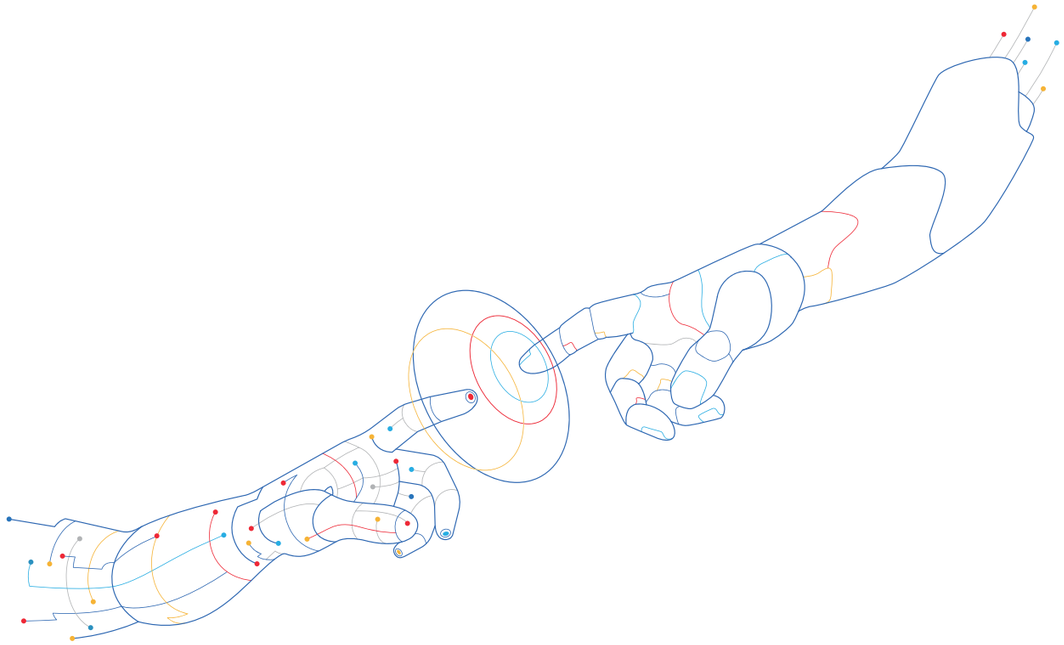
유네스코한국위원회-한국법제연구원 공동연구

# 인공지능(AI) 윤리와 법(I)
# AI 윤리의 쟁점과 거버넌스 연구

## 부록 | 유네스코 인공지능 윤리 권고

기획 | 유네스코한국위원회
이상욱, 이호영 지음



국제연합 교육과학문화기구 | 유네스코 한국위원회
한국법제연구원 KOREA LEGISLATION RESEARCH INSTITUTE

인공지능(AI) 윤리와 법(I) AI 윤리의 쟁점과 거버넌스 연구

국제연합 교육과학문화기구 | 유네스코 한국위원회
한국법제연구원 KOREA LEGISLATION RESEARCH INSTITUTE

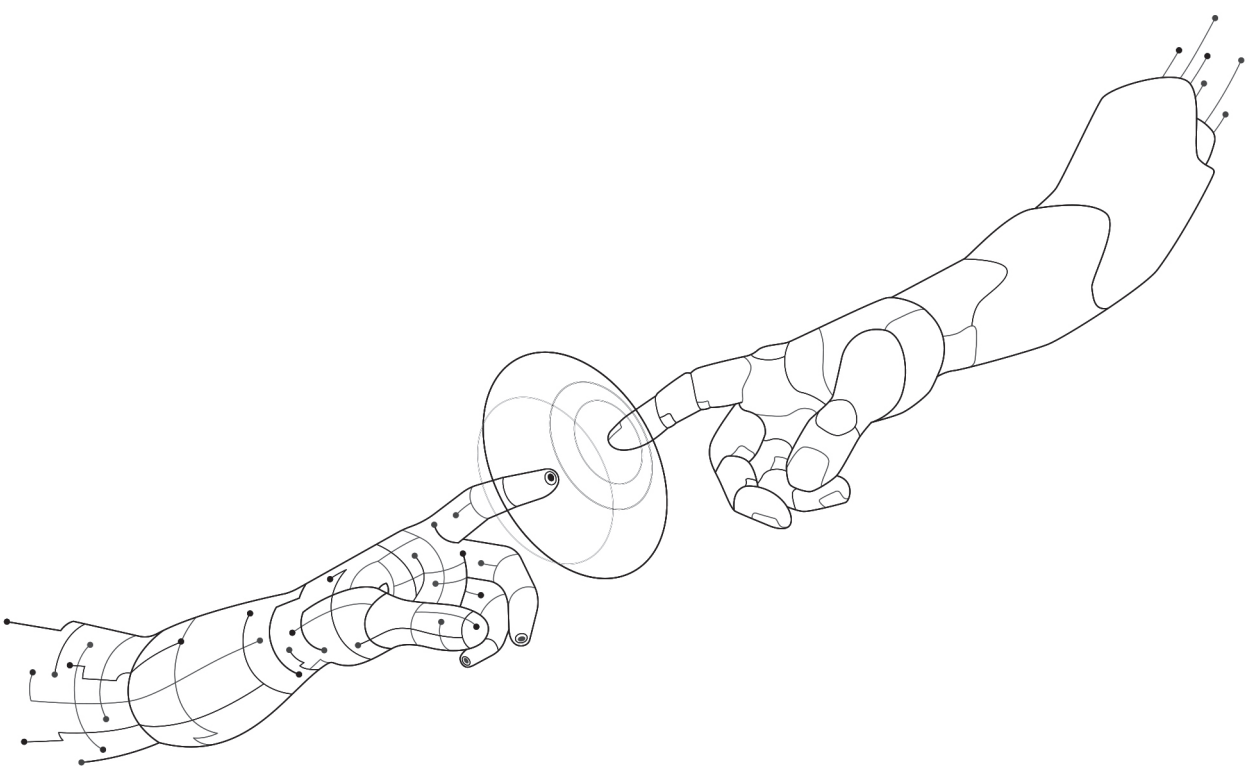
비매품/무료
03560
9 791190 615235
ISBN 979-11-90615-23-5

유네스코한국위원회-한국법제연구원 공동연구

# 인공지능(AI) 윤리와 법(I)
# AI 윤리의 쟁점과 거버넌스 연구

## 부록 | 유네스코 인공지능 윤리 권고

기획 | 유네스코한국위원회
이상욱, 이호영 지음



국제연합 교육과학문화기구 | 유네스코 한국위원회
한국법제연구원 KOREA LEGISLATION RESEARCH INSTITUTE

# AI 윤리의 쟁점과 거버넌스 연구

**기 획** 유네스코한국위원회
**지은이** 이상욱, 이호영

**펴낸곳** 유네스코한국위원회
**펴낸이** 한경구

**펴낸날** 2021년 12월 10일

유네스코한국위원회
서울 중구 명동길(유네스코길) 26
전화 (02) 6958-4164
sc.team@unesco.or.kr
www.unesco.or.kr

ISBN 979-11-90615-23-5



**일러두기**

- 동 출판물에 기재된 구체적인 내용과 방향은 유네스코한국위원회의 입장과 반드시 일치하지 않을 수도 있습니다.
- 이 책은 저작권법에 따라 보호받는 저작물이므로 무단전재와 무단복재를 금하며, 이 책 내용의 전부 또는 일부를 이용하고자 할 경우에는 유네스코한국위원회로 문의해주시기 바랍니다.

# 유네스코 헌장 전문

이 헌장의 당사국 정부는 그 국민을 대신하여 다음과 같이 선언한다.

전쟁은 인간의 마음속에서 생기는 것이므로 평화의 방벽을 세워야할 곳도 인간의 마음속이다.

인류 역사를 통해 상호간의 생활양식과 삶에 대한 무지는 사람들 사이에 의심과 불신을 가져온 공통적 원인이었으며 이러한 상호간의 차이점들이 너무도 자주 전쟁으로 이어져왔다.

이제 막 끝난 가공할 대 전쟁은 인간의 존엄, 평등, 상호존중이라는 민주주의 원리를 부정하고, 대신 무지와 편견을 통해 인간과 인종의 불평등주의를 확산시킴으로써 발생된 사건이었다.

문화의 광범한 보급과, 정의 · 자유 · 평화를 위한 인류 교육은 인간의 존엄성을 수호하기 위해 반드시 필요한 것이며, 또한 모든 국민이 상호 관심과 협력의 정신으로써 완수해야 할 신성한 의무이다.

오로지 정부 간 정치적이경제적 타협에 근거한 평화는 세계 모든 사람들의 일치되고 영속적이며 성실한 지지를 얻을 수 있는 평화가 아니다. 따라서 평화를 잃지 않기 위해서는 인류의 지적 · 도덕적 연대 위에 평화를 건설하지 않으면 안 된다.

이러한 이유에서 이 헌장의 당사국은 교육의 기회가 모든 사람에게 충분하고 평등하게 주어지고 객관적 진리가 구속받지 않고 탐구되며 사상과 지식이 자유로이 교환되어야 함을 확신하면서, 국민들 사이의 소통수단을 발전시키고 증가시키는 동시에, 서로를 이해하고 서로의 생활을 더욱 진실하고 더욱 완전하게 알기 위하여 이 소통수단을 사용할 것을 동의하고 결의한다.

그 결과 당사국은 국민들의 교육 · 과학 · 문화상의 관계를 통하여, 국제연합의 설립 목적이며 또한 그 헌장이 선언하고 있는 국제평화와 인류공동의 복리라는 목적을 촉진하기 위하여 여기에 국제연합교육과학문화기구를 창설한다.

# 발 간 사

AI는 음식 배달부터 내비게이션, 영화 예약까지 우리의 일상 속에 이미 깊숙이 퍼져 있습니다. 또한, 국제적인 기후변화 대응을 위한 방안을 연구하고, 새로운 백신을 개발하는 과정에서도 놀라운 성과를 보여주고 있습니다.

그러나 AI 기술은 전례 없는 새로운 도전을 불러일으키고 있기도 합니다. 성별 및 민족에 대한 편견의 확대재생산과 개인정보 유출, 대규모 감시의 위험 등은 AI로 인해 우리가 이전보다 더욱 우려하게 된 문제들입니다. 하지만 지금까지 이러한 문제에 대한 답을 제공할 보편적인 기준은 없었습니다.

2018년 유네스코는 인공지능(AI)의 윤리적 개발과 이용을 위한 국제적인 지침을 만들기 위한 야심찬 프로젝트를 시작했습니다. 그리고 지난 3년 여 간 전 세계 수백 여 전문가들의 헌신과 치열한 국제적인 토론을 거쳐 2021년 11월 25일 193개 유네스코 회원국의 만장일치로 'AI 윤리 권고'를 공식적으로 채택했습니다.

유네스코 AI 윤리 권고는 AI의 건전한 개발과 이용을 보장하기 위한 가치와 원칙을 담고 있습니다. AI 윤리 권고는 회원국 모두에게 권고를 적용할 책임을 부여하고 있으며, 유네스코 역시 오랫동안의 논의 끝에 지난 11월 24일에 유네스코 「AI 윤리 권고」를 채택하였고, 이 권고는 대다수 유엔회원국이 참여하였기에 향후 각국의 국내 AI 규제 방향 수립에 중요한 역할을 할 것으로 보입니다.

대한민국은 과학기술, 정보통신 분야의 선진국으로써 AI 윤리 권고 이행을 위한 모범적인 역할을 수행해야 합니다. 그러나 AI가 우리사회에 미치는 광범위한 영향력과 AI와 관련된 다양한 이해관계자들의 서로 다른 입장을 고려할 때, AI 윤리 권고를 어떻게 한국 사회에서 제도화할 것인가에 대한 구체적인 방법에 대해서는 이제부터 본격적인 토론과 숙의가 필요할 것입니다.

이러한 배경 아래 유네스코한국위원회와 한국법제연구원은 올해 공동으로 '인공지능 윤리와 법' 연구를 추진했습니다. 이번 연구에 참가한 전문가들이 국내외 AI 윤리와 법제화 동향에 대한 객관적 분석을 토대로 제시하고 있는 한국의 AI 윤리 법제화 방안이 앞으로 생산적인 토론을 불러일으키는데 기여할 수 있기를 바라마지 않습니다.

감사합니다.

유네스코한국위원회 사무총장

**한 경 구**

# 발 간 사

지난 4월 한국법제연구원과 유네스코한국위원회 간 교류협력협약서 체결 후 양 기관은 매우 활발한 교류협력활동을 전개하여 왔습니다. 그동안의 교류협력활동의 주요 성과물인 『인공지능(AI) 윤리와 법』의 발간을 진심으로 축하합니다. 코로나19로 자유로운 연구교류활동이 제한되어 어려운 공동연구 여건 속에서도 마지막까지 좋은 성과물을 만들기 위해 노력해 주신 공동연구진께 특히 감사드리고자 합니다.

지난 수년 동안 국내에서는 물론이고 글로벌 차원에서도 여러 정부, 기관, 단체 그리고 기업에 이르기까지 AI의 활용가능성과 그 부작용에 대한 규제 문제를 둘러싸고 많은 논의를 하였습니다. 이러한 논의는 대부분 AI 윤리 원칙의 제정이라는 결과로 이어졌습니다. 유네스코 역시 오랫동안의 논의 끝에 지난 11월 24일에 유네스코 「AI 윤리 권고」를 채택하였고, 이 권고는 대다수 유엔회원국이 참여하였기에 향후 각국의 국내 AI 규제 방향 수립에 중요한 역할을 할 것으로 보입니다.

최근 들어 일부 국가 또는 지역에서는 그동안의 AI 윤리 원칙 제정 논의를 넘어 AI 윤리 원칙의 법제화를 시도하고 있습니다. 대표적인 사례가 올해 초 공개된 EU의 「인공지능법안」이라 할 수 있습니다. 이 법안에 그 동안 AI 윤리 원칙 차원에서 논의 되었던 내용들이 포함되면서 AI 규제가 윤리 원칙 차원에서 나아가 법제도화의 단계로 넘어갈 가능성이 커지고 있습니다.

이와 같은 국제적인 동향 하에서 AI 윤리 이슈에 대한 전문기관인 유네스코한국위원회와 국내 입법 분야 전문국책연구기관인 한국법제연구원이 그동안의 국내외 AI 윤리와 법제화 논의를 종합·분석하여 우리나라에서의 AI 윤리 관련 법제화 방향을 제시한 이번 연구는 향후 국내 AI 관련 정책 및 법제화 논의 시 중요한 참고가 될 것으로 기대됩니다.

AI 규제 거버넌스를 둘러싸고 윤리적 수준에서 다룰 것이냐 법적 강제를 할 것이냐를 두고 아직도 많은 논의가 필요한 상황입니다. 이에 따라 양 기관 간 연구협력은 이번 한번으로는 끝낼 것이 아니라 향후에도 올해 이룩한 연구성과를 바탕으로 더욱 활발한 교류협력 활동이 이루어지기를 기원합니다. 감사합니다.

한국법제연구원 원장

**김 계 홍**

# 목차

# 목차



## 제2장 AI 윤리 거버넌스의 사회적 기반 / 43

# 제1장

# AI 윤리의 쟁점과 과제

**이 상 욱** (한양대학교 철학과 교수)

# 제1장 AI 윤리의 쟁점과 과제

## 제1절 글로벌 AI 윤리 논의의 지형[1)]

### 1. AI 윤리(Ethics)의 시대

인공지능(AI)의 시대가 도래했다. SF 영화에나 등장하는 사람과 구분되지 않는 안드로이드 로봇은 아직 먼 미래의 꿈이지만, 그보다는 훨씬 친숙하고 널리 퍼져 있는 AI가 우리에게 휴대전화라는 기술적 대상 안에 내장되어 이미 일상생활 속 깊이 파고들고 있다. 이뿐만이 아니다. 오늘 저녁 어떤 영화를 볼 것인지를 결정하거나 휴가 때 읽을 책을 선택하는 과정에서도 우리 중 많은 사람들은 이미 AI의 도움을 받고 있다. AI가 추천해 주는 선택지는 아직까지는 가끔 성가실 정도로 엉뚱한 것일 수도 있지만 상당히 많은 경우에는 꽤 쓸 만하다는 느낌이 들기도 한다. 처음에는 엉뚱해 보였던 추천 영화가 막상 보니 정말 내 취향에 딱 맞는다고 느낄 수도 있다. 이런 상황이면 조만간 기술이 더 발전해서 '나보다 나를 더 잘 아는' AI가 현실화되지 않을까 기대해 볼 수도 있다(Gans, Goldfarb, and Agrawal 2018).

AI의 일상화만큼이나 최근 국내외에서는 AI가 제기하는 여러 인문학적, 사회과학적 쟁점을 학술적으로, 실천적으로 탐색하는 연구 및 관련 활동도 활발하다. 전통적으로 인간만이 할 수 있었던 법률, 의료, 세무 등의 일자리 영역에서도 AI 활용이 늘어나면서 대량

1) 2.1 절의 일부(2.1.1, 2.1.2, 2.1.3, 2.1,4) 내용은 필자가 고등과학원 웹진 〈HORIZON〉에 2021년 5월 31일 발표한 내용을 바탕으로 작성되었다.

실업 사태가 일어날 수 있다는 종말론적 두려움과 이를 정반대로 해석해서 인간이 노동으로부터 해방된 자유를 얻게 되리라는 유토피아적 기대가 함께 제시되고 있다(Mason 2016). 어느 상황이 실현되든 하나의 해결책으로 논의되고 있는 '기본소득' 개념은 어느덧 상식적 담론이 되었다(Suskind and Suskind 2017).

AI에 대한 이런 다양한 쟁점을 통합적으로 다루는 분야를 최근 국제 논의 맥락에서는 대개 AI 윤리(ethics)라는 개념으로 포괄한다. 경제적으로 발전한 나라들의 모임인 OECD에서 AI의 개발이 가져다 줄 혜택과 위험을 고려하여 사회적으로 수용가능한 수준의 절충점을 찾으려는 노력을 할 때도 AI 윤리 원칙(ethical principles)이라는 용어를 사용하고, 최근 유엔기구 중에서 가장 활발하게 AI 윤리 논의와 규범적 틀 마련을 위해 노력하고 있는 유네스코도 AI 윤리라는 용어를 사용한다(UNESCO 2019, 2020). 특히 유네스코는 2021년 11월 총회에서 채택 예정인 AI 윤리 권고에서 AI가 개인과 사회와 맺는 다양한 접점을 포괄적으로 탐색함으로써, OECD와 EU의 AI 윤리 논의가 주로 경제적/기술적 발전에 집중하여 윤리적 쟁점을 제기한 것과 대비된다. 이는 이후 설명할 Ethics의 포괄적 의미를 온전하게 담아내려는 시도라고 평가할 수 있으며 이런 의미에서 국제적으로 다수 제안되어 있는 여러 AI 윤리 논의와 차별성을 지닌다고 볼 수 있다.

이뿐만이 아니다. 세계적으로 가장 큰 전기전자공학자단체인 IEEE(Institute of Electronical and Electronics Enginerrs)는 AI가라는 단어가 줄 수 있는 불필요한 의인화 등을 걱정하여 AI라는 용어보다는 A/IS(Autonomous Intelligent System)라는 용어를 선호한다. 그런 IEEE 또한 A/IS의 설계 단계에서부터 '윤리원칙에 일치하는 설계(Ethically Aligned Design)' 개념을 강조하며 아예 그와 관련된 국제 표준 마련에 힘을 쏟고 있다(IEEE 2020).

그런데 국내에서는 AI 윤리라는 용어 자체에 대해 어색해 하거나 불편해 하는 사람들이 꽤 있다. 이런 태도의 배경에는 가치와 무관한 과학에 윤리를 연결시키는 것이 부당하다는 직관이 있다고 볼 수도 있다. 자료를 조작하거나 다른 사람 연구를 표절하는 등 연구 부정행위를 저지르지 않는 한 과학이나 기술은 윤리와는 무관한 가치중립적 영역이라는 생각

에서 일 것이다. 실제로 우리에게는 윤리와 과학기술은 극단적인 오용 사례를 제외하고는 무관하다는, 혹은 관련이 있더라도 아주 막연한 원칙 제시 수준에서만 관련된다는 직관이 매우 강하게 있는 편이다(이상욱, 조은희 2011).

하지만 바람직한 과학은 가치와 무관해야 된다는 생각은 여러 이유로 정당화되기 어렵다(Kitcher 2001). 성공적으로 과학 연구를 수행하기 위해서는 단순성이나 설명력처럼 인식적 가치를 활용하는 것이 결정적으로 중요한데다, 연구 주제 설정에 있어서도 사회적 가치를 고려하는 것이 바람직하기 때문이다. 가치가 과학 연구에서 정확히 어떤 역할을 수행하는지에 대한 복잡한 논의를 잠시 접어 두더라도 '윤리'와 AI 사이의 밀접한 관계에 대한 설명이 가능하다.

일단 상당수의 사람들이 AI라는 기술적 대상에 대해 사람들의 개인적 행동에 적용되는 '윤리'라는 개념을 적용하는 것 자체가 이상하다고 느낀다는 점에서 출발해 보자. 이런 분들일수록 AI 윤리 논의 전체가 AI 관련 과학기술 연구의 '발목을 잡으려는' 비생산적 논의라고 규정하는 경향이 많다. 특히 과학기술 진흥과 연구개발의 효율성에 집중하는 정부 관료들 중에서는 국제적으로 이루어지는 AI 윤리 논의 자체가 형용모순이라고 생각하거나, 우리보다 AI 연구가 앞선 기술 선진국들이 윤리 논의로 우리나라와 같은 AI 기술 후발주자의 발을 묶으려는 '사다리 걷어차기' 전략이라고 의심하는 분조차 있다.

## 2. 윤리(倫理) vs. ethics

AI 윤리에 대해 국내외에서 왜 이런 차이가 발생하는 것일까? 여러 이유가 있겠지만 필자는 분리할 수 없는(incommensurate) 두 개념, 즉 윤리(倫理)와 ethics 사이의 의미 차이가 중요한 이유라고 생각한다. 일단 그 이야기부터 해보자.

우리의 일상적인 언어 직관에 따르자면, '윤리'는 지극히 개인적인 사안에만 한정된다는 느낌이 있다. 이 직관은 표준국어대사전의 '윤리'에 대한 정의, "사람으로서 마땅히 행하거나 지켜야 할 도리"와도 일치한다. 이 정의에서 연상되는 상황은 천륜을 어기고 부모를

학대하는 행위나 상식적인 허용 범위를 넘어 극단적으로 자기이익만 챙기는 행위가 될 것 같다. 즉, 우리말에서 윤리란 개인이 누구에게나 명백하게 도리에 어긋나는 행동을 하는 것과 긴밀하게 관련되는 개념이다. 표준국어대사전은 윤리 개념의 용례로 채만식의 〈낙조〉라는 소설에 등장하는 "아내가 있는 사람이 한 다른 여자와 연애를 하고 어쩌고 한다는 것은, 나의 윤리로는 허락할 수 없는 패덕이었다."는 문장을 들고 있는데, 이 문장에서도 우리의 윤리 개념이 개인적 사안과 관련된 것이며 명백한 잘잘못을 다룬다는 특징이 잘 드러나 있다.

이제 이런 윤리 개념으로 AI 윤리라는 표현을 살펴보면 누가 봐도 이상하다는 느낌을 갖지 않을 수 없다. 일단  AI 윤리에서 다루는 내용은 최근 문제가 된 AI 챗봇 이루다의 사례처럼 지극히 사회적이고 논쟁적이다. 이루다 사건의 경우에는 그것의 문제라는 점에 대해 대체적으로 사회적 합의가 이루어졌지만 많은 AI 윤리 쟁점은 그렇지 않다. 예를 들어 AI 알고리즘의 투명성을 높이거나 설명 가능성을 강하게 요구하다 보면 AI의 효율성이 저하되거나 민감 정보의 유출 가능성이 높아질 수도 있다(Mitchell 2020). 이처럼 현재 AI 윤리에서 논의되고 있는 내용은 (당연히 개인적 영역도 포함하지만) 많은 경우 사회적 수준에서 문제를 파악하고 해결책을 마련해야 하는 부분이고, 대부분의 경우 그 문제점 분석이나 해결책 마련 과정 자체가 많은 관련 집단의 이익과 다양한 가치를 종합적으로 고려해야하기 때문에 논쟁적이고 지난한 사회적 숙고를 요구한다(Fry 2019). 우리말의 '윤리' 개념으로 AI 윤리를 제대로 이해하기 어려운 것도 무리는 아니다.

그럼 이제 영어의 ethics는 어떤 의미인지 살펴보자. 어원을 따져 보면 ethics는 고대 그리스어에서 '인격(character)'을 뜻하는 단어 ethos, 그리고 라틴어에서 '관습(customs)'을 뜻하는 단어 mores와 깊은 관련이 있다. mores라는 단어는 우리가 흔히 '도덕적'이라고 번역하는 영어 'moral'의 어원이기도 하다. 우리 일상 표현에서도 윤리적과 도덕적을 서로 혼용해서 쓰듯이 영어에서도 (철학적으로 엄밀하게 구별할 때를 제외하면) 이 둘을 혼용해서 쓰는 경향이 있다. 그래서 옥스퍼드 영어사전에서 제시된 ethic의 정의는 "A set of moral principles, especially ones relating to or affirming a specified group,

field, or form of conduct"이다. 이 정의에서 주목할 점은 ethic의 정의에 특정 집단, 분야, 행위의 종류가 등장한다는 사실이다. 이는 앞서 지적했듯이 ethic의 어원에 특정 집단이나 분야마다 공유되는 올바름의 기준이 다를 수 있는, '관습'의 의미가 포함되어 있다는 점과 일맥상통한다. 그리고 이러한 특징은 우리말의 '윤리'와 달리 영어의 ethic이 특정 개인의 행동 자체만이 아니라 그 행동의 사회적 의미까지를 본질적으로 포함하고 있음을 시사한다.

서양 문명의 기원이라고 여겨지는 그리스-로마 시대의 ethic에 해당하는 개념이 이처럼 개인적 수준과 사회적 수준을 가로지르고 있다는 사실을 염두에 두면, 황우석 연구팀의 논문조작 사건으로 촉발된 '연구 윤리(research ethics)'라는 개념이나 최근 강조되고 있는 '전문직 윤리(professional ethics)' 개념이 결코 ethic 개념을 최근에 확장된 것이 아니라는 것을 짐작할 수 있다(Harris et al. 2018). 그보다는 이들 용어는 특정 집단에 고유한 내적 규범을 의미하는 ethic 본래의 의미에 충실한 것이라는 사실이 자연스럽게 이해된다. 과학 연구자가 연구만 열심히 하면 되지 따로 윤리가 왜 필요하냐는 생각은 우리말의 '윤리' 직관을 따른다면 이해될 수 있는 반응이지만 영어의 ethic을 비롯한 국제적 기준에 따른다면 부적절한 반응이라고 볼 수 있는 것이다.

이 지점에서 오해의 여지를 제거할 필요가 있다. 필자는 우리말의 '윤리' 개념이 틀렸고 서양의 ethic 개념이 올바르다고 주장하는 것이 아니다. 그런 지적은 수(number) 개념으로 자연수는 틀린 개념이고 보다 포괄적인 정수나 실수 개념만이 진정한 수 개념이라고 주장하는 것만큼이나 터무니없다. 개념은 원칙적으로 맞고 틀리고의 문제라기보다는 정의의 문제이다. 그런 의미에서 우리말의 '윤리' 개념이나 영어의 'ethics' 개념 모두 동등하게 의미있는 개념이다. 필자의 지적은, 예를 들어 AI 윤리 관련 국제 논의에서 대부분의 나라는 모두 ethic의 의미를 배경으로 참여하는데 우리만 우리말에 고유한 '윤리' 개념을 갖고 참여한다면 생산적인 의사소통이나 논의 참여가 어려울 것이라는 점이다. AI ethics와 관련하여 국제적으로 통용될 수 있는 방안을 만들거나 법제도화 등을 추진할 때 우리가 반드시 명심해야 할 부분이 바로 이것이다.

그렇다면 이렇게 개인과 사회를 가로지르는 의미의 윤리적 논의가 개인행동의 선악에 초점을 맞춘 우리의 윤리 논의와 구체적으로 어디에서 차이가 날까? 앞서 소개한 여러 AI 윤리 국제 논의에서 분명하게 부각되는 차이점은 우리가 사회적으로 추구해야 할 가치가 여럿이라는 사실, 그리고 그 가치들 사이에서는 종종 충돌이 일어난다는 사실이다. 이런 상황에서 공정하고 효율적인 윤리적 해결책은 거의 대부분의 경우 고려해야 할 여러 가치를 사회적으로 수용 가능한 방식으로 맞교환(tradeoff)하는 방식으로 얻어지게 된다. 그리고 그런 과정에서 직관적으로 '좋은 것들' 사이에 절충이나 선택을 해야 하는 경우도 발생하게 된다. 개인의 행동에 대한 선악 판단에서 암묵적으로 전제되는 '명백함'이나 '착하게 살면 윤리와 무관할 수 있다.'는 직관이 더 이상 통용되지 않는 것이다.

우리가 사회적 수준에서 추구하는 여러 가치, 예를 들어 자유와 평등 사이에는 동시에 만족하기 어려운 긴장이 존재한다. 그리고 이 여러 가치를 최대한 동시에 실현하기 위해서는 관련 이해 당사자가 모두 완벽하게 만족하는, 현실적으로 불가능한 방식이 아니라, 사회적 숙고를 통해 윤리적으로 합리적이라고 평가될 수 있는 방식으로 각각의 가치를 적절한 수준에서 절충하여 만족하는 방식을 활용하게 된다. 당연히 AI 윤리의 여러 핵심 주제에 대해서도 마찬가지 방식으로 주요 사회적 결정이 내려질 수밖에 없다.[2)]

이렇게 이해된 AI 윤리의 관점에서 보자면 AI와 관련된 다양한 개인적, 사회적, 법적, 제도적 쟁점에 대해 단순한 선악 판단을 하려고 시도하기 보다는 우리 사회에서 핵심적으로 존중되는 가치에는 어떤 것이 있으며 그 가치를 최대한 균형 있게 존중하는 방식으로 AI 개발과 활용을 하기 위해서는 어떤 점에 주의하고 어떤 제도적 장치를 마련해야 하는지를 통합적으로 탐색하려는 노력이 필요하다(이중원 외 2019, 한국인공지능법학회 2019; Sharre 2019).

이제부터는 이렇게 여러 가치를 통합적으로 고려하고 사회적으로 수용가능한 해결책을 찾아가는 과정이 정확히 무엇을 의미하는지를 보여주는 AI 윤리의 사례를 소개한다.

---

2) 앞서 소개한 유네스코 AI 윤리 권고는 이 점을 명확하게 인식하고 이 맞교환을 합리적으로 수행하는 것의 중요성을 강조하고 있다. 이는 유네스코 AI 윤리 권고가 다른 AI 윤리 관련 국제 문서와 차별점을 보이는 대목 중 하나이다.

## 3. AI는 인간보다 더 공정할까?

이루다 사건으로 AI의 공정성이 사회적 관심사로 떠오르기 전까지만 해도 AI는 인간의 편견이나 사사로운 감정으로부터 자유롭기에 인간보다 훨씬 더 공정할 것이라는 생각이 지배적이었다. 유사한 사건에 대해 그때그때 기분에 따라 다른 형량을 부과할 수 있는 인간 판사 대신 객관적인 증거와 유사 사건의 판례만을 공정하게 참조하여 판단할 수 있는 AI 판사에게 재판을 받고 싶다는 희망을 피력하는 사람도 있었다. 너무나 고려할 것이 많은 복잡한 의료 현장에서도 실수 없이 차분하게 정확한 진단이나 처방을 내리는 AI 의사를 사람 의사보다 더 신뢰한다는 의견이 언론에 보도되기도 했다. 하지만 이제 이루다 사건 이후로 사람들은 AI가 인간보다 더 공정할 수도 있지만 극단적인 방식으로 더 편견에 사로잡힐 수도 있다는 걸 알게 되었다.

그런데 정말 그럴까? 도대체 AI가 공정하거나 편견을 갖는다는 것은 정확히 무엇을 의미할까? AI가 공정해야 하는지에 대해 답하기 전에 이 문제부터 살펴보자.

AI와 관련된 윤리적 쟁점을 다룰 때 미리 분명하게 짚고 넘어가야 하는 점은 현실에 존재하는 (그리고 가까운 미래에 등장할) AI와 SF 영화에 등장하는, 인간과 구별되지 않는 수준의 감정 능력과 도덕적 판단 능력까지 발휘하는 가상의 AI 사이의 구별이다. 가까운 미래를 포함하여 당분간 우리가 경험할 AI는 인간의 특정한 능력을 '흉내'낼 목적으로 만들어진 특수지능이다. 이 사실은 중요한 함의를 갖는다(이중원 외 2018; Kaplan 2016).

첫째, '이루다'와 같은 AI가 아무리 성차별적으로 간주될 수 있는 발언을 한다고 해도 AI는 상식적 의미에서, 성차별적 의도나 감정을 갖지 않는다. 실은 '성차별'을 포함하여 자신이 산출하는 문장들의 의미를 통상적인 의미에서 이해한다고 볼 수도 없다. 예를 들어 이루다가 산출하는 문장 기호를 우리가 읽고 이루다의 '마음 상태'를 유추할 뿐이지, 실제로 이루다가 의식적 마음을 갖고 있지는 않다.

둘째, 이루다를 비롯한 챗봇 AI가 지금보다 훨씬 더 발달해서 인간과 전혀 구별할 수 없는 수준의 대화를 나눌 수 있게 되더라도, 그 AI가 평범한 인간이 하는 다른 일, 예를

들어 시각이나 음성을 통해 사람을 알아보거나 가게에 가서 물건을 사는 일까지 할 수는 없다. 물론 시각이나 음성을 통해 사람을 구별하거나 물건을 집거나 들어 올리는 일을 할 수 있는 AI 혹은 AI 로봇은 지금도 존재한다. 하지만 평범한 사람처럼 이 모든 일을 포함해 수많은 다른 일, 예를 들어 다른 사람과 협력해서 공동 작업을 수행하는 일 등을 인간 수준으로 해낼 수 있는 '일반지능(General Intelligence)'을 갖춘 AI는 아직 존재하지 않는다. 관련 연구조차 극히 초보 단계여서 가까운 시일 내에 우리 삶에서 일반 AI를 쉽게 볼 수 있을 가능성도 거의 없다.

## 4. AI '산출물'의 공정성

그러므로 이런 배경에서 AI의 공정성은 다음과 같이 이해해야 한다. 현재까지 (그리고 가까운 미래에 등장할) AI는 공정이란 단어의 의미도 알 수 없고 공정과 관련된 복잡한 의미론적, 사회적, 윤리적 관계를 이해할 수 있는 '의식적 마음'도 가질 수 없다. 그러므로 AI가 사람보다 더 혹은 덜 공정한가라는 질문은 이런 의식적 마음을 갖지 않은 복잡한 기계가 수많은 공학자들의 노력과 엄청난 양의 학습 데이터를 활용한 기계 학습을 기반으로 산출하는 결과물이 사람이 보기에 동일한 일을 수행하는 사람이 산출한 결과물보다 더 혹은 덜 공정한가를 의미한다.

이렇게 정리하고 나면 처음 제기한 문제는 너무 쉽게 답할 수 있어 보인다. 결국 AI가 최대한 공정하게 결과 값을 내도록 잘 만들면 되지 않을까? 그런데 이 지점부터 문제가 복잡해진다. 본격적인 AI 윤리 논의가 시작되는 것이다. 우리는 AI의 결과 값이 공정한 것을 항상 원하는가? 조금만 생각해 봐도 우리가 AI을 활용하는 목적에 따라 그 답은 달라질 것 같다.

우선 공정이 무엇인지 생각해 보자. 표준국어대사전은 공정을 '공평하고 올바름'으로 정의한다. 핵심은 공정이란 개념은 평가적 혹은 규범적 개념이라는 점이다. 이 말이 무엇을 의미하는지 이해하기 위해 예를 들어보자. 여성의 '평균' 키는 남성의 '평균' 키보다 약간

작다. 이는 통계적 사실이고 이 사실을 말한다고 해서 성차별적이고 공정하지 않다고 말할 사람은 없다. 하지만 국내 100대 기업의 최고경영자 중에서 남성이 여성보다 압도적으로 많다는 점 역시 통계적 사실이지만 이 사실은 많은 사람들에 의해 성차별적이고 공정하지 않은 것으로 여겨진다. 차이가 뭘까? 이 두 사례를 비교해보면, '공정함'이란 세상이 어떠하다는 사실적 주장과 관련된 것이 아니라 세상이 마땅히 어떠해야 한다는 규범적 주장과 관련됨을 알 수 있다. 대다수 사람들이 남녀 평균키가 같은 것이 윤리적으로 더 바람직하다고 보지 않는 반면, 남성과 여성이 비슷한 비율로 최고경영자가 되는 것이 윤리적으로 더 바람직하다고 보는 사람은 많기 때문이다.

그런데 이렇게 정리해도 여전히 남는 문제가 있다. 여성 최고경영자가 남성에 비해 적은 이유는 실제로 '현재 기업 환경 조건'에서 남성 최고경영자가 여성보다 더 높은 성취를 보여주기 때문일 수도 있다. 이 경우에는 남성과 여성이 최고경영자로서의 '잠재력'에 있어서는 평균적으로 완전히 동등하더라도, 기업 입장에서는 남성 최고경영자를 임용하는 것이 기업 '실제' 실적에 도움이 되기 때문에 남성 최고경영자를 선호할 수 있다. 이런 고려까지 하게 되면 결국 최고경영자 비율에서 남녀차이를 공정하지 않다고 지적하는 것은 '현재 기업 환경 조건'을 포함한 우리 사회 전체에 존재하는 여성에게 불리한 사회적 조건 전체에 대해 비판하는 것이 된다. 물론 이런 상황이 여성에게만 해당될 이유는 없다. 혹자는 현재 남성에게만 부여되는 병역의무가 남성에게 공정하지 않다고 주장하거나 남자에게 '남자다움'을 요구하는 우리 사회가 문화적 포용력을 결여하고 있다고 비판할 수 있다. 핵심은 '공정함'에 대한 규범적 판단은 단순히 여성이라는 이유로 고등교육를 받을 기회를 박탈하는 것이 부당하다는 생각처럼 광범위한 지지를 얻을 수 있는 것부터, 여성이 남성과 동등한 성취를 보이지 못하는 모든 사례가 우리 사회에 내재한 성적 불평등 탓이라는 생각처럼 논쟁적 사안처럼 다양한 스펙트럼으로 존재한다는 사실이다.

그런데 여기서 잠깐 멈추어서 우리 사회의 공정하지 못한 '측면'을 파악하고 이에 대한 대응책을 마련하기 위해 AI를 활용하는 상황을 고려해 보자. 최근 사회정책 수립이나 사회문제 해결에 AI를 활용하면 좋을 것이라는 생각이 점점 인기를 얻고 있으니 충분히 가능한

상황이다. 이런 목적이라면 우리 사회에 어떤 불평등한 모습이 있는지를 가감 없이 그대로 드러내는 AI가 필요할 것이다. 이런 AI의 산출물이 보여주는 우리 사회의 불평등한 모습이 그에 대한 교정적 정책을 시행할 수 있는 정확한 출발점이 되어야 하기 때문이다.

이처럼 AI의 용도에 따라 AI의 공정성 즉 AI 산출물의 공정성은 추구할만한 가치일 수도 있고 그렇지 않을 수도 있다. 특히 현재 사용되는 AI의 대부분이 현재까지 수집된 데이터를 기계 학습하고 그 데이터 집합에서 발견되는 규칙성 혹은 패턴이 가까운 미래에도 성립할 것이라는 전제하에 미래를 예측한다. 이는 AI의 예측이 근본적인 수준에서 '보수적'일 수밖에 없음을 의미한다. 현재까지 행해져왔던 사회적 결정과 행동의 패턴이 미래에도 그대로 실현될 것이라는 가정을 깔고 있기 때문이다. 이런 목적으로 만들어지고 활용된 AI의 산출물이 공정하지 못하다고 비판하는 것은 그 자체로는 맞는 이야기지만 초점을 잃은 비판일 수 있다.

이제 질문을 좀 더 정교하게 가다듬어 보자. AI의 제작 목적에 따라 그 산출의 공정성을 요구하지 말아야 할 AI가 분명 존재한다. 그러므로 이런 종류의 AI를 제외하고 우리가 AI의 산출물의 공정함 자체를 요구해야 할 AI가 분명 있을 것이고 그에 대해 공정하기를 요구하는 것은 윤리적으로 바람직할 것이다. 이번에 문제가 된 이루다처럼 수많은 사람들과 미리 예측하기 어려운 방식으로 상호작용하는 사회적 대화형 AI의 경우에는 당연히 그런 공정성에 대한 요구가 강해질 수밖에 없다.

하지만 이 경우조차 사안은 여전히 더 복잡하다. 이루다와 같은 사회적 파급효과가 큰 AI에 공정함을 요구하는 것이 규범적으로 타당하다는 점에는 논란의 여지가 없지만 그때 요구되는 공정함이 어느 정도 수준이어야 하는지에 대해서는 사람들 사이의 직관이 쉽게 일치하지 않기 때문이다. 예를 들어 사람들에게 웃음을 선사할 목적으로 제작된 오락 프로그램에 지나치게 강력한 도덕적 잣대나 '정치적 올바름(political correctness)'을 요구하는 것은 오락 프로그램의 본질을 훼손하는 것이라는 비판이 제기되어 왔다. 우리는 챗봇 AI가 논란의 소지가 완벽하게 제거된, 물샐틈없이 '도덕적인 문장'만을 발화하기를 원하는가? 이 부분도 고민이 필요한 주제이다.

