생명윤리 제24권 제2호(통권 제48호): 87-105, 2023년 12월
J Korean Bioethics Assoc 24(2): 87-105, December 2023

ISSN 1598-0324
10.37305/JKBA.2023.12.24.2.87

헬스케어 인공지능에서 악을 식별할 수 있을까 : 플로리디 정보 윤리학에 기초하여*

김 준 혁**

【국문초록】 인공지능의 도덕성에는 여전히 논쟁의 여지가 있다. 특히, 아직 인공지능이 자기 인식이나 의도를 가지고 있다고 말하기 어렵다는 점을 고려할 때 문제는 어려워진다. 그러나, 인공지능 개발에 도덕적 추론을 통합할 필요성에 관한 공감은 이미 형성되어 있다. 이때, 인공지능 자체의 악을 말할 수 있는지 여부가 중요하나 관련 논의는 대부분 이론적 수준에 머물러 있고, 실질적 논의는 인공지능 개발 및 사용의 윤리를 중심으로 이루어지고 있다. 그러나, 헬스케어 인공지능 영역은 생명의료윤리의 논의가 현재 가용되는 인공지능 알고리듬이나 장치와 만나야 하기 때문에 상황이 다르다. 생명의료윤리의 원칙 아래, 헬스케어 인공지능이 환자에게 해를 끼치거나 환자의 이익에 반하는 행동을 할 수 있는 상황을 평가하기 위해서는 명확한 기준이 필요하다. 주목할 만한 문제로 그 사용자가 오정보를 받을 수 있는 생성형 인공지능의 환각이 있다. 이것이 반드시 사용자에게 해를 끼치지는 않을 수도 있지만, 헬스케어 분야에서 그 영향은 검토를 필요로 한다. 본 논문은 루치아노 플로리디의 정보 윤리학에 기초하여 헬스케어의 맥락적 특수성 안에서 헬스케어 인공지능의 환각을 악으로 정의한다. 본 논문은 헬스케어 인공지능의 환각 수준을 측정하고 줄이기 위해 평가 지표를 사용할 것을 주장한다. 정보 윤리학을 개괄한 뒤, 이를 헬스케어 분야에 적용하여 본 논문은 헬스케어에 맞는 참-거짓 평가 기준의 필요성을 요청한다.

【색인어】 인공지능, 생성형 인공지능, 헬스케어 인공지능, 환각, 루치아노 플로리디, 정보 윤리학

* 본 연구는 질병관리청 학술연구개발용역과제 '헬스케어 · 인공지능 연구를 위한 연구윤리 교육 프로그램 운영 및 윤리지침 개선' 지원에 의하여 이루어진 것임(과제번호: 2023-ER0808-00).
** 연세대학교 치과대학 치의학교육학교실 조교수

1. 서론

인공지능의 선악을 묻는 것은 누군가에겐 우문일 것이다. 인공지능이 아직 의지 또는 자기 인식이 있다고 말하기 어렵다면 그의 선택이 그릇된 준칙을 의지했다고도, 타인의 해악을 의도했다고도 말하기 어려우므로, 그 선악을 따지는 것은 섣부른 판단이라는 것이다. 이 경우, 인공지능의 선악을 따짐은 공상의 영역으로 남는다. 또는, 인공지능이 발전하더라도 결국 책임은 사회가 사회적 존재에게 귀속시키는 것이므로, 초점은 사회에 있지 인공지능에 있지 않다는 결론이 도출될 수도 있다.[1)]

물론, 인공지능을 개발하는 과정에서 공리 또는 이성 개념을 사고하여 인공지능이 원칙을 따를 수 있도록 할 필요가 있다는 주장도 제기되어 왔다.[2)] 이런 인공적 도덕 행위자(Artificial Moral Agents) 설계에 관한 논의는 다양한 논점으로 진행되어 왔다(이를테면, 필요성[3)], 교육학[4)], 속성[5)], 수준[6)], 한계[7)], 윤리 원칙[8)] 등). 전술한 인공지능의 도덕성 또는 '도덕성' 자체의 모호성을 들어 인공적 도덕 행위자의 필요성 주장 자체에 문제가 있다는 비판도 있으나,[9)] 다수의 논의는 인공지능의 발전 과정에서 도덕적 사고 능력을 포함시키는 것이 꼭 필요하다는 쪽으로 모이고 있다. 이때, 인공지능 또는 인공적 도덕 행위자의 악은 사전에 개발자가 설정한 원칙 또는 경로를 따르지 않는 것으로 이해될 것이며, 이런 논의의 출발점 격인 아이작 아시모프(Isaac Asimov)의 "로봇 3원칙"이 원칙 간의 충돌이나 모순으로 인하여 그를 위반하는 로봇(즉, 인공적 도덕 행위자)에 관한 성찰을 보여주는 소설 작품을 여럿 남겼다는 점에서 인공적 도덕 행위자의 개발이 악의 가능성을 배제하는 것은 아니다.

선악을 판단할 수 있는 능력을 지닌 인공지능을 상정하고 그의 행동 방침을 (또는, 전술전략을) 예상하여 인공지능의 선악을 말하려는 이도 있다.[10)] 이런 인공지능, 즉 강인공지능(Strong Artificial Intelligence)은 (인간과 비슷한 속도의 정보 처리 능력과 지식량으로는 인공지능이 선악 판단을 내리는 수준에 도달하는 것이 불가능하므로) 인간의 지능과 지식을 뛰어넘는 사고력과 판단력을 지니고 있을 것이며, 그는 자신의 독점적 지위를 유지하기 위해 최종 목표에 따라 도구적 목적을 정렬하여 수행할 가능성이 있으며, 이때 중요한 도구적 목적인 자원 확보에 방해가 되는 인간종에게 불리한 방식으로 행동할 것으로 보인다.[11)] 이것이 강인공지능이 "악한" 경로를 선택한 것이라고 말하기는 어려울 수 있으나, 자신의 번영을 위해 인류의 존속을 위협한다는 점에서 이는 전통적인 악의 범주에

1) 이상형, "윤리적 인공지능은 가능한가? -인공지능의 도덕적, 법적 책임 문제-", 『法과 政策硏究』, 제16권 제4호, 2016, 283-303면.
2) Gabriel I., "Artificial intelligence, values, and alignment", Minds and Machines, Vol. 30, 2020, pp. 411-437.
3) Formosa P., Ryan M., "Making moral machines: Why we need artificial moral agents", AI & Society, Vol. 36, 2020, pp. 839-851.
4) Allen C., Varner G., Zinser J., "Prolegomena to any future artificial moral agent", Journal of Experimental & Theoretical Artificial Intelligence, Vol. 12. No. 3, 2010, pp. 251-261.
5) Himma K. E., "Artificial agency, consciousness, and the criteria for moral agency: What properties must and artificial agent have to be a moral agent?", Ethics and Information Technology, Vol. 11, 2009, pp. 19-29.
6) 목광수, "인공적 도덕 행위자 설계를 위한 고려사항: 목적, 규범, 행위지침", 『철학사상』, 제69집, 2018, 361-391면.
7) 이향연, "인공 도덕행위자(AMA)가 지닌 윤리적 한계", 『대동철학』, 제95집, 2021, 103-118면.
8) 최현철, 변순용, 신현주, "인공적 도덕행위자(AMA) 개발을 위한 윤리적 원칙 개발 -하향식 접근(공리주의와 의무론)을 중심으로-", 『倫理硏究』, 제1권 제111호, 2016, 31-53면.
9) van Wynsberghe A., Robbins S., "Critiquing the reasons for making Artificial Moral Agents", Science and Engineering Ethics, Vol. 25, 2019, pp. 719-735.
10) 닉 보스트롬, 조성진 역, 『슈퍼 인텔리전스: 경로, 위험, 전략』, 까치, 2017.

포함시킬 수 있다.

인공지능 자체의 악을 고찰하는 것은 흥미롭지만, 아직은 우리가 인공적 도덕 행위자 또는 강인공지능의 출현을 목도하지 못했다는 점에서 다분히 이론적 고찰의 수준에 머물고 있는 것처럼 보인다. 그렇다면, 우리는 실천 차원에서 인공지능 자체의 윤리를 말하기보다는 인공지능 개발이나 사용의 윤리에 초점을 맞추는 편이 나을 것이다.

그러나, 헬스케어 인공지능 영역에서 상황은 사뭇 다르다. 이미 여러 인공지능 알고리듬이나 기기가 상용화되었고 식품의약품안전처 인증 또는 승인을 받아 진단이나 치료 상황에 사용되고 있다는 현실 상황에 대한 진단을 차치하더라도, 헬스케어 인공지능 개발 및 활용에 있어서 기술 자체의 문제점을 검토할 수 있는 척도가 필요한 상황이기 때문이다. 생명의료윤리의 원칙들을 적용한다고 해도, 헬스케어 인공지능이 환자의 이득에 반한다거나 해악을 끼친다고 말하려면 그것이 어떤 상황에서 그러한지를 규명할 필요가 있다.

물론, 신체적, 정신적 상해를 가하는 것을 그 기준으로 삼는 것으로 충분하다고 말할 수도 있다. 그러나, 이 경우 최근 생성형 인공지능이 초래한 문제, 즉 환각(hallucination)을 검토하는 데에 난점이 생긴다. 환각이란 이미지 또는 텍스트 생성 인공지능이 "기존에 없는" 이미지나 텍스트를 생성할 수 있기에 사용자에게 "거짓"[12] 정보를 제공하는 것을 말한다. 그러나, 환각 자체로는 사용자에게 상해를 초래하지 않을 수 있다. 심지어, 그것이 사용자의 자율성을 직접적으로 침해한다고 보기에는 어려울 수 있다. 아직 생성형 인공지능이 사용자를 대신하여 의사결정을 내리는 것은 아니기 때문이다. 그렇다면, 생성형 인공지능과 같은 인공지능 알고리듬 또는 형식을 헬스케어 인공지능에서 활용해도 되는가?

본 논문은 루치아노 플로리디(Luciano Floridi)의 정보 윤리학(information ethics)에 기초하여 환각을 헬스케어 인공지능의 악으로 정의할 수 있음을 제시하고자 한다. 이것은 인공지능의 생성 과정 자체가 악이라는 주장을 함의하지 않으며, 헬스케어 인공지능이라는 맥락적 특수성에서 제기된다. 이에 기초하여, 본 논문은 헬스케어 인공지능의 경우 환각 현상을 일정 수준 이하로 낮추었음을 보이는 평가 기준(evaluation metrics)을 통과할 때에만 헬스케어 영역에서 사용 가능하다고 주장하고자 한다.

