국제법 현안 Brief

국제법 현안 Brief는 국제법 관련 현안문제에 관해 간략한 설명을 제공함으로써 국제법 연구자는 물론 일반인에게 국제법에 대한 이해를 돕는 것을 목표로 합니다. 국제사회에서 발생하는 최근 현안과 관련된 국제법 쟁점에 대한 인식과 최근 국제법 동향에 대한 지식을 공유함으로써 국제법에 대한 사회적 인지도를 향상시키는데 도움이 되기를 기대합니다.

대한국제법학회
The Korean Society of International Law

국제법 현안 Brief 편집위원회
편집위원장 최태현 교수 (한양대학교)
편집위원 오승진 교수 (단국대학교)
권현호 교수 (성신여자대학교)
김성원 교수 (한양대학교)
이기범 교수 (연세대학교)
투고문의 ksilbrief@gmail.com
웹사이트 www.ksil.or.kr

UNESCO 총회 "인공지능(AI) 윤리 권고"의 주요 내용 분석과 국제법적 함의

박 기 갑 (고려대학교 법학전문대학원 교수)

출처: UN News, "193 countries adopt first-ever global agreement on the Ethics of Artificial Intelligence"

1. 2021년 권고 채택의 배경

유엔 교육과학문화기구(이하 'UNESCO')는 2021년 11월 23일 제41차 총회 폐막일에 'AI 윤리에 관한 권고'(Recommendation on the Ethics of Artificial Intelligence, 이하 'AI 윤리 권고')를 담은 '결의 41C/Resolution 73'을 회원국 만장일치로 채택하였다. 이 권고는 UNESCO가 2019년부터 2년간에 걸쳐 AI 시스템이 인간사회에 미칠 제반 문제점을 검토하여 그에 대한 대책을 종합적으로 발표한 결과물이고, 이 권고에서 다루는 'AI 시스템' 개념은 가장 넓은 의미에서 "추론·학습·인식·예측·계획 혹은 통제와 같은 지능적 행동이 가능하고, 데이터 및 정보처리 능력이 있는 시스템이며, 학습과 인지작업 수행능력을 발생시키는 모델과 알고리즘을 통합하여 물질 및 가상 환경에서 예측, 의사결정과 같은 결과를 도출하는 정보처리기술"이다.[1]

역사적으로 볼 때 UNESCO는 과학기술의 발전이 인간사회에 상당한 영향을 미칠 수 있다는 우려가 제기될 때 그와 관련되는 권고 또는 선언을 채택하여

제시하였다. 좋은 선례가 1997년 11월 11일 채택된 “인간 게놈과 인권에 관한 보편적 선언”(Universal Declaration on the Human Genome and Human Rights)이다. 21세기에 들어서서 AI 시스템이 인류 발전에 크게 기여할 수 있고 모든 국가에 혜택을 줄 것이라는 긍정적 기대가 널리 퍼져 있다. 하지만 다른 한편에서는 AI가 편향된 정보를 학습함으로써 차별과 불평등을 심화시키고, 정보 획득과 사용의 격차를 더 벌릴 뿐만 아니라 문화적·사회적·생물학적 다양성을 위협함으로써, 한 국가 내에 그리고 국가들 사이에 존재하는 사회경제적 격차를 심화시킨다고 하는 부정적 우려 역시 적지 않다. 이러한 이유로 최근 몇 년 사이에 UNESCO를 포함한 정부간 국제기구들과 관련 시민단체는 AI 시스템과 인간의 존엄성·인권·문화적 다양성 등을 포섭하는 “인간 가치”(human value) 및 인간이 선택한 목적을 달성하기 위하여 결정하는 권한을 AI 시스템에 맡길지 여부 그리고 그 권한을 맡기는 방식 등과 관련하여 “인간에 의한 통제”(human control)의 상호관계에 대한 권고적 성격의 문서를 채택하고 있다. 이하에서 2021년 “AI 윤리 권고”의 주요 내용과 그 국제적 함의를 살펴본다.

