2024 UNESCO
ISSUE BRIEF

유네스코 「인공지능 윤리 권고」 이행을 위한 한국의 ‘준비도 평가’ 및 ‘윤리 영향 평가’ 대비 현황과 전망

UNESCO
ISSUE-
-BRIEF

2024년 제1호
유네스코 이슈 브리프

unesco
Korean
National Commission
유네스코한국위원회
1954-2024
외교부

유네스코 이슈 브리프는
유네스코와 관련된 다양한 주제에 대한
정책 제언 및 논의 확산을 위해
유네스코한국위원회가 발간하며,
집필자의 의견은 유네스코한국위원회의
공식 입장과 다를 수 있습니다.

이 글은 원고 중간 발표회에서 제시된 여러 의견을
참고하여 집필자가 작성하였습니다.

중간 발표회 2024년 6월 18일
발표 | 이상욱 (한양대학교 철학과 교수)
토론 | 김명주 (서울여자대학교 정보보호학과 교수)
토론 | 문아람 (정보통신정책연구원 디지털사회전략연구실 연구위원)

UNESCO ISSUE BRIEF

유네스코「인공지능 윤리 권고」 이행을 위한 한국의 '준비도 평가' 및 '윤리 영향 평가' 대비 현황과 전망

유네스코 「인공지능 윤리 권고」 이행을 위한 한국의 ‘준비도 평가’ 및 ‘윤리 영향 평가’ 대비 현황과 전망

이 상 욱
한양대학교 철학과 교수

Ⅰ. 유네스코 「인공지능 윤리 권고」(2021)의 의의

2021년 11월 유네스코 총회를 통과하여 국제적으로 공표된 유네스코의 「인공지능 윤리 권고」(이하 「권고」)는 인공지능 관련 국제 윤리 논의와 거버넌스 논의에서 중요한 기여를 하였다(UNESCO 2021). 유네스코의 「권고」 이전에도 유럽연합(EU) 과 경제협력개발기구(OECD)등 영향력 있는 국제 행위자들은 인공지능을 설계하고 활용하는 과정에서 지켜야 할 윤리 원칙에 대한 선언문을 발표했지만, 유네스코의 「권고」는 몇 가지 점에서 앞서 나온 선언문들과는 뚜렷한 차별성을 지닌다. 또한 「권고」가 나온 이후의 인공지능 윤리 관련 국제 흐름을 고려할 때 유네스코가 그 흐름을 예견하고 선도했다고도 볼 수 있다.

유네스코 「권고」의 가장 두드러진 차별성은 2021년까지 발표된 인공지능 윤리 관련 국제 문서 중 가장 포괄적이었다는 데 있다. 기존 국제문서는 불확실성이 큰 인공지능 기술의 성장 잠재력을 고려하여, 인공지능이 인류 복지에 기여할 잠재력을 지니지만 우리 사회의 핵심 가치를 위협할 수 있다는 점을 인정하고 그 위험에 대비하고 대응하자는 일종의 '원칙 선언'에 가까웠다. 그에 비해 유네스코의 「권고」는 인공지능의 '전 주기적whole life-cycle' 과정에 관련되는 모든 이해관계자들이 고려해야 할 다양한 쟁점을 제시함으로써 실천적 의의와 정책적 적용가능성을 높였다. 개별 인공지능 기술이 아니라 인공지능 '시스템'이라는 용어를 사용한 점과, 다른 문서에서는 그 범위의 모호성 때문에 잘 사용되지 않는 '인공지능 행위자AI actor'라는 포괄적 실천 대상을 설정한 것이 이런 경향을 잘 보여준다. 또한 투명성이나 설명 가능성처럼 통상적으로 강조되던 가치 이외에도 다양성이나 생태계 번영과 같은 참신한 가치를 제시한 점과 인공지능 윤리에 대한 교육의 중요성을 강조한 점도 주목할 만하다.

흥미로운 점은 「권고」 이후에 이루어진 국제 인공지능 거버넌스의 흐름에서 유네스코가 새롭게 강조한 다양성, 포용성, 교육의 중요성이 보다 광범위하게 인정되고 있다는 사실이다. 생성형 인공지능의 등장 이후 인공지능은 이제 막연하게 느껴지는 첨단 기술에서 개인이 실생활에서 직접 체험해 볼 수 있는 검색 엔진이나 휴대폰 애플리케이션 같은 지위를 갖게 되었다. 그러면서 자연스럽게 다양한 사용자 집단의 특성을 고려하여 인공지능이 설계되고 활용되는 것의 중요성이 부각되기 시작했다. 현재는 인공지능 윤리 논의에서 '다양성'의 가치는 투명성이나 설명 가능성보다 더 자주 언급되고 있다. 이런 추세를 반영하듯 2024년 5월 한국과 영국이 서울에서 공동주최한 2024 인공지능 정상회의

는 2023년 영국 블레츨리 파크에서 개최되었던 1차 회의에서 강조한 안전만이 아니라 다양성과 혁신을 함께 강조하기도 했다.[1]

교육은 「권고」가 주목했던 영역 중에서도 「권고」 발표 이후에 인공지능 국제 거버넌스에서 보다 많은 관심을 받은 영역이다. 물론 교육은 유네스코의 핵심 주제이기에 유네스코가 인공지능 윤리에서 교육의 역할에 주목한 것은 지극히 자연스러운 일이었다. 그런데 「권고」는 당시까지 강조되던 인공지능 개발자에 대한 교육뿐만 아니라, 인공지능 사용자와 인공지능 관련 정책입안자 및 정부 관료들도 폭넓게 인공지능 리터러시 교육을 받아야 한다는 점을 강조했다. 이 부분은 현재는 절대적인 시의성을 지니지만 2021년 11월에는 규범적으로 다소 지나치게 거창하다는 지적을 받을 여지가 있었다. 당시까지 인공지능 윤리 논의에서는 인공지능을 편향되지 않고 '신뢰 가능하도록' 잘 만드는 것이 가장 중요하다고 여겨졌고, 사용자나 정책입안자에 대한 교육의 필요성은 그다지 강조되지 않았기 때문이다.

