UNESCO ISSUE BRIEF

2022년 제2호

유네스코 이슈 브리프

유네스코 「인공지능 윤리 권고」 이행과 국제협력

unesco
유네스코한국위원회

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 다양한 전문가의
의견을 참고하여 집필자가 작성하였습니다.

중간 발표회 2022년 8월 17일
발표 | 이상욱 (한양대학교 철학과 교수)
토론 | 김형준 (한국지능정보사회진흥원 지능화법제도팀 수석연구원)
토론 | 한귀영 (한겨레경제사회연구원 사회정책센터장)

2022년 제2호
유네스코 이슈 브리프

UNESCO ISSUE BRIEF

유네스코 「인공지능 윤리 권고」 이행과 국제협력

유네스코 「인공지능 윤리 권고」 이행과 국제협력

이상욱 (한양대학교 철학과)

Ⅰ. 서론

2005년 무렵을 전후로 인공신경망 기반 기계학습 알고리즘을 활용한 인공지능이 이미지 판독을 중심으로 인간의 능력에 근접하거나 어떤 경우에는 뛰어넘는 모습을 보여주기 시작했다. 이는 1950년대 인공지능(artificial intelligence)이라는 단어를 처음으로 제안하고 인공지능 연구의 활성화를 위해 다방면으로 노력했던 인공지능 분야의 초기 선구자에 해당되는 맥카시나 민스키 등이 상상했던 상황이다. 그들은 기계를 잘 고안된 알고리즘을 통해 학습시키면 (인간과 다른 방식이긴 하지만) 결과물에 있어서는 인간이 할 수 있는 일을 '기능적으로 동등하게' 할 수 있을 것이라고 믿었다. 이런 점을 볼 때 인공지능 초기 연구자들의 기대는 어느 정도 타당했던 것으로 보인다. (Mitchell, 2020)

그럼에도 불구하고 더글라스 호프스태터처럼 인공지능의 인간능력 '흉내내기'에 대해 상당한 기대와 동시에 분명한 한계가 있으리라 예상했던 컴퓨터 공학자조차 알파고가 이세돌 9단을 물리쳤을 때 충격을 받았다고 고백했다. 인간의 지능과 컴퓨터의 지능을 동일선상에 두고 이해할 수 있다고 믿는 기능주의(functionalism)의 대표적 신봉자인 호프스태터마저 이런 반응을 보인 것으로 보아, 알파고 수준의 인공지능이 이렇게 단기간 내에 개발될 수 있다는 점에 대해 인공지능 연구자들 사이에서도 의외라는 반응이 있었던 것으로 보인다. (Hofstadter 2008)

이런 '놀라움'의 이유는 인공지능 연구의 역사에는 '인공지능 겨울'(인공지능 Winter)로 지칭되는 어려운 시기가 여러 차례 있었기 때문이다. 이 '겨울'의 주된 원인은 인공지능의 기술 발전 속도에 대한 연구자들과 대중의 무리한 기대가 후속 연구를 통해 충족되지 않았기 때문이었다. 몇몇 연구자들이 인상적인 연구 결과를 제시하고 몇 년 내로 사람과 구별되지 않는 인공지능을 만들 수 있을 것이라는 낙관적인 전망을 대중매체를 통해 주장하다가, 결국에는 그 전망을 달성하지 못하자 사회적 관심이 급격하게 줄어들고 연구 지원이 끊기면서 인공지능 연구가 침체기를 겪게 되었던 것이다.

이런 역사적 경험에 따라 인공지능 학계는 2000년대에 매우 인상적인 성과를 내는 인공지능이 등장했음에도 조심스러운 태도를 취하지 않을 수 없었다. 상당히 낙관적인 인공지능 연구 전망을 펼친 학자들조차 인간의 지능을 뛰어넘는 초지능(superintelligence)이 제기하는 인류에 대한 '실존적 위험'(existential risk)에 대해서는 되도록 언급하지 않으려는 경향이 있다. 그 이유는 현재 인공지능 연구가 주로 특정 기능을 수행하는 '특수'(special) 인공지능에 집중되어

있는 가운데, 보편적 지적 능력을 갖추어야 하는 '일반'(general) 인공지능이 초지능에 도달하는 지점인 특이점(singularity)에 대한 섣부른 논의가 기존 인공지능 연구에 또 다른 '겨울'을 가져오지 않을까 우려하기 때문이다. 이는 초지능과 실존적 위험에 대한 일반 시민과 대중매체의 높은 관심과는 분명하게 대조되는 모습이다.[1]

한편 2010년대 이후 인공지능의 잠재력과 활용도가 수많은 분야로 확장되면서, 이제는 좁은 의미에서의 인공지능 연구자만이 아니라 공학, 자연과학, 사회과학, 심지어 예술이나 체육학 연구자들도 인공지능을 활용한 다양한 연구 결과를 쏟아내고 있다. 산업계에서의 활용 역시 매우 빠르고 광범위하게 진행되고 있어서 이미 우리는 수많은 인공지능의 도움을 받으며 일상생활을 하고 있는 상황이 되었다.

이렇게 글자 그대로의 의미에서 '인공지능의 시대'에 살고 있음에도 불구하고, 우리가 그 사실을 실감하기 쉽지 않은 이유가 있다. 우리가 일상생활에서 사용하는 인공지능은 휴대전화 같은 익숙한 장치에 '숨어' 있어서, 우리가 인공지능과 상호작용하고 있다는 사실조차 인지하기 어렵기 때문이다. 이러한 이유로 인공지능의 윤리적 측면을 다루는 국제 논의에서는 '투명성'(transparency)을 중요한 가치로 인식하고 있다. 즉, 인공지능과 관련된 다양한 윤리적·사회적 쟁점에 대한 일반 시민의 올바른 이해를 위해서는 먼저 우리가 어떤 상황에서 어떤 기능을 수행하는 어떤 종류의 인공지능과 상호작용하고 있는지를 사용자가 분명히 인식할 수 있는 형태로 알려주는 사용자 환경을 마련해야 한다는 권고이다.

이상에서 설명한 맥락이 유네스코가 2021년 총회에서 만장일치로 채택한 「인공지능 윤리 권고」의 제정 배경이라 할 수 있다. 즉, 유네스코는 인공지능이 현재 우리의 삶에 끼치는 영향이 중대하고 그에 대한 국제적 대응의 필요성이 절실하다는 판단을 하고 있다. 그간 유네스코는 조직의 활동 범위에 대한 윤리적 규범 틀을 선언(declaration)과 권고(recommendation), 협약(convention)의 형태로 제시해 왔는데, 인공지능과 같은 '특정' 과학기술 분야에 대해 윤리적 규범 틀을 제시한 경우는 거의 없었다. 국제적으로 통계 처리를 통일하기 위해 제정된 70년대의 몇몇 권고(예컨대 1970년에 발표된 「도서관 통계 표준화 선언」(Recommendation concerning the International Standardization of Library Statistics)가 있었고, 1997년 「인간 유전체와 인권에 대한 선언」(Universal Declaration on Human Genome and Human

[1] 초지능과 실존적 위험에 대한 다양한 입장에 대한 소개는 Brockman 2019을, 비판적 논의는 이상욱 2020 참조.

Rights)과 2003년 「인간 유전체 데이터에 대한 국제선언」(International Declaration on Human Genetic Data)이 발표된 적이 있지만, 이는 특정 과학기술에 초점이 맞추어져 있다기보다는 인간 유전체 연구가 인류의 기본적 권리 측면에서 제기하는 광범위한 위협에 대한 일반적 관심을 반영한 것이었다고 할 수 있다.[2]

유네스코는 유엔의 여러 산하 기구 중에서 교육, 과학, 문화에 집중하는 국제기구이다. 최근 유네스코는 교육, 과학, 문화의 '오래된' 주제에 더해 기후변화에 대응하는 '지속가능한 발전'(sust인공지능nable 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 펼치고 있으며, 이 과정에서 유네스코 사무총장이 임명하는 전문가로 구성된 과학기술윤리위원회(COMEST)와 국제생명윤리위원회(IBC)가 주도적인 역할을 수행하고 있다.

