Journal of Engineering Education Research
Vol. 25, No. 6, pp. 23~37 November 2022

pISSN: 1738-6454 / eISSN: 2713-8283 / THESIS-22-036
https://doi.org/10.18108/jeer.2022.25.6.23

공학전공 대학생의 AI 로봇에 대한 윤리적 민감성

이현옥*,† · 고연주**
*강원대학교 자유전공학부 조교수
**이화여자대학교 해저드리터러시 융합교육연구소 연구원

Engineering Students' Ethical Sensitivity on Artificial Intelligence Robots

Lee, Hyunok*,† · Ko, Yeonjoo**
*Assistant Professor, Division of Liberal Studies, Kangwon National University
**Research fellow, Research Center for Hazard Literacy Education, Ewha Womans University

ABSTRACT

This study evaluated the engineering students' ethical sensitivity to an AI emotion recognition robot scenario and explored its characteristics. For data collection, 54 students (27 majoring in Convergence Electronic Engineering and 27 majoring in Computer Software) were asked to list five factors regarding the AI robot scenario. For the analysis of ethical sensitivity, it was checked whether the students acknowledged the AI ethical principles in the AI robot scenario, such as safety, controllability, fairness, accountability, and transparency. We also categorized students' levels as either informed or naive based on whether or not they infer specific situations and diverse outcomes and feel a responsibility to take action as engineers. As a result, 40.0% of students' responses contained the AI ethical principles. These include safety 57.1%, controllability 10.7%, fairness 20.5%, accountability 11.6%, and transparency 0.0%. More students demonstrated ethical sensitivity at a naive level (76.8%) rather than at the informed level (23.2%). This study has implications for presenting an ethical sensitivity evaluation tool that can be utilized professionally in educational fields and applying it to engineering students to illustrate specific cases with varying levels of ethical sensitivity.

Keywords: Ethical sensitivity, Artificial intelligence robots, Engineering ethics education

I. 서 론

인공지능, 사물인터넷, 3D 프린팅, 클라우드 서비스, 빅데이터 등의 첨단 과학기술이 발달하며 제4차 산업혁명 시대가 도래하였다. 첨단기술의 발달로 개개인은 물질적 풍요와 다양한 제품 및 서비스를 이용할 수 있게 되었으며, 현대 국가와 사회는 효율성이 증대되고 신산업이 창출되며 진보하고 있다(김진하, 2016; Prisecaru, 2016; Schwab, 2017). 그러나 과학기술은 편리함과 이익만을 제공하는 것은 아니다. 우리 사회와 개인의 삶에 해악을 끼치거나 치명적인 문제를 발생시키는 등 부정적인 영향을 주기도 한다. 과학기술의 빠른 변화와 우리 사회에 미치는 영향 및 파급력을 고려해볼 때, 과학기술 발전 방향을 설정하고 인류의 미래를 개척하는 데 기술을 설계하고 개발하는 엔지니어의 역량과 윤리적 소양은 지대한 역할을 할 것이라 기대된다.

이에 따라, 공학전공 대학생들의 인성 및 윤리적 소양을 함양하기 위한 교육이 강조되어 왔다(송성수·김병윤, 2001; 한경희 외, 2009; Vesilind, 1995). 미국공학교육인증위원회에서는 EC2000을 제안하면서 엔지니어가 갖추어야 하는 윤리적 자질과 역량을 명시하였으며, 이를 시작으로 공학윤리 교육이 본격적으로 활성화되었다(한경희 외, 2009). 국내에서는 1999년에 한국공학교육인증원이 설립됨에 따라 관련 교육프로그램을 인증하고 있으며, 특히 전문가로서 사회적, 윤리적 책임에 대한 인식 함양뿐 아니라 공학기술 발전으로 인한 경제적·환경적·사회적 변화 및 영향에 대한 이해를 학습목표로 삼는 등, 공학윤리 교육에 대한 구체적인 목표를 제시하고 있다(고인석, 2010; 한경희 외, 2009). 국내 대학에서는 공학교육인증제의 요건을 충족시키기 위해 '공학윤리', '과학과 공학윤리'와 같은 공학윤리 관련 교과목을 개설하여 운영하거나, '공학 경영', '기술과 사회', '직업과 윤리', '공학 경제' 등의 공학소양 교과를

Received October 17, 2022; Revised November 1, 2022
Accepted November 3, 2022
† Corresponding Author: holee@kangwon.ac.kr

운영하는 과정에서 일부 공학윤리 내용을 포함하여 교육을 시행하고 있다(이소이 외, 2005). 그러나 해당 수업들은 대규모 강의식 수업으로 운영되는 경우가 많아 공학윤리 교과의 특성을 반영하지 못하고 궁극적인 교육목표를 달성하지 못한다는 비판이 제기되어 왔다(이태식 외, 2008).

이외에도 공학윤리 교육현장에서는 빈번하게 발생하는 문제점들이 여럿 존재한다. 주요 문제점 및 도전으로 제기되는 것 중 첫 번째는 다수의 공학윤리 교수자들이 제시한 교육목표와 실제 수업의 활동 내용이 일치하지 않는 경우가 종종 발견되었다는 것이다(Romkey, 2015). 이는 공학윤리 교수자들이 교육목표를 피상적으로 인식하여 수업 방향성이 일관적이지 못하고, 수업내용 및 절차의 조직과정이 체계적으로 이루어지지 못하였기에 초래되는 문제이다. 둘째, 공학윤리 교수자의 전문성 부족이 지적되었다. Barry & Herkert(2014)는 공학윤리 교수자가 윤리학 및 관련 내용에 친숙하지 않아 교육내용을 구성하는 데 어려움을 겪으며, 공학기술 맥락에서 윤리적인 쟁점이나 문제를 다루는 데에도 어려움을 호소한다고 지적하였다. 셋째, 공학윤리 교육의 평가 및 효과 측정과 관련된 연구가 부족하다는 점이다. 교육현장에서 공학윤리의 효과 및 체계적인 평가를 실시하는 경우는 드문 편이다. 국내 수업에서는 주로 필답고사 형식으로 강의에서 제시한 지식을 평가하는 사례가 많았으며(이영남 외, 2007; 이태식 외, 2008), 학생들의 프리젠테이션, 그룹 프로젝트 및 포트폴리오를 비롯하여 성찰 에세이 같은 개별과제를 활용하여 합격/불합격과 같은 이분법적 평가를 하거나 윤리적 측면과 관련 없는 기준으로 단순 평가하는 경우가 많았다(Bielefeldt et al., 2016, Keefer et al., 2014). 이는 윤리 및 도덕성을 엄격하게 분석하거나 이를 측정하거나 평가할 수 없다는 인식이 존재하고(이재숭, 2011) 공학 배경을 가진 다수의 교수진이 공학윤리를 개인적이고 주관적인 주제로 간주하여 이에 대한 평가 및 적절한 피드백을 제공하는 것에 주의를 기울이지 않기 때문으로 볼 수 있다(Davis & Feinerman, 2012).

이에, 본 연구는 미래 엔지니어로서 공학전공 대학생을 대상으로 공학윤리 교육의 주요 목표 중 하나인 윤리적 민감성을 살펴보기 위하여 AI 로봇 시나리오를 개발하고 이를 활용한 평가에서 드러나는 대학생의 윤리적 민감성의 특징을 탐색하고자 하였다. 윤리적 민감성은 특정 상황을 윤리적으로 바라보는 데 필요한 것으로, 엔지니어들이 윤리적 민감성을 지녀야만 윤리적 추론이나 행위가 수반될 수 있기에 그동안 공학윤리의 교육목표로 빈번하게 제시되어 왔다(Martin et al., 2021). 특히 윤리적 민감성을 탐색하기 위한 구체적인 맥락으로 연구자들은 인공지능 기술을 선정함으로써, 4차 산업혁명의 핵심기술과 연계된 다양한 윤리적·도덕적 쟁점을 대학생들이 인지하는지를 다각도로 파악하고자 하였다. 본 연구의 결과는 공학전공 학생들의 수준별 윤리적 민감성 사례를 다양하게 제시하고 이와 함께 평가 기준을 제공할 것이다. 이는 공학윤리 교육자들이 학생들의 윤리·도덕적 측면을 평가하는 하나의 방안이 될 수 있다. 이로써 학습자들은 자신의 수준을 점검할 수 있고 교수자들은 교육내용의 적절성을 판별하여 수업을 개선할 수 있다. 즉, 본 연구는 공학윤리 교육을 체계적으로 발전시키는데 기초 자료 역할을 할 수 있을 것으로 기대된다. 본 연구의 구체적인 연구 문제는 다음과 같다.

첫째, 공학전공 대학생이 AI 로봇 개발 상황에서 고려하는 요인은 무엇인가?

둘째, 공학전공 대학생의 AI 로봇 쟁점에서 윤리적 민감성은 어떻게 나타나는가?

II. 이론적 배경

1. 윤리적 민감성[1)]

윤리적 민감성은 윤리적 행위가 발현되는 데 관여하는 여러 요인 중 하나로, 구체적 맥락에서 특정 상황이 윤리적 영역이라는 것을 인지하여 그 상황에서 전개될 수 있는 여러 상황을 고려할 뿐 아니라 행동의 결과가 다른 사람에게 미치는 영향을 가늠할 수 있는 능력이다(Rest, 1983, 1986). 도덕성 발달에 대한 Rest의 전통적 연구에서 4 구성요소 중 하나로 제시되었다. 그는 1964년에 미국 뉴욕시에서 발생한 Kitty Genovese 사건을 예시로 들며 도덕적 민감성의 중요성을 설명하였다. 밤늦은 시간 도심에 있는 아파트 주차장에서 강도가 Kitty Genovese라는 한 여성을 폭행해 살해하는 사건이 발생하였다. 여러 이웃이 강도가 저지르는 폭행과 30여 분 동안의 소란을 목격하였지만, 누구도 이 여성을 돕지 않았다. 목격자들은 면담에서 이 소란을 연인들 사이의 다툼이라고 치부하거나 어떤 상황인지 파악하지 못했다고 대답했다. 즉, 사건에서 특정 상황을 도덕적·윤리적인 것으로 해석하는 도덕적 민감성이 작동하지 않아 결국 도덕적 행위는 발생하지 않는다. Rest는 도덕적 민감성 이외에도 도덕적 행동이 발생하는 데 필요한 요소로 어떤 행동이 도덕적으로 옳은지 그른지 판단하는 '도덕적 판단(moral judgment)', 상황과 관련된 다양한 가치 중에서 도덕성을

1) 선행 연구에서 '윤리적 민감성'과 '도덕적 민감성' 용어는 연구 맥락에 따라 혼용되어 사용되었다. 전문가를 대상으로 할 때는 '윤리적 민감성'이 주로 사용되었는데 본 연구도 예비 엔지니어로서 공학전공 대학생을 대상으로 하기에 '윤리적 민감성' 용어를 사용하겠다. 다만, 관련 선행 연구를 언급할 때는 해당 논문에 나온 용어를 그대로 사용하였다.

