https://doi.org/10.15813/kmr.2024.25.3.001

AI와 윤리: 문헌의 종합적 분석과 정보시스템 분야의 향후 연구 방향[1)]

Navigating Ethical AI: A Comprehensive Analysis of Literature and Future Directions in Information Systems

민진영 (Jinyoung Min) 중앙대학교 산업보안학과[2)]

〈 국문초록 〉

AI의 사용이 일상 생활의 많은 부분에서 현실화 되어감에 따라 AI가 가져오는 긍정적인 기회와 혜택이 주목 받는 한편, AI가 초래할 수 있는 윤리적 문제들에 대한 염려도 커지고 있다. 정보시스템 분야는 기술이 비즈니스와 사회에 미치는 영향을 연구하는 분야로서 AI가 인류 사회에 바람직한 영향을 미칠 수 있도록 기여해야 한다. 따라서 AI와 윤리 관련한 다양한 연구들을 살펴보고 정보시스템 분야의 연구가 나아가야 할 방향을 탐색할 필요가 있다. 본 연구는 이를 위해 먼저 2020년부터 현재까지의 문헌을 수집하여 연구자의 코딩과 토픽 모델링을 통해 연구 주제를 범주화 하였다. 분석 결과 AI 윤리 원칙, 윤리적 AI 디자인 및 개발, 윤리적 AI 도입 및 활용, 윤리적 AI 사용의 네 가지로 연구 주제를 범주화하고, 각 범주 별로 문헌을 고찰하여 연구 현황을 짚은 후, 정보 시스템 분야에서의 AI 윤리에 대한 향후 연구 방향을 제언하였다.

주제어: AI 윤리, 윤리적 AI, AI 디자인 및 개발, AI 도입과 활용, AI 사용, 정보시스템

1) 이 논문은 2022년도 중앙대학교 학술연구비 지원에 의한 것임
2) jymin@cau.ac.kr

논문접수일 2024년 07월 31일 | 게재확정일 2024년 08월 26일

1. 서론

디지털 전환 시대를 주도하는 AI 기술의 특징은 자율(autonomy), 학습(learning), 헤아리기 어려움(inscrutability) 이라는 세 가지 특성으로 요약될 수 있다(김효은, 2022; Baird & Maruping, 2021; Berente et al., 2021). 첫 번째 특성인 자율(autonomy)은 자동(automatic)과는 구분되는 개념으로, 기계가 자동화 된다는 것은 인간의 도움을 최소화 한 상태에서 특정 과업을 미리 정해진 규칙대로 수행하게 된다는 것을 의미하지만 자율화 된다는 것은 기계가 의사 결정을 스스로 할 수 있게 되어 인간의 개입 없이 과업을 수행하게 된다는 것을 의미한다(김효은, 2022). 두 번째 특성인 학습은 현대 AI 기술의 대표적인 특징으로서 AI가 규칙이 아니라 데이터에 의해 좌우되므로, 입력 데이터가 변화하면 AI도 그것이 만들어진 시점과는 다른 존재가 될 수 있다는 것을 의미하기도 한다(김효은, 2022). 세 번째 특성인 헤아리기 어려움은 딥러닝(deep learning)으로 대표되는 현대 AI 기술의 블랙 박스와 같은 특성, 즉 입력과 출력은 존재하지만 어떻게 입력 데이터가 특정 출력 결과를 산출했는지 정확히 알기 어려움에서 기인한다. 이 세 가지 특징은 서로 연관되어 있는데 헤아리기 어려운 정도는 학습으로 인해 더 심화될 수 있고, 여기에 자율성이 가미되면 AI 기술의 헤아리기 어려움이 의사결정 결과의 헤아리기 어려움으로도 연결될 수 있기 때문이다(김효은, 2022). 앞으로 AI는 사회의 많은 부분에서 다양한 형태로 활발하게 사용되면서 그 영향력이 더 커질 것으로 예상된다. 따라서 AI의 이런 특징이 예상치 못한 부정적 결과를 가져올 수도 있다는 염려 또한 커지고 있다.

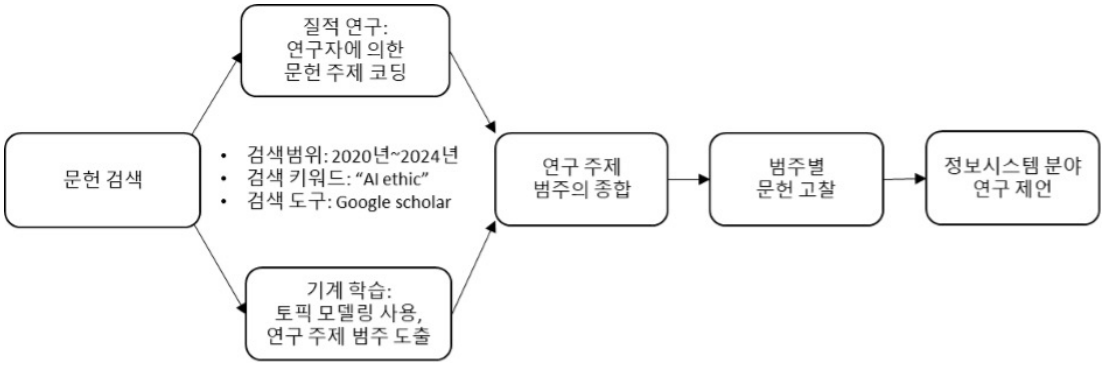
부정적인 결과에 대한 단편적인 예를 몇 가지 들면, 특정 데이터로 학습된 AI가 흑인의 사진을 인간이 아닌 고릴라로 인식하고(BBC, 2015), 신용 카드 신청 시 성별에 따라 신용도를 달리 평가하며(Nedlund, 2019), 재범 가능성 평가 AI 시스템이 특정 인종 및 지역 거주자의 재범율을 더 높게 평가(Heaven, 2020)한 예들을 들 수 있다. 이런 사례들은 AI 시스템이 인종, 성별, 사회-경제적(socio-economic) 편향(bias)을 보여준 잘 알려진 예들로써 AI가 사용하는 디지털 데이터가 편향되어 있어 나타나는 것들이다. 혹자는 현실 사회 자체가 절대적 평등과 공정이 발현되는 곳이 아니므로 AI가 현실을 오히려 잘 보여주는 것이라고도 이야기하기도 한다. 그러나, 대부분의 사람들과 연구자들은 일상 생활에서 편향된 AI 사용이 만연하게 되면 현실이 부정적으로 편향된 부분을 오히려 자연스러운 것으로 받아들이게 되고 이에 따라 이런 현상이 완화되거나 개선되지 않고 더욱 강화되는 디스토피아적인(dystopian) 미래를 염려하고 있다. AI가 가져오는 여러 긍정적 기회와 혜택에도 불구하고, AI 기술의 특징과 영향력을 고려하면 이러한 염려는 더욱 커질 수밖에 없다. 따라서 AI가 반드시 윤리(ethics)와 결합되어야 한다는 목소리가 커지고 있다.

정보시스템(IS: information systems) 분야는 기술이 비즈니스와 사회에 미치는 영향에 대해 그 어떤 분야보다 관심있게 연구해 온 분야인 만큼, AI로 인한 많은 변화가 궁극적으로 인류 사회에 바람직한 것이 되게끔 기여해야 한다. 그러기 위해서는 AI시대의 윤리란 어떤 것인가, 윤리적인 AI라는 것은 어떠한 것인가, AI를 윤리적으로 활용하기 위해서는 무엇을 해야 하는가 등에 대한 질문에 답을 제시할 수 있어야 할 것이다. 따라서 본 연구는 이를 위한 토대를 마련하기 위해, AI와 윤리에 관한 문헌들을 분석하고 고찰하여 정보시스템 분야의 바람직한 담론을 이끌어내기 위한 연구 의제(agenda)를 찾고 제언하는 것을 목표로 한다. 이를 위해 본 연구에서는 먼저 AI 윤리를 검색어로 하여 2020년부터 현재까지의 문헌들을 수집하고, 이 연구들의 주제를 연구자의 코딩과 토픽 모델링을 사용하여 범주화 하였다. 이후 각 주제별 문헌 고찰을 통해 AI와 윤리 문제의 현황을 짚은 후 정보 시스템 분야에서의 AI 윤리에 대한 향후 연구 방향에 대해 제언한다.

2. 연구 방법

본 연구에서는 체계적 문헌 연구와 주제 분석을 위하여 PRIMSA(Preferred Reporting Items for Systematic Review and Meta-Analysis) 방법(Moher, Liberati, Tetzlaff, Altman, & PRISMA Group*, 2009)을 기초로 토픽 모델링을 추가한 후 반 구조적 문헌 연구와 주제 분석 방법을 사용하였다. 이를 위해 먼저 연구 대상이 되는 문헌을 분야 제한하지 않고 수집하였다. AI가 여러 분야에서 활발하게 사용되고 이에 따른 윤리적 문제가 다양하게 나타남에 따라 AI와 윤리에 관한 연구는 필연적으로 다 학제적(interdisciplinary)인 성격을 띨 수밖에 없다. 따라서 특정 분야에 한정되는 문헌을 분석하기보다는 여러 분야의 문헌을 함께 분석하는 편이 종합적인 통찰을 제공할 수 있을 것이라 보았다. 이후, 수집한 연구의 초록을 이용하여 1) 연구자가 문헌의 주제를 식별하여 코딩, 2) 문헌의 초록을 이용한 토픽 모델링, 이렇게 두 가지 방법으로 데이터를 분석하고 결과를 종합하여 AI와 윤리 연구를 몇 개의 주제로 나눈 후 해당 주제를 잘 보여주는 논문들을 들어 연구 동향을 논의하였다. <그림 1>에 연구 과정을 도식화 하였다.

