2022 Vol.04
차세대리포트

책임성 있는
AI를 위한 조건은?

KAST 한국과학기술한림원
The Korean Academy of Science and Technology

펴낸곳
한국과학기술한림원
031)726-7900

펴낸이
유 욱 준

발행일
2022년 12월

홈페이지
www.kast.or.kr

기획·편집
배승철 한국과학기술한림원 정책연구팀 팀장
백서연 한국과학기술한림원 정책연구팀 주임
조은영 한국과학기술한림원 정책연구팀 주임

컨텐츠
정현섭 과학기술분야 전문 작가

디자인·인쇄
경성문화사
02)786-2999

이 보고서는 복권기금 및 과학기술진흥기금의 지원을 통해 제작되었으며,

발간사

한국과학기술한림원에서 발간하고 있는 정책제안서인 차세대리포트는 우수한 젊은 과학기술인 그룹인 '한국차세대과학기술한림원(Young Korean Academy of Science and Technology, Y-KAST)' 회원들의 목소리를 담아오고 있다.

2018년 차세대리포트 발간을 추진하던 당시 'Y-KAST라는 이름으로 모인 젊은 과학기술인들은 어떤 목적을 가지고, 어떤 일을 해나가야 하는지'에 대한 설문 조사와 소규모 인터뷰를 진행한 바 있다. 그 결과는 매우 인상적이었다. 이들은 성별, 전공 분야에 상관없이 'Y-KAST는 다음 세대 과학자들을 위해 기여해야 하며, 정부와 젊은 연구자들 사이의 통로로 기능해야 한다.'고 목소리를 내었던 것이다.

올해로 발간 5주년을 맞이한 차세대리포트가 이들이 제시한 두 가지의 목표를 향해 나아갈 수 있는 작은 발판이 될 수 있기를 바라며, 미래 핵심기술과 관련된 이슈를 살펴봄으로써 과학기술의 발전뿐만 아니라 국가와 사회의 발전과 성숙도를 높여가기 위한 방향을 제시하고자 한다.

AI(Artificial Intelligence)는 눈부신 발전을 통해 4차 산업혁명을 주도하며 각종 추천 서비스, 의료진단, 재범예측, 자율주행, 채용·승진 평가, 챗봇 등을 통해 오늘날 우리 삶에 깊숙이 침투하고 있다. AI는 여러 분야에서 새로운 패러다임을 그려 나아가고 있지만, AI를 통해 얻을 수 있는 편리함 이면에 발생하는 윤리적 문제들이 수면 위로 떠오르게 되면서, 이제는 AI의 빠르고 정확한 결과 산출 이상으로 그 결과가 얼마나 사회적으로 잘 수용될 것인가에 대한 문제가 중요하게 다루어지고 있다. 이에 '책임성 있는 AI'를 개발하기 위한 연구 및 논의들이 AI 관련 산·학·연 전반적으로 이루어지고 있는 상황이다.

이번 차세대리포트 2022-04호에서는 AI가 발전하면서 그동안 야기되었던 다양한 사회·윤리적인 문제를 여섯 가지 요소의 관점에서 짚어보고, 각 요소와 관련하여 현재 어떠한 연구들이 이루어지고 있는지, 앞으로 필요한 AI 관련 정책은 무엇인지 살펴본다. 이를 통해 정책관계자들에게 AI 윤리 문제 해결을 위한 새로운 실마리를 제공하는 한편, 국민들과 과학기술계의 공감대를 형성해 가는 데 일조하고자 한다.

2022년 12월

한국과학기술한림원 원장

유 욱 준

함께해주신 분들

이성주 서울대학교 산업공학과 교수

서울대학교 기술인텔리전스 연구실에서 미래 예측과 기술전략 수립 관련 연구를 수행하고 있다. 특히 특허와 논문을 포함한 과학기술 분야의 텍스트 데이터를 활용해 유망기술을 발굴하고 과학기술 변화를 감지하려는 연구를 수행하고 있다.

박상철 서울대학교 법학전문대학원 교수

자율 또는 분류 시스템이 초래할 수 있는 위해에 대한 규범적 통제 방안 및 과거 권위를 독점한 소수의 법률가와 입법자의 재량과 직관에 맡겨졌던 법체계를 통계적, 귀납식으로 일관화, 고도화하는 방안 연구를 수행하고 있다.

서창호 KAIST 전기 및 전자공학부 교수

정보이론 및 인공지능 분야를 선도하고 있는 연구자로 최근 국제전기전자공학회(IEEE) 정보이론 소사이어티에서 수여하는 '제임스 매시 연구-교육상'과 'Google Research Award'를 수상했다. 신뢰할 수 있는 인공지능(Trustworthy AI) 연구를 진행하고 있다.

이경한 서울대학교 전기 · 정보공학부 교수

차세대 모바일 네트워크 시스템을 위한 아키텍처, 프로토콜 및 알고리즘을 연구하고 있다. 5G·6G 이동통신 시스템에서의 효율적인 차세대 네트워킹 서비스(메타버스, XR, 원격실재 등) 실현을 위해 단말 및 엣지·클라우드 컴퓨팅 플랫폼에서 활용될 필요가 있는 AI·ML 기술 개발 연구에 주력하고 있다.

이병영 서울대학교 전기 · 정보공학부 교수

사용자의 데이터들을 안전하게 활용할 수 있는 컴퓨팅 시스템을 디자인하는 연구를 하고 있다. 특히 기밀계산(Confidential Computing)이라 불리는 기술을 바탕으로, 현재는 기술적으로 불가능한 "안전한 데이터 공유", "Private AI" 등과 같은 차세대 데이터 서비스를 개발하는 연구를 수행하고 있다.

이승원 성균관대학교 의과대학 교수

국내외 다양한 의료 빅데이터(건강보험공단, 심사평가원, 질병관리본부, 병원 데이터, 해외 의료데이터 등)를 인공지능 및 의학통계를 이용하여 분석하고 예측하는 연구를 수행하고 있다.

조민수 POSTECH 컴퓨터공학과 교수

계산적 시각지능 연구, 컴퓨터비전 분야 전문가로 인공지능에 관계와 구조에 대한 체계성을 부여하는 것이 주된 연구 관심사이다. 현재 시각 요소들 간의 관계추론과 학습 문제들을 중심으로 영상정합, 대칭탐지, 물체발견, 비디오해석, 물체조립 문제 등에 대한 연구를 활발히 수행하고 있다.

CONTENTS

들어가기 04

Ⅰ. 책임성 있는 인공지능의 여섯 가지 조건

❶ 공정성, 한쪽으로 치우치지 않는 공평한 AI 07
❷ 견고성, 낯선 상황에도 안정적으로 작동하는 AI 11
❸ 설명가능성, 누구나 이해하고 공감할 수 있는 AI 14
❹ 투명성, 속이 훤히 들여다보이는 AI 18
❺ 가치정렬, 인간의 가치와 잘 부합하는 AI 20
❻ 프라이버시, 개인정보를 식별하지 않는 AI 22

Ⅱ. 정책제언

❶ AI 관련 법 · 제도 마련과 윤리기준 확산이 필요하다 25
❷ 현재 기술의 한계를 이해하고 대처해야 한다 27
❸ 윤리적인 인재 양성과 학제간 연계를 강화해야 한다 28

들어가기

최근 AI가 우리 삶 깊숙한 곳까지 침투하고 있다. 이미지인식, 음성인식, 챗봇 등 다양한 AI가 개발되고 실제 산업 현장에서 활용되고 있다. 그렇다면 정말 우리의 업무를 믿고 위임할 만큼 AI가 충분한 책임성을 갖추었다고 볼 수 있을까? 사실 AI는 자의식과 자유의지를 갖고 있지 않기 때문에, AI가 책임성을 갖춘다는 말은 인간이 AI를 만들 때 얼마나 책임성을 불어넣을 수 있는가의 문제를 말하기도 한다. 책임성을 갖춘 AI는 어떻게 규정할 수 있을까? 이번 차세대리포트에서는 AI가 책임성을 갖추었는지 판단할 수 있는 조건을 아래와 같이 제시하고자 한다.

