# 인공지능 윤리원칙 분석 보고서
## : 하버드 법대 버크만 센터의 'Principled Artificial Intelligence'을 중심으로

황 선 영*

## 1. 개요

지난 5여 년 동안 우리 일상생활에 인공지능(Artificial Intelligence, 이하 AI) 기술이 더 가깝게 자리잡으며, AI 기술과 연관된 여러 사건과 사고가 뉴스를 통해 보도되었다. 2016년에 발생한 구글의 자율주행차는 교통사고는 책임과 안전에 대한 논란을 일으켰고, 2018년에 우버의 자율주행차량에 의한 사망사고는 AI 기술에 대한 큰 우려와 비난을 불러일으켰다[1]. 또한, 집과 같은 사적 공간에서 사용되는 AI 스피커의 음성인식정보가 기업들에 의해 녹음 및 분석하고 있다는 사실이 밝혀져, AI 제품이나 서비스에 의해 수집되는 개인정보의 사생활 침해가 논란이 되었다.

위와 같은 문제에 대한 인식을 바탕으로 각국의 정부, IT 기업 및 국제기구는 지난 몇 년 동안 AI 기술 개발과 사용에 대한 윤리원칙을 발표하기 시작했다. 발표된 윤리원칙은 시민사회기구가 발표한 윤리 선언문, 정부 기관이 발행한 국가 인공지능 전략의 윤리적 권고사항, 그리고 민간기업이 당사의 제품이나 서비스를 개발할 때 추구하는 원칙 등의 다양한 형태로 공개되었다.

* 디지털경제사회연구본부 연구원, 043-531-4144, sunyoung@kisdi.re.kr

1) 상세한 내용은 한겨레(2018.03.20) "우버 자율주행차 첫 보행자 사망사고…안전성 논란 증폭" 참조

이러한 배경에서 하버드 대학교의 '인터넷과 사회를 위한 버크만 클라인 센터(The Berkman Klein Center for Internet & Society at Harvard University, 이하 하버드 법대 버크만 센터)'는 그동안 발표된 AI 윤리원칙을 분석하여 2020년 1월에 보고서로 그 결과물을 발표하였다. 본 보고서는 세계의 다양한 조직이 공개한 AI 윤리문서에서 공통적으로 다뤄진 원칙과 주제를 보여줌으로써 전세계적인 AI 윤리원칙 동향을 요약해 준다. 이에 본고에서는 하버드 법학대의 버크만 센터에서 발표한 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI(원칙에 입각한 AI, 2020.01)' 보고서의 주요 내용을 소개하도록 하겠다.

## 2. 원칙에 입각한 AI 보고서의 주요 내용

하버드 법대 버크만 센터의 연구진은 2016년부터 2019년 사이에 AI 윤리와 관련되어 발표된 공식문서를 다양한 방식으로 찾아, 총 36개의 문서를 최종적으로 선별하여 분석하였다. 선정된 AI 윤리문서는 다양한 내용, 지역, 시기 등을 대변할 수 있는 영향력 있는 문서였으며 중남미, 아시아, 중동, 북아메리카, 그리고 유럽 등에서 발간되었다. 또한, 경제협력개발기구(OECD), 유럽위원회(European Commission), 텐센트(Tencent) 등의 다양한 기관에 의해 발표된 문서로 구성되었다.

