유네스코 인공지능(AI) 윤리 권고 해설서

인공지능 윤리 이해하기

유네스코한국위원회 기획 | 이상욱 지음

국제연합
교육과학문화기구

유네스코
한국위원회

유네스코 인공지능(AI) 윤리 권고 해설서

인공지능 윤리 이해하기

유네스코한국위원회 기획 | 이상욱 지음

국제연합
교육과학문화기구

유네스코
한국위원회

일러두기

- 동 출판물에 기재된 구체적인 내용과 방향은 유네스코한국위원회의 입장과 반드시 일치하지 않을 수도 있습니다.

목 차

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 수호하기위해 반드시 필요한 것이며, 또한 모든 국민이 상호관심과 협력의 정신으로써 완수해야 할 신성한 의무이다.

오로지 정부 간 정치적·경제적 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.

1945년 11월 16일 채택

발간사

인공지능의 빠른 발전은 인류가 직면한 다양한 문제를 극복하는데 큰 도움을 줄 것으로 기대되고 있습니다. 우리가 의식하지 못하는 사이에 우리는 이미 많은 혜택을 입고 있습니다. 그와 동시에 윤리, 인권, 안보 등의 측면에서 예상치 못한 편견과 불평등이 생길 수도 있습니다. 모두가 누려야 할 AI의 혜택이 자칫 특정 국가나 계층에게만 돌아가지 않도록 하고 AI 개발과 활용이 인권과 인간 존엄성을 존중할 수 있도록 하기 위해서는 모두의 공감대가 필요합니다.

유네스코는 유엔 체계 안에서 유일하게 윤리적, 지적 성찰을 도모하는 기구로, 국제적인 인공지능 윤리 지침을 제정하기 위해 지난 2년 여간 작업해 왔습니다. 그 결과로 2021년 11월에 열린 41차 유네스코 총회에서 'AI 윤리 권고'를 채택했습니다. 인공지능과 관련하여 아직 국제규범이 마련되어 있지 않은 상황 속에서 유네스코의 이러한 노력은 매우 시의적절하며 중요한 의미를 지닌다고 할 수 있습니다.

앞으로 시간이 지날수록 인공지능의 파급력은 사회 전 분야로 확산될 것입니다. 때문에 유네스코 인공지능 윤리 권고가 어떤 배경에서 준비되었으며, 어떠한 내용을 담고 있는지, 그리고 그 함의는 무엇인지를 파악하는 것은 매우 중요합니다. 이 책은 유네스코 세계과학기술윤리위원회(COMEST) 위원으로 유네스코 인공지능 윤리 권고 초안 작업에 직접 참여한 이상욱 한양대 교수가 전 세계 24명의 전문가들이 만든 권고 초안의 배경과 함의를 분석한 글을 담고 있습니다.

이 책이 인공지능 윤리를 이해하는데 도움이 되기를 바라며, 앞으로 본격적으로 시작될 인공지능과 인공지능 윤리에 대한 다양한 논의에 기여할 수 있기를 기대합니다. 앞으로 인공지능과 더불어 살아나갈 인류의 새로운 미래를 유네스코한국위원회도 함께 모색하겠습니다.

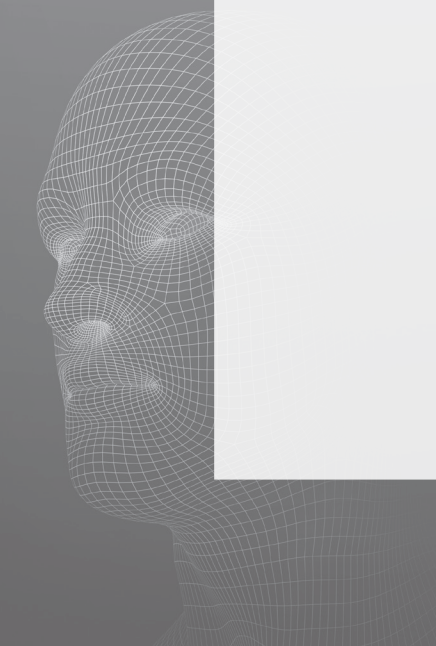
감사합니다.

유네스코한국위원회 사무총장 한 경 구

제1부

머리말

인공지능(AI) 윤리란 무엇인가?
낯선 AI의 도전
유네스코와 AI



01 머리말

인공지능(AI) 윤리란 무엇인가?

바야흐로 인공지능의 시대가 열렸다. SF 영화에나 등장하는, 사람과 구되지 않는 안드로이드 로봇은 아직 먼 미래의 꿈이지만, 그보다는 훨씬 일상적이고 널리 퍼져 있는 인공지능이 우리에게 친숙한 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 인공지능의 도움을 받고 있다. 인공지능이 추천해 주는 선택지는 아직까지는 가끔 성가실 정도로 엉뚱한 것일 수도 있지만 상당히 많은 경우 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' 인공지능이 현실화될지도 모른다고 기대할 수도 있다.[1)]

1) 인공지능 기술 개발의 현황과 미래 발전 전망에 대해 균형 잡힌 분석은 Shanahan, M., The Technological Singularity, Cambridge, MA: The MIT Press, 2015, Kaplan, Jerry,a, Artificial Intelligence:What Everyone Needs to Know, Oxford:Oxford University Press, 2016 참조.

인공지능의 일상화만큼이나 최근 국내외에서는 인공지능이 제기하는 여러 인문학적, 사회과학적 쟁점을 학술적으로, 실천적으로 탐색하는 연구 및 관련 활동도 활발하다. 전통적으로 인간만이 할 수 있었던 법률, 의료, 세무 등의 일자리 영역에서도 인공지능 활용이 늘어나면서 대량 실업 사태가 일어날 수 있다는 두려움과 이를 정반대로 해석해서 인간이 노동으로부터 해방된 자유를 얻게 되리라는 유토피아적 기대가 함께 제시되고 있고, 이와 관련된 '기본소득' 논의는 어느덧 상식적 담론의 하나가 되었다.[2)]

인공지능에 대한 이런 다양한 쟁점을 포괄적으로 다루는 분야를 최근 국제 논의에서는 점점 윤리(ethics)라는 개념으로 포괄하고 있다. 이는 국내에서 '윤리(倫理)' 개념이 사용되는 방식과 상당한 거리가 있어서 국내 정책 입안자들이나 논자들이 국제적으로 이루어지는 인공지능 '윤리' 논의를 잘못 파악하는 경우가 많다. 즉 국내에서 윤리는 지극히 개인적인 사안에만 한정되는 것으로 여겨지며 상식적인 수준에서 옳고 그름이 분명히 판단될 수 있는 사안, 예를 들어 '살인은 나쁘고, 남을 돕는 것은 좋은 일이다.' 같은 것과 관련된 것으로 여겨진다. 그래서 윤리적 삶이란 나쁜 일은 하지 않고 좋은 일을 하려고 노력하는 것으로 이해되는 경향이 있어서 기본적으로 과학과는 '무관'하다는 인식이 널리 퍼져 있다. 황우석 연구팀의 논문

2) 인공지능 기술의 보편적 보급이 전문직조차 위협할 것이라는 생각에 대한 균형잡힌 분석은 Susskind, R. and Susskind, D., The Future of Professions: How Technology Will Transform the Work of Human Experts, Oxford:Oxford University Press, 2017 참조.

조작 사건처럼, 설사 윤리적 '쟁점'이 사회적으로 부각되더라도 그런 '쟁점'은 제도적, 문화적으로 풀어가야 할 복합적 문제로 보는 것이 아니라 특정 연구자의 개인적 일탈로 이해되고 '윤리적' 선/악 구별의 이분법에 기초하여 해결될 수 있으리라 기대된다. 예를 들어 2000년대 초반 줄기세포 연구를 둘러싸고 과학윤리적 논쟁이 벌어졌을 때도 줄기세포 연구를 하는 것 자체가 선한 것인지 혹은 악한 것인지에 대한 의견 차이에 사회적 논의가 주로 집중되었고, 줄기세포 연구를 어떤 '방식'으로 어떻게 제도화하여 사회적으로 수용할 수 있을 것인지의 보다 복잡한 윤리적 논의는 그다지 주목을 받지 못했다.[3)]

하지만 인공지능에 대한 국제적 논의에서 윤리(ethics)는 훨씬 더 광범위한 주제를 포괄적으로 지칭한다. 그리고 이어질 유네스코 AI 윤리 권고(안)에 대한 설명에서도 확인할 수 있듯이 선/악 구도는 이런 의미의 포괄적 윤리 논의를 하기에 적합한 틀이 아니다. 그 주된 이유는 AI 윤리처럼 특정 주제로 한정해도 우리가 추구해야 할 가치는 하나가 아니라 일반적으로 복수이기에 특정 가치를 기준으로 선악 판단을 할 수 없기 때문이다. 실제 세계는 매우 복잡하다. 또한 우리가 사는 세계는 다양한 가치를 추구하는 개인과 집단이 서로 어울려 사는 다가치 사회이다. 이런 다가치 사회에서 존중되는 여러

3) 국내 저자들이 과학과 관련된 윤리적 쟁점을 우리말 '윤리'의 협소한 의미가 아니라 영어 ethics의 보다 포괄적인 의미에서 탐색한 내용은 이상욱, 조은희 엮음 2011, 『과학 윤리 특강 – 과학자를 위한 윤리 가이드』, 서울: 사이언스북스 참조.

가치 사이에는 충돌이 일어날 여지가 많고 대부분의 사회적 결정은 이런 가치들을 모두 만족하는 (현실적으로 불가능한) 방식이 아니라 윤리적으로 합리적이라고 평가될 수 있는 방식으로 각각의 가치를 적절한 수준에서 절충하여 만족하는 방식으로 이루어진다. 당연히 AI 윤리의 여러 핵심 주제에 대해서도 마찬가지 방식으로 주요 사회적 결정이 내려질 수밖에 없다.

이렇게 이해된 AI 윤리의 관점에서 보자면 인공지능과 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 균형있게 존중하는 방식으로 AI 개발과 활용을 하기 위해서는 어떤 점에 주의하고 어떤 제도적 장치를 마련해야 하는지를 통합적으로 탐색하려는 노력이 필요하다. 2021년 11월 유네스코 총회에서 논의된 유네스코 윤리 권고(안)은 정확히 이런 의미의 AI 윤리 개념을 염두에 두고 작성되었으며 이렇게 AI 윤리를 이해하는 것이 현재 국제적으로 널리 통용되는 방식이다.

'낯선' AI의 도전

인공지능 기술은 최근에 갑자기 등장한 기술이 아니다. 인공지능을 개념적으로 어떻게 정의하는지에 따라 정확히 언제부터 인공지

능 기술개발이 이루어졌는지에 대해 논란이 있을 수 있지만 본격적으로 인공지능이라는 용어를 사용하고 인간지능을 흉내낼 수 있는 기계지능을 만들기 시작한 것이 1950년대이므로 적어도 반세기 이상의 역사가 있다. 하지만 21세기 시작될 무렵에 등장한 인공지능은 인공지능의 윤리적 쟁점을 논할 때 결정적으로 중요한 특징을 갖는다. 그것은 바로 이 인공지능이 우리에게 '낯설다'는 점이다. 이후에 이어질 유네스코의 AI 윤리 관련 논의를 위해 이 점을 우선 지적하고 넘어가겠다.[4)]

다음과 같은 상상을 해보자. 길을 걷다가 아주 정교하게 보이는 기계 장치를 발견했다고 하자. 여러 톱니바퀴가 복잡하게 엇물려 돌아가고 사이사이 회전나사도 들어 있다. 뚜껑을 열어 보니 동그란 쇠 표면에 숫자들이 돌아가며 쓰여 있고 그 사이를 바늘이 움직인다. 이때 여러분은 이 장치가 갑자기 하늘에서 뚝 떨어졌다고 생각하지는 않을 것이다. 아마도 누군가 '의도'와 '계획'을 갖고 만들었다고짐작할 것이다.

이 사고실험은 페일리라는 19세기 영국의 자연신학자가 당시 일반인들에게 익숙한 회중시계를 사례로 창조주 신의 존재를 논증하기 위해 고안한 것이다. 탄복할만큼 복잡한 구조의 시계에 대해 우

4) 이 소절의 다음 내용은 필자가 AI 타임즈 2020년 9월 28일자에 기고한 내용에 기초하여 작성되었다. http://www.aitimes.com/news/articleView.html?idxno=132490

리가 자연스럽게 그것을 만든 시계제작자를 상정하듯, 자연의 복잡 다양한 생명체에 대해서도 의도와 계획을 갖고 그것을 창조한 신을 상정해야 한다는 생각이다.

AI 윤리에 대한 글에서 왜 엉뚱하게 시계와 시계제작자 이야기를 하는지 의아할 수 있다. 단초는 페일리의 사고실험에도 불구하고, 현재 과학계는 복잡한 생명체도 의도와 계획을 가진 초자연적 신을 구태여 상정하지 않고도, 자연적 요인이 오랜 시간 걸쳐 작용함으로써 등장할 수 있다는 다윈의 진화론을 수용하고 있다는 사실이다. 다른 말로 하자면 우리는 기가 막히게 잘 만들어진 대상에 대해서도 그것이 배후에 의도와 계획을 가진 의식적 존재가 있다고 반드시 가정할 필요는 없다는 것이다. 이 중요한 깨달음의 생생한 사례가 21세기의 인공지능이다. 즉 AI 윤리의 대상인 인공지능은 스스로 의도와 계획을 갖지 못하면서도, 우리에게는 그 행동 혹은 결과물에서는 반드시 특정 의도나 계획을 갖고 있는 것처럼 보인다는 것이다.

사진출저: https://news.microsoft.com/europe/features/next-rembrandt/

위 그림을 보자. 첫인상부터 범상치 않은 초상화다. 그림을 조금 볼 줄 아는 분이면 렘브란트의 화풍과 비슷하다고 느낄 것이다. 실제로 이 그림은 렘브란트의 작품을 기계학습한 인공지능이 렘브란트가 그렸을법한 초상화를 계산하여 3D 프린처로 출력해낸 작품이다. 하지만 이 인공지능은 렘브란트 그림을 모작해야겠다는 '의도'나 그러기 위해서는 일단 렘브란트 특유의 스타일을 학습하고, 그 스타일로 작품의 구도를 잡은 다음, 3D 프린터로(자기는 손이 없으니 붓으로 그릴 수는 없다!) 찍어내야겠다는 '계획'을 세울 수 없다. 그 의도나 계획은 모두 이 인공지능을 만든 '넥스트 렘브란트' 프로젝트에 참여한 인간 엔지니어들의 몫이었다.

최근 인공지능의 눈부신 발전에만 주로 세간의 이목이 집중되면서, 인공지능이 분명 기술적으로 탁월하지만 인간지능과 다른 방식으로 작동한다는 점이 제대로 부각되지 않고 있다. SF 영화에 등장하는 인공지능, 예를 들어 영화 〈아이언맨〉에 등장하는 자비스처럼, 인간과 구별되지 않는 정신세계를 가진 것처럼 보이는 가상의 인공지능을 떠올리는 것은 AI 윤리 논의에 거의 도움이 되지 않는다. 대신 우리가 집중해야 할 인공지능, 현재 그리고 가까운 미래에 활용될 인공지능은 압도적인 계산능력을 활용하여 주어진 데이터의 패턴을 파악하는 기계이다. 만약 이렇게 파악된 패턴이 미래에도 계속된다고 가정하면, 인공지능은 미래를 예측하는 데 활용될 수 있다. 경제추이를 예측하는 금융인공지능과 환자의 질병 유무를 의학영상 자료에서 판단하는 의료인공지능이 대표적이다.[5)]

하지만 동일한 작업을 수행하는 인간과 달리 인공지능은 자신의 계산을 증권시장 예측이나 MRI 판독 과정에서 '느낄' 수 없다. 일에 집중하느라 정신이 팔려 의식하지 못하는게 아니라(인간도 가끔씩은 그런 때가 있다), 의식하는 것 자체가 불가능하다. 그런 회로 자체가 아직은 인공지능에 없기 때문이다. 그러므로 이런 의미에서 아

5) 이 점을 인공지능이 경영 혹은 혁신의 맥락에서 사용되는 구체적 사례를 통해 잘 지적하고 있는 책으로는 Joshua Gans, Avi Goldfarb and Ajay Agrawal 2018, Prediction Machines: The Simple Economics of Artificial Intelligence, Cambridge, MA: Harvard Business School Press 참조.

직까지 인공지능은 매우 뛰어난 성능을 지닌, 하지만 본질적으로는 계산기라고 할 수 있다. 다르게 표현해 보자면, 인공지능의 도움을 받아 만들어 낸 결과물, 예를 들어 앞서 소개한 유사-렘브란트 초상화는 인간이 보기에는 인간과 유사한 마음을 가진 지적 존재에 의해 의도적으로 제작된 작품으로 보이지만, 실은 그 작품에 기여한 인공지능은 자신이 예술 작품을 창작한다는 사실은커녕 '그림'을 그리는 과정에 참여하고 있다는 사실조차 인식하지 못한 채 작업을 수행하고 있을 뿐이다. 이처럼 인공지능은 우리에게 '매우 낯선' 방식으로 지능적 결과물을 산출한다.

이 점을 정확하게 인지하는 것이 21세기 인공지능과 함께 살아가는 데 결정적으로 중요하다. 그 이유는 우리는 이미 인공지능으로 둘러싸여 살고 있기 때문이다. 인공지능 챗봇과 '친구처럼' 다정하게 이야기를 나누는 것은 좋지만, 그 챗봇이 위기상황에서 당신을 진짜 친구처럼 돌볼 수 있다고 생각하지 않아야 한다. 챗봇이 거짓으로 '친구인체' 당신을 속인 것이 아니다. 인공지능은 인간을 속일 수 있는 능력이 없다. 단지 여러분이 자신과 내밀한 대화까지 나누는 인공지능이 자신을 돌볼 수 있는 '마음'과 능력을 가졌다고 믿었기에 스스로 자기속임을 한 것이다.

해외에서는 평소에는 직원들의 어려운 사정과 가족 이야기에 그토록 관심을 보이다가 업무 평가에서는 '냉혈한으로 돌변하여' 무자비한 해고 통지를 날린 인사담당 인공지능에게 충격을 받은 사례도

보고되고 있다. 이 경우 역시 인공지능이 '이중적 태도'를 보인 것이 아니다. 직원과 '공감하는' 대화를 나누는 것은 업무 효율성을 높이는 데 도움이 되고, 일 잘 못하는 직원을 해고하는 것 역시 (그 직원의 딱한 개인 사정과 무관하게) 업무 효율성을 높이는 데 도움이 된다. 그 인사담당 인공지능은 단지 효율성을 극대화하도록 프로그램화 되었을 뿐이다. 이 점을 파악하지 못하고 인공지능을 '인간처럼' 생각했던 사람들이 뜻밖의 고통을 당하는 것이다.[6]

'낯선 인공지능'의 본성을 적절히 이해하지 못한 채 인공지능을 활용할 때 발생하는 문제는 개인적 고통에서 끝나지 않는다. 어려운 상황에도 '믿음직하게' 작동하는 인공지능에 대해 사람들은 자신보다 뛰어난 상급자를 대하듯, 다른 의견이 있어도 함부로 제시하지 못하거나 중요한 판단을 위임하기 쉽다. 이런 경향성은 국가적 규모의 에너지 체계 관리나 군사 작전에서 파국적 결과를 가져올 수 있다.[7] 물론 이런 위험성이 있으니 인공지능 사용을 중단하자고 결정하는 것은 지나친 대응이다. 그보다는 인간이 보기에는 정말 이상한 방식으로 '똑똑한' 인공지능과 서로의 단점을 보완하면서 생산적으

---

6) 이와 관련된 보다 풍부한 사례는 Fry, Hannah 2019, Hello World: How to Be Human in the Age of the Machine, London: Transworld Publishers Ltd., Mitchell, Melanie 2020, Artificial Intelligence: A Guide for Thinking Human, New York: Picador 참조.

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

로 협력하는 방식을 배워가야 한다. 네안데르탈인이 4만년 전쯤 멸종한 이후 인간만큼 똑똑한 존재를 만나본 적이 없는 인간에게 이 일은 결코 쉽지 않을 것이다. 하지만 이 어려운 일을 성공적으로 수행하는 개인과 사회는 21세기 인공지능 시대에 우위를 점하게 될 것이다.

핵심은 이것이다. 앞서 지적했듯이 인공지능과 관련된 '윤리적 논의'는 단순한 선악 판단이 아니라, 인공지능과 함께 살아갈 미래 사회의 바람직한 모습과 직접적으로 관련되어 있다. 특히 유네스코 AI 윤리 권고(안)처럼 이런 미래사회에 대한 제언을 담고 있는 문건은 실천적 함의가 매우 두드러진다. 이렇게 AI 윤리가 '낯선' 지능과 함께 살아갈 우리의 미래 삶과 직결된다는 점을 염두에 두면 우리가 인공지능과 과년되어 수행할 중요한 사회적 결정과 그에 대한 권고는 반드시 현재 개발이 진행 중인 인공지능에 대한 정확한 이해에 근거하여 이루어져야 한다. SF 영화에나 등장하는, 아직 실현되지 않은 그리고 당분간 실현될 가능성이 매누 낮은 환상적인 인공지능이 아니라 가까운 미래에 등장할 '낯선' 인공지능에 대한 정확한 이해에 기초해야 한다는 것이다.[8)]

8) 실제로 이 점은 유네스코 AI 윤리 권고(안) 작성 작업에 참여했던 다수의 전문가들에 의해 여러 차례 다양한 방식으로 언급되고 강조되었다.

유네스코와 AI

유네스코는 유엔의 여러 산하 기구 중에서 과학, 교육, 문화에 집중하는 국제기구이다. 최근 유네스코는 과학, 교육, 문화의 '오래된' 주제에 더해 기후변화에 대응하는 '지속가능한 발전(sustainable development)' 및 현대사회에서 정보기술이 갖는 중요성에 주목하고 있다. 또한 젠더 문제, 아프리카 문제 등 '새로운' 주제에 대한 특별한 관심을 보이며 활발한 관련 활동을 벌이고 있다. 특히 유네스코는 과학 연구가 사회에 미치는 다양한 방식의 영향에 대해 탐색하고 이에 대한 윤리적, 제도적, 사회적, 문화적 실천 방안을 모색하는 노력을 지속적으로 수행하고 있다. 이 과정에서 과학기술윤리위원회(COMEST)와 국제생명윤리위원회(IBC)가 주도적인 역할을 수행하고 있다.

인공지능(Artificial Intelligence)에 대해 다양한 견해가 있고, 특히 인공지능이 제기하는 '위험'의 성격과 심각성에 대해서는 상당한 의견차이가 있지만, 그럼에도 모든 논자들의 견해가 일치하는 부분은 미래 사회에서 인공지능의 영향력이 현재 사회에서 전기나 인터넷에 비견될 정도로 광범위하고 클 것이라는 점이다. 이는 인공지능에 대한 사회적 관심이 관련 기술 개발을 얼마나 효율적으로 할 것인지에만 국한되어서는 안 되고 인공지능이 사회와 맺는 여러 접점에 대한 보다 포괄적 논의(윤리, 법, 정책, 문화 등)까지 함께 진행해야 하는 당위를 제공한다. 2019년 발간된 OECD 보고서의 제목을 빌어 말하자면, 인공지능 혁신은 '사회 속의 혁신(Innovation in

Society)'이어야 한다는 말이다.[9)]

이 점은 유네스코의 AI 윤리 논의만이 아니라 EU OECD, IEEE 등 AI 윤리 논의를 수행하고 있다는 다양한 국제단체들이 일관되게 취하고 있는 입장이다. 간단하게 말하자면 이들 국제단체들은 모두 인공지능 기술이 인류에게 가져다 줄 수 있는 잠재적 혜택에 주목하고 인공지능 기술의 이런 잠재력을 적극적으로 활용해야 한다는 데 동의하면서 동시에 그러한 활용이 결코 우리 사회(국가적 수준과 국제적 수준 모두를 포함하여)가 소중하게 생각하는 핵심 가치(기본권 등)를 손상하지 않는 방식으로 이루어져야 함을 일관되게 강조하고 있는 것이다. 그러므로 '사회 속의 혁신'을 AI 윤리의 맥락에서 해석하자면, 인공지능 기술의 혁신은 사회적 가치의 테두리 내에서 이루어져야 함을 강조하는 것이라고 볼 수 있다.

물론 이 '사회적 가치의 테두리'를 어떻게 해석할 것인지, 사회적 가치는 고정된 것인지 변화될 수 있는 것인지, 사회적 가치는 서로 다른 문화에서 다르게 해석될 여지가 있는 것인지 아니면 전지구적으로 보편성을 가져야만 하는 것인지를 두고 논쟁의 여지는 있다. 실제로 AI 윤리를 다루는 여러 국제 논의들이 얼핏 보면 대충 비슷비슷하다고 생각되지만 자세한 내용을 들여다보면 상당한 차이를 발견할 수 있는 이유도 이 '사회적 가치의 테두리'를 정확하게 어떻

9) https://ec.europa.eu/jrc/communities/sites/jrccties/files/eedfee77-en.pdf

게 해석할 것인지에 대한 의견 차이를 반영했기 때문이다. 이 점에 대해서는 이후 절에서 보다 자세한 논의가 제시될 것이다.

이상의 내용을 염두에 두고 이후의 절에서는 2020년 5월에 초안이 발표되고 2020년 9월에 최종안이 발표된 유네스코의 인공지능 윤리 권고(안)에 대해 살펴보자.