하지만 2022년 11월에 출시된 생성형 오픈에이아이Open AI의 인공지능 챗지피티ChatGPT가 제시한 '환각'과 '독성'의 문제는 사용자의 인공지능 리터러시가 얼마나 현실적으로 필요한 역량인지에 대해 국제 사회가 주목하도록 만든 계기가 되었다. 마찬가지로 생성형 인공지능, 특히 초거대 기초모형foundation model 기반 생성형 인공지능이 사회에 끼칠 수 있는 잠재적 위험에 국제 사회가 보다 주목하면서, 그전까지 강조되던 '신뢰 가능한' 인공지능에 더해 '안전한' 인공지능이 강조되고 인공지능의 안전을 담보할 수 있는 각종 제도적 장치의 중요성에 국제 사회가 주목하면서 정책입안자에 대한 인공지능 윤리 및 리터러시 교육의 필요성도 커졌다.

이런 점을 고려할 때 유네스코의 「권고」는 인공지능 윤리와 국제 거버넌스 구축 과정에서 분명히 의미 있는 영향을 주었다고 볼 수 있다. 이에 더해 「권고」는 그 내용이 구체적인 정책 행동으로 이어질 수 있도록 회원국의 노력을 촉구하면서, 이를 돕기 위한 국제 협력과 더불어 두 종류의 평가를 제시했는데, 그것이 바로 준비도 평가Readiness Assessment와 윤리 영향 평가Ethical Impact Assessment이다. 이어지는 두 절에서는 각각 두 평가에 대해 2021년 11월 이후 이루어진 후속 작업과 국내 대응 현황을 살펴본다.

1 https://www.youtube.com/watch?v=q7PkxuA806Q, https://aiseoulsummit.kr/

Ⅱ. 준비도 평가의 국제 현황 및 국내 대응

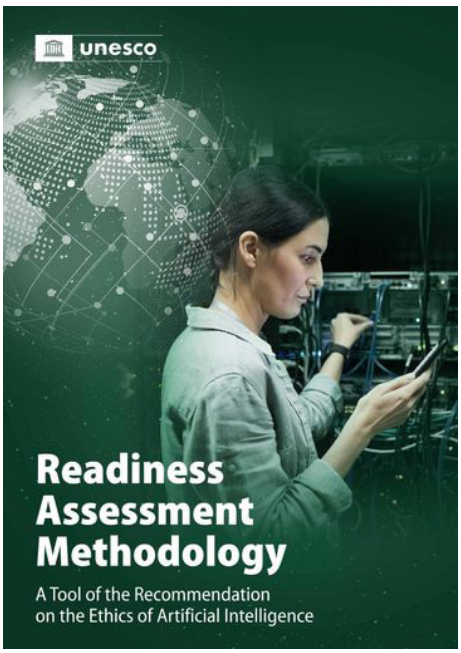
유네스코는 2021년 공표한 「권고」의 후속 작업으로 '준비도 평가'와 '윤리 영향 평가'를 제시했고 2023년 「준비도 평가 방법론(Readiness Assessment Methodology: A Tool of the Recommendation on the Ethics of Artificial Intelligence)」을 발표했다. 영어 제목에서도 알 수 있듯이 보고서는 「권고」가 내용을 제도적으로 실현하기 위한 '측정 도구'로서 발표된 것이다(UNESCO 2023b).

준비도 평가의 시작점은 「권고」 초안에 대한 의견 수렴 과정에서 유네스코 회원국들이 제기한 질문이었다. 회원국들은 「권고」 초안의 전반적인 내용에 대해 공감하였고, 추상적인 윤리 원칙만이 아니라 구체적인 '정책 행동'을 통해 인공지능 윤리가 실천되어야 한다는 데에도 큰 틀에서 동의했다. 하지만 많은 회원국들은 최종적으로 합의될 「권고」가 요구할 인공지능 윤리에 대한 정책 행동들을 실천하기에 제도적으로 준비가 안 되어 있다는 점을 지적했다. 간단하게 말해서 인공지능 윤리와 관련된 여러 실천들의 필요성에는 공감하지만, 정부의 행정력이나 재원 부족 등의 문제로 이를 실천하기에 여러 어려움이 있다는 점을 지적한 것이다. 일부 회원국들은 근본적으로 「권고」의 원칙적 내용을 각국의 사회문화적, 제도적 환경에 적합한 구체적인 정책 행동으로 어떻게 '번역'하는 것이 바람직한지에 대해서도 판단하기 어렵다고 토로하기도 했다.

이런 점을 고려하여 준비도 평가는 「권고」가 제시한 방향으로 윤리적인 인공지능 생태계를 발전시켜 나가기 위해 필요한 제도적, 규제적 조치 중에서 각국의 구체적 맥락에서 어떤 영역의 어떤 측면이 '비어 있는지'를 찾아낼 수 있는 측정 방법론을 제공한다. 그리고 이런 의미에서 준비도 평가는 다음 절에서 설명할 「윤리 영향 평가(Ethical Impact Assessment: A Tool of the Recommendation of the Ethics of Artificial Intelligence)」를 보완한다. 준비도 평가와 윤리 영향 평가 모두 2021년 발표된 「권고」의 정책적 실효성을 높이기 위한 도구라는 점에서는 공통점을 갖는다. 그러나 윤리 영향 평가가 보다 미시적 수준에서 특정 인공지능 기술의 윤리적 영향을 평가하는 방법론을 제시하고 있다면, 준비도 평가는 회원국 수준에서 제도적 윤리 대응 능력을 보다 거시적 관점에서 평가한다고 볼 수 있다.[2]

2 https://www.unesco.org/ethics-ai/en/ram

그림1 유네스코가 2023년 발표한 「준비도 평가 방법론」 표지



「준비도 평가 방법론」 보고서는 이 측정 도구에 대한 배경 설명을 제공하는 1장과, 준비도 평가 도구를 어떻게 사용할 것인지를 안내하는 2장, 그리고 본격으로 각 국가의 「권고」 이행 준비 정도를 측정하는 질문을 담은 3장으로 구성되어 있다.

본문에 해당되는 3장은 다시 5개의 일반적 질문을 담은 1절과 법적, 사회·문화적, 과학·교육적, 경제적, 기술·기반 시설적 차원 관련 점검 사항을 다루는 5개의 절이 모여 총 6개의 영역으로 구성되어 있다. 통상적으로 우리나라에서는 과학과 기술을 묶어 하나의 개념인 '과학기술'로 다루는 경우가 많은데, 유네스코는 대개 과학과 기술 사이의 차이점을 분명하게 구별하여 다루는 경우가 많다. 그런데 인공지능과 관련해서는 인공지능 관련 과학 연구와 공학 연구를 구별하는 것이 윤리적 논의 맥락에서 그다지 실효성이 있지 않다고 간주하여 인공지능 관련 '학술 연구' 모두를 '과학 연구'로 묶고, 인공지능 시스템이 작동하기 위해 필요한 기반 시설 관련 기술적 사항만을 별도의 절을 두어 점검하고 있다는 점이 특이하다.

