마

13) 2014년 OECD 통계에 따르면, 성별에 따른 임금근로자의 평균 근속기간에서 OECD 주요 10개국의 경우 남성이 여성보다 0.4-1.9년 가량 길다. 다만 스웨덴은 여성이 9.3년, 남성이 8.9년으로 여성이 약간 더 길며, 프랑스는 남녀 11.4년으로 동일하다. 한국의 경우 남성 6.7년, 여성 4.3년이다. 금재호(2015) 「노동시장 현황 및 한국적 유연안정성」. 『한국의 노동시장 평가와 유연안전성 확보 방안 토론회』. 서울: 경제사회발전노사정위원회, 27-60쪽. p.33. http://www.fki.or.kr/issue/labor/View.aspx?content_id=51dc20dd-5bf1-4de0-81df-8f72d7311684

14) 게다가 데이터 마이닝 모델을 만드는 데에 사용된 훈련 데이터는 기준 자료(ground truth)의 역할을 맡게 되므로, 레이블링 단계에서 정해진 레이블 결과들은 향후 지속적인 차별을 야기할 수 있다.

트폰 앱을 활용하여 도로의 구멍 위치를 파악하고 이 결과를 도로 보수에 활용하고자 했다. 하지만 스마트 폰 보유율이 계층별로 차이가 있고 서로 다른 계층이 주로 이용하는 도로가 다르다는 점을 감안하면, 앱에서 자동으로 수집된 정보를 바탕으로 도로 보수 지점을 결정하는 것은 시의 자원을 특정한 계층에 보다 유리하게 배정하는 결과를 낳을 수 있는 것이다(Crawford, 2010; Barocas & Selbst, 2016: 685에서 재인용).

또한 데이터 마이닝의 세 번째 단계인 특징 선택의 사용이 편향적 결과를 낳을 수도 있다. 특징 선택은 데이터가 표상하는 수많은 특징들 중에서 과업 달성에 무관하거나 중복되는 특징들을 제거하는 등의 방식으로 그 과업을 위한 최소한의 특징을 찾는 기법을 가리킨다(Liu and Motoda, 2012: xix). 예컨대 금융기관에서는 입수하기 쉽다는 이유로 주거지 정보라는 특징을 바탕으로 대출 등을 결정한다면 이 역시 차별을 낳을 수 있다. 과거에도 이런 관행은 주거지에 따라 차별적인 금융 서비스를 제공하는 레드라이닝(redlining)이라는 결과를 낳았으며,[15] 이는 불법으로 규정되어 있다. 하지만 특징 선택 문제가 해결되기는 현실적으로 어려움이 따른다. 차별적인 의사결정을 낳을 수 있는 특징 선택을 피하기 위해 더 세밀한 특징까지 수집하려면 더 많은 비용이 필요하고, 특정한 계층에 대한 오류를 줄이겠다는 명분만으로 이런 많은 비용을 정당화하기란 현실적으로 어렵기 때문이다.

15) 레드라이닝(redlining)은 미국에서 주로 흑인이 거주하는 빈곤층 거주 지역에만 금융 서비스를 제한한 행위를 말하는데, 이 명칭은 지도상에서 이 지역을 붉은 색으로 둘러싼 것에서 유래했다. 전자 레드라이닝(electronic redlining) 또는 정보 레드라이닝(information redlining)은 유색인종이나 저소득층에 대한 전자적, 정보적 차별을 뜻한다. 강준만(2014) 참조.

마지막으로 특정한 문제를 풀기 위해 설정한 판단기준 자체가 특정한 계층을 분간하는 대리지표(proxy) 역할을 할 수도 있다. 특징들이 점점 더 미세하게 세분화되고 특정한 계층이 데이터 속성들의 조합으로 표현됨으로써, 결과적으로 특정 계급이 데이터 속성으로 간주되는 효과를 낳을 수 있다.[16] 예컨대 앞서 살폈듯 아기를 낳기 위해 퇴직하는 여성들은 전체 여성들의 재직 기간 평균을 낮추며, 결과적으로 재직 기간이라는 속성은 구직자의 성별의 대리지표로 기능한다.

이와 같이 차별을 의도한 경우 뿐만 아니라 차별의 명시적인 의도가 없음에도 불구하고 데이터 마이닝의 각 단계마다 편향이 개입될 소지는 충분하다. 그렇다면 이런 단계마다 편향, 차별, 불공정성의 소지를 최소한으로 줄임으로써 알고리즘 불공정성의 문제를 개선할 수 있는 것일까? 일견 타당한 추론이지만, 여러 연구자들은 알고리즘 공정성을 확보하려는 시도가 새로운 문제를 야기할 수도 있음을 지적하고 있다. 알고리즘 공정성의 문제를 해결하려는 노력이 빅데이터 알고리즘의 본래 목적인 알려지지 않은 정보를 최대한 정확하게 추정하거나 예측하려는 노력과 충돌할 수 있기 때문이다. 예컨대 정확도를 높이기 위해 보다 세부적으로 데이터를 수집할 경우 세부 데이터의 조합이 결국 특정 계층을 지시하는 결과를 낳을 수 있다(Selbst, 2018: 137; cf. Williams et al., 2018). 알고리즘 공정성의 문제가 알고리즘 정확성의 문제와 상충 관계를 갖는다는 것이다.[17] 이런 특성들은 이후 세 개의 절

16) 앞서 대리변수(proxy variable)가 수집할 수 없지만 핵심적인 데이터 속성을 의도적으로 추론하려는 시도에 가깝다면, 판단기준 설정 단계에서의 대리지표 문제는 문제 설정 단계에서 해당 문제 설정이 차별적인 데이터 속성과 연관되어 있는지 미리 인지하지 못한다는 점이 앞서 논의한 대리변수와의 주요한 차이이다.

에서 알고리즘의 구체적인 응용을 분석하면서 보다 분명해 질 것이다.

3. 사법 영역에서의 알고리즘 공정성 논쟁: 컴파스(COMPAS)[18] 알고리즘은 흑백을 차별하는가?

알고리즘 공정성의 문제가 정확성의 문제와 상충 관계를 가진다면 알고리즘 공정성의 문제를 개선하는 데에는 불가피한 한계가 있음을 인정하지 않을 수 없게 된다. 하지만 알고리즘 공정성 문제를 개선하기 어려운 또 다른 주요한 이유는 보다 근원적인 긴장 관계에서 기인한다. 바로 알고리즘 공정성(fairness) 자체를 정의하는 지표들과 척도들이 다양하며 공정성에 대해 서로 다른 결론을 도출할 수 있기 때문이다. 대표적인 사례가 재범 확률을 예측하는 노스포인트(Northpointe)사가 개발한 알고리즘 컴파스가 형사 재판을 받는 개인의 위험성을 예측할 때 백인과 흑인에 대해 차별적인 판단을 내린다는 주장과 관련된 논쟁이다.

이 주장은 2016년에 탐사보도매체인 ProPublica에 의해서 제기되었고(Angwin, Larson, Mattu, and Kirchner, 2016), 이후 New York

17) 그렇기 때문에 데이터 과학자들이 정확한 동시에 공정한 알고리즘을 만들려면 최적의 타협점이 존재하는지 조사하여 이를 찾을 필요가 있다. 예컨대 차별을 최소화하는 알고리즘 분류기(classifier)를 만들려면 오인식률(misclassification rate)이 최대 허용 범위를 넘어서지 않는 한에서 초 매개변수(hyper-parameter)를 미세조정해야하는 과업을 수행할 수 있어야 한다(d'Alessandro, O'Neil, & LaGatta, 2017: 128).

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

Times 같은 권위있는 매체들에 의해서 보도되어 세상에 널리 알려졌다. 컴파스는 피고인의 일반범죄 재범률과 강력범죄 재범률을 추정하여 피고인의 리스크 스코어(risk score)를 1(최저위험 군)에서 10(최고위험 군)까지의 숫자로 결정해서 판사에게 제공함으로써, 판사가 형량을 선고하거나 가석방을 결정하는 보조 자료로서 널리 사용되는 알고리즘이다. ProPublica의 기자들은 미국 플로리다 주 브로워드(Broward) 카운티 법원에서 선고받은 7,000명 이상의 피의자를 대상으로 컴파스의 예측결과와 판결 후 2년 간의 실제 재범 여부를 입수하여 조사했다. 이 기자들은 특히 두 가지 통계를 보고 컴파스가 흑인과 백인을 차별하는 결과를 냈다고 주장했다. 그 중 하나는 백인의 경우에는 리스크 스코어 1에 해당되는 사람이 가장 많고, 이후 10까지 그 비율이 계속 감소했음에 반해, 흑인들의 경우는 1부터 10까지 비슷한 비율로 판정을 받았다는 것이었다. 즉, 흑인의 경우 재범 확률이 높다고 판정을 받은 사람들이 백인의 경우보다 훨씬 더 많았던 것이다.

두 번째는 재범률이 높은 것으로 예측되었지만 실제로 2년 간 범죄를 저지르지 않은 것으로 드러난 경우가 흑인의 45%로 백인의 23%에 두 배에 달한다는 점, 그리고 재범률이 낮은 것으로 예측되었지만 실제로 2년간 범죄를 저지른 경우가 백인이 48%로 흑인 28%보다 훨씬 높다는 사실이었다<표 1>. 즉 흑인의 경우 위양성(false positive)이 백인에 비해서 거의 두 배, 위음성(false negative)은 백인의 절반보다 조금 더 높은 정도였다는 것이다. 이 두 증거를 종합하면 흑인은 더 많은 수가 편파적으로 고위험 군으로 분류되었고, 백인은 더 많은 수가 부당하게 저위험 군으로 분류되었다고 할 수 있다. 이를 토대로 ProPublica는 컴파스 알고

리즘이 인종차별적인 결정을 내리고 있다고 주장했다.

〈표 1〉 백인에 비해 흑인에게 불리하게 분류된 위험점수(ProPublica, 2016)

	백인	아프리카계 미국인
위험 점수가 높다고 분류되었고 이후 실제로는 2년 안에 재범하지 않은 확률(false positive)	23.5%	44.9%
위험 점수가 낮다고 분류되었으나, 실제로는 2년 안에 재범한 확률(false negative)	47.7%	28.0%

반면 이를 반박하는 기술문서에서 노스포인트사는 앞선 통계적 사실들이 알고리즘의 의사결정이 인종 차별적이라는 주장을 뒷받침하지 않는다고 반박했다(Dieterich et al., 2016). 회사는 우선 컴파스 소프트웨어의 목적은 재범자를 예측하는 것에 있으며, 이에 있어 백인과 흑인에 차별적인 결과는 나타나지 않았다고 주장했다. 위험 점수가 높다고 분류되었고 실제로 재범한 이의 경우(진양성, true positive)는 백인 59%, 흑인 63%로 유의미한 차이가 나타나지 않았고, 위험 점수가 낮다고 분류되었고 실제로 2년 안에 재범하지 않은 이의 경우(진음성, true negative) 역시 백인 71%, 흑인 65%로 크게 차이나지 않았다는 것이다<표 2>.

〈표 2〉 백인과 흑인의 위험점수의 유불리가 뚜렷하게 구별되지 않는 경우 (Dieterich et al., 2016)

	백인	아프리카계 미국인
위험 점수가 높다고 분류되었고, 실제로 2년 안에 재범한 확률 (true positive)	59%	63%
위험 점수가 낮다고 분류되었고, 실제로 2년 안에 재범하지 않은 확률 (true negative)	71%	65%

출처: 김기태, 정재관(2016)의 정리를 참조함.

또한 노스포인트사는 ProPublica의 주장의 근거 지표 중 하나이기도 한 위양성에 대해 이 지표가 기본구성비율(base rate)의 영향을 받으며 정(正)의 상관관계를 띤다는 연구 결과(Leeang et al., 2013)를 인용하며, 기본구성비율 자체에 있어 흑인의 재범률과 백인의 재범률이 차이를 보임을 지적했다. 흑인의 일반범죄 재범률은 51%로 백인의 39%에 비해 높으며, 흑인의 강력범죄 재범률 역시 14%로 백인의 9%에 비해 높다는 것이다. 회사가 배포한 기술문서에 따르면, 높은 값의 위양성 값은 백인에 비해 높은 흑인의 재범률이라는 기본구성비율 자체의 차이에서 기인한 것일 뿐, 컴파스 알고리즘의 판정이 인종차별적임을 보이는 증거가 아니라는 것이다.

ProPublica의 주장을 보면 컴파스가 공정하지 못한 것 같지만, 노스포인트사의 주장을 보면 컴파스가 충분히 공정하다. 왜 이런 차이가 생긴 것일까? 우선 노스포인트사가 주목하는 공정성은 어떤 사람이 흑인이건 백인이건 인종에 관계없이 그를 평가한 항목들의 점수를 더한 값이 같다면 동일한 최종 리스크 스코어를 얻는다는 것을 의미한다. 이는 "캘리브레이션"(calibration)이라고 알려진 공정성의 기준이다. 그리고 회사는 컴파스가 고위험 군으로 평가된 사람들 중 실제로 재범을 저지르는 사람들의 비율(진양성) 역시 흑백과 무관하게 비슷한 값을 냈음을 강조했는데, 이는 예측적 공정성(predictive parity)이라는 기준을 만족한 것이었다(Dieterich et al., 2016).