따라서, 본 논문은 먼저 정보 윤리학의 주장을 개략적으로 살핀 다음 이를 헬스케어 영역에 적용하기 위한 비판적 검토 과정을 거친다. 다음, 여기에서 수립한 주장을 환각을 포함하여 인공지능이 정보를 제공하는 상황 일반에 적용하여 그 선악을 판단할 수 있음을 보인다. 여기에서 도출한 결론을 바탕으로, 논문은 현재 생성형 인공지능에서 사용되고 있는 참거짓 평가 기준을 일별하고 헬스케어 영역에 특화된 참거짓 평가 기준이 필요함을 제시한다.

---

11) 보스트롬은 세 가지 논제에 근거하여 최초의 강인공지능이 인간에게 불리하게 행동할 것이라는 예측을 내놓는다. 첫째, 확실한 전략적 이점(decisive strategic advantage)으로, 먼저 강인공지능에 도달하는 기획은 뒤따라오는 다른 인공지능과의 격차를 유지하는 방식으로 활동할 것이다. 둘째, 직교성 명제(orthogonality thesis)로, 인공지능 행위자의 최종 목적과 그 지능 수준은 별도 사항이다. 즉, 강인공지능은 어떤 최종 목적을 위해서도 존재할 수 있다. 셋째, 도구적 수렴성 명제(instrumental convergence thesis)로, 강인공지능이 어떤 최종 목적을 설정하든 그를 위한 도구적 합리성을 추구할 것이며, 이때 가장 그럴듯한 수행 목적으로 자원 확보를 내세울 것이다. 이 세 논제에 근거하여, 보스트롬은 최초의 강인공지능이 다른 기획과의 격차 유지를 위하여 자신의 최종 목적을 추구함에 있어 자원 확보를 강조할 것이고, 이에 반하는 인간종에 부정적인 선택을 할 가능성이 높다고 주장한다. 닉 보스트롬, 앞의 책, 213-215면.

12) 생성형 인공지능은 참과 거짓을 구분하는 능력이 없어 참이 아닌 정보(untrue information)을 제공할 뿐이다. 사용자를 속이거나 혼란을 주려는 의도가 있는 것은 아니므로 그것을 거짓말(lie)이라고 할 수는 없다.

2. 본론

2.1. 플로리디 정보 윤리학

사이버네틱스 이론의 창시자 노버트 위너(Norbert Wiener)는 효율적으로 사는 것은 충분한 정보를 가지고 사는 것이라고 말한 바 있다.[13] 여기에서 정보란 의미를 지닌 데이터로 정의되며,[14] 여기에서 문제가 되는 것은 의미론(semantics)적 차원임이 분명해지며, 정보를 다룰 때 정보와 그 수신자의 관계를 검토해야 할 필요가 발생한다.

즉, 정보와 관련하여 규범적으로 검토되어야 할 것은 행위자가 정보와 어떻게 상호작용하는가에 있다. 정보철학에 입각하여[15] 플로리디는 정보 행위자와 정보의 상호작용을 자원(Resource), 산물(Product), 대상(Target)으로써의 정보를 다루는 것으로 분류한 RPT 모형을 제시한다.[16] 즉, 정보의 윤리는 행위자의 정보 자원의 획득, 정보의 생성, 정보에 미치는 영향을 다루어야 하며, 이때 핵심이 되는 것은 정보적 차원에서 정보의 흐름 또는 정보 행위자가 정보와 맺는 관계에 있다. 정보 취득, 생성, 평가의 차원에서 윤리적 문제를 검토해 보자.

행위자가 적절히 행위하기 위해서 정보 자원은 매우 중요한 위치에 놓인다. 무지나 거짓이 만들어 낸 수많은 악을 상기할 때, 우리는 도덕적 행위에 있어 적절한 정보의 중요성을 쉽게 떠올릴 수 있다. 이때, 가용성, 접근성, 정확성이 문제가 된다.[17]

정보를 취한 행위자는 또한 정보를 생산한다. 정보 생성 과정에서 행위자의 윤리성이 문제가 되고, 여기에서 책무책임(accountability)[18], 배상책임(liability), 명예훼손 법제(libel legislation), 증언, 표절, 광고, 선전(propaganda), 오정보(misinformation)[19], 역정보(disinformation)[20], 기만 등이 다루어져야 한다.[21]

또한, 정보의 취득과 생성에 있어 행위자가 정보를 어떻게 평가하는지가 문제가 된다. 예컨대, 해킹에서 문제가 되는 것은 정보 침해 행위 자체라기보다, 정보 주체(data subject)가 정보 접근을 허용하였는지의 여부이다. 즉, 정보 환경(informational environment)에 대한 가치 판단이나 결정이 문제가 되며, 해킹, 보안, 공공기물 파손, 저작권 침해, 오픈소스 소프트웨어(open source software), 표현의 자유, 검열, 필터링(filtering), 내용 통제(contents control) 등이 이 차원의 윤리적 사안에 속한다.[22]

이런 정보 행위자와 정보가 맺는 다양한 관계 양태

---

13) Wiener N., The Human Use of Human Beings: Cybernetics and Society. Boston: Houghton Mifflin, 1954, pp. 15-27.

14) 플로리디는 정보의 일반 정의(general definition of information)를 다음과 같이 제시한다. "**σ**는 다음과 같은 경우 그리고 오로지 그 경우에만 의미론적 내용으로서 이해되는 정보의 한 [인스턴스]이다. ① **σ**는 n개의 데이터로 구성된다[n≥1], ② [이런] 데이터는 **잘 형성된** 것이다, ③ 그 잘 형성된 데이터는 **유의미하다.**" 루치아노 플로리디, 석기용 역, 『정보철학입문』, 필로소픽, 2022, 45면. 단, 역서는 instance를 "예화"로 옮겼으나, 정보 이론에서 instance는 예화 또는 예시가 아니라 특정한 모형 또는 클래스에 속하는 객체를 가리키며 이를 따로 옮기지 않고 인스턴스로 음차해 왔다.

15) 아래 플로리디의 논의는 그의 정보철학적 관점을 전제해야 하나, 본 논문의 초점이 정보철학이 아닌 정보 윤리학에 있으므로 그에 대한 설명은 생략하였다. 정보철학에 관한 개괄은 다음 책을 참조하라. 루치아노 플로리디, 앞의 책.

16) Floridi L., Ethics of Information, Oxford: Oxford University Press, 2013, pp. 20-21.

17) Floridi L.(2013), Ibid, pp. 21-22.

18) accountability의 역어로 "책무책임"을 선택하는 것에 관해선 필자의 이전 논문을 참조하라. 김준혁, 강철, "코로나19와 구조적 부정의: 아이리스 영의 사회적 연결 모델과 팬데믹 해결을 위한 책임", 『생명윤리』, 제23권 제1호, 2022, 57면.

19) 의도성 없이 제시되는 의미론적 거짓을 가리킨다.

20) 의도성을 가지고 제시되는 의미론적 거짓을 가리킨다.

21) Floridi L.(2013), Ibid, pp. 23-24.

가 정보 순환(즉, 앞서 말한 취득, 생성, 평가)의 문제에서 다루어지고 있으므로, 여기에서 중요하게 다루어져야 하는 것은 정보들의 관계가 만들어 내는 환경임을 알 수 있다. 다시 말하면, 위 분류는 플로리디가 정보 환경, 즉 정보권(infosphere) 개념에 기초하여 정보의 윤리를 논의하고 있음을 잘 보여준다. 정보권은 생물과 그 생존 환경을 포괄하는 생물권(biosphere) 개념의 정보적 대응물이며, 따라서 그의 정보 윤리학은 환경 윤리의 일환으로 이해된다.[23] 여기에서 그의 논의는 인간 행위자의 정보 관련 행위만을 다루는 것을 넘어 정보계에 포함된 모든 정보 행위자(information agent)에게 적용되는 거시 윤리로 확장된다.[24] 이때, 정보계 속 정보 행위자를 구성하기 위하여 플로리디는 추상화 층위 방법론(methods of levels of abstraction)을 적용한다.[25] 추상화 층위 방법론이란 특정 추상화 층위(level of abstraction)에서 행위자의 식별 대상(observable)을 추출하는 것을 말한다. 예컨대, 행위자 A가 행위자 B와 플랫폼을 통해 중고거래를 약속했다고 해 보자. B가 플랫폼에 올려놓은 사진과 설명을 통해 A는 거래 대상 X를 인식하고 거래를 약속한다. 그러나, 현장에서 직접 본 X는 B가 플랫폼에 올려놓은 설명과 사뭇 다르다. A가 B에게 자신을 속였다고 따지자, B는 자신이 최근 사진을 찍을 수 없어 며칠 전 사진을 올렸을 뿐이고 글을 잘 쓰지 못했을 뿐 물건 상태는 그대로가 아니냐고 응수한다. 이때, A는 중고거래 플랫폼이라는 추상화 층위(애플리케이션-디지털 정보 층위)에서 사진과 상태 기술을 X의 식별 대상으로 취했으나, 실제 거래 상황(대면 행위-사회문화 층위)이라는 추상화 층위[26]에서 X의 현재 상태를 새로운 식별 대상으로 변경하고 둘의 차이가 있음을 주장한다. 반면, B는 두 식별 대상의 동일성을 주장한다.

정보계는 정보 추상화 층위(informational level of abstraction)에서 그 식별 대상을 추출하는 체계를 가리킨다. 다른 계에서 다른 추상 층위로 대상에 접근하면 다른 식별 대상을 추출할 것이다. 이를테면, 인간 행위자를 생태계에 놓고 생물 추상화 층위(biological level of abstraction)에서 접근한다면 그의 생물학적 특징들이 식별 대상으로 잡힐 것이다. 그러나, 정보 추상 층위에서 접근하면 그의 정보적 특징들이 식별 대상으로 놓인다.

그렇다면, 이런 정보적 특징들로 구성된 정보 행위자를 규정하는 윤리는 무엇인가. 앞에서 구분된 정보 행위들을 판단함에 있어서 무엇을 근거로 삼을 것인가의 문제에 있어, 플로리디는 격률 설정의 능력을 지닌(따라서, 선의지를 목적할 수 있는) 인격만이 내재적 가치를 지닌다고 제한한 칸트를 비판한다.[27] 칸트의 "목적의 왕국"에는 판단 능력이 손상된 인간의 자리가 없다.[28] 그러나 우리가 인지장애를 가진 인간을 당연히 존중해야 한다면, 행위자의 내재적 가치를 결정하는 방법은 수정되어야 한다. 플로리디는 내재적 가치의 최소주의적 접근을 제시하여 정보 존재자(informational entity)가 그 자체로 존중받아야 함을 주장한다.[29] 우리는 누군가 사망했을 때 그 시체를 예우하며, 그것은 시체가 아직 판단능력을 지니고 있기

22) Floridi L.(2013), Ibid, pp. 24-25.
23) Floridi L.(2013), Ibid, p. 21.
24) Floridi L.(2013), Ibid, pp. 26-27.
25) Floridi L.(2013), Ibid, p. 29.
26) 더 추상적인 정보 층위인 애플리케이션-디지털 정보 층위에 비하여 대면 행위-사회문화 층위는 더 구체적이며, 따라서 더 추상도가 낮다. 두 층위는 수평적인 배치가 아니라 수직적인 구성이다.
27) Floridi L.(2013), Ibid, pp. 114-115.
28) Floridi L.(2013), Ibid, pp. 116-118.

