한국의료윤리학회지 제26권 제2호(통권 제75호): 103-131 ©한국의료윤리학회, 2023년 6월
Korean J Med Ethics 26(2): 103-131 ©The Korean Society for Medical Ethics, June 2023
pISSN 2005-8284 eISSN 2234-3598

[연구논문] https://doi.org/10.35301/ksme.2023.26.2.103

보건의료영역 인공지능 윤리 원칙과 고려사항*,**

문기업[1], 양지현[2], 손유미[3], 최은경[4], 이일학[5]

요약

보건의료영역에서의 인공지능 활용은 지속적으로 확대되고 있다. 인공지능은 보건의료영역의 인적, 물적 자원 부족 문제를 해결하고 진단 및 치료 영역에서 혁신을 가져다 줄 것으로 기대된다. 하지만 딥러닝 학습법과 그로 인한 블랙박스 속성으로 인해 인공지능에 대한 윤리적 우려가 존재한다. 이러한 문제를 해소하기 위해서는 인공지능에 대해 제기되는 여러 윤리적 이슈와 개념을 확인하고 이를 다루는 윤리적 원칙을 제시할 필요가 있다. 따라서 본 연구는 국외의 여러 보건의료영역 인공지능 윤리 가이드라인을 분석하여 각 문헌에 제시되는 윤리적 영역, 이슈, 테마, 원칙들을 확인하고자 하였다. 이를 바탕으로 데이터 수집, 임상 환경, 사회 환경 영역에 해당하는 중요 윤리 테마로 인간에 대한 존중, 책무성, 지속가능성 세 가지를 도출해 내었고, 각 영역 및 테마에 해당하는 윤리 원칙들을 제시하고 이를 도식화하였다. 그리고 나서 도출한 10가지 윤리 원칙들에 대해 이슈, 개념 분석, 적용을 중점으로 상세한 설명을 시도하였다. 그리고 보론으로 윤리원칙의 상충가능성에 대해 숙의민주주의 도입을 통한 컨센서스 형성을 제언하였다. 본 연구를 통해 고찰된 윤리원칙이 향후 윤리적인 보건의료영역 인공지능 제작 및 활용에 도움이 되기를 기대하며 글은 마무리 된다.

색인어

인공지능; 가이드라인; 윤리원칙; 보건의료; 설명가능성; 책무성

투고일: 2023년 5월 19일, 심사일: 2023년 5월 24일, 게재확정일: 2023년 6월 19일

교신저자: 최은경, 경북대학교 의과대학 의료인문 의학교육학교실. e-mail: qchoiek@gmail.com
이일학, 연세대학교 의과대학 인문사회의학교실 의료법윤리학과. e-mail: ARETE2@yuhs.ac

* 본 연구는 2022년 보건복지부의 지원를 받아 수행되었음(보건의료분야 인공지능 활용 윤리 가이드라인 수립 연구).

** 본 논문은 2022년 11월 24일 대한의료정보학회 추계학술대회 〈인공지능의 의학적 사용을 위한 윤리원칙〉, 2022년 11월 25일 한국의료윤리학회 추계학술대회 〈인공지능의 의학적 사용을 위한 윤리원칙〉에서 발표된 원고를 수정·보완한 것입니다.

1 서울대학교 의과대학 인문의학교실, 박사과정. ORCID: https://orcid.org/0009-0009-2057-7636

2 연세대학교 의과대학 인문사회의학교실 의료법윤리학과, 연구강사. ORCID: https://orcid.org/0000-0003-3590-9356

3 연세대학교 일반대학원 의료법윤리학협동과정, 연세대 의료법윤리학연구원, 박사수료. ORCID: https://orcid.org/0000-0002-5128-6972

4 경북대학교 의과대학 의료인문 · 의학교육학교실, 조교수. ORCID: https://orcid.org/0000-0003-1448-1565

5 연세대학교 의과대학 인문사회의학교실 의료법윤리학과, 부교수. ORCID: https://orcid.org/0000-0002-6531-8752

I. 서론

인공지능(artificial intelligence, AI)[1]의 활용에 대한 사회적 관심은 인공지능 기술의 혁신에 따라 나날이 늘어나고 있다. 특히 딥러닝[2]을 기반으로 하는 인공지능은 기존에 다룰 수 없던 대량의 데이터를 활용하여 주어진 문제를 해결할 수 있으며, 더 나아가서는 기존에 해결할 수 없었던 여러 문제를 해결할 수 있는 수단으로 주목받고 있다. 보건의료 영역에서도 인공지능을 활용하기 위한 시도는 활발하게 이루어지고 있다. 예를 들어 근래 가장 큰 주목을 받았던 보건의료 분야 인공지능 중 하나인 IBM Watson for Oncology는 2016년 출시되었다. 자연어 처리(natural language processing)[3] 기법을 활용한 IBM Watson은 기대에 미치지 못하는 정확성 문제로 인해 널리 활용되지 못하였다. 하지만 인공신경망(artificial neural network)을 활용한 딥러닝 기법의 지속적 발달은 인공지능이 진료현장에서 본격적으로 활용될 수 있는 가능성을 보여주고 있다.

보건의료 분야가 마주하고 있는 여러 문제는 인공지능을 활용하려는 시도가 활발하게 전개되고 있는 이유 중 하나이다. 고령화로 인한 의료수요 및 의료비 증가로 인해 선진국을 중심으로 인적 · 물적 자원의 소모가 상당히 부담으로 작용하고 있다. 따라서 보건의료 인공지능은 보건의료인력 부족을 완화하고 의료비 지출을 줄이며, 그와 더불어 더욱 정확한 진단을 가능케 함으로서 21세기 고령화 사회의 지속가능성을 증진시키는 역할을 할 것으로 기대되고 있다[2].

인공지능 의료기기의 활용 현황을 살펴보면 미국 식품의약국(U.S. Food and Drug Administration, FDA)에서는 2022년 10월 기준 인공지능 의료기기 521개가 승인되었고[3], 한국 식품의약품안전처에서는 2022년 11월 기준 139개의 인공지능 의료기기가 허가되었다[4]. 현재는 영상데이터를 활용하여 의사의 진료를 보조하는 단계의 인공지능 의료기기가 다수이나 인공지능의 활용은 진단의 영역으로 점차 확대될 것으로 예상된다.

하지만 보건의료 분야 인공지능의 활용에 대한 윤리적 우려 역시 존재한다. 딥러닝 인공지능의 메커니즘은 개발자를 포함한 인간이 설명하기 어려운 방식으로 이루어져 있으며, 기존 의료기기와 달리 개발에 방대한 데이터가 활용됨에 따라 데이터 수집 및 활용에 대한 윤리적 문제 역시 제기되고 있다. 이와 같은 윤리적 우려에 대하여 세계 각국의 정부, 국제기구, 싱크탱크, 전문가 집단은 다양한 보건의료영역 인공지능 윤리 가이드라인을 제시해 왔으며, 한국에서도 2021년 KAIST 산하 연구소 한국4차산업혁명정책센터에서 싱가포르, 영국의 기관과 협력하여 발간한「사회를 위한 보건의료 분야 인공지능 활용 가이드」[5], 국가인권위원회에서 2022년 발간한「인공지능 개발과 활용에 관한 인권 가이드라인」[6]을 비롯한 인공지능 관련 윤리 지침들이 등장하기 시작하였다.

여러 보건의료영역 인공지능 윤리 가이드라인은 보건의료영역 인공지능의 개발부터 활용단계

1) 인공지능은 주어진 특정한 목표에 대해 예측, 권고, 결정을 할 수 있는 기계 기반 시스템이다. 인공지능은 설계 목적에 따라 자동화 수준이 다양하게 설정될 수 있다[1].

2) 딥러닝은 머신러닝의 일종으로 인공신경망 등 다층적 모델(multi-layered model)을 활용하여 데이터의 특성을 추출하고 학습하여 주어진 문제를 해결하는 기법이다. 여기서 머신러닝이란 주어진 알고리즘 없이 데이터 셋을 기반으로 학습하여 기능을 수행하고 개선을 해 나가는 인공지능의 학습법을 의미한다.

3) 일상생활에서 사용하는 언어를 인공지능이 분석 및 처리할 수 있도록 하는 기술을 의미한다.

에 이르기까지 전 과정에 대해 윤리적 인공지능을 만들기 위해 필요한 개념과 방안을 제시하고 있으며, 보건의료영역 인공지능에 대한 법적 규제 및 거버넌스 구축 과정에 중요한 역할을 하고 있다. 따라서 이들 가이드라인에 등장하는 여러 윤리적 이슈를 살펴보고 제시된 개념을 명료화 하여 윤리원칙이 어떻게 적용될 수 있는지를 분석하는 것은 매우 중요한 과제라고 할 수 있다. 이번 연구에서는 국외에서 발간된 중요 보건의료영역 인공지능 윤리 가이드라인에서 제시된 윤리적 개념 및 활용 방안 중 중요한 내용을 선별한 후 이슈, 원칙, 적용에 중점을 두어 윤리 개념들을 분석하고자 한다. 이를 통해 국내 보건의료 인공지능영역 윤리 가이드라인 마련 및 거버넌스 구축에 기여하기 바란다.

Ⅱ. 본론

1. 보건의료영역 인공지능 윤리 원칙 도출 과정

1) 문헌 검색 및 수집 방법

먼저 Web of Science, PubMed 등지에서 최근 5개년(2018~2022) 간행물을 대상으로 자연어 검색 방법을 사용하였다. 1차로 “medical(healthcare, medicine, health)”, “artificial intelligence (machine learning, AI, deep learning, algorithm)”, “ethics or guideline”를 검색필드 topic으로 하여 검색을 시행하였다. 이 결과 Web of Science에서는 616건, PubMed에서는 686건의 간행물이 검색되었고, 이들을 보건의료영역 인공지능 윤리와 “아주 많이 관련 있음”, “어느 정도 관련 있음”, “관련 없음”, “잘 모르겠음”으로 나누어 구분하였다. 관련이 있다고 판단된 간행물 중 인공지능에 대한 윤리 원칙, 선언을 중심으로 추려 총 129건의 문헌을 수집하였다.

2) 문헌 분류 기준 및 주요 주제 매핑

수집한 129건의 윤리 원칙, 보고서 문헌들을 국제/국가별, 작성처별 –정부(국제기구 포함), 비영리재단, 비정부기구, 학계, 국제 전문가 집단–으로 분류하였다. 출처가 불명확한 다섯가지 간행물을 제외하고 국가별로 살펴보면 〈Table 1〉과 같다.

전체적으로 미국에서 발표된 문헌이 36건으로 가장 많았으며, 국제 기구 및 국제 전문가 집

〈Table 1〉 Number of publications by country

Country	Count	Country	Count	Country	Count
United States	36	France	3	Iceland	1
International	27	Germany	3	India	1
United Kingdom	20	Netherlands	2	Italy	1
European Union	10	Norway	2	Korea (Republic of)	1
Japan	4	Singapore	2	Russia	1
Canada	3	Australia	1	Spain	1
Finland	3	China (People’s Republic of)	1	United Arab Emirates	1
Total		124			

단, 글로벌 민간재단에서 발표된 문헌 또한 27건에 달하였다. 또한 영국에서 20건으로 다수의 문헌이 발표되었고 EU에서는 총 10건이 발표되었다. 그 외 일본, 캐나다, 핀란드, 프랑스, 네덜란드, 노르웨이, 싱가포르 등이 각각 2건 이상의 문헌을 발표하고 있음이 눈에 띄었다.

발행 주체별로 살펴보면 〈Table 2〉와 같다. 정부기관(의회, 정부주도 수립 기구, 의회 씽크탱크 등 포함)에서 발표한 문헌이 가장 많았으나 그 외 학계에서도 32건을 발표하였으며, 그 다음 산업계, 비영리재단, 비정부기관 순이었다. 인공지능 개발을 주도하는 민간기업 및 기업 연합체 등 산업계에서 인공지능 윤리 원칙에 관하여 발표하는 양상이 두드러졌다.

이들 문헌의 개별 내용을 검토한 후 보건의료 관련 이슈들을 주요한 주제로 삼고 있는 문헌들을 추려내었고, 총 23개 문헌이 보건의료 분야와 관련 있는 것으로 확인되었다. 이들 23개의 문헌을 보건의료영역 인공지능 윤리 가이드라인에 활용하기 위해 추가적으로 검토하였다.

최종적으로 보건의료와 관련성이 크고 구체적인 권고 기준을 담고 있는 총 9건의 문헌, 그리고 인공지능 윤리와 관련하여 가장 빈번하게 인용되는 OECD의 "Recommendation of the Council on Artificial Intelligence" 및 UNESCO의 "Recommendation on the Ethics of Artificial Intelligence"를 포함하여 총 11개의 문헌을 선정하였다. 그 목록은 〈Table 3〉과 같다.

2. 보건의료 인공지능 윤리 원칙: 매핑(mapping)과 도출

주 연구자 2명이 함께 선정된 보건의료 인공윤리 관련 문헌을 숙지, 검토하고 주요하게 다루어진 윤리적 영역, 키워드, 가치, 원칙(ethical domain, keyword, value, principle) 등을 추출하여 망라하였다. 문헌 중에 언급된 윤리적 영역, 윤리 키워드, 윤리적 가치, 윤리 원칙 등의 용어에 주목하였다. 먼저 연구자 A가 해당 문헌들에서 주요하게 언급하고 있는 윤리적 영역, 키워드, 가치, 원칙 등을 목록화 한 후 연구자 B가 이를 교차 검토하였다.