먼저 AI와 윤리에 대한 문헌들의 정보를 수집하기 위하여 Google Scholar 검색을 이용하였다. 검색 키워드는 윤리(ethics)뿐 아니라 윤리적(ethical), 윤리적으로(ethically)와 같은 형용사와 부사도 포함되게 하기 위하여 “AI ethic”으로 설정하였다. 문헌의 검색 기간은 AI 분야가 빠르게 변화하고 있는 분야인 만큼 최신 연구에 더 비중을 두기 위하여 최근 5년 간, 즉 2020년 1월부터 2024년 7월 현재 시점까지의 연구로 설정하였다. 보다 관련성 높은 연구들 위주로 검색되게끔 하기 위하여 2024년부터 시작 연도를 하나씩 감소시켜가며 두 개의 연도를 묶어 검색하였고 각각의 검색 기간에서 100개의 검색 결과를 수집하였다. 예를 들자면 2023년~2024년으로 첫 검색 기간을 설정하고 시작 연도를 하나 감소하여 다음 검색 기간을 2022년~2023년으로 설정하는 방식을 사용하여 각 검색 기간 마다 100개씩을 수집하였다. 수집 대상이 되는 100개의 선정은 기간 별 첫 검색 결과 100개씩으로 하였으며, 파이썬(Python) 코드를 이용하여 작성한 스크래핑(scraping) 프로그램을 통해 수집한 후 합치는 방식을 사용하였다. 그 결과 중복을 제외하고 362개의 연구가 수집되었다. 이 중에서 동료 심사(peer review)를 거치지 않은 연구나 사이트 연결 불가 등의 이유로 초록을 수집할 수 없는 연구를 제외하였다. 추가로, 인용 수에 따라서도 논문을 제외하여 연구자 커뮤니티에서 어느 정도 공감대를 얻은 연구들로 문헌 분석 대상을 구성하려 하였다. 이를 위해 2024년 출판된 연구들의 가장 적은 인용 수를 살펴보았는데 이 건수가 10건이었으므로 이 수치를 기준으로 하여 2024년 전 연구들에서 인용 수 10



〈그림 1〉 연구 과정

미만의 연구들을 제하였다. 이 결과 85개의 연구를 제하고 총 277개의 문헌이 분석 대상이 되었다. 다음으로 연구자가 연구의 초록을 읽고 연구 분야 및 주제를 직접 코딩하는 질적 연구 방법과 토픽 모델링 방법을 사용하여 문헌의 주제를 도출하고 범주화 하는 작업을 하였다. 이 분석 결과를 토대로 주제 별로 문헌에 대해 고찰하고 이를 기반으로 연구 의제를 제언하였다.

3. 데이터 분석과 결과

첫 번째 분석 방법으로는 연구자가 연구의 초록을 읽고 연구 분야 및 주제를 직접 코딩하는 질적 연구 방법을 사용하였다. 코딩 결과, 학문 분야 범주로서는 의료(예: healthcare, medical, nursing 등) 분야가 45개 논문으로 가장 많았으며, 교육(higher education, child education 등) 분야가 27개, 비즈니스 분야(marketing, hiring, corporate use 등)가 25개, 로보틱스(robots, robotics) 분야가 10개를 차지하였다. 이 외의 학문 분야들은 연구의 수가 10개 이하이거나 특정 학문 분야로 식별되기보다 다 학제적, 종합적 성격을 가지는 연구들이 대부분이었다.

다음으로 연구자가 연구들의 핵심 주제를 판별하여 세부 카테고리화하고 이들 세부 카테고리를 다시 몇 개의 요약 범주로 묶는 작업을 하였다. 그 결과 277개의 연구를 10개의 범주로 묶었다. <표 1>에 범주에 대한 설명과 각 범주 별 연구 수를 제시하였다.

두 번째 분석 방법으로는 텍스트 마이닝 기법인 Latent Dirichlet Allocation (LDA) 토픽 모델링(Topic Modeling) 방법을 이용하였다. 토픽 모델링은 텍스트의 주제를 탐색하는 여러 연구에 이용되고 있는(이소현, 김민수, & 김희웅, 2019; 홍태호, 니우한잉, 임강, & 박지영, 2018) 방법론으로서, 문서의 텍스트로 작성된 내용을 살펴 분석 대상이 되는 전체 문서들에 존재하는 주제의 수와 각 주제를 대표하는 단어를 도출한다(Bleiet al., 2003). 본 연구에서는 Python 기반의 LDA 토픽 모델링 라이브러리인 gensim을 이용하였다. 토픽 모델링을 위한 전 처리로 먼저 문헌의 초록을 토큰(token)화 하였으며, AI와 윤리 연구에 필연적으로 등장하거나, 주제를 식별하는데 의미가 있는 정보를 갖지 않는 단어들(예: AI, ethics, ethical, artificial intelligence, study, literature, research, analysis, paper, article, review, implications, data, findings, documents 등)을 식별하여 불용어(stopwords) 처리하였다. 또한 결과 해석의 명확화를 위하여 명사만을 사용하여 토픽 모델링을 진행하였다. 토픽 수를 결정하는 방법으로는 토

〈표 1〉 연구자의 코딩에 의한 연구들의 주제 범주

도출 범주	설명	연구 수
AI 윤리 원칙의 AI 디자인/개발에의 적용	AI 윤리 원칙이 실제 AI 디자인 및 개발에 어떻게 적용되어야 하는지에 관련된 이슈, 한계점, 이론과 현실의 괴리, 가이드라인 관련	51
AI 윤리 기본 원칙	AI 윤리 원칙의 제안 및 정교화	39
AI 윤리 원칙 심화 탐구	프라이버시, 책임, 편향, 설명 가능 AI(XAI), 지속 가능성, 투명성, 신뢰할 수 있는 AI, 의사결정 공정성 등에 초점	35
AI의 윤리적 도입 및 활용	AI 기술을 실제 도입하고 시행 시 관리, 규제, 전략, 법적 이슈	32
인간 AI 상호작용, 사용 윤리	AI 사용 시 윤리적 이슈, 가이드라인, 사용자 윤리	32
AI 윤리 리터러시	AI 윤리에 대한 교육 및 대중의 의식 증진	22
AI 영향과 가치	AI 가 가져올 사회적 가치, 영향의 윤리적 논의	14
AI 평가와 관리	AI의 윤리성을 평가하기 위한 감사, 사례 연구	11
기타	AI 윤리와 관련한 현상의 폭넓은 기술과 염려, 논의	41

픽의 일관성 지수인 coherence값을 이용하였다(Newman et al., 2010). coherence를 이용하는 방법은 최적의 토픽 수를 결정하기 위하여 보편적으로 쓰이는 방법으로 coherence 값이 증가하다가 갑자기 감소하는 지점을 적절한 토픽 수 값으로 사용한다. 분석 결과, 토픽 수가 4인 지점 이후 coherence 값이 급격히 감소하는 것을 확인하였으므로 토픽 수를 4개로 하여 모델링을 진행하였다. 분석 결과 도출된 키워드 모음과 이에 대해 해석한 주제명은 <표 2>에서 제시하였다.

토픽 모델링 결과와 연구자가 코딩한 결과를 비교 및 대응하면 연구자가 코딩한 결과가 토픽 모델링 결과보다 세부 카테고리로 나뉜 것을 확인할 수 있었다. 이들을 종합하여 <표 3>에서 보여지듯 통합 범주를 제시하였다. 연구자코딩의 "AI 윤리 기본 원칙"과 "AI 윤리 원칙 심화 탐구" 범주는 모두 AI가 지켜야 할 윤리 원칙에 대한 연구로 분류할 수 있었다. 연구자 코딩의 "AI 윤리 원칙의 AI 디자인/개발에의 적용'과 토픽 모델링 결과의 "윤리적 AI 디자인 및 개발"은 두 가지가 일 대 일로 대응되므로 "윤리적 AI 디자인 및 개발"로 분류하였다. 연구자 코딩 결과의 "AI의 윤리적 도입 및 활용", "AI의 영향과 가치", "AI 평가"는 토픽 모델링 결과의 "윤리적 AI 활용 관리 원칙"과 대응된다고 볼 수 있으며 조직 및 기관, 사회의 AI 도입과 활용에 관한 것이므로 "윤리적 AI 도입 및 활용"이라는 이름으로 통합하였다. 연구자 코딩의 "AI 윤리 리터러시"와 "인간 AI 상호작용, 사용 윤리", 토픽 모델링 결과의 "윤리적 AI와 사용자" 범주는 모두 개인 사용자의 AI 인지, 관련 교육, AI 사용에 관한 것이므로 "윤리적 AI 사용"이라는 범주로 통합하였다. 연구자의 코딩 결과에서 기타로 분류되는 범주는 하나의 범주로 나눌 만큼의 연구 수가 부족하거나(예: AI와 비인간 윤리), 현상에 대한 기술 등 폭넓게 일반적인 논의를 하는 연구들이므로 통합 범주에서도 기타 범주로 정리하였다.

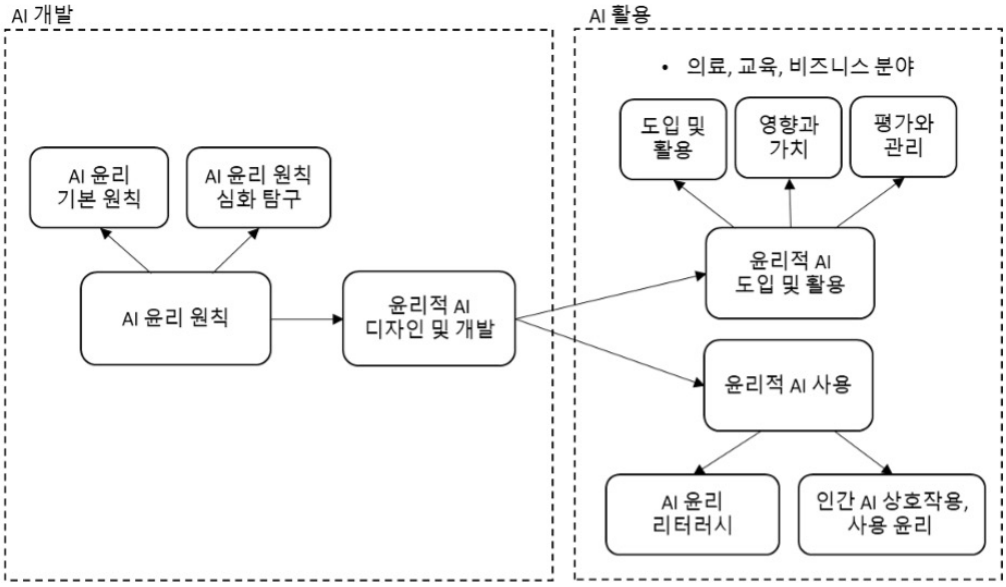
주제들의 관계를 연결하여 AI와 윤리 연구들을 종합하면 <그림 2>와 같이 도식화 될 수 있다. 최근 AI와 윤

