경제규제와 법(Journal of Law & Economic Regulation)
제13권 제1호 (통권 제25호). (Vol. 13. No. 1). 2020. 5. pp.7~36
doi :10.22732/CeLPU.2020.13.1.7
http://data.doi.or.kr/10.22732/CeLPU.2020.13.1.7

일반논문

인공지능 윤리규범과 규제 거버넌스의 현황과 과제†*

Challenges of Establishing Ethics Principles and a Governance Regime for Artificial Intelligence

고 학 수(Haksoo Ko)** / 박 도 현(Dohyun Park)*** / 이 나 래(Narae Lee)****

목차

<국문초록>

본고는 인공지능 시대에 새로이 부각되는 윤리적 이슈를 고찰하고 우리나라가 추구해야 할 윤리규범과 규제 거버넌스 체계의 방향에 대한 실마리를 제공하고자 한다. 인공지능 윤리는 그동안 관련분야와의 상호작용을 토대로 다음과 같은 방향으로 발전해왔다. 첫째, 아시모프 로봇 3원칙 이래로, 논의의 주체가 설계자, 제작자, 이용자와 같은 인간 이해관계자로 확장되어왔다. 둘째, 책임귀속에 관한 과거의 대전제가 점차 통용되지 않고, 대중의 알 권리나 이해관계자의 주체적 참여에 대한 요구가 늘어나면서 '책임성'에 대한 관념이 확장되고 있다. 셋째, '신기술' 일반의 본성을 추상적으로 규명하는 접근방식의 한계로 인해, 인공지능이 활용되는 구체적 맥락에 주목하는 방향으로 논의가 구체화되고 있다.

인공지능 윤리에 관한 국내외의 논의는 최근 몇 년 동안 급속히 진전되었다. 이러한 논의는 윤리규범을 마련하거나 규제 거버넌스 구조를 구축하는 방향으로 수렴하고 있다. 지금까지의 논의는 내용의 범위와 깊이를 기준으로 크게 세 가지 유형으로 나누어볼 수 있다. 첫째, 기본원칙 중심의 논의, 둘째, 기본원칙과 심화된 이슈를 함께 다루는 논의, 셋째, 윤리규범과 규제 거버넌스 정립방안에 대한 구체적 논의이다. 일반적으로 최신의 논의이거나 공적 주체가 마련한 논의일수록 후자에 해당하는 경향을 보이지만, 반드시 그렇지는 않다.

국내의 인공지능 윤리에 대한 최초의 논의는 2007년 발표된 '로봇윤리헌장 초안'에서 시작되었다고 볼 수 있다. 이를 기준으로 하면 세계적으로도 빠르게 논의가 시작된 편이다. 그러나 그로부터 오랜 동안 논의에 별다른 진전이 없다가, 근래에 새로이 관심이 늘고 있다. 인공지능 윤리 및 규제 거버넌스 담론은 지속적으로 국제사회와 발맞추어 진행되어야 한다. 또한, 일견 원론적이고 교과서적 내용의 단순한 나열로 보일 때에도, 배후에는 자신에게 유리한 방향으로 논의를 이끌기 위한 이해관계자들의 치열한 이익대립이 숨겨져 있다는 사실을 염두에 두고 제반 이슈에 대한 면밀한 검토와 분석이 필요하다.

† 투고일자 2020. 5. 9, 심사일자 2020. 5. 25, 게재확정일자 2020. 5. 25.
* 본고는 당초 고학수/이나래/박도현, "윤리적 인공지능의 실현과 과제", 『서울대학교 인공지능정책 이니셔티브 이슈페이퍼 01: 인공지능과 미래사회』, 2019, 70-106면으로 제작한 원고를 보완하고 새로운 논의를 추가하여 완성된 논문의 형태로 마련한 것이다.
** 서울대학교 법학전문대학원 교수, 교신저자(Professor of Law and Economics, Seoul National University School of Law).
*** 서울대학교 일반대학원 법학과 박사과정 수료·변호사, 제1저자(Attorney and Ph.D. candidate of Seoul National University School of Law).
**** 서울대학교 일반대학원 법학과 박사과정·변호사, 제2저자(Attorney and Ph.D. student of Seoul National University School of Law).

Ⅰ. 들어가는 글

인공지능 기술이 급속히 발전하고 널리 활용되면서, 부작용과 폐해를 방지하기 위한 규범적 논의도 늘어나고 있다. 미국 스탠퍼드 대학은 2014년 가을 시작된 '인공지능에 대한 100년 연구'의 첫 번째 성과로, 2016년 9월에 인공지능이 사회적 영역에서 제기하는 대표적 문제를 8가지로 정리하고 이에 대한 정책적 제언을 제시하는 보고서를 발표했다.[1] 이와 별도로 오바마 행정부는 2016년 10월에 안전의 문제와 위험 규제, 공정성 문제, 인간적 가치의 보호와 같은 윤리적·규범적 차원의 논의를 담은 보고서를 발간하였고, 같은 해 12월에는 노동 분야에 집중한 후속 보고서를 내놓았다.[2] 트럼프 행정부는 2019년 2월 기술개발에 대한 강조와 동시에 안전이나 프라이버시와 같은 윤리적 측면을 강조하는 내용이 담긴 행정명령(executive order)을 발표하였다.[3] 구글(Google), 마이크로소프트(Microsoft) 등을 포함한 많은 글로벌 기업들도 보고서, 윤리원칙, 가이드라인, 모범사례(best practice)처럼, 명칭을 불문하고 윤리적 고려사항을 반영한 실무관행을 구축해나가고 있다.

유럽의 분위기 역시 넓게 보면 크게 다르지 않다. 유럽에서는 인공지능의 윤리적 측면에 대한 관심이 일찍부터 나타났고, 근래에는 유럽연합 차원의 좀 더 본격적인 움직임을 통해 신기술 활용과 기본권 보호를 조화하려는 시도가 보이고 있다. 특히 실정법 차원에서는 2018년 5월 25일부터 시행된 유럽연합의 일반 개인정보보호규정(General Data Protection Regulation, 이하 'GDPR')에 인공지능 윤리 관점에서 중요한 의미를 지니는 조항들이 적지 않게 포함되어 있다. 대표적 사례로, GDPR은 개인정보 주체에게 프로파일링(profiling)을 포함하여 법적 효력이나 그와 유사한 중대한 효과를 미치는 경우, 오로지 자동화된 의사결정의 대상이 되지 않을 권리를 인정하고 있다. 또한 정보주체는 '설명을 요구할 권리(right to explanation)', '잊힐 권리(right to be forgotten)'로 널리 알려진 정보 제공권, 열람권, 삭제권, 반대권과 같은 다양한 권리를 행사할 수 있다. GDPR은 유럽연합 회원국에 대해 법적 구속력을 미치고, 유럽연합 역내 구성원의 개인정보를 처리할 경우 유럽연합 역내는 물론 역외에 사업장을 둔 기업체에도 적용된다.

최근에는 개별국가나 개별기업과 같은 단일 주체의 차원을 넘어, 보다 다양한 주체가 여러 가지 경로로 함께 논의를 진행하는 움직임도 보인다. 유럽연합 같은 국가연합이 개별국가 차원의 움직임을 심화하는 것처럼, 기존에 존재한 전기전자기술자협회(IEEE)나 새로이 결성된 인공지능 파트너십(Partnership on AI)과 같은 기업과 산업을 포함한 연합체 역시 인공지능 윤리 이슈에 대한 논의를 선도하고 있다. 나아가 세계적 차원의 논의를 결집한 국제기구의 목소리는 그 특성상 높은 수준의 정당성을 인정받게 되는 것이 보통이다. 대표적으로, 경제협력개발기구(OECD)의 윤리원칙은 국가, 지역, 분야 등을 넘어선 다양한 주체의 시각을 반영한 산물로 평가된다.[4] 인공지능 기술의 파급력과 복잡성을 고려할 때 각국 정부는 물론 기업과 학계, 나아가 일반 대중의 시각도 고루 반영되어야만 윤리적 문제를 적절히 해결할 수 있으므로, 이러한 분위기는 더욱 심화될 것으로 예상된다.

이러한 모습을 윤리적 인공지능을 구현하기 위한 국제사회의 거버넌스 논의라고 본다면, 자연스럽게 뒤따르는 질문은 이와 관련된 국내의 상황은 어떠한가에 대한 것이다. 우리나라에서 진행되는

1) Peter Stone et al., "Artificial Intelligence and Life in 2030", One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, 2016.

2) U.S. Executive Office of the President/National Science and Technology Council Committee on Technology, "Preparing for the Future of Artificial Intelligence", 2016; U.S. Executive Office of the President, "Artificial Intelligence, Automation, and the Economy", 2016.

3) Donald J. Trump, "Executive Order on Maintaining American Leadership in Artificial Intelligence", Federal Register: White House, 2019, pp. 3967-3972.

4) OECD, "Recommendation of the Council on Artificial Intelligence", 2019.

논의의 현황을 평가하고 올바른 방향을 제시하기 위해서는 '인공지능 윤리' 분야 자체에 대한 심도 있는 연구를 진행할 필요가 있다. 본고는 이러한 문제의식에 기초하여, 인공지능 윤리규범과 규제 거버넌스에 대한 국내외 논의 동향을 살펴보고 시사점을 모색한다. 이하에서는 먼저 인공지능 윤리 분야의 변천사를 개관하면서 그동안 지적되어온 주요 쟁점을 살펴본다(II). 다음으로는 상술한 윤리 이슈에 대한 대안으로 해외에서 마련되고 있는 윤리규범과 이를 현실에 구현하는 기제인 규제 거버넌스 논의를 유형화하여 고찰한 뒤, 의의와 한계를 도출한다(III). 끝으로, 여기에서 얻은 시사점을 국내의 윤리규범 및 규제 거버넌스 논의에 적용하고 향후 나아가야 할 방향을 모색한다(IV).

본격적 논의에 들어가기에 앞서, 본고의 몇 가지 전제사항을 밝혀두고자 한다. 첫째로, 본고에서 분석의 대상으로 삼는 '인공지능'은 소프트웨어 격에 해당하는 알고리즘뿐만 아니라, 학습의 원천인 (빅)데이터, 나아가 자율주행 자동차와 같은 하드웨어(로봇)가 함께 작용하는 경우를 포괄하는 넓은 개념이다. 둘째로, 본고에서 말하는 '인공지능'은 오늘날 국내외 학계와 실무의 연구개발 대상인 유형에 한정하도록 한다. 일부 매체에서는 인류와 똑같은 방식으로 사고하고 행동하거나 그 이상의 인지능력을 확보해 인류의 생존을 위협하는 인공지능의 모습을 그려내고 있기도 하다. 물론 초인공지능(superintelligence)이라고 불리는 이러한 인공지능 윤리에 대한 연구도 진행되고는 있다.[5] 그러나 이 글은 현시점을 기준으로 하여 실제로 일반 이용자들을 대상으로 제공되고 있는 (또는 상당히 가까운 장래에 제공될 것으로 예상되는) 유형의 기술을 전제로 논의를 전개한다. 셋째로, 본고에서의 '윤리'는 공동체를 바람직한 방향으로 이끌기 위해 구성원에게 요구되는 사회규범이라는 매우 넓은 의미로 규정한다. 여기에는 개인 차원의 '도덕'은 물론, 법규범에 비해 강제성과 명확성은 다소 낮지만 자발성과 유연성은 보다 높은 일련의 규범들이 모두 포함될 수 있다.[6]

Ⅱ. 인공지능 윤리의 변천사와 주요 쟁점

1. 인간을 중시하는 윤리관과 윤리주체의 다변화

많은 사람들은 인공지능 윤리규범의 시초로 SF 소설가 아이작 아시모프(Isaac Asimov)가 1942년 '런어라운드(Runaround)'라는 소설에서 최초로 언급한 '로봇 3원칙'을 떠올린다.[7] 인간에 대한 위해의 방지, 인간에 대한 복종, 로봇 자신의 보호라는 위계적 구조의 3원칙을 통해 (인공지능) 로

5) 초인공지능 윤리를 논의한 한 가지 사례로, Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence", Science Fiction and Philosophy: from Time Travel to Superintelligence, 2003, pp. 277-284 참조.

6) 국내 법학 분야에서 '인공지능/로봇'과 '윤리/도덕'을 제목에 명시한 일부 선행연구를 열거하면, 이상형, "윤리적 인공지능은 가능한가? – 인공지능의 도덕적, 법적 책임 문제 –", 『법과 정책연구』 제16권 제4호, 2016; 김건우, "로봇윤리 vs. 로봇법학: 따로 또 같이", 『법철학연구』 제20권 제2호, 2017; 김종호, "인공지능 시대의 윤리와 법적 과제", 『과학기술법연구』 제24권 제3호, 2018; 송상현, "인공지능과 도덕성", 『법조』 제67권 제6호, 2018; 이원태 외, 『4차산업혁명시대 산업별 인공지능 윤리의 이슈 분석 및 정책적 대응방안 연구』, 대통령 직속 4차산업혁명위원회, 2018; 정채연, "지능정보사회에서 지능로봇의 윤리화 과제와 전망 – 근대적 윤리담론에 대한 대안적 접근을 중심으로 –", 『동북아법연구』 제12권 제1호, 2018; 한희원, "인공지능(AI) 치명적자율무기(LAWs)의 법적·윤리적 쟁점에 대한 기초연구", 『중앙법학』 제20집 제1호, 2018; 한희원, 『인공지능(AI) 법과 공존윤리』, 박영사, 2018; 김효은, "인공지능과 윤리", 『인공지능과 법』, 박영사, 2019; 심민석, "로봇과 인공지능(AI)의 법적·윤리적 입법방안에 관한 연구", 『비교법연구』 제19권 제2호, 2019; John P. Sullins (권은정 감수), "미국의 인공지능(AI) 윤리 및 거버넌스 현황", 『경제규제와 법』 제12권 제2호, 2019; 김미리/윤상필/권헌영, "인공지능 전문가 윤리의 역할과 윤리 기준의 지향점", 『법학논총』 제32권 제3호, 2020 등 참조.

7) 3원칙의 내용은 다음과 같다. 첫째, 로봇은 인간에게 해를 끼치거나, 어떠한 행동도 하지 않아 인간에게 해가 가해지도록 하면 안 된다(A robot may not injure a human being, or, through inaction, allow a human being to come to harm). 둘째, 로봇은 1원칙에 위배되지 않는 한 인간의 명령에 복종하여야 한다(A robot must obey orders given it by human beings except where such orders would conflict with the first law). 셋째, 로봇은 1원칙과 2원칙에 위배되지 않는 한 자신을 보호하여야 한다(A robot must protect its own existence as long as such protection does not conflict with the first or second law).

봇이 초래할 수 있는 해악과 위험을 방지할 수 있다는 생각에서 비롯된 것이다. 문제는 로봇 3원칙을 엄격히 고수한다면, 극단적으로는 인간에게 위해의 여지가 있으면 정당방위 같은 해악을 방지하기 위한 개입조차 불가능한 반직관적 결론에 이른다는 것이다. 아시모프는 이러한 모순을 해소하고자 1985년 소설 로봇과 제국(Robots and Empire)에서 "로봇은 인류에게 해를 가하거나 어떠한 행동도 하지 않아 인류에게 해가 가해지도록 하면 안 된다(A robot may not harm humanity, or, through inaction, allow humanity to come to harm)"는 내용의 '0원칙'을 제시한다. 기존의 로봇 3원칙에 있던 '인간(human)'의 자리에 '인류(humanity)'를 도입하여 앞서 언급한 모순적 결론을 방지하려는 생각에서 비롯된 것이다.[8]

로봇 3원칙은 이하에서 볼 2006년 유럽로봇연구 네트워크 로봇윤리 로드맵의 출발점이 되는 등, 초기단계 인공지능 윤리 논의에서 일정한 역할을 담당하였다. 그러나 어디까지나 소설의 일부에 불과한 로봇 3원칙만으로 윤리적 문제를 해결하는 데는 한계가 있었다. 로봇 3원칙은 무엇이 문제였을까? 선언적 원칙이어서 구체성도 부족하였지만, 무엇보다 윤리주체가 인간이 아니라 로봇인 점이 주로 지적되었다. 로봇을 만들고 이용하는 '인간'이 준수하여야 할 윤리를 '로봇'에 전가할 수 있도록 오해될 여지가 있었기 때문이다.[9] 이러한 문제는 일본 후쿠오카에서 2004년 발표된, 인공지능·로봇에 관한 세계 최초의 윤리규범으로 평가받는 '세계로봇선언(World Robot Declaration)'에서도 마찬가지였다.[10] 세계로봇선언은 로봇이 인간에게 일방적으로 복종하도록 설정한 로봇 3원칙에 비해 '로봇과 인간의 공존'을 중시하기 때문에 진일보한 것으로 평가할 수 있지만, 여전히 인간의 윤리에 대한 언급은 없었다. 나아가, 당시는 물론 오늘날의 기술로도 구현할 수 없는 자의식과 자유의지를 전제한 듯한 '로봇의 윤리'는 정당성을 논하기 이전에 현실적 가능성조차 인정되기 어려운 일종의 사고실험에 불과한 것이었다.