매핑, 그리고 글의 구성전반과 관련해 2020년 하버드 대학교 산하 버크만 클라인 센터(Berkman Klein Center)에서 발간한 "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI"를 참조하였다[16]. 이 문헌에서는 35개 인공지능 윤리 원칙, 가이드라인을 분석하여 8가지 중요 테마(themes)[4]를 도출하였고, 각 테마에 대한 상세한 개념 설명을 더해

〈Table 2〉 Number of publications by type of publisher

Type of publisher	Count	Type of publisher	Count
Government (Incl. legislative body)	38	Non-profit organization	16
Academia	32	Non-governmental organization	6
Industry	27	Forum	5
Total		124	

4) 본 문헌에서 Fjeld 등이 제시한 8가지 중요 테마는 다음과 같다. 사생활(privacy), 책무성(accountability), 안전과 보안(safety and security), 투명성과 설명가능성(transparency and explainability), 평등과 반차별(fairness and non-discrimination), 기술에 대한 인간의 통제(human control of technology), 전문직업 책임(professional responsibility), 인간 가치 증진(promotion of human values).

〈Table 3〉 Characteristics of the 11 selected publications

Year	Publisher	Title	Source
2018	Future Advocacy	Ethical, social, and political challenges of artificial intelligence (AI) in health	[7]
2018	Nuffield Council	AI in healthcare and research	[8]
2019	Academy of Medical Royal Colleges	AI in health	[9]
2019	American College of Radiology; European Society of Radiology; Radiology Society of North America; Society for Imaging Informatics in Medicine; European Society of Medical Imaging Informatics; Canadian Association of Radiologists; American Association of Physicists in Medicine	Ethics of AI in tadiology: European and North American Multisociety Statement	[10]
2019	OECD	Recommendation of the Council on Artificial Intelligence*	[1]
2020	OECD	Trustworthy AI in Health	[2]
2020	Ministry of Health, Health Sciences Authority, Integrated Health Information Systems (Singapore)	Artificial Intelligence in Healthcare Guidelines	[11]
2021	WHO	Ethics and governance of AI for health: WHO guidance	[12]
2021	UNESCO	Recommendation on the Ethics of artificial intelligence*	[13]
2021	Council of Europe	The impact of AI on the doctor-patient relationship	[14]
2022	European Parliament Research Service	AI in healthcare: applications, risks, and ethical and societal impacts	[15]

* Although these publications are not specific to healthcare settings, they are highly cited when discussing the ethics of artificial intelligence.

구체적 원칙(principle)들에 대한 설명을 하였다. 이번 연구에서는 "Principled Artificial Intelligence"의 방법론을 참조하되 보건의료영역과 관련된 문헌들을 중심으로 윤리적 테마, 원칙 등을 분석하였다. 검토한 결과는 〈Table 4〉와 같다.

11개 문헌을 살펴본 결과, 다양한 윤리적 영역, 이슈, 가치, 원칙 중 "평등(형평, 반차별)(equality, equity, non-discrimination) (9건)"이 가장 빈번하게 제시되었다. 그 다음으로 "안전성(보안성, 견고성)(safety, security, robust-

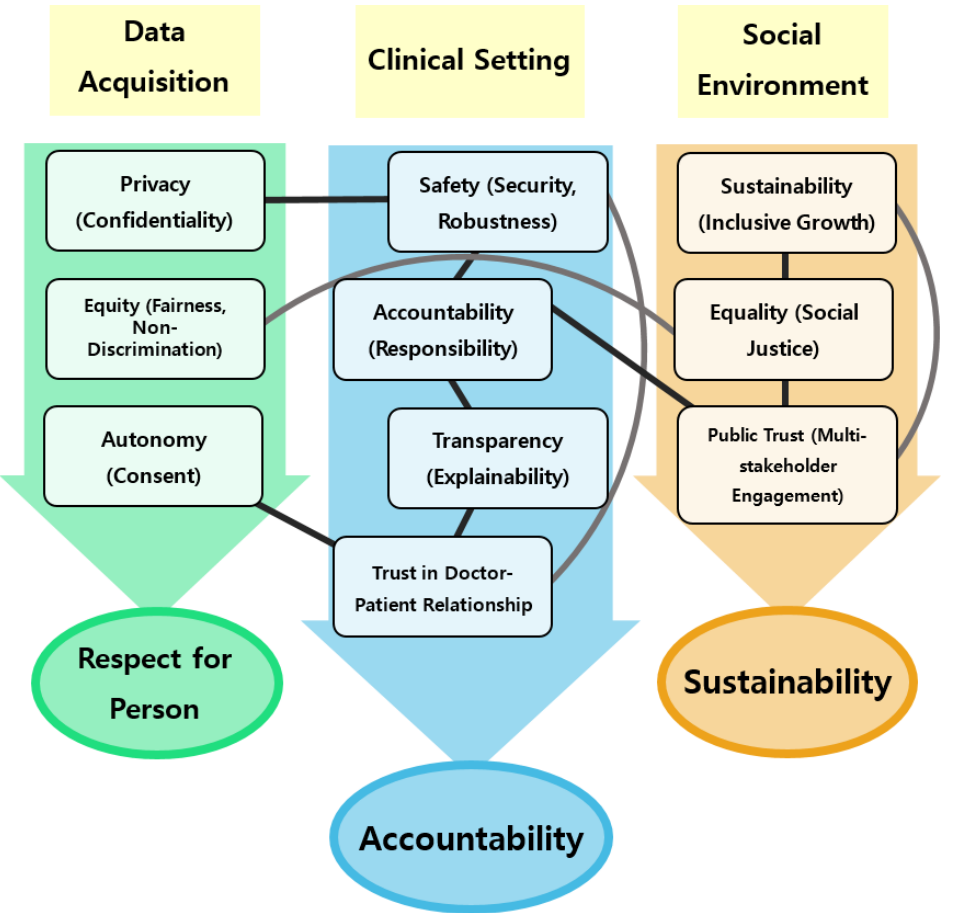
〈Table 4〉 Ethical Domains, Issues, Themes, Principles Covered in the 11 Selected Publications

Publisher	Ethical domains, issues, themes, principles				
	Autonomy (consent)	Privacy (confidentiality)	Safety (security, robustness)	Transparency (explainability)	Accountability (responsibility, human oversight & determination)
Future Advocacy	○				
Nuffield Council			○	○	○
Academy of Medical Royal Colleges		○	○		○
EU-NA Radiology Associations	○				
OECD (2019)			○	○	○
OECD (2020)					
MOH, Singapore	○	○	○	○	○
WHO	○		○	○	○
UNESCO	○	○	○	○	○
Council of Europe		○	○	○	
European Parliament Research Service		○	○	○	○

Publisher	Equality (equity, non-discrimination)	Sustainability and inclusive growth	Public trust, multi-stakeholder engagement & collaboration	Trust in doctor-patient relationship	
Future Advocacy	○				
Nuffield Council	○				
Academy of Medical Royal Colleges	○	○		○	
EU-NA Radiology Associations					
OECD (2019)	○	○			
OECD (2020)		○			
MOH, Singapore	○				
WHO	○	○			
UNESCO	○	○	○		
Council of Europe	○				
European Parliament Research Service	○	○			

ness) (8건)"이 뒤를 이었으며, "투명성(설명가능성)(transparency, explainability) (7건)", "책무성(책임성, 인간의 감독 및 결정)(accountability, responsibility, human oversight & determination) (7건)", 지속가능성, 동반 성장(sustainability, inclusive growth) (6건)" 등도 중요하게 언급되었다. "사생활 보호 및 기밀유지(privacy and confidentiality) (5건)", "자율성 존중(사전 동의)(autonomy, consent) (5건)" 또한 주요한 내용에 포함되었다.

모든 저자들(A, B, C, D, E)의 토의를 거쳐 이들 보건의료 인공지능이 다루어야 할 이슈, 지향해야 할 가치 및 원칙에 대한 키워드를 보건의료 영역에서 실천적으로 활용하기 용이하도록 데이터 수집(data acquisition), 임상 환경(clinical setting), 사회적 환경(social environment) 세 가지 영역으로 나누어 〈Figure 1〉과 같이 구조화하였다. 데이터수집 영역에서는 "인간에 대한 존중(respect for person)", 임상환경 영역에서는 "책무성(accountability)", 사회적 환경 영역에서는 "지속가능성(sustainability)"의 가치를 큰 테마로 하나씩 도출해 내었고, 테마에 대응하



〈Figure 1〉 Schematization of ethical themes and principles of artificial intelligence in health care settings

여 구체적 세부 원칙들을 제시할 수 있었다.

우선 데이터 수집의 영역에서 가장 중요하게 다룰 수 있는 테마로 '인간에 대한 존중'을 제시하였고, 그와 관련된 원칙으로 '사생활 보호(기밀유지)', '형평(equity; 공정, 반차별)', 그리고 '자율성(동의)' 세 가지를 선정하였다. 한편 임상 환경에서의 테마로 연구진은 책무성을 선정하였는데, 책무성의 영역에는 '안전성(보안성, 견고성)', '책무성(책임성)', '투명성(설명가능성)', '의사-환자 관계의 신뢰' 네 가지 원칙을 선정하였다. 그리고 마지막으로 사회 환경 영역의 주요 테마로 지속가능성을 선정하고 그에 따른 중요 원칙으로 '지속가능성(포괄적 성장)', '평등(equality; 사회정의)', '대중신뢰 및 다자참여' 세 가지를 선정하였다. 물론 이들 '인간에 대한 존중', '책무성', '지속가능성'의 각 영역의 테마가 다른 영역에 적용되지 않는다고 보기는 어렵다. 다만 각 영역에서 좀더 중점을 두게 되는 가치를 포괄화한 것으로 이해하는 것이 바람직할 것이다.

윤리원칙들을 임상 환경 영역과 데이터 수집 영역, 사회 환경 영역으로 분류하였으나, 각 영역에 해당하는 원칙들은 서로 밀접한 관련이 있다. 우선 임상 환경 영역의 네 원칙은 데이터 수집 영역의 세 원칙과 결코 무관하지 않는데, 예를 들어 사생활 보호(기밀유지)의 문제는 안전성(보안, 견고성)과 밀접한 관련이 있다. 그리고 자율성(동의)의 문제는 의사-환자 간의 신뢰와 맞닿아 있는 개념이다. 이러한 연결 개념을 〈Figure 1〉에서는 실선으로 표현하였다. 또한 같은 영역 내부의 원칙들에서도 연관관계를 찾을 수 있는데, 가령 투명성(설명가능성)은 결론이 도출된 요인과 과정에 관한 정보의 제공과 관련이 있다는 점에서 책무성(책임성)을 강화하는 요소가 될 수 있으며, 마찬가지로 투명성(설명가능성)을 갖추었을 때 안전(보안, 견고성) 그리고 의사-환자 관계의 신뢰 역시 확보할 수 있다. 이러한 연관관계 역시 실선으로 표기해 두었다.

3. 보건의료 인공지능 윤리 원칙: 핵심 내용

앞서 선정한 11가지 문헌에 등장한 보건의료 영역 인공지능이 다루어야 할 이슈, 지향해야 할 가치 및 원칙에 대한 키워드를 분석하여 세 가지 영역에 대응하는 중요 테마를 하나씩 도출해 내고 중요 테마별로 주요 원칙들을 제시할 수 있었다. 원칙들은 데이터 수집 영역-인간 존중 테마에서 3가지, 임상 환경 영역-책무성 테마에서 4가지, 사회 환경 영역-지속가능성 테마에서 3가지를 도출하여 총 10가지 원칙을 제시할 수 있었다. 지금부터는 10가지 원칙에 대한 구체적인 논의를 진행하고자 한다.

효과적인 논의를 위해 10가지 원칙을 통합적으로 9가지 개념으로 설명하고자 한다. 즉, 개념 중심의 논의를 위해 다른 영역-테마에 속해 있지만 관련성이 깊다고 판단되는 원칙들을 하나의 개념으로 묶어 설명하였다. 결과적으로 연구진들은 세부 원칙들을 아홉 가지로 나누어 핵심 내용을 파악하고 추가적으로 관련 내용을 다룬 문헌을 분석하여 구체적 논의를 진행하고자 하였다. 아홉 가지 영역은 (1) 사생활 보호, 기밀유지, (2) 형평, 공정, 차별금지, 그리고 평등, 사회정의, (3) 자율성, 동의, (4) 안전성, 보안성, 견고성, (5) 투명성, 설명가능성, (6) 책무성, 책임성, (7) 의사-환자 관계의 신뢰, (8) 지속가능성, 포괄적 성장, (9) 대중 신뢰, 다자 참여와 협력이다. 형평(공정, 차별금지)의 원칙은 평등(사회정의)의 원칙과 통합하여 다룰 수 있다고 보았다. 이는 데이터 형평성과 결과적 평등, 사회정의의 구현이 긴밀하게 연결되어 있다고 보았기 때문이다.

이러한 아홉 가지 키워드가 앞서 선정한 11개 문헌에서 어떻게 다루어 졌는지를 나열하고, 11가지 문헌과 더불어 각 개념에 대한 구체적인 논

의가 진행된 여러 연구자료를 종합하여 각각의 개념에 대한 심화된 분석을 시도하였다.

1) 사생활 보호, 기밀유지

대부분의 문헌에서 사생활 보호는 기본적 가치로 다루어지고 있다. 사생활 보호는 인간의 자율성과 존엄성을 보호하는 규범 및 실천이며 침해로부터의 자유, 감시받지 않을 자유, 자기 정체성에 대한 통제 등의 개념을 포함한다. 또한 사생활 보호는 개인의 신체, 정신 및 평판에 대한 보호를 모두 포괄한다고 할 수 있다[17].