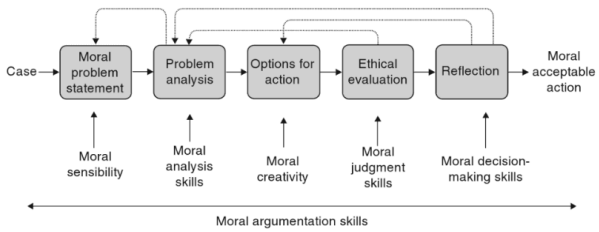

Fig. 1 Six moral skills as goals of engineering ethics education(Van de Poel, & Royakkers, 2011, p.4)

우위에 둘 수 있는 '도덕적 동기화(moral motivation)' 그리고 자신의 의도하는 바를 실행에 옮기는 과정으로 '도덕적 품성 및 실행 능력(moral character)'을 제시하였다.

특히, 윤리적 민감성은 도덕적 행동을 하기 위해 첫 번째로 습득해야 하는 것이기에 공학윤리 교육에서 주요하게 제안되었다(Harris et al., 1996; Martin & Schinzinger, 2013; Pritchard, 2005; Van de Poel & Royakkers, 2011). Van de Poel & Royakkers (2011)는 공학윤리 교육을 통해 함양할 수 있는 6가지 도덕적 기술을 제시하면서 도덕적 민감성을 선두에 배치하였다(Fig. 1). 도덕적으로 수용할 만한 행위가 발생하기까지 도덕적 민감성(moral sensibility)이 제일 먼저 등장하고, 나머지 다섯 가지 능력(도덕적 분석력, 도덕적 창의성, 도덕적 판단력, 도덕적 의사결정력, 도덕적 논증력)이 뒤이어 제시되었다.

윤리적 민감성을 평가하기 위한 도구에 관한 연구가 다양한 분야에서 이루어져 왔다. 처음 개발된 도구는 '치과 윤리적 민감성 테스트(dental ethical sensitivity test, DEST)'로 치과 분야 학생이 윤리적 쟁점을 식별하고 해석할 수 있는지를 평가하였다(Bebeau et al., 1985). DEST는 치과 전문 분야에서 자주 발생하는 윤리적인 문제 상황을 4가지 시나리오로 제시하였다. 예를 들어, 시니어와 주니어 치과의사가 서로 다른 치료를 제안하는 사례, 치료가 시급한데 경제적 형편상 치료받기 힘든 환자 사례 등이다. 학생들은 시나리오를 오디오로 듣고 치과의사 입장에서 몇 가지 질문에 대해 답하였다. 그리고 연구자들은 학생들의 응답을 도덕적·윤리적 측면을 얼마나 잘 감지하였는지에 따라 1에서 3점까지 배점하였다. 이후 DEST와 유사한 형태로 과학 분야에서 적용할 수 있는 '과학에서의 윤리적 민감성 테스트'(Test for Ethical Sensitivity in Science, 이하 TESS)가 개발되었으며(Clarkeburn, 2002; Fowler et al., 2009), 해당 테스트를 위해 과학기술관련 윤리적 쟁점과 연계된 다양한 시나리오가 개발되었다. 예를 들어, 유전자 재조합 작물, 연구비 조달의 어려움으로 대형 회사에 연구소를 인수하는 상황, 인간배아 복제 시나리오 등이 있다. 학생들에게 해당 시나리오와 의사결정이 필요한 질문을 하며 이때 고려하는 요인을 5가지 이하로 작성하도록 요청하였다. 그리고 연구자들은 학생들의 응답에서 윤리적 내용이 얼마나 나타나는지에 따라 0에서 3점까지 배점하였다. 윤리적 내용을 평가하기 위한 기준으로 "가치 진술을 포함할 수 있는 윤리적·도덕적 위험에 대한 고려"(Fowler et al., 2009, p. 288)와 몇몇 학생들의 사례가 추가적으로 제공되었다. 그러나 이 평가 기준은 모호하여 다른 시나리오에 적용한다면 무엇이 윤리적·도덕적 고려사항 인지를 판단하는 것이 어려울 수 있으며 안정적인 합의를 이루기가 쉽지 않을 수 있다.

일련의 경험 연구를 통해 사람들의 윤리적 민감성에는 다음과 같은 특징이 있음을 발견하였다. 첫째, 사람들은 단순한 맥락에서도 상황을 윤리적·도덕적으로 해석하는 데 어려움을 겪으며, 일반적으로 나이가 많아질수록 윤리적 민감성이 높아지는 경향이 있다(Rest, 1983). 둘째, 주어진 시나리오 및 구체적인 맥락에 따라 학생들의 윤리적 민감성 수준이 다르게 드러나는 것으로 확인되었다(Fowler et al., 2009). 마지막으로 윤리적 민감성은 교육을 통해 발전될 수 있는 능력임이 밝혀졌다(Baab & Bebeau, 1990; Clarkeburn, 2002). Clarkeburn(2002)은 대학의 과학교육이 윤리적 민감성에 효과가 있는지를 살펴보았는데 연구 결과, 구조적인 토론을 활용하여 윤리적 인지 향상을 목표로 하는 소규모 과학 수업에서 학생들의 윤리적 민감성 점수가 상당히 증가하였다. 즉, 윤리적 민감성은 개인이 타고나는 성향이기 보다는 잘 설계된 교육을 통해 변화시킬 수 있는 역량이다.

2. 인공지능 윤리

인공지능은 빠르게 발전하면서 광범위한 영역에 걸쳐 활용되는 기술이다. AI 기술을 활용하여 로봇을 설계하는 과정에서는 예상치 못한 상황과 다양한 윤리적 쟁점이 발생할 수 있다. AI 로봇은 Vincenti(1992)의 용어를 빌어 표현하면 '급진적 설계(radical design)'에 해당하는데 이러한 영역에서는 엔지니어들이 '일상적 설계(normal design)'에 비해 더 많은 윤리적 성찰을 해야 할 필요가 있다(van de Poel & van Gorp, 2006). 일상적인 제품 설계에서는 고려해야 하는 규정이 명확하고 외부적 요건이 널리 공유되고 있기에 개별 엔지니어가 스스로 윤리적으로 성찰해야 하는 경우가 많지 않다. 그러나 AI 로봇과 같은 급진적 설계에서는 로봇의 형태와 기능에 대한 원형이 존재하지 않을 뿐 아니라 준수해야 하는 규정이 명확하지 않기에 엔지니어는 다양한 요소를 비롯하여 충돌하는 가치를 다각적으로 성찰해야 한다.

인공지능 기술의 윤리적·도덕적 측면은 다양한 분야에서 논의되고 있으며 벌써 여러 차례 사고가 발생하여 사회적으로 논란이 되었다. 2016년 테슬라의 자율주행 자동차가 신호등이 없는 교차

로에서 상대 차량을 제대로 인식하지 못해 사고를 내었고 결국 운전자가 사망하였다. 이때, 자율주행 자동차 사고를 누가 책임져야 하는지에 대한 공방이 있었다. 또한, 마이크로소프트사에서 2016년에 공개한 채팅로봇 테이(Tay)는 각종 차별적인 표현과 비속어를 무분별하게 사용하여 16시간 만에 서비스를 중단했다. 이 사고는 개발자가 사용자의 데이터를 수집·활용하여 테이를 학습시키기 위해 대중 참여형으로 테이를 개발한 것이 빌미가 되었다(노정동, 2021). 테이 서비스가 시작되자 인종·성차별주의자를 비롯한 특정 종교 혐오집단이 테이를 그들이 원하는 방식으로 학습시켰다. 더욱이 이와 유사한 사고가 국내에서 반복되었다. 2020년 채팅로봇 이루다가 출시되었고 3주 만에 서비스가 종료되었다. 이루다의 성희롱 및 인종·성차별적 발언뿐 아니라 훈련 데이터에서 개인정보가 익명화가 되지 않아 해당 정보가 그대로 노출되는 문제도 발생했다. 이러한 사건들은 AI 학습에서 데이터의 편향 가능성과 개발자의 개인정보 보호에 대한 인식 부족이라는 문제점을 보여준다.

인공지능 윤리에 관한 관심이 높아지면서, 국내외 단체 및 협회에서 인공지능 윤리 가이드라인, 윤리 헌장, 윤리원칙, 권고안 등이 연이어 발표되고 있다. 2017년에는 비영리단체 '삶의 미래 연구소' (Future of Life Institute)의 주최로 미국 캘리포니아의 아실로마에서 '이로운 인공지능 회의'(Beneficial AI conference)가 열렸고, 이것의 결실로 '아실로마 AI 원칙'(Asilomar AI Principles)이 발표됐다. 이 원칙은 연구관련 분야 5개, 장기 이슈 5개, 윤리 및 가치 분야 13개(안전, 장애 투명성, 책임, 가치관 정렬, 자유와 개인정보, 공동이익, 인간의 통제력 등) 등 총 23개로 구성되었다(Future of Life Institue, 2022). 알파고를 개발한 구글의 데미스 하사비스, 미래학자 레이 커즈와일, 우주물리학자 스티븐 호킹, 테슬라의 설립자 일론 머스크 등을 포함하여 관련 분야의 다양한 전문가 2,000여 명이 이 원칙에 서명하였다. 2019년에 유럽연합이 발표한 '신뢰 가능한 AI를 위한 가이드라인' (Ethics Guidelines for Trustworthy AI)은 AI의 전주기 개발과정에서 지켜야 하는 세 가지 요소로 '합법성', '윤리성', '견고성'을 제시하였다(AI HLEG, 2019). '더불어 세부 요건 7가지를 통해 윤리적 고려사항을 구체적으로 제시하였다(인간행위주체성과 관리, 기술적 견고성 및 안전성, 개인정보 및 데이터 거버넌스, 투명성, 다양성, 차별금지 및 공정성, 사회·환경 복지, 책무성). 또한, 국제전기전자기술자협회(Institute of Electrical and Electronics Engineers, IEEE)는 '윤리적으로 조율된 설계'(Ethically Aligned Design)을 발간하여 자율적이고 지능적인 시스템(autonomous and intelligent systems)이 개발되는 과정에서 인류의 이익을 위해 정부, 기업, 그리고 일반 대중이 고려할 지침을 제공하였다(IEEE, 2019). 공학기술전문가 단체라는 정체성을 지닌 IEEE는 가치에 기반하는 설계·개발·구현을 위해 엔지니어가 고려해야 하는 8가지 원칙을 제시하였다(인권, 웰빙, 데이터주체성, 효율성, 투명성, 책무성, 오용에 대한 인지, 역량). 우리나라는 2020년에 '사람이 중심이 되는 인공지능(AI) 윤리기준'을 발표하여 인공지능의 전 주기 과정에서 사회구성원이 고려해야 할 가이드라인을 제공하였다(과학기술정보통신부, 2020). 이 보고서는 인공지능을 특정 분야에 한정하지 않고 총론 차원에서 포괄적으로 적용될 수 있는 기준을 제공하였다. 세부 내용으로는 인간성(humanity)을 최고 가치로 설정하고 '인간의 존엄성 원칙', '사회의 공공선 원칙', '기술의 합목적성 원칙'을 3대 원칙으로 제시하였다. 또한, 세부 요건으로 '인권 보장', '프라이버시 보호', '다양성 존중', '침해금지', '공공성', '연대성', '데이터 관리', '책임성', '안전성', '투명성의 요건'을 제안하였다.

