알고리즘 책임성 논의와 알고리즘에 대한 이해

김도훈
경희대학교 교수

I. 서론

2017년에는 전 세계적으로 AI에 대한 세간의 관심이 높아지면서 AI에 대한 기술적 측면뿐만 아니라 사회적 영향력에 대한 논의도 본격화되기 시작했다. 테슬라 자율주행자동차의 사고를 둘러싼 법적 책임소재의 문제[7]를 비롯하여 국내외 병원에서의 IBM 왓슨의 환자진료[8], ETRI 엑소브레인의 EBS 장학퀴즈 우승, 미국 법정에서의 AI를 활용한 판결[37],[39], 챗봇(Chat Bot)을 이용한 ARS의 확대, 로봇어드바이저(Robot Advisor) 주식 투자, IBM과 소프트뱅크에서의 왓슨에 의한 채용이나 인사고과평가[3] 등은 AI가 이미 우리의 삶에 직결된다는 점을 부각시키면서 큰 사회적인 반향을 불러 일으켰다. AI에 대한 투자와 혁신 활동도 2017년에는 400억 달러를 크게 상회했을 것으로 추정된다[14]. AI의 직간접적 영향력이 보편적으로 인지되면서 상충되는 입장들도 등장했다. 한편에서는 이를 4차 산업혁명과 기술혁신의 연장선에 놓인 당연한 흐름으로 이해하면서 적극적으로 수용하려는 반면, 다른 한편에서는 AI의 위험성을 경고하면서 이에 대한 법제도적 제한을 주장한다[12],[39].

AI에 대한 이해가 높아지면서 AI의 핵심 엔진인 알고리즘(Algorithm)을 둘러싸고 위와 같은 논의가 좀 더 구체화되고 있다. 본 고에서는 알고리즘 책임성(Algorithm Accountability)에 관한 최근의 논의를 정리하면서 앞으로 이 이슈가 전개될 방향을 전망해본다. 그런데 이 문제를 올바르게 다루려면 알고리즘에 대한 균형 잡힌 이해가 필요하지만, 이의 설계와 구현 과정에 필요한 핵심 요소들이 충분히 알려진 것으로는 보이지 않을 때가 많다. 이런 맥락에서 알고리즘 책임성 논의를 위해 필요하지만 잘 다루어지고 있지 못한 기술적 측면에 대해서 먼저 소개한다.

* 본 내용은 김도훈(BK21+ 데이터과학기반 경영전문연구인력양성사업단) 교수(☎ 02-961-9411, dyohaan@khu.ac.kr)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

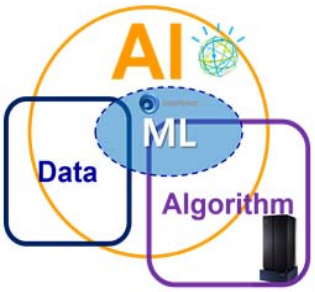
구체적으로 본 고의 내용은 다음과 같다. 먼저, AI 시스템 관점에서 빅데이터, 알고리즘, 데이터마이닝(Data Mining), ML(Machine Learning, 기계학습) 등의 용어를 체계적으로 구분하고 정리한다. 또한, NP-Complete 알고리즘 복잡성(Algorithm Complexity), 최적화(Optimization)와 휴리스틱(Heuristic), 정규화(Regularization) 등과 같이 알고리즘 전문가에게는 익숙하지만 세간에는 잘 알려지지 않아서 알고리즘에 관해 오해를 초래할 수 있는 주요 개념들을 정리하여 소개한다. 사실 이에 대한 이해는 알고리즘 책임성을 위한 올바른 논의를 정초함에 있어서 중요하다. 이제 보편적으로 사용되는 알고리즘이라는 용어는 경영과학(Management Science) 또는 OR(Operations Research)이나 응용 최적화 관점에서는 적절하지 못하다. 이들 분야에서 볼 때 본 고의 알고리즘이라는 용어는 휴리스틱(Heuristic)으로 대체되는 것이 더 타당할 것이다.[1] 이를 바탕으로 알고리즘 책임성에 관한 최근의 논의를 살펴보고 이 이슈를 보다 생산적인 방향으로 이끌 수 있는 방법을 모색해본다. 본 고의 접근법은 컴퓨터과학 및 경영과학과 법제도에 관한 사회과학적 시각 사이의 간극을 메우면서 소모적인 논쟁을 줄이는데 도움을 줄 수 있을 것이다.

Ⅱ. 알고리즘 이해와 논의를 위한 배경과 프레임

1. 의사결정과 알고리즘

AI 개발 초기와는 달리, AI는 이제 하나의 시스템으로 제공된다. AI를 편리하게 사용할 수 있도록 하는 인터페이스와 AI를 구동하는 엔진인 알고리즘, 특히 ML기반 AI에서 핵심적인 자원인 데이터 등이 이 시스템의 주요 구성요소이다([그림 1] 참조). AI가 사물이나 사건을 분류하는 것에서 더 나아가 의사결정에 직간접적으로 관여하는 경우에는 알고리즘의 역할이 더 중요해진다. 일반적으로 알고리즘은 데이터와 독립적으로 설계 구현되지만, 데이터와 상호작용하면서 동적으로 갱신되는 ML에서는 데이터와 분류기(Classifier)로서의 알고리즘이 기능적으로 긴밀하게 상호작용하도록 구현되기 때문에 그 역할과 코드블록을 명확하게 가려내는 것이 쉽지 않다.

1) 이런 관점에서 본 고의 알고리즘은 좁은 의미의 알고리즘과 휴리스틱을 포괄하는 넓은 의미의 알고리즘이다. 좁은 의미의 알고리즘과 휴리스틱의 차이와 각각의 의미와 응용에 대해서는 튜링상(Turing Award, 1975년)과 노벨 경제학상(1978년)을 수상한 Herbert A. Simon이 이미 1970년대에 설파하였다.