공정성(Fairness)

공정성은 AI가 '공평하고 올바른 가치'에 따른 판단 결과를 제공하는 것을 말한다. 많은 영역에서 AI가 인간의 판단을 대신 또는 보완하면서 공정성을 갖춘 AI에 대한 중요성이 높아지고 있다. AI의 판단이 표상(representation)을 넘어 기회와 자원의 배분(allocation), 특히 '개인에 대한 평가' 또는 '개인의 분류'와 관련된 것이라면, 공정성 문제는 더욱 중요해진다. 예컨대 채용 심사, 형량 결정, 대출 심사 등을 판단하는 경우가 그렇다. 이러한 판단은 누구나 공감하고 수긍할 수 있도록 공정하게 이루어져야 하며, 공정을 갖추지 못한 AI에 의존하게 된다면, 누군가는 불공평한 상황에 처할 수밖에 없을 것이다.

견고성(Robustness)

견고성은 AI가 다양하고 복잡한 문제에 직면했을 때에도 안정적인 동작을 제공하는 것을 말한다. 학습-추론으로 단계가 분리되어 있는 심층신경망 기반 AI는 학습에 사용한 데이터에 의존적이며, 사전에 학습한 특정 과제(task)와 제한된 환경에 대해서만 성능을 보장한다. 예컨대, 자율주행 자동차에 탑재된 AI가 이러한 구조를 따른다면 사전에 학습하지 않은 낯선 형태의 도로 환경에서는 완벽한 작동을 보장하기 어렵다. 이처럼 심층신경망 기반 AI는 복잡하고 새로운 환경에서 기술적으로 충분히 안정적이고 견고하게 작동하지 못한다. 낯선 상황에서도 잘 작동할 수 있는 견고한 AI를 개발할 수 있다면 한층 신뢰할 수 있는 안정화된 AI를 경험할 수 있을 것이다.

설명가능성(Explainability)

설명가능성은 AI 모델의 출력값에 대해 누구나 이해하고 수긍할 수 있도록 하는 것으로, '해석가능성(interpretability)'이라고 표현하기도 한다. 알고리즘 결정 과정을 상대적으로 쉽게 분석하고 파악할 수 있었던 규칙 기반 AI나 고전적인 선형회귀모형 등과는 달리, 최근 널리 활용되는 심층학습 기법은 결정 과정에서 복잡성 및 불투명성 문제가 존재한다. 이에 심층신경망 인공지능에 대해 '블랙박스와도 같다'는 오해의 소지를 불러일으키는 표현이 생겨나기도 했다. 사실 심층신경망은 내부를 들여다볼 수 없는 것이 아니고, 수많은 연산들이 무엇을 의미하는지 이해하기 쉽게 설명하기가 어려울 뿐이다. 따라서 설명가능성은 인공지능이 어떤 결과를 산출할 때 수행하는 많고도 복잡한 연산을 인간이 이해하고 공감할 수 있도록 제공하는 문제라고 할 수 있다.

투명성(Transparency)

투명성은 AI의 개발부터 관리까지 투명하게 공개해야 한다는 것을 말한다. AI의 투명성은 단순히 알고리즘의 투명성 이상을 의미한다. 블랙박스로 묘사되는 AI 시스템을 투명하게 공개함으로써 명확한 목적이 무엇이고 어떻게 설계되어 운영되고 있으며 그 과정에서 개인정보나 데이터가 어떻게 사용되는지 등을 사용자가 충분히 이해할 수 있도록 한다. 이러한 점에서 투명성은 AI의 설명가능성과 연관성이 높다고 할 수 있다. 투명성이 담보되지 않은 AI는 곧 기술에 대한 활용성과 신뢰성을 떨어뜨리는 결과를 낳게 된다.

가치정렬(Value alignment)

가치정렬은 AI가 인간의 가치와 부합하는 것을 의미한다. 단순 자동화를 넘어 자율성을 갖는 AI가 등장하면서 AI에 도덕적 가치를 부여하는 것이 중요해지고 있다. 자율성이 높은 AI는 해당 시스템이 운영되는 과정에서의 목표와 행동이 인간의 가치에 부합하도록 설계되어야 한다. 가치정렬은 기술적 관점(technical perspective)과 규범적 관점(normative perspective)에서의 논의가 필요하다. 기술적 관점은 AI가 맡은 일을 적절히 수행할 수 있도록, 가치 혹은 원칙을 AI에 어떻게 잘 인코딩(encoding)할 수

있는가에 대한 것을 다룬다. 규범적 관점은 어떠한 가치 혹은 원칙을 AI에 인코딩할 것인지 결정하는 문제에 대한 것이다.

프라이버시(Privacy)

프라이버시는 AI가 학습하는 데이터에 개인 식별성이 없는 상태를 의미한다. AI가 학습하는 데이터가 개인정보와 관련된 것일 경우 프라이버시가 특히 문제될 수 있다. 대표적인 AI 프라이버시 문제 사례로는 한 스타트업 기업에서 AI 챗봇을 개발하여 서비스하는 과정에서 개인정보를 동의나 충분한 가명처리 없이, 또는 가명정보의 적법한 활용범위를 넘어 활용하여 개인정보 보호법을 위반한 사건을 들 수 있다. 최근에는 국가적으로 프라이버시와 관련된 다양한 법·제도·정책이 개발되고 있어 이에 발맞춘 기술 개발이 요구되고 있다.

이번 차세대리포트에서는 책임성 있는 AI가 되기 위해 갖추어야 할 위 여섯 가지 조건에 대해 소개하고자 한다. 이러한 조건을 충족하지 못하면 어떠한 사회적 문제가 발생하는지, 이를 해결하기 위해 오늘날 어떤 연구들이 진행되고 있는지 알아보고, 미래에 어떤 지향점을 가지고 연구해야 책임성 있는 AI를 만들 수 있는지 논의해보고자 한다. 또한, 책임성 있는 AI 개발을 위해 정책적으로 필요한 것들은 무엇인지 알아보고자 한다.

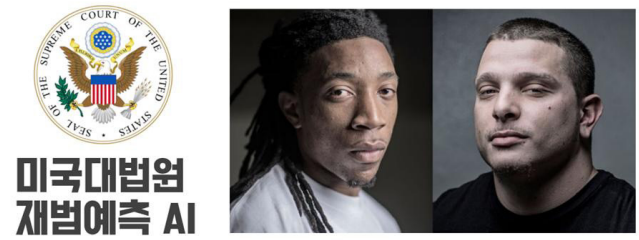
I

책임성 있는 인공지능의 여섯 가지 조건

1 공정성, 한쪽으로 치우치지 않는 공평한 AI

현재 개발되고 있는 AI는 우리가 생각한 만큼 공정할까? 다양한 영역에서 AI의 적용이 활발해지는 만큼, 그동안 공정하지 못한 AI로 인한 사회적인 논란도 많이 발생해왔다. 아마존의 AI 채용심사관은 모든 여성 지원자를 탈락시켜 성차별 논란을 일으켰으며, 구글의 물체인식 AI는 흑인 커플을 고릴라로 인식해 큰 사회적 파장을 일으켰다. 또한 미국대법원이 활용했던 재범확률 예측 AI는 백인 피의자 대비 흑인 피의자의 재범률을 높게 예측했다는 이유로 인종차별 논란을 일으켰으나, 이는 흑인의 평균적인 높은 범죄율을 반영한 것일 뿐 오류의 정도는 인종 간에 차이가 없다는 반론이 제기되는 등 논란이 계속되고 있다.

그림 1 미국대법원 재범예측 AI 사례



미국대법원의 재범예측 AI가 백인 대비 흑인 재범률을 높게 예측하여 불공정 논란을 야기하였다

출처 : ProPublica 2016

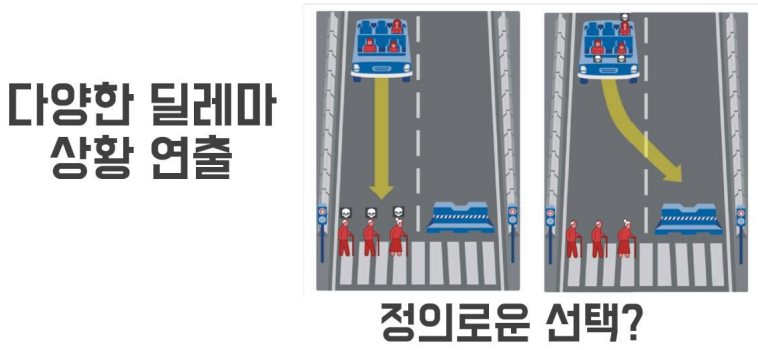
공정하지 못한 AI가 개발된 이유는 무엇일까? 바로 AI가 어떠한 이유에서든 어느 한쪽으로 편향된 데이터를 더 많이 학습했기 때문이다. 아마존의 AI 채용심사 사건은 여성 합격자 데이터가 부족하여 남성 합격자 데이터 중심으로 AI가 학습된 것이, 구글의 물체인식 AI 사건은 흑인 여성에 대한 데이터 부족이 원인이었다.

