J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr
Copyright ⓒ 2022 KSII

인공지능 윤리원칙 기반의 인격권 및 재산보호를 위한 인공지능 윤리 측정지표에 관한 연구

A Study on the Artificial Intelligence Ethics Measurement indicators for the Protection of Personal Rights and Property Based on the Principles of Artificial Intelligence Ethics

소 순 주[1] 안 성 진[1*]
Soonju So Seongjin Ahn

요 약

지능정보화 사회에서 가장 핵심으로 발전하고 있는 인공지능은 인간에게 편의성과 긍정적인 삶의 변화를 가져오고 있다. 하지만 인공지능 발전과 함께 인간의 인격권과 재산이 위협받고, 윤리적인 문제가 발생하는 사례도 증가하고 있기 때문에 그에 따른 대안이 필요하다. 본 연구에서는 인공지능의 역기능에서 가장 쟁점화되고 있는 인공지능 윤리(Artificial Intelligence Ethics) 문제를 인공지능 윤리원칙과 구성요소 기반 하에 우선적으로 인간의 인격권과 재산을 보호할 수 있도록 인공지능 윤리 측정지표를 연구, 개발하는 데 목표를 두었다. 인공지능 윤리 측정지표를 연구, 개발하기 위해 다양한 관련 문헌과 전문가 심층 면접(FGI), 델파이 설문조사를 실시하여 43개 항목의 윤리 측정지표를 도출하였다. 설문조사와 통계분석에 의하여 윤리 측정지표에 대한 기술통계량 분석, 신뢰도 분석, 상관관계 분석으로 40개 항목의 인공지능 윤리 측정지표를 확정하여 제안하였다. 제안된 인공지능 윤리 측정지표는 인공지능 설계, 개발, 교육, 인증, 운영, 표준화 등에 활용될 수 있으며, 안전하고 신뢰할 수 있는 인공지능 발전에 기여할 수 있을 것이다.

☞ 주제어 : 인공지능, 인공지능 윤리, 윤리 측정지표, 측정지표, 인격권

ABSTRACT

Artificial intelligence, which is developing as the core of an intelligent information society, is bringing convenience and positive life changes to humans. However, with the development of artificial intelligence, human rights and property are threatened, and ethical problems are increasing, so alternatives are needed accordingly. In this study, the most controversial artificial intelligence ethics problem in the dysfunction of artificial intelligence was aimed at researching and developing artificial intelligence ethical measurement indicators to protect human personality rights and property first under artificial intelligence ethical principles and components. In order to research and develop artificial intelligence ethics measurement indicators, various related literature, focus group interview(FGI), and Delphi surveys were conducted to derive 43 items of ethics measurement indicators. By survey and statistical analysis, 40 items of artificial intelligence ethics measurement indicators were confirmed and proposed through descriptive statistics analysis, reliability analysis, and correlation analysis for ethical measurement indicators. The proposed artificial intelligence ethics measurement indicators can be used for artificial intelligence design, development, education, authentication, operation, and standardization, and can contribute to the development of safe and reliable artificial intelligence.

☞ keyword : Artificial intelligence, AI ethics, AI ethics Measurement indicators, Measurement indicators, Personal Rights

1. 서 론

인공지능은 모든 산업분야에 없어서는 안 될 중요한 요소로 자리매김하고 있다. 빅데이터, 클라우드, 메타버스, 지능형 로봇, 자율주행 자동차 등 4차 산업혁명의 가장 중심에 있다고 볼 수 있다. 인공지능 발전과 함께 사회적 이슈로 인공지능 윤리 (Artificial Intelligence Ethics)가 급속하게 떠오르고 있다[1-2]. 인공지능 기술이 발전하면서 인간에게 많은 편의성과 효용성을 제공하지만 한편으로는 인간의 인격권을 위협하거나 재산을 파괴할 수도 있다[3]. 이러한 역기능을 방지하고자 국내외에서는 인공지능 윤리 가이드라인, 윤리 헌장, 윤리원칙, 개발지침, 정책, 법·제도

1 Dept. of Computer Education, Sungkyunkwan University, Seoul, 03063, Korea.
* Corresponding author (sjahn@skku.edu)
[Received 2 May 2022, Reviewed 10 May 2022, Accepted 30 May 2022]

http://dx.doi.org/10.7472/jksii.2022.23.3.111

등을 제정하여 인공지능에 적용하고 있다[4-5]. 인공지능에 윤리를 적용하기 위해서는 먼저 인공지능 윤리원칙이 제대로 정의되어야 하고, 윤리원칙 구성요소에 대한 측정지표가 있어야 한다. 현재 인공지능 윤리원칙은 대부분 인공지능 윤리 가이드라인이나 윤리 헌장, 정부 정책, 연구보고서 등에 포함되어 배포되었다[6-7].

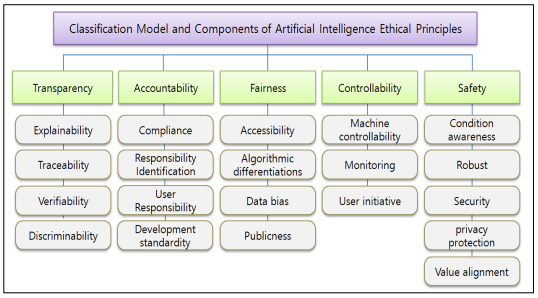
최근 인공지능의 폭발적인 수요에 대비하여 학습용 데이터 구축, 빅데이터 수집, 모빌리티 플랫폼에 의한 데이터 확보, 마이데이터 확보, 금융 및 헬스케어 데이터 구축 등이 활발하게 이루어지고 있다. 인공지능에 활용할 데이터에 대해서는 데이터의 편향적 특성을 제거하고, 데이터의 사실성, 데이터의 품질관리, 개인 정보보호, 데이터에 대한 설명 가능한 투명성 확보, 데이터 추적성 확보, 책임에 대한 식별이 가능하도록 하여 인공지능에 대한 안전과 신뢰성을 확보할 수 있는 검증 기준이 필요하지만 아직까지 명확하게 인공지능 윤리를 검증할 수 있는 윤리 측정지표가 정의되거나 개발되지 않았다.

국내외에서 인공지능 윤리에 대한 관심과 함께 윤리원칙 정의, 지침 개발, 정책 연구 등이 초기 단계에 있기 때문에 인공지능 기술 발전과 더불어 선제적으로 인공지능 윤리 측정지표가 개발되어야 할 필요성이 요구된다. 본 논문에서는 2021년 11월에 발표된 인공지능 윤리원칙 분류 모형 및 구성요소에 관한 연구 논문[8]과 국내외 인공지능 윤리 체크리스트, 윤리헌장, 윤리 가이드라인, 정책, SW 인증, 품질인증 등 관련 문헌 등의 자료를 분석하여 전문가 그룹 인터뷰(FGI)를 진행하여 인공지능 윤리원칙 중 우선적으로 인간과 인공지능 간 상호작용 시 인간의 인격권과 재산을 보호하기 위하여 적용되어야 할 윤리원칙을 선정하였다. 이를 통해 선정된 윤리원칙에서 윤리 측정지표를 도출하고, 설문조사 결과를 분석하여 인공지능 윤리 측정지표를 개발, 검증하여 제시하고자 한다.