헌법이 보장하는 표현의 자유와의 충돌 문제도 있다. 물론 표현의 자유가 다른 모든 사회적 가치를 희생하면서까지 반드시 지켜야 할 절대적 가치는 아니다. 국제적으로 여러 국가들이 자국의 역사적, 문화적, 사회적 상황에 따라 특정 종류의 혐오표현에 대해 법적으로 처벌할 근거를 마련하고 있다. 그러므로 우리는 AI 설계 단계부터 표현의 자유와 다른 사회적 가치의 맞교환 문제를 고민해야 한다. 중요한 점은 이루다와 같은 공정함을 요구할 필요가 있는 AI의 기획 및 제작 단계에서 어느 정도의 '공정함'이 적절한 수준인지를 미리 고민하고 이를 알고리즘이나 데이터 수집 및 활용 과정에 반영하는 것이다.

이상의 논의를 정리해 보면 다음과 같다. 우리는 사회적 상호작용을 비롯하여 사람들의 행동이나 가치에 큰 영향을 끼치는 AI의 '산출물'이 공정할 것을 요구해야 한다. 그런데 어느 수준의 공정함을 요구해야 할지는 AI 제작 단계에서부터 충분한 학제적 논의를 통해 결정되어야 하고 이 결정 내용이 알고리즘 자체나 훈련 데이터의 수집 및 활용 과정에 반영되어야 한다. 핵심은 AI가 단순히 공학자들이 만드는 기술이 아니라 우리 사회 전체의 윤리적 공감대를 반영해야 할 문화적 산물이라는 점을 인식하고 실천하는 것이다.

이상의 논의를 통해 우리는 AI 윤리가 무엇인지에 대해 답할 준비가 되었다. AI 윤리는 (먼 미래에 등장할 일반 지능을 갖춘 AI를 배제하면) AI의 '산출물', 특히 인간의 지속적인 통제를 받지 않는 '자동화된 결정(automated decisions)'이 우리가 소중하게 여기는 기본권 등의 다양한 사회적 가치를 최대한 존중하는 방식으로 활용되기 위해서 어떤 점에 주목하고 어떤 방식의 제도적 대응을 수행해야 하는지에 대한 논의이다. 그리고 이 논의와 그로부터 파생되는 제도적 실천은 AI 개발과 활용의 전 주기(entire lifecycle)에 적용되어야 하고 그 논의가 영향을 미치는 집단의 윤리적 공감대와 문화를 적극적으로 고려해야 한다.

## 5. 국제적 공조와 지역적 고려

앞 절에서 우리는 AI 윤리에 대한 국제적 관심은 '윤리'를 ethics의 의미로 넓게 이해하는 바탕에서 이루어지고 있고, 그런 이유로 어떤 것이 윤리적으로 타당한 결론인지가 자명하거나 모든 사람에게 동일한 호소력을 갖는 방식으로 나타나지는 않는다는 점을 알게 되었다. AI 윤리와 관련하여 자주 지적되는 '공정함'의 문제 역시 예외가 아니라는 점도 확인했다.

그래서인지 AI 관련 윤리적 주제에 대한 국제논의 흐름은 각국의 정부가 AI 윤리 관련 제도적 활동(법제화 노력 포함)을 서로 공유하되 모든 국가에 일률적으로 적용되는 표준을 강요하는 것은 바람직하지 않다는 방향으로 가고 있다. 물론 IEEE처럼 윤리적으로 설계되었다는 인증을 줄 수 있는 기술 표준을 만들려는 시도를 하고 있는 집단이 있기도 하고, 현실적으로 AI 기술 개발을 선도하는 미국이나 큰 시장을 갖고 있는 EU가 특정 AI 윤리 관련 제도를 도입하면 다른 나라들은 그 제도에 어떤 형식으로든 대응할 수밖에 없는 것은 사실이다. 하지만 그 대응 방식에 있어서도 국내 제도적 변화를 미국과 EU의 변화와 동일하게 채택해야 되는지에 대해서는 논란의 여지가 있고, 이 두 나라의 사회적, 문화적 지형의 차이 때문에 이 두 나라가 동일한 제도적 변화를 취할 가능성도 높지 않다.

## 6. 인권과 균형

이런 배경에서 AI 윤리 국제논의는 규범적 호소력이 높은 윤리 원칙을 중심으로 이루어지되 그 윤리 원칙이 구체적으로 어떻게 제도화될 지에 대해서는 각국의 여건을 고려하는 방식, 즉 글로벌 스탠다드와 맥락적 고려를 모두 중시하는 방식으로 진행되고 있다. 하지만 그렇다고 해서 국제 논의의 전체적인 흐름에 아무런 규칙성이 없는 것은 아니다. 오히려 국제 논의는 거의 예외 없이 '인권'과 '균형'을 강조하고 있다는 규칙성을 보인다.

우선 인권에 대한 강조는 유네스코 AI 윤리 권고를 비롯한 수많은 국제 AI 윤리 관련 문건에서 확인할 수 있다. 이는 물론 현실적으로 AI 기술의 활용이 인권을 훼손할 우려가

있다는 염려에서 비롯된 현상이다. 특히 유럽권 국가들은 법제화된 형태의 기본권(Basic Rights)만이 아니라 자연법적 맥락에서 법제화 이전의 불가침적 권리로 인정되는 '근본적 자유(Fundamental Freedom)'에 대한 강조가 두드러진다.

여기서 주목할 점이 두 가지 있다. 첫째는 대중매체나 AI 관련 학술적 논의에서 자주 등장하는 '로봇의 권리'나 '로봇에게 인격권을 부여할 가능성'에 대한 논의는 적어도 국가 단위의 논의에서는 찾아보기 어렵다는 사실이다. 로봇에게 법인격을 부여할 필요성이 비교적 가장 활발하게 논의된 자율주행차의 경우조차 이는 마찬가지이다. 이는 결국 AI 윤리의 추상적 원칙 수준이나 실제 법제도화 수준 모두에서 '인권'에 대한 배타적 강조가 두드러짐을 보여준다. 물론 AI 기술이 좀 더 발전해서 피터 싱어가 말하는 '(도덕적 고려의) 원 확장하기(expanding the circle)'의 필요성에 많은 사람들이 공감하게 된다면 상황이 달라질 수 있다(Singer 2011). 하지만 그런 '확장'이 사회적 공감대를 확보하기 전까지는 로봇이나 인공지능을 도덕적 논의의 맥락에서 동물에 비유하는 것은 아직 시기상조라고 판단된다.

둘째는 AI 윤리 관련 여러 문건에서 '편협한' 인간중심주의에 대한 비판이 등장함에도 불구하고 많은 경우 인권을 비롯한 인간적 가치에 대한 강조는 여전히 대부분의 논의에서 강한 호소력을 갖고 있다는 사실이다.

예를 들어, 유네스코 AI 윤리 권고에서는 생태적 번영(ecological flourishing) 등의 개념을 통해 인간이 기술적 발전 과정에서 생태계의 일원임을 기억하고 이에 합당한 실천을 할 것을 요구하고 있지만, 문건의 다른 부분에서는 인류의 기본권에 대해 여러 차례 강조하고 있다. 그나마 유네스코 권고는 이런 다양한 윤리적 고려 사이의 '맞교환(tradeoff)'의 중요성을 언급하고 있지만, 다른 AI 윤리 문건에서는 생태적 고려가 오직 AI 활용에서의 에너지 및 환경 폐기물에 대한 고려에서만 등장하는 경향이 있다. 이는 인간중심주의가 AI 윤리에서 정확히 어떻게 이해되어야 하는지에 대한 잠재적 긴장을 드러내 준다고 볼 수 있다.

인권만큼이나 AI 윤리 국제 논의에서 두드러진 경향은 AI 기술 혁신과 사회적 가치 보전 사이의 '균형'에 대한 강조이다. 이는 EU나 OECD 문건에서 두드러지는데, AI 기술이 가져올 수 있는 인류 복지에 대한 잠재적 혜택을 강조하고 그런 이유로 AI 기술 혁신이 중요하다는 점을 분명하게 지적한다. 다만 이 때 AI 기술 혁신은 '사회 속의 혁신'이 되어야 한다. 즉, 인권을 비롯하여 현재 각국의 사회적 맥락에서 중요시되는 가치를 훼손하지 않는 방식으로 혁신이 이룩되어야 한다는 것이다.

이런 AI 윤리의 국제 논의 동향은 AI 기술 자체에 대한, 혹은 적어도 그 기술의 잠재적 혜택에 대한 긍정에서 출발한다는 점에서 유전공학이나 나노기술처럼 앞선 '신기술' 유행과는 조금 다른 양상을 보인다. 예를 들어 현재 진행 중인 AI 윤리 관련 국제 논의에서 AI 기술의 잠재적 위험을 들어 AI 기술 개발 자체를 중지해야 한다고 주장하는 경우를 찾아보기 어렵다. 물론 자율형 전쟁 무기로서의 AI에 대해서는 전면 금지가 필요하다는 지적은 자주 볼 수 있지만, 이 경우 역시 전반적인 AI 윤리 논의에서 (여러 정치적 이유로) 가볍게 언급되거나 아예 다루어지지 않는 경우가 많다.

이상의 논의를 통해 우리는 AI 윤리의 국제적 지평이 다소 복잡한 양상을 보여준다는 점을 알 수 있다. 국내 논의에서, 특히 기업에 대한 규제가 논의될 때마다 자주 언급되는 '국제 기준'은 분명 존재하지만 그 내용은 인권과 균형을 강조하는 것이기에 그 구체적 제도화의 방향은 각국 정부의 판단이 작용할 여지가 상당히 많다. 하지만 그렇다고 해서 AI 기술 개발과 활용의 국제적 성격을 고려할 때 각국이 자국의 상황만을 고려해서 정책적 대응을 하는 것은 비현실적이다. 그러므로 현 단계에서 우리가 취할 적절한 제도적 대응은 국제 논의의 동향을 잘 분석하고 국내 상황을 고려한 내부 논의를 지속하면서, 가능하다면 국내 논의의 결과를 국제 기준 마련 작업에서 반영하는 방식으로 AI 윤리와 관련된 국내외 제도적 논의를 공조화할 필요성이 있다.

이를 위해서는 국제적 논의에서 비교적 큰 틀의 합의를 이룬 '인권'과 '균형' 같은 개념의 내용을 정확하게 파악하는 것도 중요하지만, 국제 논의에서도 논쟁적이고 간단한 합의점 도출이 어려운 주제에 대해 인지하고 이를 국내 논의 과정에서 탐색하여 국제 합의 형성 과정에 참여하는 것도 필요하다. 다음 절에서는 그런 문제들을 다루어 본다.

## 제2절 난제와 쟁점

### 1. Human in/on the loop?[3)]

AI 윤리에 대해 국제적으로 나라마다 이해관계에 따라 조금씩 다른 입장을 보이고 있다. 특히 어느 '수준'으로 윤리적 원칙을 제시해야 하는지, 윤리적 '원칙만' 제시할 것인지 아니면 그 원칙을 실행할 수 있는 구체적인 정책 행동도 제시할 것인지를 두고 AI 기술선진국과 후발 국가들 사이의 의견 차이가 크다. 플로리디(Floridi) 같은 학자들은 이런 상황이 AI 관련 국제 규제의 한계로 작용할 가능성을 우려하기도 한다.[4)]

유네스코는 다수의 회원국이 AI 기술발전이 성숙되지 않은 국가여서, 11월 총회에 상정될 AI 윤리 권고에서도 대체적으로 보다 구체적인 수준에서 윤리 원칙을 제시하고 그 원칙과 연동된 정책 행동을 가능한 저개발국가에게 유리한 방식으로 표현하려는 경향성이 보인다. 하지만 이에 대해 회원국 내부에서도 상당히 강한 의견 차이 또한 있는 것도 사실이다.

그럼에도 불구하고 적어도 AI 윤리가 '인간중심적(human Centered)'이어야 한다는 점에 있어서는 유네스코 회원국 사이에 의견 차이가 거의 없고, 이는 AI 윤리 관련 다른 논의에서도 마찬가지이다. 이는 기본적으로 현재까지 그리고 가까운 미래에 개발될 AI는 인간이 '도구'로서 활용하는 것이지 인간과 '동등'하게 대우하여야 할 존재가 아니라는 생각, 그리고 AI와 관련된 여러 결정에 있어서 기계에 불과한 AI에 책임을 돌리지 말고 인간이 어떤 형태로든지 책임을 져야 한다는 생각 등이 포함된다. 마지막으로 AI의 활용은 다양한 방식으로 '인간'의 복지를 증진해야 한다는 생각도 중요하게 부각되고 있다.

그런데 보다 구체적인 수준에서 '인간중심주의'를 AI 윤리에서 어떻게 실현할 것인지를

3) 이 소절의 내용에 대해 중간발표회 때 유용한 보완 제안을 해주신 아마존웹서비스(AWS)코리아 김영훈 정책협력실장님과 한국법제연구원 배건이 글로벌법제전략팀장님께 감사드린다.

4) 이에 대한 보다 자세한 논의는 본 보고서의 제2장. AI 윤리 거버넌스의 사회적 기반 참조.

생각해 보면 복잡한 쟁점이 많다는 것을 알게 된다. 우선 AI 활용에서 인간의 책임성을 요구하는 것은 자율주행차 상황처럼 인간이 실시간으로 판단하여 결정을 내리는 것이 현실적으로 어렵거나 불가능한 경우에는 적용되기 어렵다. 이 경우에 AI는 구체적인 결정에서 인간이 개입할 여지를 남겨놓지 않는다는 의미에서 '기술적으로 자율적'이기 때문이다.

물론 이런 '기술적' 의미에서 자율적이라고 해서 아직까지는 법적 지위를 갖지 못하는 자율주행차 AI에게 책임을 부과하는 것은 아니다. 그보다는 제조물책임법을 변형하여 자율주행차 제조사에게 일종의 '운행관리자' 책임을 부과하여 자율주행차의 알고리즘 설계나 운행 과정 중의 데이터 분석 알고리즘 설계, 실제 운행 과정에서 작동하는 각종 위기 상황에 대한 선제적 대응 등을 만족스러운 수준으로 수행했는지를 '사후적으로' 책임지게 하는 방식이나, 사회 보험을 통해 특정 개인이 아니라 AI 기술을 활용하는 사회가 공동으로 책임지는 방식을 채택하려는 논의가 활발하게 진행되고 있다.

이렇듯 '인간중심주의'가 실제로 AI 활용 과정에서 나타나는 방식은 점점 복잡한 의사결정 과정에서 인간이 반드시 필수불가결한 방식으로 참여하는 Human in the Loop 방식이 아니라 인간이 어떤 방식이든지 AI의 설계와 작동 등 전체 관리 체계에서 중요한 역할을 담당하고 문제가 있다고 판단될 때는 전체 과정을 중지시킬 권한을 갖는 Human on the Loop 방식이 되고 있다. 여기서 'in/on'의 차이는 구체적인 결정 과정 사슬에 인간이 '참여'하는 단계가 있는지(in) 아니면 결정 과정 전체에 대한 감사(oversight) 권한을 갖는지의 차이에 해당된다.

물론 이는 '모든' AI 활용에서 나타나는 경향성은 아니다. 자율주행차의 사례처럼 실시간 데이터 분석과 그에 기반한 실시간 '자동화된 결정'이 본질적 특징인 경우에는 Human on the Loop의 장점이 더 두드러질 것이지만, AI를 활용한 정책 결정처럼 데이터 분석과 가능한 해결책 모색 그리고 이를 제도적으로 공인된 방식으로 시행하는 과정에서 상당한 '정치적' 과정이 필수적으로 요구되는 상황이라면 당연히 Human in the Loop 방식을 채택하는 선택적 대응이 바람직해 보인다. 이를 통해 AI의 '비인간적' 특징으로 인해 필연적으로 발생하는 '이해할 수 없는' 오류의 위험에 대비하고 그 과정에서 AI 거버넌스의

민주적 정당성을 확보할 수 있을 것이다.

그러므로 중요한 점은 이제는 누구도 부인하지 않는 AI 윤리에서의 인간중심주의에 대한 선언적 강조가 아니라 이 '인간중심주의'를 각각의 AI 활용 과정에 맞게 어떤 구체적인 방식으로 실현할 것인지에 대한 논의를 보다 본격적으로 수행해야 한다. 이 과정에서 당연히 인권 존중과 AI 효율성 사이의 맞교환(tradeoff)에 대한 세밀한 고려가 이루어져야 한다. 이 부분은 특히 제도적 영향력이 큰 AI 윤리 제도화 과정에서 진지하게 고려되어야 할 내용이 될 것이다. 예를 들어, AI의 결정이 사회적으로 매우 중요한 영향을 끼칠 때 우리는 그 결정에 대해서는 높은 수준의 '해명가능성(accountability)'를 요구할텐데 이런 상황이라면 다른 가치보다 '인간중심주의'를 Human in the Loop 형식으로 요구해야 한다는 주장이 가능할 수 있다. 그러므로 AI의 자동화된 결정에 대해 인간중심주의를 어떻게 실현하는 것이 바람직한 지는 그 결정이 활용되는 구체적 맥락을 고려하여 신중하게 판단되어야 한다.

유사한 결론이 정반대 방향에서도 가능하다. 어떤 AI 활용의 맥락에서는 AI에 대한 감사 기능을 수행하는 것이 기술적으로 간단하지 않을 수 있다. 예를 들어, AI의 작동을 실시간으로 지켜보다가 AI가 문제를 일으킨다고 생각하면 AI의 작동을 중지시키는 상황을 생각해 보자. 페이스북 엔지니어들이 자신들이 만든 챗봇들이 인간의 언어가 아닌 방식으로 신호를 주고받는 것처럼 보인다고 판단하고, 그 상황을 자세하게 파악하기 전에 챗봇 중단 조치를 취한 것이 이에 해당될 것이다. 이런 상황을 흔히 '중단 스위치(Kill Switch)'를 사용했다고 표현한다. 하지만 상황에 따라 이 중단 스위치를 기술적으로 구현하는 것이 매우 어렵거나 설사 가능하더라도 그 비용이 워낙 커서 관련 AI 기술의 유용성을 심각하게 제한하는 경우가 있을 수 있다. 예를 들어, 도시 전체의 에너지 그리드를 관리하는 AI가 이해할 수 없는 방식으로 작동한다고 해서 그때마다 '중단 스위칠'를 작동한다면 에너지의 효율적 관리는 이루어지기 어려울 것이다. 이 경우에는 에너지의 효율적 관리에 필수적인 부분을 제외하고 다른 영역의 오작동 여부를 분리하여 먼저 점검하는 등의 기술적 조치를 검토해야 할 필요성이 있을 수 있다. 이는 SF 영화에서 등장하는 것처럼 파괴적 성향을 보이는

기계의 '전기 코드'를 빼서 상황을 해결하는 단순한 방식이 실제 공학적 상황에서는 그다지 현실성이 없음을 의미한다. 그러므로 이 경우에도 '인간중심주의'를 AI 거버넌스에 어떻게 적용할 것인지는 여러 가치를 종합적으로 고려하여 그 방안을 모색해야 한다.[5)]

이 지점에서 유네스코 AI 윤리 권고의 독특한 특징을 언급할 필요가 있다. 유네스코는 지속가능한 발전에 대해 환경 보호나 경제 발전만이 아니라 교육과 젠더를 포함한 우리 사회의 여러 중요한 측면에서 강조하고 있다. 이 점을 반영하여 유네스코의 AI 윤리 권고에는 인권과 기본적 자유에 대한 강조 및 책무성(accountability) 논의를 통해 '인간중심주의'를 강조하는 한편, 생물 다양성에 국한되지 않는 다양성 존중과 생태적 고려의 중요성을 문건 전체에 걸쳐 강조하고 있다. 이는 인공지능만큼이나 현재 국제적인 공조의 필요성이 강조되고 있는 기후위기 대응과도 맞물려서 인간이 자연의 주인이며 자신의 이익을 위해 자연을 마음껏 활용해도 된다는 보다 전통적인 '인간중심주의'에 대해 반대한다는 입장을 표명한 것으로 볼 수 있다. 특히 AI 기술이 현재 사용하는 막대한 에너지를 고려할 때 AI 기술의 발전은 당연히 인권과 기본적 자유를 존중해야 하지만 그 과정에서 생태적 환경에 미치는 영향도 함께 고려해야 한다는 점을 강조함으로써 '인간중심주의'를 21세기 기후위기 상황에 맞게 새롭게 이해하려는 노력을 보여주고 있다.

## 2. AI 윤리 영향 평가(AI Ethical Impact Assessment)

유네스코 AI 윤리 권고는 AI 시스템 기술이 사회에 끼치는 영향이 여러 분야에 걸쳐 나타날 수 있고 그 양상 또한 불확실성이 크고 복잡할 수 있는 반면, 사전주의적으로 대응하지 않았을 때 부작용 또한 상당히 클 수 있기에 AI 시스템의 활용 이전에 윤리영향평가를 사용 이후에는 지속적인 모니터링을 통해 보다 바람직한 방식으로 AI 기술이 활용될 수 있도록 하자는 제안을 하고 있다.

5) 본 연구의 중간발표회에서 최경진 교수님은 '중단 스위치' 관련 법규제가 이미 정보화진흥법에 포함되어 있다는 점을 지적하셨다. 이에 대해 필자는 Human on the Loop와 관련된 윤리적 고려가 '중단 스위치' 문제만으로 환원되기에는 훨씬 더 복잡하다는 점을 지적하고 싶다.

권고 초안을 작성한 전문가 작업반에서 이 윤리 영향 평가 제안이 나왔을 때 많은 전문가들은 그 방향성에는 공감했지만 두 가지 이유를 들어 이 제안이 국제적으로 실현되기 쉽지 않을 것이라는 의견을 제시했다. 첫째는 윤리 영향 평가의 원리적 필요성에는 공감하더라도 혁신과 윤리적 고려를 조화해야 하는 각국 정부의 입장에서 자국 기업들의 반발을 불러일으킬 AI 윤리 영향 평가를 제도화하는 것에 적극적이지 않을 것이라는 우려였다. 그래서 많은 경우 각국 정부는 윤리 영향 평가를 강제성 없는 자율규제의 형태로 채택할 가능성이 높고 그렇게 된다면 실질적으로 AI 기술 개발에서 윤리적 고려가 중요한 역할을 하게 될 가능성은 낮아질 수밖에 없을 것이라는 예상이었다. 둘째는 설사 윤리 영향평가의 제도적 필요성에 공감하는 국가의 경우에도 정부의 행정력이나 AI 거버넌스에 대한 전문성이 부족한 회원국의 경우 AI 윤리 평가 제도를 설계하고 이를 운영하는 데 실천적 어려움을 겪을 것이고 그래서 실제로는 시행이 불가능한 제도로 남게 될 것이라는 우려였다.

둘째, 우려에 대해서는 유네스코 권고에서 AI 윤리 정책 행동에 있어서 국제 협력과 연대성(solidarity)를 강조하는 방식으로 대응이 이루어졌다. 즉 상대적으로 제도적 자원이 풍부한 회원국들이 AI 윤리 제도화에 대한 경험을 그렇지 못한 회원국과 공유하고 필요하다면 윤리 영향 평가 제도화 과정에서 직접적인 인적, 물적 지원을 통해 국제적으로 안정적인 AI 윤리 영향 평가 거버넌스를 확립한다는 제안이다. 실제로 EU 국가 내에서 그리고 OECD나 G20 국가 내에서는 각국의 AI 윤리 제도화 경험을 공유하고 이를 바탕으로 자국의 제도화 과정을 진행하는 과정에 도움을 얻기 위해 '관측소(observatory)'라는 형식으로 정보 공유 및 모니터링 메커니즘이 작동하고 있다. 유네스코 권고의 제안은 이를 세계적으로 확대 적용하고 그에 더해 제도적 여력이 충분하지 않는 국가에서 AI 윤리 영향 평가를 비롯한 유네스코가 권고하는 정책 행동을 실현하기 위한 실질적 도움을 주자는 제안으로 이해할 수 있다. 다만 이미 작업반 내 논의를 통해, AI 윤리의 제도화는 나라마다 다른 사회문화적 환경과 역사적 경험이 고려되어야 하기에 특정 국가, 예를 들어 OECD의 경험을 그대로 수입해서 다른 나라에 이식하는 방식으로 AI 윤리의 제도화가 이루어지기는 어렵다는 점을 명심해야 한다. 그러므로 AI 윤리 영향 평가를 비롯한 AI 윤리 거버넌스의 제도화 과정에서 국제 연대와 협력은 각국의 특수성과 AI 윤리의 보편성을 동시에 존중

하는 방식으로 이루어져야 할 것이다. 그리고 이 과정에서 이제 선진국이 된 우리나라의 국제적 기여도 보다 적극적이어야 할 것이다.

AI 윤리 영향 평가가 산업계의 저항으로 실효성이 없을 것이라는 우려는 실제로는 반드시 그렇지만은 않을 수 있다는 점이 이후 각국 정부의 정책 행동에서 드러나면서 보다 복잡한 양상을 보여주고 있다. 예를 들어 AI 윤리 관련 국제 논의에 그다지 적극적이지 않았던 우리나라 정부는 최근 AI 윤리 영향 평가를 시행하기 위해 상당히 적극적으로 연구를 시작하고 있고 이는 명칭은 다르지만 미국이나 EU와 같은 기술 선진국도 마찬가지이다. 이런 상황이 정확히 어떤 이유 때문에 발생하고 있는지에 대해서는 보다 자세한 연구가 필요하지만 일단 두 가지 이유를 추측해 볼 수 있다.

첫째는 앞서 소개한 IEEE에서 이미 가장 기초적인 형태의 AI '윤리 표준'을 제정하고 있다는 사실이다. IEEE는 상당 기간의 의견 수렴을 거쳐서 추상적인 윤리 원칙을 AI 설계 과정에서 어떻게 반영할 지에 대한 연구를 진행했고 이를 Ethically Aligned Design이라는 개념으로 이미 제시한 바 있다. IEEE는 여기에 그치지 않고 이를 AI 설계의 윤리적 적합성을 평가하는 산업 기준 개발로 연결하고 있는데 현재 가장 일반적인 수준의 P7000을 발표했고 이를 보다 세분화한 기술 표준을 지속적으로 개발해서 발표할 예정이다. AI 윤리적 설계에 대한 기술 표준을 제정하려는 노력은 IEEE만 하고 있는 것이 아니다. 국제적으로 AI 윤리 분야의 기술표준을 선점하기 위해 관련 단체들이 경쟁적으로 기술 표준을 제안하거나 연구 중이다. 대부분의 첨단 기술 분야가 그러하듯, AI 분야도 다양한 기술 표준이 어떻게 제정되는 지에 따라 기술적 우위를 점하려는 여러 국가들의 이해관계가 첨예하게 대립할 수밖에 없다. 이런 상황에서 AI '윤리' 기술 표준에 대해서도 유사한 경쟁이 이미 시작되고 있다고 볼 수 있고, 이에 대한 국가적 수준의 대응이 AI 윤리 영향 평가의 제도화로 이어지고 있다고 생각할 수 있다.

이에 더해 AI 기술을 선도하는 기업일수록 AI 윤리 영향 평가를 자신들의 기술개발을 방해하는 제도적 장애물로 판단하고 무조건 이에 대해 저항하려 하기보다는 이를 경쟁적 국제 기술 개발 상황에서 자신들의 이익에 부합하도록 그 제도화 과정에 적극적으로 개입

함으로써 기업 운영의 불확실성을 해소하고 윤리적으로 책임있는 기업이라는 기업 이미지 제고를 동시에 달성하려는 전략을 보여준다. 이런 맥락에서 볼 때 AI의 윤리적 쟁점이 존재한다는 무시할 수 없는 사실과 이에 대한 사회적 대응이 완전히 기업의 자율에 맡겨지지는 않으리라는 현실적 판단이 더해지면 기업 중에서 시장지배력이나 기술적 성숙도가 높은 기업일수록 AI 윤리 영향 평가의 내용을 자신들에게 유리한 방향으로 제정하려는 노력에 동참하게 된다.

예를 들어, 젠더 데이터 편향을 비롯한 AI 훈련 데이터의 편향성이 결국에는 알고리즘 자체에 별다른 윤리적 문제가 없을 때도 AI 결과물에 공정성 논란을 일으킬 수 있다는 점이 사회적으로 부각되면서 훈련 데이터를 여러 윤리적 고려에 맞게 '준비'하는 작업의 중요성이 널리 인식되고 있다. 그런데 이러한 훈련 데이터 준비 과정은 온라인 상에서 입수가능한 모든 데이터를 그냥 모으기만 하는 것은 당연히 아니고 그 데이터 수집 과정에서 '자발적 동의'를 받기만 했다고 모든 문제가 해결되는 것도 아니다. 여러 윤리적 원칙과의 충돌 가능성을 데이터 선별 규칙을 통해 구현해서 결국에는 수집된 데이터 중에서 윤리적으로 사용가능한 데이터만을 사용하거나 필요에 따라서는 특정 데이터 집합을 과선택(oversampling)할 필요도 생기게 된다. 물론 어떤 속성을 갖춘 데이터 집합을 얼마만큼 과선택해야 기존 실제 데이터가 가진 편향성(윤리적으로 불완전한 세계에서 수집되었기에)을 윤리적으로 정당화할 수 있는 방식으로 극복할 수 있을지를 결정하는 일은 쉬운 일이 아니며 그것을 예외 없는 규칙으로 만드는 일 또한 매우 논쟁적인 사안이다. 그러므로 이런 일들을 만족스럽게 해낼 수 있는 기업은 상당한 인적, 물적 자원을 갖춘 거대 기업일 가능성이 높고 이 경우 저개발국가의 AI 신생 기업이나 기술 선진국에서조차 AI 기술 후발주자들은 산업 경쟁력에서 불리할 수밖에 없다. 데이터 편향에 대해 매우 강한 규제가 시행될 경우에 극단적인 경우에는 거대 기업이 윤리적 기준에 맞추어 '잘 준비한' 데이터 집합을 구매해서 자신들이 개발한 AI를 훈련시키는 데 사용할 수 밖에 없는 상황이 올 수도 있다. 물론 이런 상황에서도 윤리적 원칙을 AI 기술에 적용하기 위해서는 어쩔 수없이 치러야 할 대가라고 판단할 수도 있다. 한편 이 대가가 너무 크기에 데이터 준비에 대해 덜 엄격한 기준을 AI 윤리 영향 평가에서 요구하자고 결정할 수도 있다.

중요한 점은 구체적으로 어떤 판단과 결정을 내리는지는 각국 정부가 자신의 제도적, 법적 규제 영역 내에서 심사숙고하여 결정할 사안이라는 것이다. 또한 현재 AI 기술 개발이나 활용 모두 국제적인 수준에서 이루어지고 있고 산업 생태계 역시 특정 국가의 경계에 의해 나누어지지 않는 특징을 보이므로 국제적으로 AI 윤리 영향 평가에 대해 어떤 방향성과 어떤 빙식을 채택하고 있는 지에 대해 보다 면밀하게 분석하고 이에 따라 국내의 AI 윤리 영향 평가의 제도화를 점진적으로 시행하는 것이 바람직해 보인다.

### 3. **투명성**(Transparency), **설명가능성**(Explainability), **책무성**(Accountability)

AI 윤리 영향 평가를 비롯한 AI 윤리 제도화 과정에서 AI 윤리 논의에서 강조되는 윤리 원칙이 AI 시스템 전 주기를 걸쳐 추구되어야 할 가치로 언급된다. 이는 AI 윤리 제도화가 갑작스럽게 등장한 것이 아니라 국제적으로 AI 윤리 원칙에 대한 논의가 선행되었고 이를 구체적인 정책행동으로 구현하기 위해 시도되는 것이라는 맥락을 이해하면 자연스러운 현상이다. 다만 추상적 윤리 원칙으로 제시될 때와 달리 구체적인 정책 행동으로 구현될 때 AI 윤리 원칙의 몇몇 주요 개념에 대해서는 상당히 치밀한 논의가 선행되어야 한다.

우선 이미 그 제도적 실행의 어려움이 지적된 '투명성(transparency)' 윤리 원칙이 있다. AI 시스템이 정확히 그 결과를 어떤 알고리즘을 통해 도출했는지를 투명하게 공개해서 공적 검토가 가능하도록 해야 한다는 이 원칙은 그간 수많은 기술적 대상의 윤리 논의에 등장했던 원칙이다. 예를 들어 담배 회사들은 담배가 인체에 유해하다는 자신들의 내부 연구 결과를 숨긴 채 흡연과 건강 사이의 상관관계는 논쟁적이라는 입장을 오랜 기간 유지했고 이 사실은 여러 내부 고발자의 노력과 미 연방 의회 청문회를 통해서야 비로소 일반 시민에게 '투명하게' 알려졌다.

AI 윤리 관련 논의에서도 비슷한 상황에 대한 걱정이 이 '투명성' 윤리 원칙을 요구하게 만들었다고 볼 수 있다. 만약 미국의 COMPAS 사례처럼 형량이나 가석방 여부를 결정하는 AI 알고리즘이 공정하지 않다는 비판에 직면하게 되거나, 아직 COMPAS 사례의 경우

에는 논쟁이 진행 중이지만 실제로 그 알고리즘을 투명하게 공개했을 때 공정하지 않다는 점이 드러날 가능성이 분명히 존재하기에 AI 알고리즘의 투명성이 강조되는 것이라 이해할 수 있다.

하지만 많은 논자들이 지적하고 있듯이 AI 알고리즘의 코딩 내용, 즉 가장 낮은 기계적 수준의 프로그램 체계를 공개한다는 것은 영업비밀을 모두 공개하라는 것이어서 기존 법체계 상 상당한 충돌을 야기시킬 수 있다. 설사 이 부분이 '공익'을 위한다는 명분으로 넘어갈 수 있다고 하더라도 코딩 표현을 그 알고리즘을 제작하면서 습득한 '암묵지'를 갖추지 못한 사람(일반인이나 다른 분야 전문가)이 보고서 윤리적 문제점이 있는지 여부를 판단하기는 매우 어렵다. 일단 인공지능의 코딩에 사용되는 변수나 명령 체계에는 '공정'이나 '정의', '차별'과 같은 윤리적 가치 개념과 대충이라도 대응될만한 것을 찾기 어렵기 때문이다. 그러므로 수십만 줄이나 수백만 줄에 이르는 코딩 내용을 공개한다고 해서 그것에 대한 윤리 영향 평가를 수행하기는 현실적으로 불가능에 가깝다.

이 사실은 흔히 현재 주로 개발되고 있는 인공지능 기술이 인공신경망에 기초한 '암흑상자' 방식이어서 그렇다는 지적이 종종 제시된다. 하지만 실제로 깊은 학습 알고리즘을 비롯한 기계 학습 알고리즘은 '암흑 상자'라기 보다는 '회색 상자'에 더 가깝다. 즉, 실제로 노드의 업데이트가 진행되는 구체적인 규칙의 수준의 정보만으로는 그 알고리즘을 제작한 공학자조차 진행 상황을 제대로 파악하기 어려운 것이 사실이지만, 보다 높은 '구조적' 수준에서 알고리즘의 개념적 흐름도 수준에서는 알고리즘의 전체적인 얼개나 그 얼개에서 윤리적 고려가 만족되었는지 여부를 판단하는 것이 충분히 가능하기 때문이다. 물론 이 경우에도 어느 정도 수준의 개념적 흐름도 혹은 구조적 설계 내용을 공개하는 것이 영업비밀 유출에 해당되는지에 대한 논쟁은 남아 있다. 또한 추상성이 충분히 높은 개념도만을 제시한 인공지능 개발사가 실제로 코딩 수준에서 그 개념도를 충실하게 준수했는지 아니면 윤리적으로 논란이 될 만한 장치를 끼워 넣었는지를 판단하기도 어렵다. 이런 이유로 '투명성'에 대한 강조는 원칙적으로는 중요하지만 실제 AI 윤리 제도화 과정에서는 어떤 종류의 정보를 어떤 방식으로 제공하도록 요구할 것인지를 역시 다양한 고려 사항의 맞교환을 반

영하여 결정하는 것이 중요하고, AI 윤리 영향 평가를 담당하는 사람이 이해할 수 있는 수준에서 윤리적으로 문제가 없다고 판정된 알고리즘을 구체적인 코딩에서 그에 합당하게 구현하는지 여부는 결국에는 기업의 자율적 행동에 맡겨질 수밖에 없을 가능성이 있다.

투명성에 대한 앞선 논의는 관련된 원칙인 설명가능성(explainability)에도 유사하게 적용된다. 현재 활용되는 기계 학습 AI의 알고리즘이 '반투명'이기에 그 작동방식이나 특정 결과를 내놓는 과정을 사람이 이해하기 쉬운 방식으로 제시하기가 어렵다. 이런 문제를 극복하기 위해 설명 가능한 AI 개발이 적극적으로 추진되고 있으며 이는 AI 윤리의 또다른 윤리 원칙인 AI 시스템의 책무성(accountability)을 높이려는 노력과도 연관된다. AI의 결과물이 어떤 근거에서 어떤 과정을 거쳐 얻어졌는지를 사람이 이해할 수 있도록 제시되어야 그에 대해 그 근거와 과정이 윤리적, 법적으로 적절한 지를 검토할 수 있고 이런 검토 가능성을 염두에 두고 AI 개발자들은 AI 시스템의 책무성을 높이려고 노력할 것이기 때문이다.

이렇게 윤리 원칙적으로 쉽게 수긍할 수 있는 설명가능성과 책무성은 실제 AI 설계 과정이나 활용 과정에 적용하는 과정에서 복잡한 고려가 필요하다. 우선 현재까지 진행되는 설명 가능한 AI의 접근법은 시각 이미지 처리와 같이 특정 판단(즉 이 사진은 고양이 사진이다)을 내리는 과정에서 '인과적으로' 고양이의 일반적 특징(즉, 인간이 이해할 수 있는)을 활용하도록 알고리즘을 짜는 것이다. 다시 말해서 사람이 인지적으로 이해하고 이미지 판독에 활용한다고 알려져 있는 이미지의 거시적 특징을 인공지능으로 하여금 이미지 판단 과정에서 '인과적'으로 사용하도록 강제하는 것이다. 그런데 이런 방식이 이미지 판독 이외의 보다 일반적인 AI 기술에도 적용이 가능한 지는 여전히 미해결의 문제이다. 예를 들어 공공정책에 AI를 활용한다고 할 때 이미 기존 행정학 연구에서 잘 알려진 개념을 반드시 인과적으로 활용하도록 AI에게 요구한다면 AI 기계 학습의 장점을 제대로 살리기가 매우 어려울 것이다. 그 개념의 적절성 자체가 논쟁적일 수 있는데다가 워낙 추상적 개념이어서 AI의 기계 학습 및 결정 과정에서 유의미한 대응물을 미시 변수의 집합으로 정의하는 데 어려움이 있을 것이기 때문이다. 설사 이런 모든 어려움을 기술적으로 극복할 수 있다고

해도 최종적으로 얻어진 설명 가능한 정책 AI는 효율성 측면에서 설명가능성을 만족하지 않은 방식으로 작동하는 AI에 비해 현저하게 떨어질 가능성이 높다. 이는 현재 설명 가능한 AI가 비교적 성공적으로 개발되고 있는 이미지 판독 분야에서도 발생하고 있는 현상이다. 그러므로 설명가능성과 효율성은 일반적으로 음의 상관관계에 있게 되고, 이는 AI의 설명가능성을 무조건 모든 활용 환경에서 절대적으로 요구하기는 어렵다는 점을 시사한다.

결국 우리에게 필요한 것은 설명가능성과 책무성 등의 윤리적 원칙이 효율성의 가치를 압도하는 AI 활용 맥락과 그렇지 않은 경우를 구별하고 각각의 경우에 설명가능성과 책무성을 '어느 정도까지' 요구할 것인지에 대해 사람이 이해할 수 있는 방식의 기준과 이를 기계적으로 구현하는 기술적 요구조건을 정하는 일이다. 이런 구체적인 논의와 이에 따른 제도화가 이루어지지 않고서는 설명가능성과 책무성에 대한 윤리적 요구는 실천적으로 유의미한 결과를 내지 못할 수도 있다.

## 4. 적응적 거버넌스(Adaptive Governance)

이상의 논의를 통해 일반적 시사점을 얻을 수 있다. 첫째는 AI 윤리의 제도화 과정은 결코 국제적으로 대체적 합의가 이루어진 AI 관련 윤리 원칙의 내용을 가져다 그대로 법제화하는 방식으로 얻어질 수 없다는 것이다. 그 주된 이유는 앞서 설명했듯이 추상적 원칙 수준에서는 상당히 광범위한 합의가 국제적으로 형성되어 있지만 그것을 제도화, 규제 법률로 만드는 과정에서는 상당한 추가 논의가 필요하고 이는 각각의 제도화, 규제화가 이루어지는 국소적 맥락을 고려하여 이루어져야 하기 때문이다. 그리고 이러한 추가 논의 과정은 국제적 논의 흐름을 잘 모니터링하면서 수동적으로 그것을 반영하려는 소극적 태도가 아니라 그 논의 과정에 적극적으로 참여하여 바람직한 AI 윤리 제도화를 이루려는 보다 적극적 노력이 필요하다는 점 또한 중요하다.

둘째 이유는 우리가 대체적으로 합의한 윤리 원칙들은 구체적인 상황, 대개는 여러 윤리 원칙이 개입되는 복잡한 상황에서는 서로 충돌하는 경우가 많으므로 이들 사이의 합당한

방식의 맞교환을 고안해내야 하고 이를 위해서는 AI 윤리 관련 당사자를 포함한 사회적 논의와 조정 과정이 필요하다는 것이다. 앞서 지적했듯이 투명성, 설명가능성, 책무성 모두 이런 윤리 원칙에 해당되며 최근 사회적으로 관심이 집중되고 있는 데이터 편향이나 AI의 공정성 문제 역시 구체적인 정책 행동이나 제도로 번역하는 과정에서 상당한 추가 논의가 필요한 부분이다.

