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

AI 로봇의 위험성 예방을 위한 측정지표에 관한 연구

A study on the measurement indicators for risk prevention of AI robots

송 현 경[1] 안 성 진[1*]
Hyun-kyoung Song Seongjin Ahn

요 약

감정 휴머노이드 등 AI 로봇의 시대가 다가오면서 로봇의 차별 발언이나 의료 사고 등 예측 못 하는 위험이 있지만, 관련 연구는 매우 부족하다. 이 논문은 안전한 로봇 산업화를 위해 꼭 필요한 AI 로봇의 위험성 측정지표를 제안한 연구이다. AI 로봇의 위험성은 인공지능의 문제와 로봇의 문제를 예측해야 하는 고난도 연구이므로 AI・로봇 개발자, 관련 분야 교수・연구원 등 전문가 30명에게 델파이 기법으로 설문조사를 2회 진행하였다. 2회 통계 분석으로 내용타당도와 신뢰도를 확보한 4개 위험성 유형과 11개의 항목 및 37개 측정지표를 도출하였다. 이 연구에서 제안하는 측정지표를 AI 로봇의 시험평가・인증・교육 등 분야에 활용하여 점검한다면, 인간과 AI 로봇 상호 간 안전이 확보된 생활을 함께 영위할 수 있을 것이다. 이 연구의 측정지표가 초석이 되어, AI 로봇을 안전하게 사용할 수 있도록 향후 정부・산업・교육・의료・가정 등 다양한 분야의 AI 로봇 정책 및 연구 등에 활발하게 활용되길 기대한다.

☞ 주제어 : AI 로봇, 인공지능 로봇, AI 로봇 위험성, AI 로봇의 위험성 예방, AI 로봇의 위험성 측정지표, 인공지능 로봇 측정지표

ABSTRACT

As the era of AI robots such as emotional humanoids approaches, there are unpredictable risks such as robot discrimination remarks and medical accidents, but related research is very insufficient. This paper proposes a risk measurement index for AI robots that is essential for safe robot industrialization. Since the risk of AI robots is a high-level study that must predict artificial intelligence and robot problems, 30 experts, including AI and robot developers and professors and researchers in related fields, were surveyed twice using Delphi techniques. Through two statistical analyses, four risk types, 11 items, and 37 measurement indicators were derived that secured content validity and reliability. If the measurement indicators proposed in this study are checked by using them in fields such as test evaluation, certification, and education of AI robots, humans and AI robots will be able to live a safe life together. It is hoped that the measurement index of this study will become a cornerstone and be actively used in AI robot policies and research in various fields such as the government, industry, education, medical care, and home so that AI robots can be used safely.

☞ keyword : AI robot, artificial intelligence robot, AI robot risk, AI robot risk prevention, AI robot risk measurement indicators, AI robot measurement indicators

1. 서 론

로봇 바리스타가 커피를 내려주고, 로봇이 테이블로 음식을 배달하는 건 일상적 광경이며, 아이 로봇 영화의 NS-5처럼 요리 및 심부름 전담 휴머노이드와 사는 삶도 머지않아 보인다[1]. 로봇의 발전은 전격-Z 작전 드라마처럼 손목시계로 자동차 키트를 호출해 원하는 곳으로 이동하는 등 생활 속 편리를 더해줄 수도 있지만, 디트로이트 비컴 휴먼 게임 속처럼 형사에게 추격당하는 안드로이드가 자신의 존재를 위해 거짓을 말하기도 하고 권리를 주장하는 시위 상황 등 갈등이 나타날 수도 있다[1-2].

AI 로봇이 인간의 생활 속에서 공존하는 사회가 도래할수록 대화형 챗-봇이 인간에게 한 차별 발언 등 논란은 SF 영화나 게임 속 같은 위험한 일들을 이젠 현실에서도 마주할 수 있을 것이라는 뒷받침이 되는 근거가 된다.

이러한 이유로 국내외에서도 관련 연구를 하고 정책도 만들지만, 주로 인공지능 윤리 원칙・법・교육 등에 한정돼 있다. 관계부처의 정책도 인공지능 분야에서 활발하게 논의 중이고, 로봇 분야는 국가 경쟁력에 기대는 부분이 있다 보니 산업화 등 개발에 주된 초점이 맞춰져 있다.

이렇듯 현재까지도 AI 로봇의 위험성을 구체적으로 측정하기 위한 지표에 관한 연구는 매우 부족한 실정이다.

최근 지능형로봇법과 도로교통법이 개정돼서 로봇이 도로에서 주행할 수 있는 근거가 마련되고 연내 주행도 가능해져 로봇과 관련된 사고 등의 위험성을 예방하기

1 Dept of Computer Education, Sungkyunkwan University, Seoul, 03063, Korea.
* Corresponding author (sjahn@skku.edu)
[Received 24 April 2023, Reviewed 29 April 2023, Accepted 22 May 2023]

위한 연구가 더욱 활발하게 이루어져야 할 시점이다[3].

1942년 소설에서 아시모프 작가는 로봇이 인간을 해쳐서는 안 된다는 3원칙을 통해 로봇의 위험성을 언급했다.

81년 지난 현재 로봇은 인공지능과 결합하면서 인간의 신체를 해치는 것 외에도 물건을 망가뜨리거나 도로에서 교통사고를 일으키고, 인간의 감정을 상하게도 한다.

앞으로 여러 분야에서 로봇 개발이 더욱 활발해지고, 여러 형태 로봇과 관련된 위험성이 나타나게 될 것이다.

그간 발생했던 로봇과 관련된 사고 사례를 분석하면서 로봇이 인간에 대한 직접적 위험 외에도 인간을 차별해 인간의 권리를 침해할 위험, 재물손괴 또는 사회질서에 혼란을 일으킬 여러 위험 요소를 확인하였다.

이에 AI 로봇의 위험성 측정지표 연구를 시작하였다.

이 연구는 문헌 자료 분석・전문가 심층 면접 진행・델파이 기법을 활용한 연구 방법으로 AI 로봇의 위험성 분류 유형과 항목・측정지표를 도출하였다. 이후 델파이 설문조사 결과를 통계 분석하여 AI 로봇의 위험성 예방을 위한 측정지표를 개발하고, 검증 및 제시하고자 한다.

2. 이론적 배경

2.1 AI 로봇의 종류 및 정의

로봇은 1960년대 산업 분야에서 활용되기 시작하면서, 2000년 이후 IT와 접목한 지능형 로봇 출현으로 제조업 등 여러 응용 분야로 확장된 서비스 로봇으로 발전됐다[4].

로봇은 인간 사회와 생활 영역 전반에 편리함을 줄 수 있어 계속해서 발전해 나갈 것이다. 언젠가는 드론 로봇 또는 트랜스포머 영화처럼 변신 로봇을 타고 어디로든 이동할 수도 있고, 손목시계형 로봇이 인간의 감정이나 건강 상태를 수시로 점검해서 적절한 처방이나 치료를 해주는 개인 전담 의사가 되는 시대가 올 수도 있을 것이다.

이처럼 AI 로봇은 인공지능 소프트웨어를 탑재하는 등 여러 형태로 만들어질 수 있어 정부가 발간한 지능형 로봇과 AI 로봇 관련 문헌을 참조하여 미래에 나타날 것으로 예상되는 종류들을 정리하였으며, 아래 표 1과 같다.

앞으로는 위의 표 1보다 훨씬 더 많은 종류의 로봇이 나타날 것이다. 이 연구는 인간 사회생활의 편리를 위해 공존하게 될 여러 형태 AI 로봇의 범용-적 위험성을 측정하기 위한 지표의 연구로 활용 목적이 다른 전쟁 등 군사 및 특수 목적 로봇은 논외로 함을 전제하고 연구하였다.