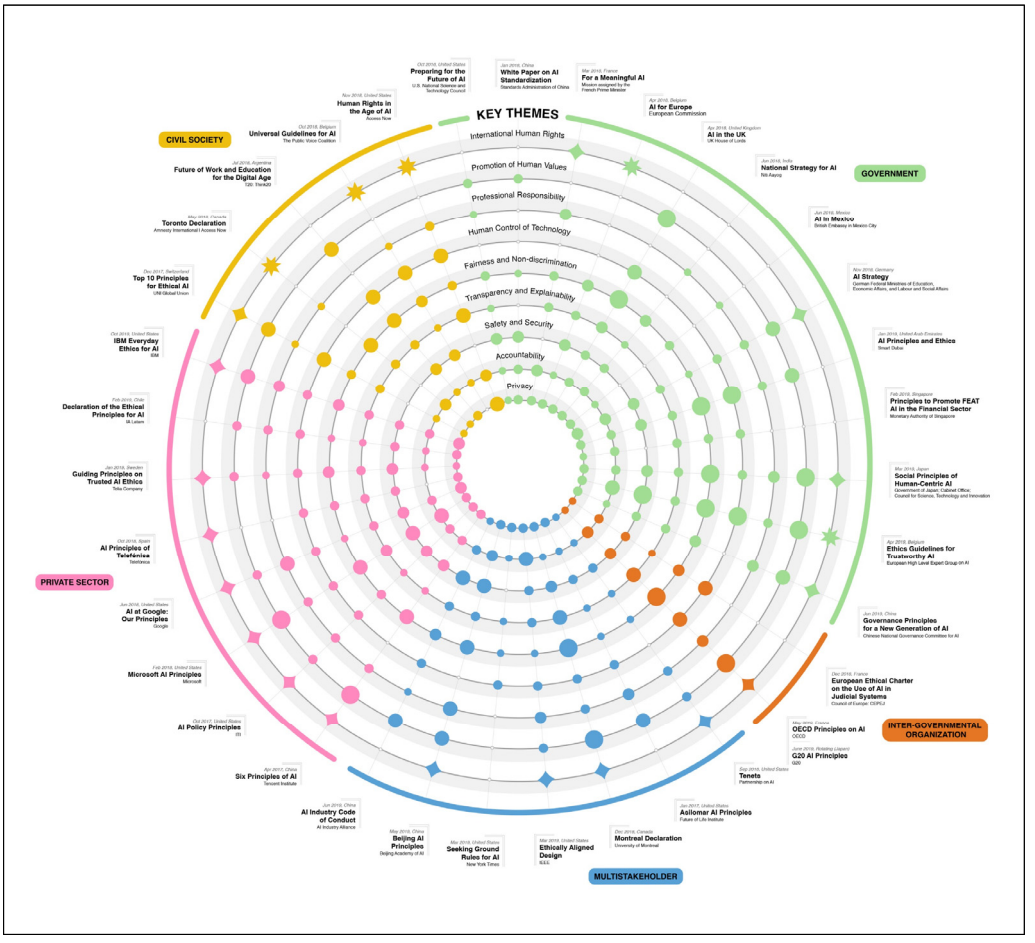
**〈표 1〉 기관 유형별 분석을 위해 선정된 문서의 개수**

| 기관 유형 | | 윤리문서 개수 | 예시 |
|---|---|---|---|
| 1 | 정부 간 기구 | 3 | 경제협력개발기구 (OECD) / 주요 20개국 (G20) |
| 2 | 정부 기관 | 13 | 미국 백악관 / 유럽위원회 |
| 3 | 민간 기관 | 8 | 구글 / 텐센트 / IBM |
| 4 | 시민사회 단체 | 5 | 국제사무직노조연합 / G20 연구·자문 네트워크 'T20' |
| 5 | 기타 이해관계자 | 7 | 뉴욕타임즈 / 미국 전기전자학회 'IEEE' |
| 모든 기관 | | 총 36개 문서 | |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

버크만 센터의 연구진은 선별된 AI 윤리문서의 내용을 48개의 원칙으로 분류하였고, 이를 8개의 주요 주제로 범주화하였다. 또한, 각 윤리문서가 언급하는 주제를 다음과 같이 바퀴모양의 그림으로 표현했다.

**[그림 1] 바퀴모양으로 시각화된 38개 윤리문서**




자료: Berkman Klein Center(2020.01)

그림 속에, 선별된 AI 윤리문서는 각각 한 개의 바퀴살로 표현되었다. 따라서 각 윤리문서를 대표하는 38개의 바퀴살을 볼 수 있다. 예시로, 3시경에 위치한 “유럽 인공지능 전략(AI for Europe)”을 살펴보면, 이 문서가 유럽위원회(European Commission)에 의해 2018년에 발표된 것을 알 수 있고, 바퀴살에 사용된 녹색을 통해 해당 문서가 정부 기관(Government) 문서라는 것을 알 수 있다.

더하여, “유럽 인공지능 전략” 바퀴살을 자세히 보면, 연구진이 범주화한 8가지 주요 주제 중 “유럽 인공지능 전략” 문서가 어떤 주제를 언급하고 있는지 알 수 있다. 해당 바퀴살에는 6개의 원형 모양이 곳곳에 자리잡고 있어[2] 8가지 주요 주제 중 해당 문서가 6가지에 대한 언급하고 있다는 것을 볼 수 있다.