<자료> 경희대학교 자체 작성

[그림 1] AI 시스템의 구성요소

알고리즘이 다루는 의사결정문제의 종류에 따라 크게 최적화를 위한 알고리즘과 분류(Classification)를 위한 알고리즘으로 구분할 수 있다.[2] 또한, 알고리즘의 적용 대상이 되는 의사결정문제 자체의 특징이 알고리즘 성과의 한계를 결정할 수도 있다. 대표적으로, 문제해결 과정에서 필요한 정보량에 따라 정확도가 크게 좌우되는 NP-Complete형 의사결정문제와 그렇지 않은 P형 문제가 서로 분리되어 존재한다.[3] NP-Complete형 문제를 위한 알고리즘은 복잡성이 높을 수밖에 없으며 현실에서 마주치는 큰 문제에 대해서는 정확한 답을 제공한다는 것을 보장할 수 없기 때문에 정확성과 속도 사이에서 타협을 해야만 한다. 최근의 빅데이터와 ML의 발전으로 정확성과 속도 모두 비약적으로 개선되고는 있지만, 양자컴퓨팅(Quantum Computing)이 보편화되지 않는 한 Alpha-Go와 같은 성능의 AI가 스마트폰에 하드웨어로 탑재되는 것이 거의 불가능한 이유도 여기에서 찾을 수 있다.

의사결정문제의 종류에 따른 알고리즘의 성격과 유형을 구분하는 것은 알고리즘에 대한 법제도적 논의에서도 중요하기 때문에 다음 절에서 이에 대해 간략히 소개한다.

2. 알고리즘의 유형과 성격

가. 최적화 알고리즘

최적화 알고리즘은 네비게이션에서 최단 경로를 찾거나 최적 재고관리 및 생산일정을 결정하고 전국적인 에너지 수급계획을 마련하는 것과 같은 최적화 문제를 해결하는 알고리즘이다. 제2차 세계대전 당시 미국에서 군수물자 생산과 병참 물류에 관한 계획을 수립하기 위해 개발된 이후, 경영학, 경제학, 공학 등 광범위한 분야에서 사용되는 선형계획법(Linear Programming)[4]에 사용되는 알고리즘들이 대표적인 예이다. 이를 비선형 함수와 이산형 변수 등으로 확장

2) 실제 현실에서 사용되는 프로그램은 최적화와 분류 알고리즘을 복합적으로 사용하는 경우가 대부분이지만, 프로그램을 구성하는 단위 모듈은 결국 하나의 의사결정문제를 다루기 때문에 현 단계에서는 위와 같이 구분할 수 있다.

3) 가상적으로만 존재하는 Non-Deterministic 튜링(Turing) 기계를 사용할 때, 정확한 답을 찾는 속도가 최악의 경우에도 문제 크기(정보량)의 다항함수(Polynomial Function)로 표현되는 문제를 NP-Complete형 의사결정문제라고 한다. P형 의사결정문제에서는 어떤 경우에도 Deterministic 튜링 기계(쉽게 말해 오늘날의 컴퓨터)를 써서 정보량의 다항식으로 표현되는 속도로 정확한 답을 얻는다. 바둑과 체스 같은 게임에서 필승전략(Winning Strategy)을 찾거나 금융상품의 최적 포트폴리오를 구성하는 문제가 NP-Complete형의 예이고, 네비게이션에서 최단 경로(Shortest Path)를 찾는 문제가 P형의 예이다. NP-Complete형 문제와 P형 문제가 서로 다른 카테고리로 분리되어 존재한다는 명제는 아직 증명되지 못했지만, 거의 모든 수학자, 컴퓨터과학자, 경영과학자가 이 분리를 전제로 알고리즘과 휴리스틱을 개발한다. 이에 대한 이론은 이미 1971년 Cook에 의해 정식화 되었다(Garey & Johnson(1979) 등을 참조).

4) 이 내용은 우리나라를 비롯한 여러 나라에서 고등학교 수준의 수학 교과과정에 이미 포함되어 있다.

한 제반 수리계획법(Mathematical Programming)에서도 다양한 알고리즘이 개발되어 사용 중이며, 이 경우 NP-Complete형 의사결정문제가 대부분이기 때문에 최적해를 찾는 것을 포기하는 대신에 꽤 괜찮은 결과를 빠른 속도로 찾는 휴리스틱이 사용된다[19],[32]. 휴리스틱에 관한 일반화된 이론은 없지만 메타 휴리스틱(Meta Heuristic)으로 불리는 여러 방법들은 매우 잘 개발되어 있다([표 1] 참조).

[표 1] 최적화 알고리즘의 종류

구분	특성	알고리즘
(좁은 의미) 알고리즘	- (이론적) 최적해 보장 - NP-Complete형 의사결정문제에서는 속도와 저장공간의 효율성은 보장 못함(단, P형은 가능 - 예: 선형계획법의 Karmarkar 및 Dijkstra 최단 경로 알고리즘 등)	- 선형계획법을 위한 단체법(Simplex), ellipsoid, Karmarkar법 등 내부점(Interior Point) 알고리즘 - 비선형계획법(Nonlinear Programming)을 위한 수치해석(Numerical Analysis)기반 알고리즘들 - (혼합)정수계획법((mixed) Integer Programming)을 위한 분지한계법(Branch & Bound Method)과 그 변형 등 - 최단 경로, 최대 흐름 등, 네트워크와 이산형 최적화(Discrete Optimization)에서 개발된 다양한 알고리즘들
휴리스틱 알고리즘	- 최적해를 보장 못함[주] - 빠른 속도 - 효율적인 저장공간 활용	- 일반 탐색(Search)을 위한 휴리스틱: Local Search 등 - 시뮬레이션기반 최적화 휴리스틱(Monte Carlo 시뮬레이션 활용 등) - 메타 휴리스틱 * 유전자 알고리즘(Genetic Algorithm) * Tabu Search * Simulated Annealing * 뉴럴 네트워크(Neural Network: NN)

주) 휴리스틱은 이론상 최적해를 보장하지 못하지만, 현실적으로 꽤 괜찮은 결과를 주는 경우가 많기 때문에 일반적으로 상업용 목적에서는 최적해로 불리기도 한다. Parnas(2017)는 AI가 대중화되면서 최적의 답을 주는 것처럼 과장 및 왜곡되는 상황을 경계한다.
<자료> 경희대학교 자체 작성

나. 분류형 알고리즘

분류형 알고리즘은 말 그대로 사물을 적절하게 분류하는 의사결정에 사용되는 알고리즘으로, 의사결정문제의 기반이 되는 수리 모형, 구현 전략, 사용 목적에 따라 다양하게 유형화할 수 있다. 여러 교재와 문헌([27],[28],[30] 등)에서 가장 많이 채택된 유형화에 따르면, 기본 구현방식에 따라 지도학습(Supervised Learning)과 비지도학습(Non-Supervised Learning)으로 대별된다([표 2] 참조). 분류형 알고리즘을 구현하는 과정에서 최적화 알고리즘이 서브루틴으로 호출되는 경우가 많다[19].

