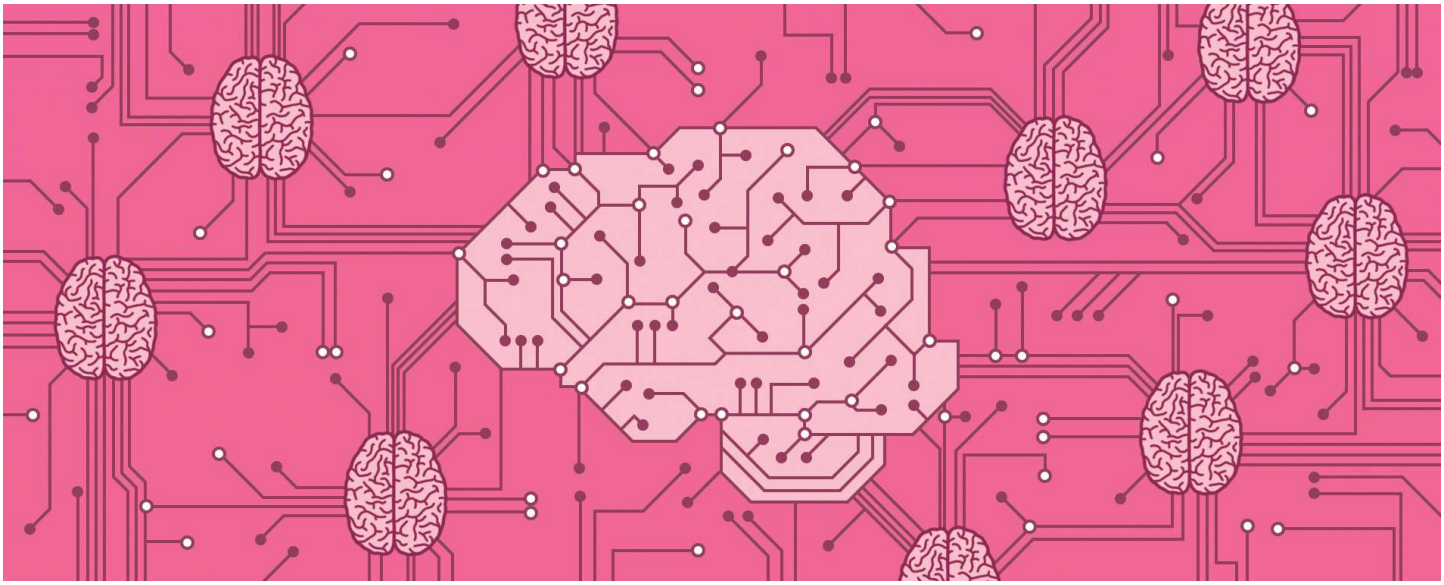
과학기술학(STS)

인공지능은 인간을 차별하는가?

2019년 8월 23일

홍성욱



2017년, MIT 대학교에서 출판하는 테크놀로지 리뷰Technology Review는 구글의 인공지능AI 연구 책임자 존 지아난드리아John Gianandrea와 인터뷰를 했다. 이 인터뷰에서 지아난드리아는 인공지능이 인간의 편견을 배우고 있다고 하면서, 인공지능의 진짜 위험은 편견에서 유래할 것이라고 예견했다. 당시 테크놀로지 리뷰Technology Review는 이 인터뷰에 대해서 “킬러 로봇은 잊어라. 편견이 진짜 위험이다”는 제목을 뽑았다. 대체 인공지능에 무슨 일이 일어나고 있는 것일까?

지난 10여 년 동안 신경망 네트워크, 머신러닝, 딥러닝 등의 방법이 인공지능에 도입되면서 인공지능의 효율성과 정확성은 놀라울 정도로 높아졌고 응용되는 영역도 넓어졌다. IBM의 인공지능 왓슨은 <Jeopardy!>의 챔피언을 이긴 뒤에 암 진단과 연금 투자 등에 응용되기 시작했으며, 구글 딥마인드의 알파고AlphaGo는 이세돌 국수에 승리를 거뒀다. 구글의 자율주행자동차는 미국 네바다주에서 최초로 라이센스를 획득했다.

