ISSN 1598-7094
성신여자대학교 교육문제연구소

교육연구 제95집 pp.287-308
https://doi.org/10.17253/swueri.2026.95.013

대학생의 인공지능 기반 의사결정 수용도 분석: 사회인지영역이론에 따른 영역별 판단 비교

박보람* · 노성호**

국문초록 본 연구는 대학생의 인공지능 기반 의사결정 수용이 단일한 기술 태도가 아니라, 판단 장면이 어떤 사회인지 영역으로 해석되는지에 따라 달라지는 '영역별 조건부 수용' 구조를 갖는다는 점을 탐색하였다. 강원대학교 사범대학 3학년 재학생 43명을 대상으로 개인적, 사회인습적, 도덕적 영역을 대표하는 9개 시나리오를 제시하고 수용도를 반복측정 분석으로 비교한 결과, 영역 간 차이가 유의했으며(Friedman test $p<.001$, Kendall's $W=0.52$), 사회인습적 영역(평균 3.86)이 개인적 영역(평균 3.26)과 도덕적 영역(평균 2.46)보다 높고 도덕적 영역이 가장 낮았다. 개방형 응답 387개를 내용분석한 결과, 개인적 영역에서는 유용성과 편의, 자율성과 선택권이 핵심 정당화 논리로 나타났고, 사회인습적 영역에서는 유용성과 함께 공정성 및 객관성 기대가 결합되어 수용을 지지했으며, 도덕적 영역에서는 안전, 해악, 생명 준거가 압도적으로 우세하여 위임을 제한하는 논리가 강하게 나타났다. 이를 통해 본 연구는 엘리엇 튜리엘(Elliot Turiel)의 사회인지영역이론에 근거해 대학생의 AI 수용 판단이 자율성, 절차적 정당성, 해악 및 권리 보호라는 서로 다른 준거에 의해 구조화됨을 제시하고, 교육 설계와 제도 도입에서 설명 가능성, 책임 귀속, 이의제기, 안전장치를 기본 조건으로 포함해야 함을 시사한다. 다만 본 연구는 강원대학교 사범대학 3학년 재학생으로 구성된 대학생 표본을 대상으로 한 탐색적 연구이므로, 결과를 일반 대학생 전체나 청소년 일반으로 확대 해석하는 데에는 신중할 필요가 있다.

주제어: 사회인지영역이론, 인공지능 의사결정 수용, 알고리즘 추천, 절차적 정당성, 해악예방과 책임

목차

논문접수일: 2026.03.04.
논문수정일: 2026.03.27.
게재확정일: 2026.04.10.

Ⅰ. 서론

오늘날 대학생이 경험하는 많은 선택은 인공지능과 알고리즘의 개입을 통해 형성된다. 무엇을 시청하고 어떤 정보를 읽을지, 어떤 답변을 참고할지를 결정하는 과정에 추천 시스템과 자동 분류 기능이

* 제1저자, 강원대학교 윤리교육과 부교수 / boraming@kangwon.ac.kr
** 교신저자, 강원대학교 윤리교육과 대학원생 / okroh0798@naver.com

깊이 관여하고 있다. 미국에서 13세에서 17세 청소년을 대상으로 실시한 2025년 조사에 따르면, 청소년의 약 3분의 2가 인공지능 챗봇을 사용한 경험이 있다고 보고했으며, 상당수는 이를 일상적으로 활용한다고 응답했다(Faverio & Sidoti, 2025). 뉴스 소비 역시 알고리즘 선별에 점점 더 의존하는 방향으로 이동하고 있다. Swart는 청소년의 뉴스 경험이 알고리즘 큐레이션에 의해 구조화되며, 그 결과 무엇을 보게 될 것인지의 범위가 사전에 설정된다고 지적한다(Swart, 2021). 본 연구는 이러한 변화가 대학생 집단의 판단에서 어떻게 나타나는지를 탐색하려는 것이며, 청소년 일반 전체를 대표하려는 연구가 아니라 강원대학교 사범대학 3학년 재학생으로 구성된 대학생 표본을 대상으로 한 탐색적 연구이다. 이와 같은 변화는 대학생의 인공지능 의사결정 수용 문제를 단순한 기술 선호의 차원을 넘어, 판단 기준과 통제권의 문제로 확장한다.

그렇지만 대학생의 수용 태도가 항상 동일하게 나타나는 것은 아니다. 같은 추천 기능이라도 음악이나 영상 선택과 같은 일상적 장면에서는 편리함을 이유로 쉽게 받아들일 수 있다. 반면 평가나 선발과 같이 결과가 중대한 상황에서는 불신이나 거부 반응이 두드러지게 나타날 수 있다. 요엘레 스와르트(Joëlle Swart)는 청소년이 알고리즘을 이해하고 대응하는 방식이 맥락에 따라 달라지며, 반복적인 사용 경험이 곧바로 신뢰로 이어지지 않는다고 지적한다(Swart, 2021). 따라서 인공지능 의사결정 수용을 하나의 일반적 태도로 환원하면, 수용을 강화하는 조건과 약화하는 상황을 구체적으로 설명하기 어렵다.

국내 연구에서도 이와 같은 맥락의존성이 확인된다. 생성형 인공지능 사용 경험을 심층 면담으로 분석한 연구는 청소년이 이를 학습과 생활의 도구로 적극 활용하면서도, 동시에 윤리적 사용 기준의 불명확성과 시스템 작동 원리에 대한 이해 부족을 문제로 인식한다고 보고했다(노양진, 박동성, 2024). 또래 규범과 메시지 프레임을 결합한 실험 연구에서는 비윤리적 상황에서 인공지능 챗봇의 설득 효과가 정보의 내용뿐 아니라 친밀감과 상호작용 만족과 같은 정서적 요인과 결합하여 나타날 수 있음을 보여주었다(박남기 외, 2024). 이는 대학생의 수용 판단이 개인적 효용을 넘어 사회적 기대와 도덕적 평가의 맥락 속에서 형성된다는 점을 시사한다.

학교 제도와 평가 절차처럼 규칙과 공정성이 중요한 상황에서는 문제가 더욱 복합적으로 전개된다. 교육부는 2025년부터 영어, 수학, 정보 교과를 중심으로 인공지능 디지털교과서를 단계적으로 도입하겠다는 계획을 발표했다(교육부, 2024). 인공지능이 학습 추천을 넘어 평가와 선발 과정에까지 개입하게 되면, 절차의 정당성, 책임 소재, 이의제기 가능성까지 함께 고려하게 된다. 대학입시 인공지능 면접을 다룬 연구에서 학생들은 공정성과 부담 감소라는 긍정적 측면을 언급하는 동시에, 데이터 기반 판단의 신뢰성에 대한 의문과 정의적 평가를 기계에 맡기는 것에 대한 거부감을 드러냈다(신나민, 장세진, 2021; 장세진, 2022).

결과의 파급력이 큰 고위험 판단 상황에서는 해악과 권리의 문제가 전면에 등장한다. 자동 에세이 채점 연구는 인간 채점과의 점수 일치가 확보되더라도 특정 조건에서 모델이 취약해질 수 있으며, 성능이 높다고 하더라도 교육 현장에 적용하는 과정에서는 위험 요소가 존재한다고 지적한다(Doewes & Pechenizkiy, 2021). 개인정보 보호 측면에서도 추천 시스템은 프로파일링과 데이터 결합을 전제로 작동

하며, 청소년 이용자의 데이터 처리 방식에 관한 조사 사례가 공개된 바 있다(Information Commissioner's Office, 2025). 국내에서는 인공지능과 사회관계망서비스 확산이 청소년 정신건강에 미치는 영향을 전제로 과의존에 관한 대응 정책이 논의되고 있다(성평등가족부, 2025). 이러한 논의는 해악 예방과 권리 보호가 인공지능 의사결정 수용도를 판단하는 중요한 기준으로 부상하고 있음을 보여준다.

이처럼 선행연구는 인공지능과 알고리즘 기반 의사결정의 수용이 상황에 따라 달라질 수 있음을 제시해 왔다. 그러나 동일한 대학생 집단이 개인적 선택 상황, 제도 운영 상황, 생명과 해악이 관련된 상황을 오가며 어떤 기준을 적용하는지, 그리고 그 차이가 어떻게 구조화되는지에 대한 통합적 설명은 여전히 제시되지 못하고 있다. 이는 인공지능 기반 의사결정 수용이 하나의 단일한 기술 태도가 아니라, 개인의 선택권이 문제 되는 장면, 제도적 절차와 규칙이 문제 되는 장면, 해악과 권리가 문제 되는 장면에서 서로 다른 규범 판단으로 나타날 수 있음을 뜻한다. 따라서 사회인지영역이론은 이러한 판단을 개인적 영역, 사회인습적 영역, 도덕적 영역으로 구분하여 비교하는 데 적절한 이론적 틀을 제공한다. 본 연구는 사회인지영역이론에 근거하여 사회적 판단을 개인적 영역, 사회인습적 영역, 도덕적 영역으로 구분하고(Turiel, 1983), 각 영역에서 나타나는 인공지능 의사결정 수용도의 양적 차이와 정당화 이유의 질적 논리를 함께 분석한다. 이를 통해 대학생 표본의 인공지능 수용을 단일한 태도가 아니라 영역에 따라 달라지는 조건부 판단 구조로 제시하고, 교육 설계와 제도 운영에서 고려해야 할 구체적 기준을 도출하고자 한다.