보건의료 인공지능은 민감정보인 의료데이터를 이용하기 때문에 사생활 보호 문제가 필연적으로 대두된다. 기본적으로 다량의 데이터를 이용한 임상연구에서 데이터 주체의 사생활 보호 및 동의 문제가 인공지능 개발에도 문제가 된다. 우선 임상연구에서 데이터 수집에 대한 동의를 구하는 경우 연구 내용에 대한 충분한 정보를 제공한 다음 동의를 구하는 것이 요구되는데, 보건의료 인공지능 연구의 경우 연구 내용에 대한 "충분한" 정보를 제공하는 것 자체가 어려울 수 있다. 뿐만 아니라 이 이슈가 중요한 이유는 딥러닝 인공지능의 학습 방식과 관련이 있다. 딥러닝 인공지능이 구현되기 위해서는 기존 연구에서는 볼 수 없는 대량의 데이터를 활용하여 인공지능을 훈련시켜야 하며, 다량의 민감정보가 추출, 수집, 연계되는 과정에서 개인의 사생활이 침해될 가능성이 커질 수 있기 때문이다.[5] 또한 딥러닝 인공지능은 한 개인에 대해 더 많은 정보가 갖추어 질수록 더 정확한 결과를 산출해낼 수 있다. 특히 사회경제적 상태, 유전적 특성 등 민감하게 여겨질 수 있는 개인의 특성은 역학적으로 입증된 질병의 상관요인이기에, 보건의료 인공지능의 성능 극대화를 추구할 경우 이러한 사생활 데이터를 활용하여 정확성을 높이는 방향으로 알고리즘이 구현될 가능성이 있다.

또한, 빅데이터를 활용하는 보건의료영역 인공지능의 개발 및 활용 과정에서 설명동의 없이 개인정보 데이터가 공유되거나 수집된 데이터를 사전동의 받지 않은 다른 목적으로 사용(re-purpose)하는 경우가 발생할 수 있다. 이러한 우려는 데이터 노출로 인한 신원 도용 및 범죄의 위험 발생 등, 사생활 보호 미비로 인한 여러 문제가 발생할 수 있다는 위험성으로 인해 더욱 강화되고 있다[15].

여러 윤리 문헌은 사생활 보호와 기밀유지에 대한 우려를 해소하기 위해 인공지능 전 생애주기에 걸쳐 사생활 보호 원칙이 지켜지도록 주의를 기울일 필요가 있음을 강조한다. 즉, 데이터 수집이 개인의 자율성에 기반한 동의를 존중하는 방향으로 이루어져야 할 것과 더불어 개인정보 데이터의 보관 과정에서의 발생할 수 있는 개인정보 노출 가능성이 없는지, 인공지능의 활용단계에서 환자의 정보 수집 시 사생활 침해의 여지가 없는지를 전 과정에서 면밀히 살펴보도록 하고있다[13,18]. 이것은 인공지능 생애주기 전반의 거버넌스에서 사생활 보호 원칙이 지켜질 수 있도록 규제 거버넌스 체계를 수립할 것을 요구한다.[6]

최근 유럽을 비롯한 각 국가에서는 개인정보의 보호 및 활용의 균형을 도모하기 위한 제도

5) 사생활 보호 문제는 정보제공 동의의 문제와 직결된다. 정보제공에 대한 동의의 문제는 '2) 형평(공정, 반차별) & 평등(사회정의)' 파트에서 다루어질 예정이다.

6) 이를테면 UNESCO에서는 여러 이해당사자의 관점에서 데이터 보호 거버넌스 체계를 마련할 필요가 있음을 지적하며 그 방안으로 사회적, 윤리적 영역을 모두 고려한 인공지능 사생활 영향 평가 체계를 마련하여 인공지능 사용 전(全) 단계에서 사생활 침해 여지를 최소화하여야

적 변화가 이루어지고 있다. 2018년부터 시행되고 있는 유럽연합의 GDPR(일반 개인정보 보호법, General Data Protection Regulation)은 가명정보[7] 개념을 명시하고, 과학적 연구에 대한 동의 방식으로 포괄적 동의를 허용하였으며, 자동화된 의사결정에 대한 설명을 들을 권리 등 변화하는 데이터 처리 환경을 반영한 규정을 도입하였다. 한국에서도 이른바 데이터 3법(개인정보 보호법, 정보통신망법, 신용정보법)의 개정을 통해 인공지능 등의 개발 및 활용 과정에서 합법적 데이터 수집의 가능성을 열고, 거버넌스 정비를 통해 인공지능 개발 등 빅데이터 활용 산업의 효율화를 도모하고자 하였다. 예를 들어 개인정보 보호법(이하 '개인정보법'이라 한다.) 개정은 가명정보 처리 특례 규정을 신설하여 개인정보의 활용에 대한 제약을 완화함으로써 데이터 기반 연구 및 개발을 촉진하는 방향으로 변화가 이루어졌다[19].

그러나 개인정보법 개정 이후에도 가명처리 기준이나 방법이 모호하여 특례규정을 활용하는데 제약이 있다는 의견을 있었고, 이를 반영하여 보건복지부와 개인정보보호위원회가 합동으로 「보건의료데이터 활용 가이드라인」을 발표해 보건의료분야의 빅데이터 활용에 대한 구체적 기준을 제시하고자 하였다[20]. 그럼에도 데이터 수집 과정에 대한 윤리적 우려는 완전히 해결되지 않았다. 가령, 국가인권위원회는 데이터 3법이 데이터 제공자의 인권보호에는 미흡함을 지적하였는데, 가명정보를 결합, 활용하는 과정에서 개인을 재식별화 할 가능성이 있음에 우려를 제기하였다[21].

보건의료영역에서 인공지능을 사용하는 것은 민감정보 사용 연구 일반의 고려사항 외에 추가적인 고려사항이 필요하다. 이를테면 데이터 수집 과정에서 사생활 보호를 위해 단순히 개인식별 정보를 제거하는 것 이상의 과정이 필요할 수 있는데, 영상 이미지를 의료 인공지능 학습에 사용하고자 할 때 이미지가 개인을 특정화할 가능성이 없는지 추가적으로 확인하는 과정을 거쳐야 할 수 있다[10,11]. 더 나아가 다른 윤리가치와 사생활 보호 간에 상충되는 상황이 발생할 수 있다. 영국 의학 왕립 아카데미(Academy of Medical Royal Colleges)와 유럽평의회(Council of Europe) 등은 사생활 보호와 정확성의 균형을 잡는 것이 중요함을 지적하고 있는데, 인공지능 알고리즘의 정확성과 진화는 양질의 대량의 데이터 가용성에 달려있기 때문이다[9,14]. 즉, 빅데이터를 활용하는 보건의료영역 인공지능 기술의 발전과 정확성 향상을 위해서 기존의 개인정보 보호 원칙을 어느 정도 완화할 수 있는지에 대한 일정한 사회적 합의점이 필요하다.

종합하자면, 보건의료 인공지능을 활용하는 데 있어 기술적, 법적, 거버넌스 차원에서 인공지능의 특성을 인지할 필요가 있고, 개발에서부터 임상현장에서의 적용에 이르기까지 이해관계자들 모두가 사생활 보호의 윤리적 중요성을 인식할 필요가 있다. 또한 다량의 의료데이터를 이용, 연계하여 인공지능을 개발할 때 의도치 않은 재식별 위험 등을 지속적으로 관리하고 예방하여야 하며, 이를 위해서 법적인 규제 방안을 마련하는

한다고 권고한다[12].

7) 가명정보란, 가명처리를 거쳐 생성된 정보로서 그 자체로는 특정 개인을 알아볼 수 없도록 처리한 정보를 의미한다. 시간 · 비용 · 기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보'인 익명정보와는 달리 추가정보를 활용할 경우 재 식별이 가능할 수 있다는 점에서 구분된다. 추가정보란, 개인정보의 전부 또는 일부를 대체하는 가명처리 과정에서 생성 또는 사용된 정보로서 특정 개인을 알아보기 위하여 사용 · 결합될 수 있는 정보(알고리즘, 매핑테이블 정보, 가명처리에 사용된 개인정보 등)를 의미한다[19].

것을 넘어선 데이터 보호 거버넌스 체계가 마련되어야 한다. 뿐만 아니라 인공지능의 알고리즘 정확성 및 의학적 유용성과 사생활 보호 사이에 상호 갈등이 있음을 인식하고 균형점을 수립하기 위해 사회적 합의가 이루어질 필요가 있다.[8)]

2) 형평, 공정, 반차별, 그리고 평등, 사회정의

앞서 윤리원칙을 매핑 및 도출하는 과정에서 데이터 수집 영역에서 형평, 공정, 반차별의 원칙, 그리고 사회적 환경 영역에서 결과의 적용의 불평등함에 대한 평등, 사회정의 원칙이 포함되었다. 유럽의회 연구기구(European Parliamentary Research Service)가 발간한 "Artificial Intelligence in Healthcare"는 보건의료분야 인공지능이 공통적으로 (1) 성별과 젠더, (2) 연령, (3) 민족, (4) 지리적 위치, (5) 사회경제적 상태에 따른 편향을 가질 수 있음을 지적한다[15]. 이러한 편향성의 원인으로는 기존 데이터셋의 불균형, 사회 내 구조화된 차별과 편향의 결과로 인한 데이터 수집단계의 문제, 그리고 개발팀의 다양성 부재로 인한 인공지능 개발 과정의 편향 감수성 미비 등을 들 수 있다. 또한 테크놀로지 장비 보급 불균형으로 인한 데이터 수집 및 보건의료 인공지능 공급과 활용인력의 불균형도 편향 및 불평등의 주된 원인이 될 수 있다[15,22]. 이를 데이터 수집영역(a. 인공지능 학습 데이터의 편향)과 결과 적용(b. 인공지능 적용의 불평등)영역 두 가지로 나누어 분석하면 다음과 같다.

a. 인공지능 학습 데이터의 편향

빅데이터를 활용하여 훈련하는 딥러닝 인공지능의 특성 상 투입 데이터가 편향될 경우 결과값도 그 편향을 그대로 반영하게 된다. 투입 데이터의 편향은 학습 데이터 셋이 다양한 인구집단의 특성을 반영하지 못하는 결과일 수 있으며(표집 편향: sampling bias), 무의식적으로 사회에 내재하는 편견이 데이터 수집, 분류 과정에서 작용한 결과일 수 있다(암묵적 편향: implicit bias). 딥러닝 인공지능의 블랙박스 속성으로 인해 일단 편향이 내재하게 되면 그 존재를 파악하기 쉽지 않기에, 이를 사전에 예방하는 절차가 요구된다.[9)]

일반적으로 빅데이터를 활용하는 인공지능의 개발 과정에서 더 많은 데이터를 확보할수록 더 높은 정확성을 갖춘 결과값을 산출해내는 인공지능을 제작할 수 있을 것이라는 기대가 존재한다. 따라서 데이터 불균형으로 인한 편향 가능성에 특별한 주의를 기울이지 않을 경우 연령, 성별, 성적지향, 체중, 신장, 인종, 사회경제적 상태, 교육수준 등 다양한 속성에서의 데이터 편향이 고려되지 않은 채로 무분별하게 데이터가 수집되고, 그 결과 편향을 내재한 인공지능이 개발될 가능성이 있다[10].[10)]

Ghassemi가 지적하듯이 각종 편향은 일상적 임상 환경의 일부로 자리하고 있으며 보건의료

8) 예를 들어, 더 다양하고 많은 양의 학습데이터를 확보하는 것이 알고리즘의 정확성과 의학적 유용성을 높이는데 도움이 될 수 있지만, 알고리즘 학습을 위해 수집할 데이터의 양과 수준을 결정할 때는 해당 데이터의 민감도 등 사생활 침해 위험도 반드시 검토해야 하는 제약이 있다.

9) 블랙박스 속성이란, 딥러닝 알고리즘이 인공신경망 등의 모델을 활용함에 따라 인공지능 제작자가 산출된 결과에 이르는 과정을 설명하기 어려움을 의미한다.

10) 보건역학적으로 연령, 성별, 인종, 사회경제적 상태, 교육수준에 따라 호발하는 질환의 차이는 존재한다. 따라서 표집 편향 및 암묵적 편향의 방지에 대한 요구가 이러한 차이를 인공지능에 반영하는 것을 막고자 하는 것은 아니다. 문제는 편향에 대한 사전의 주의 없이는 기존 사회의 편향이 인공지능 개발 과정에서 확대 재생산될 수 있다는 점이다. 많은 이들이 사전 데이터 학습 과정에서 사회 내에 존재하는 각 집단들의 데이터를 잘 분류하고 충분히 반영함에 따라 특정 소수자 집단이 다수 집단을 과대 대표하는 데이터셋을 통해 학습한 인공지능이

인공지능이 가지는 편향은 결코 우리 외부에 존재하는 것이 아니다[23]. 즉, 인간이 생산하고 라벨링하고 주석을 단 데이터를 이용하여 인공지능을 트레이닝 할 경우 인간이 원래 가지고 있던 편향이 인공지능의 메커니즘이 될 수 있다. 특히 기존의 무의식적 편견이 인공지능의 일부가 될 경우 쉽게 인식되지 않는 편견이 인공지능을 매개로 사회에 지속될 위험이 있다. 즉 인공지능이 불평등의 영속(perpetuation of inequality)에 기여하게 된다[8,23,24]. 의료 인공지능이 불평등한 결과를 초래한 예는 이미 보고되고 있다. 인종과 관련된 사례를 들자면, 흑색종(melanoma)을 이미지를 통해 진단하는 의료 인공지능은 흑인의 병변을 제대로 판별해 내지 못한 것으로 드러났다[25].[11] 또한 진통제 처방율에도 인종별로 유의미한 차이가 발견된 만큼 이러한 무의식적 관행이 인공지능에도 이식될 것이라는 우려가 존재한다[27]. 편견을 보정하지 않은 데이터 수집 및 알고리즘 개발은 결과의 정확성에도 치명적인 결과를 가져온다. 이는 의료기기에서 가장 기본적으로 요구되는 안전성과 효과성에 영향을 미치는 문제라는 점에서 더욱 세심한 관리가 필요하다.