〈표 2〉 토픽 모델링 결과에 의한 연구들의 주제 범주

주제(Topic)	주제 명(Topic Label)	키워드
1	윤리적 AI 활용 관리 원칙	principles, challenges, governance, applications, accountability, concerns, development
2	윤리적 AI와 사용자	results, framework, concerns, content, literacy, information, explainability
3	AI 윤리 원칙	development, trust, principles, fairness, transparency, society, practice
4	윤리적 AI 디자인 및 개발	principles, education, values, development, privacy, design, challenges

〈표 3〉 토픽 모델링과 연구자의 코딩 결과에 의한 문헌의 주제 범주

토픽 모델링 결과	연구자의 코딩 결과	통합 범주
AI 윤리 원칙	AI 윤리 기본 원칙	AI 윤리 원칙
	AI 윤리 원칙 심화 탐구	
윤리적 AI 디자인 및 개발	AI 윤리 원칙의 AI 디자인/개발에의 적용	윤리적 AI 디자인 및 개발
윤리적 AI 활용 관리 원칙	AI의 윤리적 도입 및 활용	윤리적 AI 도입 및 활용
	AI 영향과 가치	
	AI 평가와 관리	
윤리적 AI와 사용자	AI 윤리 리터러시	윤리적 AI 사용
	인간 AI 상호작용, 사용 윤리	
(토픽 없음)	기타	기타



〈그림 2〉 AI와 윤리 연구들의 연구 주제 분석 요약

리 관련 연구들을 종합하여 요약하면 주로 AI 개발의 기본이 되는 보편적 윤리 원칙을 기반으로 이러한 원칙을 지키기 위한 윤리적 AI 디자인 및 개발이 어떤 방식으로 되어야 할지에 대해 논의하고 있으며, 더 나아가 AI를 도입하고 윤리적으로 활용 및 관리하기 위한 가이드라인 및 전략과 일반 사용자가 AI를 윤리적으로 사용하게끔 하기 위한 가이드라인 및 교육을 여러 분야에서 논의하고 있다고 할 수 있다.

4. 문헌 고찰

문헌 고찰은 데이터 분석 결과에 기반하여 연구 주제별로 AI와 윤리에 대한 접근법을 살펴보려 하였다. 따라서 크게 1) AI 윤리 원칙, 2) 윤리적 AI 디자인 및 개발, 3) 조직 수준의 윤리적 AI 도입 및 활용, 4) 개인 수준의 윤리적 AI 사용과 관련하여 해당 주제의 연구들을 고찰하고 각 범주의 연구 경향에 대해 논하려 한다.

4.1. AI의 윤리 원칙(AI Ethics Principles)에 대한 가이드라인

AI의 윤리 원칙은 AI 개발 시에 지켜야할 보편적 윤리 원칙들을 말한다. 여러 윤리 원칙 중 AI와 밀접하게 관련되는 원칙들을 식별하고 정리한 대표적인 연구로는 Jobin et al. (2019)의 연구를 들 수 있다.[1] 윤리적 AI에 대한 나라별 가이드라인 84개를 대상으로 AI 윤리 원칙과 이를 위한 요구 사항이 전 세계적으로 수렴하는지를 살펴본 이 연구에서는 나라마다 다양한 원칙을 제시하고 있음에도 50% 이상의 문서에서 투명성(transparency), 정의(justice)와 공정성(fairness), 해악 금지(non-maleficence), 책임(responsibility), 프라이버시(privacy), 이렇게 다섯 가지가 보편적으로 AI 윤리 원칙이 되고 있다는 것을 발견하였다. 이들보다는 적은 빈도로 나타났지만 선행(beneficence), 자유와 자율(freedom and autonomy), 신뢰

1) 이 연구는 2019년 연구로 본 연구의 분석 대상이 되는 문헌 집합에 속하지는 않지만, 분석 대상이 된 AI 윤리 원칙을 다룬 다수 연구에서 이 연구를 토대로 윤리 원칙을 제시하였고, 또한 유사 개념을 비롯 포괄적인 윤리 원칙을 제시하므로 AI 윤리 원칙 주제 개괄을 위하여 제시하였다.

(trust), 지속가능성(sustainability), 존엄성(dignity), 연대(solidarity)가 발견되기도 하였다. 유사한 방법으로 진행된 Ryan and Stahl(2020)의 연구에서도 동일한 원칙을 식별하고 규정하였다. Jobin et al. 의 연구에서는 <표 4>에서 볼 수 있듯이 포괄적인 원칙 집합을 제시하였기 때문에 유사한 많은 연구들(Floridi & Cowls, 2022; Hagendorff, 2020; Mikalef et al., 2022)과 제시하는 기본 원칙에 있어 용어의 차이 정도만 있을 뿐 대부분 일치한다.

예를 들어 Hagendorff (2020)도 책무성(accountability), 프라이버시(privacy), 공정성(fairness) 관련 원칙을 분석 대상으로 삼은 가이드라인의 약 80%에서 발견하였다. 따라서 이 세가지를 윤리적 AI가 만족해야 하는 최소한의 조건으로 보기도 하였다. Floridi and Cowls(2022)는 AI 윤리 원칙 집합이 어느 정도 수렴된 상태라고 보고 있으며, 이 중 윤리적 AI를 위한 다섯 가지 핵심 원칙으로 혜택(beneficence), 해악 금지(non-maleficence), 자율(autonomy), 공정성(justice), 설명성(explicability)을 제안하였다. 이 연구에서 설명성은 투명성(transparency)과 책무성(accountability)을 포괄하는 개념이므로 Jobin et al. (2019) 연구에 대입하면 여섯 가지 원칙이라고 볼 수도 있겠다.

투명성은 특히 AI의 특징인 '헤아리기 어려움' 에서 기인하는 윤리 원칙이라고 할 수 있는데 "AI가 어떻게 작동하는가"라는 질문의 답과 관련된다. 즉, AI의 작동 방식 혹은 결과가 설명 가능하고, 이해 가능하고, 해석 가능해야 한다는 것을 이야기하고 이 설명력을 증가시키기 위한 노력을 해야 함을 뜻한다. 정의와 공정성은 AI가 연령, 성별, 장애 여부 등에 관계 없이 모든 사람에게 공정한 결과를 도출해야 하며 정의를 추구하고 공평한 분배 및 모든 종류의 차별을 제거해야 함을 말한다. 해악 금지는 기술적 혁신으로 인해 일어나는 위험으로부터 보호되어야 한다는 것을 뜻한다. 책임은 AI가 초래하는 결과를 책임져야 한다는 것을 말하는데 유사 개념으로 들어진 책무성(accountability)은 책임과 유사하면서도 다소 다른 면을 강조한다. 책임성이 책임을 질 수 있어야 한다는 것 자체를 강조한다면 책무성은 결과에 대한 세부적인 원인을 파악하여 책임 소재를 분명히 하는 것을 강조하는 것으로서 책무성을 실현하기 위해서는 AI 가 개발되는 과정과 이해관계자 등 관련된 모든 세부 내용을 명확히 하는 것이 필요하기 때문에 투명성과도 연결되는 개념이라고 할 수 있다(김효은, 2022). 이후 연구들에서 때로 책임을 포괄하는 개념으로 이야기되기도 한다. 프

〈표 4〉 AI 윤리 원칙(Jobin et al., 2019)

원칙	유사 개념 예
투명성(transparency)	설명성(explainability, explicability), 이해가능성(understandability), 해석가능성(interpretability)
정의와 공정성(justice and fairness)	일관성(consistency), 포용(inclusion), 형평성(equality), 공평(equity), 비편향성(non-bias), 비차별(non-discrimination), 다양성(diversity), 접근성(accessibility), 분배(distribution)
해악 금지(non-maleficence)	보안(security), 안전(safety), 보호(protection), 무결성(integrity)
책임(responsibility)	책무성(accountability), (법적)책임(liability)
프라이버시(privacy)	개인정보(personal information), 사적정보(private information)
선행(Beneficence)	혜택(benefits), 웰빙(well-being), 평화(peace), 사회적 선(social good), 공공선(common good)
자유와 자율(freedom and autonomy)	동의(consent), 선택(choice), 자기 결정권(self-determination), 자유(liberty), 권한(empowerment)
신뢰(trust)	
지속가능성(sustainability)	환경(자연)(environment (nature)), 에너지(energy), 자원(에너지)(resources (energy))
존엄성(dignity)	
연대(solidarity)	사회 보장(social security), 화합(cohesion)

라이버시는 AI가 광범위한 데이터 학습을 필요로 하기 때문에 이 과정에서 문제가 될 수 있는 부분을 짚는 윤리 원칙이다. 선행은 해악 금지와 유사하게 볼 수 있지만 해악 금지가 해를 끼치면 안 된다는 다소 소극적인 개념인 데 반해 적극적으로 긍정적인 행위를 해야한다는 것을 말하므로 차이가 있다. 인간 혹은 AI와 관계된 모든 생명체의 안녕 추구, 사회적 선의 추구를 위해야 한다는 것을 이야기한다. 자유와 자율은 AI가 '자율성'을 갖춘 기계로서 인간의 지시 없이 의사결정을 할 수 있게 됨에 따라 인간의 의사결정력과 AI의 의사결정력 사이에서 적절한 균형이 맞춰져서 인간의 의사 결정권, 자유 등이 침해 받아서는 안 된다는 것을 의미한다. 신뢰는 Jobin et al. (2019)의 연구에서는 하나의 윤리 원칙으로 보고 있지만 신뢰할 수 있는 AI란 궁극적으로 여러 중요 윤리 원칙들이 지켜지는 AI를 의미하므로 이후 연구들에서는 신뢰할 수 있는 AI라는 용어를 사용할 때 궁극적인 결과물로 사용하기도 한다. 지속가능성은 환경적 측면, 특히 에너지 측면에 있어 지속 가능한 AI가 되어야 한다는 것을 의미한다. 존엄성은 인간의 존엄성을 보장해야 한다는 원칙이며, 연대는 사회의 화합과 시민 권리 보장을 이야기하는 원칙인데 Floridi and Cowls(2022)에서는 선행에서 이 원칙들을 포함하였다.