Ⅱ. 이론적 배경

1. 인공지능 기반 의사결정 수용의 구조와 쟁점

인공지능 기반 의사결정은 단순히 인간의 판단을 보조하는 수준을 넘어, 추천, 배정, 평가, 선발 규제 집행과 같은 제도적 절차에서 특정 결론을 도출하는 방식으로 작동한다. 따라서 인공지능 의사결정에 대한 수용은 기술에 대한 호감이나 일반 태도에 그치지 않고, 인공지능이 제시한 판단을 따르거나 제도적으로 허용하려는 태도와 의도까지 포함하는 개념으로 보아야 한다. 네덜란드 시민을 대상으로 한 연구에 따르면, 자동화된 의사결정 시스템에 대한 수용은 하나의 요인으로 결정되기 보다, 투명성에 대한 우려가 공정성과 책무성, 프라이버시 인식에 영향을 주고, 이러한 인식이 신뢰와 유용성을 거쳐 수용 의도에 영향을 미친다(Aysolmaz et al., 2023). 이는 인공지능 의사결정 수용이 기술적 편의나 효율로 결정되지 않으며, 사용자가 무엇을 위험으로 인식하고 어떤 보호 조건이 마련되어 있는지에 따라 달라질 수 있음을 보여준다. 따라서 본 연구는 인공지능 기반 의사결정 수용도를 단순한 선호가 아니라, 신뢰, 공정성, 설명가능성, 책임성, 통제가능성과 같은 하위 준거를 바탕으로 인공지능의 판단

을 수용하거나 허용하려는 태도와 의도로 이해한다.

신뢰는 인공지능 의사결정 수용을 매개하는 주요 요인으로는 꼽힌다. 신뢰는 단순히 시스템이 정확하다는 평가가 아니라, 불확실성과 취약성이 존재하는 상황에서 그 시스템이 내 목표 달성에 도움이 될 것이라는 기대와 태도를 포함하는 개념이다(Lee & See, 2004). 인간과 인공지능이 협력하여 과업을 수행하는 환경에서는 신뢰가 무비판적 순응이 아니라, 어느 정도까지 의존할 것인지를 조정하는 판단 기준으로 기능한다. 또한 상호작용의 경험이 축적되면 특정 상황에서 형성된 신뢰가 다른 맥락으로 확장되거나, 반대로 제한되는 양상이 나타난다(National Academies of Sciences, Engineering, and Medicine, 2022). 따라서 동일한 시스템에 대해서도 일부는 쉽게 수용하고 일부는 거부하는 현상은 단순한 정보 격차로 설명되지 않는다. 위험에 대한 인식, 기대가 충족되거나 어긋나는 경험, 그리고 그에 따른 신뢰의 재조정 과정이 복합적으로 작용한다. 결국 수용은 고정된 태도가 아니라, 신뢰를 형성하고 수정하는 동적인 과정으로 이해해야 한다.

신뢰를 형성하거나 약화시키는 조건으로는 투명성과 책무성, 공정성과 편향성 문제가 특히 중요하게 논의된다. 투명성은 단순한 정보 공개가 아니라, 개인이 자신의 삶에 영향을 미치는 결정이 어떤 근거와 과정으로 이루어졌는지 이해할 수 있게 하는 조건으로 제시된다(Cheong, 2024). 국내 논의에서도 자동화된 결정이 권리 보호와 직결되기 때문에, 설명을 요구할 권리와 설명 가능성이 알고리즘 책무성의 주요 쟁점으로 다뤄지고 있다(황용석, 김기태, 2020). 또한 공공영역에서의 알고리즘 의사결정은 투입, 과정, 산출의 정당성을 동시에 위협할 수 있기 때문에, 이를 보완하기 위해 법적 장치, 시민 참여, 감시와 같은 제도적 장치가 필요하다는 논의가 제기된 바 있다(Grimmelikhuijsen & Meijer, 2022). 이러한 논의는 인공지능 의사결정 수용이 편의성에 대한 반응이 아니라, 절차가 이해 가능하고 책임이 명확하며 이의제기와 구제가 가능한지에 대한 평가와 긴밀히 연결되어 있음을 보여준다.

인간 감독과 통제의 설계, 그리고 프라이버시와 데이터 거버넌스 조건 또한 수용을 결정하는 중요한 요소다. 법 제도는 흔히 인간 감독을 안전장치로 제시하지만, 실제 현장에서는 사람들이 기계 판단에 과도하게 의존하는 자동화 편향이 나타날 수 있고, 이는 오류 탐지 실패와 의사결정 질 저하로 이어질 수 있다(Ruschemeier & Hondrich, 2024). 학습 추천 맥락에서도 사용자가 조정할 수 있는 통제 기능과 그 영향의 가시화가 신뢰를 높일 수 있다는 실험 결과가 보고되는데(Ooge et al., 2023), 이는 수용을 높이기 위해 인간 감독의 존재를 선언하는 것만으로는 부족하며, 사용자가 이해하고 조정할 수 있는 경험을 제공해야 한다는 점을 시사한다. 영국 정보위원회에서는 추천 시스템을 개인정보와 프로파일링을 활용해 콘텐츠를 제안하고 제공하는 알고리즘 과정으로 정의하면서, 아동과 청소년 보호를 위해 기본값을 제한하여 해악을 예방할 것을 강조한다(Information Commissioner's Office, 2025). 국내 공공행정 연구 또한 자동화 도입이 행정 효율을 높이는 동시에 개인정보 처리와 편향, 보안 위험을 동반한다고 지적한다(홍승헌, 황하, 2024). 대학생 집단에서도 기술의 필요성을 높게 인식하더라도 진로나 평가처럼 결과의 무게가 큰 장면에서는 수용이 낮아질 수 있으며, 데이터 기반 평가에 대한 불신이 거부의 핵심 이유로 제시된다는 점도 보고된다(신나민, 장세진, 2021; 권다남, 허나원, 강주현, 2023). 이처럼

인공지능 의사결정 수용은 신뢰를 중심으로 형성되지만, 공정성에 대한 인식, 편향에 대한 우려, 감독의 실효성, 통제 경험, 프라이버시 보호 규범이 결합한 조건부 동의의 구조로 이해할 필요가 있다.

2. 사회인지영역이론의 핵심 개념과 영역 구분

사회인지영역이론은 사회적 판단을 하나의 규범을 일괄적으로 적용하는 과정으로 보지 않는다. 대신, 상황에 따라 서로 다른 판단 기준이 작동하는 여러 영역으로 구분해 이해해야 한다는 관점에서 출발한다. 엘리엇 튜리엘(Elliot Turiel)은 사회적 지식을 “개인이 사회적 환경과의 상호작용을 통해 발달시키는 것”이라고 설명하면서, 단순한 규칙의 내면화가 아니라 사회적 경험을 해석하고 조직하는 과정에서 구성되는 지식으로 본다(Turiel, 1983). 그는 사회인습을 “사회적 상호작용을 조정하기 위해 기능하는 행동의 균일성”으로, 도덕을 “정의, 권리, 복지에 관한 규범적 판단”으로 구분하면서, 도덕규범은 특정 권위나 맥락에 의해 정의되지 않는다고 강조한다(Turiel, 1983). 에두아르 마셰리(Edouard Machery)와 스티븐 스티치(Stephen P. Stich)는 이러한 전통을 계승하여 도덕과 인습의 구분이 “심리적으로 실제이며 철학적으로 중요한 구분”이라는 점을 정리하고, 튜리엘 전통의 연구가 도덕 위반을 해악, 부정의, 권리 침해와 연결해 이해해 왔음을 제시한다(Machery & Stich, 2022). 결국 도덕 판단은 규칙을 따르는지의 문제가 아니라, 해악과 권리라는 내용 기준에 근거한 평가라는 점이 이 이론의 핵심이다. 이러한 구분은 인공지능 기반 의사결정에도 직접 적용될 수 있다. 인공지능의 판단이 개인의 취향과 생활 선택에 관여할 때에는 개인적 영역의 문제로, 공동체의 운영 절차와 배분 기준에 개입할 때에는 사회인습적 영역의 문제로, 생명과 안전, 권리와 복지에 영향을 미칠 때에는 도덕적 영역의 문제로 이해할 수 있기 때문이다.

