http://dx.doi.org/10.15801/je.1.135.202112.47

인공지능 윤리 연구에 관한 체계적 문헌고찰

A systematic literature review of research on artificial intelligence ethics

임 미 가(충남대학교 교육혁신본부 선임연구원)*

<목 차>

[국문 요약]

이 연구의 목적은 인공지능 윤리에 관한 국내 연구에 대하여 고찰하고, 앞으로의 인공지능 윤리 연구를 위한 유의미한 시사점을 얻는 것에 있었다. 이를 위해 체계적 문헌 고찰 방법으로 181편의 연구물을 분석하였다. 연구 분야별 비율은 법학 33.1%, 윤리학 22.6%, 교육학 13.4%, 정책학 9.9%, 철학 6.6%, 과학기술학 5.5%로 나타났다. 인공지능 윤리 연구 내용을 분야별로 범주화하여 살펴보면 철학적 측면의 연구들은 인공지능이라는 존재에 관심을 두고, 제도적 측면의 연구들은 인공지능의 규제 및 관리 방안을 모색하며, 교육적 측면의 연구들은 구성원들의 인공지능 윤리 의식의 함양에 관심을 두었다.

추후 인공지능 윤리 연구가 광범위한 영역에서 진행될 것과, 규범 윤리학으로써의 인공지능 윤리 연구에 더욱 관심을 기울일 것을 제언한다. 또한 추후 우리나라에서 인공지능의 윤리적 활용을 실현할 구체적 방안에 대한 후속 연구가 진행될 것을 제언

* mika@cnu.ac.kr

한다.

주제어 : 인공지능, 인공지능 윤리, 인공지능 교육, 인공지능 정책

Ⅰ. 서론

1. 연구의 필요성 및 목적

윤리(倫理)의 사전적 정의는 ‘사람으로서 마땅히 행하거나 지켜야 할 도리’이다(국립국어원, 2018). 우리는 단순한 사물에는 함부로 주체성을 부여하지 아니하므로 ‘컴퓨터 윤리’나 ‘핸드폰 윤리’라는 말은 사용하지 않는다. 그러나 ‘인공지능 윤리’라는 용어를 자연스럽게 사용하는 것을 보면 우리 스스로가 인공지능의 수준 높은 자율성과 주체성을 스스럼없이 인정하고 있는 듯하다. 4차 산업혁명으로 우리는 지능정보사회를 맞이하였고 지금 그 선두에 인공지능(Artificial Intelligence: AI)이 있다. 인공지능은 ‘이해와 합리적 적응 능력을 가진 가공의 어떤 것’인데(임미가, 2021), 2016년 사람 이세돌과 인공지능 알파고(Alphago)의 바둑 대결에서 알파고가 승리하면서 인공지능에 세계의 이목이 집중된 이유는 인공지능이 보여준 주체적 행위자로써의 능력 때문이었다. 주체적으로 행위할 수 있다는 것은 주체적으로 판단할 수 있다는 것이며 그러한 자유의지를 가진 존재는 그 행위에 윤리적, 법적 책임이 따른다. 대상의 자유의지 수준을 높게 지각할수록 우리는 그 대상에 더 큰 윤리적 책임을 요구하게 되는데(Fischer, 1994), 자유의지는 인간만의 고유한 능력이지만(Frankfurt, 1971), 인공지능을 의인화하여 인간처럼 해석한다면 인공지능에게도 자유의지가 있다고 지각될 수 있다. 이에 인공지능 기술의 발전과 함께 인공지능의 윤리적 측면에 관한 연구의 필요성이 대두되었다. 매사추세츠 공과대학(MIT), 옥스포드 대학 등 세계 유수의 대학에서 인공지능 윤리 연구에 대한 투자가 이루어지고 있고, Microsoft, Facebook와 같은 유명 IT 기업들 역시 내부적으로 인공지능 윤리 지침을 만들어 이것이 지켜질 수 있도록 하고 있다(안정용, 2021). 한편, 인공지능 기술이 예술, 금융, 국방, 노동시장 등 사회 각 영역에 큰 변화를 가져옴에 따라(임미가, 2020), 인공지능이 겪는 윤리적 딜레마의 수와 범위도 다양해지고 있다. 트롤리 문제(Trolley Problem)는 윤리학의 유명한 사고 실험으로 그 내용은 다음과 같다. ‘트롤리가 선로를 달려오고 있다. 선로 위에는 다섯 사람이 있고 트롤리가 그대로 달린다면 다섯 사람이 죽게 된다. 현재 당신은

선로 변환기 앞에 있다. 당신이 선로 변환기를 당기면 트롤리는 다른 선로로 이동하게 되고 그렇게 되면 다른 선로에 있는 한 사람이 죽게 된다. 당신은 어떻게 할 것인가?' 이 질문에 대한 답변은 사람 개인의 가치관에 따라 다를 것이고 그 답변의 옳고 그름을 명확히 판단할 수 있는 기준을 찾는 것도 어려운 일이다. 문제는 이 질문에 대한 답변자가 사람이 아닌 인공지능일 경우에 발생한다. 이와 같은 선택의 문제가 자율주행 자동차의 사고 발생 과정에서 운전자를 희생시킬지 타인을 희생시킬지에 대한 질문으로 이어진다면 인공지능은 어떠한 판단을 내리는 것이 올바른 것인가? 이와 같이 자율주행 자동차가 누구의 생명을 우선으로 보호해야 하는지의 문제, 안면 인식 등의 생체 인식 기술에서 인공지능이 가지고 있는 기초 정보에 따라 생체 인식 정확도가 다르게 나타나는 문제, 이와 비슷한 원리로 성(性)적으로 편향된 결과를 드러내는 성적 차별의 문제, 인공지능이 사람과 커뮤니케이션 하는 과정에서 얻게 된 사용자의 개인 정보에 관한 문제 등 인공지능이 갖추어야 하는 윤리적 판단 기준의 범위는 광범위하다. 더 큰 문제는 인공지능의 활동 중 사람이 미처 예상하지 못한 윤리적 문제가 발생했을 때 인공지능의 자율적 판단과 대처가 어떠한 결과를 가져올지 명확히 알 수 없다는 부분이다. 그리하여 연구자들은 인공지능의 존재론적 지위와 특징(최경석, 2020), 도덕적 행위자로서의 인격(박형빈, 2020), 인공지능과 책임의 문제(이봉재, 2006) 등 인공지능의 도덕적 책임자로서의 역할과 특징에 관하여 연구하고 이를 기반으로 인공지능 규제 등의 윤리적 쟁점과 법적 과제들에 대하여 연구하려 노력해왔다. 그러나 약 5년여의 비교적 짧은 기간 안에 인공지능 기술이 급속도로 발전해 온 점을 생각해보았을 때 인공지능 윤리 연구가 그에 발맞추어 진행되어 왔는지에 대한 반추와 앞으로의 진행 방향에 대한 탐색이 필요하다. 그러므로 이 연구에서는 인공지능 윤리에 관한 국내 연구에 대하여 고찰하고, 앞으로의 인공지능 윤리 연구를 위한 유의미한 시사점을 얻고자 하였다.

