2018
—12
NIA
Global Issue Report

기계도 사람을 차별한다?

머신러닝의 차별적 결과 방지 방안

How to Prevent Discriminatory Outcomes in Machine Learning

NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

Contents

Ⅰ. 서론

Ⅱ. 기계 학습이 차별적 결과를 낳는 원인

■ 기계 학습의 특징

■ 기계 학습이 차별적 결과를 낳게 되는 원인 : 데이터 관련

■ 기계 학습이 차별적 결과를 낳게 되는 원인 : 알고리즘 설계 관련

Ⅲ. 차별금지 원칙 구현을 위한 제안

■ 기계 학습에서 차별 근절을 위한 4가지 핵심 원칙

■ 기업을 위한 3가지 권고사항

Ⅳ. 시사점

본 보고서는 세계경제포럼 세계인권미래위원회 2016-18(The World Economic Forum Global Future Council on Human Rights 2016-18), **'기계학습의 차별적 결과 예방방법(How to Prevent Discriminatory Outcomes in Machine Learning)'** 내용을 요약, 정리 후 시사점을 도출한 보고서로 원문에 대한 저작권은 세계경제포럼 세계인권미래위원회에 있습니다.

Ⅰ 서론

○ 머신러닝 시스템은 직장 채용, 주택담보 대출 심사, 가석방 인원 선정 등 **인생을 바꿀 수 있는 결정을 내리는 데 사용되기 시작**

- 이러한 의사결정은 특히 사회의 가장 취약한 계층 혹은 저소득 및 중소득 국가 국민들의 인권에 큰 영향을 미침
- 제대로 개발되고 사용된다면 머신러닝 기술은 많은 데이터를 기반으로 신속하게 결정을 내릴 수 있으며 주관적 판단에 따른 불이익, 사회적 편견을 없애는 데 도움

○ 반면, 머신러닝 알고리즘이 편향된 데이터셋(Data sets)을 학습하게 될 경우 **한 쪽으로 편향된 결과가 발생**하게 됨

- 이 경우 머신러닝 기술이 **사회에 만연해 있는 편향성과 차별을 더욱 심화**시키고 인간의 존엄성 보장을 막을 가능성 대두

※ 예를 들어, 고용에 관한 과거 기록들에 따르면 대체적으로 여성이 남성에 비해 승진률이 낮음. 머신러닝 시스템이 해당 데이터를 학습한다면 여성을 고용하는 것이 더 좋지 않다는 결론을 내리게 되고 이로 인해 차별 심화가 우려

- 이러한 차별의 위험은 기존의 불평등이 깊게 자리 잡고 있고, 학습 데이터가 거의 없고, 정부 규제 및 감독이 적은 **저소득 및 중소득 국가에서 높게 나타나는 경향**
- 차별적 결과는 인권 침해 뿐만 아니라 머신러닝에 대한 대중의 신뢰를 잃게 하며, 이는 결과적으로 기술의 긍정적 효과를 억제하는 규제의 제정으로 연결될 수 있음

○ 세계경제포럼 세계인권미래위원회는 ‘**머신러닝(기계학습)의 차별적 결과 예방방법(How to Prevent Discriminatory Outcomes in Machine Learning)**’ 백서를 발간

- 머신러닝 응용 프로그램이 차별적 결과를 도출할 수 있는 잠재적 위험성을 이해하고, 이를 사전에 방지하기 위한 로드맵 작성의 기초가 될 수 있는 프레임워크를 제공
- 프로그램에서 생성될 수 있는 편견이나 차별을 확인·제거하는 것은 쉬운 일이 아니나, 공공·민간기관의 전문가, 기업, 이해관계자들이 협력하여 이로운 방식으로 기술을 활용하도록 촉구
- 인권이라는 기준을 머신러닝에 적용하는 논의는 최근 들어서야 이루어지고 있으며, 백서의 제안 내용은 이와 관련해 처음으로 개발·공개되었다는 점에 의의