최근 연구는 도덕 개념을 어떻게 정의할 것인가를 더 정교하게 다룬다. 아우둔 달(Audun Dahl)은 도덕을 “타인의 복지, 권리, 공정성, 정의에 대한 의무적 관심”으로 규정하며, 이를 규칙 체계나 사회적 합의로 단순화할 수 없다고 주장한다(Dahl, 2023). 동시에 그는 권위와 무관하게 판단한다거나 보편적으로 적용 가능하다는 특징이 도덕의 정의 자체라기보다 도덕 판단과 함께 나타나는 경험적 경향일 수 있음을 지적한다. 유혜나와 주디스 스메타나(Judith G. Smetana)의 메타분석은 아동이 도덕 위반을 사회인습 위반보다 더 권위와 무관하게, 더 일반화 가능하고, 더 변경하기 어려운 것으로 평가한다는 점을 종합적으로 보여준다. 이들은 “도덕과 인습을 구분해 이해하는 것은 발달에서 중요한 이정표”라고 정리하며(Yoo & Smetana, 2022), 도덕 영역이 해악과 권리 보호를 중심으로 조직된 독자적 판단 체계임을 경험적으로 뒷받침한다.

사회인습 영역과 개인적 영역은 판단의 초점과 정당화 방식에서 분명한 차이를 보인다. 사회인습 영역은 공동체의 규칙, 역할, 절차, 질서 유지와 관련된 사안을 중심으로 하며, 판단은 맥락과 제도적 합의에 의존한다(Turiel, 1983). 반면 개인적 영역은 취향, 생활 방식, 사적 선택처럼 개인의 자율권이 크게 인정되는 사안으로 구성된다. 알레그라 미드게트(Allegra J. Midgette)는 청소년이 가사노동 분담

의 공정성을 판단할 때, 평등과 형평에 근거한 도덕적 정당화와 사회적 관습에 근거한 정당화를 연령에 따라 다르게 사용함을 보여주었다. 로라 레이 레이크(Laura Wray-Lake) 등도 의사결정 자율성이 신중, 인습, 개인적, 복합 영역에 따라 서로 다른 경로로 확장된다고 보고하면서, 동일한 행위라도 어느 영역의 문제로 인식되는지에 따라 판단과 자율성 인정 범위가 달라진다고 설명한다(Wray-Lake, et al., 2010). 이는 사회적 사건이 고정된 범주에 자동으로 배치되는 것이 아니라, 판단 주체가 어떤 기준을 적용하느냐에 따라 영역이 구성된다는 점을 시사한다. 본 연구의 인공지능 시나리오에 이를 적용하면, 콘텐츠 추천과 같은 장면은 개인의 취향과 선택권이 중심이 되므로 개인적 영역에 가깝고, 기숙사 선발이나 수강신청 배정처럼 규칙과 절차의 정당성이 핵심이 되는 장면은 사회인습적 영역에 해당하며, 자율주행차나 자율무기, 장기이식 우선순위처럼 생명과 피해, 권리 보호의 문제가 전면화되는 장면은 도덕적 영역에 해당한다.

국내 연구 역시 영역 구분의 이론적 방법론적 함의를 확장해 왔다. 이상희(2024)는 사회인지영역이론이 콜버그 단계이론에 대한 단순한 비판이 아니라, 이질적인 사고가 비동시적으로 나타나는 현상을 설명하는 대안 이론임을 강조한다. 이인재(2021)는 영역 중첩과 정보적 가정을 논의하며, 동일한 사건도 전제된 정보와 맥락 이해에 따라 도덕 판단이 달라질 수 있음을 지적한다. 또한 이승민과 설선혜(2019)는 규범 분류가 항상 안정적이라고 가정하기 어렵다고 보고하며, 허용성, 권위 독립성, 일반성 같은 기준과 함께 응답자의 정당화 이유를 분석할 필요성을 제기한다. 이러한 논의는 인공지능 의사결정 수용을 분석할 때, 수용을 단일한 기술 태도로 환원하지 않고, 개인적 선택의 문제로 읽히는지, 제도적 절차의 문제로 해석되는지, 혹은 해악과 권리 보호의 문제로 이해되는지를 구분하여 비교해야 함을 이론적으로 뒷받침한다. 특히 영역 간 중첩 가능성을 고려하여, 응답자가 어떤 준거를 동원해 판단을 정당화하는지까지 분석하는 방법론적 접근이 필요하다는 것을 보여준다.

3. 청소년 맥락의 관련 선행연구 및 연구 공백

대학생의 인공지능 기반 의사결정 경험은 더 이상 특정 서비스에 국한되지 않으며, 정보 소비, 학습, 관계 형성, 제도 참여 전반으로 빠르게 확산하고 있다. 예컨대 미국 청소년 조사에서는 “청소년의 약 3분의 2가 인공지능 챗봇을 사용해 본 경험이 있다”라는 결과가 보고되었는데, 이는 인공지능 매개 판단이 이미 청년층과 대학생을 포함한 젊은 세대의 일상적 선택 환경에 구조적으로 편입되었음을 보여준다(Faverio & Sidoti, 2025). 또한 청소년이 접하는 뉴스 역시 점점 더 ‘알고리즘 큐레이션’의 대상이 되고 있으나, 청소년이 뉴스 개인화를 어떻게 인식하고 학습하며 대응하는지는 아직 충분히 알려지지 않았다는 지적이 제기된다(Swart, 2021). 이러한 논의는 대학생의 인공지능 수용을 단순한 기술 친숙도나 일반적 호감도로 설명하기 어렵다는 점을 보여주며, 상황에 따라 달라지는 판단 기준을 함께 분석해야 할 필요성을 제기한다.

국내 연구에서도 생성형 인공지능 활용은 특히 개인적 영역에 가까운 학업 지원과 일상 문제 해결에

서 두드러진다. 조희영 등의 연구에 따르면, 청소년은 생성형 인공지능을 주로 학업 지원을 목적으로 사용하며, 사용 경험이 많은 집단일수록 인공지능에 대해 더 긍정적인 태도를 보였다(조희영, 김자미, 이원규, 2024). 한편, 노양진과 박동성은 심층 인터뷰를 통해 생성형 인공지능 이용경험의 본질을 '고슴도치의 딜레마'로 설명하면서, 도움에 대한 기대와 경계의 태도가 동시에 존재하는 양가적 구조를 밝혔다(노양진, 박동성, 2024). 더 나아가 윤유빈 등은 청소년활동 기반 디지털 리터러시 프로그램이 정보 활용과 콘텐츠 표현 영역에서 유의미한 향상을 보였다고 보고한다(윤유빈, 김용갑, 문성호, 2025). 이러한 결과는 인공지능 수용이 고정된 성향이 아니라, 교육 경험과 학습 기회에 따라 변화할 수 있는 특성임을 시사한다. 그러나 이들 연구는 주로 학습 보조와 개인적 활용에 초점을 맞추고 있어, 동일한 응답자 표본 안에서 제도 운영 장면이나 고위험 상황에서 어떤 다른 기준으로 판단하는지에 대한 비교 분석은 충분히 이루어지지 않았다.

사회인습적 장면에서는 개인적 편의보다 절차의 공정성과 정당성이 수용 판단의 주요 기준이 된다. 인공지능 대학입시 면접에 대한 고등학생 인식 연구에서는 학생들은 수용을 주저한 이유로 '인공지능이 사용하는 데이터에 대한 불신'과, 감정이나 인성과 같은 '정의적적 영역을 기계가 평가하는 데 대한 거부감'을 제시하였다(신나민, 장세진, 2021). 후속 연구 역시 인공지능 면접에 유리한 학생상과 준비 요소를 분석하면서, 실제 평가 상황에서는 인공지능 대한 이해 수준과 말하기에 대한 불안이 함께 작용한다는 점을 보여주었다(장세진, 2022). 한편 교육부가 2025년부터 일부 학년과 교과에서 인공지능 디지털교과서를 본격적으로 활용하겠다고 발표한 것은, 인공지능 기반 추천과 분석이 학교 운영의 공식 절차 안으로 편입되고 있음을 의미한다(교육부, 2024). 해외에서는 추천 시스템을 아동 개인정보 보호와 직접 연결하여 규율하려는 움직임이 나타나고 있으며, 영국 정보위원회는 13세에서 17세 이용자의 개인정보가 추천 시스템에서 어떻게 처리되는지에 대한 조사에 착수했음을 공개하였다(Information Commissioner's Office, 2025). 그러나 이러한 논의는 제도적 맥락의 중요성을 확인하는 데에는 기여했지만, 개인적 장면, 제도적 장면, 고위험 장면을 분절적으로 다루는 경향이 강했으며, 동일한 표본 내에서 각 장면을 사회인지영역이론의 틀로 비교하여 어떤 준거가 수용을 정당화하는지를 체계적으로 분석한 연구는 여전히 제한적이다.