제2부

배경과 과정

02 배경과 과정

유네스코는 과학기술과 사회가 맺는 다양한 접점을 탐색하는 두 상설위원회를 갖고 있다. 국제생명윤리위원회(IBC)와 과학기술윤리위원회(COMEST)가 그것이다. 이 두 위원회를 통해 유네스코는 현대 과학기술의 여러 윤리적, 사회적 쟁점에 대한 보고서를 발간해왔고, 사안의 중요성이나 심각성이 회원국 전체의 행동을 촉구할 내용이라고 판단하는 경우에는 '규범적 틀/도구(normative framework/instrument)'를 마련하여 총회 의결을 통해 공표해왔다.[10] 이런 취지로 가장 최근에 공표된 것이, COMEST 주도로 이루어진 "기후변화윤리원칙선언"(2018)이다. 유네스코의 규범적 틀/도구는 내용의 구체성과 강제력의 정도에 따라 선언(Declaration), 권고(Recommendation), 협약(Convention)으로 나뉘는데, 이번 인공지능 윤리 관련 윤리적 틀/도구는 유네스코 총회의 결정에 따라 권고로 추진 중이다.

유네스코가 인공지능 윤리에 관심을 갖게 된 이유는 여러 가지가 있겠지만 인공지능 관련 쟁점이 유네스코가 중점 사업 분야나 핵

10) https://en.unesco.org/themes/ethics-science-and-technology/comest

심 주제 여럿과 깊이 연결되어 있다는 점이 관련 내부 논의 과정에서 지속적으로 강조되었다. 유네스코가 인공지능 윤리 권고(안) 작성 작업을 시작하기 위해서는 절차상 이에 대해 총회 의결이 필요했는데, 총회에 이 안건을 상정하기 위해서는 집행위원회 통과가 필요했다. 이때 집행위원회와 총회의 결정을 돕기 위해, COMEST 위원을 중심으로 인공지능 윤리 관련 전문가가 보강된 '확대전문가집단(Extended Experts Group)'을 꾸려져서 인공지능 윤리 관련 쟁점을 정리한 예비보고서를 만들었다.[11] 그런데 이 예비보고서 작성 과정에서 참여 전문가 위원을 제외하고 가장 많은 의견 개진과 피드백을 제공한 사람들은 유네스코의 각 분야 담당자들이었다. 이처럼 유네스코 인공지능 윤리 논의는 유네스코의 기존 활동 및 각 분야의 관심사를 반영하는 방식으로 이루어졌고, 이는 유네스코의 인공지능 윤리 논의가 국제적으로 진행되는 다른 유사한 논의와 분명한 차별성을 갖는 측면이라고 할 수 있다.

'확대전문가집단'이 인공지능 윤리에 대한 예비보고서를 2019년 봄 유네스코 잡행위원회에 제출하고, 가을 총회에서 인공지능 윤리 권고 초안 작성을 시작하는 것에 대한 의결이 이루어지자, 이제는 회원국의 추천을 받아 이 작업을 수행할 비상설전문가집단(Ad Hoc Expert Group)을 구성하게 되었다. 총 24명으로 구성된 전문가 집단은 유엔의 6개 영역에서 각 4명씩 선발되었으며, 이러한 구

11) https://ircai.org/wp-content/uploads/2020/07/PRELIMINARY-STUDY-ON-THE-ETHICS-OF-ARTIFICIAL-INTELLIGENCE.pdf

조적 특징은 각 전문가들이 자신의 국가를 '대표'하지는 않지만, 각 지역의 다양한 관심사와 다른 의견을 권고 초안에 반영하기 위한 것이라고 볼 수 있다. 절차상 비상설전문가집단이 2020년 9월에 권고(안)을 완성하고, 이에 대해 본격적으로 각 국가 대표들이 2021년 여름에 정부간협의를 통해 총회에 상정할 권고안이 결정되고, 이를 2021년 가을에 열릴 유네스코 총회에서 심의하여 채택 여부를 결정하게 된다.

원래 계획은 2020년 4월에 파리에서 비상설전문가집단이 모여 권고 초안 작성 작업을 하기로 되어 있었지만, 코로나-19 사태로 인해 온라인 회의로 전환되었고 참여 위원들의 시간대가 지구 전체에 분포되어 있다는 한계에도 불구하고 유네스코 직원들과 위원들의 적극적인 참여로 5월에 권고 초안이 완성되었다. 초안 작성 온라인 회의 과정에서도 옵저버로 참여한 회원국 대표부와 관련 단체는 서면으로 의견을 제시하였고 이 의견은 온라인 회의 중에 논의되어 초안에 반영되었다.

5월에 완성된 권고 초안에 대해 곧바로 온라인 의견수렴 과정과 각 지역별 의견수렴 회의를 통해 다양한 피드백이 수집되었다. 유네스코는 AI 윤리에 대한 유엔 수준의 대표성을 강하게 의식하고 있기에 이번 권고(안)이 되도록 많은 의견을, 되도록 많은 계층의 사람들의 이해 관계를 반영하는 방식으로 작성되기를 원했다. 그래서 5월 초안에 대한 의견 수렴 과정도 국제적 영역에서 누구나 참여할 수

있는 온라인 의견 제시 과정과 보다 공식성을 갖춘 지역별 의견수렴 회의의 두 갈래로 이루어졌다. 그리고 이 의견은 유네스코 사무국에 의해 정리되어 AHEG에게 전달되었으며 8월부터 시작된 최종안 작성 과정에서 적극적으로 활용되었다.

의견 수렴 내용이 활용된 방식은 2020년 9월에 완성된 유네스코 AI 윤리 권고안이 분명 지역별 전문가들의 노력을 통해 작성된 것이기는 하지만 순전히 전문가 위원회만의 의견이라고 보기 어렵다는 점을 잘 보여준다. 즉, 2021년 여름에 있을 정부간 회의 이전에도 각국 정부의 유네스코 대표부는 AHEG의 논의 과정에 옵저버로 참여해서 매 온라인 회의 종료 후 서면으로 논의 중인 문건의 표현이나 내용에 대해 적극적으로 의견을 제시했고 이 의견은 다음 날 회의에서 사낭별로 정리되어 AHEG 위원들에 의해 논의되고 논의 중인 문건에 최대한 반영되었다.

이 작업은 특히 2020년 8월부터 집중적으로 진행된 최종안 작성 과정에서 훨씬 두드러지게 이루어졌는데 그간 수집된 수많은 의견들이 정리되고 그에 더해 유네스코 각 영역별 의견과 유네스코 사무국의 의견까지 더해져서, 결국에는 매우 다양한 표현 및 내용 상의 제안이 고려되고 반영되었다. 특히 유네스코 정보통신 부서는 인공지능 및 빅데이터 활용과 관련하여 자신들이 이미 수행하고 있던 활동을 반영할 것을 요구했고 AHEG는 그 내용이 위원회에서 준비한 원래 문건의 내용과 잘 어울리지 않음에도 불구하고 이 요구사항을

최대한 반영하였다. 내용 상의 긴장에 있어서는 그 정도가 정보통신 부서에 비해 덜했지만, 2020년에 새로 취임한 가브리엘라 라모스 부사무총장(ADG)을 중심으로 한 유네스코 사무처의 요구도 적극적으로 고려되고 그 논의 결과가 최종안에 반영되었다. 특히 젠더 쟁점과 '행동가능한(actionable)' 정책 제시가 필요하다는 점에 대해서는 라모스 부사무총장의 의견에 위원들 사이에서 상당한 공감대가 있었고 5월에 발표한 초안과 9월의 최종안 사이에 이 주제에 대한 많은 변화가 있었다.

위원의 구성은 지역적 다양성에 더해 전문성에서도 다양성을 확보하려 노력한 흔적이 보인다. 인공지능 기술 전문가, 정책 전문가, 법률 전문가, 철학 전문가가 포함되었고, 그 중에는 여러 영역을 가로지르는 전문성을 가진 위원도 많았다. 특히 유엔 차원의 다른 국제 논의에 참여한 경험을 가진 전문가들도 많아서 하루에 두세 시간 이상 하기 어려운 온라인 회의의 한계에도 불구하고 다양한 논점 제시와 합의점 도출이 비교적 효율적으로 진행될 수 있었다. 온라인 토론 과정에서 전문가들 사이에 종종 상당한 의견 차이가 있음을 확인하였지만, 유엔 기구의 전반적 관례를 반영하여 서로 합의할 수 있는 절충점을 찾는 방식으로 권고안의 문구가 합의되었다.

하지만 여러 지역에 흩어진 위원들의 시간대를 모두 고려하여 온라인 회의를 진행해야 했기에 초안 작성 과정에서 3주, 최종안 작성 과정에서 2차례에 걸쳐 5주 이상의 강도 높은 토론이 진행되었음에

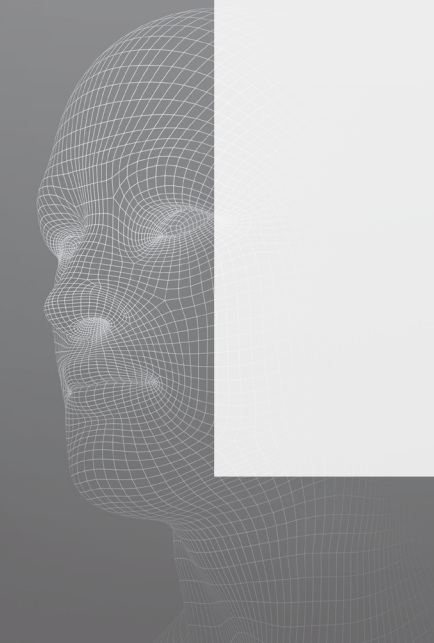
도 불구하고 위원들 사이에서 의견 차이가 컸던 여러 주제에 대해서는 완전한 합의점을 도출할 시간이 충분하지 않았다. 이런 상황에서 5월에 초안을 완성해야 했기에 AHEG는 본질적인 내용에서의 의견 차이에 대해서는 최종안 작성 과정에서 논의하기로 미루어 두는('주차해두자(Let's park this!)'는 표현이 자주 사용되었다) 경우가 많았다.

결국 이렇게 '주차되었던' 주제에 대해서는 8월부터 9월 초까지 이루어진 회의에서 집중적인 재검토가 이루어졌고 앞서 지적한 의견수렴 내용을 반영하는 동시에 재검토를 통해 도달한 합의한을 반영하는 방식으로 최종안 작성이 이루어졌다. 하지만 몇몇 민감한 주제(권고(안)의 규범성 정도와 다른 AI 윤리 논의와의 역할 분담 분제)에 대해서는 완전한 합의를 이룰 수 없었고, 소수 반대 의견을 강하게 주장하는 위원들이 자신의 다른 의견을 회의록에 남긴다는 전제로 최종안에 동의함으로써 논의가 마무리 되었다. 특별히 논쟁적이었던 주제가 무엇이었는지에 대해서는 4절에서 보다 구체적으로 논의하겠다.

제3부

구조와 내용

권고 초안(5월)의 구조와 내용
권고 최종안(9월)의 구조와 내용



03 구조와 내용

권고 초안(5월)의 구조와 내용

권고 초안의 구조는 권고가 지향하는 여러 주제를 설명하고 권고 내용의 핵심을 제시하는 전문(preamble)과 이어지는 4개의 장의로 구성되어 있다. 권고가 회원국들에게 어떻게 이해되고 사용되어야 하는지를 설명하는 첫 장과 권고의 전체적인 목적을 설명하는 둘째 장은 통상적으로 유네스코 윤리적 틀/도구에 공통되는 내용을 담고 있으며 이후 논의를 통해 비교적 제한된 내용의 수정이 이루어졌다.

그에 비해 권고가 핵심적으로 추구하는 '가치와 원칙'을 다룬 3장과 이 가치와 원칙을 회원국들이 정책에 어떻게 반영할 것인지를 다룬 4장 '정책 영역'은 초안의 핵심적인 내용이고 초안 발표 이후 논의를 통해 상당한 정도의 수정이 이루어져 최종안으로 발전된 부분이다.

3장 '가치와 원칙'의 특징은 '가치(values)'와 '원칙(principles)'을 구별하여 제시하였다는 점과 '원칙' 영역을 두 집단으로 나누어

제시했다는 점이다. 논의 과정에서 이 구조적 방식이 적절한 것인지를 놓고 여러 의견이 제시되었고, 특히 가치와 원칙을 깔끔하게 구별할 수 있는지 여부와, 인공지능 관련 윤리 원칙이 두 집단으로 간단하게 나누어지는지 여부를 두고 많은 토론이 이루어졌다. 두 사안에 대해 참여 전문가들이 완전한 의견 일치를 본 것은 아니지만 초안 단계에서는 이런 방식으로 일단 제시해 보자는 의견이 지배적이어서 이런 구조로 나오게 되었다.

우선 전문의 내용은 인공지능이 인류의 삶과 복지에 지대한 영향을 끼친다는 점을 강조하고, 그럼에도 불구하고 다양한 측면에서 우리가 소중하게 여기는 가치, 특히 인권에 위협이 될 수 있다는 점을 지적하며, 이에 대한 적절한 윤리적 대응이 필요하다는 점을 천명하고 있다.

전문에서는 이후에 나올 논의에서 사용될 주요 개념을 정의하거나 설명하고 있는데, 주목할 점은 인공지능이 개인의 삶과 사회에 끼치는 영향을 고려하기 위해서는 인공지능에 데이터가 입력되고 그 계산 결과가 출력된 후 이것이 다시 다양한 방식으로 사회에서 활용되는 전체 과정에 대해 윤리적 논의가 이루어져야 한다는 점을 강조하기 위해 통상적으로 사용하는 인공지능이라는 용어 대신 인공지능 시스템(AI System)이라는 용어를 채택했다는 점이다. 이 용어는 결국 초안만이 아니라 최종안에서도 일관성있게 채택되었는데, 이 용어의 사용은 공학자들에 의해 좁게 이해된 인공지능 기술

자체만이 아니라 인공지능이 설계되고 훈련되고 활용되고 궁극적으로는 폐기되는 인공지능 시스템의 전주기(entire life-cycle)와 관련되는 윤리적 쟁점을 다루고 있다는 점을 부각하기 위해서라고 볼 수 있다.

전문은 인공지능 윤리에 대해 기존 국가 차원이나 국제적 차원에서 제안된 여러 권고나 선언 등을 인용하면서 관련 논의들과 이후에 나올 권고의 내용을 연결 짓고 있다. 이 부분에 대한 비상설전문가집단 내의 논의에서 유엔 보편인권선언처럼 국제적으로 널리 수용되는 원칙을 포함시키는 것에 대한 이견은 없었지만, 특정 단체나 일부 국가만이 참여한 인공지능 윤리 제안에 대해서는 (반드시 그 내용에 문제가 있다고 판단해서라기보다는) 충분한 보편적 수용이 이루어지지 않아서 논쟁적일 수 있다는 이유로 포함하지 않기로 결정한 경우가 많았다. 그래서 관련 문서에 대한 언급은 유네스코 인공지능 윤리원칙 예비보고서를 비롯한 유엔 산하 기구의 관련 문서를 중심으로 국제적으로 공신력을 가진 기관이 제시한 원칙들을 주로 포함하게 되었다. 이 인용 원칙은 권고 최종안에서도 일관되게 유지되었다.

다음으로 2장에서 초안은 가치와 원칙을 구별하면서 그 차이를 구체성의 수준과 도구적 관계로 설명하고 있다. 즉 가치는 회원국들이 인공지능과 관련하여 추구해야 할 바람직한 도덕적 행위(제도를 포함한)를 제시하는 반면, 원칙은 그 가치가 구체적으로 의미하는

바를 좀 더 자세하게 설명해서 이후 제시될 정책이나 이해당사자들의 실천 수준에서 좀 더 용이하게 가치를 실현할 수 있도록 돕는다는 것이다.

주목할 점은 가치 영역 전체에 걸쳐 인공지능의 전주기에서 인간중심적인 태도를 견지해야 한다는 점이 강조되었다는 사실이다. 인간의 존엄성, 인권, 자유를 특별히 강조하는 부분도 그렇고 인공지능의 혜택이 인류 전체에 골고루 퍼져야 함을 강조하는 '어느 누구도 뒤쳐져서는 안된다(leaving no one behind)' 부분도 그렇다. 이런 수준의 '인간중심주의'에 반대할 사람은 많지 않겠지만 비상설전문가집단 내에서는 인공지능의 미래에는 인간과 인공지능 사이에 대등한 상호작용의 가능성도 배제할 수 없기에 적어도 개념적으로는 인공지능이 일정한 제도적, 도덕적 지위를 가질 가능성을 인정할 필요가 있다는 의견이 제시되었지만 소수의견이어서 초안에 반영되지는 않았다.

인간과 인공지능 사이의 관계를 '인간중심주의'가 요구하는 것보다 보다 포괄적으로 이해하는 입장은 일부 인공지능 기술전문가와 일부 철학자 위원들에 의해 제기되었는데 그밖의 위원, 특히 법률분야 위원들이 강하게 반발하여 최종안을 위한 논의 과정에서도 소수 의견으로 남았다. 다소 편협하게 생각될 수 있는 인간중심주의를 포스트휴머니즘의 맥락에서 전망할 필요성이 있다는 의견에 공감한 일부 중립적 의원들조차 유네스코 인공지능 윤리 권고가 갖는 실

천적 시사점을 약화시키지 않기 위해서는 문서 전체에 걸쳐 인간중심주의를 일관되게 견지해야 한다는 입장을 고수했다. 유네스코 사무국과 여러 영역담당자들의 의견 역시 일치했기에 결국 '인간중심주의'는 약간 '완화된' 형태이기는 하지만 최종안에도 그대로 남았다.[12)]

가치 중에서 '조화롭게 살기(leaving in harmony)' 부분에 대해서는 전문가집단 내에서 상당한 토론이 있었는데 이 가치가 다른 가치로 충분히 표현되지 않은 부분을 추가로 담고 있는지 여부와 설사 그런 내용이 있더라도 가치 영역의 다른 가치가 갖는 보편성과 중요성을 갖는지 여부에 대한 논의였다. 비상설전문가집단 내에서 이 두 사안에 대해 상당한 의견 차이가 있었지만, 이 가치를 적극적으로 옹호하는 위원이 여럿 있어서 일단 초안에서는 그대로 두고 회원국들의 반응을 검토한 후 다시 논의하기로 했는데, 초안에 대한 의견수렴 과정에서 아프리카 회원국들이 이 가치에 대해 강한 긍정 평가를 제시해서 결국에는 최종안에서도 초안보다는 다소 약화된 형태로 유지되었다.

가치 영역에서 인공지능 기술 자체에 관련성을 갖는 항목은 '신

12) 여기서 '완화된' 형태란 최종안에서는 초안과 달리 인간중심주의에 대한 강조만이 아니라 생태계(ecosystem)에 대한 고려를 일관되게 요구했다는 점을 고려한 평가이다. 하지만 이 생태계에 대한 고려가 포스트휴머니즘적 시각을 포함하는 것인지에 대해서는 의원들 사이에 명확한 합의점이 있었다고 보기는 어렵다.

뢰가능성(trustworthiness)'과 '환경보호' 항목이다. 주목할 부분은 이 두 가치를 포괄적으로 정의하고 있다는 점이다. 신뢰가능성은 인공지능 시스템 자체의 작동이 신뢰가능해야 할뿐 아니라 인공지능에 '대한' 개인과 사회의 신뢰가 확보될 수 있는 방식으로 인공지능 시스템의 전주기 관리가 이루어져야 함을 강조한다. 환경보호 역시 인공지능 시스템이 환경을 해치지 않아야 한다는 소극적 요구를 넘어서, 지속가능한 발전이나 기후변화처럼 유엔이 관심을 갖고 있는 영역에서의 해결책 마련 과정에서 인공지능 시스템이 적극적인 역할을 담당해야 한다는 점을 요구하고 있다

원칙은 두 묶음으로 나뉘어 제시된다. 첫 묶음은 인간-인공지능 접점과 관련된 주제를 다루고 있고, 둘째 묶음은 인공지능 시스템 자체의 특징과 관련된 주제를 다룬다. 이렇게 둘로 나누어 원칙을 제시한 배경에는 여러 이유가 있지만, 이 두 묶음의 논의가 적절하게 구별되지 않고 혼재되어 나타날 때 논의가 생산적으로 전개되기 어렵다는 점에 대해 상당한 공감대가 형성되었기 때문이다. 즉 인공지능 시스템의 기술적 특징에 대한 정확한 이해 없이 인공지능과 개인의 삶, 그리고 사회적 측면 사이의 상호작용에 관련된 논의가 이루어질 때 비생산적인(논의 자체가 무의미하다기 보다는 개념적 혼동으로 해결책 마련이 어려운 상황에 이르게 될 가능성이 높다는 의미에서) 논의가 될 가능성을 미연에 방지하자는 것이 두 묶음으로 원칙을 제시하게 된 취지라고 할 수 있다. 그러므로 첫 묶음에 비해 둘째 묶음 관련 주제가 좁은 의미의 인공지능 기술에 대한 적절한

조치를 통해 어느 정도 해결이 가능한 특징을 보인다고 할 수 있다.

이렇게 두 묶음으로 원칙을 제시한 근거에 대해 위원들이 생각을 바꾼 것은 아니지만 최종안에서는 가치와 원칙들을 여러 다른 고려에 의해 다시 섞어 묶는 과정을 거쳤기에 초안에서 지켜졌던 두 묶음 배열 원칙은 더 이상 유지되지 않았다. 이에 대해서는 다음 소절을 참고하기 바란다.

다음으로 앞서 제시된 가치와 원칙을 구체적으로 구현할 수 있는 정책을 설명하는 4장을 살펴보자. 4장 역시 정책이 추구하는 목적에 따라 5개의 묶음으로 나누어 정책행동을 제시하고 있는데 어떤 정책들을 어떻게 묶을 것인지와 현재 묶음에 포함되지 않은 (대개는 비상설전문가집단 내에서 광범위한 지지를 얻지 못한) 정책들을 포함시킬 것인지 여부를 두고 상당한 격론이 벌어졌다. 초안이 발표된 이후 수집된 다양한 의견 개진 내용 중에는 가치와 원칙이 너무 자세하고 세부적이라는 지적이 많았기에 최종안으로 가다듬는 과정에서는 보편적 지지를 받지 못한 가치나 원칙은 처음부터 논의에서 배제하는 방식으로 진행되었다. 그래서 초안 작성 과정에서의 다소 산만했던 논의 과정보다는 보다 집중적인 논의가 가능했다.[13)]

13) 그렇지만 초안에서 최종안으로 가면서 내용이 줄어들기만 한 것은 아니다. 초안에 대한 여러 의견 개진과 유네스코 내부 논의 결과를 반영하여 최종안에서 새롭게 추가된 내용도 여럿 있고 기본적으로 동일한 내용이라도 제시 방식을 달리한 부분도 많다. 결과적으로 (기대와 달리) 초안과 최종안 사이에 분량 차이는 크지 않다.

첫 묶음의 정책 목적은 '윤리적 지킴이(Ethical Stewardship)'이다. 인공지능 시스템의 전주기에 걸쳐 우리가 소중하게 여기는 윤리적 가치가 제대로 반영되도록 노력한다는 의미가 담겨 있다. 단어 선택에서 control보다는 약하고 governance나 manage보다는 좀 더 적극적인 용어가 필요하다는 공감대가 있었고 그래서 결국에는 stewardship으로 결정되었다.

둘째 묶음인 '영향평가(Impact Assessment)'는 인공지능 시스템의 잠재력과 위험 모두 현 단계에서 완전하게 이해하거나 예측하는 것이 불가능하므로 끊임없이 관련 쟁점과 정책의 효과를 모니터링하고 각국의 거버넌스 경험을 공유하면서 보다 바람직한 방식의 인공지능 시스템을 실현해 나가는 것이 중요함을 강조하고 있다.

셋째 묶음은 현재 우리가 여러 수준에서 인공지능 윤리를 적절하게 실천하기 위한 기반이 부족하다는 점을 직시하고 개인적, 제도적, 사회적, 국가적, 국제적 수준에서 관련 역량을 강화하기 위한 여러 정책을 담고 있다. 이는 기존 유네스코가 기후변화 등의 국제적 위기상황에 대한 대응에서 구체적인 대응 방안만이 아니라 장기적으로 위기를 적절하게 해결해 나갈 수 있는 역량 강화를 강조해 온 전통과 일관된 태도라고 판단된다.

넷째 묶음은 인공지능의 윤리적 고려가 지속가능한 발전, 특히 저개발국이나 개발도상국의 발전을 돕는 방식으로 이루어져야 하며,

이 과정에서 국제 협력이 필요하다는 점을 강조하는 정책으로 구성되어 있다. 이는 인공지능 시스템 개발의 이익이 현재 기술적 우위를 선점하고 있는 선진국에 편중될 위험에 대한 우려와 이를 막기 위한 대책의 필요성을 강조한 것으로 회원국 전체의 이익을 고려하는 유엔의 전반적 경향을 반영하고 있다.

다섯째 묶음은 어떤 관점에서 보면 첫 묶음과 겹치는 내용을 담고 있고 실제로 논의 과정에서 몇몇 전문가는 내용이 중복되므로 마지막 묶음은 삭제하거나 축소하자고 주장하기도 했다. 하지만 대다수 전문가들은 첫 묶음이 원칙 수준에서의 인공지능 시스템 개발 및 활용에서의 책임감을 강조하는 정책이었다면, 이 묶음의 정책들은 보다 구체적인 수준에서 회원국들이 바람직한 인공지능 거버넌스 체제를 확립하도록 요구하는 내용을 담고 있기에 포함되어야 한다고 판단하여 현재 초안의 내용으로 발표되었다. 하지만 다섯째 묶음의 정책에 대해서 각국의 정책 자율성을 침해할 정도로 너무 자세하다고 반대할 회원국이 상당 수 있을 수 있는 반면, 일부 전문가는 훨씬 더 강력한 정책을 권고해야 한다고 주장하고 있어서 최종안 논의 과정에서는 상당한 재조정 작업이 이루어졌다.