이런 점을 고려해서 유네스코 AI 윤리 권고는 AI 윤리 거버넌스가 미리 세세한 부분까지 완결된 형태로 규정을 제시하는 방식의 제도화가 아니라 충분한 사회적 공감대와 국제적 합의가 도출된 내용과 영역부터 차례대로 제도화를 시행하고 그 제도화도 이후 AI 기술의 발전이나 사회적 인식 변화 등을 반영하여 수정될 수 있도록 '유연한' 방식으로 이루어져야 한다는 점을 강조한다. 유네스코 AI 윤리 권고는 이를 적응적 거버넌스(adaptive governance)라는 개념으로 설명하고 있는데, AI 기술 개발의 불확실성과 각국의 제도적 대응의 불확실성을 고려할 때, AI 윤리의 제도화 과정에서 우리나라에서도 채택이 필요한 접근 방식이라고 판단된다.

## 5. AI 리터러시와 AI 윤리 교육

앞서 지적했듯이 AI 윤리(Ethics)의 쟁점은 많은 경우에 논쟁적이다. 그 이유는 우리가 소중하게 여기는 윤리 원칙들에 대해 사람들이 대체적으로는 합의하지만 구체적으로 제도화하는 단계에서는 합의하지 않는 경우가 많기 때문이다. 물론 법률적으로 보장받아야 하는 기본권에 대해 우리 사회는 이미 헌법적 가치로 합의하고 있다. 하지만 중요한 점은 이런 기본 윤리 원칙을 인공지능과 관련하여 준수하고자 할 때 많은 경우 서로 다른 윤리 원칙 사이에 충돌이 발생할 수 있다는 점이다. 그래서 유네스코 AI 윤리 권고를 비롯한 많은 국제 인공지능 윤리 논의에서는 이런 '맞교환(tradeoff)' 상황을 어떻게 해결해 나갈 것인지에 대해 사회적 논의와 현명한 결정이 필요하다고 강조하고 있다.

그러므로 우리의 AI 윤리 교육 역시 AI의 설계와 활용 과정에서 이런 문제가 발생할

수 있고 이런 문제는 이렇게 해결하면 된다는 식의 '정답'을 제시하는 방식이 아니라 인공지능의 기술적 특징과 활용 방식에 따라 왜 윤리적 문제가 발생하는 지에 대한 이해에 바탕하여 우리 사회에서 바람직한 해결책을 찾아 나가는 도덕적 사고 및 합의 도출 '역량' 교육이 되어야 할 것이다. 이 과정에서 AI와 교육에서 최근 강조되고 있는 AI 리터러시 교육과의 시너지 효과에도 주목할 필요가 있다. 결국 AI 윤리 교육은 AI 기술 자체에 대한 정확한 이해와 그 기술이 사회문화의 여러 측면과 맺는 다양한 상호작용의 성격을 올바르게 분석해 내는 역량에 기초해야 하기 때문이다. 이 부분이 AI 윤리의 성공적인 제도화와 정책 행동의 효율적 시행을 위해 결정적으로 중요하다는 점은 유네스코의 AI 윤리 권고에서 잘 지적하고 있다.

그리고 AI 리터러시 교육과 마찬가지로 AI 윤리 교육 역시 전 국민을 대상으로 하는 '보편 시민 교육'에서 AI 코딩 교육보다는 훨씬 더 핵심적인 위치를 차지하는 것이 마땅하다. 앞서 지적했듯이 AI 인터페이스 발전 방향에 따라 우리 대부분은 AI 프로그램을 하지 않고도 살 수 있을지 모르지만 AI에 기반한 자동화된 결정으로 운영되는 사회에서 살아갈 것은 거의 확실해 보이기 때문이다. 이런 점을 고려할 때 AI 윤리의 제도화 과정은 정부와 AI 개발자 및 운영자 사이의 정책적 조율로만 진행될 사안이기 보다는 대다수가 AI 사용자로 참여할 일반 시민이 포괄적 의미의 AI 리터러시를 확보하기 위한 노력과 동시에 진행되어야 한다. 이런 점에서 AI 코딩 교육에만 집중되고 있는 현재 우리나라의 AI 교육에 대한 관심과 논의는 AI 윤리의 관점에서 적극적으로 재검토되어야 한다.

## 제3절 글로벌 AI 윤리 논의의 시사점

이상의 논의를 통해 우리는 AI처럼 빠르게 발전하며 사회적 영향력을 확대하고 있는 기술에 대한 넓은 의미에서의 윤리적 고려와 그 고려를 다양한 방식으로 제도화하려는 노력은 국제적 공감대를 얻고 있음을 알 수 있다. 또한 이러한 공감대에 기초하여 국가별로 이루어지고 있는 구체적 수준의 AI 윤리 거버넌스는 개별 국가의 역사적, 사회적, 문화적

상황에 대한 치밀한 분석과 연구에 기초하여 이루어져야 하며 '경쟁적으로' 먼저 법제도화를 달성하겠다는 식의 생각은 정당하지 않다는 점도 알게 되었다.

이런 상황에 대해 현장에서 AI 기술을 개발하는 공학자나 관련 사업을 추진하는 기업가들은 껄끄러운 '규제'가 또 하나 생긴다고 불편해할 수 있다. 그들 입장에서 '규제'를 곧 '혁신 저하'로 동일시하는 경향이 있기 때문일 것이다. 하지만 이는 기술혁신의 역사에서 사실이 아닌 생각이다. 실제로 1970년대에 자동차 배기가스 규제가 도입되려 할 때 미국의 대형 자동차 회사들은 이 규제가 산업 생산력을 저하시키고 소비자의 권익을 해칠 것이라고 극렬하게 반대했지만 실제로 이 규제는 보다 친환경적인 내연기관을 개발하는 기술혁신과 배기가스 저감장치 등의 파생 기술 개발로 이어졌다. 어떤 기준으로 평가하더라도 70년대의 배기가스 규제가 기술혁신을 저하했다든지 소비자 권익을 해쳤다고 볼 근거는 없다. 이처럼 적절한 방식으로 합리적으로 운용된 규제는 기업의 산업 환경을 바꿈으로써 기업의 기술혁신 의욕을 오히려 더 고취할 수 있으며 사회적으로 유용한 방향으로 기술혁신을 유도할 수도 있다. 현재 한창 기술 개발이 이루어지고 있기에 앞으로서의 혁신 잠재력이 큰 AI 기술에서 현명한 규제가 앞서 강조한 '적응적' 방식으로 이루어진다면 70년대 배기가스 규제와 마찬가지로 기술혁신과 사회적 공익 실현을 동시에 달성할 수 있을 것이다.

그러므로 AI 윤리의 법제도화 과정에서 핵심적인 사안은 '적응적' 거버넌스를 구체적으로 어떻게 실현할 지가 될 것이다. 여기에는 몇 가지 고려사항이 있다. 첫째는 처음부터 강한 법적 규제를 도입하기 보다는 규제를 담당할 정부와 규제의 대상인 동시에 자율 규제의 주체인 기업의 '역량 강화'가 선취되어야 한다. AI의 다양한 윤리적 쟁점의 중요성을 깊이 이해하고 이를 사회적으로 풀어낼 수 있는 전문 역량을 갖춘 인력이 정부와 기업에 배치되어야 한다. 원론적 수준에서 윤리적 고려의 중요성을 공언하는 방식이 아니라 IEEE의 시도처럼 윤리적 고려나 원칙을 기술 개발 과정에서 적극적으로 고려하는 동시에 앞서 소개한 윤리 영향 평가와 같은 지속적인 모니터링과 피드백 반영을 수행할 수 있어야 하기 때문이다. 이에 더해 국제 논의에 참여하기 위해서는 대표자로 참여하는 정부와 민간기업

의 담당자들이 국제 거버넌스에 적극적으로 참여할 수 있는 역량 강화가 반드시 필요하다.

둘째로 AI 윤리 및 법제화 과정에서 AI를 단일한 기술적 대상으로 생각하기 보다는 다양한 요소 기술의 집합체인 시스템으로 이해하는 것이 중요하다. 이렇게 이해되어야, 가령 AI 기술 활용에서 핵심적인 데이터 수집, 준비, 활용, 폐기 등과 관련된 데이터 거버넌스 논의도 자연스럽게 AI 거버넌스와 함께 논의될 수 있다. 또한 AI가 성공적으로 개발되고 생산적으로 활용되기 위해 필요한 사회적 요인이나 다른 기술 시스템과의 상호작용에 대해서도 종합적인 시각으로 함께 그 윤리적 쟁점을 논의할 수 있게 될 것이다.

이는 유네스코의 AI 윤리 권고가 지속적으로 강조하고 있는 입장이기도 하다. AI를 시스템 기술로 이해함으로써 우리는 AI 기술의 광범위한 파급 효과에 대해 종합적으로 파악할 수 있으며, 전체 범위에 관련된 윤리적 쟁점에 대해 통합적인 시각으로 바라보고 대응책을 마련할 수 있다는 것이다.[6] 예를 들어 AI 윤리적 쟁점에 대해 '적응적' 거버넌스를 실천한다고 할 때 그 과정에서 민주주의적 가치를 어떻게 고려하고 반영할 것인지에 대한 문제는 추가 연구가 필요한 중요한 주제이다. AI의 자동화된 결정이 의도적이든 비의도적이든 정치적 의견 형성에 큰 영향을 끼치고 있는 상황에서 그에 대해 '적응적'으로 대응한다는 것이 정확히 무엇을 의미하는지조차 논쟁적이기 때문이다. 3장에서 이 주제에 대한 해답의 실마리가 제시되겠지만 이 주제를 포함한 AI '적응적' 거버넌스의 여러 쟁점에 대해서는 지속적인 연구와 그에 근거한 정책 실행이 이루어져야 할 것이다.

셋째 AI 기술과 그 활용의 특징을 올바로 반영하는 AI 윤리 법제도화 논의가 중요하다. 예를 들어 AI 윤리 영향 평가의 구체적인 안을 만들 때는 AI의 개발 및 활용에 참여하는 다양한 주체를 어떻게 참여시킬 것인지 그 평가를 누가 어떤 방식으로 언제 시행할 것인지에 대한 구체적인 논의를 거쳐야 한다. 그리고 이 과정에서 현재 AI 기술의 특징을 반영하는 방식으로 AI 윤리 영향 평가의 내용이 만들어져야 하고, 이에 더해 AI 기술의 지속적인 발전을 반영하여 이 평가의 형식과 내용은 지속적으로 업데이트되어야 한다. 다시 말하자

6) 이런 관점에서 볼 때 현재 개별 AI 제품에 집중되고 있는 AI 윤리 논의를 사회적 영향력이 훨씬 크다고 평가할 수 있는 AI 기반 플랫폼으로 확대할 필요가 있다.

면 AI 윤리 영향 평가의 운용 과정 역시 '적응적' 거버넌스 방식으로 이루어져야 한다는 것이다.

여기에 더해 앞서 지적하였듯이 현재 가장 많이 활용되고 있는 기계학습 기반 AI의 반투명성을 고려하여 효율성 있고 실행 가능한 법제도화가 마련되어야 한다. 중요한 점은 이런 '고려'가 AI 관련 윤리적 원칙의 내용을 약화하거나 훼손시키는 방향으로 추진되어서는 안 된다는 것이다. 그보다는 윤리적 원칙의 내용을 '현실적으로' 보다 더 잘 실현하기 위해서라도 기업이 사회적으로 공감대가 확보된 윤리적 고려를 어떻게 달성할 수 있을지에 대한 구체적 설명 없이 추상적 원칙만을 제시하고 준수를 요구하는 방식은 실효성이 없음을 지적하는 것이다. 그러므로 현재 정부가 추진 중인 AI 윤리 준수 점검리스트 작성 작업 역시 공학자와 산업계의 의견을 충분히 수렴하여 이루어져야 하고 서둘러 성과내기식으로 추진되어서는 안된다.[7)]

이 점은 윤리적 쟁점에 대한 이론적 연구와 이해당사자들의 의견을 청취하고 반영하려는 노력을 충분히 오랜 기간 수행하지 않고 성급하게 세세한 법제도화를 시도할 때 불필요한 사회적 비용이 발생할 가능성이 높다는 사실과도 연관된다. '적응적' 거버넌스의 본질적 특징은 거버넌스의 형식과 내용을 그 적용 결과에 대한 지속적인 모니터링을 통해 새로운 조건에 맞추어 수정해 나간다는 것이다. 그런데 법제도는 한번 만들어 놓으면 나중에 바꾸기가 쉽지 않다. 특히 처음에 기술적, 사회적 불확실성이 큰 상황에서 너무 세세하게 AI 윤리적 고려를 법제화할 경우 얼마 지나지 않아 AI의 기술적 특징이나 사회적 사용 현실과 괴리가 생겨 불필요한 사회적 비용이 발생할 가능성이 높다.

그러므로 이런 점을 고려할 때 AI 윤리의 법제도화는 국제적으로 상당한 공감대가 형성된 AI 윤리 원칙을 중심으로 추진하되, 충분한 기간을 두고 관련 쟁점에 대한 연구와 이해당사자와의 숙의 과정을 거쳐 실효성 있는 형태로 마련되어야 하며, 기술의 미래 발전 방

7) 예를 들어 AI 개발자에게 '설명가능성'을 준수하도록 요구하는 것은 AI 윤리적 측면에서 실효성을 갖기 어렵다. '설명가능성'이 구체적인 제품의 맥락에서 어떤 것을 만족시킬 얻어질 수 있는 개념인지에 대해 사례 등을 들어 설명하지 않으면 개발자들은 자신들이 이해하는 방식으로 설명가능성이 만족되었다고 판단해 버릴 수 있기 때문이다.

향과 AI 사용자의 문화적 대응에 대한 불확실성을 고려하여 '적응적' 거버넌스 형태로 추진되어야 한다. 이 과정에서 유네스코의 AI 윤리 권고의 내용 등 국제 사회의 AI 윤리 논의를 참고하고 국내에서 AI 윤리에 대한 공감대 및 리터러시 교육과 정부 및 기업의 대응 역량을 높이려는 노력이 동시에 추진되어야 할 것이다.

# 제2장
# AI 윤리 거버넌스의 사회적 기반

**이 호 영** (정보통신정책연구원 디지털경제사회연구본부장)

# 제2장 AI 윤리 거버넌스의 사회적 기반

## 제1절 서 론

기술에 대한 거버넌스 논의는 완전히 새로운 것이 아니지만 오늘날 인공지능(Artificial Intelligence: AI) 거버넌스에 관한 전 세계적 관심은 이례적이라고 할 만큼 뜨겁고 진지하다. 또한 특정한 기술에 대한 거버넌스를 위해서 각국이 앞다투어 물리적 기구를 설치하거나 입법을 논의하게 되는 것 역시 흔한 일은 아니다. 인공지능 거버넌스는 그 자체가 학술적 연구의 주제이자 정책 결정의 중요한 한 축으로 등장하는 중이다. 인공지능은 컴퓨터가 할 수 있는 일들을 사회 전 영역에 걸쳐 확장하였고 그 와해성으로 인해 사회에 주는 영향력은 그만큼 크고 광범위하게 받아들여지고 있다. 자율주행자동차에서 사고 시 유책 문제, 빅데이터의 차별적인 영향을 다루는 현재 법적 프레임워크의 한계, 혹은 알고리듬의 유해성을 사전에 예방하는 방법, 사회복지나 법 적용을 자동화하는 사회적 정의의 문제, 온라인에서의 미디어 편식 소비 등, 많은 영역들이 새로운 인공지능 윤리의 사회적 도전과제에 포함되었다. 동시에 이런 문제를 해결하기 위한 여정에서는 거의 모든 학문의 영역이 동원되는, 다학제간 접근이 기본값이 되었다(Fjeld et al., 2020).

이 글에서 다루려고 하는 인공지능 윤리 거버넌스 논의는 인공지능 거버넌스의 층위(layer)들 중 하나라고 할 수 있으며 인공지능의 위험(risk) 관리 논의와 밀접한 연관을 갖고 있다. 물론 문제는 그리 간단하지 않다. 인간보다 체스나 바둑을 더 잘 두는 지능형 기계라는 단순한 개념으로부터 환자를 진단하고 금융거래를 관리하며 채용과 치안, 복지 서비스 수혜자 선별, 판결에 이르기까지 인공지능의 영향력 미치는 분야가 무궁무진하다

는 사실이 알려짐에 따라 “사회적으로 선한(socially beneficial)” 인공지능의 설계에 대한 관심도 함께 높아졌다. 인공지능이 가지고 있는 복잡한 성격과 여전히 모호한 정의에도 불구하고 사회적으로 선한 인공지능이라는 문제의식은 기업 차원에서도 널리 받아들여지게 되었다. 요컨대 인공지능 윤리는 잠재적, 현재적 위험을 완화하고 사회적으로 선한 인공지능을 설계하는 문제라고 할 수 있다. 그런데 이 때 선(善)은 물론 위험 역시도 사회적으로 정의될 수밖에 없기에 인공지능 윤리의 프레임워크는 사회적으로 결정되어야 하는 문제가 된다.

OECD가 고잉 디지털(Going Digital) 프로젝트에서 인공지능의 사회적 영향력에 대한 전방위적 논의를 시작하고 나서 인공지능 시스템이 지닌 효율성과 혜택, 그리고 복잡성과 리스크에 대한 관심은 구체화되기 시작했다(OECD, 2019). 인공지능의 와해성이 커질수록 이 기술은 단순히 생산성뿐만 아니라 프라이버시, 평등, 차별, 노동의 미래, 그리고 감시 기술을 사용한 민주주의에 대한 침해 등 사회적 이슈들과 광범위하게 연결되어 있다는 연구 결과들도 많이 나왔다. 인공지능이 대변하는 기술 진보는 플랫폼 경제의 도래와 함께 비즈니스 부분을 벗어나 그 사회적 영향을 극대화했다고 볼 수 있다. 여기서 주목할 점은 인공지능이 태생적으로 그런 문제를 안고 있는 것이 아니라 일련의 사회적 선택의 결과로서 그런 문제들이 발생했다는 사실이다. 인공지능 윤리 거버넌스에 대한 이 글은 이런 관점을 견지하려고 한다.

이 글은 인공지능 윤리 거버넌스의 주요 쟁점을 사회적 관점에서 다루고 있다. 따라서 인공지능의 설계자의 의도보다는 인공지능의 사회적 영향에 더 많은 주의를 기울이고자 한다. 그리고 그 영향은 상황과 맥락, 시기와 받아들이는 주체에 따라 다른 함의를 가진다는 점에도 주목한다. 인공지능은 많은 영역에서 이용자 개인은 물론 집합적 경제사회적 후생을 증대시키고 인권을 고양할 잠재력을 갖고 있지만, 동시에 잘못 사용될 경우, 혹은 유해한 방식으로 이용될 경우 사회적 위험을 낳게 되기 때문이다.

인공지능이 점점 인간을 보조하는 범용 기술로 기대했던 것보다 빠르게 우리 사회에 스며들면서 인공지능 윤리의 문제는 비단 엔지니어나 인공지능 시스템을 설계하는 공학자

의 문제로 국한되지 않게 되었다. 세계 최대 규모의 인공지능 컨퍼런스라고 할 수 있는 NeurIPS의 조직위원회는 2020년 초, 컨퍼런스에 신청을 할 때 윤리적 측면과 미래 사회에 미칠 결과 등을 포함하는, 인공지능 연구의 광범위한 영향에 대해 진술문을 작성할 것을 의무화했다(Prunkl, Ashurst, Anderljung. 2021) 그 여파는 결코 작지 않았다. 초반에는 과학자들의 불평이 쏟아졌지만 억지로라도 이 문제를 고민하게 된 과학 커뮤니티는 인공지능의 윤리적, 사회적 측면에 대한 다양한 사고를 전개하기 시작했다. 이제 인공지능을 적용하려는 기업이나 정부는 당장의 근시안적 문제 시정에 국한되지 않고 민주주의와 인권 같은 기본 가치는 물론 중장기적인 인류의 번영과 지속가능한 미래를 동시에 고려하지 않을 수 없다. 또한 인공지능의 적용이 사회문화적 콘텍스트 밖에서, 혹은 사회적 진공상태에서 이루어지지 않는다는 점도 염두에 두어야 한다. 이것이 시스템으로서의 인공지능을 이야기하게 된 근거이기도 하다.

## 제2절 AI 시스템의 사회적 영향

### 1. AI 시스템의 적용 분야

우리 사회는 환자의 진단, 금융거래, 채용 심사처럼 점점 더 복잡하고 매우 위험한 과정을 인공지능 시스템에 위임해가고 있다. 예를 들어 미국 소재의 헤지펀드인 브리지워터에서는 알고리듬이 단순히 투자를 결정할 뿐 아니라 임원처럼 행동하고 바람직한 영업 전략을 권고하는 일을 한다(람게, 2020). 이로 인해서 데이터의 이질적 효과(boyd, Levy, and Marwick, 2014), 불평등을 유발하는 알고리듬의 통제(유뱅크스, 2018), 자율주행자동차의 사고 시 책임 소재, 프로파일 기반의 추천 서비스가 초래하는 여론 양극화, 사회복지나 법 적용을 자동화하는 문제, 맞춤형 광고에 의한 프라이버시 침해, 안면인식을 이용한 정부의 시민 통제에 이르기까지 과거에는 존재하지 않았던 새로운 질문들이 쏟아지고 있다. 전문가들은 이러한 상황이 예상치 않았던 방식으로 전개되거나 혹은 잠재적으로 유해한 방식으로 잘못 사용되지 않을지 우려를 표하고 있다.

디지털 불평등이 온라인 상의 정보 접근권의 차이에 머무르지 않고 오프라인의 불평등을 심화시키거나 가속화한다는 논의 역시 이런 맥락에서 이해될 수 있다. 단순한 디지털 기술이 주도하던 초기 정보화 시대에 정보격차의 패러다임은 기술과 기기에 대한 접근성이나 디바이스의 저렴한 가격의 문제에 불과했지만 시간의 흐름에 따라 사용과 혜택에 있어서의 불평등, 나아가 디지털 기회의 불균등한 분배 문제로 진화하였다. 인공지능 시스템은 이에 더해 사람이 이해할 수 없고 통제할 수 없는 방식의 되먹임(feedback) 프로세스의 결과로 불평등이 점점 더 확대되는 상황을 보여주고 있다. 이 문제를 논의하는 과정에서 이 프로세스를 설계하고 관리하고 감시하는 인간의 역량이 중요한 키워드가 되었고 인간이 디지털 기술의 설계와 전개에서 중심이 되어야 하며 기계와 관련한 인간의 행위는 모든 상호작용 과정에서 존중받아야 한다는 컨센서스가 이루어졌다. 그러나 단순히 개개인의 역량을 키움으로써 해결되지 않는 문제가 인공지능 시스템에 내재한다는 인식 역시 널리 받아들여지고 있다.

**<표 1> 인공지능 시스템의 적용 분야**

| | |
|---|---|
| 지능형 비서 | ▪ 사용자의 신호나 주변 환경을 인지하여 사용자가 요구하거나 실생활에서 필요한 업무(스케줄 관리 등)를 파악해 그에 맞는 대응책을 지원하는 기술 |
| 의료 진단 | ▪ 대량의 생체 데이터를 기반으로 의료 정보를 학습하고, 활용하여 환자의 상태를 추정하거나 치료를 보조하는 기술 |
| 교육 | ▪ 문제의 패턴을 인식하여 개별 학습자에게 풀이 방법을 알려주는 기술<br>▪ 채점, 평가 등 교사의 노동을 자동화하는 기술 |
| 법률 서비스 지원 | ▪ 판례에 대한 기계학습을 통해 법률 관련 지원을 하는 기술<br>▪ 법률 소비자와 변호사 사이의 매칭 서비스 |
| 금융 서비스 | ▪ 투자 분석, 예측, 대리 집행, 금융 투자 위험 예측 및 관리와 거래 이상 징후 감지 및 소비자에 대한 경고 |
| 감시시스템 | ▪ 안면인식 기술 등을 이용하여 영상 정보를 수집하고 패턴인식 및 기계학습을 바탕으로 객체 인식, 행동 인식 등을 수행하여 영상 내의 상황을 판단하고 자동으로 사용자에게 위험 상황을 알릴 수 있는 시스템 |
| 기사 작성 | ▪ 자연어 처리 등을 이용하여 신문기사를 자동으로 생성하는 로봇 저널리즘<br>▪ 반복되는 패턴 인식을 통해 주기별 기사 작성 |
| 추천 시스템 | ▪ 고객의 구매 이력이나 패턴을 분석하여 고객의 취향이나 처한 상황에 맞게 제품을 추천해주는 개인 맞춤형 추천 알고리즘의 적용 사례가 증가 |
| 지능형 로봇 | ▪ 외부환경을 인식하고, 스스로 상황을 판단하여, 자율적으로 동작하는 로봇<br>▪ 공장의 자동화 로봇 뿐 아니라, 돌봄로봇, 소셜로봇 등으로 확장 |

## 2. AI 시스템의 사회경제적 영향

인공지능 시스템의 사회적 영향을 이야기할 때 대개는 선용과 악용, 통제와 진흥의 이분법에 매몰되기 쉽다. 여기서는 좀 더 사회체계적(societal)인 변화를 살펴보고자 한다. 사실 범용기술(General Purpose Technology, GPT)로서, 또한 와해적 기술(disruptive technology)로서 인공지능이 중요한 것은 이것이 사회경제적 파급력이 있고 시간의 흐름에 따라 진화 혹은 개선되며 보완적인 혁신을 추동하기 때문일 것이다.

인공지능의 영향력에 대해서는 초기부터 수많은 논의가 있었지만 최근에는 강인공지능보다 기계학습 기반, 초대량 데이터 기반의 약인공지능이 주류화되면서 그 사회경제적 파급효과에 대한 논의가 더 많이 언급되고 있다. 그런데 이 과정은 인공지능이 점점 개별 인간으로부터 의사결정의 자율성을 획득해가는 과정이기도 하다. 인공지능이 다양한 영역의 의사결정에 도입될수록 인간이 갖는 자신의 결정에 대한 통제력은 점차 감소하게 된다. 다르게 말하자면 인공지능이 스스로 규칙을 만들어가는 기계학습의 발전에 따라 인간은 이 과정을 통제할 수 없게 되고 따라서 결과에 대해 책임질 필요도 없게 되는 상황이 만들어지게 된 것이다.

문제는 기계학습 인공지능이 스스로 학습해서 만든 규칙이 인공지능 시스템에만 영향을 미치는 게 아니고 인간이 살고 있는 사회 전반에 광범위한 영향을 미친다는 데 있다. 기본적으로 상품을 만든 제조사는 해당 상품의 결함이나 오작동에 대한 법적, 도의적 책임을 지게 되지만 정해진 모델과 데이터에 의한 것이 아닌, 대규모 기계학습을 통해 내린 인공지능의 결정의 예측 불가능성은 해당 서비스를 제공한 회사가 그 결과에 책임을 질 수 없게 만든다. 또한 인공지능 기반의 서비스에는 하나의 회사 또는 기관만 참여하는 것이 아니므로 대체 그중 누가 리스크를 책임질 것인가 하는 문제가 당연히 제기된다. 또한 지나친 법적 책임을 묻게 되면 혁신이 저해되고 산업 경쟁력을 잃게 된다는 지적 역시 경제성장을 추진해야 하는 정부로서는 무시할 수 없기에 균형 잡힌 거버넌스 구조를 만드는 일은 어려움에 직면하게 된다. 사실 예기치 않았던 윤리적 거버넌스 문제의 상당 부분은 인공지능을 사회가 받아들이는 속도와 연관이 있다.

### 1) 경제적 영향

인공지능의 경제적 영향은 생산성에 대한 기여와 고용 기회의 박탈, 그리고 저숙련자의 상대적 임금하락으로 인한 소득 불평등 심화라는 측면에서 다루어져 왔다(이상엽, 이동규, 2000). 예를 들어 경제학자 다론 아제모을루(Acemoglu)는 인공지능의 문제를 한편으로 자동화로 인한 노동의 대체와 연결시켰고 다른 한편으로 생산성을 향상시키는 기술로서 인공지능은 자동화를 통한 번영과 그 분배의 문제와도 연관되어 있다고 봤다(Acemoglu, 2021). 예컨대 산업계에서 인공지능은 오늘날 경제성장을 위한 필수적 기술로서 도입되고 있지만 '숙련편향적(skill-biased)'이라는 의미의 양극화와 밀접히 연관된다.

그간 경제학 분야에는 특히 정보통신기술(ICT)로 대변되는 기술 발전이 생각처럼 생산성에 기여하지 못했다는 연구결과가 많이 나와 있다. 이는 부분적으로 무어의 법칙으로 인해 ICT 디바이스나 소프트웨어 가격이 하락하는 이른바 생산성의 역설과 관련이 있다. 이는 기술 혁신의 결과물인 특허나 R&D의 결과물이 무형자산(intangible capital)에 속하기 때문에 벌어지는 현상이다. 그런데 브린욜프슨 등은 이제 우리는 제이 커브의 바닥에 와있으며 이륙을 준비 중이라는 말을 하고 있다(Brynjolfsson, Rock, and Syverson, 2021). 실제로 지난 몇 년간 생산되는 데이터가 폭증하고 있으며 이것이 생산성을 끌어올리는 중요한 자원이 되어줄 것이라고 보는 것이다.

고용 기회에 대해서는 인공지능이 생산의 자동화를 가속화하여 인간의 노동이 점점 더 자본으로 대체되고 업무를 수행하는 방식에까지 영향을 줌으로써 고용이 줄어든다는 식으로 논의가 이루어져 왔다. 그런데 인공지능이 반드시 노동을 대체하는 것만은 아니다. 인공지능은 과거에는 없었던 새로운 업무를 창출하는 효과, 나아가 기존 업무와 새로운 업무의 생산성을 증가시키는 효과(productivity effect)를 통해 노동의 양적 변화는 물론 질적인 변화를 동시에 가져올 수 있기 때문이다. 그런데 이런 생산성 증가 효과는 단순히 인간의 노동을 대체하기만 하는 그저 그런 자동화(so so automation)가 아닌 훨씬 더 혁신적인 자동화로 인해 가능하며 이는 대개 즉각적인 투입의 결과로 나타나지 않고 있다.

인공지능의 도입에 따른 소득 양극화 문제는 이 기술에 대한 기업의 수요가 주로 고숙련 노동에 집중되며 이 고숙련노동자는 빠른 속도로 노동시장에 공급될 수 없기에 불가피하게 임금 격차를 벌리게 된다는 것에 주로 기인한다. 동시에 인공지능을 도입하는 자동화 흐름은 반복적인 노동을 빠르게 노동시장에서 퇴출한다. 이 두 개의 흐름이 맞물리면서 특히 미국을 중심으로 불평등이 크게 확대되어왔다.

요컨대 인공지능의 경제적 영향의 문제는 지체(mismatch)의 문제와 격차의 문제로 정리해볼 수 있다. 이 중 지체는 시간의 축에서, 격차는 공간의 차원에서 논의할 수 있고 이 중에서 후자가 주로 윤리의 문제라고 볼 수 있다.

### 2) 사회적 영향

규제 메커니즘과 거버넌스 문제가 본격적으로 대두된 것은 인공지능 시스템의 사회적 영향에 대한 여러 가지 논의가 사회적으로 확산된 이후의 일이다. 투명성이나 공정성 같은 윤리적 이슈는 물론, 불평등부터 양극화, 데이터 편향으로 인한 사회적 불공정의 악화에 이르기까지 인공지능 시스템의 사회적 영향에 관한 문헌들이 속속 발간되면서 인공지능에 직간접적으로 관여하지 않는 사람들도 스스로 이러한 영향력으로부터 자유롭지 않다는 사실을 깨닫게 된 것이다.

특히 공적 영역에서 알고리즘 기반의 자동화된 의사결정과 관련된 논의는 가장 많은 주목을 끌고 있다(이호영, 2021). 기계학습 알고리듬이 사용하는 데이터 자체가 이미 존재하는 사회의 편향이나 불평등한 현실을 반영하고 있기 때문에 그 데이터를 학습한 결과가 편향된 것을 어떻게 처리할 것이냐 하는 문제나, 민감한 개인정보가 담긴 데이터를 이용하여 개인을 분류하는 것이 이미 불리한 위치에 있는 개인, 혹은 세그먼트에 부정적인 영향을 줄 수 있다는 점이 지적되곤 한다. 결과적으로 취약한 사회경제적 지위를 가지고 살아가는 것에 더해 알고리듬 기반의 자동화된 의사결정 환경에 처하게 됐을 때 더 많은 리스크에 노출되고 다른 사람이 누리는 혜택을 못 누리게 될 가능성이 있다는 것이다.

나아가 데이터에서 민감한 개인정보를 제거한다고 하더라도 기계학습의 테크닉은 확률적 추론에 의해 민감한 속성에 대한 변수를 대리변수로 사용할 수 있다. 결국 이 문제는 인공지능 거버넌스에서 자주 등장하는 공정성에 대한 논의로 이어지게 된다(Taeilagh, 2021). 실제로 인공지능의 사회적 영향에 관한 많은 논의는 공평과 효율성 사이의 트레이드오프를 어떻게 처리할 것인가, 기업이 추구하는 이윤을 위한 합리성 외에 사회적으로 고려할 가치합리성이 존재하는가, 그리고 사회 안에 이미 존재하는 윤리나 가치와 인공지능 시스템의 결과값이 충돌할 경우 이에 대한 결정을 누구의 손에 맡길 것인가 하는 문제로 이어지게 된다.

### 3) 루프 속의 사회(Society-in-the-loop:SITL)

인공지능 거버넌스의 형태를 근본적으로 규정하는 것은 단순히 인공지능 시스템이 어떻게 작동하느냐 하는 것이 아니라 이 시스템이 어떻게 이해되고 상상되는가 하는 데 달려있다.

인공지능 시스템이 협의의, 잘 정의된 기능이 아니라 그것보다 광의의 사회적 함의를 갖는 기능을 수행할 때 우리는 루프 속의 인간을 넘어서는 루프 속의 사회를 생각하지 않을 수 없다. 예를 들어 정치적 신념이나 시민들의 선호체계를 통제하는, 혹은 전체 경제에서 자원과 노동력의 배분을 결정하는 인공지능 알고리듬을 생각해보자. '루프 속의 인간(Human-in-the-loop)'이 상대적으로 협소한 범위에 영향을 미치는 인공지능 시스템의 최적화에서 개인이나 집단의 판단을 돕는 것이라면 '루프 속의 사회Society-in-the-loop'는 광범위한 함의를 갖는 사회적 결과에 대한 알고리듬 거버넌스 속에 전체 사회의 가치를 내재화하는 것이다(Rahwan, 2017).

알고리듬 거버넌스를 개개인의 편익이나 소비자주의 관점에서만 바라보는 것에는 한계가 있다. 이런 관점에서는 인공지능 거버넌스 역시 누군가가 이익을 보면 다른 누군가는 손해를 입는 제로섬 게임이 된다. 이에 대한 대안으로 사람들이 인공지능 시스템의 신뢰성이 공정하게 다양한 가치관 세트를 반영한다고 느끼는 방식으로 기계가 루프 안의 공중(public)-사회(society)에 의해 훈련받는 방법이 있다. 이는 결코 전례 없던 방식이 아니

다. 이상적인 정부 하에서는 시민들이 충분히 정보를 제공받고 정부가 스스로를 대표하고 있다고 믿으며 궁극적으로 자신들이 정부의 행동에 책임을 진다는 생각을 갖게 된다. 이러한 상황을 인공지능 거버넌스에 적용해보는 사고 실험을 구현한 것이 Rawhan(2017)이 제기한 루프 속의 사회라는 문제의식이라고 할 수 있다.

실제로 공중의 이익에 민감하게 관계되는 영역에서는 공중이 훈련시키고 공중이 믿을 만하게 충분히 투명한 인공지능을 설계할 가능성도 있다. 사실 규칙 기반의 전통적인 소프트웨어와 달리 최근 발전한 데이터 중심의(data-centric) 인공지능은 뇌와 닮아 있는 훨씬 복잡한 기계이므로 인공지능이 실제로 하는 일을 이해하기 어렵고 그런 이유로 이 결과를 신뢰하게 만들고 리스크를 관리하는 것은 공학자와 산업계가 먼저 알아서 할 일일 것이다. 하지만 오히려 바로 그런 이유로 인해 공중이, 사회가 기계의 가치와 행동을 검증하고 감사(audit)해야 할 필요성이 도출된다. 예를 들어 최근 한국에 도입된 지능정보사회기본법에 근거하여 예고된 인공지능 영향평가가 단순한 기술적 평가가 아닌, 사회적 영향평가가 되어야 하는 이유가 여기에 있다. 사회가 루프 안에 있어야 한다는 것은 인간이 루프 안에 있어야 하는 것만큼이나 인공지능 윤리 거버넌스에서 중요한 측면이라고 말할 수 있다.

## 제3절 AI 거버넌스의 윤리적 쟁점

최근 5년 간 인공지능 거버넌스나 윤리의 제도화에 대한 논의가 활발해지고는 있으나. 불행히도 여기에는 두 가지 잠재적 문제가 존재한다. 첫째, 다양한 인공지능 윤리 원칙이 알고 보면 대동소이해서 불필요한 반복과 동어반복에 그치게 되거나, 둘째, 유의미하게 서로 다 다를 경우 혼란과 모호성을 낳게 된다는 점이 그것이다(Floridi and Cowls, 2019). 그 최악의 결과는 윤리 원칙을 위한 시장이 열리게 되어 이해당사자가 가장 구미에 당기는 윤리원칙을 쇼핑하는 식으로 채택하는 것이다.

사실 인공지능은 하나의 단일 기술을 의미한다기보다는 다양한 섹터에 걸쳐진 복수의 기술들을 포함하는 느슨한 개념이다. 따라서 모든 인공지능을 규율하는 단일한 거버넌스

가 존재하기 어렵다. 게다가 인공지능에 관여하고 있는 이해 당사자들은 저마다 거버넌스에 대해 서로 다른 관념을 갖고 있으며 많은 경우 자신이 활동하는 도메인에서만 적합한 거버넌스를 좋은 거버넌스라고 여기는 경향마저 있다. 하지만 인공지능 시스템은 하나의 도메인에 머무르지 않고 이용자들이 사용하는 여러 개의 디바이스를 통해서 상호 연결되며 다양한 데이터를 기반으로 처리되기에 기본적으로 복잡성(complexity)을 띠고 있다.

어떤 사람들은 개별 기업의 자율규제를 옹호하고 또 다른 이들은 집합적인 산업별 규제를 지지하고 기업에서는 정부의 무지를 탓하기도 하지만 이는 인공지능 시스템이 갖는 복잡한 성격을 이해하지 못한 한계에서 비롯되는 것이다. 정부를 탓하는 사람들의 경우 정부는 유연성이 부족하다거나 인공지능에 대한 이해도가 떨어져 효율적인 규제 방식을 이해하지 못한다고 말한다. 또한 인공지능의 미래에 대한 우려를 과도하게 부풀린 이야기들을 무분별하게 받아들인 후 만들어내는, 섣부른 윤리적 대처가 혁신을 가로막을 것이라고 걱정하는 목소리도 존재한다.

그러나 윤리와 혁신이 언제나 대립하는 가치인 것은 아니다. 오히려 인공지능 윤리에 대한 사회적 합의를 사전적으로 잘 다져놓음으로써 혁신에 뒤따라올 리스크를 완화할 수 있고 사회적으로 지속가능한 성장을 추진할 수 있기 때문이다.

## 1. 인공지능 거버넌스의 복잡성

인공지능 거버넌스 이슈는 인공지능이 다른 기술에 비해 상대적으로 '자율성'을 띠며 많은 경우 설계자와 이용자 사이의 정보 비대칭성이 커서 이용자는 대부분 그 메커니즘을 이해하지 못하는 채 인공지능이 내리는 의사결정에 일방적으로 영향을 받게 되고 어떤 수준을 넘어가면 설계자조차 설명할 수 없게 된다는 상황(블랙박스)이 존재한다는 사실로 인해 더욱 관심을 끌고 있다. 그런데 이러한 믿음 자체가 역설적으로 인공지능을 가장 많이 사용하고 인공지능 기반으로 새로운 서비스를 만드는 데 필요한 데이터를 가장 많이 갖고 있는 빅테크 기업이 직간접적으로 규제 프로세스에 끼어들게 하는 이유를 제공하고

있다. 미국의 인공지능 거버넌스 조직은 물론이고 52인으로 이루어진 EU의 인공지능 고위 전문가 그룹 중 절반 이상이 기업 출신이다. 거버넌스에 민간 전문가나 기업 출신이 참여하는 것 자체가 문제 되는 것은 아니지만 인공지능 시스템을 만드는 사람들에 비해 인공지능 시스템의 영향을 받는 사람들이 덜 대표되는 것은 문제라고 할 수 있다.