원래 ProPublica는 흑인/백인의 리스크 스코어 분포가 크게 차이가 난다는 것을 불공정의 한 요소로 강조했지만, 이후 논의에서는 이를 다시 언급하지 않았다. 위험 분포가 같아야 한다는 것은 "통계적 공정성"(statistical parity)이라는 기준인데, 인종에 따른 범죄율과 재범률이 차이가 나는 세상에서 이는 달성하기 힘든 기준이기 때

문이다. 그렇지만 이어진 반론에서 ProPublica는 "'예측적 공정성'이라는 것은 실로 '최적의 차별'(optimal discrimination)이다"라고 평가한 시카고 대학교의 통계학자 네이튼 스레브로(Nathan Srebro)의 논평을 인용하면서, 노스포인트사가 강조한 예측적 공정성이라는 것이 결국 흑인에 대해 허위 고위험 군을 더 많이 만들어낸 결과를 낳는다고 다시 비판했다(Angwin and Larson, 2016). ProPublica는 공정한 알고리즘이라면 위양성/위음성의 비율이 흑백에 상관없이 비슷한 점수가 나와야 한다고 강조했는데, 이런 "오류 비율의 공정성"(error-rate parity)은 스레브로 그룹의 결론이기도 했다(Hardt et al., 2016).

이런 기준들을 모두 만족시키는 방식으로 공정성의 기준을 정할 수는 없을까? 2016년에 컴파스에 대한 논쟁이 시작되면서 여러 통계예측 분야의 전문가들이 이 문제를 분석했지만, 이들의 답은 부정적이었다. 일례로 클라인버그와 동료들은 캘리브레이션과 오류 비율의 공정성이라는 두 개의 기준이 아주 특정한 경우를 제외하고는 양립할 수 없는 기준임을 보여주었다(Kleinberg et al., 2016). 클라인버그의 분석에 의하면 이 중 하나의 기준을 더 잘 만족시키려고 하면, 다른 기준에서는 벗어나는 결과를 낳았던 것이다. 더 나아가 버크와 그의 동료들은 통계적으로 공정성을 가늠하는 방식이 적어도 5개는 존재한다고 지적했다(표3). 이들은 이런 5가지 조건을 동시에 만족시키는 "완전한 공정성"(total fairness)은 가능하지 않다고 하면서, 이 각각이 소수 그룹의 보호에 대해서 서로 다른 함의를 가지는 방식으로 서로 상충되며, 이 중 어느 하나를 온전히 만족시키려고 하면 통계적인 정확성을 잃는 결과가 생긴다고 지적했다(Berk et al., 2018). 반면에 또 다른 통계학자는 예측적 공정성을 조금 희생 하더라도 위양성과 위음성의 값을 같게

하는 것이 더 공정한 방식이라고 주장했다(Chouldechova, 2017).

〈표 3〉 형식 이론(formal theory)에 기반하여 다양하게 정의된 알고리즘 공정성의 지표

정의들	성립 조건	계층마다 동일해야 하는 값
전반적 정확성 균등 (Overall accuracy equality)	집단의 각 계층에(예: 백인과 흑인) 대해 전반적 절차적 정확성이 동일	(a+d)/(a+b+c+d)
통계적 동등성 (Statistical parity)	각 계층에 대해 주변 분포들 (marginal distributions)이 동일	(a+c)/(a+b+c+d) and (b+d)/(a+b+c+d)
조건부 절차적 정확성 균등 (Conditional procedure accuracy equality)	각 계층마다 조건부 절차 정확성 (Conditional procedure accuracy)이 동일	a/(a+b) and d/(c+d)
조건부 사용 정확성 균등 (Conditional use accuracy equality)	각 계층마다 조건부 사용 정확성 (Conditional use accuracy)가 동일	a/(a+c) and d/(b+d)
처리 균등 (Treatment equality)	계층마다 위음성(false negatives)과 위양성(false positives) 간의 비율이 동일	c/b or b/c
종합적 공정성 (Total fairness)	앞선 다섯 가지 공평성 조건을 모두 만족	

a: 진양성, b: 위음성, c: 위양성, d: 진음성 (Berk et al. 2018: 13–15).

컴파스의 공정성에 대한 논쟁이 생기면서 이를 법정에서 사용하는 것을 문제 삼아 대법원에 탄원한 경우도 있었다. 총격전에 사용된 차를 운전하면서 경찰의 추격을 따돌리려 한 죄로 2013년 기소된 루미스(Eric Loomis)는 성범죄 전력이 있었고, 컴파스의 알고리즘에 의해서 고위험으로 분류되었다. 그는 미국 위스컨신 주 법

원에 의해 징역 6년을 선고받았는데, 자신이 컴파스 알고리즘의 과학적 타당성을 검증할 수 없으며, 이것이 여성보다 자신과 같은 남성을 더 위험하다고 판단하는 불공정성을 내포하기 때문에 공정하게 재판받을 권리를 침해당했다고 항소했다(그는 흑백 차별을 문제 삼지는 않았다). 위스컨신 고등법원은 이 판단을 대법원에 넘겼고,[19] 위스컨신 주 대법원은 2017년 6월 이를 기각했다.[20] 대법원은 피고가 컴파스 알고리즘을 이해하지는 못하지만 설문 문항에 직접 답을 하면서 이것이 어떤 정보를 이용하는지 충분히 이해했고, 만약에 이 때 정보를 잘못 제공했다면 재판 과정에서 이견을 제시할 기회가 있었기 때문에 이것이 재판의 공정성을 침해했다고 볼 수 없다고 판결했다. 남성을 차별했기 때문에 공정하지 못하다는 주장에 대해서 대법원은 남성이 여성보다 강력범죄를 저지르는 비율이 높은 것이 사실이고 컴파스의 리스크 스코어 산출에는 이런 비율이 반영되어 있었을 뿐이며, 이렇게 나온 리스크 스코어는 판사가 판결에서 고려한 여러 요소 중 하나일 뿐이라고 반박했다.[21] 대법원의 이런 판결은 컴파스가 판사의 판결을 도와주는 알고리즘으로 미국의 법정에서 정당하게 계속 사용될 수 있다는 인가를 내준 셈이나 마찬가지였다.

19) 위스컨신 고등법원 재판부는 항소 기각 사유로 COMPAS의 분석보고서가 유일한(sole) 근거는 아니며, 6년형을 선고한 재판부가 분석 결과에 동의하지 않을 재량과 정보를 갖는다는 점을 들었다. 유영무(2017. 7. 17) 「인공지능 분석에 근거한 형사재판의 문제점」. 법률신문.

20) "Loomis v. Wisconsin – SCOTUSblog" in *SCOTUSblog - Supreme Court of the United States Blog.* http://www.scotusblog.com/case-files/cases/loomis-v-wisconsin/

21) "Brief for the United States as Amicus Curiae" (Eric L. Loomis v. State of Wisconsin, May 2017) at http://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf

대법원의 판결에서도 나타났지만 통계적 예측 프로그램의 공정성을 염려하는 사람들이 가장 어려워하는 문제는 기본구성비율이다. 우리는 강력범죄의 대다수가 젊은 남성들에 의해서 저질러진다는 것을 알고 있다.[22] 그렇다면 강력범죄에 대한 리스크 스코어를 산정한다고 했을 때 성별과 나이가 중요한 변수로 포함될 것이며, 젊은 남성의 경우 성별과 나이라는 속성 때문에 리스크 스코어가 높아질 것이다. 그런데 처음 범죄를 저질러서 재판을 받는 20대 남성의 경우에는 자신과 나이가 비슷한 다른 남성들이 저지른 범죄 때문에 자신의 리스크 스코어가 높아지는 것을 부당하다고 생각할 수 있다. 인종의 경우도 마찬가지다. 노스포인트사가 강조하듯이 미국의 경우 지금까지 강력범죄를 저지른 사람 중 백인보다 흑인이 더 높은 비율을 보인다. 기존 데이터에 기반한 리스크 스코어 산정 방식에 따라, 형량 선고를 앞둔 흑인 피의자들은 백인에 비해 더 높은 리스크 스코어를 부여받는다. 그러면 이들의 형량이 늘어나고, 백인에 비해 가석방될 가능성도 줄어든다. 이들은 더 오랜 기간 동안 교도소에 있게 되고, 이는 다시 향후 재판받을 흑인들의 리스크 스코어를 높이는 결과를 가져온다. 양의 피드백이 작동하는 이러한 "피드백 효과"는 결국 더 많은 흑인을 교도소에 가두는 결과로 이어진다. 과학철학자 이언 해킹(Ian Hacking)이 다중인격의 구성을 논하면서 제시한 "루핑 효과"(looping

22) 대검찰청이 발행한 2016년 발생 범죄 통계에 따르면 살인, 강도, 방화, 성폭력 등 "강력범죄(흉악)" 범죄자의 남녀 비율은 96.6%:3.4%이며, 가장 많은 범죄자 연령대는 19세~30세(27.0%), 31세~40세(21.4%), 51세~60세(15.0%), 18세 이하(8.7%) 순이었다. 또 다른 강력범죄 유형으로는 폭행, 상해, 공갈, 약취와 유인, 체포와 감금, 폭력 행위 등의 "강력범죄(폭력)"가 있다. 해당 분류 범죄의 2015년 통계에 따르면 남녀비율은 83.7%:16.3%이며, 범죄자 연령대 비율은 41세~50세(24.1%) 51세~60세(21.1%), 19세~30세(20.7%), 31세~40세(19.5%) 순이었다. 대검찰청, (2017). 『2017 범죄분석』. http://www.spo.go.kr/spo/info/stats/stats02.jsp

effect)[23]처럼, 알고리즘이 '교도소에 수감된 흑인 재소자'라는 범주의 인간을 더 많이 만들어 내는 것이다. 미국의 많은 민권운동가들은 흑인이 강력범죄의 비율이 높았던 것은 흑백 인종차별의 결과였다고 항변하면서 이런 차별을 고착화하는 컴파스의 사용을 반대한다. 이런 피드백 효과라는 문제는 컴파스를 지지하는 사람들조차 우려하고 있는 부분이다.

4. 치안 영역에서의 알고리즘 차별: 예측적 치안유지는 편향적인가?

영화로도 만들어져서 더 유명해진 필립 K. 딕의 SF 소설 「마이너리티 리포트」를 보면 범죄가 일어날 시간과 장소를 미리 예측해서 경찰이 범죄가 일어나기 전에 현장에 출동하는 장면이 나온다. 이렇게 범죄가 예측되어 예방되는 미래 사회는 '범죄율 제로' 사회이다. 끊임없이 발생하는 크고 작은 범죄에 시달리는 치안 담당자들에게는 이런 미래사회가 어쩌면 다름 아닌 '유토피아'일지도 모른다.

2013년의 랜드 연구소(RAND Corporation)에서 출간한 보고서에 의하면 "예측적 치안유지"(predictive policing)에는 범죄를 예측하는 것, 범인을 예측하는 것, (과거 범죄로부터) 범인의 신원을 예견

23] 루핑 효과는 특정한 정신의학적인 개념이 이에 해당하는 사람들을 만들어 내며, 이것이 다시 이런 정신의학적 개념을 정당화하는 루프(loop) 형태의 과정을 지칭한 것이다. Hacking (1995) 참조.

하는 것, 그리고 희생자를 예측하는 것의 네 가지 유형이 있다(Perry et al., 2013). 이러한 예측적 치안유지는 보통 4단계로 구성되는데, 첫 번째는 경찰이 데이터를 수집하는 것이며, 두 번째는 이를 분석해서 예측을 하는 것, 세 번째는 경찰 인력을 이용해서 작전을 짜는 것, 그리고 마지막은 범죄자의 반응에 대응하는 단계이다. 이 마지막 단계는 작전에 대한 평가를 낳으면서 다시 새로운 첫 번째 단계, 즉 데이터 수집 단계와 연결되어 있다.

여기에서는 이런 네 가지 예측 중에서 첫 번째 범주인 "범죄의 예측"을 주로 다룰 것인데, 이 첫 번째 유형의 예측은 기본적으로 범죄가 일어날 법한 시간과 장소를 예측하는 것을 의미한다. 랜드 보고서가 나오기 몇 년 전인 2008년에 미국의 범죄학자 데이비드 와이스버드(David Weisburd)는 범죄의 패러다임이 사람 중심의 패러다임(people-centric paradigm)에서 장소 중심의 패러다임(place-centric paradigm)으로 바뀌어야 한다고 주장했다. 즉, 범죄자를 프로파일링 하던 과거의 방식에서 범죄가 빈번히 일어나는 "범죄 핫스팟"(crime hotspots)을 찾아내고 이를 감시하는 방법으로의 변환을 주장했던 것이다(Weisburd, 2008). 얼마 뒤에 제임슨 툴(Jameson L. Toole)과 동료들은 수학과 물리학 공식을 사용해서 유형별로 범죄 사건의 시/공간적 특성을 찾아내어 필라델피아의 10년간 범죄 발생을 분석했고, 와이스버드의 말처럼 범죄가 빈번히 일어나는 핫스팟이 존재한다는 것을 발견했다(Toole et al., 2011).