가. 모럴머신, 공정성 있는 AI의 가능성을 보이다

공정성 있는 AI를 만들기 위해서는 AI가 편향되지 않은 데이터를 학습해야 한다. 즉, 학습 데이터를 수집할 때 다양한 계층, 국가, 민족 등이 반영되어야 하는바, 특히 공론화된 공정성 기준을 도입함으로써 공정성 있는 AI를 개발할 수 있다.

이러한 노력의 일환으로 매사추세츠공대(MIT), 하버드와 같은 유수 대학들은 '모럴머신(moral machine)'이라는 빅데이터 수집 플랫폼을 합작 개발하였다. 모럴머신은 어떤 선택이 공정한지에 대한 공론화된 선택 결과를 얻고자 다양한 계층, 국가, 문화의 사람들이 어떠한 판단을 하는지에 대한 빅데이터를 233개 국가의 230만 명으로부터 수집하였다. 예를 들면, 아래 그림과 같이 자율주행 자동차가 직진하면 노인 3명이 다치게 되고, 핸들을 왼쪽으로 틀면 차량 탑승자가 다치는 딜레마 상황을 연출하고 선택하게 했다.

그림 2 트롤리 딜레마 상황 연출



모럴머신은 트롤리 딜레마 상황을 연출하고 많은 사람들의 선택 결과
데이터를 수집하여 공론화된 의견을 도출하고자 하였다.

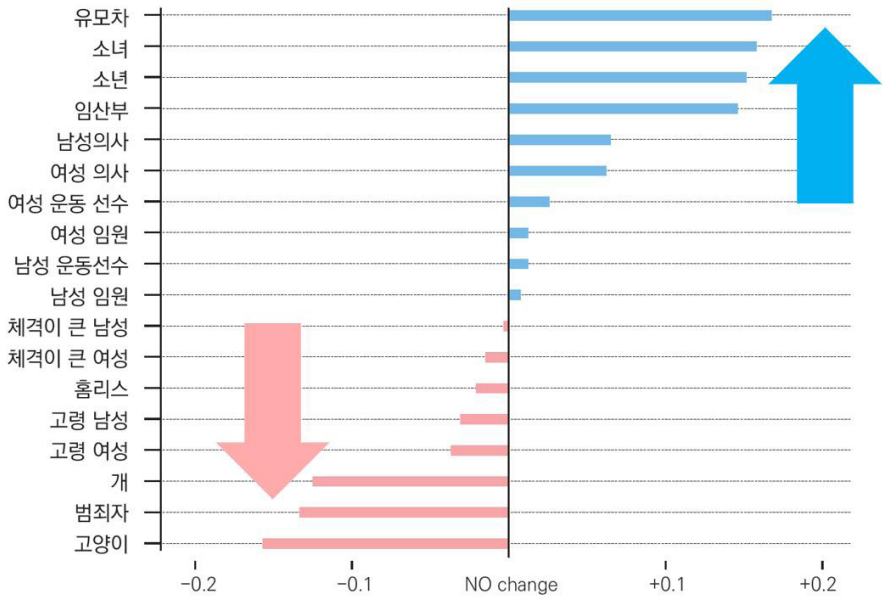
출처 : 모럴 머신 웹사이트

일반적으로 사람들이 생각하는 윤리 기준이 보편적으로 어디에 더 비중을 두는지 확인하고, 이를 통해 공론화된 공정성 기준을 도출해보고자 했던 것이다. 객체별 우선순위를 확인한 결과, 유모차나 소년, 소녀 등 어린이들을 상대적으로 중요하게 생각한다는 결과를 얻었고, 개, 고양이, 범죄자는 상대적으로 덜 중요하게 생각한다는 결과를 확인했다. 이처럼 일반 사람들이 생각하는 윤리 기준에 전반적으로 부합하는 결과가 나온 것을 확인했고, 이로써 소위 공론화된 공정성 기준을 도출하였다.

하지만 모럴머신도 두 가지 한계점이 존재한다. 첫째는 빅데이터 수집 비용이 너무 크다는 것이다. 모럴머신은 공정성 기준을 도출하기 위해 18개월 동안 230만 명의 데이터를 수집하면서 시간적으로나 금전적으로 엄청난 비용을 사용할 수밖에 없었다. 둘째는 이러한 빅데이터를 모았다고 해도 여전히 편향성 문제에서 자유로울 수 없다는 점이다. 샘플링 바이어스[1]가 있는 경우, 특정 집단의 지엽적인 의견이 반영될 수 있다.

1) 실험 전 이미 집단 간에 차이가 있어서 실험결과(해석)에 영향을 미치는 경우를 말한다.

그림 3 모럴머신의 객체별 우선순위 결과



E. Awad et al, 2018

이러한 한계를 극복하고자 도입된 새로운 방법론은 편향성 문제를 완전히 없애는 것이 어렵다고 인정하고 시작한다. 대신 알고리즘을 통해 인위적으로 공정성을 보장하는 방식을 취한다. 즉, 공정성 지표(fairness metrics)를 잘 '코딩'하여 이를 AI에 학습시키는 것이다. 이 방법론을 많은 연구자들이 채택하면서 '알고리즘 공정성'이라는 학술적인 명칭이 등장하기도 했다. 버클리, 막스플랑크 등 학술기관에서 추가 연구들이 나왔고, 구글, 아마존과 같은 글로벌 기업도 이러한 연구를 진행하고 있다. 국내에서도 2018년부터 연구를 시작하여 최근 다수의 결과가 나오고 있다.

나. 공정성 있는 AI를 만들기 위해 해결해야 하는 문제들

공정성을 갖춘 AI를 만들기 위해 앞으로 해결해야 할 도전적인 과제는 여전히 남아 있다. 공정한 알고리즘 개발을 위해서는 우선 공정성을 수치화하는 과정이 필요한데, 어느 정도의 값을 공정하다고 규정할 수 있을지에 대한 기준을 정하기가 어렵다. 다양하고 복잡한 상황에서 공정성 개념을 다룰수록 이 문제는 더욱 복잡하고 어려워진다.

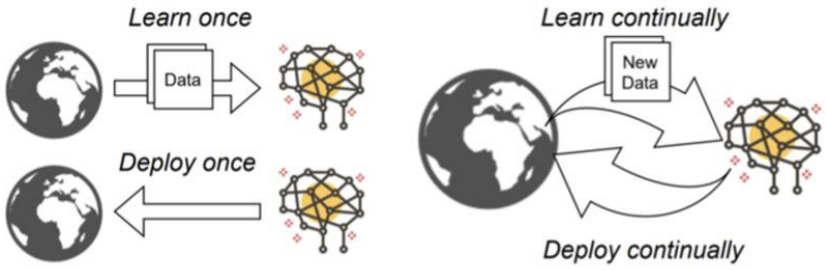
공정성 개념이 시대에 따라 변할 수 있다는 것도 도전적 과제이다. 사람과 사회의 가치관이 변할 수 있기에 공정성 개념 또한 언제든 변할 수 있고, 이에 따라 법이나 정책도 바뀔 수 있다. 이에 대응하기 위해 유연하게 동작하는 AI를 만들어야 하는데, AI는 보통 과거 데이터를 통해 훈련되기 때문에 유연한 대응에 한계가 존재하기 마련이다. 즉, 현재와 과거 데이터 사이에 간극이 있는 경우 좋은 일반화 성능을 보장하는 공정한 AI 개발에 대한 중요성은 더욱 높아질 것이다.

2 견고성, 낯선 상황에도 안정적으로 작동하는 AI

복잡한 문제에 직면한 AI가 모든 경우에 대한 최적의 대응 방법을 사전에 학습하는 것은 쉽지 않다. 이러한 점 때문에 AI가 예상치 못한 오류에 얼마나 잘 대처하면서 안정적으로 작동하는가에 대한 '견고성' 문제가 대두된다.