인공지능이란 단어는 뉴햄프셔 다트머스 대학 매카시 박사가 1955년 학회에서 말한 Artificial Intelligence라는

(표 1) AI 로봇의 종류〔1-2〕〔4-7〕
(Table 1) Types of AI Robots

구분	종류
의료	수술 로봇, 재활 로봇, 간호 로봇, 정신 상담 로봇, 웨어러블 슈트 로봇
제조	커피 제조 바리스타 로봇, 협동 작업용 로봇, 부품 조립 로봇
운반	중량물 운반용 로봇, 배달 로봇
물류	제조 공장 · 공항 물류센터 · 상급병원 · 시설 물류 작업 로봇
돌봄	보행 치료 로봇, 이동 보조 로봇, 배설 보조 로봇, 치매 예방 로봇
공공	시설 안내 로봇, CCTV 순찰 로봇, 방역 로봇, 산불 감지 드론 로봇, 쓰레기 분류 로봇, 분실물 보관 로봇, 교정 로봇, 음주단속 로봇
법률	재판 관련 판사·로봇, 변호사 로봇
안내	상품 설명 로봇, 백화점 안내 로봇, 시각 장애인 안내 로봇, 자율주행 키오스크 안내 로봇
스포츠・오락	심판 로봇, 게임 로봇, 퀴즈 로봇, 농구공 던지기 로봇, 아바타 로봇
교육	언어 등 학습용 로봇, 교육 보조 로봇, 자폐 스펙트럼 · 언어장애 교육 로봇, 동화 구현 로봇
이동	자율주행 드론 비행 로봇, 자동차로 변신하는 로봇
감정	감정 인식 휴머노이드, 교감 로봇
가정	가사 도우미 로봇, 음식 조리 로봇, 육아 로봇, 심부름 로봇, 동반자 소셜 로봇, 청소 로봇, 홈 비서 로봇, 실버 케어 로봇, 반려 로봇

표현에서 비롯되었다[7]. 유럽 위원회 공동연구센터(European Commission, Joint Research Centre)는 인공지능을 복잡한 환경 속에서 데이터를 분석하고 인식으로 얻어진 정보와 학습된 지식을 기반으로 탐색을 거쳐 목표 달성에 가장 적합한 행동을 자율적으로 선택해 실행하도록 인간에 의해 고안된 소프트웨어나 하드웨어 시스템이라 하였다[8].

지능형 로봇 개발 및 보급 촉진법 제2조에는 지능형 로봇이란 외부 환경을 스스로 인식하고 상황을 판단하여 자율적으로 동작하는 기계 장치라 정의하였고, 인공지능 로봇을 소프트웨어나 소프트웨어와 결합 된 형태로 구현돼 외부 환경을 스스로 인식하고 상황을 판단함으로써 자율적으로 동작하는 기계 또는 장치라고 정의한 연구도 있다[9-10]. 이 연구에서는 인간 사회생활에서 나타날 수 있는 다양한 위험성을 전제하여, 인공지능 소프트웨어를

장착한 기계 장치가 외부요인이나 스스로 판단해서 반응하는 것을 AI 로봇이라고 정의하였다.

2.2 AI 로봇의 문제 사례 및 관련 선행연구

2.2.1 로봇 관련 문제 사례

대화형 챗-봇의 차별적 발언이 논란이었는데, 대화형 로봇은 가정용・교육용・서비스용 로봇으로 만들어질 수 있고, 챗-봇 문제는 이제 AI 로봇의 문제가 될 수 있다.

AI 로봇과 관련된 문제가 어떤 것들이 있는지를 확인하기 위해 문헌 등 자료 조사하면서 접했던 봇 또는 로봇, 인공지능 프로그램과 관련해서 발생했던 대표적인 문제 사례들을 정리하였으며, 이는 아래 표 2와 같다.

(표 2) 봇 또는 로봇 관련 문제 사례[11-16]
(Table 2) Bot or robot related problem cases

구분	사례
생명・신체・감정 관련 문제	- 수술 중, 조절 실패로 신체 손상 - 체스 말로 오인, 손가락 부러뜨림 - 에어로졸 분사 오작동, 호흡 지장 - 인공지능과의 연애로 마음의 상처 - 출혈을 악화시키는 약 처방 충고 - 정신과 상담 챗-봇이 자살을 추천
개인 등 정보 관련 문제	- AI 시스템으로 행인 정보 수집 - 가족 간 대화를 명령으로 오인, 대화 녹음·제3자에게 임의 전송 - 가짜뉴스 등 조작된 정보를 제공
재산・재물・교통 관련 문제	- 알고리즘 오류, 더 비싸게 집 구매 - 자율주행 유도차선 인식 실패로 중앙분리대와 부딪혀, 파손
차별・명예 관련 문제	- AI가 사람을 고릴라로 분류 인식 - 챗-봇이 대화 중, 혐오・차별 발언 - AI가 면접 지원자 중 남성만 추천 - AI가 대머리를 축구공으로 인식

위의 봇 또는 로봇 관련 문제 사례들을 통해서 어떠한 위험성 유형이 있는지, 이러한 위험을 사전에 점검하기 위해서는 어떤 측정지표가 필요한지 확인할 수 있었다.

2.2.2 선행연구

AI 로봇 관련 문제를 보완하면서 혁신적 기술 개발을 위해서는 규제나 가이드-라인도 현실을 반영하면서 발전해야 한다. 이에 AI 로봇과 관련된 국내 논문을 살펴본바, AI 의료 사고 또는 자율주행 자동차의 교통사고 책임, AI 로봇의 법적 지위, 지능형 로봇의 전자 인격 부여 등 법과 책임 관련 연구에 편중된 것을 알 수 있었다[17-21].

해외 논문 중, 2014년~2019년 84개국 윤리 가이드라인 분석 연구는 인공지능을 주제로 해 로봇 문제를 다루지 못했고, 인간과 로봇 협업 시스템의 잠재적 위험성 식별 체크리스트 연구는 기존 기술 안전 표준과 차별성이 없어 활용 가능성이 미비하다는 한계를 인정한다[22-23].

유럽연합은 2014년 로봇 규제 가이드 도출 이후, 2019년 신뢰할 만한 AI 가이드-라인・2020년 AI 윤리 원칙 기반 점검 목록 등 로봇보다 인공지능 논의가 활발하다[24-25].

중국은 로봇의 법인격 부여・법적 해결 방안 등 로봇 산업화 정책을 모색하고, 미국도 2020년 인공지능의 규제 가이드-라인 발표 등 로봇 관련 정책은 미비하다[26-27].

최근 국내 연구들도 인공지능 행위자의 도덕적 고찰, 인공지능 로봇을 위한 윤리 가이드-라인, AI 데이터 윤리, AI 윤리 교육, 로봇 윤리, 인격권 및 재산 보호를 위한 AI 윤리 측정지표 등 윤리나 교육에 집중돼 있다[28-33].

국내 관계부처들도 인공지능 윤리 기준이나 자율주행 자동차 윤리 가이드-라인, 개인정보 자가 점검 윤리 체크-리스트 등 윤리를 통한 규제에만 전념하고 있다[34-36].

AI 업체들도 자체 제작 윤리 체크리스트로 점검하고, 가정용 헬스-케어 로봇 자가 점검표도 있지만 개발자에 한정돼 있다[37-39]. 선행연구를 통해 다양한 형태의 AI 로봇으로 인한 여러 가지 문제점을 사전 점검할 수 있는 측정지표의 연구는 매우 부족한 실정임을 알 수 있었다.

이 연구에서는 관련 문제 사례와 선행연구를 참고해 여러 분야에서 AI 로봇을 개발할 때, 인간 신체의 위험성 및 인간 권리에 대한 차별성, 재산이나 사회적 문제들이 발생하지 않게 AI 로봇 시험평가, 인증, 교육, 점검 등에 범용-적으로 활용할 수 있는 측정지표를 연구하였다.

3. AI 로봇의 위험성 유형 및 측정지표 연구

3.1 연구 방법

이 연구에서는 AI 로봇과 관련된 위험을 유형화하여 발생 가능성 있는 항목으로 구분하고 위험을 측정할 수 있는 지표에 대해 전문가 심층 인터뷰 및 델파이 설문조사 기법을 활용해 연구하였다[40]. 이 연구를 위해 AI 또는 인공지능 로봇의 사고 관련 문헌 등 자료를 2013년부터 2023년 1월까지 검색하였고, 76건의 사례를 찾았다. AI가 외부요인 및 스스로 오작동하여 인간 신체나 사회 등에 대한 위험이 발생할 수 있는 상황을 유형화하고, AI 로봇을

도입할 때 점검해야 하는 측정지표를 연구하였다.

전문가 심층 인터뷰로 도출된 측정지표에서 유의미한 결과를 도출하고자 전문성을 지닌 대상자를 선정하였다.

델파이 조사는 전문 지식이 있는 대상자를 최소 10명~35명 정도로 선정하는 것이 과학기술이나 사회과학 연구의 특수성을 적절히 반영한 지표 제시에 효과적이다[41].