그렇지만 더 급진적인 시도는 캘리포니아에서 일어났다. 수학자 조지 몰러(George O. Mohler)와 동료들은 큰 지진이 일어난 뒤에 여진이 생긴다는 지진의 원리와 비슷하게 하나의 범죄가 생긴 뒤에 이를 뒤따르는 범죄들이 비슷한 지역에서 뒤이어 발생한다

고 추론하고, 여진을 예측하는 프로그램을 변형해서 범죄 예측 프로그램을 만들었다. 여기에 몰러의 동료인 쇼트(Martin B. Short)는 범죄자의 심리를 포함시켜서, 어느 지역이 범죄를 저지르기에 "만만해" 보이는가의 변수를 포함시켰다. 이들은 LA의 한 지역에서 1년 동안 일어난 범죄 데이터를 가지고 알고리즘을 완성했고(Molher et al., 2011), 성공적인 테스트를 거쳤다. 이 테스트에서 인간 분석가가 형사 범죄가 일어날 지역 중 3%만을 예측한 반면, 알고리즘은 6%를 예측한 것이다.[24] 이후 몰러와 동료들은 개인투자자 및 UCLA 벤처스 투자회사로부터 총 370만 달러 규모의 투자를 받아[25] 프레드폴(PredPol)이라는 회사를 2012년 창업하여 이 기술을 상업화했다.[26] 이 회사의 범죄 예측 알고리즘 프레드폴은 출시부터 언론의 큰 주목을 받았다. 영국 경제지 Economist에 따르면, 영국 켄트 카운티(Kent County) 북부에서 2013년 4월부터 4달간 진행한 시험에 따르면, 모든 시내 형사 범죄의 8.5%가 프레드폴이 예측한 구역 안에서 일어났으며, 더 많은 수의 범죄가 그 구역과 바로 인접한 지역에서 일어났다. 이에 비해 경찰 분석가의 예측은 5%에 머물렀다.[27] 실제로 4달 실험 전후로 시내 범죄는 6% 감소하기도 했다.[28] 이후 프레드폴사의 예측 프로그램이 LA,

24) *The Economist* (2013. 7. 20), "Don't even think about it - Predictive policing"

25) CrunchBase (n.a.), "PredPol" https://www.crunchbase.com/organization/predpol

26) PredPol은 Predictive Policing의 머리글자들을 따서 만든 명칭이다. 몰러는 현재 PredPol 회사의 최고 데이터 과학자(Chief Data Scientist)로 재직중이다. PredPol (n.a.), "Company | Management Team | About Us | PredPol Predictive Policing" http://www.predpol.com/about/company/

27) *The Economist* (2013. 7. 20)

28) An Internet Archive page of Kent Police (2013. 12. 3) "Predictive Policing day of action targets

아틀란타, 시애틀, 산타 크루즈 같은 도시의 치안 유지에 도입되기 시작했다.[29] 산타 크루즈 시는 이 프로그램을 도입하고 1년 동안 절도가 11%, 강도가 27% 감소했다고 발표했다.[30]

경찰이 이 프로그램을 열성적으로 도입한 데에는 몇 가지 이유가 있다. 우선 지난 몇 십 년 동안 도시의 인구가 늘고, 따라서 절도, 강도 등 도시형 범죄가 급증했다. 반면에 여러 가지 이유로 경찰 인력은 축소되어 치안 유지에 애로점이 증가했다. 여기에 백인 경찰이 인종을 차별한다는 비판이 지속적으로 제기되었다. 미국의 여러 도시에서 백인 경찰이 흑인을 무자비하게 폭행하거나 사살하는 사건이 일어났고, 이런 사건이 일어날 때마다 경찰의 인종차별은 여론의 도마에 올랐다. 이런 상황에서 일견 중립적으로 보이면서 범죄를 예측하는 알고리즘은 경찰의 입장에서 봐도 경찰이 직면한 모든 문제를 한 번에 해결할 수 있는 매력적인 해법이었던 것이다. 초기에는 이런 알고리즘에 대해서 미래를 확실하게 알 수 있다거나, 범죄를 확연히 줄일 수 있다는 식으로 과장된 평가가 많았지만(Perry et al., 2013: 11-12), 시간이 지나면서 이런 과장은 수그러들고 현실적인 평가가 이를 대체했다.

---

burglars," archived on May 2nd, 2014. [https://web.archive.org/web/20140502001827/http://www.kent.police.uk/news/latest_news/131203_predpol_day_a.html]. Kent 카운티 경찰국의 원본 페이지는 2018년 11월 현재 열리지 않는다 [https://www.kent.police.uk/news/latest_news/131203_predpol_day_a.html].

29) *Los Angeles Times*(2010. 8. 21), "Stopping Crime Before It Starts"; *Santa Cruz Sentinel*(2012. 2. 26), "Modest Gains in First Six Months of Santa Cruz's Predictive Police Program,"; *NPR*(2011. 11. 26), "At LAPD, Predicting Crimes Before They Happen," http://www.npr.org/2011/11/26/142758000/at-lapd-predictingcrimes-before-they-happen

30) *San Francisco Weekly*(2013. 10. 30), "All Tomorrow's Crime: The Future of Policing Looks a Lot Like Good Branding."; *Santa Cruz Sentinel*(2012.02.26), "Modest Gains in First Six Months of Santa Cruz's Predictive Police Program".

범죄를 100% 정확히 예측하지는 못해도 예측적 알고리즘은 여러 이점이 있는 것으로 평가되었다. 우선 알고리즘은 특정한 지역이나 상황에 경찰의 순찰력을 집중할 수 있게 했고, 또 어떤 환경에서 범죄가 더 잘 일어나는지를 분석할 자료를 제공했다. 더 효과적으로 범죄를 예방하는 조치를 계획할 수 있게 했으며, 경찰이 특정한 전략을 사용할 수 있게 도울 수도 있었다. 이 프로그램에 비추어서 경찰이 세운 가설을 테스트해 볼 수도 있었고, 범죄들을 관통하는 패턴이나 경향을 분석할 수도 있었으며, 어떤 범죄 데이터가 분석적으로 유용한지에 대한 기준을 제공해 주기도 했다. 다른 빅데이터가 그렇듯이, 이 경우에도 범죄와 관련된 여러 데이터들이 알고리즘에 의해서 상대적으로 쉽게 상관관계로 이어질 수 있었다. 예를 들어 범죄 시간, 장소, 유형, 재산 손실, 희생자 유형 등의 상관관계를 얻어내면, 경찰이 이런저런 경우에 어떤 작전을 세워야 하는지 더 분명해졌다(Schlehahn et al,, 2015).

그렇지만 프레드폴 알고리즘의 문제점도 곧 드러났다(Perry et al., 2013: 12-13). 무엇보다 알고리즘에서 사용하는 데이터의 질이 낮다는 것에 대해 문제가 제기될 수 있었다. 예를 들어 가게 주인들이 아침 8시에 절도를 신고했다면, 이것만으로는 8시에 절도가 일어난 것인지 아니면 그 때 가게 문을 열다가 절도를 발견한 것인지 불분명했다. 경찰의 입장에서도 알고리즘이 핫스팟을 적시했다고 해도 적시된 핫스팟에서 대체 무엇에 주목해야 하는지 이해하기 힘들었고, 따라서 자신이 사용하는 알고리즘의 효력을 평가하기 쉽지 않았다. 그렇지만 가장 심각한 문제로 지적된 것은 이런 프로그램이 시민권, 특히 미국 수정헌법 제 4조(The Fourth Amendment)에 명시된 프라이버시의 권리를 침해한다는 비판이었

다. 실제로 몇 년이 지나지 않아서 이런 비판이 학계, 법조계, 인권단체에서 쏟아졌다.

『대량살상 수학무기』의 저자 캐시 오닐은 프레드폴 알고리즘을 만들 때 경범죄를 데이터에 포함시킨 것부터 잘못이라고 비판했다(오닐, 2017). 보통 경범죄는 거리에서 술을 마시면서 소란을 피우거나 대마초를 흡입하는 행위 등을 포함하는데, 중산층 이상의 백인들이 많이 사는 지역에서는 이런 행위가 집 안에서 일어나고 따라서 경찰의 순찰에 의해 적발되지 않는 반면에, 흑인들이 많이 거주하는 빈민가에서는 이런 행위가 거리에서 주로 일어나고 따라서 경찰에 의해 쉽게 적발되었다는 것이다(Ferguson, 2015: 402). 경범죄의 데이터가 포함되면서 처음부터 흑인들이 많이 거주하는 더 많은 지역이 범죄의 핫스팟으로 분류되었고, 경찰은 이런 지역을 순찰하다가 의심스럽게 보이는 사람들을 세워서 검문하게 되고, 이 중에 마약이나 수상한 물건을 가진 사람들을 검거하게 되는 식이었다. 이런 검거는 프레드폴 알고리즘의 성공률을 높이면서 다시 이 알고리즘에 입력되고 데이터로 기능하며, 이는 다시 경찰로 하여금 흑인들이 사는 지역을 집중 순찰하도록 이끈다(오닐, 2017).

또 다른 연구자인 젤프스트(Selbst, 2018)는 프레드폴의 예측이 네 가지 점에서 문제가 있다고 지적했다. 첫 번째 문제는 범죄를 컴퓨터 코드화하는 데에서 발생한다. 범죄는 다양한 방식으로 이루어지며 그 유형도 다양지만, 예측 프로그램은 이 중에서 컴퓨터가 쉽게 인지할 수 있는 범죄에 초점을 맞춘다. 그 중 대표적인 것은 절도를 포함한 재산범죄(property crime)이다. 이런 범죄는 지리학적으로 분석이 가능하기 때문에 알고리즘이 선호한다. 예를 들

어 프레드폴 알고리즘은 지역을 500피트x500피트(약 150미터x150미터)의 격자로 나눠서 분석한다. 재산범죄의 발생비율이 높은 지역을 순찰하는 경찰은 이 지역에서 살인이나 방화와 같은 다른 범죄도 더 많이 일어날 것이라고 가정한다. 그렇지만 이런 재산범죄의 발생비율이 다른 범죄의 발생비율과 항상 동일한 것은 아니다. 따라서 쉽게 지역화되는(territorialized) 재산범죄를 기준으로 순찰을 하면 재산범죄의 발생비율이 낮은 지역은 경찰 순찰에서 과소평가되는 결과가 생기고, 이는 지역 간의 차별을 불러올 수 있다. 이런 지역이 흑.백 인종의 거주 지역으로 나뉘어 있다면, 이는 인종에 대한 차별로 이어진다(ibid.: 131-133).

두 번째 편향은 빅데이터를 "훈련"시키는 과정에서 발생할 수 있다. 새로운 데이터를 처리하기 위해서는 "기준값"(ground truth)을 정하고, 이를 기준으로 이후의 데이터를 평가해야 한다. 범죄 예측 알고리즘에서는 무엇을 기준값으로 사용했을까? 앞서 나왔지만, 순찰을 하는 지역의 과거의 범죄 발생 데이터가 기준값이 되었다. 그런데 이 과거 범죄 데이터에 문제가 있다. 우선 경찰이 가진 데이터가 모든 범죄에 대한 데이터가 아니다. 신고된 범죄, 혹은 경찰이 기록한 범죄는 전체 범죄의 일부분에 불과하다. 이는 가벼운 절도 사건일 경우에는 더더욱 그러하다. 또 범인에 대한 경찰의 기록은 주로 체포에 국한된다. 경찰에 의해 체포된 수상한 인물이 나중에 무죄로 방면되었어도 경찰에는 체포 기록이 남게 되고, 이것이 범죄 예측 프로그램에 입력된다. 그런데 경찰에 의한 체포가 인종적으로 편향되었다는 것은 여러 연구에 의해서 확인된 사실이다(Beckett, et al., 2006; Gelman, et al., 2007). 즉, 프레드폴 같은 알고리즘은 인종적으로 편향된 기준 값을 사용했으며, 처음부터 흑인

이나 히스패닉이 많이 사는 지역을 범죄의 핫스팟으로 판별하고 이에 순찰력을 집중했을 가능성이 큰 것이다(Selbst, 2018: 133-135).

세 번째 편향은 데이터의 특정한 특징을 선택하는 데에서 일어난다. 경찰은 모든 범죄를 다룰 수 없듯이, 프레드폴에도 모든 변수를 입력할 수 없다. 예를 들어 경찰은 지역, 범죄 유형, 시간에 집중할지, 아니면 날씨, 휴일, 알려진 범죄자의 거주지에 집중할지를 선택해야 한다. LA 경찰이 사용하는 프레드폴은 전자를 선택했고, 시카고 경찰이 사용하는 헌치랩(Hunchlab)이라는 알고리즘은 후자를 선택했다.[31] 한 쪽을 선택하면 다른 쪽이 희생되며, 따라서 그 결과는 원칙적으로 편향될 수밖에 없다. 특히 이 과정은 경찰이 데이터 브로커(data broker)로부터 데이터를 구매해서 사용할 때 더 커진다. 대부분의 데이터 브로커는 구매력을 파악하려 하는 기업을 위해 소비자의 개인정보 데이터를 수집·가공·재판매하는데(정용찬, 2015), 범죄가 아닌 구매력에 맞춰져 있는 데이터를 범죄 예측에 사용할 때 오류의 확률은 더 커진다(ibid.: 136). 마지막 편향은 "정확도"와 "공정성" 간의 불균형이다. 알고리즘은 정확한 예측을 위해서 구역을 정하고, 우편번호나 주소 정보를 사용한다. 이렇게 더 정확하게 범죄의 발생을 예측하려고 하면, 어떤 사람들은 본인의 의지와 무관하게 그 지역 근처에 거주하는 것 때문에 불이익을 받게 된다. 예측은 더 정확해질 수 있지만, 공정성은 더 손상되는 것이다(ibid.: 137).