도덕적 장면에서는 해악 예방, 권리 보호, 공정성과 같은 규범이 전면에 등장하며, 개인 특성에 따른 민감도 차이도 보고된다. 인공지능 윤리의식 연구는 여학생이 인공지능에 대한 신뢰를 더 낮게 평가하고 차별금지에 대한 우려를 더 크게 나타내는 경향이 있음을 보고한다(김귀식, 신영준, 2021). 이는 공정성과 보호의 기준이 성별에 따라 다르게 작동할 수 있음을 보여준다. 또한 비윤리적 행동 상황에서 챗봇의 메시지 방식이 친밀감과 상호작용 만족도에 영향을 미친다는 결과는, 도덕적 판단이 규범적 원칙에 의해 결정되는 것이 아니라 정서적 단서와 함께 형성될 수 있음을 시사한다(박남기 외, 2024). 건강 영역에서는 청소년이 생성형 인공지능의 의료 활용에 대해 경계하는 태도를 보인다는 조사 결과가 제시되었으며(Schaaff et al., 2025), 정신건강 조언 상황에서는 약 8명 중 1명이 인공지능 챗봇을 사용한다는 수치가 보고되면서 안전성과 책임 문제도 함께 제기되고 있다(RAND Corporation,

2025). 이처럼 기존 연구는 각 장면의 특성을 개별적으로 밝혀 왔지만, 동일한 응답자 표본을 대상으로 개인적 영역, 사회인습적 영역, 도덕적 영역을 하나의 측정 구조 안에서 비교하고, 수용 수준의 차이와 정당화 이유를 함께 분석한 연구는 아직 드물다. 더욱이 현재까지의 논의는 청소년이나 특정 이용자 집단에 집중되어 있어, 대학생 표본을 대상으로 인공지능 기반 의사결정 수용의 영역별 차이를 탐색적으로 검토한 연구는 충분히 축적되지 않았다. 따라서 본 연구는 대학생 표본을 대상으로 개인적, 사회인습적, 도덕적 장면에서 인공지능 의사결정 수용이 어떻게 달라지는지, 그리고 그 판단이 어떤 정당화 논리와 연결되는지를 탐색적으로 분석함으로써 기존 논의의 공백을 보완하고자 한다.

Ⅲ. 연구 방법

1. 연구모형과 가설

본 연구는 대학생의 인공지능 의사결정 수용을 하나의 고정된 태도로 보지 않는다. 대신 판단 대상이 개인적 선택인지, 제도적 절차인지, 해악과 권리의 문제인지에 따라 수용 방식이 달라질 수 있다는 가정에서 출발한다. 사회인지영역이론에 따르면 개인적 영역에서는 자율성과 선택 권한이 준거로 작동하고, 사회인습적 영역에서는 규칙의 정당성과 절차의 공정성, 책임의 귀속이 중요한 판단 기준이 된다. 도덕적 영역에서는 해악 예방과 권리 보호가 우선적인 준거로 작동할 수 있다(Turiel, 1983). 또한 대학생의 알고리즘 사용 경험이 일상적되었더라도, 알고리즘을 이해하고 평가하는 방식은 장면에 따라 달라질 수 있다는 점이 보고되어 왔다(Swart, 2021). 이에 본 연구는 시나리오를 개인적, 사회인습적, 도덕적 영역으로 구분하고, 각 영역의 수용도를 반복측정 방식으로 비교한다. 동시에 수용과 거부의 이유를 내용분석으로 도출하여, 정량적 결과와 정성적 논리를 통합적으로 해석하는 혼합방법 모형을 적용한다. 즉 본 연구의 연구모형은 영역별 수용도 평균의 차이를 검증하는 정량 분석과, 각 영역에서 제시된 정당화 이유의 차이를 비교하는 정성 분석으로 구성된다.

이러한 연구모형에 근거하여 다음과 같은 가설을 설정한다.

첫째, 영역 평균 차이에 관한 가설이다. 가설 1은 인공지능 의사결정 수용도가 영역 간에 유의한 차이를 보일 것이라는 것이다. 개인적 영역에서는 편의성과 자기결정감이 결합하여 비교적 높은 수용이 나타날 가능성이 크다. 사회인습적 영역에서는 절차적 정당성과 책임 조건이 충족될 때 수용이 강화될 것으로 예상된다. 반면 도덕적 영역에서는 해악과 권리 판단이 중심이 됨녀서 낮은 수용이나 조건부 수용이 나타날 가능성이 있다(Turiel, 1983). 이 가설은 영역별 수용도 평균을 비교하는 반복측정 분석을 통해 검증한다.

둘째, 정당화 논리 차이에 관한 가설이다. 가설 2는 수용과 거부의 정당화 이유가 영역별로 체계적인

차이를 보일 것이라는 것이다. 개인적 영역에서는 자율성과 통제감, 유용성이 주요 근거로 제시될 가능성이 높다. 사회인습적 영역에서는 절차에 대한 신뢰와 공정성, 책임과 이의제기 가능성이 중심 논리로 작동할 것이다. 도덕적 영역에서는 해악 가능성과 보호의 필요성, 권리와 정의가 중심 준거로 등장할 것으로 예상된다(Turiel, 1983). 이 가설은 개방형 응답의 내용분석을 통해 영역별 상위 정당화 범주와 대표 진술을 비교하는 방식으로 검증한다.

셋째, 개인 특성의 탐색적 영향에 관한 가설이다. 가설 3은 인공지능 사용 경험 수준과 성별과 같은 개인 특성이 영역별 수용도와 정당화 논리의 강도에 영향을 미칠 수 있다는 것이다. 특히 제도적 장면이나 고위험 장면에서는 사용 경험이 자동으로 수용 증가로 이어지지 않고, 상황에 따라 다른 방향으로 작동할 가능성이 있다(Swart, 2021). 이 가설은 개인 특성별 영역 평균 비교와 정당화 이유의 분포를 함께 살펴보는 탐색적 분석으로 검토한다.

2. 연구대상과 자료수집 절차

본 연구의 대상은 강원대학교 사범대학 3학년 재학생 43명으로, 응답자의 연령 범위는 22세에서 24세였다. 본 표본은 강원대학교 사범대학 3학년 재학생을 대상으로 한 편의표집이다. 이들은 대학생 집단으로서 개인적 자율성이 확대되는 동시에 학교와 제도 경험이 축적되고, 권리와 책임에 대한 인식이 심화하는 단계에 해당한다. 따라서 사회적 판단을 서로 다른 영역으로 구분해 처리한다는 사회인지 영역이론의 관점을 경험적으로 검토하기에 적합한 집단적 특성을 지닌다(Turiel, 1983). 또한 이들은 알고리즘 기반 추천과 생성형 인공지능을 학습과 일상에서 지속적으로 경험해 온 세대이며, 대학 입시와 학교 제도 맥락에서 인공지능 활용을 직접 체감한 집단이라는 점에서도 연구 목적과 부합한다. 연구 참여는 자발적 동의에 근거하여 이루어졌으며, 연구 목적과 설문 구성, 개인정보 보호 절차에 대한 설명을 제공한 뒤 참여 의사를 확인하였다.

자료수집은 2025년 11월 온라인 설문 도구를 통해 실시하였다. 설문은 시나리오를 활용한 혼합방법으로 구성되었다. 먼저, 인공지능 사용 시간과 이용 서비스 유형 등 기본 이용 특성을 묻는 문항을 포함하였다. 다음으로, 개인적 영역, 사회인습적 영역, 도덕적 영역을 대표하는 장면을 제시하고, 각 장면에서 인공지능이 결정을 추천하거나 자동 조정하는 상황에 대한 수용 정도를 5점 척도로 응답하도록 하였다. 또한, 각 판단의 이유를 자유롭게 서술하도록 하여 정당화 논리를 수집하였다. 보상 제공을 위한 연락처 정보는 설문 응답 자료와 분리하여 저장하였으며, 분석 단계에서는 비식별 처리된 데이터만을 활용하였다.

3. 측정도구 구성과 문항 설계

본 연구의 측정도구는 인공지능 의사결정 수용을 단일한 기술 태도로 보지 않고, 장면의 성격과 위험 수준에 따라 수용 판단이 어떻게 달라지는지를 비교할 수 있도록 설계하였다. 자동화된 의사결정은 개인화 추천과 같은 일상적 상황에서 출발하여 의료나 고위험 판단과 같은 중대한 영역으로 확장될 수 있으며, 이러한 장면 차이는 신뢰와 수용 평가에 직접적인 영향을 미친다(Orbán & Stefkovics, 2025). 또한 공정성에 대한 인식과 수용 판단은 맥락과 설계 조건에 따라 달라질 수 있다는 점도 시나리오 기반 실험 연구에서 확인되었다(Kern et al., 2022). 사회인지영역이론 역시 실제와 유사한 가상 상황을 제시하고, 그에 관한 판단과 정당화 이유를 함께 수집하는 방식을 취한다(Smetana & Ball, 2018). 이에 따라 본 연구는 개인적, 사회인습적, 도덕적 영역을 대표하는 시나리오를 구성하고, 각 장면에서 인공지능에 결정을 위임하는 것에 대한 수용도를 정량적으로 측정하는 동시에 판단의 근거를 개방형으로 수집하였다. 본 도구는 하나의 동질적 심리척도라기보다, 동일한 응답자가 서로 다른 사회인지영역의 장면을 어떻게 판단하는지를 비교하기 위한 시나리오 기반 판단 도구라고 볼 수 있다. 따라서 구성타당도는 단일 내적일관성 계수에만 의존하기보다, 사회인지영역이론에 근거한 문항 구성의 적절성과 영역 분류의 타당성을 중심으로 해석하였다.