2. 2021년 권고의 주요 내용

AI 시스템과 관련된 윤리 문제는 연구·설계·개발부터 출고 및 사용뿐만 아니라 시스템의 유지·운용·교역·자금조달·모니터링 및 평가·유효성 검사·사용종료·분해·폐기에 이르는 이른바 “AI 시스템 수명주기”(life cycle)의 모든 과정과 연관되어 있다. 이처럼 다양한 윤리 문제를 다루는 2021년 “AI 윤리 권고”는 서문(preamble)과 8개의 장으로 구성되며, 총 141개 항으로 이루어져 있다. 이 중에서 국제법적 측면에서 관심을 가지고 봐야 할 부분은 “가치와 원칙”이 언급된 제3장과 국내 입법 및 행정조치를 염두에 두고 있는 “정책조치”에 관한 제4장이다.

제3장 제1절은 AI 시스템 수명주기에 관계된 국가와 공공 및 민간 AI 행위자(actors)가 준수해야 할 “가치”를 네 가지로 제시하고 있다. 첫째, AI 시스템의 수명주기 전반에서 인권과 기본적 자유 및 인간 존엄성은 존중·보호·증진되어야 한다. 둘째, AI 시스템의 수명주기와 관련된 모든 행위자는 환경 및 생태계의 보호와 복원을 위한 예방조치, 지속가능한 개발 등을 위하여 제정된 국제법과 각국의 법률, 기준과 관행을 준수해야 한다. 셋째, AI 시스템의 수명주기 전반에서 국제인권법을 포함한 국제법에 따라 다양성 및 포용성의 존중·보호·증진이 보장되어야 한다. 넷째, AI 행위자들은 인권 및 기본적 자유의 가치에 입각하여 평화롭고 정의로운 사회의 구축에 적극적으로 참여하고 이를 실현하여야 한다.

제3장 제2절은 위 가치에 기반을 두고서 AI 행위자들이 준수해야 할 “원칙”을 열 가지로 규정하고 있다. 첫째, “비례 원칙과 위해 금지”이다. AI 시스템 수명주기와 관련되는 과정에서 합법적 목표나 목적을 달성하는 데 필요 이상의 수단을 사용해서는 안 되며, 각기 상황에 비례한 결정을 내려야 하며, 특히 AI 시스템은 소셜 스코어(social scoring) 평가나 대중감시(mass surveillance)의 목적으로 사용되어서는 안 된다. 둘째, “안전 및 보안”이다. 인간·환경 및 생태계의 안전과 보안을 보장하기 위해 AI 시스템의 수명주기 전반에서 원치 않는 피해(안전 위험)와 공격에 대한 취약성(보안 위험)과 관련된 요소를 방지하고 이를 해결·예방·제거해야 한다. 셋째, “공정과 비차별”이다. AI 행위자들은 사회적 정의를 증진하고 국제법에 따라 공정과 비차별을 수호하여, 다양한 모든 사회구성원이 AI 기술이 주는 혜택을 공유할 수 있어야 한다. 넷째, “지속가능성”이다. AI 시스템이 인간·사회·문화·경제·환경 등에 미치는 영향에 대한 지속적인 평가는 “UN 지속가능한 개발목표”(UNSDGs)에 명시된 지속가능성과 조화되는지를 확인하면서 이행되어야 한다. 다섯째, “프라이버시에 대한 권리 및 데이터 보호”이다. 사생활은 AI 시스템 수명주기 전반에서 존중·보호·증진되어야 한다. 여섯째, “인간의 감독 및 결정”이다. 국가는 AI 시스템 수명주기의 모든 단계에서 AI 시스템으로부터 야기되는 윤리적·법적 책임을 물을 수 있도록 보장하여야 한다. 원칙적으로 인간의 생사와 관련된 결정은 AI 시스템에 이양될 수 없다. 일곱째, “투명성과 설명가능성(또는 석명가능성, 釋明可能性)”이다. AI 시스템의 민주적 거버넌스를 위해 수명주기 전반에서 역외 영향을 포함한 투명성 및 석명가능성을 증진하려는 노력이 계속되어야 한다. 여덟째, “책임 및 책무”이다. AI 행위자 및 국가는 인권 및 기본 자유를 준수·보호·증진하고 동시에 국가가 갖는 인권 보호 의무를 비롯한 국제법과 AI 행위자의 영역 및 통제 내에서의 문제를 포함한 AI 시스템 수명주기 전반의

