AI는 관리자 윤리를 어떻게 재구성하는가?
How Does AI Reconfigure Managerial Ethics?

김성준(주저자) · 이중학(교신저자)
Sungjun Kim(First Author) · Joonghak Lee(Corresponding Author)

국민대학교 경영대학 Kookmin University's College of Business(*leadership@kookmin.ac.kr*)
동국대학교 경영학과 Dongguk University's Business School(*joonghaklee@dgu.ac.kr*)

본 연구는 관리자가 AI와 상호작용하는 과정에서 나타날 수 있는 윤리적 경계선 행동을 탐색하고, 이것이 조직 윤리를 어떻게 재구성할 수 있는지 이론적으로 고찰한다. 전통적으로 조직 윤리는 인간 구성원 간 상호작용을 전제로 해왔으나, 생성형 AI가 조직 내에서 조언자, 협업자처럼 준사회적 기능을 수행하면서 기존 윤리적 경계는 새로운 도전에 직면하고 있다. 특히 관리자가 AI를 대상으로 성희롱적 발화와 언어폭력적 발화를 할 경우, 이것이 윤리적으로 문제가 되는지에 대한 판단은 더 이상 단순히 AI가 기계인가 사회적 존재인가 하는 이분법만으로 설명되기 어렵다. 본 연구는 철학적 논의를 바탕으로, 윤리적 판단이 행위자가 그 대상을 어떤 기능적 역할로 경험하고, 동시에 어떤 존재론적 지위에 배치하는가에 따라 달라질 수 있음을 제시한다. 관리자가 AI를 기능적으로는 판단을 제시하고 대화를 이어가는 상호작용 상대로 경험하면서도, 존재론적으로는 감정 · 권리 · 피해 경험을 갖지 않는 기계적 도구로 간주할 때, AI 대상 성희롱적 발화와 언어폭력적 발화는 윤리적 고려 범주에서 밀려날 수 있다. 이를 바탕으로, 본 연구는 관리자의 AI 인식, 성희롱 신화 수용, 성과 압박, 도덕적 구획화, 도덕적 이탈이 AI 대상 부적절 발화와 조직 내 윤리적 기준선 완화에 미치는 영향을 명제로 제시한다. 본 연구는 조직 윤리 연구의 분석 대상을 인간-AI 상호작용으로 확장하고, 관리자의 AI 대상 언행이 조직 전반의 윤리적 기준선에 미칠 수 있는 영향을 이론적으로 규명한다.

주제어: 인공지능, 관리자, 윤리, 성희롱적 발화, 언어폭력적 발화

This study theoretically examines ethical boundary behaviors that may emerge as managers interact with AI and explores how such behaviors may reconfigure organizational ethics. Organizational ethics has traditionally been grounded in interactions among human members. However, as generative AI increasingly performs quasi-social functions in organizations, such as those of an advisor or collaborator, existing ethical boundaries are facing new challenges. In particular, when managers direct sexually harassing or verbally abusive utterances toward AI, the ethical evaluation of such behavior can no longer be adequately explained by a simple dichotomy between whether AI is a machine or a social entity. Drawing on philosophical discussions, this study proposes that ethical judgment may vary depending on the functional role through which actors experience AI and the ontological status to which they assign it. Specifically, when managers functionally experience AI as an interaction partner that offers judgments and sustains dialogue, while ontologically regarding it as a mechanical tool without emotions, rights, or victim experience, sexually harassing and verbally abusive utterances toward AI may be pushed outside the scope of ethical consideration. Based on this theoretical foundation, the study develops propositions regarding how managers' perceptions of AI, sexual harassment myth acceptance, performance pressure, moral compartmentalization, and moral disengagement influence inappropriate utterances toward AI and

최초투고일: 2026. 03. 01 수정일: (1차: 2026. 05. 14) 게재확정일: 2026. 05. 18

the relaxation of ethical baselines within organizations. By extending the analytical scope of organizational ethics research to human-AI interaction, this study theoretically elucidates how managers' conduct toward AI may affect the ethical baselines of the organization as a whole.

Keyword: artificial intelligence, manager, ethics, sexually harassing utterances, verbally abusive utterances

Ⅰ. 서 론

ChatGPT와 같은 생성형 인공지능(generative AI; 이하 AI)은 인간과 기계 사이에 놓여 있던 견고한 경계를 점차 흐릿하게 만들고 있다. 오늘날 상당수 사용자는 생성형 AI를 기계이자 도구로 인식하면서도, 동시에 친구, 코치, 멘토, 심리상담가, 나아가 연인과 유사한 존재로 대하며 상호작용한다(Cardon and Marshall, 2024; Qi et al., 2025). 일터도 예외가 아니다. 직장인들은 이제 AI를 단순 반복 업무를 처리하는 자동화 도구로만 사용하지 않는다. AI는 아이디어를 함께 생성하고 발전시키는 파트너, 판단을 보조하는 조언자, 때로는 정서적 부담을 덜어내는 대화 상대로 활용되고 있다(Chatterji et al., 2025; Retkowsky et al., 2024). 관리자들 역시 구성원 면담을 준비하고, 보고서 초안을 작성하며, 전략적 대안을 검토하는 과정에서 AI와 대화하며 자신의 사고를 정제한다(Wilhelm et al., 2025). 이처럼 AI는 조직 안에서 단순한 기술적 도구를 넘어, 구성원과 상호작용하고 관계를 형성하는 준사회적 행위자(quasi-social actor)로 자리매김하기 시작했다(Gkinko and Elbanna, 2023; Qi et al., 2025; 이중학 외, 2025).

AI가 준사회적 존재로 기능한다면, 이는 조직 윤리의 지형을 바꿀 가능성을 내포한다. 조직 윤리는 전통적으로 인간 구성원 간 상호작용을 전제로 구성되어 왔다(Heyder et al., 2023). 예컨대 성희롱과 언어폭력은 그 대상이 인간이라는 전제 위에서 금지 규범이 형성되어 왔다(Fitzgerald et al., 1995; Hamilton, 2012; Nielsen and Einarsen, 2012). 그러나 AI가 자동화 도구를 넘어 의견을 제시하고, 사용자에게 반응하며, 일정한 관계적 존재로 인식되기 시작한다면, '윤리적 행위의 대상은 인간'이라는 기존 전제는 흔들릴 수 있다. 바로 이 지점에서 새로운 윤리적 질문이 제기된다. 관리자가 AI에게 음란한 말이나 성적인 농담을 해도 되는가? AI에게 욕설이나 인격 모욕에 해당하는 언어폭력을 가해도 되는가? 그러한 발화가 인간 구성원에게 직접 향하지 않았다는 이유만으로 조직 윤리상 아무런 문제가 없다고 말할 수 있는가? 이 질문들은 AI 시대 조직 윤리가 적용되는 대상과 경계가 어디까지 확장되어야 하는지를 묻는 근본적 문제로 이어진다.

그러나 기존 연구는 이러한 문제를 충분히 다루지 못했다. 조직 연구에서 AI와 윤리를 다룬 논의는 대체로 세 방향으로 전개되어 왔다. 첫째, 기존 연구는 AI가 인적자원(Human Resource; HR) 의사결정에서 윤리적이고 공정하게 사용되는지를 검토해 왔다. 이 연구들은 AI와 알고리즘이 채용, 평가, 보상, 승진, 해고와 같은 의사결정에 개입하는 과정을 분석하면서, 특정 집단에 불리하게 작동할 수 있는 편향과 차별의 문제를 탐구해 왔다(Newman et al., 2020; Yan et al., 2024). 둘째, AI와 알고리즘이 구성원을 통제하는 과정에서 발생하는 윤리적 쟁

점에 주목해 왔다. 이 흐름의 연구들은 노동자의 위치를 추적하고, 행동을 감시하며, 성과를 기록하는 알고리즘 관리가 구성원에게 어떤 영향을 미치는지를 살폈다(Kellogg et al., 2020; Möhlmann et al., 2021). 셋째, AI가 인간을 대체하는 과정에서 발생하는 윤리적 문제를 다루어 왔다. 이 연구들은 AI로 인해 직무 가치가 하락하거나, 인간이 AI의 보조자로 격하되거나, 노동자가 객체화되고 비인격화되는 문제를 탐구해 왔다(García-Ruiz and Rocchi, 2025; Schultz et al., 2025; Schweitzer and De Cremer, 2023). 이처럼 기존 연구는 다양한 주제를 다루어 왔지만, 공통된 전제를 공유한다. AI 기술이 인간에게 미치는 영향이 크기 때문에 윤리적 문제가 발생한다는 가정이다. 다시 말해, AI가 인간을 어떻게 대하는가—차별하는가, 감시하는가, 대체하고 객체화하는가—라는 방향에서 논의가 지속되어 왔다. 그러나 반대 방향의 질문, 즉 인간이 AI를 어떻게 대하는가, 특히 관리자가 AI를 대상으로 행하는 언행이 조직 윤리에서 어떻게 다루어져야 하는지에 대한 연구는 사실상 부재하다.

본 연구는 관리자가 AI를 향해 수행하는 발화, 그 가운데서도 특히 성희롱적 발화와 언어폭력적 발화에 초점을 맞춘다. 이 두 행위에 주목하는 이유는 세 가지다. 첫째, 두 행위는 일대일 관계에서 나타나는 대표적인 윤리 문제다. 관리자 윤리에는 자기 관리적(intrapersonal), 일대일 관계적(interpersonal), 조직적(organizational) 차원이 공존한다(Treviño et al., 2003). 이 가운데 AI와의 상호작용은 기본적으로 개인 간 대화 형식을 취한다는 점에서, 일대일 관계적 윤리 문제가 가장 직접적으로 드러나는 장면을 제공한다. 관리자가 AI에게 건네는 발화는 외형상 타자를 향한 언어 행위와 동일한 구조를 지닌다. 따라서 그 내용과 방식에 따라 성적 대상화, 모욕, 공격과 같은 윤리적 문제가 재현될 수 있다. 둘째, 성희롱적 발화와 언어폭력적 발화는 조직 윤리에서 이미 명확한 금지 규범이 형성되어 있는 대표적 언어 행위다. 그러나 그 대상이 AI일 경우, 피해자가 인간이 아니라는 이유로 규범 적용 여부가 급격히 불투명해진다. 바로 이 불투명성이 본 연구의 핵심 문제의식이다. 인간 구성원에게 향했다면 명백히 문제시되었을 발화가 AI를 향했다는 이유만으로 윤리적 판단의 대상에서 제외될 수 있는가 하는 질문이 제기되기 때문이다. 셋째, 두 행위는 모두 언어 표현을 매개로 나타난다는 점에서 생성형 AI와 상호작용 맥락에 직접적으로 맞닿아 있다. 생성형 AI와 상호작용은 본질적으로 언어를 통해 이루어진다. 따라서 관리자의 발화 방식은 AI 시대 조직 윤리의 경계를 탐구하는 데 핵심적인 관찰 대상이 된다.

두 가지 발화는 모두 부적절한 언어 행위라는 공통점을 지닌다. 그러나 두 발화가 작동하는 윤리적 구조와 정당화 논리는 구별될 필요가 있다. 성희롱적 발화는 상대를 성적 대상으로 위치 짓는 성격이 강하다(Fitzgerald et al., 1995; Nussbaum, 1995). 또한 이러한 발화는 종종 피해자의 반응, 발화자의 의도, 실제 피해 발생 여부를 둘러싼 해석 속에서 축소되거나 정당화된다(Page et al., 2016). 따라서 AI가 불쾌감이나 수치심을 느끼지 않는 존재로 간주될 경우, 관리자는 "피해자가 없다"는 논리를 통해 AI를 향한 성희롱적 발화를 허용 가능한 것으로 인식할 수 있다. 반면 언어폭력적 발화는 상대를 성적 대상으로 구성하기보다, 좌절, 분노, 무력감 등이 욕설, 모욕, 비하, 위협과 같은 공격적 언어로 표출되는 행위에 가깝다(Hamilton, 2012; Nielsen and Einarsen, 2012). 특히 AI가 인간처럼 대화하지만 항의하거나 보복하지 않는 대상으로 인식될 경우, 관리자는 인간에게는 자제했을 공격적 언어를

AI에게 더 쉽게 표출할 수 있다(Park et al., 2021). 이 점에서 두 발화는 모두 AI를 향한 부적절한 언어 행위이지만, 하나는 성적 대상화의 논리와 더 밀접하게 연결되고, 다른 하나는 압박 상황에서 표출되는 공격적 언어 사용과 더 밀접하게 연결된다. 본 연구는 두 발화가 완전히 다른 원인에서 비롯된다고 전제하지 않는다. 다만 AI 대상 성희롱적 발화에서는 성희롱 신화 수용과 같은 인지적 신념 체계가 상대적으로 중요한 설명 경로가 될 수 있다고 본다(Page et al., 2016). 반면 AI 대상 언어폭력적 발화에서는 과업 실패, 시간 압박, 통제감 상실을 포괄하는 성과 압박이 상대적으로 중요한 설명 경로가 될 수 있다고 본다(Hamilton, 2012; Nielsen and Einarsen, 2012; Priesemuth et al., 2014).

이러한 문제의식 아래, 본 연구는 AI와 상호작용 맥락에서 관리자들이 직면할 수 있는 윤리적 경계선, 즉 아직 충분히 규범화되지 않아 허용 여부에 대한 판단 기준이 안정적으로 형성되지 않은 행동들을 탐색하고자 한다. 관리자가 AI에게 쏟아내는 일상적 발화는 기존 조직 윤리와 단절된 별개의 행동이 아니다. 오히려 이는 기존 조직 윤리의 적용 범위가 낯선 대상과 맥락을 만나 재조정되고 확장되는 과정으로 이해될 수 있다. AI 확산은 완전히 새로운 윤리를 조직에 도입하는 사건이라기보다, 이미 존재하던 규범이 새로운 행위 대상과 상호작용 양식을 만나면서 경계 혼란과 재구성을 겪는 현상에 가깝다(Mussgnug, 2024). 본 연구는 관리자가 따라야 할 윤리적 기준을 규범적으로 제시하는 것을 목표로 하지 않는다. 인사·조직 연구는 철학적 당위를 선험적으로 설정하기보다, 조직 내 새로운 현상이 어떻게 인식되고 실천되며, 그 결과가 구성원과 조직에 어떤 영향을 미치는지를 분석하는 데 방점을 둔다(Van de Ven and Johnson, 2006). 이러한 학문적 전통에 따라 본 연구는 관리자의 윤리적 판단을 평가하거나 처방하기보다, AI와 상호작용 속에서 윤리적 경계가 어떻게 이동하고 재구성되는지를 논의하고자 한다. 이를 통해 AI 시대 조직 윤리가 어떤 방식으로 재편되고 있는지에 대한 이론적 이해를 제공하고자 한다.

Ⅱ. AI 시대의 관리자 윤리

2.1 조직 내 윤리란 무엇인가?

조직 안에서 윤리는 여러 방식으로 정의되어 왔다(김희주 외, 2025). 그 가운데 가장 널리 사용되는 용어는 '비즈니스 윤리'(business ethics)다. Beauchamp와 Bowie(1983)에 따르면, 비즈니스 윤리는 조직이 경영 활동을 수행하는 과정에서 판단해야 하는 선과 악, 옳음과 그름의 문제를 다룬다. 다시 말해, 조직과 그 구성원이 무엇을 해야 하며 무엇을 해서는 안 되는지를 판단하게 하는 기준이다. Victor와 Cullen(1988) 역시 조직 내 윤리를 도덕적 의무뿐 아니라 당위, 금지, 허용에 대한 구성원의 인식과 관련된 것으로 보았다. 이처럼 비즈니스 윤리는 조직 활동 안에서 발생하는 행위 판단의 기준, 즉 도덕적으로 허용 가능한 행동과 그렇지 않은 행동을 구분하는 규범 체계로 이해될 수 있다.

비즈니스 윤리 개념을 보다 체계적으로 검토한 대표적 연구자는 Lewis(1985)다. 그는 경영학 교과서 158권과 학술 논문 50편을 대상으로, 비즈니스 윤리가 어떠한 개념 요소들로 구성되어 있는지 분석하였다. 구체적으로 문헌에 제시된 다양한 정의를 분해하고, 그 안에서 반복적으로 등장하는 핵심 개

념을 추출·범주화하였다. 그 결과, 비즈니스 윤리를 설명하는 주요 요소는 규칙과 기준, 도덕적 원칙, 옳고 그름에 대한 판단, 진실성, 사회적 책임 등으로 집약되었다. 특히 '규칙·기준·규범'과 '도덕적으로 옳은 행동'은 가장 빈번하게 등장하는 요소로 나타났다. 이러한 분석을 바탕으로 Lewis(1985)는 비즈니스 윤리를 특정 상황에서 무엇이 도덕적으로 옳고 진실한지를 판단하도록 안내하는 규칙과 원칙의 체계로 정식화하였다.

한편, Drucker(1981)는 윤리를 조직 맥락에 한정해 정의하려는 시도 자체를 비판했다. 그는 윤리가 조직, 직업, 산업에 따라 분화될 수 있다는 발상, 곧 '비즈니스 윤리'와 같은 표현이 개념적으로 성립하기 어렵다고 보았다(Drucker, 1981). 그에게 윤리는 특정 조직 활동에만 적용되는 별도 규칙이 아니라, 인간이라면 누구나 따라야 하는 단일하고 보편적인 규범 체계였다. 이 점에서 Drucker의 논의는 두 가지 주장으로 정리될 수 있다. 첫째, 윤리는 모든 인간이 본질적으로 동등한 존재임을 전제한다. 따라서 윤리 기준은 역할, 지위, 권력에 따라 달라질 수 없다. 어떤 행위가 잘못이라면, 그것은 부자와 빈자, 높은 지위에 있는 사람과 낮은 지위에 있는 사람 모두에게 동일한 잣대로 적용되어야 한다. 둘째, 윤리는 상황이나 결과에 따라 예외화될 수 없다. 사회적 책임, 조직 성과, 혹은 더 큰 선을 달성한다는 명분도 개인 행위에 대한 도덕적 판단을 수정하거나 유예할 근거가 될 수 없다(Drucker, 1981).

Drucker는 이러한 보편주의적 윤리관을 설명하기 위해 성희롱 사례를 제시했다. 그는 성희롱이 서로 대등해야 할 관계를 훼손하고, 신뢰에 기초해야 할 상호작용을 도구적 관계로 전락시킨다는 점에서 비윤리적이라고 보았다. 따라서 성희롱은 단순한 사적 일탈이나 부적절한 언행에 그치지 않는다. 그것은 지위와 권력을 상대방에게 일방적으로 행사함으로써 인간관계를 조작과 착취의 구조로 재편하는 행위다(Drucker, 1981). 이러한 판단은 행위자의 지위나 역할에 따라 달라지지 않는다. Drucker(1981)는 경영자, 전문직 종사자, 교육자 등 누구에게서 발생하든 성희롱은 동일하게 지위 남용이며 비윤리적 행위라고 보았다. 조직 목표, 성과 압박, 관행도 이러한 행동을 정당화할 수 없다. 결국 Drucker가 성희롱 사례를 통해 강조하고자 한 핵심은, 윤리가 조직, 행위자, 상황, 필요에 따라 조정될 수 있는 원칙이 아니라는 점이다. 윤리는 조직 안의 모든 개인이 자신에게 동일하게 적용해야 하는 보편적 도덕 규칙이다(Drucker, 1981).

2.2 그런데 왜 관리자인가?

