불교와 사회 17권 2호(2025), pp. 1-48
The Journal of Buddhism and Society Vol.17 No.2(2025), pp. 1-48
https://doi.org/10.33521/jbs.2025.17.2.1

인공지능의 의미 공백과 행위성의 한계

지경선*

목차

* 이화여자대학교 일반대학원 철학과 박사과정, 차의과학대학교 일반대학원 의학과 박사

국문 초록

오늘날의 딥러닝 기반 AI는 방대한 데이터를 통해 통계적 패턴을 학습하지만, 이는 규칙에 따른 기호 조작, 즉 구문론적 처리(syntactic processing)에 머문다. 이때 본 논문에서 '의미 이해'는 모델 내부의 분포적 표상이나 기능적 정보처리 일반이 아니라, 지시·진리 조건 및 공적 규범(언어게임·삶의 형식)에 의해 성립하는 규범적 의미의 충족을 뜻한다.

본 연구는 인공지능(AI)의 의미 이해 가능성과 행위성(agency)의 한계를 철학적으로 규명하기 위해, 불교의 연기·공(空) 사상과 비트겐슈타인 후기 언어철학을 '문제-중심(problem-centered) 비교'의 틀에서 비교·분석한다. 본 논문은 현대 AI가 보이는 '의미 공백(semantic gap)'을 공통 문제로 설정하고, 비교의 기준을 (1) 의미 성립의 조건(내재성·비내재성), (2) 의미를 안정화하는 규범의 근거(언어게임·삶의 형식 vs 연기적 조건망·관습), (3) 행위성과 책임 귀속의 조건(규칙 따르기·의도 vs 무아·업(karma)과 의도(cetanā))으로 제시한다.

비교의 전략은 두 전통을 개념적으로 동일시하지 않고, 위 기준에 따라 '구조적 상응(유사성)'과 '비등가적 차이성'을 구분해 제시하는 데 있다. 비교 결과, 두 전통은 의미가 기호 내부에 자족적으로 내재하지 않고 사용·조건·맥락에 의해 성립한다는 점에서 수렴하지만, 비트겐슈타인은 언어 규범의 기술과 철학적 혼란의 치료라는 문법적 접근을 취하는 반면, 불교(중관)의 연기·공 사상은 존재와 인식의 무자성을 논증하며, 윤리적 지평에서 의미와 책임 논의를 확장한다.

이러한 비교 결과를 바탕으로, 본 논문은 AI의 결정과 행위에 대한 책임 귀속 문제를 검토하되, 불교 윤리가 강조하는 의도(cetanā) 개념을 중심으로 현존 AI가 도덕적 행위자로 성립하기 어렵다는 점을 논증한다. 이를 통해 본 연구는 AI의 의미・행위성・책임 문제를 서양 분석철학과 불교철학의 비교철학적 틀 안에서 재구성함으로써, 기술철학・윤리학・불교학의 교차 영역에 새로운 해석틀을 제시한다.

주제어 : 인공지능, 의미 공백, 행위성, 연기・공(空) 사상, 비트겐슈타인

Ⅰ. 서론

1. 연구배경

인공지능(AI)의 급속한 발전은 행위성, 의미 이해, 도덕적 책임에 관한 고전적 철학 문제들을 다시 제기하고 있다. 최근의 생성형 AI(대규모 언어 모델 및 멀티모달 기반 기초모델)는 Transformer 계열 아키텍처를 중심으로 방대한 말뭉치와 다종(多種) 모달 데이터를 자기지도 방식으로 사전학습(pretraining)하고, 손실 함수 최소화(확률적 최적화)를 통해 고차원 파라미터 공간에서 가중치를 갱신하는 방식으로 작동한다. 이러한 구조는 겉으로는 '학습'처럼 보이지만, 핵심은 세계에 대한 개념적 이해의 축적이라기보다 '분포 예측' 성능을 높이는 확률적 최적화 절차의 반복이라는 점에 있다(예: Goodfellow et al., 2016; Vaswani et al., 2017; Bommasani

et al., 2021). 중요한 점은, 이 과정에서 모델이 처리하는 정보는 형식적 패턴 및 통계적 상관관계에 한정되며, 내용적・지시적 의미에 대한 접근은 발생하지 않는다는 것이다.

이러한 구조적 특징 때문에 AI의 출력은 자연어와 유사한 형태를 갖더라도, 의미적 세계에 대한 내재적 이해를 반영하지 않는다. 이 문제는 인지과학과 철학에서 오래전부터 제기되어 온 '구문(syntax)과 의미(semantics)의 분리' 문제와 본질적으로 연결된다(Harnad, 1990; Dreyfus, 1992). 서얼의 '중국어 방' 논증(Searle, 1980)은 바로 이러한 간극을 극명하게 보여준다. 즉, 규칙에 따른 기호 조작이 가능하다는 사실은 의미 이해를 보증하지 않으며, 딥러닝 시스템의 작동 방식 또한 동일한 한계를 지닌다(Bender et al., 2021).

그 결과, 오늘날 사회 곳곳에서 AI의 판단・추천・분류가 의사결정에 직접적인 영향을 미치는 상황에서도, AI의 행위성 또는 의미 이해를 그대로 인정하기 어렵다는 철학적 문제가 대두되고 있다. 한국 사회에서도 AI 채용・신용평가・추천 알고리즘을 둘러싼 불공정성 및 책임 귀속 문제가 현실적 쟁점으로 등장하고 있으며, 법・윤리・철학의 새로운 해석틀이 요청되고 있다.

본 논문은 이러한 문제의식을 바탕으로, 현대 AI가 보여주는 의미 이해의 구조적 한계와 행위성 논쟁을 불교의 연기・공 사상 및 비트겐슈타인 후기 언어철학의 관점에서 재조명한다. 두 전통 모두 의미가 고정된 실체가 아니라, 관계적・사용 기반의 발생 구조임을 강조한다는 점에서 공통적인 통찰을 제공하며, 이는 AI 의미 공백 문제를 해석하는 데 새로운 철학적 틀을 제시할 수 있다. 본 연구는 연기・공 사상과 언어철학을 결합한 비교철학적 접근을 통해, AI 윤리・책임 논쟁을 불교적 사회담론의 지평에서 재구성

하고자 한다.

2. 연구목적

위와 같은 기술적・철학적 배경 속에서, 본 논문은 인공지능(AI)의 의미 이해 가능성과 행위성(agency)의 한계를 철학적으로 규명하는 것을 핵심 목적으로 하며, 이를 위해 다음의 연구 목적을 설정한다.

첫째, 딥러닝 기반 AI의 구문론적 정보처리와 의미론적 이해 사이에 존재하는 간극, 즉 '의미 공백(semantic gap)'을 철학적으로 정식화한다. 이를 위해 후기 비트겐슈타인의 언어철학('의미=사용')을 주된 분석 틀로 삼고, 서얼(John Searle)의 중국어 방 논증은 구문론적 처리와 의미론적 이해의 분리를 직관화・정식화하기 위한 보조 논증으로 검토한다.

둘째, 이러한 의미 공백에 대한 진단이 불교의 연기(緣起)・공(空) 사상, 특히 무자성(無自性) 개념과 어떻게 교차하는지를 밝힌다. 본 논문은 의미를 고정된 실체가 아니라 관계적 조건 속에서 발생하는 과정으로 이해하는 불교철학의 틀을 통해, AI 의미 생성의 비본질적・조건발생적 성격을 해석한다(Ames 2003).

셋째, AI의 판단과 결정이 현실에서 문제를 야기할 경우, 앞서 도출한 의미 공백과 행위성 한계가 책임 귀속(responsibility attribution) 논쟁에 함의하는 최소 결론을 검토한다. 이 과정에서 행위성, 의도성(intentionality), 책임 귀속의 경계를 재검토하되, 본 논의는 현존 AI를 도덕적 행위자(책임 보유 주체)로 간주할 수 없는 이유와 책임이 인간 및 제도적 조건망으로 귀속되어야 한다는 점을 밝히는 데 한정한다.