비슷한 시기 유럽로봇연구 네트워크(European Robotics Research Network, EURON)라는 단체는 2003년부터 3년 동안의 연구를 거쳐서 2006년 '로봇윤리 로드맵'을 발표하였는데, 여기에는 인간의 윤리를 주된 목표로 삼는다는 내용이 명시되었다.[11] 로드맵은 로봇을 제작하거나 활용할 때 중시하여야 할 13대 원칙을 구체화하고, 로봇의 3대 이해관계자인 설계자(designer), 제작자(manufacturer), 이용자(user)를 윤리주체로 특정하였다. 이듬해인 2007년 우리나라 산업자원부에서는 인간중심, 인간과 로봇의 공존, 인간과 로봇의 윤리라는 그동안의 논의를 종합적으로 반영한 '로봇윤리헌장' 초안을 공개하였다. 이와 같은 논의를 기초로, 오늘날 인공지능 윤리는 인간과 인공지능의 공존 속에서 인간의 존엄성과 기본권의 실현을 대원칙으로 두고, 윤리주체의 관점에서는 ① 인공지능 자체의 윤리, ② 설계자·제작자의 윤리, ③ 이용자의 윤리를 중심축으로 형성하게 되었다.[12]

윤리규범은 인공지능이 초래하는 해악과 위험의 방지를 주요한 목적으로 삼는 특성상 의무와 책임에 대한 내용이 많다. 이에 전통적으로 의무와 책임의 귀속이 보다 용이한 '인간의 윤리(위의 ②, ③ 유형)'가 오늘날 인공지능 윤리의 주요한 관심사에 해당하게 되었다. 그러나 최근 들어 인공지능 기술

8) 고인석, "아시모프의 로봇 3법칙 다시 보기: 윤리적인 로봇 만들기", 『철학연구』 제93집, 2012, 102면.

9) Robin R. Murphy/David D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics", IEEE Intelligent Systems Vol. 24, No. 4, 2009, pp. 14-20은 이러한 문제를 지적하고, 3원칙의 대안을 제시한다. 핵심은 인간과 로봇의 협동적 작업체계를 전제하고, 로봇의 보호를 위해 모종의 자율성을 긍정하지만, 기본적으로 로봇의 도구적 성격을 중심에 둔다는 것이다. 이에 대한 비평으로 고인석, 앞의 논문, 104면 이하 참조.

10) A. Takanishi, "World Robot Declaration", International Robot Fair, 2004. (http://prw.kyodonews.jp/prwfile/prdata/0370/release/200402259634/index.html)

11) Gianmarco Veruggio, "EURON Roboethics Roadmap(Release 1.1)", EURON Roboethics Atelier, Genua, 2006, p. 7. 이듬해 발표된 로드맵 1.2버전도 큰 차이는 없다.

12) Peter M. Asaro, "What Should We Want From a Robot Ethic?", International Review of Information Ethics Vol. 6, No. 12, 2006, pp. 9-16; 정채연, 앞의 논문, 90-93면 참조.

이 급격히 발전하면서, 로봇 3원칙 이후 실무적 관심에서 멀어져온 '인공지능(로봇)의 윤리'가 일부 조명을 받고 있다. 그러한 면에서, 인공지능의 윤리에 관한 논의는 크게 두 가지 방향으로 전개되고 있다. 하나는 인공지능의 의사결정이 인간의 윤리적 직관과 부합하도록 하는 일련의 기술적 방법론에 초점을 맞추는 것이고, 다른 하나는 인공지능에 법인격을 부여하는 등의 방식으로 (적어도 법적으로) 의무와 책임을 다하게 하려는 규범적 논의를 강조하는 것이다.

전자인 윤리적 의사결정의 주체로서의 인공지능에 대한 기술적 접근은 '인공적 도덕행위자(Artificial Moral Agent, 이하 'AMA')'로 불린다.[13] AMA를 제작하는 기술적 방식은 공리주의나 의무론을 비롯한 전통적 윤리규칙을 학습시키는 방식(하향식 접근), 덕 윤리에서 강조하듯 윤리적 행위로 볼 만한 사례를 학습시키는 방식(상향식 접근), 양자를 결합한 방식(혼합식 접근)으로 대별된다.[14] 가장 널리 알려진 하향식 접근은 자율주행 자동차가 사고에 직면한 상황에서 탑승자와 보행자 중 누구를 우선시할지와 같은 현실적 문제(일명 '트롤리 딜레마')에 적용할 경우 한계가 지적되고 있다. 예를 들어, 공리주의 원칙을 적용한 자율주행 자동차는 2명 이상의 보행자를 살릴 수 있다면 탑승자를 사망케 하는 의사결정을 할 텐데, 소비자가 그러한 자율주행 자동차를 구매하지 않을 것이기에 시장의 성립 자체가 어려울 수 있다는 것이다.[15] 하향식 접근은 모든 딜레마 상황에 특정 윤리원칙을 적용하여 유연성이 부족해지는 탓에 한계를 낳는 셈이다. 한편, 상향식과 혼합식 접근은 하향식 접근에 비해서는 발전이 더딘 상황이다.

그렇다면 이와 같은 대중의 반발과 불안감은 무엇에서 기원하는 것일까? 공공장소인 도로에서 자율주행 자동차가 산출한 결과물은 공적 의사결정의 일종으로 볼 수 있는데, 이를 대중이 의식적으로 통제할 수 없는 상황이 나타날 수 있다는 것이 문제의 발단인 것으로 보인다. 사적 주체(기업) 내지는 도덕적·법적 주체에 해당하는지조차 의문시되는 기계가 개인의 생명이라는 중대한 기본권이 달린 공적 의사결정을 좌지우지하는 이러한 상황은 적법절차나 민주적 정당성이 결여되어 있다는 문제의식이 그 배경에 있는 것으로 볼 수 있다.[16] 이에 대해, 자율주행 자동차의 의사결정이 대중의 선호를 그대로 반영하면 민주적 정당성을 확보할 수 있다는 전제 하에, 딜레마 상황에 관한 설문조사를 통해 집단 별 대중의 선호를 파악하는 연구가 이루어지기도 하였다.[17] 그러나 이러한 방식도 한계가 있는데, 가능한 딜레마 상황은 사실상 무한하지만 설문조사는 유한한 경우로 제한될 수밖에 없고, 이를 기술적으로 구현해내는 데도 한계가 있기 때문이다. 윤리원칙에 의한 접근방식은 유연성과 민주적 정당성이 부족한 반면, 윤리적 사례에 의한 접근방식은 '귀납의 문제'로 불리는 일반화에 취약한 셈이다. 결론적으로 원칙과 사례 중심의 윤리적 접근이 가진 전통적 문제점이 인공지능의 의사결정에

13) AMA의 도덕적 영향력을 기준으로 유형을 세분화하기도 한다. 무어(Moor)는 AMA를 ① 외부세계에 윤리적 영향을 끼치는 행위자(ethical-impact agent), ② 설계자가 작동과정에 어떤 도덕적 제약을 삽입한 암묵적 윤리적 행위자(implicit ethical agent), ③ 모종의 도덕규칙까지 삽입한 명시적 윤리적 행위자(explicit ethical agent), ④ 도덕적 판단을 학습하고 수립하는 완전한 윤리적 행위자(full ethical agent)로 유형화한다. James H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics", IEEE Intelligent Systems Vol. 21, No. 4, 2006, pp. 19-21 참조. 이때 유형 ②는 최근 GDPR 제25조에서 언급한 설계나 초기설정에 의해 프라이버시를 강화(data protection by design and by default)하는 방법론과 일맥상통한다. 현재 유형 ③을 완전히 구현한 AMA로 인정된 사례는 찾기 어렵다.

14) 보다 자세한 내용은, 웬델 월러치/콜린 알렌(노태복 역), 『왜 로봇의 도덕인가』, 메디치미디어, 2014, 6장-8장; 변순용, 『윤리적 AI로봇 프로젝트』, 어문학사, 2019, 2부 참조. 이러한 내용을 자율주행 자동차의 사례에 응용한 선행연구로, 이중기/오병두, "자율주행자동차와 로봇윤리: 그 법적 시사점", 『홍익법학』 제17권 제2호, 2016, 5면 이하도 참조.

15) Jean-François Bonnefon/Azim Shariff/Iyad Rahwan, "The Social Dilemma of Autonomous Vehicles", Science Vol. 354, No. 6293, 2016, pp. 1573-1576.

16) Nicholas Diakopoulos, "Accountability in Algorithmic Decision Making", Communications of the ACM Vol. 59, No. 2, 2016, pp. 58-59 참조.

17) Edmond Awad et al., "The Moral Machine Experiment", Nature Vol. 563, 2018, pp. 59-64. 이 연구는 지금까지도 한국어를 포함한 10개 언어로 진행되고 있다. 홈페이지는, http://moralmachine.mit.edu/.

서 그대로 되풀이되는 현실에 봉착하게 되었다. 이러한 문제를 해결하기 위한 다양한 시도가 이어지고 있지만, AMA를 통해 인공지능 윤리를 달성하려면 앞으로도 많은 후속연구가 필요해 보인다.

2. 책임성(accountability) 개념에 대한 이해방식의 확장

이러한 이유로 인공지능의 의사결정을 윤리적으로 만들기 위한 직접적 노력과 별개로, 인공지능 자체나 배후 인간 이해관계자에게 모종의 행위규범을 부과하여 간접적으로 규율하려는 방안이 부각되었다. 전통적으로 법규범, 윤리규범과 같은 사회규범에서 이러한 기능을 매개하는 수단으로 통용된 개념이 바로 '책임'이고, 구체적 맥락에 따라 다양한 방식으로 활용되고 있다. 대표적 사회규범인 윤리와 법 분야에서의 'responsibility'와 'liability'가 이에 대응되는 용어이다. "법은 도덕의 최소한"이라는 격언에 의하면 전자가 후자보다 더 넓은 개념으로 볼 수 있지만, 반드시 그렇지는 않다. 지금까지 양자의 관계에 대한 많은 논의가 있었고, 간단명료한 해답이 도출되지는 않았다. 다만, 법과 윤리는 때로는 법이, 때로는 윤리가 서로를 이끌어가는 상호보완적 관계에 있다는 것이 중론이다.[18]

법과 윤리는 오랫동안 비슷한 전제를 공유해왔다. 책임의 대상은 '행위(act)'로, 행위는 자율적 주체인 인간의 자유의지로부터 비롯되고, 주체는 스스로의 행위가 낳는 결과를 예견할 수 있으며, 행위와 결과의 인과관계가 명확하다는 것이 일반적으로 전제된다.[19] 이러한 관점에서 바라보면, 인공지능 책임성의 기본적 내용은 인공지능을 이용한 의사결정에 의해 사회적 문제가 발생하였을 때, 원인이 되는 행위를 한 주체에게 귀속되는 도덕적, 법적 행위책임을 의미하게 된다. 법적 책임 개념(liability)은 도덕적 책임 개념(responsibility)으로부터 많은 영향을 받은 것으로, 두 가지 책임 개념 모두 '행위책임' 개념을 대전제로 공유한다. 영국의 공학·물리학 연구위원회(Engineering and Physical Science Research Council, EPSRC)가 2010년에 발표한 '로봇윤리 5원칙'[20]에서 '법적 책임의 인간에 대한 귀속'을 명시하고 '안전과 보안'을 특별히 강조한 것은 그러한 흐름의 일환으로 파악할 수 있다.

그렇지만 오늘날 인공지능의 기술적 특징은 이러한 행위책임 개념과 자연스레 조응하지 않는 면이 많으므로, 전통적 책임관을 그대로 적용할 경우 책임격차가 발생하여 문제해결이 어려울 수 있다. 예컨대, 마이크로소프트에서 2016년 3월 개발한 채팅봇 테이(Tay)가 악성 이용자에게 인종차별적 발언을 학습하여 사회적으로 논란이 발생한 일을 떠올려보자. 만일 혐오표현을 학습시킨 이용자를 찾아낼 수 있다면, 해당 이용자를 제재하여 문제를 해결하는 것이 가능할 수도 있다. 그러나 인공지능의 훈련 데이터가 과거의 사회적 편견을 반영하여 차별적 판단을 재현하는 상황이라면 문제를 유발한 주체가 불특정 다수인 과거의 인류가 될 것이어서 책임규명이 어려워진다. 그밖에도 무수히 많은 참여자나 데이터가 개입하는 오늘날 인공지능 제작과정을 생각한다면 행위책임 관념을 전제로 책임소재를 가리는 작업은 사실상 불가능에 가까운 경우가 많다(이를 '많은 손(many hands)의 문제'라고 한다).[21] 책임을 귀속할 행위자가 없어지거나, 반대로 너무 많아지는 일이 잦기 때문이다.

행위나 인과관계와 같은 객관적 법률요건에 가해지는 책임귀속의 부담에 더해 귀책사유와 예견가능성과 같은 주관적 법률요건에 가해지는 부담도 만만치 않다. 오늘날의 '약인공지능(weak AI)'에 대해 자유의지를 인정하기 어렵다고 대체로 공감하

18) 김건우, 앞의 논문, 33면 이하 참조.

19) Merel Noorman, "Computing and Moral Responsibility", Stanford Encyclopedia of Philosophy, 2018.

20) Engineering and Physical Science Research Council, "Principles of robotics", 2010. (http://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/)

21) Helen Nissenbaum, "Accountability in a Computerized Society", Science and Engineering Ethics Vol. 2, No. 1, 1996, pp. 28-32.

는 상황에서, 데이터나 알고리즘의 불완전성에 의해 창발한 차별적 현상에 누군가의 의도나 예견가능성을 인정하기 어렵다는 문제의식이 대표적 사례다.[22] 예를 들어, 훈련과정의 과적합(overfitting)이나 과소적합(underfitting), 훈련 데이터에 내재된 역사적 차별이나 대표성 부족과 같은 차별적 현상을 유발하는 대표적 경로를 모두 차단하고서도 인공지능 의사결정의 사회적 의미가 차별적으로 인식될 경우, (간접차별은 별론으로 하고) 명시적 차별의도를 인정하기는 어렵다. '불가능한 의무를 요구할 수 없다'는 자명한 명제를 떠올려본다면, 규범적 의무가 기계의 구문론적 구조와 인류의 의미론적 구조가 상충하는 현상에까지 확장되는 결론은 과도하기 때문이다. 다른 한편, 인간이 생각하지 못한 방식의 추론을 통해 경쟁우위를 확보하는 인공지능의 본질을 생각하면 예견가능성이 낮은 점 자체를 규범적으로 문제 삼는 것 자체가 부적절해 보이는 면도 있다. 이처럼 행위책임에 대한 기존의 견해를 고수할 경우 책임공백이 발생하는 국면이 지속적으로 늘어나게 될 것이다.

나아가 최근 들어 주로 활용되는 인공신경망 기반 머신러닝, 딥러닝 인공지능 의사결정의 불투명성(opacity)은 책임공백을 더욱 확대할 수 있다.[23] 이와 같은 인공지능은 복잡하게 연결된 인공신경망에 기초하여 인과관계(causation)가 아닌 상관관계(correlation)에 의존하는 등 인간과 전혀 다른 방식의 의사결정을 하여 종종 블랙박스(black-box)에 비견되고는 한다. 모형에 따라서는 일반인은 물론 제작자조차 이해하기 어려운 오늘날의 인공지능은 사전적 규제나 사후적 책임귀속에 많은 난점을 가져오게 된다.[24] 인공지능의 불투명성을 해소하기 위한 규범적 노력은 '투명성(transparency)' 원칙의 강조로 이어지는데, 이로부터 파생된 책임 개념이 '설명책임(accountability)'이다.[25] 설명책임과 투명성 원칙은 인공지능 의사결정의 영향을 받는 당사자에게 중요한 정보가 제공되어야 한다는 관점에서 일맥상통한다. 그러나 설명책임은 인공지능의 불투명성을 극복할 수 있을 만큼의 실질적 '이해'까지를 책임의 기준으로 부과한다는 측면에서 좀 더 심화된 개념이라 할 수 있다. 위에서 언급한 GDPR의 '설명을 요구할 권리'는 인공지능의 작동과정에 대한 설명책임을 실정법에 구현한 최초의 사례라고 볼 수 있다. 설명책임을 구현하는 구체적 방법론에 대한 논의는 아직 초기 단계에 있는데, 인공지능 감사(audit)와 같은 공법적 규제에서부터 설명책임을 달성한 정도에 따라 면책의 수준을 달리 인정하는 방안과 같은 사법적 해석론까지 다양한 논의가 이루어지고 있다.

