왜 윤리인가:
현대 인공지능 윤리 논의의 조망, 그 특징과 한계

허유선* 이연희 심지원*****

【요약】
인공지능 기술 및 산업의 발전, 사회적 적용과 함께 인공지능 윤리에 대한 요청도 커지고 있다. 실제로 현재 인공지능 윤리 원리 혹은 권고안의 발간 주체는 기업, 정부, 국제기구, 학계, 종교계로 다방면을 아우른다. 통상 윤리가 기술의 발전을 저해하는 요소로 간주되었던 기술-윤리의 긴장 관계를 고려할 때, 이는 주목할 만한 지점이다. 인공지능 기술과 관련해서 왜 윤리 논의는 더욱 적극적으로 요청되며 증가하는가? 이 물음에 답하기 위해서, 현재의 인공지능 윤리 논의를 전체적으로 조망하고 비판적으로 검토하는 메타 연구가 필요하다. 우리는 현재 기술 수준의 인공지능 윤리 논의에 초점을 맞추어, 주로 3년 이내의 각국 정부, 국제기구, 기업 등 주요 인공지능 행위자가 발간한 인공지능 윤리 문헌을 검토한다. 그리고 이를 통해 오늘날 인공지능 윤리 논의의 특성을 비판적으로 고찰한다. 논문은 현재 인공지능 윤리 논의에서 윤리가 요청되는 목적, 인공지능 윤리 논의의 전반적인 경향을 밝히고, 결론적으로 인공지능 윤리의 나아갈 방향을 제안한다. 인공지능 윤리는 인공지능 기술의 특수성, 사회에 미치는 잠재적 영향력, 산업 기반 마련을 위해 다방면에서 요청된다. 그러나 이들 논의는 반복적이고 중복적인 용어로 인한 인공지능 윤리 접근성의 저하, 윤리의 도구화, 윤리 원리 및 권고안의 실효성 약화, 논의 주체의 제한과 다양성 결여, 원리론 자체의 한계, 학술 연구의 부족, 윤리와 기술 관계에 대한 제한적 이해 등의 경향을 갖는다. 인공지능 윤리는 이러한 한계를 보완하며 다양한 후속 연구 및 논의로 확장되어야 할 것이다. 이를 위해 먼저 '인공지능 윤리'가 지시하는 바가 명확해져야 한다. 향후 '인공지능 윤리학'은 윤리가 기술의 기획, 형성, 배치 등에 이미 녹아있음을 자각하고 기술의 개발 및 사회 도입에 앞서서, 그리고 기술의 전체 단계와 함께 수행되어야 할 것이다.

【주제어】 인공지능, 인공지능 윤리, 인공지능 원리, 기술 윤리, 인공지능 윤리 메타 연구

---

* 중앙대학교 인문콘텐츠연구소

** 한국교통대학교 교양학부

*** 중앙대학교 인문콘텐츠연구소

**** 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A6A3A01078538)

https://doi.org/10.34162/hefins.2020..24.006

Ⅰ.문제제기: 인공지능 시대, 왜 다시 윤리인가?

인공지능 기술에는 왜 윤리적 논의의 동반이 강력히 요청될까? 최근 몇 년간, 인공지능 기술 및 산업의 발전과 더불어 통상 '인공지능 윤리AI Ethics'에 대한 논의가 급증하고 있다. 인공지능 윤리의 급격한 성장은 이제 하나의 현상이라 할 수 있다. 이 현상은 각계의 요청에서 비롯하는데, 그 요청의 주인공이 전문적으로 윤리를 다루는 학계에 국한되지 않는 것이 흥미로운 지점이다. 인공지능 윤리의 요청 및 논의는 전 세계에 걸쳐 진행 중이다. 학계, 기술 전문가, 기업, 정부, 국제기구, NGO 등 다양한 조직이 인공지능 윤리의 원칙, 성명, 권고안을 발표하고 있다. "2017년 이후 학계, 사적 영역, 공공영역의 행위 주체들actors (……) 주요 국가들의 정부(의회), 대학 연구소, IT 기업 연합체, 싱크 탱크, 민간 이니셔티브, ITU 및 OECD 등 국제기구, World Wide Web Consortium, IEEE 등 국제 표준화 기구가 윤리·규범 정책 논의에 참여하고 있는 것이다."[1)]

인공지능 윤리는 법, 정책, 기술, 산업 등 다방면에서 다양한 주체에 의해 논의되고, 이러한 현상은 일각에서 인공지능 윤리에 대한 "전 지구적 합의"가 성장한 산물로 해석되기도 한다.[2)] 새로운 기술 산업으로서 인공지능 기술 관련 산업을 육성하려는 정부에게 인공지능 윤리는 주요한 정책 이슈이기도 하다.[3)] 무엇보다 인공지능 윤리 논의에 기술 기업tech company이 적극적으로 참여한다는 것이 주목할 만한 지점이다. 기술과 윤리를 긴장 관계로 보는 통상적 이해에 비추어 볼 때, 특히 기술 업계에서 윤리가 기술의 발전을 저해하는 것으로 간주되었음을 생각하면 이는 매우 놀라운 전환이다.[4)]

---

1) 정보통신정책연구원 (2018), p. 48.

2) Crawford et al, (2019), p. 19.

3) West, Allen (2018).

물론 현재의 인공지능 규제 논의에서 기술과 윤리를 긴장 관계로 보는 관점이 전적으로 제거된 것은 아니다.[5] 그러나 오늘날 인공지능 기술 관련 논의에서 윤리의 필요성이나 윤리적 고려가 전적으로 배제되는 경우는 없으며, 대부분의 인공지능 기술 논의는 윤리 논의를 포함한다. 윤리적 고려가 기술 발전의 저해 요인이 될 수 있다는 시각에도 불구하고, 인공지능 기술에서 윤리가 이렇게 강조되는 이유는 무엇일까? 또한 서로 다른 입장의 인공지능 윤리 논의에서 논하는 '인공지능 윤리'의 의미는 무엇일까? 그리고 작금의 인공지능 윤리 논의는 자신에게 기대되는 바 혹은 '윤리 논의'로서의 자격을 충족하기에 충분할까?

향후 기술, 산업, 정책과 함께 인공지능 윤리에 대한 요구는 더욱 커질 전망이며, 인공지능 윤리는 구체적 맥락에서의 실행, 새로운 이슈에 대한 논의 등 심화 및 후속 연구로 나아가야 할 시점이다. 그러나 이를 위해서는 먼저 위의 물음에 답할 필요가 있을 것이다. 인공지능 윤리 논의 자체에 대한 조망, 검토와 반성이 필요한 것이다. 그러나 인공지능 윤리 논의를 전체적으로 조망하는 연구는 아직 드문 실정이다.[6] 특히 인공지능 윤리 논의의 '윤리적' 성격에 초점을 맞춘 연구는 매우 드물다.

---

4) "대부분의 사람들에게 기술발전은 그 자체로 선善이었다. 그러나 최근 회자되는 인공지능이나 4차 산업혁명에 대한 논의들에서 (…) '4차 산업혁명' 같은 말을 만들어낸 장본인조차 급격한 변화가 초래할 여러 가지 문제들에 대한 우려를 표명하고 대안의 모색을 강조한다. 이렇듯 기술의 발전에 어떤 방식으로든 제한을 두어야 한다는 식의 발언이 기술발전을 선도하는 사람들에게서 나오는 것은 상당한 변화이고, 그 자체로 흥미로운 분석의 대상이다." 손화철 (2018), p. 268.

5) "대한민국 정부의 '인공지능 국가전략(2019.12)'은 'AI 경쟁력 혁신'을 달성하기 위한 과제 중 하나로 '과감한 규제 혁신 및 법제도 정비'를 꼽고 있다. "AI 시대가 도래하였으나 현행 규제와 新기술 간 괴리, AI 확산에 대응하는 기본원칙 및 각 분야별 규율체계 부재로 혁신의 지체" 가 우려되기 때문이다. (관계부처 합동, 인공지능 국가전략, p. 19) 이러한 인식의 근저에는 "윤리 문제에의 대응이 결과적으로 산업 성장을 위축시키는 규제가 되지 않도록 하는 노력이 필요하다" 는 생각이 있다." 정보통신정책연구원 (2018), p. 187.

6) 인공지능 윤리 논의에 대한 메타 연구는 세계적으로도 손에 꼽을 정도이다. 본 논문의 참고문헌을 참조하라.

우리는 '지금-여기'에서 요청되며, 앞으로 더욱 중요한 주제가 될 '인공지능 윤리'에서 논의의 혼란을 해소하고 실효성 있는 논의에 집중할 수 있도록 공통의 이해와 토대를 마련하고자 한다. 본 논문은 최근 발표된 인공지능 윤리 원리, 권고안 등 현실적으로 영향력을 발휘하고 있는 인공지능 윤리의 주요 문헌을 검토하고 이들에 대한 메타연구를 비교·분석함으로써, 인공지능 윤리 논의의 성장 및 현황을 고찰하고 현재의 인공지능 윤리 논의 전반을 조망한다. 이로부터 인공지능 기술과 관련하여 특히 윤리가 요청되는 원인을 분석하고 그 근저에 놓인 기술과 윤리에 대한 태도를 비판적으로 고찰한다. 이를 통해 논문은 1) 인공지능 기술에서 특히 더 윤리가 강조되는 이유가 있는가? 2) 현재까지의 인공지능 윤리 담론에서 윤리는 어떤 것으로 이해되고 있는가? 그 이해는 적절한 것인가? 에 답한다.

Ⅱ.연구대상 및 선정 기준

번호	제목	발간처	발간 지역	연도	비고
1	AI NOW Report 2016	AI Now Institute, New York Univers	USA	2016	메타
2	AI NOW Report 2017	AI Now Institute, New York Univers	USA	2017	메타
3	AI NOW Report 2018	AI Now Institute, New York Univers	USA	2018	메타
4	Linking Artificial Intelligence Principles	AAAI-Safe AI		2019	메타
5	The Global Landscape of AI Ethics Guidelines	Health Ethics & Policy Lab	Switzerland	2019	메타
6	신뢰 가능 AI 구현을 위한 정책 방향 -OECD AI 권고안을 중심으로	한국정보화진흥원	대한민국	2019	메타
7	Discriminating Systems - Gender, Race, and Power in AI. AI Now	AI Now Institute, New York Univers	USA	2019	메타
8	AI NOW Report 2019	AI Now Institute, New York Univers	USA	2019	메타
9	The Ethics of AI Ethics: An Evaluation of Guidelines	International Center for Ethics in the Sciences and Humanities	Germany	2019	메타

10	A Unified Framework of Five Principles for AI in Society	Harvard Data Science Review	USA	2019	메타
11	Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches	The Berkman Klein Center for Internet & Society at Harvard University	USA	2020	메타
12	OpenAI Charter	OpenAI	USA	2018	민간영역 (기업)
13	DeepMind Ethics & Society Principles	DeepMind	UK	2017 (unconfirmed)	민간영역 (기업)
14	Artificial Intelligence at Google: Our Principles	Google	USA	2018	민간영역 (기업)
15	Microsoft AI principles	Microsoft	USA	2018	민간영역 (기업)
16	Principles for Trust and Transparency	IBM	USA	2018	민간영역 (기업)
17	카카오 알고리즘 윤리 헌장	카카오(kakao)	대한민국	2018	민간영역 (기업)
18	Tenets	Partnership on AI	USA	2016	비영리 기구
19	4차 산업혁명시대 인공지능 윤리의 이슈 분석 및 정책적 대응방안 연구	정보통신정책연구원	대한민국	2018	정책
20	Preparing for the Future of Artificial Intelligence	Executive Office of the President: National Science and Technology Council: Committee on Technology	USA	2016	정책
21	Ethics Guidelines for Trustworthy AI	The European Commission's High-Level Expert Group on Artificial Intelligence	Europe	2019	정책
22	OECD Principles on Artificial Intelligence	The Organisation for Economic Co-operation and Development (OECD)	International	2019	정책
23	대한민국 2019 AI 국가 전략	대한민국 정부	대한민국	2019	정책
24	AI Rome Call for AI Ethics	The Vatican	Rome	2020	종교계
25	The Montreal Declaration for a Responsible Development of Artificial Intelligence	University of Montreal	Canada	2018	학계

26	Beijing AI Principles	Beijing Academy of Artificial Intelligence (BAAI)	China	2019	학계
27	AI Policy Principles	Information Technology Industry Council (ITI)	International	2017	학계
28	Ethically Aligned Design (version.2)	The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	International	2019	학계

[표-1] 28개의 분석 자료 목록

위의 분석 자료들은 일차적으로 문건의 성격(비고의 내용), 이차적으로 발간된 연도를 기준으로 정렬하였다.