현재 국제적으로 한창 개발 중인 인공지능 기술에 대해서는 다양한 견해가 있고, 특히 인공지능이 제기하는 '위험'의 성격과 심각성에 대해서는 상당한 의견 차이가 있다. 그럼에도 미래 사회에서의 인공지능의 영향력이 현재 사회에서의 전기나 인터넷에 비견될 정도로 광범위하고 클 것이라는 점에는 모든 논자들의 견해가 일치한다. 이는 인공지능에 대한 사회적 관심이 관련 기술의 효율적 개발에만 국한되어서는 안 되며, 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.[3]

이는 유네스코의 인공지능 윤리 논의만이 아니라 EU와 OECD, IEEE 등 인공지능 윤리 논의를 수행하는 다양한 국제단체들이 일관되게 취하고 있는 입장이다. 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 반드시 우리 사회(국가적 수준과 국제적 수준 모두를 포함하여)가 소중하게 생각하는 핵심 가치(기본권 등)를 손상

[2] 유네스코가 제정한 국제 윤리적 틀 혹은 기준 설정 도구(standard-setting instrument)에 대해서는 https://www.unesco.org/en/legal-aff인공지능rs/standard-setting/recommendations?hub=66535 참조.

[3] https://ec.europa.eu/jrc/communities/sites/jrccties/files/eedfee77-en.pdf

하지 않는 방식으로 이루어져야 함을 일관되게 강조하고 있다. 그러므로 '사회 속의 혁신'을 인공지능 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, 그것이 고정된 것인지 변화될 수 있는 것인지, 그것이 서로 다른 문화에서 다르게 해석될 여지가 있는 것인지 아니면 전지구적으로 보편성을 가져야만 하는 것인지를 두고 논쟁의 여지는 있다. 얼핏 보면 서로 비슷하게 보이는 인공지능 윤리 관련 국제 논의들 사이에서도 자세한 내용에 상당한 차이가 있는 이유 역시 이러한 '사회적 가치의 테두리'를 정확하게 어떻게 해석할 것인지에 대해 의견 차이가 있기 때문이다.

이런 상황에서 유네스코가 2021년 11월 총회에서 채택한 「인공지능 윤리 권고」는 회원국의 다양한 의견과 상황을 반영하여 초안이 작성되고 수정된 결과물이라는 점에서 현재 국제적 수준에서 가장 대표성을 갖춘 인공지능 윤리의 규범적 틀이라고 할 수 있다. 우리나라도 유네스코 회원국으로서 이 권고의 내용을 이해할 의무를 갖고 있으며, 실제로 권고 채택 이전부터 다양한 방식으로 인공지능 윤리 관련 논의와 활동을 진행하고 있다. 본 이슈 브리프에서는 유네스코의 「인공지능 윤리 권고」 채택 이후의 후속 조치와 국내의 대응 현황을 살펴보고 특히 국제협력의 관점에서의 시사점을 논의하려 한다.

그 전에 2절에서는 「인공지능 윤리 권고」(이하 「권고」)의 내용과 특징을 살펴본다. 특히 그것이 유네스코 내에서 기획되고, 준비되고, 최종적으로 채택되는 과정과 그 과정에서 어떤 핵심적인 쟁점이 부각되었는지를 중심으로 해당 「권고」의 내용과 특징을 살펴봄으로써 3절 이후의 논의에 도움을 주고자 한다.

Ⅱ. 유네스코 「인공지능 윤리 권고」의 내용과 특징

1. 배경과 과정

유네스코에는 과학기술과 사회가 맺는 다양한 접점을 탐색하는 두 상설위원회인 국제생명윤리위원회(IBC)와 과학기술윤리위원회(COMEST)가 있다. 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.[4] 이런 취지로 COMEST가 주도해 최근에 공표된 것이 「기후변화 윤리 원칙 선언」(2018)이다. 유네스코의 규범적 틀/도구는 내용의 구체성과 강제력의 정도에 따라 선언(Declaration), 권고(Recommendation), 협약(Convention)으로 나뉘는데, 인공지능 윤리와 관련된 윤리적 틀은 2019년 유네스코 총회의 결정에 따라 권고로 추진되었다. 이는 인공지능 윤리의 규범성의 적절한 수준에 대해 유네스코 회원국 사이에서 상당히 구체적인 정책 행동을 요구하는 공감대가 형성된 결과로 이해할 수 있다.

유네스코가 인공지능 윤리에 관심을 갖게 된 이유는 여럿 있겠지만, 인공지능 관련 쟁점이 유네스코의 여러 중점 사업 분야나 핵심 주제와 깊이 연결되어 있다는 점도 중요했다. 인공지능 윤리 논의가 유네스코 내부에서 논의되는 과정에서도 이러한 사실이 지속적으로 강조되었다. 유네스코가 인공지능 윤리 권고(안) 작성 작업을 시작하기 위해서는 절차상 총회 의결이 필요했고, 총회에 이 안건을 상정하기 위해서는 해당 안건을 집행이사회가 먼저 통과시켜야 했다. 이때 집행이사회와 총회의 결정을 돕기 위해 COMEST 위원을 중심으로 인공지능 윤리 관련 전문가가 보강된 '확대전문가집단'(Extended Experts Group)이 꾸려져서 인공지능 윤리 관련 쟁점을 정리한 예비보고서를 만들었다.[5] 이 예비보고서 작성 과정에서 참여 전문가 위원을 제외하고 가장 많은 의견을 내고 피드백을 제공한 주체는 유네스코의 각 분야 담당자들이었다. 이처럼 유네스코 인공지능 윤리 논의는 유네스코의 기존 활동 및 각 분야의 관심사를 반영

[4] https://en.unesco.org/themes/ethics-science-and-technology/comest

[5] https://irc인공지능.org/wp-content/uploads/2020/07/PRELIMINARY-STUDY-ON-THE-ETHICS-OF-ARTIFICIAL-INTELLIGENCE.pdf

하는 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 집행이사회에 제출하고, 같은 해 가을에 열린 유네스코 총회에서 인공지능 윤리 권고 초안 작성 개시에 대한 의결이 이루어진 뒤, 회원국의 추천을 받아 이 작업을 수행할 비상설전문가집단(Ad Hoc Expert Group, 이하 AHEG)이 구성됐다. 총 24명으로 구성된 AHEG는 유엔의 6개 지역별로 각 4명씩 선발돼 구성됐으며, 이러한 구조적 특징은 각 전문가들이 자신의 국가를 '대표'하지는 않지만, 각 지역의 다양한 관심사와 다른 의견을 권고 초안에 반영하기 위한 것이라고 볼 수 있다. AHEG는 2020년 9월에 권고안을 완성하고, 이를 각국 대표들이 논의해 2021년 여름에 정부간 협의를 통해 총회에 상정할 최종 권고안을 결정했고, 이것이 2021년 11월 유네스코 총회에서 만장일치로 채택되었다.

AHEG 위원의 구성은 지역별 다양성뿐만 아니라 전문성에서도 다양성을 확보하려 노력한 흔적이 보인다. 인공지능 기술 전문가, 정책 전문가, 법률 전문가, 철학 전문가가 위원으로 포함되었고, 그 중에는 여러 영역을 가로지르는 전문성을 가진 위원도 많았다. 특히 유엔 차원의 다른 국제 논의에 참여한 경험을 가진 전문가들도 많아서 하루 두세 시간 이상 지속하기 어려운 온라인 회의의 한계에도 불구하고 다양한 논점 제시와 합의점 도출이 비교적 효율적으로 이루어질 수 있었다. 온라인 토론에서는 전문가들 사이에서 종종 상당한 의견 차이가 드러났지만, 유엔 기구의 전반적 관례를 반영하여 서로 합의할 수 있는 절충점을 찾는 방식으로 권고안의 문구가 합의 되었다.

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 진행되었다. 그럼에도 불구하고 위원들 사이에서 의견 차이가 컸던 여러 주제에 대해서는 완전한 합의점을 도출할 시간이 충분하지 않았다. 이런 상황에서 2020년 5월까지는 초안을 완성해야 했기에 AHEG는 본질적인 내용에서의 의견 차이에 대해서는 최종안 작성 과정에서 논의하기로 미루어 두는 — '주차해두자'(Let's park this!)라는 표현이 자주 사용되었다 — 경우가 많았다.