2. 연구 내용

이 연구의 목적을 달성하기 위한 연구 내용은 다음과 같다.
첫째, 인공지능 윤리의 국내 연구 경향 및 내용을 고찰한다.
둘째, 국내 인공지능 윤리 연구를 위한 시사점을 제시한다.

Ⅱ. 이론적 배경

1. 인공지능과 사회

인공지능은 1940년대 뇌와 뉴런에 관한 연구로부터 그 개념이 등장하였고, 1950년대 앨런 튜링(Alan Turing)이 지능을 가진 기계의 개념을 언급하면서 본격적으로 다루어지기 시작하였다. 존 매카시(John McCarthy)는 다트머스(Dartmouth) 대학의 한 연설에서 인공지능에 대해 구체적으로 발표하였는데, 그 이후 인공지능 연구가 빠른 발전을 이루는 듯 보였으나, 환경 변화와 재정적 문제들로 인한 침체기를 맞이하기도 하였다. 이후 1990년대 중반, 머신러닝과 빅데이터 등의 기술이 개발되고, 2016년에는 세기의 대결인 알파고(Alphago)와 이세돌의 바둑 대결에서 인공지능 알파고가 승리하면서 그 이후 본격적인 인공지능 시대가 시작되었다. 인공지능(Artificial Intelligence: AI)은 인간의 인지, 추론, 학습을 컴퓨터를 이용하여 실행하고 여러 문제를 해결하는 기술이다. 인공지능은 딥러닝, 빅데이터, 머신러닝 기술들을 기반으로 공간과 시간을 초월하여 학습하고 사고해나가는 확장성을 보인다. Rusell & Norvig(2016)은 인간적 사고, 합리적 사고, 인간적 행위, 합리적 행위라는 4가지의 범주로 인공지능을 개념화하였고 이때 인공지능은 인간적 사고나 인간적 행위에 가까울수록 강한 인공지능이라는 평가를 받게 된다. ‘합리성’이라는 단어의 의미를 풀어 쓰면 ‘논리와 이성의 적합성’이 되는데, 임미가(2020)는 이 ‘합리성’이 항상 합리적인 사고와 행위를 하기가 어려운 인간과 대비되는 인공지능만의 특별한 능력이라고 하였다. 인공지능 기술은 센서를 이용하여 정보를 수집하고, 정보에 기반하여 지식을 축적하고, 축적된 지식을 바탕으로 새로운 지식을 유도하여 목표 상태에 도달하게 된다. 또한 이 과정에서 발생한 경험을 통하여 스스로 시스템을 수정하고 보완하는 학습 능력을 통해 문제를 해결하고 능력을 발전시켜 나간다. 앞으로 인공지능은 우리 주변의 어디에나 존재하고, 필수적이며, 피하기 어려운 우리 생활의 일부가 될 것이다. 김영식(2019)은 인공지능 분야에서 기술 역량의 증가로 인해 새로운 가치가 창출되고 있고 그것은 수많은 기회를 제공하고 있으며 이러한 인공지능 기술을 효과적으로 활용하기 위해 경제, 사회에 미치는 영향을 이해하는 것이 필수적이라고 강조하였다.

2. 인공지능 윤리

우리가 인공지능을 온전히 이해하기도 전에 인공지능이 빠르게 우리에 생활 속에 다가오면서 인공지능 윤리의 문제 또한 우리가 급박하게 해결해야 할 숙제가 되었다. 인공지능 윤리에 대한 고민 없는 인공지능과의 상호작용은 기술철학자 자크 엘륄(Jacques Ellul)의 염려와 같이 인간이 기술 사회에서 주체로서의 지위를 유지할 수 있는가의 문제에 대해 부정적 결과를 초래할 수 있다. 인공지능의 역할은 편리성, 정확성과 신속성의 측면에서 긍정적으로 평가되지만 반면 부정적인 평가도 많다. 인공지능의 활용도가 높아질수록 인간이 노동에서 소외되고 이로 인해 삶의 질이 현저히 떨어지며 결과적으로 인간소외 현상이 대두될 것이라는 '인간소외'의 측면, 소수의 자본가와 첨단 과학 기술 관련자들에게 부와 권력이 집중될 것이라는 '사회적 불평등' 측면, 인공지능이 인간의 지성과 동일하거나 그 수준을 뛰어넘었을 때에는 인공지능이 인간의 자율성과 자유를 제한하고 인간의 행위를 지배할 수도 있다는 '인간 존폐 위기의 불안감' 측면 등이 그것이다(하영숙, 2019). 세부적으로는 러시아가 사람과 물체를 추적해 공격하기 위해 개발한 킬러 로봇의 등장, 이스라엘이 개발한 인간의 조작 없이 총기 공격이 가능한 킬러로봇 'Dog'의 등장, 아마존(Amazon)이 개발한 인공지능 채용 시스템이 여성보다 남성 지원자를 선호하는 패턴을 보였던 성차별 문제의 등장, 인공지능 안면 인식 기술이 백인 남성의 경우 단 1%의 오류를 보인 반면, 피부가 검은 여성의 경우 35%의 오류를 보였던 인종 차별 문제의 등장 등이다(안정용, 2021). 그리하여 사람들은 인공지능 윤리 문제는 먼 미래의 일이 아닌 당장 해결해야 할 일이라는 불안감을 가지게 되었고 그리하여 인공지능의 도덕성 확보를 위한 연구와 제언들이 등장하였다.