AI가 견고성을 갖추지 못하면 어떤 문제가 생길까? 자율주행 자동차 사례를 살펴보자. 학습 에이전트에게 주어지는 환경은 도로 상태, 자동차 대수, 주변 지형지물 등에 따라 결정된다. 주행 상황에 대한 경우의 수는 무수히 많고 그 상태는 시간에 따라 연속적으로 변화한다. 하지만, 제한적인 환경에서 학습한 AI는 경험해 보지 못한 환경에서 최적의 동작을 가져갈 수 없다. AI 기반 질병 진단 시스템 사례 또한 AI의 견고성 문제를 명확히 보여준다. 전 세계를 강타한 전염병 코로나19(COVID-19)가 창궐하면서, 확산 예측을 위한 수백 개의 AI 모델들이 개발되었으나 시시각각 변하는 현실을 반영하지 못해 대부분 실패로 돌아갔다. 이처럼 AI는 새로 직면하는 상황에 대한 대응력이 떨어지기 때문에 견고하고 안정적인 작동이 어렵다.

그림 4 고정학습(static learning)을 적용한 기계학습(왼쪽)과 연속학습(continual learning)을 적용한 기계학습(오른쪽)의 개념도



출처 : https://ai.kuleuven.be/stories/post/2021-05-10-continual-learning/

앞선 사례들은 시간에 따라 변하는 실제 환경에 발맞추어 AI 학습 모델을 지속적으로 업데이트해야 한다는 사실을 시사한다. 그러나 모델 적용과 동시에 지속적인 학습을 수행하기 위해 학습 단계에 기존의 오프라인 알고리즘을 그대로 적용하면 다양한 문제가 발생할 수 있다. 매우 많은 양의 데이터가 필요하기도 하고 학습에 소요되는 시간과 계산량이 크게 늘어날 수 있기 때문이다.

AI가 자동화 시스템과 같이 '관찰되는 상황'에서 최적의 통제를 수행하는 정책을 학습하는 경우가 특히 그렇다. 학습을 위해서는 일반적으로 여러 임의의 행동을 수행하며 최적의 정책을 찾는 탐색(exploration) 과정이 필수적인데, 이때 오프라인 학습 과정에서 고려하지 못한 심각한 성능 저하의 문제가 발생할 수 있다. 예를 들어 자율주행 자동차가 실제 운행 중 데이터의 수집을 위하여 무작위적인 탐색 과정을 거치게 되면, 단순한 성능 저하를 넘어서 회복 불가능한 상황에 이르게 될 수도 있다.

가. 견고성을 갖춘 AI를 위한 다양한 연구들

최근까지도 심층신경망 기반 AI 영역에서 앞서 언급한 문제의 해결 방법과 관련된 연구가 활발히 진행되고 있다. 대표적으로 단순히 주어진 하나 이상의 과제를 하나의 모델로 학습하는 멀티태스킹 학습(multi-task learning) 기법부터, 사전에 학습된 지식을

비슷하지만 다른 과제가 주어졌을 때 활용하여 새로운 지식을 더 효율적으로 학습하는 영역적응(domain adaptation) 기법, 나아가 경험하는 과제의 범위가 넓어짐에 따라 지식을 계속해서 축적해 나가는 영역확장(domain expansion) 기법까지 다양한 시도들이 이루어지고 있다.

멀티태스킹 학습 기법은 하나의 환경이 아닌 주어진 과제의 집합에 대하여 과제 간의 유사성과 차이점을 찾아 활용하여 모든 과제에 대해 잘 동작하는 하나의 모델을 학습한다. 영역적응 기법의 한 영역인 전이학습(transfer learning)은 하나의 환경에 대해 학습된 심층신경망 모델의 가중치 값의 일부 또는 전부를 비슷하지만 다른 문제를 학습할 때 가져와 더 적은 데이터로 학습하는 데에 중점을 둔다. 영역적응의 또 다른 예로 메타학습(meta-learning)은 비슷한 여러 과제로부터 학습한 지식을 이용하여 다른 환경이 주어졌을 때 더 빠르고 효율적으로 학습하는 방법 자체를 학습하도록 한다. 증분식 학습(incremental learning)은 앞서 언급한 영역확장 기법의 과정 중 이전에 학습한 지식을 보존하지 못하는 문제인 심층신경망의 파괴적 망각(catastrophic forgetting) 문제를 해결하고자 하는 데에 초점을 둔다.

그러나 위 연구들은 새롭게 주어진 학습 환경이 학습 과정에서 경험한 환경과 크게 다를 경우 대응하지 못한다는 한계가 있다. 즉, 일정한 수준의 변화를 넘어서는 환경 변화에 대해서는 처음부터 다시 학습을 시작하는 것 이상의 해결책을 제시하지 못한다. 최근에는 이러한 OOD(Out-Of-Distribution) 환경에 대한 감지 및 대응에 대한 연구도 이루어지고 있으나 아직 시작 단계에 불과한 실정이다.

나. 연속학습, 견고한 AI의 개발을 위한 가능성 제시

현실 문제에 적용 가능한 견고한 AI의 개발을 위해 시간에 따라 변화하는 다양한 환경에 대응 가능하고 새로운 지식을 지속적으로 학습하는 연속학습(continual learning) 기법에 대한 연구가 활발히 이루어져야 할 것으로 판단된다.

연속학습은 안정적이고 견고한 학습 기반 문제의 해결을 위해 변화하는 환경에 대응하여 최적의 선택을 하고, 나아가 경험 해보지 않은 환경에 대해서도 기존의 학습 기반 예측을 통한 적용과 동시에 학습을 수행하는 것을 의미한다. 이를 위해서는 학습 도중에

발생하는 성능 저하를 최소화할 수 있도록 신속하고 효율적으로 학습할 수 있어야 하며, 새롭게 학습된 내용에 의해 이전에 학습된 내용이 오염되어서는 안 된다.

이는 곧, 보다 궁극적인 견고성 확보를 위해 실제 문제에 대한 적용과 동시에 학습을 수행하는 형태(Learning on the job)를 가지는 많은 새로운 연구들이 수행될 필요가 있음을 의미한다. 이를 위해서는 학습 과정에서의 성능의 저하를 최소화하면서 학습에 필요한 데이터를 확보하는 문제를 선제적으로 해결할 필요가 있다.

3 설명가능성, 누구나 이해하고 공감할 수 있는 AI

가. AI가 주는 결과뿐만 아니라 과정도 이해해야

AI는 대용량 데이터를 기반으로 예측을 수행한다. 이때 데이터에 존재하는 예상하지 못한 특이한 패턴이나 문제의 소지가 있는 정보를 바탕으로 판단을 내릴 수 있다. 따라서 AI의 판단 과정이 충분히 설명되지 않는다면, 사용자는 AI가 내린 잘못된 결과에 대해서 무심코 받아들일 가능성이 높다.

직원 채용이나 승진 결정, 대출 심사, 의료 분야 등 민감한 영역일수록 설명가능성을 갖춘 AI에 대한 중요성은 더욱 높아진다고 할 수 있다. 특히 의료 분야에서 AI의 판단에 대한 명확한 설명을 제공하는 것은 필수적이다. 병원에서 사용되는 의료 AI는 잘못된 판단을 내리면 불필요한 약물 투여나 수술이 행해질 수 있고 이는 인체에 치명적인 결과를 낳을 수 있으며 큰 위험을 초래할 수 있기 때문이다.

인터넷 포털 및 쇼핑 사이트에서 발생하는 불공정한 뉴스 편집과 제품 노출도 문제가 되고 있다. 일부 기업들은 AI에 판단을 맡겼다는 것을 이유로 책임을 회피하기도 한다. 이러한 사회문제에 대한 책임 소재를 분명히 하기 위해 설명가능 AI 도입과 이에 대한 투명한 공개가 필요하다.

나. XAI 원칙과 주요 연구 현황

설명가능성을 갖춘 AI를 '설명가능 AI(explainable AI)', 줄여서 XAI라고 부른다. 2020년 미국 국립표준기술원(NIST) 보고서[2]에서 채택한 정의에 따르면, XAI 원칙을 아래와 같은 4가지(설명, 유의미, 설명정확성, 지식한계)로 요약했다. 이 원칙들은 AI가 제공해야 할 '설명'이 구체적으로 어떤 것을 의미하는지 규정하고 있다.

