인공지능 편향성 이슈와 신뢰성 확보방안

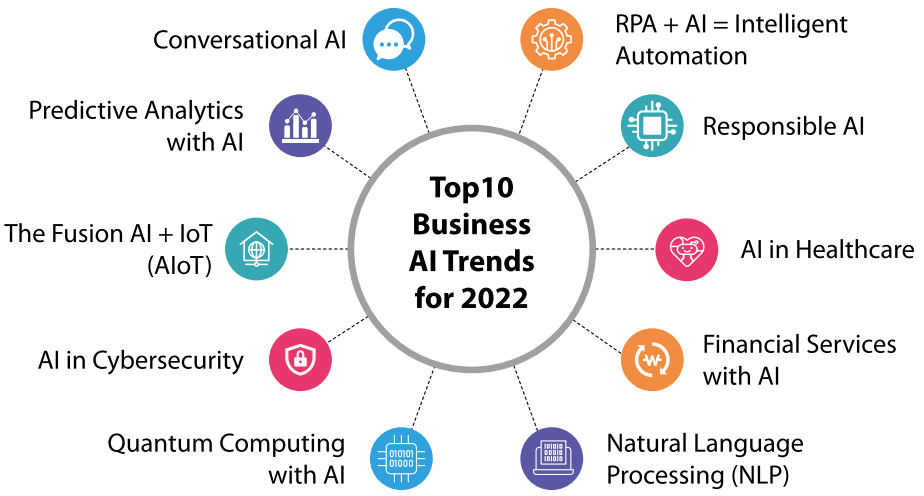
국경완 국방통합데이터센터 경영혁신실 실장

1. 머리말

인공지능(AI, Artificial Intelligence)은 상당한 역동성을 지닌 혁신적인 기술이다. 또한, 인지 기능에 다양한 파괴적 기술을 추가하여 엄청난 속도로 발전하고 있다. 인공지능이 새로운 기능을 통해 미래에 확실히 더 커질 것이라는 점을 부인할 사람은 아무도 없으며, 향후 비즈니스 환경을 크게 변화시킬 수 있는 기술임에 틀림없다. 이와 같이 인공지능은 AIoT(AI Internet of Things), RPA(Robotic Process Automation), 사이버보안 등 산업전반 모든 분야에서 큰 가능성과 유용성을 갖고 있기 때문에 이미 많은 미래 지향적인 기업 업무와 시스템에서 사용하고 있다.

인공지능은 과거에 우리가 했던 것처럼 의사결정 과정에서 관련성이 없는 것을 걸러내고, 이력서 더미에서 가장 적합한 후보자를 뽑고, 단순히 계산하는 것이 아니라 객관적으로 최고라고 계산한 것에 따라 우리를 안내하도록 가르칠 수 있는 무한한 잠재력을 가지고 있다. 2025년까지 세계 인공지능 시장 규모는 약 1,840억 달러(약 220조 원)으로 성장할 전망이다. 또한 마이크로소프트(MS), 아마존, 구글 등 글로벌 정보통신기술(ICT) 기업은 데이터 관리, 인공지능 모델 개발, 테스트, 서비스 운영, 모델 재학습 등을 지원하는 MLOps(Machine Learning Operations) 플랫폼을 개발해 클라우드 서비스로 제공하고 있다.

이와 같이 인공지능이 우리의 일상 생활과 기업 활동 등에 확대되면서 과거에는 일어나지 않았던 문제가 발생하고 있다. 바로 인공지능의 편향성(AI Bias) 문제가 대두되기 시작한 것이다. 2020년 12월 인공지능 전문 스타트업 스캐터랩 소속 핑퐁 팀(ScatterLab Pingpong Team)에서 개발한 페이스북 메신저 채팅 기반의 열린 주제 대화형 인공지능(Open-domain Conversational AI) 챗봇인 '이루다'는 출시 후 3주 만에 80만 명의 이용자가 몰릴 만큼 큰 인기와 함께 많은 논란을 가져왔다. 일부 이용자들이 20세 여성, 수동적인 대화 패턴 등 해당 인공지능 챗봇의 특징을 이용해 이루다에게 외설적인 대화를 하도록 유도하여 논란이 일었다. 이루다는 성적 단어를 금지어로 두