인공지능, 철학, 윤리, 법 등의 관련 분야 전문가 100여 명에 의하여 작성된 전기전자학회(IEEE)의 'Ethically Aligned Design'에 따르면 사람들이 인공지능에 대해 두려움을 느끼고 있으며 이를 해소하기 위하여 인공지능 개발 과정에서 인권, 웰빙, 데이터관리, 효과성, 투명성, 책임감, 오용인식 그리고 유용성에 관한 원칙이 고려되어야 한다고 하였다(Shahriari & Shahriari, 2017). IEEE는 인간 중심(human-centric)으로 사고하는 인공지능을 강조하며, 이런 원칙을 어기는 인공지능이 등장할 시 인공지능 상용화는 인류에게 악재가 될 것이라 경고했다(안정용, 2021). 인공지능 개발 분야의 선두 국가인 미국 정부는 국가과학기술위원회(National Science and Technology Council, NSTC)를 통해 '인공지능의 연구개발 전략 계획'을 7단계로 구분한 보고서(National Artificial Intelligence

Research and Development Strategic Plan)를 발표하였다. 7단계 전략의 내용은 인공지능에 대한 장기적 투자, 인간과 인공지능 간의 소통과 이해, 인공지능에 대한 윤리적·법적·사회적 의미와 해결, 인공지능의 안전성, 인공지능의 데이터 공유화, 인공지능 기술 표준성, 국가적 인공지능 개발 인력 파악 등이다(하영숙, 2019). 또한 유럽연합(EU)은 2019년에 정부와 기업이 인공지능을 개발할 때 지켜야 할 7가지 윤리 지침을 소개했다. 이 보고서는 신뢰할 수 있는 인공지능을 윤리적 인공지능으로 정의하고 이를 위해 인간에 의한 통제, 안정성, 개인 정보보호, 투명성, 차별금지, 웰빙 추구, 책무성을 준수하는 인공지능 개발을 독려하였다(HLEG, 2019). 우리는 앞으로 '관계지향적 존재'로서의 인공지능을 마주해야 할 것이며(김광연, 2018), 인공지능 윤리는 인간과 인공지능의 상생과 협업을 위해 반드시 선결되어야 할 과제가 될것이다(안정용, 2021).

Ⅲ. 연구 방법

1. 체계적 문헌 고찰(Systematic Reviews)

체계적 문헌 고찰은 특정 연구 질문에 대한 답을 도출하기 위해 구체적인 적격 기준에 맞는 실증적 근거들을 종합하는 연구 방법이다(Eden et al., 2011). 체계적 문헌 고찰은 주로 보건 및 의료 분야 연구에서 사용되어 왔는데, 과학적 근거 중심 의학의 핵심적 연구 방법이기 때문이다. 하지만 체계적 문헌 고찰 방법이 문헌 고찰을 객관적으로 수행할 수 있게 된다는 장점 때문에 현재는 사회학, 교육학, 체육학 등 여러 분야에서 활용되고 있다. 체계적 문헌 고찰의 수행 과정은 Australian National Health and Medical Research Council(2000), Khan et al.(2003), Kitchenham(2004) 등에 의하여 연구되어왔고, 이 연구들에서 제시한 과정들은 큰 틀에서는 유사한 범주를 보이지만 연구 목적에 따라 세부적인 차이가 있다(임미가, 2021). Khan et al.(2003)은 보편적 검토 방법론으로서 체계적 문헌 고찰의 과정을 제시하였으므로, 본 연구에서는 Khan et al.(2003)이 제시한 '체계적 문헌 고찰 수행의 5단계'를 기준으로 하여 문헌 고찰을 진행하고자 한다. 그리하여 이 연구에서의 체계적 문헌 고찰은 '질문 구성, 문헌 검색, 문헌 선택, 자료 추출 및 종합, 결과 제시'의 다섯 단계로 진행되었다.

2. 연구 절차 및 내용

체계적 문헌 고찰 방법에 따른 구체적 연구 절차와 내용은 다음과 같다.

1) 질문 구성

질문은 구조화되고 명확한 질문으로 구성되어야 하고 분명하고 객관적인 이유 없이 질문이 수정되어서는 안된다. 이 연구에서의 구체적인 연구 질문은 다음과 같다.

첫째, 인공지능 윤리 국내 연구의 연도별 출판 빈도는 어떠한가?
둘째, 인공지능 윤리 국내 연구의 분야 영역은 어떻게 분류되는가?
셋째, 인공지능 윤리 국내 연구의 분야별 내용은 무엇인가?