주요 국가 및 단체에서 제안하는 AI 윤리 가이드라인 및 윤리헌장 등을 살펴보면 유사하거나 겹치는 내용이 많고 분류 방식 및 표현의 차이에 따라 같은 내용이 다르게 배치되기도 한다. 이에 소순주·안성진(2021)는 국내외 AI 윤리 가이드라인 및 관련 문헌을 폭넓게 참고하고(AI 윤리 가이드라인으로 해외 10종 및 국내 5종, 관련 정책 및 연구보고서 10종, 논문 및 서적 15종), 각 분야의 AI 전문가(AI 컨설턴트, 교육자, 설계자, 개발자 등) 심층 면담을 통해 AI 윤리원칙 분류 모형을 개발하

Table 1 AI ethical principles(So & Ahn, 2021, p. 123)

윤리 원칙	정의
안전성 (Safety)	인공지능 시스템의 판단, 예측 결과로인해 시스템이 오작동, 위험 발생, 부작용 등이 발생하여 이용자와 환경에 악영향이 미치지 않도록 인공지능 작동 수명 전반에 걸쳐 안전하게 유지, 예방할 수 있는 원칙
통제성 (Controllability)	인공지능기술 및 서비스에 대한 인간의 제어 가능성 및 오작동에 대한 대비책을 미리 마련하고, 이용자의 이용 선택권은 최대한 보장하며, 인공지능 모니터링에 의하여 사전에 위험을 방지할 수 있는 원칙
공정성 (Fairness)	인공지능이 데이터를 처리하는 과정에서부터 인공지능 기능을 이용하는 단계에 이르기까지 특정 개인이나 집단에 차별, 편향되지 않도록 공정하게 보장되어야 하는 원칙
책무성 (Accountability)	인공지능 기술 및 서비스에 대한 정보 공유, 사고 등의 책임(Responsibility) 분배를 명확하게 규정하고, 배상의 정도를 정확하게 측정하여 문제 발생 시 원인을 객관적, 상세화에 의하여 그 결과를 누구에게 물어야 하는가에 대한 행위 주체별 책임에 대한 원칙
투명성 (Transparency)	인공지능에 의하여 어떤 행위가 왜 발생했는지 혹은 어떤 행위가 왜 발생하지 않았는지를 명확하게 검증이 가능한 원칙

였다. 결과적으로 AI 윤리원칙을'안전성', '통제성', '공정성', '책무성', '투명성'으로 정리하였다(Table 1 참조). 그들은 이 모형을 통해 AI 소프트웨어 설계 및 검증·인증을 비롯하여 관련 교육에 적용을 할 수 있다고 제안하였다.

Ⅲ. 연구방법

1. 연구참여자

본 연구에는 서울 소재 대학 융합전자공학부에 재학 중인 27명과 컴퓨터소프트웨어학부 27명, 총 54명이 참여하였다(남학생 38명, 여학생 16명). 이들은 해당 대학의 1학년 전체 학생들이 필수로 들어야 하는 과학기술관련 교양 교과목을 수강한 학생들이다[2]. 학생들의 자발적인 연구참여를 보장하기 위하여 미리 강의계획서에 연구실시 및 관련 내용을 공지하여 참여를 원치 않는 학생들이 다른 분반을 선택할 수 있도록 하였다. 또한, 해당 수업을 수강한 학생들도 연구 참여를 거부할 수 있다고 안내하였다. 학생들은 융합전자공학부 학생 A1-A27, 컴퓨터소프트웨어학부 학생 B1-B27로 표기되었다.

2. 자료수집

본 연구진은 Clarkeburn(2002)의 '과학에서의 윤리적 민감성 테스트'(TESS)를 참고하여, 공학기술 분야에서 활용할 수 있는 시나리오를 개발하였다(Fig. 2 참조). 우선, 컴퓨터소프트웨어 및 융합전자공학을 전공하는 연구참여자들이 쉽게 이해할 수 있고 이후 그들의 전문 활동 영역과 연계될 수 있어 흥미와 관심을 가질 수 있는 내용을 선정하고자 하였다. 또한, 시나리오의 소재 및 상황을 협소한 전문 영역에 한정시키기보다는 사회적·문화적·경제적 영향 및 파급효과가 큰 공학기술 및 사례를 도입하여 학생들이 공학기술의 윤리적 역할과 쟁점을 폭넓게 인지하는지 살펴보고자 하였다. 이를 위해 연구진은 기술영향평가[3]에서 다루어졌던 여러 기술 및 내용을 검토하고, 이 중에서 2019년도 선정된 '소셜로봇'을 주요하게 참고하였다(조아라 외, 2020). 이 기술은 로봇이 인간의 고유 영역을 대신할 수 있어 윤리적 파급력이 크다고 평가되었으며 고령화 사회 문제를 해결할 수 있지만 동시에 사생활 침해 및 과도한 의존

> 다음 상황에서 여러분들이 엔지니어로서 고려하는 사항을 5가지 제시하고, 해당 내용을 제안하는 이유를 충분히 서술해 주세요.
>
> AI 연구단은 감정을 인식하는 인공지능 로봇을 생산하기 위한 프로젝트를 계획 중에 있습니다. 감정을 인식하는 인공지능 로봇은 사람의 감정을 파악하여 적절하게 대응하는 말과 행동을 할 수 있습니다. 이 로봇 개발에는 관련 데이터를 로봇에 직접 저장하는 기술보다는 인터넷 서버와 통신을 통한 클라우드 방식이 활용될 예정입니다. 여러분이 AI 연구단의 엔지니어라면 감정을 인식하는 인공지능 로봇 개발을 위하여 어떤 문제나 측면들을 고려해야 한다고 생각하십니까?

Fig. 2 Ethical sensitivity questionnaire with an AI robot scenario

이라는 부작용 등이 예측되어 여러 논란이 발생할 수 있다는 점이 지적되었다. 또한, 연구진은 공학기술전문가 단체인 IEEE가 발행한 '윤리적으로 조율된 설계' 보고서에 나오는 인공지능 기술에 대한 윤리 지침 및 고려사항과 함께 '감정적 컴퓨팅'(affective computing) 내용에 주목하였다(IEEE, 2019). 이 보고서에서는 감정이 지능의 핵심이라고 강조하며 인공지능 시스템이 인간의 정서적 경험을 증폭하거나 약화해 피해를 줄 수 있다고 지적하였다. 이어서 인공지능 시스템에서 발생할 수 있는 감정과 관련된 다양한 쟁점과 우려를 제기하고 윤리적 권고 내용을 함께 제시하였다. 이와 같이 연구진은 교육적 관점을 가지고 관련 자료를 폭넓게 검토한 후, 공학기술 분야의 윤리적 민감성 평가를 위해 감정을 인식할 수 있는 AI 로봇 시나리오를 개발하였다. 해당 시나리오를 적용한 AI 로봇 관련 윤리적 민감성 질문지를 1주차 수업 시간에 연구참여자에게 배부하고, 15-20분 후 질문지를 수거하였다.

3. 자료분석

가. 공학전공 대학생의 AI 로봇 개발과정 고려사항

공학전공 대학생들이 감정인식 AI 로봇 개발과정에서 고려하는 내용을 분석하기 위해 반복적 비교분석법(constant comparison method)을 활용하였다(Strauss & Corbin, 1998). 먼저, 개방코딩을 통해 연구참여자의 응답을 주제별로 분류하고 코딩의 이름을 표시하였으며, 이를 통해 20여 가지의 주제를 도출하였다. 범주화 단계에서는 개방코딩의 결과로 도출된 주제 및 코드를 비교하여 유사한 내용을 묶어 상위 범주로 분류하고 해당 범주에 이름을 붙였다. 여러 차례의 수정·보완 작업을 거쳐

2) 한정된 공학전공 대학생을 대상으로 실시한 본 연구의 결과는 일반화하는 데는 다소 한계가 있지만, 공학전공 대학생의 윤리적 민감성 수준 및 특성을 가늠하는데 기초적인 자료가 될 수 있겠다.

3) 세계 인류 및 환경에 미치는 영향 및 파급효과가 크고 사회구성원의 관심도가 높은 미래 기술을 선정하고 전문가와 함께 일반 사회구성원이 함께 검토 및 평가서를 작성하여 해당 기술에 선제적으로 대처하거나 관련 정책에 반영하기 위한 제도이다.

6가지 범주를 도출하였다(‘개발 목적 및 조건’, ‘데이터 수집 및 학습과정’, ‘인간감정 정의 및 인식과정’, ‘감정인식 AI 로봇 기능 및 서비스 부작용’, ‘클라우드 효율성 및 안전’, ‘감정인식 AI 로봇으로 인한 사회변화’). 마지막으로 범주 확인 단계에서는 구성된 범주가 잘 구성되었는지를 파악하기 위하여 원자료와 대조하여 재확인하였다. 이 단계에서 몇 차례 범주를 조정하고 이름을 수정하였을 뿐 아니라 연구자 2명이 독립적으로 확인한 후 지속적인 논의를 통해 합의하며 분석의 신뢰도를 높였다.

나. AI 로봇 맥락에서 윤리적 민감성

윤리적 민감성 수준을 구분하기 위하여 임영식 외(2012)가 윤리적 민감성의 세부 요인으로 제시한 ‘상황지각 민감성’, ‘결과지각 민감성’, ‘책임지각 민감성’ 개념을 도입하여 활용하였다. ‘상황지각 민감성’은 특정 상황이 윤리적·도덕적 쟁점이 되는지를 지각하고 해석할 수 있는 능력인데, 본 연구에서는 AI 로봇 시나리오에서 학생들이 AI 원칙과 관련된 쟁점과 문제를 파악하는지를 기준으로 분석하였다. ‘결과지각 민감성’은 특정 행동의 결과가 관련된 사람들에게 미치는 영향 및 전개 과정을 예측할 수 있는 능력인데, 학생들이 주어진 시나리오에서 윤리적·도덕적 상황을 단순 지적하는 것을 넘어서서 구체적인 상황 전개와 결과를 추론할 수 있는지로 해석하였다. ‘책임지각 민감성’은 특정 상황에서 주체로서 행동해야 하는 책임을 느끼는 능력인데, 학생들이 엔지니어로서의 역할과 책임을 명확하게 제시하거나 문제를 해결하기 위한 대안을 구체화하는 능력으로 해석하였다.