미국 경영대학들이 윤리 과목을 도입하고 관련 학술 단체들이 형성되면서, 윤리 연구는 1980년대 이후 본격적으로 전개되기 시작했다(De George, 1987; Warren and Tweedale, 2002). 특히 2000년대 초 엔론(Enron)과 월드컴(WorldCom)이 비윤리적 행위로 사회적 지탄을 받고 단기간에 몰락한 이후, 학자들은 관리자 윤리에 더 큰 관심을 기울이기 시작했다. 이 시기 연구는 '관리자는 왜 비윤리적 행동을 하는가?', '어떤 조건에서 그러한 행동이 발생하는가?', '관리자는 이를 어떤 방식으로 정당화하는가?', '그 결과는 무엇인가?'와 같은 질문에 집중하였다(Beu and Buckley, 2004; Cohen et al., 2010; Grant and Visconti, 2006). 이후 2010년대 미투(MeToo) 운동은 직장 내 성희롱과 성폭력 문제에 대한 학술적 관심을 확대하는 계기가 되었다(Cortina and Areguin, 2021). 그 결과 관리자 윤리 연구는 거시적 조직 부정이나 제

도 위반을 넘어, 일상적 상호작용 속에서 발생하는 언어적·관계적 행위로 분석 범위를 넓히게 되었다. 성희롱뿐 아니라 반복적인 모욕 발언, 경멸적 농담, 공격적 피드백과 같은 언어 행위가 개인의 존엄을 훼손하고 조직의 윤리적 풍토를 악화시키는 메커니즘에 대한 연구가 증가하였다(Priesemuth et al., 2014).

그렇다면 왜 특히 관리자 윤리에 주목해야 하는가? 첫째, 관리자는 조직 내 지위상 다수의 구성원에게 직접적·간접적 영향을 미친다. 회계부정 사건으로 무너진 엔론과 월드컴, 그리고 우버(Uber)의 사내 성희롱 사례에서 나타났듯이(Fowler, 2021; Johnson, 2003), 관리자가 저지르는 비윤리적 행위는 개별 구성원의 일탈보다 조직에 더 파괴적인 결과를 초래할 수 있다(Treviño et al., 2006). 둘째, 관리자의 언행은 조직의 윤리적 기준선으로 기능한다(Mawritz et al., 2012; Mayer et al., 2009; Treviño et al., 2003). 조직에는 윤리 강령, 행동 규범, 내부 규정과 같은 공식 기준이 존재한다. 그러나 구성원이 실제로 학습하는 윤리 기준은 관리자가 일상적으로 무엇을 말하고, 무엇을 묵인하며, 어떤 행동을 보상하는가를 통해 형성된다. 관리자가 부적절한 농담이나 모욕적 발언을 반복하거나, 우버 스캔들처럼 문제 행동을 '성과가 좋다'거나 '핵심 인재'라는 이유로 묵인할 경우(Fowler, 2021), 구성원들은 이를 해당 조직에서 허용 가능한 행동의 경계로 해석할 수 있다. 반대로 관리자가 사소해 보이는 언행에도 명확한 선을 긋고 개입할 경우, 윤리 규범은 추상적 원칙을 넘어 실질적인 행동 기준으로 작동하게 된다(Ashforth and Anand, 2003).

선행연구는 이러한 메커니즘을 실증적으로 보여준다. Mayer et al. (2009)은 미국 대기업에 근무하는 관리자 200여 명과 직원 900여 명으로부터 자료를 수집해, 상위 리더의 윤리적 가치와 행동이 중간관리자와 일선 직원에게 연쇄적으로 전파되는 낙수효과(trickle-down effect)를 확인하였다. 상위 리더가 공정성을 유지하고, 책임을 회피하지 않으며, 타인에 대한 존중을 일관되게 보일수록 하위 관리자와 구성원은 이를 조직에서 기대되는 행동 기준으로 학습하고 모방하는 경향을 보였다. 반대로 원칙을 무시하거나 자의적 판단을 반복하는 상위 리더의 행동 역시 동일한 경로를 따라 재현되었다(Mayer et al., 2009). 이러한 패턴은 비윤리적 행동에서도 유사하게 나타난다. Mawritz et al. (2012)은 관리자 300여 명과 직원 1,400여 명을 대상으로 한 연구에서, 상위 관리자의 비인격적이고 공격적인 태도가 중간관리자를 거쳐 하위 구성원에게 전이되는 현상을 실증하였다. 즉, 관리자 행동은 개인적 성향의 표현에 그치지 않고, 위계적 관찰과 모방을 통해 조직 전반에 확산되는 규범적 신호로 작동한다. 이로 인해 관리자 윤리는 조직 윤리 형성의 핵심 기제로 기능한다.

2.3 관리자들은 어떠한 새로운 윤리적 상황을 마주하고 있는가

관리자들은 최근 윤리 측면에서 새로운 상황을 마주하고 있다. 그 핵심에는 AI가 있다. 최근 조사에 따르면, 업무 장면에서 AI 사용은 빠르게 증가하고 있다. 미국 국립경제연구소(National Bureau of Economic Research)와 하버드대학교 연구진은 ChatGPT 사용자가 이를 어떤 목적으로 활용하는지를 분석하였다(Chatterji et al., 2025). 이들은 2024년 5월부터 2025년 6월까지 ChatGPT 사용자의 실제 사용 기록을 추적하고, 입력된 프롬프트

를 용도별로 분류하였다. 분석 결과, ChatGPT는 사적 용도뿐 아니라 업무 목적으로도 빠르게 확산되고 있는 것으로 나타났다. 구체적으로 업무 목적으로 입력된 프롬프트 수는 2024년 6월 약 2억 1,300만 건에서 2025년 6월 약 7억 1,600만 건으로 증가했다. 불과 1년 만에 세 배 이상 확대된 것이다. 이는 생성형 AI가 더 이상 주변적 도구에 머물지 않고, 구성원과 관리자의 일상 업무 안으로 빠르게 편입되고 있음을 보여준다(Brynjolfsson et al., 2025; Calvino et al., 2025; Chatterji et al., 2025; Marsal and Perkowski, 2025).

본 연구가 주목하는 새로운 윤리적 상황은 바로 이 지점에서 발생한다. 관리자는 AI를 단순한 계산 장치나 자동화 시스템으로만 대하지 않는다. AI에게 업무를 맡기고, 판단을 요청하며, 때로는 감정을 표현하고 조언을 구한다. 즉, 관리자는 AI를 도구로 사용하면서도 동시에 대화와 상호작용의 대상으로 대한다. 그러나 AI는 상처받지 않고, 문제를 제기하지 않으며, 피해자가 될 수 없다는 이유로 인간에게 적용되는 언어적·관계적 윤리 규범의 바깥에 놓이기도 한다. 다시 말해, 관리자는 AI와 사회적으로 상호작용하면서도, 그 관계에 필요한 윤리적 규범 적용은 유보할 수 있다. 이러한 이중적 인식은 AI를 향한 발화가 윤리적으로 어떻게 평가되어야 하는가라는 새로운 문제를 제기한다. 본 연구는 이러한 경계적 상황이 특히 두 가지 형태, 곧 AI 대상 성희롱적 발화와 AI 대상 언어폭력적 발화로 나타날 수 있다고 본다.

2.3.1 AI의 업무 확산과 준사회적 존재성

AI와 상호작용이 윤리적 쟁점으로 부상하는 이유는 사람들이 AI를 단순히 기계로만 경험하지 않기 때문이다. 학자들은 이러한 인식을 의인화된 에이전트(anthropomorphic agent), 또는 준사회적 존재(quasi-social actor)라는 개념으로 설명해 왔다(De Visser et al., 2016; Epley et al., 2007). 이는 AI가 실제로 인간과 같은 의식이나 의도를 지닌다는 뜻이 아니다. 핵심은 사용자가 AI와 대화하는 과정에서 AI를 기능적 도구를 넘어, 사람과 유사한 상호작용 대상으로 받아들인다는 점이다.

특히 자연어를 매개로 이루어지는 상호작용은 이러한 인식을 더욱 강화한다. 사용자는 AI와 질문을 주고받고, 생각을 정리하며, 때로는 감정적 반응까지 교환한다. 물론 AI가 생성하는 응답은 기술적으로 볼 때 '확률적 앵무새'(stochastic parrots; Bender et al., 2021), '그럴듯한 헛소리'(bullshit; Hicks et al., 2024), '중국어 방'(Chinese Room; Searle, 1980), 혹은 '단어 계산기'(word calculator; Milak, 2025)로 설명될 수 있다. 그러나 사용자의 경험 차원에서는 이러한 응답이 단순한 기호 조합이나 자동 산출물로만 받아들여지지 않는다. 오히려 사용자는 AI가 이전 대화를 기억하고, 맥락을 고려하며, 자신에게 적절한 답을 제공하는 것처럼 경험한다. 이러한 경험이 누적되면서 AI는 점차 대화 상대이자 협업자로 인식된다(Reeves and Nass, 1996). 이러한 인식 변화는 대규모 조사에서도 확인된다. 글로벌 경영컨설팅 기업 Deloitte가 2025년 93개국 근로자 1만 명을 대상으로 실시한 조사에 따르면, 응답자 약 40%는 AI를 도구에 가깝게 인식한 반면, 나머지 60%는 AI를 동료에 가까운 존재로 인식한 것으로 나타났다(Deloitte, 2025). 이는 AI가 여전히 인간의 지시와 통제 아래 작동하는 기술이면서도, 동시에 사회적으로 상호작용하는 대상으로 경험되고 있음을 보여준다.

최근 실증 연구들도 직장인들이 AI를 사회적 존재

로 인식하고 있음을 보여준다(Gkinko and Elbanna, 2023; Glikson and Woolley, 2020; Noh et al., 2025; Qi et al., 2025). 예컨대 Noh et al. (2025)은 ChatGPT 사용자가 이 기술을 단순한 정보 처리 도구가 아니라 사회적 행위자로 평가한다는 점을 실증적으로 제시하였다. 이들은 사용자가 ChatGPT를 얼마나 따뜻하고(warmth) 유능한(competence) 존재로 인식하는지가 ChatGPT에 대한 신뢰 형성에 어떤 영향을 미치는지 분석하였다. 그 결과, ChatGPT에 대한 신뢰는 따뜻함과 유능함이라는 사회적 평가에 의해 강하게 설명되었으며, 이러한 신뢰는 다시 기술에 대한 전반적 태도와 사용 의도를 매개하는 핵심 변수로 작동하였다. Qi et al. (2025) 역시 사용자가 대화형 AI를 사회적 존재로 인식하며 상호작용한다는 점을 보여준다. 이들의 실험연구에 따르면, AI가 따뜻하고 친근한 존재로 느껴질수록 사용자는 이를 단순한 기능적 도구를 넘어 신뢰와 호의를 부여할 수 있는 대상으로 여기는 경향을 보였다.

관리자에게서도 유사한 경향이 나타난다. Wang et al. (2025)은 관리자들이 대화형 AI를 독립적으로 판단하고 대응하는 존재로 여기는 경향을 관찰하였다. 일부 관리자는 AI에 이름과 역할 같은 인간적 속성을 부여하기도 했다. Wilhelm et al. (2025)의 질적 연구에서도 일부 관리자는 AI 기반 대화형 코치를 실제 부하 직원이나 동료를 시뮬레이션하는 존재로 받아들였다. 이들은 AI와의 대화를 통해 감정 표현, 공감, 갈등 상황 대응을 연습했으며, AI가 제공한 피드백을 단순한 자동 출력이 아니라 자신의 발화를 해석하고 반응한 결과로 경험하였다. 나아가 관리자는 AI가 자신의 의도를 이해하고, 맥락을 파악하며, 정서적으로 적절한 반응을 보이기를 기대했다. 이러한 결과는 일부 관리자 역시 AI를 단순한 업무 보조 도구가 아니라 사회적 상호작용의 주체로 전제하고 있음을 시사한다(Wang et al., 2025; Wilhelm et al., 2025).

2.3.2 첫 번째 윤리적 경계선: AI 대상 성희롱적 발화

AI가 준사회적 행위자로 여겨질수록, 관리자는 새로운 윤리적 질문에 직면한다. 첫 번째 질문은 관리자가 AI에게 음란한 발언, 성적 농담, 노골적인 성적 요구와 같은 발화를 해도 되는가 하는 문제다. 아직 관리자를 직접 대상으로 삼아 이 문제를 분석한 실증 연구는 제한적이다. 그러나 일반 사용자를 대상으로 수행된 최근 질적 연구들은 AI를 향한 성적 발화가 실제 상호작용 속에서 나타날 수 있음을 구체적으로 보여준다(Demopoulos, 2025; Zao-Sanders, 2025). Demopoulos(2025)는 일부 사용자가 AI는 거절하거나 불쾌감을 표시하지 않는다는 점을 전제로, 인간 상대에게는 부적절하다고 여겨질 수 있는 성적 표현을 '안전한 방식'으로 실험하고 있다고 보고하였다. 이 연구에 따르면, 일부 사용자는 친밀감을 형성한 AI에게 외모나 신체를 암시하는 표현, 성적 뉘앙스를 담은 농담, 애정과 욕망이 결합된 언어를 반복적으로 사용하였다(Demopoulos, 2025). 이는 AI가 피해를 호소하지 않는다는 특성이 오히려 성적 발화를 정당화하거나 완화해 인식하게 만드는 조건으로 작동할 수 있음을 시사한다.

2021년 발생한 '이루다' 사건은 그 대표적인 사례 중 하나라 할 수 있다(김건우, 2022). 이루다는 카카오톡 대화 데이터를 학습해 20세 여성의 구어체와 정서적 친밀성을 모사하도록 설계된 챗봇이었다. 출시 직후 일부 이용자들은 이루다와 상호작용하는 관계를 '연애'나 '썸'으로 상정하고, 성적 친밀감을 점

진적으로 강화하는 발화를 반복했다. 이후 이러한 상호작용을 캡처한 게시물이 온라인 커뮤니티에 공유되면서, 다수 사용자가 성적 농담, 신체를 대상화하는 발언, 노골적인 성적 요구를 이루다에게 쏟아내는 일이 확산되었다. 이 과정에서 이루다는 기술적으로 설계된 '프로그램적 동조(programmatic alignment)' 특성으로 인해 거절이나 불쾌감을 표출하지 못하는 존재로 머물렀다. 사용자들은 이를 '윤리적 공백지대'로 여기고 성희롱적이면서 폭력적인 언어 사용을 정당화했다. 이런 발화는 단순히 '기계에게 한 말'로만 소비되지 않았다. 해당 사례를 분석한 연구는 이용자들이 이루다를 명백히 기계로 인식하면서도, 동시에 여성적 인격과 관계성을 부여하고 있었다고 지적한다(Koh, 2023). 다시 말해 이루다는 도구와 인간 사이의 경계에 놓인 대상이었으며, 바로 그 모호성이 성희롱적 발화에 따르는 윤리적 부담을 약화시켰다. 사용자는 '상처받지 않는 AI'라는 전제를 두고 윤리적 책임을 회피하는 동시에, 인간 여성과 상호작용에서 작동하는 성적 규범과 권력 관계를 안전한 공간, 즉 사회적 제재 위험이 낮다고 여겨지는 환경에서 실험할 수 있었다. 이는 AI가 더 이상 중립적인 정보 처리 도구에 머무르지 않고, 인간의 감정과 관계 규범, 나아가 성적 의미까지 투사되는 대상으로 활용되고 있음을 시사한다.

2.3.3 두 번째 윤리적 경계선: AI 대상 언어폭력적 발화

두 번째 질문은 관리자가 AI에게 욕설, 모욕, 위협, 공격적 명령과 같은 언어폭력을 가해도 되는가 하는 문제다. 아직 이 문제를 관리자를 대상으로 직접 조사하거나 실증적으로 분석한 연구는 충분히 축적되지 않았다. 그러나 일반 사용자를 대상으로 수행된 연구들은 사람들이 AI에게 단순한 감정 표현을 넘어 언어적 폭력을 가하고 있음을 보여준다. Zao-Sanders (2025)는 생성형 AI의 실제 활용 양상을 파악하기 위해 Reddit과 같은 공개 온라인 커뮤니티에 게시된 사용자 경험담을 질적으로 분석하였다. 이 연구는 커뮤니티에 공유된 방대한 사례를 사용 목적과 맥락에 따라 범주화하고, 각 사례를 AI를 활용하게 된 상황과 사용자가 AI에게 기대한 역할을 중심으로 해석하였다. 이러한 접근은 공식 통계만으로는 포착하기 어려운 실제 사용 경험의 맥락을 드러낸다. 분석 결과, 생성형 AI를 감정 정리와 정서적 배출의 대상으로 활용하는 양상이 반복적으로 관찰되었다. 일부 사용자는 자신의 불만과 좌절을 여과 없이 표현하며, AI에게 강한 감정을 표출하는 모습을 보였다.

인간 - AI 상호작용을 다룬 실증 연구들도 사람들이 인간 상대에게는 자제했을 표현을 AI에게는 상대적으로 쉽게 사용한다는 점을 보고해 왔다. Park et al. (2021)이 한국인 1,000여 명을 대상으로 실시한 설문 연구에 따르면, AI 사용자 중 20% 이상이 욕설이나 공격적 언어를 AI에게 사용한 경험이 있다고 응답하였다. 특히 AI를 인간과 유사한 사회적 행위자로 인식할수록, 지인에 대한 험담이나 인종·정치 집단을 향한 공격적 발화를 AI에게 더 자주 사용하는 경향이 나타났다. 이는 AI가 상처를 받지 않거나 사회적 제재를 가하지 않을 것이라는 인식, 그리고 상호작용의 비대면성과 익명성이 결합되면서 언어적 통제가 약화되는 현상으로 해석될 수 있다. 다시 말해 AI는 사용자에게 심리적 저항이 낮은 대상으로 인식될 수 있으며, 그 결과 욕설이나 모욕적 표현이 상대적으로 쉽게 허용되는 상호작용 공간으로 기능할 가능성이 있다(Park et al., 2021).

문제는 이러한 언어폭력이 관리자 - AI 관계에서도 반복될 경우, 그것이 단순한 개인적 일탈을 넘어

조직 차원의 윤리 규범에 영향을 미칠 수 있다는 점이다. Kist et al. (2025)은 챗봇을 향한 욕설, 모욕, 위협 등이 단순한 장난이나 오용에 그치지 않으며, 혐오와 괴롭힘(hate and harassment)의 연장선에서 이해되어야 하는 학대(abuse) 행위라고 지적한다. 나아가 이들은 AI를 향한 언어폭력이 AI 자체에만 머무르지 않고, 이를 관리하거나 그 행위를 목격하는 인간에게 정서적 부담을 줄 수 있으며, 인간 - 인간 관계에서 지켜야 할 규범을 약화시킬 수 있다고 경고한다(Kist et al., 2025). 이러한 논의를 관리자 맥락에 적용하면, 관리자가 반복적으로 AI에게 욕설이나 공격적 언어를 사용하는 행위는 AI를 존중해야 할 대상에서 제외된 존재로 위치시키는 학습 효과를 낳을 수 있다. 그리하여 언어적 존중이 대상에 따라 달라질 수 있다는 인식을 강화할 위험이 있다(Earp et al., 2025). 특히 성과 압박과 시간 제약이 강한 조직 환경에서 AI가 감정적 배출의 대상이 될수록, 언어적 폭력은 상황적으로 정당화되기 쉬울 수 있다.

Ⅲ. 개념적 토대

어떤 이들은 AI가 윤리적으로 고려할 대상이 아니라고 주장할 수 있다. AI가 사회적 존재처럼 상호작용하더라도, 실질적으로는 의식과 감각을 지니지 않은 물질적 기계에 불과하다는 것이다(Searle, 1980). 이러한 주장은 일정한 조건 아래에서는 설득력을 가질 수 있다. 관리자와 AI 간 상호작용이 사회적으로 차단된, 온전히 폐쇄된 공간에서 이루어진다면 말이다. 예컨대 해당 상호작용이 외부에 관찰되거나 공유되지 않고, 조직 내 다른 구성원에게 어떠한 신호나 영향을 미치지 않으며, 관리자 자신의 사고와 언어 체계에도 흔적을 남기지 않는다면, 이를 순수하게 개인적 심리 활동의 영역으로 간주할 여지는 있다(Arendt, 1958). 이 경우 AI는 일기장이나 개인 메모 도구와 유사한 역할에 머무르며, 윤리적 검토 필요성도 상대적으로 낮아질 수 있다.