31) 헌치랩은 필라델피아의 스타트업 Azavea에서 개발한 알고리즘이다. 범죄 기록과 같은 과거 통계 데이터 뿐만 아니라 인구 밀도, 인구 통계, 술집·교회·학교의 위치, 대중교통 허브, 지역 연고 스포츠 팀의 스케쥴, 달의 위상 주기 등의 현재 상태를 묘사하는 다양한 요인들을 고려하는 것으로 알려져 있다. Chammah, M. with Hansen, M. (2016. 2. 3) "Policing the Future: In the aftermath of Ferguson, St. Louis cops embrace crime-predicting software" The Verge https://www.theverge.com/2016/2/3/10895804/st-louis-police-hunchlab-predictive-policing-marshall-project

이런 알고리즘을 사용하면 다음과 같은 상황을 쉽게 예상할 수 있을 것이다. 프레드폴은 오늘 오후 4시에 특정 지역을 절도 핫스팟으로 예측했고, 이곳을 순찰하던 경찰은 주차장에서 두 명의 흑인 젊은이가 차의 내부를 들여다보면서 얘기를 나누고 있는 것을 발견했다. 경찰은 이들에게 검문을 요청해서 몸을 수색했고, 이 중 한 명에게서 대마초를, 다른 한 명에게서 드라이버를 발견하고 이들을 체포했다. 그런데 이런 경우 경찰의 검문, 몸수색은 공권력에 의한 부당한 사생활침해를 금지한 미국 수정헌법 제 4조에 위배되는가? 아마 '스몰 데이터'(small data) 시대였던 과거에는 이런 검문은 불법이라고 간주되었을 것이다. 체포된 사람들이 범죄로 간주될 수 있는 의심스러운 행동을 하지 않았기 때문이다. 미국의 경우 지금까지 법원은 경찰이 "상당한 이유"(probable cause)나 "합리적 의심"(reasonable suspicion)이 있을 경우에만 미국 수정헌법 제 4조를 어길 수 있다고, 즉 수상하다고 생각하는 사람들을 불러 세워서 수색을 할 수 있다고 판결했다. "합리적 의심"의 기준이 너무 낮다는 비판도 있지만 경찰이 "합리적 의심"을 했다는 것을 증명해야 하는 상황이 되면 그는 자신이 수상하다고 생각한 사람의 행동이 범죄와 연루되었다고 판단한 여러 가지 데이터 포인트(data point)를 제시해야 했다. 경찰이 제시한 데이터 포인트가 충분하지 않다고 해서 법원은 종종 경찰의 체포를 무효화하기도 했다. 그렇지만 빅데이터를 이용한 예측 프로그램들은 전반적으로 경찰이 "합리적 의심"을 했음을 지지하는 엄청난 데이터를 제공해 줄 수 있다. 과거에는 정당화되기 힘들었던 경찰의 검문과 수색이 충분히 정당화될 수 있다는 얘기다. 따라서 법학자들은 빅데이터와 예측 알고리즘이 그렇지 않아도 인종차별 같은 차별에 무력했

던 미국 수정헌법 4조를 더 약화시키는 결과를 낳고 있다고 판단하고 있다(Ferguson, 2015).[32]

미국 수정헌법 4조를 약화하면서 법적으로 제재를 받지 않고, 그렇지만 차별적인 결과를 낼 것이 거의 확실한 프레드폴 같은 알고리즘의 사용을 어떻게 저지할 수 있을까? 경찰이 이런 차별적인 알고리즘을 사용하는 것의 위법성을 따지기 힘들고, 또 여러 가지 내적, 외적 이유에서 이를 계속 사용한다면(Ferguson 2017), 이 차별적 체포를 줄이거나 무력화할 수 있는 방법 중 하나는 법원의 판결이다. 즉 빅데이터와 예측 알고리즘에 입각한 체포를 "합리적 의심"의 기준을 만족하지 못한 위법으로 판정하는 것인데, 퍼거슨(Ferguson, 2015)은 이를 위해 판사가 알고리즘이 예측한 수치를 요구하고, 역시 예측한 정확한 시간, 장소를 공개하라는 요청을 하는 것이 중요하다고 주장한다. 예를 들자면, '알고리즘이 예측했기 때문에 체포할 수 있었다'는 경찰의 진술과 '알고리즘이 절도 가능성을 2.06%로 예측했다'는 보고는 큰 차이가 나기 때문이다. 마찬가지로 정확한 핫스팟의 위치와 시간 정보를 경찰이 법원에 제공하는 것도 중요하다. 이렇게 되면 판사는 빅데이터라는 신탁의 예언 뒤에 가려진 충족되지 않은 이론상의 가정들, 속성들 사이의 상충적 관계(trade-off), 개인의 자결권 문제들을 비로소 볼 수 있게 되고, 이에 근거해서 경찰의 체포가 합리적 의심에 근거한 것이었는지를 더 종합적으로 판단할 수 있다.

32) 또 다른 선례로 미국 대법원이 "우범"(high-crime) 지역에서 경찰이 합리적 의심의 기준을 완화할 수 있다고 판결한 것을 들 수 있다. 예측 프로그램이 핫스팟으로 지정한 지역은 우범지역으로 충분히 해석될 수 있고, 이럴 경우에 이 지역에서는 경찰이 의심스러운 사람은 불러 세워서 몸수색을 할 수 있다(Perry et al., 2013: 13).

반면에 젤프스트(Selbst, 2018)는 경찰의 사용을 제한하는 방안을 제시했다. 미국 정부는 1970년에 제정된 국가환경정책법(National Environmental Policy Act: NEPA)에 개별사업(project)뿐만 아니라 정책, 계획, 프로그램까지 환경영향평가를 하게 규정했고, 사업의 주관 기관이 사업과 관련된 의사결정이 어떻게 내려지는지를 이해한다는 것을 보이는 환경영향평가서(Environmental Impacts Statement, EIS)를 제출하는 것을 의무화했다. 이와 비슷하게 젤프스트는 범죄 예측 알고리즘을 사용하는 경찰이 "알고리즘영향평가서"(Algorithm Impact Statements, AIS)를 만들어서 제출해야 한다고 주장했다. 그에 의하면 이 평가서에는 경찰과 같은 기관이 자신이 사용하는 알고리즘에 차별적 요소가 포함되어 있음을 인지한다는 사실과, 시민사회의 요청이 있을 때 이 사용에 대한 자료를 투명하게 공개하겠다는 선언이 담겨야 한다. 마치 환경영향평가가 산업 발전을 위해서 오염과 건강 침해는 어쩔 수 없다는 과거의 인식을 바꾸었듯이, 알고리즘영향평가 역시 데이터 산업의 발전을 위해서는 소수의 인권침해는 어쩔 수 없다는 현재의 인식을 바꿀 수 있다는 것이다.

판사에게 정량적 데이터를 요청하게 하자는 퍼거슨의 제안이나 알고리즘영향평가서를 작성하도록 하자는 젤프스트의 제안은 알고리즘이 계속 사용된다는 가정 하에 만들어진 제안이다. 이러한 가정은 인공지능 알고리즘이 인간의 판단보다 더 나은 판단을 할 것이라는 또 다른 가정에 근거하고 있다. 그런데 인공지능이 인간보다 더 나은 판단을 하는가? 인공지능 알고리즘이 사법이나 치안에서의 업무를 해결하는 가장 좋고 공평한 해결책인가? 아니면 가장 값싸고 효율적인 해결책인가? 알고리즘이 문제를 해결

하는 방식이 가장 (혹은 얼마나) 좋고 공평한지, 혹은 가장 (혹은 얼마나) 값싸고 효율적인지 판단하기 위한 조건은 무엇이며 그 조건은 누가 어떻게 정할 수 있는가? 이 문제에 대해서는 아직 충분한 연구가 부족하기 때문에 이를 본 논문에서 깊게 다루기 힘들다. 그렇지만 필자들은 논문의 결론에서 이에 대한 우리의 견해를 간략히 제시할 것이다.

5. 국가 안보 영역에서의 알고리즘 차별: 국경 관리 알고리즘의 위험 예측은 차별적인가?

최근 제주도에 입국한 예멘 난민 519명이 내전을 이유로 난민 신청을 한 것을 허용하느냐를 두고 찬반 논쟁이 크게 불거졌다.[33] 법무부 제주출입국·외국인청은 난민 여부를 심사하기 위해 "난민 전담 공무원 면접과 면접 내용에 대한 국내외 사실 검증, 국가 정황 조사, 테러 혐의 등 관계기관 신원 검증, 마약 검사, 국내외 범죄경력 조회, 중동 전문가 의견 반영 등"의 절차를 수행했다. 결과적으로 2018년 10월까지 362명에게 1년 기한의 인도적 체류 결정을 내렸으나, 아무에게도 난민 지위를 부여하지 않았다.[34] 난민

33) 실상 난민 문제는 비단 제주에 입국한 예멘인들에게 국한되는 문제는 아니며, 2013년 난민법 시행 이후로 한국 정부에 난민인정을 신청한 사람 수는 급격히 늘고 있다. 1994년부터 난민법 시행 이전인 2013년 말까지 신청자가 5,580명이었던 반면, 2013년 7월부터 2018년 5월까지 신청자 수는 34,890명에 달한다. 조선일보(2018. 7. 5) 「[난민쇼크] ③ '예멘 난민'은 시작에 불과하다는데」.

34) 세계일보(2018. 10. 18), 「"한명도 없다니, 당혹" vs "당장 추방을"… 예멘 난민심사 갈라진 민심」. 이런 찬반 논쟁을 계기로 법무부는 고시를 개정하여 2018년 6월부터 "제주도 무사증

신청자들에게 1년 기한의 인도적 체류를 허용한다는 결정은 무분별한 난민 신청을 통제하는 동시에 다소나마 난민 신청자들의 인권을 배려한 절충안으로 읽힐 수 있지만, 난민 인정을 찬성하거나 반대하는 집단 모두에게 비판을 받았다. 2019년에 난민 출국을 강행하면 새로운 종류의 논쟁이 시작되리라는 것을 쉽게 예측할 수 있다.

미국이나 유럽연합에서는 난민 심사, 국경에서의 출입국 심사에 인공지능 알고리즘이 광범위하게 사용되고 있다. 미국은 2008년 오바마 정부 이후 기준 매년 8만 명이 넘는 외국인이 난민 신청을 하며,[35] 이 중 3만 명 내외의 난민을 수용한다. 유럽은 난민의 수가 훨씬 더 많다. EU회원국가들에 망명 신청을 한 인원수는 2014년에는 62만 7,000만 명, 2015년 1월-10월 사이에는 시리아 난민 사태 등으로 인해 99만 5,800여명에 달했다.[36] 난민과 같은 사람들을 심사하는 문제는 타국의 시민들 중에서 자국이 수용할 만한 사람들과 그렇지 못한 사람들을 구별하는 문제이며, 이런 의미에서 국가 안보(national security)와 관련된 문제라고도 볼 수 있다.[37] 국가는 국민의 안전을 지킨다는 명목 하에 이런 다양한 구

입국불허국가"에 예멘을 추가하고, 이후 12개국을 더 추가하였다. 세계일보(2018. 7. 31) 「예멘 난민 논란에 정부 '제주도 입국 비자' 대상국 12개국서 24개국으로 늘려」.

35] 법무부 출입국·외국인정책본부 (2016) 「난민법 시행 3년에 대한 평가 및 향후 법 개정 방향: 합리적 난민인정절차 구축방안을 중심으로」 법무부 용역보고서. (발간등록번호: 11-12700 00-000930-01). pp. 104-105.

36] 대외경제정책연구원(2015). 「EU의 난민정책 현황과 향후 전망」. 『KIEP 오늘의 세계 경제』. Vol. 15, No. 36. 1-14쪽.

37] 안보학(security studies) 및 지정학(geographical politics) 연구자들은 이외에도 다양한 안보 안건들에서 알고리즘을 포함한 정보 기술의 사용이 불러오는 다양한 변화에 대해 주목해왔다. 예컨대 기술을 활용해 안보 위험을 예측하고 테러 위협을 선제 조치하는 방식이 안보 실행에 어떤 변화를 가져다주었는지(Aradau and Blanke, 2016; Barrett, 2017; De Goede et al., 2014; Wilson and Weber, 2008), 빅데이터와 알고리즘에 기반한 의사결정이 안보 실행에 관여된 조직

분 짓기를 수행하는데, 이 절에서는 국경을 넘는 입국자 중에서 테러리스트와 민간인을 구분하는 알고리즘과 이민과 출입국 관리에서 외국인들의 출입국 및 체류 정보를 관리하는 데이터베이스 및 자동화 기술의 문제 두 가지를 검토하겠다.[38] 이런 검토를 통해서 이질적인 타자들을 구분하는 시도들이 현대의 국가 및 사회의 정체성을 확인하는 작업들과 어떻게 연결되어 있는지를 알 수 있고, 이는 다시 예멘에서 온 난민 신청자들에게 1년간의 한시적 체류를 허용한 우리에게 실질적, 윤리적 시사점을 제공할 수 있다.

들과 실행들을 어떻게 재편성하고 있는지(Amoore, 2009a; Amoore and Raley, 2016; Dijstelbloem et al., 2011; Ulbricht, 2018), 개인들이 디지털화된 보안 실행을 통해 국가에 의해 어떻게 파편화되고 있는지(Amoore & Hall, 2009; Murphy and Maguire, 2015; Potzsch, 2015), 출입국 심사에서 활용되는 생체 인증(biometric) 기술이 어떤 새로운 프라이버시 이슈를 제기하는지(Amoore, 2006; Maguire, 2009; Muller, 2008; Walters, 2011), 시각화 기술이 기반시설과 국경을 시각적으로 이해하는 데에 어떤 변화를 가져다주었는지(Amoore, 2009b; Follis, 2017), 그리고 이러한 현상들을 어떻게 생명정치(biopolitics) 개념으로 이해할 수 있는지(Adey, 2009; Vaughan-Williams, 2010; Salter, 2006; Sparke, 2006) 등이 여러 연구자들이 주목하는 주제이다.