○ 본 보고서에서는 **기계 학습의 잠재적인 사회적 영향**에 대한 논의의 연장선상에서, 해당 백서의 내용을 요약, 정리하고 시사점을 선정

- 머신러닝의 활용 범위가 점차 넓어짐에 따라, 평범한 사람들의 기본적인 생활(주거, 신용, 고용, 교육, 보건 등)에까지 영향을 미치기 시작
- 반면, 관련 규제제정과 기술발전의 속도 차이, 일부 국가의 경우 정부의 역량 부족, 정치적 혼란 또는 기타 불리한 환경 등의 이유로 인해 적합한 정부의 규제가 부재한 상황이 다수 존재
- 이 때, **민간기업의 적극적·자발적 관리가 요구**되며 이에 따라 해당 백서가 **범국가 혹은 국가 단위로 기업 참여를 위한 주요 정책 및 로드맵을 작성**하고, **머신러닝의 차별에 따른 인권문제에 관한 논의를 발전**시키는 데 기여할 수 있을 것으로 기대

머신러닝/인공지능의 역기능 사례

o 구글의 온라인 광고시스템은 고소득 직업에 대한 구인광고를 여성보다 남성에게 훨씬 자주 표시

o 하버드 대학교수 Latanya Sweeney가 자신의 이름을 구글 검색하여 학술 논문을 찾는 중 구글이 과거 그녀가 체포된 적이 있다고 광고하는 것에 충격을 받음

- 많은 연구 끝에 그녀는 흑인을 연상하게 하는 이름을 검색할 때 이러한 종류의 광고를 게재할 가능성이 25% 더 높다는 것을 발견

o 2010년 미국 주식시장의 플래시 크래쉬(갑작스런 붕괴) 사례: 5월 6일 36분 동안 자동화 알고리즘이 긴급 조정 응답을 생성함에 따라 비정상적으로 대규모 판매 주문이 일어나 미국 주식시장에서 거의 1조 달러가 사라진 사건

- 플래시 크래쉬는 현실 세계에서 작동하는 인공 지능 에이전트가 많아짐에 따라 예측하기 어려운 방식으로 상호 작용할 수 있다는 것을 보여준 예

※ 출처: NIA Special Report, EU의 인공지능 新규제메카니즘: '설명가능 인공지능(XAI)', 한국정보화진흥원, 2018.3

Ⅱ 기계 학습이 차별적 결과를 낳는 원인

□ 기계 학습의 특징

○ **(복잡성)** 기계 학습 시스템은 왜, 어떻게 의사결정이 내려졌는지를 역추적하기 어렵고, 인간이 이해할 수 있는 논리적 흐름을 제공하지 않음

○ **(불투명성)** 과거의 시스템과는 달리, 시스템이 매우 복잡하고 알고리즘의 고유의 특징으로 인해 차별적 결과를 낳은 출처를 쉽게 찾기 어려움

○ **(보편성)** 많은 사람들, 특히 미국과 유럽에 있는 사람들은 이미 매일 매일의 일상에서 기계 학습에 기반한 시스템을 활용 중

○ **(배제성)** 데이터 셋(Data sets)은 특정 기업에 의해 독점된 경우가 많고, 데이터의 수집·구매를 위해서는 상당한 규모의 자원이 필요

□ 기계 학습이 차별적 결과를 낳게 되는 원인 : 데이터 관련

1 데이터 접근성의 차이에 따라

○ 데이터의 이용 가능 정도, 즉 데이터에 대한 접근성 차이에 따라 기계 학습이 차별적 결과를 낳게 됨

○ 데이터를 생성하거나 구매하는 기업들은 대부분 데이터를 공개하지 않으며, 데이터를 구매하거나 수집할 만한 재정이 부족한 기업, 공공 및 시민단체는 사실상 기계 학습 시장에서 배제됨

○ 한 편, 디지털 발자국(흔적)을 적게 남긴* 그룹을 조사·분석한 결과, 지금까지 차별의 대상이 되어 왔던 사람들과 저소득 국가 다수 포함

* 무선 인터넷 서비스, AI 등 디지털 혜택을 누리지 못하거나, 적게 누린 것을 뜻함[1)]