인공지능 거버넌스 논의는 관련된 많은 이슈들을 리스트에 포함시켜 왔다. 2016년 이후 급속히 증가한 인공지능 거버넌스에 대한 문헌은 단지 정부가 주도한 것들에 그치지 않고 민간 부문의 조직과 NGO를 망라한다. 이러한 문헌에는 윤리, 법안, 원칙, 가이드라인, 프레임워크, 정책 전략 등이 포함된다. Gasser & Almeida(2017)는 관련 이슈들을 계속 추가해나가는 것보다는 규제(regulation)와 좀 더 관련된 대규모의 구조적 도전을 고려해보는 것이 낫다고 제안했다. 그 도전은 다음 세 가지로 나눠볼 수 있다. 첫째, 정보 비대칭(information asymmetry) 문제다. 수많은 사람들이 인공지능에 의해 영향을 받고 인공지능에 기대어 살아가지만 극소수의 사람과 기업만이 그 기술을 이해한다. 인공지능 시스템은 이 시스템의 개발자와 소비자/정책입안자들 사이에 큰 정보 비대칭을 낳는다. 따라서 인공지능 시스템의 거버넌스는 무엇보다도 집합적 이해 수준을 제고하는 것, 즉 인공지능으로 인한 사회적 변화와 적용의 맥락을 증대시키려는 목적을 가지고 작동하는 메커니즘을 거버넌스에 포함시키는 것을 의미한다.

둘째, 규범적 컨센서스 문제는 인공지능이 사회에 주는 잠재적 발전을 지속가능한 발전목표(SDG)와 관련하여 사고하는 것을 의미한다. 거버넌스 모델은 비용편익분석 및 이해관계자 사이의 규범적 컨센서스 구축을 위한 공간을 열게 된다. Gasser & Almeida(2017)에 따르면 특히 인공지능 시스템의 설계에서 공정성과 효율성처럼 트레이드오프가 끼어드는 영역을 중심으로 규범적 컨센서스 문제가 점점 중요해진다. 따라서 규범적 차이가 존재하는 서로 다른 지역과 맥락에서 인공지능 거버넌스는 상이한 프레임워크와 접근 간에 상호운용성을 보장하는 형태로 발전되어야 한다.

셋째, 기술에 대한 상호이해가 있다고 하더라도 거버넌스의 테크닉 문제, 무엇이 바람직하지 않은가에 대한 사회적 합의, 그리고 효과적이고 효율적이며 정당한 수단의 설계 문제

는 여전히 남게 된다.

사실 인공지능 거버넌스의 복잡성은 사회가 처한 복잡성에 인공지능이 제기하는 복잡성을 곱한 것에 가깝다. 그 필요성을 사회구성원에게 설득하고 특히 이해 당사자들이 실행 가능한 방식으로 제시하는 어려움도 뒤따르게 된다. 적절한 인공지능 윤리 거버넌스 시스템을 수립하는 것이 어려운 이유는 첫째, 윤리적 관심의 다양성에 그 기원이 있다. 둘째, 이해 당사자들이 합의 가능한 선에서 적합한 규제 수단을 결정하기 어렵기 때문이다. 셋째, 경제와 시장, 개인과 사회, 환경과 정치 및 규제환경 사이의 복잡한 상호작용에 기인한다(Walz & Firth-Butterfield, 2019). 여기에 더해 윤리에 대한 국제적 논의 환경을 고려해 보면 모든 국가에 적용되는 일률적인 윤리 거버넌스가 성립할 수 없다는 결론에 도달하게 된다.

### 1) 인공지능 (시스템) 거버넌스

인공지능 시스템의 거버넌스에서 법과 윤리, 기술은 각각의 고유한 역할을 수행하지만 동시에 상호보완적이다(Cath, 2018). 인공지능 윤리 거버넌스를 이해하기 위해 우선은 인공지능 거버넌스 그 자체를 이해하는 것이 필요하다. 인공지능과 윤리적 프레임워크를 둘러싼 문헌들은 인공지능의 영향의 거버넌스를 다루는 법제적 접근, 알고리듬 영향 평가와 같은 기술적 접근, 시스템 인증을 통한 신뢰성 구축 등을 다루고 있는데 법, 윤리, 기술은 이런 의미에서 모두 연결되어 있다. 인공지능 거버넌스에 있어 중요한 것은 어떤 경우에 이 셋 중의 하나가 가장 적합성을 가지는가. 법-규제 프레임워크가 필요한 영역은 어디이며 어떤 순간에는 윤리적 혹은 기술적 접근만으로도 충분한가를 결정하는 문제다.

인공지능 거버넌스를 다룰 때 또 하나 잊지 말아야 할 것은 인공지능 시스템의 진화에 따라 주어진 사회에서 사람들이 가지고 있는 관념, 선호, 가치, 태도 등이 함께 변한다는 사실이다. 주어진 사회에서 기술 혁신을 다룰 때 핵심은 단순히 그 혁신이 지속가능한가에 그치지 않고 그 혁신은 사회적으로 바람직한가, 구성원들이 그 혁신을 기꺼이 수용하는가라는 질문과 직결되기 때문이다(Floridi, 2018). 결국 거버넌스는 이러한 인공지능 시스템

을 수용할 때 사회적으로 어떤 가치와 방법을 우선시할 것인가, 그리고 어떻게 하면 변화하는 관념과 선호, 가치와 태도를 너무 늦지 않게 반영할 수 있을 것인가 라는 물음을 동반한다.

예컨대 2021년 4월에 EU가 발표한 인공지능 법안은 적용된 기술의 위험 수준을 기준으로 거버넌스를 달리하겠다는 관점에서 작성되었다. AI의 혜택과 위험을 고려한 균형 잡히고 비례적인 규제 방식(balanced and proportionate horizontal regulatory approach)을 채택한 것은 기술의 지속가능성을 염두에 둔 것이지만 동시에 사회의 지속가능성 역시 놓치지 않으려는 시도다. 어떤 기술이나 매우 범용성이 높아지고 사회 전역으로 확산되어 가는 과정에서는 통제 불가능한 파급력에 대한 우려가 발생하기 마련이다. 자율주행자동차, 인공지능에 의한 판결, 인공지능이 심사하는 채용 서류, 인공지능에 의한 진단, 주식 거래, 임대주택의 배분 등, 사람이 담당했던 영역을 위임하는 일에 있어서 책임은 누구에게 귀속될 것인가 라는 문제는 단순히 인권 문제에 한정되지 않는 광범위한 사회적 이슈들을 제기했다.

결과적으로 인공지능 거버넌스를 둘러싼 논쟁은 윤리의 자리에 대한 논의를 전 지구적 차원에서 빠르게 정착시켰다고 할 수 있다. 실제로 인공지능 윤리 거버넌스는 데이터나 인공지능 시스템을 구동하는 알고리듬 자체에 대한 논의는 물론 사회적 맥락과 따로 떼어서 생각하기 어렵다.

### 2) 층화된 거버넌스 vs 범위에 따른 거버넌스 모형

디지털 생태계가 층화되어 있기 때문에 거버넌스도 층화될 필요가 있다. 그런 의미에서 국제법과 국내법, 규제와 윤리는 서로 다른 층위에 위치한다. 윤리와 규제 사이에는 보충성의 원리가 적용되어야 하지만 각각의 층위에서 동일한 목적을 달성할 수 있도록 함께 움직일 필요가 있다.

이러한 거버넌스 협업 체계는 매우 다양한 형태를 가질 수 있다. 정부로부터 독립된

위원회를 수립하여 정부의 역할을 대행할 수도 있고, 민간기업이나 시민사회가 주도하는 협치형 거버넌스를 만들 수도 있다. 협업체계의 모습에 대한 논의도 활발히 이루어지고 있다. 대표적으로 Gasser & Almeida(2017)는 3가지 계층으로 이루어진 거버넌스를 제시했다. 제1계층은 기술계층으로 주로 기술전문가와 기업으로 구성되는 거버넌스 계층이다. 제2계층은 윤리적 계층인데, 기술계층의 상위에 놓여 보다 광범위한 윤리적 기준과 원칙을 제시한다. 제3계층은 규제 계층인데 정부당국에 의해 주도되며 최상위의 규범이나 법률을 제정・운영한다.

<그림 1> 인공지능 거버넌스를 위한 층화 모델


A layered madel for AI governance. The interacting layers (which sit between society and AI applications) are social and legal; ethical; and technical foundations that support the ethical and social layers.

출처 : Gasser & Almeida(2017)

위의 그림은 사회와 인공지능 시스템 사이의 층화된 거버넌스 모델을 보여준다. 기술적 층위는 윤리적, 사회적 층위를 지지한다. 윤리적 층위는 기술적 층위 및 사회적 층위와

상호작용한다. 따라서 윤리적 거버넌스 논의는 인공지능 거버넌스의 일부라고 할 수 있으며 기술 층위와 사회적 논의와 따로 떨어져 존재할 수 없다.

기간에 따라서도 인공지능 거버넌스의 모델은 달라질 수 있다. 단기적으로 거버넌스는 인공지능 알고리듬의 표준과 원칙을 발전시키는 것에 집중한다. 장기적으로 국민국가는 성숙한 인공지능의 응용을 규제하기 위한 입법이나 규범을 만들 수 있다. 하지만 각각의 층위 사이에는 상호작용이 존재하며 하나의 층위 안에서도 다양한 수단이 결합될 수 있다. 거버넌스 프로세스는 시장 지향적인 솔루션으로부터 정부 기반의 구조에 이르기까지 스펙트럼이 넓으며 한 나라 안에서 혹은 국제적으로 적용될 수 있다.

Dafoe(2018)는 인공지능 거버넌스와 관련하여 기술 거버넌스, 정치 거버넌스, 이상적 거버넌스라는 분류를 제안한 바 있다. 인공지능 기술 거버넌스는 인공지능의 기술적 요소와 문제점을 다루며 주로 전문가와 기업 중심으로 이루어진다. 반면 인공지능 정치 거버넌스는 인공지능을 둘러싼 개발자, 정부, 기업, 연구자, 시민단체 등 다양한 행위 간의 관계를 다루며 이들의 이해관계를 조정하는 것을 뜻한다. 마지막으로 이상적 인공지능 거버넌스는 전 세계적인 공동체의 관점에서 인공지능의 활용과 통제에 대한 이해를 통합해 나가는 거버넌스이다. 앞서의 Gasser & Almeida(2017)가 제안한 층화 모델과는 좀 다른 방식으로 규율이 적용되는 범위와 공간을 염두에 둔 분류라고 할 수 있다.

### 3) AI의 지속가능성을 위한 글로벌 거버넌스

인공지능에 기반한 기기나 서비스의 범용성이 높기 때문에, 또 인공지능의 파급효과가 크고 미치는 범위가 넓기 때문에 인공지능 거버넌스 논의는 한 국가의 경계를 훌쩍 넘어선다. 국제기구는 인공지능 거버넌스를 논할 때 주로 '인간 중심적 접근(Human-centered approach)'을 기본적으로 전제하고 있다. 유네스코가 2021년 12월에 193개국 만장일치로 채택한 '유네스코 인공지능 윤리 권고'도 궤를 같이 한다. 권고는 모든 회원국이 AI의 건강한 발전을 위해 필요한 공통의 가치와 원칙을 규정하고 있다.

〈표 2〉 AI 분야 주요 다자간 거버넌스 이니셔티브와 행위자

| | 국가 주도 | 비국가 주도 |
|---|---|---|
| 기존 아키텍처에 배태 | G7<br>G20<br>자율살상무기 분야의 신기술에 관한 정부 전문가 그룹(GGE)<br>유럽 평의회 | UN<br>유럽 집행위원회<br>OECD<br>IEEE<br>ISO/IEC<br>AI 파트너십(PAI) |
| 새로운 수단 마련 | AI글로벌 파트너십(GPAI)<br>국방 AI 파트너십 | |

출처: Schimitt(2021)

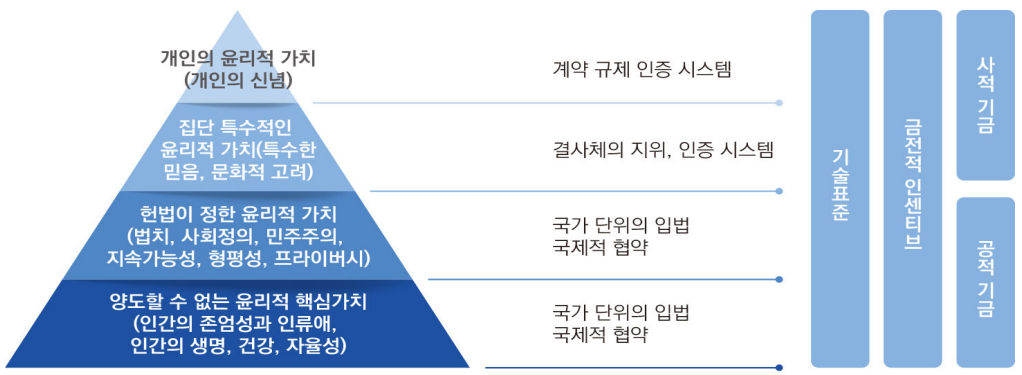
한편 Gill & Germann(2021)은 지속가능한 발전 목표(Sustainable Development Goals: SDGs)의 관점에서 인공지능 거버넌스를 살펴보았다. 실제로 2015년에 채택된 SDG나 그 이전의 밀레니엄 발전 목표에서는 ICT에 대한 접근성 위주로 정의가 되고 있었다. 2018년 7월 UN 총장이 추진한 디지털 협력에 관한 고위 패널(UN Secretary-General's High-level Panel on Digital Cooperation)은 디지털 기술의 혜택을 극대화하고 사회적, 윤리적, 법적, 경제적 영향을 그 위험을 최소화 하는 디지털 협력을 위해 만들어졌다. 공동의장으로는 멜린다 게이츠(Melinda Gates) 게이츠 재단 공동회장 및 마윈(Jack Ma) 알리바바 회장이 선임되었으며 민·관·산·학계 전문가 20명이 참여했다. 이 고위 패널이 주로 다룬 두 가지 주제는 포용성과 디지털 공공재에 관한 것이다. 이 보고서에서 패널은 광범위한 복수의 이해 당사자 연합이 UN과 함께 프라이버시를 존중하는 방식으로, 그리고 SDGs를 성취할 수 있는 분야에서 디지털 공공재의 공유, 역량의 투입, 데이터셋의 공개를 위한 플랫폼을 만들 것을 권고한 바 있다.

무역과 디지털 경제를 다루었던 2019년 6월의 G20 장관급 회담에서도 인간 중심의 인공지능의 역할을 확인한 바 있다. 이러한 인간 중심적 접근은 인공지능 윤리의 근간을 이루고 있다. 인공지능의 파급효과가 크고 상호연관성이 높으며 미치는 범위가 넓기 때문에 글로벌 인공지능 거버넌스는 매우 체계적으로, 또 지역 간의 문화적 차이를 고려하며 다루어져야 한다. 모든 지역과 분야를 아우르는 인공지능 윤리가 존재할 수 없기 때문이다.

## 2. AI 윤리 거버넌스

인공지능 윤리 거버넌스는 데이터나 인공지능 시스템을 구동하는 알고리듬에 대한 논의와 따로 떼어서 생각하기 어렵다. 인공지능 윤리가 문제시된 것은 인공지능 중에서도 기계학습의 발전과 밀접한 연관을 갖는다. 윤리적 기계학습은 설계와 개발, 전개, 실천, 이용과 이용자 등 전과정에 대한 문제를 제기하게 된다. 아래 그림은 인공지능 시스템에서 윤리적 가치들의 레벨과 그에 상응하는 위계화된 거버넌스 모형을 보여주고 있다.

<그림 2> 인공지능 시스템에 윤리적 고려를 실행하기 위한 단계별 거버넌스 모델




출처: Walz & Firth-Butterfield(2019)

플로리디는 디지털 혁명이 디지털 윤리를 요청했다고 주장하는 글에서 비윤리적인 것의 다섯 가지 위험을 열거한 바 있다. 첫째, 윤리 쇼핑, 둘째, 윤리 화장, 셋째, 윤리 로비, 다섯째, 윤리 덤핑이 그것이다. 이러한 논의는 인공지능 윤리 논의에도 적용될 수 있을 것으로 보인다.

윤리 쇼핑은 너무 많은 원칙, 규준, 가이드라인, 프레임워크가 쏟아져 나오게 되면서 일관성이 부족하고 혼란스러운 가운데 정작 이해 당사자들은 그 중에 무엇이 바람직하지를 알지 못하게 되는 것과 관계가 있다. 이 때 윤리적 위험은 이 행동과잉이 만들어낸 원칙과 가치들의 시장 그 자체이며 여기서 이해 당사자들은 실질적으로 자신들의 현재 행위들을

사회적으로 수용된 윤리적 프레임워크에 맞추려고 노력하기보다 오히려 현재 행위를 가장 잘 정당화하는 수단으로 윤리 쇼핑을 하게 된다는 것이다.

플로리디가 블루워싱이라고 표현한 윤리 화장은 환경 문제에서 통용되는 그린워싱과 비슷한 것이다. 블루워싱은 윤리적 가치와 디지털 프로세스의 가치, 제품과 서비스 혹은 다른 솔루션을 위해서 실제 그런 것보다 디지털 윤리에 더 부합하는 것처럼 보이도록 실질적이지 않은 요구를 하거나 방향 설정이 올바르지 않은 요구를 하는 것, 또는 피상적인 조치들을 시행하는 것을 뜻한다. 즉 실질적으로 윤리원칙을 적용한다기보다는 마치 적용한 것처럼 보이도록 하는 것이다.

윤리 로비는 사적인 행위자들이 자율 규제라는 이름으로 로비하려는 것을 말한다. 필요하고 바람직한 법제화를 늦추거나 수정, 교체, 회피하려는 시도다.

윤리 덤핑, 혹은 윤리 기피는 더블 스탠다드 문제다. 윤리 기피는 편향, 불공정, 이기적 이해관계에 기초하는 것이며 점점 “덜 윤리적인 작업”을 많이 함으로써 사실상 윤리적 활동을 기피하는 것을 의미한다. 특히 이들은 취약한 집단, 힘없는 기관, 법적 불확실성, 부패한 체제, 불공정한 권력 배분, 그리고 경제, 법, 정치, 사회적 악이 지배하는 곳에서 이런 일을 한다는 점에서 윤리 로비와 구별된다. 긱 이코노미가 대표적인 사례라고 할 수 있다. 결국 윤리 기피는 명백한 책임 할당의 부족에 기인한다. 에이전트는 책임을 재할당할 수 있다면 윤리를 기피한다.

이러한 비윤리성의 위험을 제어하고 인공지능의 개발과 전개에 최소한의 윤리적 한계를 정해두고자 하는 시도를 둘러싸고 다양한 행위자들이 개입하게 된다. 인공지능 거버넌스에는 정책결정자 뿐만 아니라 민간기업이나 비영리단체의 대표, 전문가를 포함해야 한다는 데 많은 사람들이 동의한다(Scherer, 2016; Cath et al., 2018; Rahwan, 2018). 인공지능이 줄 수 있는 개인적, 사회적 편익과 감수해야 할 위험에 대한 트레이드오프가 분명히 존재하기 때문에 인공지능 개발 및 이용의 원칙들을 시민사회와 기업이 공유해야 할 필요가 있다. 특히, 이러한 거버넌스 하에서 구축된 인공지능에 대한 신뢰는 인공지능의

사회적 수용성을 높이고 긍정적 영향을 극대화하는 데 필수적이다(Smith, 2018).

한편, Cath et al.(2018)의 연구는 '좋은 인공지능 사회'를 만들어나가기 위한 거버넌스 구조는 정부가 주도하는 위원회가 적합하며, 해당 위원회는 독립적인 운영과 권한을 보장받으면서 다양한 이해관계인의 참여를 보장해야 한다고 주장한다. 인공지능 윤리 거버넌스란 인공지능의 광범위한 사회경제적 영향력과 요구를 고려하면서 윤리를 체계적으로 제도화하는 것을 의미한다. 제도화는 단순히 법제화에 국한되지 않는 좀 더 광범위한 영역 즉, 원칙, 가이드라인, 권고, 활동 수칙, 혹은 프레임워크에 이르는 여러 수단을 망라한다. 또한 인공지능 윤리 거버넌스는 전통적인 규제로는 해결할 수 없는 새로운 문제를 제기하며 따라서 새로운 접근방식을 요구한다.

EU 집행위원회의 과학 및 신기술 윤리 그룹(Group on Ethics in Science and New Technologies)은 다음 그림과 같이 윤리적 문제에 대한 리스트를 발표하였다.

〈 표 3 〉 인공지능 응용에 윤리를 실행할 수단

| 과학과 신기술에 대한 윤리 그룹의 윤리적 원칙들 | | |
|---|---|---|
| 인간의 존엄성 | 정의, 공정, 연대 | 보안, 안전, 심신의 건전성 |
| 자율성 | 민주주의 | 데이터 보호와 프라이버시 |
| 책임 | 법치와 책무성 | 지속가능성 |

출처: European Group on Ethics in Science and New Technologies (2018) "Statement on Artificial Intelligence, Robotics and Autonomous Systems".

위 표에서 알 수 있듯이 윤리가 다룰 수 있는 문제는 매우 광범위하고 다양한 가치들을 포함하고 있으며 우선성 면에서 논쟁의 여지가 있다. 물론 인간의 존엄성처럼 보편적인 가치도 있지만 현실에서 인공지능 응용의 윤리적 함의를 평가하는 것은 문화적이고 경제적인 조건에 좌우될 수밖에 없다. 성차별이나 인종차별 역시 보편적으로 옹호되는 가치이지만 그것이 인공지능 스피커에 적용될 때와 채용 알고리듬에 적용될 때는 다른 효과를

가져오며 또한 그 스피커가 설치되는 주변 환경과 지역적, 문화적 맥락에 의해 평가 결과가 좌우될 수 있기 때문이다. 개인정보 문제이긴 하지만 코로나-19로 인한 팬데믹 시기 QR코드 인증이나 그린 패스를 둘러싼 인권 침해 논란은 같은 선진사회에서도 동서양의 문화적 차이를 드러냈으며 선진국과 개발도상국 사이의 수용성 차이도 큰 것으로 드러났다.

인공지능 윤리 거버넌스를 발전시키기 위해서는 우선 인공지능의 사회경제적 영향에 관한 분석, 인공지능의 확산에 따른 위험과 도전과제 등에 대한 사회적 합의가 선행되어야 한다. 둘째, 그러한 도전에 대응할 다양한 수준의 조치들을 만들어야 한다. 셋째, 이를 통해 주어진 사회에 필요한 인공지능 거버넌스를 구체화해야 한다. 넷째, 인공지능 거버넌스를 통해 지키고자 하는 사회적 가치들을 분명히 해야 한다. 이러한 접근 방식은 윤리를 통해서 발전하는 산업의 발목을 잡는 게 아니라 윤리를 고려하는, 지속가능한 혁신에 대한 사회의 비전을 제시할 것을 정부에 요구한다. 아래 표는 일반적인 의미의 거버넌스와 인공지능을 통합할 원칙에 관한 내용을 담고 있는데 윤리 거버넌스와 관련해서 중요한 통찰력을 제공해 준다.

**<표 4> 인공지능 & 선한 거버넌스를 통합할 네 가지 원칙**

| 원칙 | 내용 |
|---|---|
| 포용적 설계 | 부적합한 데이터셋과 관련한 인공지능의 차별과 편향, 소수자 배제와 덜 대표되는 집단, 설계에서 다양성 부족 등 |
| 정보가 충분한 조달 | 성실의 의무, 설계, 사용성과 리스크와 혜택의 평가 |
| 목적에 합당한 실행 | 상호운용성, 공무원 훈련 필요성, 의사결정 과정과의 통합 |
| 끈질긴 책무성 | 블랙박스 알고리즘과 관련된 책무성과 투명성 이슈, 시스템의 예측가능성과 설명가능성, 모니터링과 감사 |

출처: OxCAIGG(2020) Four Principles for integrating AI & Good Governance

# 제4절 AI 윤리 거버넌스의 구현

결국 윤리 거버넌스의 궁극적인 목적은 사회의 근간을 이루는 가치와 원칙, 그리고 윤리를 보호하도록 하는 알고리듬 기반의 사회를 설계하는 데 있다. 이는 인공지능 시대를 지속가능하게 만드는 데 필요한 중요한 집합적 역량이기도 하다. 현재 인공지능 윤리 기준이나 권고를 넘어서는 다음 단계의 인공지능 윤리 거버넌스에서 중요한 문제는 기계에 위임된 자율성의 크기에 대해 사회적으로 합의에 도달하는 문제이고 얼마나 민주적이면서도 합리적인 방식으로 그 합의를 구현할 수 있는가 하는 문제라고 할 수 있겠다. 하지만 합의 구현에 다다르는 방법은 여러 가지가 있을 수 있다. 여기서는 세계경제포럼의 백서(Walz & Firth-Butterfield, 2019)에서 시도한 인공지능 윤리 거버넌스의 유형을 접근 방식에 따라 먼저 살펴보고 이어 사회적 관점의 내재화에 대해서 논의하도록 하겠다.

## 1. 기술적 접근 방식: 설계를 통한 윤리 구현

### 1) 하향식 vs 상향식 접근

상향식(bottom-up) 접근에서는 기계가 특수한 상황에서 인간의 행위를 관찰한 뒤 이에 근거하여 인간이 윤리적 결정을 내리는 방법을 학습하지만 기계는 윤리가 무엇인지 상식이 무엇인지 알지 못한다.

### 2) 사례별 접근

보편적으로 통용되는 윤리적 의사결정 기준이 존재하지 않는다고 보고 상황별로 적용되는 윤리적 의사결정을 중시하는 방법이다. 사전 지침 및 이용자의 의지와 동의 기반으로 작동하기 때문에 명확한 지침이 부재하거나 이용자가 의지를 표명할 수 없는 상태일 때에는 무용하다.

### 3) 도그마적 접근

인공지능이 윤리적 결정을 하기 위해 모든 가능한 시나리오를 다 고려하는 대신에 특정한 사상의 조류에 맞도록 프로그램화하는 방법도 있다. 프로그램의 작동 여부는 별개의 문제라고 할 수 있다.

### 4) 기술적 메타 수준에서 인공지능 실행

인공지능에 의한 자동화된 의사결정의 관점에서 사전에 결정된 법과 윤리적 규준들에 합치되는 인공지능 주도의 모니터링 시스템(guardian AI)을 생각해볼 수 있다. 이는 인공지능 자체가 불법, 비윤리적 의사결정을 기관에 보고하도록 하는 조치다.

### 5) 기술적 수단과 메커니즘의 불충분성

인공지능 시스템은 사람과 기업이 만드는 것이므로 기술적 수단만으로는 윤리적 고려를 충분히 하지 못할 위험이 있다. 이들은 법적인 강제가 있을 때나 윤리 규준을 지킴으로써 자신의 이익이 발생할 때만 윤리적으로 합치된 방식으로 인공지능을 프로그래밍할 가능성이 있으므로 사람과 기업의 개입이 필요하다

## 2. 정책적 수단

정책적 수단을 고려할 때에도 다양한 방법이 있을 수 있다. 경제와 시장, 개인과 소비자, 정부와 국가, 기술과 혁신은 서로 밀접히 연관되어 있으며 하나의 변화가 다른 이해당사자에 직간접적 영향을 주기 때문에 이러한 상호연관성을 고려하면서 규제의 수단을 마련하는 것이 바람직하다. WEF 백서는 다음과 같은 세 가지 수단을 제시하고 있다. 첫째, 새로운 기술의 발전에도 불구하고 원칙을 천명하는 기본법 제정, 둘째, 산업별 법률 제정, 셋째 위험의 수준을 판별하여 단계적으로 법률 적용이 그것이다. 이는 앞서 지적한 것처럼

2021년에 발표된 EU의 인공지능 법안에서 취한 접근 방식이다.

입법은 윤리적 규준에 대한 최소한의 합의에 불과하며 법적 조치는 일국적 경계 안에서만 작동한다는 한계가 있다. 기술 발전 속도를 규제 메커니즘이 따라잡기가 매우 힘든 측면이 있다. 법은 혁신에 부정적 효과를 끼친다는 인식과 규제를 할 경우 규제가 약한 나라 대비 국내 비즈니스에 불이익을 준다는 논리가 존재하지만 입법이 법을 준수하는 기술과 비즈니스 모델을 만들도록 독려함으로써 오히려 혁신을 위한 인센티브를 제공할 수도 있는 것이다. 결국 정책 수단의 복수성, 즉 국제협정, 양자간 투자협정, 자율규제 및 표준화, 인증제도, 계약 규칙, 연성법(경계에 있는 법, 법과 비법의 중간단계를 인정), 애자일 거버넌스(agile governance), 금전적 인센티브 등과 같은 대안들이 있다는 것을 아는 것이 중요하다.

## 3. 사회적 관점의 내재화

아제모을루(Acemoglu, 2021)는 인공지능이 자동화를 촉진시켜 일자리를 잃게 만들고 개인의 행동을 통제하고자 하는 정부와 기업을 도울 것이라는 점에 일단 동의하지만 이것이 불가피한 미래라고 많은 사람들이 믿고 있는 것을 더 우려한다. 이러한 미래로 가는 이유는 우리가 바로 그 경로를 선택하기 때문이라는 것이다. 그는 지금이 우리가 가야할 다른 곳에 대한 이야기를 시작할 시점이라고 주장한다. 생산 영역에서 자동화를 위한 인공지능의 사용에만 초점을 맞출 것이 아니라 인공지능이 인간을 돕고 보완하도록 만들어야 한다는 것이다.

인공지능 윤리 거버넌스를 둘러싼 논의들은 단순히 인공지능 설계 단계에서의 윤리의 내재성이나 프라이버시, 인권 보호 등에 관한 문제에 국한되지 않는다. 이는 인공지능에게 얼마만큼의 자율성을 허용할 것인가 하는 문제, 즉 인공지능 시스템이 인간을 대신해서 의사결정을 할 때 그 자율성에 얼마나 우리 사회가 신뢰를 가질 수 있는가 하는 문제와 긴밀히 연관된다. 온라인 플랫폼이 개인화 알고리듬을 사용하여 데이터를 만들어낸 주체

들의 자율성을 훼손하게 되는 것을 우리는 주체의 동의를 이유로 무한히 허용할 수 있을까? 사람들의 선택지를 제한하고 과거의 이용 데이터를 기반으로 정보를 취사선택해서 보여줌으로써 민주주의가 훼손된다면, 그리고 정보의 다양성이 점점 줄어들어 사람들이 에코 체임버에 갇히게 된다면 그것이 기업의 영업 비밀이라는 이름으로 용납될 수 있을까? 물론 이러한 문제를 해결하기 위해서는 좀 더 구체적인 리스크나 위해(harms)에 대한 증거가 있어야 할 것이지만 시민들의 자율성을 침해하는 것, 그 자체가 비윤리적이라는 합의에 도달하는 것 이 문제 해결의 첫걸음이라고 할 수 있다(Taeihagh, 2021: 142). 결국 인공지능 윤리 거버넌스 논의는 처음부터 사회적 관점의 내재화로부터 출발해야 할 것이다.

## 제5절 시사점

오늘날 국제사회는 플랫폼화, 데이터화, 알고리듬 기반 자동화가 추동하는 디지털 전환 시대의 고유한 거버넌스를 찾아가는 중이다. 특히 이러한 디지털 전환에 기초가 되는 인공지능은 최근 팬데믹으로 인해 확산된 비대면 시대의 생활 수준을 고양하고 사회자본을 축적하는 사회적 기술로 각광을 받고 있지만(Gill and Germann, 2021), 같은 이유로 인공지능을 통제하고 인간에게 유리한 방식으로 사용되도록 설계하는 일의 중요성 또한 강조되고 있다. 결국 인공지능 윤리 거버넌스는 모든 리스크를 제로로 만들려는 노력이 아니고 인공지능 시스템이 만들어내는 리스크 관리를 가능하게 만들려는 인공지능 거버넌스의 한 형태라고 할 수 있다.

물론 인공지능 시스템의 발전 및 융복합 현상에 적절히 대응하지 못할 경우 규제 지체(regulatory delay) 및 규제 병목(regulatory bottleneck) 현상으로 인해 시장 확대를 저해할 수 있다는 지적은 타당하다. 혁신적 기술과 서비스가 사업 구현의 기회를 얻지 못하고 사장되면 산업적 성장은 물론 소비자 후생에도 악영향을 미칠 수 있기 때문이다. 그러나 법률은 인공지능처럼 빠르게 발전하는 동태적 기술 영역에 부적합하다. 인공지능에 대한 규율에 있어 유연한 대처 가능한 연성법(soft law) 형태의 입법이나 윤리 규범 차원

의 제재가 보다 적합하다고 보는 이유가 여기에 있다. 물론 자율주행자동차의 경우처럼 초기 시장 형성에 있어서 안전 규제가 중요하게 작용하는 영역은 선제적 법제도 구축이 중요할 수 있다.

이는 결국 새로운 디지털 시대의 사회계약에 관한 논의와 연결될 수밖에 없다. 사회 계약은 사실 진화의 결과라기보다는 상이한 계급, 다양한 이해당사자들 사이의 투쟁의 결과다. 따라서 인공지능 윤리 거버넌스를 둘러싼 논의 역시도 어찌 보면 사회의 기능을 유지하고 사회를 "보호하고" 사회의 붕괴를 막기 위한 시민사회, 정부, 민간의 노력을 조율하는 과정이라고 하겠다.

# 참고문헌


람게, 2020 누가 인공지능을 두려워하나? : 생각하는 기계 시대의 두려움과 희망. 이수영·한종혜 편역. 다섯수레.

오닐, 캐시. 2017. 대량살상수학무기. 흐름출판.

유뱅크스, 버지니아. 2018. 자동화된 불평등: 첨단 기술은 어떻게 가난한 사람들을 분석하고, 감시하고, 처벌하는가. 북트리거.

이상엽·이동규. (2020). 인공지능(AI)의 경제적 영향과 향후 정책방향에 대한 시사점: 조세 및 사회보장제도를 중심으로. 조세연구, 20(3), 61-88.

이상욱, 조은희 엮음 2011, 『과학 윤리 특강 - 과학자를 위한 윤리 가이드』, 서울: 사이언스북스.

이중원 외 2018,『인공지능의 존재론』서울: 한울.

이중원 외 2019,『인공지능의 윤리학』서울: 한울.

이호영. 2021. “알고리듬이 편향된다면”. 플랫폼 사회가 온다. 한울.

한국과학기술평가원. 2015. 기술영향평가: 인공지능.

한국인공지능법학회 2019,『인공지능과 법』서울: 박영사.

Acemoglu, D. eds. (2021) Redesigning AI. Boston Review.

boyd, danah, Karen Levy, and Alice Marwick. (2014). “The Networked Nature of Algorithmic Discrimination.” Data & Discrimination: Collected Essays (Eds. Seeta Peña Gangadharan and Virginia Eubanks), pp. 43-57.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. (2021). “The Productivity J-Curve: How Intangibles Complement General Purpose Technologies.” American Economic Journal: Macroeconomics, 13 (1):333-72.

Cath, Corinne (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges Phil. Trans. R. Soc.

Elish, Madeleine Clare and Boyd, Danah (2017). Situating Methods in the Magic of Big Data and Artificial Intelligence (September 20, 2017). Communication Monographs.

Etzioni, Amitai, and Oren Etzioni. (2017)"Incorporating Ethics into Artificial Intelligence." The Journal of Ethics 21 (4): 403-18.

European Group on Ethics in Science and New Technologies (2018) "Statement on Artificial Intelligence, Robotics and Autonomoos Systems".

Fjeld, J., N. Achten, H. Hilligoss, A. Ch. Nagy and M. Srikumar (2020). Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. The Berkman Klein Center for Internet & Society Research Publication Series: https://cyber.harvard.edu/publication/2020/principled-a.

Floridi, L. (2018). Soft ethics and the governance of the digital. Philosophy & Technology, 31(1), 1-8.

Floridi, L. and Cowls, J. (2019) A Unified Framework of Five Principles for AI in Society. Available at SSRN: https://ssrn.com/abstract=3831321 or http://dx.doi.org/10.2139/ssrn.3831321

Fry, Hannah 2019, Hello World: How to Be Human in the Age of the Machine, London: Transworld Publishers Ltd.

Gans, Joshua, Goldfarb, Avi and Agrawal, Ajay 2018, Prediction Machines: The Simple Economics of Artificial Intelligence, Cambridge, MA: Harvard Business School Press.

Gasser, U., & Almeida, V. A. (2017). A layered model for AI governance. IEEE Internet Computing, 21(6), 58-62.

Gasser, Urs, and Virgilio A.F. Almeida. 2017. "A Layered Model for AI Governance." IEEE Internet Computing 21 (6) (November): 58-62.

Google, 'AI at Google: Our Principles' (2018) <https://www.blog.google/technology

/ai/ai-principles/〉

Harris, Charles E. et al. 2018, Engineering Ethics: Concepts and Cases, 6th Edition, Boston: Sengage Learning.

IEEE 2019, Ethically Aligned Design, 1st Edition. (https://ethicsinaction.ieee.org/#series 참조)

Kaplan, Jerry 2016, *Artificial Intelligence : What Everyone Needs to Know,* Oxford : Oxford University Press.

Kitcher, Philip 2001, Science, Truth and Democracy, New York: Oxford University Press.

Mason, Paul 2016, Postcapitalism: A Guide to Our Future, London: Panguin Books.

Mitchell, Melanie 2020, Artificial Intelligence: A Guide for Thinking Human, New York: Picador.

OECD(2019) Going Digital: Shaping Policies, Improving Lives. Available at https://www.oecd.org/digital/going-digital-shaping-policies-improving-lives-9789264312012-en.htm

Prunkl, C. EA, et al. (2021).“Institutionalizing Ethics in AI through Broader Impact Requirements.” Nature Machine Intelligence, vol. 3, no. 2, Nature Research, , pp. 104-10.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5-14.

Schmitt, L. (2021). Mapping global AI governance: a nascent regime in a fragmented landscape. *AI Ethics*.

Sharre, Paul 2019, Army of None: Autonomous Weapons and the Future of War, New York: W.W. Norton & Co.

Shneiderman, B. (2021). Responsible AI: Bridging From Ethics to Practice.

*Communications of the ACM*, 64(8): 32-35.

Singer, Peter 2011, The Expanding Circle: Ethics, Evolution, and Moral Progress, Princeton, NJ: Princeton University Press.

Susskind, R. and Susskind, D. 2017, *The Future of Professions: How Technology Will Transform the Work of Human Experts,* Oxford : Oxford University Press.

Taeihagh, A. (2021) Governance of Artificial Intelligence. Policy and Society 40(2): 137-157.

UNESCO 2019, Preliminary Study on the Ethics of Artificial Intelligence (https://unesdoc.unesco.org/ark:/48223/pf0000367823)

UNESCO 2020, First Draft of the Recommendation on the Ethics of Artificial Intelligence (https://unesdoc.unesco.org/ark:/48223/pf0000373434)

Walz, A., & Firth-Butterfield, K. (2019). AI Governance: A Holistic Approach to Implement Ethics into AI. World Economic Forum.

Winfield, Alan F. T, and Marina Jirotka. 2018. "Ethical Governance Is Essential to Building Trust in Robotics and Artificial Intelligence Systems." Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences 376 (2133).