버크만 센터는 AI 윤리원칙을 ① 개인정보 보호, ② 책무성, ③ 안전과 보안, ④ 투명성과 설명가능성, ⑤ 공정성과 무(無)차별성, ⑥ 기술 통제권, ⑦ 전문가 및 이해관계자 책임, ⑧ 인적가치 증진이란 8개의 주제로 범주화하였다. 각 주제별 주요 내용은 다음과 같다.

## (1) 개인정보 보호 (Privacy) AI 윤리원칙

AI 기술의 발전은 기존의 개인정보 수집, 보관, 그리고 처리 활동에 많은 영향을 미치며, 사생활 침해와 개인정보 보호에 대한 논의가 AI 윤리원칙의 핵심적 주제로 자리 잡았다. 연구진은 이러한 내용을 다음과 같은 8개의 원칙으로 분류하였다. 개인정보 보호와 관련된 첫 번째 원칙은 ‘동의’의 원칙이며, 이는 사람들의 개인정보를 수집하거나 활용할 때 이러한 사항을 안내하는 고지 및 동의(Notice-and-consent) 등의 절차가 필요하다는 내용을 담았다. 이외에도 사람들이 자신의 정보를 통제, 정정, 또는 삭제할 수 있어야 한다는 내용을 ‘정보사용통제’, ‘정보처리 제한능력’, ‘정보정정권’, 그리고 ‘정보삭제권’ 원칙을 통해 설명하였다. 다음으로, 개인정보 보호를 위한 기술 또는 정책이 필요하다는 내용을 ‘설계에 의한 개인정보 보호’ 원칙과 ‘개인정보 보호법’ 원칙으로 소개하였다. 마지막

2) 바퀴살의 내측부터 순서대로 △ Privacy, △ Accountability, △ Safety and Security, △ Transparency and Explainability, 그리고 △ Fairness and Non-discrimination에 원 모양이 표시되어 있다.

으로, 기타적 맥락에서 사생활 보호와 관련된 설명이 있을 때 '개인정보 보호 (기타/일반)'로 이를 분류하였다.

**〈표 2〉 주요 내용: 개인정보 보호와 관련된 AI 윤리원칙**

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 동의<br>(Consent) | 개인정보 활용 시, 이 사실을 안내·통지하고 동의를 구한 후 사용해야 함 |
| 2 | 정보사용통제<br>(Control over Use of Data) | 자신의 정보가 어떠한 이유와 방식으로 사용되는지 결정할 수 있어야 함 |
| 3 | 정보처리 제한능력<br>(Ability to Restrict Processing) | 인공지능과 관련된 개인정보 처리를 제한·통제할 수 있어야 함 |
| 4 | 정보정정권<br>(Right to Rectification) | 데이터컨트롤러 (Data Controller)[3]가 보유한 정보가 부정확하거나 불완전한 경우, 사람들이 이를 수정할 권리가 있어야 함 |
| 5 | 정보삭제권<br>(Right to Erasure) | 자신의 개인정보를 삭제할 수 있는 법적 강제력이 있는 권리가 있어야 함 |
| 6 | 설계에 의한 개인정보 보호<br>(Privacy by Design) | AI 개발자와 운영자가 AI 시스템 구축과정 및 정보(사용)라이프사이클 전체를 고려하여 개인정보를 보호할 의무가 있음 (이를 '설계에 의한 데이터보호'라고도 함) |
| 7 | 개인정보 보호법<br>(Recommends Data Protection Laws) | AI 기술발전으로 인해 개인정보 보호를 위한 새로운 정부규제가 필요함 |
| 8 | 개인정보 보호 (기타/일반)<br>(Privacy – Other/General) | 인권·권리 관련 논의 중 예시로 개인정보 보호가 논의될 때 사용된 분류를 위해 원칙. 인권·권리 문제 외에도, 국제적 경쟁력과 기업적 의무 등과 같은 이유로 개인정보 보호를 옹호하는 내용 등이 담김 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

## (2) 책무성 (Accountability) AI 윤리원칙

책무성은 현재 또는 미래에 다가올 자동화된 인공지능 시대에 인간이 아닌 기계의 의사결정에 대한 책임을 누가 어떻게 짊어질지에 대한 염려를 반영한 주제다. 10개의 원칙으로 구성된 '책무성' 주제는 △ AI가 설계·개발되는 단계와 관련된 원칙, △ 현장에 AI를 배포·설치한 후 관련된 원칙, 그리고 △ AI 기술 관련 피해가 발생했을 때와 관련된 원칙