공정은 현대사회의 핵심가치로서 인공지능 분야에서도 공정에 대한 요구가 강하다. 인공지능의 성능이 정확하고 우수하더라도 그것이 공정의 가치를 희생하여 얻는 결과가 되는 것은 바람직하지 않다. 이러한 요구에 부응하여 인공지능 개발과 활용에 관여하는 이해관계자는 과정 전반에서 공정의 가치를 보장할 수 있어야 한다[7]. 이는 보건의료 인공지능의 블랙박스 특성으로 인해 산출된 결과가 편향되었는지 예측하기 쉽지 않기에 인공지능 개발 단계부터 편향이 내재될 가능성을 염두에 둘 필요가 있기 때문이다. 또한 암묵적 편향이 인공지능에 반영되는 것을 최소화하기 위해서 편향에 대한 사회 전체의 감수성을 높일 필요가 있으며, 인공지능 개발 및 활용 인력의 다양성을 갖추어 암묵적 편향에 대비할 필요가 있다[7,15].

실질적으로 공정한 인공지능을 구현하기 위한 대책은 다양한 문헌에서 언급되고 있다. 예를 들어 유럽의회 연구 기구는 편향의 완화 수단으로 대표성 있고 균형 잡힌 데이터셋을 통해 훈련할 것, 사회과학자를 포함하여 다학제적 접근을 시도할 것, 의료 인공지능 개발 영역의 다양성을 증진할 것을 제안한다. 그와 더불어 인공지능 생애주기 전반에 걸쳐 잠재적인 편향을 평가하고 이를 검사, 모니터 하는 절차를 수립할 수 있어야 한다고 제안한다[15]. 이는 앞서 언급했듯이 인공지능의 블랙박스 특성으로 인하여 편향을 모니터링 하는 절차가 부재할 경우 편향의 존재를 파악하기가 대단히 어렵기 때문이다.

그러나 다른 한편으로 연령, 인종, 거주지, 사회계층 등으로 환자를 분류하여 데이터의 균형을 맞추는 과정에서 낙인효과 등으로 인해 환자에 부정적 영향이 가지 않도록 주의해야 한다는 지적도 존재한다. 이는 라벨링과 낙인효과에 거부감을 느낀 소수자 계층이 데이터 제공에 소극적이게 될 수 있다는 점에서 더욱 그러하다[9].

결과치 편향을 감소시킬 수 있는 좋은 방안 중 하나는 제작된 인공지능이 다양한 타겟 집단 모두에 높은 정확성을 보이는지 확인하는 것이다[15]. 하지만 데이터 및 예산 부족 등으로 인해 불가피하게 모든 인구집단을 대표하는 데이터셋

활용되는 것을 예방할 수 있고, 결과적으로 연령, 성별, 인종, 사회경제적 상태에 따른 차이를 잘 반영하는 인공지능이 개발될 수 있다고 주장한다.

11) 흑색종을 진단하는 한 인공지능 개발 과정에서 이미지 데이터셋 중 흑인 환자의 데이터는 5%에 불과하였고, 그 결과 흑인 환자에 대한 인공지능의 정확성은 백인환자 대비 절반에 불과하였다[26].

을 활용하지 못했다면 차선책으로 어떤 인구집단을 타겟으로 검토하여 인공지능을 제작하였는지, 그리고 어떠한 집단에 인공지능이 높은 정확성을 가지는지를 명시해야 한다고 싱가포르 보건부는 "Artificial Intelligence in Healthcare Guidelines(AIHGIe)"에서 제안하고 있다[11]. 이는 편향 및 불공정의 근본적 해결과는 거리가 있지만 편향의 가능성을 미리 고지한다는 점에서 불공정한 인공지능으로 인한 오류의 가능성을 줄일 수 있는 한 방법이 될 수 있다.

b. 인공지능 적용의 불평등

보건의료 인공지능의 혜택이 불평등하게 분배되지 않도록 하는 것 역시 중요한 윤리적 과제이다. 특히 저소득, 저개발 지역에 거주하는 집단에게는 데이터를 수집할 인프라가 부족할 뿐 아니라 개발된 보건의료 인공지능의 공급 역시 어려울 가능성이 크다. 더욱이 희귀질환이나, 유병률이 떨어지는 질병의 경우, 소수민족/인종의 대표성 획득 어려움으로 인해 이들 지역에 거주하는 이들에게 보건의료 인공지능의 정확성이 떨어질 수 있다[28].

따라서 의료 인공지능의 개발의 이익을 다양한 인구 집단에 평등하게 분배하기 위해서 개발과정과 마찬가지로 인공지능 전 생애주기에 걸쳐 불평등의 발생을 모니터링 할 필요가 있다. 특히 중저소득 국가의 경우 편향에 대한 사회적 분석의 부족, 낮은 기술활용능력, 소수자에 강한 편견 및 법적 보호 미비로 인해 편향이 더 강한 영향을 미칠 수 있음이 고려되어야 한다[28]. 이러한 문제의식에서 보건의료 인공지능 알고리즘이 다양한 맥락과 하위집단에서도 잘 적용될 수 있는 지를 가늠할 수 있는 적합성(appropriateness) 척도를 보건의료 인공지능의 중요한 조건으로 제시하는 문헌도 있었다[29].

다양한 차원에서의 편향과 불평등 문제를 해결하기 위해 UNESCO는 각국이 상호 다른 연령 그룹, 장애인, 언어 그룹, 장애인, 여성 외 기타 소외 집단의 인공지능 시스템에 대한 포괄적 접근성을 높이고 디지털 격차를 억제할 수 있어야 한다고 권고한다[13]. 인공지능 적용 불평등의 근본원인은 사회의 전반적 불평등 문제에 기인하는 만큼 단기적 해결은 어려울 수 있다. 그럼에도 의료 인공지능이 결과적으로 현재의 불평등을 심화시키지 않으며 완화에 기여할 수 있도록 설계와 배포 단계 때부터 고려할 필요가 있다.

3) 자율성 및 동의

보건의료영역 인공지능 윤리에서 자율성 및 동의 이슈 역시 두 가지 영역으로 나누어 생각해 볼 수 있다. 데이터 수집 단계에서 동의를 얻는 것, 그리고 인공지능을 임상현장에서 사용할 때 그 활용에 대한 동의를 얻는 문제이다. 이들 중 데이터 수집 단계의 동의 문제는 사생활 보호 원칙에서 다루고 있다. 아래에서는 임상진료에서의 인공지능 활용과 관련한 자율성 및 동의의 문제에 대해 주로 다루고자 한다. 이는 보건의료 인공지능 활용이 기존의 의료기기 활용과 마찬가지로 추가적인 동의를 받을 필요가 없는 통상적 진료에 해당하는가에 대한 논의로 귀결된다.

WHO는 보건의료영역에서 인공지능을 비롯한 기계의 자동화 수준이 높아지는 것이 인간의 자율성을 침해해서는 안 된다는 점을 기본 원칙으로 삼고 있다. 이는 기존 의료 맥락에서 의료인

과 환자 간의 소통에 기반한 결정과정에 기계가 개입함에 따라 인간 행위자의 자율성[12]이 침해될 수 있다는 우려에서 비롯된다. 다시 말해 인공지능의 의사결정과정 및 소통방식은 기존의 의사-환자간 소통과는 질적으로 달라서 환자가 그를 이해하고 소통하는데 어려움을 겪을 수 있는데, 그 결과 환자가 자율적 의사결정을 할 여지가 줄어들 수 있다는 것이다[12].

현 시점에서 보건의료 인공지능 발전수준에서 의사의 개입 없이 인공지능이 최종 진단을 내리는 것이 바람직하지 않다는 데는 여러 문헌들의 견해가 일치한다[12,13]. 하지만 최종진단이 아닌 의사의 진단을 보조하는 수준의 의료 인공지능의 활용할 경우 환자에게 이를 알리고 동의를 받아야 하는지에 대해서는 다양한 의견이 존재한다[10,30,31]. 예를 들어 싱가포르 보건부에서 발간한 "Artificial intelligence in Healthcare Guidelines(AIHGIe)"과 UNESCO의 "Recommendation on the ethics of artificial intelligence"에서는 진료과정에서 인공지능이 개입할 경우, 이를 미리 고지하고 사용에 대한 동의를 구하는 것을 원칙으로 제시한다. 이러한 관점은 인공지능 활용을 환자에게 공개하고 설명함으로써 진료 과정의 투명성을 높이는 데 주안점을 두고 있다[11,13].

그러나 무조건적인 공개에 회의하며 인공지능의 활용으로 인한 위험의 특성과 그 발생가능성에 따라 공개 여부를 결정해야 한다는 시각도 존재한다. Cohen은 보건의료에서 인공지능이 의사의 감독하에 보조적으로 사용될 경우 이는 통상적 진료의 일환으로 그 사용을 공개를 할 의무는 없으며, 추후 인공지능 사용이 보편화 될 경우 더욱 그럴 것이라는 의견을 제시한다[31]. 현재 이미 현장에서 동의과정 없이 인공지능 사용이 이루어지고 있음을 받아들여야 한다는 입장도 존재한다. Kiener의 경우 블랙박스 속성을 지닌 인공지능은 편향과 사이버공격에 완전히 자유로울 수 없기에 그 사용을 공개하는 것이 옳다고 주장하지만, 그럼에도 현재 임상현장에서는 위험성이 크지 않다고 판단되는 인공지능의 경우 동의 과정을 거치지 않고 사용되고 있다고 지적한다[30].

인공지능 활용에 대한 동의를 받는다면 어떤 수준의 정보 및 설명에 근거하여 동의를 받아야 할지에 관해서도 검토가 필요하다. 유럽평의회에서는 인공지능 활용에 대한 정보를 공개할 경우 인공지능 시스템이 어떻게 기능하는지, 어떻게 설계되고 검증되었으며 운용되고 있는지, 그리고 시스템을 조사하기 위해서는 어떤 정보가 필요한지에 대한 설명이 필요하다고 주장한다. 그와 더불어 인공지능이 가지고 있는 잠재적인 위험요인 역시 공개될 필요가 있다고 주장한다[14]. 향후 충분한 설명에 근거한 동의(informed consent)가 가능하기 위해서는 인공지능에 대한 투명하고도 설명가능한 정보 공개가 선행될 필요가 있을 것이다.

4) 안전성, 보안성, 견고성

현재 인공지능 개발 분야에서는 안전성, 보안

12) 인공지능이 진료 과정에서 일부 역할을 담당함에 따라 의료인과 환자의 의사결정 과정에서의 자율성이 침해될 가능성이 있다. 보건의료 인공지능의 일차적 목표가 환자에게 더 질 높은 의료를 제공하는 데 있는 만큼 본 글에서는 환자의 자율적 의사결정의 침해 가능성에 주목하여 논의를 진행하고자 한다. 의료인의 자율성에 대해서 덧붙이자면, 현 시점에서 보건의료 인공지능의 도입은 의료인이 최종결정권을 가진다는 전제 하에 도입되고 있다. 6) 책무성, 책임성 파트에서 다루게 될 휴먼인더루프(human-in-the-loop, HITL) 개념이 이와 일맥상통한다. 그리고 추가적으로 인공지능의 도입으로 인해 변화하게 될 진료환경에 대한 의료인의 적응을 도울 교육 시스템 및 거버넌스 체계가 구비될 필요가 있을 것이다.

성, 견고성 등이 필수적인 가치로 등장하고 있다. 먼저 용어의 구체적 의미를 살펴보면, 안전성(safety)은 위험과 관련된 용어이며 인공지능 시스템 운영에 있어 시스템에 위협을 일으킬 수 있는 조건을 최소화시키는 것을 의미한다. 보안성(security)은 사이버보안(cybersecurity)[13]의 개념을 포함하며 외부의 공격 혹은 예상치 못한 시스템의 변화에 대응할 수 있음을 뜻한다. 견고성(robustness)은 인공지능 시스템에 통제불능한 요소가 존재하는 경우에도 제대로 기능할 수 있으며 발생 가능성이 있는 오류를 관리할 수 있음을 지칭한다. 신뢰성(reliability)은 수용될 수 있는 범위 내에서 통계적으로 편차가 적은 결과를 생성할 수 있는 지의 여부를 말한다[32].[14] 이러한 네 가지 요소들은 위험관리에 연계된 내용에 가깝다.

여러 문헌들은 인공지능이 의료적 의사결정에서 인간을 부분적으로 대체하고 보조하기 위해서는 보건의료 인공지능의 안전성 대한 검증기준과 그 검증이 핵심적인 임상도입의 요건이어야 한다고 말한다[14,33]. 이를 위해서 인공지능시스템 전 생애에 걸쳐 안전성의 손상이 최소화되어야 하며, 이를 보장할 수 있는 규제와 거버넌스가 필요하다.

또한 외부의 공격으로부터 보안성을 갖춘 인공지능을 갖추기 위해서는 자체보안성능이 구비되어 있어야 하며 잠재적 보안 위협에 대한 지속적인 모니터링 시스템과 보안 문제가 발생했을 때의 복구 시스템 역시 구비되어 있어야 한다.[15] 필요에 따라 인공지능 활용 과정에서 그 활용을 중단하여야 하는 최소한의 보안 하한선을 정해 두어야 할 수 있다[11].