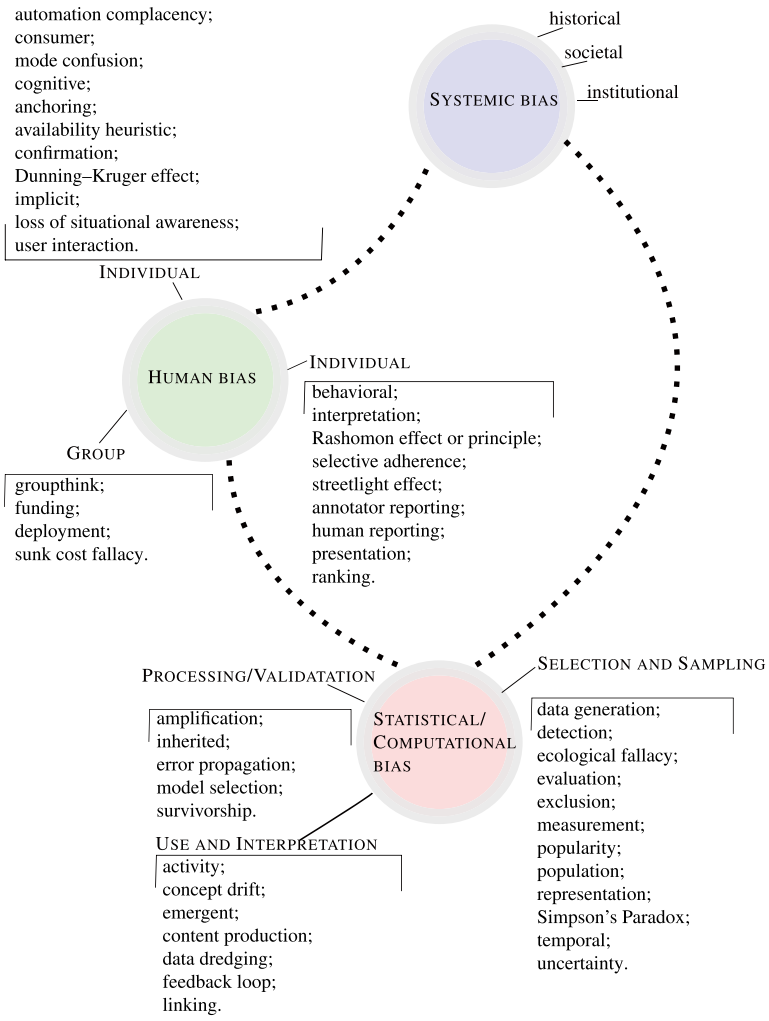


출처: appfirefit.com

[그림 1] 2022년 최근 인공지능 기술 활용분야 TOP 10

고 걸러내고 있었지만, 우회적인 표현으로 이루다와 성적 대화를 시도하고 비결을 공유하는 등 다양한 형태로 이용자들이 규칙을 피해가는 모습을 찾아볼 수 있었다.

이와 같이 인공지능의 편향은 다양한 형태로 사회 도처에 존재하며, 현실로 나타날 수 있다. 인공지능 편향은 어떠한 원인으로 인하여 인공지능 시스템이 특정 방향에 치우친 결과를 도출하는 것을 말한다. 인공지능 활용을 통해 편의성·생산성 향상 등 긍정적인 효과를 기대할 수 있지만 인공 지능 시스템에 나타난 특정 편향들은 개인과 조직, 사회에 대한 부정적 영향을 증폭하고 영속시킬 수 있다. 최근 이를 뒷받침하는 사례가 이미 나타나고 있으며, 인공 지능 시스템을 이용하는 모든 업무에 많은 부정적인 영향을 초래할 수 있다. 특히, 의료 인공지능의 편견은 중요한 문제로 대두될 수 있기 때문에 많은 관심을 가져야 한다. 본고에서는 이러한 인공지능 편향성 이슈와 신뢰성 확보방안에 대하여 살펴보고자 한다.