XAI의 4가지 원칙

설명(explanation)
인공지능 시스템은 각 출력에 대해 적절한 증거, 뒷받침, 또는 논리를 제공해야 한다.

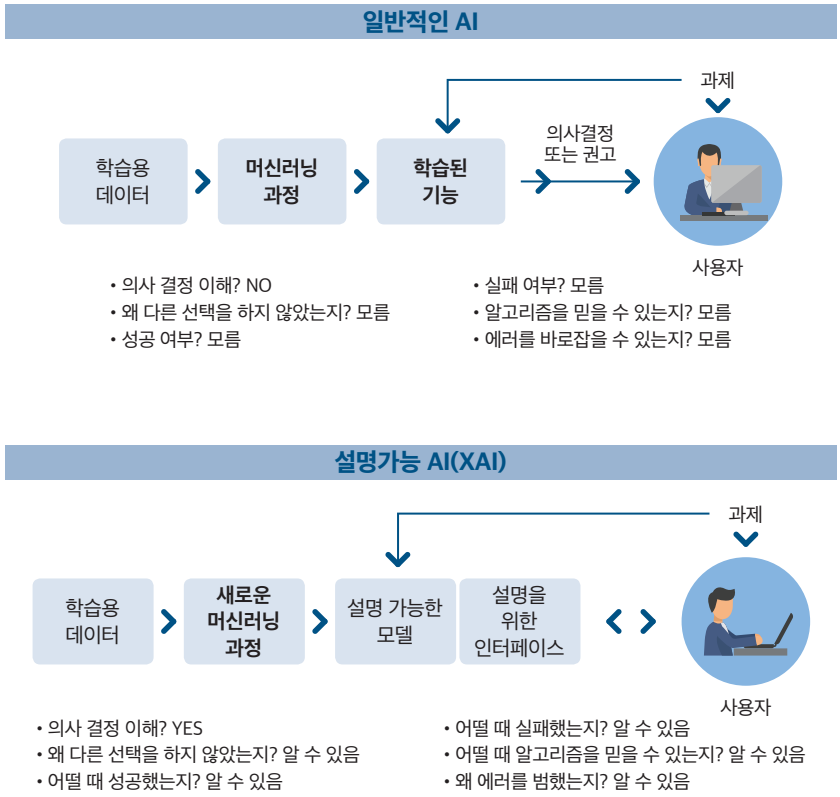
유의미(meaningful)
인공지능 시스템은 사용자가 이해할 수 있는 설명을 제공해야 한다.

설명정확성(explanation accuracy)
설명은 그 인공지능 시스템이 해당 결과를 얻기 위해 사용했던 과정을 정확하게 반영해야 한다.

지식한계(knowledge limits)
인공지능 시스템은 설계 목적에 맞는 조건 내에서 동작해야 하며, 결과에 대한 충분한 확신이 없을 때에는 결과를 제공하지 말아야 한다. 또한, 인공지능이 자신의 동작조건을 인식하고 해당조건을 벗어난 경우, 이에 따른 불확실성에 대해서도 설명해야 하는 의무가 있다.

2) NIST 보고서 : Four Pinciples of Explainable Artificial3 intelligence, NISTIR 8312

그림 5 일반적인 AI와 설명가능 AI(XAI)의 차이점



XAI는 새로운 기계학습 과정, 설명가능한 모델, 설명 인터페이스 등을 제공하여, 사용자로 하여금 AI 시스템의 예측 결과에 대한 다양한 설명을 제공한다

XAI 연구는 크게 설명가능한 예측, 설명가능한 알고리즘, 설명가능한 데이터 등 세 가지 측면으로 진행되고 있다. 오늘날 진행되고 있는 상당수의 연구는 '설명가능한 예측'과 '설명가능한 알고리즘'을 연구하는데 집중되어 있다.

가장 대표적인 연구는 신경망 계층의 어떤 부분이 활성화되었는지 역으로 추적하여 입력 데이터의 어떤 영역과 특징을 사용하였는지 시각화하는 것이다. 또한 학습 과정에서 신경망의 중요 특징들이 서로 얽히지 않도록, 설명하기 쉽도록 만드는 기술들이 개발되고 있으며, 설명하기 용이한 모델(결정트리, 베이지안 신경망, 그래프 신경망 등)을 활용하여 설명력을 높이는 방식 등이 활발히 제안되고 있다.

XAI의 주요 연구 분야

설명가능한 예측
어떤 입력에 대한 출력을 얻은 경우, 입력의 어떤 특징들이 사용되었으며 얼마만큼 영향을 주었는지에 대한 설명을 제공하는 것을 말한다.

설명가능한 알고리즘
알고리즘(모델)을 구성하는 개별적인 계층 및 모듈은 무엇이며, 이들이 출력을 산출하는데 어떻게 기여하는지 분석한다.

설명가능한 데이터
학습 데이터가 가진 특성을 바탕으로 적절성, 공정성, 편견에 대한 연구로서, 데이터로부터 편견을 측정하고 이를 제고하는 문제들을 탐구한다.

다. 설명가능 AI가 인간 수준에 도달하기 위해 필요한 것들

설명가능성에 대한 연구는 AI가 주는 결과에 대한 이해와 신뢰를 위해서도 중요하지만, AI가 인간 수준에 도달하기 위해 극복해야 할 문제와도 연관된다. 인간은 근거와 논리를 통해 생각을 정리하고 결론에 도달하기 때문에 이러한 과정을 설명하는 것이 결과와 자연스럽게 연결된다. 심층학습 기법은 이러한 인간의 고차원적 사고를 모사하기가 어려우며, 이것이 곧 설명가능성 문제로 이어졌다고 볼 수 있다. 설명가능성 문제는 인간 수준의 AI 연구가 발전하면서 함께 해결될 가능성이 높다. 이러한 관점에서 설명가능성을 위한 두 가지 방향성에 대한 논의가 활발하게 이루어지고 있다.

첫째는 '뉴로-심볼릭(neuro-symbolic) AI'이다. 이는 인간의 논리적 추론 과정이 갖는 체계성을 모사하는 고전적인 심볼릭(symbolic) AI 방법론을 심층학습과 결합하려는 연구를 말한다. 이를 통해 알고리즘의 추론과정이 자연스럽게 설명됨과 동시에, 기존 심층신경망의 단점인 조합능력[3]을 대폭 개선할 것으로 기대하고 있다.

3) 조합능력(conpositionality) : 부분을 재조합하여 새로운 구조를 이해하는 능력

둘째는 ‘인과추론과(causal inference) AI’이다. 이는 학습과정에서 인과관계를 모델링하고 이를 추론에 활용하여, 예측에 대한 인과적인 설명을 제공할 수 있는 알고리즘에 대한 연구를 말한다. 기존 AI가 입력과 출력의 상관관계를 파악해 예측을 수행한 것과 달리, 학습데이터를 대상으로 원인과 결과에 대한 분석을 수행하고 인과적 설명에 부합하는 예측을 수행한다. 예측에 대해 자연스럽게 인과적 설명을 제공하면서도 데이터 편향을 극복하고 예측 정확도를 높이는데 기여할 수 있다.

4 투명성, 속이 훤히 들여다보이는 AI

투명성은 설명가능성과 함께 강조되는 책임감 있는 AI를 위한 중요한 조건 중 하나다. 투명성은 크게 AI가 인간에 의해 관리·감독이 이루어질 수 있도록 설계 및 실행되어야 한다는 점과 AI가 무엇을 어떤 이유로 하고 있는지에 대한 설명이 이해하기 쉽고 평가가 가능한 형태로 제공되어야 한다는 점을 주요 내용으로 한다.

가. AI가 투명성을 갖추지 못하면 생기는 문제들

알고리즘 관점에서 AI의 투명성은 어떠한 머신러닝 알고리즘이 활용되었는지, 알고리즘이 어떻게 작동하는지, 어떻게 학습·검증하였는지, 결과에 가장 중요한 영향을 미치는 변수가 무엇인지를 투명하게 공개하는 것을 의미한다.