연구 대상 선정 기준	해당 분석 자료
최신의 자료일 것.	모든 자료[7]
인공지능 윤리 및 가치에 초점을 맞춘 것	모든 자료
특수한 영역과 문제가 아닌 포괄적이고 일반적인 인공지능 윤리 논의(윤리헌장, 원리, 권고안 등)일 것	모든 자료
공신력 있는 '기관'이 발간했거나 미디어를 통해 널리 알려진 것	1, 2, 3, 7, 8, 9, 10, 26, 27, 28
AI 경쟁에서 큰 영향력을 행사하는 주요 '국가 및 초국가' 기관이 발간한 인공지능 윤리를 포함한 문건	18, 20, 21, 22, 24, 25
해당 분야에서 큰 영향력을 지닌 '기업'이 발간한 인공지능 윤리 관련 문건	12, 13, 14, 15, 16
'한국 정부와 기업'의 인공지능 관련 윤리 문건	6, 17, 23
인공지능 윤리 논의를 고찰하는 '인공지능 윤리 메타 논문'	4, 5, 7, 9, 11

[표-2] 연구 대상 선정 기준과 해당분석 자료

7) 대부분이 최근 3년 간의 자료들이나, 우리의 선정 기준에 따라 예외를 허용하였다. 1, 2, 13, 18, 20, 27번 자료가 그에 해당된다.

우리는 2016년 이래 등장한 인공지능 윤리 논의의 현재를 적절하게 반영하고, 전체적 조망, 경향 및 특징을 제시하려는 목적으로 비교적 최근의 관련 문헌을 분석 대상으로 선정하였다. 그러므로 우리 연구가 인공지능 윤리를 다루는 모든 문헌을 검토하지는 않는다.[8] 우리는 다양한 분야의 행위 주체를 고려하고 상대적으로 영향력이 큰 인공지능 행위자를 중심으로 자료를 선정하였다. 또한 최신 경향을 살펴보기 위해 최근 3년간의 자료에 집중하였고 그중에서 인공지능 윤리 일반에 대한 논의, 곧 인공지능 원리 중 윤리 파트, 인공지능 윤리 원칙 및 헌장, 인공지능 윤리 권고안 등을 비교, 검토하였다. 구체적인 선정 기준은 최근의 자료일 것(AI Now의 연간 보고서 등 몇 가지 예외 자료가 있다), 인공지능 윤리 및 가치를 주로 다룰 것, 특수한 영역과 문제가 아닌 포괄적이고 일반적인 인공지능 윤리 논의(윤리 헌장, 원리, 권고안 등)일 것이다. 그중에서도 공신력 있는 기관이 발간했거나 미디어 등을 통해 널리 알려진 것 등, 실질적인 영향력을 미치는 것으로 한정하였다. 이는 공공영역과 민간영역, 초국가적 기구 및 종교계 등을 고루 포괄한다. 이러한 기준에 따라 인공지능 윤리의 특수하고 제한적 이슈를 다루는 학술 연구 문헌은 제외하였다. 또한 국가 전략, 기업 정책, 산업적 논의 등 인공지능 윤리에 초점을 맞추지 않은 것은 가능한 한 배제하였다. 그러나 AI 경쟁에서 큰 영향력을 행사하는 주요 국가 및 초국가 기관, 기업이 발간한 자료는 포함하였다. 오늘날 인공지능 윤리 논의에서 커다란 영향력을 발휘하는 주체가 인공지능 윤리를 다루는 방식과 관점을 고찰하기 위해서이다. 그리고 동일 발행처에서 발간된 관련 자료는 '인공지능 윤리 원리 혹은 권고안'이라는

8) 대신 우리는 인공지능 윤리 전반을 검토한 다른 메타 연구 자료로 다음을 추천한다. 최근 5년 간의 윤리 가이드라인 비교 분석은 Hagendorff (2019)과 Jobin, Ienca, Vayena (2019), 산업계, 정부, 학계 등 서로 다른 조직 간의 인공지능 원리의 공통점, 고유성, 관계를 분석한 것은 Yi, Lu, Huangfu (2019), 기한을 한정하지 않고 주된 윤리 원리를 선별, 분석한 것은 Floridi, Cowls (2019) 인공지능 원리 자료 전반을 분석하여 윤리 및 권리 기반 접근을 총체적으로 조망, 시각화한 것은 Fjeld et al. (2020)를 참조하라.

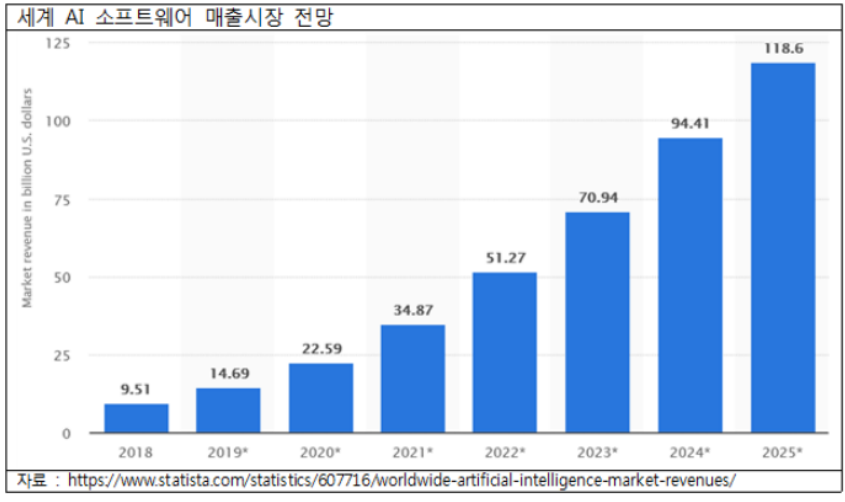
용어를 제목으로 하는 문헌, 그 중에서도 가장 최신의 자료를 검토하였다. 다만 AI Now에서 발간한 보고서의 경우는 인공지능 윤리의 메타자료로 간주하여, 인공지능 이슈의 변천을 확인하는 목적으로 2016년부터의 모든 자료를 검토하였다. 그리고 한국 내 인공지능 윤리 논의 동향의 검토를 위해 한국의 인공지능 국가전략 및 정책보고서를 포함하였다. 그러므로 인공지능 일반이 아니라 알고리즘에 국한된 것이지만 한국 기업 내 유일한 윤리헌장으로서 카카오 알고리즘 윤리 헌장도 포함하였다. 마지막으로 인공지능 윤리 논의의 양적 특성 외 일반적 경향을 이해하고 우리 분석의 타당성 검증을 위해 지금까지 발간된 인공지능 윤리 논의에 대한 메타 연구 문헌을 함께 검토하였다.

Ⅲ. 인공지능 윤리 논의의 현황과 성장

1. 인공지능 기술 및 산업 영역의 성장

인공지능 자체 기술 및 활용 기술을 통한 시장규모에 대해서 국가나 세계적인 기구 및 연구소 또는 다수의 조사기관에서 인공지능 기술 및 산업의 성장세를 보여주고 관련 시장의 잠재성에 대하여 높게 예측하고 있다. 글로벌 AI 소프트웨어 시장은 2018년 약 95억달러 규모에서 연 평균 43.4%씩 성장하여 2025년에는 1,186억 달러 규모에 이를 것으로 보인다.[9)]

9) 리테일온 (2019), “인공지능(AI) 시장 글로벌 동향”, http://www.retailon.kr/on/bbs/board.php?bo_table=r1_02&wr_id=100



[그림-1] 세계 인공지능 기술 및 산업의 성장세
"출처: 리테일온 (2019), "인공지능(AI) 시장 글로벌 동향",
http://www.retailon.kr/on/bbs/board.php?bo_table=r1_02&wr_id=1004

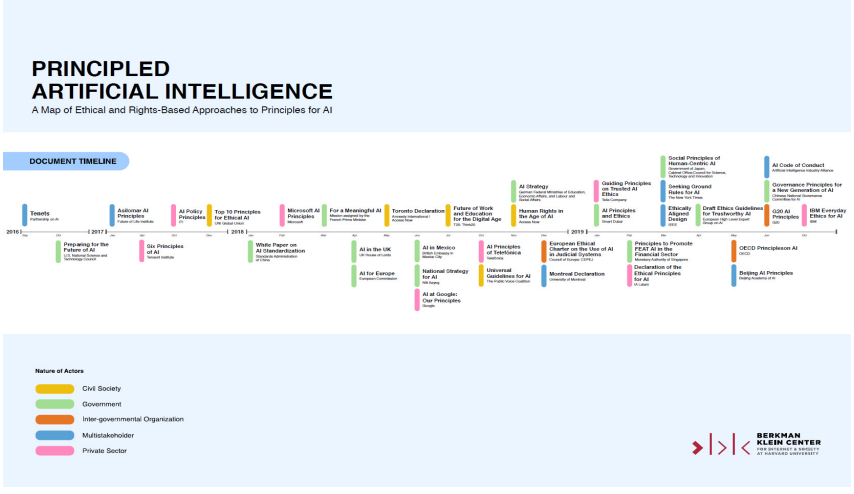
인공지능의 적용 분야는 일반적으로 인공지능 기술 그 자체를 핵심으로 하는 산업과 인공지능 기술을 응용한 산업으로 구분할 수 있으며, 단말 기반인지 클라우드 기반인지 그리고 단말 기반과 클라우드 시반을 어느 정도로 혼합하였는지에 따라서도 인공지능 기술 적용 분야를 구분할 수 있다. 인공지능 기술이 적용되는 분야로 아래와 같이 소비/오락/판매, 교통/사회기반시설, 기업, 오일 및 가스/농업, 공업/군사, 의료/헬스케어가 있으며 이외에도 '4차 산업혁명시대 산업별 인공지능 윤리의 이슈 분석 및 정책적 대응방안 연구(2019)'에 보면, 제조, 에너지, 금융, 유통, 도시, 홈, 웰니스, 법률, 사무관리, 공공분야, 교육, 문화 관광, 안정 등 인공지능이 다양한 사업에 어떻게 적용되고 있는지 잘 설명하고 있다.

[그림-2] 인공지능 기술의 적용 분야

"출처: 리테일온 (2019), "인공지능(AI) 시장 글로벌 동향",
http://www.retailon.kr/on/bbs/board.php?bo_table=r1_02&wr_id=1004

다양한 분야에서 활용될 수 있는 인공지능 기술은 기존 사회 문제를 해결할 수 있는 대안으로 부상하고 있으며 무엇보다도 복지 사각지대에 놓여 있는 사람들의 삶에 긍정적 영향을 미칠 것으로 기대되고 있다. 하지만 인공지능 기술을 주도하는 주체가 주로 기업이므로 인공지능 기술의 혜택을 누릴 수 있는 계층은 소비자로서 기능을 할 수 없는 대상에 한정되기 쉽다. 인공지능 기술이 소비자를 위한 기술이 아니라 최대한 많은 시민들을 포용할 수 있는 기술로 활용되어야 할 것이다.

2. 인공지능 윤리 논의의 성장



[그림-3] 시간에 따른 인공지능 윤리 논의의 등장
"출처: [10] Jessica Fjeld et al., (2020), pp. 18-19."

학술적 논의에서 말 그대로 "인공지능(Artificial Intelligence)"에 대한 논의는 1950년대에 최초로 등장한다.[11] "'인공지능'이라는 개념이 1956년 다트머스 컨퍼런스에서 먼저 공학자들로부터 설정되어서 개념에 대한 탐구를 과업으로 하는 철학자들이 이에 대한 본격적인 연구를 시작하였다고 말할 수도 있겠지만 인공지능의 탄생과 설계의 배경을 형성했다고도 말할 수 있다. 실제로 당시 미국을 지배하고 있던 철학은 사변철학에 대한 염증에서

10) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI(2020), pp. 18-19."

11) McCarthy et al. (2006), p. 12.

비롯된 경험주의적 실증철학이었다.”[12] 그 뒤를 이어 1960년대, 인공지능 윤리에 관한 연구가 등장한다.[13] 인공지능 기술의 발전과 산업의 성장에 따른 현실적 문제로서 인공지능 윤리 논의의 등장과 급성장은 최근 몇 년 사이의 일이다. 인공지능 윤리 논의를 분석하는 최신 보고서에 따르면[14][15], 인공지능 윤리를 직접적으로 포함하는 문건의 88%는 2016년에 발표되었으며 (Anna Jobin et. al., 2019) 인공지능 윤리는 시간에 따라 점차 증가 추세를 보이고 있다. 이러한 추세는 국내 연구 동향에서도 마찬가지이다. 무엇보다도 인공지능 윤리에 대한 논의는 인공지능 철학 내에서 활발하게 논의되고 있다. 김형주 교수가 ‘인공지능 철학 국내연구 동향 분석: 인공지능 철학의 생장점에서(2018)’ 분석한 결과를 보면, 2015년 후반기부터 발표된 논문의 수가 현저하게 증가하였다는 점과 철학 내 세부분과별 논문 발표 건수를 제시하면 인공지능과 관련한 실천윤리적 연구가 다른 분과 연구에 비해 압도적으로 많이 진행되고 있다는 사실을 나타낸다.[16] 다수의 기업에서 주도하고 있는 인공지능에 대한 논의는 상품으로써의 가능 여부, 즉 자율주행자동차가 무엇인가, 의료 인공지능이 무엇인가라는 질문보다는 자율주행자동차를 운행할 때 또는 의료인공지능을 활용할 때 발생할 수 있는 문제들과 그에 대한 대책들이 더욱 중요하다. 그렇기 때문에 지금까지는 인공지능 윤리가 철학/윤리학 영역보다 사회과학/법적/정책 영역에서 더 많이 다루어진다고 보인다. 현재 기술 발전 단계를 고려한/반영한 인공지능 윤리에 관한 논의는 법적 논의를 포함하여 정책 연구, 보고, 제안 및 해외 동향 분석 리포트 등에서 나타나기 때문이다. 또한 KCI에서 ‘인공지능*윤리’ 키워드로 검색했을 때 로봇윤리 등 특정 분야 윤리, 소위 강인공지능 윤리 문제에

12) 김형주 (2019), p. 110.