넷째, 위 두 전통의 비교가 단순한 유사성 나열에 그치지 않도록,

의미 성립의 조건, 규범과 실천의 역할, 주체·의도 개념, 책임 논의로의 확장 가능성이라는 기준에서 유사성과 차이성을 명시적으로 도출·정리하고, 그 결과를 AI 의미·행위성·책임 논쟁에 적용한다.

이상의 연구 목적을 통해, 본 논문은 AI·언어철학·불교철학을 하나의 비교철학적 틀로 재구성하되, 비교의 상응점과 비등가적 차이를 구분해 제시함으로써 AI 시대의 의미·행위성·책임 문제를 정교하게 해명하는 데 기여하고자 한다.

3. 연구 방법·비교 기준 및 범위

본 절은 본 논문이 채택하는 방법론적 절차와 논증의 범위를 제시함과 동시에, 비트겐슈타인 후기 언어철학과 불교의 연기·공 사상을 비교하기 위한 기준, 방법, 그리고 비교 결과의 제시 방식을 명시한다.

1) 비교의 기준

본 논문은 다음 세 가지 기준에 따라 두 전통을 비교한다.

첫째, 의미 성립의 조건이다. 의미가 기호 내부에 자족적으로 내재하는지, 아니면 사용·관계·조건에 의존하여 성립하는지 여부를 기준으로 삼는다.

둘째, 의미를 안정화하는 규범의 근거이다. 비트겐슈타인의 언어게임과 삶의 형식이 의미의 규범성을 어떻게 설명하는지, 그리고 불교의 연기적 조건망과 관습이 의미를 어떻게 성립시키는지를 비교한다.

셋째, 행위성과 책임 귀속의 조건이다. 규칙 따르기와 의도 개념을 중심으로 한 비트겐슈타인의 논의와, 무아·업·의도(cetanā)를

중심으로 한 불교 윤리의 관점을 대비한다.

2) 비교의 방법과 전략

본 논문은 '문제-중심(problem-centered) 비교' 전략을 취한다. 즉, AI의 의미 공백이라는 공통 문제를 설정한 뒤, 두 전통의 개념을 동일시하지 않고 각각의 고유한 문맥에서 재구성한다. 이후 비교 기준에 따라 구조적 상응점(유사성)과 비등가적 차이점을 구분하여 제시하고, 그 결과를 AI 의미 이해와 행위성·책임 논의에 적용한다. 비교의 목표는 개념적 통합이 아니라, 각 전통이 제공하는 설명 자원의 범위와 한계를 드러내는 데 있다.

3) 논증의 범위와 비교 결과의 제시 위치

본 논문은 제Ⅱ장(딥러닝 기반 AI의 정보처리 구조 분석을 통해 의미 공백을 기술적·철학적으로 정식화) - 제Ⅲ장(연기·공 사상에 따른 의미 해석을 제시하고, 장 말미에서 두 전통의 비교 결과를 유사성과 차이성으로 구분해 정리) - 제Ⅳ장(비교 결과를 전제로 행위자성(agency)·책임 귀속의 최소 조건을 점검하는 범위에서 AI 책임 문제를 제한적으로 검토) - 제Ⅴ장(전체 논증의 요약과 방법론적 의의 종합)으로 구성된다.

4) 논증(입증) 방법의 정당화

본 논문이 "현존 딥러닝 기반 AI는 의미를 이해하지 못하며(semantic gap), 도덕적 행위자로서의 행위성(agency)을 보유하기 어렵다"는 결론에 도달하는 방식은 경험적 성능 비교의 귀납이 아니라, 조건부·개념분석적 필요조건 논증이다. 논증 절차는 다음과 같이 구성된다. (a) 먼저 현존 시스템의 설계·작동 원리—확률적 토큰 예

측을 위한 통계적 최적화—가 체계 내부에 무엇을 표상·구성하는지에 대한 기술적 전제를 확정한다. (b) 다음으로 '의미 이해(semantic understanding)'가 성립하기 위한 최소 조건을 정식화한다. 본 논문은 의미 이해의 최소 조건을 지시·진리조건(또는 그에 준하는 평가 가능성)과 공적 규범에 따른 사용으로 설정하며, 이 조건은 후기 비트겐슈타인의 "의미=사용", 언어게임·삶의 형식, 규칙 따르기 논의에 의해 철학적으로 정당화된다. (c) 마지막으로 위 최소 조건이 현존 AI에 충족되지 않음을 보인다. 즉, 통계적 기호 조작이 문법적 정합성과 유사-담화적 산출을 가능하게 하더라도, 체계 내부에서 지시·진리조건 및 공적 규범에 따른 사용을 스스로 정초하거나(ground) 책임질 수 있는 방식으로 수행한다고 보기 어렵다는 점을 논증한다.

이후 행위성·책임 논증은 (b)의 결론을 전제로 확장된다. 본 논문은 책임 귀속이 가능한 행위자성을 위해 최소한 의도성(intentionality)이 요구된다는 현대 책임론의 논점을 수용하고, 이를 불교윤리에서 행위의 도덕적 성격을 규정하는 핵심 요인으로 제시되는 의도(cetanā) 개념과 접속시켜 '책임 보유 주체'의 최소 조건을 도출한다. 그 결과, 의식적 의도 및 도덕적 숙고 구조가 결여된 현존 AI는 책임 보유 주체로서의 지위를 갖기 어렵다는 결론이 산출된다. 다만 이는 "AI가 원리적으로 의미를 이해할 수 없다"는 강한 형이상학적 부정이 아니라, 본 논문이 분석 대상으로 삼는 현존 설계 패러다임(비체화·통계적 언어모델 중심) 하에서의 한계 진단임을 명시한다.

또한 본 논문은 '의미 이해'에 대한 논의를 의미 성립의 구성적 조건(constitutive conditions) 분석으로, '책임 귀속'에 대한 논의를 그 결론을 전제로 한 적용(행위자성의 최소 조건 점검)으로 층위화한다. 다시 말해, 의미(구성) 논의와 책임(행위론) 논의를 동일 층

위에서 혼합하지 않고, 전자의 결론이 후자의 판단을 제한·구속하는 연결 원리를 명시한다.

5) 용어의 작업 정의: '의미(meaning)'의 이중 용례 구분

본 논문에서 '의미(meaning)'는 AI 논의에서 빈번히 혼용되는 용례를 구분하기 위해 두 층위로 사용된다. (a) AI/컴퓨터과학 문헌에서 흔히 'semantic representation' 또는 '의미'라고 불리는 것은, 모델이 과제 수행을 위해 구성하는 분포적·벡터적 표상(예: 토큰 간 유사성 구조)으로서 '기능적 의미(functional semantics)'에 가깝다. 이는 예측·분류·번역 등에서 실용적 효용을 갖고, 일정한 수준의 구조적 안정성을 보일 수 있으나, 기호가 무엇을 지시하는지, 어떤 조건에서 참·거짓으로 평가되는지, 어떤 공적 규범 아래에서 정당한 사용으로 간주되는지에 대한 기준을 체계 내부에서 스스로 정초하지는 않는다. (b) 반면 본 논문에서 '의미 이해(semantic understanding)'가 가리키는 의미는 규범적 의미(normative semantics)이며, 지시·진리조건, 화행적 적합성, 그리고 공동체적 규범에 따른 사용(언어게임·삶의 형식)으로 구성되는 의미를 뜻한다. 이하에서 '의미 공백(semantic gap)'은 (a)의 기능적 의미가 (b)의 규범적 의미를 대체·충족하는 데 실패하는 간극, 곧 통계적 기호 조작의 성공이 규범적 의미 이해를 보증하지 못하는 구조적 비연속성을 지칭한다. 본 논문의 의미·행위성 논증은 이 구분을 전제로 전개된다.