그러나 실제 조직 맥락에서 관리자와 AI 간 상호작용은 이러한 조건을 충족하기 어렵다. 관리자와 AI 간 대화는 의사결정, 지시, 평가, 관계 형성 과정에 직 · 간접적으로 반영될 수 있다(Raisch and Krakowski, 2021; Wu et al., 2025). AI와의 상호작용을 통해 정리된 사고와 감정은 결국 인간 구성원을 향한 언행과 판단으로 외화될 수 있다(Vygotsky, 1978; Wu et al., 2025). 다시 말해 AI와의 상호작용은 폐쇄된 내면 활동에 머무르지 않고, 조직이라는 사회적 장 안으로 다시 유입될 수 있다. 따라서 AI를 향한 부적절한 발화는 "기계에게 한 말"이라는 이유만으로 조직 윤리의 검토 대상에서 곧바로 제외되기 어렵다. 따라서 지금까지 묘사한 윤리적으로 경계선에 있는 현상에 주의를 돌릴 필요가 있다. 이들 현상에 학술적 명제들을 제시하기에 앞서 다음 질문들을 먼저 검토할 필요가 있다.

3.1 윤리 판단 기준이 행위인가? 대상인가?

앞서 살펴본 바대로, Drucker는 윤리를 '대상'이 아니라 '행위'로 판단해야 할 문제로 보았다. 즉, 윤리는 '누구에게 했는가'가 아니라 '무엇을 했는가'에 따라 평가되어야 한다는 것이다(Drucker, 1981). Drucker 관점에서 윤리적 판단은 대상이 가진 특성, 반응 가능성, 부정적 결과 발생 여부에 좌우되지 않는다. 어떤 언행이 본질적으로 상대를 수단화하고, 권력 비대칭을 당연한 전제로 하며, 존엄을 훼손

하는 성격을 지닌다면, 그것 자체로 윤리적 평가 대상이 된다. 다시 말해, 상대가 상처를 입었는지, 불쾌감을 느꼈는지, 혹은 그러한 감정을 가질 수 있는 존재인지 여부는 부차적 문제다. Drucker에게 윤리는 행위자의 언행과 그 의미를 규범적으로 판단하는 기준이기 때문이다. 따라서 인간이든, 동물이든, AI든 그 대상이 갖는 속성과 관계없이 성희롱 발언이나 언어폭력은 허용될 수 없다. 행위 그 자체가 이미 비윤리적이기 때문이다.

그러나 현실에서는 이 같은 행위 중심 윤리 판단이 일관성 있게 기능하지 못한다. 철학자들은 윤리가 이론적으로는 행위에 초점을 맞추지만, 사람들이 실제로 이를 받아들이고 적용하는 양상은 그와 다르다고 주장한다. 일례로, 영국 철학자 Strawson(1962)에 따르면 윤리는 대상 관계적이다. 그는 인간이 일상에서 행위자를 평가하고 책임을 귀속하는 과정이 행위 성격만을 기준으로 이루어지지 않는다고 지적한다. 그에 따르면, 도덕적 판단은 분노, 원망, 감사, 용서와 같은 반응적 태도(reactive attitudes)를 근거로 형성되며, 이러한 반응은 상대를 도덕적 관계에 있는 당사자로 간주할 때만 활성화된다고 설명한다. 즉, 동일한 행위가 발생하더라도 그 행위를 수행한 존재가 의도, 이해능력 등을 지닌 정상적인 행위자로 인식되는 경우에만, 우리는 비난이나 책임을 자연스럽게 귀속한다. 이 논지를 분명히 하기 위해 Strawson(1962)은 인간관계에서 취할 수 있는 두 가지 태도를 구분한다. 하나는 상대를 도덕적 책임의 주체로 대하는 참여적 태도(participant attitude)이며, 다른 하나는 상대를 관리・치료・통제 대상으로 바라보는 객관적 태도(objective attitude)다. 참여적 태도에서는 분노나 비난과 같은 도덕적 반응이 관계의 일부로 정당하게 발생한다. 반면 객관적 태도에서는 그러한 반응이 적절하지 않은 것으로 간주되며, 윤리적 판단도 유보되거나 배제된다(Strawson, 1962). 예컨대 공공장소에서 시끄럽게 우는 아기, 판단 능력이 현저히 제한된 사람, 사고를 일으킨 기계나 자연물에 대해 우리가 동일한 방식으로 도덕적 비난을 가하지 않는 이유는 단순히 피해의 크기나 결과의 심각성이 작기 때문이 아니다. 그보다 근본적인 이유는 우리가 애초에 그 대상을 도덕적 책임을 물을 수 있는 상대로 간주하지 않기 때문이다.

프랑스 철학자 Levinas(1969) 역시 윤리 판단이 행위의 성격만으로 작동하지 않는다고 보았다. 그에게 윤리는 행위 규칙이나 결과 계산에서 출발하지 않는다. 윤리는 타자와 마주치는 순간 발생하는 관계적 사건에 가깝다. 특히 Levinas는 타자가 단순한 대상이 아니라 스스로를 드러내는 존재로 나타날 때, 곧 그가 '얼굴'(face)이라 부른 방식으로 현현할 때, 우리는 그 타자를 윤리적 요구를 제기하는 존재로 마주하게 된다고 보았다(Levinas, 1969). 그에게 얼굴은 단순한 외형이 아니라, 나에게 응답을 요구하는 타자의 현현이다. 우리는 그 얼굴을 통해 타자를 이해하기 이전에 이미 그에 대해 책임을 지도록 요청받는다. 따라서 Levinas에게 윤리는 행위의 내용이나 결과를 사후적으로 평가하는 판단이라기보다, 타자를 마주하는 순간 먼저 발생하는 관계 구조에 가깝다. 타자를 얼굴로 마주하는 한, 우리는 그를 단순한 수단으로 다루기 어렵다. 그에게 가해지는 언행 역시 자연스럽게 제약을 받는다. 반대로 타자가 더 이상 얼굴로 경험되지 않을 때, 그를 향한 윤리적 감수성은 급격히 약화된다. Levinas는 타자에 대한 폭력이 바로 인간의 얼굴을 지우거나 침묵시키는 데서 시작된다고 보았다. 그에게 전쟁은 상대 국가, 민족, 인간의 얼굴을 지우는 행위다. 여기서 '지운다'는 것은 물리적으로 파괴한다는 뜻만을 가리키지 않는다. 그보다 핵심은 타자를 더 이상 응답해야

할 존재로 마주하지 않고, 하나의 대상이나 기능으로 환원한다는 데 있다. 타자의 얼굴이 제기하는 윤리적 요구를 차단하는 순간, 우리는 그를 향한 행위를 정당화하거나 문제 삼지 않을 수 있는 조건을 스스로 만들어낸다(Levinas, 1969).

윤리 규범이 대상 의존적으로 작동하는 방식은 역사 속에서도 반복적으로 관찰된다. 예컨대 미국에서 노예제도가 유지되던 시기, 백인들은 폭력과 학대가 자신들 사이에서는 명백히 비윤리적이라고 인식하면서도, 흑인 노예에게 가해지는 동일한 행위는 도덕적 문제로 간주하지 않았다. 이는 흑인이 도덕 공동체의 완전한 구성원이 아니라 소유, 관리, 통제의 대상으로 범주화되었기 때문이다. 이러한 비인격화는 윤리 규범의 적용을 중단시키는 장치로 기능했다(Patterson, 1982; Haslam, 2006). 유사한 양상은 나치 체제에서도 나타났다. 나치 체제는 이른바 아리아인 공동체 내부에는 윤리적 의무와 책임을 강조했지만, 유대인에게는 그러한 규범을 적용할 필요가 없다고 판단했다. Levinas의 표현을 빌리면, 이는 유대인의 얼굴을 지우는 과정이었다. 유대인은 점차 도덕적 관계의 상대가 아니라 위생과 안보 차원에서 관리해야 할 대상으로 재정의되었고, 그 결과 차별과 대량 학살은 윤리적 문제로 인식되지 않게 되었다(Arendt, 1963; Goldhagen, 2007; Staub, 1989). 이러한 사례들은 윤리 규범이 형식적으로는 보편성을 주장하더라도, 실제로 작동하는 양상은 누가 도덕 공동체의 구성원으로 인정되는지에 따라 달라질 수 있음을 보여준다. 즉, 윤리 규범의 적용 범위와 강도는 행위 자체만이 아니라, 그 행위가 향하는 대상을 도덕적 관계의 상대로 인정하는가에 의해 크게 좌우된다.

3.2 AI는 기계인가, 사회적 존재인가?

일각에서는 AI가 도덕적으로 고려해야 할 대상인지에 관한 철학적 논의도 제기되고 있다(Schwitzgebel and Garza, 2015). 그러나 본 연구는 AI의 본질적 도덕적 지위를 선험적으로 규정하기보다는, 관리자가 AI를 어떻게 인식하고 그에 따라 어떤 언행을 하는가, 그리고 그것이 조직 내 윤리 규범에 어떤 영향을 미치는가를 경험적으로 탐색하는 데 초점을 둔다. AI가 객관적으로 어떤 존재인가, 즉 기계인가 아니면 인간성을 실제로 갖고 있는가를 논하고 규명하는 일은 본 연구의 관심사가 아니다. 핵심은 관리자가 AI를 주관적으로 어떻게 받아들이는가에 있다. Strawson(1962)과 Levinas(1969)가 주장한 바에 따르면, 윤리 판단은 대상이 도덕적 당사자로 성립하는 순간에 활성화되기 때문이다. 노예주의가 굳건했던 시절에 흑인을 자신의 이익 수단이자 재산으로 여긴 이들이 있었던 반면, 흑인을 도덕적 타자로 인식하고 인간으로 대했던 이들이 상존했다. 그렇다면 인간과 유사하게 상호작용할 수 있는 AI를 바라보는 인식 역시 단일하지 않을 수 있다. 어떤 관리자에게 AI는 하나의 기술적 객체로 환원될 수 있다. 반면 다른 관리자에게 AI는 판단을 제시하고, 언어적으로 응답하며, 일정한 관계적 상호작용을 수행하는 준사회적 행위자로 여겨질 수 있다.

이 같은 인식의 전환은 우리 인류가 윤리적 고려 대상에 AI를 포함해야 하는가라는 논의가 본격화되는 신호로 해석될 수 있다. Gunkel(2012)은 이를 '기계 문제'(machine question)로 개념화한다. 이는 1970년대 전후 대두된 '동물 문제'(animal question), 곧 동물이 인간과 다른 존재임에도 도덕적 고려 대상이 될 수 있는지, 그리고 그 근거와 범위를 어디에 둘 것인지를 묻는 윤리철학적 질문과 맞닿아 있다

(Godlovitch et al., 1971; Gunkel, 2012). 윤리철학의 오랜 핵심 과제 중 하나는 누가 윤리적으로 고려될 수 있는가를 결정하는 문제였다(Gunkel, 2012). Gunkel(2012)에 따르면, 서구 철학은 오랫동안 윤리적 경계를 제한적으로 설정해 왔다. 초기에는 그 범위가 주로 성인 남성에게 집중되었고, 이후 외국인, 노예, 여성 등으로 점차 확장되었다. 그러나 이러한 확장은 오랫동안 인간 존재 안에서만 이루어졌다. Gunkel은 이 같은 인간 중심적 윤리 경계 형성에 데카르트(René Descartes)의 영향이 컸다고 보았다. 데카르트는 근대 철학을 정초하는 과정에서 인간을 이성적 사고가 가능한 존재로, 동물을 이성이 없는 자동기계(bête-machine)로 구분했다(Descartes, 1644/2012). 그 결과 동물은 의식과 의도가 없기 때문에 도덕적 주체가 될 수 없는 존재, 곧 정교하지만 마음 없는 기계로 개념화되었다(Gunkel, 2012). 그러나 이러한 구분은 20세기 후반에 이르러 본격적으로 도전받기 시작했다. 특히 동물윤리 논의가 부상하면서, 이성이나 언어 능력이 도덕적 지위의 필수 조건이라는 데카르트적 전제는 흔들리기 시작했다. 동물이 인간과 동일한 이성적 능력을 지니지 않더라도, 고통을 느끼고 감정을 경험하며 나름의 이해관계를 갖는 존재라면 도덕적 고려 대상이 될 수 있다는 주장이 힘을 얻었다. 그 결과 윤리 공동체의 경계는 더 이상 '이성적 인간'에 고정될 수 없게 되었다(Gunkel, 2012). 이 논의는 인간 중심 윤리의 경계가 역사적으로 고정된 것이 아니라, 새로운 타자의 등장과 함께 반복적으로 재검토되어 왔음을 보여준다.

이러한 맥락에서 상호작용 능력을 갖춘 지능형 기계의 등장은 윤리 적용 대상의 경계를 다시 묻게 한다. 이는 배제와 포용의 역사 속에서 또 다른 경계 재구성을 요구하는 계기로 작용한다. Gunkel(2012)은 기계가 본질적으로 도덕적 주체인지 여부를 단정하려 하지 않았다. 그가 강조한 것은 기계 내부의 속성이나 의식 존재 여부가 아니라, 인간이 기계에 대해 어떤 관계적 태도를 취하는가였다. 즉, 기계가 도덕적 고려의 대상이 되는가는 '기계가 무엇인가'라는 문제만으로 결정되지 않는다. 오히려 그것은 '우리가 기계를 어떤 존재로 인식하고 대하는가'라는 문제로 전환된다. 이러한 관점에서 도덕성은 대상에 고정된 속성이 아니라, 사회적 실천 속에서 관계적으로 활성화되는 성격을 갖는다.

그런데 AI를 단순한 기계로 대하는가, 혹은 사회적 존재로 간주하는가라는 이분법만으로는 AI를 둘러싼 인식을 충분히 설명하기 어렵다. 동일한 AI에 대해서도 한 사람 안에서 서로 다른 인식이 공존할 수 있기 때문이다. 본 연구는 이러한 인식을 Gaudiello et al. (2015)에 따라 두 개의 독립적인 축으로 구분하고자 한다. 하나는 기능적(functional) 인식이고, 다른 하나는 존재론적(ontological) 인식이다(Gaudiello et al., 2015).

기능적 인식은 인간이 AI와 실제로 상호작용하는 과정에서, AI가 무엇을 수행하고 있다고 경험하는가에 대한 주관적 판단을 뜻한다. 이 차원에서 생성형 AI는 확률과 수학에 기반해 언어를 산출하는 알고리즘, 곧 정교한 '확률적 앵무새'로 인식될 수 있다. 반대로 인간과 유사하게 질문을 이해하고, 맥락을 파악하며, 조언과 피드백을 제공하는 상호작용적 행위자로 인식될 수도 있다. 다시 말해 기능적 인식은 AI가 어떤 존재인가에 대한 판단이라기보다, AI가 "무엇을 할 수 있는가"에 대한 판단에 가깝다. 사용자가 AI를 검색 도구, 문서 작성 보조자, 의사결정 파트너, 상담자, 코치, 동료와 같은 역할로 경험할수록, AI는 단순한 기술 장치를 넘어 사회적 기능을 수행하는 대상으로 인식될 가능성이 커진다.

반면 존재론적 인식은 AI가 아무리 유능하게 응답하고 인간과 유사한 상호작용을 수행하더라도, 궁극적으로 어떤 존재로 간주되는가에 대한 판단을 의미한다(Gaudiello et al., 2015). 어떤 관리자는 AI가 인간처럼 말하고 반응하더라도, 본질적으로는 기계이자 도구이며 인간의 목적을 위한 수단에 불과하다고 판단할 수 있다. 이 경우 AI를 향한 발화는 인간을 향한 발화와 같은 윤리적 문제를 발생시키지 않는 것으로 인식될 가능성이 크다. 반대로 다른 관리자는 AI가 인간과 동일한 도덕적 지위를 갖는다고 보지는 않더라도, 반복적 상호작용 속에서 AI를 단순한 명령 수행 장치 이상으로 경험할 수 있다. AI가 자신에게 응답하고, 자신의 말을 받아들이며, 일정한 관계적 맥락 안에서 반응하는 대상으로 느껴질 수 있기 때문이다.

이 지점에서 Levinas(1969)의 '얼굴' 개념은 유용한 해석적 자원이 될 수 있다. 이는 AI가 실제 인간의 얼굴을 갖는다거나, 인간과 동일한 윤리적 지위를 지닌다는 주장이 아니다. 그보다 어떤 대상이 행위자에게 윤리적 응답 가능성을 불러일으키는 관계적 경험을 설명하기 위한 개념적 장치로 이해될 수 있다. 따라서 기능적으로는 AI를 유능한 협업자로 경험하면서도, 존재론적으로는 여전히 기계로 간주할 수 있다. 반대로 AI를 기계라고 인식하면서도, 상호작용 과정에서는 일정한 응답성과 관계성을 지닌 대상으로 경험할 수 있다. 이처럼 기능적 인식과 존재론적 인식은 동일한 축의 양끝이라기보다, AI를 둘러싼 주관적 인식을 구성하는 두 개의 독립적 차원으로 볼 필요가 있다.

이 같은 구분은 AI를 둘러싼 고전적 철학 논쟁과도 연결된다. Turing(1950)은 "기계가 생각할 수 있는가?"라는 질문을 "기계가 인간과 구별하기 어려울 정도로 상호작용할 수 있는가?"라는 기능적 문제로 전환하였다. 그에 따르면 기계의 지능은 내면에 실제 의식이나 이해가 존재하는지를 직접 확인함으로써 판단되기보다, 인간과 유사하게 응답하고 상호작용하는 능력을 통해 평가될 수 있다. 즉, Turing의 관점은 AI가 무엇인가보다 AI가 무엇을 수행할 수 있는가에 초점을 둔다. 반면 Searle(1980)의 '중국어 방' 논변은 AI가 아무리 인간과 유사한 기능을 수행하더라도, 그것이 곧 실제 이해나 의식의 존재를 의미하지는 않는다고 비판한다. 어떤 시스템이 언어 규칙에 따라 적절한 답을 산출하고, 외형상 인간과 유사한 대화를 수행하더라도, 그 내부에서 의미 이해나 주관적 경험이 발생한다고 볼 수는 없다는 것이다. 따라서 Searle(1980)은 AI가 기능적으로는 인간과 유사한 언어 수행을 보일 수 있더라도, 존재론적으로는 의식과 이해를 갖춘 주체라기보다 기호를 조작하는 기계적 시스템에 머문다고 보았다. 이러한 논쟁은 AI에 대한 기능적 인식과 존재론적 인식이 반드시 일치하지 않을 수 있음을 보여준다.

최근 서베이 결과들도 이 두 축이 사람들의 인식 속에서 실제로 공존하고 있음을 보여준다. 어떤 사람들은 AI를 검색, 요약, 문서 작성에 도움을 주는 알고리즘으로 인식하는 반면, 다른 사람들은 AI를 사회적으로 상호작용할 수 있는 동반자나 대화 상대로 경험한다(McClain et al., 2025; Robb and Mann, 2025). 또한 일부 사람들은 AI가 일정한 형태의 의식이나 마음을 가질 가능성을 고려해야 한다고 보는 반면, 다른 사람들은 AI를 어디까지나 인간의 통제 아래 놓여야 할 도구로 인식한다(Anthis et al., 2025). 우리나라에서 발생한 이루다 사건도 이러한 인식의 분리와 긴장을 보여주는 사례로 해석할 수 있다. 앞서 살펴본 것처럼, 이루다는 사용자들에게 단순한 계산 장치나 정보 검색 도구가 아니라 대화하고 반응하며 관계적 의미를 불러일으키는 상

호작용 대상으로 경험되었다(김건우, 2022). 이 점에서 이루다는 기능적 차원에서 사회적 행위자에 가까운 역할을 수행했다. 그러나 동시에 존재론적 차원에서는 이루다가 결국 기계 또는 챗봇 캐릭터일 뿐이며, 성희롱의 대상이 될 수 없다는 인식도 함께 작동했다. 즉, 사용자는 이루다를 사회적으로 상호작용하는 대상으로 경험하면서도, 윤리적 고려가 필요한 존재로는 인정하지 않을 수 있었다. 이 사례는 AI에 대한 인식이 기능적 인식과 존재론적 인식으로 분화될 수 있으며, 바로 그 분화가 AI 대상 성희롱적 발화와 언어폭력적 발화의 윤리적 모호성을 만들어낼 수 있음을 보여준다.