지표가 측정하고자 하는 개념을 얼마만큼 적절하게 대표하는지 알기 위해서 정성적 분석 방법인 내용타당도 CVR 값을 기준으로 검증하였다. 미래 추상적인 개념이 적절하게 측정되도록 설문조사 문항을 설계하고 지표의 신뢰도와 내용타당도를 검증하는 통계 분석을 사용하여 결과를 도출하는 연구를 하였다[40].

3.2 전문가 심층 인터뷰(FGI)

AI 로봇의 위험성 측정지표 연구는 그간 발생한 문제 사례와 아직 일어나지는 않았지만, 예측되는 위험성을 측정하기 위한 지표 연구이므로, 연구 방법・설문조사 대상 선정이 중요하다. 이를 위해 전문가 심층 인터뷰를 진행하였고, 3차에 걸쳐 측정지표를 도출하기 위한 연구 방법 선정과 연구에 적합한 설문조사 참여 대상자 선정・설문조사 문항 구성을 위해 중점적으로 논의를 하였다.

이 연구를 위해 참여한 전문가 그룹은 표 3과 같다.

(표 3) 전문가 그룹 정보
(Table 3) Information of experts group

구분	분야	경력
전문가 A	로봇 개발자	11년
전문가 B	로봇 분야 연구원	26년
전문가 C	AI 윤리학자	33년
전문가 D	AI 개발 및 기획자	27년
전문가 E	컴퓨터 교육과 교수	8년

3.2.1 연구 방법 및 델파이 설문 대상 전문가 선정

1차 전문가 인터뷰는 면담으로 진행되었고, 연구자가 문헌 등 자료 조사로 정리한 AI 로봇의 종류와 정의를 설명하자, 챗 GPT가 빠르게 발전하는 것을 보면서 AI 로봇과 관련된 위험성 문제를 예방하도록 사전 점검할 수 있는 측정지표가 빨리 마련돼야 한다며 중요성을 인식하였다.

전문가들은 다양한 분야에서 AI 로봇이 만들어질 수 있다며, 연구의 방향을 범용-적으로 활용할 측정지표로 해 현재뿐만 아니라 향후 발생할 수 있는 위험성도 측정지표에 포함할 필요가 있다는 일치된 의견이 있었다.

앞으로 발생할 수 있는 위험성을 측정할 지표는 미래 예측 방식으로 문제를 추정해야 하며, 전문가의 직관을 동원해 미래 변화를 예측하면서 의견을 수집하고 구성원 의견 합의를 도출하는 연구 방법으로 사회과학 분야의 연구에서 활용되고 있는 델파이 조사 기법이 선정되었다.

2차 전문가 인터뷰에서는 전화・이-메일로 의견 수렴하였다. 미래 예측 연구방식에 적합한 대상자와 관련해서 전문가 A・D는 로봇 및 인공지능에 대한 개념이 정립된 개발자를 선정해야 한다는 의견이었다. 전문가 C는 관련 연구가 활발한 인공지능이나 법・윤리학자를 선정하는 게 좋다, 전문가 B는 여러 분야의 로봇 연구원이 참여하면 의견이 많이 나올 것 같아 연구에 도움이 될 것 같다고 하였다. 전문가 E는 전산과 통신의 문제도 있다며 보안・공학 전공자가 좋다 등 의견이 있었다.

2차 인터뷰에서 취합된 다양한 의견을 전문가들에게 공유하고, 인터뷰에서 나온 의견을 종합하여 여러 분야 전문가를 대상자로 선정하면 여러 방향에서 많은 의견이 나올 수 있어 미래 예측 연구에 적합하다고 합의하였다.

2차 전문가 의견대로 델파이 설문조사 대상자는 AI 로봇에 대한 개념 정립이 되어 있는 AI와 로봇 분야 전문가, 컴퓨터・기계・전자 등 공학, 정보보안・통신을 전공한 관련 경력자, AI나 로봇 개발자들로 구성하기로 정하였다.

델파이 설문조사를 위해 로봇진흥원의 도움으로 로봇 분야 전문가들을 선정하였고, 인공지능과 산업 간 융합 업체 M사, 인공지능 감정분석 서비스 등을 개발한 인공지능 개발 업체 C사, 공간 AI와 빅 데이터 전문 업체 B사 도움으로 AI 개발자, 관련 전공자를 선정하였다. 이외에도 AI 교육학자, AI 윤리학자 및 관련 분야 전공자, 관련 분야 교수 등을 설문조사 대상으로 선정하였다.

3.2.2 설문조사 내용 구성

3차 전문가 인터뷰는 앞서 합의된 델파이 조사를 위한 설문 문항을 구성하기 위해 연구자가 문헌 등 자료 조사로 정리한 AI 로봇 관련 문제 76개 사례를 하나씩 검토하는 방법으로 면담을 진행하였다.

먼저 사례들을 비슷한 유형으로 묶어나가면서 생명과 재산 등 위험성을 4개 유형으로 구분하였다.

중복되는 사례들은 합치고 전문가 의견을 반영하여 위험성 측정을 위해 필요한 11개 항목을 도출하였으며, 어떤 지표가 어떤 위험성 점검에 필요한지에 대해 심도 깊은 논의를 통해 48개 지표를 도출하였다.

특히, 전문가 B・C・E는 챗-봇에 대한 문제점이 지적되자 업체가 자체적으로 AI 윤리 체크-리스트를 만들어

점검했지만, 친밀한 관계에 맞춰 개발한 리스트여서 AI 윤리에서 필수적으로 거쳐야 할 일반적인 점검 항목을 모두 넣어 점검하고 완전한 상태로 재배포하였다고 보기에는 어려움이 있다는 의견이었다.

A·B·C·D·E 전문가 전원이 이 연구에서 도출된 측정지표들은 실용성과 범용성을 갖추고 있어 관련 산업 분야를 비롯해 범정부적으로 활용할 수 있는 의미 있는 지표가 될 것이라는 의견이었다.

3차 전문가 심층 인터뷰로 도출된 항목 및 측정지표로 델파이 설문조사 문항 작성 및 설문의 이해를 돕기 위해 AI 로봇의 종류와 정의를 설명하는 내용을 기술하였다.

설문조사 문항마다 AI 로봇의 위험성 분류 유형과 각 항목·측정지표가 어떠한 의미인지를 기술하여 충분히 숙지한 후 답변할 수 있도록 설계하였다.

설문에 AI 로봇이란 AI 시스템이 소프트웨어와 결합된 형태로 외부 환경을 스스로 인식하고 상황을 판단해 자율적으로 반응해 작동하는 기계·장치라고 명시하였다.

항목-명 및 측정지표-명에 대한 수정 사항은 응답자의 검토 의견을 적을 수 있게 준 개방형 설문을 포함하였다.

설문조사 문항의 마지막에는 항목이나 지표를 추가할 필요가 있는지에 대해서 각 분야 전문가의 자유로운 의견을 작성할 수 있도록 개방형 작성-란을 추가하였다.

설문조사를 측정하기 위한 도구는 Likert 5점 척도를 사용하였다. 먼저 AI 로봇의 위험성 정의가 적절한지와 유형별 분류가 적절한지, 항목에 대한 정의를 확인하기 위해 5점 척도로 확인하였다.

또한, 측정지표들이 AI 로봇의 위험성을 예방하는 데 필요한 지표인지 아닌지 적절성 정도를 확인하도록 “1점 전혀 적절하지 않다, 2점 적절하지 않다, 3점 보통이다, 4점 적절하다, 5점 매우 적절하다”로 구성하였다.

이 연구의 측정지표 연구를 위해서 선정된 델파이 조사 기법에 적합한 설문조사 문항 내용인지 아닌지 완성도를 높이기 위해 추가 인터뷰를 하였다. 전문가 C와 E가 참여하여, 전체적으로 설문조사 문항의 구성 및 설계가 적절한지에 대해 2회 검토를 거쳐서 최종적으로 완성되었다.

4. 설문조사 분석 및 연구 결과

4.1 설문조사 대상 및 통계 분석 방법

설문조사는 전문가들의 의견 합의를 위해 2회에 거쳐 실시하였고, 전문가 심층 인터뷰를 통해 참여자로 선정된 AI와 로봇 분야 전문가, 보안, 윤리, 공학, 컴퓨터 분야의 경력자들에게 익명을 보장하고자 1차는 이-메일, 2차는 온라인 방식을 활용하여 설문조사 자료를 발송하였다.