6) 이론 자원의 위계: 주축 비교와 보조 논증의 구분

본 논문의 비교철학적 주축은 비트겐슈타인 후기 언어철학과 불교(중관)의 연기·공 사상이다. 따라서 서얼의 '중국어 방' 논증

및 기타 AI 비판 논의는, 현존 AI의 '구문론적 처리 – 의미론적 이해' 간극을 정식화·직관화하기 위한 보조 논증으로만 활용되며, 본 논문의 비교축을 제3의 전통으로 확장하지 않는다. 또한 로보-다자인과 같은 미래 AI 주체성 논의는 현존 AI의 의미·행위성·책임 분석과 직접 결합되지 않으므로, 본문 논증에서는 제외하고 후속 연구 과제로만 간단히 유보한다.

Ⅱ. AI에서의 구문 처리와 의미 이해

1. 딥러닝 기반 정보처리 구조

1) 오차 최소화 절차로서의 순전파(forward propagation)와 역전파(backpropagation)

현대 AI, 특히 딥러닝 신경망은 어떤 대상을 이해한다기보다 입력된 정보를 여러 계산 단계로 변환하는 방식으로 작동한다. 학습 과정은 순전파(forward propagation)와 역전파(backpropagation)라는 두 절차로 이루어진다. 먼저 순전파에서는 입력이 여러 층을 지나면서 출력이 만들어지고, 이어서 역전파에서는 그 출력이 얼마나 틀렸는지를 계산해 가중치를 다시 조정한다. 이 조정에는 경사하강법(gradient descent)이 사용되는데, 이는 오차를 줄일 수 있는 방향으로 값을 조금씩 움직이는 반복 절차다(Goodfellow, Bengio, and Courville 2016).

이미지 분류 모델을 예로 들면 구조가 더 분명해진다. 모델은 사진 속 대상을 이해해서 답을 내는 것이 아니라, 수많은 파라미터

를 조금씩 고쳐 나가면서 특정 라벨과 더 잘 맞는 출력을 만들어 낸다. 겉으로는 학습처럼 보이지만, 실제로는 오차가 줄어드는 방향을 찾기 위해 값을 반복적으로 조정하는 계산 과정에 가깝다.

따라서 순전파는 계산을 앞쪽으로 전달하는 절차이고, 역전파는 그 계산에서 생긴 오차를 되돌려 전달해 값을 수정하는 절차다. 이 과정 어디에서도 의미를 파악하거나 개념을 이해하는 단계는 등장하지 않는다. 딥러닝의 학습은 통계적 패턴을 조정하는 반복적 최적화 작업이며, 철학적 의미의 이해나 지향성과는 성질을 달리한다. 다만 2010년대의 딥러닝 논의가 주로 지도학습 기반 분류・인식 모델(CNN/RNN 등)의 성능 문제를 중심으로 전개되었다면, 2020년대 이후에는 transformer 기반의 대규모 사전학습 모델이 생성형 AI의 기술적 토대를 이룬다(Vaswani et al., 2017; Brown et al., 2020). 이들 모델은 '다음 토큰 예측(next-token prediction)'과 같은 자기지도 목표를 통해 언어의 분포적 규칙을 학습하고, 미세조정(fine-tuning) 및 인간 피드백 기반 학습(RLHF) 등으로 출력 행태를 조정한다(Ouyang et al., 2022). 또한 텍스트뿐 아니라 이미지・오디오・영상 등을 함께 다루는 멀티모달 모델도 확산되었으나, 이러한 확장 역시 기본적으로는 통계적 대응관계 학습에 기반하며, 지시・진리조건・규범적 사용의 자기조정 메커니즘이 체계 내부에 '그 자체로' 구현되는 것은 아니라는 점에서 본 논문의 '의미 공백' 논지와 직결된다.

2) 파라미터 공간(parameter space)과 경사하강법(gradient descent)

최근의 생성형 AI(대규모 언어 모델 및 멀티모달 모델)는 일반적으로 수십억~수백억(또는 그 이상) 규모의 파라미터를 경사하강법 기반 학습으로 조정하며, 이를 안정적으로 작동시키기 위해 대규모

데이터와 계산자원을 요구한다(Goodfellow et al., 2016; Vaswani et al., 2017). 과적합(overfitting)은 모델이 훈련 데이터의 우연적 변동까지 포착하여 새로운 입력으로 일반화하지 못하는 현상을 가리키며, 현대 딥러닝에서는 정규화 · 드롭아웃 · 데이터 증강, 그리고 대규모 사전학습-미세조정과 같은 전략으로 그 양상이 상당 부분 완화되어 왔다. 따라서 본 논문의 논점은 과적합이 '해결되었는가'라는 공학적 쟁점이 아니라, 설령 일반화 성능이 향상되더라도, 학습 목표가 '기호-기호 간 분포 예측'에 놓이는 한 지시 · 진리조건 · 규범적 사용을 기준으로 출력을 스스로 교정하는 의미론적 메커니즘이 체계 내부에 자동 도입되지 않는다는 구조적 간극에 있다.

이때 흔히 파라미터 공간(parameter space)을 하나의 넓은 지형으로 비유한다. 각 파라미터 조합은 이 지형의 한 점에 해당하며, 모델은 오차가 가장 작은 지점을 찾기 위해 이 지형을 이동한다. 경사하강법은 바로 이 이동 경로를 정해 주는 규칙으로, 오차가 감소하는 방향으로 조금씩 내려가는 방식이다.

그러나 이 절차는 어디까지나 수치적으로 오차를 줄이는 계산 과정일 뿐, 모델이 대상이나 세계를 이해하고 있다는 증거는 아니다. 경사하강법은 지능적 사고를 보여주는 것이 아니라, 기울기가 낮아지는 방향을 기계적으로 따라가는 반복 규칙에 가깝다. 이런 점에서 이 과정은 철학적 의미에서 말하는 지향성이나 의미 이해와는 거리가 있다.

3) 일반화 실패의 문제

과적합(overfitting)은 전통적 딥러닝에서 반복적으로 논의된 일반화 위험이지만, 최근에는 대규모 사전학습과 정규화 기법의 발전으로 그 양상과 심각성이 크게 완화되었다. 그럼에도 생성형 AI에

서는 '환각(hallucination)'이나 분포 외(out-of-distribution) 입력에 대한 취약성처럼, 표면적 정합성의 향상이 곧 지시·진리조건에 대한 규범적 통제를 보증하지 못함을 보여주는 현상이 지속적으로 나타난다. 딥러닝 학습이 통계적 상관관계를 최적화하는 절차라는 점은 변하지 않으며, 이는 형식적 구조(syntax)의 고도화이지 의미적 세계(semantics)에 대한 규범적 접근을 그 자체로 포함하지 않는다. 데이터 규모가 확대되고 일반화가 개선되더라도, 모델은 자신의 출력을 세계 상태와 대조해 참·거짓을 판정하고 수정하는 절차를 내장하지 않기 때문에, 그럴듯한 문장·응답이 생성되더라도 그것이 곧 '이해'의 성립을 뜻하지는 않는다. 이러한 점은 '성능'과 '의미 이해'를 동일시할 수 없다는 사실, 즉 본 논문이 말하는 의미 공백(semantic gap)의 핵심을 기술적 현상 차원에서 뒷받침한다.

또한 과적합된 모델은 훈련 데이터와 구조적으로 유사한 입력에는 높은 정확도를 보이지만, 입력 조건이 미세하게 바뀌면 출력이 급격히 불안정해진다. 이는 인간이 맥락·의도·사용 등 의미론적 요인에 근거하여 판단을 조정하는 방식과, AI가 표면적 패턴 연계에 의존하는 방식 사이의 질적 차이를 극명하게 대비시킨다. 즉, 딥러닝의 오차 최소화가 형태적 정합성을 강화하더라도 의미적 정합성을 구성하는 것은 아니다.