한편, 현재 책임 개념은 어떤 구체적 문제가 발생했을 때 누군가에게 행위책임 또는 설명책임을 귀속하여 책임공백이 나타나는 것을 방지하여야 한다는 '이해관계자의 책무성' 개념으로까지 확장되어 논의되는 양상이다. 사회적으로 적절히 관리되기 어려울 정도의 커다란 위험(risk)을 창출한 주체에게 해악을 예방하거나 회복하는 작업에 대한 행위책임과 설명책임을 귀속시키자는 주장이다.[26] 이러한 이유로 오늘날 국내 문헌은 'accountability' 개념을 논의 맥락에 따라 '설명책임', '책임(성)', '책무

22) Solon Barocas/Andrew D. Selbst, "Big Data's Disparate Impact", California Law Review Vol. 104, 2016, p. 677 이하; 오요한/홍성욱, "인공지능 알고리즘은 사람을 차별하는가?", 『과학기술학연구』 제18권 제3호, 2018, 163-167면; 남중권, "머신러닝 알고리즘의 데이터 처리에 대한 법적 제한의 한계: 개인정보보호와 차별금지의 측면에서", 『과학기술과 법』 제10권 제1호, 2019, 86-87면 등 참조. 인공지능 기술이 주관적 요건의 귀속을 과거에 비해 어렵게 하는 것은 사실이지만, 현행법제에서 탄력적 법해석을 통해 주관적 요건의 귀속이 절대적으로 불가능하다고 단정하기는 어렵다는 지적도 있다. 고학수/정해빈/박도현, "인공지능과 차별", 『저스티스』 통권 제171호, 2019, 231-233면.

23) 여기서의 불투명성은 ① 인공지능 보유주체의 지식재산권이나 계약상 특약조항을 비롯한 '제도적 측면', ② 인공신경망의 복잡성과 같은 알고리즘의 '본질적 측면', ③ 검증과정의 인적, 물적 비용상의 문제로 비롯된 '현실적 측면'과 같은 다종다양한 이유로 인해 발생한다. 고학수/정해빈/박도현, 앞의 논문, 235-236면.

24) 대표적 사례로, Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", Big Data & Society, 2016, p. 9.

25) Robyn Caplan et al., "Algorithmic Accountability: A Primer", Data & Society, 2018, p. 10.

26) 신상규, "인공지능 시대의 윤리학", 『지식의 지평』 제21호, 2016, 11면 이하; 이중원, "인공지능에게 책임을 부과할 수 있는가?: 책무성 중심의 인공지능 윤리 모색", 『과학철학』 제22권 제2호, 2019, 89면 이하.

(성)'으로 달리 번역하고 있다. 여기에 더하여 인공지능 기술에 관한 법적 책임 논의에서는 두 가지 생각이 추가적으로 개입된다. 하나는 오늘날 인공지능이 보이는 증대된 (현상적) '자율성(autonomy)'[27]의 정도를 근거로 인공지능 자체에 대해 모종의 책임을 부과해야 한다는 문제의식이다. 다른 하나는, 도덕과 달리 정책적 고려가 개입되는 법의 특성상, 법적 책임분배와 인공지능 기술발전의 조화가 필요하다는 문제의식이다. 창출된 위험의 옳고 그름을 도덕적으로 판단하는 문제와는 달리, 법적으로 위험을 제거하려면 비용(cost)이 수반되게 마련이기 때문이다. 비용과 책임의 정도를 결부한다면 두 가지 논점은 사실상 하나로 연결된 셈이기도 하다. 위험을 없애는 데 드는 단기적 비용과 성장을 절감하는 장기적 비용을 고려하면, 법적 규제의 문제는 공리주의와 의무론을 아우르는 고차방정식으로 환원된다.[28]

유럽연합이 2012년부터 3년 동안 진행해온 소위 '로봇법 프로젝트(RoboLaw Project)'에 기초해 2014년 발표한 '로봇규제 가이드라인'은 이러한 문제의식을 종합하여 3가지 규제기준을 제시하였다.[29] 첫째, 가장 기술친화적 대안으로, 로봇산업의 혁신을 촉진하고 규제비용을 절감하기 위해 기술적으로 불가피한 책임의 경우에는 (적어도 단기적으로는) 면책(immunity)을 인정하자는 방안이다. 여기에 해당하는 위험은 일종의 '허용된 위험'에 해당하므로, 보험과 같은 별도의 기제를 통해 기술발전의 혜택을 누리는 공동체가 책임을 분담해야 한다는 것이다. 둘째, 일정 수준의 (현상적) 자율성을 가진 로봇에 대하여 법인격(legal personhood)을 부여하는 방안이다. 이러한 방법론은 장기적으로 로봇에 대한 '전자인(electronic person)' 지위를 고려할 필요가 있다는 유럽연합 의회의 결의안[30] 제59항 f호로 이어졌다. 다만, 이에 대해 기존의 기술적·윤리적·법적 관점과 마찰을 일으키고 인간의 책임의식을 약화할 수 있다는 관련 분야 전문가들의 문제제기가 있었고,[31] 이 논의를 반영한 법제화가 실제로 진행되지는 않았다. 셋째, 기본권 보호 정도가 가장 높은 대안으로, 특수불법행위나 제조물책임법과 같은 일부 민사특별법에 마련된 무과실책임(strict liability) 원칙을 응용하여 법제화하는 방안이다. 사고가 발생할 때 규명하기 힘든 법률요건에 대한 증명책임을 완화하거나 제작자나 설계자에게 돌리는 방안도 유사한 사고방식에 입각한 견해로 볼 수 있다.

이러한 세 가지 방안은 도덕적 책임(responsibility)의 귀속이 어려워지는 현실적 문제에 대해 민·형사상의 법적 책임(liability) 측면에서 어떻게 접근할 것인지에 관한 체계를 만들어 내기 위한 중요한 시도였다고 평가할 수 있다. 그런데 최근 들어 인공지능의 자율적 의사결정이 점차 늘어나면서, 사회적으로 새로운 차원의 문제제기가 이루어져왔다. 앞에서 언급한 인공지능의 불투명성으로 인해 사회구성원의 알 권리와 절차적 참여권이 제대로 보장되지 못하는 문제가 그것이다. 인공지능의 의사결정이 인류에게 점차 중대한 영향력을 행사하게 될 것으로 보이지만 그것이 어떠한 근거로 이루어졌는지, 어떠한 절차를 거쳐 이의제기를 할 수 있는지 등에 대한 문제의식은 상대적으로 미약하였다는 생각인 것이다.[32] 예를 들어,

27) 전통적으로 자율성(autonomy)은, 자유의지에 따라 어떠한 행위를 자유롭게 할 수 있다는 소극적 자유의 측면과, 그러한 행위가 사회적으로 정언명령 같은 보편화되는 계율에 의하여 정당화되어야 한다는 자기규율의 측면을 포괄한 도덕적 개념으로 널리 활용되어 왔다. 최근 인공지능 기술의 발달에 따라 전자의 측면이 극대화되면서 인공지능이 인간의 예측가능성과 통제가능성의 영역에서 벗어나는 현상이 증가하고 있고, 이러한 현상에 대해 '자율성'이라는 명칭을 부여하는 문헌이 점증하는 상황이다. 엄밀히 말하자면 이는 과거의 도덕적 자율성과 별개의 영역으로, 가령 '현상적 자율성'이라는 명칭을 부여하여 분리하는 것이 바람직하다. 보다 자세한 내용은, 박도현, "인공지능과 자율성의 역학관계", 『홍익법학』 제20권 제3호, 2019, 505면 이하 참조.

28) 이러한 문제의식을 집대성한 선행연구로, 이원우 외, 『4차 산업혁명 시대의 기술혁신과 규제정책』, 홍문사, 2019 참조.

29) Erica Palmerini et al., "Guidelines on Regulating Robotics", RoboLaw Project, 2014, pp. 23-24.

30) European Parliament, "European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))", 2017.

31) Nathalie Nevejans et al., "Open Letter to The European Commission Artificial Intelligence And Robotics", 2018. (http://www.robotics-openletter.eu/)

테이 사례의 피해자가 민사상 손해배상을 청구할 수 있는지 여부는 별론으로 하더라도, 기업에 인공지능의 혐오표현을 근거로 인공지능 사용을 중단하라는 취지의 청구권을 행사할 수 있는지는 사적자치의 원칙에 따라 현행법상 논란의 대상이 된다. 이처럼 '기술영역의 적법절차(technological due process)'로 불리는 일련의 문제제기가 이어지면서 참여와 민주주의 측면에서 책임성에 대한 또 다른 이해방식이 대두되었다.[33]

인공지능의 규율과 관련하여 '거버넌스(governance)'라는 용어가 종종 사용되는 이유도 유사한 맥락에서 설명될 수 있다. 제재의 부과에 초점을 둔 하향식(top-down)의 규범체계를 넘어선, 구성원의 자율적 문제해결과 대중의 능동적 참여를 강조하는 상향식(bottom-up)의 규범체계와 밀접하게 연관된 용어로 해석될 수 있기 때문이다.[34] 그렇게 보면, 'accountability'는 종래 윤리적, 법적 책임 개념(responsibility, liability)이 담당해온 역할과, 설명책임과 관련된 책임성이나 책무성의 부과, 그리고 대중의 능동적 참여기능을 더한 광의의 책임성 개념으로 볼 수 있다. 따라서 진입규제나 민·형사 제재와 같은 전통적 방식의 규제뿐만 아니라, AMA나 '설명가능 인공지능(explainable AI, 이하 'XAI')'[35]과 같은 기술적 접근방식도 얼마든지 인공지능 의사결정에서 책임성 원칙을 달성하는 적절한 수단이 될 수 있다. 이와 병행하여 사회구성원의 인공지능 문해력(literacy)을 증진시키고자 하는 장기적 노력이나, 인공지능 활용에 따른 혜택을 누리고 위험을 초래한 당사자가 비용을 분담하는 법적 책임성의 확장(위험책임, 편익책임)[36]도 책임성을 구현할 수 있는 한 가지 방편이 될 수 있다. 그렇게 보면 인공지능 시대에서의 '책임성(accountability)' 원칙이란 인공지능의 활용을 인간과 인공지능의 공존공영으로 이끄는 사회규범을 총체적으로 언급하는 것으로 재구성하여 이해할 수 있다. 이에 따르면, 인공지능의 활용에 제약을 두는 것이 책임성을 강화하는 전형적인 형태인 것으로 파악할 수도 있지만, 오히려 인공지능의 적극적인 활용을 장려하는 동시에 책임성을 강화하는 접근도 가능하게 된다. 인공지능 책임성 원칙은 오늘날 인공지능 규범 논의에서 핵심적 원칙 중 하나로 받아들여지고 있다.

3. 이중효과의 조화와 구체적 맥락 중심의 접근방식

다른 한편, 인공지능의 복잡성에 따른 창발성(emergence)과 예측불가능성은 규범적 평가를 일률적으로 행하기 어렵게 만드는 요인이 된다. 예를 들어, 인공지능 의사결정의 차별적 효과를 검증하는 과정에서 때로는 이해관계자의 개인정보나 민감정보를 수집할 필요가 있는데, 이는 프라이버시라는 또 다른 중대한 기본권과 정면으로 충돌하기 쉽다.[37] 프라이버시와 차별금지 중 어느 하나를 근거로 인공지능 의사결정을 전면적으로 긍정하거나 부정하는 이분법적 사고방식의 한계를 보여주는

32) 프랭크 파스콸레는 이러한 상황을 두고 '블랙박스 사회(black-box society)'로 표현한다. 프랭크 파스콸레(이시은 역), 『블랙박스 사회』, 안티고네, 2016; 이와 유사한 문제의식을 담은, 캐시 오닐(김정혜 역), 『대량살상 수학무기』, 흐름출판, 2017도 참조.

33) '기술영역의 적법절차 원리'를 정식화해 도입한, Danielle Keats Citron, "Technological Due Process", Washington University Law Review Vol. 85, 2008; 책임성 개념에 대한 참여 측면을 강조한, Richard Mulgan, "'Accountability': An Ever-Expanding Concept?", Public Administration Vol. 78, No. 3, 2000, pp. 569-570 참조.

34) Gerry Stoker, "Governance as Theory: Five Propositions", International Social Science Journal Vol. 50, No. 155, 1998, p. 2 참조.

35) 오늘날 인공지능의 높은 불투명성을 극복하기 위한 기술적 방법론을 포괄하는 개념으로, 알고리즘 자체를 확보하지 못한 상황을 전제하는 블랙박스 접근법(black-box approach)과, 알고리즘 자체를 확보한 상황을 전제하는 화이트박스 접근법(white-box approach)으로 대별된다. 이에 대해, Claude Castelluccia/Daniel Le Métayer, "Understanding Algorithmic Decision-Making: Opportunities and Challenges", Panel for the Future of Science and Technology, 2019, p. 47 이하 참조.

36) 새로운 책임원칙으로 대두되고 있는 위험책임주의와 편익책임주의에 대한 자세한 설명은, 오병철, "인공지능 로봇에 의한 손해의 불법행위책임", 『법학연구』 제27권 제4호, 2017, 201면 이하 참조.

37) 고학수/정해빈/박도현, 앞의 논문, 255면.

단면이다. 복수 기본권 사이의 충돌이 나타날 수 있을 뿐만 아니라 단일한 기본권으로부터 유발된 효과 사이에서도 충돌이 나타날 수 있다. 예컨대, 인공지능의 활용은 누군가에게는 노동권의 신장을, 누군가에게는 노동권의 박탈을 낳는다.[38] 노동권이라는 단일한 기본권만을 근거로 해서는 인공지능의 도입에 대한 규범적 판단에 어려움이 있을 수 있음을 암시한다. 인공지능이 가져오리라고 예측되는 수많은 긍정적 효과에도 불구하고 민주주의, 경쟁, 자율성과 같은 다양한 사회적 가치를 약화할 우려가 공존하는 상황에서, 인공지능의 도입에 따른 파급효과를 일의적으로 평가하기란 사실상 불가능하다.

사실 인류의 역사를 통틀어 보면 신기술의 출현은 주기적으로 일어난 극히 일상적 현상이었고, 그에 대한 우려와 불안의 목소리도 결코 새로운 일로 보기 어렵다. 기술의 본성을 논하는 분야인 '기술철학(Philosophy of Technology)'의 기원을 그리스 시대로 거슬러가는 이유가 여기에 있다. 다만, 현대적 의미의 기술철학 분야의 시작점은 세계대전 전후인 20세기 초중반으로 바라보는 견해가 대다수이다. 하이데거(Martin Heidegger)·엘륄(Jacques Ellul)로 대표되는 당시 기술철학은 기술 일반에 대한 부정적 시각을 견지하는 비판적 입장이 주류를 형성하였다.[39] 기술이 인간의 자율성을 앗아가고 인간이 기술에 종속되는 현상을 지적한 이와 같은 입장은 시대상을 반영한 정당하고 중요한 것이었지만, 동시에 이 입장은 기술에 대한 시각이 편협하고 개별적·구체적 기술에 대한 이해가 부족하다는 비판에 직면하기도 하였다. 그리하여 1970년대 이후의 기술철학은 이른바 '경험으로의 전환'이라는 명제로 대표되듯, 개별적·구체적 기술에 대한 이해를 강조하는 방향으로 선회하였다.[40] 이후 컴퓨터 윤리, 생명윤리와 같이 인류에게 막대한 파급효과를 낳는 신기술에 대한 윤리적 쟁점을 다루는 분야는 경험으로의 전환을 반영하여 개별적·구체적 쟁점에 초점을 두었다. 인공지능 윤리 역시 같은 관점에서 파악해야 할 것이다.

그렇기에 경제성장 같은 편익의 부산물로 나타나는 해악의 위험이라는 일종의 '이중효과(double effect)'를 공동체 구성원의 선호를 합산할 때 불가능성을 낳는 원천으로만 볼 것은 아니다. 부가가치의 창출과 같은 일반론적 이익을 넘어 자율주행 자동차는 장애인이나 노약자의 이동권을 강화하고, 드론은 물리적 접근성이 낮은 지역의 운송에 기여하는 등 신기술을 통한 순기능은 분명히 실재한다. 다른 한편, 2018년 우리나라에서 논란이 나타나기도 한 '킬러 로봇'을 포함하여 '치명적 자율무기(Lethal Autonomous Weapons, LAWs)'의 연구개발·사용에 관하여는 국제사회의 논의에 적극적으로 참여하고 부작용을 방지하기 위해 노력하여야 할 필요가 있다. 이처럼 이중효과를 개별적·구체적 맥락별로 파악하여 순기능은 활용하고 역기능은 방지함이 타당한 반면, 특정 신기술에 대한 원천적 금지를 규범적 대안으로 제시하는 접근방식은 일반적으로 바람직하지 않다.