이처럼 윤리적 AI를 개발하기 위해 지켜야 할 기본 윤리 원칙 집합은 어느 정도 논의와 합의가 된 상태이지만 이 원칙이 충분한지, 어떤 면이 부족한 지와 관련해서는 현재까지도 많은 논의가 계속되고 있다. 예를 들어 Hagendorff (2020)는 보살핌(care), 양성(nurture), 도움(help), 복지(welfare) 같은 '여성적 가치'와 사회적 책임(social responsibility)에 대한 논의와, 생태 네트워크(ecological network)와 같이 고립된 개별 시스템으로서가 아니라 더 넓은 범위의 네트워크 상에 존재하는 시스템 맥락 측면에 대한 논의가 부족하다고 하였다. 그 원인으로 현재의 AI 윤리 원칙에 관한 논의가 의무론적 윤리(deontology ethics)의 관점에 입각하여 윤리 원칙이 고정되어 있는 명제 집합으로 여겨지기 때문이라는 점을 들고 고정된 원칙 집합을 넘어 새로운 원칙들이 유연하게 제기되어야 할 필요성을 이야기하였다.

Siau and Wang (2020)은 8가지 AI 윤리 원칙을 제안하면서 각 윤리 원칙을 AI 기술의 특성에서 기인한 원칙, 인간의 사용에서 기인한 원칙, 사회적 영향과 관계된 원칙으로 분류하였다. AI 기술과 관계된 원칙으로는 투명성, 데이터 보안과 프라이버시, 자율성, 의도성(intentionality), 책임을 들었다. 즉 AI 기술이 블랙 박스의 특징을 갖고 있고 많은 데이터를 사용하며 자율화된 기술이기 때문에 이러한 특성들로 인한 결과가 비윤리적이 되지 않게끔 하기 위해 지켜야하는 원칙이라고 볼 수 있다. 인간의 사용에서 기인한 원칙으로는 책무성, 윤리적 기준, 인간의 권리에 관한 법, 사회적 영향과 관계된 원칙을 들었다. AI 사용의 결과에 대한 책임 소재를 명확히 하고 인간 사회에서 통용되는 윤리적 기준, 법적 기준, 사회적 기준을 명확히 해야 한다는 것이다. 사회적 영향과 관계된 원칙으로는 자동화와 일자리 대체, 접근성, 민주주의와 시민의 권리를 들었다. 즉, AI가 윤리적이 되기 위해서는 기술적 특징, 인간 사회에 대한 이해를 넘어 AI가 야기할 사회적 결과에 대해서 충분히 고려해야 하며 이것이 AI가 지켜야하는 윤리 원칙에도 반영되어야 한다는 것이다. 8가지 원칙 중 대부분은 기존의 원칙에서도 찾을 수 있으나 윤리적 기준이나 인간의 권리에 관한 법은 윤리나 법을 논할 때 인간이 이해하는 기준을 기계도 동일하게 이해할 수 있도록 해야 한다는 것을 의미하고 그러기 위해서는 인간이 먼저 윤리적, 법적 기준에 대해 제대로 학습해야 한다는 것을 전제하고 있다. AI가 지켜야 할 윤리적 원칙에 그 AI를 만드는 인간의 윤리적 기준이 영향을 미친다는 것을 고려한 원칙인 것이다. Munn (2023)은 AI 윤리 원칙이 비즈니스 시각에 치우쳐서 인종, 사회, 환경에 가져오는 해악을 의미 있는 방향으로 토론하는데 실패했고 그러므로 AI 윤리를 넘어 AI 정의 원칙(AI justice

principle), 즉 AI 기술의 정확성이나 감사보다는 AI로 인해 부정적인 상황에 놓이게 되는 경우에 더 초점을 맞추어야 한다고 이야기하여 과정보다 결과 중심의 원칙을 강조하기도 하였다. AI 윤리가 지나치게 기술적 요소에 초점을 맞춘 나머지 AI를 사용하는 맥락, 즉 비즈니스나 시장에 관하여는 무지한 경향이 있다는 연구(Häußermann & Lütge, 2022) 또한 비슷한 맥락의 연구라고 할 수 있겠다. 나아가 원칙들 간의 상대적 중요성에 관한 논의도 진행되고 있는데, 지역 혹은 나라마다 중시하는 AI 윤리 원칙에 차이가 있을 수 있으며 선진국에서는 프라이버시, 책무성, 투명성이 우선시되지만 개별 도상국에서는 문화적, 정치적, 경제적 요인이 먼저 고려되어야 한다는 주장도 있다(Huriye, 2023).

이 분야의 최근 연구 동향은 기본적인 AI 윤리 원칙 집합에 대한 합의를 기반으로 추가 윤리 원칙을 탐구하고 나아가 기존 윤리 원칙의 수정 필요성을 확인하기 위해 다양한 시각을 고려하며 다각화 되고 있는 것으로 보인다. 그러나 현재까지 추가로 제안된 사항들은 보편적 윤리 원칙이라기보다는 원칙과 더불어 고려해야 하는 경계 조건(boundary condition)이나 기존 윤리 원칙 범주 안에서 이해될 수도 있는 구체화 된 상세 원칙에 가까워 보인다. 이렇게 윤리 원칙의 경계 조건과 다각화가 진행되는 이유는 AI의 디자인 및 개발이 활발히 진행되면서 윤리 원칙의 현실 적용 또한 상당한 수준으로 시도된 것에서 찾을 수 있을 것이다. 원칙과 실제는 순환적 관계가 있는 만큼 실제 적용에서 도출된 피드백이 원칙에 영향을 미치는 단계에 있다고도 볼 수 있을 것이다.

4.2. 윤리적 AI 디자인 및 개발

AI 윤리 원칙은 궁극적으로 AI가 어떤 모습이어야 하느냐를 위한 것이다. 따라서 그 다음 단계에서는 그렇다면 AI의 디자인과 개발에 그 원칙을 어떻게 구현할 것인가에 대한 논의가 필연적으로 따르기 마련이다. 수집된 문헌에서 윤리적 AI 디자인 및 개발을 다룬 많은 연구들은 윤리적 AI 개발은 절대 명제이지만 원칙을 실제 기술로 구현하기가 어렵다는 것을 전제로 삼고 있으며, 그 이유와 개선 방법을 찾는데 초점을 두고 있다.

많은 연구들이 윤리원칙의 실제 구현이 어려운 이유를 윤리 원칙이 지나치게 추상적이고 상위 단의 개념이라는 데서 찾고 있다(Prem, 2023; Sanderson et al., 2023). 실제로 AI는 특정 맥락 하에서 개발되고 있는데 윤리 원칙들은 특정 맥락의 특수성이 고려되기보다는 모든 경우를 고려한 범용적인 원칙이라는 주장(Amugongo et al., 2023) 또한 같은 맥락이다. 이런 문제들의 해결을 위해 Morley et al. (2023)의 연구에서는 상위 수준의 개념으로 구성된 윤리 원칙을 시스템 요구 사항으로 이해될 수 있는 보다 구체화된 수준의 언어로 연결하는 체계를 작성하였다. 예를 들어 윤리 원칙의 설명성(explicability)을 추적가능성(traceability), 설명가능성(explainability)과 같은 시스템 요구사항으로 구체화시키고 추적가능성은 다시 “AI 시스템의 결정을 만들어내는 데 이용된 데이터와 프로세스는 문서화되어야만 한다”는 문장으로, 설명가능성은 “AI 시스템의 기술적 프로세스 및 연관된 인간의 의사결정을 설명할 수 있는 능력을 말한다”라는 문장으로 설명 및 규정하는 방식이다.

현재 AI 개발을 지원하는 도구들의 문제 또한 원인으로 논의되고 있다. 어떤 윤리 원칙의 경우에는 개발 도구나 개발 시 지켜야 할 체크리스트가 다른 원칙에 비해 상대적으로 잘 갖춰져 있지만(예: 책무성, 설명성, 프라이버시, 강건성 및 안전성(Hagendorff, 2020), 공정성, 책무성, 설명성, 투명성(Ayling & Chapman, 2022)) 대부분의 경우에는 지원 도구들과 윤리 원칙 사이에 괴리가 있다고 보고 있다(Wong et al., 2023). 예를 들어 특정 윤리 원칙의 해결책을제안하는 도구들이 존재하기는 하지만 제안된 해결책에 대한 기술적 설명력은 부족하여 개발자

들이 어려움을 겪고 있다는 것이다.

윤리적 AI 개발을 위해 윤리 원칙의 구현 문제에서 벗어나 다른 시각으로 접근하자는 연구들도 있다. 즉 윤리 원칙을 기술에 반영하는 것도 중요하지만 윤리 원칙의 기술적 구현을 넘어 다른 요소들 또한 고려해야 한다는 것이다. 예를 들어, 비즈니스 환경과 조직은 굉장히 복잡한 개체이므로 윤리 원칙 자체보다도 AI를 개발하는 회사가 어떠한 윤리적 환경과 문화를 가지고 있는가가 오히려 중요할 수 있다는 주장도 제기되었다(Lauer, 2021). Siau and Wang (2020)이 주장한 AI 개발자의 윤리적 기준과 법적 기준에 대한 학습을 중요하게 생각하는 원칙과 통한다고도 할 수 있다. 같은 맥락에서, AI와 관련된 다양한 이해관계자들의 시각이 좀 더 개발에 반영되어야 한다거나(Bélisle-Pipon et al., 2023), 기술적 강건성(technical robustness)이 고려되어야 한다거나(Choung et al., 2023), AI가 특정 가치를 내포하기 위해서는 AI가 포함되어 있는 사회적-기술적 시스템(sociotechnical system)을 이루는 사람, 기술적 인공물(artifact), 제도 혹은 기관 같은 여러 요소 중 특정 요소가 그 가치를 내포하도록 재 디자인하는 것이 상대적으로 더 쉬울 수도 있다는 연구(Van de Poel, 2020)도 있다. Van de Poel (2020)의 연구에서는 상황과 사례별로 다르기는 하지만 대게는 그 시스템을 유지하는 제도 혹은 기관이 기술적 인공물을 어떤 맥락에서 어떻게 다루는지를 규제하고 관리하기 때문에 제도 및 기관의 디자인을 논의하는 방향이 바람직하다고도 하였다.