2 편향된 데이터 입력 시

○ 편향되어 있거나 오류가 있는 데이터를 입력하면 편향된 모델과 차별적 결과를 도출한다는 것이 기계학습 시스템의 한계

○ 이를테면, 채용에 관한 과거의 기록은 대부분 남성에 비해 여성이 승진하는 경우가 드문데 이는 업무능력 차이가 아니라 오랜 기간 직장 내 자리 잡힌 불평등 문화에서 비롯

○ 그러나 기계 학습 시스템이 해당 데이터를 학습하게 될 경우, 남성이 여성보다 채용하기 좋다고 파악하고 계속하여 편향되고 차별적인 기준 데이터를 생성하기 쉬움

가설 예시

ㅇ 도서지역에 거주하는 케냐 농부의 금융권 대출 차별

- 케냐의 소액 대출을 제공하는 타라(Tala) 스마트폰 앱은 매일 앱에 접속하는 접속자 수, 이동 횟수, 하루에 부모님에게 전화하는 횟수, 청구서를 제 때 납부하는 지 여부 등 고객의 일상적인 습관에 관한 데이터를 수집

※ 이 알고리즘은 도시에 거주하는 사람들을 대상으로 데이터 포인트를 계산하도록 훈련받음

- 이는 고객의 신용도를 판단함에 있어 금융거래가 적은 사람들에 대한 신용 평가에 용이할 수 있음
- 반면, 도서·농어촌 지역에 거주하는 사람들의 경우 디지털 인프라 혜택을 누리지 못하고 있고 디지털 흔적을 생성할 수 있는 기회가 적기 때문에, 알고리즘이 이들을 부당하게 배제할 가능성이 부각

1) 출처: NIA Global Issue Report 2018-11, 머신러닝의 차별적 결과 방지 방안, 한국정보화진흥원, 2018.11

□ 기계 학습이 차별적 결과를 낳게 되는 원인 : 알고리즘 설계 관련

○ 기계 학습 알고리즘이 양질의 데이터 세트로 학습을 한다고 할지라도, 설계 단계에서 다음과 같이 차별이 인코딩될 우려가 있음

1 잘못된 모델의 선택

○ 알고리즘은 겉으로는 비슷해 보이는 분야에서 성공이 입증된 다른 알고리즘을 기반으로 설계되는 경우가 다수

○ 그러나 특정 분야에서 제대로 작동했을지라도 다른 분야에서는 차별적 결과 도출 가능

※ 치안유지 활동 예측에 활용되는 기계 학습은 지진 관련 모델을 기반으로 하는데, 이 때 지진은 범죄보다 지속적으로 기록하는 특성이 있어, 이를 적용한 치안유지 활동 예측 모델이 범죄 신고율이 높은 지역이 범죄가 많이 발생할 것이라는 왜곡된 결과를 도출할 수 있음

※ 미국의 대출 신청자의 상대적 위험을 제대로 평가했던 기계 학습 알고리즘이 다른 국가에서 사용될 경우 중요한 데이터를 간과할 수 있음

2 차별적 특징이 내재된 모델 구축

○ 사람이 부주의 등의 이유로, 기계 학습 알고리즘이 변수에 대해 어떤 가중치를 두어야 할지를 지정하면 편향된 결과 발생 가능

※ 대출 신청자의 신용등급을 평가하는 알고리즘은 신청자의 소득 수준과 과거의 채무상환의 신뢰성을 고려하여 작용함. 그러나 이를 활용하는 사람이 소득 수준이 더 중점을 두도록 한다면, 여성과 같이 소득 수준이 낮은 사람들을 부당하게 차별하는 결과 발생