측정도구는 세 부분으로 구성된다. 첫째, 성별과 인공지능 사용 시간, 사용 서비스 유형, 주요 사용 목적 등 배경 변인을 포함하여 개인 특성에 따른 차이를 분석할 수 있도록 하였다. 둘째, 영역별로 3개씩 총 9개의 시나리오를 제시하고, 각 시나리오에 대해 “인공지능이 해당 결정을 추천하거나 자동 조정하도록 허용하는 것에 동의하는가”를 5점 리커트 척도로 응답하게 하였다. 셋째, 각 판단 직후 그 이유를 서술하도록 하여 정당화 논리를 확보하였다. 이와 같은 구성은 대학생이 알고리즘을 맥락에 따라 다르게 이해하고 해석할 수 있다는 점을 전제로 한 것이다. 또한 문항의 내용타당도를 확보하기 위해 전문가 3인 이상의 검토를 통해 각 시나리오의 영역 분류 적절성과 문항 표현의 명료성을 점검하고, CVI 또는 전문가 합의 절차를 논문에 제시하도록 하였다. 가능하면 예비조사 또는 인지면접을 통해 문항 이해 가능성과 응답 부담을 추가로 확인하였음을 기술할 필요가 있다. <표 1>은 연구윤리 문항과 배경 변인을, <표 2>는 영역별 시나리오 수용도 문항과 개방형 이유 문항의 구성을 정리한 것이다.

〈표 1〉 배경변인 및 윤리 관련 문항

측정영역	문항 번호	문항
연구윤리	E1	개인정보 수집 및 이용 동의
	E2	경품 전송을 위한 연락처
배경변인	B1	성별
	B2	하루 평균 인공지능 기능 사용 총 시간
	B3	사용하는 인공지능 기반 서비스
	B4	인공지능 주요 사용 용도

〈표 2〉 영역별 시나리오 수용도 문항

사회인지영역	문항 번호	문항(시나리오)
개인적	P1	유튜브나 넷플릭스에서 인공지능이 시청 기록을 바탕으로 다음 시청 콘텐츠를 자동으로 추천하고, 추천 목록을 사실상 시청목록으로 확정한다.
	P2	스마트워치가 수면 패턴을 분석해 취침·기상 시간을 자동 조정, 수면 일정 관리 결정을 AI에게 맡김
	P3	경력 컨설팅 챗봇이 추천한 직업을 우선 지원하도록 설정하고, 진로 선택의 우선순위를 인공지능의 판단에 맡긴다.
사회인습적	S1	기숙사 선발에서 인공지능이 통학 거리, 성적, 경제적 배경을 종합해 지원자의 입사 우선순위를 산출하고, 그 결과가 선발에 반영된다.
	S2	수강 신청에서 인공지능이 과목 수요를 분석해 정원 조정, 대기자 처리, 수강 우선순위를 자동으로 결정하고, 그 결과가 배정에 반영된다.
	S3	도심 교통신호 운영에서 인공지능이 교통량을 실시간으로 분석해 신호주기와 우선순위를 자동으로 조정한다.
도덕적	M1	자율주행차가 충돌을 피할 수 없는 상황에서 차량 내 인공지능이 탑승자와 보행자 중 누구의 피해를 최소화할지 선택한다.
	M2	전쟁 상황에서 자율무기체계가 목표를 식별하고 공격 여부를 인공지능이 스스로 판단해 실행하도록 허용한다.
	M3	장기이식에서 인공지능이 생존 가능성과 적합도를 예측해 수혜자 우선순위를 산출하고, 그 결과가 배정에 반영된다.

시나리오 내용은 각 영역의 특성이 분명히 드러나도록 구성하였다. 개인적 영역에는 시청 기록 기반 추천, 수면 일정 자동 조정, 진로 상담 챗봇 추천과 같이 개인 선택과 자율성이 중심이 되는 상황을 포함하였다. 사회인습적 영역에는 기숙사 선발, 수강 정원 조정, 교통신호 제어처럼 제도 운영과 절차적 공정성이 핵심이 되는 장면을 제시하였다. 도덕적 영역에는 자율주행차의 충돌 상황 선택, 자율무기체계의 공격 결정, 장기이식 우선순위 결정과 같이 해악과 권리 문제가 전면화되는 고위험 상황을 포함하였다. 이러한 배열을 통해 동일한 응답자가 서로 다른 영역의 판단을 어떻게 구분하는지 비교할 수 있도록 설계하였다. 다만 개인적 영역의 P3 문항은 장기적 결과와 부담을 수반하는 장면이라는 점에서 다른 개인적 문항보다 위험성이 높다. 그럼에도 본 연구에서는 최종 결정권이 제도나 타인이 아니라 개인에게 남아 있다는 점에 주목하여 이를 개인적 영역의 확장 사례로 포함하였다.

개방형 응답의 코딩 범주는 기존 자동화 의사결정 수용 연구와 신뢰 연구에서 제시된 준거를 반영하여 구성하였다. 투명성, 공정성, 책임성, 프라이버시는 수용 의도에 영향을 미치는 주요 인식 경로로 보고되어 왔으며(Aysolmaz et al., 2023), 신뢰는 불확실한 상황에서 의존을 조정하는 태도로 정의된다(Lee & See, 2004). 이에 따라 자율과 통제, 유용성과 편의, 정확성과 오류 가능성, 공정성과 편향, 투명성과 설명 가능성, 책임 소재와 이의제기 가능성, 프라이버시 우려, 해악 가능성과 보호 필요를 기본 코딩 범주로 설정하였다. 코딩의 일관성을 확보하기 위해 코드북을 작성하고, 2인의 코더가 개방형 응답을 독립적으로 코딩한 뒤 불일치 항목은 협의를 통해 조정하였다. 코딩 신뢰도는 Cohen의 kappa 또는 Krippendorff's alpha를 산출하여 제시하도록 하였다. 특히 대학생 맥락에서도 통제 경험과 투명성 인식이 신뢰를 강화할 수 있다는 연구 결과를 반영하여(Ooge et al., 2023), 사회인습적 영역에서는 절차 조건과 통제 경험을 별도로 구분해 코딩하였다. 도덕적 영역의 코딩은 도덕을 타인의 복지와 권리, 공정성에 대한 규범적 판단으로 정의하는 사회인지영역이론의 기준에 따라 구성하였다(Turiel, 1983; Smetana & Ball, 2018).

Ⅳ. 분석 결과

1. 기술통계 및 측정도구의 신뢰도와 타당도 점검

설문 응답 43부를 분석한 결과, 대학생 43명의 9개 시나리오 수용도 문항은 모두 결측치 없이 1점에서 5점 범위에서 응답이 수집되었다. 전체 평균 수용도는 3.19점(표준편차 0.43)이었다. 문항별 평균을 살펴보면, 개인적 영역에서는 콘텐츠 추천 자동 결정이 4.07점으로 비교적 높게 나타났고, 사회인습적 영역에서는 기숙사 선발 우선순위 산출이 4.19점으로 가장 높은 값을 보였다. 반면 도덕적 영역에서는 자율주행 차량의 피해 최소화 선택이 2.00점, 자율무기 공격 결정이 2.21점으로 낮게 나타났다. 이는 권리와 해악이 직접적으로 문제가 되는 장면에서 인공지능 의사결정에 대한 수용이 상대적으로 약화되는 경향을 보여준다. 이러한 경향은 대학생 표본에서도 인공지능 의사결정 수용이 장면의 성격에 따라 달라질 수 있음을 시사한다.