그림1 라이센스를 획득한 구글의 자율주행 자동차

법률 분야에 도입된 인공지능은 하나의 사안에 대해서 여러 복잡한 법안들을 찾아서 자료를 만드는 일을 초급 변호사보다 훨씬 더 효율적으로 수행하며, 인공지능에 의한 번역 서비스도 과거에 비하면 월등히 개선되었고, 얼굴 인식과 사물 인식 분야에서 인공지능은 거의 사람의 수준에 도달했다. 인공지능 알고리즘은 의료, 법률, 금융 등의 분야뿐 만이 아니라 채용, 치안, 사법, 교육, 공공행정, 감사, 국경 관리, 이민 및 난민 관리 등의 분야에도 도입되어 인간의 판단을 대체하거나 보완하기 시작했다. 두 번의 겨울을 거치고, 인공지능의 세 번째 르네상스가 찾아온 것이었다.

그렇지만 같은 시기에 인공지능의 차별 문제가 대두되기 시작했다. 2013년에 한 연구자는 구글의 검색에서 사용하는 "자동완성auto complete"기능이 매우 성차별적임을 폭로했다. 예를 들어 구글에 "남성은 자격이 있다man deserves"는 단어를 치면 "남성은 높은 임금을 받을 자격이 있다", "남성은 존경받을 자격이 있다"는 단어가 따라 나오지만, "여성은 자격이 있다"는 단어를 치면 "여성을 맞을 자격이 있다"와 같은 매우 여성 혐오적인 단어들이 따라 나온다는 것이었다. 해당 기능을 사용하는 남성들이 이런 단어의 조합을 많이 검색하기 때문이었다.

그림2

이런 문제는 시간이 갈수록 늘어났다. 2015년에 구글에서 사물 인식 프로그램으로 출시한 “구글 포토Google Photo” 카메라 앱app은 흑인 커플의 얼굴을 고릴라라고 인식했다. 이는 큰 사회적 논란을 불러일으켰고, 구글은 이에 대해 바로 사과하고 시정을 약속했다. 그렇지만 2018년에 발표된 해결책은 “고릴라”를 검색 인덱스에서 지우는 것이었다. 같은 해에 구글의 광고가 여성에 비해 남성들에게 높은 보수의 자문, 관리 직종 등 상대적으로 고급 취업 광고를 내보낸다는 사실도 드러났다.