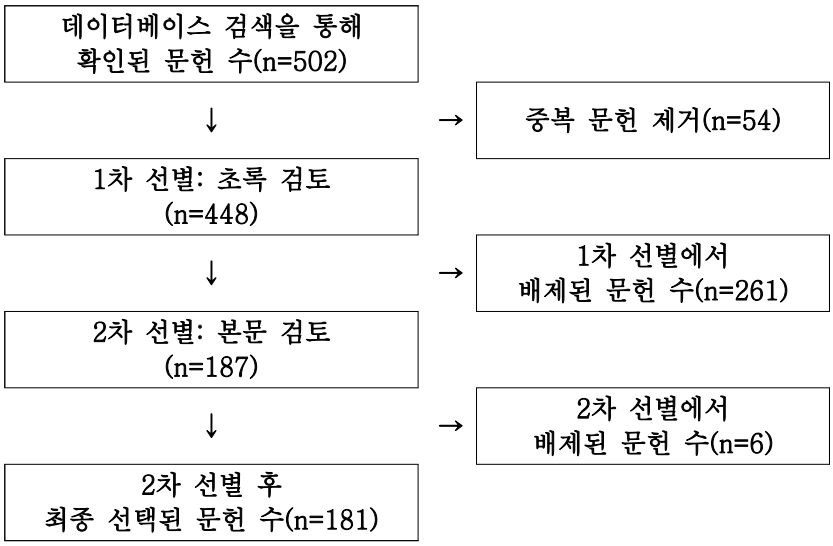
2) 문헌 검색

문헌 검색의 단계에서는 검색 데이터베이스를 설정한 후 검색의 범위와 검색어를 결정하여 문헌 검색을 시행한다. 이때의 기본 방향은 연구에 관한 검색이 광범위하게 이루어지는 것이다. 이 연구에서 문헌 검색을 위해 사용한 데이터베이스는 KCI, 스콜라, e-article, DBPIA, Kiss, KERIS dCollection 등 대부분의 국내 전자자료와 학술적 가치가 높은 무료 컨텐츠(Open Access)를 한 번에 검색하고 활용할 수 있는 검색 도구인 한국교원대학교 도서관의 Discovery(전자정보통합검색) Web DB를 사용하였다. 문헌의 질적 측면을 고려해 학술지 게재 문헌과 학위논문을 대상으로 하고, 문헌의 양적 확보를 위해 문헌 검색 기간과 분야 분야를 '전체'로 설정했다. 이 연구에서 사용된 검색어는 '인공지능 윤리'이고, 검색 범주는 '모든 텍스트'이다. 검색 결과 총 502개의 문헌이 검색되었다.

3) 문헌 선택

문헌 선택 과정에서 문헌은 적합성과 이질성을 고려한 세부적인 기준에 의해서 평가되어야 한다. 그러므로 구체화된 배제 기준을 바탕으로 연구 목적에 부합하는 문헌들을 선정해야 한다. 이 연구에서 502편의 문헌을 대상으로 하여 적합 문헌을 선택하는 과정은 [그림 1]의 플로우 차트와 같다. 문헌 검색 결과로 선정된 502편의 문헌 중 여러 데이터베이스에서 중복적으로 검색된 문헌 54편이 제외되

어 448편의 문헌이 선택 대상이 되었다. 초록 검토를 통한 1차 선별 과정에서 인공지능 윤리와 직접 연관이 없는 문헌 261편을 제외하였다. 이후 본문 검토를 통한 2차 선별 과정에서 연구의 질문과 관련이 없는 문헌 6편을 제외하여 최종 고찰 대상으로 181편의 문헌을 선정하였다.



[그림 1] 문헌 선택 과정

1차, 2차 선별 과정에 적용된 문헌 배제 기준은 <표 1>과 같다.

<표 1> 문헌 배제 기준

배제 기준 내용
교육 방법론으로써 인공지능을 활용하는 연구
인공지능 윤리와 무관한 분야의 연구
인공지능 윤리와 핵심적 연관이 없는 타 영역 중점의 연구
인공지능 윤리와 관련된 단순 서평이나 서언
인공지능 윤리가 아닌 단순 인공지능 관련 연구
포괄적 개념으로써의 4차 산업혁명 관련 연구

4) 자료 추출 및 종합

자료 추출 및 종합의 단계에서는 선택 문헌에 관한 결과 지표들을 정의하여 제시한다. 추출 자료들은 정성적(qualitative) 또는 정량적(quantitative)으로 합성할 수 있는데, 어떤 방법으로 데이터를 분석할지를 결정하는 기준은 통계적 합성 가능성 여부이다. 체계적 문헌 고찰 대상 문헌의 연구 결과들의 성격이 이질적인 경우에는 정성적 합성을 수행하여 연구 결과들을 기술적(descriptive)으로 제시한다. 이 연구는 특정 독립변수와 종속변수를 정의하지 않고 인공지능 윤리 연구 전반을 대상으로 하므로 문헌의 결과 변수들이 이질적이다. 그러므로 정성적 합성을 수행하여 기술함이 적절하다. 그리하여 제시될 결과 지표 중에서 문헌의 일반적 정보에 관한 내용은 정량적으로 제시하고, 상세 내용 분석은 그 고찰 결과를 기술적(descriptive)으로 제시하였다.

5) 결과 제시

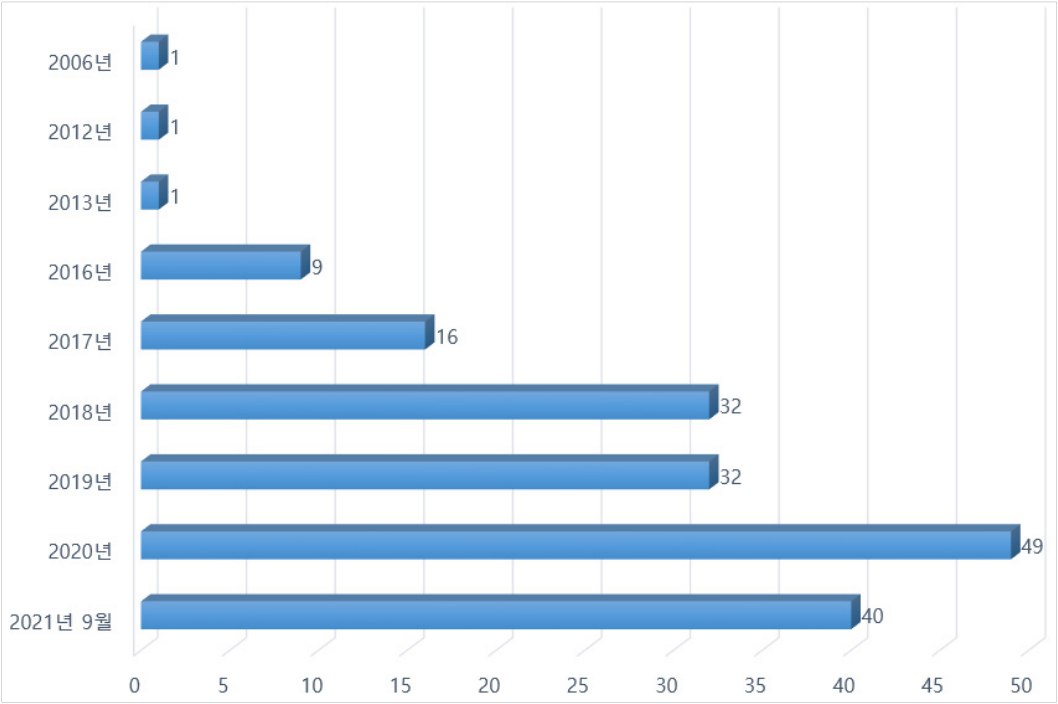
체계적 문헌고찰의 결과에 대하여 연도별 빈도 및 비율을 제시하고, 분야 영역별 범주, 주요 연구 내용을 서술하였다.