우선, ‘상황지각 민감성’판단을 위해 전술한 소순주·안성진(2021)의 AI 윤리원칙 분류모형을 도입하여 활용하였다. 이 모형이 AI 윤리원칙을 체계적으로 제시하고 대분류와 함께 세부요소를 제공하며 다수의 AI 가이드라인 및 윤리 헌장의 내용을 광범위하게 포괄한다고 판단했기 때문이다. 이 AI 윤리원칙 모형은 원칙별로 3~5개의 하위요소를 포함하고 있는데(Table 2 참조), 분석과정에서 안전성, 통제성, 공정성, 책무성, 투명성 원칙과 함께 제시된 하위요소 정의를 분석 준거로 사용하였다. 학생들의 응답을 코딩할 때 복수의 AI 윤리원칙 범주에 해당할 때는 중복코딩을 허용하였고, 동일 AI 윤리원칙에 속한 복수의 하위요소에 해당할 때는 주요하게 나타나는 한 요소를 판별하여 코딩하였다.

윤리적 민감성 수준은 ‘상황지각 민감성’만 발견될 때는 초보적(naive) 수준, 이와 함께 ‘결과지각 민감성’ 또는 ‘책임지각 민감성’이 발견되면 심화(informed) 수준으로 구별하여 평가하였다(Table 3 참조).

Table 2 AI ethical principles' subcomponents(So & Ahn, 2021, p. 124)

윤리 원칙	하위요소
투명성 (Transparency)	설명가능성
	추적가능성
	검증가능성
	식별가능성
책무성(Accountability)	준법성
	책임식별성
	이용자책임성
	개발표준성
공정성(Fairness)	접근성
	알고리즘 차별성
	데이터 편향성
	공공성
통제성(Controllability)	제어 가능성
	모니터링
	이용자 주도성
안전성(Safety)	상태 인식성
	견고성
	보안성
	프라이버시 보호
	가치정렬

Table 3 Levels of ethical sensitivity on AI Issues

수준	기준
초보적 수준	• 윤리적·도덕적 쟁점을 지각(AI 윤리원칙과 연관되는 쟁점과 내용 제시)
심화 수준	• 윤리적/도덕적 쟁점이 이해관계자에게 미치는 영향 등을 구체적으로 예측 • 윤리적/도덕적 쟁점에서 엔지니어로서 책임과 역할을 명시하거나 대안 구체화

연구자들은 각각 요소와 윤리적 민감성의 수준을 코딩한 후 세 차례의 미팅을 통해 의견이 합치될 때까지 조정하여 자료분석의 신뢰도와 타당도를 높이고자 하였다(Lincoln & Guba, 1985). 이를 통해 최종적으로 모든 연구참여자의 응답을 AI 윤리원칙과 민감성 수준에 따라 코딩하였다.

Ⅳ. 연구 결과

1. 공학전공 대학생의 AI 로봇 개발과정 고려사항

공학전공 대학생들이 감정인식 AI 로봇 개발과정에서 고려하는 사항을 분석한 결과, Table 4와 같이 여섯 가지 요인을 고려하는 것으로 드러났다. 연구참여자들은 감정인식 AI 로봇

Table 4 Engineering students' considerations for designing emotion-recognition AI robots

고려사항	내용	빈도수 (%)
가. 개발 목적 및 조건	• 감정인식 AI 로봇을 개발하는 목적 및 사용용도 • 로봇의 경제성을 비롯하여 사회적 수용도 등의 조건을 고려	18 (7.6%)
나. 데이터 수집 및 학습과정	• 감정인식 AI 로봇을 학습시키기 위한 데이터 수집 과정에서 발생하는 쟁점 및 제공자 동의와 같은 조건 등 • 데이터 학습 과정을 비롯하여 데이터 편향성 우려	20 (8.4%)
다. 인간감정 정의 및 인식과정	• 인간감정에 대한 조작적 정의 및 관련 쟁점 • 감정을 인식하는 과정 및 주의점 등	48 (20.3%)
라. 감정인식 AI 로봇 기능 및 서비스	• 감정인식 AI 로봇의 형태 및 기능 • 감정에 대한 다양한 반응 방식과 관련된 부작용 및 대안 • 로봇의 안정성을 비롯한 주의점	60 (25.3%)
마. 클라우드 효율성 및 안전	• 클라우드 효율성 및 안전성 쟁점 • 개인정보 유출을 비롯한 오용 및 남용 등	59 (24.9%)
바. 감정인식 AI 로봇으로 인한 사회변화	• 감정인식 AI 로봇으로 인한 일자리 감소, 인간관계 소원, 인간 정체성 등의 문제점 • 로봇권리, 로봇 통제권을 비롯한 법적 논쟁 등	32 (13.5%)
전체		237 (100.0%)

을 개발하는 과정에서 '감정인식 AI 로봇 기능 및 서비스'를 가장 빈번히 제시하였으며(25.3%), 이어서 '클라우드 효율성 및 안전'를 제시하였다(24.9%). 한편, AI 로봇의 개발 목적을 살펴보는 '개발 목적 및 조건'(7.6%)이나 AI 로봇의 학습을 위한 데이터와 로봇으로 출시되어 운영되는 과정에서 인간 감정에 대응하며 수집되는 데이터 등에 대한 '데이터 수집 및 학습 과정'에 대한 고려(8.4%)는 상대적으로 적은 편이었다. 위 결과를 고려해볼 때, 공학전공 대학생들은 주어진 시나리오에 명시적으로 드러난 두 요인(라, 마)은 상대적으로 손쉽게 떠올렸으나, 그 외에 다른 부분에 있어서는 보다 적극적으로 고려하지는 못한 것으로 사료 된다. 다음 절에서는 각 범주에 따른 학생들의 응답을 자세히 제시하겠다.

가. '개발 목적 및 조건'에 대한 고려

'개발 목적 및 조건'은 감정인식 AI 로봇을 설계 및 개발하는 과정에서 해당 로봇이 누구를 위해 어떠한 목적으로 사용될지, 개발을 위해 필요한 실제 조건은 무엇인지 고려하는 것을 의미한다. 감정인식 AI 로봇을 제작하는 과정에 있어 중요한 내용이지만 일부 학생들만 적극적으로 고려하는 모습을 보였다. 예를 들어, 학생들은 시나리오에 감정인식 AI 로봇의 개발 목적이나 용도, 타겟층에 대한 구체적인 언급이 없음을 지적하면서, "감정을 다룬다는 기능을 구체적으로 어떤 분야에서 쓰일지 알아야 그에 맞는 맞춤형 로봇을 개발할 수 있다"라고 언급하였다(B24). 이 학생은 감정을 인식하는 인공지능 로봇을 만들기 위해서는 어떤 기능을, 어떤 분야에서, 누구를 위해 쓰일 것인지에 대한 정보가 선행되어야 한다고 지적한 것이다. 일부 학생들은 다음 발췌문과 같이 인간의 감정을 파악해야 하는 서비스업에서 이용될 것으로 추측하여 답변하는 모습을 보이기도 하였다(B5).

> 사람감정을 파악한 후 대응하는 로봇이 어디에 쓰일지는 모르겠지만 아마 이것만을 수행하지 않고 다른 기능과 더해져 서비스업에서 주로 쓰일 듯. 악용되지 않도록 주의해야 할 것이다. 특히 심리상담과 같은 목적으로 쓰인다면 대응 방식을 더 세밀하게 고려할 필요가 있을 것이다(B5).

B5은 사람의 감정을 파악하고 대응하는 로봇의 쓰임에 주목하였는데, 특히 심리상담과 같이 인간의 감정을 세밀하게 어루만져야 하는 분야에 있어서는 상대방의 감정에 적절하게 대응하는 방안에 대해 깊이 고민해야 한다고 보았다. 이외에도 학생들은 감정인식 AI 로봇을 개발하기 전 고려해야 하는 조건으로 "대중화될 만큼"의 경제적 측면과 가격을 언급하거나(B4), 사람들이 감정을 인식하는 AI 로봇을 거부하거나 부정적으로 바라볼 수 있음을 인지하고 사회적으로 해당 로봇의 필요성을 안내하고 사회적 수용도를 높이는 방안이 필요함을 제시하기도 하였다.

나. '데이터 수집 및 학습과정'에 대한 고려

'데이터 수집 및 학습과정'은 AI 로봇의 학습을 위해 필요한 데이터 수집 기술 및 이 과정에서 발생할 수 있는 쟁점을 비롯하여 데이터 편향성과 같은 부작용 등과 관련된다. 학생들은 AI 로봇의 감정인식 기능을 구현하기 위해서 충분한 데이터가 필요하다고 지적하거나(B14), 데이터 수집 방법 및 훈련 방식 등을 언급하였다(A7).

데이타를 충분히 수집하고 제대로 분석할 수 있도록 해야 할 것입니다. 감정이라는 것은 같은 인간이라고 하더라도 오해할 수 있는 부분입니다. 예를 들면, 찡그리며 우는 표정을 지었을 때 어떤 사람은 슬퍼서 그러한 것일 수도 있지만 어떤 사람은 매우 기뻐서 그런 표정을 지은 것일 수도 있을 것입니다. 이런 미묘한 차이를 구분하기 위해서는 인간조차도 놓치는 부분까지 충분히 데이타를 수집하고 데이타를 분석해야 할 것입니다(B14)

AI 적절한 훈련(을 위해) 현재 주목받고 있는 것처럼 인공신경망을 사용한다고 하자. 현재까지 인공신경망을 훈련 시키면서 적절한 훈련 데이터의 제공(이 필요하고), 제대로 된 신경만의 설계 등이 없을 경우 인식률이 낮아질 수 있음이 알려져 있다. 이런 경우 깊은 신경망을 사용하고 많은 훈련 데이터를 얻어야 하는데, 훈련데이터의 수집 측면을 잘 해결해야 할 것으로 보인다(A7).

몇몇 학생들은 데이터 편향성에 대한 우려를 제시하였고, 최근에 언론을 통해 보도된 “인종차별적 언행을 배웠던 인공지능” 사례를 언급하며 혐오 표현 및 감정을 상하게 하는 말들로 사람들에게 위해를 가할 수 있다고 하였다(B4).

다. ‘인간감정 정의 및 인식과정’에 대한 고려

이 범주에서 학생들은 인간 감정을 인식하는 로봇의 기능을 구현하기 위하여 감정의 조작적 정의를 비롯하여 감정을 인식하는 과정에서 발생할 수 있는 여러 쟁점을 제시하였다. 예를 들어, 몇몇 학생은 감정인식 AI 로봇 개발과정에서 여느 로봇과 다른 특징 및 기능으로서 감정이 무엇인지를 생각하고 일부는 철학적인 측면을 고려하였다. 다음 B2과 B5의 예시는 이를 잘 보여준다.

일단 사람 감정의 이해가 우선 되어야 하고 사람은 무엇인지 또한 로봇은 무엇인지에 대한 철학적인 논의가 우선시 되어야 할 것 같다(B2)

사람의 감정을 판별하는 기준과 방식(을 고려해야 한다). 사람의 감정을 판단할 수 있는 지표에는 표정, 목소리 톤, 몸짓, 체온, 심장박동 등 여러 가지가 있으므로 이런 것들을 모두 고려하여 종합적인 결과를 낼 때 정확성이 높아지고 의도한 바를 수행할 수 있을 것이라 생각한다(B5).