2020년 5월에 완성된 권고 초안에 대한 피드백은 온라인 의견 수렴 및 각 지역별 의견 수렴 회의를 통해 이루어졌다. 유네스코는 인공지능 윤리에 대한 유엔 수준의 대표성을 강하게 의식하고 있기에 이번 「권고」가 되도록 많은 의견과 되도록 많은 계층의 사람들의 이해관계를 반영하는 방식으로 작성되기를 원했다. 따라서 초안에 대한 의견 수렴 과정은 국제적 영역에서 누

구나 참여할 수 있는 온라인 의견 제시 과정, 그리고 보다 공식성을 갖춘 지역별 의견 수렴 회의의 두 갈래로 이루어졌으며, 이 단계에서 우리나라는 아·태지역 회의를 개최하였다. 이렇게 수렴된 의견은 유네스코 사무국에서 정리하여 AHEG에 전달했고, 이는 8월부터 시작된 최종안 작성 과정에서 적극적으로 활용되었다.

2020년 9월 유네스코 인공지능 윤리 권고안이 완성되기까지 다양한 의견들이 수렴되고 활용된 방식을 보면, 해당 권고안이 비록 지역별 전문가들의 노력을 통해 작성된 것이기는 하지만 순전히 전문가 위원회만의 의견이 반영된 결과물은 아니라는 점을 알 수 있다. 2021년 여름에 최종안을 두고 이루어진 공식적인 정부간 회의 이전에도 각 유네스코 회원국 대표부는 AHEG의 논의 과정에 옵서버로 참여해 매 온라인 회의 종료 후 서면으로 논의 중인 문건의 표현이나 내용에 대해 적극적으로 의견을 제시했고, AHEG 위원들은 다음 날 회의에서 사안별로 정리된 의견들을 논의한 뒤 이를 문건에 최대한 반영했다.

이러한 과정은 특히 2020년 8월부터 집중적으로 진행된 최종안 작성 단계에서 두드러지게 나타났다. 그간 수집된 수많은 의견들이 정리되고, 그에 더해 유네스코의 각 부문별 의견과 담당 부서의 의견까지 더해지면서, 매우 다양한 표현 및 내용 상의 제안이 고려되고 반영됐다. 특히 유네스코의 정보·커뮤니케이션 부분은 인공지능 및 빅데이터 활용과 관련해 이미 수행되고 있던 활동도 반영할 것을 요구했고, AHEG는 그 내용이 위원회에서 준비한 원래 문건의 내용과 잘 어울리지 않음에도 불구하고 이 요구사항을 최대한 반영하였다. 내용상의 긴장에 있어서는 정보·커뮤니케이션 부문에 비해 그 정도가 덜했지만, 2020년에 새로 취임한 가브리엘라 라모스 유네스코 인문사회부문 사무총장보를 중심으로 한 유네스코 사무국 담당부서 역시 요구사항을 적극적으로 개진했고 그 논의 결과가 최종안에 반영되었다. 특히 젠더 쟁점과 '행동가능한'(actionable) 정책 제시가 필요하다는 점에 대해서는 라모스 사무총장보의 의견에 위원들 사이에서 상당한 공감대가 있었고, 따라서 5월에 발표한 초안과 9월의 최종안 사이에는 '행동가능한' 정책 제시 부분에서 많은 변화가 있었다.

2020년 8-9월의 회의에서는 초안 작성 과정에서 위원들 간의 의견 차이로 '주차되었던' 주제들에 대해서도 집중적인 재검토가 이루어졌다. 앞서 지적한 의견수렴 내용을 반영하는 동시에 재검토를 통해 도달한 합의안을 반영하는 방식으로 최종안이 마련되었지만, 몇몇 민감한 주제들, 예컨대 권고(안)의 규범성 정도와 다른 인공지능 윤리 논의와의 역할 분담 문제 등에 대해서는 완전한 합의를 이룰 수 없었다. 따라서 반대 의견을 강하게 주장하는 소수 위원들이 해당 의견을 회의록에 남긴다는 전제로 최종안에 동의함으로써 논의가 마무리되었다.

이후 정부간 회의에서도 비슷한 패턴이 반복되었다. 그 결과 정부간 회의를 거친 최종안과 AHEG가 작성한 최종안 사이의 차이는 주로 정치적으로 민감할 수 있는 부분을 '조화'(harmony)시키는 지점에 집중되었고, 회원국들의 이해관계에 따라 권고의 규범성을 약화시키거나 (경우에 따라서는) 강화시키는 방식으로 이루어졌다.

2. 내용과 특징

2021년 11월 유네스코 총회에서 채택된 「권고」의 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 측면에서 상당한 구조적 변화가 있었다.

「권고」는 지향점과 핵심 주제를 제시하는 전문(preamble)과 이어지는 8개의 장으로 구성되어 있다. 1장은 이 권고의 활용(Scope and Application)목적과 범위를 설명하고 2장은 목적과 목표(인공지능ms and Objectives)를 다룬다. 이 두 장에는 초안에 대한 의견 수렴 과정에서 나왔던 여러 우려 — 유네스코의 인공지능 윤리가 각국의 인공지능 정책을 지나치게 제한하는 것이 될 수 있다는 우려 — 에 대한 대답과, 「권고」가 제안하는 권고 정책들이 각국의 상황을 고려하는 방식으로 활용되는 것이 바람직하다는 내용이 담겨 있다.

3장과 4장에는 초안에서와 마찬가지로 「권고」의 주요 내용이 담겨 있다. '가치와 원칙'을 다루는 3장에서는 초안에 대해 제기된 여러 고려 사항을 반영하여 가치와 원칙을 일부 통합하고 가치와 원칙의 규정 방식도 상당 부분 바뀌었다. 변화의 방향은 기본적으로 가치와 원칙의 내용이 더 잘 전달되고 더 보편성을 갖도록 함이었다. 즉, 초안에 제시되었던 가치와 원칙의 내용에 여러 지역별 의견수렴 과정에서 나온 의견을 종합하여 보다 많은 사람들이 이해하고 공감할 수 있는 방식으로 재구조화되었다.

그에 비해 4장은 제목 자체가 '정책 행동'(policy actions)으로 바뀐 것에서 알 수 있듯이 초안에서 정책 내용의 배경과 기술전문적인 내용을 상당부분 제거하고, 대신 회원국 정부가 정책 입안과 실행을 통해 구체적으로 행동에 나설 수 있는 방식, 즉 '행동가능한'(actionable) 방식을 중심으로 재구조화되었다.

5장부터 8장까지는 「권고」의 효율적 활용을 위해 필요한 부가 사항과 유엔 문서로서의 형식적 필요성을 위해 포함된 내용이다. 5장 '모니터링과 평가'에는 4장의 정책 행동을 회원국들이 성

공적으로 수행하기 위해 필요한 모니터링과 인공지능 윤리 평가 방식에 대한 제언이 담겨 있다. 6장은「권고」에 대한 해석과 활용이 회원국들이 공통적으로 수용하는 기본 인권 및 핵심 가치를 부정하는 방식으로 이루어져서는 안 된다는 점을 공식적으로 재천명한 것이다. 이는 회원국 정부 혹은 다른 단체가「권고」의 내용을 문맥 그대로 사용하지 않고 악용할 여지를 형식적으로 불허하기 위해 도입된 장이다.

이 부분에 대한 예로는 초안에 대한 정부간 회의 논의 과정에서 전체주의 역사 경험이 있는 독일을 비롯한 몇몇 회원국 대표가 '조화'(harmony)라는 개념이 역사적으로 전체주의의 인권 탄압을 옹호하는 방식으로 활용되었다는 점을 지적하며 3장에서 제거할 것을 요구했으나 AHEG 전문가 상당수와 아프리카 국가들이 적극적으로 이 가치를 유지할 것을 주장하면서 드러난 의견 충돌을 들 수 있다. 그 결과「권고」는 '조화'라는 개념을 사용하지 않고 생태계와 인간이 공동 번영하는 것의 중요성을 표현하는 절충안을 채택한 바 있다. 이와 같은 배경에서 6장은 '조화' 개념 사용을 두고 벌어졌던 일과 비슷하게 핵심 윤리 기념이 그 맥락을 떠나 악용될 수 있는 위험을 해소하기 위해 제시된 장이라고 볼 수 있다.

3장 '가치와 원칙'에서는 총 4가지의 가치와 10가지의 원칙을 제시했다. 초안에서 제시했던 가치나 원칙에 비해 그 수가 다소 줄었다고 볼 수 있지만, 이를 단순히 축소로 보기는 어렵다. 대신 서로 다른 가치와 원칙을 종합해서 새로운 가치를 만들고, 초안에서 미처 다루지 않았거나 가볍게 다루어졌다고 판단된 가치를 독립적 가치나 원칙으로 보다 전면에 내세우는 방식으로 전반적 재구조화가 이루어졌다고 보는 것이 맞는다.