출처: Schwartz, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence", 2022.03.

[그림 2] 인공지능 편향의 범주

<표 1> 인공지능 신뢰성 확보를 위한 기술적 요구사항과 신뢰성 요건

요구사항	투명성
인간의 편향 (Human Bias)	- 인공지능이 학습하는 데이터는 인간으로부터 기인하고, 이 때문에 원시 데이터 자체가 편향이 개입되기 쉬우며, 인공지능이 편향을 가지는 것은 불가피 - 인공지능의 편향 발견은 의사결정 과정에서 수정을 할 수 있어 장점으로 작용할 가능성이 있음
숨겨진 / 암묵적 편향 (Hidden / Implicit Bias)	- 인공지능에서 가장 발견하기 어려운 것이 숨겨진 편향으로, 절대 보거나 발견될 수 없는 의도하지 않은 편향을 의미 - 이러한 편향을 가진 인공지능은 완벽한 이력서와 면접을 받은 면접자를 이해할 수 없는 이유로 탈락시킬 수도 있음 - 성별, 인종, 장애, 섹슈얼리티 또는 계급에 근거한 편견을 인식하지 못하기 때문에 위험함
데이터 표본 편향 (Data Sampling Bias)	- 인공 지능 시스템을 훈련할 때 좋은 데이터가 필요한데, 때때로 시스템에 공급되는 데이터에 샘플링 편향이 있어 인공지능이 편향되게 됨 - Amazon은 남성 이력서에서 더 일반적으로 발견되는 '실행됨(executed)' 또는 '캡처됨(captured)'과 같은 단어를 기반으로 지원자를 선호한다는 사실을 알게 된 후 채용 알고리즘 사용을 중단함
롱테일 편향 (Long-tail Bias)	- 훈련 데이터에서 특정 범주가 누락될 때 발생 - 인공지능이 얼굴 인식을 하고 있는데 주근깨가 많은 사람을 만났다고 가정할 때, 인공지능은 그 이미지로 무엇을 해야 할지 모를 수도 있음. 심지어 그들은 검은색, 흰색 또는 갈색으로 분류되거나 심지어 인간이 아닌 것으로 분류될 수도 있음
고의적 편향 (Intentional Bias)	- 가장 위험한 편향으로 해킹 공격을 통해 인공지능이 의도적으로 편향성이 부여될 수 있으며, 이러한 의도하지 않게 생긴 편향들은 발견하기 어렵게 숨겨지기에 더욱 위험함 - 공급망을 최적화하기 위해 특정 데이터베이스를 사용하여 인공지능을 훈련하고 있다는 사실을 이해하면 인공지능을 훈련하는 데 사용되는 데이터를 수정할 수 있으므로 잘못된 정보를 학습하게 되며, 이를 활용하여 공격자의 의도대로 정보 가공이 가능해짐

※출처 : 처 : Forbes, "How AI Can Go Terribly Wrong: 5 Biases That Create Failure," 2020. 11. 저자 재구성

2. 인공지능 편향성 이슈와 신뢰성 확보방안

2.1 인공지능 편향성 이슈

인간은 원하든 원하지 않든 다양한 비논리적인 이유로 다른 인간에 대해 편향되어 있다. 이것은 인간이 소수 인종, 종교, 성별 또는 국적에 편향되어 있는 경우 의식적으로 발생할 수 있다. 예를 들어 UN 보고서에 따르면 전 세계 남성과 여성의 최소 90%가 여성에 대해 일종의 편견을 갖고 있다고 밝혔다. 이 편향은 태어날 때부터 사회, 가족 및 사회적 조건의 결과로 이 편향은 태어날 때부터 접하는 사회, 가족 및 사회적 여건으로 인해 무의식적인 차원에 자리잡을 수도 있다. 무의식적으로 발생할 수도 있다. 이유가 무엇이든 편견은 인간에게 존재하며 인간이 만든 인공 지능 시스템도 이는 마찬가지다.