38) 안보 현장에서의 기술화(technologization) 추세는 2001년 9/11 테러 이후 뚜렷해진 것으로 이야기되는데, 그 저변에는 20세기 후반부터 개발되어 온 안보 기술의 발전이 존재한다(Ceyhan, 2008; Amoore & de Goede, 2005: 150). 80년대 초 베트남전에서 회수된 미국 보안 장비들이 1986년에 멕시코-미국 국경에 마약 밀수업자를 적발하기 위해 재설치된 것을 시점으로 하여, 90년대 불법 이민자들을 적발하기 위해 국경 통제를 강화하는 노력이 뒤따랐고, 이윽고 90년대 말 경에는 약물, 이민, 망명, 범죄, 테러리즘을 아우르는 안보 연속체(security continuum)가 성립되기에 이르렀다. 예컨대 국경 통제소에는 생체 인증 수집 장비뿐만 아니라, 다양한 센서, 움직임 탐지기, 장거리 야간 카메라 등이 설치되어 있다. 또한 비슷한 시기 유럽에서는 영국이 아일랜드공화군(IRA: Irish Republican Army)의 폭탄 테러에 맞서는 시도를 시작으로, 감시에 초점을 맞춘 범죄와의 전쟁이 확대되었다. 프랑스도 70년대부터 신분증명 문서의 안보화(securitization)를 시작으로, 문서화되지 않은 이민자를 잠재적인 신원 도용자(identity theft)로 의심하기 시작했다. 이후 기본 신분증명 정보들을 전산화함으로써 신분증의 안보 속성들을 개선해가면서, 프랑스는 이민, 국경들, 신원, 복지, 범죄, 테러리즘에 대한 다양한 이슈들로 안보 이슈들에 대한 대처 방안을 심화시켜갔다. 9/11 테러 이전까지는 20세기 후반의 사회적 맥락 속에서 개발되어온 안보 기술들이 소수 인구에 대한 파일럿 프로그램의 방식으로 시험되었다면, 9/11 이후에 이 기술들은 전 인구들을 대상으로 삼아 실행되게 되었다.

먼저 테러리스트의 판별 문제를 보자. 이 주제에 대한 연구들은 일국 내에서 사회적으로 차별적인 대우를 받아왔던 외국인 등의 소수자들이 알고리즘에 의한 입국 심사에서도 테러리스트로 분류될 가능성이 높다는 것을 보여준다. 루이스 아무르(Louise Amoore)와 마리케 드 후더(Marieke de Goede)는 "테러와의 전쟁"의 일환으로 데이터 감시 기반의 위험 관리 실행이 점차 널리 적용되고 있는 상황에서 특히 취약 집단이 이러한 새로운 감시의 주된 표적이 되고 있다고 주장했다(Amoore and de Goede, 2005). 이들이 드는 사례는 미국 국토안보부의 국경 안보 프로젝트인 "유에스비지트"(US VISIT)를 컨설팅하는 악센츄어(Accenture)사가 만든 "스마트보더솔류션"(Smart Border Solution)이라는 알고리즘이다. 스마트보더솔류션은 20개 이상의 현행 데이터베이스를 종합적으로 활용해서 입국자들 중에서 범죄자들과 테러리스트들을 걸러내는데,[39] 예를 들어 입국자가 아프가니스탄으로 국제전화를 걸었는지, 비행기 조종 훈련을 받았는지, 450kg의 비료를 구매했는지[40] 등의 여부를 바탕으로 그의 위험성을 판단한다. 이는 코드화된(encoded) 위험 프로필이 미래의 행동을 예측하는 기반으로 활용될 수 있음을 뜻한다. 달리 말해, 개인이 국경을 통과할 때 미리 그는 코드에 의해

39) 이 중 중요한 데이터베이스로는 IDENT(모든 외국 방문자의 생체정보 데이터를 저장하는 자동 지문 식별 시스템), ADIS(여행자의 입국/출국 데이터 저장), APIS(승객의 세관 정보), SEVIS(미국에 방문한 모든 외국/교환 학생의 데이터), IBIS(인터폴 및 국내범죄 데이터와 연결된 요주의 인물 목록), CLAIMS 3(복지 혜택을 청원하는 외국인 정보)이며, 그 외 지역 경찰, 재정 정보, 교육 기록 등이 참고되었다.

40) 비료는 9/11 테러 이전까지 미국 영토에 대한 가장 심각한 테러 사건이었던 1995년 오클라호마 시티 테러 사건의 폭탄으로 사용되었다. 미국 정부는 이 테러가 발생한지 16년 만인 2011년에 11kg 이상의 질산암모니아 비료를 구매하는 사람은 신고하는 법안을 통과시켰다. *USA Today*(2011. 8. 3) "Congress Seeks New Rules on Explosive Fertilizer."

쪼개지고 다시 특정하게 조합됨으로써 통과될지 혹은 거부될지가 정해진다는 것이다.[41] 요컨대, 다양한 데이터베이스를 종합하여 계량화된 위험 모델을 적용함으로써, 잘게 분해된 개인의 온갖 특징이 테러리스트에 대한 국가적/행태적/재정적 전제들에 부합할 경우 그는 위험인물로 "구성"된다.

두 번째로 이민 심사 등 일상적인 출입국 현장에서 발생하는 문제를 살펴보자. 이는 유럽 국가들의 국경을 통해 유입되는 이민자들이 알고리즘에 의해 어떻게 식별·분류·관리되고 있는지를 분석한 연구들에서 잘 보여진다. 데니스 브로더스(Dennis Broeders)는 국경 통제 목적으로 종래 사용되었던 유럽 권역의 데이터베이스들이 이제는 각국 내로 비정기적으로 드나드는 이민자들을 지속 관리하는 데에 사용되기 시작했으며, 이것이 유럽 전체에 새로운 "디지털 국경"(Digital Border)을 만들어 낸다고 주장했다(Broeders, 2007). 그는 쉥겐 지역(Schengen Area) 국가들로[42] 입국하는 것을 거부당한 사람 명단 등을 관리하는 "쉥겐정보시스템"(Schengen Information System, SIS), 난민 신청자의 신청 이력을 관리하는 유로댁(Eurodac)[43], 비자 소지자들의 입국과 과도한 체류기간 등을 관리하는 "비자정보시스템"(Visa Information System)이라는 세 가지 데이터베이스들이

41) 아무르와 드 후더는 개인이 통치되는 방식이 개인의 '훈육'으로부터 개인을 측정 가능한 위험 요인들의 집합으로 분해하는 '위험 관리'로 변모하고 있다고 하면서, 이를 가리켜 세계화로 인해 야기된 불확실성이 '표적화된 통치'(targeted governance) 방식으로 통치되고 있음을 지적했다. 세계화 시대의 표적화된 통치에 대해서는 Valverde and Mopas(2004) 참조.

42) 1985년 쉥겐 조약을 따르는 그리스, 네덜란드, 노르웨이, 덴마크, 독일 등 유럽의 26개 회원국을 일컫는다. 쉥겐 조약은 이들 26개 국가를 여행할 때는 마치 한 나라를 여행하는 것처럼 비자가 필요 없음을 확인한 조약이다.

43) "유럽 지문분석학 시스템"(European Dactylographic system)의 약자.

이민자들의 항시적 감시라는 역할을 수행한다고 지적했다. 뒤이어 글로우프치오스(Georgios Glouftsios)는 비자정보시스템이 유럽 연합 회원국들의 비자 관리 시스템으로 정착되면서, 쉥겐 비자를 신청하는 제삼국가 국민들의 인적사항이 수집되고, 여러 데이터베이스에서 공유되며, 비자 신청자의 입국 이후 디지털 기록들이라는 가상 정체성이 확립되는 등 새로운 안보 실천들이 확립되어 가고 있다고 주장했다. 특히 유럽연합이 이 시스템의 한계를 극복하기 위해 새로이 제안한 "출입국시스템"(Entry/Exit System)은 유럽연합이 비유럽인들의 실제 범죄 이력과 상관없이 그들이 어떤 국적을 지녔는가를 더 중요하게 생각하는 방식으로 변하고 있음을 보여준다(Glouftsios, 2018). 더 나아가 새롭게 제안되는 출입국시스템의 지지자들은 기존 비자정보시스템의 여러 한계를 지목하면서[44] 각 외국인에게 허용된 체류기간이 만료될 경우 각국의 치안 당국에 경보를 발급하는 신규 기능의 필요성을 역설했다. 이런 새로운 시스템은 국경 관리 실행이 개인의 기존 범죄 이력 여부로 위험도를 평가하는 예방 관점(precuation)으로부터, 기존 범죄 여부에 무관하게 기간 초과 체류자를 잠재적 위험인물로 간주하여 실제 위험이 발생하기 전 그 가능성을 차단하는 선제 관점(pre-emption)으로 옮겨가는 것을 보여준다(Glouftsios, 2018: 194). 글로우프치오스는 이러한 변화가 유럽 국가들이 비 유럽연합 국민들의 이동성(mobility) 자체를 위험스러운 것으로 재정의한다는 것을 시사하며, 따라서 우려할 만한 사례라고 평가했다.

44) 예를 들어 출국 심사원이 출국 외국인의 과도한 장기 체류를 적발하기 어렵다는 점, 각 회원국의 입출국 데이터를 중앙화된 데이터베이스에 저장하지 않기 때문에 효율적이지 않다는 점, 그리고 여전히 일부 제삼세계 국민들의 쉥겐 국가 영내 활동은 잘 관리되지 않는다는 점 등이 문제로 지목되었다 (Glouftsios, 2018: 195).

국가 안보 실천에서 알고리즘에 의한 차별은 위험한 외국인의 식별을 넘어서 주체로서의 “우리”가 객체로서의 “저들/그들/타자”를 어떻게 구분해내어 인지하는지를 규정한다. 루이스 아무르(Louise Amoore)는 국가 안보 실행의 알고리즘화가 단순히 사회의 국방화 혹은 안보의 상업화를 넘어서 상업, 군사, 민간 영역 사이를 오가며 그 경계를 흐림으로써, 전쟁의 폭력적 권력을 일상적이고 보이지 않는 것으로 만드는 결과를 낳는다고 해석했다. 그는 이러한 안보 실행의 알고리즘화를 가리켜 “알고리즘 전쟁”(algorithmic war)이라 부르면서, 이를 푸코가 지적한 “다른 수단에 의한 전쟁의 지속”(continuation of war by other means)의 한 형태로 간주했다(Foucault, 2003 [1976]: 16; Amoore, 2009a: 50에서 재인용). 푸코에 따르면, 정치적 권력은 고요한 전쟁을 지속하여, 권력 관계를 기관, 경제적 불평등, 언어, 심지어 개별 신체에 기입하는 방식으로 작동한다. 아무르는 알고리즘에 의한 보안 실행이 전쟁과 유사한 구조인 “적개심의 구조들”(architectures of enmity)을 동원하며, 푸코가 지적한대로 일상에서 자아/타자, 이곳/저곳, 안전/위험, 정상/위협의 끊임없는 구분 짓기를 수행함으로써 눈에 보이지 않는 고요한 전쟁을 수행하고 있다고 해석했다(Amoore 2009a: 51).

이런 끊임없는 일상의 피아식별은 상업, 군사, 민간 영역 사이의 경계를 흐리는 실천에 의해 지속된다. 아무르에 따르면, 국방기관들이 산업계와 유통계에서 개발되고 사용되어 온 데이터 마이닝 기법을 채택하면서 이 영역들 사이의 경계가 흐려지고 있다. 예컨대 수출 가공 공단(Export Processing Zone)이나 역외 사업장(offshore site)에서 사물을 추적(tracking)하여 뒤따라가는(tracing) 데에 사용해 온 기술들이 이제는 국경 통과 카드, 비자, 여권, 이민자 신분증

등 국가 사이의 인간의 움직임을 추적하고 그들을 타겟팅하여 배제하는 데에 고스란히 활용된다(Ibid.: 59). 상업 분야의 추적 기술의 대표적 사례인 무선 주파수 인식(Radio Frequency IDentification: 이후 RFID로 약칭)기술은 글로벌 경제에서는 곳곳에서 유동적으로 이동하는 사물들의 이동을 '읽고' 최신 도착 지점을 '기록한다'. 마찬가지로 미국 국토안보부는 2004년 컨설팅펌 악센츄어와 딜로이트(Deloitte), 그리고 필립스 반도체 사와 협력하여 RFID 기술을 미국 공항의 입국 터미널 등에 테스트한 이래로, 2006년에는 미국에 입국하는 외국인들의 비자 면제 여권에 RFID 기술을 도입했다(Ibid.: 57). 또한 미국과 멕시코 국경 횡단에 사용되는 수동 RFID 카드가 읽히면 각 개인의 위치 정보가 새롭게 갱신될 뿐만 아니라 해당 신원 정보는 과거 여행 패턴, 형사 범죄 전과, 테러리스트 감시 목록 등과 자동으로 대조된다. 이를 통해 알고리즘 전쟁은 순환을 허용하는 동시에 상이한 영역에서의 규범성들의 상호작용을 가능하게 하는 방식으로 입국자들의 이동성을 통치한다(Ibid.: 62). 즉, 알고리즘 전쟁은 '물자/인간/일용품 간의 이동이 자유롭고 개방된 세계 경제'라는 규범을 유지하면서도, 동시에 이를 위협하는 가상의, 실제의 적으로부터 이런 개방 세계가 "안전할 수 있다는 인상"(impression of securability)을 유지하는 방식으로 수행된다.