권고 최종안(9월)의 구조와 내용

권고 최종안의 구조는 초안에 비해 유네스코의 기존 윤리적 틀의

구조를 보다 충실하게 따르고 있다. 또한 초안에 대해 제기되었던 여러 의견과 수정 제안을 반영하여 특히 가치와 원칙 그리고 정책 제언 부분에서 상당한 구조적 변화가 있었다.

권고 최종안이 구조는 다음과 같다. 권고안의 지향점과 핵심 주제를 제시하는 전문(preamble)과 이어지는 8개의 장의로 구성되어 있다. 1장은 이 권고(안)의 범위와 활용(Scope and Application) 목적과 범위를 설명하고 2장은 목적과 목표(Aims and Objectives)를 다룬다. 이 두 장의 내용은 초안에 대한 의견 수렴 과정에서 나왔던 여러 우려, 즉 유네스코의 AI 윤리가 각국의 AI 정책을 지나치게 제한하는 것이 될 수 있다는 우려에 답하고 최종안이 제안하는 권고 정책들이 각국의 상황을 고려하는 방식으로 활용되는 것이 바람직하다는 내용이 담겨 있다.

3장과 4장은 초안에서와 마찬가지로 이 권고(안)의 주요 내용에 해당된다. 3장은 '가치와 원칙'을 다루는데 초안에 대해 제기된 여러 고려 사항을 반영하여 가치와 원칙을 일부 통합하고 가치와 원칙의 규정 방식도 상당 부분 바꾸었다. 변화의 방향은 기본적으로 가치와 원칙의 내용이 보다 더 잘 전달될 수 있고 보다 보편성을 갖도록 함이었다. 즉 초안에 제시되었던 가치와 원칙의 내용에 여러 지역별 의견수렴 과정에서 나온 의견을 종합하여 보다 많은 사람들이 이해하고 공감할 수 있는 방식으로 재구조화된 것이다.

그에 비해 4장은 제목 자체를 정책 행동(policy actions)로 바꾼 것에서 알 수 있듯이 초안에 비해 정책 내용의 배경과 기술전문적인 내용을 상당부분 제거하고 회원국 정부가 정책 입안과 실행을 통해 구체적으로 행동에 나설 수 있는 방식, 즉 '행동가능한(actionable)' 방식으로 재구조화되었다.

5장부터 8장까지는 본 권고(안)의 효율적 활용을 위해 필요한 부가 사항과 유엔 문서로서의 형식적 필요성을 위해 포함된 내용이다. 특히 5장 '모니터링과 평가'는 4장의 정책 행동을 회원국들이 성공적으로 수행하기 위해 필요한 모니터링과 인공지능 윤리 평가 방식에 대한 제언이 담겨 있다. 6장은 본 권고(안)에 대한 해석과 활용이 회원국들이 공통적으로 수용하는 기본 인권 및 핵심 가치를 부정하는 방식으로 이루어져서는 안 된다는 점을 공식적으로 재천명한 것인다. 이는 본 권고(안)의 내용이 문맥을 무시하는 방식으로, 유네스코 회원국 정부 혹은 다른 단체에 의해 악용될 여지를 형식적으로 불허하기 위해 도입된 장이다.

특히 초안에 대한 논의 과정에서 전체주의 역사 경험이 있는 독일의 대표부를 비롯한 몇몇 회원국 대표부가 '조화(harmony)'라는 개념이 역사적으로 전체주의의 인권 탄압을 옹호하는 방식으로 활용되었다는 점을 지적하며 3장에서 제거할 것을 요구했으나 AHEG 전문가 성덩수와 아프리카 국가들이 적극적으로 이 가치를 유지할 것을 주장하였다. 이 상황을 고려할 때 6장은 그같은 위험에 대한 해

결책으로 제시된 장이라고 볼 수 있다. 즉 이 권고(안)에서의 언급된 '조화'는 공동체의 이익을 위해 개인의 희생을 일방적으로 요구하는 것을 정당화하는 의미에서의 '조화'의 의미로 해석되는 것을 형식적으로 불허한 것이다.

하지만 이렇게 추가적인 조치를 통해 '조화로운 삶' 내용이 제기하는 우려에 대응했음에도 불구하고 이후 정부간 협의과정에서 특정 회원국들의 강한 반대에 직면한 내용은 수정될 여지가 많기에 2021년 11월 총회에 상정될 권고(안)에는 '조화'와 같은 논쟁적 개념이 사라지거나 다른 개념으로 대체될 가능성은 여전히 남아 있다.

3장 '가치와 원칙'에서 제시되는 가치는 모두 4가지이며 원칙은 10가지이다. 이는 초안에서 제시했던 가치나 원칙에 비해 숫적으로는 다소 줄었다고 볼 수 있지만 단순히 축소했다고 보기는 어렵고 서로 다른 가치와 원칙을 종합해서 새로운 가치를 만들고 초안에서 미처 다루지 않았거나 가볍게 다루어졌다고 판단된 가치를 보다 전면에, 독립적 가치나 원칙으로 내세우는 방식으로 전반적 재구조화가 이루어졌다고 보는 것이 맞다.

예를 들어 초안에는 인류의 번영만이 강조되었지만 이번 최종안에서는 인류와 그를 둘러싼 환경 그리고 다른 존재자들을 보다 적극적으로 고려하는 생태계 전체의 번역이 강조되었다. 이는 AI 윤리를 지나치게 인간중심주의적으로만 서술하지 않고 환경과 생태계 전반

에 대한 고려를 확보하는 방식으로 확장해야 한다는 전문가 위원회 내부의 견해와 초안에 대한 의견 수렴 과정에서 수집된 지적을 수용한 것이다.

비슷한 방식으로 초안에서 일방적으로 강조되던 다양성 대신 다양성이 포용가능성(inclusiveness)를 증진하는 방식으로 발휘되어야지 개인이 다양성 뒤에 숨어서 공동체에서 오히려 배제되어 버리는 부작용이 생겨서는 안 된다는 생각을 반영했다. 이는 특히 모든 다양성이 그 자체로 좋은 것이라기보다는 포용가능성처럼 우리가 중요하게 생각하는 사회적 가치를 증진하기 위해 도구적으로 좋은 것이라는 생각이 전문가 위원회 사이에서도 (초안 작성 단계에 비해) 최종안 논의 과정에서 더 우세해졌음을 의미한다.

10개 원칙 중에는 앞선 가치 영역과 마찬가지로 초안의 내용에서 통합된 것도 있지만 프라이버시처럼 새롭게 도입된 것도 있다. 초안에서와 마찬가지로 가치와 원칙 사이의 관계는 도구성과 구체성이다. 즉 원칙은 가치를 보다 구체화하고 실현 가능하게 하기 위해 집중해야 할 영역을 지시한다고 볼 수 있다. 당연히 어떤 것이 본질적으로 추구해야 할 가치이고 어떤 것이 그 가치를 실현하기 위한 구체적 영역인지에 대해서는 의견 차이가 있을 수밖에 없고 이는 초안과 최종안에서 가치에 있던 것이 원칙으로 이동하고 원칙의 내용 중 일부가 가치의 설명에 포함되는 상황이 종종 발생했다는 점에서 잘 드러난다.

예를 들어 프라이버시는 초안에서도 중요한 주제로 다루어졌지만 의견 수렴 과정에서 워낙 중요하게 많은 사람들이 언급한 내용이어서 최종안에서는 아예 독립적 원칙으로 강조하기로 의견이 모아졌다. 또한 '투명성과 설명가능성', '책임과 책무성'처럼 서로 관련성이 높은 내용을 한 원칙으로 묶고 각 개념이 서로 어떻게 연결되는지를 보다 많은 사람들이 쉽게 이해할 수 있도록 설명하려 노력했다.

최종안의 원칙 중에 눈에 뜨이는 주제어는 '적응적 거버넌스(adaptaive governance)'라는 개념이다. 비상설전문가집단 내에서 유네스코 AI 윤리가 제안하는 정책의 내용과 방식이 어떠해야 하는지를 두고 상당히 많은 격론이 있었다. 특히 모든 것을 강하게 규정할 것을 요구하는 위원들과 대부분의 것을 인공지능 개발자나 회원국의 자율적 선택에 맡기자는 위원들 사이에서 다양한 스펙트럼의 의견이 개진되었다. 초안이 발표된 후 의견 수렴 과정에서도 아프리카 국가들 상당 수는 유네스코 권고안의 규범성을 보다 강화해야 한다고 (그래서 should 대신 must를 사용해야 한다고) 목소리를 높이기도 했다. 반면 영국 등의 인공지능 기술 선진국은 권고안의 규범력을 최소한으로 하고 상식적 의미의 '권고'가 되어야 한다고 주장했는데, 이 생각은 지역 의견 수렴 과정에서 몇몇 산업계 인사들로부터 지지를 받았다.

이런 맥락에서 위원회 내에서 절충점으로 합의된 개념이 '적응적 거버넌스'이다. 우선 유네스코 권고안이 의의를 갖기 위해서는 권고(안)의 규범력을 손상시키지 않아야 한다는 데 거의 모든 위원들이 찬성했다. 근거는 그렇게 약화된 권고(안)은 이미 나와 있는 여러 다른 AI 윤리 관련 선언이나 가이드라인과 차별성이 없다는 것이다. 유네스코가 많은 노력을 기울여 새롭게 AI 윤리 권고를 제안하는 상황에서 기존 논의에 더하는 바가 없다면 그 노력 자체가 의의를 갖기 어렵다는 대다수의 위원들이 동의한 것이다. 그래서 권고안의 내용 중에 일부 should는 (그것이 타당하다고 합의한 경우에) must로 전환하는 방식으로 초안보다 최종안의 규범성을 강화했다.

하지만 그와 동시에 위원들은 유네스코 회원국들의 제도적, 법적, 사회적 하부구조가 동일하지 않다는 점에 주목했다. 즉 규범적으로 유네스코 권고(안)에 공감하는 회원국조차 이 권고(안)의 내용을 실행하기에는 제도적, 경제적으로 어려움이 많을 수 있다는 점을 공감한 것이다. 또한 인공지능 기술의 특징상 앞으로 인공지능 기술이 어떤 방향으로 어떻게 개발되고 전체 사회에 영향을 끼칠 것인지에 대해서는 상당한 불확실성이 존재한다는 점 역시 재확인했다. 이런 상황에서 너무 지나치게 자세한 정책 권고를 하거나 초기에 타당한 정책 권고가 미래에는 타장하지 않을 수 있을 가능성을 반영하지 않는 방식으로 권고(안)이 작성되는 것은 바람직하지 않다는 점에도 동의했다. 이런 합의점에 근거하여 비상설전문가집단 위원들은 우리가 제안하는 규범적 틀이 회원국들의 국소적 환경의 구체적인 내

용과 인공지능 기술의 개발 현황 및 미래 영향의 내용에 적극적으로 대응하는 방식으로 제도화될 필요가 있다는 점에 공감했다. 즉 우리가 제안하는 인공지능 국제 및 국내 거버넌스는 '적응적' 성격을 가져야 한다는 것이다. 이런 이유로 마지막 원칙은 이런 '적응적 거버넌스'가 다자주의적 고려와 협력에 입각하여 실천되어야 한다는 내용을 담게 되었다.

4장 정책 행동은 초안에서 다소 혼란스럽게 나열된 다양한 정책 행동을 그 정책이 효과를 발휘하도록 의도된 '영역'별로 재구조화되어 제시되었다. 중요한 점은 첫 영역이 인공지능에 대한 '윤리 영향 평가'라는 점이다. 즉 인공지능 윤리와 관련된 모든 정책 행동은 각 회원국들이 국소적 환경에서 인공지능이 어떤 영향을 끼치고 있는지를 지속적으로 모니터링하고 평가하는 노력에서 출발해야 한다는 점을 강조한 것이다. 그럴 때만 둘째 영역인 인공지능에 대한 윤리적 거버넌스와 돌봄이 성취될 수 있다.

나머지 8가지 정책 영역은 인공지능 관련 정책 행동이 집중해야 할 대표적인 영역을 망라한 것인데, AI 윤리에 대해 논의될 수 있는 모든 주제를 전부 포함시는 방식을 피하고, 유네스코가 강점을 지닌 분야를 강조하는 방식으로 주제를 선정하였다. 예를 들어 마지막 정책 영역인 '건강과 사회적 복지'는 2020년 등장한 COViD-19과 같은 공중 보건위기를 염두에 두고 추가된 것이다. 기존 초안에도 인공지능 연구가 인류 복지에 이바지할 잠재력에 대한 언급은 여럿 있

었지만, 이번 최종안에서는 회원국의 일반 국민에게 가장 직접적으로 다가갈 수 있는 인공지능 윤리 쟁점을 포함시키자는 생각에서 공중보건 영역에서의 인공지능 활용과 관련된 윤리적 쟁점을 따로 분리해서 제시했다. 비슷한 이유로 여러 곳에 흩어져서 논의된 데이터 관련 정책을 아예 '데이터 정책'으로 분리하여 독립 정책 영역으로 제시했다.

10개 정책 영역은 위계적 구조를 염두에 둔 것은 아니며 서로 다른 정책 영역 사이에 우선 순위를 염두에 두었다고 말하기도 어렵다. 보다 정확하게 말하자면 정책 영역 사이의 우선 순위가 필요하다는 의견 자체가 논쟁적이었다. 우선 순위를 두자는 의견이 여러 위원들에 의해 제시되었으나 논의 과정에서 우선 순위를 어떻게 두는 것이 좋을 지에 대한 의견 일치가 어렵다는 점이 분명하게 드러나서 현재 형태로 병렬적 제시를 택하게 되었다. 하지만 그렇다고 해서 영역 제시 순서에 아무런 의미가 없는 것은 아니다. 우선 앞서 설명했듯이 첫 두 정책 영역은 본 권고(안)의 지향점과 권고되는 정책의 효율성을 보장하기 위한 전제 조건이 담겨 있다. 그리고 인공지능 윤리 논의에서 데이터 관련 쟁점이 가장 두드러지고 여러 주제에 걸쳐 있다는 사실에 주목하고 '데이터 정책'을 셋째 정책 영역으로 제시했다.

또한 유네스코가 회원국의 자발적 협력을 강조하는 단체이며 본 권고(안)이 강조하는 '적응적 거버넌스'가 효과적으로 이루어지기

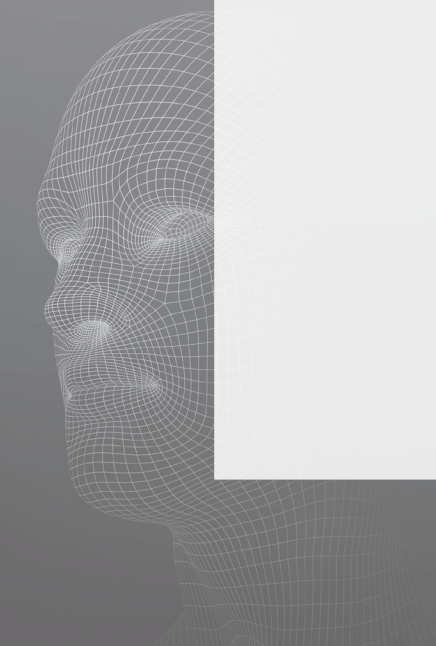
위해서는 회원국 사이의 국제협력이 무엇보다 중요하다는 점을 강조하기 위해 넷째 정책 영역으로 '발전과 국제 협력'을 제시했다. 여기서 '발전'이란 AI 기술 발전만을 의미하는 것이 아니라 AI 윤리 역량의 발전도 함께 의미한다. 즉 본 권고(안)은 유네스코 회원국 사이의 기술적 협력만이 아니라 윤리역량 제고를 위한 협력도 함께 강조하고 있는 것이다.

그러므로 전체적으로 볼 때 최종안에서 채택된 정책 행동의 구조는 회원국들이 권고(안)에 제시된 윤리 원칙과 가치를 구체적인 정책으로 실행하기 위해 필요한 일종의 가이드로 작동할 수 있도록 구성되었다고 볼 수 있다. 즉, 유네스코 AI 윤리 권고를 채택하고 이를 실천하기 위해 기본적으로 어떤 제도적 장치(예를 들어 AI 윤리영향평가)를 시행해야 하는지, 그리고 어떤 핵심 영역에서 어떤 쟁점에 특별히 주의를 기울여야 하는지, 실천 과정에서 유네스코 회원국 사이에 어떤 국제 협력이 어떤 방식으로 가능할지에 대해 설명하고 있는 것이다.

제4부

쟁점과 대응방안

AI, AI 기술, AI 시스템
인간중심주의를 어디까지 요구할 것인가?
AI 윤리 평가 실행방안



04 쟁점과 대응방안

이 절에서는 권고(안) 작성을 위한 비상설전문가집단 논의 과정에서 쟁점으로 부각된 주제들과 그것이 대한민국의 대응 방향에 갖는 시사점을 살펴본다.

AI, AI 기술, AI 시스템

비상설전문가집단 위원회 논의 초반부터 최종안을 도출하기까지 지속적으로 논쟁적이었던 사안은 AI를 어떻게 정의할 것인지의 문제와 본 권고(안)에서 AI라는 용어를 사용하는 것이 적절한지, 혹은 다른 용어로 대체해야 하는지 여부였다.

우선 AI의 정의에 대해서 위원회 내부의 기술 전문가들은 기술적으로 상세하고 정확한 정의를 본문에 제시하는 것을 선호한 반면, 다른 분야 전문가들은 이 권고(안)을 읽고 활용한 정책입안자들이나 일반 시민이 이해할 수 있고 자신의 삶이나 사회적 쟁점과 연결할 수 있는 정의를 선호했다. 이 두 정의 사이에는 상당한 간극이 있었기에 AI 정의에 대한 논쟁은 위원회 논의 과정 내내 진행되었다.

결국 최종안이 채택한 입장은 본 권고(안)의 독자를 고려하여 자나치게 전문적인 인공지능 정의를 피하되 적어도 기술전문가가 보기에 오류라고 판단할 내용은 제거하는 방식의 절충이었다. 기술전문가 위원들이 이 절충안을 수용하게 된 이유는 문서 전체에서 인공지능과 관련된 모든 문장에 AI 용어를 사용하는 대신 좀 더 구체적인 상황이 요구하는 맥락에 따라 AI technologies, Ai systems 과 같은 대안적 용어를 사용하기로 결정했기 때문이다. 이렇게 서로 연결되지만 부분적으로만 의미가 겹치는 여러 용어를 사용하는 것은 문서 전체의 일관성을 해친다는 의견도 있었지만, 관련 기술적 오류를 피하면서도 권고(안)의 이해가능성과 활용가능성을 높이기 위해 필요불가결한 방식이라는 점에 대해 의견이 모아졌다.

우리나라에서도 인공지능 혹은 AI에 대한 이해는 전문성의 내용이나 사용 맥락에 따라 상당히 다르게 나타난다. 특히 인공지능 기술 자체를 지칭하는지 인공지능 기술이 활용된 제품을 가리키는지에 대한 혼동은 관련 법률이나 제도를 만드는 과정에서 반드시 정리될 필요가 있는 쟁점이다. 그에 더해 인공지능이 사회적으로 영향력을 갖기 위해서는 반드시 인간 행위자를 포함한 다양한 대상들과 시스템으로 묶여야 된다는 점을 고려할 필요가 있다. 이 점을 고려하면 인공지능이 마치 신적인 존재처럼 스스로 알아서 인류에게 해를 줄 수도 있다는 생각의 문제점을 금방 알아챌 수 있다. 유네스코 권고안도 이 점에 주목하여 인공지능의 윤리적 쟁점을 언급할 때 거의 모든 맥락에서 AI나 AI technologies가 아니라 AI system이라는

용어를 사용했음에 주목할 필요가 있다.

이처럼 AI 개념 및 관련 개념을 두고 벌어진 권고(안) 작성 작업 과정에서의 논쟁은 우리 정부거 AI 관련 정책을 입안하고 실행하는 과정에 중요한 교훈을 제시한다.

인간중심주의를 어디까지 요구할 것인가?

앞서 설명했듯이 초안 작성 과정까지 절대 다수의 위원회 위원들은 인공지능 관련 윤리적 논의에서 가장 중요하고 강조되어야 할 점은 인간중심주의와 인권이라고 주장했다. 하지만 논의 과정에서 인간중심주의가 인공지능의 특성상 인공지능 관련 모든 결정에서 인간이 항상 주도권을 가져야 한다는 식으로 요구되기 어렵다는 점이 지적되었다. 또한 인간중심주의적 생각이 최근 유엔을 중심으로 진행되는 여러 환경 및 생태계 관련 논의와 충돌될 수 있다는 가능성도 제기되었다. 그래서 결국 최종안에서는 인간중심주의를 여전히 강조하되 여러 가치와 원칙 사이의 충돌 가능성을 명시적으로 언급하고 그럴 경우에는 관련 사안을 잘 검토하여 적절한 해결책을 책임 있는 방식으로 모색해야 한다는 내용으로 정리되었다.

인간중심주의는 인간적 가치, 혹은 인문적 가치가 인공지능의 개발 및 활용 과정에서 존중되어야 한다는 의미에서는 논란의 여지가

없이 받아들여진 생각이다. 이는 유네스코 AI 윤리 권고(안)만이 아니라 대부분의 국제 AI 윤리 논의가 수용하고 있는 입장이다. 유네스코 권고안에서는 이런 의미의 인간중심주의를 기본권에 대한 강조를 통해 보다 구체적으로 천명했다.

하지만 인간중심주의를 특정 개인이 인공지능의 작동 과정에서 언제든지 개입할 수 있고 인공지능의 작동 과정을 완벽하게 이해해야 한다는 요구조건으로 이해하는 것은 여러 의도치 않은 문제를 일으킬 수 있다. 특히 이렇게 이해된 인간중심주의 조건이 만족되지 않을 경우, 예를 들어 투명성이 완벽하게 보장되지 않을 경우는 항상 윤리적으로 바람직하지 않다는 생각은 기술적으로나 실천적으로 견지되기 어렵다는 점에 대해 위원들 사이에서 합의가 이루어졌다. 일단 인공지능 기술은 자동화 기술의 일종이라고 할 수 있는데 인류가 자동화 기술을 개발하여 사용하는 것은 정확히 인간이 자동화된 기계의 매 순간에 일일이 개입하지 않고도 원하는 결과를 얻기 위해서이다. 그러므로 인공지능의 작동의 매 단계마다 인간이 개입할 수 있는 권한을 부여해야만 윤리적으로 바람직한 인공지능이라는 생각은 자칫 인공지능 기술 발전 자체를 억제하거나, 부인하는 결과를 초래할 수 있다.

또한 최근 각광을 받고 있는 딥러닝 기법의 특징은 인공지능의 작동, 특히 왜 어떤 결과를 내놓게 되었는지에 대해 인공지능 제작자

조차도 인과적으로 상세한 이해를 하기 어렵다는 데 있다.[14] 물론 그렇다고 해서 인공지능의 작동이 완전히 불투명한 것은 아니다. 인공지능 기술자들은 인공지능이 어떻게 작동하는 지 전체적으로 파악하고 있으며 어떤 변수를 어떻게 조작해야 어떤 결과가 나오는지도 알고 있다. 다만 구체적인 결과에 대해 그 결과를 얻어내는 과정에 개입한 모든 인과 과정을 파악하기는 어렵다는 의미다. 이런 상황은 불투명성보다는 반투명성에 가깝다.

문제는 인공지능 기술과 관련해서는 반투명성이 완전한 투명성보다 대부분의 경우 더 효율적이고, 어떤 경우에는 더 윤리적일 수도 있다는 사실이다. 인공지능 기술이 최근 급속도로 발전한 것은 인간이 이해할 수 있는 방식으로 작동하는, 즉 완전히 투명한 인공지능보다 인간이 부분적으로만 이해할 수 있는 반투명의 인공지능의 효율성이 더 높기 때문이다. 그런 의미에서 모든 인공지능에게 완전한 투명성을 요구하는 방식으로 인간중심주의를 요구한다면 이는 인공지능의 효율성을 떨어뜨리는 결과를 가져올 수밖에 없다. 그리고 이런 상황은 인공지능이 높은 효율성으로 작동하는 것이 개인이나 사회의 복지에 결정적인 상황(예를 들어 의료 인공지능이나 스마트 네트워크 제어 인공지능)에서는 윤리적으로도 바람직하지 않다.

14) 1절의 '낯선' 인공지능 개념이 이 논의 맥락에서 중요하다.

그렇다고 해서 위원회가 인공지능에게 투명성이 결코 필요하지 않다고 생각한 것은 아니다. 그보다는 앞서 지적했듯이 투명성이나 설명가능성 모두 우리가 추구해야 할 가치이지만 인공지능의 효율성 등 다른 바람직한 가치와 함께 고려해야 할, 그래서 적당한 수준에서의 맞교환(trade-offs)이 불가피하다는 점을 인정해야 한다는 것이다.[15] 결국 결정적인 사안은 언제 투명성을 보다 우선적으로 요구하고, 언제 효율성을 보다 우선적으로 요구할 것인지에 대해 사회적 논의와 정책적 결정이 이루어져야 한다는 것이다. 본 권고(안)은 정확히 이 작업을 회원국이 수행할 것을 요구하고 있다. 당연히 우리 정부도 이런 논의를 당장 시작할 필요가 있다.