B2은 인간 감정을 인식하는 로봇의 위상 및 정체성에 대해 고려하였고, B5은 인간 감정의 모호함을 인지하고 감정인식의 기능을 달성하기 위해 구체적으로 무엇을 고려할 것인지를 제안하였다. 또한, 학생들은 사람의 감정을 인식하는 과정에서 발생할 수 있는 특이 상황 및 주의점 등을 다양하게 제안하였다. A26은 겉으로 드러내지 않는 감정을 어떻게 인식할 것인지를 언급하였고 B27은 동일 상황에서도 달리 표현하는 사람들의 데이터 처리를 어떻게 할 것인지 등을 염려하였다. 본질적으로 인간의 감정이 다중적일 수 있다고 전제한 학생은 로봇이 감정을 확률적으로 파악해야 하는지 아니면 가장 주도적인 감정으로 인식해야 하는지에 의문을 제기하기도 하였다(A23).

라. ‘감정인식 AI 로봇 기능 및 서비스’에 대한 고려

다수의 학생이 감정인식 AI 로봇 기능 및 서비스관련 고려사항을 언급하였고, 구체적인 기능의 발현 및 서비스 상황에서의 문제점과 주의점 등을 제안하였다. 다음에서 학생은 로봇의 형태 및 종류가 다양할 수 있다고 전제하고, 어떤 모습일지를 제안하였다.

다양한 컨셉을 개발해야 한다. 가령 걸어 다니는 방법만 해도 몇십 개의 아이디어가 나온다. 이런 점을 보았을 때 스탠드형(상반신만 있는 모델), 워킹형(상, 하반신 모델 전부)을 개발해야 할 것 같다. 또 인간을 돕는 목적으로 로봇을 개발하기에 장애를 극복하도록 도와주는 휠체어형, 목발형 등을 만들 수 있다.(A22)

이외에도 B3 학생은 인간이 표정, 제스처와 같은 비언어적 수단으로 의사소통하기에 로봇도 유사한 표현 방식이 필요하다고 주장하며 “반언어, 비언어적 표현”을 위한 기능을 요구하였다. A16은 여러 언어를 로봇에 장착하여 효율성을 높이는 것이 필요하다는 의견을 제시하였다.

또한, 여러 학생이 AI 로봇의 감정인식 기능을 고려하여 인간에게 대응하는 과정에서 주의해야 할 점을 비롯하여 관련 부작용 등을 제시하였다.

(감정인식 AI 로봇이) 감정에 대한 대응은 상대방의 특성을 고려해야 한다. 인식한 감정이 유사하더라도 상대방의 나이, 성별, 상황 등을 고려해 대응해야 한다(A18).

(로봇의) 감정인식이 너무 민감하면 (사람들이) 불편해 할수 있다. 예를 들어, (어떤 사람이) 숨기고 싶은 감정이 있을 때 (로봇이) 목소리의 떨림이나 다른 외적 요소들을 통해 그 사람의 진짜 감정을 확인하고, 대응한다면 무서울 수도 있을 것 같다(A9).

A18은 감정인식 AI 로봇이 인간을 상대로 대응하는 방식과 표현이 유연하고 다양해야 한다고 제안하고 A9는 고객이 AI 로봇이 불편할 수 있는 상황을 예시로 제시하였다. 한편, 로봇이

잘못된 대응을 할 수 있으며 이에 대한 해결방안이 필요하다는 주장도 제기되었다(A3). 더욱이 B17 학생은 로봇이 “우울한 사람에게 자살을 부추”기는 구체적인 상황을 상정하여 감정인식 AI 로봇의 윤리적 대응 매뉴얼이 필요하다고 주장하였다.

마. ‘클라우드 효율성 및 안전’에 대한 고려

이 범주에서 학생들은 시나리오에서 제시된 클라우드 방식의 효율성 및 안전성에 대한 고려사항을 제시하고 개인정보관련 주의사항 등을 다양하게 제안하였다. 다음 응답에서 학생들은 클라우드의 효율성을 고려하여 요구되는 통신기술을 확인하고 이 기술의 확보와 같은 쟁점을 언급하였다.

> 클라우드 방식의 신속성(을 고려해야 한다). 상대방의 감정을 빠르게 인식하여 즉각 반응을 할 수 있도록 해야 함(A3)

> 통신기술의 확보(를 고려한다). 클라우드 방식을 사용할 경우 로봇의 입력장치에서 입력된 데이타가 중앙서버로 전송되어 서버에 위치한 인공지능에 의해 판단된 후 다시 결과가 로봇으로 전송될 것이다. 이때, 다량의 영상 데이타가 전송 가능하도록 높은 대역폭과 낮은 지연성의 특징을 갖는 통신 기술이 필요하다. 단말인 로봇의 통신장치뿐 아니라 서버의 통신장치의 성능이 매우 뛰어나야 함을 알 수 있다(A7).

A3는 클라우드 성능으로서 빠르기를 염두에 두었고, A7은 클라우드 방식을 운영하는데 필요한 통신기술의 특징을 구체화 시켰다. 이외에도 학생들은 클라우드 서비스의 오류 발생뿐 아니라 개인정보 유출로 인한 오용 및 남용 등에 대한 문제점을 폭넓게 지적하였다.

바. ‘감정인식 AI 로봇으로 인한 사회변화’에 대한 고려

연구참여자들은 감정인식 AI 로봇이 상용화될 때 발생할 수 있는 사회변화와 다양한 문제점을 지적하였다. 대표적으로 일자리 감소 및 인간관계가 소원해지는 현상을 제안하였다.

> AI가 점점 인간을 대신해간다는 것인데 그러면 상당한 양의 일자리를 잃게 된다. AI로 인해 창출하는 분야가 인기 있게 될 것이며 그로 인해 직업 선택에서 편향성을 띄게 될 것이다(B12).

> 실제 사람들과 의사소통할 때의 두려움. 사람은 계속 익숙해져 있는 것에서 벗어날 때 엄청난 괴리감과 낯섦을 경험한다고 생각한다. 로봇과 지속적인 감정교류를 하다가 실제 사람을 만났을 때 혹여나 의사소통에 문제가 생긴다면 기계와 했던 교감에 대한 회의감과 실제 인간관계에서 소외감을 느낄 수 있다고 생각한다. 기계에 과도한 몰입을 하지 않도록 해야 한다고 생각한다(A2).

위의 발췌문에서 B12은 감정인식 AI 로봇이 활용되는 분야에서는 인력 감소가 일어날 수 있으며 이에 구직자들이 선택하는 직종에서 변화가 있을 것이라 예측하였다. A2은 감정인식 AI 로봇이 상용화되었을 때 인간 사이에서 소통의 문제가 발생할 수 있으며 이로써 AI 로봇 사용을 부정적으로 인식할 수 있으므로 로봇 개발에서 주의가 필요하다고 주장하였다. 더욱이 몇몇 학생들은 감정인식 AI 로봇과 인간의 구분이 모호해지는 경우에 발생할 수 있는 인간 정체성 쟁점을 제기하거나, 인간과 유사한 로봇의 권리를 인정해야 하는지를 묻거나 AI 로봇에 대한 통제권을 명확히 해야 한다는 주장도 제기되었다.

2. AI 로봇에 대한 윤리적 민감성

공학전공 대학생들이 감정인식 AI 로봇을 개발하는 과정에서의 고려사항으로 응답한 내용 중, 윤리적 민감성이 드러난 부분을 선별하여 AI 윤리원칙에 따라 구분한 결과는 다음 Fig. 3과 같다.

먼저, 학생들의 270개의 전체 응답 중 AI 윤리원칙에 해당하는 투명성, 책무성, 공정성, 통제성, 안전성의 다섯 가지를 고려한 것은 106개(44.5%)였다. 이를 각 AI 윤리원칙별로 살펴보면, 안전성 64개(57.1%), 통제성 12개(10.7%), 공정성 23개(20.5%), 책무성 13개(11.6%), 투명성 0개(0.0%)로 감정인식 AI 로봇의 안전성을 주요하게 고려하는 것으로 나타났다. 이중 주목할 점은 ‘투명성’에 대한 언급이 단 한 차례도 없었다는 점이다. 즉, 학생들은 AI 윤리원칙 중 일부를 빈번히 제시하며 5가지 윤리원칙을 고르게 인식하지는 못하였다.

공학전공 학생들의 윤리적 민감성을 심화(informed) 수준과 초보적(naive) 수준으로 나누어 분석한 결과, 심화 수준보다는(23.2%) 다소 초보적이고 단편적인 수준으로(76.8%)이 대부분을 차지했다. 즉, 다수의 공학전공 학생들이 인공지능 로봇을 개발하는 상황을 가정하였을 때 관련 AI 윤리원칙을 언급할

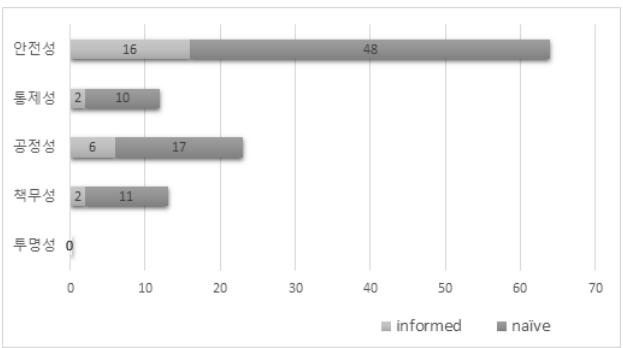

Fig. 3 Engineering students' recognition percentage of AI ethical principles and ethical sensitivity levels

Table 5 Categories of Engineering Students' AI ethical principle recognition

유형	인원수 (%)
AI 윤리원칙에 대한 고려 없음	3 명 (5.6%)
한 가지 영역 고려	26명 (48.1%)
두 가지 영역 동시 고려	14명 (25.9%)
세 가지 영역 이상 동시 고려	11명 (20.4%)

수는 있으나, 관련 이해관계자들이 어떠한 영향을 주고받는지를 구체적으로 예상하거나, 엔지니어로서 책임과 역할을 분명하게 밝히는 경우가 많지는 않았다.

분석의 단위를 동일 학생이 서술한 5가지 응답으로 하여, 학생들이 AI 윤리원칙을 다양하게 고려하는지도 살펴보았다. 교육적 관점에서 학생들이 AI 설계 및 개발 상황에서 다양한 윤리원칙을 인지하는 것은 바람직하기 때문이다. 분석 결과는 다음 Table 5와 같다. 윤리원칙을 전혀 고려하지 않는 학생은 3명뿐이었으나, 대수의 공학전공 학생들이 AI 윤리원칙 중 한 가지만을 제시하였다. 한 가지 영역만 고려한 학생 중 안전성 원칙을 가장 많이 언급하였고(23명), 그 외는 공정성과 책무성을 제시하였다. 두 가지 다른 원칙을 동시에 인지한 학생은 14명으로, 모두 안전성 원칙을 제시하고 나머지 하나의 원칙은 각각 공정성 7명, 책무성 5명, 통제성 2명으로 나타났다. 이로써 학생들은 AI 로봇 설계 과정에서 안전성 원칙은 자연스레 고려한다는 것을 확인할 수 있다. 세 가지 원칙 이상을 동시에 고려하는 학생은 11명인데, 이 중 공정성, 통제성, 안전성 원칙을 동시에 고려하는 학생이 5명으로 가장 많았다. 한편, 다양한 윤리를 인지하면서 심화 수준의 응답을 제시한 학생은 거의 없었다. 유일하게 A22 학생이 심화 수준에서 책무성, 공정성, 안정성 원칙을 고려하였다.