# 부록

# 유네스코 인공지능 윤리 권고

# THE RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE
# 인공지능 윤리 권고

## PREAMBLE
## 전문

The General Conference of the United Nations Educational, Scientific and Cultural Organization (UNESCO), meeting in Paris from 9 to 24 November 2021, at its 41st session,

2021년 11월 9일부터 24까지 파리에서 열린 제 41차 유엔 교육과학문화기구(UNESCO; 이하 유네스코) 총회는

Recognizing the profound and dynamic positive and negative impacts of artificial intelligence (AI) on societies, environment, ecosystems and human lives, including the human mind, in part because of the new ways in which its use influences human thinking, interaction and decision-making and affects education, human, social and natural sciences, culture, and communication and information,

인공지능(AI)의 사용이 인간의 사고, 상호작용, 의사 결정에 영향을 주며 교육, 인간, 사회・자연과학, 문화, 정보통신에 작용하는 새로운 방식으로 인하여, 인공지능이 사회, 환경, 생태계, 인간의 마음을 비롯한 인간의 삶에 미치는 심오하고도 역동적인 긍정・부정적 영향을 미침을 인지하고,

Recalling that, by the terms of its Constitution, UNESCO seeks to contribute to peace and security by promoting collaboration among nations through education,

the sciences, culture, and communication and information, in order to further universal respect for justice, for the rule of law and for the human rights and fundamental freedoms which are affirmed for the peoples of the world,

유네스코가 그 헌장 조항에 기반하여, 정의에 대한, 법치에 대한, 그리고 전 세계 사람들에게 보장된 인권 및 근본적 자유에 대한 보편적 존중을 강화하기 위해, 교육, 과학, 문화, 정보통신을 통한 국가간 협업을 증진함으로써 평화와 안보에 이바지하기 위해 노력한다는 사실을 상기하면서,

Convinced that the Recommendation presented here, as a standard-setting instrument developed through a global approach, based on international law, focusing on human dignity and human rights, as well as gender equality, social and economic justice and development, physical and mental well-being, diversity, interconnectedness, inclusiveness, and environmental and ecosystem protection can guide AI technologies in a responsible direction,

국제법에 기반하여 전 지구적 차원에서 개발된 기준 수립 방식으로서 제시된 것이자 성평등, 사회・경제 정의, 개발, 신체적・정신적 안녕, 다양성, 상호연결성, 포용성, 환경 및 생태계의 보호를 비롯하여 인간 존엄성 및 인권에 초점을 맞춘 본 권고가 인공지능 기술을 책임 있는 방향으로 인도할 수 있다는 점을 확신하며,

Guided by the purposes and principles of the Charter of the United Nations,

유엔 헌장의 목적과 원칙을 지침으로 삼아,

Considering that AI technologies can be of great service to humanity and all countries can benefit from them, but also raise fundamental ethical concerns, for instance regarding the biases they can embed and exacerbate, potentially resulting in discrimination, inequality, digital divides, exclusion and a threat to cultural, social and biological diversity and social or economic divides; the need for transparency

and understandability of the workings of algorithms and the data with which they have been trained; and their potential impact on, including but not limited to, human dignity, human rights and fundamental freedoms, gender equality, democracy, social, economic, political and cultural processes, scientific and engineering practices, animal welfare, and the environment and ecosystems,

인공지능 기술은 인류에게 큰 도움이 되며 모든 국가가 이로 인해 혜택을 받을 수 있지만, 이로 인해 발생・악화될 수 있는 편향이 차별, 불평등, 디지털 격차, 배제, 문화・사회・생물 다양성에 대한 위협, 사회・경제적 격차를 초래할 잠재적 가능성, 또한 알고리즘의 작동 및 그 학습에 사용된 데이터의 투명성 및 이해가능성에 대한 필요성, 인공지능이 인간 존엄성, 인권 및 근본적 자유, 성평등, 민주주의, 사회・경제・정치・문화적 프로세스, 과학・공학 실습, 동물 복지, 환경 및 생태계 등에 미칠 수 있는 잠재적 영향과 같은 근본적 윤리 문제를 제기하기도 한다는 점을 고려하고,

Also recognizing that AI technologies can deepen existing divides and inequalities in the world, within and between countries, and that justice, trust and fairness must be upheld so that no country and no one should be left behind, either by having fair access to AI technologies and enjoying their benefits or in the protection against their negative implications, while recognizing the different circumstances of different countries and respecting the desire of some people not to take part in all technological developments,

인공지능 기술이 전 세계에 존재하는 국가 내 및 국가 간 격차 및 불평등을 심화시킬 수 있다는 점, 그 어떤 국가와 사람도 인공지능 기술에 공정한 접근권을 가져 그 혜택을 누리는 것이나 인공지능 기술의 부정적인 영향으로부터 보호받는 것에서 배제되지 않도록 정의, 신뢰, 공정성이 반드시 유지되어야 한다는 점을 인지하며, 동시에 다양한 국가의 다양한 상황을 인지하고 이 모든 과학기술 발전에 참여하지 않고자 하는 일부의 바람을 존중하며,

Conscious of the fact that all countries are facing an acceleration in the use of information and communication technologies and AI technologies, as well as an increasing need for media and information literacy, and that the digital economy presents important societal, economic and environmental challenges and opportunities of benefit-sharing, especially for low- and middle-income countries (LMICs), including but not limited to least developed countries (LDCs), landlocked developing countries (LLDCs) and small island developing States (SIDS), requiring the recognition, protection and promotion of endogenous cultures, values and knowledge in order to develop sustainable digital economies,

모든 국가가 정보통신기술 및 인공지능 기술의 사용에서의 가속화와 더불어 미디어・정보 리터러시에 대한 필요성 증가를 직면하고 있다는 점, 디지털 경제는 특히 최빈개발도상국(LDC), 내륙개발도상국(LLDC), 군소도서개발도상국(SIDS)까지 포함하여 (단, 이에 국한되지 않는) 중・저소득국가(MLIC)으로의 이익 공유를 위해 중요한 사회・경제・환경적 과제 및 기회를 제공한다는 점을 의식하면서, 지속가능한 디지털 경제를 개발하기 위한 내생적 문화・가치・지식의 인식, 보호, 증진을 요구하고,

Further recognizing that AI technologies have the potential to be beneficial to the environment and ecosystems, and in order for those benefits to be realized, potential harms to and negative impacts on the environment and ecosystems should not be ignored but instead addressed,

인공지능 기술에는 환경과 생태계에 이로움을 줄 수 있는 잠재력이 있으며 이러한 이로움이 실현되기 위해서는 환경과 생태계에 미칠 잠재적 피해 및 부정적인 영향이 등한시되지 않고 오히려 해결되어야 함을 또 인지하여,

Noting that addressing risks and ethical concerns should not hamper innovation and development but rather provide new opportunities and stimulate ethically-conducted research and innovation that anchor AI technologies in human rights

and fundamental freedoms, values and principles, and moral and ethical reflection,

위험성 및 윤리적 문제를 해결하는 것이 혁신 및 개발을 방해해서는 안 되며, 오히려 새로운 기회를 제공해야 하고, 인공지능 기술을 인권 및 근본적 자유, 가치 및 원칙, 도덕·윤리적 성찰 위에 뿌리내리게 하는 윤리적으로 수행되는 연구 및 혁신을 고무시켜야 한다는 점에 주목하고,

Also recalling that in November 2019, the General Conference of UNESCO, at its 40th session, adopted 40 C/Resolution 37, by which it mandated the Director-General "to prepare an international standard-setting instrument on the ethics of artificial intelligence (AI) in the form of a recommendation", which is to be submitted to the General Conference at its 41st session in 2021,

2019년 11월, 유네스코 제40차 총회가 '40 C/Resolution 37'을 채택하여, 사무총장에게 2021년 제41차 총회 때 제출하도록 "인공지능 윤리에 관한 국제 기준 수립 도구를 권고의 형태로 준비할" 권한을 부여했다는 점을 또한 상기하면서,

Recognizing that the development of AI technologies necessitates a commensurate increase in data, media and information literacy as well as access to independent, pluralistic, trusted sources of information, including as part of efforts to mitigate risks of misinformation, disinformation and hate speech, and harm caused through the misuse of personal data,

인공지능 기술의 발전은 개인 데이터 오용으로 인한 오보, 허위 정보, 혐오 발언, 피해를 완화하기 위한 노력의 일환으로, 독립적이고 다원적이며 신뢰성 있는 정보 출처에 대한 접근과 그에 상응하는 데이터·미디어·정보 리터러시의 증가도 필요로 함을 인식하며,

Observing that a normative framework for AI technologies and its social implications finds its basis in international and national legal frameworks, human rights and fundamental freedoms, ethics, need for access to data, information and

knowledge, the freedom of research and innovation, human and environmental and ecosystem well-being, and connects ethical values and principles to the challenges and opportunities linked to AI technologies, based on common understanding and shared aims,

인공지능 기술과 그 사회적 함의에 대한 규범적 틀(framework)은 국제・국가적 법적 틀, 인권 및 근본적 자유, 윤리, 데이터・정보・지식 접근에 대한 필요성, 연구・혁신의 자유, 인류 및 환경・생태계의 안녕을 근간으로 하며, 공동의 합의와 목표에 기반하여 윤리적 가치 및 원칙을 인공지능 기술과 직결된 과제 및 기회에 결부시킨다는 점을 주시하여,

Also recognizing that ethical values and principles can help develop and implement rights-based policy measures and legal norms, by providing guidance with a view to the fast pace of technological development,

윤리적 가치 및 원칙이 빠른 기술 개발 속도를 고려한 지침을 제공함으로써 인권 기반의 정책 조치 및 법적 규범의 개발 및 시행을 도울 수 있음을 또한 인지하고,

Also convinced that globally accepted ethical standards for AI technologies, in full respect of international law, in particular human rights law, can play a key role in developing AI-related norms across the globe,

인공지능 기술에 관하여 국제법, 특히 인권법에 따라 전 지구적으로 용인된 윤리 기준이 전 지구적으로 인공지능 관련 규범을 개발하는 데에 핵심적인 역할을 수행할 수 있음을 확신하면서,

Bearing in mind the Universal Declaration of Human Rights (1948), the instruments of the international human rights framework, including the Convention Relating to the Status of Refugees (1951), the Discrimination (Employment and Occupation) Convention (1958), the International Convention on the Elimination of All Forms of Racial Discrimination (1965), the International Covenant on Civil and Political Rights

(1966), the International Covenant on Economic, Social and Cultural Rights (1966), the Convention on the Elimination of All Forms of Discrimination against Women (1979), the Convention on the Rights of the Child (1989), and the Convention on the Rights of Persons with Disabilities (2006), the Convention against Discrimination in Education (1960), the Convention on the Protection and Promotion of the Diversity of Cultural Expressions (2005), as well as any other relevant international instruments, recommendations and declarations,

'세계인권선언'(1948), '난민의 지위에 관한 협약'(1951), '고용 및 직업상의 차별에 관한 협약'(1958), '모든 형태의 인종 차별 철폐에 관한 국제 협약'(1965), '시민적 및 정치적 권리에 관한 국제규약'(1966), '경제적, 사회적 및 문화적 권리에 관한 국제규약'(1966), '모든 형태의 여성 차별 철폐에 관한 유엔 협약'(1979), '유엔 아동 권리 협약'(1989), '유엔 장애인 권리 협약'(2006), '교육상 차별금지에 관한 협약'(1960), '문화적 표현의 다양성 보호와 증진에 관한 유네스코 협약'(2005)을 비롯하여 기타 유관 국제 문서・권고・선언과 같은 국제 인권 틀에 관한  문서를 기억하며,

Also noting the United Nations Declaration on the Right to Development (1986); the Declaration on the Responsibilities of the Present Generations Towards Future Generations (1997); the Universal Declaration on Bioethics and Human Rights (2005); the United Nations Declaration on the Rights of Indigenous Peoples (2007); the United Nations General Assembly resolution on the review of the World Summit on the Information Society (A/RES/70/125) (2015); the United Nations General Assembly Resolution on Transforming our world: the 2030 Agenda for Sustainable Development (A/RES/70/1) (2015); the Recommendation Concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form (2015); the Declaration of Ethical Principles in relation to Climate Change (2017); the Recommendation on Science and Scientific Researchers (2017); the Internet Universality Indicators (endorsed by UNESCO's International Programme for the Development of Communication in 2018), including the ROAM principles (endorsed

by UNESCO's General Conference in 2015); the Human Rights Council's resolution on "The right to privacy in the digital age" (A/HRC/RES/42/15) (2019); and the Human Rights Council's resolution on "New and emerging digital technologies and human rights" (A/HRC/RES/41/11) (2019),

'개발권에 관한 유엔 선언'(1986), '미래 세대에 대한 현재 세대의 책임에 관한 유네스코 선언'(1997), '생명윤리와 인권 보편선언'(2005), '선주민의 권리에 관한 유엔 선언문'(2007), '정보사회 세계정상회의의 검토에 관한 유엔 총회 결의안'(A/RES/70/125) (2015), '세상의 변혁: 2030 지속가능발전 의제에 관한 유엔 총회 결의안'(A/RES/70/1) (2015), '디지털 형태를 포함한 기록유산의 보존과 접근에 관한 권고'(2015), '기후변화 윤리 원칙 선언'(2017), '과학 및 과학연구자에 관한 유네스코 권고'(2017), (2015년 유네스코 총회에서 승인받은) 'R.O.A.M. 원칙'을 비롯한 (2018년 유네스코 국제커뮤니케이션개발계획(IPDC)에서 승인받은) '인터넷 보편성 지표', '"디지털 시대의 프라이버시"에 관한 인권위원회 결의안'(A/HRC/RES/42/15)(2019), '"신기술과 인권"에 관한 인권위원회 결의안'(A/HRC/RES/41/11)에 주목하고,

Emphasizing that specific attention must be paid to LMICs, including but not limited to LDCs, LLDCs and SIDS, as they have their own capacity but have been underrepresented in the AI ethics debate, which raises concerns about neglecting local knowledge, cultural pluralism, value systems and the demands of global fairness to deal with the positive and negative impacts of AI technologies,

최빈개발도상국, 내륙개발도상국, 군소도서개발도상국을 비롯하여 (단, 이에 국한되지 않는) 중·저소득국가가 역량을 가지고 있음에도 인공지능 윤리 논의에서 발언권이 미약하다는 것은 토착 지식, 문화 다원주의, 가치 시스템 및 인공지능 기술의 긍정·부정적 영향을 다루기 위한 전 지구적 공정성이 무시된다는 우려를 불러 일으키기 때문에 이들 국가에 특별히 주목해야 함을 강조하고,

Also conscious of the many existing national policies, other frameworks and initiatives elaborated by relevant United Nations entities, intergovernmental organizations, including regional organizations, as well as those by the private sector, professional organizations, non-governmental organizations, and the scientific community, related to the ethics and regulation of AI technologies,

인공지능 기술의 윤리 및 규제와 관련된 것이라면, 유엔 유관 단체, (지역 기구를 포함하는) 정부간 국제기구를 비롯하여, 민간 부문, 전문 기관, 비정부기구, 과학계에서 고심한 기존의 많은 국가 정책, 기타 규범적 계획안을 또한 의식하며,

Further convinced that AI technologies can bring important benefits, but that achieving them can also amplify tension around innovation, asymmetric access to knowledge and technologies, including the digital and civic literacy deficit that limits the public's ability to engage in topics related to AI, as well as barriers to access to information and gaps in capacity, human and institutional capacities, barriers to access to technological innovation, and a lack of adequate physical and digital infrastructure and regulatory frameworks, including those related to data, all of which need to be addressed,

인공지능 기술은 중요한 이점을 가져올 수도 있지만, 이를 달성하는 것은 혁신을 둘러싼 긴장 상태, 인공지능에 관한 주제들에 참여할 수 있는 대중의 능력을 제한하는 디지털・시민 리터러시의 부족을 비롯한 지식・기술의 비대칭적 접근, 또한 정보 접근의 장애물 및 역량의 격차, 인간 및 제도적 역량의 격차, 기술 혁신 접근의 장애물, (데이터 등과 관련된) 적합한 유형・디지털 인프라 및 규제 틀의 부족을 증폭시킬 수 있으며, 이 모든 것은 해결되어야 함을 더욱이 확신하기에,

Underlining that the strengthening of global cooperation and solidarity, including through multilateralism, is needed to facilitate fair access to AI technologies and address the challenges that they bring to diversity and interconnectivity of cultures

and ethical systems, to mitigate potential misuse, to realize the full potential that AI can bring, especially in the area of development, and to ensure that national AI strategies are guided by ethical principles,

인공지능 기술에 대한 공정한 접근을 원활하게 하고 인공지능이 문화・윤리 시스템의 다양성 및 상호연결성에 초래하는 어려움을 다루며 잠재적 오용을 완화하고 인공지능이 가져올 수 있는 최대한의 잠재력을 특히 개발 영역에서 실현하며 각 국가의 인공지능 전략이 윤리 원칙을 따르도록 보장하기 위해, 다자주의를 비롯한 방식들을 통하여 전 지구적 협력과 연대의 강화가 필요함을 강조하고,

Taking fully into account that the rapid development of AI technologies challenges their ethical implementation and governance, as well as the respect for and protection of cultural diversity, and has the potential to disrupt local and regional ethical standards and values,

인공지능 기술의 급속한 발전은 이의 윤리적 구현・거버넌스를  비롯한 문화 다양성에 대한 존중 및 보호를 어렵게 한다는 점과 토착・지역적 윤리 기준 및 가치를 저해할 가능성을 가지고 있다는 점을 충분히 고려하여,

1. Adopts the present Recommendation on the Ethics of Artificial Intelligence;

1. 인공지능 윤리에 관한 현 권고를 채택하고,

2. Recommends that Member States apply on a voluntary basis the provisions of this Recommendation by taking appropriate steps, including whatever legislative or other measures may be required, in conformity with the constitutional practice and governing structures of each State, to give effect within their jurisdictions to the principles and norms of the Recommendation in conformity with international law, including international human rights law;

2. 회원국이 각 국의 헌법 관례 및 통치 구조에 적합하게, 입법 또는 기타 조치 요구를

비롯한 적절한 조치를 취함으로써 본 권고의 조항을 자발적으로 적용하여, 관할 구역 내에서 본 권고의 원칙 및 규범이 국제 인권법을 비롯한 국제법에 적합하게 효력을 발휘할 수 있도록 권고하며,

3. Also recommends that Member States engage all stakeholders, including business enterprises, to ensure that they play their respective roles in the implementation of this Recommendation; and bring the Recommendation to the attention of the authorities, bodies, research and academic organizations, institutions and organizations in public, private and civil society sectors involved in AI technologies, so that the development and use of AI technologies are guided by both sound scientific research as well as ethical analysis and evaluation.

3. 또한 회원국이 기업을 비롯한 모든 이해관계자를 참여시켜 그들이 본 권고의 이행에 있어 각자의 역할을 수행하도록 보장할 것을 권고하며, 인공지능 기술의 개발 및 사용이 건전한 과학 연구와 윤리적 분석・심사에 따르도록 하기 위해 인공지능 기술과 관여되어 있는 공공・민간・시민사회 부문의 당국, 단체, 연구・학술 단체, 기관, 조직이 본 권고에 주목할 수 있도록 할 것을 권고한다.

## I. SCOPE OF APPLICATION

## I. 적용 범위

1. This Recommendation addresses ethical issues related to the domain of Artificial Intelligence to the extent that they are within UNESCO's mandate.

1. 본 권고는 유네스코의 권한 내에서 인공지능 분야와 관련된 윤리적 사안을 다룬다.

It approaches AI ethics as a systematic normative reflection, based on a holistic, comprehensive, multicultural and evolving framework of interdependent values,

principles and actions that can guide societies in dealing responsibly with the known and unknown impacts of AI technologies on human beings, societies and the environment and ecosystems, and offers them a basis to accept or reject AI technologies.

본 권고는 인공지능 시스템이 인류, 사회, 환경 및 생태계에 미칠 수 있는 알려진 영향과 알려지지 않은 영향을 충실히 다루는 데에 있어 인간 사회의 지침이 될 수 있는 독립적인 가치 · 원칙 · 조치의 총체적, 종합적, 다문화적, 발전적 틀에 기반하여 인공지능 윤리를 하나의 체계 있는 규범적 고찰로서 접근하며, 이들에게 인공지능 기술을 수용하거나 거부할 수 있는 근간을 제공한다.

It considers ethics as a dynamic basis for the normative evaluation and guidance of AI technologies, referring to human dignity, well-being and the prevention of harm as a compass and as rooted in the ethics of science and technology.

본 권고는 윤리를 인공지능 기술에 대한 규범적 평가 및 지침을 위한 역동적인 기반으로 여기며, 인간 존엄성, 복지, 피해 방지를 나침반이자 과학 · 기술 윤리의 근본으로 간주한다.

2. This Recommendation does not have the ambition to provide one single definition of AI, since such a definition would need to change over time, in accordance with technological developments.

2. 인공지능의 정의는 기술의 진보에 따라 변화될 필요가 있기에, 본 권고는 인공지능의 단일 정의를 제공하는 데에 뜻을 두지 않는다.

Rather, its ambition is to address those features of AI systems that are of central ethical relevance.

다만, 본 권고는 인공지능 시스템의 특징 중에서 윤리와 핵심적으로 관련있는 부분을 다루는 데에 목적이 있다.

Therefore, this Recommendation approaches AI systems as systems which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control.

따라서, 본 권고는 인공지능 시스템을 일반적으로 추론, 학습, 인식, 예측, 계획, 통제를 비롯하여 지적 행위와 유사한 방식으로 데이터 및 정보를 처리할 능력이 있는 시스템이라는 측면으로 접근한다.

Three elements have a central place in this approach:

다음 세 가지 요소가 이러한 접근법의 중요한 위치를 차지한다.

(a) AI systems are information-processing technologies that integrate models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments.

(a) 인공지능 시스템은 현실・가상 환경에서 예측 및 의사 결정과 같은 결과를 도출하는 인지 과제의 학습・수행 능력을 생성하는 모델 및 알고리즘을 통합한 정보 처리 기술이다.

AI systems are designed to operate with varying degrees of autonomy by means of knowledge modelling and representation and by exploiting data and calculating correlations.

인공지능 시스템은 지식 모델링・표현의 사용 또 데이터 및 상관관계 계산에 따라서 자율성의 정도가 변화하며 동작하도록 설계되어 있다.

AI systems may include several methods, such as but not limited to:

인공지능 시스템에는 아래의 몇 가지 항목들과 같이 (단, 이에 국한되지 않는) 몇 가지 방법론이 있을 수 있다.

(i) machine learning, including deep learning and reinforcement learning;

(i) 심층학습 및 강화학습을 비롯한 기계학습.

(ii) machine reasoning, including planning, scheduling, knowledge representation and reasoning, search, and optimization.

(ii) 계획, 일정 관리, 지식 표현 및 추론, 검색, 최적화를 비롯한 기계 추론.

AI systems can be used in cyber-physical systems, including the Internet of things, robotic systems, social robotics, and human-computer interfaces, which involve control, perception, the processing of data collected by sensors, and the operation of actuators in the environment in which AI systems work.

인공지능 시스템은 이것이 작동하는 환경에서의 제어, 인지, 센서 데이터 처리, 액추에이터 조작이 수반된 사물 인터넷(IoT), 로봇 공학, 소셜 로봇, 휴먼컴퓨터인터페이스(HCI)를 비롯한 가상물리시스템에서 사용될 수 있다.

(b) Ethical questions regarding AI systems pertain to all stages of the AI system life cycle, understood here to range from research, design and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination.

인공지능 시스템 수명 주기를 관리, 운영, 매매, 재무, 모니터링 및 심사, 검증, 종료, 해체, 폐기를 비롯하여 연구, 설계, 개발로부터 배포, 사용에 이르는 과정까지 포함하는 것으로 이해할 때, 인공지능 시스템에 관한 윤리적 질문은 각 단계마다 존재한다.

In addition, AI actors can be defined as any actor involved in at least one stage of the AI system life cycle, and can refer both to natural and legal persons, such as researchers, programmers, engineers, data scientists, end-users, business enterprises, universities and public and private entities, among others.

또한, 인공지능 행위 주체는 이러한 인공지능 시스템 수명 주기에서 적어도 하나의 단계에 관련되어 있는 자로 정의될 수 있고, 특히, 연구자, 프로그래머, 엔지니어, 데이터 과학, 일반 사용자, 기업, 대학, 공공・민간 기관 등과 같은 일반인 및 법인을 지칭할 수 있다.

(c) AI systems raise new types of ethical issues that include, but are not limited to, their impact on decision-making, employment and labour, social interaction, health care, education, media, access to information, digital divide, personal data and consumer protection, environment, democracy, rule of law, security and policing, dual use, and human rights and fundamental freedoms, including freedom of expression, privacy and non-discrimination.

(c) 인공지능 시스템은 인공지능이 의사 결정, 고용 및 노동, 사회적 상호작용, 건강관리, 교육, 미디어, 정보 접근성, 디지털 격차, 개인 데이터 및 소비자 보호, 환경, 민주주의, 법치주의, 보안 및 치안유지, 군민 양용(dual use), 그리고 표현의 자유, 프라이버시, 비차별을 포함하는 인권 및 근본적 자유에 미치는 영향을 비롯하여 (단, 이에 국한되지 않는) 새로운 유형의 여러 윤리 문제를 유발한다.

Furthermore, new ethical challenges are created by the potential of AI algorithms to reproduce and reinforce existing biases, and thus to exacerbate already existing forms of discrimination, prejudice and stereotyping.

이에 더해, 기존의 편향을 재생산・강화하여 기존 형태의 차별, 편견, 고정 관념을 악화시킬 수 있는 인공지능 알고리즘의 잠재성으로 인해 새로운 윤리적 문제도 제기된다.

Some of these issues are related to the capacity of AI systems to perform tasks which previously only living beings could do, and which were in some cases even limited to human beings only.

이 중 일부 사안들은 과거에는 생명체만이 할 수 있었거나 인간에게만 국한되었던 과제를 인공지능 시스템이 수행할 능력을 갖추게 되었다는 것과 관련이 있다.

These characteristics give AI systems a profound, new role in human practices and society, as well as in their relationship with the environment and ecosystems, creating a new context for children and young people to grow up in, develop an understanding of the world and themselves, critically understand media and information, and learn to make decisions.

이러한 특징들은 인간 관습 및 사회에서 뿐만 아니라 인공지능과 환경 및 생태계와의 관계에서도 인공지능 시스템에 심오하고 새로운 역할을 부여함으로써, 어린이와 청년들이 성장하면서 세상과 자신에 대한 이해를 심화시키며 미디어 및 정보를 비판적으로 이해하고, 의사 결정하는 법을 배울 수 있도록 하는 새로운 맥락을 만들어낸다.

In the long term, AI systems could challenge humans' special sense of experience and agency, raising additional concerns about, inter alia, human self-understanding, social, cultural and environmental interaction, autonomy, agency, worth and dignity.

장기적 관점으로 보면, 인공지능 시스템은 인간 고유의 경험과 의식에 도전하여, 특히 인간의 자기 인식, 사회・문화・환경적 상호작용, 자율성, 활동(agency), 가치, 존엄성에 대한 추가적인 우려를 낳을 수 있다.

3. This Recommendation pays specific attention to the broader ethical implications of AI systems in relation to the central domains of UNESCO: education, science, culture, and communication and information, as explored in the 2019 Preliminary Study on the Ethics of Artificial Intelligence by the UNESCO World Commission on Ethics of Scientific Knowledge and Technology (COMEST):

3. 본 권고는 2019년 유네스코 세계과학지식기술윤리위원회(COMEST)가 '인공지능 윤리에 대한 예비 연구'에서 강구하였듯이 교육, 과학, 문화, 정보통신과 같은 유네스코의 핵심 영역(이하)과 관련된 인공지능 시스템의 광범위한 윤리적 함의에 각별히 주의를 기울인다.

(a) Education, because living in digitalizing societies requires new educational practices, ethical reflection, critical thinking, responsible design practices and new skills, given the implications for the labour market, employability and civic participation.

(a) 교육: 노동 시장, 고용, 시민 참여에 대한 파급 효과를 고려할 때, 디지털화되어가는 사회에서 사는 것은 새로운 교육 관행, 윤리적 성찰, 비판적 사고, 책임 있는 설계 관행, 새로운 숙련을 요구한다.

(b) Science, in the broadest sense and including all academic fields from the natural sciences and medical sciences to the social sciences and humanities, as AI technologies bring new research capacities and approaches, have implications for our concepts of scientific understanding and explanation, and create a new basis for decision-making.

(b) 과학: 가장 넓은 의미에서, 그리고 자연과학 및 의학으로부터 사회과학 및 인문학에 이르기까지 모든 학문 분야를 비롯하여 가장 넓은 의미에서, 인공지능 기술은 새로운 연구 역량 · 접근법을 제공하고 과학적 이해 · 설명에 대한 우리의 개념에 영향을 미치며, 의사 결정을 위한 새로운 토대를 마련한다.

(c) Cultural identity and diversity, as AI technologies can enrich cultural and creative industries, but can also lead to an increased concentration of supply of cultural content, data, markets and income in the hands of only a few actors, with potential negative implications for the diversity and pluralism of languages, media, cultural expressions, participation and equality.

(c) 문화 정체성 및 다양성: 인공지능 기술은 문화 · 창조적 산업을 풍요롭게 할 수도 있지만, 문화 컨텐츠 · 데이터 · 시장 · 수익의 공급이 소수의 행위 주체에게만 편중되어 언어 · 미디어 · 문화적 표현 · 참여 · 평등의 다양성 및 다원성에 부정적 영향을 줄 잠재력도 가지고 있다.

(d) Communication and information, as AI technologies play an increasingly important role in the processing, structuring and provision of information; the issues of automated journalism and the algorithmic provision of news and moderation and curation of content on social media and search engines are just a few examples raising issues related to access to information, disinformation, misinformation, hate speech, the emergence of new forms of societal narratives, discrimination, freedom of expression, privacy and media and information literacy, among others.

(d) 의사소통 및 정보: 인공지능 기술은 정보 처리・구조화・제공에 점차 중요한 역할을 하고 있고, 자동화된 기사, 알고리즘 기반 뉴스 제공, 소셜 미디어 및 검색 엔진의 콘텐츠 조정・선별 같은 문제들은 정보 접근, 허위 정보, 오보, 혐오 발언, 새로운 형태의 사회적 내러티브 출현, 차별, 표현의 자유, 프라이버시, 미디어・정보 리터러시 등과 관련된 문제를 일으키는 일부 예시에 불과하다.

4. This Recommendation is addressed to Member States, both as AI actors and as authorities responsible for developing legal and regulatory frameworks throughout the entire AI system life cycle, and for promoting business responsibility.

4. 본 권고는 인공지능 시스템 수명 주기 전반에 걸쳐 법적・규제적 틀을 개발하고 기업의 의무를 촉진시켜야 할 행위 주체이자 관계 당국인 회원국을 대상으로 한다.

It also provides ethical guidance to all AI actors, including the public and private sectors, by providing a basis for an ethical impact assessment of AI systems throughout their life cycle.

또한, 본 권고는 인공지능 시스템 수명 주기 내내 인공지능 시스템에 대한 윤리영향평가의 토대를 제공함으로써 공공・민간 부문을 비롯한 모든 인공지능 행위 주체에게 윤리 지침을 제공한다.

## II. AIMS AND OBJECTIVES

## II. 목적 및 목표

5. This Recommendation aims to provide a basis to make AI systems work for the good of humanity, individuals, societies and the environment and ecosystems, and to prevent harm.

5. 본 권고는 인공지능 시스템이 인류 · 개인 · 사회 · 환경 및 생태계의 이익을 위해 작동하며 피해를 방지하도록 하는 토대를 제공하는 것을 목적으로 한다.

It also aims at stimulating the peaceful use of AI systems.

또한, 본 권고는 인공지능 시스템의 평화적 사용을 독려한다.

6. In addition to the existing ethical frameworks regarding AI around the world, this Recommendation aims to bring a globally accepted normative instrument that focuses not only on the articulation of values and principles, but also on their practical realization, via concrete policy recommendations, with a strong emphasis on inclusion issues of gender equality and protection of the environment and ecosystems.

6. 본 권고는 전 세계 도처에 있는 인공지능에 관한 기존의 윤리적 틀을 제공하는 것에 그치지 않고, 단지 가치와 원칙의 표현뿐만 아니라, 성평등 · 환경 및 생태계 보호 문제의 수용에 대한 강조와 함께 구체적인 정책 권장사항을 통한 가치 및 원칙의 실현에 초점을 맞추는 전 지구적으로 용인되는 규범적 도구를 제공하는 것을 목적으로 한다.

7. Because the complexity of the ethical issues surrounding AI necessitates the cooperation of multiple stakeholders across the various levels and sectors of international, regional and national communities, this Recommendation aims to enable stakeholders to take shared responsibility based on a global and

intercultural dialogue.

7. 인공지능을 둘러싼 윤리적 문제의 복잡성은 다양한 층위・부문의 국제・지역・국가 공동체에 걸쳐 존재하는 이해관계자들의 협력을 필요로 하므로, 본 권고는 이해관계자들이 전 지구적・문화 간 대화를 통하여 의무를 분담할 수 있도록 하는 것을 목적으로 한다.

8. The objectives of this Recommendation are:

8. 본 권고의 구체적인 목표는 아래와 같다.

(a) to provide a universal framework of values, principles and actions to guide States in the formulation of their legislation, policies or other instruments regarding AI, consistent with international law;

(a) 국가가 인공지능에 관한 법률, 정책, 또는 기타 도구를 조성함에 있어 국제법에 부합할 수 있도록 지침이 되어주는 가치・원칙・행위에 대한 보편적인 틀을 제공하는 것.

(b) to guide the actions of individuals, groups, communities, institutions and private sector companies to ensure the embedding of ethics in all stages of the AI system life cycle;

(b) 인공지능 시스템 수명 주기의 각 단계에 윤리가 확실히 내재되도록  개인, 집단, 공동체, 기관, 민간 기업의 행동 지침을 제공하는 것.

(c) to protect, promote and respect human rights and fundamental freedoms, human dignity and equality, including gender equality; to safeguard the interests of present and future generations; to preserve the environment, biodiversity and ecosystems; and to respect cultural diversity in all stages of the AI system life cycle;

(c) 인공지능 시스템 수명 주기의 각 단계에서 인권 및 근본적 자유, 인간 존엄성, 성평등을 비롯한 평등을 보호, 촉진, 존중하고, 현재・미래 세대의 이익을 보장하며, 환경, 생물다양성, 생태계를 보존하고, 문화 다양성을 존중하는 것.

(d) to foster multi-stakeholder, multidisciplinary and pluralistic dialogue and consensus building about ethical issues relating to AI systems;

(d) 인공지능 시스템과 관련된 윤리 문제에 관해 다자간 · 다학문간 · 다원주의적 대화 및 합의 도출을 촉진하는 것.

(e) to promote equitable access to developments and knowledge in the field of AI and the sharing of benefits, with particular attention to the needs and contributions of LMICs, including LDCs, LLDCs and SIDS.

(e) 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국을 비롯한 중 · 저소득국가의 필요 및 참여에 각별히 주의하여, 인공지능 분야의 발전 및 지식에의 공평한 접근과 이익 공유를 장려하는 것.

## III. VALUES AND PRINCIPLES

## III. 가치 및 원칙

9. The values and principles included below should be respected by all actors in the AI system life cycle, in the first place and, where needed and appropriate, be promoted through amendments to the existing and elaboration of new legislation, regulations and business guidelines.

9. 이하의 가치 및 원칙은 우선 인공지능 시스템 수명 주기 내의 모든 행위 주체에 의해 존중되어야 하고, 필요하고 적절한 경우라면 기존 및 새로운 법률 · 규제 · 기업 지침의 수정 및 정교화를 통해 고취되어야 한다.

This must comply with international law, including the United Nations Charter and Member States' human rights obligations, and should be in line with internationally agreed social, political, environmental, educational, scientific and economic

sustainability objectives, such as the United Nations Sustainable Development Goals (SDGs).

이러한 존중은 유엔 헌장 및 회원국의 인권 준수 의무를 비롯하여 국제법을 준수해야 하며, 유엔 지속가능발전목표(SDGs)와 같은 국제적으로 합의된 사회・정치・환경・교육・과학・경제적 지속가능성 목표에 의거해야 한다.

10. Values play a powerful role as motivating ideals in shaping policy measures and legal norms.

10. 가치는 동기부여적 이상으로서 정책 조치와 법적 규범을 형성하는 데 있어 강력한 역할을 한다.

While the set of values outlined below thus inspires desirable behaviour and represents the foundations of principles, the principles unpack the values underlying them more concretely so that the values can be more easily operationalized in policy statements and actions.

이하 개관한 가치들이 바람직한 행동을 유도하고 원칙의 토대를 마련한다면, 원칙은 그 기저에 있는 가치를 구체적으로 풀어내어 가치가 정책 제시・조치 과정에서 더 쉽게 적용될 수 있도록 한다.

11. While all the values and principles outlined below are desirable per se, in any practical contexts, there may be tensions between these values and principles.

11. 이하 개관한 모든 가치 및 원칙은 그 자체로 바람직하지만, 실제 상황에서는 이러한 가치 및 원칙 사이에 긴장 상태가 있을 수 있다.

In any given situation, a contextual assessment will be necessary to manage potential tensions, taking into account the principle of proportionality and in

compliance with human rights and fundamental freedoms.

어떠한 상황이 주어졌을 때, 비례성 원칙을 고려하고 인권 및 근본적 자유를 지키면서 잠재적 긴장 상태를 해결하기 위해서는 주변 맥락 진단이 필수적일 것이다.

In all cases, any possible limitations on human rights and fundamental freedoms must have a lawful basis, and be reasonable, necessary and proportionate, and consistent with States' obligations under international law.

어떠한 경우에도, 인권 및 근본적 자유를 제한할 수 있는 가능성은 반드시 법률적 토대가 있어야 하며 합리적 · 필수적 · 비례적이어야 하고 국제법 하에서 회원국의 인권 준수 의무에 부합해야 한다.

To navigate such scenarios judiciously will typically require engagement with a broad range of appropriate stakeholders, making use of social dialogue, as well as ethical deliberation, due diligence and impact assessment.

이러한 시나리오들을 신중하게 탐구하는 데에는 일반적으로 적법한 이해관계자의 폭넓은 참여가 필요할 것이며, 사회적 대화와 더불어 윤리적 심의 · 실사(due diligence) · 영향평가를 활용해야 할 것이다.

12. The trustworthiness and integrity of the life cycle of AI systems is essential to ensure that AI technologies will work for the good of humanity, individuals, societies and the environment and ecosystems, and embody the values and principles set out in this Recommendation.

12. 인공지능 시스템 수명 주기에 대한 신뢰성과 무결성(integrity)은 인공지능 기술이 인류, 개인, 사회, 환경 및 생태계에 유익하게끔 작동하고 본 권고에서 제시한 가치 및 원칙을 구현하게 보장하는 데에 필수적이다.

People should have good reason to trust that AI systems can bring individual and shared benefits, while adequate measures are taken to mitigate risks.

인공지능 시스템이 개별・공유 혜택을 가져다주며 위험을 완화하기 위한 적절한 조치 또한 취해진다고 믿을 만한 좋은 근거가 있어야 한다.

An essential requirement for trustworthiness is that, throughout their life cycle, AI systems are subject to thorough monitoring by the relevant stakeholders as appropriate.

신뢰성의 필수 요건은 인공지능 시스템 수명 주기 전 영역에서 인공지능 시스템이 적절히 관련 이해관계자의 철저한 모니터링 하에 있어야 한다는 것이다.

As trustworthiness is an outcome of the operationalization of the principles in this document, the policy actions proposed in this Recommendation are all directed at promoting trustworthiness in all stages of the AI system life cycle.

신뢰성은 이곳에서 제시된 원칙들의 적용 결과이므로, 본 권고가 제시하는 정책 행동은 인공지능 시스템 수명 주기 전 단계에서의 신뢰성 증진을 겨냥한다.

## III.1 VALUES

## III.1. 가치

Respect, protection and promotion of human rights and fundamental freedoms and human dignity

인권 및 근본적 자유, 인간 존엄성의 존중, 보호, 증진

13. The inviolable and inherent dignity of every human constitutes the foundation for the universal, indivisible, inalienable, interdependent and interrelated system of

human rights and fundamental freedoms.

13. 모든 사람에게 있는 불가침의 고유한 권리는 인권 및 근보적 자유의 보편적이며 불가분이고 빼앗을 수 없으며 상호의존적이고 상호 밀접한 시스템의 토대를 구성한다.

Therefore, respect, protection and promotion of human dignity and rights as established by international law, including international human rights law, is essential throughout the life cycle of AI systems.

따라서, 국제인권법을 비롯한 국제법으로 인정받은 인간 존엄성 및 인권의 존중・보호・증진은 인공지능 시스템의 수명 주기 전 영역에서 필수적이다.

Human dignity relates to the recognition of the intrinsic and equal worth of each individual human being, regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.

인간 존엄성은 인종, 피부색, 혈통, 성, 나이, 언어, 종교, 정치적 견해, 국가・민족・사회적 출신, 출생 시 경제・사회적 조건, 장애, 또는 다른 근거와 관계없이, 인간 개개인의 내재적이고 동등한 가치를 인정함과 관련 있다.

14. No human being or human community should be harmed or subordinated, whether physically, economically, socially, politically, culturally or mentally during any phase of the life cycle of AI systems.

14. 어떤 사람 또는 공동체도 인공지능 시스템 수명 주기의 모든 단계에서 가운데 점 다른 스타일과 통일해주세 피해를 입거나 종속되어서는 안 된다.

Throughout the life cycle of AI systems, the quality of life of human beings should be enhanced, while the definition of “quality of life” should be left open to individuals or groups, as long as there is no violation or abuse of human rights

and fundamental freedoms, or the dignity of humans in terms of this definition.

인공지능 시스템의 수명 주기 전 영역에서 모든 사람의 삶의 질은 향상되어야 하는데, '삶의 질'의 정의는 그 정의에 의해서 인권 및 근본적 자유, 인간 존엄성이 침해당하거나 남용되지 않는 한 개인 또는 집단에게 열려 있어야 한다.

15. Persons may interact with AI systems throughout their life cycle and receive assistance from them, such as care for vulnerable people or people in vulnerable situations, including but not limited to children, older persons, persons with disabilities or the ill.

15. 사람은 인공지능 시스템 수명 주기 전 영역에서 인공지능과 상호작용할 수도 있으며, 어린이, 노인, 장애인, 환자 등을 비롯하여 (단, 이에 국한되지 않는) 취약 계층 또는 취약한 상황에 처한 사람에게 제공하는 돌봄과 같은 보조를 받을 수 있다.

Within such interactions, persons should never be objectified, nor should their dignity be otherwise undermined, or human rights and fundamental freedoms violated or abused.

그러한 상호작용 속에서 모든 사람은 대상화되거나, 아니면 존엄성이 훼손되거나, 인권 및 근본적 자유가 침해・남용 당하지 않아야 한다.

16. Human rights and fundamental freedoms must be respected, protected and promoted throughout the life cycle of AI systems.

16. 인권과 근본적 자유는 인공지능 시스템 수명 주기 전 영역에서 반드시 존중, 보호, 증진되어야 한다.

Governments, private sector, civil society, international organizations, technical communities and academia must respect human rights instruments and

frameworks in their interventions in the processes surrounding the life cycle of AI systems.

정부, 민간 부문, 시민사회, 국제기구, 기술 공동체, 학계는 인권 보호 도구 및 틀이 인공지능 시스템 수명 주기를 둘러싼 과정에 개입할 때, 이를 반드시 존중해야 한다.

New technologies need to provide new means to advocate, defend and exercise human rights and not to infringe them.

신기술은 인권을 침해하지 않고 지지, 보호, 실행하는 새로운 수단을 제공해야 한다.

Environment and ecosystem flourishing

환경 및 생태계의 번영

17. Environmental and ecosystem flourishing should be recognized, protected and promoted through the life cycle of AI systems.

17. 환경 및 생태계의 번영은 인공지능 시스템 수명 주기 전 영역에서 인지, 보호, 증진되어야 한다.

Furthermore, environment and ecosystems are the existential necessity for humanity and other living beings to be able to enjoy the benefits of advances in AI.

또한, 환경 및 생태계는 인류와 기타 생명체가 인공지능 발전의 혜택을 영위할 수 있기 위한 필수 존재이다.

18. All actors involved in the life cycle of AI systems must comply with applicable international law and domestic legislation, standards and practices, such as precaution, designed for environmental and ecosystem protection and restoration, and sustainable development.