보다 구체적으로 준비도 평가 내용을 살펴보자. 3.1절의 일반 질문은 우선 '당신 정부는 인공지능 국가위원회를 설립하거나 혹은 그 밖의 다른 방식으로 유네스코의 「권고」를 이

행할 계획을 갖고 있는가?'부터 시작한다. 이 질문은 준비도 평가가 「권고」의 이행을 촉진하기 위해 만들어졌음을 고려하면 자연스럽다. 그런데 이어지는 질문인 '당신 정부는 인공지능과 관련하여 정부의 개입(예를 들어, 규제, 전략, 가이드라인 등)을 통해 혜택을 얻을 수 있는 특정 영역과 관련된 우선순위를 마련했는가?'는 흥미롭다. 이는 준비도 평가에서의 「권고」 이행이 인공지능의 윤리적 개발 자체만이 아니라 윤리적 실천을 강화하는 정부의 노력을 통해 인공지능 연구 개발 및 관련 산업 발전이 촉진될 가능성을 염두에 두고 있다는 점을 시사한다. 유네스코 회원국 중 많은 기술 저개발국들이 인공지능 기술을 타 기술 선진국들을 순식간에 따라잡을 수 있는 도약 기술leapfrogging technology로 간주하고 있다는 점이 이 질문과 연결될 수 있다.

3.1절의 세 번째 질문은 정부가 시민에게 공공 서비스를 제공하는 과정에서 시민들의 유형화categorize 여부와, 자동화된 결정을 수행하는 인공지능 시스템 사용 여부를 알려주는지를 묻고 있다. 이 질문은 인공지능 활용에서 '투명성transparency'과 '책무성accountability' 가치와 관련된다. 또한 이 질문은 시민들에게 도움을 주는 것이 비교적 명백한 상황에서도 인공지능이 사용되고 있다는 사실을 고지하는 것이 필수적이라는 점을 강조하고 있다고 해석될 수 있다. 네 번째 질문은 정부의 어떤 부처에서 인공지능 거버넌스의 책임을 지고 있는지를 묻고 있는데, 이는 반드시 인공지능 국가위원회가 아니더라도 인공지능 거버넌스 전반을 책임질 수 있는 정부 기관이 설립 및 운영되는 것이 바람직하다는 「권고」의 내용과 관련된다. 마지막 질문은 '당신의 국가에서 인공지능 규제와 정책을 만드는 데 있어 가장 중요한 도전 과제는 무엇인가?'를 묻고 있는데, 이는 준비도 평가가 전반적으로 강조하는 국소성, 즉 「권고」의 실천 과정에서 개별 국가의 구체적 맥락을 고려하는 것이 바람직하다는 준비도 평가의 정신을 잘 보여준다.

3.2절은 준비도 평가의 법적 차원을 판단하기 위해 8가지 지표를 점검한다. 8가지 지표는 인공지능 정책과 규제, 데이터 보호와 사생활 법, 데이터 공유와 접근 가능성, 데이터 획득 관련법과 정책, 정보의 자유 법안 및 지식 접근 법안, 데이터 처리와 책무성, 온라인 안전과 언론의 진실성, 공공 영역 역량이다. 이 절은 전체적으로 준비도 측정 대상 국가가 정보 처리, 사생활 보호, 정보 접근성 및 언론 출판의 자유 및 진실성 확보 등에 있어서 관련 법률의 제정 여부와 해당 사안에 대해 판단하고 처리할 수 있는 공공 기관의 역량을 평가한다. 전반적으로 이 지표들은 인공지능 윤리 거버넌스가 '법의 지배Rule of Law' 체계 내에서 작동해야 한다는 「권고」의 지향점을 담고 있다고 해석될 수 있다.

3.3절은 준비도 평가의 사회·문화적 차원을 판단하기 위해 5가지 지표를 점검한다. 5가

지 지표는 1) 다양성, 포용성 그리고 평등, 2) 공공 참여와 신뢰, 3) 환경 및 지속가능성 정책, 4) 건강 및 사회적 복지, 5) 문화이다. 각각의 지표가 다루는 영역이 매우 포괄적이지만 지표의 내용을 살펴보면 「권고」의 가치, 원칙, 정책 행동에서 강조되었던 내용을 고스란히 담고 있음을 알 수 있다.

3.4절은 준비도 평가의 과학 및 교육적 차원을 판단하기 위해 10가지 지표를 점검한다. 10가지 지표는 연구 및 혁신 관련 5가지 지표와 교육 관련 5가지 지표로 나뉜다. 연구 및 혁신 관련 지표는 연구개발 지출, 연구 성과, 윤리적 인공지능 관련 연구, 인공지능 인재, 혁신 성과로 구성되어 있으며, 전체적으로 준비도 평가 대상 국가의 인공지능 관련 연구 역량 및 성과를 평가할 수 있는 내용으로 구성되어 있다. 교육 관련 지표는 교육 전략, 교육 기반, 교과 내용, 교육적 성취, 인공지능 교육에 대한 공공 접근성이다. 이 지표들은 준비도 평가 대상 국가의 전반적인 교육의 제도화 수준에 기반하여 인공지능 교육이 얼마나 잘 이루어지고 있는지, 인공지능 관련 인재들은 충분한지 등을 평가한다고 볼 수 있다.

3.5절은 준비도 평가의 경제적 차원을 판단하기 위해 3가지 지표를 점검한다. 3가지 지표는 노동 시장, 중간intermediate 소비, 투자와 생산이다. 이 지표들은 인공지능에 한정하지 않고 준비도 평가 대상 국가의 전반적인 경제력을 판단할 수 있는 내용으로 구성되어 있다. 이는 인공지능 관련 윤리적 제도화를 위해 일정한 수준의 공공 자원이 동원되어야 하고 이를 위해서는 상당한 경제적 기반이 조성될 필요가 있다는 점에 주목했기 때문으로 풀이된다.

마지막으로 3.6절은 준비도 평가의 기술·기반 차원을 판단하기 위해 4가지 지표를 점검한다. 4가지 지표는 기반 및 연결성, 응용 표준, 계산 역량, 통계적 수행이다. 이 지표들은 인공지능이 연구·개발되고 활용되기 위한 기술적 사회 기반 환경이 얼마나 잘 조성되었는지를 평가하는 것이다. 특히 최근 생성형 인공지능의 광범위한 활용으로 세계적으로 주목을 받고 있는 계산 역량을 지표로 포함시킨 점이 흥미롭다.