13) Samuel (1960), pp. 741-742 ; Wiener (1960)

14) Jobin, Ienca, Vayena (2019), p. 391.

15) Fjeld et al. (2019).

16) 김형주 (2018), pp. 154-155, p. 160.

대한 논의는 있지만 현재까지(2020.03) 인공지능 윤리 원칙/선언 자체를 분석한 논문은 없었다.

3. 인공지능 윤리 논의의 변천과 주제

학술 문헌에서 드러나는 초창기 인공지능 윤리 논의는 가상적 개념으로서 인공적 행위자와 그로 인해 초래될 잠재적 문제 등 형이상학적 논의에 가깝다. 인공지능 및 고도로 자동화된 로봇 행위자의 존재론적 지위, 윤리적 책임 귀속 가능성의 여부, 초지능Superintelligence이 야기할 인간 실존의 위협 등이 이에 포함된다. 아시모프의 로봇 3원칙 역시 가상적 개념으로서 인공 행위자인 로봇의 윤리 원칙 논의에 속하는 것이다. 2015년까지도 인공지능 윤리는 고도로 자동화된 기계와 그 윤리성, 자동화로 인한 대량 실업을 주된 논의 대상으로 삼았다.[17)]

2016년 이래, 인공지능 윤리는 현실의 인공지능 기술 단계를 고려하여, 인공지능 기술이 이미 야기했거나 야기할 수 있는 사회적 여파를 중심으로 인공지능과 연관된 윤리적 문제와 대응 방향을 다룬다. 특히 최근 3년간은 인공지능 기술 산업이 성장하면서, 각 산업 분야에서 발생할 수 있는 구체적인 문제와 대응책, 이에 대한 거버넌스 논의가 증가하였다.

현대적인 인공지능 윤리 논의 범주는 인공지능 기술 활용을 위한 원칙 및 행위 지침을 제공하는 규범 윤리 연구, 트롤리 딜레마 등 특정한 인공지능 기술-윤리 문제에 대한 경험적 연구, 인공지능 윤리에 대한 메타 연구 등을 망라한다. 이들이 다루는 윤리적 문제와 윤리적 원리, 가치는 다양하다. 그러나 대부분의 문헌이 공통으로 다루는 윤리적 문제와 가치는 대략 다음과 같이 정리할 수 있다.

이를테면 대부분의 문헌에서 다루고 있는 핵심가치로는 Humanity인간성,

17) 정보통신정책연구원(2018), p. 48.

Privacy프라이버시, Accountability책무성 혹은 해명책임, Fairness공정성, Transparency투명성, Safety안전 등을 꼽을 수 있다. 특히 Accountability, Privacy, Fairness는 전체 가이드라인의 80%에 가깝게 드러나며, '윤리적으로 건전한' 인공지능 시스템을 형성하고 사용하기 위한 최소한의 요건임을 입증하는 듯하다.[18] 이러한 가치들은 인공지능 기술과 관련된 다양한 윤리적 문제들을 함축한다. 첫째, Humanity는 인간적 가치의 지속 혹은 증진의 문제와 관련된다. 인공지능 기술의 발전은 인간의 선(본질적 가치, 안녕, 복지, 존엄, 자율성 등 이것이 정확히 무엇을 의미하는지에 대해서는 의견이 다를지라도)을 지속시키거나 증진시킬 수도 있지만, 매우 극단적으로 유연화된 노동을 강제 받는 '클릭워커Click Worker'[19]의 사례에서 알 수 있듯이 인간의 기본권을 침해할 가능성도 있다. 둘째, Privacy는 개인정보의 보호의 문제와 관련된다. 개인들에게서 수집된 빅데이터를 분석하는 인공지능 알고리즘은 개인들의 은밀한 사생활을 노출시키고 침해할 여지가 있다. 최근 DNA에서 안면 도장에 이르기까지 생체 데이터 보호에 관한 논의가 촉구되는 것은 바로 이 때문이다. 셋째, Accountability는 인공지능 작동의 사후 결과에 책임져야 하는 주체의 문제와 관련된다. 인공지능 시스템으로 인해 어떠한 문제가 발생했을 때 결과물에 대한 책임소재를 밝혀 그 이유에 대해 명료하게 설명할 수 있도록 해야 윤리적 인공지능을 기대할 수 있다. 넷째, Fairness은 차별과 편견 방지의 문제와 관련된다. 인공지능 알고리즘은 현존하는 인간 사회에서 제공하는 데이터를 기반으로 작동하게 된다. 이에 주의하지 않는다면, 인종, 젠더, 지역 등에 관한 사회적 불평등과 차별의 양상이 그대로 분석의 결과물로 나타날 수 있다. 다섯째, Transparency는 Explainability(Explicability)으로 서술되기도 하는데, 인공지능의 의사결정 과정을 인식론적 의미에서 해명하고 이해하는 문제와 관련된다. 인공지능 알고리즘의 작동 과정 및 추론 결과의

18) Hagendorff (2019), p. 3.
19) 홍석만 (2017)

기술적 이해가 투명하게 알려질 때 인공지능 알고리즘에 대한 사회적 신뢰가 쌓일 수 있다. 여섯째, Safety은 인공지능 기술의 견고성robustness 문제와 관련된다. 인공지능 시스템은 의도치 않게 나쁜 결과를 가져올 수 있다. 따라서 인공지능 시스템 내의 오류나 불일치를 처리할 수 있을 만큼의 기술적 견고함이 요구된다.

인공지능 윤리 논의에서 중요하게 다루는 핵심가치 이외에도 인공지능 기술에 대한 논의에서 새롭게 발생하는 문제들임에도 불구하고, 인공지능 기술 논의에서 간과되는 부분들이 있다. 인공지능 기술발달로 인한 다양한 긍정적인 사회적 기대 이면에 기술 발달로 인해 잘 드러나지는 않지만 지불해야하는 비용 및 대가에 해서는 의도적이든 비의도적이든 논의되지 않는 부분들이 있다. 대표적인 문제가 인공지능 기술과 환경에 관한 문제일 것이다. 기후 변화를 넘어선 기후 위기 문제에 대한 논의[20]에서 알 수 있듯이 인공지능 기술은 환경문제에 직간접적으로 피해를 줄 수 있다. 최근 연구에 따르면, 자연어 처리를 위한 AI 모델을 하나만 만들면 60만 파운드의 이산화탄소를 배출할 수 있다는 사실이 밝혀지면서, AI 개발의 기후 영향은 특히 우려되는 영역이 되었다.[21] 또한 인공지능 기술이라는 이름아래 제공되는 화려한 혜택들 뒤에는 '클릭 노동자'들이 존재한다는 사실처럼 현재 우리가 인공지능 기술을 누리기 위해서 숨기거나 잘 보이지 않는 문제들이 무엇인지 인지하고 그 문제를 발굴해야 한다. "상이한 테마들이나 방법들은 서로 관계 없이 따로따로 떨어져 놓여 있어서는 안 되며, 반대로 내면적이고 사실적 연관성 속에서 인식되고, 포괄적인 철학적 논의 속으로 통합되어야 할 것이다."[22]

20) Crawford et al. (2019), p. 47.

21) Crawford et al. (2019), p. 13.

22) 오트프리트 회페, 김시형 역 (2013), p. 12.

4. 인공지능 윤리 논의의 전체 조망 및 분석의 필요성

인공지능 기술 및 산업은 앞으로도 계속 발전할 것이다. 그와 더불어 인공지능 윤리와 관련된 문제도 증가할 것이며 정책적 요구도 높아질 것이다. 따라서 인공지능 윤리 논의와 연구는 현재 기술 발전 및 연구, 관련 상품서비스, 법 정책적 이슈, 다중 이해관계자와 그 이해 충돌 등의 현실을 잘 반영하며 실행으로 이어질 수 있는 토대를 다지고, 구체적 상황에서 실천할 수 있는 행동 방침, 이 적절한 수행을 평가할 평가 원칙 및 방법을 제공해야 할 것이다. 이를 위해서 인공지능 윤리 논의는 최초에 제안된 '원리' 혹은 방향성에 그치는 것이 아니라, 기술발전과 사회적 맥락을 고려하며 지속적인 재검토되고 개정되어야 한다. 그리고 다양한 상황을 포괄하는 대원칙의 개정과 더불어 구체적인 상황의 특수한 문제, 특수한 행위자들을 고려하는 개별적인 인공지능 윤리 이슈에 대한 연구가 동시에 진행되어야 한다.

이를 위해서는 인공지능 윤리를 아우를 수 있는 전체적 조망과 비전, 원칙이 논의되어야 할 것이다. 그러나 '인공지능 윤리'에 대한 메타적 연구는 국내는 물론이고 전 세계적으로도 많지 않은 실정이다. 그중에서도 인공지능 윤리 논의의 '윤리적' 특성에 초점을 맞추어 고찰한 메타적 연구는 더욱 드물다. 이는 인공지능 윤리에 관한 문제의식이 성장한 것에 비하여 대중에게는 물론이고 학문적으로도 '인공지능 윤리'의 정체와 목적, 과제에 대한 이해 및 합의의 기초가 충분히 마련되어 있지 않음을 보이는 증거라 하겠다. 우리는 '지금-여기'에서 요청되며, 앞으로 더욱 중요한 주제가 될 '인공지능 윤리'를 위한 이해와 논의의 토대를 마련하고자 한다. 이를 위해 본 논문에서는 현재의 인공지능 윤리 및 인공지능 거버넌스에 영향을 발휘하는 현실적이고 경험적 자료를 검토한다. 그리고 오늘날 인공지능 윤리 논의의 특성과 한계를 분석하고 인공지능 윤리론의 나아갈 바를 밝히겠다.

Ⅳ. ‘인공지능 윤리’의 정체 : 주요 인공지능 윤리원칙, 선언, 가이드라인 분석

1. ‘인공지능 윤리’에서 인공지능의 이해

인공지능 윤리 논의를 위해서 반드시 필요한 것이 ‘인공지능’ 개념의 규정이다. 그런데 인공지능 윤리 논의는 시작부터 어려움에 부딪힌다. 인공지능에 대한 통일된 규정이 없기 때문이다.[23] 인공지능 개념에 대한 학문적 논의는 20세기 이후 본격화되어 시기와 학자에 따라 다른 개념 및 이해가 제시되기 때문에 학문적으로 통일된 개념이 있다고 보기는 어렵다. 게다가 인공지능은 지금도 발전 중인 기술인데다가 적용 영역에 따라 요구되는 기술의 수준이 다르기 때문에,[24] 현실의 기술 발전에 따라 규정하려는 시도 역시 명시적이고 통일적인 규정을 찾기는 어렵다.[25] 그러므로 대부분의 문헌이 현재의 기술을 기준으로 인공지능을 논의하고 있는 것은 공통적이나 그 규정은 조금씩 다르다.

---

23) OECD (2019), p. 22.

24) Fjeld et al. (2019), p. 11 참조. “The definition of artificial intelligence, or “AI”, has been widely debated over the years, in part because the definition changes as technology advances.4[This is known as the “odd paradox” when technologies lose their classification as “AI” because more impressive technologies take their place. See, Pamela McCorduck, ‘Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence’, 2nd ed. (Natick, MA: A. K. Peters, Ltd., 2004).]

25) 김형주 (2016)는 “이를 인공지능은 “선재적先在的 개념”이라기보다 합의적 개념”이라고 서술한다. 인공지능은 “기술의 발전에 따른 프로그램 혹은 로봇, 그리고 넓게는 이를 다루는 학문의 분야를 추후적으로 가리키는 개념이기 때문”에 개념에 직접 대응하는 대상이나 의미를 규정하는 방식으로는 접근이 불가능하다고 분석한다. 따라서 오늘날 논의 속 인공지능은 “합의적 개념”이다.” 김형주 (2016), pp. 164-165.

간략하게 기계학습 프로그래밍으로 언급AI at Google하는 것부터, 인공지능을 '인공지능 시스템AI system'으로 이해하여 밀접하게 관련된 '인공지능 행위자AI actors'를 고려하는 등 보다 폭넓은 이해와 설명을 명시하는 것OECD까지, 각 문건의 인공지능 규정 서술은 다양하다. IEEE(2019)는 인공지능을 "자율적이고 지능적 시스템autonomous and intelligent systems"[26]으로 이해하면서 '인공지능'의 범위 내에 인간 이상의 능력을 지니는 초인공지능 Artificial Super intelligence까지 고려한다.