이 분야의 연구들은 크게 윤리 원칙을 그대로 AI 개발에 적용하기 위해서 어떻게 해야 하는 가에 대한 연구와 적용하기 어렵다는 전제 하에 그렇다면 어떤 다른 시도를 할 수 있을 것인가에 대한 연구로 나누어볼 수 있을 것이다. 후자의 경우에는 윤리 원칙 분야의 연구들에서 원칙을 추가로 제안하거나 경계 조건을 제안하는 것과 유사한 시도이나, 다만 이 경우에는 기술적 내용에 좀 더 초점이 맞춰져 있다고도 볼 수 있다. 특히 최근 연구들은 AI 개발의 주체가 기관 및 사람이라는 것을 강조하며 윤리 원칙에 입각한 개발을 넘어 개발 주체를 식별하고 이들의 윤리성을 고려할 것을 제안하고 있다. 윤리적 AI란 기술에 의한 통제로 오롯이 달성할 수 있는 것이 아니라는 인식과 함께 개발자, 기관, 제도의 개입 및 인식이 강조되고 있다.

4.3. 윤리적 AI 도입 및 활용

윤리적 AI의 도입 및 활용은 해당 AI 시스템이 도입되어 활용되는 맥락과 떼어 생각할 수 없기 때문에 본 연구에서는 분석 대상이 되는 문헌에서 다수를 차지한 의료, 교육, 비즈니스 분야 연구들을 기반으로 윤리적 AI의 도입 및 활용을 이해하려 한다.

먼저 의료 분야는 조기 진단 및 진단의 정확성 개선에 AI의 도입이 도움이 된다는 것을 인지하고, 그렇다면 이러한 AI가 비윤리적으로 활용되지 않으려면 어떠한 점을 간과하지 말아야 하는지에 초점을 맞추고 있다. 특히 AI 진단 시스템에 오류가 생기거나 편향이 문제가 되는 경우 그 대상이 되는 환자의 건강에 치명적인 문제가 생길 수 있음을 경고하는 연구들이 주를 이룬다(Hunkenschroer & Luetge, 2022; Zhang & Zhang, 2023). 예를 들어, 의료 분야 AI는 민감 개인정보인 환자 정보와 건강 정보를 사용하여 개발되는 것이 보통이므로 AI 시스템 도입과 활용에 있어 환자의 프라이버시 문제를 보다 중요하게 여길 것을 촉구하고 있다(Masters, 2023; Wang et al., 2023). 또한, AI의 설명성 증진을 위한 방법으로 사전동의(informed consent)와 의학 기구의 인증과 승인을 받을 것을 제안하였다(Amann et al., 2020). 의료 시스템을 개발하고 도입하는 각 단계에서 윤리적 원칙을 고려하여 평가하는 방법을 제안하기도 하였다(Char et al., 2020). 의료 분야에서 ChatGPT 사용 시 잠재적 저작권법 위반, 의료-법적

복잡성, AI가 생성한 콘텐츠의 투명성 등의 문제가 제기되기도 하였다(Dave et al., 2023).

교육 분야에서도 유사한 논의가 진행되고 있는데, AI가 강의와 학습을 위한 도구 및 교육 지원 시스템으로서 유용할 수 있음을 전제하고, 그렇다면 AI가 이들에 어떤 영향을 미치는지 윤리적 시각에서 연구해야 할 필요성이 제기되고 있다(Schiff, 2022). 교육 분야에 도입되는 AI는 학습자의 학습과정과 능력을 모니터링하고 교정하기 위하여 도입하는 경우가 많으므로 투명성과 책무성이 특히 강조되며 취약 계층의 보호와 성별에 의한 편향 제거, AI를 학습시키는 데이터셋의 질 문제가 논의되고 있다(Coghlan et al., 2021; Slimi & Carballido, 2023). 지속가능성, 프라이버시, 안전, 포용, 인간 중심적 시각이 강조되고 있기도 하다(Nguyen et al., 2023).

비즈니스 분야에서는 조직에서 어떻게 AI를 도입하여 전략적으로 활용할 것인가, 이 과정에서 윤리적인 도입 및 활용이란 어떠한 것인가에 초점을 맞추고 있다. 비즈니스 실무(business practice)에서의 AI 윤리는 AI 시스템과 관련된 조직의 자원 사용, 디자인과 개발, 배치 등 AI 도입에 대한 정치적이고 경제적인 행위로 이해되어야만 한다(Attard-Frost et al., 2023). Attard-Frost et al.(2023)는 따라서 공정성, 책무성, 지속가능성, 투명성의 네 가지 윤리 원칙을 지향하기 위한 AI 비즈니스 활동을 연결하였는데 예를 들어, 공정성의 경우 오픈 이노베이션(open innovation), 시장 공정성(market fairness), 편향(지양)과 다양성(bias & diversity in professional practices) 등의 추구를 통해 달성될 수 있으며, 책무성을 추구하기 위해서는 AI 비즈니스 관행에 대한 대중의 인식, 내부적, 외부적 감독이 필요하다고 하였다. 지속가능성을 실현하기 위해서는 지속 가능한 개발 방식과 혜택 관리 및 분배에 보다 초점을 맞추어야 하며 투명성을 위해서는 의사 결정 설명 범위, 투명한 비즈니스 관행과 문화를 추구하여야 한다고도 이야기하였다.

비즈니스의 세부분야에서 도입과 활용에 관한 내용을 좀 더 살펴보면 마케팅과 채용에 활용되는 AI의 사례를 볼 수 있다. 마케팅 분야에서는 AI가 주로 시장 조사, 전략 수립 등을 위한 데이터 수집, 시장 분석과 개인화, 고객 관계 관리를 위해 도입되어 활용되고 있다. 이에 따라 소비자 데이터가 마케팅 AI의 학습 데이터로 사용되게 되고, 그 결과 소비가 증가하고 관련된 여러 윤리적 논란이 제기될 수 있음을 주지하였다(Davenport et al., 2020). 마케팅에서의 윤리적 AI 도입에 관한 연구로서 Hermann (2022)은 어떻게 윤리적인 사회적, 법적 관리 방식과 정책을 개념화하고 시행할 것인가에 관해 질문하였다. 여기에 대한 답으로 AI의 지능화와 인간화 수준이 올라감에 따라 다섯 가지 윤리 원칙 각각과 관련한 마케팅 노력이 어떻게 달라져야 하는지를 제안하였다. 예를 들어 자율 원칙의 경우에는 고객 자율과 자기 결정권에 대한 필요가 AI의 지능화 및 인간화 수준이 높아질수록 증가할 수 있다는 것을 인지하고 이러한 경우 인간 에이전시의 필요성 또한 증가시켜야 한다는 것을 강조하였다. 윤리적 행위에 대한 보다 세부적인 가이드라인을 제공하기도 하였는데, 예를 들어 해악 금지 원칙에 따라 중독적이고 충동적인 행위에 취약한 고객 그룹을 AI를 사용해서 타겟팅하는 것을 제안해야 한다고도 하였다. 같은 맥락에서 Du and Xie (2021)는 마케팅 AI 활용에 기업의 사회적 책임을 강조하기도 하였다.

채용 절차에 AI를 활용하는 경우에도 마찬가지로 AI 도입과 활용에 따른 위험을 줄이기 위해 어떠한 노력을 기울여야 하느냐에 대한 논의가 주를 이룬다. 일례로 Hunkenschroer and Luetge (2022)의 연구에서는 정부의 규제, 인간의 관리 감독을 통한 조직의 AI 채용 기준 설정, 기술 실사, 조직원들 사이의 인식 강화가 제안되기도 하였다.

윤리적 AI 도입 및 활용과 관련된 연구들은 더 나아가 개발된 AI가 윤리적인지를 평가하는 방법과 윤리적 관리

에 관하여도 논의하였다. 개발된 AI의 윤리성을 평가하는 방법으로는 광범위한 이해관계자가 AI 개발 및 구현 각 단계에 참여하여 AI가 미치는 영향, 위험 등에 대해 보고서, 체크리스트, 설문 등을 이용하여 데이터와 모델을 확인하는 방법을 사용하고는 하는데 주로 사전에 영향력을 예측하거나 사후 영향력을 평가하는 방식을 사용한다(Ayling & Chapman, 2022). Schultz and Seele(2023)는 나아가 AI의 윤리성에 대한 관리 방법으로 비즈니스 윤리를 적용하여 이해관계자 관리, 표준화된 보고 방식, 전사적 관리와 규제 등을 이용하여야 한다고도 하였다. 그러나 AI의 평가와 감사는 각 단계가 매우 복잡하고 특정 규제 체제나 법률이 부재하기 때문에 자발적 자율 규제가 주로 쓰이고 따라서 이러한 평가와 감사가 형식적인 도구로 전락할 위험도 경고하고 있다(Ayling & Chapman, 2022).

이 분야의 연구들은 윤리적 AI의 도입과 활용 주체가 조직 및 기관에 있으므로 도입하는 조직 및 기관이 어느 분야이냐에 따라 윤리의 우선 순위 및 초점에 차이가 있었다. 그러나 모든 분야가 AI가 가져오는 혜택만큼 윤리적인 문제가 있을 수 있다는 것을 주지하고 있는 것으로 보인다. 이에 따라 비즈니스 분야를 예로 들면 AI의 활용이 비윤리적이 되지 않게 하기 위해 인지해야 할 사항들, 세부적인 비즈니스 프로세스, 통제 절차, 평가 방법 등에 대해 논의하고 있다. 즉, 도입 및 활용 단계에 이르게 되면 윤리란 조직 및 기관의 사업 계획(actionplan) 과 필연적으로 연결될 수밖에 없을 것이다.

4.4. 윤리적 AI 사용

AI의 윤리적 사용과 관련한 연구들은 사용자가 AI를 윤리적으로 사용하게끔 하기 위해 필요한 논의 및 현실 인식으로 요약될 수 있다.

AI 사용이 일상 생활에서 보편화됨에 따라 AI의 윤리적 사용에 관한 사용자의 인식 변화가 필요한 시점이 되었다. 개인의 AI와 윤리에 대한 인식 강화는 윤리적 AI 개발과도 밀접하게 연결되어 있음은 당연한 사실이다. 그렇기 때문에 올바른 AI 개발을 위해서라도 AI에 대한 사용자 교육이 국가의 정책에 들어가 있어야 한다는 주장이 제기되고 있기도 하다(Schiff, 2022; Borenstein & Howard, 2021). 그 일환으로 AI 리터러시(literacy)에 대한 필요성이 제기되며, 사람들이 어떻게 AI를 이해하는가, AI가 변화시키는 사회에서는 어떠한 새로운 역량이 필요할 것인가에 대한 연구들이 진행되고 있다. 리터러시란 원래 글의 형태로 된 언어로 자신을 표현하고 소통하는 능력을 말하는데, Long and Magerko (2020)는 AI 리터러시를 개인이 AI 기술을 비판적으로 평가할 수 있게끔 하고, AI와 효과적으로 소통하고 협업하며, AI를 온라인 도구로서 집에서 직장에서 사용할 수 있는 역량 집합이라고 정의하고 이 역량 집합에 AI 와 관련 주요 윤리 이슈에 관해 식별하고 기술할 수 있는 역량을 추가하였다.