종합적으로 볼 때 준비도 평가는 「권고」를 이행하기를 원하는 유네스코 회원국 중 희망 국가를 대상으로 각 국가의 고유한 상황적 특징 및 인공지능 윤리 관련 가용 예산의 정도 등과 같은 현실적인 측면을 평가하는 도구라고 볼 수 있다. 유네스코는 준비도 평가를 홈페이지에 소개하면서 국가별로 「권고」 내용을 잘 알고 있는 유네스코 사무국 직원, 해당 국가의 유네스코국가위원회, 해당 국가 대표자 및 학계 대표자, 민간 영역 대표자를 포함

하는 다양한 이해관계자가 참여하는 평가 팀에 의해 평가가 진행될 것이라고 설명하고 있다. 평가 결과와 평가에 포함된 설문 답변을 조합하면 국가의 준비 정도를 평가한 분석 보고서가 된다. 특히 이 분석 보고서에는 해당 국가의 다양한 영역(지표 차원)에서의 역량 수준만이 아니라 「권고」가 제시하는 인공지능 윤리를 성공적으로 실천하기 위해 '부족'하다고 판단되는 지표들을 어떻게 개선할 것인지에 대한 제안도 담기게 된다. 결론적으로 준비도 평가는 「권고」에서 강조되었던 유네스코 회원국의 인공지능 윤리 관련 역량 강화를 위한 국가 전략에 대해 컨설팅을 제공하게 된다.[3]

유네스코 홈페이지에 따르면 현재까지 준비도 평가를 받은 유네스코 회원국은 브라질, 칠레, 모로코, 세네갈 총 4개국이며, 이에 더해 아프리카 지역 18개국, 아랍 지역 3개국, 아시아-태평양 지역 10개국, 유럽과 북아메리카 지역 8개국, 라틴 아메리카 지역 12개국에서 평가가 진행 중이다. 평가를 마쳤거나 아직 진행 중인 국가 중에서 네덜란드 정도를 제외하면 기술 선진국이라고 볼 수 있는 나라는 없다. 네덜란드조차도 인공지능 기술 선도 국가라고 보기는 어렵다. 흔히 인공지능 분야에서 선도국가로 여겨지는 미국, 중국, 그리고 그 뒤를 추격하는 캐나다, 영국, 프랑스 그리고 아시아권에서 한국, 일본, 싱가포르 등은 적어도 현재까지는 평가를 받지 않고 있다. 아마도 여러 이유로 이들 국가가 유네스코의 준비도 평가를 받을 가능성은 높지 않을 것이다.

한 가지 주된 이유는 이들 국가는 인공지능 기술만이 아니라 인공지능 거버넌스에서도 나름대로 선도적인 역할을 수행하고 있거나 적어도 분명한 전략적, 정책적 입장을 갖고 있다는 점을 들 수 있다. 앞서 설명했듯이 준비도 평가는 「권고」의 내용을 이행하기 위해 필요한 국가별 전략 수립 및 거버넌스 구축에 도움을 주기 위해 마련되었다. 그런데 인공지능 기술 및 거버넌스에서 이미 선도적 위치를 점하고 있는 나라 입장에서는 준비도 평가가 불필요하다고 판단하거나, 좀 더 현실적으로 말하자면 거추장스럽다고 느낄 수 있다. 평가 결과, 자신들이 자체적으로 추진 중인 인공지능 국가 전략이나 거버넌스 구조와 어긋나는 결론이 도출될 수도 있기 때문이다.

한국은 올해 5월에 열린 제2차 인공지능 정상회의를 즈음해서 인공지능 기술 3대 강국을 목표로 하겠다는 전략을 발표했다. 인공지능 관련 여러 정책과 전략을 조정할 수 있는 인공지능 국가위원회도 올해 안으로 설립할 예정이다. 이 국가위원회는 연구개발, 산업,

3 https://www.unesco.org/en/articles/evaluating-national-ai-readiness-government-ai-readiness-index

윤리 및 안전 등 정부의 인공지능 관련 다양한 정책 행동을 모두 아우르고 조정하는 역할을 맡게 될 것으로 기대된다. 이런 상황을 고려할 때 한국이 이 단계에서 유네스코의 준비도 평가를 받아야 할 절실한 이유를 찾기는 어려워 보인다.

Ⅲ. 윤리 영향 평가의 국제 현황 및 국내 대응

유네스코는 2023년에 「윤리 영향 평가(Ethical Impact Assessment: A Tool of the Recommendation of the Ethics of Artificial Intelligence)」도 발표했다(UNESCO 2023c). 윤리 영향 평가는 준비도 평가보다 「권고」의 이행과 관련해서는 직관적으로 훨씬 더 관련이 높은 측정 도구라고 할 수 있다. 「권고」를 이행하기 위해서는 유네스코 회원국은 자국에서 연구개발 중이거나 활용 중인 인공지능 시스템이 어떤 윤리적 영향을 끼치는지를 우선 판단해야 한다. 그리고 그 판단 결과를 「권고」의 내용에 비추어 평가하고 어떤 점이 부족한지를 찾아내서 그에 대한 대응 방안을 모색하고 실천해야 한다. 그런 의미에서 윤리 영향 평가는 회원국의 「권고」 이행 노력에 결정적인 도움을 줄 수 있는 도구로 기대된다.

윤리 영향 평가 역시 준비도 평가와 마찬가지로 여러 질문을 통해 특정 종류의 인공지능 기술이 끼치는 윤리적 영향을 평가한다. 「윤리 영향 평가」 보고서의 서술은 우선 이 질문들의 범위를 설명하는 범위 지정 질문Scoping Questions 4개의 장과 유네스코가 「권고」에서 밝힌 윤리 원칙 중에서 영향 평가에서 중요하게 사용된 원칙을 점검하는 7개의 장으로 이루어져 있다.

우선 범위 지정 질문에서는 1장에서 윤리 영향 평가의 취지를 설명하고, 2장에서 비례적 선별proportional screening과 해악 금지 원칙을 설명한다. 3장 사업 운영구조에서는 평가 대상 인공지능 기술과 관련된 여러 이해당사자들의 역할과 책임을 설명한다. 4장에서는 「권고」에서도 여러 차례 강조된 다중 이해관계자 운영구조multi-stakeholders governance를 설명한다.