우리는 각 문건에서 서술되는 인공지능의 개념 규정을 확인하였다. 그리고 반복적으로 나타나는 서술을 중심으로, 인공지능 개념에 대한 서로 다른 다양한 규정을 포괄할 수 있는 일반적이고 종합적 관점을 찾았다. 이를 종합하면 현재의 인공지능 윤리에서 다루는 인공지능은 다음과 같이 이해하는 것이 적절할 것이다.

● 인공지능: 인공지능은 '인공지능 시스템'으로, 인간이 규정한 목적에 따르는 기계 기반의 자동화된 행위자이다.[27] 인공지능 시스템은 인간의 목적에 따라, 인간에 의해 설계된 소프트웨어 시스템(하드웨어를 포함할 수 있다)을 기반으로, 데이터를 받아들이고 내부적으로 분석하여 외부 환경에 영향을 주는 행위를 수행한다. 시스템은 내적 작동 논리에 따라 데이터를 분석하고 추론하여 현실적(예: 물리적, 사회적, 정신적) 혹은 가상적(예: 디지털 게임 등) 환경에 영향을 미치는 행위(예측, 권고, 의사 결정 등)를 수행한다. 인공지

26) IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019), "Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems First Edition", https://ethicsinaction.ieee.org/

27) 특히 다음의 문헌에서 제시하는 인공지능 규정을 참조하였다. European Commission's High-Level Expert Group on Artificial Intelligence (2019), "Ethics Guidelines for Trustworthy AI", https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai 강조는 저자.

능의 자율성autonomy 수준은 인공지능의 목적에 따라 다르며 다양하다. 여기서 자율성은 인간이 개입하지 않는 자동화automated를 의미하는 것으로, 철학의 오랜 탐구 대상으로서 자율성과 그 함의가 다르다.

이같은 이해를 바탕으로, 우리는 오늘날의 인공지능 논의를 둘러싼 혼란을 잠재우는 기준을 제시하고 유의미한 논의의 경계선을 그릴 수 있다. 작금의 현안으로 인공지능 윤리를 다룬다고 할 때, 인공지능은 무엇으로 이해되어야 하는가?

첫째, 인공지능은 설계, 목적 등에 연관된 인간 행위자 및 그 사용 맥락과 분리불가분의 것으로 이해되어야 한다. 둘째, 인공지능은 단일한 기술적 대상(매체)이 아니라 다양한 기술이 결합된 복합적 시스템으로 이해되어야 한다. 따라서 인공지능에 대한 이해는 인공지능 시스템의 생애주기lifecycle를 함께 고려해야 한다. 인공지능 시스템의 생애주기 단계는 계획, 설계, 데이터 수집 및 처리, 모델 생성, 시스템의 검증과 확인verification and validation, 모델화된 시스템의 실제 활용을 위한 서비스화인 디플로이먼트deployment, 운영 전반인 오퍼레이션과 모니터링operation and monitoring을 포함한다.[28]

이처럼 생애 주기를 갖는 복합적 시스템으로 인공지능을 이해하면, 인공지능에 기술적 문제 이상으로 역사적, 사회적, 문화적 맥락이 중요한 요소임을 알 수 있다. 인공지능과 연관된 인간 행위자 역시 그 범위가 크게 확장된다. OECD 권고안은 인간 행위자를 '인공지능 행위자AI actor'로 명명하여 '인공지능 시스템' 자체와 구분한다. 인공지능 행위자는 인공지능 시스템 생애주기에 적극적 역할을 수행하는 조직과 개인을 모두 포함한다. 이같은 인공지능 행위자의 확장은 인공지능 문제에 대한 도덕적 책임responsibility와 해명책임accountability[29]이 폭넓고 복합적이며, 개인적인 동시에 구조적인 것임을

28) European Commission's High-Level Expert Group on Artificial Intelligence (2019), "Ethics Guidelines for Trustworthy AI", https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. 그러나 권고안은 생애주기의 단계가 반드시 연속적인 것은 아니며 한 단계를 거치고도 다시 이전의 단계로 되돌아갈 수도 있다고 명시한다.

이해할 수 있게 한다. 그리고 우리가 인공지능에 의해 직간접적으로 영향을 주고받는 모든 사람을 포함하는 인공지능의 이해관계자Stakeholders가 누구인지 구체적으로 떠올리고 그려보는 일에 기여한다.[30] 나아가 이같은 인공지능의 이해는 최소한 현 단계의 인공지능 윤리 논의에서 인공지능이 반드시 인간과 같거나 유사한 방식의 행위자일 필요가 없고, 어떤 방식으로 행위하든지 인간 사회를 형성하고 영향을 주고받는 사회 내 행위자임을 인지하게 한다.

2. 인공지능 윤리 논의의 목적 : 인공지능 윤리는 왜 요청되는가?

흥미로운 점은 현재의 인공지능 윤리 논의에 영향력을 발휘하는 주요 문건을 다양하게 검토하였음에도 불구하고 '인공지능 윤리'에 대한 명시적 규정을 찾기 어려웠다는 것이다. 이는 오늘날 인공지능 윤리 논의의 현상적 특징일 뿐만 아니라, 인공지능 윤리를 이해하고 접근하는 인공지능 윤리 논의 주체들의 입장과 의도를 보여주는 지점이기도 하다. 이에 대해서는 각 문건의 특징을 비교, 검토하는 이후의 문단에서 상술하겠다.

그러나 논의의 진행을 위해 먼저 광의의 '인공지능 윤리' 개념을 기술하도록 하겠다. 우리는 이를 위해 인공지능 윤리를 서술하는 다양한 언급을 검토, 종합하였다. 무엇보다 해당 기술description이 인공지능 윤리를 둘러싼 다양한 이슈, 논의 주체, 논의 목적을 포괄할 수 있고 인공지능 윤리의 현재적 특성을 보이는 동시에 특정 윤리 이론에 치우치지 않도록 주의하였다.

---

29) 현재 accountability는 국내 인공지능 윤리 논의에서 주로 '책무성'이라는 용어로 번역된다. 그러나 우리는 이 용어가 '책임'을 의미하지만 인격적 주체의 도덕적 책임responsibility과는 또 다른 책임이라는 점을 강조하기 위해 '해명책임'으로 번역하였다. 이는 특정한 작동 방식 혹은 사건 등을 설명하고 대응할 책임이 누구에게 있는지를 의미한다.

30) 인공지능 행위자는 인공지능 이해관계자의 하위 집합이다.

● 인공지능 윤리AI Ethics : 인공지능 윤리는 인공지능 연구, 개발, 적용, 폐기 등과 관련한 인공지능 기술의 전체 단계에서 인공지능을 둘러싼 규범적 문제를 연구한다. 근본적인 가치, 방향성, 원칙의 문제와 더불어, 인공지능이 사회에서 윤리적으로 활용되는 것을 보장하기 위한 응용윤리적 논의 등 다양한 층위의 윤리적 연구를 포함한다. 실천적 측면에서, 인공지능 윤리는 사회에 미치는 윤리적 영향, 인간의 기본권 및 근본적 자유와 관련된 함의를 다루고, 설명틀framework과 가이드라인을 제시하는 데 초점을 맞춘다.

인공지능 윤리는 다양한 부문과 기관을 포괄한다.[31] 인공지능의 주된 논의 초점 및 구체적인 이슈는 3장(Ⅲ.3.)을 참조하라. 일반적으로 인공지능 윤리는 기술적 특성, 광범위한 사회 내 도입으로 인한 사회적 여파, 위험 및 부작용, 전대미문의 기술과 그 부정적 영향에 대한 두려움에서 출발한다.[32] 인간의 개입이 없는 자동화된 시스템, 알고리즘의 불투명성, 귀결의 불확실성 등, 인공지능 기술의 특성은 기존의 기술 논의를 다루는 근대적인 개별 책임 논의로 다루기가 어렵다.[33] 또한 사회의 다양하고 광범위한 영역에 인공지능 기술이 도입되면서, 소위 '4차 산업혁명'으로 표현될 만큼 다양하고 커다란 변화가 나타날 것이라는 예측이 팽배하였다. 그로 인해 새로 생기는 부작용, 위험, 그리고 이전에 있었던 부작용이나 위험이 더욱 커질 것이라는 사회적 우려 역시 커졌다. 이러한 문제에 대응하기 위해 규범적 논의가 요청된 것이다.

우리는 분석 대상인 문헌 내에 서술된 논의의 목적과 '윤리'라는 단어의

31) Whittlestone (2019, January), p. 195.

32) 이같은 분석은 우리가 검토한 대부분의 자료에서 공통적이다. 특히 손화철 (2018), 정보통신정책연구원 (2018), EU (2019)European Commission's High-Level Expert Group on Artificial Intelligence (2019), "Ethics Guidelines for Trustworthy AI", https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, OECD (2019)를 참조하라.

33) 허유선 (2018), pp. 165-209. 참조.

사용을 검토하여, 인공지능 윤리를 다루는 문건들이 윤리를 어떻게 이해하고 접근하는지를 고찰하고자 한다. 이를 통해 전체적인 경향과 차이점을 서술하고, 오늘날 '인공지능 윤리' 논의가 급증한 원인을 분석하려 한다. 먼저 공통적인 경향이다.

첫째, 분석 대상이 되는 문헌들은 인공지능 윤리를 다루고 요청하면서도 대부분은 '인공지능 윤리'가 무엇인지는 명확히 규정하지 않는다. 인공지능 윤리가 무엇인가에 대한 정의를 명확하게 한 보고서는 EU의 가이드라인 밖에 없다. 대부분의 보고서 및 윤리원칙들(예: Deepmind, Partnership on AI, Google, Microsoft, IBM, OECD의 원칙 등)은 인공지능 윤리에 대한 정의를 명시하지 않고 있다. 이는 분석대상 문헌이 대부분 '인공지능 윤리' 자체를 연구 대상으로 삼는 학술 문헌이 아닌 까닭도 있다. 그러나 '인공지능 윤리'를 다루는 문건에서 개념 정의가 일반적 수준에조차 명시적으로 서술되지 않는 것은 인공지능 윤리를 혼란스럽게 만드는 원인 중 하나가 된다. 인공지능 윤리 논의의 참여 주체가 '인공지능 윤리'라는 개념을 어떻게 이해하는지, 어떤 윤리적 입장을 취하는지를 파악하기 어렵기 때문이다.

둘째, 인공지능 윤리가 표방하는 가치는 대표적으로 행복, 복지, 자유 등이다. IEEE Version2 문서에서는 기술 혁신이 이룬 혜택에 대한 숙고가 엿보이며, 혜택의 대상, 혜택의 기준등이 인간의 행복과 위배되지 않도록 기술 혁신의 경로를 찾고자 한다. IEEE Version2 문서, EU 가이드라인의 목적은 지능적이고 자율적인 시스템과 기술을 윤리적이고 사회적으로 구현하여 주어진 문화적 맥락에서 인간의 복지를 우선시하는 정의된 가치와 윤리적 원칙에 맞추는 방법에 대한 공개 토론을 진행하고자 한다. IEEE Version2 문서, EU의 가이드라인, Rome Call은 인공지능 기술이 인간의 자유 및 자율의 증진에 기여하기를 기대한다.

셋째, 오늘날 인공지능 윤리 논의는 주로 실용적 논의에 초점을 맞춘다. 윤리 가이드라인은 다양한 인공지능 주체의 행위 지침으로 활용된다. 구글에서 발표한 원칙은 이론적인 개념이 아닌 실질적으로 연구하고 제품을 개발하

는데 있어서 영향을 줄 구체적인 표준들에 초점을 두고 있다.

그러나 공통점 외에 차이점도 있다. 오늘날 인공지능 윤리 논의의 네 번째 특징이기도 하다. 각 문헌들은 발간 주체에 따라 기술에 대한 입장 차이를 보인다. 기업이 주체가 된 인공지능 관련 원칙들은 인공지능 기술에 대한 긍정적 확신을 전제하고 있는 것에 비하여, 비영리단체Partnership on AI 또는 학계에서 주장된 인공지능 원칙에서는 인공지능 기술이 사회에 초래할 긍정적 영향뿐만 아니라 부정적 영향을 인지하고 논의를 시작한다. 또한 AI policy에서는 인공지능 기술이 현재 다양한 사회적 문제들을 해결해 줄 수 있는 것이라고 기대하는 바가 크다. 인공지능 기술에 대한 평가의 차이는 인공지능 윤리에 대한 인식에도 영향을 미칠 것이다. 윤리적 논의로서의 성격이 더욱 두드러지는 문건도 있었다.

다섯째, IEEE version2 문건은 다른 문건들과 달리, 인공지능 윤리의 구체적 실행을 위한 선결 조건으로서 윤리와 기술의 관계를 예리하게 파악하고 있다. 이들은 기존의 윤리 및 철학 논의를 토대로 하여, 윤리적 결과의 도출은 시스템의 결정에 영향을 받는다는 것을 전제로 한다. 그러므로 인공지능 시스템의 생애 주기 전체에 걸친 윤리적 절차의 중요함을 강조한다.