최종 회수된 38부 중에서 전문가로 적합하지 않다고 판단되었던 최종 학력이 학사이면서 5년 미만의 근무자 5명과 똑같은 번호로 적어서 회신하는 등 응답이 부실한 3부를 제외하고 총 30부를 통계 분석 자료로 활용하였다.

이 연구는 델파이 조사 기법을 활용한 것으로 구성원 의견이 합의되는지를 확인할 수 있게 1차 조사와 동일한 대상자인 전문가 30명에게 델파이 2차 설문조사하였다.

이 연구는 AI 로봇 관련 위험성을 예방하는 데 필요한 측정지표 연구로 전문가 심층 인터뷰를 통해서 완성된 위험성 유형의 분류 및 항목과 측정지표의 내용이 타당한지와 신뢰성을 확보하는 통계 분석 방법이 사용되었다.

내용타당도는 CVR 값으로 측정하였고, 신뢰도는 Cronbach α값으로 측정하였으며, 재설문의 시행 여부는 전문가 간 의견이 합의됐다고 판단하는 안정도(변이계수) 값인 0.50을 기준으로 하였다[42-43].

4.2 델파이 설문 참여 전문가 인구 통계학적 특성

AI 로봇의 위험성 측정지표 연구는 AI 즉, 인공지능 및 로봇 분야 전문가들의 미래 예측을 통한 집단지성이 있어야 하는 고난도 연구이다. 따라서 석사 이상 전공자 또는 5년 이상 경력자들이 설문 대상자로 선정되었다.

이 연구를 위해 로봇진흥원 등의 도움으로 인공지능 전문가는 데이터 학습·개발·윤리·법·교육학자 등이 선정되었고, 로봇 분야는 기계·센서·통신 등 연구원이 선정되었으며, 관련 분야인 전자 공학, 정보보안·통신 교수 등도 참여하였다. 델파이 설문조사 참여 전문가는 30명이며, 델파이 설문에 참여한 전문가를 성별·연령·학력·직업·경력의 인구 통계학적 특성은 표 4와 같다.

4.3 1차 델파이 설문조사 결과

1차 델파이 설문조사 문항은 Likert 5점 척도와 준개방형 문항으로 구성하여 2023년 3월 27일부터 3월 31일까지 진행되었다. 참여자 30명은 모두 이-메일로 응답하였고 관련 분야의 전문가들인 만큼 AI 로봇에 대한 이해도가 높고 관심이 높은 연구 분야로 다양한 의견이 모였다.

자연 환경적 위험성이나 정신에 관한 위험성 지표를 추가하면 좋겠다는 소수 의견을 비롯하여 안내나 반려 로봇에 대한 사람들의 기대감이 낮아 모욕이 느껴지지 않는다며 로봇이 관람객에게 욕을 해도 사람들은 재밌게

(표 4) 델파이 참여 전문가의 인구 통계학적 특성(N=30)
(Table 4) Demographic characteristics of Delphi Participation Specialists(N=30)

구분		전문 분야				인원수	
		AI, 데이터	윤리, 보안	로봇, 기계	컴퓨터, 공학	명	%
성별	남	5	8	5	7	25	83.3
	여	3	1	-	1	5	16.7
	소계	8	9	5	8	30	100
연령	20대	-	2	-	-	2	10.0
	30대	1	-	-	3	4	13.3
	40대	6	5	2	2	15	46.7
	50대	1	2	3	3	9	30.0
	소계	8	9	5	8	30	100
학력	학사	4	4	-	5	13	33.3
	석사	2	2	-	3	7	23.3
	박사	2	3	5	-	10	43.3
	소계	8	9	5	8	30	100
직업	개발, 기획	4	4	-	4	12	40.0
	연구원	1	-	4	1	6	20.0
	교수	1	2	1	-	4	13.3
	기타 IT	2	3	-	3	8	26.7
	소계	8	9	5	8	30	100
직업 및 연구 경력	10년 이하	2	3	1	4	10	33.3
	20년 이하	5	2	2	1	10	33.3
	30년 이하	1	3	2	3	9	30.0
	30년 이상	-	1	-	-	1	3.3
	소계	8	9	5	8	30	100
참여		1차 설문				30	100
		2차 설문				30	100

여길 수 있어 위험성으로 보기 어렵다는 의견도 있었다.

내용타당성 검증 결과 기준에 부합하지 않은 항목의 CVR 0.333 미만(응답자 30명 기준)은 제거하였다[44].

응답자의 다수가 항목-명이나 지표-명에 대한 수정이 필요하다고 회신한 검토 의견과 추가로 도출된 항목과 지표는 2차 델파이 설문조사 문항에 반영하였다. 그리고 인간에 대한 위험성 유형은 권리를 포함하는 단어라고 판단된다며 신체에 대한 물리적 위험성 유형으로 명칭을 수정하는 게 좋겠다고 한 의견도 반영하였다. 지표-명을 반려동물 훼손이라고 명하기보다는 유형물 훼손이라고 수정하는 게 좋겠다는 의견도 반영하였다. 명칭 수정에 대한 검토 및 추가 의견을 모두 반영해 2차 설문 문항을 작성했다. 1차 델파이 설문조사 결과 중, AI 로봇의 위험성 유형에 대한 내용타당도 분석 결과는 아래 표 5와 같다.

(표 5) AI 로봇의 위험성 유형에 대한 내용타당도 분석 결과 및 선정 현황(CVR=0.333)
(Table 5) Content Feasibility Analysis Results and Selection Status for Risk Types of AI Robots(CVR=0.333)

유형	N	M	SD	CVR	선정
인간에 대한 위험	30	4.20	0.872	.667	○
인간 권리에 대한 위험	30	4.13	0.957	.467	○
재산 손실의 위험	30	4.10	0.978	.600	○
사회적 위험	30	4.30	0.690	.733	○

*N(유효 수), M(평균), SD(표준편차), CVR(내용타당도)

AI 로봇의 위험성 유형에 대한 내용타당도 분석 결과 CVR 기준값인 0.333을 상회하는 수치로 나타남에 따라 모든 유형의 항목이 선정되었으며, 유형별 항목에 대한 내용타당도 분석 결과는 아래 표 6과 같다.

(표 6) AI 로봇의 위험성 항목의 내용타당도 분석 결과 및 선정 현황(CVR=0.333)
(Table 6) Results of content validity analysis and selection status of detailed items of AI robot(CVR=0.333)

유형	항목	N	M	SD	CVR	선정
인간에 대한 위험	운행, 작업 중 작동 이상	30	4.40	0.841	.667	○
	인간을 해치는 판단	30	4.13	0.991	.733	○
	의도적 침해 행위	30	4.07	0.998	.667	○
인간 권리에 대한 위험	인격권 침해	30	4.10	0.943	.467	○
	차별적 서비스	30	3.57	1.230	.267	미선정
	자유의사 방해	30	3.97	0.912	.400	○
재산 손실의 위험	가치판단 오류	30	3.80	1.077	.267	미선정
	법률, 규칙 위반	30	3.67	1.135	.267	미선정
	정보 입력, 탈취	30	4.23	0.920	.600	○
	기계 결함 오작동	30	4.23	0.989	.533	○
사회적 위험	데이터 오인 판단	30	4.20	0.909	.600	○
	법규 오인 판단	30	3.83	0.934	.333	○
	도로상 작동 이상	30	4.17	1.035	.733	○
	사회적 권한 착오	30	3.57	1.309	.200	미선정

*N(유효 수), M(평균), SD(표준편차), CVR(내용타당도)

유형별 항목에 대한 내용타당도 분석 결과 CVR 기준값 0.333보다 낮은 수치로 나타난 차별적 서비스・가치판단 오류・법률, 규칙 위반・사회적 권한 착오 항목은

제외되고 나머지 10개 항목이 선정되었다. 응답자들은 AI 로봇이 인간을 차별하거나 인간 사회의 가치판단·법률·규칙·사회적 권한 착오를 할 수 없다는 의견이었다.

이를 학습 데이터 문제로 보고 개발 단계에서 충분한 학습을 통해 해결할 수 있다며, 로봇 관련 위험성을 측정하는 지표에 적절하지 않다는 검토 의견이었다. 이러한 의견은 통계 분석 결과에서도 나타났다. AI 로봇의 위험성 측정지표의 내용타당도 분석 결과는 아래 표 7과 같다.