이어지는 7개의 장에서는 「권고」에서 언급된 10대 윤리 원칙 중에서 앞서 설명된 질문의 범위에 포함되지 않은 원칙들을 정리하고 있다. 우선 5장 안전과 보안, 6장 공정성; 비차별; 다양성, 7장 지속 가능성, 8장 사생활 및 데이터 보호, 9장 사람에 의한 감시 및 결정, 10장 투명성과 설명 가능성; 책무성과 책임성, 11장 인식과 리터러시와 관련된 질문이 제시된다. 각 장의 제목과 내용으로 판단할 때 윤리 영향 평가가 점검하는 윤리 원칙은 기본적으로 「권고」의 10대 윤리 원칙이라는 점을 알 수 있다. 대부분은 「권고」의 윤리 원칙을 그대로 사용하여 윤리 영향 평가 원칙으로 사용하고 있고 일부는 표현을 살짝 바꾸거

그림2 유네스코가 2023년 발표한 「윤리 영향 평가」의 표지

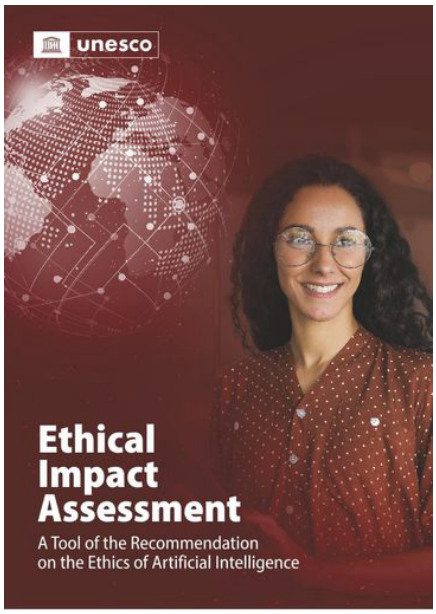
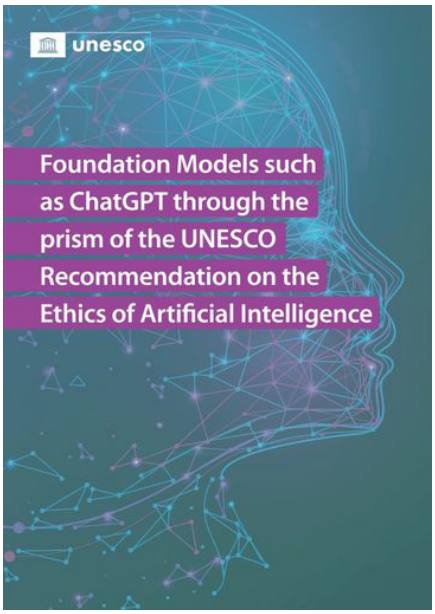


그림3 유네스코가 2023년 발표한 생성형 인공지능 관련 보고서 표지



나(“Right to Privacy”를 “Privacy”로), 두 원칙을 합쳐서 하나로 만들거나(10장), 보다 포괄적인 ‘질문 범위’ 관련 내용으로 따로 분류하는 방식을 택한다.

윤리 영향 평가는 인공지능 시스템이 일반 시민에게 공개되어 사용되기 전에도 사전 평가의 일환으로 활용될 수 있고, 사용된 후에 사후 평가로서 파급 효과를 가늠할 때도 활용될 수 있다. 하지만 두 경우 모두 핵심은 「권고」가 강조한 것처럼 인공지능 시스템이 어떤 윤리적 위험을 제기하는지, 그러한 위험을 막기 위해서 인공지능 시스템의 설계 단계부터 어떤 조치가 취해졌는지를 투명하게 평가함으로써 보다 신뢰 가능하고 안전한 인공지능 시스템을 만들고 사용하자는 것이다. 이런 취지를 고려할 때, 인공지능 시스템의 설계 단계에서 윤리 영향 평가를 시행하고 보다 바람직한 방식으로 시스템의 설계·제작·활용이 이루어지는 것이 가장 이상적일 것이다. 하지만 이미 활용되고 있는 인공지능 시스템에 대해서도 사후적으로라도 그 위험을(이 경우에는 이미 실현된 위험을 포함하여) 평가하고 드러난 문제점에 대해 효과적이고 윤리적인 사후적 대응을 추진하는 데 평가 결과를 활용할 수도 있다. 유네스코는 윤리 영향 평가 홈페이지에서 생성형 인공지능이 이러한 사전 검토를 제대로 거치지 않고 일반 시민을 상대로 ‘실험적 배포’를 진행하는 행위의 문제점을 강조하고 있다.

유네스코는 2022년 11월 등장하여 세계적인 관심사로 떠오른 챗지피티와 같은 생성형 인공지능 및 기초모형에 대해 「권고」의 관점에서 추가로 언급해야 할 필요성을 느낀 것으로 보인다. 그 결과물이 2023년 6월에 발간한 「유네스코의 인공지능 윤리 권고의 시각에서 바라본 ChatGPT와 같은 기초모형(Foundation Model such as ChatGPT through the prism of the UNESCO Recommendation on the Ethics of Artificial Intelligence)」이라는 짧은 보고서이다(UNESCO 2023a). 이 보고서를 포함하여 윤리 영향 평가는 「권고」의 파급 효과를 높이려는 유네스코의 지속적인 노력의 산물이라고 볼 수 있다.

필자가 파악하기에는 현재 우리 정부는 인공지능의 윤리적 영향 평가와 관련하여 크게 세 방향에서 정책적 노력을 기울이고 있다. 첫째는 과학기술정보통신부가 정보통신정책연구원(KISDI)을 통해 수행하고 있는 인공지능 윤리 영향 평가 프레임워크를 개발하고 시험·평가하려는 시행계획이다. 정보통신정책연구원의 윤리 영향 평가 프레임워크는 2023년 11월 초안이 발표되었고, 각계 의견 수렴을 거쳐 최종본이 2024년 발표되었다. 인공지능 기술이 아닌 서비스를 중심으로 윤리 영향 평가를 실시하고 유네스코의 윤리 영향 평가를 참조하되 우리나라가 기존에 발표한 인공지능 관련 여러 제도적 노력(윤리 원칙, 자율점검표 등)을 반영하려고 노력한 점이 눈에 뜨인다.[4] 정보통신정책연구원은 2024년 중으로 정부와 협의를 거쳐 인공지능 윤리 영향 평가 프레임워크를 완성한 후, 시범 사례로 특정 인공지능 서비스를 지정하여 영향 평가를 진행할 예정이다.