여섯째, 다른 인공지능 원칙이나 권고안에 비하여, 몬트리올 선언은 인공지능이 공동체의 공공선을 이룰 수 있도록 돕는 나침반의 바늘 역할이 되어야 함을 전제하고, 강조한다. 몬트리얼 선언의 각 원칙은 인공지능의 발달이 윤리적 도전과 사회적 위험을 포함할 수 있으며 공동체의 정치적 조직 방식에 영향을 미칠 수 있다는 인식을 기반으로 한다.

오늘날 인공지능에 관한 논의는 우선, 기업의 자발적인 참여나 협력보다는 정부나 의회의 주도 아래 진행되어 왔다. 게다가 기업이 논의에 참여·협력하더라도 인공지능 기술이 초래할 위험들은 간과하는 일반적인 경향이 있다. 인공지능 기술에 대한 맹목적 확신을 바탕으로 인공지능 기술의 도입과 발전을 당연한 것으로 전제하고 있는 것이다. 이러한 배경에는 물론 인공지능 기술이 현재 인류에게 산적한 사회적 문제들을 해결할 수 있다는 '윤리적인'

기대가 담겨 있는 것이 사실이다. 하지만 실제로는 개별 기업이나 국가의 산업적, 경제적 이익을 증진하려는 목적만이 최우선시 되고 있지는 않은지 의구심을 가질 수 있다. 인공지능에 대한 논의는 주로 법제화하기 어려운 부분을 윤리권고안에 따른 '자율규제'로 대체하려는 경향이 두드러진다. 이는 윤리를 법이나 제도와는 분리된 별개의 수단 또는 절차나 의례로 단순화·형식화하려는 발상이며 시도이다. 그러나 이 같은 발상과 시도가 인간 사회와 공존 가능한 인공지능 기술을 사회에 기여하는 방향으로 이끌 수 있는지 의문이다. 오히려 기업이 기술을 신속하게 상품화하는 데 요청되는 형식적인 통과의례로서 인공지능 윤리 개념이 소비되고 있는지 반문해보아야 한다. 인공지능에 관한 '윤리적' 논의가 인공지능 윤리 개념의 외연과 함축을 협애화하고 그 개념을 한낱 수단이나 장식으로 전락하게 하는 역설적인 상황이 벌어지고 있는 것은 아닐까? 실제로는 인공지능이 야기하는 사회적 윤리적 문제들로부터 기업이 그 책임을 회피하는 데, 나아가 책임의 면제를 당당하게 요구하는 데 기여할 수도 있는 것이다. 앞서 살펴보았듯, 전부 그렇지는 않지만 몇몇 문건을 제외하고는, 윤리 원칙 등 인공지능에 관한 논의 대다수가 윤리란 무엇인가에 대한 이해, 그리고 윤리학적 개념의 사용 등에서 한계를 노정하고 있다. 윤리적 논의가 과연 통약 가능성에 기초해서 이루어지고 있는지 의심되는 상황이며, 이것이 '인공지능 윤리는 왜 요청되는가?' 하고 다시 묻는 까닭이다.

3. 인공지능 윤리 논의의 특성에 대한 비판적 고찰

우리는 윤리적 논의라는 점에 초점을 맞추어서 현재의 인공지능 윤리 논의의 경향과 특징을 다음과 같이 고찰하였다.

- 인공지능 윤리 논의에서 용어의 중복 및 혼란
- 인공지능 윤리 논의 주체의 제한성, 다양성 결여
- 인공지능 윤리의 수단화
- 면죄부로 작용하는 인공지능 윤리, 실효성 우려
- 윤리의 제한적 이해
- 거시적, 추상적 원리의 한계
- 학술적 논의의 부족

첫째, 현재의 인공지능 윤리 논의는 비슷한 내용을 서로 다른 용어를 사용하여 논의한다. 이로 인해 인공지능 시스템에 대한 전문적 지식을 갖추지 못한 다양한 이해관계자가 인공지능 윤리 원리를 이해하기 어려울 수 있다. 현재 인공지능 윤리는 원리의 수준에서조차 유사한 내용이 다른 용어로 서술된다. 예를 들어, 인공지능의 의사 결정 근거 및 그 과정을 이해하는 과제는 문건에 따라 설명가능성Explicability, 투명성Transparency, 해명책임Accountability 등의 용어로 다르게 서술된다. 예를 들어 베이징 인공지능 원리는 “투명성”과 “설명가능성”을 모두 사용하고, IEEE는 “투명성”과 “해명책임”을 함께 사용하며, 파트너쉽은 “이해가능하고 해석가능한 understandable and interpretable”이라는 표현을 쓴다. 한편 로마콜은 “투명성”이라는 용어를 사용하며 이를 다시 “설명가능해야 한다”고 서술하고 있다. 우리는 이들이 지시하는 바와 강조하는 지점이 다르고, 따라서 이같은 용어의 차이가 유의미하다고 판단한다.[34] 그러나 설령 그 차이가 유의미하다고 할지라도 이같은 현상은 혼란과 모호함을 야기할 수 있다.[35]

비슷한 내용이 여러 주체에 의해 반복적으로, 그리고 서로 다른 개념을 통해 조망된다는 점 자체는 ‘문제적’이지 않다. 이는 오히려 이전에 없던 학술적, 실천적 논의를 위해 거쳐야 하는 과정일 수 있다. 문제가 되는 것은

34) 플로리디는 “투명성과 해명책임을 “설명가능성”으로 종합하며, 인식론적 요구와 윤리적 요구라는 문제 제기의 관점에 따라 다르게 표현되는 것이라고 해석한다.” Floridi, Cowls (2019). p. 8 참조.

35) Floridi, Cowls (2019). p. 2.

이로 인해 인공지능 윤리가 무엇을 의미하는지 전체적으로 조망하고 이해하는 일이 어려워지며, 그 결과 인공지능 윤리의 실천이 어려워진다는 점이다. 우리가 논의하는 '인공지능', '인공지능 윤리'의 정체를 선명하게 가시화하기 어려운 것이다. 서로 다른 용어를 사용하는 이해관계자 사이의 상호 소통에 많은 시간과 에너지가 요구되기 때문이다. 전문가들의 논의는 물론이고, 권고안을 실천해야 할 다양한 처지의 이해관계자가 '인공지능 윤리'에 접근하는 일조차 어렵게 만드는 것이다. 그로 인해 사람들은 인공지능 윤리에 대한 총체적 이해를 쉽게 포기하고, 윤리를 단순한 이행 '명령'의 체크리스트로 간주할 수도 있다.

그러나 이러한 문제를 제기하는 것이 인공지능 윤리에 대한 논의가 '너무 많다'거나 그렇기 때문에 향후 인공지능 윤리에 관한 연구 및 논의가 축소되어야 한다는 뜻은 아니다.[36] 윤리를 우리 삶의 결정과 행위를 지도하며 반영하는 핵심가치와 그에 관한 사유로 이해할 때, 윤리가 '너무 많다too much'는 말은 어불성설이다. 더욱이 오늘날 인공지능 윤리에 대한 연구, 제안 등은 향후 더 많은 논의와 후속 연구를 요청하는 시작 단계라 할 수 있다. 인공지능 윤리에 관한 논의는 더욱 활성화될 필요가 있으며 논의의 기회와 장은 더욱 확대되어야 한다. 논의의 접근 수준도 더욱 다양화되어야 할 것이다. 그러므로 인공지능 윤리의 위와 같은 특징은 오히려 인공지능 윤리 논의의 활성화와 실질적 영향력의 확대를 위한 과제로 이해되어야 한다. 다양한 이해관계자가 '인공지능 윤리'에 접근, 이해, 실행할 수 있도록 인공지능 윤리의 간명하고 가시적인 조망과 이해틀framework이 제공되어야 한다. 물론 이 과정 및 결과에는 전 세계적인 합의가 수반되어야 할 것이다.

둘째, 현재 인공지능 윤리 논의의 주체는 제한적이며, 논의의 다양성이

36) 인공지능 윤리 논의의 '과잉'에 대한 우려로는 다음과 같은 관점이 있다. "인공지능 윤리 이슈에 대한 선행적·사전적 논의의 필요성에도 불구하고 윤리 논의의 과잉 자체가 심리적 규제효과로 작용하여 인공지능 개발 및 산업 분야의 이해당사자들이 인공지능 윤리 문제에 자발적, 적극적인 관심과 참여를 위축시킨다는 점이다." 정보통신정책연구원(2018), p. 10.

부족하다. 따라서 전 지구적 차원의 다양한 이해관계자의 입장이 공정하게 반영되었다고 보기 어렵다. 현재 인공지능 윤리는 다양한 분야에서 논의되며, 정부, 국제기구, 종교계, 전문가 집단, 기업 등 다양한 행위 주체가 참여한다. 그리고 이 논의들은 보편적인 윤리원칙 및 가치에 기반하기 때문에 "전 지구적 합의"[37] 혹은 그러한 합의를 위한 기초 자료로 이해될 수 있다. 그러나 실제로 인공지능 윤리 논의를 주도하는 곳은 정부와 기술 기업이다. 또한 분석 대상의 발행처 목록에서 알 수 있듯이, 인공지능 윤리의 주요 문건은 대개 서구권Western의 경제적으로 발전한 국가를 중심으로 발표되었다. 이는 아프리카, 남미와 중미, 중앙아시아 등, 상대적으로 과소대표되고 있는 지역이 있음을 보여준다.[38] 지리적 과소대표는 해당 지역 내 국가의 이익이 충분히 반영되지 않을 가능성만이 아니라, 인종, 문화, 그 외 로컬의 특수성 등이 인공지능 윤리원칙에 공정하게 반영되지 않을 가능성을 보이는 증거이기도 하다.

AI Now는 2019년 보고서에서 서구권을 중심으로 발표된 인공지능 윤리 논의의 젠더 편향을 지적하고 있다. 하겐도르프Hagendorff의 인공지능 윤리 가이드라인 메타 연구는 인공지능 윤리 담론의 대부분이 백인 남성을 중심으로 형성되며, 이것이 인공지능 윤리 담론의 이슈 및 접근 방식을 다양화하지 못하게 만드는 원인 중 하나라고 분석하기도 한다.[39] 현재 인공지능 윤리 논의를 주도하는 서구권의 조직, 특히 기술 업계는 대다수가 부유한 백인 남성으로 구성되어 있으며, 이같은 주류에 속하지 않는 사람들의 목소리는 충분히 반영되기 어렵다. 문제는 인공지능 기술의 윤리적 문제에 가장 큰 피해를 입을 수 있는 집단은 이처럼 논의에 쉽게 접근할 수 없고 따라서 자신의 이해관계를 충분히 대변할 수 없는 사회적 취약계층, 경계 집단

37) Crawford et al. (2019), p. 19.

38) Crawford et al. (2019)과 Jobin, Ienca, Vayena (2019)는 이러한 문제를 지적하고 있다.

39) Hagendorff (2019), pp. 3-4.

marginalized group이라는 점이다. 인공지능 윤리 논의가 양적으로 증가하고 있으며, 원칙의 수준에서 전 세계적이고 보편적인 가치 추구를 표방하는 것과는 대조적으로 현재의 인공지능 윤리가 전 지구적 차원에서 다양한 행위자의 이해 관계를 공정하게, 그리고 충분히 고려하고 있다고 할 수는 없을 것이다.

셋째, 현재의 인공지능 윤리 논의는 윤리를 특정 목적을 달성하기 위한 수단으로 간주하는 경향이 우세하다. “윤리적 대응이 산업 성장을 위축시키지 말아야” 한다는 식이다.[40] 이는 인공지능 윤리 논의를 주도하는 주체가 주로 각국의 정부, 관련 기술 기업인 까닭이 클 것이다. 앞서 인공지능 윤리 논의의 목적과 ‘윤리’라는 용어 사용의 고찰에서 살펴본 것처럼, 국가 및 기업의 인공지능 윤리 논의는 주로 국가 경쟁력 확보, 산업의 사회적 정착을 위한 기반 조성 등에 초점을 맞춘다. 이러한 관점에서 인공지능 기술의 발전은 당연한 사실이자 당위적 가치로 간주된다. 따라서 그들이 목적하는 가치의 함의, 인공지능 기술의 발전에 따른 이익 및 사회적 대가의 분배 등은 충분히 검토되지 않는다.

윤리학은 원칙과 기준을 제시하는 동시에 그를 실천하기 위한 방법과 도구까지도 제안한다. 그러므로 원칙으로서의 윤리와 도구로서의 윤리는 양립가능하다. 그러나 원리, 원칙, 가치에 대한 근본적 검토, 정당화, 합의 절차에 대한 논의가 없는 도구는 비윤리적일 뿐만 아니라 실효성을 획득하기 어렵다. 잘못 설정된 목적에 기여하는 도구는 결과적으로 쓸모를 다하지 못할 것이기 때문이다.

기술 연관 윤리 논의는 가치의 실행을 위한 방법론뿐만 아니라, 제시된 가치의 정당성과 적절성을 검토하는 것 또한 주요 과제로 삼는다. 그러므로 윤리적 논의의 차원에서 우리는 인공지능 윤리에 대해서도 아주 단순하고 불가피한 물음을 제기할 수 있다. ‘과연 인공지능 기술은 당연히 발전해야만

40) 정보통신정책연구원 (2018), p. 224.