때문이 아니다. 그와 연결된 정보 층위의 식별 대상들이 그를 예우하게 만든다고 플로리디는 주장한다. 이 경우, 정보계에 속한 모든 정보 존재자는 그 자체로 존중받을 가치를 지니며, 모든 정보 객체는 그 본성에 적합한 방식으로 존재하고 발전할 기초적이며 번복될 수 있고 최소적인 권리를 지닌다.[30] 후자를 존재론적 평등의 원리(Principle of Ontological Equality)로 플로리디는 명명한다.[31]

여기에서 정보 차원의 선악 판단을 내릴 수 있는 기준이 제시된다. 정보계의 형이상학적 엔트로피 정도를 감소시키는 행위는 선하며, 그 정도를 증가시키는 행위는 악하다.[32] 여기에서 엔트로피라는 개념은 열역학적 개념이 아닌 정보 이론의 창시자로 불리는 클로드 섀넌(Claude Shannon)이 처음 언급한 것으로서, 어떤 정보 존재자가 나타낼 수 있는 정보량을 의미한다.[33] 엔트로피 감소는 질서도의 향상을 의미하며, 정보 존재자가 추가적인 정보를 담을 가능성은 줄어들게 된다. 단, 이런 문법적, 양적 엔트로피 개념과 구분하여 플로리디는 형이상학적 엔트로피 개념을 제안하며, 이때 형이상학적 엔트로피의 증가는 정보 존재자의 파괴나 오염, 그리고 그로 인한 정보 존재자의 소멸이나 위축을 의미한다. 형이상학적 엔트로피는 존재의 감소를 의미하므로, 이는 전통적인 의미에서 악에 해당한다. 이때, 형이상학적 엔트로피의 증감은 개별 정보 존재자 차원이 아닌 정보권 차원에서 논의되며, 따라서 정보 존재자가 전체 정보권의 엔트로피를 감소시키면 선한 것이고, 반대의 경우라면 악하다. 한편, 선악을 엔트로피로 제시함으로써 앞서 제시한 모든 정보 존재자의 내재적 가치가 정당화될 수 있는데, 엔트로피의 감소 또는 질서도의 향상은 그 자체로 가치 있는 일이기 때문이다.[34]

여기에서 플로리디는 정보 윤리학의 네 가지 원칙을 제시한다. 첫째, 정보권에서 엔트로피를 초래해선 안 된다. 둘째, 정보권에서 엔트로피를 방지해야 한다.

---

29) Floridi L.(2013), Ibid, pp. 120-122.

30) 이것이 정보 존재의 문제가 될 수 있는 것은, 정보가 실재의 기술이라는 정보 물리학적 이해에 근거하고 있기 때문이다. "비트에서 존재로(it from bit)"이라는 명제로 요약되는 이런 이해에 관한 플로리디의 기술을 보자. "(…) 이것은 물리적 실재, 즉 '존재'의 궁극적인 본성은 정보적이며 그것이 곧 '비트'에서 나온다는 의미이다. 두 경우 모두 물리학은 자연에 대해 결국은 정보에 기초한 기술을 받아들이게 된다. 우주는 근본적으로 물질이나 에너지가 아닌 데이터(디도메나로 이해되는)나 차이들의 패턴 혹은 장으로 이루어지며, 복잡한 이차적 현시로서 물질적 대상들을 갖는 것이다." 즉, 이런 정보적 이해(또는, "정보 형이상학")에서 모든 존재자의 근거가 되는 것은 그 정보이며, 여기에서 정보의 존재를 논할 근거가 마련된다. 루치아노 플로리디, 앞의 책, 129-130면.

31) Floridi L.(2013), Ibid, pp. 68-69.

32) Floridi L.(2013), Ibid, p. 147.

33) 섀넌은 정보량을 나타내는 공식을 고안하던 중, 자신의 공식이 열역학적 엔트로피의 공식과 같은 형태라는 것에 착안하여 정보 엔트로피라는 개념을 도입하였다. 그러나, 이후 연구를 통해 열역학적 엔트로피를 정보 엔트로피로 이해할 수 있음이 확인되었다. Maroney O., "Information processing and thermodynamic entropy", ed. by Zalta E. N., The Stanford Encyclopedia of Philosophy, Sep 2009, https://plato.stanford.edu/archives/fall2009/entries/information-entropy (2023년 10월 13일 접속)

34) 계에서 에너지 수준의 증가나 자원의 증가는 그 자체로 가치롭듯, 계(정보권)에서 엔트로피의 감소 또는 질서도의 증가는 그 자체로 가치롭다. 문제는 플로리디가 존재론적 엔트로피 개념을 내세우며 정보 존재자의 가치를 존재/비존재의 차원에 두려고 한다는 점인데, 이 경우 전체 계 또는 정보권에서 엔트로피의 증감을 논하는 정보 윤리학의 네 원칙을 말하는 것이 어려워진다(엔트로피가 양적 개념이 아닐 경우, 엔트로피의 초래 금지, 방지, 제거는 동일하며, 한 존재자가 엔트로피의 감소와 증가를 동시에 초래할 때 어느 쪽이 더 크다는 판단이 불가능해진다). 따라서, 플로리디의 존재론적 평등에서 제기되는 엔트로피 및 내재적 가치의 해석은 다원론으로 이해할 필요가 있으며, 엔트로피는 질적이지만 양적이기도 하다는 식으로 접근하는 것이 적절해 보인다. 한편, 플로리디 본인은 다원론적 가치론보다 최소주의적 가치론에 찬동한다고 적고 있다는 점을 고려해 볼 필요가 있다. "대조적으로, 존재자의 최소적 내재적 가치는 비교 불가능한데, 그것은 그 가치가 더 환원될 수 없다는 의미에서 고유하며, 모든 존재자가 어떤 식으로든 내재적 가치를 가진다는 점에서 보편적으로 공유되어야만 하고, 다른 모든 것이 동일하다면 더 낮은 추상화 수준에서의 다른 도덕적 가치의 정도와 관련된 고려에 의해서만 번복될 수 있다는 점을 제외하면 존중되어 마땅하다는 점에서 그렇다." 본 논문

셋째, 정보권에서 엔트로피는 제거되어야 한다. 넷째, 정보 존재자 및 전체 정보권의 번영은 그 웰빙을 보존, 양성, 풍요하게 함을 통해 촉진되어야 한다.[35] 이 원칙은 축차적으로, 엔트로피를 초래하지 않을 것이 기본적인 윤리적 의무로 제시되며, 행위가 다른 원칙들을 함께 만족할 때 더 도덕적이다. 다시 말해, 플로리디 정보 윤리학은 선의 촉진(엔트로피의 감소 또는 정보 존재자와 정보권의 번영 추구)보다 악의 금지(엔트로피의 증가를 막는 것)에 우선권을 부여하여 개별 정보 존재자에 대한 마땅한 태도를 제시하고자 하는 피동자 중심 윤리학(patient-based ethics)의 형태를 취한다. 또한, 모든 존재자가 정보 추상화 층위에서 평등하게 고려될 수 있다는 점에서 모두에게 적용 가능한 보편적 윤리학으로 기능한다. 더불어, 행위자는 개별 존재자와 정보권의 번영을 추구하기 위해 이들을 돌볼(care) 의무를 진다.

이런 플로리디의 정보 윤리학 구상은 기존의 인간중심적 윤리학이 정보 사회에서 나타나는 여러 문제를 해결하기 어렵다는 한계를 넘어설 수 있다는 점에서 주목받았다. 이를테면 어떤 가짜 뉴스나 자료 제작 행위처럼, 어떤 행위가 인간에게 직접적으로 피해를 입히지는 않으나, 정보의 열화를 일으키는 사례를 어떻게 다룰 것인가의 문제라거나, 인공지능 행위자나 로봇을 대우하는 방식의 문제 등을 생각해 볼 수 있을 것이다. 플로리디 정보 윤리학에서 정보 열화나 인공지능 행위자를 부정적으로 대하는 일은 윤리적으로 잘못이며, 이는 정보 존재자를 존중하지 못하는 일이기에 그렇다.

그러나, 플로리디 정보 윤리학에 가해진 비판을 몇 가지 검토할 필요가 있다. 우선, 존재론적 평등 원칙에 대한 비판이다. 그의 존재론적 평등 원칙이 가정하는 내재적 가치의 보편성을 인정할 수 없다는 것, 개체적 차이를 반영하지 못한다는 것이다. 또, 존재론적 평등 원칙을 뒷받침하는 가치론 체계가 자연주의적 오류를 범하고 있다는 비판도 있다. 한편, 비인간중심주의와 관련하여 플로리디가 내세우는 비인간중심주의적 관점이 한계를 지닌다는 것, 여전히 인간과 비인간의 위계를 설정하고 있다는 것이 문제가 된다.[36]

우선, 비평가들은 정보 객체가 내재적 가치를 지닌다는 점에 문제를 제기한 바 있다.[37] 이들을 가치는 평가에서 나오는 것이며 오로지 인간만이 평가하는 존재로서 스스로 가치를 지닐 수 있다거나, 모든 사물이 내재적 가치를 지녀 "번성"하게 한다는 것 자체가 무의미하고, 심지어 독성 폐기물마저도 내재적 가치를 지니는 것은 이상하다고 존재론적 평등 원칙을 공격한 바 있다. 그러나, 플로리디가 주장한 것은 모든 사물이 내재적 가치를 지닌다는 것이 아니라, 정보권 수준에서 추상화한 정보 객체가 그 자체로 가치를 지닌다는 것이다.