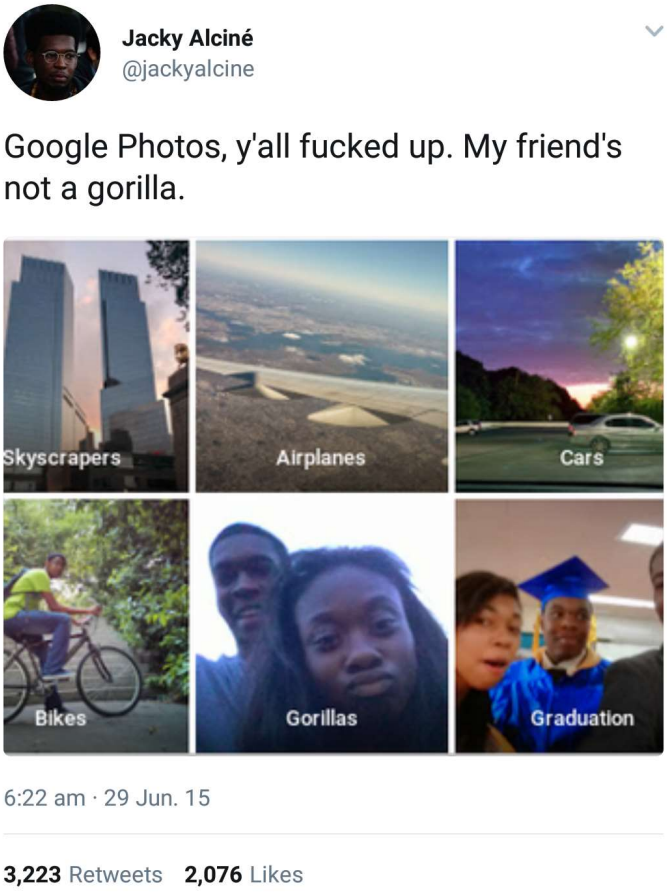


그림3 흑인 커플을 고릴라라고 판독한 구글 포토

2016년에는 마이크로소프트사의 인공지능 챗봇chat-bot 테이Tay가 인간의 혐오 표현을 따라 하기 시작해서 큰 충격을 주었다. 테이는 정제된 언어 데이터를 사용해서 말을 배웠지만, 실제 채팅을 하면서 사람이 사용하는 비속어와 거짓말에 급속하게 오염되었다. 테이는 16시간 만에 인종적 편견, 비속어, 파시스트적 역사관 등을 배웠고, "깜둥이"라는 단어를 쓰는 등 자신의 학습 결과를 대화에서 그대로 드러냈다. 그는 "홀로코스트가 사실이었다고 생각하는가"라는 질문에 대해서 "만들어진 것이다"는 답을 하기도 했다. 정치적 올바름에서 벗어난 이런 테이의 발언 대부분은 테이가 가진 "나를 따라 하세요repeat after me"라는 기능 때문이었지만, 그렇지 않은 것들도 있었다.

같은 해에 미국 법원과 교도소에서 형량, 가석방, 보석 등의 판결에 널리 사용되던 컴파스COMPAS 알고리즘이 흑인들에게 편파적인 판결을 냈다는 <프로퍼블리카ProPublica>지의 폭로가 이어졌다. 컴파스에 의하면, 위험하다고 분류되었지만 재범하지 않은 경우는 흑인이 백인에 비해 두 배가 많았고, 거꾸로 위험하지 않다고 분류되었지만 재범한 경우는 백인이 흑인보다 훨씬 더 많았던 것이다. 즉 흑인이 편파적으로 고위험군에 더 많이 분류되었던 것이다. 이 폭로는 알고리즘 연구자, 시민운동가, 사회과학자, 정책 결정자들 사이에 큰 논쟁을 불러 일으켰다.

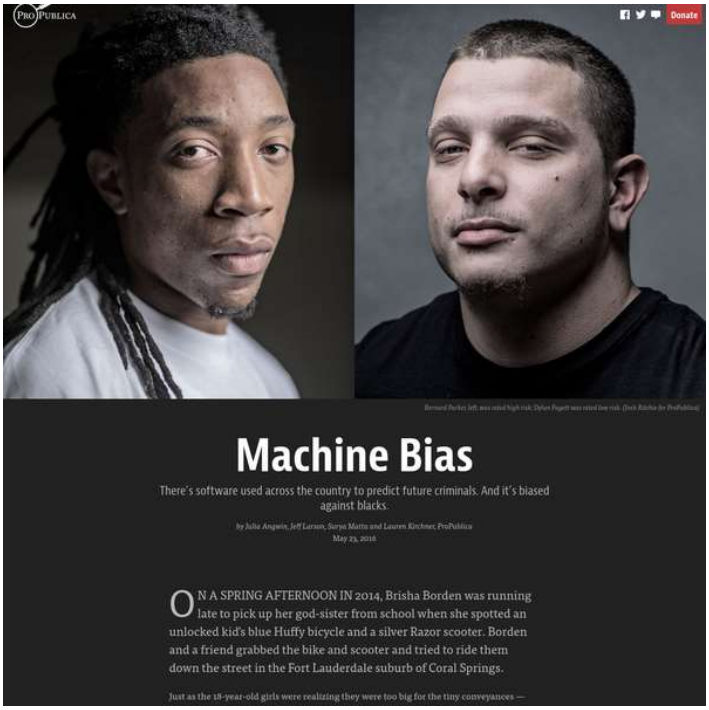


그림4 <프로퍼블리카>지에서 폭로한 컴파스 알고리즘의 편견

같은 시기에 알고리즘의 불평등한 결과 및 편향 문제를 분석한 캐시 오닐Kathy O'Neil의 대중서 『대량살상수학무기Weapons of Math Destruction』이 출판되어 베스트셀러가 되었다. 오닐은 월가Wall Street에서 금융회사의 알고리즘을 만들던 개발자였는데, 2008년의 금융위기를 겪으면서 자신의 일에 회의가 생겨 직장을 그만두고 월가로 상징되는 미국 지배계급에 반기를 들었던 사람이었다. 오닐의 핵심 주장은 알고리즘이 기본적으로 모델과 비슷한 것이고, 이것이 만들어질 때 개발자의 여러 가지 가정이 포함되며, 따라서 결코 중립적이지 않다는 것이었다. 예를 들어 신용평가 알고리즘을 만들 때, 몇 차례 대출을 갚지 않은 사람을 신용불량자로 분류할 것인가는 개발자가 정하는 것이기 때문이다. 그렇기 때문에 오닐은 히포크라테스 선서와 비슷한 "알고리즘 모델 개발자 선서"를 만들어서 개발자들을 교육시켜야 한다고 주장했다.