바로 이 지점에서 과적합 및 환각과 같은 일반화 문제는, 기술적 성능 논의와 별개로 "구문론적 최적화"가 "규범적 의미 이해"를 보증하지 못한다는 의미 공백(semantic gap)의 구조를 드러내는 사례로 해석될 수 있다. 서얼(Searle)의 중국어 방 논증이 주장하듯, 기호 조작의 정교함은 의미 이해를 보증하지 않는다. 비트겐슈타인의 후기 언어철학이 의미를 사용과 맥락(Lebensform) 속에서 발생하는 것으로 보았다는 점을 고려하면, 과적합은 AI가 '의미의 사용'에 접근할

수 없는 구조적 이유를 실증적으로 보여준다. 더 나아가 불교의 연기·공(空) 사상에서 의미는 기호 내부의 고정된 실체가 아니라, 관계적 조건에서만 발생한다는 통찰은, AI가 '본래적 의미'를 가질 수 없는 구조적 근거를 해석하는 데 중요한 철학적 자원이 된다.

따라서 본 절에서의 일반화 실패(과적합·환각 등) 논의는, 본 논문이 전개하는 AI의 구문론적 처리 - 의미 이해의 간극(의미 공백) 및 행위성 한계 논증으로 이어지기 위한 '문제 제기 장치'로 위치한다.

2. 구문론(syntax)과 의미론(semantics)의 구분

본 절에서 '의미론(semantics)'은 언어모델이 학습하는 분포적 표상(기능적 의미) 일반을 지칭하는 약한 용례가 아니라, 지시·진리조건 및 공적 규범(언어게임·삶의 형식)에 의해 성립하는 규범적 의미를 중심으로 사용한다. 따라서 이하에서 말하는 '의미 공백'은 단지 "표상이 없다"는 뜻이 아니라, 규범적 의미 이해를 성립시키는 조건이 체계 내부에 구현되어 있지 않다는 진단을 가리킨다.

1) 통계적 처리와 '언어처럼 보이는 것'

앞 절에서 보았듯, 현대 딥러닝 기반 언어 모델은 거대한 파라미터 공간에서 손실 함수를 최소화하는 방향으로 가중치를 조정함으로써, 주어진 말뭉치 위에서 특정 형태의 확률 분포를 근사한다(Goodfellow et al., 2016). 자연어 처리의 관점에서 가장 간단히 말하면, 모델은 어떤 토큰(단어/형태소) 시퀀스 $\omega_1,\ldots,\omega_t$가 주어졌을

때, 다음 토큰ω_{t+1}에 대한 조건부 확률 분포 $P(\omega_{t+1}|\omega_1, ..., \omega_t)$를 최대한 정확하게 맞추도록 학습된 함수이다. 다시 말해, 언어 모델의 "지능"은 주어진 맥락에서 다음에 올 기호의 분포를 얼마나 잘 예측하는가로 정의된다. 이 과정에서 모델 내부에서 일어나는 일은, 입력된 토큰 시퀀스를 고차원 실수 벡터로 매핑하고, 여러 층을 거쳐 비선형 변환을 적용한 뒤, 마지막 층에서 softmax 연산을 통해 각 후보 토큰에 대한 확률을 산출하는 절차이다. 전 과정은 통계적 패턴과 함수적 연산의 조합이며, 설계 수준에서 보면 "이 문장이 무엇을 가리키는가?", "이 발화가 참인가 거짓인가?"와 같은 지시・진리 조건・화행적 의도가 독립된 형식 의미 이론으로 명시 구현되어 있지 않다. 이 점에서 언어 모델이 학습하는 것은 근본적으로 언어-언어 간의 분포적 상관관계이지, 언어-세계 간의 직접적인 대응 구조가 아니다. 이는 Sellars가 말하는 "언어-진입(language-entry) 전이"와 "언어-내부(intra-linguistic) 전이"의 구분과 상응한다. Sellars의 틀에서 보면, 대규모 언어 모델은 후자의 규칙들(기호와 기호가 어떻게 이어지는가)에 대해서는 극단적으로 숙련되어 있지만, 지각・경험을 토대로 세계로부터 언어로 진입하는 규칙에 해당하는 층위는 설계 단계에서 거의 고려되지 않는다. 이러한 구조적 특징 때문에, 대규모 언어 모델이 만들어 내는 출력은 문법적으로는 정합적이고 표면 형태상 자연어 문장과 거의 구별되지 않지만, 그 생성 과정은 어디까지나 구문론적 변환과 확률적 최적화로 이해해야 한다(Harnad, 1990; Bender et al., 2021). Bender 등이 언어 모델을 "확률에 의해 움직이는 앵무새(stochastic parrots)"라고 부르는 것도 이 때문이다. 모델은 언어 공동체가 축적해 온 말뭉치의 분포를 학습해 그 패턴을 통계적으로 재생산함으로써, 마치 무언가를 이해한 것 같은 출력을 산출하지만, 참・거짓,

지시 대상, 화행 효과와 같은 의미론·화용론적 구조는 여전히 모델 외부, 곧 인간 해석자의 활동에 전적으로 의존한다(Bender et al., 2021).

요약하면, 현존 AI의 언어 능력은 구문론적 패턴의 고도화된 일반화로 보는 것이 타당하다. 모델은 "다음에 올 기호의 분포"를 정교하게 추정함으로써 언어처럼 보이는 결과물을 생산하지만, 그 결과는 기호열의 통계적 연속성을 유지하는 데 최적화된 산출물일 뿐, 기호가 가리키는 세계를 이해하고, 그에 대해 판단한 결과라고 보기는 어렵다. 이 점에서 AI의 정보처리는 인간 언어 행위가 갖는 의미론·화용론적 층위와 구조적으로 다른 차원에 머무른다.

3. 의미 공백(semantic gap)과 지시·진리 조건의 비구현

이처럼 AI가 수행하는 것이 통계적 기호 조작이라는 점을 인정하면서도, 그 능력을 곧바로 인간의 "이해"나 "지능"과 동일시하려는 시도는 철학적으로 문제가 있다(Harnad, 1990; Bender et al., 2021). 본 논문에서 말하는 '의미 공백(semantic gap)'은 언어모델이 구축하는 '기능적 의미(분포적·표상적 유사성 구조)'와 인간 언어 행위가 전제하는 '규범적 의미(지시·진리조건·화행·공적 규범에 따른 사용)' 사이의 간극을 뜻한다. 겉으로 보기에는 완결된 문장처럼 보이는 AI의 출력 속에, 그 문장이 무엇을 지시하는지, 어떤 조건에서 참·거짓이 되는지, 어떤 행위적 효과를 노리는 발화인지가 체계 내부에 의미론적으로 표상되어 있지 않다는 점 때문이다. Searle의 중국어 방 논증(Searle, 1980)은 이 문제를 선명하게 드러낸다. 규칙에 따라 기호열을 변환하는 능력 자체는, 그 기호가 나타내는 의미 내용이나 지시 대상을 "이해"하는 것과 동일하