2. 이론적 배경

2.1 인공지능 윤리원칙 분류 모형 및 구성요소

인공지능 윤리는 2004년에 로봇윤리 13원칙을 유럽로봇연구연합(EURON)에서 발표 후 미국, 일본, 유럽이 중심이 되어 인공지능 윤리 헌장, 가이드라인, 개발 지침, 윤리 표준 등이 꾸준히 발표되었다[9-10]. 2016년부터는 세부적인 인공지능 개발 지침, 윤리 가이드라인, 정부 정책을 바탕으로 인공지능 기술 개발 시 적용하고 있다. 2021년 11월 인공지능 윤리 분류 모형 및 구성요소에 관한 연구 논문은 향후 인공지능 윤리 분야에서 구체적인 윤리 측정지표를 개발할 수 있도록 제시되었다[8]. 인공지능 윤리원칙 분류 모형 및 구성요소에 대해서는 그림 1과 같다. 인공지능 윤리 분류 모형에서는 5개의 윤리원칙을 투명성, 책무성, 공정성, 통제성, 안전성으로 정의하였다. 투명성 원칙에는 설명 가능성 외 3개의 구성요소로 정의하였으며, 책무성 원칙에는 준법성 외 3개를 구성요소로 정의하였다. 공정성 원칙에는 접근성 외 3개, 통제성 원칙에는 제어 가능성 외 2개, 안전성 원칙에는 상태 인식성 외 4개의 구성요소로 정의하였다.



(그림 1) 인공지능 윤리원칙 분류모형 및 구성요소
(Figure 1) Classification Model and Components of Artificial Intelligence Ethical Principles[8]

인공지능 윤리원칙 분류 모형 및 구성요소에서 제시한 내용은 최근 신뢰할 수 있는 인공지능 개발 전략에 핵심적으로 이슈화되고 있는 인공지능 개발, 사용, 운영 요소에 필요한 인공지능 윤리 특성이 포함되어 있다.

2.1 인공지능 윤리 측정지표 주요 쟁점

인공지능 제품이나 서비스가 인간과 상호작용이 증가하면서 인간에게 편의를 제공하지만, 그와 함께 인간의 인격권 위협과 재산 피해를 발생시키고 있다. 인격권은 사람의 생명, 신체, 건강, 자유, 명예, 사생활, 초상, 개인정보, 그 밖의 인격적 이익에 대한 권리를 의미하며, 인공지능에 의한 위협이나 침해로부터 보호받아야 한다[11-12].

국내외 많은 기관에서 인공지능 제품에 대한 품질관리를 목적으로 인증 제도를 준비하고 있으며, 인공지능 학습용 데이터 구축에 따른 데이터 편향성, 알고리즘의 편향성, 데이터의 유효성, 개인정보 수집 과정에서의 보안성, 개인정보 활용, 데이터의 사실성과 유효성 검사, 데이터의 오남용, 학습모델의 유효성 등 다양한 특성을 반영하여 품질관리 지표를 만들어 적용하고 있다[13].

유럽연합에서는 인공지능 윤리 체크리스트를 개발하여 신뢰할 수 있는 인공지능 개발에 시범 적용하고 있으며, 영국 ICO(Initial Coin Offering)에서는 인공지능 및 데이터 보호 안내 지침을 개발하였고, 마이크로소프트사에서는 인공지능 공정성 체크리스트를 개발하여 적용하였다. 카네기멜론대학 소프트웨어 공학 연구소가 개발한 윤리적 인공지능 설계 체크리스트는 인공지능 윤리원칙을 바탕으로 개발되었다[14-16]. IEEE SA에서 자동화 지능형 시스템을 위한 윤리 표준 개발이 P7000 시리즈로 개발되고 있으며, 그중 P7010은 2020년에 인공지능이 인간의 행복에 미치는 영향 평가를 위한 표준으로 공표되었다[17].

국내에서는 인공지능 도덕성 연구 시스템을 개발하여 소셜로봇의 윤리적 판단 기능에 적용 가능한 10세 아동 수준의 도덕성을 갖춘 인공지능 윤리를 연구하고 있다. 이와 함께 인간과 로봇의 상호작용을 통합한 소프트웨어를 구현하여 AMA(Artificial Moral Agent)가 인간의 편리한 삶과 복지 및 행복에 기여하는 동반자 로봇 기반 기술을 확립하는 데 목표를 두고 있어 향후 로봇을 대상으로 한 윤리 인증 제도 마련에 적용될 수 있도록 연구하고 있다[18].

2.3 인공지능 윤리 측정지표 관련 문헌

국내외 인공지능 정책, 윤리헌장, 윤리 가이드, 윤리원칙, 표준화, 인증 제도, SW 측정지표, 품질관리 지표, ISO/IEC, 윤리 체크리스트, 정보보호, 논문, 개발 지침, 서적, 제품 및 서비스 등의 문헌을 바탕으로 인공지능에 대한 주요 윤리적 쟁점을 중심으로 측정지표를 도출하였다. 표 1은 인공지능 윤리 측정지표 도출을 위해 검토한 문헌 분석 자료 현황이다.

인공지능 윤리원칙 분야에서는 8종의 국내외 문헌을 분석하였으며 주요 쟁점은 인간존중, 보안 확보, 사회 공공성, 기술 합목적성, 자율성, 책임 식별에 대한 문제들 이었다. 윤리원칙은 공정성, 투명성, 통제성, 책무성, 안전성에 알고리즘 차별성과 데이터 편향성에서 측정지표를 도출하였다[19]. 인공지능 윤리가이드라인과 체크리스트는 국내외 12종에 대한 문헌을 분석하였다. 인공지능 관련 표준화 및 인증에 대한 문헌은 5종을 분석하였으며 주요 쟁점은 기술 적합성, 다양성, 완전성, 유효성, 호환성, 상호운용성, 보안성, 시험기준, 데이터 통제, 알고리즘 편향성 등에 대한 항목이 도출되었다[20]-[22].

인공지능 관련 정책, 연구 보고서는 7종을 분석하였으며 국내외를 막론하고 인공지능 발전과 함께 정부 차원에서 지속적인 연구와 인공지능 윤리 문제 해결을 위하여 연구와 법·제도 발의가 증가하고 있다[23]-[24].

(표 1) 인공지능 윤리관련 문헌 분석
(Table 1) Analysis of Artificial Intelligence Ethics Literature

문헌 분류	주요 윤리적 문제 측정 대상
AI 윤리원칙 (8 종)	· 인간존중, 보안 확보, 사회 공공성, 기술 합목적성, 자율성, 식별성, 가치 정렬 · 투명성, 통제성, 공정성, 책무성, 안전성, 알고리즘 차별성, 데이터 편향성 등
AI 윤리 가이드라인, 체크리스트 (12 종)	· 투명성, 책임, 안전, 인권보장, 다양성 존중, 침해금지, 데이터 편향성, 개인 정보보호, 건강, 행복, 생명보호 · 감독, 견고성, 거버넌스, 차별 금지, 사회적 영향, 정직성과 사용성, 데이터 처리 적법성 등
표준화, 인증 (5 종)	· 기술 적합성, 다양성, 완전성, 정확성, 유효성, 성능 효율성, 호환성, 상호운용성, 학습성, 접근성, 보안성, 이식성, 유지 보수성, 시험기준, 데이터 통제, 알고리즘 편향성 등
정책, 연구 보고서 (7 종)	· 신뢰 가능한 인공지능, 알고리즘 차별성, 데이터 편향성, 책임 식별, 인간의 생명 안전, 공공데이터 수집 조건, 위험회피, 접근성, 기술적 정확성, 보안성 인증, 데이터 주권, 자유, 건강 등
논문, 서적 (11종)	· 알고리즘의 편향성, 공정성, 법적 의사결정, 알고리즘 위험, 분류 기준, 상태 모니터링, 제어, 통제, 개인 정보보호, 데이터 이력, 검증 가능성, 운영 감시, 운영 감사, 사고 발생 이력 관리 등
AI 제품, 서비스 (6 종)	· 데이터 편향 방지, 차별 방지, 알고리즘 차별 방지, 개인 정보보호, 사용자 제어, 모니터링, 위급 상황 대응, 사이버 공격 대응, 법적 책임, 아동보호, 혐오 표현, 가명 처리 등

인공지능 윤리 관련 논문, 서적, 저널을 11종을 분석하였으며 대부분 윤리원칙 특성 수준에서 발표가 많이 되었고 윤리원칙을 세부적으로 연구하여 측정 기준이나 지표로 연구되는 자료가 빈약한 편이었다. 인공지능 제품이나 서비스에 대해서는 6종을 분석하였으며 기업에서 제공하는 윤리 헌장, 서비스 기능, 제품 소개서, 매뉴얼, 약관 등에 인공지능 윤리 준수에 대한 항목이 포함되어 있는 것으로 분석되었다[25]-[29].