18. 인공지능 시스템 수명 주기에 관여되어 있는 모든 행위 주체는 환경 및 생태계 보호·회복을 위해 고안된 예방 조치 및 지속가능한 성장과 같은, 해당 국제 및 국내법·기준·관행을 따라야 한다.

They should reduce the environmental impact of AI systems, including but not limited to its carbon footprint, to ensure the minimization of climate change and environmental risk factors, and prevent the unsustainable exploitation, use and transformation of natural resources contributing to the deterioration of the environment and the degradation of ecosystems.

인공지능 행위 주체는 기후변화와 환경 위험 요인들을 최소화를 보장하고 환경 및 생태계 파괴에 일조하는 지속가능하지 않은 착취·천연 자원의 사용 및 변환을 막기 위해, 탄소 발자국을 비롯하여 (단, 이에 국한되지 않는) 인공지능 시스템의 환경적 영향을 축소하여야 한다.

Ensuring diversity and inclusiveness

다양성 및 포용성 보장

19. Respect, protection and promotion of diversity and inclusiveness should be ensured throughout the life cycle of AI systems, consistent with international law, including human rights law.

19. 인공지능 시스템의 수명 주기 전 영역에서 인권법을 비롯한 국제법과 일치하는 선에서 다양성 및 포용성의 존중·보호·증진은 보장되어야 한다.

This may be done by promoting active participation of all individuals or groups regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.

이는 인종, 피부색, 혈통, 성, 나이, 언어, 종교, 정치적 견해, 국가・민족・사회적 출신, 출생 시 경제・사회적 조건, 장애, 또는 다른 근거와 관계없이 개인 또는 집단의 적극적 참여를 장려함으로써 이루어질 수 있다.

20. The scope of lifestyle choices, beliefs, opinions, expressions or personal experiences, including the optional use of AI systems and the co-design of these architectures should not be restricted during any phase of the life cycle of AI systems.

20. 인공지능 시스템의 선택적 사용 및 이러한 구조체(architecture)의 공동 설계를 비롯하여 생활 방식・신념・견해・표현・의사표현・개인적 경험의 범주는 인공지능 시스템 수명 주기의 어떤 단계에서도 제한되어서는 안 된다.

21. Furthermore, efforts, including international cooperation, should be made to overcome, and never take advantage of, the lack of necessary technological infrastructure, education and skills, as well as legal frameworks, particularly in LMICs, LDCs, LLDCs and SIDS, affecting communities.

21. 이에 더해, 특히 중・저소득국가, 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국에서는, 법적 틀의 부족을 극복하고 절대 악용하지 않기 위해서 국제적 협력을 비롯한 노력이 이루어져야 한다.

Living in peaceful, just and interconnected societies

조화롭고 공정하며 상호연결된 사회에서의 삶

22. AI actors should play a participative and enabling role to ensure peaceful and just societies, which is based on an interconnected future for the benefit of all, consistent with human rights and fundamental freedoms.

22. 모든 행위 주체는 평화롭고 공정한 사회를 보장하기 위해 참여적이며 상조하는 (enabling) 역할을 수행해야 하는데, 이는 인권 및 근본적 자유에 부합하도록 모든 사람의 이익이 보장되는 상호연결된 미래를 위한 것이다.

The value of living in peaceful and just societies points to the potential of AI systems to contribute throughout their life cycle to the interconnectedness of all living creatures with each other and with the natural environment.

조화롭고 공정한 사회에서의 삶의 가치는 인공지능 시스템 수명 주기 전 영역에서 모든 생명체 간의 상호연결 및 생명체와 자연환경 간의 상호연결에 기여할 수 있는 인공지능 시스템의 잠재력에서 나온다.

23. The notion of humans being interconnected is based on the knowledge that every human belongs to a greater whole, which thrives when all its constituent parts are enabled to thrive.

23. 상호연결된 사람이라는 개념은 모든 사람이 더 큰 하나의 완전체에 속하게 된다는 지식에 기반을 두고 있으며, 이는 그 구성원이 다른 사람이 번영할 수 있으면 그 자신도 번영하게 됨을 의미한다.

Living in peaceful, just and interconnected societies requires an organic, immediate, uncalculated bond of solidarity, characterized by a permanent search for peaceful relations, tending towards care for others and the natural environment in the broadest sense of the term.

평화롭고 공정하며 상호연결된 사회에서의 삶은 유기적이며 친밀하고 계산적이지 않은 일체의 유대감을 필요로 하며, 가장 넓은 의미에서 타인 및 자연 환경에 대한 관심을 지향함으로써 평화적 관계를 영구적으로 추구한다는 특징이 있다.

24. This value demands that peace, inclusiveness and justice, equity and interconnectedness should be promoted throughout the life cycle of AI systems, in so far as the processes of the life cycle of AI systems should not segregate, objectify or undermine freedom and autonomous decision-making as well as the safety of human beings and communities, divide and turn individuals and groups against each other, or threaten the coexistence between humans, other living beings and the natural environment.

24. 인공지능 시스템 수명 주기 프로세스가 자유 및 자율적 의사 결정과 인간 및 공동체의 안전을 분리, 대상화・약화하거나 개인 및 집단이 분열・반목하게 하거나 인간, 다른 생물, 자연환경의 공존을 위협하지 않는 한, 이 가치는 인공지능 시스템 수명 주기 전 영역에서 평화, 포용성 및 공정성, 형평성 및 상호연결성이 증진되도록 해야 한다.

## III.2 PRINCIPLES

## III.2. 원칙

Proportionality and Do No Harm

과잉금지의 원칙 및 위해 금지('Do No Harm') 원칙

25. It should be recognized that AI technologies do not necessarily, per se, ensure human and environmental and ecosystem flourishing.

25. 인공지능 기술이 그 자체만으로 인간, 환경 및 생태계의 번영을 반드시 보장하는 것은 아니라는 점이 인지되어야 한다.

Furthermore, none of the processes related to the AI system life cycle shall exceed what is necessary to achieve legitimate aims or objectives and should be appropriate to the context.

이에 더해, 인공지능 시스템 수명 주기와 관련된 어떤 프로세스도 적법한 목적·목표를 달성하는 데에 필요한 수준을 초과할 수 없으며 주변 맥락에 적합해야 한다.

In the event of possible occurrence of any harm to human beings, human rights and fundamental freedoms, communities and society at large or the environment and ecosystems, the implementation of procedures for risk assessment and the adoption of measures in order to preclude the occurrence of such harm should be ensured.

인간, 인권 및 근본적 자유, 공동체 및 사회 전반, 환경 및 생태계가 피해를 입을 수 있는 경우, 위험 평가 절차가 실행되고 그러한 피해를 방지할 수 있는 조치의 채택이 보장되어야 한다.

26. The choice to use AI systems and which AI method to use should be justified in the following ways: (a) the AI method chosen should be appropriate and proportional to achieve a given legitimate aim; (b) the AI method chosen should not infringe upon the foundational values captured in this document, in particular, its use must not violate or abuse human rights; and (c) the AI method should be appropriate to the context and should be based on rigorous scientific foundations.

26. 인공지능 시스템의 사용 여부 및 어떤 인공지능 기법을 사용할 지는 다음과 같은 방식에 따라 정당성을 가져야 한다. (a) 선택된 인공지능 기법은 주어진 타당한 목표를 달성하기에 바람직하고 비례적이어야 한다. (b) 선택된 인공지능 기법은 본 권고가 지니는 근본적인 가치에 대해 침해가 없어야 한다. 특히, 이의 사용은 절대로 인권을 침해하거나 남용하지 않아야 한다. (c) 인공지능 기법은 사용 맥락에 적합해야 하고 철저히 과학적 근거에 토대를 두어야 한다.

In scenarios where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life and death decisions, final human

determination should apply.

어떤 결정이 돌이킬 수 없거나 번복하기 어려운 영향력을 가지는 것으로 판단되는 경우 또는 생사 결정을 논하는 상황에는 사람이 최종적인 결정을 내려야 한다.

In particular, AI systems should not be used for social scoring or mass surveillance purposes.

특히, 인공지능 시스템은 사회적 점수 평가(social scoring) 또는 대중 감시 목적으로는 사용되지 않아야 한다.

Safety and security

안전 및 보안

27. Unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be avoided and should be addressed, prevented and eliminated throughout the life cycle of AI systems to ensure human, environmental and ecosystem safety and security.

27. 인간, 환경 및 생태계의 안전과 보안을 보장하기 위해, 원치 않은 피해(안전 위험)와 공격에 대한 취약성(보안 위험)은 인공지능 수명 주기 내내 지양되어야 하며 해결, 예방, 제거되어야 한다.

Safe and secure AI will be enabled by the development of sustainable, privacy-protective data access frameworks that foster better training and validation of AI models utilizing quality data.

안전하고 보안이 철저한 인공지능은 지속가능하며 프라이버시가 보장되는 데이터 접근 틀의 개발을 통해 가능한데, 이는 양질의 데이터를 활용함으로써 인공지능 모델이 더 뛰어난 학습・검증을 할 수 있도록 한다.

### Fairness and non-discrimination

### 공정성 및 차별 금지

28. AI actors should promote social justice and safeguard fairness and non-discrimination of any kind in compliance with international law.

28. 인공지능 행위 주체는 사회 정의를 증진해야 하며 국제법에 따라 공정성 및 어떤 종류의 차별 금지도 수호해야 한다.

This implies an inclusive approach to ensuring that the benefits of AI technologies are available and accessible to all, taking into consideration the specific needs of different age groups, cultural systems, different language groups, persons with disabilities, girls and women, and disadvantaged, marginalized and vulnerable people or people in vulnerable situations.

이는 각 연령 집단, 문화 체계, 언어 집단, 장애인, 소녀 및 여성, 빈민・소외・취약 계층 또는 취약한 상황에 처한 사람의 특정 필요까지도 고려함으로써 인공지능 기술의 혜택이 모든 사람에게 미치고 접근될 수 있도록 하는 포용적 접근법을 내포한다.

Member States should work to promote inclusive access for all, including local communities, to AI systems with locally relevant content and services, and with respect for multilingualism and cultural diversity.

회원국은 지역 공동체를 비롯하여 누구든지 그 지역에 관련 있는 컨텐츠 및 서비스를 가지고 있는, 다언어성 및 문화 다양성을 존중하는 인공지능 시스템에 접근할 수 있도록 장려해야 한다.

Member States should work to tackle digital divides and ensure inclusive access to and participation in the development of AI.

회원국은 디지털 격차를 해결하며 인공지능 개발에서의 포용적 접근 및 참여를 보장하기 위해 노력해야 한다.

At the national level, Member States should promote equity between rural and urban areas, and among all persons regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds, in terms of access to and participation in the AI system life cycle.

국가적 차원에서, 인공지능 수명 주기에 대한 접근 및 참여에 대하여 회원국은 농촌과 도시 지역 간에, 그리고 사람들 사이에서 인종, 피부색, 혈통, 성, 나이, 언어, 종교, 정치적 견해, 국가 · 민족 · 사회적 출신, 출생 시 경제 · 사회적 조건, 장애, 또는 다른 근거에 상관없이 형평성을 증진해야 한다.

At the international level, the most technologically advanced countries have a responsibility of solidarity with the least advanced to ensure that the benefits of AI technologies are shared such that access to and participation in the AI system life cycle for the latter contributes to a fairer world order with regard to information, communication, culture, education, research and socio-economic and political stability.

국제적 차원에서는, 기술선도국은 기술후발국과 연대함으로써, 인공지능 기술의 혜택이 공유되어 기술후발국의 인공지능 시스템 수명 주기에 대한 접근 및 참여가 정보, 통신, 문화, 교육, 연구, 사회-경제적 · 정치적 안정에 있어 더 공정한 세계 질서에 기여하도록 보장할 의무가 있다.

29. AI actors should make all reasonable efforts to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system to ensure fairness of such systems.

29. 인공지능 행위 주체는 차별적이거나 편향된 응용 및 결과물이 강화되거나 영속하는 것을 최소화하기 위해 모든 합리적인 노력을 기울여서 인공지능 시스템 수명 주기 전 영역에서 이 시스템의 공정성을 보장해야 한다.

Effective remedy should be available against discrimination and biased algorithmic determination.

차별 및 편향된 알고리즘 결정에 대해서 사용할 수 있는 효과적인 방안이 있어야 한다.

30. Furthermore, digital and knowledge divides within and between countries need to be addressed throughout an AI system life cycle, including in terms of access and quality of access to technology and data, in accordance with relevant national, regional and international legal frameworks, as well as in terms of connectivity, knowledge and skills and meaningful participation of the affected communities, such that every person is treated equitably.

30. 이에 더해, 모든 사람이 평등하게 대우받을 수 있도록, 국가 내, 국가 간 디지털・지식 격차는 관련 국가・지역・국제적 법적 틀에 따라 과학기술・데이터에 대한 접근 및 접근의 질이라는 관점에서, 또  연결성・지식・숙련 및 영향받는 공동체(affected communities)의 유의미한 참여의 관점에서, 인공지능 시스템 수명 주기 전 영역에서 해결되어야 한다.

## Sustainability

## 지속가능성

31. The development of sustainable societies relies on the achievement of a complex set of objectives on a continuum of human, social, cultural, economic and environmental dimensions.

31. 지속가능한 사회의 발전은 인간 · 사회 · 문화 · 경제 · 환경적 차원의 지속이라는 복잡한 목표의 달성에 달려있다.

The advent of AI technologies can either benefit sustainability objectives or hinder their realization, depending on how they are applied across countries with varying levels of development.

인공지능 기술의 도래는 발전 수준이 다양한 국가들에 이것이 어떻게 적용이 되는지에 따라, 지속가능성 목표에 도움을 줄 수도, 이의 실현에 장애물이 될 수 있다.

The continuous assessment of the human, social, cultural, economic and environmental impact of AI technologies should therefore be carried out with full cognizance of the implications of AI technologies for sustainability as a set of constantly evolving goals across a range of dimensions, such as currently identified in the Sustainable Development Goals (SDGs) of the United Nations.

따라서 인공지능 기술의 인간 · 사회 · 문화 · 경제 · 환경적 영향력에 대한 지속적인 평가는, 지속가능성을 위한 인공지능 기술의 함의에 대한 완전한 이해를 바탕으로, 현재 유엔 지속가능발전목표(SDGs)에서 확인할 수 있는 것처럼, 다양한 차원에서 지속적으로 진전되는 목표로서 수행되어야 한다.

### Right to Privacy, and Data Protection
### 프라이버시 권리 및 정보 보호

32. Privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of AI systems.

32. 인간 존엄성, 인간 자율성, 인간 활동의 보호에 핵심적 권리인 프라이버시는 인공지능 시스템 수명 주기 전 영역에서 반드시 존중, 보호, 증진되어야 한다.

It is important that data for AI systems be collected, used, shared, archived and deleted in ways that are consistent with international law and in line with the values and principles set forth in this Recommendation, while respecting relevant national, regional and international legal frameworks.

인공지능 시스템을 위한 데이터가 국제법에 부합하고 본 권고에서 명시된 가치와 원칙에 맞게, 그리고 또한 유관 국가・지역・국제적 법률 틀을 존중하는 식으로 수집, 사용, 공유, 보관, 삭제되는 것은 중요하다.

33. Adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach at the national or international level, protected by judicial systems, and ensured throughout the life cycle of AI systems.

33. 적절한 데이터 보호 틀 및 거버넌스 메커니즘은 국가・국제적 차원의 다자적 접근법으로 확립되어야 하고, 사법 체계에 의해 보호받아야 하며, 인공지능 시스템 수명 주기 전 영역에서 보장되어야 한다.

Data protection frameworks and any related mechanisms should take reference from international data protection principles and standards concerning the collection, use and disclosure of personal data and exercise of their rights by data

subjects while ensuring a legitimate aim and a valid legal basis for the processing of personal data, including informed consent.

데이터 보호 틀 및 관련 메커니즘은 개인 정보의 수집・사용・공개 및 데이터 주체의 권리 행사와 관련된 국제 데이터 보호 원칙・기준을 참고해야 하며, 동시에 인지동의를 비롯하여 개인 데이터 처리를 위한 적법한 목적 및 타당한 법적 토대도 확고히 해야 한다.

34. Algorithmic systems require adequate privacy impact assessments, which also include societal and ethical considerations of their use and an innovative use of the privacy by design approach.

34. 알고리즘 시스템은 이의 개인정보 사용과 설계 단계에서의 개인정보의 혁신적 사용에 대한 사회・윤리적 고려를 포함하여 적절한 프라이버시 영향평가를 요구한다.

AI actors need to ensure that they are accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system.

인공지능 행위 주체는 가령 개인 정보가 인공지능 시스템의 전 영역에서 보호받음을 보장하는 식으로 그들이 인공지능 시스템의 설계 및 구현에 책임을 짐을 보장해야 한다.

### Human oversight and determination

### 인간의 감독 및 결정

35. Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems, as well as in cases of remedy related to AI systems, to physical persons or to existing legal entities.

35. 회원국은 인공지능 시스템 수명 주기의 어느 단계에서도, 또한 인공지능 시스템에 연관된 배상 문제에서, 개인 또는 법인에 윤리적・법적 책임을 묻는 것이 항상 가능하도록 해야 한다.

Human oversight refers thus not only to individual human oversight, but to inclusive public oversight, as appropriate.

따라서 인간의 감독이란 1인 감독뿐만 아니라, 적절한 경우 포괄적으로 대중의 감독까지도 지칭한다.

36. It may be the case that sometimes humans would choose to rely on AI systems for reasons of efficacy, but the decision to cede control in limited contexts remains that of humans, as humans can resort to AI systems in decision-making and acting, but an AI system can never replace ultimate human responsibility and accountability.

36. 인간이 의사결정 및 행동에 있어 인공지능 시스템에 의지할 수는 있지만, 인공지능 시스템이 인간의 절대적인 책임과 책무(accountability)를 대신할 수 없기 때문에, 때때로 인간이 효율성을 위해 인공지능 시스템에 의지하길 선택하는 경우가 있더라도, 제한적인 맥락에서 제어권을 양도할지 결정하는 것은 여전히 인간의 몫이다.

As a rule, life and death decisions should not be ceded to AI systems.

원칙적으로, 삶과 죽음에 대한 결정은 인공지능 시스템에게 양도되어서는 안 된다.

Transparency and explainability

투명성 및 설명가능성

37. The transparency and explainability of AI systems are often essential preconditions to ensure the respect, protection and promotion of human rights, fundamental freedoms and ethical principles.

37. 인공지능 시스템의 투명성 및 설명가능성은 인권 및 근본적 자유, 윤리적 원칙에 대한 존중・보호・증진을 확고히 함에 있어 필수 선결조건이다.

Transparency is necessary for relevant national and international liability regimes to work effectively.

투명성은 관련 국내·국제 법적 책임 체제가 효과적으로 작동하는 데에 필수적이다.

A lack of transparency could also undermine the possibility of effectively challenging decisions based on outcomes produced by AI systems and may thereby infringe the right to a fair trial and effective remedy, and limits the areas in which these systems can be legally used.

또한 투명성의 부족은 인공지능 시스템이 산출한 결과물에 기반한 의사결정에 대한 실질적 이의제기 가능성을 저해할 수 있고, 이에 따라 공정한 재판 및 배상을 받을 권리를 침해할 수 있으며, 이러한 시스템이 법적으로 사용될 수 있는 영역을 제한할 수 있다.

38. While efforts need to be made to increase transparency and explainability of AI systems, including those with extra-territorial impact, throughout their life cycle to support democratic governance, the level of transparency and explainability should always be appropriate to the context and impact, as there may be a need to balance between transparency and explainability and other principles such as privacy, safety and security.

38. 인공지능 시스템 수명 주기 전 영역에서 민주적 거버넌스를 지원하기 위해서는 (법역외 영향까지도 포함하여) 이의 투명성 및 설명가능성을 향상시키려는 모든 노력이 이루어져야 하는 반면, 투명성 및 설명가능성의 수준은 항상 해당 맥락과 영향 정도에 따라 적절한 수준이어야 하는데, 이는 투명성 및 설명가능성과 프라이버시, 안전 및 보안과 같은 다른 원칙 사이에서 균형을 이루어야 할 필요성이 있을 수 있기 때문이다.

People should be fully informed when a decision is informed by or is made on the basis of AI algorithms, including when it affects their safety or human rights, and in those circumstances should have the opportunity to request explanatory

information from the relevant AI actor or public sector institutions.

사람들은 의사결정이 인공지능 알고리즘으로부터 정보를 얻거나 이에 기반하여 내려지는 경우, 특히 이로써 그들의 안전 또는 인권에 영향이 가는 경우에 사전에 알 수 있어야 하고, 그러한 상황에서 관련 인공지능 행위 주체 또는 공공 기관에서 정보 설명을 요구할 수 있는 기회를 가지고 있어야 한다.

In addition, individuals should be able to access the reasons for a decision affecting their rights and freedoms, and have the option of making submissions to a designated staff member of the private sector company or public sector institution able to review and correct the decision.

더욱이, 개인은 자신의 권리 및 자유에 영향을 미치는 결정에 대한 근거에 접근할 수 있어야 하고, 이 결정을 검토・정정할 수 있는 민간 기업 또는 공공 기관의 담당 직원에게 의견을 개진할 수 있는 선택권을 가지고 있어야 한다.

AI actors should inform users when a product or service is provided directly or with the assistance of AI systems in a proper and timely manner.

인공지능 행위 주체는 제품이나 서비스가 직접적으로 또는 인공지능 시스템의 보조를 통해 제공될 때 사용자에게 이를 시기적절하게 사전 통보해야 한다.

39. From a socio-technical lens, greater transparency contributes to more peaceful, just, democratic and inclusive societies.

39. 사회・기술적 관점에서, 더 높은 투명성은 더 평화롭고 공정하고 민주적이며 포용적인 사회에 기여한다.

It allows for public scrutiny that can decrease corruption and discrimination, and can also help detect and prevent negative impacts on human rights.

이는 부패 및 차별을 감소시킬 수 있는 공개 조사를 가능하게 하고, 또한 인권에 대한 부정적 영향을 감지·예방하는 데에 도움이 될 수 있다.

Transparency aims at providing appropriate information to the respective addressees to enable their understanding and foster trust.

투명성은 정보를 받는 각 사람의 이해를 돕고 신뢰를 형성하기 위하여 알맞은 정보를 제공하는 것을 목적으로 한다.

Specific to the AI system, transparency can enable people to understand how each stage of an AI system is put in place, appropriate to the context and sensitivity of the AI system.

특히 인공지능 시스템에서, 투명성은 인공지능 시스템의 각 단계가 어떻게 상황과 인공지능 시스템의 맥락과 민감성에 적합하게 실행되는지 사람들이 이해할 수 있게 한다.

It may also include insight into factors that affect a specific prediction or decision, and whether or not appropriate assurances (such as safety or fairness measures) are in place.

투명성은 특정 예측 또는 결정에 영향을 미치는 요인들에 대한 통찰력과, 안전 또는 공정성 조치와 같은 타당한 보증이 준비가 되어있는지 여부 또한 포함할 수도 있다.

In cases of serious threats of adverse human rights impacts, transparency may also require the sharing of code or datasets.

심각한 반인권적 위협이 예상되는 경우, 투명성은 코드 또는 데이터셋의 공유를 요구할 수도 있다.

40. Explainability refers to making intelligible and providing insight into the outcome of AI systems.

40. 설명가능성은 인공지능 시스템의 결과가 이해될 수 있게 하고 이에 타당한 통찰력을 제공하는 것을 의미한다.

The explainability of AI systems also refers to the understandability of the input, output and the functioning of each algorithmic building block and how it contributes to the outcome of the systems.

인공지능 시스템의 설명가능성은 각 알고리즘 구성 요소의 입력・출력・기능 및 이것이 시스템의 결과물에 기여하는 방식에 대한 이해가능성을 의미한다.

Thus, explainability is closely related to transparency, as outcomes and sub-processes leading to outcomes should aim to be understandable and traceable, appropriate to the context.

따라서, 결과물 및 결과물로 이어지는 하위 과정은 이해 및 추적이 가능하고 사용 맥락에 적합해야 함을 목적으로 한다는 점에서, 설명가능성은 투명성과 매우 밀접하게 연관되어 있다.

AI actors should commit to ensuring that the algorithms developed are explainable.

인공지능 행위 주체는 개발된 알고리즘이 설명가능해야 한다는 점에 약속해야 한다.

In the case of AI applications that impact the end user in a way that is not temporary, easily reversible or otherwise low risk, it should be ensured that the meaningful explanation is provided with any decision that resulted in the action taken in order for the outcome to be considered transparent.

일반 사용자에 미치는 영향이 임시적이지 않거나 쉽게 되돌릴 수 있거나 아니면 위험도가

낮은 것이 아니라면, 인공지능의 응용에 있어서 어떤 행동을 야기한 결정에 대해 그 결과물이 투명하다고 여겨질 수 있도록 유의미한 설명의 제공이 보장되어야 한다.

41. Transparency and explainability relate closely to adequate responsibility and accountability measures, as well as to the trustworthiness of AI systems.

41. 투명성과 설명가능성은 적정한 책임・책무 조치뿐만 아니라 인공지능 시스템의 신뢰성과도 깊은 연관이 있다.

Responsibility and accountability

책임 및 책무

42. AI actors and Member States should respect, protect and promote human rights and fundamental freedoms, and should also promote the protection of the environment and ecosystems, assuming their respective ethical and legal responsibility, in accordance with national and international law, in particular Member States' human rights obligations, and ethical guidance throughout the life cycle of AI systems, including with respect to AI actors within their effective territory and control.

42. 인공지능 행위 주체 및 회원국은 인공지능 시스템 수명 주기 내내 국내・국제법, 특히 국제 회원국의 인권 준수 의무와 인공지능 시스템의 수명 주기 전 영역에서의 윤리 지침에 따라 인공지능 행위 주체를 효과적인 영역과 통제 하에 둠으로써, 각각의 윤리적・법적 책임를 지고 인권 및 근본적 자유를 존중・보호・증진하며 환경 및 생태계 보호를 장려해야 한다.

The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system.

일단 인공지능 시스템에 기반하여 내린 결정 및 행동에 대한 윤리적 책무와 법적 책임은 항상 궁극적으로 인공지능 시스템 수명 주기에서의 해당 역할의 인공지능 행위 주체에게 귀속된다.

43. Appropriate oversight, impact assessment, audit and due diligence mechanisms, including whistle-blowers' protection, should be developed to ensure accountability for AI systems and their impact throughout their life cycle.

43. 인공지능 시스템과 이의 수명 주기 전 영역에서의 영향력에 대한 책임을 보장하기 위해서는 제보자 보호를 비롯하여 적절한 감독, 영향 평가, 감사, 실사 메커니즘이 개발되어야 한다.

Both technical and institutional designs should ensure auditability and traceability of (the working of) AI systems in particular to address any conflicts with human rights norms and standards and threats to environmental and ecosystem well-being.

특히 인권 개념・기준과의 충돌 및 환경 및 생태계의 안녕에 대한 위협이 발생하는 경우, 기술적, 제도적 설계는 이를 해결할 수 있도록 인공지능 시스템(또는 이의 작동)에 대한 감사가능성 및 추적가능성을 보장해야 한다.

Awareness and literacy

인식 및 리터러시

44. Public awareness and understanding of AI technologies and the value of data should be promoted through open and accessible education, civic engagement, digital skills and AI ethics training, media and information literacy and training led jointly by governments, intergovernmental organizations, civil society, academia, the media, community leaders and the private sector, and considering the existing

linguistic, social and cultural diversity, to ensure effective public participation so that all members of society can take informed decisions about their use of AI systems and be protected from undue influence.

44. 인공지능 기술 및 데이터의 가치에 대한 대중의 인식과 이해는 개방적이고 접근이 편리한 교육, 시민 참여, 디지털 숙련 및 인공지능 윤리 교육, 그리고 정부, 정부간 국제기구, 시민사회, 학계, 미디어, 사회지도자, 민간 부문이 공동으로 추진하는 미디어·정보 리터러시 및 교육, 언어·사회·문화적 다양성의 고려를 통하여 제고되어야 하는데, 이로써 대중의 실제적 참여가 보장되어 사회 모든 구성원이 인공지능 시스템의 사용에 대해 올바른 결정을 내리고 과도한 영향으로부터 보호받을 수 있게 된다.

45. Learning about the impact of AI systems should include learning about, through and for human rights and fundamental freedoms, meaning that the approach and understanding of AI systems should be grounded by their impact on human rights and access to rights, as well as on the environment and ecosystems.

45. 인공지능 시스템의 영향에 대한 학습에는 인권 및 근본적 자유와 이에 대한, 이를 통한, 이를 위한 학습이 포함되어야 하는데, 이는 환경 및 생태계, 인권과 이에 대한 접근에 미치는 인공지능의 영향이 인공지능 시스템에 대한 접근방식과 이해의 기초가 되어야 함을 의미한다.

Multi-stakeholder and adaptive governance and collaboration

다자적 및 적응적 거버넌스 및 협업

46. International law and national sovereignty must be respected in the use of data.

46. 데이터의 사용에 있어서 국제법과 국가의 주권은 반드시 존중되어야 한다.

That means that States, complying with international law, can regulate the data generated within or passing through their territories, and take measures towards effective regulation of data, including data protection, based on respect for the right to privacy in accordance with international law and other human rights norms and standards.

이는 국가가 국제법을 준수하면서, 자국 내에서 생성・통용되는 데이터를 규제하고, 국제법 및 다른 인권 개념・기준에 따른 프라이버시권 및 인권에 대한 존중을 바탕으로 데이터 보호를 비롯한 효과적인 데이터 규제 조치를 취할 수 있음을 의미한다.

47. Participation of different stakeholders throughout the AI system life cycle is necessary for inclusive approaches to AI governance, enabling the benefits to be shared by all, and to contribute to sustainable development.

47. 인공지능 시스템 수명 주기 전 영역에서 다양한 이해관계자의 참여는 인공지능 거버넌스에 대한 포용적 접근방식을 위해 필수적인데, 이로써 혜택이 모두에게 공유될 수 있고 지속가능한 발전에 기여할 수 있게 된다.

Stakeholders include but are not limited to governments, intergovernmental organizations, the technical community, civil society, researchers and academia, media, education, policy-makers, private sector companies, human rights institutions and equality bodies, anti-discrimination monitoring bodies, and groups

for youth and children.

이해관계자는 정부, 정부간 국제기구, 과학기술 공동체, 시민사회, 연구자 및 학계, 미디어, 교육계, 정책결정자, 민간 기업, 인권단체 및 평등단체, 차별금지감시기관, 청소년 및 어린이보호단체 등을 포함한다.

The adoption of open standards and interoperability to facilitate collaboration should be in place.

협업을 촉진하기 위해서는 개방적 기준과 호환작업성(interoperability)이 도입되어야 한다.

Measures should be adopted to take into account shifts in technologies, the emergence of new groups of stakeholders, and to allow for meaningful participation by marginalized groups, communities and individuals and, where relevant, in the case of Indigenous Peoples, respect for the self-governance of their data.

과학기술의 변화, 새로운 이해관계자 집단의 출현을 고려하는, 그리고 소외 집단, 사회, 개인의 유의미한 참여를 가능하게 하고 (어떤 경우) 원주민의 자율적 데이터 거버넌스를 존중하는 조치가 채택되어야 한다.

## IV. AREAS OF POLICY ACTION

## IV. **정책 행동 영역**

48. The policy actions described in the following policy areas operationalize the values and principles set out in this Recommendation.

48. 이하의 정책 분야 내에 기술된 정책 행동은 본 권고에서 제시한 가치와 원칙을 활용한다.

The main action is for Member States to put in place effective measures, including, for example, policy frameworks or mechanisms, and to ensure that other stakeholders, such as private sector companies, academic and research institutions, and civil society adhere to them by, among other actions, encouraging all stakeholders to develop human rights, rule of law, democracy, and ethical impact assessment and due diligence tools in line with guidance including the United Nations Guiding Principles on Business and Human Rights.

모든 행동 중에서 가장 핵심은 회원국이 가령 정책 틀 또는 메커니즘과 같은 효과적인 조치를 마련하는 것이고, 또 모든 이해관계자로 하여금 인권, 법치주의, 민주주의, 그리고 '유엔 기업과 인권 이행원칙'을 비롯한 지침에 따르는 윤리영향평가 및 실사 도구를 개발하도록 장려함으로써 민간 기업, 학술 및 연구 기관, 시민사회와 같은 다른 이해관계자들이 이러한 노선을 견지할 수 있도록 하는 것이다.

The process for developing such policies or mechanisms should be inclusive of all stakeholders and should take into account the circumstances and priorities of each Member State.

그러한 정책 또는 메커니즘을 개발하는 절차는 모든 이해관계자들을 포함해야 하며, 각 회원국의 상황 및 우선순위를 고려해야 한다.

UNESCO can be a partner and support Member States in the development as well as monitoring and evaluation of policy mechanisms.

유네스코는 정책 메커니즘의 모니터링 및 심사뿐만 아니 개발에 있어서도 협력자로서 회원국을 지원할 수 있다.

49. UNESCO recognizes that Member States will be at different stages of readiness to implement this Recommendation, in terms of scientific, technological, economic, educational, legal, regulatory, infrastructural, societal, cultural and other

dimensions.

49. 유네스코는 회원국마다, 과학・기술・경제・교육・법・규제・인프라・사회・문화 및 기타 영역에서 본 권고를 이행할 수 있는 준비 정도가 각각 다름을 인정한다.

It is noted that “readiness” here is a dynamic status.

여기서 ‘준비 정도’는 변동될 수 있는 상태라는 점을 유의해야 한다.

In order to enable the effective implementation of this Recommendation, UNESCO will therefore: (1) develop a readiness assessment methodology to assist interested Member States in identifying their status at specific moments of their readiness trajectory along a continuum of dimensions; and (2) ensure support for interested Member States in terms of developing a UNESCO methodology for Ethical Impact Assessment (EIA) of AI technologies, sharing of best practices, assessment guidelines and other mechanisms and analytical work.

따라서 본 권고가 효과적으로 이행되도록 하기 위해서, 유네스코는 (1) 관심 있는 회원국이 일련의 영역에서 준비 궤도의 어느 특정 시점에 준비 상태를 판별할 수 있도록 돕는 준비 정도 평가방법론을 개발할 것이며, (2) 관심 있는 회원국에게는 인공지능 기술의 윤리영향평가(EIA), 모범 사례 공유, 평가 지침, 기타 메커니즘 및 분석작업에 대하여 유네스코 방법론의 개발 측면에서 지원을 보장할 것이다.

## POLICY AREA 1: ETHICAL IMPACT ASSESSMENT

## 정책 영역 1: 윤리영향평가

50. Member States should introduce frameworks for impact assessments, such as ethical impact assessment, to identify and assess benefits, concerns and risks of AI systems, as well as appropriate risk prevention, mitigation and monitoring measures, among other assurance mechanisms.

50. 회원국은 위험 예방・완화・모니터링 조치에 더하여 인공지능 시스템의 이익, 우려, 위험성을 파악 및 평가하기 위해, 다른 보증 메커니즘 보다도 윤리영향평가와 같은 영향평가 틀을 도입해야 한다.

Such impact assessments should identify impacts on human rights and fundamental freedoms, in particular but not limited to the rights of marginalized and vulnerable people or people in vulnerable situations, labour rights, the environment and ecosystems and ethical and social implications, and facilitate citizen participation in line with the values and principles set forth in this Recommendation.

윤리영향평가는 인권 및 근본적 자유, 특히 소외・취약계층 및 취약한 상황에 처한 사람의 권리, 노동자 인권, 환경 및 생태계, 윤리・사회적 함의 등에 대한 영향을 평가해야 하며, 본 권고에서 제시하는 가치 및 원칙에 따라 시민 참여를 활성화해야 한다.

51. Member States and private sector companies should develop due diligence and oversight mechanisms to identify, prevent, mitigate and account for how they address the impact of AI systems on the respect for human rights, rule of law and inclusive societies.

51. 회원국과 민간 기업은 인공지능 시스템이 인권, 법치주의, 포용 사회에 주는 영향을 어떻게 다루고 있는지 판별・예방・완화・설명할 수 있는 실사 및 감독 메커니즘을 개발해야 한다.

Member States should also be able to assess the socio-economic impact of AI systems on poverty and ensure that the gap between people living in wealth and poverty, as well as the digital divide among and within countries, are not increased with the massive adoption of AI technologies at present and in the future.

또한, 회원국은 인공지능 시스템이 빈곤에 미치는 사회・경제적 영향을 평가할 수 있어야

하며, 현재 및 미래에 인공지능 기술의 대규모 도입으로 인하여 빈부격차 및 국내·국가간 디지털 격차가 더 커지지 않도록 보장해야 한다.

In order to do this, in particular, enforceable transparency protocols should be implemented, corresponding to the access to information, including information of public interest held by private entities.

이를 위하여 특별히 강제성을 띤 투명성 규약이 마련되어야 하는데, 이는 정보, 심지어 민간 기관이 보유하였어도 공익을 위한 정보라면 접근할 수 있는 권한을 의미하는 것이다.

Member States, private sector companies and civil society should investigate the sociological and psychological effects of AI-based recommendations on humans in their decision-making autonomy.

회원국, 민간 기업, 시민사회는 인공지능 기반 추천이 인간의 의사 결정에 관한 자율성에 미치는 사회학적, 심리학적 효과를 조사해야 한다.

AI systems identified as potential risks to human rights should be broadly tested by AI actors, including in real-world conditions if needed, as part of the Ethical Impact Assessment, before releasing them in the market.

인권에 잠재적인 위험일 것으로 파악되는 인공지능 시스템은 시장에 출시되기 전에 윤리 영향평가의 일환으로 인공지능 행위 주체에게 광범위한 점검을 받아야 하며, 필요하다면 실제 조건 하에서 그러해야 한다.

52. Member States and business enterprises should implement appropriate measures to monitor all phases of an AI system life cycle, including the functioning of algorithms used for decision-making, the data, as well as AI actors involved in the process, especially in public services and where direct end-user interaction is needed, as part of ethical impact assessment.

52. 회원국 및 기업은 윤리영향평가의 일환으로 인공지능 시스템 수명 주기의 전 단계를 모니터링할 수 있는 적절한 조치를 마련해야 하며, 이런 단계들은 특히 공공 서비스에서와 최종 사용자와의 직접적인 상호작용이 필요한 곳의 프로세스와 관여되어 있는 인공지능 행위 주체와 더불어 의사결정에 사용되는 알고리즘의 작용, 데이터를 포함한다.

Member States' human rights law obligations should form part of the ethical aspects of AI system assessments.

회원국의 인권법 준수 의무는 인공지능 시스템 평가의 윤리적 측면의 일부로 자리해야 한다.

53. Governments should adopt a regulatory framework that sets out a procedure, particularly for public authorities, to carry out ethical impact assessments on AI systems to predict consequences, mitigate risks, avoid harmful consequences, facilitate citizen participation and address societal challenges.

53. 정부는 특히 공공기관이 결과 예측, 위험 완화, 피해 예방, 시민 참여 확대, 사회 문제 해결을 위하여 만든 인공지능 시스템에 대하여 윤리영향평가를 수행하는 절차를 제시하는 규제적 틀을 채택해야 한다.

The assessment should also establish appropriate oversight mechanisms, including auditability, traceability and explainability, which enable the assessment of algorithms, data and design processes, as well as include external review of AI systems.

또한 이 평가는 인공지능 시스템에 대한 외부 검토를 포함함과 더불어, 알고리즘・데이터・설계과정에 대한 평가를 가능하게 하는 감사가능성, 추적가능성, 설명가능성을 비롯한 적절한 감독 메커니즘을 확립해야 한다.

Ethical impact assessments should be transparent and open to the public, where appropriate.

윤리영향평가는 투명해야 하며 적절한 경우 대중에게 공개되어야 한다.

Such assessments should also be multidisciplinary, multi-stakeholder, multicultural, pluralistic and inclusive.

이러한 평가는 다학문적・다자적・다문화적・다원주의적・포용적이어야 한다.

The public authorities should be required to monitor the AI systems implemented and/or deployed by those authorities by introducing appropriate mechanisms and tools.

공공기관은 적절한 메커니즘과 도구를 도입함으로써 자체적으로 구현 및/또는 배치한 인공지능 시스템을 모니터링해야 한다.

## POLICY AREA 2: ETHICAL GOVERNANCE AND STEWARDSHIP

## 정책 영역 2: 윤리적 거버넌스 및 감독의무(stewardship)

54. Member States should ensure that AI governance mechanisms are inclusive, transparent, multidisciplinary, multilateral (this includes the possibility of mitigation and redress of harm across borders) and multi-stakeholder.

54. 회원국은 인공지능 거너번스 메커니즘이 포용적이고 투명하며 다학문적이고 다국적(국경을 초월하는 위험 완화 및 해결 가능성까지 의미함)이고 다자적이도록 보장해야 한다.

In particular, governance should include aspects of anticipation, and effective protection, monitoring of impact, enforcement and redress.

특히, 거버넌스는 예상, 효과적인 보호, 영향 모니터링, 강제이행, 시정 같은 측면을 포함해야 한다.

55. Member States should ensure that harms caused through AI systems are investigated and redressed, by enacting strong enforcement mechanisms and remedial actions, to make certain that human rights and fundamental freedoms and the rule of law are respected in the digital world and in the physical world.

55. 회원국은 인권 및 법치가 현실 및 디지털 세계에서 존중될 수 있도록 강력한 강제이행 메커니즘 및 시정 조치를 제정함으로써, 인공지능 시스템으로 인해 발생하는 피해가 조사 및 해결되도록 보장해야 한다.

Such mechanisms and actions should include remediation mechanisms provided by private and public sector companies.

이러한 메커니즘과 조치에는 민간・공공 기업이 제공하는 시정 메커니즘이 포함되어야 한다.