윤리 지침에 따라 각각 윤리적·법적 책임을 다함으로써 환경 및 생태계의 보호를 장려해야 한다. 아홉째, “의식 및 문해력(文解力, literacy)”이다. 사회구성원들이 AI 시스템 사용에 대한 충분한 정보를 가지고 의사결정을 내리며, 부당한 영향으로부터 보호받을 수 있도록 대중의 효과적인 참여를 보장하고 언어・사회・문화적 다양성을 고려해야 한다. 마지막 열 번째, “다자적이고 조정 가능한 거버넌스 및 협력”이다. 데이터 사용에서 국제법과 국가 주권이 존중되어야 한다. 이는 국가가 국내에서 생성되거나 자국의 영토를 거치는 데이터를 국제법을 준수하는 방식으로 규제하고 데이터의 효과적인 규제를 위해 국제법과 기타 인권 보호 관련 규정 및 기준에 따라 프라이버시 보장을 위한 데이터 보호와 같은 조치를 할 수 있음을 의미한다.

제4장은 “정책조치 분야”를 다루고 있다. 그 내용은 위에 언급한 가치와 원칙을 바탕으로 각 국가가 자발적으로 관련 정책 또는 메커니즘을 확립하고, 공공 및 민간 AI 행위자가 관련 조치를 준수하게 하는 것을 목표로 한다. 여기에서는 윤리영향평가, 윤리적 거버넌스 및 직무, 데이터 정책, 발전 및 국제협력, 환경 및 생태계, 젠더, 문화, 교육 및 연구, 통신과 정보, 경제 및 노동, 의료 및 사회복지 등 총 11개의 정책분야를 규율하고 있다.

위에 언급한 “가치”(제2장), “원칙”(제3장)과 “정책조치 분야”(제4장)는 2021년 “AI 윤리 권고”의 대부분을 차지한다. 이에 대하여 몇 가지 선결적 질문을 던져본다. 첫째, AI 시스템의 운용과 규율 문제는 마치 사이버 공간에서의 국가 또는 사인(私人)의 행위 규율처럼 새로운 분야인데 과연 기존의 법체계가 이들을 모두 포섭할 수 있을까? 또한 여기서 준거법으로 언급되는 국제법은 그 실체가 분명한가? 이 권고는 국제법과 국제인권법을 상당히 많은 데에서 준거법으로 인용하고 있다. 가령 “국제법에 입각한 본 권고”(서문 세번째 항), “국제인권법과 국제법에 따라 채택된 AI 기술에 관한 글로벌 윤리 기준”(제2장(목표) 제8항), “국제법에 입각하여” 또는 “국제법에 또는 국제법과 국제인권법에 따라”(제2장(목표) 제11항. 제3장(가치) 제13항, 제18항, 제19항, 제28항, 제32항, 제42항, 제46항. 제4장(정책조치 분야) 제63항, 제72항, 제107항, 제121항. 제5장(모니터링 및 평가) 제131항, 제133항 등)이다. 그러나 2021년 “AI 윤리 권고”에서 말하는 국제법의 범주는 인용된 국제문서의 법적 효력과는 무관하게 국제조약뿐만 아니라 선언, 결의, 권고 내지 지침 등까지 확대되고 혼재되어 있다.

둘째, 권고에 명시된 각각의 “가치”와 “원칙”은 하나하나가 중요하고 필요하다고 보이지만, 현재 COVID-19 대확산 상황에서 공공이익 보호를 위하여 개인의 기본적 인권을 정지 내지 침해할 수 있는지에 대한 논쟁처럼, 실제 적용시 개별 “가치”와 “원칙” 상호 간에 충돌이 발생할 수 있다. 이 경우 무엇을 우선할 것인가? 이에 대하여 권고 제2장(목표) 제11항은 “…명시된 가치들은 그 자체로 바람직하지만, 실무적 맥락에서 가치와 원칙 간에 부조화가 발생할 수 있다. 따라서 잠재적 긴장을 완화하기 위해 비례성과 인권 및 기본적 자유의 원칙에 입각한 상황별 우선시가 필요할 것이다”라고 나름대로 해법을 제시하고 있다. 여기서 말하는 “상황별 우선시”는 권고에 담긴 가치 또는 국제법 원칙 상호 간에 충돌이 발생할 경우, 사례별로 구체적 해결방안을 모색하는 것을 의미하므로, 관련된 원칙을 구체화하는 과정에서 발생할 수 있는 자의적이고 임의적인 판별을 최대한 줄이는 것이 중요한 관건이 될 것 같다.