예를 들어 초안에는 인류의 번영을 강조했지만 최종적으로 채택된「권고」는 인류와 그를 둘러싼 환경, 그리고 다른 존재들을 보다 적극적으로 고려하는 생태계 전체의 번영을 강조했다. 이는 인공지능 윤리를 지나치게 인간중심주의적으로만 서술하지 않고 환경과 생태계 전반에 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, 다양성을 포용가능성(inclusiveness)을 증진하는 방식으로 발휘함으로써 개인이 다양성 뒤에 숨어서 공동체로부터 오히려 배제되어 버리는 부작용을 막아야 한다는 생각이 반영되었다.

10개 원칙 중에는 가치 영역과 마찬가지로 초안의 내용에서 통합된 것도 있지만 '프라이버시'처럼 새로 도입된 것도 있다. 초안에서와 마찬가지로 가치와 원칙 사이의 관계는 도구성과 구

체성이다. 즉, 원칙은 가치를 보다 구체화하고 실현 가능하게 하기 위해 집중해야 할 영역을 지시한다고 볼 수 있다. 그런데 어떤 것이 본질적으로 추구해야 할 가치이고, 어떤 것이 그 가치를 실현하기 위한 구체적 영역인지에 대해서는 당연히 의견 차이가 있을 수밖에 없다. 이는 초안과 「권고」에서 가치에 있던 것이 원칙으로 이동하고, 원칙의 내용 중 일부가 가치의 설명에 포함되는 사례가 다수 발생했다는 점에서도 잘 드러난다.

예를 들어 '프라이버시'는 초안에서도 중요한 주제로 다루어졌지만 의견 수렴 과정에서 워낙 많은 사람들이 중요하게 언급한 내용이어서 정부간 회의안부터는 아예 독립적 원칙으로 강조하기로 의견이 모아졌고, 이는 최종 채택된 「권고」에도 그대로 반영되었다. 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.

「권고」의 원칙 중에 눈에 띄는 주제어는 '적응적 거버넌스'(adaptive governance)라는 개념이다. AHEG 내에서는 유네스코 인공지능 윤리가 제안하는 정책의 내용과 방식이 어떠해야 하는지를 두고 상당히 많은 격론이 있었다. 특히 모든 것을 강하게 규정할 것을 요구하는 위원들과, 인공지능 개발자나 회원국의 자율적 선택에 대부분을 맡기자는 위원들 사이에서 다양한 스펙트럼의 의견이 개진되었다. 초안 발표 후 의견 수렴 과정에서도 상당수 아프리카 국가는 유네스코 권고안의 규범성을 보다 강화해야 한다고 — 예컨대 'should' 대신 'must'를 사용해야 한다고 — 목소리를 높이기도 했다. 반면 영국 등의 인공지능 기술 선진국은 권고안의 규범력을 최소한으로 하고 문자 그대로 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 바로 '적응적 거버넌스'다. 우선 유네스코 「권고」가 의의를 갖기 위해서는 「권고」의 규범력을 손상시키지 않아야 한다는 데 거의 모든 위원들이 찬성했다. 「권고」가 그렇게 약화된 형태로 공표된다면 이미 나와 있는 여러 다른 인공지능 윤리 관련 선언이나 가이드라인과 차별성이 없다는 이유에서였다. 유네스코가 많은 노력을 기울여 새롭게 제안한 인공지능 윤리 권고가 기존 논의에 더하는 바가 아무 것도 없다면 그 노력 자체가 의의를 갖기 어렵다는 데 대다수의 위원들이 동의한 것이다. 그래서 정부간 회의에 제출된 AHEG의 최종안은 일부 'should'를 (그것이 타당하다고 합의한 경우에) 'must'로 전환하는 방식으로 규범성을 강화했다.

그와 동시에 위원들은 유네스코 회원국들의 제도적·법적·사회적 하부구조가 동일하지 않다는 점에도 주목했다. 즉, 규범적으로 유네스코 「권고」에 공감하는 회원국조차 이 「권고」의 내용을

실행하기에는 제도적·경제적으로 어려움이 많을 수 있다는 점을 공감한 것이다. 인공지능 기술의 특징상 앞으로 인공지능 기술이 어떤 방향으로 어떻게 개발되고 전체 사회에 영향을 끼칠 것인지에 대해서는 아직 불확실성이 크다는 점 역시 재확인했다. 이런 상황에서 지나치게 자세한 정책 권고를 하거나, 초기에는 타당할지라도 미래에는 타당하지 않을 수도 있다는 가능성을 「권고」에 반영하지 않는 것은 바람직하지 않다는 점에도 동의했다.

이런 합의점에 근거하여 AHEG 위원들은 우리가 제안하는 규범적 틀이 회원국의 국소적 환경의 구체적인 내용과 인공지능 기술의 개발 현황 및 미래 영향의 내용에 적극적으로 대응하는 방식으로 제도화될 필요가 있다는 점에 공감했다. 달리 말해 우리가 제안하는 인공지능의 국제 및 국내 거버넌스는 '적응적' 성격을 가져야 한다는 뜻이다. 이런 이유로 마지막 원칙은 이런 '적응적 거버넌스'가 다자주의적 고려와 협력에 입각하여 실천되어야 한다는 내용을 담게 되었다.

4장 '정책 행동'은 초안에서 다소 혼란스럽게 나열된 다양한 정책 행동을 그 정책이 효과를 발휘하도록 의도된 '영역'별로 재구조화되어 제시되었다. 중요한 것은 인공지능에 대한 '윤리 영향 평가'(Ethical Impact Assessment)가 그 첫 번째 영역이라는 점이다. 즉, 인공지능 윤리와 관련된 모든 정책 행동은 각 회원국이 국소적 환경에서 인공지능이 어떤 영향을 끼치고 있는지를 지속적으로 모니터링하고 평가하는 노력에서 출발해야 한다는 점을 강조한 것이다. 그렇게 해야만 두 번째 영역인 인공지능에 대한 윤리적 거버넌스와 돌봄이 성취될 수 있다.

이어지는 8가지 정책 영역은 인공지능 관련 정책 행동이 집중해야 할 대표적인 영역을 망라한 것들이다. 여기서는 인공지능 윤리에 대해 논의될 수 있는 모든 주제를 전부 포함시키기보다는 유네스코가 강점을 지닌 분야를 강조하는 방식으로 주제를 선정하였다. 예를 들어 마지막 정책 영역인 '건강과 사회적 복지'는 2020년 등장한 코로나19와 같은 공중 보건위기를 염두에 두고 추가된 것이다. 초안에도 인공지능 연구가 인류 복지에 이바지할 잠재력에 대한 언급은 여럿 있었지만, 최종안에서는 회원국의 일반 국민에게 가장 직접적으로 다가갈 수 있는 인공지능 윤리 쟁점을 포함시키자는 의도로 공중보건 영역에서의 인공지능 활용과 관련된 윤리적 쟁점을 따로 분리해서 제시했다. 비슷한 이유로 여러 영역에서 산발적으로 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 사이에서 우선순위를 염두에 두었다고 말하기도 어렵다. 보다 정확하게 말하자면 정책 영역 사이의 우선순위를 두는 것 자체가 필요한지에 대한 논쟁도 있었다. 여러 위원들이 우선순위를 두자는 의견을 제시했지만, 논의 과정에서 그 순위를 어떻게 두는 것이 좋을지에 대한 의견 일치가 어렵다

는 점이 분명하게 드러나면서 현재의 형태로 병렬적 제시를 택하게 되었다.

그렇다고 해서 영역 제시 순서에 아무런 의미가 없는 것은 아니다. 앞서 설명했듯이 첫 두 정책 영역은 「권고」의 지향점과 권고되는 정책의 효율성을 보장하기 위한 전제 조건을 담고 있다. 이어서 인공지능 윤리 논의에서 데이터 관련 쟁점이 가장 두드러지며 여러 주제에 걸쳐 있다는 사실에 주목해 '데이터 정책'을 세 번째 정책 영역으로 제시했다.