보다 구체적으로 살펴보자면, 정보통신정책연구원의 프레임워크 연구팀은 윤리 영향 평가 프레임워크의 목적을 "(1) 인공지능 윤리·신뢰성 실천을 위한 기업의 자율적 노력을 지원하고, 사용자가 인공지능을 윤리적이고 주체적으로 활용하기 위한 기준 제시, (2) 인공지능 제품·서비스의 윤리적 영향력을 사전에 평가함으로써 긍정적 영향 극대화 및 부정적 영향 최소화를 위한 관리·제도·정책적 조치 방안 등 시사점 도출, (3) 인공지능의 윤리적 영향력을 체계적으로 파악할 수 있는 참고 자료를 기업, 시민사회, 학계, 공공부문(정부) 등에 제공하고, 인공지능 제품·서비스를 보다 윤리적인 방식으로 개발·배포·활용하도록 장려"하는 것으로 밝히고 있다. 프레임워크는 정부를 시행 주체로 명시하여 인공지능 제품·서비스군을 대상으로 "(1) 국가 인공지능 윤리기준 10대 핵심 요건에 대한 인공지능 제품·서비스의 영향, (2) 윤리·신뢰성 측면에서 인공지능 제품·서비스의 긍정적 영향과 부정적 영향, (3) 적용 가능한 긍정적 영향 촉진 전략 및 부정적 영향 완화 전략"의

4 https://www.youtube.com/watch?v=wFgsrfNbPWs

3대 평가 요소를 중심으로 윤리 영향 평가를 진행하게 된다. 평가 체계는 평가 대상 선정, 기초 분석, 윤리 영향 평가, 정책제언·보고서(작성)의 4단계로 이루어진다(정보통신정책연구원 2024).

이상에서도 알 수 있듯이 정보통신정책연구원이 개발한 윤리 영향 평가 프레임워크나 올해 2024년 추진 중인 시범 평가 시행체계는 유네스코의 윤리 영향 평가의 특징을 상당 부분 준용하되, 우리나라에서 2020년부터 독립적으로 진행되던 인공지능 윤리 거버넌스의 흐름과 정책 수요를 반영하여 만들어졌음을 알 수 있다. 이런 의미에서 정보통신정책연구원이 현재 진행 중인 인공지능 윤리 영향 평가는 유네스코의 「권고」 이행의 대표적인 사례라고 할 수 있다.

두 번째 흐름은 한국과학기술기획평가원(KISTEP)이 과학기술기본법에 따라 매년 실시하는 기술영향평가technology assessment의 2024년도 평가대상 기술로 신뢰-안전 인공지능 기술이 선정된 것과 관련이 있다. 이는 인공지능 기술에 대해 '윤리' 영향 평가만을 시행하는 것이 아니라 과학기술 기본권에 명시된 다양한 항목, 예를 들어 경제적, 사회적, 법적, 환경적, 젠더 특성적 측면을 다양하게 평가하게 된다. 현재 평가가 진행 중이기에 평가가 구체적으로 어떻게 이루어질지에 대해서는 분명하게 공표된 내용이 없지만, 전문가 중심의 도구적 기술영향평가 방법론과 일반 시민이 참여하는 담론 구성적, 사회참여 학습적 기술영향평가 방법론이 복합적으로 활용될 것으로 예상된다.[5] 기술영향평가의 구체성과 논의 결과의 실효성을 높이기 위해 뇌 인공지능, 의료 인공지능, 휴머노이드 인공지능처럼 일반 시민이 실생활에서 해당 파급 효과를 직관적으로 이해하기 쉬운 기술을 대상으로 평가를 추진 중이다.

한국과학기술평가원이 수행하는 기술영향평가와 별개로 한국과학기술단체총연합회 산하 젠더혁신센터에서도 2024년 젠더 다양성을 중심으로 인공지능 기술에 대한 기술영향평가를 실시한다. 특히 젠더혁신센터의 기술영향평가는 인공지능 기술을 대상으로 젠더 다양성을 비롯한 다양성과 포용성에 초점을 둔 대안적 기술영향평가방법론도 함께 개발할 예정이다.[6]

---

5 현재 KISTEP 홈페이지에는 2024년도 기술영향평가 시민포럼 시행주체에 대한 입찰 공고가 게시되어 있다. https://www.g2b.go.kr:8081/ep/invitation/publish/bidInfoDtl.do?bidno=20240712428&bidseq=00&releaseYn=Y&taskClCd=5 참조.

6 젠더혁신센터가 추진하는 포용적 혁신과 기술영향평가의 접점에 대해서는 https://www.youtube.com/watch?v=7eX8PJCBafw 참조.

세 번째 흐름은 지능정보화 기본법 제56조가 권고하여, 인공지능을 비롯한 지능정보화 서비스에 대해 실시하는 사회영향평가이다. 이 또한 큰 흐름에서 유네스코 「권고」가 규정한 윤리 영향 평가의 일환으로 볼 수 있다. 인공지능 서비스에 대한 사회영향평가의 일환으로 한국지능정보사회진흥원(NIA)은 2023년 인공지능 서비스 제공자와 전문가를 대상으로 설문 및 심층인터뷰 방식의 보건의료와 금융 분야의 인공지능 서비스에 대한 시험평가를 수행했다. 한국지능정보사회진흥원은 2024년에는 교육과 교통 분야의 인공지능 서비스를 대상으로 일반시민까지 대상을 확대하여 설문과 심층인터뷰 방식의 사회영향평가를 수행할 계획이다.

마지막으로 윤리 영향 평가와 직접적으로 관련되어 있지는 않지만, 넓은 의미에서 「권고」의 이행 관련 활동에서 두드러진 역할을 수행하는 곳은 당연하게도 유네스코한국위원회이다. 유네스코한국위원회는 인공지능 윤리 관련 다양한 교육 콘텐츠를 만들어 소책자나 유튜브를 비롯한 다양한 매체를 통해 알리고, 「권고」가 강조하는 여러 윤리 원칙과 쟁점을 소개하며, 인공지능 윤리와 관련된 여러 학술 활동과 다자간 협력도 강화하고 있다.

이상에서 알 수 있듯이 한국의 인공지능 윤리 영향 평가는 유네스코의 윤리 영향 평가를 참고하는 수준에서 2020년부터 독자적으로 진행된 것으로 보는 게 적절하다. 또한 「권고」를 이행하며 인공지능 윤리 논의를 반영하고 연구개발 및 활용 방안을 모색해 온 제도적 경험에 근거한 방식으로 이루어지고 있다는 점을 알 수 있다. 다만 다음 절에서 논의할 인공지능 관련 국제 거버넌스의 최근 동향을 고려할 때 윤리 영향 평가에 대한 우리의 독자적 대응이 어떤 방식으로 이루어져야 할지에 대해서는 보다 복잡한 고려가 필요해 보인다.