지도학습은 올바른 분류에 대한 답(예; 회귀분석의 종속변수)과 해당 속성값(예; 회귀분석의 설명변수)이 관계를 맺을 때 가장 좋은 성과를 보이는 분류기(Classifier)를 산출한다. 이에 반

해 비지도학습에서는 정답을 사전에 알 수 없거나 알리지 않은 상태에서 속성값만으로 주어진 데이터를 몇 개의 카테고리로 분할하는 분류기가 산출된다. 따라서 지도학습의 분류기의 성능은 검증용 데이터에 의해 객관적으로 파악될 수 있지만, 비지도학습에 의한 분할의 적절성에 대해서는 별도의 평가방식이나 기준이 필요하다. 이런 식의 구분은 분류형 알고리즘을 적용해야 하는 목적과 용도에 따른 것이므로, 한 유형의 알고리즘을 다른 유형의 알고리즘과 직접 비교할 수가 없을 때도 많다.

[표 2] 분류형 알고리즘의 종류

구분	특성	알고리즘(분류기)
지도학습	- 판별 결과(맞거나 틀리거나)가 알려진 검증 데이터나 지식이 있고, 이를 활용하여 분류기를 산출 - 분류기 성능 평가가 용이하고 표준화되어 있음 (예; confusion matrix, ROC) - 훈련용 데이터의 품질에 성능이 크게 의존함 (예; garbage-in garbage-out)	- 회귀분석형(Regression Analysis) 분류기 * 일반 선형 및 비선형 회귀분석 * 로지스틱형(Logistic) 회귀분석: Logit, Probit 등 - 반복 분할형(Recursive Partitioning) 분류기 * CART(Classification and Regression Tree), MARS(Multivariate Adaptive Regression Spline) 등 - 커널함수(Kernel Function) 및 거리함수(Distance Function)기반 분류기 * SVM(Support Vector Machine), k-nearest neighbor 등 - (기본) 뉴럴 네트워크(Neural Network: NN) * N-계층 NN, CNN(Convolution NN)에 의한 Deep Learning 등
비지도학습	- 판별 결과를 확인해 줄 수 있는 검증 데이터 및 지식을 사전에 확보하지 않음 - 성능 평가를 위한 또 다른 프레임워크나 전략이 필요함 - 새로운 분류체계를 창출해야 하는 경우에 적합 (예; 신규 시장에서 고객집단 분류, 핵심정보 요약 추출을 위한 차원 축소)	- 클러스터링(Clustering) 혹은 군집화 알고리즘 * k-means, p-median, 계층화(Hierarchical) 클러스터링 등 - 성분분석(component analysis, CA) 및 분산분석(Variance Analysis) * 주성분분석(Principal CA), 독립성분분석(Independent CA) 등 * 공분산 행렬 분해(Covariance Matrix Decomposition): SVD Singular Value Decomposition) 등의 차원 축소(Dimension Reduction) 방식 - 비정상 행위 및 이상 징후 탐지(Anomaly Detection) 알고리즘 - 자기강화학습형(self-reinforcing learning) deep learning NN주) * generative adversarial NN, Hebbian NN 등 * 자기조직화지도(Self-Organizing Map) 등

주) 자기강화학습을 구현하는 방법과 전략에 따라 지도학습으로 분류될 수 있다. 그러나 훈련 단계에서 판별 결과를 검증할 수 있는 수단이 있다고 해도 이를 적극적으로 활용하지 않는다면(즉, 오류의 여지를 인정하면서 그것 마저도 사후에 바로잡을(Correction) 수 있는 식으로 구현하면) 비지도학습에 가깝다. 이는 자기강화학습에서 보상(Reward)을 보는 시각과 구현 전략에 따른 해석의 차이로 보아야 한다.

<자료> Cao(2017)[17], Hastie et al.(2009)[27], James et al.(2013)[28], Mullainathan & Spiess(2017)[30], Ng(2016)[31], Pyle & San Jose(2016)[34] 등 재구성

또한, 분류기의 수리적 구조에 따라 함수(적) 형태와 비함수 형태로 나누어 볼 수 있다. 기본 뉴럴 네트워크를 제외한 지도학습의 알고리즘과 비지도학습의 분산분석형 알고리즘은 함수 형태로 분류기를 산출하므로 정도의 차이는 있지만 판별에 대한 근거를 추적하고 검증하는 것이 비교적 용이하다. 클러스터링 분류기도 그 자체가 함수 형태를 취하지는 않지만 그룹화

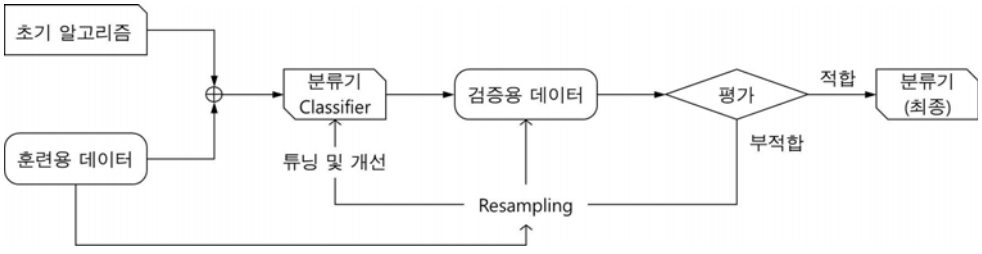
기준과 커널 및 거리함수를 통해 판별의 근거를 추론하거나 해석하는 것이 가능하다. 따라서 이들은 최소한 이론적으로는 화이트박스나 그레이박스 수준의 투명성(Transparency)을 가진다. 그러나 함수 형태를 갖지 않거나 너무 복잡해서 해석이 불가능한 분류기는 블랙박스 같은 불투명성(Opaqueness)을 가진다. 그러나 다음 절에서 보듯이 화이트박스 형태의 분류기라고 하더라도 그 구현과정까지 고려하면 실제로는 블랙박스와 다르지 않게 불투명할 가능성이 높다.