유네스코가 회원국의 자발적 협력을 강조하는 단체이며, 「권고」가 강조하는 '적응적 거버넌스'가 효과적으로 이루어지기 위해서는 회원국 사이의 국제협력이 무엇보다 중요하다는 점을 강조하기 위해 네 번째 정책 영역으로는 '발전과 국제협력'을 제시했다. 여기서 '발전'이란 인공지능 기술 발전만을 의미하는 것이 아니라 인공지능 윤리 역량의 발전도 함께 의미한다. 유네스코의 인공지능 윤리 권고는 유네스코 회원국 사이의 기술적 협력만이 아니라 윤리역량 제고를 위한 협력도 함께 강조하고 있다는 뜻이다.

그러므로 전체적으로 볼 때 「권고」의 정책 행동 구조는 각 회원국이 제시된 윤리 원칙과 가치를 구체적인 정책으로 실행하는 데 필요한 일종의 가이드로 작동할 수 있도록 구성되었다고 볼 수 있다. 유네스코 인공지능 윤리 권고를 채택하고 이를 실천하기 위해 기본적으로 어떤 제도적 장치(예를 들면 인공지능 윤리영향평가)를 시행해야 하는지, 그리고 어떤 핵심 영역에서 어떤 쟁점에 특별히 주의를 기울여야 하는지, 실천 과정에서 어떤 국제협력이 어떤 방식으로 회원국 사이에서 가능할지에 대해 설명하고 있는 것이다.

Ⅲ. 유네스코의 후속 조치 및 국내 대응

1. 유네스코의 후속 조치

「권고」가 실질적인 영향력을 발휘하기 위해서는 인공지능과 관련된 여러 사안에 대해 '바람직한' 윤리 원칙을 제시하는 것을 넘어서는, 각국 정부를 포함한 다양한 수준의 이해당사자(stakeholders)들이 실행에 옮길 수 있는 수준의 구체적인 정책 제안을 해야 한다는 점에 대한 광범위한 공감대가 「권고」 초안 작성 과정에서부터 형성되어 있었다. 그런데 구체적인 정책 제안을 어떻게 정식화할 것인지 고민하는 과정에서는 또 다른 현실적인 문제에 직면했다. 그것은 바로 회원국마다 정책적·제도적·경제적 환경이 매우 다르기에, 설령 「권고」가 제시하는 구체적인 정책적 제안을 원론적으로 수용하는 회원국이라도 그 내용을 실행할 수 있는 전문가 집단이나 제도적 장치, 행정력 등을 갖추지 못한 경우가 많다는 사실이었다.

구체적으로 쟁점이 된 내용이 바로 「권고」가 다른 국제 인공지능 윤리 선언들과 분명한 차별점을 보인 '인공지능 윤리 영향 평가' 부분이었다. 인공지능의 전주기적 윤리적 고려에 초점을 둔 유네스코 「권고」는 인공지능의 설계 단계부터 인공지능의 윤리적 영향을 평가하고 그에 따라 예상되는 부작용에 사전주의적으로 대응할 수 있도록 하는 윤리영향평가를 시행할 것을 권고하고 있다.[6] 이는 IEEE를 비롯한 많은 인공지능 연구자들이 말하는 '윤리적 설계 혹은 설계를 통한 윤리'(Ethics by Design) 개념과 일맥상통하는 것이지만, 윤리영향 '평가'라는 구체적 제도를 회원국에서 시행할 것을 권고하고 있다는 점에서 특별한 의미가 있다.

「권고」 초안을 완성하는 과정에서 유네스코 회원국 중에는 이 윤리영향평가를 하고 싶어도 역량 부족으로 할 수 없는 수많은 회원국이 있다는 점이 지적되었고, 국제협력을 통해 이를 해결할 방안을 모색해야 한다는 의견이 많은 공감대를 확보했다. 그러한 협력을 위한 준비 단계로

[6] 우리말의 '윤리'의 의미와 영어의 'ethics' 의미 사이에는 중요한 차이가 있는데, 이 차이점에 주목하지 않으면 '그런 내용을 왜 윤리 논의에서 다루느냐'는 식의 오해를 하기 쉽다. 따라서 우리말의 '윤리' 개념도 명확한 의미와 범위를 갖는 정당한 개념이지만, 국제적 인공지능 윤리 논의에서의 '윤리'는 'ethics'의 개념임을 명확하게 인식하는 것이 중요하다. 관련 논의는 이상욱 2021 참조.

우선 각 회원국이 「권고」의 내용을 이행하는 데 얼마나 '준비'가 되어 있는지를 평가하는 '준비 정도 평가'(readiness assessment)를 실시하기로 하였고, 2022년 8월 현재 국제적으로 이 '준비 정도 평가'의 방법론에 대한 전문가 설문이 진행되고 있다.

이와 별도로 인공지능 '윤리영향평가'의 방법론을 정하기 위한 전문가 집단이 2022년 초반에 유네스코 내에 만들어졌고 현재 평가 방법론에 대한 논의가 진행 중인 것으로 알려져 있다. 다만 이 평가 방법론이 모든 회원국에게 일률적으로 적용될 보편적 성격의 방법론이 될지, 각 회원국이 자국의 특수한 상황을 고려해 자체적으로 평가 방법론을 개발하는 데 도움이 될 '지침'의 성격을 가질지는 알려지지 않은 상태이다.

'준비 정도 평가'나 '윤리 영향 평가'와 달리 「권고」 초안 작성 단계에서 논의가 되었음에도 본격적으로 추진되지 않고 있는 제안은 유럽연합이 실시하고 있는 '관측소'(observatory)를 유네스코 회원국들을 대상으로 설치하고 운용하자는 것이다. 유럽연합은 2020년 고위 전문가 집단의 작업을 통해 인공지능 윤리원칙을 발표한 후 구체적인 시행 방안을 만들고 있는데, 여러 이해관계 조정이 쉽지 않아 최종안 발표에는 조금 더 시간이 걸릴 것으로 예상된다. 하지만 이 최종안과 독립적으로 인공지능의 위험도에 따라 다른 규제 방식을 택하는 인공지능 윤리 법안이 입법 과정 중에 있다. 따라서 유럽연합에서는 GDPR 규정 등 윤리적으로 민감한 정보 혹은 상황과 관련된 '고위험 인공지능'과 상대적으로 윤리적 민감도가 적은 '저위험 인공지능'에 대해 다른 규제를 적용하겠다는 큰 틀의 원칙은 정해진 것으로 보인다.

하지만 유럽연합의 이러한 큰 틀에서의 규제 원칙이 구체적으로 유럽연합 회원국의 국내법에 어떻게 적용될 것인지는 각국의 정치적·경제적·사회적 조건에 따라 상당한 차이가 있을 것으로 예상된다. 유럽연합의 '관측소'는 바로 이런 점을 고려하여 각국이 인공지능 윤리 관련 제도를 도입할 때 다른 유럽연합 국가의 사례를 참조할 수 있도록 하는 것으로, 유네스코 「권고」 초안 작성 과정에서도 이런 목적의 '관측소'를 운영하는 것이 좋겠다는 제안에 많은 전문가들이 공감했다. 특히 유네스코와 유럽연합의 차이점을 반영하여 유네스코 '관측소'는 법률과 같은 '강한' 제도적 장치를 도입하는 것뿐만 아니라 유네스코가 강조하는 인공지능 윤리 교육이나 다양성 증진 활동과 같이 인공지능 윤리와 관련된 회원국들이 자국의 다양한 실천 경험을 서로 공유함으로써 「권고」의 실행력을 높이자는 의도도 강조되었다. 아직까지는 유네스코 본부에서 이 '관측소' 관련 활동에 대한 움직임이 없지만, 이후 「권고」 이행 관련 국제 협력이 활발해지면 '관측소' 혹은 그와 유사한 제도가 도입될 가능성은 높다고 판단된다.

2. 국내 대응

엄격하게 말하자면 유네스코 「권고」에 대한 국내 대응은 이미 유네스코 「권고」 채택 이전부터 시작되었다고 볼 수 있다. 여기에 중요한 자극제 역할을 한 것이 2019년 발표된 「OECD의 인공지능 윤리원칙 선언」과 EU의 인공지능 윤리 관련 논의들이었다. 인공지능 기술의 잠재력과 위험성에 대한 국제 사회의 이러한 관심은 대중매체를 통해 널리 알려졌고, 국가 주도 과학기술 개발의 성공적 경험에 익숙한 우리나라 정부는 일찍부터 인공지능 기술의 경제적·산업적 영향력에 주목하여 이를 집중육성 대상 기술로 여기고 있었다. 이에 더해 2021년 초에 불거진 '이루다 사태' 등으로 인해 인공지능의 윤리적 측면에 대한 사회적 관심도 높아졌다.