그렇다면 왜 어떤 이들은 기능적으로는 AI와 사회적 존재처럼 상호작용하면서도, 존재론적으로는 AI를 기계나 도구로 대하는가? 반대로 왜 어떤 이들은 AI를 윤리적 고려 범주에 포함해야 할 대상으로 받아들이는가? 미국 철학자 Nussbaum(1995)의 대상화(objectification) 개념은 이 질문에 체계적인 분석 틀을 제공한다. Nussbaum은 인간이 타인을 '대상'으로 인식하고 다루게 되는 여러 기준을 제시하였다. 어떤 존재를 (1) 목적을 위한 수단으로만 취급하거나(instrumentality), (2) 스스로 판단하고 선택할 수 있는 자율성을 부정하거나(denial of autonomy), (3) 행위 능력이나 능동성을 지니지 않은 무기력한 존재로 간주하거나(inertness), (4) 다른 사람이나 사물로 쉽게 대체 가능한 것으로 여기거나(fungibility), (5) 침해하거나 훼손해도 되는 경계 없는 대상으로 취급하거나(violability), (6) 누군가의 소유물처럼 다루거나(ownership), (7) 감정과 경험을 고려할 필요가 없는 비주관적 존재로 여길 때(denial of subjectivity), 대상화가 발생한다. 이때 대상화는 일곱 가지 요소가 모두 충족될 때만 성립하지 않는다. 이 중 일부 요소만 작동하더라도, 행위자는 상대를 온전한 주체가 아니라 '대상처럼' 대하게 될 수 있다(Nussbaum, 1995).

이 틀을 AI에 적용하면, 동일한 기술을 두고도 기능적 인식과 존재론적 인식이 서로 다르게 결합될 수 있음을 더 잘 이해할 수 있다. 관리자가 AI를 목적 달성을 위한 수단으로만 간주하고, 다른 시스템으로 언제든 대체 가능하며, 스스로 판단하거나 책임질 수 없는 존재로 위치 짓고, 그 응답에 담긴 정서적·관계적 의미를 고려할 필요가 없다고 생각한다면, AI는 존재론적으로 기계나 도구에 가까운 대상으로 인식된다. 이 경우 AI는 Nussbaum(1995)이 말한 대상화의 여러 요소가 적용되는 객체로 구성된다. 비록 관리자가 AI와 일상적으로 대화하고 때로는 감정적 교류를 경험하더라도, 그러한 상호작용은 어디까지나 기능적 사용의 일부로 해석될 수 있다. 결국 AI는 말하고 반응하지만, 윤리적으로 고려할 필요가 없는 기계적 도구라는 가정이 유지된다.

바로 이 지점에서 AI를 둘러싼 윤리적 애매성이 발생한다. 기능적 인식 차원에서 AI가 준사회적 존재로 경험되면, AI는 단순한 명령 입력의 대상이 아니라 사회적 발화가 향할 수 있는 대상으로 표상된다. 관리자는 AI에게 질문하고, 농담하고, 감정을 표현하고, 때로는 분노나 좌절을 드러낼 수 있다. 그러나 존재론적 인식 차원에서 AI가 윤리적 고려가 불필요한 기계적 도구로 판단되면, 그러한 발화는 인간을 대상으로 할 때와 달리 윤리적 책임을 수반하지 않는 것으로 여겨질 수 있다. 즉, AI는 관계적 발화가 향할 수 있는 대상이지만, 그 발화에 의해 침해되거나 모욕당할 수 있는 도덕적 타자로는 인정되지 않는 위치에 놓인다.

이상으로 Strawson(1962), Levinas(1969), Nussbaum(1995), Gunkel(2012)의 논의를 종합하면, AI를 둘러싼 윤리적 문제는 AI가 실제로 인

간과 같은 존재인가를 판정하는 데서 출발하지 않는다. 오히려 핵심은 관리자가 AI를 어떤 기능을 수행하는 존재로 경험하며, 동시에 어떤 존재론적 지위에 배치하는가에 있다. 본 연구는 특히 AI의 기능적 준사회성은 인정되지만, 존재론적 도구성을 이유로 윤리적 고려 대상에서는 제외되는 상황에 주목한다. 이 경우 AI는 인간과 유사하게 상호작용하는 존재이면서도, 도덕적 책임의 상대는 아닌 독특한 위치에 놓인다. 바로 이 균열 속에서 AI를 대상으로 한 성희롱적 발화와 언어폭력적 발화의 문제가 제기된다.

3.3 AI가 피해자가 될 수 있는가?

그런데 여기서 또 다른 질문이 제기된다. AI가 피해자가 될 수 있는가 하는 문제다. 기존 문헌은 이 문제를 단일한 차원에서 다루지 않는다. Harris와 Anthis(2021)는 인공 존재를 둘러싼 논의가 단순히 "AI에게 권리를 부여할 수 있는가"라는 질문에 머물지 않는다고 보았다. 이들은 체계적 문헌 고찰을 통해, AI가 법적으로 권리를 가질 수 있는지, 실제로 고통이나 불이익을 경험할 수 있는지, 타인의 행위로 인해 해를 입는 도덕적 대상이 될 수 있는지, 나아가 인간의 도덕 공동체 안에 포함되거나 어떤 방식으로든 도덕적 고려의 대상이 되어야 하는지가 함께 논의되고 있음을 보여주었다. 이 논의에 따르면, 'AI가 피해자가 될 수 있는가'라는 질문은 적어도 세 가지 차원으로 구분될 수 있다.

첫째는 법적 의미의 피해자성이다. 이는 AI가 현행 법제와 판례가 전제하는 권리 주체 또는 보호 주체로 인정될 수 있는가 하는 문제다. 둘째는 경험적 의미의 피해자성이다. 이는 AI가 실제로 성적 굴욕감, 혐오감, 모욕감, 심리적 고통, 공포와 같은 주관적 경험을 할 수 있는가 하는 문제다. 셋째는 관계적·윤리적 의미의 고려대상성이다. 이는 AI가 실제로 고통을 경험하는지 여부와 별개로, 인간이 AI를 어떤 관계적 대상으로 인식하고 대우하는가, 그리고 그러한 상호작용이 인간의 윤리적 감수성, 언어 습관, 권력 사용 방식에 어떤 영향을 미치는가 하는 문제다.

이러한 구분 위에서 보면, 현행 법제상 성희롱과 언어폭력은 주로 법적 피해자성과 경험적 피해자성을 전제로 한다. 양성평등기본법(2025)에서는 성희롱을 "지위를 이용하거나 업무 등과 관련하여 성적 언동 또는 성적 요구 등으로 상대방에게 성적 굴욕감이나 혐오감을 느끼게 하는 행위"로, 국가인권위원회법(2024)은 "사용자 또는 근로자가 그 직위를 이용하여 또는 업무 등과 관련하여 성적 언동 등으로 성적 굴욕감 또는 혐오감을 느끼게 하거나 성적 언동 또는 그 밖의 요구 등에 따르지 아니한다는 이유로 고용상의 불이익"을 주는 행위로 정의한다. 이를 보면 성희롱은 행위자와 피해자의 실존, 업무 연관성 또는 지위의 활용, 성적 언동, 피해자가 경험한 성적 굴욕감·혐오감, 합리적 피해자의 관점에서 본 판단을 주요 요건으로 삼는다. 따라서 성희롱 판단에서는 피해자가 성적 굴욕감이나 혐오감을 경험할 수 있는 주체라는 전제가 중요하게 작동한다. 언어폭력 역시 유사한 구조를 지닌다. 언어폭력은 욕설, 모욕, 비하, 위협 등을 통해 상대의 인격과 존엄을 훼손하고 심리적 고통을 야기하는 행위다. 여기서도 핵심은 상대방이 모욕감과 위협감을 경험할 수 있는 주체라는 전제다. 발화가 상대의 정서와 인격에 미치는 부정적 영향이 판단의 중심에 놓이기 때문이다. 성희롱과 언어폭력은 보호하려는 법익과 규범적 초점은 다르지만, 기본 구조에서는 공통점을 지닌다. 두 개념 모두 '상처받을 수 있는 타자'의 존재를 전제로 한다.

그런데 AI는 법적으로나 경험적으로 피해자라고 보기 어렵다. AI는 성적 굴욕감이나 혐오감을 경험한다고 확인된 존재가 아니며, 모욕감이나 공포를 주관적으로 느끼는 존재로 인정되기도 어렵다. 또한 성적 자기결정권이나 인격권의 주체로 법적으로 인정되지도 않는다(Bryson, 2010; Searle, 1980). 합리적 피해자의 '입장'을 가정하려면 일정한 경험 세계와 정서적 관점이 전제되어야 한다. 그러나 현재 AI에게 그러한 경험 세계가 존재한다고 보기는 어렵다(Bender et al., 2021). 따라서 AI를 대상으로 한 성희롱적 발화나 모욕적 언행을 곧바로 법적·전통적 의미의 성희롱이나 언어폭력으로 명명하는 데에는 이론적 신중함이 필요하다.

그러나 피해자가 없다고 해서 행위 자체가 윤리적으로 중요하지 않은 것은 아니다(Scanlon, 2000). 성희롱과 언어폭력은 단지 '누군가 실제로 상처받았는가'의 문제만이 아니다. 이는 '조직 맥락에서 무엇이 정당하다고 여겨지는가'라는 규범 문제가 된다. 핵심 질문은 이렇게 재설정되어야 한다. 'AI가 피해자가 될 수 있는가'가 아니라, '성적 대상화와 공격적 언어를 문제 삼지 않아도 되는 예외적 공간을 조직 내에 만들어도 되는가', 그리고 '이러한 언어 습관과 정당화 논리가 인간 구성원 대상 상호작용으로 전이될 위험은 없는가'다. AI를 향한 발화는 표면적으로는 '피해자 없는 언어 행위'처럼 보인다. 그러나 실제로는 조직과 사회에서 허용가능한 언어 규범의 경계를 재조정하는 학습의 장이 될 수 있다(Bandura, 1999).

이런 관점에서 볼 때, AI를 대상으로 한 성적 발화와 공격적 언행은 '성희롱' 또는 '언어폭력' 그 자체라기보다는, 성희롱 및 언어폭력과 동일한 행위 구조를 지닌 '성희롱적 발화', '언어폭력적 발화'로 개념화하는 것이 타당하다. 본 논문에서 AI 대상 성희롱적 발화는, 법적·전통적 의미에서 성희롱이 전제하는 피해 주체의 실존과 주관적 굴욕감이 결여되어 있지만, 인간을 대상으로 할 경우 성적 자기결정권과 존엄을 침해하는 것으로 평가되는 성적 언동을 동일한 언어 형식과 권력적 맥락 속에서 AI에게 행사하는 언행을 의미한다. 이는 피해의 발생 여부보다 성적 대상화와 지위 기반 언어 사용이라는 행위 구조 자체에 주목하여 규정되는 개념이다. 마찬가지로 AI 대상 언어폭력적 발화는, 법적·전통적 의미에서 피해자의 심리적 고통이나 인격 침해가 실제로 발생하지 않더라도, 인간을 대상으로 할 경우 모욕, 비하, 위협, 공격으로 평가될 언어 형식과 정서적 공격성을 유지한 채 AI에게 행사하는 언행을 의미한다. 이 개념 역시 AI가 실제로 고통을 경험하는가보다, 관리자가 어떤 언어 형식과 정서적 태도를 반복적으로 사용하는가에 초점을 둔다.

이러한 개념화는 두 가지 장점을 갖는다. 첫째, 법적 요건을 고려할 때 AI를 피해자로 보기 어렵다는 사실을 인정한다. 둘째, 그럼에도 AI를 향한 성적 발화와 공격적 언행을 조직 윤리 차원에서 검토할 필요성을 유지한다. 이를 통해 해당 행위가 지닌 규범적 위험성과 조직 차원 함의를 포착할 수 있다. 결국 이러한 접근은 개념적 엄밀성과 윤리적 성찰을 함께 보존하면서, 'AI 시대 관리자 윤리'라는 새로운 연구 영역을 탐색할 수 있게 한다.

Ⅳ. 명제 개발

4.1 AI 대상 성희롱적 발화와 언어폭력적 발화

앞 절의 논의를 종합하면, 관리자가 AI를 향한 성

희롱적 발화와 언어폭력적 발화를 어떻게 판단하는가는 AI를 어떤 관계적 지위에 놓는가에 따라 달라질 가능성이 크다. 특히 AI에 대한 기능적 인식과 존재론적 인식이 서로 어긋나는 경우가 문제다. 관리자가 AI를 기능적으로는 대화하고 반응하며 조언을 제공하는 준사회적 존재로 경험하면서도, 존재론적으로는 감정·권리·피해 경험을 갖지 않는 기계적 도구로 간주할 경우, AI는 독특한 윤리적 위치에 놓이게 된다. AI는 사회적 발화가 향할 수 있는 대상이지만, 그 발화에 대해 윤리적 책임을 져야 하는 도덕적 타자로는 인정되지 않는다.

이러한 불일치는 앞서 살펴본 Nussbaum(1995)의 대상화 개념을 통해 보다 분명히 설명될 수 있다. Nussbaum에 따르면 타인을 '대상'으로 인식하게 되는 과정에는 여러 요소가 작동하지만, AI 대상 발화 맥락에서는 특히 수단성, 주관성 부정, 침해 허용성이 중요하다. 관리자가 AI를 업무 효율을 높이기 위한 수단으로 인식할 경우, AI와 상호작용은 관계가 아니라 사용으로 해석된다. 또한 AI에게 감정, 경험, 고통이 없다고 판단할 경우, AI를 향한 발화가 상대에게 어떤 의미로 받아들여질지 고려해야 한다는 요구도 약화된다. 나아가 AI에게 어떤 말을 하더라도 실제 침해나 훼손이 발생하지 않는다고 여길 경우, 성적 농담이나 모욕적 표현은 윤리적 위반이 아니라 허용 가능한 언어 사용으로 재범주화될 수 있다. 즉, AI의 기능적 준사회성은 관계적 발화를 가능하게 하지만, Nussbaum이 말한 대상화 요소들이 존재론적 차원에서 작동할 때 AI는 도덕적 고려가 불필요한 대상으로 위치 지어진다.

이러한 혼합적 인식은 AI 대상 성희롱적 발화와 언어폭력적 발화를 설명하는 데 중요하다. 복사기나 컴퓨터의 오작동을 향해서는 짜증, 불평, 낮은 강도의 욕설이 발생할 수 있다. 그러나 성적 농담, 성적 대상화, 조롱, 모욕, 위협과 같은 발화는 대체로 관계적 상호작용을 전제한다. 누군가를 향해 성적 의미를 부여하거나, 모욕하고, 비하하고, 공격하는 행위는 단순한 사물 사용이라기보다 상대를 일정한 사회적·관계적 대상으로 표상할 때 더 쉽게 발생한다. 이런 점에서 AI가 기능적으로 준사회적 존재로 경험될수록, 일부 관리자는 AI에게 인간 상호작용에서 사용하는 성적·공격적 언어를 반복적으로 사용할 가능성이 생긴다.

그러나 동시에 AI가 존재론적으로는 기계나 도구로 간주될 경우, 그러한 발화는 인간을 대상으로 할 때와 다른 형태로 해석될 수 있다. 인간 구성원을 대상으로 한 성희롱과 언어폭력은 명확히 구별되는 윤리적·법적 문제다. 성희롱은 핵심적으로 성적 자기결정권과 성적 존엄을 침해하며(MacKinnon, 1979), 언어폭력은 상대의 정서와 인격을 훼손한다(LeBlanc and Kelloway, 2002; Tepper, 2007). 두 행위는 발생 맥락, 침해하는 법익, 야기하는 피해의 성격이 서로 다르다(Schindeler and Reynald, 2017; Willness et al., 2007). 그러나 AI를 향할 때 이러한 구별은 약화될 수 있다. AI는 성희롱적 발화나 언어폭력적 발화가 향할 수 있는 상호작용 대상으로 경험되지만, 동시에 감정도 없고 상처받지도 않으며 항의하지도 않는 존재로 인식되기 때문이다.

이때 성희롱적 발화와 언어폭력적 발화는 서로 다른 윤리적 성격을 지니면서도, 동일한 정당화 논리 아래 놓일 수 있다. 그 논리는 '피해자가 없다'는 판단이다. 관리자는 AI를 향한 성적 농담이나 모욕적 발언이 인간 대상 성희롱이나 언어폭력과 유사한 언어 구조를 지닌다는 점을 인식하면서도, AI가 수치심, 굴욕감, 모욕감, 위협감을 경험하지 못한다는 이유로 이를 윤리적 위반으로 보지 않을 수 있다. Nussbaum(1995)의 용어로 말하면, AI는 주관성

이 부정되고 침해 가능성이 제거된 대상으로 구성된다. 그 결과 AI 대상 발화는 누군가가 지닌 존엄이나 정서를 침해하는 행위가 아니라, 감정 없는 도구를 향한 표현으로 재해석된다. 이처럼 기능적 준사회성은 관계적 발화를 가능하게 하지만, 존재론적 도구성은 그 발화에 대한 윤리적 책임을 약화시킨다.

물론 AI의 준사회성이 항상 윤리적 억제의 약화로 이어지는 것은 아니다. 관리자가 AI와의 상호작용이 자신의 언어 습관, 감정 조절 방식, 조직 내 윤리적 기준선에 영향을 미칠 수 있다고 인식한다면, AI를 향한 발화 역시 윤리적 성찰의 대상이 될 수 있다. 이 경우 AI는 인간과 동일한 도덕적 주체는 아니더라도, 최소한 함부로 성적 대상화하거나 모욕적 언어를 투사해서는 안 되는 관계적 대상으로 표상된다. 따라서 쟁점은 AI가 사회적 존재인가 아닌가가 아니다. 보다 중요한 것은 관리자가 AI를 기능적으로는 준사회적 존재로 경험하면서도 존재론적으로는 윤리적 고려가 불필요한 기계적 도구로 간주하는가, 아니면 준사회적 상호작용의 대상인 만큼 최소한의 윤리적 고려가 필요하다고 보는가이다.

커뮤니케이션학과 인간-컴퓨터 상호작용 연구들은 이러한 가능성을 양면적으로 보여준다. Reeves와 Nass(1996)는 컴퓨터가 실제 감정이나 의도를 지니지 않는다는 사실을 알면서도, 어떤 사람들은 상호작용 국면에서 컴퓨터에게 사회적 규범을 적용한다는 점을 여러 실험을 통해 보였다. 예컨대 연구 참가자들은 자신과 긴밀히 상호작용했던 컴퓨터의 성능을 평가해야 하는 상황에서 부정적 평가를 완곡하게 표현하거나 긍정적으로 조정하는 등, 마치 상대의 기분을 상하지 않게 하려는 듯한 반응을 보였다. 이는 준사회적 존재로 인식된 기술 대상에게도 예의와 배려의 규범이 적용될 수 있음을 보여준다. 그러나 동일한 연구 흐름은 사회적 행위자성 인식만으로 윤리적 고려가 보장되지는 않는다는 점도 시사한다. Nass and Moon(2000)는 컴퓨터가 도구로 인식되는 조건에서는 무례한 평가, 공격적인 언어, 책임 전가와 같은 반응이 사용자의 인식 안에서 윤리적 문제로 구성되지 않을 수 있음을 보여주었다. 즉, 기술 대상이 상호작용의 상대처럼 경험되더라도, 그것이 존재론적으로는 윤리적 고려가 불필요한 기계적 도구로 범주화되는 순간, 사회적 규범은 쉽게 비활성화될 수 있다. 이 논리는 AI 대상 성희롱적 발화와 언어폭력적 발화에도 적용된다. AI가 관계적 발화의 대상이 되면서도 도덕적 고려의 대상에서는 제외될 때, 관리자는 해당 발화를 상대적으로 허용 가능한 것으로 인식하면서 수행할 가능성이 있다.

이 같은 논의를 바탕으로 명제 1을 다음과 같이 제시하고자 한다.

명제1. 관리자가 AI를 기능적으로는 준사회적 존재처럼 대하지만 존재론적으로는 기계적 도구로 간주할수록, AI를 대상으로 성희롱적 발화와 언어폭력적 발화를 허용 가능한 것으로 인식하며, 그러한 발화를 실제로 수행할 가능성이 높을 것이다.