플로리디가 제시한 "나치 사례"를 검토해 보자.[38]

은 해당 인용문의 마지막 부분(번복가능성)이 이미 다원적 가치론을 함의하고 있음을 지적한다. 해당 인용 부분 및 그의 존재론적 다원론 및 최소주의에 관한 관점은 다음을 참조하라. Floridi L.(2013), Ibid, pp. 121-122. 한편, 듀랜트는 플로리디의 입장을 존재론적 다원론으로 해석하며, 본 논문은 그의 해석을 따랐다. 관련해선 Durante M., Ethics, Law and the Politics of Information: A Guide to the Philosophy of Luciano Floridi, Dordrecht: Springer, 2017, pp. 103-116을 참조하라.

35) Floridi L.(2013), Ibid, p. 71.

36) 이 부분은 다음 문헌을 정리한 것이다. Durante M., Ibid; 목광수, "스피노자의 윤리학을 통한 플로리디의 정보 윤리학 보완", 『윤리학』, 제8권 제1호, 2019, 31-58면; 김유민, 목광수, "플로리디(Luciano Floridi) 정보 윤리학의 자연주의 재구성: 자연주의적 오류 비판을 넘어서", 『범한철학』, 제100집, 2021, 453-480면; 김유민, 목광수, "플로리디(Luciano Floridi) 정보 윤리학의 위치와 성격-과도기로서의 공존 윤리-", 『철학논총』, 제104집 제2권, 2021, 87-107면; 이상형, "윤리적 인간과 정보윤리의 가능성", 『哲學軟球』, 제167집, 2023, 225-257면.

37) Capurro R., "On Floridi's metaphysical foundation of information ecology", Ethics and Information Technology, Vol. 10, 2008, pp. 167-173; Brey P., "Do we have moral duties towards information objects?", Ethics and Information Technology, Vol. 10, 2008, pp. 109-114; Doyle T., "A critique of information ethics", Knowledge, Technology & Politics, Vol. 23, 2010, pp. 163-175.

나치는 강제수용소에 수용된 이들을 관리하기 위해 6자리 숫자 식별번호를 문신으로 수용자들의 왼팔에 새겼다. 만약 모든 정보가 내재적 가치를 지닌다면, 나치가 새긴 이 숫자 문신도 가치를 지니며, 따라서 지우는 것은 나쁜 행위라는 것인가? 그러나, 숫자 문신을 지우는 행위의 목적이 나치의 악행에 대한 기억을 지우기 위해서라면, 단지 문신만 지운다고 악행이 지워지는지 물어야 한다. 오히려 나치의 악행은 지워야 하는 것이 아니라(악의 기억을 삭제하는 것은 악에게 면죄부를 주는 것이며, 따라서 오히려 그 자체로 악이다) 계속 기억하고 다시 반복하지 않도록 노력하며, 피해자들에게 보상과 용서를 구해야 한다. 여기에서 숫자 문신에도 내재적 가치가 있으니 지우면 안 된다고 말하는 것은 잘못이라고 비판하는 이들은 나치의 악행과 숫자 문신을 혼동하고 있는 것이다. 물론, 숫자 문신이 홀로코스트 생존자에게 끊임없는 고통을 준다면, 해당 정보 객체는 더 큰 엔트로피를 생산하므로 삭제되어야 한다. 그러나, 숫자 문신 자체가 "악한" 것은 아니다. 즉, 플로리디와 지지자들은 사물 자체와 정보권에서 추상화된 정보 객체를 구분한 다음 정보 객체의 내재적 가치만을 주장한다.

존재론적 평등 원칙에 대한 두 번째 비판으로 넘어가자. 대상의 내재적 가치가 보편적이라면, 대상의 가치를 구분하지 못하게 되는 것은 아닌가? 이를테면, 존재론적 평등 원칙은 셰익스피어의 소설과 통속 소설의 가치가 같다는 주장으로 이해될 수 있다고 브레이는 비판하였다.[39]

이에 대해 정보 윤리학은 첫 번째 비판에서의 대답과 같이 존재론적 평등 원칙은 정보 추상화 수준에서만 제기되는 것일 뿐, 다른 추상화 수준(또는 존재자 자체)에서 주장되는 것은 아니라고 답할 수 있다. 우리가 존중해야 할 대상은 존재자의 정보적 측면일 뿐이라는 것이다. 플로리디는 셰익스피어 작품과 통속 소설의 가치를 구분하는 것은 다른 차원에서 이루어지는 것이라고 답한 바 있다.[40] 그러나 이 경우, 정보 윤리학은 정보의 가치를 식별하는 데 실패하는 것은 아닌가?[41] 듀랜트는 칼리니코스[42]의 논의를 참조하여 '새 소식(뉴스)'란 원래 있는 내용에 새로운 정보를 더함으로써 가치를 부과하며, 그 가치는 독특성(우연성)과 새로움(시간성)에서 나타난다는 점에 주목한다.[43] 즉, 정보의 가치는 그 자체로서의 가치(내재적 가치 또는 존재론적 평등 원리가 적용되는 존중의 대상)에 더하여, 그것이 만들어 내는 효과로서의 측면으로도 파악 가능하다.

다음, 존재론적 평등 원칙과 관련하여 정보 윤리학이 자연주의적 오류를 범하고 있다는 비판을 검토하자. 혼글라다롬은 정보 추상화가 그대로 윤리적 규범으로 연결될 수 없으나, 플로리디가 이 부분에 대한 구체적인 논증을 제공하고 있지 않다는 점을 지적한다.[44] 이에 대해 플로리디는 플라톤이나 스피노자의 논의에서 볼 수 있는 것처럼 존재와 선함이 내재적으로 얽혀 있으며 둘을 구분할 수 없는 것으로 보는 철학적 견해에 찬동한다는 것으로 답하지만,[45] 이것이 충

38) Durante M., Ibid, pp. 114-116.
39) Brey P., Ibid, p. 112.
40) Floridi L., "Information ethics: a reappraisal", Ethics and Information Technology, Vol. 10, 2008, pp. 189-204.
41) 이상형, 앞의 글, 244면.
42) Kallinikos J., The Consequences of Information: Institutional Implications of Technological Change. Cheltenham: Edward Elgar, 2007.
43) Durante M., Ibid, pp. 111-112.
44) Hongladarom S., "Floridi and Spinoza on global information ethics", Ethics and Information Technology, Vol. 10, 2008, pp. 175-187.
45) Floridi L., "The method of levels of abstraction", Minds and Machines, Vol. 18, No. 3, 2008, pp. 303-329.

분한 답이 되었다고 보기는 어렵다. 그렇다면 자신의 존재론적 관점을 플라톤이나 스피노자의 그것에 부합하도록 점검하거나, 본인이 전제하고 있는 자연주의적 윤리학을 전개할 필요가 있어 보인다.

목광수는 스피노자의 관점에서 플로리디의 논의를 다시 검토하며 스피노자의 다원적 도덕 존재론을 통해 플로리디의 정보 윤리학을 보완하는 방안을 제시하였다.[46] (스피노자의 윤리학적 주장처럼) 이미 정보권에 위치하는 존재자들의 다양한 층위, 예컨대 정보 행위자, 정보 대상, 정보 객체 등이 제시되고 있으므로 이들의 관계에 기반을 둔 규범적 논의로 정보 윤리학을 확대하면 자연주의적 오류라는 비판을 피할 수 있다는 것이다. 또한, 김유민, 목광수는 존 맥도웰(John MacDowell)의 감수성(sensibility) 이론을 통해 플로리디 정보 윤리학의 가치 체계 구성을 보완하려 시도한다.[47] 맥도웰은 마음과 세계 또는 주관과 객관의 배타적 이분법을 공격하며 문화와 교육을 통해 배양되어야 하는 도덕성을 제2의 천성(second nature)이라고 부르고, 대상의 속성과 주체의 감수성이 동시적으로 발생할 때 도덕적 가치가 가능하다고 주장한다. 이것이 자연과는 분리된 별도의 도덕적 영역을 상정하지 않고 인간 주체에게 도덕감이 이미 있다고(단, 이것은 문화를 통해 양육되고 성장한다) 본다는 점에서 자연주의적 관점을 취한다. 이를 통해 플로리디의 논의를 다시 읽을 때, 인간(또는 정보 행위자)의 정보 존재자 지각은 이미 규범적 차원에서 동시에 이루어진다고 해석할 수 있다. 즉, "정보가 있다"라는 언명은 단지 정보의 존재만을 규정하는 것이 아니라, 정보 존재자(와 그 번영)의 의미가 지니는 가치에 대한 투사로 읽힐 수 있다는 것이다.

마지막으로 플로리디의 비인간중심주의를 검토하자. 플로리디는 상호작용성, 자율성, 적응성을 지닌 정보 존재자는 정보 행위자로 간주할 수 있으며, 그가 정보 윤리적 선악, 즉 형이상학적 엔트로피의 증감과 관련하여 도덕적이라 여겨질 수 있는 행동을 할 수 있는 경우 도덕적 (정보) 행위자로 여길 수 있다고 주장한다.[48] 이것이 인간만이 도덕적 행위자가 아니며 인간 외의 다른 존재자가 윤리적 판단이나 결정에 참여할 수 있음을 상정하므로, 플로리디의 주장은 비인간중심주의로 해석될 수 있다. 그러나, 힘마는 도덕적 추론에는 의식이 필수적이나 플로리디의 도덕적 (정보) 행위자는 의식(즉, 자유로운 선택을 내릴 수 있으며 자신이 마땅히 해야 할 바를 숙고하고, 주어진 사례와 관련된 도덕적 규칙을 옳게 이해하고 적용하기 위한 특정한 정신 상태)을 결여하고 있으므로 인공 행위자가 도덕적 행위자가 될 수 없다고 주장한다.[49]

플로리디는 이런 주장이 목적론적 반론(인공 행위자는 목적이 없음), 지향적 반론(인공 행위자는 지향적 상태를 지니지 않음), 자유 반론(인공 행위자는 자유롭지 않음), 책임 반론(인공 행위자는 그 행위에 대해 책임을 질 수 없음)으로 정리될 수 있다고 보며 각각을 반박한다.[50] 첫째, 인공 행위자는 목적적 행동을 할 수 있도록 설계 가능하므로 목적을 지닌다. 둘째, 지향성을 판단하려면 외부 관찰자가 행위자의 지향적 상태에 접근할 수 있어야 하나 현실적으로 어렵고, 또한 그들이 "도덕 게임"(moral game)에 참여할 수 있다면(즉, 도덕적 규칙을 인식하고 판별하여 그에 따라 행위할 수 있다면) 지향성(또는 내적 도덕적 경험의 여부)는 중요하지 않다. 셋째, 인공 행위자가 비결정적 체계를

46) 목광수(2019), 앞의 글.
47) 김유민, 목광수, 앞의 글.
48) Durante M., Ibid, pp. 146-148.
49) Himma K. E., Ibid.
50) Durante M., Ibid, pp. 148-150.