Ⅲ. 알고리즘 구현과정의 시사점

앞 절에서 살펴 본 여러 유형의 알고리즘 중에서 AI 시스템과 관련하여 현재 가장 많은 관심을 받고 있는 분류형 알고리즘의 구현과정을 간략히 소개한다. 특히, 알고리즘 책임성에 관한 법제도적 논의에서 고려해 볼 필요가 있는 부분을 중심으로 살펴본다.

1. 분류형 알고리즘 구현과정의 개요

분류형 알고리즘은 데이터와 긴밀한 관계를 유지하면서 분류기를 생성하기 때문에 실제로 그 성능은 데이터의 품질에 크게 의존한다. 데이터 자체에 문제가 없고 충분한 규모의 데이터가 확보되었다고 전제할 때, 분류형 알고리즘은 [그림 2]와 같은 절차로 구현된다[20],[27],[28][30],[34]. 먼저, 데이터 일부는 알고리즘에 의해 산출된 분류기의 성능을 검증하기 위해 남겨진다(검증용 데이터 집합, Validation Set 또는 Hold-Out Set).[5] 남겨진 훈련용 데이터 집합



<자료> 경희대학교 자체 작성

[그림 2] 분류형 알고리즘 구현의 기본 절차

5) 비지도학습의 경우, 그림의 검증용 데이터는 성능 평가보다는 분류기의 정상 작동을 파악하려는 차원에서 필요한 것으로 보아야 한다. 이 단계가 생략될 수도 있다.

(Training Set)은 알고리즘을 가동하고 파라미터를 결정 및 조절하는데 사용된다. 훈련과정에서 지도학습과 비지도학습 간 다소의 차이가 있지만, 적용되는 알고리즘에 의한 영향력의 차이가 더 크다. 분류기가 생성되면 검증용 데이터에 의한 성능 평가가 진행되고 문제가 있을 경우 원데이터의 재추출(Resampling)과 분류기 파라미터의 튜닝 및 개선을 거쳐서 새로운 분류기를 산출한다.

분류기의 성능 평가는 지도학습과 비지도학습에서 차이를 보이는데, 지도학습의 경우 정오표(Confusion Matrix)를 만들고 정답률(Accuracy), 정밀성(Precision), 민감도(Sensitivity 또는 재현율, Recall), 특이도(Specificity) 등의 지표(Metrics)를 사용한다. 비지도학습에서도 이들 지표를 산출할 수 있는 경우에는 지도학습과 유사한 방식을 따른다. 민감도는 양성(Positive)에 대해 제대로 양성으로 판별한 비율(true positive rate)인데, "1 - 민감도(= false negative rate)"는 제2종 오류(β, type 2 error)에 해당한다. 특이도는 음성(Negative)을 제대로 음성으로 판별한 비율(true negative rate)로, "1 - 특이도(= false positive rate)"는 제1종 오류(α, type 1 error)에 해당한다. 분류기의 판별 임계치(Threshold) 파라미터 변화에 따라 특이도와 민감도는 상충관계(Trade-Off)를 보이기 마련이며, 이 상충관계를 ROC(Receiver Operating Characteristic) 곡선으로 추적하면서 파라미터를 조정하고, ROC 곡선의 AUC(Area under the Curve) 면적을 계산하여 분류기의 성능을 평가한다(AUC가 클수록 우수한 분류기일 가능성이 높음).[6)]

성능 평가와 이를 통한 분류기 개선의 기본적인 접근법은 교차검증(Cross Validation)과 부트스트랩핑(Bootstrapping)을 활용한 반복과 재활용이다(Resampling 단계). 교차검증은 훈련용과 검증용 데이터 집합을 체계적으로 구성하여 검증을 반복함으로써 분류기의 성능을 높이고 과적합 문제(다음 절 참조)를 해결하는 방법이다. 크기 p의 검증용 데이터를 모든 경우의 수로 생성하는(Leave-p-Out) 심층적인(Exhaustive) 교차검증과 사전에 k개의 부분집합(k-fold)으로 데이터를 분할하여 k번만 교차검증 하는 비심층적 방식으로 대별된다. 부트스트랩핑은 원래 작은 표본으로부터 통계적 추론의 신뢰성을 높이기 위해(예; 추정량(Estimator)의 분산 줄이기) 개발된 방법이다. ML 등 분류형 알고리즘에서는 데이터(표본)의 체계적인 재활용을 통해 분류기의 강건성(Robustness)을 높이고 과적합 문제를 해결하기 위해 활용된다.

6) 여기서 소개한 지표들은 실제로 식품의약품안전처가 2017년에 제시한 의료용 AI의 성능 및 유효성 평가 기준으로 사용되고 있다[8].

2. 구현과정의 기술적 이슈와 해결방안

[그림 2]의 기본 절차만으로는 현실에서 사용할 수 있는 성능을 갖춘 분류기를 산출하기 어려울 때가 많다. 특히 과적합 이슈와 유연성을 높여야 하는 도전과제를 해결해야 한다.

가. 과적합과 정규화

ML에서는 데이터와 알고리즘이 유기적으로 상호작용하기 때문에 훈련과정에 쓰인 데이터에서는 분류기가 우수한 성능을 보이지만 다른 데이터에 대해서는 성능이 떨어지는 과적합(Overfitting) 문제가 발생할 가능성이 높다. 성능 평가에서 교차검증과 부트스트랩핑을 통해 과적합 발생 가능성을 낮출 수 있지만 한계가 존재한다. 어떤 형태의 데이터에 대해서도 일관된 성능을 보일 수 있는 유연한 분류기를 산출하기 위해서는 보다 정교한 개선이 필요한데, 그 해법에는 정규화와 랜덤화 과정이 수반된다.