그런데 AI 리터러시 교육을 위해서는 사용자가 AI에 대해 가지고 있는 윤리적 기반을 먼저 이해할 필요가 있다. 따라서 AI 윤리에 대한 사용자 인식 연구들 또한 진행되었다. 이와 관련된 연구들은 인종, 전문 분야, 정치적 선호 등이 서로 다른 그룹의 사람들은 AI 윤리 관련해서도 각각 다르게 인식할 수 있기 때문에 다양한 그룹의 윤리적 이슈를 만족시키는 AI 개발을 위해 사용자 인식 연구가 필요하다는 것을 전제로 하고 있다. 일례로 Jakesch et al. (2022)의 연구에서는 전문가 그룹과 대중, AI에 의해 자동화된 노동에 종사하는 그룹은 AI 윤리에 대해 서로 다른 우선 순위를 가지고 있다는 것을 밝혔다. AI 전문가들이 프라이버시, 안전, 책무성, 투명성, 공정성에 우선순위를 높게 둔 반면 크라우드 플랫폼에서 일하고 있는 그룹과 일반 대중들은 안전, 프라이버시, 성과, 투명성, 책무성에 더 높은 우선순위를 두는 것으로 나타났다. 또한 AI와 사람을 비교할 경우 AI가 비윤리적일 경

우 사람보다 더 비윤리적이라고 느낀다는 연구도 있었다(Schelble et al., 2024).

AI 사용의 윤리적 문제에 관한 대표적인 예로는 학술 분야에서 AI를 사용하는 것이 어느 수준까지 윤리적으로 허용되는 것인지, 특히 공저자로도 허용되는지에 대한 논의일 것이다. 현재의 논의는 저자로서 가져야 하는 책임성과 능력 사이의 갈등으로 이해될 수 있다. 즉 책임의 측면에서 AI는 저자로서 한계를 가지지만 능력의 측면에서는 AI가 인간의 그것에 가까운 수준을 보인다는 것이다(Teixeirada Silva and Tsigaris, 2023). 따라서 Teixeira da Silvaand Tsigaris (2023)는 AI가 저자로서 등재되어야 하는 기준을 논의하여 명확히 할 필요성에 대해 이야기하고 있기도 하다. 더불어, 챗봇 사용과 관련한 윤리적 이슈에 대한 논의도 있는데, 예를 들어 Murtarelli et al. (2021)는 챗봇과 같이 인간과 상호작용하는 AI는 일반 사용자의 입장에서는 기술적 시각과 기술이 의인화(anthropomorphism)된 시각이 모두 개입된 사용 환경이라는 것을 강조하고, 이런 환경에서는 정보 비대칭, 챗봇을 인간과 동일시화하는 현상, 사용자 프라이버시 모두가 윤리와 관련된 문제가 될 수 있음을 이야기했다.

비윤리적 AI 사용의 결과에 관해서는 Schelble et al. (2024)이 팀 협업 과정에서 비윤리적인 AI 팀원과 같이 일하게 된다면 팀원에 대해 어떻게 느끼고 되고 이것이 팀 성과에 어떠한 영향을 주는 지와 관련된 연구를 통해 재미있는 결과를 보였다. 비윤리적인 AI 팀원과 협업한 경우 AI 팀원 뿐 아니라 전체 팀 신뢰까지 낮아진 것을 확인할 수 있었으며, 그 과정에서 현재 생성형 언어 모델이 사용하는 전략인 부정이나 사후 사과와 같은 전략은 신뢰를 회복하는데 도움이 되지 않은 것으로 밝혀졌다.

이 분야의 연구들은 AI 윤리 원칙의 개발에서 가장 멀리 떨어져 있는 단계를 다루기 때문에 보편적인 윤리 원칙을 다루기보다는 개별 사용 행태의 윤리성을 판단하거나, 비윤리적 사용 행태 식별, 윤리적 사용을 위한 윤리 교육과 사용자 인식을 강조하는 경향을 보인다. 본 연구에서 고찰한 연구 범주들 중 윤리적 AI 도입의 활용 분야와 함께 가장 다양한 소재에 대해 이야기할 수 있는 주제일 것이고, 또한 AI의 사용이 일상화 됨에 따라 앞으로 더욱 많이 연구될 분야로 보인다.

5. 정보시스템 분야의 향후 연구 방향

분석 결과를 토대로 정보시스템 분야 연구 의제를 도출하기 앞서 추가로 국내 정보시스템 분야 논문에서 AI 윤리 연구를 살펴보았다. 본 연구에서는 Google Scholar를 검색 도구로 사용하였으므로 영문 논문이 검색되었고, 또한 연구자의 주관을 보완하기 위해 사용한 토픽 모델링의 방법론적 특성 상 하나의 언어만을 대상으로 할 수밖에 없었다. 따라서 국내 논문은 연구 주제 범주를 나누기 위하여 이용하지 않고, 주제 도출 후 국내 경영정보학 분야 저널에서 “AI와 윤리”라는 키워드를 사용하여 별도로 검색하여 살펴보았다. 경영정보시스템 분야의 저널(예: 경영정보학연구, 지식경영연구, 지능정보연구)과 경영학 분야의 국내 대표 저널인 경영학연구까지 검색하였음에도 소수의 논문만이 검색되어 국내 정보시스템 분야에 AI와 윤리 관련 많은 연구의 여지가 있음을 확인할 수 있었다. 구체적으로 살펴보면 ChatGPT 사용 연구 윤리(손화철, 2023), 거대언어모델의 차별 문제 연구(이위 등, 2023; 최지애, 2023), 회계 분야에서 AI 사용으로 인한 윤리적 문제(윤소라, 2020)를 다룬 연구들이 있었다. 따라서 국내 정보시스템 분야 연구에서 다루고 있는 주제도 본 연구에서 분석한 주제 범주에서 벗어나지 않는 것으로 보인다.

정보시스템 분야 연구 의제를 도출하기 위하여 문헌 분석 대상이 된 논문 중 먼저 정보시스템 분야의 연구를 살펴보았다. 국내 연구와 마찬가지로 매우 드물었으

며 AI와 윤리 문헌 분석 연구 세 건(Ashok et al., 2022; Mikalef et al., 2022; Mirbabaie et al., 2022), ChatGPT 관련 윤리 논의 1건(Stahl & Eke, 2024), 책임성 있는 AI 혁신 생태계에 대한 함의를 제공하는 연구 1건(Stahl, 2022), 윤리적 AI 도입과 관리 방안에 대해 논의하는 연구 1건(Koniakou, 2023)이 검색되었다. 이는 정보시스템 분야에서 앞으로 AI와 윤리 관련하여 많은 논의가 필요함을 시사한다.

이 중 Mirbabaie et al. (2022)의 연구는 정보시스템 분야에서 AI와 윤리 관련 제언을 도출하기 위해 작성된 논문으로 본 연구와 그 목표가 유사하다. 그러나 이 연구에서는 해당 주제에 대한 정보시스템 분야의 담론을 설정하기 위하여 2020년 6월에 논문을 검색하였으므로 분석의 대상이 되는 연구들이 대부분 2020년 이전 연구들이고, 논문 심사를 거치면서 일부 추가 되었지만 2020년 이후 논문은 소수만 포함되어 있다. 본 연구는 2020년 이후 논문들을 대상으로 하였고, 이 시기에 거대 언어 모델을 사용한 AI가 등장하여 산업계와 학계뿐 아니라 일반인의 AI에 대한 관심이 급격히 증가하였다. 따라서 어떤 의미에서 본 연구는 Mirbabaie et al. (2022) 연구의 연장선이자 차별화된 지점을 다루는 연구라고 할 수도 있겠다. Mirbabaie et al. (2022)의 연구에서는 AI와 윤리 관련하여 근본이 되는 논문이라고 할 수 있는 정보시스템 분야 논문을 발견하지 못하였고, 그렇기 때문에 정보시스템 윤리와 AI 윤리 원칙을 연결시키는 시도를 하였다. 그 결과 대부분의 AI 윤리 원칙과 정보시스템 윤리 원칙이 일맥 상통하는 것을 보였다. 나아가 이 연구에서는 각각의 윤리 원칙 측면에서 가능한 연구 질문 예제를 제공하였는데 특정 해답을 도출하는 매우 구체적 수준의 연구 문제들로 유용한 의미를 갖는다. 그러나 윤리 원칙들이 서로 연결되어 있고, 윤리적 AI 관련 이슈들은 최근 들어 매우 복잡한 양상을 띠기 때문에 개별 윤리 원칙 마다 연구 질문을 정확히 분리해내기는 어렵다고 여겨진다. 따라서 본 연구는 정보시스템 분야에서 AI와 윤리 관련 문헌을 고찰한 몇 안 되는 논문들 중 하나로서 의미를 가지며, 그러한 논문들 중에서도 다양한 분야의 논문들을 망라하여 살펴보았다는 점, 또한 연구자 코딩을 사용하는 질적 연구 방법과 기계학습 기반 분석 방법인 토픽 모델링을 함께 사용하여 문헌의 연구 주제를 도출 및 범주화 하였다는 점에서 기존 연구들과 차별되는 시사점을 가진다. 이에 본 연구에서는 본 연구에서 분석한 연구 분야별로 정보시스템 분야 연구들의 향후 연구 방향을 제언하려 한다.

먼저 AI 윤리 원칙과 관련해서는 특정 기술과 그 사용 맥락과 관련한 경계 조건을 통해 AI 윤리 원칙을 정제하는 데 기여할 수 있을 것이다. 예를 들어 음식 배달 비즈니스에서는 디지털 플랫폼이 등장하면서 배달 노동자가 플랫폼과 계약을 맺는 개인 사업자 형태로 배차 AI의 지시에 따라 일하게 되었다. 이에 따라 AI 사용 기관인 플랫폼과 개인 노동자 사이의 공정성과 정의 문제, AI 알고리즘이 작동하는 방식이 모호함에 따라 노동자들이 느끼는 투명성 문제 등 다수의 윤리 관련 문제가 있을 수 있다. 디지털 전환 시대에 다양한 비즈니스가 다양한 방식으로 AI를 사용할 것이 예상되므로 기술을 기반으로 한 비즈니스와 AI 윤리 문제 또한 여러 가지 양상으로 나타날 것이며 따라서 중요시되는 윤리 원칙의 우선 순위도 다를 것이다. 정보시스템 분야는 이러한 연구들을 통해 윤리 원칙이 추상적 개념에 머무르지 않고 현실을 잘 반영하는 원칙이 되는 것에 기여할 수 있을 것이다.