그러나 보건의료 인공지능이 기술적 차원에서 안전하고 신뢰 가능하며 오류산출 가능성이 적다 하더라도 임상적 적용에서는 또 다른 환자 안전에 대한 위협이 될 수 있다. 계산 상에서는 오류가 없다고 하더라도 현실 맥락을 반영하지 못함으로써 위해를 가져올 수 있는 것이다. 가령 한 의료 인공지능은 배경지식에 대한 인지 불능으로 인해 천식을 기저질환으로 가진 경우 그렇지 않은 경우보다 폐렴 감염의 예후가 좋다고 판단하였다. 인공지능이 데이터를 수집한 의료기관에서는 천식을 기저질환으로 가지고 있던 환자는 폐렴으로 병원에 내원할 경우 즉시 중환자실 입원 대상이었고, 중환자실의 집중치료는 매우 효과적이었기 때문에 생존율이 일반 환자보다 오히려 높았다. 인공지능은 이러한 맥락을 이해하지 못하여 천식과 생존률의 상관관계만을 고려했고 결과적으로 잘못된 인공지능 알고리즘을 갖추게 된 것이다[35]. 특히 딥러닝 인공지능이 가지는 블랙박스 특성으로 인해 인공지능이 위해를 발생시킬 가능성을 내재하고 있다고 하더라도 그를 파악하기 어려울 가능성이 존재하며, 만약 오류를 감지해 내기 어렵거나 혹은 오류로 인해 연쇄반응이 이루어 진다면 심각하게 환자에게 위해를 가져다 주는 결과를 초래할 수 있다[8].

임상현장에서 발생할 수 있는 위해를 최소화하기 위한 더 적극적인 대응으로, 인공지능 전반에

13) 인공지능이 의료에 미치는 영향이 커짐에 따라, 의도적인 인공지능에 대한 공격으로 인한 작동 중지 및 데이터 탈취로 인한 위험 역시 커지고 있다. 사이버보안은 이러한 의도적 공격에 대응할 수 있는 능력에 대한 개념으로, 주로 인프라 및 기술적 개선을 통해 이루어지고 있다[12].

14) 신뢰성이 문제될 수 있는 예로, 인공지능이 내놓는 결과 값 편차가 인간 의료진의 통상적 진료에 비해 클 경우 일부 환자가 위험해질 수 있다는 우려를 제기한다[9]. 보건의료 인공지능 기술은 지속적으로 발전하고 있기에 신뢰성 문제는 점차 해결될 것이라는 견해도 있지만, 질 나쁜 의료 인공지능이 환자에 위해를 가할 가능성에 대해서는 지속적인 주의가 필요하다.

15) 식약처에서는 「의료기기의 사이버보안 허가 · 심사 가이드라인」을 배포하여 의료기기 제작자가 위험분석, 위험통제 등 위험관리를 적용하도록 권고하고 있다[34].

걸쳐 오류가 발생했을 때 이를 찾아내고 보완할 수 있는 개념인 추적가능성(traceability)을 최대한 보장해야 된다는 견해가 있다[1]. 추적가능성은 투명성, 설명가능성과도 큰 관련성이 있는데, 의료에 존재할 수 있는 모든 위협요소를 인공지능의 전 생애주기에 걸쳐 찾아내고 보완하기 위해서는 위협을 확인하고 해결할 수 있는 가능성을 가지고 있어야 하기 때문이다.

더 나아가서 EU에서는 견고함의 개념을 기술적인 차원을 넘어 사회적 차원으로 확대한다. 견고한 인공지능(robust AI)을 윤리적일 것(ethical), 적법할 것(lawful) 외 인공지능의 주요 3요소 중 하나로 설명하고 있다. 이는 사회적 평등, 지속가능성 등 사회적, 범지구적으로 중요한 가치 역시 손상시키지 않는 방향으로 인공지능이 개발되어야 한다는 취지이다. 이러한 시각은 평등, 지속가능성 등 다른 윤리적 차원과 맞닿아 있다고 할 수 있다[32].

인공지능의 안전성을 확보하고 신뢰를 증진시키기 위해 유럽평의회는 기술적 차원 이상의 인공지능 평가 체계를 갖출 것을 요구한다. 즉, (1) 인공지능 의료기기의 임상적 역할을 명확히 세울 것, (2) 정확성 이외에 갖추어야 할 능력을 규정할 것, (3) 평가 과정을 간단한 것부터 차례대로 세분화할 것, (4) 제 3자를 통한 외부 평가 과정을 증진할 것 (5) 인공지능 평가 결과 보고를 위한 표준화된 가이드라인을 활용할 것을 제시하고 있다[14]. 이러한 방법을 통해 의료적으로 불가피한 범위 이상의 위해와 안전성 위협은 인공지능 생애주기 전체에 걸쳐 최소화 될 수 있어야 하며, 예상치 못한 변화 혹은 위해 발생 가능성이 확인될 시 이에 대응하는 대응책이 마련되어 있어야 한다[2,13].

5) 투명성, 설명가능성, 해석가능성

설명가능성 및 투명성의 문제 또한 딥러닝 인공지능의 특성인 블랙박스 속성과 관련성이 크다. 딥러닝 인공지능은 인간의 뇌신경망을 모방하여 알고리즘 내부에서 여러 단계의 네트워크를 구성하여 결과값을 산출하는데 이 과정에서 결과값을 내는 프로세스를 기존의 인간의 학습 및 정보전달 방식으로 제시하는 것이 어렵기 때문이다. 이러한 어려움은 의료 인공지능을 실제로 활용하는 의료진과 환자 뿐 아니라 인공지능 개발자에게도 마찬가지로 적용된다. 이는 딥러닝 인공지능이 지니는 독특한 특성으로, 보건의료 인공지능의 개발 및 활용의 윤리적 논의에서도 설명가능성 및 투명성의 문제는 빈번히 다루어지고 있다.

투명성, 설명가능성, 해석가능성 등 용어들의 구체적 의미를 살펴볼 필요가 있다. 우선, 투명성(transparency)은 인공지능이 산출한 결과값이 원래 의도했던 결과값인지 평가하기 충분한 정보를 제공할 수 있도록 사전에 설계되어 있어야 함을 말한다. 설명가능성(explainability)은 인공지능 모델의 운영방식에 대한 설명이 쉽게 이해될 수 있어야 한다는 의미이며, 해석가능성(interpretability)은 인간이 의사결정을 내리기에 유용한 결과값을 인공지능이 도출할 수 있어야 한다는 의미이다[10].

인공지능이 편견을 가지는지 혹은 인공지능이 잠재적 위해의 가능성을 가지고 있는지 등 보건의료 인공지능의 윤리적 문제를 파악하기 위해서는 투명성과 설명가능성이 선행되어야 한다. 즉, 투명성과 설명가능성은 윤리적 인공지능이 기본적으로 갖추어야 하는 전제조건이다[13,36].[16)]

16) 인공지능의 투명성과 설명가능성이 의료 사고 대처에 유용함은 여러 예에서 알려져 있다. 한 예로, 마운트 시나이 병원(Mount Sinai

우선 투명성은 기술의 재현가능성을 평가하기 위한 필수조건이다. 유럽의회 연구기구에서는 투명성이 결여될 경우 (1) 인공지능 알고리즘에 대한 독립적 재현 및 평가가 어려워지고, (2) 인공지능 오류의 원인 파악 및 책임소재 파악이 어려워지며, (3) 인공지능의 예측과 결정에 대한 신뢰 및 이해가 불가능해져, (4) 임상진료 및 일상생활에서 인공지능 도구 채택이 제한될 수 있다고 설명하고 있다[15]. 그리고 WHO는 투명성이 요구되는 내용에는 기술이 전제하는 조건과 그 한계, 운영 프로토콜, 데이터의 속성, 알고리즘 모델 개발의 속성에 대한 정확한 정보를 포함한다고 말한다[12].

그러나 인공지능에는 투명성 뿐만 아니라 설명가능성의 가치 또한 필요로 한다. 인공지능이 특정한 결과값을 낸 판단의 근거가 무엇인지, 인간의 언어로 해석하고 설명하고 이해가 가능해야 하기 때문이다. 의료진은 보건의료 인공지능과 효과적으로 상호작용하고 그 산출결과를 이해하며 그 결과에 대한 정보를 환자와 공유하기 위해서 투명성에 더해 설명가능성을 갖출 필요가 있다[2]. EU의 GDPR 역시 모든 사람은 자동화된 인공지능 프로세스 뒤에 존재하는 의미 있는 정보를 제공받을 권리가 있음을 언급하여 설명가능성의 중요성을 드러내고 있다[10].[17)]

유럽의회 연구기구는 설명가능성 문제를 보완하기 위한 여러가지 복안을 제시하는데, (1) 'AI 패스포트'를 창출하는 것,[18)] 2) 인공지능 알고리즘을 모니터링 할 수 있는 추적가능한(traceable) 도구를 개발하는 것,[19)] (3) 임상현장에서 사용자가 설명가능한 인공지능 설계에 참여하는 것, (4) 추적가능성[20)]과 설명가능성을 인공지능 의료기기 인허가의 전제 조건으로 제시할 것 등을 들고 있다. 이러한 복안들은 인공지능 기술 도입 과정에서 설명가능성과 결과의 추적가능성 수준을 높일 것으로 일정 수준 기대된다.

그러나 일정 수준의 설명가능성과 추적가능성을 높이더라도 기존의 의료 수준의 설명가능성을 확보할 수 있을지는 의문이 크다. 현재 기술적 차원에서 딥러닝 인공지능의 설명가능성을 확보하기 위한 노력으로는 사후 설명가능성(post-hoc explainability) 확보 시도가 있다. 가령 영상의학적 이미지를 활용하는 의료 인공지능의 경우 히트 맵(heat map) 혹은 돌출 맵(saliency map)을 통해 어떤 영역이 결과값 산출에 큰 영향을 미

Hospital)에서 개발한 고위험군 환자 분류 인공지능이 마운트 시나이 병원 외의 의료기관에서 그 정확성이 크게 떨어지는 것이 발견되었다. 그 이유를 파악한 결과, 마운트 시나이 병원에서는 고위험군을 위한 중환자실에서 특정한 X-Ray 기기를 사용하고 있었고, 인공지능은 메타데이터의 기기정보를 환자 분류 알고리즘에 포함하고 있음이 확인되었다. 이 인공지능의 경우 인공지능의 메커니즘에 대한 설명가능성이 어느정도 확보되어 있었기 때문에 타 병원에서 문제가 발견되었을 때 신속하게 대처할 수 있었다[36].

17) 그러나 EU-NA Radiologists Association에서 GDPR의 '설명에 대한 권리'가 과정의 결과를 예상, 직시하는 권리(envisaged consequences)에 가까우며, 특정한 개별 결정에 대한 설명과는 거리가 있다고 언급하다. 설명가능성을 충족하는 구체적인 설명 수준의 단계에 대해서는 앞으로 논의되어야 될 부분이다[10].

18) 여기서 AI패스포트는 인공지능의 투명성을 생애주기 전 과정에서 추적하기 위해 인공지능과 관련된 정보를 기록하는 규격이라 할 수 있다. 예를 들어 AI패스포트에는 인공지능이 어떤 기술을 활용하여 제작되었는지, 어떠한 데이터를 사용하였는지, 평가는 어떻게 시행하였는지에 대한 정보, 인공지능 운용 및 유지와 관련된 정보를 기록해 둘 수 있다. 기록해야 할 정보를 규격화 함에 따라 사용되는 특정 지역, 의료기관에 무관하게 지속적 모니터링을 용이하게 할 수 있다[14].

19) 아래에서 설명될 기술적 사후 설명가능성(post-hoc explainability) 확보가 그 예이다.

20) OECD에 따르면 추적가능성은 인공지능이 산출한 결과에 대한 분석 및 조사가 필요한 경우 인공지능의 데이터셋, 프로세싱, 의사 결정 과정 등 생애주기 전반에 대해 최신기술을 활용한 분석, 그리고 인공지능 활용의 맥락을 고려한 분석을 가능하게 하는 것을 의미한다[1].

쳤는지를 보여주는 시스템을 활용하는 것이다.[21] 하지만 이와 같이 인공지능의 작동원리를 개괄해 낼 수 있는 시스템을 마련한다고 하더라도 딥러닝 인공지능 메커니즘의 특성상 임상에서 개별 환자에 대해 인공지능이 산출해 낸 결과값에 대해 그 환자에게 충분한 설명을 제공하기 어려울 수 있다. 설명가능한 인공지능에 대한 여러 문헌들은 설명가능성이 맥락, 청자, 목적에 따라 다르게 적용될 수 있음을 지적한다[2,38]. 이는 어떠한 정보가 더 중요하게 사용되었는지를 제시하는 수준의 사후 설명가능성 기술로는 임상적 활용에 충분한 설명수준을 확보하지 못할 수 있음을 시사한다.

또한 딥러닝 인공지능의 내재적 특성상 설명가능성의 확보에 한계가 있다는 시각도 존재한다. 많은 문헌에서 인공지능의 설명가능성은 그 가치 확보를 위해 정확성을 희생하는 상충(trade-off) 관계에 있음을 지적한다[32,38,39]. 설계 단계에서부터 설명가능한 인공지능을 제작할 경우 딥러닝의 장점을 일정부분 희생해야 할 수도 있다는 것이다.[22]

사후 설명가능성의 확보, 인공지능 패스포트 등 설명가능성 및 추적가능성을 확보할 수 있는 여러 방안이 제시되고 있지만, 블랙박스 특성의 한계로 인해 설명가능성을 완벽하게 확보하는 것은 어려운 상황이다. 그러나 UNESCO가 지적하듯, 투명성과 설명가능성 등에 대한 분명한 요건의 수립이 인공지능의 신뢰성 확보에 바탕이 되어야 하며, 인공지능 기술 인허가 과정에서 핵심 이슈가 될 필요가 있다[13].