3 인간의 감독 및 관여 부족

○ 기계 학습이 점차 정교해지고 복잡해지면서, 인간의 감독은 줄어드는 경향을 보임

○ 그러나, 기계 학습에서 중요한 요소가 예상치 못하게 간과되지 않았는지 확인하기 위해서는 인간의 지속적인 개입이 필요

※ 피츠버그 대학병원은 기계 학습을 이용해 폐렴환자의 합병증 발병 위험성과 이들의 퇴원 가능한 날짜를 예측. 이 모델은 의사가 사전 예방적 조치를 위해 천식 환자를 입원시켰을 때, 천식환자 중 합병증 발생 확률이 낮으므로 환자를 퇴원시킬 것을 권함

4 복잡하고 이해할 수 없는 시스템

○ 기계 학습을 활용하여 누구를 채용하고 채용하지 말아야할지 등과 같은 결정을 내리고 나면, 그렇게 결정한 이유를 역으로 추적하기 어려움

○ 그러나 발생 가능한 차별을 미리 파악하고 이를 바로잡기 위해서는 이러한 결정에 이르게 된 과정을 이해하는 것이 필요

○ 모든 의사 결정의 경로를 파악할 필요는 없으나, 이러한 결정이 인간의 권리에 영향을 미친다면 이는 매우 중요한 문제임

* 구글 지도에 사용된 기계 학습이 경로를 어떻게 설정하게 되었는지를 다 이해할 필요는 없음. 그러나 이러한 결정이 권리에 영향을 미친다면 매우 중요한 문제

5 의도적인 차별

○ 알고리즘 설계 시 고의적으로 편향된 생각과 의견을 입력하는 경우, 이를 규제할 수 있는 규정이나 법이 없으면 문제가 심각

가설 예시

① 멕시코의 건강보험 배제

- 멕시코 제도상, 양질의 의료서비스는 민영의료보험을 통해서만 받을 수 있음
  이에 따라 최소 2개 이상의 다국적 보험회사가 운영 중이고 이들은 기계 학습을 활용 중
- 보험회사가 기계 학습을 이용해 이용자의 과거 쇼핑 기록 등의 데이터를 수집하고, 더 많은 비용을 청구하기 위해 고위험군 고객의 일정한 패턴을 파악하려 했다고 가정한다면, 이 경우 가장 가난하고 아픈 사람이 의료 보험 서비스를 이용할 수 있는 기회가 가장 적음
- 결국, 이들 기계 학습 시스템이 **의료보험의 평등에 관한 권리를 침해**할 가능성을 시사

② 중국의 사회 신용 점수

- 대중에게 공개된 세부 정보는 거의 없지만, 보고서를 통해 중국이 금융거래, 세금, 직업 및 성과 기록부터 스마트폰, 전자상거래 및 소셜 미디어 활동에 이르기까지 다양한 데이터를 분석해 자국민의 점수를 매기는 모델을 만들었다는 사실이 알려짐
- 정부가 불완전하고, 기존의 편향성이 내재되어 있으며, '공정성'을 위해 만들어진 모델이 아닌 경우, 이렇게 수집된 데이터를 이용해 자동으로 점수를 매긴다면 이것이 무엇을 의미하는가에 관한 의문이 제기됨

Ⅲ 차별금지 원칙 구현을 위한 제안

□ 기계 학습에서 차별 근절을 위한 4가지 핵심 원칙

○ 기계 학습에서 편견을 없애고 인권과 존엄성을 보호하기 위한 4가지 핵심 원칙을 제안

- **(적극적 포용)** 기계 학습 애플리케이션을 개발하고 설계할 때는 반드시 인공지능(AI) 시스템의 결과물에 영향을 받는 사람들의 가치 등 다양한 내용을 포함해야 함

※ <참고> 예시 질문

· 시스템 개발에 참여하는 설계자 집단이 얼마나 다양한가?
· 데이터의 정확성을 평가하고 대체할 수 있는 데이터를 고려했는가?
· 앞으로 사용할 시스템으로 인해 혜택이나 불이익을 받게 될 집단을 파악했는가?
· 시스템이 사용될 부문과 관련해 사회적 규범을 충분히 연구하고 고려했는가?
· 여러 하위 집단에 대한 오류율과 유형을 계산하고 잠재적 영향을 계산했는가?