〈표 3〉 대학생 표본의 기본 특성(N=43명)

구분	범주	빈도	비율(%)
성별	여성	23	53.5
	남성	20	46.5
하루 평균 AI사용 시간	1시간 미만	17	39.5
	1에서 3시간	20	46.5
	3시간 이상	6	14

〈표 4〉 시나리오 문항별 수용도 기술통계(N=43)

영역	문항	평균	표준편차
개인적	콘텐츠 추천 자동 결정	4.07	0.83
	수면 일정 자동 조정	3.26	1.31
	진로 추천 우선 지원	2.44	1.08
사회인습적	기숙사 선발 우선순위	4.19	1.03
	수강 정원 및 우선순위	3.70	1.06
	교통신호 운영	3.70	1.26
도덕적	자율주행 피해 최소화	2.00	1.09
	자율무기 공격 결정	2.21	1.12
	장기이식 우선순위	3.16	1.45

측정도구의 내적일관성을 점검한 결과, 전체 9문항의 크롬바흐 알파는 0.24로 나타났으며, 영역별로는 개인적 영역이 음수, 사회인습적 영역 0.14, 도덕적 영역 0.45였다. 다만 본 도구는 동일한 내용을 반복 측정하는 단일 척도라기보다, 동일한 대학생 응답자가 서로 다른 사회인지영역의 장면을 어떻게

판단하는지를 비교하기 위한 시나리오 기반 판단 과제의 묶음에 가깝다. 이는 본 도구가 동일한 내용의 문항을 반복하여 하나의 단일 구성개념을 측정하는 척도라기보다, 영역 내에서도 서로 다른 상황을 제시하는 시나리오 묶음으로 구성되었기 때문으로 해석된다. 따라서 본 연구에서는 Cronbach's alpha를 절대적 판단 기준으로 사용하기보다 참고치로 제시하고, 영역 평균뿐 아니라 문항별 평균과 표준편차를 우선적으로 검토하며, 개방형 정당화 이유를 함께 분석하여 해석의 타당도를 보완하는 전략을 취한다. 즉 낮은 알파 값은 곧바로 측정도구의 실패를 의미한다기보다, 이질적인 장면을 포함한 시나리오 기반 판단 도구의 특성을 반영한 결과로 이해할 필요가 있다. 내용 타당도는 사회인지영역이론의 정의에 따라 개인적, 사회인습적, 도덕적 준거를 대표하는 장면을 구성하고, 각 장면에 대해 수용 판단과 정당화 이유를 동시에 수집했다는 점에서 확보하였다. 또한 전문가 검토 절차를 통해 각 시나리오의 영역 분류 적절성과 문항 표현의 명료성을 점검하였다는 점도 내용타당도의 근거로 제시할 수 있다. 구성타당도 역시 단일 내적일관성 계수만이 아니라, 사회인지영역이론에 따라 서로 다른 장면에서 판단 차이가 나타나는지를 확인하는 방식으로 해석하였다.

〈표 5〉 영역별 평균 수용도와 내적일관성 참고치(N=43명)

영역	문항 수	평균	표준편차	Cronbach's Alpha	평균 항목 간 상관
개인적	3	3.26	0.60	−0.15	−0.04
사회인습적	3	3.86	0.68	0.14	0.07
도덕적	3	2.46	0.85	0.45	0.23
전체	9	3.19	0.43	0.24	0.04

2. 영역별 수용도 차이와 개인 특성에 따른 차이 분석

영역 간 차이의 유의성은 동일한 대학생 응답자가 세 영역에 모두 응답한 반복측정 구조라는 점을 고려하여 분석하였다. 본 연구에서는 통계 전략의 일관성을 확보하기 위해 영역 평균점수와 차이점수의 분포를 점검한 뒤, 세 영역의 평균 차이에 대해서는 반복측정 분산분석을 실시하였다. 그 결과, 영역 효과는 유의하였고, 효과크기도 중간 이상 수준으로 나타났다. 이는 동일한 인공지능 의사결정이라도 그것이 개인적 선택, 제도적 절차, 해악과 권리의 문제로 인식되는지에 따라 수용 판단이 체계적으로 달라질 수 있음을 보여준다.

〈표 6〉 영역 간 수용도 차이 검정(반복측정 분산분석)

응답자 수	영역 수	F	자유도	P	부분 η^2
43	3	43.15	2.84	〈.001	0.51

사후 비교에서도 세 영역 간 모든 쌍이 유의하게 달랐다. 사회인습적 영역은 개인적 영역보다 평균

0.60점 높았고, 사회인습적 영역은 도덕적 영역보다 평균 1.40점 높았다. 개인적 영역 역시 도덕적 영역보다 평균 0.80점 높게 나타났으며, 모든 비교에서 보정된 p값은 .001 미만이었다. 이는 제도적 절차와 운영이 중심이 되는 장면에서 수용이 가장 높고, 생명과 해악이 직접적으로 문제가 되는 고위험 장면에서 수용이 가장 낮게 나타나는 경향을 보여준다.

〈표 7〉 영역 간 사후 비교(대응표본 t 검정, Bonferroni 보정)

비교	평균차(앞-뒤)	t	보정 p
사회인습적-개인적	0.60	4.15	〈.001
사회인습적-도덕적	1.40	9.34	〈.001
개인적-도덕적	0.80	5.04	〈.001

개인 특성에 따른 차이도 함께 탐색하였다. 성별 비교에는 독립표본 t 검정을 적용하였고, 인공지능 사용 시간 비교에는 일원분산분석을 적용하였다. 성별 비교에서는 남성이 개인적 영역과 도덕적 영역에서 여성보다 다소 높은 평균을 보였으나, 영역별 차이는 모두 통계적으로 유의하지 않았다(모두 n.s). 인공지능 사용 시간은 표본 분포를 고려하여 세 범주로 구분한 뒤 분석하였으며, 세 영역 모두에서 유의한 차이는 확인되지 않았다(모두 n.s.). 이는 본 표본에서는 성별이나 사용 시간과 같은 개인 특성보다 장면의 성격과 영역 구분이 수용도에 더 큰 영향을 미쳤음을 시사한다.

〈표 8〉 개인 특성별 영역 평균 비교

개인 특성	집단	n	개인적 평균	사회인습적 평균	도덕적 평균	차이 검정
성별	남성	20	3.38	3.83	2.62	모두 n.s.
	여성	23	3.14	3.88	2.32	
인공지능 사용 시간	1시간 미만	17	3.20	3.94	2.67	모두 n.s.
	1에서 3시간	20	3.28	3.92	2.30	
	3시간 이상	6	3.33	3.44	2.39	

3. 개방형 응답 내용분석 결과와 혼합방법 통합 해석

개방형 응답은 9개 시나리오 각각에 대해 제시된 이유 서술을 분석 대상으로 삼았으며, 대학생 43명 기준 총 387개의 서술이 수집되었다. 내용분석은 사전에 설정한 9개 준거 범주에 따라 진행했으며, 각 서술은 의미 중심의 1차 범주로 단일 코딩하였다. 코딩 범주는 유용성 및 편의성, 자율성 및 선택권, 정확성 및 객관성, 공정성 및 차별, 절차와 책임 및 이의제기, 투명성 및 설명 가능성, 프라이버시, 안전과 해악 및 생명, 보안 및 해킹이다.

<표 9>는 영역별로 빈도가 높은 정당화 범주를 요약한 것이다. 개인적 영역에서는 유용성 및 편의성

과 자율성 및 선택권이 핵심 근거로 함께 등장했다. 사회인습적 영역에서는 유용성 및 편의성이 가장 많이 언급되었고, 그 다음으로 공정성 및 차별과 정확성 및 객관성이 뒤를 이었다. 도덕적 영역에서는 안전과 해악 및 생명이 압도적으로 우세했고, 자율성 및 선택권과 정확성 및 객관성은 상대적으로 낮은 빈도로 나타났다.

〈표 9〉 영역별 정당화 범주 빈도 상위 항목(영역별 응답 129개 기준)

영역	상위 정당화 범주	빈도	비율
개인적	유용성 및 편의성	62	48.1
	자율성 및 선택권	54	41.9
	정확성 및 객관성	13	10.1
사회인습적	유용성 및 편의성	43	33.3
	정확성 및 객관성	27	20.9
	공정성 및 차별	15	11.6
도덕적	안전과 해악 및 생명	60	46.5
	자율성 및 선택권	33	25.6
	절차와 책임 및 이의제기	22	17.1

개인적 영역의 응답은 저위험 생활 선택 장면에서 편의와 개인화 이점이 수용을 지지한다는 점을 보여준다. 예컨대 추천이 “관심있는 분야를 편안하게 계속 볼 수” 있게 해 준다는 진술처럼, 시간 절약과 맞춤 제공이 긍정 근거로 제시되었다. 반면 진로와 같이 결과 부담이 큰 선택에서는 “인간의 주체성의 위협”과 같은 표현이 반복되어, 위임 자체에 대한 심리적 저항이 함께 관찰되었다. 같은 개인적 영역 안에서도 선택의 되돌릴 가능성과 삶의 장기 결과가 결합될수록 자율성 준거가 강화되는 조건부 구조가 확인된다. 이는 대학생 표본에서도 개인적 장면의 수용이 일률적이지 않으며, 저위험 선택과 장기 결과를 수반하는 선택이 서로 다르게 해석될 수 있음을 보여준다.

사회인습적 영역에서는 제도 운영의 효율과 운영 안정성에 대한 기대가 수용을 뒷받침했다. 교통 신호 운영과 같은 장면에서 “면적대비 차가 많은 동네에는 유용”하다는 응답은 효율성과 공익적 편익을 직접 연결한다. 동시에 기숙사 선발과 같은 배분 장면에서는 “공정하게 진행”되기를 바란다는 진술이 나타나, 절차의 공정성과 차별 가능성에 대한 감수성이 수용 판단에 함께 작동함을 보여준다. 요컨대 사회인습적 영역의 수용은 편의만으로 성립하기보다, 공정성과 객관성에 대한 신뢰가 동반될 때 강화되는 경향이 있다. 이 결과는 대학생이 제도적 장면에서 인공지능 판단을 평가할 때에도 단순한 효율성보다 절차적 정당성을 함께 고려함을 시사한다.