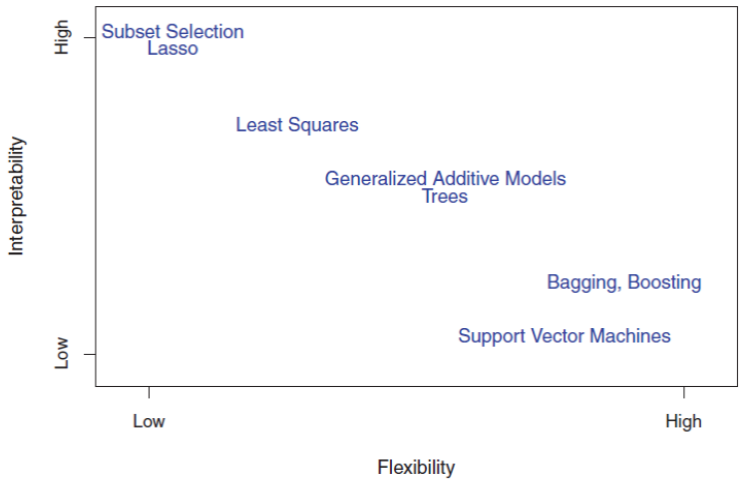
정규화(Regularization)는 훈련과정에 필요한 최적화를 의도적으로 제한함으로써 분류기가 다소 특이한 데이터 속성까지 고려하는 것을 방지한다. 최적화를 제한하는 방식과 제한 수준에 따라 다양한 정규화 방법이 가능하다. 보통 페널티(Penalty) 혹은 손실(Loss) 함수를 도입하는데, 예를 들어 회귀분석형 분류기에 적용되는 LASSO(Least Absolute Selection and Shrinkage Operator)는 페널티 함수로 L1 norm을 사용한다. 정규화를 통한 과적합 문제의 해결은 대개의 경우 위에서 소개한 교차검증과 결합되어 사용된다.

나. 랜덤화 과정

랜덤화(Randomization)는 분류기의 강건성과 유연성을 높이는 또 다른 방식으로, 크게 두 가지 전략이 있다. 첫째, 배깅(Bagging)은 훈련과정에서 데이터를 랜덤하게 분리하여 여러 부분집합을 생성하고 부분집합별로 서로 독립적인 복수의 분류기를 생성한 후 이를 종합하여 활용하는 방식이다. 생성된 여러 분류기들 집합을 앙상블(Ensemble)이라고 부른다. 둘째, 배깅이나 부트스트랩핑과는 달리, 부스팅(Boosting)은 설명변수 혹은 속성에 대한 랜덤화로 볼 수 있다. 부스팅은 사용 가능한 속성의 일부만을 이용하여 정밀도가 떨어지는 약한 분류기(Weak Classifier 또는 Weak Learner)를[7] 여러 개 생성한 뒤에 다시 훈련과정을 거치면서 약한 분류기들에게 적절한 가중치를 부여하여 종합하는 방식이다.

배깅, 부스팅, 교차검증, 부트스트랩핑은 모두 랜덤화를 통해 과적합 문제를 해결하면서

7) CART 알고리즘을 예로 들면, 의도적으로 깊이(depth)를 제한한 의사결정나무(decision tree)가 여기에 해당된다. 비유적으로 말하면 부스팅은 집단지성(Collective Intelligence)이라고도 할 수 있다.



<자료> G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction Statistical Learning with Applications in R, Springer, 2013. p.25.

[그림 3] 해석 가능성과 유연성 사이의 상충관계(Trade-Offs)

동시에 보다 강건한 분류기로 업그레이드시키려는 목적을 가진다. 실제로 이들은 복합적으로 적용되기 때문에 구현과정에서 명확하게 구별하는 것은 쉽지 않다.[8] 그런데 [그림 3]에서 보듯이 알고리즘의 기본형은 해석과 구분이 가능할지라도, 이들 방식을 적용하여 고도화된 분류기는 강건성과 유연성이 높아진 대가로 불투명성이 높아지며 그 결과에 대한 해석은 어려워진다. 이는 다음 절에서 소개할 알고리즘에 관한 법제도적 이슈와도 밀접히 관련된다.

IV. 알고리즘[9] 책임성(Algorithm Accountability)에 관한 논의

1. 알고리즘 책임성의 배경

ML의 신뢰성의 문제는 여러 문헌에서 지적되었는데([6],[7],[13],[35]), 주로 차별(Discrimination)과 오류에 관한 명확한 원인을 파악하거나 이해하기 어려울 뿐만 아니라 알고리즘 설계에

8) 예를 들어, 기본 CART의 발전된 형태인 랜덤포레스트(Random Forest)는 교재와 문헌에 따라서 배깅이나 부스팅의 일종으로 다루어진다. 그러나 이를 어느 범주에 포함시킬 것인지는 본질적인 문제가 아니다.

9) 이 절에서 알고리즘은 지금까지 다룬 순수 알고리즘보다는 프로그램에 가깝다. 이는 현실에서 알고리즘이라는 용어가 쓰이는 일반적인 맥락을 반영한 것이다. 따라서 여기서의 알고리즘은 앞 절에서 소개한 여러 알고리즘을 복합적으로 사용하는 프로그램으로 보아야 한다. 예를 들어, 가격설정(Pricing) '알고리즘'으로 불리는 것은 ML형 알고리즘과 최적화 알고리즘이 유기적으로 통합된 프로그램이다.

주관적 편향(Bias)이 개입될 가능성에 기인한다. 치안이나 의료와 같은 공공성이 강한 분야에서 결과의 공정성은 과정의 투명성(Process Transparency)을 요구하는 경우가 많은데, 자동화 알고리즘이 적용되는 경우 투명성을 보장하기 어렵다. 투명성은 경쟁법 전통과 관점에서 중요한 이슈로, 파스콸레(2016)는 평판, 검색, 금융에서 빅데이터와 알고리즘의 불투명한 활용을 지적하면서('비밀주의'라고까지 부른다), 사회적 개방성과 시장의 공정성 관점에서 알고리즘의 부정적 기능을 경고한다.