AI 윤리 평가 실행 방안

앞서 설명했듯이 최종안에서 '적응적 거버넌스'와 '맞교환' 그리고 지속적인 모니터링의 필요성이 강조되면서 초안에서는 회원국들에게 다소 부담스러울 수도 있겠다고 평가되었던 AI 윤리 평가에 대한 구체적인 논의가 담기게 되었다. 핵심은 본 권고(안)에서는 AI 윤리 평가의 형식과 내용을 구체적으로 제시하지 않고 방향성과 국제협력을 강조하되, 이후에 유네스코가 이 논의를 본격적으로 시행하기로 결정하는 방식으로 마무리 되었다.

---

15) 우리가 추구하는 다양한 가치 사이의 '적절한 맞교환'의 중요성에 대한 강조는 유네스코 AI 윤리 권고(안) 전체에 걸쳐 일관되게 유지된 입장이다.

이는 규제의 강도에 대한 논쟁에서 강한 규제와 부드러운 규제 중 어느 것을 택할 지에 대한 의견 차이와도 관련되지만, 그보다는 회원국들이 규범적 수준에서 권고(안)에 동의하더라도 제도적 기반이 잘 마련된 일부 회원국을 제외하면 AI 윤리 영향 평가를 실제로 수행할 역량이 부족하다는 점을 위원들이 함께 인식한 이유가 더 크다. 즉 설사 특정 '강도'의 규제가 가장 바람직하다는 데 회원국들의 의견이 모이더라도 그 규제를 제도적으로 실천할 수 있는 역량을 갖춘 회원국이 많지 않을 수 있다는 점을 고려해야 했던 것이다.

그러므로 위원회는 현 단계에서 유네스코 다수 회원국의 역량에 비추어 다소 무리한 요구가 될 수 있는 AI 윤리 영향평가를 규범적 수준에서만 요구하되 거기에서 멈추지 않고 그 영향 평가를 수행할 수 있는 도구와 방법 등을 이후 독자적인 연구를 통해 회원국들에게 제공하고 국제 협력을 통해 회원국들이 AI 윤리 영향 평가를 생산적인 방식으로 활용하여 유네스코 AI 윤리 권고를 보다 충실하게 이해하도록 노력하겠다는 의도를 밝혔다. 이 점에 대해서는 특히 유네스코 사무국과 라모스 부사무총장의 의지가 분명해서 2021년 이후 AI 윤리에 대한 유네스코 활동의 중요한 한 축이 될 것으로 예상된다.

이 점을 고려할 때 우리 정부도 유네스코가 이후에 시작할 AI 윤리 영향 평가 관련 논의에 적극적으로 참여할 필요가 있다. 이 과정에서 우리 정부는 〈과학기술기본법〉에 입각하여 수행 중인 기술영향평가나 다른 여러 부처에서 시행 중인 영향평가의 경험을 유네스

코 회원국과 공유하면서 AI 윤리 영향 평가의 형식과 방법론 개발에 적극적으로 참여하는 것이 바람직해 보인다. 특히 우리나라의 인공지능 기술 개발 및 산업화 정도가 미국, 중국, 영국, 캐나다 등 기술 선진국에 비해서는 부족하지만 관련 인프라조차 부족한 대다수 유네스코 회원국에 비해서는 상당히 좋은 조건이라는 점을 고려할 필요가 있다. 이렇게 인공지능 기술 선진국의 입장과 기술 저개발국의 입장을 동시에 이해할 수 있는 우리나라가 AI 윤리 평가의 방법론과 틀 마련에 있어 중요한 역할을 수행할 수 있는 좋은 조건을 갖추고 있다고 생각할 수 있다.

제5부

제언: 대한민국의 대응 방안

한국의 국가경쟁력을 고려해야
데이터 사용에 대한 제도적 규제 수준
인공지능 윤리 관련 규제 방식
인공지능의 윤리적, 사회적 영향 측정

05 제언: 대한민국의 대응 방안

유네스코 AI 윤리 권고(안)은 코로나-19 사태로 공동숙의가 쉽지 않은 상황에서 도출된 결과이지만 나름대로 다양한 관점과 회원국들의 이해를 담아내려 노력이 확인된다고 볼 수 있다. 하지만 최종안 도출 과정에서조차 몇몇 쟁점에 대해서는 AHEG 전문가 사이에서도 상당한 의견 차이가 남아 있다는 점에 주목할 필요가 있다. 물론 그 의견 차이는 대부분 윤리적 가치나 원칙에 대한 것이기보다는 그 가치와 원칙을 실행하는 구체적인 정책 내용, 특히 그 정책 내용이 얼마나 강한 규범적 힘을 가질 것인지에 집중되어 있기는 하다. 하지만 유엔 산하 기구로서의 유네스코가 다루어야 할 내용이 다른 유엔기구에서 AI와 관련하여 이루어지고 있는 논의와 어떤 관련을 맺어야 하는지처럼 보다 정치적인 측면이 강한 주제도 있다.

특히 유네스코의 규범적 틀 논의에서 항상 최소주의적 입장을 견지하던 미국이 아예 유네스코를 탈퇴한 현 상황에서 다수 회원국의 이해가 다양한 방식으로 조정되고 전문가집단 내에서 충분한 공감대를 확보하는 방식으로 최종안이 작성되었다는 점에 주목할 필요가 있다. 즉 이 최종안은 위원회의 주도하에 작성된 것은 맞지만 5월 초안에 비해 다양한 이해당사자들의 의견이 반영된 '절충안'이라

는 것이다. 게다가 이 최종안이 그대로 2021년 총회에 상정되는 것도 아니다. 이 최종안은 그야말로 '안'일 뿐이고 2021년 여름에 열릴 정부간 회의에서 내용이 최종적으로 조율되어 총회에 상정할 안을 확정하게 된다.

그러므로 우리나라도 일단은 우리 정부가 생각하기에 바람직한 방향에 대해 이후 진행 과정에서 의견을 적극적으로 개진할 필요가 있다. 의견을 개진한다고 해서 반영된다는 보장은 없지만 의견을 개진하지 않으면 아예 반영될 가능성조차 없기 때문이다. 그리고 우리나라의 이해관계와 직접적으로 관련된 주제에 대해서는 우리의 입장을 부처간 협력을 통해 우선 정하고 이것을 국제적 공감대를 얻어낼 수 있는 형태로 조정한 다음, 정부간 협의과정에서 다른 회원국들을 설득할 논리를 적극적으로 개발하고 이를 바탕으로 적극적으로 의견을 개진하고 합의 도출 과정에 참여할 필요가 있다.

특히 권고(안)의 내용 중에서 우리 정부에게 여러 의미에서 중요한 내용이 있다면, 이번 기회에 정부간 관련 부처 협의를 통해 어떤 방식의 대응이 가장 적절할 것인지에 대한 논의가 필요하다. 인공지능 기술과 그 사회적 파급효과는 현재도 교육부, 과학기술부, 산업부, 보건복지부 등 여러 부처에서 각기 다양한 방식으로 진흥 정책을 쏟아내고 있는 상황이다. 이런 상황에서는 유네스코 회원국 전체에 영향을 주는 권고(안)에 대한 효율적 대응이 어려울 수 있다. 그러므로 내년 정부간 회의에서 적극적으로 관련 내용에 대한 논리를

만들어 총회에 상정될 최종안에 반영될 수 있도록 노력하는 것이 물론 중요하지만 그 전에 정부 부처간 협의를 통해 합리적인 우리의 최종 의견을 만드는 작업에 충분한 시간과 노력을 투여할 필요가 있다 다음은 보다 구체적인 제언 내용이다.

AI 분야에서 한국의 국가경쟁력을 고려한 의견 개진 필요

유네스코 회원국은 2021년 여름에 개최될 정부간 회의에서 대체적인 윤리 원칙에는 동의하면서 자국 상황에 비추어 불리한 내용에 대해 변경 혹은 삭제를 요구하거나 보다 적극적으로는 자국에게 이익이 되는 내용을 포함할 것을 제안할 가능성이 높다. 다만 자국의 이익을 전면으로 내세우는 방식이 아니라 나름의 논리를 세워서 왜 자신들이 제안하는 변경/삭제/추가 내용이 타당한지를 타 회원국에게 설득하려 할 것이다. 이런 점을 고려하여 우리 정부도 우리나라의 법제도, 산업환경, 국민 정서 등에 비추어 민감한 부분에 대한 변경/삭제/추가를 요청하되 그 논리를 잘 만들어서 궁극적으로 내년 유네스코 총회에서 심의될 최종 권고안에 반영되도록 노력해야 할 것이다.

이 점이 반드시 대한민국의 의견 개진에 불리한 점은 아니라는 사실을 짚고 넘어갈 필요가 있다. AI 기술 개발에서 대한민국의 위치가 중간적이기에 미국이나 유럽 같은 기술 선진국의 입장과 아프리

카처럼 기술 저개발국의 입장 사이에서 적절한 절충안을 제시함으로써 대한민국이 AI 윤리를 구체화하는 과정에서 중요한 역할을 수행할 수도 있기 때문이다. 이러한 가능성을 염두에 둘 때 대한민국 정부는 AI 윤리에 대한 국제 논의에 적극적으로 참여하는 것이 무엇보다 중요하다.

데이터 사용에 대한 제도적 규제 수준 관련

이 주제는 전문가 회의 중에서도 매우 민감한 쟁점이었다. 다만 인공지능 기술발전을 위해 데이터 관련 규제의 강도를 낮춰야 한다는 강한 의견을 주장하는 전문가가 없었기에,[16] '데이터 주권'이라는 매우 강한 개념까지 사용하면서 데이터 사용에 대한 제도적 규제 정책이 권고되어 있다.

데이터 주권 관련 논의는 여러 전문가, 특히 인공지능 기술 선진국으로부터 데이터 수집 및 활용 과정에서 착취당하고 있다고 느끼는 회원국 출신 전문가들에 의해 주도되었고 어떤 방식으로 권고안에 이 내용을 담을 것인지를 두고 상당한 시간 논쟁이 이어졌다. 결국 최종안에서 데이터 주권은 규범적인 측면을 강조하되 지나치게

16) 보다 정확하게 이야기하자면 세 명 정도의 기술전문가가 권고(안)의 데이터 윤리 관련 부분이 산업적 이해나 기술발전을 막을 가능성에 대해 의견을 내긴 했지만 전체 논의의 흐름에 거의 영향을 미치지 못했다.

자세하게 제도적 행동을 요구하지는 않는 방식으로 절충이 이루어졌지만 여전히 필자 개인적으로는 '데이터 주권' 개념 자체에 대한 우려를 갖고 있다. 필자는 논의 과정에서 데이터 주권이 독재국가 등에 의해 오용될 가능성을 지적했고 대부분의 전분가 위원들이 이 가능성에 공감했지만 그럼에도 불구하고 '데이터 주권'이 기술 저개발국에서는 기술 선진국에 대해 사용할 수 있는 일종의 제도적 대응책이라는 점이 강조되어 결국에는 데이터 주권에 대한 명시적 언급이 최종안에 포함되었다.

필자 개인의 이런 생각과 별도로 대한민국 정부의 입장은 다를 수 있다. 특히 우리나라 인공지능 기술이나 데이터 활용 수준이 미국/중국처럼 다른 나라의 데이터를 적극적으로 활용하는 인공지능 기술 선진국이나 아프리카처럼 자신의 데이터를 '빼앗기는' 상황을 걱정하는 기술 후진국의 측면을 동시에 갖고 있는 중간적 위치라는 점에 주목할 필요가 있다. 또한 우리나라 기업들이 해외에서 적극적으로 데이터를 수집하고 있는지(예를 들어 웹툰 시장의 발전으로 네이버 등은 동남아시아에서 상당한 데이터를 수집하고 있을 것으로 예상된다.), 그 과정에서 데이터 수집 대상국 정부와 어떤 조정을 거쳤는지, 앞으로의 향후 전망은 어떤지에 대한 정보를 수집해서 차분하게 판단해야 한다. 그리고 이 과정에서 우리가 제안하는 내용이 미국 등의 기술 선진국이 얻을 이익과 비교해 충분히 큰지 여부도 고려해야 한다. 만약 그 이익이 크지 않다면 데이터 주권 등을 강조하는 기술 후진국과 외교적으로 껄끄러운 상황을 구태여 만들 이유는

없기 때문이다.

또한 여기에 더해 최근 통과된 데이터3법이 유럽의 GDPR 규정과 어떤 공통점과 차이점이 있는지, 그리고 우리가 유럽 시장에 진출하는 과정에서 데이터3법이 유럽 규정과 충돌을 빚을 가능성은 없는지 등에 대한 법률적 검토를 철저하게 해야 한다. 현재 유네스코 인공지능 윤리가 채택한 데이터 윤리 관련 정책은 대체적으로 유럽의 GDPR에 인공지능 기술 저개발국의 데이터 주권 논의가 더해진 것으로 볼 수 있기 때문이다. 그러므로 유럽의 GDPR의 배경 논리를 잘 검토해서 우리 입장에서 우리의 제도(데이터3법까지 포함해서)가 GDPR의 규제망을 충분히 넘을 수 있도록 논리를 개발할 필요가 있고, 이 논리를 그대로 유네스코 인공지능 윤리 권고안에 대한 우리 정부의 의견 개진에 활용할 수 있을 것이라 판단한다.

마지막으로 데이터 사용 관련 '이익 공유' 문제를 검토할 필요가 있다. 현재 유럽 법원이나 유럽연합 집행위원회에서 잇달아 구글이나 페이스북 같은 미국 기업에 대해 벌금이 부과되고 있는데, 이는 표면적으로는 관련 규정을 어겨서라는 이유를 달고 있지만 보다 근본적으로는 유럽에서 영업하면서 유럽인의 데이터로 얻은 이익을 '세금'의 형태로 제대로 이익 공유하지 않았다는 인식에서 출발하고 있다. 그러므로 우리나라 기업도 현재 동남아시아나 추후 유럽에서 데이터 기반 인공지능 사업을 하기 위해서는 이 이익 공유를 어떤 방식으로 하는 것이 각국의 문화적 전통과 사회적 직관에 비추어 적

절할 것인지를 고민할 필요가 있다.

그런데 이 고민은 실은 우리 정부각 국내에서 이루어지는 여러 인공지능 활용과 관련해서도 선제적으로 할 필요가 있다. 국내에서도 아마 조만간 인공지능이 활용하는 데이터 사용으로부터 얻는 이익 공유에 대한 문제 제기가 (아마도) 기업 이익의 사회적 환원 형태로 제기될 가능성이 높기 때문이다. 우리가 관련 서비스를 활용하기 위해 '할수 없이' 동의를 끊임없이 눌러대고 있다고 해서 사람들이 이 주제에 대해 둔감하다고 생각하는 것은 너무 상황을 쉽게 생각하는 것이다. 생물다양성 협약을 통해 생물자원에 대한 이익 공유 문제가 제도화되었듯이, 장기적으로는 인공지능 관련 데이터 활용과 관련된 이익 공유 문제가 제도화될 가능성이 있고 이에 대해 정부가 미리 고민하여 이번 유네스코 AI 윤리 권고안 작업 등에서 의견을 개진하는 것이 좋을 것이다.

사실 필자의 개인적 생각으로 이 사안은 이익 공유보다는 인공지능 데이터 활용에 대한 시민의 신뢰를 충분히 확보할 수 있는 거버넌스를 만든 후에 일종의 데이터 공공재 개념을 적용하는 것이 적절하다고 본다. 하지만 이는 필자의 개인적 견해인데다, 현재 인공지능 관련 국제 동향을 고려할 때 실현될 가능성은 높지 않다고 판단한다.

AI 관련 규제 방식에 대한 고려

이 문제는 전문가집단 내부에서도 매우 긴 논쟁이 이어졌던 주제이다. 핵심은 인공지능 윤리 관련 제도를 만들 때 'hard'하게 만들 것인지, 'soft'하게 만들 것인지의 선택이다. 일반적으로 과학기술과 관련된 인문학적 논의 맥락에서는 (실제로 실천하기는 쉽지 않지만) '자율 규제'를 최대한 시도하는 것이 바람직하고, 특정 윤리 쟁점에 대해 사안마다 판단하고 강한 규제 권한을 가진 기관의 설치는 필수 불가결한 경우에만 허용하는 것이 바람직하다고 판단한다. 이는 인문학자들이 '작은 정부'를 지향하는 보수주의자여서가 아니라, 인공지능 관련 이해당사자들이 윤리적 규범을 내재화해서 자율적으로 실천하기 위해 서로 협력하지 않는다면 인공지능 윤리가 사회적으로 진정한 효력, 즉 행위자의 '자율성'에 근거한 효력을 발휘하기 어렵다고 생각하기 때문이다.

예를 들어 과학자들의 연구 부정행위와 관련된 연구윤리 제도화 과정에서 미국과 유럽은 서로 다른 방향으로 접근한다. 미국은 '법률가의 나라'답게 연구자가 지켜야할 규정을 매우 세세하고 만들고 이를 바탕으로 판단하여 적절한 처분을 내리는 상설기구와 같은 규제 제도를 선호한다. 이에 비해 유럽은 사안마나 자율적 해결을 먼저 시도하고 그것이 실패할 경우 중재를 하거나 법적 처분을 '제안'하는 전문가 위원회를 활용한다. 두 제도 모두 장단점이 있지만 인공지능 윤리 쟁점처럼 미래에 어떤 것이 쟁점이 될지 여부 자체가

이후에 전개될 기술 개발과 사회적 논의에 결정적으로 의존하는 사안에 대해 미리 제도적으로 세세하게 규정하려 시도하는 것은 그다지 생산적이라고 생각되지 않는다.

그보다는 인공지능이 제기하는 윤리적 쟁점의 중요성을 회원국 정부가 분명하게 인식하고 관련 연구와 교육을 통해 인공지능 윤리에 대한 사회적 공감대를 충분히 확보한 후에 차근차근 합의할 수 있는 제도적 규제를 도입하는 것이 더 생산적일 것이다. 이런 방식이 위원회 내에서 이야기했던 '부드러운 규제' 혹은 '적응적 규제'의 방식이다.

위원회 내부의 논의 과정에서 초기에는 강력한 규제를 원하는 전문가가 조금 더 많은 편이었지만, 논의가 진행될수록 인공지능처럼 기술내용이나 사회적 파급효과에 있어 불확실성이 많은 기술을 미리 너무 자세하게 규제하는 것은 윤리적으로도 바람직하지 않다는 점에 많은 전문가들이 공감하게 되었다. 결국 대다수 위원들이 '적응적 거버넌스'가 윤리적 무책임이나 경제 논리에 윤리적 고려를 희생시키는 것이 아니라 인공지능 기술 자체 발전과 사회와의 상호작용이 갖는 근본적인 불확실성을 고려한 것이라는 점을 인식하게 된 것이다. 그러므로 이런 방향으로 인공지능에 대한 윤리적 거버넌스 제도화에 대한 의견을 정부간 협의 과정에서 한국 정부가 낸다면 회원국 사이에서 상당한 공감대를 얻을 수 있을 것으로 기대된다.

AI 윤리적, 사회적 영향 측정 관련

마지막으로 초안에 비해 최종안에서 더욱 확장되고 구체적인 정책 제언을 포함하게 된 인공지능의 윤리적, 사회적 영향평가와 관련된 사안이 중요하다. 개인적으로 필자는 2003년부터 국내에서 진행중인 기술영향평가(Technology Assessment)에 참여해 왔는데 이때 느꼈던 문제의식과 유네스코 AI 윤리(안)에서 제시한 윤리적 평가와 상당한 공통점이 있다.

필자의 문제의식은 인공지능과 같은 첨단 과학기술의 '영향(impact)'은 예측하거나 측정하기가 매우 어렵다는 사실에서 출발한다. 현재 기술영향평가는 과학기술기본법에 따라 우리 사회에 5~10년 내에 큰 영향을 끼칠 중간 수준 기술을 대상으로 그 영향을 '긍정적', '부정적'으로 나누어 평가하고 긍정적 영향은 극대화하고 부정적 영향을 최소화하는 정책방안을 마련하는 방법론을 채택하고 있다. 하지만 이런 방법론은 첨단기술처럼 그 영향의 성격을 어떻게 규정할 것인지 자체가 논쟁적인 기술에 대해서는 실질적으로 잘 작동하지 않는다. 쉬운 예를 들면 정보통신기술의 광범위한 사용으로 사람들이 책을 점점 덜 읽거나 적어도 종이책 대신 전자책을 더 많이 읽게 되는 현상은 '부정적' 영향인가? 정보통신기술의 보편적 보급 이전의 사회적 기준에 따르자면 부정적이겠지만 정보통신기술 기반사회의 새로운 핵심역량 기준에서 보면 '긍정적'일 수 있다. 실제 이 쟁점은 학계나 시민사회에서 세계적으로 활발하게 논의되고

있는 주제이고 관련 법 제정에서 이 논의 결과가 반영되는 '실천적' 주제이기도 하다.

그러므로 우리는 인공지능 기술이 끼치는 영향을 정확히 어떻게 규정하고 어떻게 측정하며, 그 결과를 어떤 방식으로 다른 나라와 공유하고 역으로 다른 나라의 측정 결과로부터 어떻게 배울 것인지에 대해 제도적인 노력을 기울일 필요가 있다. 이 노력이 담보되어야 인공지능 관련 윤리적 고려의 결과가 실질적으로 사회에 반영될 수 있기 때문이다. 또한 정부간 협의 과정에서 한국 정부가 의견을 내면서 자국의 이해와 관련된 사안에만 의견을 낸다는 인상을 주지 않으려면 이런 '건설적인' 제안도 포함시키는 것이 정치적으로 필요하다고 판단한다.

참고문헌

고학수 • 정해빈 • 박도현, 「인공지능과 차별」, 「저스티스」 통권 제 171호, 2019, 1992~77쪽.

오요한 • 홍성욱, 「인공지능 알고리즘은 사람을 차별하는가?」, 「과학기술학연구」 18(3), 2018, 153 ~ 215쪽.

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

이상욱, 「인간, 낯선 인공지능과 마주하다」, 이중원 엮음, 「인공지능의 존재론」, 한울, 2018, 285 ~ 316쪽.

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

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

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

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

Shanahan, M., The Technological Singularity, Cambridge, MA: The MIT Press, 2015.

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

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

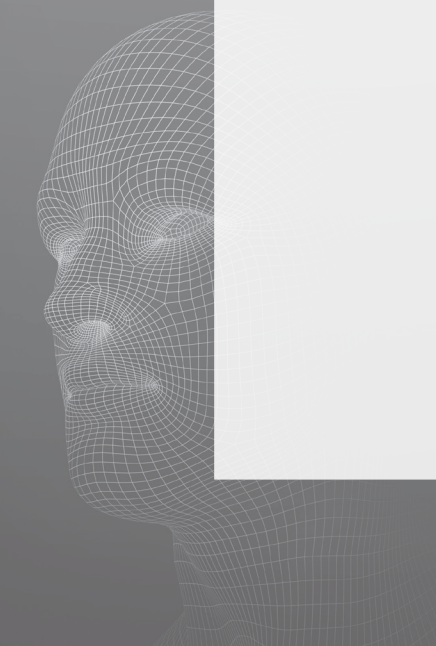
Tegmark, M., Life 3.0: Being Human in the Age of Artificial Intelligence, New York: Penguin Books, 2016.