지 않다. 중국어 방 안의 행위자가, 혹은 오늘날의 언어 모델이, 주어진 입력에 대해 문법적으로 적절한 기호열을 응답으로 내놓을 수 있다고 하더라도, 그 체계가 “이 언어가 무엇에 대한 언어인지”, “어떤 세계 상태를 기술하는지”를 파악하고 있다는 근거는 없다. 내부에서 진행되는 것은 기호-기호 간 연산이며, 기호-세계-사용자 사이의 삼항적 관계는 체계 설계의 일부로 구현되어 있지 않다는 것이 Harnad가 제기한 ‘symbol grounding problem’의 핵심 진단이다(Harnad, 1990). Bender 등도 이 점을 강조하며, 의미는 기호열의 형식 그 자체에 들어 있는 것이 아니라, 그것을 사용하는 사람들의 실천과 맥락 속에서 생성된다고 주장한다(Bender et al., 2021). 언어 모델은 방대한 텍스트에서 기호 간 분포적 연관성을 포착해 이를 통계적으로 재조합하지만, 인간 화자가 언어 사용을 통해 전제하는 세계에 대한 관여, 공동체적 규범, 신체적·정동적 경험은 모델의 구조 속에 포함되어 있지 않다. 이런 의미에서 AI의 출력은 형식적·구문론적 층위에서는 충실하지만, 의미론적·화용론적 층위에서는 구조적인 공백을 지닌다. 명제의 진리 조건을 중심으로 언어와 세계의 대응을 분석한 비트겐슈타인의 초기 그림 이론(Wittgenstein, 1922)이나, 의미를 “사용”과 “삶의 형식(Lebensform)” 속에서 파악한 후기 언어철학(Wittgenstein, 1953/2009)을 기준으로 할 때, 현존 AI는 의미가 자리 잡을 수 있는 세계–사용자–규범의 장(field)을 체계 내부에 갖추지 못한다. Sellars의 용어를 빌리면, AI는 “언어-내부 전이(language-language transitions)”에는 탁월하지만, 세계에서 언어로 들어가는 “언어-진입 전이(language-entry transitions)”, 그리고 언어에서 행위로 나가는 “언어-출구 전이(language-exit transitions)”를 구성하지 못하는 체계에 가깝다. 따라서 ‘의미 공백’은 단순히 “아직 충분히 고도화되지 않은 기술 수준”

의 문제가 아니라, 현재의 설계 패러다임이 원칙적으로 포괄하지 못하는 층위를 가리키는 개념이다. 곧, 기호가 세계・사용자와 맺는 의미론적・실천적 관계가 체계 내부에서 비어 있다는 구조적 진단이다. 이 진단은 뒤에서 다룰 Searle의 중국어 방 논증과 비트겐슈타인의 "의미=사용" 논의를 이어 주는 동시에, 불교 연기・공 사상에서 말하는 무자성(無自性), 다시 말해 기호와 정보가 그 자체로 고정된 의미를 지니지 않고, 관계적 조건망 속에서만 의미를 부여받는다는 통찰과 상응한다. 이런 관점에서 보면, AI의 '의미 공백'은 단지 기술적 결함이 아니라, 기호-세계-사용자 연기망 안에서만 의미가 성립한다는 사실을 드러내는 현대적 사례로 해석될 수 있다.

요컨대 본 장의 기술적 분석은 (1) 현존 언어 모델이 수행하는 연산이 '기호-기호 간 확률적 변환'이라는 점, (2) 그 과정에 지시・진리 조건 및 규범적 사용의 자기조정 메커니즘이 체계 내부에 구현되어 있지 않다는 점을 확인한다. 따라서 다음 절에서는 비트겐슈타인의 후기 언어철학('의미=사용', 규칙 따르기)을 통해 의미 성립의 공적・규범적 조건을 검토하고, 그 관점에서 AI 출력의 의미가 기계 내부에 내재하기보다 해석자・언어공동체・사회적 맥락이라는 사용 조건망에서 성립함을('의미 공백') 정식화한다.

4. 비트겐슈타인의 후기 언어철학과 의미의 사용론

이 통찰은 비트겐슈타인의 언어철학에서도 드러난다. 초기 비트겐슈타인은 언어를 세계의 논리적 구조를 반영하는 체계로 이해하며, 명제는 사실(fact)의 형식을 '그려내는(picturing)' 논리적 표현이라고 보았다(Wittgenstein 1922). 이러한 관점은 언어의 의미를

형식적 규칙과 구문(syntax)의 조합으로 환원하는 입장을 따른다. 비트겐슈타인의 후기 언어철학은 의미를 고정된 대상이나 내적 표상에 두던 관점에서 벗어나, 언어의 실제 사용 속에서 의미를 찾는 전환을 보여준다. 그는 “한 단어의 의미는 그 사용에 있다”고 보며, 개별 표현의 의미가 언어게임과 삶의 형식 속에서 결정된다고 설명한다(Wittgenstein 1953/2009). 의미는 기호 자체에 담긴 속성이 아니라, 기호와 세계, 그리고 사용자의 실천이 서로 얽혀 있는 관계적 맥락에서만 성립한다. 크립키(Kripke, 1982)는 이러한 문제의식을 이어 규칙 따르기 문제를 제기하면서, 규칙의 의미가 개인의 내적 표상에서가 아니라 공동체적 합의와 반복된 실행 속에서 안정된다고 보았다. 이 관점에서 보면, 사회적 맥락에 참여하지 못한 채 형식 알고리즘에 따라 작동하는 AI가 생성하는 언어 출력이 과연 규칙을 “따른” 것인지에 대한 의문이 제기된다. 비슷한 지점을 플로리디(Floridi, 2011)와 브랜덤(Brandom, 1994) 역시 강조한다. 플로리디는 정보가 해석을 통과해야 비로소 ‘의미 있는 것’이 되며, 의미는 언제나 맥락적 관계망 속에서 발생한다고 말한다. 브랜덤은 언어의 의미가 사회적 추론과 규범적 관계 구조에서 형성된다고 보며, 단순한 기호 처리만으로는 의미에 접근할 수 없음을 지적한다. 이 둘 모두 의미를 기호 내부가 아니라, 관계적·사회적 맥락에서 설명하려는 공통점을 갖는다. 거대 언어모델에 대한 최근 비판도 같은 결론으로 이어진다. 벤더(Bender, 2021) 등은 대규모 언어모델을 “확률적 앵무새”로 규정하며, 이러한 시스템이 자연어를 생성하더라도 이는 통계적 상관성을 모방하는 결과일 뿐, 산출된 문장의 참·거짓이나 상황 적합성을 스스로 판단하는 것은 아니라고 지적한다. 즉, 현재의 AI는 인간 언어 공동체가 만들어 내는 의미의 규범적 구조에 실천적으로 참여하지 못한다는 한

계가 다시 확인된다. 결국 의미는 고정된 실체가 아니라 관계 속에서만 발생하는 개념이며, 이러한 비본질적·관계적 의미관은 관계적 실재론 혹은 의미의 비고정성으로 설명될 수 있다(허남결, 2024a; 이승종, 2002). 이러한 관점은 불교의 연기론과 무자성 사상과도 상응하며, 다음 장에서 두 전통 간의 구조적 연결점을 보다 상세히 논의할 것이다(허남결, 2024b; 이승종, 2024).

III. 연기(緣起, pratītyasamutpāda)와 의미의 공(空)

1. 중관철학(中觀哲學)의 무자성(無自性)과 관계적 존재론

1) 제법무자성(諸法無自性)의 철학적 의미

AI의 구문(syntax)과 의미(semantics) 사이의 간극은 불교의 인식론적 통찰과 깊이 공명한다. 앞서 보았듯, 현대 인공지능은 기호를 규칙적으로 조작할 수는 있지만 그 기호가 지시하는 '의미의 세계'에는 도달하지 못한다. 이와 유사하게 대승불교, 특히 중관(中觀) 사상은 모든 현상이 고정된 실체를 지니지 않으며, 오직 연기(緣起, pratītyasamutpāda) – 곧 조건과 상호의존 – 속에서만 성립한다고 본다(Garfield, 1995; Siderits, 2007; Westerhoff, 2009). 사물이란 독립된 자아나 본성을 갖지 않고, 관계적 작용의 그물망 속에서만 드러난다. 불교의 연기 사상은 데이터와 알고리즘이 상호의존적으로 작동하는 AI의 구조와도 철학적으로 공명하는 측면이 있다. 이는 단순한 비유라기보다, 세계를 관계적 과정으로 파악한다는 점에서 두 사유가 상당한 상응성을 지닌다는 점을 시사한다. 씨앗

이 흙・물・햇빛・계절 등 다양한 조건의 결합 속에서만 발아하듯, 모든 존재는 고립된 실체가 아니라, 상호의존적 조건망 속에서 성립한다. 용수(龍樹, 약 150~250)는 이를 "모든 법(法)은 스스로의 성품이 없다(諸法無自性)"고 표현하며, 속성의 실재성을 부정했다(Garfield, 1995; Ames, 2003). 이 '무자성(無自性)' 개념은 단순한 존재론 주장 이상의 의미를 갖는데, 모든 존재가 그 자체로는 아무것도 결정되지 않으며, 언제나 조건적으로만 성립한다는 근본적 인식론을 드러낸다. 이는 AI의 의미 처리 구조에서 발견되는 '의미 비고정성'과 깊이 연결될 수 있다(Thompson, 2007; Ladyman et al., 2007).