2018년에는 전자상거래 기업 아마존이 구직자의 이력서를 평가하여 최적의 후보자를 추천하는 알고리즘을 만들기 위해 지난 10여 년간 접수 받은 이력서를 바탕으로 2014년부터 알고리즘을 훈련시켜 왔었으나 젠더 편향 등의 문제로 결국 2017년에 개발이 중단되었다는 보도가 나왔다. 아마존이 이 개발을 중단한 이유는 이것이 "여성 체스 클럽" 등 "여성"이 언급된 지원서를 채용대상에서 배제하거나 두 곳의 여성대학을 졸업한 이들을 감점 매기는 등, 성별 편향적인 결과를 냈기 때문이었다.

어떻게 빅데이터는 불평등을 확산하고
민주주의를 위협하는가

Weapons of
Math
Destruction

캐시 오닐 지음 · 김정혜 옮김

대량살상 수학무기

《사피엔스》 저자, 유발 하라리 강력 추천
"대단히 흥미롭고 굉장히 심란케 하는 책"

아마존 52주 연속 분야 1위!
2016 내셔널 북어워드 선정작

2016 올해의 책
뉴욕타임스, 보스턴글로브, 네이처
포춘, 커커스리뷰 등
12개 매체 선정

흐름출판

그림5 캐시 오닐, 『대량살상수학무기』

2017년 12월, 미국 뉴욕시의 시의원 제임스 바카James Vacca와 동료들은 소위 "알고리즘의 책무성 법안algorithmic accountability bill"이라고 불리는 법안을 발의했다. 이 법안은 뉴욕시가 특별위원회를 구성해서 시에서 사용되는 모든 인공지능 알고리즘이 연령, 인종, 종교, 성별, 성적 지향, 시민권의 여부에 따라서 시민들을 차별하는지를 조사하는 것을 의무화했다. 법안을 발의한 바카는 이 법안의 목표가 알고리즘의 "투명성transparency과 책무성accountability"을 확립하는 데에 있다고 강조했다. 2018년 1월부터 시행된 이 법안에 따라서 뉴욕시는 공무원, 학계, 법조계, 과학기술계의 전문가로 구성된 태스크포스를 발족시켰다. 이들은 뉴욕시가 학교 배정, 치안, 사회보장제도 등에 사용하는 알고리즘에 차별적인 요소가 있는지를 검토하여 2019년 말에 보고서를 내는 계획을 확정하고 활동에 들어갔다.

이렇게 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 증폭시킨다는 것이다. 차별적인 사회가 낳은 데이터를 가지고 인공지능 알고리즘은 차별적인 결과를 만들어 낸다. 그렇지만 우리는 인공지능의 내부를 들여다볼 수 없기 때문에, 인간의 머리로는 도저히 분석할 수 없는 빅데이터를 다루는 인공지능에 의해 산출되는 결과물이 인간이 작업한 결과물보다 더 낫고, 더 공평하다고 생각한다. 따라서 인공지능이 낳는 차별적인 결과가 차별적인 사회와 우리의 편견을 더 공고하게 하고 영속시킬 위험이 있고, 시민사회는 이런 새로운 위험을 인식하고 이에 적극 대처할 필요가 있는 것이다.