부록

유네스코 인공지능 윤리 권고(초안)

FIRST VERSION OF A DRAFT TEXT OF
A RECOMMENDATION ON THE ETHICS OF
ARTIFICIAL INTELLIGENCE

인공지능 윤리 권고안 첫 번째 초안



PREAMBLE
전문

The General Conference of the United Nations Educational, Scientific and Cultural Organization (UNESCO), meeting in Paris from xx to xx, at its xx session,

xx부터 xx까지 파리에서 열린 제 xx차 유엔 교육과학문화기구(UNESCO; 이하 유네스코) 총회는

Recalling that, by the terms of its Constitution, UNESCO seeks to construct the defences of peace in the minds of human beings and aims to promote cooperation among the nations through education, science, 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 헌장 조항에 기반하여, 정의에 대한, 법치에 대한, 그리고 전세계 사람들에게 보장된 인권 및 근본적 자유에 대한 보편적

존중을 강화하기 위해, 인류의 마음에 평화의 방벽을 구축하고자 노력하고 교육, 과학, 문화, 정보통신을 통한 국가간 협력을 증진하는 것을 목적으로 한다는 사실을 상기하면서,

Reflecting on the profound influence that Artificial Intelligence (AI) may have on societies, ecosystems, and human lives, including the human mind, in part because of the new ways in which it influences human thinking and decision-making, and affects education, science, culture, and communication and information,

인공지능(AI)이 인간의 사고 및 의사 결정에 영향을 주며 교육, 과학, 문화, 정보통신에 작용하는 새로운 방식으로 인하여, 인공지능이 사회, 생태계, 인간의 마음을 비롯한 인간의 삶에 미치는 심오한 영향을 고찰하며,

Considering that AI systems can be of great service to humanity but also raise fundamental ethical concerns, for instance regarding the biases they can embed and exacerbate, potentially resulting in inequality, exclusion and a threat to cultural and social diversity and gender equality; 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 privacy, freedom of speech, social, economic and political processes, and the environment,

인공지능 기술은 인류에게 큰 도움이 되기도 하지만, 이로 인해 발생·악화될 수 있는 편향이 불평등, 배제, 문화·사회적 다양성 및 성평등에 대한 위협을 초래할 잠재적 가능성, 또한 알고리즘의 작동 및 그 학습에 사용된 데이터의 투명성 및 이해가능성에 대한 필요성, 인공지능이 프라이버시, 표현의 자유, 사회·경제·정치적 과정, 환경에 미치는 잠재적 영향과 같은 근본적 윤리 문제를 제기하기도 한다는 점을 고려하고,

Recognizing that the development of AI can deepen existing divides and inequalities in the world, and that no one should be left behind who does not want to, either in enjoying the benefits of AI or in the protection against its negative implications, while recognizing the different circumstances of different countries,

인공지능의 발전이 전세계에 존재하는 격차 및 불평등을 심화시킬 수 있음과 그 누구도 인공지능 기술의 혜택을 누리거나 인공지능 기술의 부정적인 영향으로부터 보호받는 것에서 배제되길 원치 않는다면 배제되지 않아야 함을 인지하며, 동시에 다양한 국가의 다양

한 상황을 인지하고,

Conscious of the fact that low and middle income countries (LMICs), including but not limited to those in Africa, Latin America and the Caribbean, and Central Asia, as well as Small Island Developing States, are facing an acceleration of the use of information technologies and AI and that the digital economy presents important societal challenges and opportunities for creative societies, requiring the recognition of endogenous cultures, values and knowledge in order to develop economies,

아프리카, 라틴아메리카 및 카리브해, 중앙아시아에 있는 국가와 더불어 군소도서개발도상국(SIDS)까지 포함하여 (단, 이에 국한되지 않는) 중·저소득국가(LMIC)가 정보기술 및 인공지능 사용의 가속화를 직면하고 있다는 점, 그리고 디지털 경제는 창조적 사회로 가기 위한 중요한 사회적 과제 및 기회를 제공한다는 점을 의식하면서, (디지털) 경제를 개발하기 위한 내생적 문화·가치·지식의 인식을 요구하고,

Recognizing that AI has the potential to be beneficial to the environment, via its roles in ecological and climate research, disaster risk management, and agriculture, but that

for those benefits to be realized, fair access to the technology is required and the potential benefits need to be balanced against the environmental impact of the entire AI and information technology production cycle,

인공지능은 생태계·기후 연구, 재난위기관리, 농업에서의 그 역할을 통해 환경에 이로움을 줄 수 있는 잠재력이 있지만, 이러한 이로움이 실현되기 위해서는 공정한 접근이 필요하고 잠재적 이익이 인공지능 및 정보 기술 생산주기의 환경에 대한 영향과 균형을 이루어야 한다는 점을 인지하며,

Noting that addressing risks and ethical concerns should not hamper innovation but rather stimulate new practices of responsible research and innovation in which the research, design, development, deployment, and use of AI is anchored in moral values and ethical reflection,

위험성 및 윤리적 문제를 해결하는 것이 혁신을 방해해서는 안되며, 오히려 인공지능의 연구 · 설계 · 개발 · 보급 · 사용이 도덕 · 윤리적 성찰 위에 뿌리 내린 책임 있는 연구 및 혁신이라는 새로운 관행을 고무시켜야 한다는 점에 주목하고,

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, 사무 총장에게 2021년 제 41차 총회 때 제출하도록 "인공지능 윤리에 관한 국제 기준 수립 도구를 권고의 형태로 준비할" 권한을 부여했다는 점을 상기하면서,

Convinced that the standard-setting instrument presented here, based on a global normative approach, and focusing on human dignity and human rights, including diversity, interconnectedness, inclusiveness and fairness, can guide the research, design, development, deployment, and use of AI in a responsible direction,

전지구적 규범적 접근에 기반하여 본 안에서 제시하는 기준 수립 방식과 다양성, 상호연결성, 포용성, 공정성을 비롯하여 인간의 존엄성 및 인권에 초점을 맞추는 것이 인공지능 연구 · 설계 · 개발 · 보급 · 사용을 책임 있는 방향으로 인도할 수 있다는 점을 확신하며,

Observing that a normative framework for AI and its social implications finds itself at the intersection of ethics, human rights, international and national legal frameworks, and the freedom of research and innovation, and human well-being,

인공지능의 규범적 프레임워크와 그 사회적 함의는 윤리, 인권, 국제·국가 법률 프레임워크, 연구·혁신의 자유, 인류 복지의 교차점에 있다는 점을 주시하며,

Recognizing that ethical values and principles are not necessarily legal norms in and of themselves, but can powerfully shape the development and implementation of policy measures and legal norms, by providing guidance where the ambit of norms is unclear or where such norms are not yet in place due to the fast pace of technological development combined with the relatively slower pace of policy responses,

윤리적 가치 및 원칙은 반드시 법적 규범은 아니지만, 규범의 범위가 불분명하거나 빠른 기술 개발 속도에 비해 상대적으로 느린 정책 대응 속도로 인해 규범이 아직 확립되지 않은 경우에 지침을 제공함으로써 정책 조치 및 법적 규범의 개발 및 구현을 강력하게 형

성할 수 있음을 인지하고,

Convinced that globally accepted ethical standards can play a helpful role in harmonizing AI-related legal norms across the globe, and responsible application of existing international law, if this application is in line with ethical frameworks and does not cause harm locally,

전지구적으로 용인된 윤리 기준이 전지구적으로 인공지능 관련 법적 규범을 조화롭게 하는 데에 유용한 역할을 수행할 수 있음을 확신하고, 또한 기존 국제법의 책임 있는 적용이 윤리적 프레임워크에 부합하고 어떤 곳에도 피해가 되지 않는다면 이 또한 그러한 역할을 수행할 수 있음을 확신하면서,

Recognizing the Universal Declaration of Human Rights (1948), including Article 27 emphasizing the right to share in scientific advancement and its benefits; the instruments of the international human rights framework, including the United Nations Convention on the Elimination of All Forms of Discrimination against Women (1979), the United Nations Convention on the Rights of the Child (1989), and the United Nations Convention on the Rights of Persons with Disabilities (2006); the UNESCO Convention on the Protec-

tion and Promotion of the Diversity of Cultural Expressions (2005),

과학 진보 및 그 혜택을 공유한 권리를 강조하는 제27조를 비롯한 '세계인권선언'(1948), 그리고 '모든 형태의 여성 차별 철폐에 관한 유엔 협약'(1979), '유엔 아동 권리 협약'(1989), '유엔 장애인 권리 협약'(2006), '문화적 표현의 다양성 보호와 증진에 관한 유네스코 협약'(2005)을 비롯하여 국제 인권 프레임워크의 도구들을 인지하며,

Noting the UNESCO Declaration on the Responsibilities of the Present Generations Towards Future Generations (1997); the United Nations Declaration on the Rights of Indigenous Peoples (2007); the Report of the United Nations Secretary-General on the Follow-up to the Second World Assembly on Ageing (A/66/173) of 2011, focusing on the situation of the human rights of older persons; the Report of the Special Representative of the United Nations Secretary-General on the issue of human rights and transnational corporations and other business enterprises (A/HRC/17/31) of 2011, outlining the 'Guiding Principles on Business and Human Rights: Implementing United Nations "Protect, Respect and Remedy" Framework'; the Human Rights Coun-

cil's resolution on 'The right to privacy in the digital age' (A/HRC/RES/42/15) adopted on 26 September 2019; the UNESCO Recommendation on Science and Scientific Researchers (2017); the UNESCO Internet Universality Indicators (2019), including the R.O.A.M. principles; the Report of the United Nations Secretary-General's High-level Panel on Digital Cooperation on 'The Age of Digital Interdependence' (2019); and the outcomes and reports of the ITU's AI for Good Global Summits,

'미래 세대에 대한 현재 세대의 책임에 관한 유네스코 선언'(1997); '선주민의 권리에 관한 유엔 선언문'(2007), 고령자의 인권 상황을 다룬 2011년의 '제 2세계 고령화 후속조치에 대한 사무총장 보고서'(A/66/173), '기업과 인권에 대한 지침 원칙: 유엔의 "보호, 존중 및 구제" 프레임워크에 대한 이행'의 윤곽을 그린 2011년의 '인권과 다국적 기업 및 기타 비즈니스 기업 문제에 대한 사무총장 특별대표의 보고서'(A/HRC/17/31), 2019년 9월 26일에 채택된 '"디지털 시대의 개인정보보호권리"에 대한 인권이사회 결의안'(A/HRC/RES/42/15), '과학 및 과학연구자에 관한 유네스코 권고'(2017), 'R.O.A.M. 원칙'을 비롯한 '유네스코 인터넷 보편성 지표'(2019), '"디지털 상호 의존 시대"에 대한 유엔 사무총장의 디지털 협력에 관한 고위급 패널 보고서'(2019), 국제전기통신연합(ITU)의 '착한 인공지능 세계회의'의 결과 및 보고서에 주목하고,

Noting also existing frameworks related to the ethics of AI of other intergovernmental organizations, such as the relevant human rights and other legal instruments adopted by the Council of Europe, and the work of its Ad Hoc Committee on AI (CAHAI); the work of the European Union related to AI, and of the European Commission's High-Level Expert Group on AI, including the Ethical Guidelines for Trustworthy AI; the work of the OECD Expert Group on AI (AIGO), and the OECD's Recommendation of the Council on AI; the G20 AI Principles, drawn therefrom, and outlined in the G20 Ministerial Statement on Trade and Digital Economy; the G7's Charlevoix Common Vision for the Future of AI; the work of the African Union's Working Group on AI; and the work of the Arab League's Working Group on AI,

또한 유럽평의회에서 채택한 유관 인권 및 기타 법적 도구, 인공지능 특별위원회(CAHAI)의 작업, 인공지능과 관련한 유럽연합(EU)의 작업 및 '신뢰할 수 있는 인공지능에 대한 윤리 지침'을 비롯한 유럽집행위원회(EC)의 인공지능 고위급 전문가 그룹(AI HLEG)의 작업, 경제협력개발기구(OECD)의 인공지능 전문가 그룹(AIGO)의 작업 및 '인공지능에 관한 OECD 이사회 권고안', 'G20 무역 및 디지털경제 장관선언문'에서 윤곽이 드러난 'G20 인공지능 원칙'(G20 AI Principles), G7의 '인공지능의 미래를 위한 샤를부아

공동비전', 아프리카연합(AU) 인공지능실무단(AIWG)의 작업, 아랍연맹(Arab League) 인공지능실무단의 작업과 같은 기타 정부간 국제기구의 인공지능 윤리와 관련된 기존 프레임워크에 주목하며,

Emphasizing that specific attention must be paid to LMICs, including but not limited to those in Africa, Latin America and the Caribbean, and Central Asia, as well as Small Island Developing States, as they have been underrepresented in the AI ethics debate, which raises concerns about neglecting local knowledge, cultural and ethical pluralism, value systems and the demands of global fairness,

아프리카, 라틴아메리카 및 카리브해, 중앙아시아에 있는 국가와 더불어 군소도서개발도상국(SIDS)까지 포함하여 (단, 이에 국한되지 않는) 중·저소득국가(LMIC)가 인공지능 윤리 논의에서 발언권이 미약하고 이는 토착 지식, 문화·윤리적 다원주의, 가치 시스템 및 전 지구적 공정성이 무시된다는 우려를 불러 일으키기 때문에 이들 국가에 특별한 관심이 필요함을 강조하고,

Conscious of the many national frameworks related to the ethics and regulation of AI,

인공지능 윤리 및 규제와 관련된 많은 국가적 프레임워크를 의식하며,

Conscious as well of the many initiatives and frameworks related to the ethics of AI developed by the private sector, professional organizations, and non-governmental organizations, such as the IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems and its work on Ethically Aligned Design; the World Economic Forum's 'Global Technology Governance: A Multistakeholder Approach'; the UNI Global Union's 'Top 10 Principles for Ethical Artificial Intelligence'; the Montreal Declaration for a Responsible Development of AI; the Harmonious Artificial Intelligence Principles (HAIP); and the Tenets of the Partnership on AI,

또한 전기전자기술자협회(IEEE)의 '자율·지능시스템 윤리에 관한 전 지구적 계획안' 및 '윤리적 설계'(EAD)에 대한 작업, 세계경제포럼(WEF)의 '글로벌 기술 거버넌스: 다자간 접근' 정책 보고서, 국제사무금융서비스노동조합연맹(UNI) 글로벌 유니온(Global Union)의 '윤리적 인공지능 10대 원칙', '책임있는 인공지능 개발을 위한 몬트리올 선언', '조화로운 인공지능 원칙'(HAIP), 인공지능 파트너십(Partnership on AI)의 '교의'(Tenets)와 같이 민간 부문, 전문가 단체, 비정부 기구(NGO)에서 개발한 AI 윤리와 관련된 많은 계획안 및 프레임워크를 의식하고,

Convinced that AI can bring important benefits, but that achieving them can also be under tension of innovation debt, asymmetric access to knowledge, barriers of rights to information and gaps in capacity of creativity in developing cycles, human and institutional capacities, barriers to access technological innovation, and a lack of adequate infrastructure and regulations regarding data,

AI 기술이 중요한 혜택을 가져올 수도 있지만, 이를 달성하는 것은 혁신 비용, 비대칭적 지식 접근, 정보권의 장애물, 그리고 개발 주기에서의 창의적 역량의 격차, 인간 및 제도적 역량의 격차, 기술 혁신 접근성에 대한 장애물, 데이터와 관련한 적합한 인프라 및 규제의 부족으로 인한 긴장 하에 있음 또한 확신하기에,

Recognising that economic competition is taking place within and between states and also between multinational companies, potentially causing AI strategies and regulatory frameworks to be focused on national and commercial interests, while global cooperation is needed to address the challenges that AI brings in a diversity and interconnectivity of cultures and ethical systems, and to mitigate potential misuse,

국가 간 및 다국적 기업 간의 경제적 경쟁으로 인해 잠재적으로 인공지능 전략 및 규제 프레임워크가 국가·상업적 이익에 초점을 맞추게 되는 상황에서, 인공지능이 문화·윤리 시스템의 다양성 및 상호연결성에 초래하는 어려움을 다루고 잠재적 오용을 완화하기 위해 전지구적 협력이 필요함을 강조하고,

Taking fully into account that the rapid development of AI systems encounters barriers to understand and implement AI, because of the diversity of ethical orientations and cultures around the World, the lack of agility of the law in relation to technology and the information society, and the risk that local and regional ethical standards and values be disrupted by AI,

전세계의 윤리적 지향 및 문화의 다양성, 기술과 정보 사회와 관련한 법률의 민첩성 부족, 토착·지역적 윤리 기준 및 가치가 인공지능에 의해 지장을 받을 위험성으로 인해, 인공지능 기술의 급속한 발전은 인공지능의 이해 및 구현을 가로막는 걸림돌이 된다는 점을 충분히 고려하여,

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

1. 인공지능 윤리에 관한 현 권고안을 채택하고,

2. Recommends that Member States apply 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;

2. 회원국은 각 국의 헌법 관례 및 통치 구조에 적합한 입법 또는 기타 조치 요구를 비롯한 적절한 조치를 취함으로써 본 권고안의 조항을 적용하여, 관할 구역 내에서 본 권고안의 원칙 및 규범이 효력을 발휘할 수 있도록 권고하며,

3. Also recommends that Member States bring the Recommendation to the attention of the authorities, bodies, institutions and organizations in public, commercial and non-commercial sectors involved in the research, design, development, deployment, and use of AI systems.

3. 또한 회원국이 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용과 관련된 공공·영리·비영리 부문의 공공기관, 단체, 기관 및 조직이 본 권고안에 주목할 수 있도록 할 것을 권고한다.

I SCOPE OF APPLICATION
적용 범위

1. This Recommendation addresses ethical issues related to AI. It approaches AI ethics as a holistic framework of interdependent values, principles and actions that can guide societies in the AI system lifecycle, referring to human dignity and well-being as a compass to deal responsibly with the known and unknown impacts of AI systems in their interactions with human beings and their environment. The AI system lifecycle refers to the research, design, development, deployment, and use of AI systems, and the use of AI systems can be understood to include the maintenance, operation, end-of-use, and disassembly of AI systems. It is not within the ambition of this instrument to provide one single definition of AI, since such a definition would need to change over time, in accordance with technological developments. Rather, its ambition is to address those features of AI systems that are of central ethical relevance and on which there is large international con-

sensus. For the purposes of this Recommendation, AI systems can be approached as technological systems which have the capacity to process information in a way that resembles intelligent behaviour, and typically includes aspects of learning, perception, prediction, planning or control. This Recommendation approaches AI systems along the following lines:

1. 본 권고안은 인공지능과 관련된 윤리적 사안을 다룬다. 본 안은 인공지능 시스템이 인류 및 환경에 미칠 수 있는 알려진 영향과 알려지지 않은 영향을 충실히 다루는 데에 있어 인간 존엄성과 복지를 나침반으로 여김으로써, 인공지능 윤리를 인공지능 시스템 수명 주기 동안 인간 사회의 지침이 될 수 있는 독립적인 가치·원칙·조치의 총체적·발전적 프레임워크로서 접근한다. 인공지능 시스템 수명 주기는 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용을 가리키며, 인공지능 시스템의 사용은 이의 관리, 운영, 사용 종료, 해체를 모두 포함하는 것으로 이해될 수 있다. 인공지능의 정의는 기술의 진보에 따라 변화될 필요가 있기에, 인공지능의 단일 정의를 제공하는 것은 본 권고안의 뜻에 맞지 않는다. 다만, 본 권고안은 윤리와 핵심적으로 관련되어 있고 충분한 국제적 합의가 이루어진 인공지능 시스템의 특징들을 다루는 데에 목적이 있다. 본 권고안의 목적에 따라서, 인공지능 시스템은 일반적으로 학습, 추론, 인식,

예측, 계획, 통제를 비롯하여 지적 행위와 유사한 방식으로 정보를 처리할 능력이 있는 기술적 시스템이라는 측면에서 접근될 수 있다. 본 권고안은 인공지능 시스템을 접근하는 방식에 있어 아래의 내용을 따른다.

a. First of all, AI systems embody models and algorithms that produce a capacity to learn and to perform cognitive tasks, like making recommendations and decisions in real and virtual environments. AI systems are designed to operate with varying levels of autonomy by means of knowledge modeling and representation and by exploiting data and calculating correlations. AI systems may include several approaches and technologies, such as but not limited to:

a. 첫 번째로, 인공지능 시스템은 현실 및 가상 환경에서의 추천 및 의사 결정과 같은 인지 과제의 학습 및 수행 능력을 생성하는 모델과 알고리즘을 구현한다. 인공지능 시스템은 지식 모델링 및 표현을 통해, 또 데이터 및 상관관계의 계산을 활용하여, 서로 다른 수준의 자율성을 가지고 동작하도록 설계되어 있다. 인공지능 시스템에는 아래의 몇 가지 항목들과 같이 (단, 이에 국한되지 않는) 몇 가지 접근법이 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, search, and optimization, and

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

iii. cyber-physical systems, including internet-of-things and robotics, which involve control, perception, the processing of data collected by sensors, and the operation of actuators in the environment in which AI systems work.

iii. 인공지능 시스템이 작동하는 환경에서의 제어, 인지, 센서 데이터 처리, 액추에이터 조작이 수반된 사물 인터넷(IoT) 및 로봇 공학을 비롯한 가상물리시스템.

b. Second, besides raising ethical issues similar to the ones raised by any technology, AI systems also raise new types of issues. Some of these issues are related to the fact that AI systems are capable of doing things which previously only living beings could do, and which were in some cases even limited to human beings only. These characteristics give AI systems a profound, new role in human practices and society. Going even further, in the long term, AI systems could challenge human's special sense of experience and consciousness, raising additional concerns about human autonomy, worth and dignity, but this is not yet the case.

b. 두 번째로, 인공지능 시스템은 여느 과학기술에서 제기되었던 것과 같은 윤리적 문제뿐만 아니라, 또한 새로운 유형의 문제를 제기한다. 이 중 일부 사안들은 과거에는 생명체만이 할 수 있었거나 인간에게만 국한되었던 과제를 인공지능 시스템이 수행할 능력을 갖추게 되었다는 사실과 관련이 있다. 이러한 특징들은 인간 관습 및 사회에서 인공지능 시스템에게 심오하고 새로운 역할을 부여한다. 장기적 관점으로 보면, 물론 아직 그렇지는 않지만, 인공지능 시스템은 인간 고유의 경험과 의식에 도전하여, 인간의 자율성, 가치, 존엄성에 대한 추가적인 우

려를 낳을 수도 있다.

c. Third, even though ethical questions regarding AI are generally related to the concrete impact of AI systems on human beings and societies, another set of ethical issues is directed at the interactions between AI systems and human beings and its implications for our understanding of both human beings and technologies. This Recommendation acknowledges that both types of questions are closely related and are necessary elements of an ethical approach to AI.

c. 세 번째로, 인공지능에 관한 윤리적 질문은 일반적으로 인공지능 시스템이 인간 및 사회에 미치는 구체적인 영향과 관련되어 있지만, 또 다른 몇몇 윤리적 문제는 인공지능 시스템과 인간 간의 상호작용 및 이것이 우리가 인간과 과학기술을 이해하는 방식에 미치는 영향에 관한 것이다. 본 권고안은 이 두 가지의 질문이 서로 긴밀하게 연결되어 있으며 인공지능에 대한 윤리적 접근에 있어서 필요한 부분이라는 점을 인정한다.

2. This Recommendation pays specific attention to the broader ethical implications of AI in relation to the central domains of UNESCO: education, science, cul-

ture, 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):

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

a. AI systems are connected to education in many ways: they challenge the societal role of education because of their implications for the labour market and employability; they might have impact on educational practices; and they require that education of AI engineers and computer scientists creates awareness of the societal and ethical implications of AI.

a. 인공지능 시스템은 다양한 방식으로 교육과 연결되어 있다. 인공지능 시스템은 노동시장과 고용에 미치는 파급효과로 인해 교육의 사회적 역할에 도전을 제기한다. 인공지능 시스템은 교육의 관행에도 영향을 미칠 수 있다. 또한 인공지능 시스템으

로 인해 인공지능 기술자 및 컴퓨터 과학자 교육은 인공지능의 사회적 · 윤리적 파급효과에 대한 인식을 형성해줄 수 있어야 한다.

b. In all fields of the sciences, social sciences and humanities, AI has implications for our concepts of scientific understanding and explanation, and for the ways in which scientific knowledge can be applied as a basis for decision-making.

b. 모든 분야의 과학, 사회과학, 인문학에서, 인공지능은 과학적 이해 및 설명에 대한 우리의 개념에 영향을 미치며 과학 지식이 의사 결정을 위한 토대로서 적용될 수 있는 방법에도 영향을 미친다.

c. AI has implications for cultural identity and diversity. It has the potential to positively impact the cultural and creative industries, but it may also lead to an increased concentration of supply of cultural content, data and income in the hands of only a few actors, with potential negative implications for the diversity of cultural expressions and equality.

c. 인공지능은 문화 정체성 및 다양성에 파급효과를 가진다. 인공지능은 문화 및 창조적 산업에 긍정적 영향을 줄 가능성도 있지만, 또한 문화 컨텐츠, 데이터, 수익의 공급이 소수의 종사자들에게만 편중되게 하여 문화적 표현의 다양성 및 평등에 부정적 영향을 줄 수도 있다.

d. In the field of communication and information, machine-powered translation of languages is likely to play an increasingly important role. This might have a substantial impact on language and human expression, in all dimensions of life, bringing a responsibility to deal carefully with human languages and their diversity. Moreover, AI is challenging practices of journalism, and the social role of journalists, media workers, and social media producers who are engaged in journalistic activities, and is connected to both the spreading and the detection of disinformation or misunderstanding.

d. 정보통신 분야에서 언어의 기계 번역은 점차 중요한 역할을 할 것이다. 이는 일상의 모든 방면에서 언어와 인간의 표현에 상당한 영향을 미칠 것이며, 이로써 인간의 언어 및 그 다양성을 주의 깊게 다룰 책임을 불러일으킨다. 더욱이 인공지능은 언론의 관행에 도전하고 있고, 언론 활동에 종사하며 허위 정

보 및 사실 왜곡의 유포와 발견 모두에 연관이 되어있는 기자, 미디어 종사자, 소셜미디어 생산자의 사회적 역할에도 도전하고 있다.

3. This Recommendation is addressed to States. As appropriate and relevant, it also provides guidance to decisions or practices of individuals, groups, communities, institutions and corporations, public and private, particularly AI actors, understood as those who play an active role in the AI system lifecycle, including organizations and individuals that research, design, develop, deploy, or use AI.

3. 본 권고안은 국가를 대상으로 한다. 적절하고 관련된 경우, 본 권고안은 또한 개인, 집단, 공동체, 민간 · 공공 기관 및 기업, 특히 인공지능을 연구 · 설계 · 개발 · 보급 · 사용하는 조직 및 개인을 비롯하여, 인공지능 시스템 수명 주기에서 적극적인 역할을 수행하는 인공지능 행위 주체의 의사 결정 또는 행동에 대한 지침을 제공한다.

II AIMS AND OBJECTIVES
목적 및 목표

4. This Recommendation aims for the formulation of ethical values, principles and policy recommendations for the research, design, development, deployment and usage of AI, to make AI systems work for the good of humanity, individuals, societies, and the environment.

4. 본 권고안은 인공지능 연구 · 설계 · 개발 · 보급 · 사용에 필요한 윤리적 가치, 원칙, 정책 권고사항을 수립하여, 인공지능 시스템이 인류, 개인, 사회, 환경의 이익을 위해 작동하게 하는 것을 목적으로 한다.

5. The complexity of the ethical issues surrounding AI requires equally complex responses that necessitate the cooperation of multiple stakeholders across the various levels and sectors of the international, regional and national communities.

5. 인공지능을 둘러싼 윤리적 문제의 복잡성은 다양한 층위·부문

의 국제·지역·국가 공동체에 존재하는 많은 이해관계자의 협력이 수반되어야 하는 복잡한 대응을 요구한다.