이처럼 인공지능에 대한 선험적 판단 대신 구체적 의사결정 상황별로 규범적 판단이 달라져야 한다는 사고를 '맥락(context) 중심의 접근방식'이라고 부른다. 인공지능은 자율성을 증진하거나 억제할 수도, 인류에게 혜택이 되거나 해가 될 수도 있음을 전제로, 개별적·구체적 맥락을 고려하여 규범적·정책적 판단을 내려야 한다는 것이다. 맥락 중심의 접근방식은 인공지능 규범논의에서 점차 강조되고 있다. 앞서 본 유럽연합의 로봇규제 가이드라인은 일반론 대신 사례별 접근을 택하고, 규제와 산업발전이 공존 가능함을 전제로 하며, 기술을 바탕으로 한 자율규제, 윤리원칙, 엄격한 법제 사이의 상호 조화를 추구하는 프레임워크를 제시하여 이러

38) U.S. Executive Office of the President, 앞의 글, p. 10 이하 참조.

39) Maarten Franssen/Gert-Jan Lokhorst/Ibo van de Poel, "Philosophy of Technology", Stanford Encyclopedia of Philosophy, 2018.

40) 손화철, "기술철학에서의 경험으로의 전환: 그 의의와 한계", 『철학』 제87집, 2006, 145면 이하 참조.

한 사고를 반영하였다.[41] '인공지능'이라는 어떤 단일의 실체를 전제로 하여 그 본성을 규명하고 개념을 정의하려는 시도보다는, 자율주행 자동차나 로보 어드바이저처럼 현실에서 실제로 활용되는 사례에 주목하여야 한다는 견해와도 일맥상통한다.

지금까지 인공지능 윤리 분야의 변천사를 추적하면서 주요 쟁점을 살펴보았다. 오늘날의 인공지능 거버넌스는 큰 틀에서 이와 같은 이해방식을 공유하고 그에 기초한다. 그러나 주체의 특성과 같은 기준 하에 비교해보면, 강조점과 구속력의 정도 등에 있어 어느 정도 차이가 발견된다. 국제사회에서 이루어지는 논의의 공통점과 차이점을 분명히 파악하였을 때, 비로소 우리나라의 실정에 맞는 특유한 규범을 확립할 수 있고, 국가적 차원에서 이러한 국제사회의 논의에 적극적으로 참여하고 주도해야 할 필요도 있다. 이상의 필요성에 비추어, 이하에서는 해외의 최신 윤리규범 내용과 그에 기초한 거버넌스 구조를 비교·분석해보도록 한다.

Ⅲ. 해외의 인공지능 윤리규범 및 규제 거버넌스 비교·분석

1. 논의의 출발점

이하에서는 앞서 소개한 인공지능 윤리 영역의 주요 쟁점에 관한 국제사회의 최신 윤리규범과 규제 거버넌스를 고찰한다. 인공지능 윤리 영역의 주요 쟁점이 윤리규범과 규제 거버넌스 형태로 수렴하는 이유는, 윤리규범을 마련하여 중요한 윤리 이슈를 선별하거나 가이드라인을 제시할 수 있고, 규제 거버넌스라는 기제를 통해 그러한 내용을 현실에 확립해야 할 필요가 있기 때문이다. 윤리규범과 규제 거버넌스는 윤리 이슈에 대한 단순한 문제제기를 넘어, 구체적 해결책을 현실에 제시하는 첫걸음인 셈이다.[42]

서론에서 언급하였듯, 인공지능 기술이 점차 상용화되고 고도화되면서 정부와 산업계를 불문하고 인공지능 윤리규범과 규제 거버넌스에 대한 논의가 활발하게 이루어지고 있다. 그러나 논의의 구체적 내용을 살펴보면, 다루는 이슈의 범위나 깊이의 양적·질적 측면에서는 적지 않은 차이가 있다. 인공지능을 연구개발하고 활용하는 과정에 적용되는 일반적·기본적 원칙을 천명한 수준에 그치는 경우도 있고, 인공지능이 이용되고 있거나 가까운 시일 이내에 이용될 것으로 예측되는 분야의 상세한 현실적 이슈까지 다룬 경우도 있다. 어떤 경우는 인공지능 윤리규범이나 규제 거버넌스를 어떻게 정립할지에 대한 구체적 방법론까지 언급하기도 하였다. 이하에서는 논의의 범위와 깊이를 기준으로, ① 기본원칙 중심의 논의 유형, ② 기본원칙과 심화된 이슈를 함께 다루는 논의 유형, ③ 윤리규범과 규제 거버넌스 정립방안에 대한 구체적 논의 유형으로 구분하여 구체적 사례를 분석한다. 이와 같은 구분은 상당히 작위적인 것이고 개별 논의에 따라서는 어느 한 유형에 편입하여 고려하는 것이 어색한 경우도 있지만, 본격적인 검토를 위한 출발점을 제공해주는 편리함이 있다.

2. 기본원칙 중심의 논의

기본원칙 중심의 논의는 인공지능을 연구개발하거나 활용할 때 준수하여야 할 대전제나 주요 윤리원칙을 간략하게 담고 있는 선언 유형의 논의이다. 규제 거버넌스와 같은 구체화된 대안은 제시하지 않기에, 다소 추상적이고 규범력이 떨어지는 논의로 여겨질 수 있는 경우가 적지 않다. 현실에서도 논의의 양과 질이 어느 정도는 비례하는 편이기도 하다. 다만, 이러한 유형의 논의를 폄하할 일만은 아닌 것이, 이와 같은 방식의 논의도 어느 정도는

41) Erica Palmerini et al., 앞의 글, p. 8 참조.

42) Alan F. T. Winfield/Marina Jirotka, "Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence Systems", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences Vol. 376, No. 2133, 2018, p. 3 참조.

실질적 유용성을 가질 수 있기 때문이다. 첫째, 꾸준히 발전하고 변화하는 '인공지능' 기술의 특성상, 지나치게 구체화된 논의에 돌입할 경우 변화된 현실과 쉽게 괴리될 수 있는 위험에 직면한다. 향후 꾸준히 문제될 대전제나 주요 윤리원칙에 대해 합의하여 명문화하는 일만으로도 상당한 유용성을 갖는 이유가 여기에 있다. 둘째, 인공지능 윤리 분야가 새로이 대두된 탓에 어떤 이슈가 존재하고 나타날지에 대한 정보가 충분하지 않은 편이다. 다양한 영역에 걸쳐 파편화된 정보를 수집하여 개별 문서에 반영하는 작업만 하더라도 초기에는 논의를 상당히 진전시키는 요인이 된다. 셋째, 인공지능 기술이 가져올 것으로 예상되는 막대한 부가가치로 인해 다양한 이해관계자들이 개입되어 있는 현실도 중요한 요인으로 꼽힌다. 원칙별로 수많은 이해관계자가 제각기의 생각을 가지고 있고, 극단적으로는 인공지능 윤리에 대한 논의 자체가 무의미하다고 보는 이해관계자도 존재할 수 있다. 자연스레 어떤 식의 합의점을 도출하기 위해서는 누구도 부정하기 어려운 극히 추상적인 차원의 대원칙 정도만이 문서에 반영될 수도 있는 것이 국제사회의 현실이다. 이 중에서 대체로 초기 논의에서는 첫째·둘째 이유가 지배적이었던 반면, 최근 들어서는 셋째 이유가 기본원칙 중심의 논의를 낳는 보다 주요한 원인이 되고 있는 것으로 보인다. 그 이외에, 이처럼 발전적 이유보다는 국제사회나 시민사회에서 가하는 비판을 피하기 위해 장식적 성격이 강한 피상적인 논의가 나타나는 경우도 있는 것으로 보인다. 이러한 피상적 논의를 일의적으로 규정하기란 쉽지 않지만, 최신의 논의인데도 구체성이 떨어지고 후속적 검증이나 거버넌스 구축에 대한 인적·물적 지원이 불명확한 경우는 의심해볼 만하다.

첫째·둘째에 관한 사례는 비영리 연구단체 '삶의 미래 연구소(Future of Life Institute)'가 2017년 1월 17일 발표한 '아실로마 원칙(Asilomar Principles)'이 손꼽힌다. 로봇 또는 (빅)데이터 윤리와는 별개로 '인공지능 윤리'를 독립적으로 중요한 쟁점으로 규정하는 목소리가 제기되기 시작한 시점은 대략적으로 2015년 전후로 보인다. 삶의 미래 연구소에서는 2015년 7월 28일 앞서 언급한 치명적 자율무기(LAWs)의 개발을 금지하는 국제적 노력을 촉구하는 '인공지능에 대한 공개서한(Open Letter on Artificial Intelligence)'을 발표하였고, 일론 머스크, 빌 게이츠, 스티븐 호킹과 같은 유명인사는 물론 인공지능 연구자 수천 여 명의 서명을 받았다.[43] 일론 머스크(Elon Musk)가 10억 달러를 투자한다고 발표하여 유명해진 인공지능의 사회적 문제에 대한 비영리 연구단체 '오픈AI(OpenAI)'도 같은 해 10월 출범하였다. 인공지능 분야를 선도하는 아마존, 구글, 페이스북, IBM, 마이크로소프트 5개사는 2016년 9월 '인공지능 파트너십(Partnership on AI)'이라는 비영리 연구단체를 설립하였고, 지금은 13개국의 100개 이상의 업체 및 기관이 참여하는 대규모 단체로 발돋움했다. 이러한 사회적 분위기에 힘입어, 삶의 미래 연구소는 1975년 생명의료윤리 분야를 탄생케 한 상징적 장소인 미국 캘리포니아주 아실로마에서 2017년 1월 6일부터 3일 간 'Beneficial AI 2017' 컨퍼런스를 개최하여 인공지능 윤리 이슈를 논의한 뒤, 같은 달 17일 '아실로마 원칙'이라는 명칭으로 정리하여 발표하였다.[44]

아실로마 원칙은 연구 분야 이슈(Research Issues), 윤리와 가치(Ethics and Values), 장기적 이슈(Longer-term Issues)라는 세 가지 영역에 걸친 23가지의 사항으로 구성되어 있다.[45] 인공지능 연구의 목표로서 인간에게 이로움을 주는 방향을 명시한 부분은 이전 논의의 연장선에 놓여 있는 내용이다. 연구개발 팀이 지나친 경쟁으로 안전기준을 회피할 수 있는 위험성을 지적한 부분은 컨퍼런스 참여자가 주로 공학자라는 점에서 새로운

43) Future of Life Institute, "Autonomous Weapons: An Open Letter from AI & Robotics Researchers", 2015. (http://futureoflife.org/open-letter-autonomous-weapons/)
44) 양희태, "인공지능의 위험성에 대한 우려로 제정된 아실로마 인공지능 원칙", 『과학기술정책』 제27권 제8호, 2017, 4면.
45) Future of Life Institute, "Asilomar AI Principles", 2017. (https://futureoflife.org/ai-principles/)

시사점을 제공한 원칙으로 볼 수 있다. 연구비 지원을 '윤리원칙'의 일환으로 포함했는데, 이는 다소 어색하기는 하지만 참여자 구성을 통해 이해할 수 있는 사항이다. 다른 한편, 주로 공학자의 시각에서 바라본 윤리원칙인 탓에 사회적 차원에서 일종의 사각지대가 있었던 것으로 볼 수 있는 내용도 있다. 예컨대, 투명성 원칙을 기능상 장애와 사법 영역에 국한하였는데, 반드시 그래야 할 필요는 없을 것이다. 여기에 해당하지 않더라도 예를 들어 이용자의 자율성을 신장하기 위해 인공지능의 투명성을 강화해야 할 필요가 있기 때문이다. 나아가, 윤리와 가치에 관해 키워드 위주로 나열할 뿐이고, 구체적으로 어떻게 그러한 윤리와 가치를 확립할지에 대한 규제 거버넌스 논의가 부족한 부분도 아쉬움이 있다. 그 이외에, 장기적 이슈에서 지금은 별로 논의되지 않는 초지능(super-intelligence)을 다룬 부분은 최근의 논의와는 결을 달리 한다. 인공지능 능력의 상한선을 함부로 가정하지 않아야 한다든가 재귀적 향상을 거듭하여 초지능으로 도약할 위험성을 통제하여야 한다는 내용(이른바 'kill switch' 또는 'big red button')은 공학자의 시선을 통해 흥미로운 시사점을 주는 또 다른 지점이다. 이처럼 아실로마 원칙은 구체성이 떨어지는 편이기는 하지만, 인공지능 윤리 분야에 지침이 되는 핵심적 토대를 마련해준 초창기의 중요한 논의라는 평가를 내릴 수 있다.

미국의 컴퓨터 전문가 단체인 ACM(Association for Computing Machinery)의 논의 역시 유사한 맥락에서 이해될 수 있다. 공학자들이 전문가 단체를 설립하여 윤리규범 논의를 주도하는 작업의 기원은 20세기 초로 거슬러가지만, 제2차 세계대전에서 원자폭탄이 투하된 일이 문제의식을 본격적으로 촉발한 계기가 되었다. 이후부터는 노버트 위너(Norbert Wiener)가 정립한 사이버네틱스(Cybernetics) 분야를 필두로 여러 공학단체들에서 기술전문가 윤리를 논의하기 시작하였다.[46] 전문가 윤리가 특히 강조되는 이유는 새로운 위험원에 대한 결정권을 부여한 데 대한 반대급부인 책무성과, 그들의 전문성이라는 역량에 대한 믿음 때문이다.[47] 1947년 설립된 ACM은 1966년 최초의 윤리강령(code of ethics)을 제정한 뒤 컴퓨터 윤리에 대한 논의를 지속적으로 진행해왔다. 대중, 고용주와 고객, 여타의 전문가 주체에 대한 윤리라는 3가지의 항목으로 구성된 최초의 윤리강령은 이후 발전하는 기술수준과 사회적 요구사항에 맞춰 꾸준히 수정되었다. 특히 2017년 1월에는 발전한 알고리즘 기술에 대한 문제의식으로부터 출발하여, 알고리즘의 해악과 편향에 대한 인식, 접근과 구제, 책임성, 설명, 데이터 출처, 감사가능성, 검증과 테스트라는 7가지 항목에 대한 윤리원칙을 발표했다.[48] ACM의 논의가 어느 정도 진정성이 있다고 볼 이유는 원칙을 발표한 이후의 행보 때문인데, 이듬해인 2018년 윤리강령을 개정하면서 직전 해 발표한 알고리즘에 대한 7가지 윤리원칙 논의를 발전적으로 수용하는 모습을 보였기 때문이다.[49]

셋째 요인의 대표적 사례는 경제협력개발기구(이하 'OECD')의 논의를 들 수 있다. OECD는 2016년 인공지능에 대한 포럼을 개최한 이후로 개별국가와 국제사회 차원의 논의를 조사하는 작업에 착수하게 된다. 이후 OECD의 디지털 경제정책 위원회(Committee on Digital Economy Policy, 이하 'CDEP')는 인공지능 기술 채택과 신뢰를 제고하기 위한 원칙을 제시하기 위하여 2018년 5월 각계각층의 전문가를 초빙한 'AIGO(AI Expert Group at OECD, 이하 'AIGO')'라는 그룹을 설립했다. AIGO에서는 2018년 9월부터 이듬해 2월까지 네 차례에 걸쳐 모임을 개최하여 얻은 정보를 통해 인공지능 가이드라인 권고안을 마련하였다.

46) 송성수, "공학단체의 윤리강령에 대한 비교분석: 미국과 한국의 사례를 중심으로", 『공학교육연구』 제11권 제3호, 2008, 79-81면; Terrell Bynum, "Computer and Information Ethics", Stanford Encyclopedia of Philosophy, 2015.

47) 김미리/윤상필/권헌영, 앞의 논문, 18-21면.

48) Association for Computing Machinery US Public Policy Council, "Statement on Algorithmic Transparency and Accountability", 2017.

49) Don Gotterbarn et al., "ACM Code of Ethics and Professional Conduct", Association for Computing Machinery, 2018.

2019년 5월 각료이사회(Ministerial Council Meeting)에 제출된 권고안에서는 신뢰할 수 있는 인공지능을 책임성 있게 달성하기 위해 필요한 기본원칙과, 신뢰할 수 있는 인공지능을 달성하기 위해 필요한 국가정책과 국제협력의 방향이라는 두 차원의 논의를 포괄하여 담았다. 전자의 기본원칙은 아실로마 원칙이나 그 이전 논의의 흐름을 이어가는 것으로, 포용적 성장과 지속가능한 발전 그리고 삶의 질, 인간 중심의 가치와 공정성, 투명성과 설명가능성, 견고성과 보안 그리고 안전, 책임성이라는 내용으로 구성되어 있다. 후자의 정책적 논의는 OECD의 고유한 특성을 반영한 것으로, 인공지능에 대한 개별국가 차원에서의 정책과 국제적 협력을 논의한다. 인공지능의 연구개발 투자와 디지털 생태계의 조성, 혁신을 위한 밑바탕이 되는 정책적 환경의 형성, 노동시장의 구조적 전환을 대비한 인간 능력 개발, 신뢰할 수 있는 인공지능을 위한 국제사회의 협력을 포괄하는 내용이 여기에 담겨있다.