하는 것인가? 당연히 발전할 수밖에 없다고 말하며 그 수용을 주도하는 주체는 누구인가? 목적의 내용, 범위, 설정 과정은 충분히 윤리적인가?' 우리는 수단을 정당화하는 목적을 검토해야 할 것이다.

인공지능 윤리를 수단화하는 태도는 역설적으로 '인공지능 윤리'를 수단으로 활용하는 일조차 어렵게 만든다. 그러한 입장은 인공지능 윤리원칙 및 권고안의 실질적 영향력을 약화시키기 때문이다. 이들이 향후 더욱 구체적이고 상세하게 제시된다고 할지라도, 명시적으로 혹은 숨겨진 다른 목적을 우선한다면 인공지능 윤리는 간과될 것이다. 다른 목적이 우선일 때, 윤리라는 '수단'을 변경하는 것은 어렵지 않은 일이다. 그러므로 인공지능 윤리는 그 수단적 역할마저도 충실히 이행하기 어려워지며, 실효성을 의심받게 된다.

넷째, 지금까지 서술한 주요 참여 주체 및 논의 목적과 관련하여 인공지능 윤리원칙 및 권고안은 특히 기술 기업에게 일종의 면피적 장치로 기능할 위험을 안고 있다. 기술 기업이 인공지능 윤리 논의를 주도하며 적극적으로 참여하는 것은 그만큼 인공지능 기술의 윤리적 문제에 대한 우려가 일반적이며 충분히 숙고할 필요가 있음을 보여준다. 그러나 최근의 분석[41] 은 기업의 인공지능 윤리에 대한 적극 참여가 기업의 실질적 책임 이행을 강제하지 못하고 오히려 면죄부로 작용하고 있다고 지적한다. 기업은 국가의 직접 규제 및 경성 규범hard law을 피하고 기업의 자율 규제 등 해당 기업에 유리한 것만을 취한다는 것이다. 기업은 윤리원칙을 발표하고 윤리 위원회를 구성하기도 하지만 기업이 실제로 윤리 가이드라인을 어떻게 실천하고 있는지, 어떤 과정을 거쳐 어떤 점을 문제삼아 대응책을 정하게 되는지는 공중에게 투명하게 드러나지 않는다.[42] 이들을 감시하고 관리할 실행력 있거나 독립적

41) Jobin, Ienca, Vayena (2019); Crawford et al. (2019) ; Vincent (2019), "The problem with AI ethics", https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech.

42) "구글은 2019년 "책임있는 인공지능 개발"을 위해 외부인이 참여하는 윤리위원회를 만든다고 발표하였다. 그러나 위원회에는 LGBTQ 차별금지법에 반대하는

인 기관도 없는 실정이다. 또한 기업은 인공지능의 윤리적 문제의 기술적인 것에 초점을 맞추면서, 기술 발전에 시간과 수고가 필요함을 이유로 자신들의 책임 이행을 연기한다. 그리고 광범위하고 복합적인 사회적 맥락을 고려해야 하는 인공지능 윤리의 문제를 단지 기술적인technical 문제로 축소시킨다. 기업의 자율 규제 이외에 다른 법적 규제를 만드는 것이 불필요하다는 주장에 힘을 싣는 것은 물론이다. 이는 실제로는 정치적이고 사회적인 성격을 띄는 관련 결정 및 권한이 기술 기업에게 이양되는 것을 자연스러운 일처럼 생각하게 만든다. 인공지능 윤리가 단지 기술적 문제라면, 기술전문가만이 혹은 기술전문가를 중심으로 논의가 주도될 수밖에 없기 때문이다. 해당 기술의 개발과 활용이 기본 인권, 시민권, 정치 권력의 배분에 미치는 영향은 가리워진다.

플로리디는 최악의 경우, 인공지능 윤리 논의의 증대가 일종의 윤리 시장The Ethics Market으로 이어질 수 있다고 경고한다. 인공지능 윤리 논의의 양적 증가는 이해관계자가 가장 매력적으로 느끼는 원리를 '구매shop'하는 '원리 시장market for principles'으로 전락할 위험을 안고 있다는 것이다.[43] 이는 윤리 원리 및 권고안이 기업이 실제로 이행하게끔 관리, 감독할 수 있는 구속력 있는 규제로 이어지지 않고 있기 때문이다. 이제 인공지능의 윤리 논의는 윤리 원리 혹은 선언이 기술 기업의 "윤리적 세탁ethics washing" 수단이 되지 않도록 경계할 수 있는 실질적 장치를 고민해야 할 것이다.[44]

캠페인을 벌이는 멤버[Kay Coles James]가 포함되었으며 많은 연구자와 참여자들이 이를 공개적으로 반대하는 성명을 발표하였다. 그러나 구글은 이에 대해 어떤 언급도 하지 않았다. 마이크로소프트는 자사의 인공지능윤리 감독위원회의 권고를 받아 중요한 판매를 중단한다고 밝혔지만, 무엇이 어떠한 이유로 윤리적 문제가 되는지는 밝히지 않았다. 이들 위원회가 기업의 의사결정에 어떤 방식으로 얼마만큼 영향을 미치는지는 공개되지 않는다. 생명윤리분야의 유사한 활동에 비하면 기업의 인공지능 윤리 위원회의 활동 및 영향력은 놀랄 만큼 불투명한 것이다. "Vincent (2019), "The problem with AI ethics", https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech. 참조.

43) Floridi (2019)

다섯째, 이는 인공지능 윤리 논의에서 윤리 원리의 한계를 보여준다. 인공지능 윤리 원리 혹은 권고안이 실제로 기업의 변화를 강제할 수 없고 오히려 기업의 면죄부로 활용된다는 것은 원리와 실행 사이의 간극을 드러낸다. 최근의 연구는 공통적으로 원리와 실행 사이의 간극을 줄이기 위한 논의가 필요함을 지적하고 있다.[45] 원리만으로는 실천을 보장할 수 없기 때문이다.

그러나 이것이 원리 자체의 무효성을 의미하는 것은 아니다.[46] 원리는 구체적이고 다양한 맥락, 행위자를 모두 포괄해야 하는 것이고, 따라서 추상적일 수밖에 없다.[47] 원리의 본성상, 다양한 구체적 상황과 그 행위자에게 딱 맞는 방침을 전부 제시할 수는 없는 것이다. 이런 이유로 인공지능 윤리의 실효성은 의심받지만, 이것은 포괄적 원리와 맥락 및 상황의 특수성 사이에서 항상 발생하는 간극이며 비단 '윤리'만의 문제는 아니다. 그러나 원리와 실행 사이의 간극을 최대한 줄이는 일이 추후 인공지능 윤리 논의의 중요한 과제임은 틀림없다. 인공지능 윤리 논의는 향후 원리 자체에 대한 논의 및 합의, 그리고 그 원리를 구체적 맥락에서 현실의 이해관계자가 이행하는 구체적 실천 방법에 대한 연구를 동시에 진행해야 할 것이다.[48]

여섯째, 최근의 양적 성장에 비해 인공지능 윤리 논의에서 학술적 연구가 차지하는 비중은 크지 않다. 인공지능에 대한 논문 및 컨퍼런스는 계속 증가

---

44) Vincent (2019), "The problem with AI ethics", https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech.

45) Jobin, Ienca, Vayena (2019). Crawford et al. (2019), Vincent (2019), "The problem with AI ethics", https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech., Hagendorff (2019) 등을 참조하라.

46) Whittlestone et al. (2019), p. 195 참조.

47) Hagendorff (2019), p. 9.

48) 유럽의회는 "먼저 발표된 인공지능 윤리 지침을 보완하는 것으로서, 인공지능 기술 제품의 설계, 배포, 개발, 구현과 관련한 해설인 인공지능 윤리 가이드의 맥락과 실행에 관한 보고서를 발간하기도 하였다." Madiega, T. (2019) 참조.

추세이며 AI 윤리에 관한 논의 역시 1960년대 이래로 계속 진행 중이다. 그러나 본 논문에서 다루는 대상인 현실적인 기술 수준에서의 인공지능 윤리 일반 혹은 인공지능의 사회적, 윤리적 여파를 연구하는 학술 논문은 여전히 많지 않은 상황이다. 이는 인공지능 윤리 논의가 기술과 정책적 필요성에 의해 추동됨을 보여준다. 현재 인공지능의 윤리적 문제에 대한 해법으로 주로 기술적 접근이 제안된다. 위에서 서술한 '윤리'의 실질적 영향력과 유효성에 대한 의심에 더해 기술의 문제는 기술로 해결할 수 있다는 입장이 결합된 것이다. 특히 프라이버시, 해명책임, 안전 등의 문제에 대한 기술적 접근은 더욱 두드러지며, 이 문제에 대응하는 기술적 해법에 대한 연구는 이미 착수되었다.[49]

그러나 '인공지능 시스템'의 이해에서 알 수 있듯이, 인공지능과 그 사회적 여파는 기술적인 문제로만 국한할 수 없으며 구체적인 적용 및 사용의 정황, 사회문화적 맥락, 권력 관계 등 보다 광범위한 관점을 요구하는 복합적 문제이다. 기술이 발전 중임을 고려할 때, 기술을 둘러싼 혹은 기술이 야기하는 문제는 기술적 대응으로 해결하여 '제거'할 수 없다. 이들은 끊임없이 새로운 문제를 야기하며 부단하고 다각적인 대응을 요청할 것이기 때문이다.

기술의 발전과 동시에 위험 및 부작용도 생겨난다는 것, 그리고 인공지능 기술의 복합적 성격과 사회에 대한 잠재적 영향력을 고려할 때 인공지능 윤리에 대한 학술적 연구 자료의 축적과 활발한 논의는 시급히 그리고 지속적으로 요청되는 것이다.[50] 어떤 것이 '문제시'할 만한 것이고, 그동안 논의의 사각지대였으나 시급하고 중요하게 논의되어야 할 것, 문제의 복합적 성격을 밝히고 공중의 이해를 위한 토대를 제시하는 것은 기술에 대한 사실 판단이 아니라 기술-사회-가치의 연관 및 그 관계를 읽어내는 해석이자 가치 판단이기

49) Hagendorff (2019), p. 3.

50) 울리히 벡은 "현대기술의 위험은 단지 기술적인 문제가 아니라 사회적으로 위험을 어떻게 규정하고 구성하는지의 문제임을 지적한 바 있다." 울리히 벡 (1998) 참조.

때문이다.[51] 또한 인공지능 논의의 중복 및 서로 다른 용어의 혼란 등을 해소하기 위해서도 인공지능 윤리 논의에 필수적인 기초 개념의 정련, 전체적 조망의 제공, 다양한 관점의 적절성과 정당성을 검토하는 학술 연구가 더 많아져야 할 것이다.

마지막으로, 현재 인공지능 윤리 논의는 '윤리Ethics' 및 기술-가치 관계에 대한 제한적인 이해를 보여준다. 먼저 현재의 인공지능 윤리 논의는 그 이름과 달리, '윤리'적 고찰이 충분히 반영되지 않은 것처럼 보인다. 인공지능 윤리 논의는 일부(EU의 가이드라인 등)를 제외하고 '윤리' 혹은 '윤리적인 것'이 무엇인지 명확하게 규정하지 않는다. 대신 인공지능 윤리의 대다수는 '사회를 위한[for society]' 혹은 '인간 중심[Human-centered]'의 기술을 말하지만 그것이 구체적으로 무엇인지를 제시하는 문건은 많지 않다. 국가 전략이나 기업의 논의에서는 이같은 특징이 더욱 두드러진다. 또한 주된 윤리적 가치를 누가, 어떤 과정을 거쳐 결정하고 있는지, 인공지능 윤리 원리의 민주적이고 절차적으로 정당한 논의 도출 과정이라는 문제를 어떻게 다뤄야 하는지에 대한 언급 및 고려는 극히 드물다.

이미 제시된 가치의 적절성과 정당성 역시 재검토될 필요가 있다. 대부분의 문헌은 인공지능 기술의 발전 및 사회 도입으로 인한 세계적인 변화를 곧 도래할 '사실'로 간주하지만 그러한 변화와 더불어 인간 혹은 사회에 대한 규정이 변화할 가능성, 추구 가치의 변형 및 확장 가능성 등은 충분히 고찰하지 않는다.[52] 이들이 주로 제시하는 가치는 행복, 복지, 자유 등의

51) 허유선은 "인공지능 알고리즘에 의한 편향과 차별에 대한 논의에서 이 문제가 단지 기술적인technical 문제가 아니며, 이 문제는 차별을 어떻게 규정하고 해석할 것인지, 누구의 어떤 차별 경험을 가시화할 것인지 등의 문제화 및 해석에 따르는 철학적, 윤리적, 정치적 문제임을 논한다." 허유선 (2018), pp. 165-209 참조.