다음으로 AI 디자인 및 개발 관련해서는 현재 연구 동향이 개발자에 의한 기술 통제가 아니라 개발자, 기관, 제도 및 다양한 이해 관계자들의 협업을 강조하고 있으므로 이해 관계자 별 다양한 윤리 지향점이 어떻게 합의되어야 할 지에 대해 연구하는 것이 의미 있을 것이다. 다양한 이해 관계자들은 AI에 대한 이해도, 기술과 윤리 및 사회에 대한 이해도가 모두 다를 것이므로 이 차이와

각각의 장단점을 모두 이해할 수 있는 시각이 필요할 것이며 이는 정보시스템 분야의 장점이기도 하다. 또한 AI 사용이 일상 생활에서 보편화됨에 따라 이해 관계자의 폭이 AI 서비스를 위해 일하는 노동자, AI 제품의 소비자까지 확대되어야 할 것이며 이들의 시각이 디자인에 반영될 수 있는 방법론 개발 또한 큰 의미를 갖는 연구 의제가 될 수 있을 것으로 보인다.

윤리적 AI 도입과 활용 관련해서는, 윤리적 도입과 활용을 지원하는 연구와 비윤리적 도입과 활용을 방지하는 연구의 두 가지 방향으로 연구를 진행할 수 있을 것이다. 모두 절차 및 결과 상의 장치를 마련하는 것과 관계될 수 있는데 특히 도입과 활용 시 평가 및 관리 방법에 있어서는 많은 연구 기회 및 논의의 여지가 있을 것으로 보인다. AI의 윤리적 평가 및 관리 방법은 기술적 절차뿐 아니라 사회적 영향을 반드시 고려해야 하기 때문에 역시 정보시스템 분야의 장점이 잘 발휘될 수 있는 분야일 것이다.

마지막으로 사용자 윤리 관련해서는 AI가 여러 다양한 사용자 층을 포용할 수 있도록 AI 윤리 방향을 이끌 수 있는 연구가 필요할 것이다. AI의 개발은 궁극적으로 인류에 혜택이 되기 위함일 텐데 기술의 한계 및 자본주의의 탐욕 때문에 오히려 소외되는 계층이 생겨나면 원취지에 어긋나는 것일 것이다. 그러므로 이 주제의 정보시스템 분야 연구들이 AI가 미치는 부정적 영향과 그 영향의 형태 등을 상세한 수준에서 기술한다면 문제점을 개선하고 더 나은 방향으로 나아가는데 기여할 수 있을 것이다.

현재까지 예로 든 연구들은 모두 정보시스템 분야의 강점인 기술, 인간, 사회, 비즈니스를 아우를 수 있는 종합적인 시각을 필요로 할 것이며, 이를 이용하여 탐구해야 하는 디지털 전환 시대의 연구 질문은 그 한계가 없다고 해도 과언이 아닐 것이다.

6. 연구의 한계

본 연구는 AI와 윤리에 관해 연구한 문헌 분석을 통해 AI와 윤리 연구들의 주제를 분석하고 연구 동향을 파악하며, 정보시스템 분야의 연구 의제를 제언하려 하였다. 그러나 본 연구의 제언은 다음과 같은 한계를 감안하여 이해되어야 한다. 첫째, 본 연구는 정보시스템 분야의 연구 의제를 도출하기 위하여 문헌 연구 방법을 사용하였다. 이 과정에서 연구자가 논문의 초록을 읽고 연구 주제를 코딩하는 질적 연구 방법을 사용하였다. 이 방법은 연구자의 주관이 들어갈 수 있다는 한계를 가진다. 그러나 토픽 모델링 방법을 사용하여 연구자의 주관을 보완하려 하였으며, 도출된 주제의 범주가 AI와 윤리를 포괄하기 때문에 연구자의 주관에서 기인하는 오류는 연구에서 제안하는 의제에 큰 영향을 미치지 않는 수준일 것으로 추측된다. 향후 연구에서는 이 점을 보완하기 위하여 다수 연구자가 코딩을 상호 검증하는 방법을 사용할 수 있을 것이다. 둘째, 연구의 대상이 되는 논문이 2020년 이후 논문만을 포함하고 있다는 한계를 가진다. 이는 AI 분야가 매우 빠르게 변화하는 분야인 만큼 최근 논문 동향을 파악하기 위한 선택이었고, AI의 기본 윤리 원칙이 2019년과 2020년 즈음 연구들에서 대체로 정리 및 정립되었기에 원칙 정립 이후의 연구를 대상으로 한다는 의미를 가진다. 그러나 후속 연구에서는 이 기간을 더 길게 하여 AI와 윤리 연구의 변화 추이까지 살핀다면 더 많은 시사점을 제공할 수 있을 것이다.

7. 결론 및 제언

윤리가 적용되는 대상을 주체와 객체로 구분하여 살펴본다면 AI가 주체가 되었을 때의 윤리적 쟁점은 과연 AI가 어떠한 방식으로 작동해야 하는가에 대한 것이 될 것

이고, AI를 객체로 놓고 보았을 때의 윤리적 쟁점은 AI를 어떻게 대해야 할 것인가, 즉 사용자 윤리로 구분하여 논할 수 있을 것이다. AI를 객체로 놓는 경우에 대해 더 정확히 이야기하면 세상의 윤리 규범이 지켜져야 할 대상에 AI 또한 포함시키는 것이므로, 분배의 문제, 평등의 문제 등에 AI를 포함시키는 것이 되지만 이는 강인공지능 혹은 범용 인공지능이라고 불리는 AGI(artificial general intelligence)가 현실이 되어 사회 구성원에 준하는 위치를 차지하였을 때를 위한 논의일 것이다(김효은, 2022). 현재의 AI는 좁은 인공지능이라 불리기도 하는 특정한 과업을 수행하는 ANI(artificial narrow intelligence) 로서의 위치가 더 강하므로 AI를 객체로 놓고 보는 경우에도 도구로서 논하는 것이 현재 시점에서는 더 타당할 것이다. 따라서 도구를 이용하는 사용자 윤리에 대한 논의가 이를 대신할 수 있을 것이다. 본 연구의 문헌 분석 결과 또한 윤리 주체로서의 AI에 대한 관점이라고 볼 수 있는 AI의 윤리 원칙과 디자인 및 개발 원칙에 대한 연구들이 주를 이루고 있고, 윤리 객체로서 AI에 대한 관점이라고 볼 수 있는 AI 사용자 윤리에 대한 연구들이 그 다음으로 주를 이루고 있다. 그러나 AI 윤리 원칙, 디자인 및 개발을 초점으로 하는 연구들이 대원칙에 덧붙여 다양한 세부 원칙을 추가로 도출하기도 하고, 문화나 상황에 따라 이러한 원칙들이 다르게 적용되어야 한다는 연구들도 다수 존재하는 것을 보면 AI의 디자인 및 개발에 보편타당하게 적용되어야 하는 총체적 원칙 집합은 현재에도 여전히 개발 중인 것으로 보인다. 따라서 다양한 분야에서 다양한 시각으로 원칙에 대해 논하는 연구들이 앞으로도 다수 생성될 것이라 예상된다. 특히 정보시스템 분야에서는 기술과 비즈니스, 사회에 대한 융합적 시각으로 여러 분야의 문제를 포괄할 수 있는 논의가 가능할 것이다. 디지털 전환 시대에 기술이 변화시키는 사회가 인류에 바람직한 모습이기 위해서는 AI가 윤리 주체로서 지켜야 할 원칙이 어떤 것인지 보다 활발하게 논하는 것이 필요할 것이다.

AI가 헤아리기 어렵다는 특성은 AI 기술의 작동이 블랙박스와 같기 때문이라는 것은 잘 알려진 사실이다. 재미있는 것은 비즈니스 분야에서는 이미 기술의 작동 메커니즘을 블랙박스로 놓고 그 입력과 결과, 즉 기술에 영향을 미치는 개인 및 조직과 사회의 특정 요인들과 그 기술로 인해 영향을 받는 사회 현상 및 변화를 연구하는 것에 이미 익숙하다는 것이다(Berente et al., 2021). 따라서 정보시스템 분야에서 “윤리적 AI는 어떠한 모습이 되어야 할 것인가”와 같은 질문에 답하기 위해 “AI의 디자인 및 기술 개발이 어떤 원칙을 지켜야 할 것인가”와 같은 기술 중심의 세부 질문으로 접근하기보다는 “사회가 요구하는 윤리적인 AI는 어떠한 모습이어야 하는가”와 같은 사회와 인류 중심의 세부 질문으로 접근하는 것이 더 적합하며 효과적일 것이다. 바꾸어 말하면 AI의 윤리성을 평가하는 사회적, 인류적 기준에 대한 연구들이 필요하다는 것이다. 전통적인 튜링 테스트(Turing test)는 만들어진 기술이 정말 인간 같으냐에 대한 테스트 방법으로 인간이 AI를 기계라는 것을 못 느낀 채 인간으로 인식한다는 것에 초점을 맞추지 AI가 어떻게 인간처럼 만들어졌느냐에 초점을 맞추지는 않는다. 즉, AI와 윤리 관련해서도 AI가 어떠한 윤리 원칙을 지켜 개발되었느냐가 아니라 AI가 얼마나 윤리적 결과 기준에 부합하게끔 작동하느냐가 초점인 AI 윤리 튜링 테스트가 만들어져야 할 것이다. AI 개발 과정을 제어할 수 있는 다양한 장치들이 오히려 AI의 비윤리적 작동에 대한 면죄부가 될 수 있다는 연구(Ayling & Chapman, 2022)는 결과에 대한 기준이 과정에 대한 기준보다 훨씬 중요할 수 있음을 시사한다. 물론 이것은 매우 어려운 문제로 자율주행차를 대상으로 한 윤리 기준을 개발하기 위한 윤리적 기계(moral machine) 테스트(Awad et al., 2018)가 오히려 기계의 행위에 일관적이고 보편적인 윤리가 적용되기 어렵다는 것을 보여주는 예가 되는 것을 보아도 알 수 있다. 본 연구의 분석 결과에서도 이러한 평가 기준을 연구한