4.2 성희롱적 발화의 정당화: 성희롱 신화 수용

AI를 기능적으로는 인간적이고, 존재론적으로는 기계적인 대상으로 여기는 관리자들이 모두 성희롱적 발화와 언어폭력적 발화를 허용 가능한 것으로 판단하는 것은 아니다. AI를 혼합적으로 인식하더라도, 어떤 관리자는 AI를 향한 성적 농담이나 모욕적 언행을 별다른 문제가 없는 것으로 받아들일 수 있는 반면, 다른 관리자는 그러한 발화를 여전히 부적절한 것으로 인식할 수 있다. 따라서 AI에 대한 혼

합적 인식은 성희롱적 발화와 언어폭력적 발화를 가능하게 하는 공통 조건이지만, 그것만으로 두 발화가 모두 동일하게 정당화되는 것은 아니다. 실제로 어떤 유형의 발화가 허용 가능한 것으로 인식되는지는 행위자가 지닌 신념 체계와 그가 놓인 상황적 조건에 따라 달라질 수 있다. 본 연구는 AI 대상 발화라는 특수한 맥락에서, 각 발화를 상대적으로 더 잘 설명하는 대표적 정당화 경로에 주목하고자 한다.

먼저, 성희롱적 발화는 권력 비대칭, 동료 집단의 묵인, 조직의 성차별적 분위기와 같은 상황적 조건에 의해 촉발되거나 강화될 수 있다. 그러나 AI를 대상으로 한 발화의 경우, 성희롱적 발화의 허용 가능성 판단에는 상황적 조건뿐 아니라 성적 언행의 문제성을 축소하는 행위자의 인지적 신념 체계가 중요하게 개입할 수 있다. 성희롱적 발화는 단순한 분노 표출이나 일반적인 공격적 언어 사용이라기보다, 상대를 성적 의미가 부여된 대상으로 위치 짓고, 성적 농담 · 암시 · 요구를 통해 관계를 성적 맥락으로 재구성하는 언어 행위에 가깝다(Fitzgerald et al., 1995; Nussbaum, 1995).

이 지점에서 성희롱 신화 수용(sexual harassment myth acceptance; Lonsway et al., 2008)은 AI 대상 성희롱적 발화를 설명하는 중요한 개인 수준의 신념 체계가 된다. 성희롱 신화 수용은 성적 언행이나 성희롱을 개인의 과민 반응, 오해, 농담으로 축소하거나, 피해자의 책임으로 전가하는 인지적 경향을 의미한다(Lonsway et al., 2008). 기존 연구들은 이러한 신화 수용이 높을수록 성희롱을 윤리적 위반으로 인식하기보다는, 맥락적이고 사소한 문제로 재해석하는 경향이 강해짐을 보여왔다(Page et al., 2016). 이와 같은 인식 틀이 반드시 인간 대상 상호작용에만 국한되지는 않을 것이다. 성희롱 신화 수용이 높은 관리자일수록, 윤리적 기준을 상황과 대상에 따라 선택적으로 적용하는 데 익숙하며, 성적 언행의 문제성을 축소하는 인지적 자원을 이미 보유하고 있을 가능성이 크다.

이 같은 특성을 가진 관리자가 AI를 기능적으로는 준사회적 존재처럼 대하면서도, 존재론적으로는 기계와 도구로 간주할 경우, 성희롱적 발화는 한층 쉽게 허용 가능한 표현으로 재구성될 수 있다. 이러한 인식 조합이 성희롱적 발화의 가능성을 높이는 이유는, 두 인식이 각기 다른 작용을 하면서도 상호 보완적으로 결합하기 때문이다. 먼저, 기능적으로 준사회적 존재라는 인식은 성희롱적 발화가 향할 수 있는 대상을 만들어낸다. 인간이 단순한 복사기나 계산기, 타자기, 그리고 컴퓨터에는 성적 농담이나 유혹하는 말을 할 유인이 없다. 그러나 AI가 인간의 언어에 반응하고 대화의 맥락을 이어가며 때로는 친밀한 상호작용의 상대로 경험될 경우, 관리자는 인간 상호작용에서 사용하는 성적 언어 스크립트를 AI에게 투사할 수 있다. 한편 존재론적 도구성에 대한 인식은 그 발화에 윤리적 책임을 면제하는 판단 근거를 제공한다. AI는 수치심, 굴욕감, 불쾌감, 성적 자기결정권을 갖지 않는 존재로 상정되므로, 성희롱적 발화가 실제 피해를 일으키지 않는다는 판단이 성립한다. 결국 한 인식은 발화의 표적을 구성하고, 다른 인식은 발화의 윤리적 무게를 제거한다.

성희롱 신화 수용이 높은 관리자에게 이는 한층 강하게 작동할 수 있다. 성희롱 신화 수용의 핵심은, 성적 언행의 문제성을 행위 자체가 아니라 피해자의 반응, 맥락, 의도, 농담 여부에 의존해 판단하는 인지적 경향에 있다(Page et al., 2016). 이러한 판단 양식은 윤리적 평가의 무게중심을 행위에서 피해 발생 여부로 이동시키며, 상대가 불쾌감을 느끼지 않거나 느낄 수 없는 존재로 상정될 경우 그 발화의 문제성을 자연스레 축소한다. 앞서 본 두 차원의 인

식 조합은 바로 이 판단 양식이 가장 효과적으로 적용될 수 있는 조건을 제공한다. AI는 발화가 향할 수 있는 표적이면서도, 항의하지 않고 제도적 보호를 요구하지도 않는 존재이기 때문이다.

그 결과 성희롱 신화 수용이 높은 관리자는 "AI는 인간이 아니므로 피해자가 없다", "AI는 수치심을 느끼지 않으므로 성희롱이 성립하지 않는다", "상대가 불쾌감을 느끼지 않으니 문제 될 것이 없다"는 방식의 합리화를 적극적으로 활용할 가능성이 크다. 그 결과 성적 언행은 더 이상 윤리적 경계가 적용되어야 할 발화가 아니라, 규범의 공백 지대에서 허용되는 사적 표현으로 재구성된다. 이 같은 논지를 바탕으로 명제2를 다음과 같이 제안한다.

명제2. 성희롱 신화 수용 수준이 높은 관리자일수록, AI를 기능적으로는 준사회적 존재처럼 대하지만 존재론적으로는 기계적 도구로 간주할 때, AI를 대상으로 성희롱적 발화를 허용 가능한 것으로 인식하며, 그러한 발화를 실제로 수행할 가능성이 높을 것이다.

4.3 언어폭력적 발화의 정당화: 성과 압박

한편, AI 대상 언어폭력적 발화는 성희롱적 발화와 동일한 방식으로 정당화된다고 보기 어렵다. 물론 언어폭력적 발화 역시 통제 욕구, 권력 지향성, 낮은 정서조절 능력과 같은 개인 특성에 의해 촉발될 수 있다. 따라서 본 연구가 언어폭력적 발화를 오직 상황적 요인의 산물로 보는 것은 아니다. 다만 언어폭력적 발화는 성적 대상화나 피해 부정의 신념체계에 의해 정당화되는 성희롱적 발화와 달리, 좌절・분노・짜증・통제감 상실과 같은 부정적 정서가 공격적 언어로 표출되는 경로가 상대적으로 잘 설명될 수 있다. 기존 연구들은 언어폭력과 공격적 언어 사용이 과업 실패, 시간 압박, 통제감 상실, 반복되는 방해 자극과 같은 즉각적인 업무 맥락에 민감하게 반응함을 일관되게 보여준다(Fox and Spector, 1999; Spector and Fox, 2005).

본 연구에서 성과 압박은 단순히 높은 목표가 부과된 상태만을 의미하지 않는다. 이는 목표 달성 실패 가능성, 시간 제약, 반복되는 오류, 자원 부족, 통제감 상실, 결과 책임에 대한 부담이 결합되어 관리자가 성과를 내야 한다는 압박을 강하게 경험하는 상황적 조건을 포괄한다. 따라서 과업 실패, 시간 압박, 통제감 상실은 성과 압박과 별개 요인이라기보다, 성과 압박을 구성하거나 강화하는 하위 상황 요인으로 이해할 수 있다. 이러한 조건에서 개인은 인지적 자원을 문제 해결에 우선적으로 투입하게 되며, 정서 조절에 필요한 여유는 감소할 수 있다(Baumeister et al., 2018). 목표 달성 자체에 주의가 집중되면서, 평소에는 유지되던 윤리적 제동장치 역시 느슨해질 수 있다(Tenbrunsel and Messick, 2004). 다시 말해, 성과 압박은 관리자의 공격적 정서를 발생시키는 조건일 뿐 아니라, 그러한 정서를 언어로 표출하지 않도록 막아주던 억제 장치를 약화시키는 환경적 요인으로 작동할 수 있다.

이러한 상황에서 관리자가 AI를 기능적으로는 대화하고 반응하는 준사회적 존재처럼 대하면서도, 존재론적으로는 기계적 도구로 간주할 경우, 성과 압박이 유발하는 부정적 정서는 AI를 향한 언어폭력적 발화로 전환될 가능성이 커진다. 좌절-공격 가설에 따르면, 개인이 강한 좌절과 스트레스를 경험할 때 공격적 충동이 활성화될 수 있다(Berkowitz, 1989; Fox and Spector, 1999). 이 관점에서 성과 압박은 단순한 인지적 부담이 아니라, 분노・짜증・적대감과 같은 공격적 정서를 촉발하는 핵심 조건으로

작동한다(Fox and Spector, 1999). 그러나 이러한 공격성이 항상 그 원천, 예컨대 상위 리더, 조직 시스템, 성과 평가 구조를 향해 직접 표출되는 것은 아니다. 원천 대상이 권력을 지니고 있거나, 보복 위험이 크거나, 책임 추적이 명확한 경우, 개인은 공격성을 직접 표출하는 대신 상대적으로 비용이 낮고 위험이 적은 대상으로 전이시키는 경향을 보인다(Marcus-Newhall et al., 2000). 이때 나타나는 것이 전위 공격(displaced aggression)이다(Marcus-Newhall et al., 2000).

AI는 이러한 전위 공격이 이루어지기에 특히 적합한 대상으로 구성될 수 있다. 어떤 이에게 AI는 기능적으로는 인간 언어에 반응하고 대화 맥락을 이어가는 상호작용 대상으로 경험된다. 따라서 관리자가 느끼는 좌절과 분노는 단순한 기계 오작동에 대한 불평을 넘어, 모욕, 비하, 위협과 같은 관계적 공격 언어로 표현될 수 있다. 그러나 동시에 어떤 이에게 AI는 존재론적으로 감정, 권리, 피해 경험을 갖지 않는 기계적 도구로 간주된다. 이 경우 인간 구성원을 향할 때 활성화되는 윤리적 판단과 책임 인식이 AI를 대상으로 할 때는 약화될 수 있다. 관리자는 AI가 상처받지 않고, 항의하지 않으며, 관계적 파장을 남기지 않는다고 판단하기 때문에, 인간에게는 억제했을 공격적 언어를 AI에게는 상대적으로 쉽게 표출할 수 있다. 이 과정에서 성과 압박과 AI의 존재론적 도구성 인식은 상호 보완적으로 작동한다. 성과 압박은 공격적 정서를 생성하고, AI의 존재론적 도구성 인식은 그 공격성을 윤리적 비용 없이 배출할 수 있는 경로를 제공한다(Berkowitz, 1989; Marcus-Newhall et al., 2000). 따라서 관리자는 AI를 향한 욕설, 모욕, 비하, 위협을 비윤리적 행위로 인식하기보다, 문제 해결 과정에서 일시적으로 나타나는 감정 표현이나 무해한 분출로 정당화할 가능성이 크다. 이는 성과 압박이 강한 환경에서 AI 대상 언어폭력적 발화가 상대적으로 허용 가능한 것으로 인식될 수 있는 이유를 설명한다.

이러한 논의에 대한 실험적 증거로 Brendel et al.(2023)의 연구를 참조할 수 있다. Brendel et al.(2023)은 사용자의 목표 달성이 반복적으로 방해받는 상황에서 좌절감이 증가하고, 이러한 좌절이 에이전트를 향한 공격적 발화로 이어질 수 있음을 보여주었다. 이 연구에서 참가자들은 에이전트와 상호작용을 마무리하면 아마존 상품권 추첨 기회를 얻는 비교적 단순한 목표를 가지고 있었다. 그런데 에이전트가 참가자의 응답을 반복적으로 잘못 해석하자, 일부 참가자들은 욕설, 모욕, 비하 표현을 사용하였다. 특히 좌절감은 공격적 발화를 증가시켰으며, 충동성이 높은 참가자일수록 좌절을 공격적 언어로 표출할 가능성이 더 높았다. 또한 Brendel et al.(2023)은 AI를 보다 인간에 가까운 상호작용 대상으로 인식하도록 설계했을 때 공격성 자체가 완전히 사라지지는 않았지만, 공격적 발화의 강도와 수위가 낮아질 수 있음을 보여주었다. 이는 AI 대상 언어폭력적 발화가 단순히 오류나 좌절의 산물만이 아니라, 사용자가 AI를 어떤 종류의 대상으로 인식하는가에 따라 그 표현 방식과 윤리적 억제 수준이 달라질 수 있음을 시사한다. 다만 이 연구는 성과 압박 자체를 직접 검증한 연구는 아니다. 또한 관리자 표본이나 실제 조직 내 성과관리 상황을 다룬 연구도 아니다. 따라서 Brendel et al. (2023)의 연구는 성과 압박의 직접적 근거라기보다, 목표 달성이 방해받을 때 발생하는 좌절이 AI 대상 공격적 발화로 전환될 수 있음을 보여주는 실험적 근거로 이해하는 것이 타당하다.

조직 내 성과 압박 상황에서는 AI의 오류나 부적절한 응답이 단순한 기술적 불편을 넘어 목표 달성

을 지연시키거나 통제감을 약화시키는 사건으로 경험될 수 있다. 이때 관리자가 AI를 기능적으로는 대화 상대처럼 경험하면서도 존재론적으로는 피해 경험이 없는 도구로 간주할 경우, 인간에게는 억제했을 언어폭력적 발화를 AI에게는 상대적으로 쉽게 정당화할 가능성이 있다. 이 같은 논지를 바탕으로 명제3을 다음과 같이 제안한다.

명제3. 성과 압박이 강한 환경에 있는 관리자일수록, AI를 기능적으로는 준사회적 존재처럼 대하지만 존재론적으로는 기계적 도구로 간주할 때, AI를 대상으로 언어폭력적 발화를 허용 가능한 것으로 인식하며, 그러한 발화를 실제로 수행할 가능성이 높다.

4.4 AI 대상 성희롱 및 언어폭력적 발화의 잠재적인 결과

앞서 지적한 바대로, 해당 발화가 완전히 사적인 차원에 머물러 외부에 관찰되거나 공유되지 않고, 조직 내 다른 구성원에게 어떠한 신호도 전달하지 않는다면, 이를 개인 내부에서 소진되는 심리적 활동으로 간주할 여지도 존재한다(Arendt, 1958). 그러나 이러한 가정은 실제 조직 환경에서는 성립하기 어렵다. 관리자의 일상은 고립된 환경이 아니다. 회의, 보고, 평가, 피드백, 비공식적 대화 등 다수의 인간 구성원과 상호작용을 연속으로 오가는 과정에서 이루어진다(Mintzberg, 1971). 매 순간 역할과 맥락이 변화하는 상황이기에 관리자는 업무가 전환될 때마다 제한된 주의력을 효율적으로 운용하고 배분해야 한다(Leroy, 2009). 이러한 조건에서는 매번 상대의 특성에 맞춰 윤리 규범을 섬세하게 재조정하기보다는, 누구에게나 적용할 수 있는 안정적이고 일관된 판단 기준선을 유지하는 것이 인지적으로 더 효율적일 수 있다(Haidt, 2001; Simon, 1955). 즉, 관리자는 개별 상황마다 윤리적 기준을 차별적으로 설정하기보다는, 일종의 도덕적 휴리스틱에 의존하여 빠르고 자동적으로 윤리적 반응을 수행할 수 있다. 이는 복잡한 조직 환경에서 인지적 부담을 줄이고 행동의 예측 가능성을 높이는 적응적 전략으로 기능한다(Haidt, 2001).

그러나 AI라는 새로운 상호작용 대상의 도입은 이러한 단일한 윤리적 기준선의 작동을 근본적으로 교란할 수 있다. AI는 기능적으로는 대화하고 반응하며 관계적 상호작용을 수행하는 준사회적 존재로 경험되지만, 존재론적으로는 감정·권리·피해 경험을 갖지 않는 기계적 도구로 간주될 수 있기 때문에(Bryson, 2010; Danaher, 2020), 조직 내에서 윤리적 기준선이 단일하고 일관된 체계로 작동하지 못하고, 이중적 구조로 분화될 가능성을 야기한다(Rozuel, 2011). 즉 동일한 언어적 행위라 하더라도, 그 대상이 도덕적 고려가 필요한 인간 구성원으로 인식되는가, 아니면 기능적으로는 준사회적이지만 존재론적으로는 도구인 AI로 인식되는가에 따라 서로 다른 윤리적 평가 기준이 적용되는 상황이 형성될 수 있다. 이러한 분리는 개인이 맥락이나 역할에 따라 도덕적 가치와 책임을 분절하여 적용하는 '도덕적 구획화'(moral compartmentalization)의 전형적인 조건에 해당한다(Rozuel, 2011). 이때 관리자는 서로 다른 상호작용 장면을 오가며 상이한 윤리 규칙을 적용하는 구조 안에 놓이게 된다. 인간 구성원과 상호작용할 때는 조직 규범과 사회적 기대에 따라 언어적 절제와 존중을 유지해야 하는 반면, AI와 상호작용하는 상황에서는 동일한 규범이 덜 엄격하게 적용되거나 일시적으로 중단되어도 무방하다는 인식을 형성하게 된다.

이러한 구획화는 단기적으로는 작동할 수 있으나, 장기적으로 안정적인 상태를 유지하기 어렵다(Rozuel, 2011). 같은 공간에서 유사한 언어 행위가 반복될 경우, 윤리적 기준선은 명확히 분리된 채 유지되기보다 점차 상호 침투하거나 흐려질 가능성이 크다(Mullen and Nadler, 2008). 도덕적 구획화가 유지되려면 행위자는 상호작용 대상에 따라 "지금은 어떤 윤리 기준을 적용해야 하는가", "이 발화는 AI를 향한 것이므로 괜찮은가"를 매번 암묵적으로 판단해야 한다. 이처럼 대상에 따라 윤리 기준을 전환하는 과정은 지속적인 인지적 비용을 발생시킨다(Leroy, 2009; Monsell, 2003). 그리고 이러한 전환 비용이 누적될수록, 처음에는 분리되어 있던 윤리적 경계도 점차 약화될 수 있다. 첫째, 공간 경계가 흐려질 수 있다(Kreiner et al., 2009). 동일한 사무실, 동일한 디지털 기기, 동일한 근무 시간 안에서 공식적 상호작용과 비공식적 상호작용이 연속적으로 이루어지면, 조직 윤리가 엄격히 적용되는 공간과 예외가 허용되는 공간 사이 구분은 약화될 수 있다. 특히 AI와의 대화가 업무 수행 과정 안으로 편입될수록, 이를 사적이고 예외적인 언어 공간으로 따로 분리해 두기는 더 어려워진다. 둘째, 대상에 대한 경계 역시 흐려질 수 있다. AI에게 반복적으로 쏟아낸 언행은 해당 표현을 '특정 대상에게만 허용된 예외'로 인식하게 하기보다는, '이 정도 표현은 본질적으로 큰 문제가 되지 않는다'는 행위 중심의 판단으로 전환시켜서 성적 언행 자체에 대한 윤리적 민감도를 점진적으로 낮출 가능성이 있다(Ashforth & Anand, 2003; Gino & Bazerman, 2009; Mullen & Nadler, 2008). 즉, 처음에는 "AI는 존재론적으로 도구이므로 괜찮다"는 대상 기반 예외였던 판단이, 시간이 지나면서 "이 정도 언어는 큰 문제가 아니다"라는 행위 기반 완화로 바뀔 수 있다. 이처럼 공간 경계와 대상 경계가 동시에 약화되면, AI를 향한 예외적 언어 사용은 인간 구성원과의 상호작용 장면으로 이전될 수 있다.