이전 절들에서 사법과 치안 분야에서의 알고리즘의 적용을 검토하면서 필자들은 인공지능 알고리즘이 판단의 공평성과 효율성에 대해 꼬리를 무는 질문을 던지고 있음을 살펴보았다. 이 절에서 검토한 국가 안보의 알고리즘화와 이를 둘러싼 차별은 그 질문의 폭을 더욱 넓혀, 위험한 인물, 시공간, 그리고 국적이 어떻게 정의되고 식별되고 있는지, 이 과정에서 평가받는 대상들을 공

정하게 판단한다는 것은 무엇이며, 동시에 그 판단을 효율적으로 내리는 것이 무엇을 뜻하는지에 대한 질문을 제기한다. 사법 실행에서 기존 범죄 통계가 개인의 미래 재범 확률에 투사되고, 치안 실행에서 기존 범죄 발생 통계가 특정한 시공간의 범죄 발생 확률에 투사되었던 반면, 안보 실행에서는 개인의 기존 범죄 발생 여부와 무관하게 국적이 개인의 미래 범죄 확률이 예측되거나 그 사람의 국경 간의 이동이 관리되어야 하는지 여부를 결정하는 데에 핵심적인 요인으로 작동한다.[45] 달리 말해 과거의 통계 데이터, 그리고 해당 사회 현상 및 그 데이터 수집에 연루된 사회적 선입견이 알고리즘에 의해 생생하게 소환되어 현재 행위자의 활동에 왕성하게 개입하여 현재의 양태를 만들어내고, 미래 시점에서 참고할 새로운 데이터를 만들어내는 데에 참여하고 있는 것이다.

6. 결론

지금까지 우리는 사법, 치안, 안보의 영역에서의 인공지능 알고리즘이 어떻게 민감한 그룹에 대한 차별적으로 보이는 판단을 낳고, 그 재생산에 기여하는지를 논한 연구들을 분석했다. 사실 인공지능 알고리즘이 낳는 차별은 이런 영역에 그치는 것이 아니다. 오히려 이런 분야보다 더 일상적으로 경험하는 것은 구글의 검색에

45) 국가 권력 실행에 도입되는 알고리즘과 소프트웨어를 개발하는 데에 연루되는 민간 사업체의 복잡도 역시 적용 분야에 따라 달라진다. 사법 실행에서는 노스포인트, 치안 실행에서는 프레드폴 같은 특정 기술 업체가 관여한 반면, 국가 안보 실행에서는 악센츄어와 딜로이트 같은 컨설팅펌과 기술회사의 컨소시엄처럼 한층 커진 규모의 연합체가 관여하고 있다.

서 사용되는 '단어 끼어 넣기'(word embedding), 네이버가 뉴스를 보여주는 알고리즘에서 보인 편향, 차별이다. 또한 구글 이미지 검색, 번역 서비스에서도 남/녀 차별이 존재한다는 사실이 지적되었다(Noble, 2018). 미국에서는 학생을 학교에 배정하는 알고리즘, 입사 서류를 평가하는 알고리즘, 면접을 수행하는 인공지능도 차별적이라는 주장이 제기되었고, 의료 알고리즘에도 비슷한 문제가 있을 수 있다는 지적도 나왔다. 인공지능이 우리의 일상생활에 더 많이 사용되면서, 알고리즘 속에 숨겨진 차별이 속속 드러나고 있는 것이다.

문제는 이런 빅데이터 인공지능 알고리즘이 우리 사회에 만연한 차별을 반영하는 데 그치지 않고 이를 영속시키고 증폭시킬 수 있다는 것이다. 차별적인 사회가 낳은 데이터를 가지고 인공지능 알고리즘은 차별적인 결과를 만들어 낸다. 그렇지만 우리는 인공지능의 내부를 들여다볼 수 없다. 즉 우리는 왜 알고리즘이 특정한 판단을 내린 것인지 그 내부 기제나 작동 원리를 볼 수 있는 허가를 받지 못하거나, 매우 개략적인 수준의 설명만 제공되어 어렴풋하게 추정하는 정도만 가능하거나, 상세한 방식이 공개되더라도 알고리즘을 동작시키는 패턴 인식이나 통계적 방법론 등을 전문적으로 이해하기 어렵다. 설령 이에 대한 전문가들의 분석이 있어도 알고리즘이 자동적으로 생성해낸 분류 기준들이 때때로 인간의 판단 기준이나 직관에 배치되기도 한다(Burrell, 2016). 이런 여러 이유 때문에 우리는 인간의 머리로는 도저히 분석할 수 없는 빅데이터를 순식간에 분석해 내는 인공지능 알고리즘에 의해서 산출되는 결과물들을 인간이 낳은 결과물보다 더 낫고, 더 공평하다고 생각한다.

알고리즘이 인간의 판단보다 더 나은 판단을 제공하는가? 컴파스 알고리즘은 재범률을 따지기 위해서 137개의 문항에 대한 답을 검토한다고 알려져 있다(Angwin et al., 2016). 반면에 사람은 보통 10개 내외의 항목을 검토해서 판단을 내린다고 알려져 있다. 사람들은 137개의 데이터 포인트를 연결하고 분석하고 종합해서 내린 결론이 10개의 데이터를 검토해서 얻어진 결론보다 더 낫다고 생각할 것이다. 그렇지만 2018년에 Science Advances 지에 출판된 연구는 컴파스의 판단이 법률 비전문가 인간의 판단보다 더 나은 것은 아니라는 점을 보여준다.[46] 재범 확률 예측의 정확도에서 인간의 판단이(흑인 피의자: 68.2%, 백인 피의자: 67.6%) 알고리즘의 판단보다(흑인 피의자: 64.9%, 백인 피의자: 65.7%) 미세하게 더 정확했고, 인간이 내린 판단에서 흑인에 대한 위양성의 비율도(흑인 피의자: 37.1%, 백인 피의자: 27.2%) 알고리즘의 판단에 비해 (흑인 피의자: 40.4%, 백인 피의자: 25.4%) 약간 줄어들었고, 흑인/백인의 위양성율의 차이도 줄어들었다. 인간의 판단에서 백인에 대한 위음성도(흑인 피의자: 29.2%, 백인 피의자: 40.3%) 알고리즘의 판단에 비해 (흑인 피의자: 30.9%, 백인 피의자: 47.9%) 줄어들었고, 백인/흑인의 위음성율의 차이도 줄었다. 모든 경우에 인간의 판단이 아주 조금이나마 더 바람직한 결과를 낳았던 것이다(Dressel and Farid, 2018). 이런 연구가 다른 영역으로도 확장된다면 우리 사회가 빅데이터, 인

46) 인간 참여집단과 알고리즘은 모두 2013-2014년 플로리다 주의 브로워드(Broward) 카운티의 데이터베이스에 대해 판단을 내렸다. 알고리즘은 7214 명의 피의자 전원에 대해 판단했고, 인간 참여자들은 랜덤하게 구성되어 전체 모집단 데이터와 유사한 통계 특성을 갖는 1000건의 피의자 사안에 대해 판단했다. 총 2회차의 실험 중 피의자의 인종을 고려하지 않은 1차 실험에 참여한 462명 중 400명의 답변이 유효답변으로 연구에 활용되었고, 피의자의 인종을 고려한 2차 실험의 449명의 참여자 중 400 명의 답변이 유효하게 연구에 활용되었다(Dressel and Farid, 2018).

공지능에 대한 과학주의적 환상을 걷어냄으로써 그 가능성과 역량에 대해 보다 현실적인 기대치를 설정하고 그 목표를 꾸준하게 달성해가는 데 일조할 수 있을 것이다.

인공지능 알고리즘의 문제가 드러난 뒤에 여러 인권단체들은 인공지능에 의한 차별을 비판하고 알고리즘에 대한 검사(auditing)를 주장해 왔다. 인공지능을 검사하는 방법으로는 알고리즘 자체를 들여다보는 방법이 있을 수 있는데, 이는 알고리즘에 대한 화이트박스 테스팅(white-box testing)이라고 알려진 방법이다. 그렇지만 기업이 지적 재산권을 주장하는 알고리즘을 공개하지 않는 경우가 많기 때문에, 이런 방법이 불가능한 경우가 많이 있다. 또한 공개가 가능하다고 가정하더라도, 설명하기 쉽도록 만든 예측 모델을 실제에 적용할 수 있는 수준으로 유연하게 개선시켜 가다보면 불가피하게 모델의 복잡성이 증가한다. 일례로 분류 알고리즘이 훈련 데이터에 지나치게 최적화되는 것을 피하고, 학습 과정에서 관찰되지 않았던 새로운 패턴에 적응하기 위해서는 최적화 수준을 의도적으로 제한하는 정규화(regularization), 그리고 훈련 데이터를 소집단으로 쪼개거나 데이터의 속성을 소집단으로 쪼개어 랜덤화(randomization)하는 과정이 필요하다. 하지만 이들 과정이 분류 알고리즘 단계부터 관여하므로, 결국 유연성을 얻은 대가로 모델에 대한 해석가능성이 희생되게 된다. 다시 말해 투명성을 일정 수준 달성하기 위해서는 어느 정도 모델의 효과성을 희생시켜야 하는 상충관계가 있고, 그 역도 역시 성립한다. 이런 이유 때문에 잘 작동하는 인공지능의 경우에는 만족할 만한 검사가 가능한 정도의 투명성을 확보하기가 힘들게 된다(김도훈, 2018: 17-22).

따라서 알고리즘 검사를 주장하는 인권단체는 인공지능 알고리즘에 대한 블랙박스 테스팅(black-box testing)에 의존하는 경우가 많다. 이는 다양한 집단과 계층을 표상하는 편향되지 않은 데이터를 확보해서 이를 알고리즘을 통해 처리하게 한 뒤에, 그 결과가 특정 집단이나 계층에 편향적이지 않고 중립적으로 나왔는지를 보는 것이다. 이 때 사용하는 비편향적 데이터는 연구자들이 활용할 수 있도록 자신의 메타정보, 입력 데이터, 알고리즘의 출력 데이터를 자발적으로 제공하는 사용자들을 모집하는 방식 등을 통해서 얻어질 수 있다(Sandvig et al., 2016). 하지만 이런 방법으로 중립적인 데이터를 얻는다는 게 현실적으로 쉽지 않은 경우가 많기 때문에, 이 역시 손쉬운 방법은 아니다.

따라서 알고리즘에 대한 검사는 사법권을 가진 정부기관이나 위원회에 의해서 이루어지는 것이 현실적으로 가능한 방법이다. 서론에서 언급했듯이 알고리즘의 투명성을 요구하는 뉴욕시의 법안 같은 조치가 이런 검사를 실질적으로 효과적인 것으로 만들어 준다. 비슷한 문제의식을 가지고 유럽연합이 2017년에 제정한 개인정보보호법(General Data Protection Regulation, GDPR)은 개인이 자동화된 프로파일링의 결정 대상이 되지 않을 권리를 보장하고 있으며, 인공지능 알고리즘이 내린 결정에 대해서 개인이 설명을 요구할 권리를 명시하고 있다. 덧붙여 알고리즘뿐만 아니라 공공 데이터가 공개되는 것도 알고리즘 설명 문제를 개선하는 데에 도움이 될 수 있다. 뉴욕시는 2012년 3월 세계 최초로 기술 정책이나 행정 명령보다 상위 수준에서 오픈 데이터 법(Open Data Law)을 지방법으로 제정하여[47] 2018년 말까지 단일 웹 포털에서 모든 공공

47] 전자신문(2013. 4. 24), 「금융의 중심 뉴욕, IT 메카까지 넘본다」.

데이터가 접근 가능해야 한다고 명시했다.[48] 그 결과물인 뉴욕시의 공공 데이터 포털에서는 비단 보건, 위생, 교통 등의 일상 데이터뿐만 아니라 경찰국의 시내 범죄, 공원 범죄, 범죄 진압 활동 등의 다양한 데이터가 공개되어 있다(Poirier, et al., 2018).[49] 한국의 경우 행정안전부 주도로 범정부 공공데이터 개방 창구인 공공데이터포털(data.go.kr), 그리고 정부 및 민간 사이의 데이터 활용 관련 협의단체인 "공공데이터 기반 사회혁신을 위한 오픈데이터포럼"(odf.or.kr) 등이 유사한 역할을 수행하고 있다.[50] 이러한 데이터 공개의 범위와 수가 늘어날수록 향후 치안, 사법 등에 관련한 알고리즘의 판단 결과를 두고 논쟁이 불거졌을 때 알고리즘의 판단 결과를 설명하는 것을 도울 수 있는 참고자료가 많아질 수 있다.

사회적 압력이 거세지면 문제의 해결에 시장의 논리가 긍정적으로 기능할 수도 있다. 최근에 인공지능에 의한 차별의 문제가 인공지능의 사회적 확산에 장애가 되는 심각한 문제로 부상함에 따라 이를 극복하려는 노력이 실리콘 밸리의 거대 기업과 스타트업에 의해서 이루어지고 있다. 알고리즘의 설명가능성(explainability)[51] 및 해석가능성(interpretability)을 담보하려는 노력, 모델 구축, 추론, 예측,

48) City of New York (n.a.) "Laws and Reports" https://opendata.cityofnewyork.us/open-data-law/

49) City of New York (n.a.) "NYC Open Data" https://opendata.cityofnewyork.us/ ; City of New York (n.a.) "Results matching of Policy Department (NYPD) | NYC Open Data" https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Police+Department+%28NYPD%29

50) 연합뉴스(2017. 7. 27), 「행안부, 오픈데이터포럼 발족」.

51) "알고리즘 설명가능성" 연구는 비단 기업뿐만 아니라 자동화된 의사결정이 인명 살상 등의 치명적 결과로 이어질 수 있는 DARPA 같은 미국 국방부의 기관에서 선도적으로 수행되고 있다. 이런 군부-정보산업의 연관의 배경에 관련해서는 군산복합체의 정보기술에 대한 영향(Leslie, 1993; Der Derian, 2009), 상업 영역과 군사·안보 영역에서 알고리즘 기술에 의한 인지처리적 상호 균질화(Amoore, 2009a) 등의 연구를 참조할 것.