52) 대한민국 정부 (2019)는 인공지능 기술로 인한 변화를 "문명사적 변화", "모든 영역의 패러다임 변화 초래"로 규정하고, 3대 과제 중 하나로 '사람 중심의 인공지능 구현'을 제시한다. 그러나 문명사적 변화 및 모든 영역의 패러다임 변화는 당연한 것으로 전제하고 거대한 변화 앞에 놓인 공동체의 가치 지향에 대해서

전통적 가치이다. 이는 여전히 중요한 가치이지만 이제는 전통적인 가치의 변형, 확장을 고려할 때이다. 사회와 삶의 방식이 그들 문건이 예상한 만큼 크게 달라진다면 인간에 대한 이해 및 사회를 위한 가치의 규정과 내용 역시 달라질 수밖에 없기 때문이다. 기술의 발전과 함께 인간의 정체성을 형성하고 규정하는 방식 역시 변할 수 있다. '인간의 자율성' 등 기존 가치의 외연 및 구현 방식 또한 확장되거나 변형될 것이다. 그러므로 전통적인 인간의 가치는 기술과 함께 상호영향을 주고받으며 형성된 인간과 사회를 고려하며 기술 등 주요한 비인간 행위자, 생태계 전체를 포함하는 등, 확장된 가치로서 숙고되어야 할 것이다.

또한 현재의 인공지능 윤리 논의는 주로 기술, 법・정책, 윤리 분야를 중심으로 구성된다. 이는 윤리를 기술 및 법(정책)과는 전적으로 구분가능한 것으로 간주하고, 나아가 윤리를 기술과 법, 정책, 원활한 거버넌스를 위한 도구로 이해하는 경향을 드러낸다. 이런 맥락에서 인공지능 윤리는 기술의 발전 및 적용 양상을 예측하기 어렵고 기술의 발전에 부담을 줄 수 있을 것이라는 이유로, 성문화된 법이나 강력한 규제보다 적용하거나 준수하기 쉬운 장치로 간주된다. 곧, 인공지능 기술의 개발, 도입, 적용 과정에서 핵심 동력은 경제적, 국가경쟁 차원의 이익이며 법・정책 등은 핵심 운용 장치이다. 그리고 윤리는 핵심 운용 장치가 원활하게 기능하기 어렵거나 핵심 동력을 저해할 것이 우려될 때 대안으로 활용하는 일종의 보조 장치인 것이다.

우리는 여기서 윤리에 대한 제한적 이해만이 아니라, 기술에 대한 전통적 관점의 혼재 역시 읽어낼 수 있다. 먼저 이같은 도식화는 기술 자체는 가치 중립적이며, 나아가 가치 중립적 기술을 인간의 통제하에 둘 수 있다고 보는

는 고민하지 않으며, '적용'을 위한 전략을 중심으로 한다. 또한 '사람 중심의 인공지능'은 '사람 중심'이라는 가치를 포함하지만, 이는 그 이상으로 논의되지 않고 대신 포용적 일자리 안전망 구축, 직업교육 재편, 역기능 방지 및 AI 윤리 체계 마련으로 압축된다. 특히 AI 윤리 체계의 마련은 다른 과제 추진과 비교할 때 글로벌 수준에 도달하겠다는 목표 외에 구체적인 서술을 찾기는 어렵다. 가치의 논의가 이러한 과제로만 다루어질 수 있는지도 의문이다.

기술 도구론의 관점을 전제한다. 이는 기술을 가치와 분리 가능한 것으로 간주하는 태도이다. 그럼에도 불구하고 가치를 논하며 가이드라인을 제시하는 것은 기술이 가치와 무관하지만 인간 통제를 전적으로 벗어나는 것은 아니라는 믿음을 전제하는 것이다. 이같은 입장에서 기술은 가치 중립적이지만 인간의 목적과 통제하에 놓인다. 흥미로운 점은 인공지능 윤리 논의에서 기술 도구론과 대조적인 관점인 기술 결정론적 태도 역시 발견된다는 점이다. 인공지능 기술에 의해 인간 사회가 불가피하게 변화할 수밖에 없다고 받아들이는 것은 기술이 사회의 변화를 규정한다는 기술결정론의 관점이다.[53] 인공지능 윤리 논의에는 이처럼 대조적인 기술 도구론과 기술결정론의 관점이 혼재되어 있다. 기술 도구론은 인간의 책임 소재를 강조할 수 있지만, 동시에 기술의 핵심을 목적 성취에 기여하는 도구적 효용성으로 간주하기 때문에 기술의 법적, 사회적, 정치적 여파는 쉽게 간과할 수 있다. 기술결정론은 기술의 강력한 힘과 영향력을 경계할 것을 촉구할 수 있지만, 인간의 책임 소재는 묻지 않는 귀결로 향하기 쉽다.

그러나 이는 현대기술, 특히 인공지능 기술을 이해하기에는 적절한 관점이라 볼 수 없다. 인공지능 기술은 알고리즘의 불투명성 등, 기술의 특성상 인간에 의해 전적으로 예측, 이해되거나 통제될 수 없다. 인공지능 기술이 충분히 통제 가능한 도구가 아니기 때문에 인공지능 윤리가 이토록 요청되고 있는 것이다. 그러나 다른 한편, 인공지능 기술은 인간에 의해 개발되고 선택되며 수용된다. 기술이 일방적으로 인간의 삶과 사회의 구조를 결정하는 것은 아니다. 기술과 인간은 상호영향을 미치며 서로를 함께 구성한다.

한편 윤리에 대한 이해를 다시 생각해보자. 우리 삶에서 옳은 것, 좋은

53) 손화철은 “인공지능 기술 발전과 그로 인한 사회적 변화를 정해진 ‘사실’처럼 받아들이는 이같은 태도를 날씨 예보를 대하는 사람들의 자세와 등치시킨다. 그리고 이같은 태도는 기술이 인간의 통제를 벗어나 자체적으로 발전한다는 ‘기술 자율성’에 대한 믿음을 바탕으로 한다고 지적하며, 기술을 인간이 통제할 수 없다고 보면서도 기술로 인한 문제 해결의 주체는 인간으로 간주하는 모순이 있음을 논한다.” 손화철 (2018), pp. 282-283 참조.

것의 가치 전반에 대한 물음과 탐구이다. “소크라테스의 말을 빌자면 “우리가 어떻게 살아야만 하는가”와 “왜 그러한가”에 대한 것이다.[54] 세부적으로는 가치, 원리, 규범의 내용에 대한 탐구이자, 우리가 윤리적 용어를 어떻게 이해하고 기술description하며 논증하는지에 대한 탐구이고, 실제 경험 상황에서 추구해야 하는 최선의 가치와 행동이 무엇이고 어떻게 실행할 수 있는지에 대한 탐구이기도 하다. 그러므로 윤리는 우리 삶, 결정, 행위 등과 불가분의 관계이다. 다시 말해 윤리는 법, 정책, 제도, 기술에 추가적으로 덧붙여지는 것이 아니라 이들의 기획, 배치, 형성, 작동에 이미 녹아들어 있는embeded 것이다. 가치에 대한 숙고 및 탐구로서 윤리, 윤리학은 가장 기초적인 삶의 기획이기 때문이다.[55]

결론적으로 ‘인공지능 윤리[AI Ethics]’ 역시 단순히 법(정책)을 보조하는 것, 기술의 발전을 저해하는 것으로 간주될 수 없으며 오히려 우리 사회의 기획, 그리고 사회를 구성하는 요소 중 하나인 기술 기획을 위한 대전제이자 불가결의 구성 요소로 이해되어야 할 것이다. 특정 기술의 형식, 구성, 사회 내 도입을 당연한 사실로 받아들인 후 “어떻게 좋게 사용할 것인가?”를 묻는 것이 아니라 기술 연구, 개발, 설계, 사회 내 도입 단계부터 그것이 우리 사회가 추구하는 가치에 부합하고 이를 증진하는 것인지를 철저히 검토해야 하는 것이다.[56] 윤리는 기술의 모든 단계의 결정 및 행동, 작용과 함께한다. 따라서 우리는 인공지능 윤리에서 윤리의 자리를 기술 발전을 저해하지 않지만 나쁜 사용이나 부작용은 통제하도록 조종하는 것에서 찾는

54) 제임스 레이첼즈, 노혜련, 김기덕, 박소영 역 (2006), p. 29.

55) Charles (1989), pp. 53-90 참조.

56) 손화철은 인공지능 기술이 사회에 끼칠 영향을 고려할 때, 기술의 다양한 결과를 고루 예상하고 고려하는 일보다 기술 개발 자체의 정당성을 먼저 물어야 한다고 주장한다. “...인공지능이 초래할 손해와 그 손해를 입는 사람들을 염두에 둔다면, 도대체 인공지능을 개발해야 하 는 이유가 무엇인가? 대규모 실업이나 권력의 집중화처럼 해결하기 힘들어 보이는 부작용이 예상되는 데다 어디로 튈지 모르는 불확실성마저 큰 기술을 굳이 개발해야 하는가?” 손화철 (2018), p. 288.

것을 경계해야 한다. 우리는 그 이상으로 나아가, 윤리를 기술 발전을 기획하고 주도하며 제시되고 협의된 가치에 따라 기술 혁신을 추동하는 힘으로 간주해야 할 것이다.

Ⅴ. 인공지능 윤리학의 과제: '인공지능' 윤리학과 인공지능 '윤리학', 그 사이

이상의 논의를 토대로 인공지능 윤리학의 향후 과제를 다음과 같이 정리할 수 있다.

첫째, 다양한 이해관계자가 '인공지능 윤리'에 접근, 이해, 실행할 수 있도록 인공지능 윤리의 간명하고 가시적인 조망과 이해틀framework이 제공되어야 한다. 인공지능 윤리와 관련된 기초 지식, 개념, 용어 역시 더욱 정련되어야 할 것이다. 물론 이 과정 및 결과에는 전 세계적인 합의가 수반되어야 한다.

둘째, 인공지능 윤리 논의 참여 주체의 다양성 확보에 노력을 기울여야 한다. 인공지능 윤리 논의 주체의 편향성, 제한적 특성이 현재의 불공정한 권력 관계를 무비판적으로 반영하거나 강화하지 않도록 주의해야 할 것이다. 이를 위해서는 다양한 입장, 특히 지금까지 과소대표된 집단의 입장이 충분히 반영되어야 한다. 따라서 인공지능 윤리 논의 주체의 다양성과 관련한 문제 제기, 로컬의 특수성 등 다양한 층위의 특수성 논의 및 다양성의 기준과 검사 방법 등 구체적인 방법론에 대한 논의가 지속적으로 생산되어야 할 것이다. 기술 전문가 및 행정 관료 등 관련 전문가가 아닌 시민사회의 참여 확대 및 그 방식 역시 구체적으로 논의되어야 할 것이다.

셋째, 원리와 실행의 간극을 줄여 인공지능 윤리 원리의 실효성을 높일 수 있는 후속 연구가 필요하다. 현재까지의 인공지능 윤리는 주로 다양한

경우, 행위자 등을 포괄할 수 있는 추상적인 원리의 차원에서 논의되었다. 동시에 경제적 이익 등 다른 목적을 위해서 쉽게 무력화될 수 있는 '약한' 경고 표지처럼 간주되기도 한다. 이 두 가지 요인은 인공지능 윤리 원리의 실행력을 약화시킨다. 그러므로 향후 인공지능 윤리는 구체적인 맥락과 실천을 고려하며, 합의된 원리 및 규범의 실행을 장려하는 실질적이고 구속력 있는 방법을 논의해야 할 것이다.

넷째, 인공지능 윤리 논의는 기술과 윤리에 대한 제한적이고 전통적인 이해를 확장하고 전환할 필요가 있다. 그 개발의 맥락, 실행, 여파를 고려할 때 가치와 무관한 기술은 있을 수 없으며 동시에 기술은 오늘날의 인간 정체성, 관계, 사회를 구성하고 영향을 미치는 주요 행위자로 간주되어야 한다. 인공지능 윤리 논의는 인간의 삶에 기여하고 인간의 책임을 요청한다는 점에서 인간 중심적인 특성을 포기할 수 없지만, 그 중심성은 과거와 달리 더욱 개방적, 포용적이 될 과제를 갖는다. 그러므로 향후 인공지능 윤리는 인간 개념 및 가치의 변형과 확장, 기술적 행위자와의 관계 및 공존, 전체 생태계 등을 포함하는 가치 패러다임의 전환을 더욱 진지하게 고려해야 할 것이다.

그리고 이 모든 과제를 위해서는 인문과학, 사회과학 등 가치와 사회를 연구하는 학계의 적극적인 연구 및 논의 활동이 요청된다. 우리는 가치와 기술에 대해 더 많은 문제를 제기하고, 더 많이 참여해야 한다. 이를 위해서는 기술적 지식에 대해 최소한의 이해가 필수적으로 요청될 것이다. 이 또한 인공지능 윤리 논의에 참여하는 학자들의 과제가 될 것이다.