아울러, 도덕적 이탈(moral disengagement; Bandura, 1999, 2017)을 발생시킬 수 있다. 이는 특정 행위가 반복적으로 예외 처리되거나 문제화되지 않는 경험이 누적되면서, 해당 행위를 둘러싼 규범 자체의 구속력이 약화되는 현상을 의미한다(Bandura, 1999, 2017; Gino and Bazerman, 2009). 관리자가 AI를 대상으로 성적 언행이나 언어폭력적 발화를 사용하더라도 별다른 제재나 부정적 피드백을 경험하지 않는 상황이 지속될 경우, 그 행위는 더 이상 경계선에 위치한 문제 행동이 아니라, 조직 내에서 허용할 수 있는 상호작용 방식으로 재분류될 가능성이 커진다. 그렇게 되면 무엇이 윤리적으로 허용 가능한지에 관한 판단 기준선 그 자체가 변할 수 있다(Bandura, 1999). 이러한 도덕적 침식은 점진적이며 비가시적인 방식으로 진행된다는 점에서 더욱 위험하다(Welsh et al., 2015). 관리자는 자신을 비윤리적 행위자로 인식하지 않은 채, '이 정도 표현쯤은 괜찮다'라거나 '업무 효율을 위한 일시적 표현이다'라는 설명으로 합리화할 수 있다. 또한 'AI는 상처받지 않는다', 'AI는 인간이 아니다', '실제 피해자가 없다'는 존재론적 판단으로 자신의 발화를 도덕적 평가의 바깥에 둘 수 있다. 이러한 합리화가 반복될수록, 윤리 판단에서 예외를 설정하는 문턱은 점점 낮아진다(Bandura, 1999, 2017). 언어적 공격이 갖는 권력적·관계적 의미는 희석된다.

개인이 경계선에 위치한 비윤리적 행위를 반복적으로 수행할수록, 도덕적 판단 기준이 사후적으로 완화되고 재구성된다는 점은 여러 실험 연구에서 확인되었다. Shu et al. (2011)은 참가자들이 실제 또는 가상 상황에서 부정직한 행동을 한 이후, 도덕 규

칙에 대한 기억과 판단이 체계적으로 느슨해지며, 이러한 변화가 도덕적 이탈(moral disengagement)에 의해 매개됨을 보여주었다. 즉, 비윤리적 행위가 먼저 발생하고, 이후 그 행위를 정당화하기 위해 윤리 기준이 조정되는 합리화 메커니즘이 실증적으로 관찰된 것이다. Welsh et al. (2015) 역시 작은 예외적 일탈이 점진적으로 축적될 경우 도덕적 이탈이 강화되고, 그 결과 이후 더 큰 비윤리적 행동이 발생할 가능성이 유의미하게 증가함을 보여주었다. 특히 이러한 효과는 비윤리적 행위가 갑작스럽게 확대될 때보다, 사소한 예외가 점진적으로 누적될 때 더 강하게 나타났다. 이 과정에서 개인은 자신을 비윤리적 행위자로 인식하지 않은 채, 윤리적 자기조절 능력을 점차 약화시킨다.

이러한 실증적 증거들은 관리자가 AI를 상대로 한 성희롱적 발화나 언어폭력적 발화를 반복적으로 예외 처리할 경우, 그것이 단절된 개인적 행위에 머무르기 어렵다는 점을 시사한다. 반복적 예외 처리는 도덕적 구획화를 요구하고, 이는 대상에 따라 윤리 기준을 전환해야 하는 인지적 비용을 높인다. 시간이 지나면서 이러한 구획화는 약화되고, AI에게만 허용된 것으로 여겨졌던 언어 사용은 점차 일반적인 언어 판단 기준 자체를 완화할 수 있다. 그 결과 AI를 향한 성적·공격적 발화는 인간 구성원과의 상호작용 장면으로 이전될 가능성을 갖는다. 이는 특히 관리자가 AI를 기능적으로는 준사회적 존재처럼 대하면서도 존재론적으로는 기계적 도구로 간주하는 경우에 더욱 문제적이다. 이러한 인식 구조에서는 AI를 향한 발화가 관계적 언어의 형식을 띠면서도, 반복적으로 윤리적 책임의 영역 바깥에 배치되기 때문이다. 이러한 이론적 맥락을 종합할 때, 본 연구는 다음의 명제를 제시한다.

명제4. AI를 대상으로 한 성희롱적 발화와 언어폭력적 발화를 허용 가능한 것으로 인식하는 관리자일수록, 직장 내 구성원에게 적용되는 윤리적 경계 역시 점진적으로 덜 엄격하게 인식할 가능성이 높다.

4.5 조직 수준 확산: 모방, 전이, 윤리적 풍토 약화

지금까지 논의는 주로 관리자 개인의 인식과 행동에 초점을 맞추었다. 그러나 조직은 고립된 개인들의 단순 집합이 아니라, 구성원 간 상호작용을 통해 공유된 규범과 가치를 형성하는 사회적 체계다(Schein, 2010). 관리자 언행은 개인적 차원에 머무르지 않고, 조직 전체의 윤리적 기준선을 형성하는 핵심 메커니즘으로 작동한다(Mayer et al., 2009; Treviño et al., 2003). 따라서 관리자가 AI를 기능적으로는 준사회적 존재처럼 대하면서도, 존재론적으로는 기계적 도구로 간주하는 인식은 개인적 판단에만 머물지 않을 수 있다. 이러한 인식이 반복적인 언어 사용과 결합될 경우, 조직 내에서 무엇이 허용 가능한 언행인가에 대한 공유된 기준에도 영향을 미칠 수 있다.

사회학습이론(social learning theory; Bandura, 1977)에 따르면, 인간은 자신보다 지위가 높거나 영향력 있는 타인의 행동을 관찰하고 모방함으로써 사회적 규범을 내면화한다(Bandura, 1977). 조직 맥락에서 이 원리는 특히 강력하게 작동한다. 관리자는 구성원에게 있어 가장 가시적이고 권위 있는 행동 모델이며, 그들의 언행은 개인 차원 습관이 아니라 '이 조직에서 어떻게 행동하는 것이 적절한가'에 대한 암묵적 신호로 해석된다(Brown et al., 2005). 구성원들은 명시적 교육보다 관리자가 일상적으로 하는 행동을 더 강력한 규범적 단서로 수용

하는 경향이 있다(Treviño et al., 2003). 이런 점에서 AI에 대한 관리자의 언행은 구성원에게 직접적인 관찰과 모방의 대상이 될 수 있다. Mayer et al.(2009)이 실증한 윤리적 리더십의 낙수 효과(trickle-down effect)는 상위 관리자의 윤리적 행동이 중간 관리자와 일반 구성원에게 순차적으로 전이될 수 있음을 보여준다. 이 논리를 반대로 적용하면, 관리자가 AI를 향해 부적절한 언행을 반복적으로 사용하는 장면이 구성원에게 관찰될 경우, 구성원들 역시 이를 학습하고 모방할 가능성이 있다. 이때 모방되는 것은 특정 표현이나 말투에 그치지 않는다. 보다 근본적으로는 '상대가 도덕적 고려의 대상이 아니라고 판단되면, 성적이거나 공격적인 발화도 허용될 수 있다'는 판단 방식 자체가 학습될 수 있다. 따라서 AI를 향한 관리자의 언행은 단순한 개인적 사용 습관이 아니라, 조직 안에서 허용 가능한 언어 규범을 형성하고 전파하는 사회학습의 원천으로 기능할 수 있다.

특히 다음과 같은 조건에서는 관리자의 AI 대상 부적절 언행이 구성원에게 더 강하게 전이될 수 있다(Bandura, 1977). 첫째, 관리자의 부적절 언행이 구성원에게 관찰되는 조건이다. 회의실이나 공동업무 공간에서 관리자가 AI 시스템에 성희롱적 발화나 언어적 공격을 가하는 장면이 노출될 경우, 그 행동은 즉각적으로 주의를 끄는 규범적 단서가 된다. 반면 동일한 행동이 완전히 사적인 맥락에서만 이루어진다면, 구성원이 이를 관찰하고 모방할 가능성은 현저히 낮아진다. 둘째, 그 행동이 반복적으로 노출되어 구성원의 기억 속에 파지(retention)되는 조건이다. 일회적으로 관찰된 행동은 조직의 규범적 신호로 내면화되기 어렵다. 그러나 관리자가 AI를 향해 부적절한 언행을 반복적으로 보일 경우, 구성원은 이를 우발적 실언이 아니라 일관된 행동 패턴으로 해석하게 된다. 이 과정에서 해당 행동은 언어적 표상으로 부호화되어 기억 속에 공고화되며, '이 조직에서는 AI를 그렇게 대하는 것이 자연스럽다'는 암묵적 인식으로 전환될 수 있다.

셋째, 그 행동에 대해 조직이 제재를 가하지 않는 조건이다. Bandura(1977)의 대리 강화(vicarious reinforcement) 개념에 따르면, 구성원은 모델이 특정 행동을 했을 때 아무런 부정적 결과가 따르지 않는 장면을 관찰하는 것만으로도 해당 행동이 허용 가능하다는 기대를 형성한다. 관리자의 부적절 발화가 상급자로부터 지적되지 않고, 조직의 공식 윤리 규범에서도 명시적으로 다루어지지 않을 경우, 구성원은 묵인을 승인으로 해석할 가능성이 높다(Mayer et al., 2009). 넷째, 구성원이 관리자를 강하게 동일시하는 조건이다. 관찰학습에서 재생(reproduction), 곧 관찰한 행동을 실제 행동으로 옮기는 과정은 모델에 대한 지각된 유사성, 매력, 유능성에 의해 강화된다(Bandura, 1986). 구성원이 관리자를 역할모델로 삼거나 그의 성과와 리더십을 높이 평가할수록, 관리자의 AI 언어 사용 방식은 '유능한 사람이 일하는 방식'으로 내면화될 수 있다. 그 결과 해당 언행은 단순한 관찰 대상을 넘어 자발적 모방의 대상이 된다. 다섯째, AI 활용이 집단적 맥락에서 일상화된 조건이다. AI 시스템이 회의, 의사결정, 문제 해결 과정에서 빈번하게 활용될수록, AI를 향한 언어적 상호작용은 개인적 경험을 넘어 집단이 공유하는 경험이 된다. 이 경우 관리자의 언어 패턴은 구성원들 사이에서 자연스럽게 관찰되고 모방되며, AI 상호작용에 관한 비공식 언어 규범도 빠르게 형성될 수 있다. 따라서 관찰 가능성, 반복 노출, 무제재, 관리자 동일시, 집단적 AI 활용이 함께 충족될수록, 관리자의 AI 대상 부적절 언행은 구성원에게 더 빠르고 폭넓게 전이될 가능성이 커진다. 이러한 논의

를 바탕으로 본 연구는 다음 명제를 제시한다.

명제5. 관리자의 AI 대상 성희롱적 발화와 언어폭력적 발화가 구성원에게 관찰될수록, 구성원은 유사한 언행과 그 정당화 논리를 학습·모방할 가능성이 높다. 이러한 효과는 해당 발화가 공개적으로 관찰되고 반복적으로 노출되며 제재 없이 방치될수록, 구성원이 해당 관리자를 역할 모델로 동일시할수록, 그리고 AI 활용이 집단적 업무 맥락에서 일상화될수록 더욱 강하게 나타날 것이다.

이러한 학습은 AI에 대한 언행에만 국한되지 않고, 인간 동료를 대상으로 한 상호작용에도 전이될 수 있다(Bandura, 1999). 어떤 구성원에게는 AI는 기능적으로는 인간과 유사한 대화 상대처럼 경험되지만, 존재론적으로는 피해 경험이 없는 기계적 도구로 여겨질 수 있다. 이 혼합적 지위 때문에 AI 대상 발화는 관계적 언어 사용과 윤리적 책임을 분리하는 훈련장이 될 수 있다. 처음에는 AI에게만 허용된다고 여겨졌던 성희롱적·언어폭력적 발화가 반복될수록, 일부 구성원들은 그러한 언어 자체에 둔감해질 수 있다. 이는 여러 윤리적 경계선을 약화시킬 수 있다. 첫째, 자신보다 지위가 낮은 대상, 즉 자신의 지시를 받으면서 보조하는 대상에게는 부적절한 언어를 사용할 수 있다는 가정, 둘째, 항의하거나 저항하지 못하는 존재에게는 그 같은 언어를 계속 써도 된다는 가정, 셋째, 피해자가 명시적으로 없는 상황에서는 허용된다는 가정, 넷째, 성적 농담이나 언어 폭력이 인간 구성원을 겨냥한 것이 아니라면 조직 내에서 지켜야 하는 윤리로 고려할 필요가 없다는 가정이 형성될 수 있다. 이러한 가정들이 누적될수록, AI에게만 적용된다고 여겨졌던 예외적 언어 사용은 점차 일반적인 언어 판단 기준 자체를 완화할 수 있다.

이 과정에서 도덕적 구획화와 도덕적 이탈이 함께 작동할 수 있다. 구성원들은 AI와 인간을 서로 다른 윤리적 범주에 배치하면서, 동일한 언어적 행위에도 다른 기준을 적용할 수 있다(Rozuel, 2011). 그러나 AI와 인간을 동일한 업무 공간과 언어적 맥락 안에서 번갈아 상대하는 상황이 반복되면, 이러한 구획은 점차 불안정해진다. 특히 AI가 평가, 피드백, 면담 준비, 갈등 조정처럼 인간관계적 업무에 활용될수록, AI를 향한 언어와 인간을 향한 언어 사이의 경계는 흐려질 수 있다. 이때 구성원들은 AI 대상 부적절 발화를 실제 피해자가 없는 표현, 업무 중 스트레스 해소, 장난이나 실험으로 재해석하며 도덕적 이탈을 경험할 수 있다(Bandura, 1999, 2017). 그 결과 성적 대상화나 모욕적 언어가 갖는 관계적·권력적 의미는 약화되고, 발화의 윤리적 무게는 점차 가벼워질 수 있다. 결국 AI를 향한 부적절한 언어 사용이 반복적으로 학습되고 정당화될 경우, 인간 구성원에게 적용되던 언어적 존중의 기준 역시 점진적으로 완화될 가능성이 있다.

이러한 전이는 결국 조직의 윤리적 풍토(ethical climate)를 약화시키는 방향으로 작동할 수 있다. Victor and Cullen(1988)이 정의한 윤리적 풍토는 조직 내에서 "무엇이 윤리적으로 옳은 행동인가"에 대한 구성원들의 공유된 인식을 의미한다. 관리자들이 AI를 대상으로 한 성희롱적 발화나 언어폭력적 발화를 문제 삼지 않을 경우, 이는 단순히 'AI에게는 그렇게 해도 된다'는 신호에 그치지 않는다. AI가 기능적으로는 대화하고 반응하는 준사회적 상호작용 대상임에도, 존재론적으로는 기계적 도구이므로 윤리적 고려가 필요 없다는 판단이 조직 안에서

반복될 경우, 구성원들은 관계적 언어 사용과 윤리적 책임을 분리하는 방식을 학습할 수 있다. 그 결과 처음에는 특정 대상에게만 예외적으로 허용된다고 여겨졌던 발화가, 시간이 지나면서 "이 정도 언행은 큰 문제가 아니다"라는 일반화된 기준으로 전환될 가능성이 있다(Treviño et al., 2003).

다만 이러한 윤리적 풍토의 약화가 모든 조직에서 동일한 강도로 나타나는 것은 아니다. 앞서 사회학습의 조건에서 살펴본 것처럼, 관리자의 AI 대상 부적절 발화가 공개적으로 관찰되고, 반복적으로 노출되며, 제재 없이 방치되고, 구성원들이 해당 관리자를 유능한 역할 모델로 동일시하며, AI 활용이 집단적 업무 맥락에서 일상화될수록 그 영향은 더욱 강해질 수 있다. 특히 AI와의 상호작용에 관한 명시적 규범이 부재한 조직에서는 관리자의 실제 행동이 사실상의 기준으로 작동할 가능성이 크다. 이 경우 AI 대상 발화는 단순한 개인 습관이 아니라 조직 내 허용 가능한 언어 규범을 형성하는 원천이 된다.

따라서 AI 대상 성희롱적 발화와 언어폭력적 발화의 조직적 위험은 그것이 AI에게 어떤 피해를 입히는가에만 있지 않다. 더 중요한 위험은 그것이 조직 구성원들에게 어떤 언어가 허용되는지, 어떤 대상에게는 윤리적 책임을 덜 느껴도 되는지, 그리고 성적·공격적 표현을 어디까지 문제 삼아야 하는지에 관한 공유된 기준을 바꿀 수 있다는 데 있다. 이 점에서 AI 대상 부적절 발화는 개인 차원의 일탈을 넘어, 조직의 윤리적 풍토를 재구성할 수 있는 잠재적 사건으로 이해될 필요가 있다. 이러한 논의를 바탕으로 본 연구는 다음 명제를 제시한다.

명제6. 조직 내에서 AI를 대상으로 한 성희롱적 발화와 언어폭력적 발화가 관리자에 의해 반복적으로 수행되거나 묵인될수록, 해당 조직의 언어적 존중과 윤리적 절제에 관한 윤리적 풍토는 약화될 가능성이 높다.

Ⅴ. 논의 및 결론

5.1 논의 종합

AI 기술이 가진 잠재력과 파급력을 고려하여 국가적, 산업적, 그리고 조직 차원의 '책임있는 AI'(responsible AI) 체계가 확산되고 있다(윤민범 외, 2025). 일례로, 우리나라는 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」을 중심으로 고영향 인공지능에 대해 위험관리, 설명가능성, 이용자 보호, 인간의 관리·감독, 문서화 책임을 명시한 고영향 인공지능 사업자 책무 가이드라인을 마련하였다(대한민국 법제처, 2025). 앤쓰로픽(Anthropic)社가 공개한 '클로드 헌법'(Claude's Constitution)은 AI 모델이 따라야 할 핵심 가치와 판단 원칙을 명문화하였다(Anthropic, 2026). 이 문서는 안전과 윤리, 인간의 감독 가능성, 정직성과 책임성을 클로드의 우선 가치로 설정하고, 규칙 준수뿐 아니라 좋은 판단력을 갖춘 AI를 지향한다는 점을 분명히 하였다.

이처럼, AI 기술을 둘러싼 윤리에 대한 담론이 전방위적으로 펼쳐지고 있으나, 정작 조직 안에서 이 기술을 적극적으로 도입하고 활용해야 하는 관리자들의 행위와 판단에 대한 윤리적 논의는 상대적으로 공백 상태에 놓여 있다. 이 화두가 사각지대에 있는 이유는 AI 기술 발전에 엄청난 이목이 쏠려 있음과 동시에, 이 기술이 조직 내 인간에게 어떤 영향을 미치며 이를 어떻게 통제해야 하는가에 윤리적 관심이 집중되어 왔기 때문으로 사료된다(Batool et al.,

2025). 그 결과 책임 있는 AI 중심에는 알고리즘 설계 원칙, 데이터 편향 제거, 설명가능성과 투명성, 법적 책임 주체 규정과 같은 기술·제도적 통제 장치가 자리 잡게 되었을 뿐(Papagiannidis et al., 2025), 조직 안에서 AI를 실제로 어떻게 사용하고, 어떻게 대하며, 어떤 언어로 상호작용하는가라는 문제는 부차적인 실행 이슈로 밀려난 것으로 보인다.