한편, 이러한 제언의 후속작업으로, AIGO에서는 권고안이 채택된 날로부터 5년 동안 OECD(CDEP)에 진행상황을 보고하고 실무 가이드를 제공하도록 하여 권고안이 제대로 이행되는지 여부를 감독하는 절차를 마련하였다. CDEP에는 각국과 국제기구를 포괄하는 다양한 분야의 이해관계자가 학제간 대화를 통해 인공지능 정책에 대한 정보를 교환하는 포럼을 조직하고 감독하는 기구를 2019년 말까지 설립하도록 하였다.[50] OECD 권고안은 다양한 이해관계를 가진 회원국의 협의를 도출해내기 위한 목적으로 기존 논의와 크게 다르지 않은 추상적 원칙 위주의 내용을 담았다는 비판도 받지만, 구체적 후속작업과 규제 거버넌스 논의를 반영하였다는 측면에서 진일보하였다는 긍정적 평가도 공존한다. OECD 보고서가 발간된 다음 달 진행된 G20 회의에서 발표된 인공지능 원칙은 OECD 권고안을 거의 그대로 수용하였는데,[51] 기본원칙 중심의 논의일지라도 현실적 파급력을 가질 수 있음을 보여주는 단면이다. 특히 지금까지 인공지능 윤리 이슈에 대해 미온적 태도를 보인다고 비판받아온 중국이 일원으로 참여한 것에서부터 내용의 추상성과는 별개로 다양한 국제사회 구성원의 참여를 이끌어낸 것 자체가 중요한 성과라는 평가도 있을 수 있다.

3. 기본원칙과 심화된 이슈를 함께 다루는 논의

아실로마 원칙이 발표된 이후부터 일정한 기간 동안 적지 않은 논의가 기존의 내용에서 일부를 변형하거나 추가하는 선에서 이루어졌다. 이를 긍정적으로 해석하자면 초창기의 원칙이 충분한 논의의 기초를 제공할 정도의 큰 파급력을 낳았다고 볼 수 있고, 부정적으로 해석하자면 관련 이해관계자들이 인공지능 윤리 영역의 논의에 관해 그다지 진지하게 여기지 않았다고 볼 수도 있다. 물론 그와 별개로 유럽연합의 경우 초기 단계부터 양적·질적으로 풍부한 논의를 지속적으로 선도해왔고, 아래에서 별도로 살펴볼 전기전자기술자협회(Institute of Electrical and Electronics Engineers, 이하 'IEEE')의 경우 많은 논의를 거쳐 매우 상세하고 풍부한 내용의 논의를 담아내기도 했다. 하지만 상당수의 논의는 주요 윤리원칙을 열거하면서 그에 대한 설명을 부가하는 방식을 택하였다. 다만, 원칙에 대한 설명이나 제시된 예시의 구체성은 개별 논의에 따라 다르게 나타났다. 이러한 논의는 점차 일반론적인 원칙의 제시를 넘어, 좀 더 구체적이거나 심화된 사안들에 대한 다양하고 차별화된 논의로 변모하였다.

이 시기의 논의는 인공지능 서비스를 제공하는 몇몇 기업들 사이에서 활발하게 이루어진 편이다. 논의를 선도한 주요 사례로 구글(Google)을 꼽을 수 있다. 구글은 2018년 6월 선다 피차이(Sundar Pichai) CEO가 인공지능에 대한 윤리원칙을 직접

50) OECD, 앞의 글.
51) G20, "G20 AI Principles", 2019.

발표하였다. 구글이 제작한 인공지능은 사회적 혜택, 편향성의 생성과 강화 방지, 테스트를 통한 안전성의 확립, 책임성 강화, 프라이버시를 고려한 설계, 높은 과학적 탁월성 기준이라는 6가지 원칙을 제작·설계 단계는 물론이고 현실적으로 이용되는 상황에서까지 지켜질 수 있게 해야 한다는 것이다. 한편, 추구하지 말아야 할 4가지 사항도 언급되었는데 혜택을 중대하게 초과한 해악과 위험 초래, 사상자를 낳을 수 있는 무기의 개발, 국제규약을 위반하는 감시를 위한 정보의 수집과 활용, 국제법과 인권의 침해가 그것이다.[52] 이러한 원칙의 제시가 출발점이 되어, 아래에서 보듯 더욱 세분화된 영역에 대한 후속 논의가 진행되었다.

마이크로소프트의 경우, 앞에서 언급한 테이 사건이 발생한 이후부터 몇 가지 가이드라인을 마련하여 발표하였다. 2018년 1월 '인공지능으로 변화될 미래(The Future Computed)'라는 제목의 책자에서 공정성, 신뢰성과 안전, 프라이버시와 보안, 포용성, 투명성, 책임성이라는 6가지 윤리원칙을 언급한 사례를 가장 먼저 꼽을 수 있다.[53] 이후에는 구체적 상품에 대한 윤리원칙이 발표되었다. 2018년 11월 대화형 인공지능(챗봇) 개발자를 대상으로 한 10가지 사항의 가이드라인을 발표한 뒤,[54] 2018년 12월에는 안면인식 기술의 개발과 활용에 관련된 6가지의 원칙을 발표했다.[55] 여기에는 투명성이나 책임성처럼 앞서 언급한 6가지 일반적 원칙과 겹치는 내용도 있지만, 챗봇이나 안면인식과 같은 구체적 기술의 특성에 비추어서 일부 내용을 변형하거나 새로운 내용을 첨가하기도 하였다. 그밖에도 IBM, 소니(Sony), 인텔(Intel)과 같은 여러 글로벌 기업들이 2017년부터 2018년 사이에 인공지능 윤리 맥락의 원칙을 정리하여 발표하였다.

다른 한편, 일부 기업들은 인공지능 윤리원칙에 대한 구체화 및 이행을 위한 별도의 노력을 기울이는 모습을 보이기도 하였다. 예를 들어, 마이크로소프트에서는 윤리원칙을 마련하는 것 이외에, 이를 실현하기 위한 목적의 사내조직으로 엔지니어링과 연구 분야의 인공지능과 윤리 위원회(AI and Ethics in Engineering and Research, AETHER)를 출범하였다. 위원회는 모범사례의 확립과 안내지침의 제공과 같은 사전예방과 사후대처를 담당하는 조직이다. 구글의 경우에는 2018년 6월 최초로 윤리원칙을 발표하고, 그로부터 6개월과 1년이 지난 시점에서 새로운 노력을 반영하는 후속작업을 수행하였다. 예를 들어, 내부적 교육을 강화하고, 인공지능 모델들 사이의 편향성과 공정성의 측면을 시각적으로 비교해주는 'What-If' 툴과 같은 보조도구를 제작하였으며, 윤리원칙이 제대로 적용되는지 여부를 감독하는 절차를 마련하여 발표하였다.[56] 특히 구글에서 2019년 1월 발표한 백서는 설명성 표준, 공정성 평가, 안전성 고려, 인간과 인공지능의 협동, 책임 프레임워크라는 5가지 측면에서 구글이 활용하는 구체적 예시를 포함하여 설명한 문건으로, 논의를 더 구체화하게 해주었다.[57] 그밖에도 구글은 인공지능 개발, 오픈 데이터 공유, 공공선의 강화와 같은 부문에서 책임성 있는 인공지능의 실현에 기여하기 위해 백서를 제작하거나 기술적 도구를 만들어 배포해오고 있다. 아마존, 페이스북 같은 다른 글로벌 기업에서도 점차로 유사한 맥락에서 해석될 수 있는 움직임이 나타나고 있는 추세이다.[58]

52) Sundar Pichai, "AI at Google: our Principles", 2018. (https://www.blog.google/technology/ai/ai-principles/)
53) Brad Smith/Harry Shum, *The Future Computed*, Microsoft, 2018, pp. 50-74.
54) Microsoft, "Responsible Bots: 10 Guidelines for Developers of Conversational AI", 2018.(http://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf)
55) Microsoft, "Six Principles for Developing and Deploying Facial Recognition Technology", 2018. (https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2018/12/MSFT-Principles-on-Facial-Recognition.pdf)
56) Jeff Dean/Kent Walker, "Responsible AI: Putting our principles into action", 2019. (https://www.blog.google/technology/ai/responsible-ai-principles/)
57) Google, "Perspectives on Issues in AI Governance", 2019.
58) 아마존과 페이스북의 동향에 대하여는, 박도현, "미래를 향한 인공지능 정책: 우리는 AI를 신뢰할 수 있을까?", 『2019 국제학술

인공지능 윤리규범에 대한 기본원칙을 제시하면서 해당 주체가 중시하는 일부 쟁점에 대한 상세한 설명을 더하는 방식의 논의는 비단 개별 기업체 차원에만 국한되지는 않는다. 여타 주체의 경우에도 이러한 방식의 보고서를 배포한 사례를 종종 찾아볼 수 있다. 일례로 일본의 경우는 정부 차원에서 윤리규범에 대한 논의를 꾸준히 진행하고 보완하는 모습을 보였다. 일본 총무성은 2017년 7월 '인공지능 연구개발 가이드라인 초안'이라는 공식문건을 발표했다. 가이드라인은 인간중심 사회를 최우선시하는 목표로 삼고 개발자에 대한 과도한 부담이 없는 한도 내에서의 꾸준한 개정을 전제로, 국제적 이해관계자들과 모범세칙을 공유한다는 기치를 내걸었다. 인공지능이 가져다주는 편익을 증진하기 위해 연구개발을 중시하면서, 동시에 위험방지를 추구하는 9가지 원칙을 포함한 것이 특징적이다.[59] 여기에는 종래 지속적으로 논의된 투명성, 통제가능성, 안전, 보안, 프라이버시, 윤리, 책임성과 같은 내용이 포함된다. 가이드라인의 또 다른 특징적 내용으로 인공지능 시스템의 상호연결성과 상호운용성을 강조하는 연계의 원칙, 이용자가 인공지능의 도움을 받아 적절하게 의사결정을 할 수 있게 하는 이용자 지원의 원칙을 꼽을 수 있다. 한편, 가이드라인이 논의의 대상이 되는 인공지능을 현재 사용되고 있는 좁은 인공지능(Narrow AI)으로 상정하고, 일반 인공지능(Artificial General Intelligence, AGI)은 논의의 대상에서 제외한 부분도 특기할만하다.[60] 아실로마 원칙 이후의 논의에서는 현실적으로 발생할 수 있는 문제에 초점을 맞추는 경향이 일반적인데, 그러한 맥락에서 이해할 수 있다.

2017년 가이드라인은 연구개발을 대상으로 한 것이었기 때문에 이용자 맥락의 논의는 부족한 편이었다. 일본 총무성은 이듬해인 2018년 7월 '인공지능 이용원칙 초안'을 발표하여 공백을 해소하고자 하였다. 10개 항목으로 구성된 이용원칙은 연구개발 원칙과 중복되는 내용을 상당부분 포함하고 있지만, 이용자에 의한 적절한 이용이라는 원칙을 언급하여 개발자의 의무만을 강조해온 그동안의 논의와 차별화되는 내용을 포함하였다. 테이 사례처럼 이용자가 악의를 품을 경우에는 제작자가 악의를 가지는 경우 못지않은 문제를 야기할 수 있기 때문에 이용자에 대해 언급하는 것은 유용한 의미를 가질 수 있다. 그밖에 데이터의 품질과 공정성 원칙은 연구개발 가이드라인에는 없던 새로운 내용으로, 이용자에 초점을 맞춘 이용원칙의 특징을 반영한 내용이다.[61] 이와 같은 논의에 기초하여, 일본 정부는 2019년 3월 인공지능 전략을 내놓으면서, 인간중심의 원칙, 교육과 리터러시 원칙, 프라이버시 확보 원칙, 보안 확보 원칙, 공정한 경쟁 확보 원칙, 공정성과 설명책임 및 투명성 원칙, 혁신의 원칙이라는 7가지의 '인공지능 사회 원칙'을 정리하여 발표하였다.[62]

미국 정부의 최근 논의도 유사한 면모를 보인다. 오바마 행정부 당시 인공지능 규제 논의에서 상당히 앞서가던 미국 정부는 트럼프 행정부가 출범한 직후에는 인공지능 분야 전반에 대해 상대적으로 무관심한 태도를 보였지만 2018년 무렵부터 이와 같은 입장에서 선회하였고, 2019년 2월에는 트럼프 대통령이 직접 행정명령을 발표하기도 하였다. 행정명령은 대체로 연구개발에 초점을 맞추었지만, 5가지 원칙 중 네 번째 원칙에 시민의 자유와 프라이버시를 보호하여야 한다는 내용을 포함하였고, 이후 인공지능 규제 원칙을 발표하도록 하는 내용을 명시했다.[63] 관리예산실(Office of Management

대회 보고서』, 서울대학교 법과경제연구센터, 2019, 15, 23-24면 참조. (http://ai.re.kr)

59) 황현주, "AI 연구개발과 활용 촉진을 위한 'AI 개발 가이드라인(안)'", 『NIA Special Report』 2017-9, 2017, 1면.

60) The Conference toward AI Network Society, "Draft AI R&D Guidelines for International Discussions", 2017. (www.soumu.go.jp/main_content/000507517.pdf)

61) The Conference toward AI Network Society, "Draft AI Utilization Principles", 2018. (www.soumu.go.jp/main_content/000581310.pdf)

62) 유재홍, "일본의 인공지능 전략 동향: AI 전략", 『월간 SW 중심사회』 통권 제60호, 2019, 22-23면 참조.

63) Donald J. Trump, 앞의 글, Section 1, 6 참조.

and Budget, OMB)을 중심으로, 백악관은 2020년 1월 공적 신뢰, 대중의 참여, 과학적 무결성과 양질의 정보, 위험평가와 관리, 혜택과 비용, 유연성, 공정성과 차별금지, 공개와 투명성, 안전과 보안, 기관 간 공조라는 10가지의 항목에 걸친 인공지능 규제 원칙을 발표하였다.[64] 한편, 백악관 과학기술정책국이 그 다음 달 발표한 인공지능 이니셔티브에서는 위 규제 원칙 그리고 앞서 언급한 2019년의 OECD 및 G20의 윤리원칙을 준수한다는 내용을 명시하였다.[65]

이러한 접근방식과 성격을 달리하는 것으로, 인공지능 윤리 전반이 아닌, 특정한 영역에서의 윤리적 쟁점에 집중하여 심화된 논의를 전개한 사례도 종종 찾아볼 수 있다. 독일 정부가 2017년 6월 자율주행 자동차 분야에 대해 발간한 보고서를 대표적 사례로 꼽을 수 있다. 보고서는 자율주행 자동차의 운행과 관련된 주요한 쟁점을 20가지로 정리하여 제시하였다. 원칙이 아닌 규칙(rule)의 형태로 구체적 상황과 행동지침을 제시한 데서 여타의 논의와 차별화된다. 보고서는 이용자의 안전과 혜택의 창출이라는 두 가지 목표를 병기하면서, 개인의 보호가 여타 공리주의적 고려에 앞선다고 하여 인간의 생명에 대한 우월성을 강조하였다. 현실적으로 가능한 선에서 위험과 딜레마 상황을 최소화하면서 자율주행 자동차의 도입 자체는 긍정하되, 예상되는 딜레마 상황별로 우선시해야 할 가치를 지침화하는 형태를 취하였다.[66]

노동 분야에서 유니 글로벌 유니온(UNI Global Union)이 2017년 12월 발표한 10가지 윤리원칙은 내용 자체는 대체로 기존의 논의와 대동소이하지만 인공지능 시대의 노동자의 권리라는 특수한 쟁점을 다룬 자료이다. 한 가지 특이한 점은 노동자의 권리를 논의하는 10가지 윤리원칙에 로봇에 대한 책임귀속을 금지하라는 내용이 포함되어 있는 부분이다. 아마도 노동자가 로봇에 의해 상해를 입거나 피해를 당한 경우 고용주가 로봇에 대하여 책임을 전가하는 상황을 염두에 둔 내용으로 보인다. 또 다른 특수한 윤리원칙 사례로, '사법 시스템과 사법 환경에서 인공지능의 활용에 대한 유럽 윤리헌장(European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment)'을 꼽을 수 있다. 유럽평의회 산하 사법 효율성을 위한 유럽위원회(European Commission for the Efficiency of Justice)에서 2018년 12월 채택한 이 헌장은, 기본권 존중, 차별금지, 품질과 보안, 투명성과 불편부당성 및 공정성, 이용자의 통제라는 5가지의 원칙으로 구성되어 있다.[67] 원칙의 내용 자체는 여타 논의와 크게 다르지 않지만, 사법이라는 특수한 활용맥락을 고려해 부록(appendix)에 그에 관한 상세한 내용을 포함한 것이 특징적이다.[68]

다른 한편, 인공지능의 소프트웨어에 해당하는 알고리즘과 하드웨어에 해당하는 로봇의 윤리에 주로 초점을 맞춘 논의 이외에도, 인공신경망 방식의 머신러닝을 실질적으로 가능하게 하는 '(빅)데이터 윤리'에 대한 논의도 꾸준히 진행되어 왔다. 빅데이터 윤리에 대한 관심이 본격적으로 나타난 초기 논의로 오바마 행정부가 2012년 2월 발표한 소비자 데이터 프라이버시 보고서를 들 수 있다. 2000년대 들어 새로이 등장한 빅데이터(big data)라는 개념이 프라이버시에 대한 종래의 관념을 상당 부분 뒤바꾸고 있다는 문제의식에서 출발하여 관련 논의를 정리한 결과물이다. 데이터의 총량과 지속

64) Russell T. Vought, "Guidance for Regulation of Artificial Intelligence Applications(Draft)", 2020, pp. 3-6.