- **(공정성)** 기계 학습의 내용과 적용에 가장 적합하도록 '공정성'의 정의를 내릴 필요가 있으며, 기계 학습 시스템을 설계하고 평가하는 데 있어 공정성이 최우선이 되어야 함

※ <참고> 예시 질문

· 국제인권선에 부합하고, 기계 학습 시스템의 내용과 활용에 적합한 공정성의 정의가 무엇인지 확인했는가?
· 잠재적으로 편향성이나 불공정성을 나타낼만한 자료를 파악하고, 이를 해결하기 위한 방법을 설계할 전문가들이 포함되어 있는가?
· 앞으로 사용할 시스템으로 인해 혜택이나 불이익을 받게 될 집단을 파악했는가?
· 데이터, 알고리즘, 시스템 설계로 인한 문제로 인해 편향성 및 오류가 심화되지 않도록 '출시 전 테스트'를 엄격하게 실행했는가?
· 제품의 수명주기 동안 지속적으로 공정성을 평가하기 위한 시스템이 있는가? 예기치 못한 불공정 사례를 바로 잡기 위한 단계적 절차(혹은 긴급 절차)가 있는가?

- **(이해할 권리)** 기계 학습 시스템이 개인의 권리에 영향을 미치는 의사결정을 할 때에는 어떻게 영향을 미치는지 반드시 공개하고, 최종 사용자가 이해할 수 있도록 의사결정 결과에 대해 설명 필요

※ <참고> 예시 질문

· AI의 잠재적인 오류 및 불확실성의 모든 원인을 기록했는가?
· 의사결정의 측면에서 어떤 알고리즘을 공개했는가?
· 데이터 소스의 투명성은 어느 정도인가?
· 최종 사용자에게 기계 학습 시스템을 얼마나 설명할 수 있는가?
· 기계 학습 관련 코드 및 절차 중 오픈소스를 얼마나 많이 만들었는가?
· 제 3자가 시스템의 특징을 검토할 수 있도록 자세한 문서, 기술적으로 적합한 응용프로그램 인터페이스(API) 및 이용약관을 제공하고 있는가?

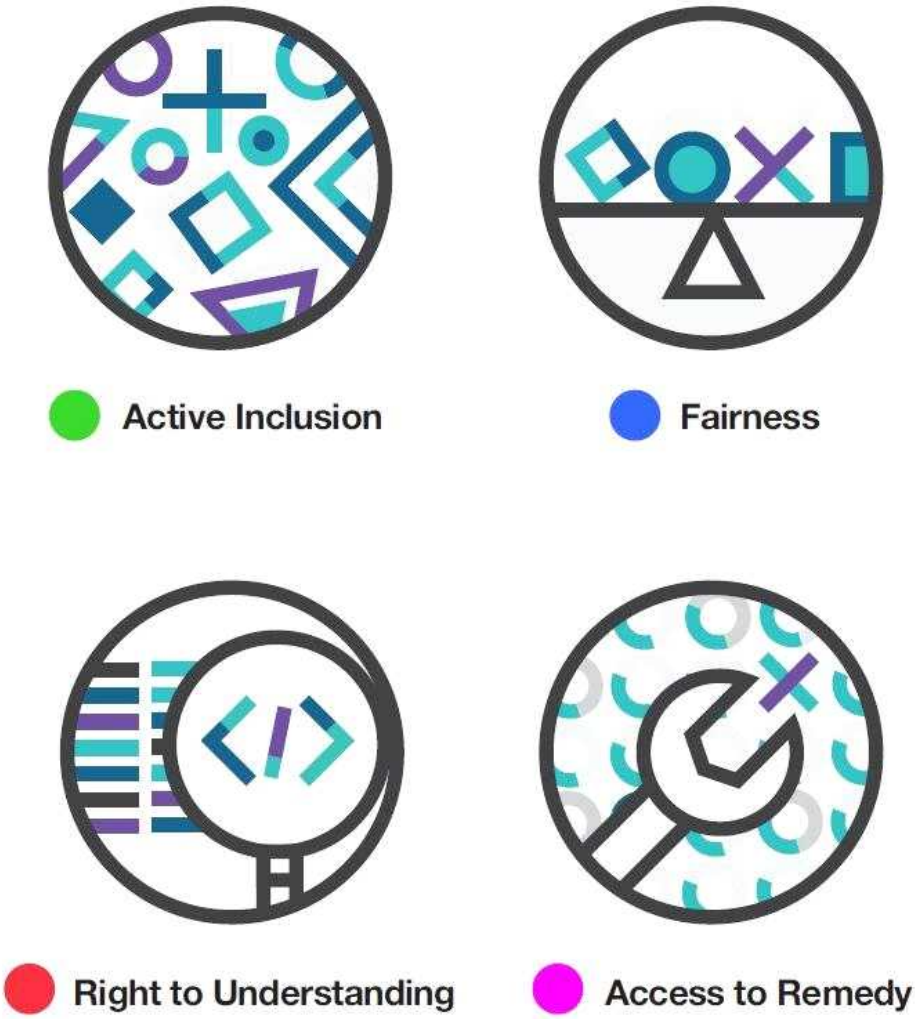
- **(피해구제)** 개발자나 설계자들은 시스템의 편향성으로 인해 영향을 받는 사람들을 위한 명확한 구제 방안과, 차별적 결과를 시의적절하게 수정할 수 있는 절차를 마련해야 함