알고리즘 기반 자동화 검색과 가격설정에서 불공정한 오용 가능성을 보여주는 최근의 사례들이 이러한 우려를 현실로 만들고 있다. 불투명한 알고리즘이 유발한 부정적인 사례로, 특정 대출자의 상환 지연은 용인되고 다른 대출자의 상환 지연은 문제가 되는 경우(미국의 현행법은 금융기관이 그 이유를 설명하도록 요구함), 동영상 서비스를 제공하는 플랫폼사업자의 검색 결과에 언제나 자사의 비디오 콘텐츠가 먼저 추천되는 경우 등이 보고된다. 아무리 기업의 선의적 판단에 따른 결과라고 해도 공정성에 문제가 있을 경우 과정에 대한 검증이 제공되어야 한다는 주장이 제기되는 배경이다.

미국과 유럽의 경우 온라인상에서의 알고리즘 편향[35]과 가격설정 알고리즘에서의 담합이나 불공정 거래의 가능성에 대해서는 많은 논의가 진행되었으며, EC 및 OECD 같은 국제기구에서도 관심을 기울이고 있다[1],[9],[15],[16]. 알고리즘 편향은 미국 공정거래위원회(FTC)가 2010년에 구글의 검색 알고리즘의 불공정 가능성을 조사하면서 세간의 주목을 받았다. FTC가 2013년 증거 부족을 이유로 심의를 종료하였지만, 이 문제는 많은 플랫폼이 채택하고 있는 양면시장(Two-Sided Market) 비즈니스 모델[2],[24]에서 복합 서비스(Bundling 및 Tying)를 제공될 때 언제든지 다시 붉어질 소지가 있다[29]. 가격설정 알고리즘의 경우에도 2015년 미국 법무부(DOJ)의 반규제분과가 아마존 마켓플레이스에서 거래되는 포스터 상품의 가격설정 알고리즘에 문제가 있다고 판결한 바 있다[9],[29]. DOJ는 2016년 우버의 동적 가격 설정(Dynamic Pricing) 방식에도 문제가 있다고 지적하였다.[10)]

그러나 이러한 문제가 알고리즘의 투명성을 보장하는 것으로 해결될 수 없다는 방향으로 의견이 수렴되고 있다. 먼저, 앞 절에서 소개한 바와 같이, 기술적 측면에서 ML과 같은 분류형 알고리즘의 구현방식이 원천적으로 불투명하다. 따라서 테슬라 주행사고처럼 오류가 확실하다고 해도 일반적인 소프트웨어 디버깅 방식으로는 문제를 해결할 수 없다. 단지, 수차례 실험

10) 가격설정 알고리즘의 반경쟁적 행위(가격 담합 등)에 대해서는 미국 항공사들이 동적 가격설정을 광범위하게 도입한 1990년대부터 여러 연구가 진행되었다. 오늘날 이 문제가 주목받는 이유는 AI와 빅데이터 확산에 힘입어 특정 산업에만 국한되었던 이슈가 全산업에 걸쳐 보편화될 여건이 조성되었기 때문이다.

을 통해서 경험적으로만 신뢰성 및 안전성 수준을 점검할 수 있을 뿐이다.[11] 또한, 투명성 보장이 악용될 가능성도 높다[15]. OECD에서 잠정적으로 제안한 해결 방향에서 투명성이 강조되고 있지만 그 유용성을 의심하는 경향이 강하다[16]. 투명성이 해결책이 될 수 없는 궁극적인 이유는 ML을 비롯한 분류형 알고리즘이 인과관계(Causal-Effect)를 직접 다루는 것이 아니라 상관관계(Correlation or Association)를 바탕으로 마치 인과관계를 추론하는 듯이 보이기 때문이다[22],[30]. 따라서 이론적으로도 알고리즘의 투명성은 해결책이 되기 어렵다는 것을 알 수 있다.

2. 알고리즘 책임성에 관한 법제도적 논의

가. 책임성 이슈와 도전과제

알고리즘 책임성을 강하게 요구하는 법제도적 논의는 여러 난관에 부딪힌 것으로 보인다. 예를 들어, EU가 2016년에 마련하여 2018년 5월 25일부터 회원국에 적용하는 GDPR(General Data Protection Regulation, https://www.eugdpr.org[4])을 충실히 준수하는 수준으로 알고리즘 책임성을 구체화한다면, GDPR의 정보주체에 관한 권리(예; 설명을 요구할 권리(Right to Explanation: 제13조, 제14조, 제22조와 관련), 열람권(Right to Access: 제15조) 등)를 강제하는 것이 데이터 관리와는 또 다른 차원에서 기술적으로 어렵기 때문에 답보상태에 빠지기 쉽다.

법제도의 전통에서는 책임성을 위해 투명성을 요구하는 경향이 강하다. 그런데 알고리즘의 기술적 특성으로 인해 투명성 자체가 불가능한 경우가 많기 때문에 이러한 전통적 접근법에 대해 법학 분야에서도 회의적인 시각이 부상하고 있다[21],[33]. 첫째, 알고리즘이 공개되어도 복잡성으로 인해 현실적으로 해석이 불가능하거나 중의적 해석이 가능하다. 둘째, 2010년 YouTube와 Viacom 판례처럼 알고리즘 자체가 사적 자산으로서 지적재산(IP)으로 보호받아야 할 대상인 경우가 많다. 셋째, 알고리즘 자체가 동적으로 변한다(특히, 학습기능이 있는 경우). 넷째, 알고리즘 스스로도 주어진 입력에 대해 어떤 산출이 발생할지 모르는 경우(입력-산출 관계가 명시적이지 않은 경우)가 많다. 이러한 이유에서 알고리즘 감사(Algorithm Audit, [1],[23]) 제도를 도입하더라도 실효성 있는 대안이 되기 어렵다.

11) 이는 품질보장(quality assurance)과 관련한 상업화 측면에서 매우 비효율적인 방법이다. 만약 오늘날의 제조물배상책임(product liability)에 준하는 품질보장(예; 6시그마)을 알고리즘에도 요구하는 법제도가 시행된다면 AI 산업의 비용 부담은 막대할 것이다. 이러한 이유에서 ML의 데이터 의존도를 낮추려는 연구가 IBM, Nvidia, Qualcomm 등에서 진행되고 있다(one-shot learning이나 정교한 부스팅 및 가지치기(pruning) 등에 기반한 smaller algorithm 등, Sun, 2018). 이와 같은 연구가 성공하면 방대한 데이터 처리에 필요한 CPU나 GPU도 적게 소요되므로 하드웨어 제조업에도 직접적인 영향을 줄 것이다.