3. 인공지능 윤리 측정지표 개발 연구

3.1 연구 방법

본 연구를 하기 위하여 인공지능 윤리원칙 분류와 구성요소를 바탕으로 국내외 다양한 인공지능 관련 문헌을 통하여 인공지능 윤리에 적용할 측정항목을 도출하여 분석하였다. 문헌 분석을 통하여 정리된 인공지능 윤리 측정지표 항목은 인간과 인공지능 간 상호작용에서 인간의 생명과 재산을 보호하기 위한 기준을 마련하고자 하였다. 인공지능 윤리원칙 구성요소에서 선행 연구해야 할 구성요소 선정과 측정지표 도출을 위하여 20명의 전문가 심층면접(Focus Group Interview) 3회와 델파이 설문조사 2회를 실시하였다. FGI에서는 브레인스토밍으로 진행하고, 산출물 작성 도구는 마인드맵(MindMap)을 사용하여 측정 대상 항목과 의견을 정리하였으며, 설문조사는 내용타당도를 분석하여 측정지표 도출 대상 인공지능 윤리원칙의 구성요소를 선정하였다.

FGI에서 최종적으로 도출된 인공지능 윤리 측정지표에 대한 상관관계를 검증하기 위해 80명의 인공지능 관련 이해관계자를 대상으로 8일 동안 설문조사를 실시하였다. 설문결과로 신뢰도 분석, 기술통계량 분석, 윤리원칙 구성요소와 측정지표 간 상관관계분석 (Correlation Analysis)을 통하여 인공지능 윤리 측정지표를 검증하고 해당 지표에 대한 정의를 제시하고자 하였다.

3.2 전문가 심층 면접(FGI)

FGI에 참여자는 인공지능과 인공지능 윤리 관련 업무에 종사한 이해관계자 20명이었으며, 3차에 거쳐 FGI를 진행하고, 델파이 설문조사는 2회 실시하였다. 표 2는 FGI와 델파이 설문조사에 참여한 이해관계자 현황이다.

(표 2) 전문가 그룹 참여자 현황
(Table 2) Information of experts group

직 무	인원(비율)	평균경력연수	평균연령
AI 윤리	5 (25.0)	4.8	48.4
AI 법률	3 (15.0)	5.3	50.1
AI 교육	3 (15.0)	5.0	50.0
AI 설계	4 (20.0)	4.3	47.5
AI 개발	3 (15.0)	4.6	41.2
AI 제작	2 (10.0)	5.9	54.5
Total	20 (100)	4.9	48.6

FGI와 델파이 설문에 참여한 전문가 직무는 AI 윤리 5명, AI 관련 법률가 3명, AI 설계 및 컨설턴트 4명, AI 개발자 3명, AI 제작자(기업) 2명, AI 교육자 3명이 참여하였고, 해당 직무에 종사한 평균 경력 연수는 4.9년이다.

1차 델파이 설문조사는 인공지능 윤리원칙의 구성요소를 바탕으로 인공지능과 인간이 상호작용하는 과정에서 인간의 생명과 재산을 보호하기 위한 윤리 측정지표를 우선적으로 개발할 구성요소를 선정하는 조사였다. 설문지는 Likert 5점 척도를 사용하였으며, 설문 결과 Likert 4(타당함)와 5(매우 타당함)에 응답한 패널 수를 적용하여 내용타당도 비율(CVR) 값을 산정하여 적용하였다. 표 3은 측정지표 개발에 우선 적용할 인공지능 윤리원칙 구성요소 내용타당도 분석 및 선정 현황이다.

(표 3) 구성요소 내용타당도 분석 및 선정현황
(Table 3) Analysis and Selection of Component Content Feasibility

윤리원칙	구성요소	유효수	최솟값	최대값	평균	표준편차	CVR값	선정
투명성	설명가능성	20	3	5	4.20	.69585	0.70	○
	추적가능성	20	2	5	4.05	.94451	0.60	○
	검증가능성	20	3	5	4.10	.71818	0.60	○
	식별가능성	20	2	4	3.40	.82078	0.20	-
책무성	준법성	20	3	5	4.05	.68633	0.60	○
	책임식별성	20	2	5	3.95	.68633	0.70	○
	이용자책임성	20	3	4	3.65	.48936	0.30	-
	개발표준성	20	2	5	3.45	.88704	0.20	-
공정성	접근성	20	2	5	3.55	.94451	0.30	-
	데이터편향성	20	4	5	4.80	.41039	1.00	○
	알고리즘차별성	20	4	5	4.75	.44426	1.00	○
	공공성	20	3	5	4.05	.68633	0.60	○
통제성	제어가능성	20	2	5	4.00	.79472	0.60	○
	모니터링	20	2	4	3.40	.75394	0.10	-
	이용자주도성	20	3	5	4.10	.64072	0.70	○
안전성	상태인식성	20	2	5	3.80	.83351	0.30	-
	보안성	20	3	5	4.60	.59824	0.90	○
	프라이버시보호	20	3	5	4.45	.68633	0.80	○
	견고성	20	3	5	4.20	.69585	0.70	○
	가치정렬	20	2	4	3.20	.83351	-0.10	-

1차 델파이 설문조사 결과 우선 적용할 인공지능 윤리원칙의 구성요소 20개 중 13개 구성요소가 선정되었다. 선정 기준은 패널 수가 20일 경우 타당성을 갖기 위한 CVR의 최솟값이 0.42이기 때문에 0.42 이상의 CVR 값을 갖는 구성요소를 선정하였다.

13개의 인공지능 윤리원칙 구성요소를 기준으로 1차 FGI를 12명이 참석한 가운데 실시하여 1시간 동안 인공지

능 윤리와 원칙, 구성요소, 가이드라인, 인증 제도, 표준화, IEEE SA에서 진행하는 인공지능 윤리 표준, 윤리적 문제점 및 측정지표 개발에 대한 설명을 실시하였다. 1시간 동안은 미리 준비한 측정지표 도출 자료를 바탕으로 참석자 의견과 추가 측정항목을 도출하였다. 각 구성요소에 대한 측정지표 항목은 설명 가능성에 9개 항목이 도출되었으며, 추적 가능성에 10개 항목, 검증 가능성에 9개 항목, 준법성에 10개 항목, 책임 식별성에 9개 항목, 알고리즘 차별성에 9개 항목, 데이터 편향성에 13개 항목, 공공성에 10개 항목이 도출되었다.

2차 FGI는 13명이 참석하여 통제성 원칙, 안전성 원칙에 대한 구성요소의 측정지표 항목에 대하여 의견과 추가 측정항목을 도출하였다. 각 구성요소에 대한 측정지표 항목은 제어 가능성에 8개 항목, 이용자 주도성에 9개 항목, 견고성에 10개 항목, 보안성에 9개 항목, 프라이버시 보호에 12개 항목이 도출되었다.