의사 결정에서 상대적으로 보다 통합적인 프레임워크를 제공하는 베이지안 추론(Bayesian reasoning) 등의 분야에서부터 알고리즘 해석 가능성의 실질적인 개선을 이루어가려는 노력(Chakraborty et al., 2017: 5), 더 공정한 알고리즘을 만들려는 노력, 차별 같은 문제가 드러나면 이를 고칠 수 있게 만들려는 노력이 가시화되는 것이다. 타 경쟁사로 핵심 정보가 유출되지 않는 한도 내에서 기업이 알고리즘 설명 활동을 적극적으로 한다던가(Michael and Lupton, 2016; 오요한, 2018), 학술단체의 윤리 위원회에서 "자율 지능 시스템의 윤리에 대한 국제 이니셔티브"[52]와 "알고리즘 투명성 및 설명책임에 대한 성명서"[53] 등의 활동을 통해 학계와 산업계를 향한 권고 사항을 제안한다거나, 연구자들이 자생적으로 머신러닝의 공정성, 설명책임, 투명성(Fairness, Accountability, and Transparency in Machine Learning, FAT/ML)을 추구하는 공동체를[54] 만든 사례도 이러한 맥락에서 이해될 수 있다.[55]

52] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017).

53] ACM U.S. Public Policy Council and ACM Europe Policy Committee (2017).

54] www.fatml.org/

55] 한편 컴퓨터 과학 전문가들이 머신 러닝, 인공 지능의 공정성을 위해 조직한 학술단체 뿐만 아니라 기술 기업들이 사회·윤리적 문제에 공동대응하고자 결성한 연합체인 The Partnership on AI to Benefit People and Society, 혹은 그러한 목적으로 설립된 회사 OpenAI 등의 활동들이 공통적으로 전제한 가정에 대해 비판적인 견해도 존재한다. 대니얼 그린 (Daniel Greene) 등의 저자들의 지적에 따르면, 이러한 활동을 벌이는 전문가 단체들 및 기술기업들은 윤리와 가치들이 기술 및 디자인 전문가에 의해 다뤄지는 것이 최선이라는 기술 결정론적 시각을 전제로 하고 있다. 또한 이 전문가 단체들은 과학기술학에서의 비판적 방법론으로부터 차용한 절차적 요소들과 맥락적 프레임을 공유하면서도 정작 과학기술학자들의 궁극적 목표 중의 하나인 사회 정의와 균등성에 대한 추구를 논외로 두고, 사업 윤리(business ethics)를 "도덕적 배경"(moral background)으로 삼고 있다고 볼 수 있다 (Greene, Hoffmann & Stark, 2019).

그렇지만 이런 노력이 계속된다고 해도 절대적으로 공정한 알고리즘을 만들 수 없는 경우가 있음을 염두에 두어야 한다. 컴파스의 경우에 보듯이 알고리즘 공정성에 일종의 상충적 관계(trade-off)가 있음을 인식해야 한다. 즉 완벽한 공정성이란 존재하지 않는다는 얘기다. 그렇다면 결국 그때그때 맥락과 적용 분야에 맞게 공정성을 찾아가야 한다는 것이 중요해진다. 이 알고리즘이 어떤 목표를 가지고 있는지, 그 목표를 달성하기 위해서 어떤 기준을 정했는지, 사용하는 데이터가 비교적 공정하게 얻어진 것인지, 사회적으로 시급한 정의(justice)가 어느 영역에서 찾아져야 하는지, 특정하게 정의된 정의를 실현하기 위해 손쉽게 수치화될 수 있는 요인들은 무엇이며 이 요인들에 담지 못하는 정의의 포괄적인 사례가 무엇인지와 같은 문제를 고려해야 한다는 것이다. 알고리즘은 보편적이고 중립적인 것이 아니라 상황지어지고(situated), 사회기술적 맥락에 의해서 구성된(configured) 것임을 고려하면서 문제에 접근해야 한다.

우리나라에서는 아직 이런 알고리즘에 의한 차별이 큰 사회적 문제가 되지 않는다. 그렇지만 이런 알고리즘이 사용이 안 되는 것은 아니다. 잘 드러나지 않지만 정부 규제기관이나 금융기관에서는 여러 종류의 자동화된 결정 알고리즘을 쓰고 있다. 특히 정부가 4차 산업혁명을 추진하고 스타트업을 장려하면서 이런 알고리즘이 확산될 것이고, 이것이 낳는 차별의 문제가 표면 위로 부상할 것임이 확실하다. 2018년부터 대기업과 중견기업에서 취업 면접에 인공지능 알고리즘이 도입되기 시작했다. 대검찰청과 경찰청에서는 각각 사법 알고리즘과 범죄 예측 알고리즘을 개발하기 시작했다.[56] 미국과 유럽을 중심으로 인공지능 알고리즘이 불러일

56) 전자신문(2016. 8. 9), 「인공지능으로 범죄 막을 수 있을까…대검, 범죄예방체계 컨설팅

으키는 차별에 대한 사회적 논란을 분석한 본 연구는 가까운 미래에 이런 문제가 우리에게 닥치기 전에 우리가 이를 어떻게 사전 대응할 수 있는지에 대한 여러 가지 정책적 시사점을 제공한다. 남녀간, 연령간, 지역간, 자산 및 소득 계층간의 갈등, 편견, 혐오의 뿌리가 깊을 뿐만 아니라 그 가지가 넓게 퍼져있고, 점차 다문화사회로 변하면서 인종간의 갈등도 표면화되는 한국 사회에서 인공지능 알고리즘의 확산은 사회적 차별을 반영하고 증폭시킬 수 있다. 알고리즘의 차별 가능성에 대해서 적극적으로 개입하고 알고리즘의 반민주주의적 사용을 반대하고 저지하는 "알고리즘 시민권"(algorithmic citizenship)의 형성이 어느 때보다도 더 절실하다.

착수」; 매일경제(2016. 12. 29), 「범죄 용의자 AI가 찾는다」.

참고문헌

- 강준만 (2014), 「왜 흑인이 사는 빈곤층 거주 지역에 붉은 줄을 긋는가? - redlining」, 『인문학은 언어에서 태어났다』, 서울 : 인물과사상사, 288-290쪽.
- 김기태·정재관 (2018), 「알고리즘이 이용자에게 미친 영향과 방법론적 쟁점: 알고리즘 편향(bias)과 개인정보 보호(privacy protection)를 중심으로」, 정보통신정책연구원, 『사이버 커뮤니케이션학회 공동주최 지능정보화 이용자 보호 학술 세미나 '알고리즘 시대 이용자 연구 어떻게 할 것인가' 자료집 (2018.6.14)』, 17-35쪽.
- 김도훈 (2018), 「알고리즘 책임성 논의와 알고리즘에 대한 이해」, 정보통신기술진흥센터 (IITP), 『주간기술동향』, 제16권, 제1호, 제1848호 (2018. 5. 30). 14-28쪽. [http://www.itfind.or.kr/WZIN/jugidong/1848/file41364086823808863 23-184802.pdf]
- 안중호·양지윤 (2006), 「기업 거버넌스 측면에서의 IT 거버넌스」, 『경영정보논총』, 제16권, 제1호, 97-119쪽.
- 에릭 브린욜프슨, 앤드루 맥아피, 이한음 번역 (2014), 『제2의 기계 시대 - 인간과 기계의 공생이 시작된다』, 서울 : 청림출판. [Brynjolfsson, E., and A. McAfee (2016), *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York, NY: W. W. Norton & Company.]
- 오요한 (2018), 「알고리즘 조사 활동과 그 제약사항의 설정: 네이버 실시간급상승검색어 알고리즘에 대한 검증 논쟁을 중심으로」, 서울대학교 대학원 석사학위논문.
- 정용찬 (2015). 「빅데이터 산업과 데이터 브로커」, 『KISDI Premium

Report』, 15-04. 진천: 정보통신정책연구원. 1-23쪽.

- 캐시 오닐, 김정혜 번역 (2017), 『대량살상 수학무기: 어떻게 빅데이터는 불평등을 확산하고 민주주의를 위협하는가』, 서울 : 흐름출판. [O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, NY: Crown.]
- 프랭크 파스콸레, 이시은 번역 (2016), 『블랙박스 사회: 당신의 모든 것이 수집되고 있다』, 안양 : 안티고네. [Pasquale, F. (2015), *The Black Box Society: The secret algorithms that control money and information. Cambridge*, MA: Harvard University Press.]
- 필립 K. 딕, 조호근 번역 (2015), 『마이너리티 리포트』, 서울 : 폴라북스 [Dick, P. K. and Triptree Jr., J. (2002), *The Minority Report and Other Classic Stories*, New York, NY: Citadel.]
- ACM (Association for Computing Machinery) U.S. Public Policy Council and ACM Europe Policy Committee (2017), "Statement on Algorithmic Transparency and Accountability", Accessed on: Oct. 14, 2018. [Online]. Available: https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf
- Adey, P. (2009), "Facing Airport Security: Affect, Biopolitics, and the Preemptive Securitisation of the Mobile Body", *Environment and Planning D: Society and Space*, Vol. 27, No. 2, pp. 274-295.
- Amoore, L. (2006), "Biometric Borders: Governing Mobilities in the War on Terror", *Political Geography*, Vol.25, No. 3, pp. 336-51.
- Amoore, L. (2009a), "Algorithmic War: Everyday Geographies of the War on Terror", *Antipode*, Vol. 41, No.1, pp. 49-69.
- Amoore, L. (2009b), "Lines of Sight: on the Visualization of Unknown

Futures", *Citizenship Studies*, Vol 13, No. 1, pp. 17-30.

- Amoore, L. and Hall, A. (2009), "Taking People Apart: Digitised Dissection and the Body at the Border", *Environment and Planning D: Society and Space*, Vol. 27, No. 3, pp. 444-464.
- Amoore, L. and de Goede, M. (2005), "Governance, Risk and Dataveillance in the War on Terror", *Crime, Law and Social Change*, Vol. 43, No. 2-3, pp. 149-173.
- Amoore L. and Raley, R. (2016), "Securing with Algorithms: Knowledge, Decision, Sovereignty", *Security Dialogue*, Vol. 48, No. 1, pp.3-10.
- Angwin, J. and Larson, J. (2015), "The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review", *ProPublica*. [https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review]
- Angwin, J. and Larwon, J. (2016), "Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say", *ProPublica*. [https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say]
- Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016.05.23.), "Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks", *ProPublica*. [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing]
- Angwin, J., and Larson, J. (2016.07.29), "ProPublica Responds to Company's Critique of Machine Bias Story", *ProPublica*. [https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story]

- Aradau, C. and Blanke, T. (2016), "Politics of Prediction: Security and the Time/Space of Governmentality in the age of Big Data", *European Journal of Social Theory*, Vol. 20, No. 3, pp. 373-391.
- Aradau, C. and Blanke, T. (2018), "Governing Others: Anomaly and the Algorithmic Subject of Security", *European Journal of International Security*, Vol. 3, No.1, pp. 1-21.
- Asaro, P. M. (2013), "The Labor of Surveillance and Bureaucratized Killing: New Subjectivities of Military Drone Operators", *Social Semiotics*, Vol. 23, No. 2, pp. 196–224.
- Barocas, S. and Selbst, A. D. (2016), "Big Data's Disparate Impact", *California Law Review*, Vol. 104, No. 3, pp. 671-732.
- Barrett, L. (2017), "Reasonably Suspicious Algorithms: Predictive Policing at the United States Border", *NYU Review of Law & Social Change*, Vol. 41, pp. 327-363.
- Beckett, K., Nyrop, K., & Pfingst, L. (2006), "Race, drugs, and policing: Understanding disparities in drug delivery arrests." *Criminology*, Vol. 44, No. 1, pp. 105-137.
- Berk, R. Hoda, H. Shahin, J. Michael K. and Aaron R, (2018), "Fairness in Criminal Justice Risk Assessments: The State of the Art", *Sociological Methods & Research*, Vol. 20, No. 10, pp. 1-42.
- Burrell, J. (2016), "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data & Society*, Vol. 3, No. 1.
- Ceyhan, A. (2008), "Technologization of Security: Management of Uncertainty and Risk in the Age of Biometrics", *Surveillance & Society*, Vol. 5, No. 2, pp. 102-123.
- Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M.,

Cerutti, F., Srivastavaz, M., Preecey, A., Julieryy, S., Rao, R. M., Kelley, T. D., Brainesx, D., Sensoyk, M., Willis, C. J., & Gurram, P. (2017), "Interpretability of deep learning models: a survey of results." In *IEEE Smart World Congress 2017 Workshop: DAIS.*

- Constantiou, I. D. and Kallinikos, J. (2015), "New Games, New Rules: Big Data and the Changing Context of Strategy", *Journal of Information Technology*, Vol. 30, No. 1, pp. 44-57.
- Corbett-Davies, S. Pierson, E. Feller, A. Goel, S. and Hug, A. (2017), "Algo rithmic Decision Making and Cost of Fairness", KDD '17 (Kno wledge Discovery and Data mining). [https://arxiv.org/abs/1701.0 8230]
- Crawford, K. (2013.05.10), "Think Again: Big Data - Why the Rise of Machines isn't All it's Cracked Up to Be", *Foreign Policy*. [https://foreignpolicy.com/2013/05/10/think-again-big-data]
- Curtis, N. (2016), "The Explication of the Social: Algorithms, Drones and(counter-) Terror", *Journal of Sociology*, Vol. 52, No.3, pp. 522-536.
- d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017), "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification." *Big data*, Vol. 5, No. 2, pp. 120-134.
- Dastin, J. (2018), "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, (2018. 10. 9) https://www.reuters.com/ar ticle/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-r ecruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
- de Goede, M., Stephanie S., and Hoijtink, M. (2014), "Performing preemption", *Security Dialogue*, Vol. 45, pp. 411-422.