65) The White House Office of Science and Technology Policy, "American Artificial Intelligence Initiative: Year One Annual Report", 2020, pp. 13-15, 21-22.

66) Udo Di Fabio et al., "Ethics Commission Automated and Connected Driving", Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany, 2017, pp. 9-12.

67) European Commission for the Efficiency of Justice, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment", 2018, pp. 7-12.

68) 윤리헌장을 상세히 분석한 선행연구로, 한애라, ""사법시스템과 사법환경에서의 인공지능 이용에 관한 유럽 윤리헌장"의 검토 – 민사사법절차에서의 인공지능 도입 논의와 관련하여 –", 『저스티스』 통권 제172호, 2019, 46면 이하 참조.

시간이 급격히 늘어나면서 상관관계로 인과관계 못지않은 통찰력을 얻을 수 있고, 데이터 활용에 대한 동의 시 미래의 파급효과를 정확히 예측하기가 어렵게 된 것이 그 배경으로 꼽힌다.[69] 보고서는 7가지의 하부 권리로 이루어진 '소비자 프라이버시 권리장전(Consumer Privacy Bill of Right)'을 이에 대한 대안으로 제시하였다. 권리장전은 개인의 통제, 투명성, 보안, 접근과 정확성, 책임성과 같은 지금까지도 널리 원용되고 있는 윤리원칙뿐 아니라, 최소수집이라는 프라이버시 분야의 전통적 원칙까지 포괄하여 언급하였다.[70] 오바마 정부는 2014년부터 2016년까지 매년 꾸준히 빅데이터 정책과 윤리적 이슈에 대한 보고서를 발표하여 이와 같은 흐름을 이어왔다.[71]

최근 들어서는 빅데이터 윤리 이슈를 별도로 고찰하는 대신 인공지능 윤리의 일환으로 보는 경우가 많지만, 두 가지 이슈를 분리하여 수행한 논의도 여전히 존재한다. UNGP(UN Global Pulse)와 국제프라이버시전문가협회(International Association of Privacy Professionals, 이하 'IAPP')가 2017년 5월에 개최한 데이터 프라이버시 윤리 포럼에서의 논의를 바탕으로 2018년 10월에 발간한 빅데이터와 인공지능 분야의 윤리적 프라이버시 프레임워크 구축에 관한 보고서가 이러한 맥락의 주요 사례이다.[72] 보고서에서는 GDPR에서 규정하는 것처럼 프라이버시 문제를 개발단계에서부터 선제적으로 고려(privacy by design)하도록 강조한다. 나아가 빅데이터나 인공지능의 활용 과정에서의 오용(misuse)뿐만이 아니라 사용하지 않음으로 인해(missed use) 생기는 영향도 균형 있게 판단하여야 한다는 내용을 덧붙였다. 보고서에서는 빅데이터 윤리를 확립하기 위한 구체적 방안으로 내부적·외부적 프레임워크를 융합하는 방안을 제시하면서, 구체적 선례를 함께 언급했다. UNGP와 IAPP의 보고서는 프라이버시 분야에 국한된 것이기는 하지만, 민간 전문가협회와 국제기구의 협력을 통해 이루어진 구체적 성과라는 점에서 주목할 만한 사례로 볼 수 있다.

4. 윤리규범과 규제 거버넌스 정립방안에 대한 구체적 논의

끝으로, 인공지능의 윤리적 쟁점에 관한 기본원칙과 일부의 쟁점을 심화시키는 차원을 넘어선, 규제 거버넌스의 정립방안에 대한 어느 정도의 구체적 논의까지 이루어진 경우도 찾아볼 수 있다. 이와 같은 논의는 각국 정부가 발표한 자료에서 가장 빈번하게 발견된다. 공적 주체의 특성상 규제에 대한 논의를 진행해야 할 당위를 갖고 있고, 물적·인적 자원이 풍부하며, 이해의 충돌에서 어느 정도 자유롭다는 측면이 구체화된 논의를 가능하게 한 배경인 것으로 보인다. 인공지능 거버넌스에 대한 정부 차원의 대표적 논의로는 초창기의 미국 정부의 보고서가 손꼽힌다. 오바마 정부 국가기술과학위원회(National Science and Technology Council)의 머신러닝 및 인공지능 소위원회는 2016년 5월부터 5차례의 워크숍을 개최한 뒤, 같은 해 10월 안전과 위험의 규제, 공정성 문제, 자율무기 이슈, 인간적 가치를 비롯한 윤리적·규범적 차원의 논의를 다룬 보고서를 발간하였다. 보고서는 기본원칙을 제시하고 그에 관한 설명을 덧붙인 수준을 넘어 부문 별로 구체적 사례까지 포함하는 23가지에 달하는 권고안을 제시했고, 관리감독

69) Catherine Tucker, "Privacy, Algorithms, and Artificial Intelligence", in Ajay Agrawal/Joshua Gans/Avi Goldfarb (eds.), *The Economics of Artificial Intelligence*, University of Chicago Press, 2019, pp. 423 이하 참조.

70) The White House, "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy", 2012, p. 9 이하 참조.

71) U.S. Executive Office of the President, "BIG Data: Seizing Opportunities, Preserving Values", 2014; U.S. Executive Office of the President, "BIG Data: Seizing Opportunities, Preserving Values Interim Progress Report", 2015; U.S. Executive Office of the President, "Big Data: A Report on Algorithmic Systems", Opportunity, and Civil Rights, 2016.

72) United Nations Global Pulse/International Association of Privacy Professionals, "Building Ethics into Privacy Frameworks for Big Data and AI", 2018.

체계와 후속연구에 대한 제언과 국제사회와의 공조라는 구체적 거버넌스 체계도 언급하였다. 윤리적 인공지능을 실현하기 위한 규제정책의 일환으로 기술공동체와의 협업, 기초적·장기적 연구, 인공지능 문해력(literacy) 향상을 위한 대중적 교육을 강조하는 부분은 정부의 논의가 가진 고유한 특색을 보인 대목이다.[73]

이듬해인 2017년 말 프랑스 정보자유 국가위원회(Commission Nationale de I'informatique et des Libertés, 이하 'CNIL')는 인공지능의 윤리이슈를 직접적 논의대상으로 삼은 보고서를 발표했다. CNIL은 인공지능 윤리를 다루어야 하는 이유에 대하여 법적 규준을 마련하기 위한 사전 단계의 논의라는 점을 천명하고, 규범의 적용대상인 당사자가 논의에 주체적으로 참여한다는 측면을 강조하였다. 그런 점에서 거버넌스와 책임성 개념의 새로운 단면이 현실에 반영된 사례라고 볼 수 있을 것이다. 보고서에서 언급한 윤리이슈의 목록은 자율성과 책임, 편향과 차별, 프로파일링, 프라이버시, 데이터의 품질과 같이 기존 내용과 크게 다르지 않지만, 인간의 정체성에 대한 위협과 같은 미래지향적 내용도 포함되어있다. 인공지능이 인간과 정서적 교감을 공유하는 과정에서 인간만의 고유한 특질을 뒤바꿀지도 모른다는 문제의식을 반영한 것이다. 이후 보고서는 현행 법제의 내용과 한계를 검토하고 규제가 적용되지 말아야 할 예외적 분야를 살펴본 뒤, 공정성과 지속적 감시라는 개발과정의 두 가지 근본원칙과 여기에 기초하는 이해가능성, 투명성, 인간의 개입이라는 세 가지 공학적 원칙을 도출하였다. 여기까지 논의는 기존의 논의와 비슷하지만, 일련의 원칙을 실현하기 위한 6가지의 구체적인 정책제언을 덧붙여 후속절차의 이행에 도움을 제공하고자 하였다.[74] CNIL 보고서는 프랑스 정부가 이듬해 3월 인공지능 전략[75]을 발간하는 데 중요한 밑바탕이 된 것으로 보인다.

유럽연합의 일원이지만 브렉시트(Brexit)를 선언한 영국에서도 유럽연합 차원과는 별도의 논의가 진행되었다. 영국 상원의 인공지능 특별위원회에서는 2018년 4월 영국이 인공지능 관련 산업의 발전을 선도할 수 있는 토대를 만들고 이를 위해 검토·준비할 사항을 영국 정부에 권고하는 내용을 담은 보고서를 발간했다.[76] 보고서는 기본적으로 인공지능의 설계와 개발에 초점을 맞추고 있지만, 인공지능 산업이 유발할 것으로 예상되는 사회적 문제와 규제방안에 대한 검토도 포함하였다. 영국 정부는 인공지능이 준수해야 할 5가지 대원칙으로 인류의 공동선과 이익, 이해가능성과 공정성, 프라이버시, 교육, 자율성을 제시하면서, 여기에서 그치지 않고 대기업에 의한 데이터의 독점이나 편향성, 비식별화 조치와 같은 데이터와 관련된 제반 이슈에 대한 논의 또한 병행하였다. 보고서는 법적 규제와 관련하여 법사위원회를 통해 현행법이 인공지능의 법적 책임을 묻는 데 적절한지 여부를 검토하도록 하면서도 인공지능에만 특화된, 윤리적 준칙을 넘어서는 법적 규제를 도입하는 것은 현 단계에서는 적절치 않다고 바라보았다.

한편, 싱가포르는 최근 인공지능 거버넌스 분야에서 많은 노력을 들인 중요한 사례다. 싱가포르 개인정보보호위원회(Personal Data Protection Commission, PDPC)는 2019년 1월 다보스의 세계경제포럼에서 설명가능성, 투명성, 공정성과 인간중심의 인공지능이라는 대원칙을 바탕으로 하여, 조직 내부의 전반적 구조, 인공지능 모델의 결정, 작업과정의 관리, 이용자 관리라는 네 가지 영역에 걸친 인공지능 거버넌스 프레임워크를 발표하고, 부록(annex)에서 감사(audit)를 위한 방안을 제시하였다.[77] 싱가포르 정부는 이후 1년 동안 축적

73) U.S. Executive Office of the President/National Science and Technology Council Committee on Technology, 앞의 글, p. 13 이하 참조.

74) Victor Demiaux/Yacine Si Abdallah, "How can Humans keep the Upper Hand: The Ethical Matters Raised by Algorithms and Artificial Intelligence", 2017, p. 24 이하.

75) Cédric Villani et al., "For A Meaningful Artificial Intelligence: Towards A French and European Strategy", 2018.

76) UK House of Lords Select Committee on Artificial Intelligence, "AI in the UK: ready, willing and able?", Report of Session 2017-19, 2018.

된 경험과 유럽연합, OECD를 비롯한 국제사회에서 발표된 새로운 논의를 덧붙여, 이듬해 1월 새로운 인공지능 거버넌스 프레임워크 버전을 내놓았다. 새로운 보고서는 그동안 누적된 선례나 연구를 보완하는 한편, 프레임워크의 두 번째 영역을 인공지능의 모델을 결정하는 행위를 넘어 인간의 전반적 개입수준에 대한 것으로 변경하고, 네 번째 영역을 이용자를 넘어서는 이해관계자 전반의 상호작용과 대화로 확장하였다.[78]

이러한 사례에서 볼 수 있듯, 인공지능 윤리·거버넌스 분야의 구체적 논의를 진행하는 주체는 각국의 정부인 경우가 많지만 여기에 국한되지는 않는다. 이에 대한 두 가지의 대표적 예외사례로 IEEE와 유럽연합을 제시할 수 있다. IEEE는 세계적 기술전문가 집단이면서 국제 표준화기구이기도 하다. 인공지능 및 자율시스템의 윤리적 고려사항에 대한 IEEE 글로벌 이니셔티브에서는 2016년 12월과 2017년 12월 두 차례에 걸쳐 윤리적 인공지능에 대한 보고서를 발간한 뒤,[79] 이에 대한 의견수렴을 거쳐 2019년 3월 최종 버전의 보고서를 발표하였다. IEEE 보고서는 기술전문가가 인공지능을 설계할 때 발생할 만한 윤리적 문제에 집중하고 있고, 약 300페이지에 달하는 방대한 분량을 자랑한다. 보고서에서는 윤리적 인공지능 설계의 3대 축으로 인간의 보편적 가치, 인간의 정치적 자기결정권과 데이터에 대한 주체성, 기술적 신뢰성을 제시하고, 그로부터 8가지의 일반원칙을 제시한다. 원칙적 차원에서는 대체로 기존에 논의된 내용과 큰 차이는 없지만, 인공지능 시스템이 목적에 맞게 제대로 기능해야 한다는 효과성(effectiveness), 제작자가 안전성과 효과성을 갖출 수 있을 만큼의 기술적 능력(competence)을 구비하여야 한다는 부분은 기술전문가 조직인 IEEE의 고유한 특성을 보여주는 대목이다. 3대 축을 8원칙으로, 8원칙을 설계의 과정으로 반영해나가야 한다는 실무지향적 태도와, 윤리적 인공지능을 제작하는 AMA 분야에 대한 논의에 많은 분량을 할애한 것도 같은 맥락에서 이해될 수 있다.[80] 이후 보고서는 고전윤리에 대한 검토는 물론 윤리적 개념을 어떻게 기술적 언어로 풀어낼지, 기존의 논의가 서양의 시각에 지나치게 경도된 것은 아닌지와 같은 광범위한 쟁점을 심도 있게 논의하였다. 이러한 논의에 기초하여, 보고서는 삶의 질(well-being)이라는 질적 요인이 인공지능에 반영될 수 있게 하는 지표(metric)의 개발, 감성 컴퓨팅(affective computing)의 개발이 인류의 정서에 미치는 영향력, 인공지능이 개인의 선택을 유도함으로써 자율성을 저해하는 넛징(nudging)의 문제와 같은 심화된 논의를 진행했다.[81] 이러한 점을 통해 알 수 있는 것은 표면적으로 드러난 윤리원칙은 추상적이고 당연한 내용에 불과하더라도, 이로부터 얼마든지 광범위하고 상당히 심층적인 논의가 파생될 수 있다는 사실이다. IEEE는 이러한 일련의 윤리적 문제에 대한 대안으로, 학제 간 교육과 연구, 조직 내의 관행 형성, 책임과 평가라는 세 가지 차원의 거버넌스를 제시했다. 일견 원론적 논의로 그칠 수 있는 내용이지만 IEEE는 여기서 한 걸음 더 나아가 구체적 예시를 포함하는 정책적·법적 제언까지 덧붙였다. 예를 들어, 인공지능에 대한 완전한 법인격의 부여는 시기상조라고 바라보면서도, 다른 한편 (현상적) 자율성을 나타내는 인공지능의 규율공백에 유의하여야 하고 미래지향적 인공지능이 출현할 경우에 대

77) Personal Data Protection Commission Singapore, “A Proposed Model AI Governance Framework”, 2019.

78) Infocomm Media Development Authority/Personal Data Protection Commission Singapore, “Model Artificial Intelligence Governance Framework Second Edition”, 2020.

79) Institute of Electrical and Electronics Engineers, “Ethically Aligned Design Version 1: A Vision for Prioritizing Human Well-being with Artificial Intelligence and Autonomous Systems”, 2016; Institute of Electrical and Electronics Engineers, “Ethically Aligned Design Version 2: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems”, 2017.

80) Institute of Electrical and Electronics Engineers, “Ethically Aligned Design First Edition: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems”, 2019, pp. 10-14.

81) Institute of Electrical and Electronics Engineers, 앞의 글, pp. 36-123.