3) 데이터 컨트롤러는 개인정보 처리의 목적과 수단을 결정하는 주체를 말한다.
〈출처: 한국인터넷진흥원(2018.05), "우리 기업을 위한 EU 일반 개인정보보호법 가이드북"〉

으로 나누어 볼 수 있다. 먼저, AI 시스템을 설계할 때 해당 시스템이 사회에 이로운 영향을 미치며 정상적으로 작동하기 위해선 시스템의 '검증가능성 및 복제가능성'을 보장하고, '영향평가'를 체계적으로 수행해야 하며, AI와 관련된 '환경에 대한 책임'을 고려해야 한다. 그 후, AI 시스템을 현장에 설치하고 배포한 후 시스템을 지속적이며 체계적으로 모니터링하고 개선하기 위해선 '평가·감사를 요구'하고, '관리감독기관을 창설' 해야 하며, '이의제기 가능성'을 보장해야 한다. 마지막으로, AI 시스템으로 인해 피해가 발생한 후 이에 대응하기 위해서 '배상·구제조치'를 마련하고, 관련 개인이나 조직이 이에 대한 '법적 책임'을 지도록 하며, '새로운 규제를 도입'하여 관련 조치와 절차를 명확히 해야 한다. 그리고 마지막으로, 연구진은 '책무성' 또는 '책무'라는 구체적 단어를 쓴 기타 내용은 '책무성'이란 원칙으로 분류하였다.

**〈표 3〉 주요 내용: 책무성과 관련된 AI 윤리원칙**

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 검증 및 복제 가능성 (Verifiability & Replicability) | AI 시스템의 정상적 작동을 위해 (1) 실험 시 같은 조건에 같은 결과를 산출하고, (2) 타당성 검증이 가능하도록 운영체계에 대한 충분한 정보가 제공되어야 함 |
| 2 | 영향평가 (Impact Assessments) | 인권영향평가 및 AI 기술의 부정적 영향을 예측, 방지 및 완화를 위한 평가가 있어야 함 |
| 3 | 환경에 대한 책임 (Environmental Responsibility) | AI는 미래에 널리 사용될 기술로 큰 환경적 영향이 있을 것이며, 해당 기술을 설계·구현하는 사람들은 이에 대한 책임을 짊어야 함 |
| 4 | 평가·감사 요구 (Evaluation & Auditing Requirement) | 감사를 실시할 수 있는 기술 구축뿐만 아니라 평가를 통해 얻은 지식을 다시 시스템에 적용하여 끊임없이 시스템을 개선해야 함 |
| 5 | 관리감독기관 창설 (Creation of a Monitoring Body) | AI 기술발전으로 인해 관련 기술표준과 모범사례를 발굴·감독하기 위한 새로운 기관 또는 체계가 필요함 |
| 6 | 이의제기 가능성 (Ability to Appeal) | AI 의사결정의 대상인 사람은 해당 결정에 대한 이의를 제기할 수 있어야 함 |
| 7 | 배상·구제조치 (Remedy for Automated Decisions) | 인공지능 기술이 더 널리 활용되며 실질적 피해를 야기할 수 있고 이에 대한 배상·구제조치를 준비해야 함 |
| 8 | 법적 책임 (Liability & Legal Responsibility) | 개인 또는 조직의 잘못으로 인해 발생한 AI 관련 피해에 대한 책임을 해당 개인·조직이 져야 함 |

| 원칙 | | 세부 내용 |
|---|---|---|
| 9 | 새로운 규제 도입<br>(Recommends Adoption of New Regulations) | AI 기술을 위한 새로운 규제체계가 필요함. 특히 관련 관계자의 책임의무을 명확히 하며 맥락과 분야에 맞는 맞춤형 규제가 필요함 |
| 10 | 책무성<br>(Accountability Per Se) | "책무성 (Accountability)" 또는 "책임이 있는 (Accountable)" 이란 용어를 사용한 내용을 분류하기 위한 원칙. 다양한 관계자의 책무를 구체화할 "책무성 프레임워크" 개발이 필요하다는 내용 등을 담음 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