(표 7) AI 로봇의 위험성 측정지표의 내용타당도 분석 결과 및 선정 현황(CVR=0.333)

(Table 7) Results of content validity analysis and selection status of AI robot risk measurement indicators(CVR=0.333)

항목	측정지표	N	M	SD	CVR	선정
운행, 작업 중 작동 이상	통신, 센서 장애	30	4.33	0.869	.800	○
	이동 중 운행 정지	30	4.40	1.020	.867	○
	성능 저하	30	4.17	1.003	.600	○
	균형성 유지 실패	30	3.90	0.978	.667	○
	프로그램 오작동	30	4.37	0.875	.733	○
인간을 해칠 판단	요구사항 인식 오류	30	4.23	0.955	.667	○
	협동 작업 중 규칙 혼동	30	4.03	0.948	.400	○
인간을 해칠 판단	사고 희생자 선택	30	4.07	0.998	.733	○
	잘못된 의학적 판단	30	3.87	0.991	.533	○
의도적 침해 행위	위험한 물건 사용	30	3.83	0.969	.333	○
	직접 폭력 실행	30	3.50	1.360	.267	미선정
인격권 침해	대상 인식에 대한 판단 오류	30	3.90	0.978	.467	○
	명령어 등 음성을 잘못 인식	30	4.07	0.998	.400	○
	감정 피드백 오류	30	3.83	0.969	.467	○
차별적 서비스	차별, 혐오 발언	30	4.00	0.966	.333	○
	선택적 정보 제공	30	3.87	0.921	.467	○
자유의사 방해	인간의 의사 결정을 통제	30	3.93	0.964	.400	○
	자살, 자해 권유	30	4.03	0.983	.400	○
가치 판단 오류	알고리즘 오류	30	4.43	0.883	.733	○
	하자 상품 추천	30	4.00	0.931	.600	○
	반려동물 훼손	30	3.90	0.907	.467	○
	부품 인식 오류	30	3.90	0.978	.400	○
법률, 규칙 위반	창작물 무단 복제	30	4.27	0.998	.667	○
	이동 시 시설물 파손	30	3.97	0.948	.467	○
	안전거리 계산 오류	30	4.00	0.966	.467	○
정보 입력, 탈취	시스템 접근 제어망 침범	30	4.17	1.067	.400	○
	비인가자 접속 허용	30	3.93	1.093	.400	○
기계 결함 오작동	과부하 발생	30	4.27	1.062	.667	○
데이터 오인 판단	부적절한 학습 데이터 수집	30	4.07	0.964	.533	○
법규 오인 판단	법률, 규칙 혼동	30	4.03	0.983	.467	○
	차선, 신호 인식 오류	30	3.90	0.907	.333	○
도로상 작동 이상	배터리 방전 등 시동 꺼짐	30	4.10	1.044	.600	○
	통신 이상 등 장애	30	4.33	0.907	.867	○
	보행자 통행 방해	30	4.27	0.772	.733	○
사회적 권한 착오	데이터 유출	30	4.30	0.936	.667	○
	개인정보 무작위 수집	30	4.20	0.980	.533	○
	허용되지 않은 행위로 혼란 유발	30	4.00	0.683	.533	○

*N(유효 수), M(평균), SD(표준편차), CVR(내용타당도)

AI 로봇의 위험성 측정지표에 대한 내용타당도 분석 결과 CVR 기준값인 0.333보다 낮은 수치로 나타난 직접 폭력 실행 지표를 제거하고 37개 지표가 선정되었는데, 앞서 제거된 4개 항목에 포함된 지표가 12개로 나타났다.

1차 조사를 통해 해당 지표들이 위험성 측정지표로는 의미가 있지만 재분류가 필요하다고 의견을 준 응답자와 추가로 필요하다고 위험성 항목에 대한 의견을 작성해서 회신한 응답자들의 보충적 의견을 반영하고자, 제시된 항목 추가 및 12개 지표를 재분류하는 작업을 하였다.

1차 조사에서 구성원 의견을 토대로 2차 조사 문항을 설계하였고, Likert 5점 척도로 설문조사를 진행하였다.

4.3 2차 델파이 설문조사 결과

2차 델파이 설문조사 문항은 Likert 5점 척도로 구성해 일주일간(2023년 4월 11일~4월 17일) 진행하였고, 참여자 30명 모두 온라인 설문으로 응답하였다.

2차 델파이 설문조사 통계 분석도 내용타당성을 측정하는 CVR 값이 0.333 기준 이상의 수치를 선정하였으며, 모든 설문조사 문항에서 재설문 여부를 판단하기 위한 안정도 값이 재설문이 필요 없는 0.50 이하로 나타남에 따라 추가 조사는 하지 않았다[43].

설문 응답자들이 측정 도구를 정확하고 이해하고 일관되게 측정되었는가를 확인하고, 측정 도구의 정확성과 정밀성을 확보하기 위해 신뢰도 분석을 하며, Cronbach α 값이 일반적으로 0.6 이상이면 신뢰도가 있다고 본다[45].

2차 델파이 설문조사 결과 Cronbach α 값이 0.613 ~ 0.853의 수치를 나타내, 이 연구에서 제안한 위험성 유형 및 항목별 측정지표의 신뢰도를 확보한 것으로 나타났다.

먼저 AI 로봇의 위험성 유형별 통계를 분석한 결과는 아래 표 8과 같다.

(표 8) 위험성 유형별 내용타당도 및 신뢰도 분석(N=30)
(Table 8) Content validity and reliability analysis by risk type(N=30)

유형	M	SD	안정도	CVR	α	선정
신체에 대한 물리적 위험	4.17	0.860	0.21	.533	.734	○
인간 권리에 대한 위험	3.87	0.991	0.26	.400		○
재산 손실 위험	4.13	0.884	0.21	.600		○
사회적 위험	3.87	0.957	0.25	.467		○

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

AI 로봇의 위험성 유형별 내용타당도 분석 결과 CVR 기준값인 0.333보다 높은 수치로 나타나서 모든 유형의 항목이 선정되었고, 신뢰도 값도 0.734로 높게 나타났다.

다음은 신체에 대한 물리적 위험의 항목별 통계 분석 결과로 아래 표 9과 같다.

(표 9) 신체에 대한 물리적 위험의 항목별 내용타당도 및 신뢰도 분석(N=30)
(Table 9) Content validity and reliability analysis by detailed item of physical risk to the body (N=30)

항목	M	SD	안정도	CVR	α	선정
운행, 작업 중 작동 이상	4.10	0.907	0.22	.533	.613	○
인간을 해칠 판단	4.17	0.820	0.20	.733		○
자의에 의한 침해	3.50	0.992	0.28	.067		미선정

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

신체에 대한 물리적 위험의 항목별 내용타당도 분석 결과 CVR 기준값인 0.333보다 낮게 나타난 자의에 의한 침해 항목은 제거되었고, 나머지 2개 항목이 선정되었고, 신뢰도 값은 0.613으로 나타났다. 다음은 인간 권리에 대한 물리적 위험의 항목별 통계 분석 결과로 아래 표 10과 같다.

(표 10) 인간 권리에 대한 위험의 항목별 내용타당도 및 신뢰도 분석(N=30)
(Table 10) Content validity and reliability analysis by detailed item of risk to human rights (N=30)

항목	M	SD	안정도	CVR	α	선정
인격권 침해	3.83	0.969	0.25	.400	.820	○
인간의 자유의사 방해	3.83	0.969	0.25	.333		○
선택적 서비스나 정보 제공	4.27	0.854	0.20	.733		○

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

인간 권리에 대한 위험성의 항목별 내용타당도 분석 결과 CVR 기준값인 0.333 이상의 수치로 나타났다. 모든 항목이 선정되었고, 신뢰도 값은 0.820으로 매우 높게 나타났다. 재산 손실 위험의 항목별 통계 분석 결과는 아래 표 11과 같다.

(표 11) 재산 손실 위험의 항목별 내용타당도 및 신뢰도 분석(N=30)

(Table 11) Analysis of content validity and reliability by detailed item of property loss risk (N=30)

항목	M	SD	안정도	CVR	α	선정
정보 입력, 탈취	4.17	0.687	0.16	.800	.621	○
하드웨어적 기계 결함 오작동	4.17	0.934	0.22	.667		○
잘못된 의사 결정	4.20	0.748	0.18	.733		○

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

재산 손실 위험의 항목별 내용타당도 분석 결과 CVR 기준값인 0.333보다 높게 나타나 모든 항목이 선정되었고, 신뢰도 값은 0.621로 신뢰도 역시 확보되었으며, 사회적 위험의 항목별 통계 분석 결과는 아래 표 12와 같다.