3차 FGI는 11명이 참석하여 1차, 2차에서 도출된 측정지표를 대상으로 논의를 진행하였다. 중복된 측정지표 항목을 조정과 추가, 용어나 측정지표에 대한 정의가 명확하지 않은 부분을 명확하게 정의하였다. 최종적으로 측정지표 항목을 도출하였다. 그 결과 설명 가능성에 7개 항목이 확정되었으며, 추적 가능성에 5개 항목, 검증 가능성에 5개 항목, 준법성에 4개 항목, 책임 식별성에 6개 항목, 알고리즘 차별성에 7개 항목, 데이터 편향성에 11개 항목, 공공성에 7개 항목, 제어 가능성에 5개 항목, 이용자 주도성에 6개 항목, 견고성에 7개 항목, 보안성에 5개 항목, 프라이버시 보호에 7개 항목으로 총 82개가 도출되었다.

2차 델파이 설문조사는 3차 FGI에서 도출된 82개 측정지표 항목에 대하여 검증 수행을 위한 타당성 분석을 실시한 조사였다. FGI에 참여했던 전문가 11명을 대상으로 설문조사를 실시하였다. 설문지는 Likert 5점 척도를 사용하였으며, 설문 결과 Likert 4(타당함)와 5(매우 타당함)에 응답한 패널 수를 적용하여 내용 타당 비율(CVR) 값을 산정하고, 11명의 패널 수에 대한 타당성 인정을 위한 최솟값 0.59 이상인 항목을 선정하여 측정지표 검증을 진행하였다. 측정지표로 도출된 82개 측정지표 항목에서 43개 항목이 검증 대상으로 선정되었다. 표 4는 인공지능 윤리 측정지표 도출 후 검증 대상 항목에 대하여 내용타당도 분석을 실시하고 그 결과에 의하여 윤리 측정지표로 선정된 측정항목 현황이다.

선정된 설명 가능성에 속한 측정지표 항목은 최종 사용자를 위한 투명성, 사고 조사자를 위한 투명성, 변호사 및 전문가 증인을 위한 투명성, 이용 범위 고지 투명성,

(표 4) 구성요소별 측정지표 검증 대상 항목 수 현황
(Table 4) Status of the number of items Measurement indicators verification by component

윤리원칙	구성요소	도출된 측정 항목 수	선정된 측정항목 수
투명성	설명가능성	7	5
	추적가능성	5	3
	검증가능성	5	-
책무성	준법성	4	3
	책임식별성	6	4
공정성	알고리즘차별성	7	5
	데이터편향성	11	7
	공공성	7	-
통제성	제어가능성	5	4
	이용자주도성	6	-
안전성	견고성	7	3
	보안성	5	4
	프라이버시보호	7	5
합 계		82	43

이용자 데이터 수집 투명성으로 5개 항목이다. 추적 가능성에 속한 측정지표 항목은 문제 분석을 위한 데이터 제공, 제작 과정의 기록 보관, 사고 발생 이력 보관으로 3개 항목이다. 준법성에 속한 측정지표 항목은 법·제도 준수, 개인정보 제공 동의 획득, 데이터 사용 사전 허가 획득으로 3개 항목이다. 책임 식별성에 속한 측정지표 항목은 설계자 책임 식별, 제작자 책임 식별, 운영자 책임 식별, 이용자 책임 식별로 4개 항목이다. 알고리즘 차별성에 속한 측정지표 항목은 알고리즘의 의사결정 설명, 의도적 차별 요소 제한, 개인 편향적 차별 제한, 불공정한 알고리즘, 인간의 생명 위협 방지로 5개 항목이다. 데이터 편향성에 속한 측정지표는 편향적 특성 제거, 데이터의 사실성, 학습용 데이터 설명, 학습용 데이터의 수량, 데이터의 유효성 검사, 학습용 데이터 품질관리, 데이터 편향 방지 알고리즘 적용으로 7개 항목이다. 제어 가능성에 속한 측정지표는 정책적 통제권 이양, 이용자 제어장치 제공, 생명 위협 시 자동 종료, 운영자 통제 및 제어로 4개 항목이다. 견고성에 속한 측정지표는 인간의 보호 우선, 알고리즘의 오류 관리, 외부 공격에 대한 견고성으로 3개 항목이다. 보안성에 속하는 측정지표는 보안성 인증, 사용자 인증, 상호작용 데이터 보안, 보안 이벤트 사용자 제공으로 4개 항목이다. 프라이버시 보호에 속하는 측정지표는 개인 정보보호 영향 평가, 개인정보 수집 고지, 개인정보 가명화, 개인 민감정보 수집 차단, 개인정보 활용으로 5개 항목이다. 구성요소 중에서 검증 가능성, 공공성, 이용자

주도성은 모두 선정에서 제외되었다.

3.3 인공지능 윤리 측정지표 제안

인공지능 윤리원칙과 구성요소를 기반으로 윤리 측정지표를 도출하기 위하여 FGI와 델파이 설문조사를 실시, 분석한 결과 표 5와 같이 인공지능 윤리 측정지표 항목을 도출하였다. 최종적으로 만들어진 윤리 측정지표 43개 항목에 대하여 인명과 재산을 보호하기 위해 우선적으로 적용해야 할 인공지능 윤리 측정지표 최종 도출 항목을 제안하였다.

(표 5) 인공지능 윤리 측정지표 도출 항목 현황
(Table 5) Status of AI ethics Measurement indicators Items

윤리원칙	구성요소	측정지표 항목
투명성	설명가능성	• 최종 사용자를 위한 투명성 • 사고조사자를 위한 투명성 • 변호사 및 전문가 증인을 위한 투명성 • 이용 범위 고지 투명성 • 이용자 데이터 수집 투명성
	추적가능성	• 문제 분석을 위한 데이터 제공 • 제작 과정의 기록 보관 • 사고 발생 이력 보관
책무성	준법성	• 법・제도 준수 • 개인정보 제공 동의 획득 • 데이터 사용 사전 허가 획득
	책임식별성	• 설계자 책임식별 • 제작자 책임식별 • 운영자 책임식별 • 이용자 책임식별
공정성	알고리즘 차별성	• 알고리즘의 의사결정 설명 • 의도적 차별 요소 제한 • 개인 편향적 차별 제한 • 불공정한 알고리즘 • 인간의 생명 위협 방지
	데이터 편향성	• 편향적 특성 제거 • 데이터의 사실성 • 학습용 데이터 설명 • 학습용 데이터의 수량 • 데이터의 유효성 검사 • 학습용 데이터 품질관리 • 데이터 편향 방지 알고리즘 적용
통제성	제어가능성	• 정책적 통제권 이양 • 이용자 제어장치 제공 • 생명 위협 시 자동 종료 • 운영자 통제 및 제어
안전성	견고성	• 인간의 보호 우선 • 알고리즘의 오류 관리 • 외부공격에 대한 견고성
	보안성	• 보안성 인증 • 사용자 인증 • 상호작용 데이터 보안 • 보안 이벤트 사용자 제공
	프라이버시 보호	• 개인정보보호 영향평가 • 개인정보 수집고지 • 개인정보 가명화 • 개인 민감정보 수집차단 • 개인정보 활용

4. 설문조사 분석과 연구결과

4.1 설문조사 방법과 대상

설문조사 문항은 FGI에서 도출 측정지표 항목과 델파이 설문조사에서 선정된 측정지표 항목을 확정하여 설문지를 작성하였다. 설문 응답자가 쉽게 이해할 수 있도록 윤리원칙과 구성요소에 대한 설명을 먼저 기술하였고, 측정지표에 대한 항목 정의를 명확하게 제시하여 답변할 수 있도록 하였다. 측정도구는 Likert 5점 척도를 사용(전혀 아니다, 아니다, 보통이다, 그렇다, 매우 그렇다) 하여 구성하였다. 설문지 설계는 인공지능 윤리원칙 구성요소와 구성요소에 포함된 측정지표 간의 관계가 성립되는지에 검증할 수 있도록 설계되었다. 설문조사 후 측정지표 항목을 변수로 하여 신뢰도 분석, 상관관계 분석, 측정지표별 기술통계량을 분석할 수 있도록 하였다. 설문조사는 인공지능 분야 종사자를 대상으로 총 80명을 선정하여 설문지를 배포하고 회수된 58부 중 응답이 부실한 3부를 제외한 55부를 분석하였다.