셋째, “가치”와 “원칙”을 반영한 “정책조치 분야”는 각 국가가 자발적으로 국내 입법・행정 시스템에 적용할 것을 목표로 삼고 있다. 그러나 개별 국가의 기술・재정 능력과 사회환경은 천차만별이기 때문에 “정책조치 분야”에 언급된 많은 제언을 국가가 국내적으로 반영하는 정도에서는 격차가 발생하고 그 형태에서는 취사선택 또는 변형이 일어날 수 있다. 이렇게 된다면 정작 2021년 “AI 윤리 권고”가 달성하고자 했던 소기의 목표는 불균등하고 파편화되고 희석화된 결과로 가시화될 수 있다. 이에 대하여 권고 제4장(정책조치 분야) 제49항은 “UNESCO는 회원국이 본 권고의 이행에 있어서 과학적・기술적・경제적・교육적・법적・규제적・인프라적・사회적・문화적 면에서의 준비가 상이하고 각기 다른 단계에 있음을 인정하면서 “준비”란 유동적이다”라고 설명한다. 이는 결국 자발적 이행과정에 한계가 있음을 인정한 셈이다. 비록 UNESCO가 일단은 회원국들에 관련 기술지원 제공 등을 제안하고 있지만, 그 제공이 의무가 되려면 중장기적으로는 법적 구속력을 갖는 지역적 또는 보편적 국제협약 채택 노력이 요구된다. 짧은 시간 내에 많은 국가가 참여할 수 있는 관련 국제협약 채택을

원한다면 그 형태는 구체적이며 세부적인 측면을 다루기보다는 현존하는 주요 국제인권협약의 관련 내용을 차용 또는 준용하는 기본골격(framework) 체제가 바람직할 것이다.

3. 2021년 권고의 국제법적 함의

2021년 "AI 윤리 권고"는 다음과 같은 몇 가지 국제법적 함의를 갖는 것으로 보인다. 첫째, AI 시스템의 연구개발과 그것의 활용 과정 전반에 적용돼야 할 전 세계적 윤리 가이드라인으로서의 역할이다. UNESCO의 세계과학기술윤리위원회(COMEST)가 사전 연구를 통해 밝혔듯이, 현재까지 AI 시스템의 개발과 적용 과정에 윤리적이며 인간 중심적인 접근법을 강조하는 전 세계적 문서가 없었다. 따라서 이 권고는 현재와 미래에 교육과 행정, 네트워크 등 광범위한 영역에서 활용되며, 수십억 명의 사람들의 생활에 직간접적인 영향을 미칠 AI 시스템에 대한 긍정적인 가능성에 대한 기대와 오남용으로 인한 인권침해나 불평등 심화 등의 부작용을 우려하는 목소리를 포괄적이며 종합적으로 다루고 있으므로 그 유용성은 높게 평가될 것이다.

둘째, 이 권고가 적용되는 일차적 대상은 UNESCO 회원국들이다. 국가는 AI 시스템 수명주기 전반과 관련된 법적 규제체제를 개발하고 AI 시스템 개발과 활용을 촉진할 권리와 의무를 행사하므로, 그 과정에서 국가는 이 권고를 AI 시스템 관련 국내 법률의 제정, 정책의 입안 그리고 기타 수단에 관한 지침으로써 자발적으로 활용할 수 있을 것이다. 이 권고가 과거에 유네스코가 채택한 "인간게놈과 인권에 관한 보편적 선언"(1997년), "생명윤리와 인권에 관한 보편적 선언"(2005년)처럼 과학 분야의 발전을 더 인간적이고 윤리적이며 포용적인 방향으로 이끌어 줄 것으로 기대한다.