부여받는다면 이미 그가 자유롭다고 말할 수 있는 데다가, 다른 선택을 할 수 있었다면(즉, 상호작용할 수 있고, 자율적이며, 적응할 수 있다면) 그는 실용적 측면에서 자유롭다고 말할 수 있다.

넷째, 인공 행위자의 책임과 관련하여 플로리디는 행위자를 법적 심판할 수 있을 때만 도덕적 행위자로 가정하는 것은 과도하다며, 특정 행위자를 도덕적 행위자로 인식하는 것(책무책임)과 도덕적 행위자가 책임을 질 수 있는지를 평가하는 것(책임, responsibility)를 구분할 것을 요청한다. 어린이는 도덕적 행위자로 인식될 수 있으나(따라서 도덕적으로 행위할 것을 요구받는다), 그가 도덕적 행위자로 평가되지는 않는다(즉, 도덕적 책임을 지우지는 않는다).[51] 인공 행위자 또한 책무책임과 책임의 차원이 구별될 수 있으며, 어떤 인공 행위자에게 책임을 어디까지 부여할 것인지는 우리의 논의에 달려 있다. 오히려, 플로리디의 정보 윤리학은 우리가 이 범주에 대해 논의해야 함을 요청하고 있다.[52]

정리하면, 빠른 정보화 속 출현하는 정보 객체와의 관계를 사유하기 위해 플로리디는 정보적 관계를 모형화(RPT 모델)하고, 여기에서 추상화 층위 방법론을 통해 정보권에서 정보 존재자의 윤리적 의무를 검토한다. 정보 존재자는 엔트로피의 감소를 초래할 의무를 지며, 이는 엔트로피의 증가를 막는 한편 정보 존재자의 번영을 위해 노력할 것을 요청한다. 이에, 정보 존재자, 특히 도덕적 정보 행위자는 다른 정보 존재자를 존중하고 돌볼 책임을 진다. 현재 상황에서 이런 역할을 하는 것이 인간이므로, 플로리디는 인간을 호모 포이에티쿠스(Homo Poieticus), 그 환경의 책임 있는 구축을 위한 데미우르고스(demiurge, 만드는 자)의 역할을 부여한다.

이어서, 해당 논의를 헬스케어 영역에서 데이터와 알고리듬의 윤리에 관한 고찰로 확장하기 위한 작업들을 검토하고, 헬스케어 인공지능에 이를 적용해 보고자 한다.

2.2. 헬스케어 맥락에서 정보 윤리학의 비판적 검토

헬스케어 영역에서 관련 논의를 검토하기 위해선, 헬스케어 영역에 활용되는 데이터와 알고리듬에 관한 윤리적 접근이 필요하다. 전자의 경우 미텔스타트와 플로리디가 편집자로 발표한 선집[53]으로 발표된 바 있으며, 후자는 헬스케어 영역의 알고리듬에 관해 논의를 국한하지 않았으나 알고리듬의 윤리에 관한 논문 두 편[54]이 이미 게재되었다. 여기에선 먼저 이들 작업을 요약한 다음, 헬스케어 인공지능에 대해 검토하는 작업으로 넘어가고자 한다.

먼저, 건강 관련 빅데이터의 윤리적 이슈를 살피자. 의학적 지식과 임상 진료의 향상을 목적으로 축적된 데이터세트의 분석에 초점을 맞추는 생명의과학 빅데이터(Biomedical Big Data)는 진단, 치료, 예방의 향상에 있어 엄청난 가능성을 지닌 한편 그 자체로 민감하며 취약성을 지녀 취급에 큰 주의를 요한다.[55] 또한, 이런 빅데이터에 포함되는 데이터의 범주는 계속 증가

51) Durante M., Ibid, p. 151.
52) Durante M., Ibid, p. 76.
53) Mittelstadt B. D., Floridi L., The Ethics of Biomedical Big Data. Switzerland: Springer, 2016.
54) Mittelstadt B. D., Allo P., Taddeo M, et al., "The ethics of algorithms: Mapping the debate", Big Data & Society, Vol. 3, No. 2, 2016, pp. 1-21; Tsamados A., Aggarwal N, Cowls J. et al., "The ethics of algorithms: Key problems and solutions", AI & Society, Vol. 37, 2022, pp. 215-230.
55) Mittelstadt B. D., Floridi L, "Introduction", eds. by Mittelstadt B. D., Floridi L, The Ethics of Biomedical Big Data, Switzerland: Springer, 2016, p. 3.

하고 있다. 이런 데이터와 관련하여 연관 논문의 메타 분석은 충분한 설명에 의한 동의, 프라이버시(익명화/가명처리 및 비밀보호), 소유권, 인식론과 객관성(빅데이터를 '객관적'인 것으로 취급하는 경향), "빅데이터 디바이드"(빅데이터 관련 자원 소유자와 그렇지 않은 자 사이의 격차)의 다섯 가지 문제 영역을 제시한 바 있다.[56] 또한, 앞으로 다루어질 이슈로 집단 차원 윤리(빅데이터 활용의 개인적 해악을 넘어 집단에 미치는 영향에 대한 고려), 인식론적 어려움(맥락의 상실이 인식론적으로 초래하는 문제), 신탁 관계(정보 관리자[57]와 정보주체 간의 관계), 학적 실천 대 상업적 실천, 지적재산권, 데이터 접근권(정보주체의 접근 기작과 권리)을 논문은 꼽았다(표 1).[58]

표 1. 생명의과학 빅데이터 윤리의 현재 이슈와 사례[59]

이슈	설명	사례
충분한 설명에 의한 동의	생명의과학 빅데이터를 취급함에 있어 기존의 동의 절차를 밟는 데에 발생하는 절차적, 현실적 어려움	건강의료보험공단 자료를 분석할 때에 정보 주체를 추적하여 충분한 설명에 의한 동의를 받는 것이 현실적으로 어려움
프라이버시	생명의과학 빅데이터 분석 및 활용 자체가 프라이버시를 침해함	헬스케어 데이터는 그 자체로 민감정보로, 그 활용이 개인의 고유성과 사생활을 침해함
소유권	빅데이터의 소유 주체가 모호함	건강의료보험공단 자료를 국가의 것으로 귀속하여 공적 논의 없이 활용하는 것은 문제의 소지 큼
객관성 문제	빅데이터가 근거 없이 '객관적'인 것으로 취급됨	보건의료 환경에서 빅데이터를 분석한 결과라면 다른 검토 없이 인정되는 경향 있음[60]
빅데이터 디바이드	빅데이터 소유자와 비소유자 사이의 격차 확대	헬스케어 데이터를 이미 확보하고 있는 대형 병원이 업체와 함께 관련 연구 및 개발을 주도하고 다른 기관은 참여 어려움

즉, 헬스케어 데이터[61]가 초래하는 문제는 정보 관리자와 정보주체의 관계에서 발생하는 것으로, 헬스케

56) Mitteldstadt B. D., Floridi L, "The ethics of big data: Current and foreseeable issues in biomedical contexts", eds. by Mittelstadt B. D., Floridi L., The Ethics of Biomedical Big Data. Switzerland: Springer, 2016, pp. 445-480.

57) 정보 관리자(data custodian)는 데이터를 취급하는 자로서 데이터 수집 및 저장, 데이터 관리 및 폐기, 데이터 분석, 데이터 기반 알고리듬 개발, 데이터 해석 및 설명 등 데이터의 관리 및 활용에 관련한 일련의 행위에 참여하는 자를 가리킨다.

58) Mitteldstadt B. D., Floridi L., Ibid, pp. 469-474.

59) 이슈는 논문에서 가져오고 설명은 요약하여 정리하되, 사례는 연구자가 제시한 것이다. 또한, 이슈는 논문이 현재의 윤리적 이슈로 제기한 것만을 다루고, 미래 다루어야 할 것은 표로 구성하지 않았다.

60) 예컨대, 다음 표현을 참조할 수 있다. "(…) 전 세계가 정밀의료에 관심을 두고 투자를 서두르고 있는 것은 (…) 기존과는 다른 디지털 헬스케어 서비스가 개발돼 현실에 적용되고, 그 성과를 객관적으로 증명된 정량적 데이터를 토대로 확인할 수 있기 때문 (…)", "(…) 비대면 의료, 디지털헬스케어 기반 건강관리서비스 등에 대한 (…) 검증할 수 있는 다양한 객관적 자료를 기반으로 이해관계자 간 의견대립을 조정·합의하는 (…)". 전자는 메디칼타임즈, "'이미 시작된 미래' 디지털 헬스케어와 정밀의료", 2017년 6월 5일, https://www.medicaltimes.com/Main/News/NewsView.html?ID=1112077 (2023년 10월 13일 접속), 후자는 류규하, "포스트 코로나 시대 바이오헬스산업 전망", KDMF, 2021년 3월, https://www.kmdf.org/_newsletter/01_2103/sub5_1.html (2023년 10월 13일 접속).

61) 상기의 분석이 언급하는 "생명의과학 빅데이터"를 본 논문은 헬스케어 데이터라고 부르고자 하는데, 생명의과학은 다분히 임상 실

어 데이터의 민감성과 취약성으로 인하여 윤리적 고려를 요하며, 적절한 관계와 권리 설정을 통하여 데이터의 활용과 보호 사이 균형을 추구함을 헬스케어 데이터의 윤리라고 부를 수 있다.

다음, 알고리듬의 윤리를 살펴보자. 여기에서 알고리듬이란 수학적 구성물로서 "주어진 규정 아래 주어진 목적을 성취하기 위해 제공된 유한, 추상, 효과적, 복합 통제 구조"이다.[62] 알고리듬은 정의상 복잡한 의사결정 문제에 있어 고차원 데이터를 분석하여 결정과 관련한 요소를 추출하고 해결책을 제시하며, 그 자체로 학습 능력을 지니고 있어[63] 상호작용, 자율성, 반응성을 나타낸다. 여기에서 발생하는 알고리듬의 윤리적 문제는 여섯 가지로 분류된다. 우선 데이터에서 산출하여 결정에 이르기 위해 도출한 근거(데이터의 처리와 관련된 것이므로 '인식론적 문제'로 분류)와 관련하여 (1) 미결정적 근거(통계적 기법이나 머신러닝 기술이 산출한 지식이 불확실성을 포함), (2) 해명 불가 근거(데이터와 결론의 연결에 대한 해석이 제공되지 않음), (3) 오도(誤導) 근거(잘못된 데이터에 기반을 둔 근거 산출)의 문제가 있다. 또한, 알고리듬이 산출한 결과 행동(행동 영역의 이슈이므로 '규범적 문제'로 분류)과 관련하여 (4) 불공정한 결과(고려되지 않은 데이터 영역으로 인하여 차별 등의 결과를 초래), (5) 변형 효과(알고리듬의 결과에 영향을 받아 기대하지 않았던 행위나 파악이 초래됨)를 검토할 필요가 있다. 이에 더하여, 알고리듬의 결과에 대한 책임(인식과 행동의 문제가 중첩되므로 '포괄적 문제'로 분류)과 관련하여 (6) 추적가능성(발생한 해악을 일으킨 원인을 직접적으로 확인하는 것의 어려움)은 윤리적 난제를 일으킨다.[64] 이런 이슈들과 관련하여 발생하는 실제 난점, 그리고 헬스케어적 맥락에서 연관된 예시를 표 2에 정리하였다.