The auditability and traceability of AI systems should be promoted to this end.

이를 위해서는 감사가능성과 추적가능성이 증진되어야 한다.

In addition, Member States should strengthen their institutional capacities to deliver on this commitment and should collaborate with researchers and other stakeholders to investigate, prevent and mitigate any potentially malicious uses of AI systems.

더욱이, 회원국은 이러한 약속을 실천할 제도적 역량을 강화해야 하며, 인공지능 시스템의 잠재적 악용을 조사・예방・완화하기 위하여 연구자 및 기타 이해관계자들과 협업해야 한다.

56. Member States are encouraged to develop national and regional AI strategies and to consider forms of soft governance such as a certification mechanism for AI systems and the mutual recognition of their certification, according to the sensitivity of the application domain and expected impact on human rights, the

.environment and ecosystems, and other ethical considerations set forth in this Recommendation.

56. 응용 분야의 민감성, 그리고 인권, 환경 및 생태계, 본 권고에서 제시하는 기타 윤리적 고려사항에 미칠 것으로 예상되는 영향에 따라, 회원국은 국가 · 지역적 인공지능 전략을 개발하고 인공지능 시스템의 인증 메커니즘 및 그러한 인증의 상호적 인지와 같은 유연한 거버넌스의 유형을 고려하도록 권장된다.

Such a mechanism might include different levels of audit of systems, data, and adherence to ethical guidelines and to procedural requirements in view of ethical aspects.

윤리적 측면을 고려하여 이러한 메커니즘에는 시스템, 데이터, 윤리 지침 준수 및 절차적 필수조건 준수에 대한 다양한 수준의 감사가 포함될 수 있다.

At the same time, such a mechanism should not hinder innovation or disadvantage small and medium enterprises or start-ups, civil society as well as research and science organizations, as a result of an excessive administrative burden.

이와 동시에, 이러한 메커니즘은 과도한 행정 부담으로 혁신을 가로막거나, 중소기업 또는 스타트업, 시민사회, 연구 · 과학 기관에 불이익을 주어서는 안 된다.

These mechanisms should also include a regular monitoring component to ensure system robustness and continued integrity and adherence to ethical guidelines over the entire life cycle of the AI system, requiring re-certification if necessary.

이 메커니즘은 필요하다면 재인증도 요구함으로써, 인공지능 시스템의 전체 수명 주기 동안 시스템 견고성, 지속적인 무결성 및 윤리 지침 준수를 보장하기 위한 정기적인 모니터링 요소도 포함해야 한다.

57. Member States and public authorities should carry out transparent self-ssessment of existing and proposed AI systems, which, in particular, should include the assessment of whether the adoption of AI is appropriate and, if so, should include further assessment to determine what the appropriate method is, as well as assessment as to whether such adoption would result in violations or abuses of Member States' human rights law obligations, and if that is the case, prohibit its use.

57. 공공기관은 현재 및 새로운 인공지능 시스템에 대해 투명하게 자체 평가를 수행해야 하며, 이는 특히 인공지능의 도입이 적절한지, 그렇다면 적절한 방식이 무엇이고 그 도입이 회원국의 인권법 준수 의무 위반 또는 남용을 초래할 것인지에 대한 평가를 포함해야 하며, 만약 그럴 경우 인공지능의 사용을 금지해야 한다.

58. Member States should encourage public entities, private sector companies and civil society organizations to involve different stakeholders in their AI governance and to consider adding the role of an independent AI Ethics Officer or some other mechanism to oversee ethical impact assessment, auditing and continuous monitoring efforts and ensure ethical guidance of AI systems.

58. 회원국은 공공 단체, 민간 기업, 시민사회로 하여금 인공지능 거버넌스에 다른 이해관계자를 참여시키도록, 그리고 윤리영향평가, 감사, 지속적인 모니터링을 감독하고 인공지능 시스템에 대한 윤리적 인도를 확고히 하기 위해 '인공지능윤리책임자'와 같은 역할 또는 기타 메커니즘의 추가를 고려하도록 장려해야 한다.

Member States, private sector companies and civil society organizations, with the support of UNESCO, are encouraged to create a network of independent AI Ethics Officers to give support to this process at national, regional and international levels.

유네스코의 지지 하에, 회원국, 민간 기업, 시민사회 조직은 이러한 프로세스를 국가・지역・국제적 수준에서 지원할 수 있는 별도의 '인공지능윤리책임자' 네트워크의 생성이 권장된다.

59. Member States should foster the development of, and access to, a digital ecosystem for ethical and inclusive development of AI systems at the national level, including to address gaps in access to the AI system life cycle, while contributing to international collaboration.

59. 회원국은 인공지능 시스템 수명 주기에 대한 접근의 격차를 해결함을 비롯하여 국가적 차원에서 인공지능의 윤리적・포용적 발전을 위한 디지털 생태계의 개발 및 이에 대한 접근을 촉진해야 하며, 동시에 국제적 협업에 기여해야 한다.

Such an ecosystem includes, in particular, digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate.

이러한 디지털 생태계에는 특히 디지털 기술・인프라와 적절한 경우 인공지능 지식 공유 메커니즘까지 포함된다.

60. Member States should establish mechanisms, in collaboration with international organizations, transnational corporations, academic institutions and civil society, to ensure the active participation of all Member States, especially LMICs, in particular LDCs, LLDCs and SIDS, in international discussions concerning AI governance.

60. 회원국은 국제기구, 다국적 기업, 학술 단체, 시민사회와 협업하여 모든 회원국, 특히 중・저소득국가, 그 중에서도 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국이 인공지능 거버넌스에 관한 국제적 논의에 적극적으로 참여하도록 보장하는 메커니즘을 확립해야 한다.

This can be through the provision of funds, ensuring equal regional participation, or any other mechanisms.

이는 모든 지역의 평등한 참여를 보장할 수 있는 자금 지원을 통해, 또는 기타 메커니즘을 통해 가능하다.

Furthermore, in order to ensure the inclusiveness of AI fora, Member States should facilitate the travel of AI actors in and out of their territory, especially from LMICs, in particular LDCs, LLDCs and SIDS, for the purpose of participating in these for a.

이에 더해, 인공지능 포럼의 포용성을 보장하기 위해서 회원국은 인공지능 행위 주체, 특히 중·저소득국가, 그 중에서도 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국 출신이 포럼 참여의 목적으로 자국영토 안팎에서 용이하게 통행할 수 있도록 해야 한다.

61. Amendments to the existing or elaboration of new national legislation addressing AI systems must comply with Member States' human rights law obligations and promote human rights and fundamental freedoms throughout the AI system life cycle.

61. 인공지능 시스템을 다루는 기존 국내법의 수정 또는 새로운 국내법 정교화는 회원국의 인권법 준수 의무를 반드시 따라야 하며 인공지능 시스템 수명 주기 내내 인권과 기본적 자유를 고취시켜야 한다.

Promotion thereof should also take the form of governance initiatives, good exemplars of collaborative practices regarding AI systems, and national and international technical and methodological guidelines as AI technologies advance.

이를 증진하는 것은 거버넌스 계획, 인공지능 시스템에 관한 협업 관행의 모범 사례, 인공지능 기술의 발전에 따른 국내·국제적 기술·방법론 지침의 형태를 띄어야 한다.

Diverse sectors, including the private sector, in their practices regarding AI systems must respect, protect and promote human rights and fundamental freedoms using existing and new instruments in combination with this Recommendation.

인공지능 시스템에 관한 사용 관행에서, 민간 부문을 비롯한 다양한 부문은 반드시 본 권

고와 함께 기존 및 새로운 도구를 병용함으로써 인권 및 근본적 자유를 존중・보호・증진해야 한다.

62. Member States that acquire AI systems for human rights-sensitive use cases, such as law enforcement, welfare, employment, media and information providers, health care and the independent judiciary system should provide mechanisms to monitor the social and economic impact of such systems by appropriate oversight authorities, including independent data protection authorities, sectoral oversight and public bodies responsible for oversight.

62. 법 집행, 복지, 고용, 미디어 및 정보 제공자, 건강 관리, 독립 사법 체제와 같은 인권 민감 사안에 대해 인공지능 시스템을 사용하는 회원국은 독립된 데이터 보호기관, 부문별 감시, 감독 책임이 있는 공공 단체를 비롯한 적절한 감독기관을 통해 이의 사회・경제적 영향을 모니터링하도록 메커니즘을 제공해야 한다.

63. Member States should enhance the capacity of the judiciary to make decisions related to AI systems as per the rule of law and in line with international law and standards, including in the use of AI systems in their deliberations, while ensuring that the principle of human oversight is upheld.

63. 회원국은 사법부가 심의에 인공지능 시스템을 사용하는 것을 비롯하여 사법부가 법치주의 및 국제법・기준에 따라 인공지능 시스템에 대한 결정을 내릴 수 있는 역량을 강화해야 하며, 동시에 인간 감독의 원칙이 유지되도록 보장해야 한다.

In case AI systems are used by the judiciary, sufficient safeguards are needed to guarantee inter alia the protection of fundamental human rights, the rule of law, judicial independence as well as the principle of human oversight, and to ensure a trustworthy, public interest-oriented and human-centric development and use of AI systems in the judiciary.

인공지능 시스템이 사법제도에서 사용되는 경우, 인간 감독과 더불어 무엇보다도 근본적 인권, 법치주의, 독립된 사법부의 보호를 보장하기 위하여, 그리고 사법제도에서의 인공지능의 개발과 사용이 신뢰받을 수 있고 공익을 위하며 인간 중심적이도록 하기 위하여 충분한 안전장치가 필요하다.

64. Member States should ensure that governments and multilateral organizations play a leading role in ensuring the safety and security of AI systems, with multi-takeholder participation.

64. 회원국은 정부와 다국적 조직이 다자적 참여를 통해 인공지능 시스템의 안전과 보안을 확고히 하는 데에 주도적 역할을 하도록 보장해야 한다.

Specifically, Member States, international organizations and other relevant bodies should develop international standards that describe measurable, testable levels of safety and transparency, so that systems can be objectively assessed and levels of compliance determined.

특히, 회원국, 국제기구, 기타 유관 단체는 시스템을 객관적으로 평가하고 규정 준수 수준을 결정할 수 있도록, 측정 가능하고 검 가능한 수준의 안전 및 투명성을 명시하는 국제 기준을 개발해야 한다.

Furthermore, Member States and business enterprises should continuously support strategic research on potential safety and security risks of AI technologies and should encourage research into transparency and explainability, inclusion and literacy by putting additional funding into those areas for different domains and at different levels, such as technical and natural language.

이에 더해, 회원국은 인공지능 기술의 잠재적인 안전・보안 위험에 대한 전략적 연구를 지속적으로 지원하여야 하며, 과학기술 및 자연어 등과 같이 다양한 분야 및 수준에서 투명성・설명가능성 및 포용성・리터러시에 추가 자금을 투입함으로써 이러한 영역에 대한 연구를 장려해야 한다.

65. Member States should implement policies to ensure that the actions of AI actors are consistent with international human rights law, standards and principles throughout the life cycle of AI systems, while taking into full consideration the current cultural and social diversities, including local customs and religious traditions, with due regard to the precedence and universality of human rights.

65. 회원국은 인공지능 행위 주체의 행동이 인공지능 시스템 수명 주기 전 영역에서 국제 인권법・기준・원칙을 따르면서도, 인권의 우선성・보편성에 따라 현지 관습 및 종교적 전통을 비롯한 오늘날의 문화・사회적 다양성에 대해 충분한 고려를 하도록 보장해야 한다.

66. Member States should put in place mechanisms to require AI actors to disclose and combat any kind of stereotyping in the outcomes of AI systems and data, whether by design or by negligence, and to ensure that training data sets for AI systems do not foster cultural, economic or social inequalities, prejudice, the spreading of disinformation and misinformation, and disruption of freedom of expression and access to information.

66. 회원국은 고의 또는 부주의로 인해 인공지능 시스템 및 데이터의 결과물에 존재하는 모든 종류의 고정관념을 인공지능 행위 주체가 공개하고 대응하도록 요구하기 위한, 그리고 인공지능 시스템을 위한 훈련데이터 집합이 문화・경제・사회적 불평등, 편견, 허위정보・오보의 확산, 표현의 자유 및 정보 접근에의 지장을 조장하지 않기 보장하기 위한 메커니즘을 마련해야 한다.

Particular attention should be given to regions where the data are scarce.

데이터가 부족한 지역에는 더 각별한 주의가 필요하다.

67. Member States should implement policies to promote and increase diversity and inclusiveness that reflect their populations in AI development teams and training datasets, and to ensure equal access to AI technologies and their benefits,

particularly for marginalized groups, both from rural and urban zones.

67. 회원국은 인공지능 개발팀 인원 및 훈련데이터 집합을 나타내는 다양성 및 포용성을 증진 및 증가시키고 농촌・도시 지역에서, 특히 소외집단이 인공지능 기술 및 이의 혜택에 평등하게 접근할 수 있도록 보장하는 정책을 구현해야 한다.

68. Member States should develop, review and adapt, as appropriate, regulatory frameworks to achieve accountability and responsibility for the content and outcomes of AI systems at the different phases of their life cycle.

68. 회원국은 인공지능 시스템 수명 주기 각 단계에서 이의 내용물과 결과물에 대한 책무 및 책임을 확보하기 위해 규제적 틀을  적절하게 개발・검토・조정해야 한다.

Member States should, where necessary, introduce liability frameworks or clarify the interpretation of existing frameworks to ensure the attribution of accountability for the outcomes and the functioning of AI systems.

회원국은 인공지능 시스템의 결과물 및 작용에 대한 책임소재를 명확히 할 수 있도록, 필요하다면 법적 책임 틀을 도입하거나 기존 틀에 대한 해석을 명료하게 해야 한다.

Furthermore, when developing regulatory frameworks, Member States should, in particular, take into account that ultimate responsibility and accountability must always lie with natural or legal persons and that AI systems should not be given legal personality themselves.

이에 더해, 규제적 틀을 개발할 때, 회원국은 언제나 궁극적인 책임 및 책무가 반드시 개인 또는 법인에게 있어야 하며 인공지능 시스템에 법인격이 부여되지 않아야 한다는 점을 고려해야 한다.

To ensure this, such regulatory frameworks should be consistent with the principle of human oversight and establish a comprehensive approach focused on AI actors and the technological processes involved across the different stages of the AI system life cycle.

이를 보장하기 위하여, 이러한 규제 프레임워크는 인간 감독의 원칙과 일치되어야 하며, 인공지능 시스템 수명 주기의 각 단계에 이에 관여되어 있는 행위 주체 및 기술적 프로세스에 초점을 맞춘 통합적 접근 방식을 수립해야 한다.

69. In order to establish norms where these do not exist, or to adapt the existing legal frameworks, Member States should involve all AI actors (including, but not limited to, researchers, representatives of civil society and law enforcement, insurers, investors, manufacturers, engineers, lawyers and users).

69. 규범이 존재하지 않은 영역에서 이를 확립하거나 기존의 법적 틀을 조정할 때, 회원국은 (연구자, 시민사회 및 법 집행기관의 대표자, 보험사, 투자자, 제조업자, 공학자, 변호사, 사용자를 포함하되 이에 국한되지는 않는) 모든 인공지능 행위 주체가 참여하도록 해야 한다.

The norms can mature into best practices, laws and regulations.

이 규범은 모범 관행, 법률, 규정으로 발전할 수 있다.

Member States are further encouraged to use mechanisms such as policy prototypes and regulatory sandboxes to accelerate the development of laws, regulations and policies, including regular reviews thereof, in line with the rapid development of new technologies and ensure that laws and regulations can be tested in a safe environment before being officially adopted.

이에 더해, 신기술의 급속한 발전에 따라, 회원국은 정기적 검토가 포함된 법률・규정・정책의 개발을 가속화하고 법률 및 규정을 공식적으로 도입하기 전 안전한 환경에서 점검될

수 있도록 정책 프로토타입 및 규제 실험장 같은 메커니즘을 사용하도록 더욱 권장된다.

Member States should support local governments in the development of local policies, regulations and laws in line with national and international legal frameworks.

회원국은 국내・국제적 법적 틀에 따른 지역 정책・규정・법률의 개발에 있어서 지방행정기관을 지원해야 한다.

70. Member States should set clear requirements for AI system transparency and explainability so as to help ensure the trustworthiness of the full AI system life cycle.

70. 회원국은 인공지능 시스템 수명 주기 전 영역에서 신뢰성을 보장하기 위해 인공지능 시스템의 투명성 및 설명가능성에 대한 명확한 요구사항을 설정해야 한다.

Such requirements should involve the design and implementation of impact mechanisms that take into consideration the nature of application domain, intended use, target audience and feasibility of each particular AI system.

이러한 요구 조건에는 각 인공지능 시스템의 응용 영역의 성격, 사용 의도, 사용대상자, 타당성에 대해 고려한 영향 메커니즘의 설계 및 구현이 수반된다.

## POLICY AREA 3: DATA POLICY

## 정책 영역 3: 데이터 정책

71. Member States should work to develop data governance strategies that ensure the continual evaluation of the quality of training data for AI systems including the adequacy of the data collection and selection processes, proper data security and protection measures, as well as feedback mechanisms to learn from mistakes and share best practices among all AI actors.

71. 회원국은 데이터 수집 및 선택 프로세스의 타당성, 적절한 데이터 보안・보호 조치를 비롯하여 인공지능 시스템의 학습 데이터 품질에 대한 지속적인 심사를 보장하는 데이터 거버넌스 전략을 개발하며, 실수로부터 교훈을 얻고 인공지능 행위 주체들 사이에서 모범 사례를 공유하기 위한 피드백 메커니즘을 개발하기 위해 노력해야 한다.

72. Member States should put in place appropriate safeguards to protect the right to privacy in accordance with international law, including addressing concerns such as surveillance.

72. 국제법에 따라, 회원국은 감시에 대한 우려를 해소하는 것을 비롯하여 프라이버시권을 보호하기 위해 적절한 안전장치를 마련해야 한다.

Member States should, among others, adopt or enforce legislative frameworks that provide appropriate protection, compliant with international law.

무엇보다도, 회원국은 국제법에 따라 적절한 보호를 제공하는 입법 입법 틀을 도입하거나 집행해야 한다.

Member States should strongly encourage all AI actors, including business enterprises, to follow existing international standards and, in particular, to carry out adequate privacy impact assessments, as part of ethical impact assessments, which take into account the wider socio-economic impact of the intended data processing, and to apply privacy by design in their systems.

회원국은 기업을 비롯한 모든 인공지능 행위 주체가 기존의 국제적 기준을 따르도록, 특히 윤리영향평가의 일환으로서 의도된 데이터 프로세스의 광범위한 사회・경제적 영향을 고려하는 프라이버시 영향평가를 수행하도록, 그리고 인공지능 시스템의 설계에 프라이버시를 적용하도록 강력히 권장해야 한다.

Privacy should be respected, protected and promoted throughout the life cycle of AI systems.

프라이버시는 인공지능 시스템 수명 주기 전 영역에서 존중, 보호, 증진되어야 한다.

73. Member States should ensure that individuals retain rights over their personal data and are protected by a framework, which notably foresees: transparency; appropriate safeguards for the processing of sensitive data; an appropriate level of data protection; effective and meaningful accountability schemes and mechanisms; the full enjoyment of the data subjects' rights and the ability to access and erase their personal data in AI systems, except for certain circumstances in compliance with international law; an appropriate level of protection in full compliance with data protection legislation where data are being used for commercial purposes such as enabling micro-targeted advertising, transferred cross-border; and an effective independent oversight as part of a data governance mechanism which keeps individuals in control of their personal data and fosters the benefits of a free flow of information internationally, including access to data.

73. 회원국은 개개인이 자신의 개인 데이터에 대한 권리를 보유하고 보호받도록 보장해야 하는데, 이를 보호하는 틀은 투명성, 민감한 데이터 처리에 대한 적절한 안전장치, 적정 수준의 데이터 보호, 효과적이고 유의미한 책임 제도 및 메커니즘, 국제법에 부합하는 특정 상황이 아닌 한에서 인공지능 시스템 내 개인 데이터 접근・삭제에 대한 데이터 주체의 권리・능력 향유, 데이터가 국경을 넘어 개인맞춤형 광고 같은 상업적 목적으로 쓰이는 경우 데이터 보호법에 철저히 따른 적정한 수준의 보호, 그리고 개인의 개인 데이터 통제권을 지켜주고 데이터 접근을 비롯한 정보의 자유로운 국제적 이동의 이익을 증진하는 데이터 거버넌스 메커니즘의 일환으로서 효과적이고 독립적인 감독을 주목하여 예상한다.

74. Member States should establish their data policies or equivalent frameworks, or reinforce existing ones, to ensure full security for personal data and sensitive data, which, if disclosed, may cause exceptional damage, injury or hardship to individuals.

74. 회원국은 유출될 경우 개인에게 상당한 피해・손상・곤란을 초래할 수도 있는 개인 데이터 및 민감한 데이터의 철저한 보안을 확고히 하기 위하여 데이터 정책 또는 이에 상응하는 틀을 수립하거나 기존의 것을 강화해야 한다.

Examples include data relating to offences, criminal proceedings and convictions, and related security measures; biometric, genetic and health data; and -personal data such as that relating to race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other characteristics.

이러한 데이터의 예로는 범죄・형사소송・전과 및 이와 관련된 보안조치 데이터, 생체・유전・건강 데이터, 그리고 인종, 피부색, 혈통, 성, 나이, 언어, 종교, 정치적 견해, 국가・민족・사회적 출신, 출생 시 경제・사회적 조건, 장애, 기타 다른 특징과 관련된 개인 데이터 등이 있다.

75. Member States should promote open data.

75. 또한 회원국은 열린 데이터를 장려해야 한다.

In this regard, Member States should consider reviewing their policies and regulatory frameworks, including on access to information and open government to reflect AI-specific requirements and promoting mechanisms, such as open repositories for publicly funded or publicly held data and source code and data trusts, to support the safe, fair, legal and ethical sharing of data, among others.

이와 관련하여, 회원국은 정부 정책 및 규제적 틀의 검토를 고려해야 하는데, 여기에는

정보 접근에 관한 것, 특정 인공지능 요구 조건을 반영하기 위한 열린 정부, 무엇보다도 안전하고 공정하며 합법적이고 윤리적인 데이터 공유를 지원하기 위하여 공공 지원・보유 데이터, 소스 코드, 데이터 신뢰를 위한 열린 저장소 같은 메커니즘을 장려하는 것이 포함된다.

76. Member States should promote and facilitate the use of quality and robust datasets for training, development and use of AI systems, and exercise vigilance in overseeing their collection and use.

76. 회원국은 양질의 견고성이 뛰어난 학습데이터 집합의 사용 및 이러한 인공지능 시스템의 개발 및 사용을 장려・촉진해야 하며, 이의 수집・활용을 감독함에 있어 주의해야 한다.

This could, if possible and feasible, include investing in the creation of gold standard datasets, including open and trustworthy datasets, which are diverse, constructed on a valid legal basis, including consent of data subjects, when required by law.

현실적으로 실행 가능하다면, 이는 신뢰할 수 있는 공개된 데이터 집합을 비롯한 '바람직한 표준' 데이터 집합 생성에 투자하는 것을 포함할 수 있는데, 이 데이터 집합은 다양하며 법이 요구하는 경우 데이터 주체의 동의 하에 타당한 법적 근거 위에 구성된 것을 말한다.

Standards for annotating datasets should be encouraged, including disaggregating data on gender and other bases, so it can easily be determined how a dataset is gathered and what properties it has.

데이터를 성 및 다른 요소 별로 해체하는 것을 비롯하여 데이터 집합 주석 표준을 권장함으로써, 데이터 집합의 수집 방식과 특성을 쉽게 확인할 수 있어야 한다.

77. Member States, as also suggested in the report of the United Nations Secretary-General's High-level Panel on Digital Cooperation, with the support of the United Nations and UNESCO, should adopt a digital commons approach to data where appropriate, increase interoperability of tools and datasets and interfaces of systems hosting data, and encourage private sector companies to share the data they collect with all stakeholders, as appropriate, for research, innovation or public benefits.

77. 또한, '유엔 사무총장의 디지털 협력에 관한 고위급 패널 보고서'에서 제안되었듯, 회원국은 유엔과 유네스코의 지원 하에 적합한 곳에 디지털 커먼즈(digital commons) 접근 방식을 채택해야 하며, 도구 및 데이터 집합, 데이터를 주관하는 시스템의 인터페이스의 호환성을 향상시켜야 하고, 민간 기업이 수집한 데이터를 연구, 혁신, 공익을 위해 모든 이해관계자와 적절히 공유하도록 권장해야 한다.

They should also promote public and private efforts to create collaborative platforms to share quality data in trusted and secured data spaces.

또한, 회원국은 믿음직하고 안전한 데이터 공간에서 양질의 데이터를 공유할 수 있는 협업 플랫폼을 만들기 위해 공공・민간 노력을 장려해야 한다.

## POLICY AREA 4: DEVELOPMENT AND INTERNATIONAL COOPERATION

## 정책 영역 4: 개발 및 국제 협력

78. Member States and transnational corporations should prioritize AI ethics by including discussions of AI-related ethical issues into relevant international, intergovernmental and multi-stakeholder for a.

78. 회원국과 다국적 기업은 인공지능과 관련된 윤리적 사안에 관한 논의를 관련 국제・국가간・다자간 포럼에 포함시킴으로써 인공지능 윤리를 주류화해야 한다.

79. Member States should ensure that the use of AI in areas of development such as education, science, culture, communication and information, health care, agriculture and food supply, environment, natural resource and infrastructure management, economic planning and growth, among others, adheres to the values and principles set forth in this Recommendation.

79. 회원국은 무엇보다 교육, 과학, 문화, 정보통신, 건강관리, 농・식품 공급, 환경, 천연자원・인프라 관리, 경제 계획 및 성장 등의 개발 분야에 있어서 인공지능의 사용이 본 권고에 제시된 가치 및 원칙을 따르도록 보장해야 한다.

80. Member States should work through international organizations to provide platforms for international cooperation on AI for development, including by contributing expertise, funding, data, domain knowledge, infrastructure, and facilitating multi-takeholder collaboration to tackle challenging development problems, especially for LMICs, in particular LDCs, LLDCs and SIDS.

80. 회원국은 국제기구를 통하여 전문지식, 자금 지원, 데이터, 각 분야 지식, 인프라를 기부하고 특히 중ㅇ저소득국가, 그 중에서도 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국의 힘겨운 개발 문제를 해결할 수 있도록 다자간 협업을 원활하게 함으로써, 개발에 사용되는 인공지능에 대한 국제적 협력을 위한 플랫폼을 제공하기 위해 힘써야 한다.

81. Member States should work to promote international collaboration on AI research and innovation, including research and innovation centres and networks that promote greater participation and leadership of researchers from LMICs and other countries, including LDCs, LLDCs and SIDS.

81. 회원국은 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국를 비롯한 중・저소득국가 및 기타 국가 연구자의 더 많은 참여와 주도를 장려할 수 있는 연구 ・혁신 센터 및 네트워크를 비롯한 인공지능 연구 및 혁신에 있어서 국제적 협업을 증진하기 위해 노력해야 한다.

82. Member States should promote AI ethics research by engaging international organizations and research institutions, as well as transnational corporations, that can be a basis for the ethical use of AI systems by public and private entities, including research into the applicability of specific ethical frameworks in specific cultures and contexts, and the possibilities to develop technologically feasible solutions in line with these frameworks.

82. 회원국은 공공・민간 단체가 인공지능 시스템을 윤리적으로 사용하는 기반이 될 수 있는 다국적 기업, 국제기구 및 연구 기관을 인공지능 윤리 연구에 관여시킴로써 이를 장려해야 하는데, 이는 특정 문화 및 배경에 특정 윤리적 틀을 적용할 수 있는 가능성 및 이러한 틀에 따라 기술적으로 구현 가능한 솔루션을 개발할 수 있는 가능성 연구를 포함한다.

83. Member States should encourage international cooperation and collaboration in the field of AI to bridge geo-technological lines.

83. 회원국은 인공지능 분야에서 국제적인 협력・협업을 장려하여 지정학적 기술연결로(geo-technological line)를 확보해야 한다.

Technological exchanges and consultations should take place between Member States and their populations, between the public and private sectors, and between and among the most and least technologically advanced countries in full respect of international law.

기술 교류 및 협의는 회원국과 자국민 사이, 공공 부문과 민간 부분 사이, 회원국 사이, 그리고 국제법에 따라 기술선도국과 기술후발국 사이에 이루어져야 한다.

## POLICY AREA 5: ENVIRONMENT AND ECOSYSTEMS

## 정책 영역 5: 환경 및 생태계

84. Member States and business enterprises should assess the direct and indirect environmental impact throughout the AI system life cycle, including, but not limited

to, its carbon footprint, energy consumption and the environmental impact of raw material extraction for supporting the manufacturing of AI technologies, and reduce the environmental impact of AI systems and data infrastructures.

84. 회원국은 인공지능 시스템 수명 주기 전 영역에서 그 자신의 탄소발자국, 에너지 소비, 인공지능 기술의 제조를 지원하기 위한 원료 추출의 환경적 영향을 비롯한 (단, 이에 국한되지 않는) 직·간접적 환경 영향을 평가하고 인공지능 시스템 및 데이터 인프라의 환경 영향을 줄이기 위하여 노력해야 한다.

Member States should ensure compliance of all AI actors with environmental law, policies and practices.

회원국은 모든 인공지능 행위 주체가 환경법·정책·관행을 준수하도록 보장해야 한다.

85. Member States should introduce incentives, when needed and appropriate, to ensure the development and adoption of rights-based and ethical AI-powered solutions for disaster risk resilience; the monitoring, protection and regeneration of the environment and ecosystems; and the preservation of the planet.

85. 필요하고 적절한 경우, 회원국은 재난 위험에 대한 회복탄력성, 환경 및 생태계 모니터링·보호·회복, 지구의 보존을 위한 인권 기반 및 윤리적 인공지능 기반의 솔루션 개발 및 채택을 보장하기 위하여 유인책을 도입해야 한다.

These AI systems should involve the participation of local and indigenous communities throughout the life cycle of AI systems and should support circular economy type approaches and sustainable consumption and production patterns.

이러한 인공지능 시스템은 수명 주기 전 영역에서 지역·토착 공동체의 참여가 수반되어야 하며, 순환 경제 유형 접근법 및 지속가능한 소비·생산 패턴을 지원해야 한다.

Some examples include using AI systems, when needed and appropriate, to:

필요하고 적절한 경우 인공지능 시스템을 사용하는 몇 가지 예는 다음과 같다.

(a) Support the protection, monitoring and management of natural resources.

(a) 천연 자원의 보호・모니터링・관리 지원.

(b) Support the prediction, prevention, control and mitigation of climate- related problems.

(b) 기후문제의 예측・예방・통제・완화 지원.

(c) Support a more efficient and sustainable food ecosystem.

(c) 더 효율적이고 지속가능한 식량 생태계 지원.

(d) Support the acceleration of access to and mass adoption of sustainable energy.

(d) 지속가능한 에너지에 대한 접근 및 대규모 도입의 가속화 지원.

(e) Enable and promote the mainstreaming of sustainable infrastructure, sustainable business models and sustainable finance for sustainable development.

(e) 지속가능한 발전에 있어 지속가능한 인프라, 지속가능한 사업 모델, 지속가능한 재정 지원의 주도 활성화 및 증진.

(f) Detect pollutants or predict levels of pollution and thus help relevant stakeholders identify, plan and put in place targeted interventions to prevent and reduce pollution and exposure.

(f) 오염 물질 탐지 또는 오염 수준 예측, 또한 오염・위험 노출의 방지 및 경감을 위한 표적 개입의 식별・계획・시행에 대한 관련 이해관계자 보조.

86. When choosing AI methods, given the potential data-intensive or resource-intensive character of some of them and the respective impact on the environment, Member States should ensure that AI actors, in line with the principle of proportionality, favour data, energy and resource-efficient AI methods.

86. 인공지능 방법론을 선택할 때, 이의 잠재적인 데이터・자원 집약적 특성 및 환경에 대한 각 영향을 고려하여, 회원국은 인공지능 행위 주체가 비례성 원칙에 따라 데이터・에너지・자원 효율적인 인공지능 방법론을 선호하게끔 보장해야 한다.

Requirements should be developed to ensure that appropriate evidence is available to show that an AI application will have the intended effect, or that safeguards accompanying an AI application can support the justification for its use.

인공지능의 적용이 의도한 효과를 불러올 것을 보여줄 수 있도록, 또는 인공지능의 적용에 수반되는 안전장치가 이의 사용에 있어 정당성을 뒷받침할 것을 보여줄 수 있도록 적절한 증거를 보장하는 요구 조건이 개발되어야 한다.

If this cannot be done, the precautionary principle must be favoured, and in instances where there are disproportionate negative impacts on the environment, AI should not be used.

이것이 불가능할 경우, 사전 주의 원칙이 반드시 우선되어야 하며 환경에 과도하게 부정적인 영향이 있을 경우 인공지능은 사용되어선 안 된다.

## POLICY AREA 6: GENDER

## 정책 영역 6: 성

87. Member States should ensure that the potential for digital technologies and artificial intelligence to contribute to achieving gender equality is fully maximized, and must ensure that the human rights and fundamental freedoms of girls and

women, and their safety and integrity are not violated at any stage of the AI system life cycle.

87. 회원국은 디지털 기술 및 인공지능이 성평등의 달성에 이바지할 잠재력이 극대화되도록 보장해야 하며, 소녀 및 여성의 인권 및 근본적 자유, 그들의 안전과 무결성이 인공지능 시스템 수명 주기의 어느 단계에서도 침해되지 않도록 반드시 보장해야 한다.

Moreover, Ethical Impact Assessment should include a transversal gender perspective.

이에 더해, 윤리영향평가에는 폭넓은 성인지적 관점이 포함되어야 한다.

88. Member States should have dedicated funds from their public budgets linked to financing gender-responsive schemes, ensure that national digital policies include a gender action plan, and develop relevant policies, for example, on labour education, targeted at supporting girls and women to make sure they are not left out of the digital economy powered by AI.

88. 회원국은 성 관련 문제에 대응하는 계획을 지원하는 공공 예산에서 자금을 확보하고, 국가 디지털 정책에 성 행동 계획이 포함되도록 보장해야 하며, 가령 노동 교육에서 소녀 및 여성이 인공지능이 주도하는 디지털 경제에서 배제되지 않도록 지원하는 데에 목적을 두는 관련 정책을 개발해야 한다.

Special investment in providing targeted programmes and gender-specific language, to increase the opportunities of girls' and women's participation in science, technology, engineering, and mathematics (STEM), including information and communication technologies (ICT) disciplines, preparedness, employability, equal career development and professional growth of girls and women, should be considered and implemented.

정보통신기술(ICT) 분야를 비롯한 과학・기술・공학・수학(STEM) 분야에서의 소녀와 여

성의 참여 기회를 증진시키고 소녀 및 여성의 준비성, 취업능력, 평등한 경력 개발 및 전문가로서의 성장을 증진시키기 위하여, 표적 프로그램 및 성을 고려한 언어를 제공함에 있어 특별한 투자를 고심하여 마련해야 한다.

89. Member States should ensure that the potential of AI systems to advance the achievement of gender equality is realized.

89. 회원국은 성평등의 달성을 앞당길 수 있는 인공지능의 잠재력이 실현되도록 해야 한다.

They should ensure that these technologies do not exacerbate the already wide gender gaps existing in several fields in the analogue world, and instead eliminate those gaps.

회원국은 이러한 과학기술이 기존의 아날로그 세상에서 여러 분야에 이미 크게 존재하는 성별 격차를 악화시키지 않도록 보장해야 하며 오히려 이를 근절해야 한다.

These gaps include: the gender wage gap; the unequal representation in certain professions and activities; the lack of representation at top management positions, boards of directors, or research teams in the AI field; the education gap; the digital and AI access, adoption, usage and affordability gap; and the unequal distribution of unpaid work and of the caring responsibilities in our societies.

성 격차란 임금격차, 특정 직업 및 활동의 대표 성별의 불균형, 인공지능 분야의 최고 경영진・이사회・연구팀에서의 성별 대표성 부족, 교육 격차, 디지털・인공지능에 대한 접근・도입・사용・소비 격차, 우리 사회에서의 무임노동 및 양육의무의 편중성을 포함한다.

90. Member States should ensure that gender stereotyping and discriminatory biases are not translated into AI systems, and instead identify and proactively redress these.

90. 회원국은 성별 고정 관념과 차별적 편향이 인공지능 시스템으로 전이되지 않도록 보장해야 하며, 오히려 이를 판별하여 적극적으로 바로잡아야 한다.

Efforts are necessary to avoid the compounding negative effect of technological divides in achieving gender equality and avoiding violence such as harassment, bullying or trafficking of girls and women and under-represented groups, including in the online domain.

성평등을 달성하고 소녀, 여성, (온라인도 포함하는) 소외 계층에 대한 괴롭힘, 따돌림, 인신매매와 같은 폭력을 피함에 있어서, 기술 격차의 복합적이고 부정적인 효과를 피하기 위한 노력이 필요하다.

91. Member States should encourage female entrepreneurship, participation and engagement in all stages of an AI system life cycle by offering and promoting economic, regulatory incentives, among other incentives and support schemes, as well as policies that aim at a balanced gender participation in AI research in academia, gender representation on digital and AI companies' top management positions, boards of directors and research teams.

91. 회원국은 다른 유인책 및 지원 계획보다도 경제・규제적 유인책을 제공, 장려함으로써, 그리고 학계 인공지능 연구에서의 균형 잡힌 성별 참여 및 디지털・인공지능 기업의 최고경영진・이사회・연구팀에 대한 균형 잡힌 성별 대표성을 지향하는 정책을 제공, 장려함으로써, 인공지능 시스템 수명 주기의 모든 단계에서 여성 기업가정신・참여・개입을 장려해야 한다.

Member States should ensure that public funds (for innovation, research and technologies) are channelled to inclusive programmes and companies, with clear gender representation, and that private funds are similarly encouraged through affirmative action principles.

정부는 (혁신, 연구, 과학기술에 관한) 공공 기금이 분명한 성별 대표성을 가진 포용적인 프로그램・기업에 전달되도록 보장해야 하며, 사적 기금도 마찬가지 방식으로 적극적 행동 원칙을 통해 장려되도록 해야 한다.

Policies on harassment-free environments should be developed and enforced, together with the encouragement of the transfer of best practices on how to promote diversity throughout the AI system life cycle.

인공지능 시스템 수명 주기 전 영역에서 다양성을 증진할 방법에 대한 모범 사례의 전파를 장려함과 동시에, 괴롭힘 없는 환경에 대한 정책이 개발 및 시행되어야 한다.

92. Member States should promote gender diversity in AI research in academia and industry by offering incentives to girls and women to enter the field, putting in place mechanisms to fight gender stereotyping and harassment within the AI research community, and encouraging academic and private entities to share best practices on how to enhance gender diversity.

92. 회원국은 소녀 및 여성이 인공지능 분야에 진출할 수 있도록 유인책을 제공하고, 인공지능 연구 공동체 내의 성 고정관념 및 괴롭힘에 대항할 수 있는 메커니즘을 마련하고, 학술・민간 단체가 성별 다양성을 강화하는 방법에 대한 모범 사례를 공유하도록 장려함으로써, 학계 및 산업의 인공지능 연구에서 성 다양성을 증진해야 한다.

93. UNESCO can help form a repository of best practices for incentivizing the participation of girls, women and under-represented groups in all stages of the AI system life cycle.

93. 유네스코는 인공지능 수명 주기의 모든 단계에서 소녀, 여성, 소외 계층의 참여에 유인책을 제공하는 모범 사례 저장소의 구축을 도울 수 있다.

## POLICY AREA 7: CULTURE

## 정책 영역7: 문화

94. Member States are encouraged to incorporate AI systems, where appropriate, in the preservation, enrichment, understanding, promotion, management and accessibility of tangible, documentary and intangible cultural heritage, including endangered languages as well as indigenous languages and knowledges, for example by introducing or updating educational programmes related to the application of AI systems in these areas, where appropriate, and by ensuring a participatory approach, targeted at institutions and the public.

94. 회원국은 토착어 및 토착지식, 희소언어를 비롯한 유형・무형문화유산의 보존・강화・이해・홍보・관리・접근성에 있어 적합한 경우 인공지능 시스템을 활용하도록 권장되는데, 가령 기관 및 대중을 대상으로 이러한 분야에 인공지능 시스템의 응용과 관련된 하는 교육 프로그램을 도입 또는 갱신하며 참여적 접근을 보장하는 것이다.

95. Member States are encouraged to examine and address the cultural impact of AI systems, especially natural language processing (NLP) applications such as automated translation and voice assistants, on the nuances of human language and expression.

95. 회원국은 인공지능 시스템, 특히 인간 언어・표현의 뉘앙스에 대한 자동 번역 및 음성 도우미와 같은 자연어처리(NLP) 응용의 문화적 영향을 조사하고 그에 대응하도록 장려된다.

Such assessments should provide input for the design and implementation of strategies that maximize the benefits from these systems by bridging cultural gaps and increasing human understanding, as well as addressing the negative implications such as the reduction of use, which could lead to the disappearance of endangered languages, local dialects, and tonal and cultural variations associated

with human language and expression.

이러한 평가는 문화적 격차를 해소하고 인간에 대한 이해를 증가시킴으로써, 그리고 희소 언어, 지역 방언, 인간 언어・표현과 관련된 음성적・문화적 변주의 소실로 이어질 수 있는 언어 사용 감소와 같은 부정적 영향을 해결함으로써, 이런 시스템의 혜택을 극대화하는 전략의 설계 및 구현을 위한 정보를 제공해야 한다.