2) Ames의 해석: 공성과 관계적 실재론

이는 윌리엄 에임스(William L. Ames, 2003)가 지적하듯, 용수의 『중론(中論)』 제24장 18게 – "因緣所生法 我說卽是空(연기로 생겨난 모든 법을 나는 곧 공이라 설한다)" – 의 핵심 사유를 현대 과학철학적으로 해석한 것으로 볼 수 있다(Ames, 2003). 따라서 탁자가 탁자일 수 있는 것도 그 자체의 고유한 본성 때문이 아니라, 인간의 감각・의도・사용 맥락 속에서만 "탁자다움"이 발생하기 때문이다. 에임스(Ames, 2003)는 공성(空)을 '관계적 실재론(relational realism)'이라는 틀로 읽어낼 수 있다고 제안하면서, 연기・공 사상을 존재가 독립적 실체가 아니라, 관계 속에서만 드러난다는 통찰로 해석한다. 본 논문은 그의 제안을 중관학파 해석 가운데 하나의 현대적 시도로 간주하고, AI 의미 생성의 비고정성을 설명하는 비교철학적 틀로 부분적으로 수용한다. 이러한 해석을 통해 연기・공 사상이 과학철학의 논의와 접속할 수 있는 가능성이 드러나지만, 이는 중관 전통 전체를 대표하는 유일한 독해라기보다 특정 방향의 재구성으로 이해되어야 한다(Ames, 2003; Westerhoff, 2009).

2. 의미의 연기(緣起)

1) 기호-해석자-맥락의 상호조건성

이 관점은 정보・의미에도 적용될 수 있다. 기호 자체에는 의미가 없고, 언어 관습・화자・맥락에 의존해 의미를 지닌다(Putnam, 1975; Kripke, 1982; Wittgenstein, 1953/2009). 불교적으로는 의미가 자성적으로 공(空)하므로, 언어공동체・사용 맥락에서 연기(緣起)로 생겨날 뿐이다(Garfield, 1995; Siderits, 2007). 따라서 AI 출력의 의미는 AI 내부에 '본래' 있는 것이 아니라, 해석자(인간)와 사회언어적 관계망 속에서 연기(緣起)된다(Harnad, 1990; Bender et al., 2021). 구문 - 의미의 분리는 의미가 외재적・의존적 성질임을 재확인한다. 즉, 의미는 기호 자체에 '붙어 있는 성질'이 아니라, 기호와 해석자, 그리고 사회적 맥락의 관계적 작용 속에서 발생한다(Quine, 1951; Wittgenstein, 1953/2009). 후기 비트겐슈타인에서 '사용'은 단순한 빈도나 관찰 가능한 반응의 누적이 아니라, 옳고 그름을 가르는 규칙과 기준(criteria)이 작동하는 공적 실천을 가리킨다. 언어게임은 발화가 작동하는 과업・규칙・역할・맥락의 총체이며, 삶의 형식은 이러한 규칙들이 성립・유지되는 생활양식이다. 따라서 의미 이해는 단순히 어떤 기호열을 산출하는 능력이 아니라, 공동체적 규범 아래에서 "무엇이 올바른 적용이고 무엇이 오류인지"를 판정・교정할 수 있는 규범적 자리를 포함한다.

또한 규칙 따르기 논의가 보여주듯, 규칙은 개인의 내적 해석만으로 고정되지 않는다. '규칙을 따른다'는 것은 공동체적 훈련과 상호 교정 속에서 행위가 평가되고 안정화되는 것을 뜻한다. 이 관점에서 보면 언어모델이 외형상 규칙에 부합하는 문장을 산출하더라도, (1) 자신의 발화가 어떤 언어게임에서 어떤 규칙을 수행하

는지, (2) 어떤 경우에 자신의 발화가 잘못되었고, 어떻게 수정되어야 하는지를 공적 기준에 따라 스스로 조정하는지, (3) 그 조정에 대해 이유 제시와 책임을 질 수 있는지는 별개의 문제로 남는다. 본 논문은 이 간극을 기능적 의미와 규범적 의미 사이의 '의미 공백'으로 정식화하며, 바로 그 점에서 현존 AI의 의미 이해를 제한적으로 부정한다. 이와 같은 의미의 비내재성과 공적 규범에 따른 사용 구조는, 불교의 연기론이 주장하는 상호조건성의 핵심과 가까운 유사성을 지니는 것으로 볼 수 있다(Garfield, 1995).

2) 의미의 발생과 언어적 의존성

물리학에서 전자(electron)는 관측 맥락에 따라 파동으로도, 입자로도 드러난다. 이처럼 대상의 성질은 관찰 조건에 의존한다. 마찬가지로 컴퓨터 속의 기호 역시 독립적 의미를 지닌 실체가 아니라, 거대한 정보 체계 속 해석 작용을 통해서만 의미를 부여받는다(Floridi 2011).

에임스(Ames, 2003)는 공성과 양자론의 관계를 논하면서, 중관 철학이 제시하는 "모든 법은 자성으로부터 공하다"는 명제를 관계적 실재론의 통찰로 재해석할 수 있다고 본다. 그의 해석에 따르면, 존재와 인식은 독립된 실체가 아니라, 관계적·상호의존적 과정으로 성립하는 것으로 이해될 수 있다(Garfield, 1995; Ames, 2003). 이 논문은 이러한 관계적 해석을 참고하여, 의미 발생을 고정된 실체의 속성이 아니라 관계적 상호작용의 산물로 파악하는 관점을 채택한다. 그럴 때 인간 언어의 의미가 관용·관습·실천 속에서 고정되듯, AI가 생성하는 의미도 해석자와 사회적 사용 형태를 떠나 독립적으로 존재할 수 없다는 점이 보다 분명해진다(Kripke, 1982; Baker & Hacker, 2009).

3. AI 의미 생성의 비고정성

1) 불교적 조건망(緣起)으로 본 AI의 의미 형성

AI가 생성한 텍스트의 의미 형성 과정을 불교적 관점에서 보면, AI는 '본래적 이해'를 스스로 지니지 않으며, 그 이해는 본질적으로 공(空)하다(Siderits, 2007; Garfield, 1995). AI의 출력에 이해나 의도처럼 보이는 것이 있다면, 그것은 설계자(architecture), 데이터(corpus[1]), 사용자(interpretation). 그리고 이 모든 것이 얽혀 있는 연기적 조건망(緣起) 속에서만 드러난다(Floridi, 2011; Ladyman et al., 2007). 즉, AI가 출력한 문장은 고정된 의미를 갖지 않으며, 언제나 데이터・알고리즘・사용자 해석이라는 삼중 조건의 작용 속에서만 의미가 발생한다. 이는 불교가 말하는 "조건 발생적 존재론"과 구조적 유사성을 보이며, 양자 간의 비교 가능성을 보여준다(Garfield, 1995; Ames, 2003; Thompson, 2007).

2) 해석자와 사회언어적 맥락의 역할

기계 내부에는 '영혼'이라 부를 만한 고정된 자아가 자리할 여지가 없다. 불교가 인간을 오온(五蘊)—형태・감각・지각・의지・의식—의 조건적 결합과 흐름으로 파악하듯이(Garfield, 1995; Siderits, 2007), AI의 내부 역시 알고리즘・하드웨어・데이터・피드백이 얽힌 연산적 상호작용의 집합일 뿐이며, 그 어디에도 '이해'라는 자성(自性)이 독립적으로 존재하지 않는다. 이러한 구조적

1) Goodfellow, I., Y. Bengio, & A. Courville, *Deep Learning* (Cambridge, MA: MIT Press, 2016), p.446. 저자들은 "the TIMIT corpus"라는 표현을 사용하여 언어 및 음성 인식 모델의 훈련 데이터를 corpus로 지칭한다. 본 논문에서는 이러한 용례를 따라 corpus를 모델이 학습하는 대규모 언어 데이터 집합(training corpus)으로 확장하여 사용한다.