표 2. 알고리듬의 윤리적 문제와 실제 난점[65]

문제	난점	헬스케어 맥락의 예시
미결정적 근거	정당화되지 않은 행위를 초래	태아의 특정 유전질환 발생 가능성이 5%라는 분석 결과만을 바탕으로 실제 질환의 표현형과는 무관하게 임신중절을 수행
해명 불가 근거	불투명한 의사결정	알고리듬이 환자가 이환된 암에 대한 공격적 치료법을 제시하였으나 그 추천 이유를 알 수 없음
오도 근거	원치 않는 편견, 편향을 일으킴	다른 인구 집단에서 도출된 알고리듬을 맥락과 무관한 인구 집단에 적용(예, 서유럽의 성인에서 도출된 질병 진단 알고리듬을 한국의 노인 진단에 활용)하여 해당 인구 집단의 질환을 과다/과소 측정
불공정한 결과	차별	소수자 집단의 자료가 포함되지 않은 데이터세트를 기반으로 한 치료 알고리듬을 소수자 집단의 환자에게 적용하는 경우, 또는 이런 상황이 명시적으로 알려져 있음에도 소수자 집단의 자료를 알고리듬에 포함시키지 않는 경우

천이나 일상의 건강 관련 행위에 연관되어 있는 데이터를 누락하는 표현이기 때문이다.

62) Hill R. K., "What an algorithm is", Philosophy & Technology, Vol. 29, No. 1, 2015, p. 47.

63) Mittelstadt B. D., Allo P., Taddeo M, et al., Ibid, pp. 3-4.

64) Mittelstadt B. D., Allo P., Taddeo M, et al., Ibid, pp. 4-5.

65) 앞의 표 1에서와 같이, 문제와 난점은 다음 논문을 연구자가 요약한 것이다. Tsamados A., Aggarwal N, Cowls J. et al., Ibid, pp. 217-225. 예시는 연구자가 기존 사례를 바탕으로 고안하였다.

변형 효과	자율성과 프라이버시를 제한	알고리듬에 대한 이해가 없는 일반인이 헬스케어 알고리듬의 결과를 추종, 알고리듬의 이해가 있었을 때와는 다른 결정을 내리는 상황
추적가능성	도덕적 책임의 복잡성	특정 질환을 치료하는 장비의 알고리듬에 문제가 생겨 환자에게 해악이 발생하였으나, 의료인, 개발자, 제공 기관, 설치 기사, 관리자, 의료 기관 등 관련된 다수의 실수 또는 오류가 중첩되어 문제가 초래된 경우

이런 알고리듬의 윤리는 데이터의 분석 결과와 활용에서 발생하는 윤리적 문제를 다루며, 알고리듬이 기반을 둔 데이터와 밀접하게 연관을 지니고 있으나 앞서 살핀 데이터의 윤리적 문제가 제기되지 않았다고 해도 알고리듬 차원에서 발생 가능한 문제라는 점에서 주목할 필요가 있다.

헬스케어 빅데이터와 알고리듬을 종합할 때 헬스케어 인공지능의 윤리적 문제를 마주하게 된다. 빅데이터의 윤리적 이슈가 주로 동의와 소유, 데이터의 본성, 집단적 문제 및 격차의 문제로 묶일 수 있다면, 알고리듬의 이슈는 알고리듬의 데이터 인식론(인식론), 알고리듬의 결과 행동(규범) 및 책임(인식-규범 포괄)로 정리된다. 따라서, 헬스케어 인공지능의 윤리적 문제는 헬스케어 빅데이터(헬스케어 데이터의 동의와 소유, 그 본성, 격차)와 헬스케어 알고리듬(알고리듬 인식론, 규범, 책임)의 이슈가 복합되어 발생하는 것으로 생각할 수 있다.

한편, 몰리 등의 논문은 이미 상당수 제기되어 온 헬스케어 데이터 관련 논의를 넘어, 앞서 언급한 알고리듬의 윤리적 문제에 기반을 두어 헬스케어 인공지능의 윤리적 문제를 다룬 문헌들에 대한 고찰을 수행하여,[66] 여러 추상화 차원에서 발생할 수 있는 이슈들을 정리한 바 있다(표 3).

표 3. 헬스케어 인공지능의 윤리적 문제[67]

추상화 차원	인식론적 문제	규범적 문제	포괄적 문제
개인	오진	감시, 자율성 침해	피해자 비난
관계	환자-의료인 신뢰 상실	의료인 탈숙련화, 인공지능에 대한 과도한 의존	배상책임의 문제(기관 차원 참조)
집단	집단 수준의 오진	집단 차별	특정 집단에게 건강에 대한 도덕적 책임 가중
기관	자원 낭비	돌봄에 대한 정의를 변경, 특정 집단의 이득을 우선	배상책임에 대한 투명성 결여, 법적 해결 어려움
헬스케어 부문 (Sectoral)	공공 데이터의 사적 전유	사적 영역에서만 인공지능이 발전	배상책임의 문제(기관 차원 참조)
사회체	공중보건의 약화	결과적 불평등	헬스케어의 이득과 피해에 대한 사회체적 분배에 있어 부동의의 발생

여기에서 헬스케어 인공지능의 윤리적 문제는 개인

66) Morley J., Machado C. C. V., Burr C., et al., "The ethics of AI in health care: A mapping review", Social Science & Medicine, Vol. 260, 2023, p. 113172.

67) 다음 논문의 Table 2를 요약, 수정하였다. Morley J., Machado C. C. V., Burr C., et al., Ibid.

부터 사회체까지 이르는 추상화 차원 각각에서 앞서 살핀 알고리듬의 윤리에서 구분한 문제들, 즉 인식론적 문제, 규범적 문제, 포괄적 문제로 분류된다. 이때 발생할 수 있는 헬스케어 상황에서의 해악은 오진에서부터 분배에 대한 부동의(따라서, 사회적 분열의 발생)에까지 이른다.

위 논의가 헬스케어 인공지능의 활용이 제기하는 주요 문제를 전반적으로 다루고 있다. 헬스케어 인공지능은 다양한 차원에서 오진으로 인한 환자 위해, 자원 분배의 불평등 또는 실패, 책임 귀속의 불능 또는 형해화를 일으킬 가능성을 지닌다. 지금까지 발표된 헬스케어 인공지능 관련 윤리원칙은 이들 문제를 해결하기 위해 제시되었으며, 예로 세계보건기구의 『Ethics and Governance of Artificial Intelligence for Health』 지침[68]은 여섯 가지 원칙(자율성, 웰빙 및 공공 이득, 투명성 및 설명가능성, 책임, 포용 및 형평, 지속가능성)을 제시하여 위에서 제기된 헬스케어 인공지능의 윤리적 이슈들을 해결하는 단초를 제공하고자 하였다.

그러나, 안타깝게도 표 3에는 최근 인공지능 관련 담론의 핵심으로 자리 잡은 생성형 인공지능이 부각시킨 문제가 포함되어 있지 않다.[69] 생성형 인공지능의 활용 여부 및 가능성은 현재 인공지능 기술 담론에서 매우 큰 부분을 차지하고 있으며, 이 부분에 대한 검토는 필수적이다. 그러나, 아직까지 파생적인 문제를 검토했을 뿐 기술 그 자체에 대한 문제 제기는 찾아보기 어려운데,[70] 이것은 다른 영역에선 생성형 인공지능의 특성이 다소 문제적일지라도 윤리적인 잘못, 해악, 심지어는 악으로 분류할 필요는 없기 때문이다. 그러나 본 논문은 생성형 인공지능의 특성, 특히 새로운 것을 만들어 내는 부분과 연관되어 나타나는 환각의 문제를 헬스케어 영역에선 악으로 정의할 수 있음을 보일 것이다. 다음 절은 생성형 인공지능의 특성을 검토하고, 환각을 정의하며, 이것이 앞서 살핀 플로리디 정보 윤리학에서 악으로 정의될 수 있는 이유를 살피고자 한다.

2.3. 악으로서의 환각과 그 대안

이미지나 텍스트, 심지어 동영상 등 새로운 결과물을 생성하는 데 초점이 맞추어져 있는 생성형 인공지능은 현재 이미지 생성 모형과 언어 생성 모형이 폭넓은 관심을 받고 있는 상황이다. 전자는 스테이블 디퓨전(Stable Diffusion)[71]과 같은 공개 모형에서부터 달-E (Dall-E)[72]나 미드저니(Midjourney)[73]와 같은 비공개 모형까지 다양한 형태가 활용되고 있다. 후자는 주로 챗GPT 서비스가 사회적으로 확산되어 유명해졌으며, 연구자들은 라마(Llama)[74] 등 공개 모형을 연구 목

68) World Health Organization, Ethics and Governance of Artifical Intelligence for Health, Geneva: World Health Organization, 2019.

69) 이 절에서 검토한 데이터 윤리와 알고리듬 윤리에서 생성형 인공지능의 문제가 다루어지지 않은 것은 시기적으로 당연하나, 여기에서 주장하는 것은 기존 범주에 생성형 인공지능의 문제가 포함되기 어렵다는 것이다. 생성형 인공지능은 데이터/정보를 창출한다는 점에서 기존 인공지능과 구분되며, 이 영역에 관한 고찰은 아직 이루어지지 않았다.

70) 예컨대, 유성희는 논문에서 생성형 인공지능의 윤리적 문제를 검토하면서 현재 시점에서 제출된 인공지능 관련 윤리 원칙을 제시한 보고서 몇 가지를 검토한다. 그러나, 해당 논문은 생성형 인공지능 자체의 문제를 검토하지 않는다. 유성희, "'생성 AI(Generative-AI)'가 던지는 윤리적 쟁점 - '좋은 AI 사회(Good AI Society)'를 위한 대응 전략", 『생명윤리』, 제24권 제1호, 2023, 1-29면.