예들은 매우 드문데 이것이 얼마나 어렵고 현재로서는 모호한 일인지 보여주는 반증이 될 수 있다. 그러나 AI가 얼마나 윤리적으로 작동하는지 평가할 수 있는 기준 혹은 테스트 체계의 설정은 AI의 활용이 보편화될 미래에 반드시 필요하며 따라서 앞으로 중요한 연구 담론 및 의제가 될 수 있을 것이다. 현재의 AI와 윤리와 관련한 개별 연구 활동은 궁극적으로는 “과연 윤리적 AI와 그 AI가 사용되는 사회의 모습은 어떠한 모습이어야 하는가”로 연결될 것이다. 즉, 현재의 많은 연구들은 Kuhn이 이야기하는 정상 과학(normal science) 활동으로서 이러한 궁극적 질문에 대한 퍼즐 풀이를 하고 있다고도 볼 수 있다. 본 연구에서 살펴본 바와 같이 다양한 분야에서 다양한 시각으로 AI와 윤리 문제에 접근하고 있는 현재의 추세는 이와 잘 부합한다고 말할 수 있다. 정보시스템 분야는 이미 다학제적 전통을 가지고 있는 분야로서 기술, 인간, 사회, 비즈니스를 융합하여 조망할 수 있는 강점을 가지고 있으므로 AI와 사회의 윤리 문제에 대한 바람직한 방향을 설정하는 데 크게 기여할 수 있을 것이다.

〈참고문헌〉

[국내 문헌]

1. 김효은. (2022). **인공지능과 윤리**. 커뮤니케이션북스.
2. 손화철. (2023). ChatGPT와 연구윤리. **지식경영연구, 24**(3), 1-15.
3. 윤소라. (2020). The impact of new technology on ethics in accounting: Opportunities, threats, and ethical concerns. **경영학연구, 49**(4), 983-1010.
4. 이소현, 김민수, 김희웅. (2019). 워라밸 이슈 비교 분석: 한국과 미국. **정보시스템연구, 28**(2), 153-179.
5. 이위, 황경화, 최지애, 권오병. (2023). 거대언어모델의 차별문제 비교 연구. **지능정보연구, 29**(3), 125-144.
6. 최지애. (2023). 거대언어모델(LLM)이 인식하는 공연예술의 차별 양상 분석: ChatGPT를 중심으로. **지능정보연구, 29**(3), 401-418.
7. 홍태호, 나우한잉, 임강, 박지영. (2018). LDA를 이용한 온라인 리뷰의 다중 토픽별 감성분석: TripAdvisor 사례를 중심으로. **정보시스템연구, 27**(1), 89-110.

[국외 문헌]

8. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Consortium, P. Q. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. ***BMC Medical Informatics and Decision Making, 20***, 1-9.
9. Amugongo, L. M., Kriebitz, A., Boch, A., & Lütge, C. (2023). Operationalising AI ethics through the agile software development lifecycle: A case study of AI-enabled mobile health applications. ***AI and Ethics***, 1-18.
10. Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. ***International Journal of Information Management, 62***, 102433.
11. Attard-Frost, B., De los Ríos, A., & Walters, D. R. (2023). The ethics of AI business practices: A review of 47 AI ethics guidelines. ***AI and Ethics, 3***(2), 389-406.
12. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J., & Rahwan, I. (2018). The moral machine experiment. ***Nature, 563***(7729), 59-64.
13. Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? ***AI and Ethics, 2***(3), 405-429.
14. Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. ***MIS Quarterly, 45***(1), 315-341.
15. BBC (2015). ***Google apologises for Photos app's racist blunder***. Retrieved from https://www.bbc.com/news/technology-33347866
16. Bélisle-Pipon, J. C., Monteferrante, E., Roy, M. C., & Couture, V. (2023). Artificial intelligence ethics has a black box problem. ***AI & Society***, 1-16.
17. Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. ***MIS Quarterly, 45***(3), 1433-1450.
18. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. ***The Journal of Machine Learning Research, 3***, 993-1022.
19. Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. ***AI and Ethics, 1***, 61-65.
20. Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying ethical considerations for machine learning healthcare applications. ***The American Journal of Bioethics, 20***(11), 7-17.
21. Choung, H., David, P., & Ross, A. (2023). Trust and ethics in AI. ***AI & Society, 38***(2), 733-745.
22. Coghlan, S., Miller, T., & Paterson, J. (2021). Good proctor or "big brother"? Ethics of online exam supervision technologies. ***Philosophy & Technology, 34***(4), 1581-1606.
23. Dave, T., Athaluri, S. A., & Singh, S. (2023). ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. ***Frontiers in Artificial Intelligence, 6***, 1169595.
24. Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. ***Journal of the Academy of Marketing Science, 48***, 24-42.

25. Du, S., & Xie, C. (2021). Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. *Journal of Business Research, 129*, 961-974.

26. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine Learning and the City: Applications in Architecture and Urban Design*, 535-545.

27. Häußermann, J. J., & Lütge, C. (2022). Community-in-the-loop: Towards pluralistic value creation in AI, or—why AI needs business ethics. *AI and Ethics*, 1-22.

28. Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99-120.

29. Heaven, W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*, Retrieved from https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

30. Hermann, E. (2022). Leveraging artificial intelligence in marketing for social good—An ethical perspective. *Journal of Business Ethics, 179*(1), 43-61.

31. Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics, 178*(4), 977-1007.

32. Huriye, A. Z. (2023). The ethics of artificial intelligence: Examining the ethical considerations surrounding the development and use of AI. *American Journal of Technology, 2*(1), 37-44.

33. Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How different groups prioritize ethical values for responsible AI. Paper presented at *the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

34. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399.

35. Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in Artificial Intelligence. *Information Systems Frontiers, 25*(1), 71-102.

36. Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics, 1*(1), 21-25.

37. Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. Paper presented at *the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

38. Masters, K. (2023). Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. *Medical Teacher, 45*(6), 574-584.

39. Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems, 31*(3), 257-268.

40. Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in information systems research. *Communications of the Association for Information Systems, 50*(1), 38.

41. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269.

42. Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: Barriers, enablers and next steps. *AI & Society*, 1-13.

43. Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics, 3*(3), 869-877.

44. Murtarelli, G., Gregory, A., & Romenti, S. (2021). A conversation-based perspective for shaping ethical human-machine interactions: The particular challenge of chatbots. *Journal of Business Research, 129*, 927-935.

45. Nedlund, E. (2019). Apple Card is accused of gender bias. Here's how that can happen. *CNN Businesss*, Retrieved from https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html

46. Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. Paper presented at *the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

47. Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B. P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies,*

28(4), 4221-4241.

48. Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. ***AI and Ethics, 3***(3), 699-716.
49. Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. ***Journal of Information, Communication and Ethics in Society, 19***(1), 61-86.
50. Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajikowicz, S., Robinson, C., & Hansen, D. (2023). AI ethics principles in practice: Perspectives of designers and developers. ***IEEE Transactions on Technology and Society, 4***(2), 171-187.
51. Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2024). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. ***Human Factors, 66***(4), 1037-1055.
52. Schiff, D. (2022). Education for AI, not AI for education: The role of education and ethics in national AI policy strategies. ***International Journal of Artificial Intelligence in Education, 32***(3), 527-563.
53. Schultz, M. D., & Seele, P. (2023). Towards AI ethics' institutionalization: Knowledge bridges from business ethics to advance organizational AI ethics. ***AI and Ethics, 3***(1), 99-111.
54. Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. ***Journal of Database Management, 31***(2), 74-87.
55. Slimi, Z., & Carballido, B. V. (2023). Navigating the ethical challenges of artificial intelligence in higher education: An analysis of seven global AI ethics policies. ***TEM Journal, 12***(2), 590-602.
56. Stahl, B. C. (2022). Responsible innovation ecosystems: Ethical implications of the application of the ecosystem concept to artificial intelligence. ***International Journal of Information Management, 62***, 102441.
57. Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT-Exploring the ethical issues of an emerging technology. ***International Journal of Information Management, 74***, 102700.
58. Teixeira da Silva, J. A., & Tsigaris, P. (2023). Humanrand AI-based authorship: Principles and ethics. ***Learned Publishing, 36***(3), 453-462.
59. Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. ***Minds and Machines, 30***(3), 385-409.
60. Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. ***Journal of Medical Internet Research, 25***, e48009.
61. Wong, R. Y., Madaio, M. A., & Merrill, N. (2023). Seeing like a toolkit: How toolkits envision the work of AI ethics. ***Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1)***, 1-27.
62. Zhang, J., & Zhang, Z. M. (2023). Ethics and governance of trustworthy medical artificial intelligence. ***BMC Medical Informatics and Decision Making, 23***(1), 7.

저 자 소 개

민 진 영 (Junyoung Min)

중앙대학교 경영경제대학 산업보안학과에서 부교수로 재직 중이며 KAIST 경영대학에서 경영정보시스템 전공으로 박사학위를 취득하였다. 주요 관심 분야는 Privacy, Algorithm Automation and Platform 등이다. Computers in Human Behavior, International Journal of Information Management, Journal of the Association for Information Science and Technology, Communication of the ACM, 경영학연구, 경영정보학연구 등에 논문을 게재한 바 있다.

〈 Abstract 〉

Navigating Ethical AI: A Comprehensive Analysis of Literature and Future Directions in Information Systems*

Jinyoung Min**

As the use of AI becomes a reality in many aspects of daily life, the opportunities and benefits it brings are being highlighted, while concerns about the ethical issues it may cause are also increasing. The field of information systems, which studies the impact of technology on business and society, must contribute to ensuring that AI has a positive influence on human society. To achieve this, it is necessary to explore the direction of research in the information systems field by examining various studies related to AI and ethics. For this purpose, this study collected literature from 2020 to the present and analyzed their research topics through researcher coding and topic modeling methods. The analysis results categorized research topics into AI ethics principles, ethical AI design and development, ethical AI deployment and application, and ethical AI use. After reviewing the literature in each category to grasp the current state of research, this study suggested future research directions for AI ethics in the field of information systems.

Key words: AI ethics, Ethical AI, AI design and development, AI deployment and application, AI use, Information systems

* This research was supported by the Chung-Ang University Research Grants in 2022

** Department of Industrial Security, Chung-Ang University