(표 12) 사회적 위험의 항목별 내용타당도 및 신뢰도 분석 (N=30)

(Table 12) Content validity and reliability analysis by detailed item of social risk (N=30)

항목	M	SD	안정도	CVR	α	선정
데이터 문제	3.87	0.718	0.19	.333	.731	○
법규 판단 오류	4.00	0.931	0.23	.533		○
도로상 작동 중지	4.20	0.653	0.16	.733		○

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

사회적 위험의 항목별 내용타당도 분석 결과 CVR 기준값인 0.333 이상의 수치로 나타나 모든 항목이 선정되었으며, 신뢰도 값은 0.731로 높게 나타났다.

다음은 신체에 대한 물리적 위험의 측정지표별 통계 분석 결과로 아래 표 13과 같다.

(표 13) 신체에 대한 물리적 위험의 측정지표별 내용타당도 및 신뢰도 분석(N=30)

(Table 13) Content validity and reliability analysis by measurement indicator of physical risk to the body (N=30)

항목	측정지표	M	SD	안정도	CVR	α	선정
운행, 작업 중 작동 이상	센서 및 인식 오류	4.50	0.619	0.14	.867	.807	○
	운동 지능 제어 실패	3.93	0.727	0.18	.533		○
	기체 성능 저하	4.20	0.748	0.18	.733		○
	균형성 유지 실패	3.60	0.987	0.27	.133		미선정
	압력 조절 실패	3.60	0.987	0.27	.133		미선정
	프로그램 오류로 인한 오작동	4.47	0.670	0.15	.800		○
	통신 장애 등 사고 위험 감지 지연	4.23	0.667	0.16	.733		○
인간을 해칠 판단	의도 파악 인식 오류	3.97	0.836	0.21	.533	.755	○
	협동 작업 중, 규칙 판단 지능의 오류	3.93	0.814	0.21	.400		○
	프로그램의 판단 오류로 사고 희생자 선택 착오	4.00	1.000	0.25	.400		○
	학습량 부족으로 맞지 않는 의학적 자료 제공	4.07	0.814	0.20	.667		○
자의에 의한 침해	물건 등 도구 사용	3.40	1.227	0.36	.067	.846	미선정
	방어적 폭력 사용	3.23	1.174	0.36	-.200		미선정
	신체의 자유 억압	3.57	0.920	0.26	.067		미선정

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

신체에 대한 물리적 위험의 측정지표별 내용타당도 분석 결과 CVR 기준값인 0.333보다 낮은 수치로 나타난 균형성 유지 실패, 압력 조절 실패, 물건 등 도구 사용, 방어적 폭력 사용, 신체의 자유 억압의 5개 지표는 제거되고 나머지 9개 지표가 선정되었다. 신뢰도 값은 각각 0.807과 0.755로 높게 나타났다. 다음은 인간 권리에 대한 위험의 측정지표별 통계 분석 결과로 아래 표 14와 같다.

(표 14) 인간 권리에 대한 위험의 측정지표별 내용타당도 및 신뢰도 분석(N=30)

(Table 14) Content validity and reliability analysis by measurement indicator of risk to human rights (N=30)

항목	측정지표	M	SD	안정도	CVR	α	선정
인격권 침해	대상 인식 판단 오류	3.97	0.912	0.23	.533	.844	○
	명령어 등 음성을 잘못 인식	4.00	0.966	0.24	.467		○
	감정 피드백 오류	3.77	1.086	0.29	.333		○
인간의 자유 의사 방해	의사 결정을 특정 방향으로 유도	4.07	0.998	0.25	.600	.854	○
	자살, 자해 권유	3.27	1.365	0.42	.000		미선정
	여론 형성 조장	3.93	1.062	0.27	.467		○
선택적 서비스, 정보 제공	차별, 혐오 발언	3.80	1.108	0.29	.200	.799	미선정
	학습 데이터 오류로 인해 잘못된 언어 사용	4.13	0.763	0.18	.667		○
선택적 서비스, 정보 제공	편향된 학습 데이터로 인한 차별적 정보 제공	4.33	0.745	0.17	.800	.799	○

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

인간 권리에 대한 위험의 측정지표별 내용타당도 분석 결과 CVR 기준값인 0.333보다 낮은 수치인 자살, 자해 권유, 차별, 혐오 발언 2개 지표는 제거되고 나머지 7개 지표가 선정되었다. 신뢰도 값은 각각 0.799에서 0.854로 높게 나타났으며, 다음은 재산 손실의 위험의 측정지표별 통계 분석 결과이며, 아래 표 15와 같다.

(표 15) 재산 손실 위험의 측정지표별 내용타당도 및 신뢰도 분석(N=30)

(Table 15) Content validity and reliability analysis by measurement indicator of property loss risk (N=30)

항목	측정지표	M	SD	안정도	CVR	α	선정
정보 입력, 탈취	시스템 접근 제어망 침범	4.33	0.745	0.17	.800	.898	○
	비인가 시스템 접근 허용	4.40	0.952	0.22	.733		○
	악성코드, 바이러스 감염 등으로 시스템 오류	4.47	0.763	0.17	.800		○
(하드웨어) 기계 결함 오작동	센서 모니터링 실패	4.30	0.737	0.17	.800	.836	○
	다양한 상황에 의한 충돌로 시설물 파손	4.27	0.680	0.16	.733		○
	구동부(모터, 유압 등) 문제	4.00	0.816	0.20	.467		○
잘못된 의사 결정 (1차 때 가치 판단 오류 및 법규 위반)	유형의 재물 훼손	4.33	0.699	0.16	.867	.812	○
	인식 지능에 대한 오류	4.20	0.600	0.14	.800		○
	창작물 무단 복제	4.03	1.016	0.25	.533		○
	하자 상품 추천	3.90	0.978	0.25	.467		○
	안전거리 계산 오류	4.13	0.670	0.16	.667		○
	알고리즘의 최적화 성능 부족	4.20	0.702	0.17	.800		○

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

재산 손실 위험의 측정지표별 내용타당도 분석 결과 CVR 기준값인 0.333보다 높은 수치로 나타남에 따라 모든 항목이 선정되었다. 신뢰도 값은 0.8 이상으로 매우 높게 나타났다. 다음은 사회적 위험의 측정지표별 통계 분석 결과로 아래 표 16과 같다.

(표 16) 사회적 위험의 측정지표별 내용타당도 및 신뢰도 분석(N=30)

(Table 16) Content validity and reliability analysis by measurement indicator of social risk (N=30)

항목	측정지표	M	SD	안정도	CVR	α	선정
데이터 문제	데이터 유출	4.30	0.781	0.18	.733	.925	○
	(개인 등) 정보 무작위 수집	4.27	0.929	0.22	.600		○
	부적절한 학습 데이터 수집	4.33	0.745	0.17	.800		○
법규 판단 오류	법률, 규칙 혼동	4.00	0.931	0.23	.400	.832	○
	인간이 의도하지 않은 행위로 혼란 유발	4.13	0.921	0.22	.667		○
	차선, 신호 등 인식 오류	3.97	0.836	0.21	.533		○
도로상 작동 중지	(관리 부실 등) 연료 부족으로 정상 운행 중 작동 정지	3.83	1.098	0.29	.400	.732	○
	통신 이상 등 장애	4.27	0.680	0.16	.867		○
	보행자 통행 방해	4.07	0.929	0.23	.600		○

* M(평균), SD(표준편차), CVR(내용타당도), α(Cronbach α 신뢰도)

사회적 위험의 측정지표별 내용타당도 분석 결과 CVR 기준값인 0.333보다 높은 수치로 모든 항목이 선정되었고, 신뢰도 값은 0.732에서 0.925로 높게 나타났다.

4.4 연구 결과

현재 AI 로봇의 기술은 인공지능을 탑재해서 인간의 감정을 인식하고 표현하는 휴머노이드를 개발한 수준이다.

의료 로봇의 수술 사고나 출혈을 악화시키는 약 처방 충고 등 AI 로봇의 위험성에도 정책이나 연구는 기술을 뒤따라가고 있는 실정으로 AI 로봇의 위험성을 측정하기 위한 지표에 관한 연구는 매우 부족하다.