6. Even though this Recommendation is addressed primarily to policy-makers in and outside UNESCO Member States, it also aims to provide a framework for international organizations, national and transnational corporations, NGO's, engineers and scientists, including representatives of humanities, natural and social sciences, non-governmental organizations, religious organizations, and civil society, stimulating a multi-stakeholder approach, grounded in a globally accepted ethical framework that enables stakeholders to collaborate and take common responsibility based on a global, intercultural dialogue.

6. 본 권고안은 유네스코 회원국 안팎의 정책 입안자들을 대상으로 하지만, 또한 인문학 · 자연과학 · 사회과학, 비정부기구, 종교 조직, 시민 사회의 대표자들을 비롯하여 국제 기구, 국내 및 다국적 기업, 비정부기구, 공학자, 과학자에게 프레임워크를 제공하는 것을 목적으로 함으로써, 이해관계자들이 전 지구적 · 문화 간 대화를 통하여 협력하고 공통의 책임을 질 수 있도록 하는 전지구적으로 용인되는 윤리 프레임워크에 기반한 다자적 접근도 촉진한다.

III VALUES AND PRINCIPLES
가치와 원칙

7. Values and principles are not necessarily legal norms in and of themselves, as stated in the preamble to this Recommendation. They play a powerful role in shaping policy measures and legal norms, because values encompass internationally agreed expectations of what is good and what is to be preserved. As such, values underpin principles.

7. 가치와 원칙은 본 권고안의 전문에 명시된 바와 같이, 반드시 그 자체가 법적 규범이거나 법적 규범에만 관련된 것은 아니다. 가치는 무엇이 좋은지와 무엇이 보존되어야 하는지에 대한 국제적 합의를 아우르기 때문에, 가치와 원칙은 정책 조치와 법적 규범을 형성하는 데 있어 강력한 역할을 한다. 따라서 가치는 원칙의 토대가 된다.

8. Values thus inspire good moral behaviour in line with the international community's understanding of such behaviour and they are the foundations of principles,

while principles unpack the values underlying them more concretely so that values can be more easily actualised in policy statements and actions.

8. 따라서 가치는 좋은 도덕적 행동이 무엇인지에 대한 국제 사회의 이해에 맞게 그러한 행동을 고취시키며 원칙의 토대가 되는 반면, 원칙은 그 기저에 있는 가치를 구체적으로 풀어내어 가치가 정책 제시 · 조치 과정에서 더 쉽게 실현될 수 있도록 한다.

III.1. VALUES
III.1. 가치

Human dignity
인간 존엄성

9. The research, design, development, deployment, and use of AI systems should respect and preserve human dignity. The dignity of every human person is a value that constitutes a foundation for all human rights and fundamental freedoms and is essential when developing and adapting AI systems. Human dignity relates to

the recognition of the intrinsic worth of each individual human being and thus dignity is not tied to national origin, legal status, socio-economic position, gender and sexual orientation, religion, language, ethnic origin, political ideology or other opinion.

9. 인공지능의 연구 · 설계 · 개발 · 보급 · 사용은 인간 존엄성을 존중하고 보호해야 한다. 모든 인간의 존엄성은 모든 인권 및 근본적 자유의 토대를 구성하며, 인공지능 시스템을 개발하고 적용하는 과정에서 매우 중요하다. 인간 존엄성은 인간 개개인의 내재적 가치를 인정함과 관련 있기에, 존엄성은 국적, 법적 상태, 사회 · 경제적 지위, 성 및 성적 지향, 종교, 언어, 인종적 태생, 정치적 이념 또는 기타 견해에 구속되지 않는다.

10. This value should be respected by all actors involved in the research, design, development, deployment, and use of AI systems in the first place; and in the second place, be promoted through new legislation, through governance initiatives, through good exemplars of collaborative AI development and use, or through government-issued national and international technical and methodological guidelines as AI technologies advance.

10. 첫 번째로, 이 가치는 인공지능 연구 · 설계 · 개발 · 보급 · 사용에 관련된 모든 행위 주체에게 존중 받아야한다. 두 번째로, 이 가치는 새 법안, 거버넌스 계획, 인공지능 공동 개발 · 사용의 모범, 또는 인공지능 기술 발전에 대하여 정부가 발행하는 국내 및 국제 기술 · 방법론 지침을 통해 제고되어야 한다.

Human rights and fundamental freedoms
인권과 근본적 자유

11. The value of the respect for, and protection and promotion of, human rights and fundamental freedoms in the AI context means that the research, design, development, deployment, and use of AI systems should be consistent and compliant with international human rights law, principles and standards.

11. 인공지능이라는 맥락에서 인권 및 근본적 자유에 대한 존중 · 보호 · 증진의 가치는 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용이 국제 인권법 · 원칙 · 기준에 부합하고 이를 준수해야 함을 의미한다.

Leaving no one behind
누구도 소외되지 않음

12. It is vital to ensure that AI systems are researched, designed, developed, deployed, and used in a way that respects all groupings of humanity and fosters creativity in all its diversity. Discrimination and bias, digital and knowledge divides and global inequalities need to be addressed throughout an AI system lifecycle.

12. 인공지능 시스템이 모든 인류 집단을 존중하고 그 모든 다양성 안에서 창의성을 촉진하는 방식으로 연구 · 설계 · 개발 · 보급 · 사용되도록 하는 것이 중요하다. 차별 및 편향, 디지털 · 지식 격차, 전지구적 불평등은 인공지능 시스템 수명 주기 전 영역에서 해결되어야 한다.

13. Thus, the research, design, development, deployment, and use of AI systems must be compatible with empowering all humans, taking into consideration the specific needs of different age groups, cultural systems, persons with disabilities, women and girls, disadvantaged, marginalized and vulnerable populations; and should not be used to restrict the scope of

lifestyle choices or personal experiences, including the optional use of AI systems. Furthermore, efforts should be made to overcome the lack of necessary technological infrastructure, education and skills, as well as legal frameworks, particularly in low- and middle-income countries.

13. 따라서, 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용은 다양한 연령 집단, 문화적 체제, 장애인, 여성 및 소녀, 빈민 · 소외 · 취약 계층의 특정 필요를 고려함으로써 모든 인류의 역량을 강화하는 것과 양립되어야 한다. 그리고 이는 인공지능 시스템의 선택적 사용을 비롯한 생활 방식 또는 개인적 경험의 범주를 제한하기 위해 사용되어서는 안된다. 더욱이, 특히 중 · 저소득국가에서 필요한 기술 인프라 · 교육 · 숙련 및 법적 프레임워크의 부족을 극복하기 위한 노력이 이루어져야 한다.

Living in harmony
조화로운 삶

14. The value of living in harmony points to the research, design, development, deployment, and use of AI systems recognising the interconnectedness of all hu-

mans. The notion of being interconnected is based on the knowledge that every human belongs to a greater whole, which is diminished when others are diminished in any way.

14. 조화로운 삶의 가치는 모든 인간의 상호연결성을 인지하고 있는 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용에 있다. 상호연결성이라는 개념은 인간이 모든 사람이 더 큰 하나의 완전체에 속하게 된다는 지식에 기반을 두고 있으며, 이는 다른 사람이 약화되면 그 자신도 약화됨을 의미한다.

15. This value demands that the research, design, development, deployment, and use of AI systems should avoid conflict and violence, and should not segregate, objectify, or undermine the safety of human beings, divide and turn individuals and groups against each other, or threaten the harmonious coexistence between humans and the natural environment, as this would negatively impact on humankind as a collective. The purpose of this value is to recognise the enabling role that AI actors should play in achieving the goal of living in harmony, which is to ensure a future for common good.

15. 갈등과 폭력이나, 인류의 안전을 분리·대상화·약화시키는 것이나, 개인 및 집단이 서로 분열·반목하는 것이나, 인간과 자연 환경 간의 조화로운 공존을 위협하는 것 모두 공통적으로 인류에 부정적인 영향을 끼칠 것이기에, 이 가치는 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용이 이렇게 되지 않도록 해야 한다. 이 가치의 목적은 공익의 미래를 보장하는 조화로운 삶이라는 목표를 달성하기 위해 인공지능 행위 주체가 수행해야 하는 역할을 인지하는 것이다.

Trustworthiness
신뢰성

16. AI systems should be trustworthy. Trustworthiness is a socio-technical concept implying that the research, design, development, deployment, and use of AI systems should inspire, instead of infringing on, trust among people and in AI systems.

16. 인공지능 시스템은 신뢰할 수 있어야 한다. 신뢰성은 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용이 사람들 중에서, 그리고 인공지능 시스템 내에서 권리 침해가 아닌 신뢰를 불어넣어야 한다는 사회-기술적 개념이다.

17. Trust has to be earned in each use context and more broadly is a benchmark for the social acceptance of AI systems. Therefore people should have good reason to trust that AI technology brings benefits while adequate measures are taken to mitigate risks.

17. 신뢰는 모든 사용 맥락에서 존재해야 하며, 보다 광범위하게는 인공지능 시스템의 사회적 수용을 위한 기준점이다. 따라서, 인공지능 기술이 혜택을 가져다 주며 위험을 완화하기 위한 적절한 조치 또한 취해진다고 믿을 만한 좋은 근거가 있어야 한다.

Protection of the Environment
환경 보호

18. The aim of this value is to ensure that the research, design, development, deployment, and use of AI systems recognise the promotion of environmental well-being. All actors involved during the lifecycle of AI systems should follow relevant international and domestic laws in the field of environmental protection and sustainable development to ensure the minimisa-

tion of climate change risk factors, including carbon emission of AI systems, and prevent the exploitation and depletion of natural resources contributing to the deterioration of the environment.

18. 이 가치의 목적은 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용이 자연의 안녕을 증진하는 것의 중요성을 인식하게 하는 것이다. 인공지능 시스템 수명 주기에 연관된 모든 행위 주체는 인공지능 시스템의 탄소 배출을 비롯한 기후 변화 위험 요인을 최소화하고 환경 파괴에 일조하는 천연 자원의 사용과 소모를 막기 위하여 환경 보호 및 지속가능한 개발 분야에서 유관 국제 · 국내법을 준수해야 한다.

19. At the same time, AI systems should be used to provide solutions to protect the environment and preserve the planet by supporting circular economy type approaches.

19. 동시에 인공지능 시스템은 순환 경제 유형의 접근 방식을 지원함으로써 환경을 보호하고 지구를 보존하기 위한 방침을 제공하는 데 사용되어야 한다.

III.2. PRINCIPLES
III.2. 원칙

20. Bearing in mind that any AI system has a number of essential evolving human and technology dependent situational characteristics, principles are presented in two groups.

20. 모든 인공지능 시스템이 진화하는 인간과 기술이 의존하는 상황적 특징과 관련된 수많은 핵심 요소를 갖는다는 점을 기억하면, 원칙은 두 그룹으로 제시될 수 있다.

21. The first group consists of principles reflecting characteristics that are associated with the human-technology interface, i.e. human-AI systems interaction. Note that the research, design, development, deployment, and use of AI systems influence human agency in two ways: First, in terms of expanding the scope for machine autonomy and decision-making, and second, by influencing the quality of human agency in both positive and negative ways.

21. 첫 번째 그룹은 인간-인공지능 상호작용과 같은 인간-기술 인터페이스와 연관된 특징을 반영하는 원칙으로 구성된다. 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용이 두 가지 방식으로 인간 활동에 영향을 준다는 점을 주목해야 한다. 첫 번째는 기계의 자율성과 의사 결정에 대한 범위 확장이라는 측면이고, 두 번째는 긍정적 방식, 부정적 방식 모두로 인간 활동의 질에 영향을 준다는 측면이다.

22. The second group of principles consists of principles reflecting characteristics associated with the properties of AI systems themselves that are pertinent to ensuring the research, design, development, deployment, and use of AI systems happen in accordance with internationally accepted expectations of ethical behaviour.

22. 원칙의 두 번째 그룹은 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용이 국제적으로 용인된 윤리적 행동 방식에 따라 이루어지도록 적절히 보장하는 인공지능 시스템 그 자체의 속성과 연관된 특징을 반영하는 원칙으로 구성된다.

GROUP 1
그룹 1

For human and flourishing
인간과 번영을 위하여

23. AI systems should be researched, designed, developed, deployed, and used to let humans and the environment in which they live, flourish. Throughout the lifecycle of AI systems the quality of life of every human being should be enhanced and the enjoyment of all human rights for every human being should be promoted, while the definition of 'quality of life' should be left open to individuals or groups, as long as no human being is harmed physically or mentally, or their dignity diminished as a rsult this definition.

23. 인공지능 시스템은 인간과 인간이 살고 있는 환경이 번영하게끔 연구 · 설계 · 개발 · 보급 · 사용되어야 한다. 인공지능 시스템의 수명 주기 전 영역에서 모든 사람의 삶의 질은 향상되어야 하고 각사람은 모든 인권을 더욱이 향유할 수 있어야 하는데, '삶의 질'의 정의는 어떤 사람도 육체적으로 또는 정신적으로 피해를 입지 않거나 이 정의에 의해서 존엄성이 훼손되지 않는 한 개인 또는 집단에게 열려 있어야 한다.

24. AI systems may be researched, designed, developed, deployed or used to assist in interactions involving vulnerable people, including, but not limited to children, the elderly or the ill, but should never objectify humans or undermine human dignity, or violate or abuse human rights.

24. 인공지능 시스템은 아동, 노인, 환자를 포함하는 (단, 이에 국한되지 않는) 취약 계층과 관련된 상호작용을 지원하기 위해 연구 · 설계 · 개발 · 보급 · 사용될 수는 있지만, 결코 인간을 대상화하거나 인간 존엄성을 훼손하거나 인권을 침해 · 남용해서는 안된다.

Proportionality
비례성원칙[17)]

25. The research, design, development, deployment, and use of AI systems may not exceed what is necessary to achieve legitimate aims or objectives and should be appropriate to the context.

17) 참고: https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE02431181 (25번 항목은 이에 대한 설명)

25. 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용은 적법한 목적 또는 목표를 달성하는 데에 필요한 수준을 초과할 수 없으며, 상황에 적합해야 한다.

26. The choice of an AI method should be justified in the following ways: (a) The AI method chosen should be desirable and proportional to achieve a given aim; (b) The AI method chosen should not have an excessive negative infringement on the foundational values captured in this document; (c) The AI method should be appropriate to the context.

26. 인공지능 방법론의 선택은 다음과 같은 방식에 따라 정당성을 가져야 한다. (a) 선택된 인공지능 방법론은 주어진 목표를 달성하기에 바람직하고 비례적이어야 한다. (b) 선택된 인공지능 방법론은 본 안이 지니는 근본적인 가치에 대해 과도한 부정적 침해가 없어야 한다. (c) 인공지능 방법론은 상황에 적합해야 한다.

Human oversight and determination
인간의 감독 및 결정

27. It should always be possible to attribute both ethical

and legal responsibility for the research, design, development, deployment, and use of AI systems to a physical person or to an existing legal entity. Human oversight refers thus not only to individual human oversight, but to public oversight.

27. 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용에 대하여 개인 또는 법인에게 윤리적 · 법적 책임을 묻는 것이 언제든지 가능해야 한다. 따라서 인간의 감독이란 1인 감독뿐만 아니라 대중의 감독까지도 지칭한다.

28. It may be the case that sometimes humans would have to share control with AI systems for reasons of efficacy, but this decision to cede control in limited contexts remains that of humans, as AI systems should be researched, designed, developed, deployed, and used to assist humans in decision-making and acting, but never to replace ultimate human responsibility.

28. 인간이 효율성을 위해 인공지능 시스템과 제어권을 공유해야 할 경우가 있을 수도 있는데, 인공지능 시스템은 의사결정 및 행동에 있어 인간의 궁극적인 책임을 대체하기 위함이 아니라 인간을 보조하기 위해 연구 · 설계 · 개발 · 보급 · 사용되

어야 하기 때문에, 한정된 상황에서 제어권을 양도할지 결정하는 것은 여전히 인간의 몫이다.

Sustainability
지속가능성

29. In the context of promoting the development of sustainable societies, AI actors should respect the social, economic and environmental dimensions of sustainable development of all of humanity and the environment. AI systems should be researched, designed, developed, deployed, and used to promote the achievement of sustainability related to globally accepted frameworks such as the sustainable development goals.

29. 지속가능한 사회의 발전을 촉진하려는 상황에서, 인공지능 행위 주체는 모든 인류 및 환경의 지속가능한 개발에서 사회 · 경제 · 환경적 측면을 존중해야 한다. 인공지능 시스템은 지속가능발전목표(SDG)와 같이 전지구적으로 용인된 프레임워크와 관련된 지속가능성의 달성을 촉진하기 위해 연구 · 설계 · 개발 · 보급 · 사용되어야 한다.

Diversity and inclusiveness
다양성 및 포용성

30. The research, design, development, deployment, and use of AI systems should respect and foster diversity and inclusiveness at a minimum consistent with international human rights law, standards and principles, including demographic, cultural and social diversity and inclusiveness.

30. 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용은 국제법 · 기준 · 원칙과 최소한 일치하는 선에서 인구 · 문화 · 사회적 다양성 및 포용성을 비롯하여 다양성 및 포용성을 존중하고 장려해야 한다.

Privacy
프라이버시

31. The research, design, development, deployment, and use of AI systems should respect, protect and promote privacy, a right essential to the protection of human dignity and human agency. Adequate data governance mechanisms should be ensured throughout the lifecycle of AI systems including as concerning the collection of data, control over the use of data through

informed consent and permissions and disclosures of the application and use of data, and ensuring personal rights over and access to data.

31. 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용은 인간 존엄성과 인간 활동의 보호에 있어 핵심적인 권리인 프라이버시를 존중 · 보호 · 증진하여야 한다. 데이터 수집, 사전 동의 · 허가 및 데이터의 응용 · 사용에 대한 공개를 통한 데이터 사용의 통제, 데이터에 대한 개인의 권리 및 접근 보장을 비롯한 적절한 데이터 거버넌스 메커니즘은 인공지능 시스템 수명 주기 전 영역에서 보장되어야 한다.

Awareness and literacy
인식 및 활용능력

32. Public awareness and understanding of AI technologies and the value of data should be promoted through education, public campaigns and training to ensure effective public participation so that citizens can take informed decisions about their use of AI systems. Children should be protected from reasonably foreseeable harms arising from AI systems, should

have access to such systems through education and training, and children should not be disempowered by their interaction with AI systems.

32. 인공지능 기술 및 데이터의 가치에 대한 대중의 인식과 이해는 교육, 공공 캠페인, 훈련을 통해 제고되어야 하는데, 이로써 대중의 실제적 참여가 보장되어 시민들이 인공지능 시스템의 사용에 대해 올바른 결정을 내릴 수 있게 된다. 아동은 인공지능 시스템으로 인해 발생하는 합리적으로 예측 가능한 피해로부터 보호받아야 하며, 교육훈련을 통해 이런 시스템에 접근할 수 있어야 하고, 인공지능 시스템과의 상호작용으로 인해 역량을 상실해서는 안된다.

Multi-stakeholder and adaptive governance
다자적 및 유동적 거버넌스

33. Governance of AI should be responsive to shifts in technology and associated business models, inclusive (with the participation of multiple stakeholders), potentially distributed across different levels, and ensure through a cross-domain systems approach, fit-for-purpose governance responses.

33. 인공지능 거버넌스는 기술 및 유관 비즈니스 모델의 변화에 민감해야 하며, (다자적 참여에) 포용적이어야 하고, 다양한 수준에 잠재적으로 분산되어 있어야 하며, 분야를 넘나드는 시스템 접근방식을 통해 목적에 맞는 거버넌스 대응을 보장해야 한다.

34. Governance should consider a range of responses from soft governance through self-regulation and certification processes to hard governance with national laws and, where possible and necessary, international instruments. In order to avoid negative consequences and unintended harms, governance should include aspects of anticipation, protection, monitoring of impact, enforcement and redressal.

34. 거버넌스는 자기 관리 및 인증 절차를 통한 유연한 거버넌스부터, 국내법, 또는 가능하고 필요한 경우 국제적 수단을 통한 엄격한 거버넌스까지 다양한 형태의 대응을 고려해야 한다. 부정적 결과 및 의도치 않은 피해를 방지하기 위해서는, 거버넌스가 예상, 보호, 영향 모니터링, 집행, 시정 같은 측면을 포함해야 한다.

GROUP 2
그룹 2

Fairness
공정성

35. AI actors should respect fairness, equity and inclusiveness, as well as make all efforts to minimize and avoid reinforcing or perpetuating socio-technical biases including racial, ethnic, gender, age, and cultural biases, throughout the full lifecycle of the AI system.

35. 인공지능 행위 주체는 인공지능 시스템의 수명 주기 전 영역에서, 공정성, 형평성, 포용성을 존중함과 더불어, 인종 · 민족 · 성 · 연령 · 문화 편향을 비롯한 사회-기술적 편향의 강화 또는 지속을 최소화 및 예방하기 위한 모든 노력을 기울여야 한다.

Transparency and explainability
투명성 및 설명가능성

36. While, in principle, all efforts need to be made to increase transparency and explainability of AI systems

to ensure trust from humans, the level of transparency and explainability should always be appropriate to the use context, as many trade-offs exist between transparency and explainability and other principles such as safety and security.

36. 원칙적으로, 인간의 신뢰를 확고히 하기 위해서는 인공지능 시스템의 투명성과 설명가능성을 향상시키려는 모든 노력이 이루어져야 하는 반면, 투명성 및 설명가능성의 수준은 항상 상황에 적합해야 하는데, 이는 투명성 및 설명가능성과 안전 및 보안과 같은 다른 원칙 사이에는 많은 맞교환관계가 존재하기 때문이다.

37. Transparency means allowing people to understand how AI systems are researched, designed, developed, deployed, and used, appropriate to the use context and sensitivity of the AI system. It may also include insight into factors that impact a specific prediction or decision, but it does not usually include sharing specific code or datasets. In this sense, transparency is a socio-technical issue, with the aim of gaining trust from humans for AI systems.

37. 투명성은 인공지능 시스템이 사용 맥락 및 민감도에 적합하게 어떻게 연구 · 설계 · 개발 · 보급 · 사용되어지는지를 사람들이 이해할 수 있게끔 하는 것을 의미한다. 또한 이러한 이해는 특정한 예측 또는 결정에 영향을 미치는 요인에 대한 통찰력을 포함할 수도 있지만, 일반적으로 특정 코드 또는 데이터 집합을 공유하는 것까지는 포함하지 않는다. 이런 의미에서, 투명성은 인공지능 시스템에 대한 인간의 신뢰를 얻기 위한 목적의 사회-기술적 사안이다.

38. Explainability refers to making intelligible and providing insight into the outcome of AI systems. The explainability of AI models also refers to the understandability of the input, output and behaviour of each algorithmic building block and how it contributes to the outcome of the models. Thus, explainability is closely related to transparency, as outcomes and sub processes leading to outcomes should be understandable and traceable, appropriate to the use context.

38. 설명가능성은 인공지능 시스템의 결과에 타당한 통찰력을 제공하는 것을 의미한다. 또한 인공지능 모델의 설명가능성은 각 알고리즘 구성 요소의 입력·출력·동작에 대한 이해가능성

과 이것이 모델의 결과물에 기여하는 방식을 가리킨다. 따라서, 결과물 및 결과물로 이어지는 하위 과정은 이해 및 추적이 가능해야 하며 사용 맥락에 적합해야 한다는 점에서, 설명 가능성은 투명성과 매우 밀접하게 연관되어 있다.

Safety and security
안전 및 보안

39. The research, design, development, deployment, and use of AI systems should avoid unintended harms (safety risks) and vulnerabilities to attacks (security tasks), so as to ensure safety and security throughout the lifecycle of the AI system.

39. 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용은 인공지능 시스템의 수명 주기 전 영역에서 안전과 보안을 보장하기 위해 의도치 않은 피해(안전 위험)와 공격에 대한 취약성(보안 위험)을 예방하여야 한다.

40. Governments should play a leading role in ensuring safety and security of AI systems, including through establishing national and international standards and

norms in line with applicable international human rights law, standards and principles. Strategic research on potential safety and security risks associated with different approaches to realize long-term AI should be continuously supported to avoid catastrophic harms.

40. 정부는 적용할 수 있는 국제 인권법 · 기준 · 원칙에 따라 국내 · 국제 기준 및 규범을 수립함으로써 인공지능 시스템의 안전 및 보안을 보장하는 데 주도적인 역할을 해야 한다. 치명적 피해를 예방하기 위해서는, 장기적 인공지능을 실현하기 위한 다양한 접근법에서 발생하는 잠재적 안전 및 보안 위험에 대한 전략적 연구를 지속적으로 지원해야 한다.

Responsibility and accountability
책임 및 책무성

41. AI actors should assume moral and legal responsibility in accordance with extant international human rights law and ethical guidance throughout the lifecycle of AI systems. The responsibility and liability for the decisions and actions based in anyway on an AI system should always ultimately be attributable to AI actors.

41. 인공지능 행위 주체는 인공지능 시스템의 수명 주기 전 영역에서 현존 국제인권법 및 윤리 지침에 따른 윤리적 · 법적 책임을 져야 한다. 어떤 방식으로든지 일단 인공지능 시스템에 기반하여 내린 결정 및 행동에 대한 책임과 법적 손해배상 책임은 항상 궁극적으로 인공지능 행위 주체에게 귀속되어야 한다.

42. Appropriate mechanisms should be developed to ensure accountability for AI systems and their outcome. Both technical and institutional designs should be considered to ensure auditability and traceability of (the working of) AI systems.

42. 인공지능 시스템과 이의 결과물에 대한 책무성을 확고하게 하기 위해서는 적절한 메커니즘이 개발되어야 한다. 인공지능 시스템(또는 이의 작동)에 대한 감사 및 추적을 가능하게 하는 기술 및 제도적 설계가 확립되어야 한다.