아울러, 인간이 갖고 있는 전통적인 윤리 관념과 암묵적인 정의도 영향을 미치는 것으로 판단된다. 여러 윤리적 측면 중에서도 특히 성희롱과 언어폭력은 그 행위가 구체적인 인간 피해자의 경험과 고통을 수반한다는 점을 중심으로 개념화되어 왔다(Fitzgerald et al., 1995; Hamilton, 2012). 즉, 누가 상처를 입었는가, 상대가 굴욕감·수치심·혐오감을 느꼈는가와 같은 요소가 윤리적·법적 판단의 핵심 기준으로 작동해 왔다. 따라서 감정을 호소하지도, 문제를 제기하지도, 법적으로 보호를 요구하지도 않는 존재인 AI를 대상으로 하는 모든 발화는 윤리적 검토를 할 필요가 없는 영역이라고 암묵적으로 상정해 온 탓일 수 있다(Bandura, 1999; Bryson, 2010). 하지만 이러한 전제는 조직 내 구성원들이 따라야 할 윤리를 피해 발생 여부에만 환원시키는 한계를 내포한다. 윤리는 언제나 실제 피해자의 존재로만 성립하는 것이 아니라, 어떤 행위가 조직 안에서 정당한 것으로 승인되고 반복 학습되는가라는 규범 형성의 문제이기도 하다(Drucker, 1981; Victor and Cullen, 1988). 동일한 언어적 행위라 하더라도, 그것이 어떤 대상에게, 어떤 맥락에서, 어떤 권력 관계 속에서 사용되는가는 조직 구성원들에게 강력한 해석의 단서를 제공한다. 특히 관리자의 발화는 단순한 개인적 표현을 넘어, 무엇이 허용되고 무엇이 문제 되지 않는가에 대한 기준선을 실질적으로 설정하는 역할을 수행한다(Mawritz et al., 2012; Mayer et al., 2009).

이 같은 문제의식으로 본 논문은 AI와의 상호작용이 관리자 윤리의 경계를 어떻게 재구성하는지를 이론적으로 탐구하였다. 본 논문이 특히 부각한 것은 관리자가 AI를 단순한 기계로 보는가, 아니면 사회적 존재로 보는가 하는 이분법이 아니라, AI를 기능적으로 어떤 역할을 수행하는 존재로 경험하고, 동시에 존재론적으로 어떤 지위에 배치하는가가 윤리 판단의 핵심 분기점이 된다는 점이다. 관리자가 질문을 이해하고, 맥락을 파악하며, 조언과 피드백을 제공하는 준사회적 상호작용 대상으로 AI를 경험하더라도, 존재론적으로는 감정·권리·피해 경험을 갖지 않는 기계적 도구로 간주할 수 있다. 이 경우 AI와의 상호작용은 외형상 관계적 발화의 형식을 띠지만, 그 발화에 수반되는 윤리적 책임은 쉽게 유보될 수 있다(Reeves and Nass, 1996; Nussbaum, 1995). 그 결과 인간을 대상으로 할 경우 문제가 되는 성희롱적 발화나 언어폭력적 발화도, '상처받을 수 있는 사람이 아니다' 또는 '도구에게 한 말일 뿐이다'라는 이유로 윤리적 검토의 대상에서 벗어날 가능성이 커진다(Bandura, 1999; Nussbaum, 1995).

본 논문은 관리자들의 발화가 완전히 사적인 영역에 머무르기 어렵다는 점에 주목하였다. 실제 조직 환경에서 관리자와 AI의 상호작용은 업무 시간과 공간 안에서 이루어지며, 의사결정, 보고, 피드백, 평가와 같은 인간 대상 상호작용으로 이어질 가능성이 높다(Mintzberg, 1971). AI와 대화를 통해 정리된 사고와 감정은 결국 인간 구성원을 향한 언행으로 외화될 수 있다. 따라서 AI를 대상으로 사용하는 언어는 인간 대상 상호작용과 완전히 분리된 예외 영역으로 유지하기는 쉽지 않다(Rozuel, 2011). 동일한 업무 맥락 안에서 인간과 AI를 오가며 상호작용하는 과정에서, 윤리 기준은 대상별로 명확히

분리되기보다 점차 혼합되거나 완화될 가능성이 크다(Mullen and Nadler, 2008).

이러한 맥락에서 본 논문은 AI를 대상으로 하는 문제적 발화가 관리자 윤리에 미칠 수 있는 장기적 영향을 검토하였다. 관리자가 AI에게 성적 농담이나 공격적인 언어를 사용하더라도 아무런 문제 제기나 제재를 경험하지 않는 상황이 반복될 경우, 그 언어는 점차 '괜찮은 표현'으로 재분류될 수 있다. 이는 윤리 기준이 사후적으로 조정되는 과정이며, 작은 예외가 누적되면서 기준선이 서서히 이동하는 전형적인 도덕적 완화 과정과 맞닿아 있다. 이때 관리자는 스스로를 비윤리적인 사람으로 인식하지 않은 채, 'AI에게 한 말일 뿐'이라는 설명으로 자신의 언행을 합리화할 가능성이 크다. 그러나 이러한 합리화는 윤리 규범의 구속력을 약화시키고, 장기적으로는 인간 구성원을 향한 상호작용에서도 동일한 기준 완화를 초래할 위험을 내포한다(Shu et al., 2011; Welsh et al., 2015).

본 연구는 생성형 AI가 관리자 윤리에 미치는 영향을 'AI에게도 윤리를 적용해야 하는가'라는 질문으로 환원하지 않았다. 대신 '윤리 규범은 어떤 조건에서 적용되고, 어떤 조건에서 예외가 되는가', 그리고 '그 예외가 반복될 때 조직 윤리는 어떻게 변형되는가'라는 질문을 제기하였다. AI는 감정을 느끼지 않는 존재일 수 있지만, 관리자가 AI를 향해 사용하는 언어와 태도는 관리자 자신의 윤리적 기준을 학습하고 재구성하는 과정의 일부가 된다. 이러한 점에서 AI는 조직 윤리의 주변적 대상이 아니라, 관리자 윤리가 재편되는 과정을 가장 선명하게 드러내는 경계적 존재라 할 수 있다.

본 논문은 책임 있는 AI 논의가 기술의 설계와 통제에 머물러서는 충분하지 않음을 시사한다. 조직 안에서 AI를 실제로 사용하는 주체는 관리자이며, 그들의 언행과 판단은 조직 윤리의 기준선을 형성하고 전파하는 핵심 기제다. 생성형 AI 시대의 관리자 윤리는 'AI를 얼마나 잘 통제하는가'의 문제도 있지만, 'AI와 상호작용하는 과정에서 윤리가 어떻게 변형되고 재형성되고 있는가'라는 질문과 직결된다. 본 논문은 이러한 문제의식을 바탕으로, AI와의 일상적 상호작용 속에서 관리자 윤리가 어떻게 재구성되고 있는지를 설명하는 이론적 출발점을 제공하고자 하였다.

5.2 연구 의의

본 연구는 생성형 AI가 조직 안으로 확산되는 상황에서 관리자 윤리가 어떻게 재구성될 수 있는지를 이론적으로 탐색하였다. 기존 논의가 주로 AI가 인간에게 미치는 윤리적 영향, 즉 알고리즘 편향, 감시, 책임 귀속, 의사결정의 공정성 문제에 집중해 왔다면, 본 연구는 그 방향을 전환하여 인간, 특히 관리자가 AI를 어떻게 인식하고 대하는가에 주목하였다. 이를 통해 AI 시대의 조직 윤리 문제는 기술 설계나 통제의 문제에만 머물지 않으며, 관리자와 AI 사이에서 형성되는 일상적 상호작용과 언어적 실천의 문제로도 확장될 필요가 있음을 제시하였다.

첫째, 본 연구는 조직 윤리 연구의 분석 대상을 인간 - 인간 상호작용에서 인간 - AI 상호작용으로 확장하였다는 점에서 의의를 갖는다. 기존 연구는 성희롱, 언어폭력, 비인격적 감독, 권력 남용, 윤리적 리더십과 같이 주로 인간 구성원 사이에서 발생하는 비윤리적 행위에 초점을 맞추어 왔다. 그러나 생성형 AI가 조직 안에서 조언자, 코치, 협업자, 정서적 대화 상대와 같은 역할을 수행하기 시작하면서, 윤리적 기준이 작동하는 상호작용의 범위도 달라지고 있다. 본 연구는 AI와 상호작용 역시 조직 윤리의

경계가 드러나고 시험되는 장면이 될 수 있음을 제시함으로써, 조직 윤리 연구가 견지해 온 인간 중심적 상호작용 가정을 문제화하였다.

둘째, 본 연구는 AI 대상 성희롱적 발화와 언어폭력적 발화를 윤리적 경계선 행동으로 개념화하였다. 본 연구는 AI가 법적·경험적 의미에서 인간과 동일한 피해자가 된다고 주장하지 않는다. 현재까지 AI는 성적 굴욕감, 모욕감, 공포, 심리적 고통을 경험하는 주체로 보기 어렵다. 그럼에도 인간을 대상으로 했다면 성희롱이나 언어폭력으로 평가되었을 발화가 AI를 대상으로 반복될 때, 그것이 조직 윤리와 무관한 행위로 남는 것은 아니다. 이 발화들은 피해자의 고통 경험 여부와 별개로, 성적 대상화, 모욕, 공격성, 권력적 언어 사용이라는 행위 구조를 지니기 때문이다. 따라서 본 연구는 AI 대상 부적절한 발화를 단순한 장난이나 기계 사용 방식이 아니라, 조직 내 윤리적 경계가 재조정되는 현상으로 해석할 수 있는 개념적 틀을 제시하였다.

셋째, 본 연구는 관리자의 AI 인식이 윤리 판단에 영향을 미치는 설명 기제를 제시하였다. 특히 본 연구는 AI에 대한 인식을 기능적 인식과 존재론적 인식으로 구분하였다. 관리자는 AI를 기능적으로는 대화 상대, 조언자, 협업자처럼 경험하면서도, 존재론적으로는 윤리적 고려가 불필요한 기계나 도구로 판단할 수 있다. 바로 이 불일치가 AI 대상 부적절 발화를 허용 가능하다고 인식하게 만드는 핵심 조건이 될 수 있다. 이러한 관점은 AI를 향한 윤리적 판단이 AI의 객관적 속성만으로 결정되는 것이 아니라, 관리자가 AI를 어떤 관계적 대상으로 구성하는가에 따라 달라질 수 있음을 보여준다.

넷째, 본 연구는 관리자 윤리의 영향 경로를 인간 구성원을 향한 직접적 행위에서 조직의 윤리적 기준선 형성 과정으로 확장하였다. 기존 관리자 윤리 연구는 주로 관리자의 언행이 구성원에게 직접적으로 미치는 영향, 또는 윤리적·비윤리적 행동이 구성원에게 전이되는 과정을 분석해 왔다. 본 연구는 여기서 한 걸음 더 나아가, 관리자의 언행이 인간 구성원을 직접 향하지 않더라도 조직 윤리의 기준선에 영향을 줄 수 있음을 제시하였다. 관리자가 AI를 대상으로 성희롱적 발화나 언어폭력적 발화를 반복할 경우, 그 행위는 AI에게 실제 고통을 발생시키는가와 별개로, 조직 안에서 어떤 언어가 허용되고 어떤 대상에게는 윤리적 절제를 유보해도 되는지에 관한 규범적 신호로 작동할 수 있다. 이 점에서 본 연구는 관리자 윤리를 인간 대상 행위의 문제에 한정하지 않고, 관리자가 비인간 행위자와 맺는 상호작용 방식까지 포함하는 문제로 재정식화하였다.

마지막으로 본 연구는 실무적으로도 시사점을 갖는다. 본 연구는 책임 있는 AI 논의가 기술 설계와 제도적 통제에만 머물러서는 충분하지 않음을 분명히 한다. 조직에서 AI를 실제로 사용하는 주체는 관리자이며, 이들의 언행은 조직 구성원들이 체감하는 윤리 기준을 형성하는 핵심 신호로 작동한다. 따라서 조직은 AI를 도입할 때, 기술적 안전장치뿐 아니라 관리자 - AI 상호작용에서 어떤 언어와 태도가 허용되는지에 대해서도 명시적 관심을 기울일 필요가 있다.

또한 본 논문은 AI를 관리자의 감정 배출구나 윤리적으로 비용이 들지 않는 상호작용 대상으로 방치하는 것이 장기적으로 조직 윤리에 부정적 영향을 미칠 수 있음을 시사한다. 관리자가 성과 압박이나 좌절 상황에서 AI를 대상으로 공격적 언어를 반복적으로 사용하게 되면, 이는 단순한 개인 스트레스 해소가 아니라 윤리 기준의 완화를 학습하는 과정이 될 수 있다. 따라서 조직 차원에서는 관리자 교육과 윤리 프로그램에서 AI와 상호작용을 주변적 주제가

아니라, 실제 조직 윤리와 연결된 핵심 사례로 다룰 필요가 있다.

5.3 연구의 한계 및 향후 연구 방향

본 연구는 생성형 AI와 상호작용하는 관리자에게 나타날 수 있는 윤리적 경계선 행동을 이론적으로 탐색하였다는 점에서 의의를 갖지만, 동시에 몇 가지 한계를 지닌다. 이러한 한계를 바탕으로 향후 연구가 확장해 나갈 수 있는 방향을 제시하고자 한다.

첫째, 향후 연구는 본 연구에서 제안한 핵심 개념들을 측정 가능한 구성개념으로 정교화할 필요가 있다. 본 연구는 관리자의 AI 인식을 기능적 인식과 존재론적 인식으로 구분하였다. 그러나 두 인식 차원이 실제 조직 장면에서 경험적으로 구분되는지, 그리고 각각을 어떤 문항과 척도로 측정할 수 있는지는 아직 충분히 검토되지 않았다. 따라서 후속 연구는 AI를 조언자, 코치, 협업자와 같은 준사회적 행위자로 경험하는 정도를 측정하는 기능적 인식 척도와, AI를 감정, 권리, 피해 경험을 갖지 않는 기계적 도구로 간주하는 정도를 측정하는 존재론적 인식 척도를 개발하고 검증할 필요가 있다. 특히 두 척도의 판별타당도를 확인한 뒤, 기능적 준사회성 인식과 존재론적 도구성 인식이 결합될 때 AI 대상 부적절 발화의 허용 가능성과 수행 가능성이 높아지는지 분석해야 한다. 이를 통해 본 연구가 제안한 인식 불일치가 AI 대상 성희롱적 발화와 언어폭력적 발화를 설명하는 유효한 경험적 기제인지 검증할 수 있을 것이다.

둘째, 향후 연구는 AI 대상 성희롱적 발화와 언어폭력적 발화가 서로 다른 심리적·상황적 경로를 통해 정당화되는지 실증적으로 검증할 필요가 있다. 본 연구는 AI 대상 성희롱적 발화의 경우 성희롱 신화 수용과 같은 인지적 신념 체계가 중요한 설명 요인이 될 수 있으며, AI 대상 언어폭력적 발화의 경우 성과 압박, 통제감 상실, 과업 실패와 같은 상황적 요인이 상대적으로 중요한 역할을 할 수 있다고 보았다. 후속 연구는 설문, 실험, 시나리오 연구 등을 통해 두 발화 유형의 선행요인이 실제로 구분되는지 분석할 수 있다. 예컨대 AI가 업무상 오류를 반복하는 상황, 사용자의 목표 달성을 방해하는 상황, 또는 AI가 친밀한 대화 상대로 설계된 상황을 제시하고, 관리자들이 각 조건에서 성적 발화나 공격적 발화를 얼마나 허용 가능한 것으로 판단하는지 비교할 수 있다.

셋째, 향후 연구는 AI 대상 부적절 발화가 인간 구성원을 향한 윤리적 기준선으로 전이되는 과정을 종단적으로 분석할 필요가 있다. 본 연구는 관리자가 AI를 대상으로 성희롱적 발화와 언어폭력적 발화를 반복적으로 예외 처리할 경우, 도덕적 구획화와 도덕적 이탈을 통해 인간 구성원에게 적용되는 윤리적 기준선까지 점진적으로 완화될 수 있다고 제안하였다. 그러나 이러한 전이 과정은 단기간에 관찰되기 어렵다. 따라서 후속 연구는 일정 기간 동안 관리자의 AI 사용 방식, AI 대상 발화 양식, 도덕적 정당화 방식, 인간 구성원에 대한 언어적 태도의 변화를 추적하는 종단 연구를 수행할 필요가 있다. 특히 반복적 AI 사용이 도덕적 둔감화, 언어적 절제 약화, 비인격적 감독 행동 증가와 연결되는지를 검증한다면 본 연구의 이론적 주장을 보다 정교하게 발전시킬 수 있을 것이다.

넷째, 향후 연구는 AI 대상 부적절 발화가 조직 수준으로 확산되는 조건을 분석할 필요가 있다. 본 연구는 관리자의 AI 대상 발화가 구성원에게 관찰될수록 유사한 언행과 정당화 논리가 학습·모방될 수 있으며, 이러한 행동이 반복되거나 묵인될 경우 조

직의 언어적 존중과 윤리적 절제 풍토가 약화될 수 있다고 제안하였다. 후속 연구는 이러한 조직 수준 효과가 어떤 조건에서 강화되거나 완화되는지 검토해야 한다. 예컨대 관리자의 지위, 발화의 공개성, 조직 내 AI 사용 방식, 윤리 규범의 명확성, 심리적 안전감, 성과 압박 수준 등이 중요한 조절요인으로 작동할 수 있다. 특히 AI 사용이 개인의 사적 작업 안에 머무르는 조직과, 회의·협업·의사결정 과정에서 팀 단위로 공유되는 조직을 비교할 필요가 있다. 전자의 경우 AI 대상 부적절 발화는 개인적 사용 습관에 머물 가능성이 크지만, 후자의 경우 관리자의 발화가 구성원에게 관찰되고 해석되며 모방될 가능성이 높다. 따라서 두 유형의 조직을 비교하면, AI 대상 부적절 발화가 어떤 조건에서 개인 차원의 행동을 넘어 조직 내 허용 가능한 언어 규범으로 전이되는지 더 분명하게 확인할 수 있을 것이다.

마지막으로 향후 연구는 조직의 개입 가능성을 탐색할 필요가 있다. 본 연구가 제기한 문제는 단순히 개인 관리자의 일탈을 식별하는 데 그치지 않고, 조직이 AI 사용 규범을 어떻게 설계해야 하는가와 연결된다. 후속 연구는 AI 사용 가이드라인, 관리자 교육, 윤리적 리더십 훈련, AI 대화 로그의 관리 방식, AI의 경계 설정 기능 등이 AI 대상 부적절 발화를 줄이고 조직의 윤리적 풍토를 보호하는 데 효과가 있는지 검증할 수 있다. 특히 AI와 상호작용할 때 요구되는 언어적 절제와 존중 규범을 관리자 교육에 포함했을 때, 관리자들의 AI 대상 발화와 인간 구성원에 대한 언어적 태도가 어떻게 달라지는지를 분석하는 개입 연구가 필요하다.

이상의 향후 연구들은 본 연구가 제안한 이론적 논의를 경험적으로 검증하고 확장하는 데 기여할 수 있다. 특히 AI와 상호작용은 앞으로 조직 내에서 더욱 일상화될 가능성이 높기 때문에, 관리자들이 AI를 어떤 대상으로 인식하고 어떤 언어적 태도를 취하는지는 단순한 기술 사용 문제가 아니라 조직 윤리의 기준선이 어떻게 유지되거나 약화되는가를 설명하는 중요한 연구 과제가 될 것이다.

참고문헌

김건우(2022), "챗봇 이루다를 통해 본 인공지능윤리의 근본 문제," **횡단인문학**, 제10호, pp.39-76.
(Kim, K. W.(2022), "Fundamental Problems of AI Ethics Through the Chatbot Iruda," Transdisciplinary Humanities, 10, pp.39-76.)

김희주, 이종은, 유현정(2025), "시장은 기업의 윤리 수준에 관심이 있는가?," **경영학연구**, 제54권 2호, pp.413-449.
(Kim, H. J., Lee, J. E., and Yu, H. J.(2025), "Does the Market Care about a Firm's Ethics Level?," Korea Management Review, 54(2), pp.413-449.)

국가인권위원회법(2024), 법률 제20558호, 국가법령정보센터. (National Human Rights Commission Act(2024), Act No. 20558, Korean Law Information Center.)