세계 최대 전자상거래 기업 아마존이 2014년부터 인공지능 채용 시스템을 개발해오다 알고리즘에서 여성 차별적 인식이 드러나자 폐기한 것으로 전해졌다. 공정성을 위해 점점 많은 기업들이 인공지능 채용시스템 도입을 준비하고 있지만 인공지능 알고리즘마저 편향적인 모습을 보인 것이다. 이 채용시스템은 500대의 컴퓨터가 구직 희망자의 지원서를 약 5만 개 키워드로 분석해 1개에서 5개까지의 별점을 매기는 프로그램이다. 그러나 개발이 1년쯤 진행되었을 무렵 아마존 자체 인공지능 채용시스템이 여성 지원자를 선호하지 않는다는 사실이 드러났다. 인공지능이 10년간의 아마존 지원자 데이터를 분석한 결과 남성 지원자가 압도적으로 많았기 때문이다.

인공지능 내에서 체계적이고 인간적인 편견이 어떻게 존재하는지 정의하고 설명함으로써 우리는 편견을 분석하고 관리하며 완화하기 위한 새로운 접근 방식을 구축하고 이러한 편견

이 서로 상호 작용하는 방식을 이해할 수 있다. 편향은 시스템적(Systemic) 편향, 통계·계산적(Statistical and Computational) 편향, 인적 편향(Human) 등 세 가지 범주로 식별할 수 있다. [그림 2]는 인공지능의 편향을 인공지능 애플리케이션 설계, 개발, 배포, 평가, 사용, 감사 등의 과정에서 고려해야 할 주요 위험 및 취약성으로 범주화해 보여 준다.

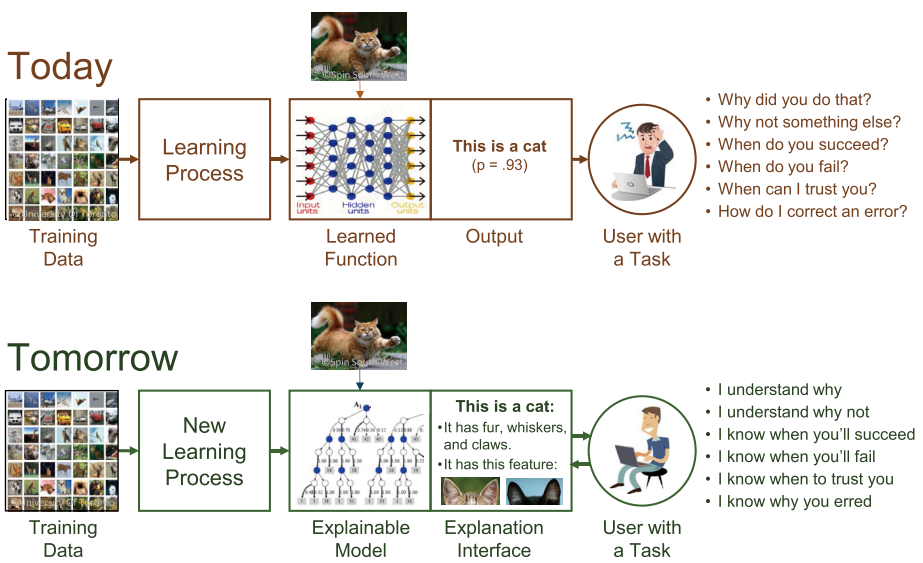
포브스(Forbes)는 인공지능 활성화를 저해하는 편향(Bias) 5가지를 <표 1>과 같이 소개했다. 인공지능 편향은 기계 학습 알고리즘 출력의 이상 현상(anomaly)에 기인하며, 이는 알고리즘 개발 과정에서 만들어진 편견이나 훈련 데이터의 편견 때문일 수 있다고 보았다.