그러므로 유네스코에서 「권고」 초안 작성 작업이 진행되던 2020년에도 국내에서는 인공지능 윤리 관련 기준안을 만들고자 하는 노력이 과학기술정보통신부와 정보통신정책연구원을 통해 진행되고 있었고, 이는 2020년 12월 인간을 중심에 둔 「인공지능 윤리 기준」으로 발표되었다.[7] 과학기술정보통신부는 마찬가지로 정보통신정책연구원을 통해 「인공지능 윤리 기준」의 실천적 활용을 높이기 위한 다양한 후속 작업도 2021년부터 시행하고 있는데, 이 과정에서 해당 작업을 유네스코 「권고」의 내용과 연계하는 등의 국제 협력의 가능성도 함께 모색하고 있다.

그 결과 2022년에 발표된 것이 「신뢰할 수 있는 인공지능 개발 안내서(안)」과 「인공지능 윤리 기준 실천을 위한 자율점검표(안)」이다. 이 두 안 모두 다양한 전문가와 산업계의 의견을 반영하여 작성되었으며, 모두 2020년에 발표된 윤리기준의 활용도를 높이면서 특히 인공지능 개발자들에게 직접적이고 효율적인 방식으로 도움을 주려는 의도로 개발되었다.

하지만 국내 산업계에는 '자율점검표'라는 명칭에도 불구하고 그 내용이 궁극적으로 인공지능 기술개발 및 산업화에 대한 정부 규제로 이어질 것이라는 의심이 퍼져 있고, 실제로 이런 이유로 두 안이 얼마나 실효성 있게 활용될지 여부는 아직 불확실한 상황이다. 따라서 두 안이 최종 확정되어 국내에서 인공지능 설계 단계에서부터 인공지능 윤리가 활발하게 논의되고 적용되기 위해서는 산업계 자율 규제 능력의 확보와 더불어 정부 정책에 대한 신뢰 확보 역시 필요할

[7] 발표 시점을 보면 과기정통부의 「인공지능 윤리 기준」은 유네스코 「권고」의 최종안이 마련되는 중에 공표되었음을 알 수 있다. 실제로 당시 관련 보도자료 역시 유네스코 「권고」 작업에 대한 언급 없이 OECD와 EU의 인공지능 윤리 관련 논의를 참고로 언급하고 있다.

것으로 예상된다. 이와 더불어 빠르게 변화하는 인공지능 기술 개발 현실에 적합한 형태로 '자율점검표'의 내용을 어떻게 구성할 것인지에 대한 기술적 고민도 함께 이루어져야 할 것이다.

이와 별도로 이용자 보호를 위한 가이드라인 제시 등과 같이 정보통신기술 사용자의 권익 보호를 위한 다양한 방안도 방송통신위원회를 중심으로 모색되고 있다. 이러한 이용자 보호 노력 역시 유네스코의 「권고」 작업으로부터 직접적 영향을 받았다기 보다는 독자적으로 평행하게 진행되었지만, 유네스코 「권고」 채택 이후 그 내용을 비롯한 여러 국제 논의를 여기에 반영하려는 노력은 꾸준히 진행되고 있다. 특히 국내의 정보통신 서비스 사업자와 플랫폼 기업들의 의견을 지속적으로 반영하여 현실적으로 효과를 낼 수 있는 지능정보기술(인공지능 기술을 포함한 보다 보편적인 개념)이용자 보호 방안이 모색될 것으로 기대된다. 예를 들어, 방송통신위원회가 정보통신정책연구원과 협력하여 2021년 6월에 발표한 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙」은 넷플릭스나 유튜브처럼 인공지능 기반 추천 시스템을 활용하는 기업들이 인공지능을 이용자의 권익을 보호하는 방식으로 활용하기 위해 고려해야 할 윤리 원칙을 제시하고 있다.[8]

마지막으로 유네스코한국위원회는 유네스코 「권고」 초안 및 최종안의 국문 번역을 제공하는 한편, 「권고」에서 강조하고 있는 여러 윤리적 쟁점을 일반 시민에게 보다 널리 알리는 교육 콘텐츠를 제작했다. 특히 인공지능 윤리의 여러 쟁점을 소개하는 동영상을 제작하여 유튜브를 통해 배포하고 관련 자료를 책자로 제작하여 배포하는 등, 유네스코 「권고」가 강조한 교육 관련 정책 행동을 충실하게 수행하고 있다.[9] 이번 이슈 브리프 역시 이런 유네스코한국위원회의 노력의 일환이라고 할 수 있다.

이처럼 현재까지 이루어진 국내의 인공지능 윤리 관련 활동은 엄격한 의미에서는 유네스코한국위원회를 제외하고는 유네스코 「권고」 초안 작성 전후의 국내외적 흐름에 대한 대응으로서 독립적으로 시작되었으며 상당히 활발한 성과를 내고 있다고 판단된다. 특히 「권고」가 강조하고 있는 국내 이해관계자와의 협의를 통해 실천 가능하고 효과적인 실행 방안을 모색한다는 점에서 국내 여러 부처와 유네스코한국위원회가 추진한 여러 사업의 성과는 충분히 인상적이라고 할 수 있다.

[8] https://www.korea.kr/news/pressReleaseView.do?newsId=156459220

[9] 유네스코한국위원회의 유튜브 계정에는 인공지능 윤리와 관련된 좋은 교육 콘텐츠가 있다. (https://www.youtube.com/playlist?list=PLZXvLuqePFUoqZGxwgrlWm2BiTi2YV85S)

특히 유네스코한국위원회가 한국법제연구원과 함께 수행한 인공지능 윤리 법제화 연구는 인공지능 윤리가 보다 실천적 함의를 갖기 위해 필요한 다양한 규제 방안을 사회학, 철학, 법학 전문가들이 학제적으로 탐색했다는 데 큰 의미가 있다. 그 구체적인 의견에는 서로 차이가 존재하지만, 세 전문가들은 대체적으로 기존 법률로 이미 보호받고 있는 개인정보 침해 등의 사안이 아니라면 아직 기술개발의 초기 단계인 인공지능 기술에 대해 강한 법적 규제를 서둘러 만드는 것보다는 규제 샌드박스나 적응적 거버넌스(adaptive governance)의 관점에서 적절하고 지속적인 모니터링과 제도적 대응을 수행하는 방식이 더 적절하다는 데 의견을 모았다는 점이 중요하다. 또한 이는 인공지능 기술의 사회적 영향력에 대한 지속적 모니터링에 기반한 적응적 거버넌스를 강조한 유네스코 「권고」의 내용과도 일치한다는 점에서도 중요하다.

3. 국제협력의 가능성

유네스코 「권고」는 정책 행동의 실천 과정에서 회원국 사이의 협력을 매우 강조한다. 물론 이는 OECD나 EU의 인공지능 관련 문건에서도 동일하게 나타나는 내용이다. 국제협력의 강조는 일차적으로는 사이버 공간에서 주로 작동하는 인공지능 기술의 특성상 국제협력이 반드시 이루어져야만 인공지능 윤리 관련 정책 행동이 효과를 발휘할 수 있다는 현실적 고려에서 나온 것으로 볼 수 있다. 이에 더해 EU와 같은 국가 연합체나 유네스코 및 OECD와 같은 국제기구들이 모두 회원국 사이의 이익 증진과 상호협력을 기본 가치로 내세우고 있으며, 따라서 인공지능 윤리 관련 실천적 협력에도 이러한 특징이 반영된 것이라고 볼 수 있다.

이런 두 가지 특징, 즉 인공지능의 기술적 특징과 윤리 규범 및 정책 행동을 제안하는 국가 연합체 혹은 국제기구의 특징을 배경으로 하여 인공지능 윤리의 실천적 힘을 극대화하기 위해 제기되는 것이 바로 기후변화 대응에서 결정적인 역할을 했던 IPCC와 같은 인공지능 윤리 국제기구 — 가령 IP인공지능(가칭)와 같은 — 를 만들어야 한다는 제안이다.[10]

실제로 인공지능과 기후변화는 그 국제적 대응에 있어 많은 공통점이 있다. 기후변화 문제가 특정 국가에 국한될 수 없듯이 세계적 수준에서 운용되는 인공지능 역시 특정 국가의 제도적

[10] 2020년 1월 두바이에서 개최된 세계정부포럼에서는 이 제안에 대한 특별 세션이 구성되어 IP인공지능(가칭)의 장단점이 논의되기도 했다.