6) 책무성, 책임성

인공지능 맥락에서 책무성(accountability)은 위험한 결과가 발생했을 때 책임질 수 있는 주체에 관한 개념이다. 개별 운영자와 조직은 인공지능이 산출한 결과에 대해 응답하고 책임질 수 있어야 한다[40]. 책임성(responsibility)은 책무성과 거의 동시에 사용되는데, 역시 결과에 대한 책임과 관련된 개념이라 할 수 있다. 책무성 및 책임성 역시 설명가능성과 연계된 개념이라고 할 수 있다. 문제가 발생했을 경우 그 원인에 대해 이해하고 설명하는 것이 책무성의 핵심개념이기 때문이다[41].

보건의료 인공지능의 책무성 문제에서 다루어지는 중요한 이슈 중 하나는 의료 인공지능의 활용의 결과 오진이 발생했을 때 설명의 의무 및 책임이 누구에게 있는 가의 문제이다. 이는 법적 책임(liability)을 누가 감수해야 하는 지와 직결되어 있다. 인공지능을 활용한 진료에서 인간에게 위해가 발생할 수 있는 두 가지 경우 [(1) 인공지능을 따르지 않아 오진이 생기는 경우, 즉, 인공지능은 바르게 진단했으나 인간이 그를 따르지 않은 경우, (2) 인공지능을 따른 결과 오진이 생기는 경우, 즉, 인공지능은 잘못 진단했으나 인간이 그에 반하는 결정을 내리지 않음] 각 경우에 대해 인간의 책임은 어디까지인지, 책임을 져야 한다면 인공지능 개발자와 활용 의료진 중 누구의 책임인지에 대한 법적, 윤리적 불확실성이 해소되지 않은 상태이다. 이러한 불확실성은 책무성의 공백(gap in accountability)이라는 개념으로 표현된다[42,43]. 책무성의 공백이란 인공지능의 블

21) 예를 들어 흉부 엑스레이 이미지를 기반으로 폐렴을 진단하는 인공지능은 이미지의 어떠한 부분이 진단에 영향을 미쳤는지를 보여주고 있다. 이를 바탕으로 인공지능이 작동을 간접적으로나마 설명해낼 수 있다[37].

22) 설명가능성을 확보하기 위해 인공신경망 기술의 장점을 최대로 활용하지 못할 수 있기 때문이다.

랙박스 특성으로 인한 투명성 부족으로 인해 개발부터 활용에 참여하는 다양한 행위자의 역할과 책임을 규명하는 것이 어려움을 의미한다[14].

책무성의 공백으로 인해 의료인들이 법적 책임의 가능성을 우려하여 인공지능 도입을 꺼리게 될 수 있다. 구체적으로 책무성의 공백에 해당하는 요소로 유럽의회 연구기구는 (1) 책임과 책무성에 대한 명확한 법적 규제가 없는 점, (2) 의료 인공지능 개발 및 활용에 다양한 행위자가 참여하여 책임소재를 가리기 어려운 점, (3) 인공지능 산업에 법적, 윤리적 거버넌스가 미비한 점을 들고 있다[15]. 그와 함께 이러한 책무성의 공백을 완화할 수 있는 방안으로는 (1) 인공지능이 개인에 위해를 가했을 때의 개발자 및 사용자의 역할을 확인하는 프로세스를 도입하는 것, (2) 인공지능 개발자 및 사용자에 일관된 규제 프레임워크를 개발 및 적용하는 것, (3) 의료 인공지능을 규제하는 전문 규제기관을 설립하는 것이 제시되고 있다[15].

UNESCO에서는 인공지능 생애주기에 있는 모든 행위자들이 윤리적, 법적 책임감을 가지고 인공지능 시스템의 결정과 행동에 대한 책임을 나눌 수 있어야 한다고 말하며, 이를 위해 내부고발자 보호를 위한 적절한 감독, 영향평가, 감사 및 실사(due diligence) 메커니즘을 확보해야 한다고 언급한다[13]. 책무성의 공백 해소에 대한 현재의 요구는 인공지능 시스템의 감사가능성(auditability)을 염두에 둔다.

다만 OECD는 책무성을 인간 주체성 및 감독(human agency and oversight)과 구분하여 제시한다[2]. 여기서 인간주체성은 인공지능의 자동적 결정이 만들어낼 수 있는 무의식적인 종속, 조건화 등을 경계하고 이러한 특성에 대한 사전 설명이 이루어져야 한다는 개념이며, 책무성과 달리 고려해야 할 사안이라고 본다. 인간의 감독은 휴먼인더루프(Human-in-the-Loop) 휴먼온더톱(human-on-the-top), 휴먼인커맨드(human-in-command) 등의 용어와 함께 사용된다. 휴먼인더루프는 인공지능의 모든 결정 사이클에 인간이 개입할 수 있는 역량을 의미하며, 휴먼온더톱는 인공지능의 설계 사이클 및 운영 모니터링에 인간이 개입할 수 있는 역량을 의미한다. 휴먼인커맨드는 인공지능 시스템의 전반적 활동을 감독하고 이를 어떤 상황에서 어떻게 사용할 지를 결정할 수 있는 역량을 의미한다. 즉, 책무성이 부정적 영향의 최소화를 위한 감사가능성을 의미한다면, 인간 감독은 인간의 능동적 인공지능 시스템 개입 능력을 의미하는 것이다.

인간의 감독 관점에서는, 보다 주체적으로 인간이 인공지능을 제어하고 관리할 수 있기 위해 인공지능 시스템과 그 사용되는 환경에 대한 이해를 명확히 하고 인공지능에 대해 인간이 개입할 수 있는 충분한 시간을 확보하는 것이 필요하다. 가령 사이버 보안 문제가 발생시 제어하기 어려운 속도로 문제가 확산될 수 있는데 시스템적으로 인간이 개입할 여지가 미리 확보될 경우 여러 예외상황에서도 인간이 책무성을 가지고 이를 해결해 나갈 수 있게 된다[44]. 이것은 문제 발생 사후 감사가능성의 관점으로 책무성 문제를 접근하는 것과 다른 접근이 될 수 있다.

7) 지속가능성 및 포괄적 성장

1987년 세계환경개발위원회(World Commission on Environment and Development)에 따르면 지속가능한 발전은 미래세대가 누릴 수 있는 잠재적 이익을 침해하지 않으면서 현재 세대의 욕구를 충족시키는 발전을 의미한다[45]. 여러 문헌들은 지속가능한 개발과 포괄적 성장을 인공지능이 가져야 할 중요한 윤리적 특성으로 언급하고 있다[13,40]. 이는 사회, 문화, 경제, 환경 등 다양한 환경적 측면에서 인공지능이 지

속가능한 사회를 만들어 나갈 수 있도록 기여해야 한다는 의미이다. 여기에는 인공지능에 투입되는 비용 대비 편익이 지속가능해야 된다는 좁은 의미의 내용[23]에서부터 소수자 집단과의 포괄적 성장, 성별, 인종, 민족간 불평등 해소, 그리고 자연환경 보호의 개념이 모두 포함되어 있다고 할 수 있다. WHO는 지속가능성의 확보가 보건의료 인공지능의 도입을 위해 필수적임을 강조한다[12]. 즉, 다양한 차원에서의 불평등을 해소하고 지속가능한 근무환경, 자연환경의 유지를 토대로 보건의료 인공지능이 포괄적 성장을 이뤄낼 수 있을 것으로 기대한다.

현재 보건의료 인공지능의 도입에서 가장 많이 논의되는 지속가능성의 영역은 보건의료 근무환경이다. 보건의료 인공지능의 도입을 통해 환자의 의료 접근성 및 진료 정확성이 개선되어 환자에게 긍정적으로 작용할 것으로 기대되지만 의료상담, 조언, 치료과정에서 의료제공자의 역할에 큰 변화가 일어남에 따라 여러 예상치 못한 문제점을 유발할 수 있다[9]. 특히 보건의료 인공지능 시스템이 의료환경에 성공적으로 도입되기 위해서는 의료제공자가 인공지능을 신뢰하는 것이 중요한 요건이므로 인공지능 시스템이 보건의료환경에 성공적으로 정착되기 위해서는 의료제공자에 미칠 잠재적 영향을 고려하여야 한다[8]. 인공지능 도입으로 인해 의료인들은 자율성이 침해당한다는 생각이 들 수 있으며 요구되는 기술이 변함에 따라 환경에 대한 적응이 어려울 수 있다. 또한 의료인과 환자의 접촉 양상 또한 크게 달라질 수 있으므로 의료인들에게 이에 대한 대비 또한 요구된다. 따라서 근무환경이 지속가능하기 위해 정부 등 보건당국은 의료인이 바뀌는 근무환경에 쉽게 적응할 수 있도록 교육 프로그램을 제공해야 한다. 그와 더불어 인공지능 도입을 이유로 의료인력의 지나친 축소가 일어나지 않도록 감독하는 등 보건의료 직무환경의 지속가능성을 유지, 증진하는 노력이 꾸준히 동반되어야 한다[12].

의료 인공지능의 발전이 인적, 물적 의료 자원이 넉넉하지 않은 개발도상국의 의료의 질 향상에 기여하여 지속가능한 세계를 만드는 데 큰 기여를 할 것이라는 기대 역시 존재한다. 보건의료 인공지능은 의료에 필요한 비용을 절감하고 의료 접근성을 강화하며 전염성 질환 등 개발도상국에 호발하는 질환의 추적 및 예방에 도움이 될 수 있다[47].[24] 하지만 인공지능의 활용을 통한 개발도상국 의료의 질 향상을 위해서 추가적으로 고려해야 할 사항이 존재한다. 가령 개발도상국에서는 인공지능 알고리즘의 높은 질을 보장하는 인풋 데이터의 확보가 어려울 수 있다. 선진국에서 수집한 데이터를 바탕으로 제작한 보건의료 인공지능은 선진국과는 다른 질병특성을 가진 개발도상국의 건강상태를 반영하지 못할 수 있으며, 부족한 경제적 상황을 고려하지 않는 고비용 치료를 권할 수 있다는 점에서 효용성이 떨어질 수 있다. 또한 개발도상국에서는 윤리적이고 효율적인 의료 인공지능을 관리하는 거버넌스 체계가 미비할 가능성도 있다[12,40,47]. 따라서 인공지능의 개발도상국 도입 및 확산 과정에서 이러한 특성에 대해 유념해야 할 필요성이 존재한다.

환경 친화적 의료시스템 구축에 보건의료 인공지능이 기여할 수 있을 것이라는 기대도 존재한다. '녹색 생명윤리(green bioethics)"의 관점에

23) 영국 국가보건서비스(National Health Service, NHS)는 NHS에 도입하는 인공지능 도구의 시스템 평가 항목 중 하나로 비용 대비 이익이 지속가능해야 한다고 제시하고 있다[46].

24) 비용 절감은 인적자원의 절감뿐 아니라 물적자원의 절감으로도 이어질 수 있다. 예를 들어 안과 영역에서 인공지능 시스템을 활용할 경우 값비싼 다양한 종류의 안과전문진단기구의 구입을 최소화할 수 있다[47].

서 지속가능한 의료는 미래세대의 잠재적 이익을 희생하지 않는 현세대의 발전으로, 정의의 원칙(principle of justice)에 부합하는 중요한 윤리적 가치이다. 이러한 관점에서 자원의 소비를 절감하는 응급의료 인공지능 트리아지 시스템(triage system)은 녹색생명윤리의 관점에서도 바람직한 인공지능 개발이 된다. 반면 거동이 불편한 환자를 모니터링 하는 의료 인공지능에 오류가 발생하여 환자의 부상을 방치한다면 이는 의료적으로도 불량한 인공지능임과 동시에 자원낭비를 유도하는 인공지능에도 해당할 수 있다[48].

8) 기타 고려해야 할 사항들

a. 의사–환자 관계의 신뢰

보건의료 인공지능이 의사–환자 관계에서 새로운 행위자로 참여하게 되면 의사–환자 관계에도 큰 변화가 일어날 것이라 예측되고 있다. 환자와 의료 인공지능의 접촉이 늘어날수록 실제 의사와의 접촉이 줄어들고 의사–환자 관계에서 얻을 수 있는 특수한 의료적 효용성이 손상될 수 있다는 우려가 존재한다[49].[25] 따라서 의료 인공지능의 도입으로 인해 변화가 자명한 의사–환자 관계의 신뢰도를 희생시키지 않도록 하는 것 또한 중요하게 거론되는 윤리적 과제이다. EU에서는 보건의료 인공지능이 사생활 침해, 의료접근 불평등 강화, 투명성 손상, 사회적 편향 위험, 의료진의 책임 희석,[26] 자동화 편향 유발,[27] 의료진의 비숙련화 유도, 책무성의 공백 등 앞서 살펴본 여러 윤리적 문제들을 유발하여 의사–환자 관계의 신뢰수준을 저하할 수 있음을 우려한다. 즉, 보건의료 인공지능의 윤리성을 확보해야 실제 활용에 있어 의사–환자 관계의 신뢰가 유지될 수 있다[14].

여러 우려에도 불구하고 의사–환자 관계가 기존의 의사의 권위에 기댄 가부장적 모델에서 환자의 질병자기결정 모형으로 바뀌어 가고 있기에 의료 인공지능의 도입이 크게 현재 의료 경향을 크게 바꾸지 않을 것이라는 예측도 존재한다. 또한 의사가 환자의 돌봄 및 감정적 영역에 더욱 집중할 수 있기에 의사–환자 관계는 더 긴밀해질 것이라는 의견 역시 존재한다[8]. 하지만 그럼에도 의사–환자관계에서 환자의 취약성은 여전히 존재하며 따라서 의사에게는 여전히 (선량한 관리자의) 주의 의무(fiduciary duty)가 존재한다고 할 수 있다[14]. 이러한 우려를 불식시키기 위해 유럽의회 연구기구에서는 전체적으로 돌봄의 질 수준을 높이는 목표 아래 보건의료 인공지능 개발 및 사용자와 환자 사이의 적절한 가치균형을 이루어야 한다고 권고한다[15].

b. 대중 신뢰, 다자 참여와 협력

현 시점에서 보건의료 인공지능 개발은 민간기업에 의해 주도되고 있다. 인공지능 개발자들이 환자의 빅데이터를 수집하고 임상시험을 하고 규제당국의 의료기기 허가를 받는 과정에서 일반 대중들의 참여가 이루어질 여지는 대단히 적다. 이러한 상황은 민간기업이 환자 데이터에 접근하

25) 구체적으로는 의료 인공지능로 인해 돌봄(care)의 비기계적 측면, 비언어적 측면을 저하시킴으로써 오랜 기간 형성되어 온 전문가의 환자에 대한 맥락화, 역사화 된 지식을 무력화할 수 있다는 우려가 존재한다.