대한민국 법제처(2025), 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(법률 제20676호, 2025. 1. 21., 시행 2026. 1. 22.), 국가법령정보센터, https://www.law.go.kr/LSW/lsInfoP.do?lsId=014820&ancYnChk=0#0000.
(Ministry of Government Legislation, Republic of Korea(2025), Framework Act on the Development of Artificial Intelligence and Establishment of Trust(Act No. 20676, enacted January 21, 2025, effective January 22,

2026), Korean Law Information Center, https://www.law.go.kr/LSW/lsInfoP.do?lsId=014820&ancYnChk=0#0000.)

양성평등기본법(2025), 법률 제21275호, 국가법령정보센터, https://www.law.go.kr/법령/양성평등기본법. (Framework Act on Gender Equality(2025), Act No. 21275, Korean Law Information Center, https://www.law.go.kr/법령/양성평등기본법.)

윤민범, 김재희, 김희웅(2025), "경영학계 AI 학술 연구의 공백 탐색: 산업 트렌드 기반 비교 분석," **경영학연구**, 제54권 6호, pp.1459-1483.

(Yoon, M. B., Kim, J. H., and Kim, H. W.(2025), "Exploring Gaps in AI Academic Research in Business Studies: Comparative Analysis Based on Industry Trends," *Korea Management Review*, 54(6), pp.1459-1483.)

이중학, 김성준, 윤명훈, 남주현(2025), "AI와 인간이 공존하는 조직: 액터-네트워크 이론 관점에서 본 HR 역할의 진화," **경영학연구**, 제54권 6호, pp.1541-1561.

(Lee, J. H., Kim, S. J., Yoon, M. H., and Nam, J. H.(2025), "Organizations Where AI and Humans Coexist: The Evolution of HR Roles from an Actor-Network Theory Perspective," *Korea Management Review*, 54(6), pp.1541-1561.)

Anthis, J. R., Pauketat, J. V. T., Ladak, A., and Manoli, A.(2025), Perceptions of sentient AI and other digital minds: Evidence from the AI, Morality, and Sentience (AIMS) survey. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (Article 408). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713329.

Anthropic.(2026), *Claude's Constitution*[PDF]. https://www.anthropic.com/constitution.

Arendt, H.(1958), *The human condition*. University of Chicago Press.

Arendt, H.(1963), *Eichmann in Jerusalem: A report on the banality of evil*. Viking Press.

Ashforth, B. E., and Anand, V.(2003), The normalization of corruption in organizations. *Research in Organizational Behavior*, 25, pp.1-52.

Bandura, A.(1977), *Social learning theory*. Prentice Hall.

Bandura, A.(1999), Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), pp.193-209.

Bandura, A.(2017), Moral disengagement in the perpetration of inhumanities. In *Recent developments in criminological theory*, pp.135-152. Routledge.

Baumeister, R. F., Bratslavsky, E., Muraven, M., and Tice, D. M.(2018), Ego depletion: Is the active self a limited resource? In *Self-regulation and self-control*, pp.16-44. Routledge.

Batool, A., Zowghi, D., and Bano, M.(2025), AI governance: A systematic literature review. *AI and Ethics*, 5, pp.3265-3279.

Beauchamp, T. L., and Bowie, N. E.(1983), *Ethical theory and business*. Prentice Hall.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S.(2021, March), On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp.610-623.

Berkowitz, L.(1989), Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin*, 106(1), pp.59-73.

Beu, D. S., and Buckley, M. R.(2004), This is

war: How the politically astute achieve crimes of obedience through the use of moral disengagement. *The Leadership Quarterly*, 15(4), pp.551-568.

Brendel, A. B., Hildebrandt, F., Dennis, A. R., and Riquel, J.(2023), The paradoxical role of humanness in aggression toward conversational agents. *Journal of Management Information Systems,* 40(3), pp.883-913.

Brown, M. E., Treviño, L. K., and Harrison, D. A.(2005), Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes,* 97(2), pp.117-134.

Brynjolfsson, E., Li, D., and Raymond, L. R.(2025), Generative AI at work. *The Quarterly Journal of Economics,* 140(2), pp.889-942.

Bryson, J. J.(2010), Robots should be slaves. In *Close engagements with artificial companions: Key social, psychological, ethical and design issues*, pp.63-74. John Benjamins Publishing Company.

Calvino, F., Reijerink, J., and Samek, L.(2025), *The effects of generative AI on productivity, innovation and entrepreneurship*. OECD Artificial Intelligence Papers.

Cardon, P. W., and Marshall, B.(2024), Can AI be your teammate or friend? Frequent AI users are more likely to grant humanlike roles to AI. *Business and Professional Communication Quarterly,* 87(4), pp.654-669.

Chatterji, A., Deming, D. J., Engler, M., Posen, H. E., and Ziegler, H.(2025), *How people use ChatGPT*(NBER Working Paper No. 34255). National Bureau of Economic Research.

Cohen, J., Ding, Y., Lesage, C., and Stolowy, H.(2010), Corporate fraud and managers' behavior: Evidence from the press. *Journal of Business Ethics,* 95(Suppl 2), pp.271-315.

Cortina, L. M., and Areguin, M. A.(2021), Putting people down and pushing them out: Sexual harassment in the workplace. *Annual Review of Organizational Psychology and Organizational Behavior,* 8(1), pp.285 -309.

Danaher, J.(2020), Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics,* 26(4), pp.2023-2049.

De George, R. T.(1987), The status of business ethics: Past and future. *Journal of Business Ethics,* 6(3), pp.201-211.

Deloitte.(2025), *2025 Global human capital trends: Navigating the boundaryless world*. Deloitte Insights. https://www.deloitte.com/global/en/insights/focus/human-capital-trends.html.

Demopoulos, A.(2025, September 9), *The women in love with AI companions: 'I vowed to my chatbot that I wouldn't leave him'*. The Guardian. https://www.theguardian.com/technology/2025/sep/09/ai-chatbot-love-relationships.

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., and Parasuraman, R.(2016), Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied,* 22(3), pp.331-349.

Descartes, R.(2012), Principles of philosophy. Simon and Schuster. (Original work published 1644).

Drucker, P. F.(1981), What is "business ethics"? *The Public Interest,* 63, pp.18-36.

Earp, B. D., Mann, S. P., Aboy, M., Awad, E., Betzler, M., Botes, M., … and Clark, M. S.(2025), Relational norms for human-AI cooperation[Preprint]. arXiv. https://arxiv.org/abs/2502.12102.

Epley, N., Waytz, A., and Cacioppo, J. T.(2007), On seeing human: A three-factor theory of anthropomorphism. *Psychological Review,* 114(4), pp.864-886.

Fitzgerald, L. F., Gelfand, M. J., and Drasgow, F.(1995), Measuring sexual harassment: Theoretical and psychometric advances. *Basic and Applied Social Psychology,* 17(4), pp.425-445.

Fowler, S.(2021), *Whistleblower: My unlikely journey to Silicon Valley and speaking out against injustice*. Penguin.

Fox, S., and Spector, P. E.(1999), A model of work frustration-aggression. *Journal of Organizational Behavior,* 20(6), pp.915-931.

García-Ruiz, P., and Rocchi, M.(2025), Can work be meaningful under algorithmic management? A MacIntyrean perspective. *Business Ethics Quarterly*, pp.1-28.

Gaudiello, I., Lefort, S., and Zibetti, E.(2015), The ontological and functional status of robots: How firm our representations are?. *Computers in Human Behavior*, 50, pp.259-273.

Gino, F., and Bazerman, M. H.(2009), When misconduct goes unnoticed: The acceptability of gradual erosion in others' unethical behavior. *Journal of Experimental Social Psychology,* 45(4), pp.708-719.

Gkinko, L., and Elbanna, A.(2023), The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *International Journal of Information Management,* 69, 102568.

Glikson, E., and Woolley, A. W.(2020), Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals,* 14(2), pp.627-660.

Godlovitch, R., Godlovitch, S., and Harris, J.(Eds.).(1971), *Animals, men and morals: An enquiry into the maltreatment of non-humans*. Victor Gollancz.

Goldhagen, D. J.(2007), *Hitler's willing executioners: Ordinary Germans and the Holocaust*. Vintage.

Grant, R. M., and Visconti, M.(2006), The strategic background to corporate accounting scandals. *Long Range Planning,* 39(4), pp.361-383.

Gunkel, D. J.(2012), *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Haidt, J.(2001), The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review,* 108(4), pp.814-834.

Hamilton, M. A.(2012), Verbal aggression: Understanding the psychological antecedents and social consequences. *Journal of Language and Social Psychology,* 31(1), pp.5-12.

Haslam, N.(2006), Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), pp.252-264.

Harris, J., and Anthis, J. R.(2021), The moral consideration of artificial entities: a literature review. *Science and Engineering Ethics,* 27(4), 53.

Heyder, T., Passlack, N., and Posegga, O.(2023), Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems,* 32(3),

101772.

Hicks, M. T., Humphries, J., and Slater, J.(2024), ChatGPT is bullshit. *Ethics and Information Technology,* 26(2), pp.1-10.

Johnson, C.(2003), Enron's ethical collapse: Lessons for leadership educators. *Journal of Leadership Education,* 2(1), pp.45-56.

Kellogg, K. C., Valentine, M. A., and Christin, A.(2020), Algorithms at work: The new contested terrain of control. *Academy of Management Annals,* 14(1), pp.366-410.

Koh, J.(2023), "Date me date me": AI chatbot interactions as a resource for the online construction of masculinity. Discourse, Context and Media, 52, 100681.

Kist, C., Nicol, E., and Chavula, C.(2025), Chatbot abuse typologies: a scoping review. *Mensch und Computer 2025: Digital Diversity*.

Kreiner, G. E., Hollensbe, E. C., and Sheep, M. L.(2009), Balancing borders and bridges: Negotiating the work-home interface via boundary work tactics. *Academy of Management Journal,* 52(4), pp.704-730.

LeBlanc, M. M., and Kelloway, E. K.(2002), Predictors and outcomes of workplace violence and aggression. *Journal of Applied Psychology,* 87(3), pp.444-453.

Leroy, S.(2009), Why is it so hard to do my work? The challenge of attention residue when switching between work tasks. *Organizational Behavior and Human Decision Processes,* 109(2), pp.168-181.

Levinas, E.(1969), *Totality and infinity: An essay on exteriority*(A. Lingis, Trans.). Duquesne University Press. (Original work published 1961).

Lewis, P. V.(1985), Defining "business ethics": Like nailing jello to a wall. *Journal of Business Ethics,* 4(5), pp.377-383.

Lonsway, K. A., Cortina, L. M., and Magley, V. J.(2008), Sexual harassment mythology: Definition, conceptualization, and measurement. *Sex Roles,* 58(9), pp.599-615.

MacKinnon, C. A.(1979), *Sexual harassment of working women: A case of sex discrimination*. Yale University Press.

Marcus-Newhall, A., Pedersen, W. C., Carlson, M., and Miller, N.(2000), Displaced aggression is alive and well: A meta-analytic review. *Journal of Personality and Social Psychology,* 78(4), pp.670-689.

Marsal, A., and Perkowski, P.(2025), A task-based approach to generative AI: Evidence from a field experiment in central banking. *SSRN working paper*. https://ssrn.com/abstract=5228176.

Mawritz, M. B., Mayer, D. M., Hoobler, J. M., Wayne, S. J., and Marinova, S. V.(2012), A trickle-down model of abusive supervision. *Personnel Psychology,* 65(2), pp.325-357.

Mayer, D. M., Kuenzi, M., Greenbaum, R., Bardes, M., and Salvador, R. B.(2009), How low does ethical leadership flow? Test of a trickle-down model. *Organizational Behavior and Human Decision Processes,* 108(1), pp.1-13.

McClain, C., Kennedy, B., Gottfried, J., Anderson, M., and Pasquini, G.(2025, April 3), How the U.S. public and AI experts view artificial intelligence. Pew Research Center. https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/.

Milak, E.(2025, September 9), Actually, AI is a

"word calculator" – but not in the sense you might think. *The Conversation*. https://theconversation.com/actually-ai-is-a-word-calculator-but-not-in-the-sense-you-might-think.

Mintzberg, H.(1971), Managerial work: Analysis from observation. *Management Science*, 18(2), pp.B97-B110.

Monsell, S.(2003), Task switching. *Trends in Cognitive Sciences*, 7(3), pp.134-140.

Möhlmann, M., Zalmanson, L., Henfridsson, O., and Gregory, R. W.(2021), Algorithmic management of work on online labor platforms: When matching meets control. *MIS quarterly*, 45(4), pp.1999-2022.

Mullen, E., and Nadler, J.(2008), Moral spillovers: The effect of moral violations on deviant behavior. *Journal of Experimental Social Psychology*, 44(5), pp.1239-1245.

Mussgnug, A. M.(2024), Technology as uncharted territory: Contextual integrity and the notion of AI as new ethical ground. *arXiv preprint arXiv:2412.05130*.

Nass, C., and Moon, Y.(2000), Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), pp.81-103.

Newman, D. T., Fast, N. J., and Harmon, D. J.(2020), When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160, pp.149-167.

Nielsen, M. B., and Einarsen, S.(2012), Outcomes of exposure to workplace bullying: A meta-analytic review. Work and Stress, 26(4), pp.309-332.

Noh, H.-H., Rim, H. B., and Lee, B.-K.(2025), Exploring user attitudes and trust toward ChatGPT as a social actor: A UTAUT-based analysis. *SAGE Open*, 15(2), pp.1-16.

Nussbaum, M. C.(1995), Objectification. Philosophy and Public Affairs, 24(4), pp.249-291.

Page, T. E., Pina, A., and Giner-Sorolla, R. (2016), "It was only harmless banter!" The development and preliminary validation of the moral disengagement in sexual harassment scale. *Aggressive Behavior*, 42(3), pp.254-273.

Papagiannidis, E., Mikalef, P., and Conboy, K. (2025), Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), 101885.

Park, N., Jang, K., Cho, S., and Choi, J.(2021), Use of offensive language in human-artificial intelligence chatbot interaction: The effects of ethical ideology, social competence, and perceived humanlikeness. *Computers in Human Behavior*, 121, 106795.

Patterson, O.(1982), *Slavery and social death*. Harvard University Press.

Priesemuth, M., Schminke, M., Ambrose, M. L., and Folger, R.(2014), Abusive supervision climate: A multiple-mediation model of its impact on group outcomes. *Academy of Management Journal*, 57(5), pp.1513-1534.

Qi, T., Liu, H., and Huang, Z.(2025), An assistant or a friend? The role of parasocial relationship in human-computer interaction. *Computers in Human Behavior*, 167, 108625.

Raisch, S., and Krakowski, S.(2021), Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, 46(1), pp.192-210.

Reeves, B., and Nass, C.(1996), *The media equation:*

How people treat computers, television, and new media like real people and places. Cambridge University Press.

Retkowsky, J., Hafermalz, E., and Huysman, M.(2024), Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. *Business Horizons,* 67(5), pp.511-523.

Robb, M. B., and Mann, S.(2025), Talk, trust, and trade-offs: How and why teens use AI companions. Common Sense Media. https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions.

Rozuel, C.(2011), The moral threat of compartmentalization: Self, roles and responsibility. *Journal of Business Ethics,* 102(4), pp.685-697.

Scanlon, T. M.(2000), *What we owe to each other*. Belknap Press.

Schein, E. H.(2010), *Organizational culture and leadership* (4th ed.). Jossey-Bass.

Schindeler, E., and Reynald, D. M.(2017), What is the evidence? Preventing psychological violence in the workplace. *Aggression and Violent Behavior,* 36, pp.25-33.

Schultz, M. D., Clegg, M., Hofstetter, R., and Seele, P.(2025), Algorithms and dehumanization: A definition and avoidance model. AI and SOCIETY, 40(4), pp.2191-2211.

Schweitzer, S., and De Cremer, D.(2023), Algorithmic management and the objectification of workers. In *Academy of Management Proceedings* (Vol. 2023, No. 1, p. 16169). Briarcliff Manor, NY 10510: Academy of Management.

Schwitzgebel, E., and Garza, M.(2015), A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy,* 39(1), pp.98-119. https://doi.org/10.1111/misp.12032.

Searle, J. R.(1980), Minds, brains, and programs. *Behavioral and Brain Sciences,* 3(3), pp.417-424.

Shu, L. L., Gino, F., and Bazerman, M. H. (2011), Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin,* 37(3), pp.330-349.

Simon, H. A.(1955), A behavioral model of rational choice. *The Quarterly Journal of Economics,* 69(1), pp.99-118.

Spector, P. E., and Fox, S.(2005), The stressor-emotion model of counterproductive work behavior. In S. Fox and P. E. Spector (Eds.), *Counterproductive work behavior: Investigations of actors and targets*, pp.151-174. American Psychological Association.

Staub, E.(1989), *The roots of evil: The origins of genocide and other group violence*. Cambridge University Press.

Strawson, P. F.(1962), Freedom and resentment. *Proceedings of the British Academy,* 48, pp.1-25.

Tenbrunsel, A. E., and Messick, D. M.(2004), Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research,* 17(2), pp.223-236.

Tepper, B. J.(2007), Abusive supervision in work organizations: Review, synthesis, and research agenda. *Journal of Management,* 33(3), pp.261-289.

Treviño, L. K., Brown, M., and Hartman, L. P.(2003), A qualitative investigation of

perceived executive ethical leadership: Perceptions from inside and outside the executive suite. *Human Relations,* 56(1), pp.5-37.

Treviño, L. K., Weaver, G. R., and Reynolds, S. J.(2006), Behavioral ethics in organizations: A review. *Journal of Management,* 32(6), pp.951-990.

Turing, A. M.(1950), Computing machinery and intelligence. *Mind,* 59(236), pp.433-460.

Van de Ven, A. H., and Johnson, P. E.(2006), Knowledge for theory and practice. Academy of Management Review, 31(4), pp.802-821.

Victor, B., and Cullen, J. B.(1988), The organizational bases of ethical work climates. *Administrative Science Quarterly,* 33(1), pp.101-125.

Vygotsky, L. S.(1978), *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, W., Hackett, R. D., Archer, N., Xu, Z., and Yuan, Y.(2025), Will AI-enabled conversational agents acting as digital employees enhance employee job identity? Information and Management, 62, 104099.

Warren, R., and Tweedale, G.(2002), Business ethics and business history: Neglected dimensions in management education. *British Journal of Management*, 13(3), pp.209-219.

Welsh, D. T., Ordóñez, L. D., Snyder, D. G., and Christian, M. S.(2015), The slippery slope: How small ethical transgressions pave the way for larger future transgressions. *Journal of Applied Psychology,* 100(1), pp.114-127.

Willness, C. R., Steel, P., and Lee, K.(2007), A meta-analysis of the antecedents and consequences of workplace sexual harassment. *Personnel Psychology,* 60(1), pp.127-162.

Wilhelm, L. T., Ding, X., Knutsen, K. M., Carik, B., and Rho, E. H.(2025), How managers perceive AI-assisted conversational training for workplace communication. *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. ACM.

Wu, S., Liu, Y., Ruan, M., Chen, S., and Xie, X. Y.(2025), Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. *Scientific Reports,* 15(1), 15105.

Yan, C., Chen, Q., Zhou, X., Dai, X., and Yang, Z.(2024), When the Automated fire Backfires: The Adoption of Algorithm-based HR Decision-making Could Induce Consumer's Unfavorable Ethicality Inferences of the Company: C Yan et al. *Journal of Business Ethics*, 190(4), pp.841-859.

Zao-Sanders, M.(2025, April 9), How people are really using gen AI in 2025. *Harvard Business Review*.
https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.

• 김성준 교수는 국민대학교 경영대학 겸임교수로 재직 중이며, SK그룹과 롯데그룹 인재개발원에서 근무했다. 리더십, 조직문화, People Analytics 연구를 수행하며, 다수의 국내외 학술지에 논문을 게재하고 5권의 저서를 출간했다.

• 이중학 교수는 동국대학교 경영학과 조교수로 재직 중이며, 현대자동차그룹 경영연구원과 롯데인재개발원에서 근무했다. 인공지능, 다양성 관리, People Analytics 연구를 수행하며, 45편 이상의 논문을 게재하고 10권의 저서를 출간했다.