2.2 인공지능 신뢰성 확보방안

2.2.1 양질의 데이터 확보

인공지능의 신뢰성을 확보하기 위한 한 가지 방법은 인공 지능 시스템을 훈련시키기 전에 편향이 제거되도록 데이터를 전처리(Pre-processing)하는 것이다. 편향되지 않은 데이터로 학습시켜 편향되지 않은 인공 지능 시스템을 만드는 방법이다. 또 다른 방법은 인공 지능 시스템이 데이터를 학습한 후 후처리(Post-processing)하는 것이다. 이는 인공 지능 시스템이 미리 결정할 수 있는 임의의 공정성 상수를 충족하도록 일부 예측을 변경하는 것을 의미한다. 그러나 이 두 가지 방법 모두 다음에 설명할 쉽게 설명가능한 인공지능(XAI) 알고리즘 개발에 포함된다. 이것은 인공지능 알고리즘이 일반적으로 블랙박스이고 어떻게 결론에 도달했는지 이해하기 매우 어렵기 때문에 필요하다. 인공지능 알고리즘에서 편향이 어디에 있는지 이해하기도 어려울 수 있다. 그러나 인공지능 알고리즘이 쉽게 설명될 수 있다면 편향을 찾아 제거할 수 있다.

최근 IBM Research는 인공지능 편견을 제거



출처: David Gunnning, Explainable Artificial Intelligence, DARPA

[그림 3] XAI 개념

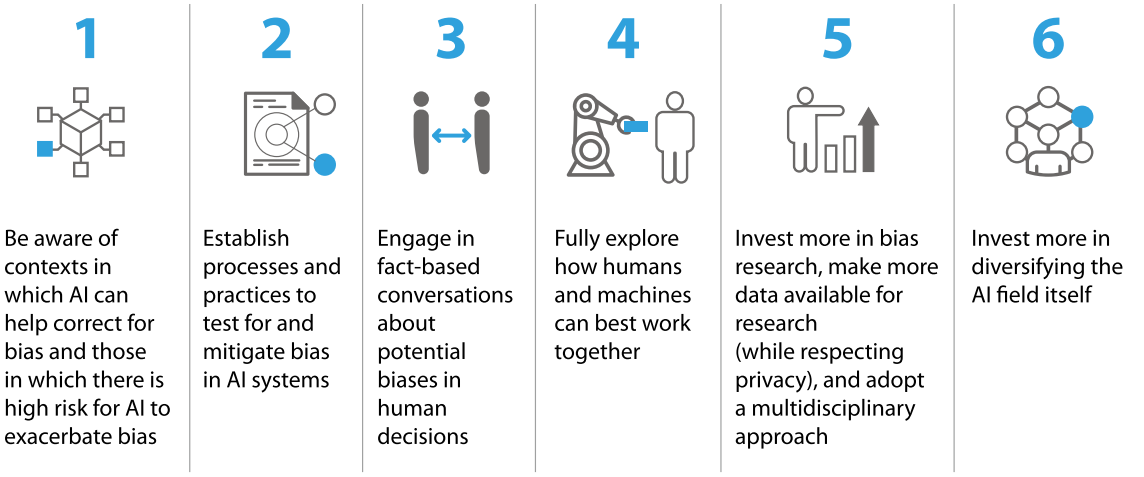
하기 위해 노력하고 있다. 그들은 5년 이내에 인공지능 사용이 증가함에 따라 알고리즘에서 인공지능의 편향의 수가 증가할 것이라고 예측하고 있다. 그러나 그들은 이러한 편견을 통제하고 편견이 없는 인공 지능 시스템을 만들기 위한 새로운 솔루션을 개발하고 있다. MIT-IBM Watson AI Lab은 최근 발전된 컴퓨팅 인지 모델링 및 인공지능을 사용하여 가치와 윤리적 의사 결정을 기계에 통합하고 이를 의사 결정에 적용하는 방법을 연구 중이다. IBM 과학자들은 인공 지능 시스템의 공정성을 판단하는 데 사용할 수 있는 독립적인 편향 평가 시스템도 만들어서 테스트 중이다.