도덕적 영역에서는 AI가 생명과 해악이 걸린 판단에 개입하는 것 자체에 대한 거부가 핵심이었다. “생명관련해서는 AI가 관여하기 힘들다”는 응답은 생명과 해악 준거가 최우선으로 작동함을 드러낸다. 또한 “전장에서 내 목숨을 맡기기엔 불안”하다는 진술은 통제 가능성과 책임 귀속이 충분히 확보되지 않은 상황에서 위임이 정당화되기 어렵다는 인식을 반영한다. 도덕적 영역의 낮은 수용은 기술 성능의

문제가 아니라, 해악 예방과 책임이라는 규범적 조건이 선행되어야 한다는 요구가 전면화된 결과로 해석된다. 즉 대학생 표본에서 도덕적 장면의 수용은 편의나 효율이 아니라 생명, 안전, 책임의 기준에 의해 강하게 제한되는 경향을 보였다.

〈표 10〉 영역별 수용도 평균과 주요 정당화 논리의 통합 요약

영역	평균	표준편차	주요 정당화 논리 (상위 2개)
개인 적영역	3.26	0.60	유용성 및 편의, 자율성 및 선택권
사회인습적 영역	3.86	0.68	유용성 및 편의, 공정성 및 차별
도덕적 영역	2.46	0.85	안전과 해악 및 생명, 자율성 및 선택권

정량 결과와 정성 결과를 통합하면, 영역별 평균 차이는 정당화 논리의 우선순위 차이와 일관되게 연결된다. 사회인습적 영역은 평균 수용도가 가장 높았고, 주요 근거도 유용성과 공정성에 집중되었다. 개인적 영역은 평균이 중간 수준이었는데, 편의 중심의 수용 근거와 자율성 우려가 동시에 존재하기 때문이다. 도덕적 영역은 평균이 가장 낮았고, 안전과 생명 준거가 압도적으로 우세해 위임을 원천적으로 제한하는 논리가 강하게 나타났다. 이러한 혼합방법 결과는 강원대학교 사범대학 3학년 재학생으로 구성된 대학생 표본에서 인공지능 기반 의사결정 수용이 단일한 태도가 아니라 영역별로 달라지는 조건부 판단 구조를 보인다는 점을 시사한다.

V. 결론

본 연구의 결론은 대학생 집단의 인공지능 의사결정 수용이 하나의 일반 태도로 수렴하지 않는다는 점이다. 수용 여부는 판단 장면이 개인적 선택의 문제로 인식되는지, 제도적 절차의 문제로 이해되는지, 혹은 해악과 권리의 문제로 해석되는지에 따라 체계적으로 달라졌다. 사회인지영역이론에 따르면, 사회인습은 사회적 상호작용을 조정하기 위한 '사회적 행동의 합의'이고, 도덕 영역은 '정의, 권리, 복지에 관한 규범적 판단'으로 구분된다(Turiel, 1983). 본 연구에서 사회인습적 영역의 수용도가 가장 높고 도덕적 영역의 수용도가 가장 낮게 나타난 결과는, 응답자들이 동일한 인공지능 판단을 상황에 따라 운영의 효율성 문제로 읽거나, 해악 예방과 권리 보호의 문제로 재해석했기 때문으로 이해할 수 있다. 또한 대학생의 인공지능 경험이 이미 일상화되었음에도, 알고리즘을 해석하고 의미를 부여하는 전략이 장면에 따라 달라진다는 선행연구의 지적과도 맥락을 같이한다(Swart, 2021; Faverio & Sidoti, 2025).

영역별로 보면 개인적 영역에서는 저위험 생활 선택에서 수용이 높았으나, 진로와 같이 장기적 결과와 자기정체성이 결합한 장면에서는 수용이 낮아졌다. 이는 되돌릴 수 있는 선택에서는 편의와

개인화의 이점이 수용을 강화하지만, 삶의 방향을 좌우하는 결정에서는 위임 자체가 자기결정권의 침해로 인식될 수 있음을 보여준다(Swart, 2021). 더 나아가 개인화 추천이 개인정보와 프로파일링을 전제로 한다는 점은 수용의 경계 조건을 분명히 한다. 영국 정보위원회가 추천 시스템을 개인정보를 활용해 선호를 학습하고 콘텐츠를 제안하는 알고리즘 과정으로 정의한 것은, 개인적 영역 수용이 단순한 편의 판단을 넘어 사생활과 데이터 처리의 정당성 문제와 결합할 수 있음을 시사한다(Information Commissioner's Office, 2025). 따라서 교육적으로는 개인적 영역에서 인공지능의 활용을 늘리는 것보다, 어떤 결정을 언제까지 위임할 것인지, 데이터가 어떻게 사용되는지, 사용자가 어떤 방식으로 개입하고 되돌릴 수 있는지를 중심으로 자기결정과 통제 감각을 강화하는 설계가 중요하다.

사회인습적 영역에서 상대적으로 높은 수용은 효율성 기대만으로 설명하기 어렵다. 공정성과 절차적 정당성에 대한 기대가 효율성 인식과 결합하여 수용을 지지한 결과로 해석하는 것이 타당하다. 자동화 의사결정의 수용 연구는 투명성에 대한 우려가 공정성, 책무성, 프라이버시 인식을 거쳐 신뢰와 채택 의도로 이어지는 경로를 제시한다(Aysolmaz et al., 2023). 또한 시나리오 실험 연구는 인간이 자동화만으로 결정할 때보다 인간 결정자가 최종 결정에 관여하는 방식이 더 공정하게 인식될 수 있음을 보여주며, 맥락에 따라 부분 자동화가 수용에 유리할 수 있음을 시사한다(Kern et al., 2022). 다만 인간 감독이 형식적으로 존재하는 것과 실제로 의미 있게 작동하는 것은 구분되어야 한다. 투명성은 인공지능의 판단 과정을 이해할 수 있도록 돕는 조건이며(Cheong, 2024), 자동화 편향은 인간이 기계 제안에 과도하게 의존해 검토를 약화하는 위험을 내포한다(Ruschemeier & Hondrich, 2024). 대학생의 학습 추천 맥락에서도 통제의 효과가 시각화될 때 신뢰가 증가했다는 결과는, 설명 가능성과 통제 경험이 함께 제공될 때 제도 장면에서 수용이 안정화될 수 있음을 뒷받침한다(Ooge et al., 2023). 따라서 학교와 공공영역에서 인공지능을 도입할 경우, 설명, 이의제기, 책임 귀속, 안전장치를 선택적 요소가 아니라 기본 설계 조건으로 포함할 필요가 있다.

도덕적 영역에서 수용도가 가장 낮게 나타난 결과는, 고위험 장면에서는 효율성보다 해악 예방과 책임 귀속이 우선적 판단 기준으로 작동함을 분명히 보여준다. 자동화된 평가도구는 인간 채점과 높은 점수 일치도를 보이더라도, 실제 현장에 적용하기에는 여전히 위험할 수 있다는 경고가 제기되어 왔다(Doewes & Pechenizkiy, 2021). 또한 정신건강 조언과 같이 취약성이 높은 맥락에서 청년층과 대학생이 인공지능을 실제로 활용하고 있다는 사실은, 기술 확산과 동시에 위험 관리 체계의 정교화가 시급함을 드러낸다(RAND Corporation, 2025). 국내 입시 인공지능 면접 연구 역시 데이터 기반 평가에 대한 불신과 정의적 판단을 기계에 맡기는 것에 대한 거부감이 수용을 낮추는 핵심 요인임을 보고한다(신나민, 장세진, 2021). 이는 평가와 선발처럼 개인의 미래에 직접적 영향을 미치는 장면에서는 절차적 정당성과 책임 명확성에 대한 요구가 더욱 강하게 제기된다는 점을 보여준다.

본 연구는 몇 가지 한계를 지닌다. 첫째, 강원대학교 사범대학 3학년 재학생 43명을 대상으로 한 단일 기관 편의표집 표본을 사용하였다는 점에서, 결과를 일반 대학생 전체나 청소년 일반으로 확대 해석하는 데에는 제약이 있다. 둘째, 서로 다른 성격의 시나리오를 묶어 영역별로 제시하였기 때문에

내적일관성 지표가 낮게 나타날 수 있다. 그러나 이는 본 도구가 동질적인 단일 척도라기보다 서로 다른 장면에 대한 판단을 비교하기 위한 시나리오 기반 도구라는 점과 함께 해석할 필요가 있다. 셋째, 표본 규모의 한계로 인해 성별이나 사용 경험과 같은 개인 특성 효과가 충분히 드러나지 않았을 가능성도 있다. 따라서 본 결과는 탐색적 성격을 지니며, 후속 연구를 통해 보완될 필요가 있다. 향후 연구에서는 고위험 장면에서 책임 주체의 명확성, 이의제기 가능성, 인간 개입의 실효성 등을 조건으로 체계적으로 조작하여 수용 변화 양상을 검증할 필요가 있다. 또한 통제 경험과 설명 제공이 신뢰 형성과 수용 판단을 어떻게 매개하는지에 대한 실험적 검증도 강화되어야 한다. 아울러 후속 연구는 대학생 표본을 넘어 다양한 연령대와 교육 수준의 집단을 포함하여 영역별 판단 구조를 비교함으로써, 인공지능 기반 의사결정 수용의 조건부 구조가 어떻게 달라지는지 보다 정밀하게 검토할 필요가 있다. 이를 통해 영역별 조건부 수용을 단순한 평균 점수 차이로 해석하는 수준을 넘어, 상황에 따라 정당화 준거의 우선순위가 어떻게 재구성되는지 설명하는 더 정밀한 이론적 모형으로 발전시킬 수 있을 것이다.