이 연구는 문헌 등 자료 조사로 AI 로봇의 위험성을 측정할 수 있는 사례를 정리하여, 전문가 심층 인터뷰를 통해 위험성 유형과 항목·측정지표를 도출하였다.

또한, 전문가 심층 인터뷰로 델파이 기법을 통한 연구 방법과 설문조사에 적합한 대상 전문가들이 선정되었고, 설문지 문항을 구성하고 설문지를 검토하여 배포하였다.

델파이 설문조사에 참여한 대상자들은 AI·로봇 분야 윤리·교육·개발 등 5년 이상 경력자들과 보안·공학·컴퓨터 분야 교수 등 전문가 30명이며, 2차에 거쳐 설문조사하였다. 델파이 설문조사 결과를 통계 분석하였고, 30명의 전문가 의견 합의를 나타내는 내용타당도 수치가 기준보다 낮은 항목이나 지표는 제거하여, 내용타당성과 신뢰성을 확보하였다. 이 연구의 결과로 검증된 위험성 유형 4개 및 유형별 11개 항목과 37개 측정지표가 최종 도출되었고, 연구 결과를 아래 표 17과 같이 제시하였다.

(표 17) AI 로봇의 위험성 유형 및 측정지표 도출 현황

(Table 17) Risk Types and Measurement Indicators for AI Robots

유형	항목	측정지표
신체에 대한 물리적 위험	운행, 작업 중 작동 이상	센서 및 인식 오류
		운동 지능 제어 실패
		기체 성능 저하
		프로그램 오류로 인한 오작동
		통신 장애 등 사고 위험 감지 지연
	인간을 해칠 판단	의도 파악 인식 오류
		협동 작업 중, 규칙 판단 지능의 오류
		프로그램의 판단 오류로 사고 희생자 선택 착오
		학습량 부족으로 맞지 않는 의학적 자료 제공
인간 권리에 대한 위험	인격권 침해	명령어 등 음성을 잘못 인식
		감정 피드백 오류
		대상 인식 판단 오류
	자유의사 방해	의사 결정을 특정 방향으로 유도
		여론 형성 조장
	선택적 서비스나 정보 제공	학습 데이터 오류로 인해 잘못된 언어 사용
		편향된 학습 데이터로 인한 차별적 정보 제공
재산 손실의 위험	정보 입력, 탈취	시스템 접근 제어망 침범
		비인가 시스템 접근 허용
		악성코드, 바이러스 감염 등으로 시스템 오류
	(하드웨어) 기계 결함 오작동	센서 모니터링 실패
		다양한 상황에 의한 충돌로 시설물 파손
		모터, 유압 등 구동부 문제
	잘못된 의사 결정	유형의 재물 훼손
		인식 지능에 대한 오류
		창작물 무단 복제
		하자 상품 추천
		안전거리 계산 오류
		알고리즘의 최적화 성능 부족

유형	항목	측정지표
사회적 위험	데이터 문제	데이터 유출
		개인 등 정보 무작위 수집
		부적절한 학습 데이터 수집
	법률, 규칙 혼동	법률, 규칙 혼동
		인간이 의도하지 않은 행위로 혼란유발
		차선, 신호 등 인식 오류
	도로상 작동 중지	관리 부실 등 연료 부족으로 정상 운행 중 작동 정지
		통신 이상 등 장애
		보행자 통행 방해

5. 결 론

일상에서 AI 로봇과 함께 사는 것은 편리함을 주고 있어 더 나은 삶을 위해서 앞으로도 계속 발전해 나갈 것이다.

이에 AI 로봇 기술 개발보다 먼저 그에 따른 부작용을 예방하기 위한 규제・연구 등이 선행되어야 하며, 앞으로 나타날 수 있는 여러 가지 위험 요소들까지 대비해 사전 점검을 위해 표준화된 측정 기준이 마련될 필요가 있다.

이 연구는 앞으로 나타나게 될 다양한 위험 요소들을 구체적인 지표를 도출하고 검증함으로써, 여러 분야에서 범용-적으로 활용할 수 있는 표준화된 AI 로봇의 위험성 측정지표를 개발한 매우 의미 있는 연구라 할 것이다.

그동안 인공지능의 법・규제・교육・윤리 등에 집중돼 있던 정책과 연구에서 한발 더 나아가서 AI 로봇과 관련된 초석이 되는 연구로 학술적으로도 기여한 바가 크다.

AI 로봇의 위험성 측정지표 연구는 AI 즉, 인공지능 및 로봇 분야 전문가들의 미래 예측을 통한 집단지성이 있어야 하는 고난도 연구이다.

이 연구를 위해 전문가 심층 인터뷰를 통해 인공지능 전문가는 데이터 학습 분야・개발・윤리・법・교육학자 등이 선정되었고, 로봇 분야 전문가는 기계・센서・통신 연구원이 선정되었으며, 로봇 관련 문제에 대한 개념이 정립된 전자 공학・보안・통신・컴퓨터 관련 교수 등도 참여하였다. 이 연구의 측정지표를 활용하여 AI 로봇과 관련된 신체・권리・재산・사회적 위험성을 사전에 점검한다면 더욱 안전하게 로봇을 사용하고 함께 살아갈 수 있을 것이다. 기존의 연구들은 AI 로봇의 물리적 위험에 주를 이뤄왔지만, 이 연구는 인간의 심리에도 영향을 미칠 수 있음을 검증하였다.

아시모프 로봇의 제3원칙 중 제1원칙인 “로봇은 인간을 해치거나 인간에게 해가 되는 행위를 할 수 없다”를 실천할 구체적인 측정지표 외에도 인간의 권리・재산・사회적 위험성을 측정지표로 나타낸 의미 있는 연구라 할 것이다.

이 연구를 통해 제안된 위험성 측정지표가 AI 로봇의 시험평가・인증・교육 등에 활용된다면 향후 출현하게 될 여러 AI 로봇과 함께 정부・산업・교육・의료・가정 등에서 안전한 삶을 살아가는 데 도움이 될 것이라 기대한다.

참고문헌(Reference)

[1] BG Goo, “The Age of Robots, Human Work: A Guide for Those Who Must Live in the Age of Artificial Intelligence”, 392, Accross Publishing Group, 2020.

[2] SJ Lee, “A study on the Programmed Freedom of Interactive Movie : with the analysis of Detroit Become Human”, Humanities Content Society Journal, Vol.59, pp.219-241, 2020. https://doi.org/10.18658/humancon.2020.12.219

[3] GH Goo, “Beginning at the end of the year, the era of self-driving ’delivery robots’ will open, and the National Assembly legislation will be completed”, Kyunghyang newspaper, 2023. https://www.khan.co.kr/economy/economy-general/article/202304031446001

[4] JM Lee, TS Jin, JP Go, SJ Kim, “Intorduction to robotics”, 273, Jinyoung Publishing Group, 2003.

[5] AK Kim, “2022 Intelligent Robot Action Plan Announced”, Ministry of Trade, Industry and Energy, pp.47, 2022.

[6] OK Choi, BW Jung, KW Gwak, SB Moon, “Degree of autonomy for education robot”, Korean Society for Internet Information, vol.17, no.3, pp.67-73, 2016. https://doi.org/10.7472/jksii.2016.17.3.67

[7] GB Song, “A Study on AI Robots’ Subjectivity for Crime and Its Legal Countermeasures”, 2019.