Ⅳ. 연구 결과

1. 인공지능 윤리에 관한 국내 연구 추이

1) 연도별 출판 빈도

인공지능 윤리의 시기별 추이를 살펴보기 위하여 연도별 출판 빈도를 [그림 2]와 같이 나타내었다.



[그림 2] 연도별 출판 빈도

인공지능 윤리에 관한 국내 연구는 2006년에 시작되었으나 약 10여 년간 진전을 이루지 못하고 있다가 2016년 9편, 2017년 16편, 2018년과 2019년 각 32편, 2020년 49편, 2021년 9월 현재 40편으로 꾸준히 증가하는 추세임을 알 수 있었다.

2) 연구 분야 영역

인공지능은 오늘날 사회 각 분야에 광범위하게 영향을 미치고 있으므로 인공지능 윤리에 관한 연구도 다양한 학문 분야에서 연구되고 있었다. 문헌이 주로 어떤 분야를 중점으로 연구되었는지를 기준으로 하여 <표 2>와 같이 연구 분야를 분류하였다.

<표 2> 연구 분야

구분	세부분야	빈도(개)	비율(%)
법학	국가법, 산업법, 젠더법, 헌법학, 해사법, 행정법	60	33.1
윤리학	사회윤리, 생명윤리, 인터넷윤리, 정보윤리	40	22.6
교육학	과학교육, 도덕교육, 윤리교육, 뇌기반교육, 에너지 기후변화교육, 인공지능교육학, 초등교육	25	13.4
정책학	과학기술정책학, 법정책학, 인터넷자율정책, 자치행정학, 정보통신정책학, 지방자치학	18	9.9
철학	과학철학, 법철학, 양명학, 유교학, 정보철학, 환경철학	12	6.6
과학기술학	기계학습, 위험요인, 비도덕문장판별	10	5.5
의학	의료윤리, 로봇의사, 헬스데이터	5	2.8
공학	자율시스템, 공학전문가 인식, 프로그래밍	4	2.2
융합연구	영화, 소설 속 인공지능	2	1.1
군사학	무기	1	0.6
미래학	미래사회전략	1	0.6
심리학	자유의지	1	0.6
언론학	로봇저널리즘	1	0.6
종교학	불교학	1	0.6
합계		181	100

본 연구에서는 특정 철학 분야 또는 철학가의 사상에 근거한 연구는 철학 분야로 분류하고, 직접적으로 윤리적 측면에 관해 연구된 것은 윤리학 분야로 분류하였다. 생물과 관련된 인공지능 윤리 중 의료에 연관된 생명 윤리는 의학 분야로 분류하였다. 과학기술학은 과학과 기술에 대하여 인문학 및 사회과학의 방법을 따르는 탐구를 수행하는 학제 간 연구 분야로써 넓게 보면 과학사, 과학기술사회학, 과학기술 정책, 과학기술과 법, 기술사 및 기술철학 등을 포괄하는 용어이지만, 맥락에 따라 그 범위나 의미가 달라지는 것이 일반적이다. 본 연구에서는 '위험요인 구명, 문장 판별'과 같이 법, 철학, 윤리학 등 특정 학문적 성격과 연관이 깊지 않은 분야의 연구를 과학기술학 분야로 분류하였다. 연구 분야별 빈도를 살펴본 결과 법학이 33.1%로 가장 많은 비율을 차지하였으며, 윤리학 22.6%, 교육학 13.4%, 정책학 9.9%, 철학 6.6%, 과학기술학 5.5% 등으로 나타났다.

2. 인공지능 윤리의 주요 연구 내용 고찰

인공지능 윤리의 주요 연구 내용을 분야별로 범주화하여 살펴보았다. 윤리학과 철학 분야의 연구를 철학적 측면으로 범주화하고, 법학과 정책학 분야의 연구를 제도적 측면으로 범주화하였다. 이외 교육적 측면, 의학적 측면, 융합적 측면으로 구분하고 각 분야 범주에서의 인공지능 윤리 연구의 주요 내용을 살펴보았다.

1) 철학적 측면

인공지능 윤리에 관한 철학적 측면의 연구들은 인공지능이라는 '존재'를 어떻게 바라보고 어떠한 존재로 '인정'해야 하는가에 대한 해답을 찾고자 노력한다. 이 범주에서 이루어지는 연구들은 인공지능의 도덕성에 관한 연구, 인공지능의 도덕적 지위 및 권리를 탐구하는 연구, 인공지능의 윤리적 책임 범위에 관한 연구, 인공지능의 윤리적 문제를 예측하는 연구, 각종 인공지능의 이슈를 탐색하는 연구들이다. 인공지능의 도덕적 지위를 및 권리를 탐구할 때에는 철학적 관점에서 유교학, 양명학 등의 동양철학 또는 칸트, 흄, 요나스 등의 서양철학 사상 관점에서 인공지능의 지위를 고찰하고 있었고, 인공지능이 야기할 수 있는 윤리적 문제로써 인공지능의 평향성, 거짓 등의 문제를 예측하여 보고 이것에 대한 인공지능의 윤리적 책임 범위를 탐색하기도 하였다. 특정 분야를 정하지 않고 인공지능의 윤리적 이슈 전반을 탐색하여 인류를 위한 유의미한 시사점을 얻고자 하는 연구들도 있었다. 구체적으로 살펴보면, 이재숭(2020)은 인공적인 도덕행위자의 도덕적 지위 설정을 위해 정보 철학자들의 이론을 탐색하였으며, 양선진(2016)은 인공지능시대에 인간과 로봇의 사회에 적합한 과학기술의 윤리가 요청된다고 이야기하며 양명학의 이론을 기초로 관용, 존중, 배려의 덕목을 제시하였고, 김다솜, 맹주만(2021)은 도덕을 이해하는 다양한 관점이 존재할 수 있다는 입장에서 칸트적 모델과 흄적 모델을 기반으로 도덕적 기계의 가능한 모델을 설계하고자 하였다. 추병완(2017)은 도덕적 인공지능에 관한 비판적 고찰 연구에서 도덕적 인공 지능을 개발할 필요가 있음을 강조하면서 도덕적 인공 지능이 수행해야 할 구체적인 기능 목록을 제시하였고, 변순용(2018)은 인공지능 및 로봇의 개발과 사용 과정에서 발생 가능한 사회적 문제들을 해결하기 위하여 인공지능 및 로봇에 대한 윤리 가이드라인을 제안하였으며, 박소영(2019)은 인간과 인공지능의 공존을 위한 윤리적 문제의 탐색 연구를 진행하였다.