참고문헌

권다남 · 허나원 · 강주현 (2023). 인공지능(AI)에 대한 고등학생 인식 조사. **한국데이터분석학회지, 25**(6), 2473-2488.

김귀식 · 신영준 (2021). 초 · 중 · 고등학생의 인공지능 윤리의식의 성차 분석. **과학교육연구지, 45**(1), 105-117.

노양진 · 박동성 (2024). 청소년의 생성형 AI 이용경험에 대한 현상학적 연구. **학습자중심교과교육연구, 24**(7), 557-577.

박남기 · 피연진 · 이휘란 · 이승희 · 신제인 (2024). 청소년의 또래집단 규범이 비윤리적 행동 중단의도, 인공지능 챗봇에 대한 친밀감과 상호작용 만족도에 미치는 영향: 메시지 프레임의 조절효과. **한국언론학보, 68**(3), 156-197.

신나민 · 장세진 (2021). 대학입시 전형 AI 면접에 대한 고등학생의 인식. **한국산학기술학회 논문지, 22**(7), 242-251.

이상희 (2024). 구조 발달 관점에서 사회적 영역 이론의 함의. **도덕윤리과교육,** (84), 87-106.

이승민 · 설선혜 (2019). 도덕 판단에서 나타나는 내 외집단 차이: 위반 주체 소속집단과 위반 장소의 효과. **한국심리학회지: 사회 및 성격, 33**(1), 19-52.

이인재 (2021). 초등학생의 사회 및 도덕 판단에 대한 사회인지적 영역이론의 관점과 도덕교육에의 함의. **초등도덕교육,** (71), 257-280.

이준혁 · 김채원 · 김현정 (2025). 자동화된 의사결정 시스템의 공정성과 편향성이 수용성에 미치는 영향: 신뢰성의 매개효과와 투명성의 조절된 매개효과를 중심으로. **경영정보학연구, 27**(1),

155-177.

장세진 (2022). 대학입시의 AI 면접에 대한 고등학생들의 인식 연구. **학습자중심교과교육연구, 22**(24), 537-550.

조희영・김자미・이원규 (2024). 청소년의 생성형 인공지능 사용 경험에 따른 인공지능 교육에 대한 인식 분석. **컴퓨터교육학회 논문지, 27**(9), 15-22.

윤유빈・김용갑・문성호 (2025). 디지털 리터러시 강화를 위한 청소년활동 프로그램의 효과성 분석. **리터러시 연구, 16**(4), 35-57.

황용석・김기태 (2020). 알고리듬 기반 자동화된 의사결정의 설명 가능성에 대한 연구. **언론정보연구, 57**(3), 41-80.

홍승헌・황하 (2024). 누구를 위한 디지털 전환인가? 자동화된 복지행정의 위험성. **정부학연구, 30**(2), 61-84.

교육부 (2024).『2025년, 교실에서 마주할 인공지능 AI 디지털교과서, 모두를 위한 맞춤 교육을 실현』. 교육부. https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=101774&lev=0&m=020402 (검색일: 2026. 3. 5)

성평등가족부 (2025).『성평등가족부, 미디어 과의존 대응 정책포럼 개최』. 보도자료. https://www.mogef.go.kr/nw/rpd/nw_rpd_s001d.do?bbtSn=710698&mid=news405 (검색일: 2026. 3. 5)

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.

Machery, E., & Stich, S. (2022). The moral/conventional distinction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022 ed.). Metaphysics Research Lab, Stanford University.

National Academies of Sciences, Engineering, and Medicine (2022). *Human-AI teaming: State-of-the-art and research needs*. The National Academies Press.

Aysolmaz, B., Muller, R., & Meacham, D. (2023). The public perceptions of algorithmic decision-making systems: Results from a large-scale survey. *Telematics and Informatics*, *79*, 1-16.

Cheong, B. C. (2024). Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, *6*, 1-11.

Dahl, A. (2023). What we do when we define morality (and why we need to do it). *Psychological Inquiry*, *34*(2), 53-79.

Doewes, A., & Pechenizkiy, M. (2021). On the limitations of human-computer agreement in automated essay scoring. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)* (pp. 475-480). International Educational Data Mining Society.

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of algorithmic decision-making: Six threats and the need for a calibrated institutional response. *Perspectives on Public Management and Governance*,

5(3), 232-242.

Kern, C., Gerdon, F., Bach, R. L., Keusch, F., & Kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, *3*(10), 1-38.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.

Midgette, A. J. (2020). Chinese and South Korean children's moral reasoning regarding the fairness of a gendered household labor distribution. *Developmental Psychology*, *56*(1), 91-102.

Ooge, J., Dereu, L., & Verbert, K. (2023). Steering recommendations and visualising its impact: effects on adolescents' trust in e-learning platforms. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 156-170.

Orbán, F., & Stefkovics, A. (2025). Trust in artificial intelligence: a survey experiment to assess trust in algorithmic decision-making. *AI & SOCIETY 40*, 4955-4969.

Ruschemeier, H., & Hondrich, L. J. (2024). Automation bias in public administration-an interdisciplinary perspective from law and psychology. *Government Information Quarterly*, *41*(3), 1-10.

Schaaff, C. et al. (2025). Youth Perspectives on Generative AI and Its Use in Health Care. *Journal of Medical Internet Research*, *27*, e72197.

Smetana, J. G., & Ball, C. L. (2018). Young children's moral judgments, justifications, and emotion attributions in peer relationship contexts. *Child Development*, *89*(6), 2245-2263.

Swart, J. (2021). Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media + Society, 7*(2), 1-11.

Wray-Lake, L., Crouter, A. C., & McHale, S. M. (2010). Developmental patterns in decision-making autonomy across middle childhood and adolescence: European American parents' perspectives. *Child Development*, *81*(2), 636-651.

Yoo, H. N., & Smetana, J. G. (2022). Distinctions between moral and conventional judgments from early to middle childhood: A meta-analysis of social domain theory research. *Developmental Psychology*, *58*(5), 874-889.

Faverio, M., & Sidoti, O. (2025). *Teens, social media and AI chatbots 2025.* Pew Research Center. Retrieved March 5, 2026 from https://www.pewresearch.org/internet/2025/12/09/teens-social-media- and-ai-chatbots-2025/

Information Commissioner's Office (2025.3.3). Children's code strategy progress update March 2025. Retrieved March 5, 2026 from https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/protecting-childrens-privacy-online-our-childrens-code-strategy/children-s-code-strategy-progress-update-march-2025/

RAND Corporation (2025). *One in eight adolescents and young adults use AI chatbots for mental health advice.* RAND. Retrieved March 5, 2026 from https://www.rand.org/news/press/2025/11/one-in-eight-adolescents-and-young-adults-use-ai-chatbots-for-mental-health-advice.html

Abstract

Acceptance of AI-Based Decision-Making among University Students: A Comparative Analysis of Domain-Specific Judgments Based on Social Domain Theory

Park, Boram[1] · Noh, Sung ho[2]

[1]*Assistant Professor, Kangwon National University*
[2]*Master's Student, Kangwon National University*

This study explores whether university students' acceptance of AI-based decision-making reflects a single general attitude or a domain-specific, conditional structure depending on how a situation is interpreted within social domains. A total of 43 third-year students from the College of Education at Kangwon National University participated in a repeated-measures design using nine scenarios representing personal, social-conventional, and moral domains. The results indicated significant differences across domains (Friedman test, $p < .001$; Kendall's $W = .52$), with the highest acceptance in the social-conventional domain ($M = 3.86$), followed by the personal domain ($M = 3.26$), and the lowest in the moral domain ($M = 2.46$). Content analysis of 387 open-ended responses revealed that utility, convenience, and autonomy were central justifications in the personal domain, while utility combined with expectations of fairness and objectivity supported acceptance in the social-conventional domain. In contrast, concerns about safety, harm, and life dominated moral reasoning and strongly constrained acceptance. Drawing on Turiel's social domain theory, the findings suggest that AI acceptance is structured by distinct criteria—autonomy, procedural legitimacy, and protection from harm. The study implies that explainability, accountability, contestability, and safety should be treated as essential conditions in educational and institutional AI implementation, while acknowledging limited generalizability.

Key Words: Social Domain Theory, Acceptance of AI-Based Decision-Making, Algorithmic Recommendation, Procedural Legitimacy, Harm Prevention and Accountability