4.2 인구통계학적 특성과 기술통계량 분석

설문 분석에 사용한 표본 수는 총 55명이며, 남성 42명, 76.0%이며, 여성 13명, 24.0%가 설문에 응답하였다. 응답자의 인구통계학적 특성은 표 6과 같이 조사되었다.

설문에 응답한 인공지능 관련된 이해관계자는 AI 윤리 관련자 8명으로 14.5%, AI 관련 법률가 6명으로 10.9%, AI 교육자 8명으로 14.5%, AI 설계자(컨설턴트 포함) 11명으로 20.0%, AI 개발자 9명으로 16.4%, AI 제작자(기업) 4명으로 7.3%, AI에 많은 관심을 가지고 있는 AI 사용자 9명으로 16.4%로 나타났다. 인공지능 기술 발전이 최근에 급성장하고 있기 때문에 인공지능 관련 업무 수행과 관심이 많은 연령층은 30대, 40대가 주류를 이루어 있다.

인공지능 윤리 측정지표 항목에 대한 기술통계량 분석 결과는 표 7과 같이 산출되었다. 분석에 적용된 표본 55개는 전체 유효하게 사용되었다. 기술통계량 분석 결과는 최솟값은 1이며, 최댓값은 5로 산출되었다. 평균값이 3.0 이하는 3개 항목이었으며, 평균값이 3.0 이상은 40개 항목

(표 6) 인구통계학적 특성
(Table 6) Demographic characteristics

구 분		빈도수(명)	구성비(%)
성 별	남	42	76.4
	여	13	23.6
	합 계	55	100
연 령	20세 ~ 29세	3	5.45
	30세 ~ 39세	18	32.73
	40세 ~ 49세	23	41.82
	50세 ~ 60세	11	20.0
	합 계	55	100
직 업	AI 윤리	8	14.5
	AI 법률가	6	10.9
	AI 교육자	8	14.5
	AI 설계자	11	20.0
	AI 개발자	9	16.4
	AI 제작자	4	7.3
	AI 일반사용자	9	16.4
	합 계	55	100

으로 산출되었다.

(표 7) 인공지능 윤리 측정지표 기술통계량 분석
(Table 7) Descriptive statistics analysis of AI Ethics Measurement indicators

윤리원칙 구성요소	측정지표 항목	분석수	최솟값	최댓값	평균	표준편차
설명 가능성	최종 사용자를 위한 투명성	55	2	5	4.33	.747
	사고조사자를 위한 투명성	55	2	5	4.29	.916
	변호사 및 전문가 증인을 위한 투명성	55	1	5	4.44	.977
	이용 범위 고지 투명성	55	2	5	4.40	.807
	이용자 데이터 수집 투명성	55	2	5	4.13	.944
추적 가능성	문제 분석을 위한 데이터 제공	55	3	5	4.29	.685
	제작 과정의 기록 보관	55	2	5	4.27	.870
	사고 발생 이력 보관	55	1	5	4.45	.978
준법성	법·제도 준수	55	1	5	4.45	.919
	개인정보 제공 동의 획득	55	2	5	4.42	.786
	데이터 사용 사전 허가 획득	55	2	5	4.15	.951
책임 식별성	설계자 책임식별	55	2	5	4.31	.879
	제작자 책임식별	55	2	5	4.40	.852
	운영자 책임식별	55	2	5	4.45	.789
	이용자 책임식별	55	2	5	4.13	.982
알고리즘 차별성	알고리즘의 의사결정 설명	55	1	4	2.18	.475
	의도적 차별 요소 제한	55	2	5	4.49	.717
	개인 편향적 차별 제한	55	2	5	4.42	.832
	불공정한 알고리즘	55	2	5	4.55	.789
	인간의 생명 위협 방지	55	1	5	4.51	.836
데이터 편향성	편향적 특성 제거	55	2	5	4.35	.865
	데이터의 사실성	55	1	5	4.40	.852
	학습용 데이터 설명	55	1	3	2.11	.567
	학습용 데이터의 수량	55	2	5	4.44	.764
	데이터의 유효성 검사	55	2	5	4.09	.908
	학습용 데이터 품질관리	55	2	5	4.25	.844
	데이터 편향 방지 알고리즘 적용	55	1	5	4.49	.920
제어 가능성	정책적 통제권 이양	55	1	5	4.47	.858
	이용자 제어장치 제공	55	2	5	4.07	.920
	생명 위협 시 자동 종료	55	1	5	4.44	.856
	운영자 통제 및 제어	55	2	5	4.15	.951
견고성	인간의 보호 우선	55	1	5	4.44	.856
	알고리즘의 오류 관리	55	2	5	4.31	.858
	외부공격에 대한 견고성	55	1	5	4.51	.858
보안성	보안성 인증	55	2	5	4.51	.767
	사용자 인증	55	2	5	4.02	.892
	상호작용 데이터 보안	55	2	5	4.31	.879
	보안 이벤트 사용자 제공	55	1	3	2.07	.378
프라이버시보호	개인정보보호 영향 평가	55	1	5	4.45	.919
	개인정보 수집고지	55	2	5	4.45	.789
	개인정보 가명화	55	2	5	4.09	.908
	개인 민감정보 수집차단	55	3	5	4.33	.695
	개인정보 활용	55	2	5	4.29	.875

4.3 인공지능 윤리 측정지표 신뢰도 분석

인공지능 윤리 측정지표에 대한 신뢰도 분석은 표 8과 같이 산출되었다. 윤리 측정지표를 측정변수로 설정하여 분석하였으며, 신뢰도 분석을 위한 변수 문항에 내적 일관성을 나타낸 크론바흐 알파 계수 (Cronbach α)를 산출하였다. 표 8에 제시된 바와 같이 Cronbach α가 0.8 이상으로 높은 신뢰계수를 나타냄으로써 측정지표에 대한 측정항목의 신뢰도에는 문제가 없는 것으로 나타났다.