무자성은 곧 AI가 마음의 공성(空性), 즉 인식과 존재가 본래 독립 실체가 아니라, 관계 속에서만 성립한다는 불교적 통찰을 드러내는 현대적 사례가 될 수 있음을 시사한다(Ames, 2003; Westerhoff, 2009). 더 나아가 연기론의 관점에서 보면, AI가 사회와 사용자, 데이터, 설계라는 조건들에 의해 끊임없이 규정·재구성된다는 사실도 명확해진다. 이승종(2024) 역시 인간은 스스로의 무규정성 속에서 자신을 투사한 창조물을 만들어내며, AI는 결국 인간 지능의 거울·확장으로 기능한다는 점을 강조한다. 따라서 AI가 산출하는 의미는 기계 내부의 고정 속성이 아니라, 해석자·데이터·설계·사회적 사용이라는 조건망에서 성립한다. 이 점은 연기·공이 강조하는 '비내재성·조건발생'의 구조를 현대 기술 사례에서 다시 보게 해주는 비교의 계기로 제시될 수 있다(Garfield, 1995; Ames, 2003; Siderits, 2007).

4. 비교 분석: '의미=사용'과 연기·공(空)의 유사성과 차이

본 절은 제Ⅱ장에서 정식화한 비트겐슈타인적 문제설정(의미=사용, 규칙 따르기, 삶의 형식)과 제Ⅲ장에서 제시한 중관의 연기·공(무자성) 해석틀을, 서론에서 제시한 비교 기준에 따라 '유사성'과 '차이성'으로 구분하여 정리한다. 비교의 목적은 두 전통을 동일시하는 것이 아니라, AI 의미 공백을 해명하는 데 필요한 철학적 자원을 각각의 강점과 한계까지 포함해 분해·재구성하는 데 있다(Wittgenstein, 1953/2009; Garfield, 1995; Westerhoff, 2009).

1) 유사성: 의미의 비내재성과 관계적 발생

첫째, 두 전통 모두 의미를 기호 내부의 자족적 속성으로 보지

않고, 관계적 조건 속에서 성립하는 것으로 본다. 비트겐슈타인의 "의미=사용"은 단어의 의미가 언어게임과 삶의 형식이라는 공적 실천 속에서만 규정됨을 강조하며(Wittgenstein, 1953/2009), 중관의 연기·공 사상은 어떠한 법도 자성(自性)을 갖지 않고 조건적 결합 속에서만 성립함을 논증한다(Garfield, 1995; Siderits, 2007). 이 점에서 '의미 공백'은 단순 기술 결함이 아니라, 의미가 원칙적으로 외재적·관계적 조건에 의존한다는 구조를 드러내는 사례로 재기술될 수 있다.

둘째, 두 전통 모두 '규범·관습·맥락'의 역할을 핵심으로 둔다. 비트겐슈타인은 규칙 따르기 문제를 통해 의미의 규범성이 개인의 내적 표상이 아니라 공동체적 관행에 의해 유지됨을 부각시키며(Kripke, 1982), 불교는 언어·개념·관습이 조건망 속에서 작동하면서 의미가 성립한다는 점을 전제한다(연기). 따라서 AI 출력이 "언어처럼 보임"에도 불구하고, 의미의 주체가 될 수 없는 이유는, 의미를 안정화하는 규범적 장(공동체적 실천·조건망)에 AI가 스스로 참여하지 못한다는 점에서 공통적으로 설명된다.

셋째, 주체(자아)에 대한 비실체화 경향이 공명한다. 비트겐슈타인은 의미를 '내면의 표상'에 두는 설명을 경계하고 공적 기준을 강조하는 반면, 불교는 무아·오온의 관점에서 고정된 자아를 부정한다(Garfield, 1995; Siderits, 2007). 이 공명은 AI의 '행위자성' 논의에서, 내부에 고정된 의도 주체를 상정하는 방식 자체가 문제일 수 있음을 시사한다.

2) 차이성: 분석의 범위·방법·지향의 비등가

첫째, 분석 대상의 범위가 다르다. 비트겐슈타인의 논의는 주로 언어 의미와 규범(문법)의 작동을 기술·정리하는 데 집중하는 반

면(Wittgenstein, 1953/2009), 중관의 연기・공은 의미를 포함하되 존재・인식 전반의 무자성을 다루는 더 넓은 존재론・인식론적 논증을 수행한다(Garfield, 1995; Westerhoff, 2009). 따라서 '공=의미=사용'처럼 단순 등치하면 범주 오류가 된다. 본 논문은 공을 '언어 이론'이 아니라 '비내재성・조건성'이라는 상위 구조로 읽는 수준에서만 비교한다.

둘째, 방법과 지향이 다르다. 비트겐슈타인은 철학의 과제를 설명・이론화보다 '혼란의 치료'로 이해하는 경향이 강하고(문법적・치료적 전략), 중관은 논증을 통해 자성 집착을 해체하고 수행론적・윤리적 지평에서 세계 이해를 전환시키는 목표를 가진다(Siderits, 2007; Westerhoff, 2009). 즉, 전자는 언어 사용의 기술과 정리, 후자는 집착 해체와 해탈론적 함의를 포함한다는 점에서 지향이 비등가적이다.

셋째, 규범의 근거가 다르다. 비트겐슈타인의 규범성은 주로 언어공동체의 관행과 합의에 뿌리를 두지만, 불교는 업・의도(cetanā)・자비와 같은 윤리적 구조를 통해 규범적 판단이 확장된다(Sirimanne, 2018). 이 차이는 제Ⅳ장의 책임 논의에서 결정적으로 작동한다.

3) AI 논의에 대한 함의: '의미의 외재성'은 같고, '책임의 재구성 방식'은 다르다.

요컨대 두 전통은 AI 의미 공백을 '의미의 외재성・관계성'이라는 공통 구조로 해명하는 데 수렴하지만, 비트겐슈타인이 주로 의미 성립의 규범적 조건(삶의 형식・언어게임)을 정리하는 데 유용하다면, 불교(연기・공, cetanā)는 의미와 행위성・책임을 연기적 조건망 속에서 재분배・재구성하는 윤리적 지평을 제공한다. 본

논문은 이 유사성과 차이성을 구분해 제시함으로써, “비교가 모호하다”는 비판을 ‘비교 결과의 명시’로 해소한다.

Ⅳ. AI 결정과 도덕적 책임

앞 장에서 AI 의미 생성이 자성(自性)을 갖지 않고 연기적 조건망 속에서만 성립한다는 점을 살펴보았다. 이제 같은 관점을 책임 논의에 적용할 때, AI 결정으로 인한 피해에서 ‘누가, 무엇에 대해, 어떻게 책임을 져야 하는가’라는 문제가 새롭게 구성된다. 이때 본 장은 책임 문제를 독립적인 행위론·규범윤리 이론으로 전면 전개하기보다, 제Ⅱ–Ⅲ장에서 정식화한 ‘의미 이해의 구성적 조건’이 현존 AI에서 충족되지 않는다는 결론을 전제로, 책임 귀속이 요구하는 행위자성의 최소 조건(의도성, cetanā)을 점검하는 범위에서 논의를 제한한다. 본 장은 AI 책임 문제를 독립적인 윤리학·법철학 논쟁으로 전면 전개하지 않는다. 여기서의 목표는 (1) 제Ⅱ–Ⅲ장에서 도출한 의미 공백·행위성 한계가 ‘현존 AI의 책임 보유 주체성’을 배제한다는 점을 확인하고, (2) 그 결과 책임이 인간 행위자와 제도 설계의 조건망으로 귀속되어야 한다는 최소 원리를 제시하는 데 한정된다. 배상·처벌·규제와 같은 가치론적·정책적 설계의 상세화는 후속 연구 과제로 유보한다.