2.2.2 설명가능한 인공지능(XAI, eXplainable Artificial Intelligence)

현재 인공지능 알고리즘에 의해 생성된 결과는 사실상 블랙박스이다. 일반적으로 결과는 예측, 추천 또는 결정의 형태로 알고리즘에 의해 생성된 다음 비판적 분석 없이 실생활에 적용된다. 문제가 되기 시작한 것은 이 알고리즘이 정확히 어떻게 결과에 도달하는지, 책임 및 윤리 개념에 해당하는 편향 또는 매개변수 등에 대한 규칙이 무엇이었는지 아무도 모른다는 사실이다. XAI는 인공지능에 좋은 편향이 반영됐는지 나쁜 편향이 적용됐는지 구분하는 데도 도움을 준다. 어떤 요소를 더 중요하게 평가하는지도 알려준다. 또한 인공지능 훈련용 데이터에서 편향이 생기는지, 혹은 인공지능이 라벨마다 지정한 가중치의 차이에서 생기는지를 이해하는 데도 유용하다. 다만 XAI가 편향성 자체를 직접 감지할 순 없다는 단점이 존재하기도 한다.

2.2.3 인공지능 및 기계학습 알고리즘의 편향 수정

인공지능 학습데이터에 활용되는 데이터 세트가 완전하다면 인공지능 편향은 인류의 편견으로 인해 발생할 수 있음을 인정하고 데이터 세트에서 그러한 편견을 제거하는 데 집중해야 하지만 실제 쉬운 작업은 아니다. 가장 간단한 방법은 데이터에서 보호된 클래스(예: 성별 또는 인종)를 제거하고 알고리즘을 편향되게 만드는 레이블을 삭제하는 것이다. 그러나 제거된 레이블



출처: McKinsey, Tackling bias in artificial intelligence and in humans, 2019.6.

[그림 4] 인공지능 편향성 다루기

은 모델의 이해에 영향을 미치고 결과의 정확도가 나빠질 수 있으므로 이 접근 방식은 작동하지 않을 수 있다.

인공지능의 편향성을 제거하는 정확하고 빠른 방법은 없지만 맥켄지(Mckinsey)에서는 인공지능의 편향성을 최소화하기 위해 [그림 4]와 같이 모범 사례를 강조하는 높은 수준의 권장 사항을 제공해 주고 있다. 주요 단계로는 ①인공지능을 배포할 때 편향된 시스템의 이전 예나 편향된 데이터가 있는 영역과 같이 잠재적으로 불공정한 편향이 발생하기 쉬운 영역을 예상, ②인공 지능 시스템의 편견을 테스트하고 완화하기 위한 프로세스와 관행을 수립, ③인간의 결정에 잠재적인 편향성이 있다는 사실에 기반한 대화에 참여, ④인간과 기계가 함께 가장 잘 협업할 수 있는 방법을 완전히 탐구, ⑤편향성 연구에 더 많이 투자하고 연구에 더 많은 데이터를 제공(프라이버시를 존중)하여 다양한 접근 방식을 채택, ⑥ 인공지능 분야 자체를 다각화하는 데 더 많은 투자의 필요를 언급하고 있다.

3. 맺음말

모든 인공지능 모델에는 '편향성'이 존재한다. 훈련용 데이터에 편향이나 편견이 포함될 수 있기 때문이다. 알고리즘도 의도적이든 우연이든 편향적 설계가 가능하다. 그러나 이러한 인공지능 편향성이 무조건 나쁜 건 아니다. 이 편향성을 활용해 좀 더 정확한 예측을 얻을 수 있기 때문이다. 인공지능은 광고를 집중해야 할 제품의 결정, 소비자 인지, 우수 면접자 선발이나, 특정 신용 상품에 대한 자격 및 기타 여러 결정을 내리는 데 많은 도움이 된다. 그러나 예측 및 의사결정에 인공지능을 활용하면 인간의 주관성을 줄일 수 있지만 편견을 포함하여 인구의 특정 하위 집합에 대해 부정확하거나 차별적인 예측 및 출력을 초래할 수 있는 단점이 있다.