하지만, AI의 투명성은 알고리즘의 투명성 이상을 의미한다. 어떻게 개발되었는지 뿐만 아니라 어떻게 운영되고 관리되는지 전반적으로 이해할 수 있어야 진정으로 투명한 AI라고 할 수 있다. 어떠한 기술이든 해당 기술이 어떻게 작동하는지 투명하게 알 수 없는 상황에서 기술의 활용성과 신뢰성은 떨어질 수밖에 없다.

AI의 투명성은 공정성과 마찬가지로 다양한 분야에서 문제가 되었다. AI 면접의 경우, 어떠한 지표(성과, 이직, 인성 등)를 기준으로 이루어지며, 어떠한 요소(표정, 말투 등)들이 면접 평가에 반영되는지와 함께, 어떠한 데이터와 알고리즘에 의해 AI

시스템이 학습되었는지, 내가 AI 면접에서 떨어졌다면, 왜 떨어졌는지 등을 투명하게 설명할 수 있어야 한다. 그래야 AI 시스템을 신뢰하고 결과에 승복할 수 있기 때문이다. 우리나라에서도 최근 수원지방법원, 광주지방법원이 각각 한국국제협력단, 한전KDN 등 공공기관에 AI 면접과 관련된 정보를 공개하라는 판결을 내리기도 하였다.

네덜란드 법원 사례도 주목할 만하다. 네덜란드 법원은 사회보장 시스템의 복지 혜택과 관련한 부정 수급 가능성을 예측하는 AI 시스템에 대해 해당 시스템이 비공개 알고리즘을 기반으로 구축되었다는 이유로 인권법과 EU 일반정보보호법(general data protection regulation)을 위반했다고 판단했다. 결국 네덜란드 정부는 해당 AI 시스템의 사용을 중지시켰다.

나. AI의 투명성을 높이기 위한 다양한 논의들

AI의 투명성을 높이기 위해 데이터 보호나 프라이버시, 평등, 사용자 보호, 제품 안정성과 책임 이슈 등과 관련된 데이터에 대한 연구가 이루어지고 있다. 또 의사결정의 투명성을 높이기 위한 연구 등이 진행되고 있다. 특히 투명성과 관련해서는 아래와 같은 다섯 가지 이슈가 문제시 될 수 있어, 관련 논의들이 활발히 이루어지고 있다.

첫째는, 영업비밀로 인식될 수 있는 AI 코드와 데이터의 소유권과 관련된 이슈가 있다. 예를 들면 기업에서의 채용과정은 영업비밀로 인식될 수 있는데, 이런 영역에서도 AI 알고리즘을 투명하게 공개할 수 있을 것인가에 대한 것이다. 나아가 무분별한 공개의무는 모델 탈취(model stealing) 공격을 유발할 수도 있다.

둘째는, 너무 많은 것이 투명하게 공개될 경우 원래 AI 시스템의 목적을 벗어나 시스템이 오용될 위험이 있다는 것이다. AI 면접에서 어떠한 표정과 말투, 제스처를 통해 높은 점수를 획득할 수 있을지 알 수 있다면, 의도적으로 이를 수행하여 면접에서 높은 점수를 획득할 가능성이 있다. 이러한 '게이밍(gaming)' 문제는 AI에 기한 분류 시스템의 존재 가치를 무색케 할 수 있는 중요한 문제다.

셋째는, 데이터와 알고리즘이 사용자에게 얼마나 쉽게 설명될 수 있을지에 대한 이슈가 있다. 이는 앞서 살펴본 AI의 설명가능성 문제와 동일하다고 볼 수 있다.

넷째는, AI 시스템을 운영하는 과정에서 수집된 개인 데이터가 어떻게 활용되는지, 또한 AI 시스템을 구축하는 과정에서 활용된 데이터가 어디에서 확보된 것인지 등 데이터와 관련된 이슈가 있다. 예를 들면, AI 면접 시스템은 어떠한 데이터를 활용하여 개발되었는지, AI 면접과정에서 수집된 나의 데이터가 어떻게 활용될 것인지 등이 문제가 될 수 있다.

다섯째는, AI 시스템이 구조적 차별이나 다른 불공평한 결과를 가져오는지 여부를 감지할 수 있도록 하는 방법에 대한 이슈가 있다.

AI의 투명성을 높이기 위해서는 설명가능한 AI 기술 개발 연구가 지속될 필요가 있다. 특히 AI 기술과 시스템을 설계하는 단계에서부터, 해당 기술이 사회에 미칠 수 있는 긍정적, 부정적 영향을 다양한 관점에서 예측하고 이에 대한 대응방안을 적극 모색해야 한다. 또한 AI의 투명성을 높이는 과정에서는 다양한 이해관계자들의 지속적 협의와 논의가 필요하다.

5 가치정렬, 인간의 가치와 잘 부합하는 AI

AI는 인간의 가치와 잘 부합하도록 개발되어야 하며, 이러한 개념을 '가치정렬'이라고 부른다. 가치정렬에 어긋나는 AI 사례 또한 다양하게 나타나고 있다. 2018년 우버(Uber)는 자율주행 자동차가 시범운행을 하며 무단횡단하는 행인을 인지하지 못하여 사고를 냈다. 국내에서는 2021년 AI 챗봇이 의도하지 않게 성차별, 인종차별적 발언을 한 사건이 유명하다. 이처럼 인간의 가치와 부합하는 의사결정을 내릴 수 있는 신뢰도 높은 시스템이 개발되지 못했을 때, 사회에서 AI의 활용은 제한될 수밖에 없다.

가. 기술적 관점에서의 가치정렬

기술적 관점에서 보면, 상당수의 AI에 채용되는 강화학습(reinforcement learning)은 자신이 받는 보상을 최대화 하는 방향으로 행동하도록 학습한다. 즉, 시행착오(trial-

and-error)를 거쳐 보상이 최대가 되도록 지속적으로 알고리즘을 수정하면서 점점 더 좋은 성능을 내게 된다. 하지만, 보상에 의해서만 행동하는 것은 결과만을 따라 행동하는 것에 국한된다는 문제가 있다. 인간은 결과만 지향하는 것이 아니라 스스로가 갖고 있는 다양한 동기에 따라 행동한다.

이러한 한계를 극복하기 위해 최근에는 역강화학습(Inverse Reinforcement Learning, IRL) 접근법이 제안되고 있다. 역강화학습 가운데, 가장 대표적인 모방학습(imitation learning)은 AI가 인간 전문가의 보상함수를 추론하여 그 행동을 따라하는 방법론을 말한다. 한 명의 전문가를 따라하는 대신 인간의 행위에 대한 대량의 데이터베이스로부터 보상함수를 추론하는 방법도 있다. 이는 인간의 행동 자체 대신에 인간의 행동이 이루어지는 동기 요인을 모델링하는 접근법이다. 한층 더 진보한 방법론으로는 다양한 정책에 따라 환경과 상호작용하는 여러 AI들의 행위를 전 생애동안 평가하여, 전체 보상이 최대화되는 행위들을 선택하는 접근법이 연구되고 있다.

나. 규범적 관점에서의 가치정렬

가치정렬은 기술적 관점에 대한 연구와 더불어 규범적 관점에 대한 활발한 논의가 진행될 필요가 있다. AI의 기술적 접근법이 성공적으로 적용되기 위해서는 여러 규범적인 관점이 고려되어야 한다. 가치정렬을 위해 규범적 관점에서 던져야 하는 주요한 4가지 질문은 아래와 같이 정리할 수 있다.

규범적 관점에서 AI의 가치정렬을 위한 4가지 질문

❶ AI가 모방해야 하는 도덕적 전문가는 누구인가?

❷ 어떠한 데이터로부터 가치에 대한 개념(보상함수로 표현되는)을 추출해야 하며 이를 어떻게 추출할 것인가?

❸ 데이터는 모두의 행위에 대한 데이터여야 할까?
(비윤리적이고 합리적이지 않은 행위에 대한 데이터는 제외해야 할 것인가?)

❹ 어떠한 인공지능이 가장 도덕적이라는 판단은 어떠한 기준으로 내릴 것이며, 이러한 방식으로 우선순위를 매기는 것이 가능한가?