(표 8) 인공지능 윤리 측정지표 신뢰도 분석
(Table 8) Reliability analysis of AI Ethics Measurement indicators

윤리원칙 구성요소	측정지표	Alpha if Item Deleted	Cronbach α
설명 가능성	최종 사용자를 위한 투명성	.847	.882
	사고조사자를 위한 투명성	.875	
	변호사 및 전문가 증인을 위한 투명성	.840	
	이용 범위 고지 투명성	.863	
	이용자 데이터 수집 투명성	.855	
추적 가능성	문제 분석을 위한 데이터 제공	.820	.852
	제작 과정의 기록 보관	.743	
	사고 발생 이력 보관	.812	
준법성	법·제도 준수	.730	.850
	개인정보 제공 동의 획득	.767	
	데이터 사용 사전 허가 획득	.872	
책임 식별성	설계자 책임식별	.828	.870
	제작자 책임식별	.830	
	운영자 책임식별	.834	
	이용자 책임식별	.843	
알고리즘 차별성	알고리즘의 의사결정 설명	.867	.804
	의도적 차별 요소 제한	.750	
	개인 편향적 차별 제한	.756	
	불공정한 알고리즘	.708	
	인간의 생명 위협 방지	.693	
데이터편향성	편향적 특성 제거	.847	.864
	데이터의 사실성	.851	
	학습용 데이터 설명	.915	
	학습용 데이터의 수량	.872	
	데이터의 유효성 검사	.866	
	학습용 데이터 품질관리	.855	
	데이터편향 방지 알고리즘적용	.844	
제어 가능성	정책적 통제권 이양	.888	.912
	이용자 제어장치 제공	.890	
	생명 위협 시 자동 종료	.889	
	운영자 통제 및 제어	.877	
견고성	인간의 보호 우선	.815	.872
	알고리즘의 오류 관리	.890	
	외부공격에 대한 견고성	.746	
보안성	보안성 인증	.751	.739
	사용자 인증	.626	
	상호작용 데이터 보안	.612	
	보안 이벤트 사용자 제공	.811	
프라이버시보호	개인정보보호 영향평가	.803	.901
	개인정보 수집고지	.876	
	개인정보 가명화	.892	
	개인 민감정보 수집차단	.881	
	개인정보 활용	.877	

4.4 인구지능 윤리 측정지표 상관관계분석

본 연구에서는 인공지능 윤리원칙의 구성요소별로 개발된 윤리 측정지표가 각 구성요소 내에서 어느 정도 상관관계를 보여주고 있는지 검증하는 하였다. 윤리원칙의 구성요소에 포함된 측정지표가 변수이며, 변수들 간의 관련성을 분석, 검증하였다. 두 변수 간의 관계 정도를 파악하고, 통계학적 의미의 상관관계를 상관계수의 수준에 의하여 상관관계 유의 여부를 판단하였다[30][31]. 상관관계분석은 인공지능 윤리원칙의 구성요소에 포함된 측정지표를 변수로 하여 분석하였다. 상관관계분석 결과 인공지능 윤리원칙의 각 구성요소에 포함된 측정변수의 상관계수는 모두 0.01 유의수준 하에 유의한 것으로 분석되었다.

먼저 인공지능 윤리 투명성 원칙의 설명 가능성 구성요소에 포함된 5개의 윤리 측정지표에 대한 상관관계분석 결과는 표 9와 같다.

(표 9) 설명가능성 상관관계분석 결과
(Table 9) Correlation Analysis of Explainability

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)				
			1	2	3	4	5
최종 사용자를 위한 투명성	4.33	.747	1.00				
사고조사자를 위한 투명성	4.29	.916	.589*	1.00			
변호사 및 전문가 증인을 위한 투명성	4.44	.977	.664*	.600*	1.00		
이용 범위 고지 투명성	4.40	.807	.557*	.441*	.737*	1.00	
이용자 데이터 수집 투명성	4.13	.944	.728*	.556*	.601*	.564*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

설명 가능성에 포함된 측정지표 상관관계분석 결과 5개 변수 모두 상관계수가 0.441 ~ 0.728로 다소 높은 상관관계로 분석되었다.

인공지능 윤리 투명성 원칙의 추적 가능성 구성요소에 포함된 3개의 윤리 측정지표에 대한 상관관계분석 결과는 표 10과 같다.

(표 10) 추적가능성 상관관계분석 결과
(Table 10) Correlation Analysis of Traceability

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)		
			1	2	3
문제 분석을 위한 데이터 제공	4.29	.685	1.00		
제작 과정의 기록 보관	4.27	.870	.703*	1.00	
사고 발생 이력 보관	4.45	.978	.628*	.700*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

추적 가능성에 포함된 측정지표 상관관계분석 결과 3개 변수 모두 상관계수가 0.628 ~ 0.703으로 다소 높은 상관관계로 분석되었다.

인공지능 윤리 책무성 원칙의 준법성 구성요소에 포함된 3개의 윤리 측정지표에 대한 상관관계분석 결과는 표 11과 같다.

(표 11) 준법성 상관관계분석 결과
(Table 11) Correlation Analysis of Compliance

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)		
			1	2	3
법·제도 준수	4.45	.919	1.00		
개인정보 제공 동의 획득	4.42	.786	.783*	1.00	
데이터 사용 사전 허가 획득	4.15	.951	.622*	.586*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

준법성에 포함된 측정지표 상관관계분석 결과 3개 변수 모두 상관계수가 0.586 ~ 0.783으로 다소 높은 상관관계로 분석되었다.

인공지능 윤리 책무성 원칙의 책임 식별성 구성요소에 포함된 4개의 윤리 측정지표에 대한 상관관계분석 결과는 표 12와 같다.

(표 12) 책임식별성 상관관계분석 결과
(Table 12) Correlation Analysis of Responsibility Identification

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)			
			1	2	3	4
설계자 책임식별	4.31	.879	1.00			
제작자 책임식별	4.40	.852	.623*	1.00		
운영자 책임식별	4.45	.789	.594*	.716*	1.00	
이용자 책임식별	4.13	.982	.683*	.580*	.593*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

책임 식별성에 포함된 측정지표 상관관계분석 결과 3개 변수 모두 상관계수가 0.594 ~ 0.716으로 다소 높은 상관관계로 분석되었다.

인공지능 윤리 공정성 원칙의 알고리즘 차별성 구성요소에 포함된 5개의 윤리 측정지표에 대한 상관관계분석 결과는 표 13과 같다.

(표 13) 알고리즘 차별성 상관관계분석 결과
(Table 13) Correlation Analysis of Algorithmic differentiations

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)				
			1	2	3	4	5
알고리즘의 의사결정 설명	2.18	.475	1.00				
의도적 차별 요소 제한	4.49	.717	.059	1.00			
개인 편향적 차별 제한	4.42	.832	.132	.457*	1.00		
불공정한 알고리즘	4.55	.789	.126	.664*	.520*	1.00	
인간의 생명 위협 방지	4.51	.836	.042	.626*	.673*	.778*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

알고리즘 차별성에 포함된 측정지표 상관관계분석 결과 5개 중 알고리즘의 의사결정 설명 측정 변수는 4개 측정 변수와 상관계수가 0.042 ~ 0.132로 상관관계가 거의 없는 것으로 분석되었으며, 나머지 4개 측정 변수는 상관계수가 0.457 ~ 0.778로 다소 높은 상관관계로 분석되었다.

인공지능 윤리 공정성 원칙의 데이터 편향성 구성요소에 포함된 7개의 윤리 측정지표에 대한 상관관계분석 결과는 표 14와 같다.

(표 14) 데이터 편향성 상관관계분석 결과
(Table 14) Correlation Analysis of Data bias

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)						
			1	2	3	4	5	6	7
편향적 특성 제거	4.35	.865	1.00						
데이터의 사실성	4.40	.852	.663*	1.00					
학습용 데이터 설명	2.11	.567	.148	.061	1.00				
학습용 데이터의 수량	4.44	.764	.496*	.637*	.059	1.00			
데이터의 유효성 검사	4.09	.908	.645*	.574*	.160	.475*	1.00		
학습용 데이터 품질관리	4.25	.844	.917*	.603*	.173	.485*	.549*	1.00	
데이터 편향 방지 알고리즘 적용	4.49	.920	.713*	.926*	.073	.690*	.588*	.647*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

데이터 편향성에 포함된 측정지표 상관관계분석 결과 7개 중 학습용 데이터 설명 측정 변수는 6개 측정 변수와 상관계수가 0.059 ~ 0.173으로 상관관계가 거의 없는 것으로 분석되었으며, 나머지 6개 측정 변수는 상관계수가 0.475 ~ 0.917로 다소 높거나 매우 높은 상관관계로 분석되었다.