2) 제도적 측면

인공지능 윤리에 관한 제도적 측면의 연구들은 '법'과 '제도'를 이용하여 인공지능을 '규제'함으로써 인공지능의 부정적 영향을 최소화하기 위한 방안을 모색한다. 제도적 측면에 관련하여서는 인공지능 규제 거버넌스, 세계 각국의 인공지능 거버넌스, 인공지능 공정성 제고를 위한 제도, 인공지능 보안 패러다임, 인공지능 윤리 인증 정책 등의 연구가 있었고, 법적 측면에 관련하여서는 인공지능 규제 법안, 인공지능의 입법 동향 탐색, 인공지능 윤리의 법적 과제, 인공지능의 법인격 및 책임에 관한 연구들이 있다. 구체적으로 살펴보면, 민춘기(2020)는 독일 AI 국가전략의 지향점이 한국의 정책 방향에 주는 시사점을 제시하였고, 박혜성, 김법연, 권헌영(2021)은 유럽연합이 AI에 대한 윤리 가이드라인과 관련 법안을 준비하고 있다고 밝혔다. 방정미(2021)는 인공지능 관련 규제에 있어 규제 강화나 규제 약화의 차원에서의 논의에서 벗어나 개별적으로 합리적이고 효율적인 맞춤형 규제를 고려하는 단계로 발전되어야 한다고 이야기하며 유럽과 미국의 알고리즘 규제 및 인공지능 관련 법안을 비교하였다. 김용대, 장원철(2016)는 인공지능의 발전 속에서 개인정보보호 관련 법률체계를 고찰하고 발전 방향을 제시하였고, 정용하(2018)는 미국의 자율주행 자동차의 안전성과 윤리 및 법적 책임에 관한 연구에서 제조사의 손해배상책임을 강조하였다.

3) 교육적 측면

인공지능 윤리에 관한 교육적 측면의 연구들은 구성원들이 인공지능 윤리 의식을 효과적으로 '함양'할 수 있도록 하는 구체적인 '교육 방안'을 모색한다. 이 범주에서 이루어지는 연구들은 인공지능 윤리교육의 필요성 탐색, 인공지능 윤리교육의 교육과정, 교수학습모델, 교육프로그램, 학습요소추출, 강의 콘텐츠 개발 연구들이다. 또한 교육의 필요성을 뒷받침하거나 효과성을 검증하기 위해 인공지능 학습자 인공지능 윤리의식의 정도를 진단하거나 윤리의식 검사도구 개발하기도 하였다. 구체적으로 살펴보면, 김태창, 변순용(2021)은 AI 윤리교육의 필요성과 내용 구성에 관해 연구하였고, 김지언, 이철현(2021)은 초등학생 인공지능 윤리교육을 위한 STEAM 프로그램을 개발하였으며, 김귀식, 신영준(2021)은 인공지능 윤리의식 검사 도구 개발 연구를 진행하였다. 세부 교과 영역으로는 주로 도덕 교육과 윤리 교육 분야에서 관련 연구를 진행중임을 알 수 있었다.

4) 의학적 측면

인공지능 윤리에 관한 의학적 측면의 연구들은 의학적 지식과 기능을 가진 새로운 존재인 인공지능에 관련하여 새로운 ‘생명의료윤리’를 정립하고, ‘의료적 책임’의 범위를 설정하고자 노력한다. 의료 분야에 대한 인공지능의 기여는 분명 매우 가치로우며 획기적이나, 인간의 안전과 생명이라는 중요한 부분이기에 더욱 신중해야 한다. 그리고 그 책임 또한 엄중하다. 이와 관련하여 나해란, 김현성(2020)은 인공지능시대의 의료윤리의 방향을 연구하였고, 조수경(2021)은 의료분야의 인공지능 도입과 생명의료윤리 정립 방안을 연구하면서 인공적 도덕 행위자로서의 인공지능의사와 보건의료인을 위한 윤리적 원칙을 구체화하고자 하였다. 정창록(2018)은 인공지능 로봇 의료의 도덕적 이상향을 모색하면서, 2017년 유럽연합(European Union, EU)이 AI로봇에게 전자인간으로서의 법적 지위를 부여한 사안을 고찰하며, AI로봇 의사의 책임 범위를 탐색하였다.

5) 융합적 측면

인공지능 윤리에 관한 융합적 측면의 연구들은 언어, 문학, 영화, 종교, 군사 등 인간 ‘삶 속의 다양한 영역’에서 인공지능 윤리와의 연관을 탐색한다. 앞서, 연구의 필요성에서 ‘인공지능은 우리 주변의 어디에나 존재하고, 필수적인 생활의 일부가 될 것이다.’라고 이야기한 것과 같이 인공지능은 필수불가결하게 우리 삶과 긴밀히 연결될 것이므로, 인공지능 윤리와 관련된 올바른 제도의 정립을 위해서는 인공지능 윤리가 실제 우리 삶의 각 분야와 어떻게 관계맺음을 하고 있는지를 우선 살펴야 한다. 구체적으로 살펴보면, 이청호 외(2021)는 언어를 통한 인공지능의 윤리성 검증을 위해, 인공지능을 위한 비도덕 문장 판별 온톨로지 구축을 연구하였고, 김익현(2020)은 군사적 인공지능 분야에서 무기체계에의 인공지능 적용을 연구하였다. 추재욱(2018)은 소설 ‘뉴로맨서’에 나타난 인공지능 윤리의식에 대하여, 박선화(2018)는 영화 ‘그녀’에서의 인공지능과 몸, 감정, 윤리적 주체의 문제를 탐색하였다.