IV AREAS OF POLICY ACTION
정책 행동 영역

ACTION GOAL I: ETHICAL STEWARDSHIP
행동 목표 I: 윤리적 책무(stewardship)

43. Ensure alignment of AI research, design, development, deployment, and use with foundational ethical values such as human rights, diversity and inclusiveness, etc.

43. 인공지능 연구 · 설계 · 개발 · 보급 · 사용이 인권, 다양성, 포용성 등과 같은 기초적인 윤리적 가치와 일치하는지 확인해야 한다.

Policy Action 1: Promoting Diversity & Inclusiveness
정책 행동 1: 다양성 및 포용성 증진

44. Member States should work with international organizations to ensure the active participation of all Member States, especially LMICs in international

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

44. 회원국은 국제기구와 협업하여 모든 회원국, 특히 중·저소득 국가가 인공지능에 관한 국제적 논의에 적극적으로 참여할 수 있도록 해야 한다. 이는 모든 지역의 평등한 참여를 보장할 수 있는 자금 지원을 통해, 또는 기타 메커니즘을 통해 가능하다.

45. Member States should require AI actors to disclose and combat any cultural and social stereotyping in the workings of AI systems whether by design or by negligence, and ensure that training data sets for AI systems should not foster cultural and social inequalities. Mechanisms should be adopted to allow end users to report such inequalities, biases and stereotypes.

45. 회원국은 고의 또는 부주의로 인해 인공지능 시스템의 작동 내에 존재하는 모든 문화 · 사회적 고정관념을 인공지능 행위 주체가 공개하고 대응하도록 요구해야 하며, 인공지능 시스템을 위한 훈련데이터 집합이 문화·사회적 불평등을 조장하지 않도록 해야 한다. 최종 사용자가 그러한 불평등, 편

향, 고정관념을 보고할 수 있도록 하는 메커니즘이 도입되어야 한다.

46. Member States should ensure that AI actors demonstrate awareness and respect for the current cultural and social diversities including local customs and religious traditions, in the research, design, development, deployment, and use of AI systems while being consistent with international human rights standard and norms.

46. 회원국은 인공지능 행위 주체가 인공지능 시스템의 연구 · 설계 · 개발 · 보급 · 사용에서 국제인권기준 및 규범을 따르면서도 현지 관습 및 종교적 전통을 비롯한 오늘날의 문화·사회적 다양성에 대해 인식과 존중을 보이도록 해야 한다.

47. Member States should work to address the diversity gaps currently seen in the development of AI systems, including diversity in training datasets and in AI actors themselves. Member States should work with all sectors, international and regional organizations and other entities to empower women and girls to participate in all stages of an AI system lifecycle by offering

incentives, access to mentors and role models, and protection from harassment. They should also work to make the domain of AI more accessible to people from diverse ethnic backgrounds as well as people with disabilities. Moreover, equal access to AI system benefits should be promoted, particularly for marginalized groups.

47. 회원국은 훈련 데이터 집합 및 인공지능 행위 주체의 다양성을 비롯하여 현재 인공지능 시스템의 개발에서 발견되는 다양성 격차를 해소하기 위해 노력해야 한다. 회원국은 모든 부문, 국제 · 지역 기구, 기타 단체와 협력하여, 여성 및 소녀에게 인센티브, 멘토 · 역할 모델과의 만남, 괴롭힘으로부터의 보호를 제공함으로써 그들이 인공지능 시스템 수명 주기의 모든 단계에 참여할 수 있도록 힘을 실어주어야 한다. 또한 회원국은 다양한 민족 배경의 사람에 더하여 장애인이 인공지능 분야에 더 쉽게 접근할 수 있도록 노력해야 한다. 더욱이, 회원국은 특히 소외집단이 인공지능 시스템의 혜택에 평등하게 접근할 수 있도록 해야 장려해야 한다.

48. Member States should work with international organizations to mainstream AI ethics by including discussions of AI-related ethical issues into relevant interna-

tional, intergovernmental and multi-stakeholder fora.

48. 회원국은 국제기구와 협력하여 인공지능과 관련된 윤리적 사안에 대한 논의를 유관 국제 · 국가간 · 다자간 포럼에 포함시킴으로써 인공지능 윤리를 주류화시켜야 한다.

ACTION GOAL II: IMPACT ASSESSMENT
행동 목표 II: 영향 평가

49. Build observatory and anticipatory capacities to respond in time to negative or other unintended consequences arising from AI systems.

49. 인공지능 시스템에서 발생하는 부정적인 또는 기타 의도치 않은 결과에 대하여 적시에 대응할 수 있는 관측 및 예측 능력을 구축해야 한다.

Policy Action 2: Addressing Labour Market Changes
정책 행동 2: 노동 시장 변화에 대한 대응

50. Member States should work to assess and address the impact of AI on labour markets and its implications for ed-

ucation requirements. This can include the introduction of a wider range of 'core skills' at all education levels to give new generations a fair chance of finding jobs in a rapidly changing market and to ensure their awareness of the ethical aspects of AI. Skills such as 'learning how to learn', communication, teamwork, empathy, and the ability to transfer one's knowledge across domains, should be taught alongside specialist, technical skills. Being transparent about what skills are in demand and updating school curricula around these is key.

50. 회원국은 인공지능이 노동 시장에 미치는 영향과 교육 요구 조건에 갖는 시사점을 평가하고 이에 대응해야 한다. 급변하는 시장에서 신세대에게 공정한 구직 기회를 제공하고 인공지능의 윤리적 측면에 대한 인식을 제고하기 위해서, 여기에는 모든 교육 수준에서 광범위한 '핵심적 능력'의 도입이 포함될 수 있다. '학습 방법 습득', 의사소통, 팀워크, 공감, 분야에 구애 받지 않는 지식 전달력과 같은 능력은 전문가, 전문 기술과 함께 가르쳐야 한다. 어떤 숙련이 수요가 있는지 투명하게 공개하고 이를 중심으로 학교 커리큘럼을 갱신하는 것이 중요하다.

51. Member States should work with private entities,

NGOs and other stakeholders to ensure a fair transition for at-risk employees. This includes putting in place upskilling and reskilling programs, finding creative ways of retaining employees during those transition periods, and exploring 'safety net' programs for those who cannot be retrained.

51. 회원국은 불안정성에 노출된 근로자의 정당한 직업 전환을 보장하기 위해 민간 단체, 비정부기구, 기타 이해관계자들과 협력해야 한다. 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.

52. Member States should encourage researchers to analyze the impact of AI on the local labour market in order to anticipate future trends and challenges. These studies should shed light on which economic, social and geographic sectors will be most affected by the massive incorporation of AI.

52. 회원국은 미래의 트렌드 및 어려움을 예측하기 위해, 연구자들이 인공지능이 지역 노동 시장에 미치는 영향을 분석하도

록 장려해야 한다. 이러한 연구는 인공지능의 대규모 편입에 가장 영향을 받을 경제 · 사회 · 지리적 부문이 어디인지 밝혀야 한다.

53. Member States should develop labour force policies targeted at supporting women and underrepresented populations to make sure no one is left out of the digital economy powered by AI. Special investment in providing targeted programs to increase the preparedness, employability, career development and professional growth of women and underrepresented populations should be considered, and implemented if feasible.

53. 회원국은 여성 및 소외 인구를 선별 지원하는 노동력 정책을 개발하여, 인공지능 주도의 디지털 경제에서 낙오자가 없게 해야 한다. 준비성, 취업능력, 여성 및 소외 인구의 경력 개발 및 전문가로서의 성장을 증진시키기 위하여, 표적 프로그램을 제공함에 있어 특별한 투자를 고심해야 하며, 가능한 한 마련해야 한다.

Policy Action 3: Addressing the social and economic impact of AI
정책 행동 3: 인공지능의 사회 · 경제적 영향에 대한 대응

54. Member States should devise mechanisms to prevent the monopolization of AI and the resulting inequalities, whether these are data, research, technology, market or other monopolies.

54. 데이터·연구·과학기술·시장 및 기타 어떤 독점이든지 간에, 회원국은 인공지능 시스템 수명 주기 내내 인공지능 시스템의 독점 및 이로 인한 불균형을 방지하기 위한 메커니즘을 고안해야 한다.

55. Member States should work with international organizations, private and non-governmental entities to provide adequate AI literacy education to the public especially in LMICs in order to reduce the digital divide and digital access inequalities resulting from the wide adoption of AI systems.

55. 인공지능 시스템의 광범위한 도입에서 파생되는 디지털 격차 및 디지털 접근불평등을 줄이기 위해서, 대중, 특히 중 ·

저소득국가의 대중에게 적절한 인공지능 활용능력 교육을 제공할 수 있도록 국제 기구, 민간 단체, 비정부단체와 협력해야 한다.

56. Member States should establish monitoring and evaluation mechanisms for initiatives and policies related to AI ethics. Possible mechanisms include: a repository covering ethical compliance initiatives across UNESCO's areas of competence, an experience sharing mechanism for Member States to seek feedback from other Member States on their policies and initiatives, and a guide for developers of AI systems to assess their adherence to policy recommendations mentioned in this document.

56. 회원국은 인공지능 윤리와 관련된 계획안 및 정책에 대한 모니터링·심사 메커니즘을 확립해야 한다. 메커니즘의 예는 다음을 포함한다. 유네스코의 권한 범위 내 영역의 모든 윤리 준수 계획안을 보관하는 저장소, 정책 및 계획안에 관하여 회원국이 다른 회원국으로부터 피드백을 구할 수 있는 경험 공유 메커니즘, 인공지능 시스템 개발자가 스스로 본 안에서 언급된 정책 권장사항의 준수 여부를 평가할 수 있는 지침이다.

57. Member States are encouraged to consider a certification mechanism for AI systems similar to the ones used for medical devices. This can include different classes of certification according to the sensitivity of the application domain and expected impact on human lives, the environment, ethical considerations such as equality, diversity and cultural values, among others. Such a mechanism might include different levels of audit of systems, data, and ethical compliance. At the same time, such a mechanism must not hinder innovation or disadvantage small enterprises or startups by requiring large amounts of paperwork. These mechanisms would also include a regular monitoring component to ensure system robustness and continued integrity and compliance over the entire lifetime of the AI system, requiring re-certification if necessary.

57. 회원국은 의료 기기에 사용되는 것과 같은 종류의 인공지능 시스템에 대한 인증 메커니즘을 두는 것을 고려해야 한다. 여기에는 응용 분야의 민감성과 인간의 삶, 환경, 특히 평등 · 다양성 · 문화적 가치와 같은 윤리적 고려대상에 미칠 것으로 예측되는 영향에 따라서 다양한 등급의 인증이 포함될 수 있

다. 이러한 메커니즘에는 시스템 · 데이터 · 윤리 준수에 대한 다양한 수준의 감사가 포함될 수 있다. 동시에, 이러한 메커니즘은 많은 서류 작업을 요구하여 혁신을 방해하거나, 중소기업 및 스타트업에게 불이익을 주어서는 안된다. 또한 이 메커니즘은 정기적인 모니터링 요소도 포함하여, 필요한 경우 재인증도 요구함으로써, 인공지능 시스템의 전체 수명 동안 시스템 견고성, 지속되는 일관성 및 윤리 준수를 보장한다.

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

58. 회원국은 민간 기업으로 하여금 인공지능 거버넌스에 다른 이해관계자를 참여시키도록, 그리고 영향평가, 감사, 지속적인 모니터링을 감독하고 인공지능 시스템의 윤리 준수를 보장하기 위한 별도의 '인공지능윤리책임자'와 같은 역할 또는 기타 메커니즘의 추가를 고려하도록 장려해야 한다.

59. Member States should work to develop data governance strategies that ensure the continuous evaluation

of the quality of training data for AI systems including the adequacy of the data collection and selection processes, proper security and data protection measures, as well as feedback mechanisms to learn from mistakes and share best practices among all AI actors. Striking a balance between metadata and users' privacy should be an upfront concern for such a strategy.

59. 회원국은 데이터 수집 및 선택 프로세스의 타당성, 적절한 보안 및 데이터 보호 조치를 포함하는 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. 메타데이터와 사용자 프라이버시 사이의 균형을 맞추는 것은 이러한 전략의 선행 목표가 되어야 한다.

Policy Action 4: Impact on Culture and on the Environment
정책 행동 4: 문화 및 환경에 대한 영향

60. Member States are encouraged to incorporate AI sys-

tems where appropriate in the preservation, enrichment and understanding of cultural heritage, both material and intangible, including rare languages, for example by introducing or updating educational programs related to the application of AI systems in these areas, targeted at institutions and the public.

60. 회원국은 희소언어를 비롯한 유형·무형문화유산의 보존 · 강화 · 이해에 있어 적합한 경우 인공지능 시스템을 활용하도록 권장되는데, 가령 이러한 분야에 인공지능 시스템의 응용과 관련된 기관 및 대중을 대상으로 하는 교육 프로그램을 도입 또는 갱신하는 것이다.

61. Member States are encouraged to examine and address the impact of AI systems, especially Natural Language Processing applications such as automated translation and voice assistants on the nuances of human language. Such an assessment can include maximizing the benefits from these systems by bridging cultural gaps and increasing human understanding, as well as negative implications such as the reduced pervasiveness of rare languages, local dialects, and the tonal and cultural variations associated with human

language and speech.

61. 회원국은 인공지능의 시스템, 특히 인간 언어의 뉘앙스에 대한 자동 번역 및 음성 도우미와 같은 자연어처리(NLP) 응용의 영향을 조사하고 그에 대응하도록 장려된다. 그러한 평가는 문화적 격차 해소와 인간에 대한 이해 증가를 통해 이런 시스템의 혜택을 극대화하는 것은 물론, 희소언어, 지역 방언, 인간 언어 · 발화와 관련된 음성적·문화적 변주의 소실과 같은 부정적 영향까지도 포함할 수 있다.

62. Member States should encourage and promote collaborative research into the effects of long-term interaction of people with AI systems. This should be done using multiple norms, principles, protocols, disciplinary approaches, and assessment of the modification of habits, as well as careful evaluation of the downstream cultural and societal impacts.

62. 회원국은 사람과 인공지능 시스템 간 장기적 상호작용의 영향에 대한 공동 연구를 장려 및 촉진해야 한다. 이는 하위문화의 문화·사회적 영향에 대한 신중한 심사를 비롯하여 여러 가지 규범, 원칙, 프로토콜, 징계 수단, 습관의 시정 평가를 사용함으로써 가능하다.

63. Member States should promote AI education for artists and creative professionals to assess the suitability of AI for use in their profession as AI is being used to create, produce, distribute and broadcast a huge variety of cultural goods and services, bearing in mind the importance of preserving cultural heritage and diversity.

63. 인공지능 기술이 매우 다양한 문화 상품 및 서비스를 창출 · 생산 · 유통 · 방송하는 데에 사용되고 있기에, 회원국은 문화유산 및 다양성의 보존이 중요함을 유념하여, 예술가 및 창조적 직업군이 자신의 직군에서의 인공지능 사용의 적합성을 평가할 수 있도록 인공지능 교육을 장려해야 한다.

64. Member States should promote awareness and evaluation of AI tools among local cultural industries and startups working in the field of culture, to avoid the risk of greater concentration in the cultural market.

64. 회원국은 문화계에 종사하는 지역문화기업 및 스타트업의 인공지능 도구에 대한 인식 및 평가를 제고하여 문화 시장에서 편중화가 심화될 위험성을 예방하여야 한다.

65. Member States should work to assess and reduce

the environmental impact of AI systems, including but not limited to, its carbon footprint. They should also introduce incentives to advance ethical AI-powered environmental solutions and facilitate their adoption in different contexts. Some examples include using AI to:

65. 회원국은 그 자신의 탄소발자국을 비롯한 (단, 이에 국한되지 않는) 인공지능 시스템의 환경적 영향을 평가하고 이를 줄이기 위하여 노력해야 한다. 또한 회원국은 윤리적 인공지능 기반의 환경 솔루션을 발전시키기 위해서 인센티브를 도입해야 하며, 다양한 상황에서의 채택을 촉진해야 한다. 인공지능을 사용하는 몇 가지 예는 다음과 같다.

a. Accelerate the protection, monitoring and management of natural resources.

a. 천연 자원의 보호 · 모니터링 · 관리 가속화.

b. Support the prevention, control and management of climate-related problems.

b. 기후문제의 예방·통제·관리 지원.

c. Support a more efficient and sustainable food eco-system.

c. 더 효율적이고 지속가능한 식량 생태계 지원.

d. Accelerate the access to and mass adoption of green energy.

d. 친환경 에너지에 대한 접근 및 대규모 채택 가속화.

ACTION GOAL III: CAPACITY BUILDING FOR AI ETHICS

행동 목표 III: 인공지능 윤리를 위한 역량 구축

66. Develop human and institutional capacity to enable ethical impact assessment, oversight and governance.

66. 윤리적 영향 평가, 감독 및 거버넌스를 가능하게 하는 인적 및 기관 역량을 계발해야 한다.

Policy Action 5: Promoting AI Ethics Education & Awareness
정책 행동 5: 인공지능 윤리 교육 및 인식 증진

67. Member States should encourage in accordance with their national education programmes and traditions the embedding of AI ethics into the school and university curricula for all levels and promote cross collaboration between technical skills and social sciences and humanities. Online courses and digital resources should be developed in local languages and in accessible formats for people with disabilities.

67. 회원국은 자국의 교육 프로그램 및 전통에 따라 학교 및 대학의 전 학년 교육과정에 인공지능 윤리 과목 포함을 장려하고, 기술 숙련 · 사회과학 · 인문학 간의 교차협력을 증진해야 한다. 온라인 과정 및 디지털 자료는 토착 언어로, 그리고 장애인도 접근 가능한 형태로 개발되어야 한다.

68. Member States should promote the acquisition of 'prerequisite skills' for AI education, such as basic literacy, numeracy, and coding skills, especially in countries where there are notable gaps in the education of these skills.

68. 회원국은 기본언어능력, 산술능력, 코딩 능력과 같은 인공지능 교육을 위한 '기초 소양'의 습득을 특히 이러한 기초소양 교육의 격차가 두드러지는 국가에서 장려하여야 한다.

69. Member States should introduce flexibility into university curricula and increase ease of updating them, given the accelerated pace of innovations in AI systems. Moreover, the integration of online and continuing education and the stacking of credentials should be explored to allow for agile and updated curricula.

69. 인공지능 시스템 혁신의 속도가 가속화됨에 따라, 회원국은 대학 교육과정에 유연성을 도입하고 갱신의 용이성을 높여야 한다. 더욱이, 교육과정이 동향에 민감한 최신의 상태를 유지하기 위해서는 지속적 온라인 교육 통합 및 자격증 발행이 강구되어야 한다.

70. Member States should promote general awareness programs of AI and the inclusive access to knowledge on the opportunities and challenges brought about by AI. This knowledge should be accessible to technical and non-technical groups with a special focus on un-

derrepresented populations.

70. 회원국은 인공지능에 대한 보편 인식 프로그램을 장려하고, 인공지능 기술이 초래하는 기회 및 어려움에 대한 지식을 누구나 습득할 수 있도록 해야 한다. 이러한 지식은 특히 소외 인구에 중점을 맞추어 기술전문가집단과 비전문집단에게도 이용이 쉬워야 한다.

71. Member States should encourage research initiatives on the use of AI in teaching, teacher training and e-learning, among other topics, in a way that enhances opportunities and mitigates the challenges and risks associated with these technologies. This should always be accompanied by an adequate impact assessment of the quality of education and impact on students and teachers of the use of AI and ensure that AI empowers and enhances the experience for both groups.

71. 회원국은 다른 주제보다도 교육, 교사 연수, 온라인 교육에서 인공지능의 사용에 대한 연구 계획을, 이 기술과 관련된 기회를 향상시키고 도전 및 위험성은 완화하는 방식으로 장려해야 한다. 이 계획에는 교육의 질에 대한 적절한 영향 평가와,

인공지능의 사용이 학생 및 교사에 미치는 영향에 대한 적절한 평가가 수반되어야 하며, 인공지능 기술이 학생 및 교사 모두의 역량을 강화하고 그들의 경험을 향상시킬 수 있도록 보장해야 한다.

72. Member States should support collaboration agreements between academic institutions and the industry to bridge the gap of skillset requirements and promote collaborations between industry sectors, academia, civil society, and the government to align training programs and strategies provided by educational institutions, with the needs of the industry. Project-based learning approaches for AI should be promoted, allowing for partnerships between companies, universities and research centers.

72. 회원국은 학술 기관과 업계 간의 협업 계약을 지원하여 기술 요건의 격차를 줄이고, 산업 부문, 학계, 시민 사회, 정부 간의 협업을 증진하여 업계의 요구에 맞추어 교육 기관이 제공하는 훈련 프로그램 · 전략을 조정해야 한다. 기업, 대학, 연구소 간의 파트너십을 허용함으로써, 프로젝트 기반의 인공지능 학습 접근법을 장려해야 한다.

73. Member States should particularly promote the participation of women, diverse races and cultures, and people with disabilities, in AI education programs from basic school to higher education, as well as promote the monitoring and sharing of best practices with other Member States.

73. 회원국은 초등 교육기관부터 고등 교육기관까지의 인공지능 교육 프로그램에서 특히 여성, 다양한 인종 및 문화, 장애인의 참여를 증진해야 하며, 다른 회원국과 모범 사례의 모니터링 및 공유를 장려해야 한다.

Policy Action 6: Promoting AI Ethics Research
정책 행동 6: 인공지능 윤리 연구 장려

74. Member States should promote AI ethics research either through direct investments or by creating incentives for the public and private sectors to invest in this area.

74. 회원국은 직접투자를 통하여 또는 공공 및 민간 부문으로 하여금 이 분야에 투자하게 하는 인센티브를 만듦으로써 인공

지능 윤리연구를 장려해야 한다.

75. Member States should ensure that AI researchers are trained in research ethics and require them to include ethical considerations in their research design and end products, particularly analyses of the datasets they use, how they are annotated and the quality and the scope of the results.

75. 회원국은 인공지능 연구자가 연구 윤리를 체화하도록 해야 하며, 그들이 연구 설계 및 최종 생산, 특히 사용되는 데이터 집합의 분석, 즉 주석표기방식 및 결과의 품질·범위에 대하여 윤리적 사항을 고려하도록 요구해야 한다.

76. Member States and private companies should facilitate access to data for research for the scientific community at the national level where possible to promote the capacity of the scientific community, particularly in developing countries. This access should not be at the expense of citizens' privacy.

76. 회원국과 민간 기업은 특히 개발도상국에서 과학계의 역량이 증진될 수 있도록 국가적 차원에서 과학계가 연구를 위해 데

이터에 쉽게 접근할 수 있도록 해야 한다. 이 접근으로 인하여 시민의 프라이버시의 침해가 일어나서는 안 된다.

77. Member States should promote gender diversity in AI research in academia and industry by offering incentives to women to enter the field, put 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 promote diversity.

77. 회원국은 여성이 인공지능 분야에 진출할 수 있도록 인센티브를 제공하고, 인공지능 연구 공동체 내의 성 고정관념 및 성적 괴롭힘에 대항할 수 있는 메커니즘을 마련하고, 학술 및 민간 단체가 성별 다양성을 강화하는 방법에 대한 모범 사례를 공유하도록 장려함으로써, 학계 및 산업의 인공지능 연구에서 성 다양성을 증진해야 한다.

78. Member States and funding bodies should promote interdisciplinary AI research by including disciplines other than science, technology, engineering, and mathematics (STEM), e.g. law, international relations, political sciences, education, philosophy, culture, and

linguistic studies to ensure a critical approach to AI research and proper monitoring of possible misuses or adverse effects.

78. 회원국은 인공지능 연구의 비판적 접근법과 잠재적 손실 또는 악영향에 대한 적절한 모니터링을 보장하기 위해서는, 법학 · 국제관계학 · 정치학 · 교육학 · 철학 · 문화학 · 언어학과 같이 과학·기술·공학·수학(STEM) 외 다른 학문분야를 포함시킴으로써 학제적 인공지능 연구를 장려해야 한다.

ACTION GOAL IV: DEVELOPMENT AND INTERNATIONAL COOPERATION
행동 목표 IV: 개발 및 국제 협력

79. Ensure a cooperative and ethical approach to using AI in development applications, given the great opportunity this technology affords towards the acceleration of development efforts.

79. 인공지능 기술이 개발 노력을 가속화할 수 있는 좋은 기회를 제공하는 경우, 개발 응용에서 인공지능의 사용에 대한 협력적 · 윤리적 접근을 보장해야 한다.

Policy Action 7: Promoting Ethical Use of AI in Development
정책 행동 7: 개발에서 인공지능의 윤리적 사용 권장

80. Member States should encourage the ethical use of AI in areas of development such as healthcare, agriculture/food supply, education, culture, environment, water management, infrastructure management, economic planning and growth, and others.

80. 회원국은 건강관리, 농·식품 공급, 교육, 문화, 환경, 수도 관리, 인프라 관리, 경제 계획 및 성장 등의 개발 분야에 있어서 인공지능의 사용을 장려해야 한다.

81. Member States and international organizations should strive to provide platforms for international cooperation on AI for development, including by contributing expertise, funding, data, domain knowledge, infrastructure, and facilitating workshops between technical and business experts to tackle challenging development problems, especially for LMICs and LDCs.

81. 회원국과 국제 기구는 전문지식, 자금 지원, 데이터, 각 분야 지식, 인프라를 기부하고 특히 중·저소득국가 및 최빈개발도

상국의 힘겨운 개발 문제를 해결할 수 있도록 기술 및 사업 전문가 간의 공동작업을 원활하게 하는 것을 비롯하여, 개발에 사용되는 인공지능에 대한 국제적 협력을 위한 플랫폼을 제공하기 위해 힘써야 한다.