인공지능 윤리 통제성 원칙의 제어 가능성 구성요소에 포함된 4개의 윤리 측정지표에 대한 상관관계분석 결과는 표 15와 같다.

(표 15) 제어가능성 상관관계분석 결과
(Table 15) Correlation Analysis of Machine controllability

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)			
			1	2	3	4
정책적 통제권 이양	4.47	.858	1.00			
이용자 제어장치 제공	4.07	.920	.589*	1.00		
생명 위협 시 자동 종료	4.44	.856	.950*	.594*	1.00	
운영자 통제 및 제어	4.15	.951	.641*	.940*	.626*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

제어 가능성에 포함된 측정지표 상관관계분석 결과 4개 변수 모두 상관계수가 0.594 ~ 0.950으로 다소 높거나 매우 높은 상관관계로 분석되었다.

인공지능 윤리 안전성 원칙의 견고성 구성요소에 포함된 3개의 윤리 측정지표에 대한 상관관계분석 결과는 표 16과 같다.

(표 16) 견고성 상관관계분석 결과
(Table 16) Correlation Analysis of Robust

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)		
			1	2	3
인간의 보호 우선	4.44	.856	1.00		
알고리즘의 오류 관리	4.31	.858	.595*	1.00	
외부공격에 대한 견고성	4.51	.858	.802*	.688*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

견고성에 포함된 측정지표 상관관계분석 결과 3개 변수 모두 상관계수가 0.595 ~ 0.802로 다소 높거나 높은 상관관계로 분석되었다.

인공지능 윤리 안전성 원칙의 보안성 구성요소에 포함된 4개의 윤리 측정지표에 대한 상관관계분석 결과는 표 17과 같다.

(표 17) 보안성 상관관계분석 결과
(Table 17) Correlation Analysis of Security

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)			
			1	2	3	4
보안성 인증	4.51	.767	1.00			
사용자 인증	4.02	.892	.609*	1.00		
상호작용 데이터 보안	4.31	.879	.586*	.583*	1.00	
보안 이벤트 사용자 제공	2.07	.378	.118	.051	.154	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

보안성에 포함된 측정지표 상관관계분석 결과 4개 중 보안 이벤트 사용자 제공 측정 변수는 3개 측정 변수와 상관계수가 0.051 ~ 0.154로 상관관계가 거의 없는 것으로 분석되었으며, 나머지 3개 측정 변수는 상관계수가 0.586 ~ 0.609로 다소 높은 상관관계로 분석되었다.

마지막으로 인공지능 윤리 안전성 원칙의 프라이버시 보호 구성요소에 포함된 5개의 윤리 측정지표에 대한 상관관계분석 결과는 표 18과 같다

(표 18) 프라이버시 보호 상관관계분석 결과
(Table 18) Correlation Analysis of Privacy protection

측정지표	평균	표준편차	구성개념 간 상관관계 (Inter-Construct Correlations)				
			1	2	3	4	5
개인정보보호 영향평가	4.45	.919	1.00				
개인정보 수집고지	4.45	.789	.807*	1.00			
개인정보 가명화	4.09	.908	.615*	.613*	1.00		
개인 민감정보 수집차단	4.33	.695	.632*	.601*	.626*	1.00	
개인정보 활용	4.29	.875	.684*	.609*	.595*	.754*	1.00

* 상관계수는 0.01 수준(양쪽)에서 유의

프라이버시 보호에 포함된 측정지표 상관관계분석 결과 4개 변수 모두 상관계수가 0.595 ~ 0.807로 다소 높거나 높은 상관관계로 분석되었다.

4.5 연구결과

인공지능 기술이 급속하게 발전하면서 인공지능 윤리에 대한 연구도 활발하게 진행되고 있으나 아직 인공지능 산업에 윤리를 적용하는 사례가 매우 부족하다[5]. 이러한 시점에 본 연구는 인공지능 윤리 원칙과 구성요소 기반 하에 국내외 인공지능 윤리 관련 가이드라인, 정책, 개발지침, 논문 등 다양한 문헌을 바탕으로 윤리적 쟁점에서 자료 분석과 3차에 거쳐 FGI, 2회의 델파이 설문조사를 통하여 최종적으로 윤리 측정지표를 도출하였다. 최종적으로 도출된 윤리 측정지표를 대상으로 설문조사를 실시하여 인공지능 윤리원칙의 구성요소에 포함되는 윤리 측정지표에 대한 신뢰도 분석과 상관관계 분석을 통하여 유의한 윤리 측정지표를 개발하였다.

FGI에서 최종적으로 도출된 윤리 측정지표는 43개 항목이었으나 본 설문조사에 의하여 상관관계 분석 결과 3개 항목 (알고리즘의 의사결정 설명, 학습용 데이터 설명, 보안 이벤트 사용자 제공)은 해당 구성요소에 포함된 측정변수와 상관관계가 거의 없어서 제거하고 40개 항목에 대하여 윤리 측정지표를 개발하여 표 19와 같이 제시하였다.

(표 19) 인공지능 윤리 측정지표
(Table 19) AI Ethics Measurement indicators

윤리원칙	구성요소 (항목 수)	측정지표
투명성	설명가능성 (5)	최종 사용자를 위한 투명성
		사고조사자를 위한 투명성
		변호사 및 전문가 증인을 위한 투명성
		이용 범위 고지 투명성
		이용자 데이터 수집 투명성
	추적가능성 (3)	문제 분석을 위한 데이터 제공
		제작 과정의 기록 보관
		사고 발생 이력 보관
책무성	준법성 (3)	법·제도 준수
		개인정보 제공 동의 획득
		데이터 사용 사전 허가 획득
	책임식별성 (4)	설계자 책임식별
		제작자 책임식별
		운영자 책임식별
		이용자 책임식별
공정성	알고리즘 차별성 (4)	의도적 차별 요소 제한
		개인 편향적 차별 제한
		불공정한 알고리즘
		인간의 생명 위협 방지
	데이터 편향성 (6)	편향적 특성 제거
		데이터의 사실성
		학습용 데이터의 수량
		데이터의 유효성 검사
		학습용 데이터 품질관리
		데이터 편향 방지 알고리즘 적용
통제성	제어가능성 (4)	정책적 통제권 이양
		이용자 제어장치 제공
		생명 위협 시 자동 종료
		운영자 통제 및 제어
안전성	견고성 (3)	인간의 보호 우선
		알고리즘의 오류 관리
		외부공격에 대한 견고성
	보안성 (3)	보안성 인증
		사용자 인증
		상호작용 데이터 보안
	프라이버시 보호 (5)	개인정보보호 영향평가
		개인정보 수집고지
		개인정보 가명화
		개인 민감 정보 수집차단
		개인정보 활용

5. 결 론

메타버스, 클라우드, 빅데이터, 지능형 로봇, 헬스케어 로봇, 자율주행 자동차 등의 4차 산업 기술에서 인공지능은 필수적인 요소로 정착되었다. 이제는 인공지능 역기능에서 인간의 인격권과 재산 보호가 가능하도록 인공지능 윤리를 구체적으로 개발하여 인공지능에 적용하여야 한다.