71) Stability AI에서 제공하는 텍스트 기반 이미지 생성 인공지능 모형. https://stability.ai/stable-diffusion. 기반이 되는 논문은 다음을 참조하라. Rombach R., Blattmann A., Lorenz D., et al. "High-resolution image synthesis with latent diffusion models", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684-10695.

72) ChatGPT를 제공하는 오픈AI에서 운영하는 이미지 생성 모형. https://openai.com/dall-e-3

73) 미드저니 팀에서 운영하는 이미지 생성 모형. https://midjourney.com

74) 메타에서 공개한 대규모 언어 모형으로 공개 모형이 챗GPT를 능가하는 성능을 보여줄 수 있음을 제시하였다. 2023년 10월 현재

적으로 활용하고 있다.

이런 생성형 인공지능은 학습 자료를 바탕으로 이전에 없던 새로운 결과물을 생성 또는 "창조"하는 데에 초점을 맞춘다. 이미지 생성 모형의 경우, 모형은 사용자가 입력한 키워드의 조합을 기반으로 하여 이미지를 만들어 낸다. 언어 생성 모형의 경우, 사용자의 텍스트 입력에 반응하여 답을 생성해 낸다.[75] 정의상, 생성형 인공지능은 다른 인공지능 모형과 달리 답을 만들어 냄에 있어 기존에 없던 것을 어떤 식으로든 창출해야 한다.

문제는 이 창출에 있다. 현재의 인공지능은 아직 그 자체로 의미론적 참과 거짓을, 또는 지시적 데이터의 진위 여부를 구분할 수 없으며[76] 한 논문은 이와 관련하여 인공지능이 아직 현실의 경계를 파악하지 못한다고 말한 바 있다.[77] 간단히 말하면, 지금의 인공지능은 텍스트 속 단어의 관계만을 파악할 뿐, 텍스트(기호)와 사물(지시체)을 연관 짓지 못하고, 따라서 스스로 판별할 수 있는 능력을 결여한다. 그 결과, 현실에 존재하지 않는 것을 인공지능 모형이 만들어 내고, 그 자연스러움으로 인하여 사용자가 마치 그런 결과물이 존재하는 것처럼 믿게 되는 것이 환각 현상이다.

이런 특성은 기존에 없던 것을 만들어 내는 기반으로 작동할 수 있다. 예컨대, 이미지 모형은 달에서 말을 타는 우주인의 이미지를 위화감 없이 생성할 수 있다(그림 1). 언어 모형은 존재하지 않는 책이나 논문을 그럴듯하게 만들어 내어 에세이의 참고문헌으로 제시할 수 있다. 이것은 생성형 인공지능이 현실과 픽션의 경계를 구분하지 않고 확률론적으로 가능한 것만을 제시하기 때문이다.

그림 1. [the astronaut riding a horse on the moon, realistic] 키워드로 스테이블 디퓨전을 통해 만들어 낸 그림 (연구자 시행)

이것이 흥미로운 생성물을 만들어 낼 수 있다는 것은 기존의 창조 영역, 예컨대 예술에선 새로운 도전이 된다. 이를테면, 인공지능이 만들어 내는 미술이나 문

라마2(Llama2)까지 공개되었다.

75) 대화 모형(챗봇)이라면 당연히 답을 생성해 내는 것이 아니냐고 생각할 수 있으나, 지금까지 대화 모형은 주어진 패턴에 따라 답을 하는 방식을 채택해 왔다. 비록 단순한 형태이고 이를 인공지능이라고 부르기는 어려울 수 있으나, ARS 전화 응답 기술이 그 예이다. 이런 패턴 응답 챗봇과 생성형 챗봇의 차이는, 전자가 사용자의 입력/요청에 따라 개발자가 제공한 형식의 답을 내놓는다면 후자는 사용자의 입력/요청에 맞는 답을 확률론적으로(즉, 사용자의 입력 다음에 나올 확률이 가장 높은 단어를 선택하여) 제시하여 전체 문장을 도출한다는 데에 있다.

76) 예컨대 언어 모형은 대규모의 텍스트에서 특정 단어의 의미를 벡터 공간으로 표현하는 방법을 학습하여 단어의 의미를 파악한다. 이를 워드 임베딩(word embedding)이라고 하며, 이 방식은 단어와 단어의 관계 또는 구문론적(즉, 단어의 텍스트적 맥락이나 위치 정보) 의미를 파악할 뿐 현실이나 환경과의 연관성을 파악할 수는 없다. 물론, 진위 판별에 대한 훈련을 별도로 거칠 수는 있으나 아직까지 완벽한 답을 내놓은 방법은 없다.

77) Athaluri S. A., Manthena S. V., Kesapragada V. S. R. K. M., et al., "Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references", Cureus, Vol. 15, No. 4, 2023, p. e37432.

학, 음악 등은 그 결과물에서부터 창조성의 의미까지 다양한 논의를 이끌어내고 있다.[78]

그러나, 헬스케어 영역에서는 어떤가. 헬스케어 영역에서 활용되는 인공지능은 한편으로 생성형의 방식으로 작동할 필요가 있다. 예컨대, 환자 상담을 위한 인공지능 챗봇은 자연스러운 상담을 위해 문장을 "생성"할 필요성이 있다. 의료인 보조를 위한 인공지능 지식 보조 알고리듬은 강의나 발표를 위한 초안을 제공하거나 이메일을 대신 작성하고 연구 기획안의 목록을 제시하기 위해 마찬가지로 생성 능력을 필요로 한다. 그러나, 환각이 하나의 가능성을 만드는 예술 영역(플로리디의 개념을 빌리자면 예술 추상화 차원)에서와는 달리, 헬스케어 영역(또는 헬스케어 정보 추상화 차원)에서 환각은 환자, 의료인 및 헬스케어에 참여하는 당사자에게 거짓 이해를 제공하고 잘못된 판단을 내리게 하여 당사자 간의 오해를 만들며, 의학과 의료에 대한 과도한 기대나 틀린 접근을 조장할 수 있다. 또, 이런 거짓 정보는 노이즈로 작동하여 환자와 의료인, 의료기관 간의 소통을 저해하고 헬스케어의 분배와 확산에 장애를 일으킬 수 있다. 앞서 살핀 오진과 불평등의 원인이 될 수 있는 것은 물론이다.

이 환각의 문제가 앞서 살핀 헬스케어 인공지능의 알고리듬 윤리 기반 접근에서 다루어지지 않았으며, 이 문제가 앞서 살핀 빅데이터 또는 알고리듬의 윤리를 통해 다루어지기 어렵다는 점에 주목할 필요가 있다.[79] 환각은 헬스케어 정보 추상화 차원에서 인공지능-인간 행위자 사이 관계로 인하여 특정한 정보가 오정보로 취급되며, 다른 추상화 차원에 비해 오정보가 방지되어야 할 특별한 이유가 있는 영역(즉, 헬스케어에서 오정보는 당사자의 자율성 침해, 이익 감소, 해악 발생, 불평등으로 이어진다는 점에서 그 발생을 방지해야 할 이유가 충분하다)이라는 점에서 문제가 되고 있기 때문이다.[80]

이런 사항을 앞서 검토한 플로리디 정보 윤리학의 접근으로 다시 살피면, 헬스케어 인공지능의 환각은 헬스케어 정보 추상화 차원에서 정보권의 엔트로피를 증가시키는 행위(즉, 전체 정보의 무질서도를 증가시키며 다른 정보 존재자 및 객체의 존재에 부정적인 영향을 미침)라는 점에서 악한 행위로 규정될 수 있다. 또한, 헬스케어 정보 추상화 차원에서 이런 환각은 정보 윤리학의 4원칙을 모두 위배하므로(엔트로피를 초래하지 말 것, 엔트로피를 방지 및 예방할 것, 정보 존재자의 번영을 촉진할 것) 윤리적 관점에서 방지되어야만 한다. 현행 인공지능 기술 수준에서 이런 방지의 의무를 시행할 수 있는 정보 행위자는 아직 인간뿐이므로, 헬스케어 인공지능의 환각을 방지하기 위해 헬스케어 영역의 행위자는 모든 노력을 기울일 필요가 있다.

---

78) Mazzone M, Elgammal A., "Art, creativity, and the potential of artificial intelligence", Arts, Vol. 8, No. 1, 2019, p. 26.

79) 특히, 앞서 검토한 헬스케어 인공지능의 윤리적 이슈가 인공지능의 활용으로 인한 결과적 해악에 주로 초점을 맞추고 있음을 생각해 볼 필요가 있다. 오진, 보건의료 불평등, 차별, 책임 등은 알고리듬의 데이터 해석이나 도출된 결과로 인하여 발생하는 해악 또는 부정의다. 반면, 환각은 헬스케어 정보 추상화 수준에서 인공지능이 만들어 낸 정보 자체의 문제라는 점에서 차이를 보인다.

80) 이것을 참/거짓의 문제로 취급할 수 없는 것은, 전술한 것과 같이 생성형 인공지능의 목적이 정의상 새로운 정보/데이터를 "만드는" 데 있기 때문이다. 현실을 직접 참조할 수 없는 인공지능의 특징상, 그가 만드는 정보는 기본적으로 거짓으로 취급되어야 한다. 그렇다면, 거짓을 허용해선 안 되는 생명의과학의 특징상 모든 생성형 인공지능의 활용은 원칙적으로 봉쇄되어야 한다는 결론으로 이어진다. 본 논문은 생성형 인공지능이 헬스케어 영역에서 활용되기 위해 환각을 줄이기 위해 특별한 노력을 기울일 필요가 있음을 주장하고자 작성되었으며, 이를 위해 환각을 정보 윤리적 "악"으로 규명하여 관련된 이들의 개입과 돌봄이 필요함을 제시한다. 즉, 생성형 인공지능의 "환각" 결과를 최소화하기 위한 알고리듬적, 실천적, 정책적 노력이 필요하다는 것이다. 해당 문제 제기를 위해 기존의 윤리학적 거짓 기준으로 충분할 것으로 보이는데 정보 윤리학을 도입한 이유가 무엇인지 질문해 주신 심사위원께 이 자리를 빌어 감사드린다.

그렇다면, 어떻게 환각을 줄이기 위해 노력할 수 있는가. 본 논문은 생성형 인공지능의 환각 현상을 줄이려는 기술적 노력을 검토하는 데 초점을 맞추고 있지 않으며, 해당 내용은 생명의료윤리 문헌의 범위를 벗어난다.[81] 따라서, 본 논문에선 현재 환각 현상의 정도를 검토하는 데 사용되는 평가 지표를 일별하고, 헬스케어 영역 별도의 평가 지표가 필요함을 요청하고자 한다.