## (3) 안전과 보안(Safety and Security) AI 윤리원칙

초기에 도입된 AI 기술과 관련된 여러 안전사고로 인해 수많은 AI 윤리문서에 안전과 보안에 대한 논의가 많이 있었고, 연구진은 이를 '안전과 보안'이라는 주제로 분류하였다. 주제의 이름에서 이미 나타나는 '안전성'과 '보안성' 원칙이 먼저 소개되었다. 그리고 AI 시스템의 개발 단계에서부터 보안성을 보장해야 한다는 '설계에 의한 보안' 원칙이 소개되었고, AI 시스템의 안전성을 위해선 해당 시스템이 예측 가능한 방식으로 작동해야 한다는 '예측가능성' 원칙이 설명되었다.

**〈표 4〉 주요 내용: 안전과 보안과 관련된 AI 윤리원칙**

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 안전성<br>(Safety) | AI 시스템은 생명체와 환경을 해치지 않고 의도한 데로 작동하는 함 |
| 2 | 보안성<br>(Security) | AI 시스템은 외부 공격·위협에 대처할 수 있는 능력이 있어야 함 |
| 3 | 설계에 의한 보안<br>(Security by Design) | AI 시스템은 개발 단계에서부터 보안성을 보장해야 함 |
| 4 | 예측가능성<br>(Predictability) | AI 시스템은 입력된 정보(Input)에 따라 일관된 결과가 나와야 함 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

## (4) 투명성과 설명가능성(Transparency and Explainability) AI 윤리원칙

투명성과 설명가능성은 AI 기술을 관리하는 구조적 체계인 AI 거버넌스와 밀접히 연관된 주제이다. 다른 기술에 비해 더 넓은 범위와 형식의 데이터를 기반으로 작동하는 AI 기술은 다양한 이해관계자에게 큰 영향을 미칠 수 있는 매우 복잡한 기술이다. 따라서, 상황과 때에 따라, AI 시스템과 관련된 데이터나 정보를 투명한 방식으로 설명하고 공개할 수 없는 상황이 종종 발생한다. '투명성과 설명가능성'이란 주제는 이러한 상황에 발생할 수 있는 문제에 대처하기 위한 8가지 원칙으로 구성되어 있다. 적절한 투명성과 설명가능성을 목표로 삼아야 한다는 '투명성'과 '설명가능성' 원칙이 가장 먼저 논의된다. 그다음, 다양한 이해관계자가 공동으로 협력하여 AI 기술을 발전시키자는 '오픈소스 데이터 및 알고리즘' 원칙과 정부의 투명한 인공지능 기술 조달을 요구하는 '공개조달' 원칙이 설명된다. 이외에도 AI에 대한 일반 시민들의 권리에 대한 3가지 원칙인 '정보에 대한 권리', 'AI 의사결정에 대한 알림', 그리고 '상호작용 중인 AI에 대한 알림' 원칙이 제시된다. 마지막으로, AI 시스템에 대한 체계적 정보 공개를 권장하는 '정기적인 보고' 원칙이 소개된다.

**〈표 5〉 주요 내용: 투명성과 설명가능성과 관련된 AI 윤리원칙**

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 투명성<br>(Transparency) | AI 시스템의 작동·운영에 대한 관리·감독이 가능하도록 시스템을 설계 및 실행해야 함 |
| 2 | 설명가능성<br>(Explainability) | 사람이 비슷한 의사결정을 할 때 요구되는 설명과 같이, AI 시스템이 무엇을 어떤 이유로 하고 있는지에 대한 명확하며 완전한 설명이 이해하기 쉬운 형태로 제공되어야 함 |
| 3 | 오픈소스 데이터 및 알고리즘<br>(Open Source Data and Algorithms) | 데이터·플랫폼 독점을 방지하며 공평하게 AI의 혜택을 나눌 수 있도록 다양한 관계자가 공동으로 알고리즘을 개발하고 개방된 연구협력을 할 수 있어야 함 |
| 4 | 공개조달<br>(Open Government Procurement) | 정부기관이 AI 시스템이나 구성요소를 조달하려고 할 때 공개적 조달 기준에 맞는 투명한 방식으로 해야 함 |