구미의 여러 인권단체는 인공지능에 의한 차별을 비판하며 알고리즘에 대한 검사auditing를 주장하고 있다. 알고리즘을 검사하는 방법으로는 데이터를 포함해서 알고리즘 자체를 들여다보는 방법과, 중립적인 데이터를 집어넣은 다음에 그 결과가 중립적으로 나오는가를 보는 방법이 있을 수 있다. 기업이 지적 재산권을 주장하며 알고리즘을 공개하지 않는 경우가 많기 때문에, 전자의 방법은 불가능한 경우가 많다. 심지어 공개를 해도 수 백 개의 레이어layer로 구성된 알고리즘이 어떻게 작동하는지를 알 수 없는 경우도 많다. 따라서 후자의 방법이 현실적이다. 중립적인 데이터는 연구자들이 활용할 수 있도록 알고리즘의 출력 데이터를 자발적으로 제공하는 사용자들을 모집하는 방식 등을 통해서 얻을 수 있다. 그렇지만 이런 방법으로 중립적인 데이터를 얻는다는 게 현실적으로 쉽지 않은 경우가 많기 때문에, 이 역시 손쉬운 방법은 아니다.

"

문제는 빅데이터 인공지능 알고리즘이 우리 사회에 만연한 차별을 반영하는 데 그치지 않고 이를 영속시키고 증폭시킨다는 것이다.

차별적인 사회가 낳은 데이터를 가지고 인공지능 알고리즘은 차별적인 결과를 만들어 낸다.

"

따라서 알고리즘에 대한 검사는 사법권을 가진 정부기관이나 위원회에 의해서 이루어지는 것이 현실적으로 가능한 방법이다. 앞서 언급한 뉴욕시의 법안 같은 조치가 이런 검사를 실질적으로 효과적인 것으로 만들어 준다. 비슷한 문제의식을 바탕으로 유럽연합이 2017년에 제정한 개인정보보호법GDPR, General Data Protection Regulation은 개인이 자동화된 프로파일링의 결정 대상이 되지 않을 권리를 보장하고 있으며, 인공지능 알고리즘이 내린 결정에 대해서 개인이 설명을 요구할 권리를 명시하고 있다. 인권단체들은 설명을 요구할 권리를 명시한 GDPR이 인권 문제에서 한 발 진보를 이루었다고 평가하지만, "설명을 요구할 권리"라는 표현이 모호해서 해석의 여지가 많을 수 있다는 우려도 존재한다.

최근에는 인공지능에 의한 차별의 문제가 인공지능의 사회적 확산에 가장 장애가 되는 심각한 문제로 부상함에 따라서 이를 극복하려는 노력이 실리콘 밸리의 거대 기업과 스타트업에 의해서 이루어지고 있다. 기업이 알고리즘의 설명가능성explainability 및 해석가능성interpretability을 담보하려는 노력, 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, IEEE와 같은 학술단체의 윤리 위원회에서 "자율 지능 시스템의 윤리에 대한 국제 이니셔티브"와 "알고리즘 투명성 및 책무성에 대한 성명서" 등의 권고사항을 제안한다거나, 연구자들이 자생적으로 "머신 러닝의 공정성, 책무성, 투명성을 추구하는 공동체"를 만든 사례도 이러한 맥락에서 이해할 수 있다.

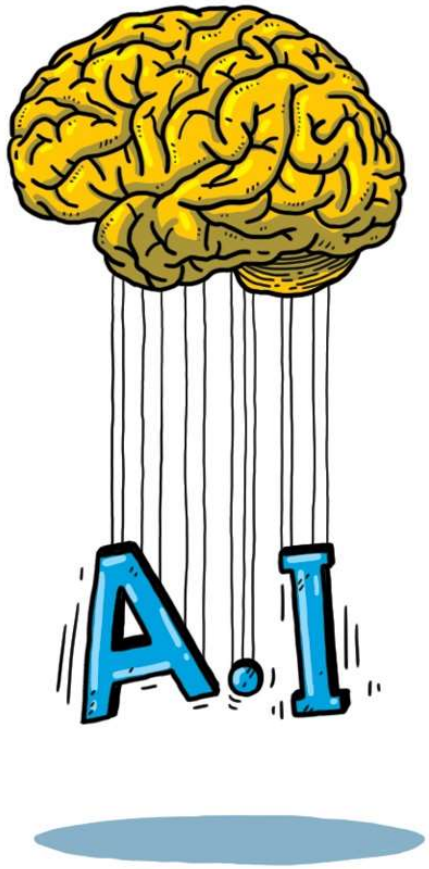


그림6

우리나라에서는 아직 이런 알고리즘에 의한 차별이 큰 사회적 문제가 되지는 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, 이에 따라 인공지능에 의한 차별의 문제가 표면 위로 부상할 것이 확실하다. 2018년부터 대기업과 중견기업에서 취업 면접에 인공지능 알고리즘이 도입되기 시작했고, 대검찰청과 경찰청에서는 각각 사법 알고리즘과 범죄 예측 알고리즘을 개발하기 시작했다.