셋째, 이 권고는 이차적으로 공공 및 민간 AI 행위자 또는 이해관계자들에게도 관련된다. 이들은 AI 시스템 수명주기 중 적어도 한 단계에 관련된 행위자로 정의될 수 있으며, 연구자, 프로그래머, 엔지니어, 데이터 과학자, 최종 사용자, 기업, 대학, 민간단체와 공공단체 등을 비롯한 자연인과 법인 전체를 포함한다. 가령 권고의 제50항에서 제53항에서 언급하고 있는 AI 시스템 수명주기 전반에 대한 "윤리영향평가"(ethical impact assessment)는 국가가 아닌 개인에게도 윤리적 지침을 제공하는 역할을 담당할 것이다. 참고로 "윤리영향평가"라 함은 AI 시스템의 혜택, 우려 및 위험을 식별 평가하고 적절한 위험 예방, 완화, 모니터링 조치 등을 행하는 것을 말한다. 이러한 영향평가는 소외계층과 취약계층 혹은 취약한 상황에 노출된 개인의 권리, 노동권, 환경·생태계 및 윤리·사회적 연관성처럼 인권 및 기본적 자유에 영향을 미치는 요소를 식별하는 기능을 갖는다.

넷째, 2021년 "AI 윤리 권고"에는 물적 적용 범위(*ratione materiae*)가 존재한다. 이 권고의 제1장(적용 범위) 제1항은 "본 권고는 UNESCO의 권한 내에서 AI 영역과 관련된 윤리 문제를 다룬다"라고 밝히고 있다. 이는 UNESCO라는 국제기구의 법인격의 한도 내에서 허용되는 활동 범위, 즉 교육·과학·문화 협력 측면에 충실하겠다는 의미이다. 따라서 이 권고는 원칙적으로 평화 시의 AI 시스템 관련 문제에 맞춰져 있으며, 전시(戰時)의 AI 시스템 사용 문제는 권고의 직접적 규율 범위 밖에 있는 것으로 봐야 할 것이다. 이는 제2장(목표) 제5항의 "본 권고는 … AI 시스템의 평화로운 사용의 촉진을 권고한다"라는 문안에서 암묵적으로 추정할 수 있다. 그렇다고 해서 AI 시스템의 활용이 전시(戰時)에 금지된다는 의미는 아니며, 그의 적법성 여부는 전쟁법 또는 국제인도법이라는 특별법(*lex specialis*)에 의해 규율될 것이다. AI 시스템을 활용하는 신무기와 관련해서는 "신무기, 전투 수단 또는 방법의 연구·개발·획득 및 채택에 있어서 체약당사국은 동 무기 및 전투 수단의 사용이 본 의정서 및 체약당사국에 적용 가능한 국제법의 다른 규칙에 의하여 금지되는지 여부를 결정할 의무가 있다"라고 규정한 1977년 제1추가 의정서 제36조를 참고로 할 수 있다. 현존하는 전쟁법 또는 국제인도법의 규정과 원칙이 과연 나날이 진보하는 AI 시스템 또는 자신의 동작을 스스로 개선·증강할 수 있는 슈퍼컴퓨터의 능력을 일컫는 "기계학습"(machine learning)에 적절하게 대처하는 데 충분한지 아니면 보완되어야 하는지에 대해서는 다른 논의의 장을 필요로 한다. 왜냐하면 자율적 무기체제(autonomous weapon systems) 중에서도 AI 시스템과 기계학습을 갖춤으로써 공격목표를 선택하고 스스로 공격하는 결정적 기능에 있어서 자율성을 갖는 무기체제는 무력사용에 대한 인간의 통제 상실의 위험을 안고 있으므로 인도적·법적 그리고 윤리적 관점에서 큰 우려를 자아내기 때문이다.

필자 소개

박기갑 교수는 고려대학교 법학전문대학원 교수로 재직 중이다.

국제법 현안 Brief의 내용은 필자 개인의 견해이며 **대한국제법학회**의 공식적인 입장은 아닙니다.

---

1 2021년 “AI 윤리 권고”의 한글번역은 유네스코 한국위원회의 번역본을 최대한 따랐음을 밝힌다.