82. Member States should work to promote international collaborations on AI research, including research centers and networks that promote greater participation of researchers from LMICs and other emerging geographies.

82. 회원국은 중·저소득국가 및 기타 부상하는 지역 출신의 연구자의 더 많은 참여를 장려할 수 있는 연구 및 네트워크를 비롯하여, 인공지능 연구에 있어서 국제적 협업을 증진하기 위해 노력해야 한다.

Policy Action 8: Promoting International Cooperation on AI Ethics
정책 행동 8: 인공지능 윤리에 대한 국제적 협력 증진

83. Member States should work through international organizations and research institutions to conduct

AI ethics research. Both public and private entities should ensure that algorithms and data used in a wide array of AI areas – from policing and criminal justice to employment, health and education – are applied equally and fairly, including investigations into what sorts of equality and fairness are appropriate in different cultures and contexts, and exploring how to match those to technically feasible solutions.

83. 회원국은 국제 기구 및 연구 기관을 통해 인공지능 윤리 연구를 수행하려 노력해야 한다. 공공 및 민간 단체 모두 치안 유지 및 사법 제도부터 고용 · 건강 · 교육까지 다양한 종류의 인공지능 분야에서 알고리즘과 데이터가 공평하고 공정하게 사용될 수 있도록 해야 하는데, 이는 어떤 형태의 공평성 및 공정성이 각 문화 및 맥락에서 적절한지 조사하고 이를 기술적으로 구현 가능한 해결책에 접목시키는 것을 포함한다.

84. Member States should encourage international co-operation in AI development and deployment to bridge geo-technological lines. This necessitates a multi-stakeholder effort at the national, regional and international levels. Technological exchanges/ consultations should take place between Member States

and their populations, between the public and private sectors, and between and among Member States.

84. 회원국은 인공지능 개발 및 보급에서 국제적인 협력을 장려하여 지정학적 기술연결로(geo-technological line)를 확보해야 한다. 이를 위해서는 국가 · 지역 · 국제적 수준의 다자적 노력이 필요하다. 기술 교류 및 협의는 회원국과 자국민 사이, 공공 부문과 민간 부분 사이, 회원국 사이에 이루어져야 한다.

ACTION GOAL V: GOVERNANCE FOR AI ETHICS
행동 목표 V: 인공지능 윤리를 위한 거버넌스

85. Promote and guide the inclusion of ethical considerations in the governance of AI systems.

85. 인공지능 시스템의 거버넌스에서 윤리적 고려사항을 포함하도록 장려하고 안내해야 한다.

Policy Action 9: Establishing Governance Mechanisms for AI Ethics
정책 행동 9: 인공지능 윤리를 위한 거버넌스 메커니즘 구축

86. Member States should ensure that any AI governance mechanism is:

86. 회원국은 모든 인공지능 거버넌스 메커니즘이 이하의 특징을 갖도록 해야 한다.

a. Inclusive: invites and encourages participation of representatives of indigenous communities, women, young and elderly people, people with disabilities, and other minority and underrepresented groups.

a. 포용성: 현지 커뮤니티, 여성, 청년 및 노인, 장애인, 기타 소수 · 소외 집단의 대표를 끌어들여 참여를 장려해야 한다.

b. Transparent: accepts oversight from relevant national structures or trusted thirdparties. For the media, this could be a cross-sectoral taskforce that fact-checks sources; for technology companies, this could be

external audits of design, deployment and internal audit processes; for Member States, this could be reviews by human rights forums.

b. 투명성: 관련 국가 구조 또는 신뢰할 수 있는 제 3자의 감독을 허용해야 한다. 이는 미디어의 경우 모든 부문을 넘나들며 출처의 사실 확인을 담당하는 다분야 전담반일 수 있고, IT 기업의 경우 설계 · 보급 · 내부감사과정의 외부 감사일 수 있고, 회원국의 경우 인권 포럼에 의한 검토일 수 있다.

c. Multidisciplinary: any issue should be viewed in a holistic way and not only from the technological point of view.

c. 학제성: 모든 사안은 단지 기술적 관점에서만 아니라 전체적인 방식으로 봐야 한다.

d. Multilateral: international agreements should be established to mitigate and redress any harm that can appear in a country caused by a company or user based in another. This does not negate different countries and regions developing their own rules as appropriate to their cultures.

d. 다자성: 한 국가에서 외국계 기업 또는 외국인 사용자에 의해 발생할 수 있는 모든 피해를 완화 및 해결하기 위한 국제 협약이 수립되어야 한다. 이는 다양한 국가 및 지역이 그 문화에 적합하게 자체적으로 개발하는 규칙을 무효화하지 않는다.

87. Member States should foster the development of, and access to, a digital ecosystem for ethical AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. 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 data trusts, to support the safe, fair, legal and ethical sharing of data, among others.

87. 회원국은 윤리적 인공지능을 위한 디지털 생태계의 개발 및 접근을 촉진해야 한다. 이러한 디지털 생태계에는 특히 디지털 기술 · 인프라와 적절한 경우 인공지능 지식 공유 메커니즘까지 포함된다. 이와 관련하여, 회원국은 정부 정책 및 규제 프레임워크 검토를 고려해야 하는데, 여기에는 특정 인공

지능 요구사항을 반영하기 위한 정보 및 열린 정부에 대한 것과 안전하고 공정하며 합법적이고 윤리적인 데이터 공유를 지원하기 위해 데이터 신뢰와 같은 메커니즘을 장려하는 것이 포함된다.

88. Member States should encourage development and use of comparable AI guidelines, including ethical aspects at global and regional levels, and gather the required evidence to evaluate, monitor and control the progression in the ethical implementation of AI systems.

88. 회원국은 전지구적 및 지역적 수준에서의 여러 윤리적 측면들과 같이 비교 가능한 인공지능 지침서의 개발 및 사용을 장려해야 하며, 인공지능 시스템의 윤리적 구현 진행과정을 평가 · 모니터링 · 통제하는 데 필요한 증거를 수집해야 한다.

89. Member States should consider the development and implementation of an international legal framework to encourage international cooperation between States and other stakeholders.

89. 회원국은 국가와 다른 이해관계자 간의 국제적 협력을 장려하기 위한 국제법 프레임워크의 개발 및 구현을 고려해야 한다.

Policy Action 10: Ensuring Trustworthiness of AI Systems
정책 행동 10: 인공지능 시스템의 신뢰성 보장

90. Member States and private companies should implement proper measures to monitor all phases of an AI system lifecycle, including the behaviour of algorithms in charge of 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.

90. 회원국 및 민간 기업은 인공지능 시스템 수명 주기의 전 단계를 모니터링할 수 있는 적절한 조치를 마련해야 하며, 이런 단계들은 의사결정을 내리는 알고리즘의 동작, 데이터뿐만 아니라 어떤 프로세스, 특히 공공 서비스에서 및 최종 사용자와의 직접적인 상호작용이 필요한 곳의 프로세스와 연관된 인공지능 행위 주체를 포함한다.

91. Member States should work on setting clear requirements for AI system transparency and explainability based on:

91. 회원국은 인공지능 이하의 항목에 기반한 시스템의 투명성 및 설명가능성에 대한 명확한 요구사항을 설정하기 위해 노력해야 한다.

a. Application domain: some sectors such as law enforcement, security, education and healthcare, are likely to have a higher need for transparency and explainability than others.

a. 응용 분야: 법 집행, 보안, 교육, 의료와 같은 일부 부문은 다른 부문보다 투명성과 설명가능성에 대한 요구가 더 높을 수 있다.

b. Target audience: the level of information about an AI system's algorithms and outcome and the form of explanation required may vary depending on who are requesting the explanation, for example: users, domain experts, developers, etc.

b. 대상자: 인공지능 시스템의 알고리즘 및 결과에 대한 정보 수준과 필요한 설명의 형식은 사용자, 각 분야 전문가, 개발자 등과 같이 설명을 요청하는 사람이 누구인지에 따라 다를 수 있다.

c. Feasibility: many AI algorithms are still not explainable; for others, explainability adds a significant implementation overhead. Until full explainability is technically possible with minimal impact on functionality, there will be a trade-off between the accuracy/quality of a system and its level of explainability.

c. 실현가능성: 많은 인공지능 알고리즘은 여전히 설명이 불가능하다. 설명이 가능하다고 하더라도 설명가능성은 상당한 구현 비용을 요구한다. 기능에 미치는 영향의 최소화와 함께 완전한 설명가능성이 기술적으로 가능해지기 전까지는 시스템의 정확도 · 품질과 이의 설명가능성 정도 사이에는 맞교환 관계가 존재할 것이다.

92. Member States should encourage research into transparency and explainability by putting additional funding into those areas for different domains and at

different levels (technical, natural language, etc.).

92. 회원국은 기술 및 자연어 등과 같이 다양한 분야·수준에서 투명성 및 설명가능성에 추가 자금을 투입함으로써 이에 대한 연구를 장려해야 한다.

93. Member States and international organizations should consider developing international standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined.

93. 회원국과 국제 기구는 시스템을 객관적으로 평가하고 규정 준수 수준을 결정할 수 있도록, 측정 가능하고 평가 가능한 수준의 투명성을 명시하는 국제 기준을 개발해야 한다.

Policy Action 11: Ensuring Responsibility, Accountability and Privacy
정책 행동 11: 책임, 책무성 및 프라이버시

94. Member States should review and adapt, as appropriate, regulatory and legal frameworks to achieve

accountability and responsibility for the content and outcomes of AI systems at the different phases of their lifecycle. Governments should introduce liability frameworks or clarify the interpretation of existing frameworks to make it possible to attribute accountability for the decisions and behaviour of AI systems. When developing regulatory frameworks governments should, in particular, take into account that responsibility and accountability must always lie with a natural or legal person; responsibility should not be delegated to an AI system, nor should a legal personality be given to an AI system.

94. 회원국은 인공지능 시스템 수명 주기 각 단계에서 이의 내용물과 결과물에 대한 책무성 및 책임을 확보하기 위해 규제 및 법적 프레임워크를 적절하게 검토하고 조정해야 한다. 정부는 인공지능 시스템의 결정 및 작동에 대한 책임소재를 명확히 할 수 있도록 법적 책임 프레임워크를 도입하거나 기존 프레임워크의 해석을 명료하게 해야 한다. 규제 프레임워크를 개발할 때, 정부는 언제나 책임과 책무성은 반드시 개인 또는 법인에게 있어야 하며 책임이 인공지능 시스템에 위임되거나 인공지능 시스템에 법인격이 부여되지 않아야 한다는 점을 특히 고려해야 한다.

95. Member States are encouraged to introduce impact assessments to identify and assess benefits and risks of AI systems, as well as risk prevention, mitigation and monitoring measures. The risk assessment should identify impacts on human rights, the environment, and ethical and social implications in line with the principles set forth in this Recommendation. Governments should adopt a regulatory framework that sets out a procedure for public authorities to carry out impact assessments on AI systems acquired, developed and/or deployed by those authorities to predict consequences, mitigate risks, avoid harmful consequences, facilitate citizen participation and address societal challenges. As part of impact assessment, the public authorities should be required to carry out self assessment of existing and proposed AI systems, which in particular, should include the assessment whether the use of AI systems within a particular area of the public sector is appropriate and what the appropriate method is. The assessment should also establish appropriate oversight mechanisms, including auditability, traceability and explainability which enables the assessment of algorithms, data and design

processes, as well as include external review of AI systems. Such an assessment should also be multidisciplinary, multi-stakeholder, multicultural, pluralistic and inclusive.

95. 회원국은 위험 예방·완화·모니터링 조치에 더하여 인공지능 시스템의 이익 및 위험성을 파악 및 평가하기 위해 영향평가를 도입하도록 권장된다. 위험성 평가는 본 권고안에 명시된 원칙들에 따라 인권, 환경, 윤리 · 사회적 함의 등에 대한 영향을 파악해야 한다. 정부는 공공기관이 결과 예측, 위험 완화, 피해 예방, 시민 참여 확대, 사회 문제 해결을 위하여 자체적으로 획득, 개발 및(또는) 배치한 인공지능 시스템에 대하여 영향 평가를 수행하는 절차를 제시하는 규제 프레임워크를 채택해야 한다. 영향 평가의 일환으로서, 공공기관은 자체 평가를 수행해야 하며, 이는 특히 공공 부문의 특정 영역 내에서 인공지능 시스템의 사용이 적절한지, 또 적절한 방법론이 무엇인지에 대한 평가를 포함해야 한다. 또한 이 평가는 인공지능 시스템에 대한 외부 검토를 포함함과 더불어, 알고리즘 · 데이터 · 설계과정에 대한 평가를 가능하게 하는 감사 가능성, 추적가능성, 설명가능성을 비롯한 적절한 감독 메커니즘을 확립해야 한다. 이러한 평가는 다학문적 · 다자적 · 다문화적 · 다원주의적 · 포용적이어야 한다.

96. Member States should involve all actors of the AI ecosystem (including, but not limited to, representatives of civil society, law enforcement, insurers, investors, manufacturers, engineers, lawyers, and users) in a process to establish norms where these do not exist. The norms can mature into best practices and laws. Member States are further encouraged to use mechanisms such as regulatory sandboxes to accelerate the development of laws and policies in line with the rapid development of new technologies and ensure that laws can be tested in a safe environment before being officially adopted.

96. 회원국은 존재하지 않는 규범을 확립하는 과정에 (시민사회 및 법 집행기관의 대표자, 보험사, 투자자, 제조업자, 공학자, 변호사, 사용자를 포함하되 이에 국한되지는 않는) 모든 인공지능 생태계의 행위 주체가 참여하도록 해야 한다. 이 규범은 모범 관행 및 법률로 발전할 수 있다. 회원국은 신기술의 급속한 발전에 따라 법률 및 정책의 개발을 가속화하고 법을 공식적으로 도입하기 전 안전한 환경에서 시험해볼 수 있도록 규제 실험장 같은 메커니즘을 사용하도록 더욱 권장된다.

97. Member States should ensure that harms caused to

users through AI systems can be investigated, punished, and redressed, including by encouraging private sector companies to provide remediation mechanisms. The auditability and traceability of AI systems, especially autonomous ones, should be promoted to this end.

97. 회원국은 민간 기업이 손해배상 메커니즘을 제공하도록 장려하는 것을 비롯하여, 인공지능 시스템으로 인해 사용자에게 야기된 피해에 대하여 조사 · 처벌 · 배상이 이루어질 수 있도록 해야 한다. 인공지능 시스템 중에서 특히 자율적인 시스템의 경우, 이러한 목적을 위해 감사가능성 및 추적가능성이 증진되어야 한다.

98. Member States should apply appropriate safeguards of individuals' fundamental right to privacy, including through the adoption or the enforcement of legislative frameworks that provide appropriate protection, compliant with international law. In the absence of such legislation, Member States should strongly encourage all AI actors, including private companies, developing and operating AI systems to apply privacy by design in their systems.

98. 회원국은 국제법에 따라 적절한 보호를 제공하는 입법 프레임워크의 채택 또는 집행 등을 통하여, 개인 기본권에 대한 적절한 보호장치를 프라이버시에 적용해야 한다. 이러한 법률이 없는 경우, 회원국은 민간 기업을 비롯하여 인공지능 시스템을 개발 및 운용하는 모든 인공지능 행위 주체가 시스템의 설계에 프라이버시를 적용하도록 강력히 권장해야 한다.

99. Member States should ensure that individuals can oversee the use of their private information/data, in particular that they retain the right to access their own data, and "the right to be forgotten".

99. 회원국은 개인이 자신의 개인 정보 및 데이터의 사용을 감시할 수 있도록, 특히 개인이 자신의 데이터에 접근할 권리 및 '잊혀질 권리'를 보유하도록 해야 한다.

100. Member States should ensure increased security for personally identifiable data or data, which if disclosed, may cause exceptional damage, injury or hardship to a person. Examples include data relating to offences, criminal proceedings and convictions, and related security measures; biometric data; personal data relating to "racial" or ethnic origin, polit-

ical opinions, trade-union membership, religious or other beliefs, health or sexual life.

100. 회원국은 유출될 경우 개인에게 상당한 피해, 손상, 곤란을 초래할 수도 있는 개인 식별 데이터 및 일반 데이터의 보안을 강화해야 한다. 이러한 데이터의 예로는 범죄, 형사소송 · 전과, 유관 보안조치 데이터, 생체 데이터, 인종·민족적 태생, 정치적 견해, 노동조합가입여부, 종교 및 기타 신념, 건강 및 성생활과 관련된 개인 데이터 등이 있다.

101. Member States should work to adopt a Commons approach to data to promote interoperability of datasets while ensuring their robustness and exercising extreme vigilance in overseeing their collection and utilization. This might, where possible and feasible, include investing in the creation of gold standard datasets, including open and trustworthy datasets, that are diverse, constructed with the consent of data subjects, when consent is required by law, and encourage ethical practices in the technology, supported by sharing quality data in a common trusted and secured data space.

101. 데이터 집합의 호환성을 증진시키기 위해서 회원국은 데이터에 '커먼즈'(Commons) 접근 방식을 도입하기 위해 노력해야 함과 동시에, 데이터 집합의 견고성을 보장하고 이의 수집 및 활용을 감독함에 있어 극도로 주의해야 한다. 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 구성된 것을 말한다. 또한, 회원국은 믿음직하고 안전한 데이터 공간에서 양질의 데이터를 공유함으로써, 기술에서 윤리적 실천을 장려해야 한다.

V MONITORING AND EVALUATION
모니터링 및 평가

102. Member States should, according to their specific conditions, governing structures and constitutional provisions, monitor and evaluate policies, programmes and mechanisms related to ethics of AI using a combination of quantitative and qualitative approaches, as appropriate. Member States are encouraged to consider the following:

102. 회원국은 자국의 특정 조건, 지배 구조, 헌법 조항에 따라 양적·질적 접근법을 적절하게 조합하여, 인공지능 윤리와 관련된 정책 · 프로그램 · 메커니즘을 모니터링하고 평가해야 한다. 회원국에게는 이하의 항목을 고려할 것을 권장한다.

a. deploying appropriate research mechanisms to measure the effectiveness and efficiency of ethics of AI policies and incentives against defined objectives;

a. 인공지능 윤리 정책 및 분명한 목표에 대한 인센티브의 효율성 및 효과성을 측정하기 위한 적절한 연구 메커니즘을 배치함.

b. collecting and disseminating progress, good practices, innovations and research reports on ethics of AI and its implications with the support of UNESCO and international ethics of AI communities.

b. 유네스코와 국제 인공지능 윤리 공동체의 지원을 받아 인공지능 윤리 및 그 함의에 대한 진행 상황, 우수 사례, 혁신, 연구 보고서를 수집하고 배포함.

103. The possible mechanisms for monitoring and evaluation may include an AI observatory covering ethical compliance across UNESCO's areas of competence, an experience sharing mechanism for Member States to provide feedback on each other's initiatives, and a 'compliance meter' for developers of AI systems to measure their adherence to policy recommendations mentioned in this document.

103. 가능한 모니터링 · 평가 메커니즘에는 유네스코의 권한 범위 내 영역에서 윤리 준수를 책임지는 인공지능 윤리 관측

기구, 회원국이 서로의 계획안에 피드백을 제공하는 경험 공유 메커니즘, 인공지능 시스템 개발자가 본 안에 언급된 정책 권장사항의 준수 여부를 스스로 측정할 수 있는 '규정 준수 측정기'가 포함된다.

104. Appropriate tools and indicators should be developed for measuring the effectiveness and efficiency of polices related to ethics of AI against agreed standards, priorities and targets, including specific targets for disadvantaged and vulnerable groups. This could involve evaluations of public and private institutions, providers and programmes, including self-evaluations, as well as tracer studies and the development of sets of indicators. Data collection and processing should be conducted in accordance with legislation on data protection.

104. 합의된 기준, 우선순위, 대상자, 특히 빈민 · 취약 계층에 속하는 특정 대상자에 대하여, 인공지능 윤리에 관한 정책의 효율성 및 효과성을 측정하기 위해서는 적절한 도구 및 지표가 개발되어야 한다. 이는 자체 평가를 비롯한 공공 · 민간 기관, 공급자, 프로그램에 의한 평가 뿐만 아니라, 추적 연구 및 지표 개발을 수반할 수 있다. 데이터 수집 및 처리

는 데이터 보호에 대한 법률에 따라 이행되어야 한다.

105. Processes for monitoring and evaluating should ensure broad participation of relevant stakeholders, including, but not limited to, people of different age groups, persons with disabilities, women and girls, disadvantaged, marginalized and vulnerable populations, and respecting social and cultural diversity, with a view to improving learning processes and strengthening the connections between findings, decision-making, transparency and accountability for results.

105. 모니터링 및 평가 과정은 다양한 연령 집단, 장애인, 여성 및 소녀, 빈곤·소외·취약 계층 등을 포함하여 (단, 이에 국한되지 않는) 유관 이해관계자들의 광범위한 참여를 보장해야 하며, 학습 과정을 개선하고 결과에 대한 시사점 · 의사결정 · 투명성 · 책무성 간의 연계성을 강화하기 위하여 사회 · 문화적 다양성을 존중해야 한다.

VI UTILIZATION AND EXPLOITATION OF THE PRESENT RECOMMENDATION

현 권고안의 활용 및 이용

106. Member States should strive to extend and complement their own action in respect of this Recommendation, by cooperating with all national and international governmental and non-governmental organizations whose activities fall within the scope and objectives of this Recommendation.

106. 회원국은 본 권고안의 범위 및 목표 내에서 활동하는 국내 · 국제 정부 및 비정부기구와 협력함으로써, 본 권고안에 관해 자국의 행동을 확장 및 보완하기 위해 노력해야 한다.

107. Member States and stakeholders as identified in this Recommendation should take all feasible steps to apply the provisions spelled out above to give effect to the foundational values, principles and actions set forth in this Recommendation.

107. 본 권고안에서 확인된 회원국과 이해관계자는 본 권고안에 명시된 기초적 원칙·기준 · 시행사항을 실행하기 위하여, 위에서 상세히 기술한 조항들을 적용하기 위한 가능한 모든 조치를 취해야 한다.

VII PROMOTION OF THE PRESENT RECOMMENDATION
현 권고안의 홍보

108. UNESCO has the vocation to be the principal United Nations agency to promote and disseminate this Recommendation, and accordingly shall work in collaboration with other United Nations entities, including but not limited to the United Nations Secretary-General's High-level Panel on Digital Cooperation, COMEST, the International Bioethics Committee (IBC), the Intergovernmental Bioethics Committee (IGBC), the International Telecommunication Union (ITU), and other relevant United Nations entities concerned with the ethics of AI.

108. 유네스코는 본 권고안의 홍보 및 확산에 앞장서는 유엔 기구라는 사명이 있으며, 그에 따라 유엔 사무총장의 디지털 협력에 관한 고위급 패널, 세계과학기술윤리위원회(COMEST), 국제생명윤리위원회(IBC), 정부간생명윤리위원회(IGBC), 국제전기통신연합(ITU) 및 인공지능 윤리와

관련된 기타 유관 유엔기구를 포함하여 (단, 이에 국한되지 않는) 다양한 유엔 기구들과 협업할 것이다.

109. UNESCO shall also work in collaboration with other international organizations, including but not limited to the African Union (AU), the Association of Southeast Asian Nations (ASEAN), the Council of Europe (CoE), the Eurasian Economic Union (EAEU), the European Union (EU), the Organisation for Economic Co-operation and Development (OECD) and the Organization for Security and Co-operation in Europe (OSCE), as well as the Institute of Electrical and Electronic Engineers (IEEE) and the International Organization for Standardization (ISO).

109. 또한 유네스코는 아프리카연합(AU), 동남아시아국가연합(ASEAN), 유럽평의회(CoE), 유라시아경제연합(EAEU), 유럽연합(EU), 경제협력개발기구(OECD), 유엔안보협력기구(OSCE)와 더불어 전기전자기술자협회(IEEE), 국제표준화기구(ISO)와 같은 국제금융기관을 포함하여 (단, 이에 국한되지 않는) 다양한 국제 기구와 협업할 것이다.

VIII FINAL PROVISIONS
최종 규정

110. The Recommendation needs to be understood as a whole, and the foundational values and principles are to be understood as complementary and interrelated. Each principle is to be considered in the context of the foundational values.

110. 본 권고안은 그 전체로서 이해되어야 하며, 근본 가치 및 원칙은 상호보완적이며 서로 밀접한 관계가 있는 것으로 이해되어야 한다. 각 원칙은 이의 토대가 되는 가치의 맥락에서 고려되어야 한다.

111. Nothing in this Recommendation may be interpreted as approval for any State, other social actor, group, or person to engage in any activity or perform any act contrary to human rights, fundamental freedoms, human dignity and concern for life on Earth and beyond

111. 본 권고안의 어떤 내용도 국가나 기타 사회적 행위 주체, 집단, 개인이 인권, 근본적 자유, 인간 존엄성, 지구 안팎의 생명체에 대한 관심에 반하는 활동에 참여하는 것이나 그러한 행위를 하는 것을 승인한다는 뜻으로 해석될 수 없다.

유네스코 인공지능(AI) 윤리 권고 해설서
인공지능 윤리 이해하기

지은이 이상욱
펴낸이 한경구

펴낸날 2021년 12월 3일
펴낸곳 유네스코한국위원회
주소 서울시 중구 명동길(유네스코길) 26
전화 02-6958-4164
팩스 02-6958-4250
전자우편 science@unesco.or.kr
홈페이지 www.unesco.or.kr
디자인 디자인프리즘 02-2264-1728

ISBN 979-11-90615-22-8

비매품/무료
03500
9 791190 615228
ISBN 979-11-90615-22-8