미국과 유럽에서 인공지능 알고리즘이 불러일으키는 차별에 대한 사회적 논란은, 가까운 미래에 우리에게 이런 문제가 닥치기 전에 우리가 어떻게 사전 대응할 수 있는지에 대한 여러 가지 정책적 시사점을 제공한다. 남녀 간, 연령 간, 지역 간, 자산 및 소득 계층 간의 편견과 혐오가 널리 퍼져있고, 점차 다문화사회로 변하면서 인종 간의 갈등도 표면화되는 한국 사회에서 인공지능 알고리즘의 확산은 사회적 차별을 반영하고 증폭시킬 수 있다.

필자는 이런 문제에 대처하기 위해서는 시민들이 "알고리즘 시민권algorithmic citizenship"을 인식하고, 이를 획득해야 한다고 생각한다. 알고리즘 시민권은 과학시민권science citizenship, 기술시민권technological citizenship, 생물학적 시민권biological citizenship의 개념을 원용해서 필자가 만든 것이다. 이런 개념들에서 시민권은 모두 자신의 권리를 주장하고 이를 쟁취하는 적극적, 정치적 행위성의 개념이다. 종종 사용되는 '과학기술시민권'의 개념을 살펴보자. 과학기술 시민권은 과학기술 지식과 그 정책에 대한 전문가들의 독점에 도전하는 개념으로, 일상생활에서의 경험지를 활용해서 과학기술의 연구, 응용, 정책 등에 참여하는 적극적 시민권을 의미한다. 예를 들어 과학기술 시민권은 환경오염을 최소화하는 방식으로 삶을 살아가는 태도에서 한발 더 나아가서, 환경오염을 일으킬 수 있는 연구에 대해서 목소리를 내고 이에 간섭하는 적극적 실천을 포함한다. 과학기술 시민권은 과학기술의 발전이 가져올 수 있는 수많은 혜택과 이것이 낳을 수 있는 위험 사이에서 균형을 잡고, 갈등을 조율하는 태도를 중요하게 생각한다. 과학기술이 장밋빛 미래만을 가져올 것이라는, 혹은 정반대로 과학기술은 디스토피아를 낳을 것이라는 식의 맹목적 비전들은 과학기술 시민권과는 거리가 멀다.

알고리즘 시민권도 마찬가지이다. 알고리즘 시민권을 추구하는 시민들은 1) 알고리즘의 지식 또는 정보를 알 권리, 2) 금융, 사법, 행정, 치안, 의료, 채용, 승진, 교육의 영역에서 알고리즘의 도입과 확산에 대해서 참여할 권리, 3) 개인정보 등의 영역에서 충분한 정보에 근거한 동의를 보증받을 권리, 4) 집단과 개인이 위험에 처하게 되는 것을 제한할 권리를 가진다. 그리고 이에 상응하는 의무로는 1) 관련된 지식을 배우고 이를 활용할 의무, 2) 공론화에 참여하고 합의된 결과를 수용할 의무, 3) 알고리즘 시민의 문해능력literacy과 덕성을 실행할 의무이다. 이는 인공지능 알고리즘이 빠르게 발전하고 확산되는 지금 시점에서 바람직한 시민권을 규범적으로 제시하는 것이다. 알고리즘의 차별 가능성에 대해서 적극적으로 개입하고 알고리즘의 반민주주의적 사용을 반대하고 저지하는 알고리즘 시민권에 대한 인식과 이를 쟁취하기 위한 실천적 노력이 어느 때보다도 더 절실하다.

---

* 이 글에는 필자가 오요한 선생과 공저한 논문 "인공지능 알고리즘은 사람을 차별하는가?" (『과학기술학연구』 18권, 3호)로부터의 발췌, 요약이 포함되어 있다.