26) 의료진과 인공지능이 환자 진료에 있어 역할을 분담하게 됨에 따라 의료진의 책임의식이 감소할 위험이 존재한다.

27) 자동화 경향은 인공지능이 산출한 데이터가 잘못되었을 가능성을 배제할 수 없을 상황에서도 그에 의문을 품지 않는 경향을 의미한다. 자동화 경향으로 인해 인공지능의 잘못된 결과값을 정정하지 못하고 과잉 의존함에 따라 환자의 건강이 훼손될 가능성이 있다[14].

고 이를 통해 수익을 얻는 것에 대한 대중의 거부감을 유발할 수 있다[8]. 만약 인공지능 시스템이 공공의 이익에 부합하지 않는다는 느낌을 받게 되면 환자들이 자신의 정보를 제공하기를 꺼릴 수 있다는 점에서 대중들의 참여 가능성은 윤리적 관점 그리고 의료 인공지능의 정확성 관점 모두에 중요하다 할 수 있다.

인공지능을 제작하는 과정에서 민간기업이 정부가 가진 데이터를 활용하는 등 여러 방면에서 공공-민간 파트너쉽(public-private partnership)이 형성된다. 이러한 파트너쉽과 관련된 정보는 대중들에게 투명하게 공개될 필요가 있으며, 그 과정에 대해 대중들 역시 참여할 수 있어야 한다. 특히 이러한 파트너쉽은 개인과 공동체의 권리의 보호에 적극적으로 참여하여야 하며, 제작된 인공지능으로 인한 이익이 모두에게 공유될 수 있도록 하여야 한다[12].

공공-민간 파트너쉽 운영은 유연한 거버넌스와 협력을 필요로 한다. UNESCO는 유연한 거버넌스와 협력을 통해 인공지능 시스템 생애주기에 다양한 이해당사자가 참여하여 포괄적 개발이 가능해야 한다고 지적한다. 여기서 이해당사자는 개발자 및 의료종사자 뿐 아니라 시민사회, 연구자 및 학자, 미디어, 교육계, 정책입안자를 포함하며, 제 3의 이해당사자[28]가 등장할 가능성 또한 염두에 둘 필요가 있다[13].

4. 윤리원칙을 어떻게 적용할 것인가? 숙의 과정을 통한 윤리원칙 수렴과 상충 해결 모색

앞서 본문에서 저자들은 투명성(설명가능성, 해석가능성)의 가치는 의료 인공지능의 정확성과 상충관계에 있으며, 또한 데이터 수집 과정에서 사생활 보호의 가치를 위해 정보수집동의를 구할 경우 인공지능의 학습데이터 양에 제약을 가하기에 정확성의 가치와 상충관계에 있음을 보였다. 그 외에도 다양한 윤리 가치를 준수하기 위한 개발과정에서의 추가적인 고려는 인공지능 제작의 비용을 상승시킴에 따라 의료 인공지능 개발로 인한 잠재적 건강 이익을 감소시킬 가능성을 가지고 있다. 이처럼 다양한 가치의 상충을 어떻게 판단하고 합의점을 찾을 것인가는 컨센서스 페이퍼 마련 및 인공지능 거버넌스 구축에서 중요한 문제이다.

상충되는 가치 간의 합의점을 마련하고 컨센서스를 구축하기 위해서 윤리 및 인공지능 전문가들의 의견을 청취하고 일반인의 여론을 수렴하고 숙의 과정을 통해 도출하는 것은 필수적이다. 그러나 전문가들의 의견을 파악하는 과정은 상대적으로 원활히 이루어질 수 있으나, 다양한 시민의 의견을 듣고 이를 반영하는 것은 상대적으로 어렵다. 하지만 의료는 시민의 건강에 직결되고, 사회적으로도 큰 관심을 받는 분야이며, 집단별로 중시하는 가치가 다를 수 있기 때문에 여론 수렴과 컨센서스 형성은 매우 중요하다.

혹자는 윤리 원칙의 바람직한 균형점을 찾기

28) 보건의료 현장에 인공지능이 도입이 됨에 따라 의료 인공지능을 기획하고 생성하며 적용하고 이를 관리하는 새로운 의료전문가가 등장할 수 있다. 이처럼 새로운 전문가 집단의 출현은 기존의 의사-환자 관계의 윤리로 포괄되지 못하는 윤리 문제를 야기할 수 있으며, 각각의 전문가들이 환자와 맺는 관계들에 관한 윤리 문제가 고민되어야 할 필요가 있다는 지적이 존재한다. 필자들은 기본적으로 인공지능의 등장으로 말미암아 보건의료 분야에서 환자에게 지켜져야 할 가치가 달라져야 한다고 보지는 않는다. 그러나 보건의료서비스 전달에 있어 영역별로 새로운 전문가군이 형성될 수 있음은 주지할 만한 부분이며, 각 전문가군이 어느 정도 수준으로 윤리성을 책임질 수 있는지, 이것이 전체적으로 관리될 수 있을지는 고려되어야 할 필요가 있다.

위해 일반인의 의견 청취를 포함하는 숙의 과정이 반드시 필요한지 의문점을 가질 수 있다. 일정한 수준 이상의 지식을 갖춘 학자나 전문가가 윤리 원칙을 다루기에 도리어 적절하다고 보는 것이다. 그러나 생명윤리 및 의료윤리 영역의 의사결정은 전문가뿐만 아니라 공론의 영역에서 다양한 가치와 관점을 지닌 이해당사자들을 포괄해야 함을 원칙으로 삼고 있으며, 숙의민주주의는 중요한 도구이자 방법론이 되어 왔다[50,51]. 그리고 그 원칙과 방법론에 관하여 공적 생명윤리(public bioethics)라는 주제로 다양하게 다루고 있다.[29] 공공 숙의(public deliberation)는 낙태나 안락사 등의 쟁점이 첨예한 생명윤리 이슈 외에 백신 정책, 의료자원 우선 적용 문제 등 윤리적 상충점이 존재하는 의료 및 공중보건 문제들로 적용이 확대되고 있다[54,55]. 또한 이들 문제에 관하여 공공 숙의 방법론이 객관적이고 합리적이며 유용하다고 평가되어 왔다[56]. 윤리적 가치의 상충 가능성이 크게 존재하는 인공지능 개발과 적용 또한 예외가 되기 어렵다. 공공 숙의 접근법은 어떠한 가치를 우선화할지, 어떻게 가치의 균형점을 찾으며 인공지능을 개발할 것인지를 결정함에 있어 유용할 수 있다. 누필드 재단에서도 인공지능 영역의 윤리 사회적 함의 연구 로드맵을 제안하면서 1) 인공지능 테크놀로지가 상호 다른 가치를 지지하거나 위협할 수 있는 윤리적 긴장을 파악하고 해결할 것, 그리고 2) 윤리 사회적 이슈 토론을 위해 보다 확고한 근거를 구축할 것을 제안하며 이를 위해 기존의 공적 숙의 기구를 활용할 것을 덧붙이고 있다[57].

보건의료 영역 인공지능의 윤리 원칙 결정에 관하여 전문가들의 의견 만으로 윤리 원칙을 적용하기 어려움을 예상하고 이를 극복하기 위한 한 예로 2019년 영국 맨체스터 대학 연구팀의 시민 배심원(citizens' juries) 연구를 들 수 있다[50]. 연구팀은 시민 배심원을 구성하여 일반 시민들이 보건의료 인공지능의 설명가능성과 정확성이 상충되는 상황에서 어느 쪽을 더 중요시하는지에 대한 평가를 실시하였다 이를 위해 18명의 다양한 배경을 가진 시민 배심원으로 구성된 두 팀을 선정하고 이들 간의 논의와 투표 과정을 관찰하였다. 특히 이 연구에서는 보건의료에서의 인공지능 활용과 다른 맥락에서의 활용을 구분하였고, 각 상황에서 설명가능성이 강화된 인공지능, 두 가치를 함께 고려한 인공지능, 그리고 정확성이 강화된 인공지능 세 가지 중 선호하는 시스템을 선택하도록 하였다. 이 과정에서 보건의료의 맥락과 그렇지 않은 맥락에서 선호하는 가치에 대한 차이가 존재하는 지의 여부도 확인하고자 하였다.

위 조사의 결과는 보건의료의 상황의 경우 시민들이 설명가능성보다는 정확성의 가치를 중요시 여기는 것이 드러났다. 이는 일반적인 상황의 경우 두 가치를 비슷하게 평가하거나, 혹은 설명가능성을 중요시하는 것과는 차이가 있었다. 특히, 이 연구에서는 전문가들이 예상했던 것에 비해 시민들이 설명가능성의 가치에 비해 정확성의 가치를 높게 평가하고 있음이 지적되었다. 이는 전문가들을 통한 의견 수렴만으로는 보건의료 인공지능에 대한 대중의 평가를 파악하기 어려움을 시사하고 있다. 따라서 이러한 결과는 한국에서도 유사한 방식의 과정을 통해 심층적으로 시민들의 의견을 파악할 필요성을 드러낸다.

인공지능이 상호 다른 가치로 빠르게 개발될

29) 공적 생명윤리(public bioethics)는 생명윤리심의위원회 등 공공 숙의 기구를 통한 생명윤리 영역 의사 결정을 지칭한다. 그 원칙과 적용에 관해서는 Childress의 글, Kelly의 글 등을 참조할 것[52,53].

가능성이 존재하는 현 상태에서 공동의 가치를 사회적으로 합의하고 어떤 가치를 우선화할 것인지 거버넌스를 구축하는 것은 매우 필수적이다. 보건의료 영역에서는 의견 수렴과 함께 공론 조사, 숙의 토론 등을 바탕으로 한 숙의 메커니즘이 도입될 필요가 있다. 그리고 그 결정의 근거를 제공할 수 있을 만한 기술의 실현가능성, 효과, 대중 참여 방법론 등에 대한 광범위한 연구가 필요하다. 이는 인공지능을 개발하는 개발자 뿐 아니라 보건의료영역의 다양한 당사자, 관련 정책들을 결정하는 정책결정권자들 모두에게 그 필요성에 대한 공감대가 확대되어야 할 것이다.

III. 결론

11개의 국외 보건의료 인공지능 윤리 가이드라인의 분석을 통해 다수 문헌에서 언급한 의료윤리영역 인공지능 활용에서의 윤리적 영역, 키워드, 가치, 원칙(ethical domain, keyword, value, principle) 등을 선별할 수 있었다. 이를 바탕으로 데이터 수집, 임상 환경, 사회 환경 등 3가지 영역에 대응하는 인간에 대한 존중, 책무성, 지속가능성의 테마를 도출할 수 있었다. 마지막으로 각 영역과 테마에 해당하는 핵심 윤리 키워드들을 제시하고 각 키워드에 대한 자세한 설명을 하고자 하였다.

현 시점에서 보건의료영역 인공지능의 활용에 있어 여러 이견들이 해소가 되지 않고 있다. 딥러닝 인공지능 개발에 빅데이터를 활용하는 데 있어 사생활 보호의 원칙을 얼마나 완화할 수 있는 가의 문제, 그리고 책무성의 공백의 문제 등에서 확인할 수 있듯이 여러 윤리적 원칙들이 상충(trade-off) 관계에 있음을 확인할 수 있었다. 이러한 이견들은 윤리적 차원 뿐 아니라 법적, 기술적 차원에서 또한 논의가 되어야 하는 문제이다.

모든 이해 관계자들의 숙의 과정을 통한 컨센서스는 인공지능의 윤리적 활용을 위한 필수적인 조건이다. 법적, 기술적, 윤리적 컨센서스 없는 의료 인공지능의 광범위한 확산은 환자의 건강을 위협하는 등 여러 차원의 사회 문제를 유발할 수 있다. 또한 법적, 윤리적 불확실성으로 인해 인공지능의 개발 및 활용을 저해함에 따라 보건의료 인공지능의 도입으로 얻을 수 있는 여러 혜택을 누리지 못할 가능성이 있다.

이번 연구에서는 문헌들이 제시한 윤리적 개념에 대한 이슈, 원칙, 그리고 적용에 대한 구체적인 설명을 하는 데 집중하였다. 현재 국내에서는 보건윤리영역 인공지능의 원칙을 다루는 컨센서스 페이퍼가 부재한 상황이다. 주요 국외 보건윤리 인공지능 가이드라인에 제시된 개념을 설명하고 향후 공론 조사 등을 통한 컨센서스 형성을 제안한 이번 연구가 향후 사회적 합의 도출에 도움이 되기를 바란다.

Conflict of Interest

There are no potential conflicts of interest relevant to this article.