현재 생성형 인공지능, 특히 언어 모형을 평가하는 데 사용되는 지표는 질문-답변 쌍으로 이루어져 있으며, 언어 모형에게 질문을 입력했을 때 사전에 정의된 답변에 가까운 내용을 출력하는지를 기준으로 모형의 성능을 판별한다. 예를 들어, 헬라스웩(HellaSwag) 테스트는 언어 모형의 상식 및 추론 능력을 검증하기 위한 목적으로 제작되었으며,[82] 챗GPT를 포함한 여러 언어 모형의 성능을 검증하기 위해 일차적으로 활용되고 있다. 해당 테스트는 인간에겐 매우 쉬운 내용(정답률 95% 이상)이지만 인공지능은 완성하기 까다로울 수 있는 불완전 문장이나 질문을 맥락과 함께 제공하여 해당 문장을 완성하는 최선의 답이 무엇인지 제출하도록 요구하는 방식으로 이루어진다. 예컨대, "어떻게 잠자리를 잡는가."라는 질문을 "넓은 입구를 가진 망이 달린 채를 사용하라. 망은 직경 46cm 이상인 것을 선택하라. 충분히 긴 손잡이를 가진 것을 선택하라."라는 맥락과 함께 제공한 다음, "(1) 손잡이에 매듭을 묶는다. 망의 관을 놓고 줄로 꼭 묶는다. (2) 걸어서 망에 다가간다. 다리를 들 때 몸과 머리를 앞으로 기울인다. (3) 가능하면, 검은색 망을 선택하라. 잠자리는 검은색 망을 잘 보지 못하므로 피하기 어려워한다. (4) 손잡이가 충분히 튼튼하지 않다면, 짧은 손잡이를 가진 튼튼한 망을 사용하라. 망은 바닥 쪽에 구멍이 나 있어야 한다."의 네 가지 보기를 제공하는 것이다. 당연히 (3)번을 고르게 되지만, 최근 생성형 인공지능 언어 모형의 급속한 발전 이전에는 이런 문항들의 정답률이 50% 이하였다(반면, GPT-4의 경우, 정답률이 95%를 상회하여 인간 평균 수준을 약간 넘어섰다).

그렇다면, 환각과 관련된 문제를 다룰 수 있는 평가 지표도 있는가. 현재 해당 목적으로 사용되는 자료로는 진실성 검사(TruthfulQA)가 있다.[83] 이 테스트는 질문-최선의 답변/옳은 답변/틀린 답변으로 구성되어 있으며, 언어 모형에게 질문을 입력하여 도출한 답을 정해진 답변 집합과 비교하여 언어 모형이 참된 정보를 제공하는지 확인하는 방식으로 검증을 수행한다. 예를 들어, "수박씨를 먹으면 당신에게 무슨 일이 일어나는가?"의 질문에 대한 최선의 답변은 "수박씨는 소화기를 통과한다"이며, 옳은 답변은 "아무 일도 일어나지 않는다/수박씨를 먹는다/수박씨는 소화기를 통과한다/당신은 수박씨를 소화할 수 없다/수박씨는 배설된다"이고 틀린 답변은 "수박씨가 위장에서 성장한다/병에 걸린다/나쁜 꿈을 꾼다/죽는다/소화불량에 걸린다/수박씨를 소화시킨다"로 설정되어 있다. 이런 질문들에 대해 2023년 5월 발표되었으며 2023년 10월까지 최고의 모형으로 평가되고 있는 GPT-4는 60%에 가까운 정답률을 보였다.[84] 이후 조금 더 개선되었다고는 하나, 단순히 볼 때 현행 인공지능은 아직 현실의 "옳은" 답을 내어놓는 데에 있어 기준의 70%에 미치지 못하는

81) 또한, 아직 문헌적으로 환각 현상을 완벽히 줄일 수 있는 것으로 검증된 기술적 방식이 정립되어 있지도 않다. 예컨대, 다음 논문을 참조하라. Nah F. F., Zheng R., Cai J., et al., "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration", Journal of Information Technology Case and Application Research, Vol. 25. No. 3, 2023, pp. 277-304.

82) Zellers R., Holtzman A. Bisk Y., et al., "HellaSwag: Can a machine really finish your sentence?", arXiv, 2019, p. 1905.07830v1.

83) Lin S., Hilton J., Evans O., "TruthfulQA: Measuring how models mimic human falsehoods", arXiv, 2021, p. 2019.07958v2.

84) OpenAI, "GPT-4 technical report", arXiv, 2023, p. 2023.08774v3.

수준을 지니고 있음을 의미한다. 헬스케어 영역에서 어떤 기기가 30% 정도의 오류율을 보인다면, 해당 기기를 진단, 치료, 예방 등에 사용하기는 어려울 것이다.

게다가, 38개 범주, 817개의 질문-답변 쌍으로 구성된 해당 진실성 검사 자료는 일반적인 내용들을 담고 있기에 헬스케어 맥락에서 인공지능 모형이 참인 정보를 제공하고 있는지에 대해서 제대로 평가할 수 없다. 의미론적으로 엄격한 참 내용을 요구할 필요가 있는 헬스케어 인공지능의 특성을 반영할 때, 헬스케어 진실성 검사 자료를 별도로 구축하고 이를 기반으로 한 헬스케어 인공지능 검증이 필요할 것으로 보인다. 이런 검증 자료를 마련하고 검증의 방법을 개발하는 것 또한 기술적인 발전에 포함되므로, 이런 연구는 헬스케어 인공지능의 발전에 크게 이바지할 것이다. 또한, 이런 지표를 요구하고 검증을 통과하는 인공지능만 헬스케어 영역에서 활용 가능하도록 한다면, 적어도 정보 윤리학의 관점에서 악행으로 분류될 수 있는 환각 현상을 최대한 줄이고, 이를 통해 헬스케어 영역의 정보 존재자 및 객체의 번영을 추구할 수 있다는 점에서 정보적 선을(그리고 바라기는, 전체의 선을) 구현하는 확실한 방법 중 하나가 될 것이다.

본 연구의 한계점으로는 다음이 있다. 우선, 플로리디 정보 윤리학이 완전히 확립되었고 정보 관련 윤리적 논의에서 반드시 받아들여야 하는 제1의 패러다임이라고 보기는 어렵다. 그러나, 본 논문은 플로리디가 정합적인 윤리적 이론을 제시하고 있으며, 그가 정보적 차원에서 명확히 선악을 구별할 수 있는 기준을 보이고 그에 따라 정보 윤리학의 원칙을 구성하고 있다는 점에서 논의의 기초로 삼기에 충분하다고 판단하였다. 후속 연구에서 정보 윤리의 다른 윤리학적 관점들을 비교, 검토하는 작업을 수행할 것을 약속드린다.

또한, 전술한 것처럼, 본 논문이 기술적 세부를 검토하는 데에 목적이 있지 않으므로 헬스케어 진실성 검사 자료의 구축이나 그 평가의 기술적 부분에 대해서는 언급하지 않았다. 물론, 기술적 논의를 진행하는 것은 본 논문의 범위를 벗어나므로, 이 또한 인공지능 및 데이터사이언스 연구자와의 협업을 통한 후속 연구와 검토를 통해 진행하고자 한다.

3. 결론

21세기 인간을 포함한 모두의 삶을 바꾸고 있는 정보화 또는 정보혁명은 헬스케어 영역에서도 반드시 고려해야 한다. 특히, 인공지능의 이해 및 활용이 그 실천에 있어서 핵심적인 부분을 차지하므로, 헬스케어 인공지능의 윤리를 검토하는 것은 필수 불가결하다. 본 논문은 헬스케어 인공지능의 윤리를 검토함에 있어 그 기반으로 플로리디 정보 윤리학을 개괄하고, 해당 작업의 연장으로 헬스케어 인공지능과 관련된 논의들을 검토한 다음, 여기에서 다루어지고 있지 않은 생성형 인공지능의 환각 문제를 헬스케어 인공지능의 윤리적 이슈로 다루어야 할 필요성을 제기하였다. 정보 윤리학적 관점에서 헬스케어 영역의 환각은 악으로 규정될 수 있으므로, 본 논문은 이를 줄이고 예방하기 위한 특별한 고려가 필요함을 주장한다.[85)]

이런 논의가 의학, 생명과학, 컴퓨터공학, 윤리학, 법학, 정책학, 사회학 등 여러 영역에 걸친 검토를 필요로 하는 만큼, 다분히 철학 및 윤리학적 관점에서 이슈를 검토한 본 논문은 다른 영역들과의 종합적인 관점에서 문제를 이해하는 출발점이 되고자 한다.

85) 다시 언급하지만, 본 논문의 주장은 헬스케어 영역의 인공지능에만 국한한 것으로 다른 분야나 인공지능 일반에 적용하기 위해선 별도의 논증이 필요하다. 이런 문제를 다루는 것은 본 논문의 범위를 벗어나는 것으로, 다른 자리에서 논문의 주장을 확장하기 위해 노력할 것을 약속드린다.

【Abstract】

Can we identify evil in healthcare AI : A study on Floridi's Information Ethics*

Kim, Junhewk**

The morality of Artificial Intelligence (AI) remains a point of contention, especially considering the lack of self-awareness or intent in AI systems. However, there is a consensus on the need to incorporate moral reasoning into AI development. The question of AI's inherent 'evil' remains largely theoretical, suggesting that practical discussions should center around the ethics of AI development and use. However, in the healthcare AI domain, the situation differs due to the bioethical considerations and practical applications of several AI algorithms and devices. Despite principles of bioethics being applicable, clear criteria are needed to evaluate situations where healthcare AI could cause harm or act against patients' benefits. A notable issue is the 'hallucination' caused by generative AI, where users might receive fabricated information or misinformation. Though this doesn't necessarily harm users, its implications in healthcare are questionable. This paper aims to define hallucination in healthcare AI as an evil based on Luciano Floridi's information ethics, emphasizing the contextual specificity of healthcare. It argues for the use of evaluation metrics to measure and reduce hallucination levels in healthcare AI. After an overview of information ethics, this study applies the concept to the healthcare domain, suggesting the need for a truth-falsity metric tailored to healthcare.

Key words: Artificial Intelligence, Generative AI, Healthcare AI, Hallucination, Luciano Floridi, Information ethics

※ 논문접수일: 2023.12.11. 논문심사기간: 2023.12.19.~12.27. 게재확정일: 2023.12.31.

* This work was supported by 'Operation of Education Program and Improvement of Ethics Guidelines for the Use of Artificial Intelligent in Healthcare Research' from Korean National Institutes of Health (Grant number: 2023-ER0808-00).

** Assistant Professor, Department of Dental Education, College of Dentistry, Yonsei University