Ⅳ. 인공지능 국제 거버넌스 현황과 「권고」 대응 전략[7]

인공지능 국제 거버넌스의 최근 동향 논의에서 2022년 11월 공개된 오픈에이아이의 챗지피티는 빠질 수 없다. 생성형 인공지능은 그전에도 존재했지만, 매개변수 개수를 엄청나게 늘려서 만든 초거대 거대언어모형Hyper-scale Large Language Model은 수행 가능한 과업의 범위와 품질에서뿐만 아니라 그것이 끼칠 수 있는 잠재적, 실질적 위험에 있어서도 세계적으로 큰 충격을 주었다.

그전까지 인공지능 국제 거버넌스는 인공지능 규제 법안을 준비 중이던 유럽연합과 혁신을 강조하던 미국으로 양분되어 있었고, 두 접근의 차이를 좁힐 여지는 많지 않아 보였다. 하지만 2023년에 접어들면서 적어도 초거대 생성형 인공지능에 관한 한 기업의 자율규제에만 맡기는 것은 치명적인 위험이 많다는 인식이 세계적인 인공지능 기술 선도국 사이에서 공유되기 시작했다. 미국도 상원 청문회를 통해 인공지능 빅테크 기업에 적절한 안전조치를 주문하기 시작했고, 미 백악관은 빅테크 기업에 미국의 기술적 주도권을 계속 유지하면서도 안전하게 인공지능을 개발하고 활용할 수 있는 방안을 주문한 것으로 알려졌다. 유럽연합은 준비 중이었던 「인공지능 법안」에 생성형 인공지능 범주를 새롭게 만들고, 안전 연구소 설립 방안을 비롯한 법안 통과 후의 후속 조치에 대해서도 공을 들이기 시작했다. 대부분의 국내 기업들도 인공지능 기술 개발에 있어 기업의 자율규제 촉구 등 정부 차원의 개입은 혁신을 저해한다고 목소리를 높였다. 그러나 2023년 미국 행정부가 인공지능 거버넌스 관련 행정명령을 발표하고, 2024년에는 유럽의회에서 강력한 처벌 조항을 담은 「인공지능 법」이 통과되면서, 국제 거버넌스에 일정한 규제 흐름이 상수로 자리 잡고 있다.[8] 따라서 현재 인공지능 기술과 서비스를 개발하고 제공하려는 글로벌 기업들은 이런 규제 흐름에 대응할 수 있는 윤리 역량을 갖추어야 할 필요성

---

7 이 절은 2024년 6월 18일 유네스코한국위원회에서 열렸던 본 이슈 브리프의 중간발표에서 귀중한 논평을 해 주신 서울여대 김명주 교수님과 정보통신정책연구원의 문아람 박사님 덕분에 크게 보완되었다. 이 점에 대해 두 분께 감사드린다.

8 예를 들어 미 행정부는 2022년 <AI Training Act>(정식 명칭은 '인력 획득 및 기타 목적을 위해 인공지능 훈련 프로그램을 수립하거나 다른 방법으로 제공하도록 행정관리예산국장에게 요구하는 법률')를 통과시켰다. 이 법은 미국 행정부가 인공지능 규제만이 아니라 인공지능 활용에 있어서도 매우 적극적으로 거버넌스를 확립하려는 노력을 기울이고 있음을 잘 보여준다. 관련 내용은 https://www.congress.gov/bill/117th-congress/senate-bill/2551 참조.

을 느끼고 있다.

이런 배경에서 필자는 2023년 영국이 주도한 제1차 인공지능 안전성 정상회의AI Safety Summit와 올해 한국과 영국이 공동 주최한 제2차 인공지능 정상회의에 주목할 필요가 있다고 생각한다. 두 회의에서 산출된 합의문은 법적 구속력을 가지지도 않고, 그 내용도 원론적 수준에서 크게 벗어나지 않으며, 인공지능 기업의 자율적인 노력을 강조하였다. 하지만 중요한 점은 이 정상회의가 실제로는 각국 행정부 간의 논의에 그치지 않고 세계적으로 인공지능 개발을 선도하는 빅테크 및 한국 기업들도 대거 참여하여 합의 가능한 국제 인공지능 거버넌스를 모색한 회의였다는 점이다.

필자는 개인적으로 이 추세는 앞으로도 지속될 가능성이 높다고 본다. 인공지능 기술을 선도하는 빅테크 입장에서 보면 자신들 정도만 감당할 수 있는 높은 규제 장벽은 (일단 규제가 아예 없는 것보다는 못하지만) 후발 주자들의 추격을 따돌릴 수 있는 '사다리 걷어차기' 전략으로 활용될 수 있다. 그런 치밀한 계산이 아니더라도 생성형 인공지능의 등장으로 여러 사회적 쟁점과 위험성이 보다 분명해진 현 상황에서 어차피 규제가 도입될 것이라면 국제 규제 거버넌스의 논의 단계에 처음부터 참여하여 빅테크 기업들이 수용 가능한 내용과 형식으로 도입되도록 유도하는 것이 경영의 불확실성을 줄이는 똑똑한 대비책일 수 있다. 아마도 오픈에이아이를 비롯한 여러 인공지능 관련 주요 기업들이 '과도한 규제'에는 여전히 분명한 반대 입장을 밝히면서도 '적절한 수준'의 규제가 필요하다는 점에는 큰 틀에서 이견을 달지 않는 데는 이런 배경이 있을 것이다.

이와 같은 인공지능 국제 거버넌스의 현황을 고려할 때, 한국 정부의 유네스코 「권고」 이행과 관련된 정책 행동은 어떤 방향이 바람직할까? 섣불리 단정하기는 어렵지만 적어도 두 가지는 분명해 보인다.

첫째, 우리는 「권고」의 이행에 있어서는 어떤 기준으로 평가해도 비교적 모범적인 국가로 평가될 수 있다는 사실이다. 이는 「권고」 제정 작업과 거의 비슷한 시기에 국내에서 정부 주도의 인공지능 윤리 원칙에 대한 논의가 시작되었고, 자율점검표나 신뢰 가능한 인공지능에 대한 인증처럼 관련 후속 조치도 차근차근 진행된 덕분이라고 할 수 있다. 그런 이유로 앞서 지적했듯이 우리는 준비도 평가를 따로 받을 이유가 크지 않고, 윤리 영향 평가도 유네스코가 제시한 방향과 대체적으로 일치하는 방식하에 독자적으로 추진 중이라는 점을 내세울 수 있다. 유네스코의 윤리 영향 평가와 별도로 2003년부터 과학기술기본법에 의해 시행되어 온 기술영향평가의 2024년 대상 기술이 신뢰-안전 인공지능 기술

이라는 점을 「권고」에 대한 우리 정부의 이행 실적에 포함시킬 수 있다.