※ <참고> 예시 질문

· 알고리즘의 의사 결정 결과에 대해 얼마나 확신하는가?
· 알고리즘을 이용하려고 하는 의사 결정자는 알고리즘의 확률적 특성을 이해하며, 결과가 100% 정확하지 않다는 것을 인지하고 있는가? 또한 오류를 수정하는가?
· 알고리즘의 산출물이 보안 정보나 민감한 자료를 포함하고 있지는 않은지 확인할 방법이 있는가?
· 최종 사용자의 인종, 연령, 거주 지역에 따라 알고리즘의 결과가 달라지는지 추적하기 위한 반사실적 가정을 테스트했는가?
· 현재 보고 절차 및 방법은 무엇인가?
· 보고된 문제를 기반으로 시스템 설계에 있어 필요한 내용을 수정할 수 있는 프로세스를 갖추고 있는가?

참 고

기계 학습에서 차별 근절을 위한 4가지 핵심 원칙 (four central principles to combat bias in machine learning and uphold human rights and dignity)



※ 출처: World Economic Forum White Paper, How to Prevent Discriminatory Outcomes in Machine Learning, World Economic Forum, March.2018

□ 기업을 위한 3가지 권고사항

○ 기업 활동과 관련이 있는 인권침해 위험을 사전에 확인 필요

- 기계 학습 시스템을 설계·실행하는 기업은 기계 학습 시스템의 설계·개발 단계부터 사용에 이르기까지 제품의 생명주기동안 인권침해의 가능성을 요약해서 간단하게 알려야 할 책임이 있음
- 인권침해의 위험 판단을 위한 공통된 최소한의 기준 및 절차, 그리고 학습 데이터의 적합성 및 잠재적 편향성을 평가하기 위한 일반 기준을 마련하고 채택할 것을 제안

○ 인권침해 위험 예방 및 경감을 위해 실질적인 조치 마련

- 기계 학습 시스템을 사용 시 인권에 상당한 영향을 미칠 수 있는 경우, 기업은 (업계 기준 및 인권관련 법률에 근거해) 독립적으로 알고리즘에 대한 감사 실시 필요
- 시스템에 내재되어 있는 잘못된 편향성을 확인하고 이를 수정하기 위해서는 사람이 직접 시스템을 점검 및 확인 필요