| 원칙 | | 세부 내용 |
|---|---|---|
| 5 | 정보에 대한 권리<br>(Right to Information) | 일상생활에 사용되는 AI 시스템과 이로 인해 영향받는 사람들이 관련 정보를 알 권리가 있으며 해당 조직은 이러한 정보를 제공할 의무가 있음 |
| 6 | AI 의사결정에 대한 알림<br>(Notification When AI Makes a Decision about an Individual) | AI가 사용될 때, 활용된 AI의 '대상자'가 이 사실을 알아야 함 |
| 7 | 상호작용 중인 AI에 대한 알림<br>(Notification when Interacting with AI) | 인간과 구분하기 힘든 인공지능이 나타날 것을 염두에 두며, 사람들이 인간이 아닌 인공지능과 상호작용할 때 해당 사실을 반드시 알려야 함 |
| 8 | 정기적인 보고<br>(Regular Reporting) | AI 시스템을 사용하는 기관이 이에 대한 주요 정보를 체계적으로 공개해야 함 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

### (5) 공정성과 무(無)차별성(Fairness and Non-discrimination) AI 윤리원칙

버크만 센터의 연구진이 분류한 8가지 AI 윤리주제 중 가장 많이 언급한 주제는 바로 '공정성과 무차별성'이란 주제이다. 특정 집단이나 개인에 편향되거나 차별적인 의사결정을 내리는 AI 알고리즘 편향성 문제는 심각한 사회문제를 야기할 수 있으며, 이를 방지하기 위해 6가지 원칙이 소개되었다. 첫 번째로, AI 시스템의 편견과 차별을 방지해야 한다는 '무차별성과 편견방지' 원칙이 제시되었다. 이러한 목표를 실제로 실현하기 위한 구체적 방식으로는 '대표성 있는 양질의 데이터'를 AI 시스템에 사용해야 한다는 내용이 있었다. AI 시스템이 추구해야 할 전반적 가치로는 '공정성'과 '평등성'도 언급되었다. 이외에도 AI 시스템의 혜택이 소외 집단을 포함한 다양한 사람들에게 정당히 분배되어야 한다는 '혜택의 포괄성' 원칙과 포괄적이며 다양한 배경을 가진 사람들을 기반으로 AI 시스템을 설계해야 한다는 '설계과정의 다양성·포괄성' 원칙이 포함되었다.

〈표 6〉 주요 내용: 공정성과 무(無)차별성과 관련된 AI 윤리원칙

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 무(無)차별성 및 편견방지 (Nondiscrimination & the Prevention of Bias) | 인공지능을 훈련하기 위해 사용되는 데이터, 관련 기술 설계 또는 배포 등으로 인해 발생하는 인공지능 편향성을 완화하여 차별을 방지해야 함 |
| 2 | 대표성 있는 양질의 데이터 (Representative & High Quality Data) | '쓰레기를 넣으면 쓰레기가 나온다'라는 말과 같이, 예측대상에 맞는 적절한 데이터를 AI 시스템에 사용해야 함 |
| 3 | 공정성 (Fairness) | AI 시스템은 사람들을 공정하고 편견 없이 대우해야 함 |
| 4 | 평등성 (Equality) | 사람들은 AI 기술의 발전과 관련하여 동등한 기회와 보호를 받을 자격이 있음 |
| 5 | 혜택의 포괄성 (Inclusiveness in Impact) | AI 기술의 혜택을 정당히 분배할 해야 함. 특히 과거에 소외된 집단도 혜택을 누릴 수 있도록 보장해야 할 것 |
| 6 | 설계과정의 다양성·포괄성 (Inclusiveness in Design) | AI 시스템 개발과정에 현재보다 더 다양한 사람들이 참여해야 함. 다양한 사람들로 구성된 AI 설계팀을 구성해야 하며, AI 시스템에 대한 사회구성원의 의견수렴을 진행하고 이를 반영해야 함 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

## (6) 기술통제권(Human Control of Technology) AI 윤리원칙

현재와는 동떨어진 이야기 같지만, 이미 일부 AI 윤리문서는 미래 기술발전으로 인해 인간이 AI 시스템에 대한 통제권을 잃을 수 있다는 가능성을 염두에 두고 있었다. 이러한 디스토피아적 미래를 방지하기 위해 3가지 원칙이 논의되었다. 첫 번째로 논의된 원칙은 AI의 자동화된 의사결정에 의해 부당한 결과가 발생하면 사람들이 이에 대한 재검토를 요청할 수 있다는 내용을 담은 원칙이다. 두 번째 원칙은 애초에 AI의 자동화된 결정에서 제외될 수 있는 선택권을 사람들에게 제공해야 한다는 원칙이며, 마지막 원칙은 전반적인 기술 통제권에 대한 논의를 분류한 원칙이므로 기술 통제권을 구현할 수 있는 다양한 방식과 이와 관련된 인간의 역할 등을 설명하였다.