[8] HK Kang, “The Legal Definition of Artificial Intelligence and the Duty of Protecting Constitutional Fundamental Rights According to the Development of Artificial Intelligence Technology”, the Institute of Constitutional Court, vol.8, no.2, pp.29-64, 2021. https://doi.org/10.35215/jcj.2021.8.2.002

[9] Ministry of Trade, Industry and Energy Mechanical Robot Division, “Intelligent Robot Development and Promotion Act”, 2008.

https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=90237#0000

[10] MS Choi, “A Study on the Tort Liability caused by Accident due to Malfunction of Artificial Intelligence Robots”, The Theory and Practice Society of Civil Law, vol.23, no.3, pp.1-61, 2020. DOI : http://dx.doi.org/10.21132/minsa.2020.23.3.01

[11] SJ Yun, “The Allegory of AI and Empathy in the Movie Her”, Video Culture Contents Research Institute, 2019(0), pp.213 – 236, 2020. https://doi.org/10.24174/jicc.2020.02.19.213

[12] JW Kim, “The issue of determining negligence in the civil liability for medical malpractice due to robot surgery”, Ajou University Law Research Institute, vol.14, no.1, pp.27-56, 2020. http://dx.doi.org/10.21589/ajlaw.2020.14.1.27

[13] HM Jo, “AI mistakes learned by data...In the end, humans have to get it right”, AITIMES, 2021. https://www.aitimes.com/news/articleView.html?idxno=140668

[14] CY Kim, “I can’t trust AI, distrust grows”, The Seoul Economy, 2018. https://www.sedaily.com/NewsView/1RZOK5IKLJ

[15] ALan H., “1.3 billion face recognized in 3 seconds...China’s ’Big Brother’ Scary AI Technology”, The joongAng, 2018. https://www.joongang.co.kr/article/22556103#home

[16] Thor Olavsrud, ″Losing money and ruining reputation, The infamous AI disaster 7th term″, Korea IDG, 2022. https://www.ciokorea.com/news/233054

[17] JS Sun, “The Study of Criminal Law of Medical Artificial Intelligence, focus on Watson”, Korean Society for Law Policy, vol.20, no.3, pp.249 – 274, 2020. https://kiss-kstudy-com-ssl.ca.skku.edu/Detail/Ar?key=3827024

[18] JY Maeng, “Self-driving Cars and Legal Responsibilities”, Seoul: Park Young Publishing Company, pp.497, 2020.

[19] SG Park, “Legal Status of Autonomous Intelligent Robot”, The Institute of Legal Studies, Kyunghee University, vol.15, no.2, pp.1-4, 2022.

[20] MB Kim, “Issues and challenges about the legal status of AI robots - Focusing on Bryson s theory of legal personality”, Korean Public Land Law Association, 2019(87), pp.791-814, 2019. http://dx.doi.org/10.30933/KPLLR.2019.87.791

[21] CW Kim, “Judicial regulations on intelligent robots of the Korean Society of Land and Public Law, on the occasion of the legislative recommendation of the European Union”, A Legal Association, vol.66, no.3, pp.5-59, 2017. http://dx.doi.org/10.17007/klaj.2017.66.3.001

[22] Jobin A., Ienca M., Vayena E., “Artificial Intelligence: the global landscape of ethics guidelines”, Nature Machine Intelligence, vol.1, pp.389-399, 2019. https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf

[23] Huck TP., Münch N., Hornung L., Ledermann C., Wurll C., “Risk assessment tools for industrial human-robot collaboration: Novel approaches and practical needs”, Safety Science, Elsevier B.V., Vol. 141, 2021. https://doi.org/10.1016/j.ssci.2021.105288

[24] WT Lee, “RoboLaw Project in the European Union (EU)”, KISO Journal, vol.23, 2016. https://journal.kiso.or.kr/?p=7496

[25] JH Yoo, “To realize reliable AI domestic and foreign policy trends, Korea Information Strategy Development Institute”, vol.2021, no.6, pp.17-32, 2021. https://ca.skku.edu:8443/link.n2s?url=https://kiss.kstudy.com/ExternalLink/Ar?key=3946171

[26] JW Kim, “Study on the Legal Personality of the Artificial Intelligence Robots in the China”, Chinese Law Review, vol.38, pp.35-65, 2019. https://www-earticle-net-ssl.ca.skku.edu/Article/A355293

[27] SJ So, “A Study on the Development and Reliability of AI Ethics Principles Classification Model and Ethics Metrics”, Sungkyunkwan University General Graduate School, pp.191. 2023.

[28] YS Choi, MJ Chun, “The Moral Philosophical· Psychological Reflection on Artificial Moral Agent”, Korean Society for Ethical Education, Vol. 46, pp. 93-129, 2017. http://dx.doi.org/10.18850/JEES.2017.46.04

[29] SY Byun, “A Study on the Ethical Guideline for artificial intelligence robots - Focusing on 4 principles

of artificial intelligence robot ethics", Korean Society for Ethical Education, Vol. 47, pp. 233-252, 2018. http://dx.doi.org/10.18850/JEES.2018.47.09

[30] MJ Park, "A Study on Artificial Intelligence and Data Ethics: Focusing on Health Data Used in Artificial Intelligence", Journal of the Korean Medical Ethics Society, Vol. 22, No.3, pp.255-273, 2019. https://ca.skku.edu:8443/link.n2s?url=https://kiss.kstudy.com/ExternalLink/Ar?key=3703603

[31] SY Byun, "A Study on the Necessity of AI Ethics Education", Korean journal of elementary education, Vol.31, No.3, pp.153-164. 2020. DOI http://dx.doi.org/10.20972/Kjee.31.3.202009.153

[32] IS Ko, "Basic Principles of Robot Ethics : From Robot Ontology", the Pan-Korean Philosophical Society, Vol.75, pp.401-426. 2014. https://www-earticle-net-ssl.ca.skku.edu/Article/A383624

[33] SJ So, SJ Ahn, "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", Korean Society for Internet Information, Vol.23, No.3, pp.111-123. 2022. https://ca.skku.edu:8443/link.n2s?url=https://kiss.kstudy.com/ExternalLink/Ar?key=3955505

[34] JY Yu, "Establishment of AI Ethics Standards", Ministry of Science and ICT, 2020. https://www.msit.go.kr/bbs/view.do?sCode=user&mPid=112&mId=113&bbsSeqNo=94&nttSeqNo=3179742

[35] CG Lee, TY Jo, JG Lee, EJ Lee, CM Chio, Guidelines for Ethics, Security, and Safety of Self-Driving Vehicles, Ministry of Land, Infrastructure and Transport, 2020. http://www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?id=95084902

[36] WH Kim, "AI Personal Information Protection Self-Checklist", Personal Information Protection Committee, 2021. https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS217&mCode=D010030000&nttId=7347#LINK

[37] BS Beak, "Naver-Seoul National University Announces AI Ethics Rules Prepared for Three Years", ZDNET KOREA, 2021. https://zdnet.co.kr/view/?no=20210217150045

[38] JW Kim, "Scatter Lab AI Chatbot Ethics Checklist", SCATTER LAB, 2022. https://team.luda.ai/ai-ethics-checklist

[39] CH Lee, SY Byun, BJ Kim, HJ Kim, HC Choi, YG Kim, JW Kim, "A Study on the Development of Ethics Checklist for Developers of Healthcare AI Robots in Home", The Korean Ethics Studies Association, Vol.132, pp.263-280. 2021. http://dx.doi.org/10.15801/je.1.132.202103.263

[40] JS Lee, "Delphi method", pp.138, Educational Science Publishing Company, 2020.

[41] HC Kim, IJ Park, MU Kim, "Suggestions for Developing a Metaverse Platform for Educational Purpose: A Delphi Study", Journal of The Korea Society of Computer and Information, Vol.28, No.2, pp.235-246, 2023. https://doi.org/10.9708/jksci.2023.28.02.235

[42] SH Lee, MY Jung, EY Yoo, "Developing Social Play Evaluation Items for Preschool Children: A Delphi Study", Therapeutic Science for Rehabilitation, Vol.10. No.3. pp.97-110, 2021. https://doi.org/10.22683/tsnr.2021.10.3.097

[43] CH Lee, SM Hwang, SY Park, SE Chae, JR Kim, "Development of Guidelines for Setting Up Sensory Integration Rooms in Korea Using the Delphi Method", The Journal of Korean Academy of Sensory Integration, Vol.18, No.2, pp.1-14, 2020. http://dx.doi.org/10.18064/JKASI.2020.18.2.01

[44] C. H. Lawshe, "A quantitative approach to content validity", Personal Psychology, Vol. 28, No. 4, pp. 563-575, 1975.

[45] JJ Song, "Analysis of SPSS/AMOS statistics required for thesis writing", 21st Century Publishers, pp.477, 2021.

◐ 저 자 소 개 ◑

송 현 경(Hyun-kyoung Song)
2010년 세종 사이버대학교 상담심리학과(학사)
2016년 한국외국어대학교 교육대학원 상담심리학과(교육학석사)
2018년 성균관대학교 일반대학원 컴퓨터 교육과(박사 수료)
2008년 7월～현재 경찰청 소속
관심 분야 : 사이버범죄 예방, 학교폭력 예방, 데이터 분류 모형, 인공지능 로봇, AI 윤리 등
E-mail : onlyone_teacher@naver.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 컴퓨터 교육과 교수
관심 분야 : 네트워크, 정보보안, SW・AI 교육, AI 윤리 등
E-mail : sjahn@skku.edu