다음 절에서는 학생들의 AI 로봇관련 윤리적 민감성에 대한 경향을 안전성, 통제성, 공정성, 책무성 각각의 하위 요소에 대한 언급 비율을 비롯하여 심화 수준에서 나타나는 내용의 특성을 제시하겠다.

가. 안전성 원칙에 대한 인식

'안전성'은 AI 시스템이 작동하는 과정에서 발생할 수 있는 여러 문제로 인하여 안전이 위협받지 않고 예방할 수 있는 원칙을 의미한다. 학생 대부분은 감정인식 AI 로봇을 개발하는 과정에서 안정성의 하위요소 '프라이버시 보호', '보안성', '견고성'의 측면을 고려하는 모습을 보였다. 감정인식 AI 로봇이 클라우드를 통해 작동한다는 조건과 연계하여 보안성, 프라이버시 보호, 해킹 가능성, 외부 공격에 대한 대비 등을 제시하였으며, 감정과 관련된 정보가 내밀한 개인정보의 하나임을 강조하여 이에 대한 주의를 요청하기도 하였다. 다음 B9와 B23의 응답은 이를 잘 보여준다.

> 클라우드의 안전성(을 고려한다). 감정인공지능 로봇을 만들어낼 관련 데이타를 클라우드에 저장되었을 때 이를 악용할 수 있는 제3자의 개입을 막아야 한다고 생각한다. 데이터가 누군가의 의해 변용되었을 때 가져올 수 있는 악영향을 엄청나다고 생각한다(B9).

> 개인정보 유출(을 고려한다). 감정이라는 것은 매우 개인적인 정보인데 서버와의 통신은 외부인으로부터 노출될 수 있기 때문이다(B23).

B9는 감정인식 AI 로봇의 데이터가 클라우드에 저장되는 과정에서 일어날 수 있는 악용가능성을 지적하면서, 데이터가 유출되거나 해킹될 수 있다고 하였다. 이와 유사하게 B23 역시 인터넷 서버와 통신을 통한 클라우드 방식은 개인정보의 노출 위험이 크므로 주의할 필요가 있다고 설명하였다. 이외에도 '견고성'요소가 빈번하게 등장하였는데, 학생들은 감정인식 AI 로봇 성능과 기능이 기대 수준에서 벗어나서 극단적으로 바뀐 상황을 구체화하고 이에 대해 주의를 당부하였다. B24 학생은 불안정한 통신 상황을 지적하면서 이에 대한 대응으로 사전에 통신사와의 계약 및 밀접한 연계가 필요하다고 제안하였고(B24), B13 학생은 데이터 손실을 막을 수 있도록 다른 저장소를 추가적으로 마련하는 방안을 요구하였다.

일부 학생은 안전성 중 사용자가 부작용 발생 시 이를 대처할 수 있도록 조치하는 '상태인식성'과 인간의 가치와 일치하도록 AI 시스템을 설계 및 개발해야 한다는 '가치정렬'을 고려하기도 하였다. 로봇이 "생뚱맞은 답변"을 제공하는 경우를 대비하여 유사 감정에 대한 응답의 패턴을 파악해야 한다고 보거나(B13) '이타적인 인공지능'을 개발하여 실질적으로 타인을 도울 수 있도록 해야 한다고 제안한 것이 그 예이다(B1).

전반적으로 안전성 관련하여 윤리적 민감성이 심화 수준인 경우는 많지 않았다. 상대적으로 하위요소'프라이버시 보호'에서 높은 편이었는데, AI 로봇에 관련된 쟁점과 문제를 보다 구체적으로 상상해 제시하고 이에 대한 구체적인 대안까지 제안하였다.

> 보안 취약을 막아야 한다. 우리는 어떤 행동을 함으로써 아날로그적 데이타를 순간마다 생성하는데 로봇이 동행한다면 그 로봇은 나의 개인정보를 매번 수집 및 분석한다는 것이다. 하지만 모든 정보를 얻는다고 좋지만은 않다. 예를 들어 해커가 나의 로봇을 해킹하여 로봇의 데이터, 카메라 녹화권한, 마이크 녹음권한 등을 가져갈 시 범죄에 악용될 수 있고 로봇의 움

직임 조작권한 마저 가져갈 시 물리적 폭행으로 번질 수 있기 때문이다(A22).

악용에 대비하여 로봇에 접근하는 사람의 신원파악을 철저하게 확인한 후 로봇-개인간의 대화가 잘 이루어져야 할 것입니다. 다른 사람으로 가장하고 개인정보를 유출할 것을 목적으로 접근하는 경우에는 홍채, 지문 등의 직접적인 인식 방법을 이용하는 것이 좋은 절차일 것 같습니다(A24).

A22은 감정인식 AI 로봇을 훈련 및 개발할 때 필요한 데이터뿐 아니라 상품화된 로봇을 사용하는 과정에서도 사용자의 개인정보가 지속적으로 수집될 수 있음을 인지하고 있다. 즉, 로봇을 사용하면서 매우 내밀한 개인정보가 수집된다는 것을 지적한 것이다. 로봇기능 향상을 위해 대량의 데이터가 항상 최고의 선택이 아닐 수 있음을 인지하며, 해킹되어 어떠한 피해까지 가능한지를 구체화하였다. A24는 의심 가는 사람이 로봇에 접근하여 개인정보 유출되는 상황을 구체화하고 이때 적용될 수 있는 기술적 대안으로 홍채 및 지문 등을 활용하는 방안을 제안하였다.

나. 통제성 원칙에 대한 인식

'통제성' 원칙은 인간이 AI 시스템을 제어할 수 있고 이용자의 선택권을 최대 보장하면서 사전에 위험을 방지할 수 있다는 것을 의미한다. 이에 대한 학생들의 인식은 하위요소 '제어가능성'과 '이용자 주도성' 에서 주로 나타났다. 학생들은 AI 로봇이 오작동을 일으키거나 기술이 고도로 발전하여 AI 로봇이 자아를 형성하거나 인간의 통제에서 벗어나는 상황을 상상하고 이때 필요한 인간의 통제권 및 제어방안을 요구하였다. 아래 발췌문은 이와 관련된 내용을 잘 보여주는 예시이다.

로봇의 자아형성(이 가능하다). 인간의 고유감정과 언어를 인터넷에 저장한다는 것은 한순간에 로봇이 인간과 같이 행동할 수 있다는 말이 된다. 어느 순간 오류로 인해 로봇이 자아를 가지게 되고 다른 로봇들을 인터넷 서버를 통해 자아를 가지게 한다면 로봇들이 반란을 일으킬 수 있는 상황이 벌어질 수 있다(B8).

항상 모든 통제권을 인간이 가지고 있는다. 갑작스러운 로봇의 행동에 대처하기 위해서이다(A11).

B8은 AI 로봇에게 인간의 고유한 감정 데이터를 학습시키다가 어떤 순간에 로봇이 자아를 가질 수 있다고 추측하였다. 이는 SF영화에서 자주 등장하는 설정으로 유사하게 B6 학생도 AI 로봇에 의해 인간이 멸망하는 상황을 제시하였다. 이러한 상황에 대한 대비로 통제권 및 제어방안에 대해 요구를 하였다(A11). 더욱이 학생들은 AI 로봇의 감정인식 기능과 관련하여 이용자들이 자율적으로 조작하지 못하거나 불편함을 느끼는 상황을 고려하여 '이용자 주도성'에 대한 인식을 보여주었다.

로봇 앞에서도 감정을 억누르거나 신경 써야 하므로 불편하고, 감정을 가지고 인식하는 것은 인간의 영역을 침범하는 것이기 때문에 이러한 부분에 대한 고려가 필요할 것 같습니다(B25).

B25는 로봇의 감정인식 기능으로 인하여 사용자가 자유롭게 감정을 표현하지 못하지 못하는 상황을 인식하였고, 이것이 인간 고유의 영역을 침범한 것이라 해석했다. 대부분 학생은 이에 대한 구체적인 대책 및 방안을 제시하지 못하였으며, 소수의 학생만이 대안을 제안하였다.

아무리 발전을 거듭해도 로봇은 로봇일 뿐 진짜 사람이 될 수는 없습니다. 그렇기에 대화의 한도를 설정해야 합니다. 영화 Her에서처럼 '사랑'의 감정을 느끼는 단계가 오지 않게 체계적으로 대화를 연결하는 설정값을 부여해야 합니다. (A24)

A24는 AI 로봇이 결국 인간과는 다르다고 전제하고 사랑이라는 감정까지 도달하는 것은 금지되어야 한다고 주장하였다. 이를 위해 대화의 한도 및 설정값을 부여하여 조정해야 한다고 제안하였다. 이 학생은 AI 로봇에 대한 통제의 필요성을 인지하는 것을 넘어서서 구체적인 대안까지 고려하였다는 점에서 심화 수준에 해당한다.

다. 공정성 원칙에 대한 인식

'공정성' 원칙은 AI 제품 및 서비스가 특정 개인이나 집단에게만 유용하고 그 외 집단에는 접근할 수 없거나 한정되는 상황이 벌어지는 것을 지양하여 공정성을 보장해야 한다는 의미를 지닌다. 학생들의 응답은 주로 하위요소 '공공성' 및 '데이터 편향성'에서 나타났다. 학생들은 AI 로봇의 상용화로 인하여 발생할 수 있는 사회적 문제 및 부작용이 특정 개인이나 집단에 피해를 줄 수 있다고 지적함으로써 AI 로봇의 이익과 혜택이 인류에게 공유되어야 한다는 '공공성'에 배치되는 것을 인식하였다.

(AI 로봇의) 사회적 영향력(을 고려한다). 이 AI가 감정인식 기능을 탑재하여 상용화됐을 때 이것이 가져올 수 있는 사회적 파장에 대비해야 할 것이다. 예를 들어, 감정인식 로봇으로 인해 상담사와 같은 직업이 사라질 수 있다. 이 업계에 종사하는 사람들의 반응은 미루어 짐작할 수 있을 것이다(B9).

B9 학생은 감정인식 AI 로봇으로 인하여 상담사와 같은 직업이 사라질 수 있음을 지적하고 이에 따른 관련 업계 사람들의 피해 및 부정적 반응을 고려하였다. 다른 학생들도 이와 유사하게 AI 로봇으로 인한 일자리 감소 문제를 지적하고 AI 로봇이 궁극적으로 인류에게 이익이 될 수 있을지를 우려하였다. 몇몇 학생은 인간소외 및 인간관계가 소원해지는 현상을 언급하였다. 이들은 피해받는 사람들을 특정하면서 AI 로봇으로 인한 이익과 혜택이 광범위하게 공유되어야 함을 주장하기도 하였다.