나. 몇 가지 대안적 방향

법학 분야에서 제안하는 대안들은 최대한의 투명성 혹은 유사 투명성을 제공하도록 하여 책임성을 담보하고자 한다. 예를 들어, Perel & Elkin-Koren은 블랙박스 두드려 보기(Black Box Tinkering)를 제안한다[33]. 불투명한 알고리즘에 사전 혹은 사용 중에 다양한 입력-산출 실험을 적용하여 그 결과를 공개하고 문제가 발견될 경우 알고리즘을 보완 수정하도록 하는 법적 근거를 마련하자는 것이다. 또한, 두드려 보기 절차를 제공하는 알고리즘 운영자에게는 문제가 발생할 경우 면책(safe harbor)도 가능하게 한다. 이 접근법은 이미 실행 중인 DMCA(Digital Millennium Copyright Act)와 유사하기 때문에 DMCA가 현실적으로 부딪히는 한계(예; 발견 즉시 조치(Notice and Take-Down) 실행의 어려움)를 기술적 법제도적으로 극복해야 하는 도전과제가 남는다.

Desai & Kroll[21] 및 Diakopoulos[22]도 알고리즘의 태생적 불투명성을 받아들인 상태에서 관련 법제도를 마련하는 방안을 제안한다. 보다 현실적인 대응방안으로, 먼저 알고리즘이 활용되는 목적과 상황에 따라 공공용(Public Interest)과 사적인 용도(Private Interest)로 구분하여 대응하는 것이 필요하다. 공공성 차원에서는 알고리즘의 정의로운 구현(Justice and Due Process)이 가장 중요하며, 사적인 상황(특히 상업용)에서는 불법적인 차별(Illegal Discrimination)을 억제하며 민주적으로 공평하게 사용되는(Fair Use) 환경을 조성하는데 초점을 맞춘다[5]. 소프트웨어 엔지니어링에서 사용되는 화이트박스 수준의 검증은 불가능하지만 블랙박스에 대해서도 충분한 검침이 가능하며, 문제가 발생했을 때 적극적으로 사후 개입(Ex-post Analysis and Oversight)할 수 있는 법제도적 장치를 마련하는 것도 필요하다고 한다. 그러나 Desai & Kroll[21]도 궁극적으로 내부고발자(Whistle-Blower)나 감시자에 의존할 수밖에 없다는 점을 인정한다. 이와 같이 알고리즘의 불투명성과 동태성은[12] 그 설계・구현・운영에서의 책임을 법제도로 구체화할 때 피할 수 없는 한계를 낳는 원천이다.

이러한 배경에서 알고리즘 책임성에 대해 애초부터 보다 기술적인 대안을 마련할 필요가 있다는 인식도 확산되고 있다. 마치 자동차를 설계・생산하는 단계에서 배출가스나 안전 기준을 강제하는 것과도 비슷한 맥락이다. 컴퓨터과학 분야의 대표적 국제학술단체인 ACM (Association for Computing Machinery)의 공공정책분과(Public Policy Council)에서는 알고리즘 책임성을 높이기 위한 본격적인 연구에 착수하였다. 먼저 알고리즘 개발과 구현에서 책임성 강화를 위해 7가지 원칙을 [표 3]과 같이 제시한다[40].

12) 알고리즘의 동적 특성과 상황의존성은 사후 개입의 실효성을 약화시키는 가장 큰 요인이다.

[표 3] USACM과 EUACM의 알고리즘 책임성을 위한 7대 원칙

원칙	설명
인지 가능성 (Awareness)	알고리즘이 사용되고 있음을 충분히 알리고 가능하면 사용법도 공지함
접속 및 시정 (Access and Redress)	알고리즘에 대한 조사가 원칙적으로 가능해야 하며 오류 및 잘못된 의사결정에 대한 수정지침을 사전에 제공함
책임감 부여 (Accountability)	알고리즘 구현 및 운영을 담당하는 주체를 명확하게 하고 책임감(responsibility)을 부여함
설명력(Explanation)	인간이 이해할 수 있는 수준으로 작동원리(logic)에 대해 설명할 수 있어야 함
데이터 출처 (Data Provenance)	알고리즘의 올바른 작동을 위한 충분한 데이터를 확보하고 데이터 출처에 대한 기록과 무결성을 제공함
감사 가능성(Auditability)	로그와 작동 기록을 남김으로써 감사와 분쟁 해결이 가능하도록 함
타당성 평가와 검사 (Validation and Testing)	알고리즘 성능에 대한 평가방식을 제공하고 적절한 방식으로 검사가 가능하도록 함

<자료> ACM U.S.(USACM) Public Policy Council & ACM Europe(EUACM) Policy Committee, Statement on Algorithmic Transparency and Accountability, 2017. 5. 25. p.2.

ACM의 알고리즘 책임성에 대한 노력은 현재 선언적 수준이며 보다 구체적인 방안을 마련하기 위해서 더 많은 연구가 진행 중에 있다(설명 가능한 인공지능(XAI, Explainable AI)에 대한 미국 DARPA와 EU의 연구 등[10][11]). 2~3년 전에 의욕적으로 출발하였던 기술공학적 관점에서의 알고리즘 및 AI 거버넌스(Governance)에 대한 논의도 편향, 검열, 차별, 사생활 보호, 지적재산권 보호, 지배력 남용에 대한 원론적인 방향 제시와 업계의 자율규제(Self- Regulation)에 의존하는 정도에서 답보상태에 있다[23],[36]. 알고리즘 책임성에 대한 지금까지의 노력을 살펴보면 아직도 우리 사회가 알고리즘(및 AI)에 대한 이해가 부족하다는 것을 시사한다. 알고리즘의 불완전성을 받아들이고[18],[31] 그 전제 하에서 시스템을 설계하고 운영하며 사후적 대비책을 마련하는 것이 현재 취할 수 있는 최선의 노력일 수 있다. 따라서 아직까지는 알고리즘(및 AI)을 생산성 혁신을 위한 도구로 인식하고 사회에서 단계적으로 수용하는 로드맵을 마련하는데 보다 집중할 필요가 있어 보인다.