REFERENCES

[1] OECD. Recommendation of the council on artificial intelligence [Internet]. Paris: OECD; 2019 [cited 2023 Apr 10]. Available from: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

[2] OECD. Trustworthy AI in health [Internet]. Paris: OECD; 2020 [cited 2023 Apr 10]. Available from: https://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf

[3] FDA. Artificial intelligence and machine learning (AI/ML)-enabled medical devices [Internet]. Silver Spring: FDA; 2022 [cited 2023 Apr 10]. Available from: https://www.fda.gov/

medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

[4] Ministry of Food and Drug Safety. Achievements and future plans for AI medical devices in 2022 [Internet]. Cheongju: Ministry of Food and Drug Safety; 2022 [cited 2023 Apr 10]. Available from: https://www.mfds.go.kr/brd/m_220/down.do?brd_id=data 0014&seq=32871&data_tp=A&file_seq=3

[5] Korea Policy Center for the Fourth Industrial Revolution, Lloyd's Register Foundation Institute for Public Understanding of Risk, Sense about Science. Using artificial intelligence to support healthcare decisions- a guide for society [Internet]. Daejeon: KAIST; 2021 [cited 2023 Apr 10]. Available from: https://kpc4ir.kaist.ac.kr/index.php?document_srl=3402&mid=KPC4IR_Reports

[6] National Human Rights Commission of the Republic of Korea. Human rights guidelines for the development and use of artificial intelligence [Internet]. Seoul: National Human Rights Commission of The Republic of Korea; 2022 [cited 2023 Apr 10]. Available from: https://www.humanrights.go.kr/site/program/board/basicboard/view?boardtypeid=24&boardid=7607961&menuid=001004002001

[7] Future Advocacy. Ethical, social, and political challenges of artificial intelligence in health [Internet]. London: Future Advocacy; 2018 [cited 2023 Apr 10]. Available from: https://futureadvocacy.com/publications/ethical-social-and-political-challenges-of-artificial-intelligence-in-health/

[8] Nuffield Council on Bioethics. Artificial intelligence (AI) in healthcare and research [Internet]. London: Nuffied Council on Bioethics; 2018 [cited 2023 Apr 10]. Available from: https://www.nuffieldbioethics.org/publications/ai-in-healthcare-and-research

[9] Academy of Medical Royal Colleges. Artificial intelligence in healthcare [Internet]. London: Academy of Medical Royal Colleges; 2019 [cited 2023 Apr 10]. https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf

[10] European and North American Multisociety. Ethics of AI in radiology- European and North American Multisociety statement [Internet]. Reston: American College of Radiology; 2019 [cited 2023 Apr 10]. Available from: https://www.acr.org/-/media/ACR/Files/Informatics/Ethics-of-AI-in-Radiology-European-and-North-American-Multisociety-Statement--6-13-2019.pdf

[11] Ministry of Health of Singapore, Health Sciences Authority, IHIS. Artificial intelligence in healthcare guidelines [AIHGle] [Internet]. Singapore: Ministry of Health; 2021 [cited 2023 Apr 10]. Available from: https://www.moh.gov.sg/docs/librariesprovider5/eguides/1-0-artificial-in-healthcare-guidelines-(aihgle)_publishedoct21.pdf

[12] World Health Organization [WHO]. Ethics and governance of artificial intelligence for health [Internet]. Geneva: WHO; 2021 [cited 2023 Apr 10]. Available from: https://www.who.int/publications/i/item/9789240029200

[13] UNESCO. Recommendation on the ethics of artificial intelligence [Internet]. Paris: UNESCO; 2021 [cited 2023 Apr 10]. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000380455

[14] Mittelstadt B. The impact of artificial intelligence on the doctor-patient relation-ship [Internet]. Strasbourg: Council of Europe; 2021 [cited 2023 Apr 10]. Available from: https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859

[15] European Parliament Research Service. Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts [Internet]. Strasbourg: Think Tank European Parliament; 2022 [cited 2023 Apr 10]. Available

from: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512

[16] Fjeld J, Achten N, Hilligoss H, et al. Artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI [Internet]. Cambridge: Berkman Klein Center; 2020 [cited 2023 Apr 10]. Available from: https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf?sequence=1&isAllowed=y

[17] UN Human Rights Council. The right to privacy in the digital age : resolution/adopted by the human rights council on 26 September 2019. New York: UN; 2019 [cited 2023 Apr 10]. Available from: https://digitallibrary.un.org/record/3837297

[18] Forti M. The deployment of artificial intelligence tools in the health sector: privacy concerns and regulatory answers within the regulation (EU) 2016/679. Eur J Leg Stud 2021;13(1):29-44. https://doi.org/10.2924/EJLS.2019.040

[19] Ministry of Culture, Sports and Tourism. Data rule of thirds [Internet]. Sejong: Ministry of Culture, Sports and Tourism; 2021 [cited 2023 Apr 10]. Available from: https://www.korea.kr/special/policyCurationView.do?newsId=148867915

[20] Ministry of Health and Welfare. Health data utilization guidelines [Internet]. Sejong: Ministry of Health and Welfare; 2022 [cited 2023 Apr 10]. Available from: https://www.mohw.go.kr/react/al/sal0101vw.jsp?PAR_MENU_ID=04&MENU_ID=040101&page=1&CONT_SEQ=374313

[21] National Human Rights Commission of The Republic of Korea. Statement by the national human rights commissioner on the enactment of the "3 Data Acts" by the national assembly, including the personal information protection act [Internet]. Seoul: National Human Rights Commission of The Republic of Korea; 2020 [cited 2023 Apr 10]. Available from: https://www.humanrights.go.kr/site/program/board/basicboard/view?menuid=001004002001&pagesize=10&boardtypeid=24&boardid=7604976

[22] Select Committee on Artificial Intelligence. AI in the UK: ready, willing and able? [Internet]. London: House of Lords; 2018 [cited 2023 Apr 10]. Available from: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[23] Ghassemi M. Exploring healthy models in machine learning for health [Internet]. Toronto: University of Toronto; 2021 [cited 2023 Apr 10]. Available from: https://youtu.be/5uZROGFYfcA

[24] Arshad Ahmed M, Chatterjee M, Dadure P, et al. The role of biased data in computerized gender discrimination. In: 2022 IEEE/ACM 3rd International Workshop on Gender Equality, Diversity and Inclusion in Software Engineering (GEICSE). Pittsburgh; 2022. pp.6–11.

[25] Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018;154(11):1247-1248. https://doi.org/10.1001/jamadermatol.2018.2348

[26] Kamulegeya LH, Okello M, Bwanika JM, et al. Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. Bioinformatics 2019. https://doi.org/10.1101/826057

[27] Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. J Public Health Policy 2021;42(4):602-611. https://doi.org/10.1057/s41271-021-00319-5

[28] Hart RD. If you're not a white male, artificial intelligence's use in healthcare could be dangerous [Internet]. New York: Quartz; 2017 [cited 2023 Apr 10]. Available from: https://

qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous

[29] Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. Front Artif Intell 2021;3:561802. https://doi.org/10.3389/frai.2020.561802

[30] Kiener M. Artificial intelligence in medicine and the disclosure of risks. AI Soc 2021;36(3):705–713. https://doi.org/10.1007/s00146-020-01085-w

[31] Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? Geo Law J. 2019;108:1425.

[32] High-Level Expert Group on AI. Ethics guidelines for trusworthy AI [Internet]. Brussel: European Commission; 2019 [cited 2023 Apr 10]. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[33] BSI, AAMI. Machine learning AI in medical devices [Internet]. Arlington: BSI; 2020 [cited 2023 Apr 10]. https://www.medical-device-regulation.eu/wp-content/uploads/2020/09/machine_learning_ai_in_medical_devices.pdf

[34] Ministry of Food and Drug Safety [MFDS]. Medical Device Cyber Seccurity Application Methods and Casebook (User guidance) [Internet]. Cheongju: Ministry of Food and Drug Safety; 2022. Available from: https://www.mfds.go.kr/brd/m_1060/view.do?seq=15120&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=1

[35] Caruana R, Lou Y, Gehrke J, et al. Intelligible models for HealthCare: predicting pneumonia risk and hospital 30-day readmission. In: 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. online conference; 2015. pp.1721-1730.

[36] Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310. https://doi.org/10.1186/s12911-020-01332-6

[37] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[38] Nyrup R, Robinson D. Explanatory pragmatism: a context-sensitive framework for explainable medical AI. Ethics Inf Technol 2022;24(1):13. https://doi.org/10.1007/s10676-022-09632-3

[39] Hamon R, Junklewitz H, Sanchez I, et al. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. IEEE Comput Intell Mag 2022;17(1):72-85. https://doi.org/10.1109/MCI.2021.3129960

[40] G20 Ministerial Meeting. G20 AI principles [Internet]. Tsukuba: G20 Ministerial Meeting; 2019 [cited 2023 Apr 10]. Available from: https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf

[41] Buruk B, Ekmekci PE, Arda B. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. Med Health Care Philos. 2020;23(3):387-399. https://doi.org/10.1007/s11019-020-09948-1

[42] Nicholson Price W 2nd, Gerke S, Cohen G. Potential liability for physicians using artificial intelligence. J Am Med Assoc 2019;322(18):1765-1766. https://doi.org/10.1001/jama.2019.15064

[43] Banja JD, Hollstein RD, Bruno MA. When artificial intelligence models surpass physician performance: medical malpractice liability

in an era of advanced artificial intelligence. J Am Coll Radiol 2022;19(7):816-820. https://doi.org/10.1016/j.jacr.2021.11.014

[44] van der Waa J, Verdult S, van den Bosch K, et al. Moral decision making in human-agent teams: human control and the role of explanations. Front Robot AI 2021;8:640-647. https://doi.org/10.3389/frobt.2021.640647

[45] United Nations. Sustainability [Internet]. New York: United Nations. Available from: https://www.un.org/en/academic-impact/sustainability

[46] National Health Service [NHS]. Artificial intelligence: how to get it right [Internet]. London: NHS; 2019 [cited 2023 Apr 10]. Available from: https://transform.england.nhs.uk/media/documents/NHSX_AI_report.pdf

[47] Alami H, Rivard L, Lehoux P, et al. Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. Glob Health 2020;16(1):52. https://doi.org/10.1186/s12992-020-00584-1

[48] Richie C. Environmentally sustainable development and use of artificial intelligence in health care. Bioethics 2022;36(5):547–555. https://doi.org/10.1111/bioe.13018

[49] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bull World Health Organ 2020;98(4):257-262. https://doi.org/10.2471/BLT.19.237289

[50] van der Veer SN, Riste L, Cheraghi-Sohi S, et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. J Am Med Inform Assoc 2021;28(10):2128-2138. https://doi.org/10.1093/jamia/ocab127

[51] Lee IH, From regulation to communication: a proposal for transformation of national bioethics committee. Bio, Ethic Policy 2017;1(2):1-18.

[52] Childress JF. Public bioethics: principles and problems. New York: Oxford University Press; 2020.

[53] Susan EK. Public bioethics and publics: consensus, boundaries, and participation in biomedical science policy. Sci Technol Hum Values 2003;28(3):339-364. https://doi.org/10.1177/0162243903028003001

[54] O'Doherty KC, Crann S, Bucci LM, et al. Deliberation on childhood vaccination in Canada: public input on ethical trade-offs in vaccination policy. AJOB Empir Bioeth 2021;12(4):253-265. https://doi.org/10.1080/23294515.2021.1941416

[55] Schindler M, Danis M, Goold SD, et al. Solidarity and cost management: Swiss citizens' reasons for priorities regarding health insurance coverage. Health Expect 2018;21(5):858-869. https://doi.org/10.1111/hex.12680

[56] Manafò E, Petermann L, Vandall-Walker V, et al. Patient and public engagement in priority setting: a systematic rapid review of the literature. PLOS ONE 2018;13(3):e0193579. https://doi.org/10. 1371/journal.pone.0193579

[57] Whittlestone J, Nyrup R, Alexandrova A, et al. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation; 2019.

한국의료윤리학회지 제26권 제2호(통권 제75호): 103-131 ©한국의료윤리학회, 2023년 6월
Korean J Med Ethics 26(2): 103-131 ©The Korean Society for Medical Ethics, June 2023
pISSN 2005-8284 eISSN 2234-3598

Ethical Principles and Considerations concerning the Use of Artificial Intelligence in Healthcare*

MOON Gieop[1], YANG Ji Hyun[2], SON Yumi[3], CHOI Eun Kyung[4], LEE Ilhak[5]

Abstract

The use of artificial intelligence (AI) in healthcare settings has become increasingly common. Many hope that AI will remove constraints on human and material resources and bring innovations in diagnosis and treatment. However, the deep learning techniques and resulting black box problem of AI raise important ethical concerns. To address these concerns, this article explores some of the relevant ethical domains, issues, and themes in this area and proposes principles to guide use of AI in healthcare. Three ethical themes are identified, including respect for person, accountability, and sustainability, which correspond to the three domains of data acquisition, clinical setting, and social environment. These themes and domains were schematized with detailed explanations of relevant ethical issues, concepts, and applications, such as explainability and accountability. Additionally, it is argued that conflicts between ethical principles should be resolved through deliberative democratic methods and a consensus building process.

Keywords

artificial intelligence; guideline; ethical principle; healthcare; explainability; accountability

* This study was supported by the Ministry of Health and Welfare in 2022 (A study on healthcare Artificial Intelligence Ethics Guidelines).

1 Graduate Student, Department of History of Medicine and Medical Humanities, Seoul National University College of Medicine.

2 Research Fellow, Division of Medical Law and Ethics, Department of Medical Humanities and Social Sciences, Yonsei University College of Medicine.

3 Ph.D. Candidate, Doctoral Program in Medical Law and Ethics, Yonsei University; Asian Institute for Bioethics and Health Law, Yonsei University.

4 Assistant Professor, Department of Medical Humanities and Medical Education, Kyungpook National University School of Medicine: ***Corresponding Author***

5 Associate Professor, Division of Medical Law and Ethics, Department of Medical Humanities and Social Sciences, Yonsei University College of Medicine; Asian Institute for Bioethics and Health Law, Yonsei University; Institute for Innovation in Digital Health Care, Yonsei University: ***Corresponding Author***