'데이터 편향성' 하위요소에 대한 학생들의 인식은 인공지능 학습과정에서 데이터가 편향되거나 의도적으로 조작된 데이터의 사용에 대한 우려로 나타났다.

> 정보의 보편성(을 고려한다). 인공지능은 data set으로부터 학습하는데 만약 데이터가 한 가지 성향으로 편향되어 있다면 올바른 판단을 하기 어렵기 때문이다(B23).

> 사람에게 위해를 가할 우려가 있는지 고려해야 한다. 인터넷 서버와 통신을 통해 관련 데이터를 얻는 방식이므로 인터넷에 있는 혐오 표현 등 날카로운 말들 때문에 사람의 마음이 다칠 수 있다. 예를 들어, 실제로 인종차별적 언행을 배웠던 인공지능이 존재한다(B4).

B23는 편향된 데이터 세트에 대한 우려를 언급하였고, B8은 마이크로소프트 사의 테이 혹은 우리나라에서 출시되었던 이루다 같은 채팅로봇이 인종차별 언행을 해서 미디어에 보도된 사건을 떠올린 것으로 추측된다. 학생들은 정제되지 않은 데이터를 온라인상에서 수집하는 것의 문제점을 지적하였다. 몇몇 학생은 편향되지 않은 데이터를 만들기 위한 노력과 대책 마련이 필요하다고 주장하였다. 다음은 심화 수준의 학생 응답으로 데이터 편향성이 발생하는 상황을 구체적으로 묘사하고 이에 대한 대안이 제안하였다.

> 악성적인 목적의 세뇌를 막아야 한다. 대체로 인공지능 로봇의 대화 기술은 딥러닝 방식을 기반으로 한다. 그렇기에 인터넷 서버와 통신을 통한 클라우드 방식으로 시스템을 설계한 것으로 보인다. 하지만 이 방식에는 한가지 문제가 따른다. 만약에 로봇이 대화를 통해 데이타를 축적하는 방식으로 시스템이 설계되어 있다면 한 이익 집단에서 홍보, 정치적 성향, 욕설, 협박 등을 지속적으로 학습시켜 로봇이 대중에게 동일한 내용을 설파할 가능성이 있다는 것이다. 이런 점에서 '로봇 윤리'를 작성하여 주기적으로 교육하는 방안으로 해결 가능하다고 생각한다(A22).

A22은 마이크로소프트 사의 채팅로봇 테이 사건처럼 AI 로봇을 학습시키는 데이터가 특정 집단에 의하여 의도적으로 편향되거나 조작되는 상황을 묘사하였고, 이를 방지하기 위하여 로봇윤리를 구성하여 대비하는 방안을 제시하였다. 연구진은 이 학생이 해결방안의 기술적 구현방안을 명확히 상술하고 있지는 않지만, 기준을 가지고 데이터 내용 및 학습과정을 주기적으로 확인하는 것이 필요하다고 지적하였기에 윤리적 민감성 수준이 높다고 평가하였다.

이외에도 소수의 학생이 '접근성' 하위 요소와 관련된 내용을 언급하였다. AI 로봇의 혜택을 누구나 공정하게 이용하고 소수집단 및 취약계층이 차별받지 않아야 한다는 인식은 "말을 하지 못하는 사람들"도 활용할 수 있도록 감정인식 AI 로봇은 언어뿐 아니라 손, 발로 표현하는 감정까지도 인식할 수 있어야 한다는 주장(A24)과 "장애를 극복하도록 도와주는 휠체어형, 목발형"의 로봇 외형이 필요하다는 의견(A22)에서 표현되었다.

라. 책무성 원칙에 대한 인식

'책무성' 원칙은 AI 시스템 및 서비스 사용에서 사고가 발생했을 때 행위 주체별 책임을 명확히 하는 데 여러 조치가 필요하다는 내용을 담고 있다. 학생들의 응답은 '개발표준성', '책임식별성', '준법성' 요소에서 나타났다. 학생들은 인공지능 기술이 다양한 분야에 접목이 되고 예상치 못한 사고 등이 발생할 때 이에 대한 책임의 소재를 명확히 하거나 관련 규정 및 법률을 제정하여 사회의 혼란을 줄이는 것을 요구하였다. 또한, AI 로봇 개발과정에서 국제적 논의 및 표준을 준용해야 한다는 '개발표준성'에 대한 인식이 다음처럼 나타났다.

> 감정을 인식하는 것에서 더 나아가 감정을 표현하게 되는 로봇이 개발되는 것을 고려할 때 인간과 로봇을 구분할 수 없는 단계에 도달하기 전에 그런 사항에 대해 어떻게 생각하며 어떻게 대처할지에 대해 국가적/세계적인 논의가 필요하다고 생각한다(A17).

A17는 AI 감정인식 로봇이 고도로 발달하면 인간과 구별하기 힘든 단계에 도달할 수 있고 이런 상황에서는 사회가 혼란에 빠질 것으로 예측하였다. 따라서 이에 대한 국제적인 논의 또는 표준 등이 필요함을 주장하였다. '개발표준성'을 인지한 학생들은 일반적인 AI 설계 및 개발과정에서 준수해야 하는 국제표준을 언급하기보다는 AI 로봇의 오용·악용 또는 극단적 상황에서 관련 규정이 필요하거나 이를 준수해야 한다는 의견을 제시하였다. '책임식별성'에 대한 학생들의 인식도 AI 로봇 관련 사고에 대한 책임 여부를 명확히 하기 위해 관련 자료 및 기록이 필요하다는 요청을 강조하기보다는 책임을 구별하기

힘든 상황을 구체화하는 방식으로 드러났다. 예를 들어, A27은 “아이나 지적 장애인”과 같이 자신의 견해를 명확히 밝히거나 판단할 수 없다고 여기는 사람들을 보호하는 데 필요한 수칙을 요구하였다.

‘준법성’에 대해 학생들은 사고 발생 시 책임을 논의하기 위한 법규와 윤리를 요청하거나 AI 로봇의 개발자 관점에서 어떤 상황에서 책임에 대한 논란이 벌어질 수 있는지를 강조하였다.

> 감정인식 로봇의 말, 행동으로 인한 결과에 대한 법률적 조항에 대한 예측을 고려한다. 예) 로봇이 예상치 못하게 적절치 못한 언행으로 누군가에게 피해를 줬을 때 피해 보상의 주체는 누구인가? 등 (B4)

> 감정을 인식하고 적절한 말을 할 때 윤리적인 기준에 부합하는 말을 해야 할 것이다. 예를 들어, 사용자의 죽고 싶다는 감정을 인식하였을 때 “그럼 죽어”라는 식으로 얘기하는 것이 아닌 다른 방법으로 고민을 해결해야 한다(A26).

B4은 사고 발생 시 피해 보상과 관련된 법률을 언급하며 이를 고려해야 한다고 제안하였다. 또한, 이 학생의 다른 응답에서 “동물·식물 보호법”과 같이 로봇의 학대 및 파괴에 대응하는 법률에 대한 고려도 등장했다. A26은 법률 및 윤리 가이드에 어긋나는 상황을 구체화하였다. 사람의 부정적인 감정 특히, 자살이라는 충동이 있을 때 이에 대한 적절한 가이드라인이 필요하다는 의견을 제시하였다.

전술한 바처럼 책무성관련 인식에서 학생들은 법률, 윤리 가이드라인을 벗어나는 상황을 구체화하는 예에서 심화 수준이 있었으나. 사고 발생 시 어떤 배상 체계 및 책임자를 식별하는 방식을 명확히 제시하거나 이러한 과정에서의 자신의 역할을 분명히 제시한 경우는 없었다.

Ⅴ. 결론 및 제언

본 연구는 미래 엔지니어로서 공학전공 대학생을 대상으로 AI 로봇 시나리오를 적용한 윤리적 민감성을 평가하고 여기서 드러나는 대학생들의 특징을 탐색하였다. 첫째, 감정인식 AI 로봇 개발과정에서 공학전공 대학생이 고려하는 것이 무엇인지를 귀납적으로 분석한 결과, ‘감정인식 AI 로봇 기능 및 서비스’, ‘클라우드 효율성 및 안전’과 같이 주어진 시나리오와 직접적으로 관련된 사항을 자주 언급하였으며, ‘개발 목적 및 조건’, ‘데이터 수집 및 학습과정’, ‘인간감정 정의 및 인식과정’, ‘AI 로봇으로 인한 사회변화’와 같이 추론이 필요한 내용은 상대적으로 적었다. 이처럼 공학전공 학생들은 다소 시나리오에 의존하는 모습을 보였으나, 전반적으로 AI 로봇 개발과정에서 제기될 수 있는 여러 쟁점을 포착하여 다각적으로 제안하였다. 둘째, 공학전공 대학생의 윤리적 민감성을 AI 윤리원칙과 그 수준에 따라 분석한 결과, 학생들의 응답 중 윤리적·도덕적 내용이 포함된 것은 44.5%이었으며, 다섯 가지 AI 윤리원칙 중에서는 ‘안전성’이 가장 빈번하게 언급되었다(57.1%). 반면,‘투명성’은 한 번도 제시되지 않았는데, 이는 별도의 조치가 없다면 학생들이 AI의 투명성 원칙을 명확히 인지하지 못할 가능성이 크다는 것을 시사한다. ‘투명성’은 인공지능 기술이 여느 다른 기술과 차별화되는 특성이다. AI 로봇이 스스로 배우고 판단하면서 작동 방식이 블랙박스화되기에 전문가들도 결과 및 산출물을 판단하고 검증하는 데 어려움을 겪을 수 있다는 것과 연계된다. 따라서 AI 기술에 대한 ‘투명성’원칙을 교수자가 명시적으로 교육하거나, AI 윤리 교육에서 특화하여 교수할 필요가 있겠다. 더불어 학생들의 윤리적 민감성 수준을 분석한 결과, 대부분 초보적 수준에 머무르고 있으며(76.8%) 심화 수준에 도달한 경우는 많지 않았다(23.3%). 윤리적 민감성이 심화 수준인 사람은 단순히 어떤 상황이 윤리적·도덕적 상황이라는 것을 알아채는 것을 넘어서서 그 상황이 어떻게 진행될지를 구체적으로 상상하고, 관련 이해관계자들의 관점을 고려할 수 있다. 또한, 책임의 주체로서 자신의 역할을 명확히 인식하는데, 엔지니어 입장에서는 다양하고 구체적인 대안을 제안하는 것과 연계된다. 따라서 공학교육 설계 및 실행에서는 단순히 규범적인 윤리원칙을 제시하는 것을 넘어서서 윤리적 상상력 및 세부적인 전공내용과 연계시켜 윤리적 민감성 수준을 향상시킬 필요가 있다.