비하여야 할 필요성도 지적했다. IEEE는 'P7000 TM 시리즈'라고 불리는 신기술에 관한 윤리적 고려를 담은 13가지 표준(standard)을 제작해왔는데, 이러한 국제표준은 규제 거버넌스 논의에 많은 시사점을 준다.[82]

개별 정부 이외의 주체가 진행한 또 다른 중요한 논의의 사례로는 유럽연합을 제시할 수 있다. 유럽연합 집행위원회(European Commission)는 2018년 4월 25일 인공지능 기술에 관한 성장과 윤리를 아우르는 새로운 비전을 발표한 뒤, 같은 해 6월 인공지능 고위급 전문가 그룹을 출범하였다. 이들의 주요한 목적은 윤리 가이드라인과 정책 및 투자 권고안을 제작하는 작업이었다. 본고의 주제인 윤리 가이드라인의 경우, 2018년 12월 18일 초판이 발표되고 몇 가지의 수정을 거쳐 2019년 4월 8일 최종본이 발표되었다.[83] 가이드라인은 유럽연합이 추구하는 신뢰할 수 있는(trustworthy) 인공지능을 형성하는 3가지 축으로 합법성(lawful), 윤리성(ethical), 기술적·사회적 견고성(robustness)을 제시하면서, 법적 측면은 제외하고 나머지 두 가지 측면에 초점을 맞추겠다고 밝혔다. 보고서에서는 3대 축의 핵심요소로 기본권 존중을 들면서, 자율성 존중, 해악금지, 공정성, 설명가능성이라는 4가지 하위 윤리원칙을 제시했다. 다음으로 이와 같은 원칙을 실현하기 위한 7가지 요구사항인 인간의 주체성과 감시, 기술적 견고성과 안전성, 프라이버시와 데이터 거버넌스, 투명성, 다양성과 차별금지 및 공정성, 사회·환경적 삶의 질, 책임성을 이끌어냈다. 신뢰할 수 있는 인공지능을 실현하기 위한 윤리원칙과 요구사항을 달성하기 위하여 기술적·비기술적 방법론 모두가 중요하고 함께 고려될 필요가 있다고 강조한 점이 특징적이다. 가이드라인은 또한 인공지능의 개발·배치·이용 과정에서 신뢰할 수 있는 인공지능에 해당한다고 볼 수 있는지를 평가하기 위한 구체적 평가항목을 제작하여 예시로 제시하였다. 가이드라인은 이러한 사항의 후속절차로, 모든 이해관계자들이 평가목록에 대한 파일럿 테스트를 진행하고 유럽연합 집행위원회가 여기서 얻은 피드백을 토대로 2020년 초까지 평가기준에 대한 개정판을 마련하도록 명시하였다.

Ⅳ. 해외의 논의가 국내에 주는 시사점

1. 해외 논의의 특징과 경향성

지금까지 살펴본 해외 논의의 대략적 특징과 경향성은 다음과 같이 요약될 수 있다. 첫째, 논의의 주요 방향이나 원칙은 점차로 수렴되는 모습을 보인다. 먼저 대부분의 논의가 인간 중심의 관점이나 인간과 인공지능의 공존을 대원칙으로 설정한 뒤, 하위 원칙이나 구체적 실현방식을 제시해나가는 모습을 보여준다. 주요 윤리원칙 역시 대체적으로는 수렴하는 것으로 보인다. 헌법에서 중요하게 다루는 존엄성, 공정성, 기본권과 같은 개념에 더하여, 프라이버시, 보안, 투명성, 견고성과 같은 인공지능 기술이 갖는 특성에서 비롯된 개념이 주류적 원칙으로 부각되고 있다. 또한 최근의 인공신경망 기반 인공지능이 가져다주는 불투명성의 문제를 극복하기 위한 투명성, 설명가능성, 책임성과 관련된 내용은 대다수 윤리원칙에 빠지지 않고 등장하는 사항이다.

둘째, 어느 정도는 논의를 이끄는 주체가 가진 특수성이 반영되고 있는 모습도 나타난다. 대체적으로는 사적 주체가 내놓은 규범의 형태가 일반론적이거나 원론적 내용이 담긴 경우가 많은 반면, 공적 주체를 통해 도출된 규범은 비교적 다루는 분야가 넓은 편이고 규범의 구체성도 높은 경우가 많다. 다만, 이와 같은 해석을 무분별하게 일반화하기는 어렵다. 공적 주체일 경우에도 논의를 막 시작한 초기 단계에서는 원론적 내용이 포함될 수 있고,

82) Institute of Electrical and Electronics Engineers, 앞의 글, pp. 256-257, 285-286.

83) European Commission High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI", 2019.

다른 한편 몇몇 기업체의 경우는 상당한 구체성을 가지고 지속가능한 논의를 내놓는 모습을 보여주기 때문이다. 이와 별개로, 다양한 이해관계를 가진 주체가 공동으로 마련한 규범은 추상적 내용 위주로 담겨있는 편이다. 공적 주체의 대표 격인 국제기구 차원의 논의가 특히 그렇다. 예를 들어, OECD에는 다양한 이해관계를 가진 주체의 협상을 통한 합의가 반영되기에 구체적 내용이 많이 담겨있지 않다는 평가를 받기도 한다. 그렇다고 해서 논의의 양과 규범력이 반드시 비례하는 것도 아니다. OECD의 규범은 후속조치와 정기적 모니터링에 관한 내용을 담고 있고, G20 선언으로 논의가 확장되는 모습을 보여주기도 했다. IEEE는 다수의 주체가 참여한 단체이기는 하지만, 기술전문가로서 동질적 정체성을 가진 덕분에 상당히 깊이 있는 논의를 진행할 수 있기도 하였다.

셋째, 개별주체 사이의 지역적·문화적 차이나 규제를 바라보는 관점의 차이도 어느 정도 볼 수 있다. 유럽과 미국을 비교하자면, 유럽연합은 윤리 규범의 내용이 훨씬 구체화된 상태이다. 반면, 미국의 경우 대체로 개별 기업이나 학계를 통해 활발한 논의가 이루어지는 것과 달리, 정부 차원에서는 논의가 계속 진행되기는 하더라도 윤리 이슈 맥락에서 정부의 역할은 아직까지 제한적인 것으로 보인다. 이러한 차이가 나타나는 한 가지 이유로, 유럽은 GDPR과 같은 일반적 법규범을 통한 통일적 규율이 자연스럽게 받아들여지는 반면, 미국은 개별 영역별 접근과 자율규제의 역할을 상대적으로 강조하는 면이 있다는 점을 꼽을 수 있다. 다만, 근래에는 산업의 발전과 윤리적 규율을 별도로 논의하는 대신, 합일된 관점 하에 두 가지 모두를 고려하는 흐름이 나타나는 모습도 보여준다. 이러한 경향을 국내 인공지능 규범 논의에 참조할 때는 어느 한 단면만을 강조하는 대신 우리나라의 현실적 여건과 국제적 논의의 전개과정을 종합적으로 고려하여야 할 것이다.

2. 우리나라 논의의 현황과 과제

그렇다면 우리나라의 논의는 어떤 모습으로 변화해왔고, 어떤 상황에 있을까? 앞서 언급한 것처럼, 2007년에 발표된 로봇윤리헌장 초안을 가장 먼저 떠올려볼 수 있다. 이는 '지능형 로봇 개발 및 보급 촉진법' 제18조에 법적 근거를 두었고, 2016년에는 개선안까지 마련되었지만 시행단계에 이르지는 못하였다. 그러다가 알파고가 큰 사회적 파장을 낳은 뒤부터 인공지능 기술에 대한 사회적인 관심 및 관련된 규범 마련에 대한 논의가 좀 더 본격화되었다. 2017년부터는 '지능정보사회 기본법안'이나 '로봇기본법안'과 같은 관련 분야의 법률안이 마련되기도 하였다. 나아가 정보문화포럼과 한국정보화진흥원은 2018년 초 공공성(Publicness), 책무성(Accountability), 통제성(Controllability), 투명성(Transparency)으로 이루어진 'PACT 원칙'이라고 일컬어지는 윤리원칙을 마련하였고, 이 원칙을 개발자, 공급자, 이용자에 대해 적용한 지능정보사회 윤리 가이드라인 및 윤리헌장을 발표하였다.[84] 윤리헌장은 6가지의 짤막한 문장으로 구성된 다소 원론적 내용을 담고 있지만, 가이드라인은 원칙에 대한 세부지침을 포함한 20여 페이지의 분량으로 되어 있다. 2020년 5월 20일에 통과된 '국가정보화 기본법' 전부개정 법률안 제62조는, 국가기관과 지방자치단체가 인공지능 기술을 포함한 지능정보기술을 개발·활용·제공·이용함에 있어 인간의 존엄과 가치, 공공성·책무성·통제성·투명성 등의 윤리원칙을 담은 지능정보사회윤리를 확립하여야 한다고 규정하고 있어서 향후 더욱 심도 있는 논의가 이루어질 것이 예상된다. 이와 별도로 방송통신위원회는 2018년 2월부터 2019년 8월까지 여러 차례 포럼과 간담회를 거쳐 논의를 수렴한 뒤, 2019년 11월 '이용자 중심의 지능정보사회를 위한 원칙'을 발표하였다. 방송통신위원회의 원칙

84) 정보문화포럼/한국정보화진흥원, "지능정보사회 윤리 가이드라인", 2018.

에는 사람 중심의 서비스 제공, 투명성과 설명가능성, 책임성, 안전성, 차별금지, 참여, 프라이버시와 데이터 거버넌스라는 7가지 기본원칙에 더하여, 이용자 보호를 위한 공동의 노력이라는 별도의 항목이 덧붙여졌다.[85]

민간기업의 논의로는 카카오에서 2018년 1월 31일 발표한 '카카오 알고리즘 윤리 헌장'을 들 수 있다. 카카오는 2019년 8월에 기본원칙, 차별금지, 학습 데이터, 알고리즘 독립성, 설명가능성으로 구성된 기존 5원칙에 포용성을 덧붙인 새로운 윤리 헌장을 발표하였다.[86] 꾸준히 윤리 이슈에 관심을 기울이고 새로운 내용을 업데이트한다는 점에서 긍정적인 평가를 할 수 있지만, 추상적인 원칙의 제시 차원에 그치고 있고 이에 관한 구체적인 이행이나 거버넌스 차원의 내용은 담겨있지 않다는 한계가 있다. 삼성전자는 2019년 공정성, 투명성, 책임성으로 구성된 '인공지능 윤리 핵심원칙'을 발표하였는데,[87] 마찬가지로 원칙의 제시 수준에 머물러 있다. 이외에도 몇몇 기업이 관련 논의를 진행하고 있으나, 아직까지 명시적인 규범이나 거버넌스에 관한 내용을 제시하는 단계에 이르지는 못한 것으로 보인다.

국내의 논의는 인간 중심의 대원칙과 국제사회에서 주로 논의되는 윤리원칙을 상당 부분 수용해 나가고 있다는 점에서, 어느 정도 해외의 논의와 일맥상통하는 모습으로 나아가고 있다고 평가할 만하다. 향후 논의가 계속된다면, 이제부터는 규범의 구체화나 거버넌스 구조의 구비, 이행을 위한 유인 제공과 같은 더 현실적인 사항에 대한 작업이 행해질 것으로 예상된다. 다만, 향후 진행될 논의는 국내의 실정에 부합하는 동시에 국제사회의 흐름에 발맞추어 진행되어야 할 것이다. IEEE의 사례처럼 전문가가 만든 국제표준은 실무 개발자에게는 사실상의 행동지침으로 작용할 수 있다. 다른 한편, OECD나 유럽연합 등을 통한 논의는, 참가 당사국을 통해 후속논의와 후속조치에 대한 장치를 마련하고 있어서, 이를 통해 현실적인 집행력이 확보될 것이고 정책적으로도 적지 않은 파급력을 미칠 것이다.

인공지능 윤리규범 논의가 경우에 따라서는 일견 원론적이고 교과서적 차원의 담론에 그친다고 볼 수 있지만, 간단한 추상적 내용만을 담고 있는 규범일지라도 이해관계자 사이의 치열한 이익충돌이 기저에 숨겨져 있게 마련임을 간과해서는 곤란하다. 윤리원칙의 내용이 매우 간략할 경우, 개별 규범에 담긴 내용은 물론 규범에 담기지 않은 이면적 사항이 무엇인지를 파악하고 함의를 분석하는 일이 보다 중요할 수 있고, 추상적 원칙을 분석하여 '행간'을 읽어낼 필요도 있다. 그 이외에도, 선언적인 내용이 선언 자체에 그칠 경우와 실제로 집행력을 확보하는 경우 사이의 구분도 중요하다. 이처럼 면밀한 분석과 판단이 이루어지지 않으면 우리나라 실정에 부합하지 않거나 부작용을 낳을 수 있는 규범임에도 '사실상의 국제표준(de facto global standard)'이라는 이유만으로 억지로 수용하는 결과를 낳을 수도 있다. 우리나라가 국제사회의 치열한 '규범전쟁(norm war)'에서 주체적 목소리를 내고 선도적 지위를 점유하기 위해서는 그와 같은 깊이 있는 분석과 논의가 지속되어야 한다.

Ⅴ. 윤리적 인공지능의 실현과제: 결론을 대신하여

이하에서는 결론을 대신하여 윤리적 인공지능의 실현과 정착을 위한 몇 가지 제언을 정리하도록 한다. 지금까지의 논의를 토대로 공감대를 얻은 가장 핵심적 원칙은 인류가 지향해야 할 목표는 인공지능이 인간의 존엄성과 권리의 실현을 위하여 활용되어야 한다는 점이다. 그러한 대원칙을 실현하는데 적절하지 않은 관념, 특히 아래에서 언급할 이

85) 방송통신위원회, "이용자 중심의 지능정보사회를 위한 원칙", 2019.
86) 카카오, "카카오 알고리즘 윤리 헌장", 2018; 카카오, "카카오 새로운 알고리즘 윤리 헌장", 2019.
87) 삼성전자, "지속가능경영보고서 2019", 2019, 70면.

분법적 관념은 인공지능 시대와 부합할 수 있도록 변화를 모색할 필요가 있다.

우선, 법규범과 윤리와 같은 여타 사회규범의 유형을 엄밀히 구별하는 태도를 들 수 있다. 앞서 언급한 것처럼 인공지능 기술은 지속적으로 발전하고 변화하는 상태에 있기 때문에, 법규범을 신속하게 마련하여 모든 것을 규제하여야 한다는 전통적 관념은 적절하지 않다. CNIL의 보고서가 강조하듯 윤리규범을 법규범을 마련하는 데 필요한 정보를 획득하는 선행단계의 논의로 이해하는 경우, 양자는 반드시 배타적 관계에 놓이지 않는다. 둘째로 기술과 규범을 엄밀하게 분리하여 파악하려는 태도도 바람직하지 않다. 과거 일부 기술철학 영역에서 기술 전반을 적대시한 결과 현실에 맞지 않은 규범이 주창되기도 했던 사례를 되풀이하지 않아야 한다. 인공지능 윤리의 맥락에서는 AMA나 XAI처럼 기술을 통해 규범을 실현하거나, 규범을 통해 기술발전에 대한 대중의 신뢰를 확보하는 방안과 같이 기술과 규범이 서로 보완적 역할을 할 수 있고, 그래야 한다. 셋째, 법규범 내부에서 해악의 위험성에 대한 사전적, 공적 규율방식을 보다 중시하는 영역이나, 발생한 해악에 대한 사후적, 사적 규율방식을 보다 중시하는 영역을 엄밀하게 구분하는 태도도 재고되어야만 한다. 인공지능에 의한 존재론적 위협에 대처하기 위해서는 기존 규율방식을 넘어선 새로운 사고가 요청되고, '책임성(accountability)'이라는 종전과 다른 책임 관념이 지향해야 할 방향 역시 그와 다르지 않다.

전반적으로 현재 인공지능 영역에서 진행되고 있는 국제사회의 논의는 전통적 이분법의 관념과는 크게 다르다. 예를 들어, 오늘날 선진적 인공지능 방법론의 전형으로 일컬을 만한 것은 딥러닝(Deep Learning) 기술일 텐데, 실무적으로는 '딥러닝' 기술만을 배타적으로 이용하는 경우는 오히려 흔치 않을 것이다. 딥러닝 방식과 대비된 규칙기반 전문가 시스템도 여전히 사용되고 있고, 여러 유형의 인공지능 모델이 조합하여 이용되는 경우가 빈번하다. 딥러닝 알고리즘도 수많은 변이가 이루어지고 있고, 딥러닝 패러다임 자체가 전환될 가능성도 배제할 수 없다. '인공지능'과 인공지능에 해당하지 않는 영역을 이분법적으로 나누는 일도 현실에서는 쉽지 않다. 개발된 기술을 상용화하는 과정에서는 흔히 기존의 제품과 서비스에 인공지능 기능을 약간씩 포함하고, 이를 점차 확대하고 고도화하는 과정을 거치게 되는 것이 일반적인 과정이다. 이러한 경우에, 개별적 제품과 서비스에서 인공지능에 해당하는 요소를 별도로 떼어내어 법규범적 검토를 하는 것은 현실적이지 않다.

인공지능 기술이 변화하는 속도가 빠르고 방향에 대한 예측이 어렵다는 점은 입법자가 인공지능의 단일한 본성을 상정하고 사전적 규제 체계를 마련한다거나 사후적 민·형사 책임 위주의 법체계를 입법하는 '경성법(hard law)'으로 불리는 전통적 접근을 어렵게 한다. 인공지능은 국제사회의 이해관계자들 사이의 첨예한 대립이 빈번하다는 점에서, 구속력과 강제력을 지나치게 강조할 경우에는 도리어 무규범(anomie) 상태로 흐르거나, 상충하는 법체계가 복잡하게 혼재하는 상태가 초래될 가능성도 있다. 이와 달리, 실정법적인 구속력과 강제력은 없지만, 행위규범의 일종으로 사회 구성원에게 사실상의 영향력을 미치는 '연성법(soft law)'을 통해 경성법 체계와 조화를 시도하는 방안을 대안으로 생각해볼 수 있다. 앞서 보았듯 여러 사적·공적 주체가 가이드라인, 원칙, 행동강령과 같은 이름으로 만들어내고 있는 연성법은 미래사회의 방향성을 제시하고, 현실과 이상 간의 간극을 메우는 등 보다 유연한 대처가 가능한 장점을 가지기 때문이다.[88] 연성법은 기술규제, 자율규제와 같은 최근의 논의와 부합하는 측면이 있어 인공지능 거버넌스 담론에서 많은 지지를 획득하고 있다.

다만, 연성법 체제를 조금 더 본격적으로 도입하기 위해서는 이를 위한 충분한 준비가 필요하다. 가장 먼저 우리나라 법체계 구조와 관행상, 경성법

88) 최난설헌, "연성규범(Soft Law)의 기능과 법적 효력 : EU 경쟁법상의 논의를 중심으로", 『법학연구』 제16집 제2호, 2013, 96-98면 참조.

이 연성법에 비해 높은 예측가능성을 가진다는 점이 지적될 수 있다. 또한 연성법이 경성법의 보완재라기보다는 예측가능성 낮은 '추가적 규제'로 무분별하게 남용될 경우에는, 수범자와 사회의 커다란 혼란을 불러일으킬 가능성도 있다. 행정, 사법, 산업 영역에 대한 사회구성원들의 신뢰(trust) 정도도 중요한 변수가 될 수 있다. 신뢰수준이 낮은 영역에 연성법을 통한 독자적 또는 자율적 규율권한을 과하게 부여할 경우 상당한 사회적 비용이 유발될 수도 있을 것이기 때문이다. 다른 한편, 연성법이 그저 대원칙에 대한 선언으로만 비쳐지고 현실적이고 실질적 구속력이 확보되지 못한다면, 많은 경우에 이는 불필요한 낭비만 초래할 수도 있다. 그런 면에서, 위에서 살펴본 다양한 규범들도 실효성이 있는 규범과, 실효성의 확보가 어려운 규범으로 나누어 살펴볼 수 있다. 향후의 인공지능 윤리 담론은 이와 같은 다양한 측면을 함께 고려한 풍부하고 깊이 있는 논의가 되어야 할 것이다.

<참고문헌>

국내문헌

변순용, 『윤리적 AI로봇 프로젝트』, 어문학사, 2019.

웬델 월러치/콜린 알렌(노태복 역), 『왜 로봇의 도덕인가』, 메디치미디어, 2014.

이원우 외, 『4차 산업혁명 시대의 기술혁신과 규제정책』, 홍문사, 2019.

이원태 외, 『4차산업혁명시대 산업별 인공지능 윤리의 이슈 분석 및 정책적 대응방안 연구』, 대통령 직속 4차산업혁명위원회, 2018.

캐시 오닐(김정혜 역), 『대량살상 수학무기』, 흐름출판, 2017.

프랭크 파스콸레(이시은 역), 『블랙박스 사회』, 안티고네, 2016.

한희원, 『인공지능(AI) 법과 공존윤리』, 박영사, 2018.

고인석, "아시모프의 로봇 3법칙 다시 보기: 윤리적인 로봇 만들기", 『철학연구』 제93집, 2012.

고학수/정해빈/박도현, "인공지능과 차별", 『저스티스』 통권 제171호, 2019.

김건우, "로봇윤리 vs. 로봇법학: 따로 또 같이", 『법철학연구』 제20권 제2호, 2017.

김미리/윤상필/권헌영, "인공지능 전문가 윤리의 역할과 윤리 기준의 지향점", 『법학논총』 제32권 제3호, 2020.

김종호, "인공지능 시대의 윤리와 법적 과제", 『과학기술법연구』 제24권 제3호, 2018.

김효은, "인공지능과 윤리", 『인공지능과 법』, 박영사, 2019.

남중권, "머신러닝 알고리즘의 데이터 처리에 대한 법적 제한의 한계: 개인정보보호와 차별금지의 측면에서", 『과학기술과 법』 제10권 제1호, 2019.

박도현, "인공지능과 자율성의 역학관계", 『홍익법학』 제20권 제3호, 2019.

박도현, "미래를 향한 인공지능 정책: 우리는 AI를 신뢰할 수 있을까?", 『2019 국제학술대회 보고서』, 서울대학교 법과경제연구센터, 2019.

방송통신위원회, "이용자 중심의 지능정보사회를 위한 원칙", 2019.

삼성전자, "지속가능경영보고서 2019", 2019.

손화철, "기술철학에서의 경험으로의 전환: 그 의의와 한계", 『철학』 제87집, 2006.

송상현, "인공지능과 도덕성", 『법조』 제67권 제6호, 2018.

송성수, "공학단체의 윤리강령에 대한 비교분석: 미국과 한국의 사례를 중심으로", 『공학교육연구』 제11권 제3호, 2008.

신상규, "인공지능 시대의 윤리학", 『지식의 지평』 제21권, 2016.

심민석, “로봇과 인공지능(AI)의 법적·윤리적 입법방안에 관한 연구”, 『비교법연구』 제19권 제2호, 2019.

양희태, “인공지능의 위험성에 대한 우려로 제정된 아실로마 인공지능 원칙”, 『과학기술정책』 제27권 제8호, 2017.

오병철, “인공지능 로봇에 의한 손해의 불법행위책임”, 『법학연구』 제27권 제4호, 2017.

오요한/홍성욱, “인공지능 알고리즘은 사람을 차별하는가?”, 『과학기술학연구』 제18권 제3호, 2018.

유재홍, “일본의 인공지능 전략 동향: AI 전략”, 『월간 SW 중심사회』 통권 제60호, 2019.

이상형, “윤리적 인공지능은 가능한가? – 인공지능의 도덕적, 법적 책임 문제 –”, 『법과 정책연구』 제16권 제4호, 2016.

이중기/오병두, “자율주행자동차와 로봇윤리: 그 법적 시사점”, 『홍익법학』 제17권 제2호, 2016.

이중원, “인공지능에게 책임을 부과할 수 있는가?: 책무성 중심의 인공지능 윤리 모색”, 『과학철학』 제22권 제2호, 2019.

정보문화포럼/한국정보화진흥원, “지능정보사회 윤리 가이드라인”, 2018.

정채연, “지능정보사회에서 지능로봇의 윤리화 과제와 전망 – 근대적 윤리담론에 대한 대안적 접근을 중심으로 –”, 『동북아법연구』 제12권 제1호, 2018.

최난설헌, “연성규범(Soft Law)의 기능과 법적 효력 : EU 경쟁법상의 논의를 중심으로”, 『법학연구』 제16집 제2호, 2013.

카카오, “카카오 알고리즘 윤리 헌장”, 2018.

카카오, “카카오 새로운 알고리즘 윤리 헌장”, 2019.

한애라, ““사법시스템과 사법환경에서의 인공지능 이용에 관한 유럽 윤리헌장”의 검토 – 민사사법절차에서의 인공지능 도입 논의와 관련하여 –”, 『저스티스』 통권 제172호, 2019.

한희원, “인공지능(AI) 치명적자율무기(LAWs)의 법적·윤리적 쟁점에 대한 기초연구”, 『중앙법학』 제20집 제1호, 2018.

황현주, “AI 연구개발과 활용 촉진을 위한 ‘AI 개발 가이드라인(안)’”, 『NIA Special Report』 2017-9, 2017.

외국문헌

Smith, Brad/Shum, Harry, *The Future Computed*, Microsoft, 2018.

Asaro, Peter M., “What Should We Want From a Robot Ethic?”, International Review of Information Ethics Vol. 6, No. 12, 2006.

Association for Computing Machinery US Public Policy Council, “Statement on Algorithmic Transparency and Accountability”, 2017.

Awad, Edmond et al., “The Moral Machine experiment”, Nature Vol. 563, 2018.

Barocas, Solon/Selbst, Andrew D., “Big Data’s Disparate Impact”, California Law Review Vol. 104, 2016.

Bonnefon, Jean-François/Shariff, Azim/Rahwan, Iyad, “The Social Dilemma of Autonomous Vehicles”, Science Vol. 354, No. 6293, 2016.

Bostrom, Nick, “Ethical Issues in Advanced Artificial Intelligence”, Science Fiction and Philosophy: from Time Travel to Superintelligence, 2003.

Burrell, Jenna, “How the machine ‘thinks’: Understanding Opacity in Machine Learning Algorithms”, Big Data & Society, 2016.

Bynum, Terrell, “Computer and Information Ethics”, Stanford Encyclopedia of Philosophy, 2015.

Caplan, Robyn et al., “Algorithmic Accountability: A Primer”, Data & Society, 2018.

Castelluccia, Claude/Le Métayer, Daniel, “Understanding Algorithmic Decision-Making: Opportunities and Challenges”, Panel for the Future of Science and Technology, 2019.

Citron, Danielle Keats, “Technological Due Process”, Washington University Law Review Vol. 85, 2008.

Dean, Jeff/Walker, Kent, “Responsible AI: Putting our principles into action”, 2019.

Demiaux, Victor/Abdallah, Yacine Si, “How can Humans keep the Upper Hand: The Ethical Matters Raised by Algorithms and Artificial Intelligence”, 2017.

Diakopoulos, Nicholas, “Accountability in Algorithmic Decision Making”, Communications of the ACM Vol. 59, No. 2, 2016.

Di Fabio, Udo et al., “Ethics Commission Automated and Connected Driving”, Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany, 2017.

Engineering and Physical Science Research Council, “Principles of robotics”, 2010.

European Commission for the Efficiency of Justice, “European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment”, 2018.

European Commission High-Level Expert Group on Artificial Intelligence, “Ethics Guidelines for Trustworthy AI”, 2019.

European Parliament, “European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))”, 2017.

Franssen, Maarten/Lokhorst, Gert-Jan/van de Poel, Ibo, “Philosophy of Technology”, Stanford Encyclopedia of Philosophy, 2018.

Future of Life Institute, “Autonomous Weapons: An Open Letter from AI & Robotics Researchers”, 2015.

Future of Life Institute, “Asilomar AI Principles”, 2017.

G20, “G20 AI Principles”, 2019.

Google, “Perspectives on Issues in AI Governance”, 2019.

Gotterbarn, Don et al., “ACM Code of Ethics and Professional Conduct”, Association for Computing Machinery, 2018.

Infocomm Media Development Authority/Personal Data Protection Commission Singapore, “Model Artificial Intelligence Governance Framework Second Edition”, 2020.

Institute of Electrical and Electronics Engineers, “Ethically Aligned Design Version 1: A Vision for Prioritizing Human Well-being with Artificial Intelligence and Autonomous Systems”, 2016.

Institute of Electrical and Electronics Engineers, “Ethically Aligned Design Version 2: A Vision for Prioritizing Human Well-being with Autonomous

and Intelligent Systems", 2017.
Institute of Electrical and Electronics Engineers, "Ethically Aligned Design First Edition: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems", 2019.
Microsoft, "Responsible Bots: 10 Guidelines for Developers of Conversational AI", 2018.
Microsoft, "Six Principles for Developing and Deploying Facial Recognition Technology", 2018.
Moor, James H., "The Nature, Importance, and Difficulty of Machine Ethics", IEEE Intelligent Systems Vol. 21, No. 4, 2006.
Mulgan, Richard, "'Accountability': An Ever-Expanding Concept?", Public Administration Vol. 78, No. 3, 2000.
Murphy, Robin R./Woods, David D., "Beyond Asimov: The Three Laws of Responsible Robotics", IEEE Intelligent Systems Vol. 24, No. 4, 2009.
Nevejans, Nathalie et al., "Open Letter to The European Commission Artificial Intelligence And Robotics", 2018.
Nissenbaum, Helen, "Accountability in a Computerized Society", Science and Engineering Ethics Vol. 2, No. 1, 1996.
Noorman, Merel, "Computing and Moral Responsibility", Stanford Encyclopedia of Philosophy, 2018.
OECD, "Recommendation of the Council on Artificial Intelligence", 2019.
Palmerini, Erica et al., "Guidelines on Regulating Robotics", RoboLaw Project, 2014.
Personal Data Protection Commission Singapore, "A Proposed Model AI Governance Framework", 2019.
Pichai, Sundar, "AI at Google: our Principles", 2018.
Stoker, Gerry, "Governance as Theory: Five Propositions", International Social Science Journal Vol. 50, No. 155, 1998.
Stone, Peter et al., "Artificial Intelligence and Life in 2030", One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, 2016.
Sullins, John P.(권은정 감수), "미국의 인공지능(AI) 윤리 및 거버넌스 현황", 『경제규제와 법』 제12권 제2호, 2019.
Takanishi, A., "World Robot Declaration", International Robot Fair, 2004.
The Conference toward AI Network Society, "Draft AI R&D Guidelines for International Discussions", 2017.
The Conference toward AI Network Society, "Draft AI Utilization Principles", 2018.
The White House, "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy", 2012.
The White House Office of Science and Technology Policy, "American Artificial Intelligence Initiative: Year One Annual Report", 2020.
Trump, Donald J., "Executive Order on Maintaining American Leadership in Artificial Intelligence", Federal Register: White House, 2019.

Tucker, Catherine, “Privacy, Algorithms, and Artificial Intelligence”, in Agrawal, Ajay/Gans, Joshua/Goldfarb, Avi (eds.), *The Economics of Artificial Intelligence*, University of Chicago Press, 2019.

U.S. Executive Office of the President, “BIG Data: Seizing Opportunities, Preserving Values”, 2014.

U.S. Executive Office of the President, “BIG Data: Seizing Opportunities, Preserving Values Interim Progress Report”, 2015.

U.S. Executive Office of the President, “Big Data: A Report on Algorithmic Systems”, Opportunity, and Civil Rights, 2016.

U.S. Executive Office of the President, “Artificial Intelligence, Automation, and the Economy”, 2016.

U.S. Executive Office of the President/National Science and Technology Council Committee on Technology, “Preparing for the Future of Artificial Intelligence”, 2016.

UK House of Lords Select Committee on Artificial Intelligence, “AI in the UK: ready, willing and able?”, Report of Session 2017–19, 2018.

United Nations Global Pulse/International Association of Privacy Professionals, “Building Ethics into Privacy Frameworks for Big Data and AI”, 2018.

Veruggio, Gianmarco, “EURON Roboethics Roadmap(Release 1.1)”, EURON Roboethics Atelier, Genua, 2006.

Veruggio, Gianmarco, “EURON Roboethics Roadmap(Release. 1.2)”, 2007.

Villani, Cédric et al., “For A Meaningful Artificial Intelligence: Towards A French and European Strategy”, 2018.

Vought, Russell T., “Guidance for Regulation of Artificial Intelligence Applications(Draft)”, 2020.

Winfield, Alan F. T./Jirotka, Marina, “Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence Systems”, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences Vol. 376, No. 2133, 2018.

<ABSTRACT>

Challenges of Establishing Ethics Principles and a Governance Regime for Artificial Intelligence

Haksoo Ko / Dohyun Park / Narae Lee

This paper examines various ethical issues related to Artificial Intelligence (AI). Major characteristics of discussions on AI ethics can be explained as follows. First, since Asimov's three laws of robotics were first proffered, a main focus of discussions has shifted from robot actors toward human actors. Second, the role of "accountability" has expanded as the gap widened between granting redress through legal liability and calling for responsibility for undesirable happenstance. Third, AI-related technologies as well as their applications are changing constantly and, as such, specific contexts and circumstances should be considered in developing relevant norms.

Discussions on AI ethics principles have progressed rapidly in recent years. There are several types of discussions, with varying degree of the scope and depth of the discussions. First, certain discussions focused on establishing consensus and deriving basic principles. Second, other discussions dealt more specific issues, often accompanied with in-depth analyses of relevant issues. Third, still other types of discussions focused on establishing ethical norms as well as a governance structure in order to carry out the ethical norms.

While discussions on AI ethics in Korea first began in 2007, more serious discussions did not resume until recent years. Generally speaking, discussions in Korea are in conformity with discussions that have been taking place in other countries and the international community.

Keywords : 인공지능(Artificial Intelligence), 인공지능 윤리(AI Ethics), 규제(Regulation), 거버넌스(Governance), 책임성(Accountability)