우리의 연구는 현재 인공지능 윤리 논의의 일반적 경향과 특징을 조망하고, 그 안에서 인공지능 윤리가 어떤 목적으로 요청되고 있는지를 밝히는 것이었다. 이 연구의 목적 및 초점은 향후 인공지능 윤리 논의의 발전과 활성화를 위한 기초적 토대를 마련하는 것에 있다. 우리의 논의는 이에 따라 인공지능 원리 및 윤리 문헌을 총괄하여 분석하거나 구체적인 이슈에 대한 분석을 제시하지는 않았다. 그러므로 연구대상을 확장하고, 인공지능 윤리의

일반적 경향과 특징을 다양한 관점에서 논의하며, 각각의 인공지능 윤리 논의 자료에서 주목할 만한 지점을 고찰하고, 현재 인공지능 윤리의 한계와 향후 과제에 대한 상세하고 구체적으로 다루는 논의들이 후속 연구로 뒤따라야할 것이다. 인공지능 윤리에 대한 심화 이해, 구체적 이슈에 대한 논의, 그리고 특수한 국내 상황을 철저하게 연구하는 후속 연구는 우리 연구팀이 남겨둔 과제이기도 하다.

그러나 앞으로의 과제 수행을 위해서 인공지능 윤리는 무엇보다 '인공지능 윤리'라는 말로 지시하는 바를 명료하게 할 필요가 있다. 현재까지의 인공지능 윤리 논의는 살펴본 바와 같이 목적, 이슈, 범위가 다양하며 이들이 서로 분명하게 구분되지 않은 채로 진행 중이다. 이는 불필요한 오해, 혼란, 소통의 어려움을 낳는다. 우리는 이 논문을 통해 인공지능 윤리의 의미, 범위, 목적에 대한 기초적 이해가 마련되고 향상되기를 희망한다.

인공지능 윤리는 단지 인공지능 기술의 문제도 아니며, 전통적인 가치를 고수하며 이를 기술에 단순 적용하는 문제도 아니다. 인공지능을 둘러싼 윤리적 연구가 '인공지능'이라는 기술에 매몰되지 않으면서 동시에 새로운 기술 및 변화를 이해하는 윤리 논의가 되려면 이같은 이원론적 사고방식의 전환이 필요할 것이다. 그러므로 우리는 새로운 '인공지능 윤리학'의 정초기에 서 있다. 우리의 오랜 윤리학은 이제 새로운 기술을 더욱 확장된 관점에서 전통적인 방법과 새로운 방법을 함께 사용하며 가치, 방향성, 그 정당화를 고민해야 할 것이다.[57] 인공지능 윤리학은 기술의 특수성 및 실제 적용 맥락, 다양한 이해관계자를 고려하는 윤리론이어야 하며, 그러나 '새로운 기술'이라는 이름으로 다양한 사회적 결정 및 변화가 기술적으로 이미 정해진 '사실'인 것처럼 다루는 적응 방법론에 그쳐서는 안 된다. 윤리가 기술의 기획, 형성,

57) 예를 들면 우리에게는 다음의 질문이 과제로 놓여있다. 인공지능 기술로 기여하려는 인간 삶은 무엇을 지향하는가? 이것이 지금까지 역사적으로 쌓아온 부정적 편향성을 강화하지 않는가? 이것을 고려하는 가치 및 그 우선순위에 대한 사회적 합의는 어떻게 수행될 수 있겠는가?

배치 등에 이미 녹아있는 것임을 받아들인다면 인공지능 윤리는 기술의 개발 및 사회 도입에 앞서 더욱 선제적이고 전향적으로, 그리고 기술의 전체 단계와 함께 수행되어야 할 것이다. 철학자 회페는 윤리학이 진정으로 책임있고 비판적인 사유를 수행하는 학문이 되기 위해서는 공포와 희망, 어느 쪽에도 편향되지 않고 실천적 판단력을 '적시에' 발휘하는 것이 중요하다고 강조한다. 인공지능 윤리학의 과제 역시 미래를 바라보며 현재적으로 그 자신의 역할을 지금-여기에서 수행할 수 있어야 할 것이다. "책임 담론은 연구를 동반하며, 심지어 미래를 먼저 내다보며 실행되어야 한다. 만일 아테네의 부엉이가 단지 저녁에만 날기 시작한다면, 그렇다면 그 전날 저녁이면 왜 안 되는가?" 58)

58) 오트프리트 회페, 김시형 역 (2013), p. 409.

참고문헌

김동현 · 장준희(2019), 「신뢰 가능 AI 구현을 위한 정책 방향-OECD AI 권고안을 중심으로」, 『IT&Future Strategy』, 2: 한국정보화진흥원.

김형주 (2019), 「인공지능인문학: If 의 미래학에서 As-If 의 철학으로」, 『철학연구』, 151: 109-134.

김형주 (2018), 「인공지능 철학 국내연구 동향 분석—인공지능 철학의 생장점에서—」, 『인공지능인문학연구』, 1: 149-170.

김형주(2016). 「`인공지능`과 `인간지능` 개념에 대한 철학적 분석 시도」, 『철학탐구』, 43 : 161-190.

대한민국 정부(2019), 「대한민국 2019 AI 국가 전략」, 『대한민국 정책브리핑』.

손화철 (2018), 「인공지능 시대의 과학기술 거버넌스」, 『철학사상』, 68: 267-299.

정보통신정책연구원 (2018), 「4차산업혁명시대 산업별 인공지능 윤리의 이슈 분석 및 정책적 대응방안 연구」

허유선 (2018), 「인공지능에 의한 차별과 그 책임 논의를 위한 예비적 고찰: 알고리즘의 편향성 학습과 인간 행위자를 중심으로」, 『한국여성철학』, 29: 165-209.

오트프리트 회페 (1993), 김시형 역 (2013), 『학문윤리학』, 서울: 시와진실.

울리히 벡 (1998), 박미애 및 이진우 역 (2010), 『글로벌 위험사회』, 서울: 도서출판 길.

제임스 레이첼즈 (1986), 노혜련, 김기덕 및 박소영 역 (2006), 『도덕철학의 기초』, 서울: 나눔의집.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar(2020), "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI." Berkman Klein Center for Internet & Society.

Floridi, Luciano (2019) “Translating principles into practices of digital ethics: five risks of being unethical”, *Philosophy & Technology*, 32(2): 185-193.

Floridi, Luciano, and Josh Cowls (2019) “A unified framework of five principles for AI in society”, *Harvard Data Science Review*. 1(1),

Jobin, Anna, Marcello Ienca, and Effy Vayena (2019) “Artificial Intelligence: the global

landscape of ethics guidelines", arXiv preprint arXiv:1906.11668.

McCarthy, John, Minsky, Marvin L, Rochester, Nathaniel and Shannon, Claude E (2006) "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955", *AI magazine*, 27(4): 12-12.

Samuel, Arthur L. (1960) "Some moral and technical consequences of automation—a refutation", *Science*, 132(3429): 741-742.

West, Sarah Myers, Meredith Whittaker and Kate Crawfor(2019), "Discriminating Systems - Gender, Race, and Power in AI", AI Now Institute, 1-33.

Wiener, Norbert (1960) "Some moral and technical consequences of automation", *Science*, 131(3410): 1355-1358.

Hagendorff, Thilo (2019) "The ethics of AI ethics--an evaluation of guidelines", arXiv preprint arXiv:1903.03425.

Whittlestone, Jess, Nyrup, Rune, Alexandrova, Anna and Cave, Stephen (2019, January) "The role and limits of principles in AI ethics: towards a focus on tensions", In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 195-200).

Fjeld, Jessica, Achten, Nele, Hilligoss, Hannah, Nagy, Adam and Srikumar, Madhulika (2020) "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI", Berkman Klein Center Research Publication,

Taylor, Charles (1989), *Sources of the Self: The Making of the Modern Identity*, Cambridge, Mass.: Harvard University Press.

The Vatican (2020), AI Rome Call for AI Ethics, Rome: The Vatican.

Zeng, Yi, Enmeng Lu, Cunqing Huangfu(2019), "Linking Artificial Intelligence Principles", In the Proceedings of the AAAI Workshop on Artificial Intelligence Safety.

Madiega, T. (2019), "EU guidelines on ethics in artificial intelligence: Context and implementation", European Parliamentary Research Service.

Campolo, Alex, Sanfilippo, Madelyn, Whittaker, Meredith, Crawford, Kate (2017), AI NOW Report 2017, New York: AI Now Institute.

Crawford, Kate, Whittaker, Meredith, Elish, Madeleine Clare, Barocas, Solon, Plasek, Aaron, Ferryman (2016), AI NOW Report 2016, New York: AI Now Institute.

Crawford, Kate, Dobbe, Roel, Dryer, Theodora, Fried, Genevieve, Green, Ben, Kaziunas, Elizabeth, Kak, Amba, Mathur, Varoon, McElroy, Erin, Sánchez, Andrea Nill, Raji, Deborah, Rankin, Joy Lisi, Richardson, Rashida, Schultz, Jason, West,

Sarah Myers, and Whittaker, Meredith (2019), AI Now 2019 Report, New York: AI Now Institute.

Executive Office of the President National Science and Technology Council Committee on Technology(2016), “Preparing for the Future of Artificial Intelligence”, Preparing for the future of artificial intelligence. Executive office of the president.

OECD (2019), Artificial Intelligence in Society, Paris: OECD Publishing.

McCorduck, Pamela, and Cli Cfe (2004), *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, Florida: CRC Press

Whittaker, Meredith, Crawford, Kate, Dobbe, Roel, Fried, Genevieve, Kaziunas, Elizabeth, Mathur, Varoon, West, Sarah Mysers, Richardson, Rashida, Schultz, Jason, Schwartz, Oscar (2018), AI NOW Report 2018, New York: AI Now Institute.

홍석만 (2017), “디지털 전환과 노동의 미래 로봇, 디지털 경제와 자본주의의 미래(4)”, http://workers-zine.net/27609. (검색일: 2020.03.28.)

리테일온 (2019), “인공지능(AI) 시장 글로벌 동향”, http://www.retailon.kr/on/bbs/board.php?bo_table=r1_02&wr_id=100. (검색일: 2020.03.27.)

카카오(2018), “카카오 알고리즘 윤리 헌장”, https://www.kakaocorp.com/kakao/ai/algorithm. (검색일: 2020.03.28.)

Beijing Academy of Artificial Intelligence (2019), “Beijing AI Principles”. https://www.baai.ac.cn/blog/beijing-ai-principles. (검색일: 2020.03.28.)

Darrell West, John Allen (2018), “How Artificial Intelligence is Transforming the World”, https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/ (검색일: 2020.03.28.)

DeepMind(2017), “DeepMind Ethics & Society Principles”, https://deepmind.com/about/ethics-and-society. (검색일: 2020.03.28.)

European Commission's High-Level Expert Group on Artificial Intelligence (2019), “Ethics Guidelines for Trustworthy AI”, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (검색일: 2020.03.28.)

Google(2018), “Artificial Intelligence at Google: Our Principles”, https://www.blog.google/technology/ai/ai-principles/. (검색일: 2020.03.28.)

Microsoft(2018), “Microsoft AI principles”, https://www.microsoft.com/en-us/ai/responsible-ai. (검색일: 2020.03.28.)

OpenAI(2018), “OpenAI Charter.”, https://openai.com/charter/. (검색일: 2020.03.28.)

IBM(2018), “Principles for Trust and Transparency”, https://www.ibm.com/blogs/policy/trust-principles/. (검색일: 2020.03.28.)

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019), “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems First Edition“, https://ethicsinaction.ieee.org/ (검색일: 2020.03.28.)

ITI(2017), “AI Policy Principles”, https://www.itic.org/public-policy/ITIAIPolicy PrinciplesFINAL.pdf. (검색일: 2020.03.28.)

Partnership on AI(2016), “Tenets”, https://www.partnershiponai.org/tenets/. (검색일: 2020. 03.28.)

University of Montreal(2018), “The Montreal Declaration for a Responsible Development of Artificial Intelligence”, University of Montreal. https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/. (검색일: 2020.03.28.)

Vincent, James (2019), “The problem with AI ethics”, https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech. (검색일: 2020.03.28.)

Abstract

Why Ethics is:

A Landscape of Modern AI Ethics Debate, Its Features and Limitations

Heo, Eu-Sun Lee, Yeon-Hee Shim, Ji-Won

Despite the tension of a technology-ethics that ethics generally regard as a inhibitor to the development of technology, why is ethical discussions more actively requested and increased regarding artificial intelligence technology? In order to answer this question, a meta-research is needed to look at and critically review the current discussion of artificial intelligence ethics as a whole. Focusing on discussing artificial intelligence ethics at the current technology level, we review artificial intelligence ethics literature published mainly by major artificial intelligence actors such as governments, international organizations and companies in each country mainly the last three years. And through this, we critically examine the characteristics of today's discussions on artificial intelligence ethics. The paper reveals the purpose for which ethics is requested at present in the discussion of artificial intelligence ethics, the overall trend of artificial intelligence ethics discussion, and in conclusion, ‘artificial intelligence ethics’ will have to be carried out with the whole stage of technology, recognizing that ethics are already melted in the planning, formation and deployment of technology, prior to the development and application of the society of technology.

【Keywords】 AI, AI Ethics, AI Principle, Ethics of Technology, Meta Study of AI Ethics

논문 투고일: 2020. 03. 31
심사 완료일: 2020. 04. 13
게재 확정일: 2020. 04. 13