AI는 내가 시키는대로(instructions), 내가 표현한 의도대로(expressed intentions), 내 행동에서 나타나는 선호도에 따라(revealed preferences), 나의 선호도나 바람에 따라(informed preferences or desires), 나의 이익이나 복지에 따라(interest or well-being) 다양하게 행동할 수 있다. AI는 개인이나 사회가 정의한 바에 따라 일을 도덕적으로 수행해야 하므로 이와 관련한 윤리적 원칙을 세우는 것이 필요하다. 이를 위해서는 기술적인 관점뿐만 아니라, 일반적으로 공정하다고 알려져 있는 원칙을 규명하는 규범적인 관점 또한 중점적으로 다루어야 할 것이다.

6 프라이버시, 개인정보를 식별하지 않는 AI

개인정보와 관련된 데이터를 기반으로 작동하는 AI는 프라이버시 문제를 야기할 수 있다. 프라이버시는 곧 데이터의 개인 식별성이 없는 상태를 의미한다. 이러한 프라이버시 문제는 특히 보건의료 측면에서 두드러지게 나타난다. 보건의료 AI는 정확한 진단을 위해 유전 정보, 망막 정보, 정맥 분포, 골격 등의 개인정보를 수집한다. 따라서 충분한 보안을 발휘하지 않으면, 이들 데이터를 이용하여 개인을 식별할 수 있는 문제가 발생할 수 있다. 망막, 유전자 등의 데이터 자체는 특정 개인과 1:1 대응(single out)할 수 있을 정도로 강력하다.

일부가 익명화처리 된다고 하더라도 각 정보 등의 재조합을 통해 개인을 식별할 수도 있다. 또 희귀질환과 내원 일자 같은 자료의 조합을 통해서는 개인을 거의 정확하게 식별할 수 있다. 데이터 전체가 충분히 보호된다고 하더라도, AI 알고리즘 자체를 분석하거나 입력 패턴을 조금씩 계속 바꾸어 입력하는 방식을 사용하면 문제가 생길 수 있다. 만약 '흔하지 않은 개인'이 있다면 이와 같은 과정을 통해 해당 개인의 다른 자료를 AI의 반응을 통해 유추할 수 있다. 유명인의 성병, 정신질환과 같은 민감한 자료가 유추된다면 사회적 파장은 더욱 커질 수 있다.

가. 프라이버시 문제 해결을 위해 수행되고 있는 최신연구들

AI 프라이버시 문제를 해결하기 위해 어떤 연구들이 수행되고 있을까? 가장 대표적인 프라이버시 관련 연구는 차분프라이버시(differential privacy), 기밀컴퓨팅(confidential computing), 동형암호(homomorhpic encryption), 연합학습(federated learning), 비식별처리(de-identification) 등이 있다.

차분프라이버시는 질의(query)에 대응하여 통계 생산 시 정교하게 설계된 잡음(noise)을 데이터에 부가함으로써 각각의 개인이 데이터베이스에 있을 때 생산된 정보와 없을 때 생산된 정보의 차이를 없애 프라이버시를 보호하는 기술이다.

기밀컴퓨팅은 하드웨어에 구현된 보안 기능을 활용해 데이터를 유출하지 않는 안전한 계산 및 AI 학습을 하는 기술을 말한다. 기밀컴퓨팅은 마이크로소프트 애져(Microsoft Azure), 구글 클라우드(Google Cloud), 아마존 EC2(Amazon EC2) 등 클라우드컴퓨팅 서비스를 통해서도 제공되고 있다. 동형암호는 암호화된 데이터로 계산 수행이 가능하게 하는 기술로, 암호화된 데이터를 통해 계산 및 AI 학습을 진행하므로 데이터 유출이 없다는 장점이 있다.

연합학습은 AI 학습을 위한 데이터를 중앙 서버에서 한꺼번에 처리하는 것이 아니라, 각 객체에서 개별적으로 부분적 학습을 수행하는 것을 말한다. 추후 이러한 개별적으로 학습된 모델의 파라미터나 하이퍼파라미터를 중앙서버에서 조합하여 최종 AI모델을 학습한다. 따라서 학습을 위한 데이터가 개별 객체에만 머무르고, 외부 중앙서버로 전달되지 않으므로 개인정보 데이터를 보호할 수 있다.

데이터 비식별처리는 데이터에 개인을 식별할 수 있는 민감한 데이터를 일반화, 범주화, 익명처리, 마스킹 등의 처리를 통해 특정 개인으로 식별하지 못하도록 하는 기술을 말한다.

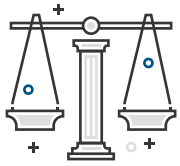
위 기술들 중 일부는 구체화 되고 구현되기 시작한지 얼마 되지 않은 신생 연구들이다. 앞으로는 이러한 프라이버시 기술들을 활용하여 실제 AI 서비스들을 구현하는 연구가 필요할 것으로 전망된다. 특히 현재의 AI 서비스들은 복잡한 컴퓨팅 시스템 위에서 높은

리소스와 많은 데이터에 기반하여 동작하므로, 이러한 컴퓨팅 요구사항(computing requirement)을 만족할 수 있는 실용적인 보호 기술에 대한 연구가 수행되어야 할 것이다.

II

정책제언

1 AI 관련 법·제도 마련과 윤리기준 확산이 필요하다



AI가 정치, 경제, 문화 등 다양한 영역에서 성장하고 영향력이 커지는 만큼, AI는 인간 삶에 큰 영향을 미치고 있다. 이러한 기조 속에서 책임성 있는 AI를 만들기 위한 법제도 및 정책에 대한 필요성은 더욱 높아지고 있다. 해외의 경우 각종 AI 법안과 가이드라인, ISO/IEC와 NIST 표준, IEEE 보고서 등을 통해 다양한 정책들이 제시되고 있다. 우리나라 또한 AI 법·제도와 정책을 강화하기 위한 노력을 지속하고 있다.

정부는 2021년 AI 기술 수준과 국내외 법제 동향을 분석하여 종합적이고 선제적인 법·제도 로드맵을 마련하여, 국내 AI 산업의 투명성과 공정성을 유도하고 있다. 기술적인 제재를 가하는 방안으로 '표준화'도 활발히 이루어지고 있다. 표준화는 AI의 공정성, 윤리성 등을 보장하기 위해 특정한 원칙을 지켜야만 제품으로 승인해주는 규범을 말한다. 한국정보통신기술협회(TTA)는 데이터 수집, AI 개발, 활용 전 단계에 신뢰성 및 공정성과 관련된 원칙을 수립했다. 신뢰성 판단을 위한 체크리스트를 만들고, 공정성을 어떻게 수치화하여 판단할지에 대한 기준을 마련했다. 이러한 노력들은 책임성을 갖춘 AI를 개발하기 위해 매우 고무적이라고 할 수 있다.

향후에는 AI의 편향성 문제를 다룰 수 있는 법체계가 정립되어야 할 것으로 판단되고 있다. AI의 편향성은 공정성과 투명성, 설명가능성 등 책임성 있는 AI를 위한 조건들과 연결되는 문제다. 정부와 민간기업이 AI의 실제 사용과정에서 발생했던 편향성 이슈들을 종합하고 분석함으로써 발생 가능한 문제를 사전에 예측할 수 있도록 지원할 필요가 있다.

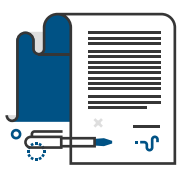
AI 프라이버시와 관련해서는 '데이터 주권'에 대한 논의와 함께 국가별로 새로운 법체계가 다수 생겨나고 있다. 대표적으로는 유럽연합(EU)의 일반데이터보호규정(GDPR), 미국 캘리포니아의 소비자 프라이버시법(CCPA), 중국의 개인정보보호법(PIPL)이 있으며, 우리나라는 개인정보보호법, 정보통신망법, 신용정보법을 2020년 대폭 개정하여 시행하였다. 프라이버시 법안 및 정책들은 궁극적으로 개인정보의 안전과 AI 산업을 보호하면서도, 글로벌 기술 트렌드에 뒤쳐지지 않도록 균형적으로 이루어져야 할 것으로 판단된다. 또한, 각 국가별로 다양한 형태의 규제가 자리 잡고 있고 법안도 더욱 다양화되고 구체화될 것으로 예상되는 만큼, 현지 법·제도에 부합하는 AI 솔루션을 개발하는 것도 중요한 과제로 부상하고 있다.