- Dennis, B. (2007), "The New Digital Borders of Europe: EU Databases and the Surveillance of Irregular Migrants", *International Sociology*, Vol. 22, No. 1, pp. 71-92.
- Der Derian, J. (2009), *Virtuous war: Mapping the military-industrial-media-entertainment-network*. Routledge.
- Dieterich, W., Mendoza, C., and Brennan, T. (2016), "COMPAS Risk Scale s: Demonstrating Accuracy Equity and Predictive Parity", North pointe Inc. Research Department. Published by Northpointe (20 16. 7. 8.). [http://www.northpointeinc.com/northpointe-analysis]
- Dijstelbloem, H., Meijer, A., and Brom, F. (2011), "Reclaiming Control over Europe's Technological Borders", in H. Dijstelbloem and A. Meijer eds., *Migration and the new Technological Borders of Europe*, Palgrave Macmillan, pp. 170-185.
- Dourish, P. (2016), "Algorithms and Their Others: Algorithmic Culture in Context", *Big Data & Society*, Vol 3, No. 2, pp. 1-11.
- Dressel, J. and H. Farid. (2018), "The Accuracy, Fairness, and Limits of Predicting Recidivism", *Science Advances*, Vol. 4, No.1 [DOI: 10.1126/sciadv.aao5580].
- Ferguson, A. G. (2015), "Big Data and Predictive Reasonable Suspicion", *University of Pennsylvania Law Review*, Vol. 163, No. 2, pp. 327-410.
- Ferguson, A. G. (2017), *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement.* NYU Press.
- Follis, K. S. (2017), "Vision and Transterritory: The Borders of Europe", *Science, Technology, & Human Values.* Vol. 42, No. 6, pp. 1003-1030.

- Foucault, M. (2003), *Society Must be Defended: Lectures at the C`ollege de France* New York: Picador.
- Friedler, S., Scheidegger, C., and Venkatasubramanian, S. (2016), "On the (im)possibility of Fairness", [https://arxiv.org/abs/1609.07236]
- Gelman, A., Fagan, J., & Kiss, A. (2007), "An analysis of the New York City police department's 'stop-and-frisk' policy in the context of claims of racial bias." *Journal of the American Statistical Association,* Vol. 102, No. 479, pp. 813-823.
- Gibbs, S. (2015.07.08), "Women less likely to be Shown Ads for High-paid Jobs on Google, Study Shows", *The Guardian.* [https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study]
- Gillespie, T. (2014), "The Relevance of Algorithms", in T. Gillespie, Boczkowski, P.J., and Foot, K.A. eds., *Media Technologies: Essays on Communication, Materiality, and Society*, pp. 167-193, MIT Press.
- Glouftsios, G. (2018), "Governing Circulation Through Technology within EU Border Security Practice-networks", *Mobilities*, Vol. 13, No.2, pp. 185-199.
- Greene, D., Hoffmann, A. L., and Stark, L. (2019), "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning", *The 52nd Annual Hawaii International Conference on System Sciences (HICSS)*, Maui, HI.
- Hacking I. (1995), "The looping effect of human kinds", In Sperber. D. Premack, D ., and Premack, A. J. eds., *Causal cognition: a multidisciplinary debate*, Oxford: Oxford University Press, pp.

351-83.

- Hand, D. J. (2006), "Classifier Technology and the Illusion of Progress", *Statistical science*, Vol. 21, No.1, pp. 1-14.
- Hardt, M., Price, E., and Srebro, N. (2016), "Equality of Opportunity in Supervised Learning", *Advances in Neural Information Processing Systems 29* (NIPS 2016), [https://arxiv.org/abs/1610.02413]
- The IEEE (Institute of Electrical and Electronics Engineers) Global Initiative on Ethics of Autonomous and Intelligent Systems (2017), "Ethic ally aligned design: A vision for prioritizing human well-being w ith autonomous and intelligent systems", version 2. Accessed on: Oct. 14, 2018. [http://standards.ieee.org/develop/indconn/ec/ead_v 2.pdf]
- Jasanoff, S. (2017), "Science and democracy", in Hackett, E. J., Amsterdamska, O., Lynch, M., and Wajcman, J. Eds. *The handbook of science and technology studies*, The Fourth Edition, pp. 259-288, Cambridge, MA: MIT Press.
- Jeandesboz, J. (2014), "EU Border Control: Violence, Capture and Apparatus", In Jansen, Y., Celikates, R., and De Bloois, J. eds., *The Irregularization of Migration in Europe*, London: Rowman & Littlefield International. pp. 87–103,
- Kahn, C. (2011.11.26.), "At LAPD, Predicting Crimes Before They Happen", *National Public Radio*. [https://www.npr.org/2011/11/26/1427580 00/at-lapd-predicting-crimes-before-they-happen]
- Kirchner, L. (2017.12.18), "New York City Moves to Create Accountability f or Algorithms", *Propublica* [https://www.propublica.org/article/new -york-city-moves-to-create-accountability-for-algorithms]
- Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016.09.11), "Inherent

Trade-Offs in the Fair Determination of Risk Scores", in *Proc. 8th Conf. on Innovations in Theoretical Computer Science* (ITCS), 2017. [https://arxiv.org/abs/1609.05807]

- Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., and Yu, H. (2017), "Accountable algorithms", *University of Pennsylvania Law Review*, Vol. 165, No.3, pp. 633-705.
- Kuhn, T. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.
- Larson, J. and Angwin, L. (2016.07.29), "Technical Response to Northpointe", *ProPublica.* [https://www.propublica.org/article/technical-response-to-northpointe]
- Leeang, M. M., Rutjes, A. W., Reitsma, J. B., Hooft, L., and Bossuyt, P. M. (2013), "Variation of a Test's Sensitivity and Specicity with Disease Prevalence", *Canadian Medical Association Journal*, Vol. 185, No. 11, pp. 537-544.
- Leslie, S. W. (1993), *The Cold War and American science: The military-industrial-academic complex at MIT and Stanford*. Columbia University Press;
- Liu, H. and Motoda., H. (2012), *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media.
- Liu, Y. (2017.01.17), "The Accountability of AI — Case Study: Microsoft's Tay Experiment", *Medium.* [https://chatbotslife.com/the-accountability-of-ai-case-study-microsofts-tay-experiment-ad577015181f]
- Maguire, M. (2009), "The Birth of Biometric security", *Anthropology Today*, Vol. 25, No. 1, pp. 9-14.
- Martin, B. and Richards, E. (1995), "Scientific Knowledge, Controversy, and

Public Decision Making", In Jasanoff, S., Markle, G. E., Petersen, J. C., and Pinch, T. eds., *Handbook of Science and Technology Studies*. The Second Edition, pp. 506-526, Cambridge, MA: MIT Press.

- Michael, M., & Lupton, D. (2016), "Toward a manifesto for the 'public understanding of big data'." *Public Understanding of Science*, Vol. 25, No. 1, pp. 104-116.
- Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita. G. E. (2011), "Self-exciting Point Process Modeling of Crime", *Journal of the American Statistical Association*, Vol. 106. (Issue 493), pp. 100-108.
- Muller, B. J. (2008), "Travellers, Borders, Dangers: Locating the Political at the Biometric Border", in M. B. Salter Ed., *Politics at the Airport*, University of Minnesota Press, pp. 127-143.
- Murphy, E. and Maguire, M. (2015), "Speed, Time and Security: Anthropological Perspectives on Automated Border Control", *Etnofoor*, Vol. 27, No. 2, pp. 157-177.
- Nelkin, D. (1995), "Science Controversies: The Dynamics of Public Disputes in the United States", In Jasanoff, S., Markle, G. E., Petersen, J. C., and Pinch, T. eds., *Handbook of Science and Technology Studies*. The Second Edition, pp. 444-456, Cambridge, MA: MIT Press.
- Noble, S. U. (2018), *Algorithms of Oppression: How search engines reinforce racism*. New York: NYU Press.
- Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., and Hollywood, J. S. (2013), *Predictive policing: The Role of Crime Forecasting in Law Enforcement Operations*, Rand Corporation.

- Poirier, L., Hidalgo, N., & Goldman, E. (2018), "Data Design Challenges and Opportunities for NYC Community Boards." *BetaNYC.* [https://beta.nyc/publications/data-design-challenges-and-opportunities-for-nyc-community-boards/]
- Potzsch, H. (2015), "The emergence of Border: Bordering Bodies, Networks, and Machines", *Environment and Planning D: Society and Space*, Vol. 33, No. 1, pp. 101-118.
- Salter M. B. (2006), "The Global Visa Regime and the Political Technologies of the International Self: Borders, Bodies, Biopolitics", *Alternatives: Global, Local, Political*, Vol. 31, pp. 167-189.
- Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014), "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms", In Data and Discrimination: Converting Critical Concerns into Productive: A preconference at the 64th Annual Meeting of the International Communication Association. Seattle, WA, 2014.
- Schlehahn, E., Aichroth, P., Mann, S., Schreiner, R., Lang, U., Shepherd, I. D. H., and Wong, B. L. W. (2015), "Benefits and Pitfalls of Predictive Policing", *Proceedings of 2015 European Intelligence and Security Informatics Conference* (EISIC 2015). [https://ieeexplore. ieee.org/document/7379738]
- SearchEnterpriseAI (2015), "algorithmic transparency", by Matthew Haughn. https://searchenterpriseai.techtarget.com/definition/algorithmic-transparency
- SearchEnterpriseAI (2017), "algorithmic accountability", by Matthew Haughn. https://searchenterpriseai.techtarget.com/definition/algorithmic-acco

untability

- Selbst, A. D. (2018), “Disparate Impact in Big Data Policing”, *Georgia Law Review*, Vol. 52, pp. 109-195.
- Simonite, T. (2018.11.01), “When It Comes to Gorillas, Google Photos Rem ains Blind”, *Wired.* [https://www.wired.com/story/when-it-comes-t o-gorillas-google-photos-remains-blind]
- Sparke, M. B. (2006), “A Neoliberal Nexus: Economy, Security and the Biopolitics of Citizenship on the Border”, *Political Geography*, Vol. 25, No. 2, pp. 151-180.
- Toole, J. L., Eagle, N., and Plotkin, J. B. (2011), “Spatiotemporal correlations in Criminal Oense Records.” *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 4, pp. 1-38.
- Rubin, J. (2010.08.21.), “Stopping Crime Before It Starts”, *Los Angeles Times.* [http://articles.latimes.com/2010/aug/21/local/la-me-predictcrime-20 100427-1]
- Ulbricht, L. (2018), “When Big Data Meet Securitization. Algorithmic Regulation with Passenger Name Records”, *European Journal for Security Research*, Vol. 3, No.2, pp. 1-23.
- Valverde, M. and Mopas, M. (2004), “Insecurity and the Dream of Targeted Governance”, in W. Larner and W. Walters eds., *Global Governmentality: Governing International Spaces*, London: Routledge. pp. 245-62.
- Vaughan-Williams, N. (2010), “The UK Border Security Continuum: Virtual Biopolitics and the Simulation of the Sovereign Ban”, *Environment and Planning D: Society and Space*, Vol. 28, pp. 1071-1083.

- Walters, W. (2011), “Rezoning the global: technological zones, technological work and the(un-)making of biometric borders”, in V. Squire Ed., *The Contested Politics of Mobility: Borderzones and Irregularity*, Routledge, pp. 51-73,
- Weisburd, D. (2008), “Place-based policing”, *Ideas in American policing*, Vol. 9 (January 2008), pp. 1-15.
- Wilke, C. (2017), "Seeing and Unmaking Civilians in Afghanistan: Visual Technologies and Contested Professional Visions", *Science, Technology, & Human Values*, Vol. 42, No. 6, pp. 1031-1060.
- Williams, B. A., Brooks, C. F., and Shmargad, Y. (2018), “How Algorithms Discriminate Based on Data They Lack”, *Journal of Information Policy*, Vol. 8, No. 1, pp. 78-115.
- Wilson, D. and Weber, L. (2008), “Surveillance, Risk and Preemption on the Australian Border”, *Surveillance & Society*, Vol. 5, No. 2, pp. 124-141.
- Wilson, M. (2017), “Algorithms(and the) Everyday”, *Information, Communication & Society*, Vol. 20, No, 1, pp. 137-150.
- Wirth, N. (1975), *Algorithms + Data Structures = Programs.* Englewood Cliffs, Prentice-Hall.

논문 투고일	2018년 10월 14일
논문 수정일	2018년 11월 05일
논문 게재 확정일	2018년 11월 19일

Does Artificial Intelligence Algorithm Discriminate Certain Groups of Humans?

Oh, Yoehan · Hong, Sungook

ABSTRACT

The contemporary practices of Big-Data based automated decision making algorithms are widely deployed not just because we expect algorithmic decision making might distribute social resources in a more efficient way but also because we hope algorithms might make fairer decisions than the ones humans make with their prejudice, bias, and arbitrary judgment. However, there are increasingly more claims that algorithmic decision making does not do justice to those who are affected by the outcome. These unfair examples bring about new important questions such as how decision making was translated into processes and which factors should be considered to constitute to fair decision making. This paper attempts to delve into a bunch of research which addressed three areas of algorithmic application: criminal justice, law enforcement, and national security. By doing so, it will address some questions about whether artificial intelligence algorithm discriminates certain groups of humans and what are the criteria of a fair decision making process. Prior to the review, factors in each stage of data mining that could, either deliberately or unintentionally, lead to discriminatory results will be discussed. This paper will conclude with implications of this theoretical and practical analysis for the contemporary Korean society.

Key terms | artificial intelligence, algorithm, big data, discrimination, COMPAS algorithm, PredPol algorithm, border control algorithm