한계를 벗어나는 경우가 많다. 또한 그 구체적인 위협의 정도에 대해서는 전문가들 사이에서도 이견이 있지만, 기후변화가 인류에게 '실존적 위험'(existential risk)을 제기하고 있듯이 초지능에 도달한 인공지능이 인류에게 또 다른 '실존적 위험'을 제기할 가능성에 대해서도 스티븐 호킹이나 일론 머스크 같은 유명 과학자나 기술자들이 경고한 바 있다.

하지만 IPCC와 유사한 형태의 인공지능 거버넌스를 총괄하는 국제기구가 가까운 장래에 등장할 가능성은 높지 않다. 이는 인공지능과 기후변화 사이에는 공통점만큼이나 차이점도 많으며, 이러한 차이점 때문에 많은 전문가들과 정책입안자들은 인공지능에 대한 바람직한 국제협력은 IPCC와 같은 정부간 패널의 형태보다는 각국 정부가 자국의 특성을 고려한 인공지능 윤리 정책을 펴고 그 바탕에서 국제협력이 이루어지는 방식이 더 바람직하다고 보기 때문이다.

기후변화와 인공지능의 결정적 차이점이란 인공지능 기술이 정확히 어떤 영향을 어떤 방식으로 미칠 지에 대한 '불확실성', 즉 확률값조차 부여할 수 없는 진정한 의미에서의 '열려진 미래'의 시나리오를 갖고 있다는 사실이다. 이는 인공지능 기술이 아직 완성된 기술이 아니어서 우리가 기술 개발을 어떻게 하는지, 특히 앞서 소개한 '윤리적 설계'(Ethics by Design)가 얼마나 성공적으로 작동하는지에 따라 그 사회적 영향의 내용이 크게 달라질 것이기 때문이다. 인공지능 기술 개발에 대해 우리가 어떤 거버넌스를 채택할 것이며, 사회적 수용성이나 이용자들의 대응이 어떻게 전개되는지에 따라서도 인공지능이 끼칠 영향의 내용이 달라질 것이라는 점도 중요하다. 물론 기후변화의 경우에도 우리가 얼마나 많은 양의 온실가스를 배출할 것인지에 따라 여러 시나리오가 가능하며 IPCC 역시 각 시나리오별로 예측 상황을 발표하고 있지만, 기후변화의 메커니즘이나 원인에 대해서는 과학적으로 알려지고 합의된 사실이 더 많기 때문에 이런 시나리오 기반 예측에서의 불확실성은 인공지능의 그것에 비해 상대적으로 적다고 볼 수 있다.

이런 차이점을 고려할 때 현 단계에서 섣불리 IPCC와 유사한 인공지능 윤리 관련 정부간 패널을 출범시킨다면 그 영향력이 모든 나라에 공정한 방식으로 배분될지 여부에 대해 많은 국가, 특히 인공지능 기술을 지금 막 개발하려고 노력하는 저개발국들이 회의적인 시각을 보이고 있다. 자칫 그들에게는 IP인공지능(가칭)와 같은 국제 기구의 출범이 미국과 유럽, 중국이 차지하고 있는 인공지능 기술의 주도권을 공고히 하는 데 사용될 수 있을 가능성이 높다고 여겨질 수도 있기 때문이다.

이는 인공지능 윤리 관련 국제 협력을 진행할 때 주의해야 할 중요한 측면을 잘 보여준다. 유네스코 「권고」의 총회 상정안을 만들기 위해 열린 정부간 회의에서 회원국들은 「권고」의 윤리 원칙에 대해 대체적으로 크게 공감했다. 하지만 구체적인 정책 행동의 범위와 내용에 있어서는

회원국 사이의 의견 차이가 상당했으며, 특히 「권고」가 너무 구체적인 정책 행동을 제안하는 것은 각국의 자율권을 해치는 것이므로 바람직하지 않다는 러시아와 중국의 입장과 같은 일종의 '최소주의적' 태도를 견지한 나라도 있었다.

그에 비해 회원국 중에서 인공지능을 일종의 '도약 기술'(leapfrogging technology)로 활용하려는 열망을 가진 인공지능 기술 저개발국들은 다소 복잡한 입장을 견지했다. 그들은 인공지능 기술 개발에서 현재 앞서가고 있는 나라들이 기술 패권을 유지하기 위해 자국으로부터 데이터를 가져가면서도 자국의 기술개발에는 별다른 도움을 주지 않는다는 점을 비판하면서, 인공지능 윤리 기준이 지나치게 엄격하게 적용되면 제도적 기반이 약한 자국의 인공지능 기술개발이 정체되고 기술 강대국의 기술패권 전략은 더욱 공고화될 것이라는 점을 지적했다. 따라서 그런 입장을 가진 회원국일수록 인공지능 윤리의 정책 행동 실천을 위한 다양한 국제 협력의 필요성을 강조했으며, 인류 전체에 도움이 될 인공지능 기술 개발의 공동 협력자로서 기술 저개발국을 참여시키는 방향으로 국제협력이 이루어져야 한다는 점을 역설했다.

IV. 정책적 함의

이상의 논의를 통해 우리는 인공지능처럼 빠르게 발전하며 사회적 영향력을 확대하고 있는 기술에 대한 넓은 의미에서의 윤리적 고려와 이를 다양한 방식으로 제도화하려는 노력이 필요하다는 주장이 국제적 공감대를 얻고 있음을 알 수 있다. 또한 이러한 공감대에 기초하여 국가별로 이루어지고 있는 구체적 수준의 인공지능 윤리 거버넌스는 개별 국가의 역사적·사회적·문화적 상황에 대한 치밀한 분석과 연구에 기초하여 이루어져야 하며, '경쟁적으로' 먼저 법제도화를 달성하겠다는 식의 생각은 정당하지 않다는 점도 알게 되었다.

이런 상황에 대해 현장에서 인공지능 기술을 개발하는 공학자나 관련 사업을 추진하는 기업가들은 껄끄러운 '규제'가 또 하나 생긴다고 불편해할 수 있다. '규제'를 곧 '혁신 저하'로 동일시하는 경향이 있기 때문일 것이다. 하지만 이는 기술혁신의 역사의 사실과 어긋난 생각이다. 1970년대에 자동차 배기가스 규제가 도입되려 할 때 미국의 대형 자동차 회사들은 이 규제가 산업 생산력을 저하시키고 소비자의 권익을 해칠 것이라고 극렬하게 반대했지만, 실제로 이 규제는 보다 친환경적인 내연기관을 개발하는 기술 혁신과 배기가스 저감장치 등의 파생 기술 개발로 이어졌다. 어떤 기준으로 평가하더라도 70년대의 배기가스 규제가 기술혁신을 저하했다든지 소비자 권익을 해쳤다고 볼 근거는 없다.

이처럼 적절한 방식으로 합리적으로 운용된 규제는 산업 환경을 바꿈으로써 기업의 기술혁신 의욕을 오히려 더 고취할 수 있으며 사회적으로 유용한 방향으로 기술혁신을 유도할 수도 있다. 현재 한창 기술개발이 이루어지고 있기에, 앞으로 혁신 잠재력이 큰 인공지능 기술에서 현명한 규제가 앞서 강조한 '적응적' 방식으로 이루어진다면 70년대 배기가스 규제와 마찬가지로 기술혁신과 사회적 공익 실현을 동시에 달성할 수 있을 것이다.

그러므로 인공지능 윤리의 제도화와 국제협력 과정에서 핵심적인 사안은 '적응적' 거버넌스를 구체적으로 어떻게 실현할지가 될 것이다. 여기에는 몇 가지 고려사항이 있다. 첫째는 처음부터 강한 법적 규제를 도입하기보다는 규제를 담당할 정부, 그리고 규제의 대상인 동시에 자율규제의 주체인 기업의 '역량 강화'가 선취되어야 한다. 인공지능의 다양한 윤리적 쟁점의 중요성을 깊이 이해하고, 이를 사회적으로 풀어낼 수 있는 전문 역량을 갖춘 인력이 정부와 기업에 배치되어야 한다. 이는 원론적 수준에서 윤리적 고려의 중요성을 공언하는 방식이 아닌, IEEE의 시도처럼 윤리적 고려나 원칙을 기술 개발 과정에서 적극적으로 고려하는 동시에 앞서 소

개한 윤리영향평가와 같은 지속적인 모니터링과 피드백 반영을 수행할 수 있어야 하기 때문이다. 이에 더해 국제 논의에 대표자로 참여하는 정부와 민간기업의 담당자들이 국제 거버넌스에 적극적으로 참여할 수 있는 역량강화가 반드시 필요하다.