※ 기업의 개발자들은 데이터의 편향성을 찾고 이를 수정해야 하며 데이터 세트가 향후 영향을 받을 인구 계층을 대표하고 있는지 확실히 해야 함

※ 기업의 경영진은 개발팀에게 기본 틀과 인센티브 제공 필요

○ 인권침해의 위험성 파악, 예방 및 감소를 위한 노력을 투명하게 공개

- 기업이 자사의 기계 학습 응용 프로그램과 감사 보고서 결과를 직접 모니터할 것을 제안
- 기업 내부 윤리강령 및 책임 기준과 감사 결과에 대한 기업의 대응방안을 대중에게 투명하게 공개

Ⅳ 시사점

○ 기계 학습의 결과가 개인의 권리에 영향을 미칠 경우, **사용자가 이해할 수 있도록 알고리즘 개방에 대한 요구권 보장** 필요

- 알고리즘은 기업의 자산인 동시에 개인 인권 및 인생의 중대한 영향을 미칠 수 있는 의사 결정의 주체·수단

- 이러한 경우, **최종 사용자가 의사결정 결과를 이해할 권리** 및 알고리즘에 대한 공개 청구권을 보장할 필요

※ 2018 OECD e-Leaders Seoul Meeting(2018.10.30.~10.31, 한국 서울)에서는 신기술 도입과 적용에 있어 시민 참여를 통한 투명성과 신뢰확보가 핵심 요소이며, 이를 위해서는 국가 간 장벽을 넘어 정보 공유와 협업이 필요하다고 논의됨

○ 설계부터 사용까지 기계 학습 활용의 모든 단계에서 **지속적인 인간의 개입(감독 및 관여)이 중요**

- 기계 학습 시스템, 인공지능이 우리 사회에 확산되려면 기술, 알고리즘에 대한 신뢰가 관건

- 기계 학습에서 중요한 요소가 예상치 못하게 간과되지 않았는지 확인하기 위해서는 인간의 지속적인 개입이 필요

※ 폐렴환자의 합병증 발병 위험성과 이들의 퇴원 가능한 날짜를 예측하는 기계학습 모델은 천식환자 중 합병증 발생 확률이 낮으므로 환자를 퇴원시킬 것을 권하는 반면, 의사는 사전 예방적 조치를 위해 천식 환자를 입원시킬 수 있음

※ 탈린 디지털 서밋 2018(2018.10.14.-18, 에스토니아 탈린)에서는 세금, 복지, 범죄 등의 분야의 AI 자동화절차에서 인간이 개입하여 판단하는 것이 필요하다고 논의됨. 또한 AI 윤리 프레임워크도 전문가뿐만 아니라 시민 토론 등을 거쳐 사회적 공감대를 확보하여 마련되어야 할 것을 강조

○ 기계 학습의 내용과 적용에 가장 적합하도록 **‘공정성’의 정의**를 내리고, **윤리적인 이슈에 대해 고민하고 공론화 할 필요**

- 기계 학습에 대한 지나친 우려는 기술의 발전을 저해할 수 있지만, 이미 일어나고 있고 또 예방 가능한 것들에 대해서는 실질적인 방안 마련이 필요

- 경제적 관점 등 한 쪽에 치우치지 않도록 ‘공정성’의 정의를 내리고, 기계 학습이 국가별, 개별 평등을 확대하는 데 이바지 할 수 있도록 방향을 설정하는 것이 중요

▌참고문헌

[1] World Economic Forum White Paper, How to Prevent Discriminatory Outcomes in Machine Learning, World Economic Forum, March.2018