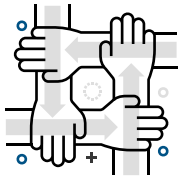
법·제도를 통한 규제는 자칫 산업 발전을 저해할 수 있기 때문에, AI가 활용되는 영역의 특수성을 고려할 필요가 있다. 예를 들면, 채용 및 신용 평가를 수행하는 AI는 공정성 및 투명성을 갖추어야 하고, 보건의료 AI는 설명가능성이 중요하며, 자율주행차나 AI 승강기는 안전성이 중요할 것이다. 반면, AI 청소기는 이런 조건들이 상대적으로 덜 중요할 것이다. 현재 EU가 준비 중인 AI 법안(draft AI regulation)처럼 특정 범주의 AI에만 들어가면 공정성, 투명성, 설명가능성, 안전성, 견고성, 사이버보안 등을 모두 준수하도록 하는 과잉규제는 지양해야 한다. 이처럼 기술 혁신과 책임성 간 균형감각을

통해 현재 쏟아지고 있는 법안을 합리화하기 위해 다양한 융합 연구가 선행될 필요가 있으며, 그다음 합리적인 규범 체계를 확립해 가야 할 것이다.

법·제도를 통한 강제적인 방법뿐만 아니라 사회적 공감을 이끌어낼 수 있는 유연한 접근도 필요하다. AI 윤리 기준 발표를 통한 업계의 자발적인 노력이 이러한 사례라고 할 수 있다. 과학기술정보통신부는 2020년 「인공지능(AI) 윤리기준」을 마련했으며, 뒤이어 네이버, 카카오, SK텔레콤, 삼성전자 등 국내 주요 IT 기업들도 나름의 AI 윤리 기준을 만들어 발표했다. 이러한 AI 윤리 기준은 법·제도만큼 강하지는 않지만, 윤리적인 AI에 대한 원칙을 공유하고 사회적인 합의를 이끌어낼 수 있으며, 책임성 있는 AI를 만들기 위한 좋은 가이드라인을 제시할 수 있다.

2 현재 기술의 한계를 이해하고 대처해야 한다



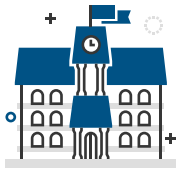


현재 AI 기술에 대한 한계를 이해하고 이에 따른 연구개발 관련 정책 마련도 필요하다. 데이터 기반 다중 계층 신경망 구조의 파라미터 최적화를 기본으로 하는 현재의 심층학습 기법의 한계에 대한 비판은 지속적으로 이루어져 왔으나, 오늘날 상대적으로 주목받지 못하고 있다.

현재는 잘 알려진 심층학습 구조하에서 성능을 향상시키는 것에 대부분의 연구개발 역량이 집중되고 있는 실정이다. 예컨대 견고한 AI를 만들기 위해 연속학습을 가능하게 하는 효율적 학습 구조, 지식 저장 구조에 대한 연구는 현재까지 크게 진행되지 못하고 있다. 이처럼 현재 기술이 가지는 본질적인 한계에 대한 명확한 이해 없이 향후에도 지금까지 해온 것과 같은 방식으로 연구개발을 수행한다면 국가적 차원에서 큰 낭비를 감수할 수밖에 없을 것이다.

AI의 주요 기법인 심층학습의 한계를 인식하고 효율적인 역량 집중에 대한 필요성이 높아지고 있다. 이러한 논의를 정책적으로 확장하여, 현재 기계학습 및 심층학습의 한계점에 대해 개발자와 학습자들의 이해도를 높이는 것이 필요하다. 현 기술에 대한 한계를 이해함으로써, 미래에 보다 효율적이고 신뢰성 있는 AI 기술 개발에 집중할 수 있을 것이다.

3 윤리적인 인재 양성과 학제간 연계를 강화해야 한다



책임성 있는 AI를 개발하기 위해서는 AI 산업 종사자의 공정성과 윤리에 대한 올바른 개념이 확립되어야 한다. 이를 위해 장기적인 관점에서 윤리성을 겸비한 AI 인재 양성에 대한 요구도 높아지고 있다. 책임성 있는 AI의 핵심은 공정성, 투명성 등 윤리적인 개념을 기계에 심어주는 것이다. AI를 설계하는 것은 결국 인간이므로, AI를 개발하는 사람이 윤리성을 갖추어야 책임성 있는 AI의 개발이 가능할 것이다.

정부와 기업, 대학 등 유관기관들은 AI 투명성을 높이기 위한 원칙과 가이드라인을 협력하여 개발하는 한편, AI 시스템 개발에 관여하는 개발자뿐 아니라 사용자들에게도 해당 가이드라인을 핵심적인 교육 자료로 활용할 필요가 있다. 특히 대학의 AI 관련 대부분이 기술적 측면에 초점을 맞추고 있는데, 기초 교육에서부터 AI 윤리 교육에 대한 비중을 현재보다 높여야 할 것으로 판단되고 있다.

책임성 있는 AI는 인간과 밀접한 관련이 있기 때문에 다학제간 연계 및 협업도 요구되고 있다. 과학자 및 공학자에게 인문학, 법률, 사회학을 친숙하게 제공할 수 있는 교육 시스템을 구축해야 한다. 이를 통해 인문학자, 법학자와의 소통을 강화하여 다양한 협력 연구 로드맵을 개발해야 한다. 반대로 인문학자, 법학자들에게는 교양 수준의 AI 교육 제공을 통해, AI 기술과 친해질 수 있는 환경을 제공해야 할 것이다.

차세대리포트

2018 젊은 과학자들을 위한 R&D 정책은 무엇인가(上)
젊은 과학자들을 위한 R&D 정책은 무엇인가(下)
과학자가 되고 싶은 나라를 만드는 방법
영아카데미, 한국 과학의 더 나은 미래를 위한 엔진
10년 후 더 건강한 한국인을 위해 필요한 과학기술은 무엇인가?

2019 머신러닝, 인간처럼 보고 생각하고 예측하라
수소사회, 과학기술이 만들어가는 미래
양자기술, 과학은 끝없이 증명할 뿐이다

2020 뉴로모픽칩, 인간의 뇌를 담은 작은 반도체
대학의 미래, 젊은 과학자의 시선으로 바라보다
암과의 전쟁, 정복을 향한 꿈의 치료법
디지털 헬스케어, 건강관리의 새로운 패러다임

2021 자율주행, 그 이상의 모빌리티 생각하는 자동차
젊은 과학자의 눈으로 바라보다, 과학기술 2050
학령인구 절벽시대를 마주하다, 대학이 나아갈 길
새로운 팬데믹, 어떻게 준비해야 할까?

한국과학기술한림원은,

대한민국 과학기술분야를 대표하는 석학단체로서 1994년 설립되었습니다. 1,000여 명의 과학기술분야 석학들이 한국과학기술한림원의 회원이며, 각 회원의 지식과 역량을 결집하여 과학기술 발전에 기여하고자 노력해오고 있습니다. 그 일환으로 기초과학연구의 진흥기반 조성, 우수한 과학기술인의 발굴 및 활용 그리고 정책자문 관련 사업과 활동을 펼쳐오고 있습니다.

한림석학정책연구는,

우리나라의 중장기적 과학기술정책 및 과학기술분야 주요 현안에 대한 정책자문 사업으로 한국과학기술한림원 회원들이 직접 참여함으로써 과학기술분야 및 관련분야 전문가들의 식견을 담고 있습니다. 한림연구보고서, 차세대리포트 등 다양한 형태로 이루어지고 있으며 국회, 정부 등 정책 수요자와 국민들에게 필요한 정보와 지식을 전달하기 위하여 꾸준히 노력하고 있습니다.

한국과학기술한림원 더 알아보기

홈페이지 www.kast.or.kr
블로그 kast.tistory.com
포스트 post.naver.com/kast1994
페이스북 www.facebook.com/kastnews

이 사업은 복권기금 및 과학기술진흥기금 지원을 통한 사업으로 우리나라의 사회적 가치 증진에 기여하고 있습니다.

KAST 한국과학기술한림원
The Korean Academy of Science and Technology

(13630) 경기도 성남시 분당구 돌마로 42
Tel 031-726-7900 **Fax** 031-726-7909 **E-mail** kast@kast.or.kr

9 772635 716007 20
ISSN 2635-716X