96. Member States should promote AI education and digital training for artists and creative professionals to assess the suitability of AI technologies for use in their profession, and contribute to the design and implementation of suitable AI technologies, as AI technologies are being used to create, produce, distribute, broadcast and consume a variety of cultural goods and services, bearing in mind the importance of preserving cultural heritage, diversity and artistic freedom.

96. 인공지능 기술이 다양한 문화 상품・서비스를 창출・생산・유통・방송・소비하는 데에 사용되고 있기에, 회원국은 문화 유산, 다양성, 예술적 자유의 보존이 중요함을 유념하여, 예술가 및 창조적 직업군이 자신의 직군에서의 인공지능 기술 사용 적합성을 평가할 수 있도록 인공지능・디지털 교육을 장려해야 한다.

97. Member States should promote awareness and evaluation of AI tools among local cultural industries and small and medium enterprises working in the field of culture, to avoid the risk of concentration in the cultural market.

97. 회원국은 문화계에 종사하는 지역문화기업 및 중소기업의 인공지능 도구에 대한 인식 및 평가를 제고하여 문화 시장에서 편중화의 위험성을 예방하여야 한다.

98. Member States should engage technology companies and other stakeholders to promote a diverse supply of and plural access to cultural expressions, and in particular to ensure that algorithmic recommendation enhances the visibility and discoverability of local content.

98. 회원국은 문화적 표현에 대한 다양한 공급 및 다원주의적 접근을 증진하는 일과, 특히 알고리즘의 추천이 지역 컨텐츠를 더 드러내고 나타내게끔 하는 일에 IT기업 및 다른 이해관계자를 끌어들여야 한다.

99. Member States should foster new research at the intersection between AI and intellectual property (IP), for example to determine whether or how to protect with IP rights the works created by means of AI technologies.

99. 회원국은, 가령 인공지능 기술이 생성한 결과물의 지적재산권의 보호 여부 및 방식에 대해 결정하는 경우와 같이, 인공지능과 지적재산권의 접점에서 새로운 연구를 촉진해야 한다.

Member States should also assess how AI technologies are affecting the rights or interests of IP owners, whose works are used to research, develop, train or implement AI applications.

회원국은 인공지능 응용 연구・개발・학습・구현에 사용되는 저작물에 대한 지적재산권 소유자의 권리 또는 이해에 인공지능 기술이 어떤 식으로 영향을 미치는지도 또한 평가해야 한다.

100. Member States should encourage museums, galleries, libraries and archives at the national level to use AI systems to highlight their collections and enhance their libraries, databases and knowledge base, while also providing access to their users.

100. 회원국은 박물관, 미술관, 도서관, 기록보관소가 인공지능 시스템을 개발 및 사용하여 소장품을 부각시키고 도서관, 데이터베이스, 지식 베이스의 수준을 높이도록 국가적 차원에서 장려해야 함과 동시에 그 사용자에게 접근권을 부여해야 한다.

## POLICY AREA 8: EDUCATION AND RESEARCH

## 정책 영역 8: 교육 및 연구

101. Member States should work with international organizations, educational institutions and private and non-governmental entities to provide adequate AI literacy education to the public on all levels in all countries in order to empower people and reduce the digital divides and digital access inequalities resulting from the wide adoption of AI systems.

101. 인간의 역량을 강화하고 인공지능 시스템의 광범위한 도입에서 파생되는 디지털 격차 및 디지털 접근불평등을 줄이기 위해서, 모든 국가의 모든 수준의 대중에게 적절한 인공지능 리터러시 교육을 제공할 수 있도록 국제기구, 교육 기관, 민간 단체, 비정부단체와 협력해야 한다.

102. Member States should promote the acquisition of "prerequisite skills" for AI education, such as basic literacy, numeracy, coding and digital skills, and media and information literacy, as well as critical and creative thinking, teamwork, communication, socio-emotional and AI ethics skills, especially in countries and in regions or areas within countries where there are notable gaps in the education of these skills.

102. 회원국은 기본언어능력, 산술능력, 코딩・디지털 능력, 미디어・정보 리터러시뿐만 아니라 비판적 사고, 팀워크, 사회정서적 능력, 인공지능 윤리 능력과 같은 인공지능 교육을 위한 '기초 소양'의 습득을 특히 이러한 기초소양 교육의 격차가 두드러지는 국가 및 이러한 국가 내 지역・구역에서 장려하여야 한다.

103. Member States should promote general awareness programmes about AI developments, including on data and the opportunities and challenges brought

about by AI technologies, the impact of AI systems on human rights and their implications, including children's rights.

103. 회원국은 데이터, 인공지능 기술이 초래하는 기회 및 어려움, 유아의 권리를 비롯한 인권에 대한 인공지능 시스템의 영향 및 함의를 비롯하여 인공지능 개발에 대한 보편 인식 프로그램을 장려해야 한다.

These programmes should be accessible to non-technical as well as technical groups.

이러한 프로그램은 기술전문가집단뿐만 아니라 비전문집단에게도 이용이 쉬워야 한다.

104. Member States should encourage research initiatives on the responsible and ethical use of AI technologies in teaching, teacher training and e-learning, among other issues, to enhance opportunities and mitigate the challenges and risks involved in this area.

104. 회원국은 다른 사안보다도 교육, 교사 연수, 온라인 교육에서 인공지능의 윤리적이고 책임 있는 사용에 대한 연구 계획을 장려하여 이 분야와 관련된 기회는 향상시키고 어려움 및 위험성은 완화해야 한다.

The initiatives should be accompanied by an adequate assessment of the quality of education and impact on students and teachers of the use of AI technologies.

이 계획에는 교육의 질에 대한 적절한 평가, 인공지능 기술의 사용이 학생 및 교사에 미치는 영향에 대한 적절한 평가가 수반되어야 한다.

Member States should also ensure that AI technologies empower students and teachers and enhance their experience, bearing in mind that relational and social aspects and the value of traditional forms of education are vital in teacher-student and student-student relationships and should be considered when discussing the

adoption of AI technologies in education.

회원국은 전통적 교육 형태의 관계 · 사회적 측면 및 가치가 교사-학생 및 학생-교사 관계에서 매우 중요하며 이것이 교육으로의 인공지능 기술의 도입을 논할 때 고려되어야 한다는 점을 유념한 채로, 인공지능 기술이 학생 및 교사의 역량을 강화하고 경험을 증대시킬 수 있도록 보장해야 한다.

AI systems used in learning should be subject to strict requirements when it comes to the monitoring, assessment of abilities, or prediction of the learners' behaviours.

모니터링, 능력 평가, 학습자의 행동 예측에 관한 것이라면, 학습에 사용되는 인공지능 시스템은 엄격한 요구조건 하에 있어야 한다.

AI should support the learning process without reducing cognitive abilities and without extracting sensitive information, in compliance with relevant personal data protection standards.

인공지능은 관련 개인 데이터 보호 기준에 따라 인지능력의 약화 및 민감정보의 유출 없이 학습과정을 보조해야 한다.

The data handed over to acquire knowledge collected during the learner's interactions with the AI system must not be subject to misuse, misappropriation or criminal exploitation, including for commercial purposes.

학습자와 인공지능 간의 상호작용 동안 축적된 지식을 습득하기 위해 양도된 데이터는 상업적 사용을 비롯하여 오용, 유용, 또는 범죄적으로 사용되지 않아야 한다.

105. Member States should promote the participation and leadership of girls and women, diverse ethnicities and cultures, persons with disabilities, marginalized and vulnerable people or people in vulnerable situations, minorities and all persons not

enjoying the full benefits of digital inclusion, in AI education programmes at all levels, as well as the monitoring and sharing of best practices in this regard with other Member States.

105. 회원국은 모든 수준의 인공지능 교육 프로그램에서 소녀 및 여성, 다양한 민족 및 문화, 장애인, 소외・취약 계층 또는 취약한 상황에 처한 사람, 소수자, 디지털 포용의 완전한 혜택을 향유하지 못하는 모든 사람의 참여 및 주도를 장려함과 더불어 이에 대한 다른 회원국과의 모범 사례 모니터링・공유를 증진해야 한다.

106. Member States should develop, in accordance with their national education programmes and traditions, AI ethics curricula for all levels, and promote cross-collaboration between AI technical skills education and humanistic, ethical and social aspects of AI education.

106. 회원국은 자국의 교육 프로그램 및 전통에 따라 전 학년에 인공지능 윤리 교육과정을 개발하고, 기술 숙련 교육과 인공지능 교육의 인문학・윤리・사회적 측면 간의 교차협력을 증진해야 한다.

Online courses and digital resources of AI ethics education should be developed in local languages, including indigenous languages, and take into account the diversity of environments, especially ensuring accessibility of formats for persons with disabilities.

인공지능 윤리 교육의 온라인 과정 및 디지털 자료는 원주민 언어를 비롯한 토착 언어로 개발되어야 하며 환경의 다양성을 고려해야 하는데, 특히 장애인이 접근 가능한 형태로 해야 한다.

107. Member States should promote and support AI research, notably AI ethics research, including for example through investing in such research or by creating incentives for the public and private sectors to invest in this area, recognizing that

research contributes significantly to the further development and improvement of AI technologies with a view to promoting international law and the values and principles set forth in this Recommendation.

107. 국제법 및 본 권고에서 제시한 가치・원칙을 증진하기 위한 인공지능 기술의 향후 발전 및 개선에 연구가 상당히 이바지한다는 점을 인지하여, 회원국은 연구투자를 통하여, 또는 공공・민간 부문으로 하여금 이 분야에 투자하게 하는 유인책을 만듦으로써, 인공지능 연구, 특히 인공지능 윤리연구를 장려・지원해야 한다.

Member States should also publicly promote the best practices of, and cooperation with, researchers and companies who develop AI in an ethical manner.

회원국은 또한 윤리적으로 인공지능을 개발하는 연구자・기업의 모범 사례를 홍보하고 이와의 협력을 증진해야 한다.

108. Member States should ensure that AI researchers are trained in research ethics and require them to include ethical considerations in their designs, products and publications, especially in the analyses of the datasets they use, how they are annotated, and the quality and scope of the results with possible applications.

108. 회원국은 연구 윤리가 인공지능 연구자에게 체화되도록 보장해야 하며, 그들이 설계, 생산, 공개의 단계에서, 특히 사용되는 데이터 집합의 분석, 주석표기방식, 가능한 응용결과의 품질 및 범위에 대하여 윤리적 사항을 고려하도록 요구해야 한다.

109. Member States should encourage private sector companies to facilitate the access of the scientific community to their data for research, especially in LMICs, in particular LDCs, LLDCs and SIDS.

109. 회원국은 특히 중・저소득국가, 그 중에서도 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국에서 민간 기업으로 하여금 과학계가 연구를 위해 데이터에 쉽게 접근하는 것을 허용하도록 장려해야 한다.

This access should conform to relevant privacy and data protection standards.

이러한 접근은 관련 프라이버시・데이터 보호 기준에 따라야 한다.

110. To ensure a critical evaluation of AI research and proper monitoring of potential misuses or adverse effects, Member States should ensure that any future developments with regards to AI technologies should be based on rigorous and independent scientific research, and promote interdisciplinary AI research by including disciplines other than science, technology, engineering and mathematics (STEM), such as cultural studies, education, ethics, international relations, law, linguistics, philosophy, political science, sociology and psychology.

110. 인공지능 연구의 비판적 심사와 잠재적 오용 또는 악영향에 대한 적절한 모니터링을 보장하기 위해서, 회원국은 인공지능과 관련된 향후 어떤 개발도 엄격하고도 독립적인 과학적 연구에 기반을 두게 보장해야 하며, 문화학・교육학・윤리학・국제관계학・법학・언어학・철학・정치학・사회학・심리학과 같이 과학・기술・공학・수학(STEM) 외 다른 학문분야를 포함시킴으로써 학제적 인공지능 연구를 장려해야 한다.

111. Recognizing that AI technologies present great opportunities to help advance scientific knowledge and practice, especially in traditionally model-driven disciplines, Member States should encourage scientific communities to be aware of the benefits, limits and risks of their use; this includes attempting to ensure that conclusions drawn from data-driven approaches, models and treatments are robust and sound.

111. 인공지능 기술이 특히 전통적 모델이 주도하는 학문분야에서 과학 지식・실습의 발전을 돕는 데에 큰 기회를 제공한다는 점을 인지하여, 회원국은 과학계가 인공지능 기술 사용의 혜택, 한계, 위험성을 인식하도록 장려해야 한다. 이는 데이터 주도의 접근・모델・방안으로부터 도출된 결론이 타당하고 건전하게끔 보장하는 노력을 포함한다.

Furthermore, Member States should welcome and support the role of the scientific community in contributing to policy and in cultivating awareness of the strengths and weaknesses of AI technologies.

이에 더해, 회원국은 정책 기여, 그리고 인공지능 기술의 장단점에 대한 인식을 배양하는 데에 있어 과학계의 역할을 환영하고 지지해야 한다.

## POLICY AREA 9: COMMUNICATION AND INFORMATION

## 정책 영역9: 정보통신

112. Member States should use AI systems to improve access to information and knowledge.

112. 회원국은 정보 및 지식에 대한 접근을 향상시키기 위해 인공지능 시스템을 사용해야 한다.

This can include support to researchers, academia, journalists, the general public and developers, to enhance freedom of expression, academic and scientific freedoms, access to information, and increased proactive disclosure of official data and information.

여기에는 연구자, 학계, 기자, 대중, 개발자의 표현의 자유, 학술・과학적 자유 정보 접근을 향상시키기 위한 지원과 공공 데이터・정보의 사전 공개가 포함될 수 있다.

113. Member States should ensure that AI actors respect and promote freedom of expression as well as access to information with regard to automated content generation, moderation and curation.

113. 회원국은 인공지능 행위 주체가 표현의 자유를 존중 및 증진함과 더불어 자동 컨텐츠 생성・조정・선별에 관한 정보에도 접근하도록 보장해야 한다.

Appropriate frameworks, including regulation, should enable transparency of online communication and information operators and ensure users have access to a diversity of viewpoints, as well as processes for prompt notification to the users on the reasons for removal or other treatment of content, and appeal mechanisms that allow users to seek redress.

규제를 비롯한 적절한 틀은 온라인 정보통신 운영을 투명하게 만들어야 하고, 컨텐츠의 제거 또는 기타 처리를 이유로 사용자에게 즉각 통보하는 프로세스 및 사용자의 배상 요구를 가능하게 하는 상고 메커니즘과 더불어 사용자가 관점의 다양성을 접할 수 있도록 보장해야 한다.

114. Member States should invest in and promote digital and media and information literacy skills to strengthen critical thinking and competencies needed to understand the use and implication of AI systems, in order to mitigate and counter disinformation, misinformation and hate speech.

114. 회원국은 인공지능 시스템의 사용 및 함의를 이해하는 데에 필요한 비판적 사고 및 역량을 강화하기 위해, 디지털 · 미디어 · 정보 리터러시 숙련에 투자하고 이를 장려하여서 허위 정보, 오보, 혐오 발언을 경감시키고 이에 대응하여야 한다.

A better understanding and evaluation of both the positive and potentially harmful effects of recommender systems should be part of those efforts.

추천 시스템의 긍정적 효과 및 잠재적 해악 효과에 대한 더 나은 이해 및 심사는 이러한 노력의 일환이 되어야 한다.

115. Member States should create enabling environments for media to have the rights and resources to effectively report on the benefits and harms of AI systems, and also encourage media to make ethical use of AI systems in their operations.

115. 회원국은 미디어가 인공지능 시스템의 장단점을 실제적으로 보도할 권리 및 자원을

가질 수 있는 환경을 조성해주어야 하며, 그 운영에 인공지능 시스템을 윤리적으로 활용하도록 장려해야 한다.

## POLICY AREA 10: ECONOMY AND LABOUR

## 정책 영역 10: 경제 및 노동

116. Member States should assess and address the impact of AI systems on labour markets and its implications for education requirements, in all countries and with special emphasis on countries where the economy is labour-intensive.

116. 회원국은 모든 국가에서, 특히 노동집약적 국가에 중점을 두고, 인공지능 시스템이 노동 시장에 미치는 영향과 교육 요구조건에 갖는 함의를 평가하고 이에 대응해야 한다.

This can include the introduction of a wider range of "core" and interdisciplinary skills at all education levels to provide current workers and new generations a fair chance of finding jobs in a rapidly changing market, and to ensure their awareness of the ethical aspects of AI systems.

급변하는 시장에서 신세대에게 공정한 구직 기회를 제공하고 인공지능 시스템의 윤리적 측면에 대한 인식을 제고하기 위해서, 이 대응에는 모든 교육 수준에서 광범위한 '핵심적' 및 학제적 능력의 도입이 포함될 수 있다.

Skills such as "learning how to learn", communication, critical thinking, teamwork, empathy, and the ability to transfer one's knowledge across domains, should be taught alongside specialist, technical skills, as well as low-skilled tasks.

'학습 방법 습득', 의사소통, 비판적 사고, 팀워크, 공감, 분야를 횡단하여 지식을 전달할 수 있는 능력과 같은 능력은 전문가, 전문 기술은 물론 저숙련 노동과도 함께 가르쳐야 한다.

Being transparent about what skills are in demand and updating curricula around these are key.

어떤 숙련이 수요가 있는지 투명하게 공개하고 이를 중심으로 커리큘럼을 갱신하는 것이 중요하다.

117. Member States should support collaboration agreements among governments, academic institutions, vocational education and training institutions, industry, workers' organizations and civil society to bridge the gap of skillset requirements to align training programmes and strategies with the implications of the future of work and the needs of industry, including small and medium enterprises.

117. 회원국은 학술 기관, 직업교육훈련기관, 업계, 근로자단체, 시민사회 간의 협업 계약을 지원하여, 노동의 미래 및 중소기업을 비롯한 산업의 요구에 맞추어 교육 프로그램 및 전략을 조정하기 위해서 기술요구요건의 격차를 줄여야 한다.

Project-based teaching and learning approaches for AI should be promoted, allowing for partnerships between public institutions, private sector companies, universities and research centres.

공공기관, 민간기업, 대학, 연구소 간의 파트너십을 허용함으로써, 프로젝트 기반의 인공지능 교육・학습 접근법을 장려해야 한다.

118. Member States should work with private sector companies, civil society organizations and other stakeholders, including workers and unions to ensure a fair transition for at-risk employees.

118. 회원국은 불안정성에 노출된 근로자의 정당한 직업 전환을 보장하기 위해 민간 기업, 시민사회, 및 노동자・노동조합을 비롯한 기타 이해관계자들과 협력해야 한다.

This includes putting in place upskilling and reskilling programmes, finding effective mechanisms of retaining employees during those transition periods, and exploring “safety net” programmes for those who cannot be retrained.

이는 숙련향상・재숙련 교육 프로그램을 마련하는 것, 전환 기간 동안 근로자를 보존하는 효과적인 방식을 찾아내는 것, 재훈련이 어려운 사람들을 위한 ‘안전망’ 프로그램을 강구하는 것을 포함한다.

Member States should develop and implement programmes to research and address the challenges identified that could include upskilling and reskilling, enhanced social protection, proactive industry policies and interventions, tax benefits, new taxation forms, among others.

회원국은 무엇보다 숙련향상・재숙련 교육, 사회 보호 강화, 적극적 산업 정책・개입, 세금 혜택, 새로운 과세 양식 등을 포함할 수 있는 것으로 알려진 과제를 연구 및 해결할 수 있는 프로그램을 개발 및 시행해야 한다.

Member States should ensure that there is sufficient public funding to support these programmes.

회원국은 이러한 프로그램을 지원할 공적 기금이 충분하도록 보장해야 한다.

Relevant regulations, such as tax regimes, should be carefully examined and changed if needed to counteract the consequences of unemployment caused by AI-based automation.

인공지능 기반 자동화가 초래한 실업의 결과에 대응하기 위해서, 필요한 경우 세제와 같은 관련 규제를 면밀히 검토하고 변경해야 한다.

119. Member States should encourage and support researchers to analyse the impact of AI systems on the local labour environment in order to anticipate future

trends and challenges.

119. 회원국은 미래의 트렌드 및 어려움을 예측하기 위해, 연구자들로 하여금 인공지능 시스템이 지역 노동 시장에 미치는 영향을 분석하도록 장려・지원해야 한다.

These studies should have an interdisciplinary approach and investigate the impact of AI systems on economic, social and geographic sectors, as well as on human-robot interactions and human-human relationships, in order to advise on reskilling and redeployment best practices.

이러한 연구는 재숙련 교육 및 직업이동의 모범 사례에 대해 의논하기 위해, 인공지능이 경제・사회・지리 부문에 미치는 영향뿐만 아니라 인간로봇상호작용(HRI) 및 인간간 관계에도 미치는 영향을 조사해야 한다.

120. Member States should take appropriate steps to ensure competitive markets and consumer protection, considering possible measures and mechanisms at national, regional and international levels, to prevent abuse of dominant market positions, including by monopolies, in relation to AI systems throughout their life cycle, whether these are data, research, technology, or market.

120. 인공지능 시스템 수명 주기 전 영역에서, 데이터・연구・과학기술・시장 어떤 것이든지 간에 인공지능 시스템과 관련이 있는 것은 독점을 비롯하여 시장에서의 지배적 지위 남용을 방지하기 위해서, 회원국은 국가・지역・국제적 수준의 가능한 조치 및 메커니즘을 고려함으로써 경쟁력 있는 시장 및 소비자 보호를 보장하기 위한 적절한 대책을 세워야 한다.

Member States should prevent the resulting inequalities, assess relevant markets and promote competitive markets.

회원국은 이로 발생하는 불평등을 예방하여 관련 시장은 과세하고 경쟁 시장은 장려해야 한다.

Due consideration should be given to LMICs, in particular LDCs, LLDCs and SIDS, which are more exposed and vulnerable to the possibility of abuses of market dominance as a result of a lack of infrastructure, human capacity and regulations, among other factors.

다른 요인 보다도, 인프라, 인적역량, 규제 부족에 대한 결과로 시장 지배의 남용 가능성에 더 노출되고 취약한 중・저소득국가, 특히 최빈개발도상국, 내륙개발도상국, 군소도서개발도상국에 대한 고려가 어루어져야 한다.

AI actors developing AI systems in countries which have established or adopted ethical standards on AI should respect these standards when exporting these products, developing or applying their AI systems in countries where such standards may not exist, while respecting applicable international law and domestic legislation, standards and practices of these countries.

인공지능에 대한 윤리 기준을 수립・도입한 국가에서 인공지능 시스템을 개발하는 인공지능 행위 주체는 상품을 수출하거나 이러한 기준이 부재하는 국가에서 인공지능 시스템을 개발・적용할 때 이러한 기준을 존중해야 하며, 또한 적용 가능한 국제법 및 해당 국가의 국내법・기준・관행을 존중해야 한다.

## POLICY AREA 11: HEALTH AND SOCIAL WELL-BEING

## 정책 영역 11: 건강 및 사회 복지

121. Member States should endeavour to employ effective AI systems for improving human health and protecting the right to life, including mitigating disease outbreaks, while building and maintaining international solidarity to tackle global health risks and uncertainties, and ensure that their deployment of AI systems in health care be consistent with international law and their human rights law obligations.

121. 회원국은 전 지구적 건강 위험성·불확실성을 해결하기 위해 국제적 연대를 구축 및 유지함과 동시에 인류의 건강을 개선하고 전염병 발생을 완화하며 생명권을 보호하기 위해 효과적인 인공지능 시스템을 활용하려고 노력해야 하며, 건강관리에서의 인공지능 시스템 활용이 국제법 및 인권준수의무에 부합하도록 해야 한다.

Member States should ensure that actors involved in health care AI systems take into consideration the importance of a patient's relationships with their family and with health care staff.

회원국은 건강관리 인공지능 시스템에 관여되어 있는 행위 주체가 환자와 가족 간 관계 및 환자와 의료 직원 간 관계의 중요성을 고려하도록 해야 한다.

122. Member States should ensure that the development and deployment of AI systems related to health in general and mental health in particular, paying due attention to children and youth, is regulated to the effect that they are safe, effective, efficient, scientifically and medically proven and enable evidence-based innovation and medical progress.

122. 회원국은 아동 및 청년에 마땅히 관심을 가지고 일반 건강 및 특히 정신 건강과 관련된 인공지능 시스템의 개발 및 활용을 규제하여, 이것이 안전하고 효과적이고 효율적이며 과학·의학적으로 증명된 것임을 보장해야 하고, 증거 기반 혁신 및 의학적 진보가 가능하도록 해야 한다.

Moreover, in the related area of digital health interventions, Member States are strongly encouraged to actively involve patients and their representatives in all relevant steps of the development of the system.

더욱이, 디지털 보건 개입의 관련 영역에서, 회원국은 적극적으로 환자 및 보호자를 시스템 개발의 모든 관련된 단계에 참여시키도록 강력히 권장된다.

123. Member States should pay particular attention in regulating prediction, detection and treatment solutions for health care in AI applications by:

123. 회원국은 인공지능 응용에 있어, 건강관리를 위한 예측・탐지・치료 솔루션을 규제할 때 다음과 같이 각별한 주의를 기울여야 한다.

(a) ensuring oversight to minimize and mitigate bias;

(a) 편향을 최소화・축소하도록 감독을 보장함.

(b) ensuring that the professional, the patient, caregiver or service user is included as a “domain expert” in the team in all relevant steps when developing the algorithms;

(b) 알고리즘 개발시, 의료전문가, 환자, 간병인 또는 서비스 사용자가 관련 있는 모든 단계에서 개발팀의 ‘분야 전문가’로 포함되도록 보장함.

(c) paying due attention to privacy because of the potential need for being medically monitored and ensuring that all relevant national and international data protection requirements are met;

(c) 치료 모니터링의 잠재적 필요성으로 인해 생기는 프라이버시에 대해 적절한 주의를 다하고, 관련 있는 모든 국가・국제적 데이터 보호 요구 조건의 충족을 보장함.

(d) ensuring effective mechanisms so that those whose personal data is being analysed are aware of and provide informed consent for the use and analysis of their data, without preventing access to health care;

(d) 개인 데이터 분석 대상자들이 자신의 데이터 사용・분석에 대해 인지하고 인지 동의를 제공하도록 하는 효과적인 메커니즘을 보장함.

(e) ensuring the human care and final decision of diagnosis and treatment are taken always by humans while acknowledging that AI systems can also assist in their work;

(e) 인공지능 시스템이 의료에 도움이 될 수 있음을 인정함과 동시에, 인간이 간호 및 진단・치료에 대한 최종 결정을 내리도록 보장함.

(f) ensuring, where necessary, the review of AI systems by an ethical research committee prior to clinical use.

(f) 필요한 경우, 의료적 사용에 앞서 윤리연구위원회의 인공지능 시스템 검토를 보장함.

124. Member States should establish research on the effects and regulation of potential harms to mental health related to AI systems, such as higher degrees of depression, anxiety, social isolation, developing addiction, trafficking, radicalization and misinformation, among others.

124. 회원국은 심한 우울증, 불안, 사회적 고립, 중독・인신매매・과격화・오보 발현 등 인공지능 시스템이 미칠 수 있는 정신 건강과 관련된 잠재적 유해성의 영향 및 규제에 관한 연구를 확립해야 한다.

125. Member States should develop guidelines for human-robot interactions and their impact on human-human relationships, based on research and directed at the future development of robots, and with special attention to the mental and physical health of human beings.

125. 회원국은 연구에 기반을 두고 향후 로봇 개발을 겨냥하여서, 인간의 정신・신체 건강에 각별한 주의를 가지고 인간로봇상호작용 및 이것이 인간 간 관계에 미치는 영향에 대한 지침을 개발해야 한다.

Particular attention should be given to the use of robots in health care and the care for older persons and persons with disabilities, in education, and robots for use by children, toy robots, chatbots and companion robots for children and adults.

특히 노인 및 장애인을 위한 건강관리・간호 로봇, 교육용 로봇, 장난감 로봇, 챗봇, 유아 및 성인용 반려로봇의 사용에 관해서는 각별한 주의를 기울여야 한다.

Furthermore, assistance of AI technologies should be applied to increase the safety and ergonomic use of robots, including in a human-robot working environment.

이에 더해, 인간-로봇 작업환경 등에서 로봇의 안전하고 인체공학적 사용을 높이기 위하여 인공지능 기술이 보조되어야 한다.

Special attention should be paid to the possibility of using AI to manipulate and abuse human cognitive biases.

인간인지편향을 조작・남용하는 데에 인공지능이 사용될 가능성에 관해 각별한 주의를 기울여야 한다.

126. Member States should ensure that human-robot interactions comply with the same values and principles that apply to any other AI systems, including human rights and fundamental freedoms, the promotion of diversity, and the protection of vulnerable people or people in vulnerable situations.

126. 회원국은 인간로봇상호작용이 인권 및 근본적 자유, 다양성 촉진, 취약계층 및 취약한 상황에 처한 사람 보호 등 다른 모든 인공지능 시스템에 동일하게 적용되는 가치 및 원칙을 준수하도록 보장해야 한다.

Ethical questions related to AI-powered systems for neurotechnologies and brain-computer interfaces should be considered in order to preserve human

dignity and autonomy.

인간 존엄성・자율성을 보존하기 위하여 회원국은 인공지능 기반 신경공학 시스템 및 뇌-컴퓨터 인터페이스와 관련된 윤리적 질문을 숙고해야 한다.

127. Member States should ensure that users can easily identify whether they are interacting with a living being, or with an AI system imitating human or animal characteristics, and can effectively refuse such interaction and request human intervention.

127. 회원국은 사용자가 자신이 생명체와 상호작용하는지, 아니면 인간 또는 동물적 특성을 모방한 인공지능 시스템과 상호작용하는지 쉽게 파악할 수 있도록 해야 하며, 사용자가 그러한 상호작용을 거부하고 인간 개입을 실제적으로 요청할 수 있도록 해야 한다.

128. Member States should implement policies to raise awareness about the anthropomorphization of AI technologies and technologies that recognize and mimic human emotions, including in the language used to mention them, and assess the manifestations, ethical implications and possible limitations of such anthropomorphization, in particular in the context of robot-human interaction and especially when children are involved.

128. 회원국은 인공지능을 언급할 때 사용되는 언어 등에서 인공지능 기술 및 인간 감정을 인지・모방하는 과학기술의 인격화에 관한 인지도를 높이는 정책을 시행해야 하며, 인간 로봇상호작용의 맥락에서, 특히 유아가 연관되어있는 경우, 그러한 인격화의 표면적 현상, 윤리적 함의, 한계를 평가해야 한다.

129. Member States should encourage and promote collaborative research into the effects of long-term interaction of people with AI systems, paying particular attention to the psychological and cognitive impact that these systems can have on children and young people.

129. 회원국은 인공지능 시스템이 아동 및 청년에게 미칠 수 있는 정신적 · 인지적 영향에 각별히 주의를 기울여, 회원국은 사람과 인공지능 시스템 간 장기적 상호작용의 영향에 대한 공동 연구를 장려 및 촉진해야 한다.

This should be done using multiple norms, principles, protocols, disciplinary approaches, and assessment of the modification of behaviours and habits, as well as careful evaluation of the downstream cultural and societal impacts.

이는 하위문화의 문화 · 사회적 영향에 대한 신중한 심사를 비롯하여 여러가지 규범, 원칙, 프로토콜, 징계 수단, 행동 · 습관의 시정 평가를 사용함으로써 가능하다.

Furthermore, Member States should encourage research on the effect of AI technologies on health system performance and health outcomes.

이에 더해, 회원국은 건강 시스템 성능 및 건강 결과물에 관한 인공지능 기술의 영향에 관한 연구를 장려해야 한다.

130. Member States, as well as all stakeholders, should put in place mechanisms to meaningfully engage children and young people in conversations, debates and decision-making with regard to the impact of AI systems on their lives and futures.

130. 모든 이해관계자와 더불어 회원국은 아동 및 청년이 삶과 미래에 미치는 인공지능 시스템의 영향에 관한 대화 · 토론 · 의사결정에 유의미하게 참여할 수 있도록 하는 메커니즘을 마련해야 한다.

## V. MONITORING AND EVALUATION

## V. 모니터링 및 평가

131. Member States should, according to their specific conditions, governing structures and constitutional provisions, credibly and transparently monitor and evaluate policies, programmes and mechanisms related to ethics of AI, using a combination of quantitative and qualitative approaches.

131. 회원국은 자국의 특정 조건, 지배 구조, 헌법 조항에 따라 양적・질적 접근법을 조합하여, 인공지능 윤리와 관련된 정책・프로그램・메커니즘을 신뢰할 수 있고 투명하게 모니터링하고 평가해야 한다.

To support Member States, UNESCO can contribute by:

유네스코는 다음과 같이 회원국을 지원할 수 있다.

(a) developing a UNESCO methodology for Ethical Impact Assessment (EIA) of AI technologies based on rigorous scientific research and grounded in international human rights law, guidance for its implementation in all stages of the AI system life cycle, and capacity-building materials to support Member States' efforts to train government officials, policy-makers and other relevant AI actors on EIA methodology;

(a) 엄격한 과학적 연구와 국제인권법을 기반으로 하는 인공지능 기술의 윤리영향평가(EIA) 유네스코 방법론, 이를 인공지능 시스템 수명 주기의 모든 단계에서 실행하기 위한 지침, 그리고 윤리영향평가 방법론에 대하여 정부 공직자, 정책입안자, 기타 관계된 인공지능 행위 주체를 교육하기 위한 회원국의 노력을 지원할 역량강화도구를 개발함.

(b) developing a UNESCO readiness assessment methodology to assist Member States in identifying their status at specific moments of their readiness trajectory

along a continuum of dimensions;

(b) 일련의 영역에서 준비 궤도의 어느 특정 시점에 회원국이 자신의 준비 상태를 판별할 수 있도록 돕는 유네스코 준비 정도 평가방법론을 개발함.

(c) developing a UNESCO methodology to evaluate ex ante and ex post the effectiveness and efficiency of the policies for AI ethics and incentives against defined objectives;

(c) 인공지능 윤리 정책 및 명확한 대상을 위한 인센티브의 효과성 및 효율성을 전후로 심사하기 위해 유네스코 방법론을 개발함.

(d) strengthening the research- and evidence-based analysis of and reporting on policies regarding AI ethics;

(d) 인공지능 윤리에 관한 정책에 대하여 연구・증거 기반의 분석을 강화하고 이에 대한 정책을 보고함.

(e) collecting and disseminating progress, innovations, research reports, scientific publications, data and statistics regarding policies for AI ethics, including through existing initiatives, to support sharing best practices and mutual learning, and to advance the implementation of this Recommendation.

(e) 모범 관행의 공유 및 상호 학습을 지원하고 본 권고의 이행을 추진하기 위해, 기존의 계획 등을 통하여 AI 윤리 정책에 관한 진행 상황, 혁신, 연구 보고서, 과학 간행물, 데이터 및 통계를 수집・배포함.

132. Processes for monitoring and evaluation should ensure broad participation of all stakeholders, including, but not limited to, vulnerable people or people in vulnerable situations.

132. 모니터링・심사 과정은 취약 계층 및 취약한 상황에 처한 사람을 비롯하여, (단, 이에 국한되지 않는) 모든 이해관계자들의 광범위한 참여를 보장해야 한다.

Social, cultural and gender diversity should be ensured, with a view to improving learning processes and strengthening the connections between findings, decision-making, transparency and accountability for results.

학습 과정을 개선하고 연구 결과에 대한 시사점・의사결정・투명성・책무성 간의 연계성을 강화하기 위하여 사회적・문화적・성별 다양성이 보장되어야 한다.

133. In the interests of promoting best policies and practices related to ethics of AI, appropriate tools and indicators should be developed for assessing the effectiveness and efficiency thereof against agreed standards, priorities and targets, including specific targets for persons belonging to disadvantaged, marginalized populations, and vulnerable people or people in vulnerable situations, as well as the impact of AI systems at individual and societal levels.

133. 인공지능 윤리와 관련된 모범 정책・관행을 장려하기 위해서는, 합의된 기준, 우선순위 및 대상자, 특히 빈민・소외・취약 계층에 속한 사람 또는 취약한 상황에 처한 사람들과 같은 특정 대상자에 대하여, 인공지능의 효율성 및 효과성과 더불어 개인・사회 차원에서 인공지능 시스템의 영향을 평가하기 위한 적절한 도구 및 지표가 개발되어야 한다.

The monitoring and assessment of the impact of AI systems and related AI ethics policies and practices should be carried out continuously in a systematic way proportionate to the relevant risks.

인공지능 시스템의 영향 모니터링・평가, 그리고 해당 인공지능 정책・관행은 관련된 위험성에 비례하여 체계적인 방식으로 지속적으로 수행되어야 한다.

This should be based on internationally agreed frameworks and involve evaluations of private and public institutions, providers and programmes, including self-evaluations, as well as tracer studies and the development of sets of indicators.

이는 국제적으로 합의된 프레임워크에 기반을 두어야 하며, 자체 평가를 비롯한 공공・민간 기관, 공급자, 프로그램에 의한 평가뿐만 아니라, 추적 연구 및 지표 개발을 수반할 수 있다.

Data collection and processing should be conducted in accordance with international law, national legislation on data protection and data privacy, and the values and principles outlined in this Recommendation.

데이터 수집・처리는 데이터 보호 및 데이터 프라이버시에 관한 국제・국내법, 그리고 본 권고에서 간략히 제시한 가치 및 원칙에 따라 이행되어야 한다.

134. In particular, Member States may wish to consider possible mechanisms for monitoring and evaluation, such as an ethics commission, AI ethics observatory, repository covering human rights-compliant and ethical development of AI systems, or contributions to existing initiatives by addressing adherence to ethical principles across UNESCO's areas of competence, an experience-sharing mechanism, AI regulatory sandboxes, and an assessment guide for all AI actors to evaluate their adherence to policy recommendations mentioned in this document.

134. 윤리 위원회, 인공지능 윤리 관측기구, 인공지능 시스템의 인권합치적이며 윤리적인 개발을 다루는 저장소, 즉 유네스코의 모든 권한 내 분야에서 윤리 원칙을 준수하게 함으로써 기존 계획에 이바지하는 것, 경험 공유 메커니즘, 인공지능 규제 실험장, 모든 인공지능 행위 주체가 본 권고에서 언급된 정책 권장사항 준수 정도를 스스로 심사할 수 있는 평가 지침을 포함한다.

## VI. UTILIZATION AND EXPLOITATION OF THE PRESENT RECOMMENDATION

## VI. 현 권고의 활용 및 이용

135. Member States and all other stakeholders as identified in this Recommendation should respect, promote and protect the ethical values, principles and standards regarding AI that are identified in this Recommendation, and should take all feasible steps to give effect to its policy recommendations.

135. 본 권고에서 확인된 회원국과 다른 모든 이해관계자는 본 권고에서 확인된 인공지능에 관한 윤리적 가치・원칙・기준을 존중, 증진, 보호해야 하며, 정책 권장사항을 적용하기 위한 가능한 모든 조치를 취해야 한다.

136. Member States should strive to extend and complement their own action in respect of this Recommendation, by cooperating with all relevant national and international governmental and non-governmental organizations, as well as transnational corporations and scientific organizations, whose activities fall within the scope and objectives of this Recommendation.

136. 회원국은 본 권고의 범위 및 목표 내에서 활동하는 모든 다국적 기업 및 과학 조직, 유관 국내・국제 정부 및 비정부기구와 협력함으로써, 본 권고에 관해 자국의 행동을 확장 및 보완하기 위해 노력해야 한다.

The development of a UNESCO Ethical Impact Assessment methodology and the establishment of national commissions for the ethics of AI can be important instruments for this.

유네스코 윤리영향평가 방법론의 개발 및 인공지능을 위한 국가위원회의 설립이 이를 위한 중요한 도구가 될 수 있다.

## VII. PROMOTION OF THE PRESENT RECOMMENDATION

## VII. 현 권고의 홍보

137. UNESCO has the vocation to be the principal United Nations agency to promote and disseminate this Recommendation, and accordingly will work in collaboration with other relevant United Nations entities, while respecting their mandate and avoiding duplication of work.

137. 유네스코는 본 권고의 홍보 및 확산에 앞장서는 유엔기구라는 사명이 있으며, 그에 따라 기타 유관 유엔기구와 협업함과 동시에, 그들의 권한을 존중하고 중복된 일을 피할 것이다.

138. UNESCO, including its bodies, such as the World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), the International Bioethics Committee (IBC) and the Intergovernmental Bioethics Committee (IGBC), will also work in collaboration with other international, regional and sub-regional governmental and non-governmental organizations.

138. 세계과학기술윤리위원회(COMEST), 국제생명윤리위원회(IBC), 정부간생명윤리위원회(IGBC)와 같은 유네스코 산하위원회를 포함하여 유네스코는 또한 다른 국제, 지역, 하위 지역 정부・비정부 기구와도 협업할 것이다.

139. Even though, within UNESCO, the mandate to promote and protect falls within the authority of governments and intergovernmental bodies, civil society will be an important actor to advocate for the public sector's interests and therefore UNESCO needs to ensure and promote its legitimacy.

139. 비록 유네스코 내에서 증진・보호 권한이 정부・정부간 기구의 권한에 속하더라도, 시민사회가 공공 부문의 이익을 대변하는 중요한 행위 주체일 것이므로 따라서 유네스코는 그 정당성을 보장・증진할 필요가 있다.

## VIII. FINAL PROVISIONS

## VIII. 최종 규정

140. This Recommendation needs to be understood as a whole, and the foundational values and principles are to be understood as complementary and interrelated.

140. 본 권고는 그 전체로서 이해되어야 하며, 근본 가치 및 원칙은 상호보완적이며 서로 밀접한 관계가 있는 것으로 이해되어야 한다.