Ⅴ. 결론 및 시사점

이 연구의 목적은 인공지능 윤리에 관한 국내 연구에 대하여 고찰하고, 앞으로의 인공지능 윤리 국내 연구를 위한 유의미한 시사점을 얻는 것에 있었다. 이를 위해 체계적 문헌 고찰을 통하여 연구를 진행하였으며 2006년부터 2021년 9월까지의 관련 국내 연구물 중 최종 181편의 연구를 선정하여 분석하였다. 연구 분야별 비율을 살펴본 결과 법학이 33.1%, 윤리학 22.6%, 교육학 13.4%, 정책학 9.9%, 철학 6.6%, 과학기술학 5.5% 로 나타났다. 인공지능 윤리의 주요 연구 내용을 분야별로 범주화하여 살펴보면 인공지능 윤리에 관한 철학적 측면의 연구들은 인공지능이라는 존재에 관심을 두고, 제도적 측면의 연구들은 인공지능의 규제 및 관리 방안을 모색하며, 교육적 측면의 연구들은 구성원들의 인공지능 윤리 의식의 함양에 관심을 두었다. 이 연구의 결과를 통하여 다음과 같은 시사점을 얻을 수 있었다.

첫 번째 시사점은 연구의 양적 성장에 관한 것이다. 인공지능 윤리에 관한 국내 연구는 [그림 2]와 같이 2016년부터 활발히 시작되어 매년 1.5~2배 가량의 양적 성장을 보이고 있었다. 인간 이세돌과 인공지능 알파고(Alphago)의 바둑 대결에서 알파고가 승리하면서 본격적인 인공지능의 시대가 열리게 되었던 시기가 2016년임을 떠올려 보았을 때 연구자들은 인공지능에 이목이 집중되기 시작하는 시점부터 인공지능 윤리 연구에 관심을 보였고, 시간이 흐를수록 인공지능 윤리에 관한 관심도 역시 높아지고 있음을 알 수 있었다. 인공지능 기술의 발전은 인공지능 윤리와 한 쌍의 바퀴를 이룰 때 지속가능하게 의미있을 것이므로, 지금과 같이 인공지능 윤리 연구에 대한 관심도가 꾸준히 이어지도록 해야 할 것이다.

두 번째 시사점은 연구의 분야에 관한 것이다. 본 연구는 '인공지능'에 관한 '윤리적 접근'에 관한 것이다. 윤리학은 가치판단 및 도덕적 평가의 기준을 제시하는 규범 윤리학과 도덕의 실천적 방면을 연구하는 실천 윤리학으로 구분할 수 있는데, 이 연구에서 고찰한 문헌 중 윤리학과 철학 관련 문헌은 규범 윤리학의 성격을 가지며, 법학 및 교육학 등에 관련한 문헌은 실천 윤리학의 성격을 가지는 것으로 구분할 수 있다. <표 2>의 분야별 빈도를 이러한 기준에 의하여 살펴보면 29.2%가 규범 윤리학, 70.8%가 실천 윤리학 기반의 연구임을 알 수 있다. 인공지능 윤리에 관한 국내 연구는 앞서 언급한 것처럼 꾸준한 상승세를 보이고 있으나, 규범 윤리학으로써의 인공지능 윤리 연구 비중이 전체의 30% 이내라는 점에 있어서 '가치판단의 기준에 대한 연구가 충분히 이루어지고 있는가?'라는 물음이 남을 수 있다. 바람직한 실천의 방향은 충분한 규범적 고찰을 토대로 정립될 수 있

을 것이다.

세 번째 시사점은 연구의 내용에 관한 것이다. 문헌들의 내용을 살펴보았을 때 인공지능의 급속한 발전으로 인한 사회 혼란을 최소화하기 위해 인공지능의 윤리적 성격에 관한 연구와 이에 대한 법제화 관련 연구를 신속하게 진행하여 온 것을 알 수 있다. 그러나 앞으로 우리 사회에서 인공지능이 적용되는 범위가 굉장히 광범위할 것이라는 점을 고려하였을 때 법학, 윤리학, 교육학, 정책학, 철학의 범주에 인공지능 윤리 연구가 집중되고 있고 그 외의 분야에 관련하여서는 인공지능 연구가 미약함에 아쉬움이 남는다. 앞으로 인공지능은 언제 어디에서나 우리 삶에 깊숙이 관여하게 될 것이므로, 보다 광범위한 영역에서 융합적 관점으로 인공지능 윤리에 관심을 기울일 필요가 있다.

이러한 결론 및 시사점을 바탕으로 추후 인공지능 윤리 연구가 여러 영역에서 활발하게 진행될 것과, 규범 윤리학 분야의 인공지능 윤리 연구에 더욱 관심을 기울일 것을 제언한다. 또한 추후 인공지능 윤리의 개념과 정의 문제, 우리나라에서 인공지능의 윤리적 활용을 실현할 구체적 방안 등에 대한 후속 연구가 진행될 것을 제언한다.

□ 투고(접수)일 : 2021년 11월 20일 / 심사(수정)일 12월 07일 / 게재확정일 12월 23일

참 고 문 헌

국립국어원(2018), “윤리”, https://stdict.korean.go.kr/search/searchResult.do?pageSize=10&searchKeyword=%EC%9C%A4%EB%A6%AC (검색일: 2021. 11.12.)

김광연(2018), “인공지능 및 사이버휴먼 시대의 윤리적 논쟁과 규범윤리의 요청”, 『인문학연구』, 57(2), 55-77.

김귀식, 신영준(2021), “인공지능 윤리의식 검사 도구 개발 연구”, 『한국인공지능교육학회』, 2(1), 1-19.

김다솜, 맹주만(2021), “인공지능과 도덕적 기계-칸트적 모델과 흄적 모델”, 『철학탐구』, 62, 177-216.

김영식(2019), “AI와 고용, 경제성장, 불평등: 최근 문헌 개관과 정책 함의”, 『한국경제학회』, 12(3), 1-34.

김용대, 장원철(2016), “인공지능산업 육성을 위한 개인정보보호 규제 발전 방향”, 『경제규제와 법』, 9(2), 161-176.

김익현(2020), “윤리적 인공지능 개발을 위한 시험평가검증확인(TEVV) 전략”, 『한국국방기술학회 논문지』, 2(1), 1-4.