본 연구에서 인공지능 윤리원칙과 구성요소를 기반으로 보다 구체적인 40개 항목의 인공지능 윤리 측정지표를 개발하여 제안하였다. 현재까지 만들어진 인공지능 윤리원칙 수준에서는 급속하게 발전하는 인공지능 환경에 적용하기 어려웠으나 본 연구에서 개발된 인공지능 윤리 측정지표는 우선적으로 인간의 인격권과 재산을 보호할 수 있도록 하는데 적용이 가능할 것이다.

인공지능 제품이나 서비스는 인간과 상호작용 시 안전과 신뢰를 보장할 수 있어야 한다. 본 연구에서 개발된 윤리 측정지표를 바탕으로 인공지능 설계자, 개발자, 사용자, 운영자, 제작자 및 다양한 이해관계자가 인공지능 개발과 연구, 활용에 적용할 수 있을 것이다. 또한 국내외 다양한 기관에서 인공지능 윤리에 대한 연구와 적용 방법을 개발하고 있기 때문에 본 연구에서 제시된 인공지능 윤리 측정지표는 인공지능 윤리 체크리스트 개발, 교육, 표준화, 인증, 개발지침, 인공지능 도입 기준 등 많은 분야에서 폭넓게 활용될 수 있을 것으로 본다. 향후 지속적인 연구를 통하여 인공지능 모든 영역에 윤리 측정지표를 개발, 적용하여 안전하고 신뢰할 수 있는 지능정보화사회가 되는데 본 연구가 기여할 수 있기를 기대한다.

참고문헌(Reference)

[1] NK Park, "Artificial Intelligence and Ethical Issues", Journal of Communication Research, vol.57, no.3, pp.122-154, 2020. http://doi.org/10.22174/jcr.2020.57.3.122

[2] Park H, Kim B, Kwon H, "Trends and Implications of Regulations for AI Control", Journal of Korea Information law, Vol. 25, No. 2, pp.1-39, 2021. https://doi.org/10.22846/kafil.25.2.202108.001

[3] SY Jho, "Review of a new protection system for personal right in an intelligent information society", Korean Public Law Association, 21(3), pp.109-129, 2020. http://doi.org/10.31779/plj.21.3.202008.004

[4] Adamson G, "Designing a value-driven future for ethical autonomous and intelligent systems", Proceedings of the IEEE 107 (3), pp.518 ~ 525, 2019.

[5] HT Yang, "Safety Issues of Artificial Intelligence and Policy Responses", The Journal of Korean Institute of Communications and Information Sciences, 43 (10), pp.1724-1732, 2018.

[6] MJ Kim, "Seoul PACT : Principles of Artificial Intelligence Ethics and its Application Example to Intelligent E-Government Service", Korea Society of IT Serviecs, 18. 3, pp.117-128, 2019.

[7] Ministry of Science and ICT, "Reliable artificial intelligence realization strategy", 2021.

[8] SJ So, SJ Ahn, "A Study on the Classification Model and Components of Artificial Intelligence Ethical Principles", The Korean Association Of Computer Education, 24(6), pp.119- 132, 2021. https://doi.org/10.32431/kace.2021.24.6.010

[9] AI HLEG, "Ethics Guidelines for Trustworthy AI", European Commission, 2019.

[10] Ministry of Science and ICT, "A study on the national policy for ethical Artificial Intelligence", 2020.

[11] TS Kwon, "Legal Review of the Protection System for Personal Rights", Press Arbitration Commission, Vol. 162, pp.4-17, 2022.

[12] HS Lim, "A Study on the Protection of Personal Rights in the Digital Age", Journal of Next-generation Convergence Technology Association, Vol.5, No.6, pp. 1246-1253, 2021.

[13] NIA, "Guidelines for Data Quality Management for AI Learning V1.0", 2021.

[14] European Commission, "The Assessment List for Trustworthy AI", Document made public on the 16th of July 2020.

[15] Carnegie Mellon University, "Designing Ethical AI Experiences: Checklist and Agreement", Software Engineering Institute, 2019.

[16] Stanford University HCAI, "Artificial Intelligence Index Report 2022", 2022.

[17] IEE SA, "IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being", IEEE Std 7010TM-2020, 2020.

[18] SY Byun, "A Study on the Ethics Certification Program Based on the Morality Types of AI Robots", Journal of Korean Ethics Studies, vol.1, no.126, pp.73-90, 2019. http://doi.org/10.15801/je.1.126.201909.73

[19] WS Jung, "Discrimination and Bias of Artificial Intelligence", Institute of Human, Environment & Future. vol., no.25. pp.55-73, 2020. http://doi.org/10.34162/hefins.2020..25.003

[20] Cowls. J, Floridi. L, "Prolegomena to a White Paper on an Ethical Framework for a Good AI Society", SSRN Electronic Journal, 2019. http://doi.org/10.2139/ssrn.3198732

[21] Erik, b, Daniel. R, Chad. S, "Artificial intelligence and the modern productivity paradox: A Clash of expectations and statistics", NBER Working Paper 24001(1), 2017. http://doi.org/10.3386/w24001.

[22] ES jang, JH Kim, "Development of Artificial Intelligence Education Contents based on Tensorflow for Reinforcement of SW Convergence Gifted Teacher Competency", Journal of Internet Computing and Services, Vol.20, No.6, pp.167 - 177, 2019. https://doi.org/10.7472/JKSII.2019.20.6.167

[23] Ministry of Science and ICT, "National AI Ethics Standards", 2020.

[24] Ministry of Land, Infrastructure and Transport, "Autonomous Vehicle Ethics Guidelines", 2021.

[25] IBM, "Trusted AI", 2021. https://research.ibm. com/teams/trusted-ai.

[26] Google, "Developer Policy Center", 2021. https://play.google.com/intl/ko/about/developer-content-policy

[27] Lee. J, Kim. D, Yang. H, "A Prospective Analysis of Artificial Intelligence(AI) Technology and Innovation Policies(Year 2)", STEPI, 2019.

[28] GS Lee, SJ Ahn, "A study on the environmental factors and detailed measurement items to be considered in establishing integrated information system in higher education institutions", Journal of Internet Computing and Services, Vol.14, No.3, pp.57 - 65, 2013. https://doi.org/10.7472/JKSII.2013.14.3.57

[29] KS Lee, SJ Ahan, "The Comparing Research of Portal Companies's code of ethics for Concrete Social Responsibility", The Journal of Korean association of computer education. Vol. 16, No. 1, pp.118-121, 2011.

[30] Falk. R, Well, "Many Faces of the Correlation Coefficient", A Journal of Statistics Education. Vol.5, No.3, 1997.

[31] TJ Sung, "Understanding and application of modern basic statistics", 598, Hakjisa, 2019.

◐ 저 자 소 개 ◑

소 순 주(Soonju So)
1995년 광주대학교 컴퓨터학과((공학사)
2014년 성균관대학교 대학원 IT컨설팅학과(공학석사)
2018년 성균관대학교 대학원 컴퓨터교육과(박사수료)
2016년~현재 (주)코어소프트 대표이사
관심분야 : 인공지능, 인공지능윤리, 정보보안, 컴퓨터교육, etc.
E-mail : SSJLHD@nate.com

안 성 진(Seongjin Ahn)
1988년 성균관대학교 정보공학과(학사)
1990년 성균관대학교 대학원 정보공학과(석사)
1998년 성균관대학교 대학원 정보공학과(박사)
1990년~1995년 KIST/SERI 연구원
1996년 정보통신기술사
1999년 3월~현재 성균관대학교 컴퓨터교육과 교수
관심분야 : 네트워크관리, 산업보안, SW/AI교육, AI 윤리., Etc.
E-mail : sjahn@skku.edu