〈표 7〉 주요 내용: 기술통제권과 관련된 AI 윤리원칙

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 자동화된 의사결정에 대한 재검토 (Human Review of Automated Decisions) | AI 시스템 의사결정의 대상자는 해당 결정에 대한 재검토를 요청할 수 있으며 이 검토는 기계가 아닌 사람이 해야 함 |
| 2 | 자동화된 결정에서 제외될 선택권 (Ability to Opt out of Automated Decisions) | AI 시스템이 사용되는 경우, 해당 시스템의 대상자에서 제외될 수 있는 선택권(옵트아웃할 기회)을 제공해야 함 |
| 3 | 기술 통제권 (기타/일반) (Human Control of Technology - Other/General) | 사람들이 AI 시스템의 의사결정과 행위에 개입할 수 있도록 시스템을 설계 및 배포해야 함 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

## (7) 전문가 및 이해관계자 책임(Professional Responsibility) AI 윤리원칙

'전문가 및 이해관계자 책임'이란 주제는 AI 기반 제품이나 서비스를 설계, 개발, 또는 배포하는 사람들을 대상 한 원칙을 분류한 주제이다. 이 주제는 개별 상황에 따라 특정 제도나 조직보다 직접 AI 시스템을 설계하거나 배포하는 개개인 관계자가 윤리적 문제에 더 크고 직접적인 영향을 미칠 수 있다는 인식을 반영한 주제이다. 이를 바탕으로 AI 전문가와 이해관계자에게 다음과 같은 각별한 책임감을 요구하는 원칙들을 '전문가 및 이해관계자 책임'이란 주제로 묶은 것이다. 가장 먼저, 전문가는 신뢰할 수 없는 정보의 확산을 방지하고 정확한 정보를 다루는 AI 시스템을 설계 및 활용해야 한다는 내용이 '정확성'의 원칙을 통해 제시되었다. 이러한 목표를 위해, '책임성 있는 설계'를 추구하고, AI 시스템의 '장기적 영향 고려'하며, '다중-이해당사자와의 협력'을 도모하고, '연구윤리 준수' 해야 한다는 원칙들이 논의되었다.

〈표 8〉 주요 내용: 전문가 및 이해관계자 책임과 관련된 AI 윤리원칙

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 정확성 (Accuracy) | AI 기술은 정보를 알맞은 부류·범주로 분류할 수 있어야 하며, 데이터와 모델을 기반으로 정확한 예측, 추천, 또는 의사결정을 할 수 있어야 함 |
| 2 | 책임성 있는 설계 (Responsible Design) | AI 시스템 설계에 종사하는 관계자들은 미래 인공지능 기술에 강한 영향력을 행사할 수 있는 독특한 위치에 있으며, 양심적이게 행동하며 맥락 및 사회적 가치를 섬세하게 고려해야 함 |
| 3 | 장기적인 영향 고려 (Consideration of Long Term Effects) | AI 시스템 설계 및 배포 단계에서 해당 시스템으로 인해 먼 미래에 발생할 수 있는 영향을 의도적으로 고려해야 함 |
| 4 | 다중-이해당사자와 협력 (Multi-stakeholder Collaboration) | AI 애플리케이션을 개발하고 관리할 때, 설계자와 사용자들이 관련 이해관계자 단체의 자문을 구할 것을 요구 또는 권장해야 함 |
| 5 | 연구윤리 (Scientific Integrity) | AI 시스템을 설계하고 상용화하는 종사자는 해당 분야에 이미 존재하는 가치와 규범을 따라야 함 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