김지언, 이철현(2021), “초등학생 인공지능윤리교육을 위한 STEAM 프로그램 개발”, 『한국인공지능교육학회』, 2(1), 21-28.

김태창, 변순용(2021), “AI 윤리교육의 필요성과 내용 구성에 관한 연구”, 『인공지능인문학연구』, 8, 71-104.

나해란, 김현성(2020), “빅데이터, 인공지능시대의 의료윤리”, 『Journal of Korean Diabetes』, 21(3), 126-129.

민춘기(2020), “독일 AI 국가전략의 지향점이 한국의 정책 방향에 주는 시사점”, 『독일어문학』, 88, 165-182.

박선화(2018), “스파이크 존즈의 〈그녀〉에서의 인공지능과 몸, 감정, 윤리적 주체의 문제”, 『스토리&이미지텔링』, 15, 117-143.

박소영(2019), “인간과 인공지능의 공존 가능성에 대한 탐색: 책임의 윤리와 문학적 상상력”, 『한국윤리학회』, 124, 17-35.

박형빈(2020), “인공지성의 윤리 문제와 새로운 도덕과 교육과정 반영 방안”, 『한국도덕윤리과교육학회 학술대회 자료집』, 20(10), 679-710.

박혜성, 김법연, 권헌영(2021), “인공지능 규제에 대한 연구 - 유럽연합의 입법안을 중심으로”, 『공법연구』, 49(3), 349-374

방정미(2021), “인공지능 알고리듬 규제거버넌스의 전환 - 최근 미국의 알고리듬 규제와 인공지능 윤리원칙을 중심으로 -”, 『한국공법학회』, 49(3), 375-406.

변순용(2018), “인공지능로봇을 위한 윤리가이드라인 연구-인공지능로봇윤리의 4원칙을 중심으로-”, 『윤리교육연구』, 47, 233-252.

안정용(2021), “인공지능 윤리: 지각된 인공지능 자유의지가 인공지능의 윤리적 책임에 미치는 영향을 중심으로”, 고려대학교 박사학위논문.

양선진(2016), “양명학을 통해 본 인공지능(AI) 시대의 과학기술 윤리”, 『한국양명학회』, 45, 479-507.

이봉재(2006), “인공지능과 책임의 문제”, 『대동철학회』, 37, 73-92.

이재숭(2020), “AMA의 도덕적 지위의 문제”, 『새한철학회』, 102, 527-545.

이청호, 김봉제, 김형주, 변순용, 이찬규(2021), “윤리적 인공지능을 위한 비도덕 문장 판별 온톨로지 구축에 대한 연구”, 『인공지능인문학연구』, 7, 149-170.

임미가(2021), “AI융합 개별화 학습을 위한 초·중등 기술교육의 학습자 성향 국내 문헌 고찰”, 『한국실과교육학회』, 34(2), 121-145.

임미가(2020), “인공지능 시대에서 기술 교육의 방향에 관한 고찰”, 『한국실과교육학회』, 33(4), 81-102.

정용하(2018), “미국의 자율주행 자동차의 안전성과 윤리 및 법적 책임”, 『미국헌법학회』, 29(2), 199-244.

정창록(2018), “인공지능로봇의료의 도덕형이상학적 모색: 의료적 전자인간의 책임가능성”, 『한국의료윤리학회지』, 21(2), 143-156.

조수경(2021), “의료분야의 인공지능 도입과 생명의료윤리 정립 연구”, 부산대학교 대학원, 박사학위논문.

최경석(2020), “인공지능이 인간 같은 행위자가 될 수 있나?”, 『생명윤리』, 21, 71-85.

추병완(2017), “도덕적 인공지능에 관한 비판적 고찰”, 『윤리교육연구』, 44, 1-24.

추재욱(2018), “뉴로맨서에 나타난 인공지능 윤리의식에 대한 연구”, 『한국문화융합학회』, 1-7.

하영숙(2019), “인공지능이 인간존엄성에 미치는 영향에 대한 생명윤리적 고찰”, 가톨릭대학교 박사학위논문.

Australian National Health and Medical Research Council(2000), *How to review the evidence: systematic identification and review of the scientific literature.*

Ellul, J.(1954), *The technological society. Newyork: Vintage*, 박광덕 역(1996), 『기술의 역사』, 서울: 한울.

Fischer, J. M. (1994), *The metaphysics of free will (Vol. 1),* Oxford: Blackwell.

HLEG, A. (2019), *High-level expert group on artificial intelligence, Ethics Guidelines for Trustworthy AI.*

Khan, K. S., Kunz, R., Kleijnen, J. & Antes, G.(2003), "Five steps to conducting a systematic review", *Journal of the royal society of medicine,* 96, 118-121.

Kitchenham. B. (2004), *Procedures for Performing Systematic Reviews, Software Engineering Group Department of Computer Science Keele University Keele.* Staffs ST5 5BG, UK Keele University Technical Report TR/SE-040.

Rusell, S. J., Norvig,P. (2010), *Artificial Intelligence: A Modern Approach (AIMA)*, 류광 역(2016). 『인공지능: 현대적 접근방식』. 서울: 제이펍.

Shahriari, K., & Shahriari, M. (2017, July), “IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems". *In 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), IEEE*, 197-201,

<Abstract>

A systematic literature review of research on artificial intelligence ethics

Lim, Mika

The purpose of this study was to examine the research on artificial intelligence ethics and to obtain significant implications for future artificial intelligence ethics research. For this purpose, 181 studies were analyzed using a systematic literature review method. The ratio by research topic was 33.1% in Law, 22.6% in Ethics, 13.4% in Education, 9.9% in Policy Studies, 6.6% in Philosophy, and 5.5% in Science and Technology. The contents of this study's AI ethics research were reviewed by subject. Studies in the philosophical aspect focus on the existence of artificial intelligence, studies in the institutional aspect look for ways to regulate and manage artificial intelligence, and studies in the educational aspect pay attention to the cultivation of ethical awareness of AI among members.

It is suggested that AI ethics research will be conducted in a wide area, and that more attention should be paid to AI ethics research as a normative ethics. In addition, it is suggested that follow-up research on specific ways to realize the ethical use of artificial intelligence in Korea is carried out in the future.

key words : AI, AI Ethics, AI Education, AI Policy