둘째로 인공지능 윤리 및 법제화 과정에서 인공지능을 단일한 기술적 대상으로 생각하기보다는 다양한 요소 기술의 집합체, 즉 시스템으로서 이해하는 것이 중요하다. 이렇게 이해될 때 가령 인공지능 기술 활용에서 핵심적인 데이터 수집, 준비, 활용, 폐기 등과 관련된 데이터 거버넌스 논의도 자연스럽게 인공지능 거버넌스와 함께 논의될 수 있다. 또한 인공지능가 성공적으로 개발되고 생산적으로 활용되기 위해 필요한 사회적 요인이나 다른 기술 시스템과의 상호작용에 대해서도 종합적인 시각에서 함께 그 윤리적 쟁점을 논의할 수 있게 될 것이다.

이는 유네스코 「권고」가 지속적으로 강조하고 있는 입장이기도 하다. 인공지능을 시스템 기술로 이해함으로써 우리는 인공지능 기술의 광범위한 파급 효과에 대해 종합적으로 파악할 수 있으며, 전체 범위에 관련된 윤리적 쟁점에 대해 통합적인 시각으로 바라보고 대응책을 마련할 수 있다는 것이다.[11] 예를 들어 인공지능의 윤리적 쟁점에 대해 '적응적' 거버넌스를 실천한다고 할 때, 그 과정에서 민주주의적 가치를 어떻게 고려하고 반영할 것인지에 대한 문제는 추가 연구가 필요한 중요한 주제이다. 인공지능의 자동화된 결정이 의도적이든 비의도적이든 정치적 의견 형성에 큰 영향을 끼치고 있는 상황에서 그에 대해 '적응적'으로 대응한다는 것이 정확히 무엇을 의미하는지조차 논쟁적이기 때문이다.

셋째로는 인공지능 기술과 그 활용의 특징을 올바로 반영하는 인공지능 윤리 법제도화 논의가 중요하다. 예를 들어 인공지능 윤리영향평가의 구체적인 안을 만들 때 인공지능의 개발 및 활용에 참여하는 다양한 주체를 어떻게 참여시킬 것인지, 그 평가를 누가 어떤 방식으로 언제 시행할 것인지에 대한 구체적인 논의를 거쳐야 한다. 이 과정에서 현재 인공지능 기술의 특징을 반영하는 방식으로 인공지능 윤리영향평가의 내용이 만들어져야 하고, 이에 더해 인공지능 기술의 지속적인 발전을 반영하여 이 평가의 형식과 내용을 지속적으로 업데이트해야 한다. 다시 말하자면 인공지능 윤리영향평가의 운용 과정 역시 '적응적' 거버넌스 방식으로 이루어져야 한다는 것이다.

[11] 이런 관점에서 볼 때 현재 개별 인공지능 제품에 집중되고 있는 인공지능 윤리 논의를 사회적 영향력이 훨씬 크다고 평가할 수 있는 인공지능 기반 플랫폼으로 확대할 필요가 있다.

여기에 더해 현재 가장 많이 활용되고 있는 기계학습 기반 인공지능의 반투명성을 고려하여 효율성 있고 실행가능한 법제도화가 필요하다. 중요한 점은 이러한 '고려'가 인공지능 관련 윤리적 원칙의 내용을 약화하거나 훼손시키는 방향으로 추진되어서는 안 된다는 것이다. '실행가능한'이라는 말은 기업이 사회적으로 공감대가 확보된 윤리적 고려를 어떻게 달성할 수 있을지에 대한 구체적 설명 없이 추상적 원칙만을 제시하고 준수를 요구하는 방식이 윤리적 원칙의 내용을 '현실적으로' 보다 더 잘 실현하는 데 실효성이 없음을 지적하는 것이다. 그러므로 현재 정부가 추진 중인 인공지능 윤리 준수 점검리스트 작성 작업 역시 공학자와 산업계의 의견을 충분히 수렴하여 이루어져야 하고, 성과내기식으로 서둘러 추진되어서는 안 된다.[12]

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

그러므로 이런 점을 고려할 때 인공지능 윤리의 제도화는 국제적으로 상당한 공감대가 형성된 인공지능 윤리 원칙을 중심으로 추진하되, 충분한 기간을 두고 관련 쟁점에 대한 연구와 이해당사자와의 숙의 과정을 거쳐 실효성 있는 형태로 마련되어야 하며, 기술의 미래 발전 방향과 인공지능 사용자의 문화적 대응에 대한 불확실성을 고려하여 '적응적' 거버넌스 형태로 추진되어야 한다. 이 과정에서 유네스코의 인공지능 윤리 권고의 내용 등 국제 사회의 인공지능 윤리 논의를 참고하고, 국내에서 인공지능 윤리에 대한 공감대 및 리터러시 교육과 정부 및 기업의 대응 역량을 높이려는 노력이 동시에 추진되어야 할 것이다.

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

참고문헌

· 과학기술정보통신부 2020, 「인공지능(인공지능) 윤리 기준」, https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=112&pageIndex=&bbsSeqNo=94&nttSeqNo=3179630&searchOpt=ALL&searchTxt=.

· 과학기술정보통신부 2022a, 「인공지능(인공지능) 윤리 기준 실천을 위한 자율점검표(안)」, 정보통신정책연구원.

· 과학기술정보통신부 2022b, 「신뢰할 수 있는 인공지능(인공지능) 개발안내서」, 정보통신정책연구원.

· 개인정보보호위원회 2021, 「인공지능(인공지능) 개인정보보호 자율점검표」, https://privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000842517&fileSn=0

· 법제연구원 기획, 최경진, 이기평 지음 2021, 『인공지능 윤리와 법(2): 인공지능 윤리 관련 법제화 방안 연구』, 한국법제화연구원.

· 유네스코한국위원회 기획, 이상욱 지음 2021a, 『유네스코 인공지능(인공지능) 윤리 권고 해설서: 인공지능 윤리 이해하기』, 유네스코한국위원회.

· 유네스코한국위원회 기획, 이상욱 지음 2021b, 『인공지능 윤리 함께 생각하기』, 유네스코한국위원회.

· 유네스코한국위원회 기획, 이상욱, 이호영 지음 2021, 『인공지능 윤리와 법(1): 인공지능 윤리의 쟁점과 거버넌스 연구』, 유네스코한국위원회.

· 이상욱 2020, 「인공지능과 실존적 위험 – 비판적 검토」, 『인간연구』 40: 1-30.

· 이상욱 2021, 「인공지능 윤리란 무엇인가?」, 『HORIZON』(고등과학원웹진), 2021년 5월 31일 게재. (https://horizon.kias.re.kr/17815/_)

· 이중원 외 2018, 『인공지능의 존재론』 서울: 한울.

· 이중원 외 2019, 『인공지능의 윤리학』 서울: 한울.

· 이중원 외 2021, 『인공지능 시대의 인간학』 서울: 한울.

· 한국인공지능법학회 2019, 『인공지능과 법』 서울: 박영사.

· Brockman, J. (eds.) 2019, *Possible Minds: 25 Ways of Looking at 인공지능*, New York: Penguin Press.

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

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

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

· Hofstadter, Douglas 2008, *I am a Strange Loop*, New York: Basic Books.

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

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

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

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

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

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

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

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

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

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

2022년 제2호
유네스코 이슈 브리프

기 획 유네스코한국위원회
지은이 이상욱
편 집 김은영 백영연 김혜수
발간일 2022년 11월 16일
펴낸곳 유네스코한국위원회
교 열 김보람
디자인 수카디자인
주 소 서울특별시 중구 명동길(유네스코길) 26
전자우편 ir.team@unesco.or.kr

간행물 등록번호 IR-2022-RP-3

유네스코 이슈 브리프는 외교부의 지원으로
발간되었습니다.

www.unesco.or.kr

유네스코 이슈 브리프

UNESCO ISSUE BRIEF