둘째, 최근 인공지능 국제 거버넌스 현황을 고려할 때 우리의 인공지능 윤리 노력은 이제 원칙을 제시하거나 평가 도구를 제공하는 수준을 넘어서, 유네스코가 말하는 '인공지능 행위자'들이 윤리 역량을 높일 수 있도록 도움을 주는 역할로 발돋움하도록 관련 정책 행동을 적극적으로 추진해야 한다는 것이다. 예를 들어 정부에서 현재 추진 중인 인공지능 서비스에 대한 윤리 영향 평가는 평가 후 점수 부여보다는, 유네스코의 「준비도 평가 방법론」과 「윤리 영향 평가」 보고서에서 강조하듯이 기업이 보다 효율적으로 점차 강화되는 인공지능 국제 규제 환경에 대응할 수 있도록 부족한 지점과 보완 방안을 제시하는 '도우미' 역할을 해야 한다. 궁극적으로 중요한 것은 평가 결과 자체가 아니라 그 결과의 환류를 통해 한국의 전반적인 인공지능 윤리 역량이 높아지고 시민의 인공지능 리터러시 수준이 올라가는 것이기 때문이다.

또한 현재 국회에서 추진 중인 「인공지능 기본법(가칭)」의 내용에 유네스코의 윤리 영향 평가 내용과 더불어 세계은행의 디지털 정부 준비도 평가(Digital Government Readiness Assessment)의 내용을 반영할 필요가 있다(World Bank 2020, 2022a, 2022b). 이는 우리의 규제 제도가 국제적으로 이미 형성되고 있는 거버넌스의 큰 틀을 따라가되, 우리나라에만 존재하여 한국 기업이 상대적으로 더 높은 규제 비용을 지불해야 하는 추가적인 규제 요소 도입에는 신중할 필요가 있다는 점과도 관련된다. 글로벌 인공지능 기술 및 서비스 시장을 염두에 두어야 하는 우리나라 기술 산업 환경을 고려할 때, 현재 빠르게 형성되고 있는 국제 인공지능 (윤리) 거버넌스의 핵심적 내용을 우리 법과 제도에 적절하게 반영하는 것은 우리 기업의 인공지능 윤리 역량 및 제도 역량 강화에 분명 도움이 될 것이다. 또한 유네스코의 준비도 평가와 세계은행의 준비도 평가를 활용해서 우리가 현재 어떤 부분이 부족하고 보완이 필요한지를 판단할 수도 있을 것이다.

마지막으로 윤리 영향 평가나 준비도 평가 자체와 직접적으로 관련이 있는 것은 아니지만, 유네스코의 「권고」 이행 과정에서 인공지능 리터러시를 현재 우리나라에서 이해되고 있는 것보다 보다 포괄적이고 미래지향적으로 규정하고, 이에 대한 교육과 훈련을 실천할 필요성에 주목해야 한다. 신기술에 대한 수용성이 비교적 높은 우리나라 교육계는 인공지능, 특히 생성형 인공지능이 교육 환경에서 갖는 장점에 주목하고 이를 적극적으로 도입하려는 다양한 노력을 펼치고 있다. 물론 미래 시민의 핵심 역량에서 인공지능과 같은 강력한 범용적 기술을 능숙하게 활용하는 능력은 무척 중요할 것이다. 앞으로의 직업이 요구하는 다양한 직능 중 상당수가 인공지능의 활용을 필수적으로 전제할 가능성이

많기에 더욱 그러하다. 하지만 많은 국제기구의 관련 보고서가 지적하듯 인공지능의 교육적 활용은 그것의 한계에 대한 정확한 이해와 그 결과물에 대한 비판적 검토 능력을 함께 교육할 때만 의미가 있다(UNESCO 2024). 이 점은 유엔이 2024년 3월에 채택한 <인공지능 결의(AI Resolution)>에서도 강조되고 있다(UN 2024). 그러므로 우리나라의 유네스코 「권고」의 교육 분야 이행에서 바람직한 미래에 대한 성찰에 기반을 둔, 보다 포괄적인 인공지능 리터러시 교육에 대한 로드맵 제시와 실천이 이루어져야 한다.

참고문헌

· UN 2024, *AI Resolution*.

· UNESCO 2021, *Recommendations on the Ethics of Artificial Intelligence*.

· UNESCO 2023a, *Foundation Models such as ChatGPT through the prism of the UNESCO Recommendations on the Ethics of Artificial Intelligence*.

· UNESCO 2023b, *Readiness Assessment Methodology: A Tool of the Recommendation of the Ethics of Artificial Intelligence*.

· UNESCO 2023c, *Ethical Impact Assessment: A Tool of the Recommendation of the Ethics of Artificial Intelligence*.

· UNESCO 2024, *Guidance for Generative AI in Education and Research*.

· US Government 2022, *Artificial Intelligence Training for the Acquisition Workforce Act*.

· World Bank 2020, *Digital Government Readiness Assessment*, Version 1.

· World Bank 2022a, *Digital Government Readiness Assessment*, Version 2 Survey Toolkit.

· World Bank 2022b, *Digital Government Readiness Assessment*, Version 2 Web Toolkit.

· 정보통신정책연구원 2024, 『인공지능 윤리영향평가 프레임워크』

MEMO

MEMO

유네스코 이슈 브리프 – 2024년 1호

유네스코 「인공지능 윤리 권고」 이행을 위한 한국의 '준비도 평가' 및 '윤리 영향 평가' 대비 현황과 전망

기 획	유네스코한국위원회
지 은 이	이상욱
편 집	김은영 백영연 김초연
발 간 일	2024년 8월 9일
펴 낸 곳	유네스코한국위원회
디 자 인	수카디자인
주 소	서울특별시 중구 명동길(유네스코길) 26
전자우편	ap.center@unesco.or.kr

간행물 등록번호
IR-2024-RP-5

유네스코 이슈 브리프는
외교부의 지원으로 발간되었습니다.

www.unesco.or.kr

유네스코 이슈 브리프

UNESCO ISSUE BRIEF

비매품/무료
94300
9 791190 615570
ISBN 979-11-90615-57-0
ISBN 979-11-90615-56-3 (세트)