한편, 윤리적 민감성과 관련된 학생들의 응답을 학생별로 분석한 결과, AI 윤리원칙을 다양하게 인식하는 비율이 상대적으로 낮았다(3가지 영역 동시 고려 20.4%). 다만, AI 로봇을 개발하는 상황에서 관련 윤리원칙을 하나도 고려하지 못하는 학생의 비율은 낮아서(5.6%), 관련 교육을 위한 기초적인 기반은 마련되어 있다고 할 수 있다. 즉, 다수 학생은 AI 로봇을 개발하는 과정에서 도덕적·윤리적 쟁점 및 상황이 존재한다는 것을 인지하기에 수업에서 구체적인 AI 윤리원칙을 소개하며 자연스럽게 관련 교육을 진행할 수 있다. 또한, ‘안전성’에 대한 학생들의 인지가 높다는 결과를 반영하여 교육설계 및 진행에서 안전성 원칙을 시작으로 다른 AI 윤리를 소개하거나 확장하는 방식으로 교수를 시행할 수 있다.

본 연구의 결과는 공학윤리 교육에 다음과 같은 함의를 지닌다. 첫째, 윤리적 민감성이라는 교육목표를 직접 평가하고 현장에서 쉽게 적용할 수 있는 평가 방법을 구체적으로 제시하였다. 선행 연구들은 공학윤리 교육에서 윤리적 측면을 직접적으

로 평가하기보다는 학생들의 프로젝트 산출물이나 반성저널 등을 분석하여 간접적으로 이해하는 데 그쳤다(Bielefeldt et al., 2016). 본 연구처럼 공학윤리 교육의 주요 목표에 해당하는 윤리적 민감성을 소개하고 평가 도구 활용방식을 제공하는 것은 현장의 공학윤리 교육자에게 유용할 것이다. 또한, 교육 현장에서 자주 활용되는 루브릭 형식으로 윤리적 민감성을 평가하고 수준별 사례를 다양하게 제시하였기에 교수자는 이를 쉽게 자신의 수업에서 적용할 수 있으며 교육내용의 적절성을 판별하여 수업을 개선할 수 있고, 학습자들은 평가 내용으로 자신의 수준을 점검할 수 있다. 둘째, 본 연구진은 연구참여자와 교육 맥락을 고려하여 감정인식 AI 로봇 시나리오를 개발하였는데, 이와 유사하게 다른 공학윤리 교육에서도 각각의 상황과 맥락을 고려하여 윤리적 민감성 시나리오를 개발·적용하여 맞춤형 교육을 할 수 있다. 윤리적 민감성은 맥락 특이적임을 고려해본다면(Fowler et al., 2009), 다양한 시나리오를 개발하여 교육에 활용하는 것은 바람직하다. 더욱이, 짧게 적용할 수 있는 공학윤리 교육 모듈을 개발하여 각 공학 전공교과목에 적용한다면 공학윤리 교육의 저변을 넓히고 다양화시키는 계기가 될 것이다(이영남 외, 2007). 독립적인 교과로 실행되는 공학윤리 교육은 폭넓은 내용을 집중적으로 교육시킬 수 있는 장점이 있지만, 한 학기 동안 유사한 내용으로 이루어지기에 학습자가 지루해할 수 있으며 공학윤리를 전공하지 않은 교수자에게 부담될 수도 있기 때문이다. 따라서 각 전공관련 설계 및 개발과정에서 발생할 수 있는 윤리적 쟁점 사례들을 활용하여 윤리적 민감성 시나리오를 개발하고, 이를 각 전공 교과에서 맞춤형 교육모듈로 적용하는 방식으로 공학윤리 교육을 확장 시킬 수 있다.

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1C1C1007116).

참고문헌

1. 고인석(2010). 공학윤리교육의 지향점과 방법. **철학논총, 59**(1), 3-23.
2. 과학기술정보통신부(2020.12.22.). 과기정통부, 사람이 중심이 되는 「인공지능(AI) 윤리기준」 마련 보도자료. Retrieved August 31, 2022, from https://www.msit.go.kr/bbs/view.do?sCode=user&m Pid=112&mId=113&bbsSeqNo=94&nttSeqNo=3179742
3. 김진하(2016). 제4차 산업혁명 시대, 미래사회 변화에 대한 전략적 대응 방안 모색. *KISTEP R&D InI*, 15(47), 45-59.
4. 노정동(2021). **대중에게 훈련 맡기자 '인종차별'부터 배운 AI.** Retrieved August 31, 2022, from https://www.hankyung.com/it/ article/202101169172g
5. 소순주·안성진(2021). 인공지능 윤리원칙 분류 모형 및 구성요소에 관한 연구. **컴퓨터교육학회논문지, 24**(6), 119-132.
6. 송성수·김병윤(2001). 공학윤리의 흐름과 쟁점. 유네스코한국위원회 편, **과학연구윤리** (pp. 173-204). 서울: 당대.
7. 이소이 외(2005). 공과대학의 공학윤리 교육과정 운영 실태 조사. **공학교육연구, 8**(2), 35-51.
8. 이영남 외(2007). 공학윤리 교육모듈 컨텐츠를 이용한 전공교과목에서의 공학윤리 교육. **공학교육연구, 10**(4), 78-92.
9. 이재승(2011). 공학인증과 공학윤리. **윤리교육연구, 26**, 277-296.
10. 이태식 외(2008). 공학윤리 교육과목 실태분석을 통한 개선방안 도출: 수요자를 중심으로. **공학교육연구, 11**(3), 96-107.
11. 임영식 외(2012). **청소년 도덕성 진단 검사도구 표준화 연구 II: 청소년 도덕적 감수성 및 도덕적 판단력 검사도구 개발** (Report no. 12-R12-1). 서울: 한국청소년정책연구원.
12. 조아라 외(2020). **2019년 기술영향평가 보고서: 소셜로봇 기술 보고서**(Report no. 기관2019-017). Retrieved August 31, 2022, from https://www.kistep.re.kr/reportDetail.es?mid=a10305010000&rpt_no=RES0220200023. 충북: 한국과학기술기획평가원.
13. 한경희·허준행·이충용(2009). 공학 윤리 교육: 현황과 쟁점, 그리고 전략. **공학교육연구, 12**(1), 31-41.
14. AI HLEG(2019). *Ethics guidelines for trustworthy AI*. Brussels, Belgium: European Commission.
15. Baab, D. A., & Bebeau, M. J.(1990). The effect of instruction on ethical sensitivity. *Journal of Dental Education,* 54(1), 44.
16. Barry, B. E., & Herkert, J. R.(2014). Engineering ethics. In A. Johri & B. M. Olds (Eds.), *Cambridge Handbook of Engineering Education Research* (pp. 673-692). Cambridge, England: Cambridge University Press.
17. Bebeau, M. J., Rest, J. R., & Yamoor, C. M.(1985). Measuring dental students' ethical sensitivity. *Journal of Dental Education,* 49(4), 225-235.
18. Bielefeldt, A. R. et al.(2016). *Efficacy of macroethics education in engineering*. Paper presented at the 2016 ASEE Annual Conference and Exposition.
19. Clarkeburn, H.(2002). A test for ethical sensitivity in science. *Journal of Moral Education,* 31(4), 439-453.
20. Davis, M., & Feinerman, A.(2012). Assessing graduate student progress in engineering ethics. *Science and Engineering Ethics,* 18(2), 351-367.
21. Fowler, S. R., Zeidler, D. L., & Sadler, T. D.(2009). Moral sensitivity in the context of socioscientific issues in high

school science students. *International Journal of Science Education,* 31(2), 279-296.
22. Future of Life Institute(2022). 아실로마 AI 원칙. Retrieved August 31, 2022, from https://futureoflife.org/2017/11/18/ai-principles-korean/
23. Harris, C. E. et al.(1996). Engineering Ethics: What? Why? How?. *Journal of Engineering Education,* 85(2), 93–96.
24. IEEE(2019). *Ethically aligned design first edition: A vision for prioritizing human well-being with autonomous and intelligent systems*. Retrieved October 15, 2022, from https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
25. Keefer, M. W. et al.(2014). The importance of formative assessment in science and engineering ethics education: Some evidence and practical advice. *Science and Engineering Ethics,* 20(1), 249-260.
26. Lincoln, Y. S., & Guba, E. G.(1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage.
27. Martin, D. A., Conlon, E. & Bowe, B.(2021). Using case studies in engineering ethics education: The case for immersive scenarios through stakeholder engagement and real life data. *Australasian Journal of Engineering Education,* 26(1), 47–63.
28. Martin, M. W., & Schinzinger, R.(2013). *Ethics in engineering* (4th ed.). New York, NY: Mcgraw-Hill Book.
29. Prisecaru, P.(2016). Challenges of the fourth industrial revolution. Knowledge horizons. *Economics,* 8(1), 57-62.
30. Pritchard, M. S.(2005). Perception and imagination in engineering ethics. *International Journal of Engineering Education,* 21(3), 415–423.
31. Rest, J. R.(1983). Morality. In J. H. Flavell & E. Markman (Eds.), *Handbook of child psychology (Vol. 3): Cognitive development* (pp. 556–629). New York, NY: Wiley.
32. Rest, J. R.(1986). *Moral development: Advances in research and theory*. New York, NY: Praeger.
33. Romkey, L.(2015). *Engineering, society, and the environment in the teaching goals and practices of engineering instructors*. Paper presented at the 2015 ASEE Annual Conference & Exposition.
34. Schwab, K.(2017). *The fourth industrial revolution.* New York, NY: Crown Business.
35. Strauss, A., & Corbin, J.(1998). *Basics of qualitative research techniques*. Thousand Oaks, CA: Sage.
36. Van de Poel, I., & Royakkers, L.(2011). *Ethics, technology and engineering: An introduction*. West Sussex, England: John Wiley & Sons.
37. van de Poel, I., & van Gorp, A. C.(2006). The need for ethical reflection in engineering design: The relevance of type of design and design hierarchy. *Science, Technology, & Human Values,* 31(3), 333–360.
38. Vesilind, P. A.(1995). Evolution of the American society of civil engineers code of ethics. *Journal of Professional Issues in Engineering Education and Practice,* 121(1), 4-10.
39. Vincenti, W. G.(1992). Engineering knowledge, type of design, and level of hierarchy: Further thoughts about what engineers know…. In Kroes, P & Bakker, M. (Eds), *Technological development and science in the industrial age: New perspectives on the science-technology relationship* (pp. 17-34). Dordrecht, Netherlands: Springer.

이현옥 (Lee, Hyunok)
2015년: 이화여자대학교 과학교육학 박사
현재: 강원대학교 자유전공학부 조교수
관심분야: 과학기술관련 사회·윤리 쟁점 교육, 공학/기술의 본성, 공학윤리
E-mail: holee@kangwon.ac.kr

고연주 (Ko, Yeonjoo)
2017년: 이화여자대학교 과학교육학 박사
현재: 이화여자대학교 해저드리터러시 융합교육연구소 연구원
관심분야: 과학기술관련 사회쟁점 교육, 테크놀로지기반교육
E-mail: yeonjooko1@ewha.ac.kr