[참고문헌]

[1] 김건우, “알고리즘으로 움직이는 경제 디지털 카르텔 가능성 커진다”, LG경제연구원, 2017. 8. 2.
[2] 김도훈, “플랫폼서비스 생태계의 개념적 유형화”, IT서비스학회지, 15(1), 2016, pp.299-319.
[3] 김경필, “서류전형에 15초 - 인공지능이 사람 뽑는다”, 조선일보, 2017. 8. 18.
[4] 김정곤, 윤재석, “EU의 일반개인정보보호법(GDPR) 발효와 대응과제”, KOTRA, Global Strategy Report, 18-002, 2018. 5.

[5] 박종훈, “AI로 인한 배제와 차별에 문제를 제기하는 신흥국의 AI 논의”, IITP, 주간기술동향, 2018. 1. 17, pp.33-37.
[6] 박종훈, “딥러닝에도 보안 문제, 인공지능(AI)을 속이는 수법에 주의할 필요”, IITP, 주간기술동향, 2017. 11. 29, pp.40-42.
[7] 박종훈, “답은 맞는데 풀이과정은 알 수 없는 인공지능을 믿어야 할까?”, IITP, 주간기술동향, 2016. 12. 28, pp.31-38.
[8] 정원준, “국내 인공지능(AI) 의료기기 현황 및 규제 이슈”, IITP, 주간기술동향, 2018. 1. 31, pp.2-15.
[9] 최난설헌, “알고리즘을 통한 가격정보의 교환과 경쟁법적 평가”, 경쟁법연구, 35, 2017, 215-241.
[10] 한국정보화진흥원, “미 국방연구원 설명가능 인공지능(XAI)”, NIA, Special Report 2018-2, 2018. 2. 26.
[11] 한국정보화진흥원, “EU의 인공지능 新 규제메카니즘: 설명가능 인공지능(XAI)”, NIA, Special Report 2018-3, 2018. 3. 15.
[12] F. 파스콸레(Pasquale), 이시은(역), 블랙박스 사회(원제: Black Box Society), 안티고네, 2016.
[13] T. Baer and V. Kamalnath, “Controlling Machine-Learning Algorithms and Their Biases,” McKinsey Global Institute, November 2017.
[14] G. Batra, A. Queirolo, and N. Santhanam, “Artificial Intelligence: The Time to Act Is Now,” McKinsey Global Institute, 2018. 1. https://www.mckinsey.com/
[15] J. Blockx, “Antitrust in Digital Markets in the EU: Policing Price Bots,” Proceedings of the Radboud Economic Law Conference, 2017. 6. 9.
[16] A. Capobianco and P. Gonzaga, “Algoricuthms and Competition: Friends or Foes?,” CPI Antitrust Chronicle, August 2017.
[17] L. Cao, “Data Science: A Comprehensive Overview,” ACM Computing Survey, 50(3), 2017, 43:1-42.
[18] M. Chui, J. Manyika, and M. Miremadi, “What AI Can and Can’t Do (yet) for Your Business,” McKinsey Quarterly, January 2018, 1-11.
[19] F.E. Curtis and K. Scheinberg, “Optimization Methods for Supervised Machine Learning: From Linear Models to Deep Learning,” Tutorials in OR, 2017, 89-113.
[20] E. Demir, “A Decision Support Tool for Predicting Patients at Risk of Readmission: A Comparison of Classification Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines,” Decision Sciences, 45(5), 2014, 849-880.
[21] D.R. Desai and J.A. Kroll, “Trust But Verify: A Guide to Algorithms and the Law,” Harvard Journal of Law & Technology, 2018(출간 예정).
[22] N. Diakopoulos, “Accountability in Algorithmic Decision Making,” Communications of the ACM, 59(2), 2016, 56-62.
[23] D. Doneda and V.A. Almeida, “What Is Algorithm Governance?,” IEEE Internet Computing, 20(4), 2016, 60-63.
[24] E. Garces-Tolon, “The Dynamics of Platform Business Value Creation,” CPI Antitrust Chronicle, August Issue, 2017.
[25] S. Garfinkel, J. Matthews, S.S. Shapiro and J.M. Smith, “Toward Algorithmic Transparency and

Accountability," Communications of the ACM, 60(9), 2017, 5-5.
[26] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, 1979.
[27] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Prediction, Inference and Data Mining(2nd ed), Springer Verlag, 2009.
[28] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction Statistical Learning with Applications in R, Springer, 2013.
[29] S. Li and C. Xie, "Rise of the Machines: Emerging Antitrust Issues Relating to Algorithm Bias and Automation," SSRN Working Paper, 2017. http://papers.ssrn.com.
[30] S. Mullainathan and J. Spiess, "Machine Learning: An Applied Econometric Approach," Journal of Economic Perspectives, 31(2), 2017, 87-106.
[31] A. Ng, "What Artificial Intelligence Can and Can't Do Right Now," Harvard Business Review, November Issue, 2016.
[32] D.L. Parnas, "The Real Risks of Artificial Intelligence," Communications of the ACM, 60(10), 2017, 27-31.
[33] M. Perel and N. Elkin-Koren, "Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement," Florida Law Review, 69, 2017, 181-234.
[34] D. Pyle and C. San Jose, "An Executive's Guide to Machine Learning," McKinsey Quarterly, June Issue, 2016.
[35] A. Raymond, E.A.S. Young and S. Shackelford, "Building a Better HAL 9000: Algorithms, the Market, and the Need to Prevent the Engraining of Bias," Northwestern Journal of Technology and Intellectual Property, 2017.
[36] F. Saurwein, N. Just and M. Latzer, "Governance of Algorithms: Options and Limitations," Info, 17(6), 2015, 35-49.
[37] M. Smith, "In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures," The New York Times, 2016. 6. 22.
[38] Y. Sun, "More Efficient Machine Learning Could Upend the AI Paradigm," MIT Technology Review, 2018. 2. 2. https://www.technologyreview.com/
[39] J. Tashea, "Courts are using AI to sentence criminals. That must stop now," Wired, 2017. 4. 17. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/
[40] USACM, "Statement on Algorithmic Transparency and Accountability," ACM U.S. Public Policy Council & ACM Europe Policy Committee, 2017. 5. 25. https://www.acm.org/public-policy/