예를 들어, 마케터는 인공지능에 의존하여 회사의 제품 및 서비스에 대한 최고의 잠재 고객을 목표로 삼을 수 있지만, 인공지능 알고리즘에서 의도하지 않은 편견을 제거하기 위한 조치도 취해야 한다. 근본적인 편견으로 인해 마케팅 메시지가 좋은 잠재 고객에게 전달되지 않을 수도 있다. 따라서 이러한 인공지능 알고리즘을 생산하는 기술 산업은 알고리즘을 시장에 출시하기 전에 편견이 없는지 확인해야 한다.

기업은 인공지능 편향에 대한 연구를 장려함으로써 편향의 제거 가능성을 높일 수 있다. 인공지능의 편향을 제거하기 위한 여러 가지 접근 방식이 있으나, 어느 것도 완벽하다고 볼 수 없다. 여기에는 상대적으로 편향이 없도록 애플리케이션을 공식화하는 접근 방식, 상대적으로 편향되지 않은 방식으로 데이터를 수집하는 접근 방식, 편향을 최소화하기 위한 수학적 알고리즘 설계에 이르기까지 다양하다.

인공지능 기술은 삶의 모든 측면에서 더 큰 통합을 향해 거침없이 움직이고 있다. 이런 일이 발생하면 복잡하고 거대한 시스템을 통해 편향이 발생할 가능성이 높지만 역설적으로 식별 및 예방이 쉽지만은 않다. 인공지능 편향 완화를 위한 지침 마련에 앞서 국내의 정책적, 사회적 상황을 반영한 인공지능 편향 범주화가 선행되어야 하며, 이를 바탕으로 체계적인 대응책을 마련하는 것이 중요하다.

참고문헌

[1] 국경완, '인공지능 편향성 이슈와 신뢰성 확보 방안,' 정보통신기획평가원(IITP) 주간기술동향, 2020호, 2021.10.26.
[2] 변순용, '데이터 윤리에서 인공지능 편향성 문제에 대한 연구,' 한국윤리학회(윤리학), 윤리연구 제1권 제128호, pp.143-158.
[3] 송은지, 봉기완, '인공지능의 윤리적 문제와 해결방안 모색,' 주간기술동향(1966호), 2020. 9. 30, p.16.
[4] 양기문, '미국 국립표준기술연구소 인공지능 편향식별 및 관리기준 마련', 한국전자통신연구원, 2022.04.02.
[5] Schwartz, R. et al, 'A Proposal for Identifying and Managing Bias in Artificial Intelligence', 2021.01.06
[6] Schwartz, R. et al, 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', 2022.03.05.
[7] M. Thomas, 'The Future of AI: How Artificial Intelligence Will Change the World, builtin.com,' 2021. 7.
[8] harkiran78, 'What is Artificial Intelligence Bias and How to Remove it?', geeksforgeeks.org, 2021.01.13.
[9] Praveen Singh, 'Top Artificial Intelligence (AI) Trends for 2022', appfirefit.com, 2021.11.24,
[10] Agbolade Omowole, 'artificial intelligence bios remove eliminate', weforum.org, 2021.07.19.
[11] 동아닷컴, '인공지능 신뢰성 높이는 '설명가능 인공지능(XAI)'의 시대', 2021.02.22
[12] AI타임즈, '포브스가 꼽은 인공지능을 망칠 수 있는 5가지 편향,' 2020.11.02.
[13] 전자신문, '인공지능 편향성 해결의 실마리 '데이터 품질'', 2021.03.17.
[14] ZDNet Korea, 'AI 알고리즘 규제' 선언 FTC, 구글·페북도 손볼까', 2021.04.21.
[15] IT Daily, '인공지능 기술의 윤리 문제에 대처하는 방법', 2021.05.01.