## (8) 인적가치 증진(Promotion of Human Values) AI 윤리원칙

마지막 주제인 '인적가치 증진'은 AI 기술의 목적과 구현방식이 현재의 사회규범과 부합해야 한다는 인식을 바탕으로 한다. 다시 말해, AI 기술은 궁극적으로 인간의 번영과 공익에 기여하는 기술이어야 한다는 내용을 담은 주제이다. 첫 번째 원칙으로 소개된 '인적가치 및 인간의 번영' 원칙은 AI 기술이 인간의 존엄성을 존중하고 사람들의 인권과 자율성을 보호하며 개개인의 역량 발휘를 도모하는 방식으로 활용되어야 한다는 원칙이다. 이를 이어, 경제적·사회적 불평등을 심화하지 않는 방식으로 AI 기술의 접근성을 보장해야 한다는 '기술 접근성' 원칙이 설명되었다. 마지막 원칙인 '사회에 이롭게 사용' 원칙은 AI 기술이 수익성, 안정성 또는 합법성뿐만 아니라 사람들의 행복·안녕(Human Well-being) 및 공익을 기준으로 개발되고 활용되어야 한다는 내용을 담았다.

〈표 9〉 주요 내용: 인적가치 증진과 관련된 AI 윤리원칙

| 원칙 | | 세부 내용 |
|---|---|---|
| 1 | 인적가치 및 인간의 번영 (Human Values & Flourishing) | AI 개발과 활용 방식이 기존 사회규범, 문화적 신념, 그리고 인류의 번영을 염두에 두어야 함 |
| 2 | 기술 접근성 (Access to Technology) | 인공지능이 (사회적) 불평등을 심화할 수 있으므로 많은 사람이 AI 기술을 사용할 수 있도록 보장하여 혜택을 누릴 수 있어야 함 |
| 3 | 사회에 이롭게 사용 (Leveraged to Benefit Society) | AI 시스템은 공공의 이익에 기여하도록 사용되어야 함 |

자료: Berkman Klein Center(2020.01)를 참고하여 작성

# 3. 결어

현재 많은 연구가, 전문가, 개발자, 그리고 기업이 AI 시스템을 설계하고 개발하고 있으며, 일반 사람들도 다양한 기기와 서비스를 통해 알게 모르게 AI 기술을 사용하고 있다. 우리나라 정부도 2019년 12월에 발표한 '인공지능 국가전략'을 통해 AI 기술이 한국 경제의 활력 제고와 사회문제 해결에 기여할 수 있는 유력한 방안이라 설명하였다.[4] 하지만 과거 신기술과 같이 AI 기술에 대한 명암도 엇갈릴 것이며, AI 기술의 역기능과 부정적 영향에 대응하고 대처하기 위한 장치가 마련되어야 할 것이다.

하버드 법대 버크만 센터의 보고서는 이러한 장치를 마련하기 위한 다양한 이해관계자와 국가의 노력을 분석하고 분류한 결과물이라 이해할 수 있다. 아직 구체적 규제나 법체계가 없는 과도기적인 상황에서 미래를 대비하기 위해 AI 윤리에 대한 논의와 세계적 합의를 이끌 수 있는 기준을 설립하는 것인 매우 중요한 과제라 생각한다. 우리나라 정부도 앞으로 다가올 AI 시대를 뒷받침할 법제도를 정립하기 위해 AI 법제정비단을 발족하여 AI 시대의 기본원칙과 윤리 기준을 마련하려 노력하고 있다.[5] 이러한 상황을 종합적으로

4) 관계부처 합동(2019.12), "인공지능 국가전략"
5) 관계부처 합동(2019.12), "인공지능 국가전략"

고려해봤을 때, 버크만 센터의 'Principled Artificial Intelligence'는 세계적 AI 윤리원칙 동향의 요약서로써 우리나라 법제도 정립에 참고자료가 될 수 있으며, 앞으로 글로벌 수준의 AI 윤리규범을 확립하는데 도움일 될 것으로 예상된다. 물론 이를 위해서 한국 특유의 사회적·기술적 상황에 맞게 원칙을 적용하는 과정이 선행되어야 할 것이다.

## 〈참고문헌〉

관계부처 합동(2019.12), "인공지능 국가전략"

한겨레(2018.3.20) "우버 자율주행차 첫 보행자 사망사고…안전성 논란 증폭" http://www.hani.co.kr/arti/international/international_general/836834.html

한국인터넷진흥원(2018.05), "우리 기업을 위한 EU 일반 개인정보보호법 가이드북"

Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika(2020.01). "Principled Artificial Intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI." Berkman Klein Center Research Publication No. 2020-1.