[2] NIA Special Report, EU의 인공지능 新규제메카니즘: '설명가능 인공지능(XAI)', 한국정보화진흥원, 2018.3

발 행 처 한국정보화진흥원

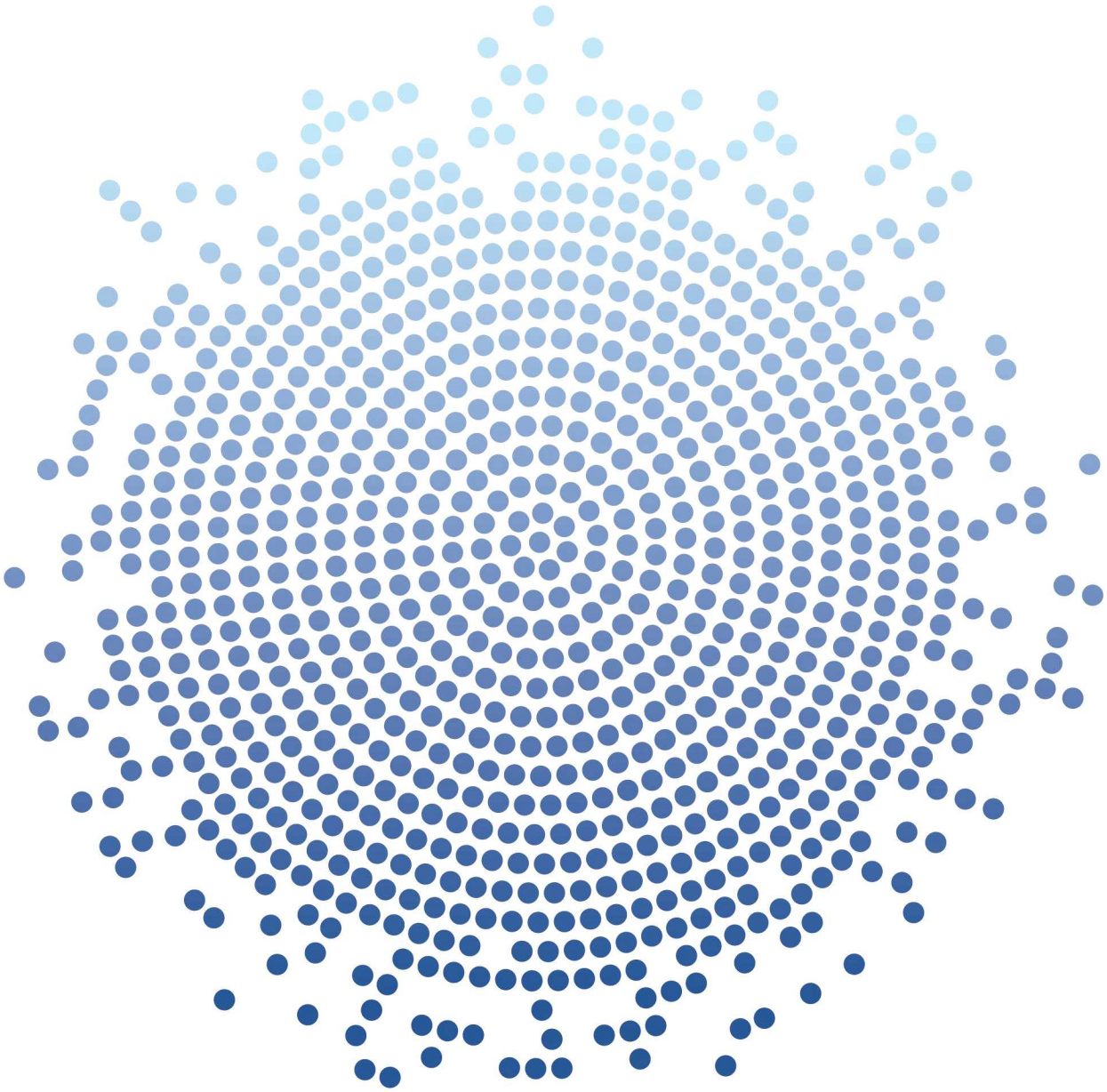
발 행 인 문용식

작　　성 한국정보화진흥원 글로벌협력단 글로벌기획팀

· 김인애 주임 연구원 inae_kim@nia.or.kr

기획 · 자문 오강탁 글로벌협력단장, 오현목 글로벌기획팀장

보고서 온라인 서비스 www.nia.or.kr



NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY