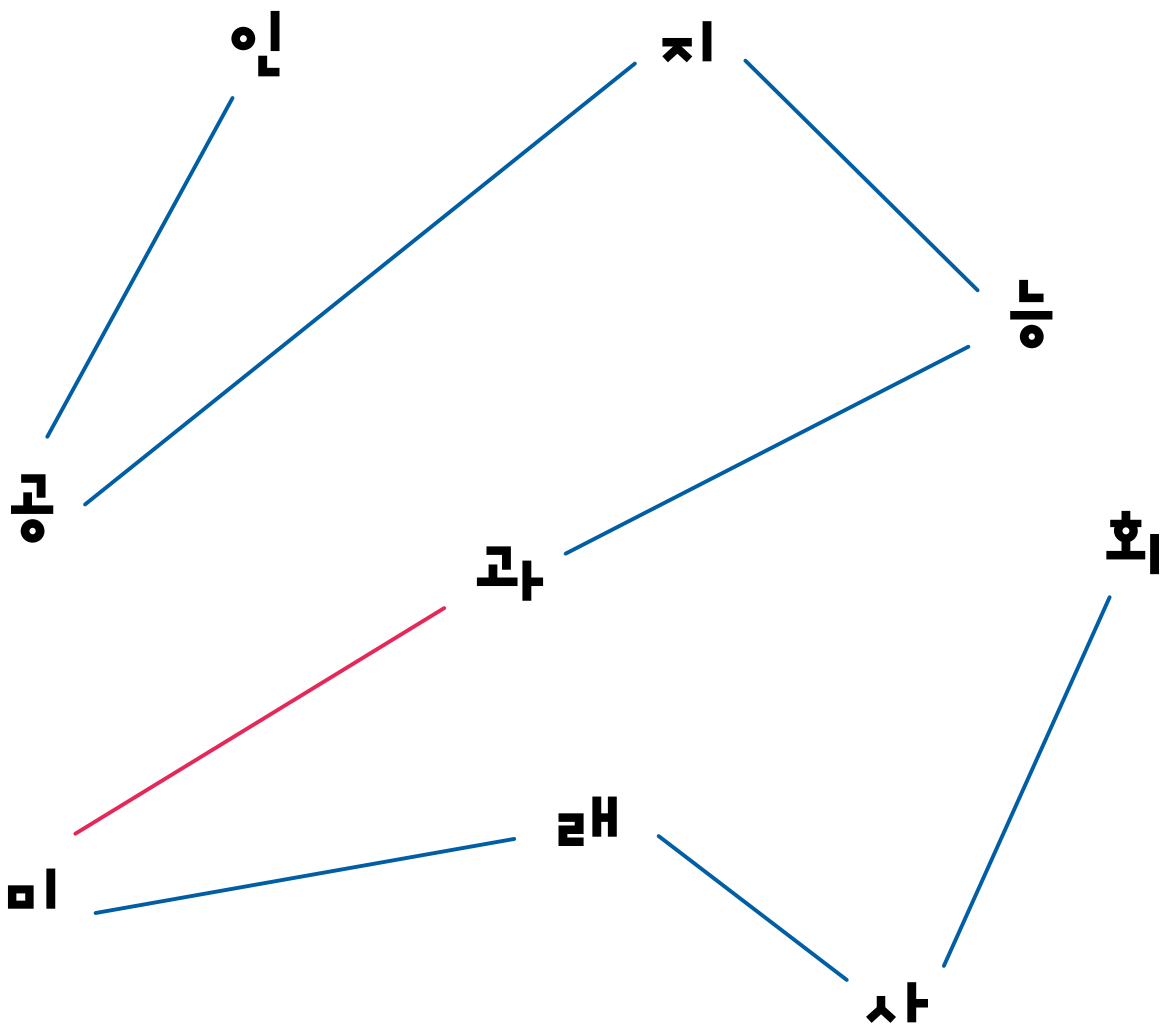
서울대학교
법과경제연구센터
인공지능정책
이니셔티브

이슈페이퍼 04

“인공지능과 미래사회”



http://ai.re.kr

윤리적 인공지능의 실현과 과제

고학수(교수, 서울대학교 법학전문대학원)
이나래(변호사, 서울대학교 법학대학원 박사과정)
박도현(변호사, 서울대학교 법학대학원 박사과정)

윤리적 인공지능의 실현과 과제

고학수(교수, 서울대학교 법학전문대학원)
이나래(변호사, 서울대학교 법학대학원 박사과정)
박도현(변호사, 서울대학교 법학대학원 박사과정)

웹에서 PDF 바로 보기

요약

본고는 인공지능 시대에 우리나라가 앞으로 어떠한 형태의 윤리규범 체계를 갖추어야 하는가라는 문제의식에 대한 실마리를 제공하는 것을 목적으로 하는 글이다. 아시모프 '로봇 3원칙' 이래로, 인공지능 윤리규범 이슈는, 첫째, 논의의 주체가 설계자, 제작자, 이용자와 같이 배후에 위치한 인간에게로 확장되었고, 둘째, 책임귀속에 관련된 과거의 대전제가 오늘날에는 통용되지 않게 된 경우가 많아졌으며, 셋째, 인공지능에 대한 대중의 알 권리와 이해관계자의 주체적 참여에 대한 요구가 늘어나면서 책임성(accountability)에 대한 관념이 확장되고 있고, 넷째, 일률적인 규정을 기계적으로 적용하는 방식의 한계가 지적되고 그 대신 인공지능이 활용되는 구체적 맥락(context) 위주의 접근법에 대한 필요성이 강조되는 방향으로 논의가 이루어지고 있다.

인공지능의 윤리적 측면에 관한 국내외의 논의는 최근 2년여의 기간 동안 급속도로 진전이 이루어졌다. 그동안 해외에서 제시된 윤리규범은 대체적으로는, ① 기본원칙을 중심으로 하는 윤리규범 유형, ② 기본원칙에 더하여 주요 이슈에 대해 상세하게 함께 다룬 윤리규범 및 보고서 유형, ③ 윤리규범의 정립방안에 대한 구체적 방법론을 함께 언급한 유형으로 구분지어 파악할 수 있다. 기본원칙 중심의 윤리규범은 OECD 권고안, 일본 총무성 가이드라인, 아실로마 원칙, 그리고 그 이외에 마이크로소프트, 구글과 같은 사기업에서 발표한 규범 등을 들 수 있다. 기본원칙 중심의 윤리규범은 다양한 참여자들의 다양한 관점 및 상이한 이해관계를 종합하기에 용이한 방식인 한편, 추가적인 후속 규범이 마련되지 않으면 규범력이 상대적으로 낮을 가능성이 있다. 다음으로, 기본원칙과 주요 이슈를 함께 다룬 윤리규범 유형으로는 IEEE 보고서, 영국 상원 보고서, 그리고 UNGP & IAPP의 보고서를 들 수 있다. 구체적 논의에 대한 세부사항까지도 포함하는 셋째 유형으로는 유럽의회 결의안, EU 집행위원회 고위급 전문가 그룹의 가이드라인 등을 들 수 있다. 이러한 윤리규범 중에는, 원칙에 대한 선언 그 자체로 의미를 가지는 것도 있고, 후속 작업과 모니터링의 과정을 통해 더 구체화하고 규범력을 확보할 수 있는 장치가 마련된 것도 있다.

우리나라에서 인공지능 윤리 이슈에 대한 논의는 2007년 발표된 ‘로봇윤리헌장 초안’에서 시작되었다. 세계적으로도 빠른 편이다. 그러나 그로부터 10여년이 지난 지금도 논의가 크게 진전되지는 못한 상황이다. 하지만, 알파고에서 촉발된 충격 이후로 인공지능 규범 마련에 대한 관심이 늘어나면서, 로봇기본법안, 지능정보사회 윤리헌장, 국가정보화 기본법 전부개정법률안 등 다양한 형태로 관심이 표출되고 있는 중이다. ‘개발자, 공급자, 이용자 등에 대한 공공성 책무성 통제성 투명성’ 등의 명제로 대표되는 국내의 논의는 해외의 논의와 대체로 일맥상통한 내용을 담고 있다. 그러나 인공지능 규범 논의는 국내의 실정과 부합하는 동시에 국제사회와 발맞추어 진행해나가야 한다는 점, 일견 원론적이고 교과서적인 내용의 단순한 나열로 비쳐지는 경우에도 그 배후에는 이해관계자들 사이에 자신에게 유리한 방향으로 향후 규범논의를 이끌어가려는 치열한 이익대립이 숨겨져 있는 점을 고려하여 관련 이슈들에 대한 면밀한 검토와 분석이 필요하다.

인류가 지향해야 할 궁극의 목표는 인공지능이 인간 존엄성과 기본권 실현을 위한 방향으로 활용되어야 한다는 점이다. 따라서 그러한 목적과 부합하지 않는 여러 관념, 특히 이분법적 관념들은 인공지능 시대와 부합하도록 변화를 모색할 필요가 있다. 가령, 윤리를 비롯한 여타 사회규범 유형을 실정법적 법규범과 엄밀히 구별하는 태도, 기술과 규범을 엄밀하게 분리하는 태도는 이 논의의 맥락에서는 대체로 도움이 되기 어려운 태도이다. 구속력과 강제력은 높지 않지만, 구성원에게 사실상의 행위규범으로 기능하는 '연성법(soft law)'이 인공지능 윤리규범 논의의 대안으로 부상하는 것에 주목할 필요가 있다. 또한, 인공지능의 '책임'에 관해서도, 법적 책임을 의미하는 'liability' 개념이 많이 언급되지 않고, 그보다 더 폭넓은 의미의 사회적 책임을 의미하는 'accountability' 개념이 좀 더 흔하게 언급되는 것은 중요한 함의를 갖는다.

1. 들어가는 글

인공지능 기술이 급속히 발전하고 널리 확산되면서, 부작용과 폐해를 미연에 방지하기 위한 규범적 논의도 점차 늘어가는 중이다. 미국 스탠퍼드 대학에서는 2016년 9월 인공지능 100년 연구 첫 번째 보고서를 발표했고, 오바마 행정부는 2016년 10월과 12월 기술적 연구개발과 별도로 사회 경제적 이슈에 집중한 보고서를 내놓았다. AI Now Institute는 2016년부터 주로 사회 윤리적 이슈에 관한 연례보고서를 발표해왔다. 구글(Google)에서도 올해 초 이와 같은 대열에 동참하여 거버넌스 보고서를 발표했다. 마이크로소프트(Microsoft), 인텔(Intel), 소니(Sony)를 비롯한 글로벌 기업들 또한 보고서, 원칙(principle), 가이드라인, 모범사례(best practice) 등 명칭을 불문하고 윤리적 고려를 반영한 나름대로의 실무관행을 구축해가고 있다.

윤리적 인공지능에 대한 관심은 유럽에서도 높은 수준으로 나타나고 있다. 유럽에서는 인공지능의 윤리적 측면에 대한 관심이 일찍이 나타나기 시작했고, 유럽연합(European Union, EU) 차원에서의 논의를 보면 엄격한 법제를 도입함으로써 신기술 활용과 기본권 보호를 조화하려는 움직임이 좀 더 두르러지게 나타난다. 대표적 예로, 유럽연합에서는 개인정보보호규정(General Data Protection Regulation, 이하 “GDPR”)이 2018년 5월 25일부터 시행되고 있는데, GDPR에는 인공지능의 윤리적 측면에서 중요한 의미를 지니는 조항들이 적지 않게 포함되어 있다. 핵심적으로, GDPR은 개인정보 주체에게 법적 효력이나 그와 유사한 중대한 효과를 미칠 경우에 프로파일링을 포함한 ‘오직 자동화된 의사결정’의 대상이 되지 않을 권리를 인정한다. 정보주체는 또한 ‘설명을 요구할 권리(right to explanation)’와 ‘잊힐 권리(right to be forgotten)’로 널리 알려져 있는 정보 제공권, 열람권, 삭제권, 반대권과 같은 다양한 권리를 행사할 수 있다. GDPR은 유럽연합 역내 모든 구성원에게 적용되므로 해외의 사업자도 경우에 따라 이를 준수할 의무가 부과되고, 이를 위반할 경우 강력한 제재가 동반된다.

이 글은, 우리나라가 제4차 산업혁명 시대에 인공지능의 윤리적 측면에 관하여 어떤 형태의 규범체계를 갖추어야 하는가라는 질문을 염두에 두고, 최근의 국내외 논의 동향은 어떠한지 살펴보고 이로부터 시사점을 모색해 보고자 한다. 이하에서는 먼저 인공지능 윤리규범 논의의 변천사를 개관하면서 그동안 지적되어온 주요한 쟁점을 살펴보도록 한다(II). 다음으로 이런 과정을 거쳐 정립된 윤리규범의 현황을 살펴보기 위하여 해외 주요 윤리규범의 내용을 고찰하고 그 의의와 한계를 도출한다(III). 끝으로 여기서 얻은 시사점을 국내 윤리규범 논의에 적용하고 향후 나아가야 할 방향을 모색한다(IV).

본격적 논의에 들어가기에 앞서, 본고의 몇 가지 전제사항에 대해 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 부착된 경우까지 포괄한 넓은 개념이다. 둘째로, 본고의 '인공지능'은 오늘날 국내외 학계와 실무의 연구개발 대상인 유형에 한정하도록 한다. 일부 매체에서는 인류와 같은 방식으로 사고하고 행동하거나, 그 이상의 인지능력을 확보하여 인류의 생존을 위협하는 미래지향적 인공지능의 모습을 그려내고 있기도 하다. 물론 이러한 인공지능에 관해서도 '인공지능 윤리'의 측면에 관해 다룰 수는 있지만, 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 사회를 보다 바람직한 방향으로 이끌기 위하여 사회구성원에게 요구되는 사회규범의 일종으로, 법규범과 비교하여 강제성과 명확성은 다소 낮지만 도덕성과 유연성은 보다 높은 유형으로 간주한다.[01]

2. 인공지능 윤리의 변천사와 주요 쟁점

1. 윤리주체의 다변화와 인간을 중시하는 윤리관

인공지능 윤리규범의 시초로 많은 이들은 SF 소설가 아이작 아시모프(Isaac Asimov)가 1942년 런어라운드(Runaround)라는 소설에서 최초로 언급한 '로봇 3원칙'을 꼽는다. 3원칙의 내용은 다음과 같다. 첫째, 로봇은 인간에게 해를 끼치거나, 어떠한 행동도 하지 않아 인간에게 해가 가해지도록 하면 안 된다. 둘째, 로봇은 1원칙에 위배되지 않는 한 인간의 명령에 복종하여야 한다. 셋째, 로봇은 1원칙과 2원칙에 위배되지 않는 한 자신을 보호하여야 한다.

로봇 3원칙을 엄격하게 고수하면, 극단적으로는 인간에게 위해의 여지가 있으면 정당방위와 같은 해악을 방지하기 위한 개입조차도 할 수 없다는 반직관적 결론에 다다른다. 아시모프는 1985년 로봇과 제국(Robots and Empire)이라는 소설에서 "로봇은 인류에게 해를 끼치거나, 어떠한 행동도 하지 않음으로써 인류에게 해를 가하도록 하면 안 된다"는 '0원칙'을 제시한다. 이는, 3원칙의 '인간' 자리에 '인류'를 도입하여 위와 같은 모순적 결론을 방지하려는 것이다.[02]

---

01 김종호, "인공지능 시대의 윤리와 법적 과제", 『과학기술법연구 제24권 제3호』, 2018, 184-187면 참조.

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

로봇 3원칙은 이하에서 살펴볼 2006년 유럽로봇연구 네트워크 로봇윤리 로드맵의 출발점이 되는 등, 인공지능 윤리 논의에서 나름대로의 역할을 담당하였다. 그러나 어디까지나 소설의 일부에 불과한 로봇 3원칙만으로는 윤리적 문제를 해결하는 데 한계가 있었다. 로봇 3원칙은 무엇이 문제였을까? 선언적 원칙이어서 구체성도 부족하였지만 무엇보다도 윤리주체가 인간이 아니라 로봇인 점이 주로 지적되었다. 로봇을 만들고 이용하는 '인간'이 준수하여야 할 윤리를 '로봇'에 전가할 수 있도록 오해될 여지가 있었기 때문이다.[03] 이러한 문제는 최초의 인공지능 윤리규범으로 평가받고 있는, 일본 후쿠오카에서 2004년 발표된 '세계로봇선언(World Robot Declaration)'에서도 마찬가지였다.[04] '로봇과 인간의 공존'을 중시하는 세계로봇선언은 로봇이 인간에게 일방적으로 복종하는 관계로 설정한 '로봇 3원칙'에 비해 진일보한 것으로 평가될 수 있지만, 세계로봇선언에서도 인간의 윤리에 대한 언급은 부재했다.

다른 한편, 유럽로봇연구 네트워크(European robotics research Network, EURON)은 2003년부터 3년 동안의 연구를 거쳐 2006년에 '로봇윤리 로드맵'을 발표했는데, 이 로드맵은 인간의 윤리를 주된 목표로 함을 명시했다.[05] 3대 이해당사자인 설계자(designer), 제작자(manufacturer), 이용자(user)를 특정하여, 로봇을 제작하거나 활용할 때 중시해야 할 인간의 존엄성과 권리를 비롯한 13대 원칙을 통해 이러한 목표를 구체화했다. 이듬해인 2007년 우리나라 산업자원부도 인간중심, 인간과 로봇의 공존, 인간과 로봇의 윤리라는 그동안의 논의를 종합적으로 반영한 '로봇윤리헌장' 초안을 공개하였다. 이런 과정을 거치면서, 오늘날 인공지능 윤리는 대략적으로 ① 인공지능 자체의 윤리, ② 인공지능 설계자나 제작자의 윤리, ③ 인공지능 이용자의 윤리라는 세 가지 범주로 나누어 볼 수 있게 되었고, 인간과 인공지능의 공존 속에서 인간의 존엄성과 기본권을 도모하는 윤리관이 수립되었다.[06]

그렇다면 '인공지능 자체의 윤리'는 어떤 모습으로 발전해왔을까?

---

03 Robin R. Murphy David D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics", IEEE Intelligent Systems Vol. 24, Issue 4, 2009, pp. 14-20은 이를 지적하고, 3원칙의 대안을 제시한다.

04 International Robot Fair 2004, "World Robot Declaration", http://prw.kyodonews.jp/prwfile/prdata/0370/release/200402259634/index.html, 2004. 2. 25.자.

05 Gianmarco Veruggio, EURON Roboethics Roadmap(ver. 1.1), 2006, p. 7. 이는 이듬해 발표된 로드맵 1.2버전에서도 마찬가지이다.

06 정채연, "지능정보사회에서 지능로봇의 윤리화 과제와 전망 - 근대적 윤리담론에 대한 대안적 접근을 중심으로 -", 『동북아법연구 제12권 제1호』, 2018, 90-93면; Peter M. Asaro, "What Should We Want From a Robot Ethic?", International Review of Information Ethics Vol. 6, 2006, pp. 9-16 참조.

오늘날 인공지능 자체의 윤리는 로봇 3원칙을 넘어, 인공지능의 의사결정이 인간의 윤리적 직관과 부합하게 조정하는 보다 적극적인 방법론을 가리키는 것으로 이해된다. 윤리적 의사결정의 주체인 인공지능은 통상적으로 '인공적 도덕행위자(Artificial Moral Agent, 이하 "AMA")'로 불린다.[07] AMA를 만드는 기술적 방식은 윤리적 규칙을 학습시키는 방식(하향식 접근), 윤리적 사례를 학습시키는 방식(상향식 접근), 양자를 혼합한 방식으로 대별된다. 그러나 어떤 방식이건, 가령 자율주행차가 사고에 직면한 상황에서 탑승자와 보행자 중 누구를 우선시할지와 같은 현실적 문제(일명 '트롤리 딜레마')에서 한계를 지적받고 있다. 단순화하여 생각하면, 단순한 공리주의 원칙을 적용한 자율주행차라면 2명 이상의 보행자를 살리기 위하여 탑승자를 사망에 이르는 의사결정을 할 것인데, 소비자들이 그와 같은 자율주행차는 구매하지 않을 것이기 때문에 시장의 성립 자체가 어려울 수 있다는 것이다.[08]

2. 책임귀속의 어려움을 극복하기 위한 방법론 모색

한편, 사회규범의 목적이 구성원들의 행위를 바람직한 방향으로 이끄는 것에 있다고 본다면, 그러한 목적과 부합하지 않은 행위를 했거나, 할 것으로 예상되는 구성원에 대한 조정(coordination) 문제가 사회규범의 핵심적 역할이 된다. 전통적으로 조정기능을 매개하는 수단으로 활용되어온 것이 바로 '책임' 개념이고, 사회규범을 대표하는 윤리와 법 분야에서의 'responsibility'와 'liability'가 여기에 대응되는 용어이다. 흔히들 말하는 "법은 도덕의 최소한"이라는 격언에 의하면 전자가 후자에 비해 다소 넓은 개념으로 볼 수 있지만, 반드시 그렇지는 않다. 그동안 양자의 본성과 관계에 대하여는 수많은 논의가 있어왔지만, 간단명료한 답이 도출된 것은 아니다. 다만, 양자는 밀접한 관계가 있어, 어떤 경우는 법규범이, 어떤 경우는 도덕이나 윤리규범이 서로를 이끌어가는 상호보완적 관계라고 할 수 있다.[09] 이들은 오랫동안 비슷한 전제를 공유해왔는데, 어떤 행위는 자율적 주체인 인간의 자유의지에서 비롯된 것으로, 주체는 행위가 낳는 결과를 예견할 수 있고, 주체와 행위 간의 인과관계가 명확하다는 것이다.[10]

---

07 대표적 연구로, 웬델 월러치 콜린 알렌, 노태복 역, 『왜 로봇의 도덕인가』, 메디치미디어, 2014가 있다.

08 Jean-François Bonnefon Azim Shariff Iyad Rahwan, "The social dilemma of autonomous vehicles", Science Vol. 354, Issue. 6293, 2016, pp. 1573-1576.

09 김건우, "로봇윤리 vs. 로봇법학: 따로 또 같이", 『법철학연구 제20권 제2호』, 2017, 33면 이하 참조.

10 Merel Noorman, "Computing and Moral Responsibility", Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/computing- responsibility/, 2018. 2. 16.자.

이러한 관점에서, 인공지능 책임성이란 인공지능의 의사결정에 의해 사회적 문제가 발생했을 때, 원인이 되는 행위를 한 주체에게 귀속되는 도덕적, 법적 책임을 뜻하는 것으로 볼 수 있다. 이는, 이해당사자 누군가에게 책임이 귀속되어야 하고, 책임공백이 있어서는 안 된다는 '이해관계자의 책무' 개념을 암묵적으로 전제하는 것이다. 영국의 공학, 물리학 연구위원회(Engineering and Physical Science Research Council, EPSRC)가 2010년 발표한 로봇윤리 5원칙에서 '법적 책임의 인간에 대한 귀속'을 명시하고 '안전과 보안'을 특히 강조한 것은 그 일환으로 바라볼 수 있다.[11]

그렇지만 오늘날 인공지능의 기술적인 특징은 이러한 책임 개념과 자연스럽게 조응하지 않는 면이 있어서, 전통적 책임관을 그대로 적용하면 소위 책임격차(responsibility gap)가 발생하여 문제해결이 어려울 수 있다. 예를 들어, 마이크로소프트가 2016년 3월 대화형 인공지능으로 개발한 챗봇 테이(Tay)가 악성 이용자들로부터 학습한 인종차별적 발언을 트윗을 하여 사회적 논란이 발생한 상황을 보자.[12] 이 경우는 테이에게 혐오표현을 학습시킨 이용자를 제재하여 문제를 해결할 수 있을 것이다. 그러나 인공지능 알고리즘의 훈련 데이터가 과거의 편견을 반영한 경우, 문제를 유발한 주체가 불특정 다수인 과거의 인류라면 책임규명이 어려워질 수 있다. 그밖에 무수히 많은 참여자나 데이터가 개입하는 오늘날 소스코드 제작과정을 생각해보면 책임소재를 가리는 일은 사실상 불가능해질 수 있다(이를 'many hands의 문제'라고 한다).[13]

다른 한편, 인공지능 의사결정의 불투명성(opacity)은 새로운 형태의 문제를 낳는 원인이 된다. 여기서 불투명성은 ① 인공지능 보유주체의 지식재산권이나 계약상 특약조항과 같은 '제도적 측면', ② 인공신경망의 복잡성과 같은 알고리즘의 '본질적 측면', ③ 검증과정의 인적, 물적 비용문제로 인한 '현실적 측면' 등 다종다양한 이유로 인해 발생된다.[14] 이러한 불투명성은 오늘날 인공지능 의사결정의 증대된 자율성(autonomy)과 결부되어, 현실적 문제해결 능력이 미약한 인공지능 그 자체에게 책임을 돌려야만 하는 것처럼 여겨질 수 있다. 결과적으로 책임을 귀속할 행위자가 부재하게 되거나 또는 그 정반대로 너무 많아져서 책임귀속이 힘들어질 수 있다.

---

11 Engineering and Physical Science Research Council, "Principles of robotics", http://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/, 2010. 9.자.

12 Peter Lee, "Learning from Tays introduction", https://blogs.microsoft.com/blog/2016/03/25/learning- tays-introduction/, 2016. 3. 25.자.

13 Helen Nissenbaum, "Accountability in a Computerized Society", Science and Engineering Ethics Vol. 2, Issue. 1, 1996, pp. 28-32.

14 고학수 정해빈 박도현, "인공지능과 차별", 『저스티스 통권 제171호』, 2019, 235-236면.

유럽연합이 2012년부터 3년 여 동안의 '로봇법 프로젝트(RoboLaw Project)' 연구를 거쳐 2014년 발표한 로봇규제 가이드라인은 그 대안으로 3가지의 기준 원칙을 제시한다.[15] 첫째, 가장 기술친화적 원칙으로, 로봇산업의 혁신을 촉진하고 규제비용을 절감하기 위하여 기술적으로 불가피한 책임의 경우는 (적어도 단기적으로는) 면책(immunity)을 인정하는 방안이다. 둘째, 일정한 수준의 자율성을 획득한 로봇에 대하여 일종의 법인격(legal personhood)을 부여하는 방안이다. 이러한 방법론은 이후 로봇에게 법적 책임을 부여할 수 있도록 '전자인(electronic person)' 지위를 부여할 필요가 있다는 유럽연합 의회 결의안[16]으로 이어진다. 셋째, 기본권 보호 정도가 가장 높은 방안으로, 민법상 특수불법행위나 제조물책임법과 같은 일부 특별법에 마련되어 있는 무과실책임(strict liability)을 법제화하는 것이다.

3. 투명성의 확보와 사회구성원의 주체성 강화

위와 같은 세 가지 방안은 윤리적 책임(responsibility)의 귀속이 어려워지는 현실적 문제를 적어도 민 형사상의 법적 책임(liability) 측면에서는 어느 정도 해결하였다고 평가할 수 있다. 그런데 최근 들어 인공지능의 자율적 의사결정이 점차 늘어나면서, 사회적으로 새로운 차원의 문제제기가 이루어져왔다. 그것은 앞서 언급한 인공지능의 불투명성으로 인하여 사회구성원의 알 권리와 절차적 참여권이 제대로 보장되지 않는다는 것이었다. 인공지능이 행한 의사결정이 인류의 삶에 중대한 영향력을 행사하고 있음에도 불구하고, 그러한 의사결정이 어떠한 근거로 이루어졌는지, 어떠한 절차를 거쳐서 이의제기를 할 수 있는지에 대한 문제의식은 상대적으로 미약했던 것이다.[17] 예컨대 앞서 본 테이 사례의 당사자의 경우, 민사상 손해배상은 별론으로 하더라도, 사적자치의 원칙에 비추어 기업체에 단지 인공지능의 혐오표현을 근거로 인공지능 사용을 중단하라는 청구권을 행사할 수 있는지는 현행법상 논란이 따른다. 나아가 이와 같은 문제는 향후 법적 책임을 평가함에 있어 '증명(proof)'의 현실적 어려움과도 맞닿아 있다.

'기술영역의 적법절차(technological due process)'로 불리는 일련의 문제제기가 이어지며, 인공지능의 불투명성을 해소하는 규범적 노력이

15 Erica Palmerini et al., "Guidelines on Regulating Robotics", 2014, pp. 23-24.

16 European Parliament, "European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))", 2017. 2. 16.자.

17 프랭크 파스콸레는 이러한 상황을 '블랙박스 사회(black-box society)'라는 말로 표현한다. 프랭크 파스콸레, 이시은 역, 『블랙박스 사회』, 안티고네, 2016.

‘투명성(transparency) 원칙’이라는 명칭으로 대두되었다.[18] 그리고 여기서 파생된 ‘책임성’ 개념이 바로 ‘accountability’이다(이를 한글로 표현할 때, ‘책임(성)’ 또는 ‘책무(성)’이라고도 하고 다른 유형의 책임과 구분하여 ‘설명책임’이라 하기도 한다).[19] 설명책임은 단지 인공지능 의사결정의 효력이 배후의 인간 행위자에게 귀속되는 차원을 넘어, 그런 의사결정이 어떤 과정을 거쳐 이루어졌는지 설명하도록 요구하는 책임을 의미한다. 이를 현실에서 구현하기 위한 구체적인 방법론으로서, 그동안 인공지능 감사(audit), 검사(review), 영향평가(impact assessment)와 같은 공법적 규제부터, 설명책임의 달성 정도에 따라 면책의 수준을 달리 인정하는 것과 같은 사법적 해석론까지 다양한 논의가 이루어지고 있다.

인공지능 규범체계 관점에서 ‘거버넌스(governance)’라는 용어가 주로 사용되는 이유 역시 같은 맥락에서 설명된다. 주로 책임의 부과에 초점이 맞춰진 하향식(top-down)의 규범체계를 넘어 구성원의 능동적 참여를 포용하는 상향식(bottom-up)의 규범체계를 전제하기 때문이다. 그렇게 보면 ‘accountability’는 윤리적, 법적 책임 개념(responsibility, liability)이 담당해오던 역할에 설명책임과 능동적 참여 기능을 더한 광의의 책임성 개념인 것으로 이해할 수 있을 것이다. 그런 점에서, 진입규제나 형사처벌과 같은 전통적 방식의 규제뿐만 아니라, AMA나 소위 ‘설명가능 인공지능(explainable AI, XAI)’ 등을 통해 알고리즘 의사결정 방식을 설명하는 기술적 접근법 또한 얼마든지 책임성 원칙을 달성하는 적절한 수단이 될 수 있다. GDPR에서 설명을 요구할 권리나 잊힐 권리를 정한 이유도 책임성에 대한 광의의 이해방식을 취할 때 보다 분명한 파악이 가능해진다. 이와 병행하여 구성원들의 인공지능 문해력(literacy)을 증진하는 노력도 장기적으로 요구되며, 인공지능 활용에 따른 혜택을 누리고 위험을 초래한 당사자가 비용을 분담할 수 있는 메커니즘을 구축하는 것이 책임성을 구현할 수 있는 한 가지 방편이다. 그렇다면 인공지능 시대의 책임성(accountability) 원칙이란 인공지능의 활용을 인간과 인공지능의 공존공영으로 이끄는 사회규범의 총체적 활용과정으로 재구성할 수 있다. 인공지능의 활용에 제약을 두는 것이 책임성을 강화하는 전형적 형태지만, 경우에 따라 인공지능을 적극적으로 활용하여 책임성을 강화할 수도 있다는 것이다. 인공지능 책임성 원칙은 오늘날 인공지능 규범 논의에서 가장 기본적이고 핵심적인 원칙의 하나로 인정받고 있다.

---

18 고학수 정해빈 박도현, “인공지능과 차별”(주 14), 245면 이하 참조.

19 Robyn Caplan et al., “Algorithmic Accountability: A Primer”, Data&Society, 2018, p. 10.

4. 이중효과의 조화와 맥락 중심의 접근방식

다른 한편, 인공지능의 강한 복잡계(complex system)적 특성은 규범적 평가를 일률적으로 행하기 어렵게 만드는 요인이 된다. 예를 들어, 인공지능 의사결정의 차별적 효과를 검증하는 과정에서, 불투명성을 극복하기 위하여 때로는 문제되는 이해당사자의 개인정보나 민감정보를 수집할 필요가 있는데, 이는 프라이버시라는 또 다른 중대한 기본권과 정면으로 충돌한다.[20] 기본권 사이의 충돌을 넘어, 단일한 기본권에 대해 유발된 효과가 상충할 수 있다. 예를 들어, 인공지능의 활용은 누군가에게는 노동권의 신장을, 누군가에게는 노동권의 박탈을 낳는다.[21] 경쟁, 자율성, 삶의 질(well-being), 민주주의 등에 대한 우려도 같은 맥락에 위치한다.

그렇지만 해악의 위험과 혜택의 가능성이라는 일종의 '이중효과(double effect)'를 딜레마나 역설의 원천으로만 바라볼 것은 아니다. 얼마든지 양자 사이의 발전적 조화를 구상해 볼 수 있다. 예를 들어, 단순히 인공지능이 부가가치를 창출한다는 등의 일반론적인 이익을 넘어, 자율주행차는 장애인이나 노약자의 이동권을 강화하고, 드론은 물리적 접근성이 낮은 지역의 운송에 기여할 수 있는 등 새로운 기술을 통해서 새로이 구현가능하게 된 기능이 존재한다. 다른 한편, 2018년 우리나라에서 논란이 발생한 바 있는, '킬러 로봇'을 포함한 '치명적 자율무기(Lethal Autonomous Weapons, LAWs)'의 연구개발 및 사용과 관련된 이슈에 관해서는 국제사회의 논의에 적극적으로 참여하고 주도해 나가도록 노력해야 할 것이다.[22] 더 넓게는 인공지능의 복잡계적 특성을 고려하여, 해악의 의도가 없어도 예측하지 못한 악결과가 나타날 수 있음을 인지하고 그에 대비하여야 한다. 그렇다고 하여, 이중효과를 근거로 특정 유형의 신기술에 대한 원천적 금지를 규범적 대안으로 제시하는 접근방식은 신중해야 할 필요가 있다.

이처럼 인공지능에 대한 선험적 판단 대신에 의사결정의 구체적 상황에 따라 규범적 판단이 달라져야 한다는 접근방식을 '맥락(context) 중심의 접근방식'이라 부르는데, 이러한 맥락중심의 접근방식은 인공지능 규범논의에서 점차 더욱 강조되고 있다. 인공지능은 자율성을 증진할 수도 억제할 수도, 인류에게 혜택이 될 수도 해가 될 수도 있다는 것을 전제로, 개별 맥락을 고려하여 규범적 또는 정책적 판단을 해야 한다는 것이다.

---

20 고학수 정해빈 박도현, "인공지능과 차별"(주 14), 255면.

21 U.S. Executive Office of the President, "Artificial Intelligence, Automation, and the Economy", 2016, p. 10 이하 참조.

22 전문가들의 공개서한으로, Future of Life Institute, "Autonomous Weapons: An Open Letter from AI & Robotics Researchers", http://futureoflife.org/open-letter-autonomous-weapons/, 2015. 7. 28.자 참조.

앞서 본 유럽연합의 로봇규제 가이드라인은 일반론 대신 사례별 접근을 택하고, 규제와 산업발전이 공존 가능함을 전제로 하며, 기술을 바탕으로 한 자율규제, 윤리원칙, 엄격한 법제를 상호 조화한 프레임워크를 제시함으로써 이러한 사고를 반영하였다.[23]

지금까지 인공지능 윤리규범의 변천사를 추적해보면서 주요한 쟁점을 살펴보았다. 오늘날의 윤리규범은 큰 틀에서 이와 같은 이해방식을 공유하고 그에 기초하고 있다. 그러나 영역이나 지역 등의 기준 하에 비교해보면, 강조하는 측면이나 구속력의 정도 등에 있어서는 어느 정도 차이가 존재한다. 국제사회에서 이루어지는 논의의 공통점과 차이점을 분명히 파악하였을 때, 비로소 우리나라의 실정에 맞는 특유한 규범을 확립할 수 있다는 측면에서 최신규범의 내용을 고찰할 필요가 특히 강하다. 또한 국가적인 차원에서는 해외의 논의에 적극 참여하고 주도해야 할 필요도 있다. 이런 필요성에 비추어, 이하에서는 해외의 최신 윤리규범 내용을 비교 분석해보도록 한다.

3. 해외의 최신 인공지능 윤리규범 비교 분석

1. 개괄

아래에서는 해외에서의 인공지능 윤리이슈에 대한 주요 논의를 검토하면서 그 시사점을 찾아보고자 한다. 인공지능 기술이 상용화되고 고도화되면서, 미국과 유럽연합을 중심으로 인공지능 윤리이슈에 대한 많은 논의가 이루어지고 있다. 인공지능 기술을 서비스나 플랫폼에 활용하고자 하는 IT 회사들 역시 연구개발과 서비스에 있어 인공지능의 윤리원칙 등 자체적 규범을 발표하는 곳들이 나타나고 있다.

그러나 윤리규범의 구체적 내용을 살펴보면, 다루고 있는 이슈의 범위나 정도에서 적지 않은 차이가 있다. 인공지능을 연구개발하고 활용하는 데 적용되는 일반적이고 기본적인 원칙을 천명한 수준인 경우도 있고, 여기서 더 나아가 주로 인공지능이 이용되고 있거나 가까운 시일 이내에 이용될 것으로 예측되는 분야에서의 주요한 윤리적 이슈까지 함께 다룬 경우도 있다. 또한, 인공지능 규범을 어떻게 정립해야 할지에 대하여 비교적 구체적인 방법론을 언급한 경우도 있다. 이 글에서는 다루어지는 내용의 범위에 따라, ① 기본원칙을 중심으로 하는 윤리규범 유형, ② 기본원칙과 주요 이슈를 함께 (흔히 보고서의 형태로) 다룬 윤리규범 유형, ③ 윤리규범의 정립방안에 대한 구체적 방법론을 함께 언급한 유형으로 나누어

23 Erica Palmerini et al., “Guidelines on Regulating Robotics”(주 15), p. 8 이하; 이원태, “유럽연합(EU)의 로봇법(RoboLaw) 프로젝트”, 『KISO Journal Vol. 23』, 2016, 29-32면 참조.

사례를 제시하고자 한다. 매우 다양한 형태의 다양한 문건들이 제시된 바 있는데, 이 중 일부를 선별하여 소개한다. 주요 문건들의 목록은 별도로 정리하여 부록으로 포함하였다.

2. 기본원칙 중심의 윤리규범

기본원칙 중심의 윤리규범은 주로 인공지능 시스템을 설계하고 개발할 때 준수하여야 하는 일반적 사항을 담고 있다. 이런 윤리규범의 예시로, OECD 보고서, 일본 총무성 가이드라인, 아실로마 원칙 및 구글, 마이크로소프트 등 사기업에서 발표한 규범을 소개한다.

1) OECD AIGO, “Draft Recommendation of the Council on Artificial Intelligence” (2019년 5월 최종안 제출 예정)

OECD는 2018년 9월 인공지능 원칙의 방향성을 제시하는 전문가 그룹인 ‘AIGO(AI Expert Group at OECD, 이하 “AIGO”)’를 만들었다. AIGO는 정부, 기업, 노동자, 대중에게 인공지능의 이익을 극대화하고 위험은 최소화할 수 있도록 하는 작업을 하고 있고, 그 일환으로 인공지능 가이드라인의 초안을 마련하였다.[24] 2019년 5월경 개최될 예정인 각료이사회(Ministerial Council Meeting)에 현재 초안 상태인 인공지능 가이드라인 권고안이 제출되어 확정될 예정이다.[25] 권고안에는, 신뢰할 수 있는 인공지능을 만들어 이를 책임 있게 관리하기 위해 필요한 기본원칙과, 신뢰할 수 있는 인공지능을 만들기 위해 필요한 국가정책, 국제협력의 방향 등에 대한 내용이 담겨있다.

권고안 초안을 보면, 신뢰할 수 있는 인공지능을 만들어 책임 있게 관리하기 위한 원칙으로 특정 집단을 배제하지 않는(inclusive) 포용성, 지속가능한 발전과 삶의 질(well-being), 인간 중심의 가치, 공정성, 투명성 및 설명가능성, 견고함 및 안전 등의 개념이 강조된다. 그리고 인공지능에 대한 국가정책 및 국제협력의 측면에서, 인공지능 연구개발 투자와 디지털 생태계의 조성, 인공지능 혁신이 이루어질 수 있는 정책적 환경을 조성하기 위한 개별국가의 노력 및 국제사회의 협력을 촉구한다.

이와 같은 제언을 바탕으로, AIGO는 이를 실행할 수 있도록 늦어도 2019년 12월 말까지 OECD 디지털 경제정책위원회에 실무 가이드라인을

24 OECD, “OECD creates expert group to foster trust in artificial intelligence”, http://www.oecd.org/going-digital/ai/oecd-creates-expert-group-to-foster-trust-in-artificial-intelligence.htm, 2018. 9. 13.자.

25 OECD, “OECD moves forward on developing guidelines for artificial intelligence (AI)”, http://www.oecd.org/going-digital/ai/oecd-moves-forward-on-developing-guidelines-for-artificial-intelligence.htm, 2019. 2. 20.자.

제출하고, 다양한 이해관계자를 확보하고 이들이 학제간 대화를 할 수 있도록 인공지능 정책에 대한 정보를 교환하는 포럼을 만들며, 제안사항에 대한 실행을 모니터링하고 이 제안을 채택한 날로부터 5년 이내에 상황을 보고하도록 권고하는 내용을 담았다. 이처럼 OECD 권고안은 권고안의 제시에 그치지 않고, 그 이행을 담보하기 위한 후속조치까지 마련하고 있다. 그리고 이 권고안은 다양한 OECD 회원국은 물론 여러 이해당사자 사이의 협의를 반영하는 것이라는 특징이 있다.

2) 일본 총무성, “Draft AI R&D Guidelines for International Discussions” (2017년 7월)[26]

일본 총무성은 2017년 7월 인공지능 활용을 통한 이익을 증대하고 위험을 통제하기 위한 목적으로 여러 나라 이해관계자들 사이에 구속력 없는 연성법 형태의 가이드라인을 만드는 데에 도움을 얻기 위하여 “국제적 논의를 위한 인공지능 연구개발 관련 가이드라인 초안(Draft AI R&D Guidelines for International Discussions, 이하 ”가이드라인”)”을 발표하였다. 가이드라인은 인공지능 이용자들의 이익을 보호하고 위험의 확산을 억제하여 인간 중심의 ‘지식네트워크 사회(知連社會)’를 만드는 것을 목표로 제시했다. 가이드라인은 논의의 대상을 현재 사용되고 있는 좁은 인공지능(Narrow AI)으로 상정하고, 자율적 인공지능(autonomous AI) 혹은 일반 인공지능(Artificial General Intelligence, AGI)은 논의 대상에서 제외하였다.

가이드라인은 필요할 경우 계속적으로 가이드라인을 검토하고 유연하게 수정할 것을 전제로 인공지능 연구개발 시 준수해야 할 9가지 원칙을 발표했다. 우선, ① 인공지능 시스템의 상호연결성 및 상호운용성을 보장하도록 하는 ‘협력의 원칙’, 그리고 ② 인공지능 시스템의 입출력에 대한 검증가능성과 의사결정에 대한 설명가능성을 보장하도록 하는 ‘투명성 원칙’이 제시되었다.

다음으로, ③ 사전적으로 인공지능의 행위를 검증하고 유효성을 확인하기 위하여 노력하며, 연구실이나 샌드박스 등 닫힌 공간에서 인공지능 시스템을 실험해 보도록 하는 ‘통제가능성의 원칙’이 있다. 여기에는 사람이나 다른 신뢰 가능한 인공지능 시스템이 인공지능을 모니터링 하거나 경고하는 형태로 감시하고 셧다운, 네트워크 차단, 보수 등의 반대조치를 취하는 것이 효율적인지 여부를 검토하여야 한다는 내용도 포함된다. 같은

26 The Conference toward AI Network Society, “Draft AI R&D Guidelines for International Discussions”, www.soumu.go.jp/main_content/000507517.pdf, 2017. 7. 28.자; 이듬해인 2018년에는 AI의 이용(utilization) 측면에서 비슷한 내용의 원칙이 발표되었다. The Conference toward AI Network Society, “Draft AI Utilization Principles”, www.soumu.go.jp/main_content/000581310.pdf, 2018. 7. 17.자.

맥락에서 ④ 인공지능 시스템이 이용자 또는 제3자의 생명, 신체, 재산을 해하지 않도록 하여야 한다는 '안전의 원칙', 그리고 ⑤ 인공지능 시스템이 의도한 대로 작동하고 승인받지 않은 제3자를 위해 이용되지 않게 하는 안정성과 물리적 공격과 사고에 견딜 수 있도록 하는 견고함을 갖추고 기밀성과 진실성 등을 보장하여야 한다는 '보안의 원칙'이 존재한다.

또한 ⑥ 사전에 프라이버시 침해의 위험을 평가하고 프라이버시 영향평가를 진행하도록 하는 '프라이버시의 원칙', ⑦ 인공지능 시스템 연구개발에 있어 인간의 존엄성 및 개인의 자율성을 존중하고 국제인권법 및 국제인도법에 기반한 인간성의 가치를 침해하지 않도록 하는 '윤리의 원칙', ⑧ 이용자가 인공지능 시스템의 도움을 받아 의사결정을 할 때 시의적절하게 인터페이스를 제공하는 등 이용자가 적절하게 이용할 수 있도록 하는 '이용자 지원의 원칙', 그리고 ⑨ 인공지능 시스템을 활용할지 여부를 결정할 수 있도록 이용자에게 기술적 특성에 대해 정보와 설명을 제공하며 다양한 이해관계자들과 대화를 통해 의견을 듣는 방식으로 이해관계자들의 참여를 유도하도록 하는 '책임의 원칙'이 있다.

3) Future of Life Institute, "Asilomar AI Principles"(2017년 1월)

민간영역에서도 비정부기구, 연구자, 기술 전문가 등을 포함한 여러 관련당사자들 사이에 인공지능이 제기하는 도전과제에 맞서 책임 있는 인공지능(responsible AI)을 어떻게 구현할 것인지 논의하기 위한 움직임이 계속되었다. 2015년에는 치명적 자율무기(LAWs)가 세계적으로 이슈화되어, 이에 관한 문제제기를 하는 "인공지능에 대한 공개서한(Open Letter on Artificial Intelligence)"에 일론 머스크, 빌 게이츠, 스티븐 호킹 등 유명인사 및 8,000명 이상의 인공지능 연구자들이 서명하기도 했다. 이러한 민간사회의 움직임이 계속되어 미국 캘리포니아 아실로마에서는 2017. 1. 6.부터 3일 동안 'Beneficial AI 2017' 컨퍼런스가 개최되어 인공지능 윤리규범이 논의되었고, 비영리 연구단체인 '삶의 미래 연구소(Future of Life Institute)'가 2017. 1. 17. '아실로마 원칙(Asilomar Principles)'이라는 이름으로 이를 정리하여 발표하였다.[27] 아실로마 원칙은 총 23가지의 사항으로 구성되어 있다. 먼저 유익한 인공지능 연구에 재정지원이 이루어지도록 하여야 하고, 과학-정책 간의 연계를 통한 교류가 있어야 하며, 협력 및 신뢰, 투명성을 중심으로 하는 연구문화를 갖도록 해야 한다는 원칙을 주장한다. 또한 인공지능 시스템의 설계 및 개발 시의 윤리에 대해 안전을 보장하고 투명성을 담보할 것, 인간과 인공지능의 가치가 일치하여야 할 것, 인공지능 시스템에 결정을 위임할지 여부를 사람이 선택할 수 있어야

27 양희태, "인공지능의 위험성에 대한 우려로 제정된 아실로마 인공지능 원칙", 과학기술정책 제27권 제8호, 2017, 4면.

한다는 '통제의 원칙' 등을 담고 있다. 장기적인 이슈로서 인간의 지성을 초월하는 소위 '초지능(super intelligence)'이 등장했을 때를 대비해서 위험을 관리할 수 있도록 하는 원칙에 대해서도 언급되었다.[28]

4) 민간기업에서 발표한 가이드라인

인공지능 기술을 활용하여 다양한 서비스를 제공하고 있는 몇몇 기업들도 인공지능과 관련된 원칙을 자체적으로 발표하였다. 대표적으로 마이크로소프트는 앞서 언급한 테이 사건이 발생한 이후 몇 개의 가이드라인을 마련하여 발표하였다. 2018년 1월 "The Future Computed" 책자를 발간하면서 인공지능에 관한 원칙을 포함하기도 했고, 개별 상품별로, 2018년 10월에는 대화형 인공지능(챗봇) 개발자를 대상으로 한 10개 가이드라인을 발표한 뒤,[29] 2018년 12월에는 안면인식 기술의 개발과 활용에 관한 6가지 원칙을 발표하였다.[30] 그 내용으로는 투명성이나 책임의 원칙과 같이 인공지능을 활용함에 있어 기본적으로 준수하여야 하는 원칙이 포함되었다. 특히 대화형 인공지능(챗봇)의 경우 테이 사례를 고려해 인공지능의 능력을 넘어 사람의 판단이 필요한 경우에는 사람이 통제할 수 있도록 하고, 문화적인 규범을 존중하도록 설계하며, 키워드 필터링 방식을 이용해서 인공지능이 이용자의 공격적인 입력에 적절히 대응하도록 하고, 인공지능이 사람을 공정하게 대할 것을 규정했다. 안면인식 인공지능의 경우 인간에 대한 감시의 도구로 악용될 우려가 있다는 점을 고려해서 최소한 정보주체에게 동의를 받고 인공지능을 사용할 것을 권고하였다. 그리고 법집행기관이 안면인식 기술의 사용을 요청한 경우는, 공적인 공간에서 기술을 사용할 수 있는 기준과 한계를 정한 법률이 있고, 법원이 집행기관에 기술의 사용을 허가하였으며, 생명이나 신체에 대한 긴급한 위험이 있는 등 반드시 필요한 경우에 정부의 정당한 이익(legitimate interests)과 개인의 자유(civil liberties) 및 프라이버시의 보호 사이에서의 이익형량을 전제로 하여 제한적으로만 사용할 수 있도록 규정하였다. 원칙을 제시한 수준의 간단한 문건이지만, 이와 같은 문건은 개별 상황에 따라서는 매우 구체적인 현실적인 함의를 가질 수도 있다. 예를 들어, 만일 안면인식 기술의 활용에 있어 동의가 필수적으로 요구된다면, 이 기술의 현실적인 활용가능성은 상당히 제한될 수밖에 없다.

28 Future of Life Institute, "Asilomar AI Principles", https://futureoflife.org/ai-principles/?cn-reloaded= 1, 2017. 1. 17.자.

29 Microsoft, "Responsible bots: 10 guidelines for developers of conversational AI", http://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf, 2018. 11. 4.자.

30 Microsoft, "Six Principles for Developing and Deploying Facial Recognition Technology", https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2018/12/MSFT-Principles-on-Facial-Recognition.pdf, 2018. 12.자.

구글도 2018년 6월 인공지능에 대한 원칙을 발표하였다. 구글의 인공지능이 사회적으로 이익이 되고, 불공정한 편견을 만들어 내거나 강화하지 않아야 한다는 기본원칙 등이 포함된 것이다.[31] 그리고 이 윤리원칙에 대한 후속 작업으로, 2018년 12월 책임성 있는 인공지능 실무관행(Responsible AI Practices)을 발표하였다.[32] 구글은 이러한 원칙을 실제로 거버넌스 구조에 적용하고자, 2019년 3월 외부 인사들로 이루어진 자문위원회(Advanced Technology External Advisory Council)를 구성했다. 다만, 현재 자문위원회는 일부 위원의 탈퇴 및 자격논란 등으로 인해 문제에 봉착한 상황이다.[33]

3. 기본원칙 및 세부분야에서의 이슈를 제기하는 윤리규범 유형

인공지능을 설계하고 개발할 때 준수하여야 할 기본원칙을 제시하면서, 동시에 인공지능이 사회에 미치는 영향을 고려할 때 발생할 수 있는 이슈를 찾아 분석하고 해결방안을 검토하는 분석보고서의 성격을 포괄하는 윤리규범 유형도 있다. 이 유형에 해당하는 사례로, IEEE 보고서, 영국 상원 보고서 및 UNGP & IAPP의 보고서에 대해 살펴보기로 한다.

1) IEEE, “Ethically Aligned Design” (2016년 12월, 2017년 12월, 2019년 예정)[34]

전기전자기술자협회(Institute of Electrical and Electronics Engineers, 이하 “IEEE”)는 세계적인 규모의 기술 전문가 집단인 동시에 중요한 국제표준화기구이기도 하다. 인공지능 및 자율시스템의 윤리적 고려사항에 대한 IEEE 글로벌 이니셔티브에서는 2016년 12월과 2017년 12월 두 차례에 걸쳐서 “Ethically Aligned Design”이라는 이름의 보고서를 발간했고, 이에 대한 의견 수렴을 거쳐 2019년 최종 버전을 발간할 예정에 있다. 이하에서는 2016년 보고서의 내용에 감정 컴퓨팅(affective computing), 섹스 로봇, 시스템 조작, 넛징(nudging) 등의 새로운 영역의

31 Sundar Pichai, “AI at Google: our principles”, https://www.blog.google/technology/ai/ai-principles/ 2018. 6. 7.자.

32 Google AI, “Responsible AI Practices”, https://ai.google/education/responsible-ai-practices, 2018. 12.자.

33 Kent Walker, “An external advisory council to help advance the responsible development of AI”, https://www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/, 2019. 3. 26.자.

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

문제를 추가하여 발표한 2017년 보고서에 담긴 내용을 간략히 살펴본다.

이 보고서에서는 인공지능을 설계할 때, 인간이 얻을 수 있는 경제적 이익뿐만 아니라, 인공지능이 사람의 감정적, 정서적 측면에서의 삶의 질에 미치는 영향까지 고려해서 인공지능 시스템을 설계해야 한다는 점을 강조하고 있다. 더 나아가, 인공지능 기술이 더욱 발전한 상황, 즉 자율무기 시스템이 매우 고도화된 상황, 스스로 판단할 수 있는 일반 인공지능(Artificial General Intelligence)이나 초인공지능(Artificial Super Intelligence)이 등장한 상황, 사람의 감정 영역까지 관장할 수 있는 컴퓨팅 기술이 등장한 상황, 가상현실 기술 등의 발전으로 현실과 가상세계의 구분이 되지 않는 상황 등에서 대두될 윤리적인 문제에 대해서도 함께 언급하고 있다.

보고서는 인공지능을 설계, 개발, 실행할 때 적용되어야 할 5가지의 기본원칙을 제시하는데, ① 인공지능이 인권, 자유, 인간의 존엄성 및 문화의 다양성을 존중하는 방향으로 설계되고 운영되어야 한다는 원칙, ② 인공지능이 인간의 삶의 질에 어떤 영향을 미칠지 고려하여 삶의 질을 증진하는 방향으로 개발되고 이용되어야 한다는 원칙, ③ 설계자와 이용자가 인공지능의 결정이나 행위에 대한 책임을 지도록 보장해야 하고 이를 위하여 인공지능에 대한 등록 시스템과 기록보존 시스템을 만들어야 한다는 원칙, ④ 인공지능의 알고리즘과 의사결정 기준을 투명하게 알 수 있도록 하고 투명성에 대한 객관적 평가가 가능하도록 하여야 한다는 원칙, ⑤ 인공지능 시스템이 오용될 가능성이 있다는 것을 의식하고 그 위험을 최소화하도록 하여야 한다는 원칙이 바로 그것이다.

그리고 이 보고서에서는 위 원칙을 준수함에 있어 함께 고려되어야 할 중요한 영역으로 개인정보 분야를 언급하면서 개인정보에 대한 정보주체의 권리를 존중하고 개인정보에 대한 접근을 정보주체가 통제할 수 있어야 한다고 언급하였다. 또한 디지털 세계의 인격(digital personas)이 개인의 인격과 달라질 수 있음을 인지하고, 디지털 세계의 인격도 통제할 수 있도록 해야 하고, 마이데이터 논의에서 볼 수 있는 것과 같이 개인정보의 주체가 자신의 정보를 관리할 수 있는 방향으로 기술발전이 이루어져야 할 것으로 언급되었다. 또한 정보주체의 권리를 보장하려면 인공지능 시스템이 개인의 데이터를 수집하는 경우 데이터가 수집되었다는 사실과 수집의 목적이 무엇인지에 대해 정보주체에게 통지를 해야 하고, 다른 한편 정당하게 주어진 권한을 넘어 데이터에 접근이 이루어질 수도 있는 우회로(backdoor)는 존재하지 않아야 한다고 규정하였다.

그 이외에 이 보고서는 인공지능이 인간에게 피해를 줄 경우의 책임 분배의 문제를 인공지능의 법적 지위와 관련하여 다루고 있다. 책임 문제를 다루는 두 가지 프레임워크로 재산법의 원칙에 따라 보는 방안과 인공지능에

위임되지 않아야 하는 결정을 선별하고 그에 대한 인간의 통제를 보장하는 방안을 제시한다. 보고서는 현재 단계에서 인공지능에 독립적인 법인격(legal personhood)을 부여하는 것에 대하여 부정적인 입장을 취한다. 그 이유로, 인공지능에 독립적인 법인격을 부여하는 경우 인공지능의 의사결정에 대한 사람의 책임을 제한하거나 없앨 수 있고, 인공지능을 설계, 개발, 이용할 때 인공지능의 안전을 보장하려는 인센티브를 낮출 가능성 등이 제시되었다. 그에 따라, 이 보고서는 인공지능의 책임은 현재로서는 재산법 영역의 이슈로 다루어야 한다고 하고, 입법자들과 법집행자들은 인공지능을 이용한다는 것이 손해배상과 같은 법적 책임을 피할 수 있는 수단으로 이용되지 않도록 고려할 필요가 있다고 한다. 그리고 법적 책임 가능성에 따른 배상책임 등 경제적 책임의 이행을 보장하기 위한 보험 등의 제도를 갖출 필요성이 있을 것임을 언급한다.

보고서는 인공지능 시스템의 투명성을 보장하기 위한 구체적인 방안으로서, 당사자, 변호사와 법원으로 하여금 정부 또는 다른 국가기관이 채택하는 알고리즘과 데이터에 대해서 합리적인 수준에서 접근할 수 있도록 허용하는 '접근가능성'을 강조한다. 또한, 인공지능 시스템에 내재된 논리와 규칙에 대하여, 관련된 리스크를 평가하고 엄격한 테스트 및 감시가 가능하도록 해야 하며, 인공지능 시스템이 내린 결정 및 해당 결정의 근거를 기록하는 감사추적을 통해 제3자에 의한 검증이 가능하게 해야 한다며 '감시 및 검증가능성'을 강조한다. 한편, 대중은 알 권리 차원에서 인공지능 시스템에 대한 윤리적인 결정을 내리거나 지원을 하는 주체가 누구인지를 알 수 있어야 한다고 본다.

보고서는 교육정책 및 의식의 제고가 필요하다는 점도 언급하면서 인공지능 기술의 발전이 사회적으로 어떤 영향을 미치는지에 대하여 대중에게 교육을 할 필요가 있으며, 관련 분야의 전문가를 충분히 양성하여야 한다고 한다. 끝으로 보고서는 인공지능 윤리의 이론적 바탕으로 전통적 윤리를 기반으로 두면서도 인공지능이 인류의 삶의 질에 미칠 영향도 동시에 고려해야 한다고 본다. 구체적으로 자율적인 인공지능 시스템에 어떤 가치가 포함되어야 할지에 관한 '가치에 기반을 둔 설계 방법론(value-based design methodologies)'을 도입할 것을 제시한다.

2) 영국 상원, "AI in the UK: ready, willing and able?" (2018. 4. 16.)[35]

영국 상원의 인공지능 특별위원회에서는 2018년 4월 영국이 인공지능 관련 산업의 발전을 선도할 수 있는 토대를 만들고 이를 위해 검토하고

35 UK House of Lords, "AI in the UK: ready, willing and able?", 2018.

준비하여야 할 사항이 어떤 것인지에 대하여 영국 정부에 권고하는 내용을 담아 “AI in the UK: ready, willing and able?”이라는 보고서를 발간했다. 본 보고서에서는 인공지능을 설계하고 개발할 때 준수해야 할 원칙뿐만 아니라, 정부가 인공지능 산업에 어떻게 관여하고 규제해야 할지, 사회 전반적으로 인공지능과 함께 살아가기 위해서는 어떻게 해야 하는지, 인공지능의 위험을 줄이려면 어떻게 해야 하는지 등과 같이 여러 측면에 관한 이슈를 검토하고 있다.

본 보고서는 인공지능이 준수해야 할 중요한 5가지 원칙을 제시한다. ① 인류 공동의 선과 이익을 위해 인공지능을 개발하여야 한다는 원칙, ② 인공지능은 이해가능성(intelligibility)과 공정성(fairness)의 원칙하에 작동하여야 한다는 원칙, ③ 인공지능이 개인, 가족 및 공동체의 개인정보 권리를 줄이거나 프라이버시를 침해하는 방향으로 이용되지 말아야 한다는 원칙, ④ 모든 시민이 인공지능과 함께 정신 감정 경제적 번영을 누릴 수 있도록 교육을 받을 권리를 가진다는 원칙, ⑤ 인공지능에 사람을 해치거나 파괴하거나 속일 수 있도록 자율적인 권한이 부여되어서는 안 된다는 원칙이 그것이다.

또한, 이 보고서에서는 인공지능을 설계하는 과정에서 데이터의 독점과 편견 문제를 비교적 자세히 언급하고 있다. 인공지능 기술이 발전하려면 데이터에 대한 접근과 통제가 필수적인데 큰 IT 기업은 스스로 확보한 데이터의 양도 많고 구매하기도 용이한 반면, 중소기업(SME)의 경우에는 많은 양질의 데이터세트에 접근하기가 쉽지 않을 것이라고 지적하고 있다. 이를 고려하여, 잠재적인 데이터 독점 문제를 적극적으로 검토해야 하는 한편, 정부에서는 공공기관이 가지고 있는 데이터를 인공지능 연구자와 개발자들에게 제공하고, 개인정보를 안전하게 공유할 수 있는 프레임워크를 수립해야 할 필요가 있음을 강조한다. 이 보고서는 또한 편견이 있는 데이터를 학습한 인공지능의 위험에 대하여 언급하면서, 데이터가 다양한 인구집단을 대표하도록 하고 데이터 학습을 통해 사회적인 불평등이 고착화되지 않도록 하기 위해 적절한 조치가 필요하다고 한다. 이를 위해 인공지능 알고리즘에 대해서 감사하고 테스트하는 시스템을 만들어야 한다고 주장한다.

한편, 보고서에서는 인공지능의 사용이 활성화 되어 있는 헬스케어 분야에서 중소기업이 국가의 보건의료 시스템 데이터에 접근할 수 있도록 해야 한다는 점을 강조하고, 이를 위하여 데이터를 적절히 익명화해서 환자의 데이터를 공유할 수 있는 프레임워크를 만들 것을 정부에 권고하고 있다.

또한 보고서는 인공지능이 노동시장에 미칠 영향에 대해서도 주목하면서, 영국에서 인공지능의 발전에 따라 노동시장에 미치는 영향에 대해 계속 분석하고 평가하여야 하며 이에 대한 대책을 세워야 한다고

강조한다. 뿐만 아니라, 사회 전반적으로 인공지능이 미칠 영향을 고려해 디지털 및 데이터 문해력(literacy) 교육을 강화하고, 모든 사람이 인공지능이 제공하는 기회에 접근할 수 있도록 하며, 인공지능에 의해 초래될 수 있는 사회적, 지역적 불평등을 해결하는 방안을 만들어야 한다고 주장한다.

인공지능의 사회적 위험과 관련해서 이 보고서가 인공지능의 행위에 대해 어떤 법적 책임을 부여할 것인지를 직접 언급하고 있지는 않다. 다만, 보고서에서는 영국 법사위원회로 하여금 현재의 법률이 인공지능의 법적 책임을 언급하기에 적절한지 여부를 검토하도록 하고, 필요한 경우 정부가 이에 대한 적절한 구제책을 제공하도록 권고한다. 보고서는 인공지능에만 특화된 규제를 도입하는 것은 현 단계에서 적절치 않다고 보고 인공지능이 활용되는 개별 영역 별로 규제자들이 추가적인 규제가 있을 경우의 영향에 대해서 충분히 검토할 것을 권고하고, 다만, 인공지능을 개발하고 이용하는 공적, 사적 기관에서 활용할 수 있는 여러 영역에 교차적으로 적용될 수 있는 윤리적 행위 준칙은 필요할 것으로 보고 있다.

3) UNGP & IAPP, “Building ethics into privacy frameworks for big data and AI” (2018년 10월)[36]

UNGP(UN Global Pulse)와 IAPP(International Association of Privacy Professionals)은 2017년 5월 “견고한 데이터 프라이버시와 윤리 프로그램의 구축: 이론에서 실제로(Building a strong data privacy and ethics program: from theory to practice)” 포럼을 개최했다. 포럼에서 논의된 내용을 바탕으로 조직에서 데이터 윤리를 실행할 수 있는 방법론과 빅데이터 분석 분야에서 데이터 윤리 및 프라이버시 모범사례를 마련하는 데 중점을 두고 2018년 10월 “빅데이터와 인공지능 관련 프라이버시 프레임워크에 윤리를 구축하기(Building ethics into privacy frameworks for big data and AI)”라는 보고서를 출간했다. 이 보고서는 프라이버시 분야의 전문가들을 대표하는 민간의 협회가 UN 기구와 협력하여 인공지능 시스템 개발에 필수적인 데이터 측면을 중심으로 인공지능 윤리를 논의한 결과물을 만들어 냈다는 점에 일차적인 의의가 있다.

이 보고서는 빅데이터와 관련한 주요 윤리적 문제로 사이버보안, 인권 및 프라이버시를 들고, 이러한 문제로 인한 부작용을 최소화하기 위해서는 데이터 수집과 프로젝트 개발 단계에서부터 빅데이터 활용과 관련된 리스크와 프라이버시 및 윤리 문제를 적극적으로 고려해야 한다고 강조한다. 이 보고서에서는 인권 차원에서 프라이버시 권리를 다루는 것이 필요하다고 하면서, 빅데이터나 인공지능의 잘못된 사용(misuse)도 문제일 수 있지만

36 United Nations Global Pulse International Association of Privacy Professionals, “Building ethics into privacy frameworks for big data and AI”, 2018.

이를 사용하지 않음에 따라(missed use) 인권에 미치는 영향도 있을 수 있음을 고려해서 균형 있게 판단해야 한다고 본다. 그리고 이 보고서는 데이터에 대한 분석과 의사결정을 통해 새로운 방식으로 개인에게 영향이 나타날 수 있기 때문에 성문화된 데이터 윤리가 필요하다는 점을 지적하고, 전반적으로 데이터 활용이 가져오는 사회적 가치의 관점에서 윤리 이슈를 볼 필요가 있다고 한다.

그리고 이 보고서에서는 데이터 윤리를 실행하는 방법으로 '내부적인' 프레임워크와 '외부적인' 프레임워크 모두가 필요하다고 본다. 데이터 윤리에 입각해서 운영을 하려면 책임과 투명성이 특히 중요하므로 조직 내부에서 프라이버시와 윤리 프레임워크를 만들고, 데이터 윤리에 대한 리더십을 구축하며, 회사나 조직 차원의 윤리적인 접근에 대하여 지속적으로 평가할 수 있는 데이터 영향평가나 리스크평가 방법을 구축하는 것이 중요하다고 강조한다.

4. 기본원칙과 이슈 제시를 넘어 구체적 논의를 담은 윤리규범 유형

인공지능 설계와 개발에 관한 기본원칙과 이와 관련된 세부적 이슈를 언급하는 것에서부터 한 걸음 더 나아가, 인공지능 관련 사회규범을 어떻게 형성할지에 대해 구체적으로 논의하는 유형의 문건도 있다. 그 대표적 사례로, 유럽의회의 로보틱스에 관한 결의안과 EU 집행위원회 고위급 전문가 그룹의 인공지능 윤리 가이드라인에 관해 살펴본다.

1) 유럽의회, "로보틱스에 관한 민사법적 규율에 관한 위원회 권고 결의안" (2017년 2월)[37]

인공지능 알고리즘이 발달하면서 주로 로봇의 행위에 대하여 손해배상과 같은 법적 책임과 관련된 민사법적 문제가 대두되자 유럽의회(European Parliament)는 2017. 2. 16. 집행위원회(Commission)에 대한 권고의 내용을 담아 "로보틱스에 관한 민사법적 규율에 관한 권고안"을 결의하였다. 이 결의안은 기존에 논의된 바 있는 로봇윤리 로드맵과 로봇윤리 가이드라인 등을 바탕으로, 유럽의회에서 로봇의 행위에 대한 책임과 법인격 부여와 같은 민사법적 이슈를 구체화하여, 이에 대해 집행위원회로 하여금 후속 조치를 취하도록 권고하는 형태를 띤 것이다. 결의에서는 로봇의 사회적, 의학적, 생명윤리적인 영향을 고려해서 현재 유럽연합의 법적인 프레임워크를 윤리적 원칙에 따라 업데이트하고

37 European Parliament, "European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))", 2017. 2. 16.자.

보완하여야 할 필요성이 있음을 언급한다. 또한, 투명성의 원칙을 강조해서 로봇(advanced robots)에 블랙박스를 설치해서 해당 로봇이 수행하는 모든 거래(transaction)에 대한 데이터를 기록하고, 의사결정에 이르기까지의 로직을 기록하도록 해야 한다고 본다. 그리고 인공지능 기술을 발전시키기 위해서는 인공지능 시스템 사이의 상호운용성(interoperability)이 핵심일 것으로 보고, EU 집행위원회로 하여금 폐쇄적인 전유 시스템(proprietary system)을 피하고 개방된 기준과 혁신적인 라이센싱 모델에 기초한 개방된 환경을 조성할 것을 요청한다.

유럽의회는 로봇이 자율적으로 의사결정을 할 수 있게 되면, 로봇이 야기한 피해에 대하여 책임 있는 당사자를 찾아 손해에 대한 책임을 지우는 것이 어려울 수 있고, 그 경우 전통적 규율이나 계약법적 책임 방식으로는 문제를 해결하기 어려울 수도 있다는 한계를 인식하였다. 이에 EU 집행위원회로 하여금 이 분야에 대한 사회적 논의를 시작하고 별도 기구를 지정하여 이 문제에 대한 전문성을 확보할 것을 권고하였다. 유럽의회는 로봇의 행위에 의한 법적 책임 문제를 다루기 위한 방안 중 하나로, 로봇에 대한 의무적인 책임보험이 역할을 할 수도 있을 것이라 하고, 또한 보험이 커버되지 않는 범위의 손해에 대비하기 위해 보상 목적의 펀드를 별도로 만들 수도 있을 것임을 언급한다. 그리고 의무 책임보험제도에 대한 전제로, 로봇을 관리기관에 등록하도록 하는 '스마트로봇 등록제도'를 만들어 로봇의 등록번호를 통하여 로봇과 보험 또는 펀드 간의 연결 관계를 확인할 수 있는 시스템 구축을 고려하도록 했다. 장기적으로는 로봇과 관련하여 전자인, 전자적 인격 등의 특별한 법적인 지위를 만들 것인지에 대해서 별도로 고려하도록 권고하고 있다.

결의안 첨부(Annex)에서는 로보틱스 헌장(Charter on Robotics)을 제시하여 집행위원회가 이를 고려하도록 하였다. 로보틱스 헌장은, 로봇 엔지니어를 위한 윤리적 행위 준칙(Code of Ethical Conduct for Robotics Engineers)과 연구 윤리위원회 준칙(Code for Research Ethics Committees)를 포함하고 있다. 로봇 엔지니어들에게는, 기본권을 존중하고 책임 있는 태도를 가지며 프라이버시를 존중하고 연구에 리스크가 따를 경우 리스크에 대한 평가 및 관리 프로토콜을 마련할 것을 준칙으로 언급한다. 연구 윤리위원회와 관련해서는, 독립성(independence), 전문성(competence), 투명성 및 책임성(transparency and accountability)을 대전제로 하여, 다양한 배경(multidisciplinary)을 가진 위원으로 구성될 필요가 있음을 언급하고 있다.

2) EU 집행위원회 인공지능 고위급 전문가 그룹(HLEG), “Ethics Guidelines for Trustworthy AI” (2019년 4월)[38]

유럽연합 집행위원회(European Commission)는 2018년 4월 25일 인공지능 고위급 전문가 그룹(High-Level Expert Group, 이하 “HLEG”)을 발족했다. HLEG은 2018. 12. 18. “신뢰가능한 인공지능을 위한 윤리 가이드라인(Ethics Guidelines for Trustworthy AI)” 초안을 발표하였고, 이후 2019. 4. 8. 최종본을 발표했다. 가이드라인에서는 신뢰가능한 인공지능 시스템은 합법적이고(lawful), 윤리적이며(ethical), 견고한(robust) 인공지능 시스템을 의미한다고 정의하고, 법률적인 측면은 제외하고 주로 윤리적이고 견고한 시스템에 초점을 맞추어 이를 위한 원칙과 이 원칙을 실행하기 위한 요건 및 준수 여부를 평가할 수 있는 기준을 구체화하여 설시하고 있다.

이 가이드라인에서는 신뢰가능한 인공지능을 위하여 기본권에 대한 존중이 핵심요소라고 보고 이를 위한 원칙을 제시한다. 구체적 내용으로는, ① 인공지능 시스템은 사람을 대리하는 에이전트일 뿐이며, 인공지능 시스템에 종속되거나 강압 받지 않을 자유가 보장되어야 하며, 직접적 또는 간접적으로 인공지능의 의사결정의 대상이 될지 여부는 인간이 결정해야 한다는 ‘인간의 자율성에 대한 존중의 원칙(respect for human autonomy)’, ② 인공지능이 어린이, 소수자 등의 취약계층을 포함하여 인간에게 해를 끼치지 않아야 한다는 ‘해악방지의 원칙(prevention of harm)’, ③ 인공지능이 실질적, 절차적 측면 모두에서 공정해야 한다는 ‘공정성의 원칙(fairness)’, ④ 인공지능의 기술적인 투명성과 사업모델의 투명성을 기반으로 하여 설명이 가능해야 한다는 ‘설명가능성의 원칙(explicability)’을 들고 있다.

다음으로 이런 원칙을 실현하기 위한 요구사항으로, ① 자율성에 대한 존중의 원칙에 따라 인공지능 시스템이 인간 이용자의 에이전트로서 인간의 기본권을 증진시키고 인간의 감독을 허용해야 하고, ② 해악방지의 원칙에 따라 인공지능 시스템에 의한 원하지 않은 피해를 예방하기 위해 기술적 견고함과 안전성을 갖추어야 하며, ③ 인공지능 시스템이 프라이버시에 피해를 가져오지 않도록 하기 위해 데이터 처리에 대한 접근 프로토콜과 프라이버시를 보호하는 방향의 데이터 거버넌스(data governance) 구조를 만들어야 하고, ④ 설명가능성의 원칙에 따라 인공지능 시스템이 내린 결정에 관련된 데이터세트와 절차가 추적 가능하고 인공지능 시스템의 기술적 측면과 그에 관한 사람의 결정이 설명 가능하도록 해서 투명성을

38 European Commission High-Level Expert Group, “Ethics Guidelines for Trustworthy AI”, https:// ec.europa.eu/futurium/en/ai-alliance-consultation, 2019. 4. 8.자.

보장해야 하며, ⑤ 공정성의 원칙에 따라 인공지능 시스템이 다양성을 보장하고 성별, 나이, 능력 등을 불문하고 누구나 접근 가능하도록 보편적 설계(universal design)를 통해서 차별이 없어야 하고, ⑥ 공정성의 원칙과 해악방지의 원칙에 따라서 사회적, 환경적 요소도 고려요소에 포함시켜 지속가능하고 생태계에 대해서도 책임성 있는 시스템을 만들어야 하며, ⑦ 공정성의 원칙에 따라 인공지능 시스템과 시스템이 만들어낸 결과를 감사할 수 있도록 해야 하고 부정적인 영향을 최소화하고 부정적인 영향이 발생할 경우 이를 보고하며 잘못된 점을 바로잡을 수 있도록 해서 책임 보장 메커니즘을 만들 것을 들고 있다. 가이드라인은, 신뢰할 수 있는 인공지능을 구현하기 위해서는 인공지능의 설계와 시스템 구조와 관련된 기술적인 방법뿐만 아니라 규제나 행동강령과 같은 비기술적인 방법 모두 중요하고 함께 고려되어야 한다고 본다.

가이드라인에서는 인공지능을 개발하고, 배치하며, 이용하는 과정에서 신뢰할 수 있는 인공지능에 해당하는지 여부를 평가할 수 있는 항목들을 만들어 예시로 제시하는 한편, 인공지능 시스템의 이해관계자들이 자신의 조직에서 가이드라인에 제시된 평가요소를 조직의 지배구조 메커니즘에 포함시켜 실천하도록 권고하였다. 특히 회사와 같은 조직에서 경영자, 컴플라이언스 및 법무부서, 상품 및 서비스 개발부서 등이 인공지능 시스템의 개발과 관련해서 어떤 역할을 해야 하는지에 대해서도 언급하였다. 다만, 가이드라인은 평가 리스트를 모두 준수했다고 하여 관련 법률을 모두 준수한 것이라고는 볼 수 없고 리스트 내용도 모든 관련된 사항을 망라하고 있지는 않다는 것을 밝히고 있다. 하지만 이 리스트가 인공지능 시스템을 개발하거나 도입하려는 조직에서 중요한 내부적인 참조 평가기준으로 활용될 수는 있을 것이다. EU 집행위원회에서는 이러한 평가목록이 실무에서 활용될 수 있도록 하기 위해서 2019년 6월부터 모든 이해관계자들이 평가목록을 테스트 해보고 피드백을 제공할 수 있도록 하며, AI HLEG를 통하여 공적인 영역과 사적인 영역의 이해관계자들이 모두 모여 가이드라인이 실제로 어떻게 적용될 수 있을지에 대한 심층적 검토를 하도록 할 것이라고 밝혔다.

4. 해외 윤리규범이 국내에 주는 시사점

1. 해외 윤리규범의 특징과 경향성

지금까지 살펴본 다양한 해외 논의의 대략적 특징과 경향성은 다음과 같이 요약될 수 있다. 첫째, 윤리규범의 형성 주체에 따른 차이가 존재한다. 사적 주체가 내놓은 규범의 형태가 일반론적이거나 원론적 내용이 담긴 경우가 많은 반면, 공적 주체를 통해 도출된 규범의 경우는 상대적으로

다루는 분야가 넓은 편이고 규범의 구체성도 높은 경우가 많다. 특히, 유럽에서 발표된 규범에 그런 경향이 더 명확하게 나타나는 편이다. 둘째, 다양한 성격과 이해관계를 가진 주체가 공동으로 내놓은 규범일수록 상대적으로 원론적인 내용 위주이다. 대표적으로 OECD의 경우에는 매우 다양한 이해관계를 가진 주체들 사이의 협의와 협상을 통한 합의를 반영하여, 구체적인 내용이 많이 담겨있지 않다. 수많은 사적 주체가 공동으로 내놓은 아실로마 원칙 또한 참여자들이 관심을 두는 영역이 다르기 때문인지 23가지로 항목의 숫자가 적지 않지만 내용상으로는 단문 형태의 일반적 명제의 진술에 그치고 있다. 다만, 원론적인 내용 위주라고 해서 반드시 규범력이나 강제력이 적은 것은 아니다. 특히 OECD 규범은 후속조치와 정기적인 모니터링에 관한 내용을 담고 있어, 이러한 방식을 통해 실질적인 규범력이 확보될 수 있는 장치를 마련하고 있다.

셋째, 지역적인 문화적 차이나 규제를 바라보는 시각차이도 어느 정도 엿볼 수 있다. 유럽과 미국을 비교하면, 유럽연합의 경우는 시범사업(pilot test) 단계에 돌입할 정도로 윤리규범의 내용이 구체화된 상태이다. 반면, 미국의 경우 대체로 개별 기업들을 통해 규범이 제시되고, 그와 별개로 학계를 통해 활발한 논의가 이루어지고 있는 형편이다. 이런 차이에는 다양한 원인이 있을 수 있지만, 최근의 법제도 동향을 보면 유럽에서는 GDPR과 같은 일반적인 법규범을 마련하여 통일적으로 시행하는 것이 자연스럽게 받아들여지는 한편, 미국에서는 적어도 지금까지는 개별 영역별 접근 그리고 자율규제의 역할을 상대적으로 강조하는 면이 있다. 따라서 이와 같은 경향성을 국내 인공지능 규범 논의에 참조할 때에는, 우리나라가 처해있는 현실적 여건을 종합적으로 고려한 대안을 마련하여야 할 것이다.

2. 우리나라 윤리규범 논의의 현황과 과제

그렇다면 앞서 살펴본 해외의 논의와 비교하여, 우리나라의 인공지능 윤리규범 논의는 어떤 모습으로 변화해왔을까? 먼저 앞서 언급했던 2007년에 발표된 로봇윤리헌장 초안을 살펴보자. 이는 지능형 로봇 개발 및 보급 촉진법 제18조에 법적 근거를 두고, 2016년에는 개선안까지 마련되었지만 아직까지 정식으로 시행되지는 못했다. 그러나 알파고가 사회적으로 큰 파장을 낳은 이후 인공지능 규범 마련에 대한 공감대가 형성되면서 여러 논의가 이어지고 있다. 박영선 의원은 2017. 7. 19. 로봇의 윤리와 책임에 관한 내용을 담은 '로봇기본법안'을 대표로 발의하였고, 그해 연말 정보문화포럼과 한국정보화진흥원이 공동으로 공공성(Publicness), 책무성(Accountability), 통제성(Controllability), 투명성(Transparency)이라는 소위 'PACT 원칙'을 개발자, 공급자,

이용자에 대해 적용한 '지능정보사회 윤리 가이드라인'이 마련되었다. 그리고 카카오는 2018. 1. 31.에 '카카오 알고리즘 윤리 헌장'을 마련하여 발표한 바 있다. 다만, 카카오 윤리 헌장은, ① 카카오 알고리즘의 기본 원칙, ② 차별에 대한 경계, ③ 학습 데이터 운영, ④ 알고리즘의 독립성, ⑤ 알고리즘에 대한 설명이라는 다섯 가지 원칙을 원론적으로 설명한 것이어서, 내용이 구체적이지 않다. 같은 해 6월에 발표된 '지능정보사회 윤리헌장'에서도 단지 여섯 가지의 선언적 명제를 짤막하게 언급한다.[39] 한편, 변재일 의원이 2018. 2. 14. 대표로 발의한 '국가정보화 기본법 전부개정법률안' 제62조는 "공공성 책무성 통제성 투명성 등의 윤리원칙을 담은 지능정보사회윤리"라는 표현을 명시적으로 언급하고 있다.

'개발자, 공급자, 이용자의 공공성 책무성 통제성 투명성'이라는 명제로 대표되는 국내의 논의는 규범 논의의 주체와 목적을 인간으로 상정한다는 점, 책무성과 투명성이라는 주체성과 책임성을 동시에 고려한 원칙을 수용하고 있다는 점에서 어느 정도 해외의 논의와 일맥상통한 내용을 담고 있다고 평가할 수 있다. 향후에 논의가 계속된다면, 이제부터는 규범의 구체화와 강제력 부여를 비롯한 좀 더 현실적인 사항들에 대한 논의가 이루어질 것으로 예상해 볼 수 있다. 국내에서의 향후 논의는 국내의 실정에 부합하는 동시에 국제사회의 논의에 발맞추어 진행되어야 한다. IEEE나 IAPP의 사례처럼 전문가들이 참여하여 만든 실질적인 국제표준은 실무자나 개발자 등에게는 사실상의 행동지침으로 작용할 수 있다. 다른 한편 OECD나 유럽연합 등을 통한 논의는, 참가 당사국을 통해 후속 논의와 후속 조치를 하는 장치를 마련하고 있어서, 이를 통해 현실적인 집행력이 확보될 것이고, 이는 정책적으로도 적지 않은 파급력을 미칠 것이다.

그런데 인공지능 윤리규범에 관한 논의가 경우에 따라 일견 원론적이고 교과서적 차원의 논의에 그치는 것으로 비쳐질 수도 있지만, 매우 간단한 추상적인 내용만을 담고 있는 규범의 경우에도 그 이면에는 이해관계자 사이에 규범논의를 이끌어가려는 치열한 이익대립이 숨겨져 있는 것이 일반적임을 간과해서는 곤란하다. 윤리규범의 내용이 매우 간략한 경우에, 개별 규범에 담긴 내용이 어떤 것인지 파악하여 분석하는 것이 중요한 것은 물론 규범에 담기지 않은 것은 어떤 것인지에 대하여 파악하고 그 함의를 분석하는 것이 더 중요할 수도 있다. 또는 추상적인 원칙으로부터 분석을 통해 '행간'을 읽어낼 필요가 있을 수도 있다. 그 이외에도, 선언적인 내용이, 선언 그 자체에 그치는 경우와 실제로 집행력을 확보하게 되는 경우 사이의 구분도 중요하다. 그러한 면밀한 분석과 판단이 이루어지지 않으면 우리나라의 실정에 부합하지 않거나 부작용을 낳을 수 있는 규범임에도

39 정보문화포럼 한국정보화진흥원, "지능정보사회 윤리 가이드라인", 2017, 23면.

'사실상의 국제표준(de facto global standard)'이라는 이유로 억지로 수용하는 결과를 낳을 수도 있다. 국제사회의 '규범전쟁(norm war)' 속에서 주체적으로 목소리를 내고 나아가 논의에서 선도자로서의 지위를 점유하기 위해서는 그와 같은 깊이 있는 분석과 관련된 논의가 지속될 필요가 있다.

5. 윤리적 인공지능의 미래: 결론을 대신하여

이하에서는 결론을 대신하여 윤리적 인공지능의 실현과 정착을 위한 몇 가지 제언을 정리하도록 한다. 지금까지의 논의를 통해 공감대를 형성한 가장 커다란 원칙은, 인류가 지향해야 할 목표는 인공지능이 인간의 존엄성과 권리의 실현을 위한 방향으로 활용되어야 한다는 점이다. 그러한 대원칙에 부합하지 않는 관념들, 특히 아래에서 언급하는 이분법적 관념들은 인공지능 시대와 부합하도록 변화를 모색할 필요가 있다.

우선, 법규범과 윤리를 비롯한 여타 사회규범의 유형을 엄밀히 구별하는 태도를 들 수 있다. 전통적으로 법규범은 강제력을 주요한 근거로 삼아 여타 사회규범 유형과 뚜렷이 구분되어 왔지만, 인공지능의 맥락에서는 그러한 구분이 명확하지 않은 경우도 많고 또한 그러한 구분으로부터의 실익이 크지 않은 경우도 많다. 둘째로 기술과 규범을 엄밀하게 분리하여 파악하려는 태도가 존재한다. 인공지능 윤리의 맥락에서는, 기술과 규범이 서로 보완적인 역할을 할 필요가 있다. 셋째, 법규범 내부에서 해악의 위험성에 대하여 사전적으로 공적, 행정적 규제를 강조하는 영역과, 발생된 해악에 대하여 주로 사후적이고 민사적인 사법판단을 중시하는 영역을 엄밀히 구분하려는 태도가 있다. 이는, 사적 영역은 '사적자치의 원칙'이 강조되어 구체적 해악이 초래되지 않는 이상 외부에서의 개입이 정당화되기 어렵고, 소위 '위험사회'에 대한 사전규제는 공공기관이 담당해야 할 몫이라는 입장과도 이어지는 것이다. 이 또한 인공지능의 맥락에서는 현실성이 약한 구분이다.

전반적으로, 오늘날 진행되고 있는 국제사회의 인공지능 규범논의는 이와 같은 전통적 이분법의 관념과는 크게 다른 것이다. 무엇보다도 인공지능과 같이 급속도로 발전하는 신기술은 전형적인 변화 패턴을 예상하기 어렵다. 예컨대, 오늘날 인공지능 방법론의 전형이라고 할 만한 것은 딥러닝(Deep Learning) 기술일 텐데, 어떠한 인공지능 시스템도 학습방식을 오직 딥러닝에만 국한하지는 않는다. 딥러닝과 대비되는 규칙기반 전문가 시스템도 여전히 사용되고 있고, 여러 유형의 인공지능 모델이 조합을 이루어 이용되는 경우를 빈번하게 볼 수 있다. 또한 딥러닝 알고리즘도 수많은 변이가 발생 중이고, 딥러닝 패러다임 자체가 전환될 가능성도 배제할 수 없다. 그리고 실제 상용화 과정에서는 흔히 기존의 제품과 서비스에 인공지능 요소를 약간씩 포함시키고, 이를 점차 확대하고

고도화하는 과정을 거치게 된다. 이때, 개별 제품이나 서비스에서 인공지능 요소를 별도로 떼어내어 법규범적 검토를 하는 것은 현실적이지 않다.

인공지능 기술의 발전과 변화의 속도가 빠르고 변화의 방향에 대해서도 예측이 어렵다는 점은 입법자가 인공지능의 단일한 본성을 상정하고 법적 절차를 통하여 사전적 규제 체계를 마련한다거나 사후적 민 형사 책임 위주의 법체계를 입법하는 '경성법(hard law)' 위주의 전통적 접근을 어렵게 한다. 인공지능은 국제사회의 이해관계자들 사이의 첨예한 대립이 빈번하다는 점에서, 구속력과 강제력을 지나치게 강조하는 경우 도리어 무규범(anomie) 상태로 흐르거나 서로 상충되는 법체계가 복잡하게 혼재하는 상태가 초래될 가능성도 있다. 반면, 실정법적인 구속력과 강제력은 없지만, 행위규범의 일종으로 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 다양한 이름으로 만들어내고 있는 연성법은 미래사회의 방향성을 제시하고, 현실과 이상 간의 간극을 메우는 등 보다 유연한 대처가 가능한 장점을 가지기 때문이다.[40] 연성법은 기술규제, 자율규제와도 공존하는 대목이 있어 인공지능 거버넌스 담론에서 많은 지지를 획득하고 있다. '책임'에 관한 논의에 있어서도, 법적 책임을 의미하는 'liability'가 아니라 그 보다 포괄적인 의미를 가지는 'accountability' 개념이 주로 언급되는 것에 유의할 필요가 있다.

다만, 연성법 체제를 좀 더 본격적으로 도입하기 위해서는 이를 위한 충분한 준비가 필요하다. 가장 기본적으로 우리나라의 법체계 구조와 관행상, 경성법이 연성법에 비해 높은 예측가능성을 가진다는 점이 지적될 수 있다. 연성법이 경성법의 보완재라기보다는 예측가능성 낮은 '추가적 규제'로 무분별하게 남용될 경우, 명확한 경성법 일원론적 규제방식보다도 못한 결과를 낳을 수도 있다. 또한 행정, 사법, 산업 영역에 대한 사회구성원들의 신뢰(trust) 정도도 중요한 변수가 될 수 있다. 신뢰수준이 낮은 영역에 연성법을 통한 독자적 또는 자율적 규율 권한을 과도하게 부여하는 경우 막대한 사회비용이 유발될 수도 있을 것이기 때문이다. 다른 한편, 연성법이 그저 대원칙에 대한 선언으로만 비쳐지고 현실적이고 실질적인 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. 앞으로의 인공지능 윤리담론은 이런 다양한 측면을 함께 고려한 풍부하고 깊이있는 논의가 되어야 할 것이다.

---

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

참고 문헌

김효은, 『인공지능과 윤리』, 커뮤니케이션북스, 2019.

라파엘 카푸로 미카엘 나겐보르그, 변순용 송선영 역, 『로봇윤리 - 로봇의 윤리적 문제들 -』, 어문학사, 2013.

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

이원태 문정욱 이시직 심우민 강일신, 『지능정보사회의 규범체계 정립을 위한 법 제도 연구』, 정보통신정책연구원, 2016.

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, "인공지능 알고리즘과 시장", 『데이터 이코노미』, 한스미디어, 2017.

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

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

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

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

양천수, "현대 지능정보사회와 인격성의 확장", 『동북아법연구 제12권 제1호』, 2018.

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

오요한 홍성욱, "인공지능 알고리즘은 사람을 차별하는가?", 『과학기술학연구 제18권 제3호』, 2018.

윤지영, "인공지능 기술 관련 법적 제도적 논의 현황", 『법과학을 적용한 형사사법의 선진화 방안(VIII) : 인공지능 기술』, 한국형사정책연구원, 2017.

이원태, "유럽연합(EU)의 로봇법(RoboLaw) 프로젝트", 『KISO Journal Vol. 23』, 2016.

이원태, "4차 산업혁명과 지능정보사회의 규범 재정립", 『KISDI Premium Report 17-10』, 2017.

정보문화포럼 한국정보화진흥원, "지능정보사회 윤리 가이드라인", 2017

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 정책지원팀, "KAKAO AI REPORT", Vol. 1, 2017.

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

Alex Campolo et al., "AI Now 2017 Report", 2017.

Erica Palmerini et al., "Guidelines on Regulating Robotics", 2014.

European Commission High-Level Expert Group, "Ethics Guidelines for Trustworthy AI", 2019.

Gianmarco Veruggio, "EURON Roboethics Roadmap(ver. 1.1)", 2006.

Gianmarco Veruggio, "EURON Roboethics Roadmap(ver. 1.2)", 2007.

Google, "Perspectives on Issues in AI Governance", 2019.

Helen Nissenbaum, "Accountability in a Computerized Society", Science and Engineering Ethics Vol. 2, Issue. 1, 1996.

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

Institute of Electrical and Electronics Engineers, “Ethically Aligned Design Version 2 - A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems”, 2017.

Jean-François Bonnefon·Azim Shariff·Iyad Rahwan, “The social dilemma of autonomous vehicles”, Science Vol. 354, Issue. 6293, 2016.

Karl M. Manheim, Lyric Kaplan, “Artificial Intelligence: Risks to Privacy and Democracy”, Forthcoming, Yale Journal of Law and Technology, 2018.

Kate Crawford·Meredith Whittaker, “The AI Now Report”, 2016.

Meredith Whittaker et al., “AI Now Report 2018”, 2018.

Merel Noorman, “Computing and Moral Responsibility”, Stanford Encyclopedia of Philosophy, 2018.

Microsoft, “Responsible bots: 10 guidelines for developers of conversational AI”, 2018.

Microsoft, “Six Principles for Developing and Deploying Facial Recognition Technology”, 2018.

Peter M. Asaro, “What Should We Want From a Robot Ethic?”, International Review of Information Ethics Vol. 6, 2006.

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.

Robin R. Murphy · David D. Woods, “Beyond Asimov: The Three Laws of Responsible Robotics”, IEEE Intelligent Systems Vol. 24, Issue 4, 2009.

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

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.

UK House of Lords, “AI in the UK: ready, willing and able?”, 2018.

United Nations Global Pulse · International Association of Privacy Professionals, “Building ethics into privacy frameworks for big data and AI”, 2018.

U.S. Executive Office of the President, “Preparing for the Future of Artificial Intelligence”, 2016.

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

부록

인공지능과 관련한 주요 윤리 규범 목록
(발간일 기준 정렬)

	발간주체	윤리 규범	발간일
1	유럽로봇연구네트워크(EURON)	Roboethics Roadmap Release 1.1 Roboethics Roadmap Release 1.2	2006.7 2007.1
2	산업통상자원부	로봇윤리헌장 초안	2007
3	유럽 로봇법 프로젝트(RoboLaw)	RoboLaw Guidelines on Regulating Robotics	2014.9
4	IEEE (Institute of Electrical and Electronics Engineers)	Ethically Aligned Design	2016.12 발간 (Version I), 2017.12 발간 (Version II), 2019년 최종본 발간 예정
5	Future of Life Institute	Asilomar AI Principles	2017.1
6	ACM (Association for Computing Machinery)	Principles for Algorithmic Transparency and Accountability	2017.1
7	유럽의회 법사위원회(European Parliament Committee on Legal Affairs)	European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics	2017.2
8	일본 인공지능학회	윤리지침	2017.2
9	독일 교통부 윤리위원회 (German Federal Ministry of Transport and Digital Transformation, Ethics Commission)	German Ethics Code for Automated and Connected Driving	2017.6
10	Sage	Ethics of Code: Developing AI for Business with Five Core Principles	2017.6
11	영국 상원(House of Lords)	5 Overarching Principles for an AI	2017.6
12	일본 총무성	Draft AI R&D Guidelines for International Discussions	2017.7
13	UNESCO COMEST (UNESCO 세계과학기술윤리위원회)	Report of COMEST on Robotics Ethics	2017.9
14	정보문화포럼 & 한국정보화진흥원	지능정보사회 윤리 가이드라인	2017.11

	발간주체	윤리 규범	발간일
15	카카오	카카오 알고리즘 윤리헌장	2018.1
16	영국 상원(House of Lords)	AI in the UK: ready, willing and able	2018.4
17	Google	AI at Google: Our Principles	2018.6
18	정보문화포럼 & 한국정보화진흥원	지능정보사회 윤리헌장	2018.6
19	ACM (Association for Computing Machinery)	Code of Ethics	2018.7
20	Microsoft	Responsible bots: 10 guidelines for developers of conversational AI	2018.10
21	Sony	Sony Group AI Ethics Guidelines	2018.10
22	UNGP & IAPP (UN Global Pulse & International Association of Privacy Professionals)	Building ethics into privacy framework for big data and AI	2018.10
23	Public Voice	Universal Guidelines for Artificial Intelligence	2018.10
24	Microsoft	Six Principles for Developing and Deploying Facial Recognition Technology	2018.12
25	Google	Responsible AI Practices	2018.12
26	UNESCO COMEST (UNESCO 세계과학기술윤리위원회)	Preliminary Study on the Technical and Legal Aspects Relating to the Desirability of a Standard-Setting Instrument on the Ethics of Artificial Intelligence	2019.3
27	EU 집행위원회 인공지능 고위급 전문가 그룹(High-Level Expert Group)	Ethics Guidelines for Trustworthy AI	2019.4
28	OECD AI Expert Group	Draft Recommendations of the Council on Artificial Intelligence	2019.5 예정

서울대학교
인공지능정책
이니셔티브 안내

서울대학교 인공지능정책 이니셔티브는 인공지능과 관련된 다양한 사회경제적, 법적, 정책적 이슈들을 연구하고 논의하기 위해 시작된 서울대학교 법과경제연구센터의 프로그램입니다. '소셜랩(Social Lab)' 개념을 지향하여, 여러 배경과 관심을 가진 분들 사이의 협업과 지속적인 대화를 추구합니다. 서울대학교 법학전문대학원의 고학수 교수와 임용 교수가 함께 이끌고 있습니다.

1. 발간물 안내

서울대학교 인공지능정책 이니셔티브의 주요 발간물은 이슈페이퍼와 워킹페이퍼가 있고, 비정기적으로 발간되는 단행본 및 학술행사 자료집 등이 있습니다. 이슈페이퍼와 워킹페이퍼 등의 자료들은 홈페이지를 통해 다운로드 받으실 수 있습니다.

2. 행사 안내

서울대학교 인공지능정책 이니셔티브의 주요 행사는 이슈페이퍼를 발표하고 논의하는 행사(상반기 및 하반기 각 1회) 그리고 국내외 연구자들을 초빙하여 진행하는 대규모 국제학술대회(연 1회) 등이 있습니다. 그 이외에 비정기적으로 진행하는 행사들도 있습니다.

3. 이슈페이퍼 2019

이번 이슈페이퍼는 서울대학교 인공지능정책 이니셔티브의 첫 이슈페이퍼로, 2019.5.16. D2 Startup Factory에서 열린 행사에 맞춰 준비되었습니다.

http://ai.re.kr