1. 책임의 간극(responsibility gap) 문제

AI가 자율주행 사고, 채용·보석(가석방) 알고리즘의 불공정 같

은 피해를 야기할 때 누가 책임을 지는가? 이는 현대 윤리·법에서 "책임의 간극(responsibility gap)"으로 논의되어 왔다(Matthias, 2004). 지배적 견해는 현재의 AI는 도덕 행위자(moral agent)가 아니며 법적·도덕적 책임을 질 수 없다는 것이다(Tigard, 2021; Bryson, 2018). 인간 설계자·운영자·기관 등에게 책임이 귀속되어야 한다(Johnson, 2009). 오늘날 가장 진보한 시스템도 행위의 도덕적 의미를 이해하거나 의도·의식을 갖추지 못했다(Searle, 1980; Dreyfus, 1992). 윤리 논의에서도, 설령 편의상 AI를 "에이전트(agent, 행위자)"라 부르더라도 의식과 진정한 자율성이 결여된 한 "책임 귀속의 적격자"가 될 수 없다는 점이 반복해서 강조된다(Tigard, 2021; Bryson, 2018). 카메라에 과속을 '탓'하지 않는 것과 같은 이치다(Danaher, 2016). 문제는 AI의 결정 경로가 복잡하고 불투명하다는 점이다(Marcus, 2020; Bender et al., 2021). 결과에는 개발자, 데이터 제공자, 사용자 등 인간 행위자와, 이를 둘러싼 제도·규제 구조가 얽혀 있다(Johnson, 2009; Crawford, 2021). 이에 따라 AI는 '책임의 대상'이 아니라 '책임을 재분배해야 하는 문제의 출발점'으로 간주된다(Bryson, 2018). 특히 AI가 인간 의사결정의 일부를 대체하면서, 기존 법·윤리 체계가 상정했던 책임 모델이 더 이상 충분하지 않다는 점이 명확해졌다(Matthias, 2004; Elish, 2019). 책임의 간극은 기술의 복잡성에서 오는 불가해성과, 인간-기계 행위가 얽히는 다층적 사고 구조에서 발생하며(Danaher, 2016; Crawford, 2021), 이는 새로운 철학적 해석틀의 필요성을 제기한다(Kuhn, 1962; Ladyman et al., 2007). 예컨대 한국의 대형 플랫폼 기업들이 활용한 AI 채용 필터는 특정 대학·연령·언어 패턴을 가진 지원자를 유리하게 선별하며, '알고리즘 차별'을 일으켰다는 지적을 꾸준히 받아왔다(박태우 2020; 홍진수 2021; 이지윤 외

2023). 국가인권위원회의 연구보고서(2021) 역시 AI 채용 시스템이 학력・지역・언어 패턴 등의 기존 편향을 그대로 학습해 구조적 차별을 강화할 위험이 크다고 분석하며, “AI가 평가했다”는 기업의 주장으로 책임이 면제되지 않는다는 점을 명확히 한다. 실제로 채용 알고리즘이 작동하는 전 과정－데이터 구성, 가중치 설계, 면접 스크립트 작성, 결과 해석－은 모두 인간의 선택이 개입한 조건의 산물이며, 알고리즘은 그 조건을 기계적으로 재현할 뿐이다. 불교적 관점에서 보면, 이는 단일 행위자의 책임이라기보다 연기적 조건의 결합 속에서 발생한 구조적 결과이며, 책임 또한 그 조건을 형성한 인간・제도・조직에 분배되어야 한다.

2. 불교 윤리에서의 의도(cetanā)

불교 윤리는 행위의 도덕적 성격을 결정하는 요인을 의도(cetanā)와 인과적 책임으로 본다(Sirimanne, 2018; Siderits, 2007). Sirimanne(2018)은 이러한 의도 개념을 “불교에서 의도를 곧 업(karma)이라 부른다”는 초기 경전의 가르침으로부터 해석하며, 모든 도덕적 판단은 행위의 결과가 아니라 의도의 구조에 달려 있다고 지적한다. 이에 따르면 인공지능은 의식적 의도나 도덕적 숙고 능력이 없기 때문에 도덕적 행위자가 될 수 없으며(Searle, 1980; Dreyfus, 1992), 책임은 오롯이 의도를 가진 인간(설계자, 사용자, 사회 제도)에 귀속된다(Sirimanne, 2018; Johnson, 2009). 불교적 관점에서 보면, AI의 윤리 문제는 인공지능 자체의 문제가 아니라 ‘무의도적 시스템에 의도를 투사하는 인간의 집착’ 문제로 해석될 수 있다(Siderits, 2007; Westerhoff, 2009). 즉, 문제의 핵심은 AI의 결함이 아니라 ‘AI가 의도를 가진 것처럼 착각하는 우리의 해석

방식'이며, 불교 윤리는 이러한 투사의 위험성을 반복적으로 경고한다(Sirimanne, 2018).

1) 의도의 구조와 업(karma)의 발생

불교에서는 행위의 결과보다 행위자의 마음상태－욕망・집착・무지－가 행위의 도덕적 질을 규정한다(Sirimanne, 2018; Siderits, 2007). 의도가 선하면 결과가 우연히 해롭더라도 업의 성격은 선한 것으로 분류된다. 반대로, 의도가 악하면 겉보기에는 규범을 지킨 행동일지라도 업은 악한 것으로 남는다. 이 관점에서 보면, AI는 의도를 가질 수 없기 때문에 "업의 주체"가 될 수도 없다(Searle, 1980; Tigard, 2021). 업(karma)은 행위자의 도덕적 의지에서 발생하는 것이고, AI는 의식・의도・지향성이 없으므로 업을 생성하지 않는다(Sirimanne, 2018). 따라서 AI 행동으로 인한 피해의 업적 귀속은 전적으로 AI를 설계하고 사용한 인간에게 돌아간다(Johnson, 2009; Bryson, 2018).

2) 무의도적 시스템에 대한 책임 판단

동시에 불교는 연기의 복잡성을 인정한다(Garfield, 1995; Ames, 2003). 사고는 다수 조건의 결합으로 생긴다. 이는 버그・불량 데이터・인간 감시 실패 등 복합 요인으로 생기는 AI 사고의 현대적 분석과 상응한다(Crawford, 2021; Marcus, 2020). 그렇더라도 불교는 개인적 책임 자체를 부정하지 않는다(Siderits, 2007; Sirimanne, 2018). 가령 허위정보를 증폭해 사회 혼란을 유발할 것을 알면서 추천 시스템을 설계・배포했다면, 그 의도와 예견 가능한 결과에 대해 도덕적 책임을 진다(Johnson, 2009; Bryson, 2018). 칼을 탓하지 않듯, 비의식적 알고리즘에 비난을 돌릴 수는 없다(Searle, 1980;

Dreyfus, 1992). 즉, 불교 윤리의 책임 판단은 “무엇이 직접 행위를 수행했는가?”가 아니라 “그 행위가 어떤 의도와 어떤 조건에서 비롯되었는가?”에 달려 있다(Siderits, 2007; Sirimanne, 2018). 이 기준은 복잡한 AI 시스템에도 일관되게 적용될 수 있다(Tigard, 2021; Johnson, 2009). 따라서 AI 행동의 도덕적 책임을 기계 자체에게 돌리는 것은 논리적·윤리적 오류이며, 불교적 관점에서는 이러한 의인화/주체화가 자아를 투사해 붙잡는 ‘아집(我執, ātma grāha)’의 한 양상으로 이해될 수 있다(Westerhoff 2009; Garfield 1995).

한국의 포털 및 소셜미디어에서 추천 알고리즘이 특정 뉴스·게시물을 반복적으로 노출함으로써 혐오 표현과 정치적 양극화를 심화시킨다는 문제는 이미 여러 공공기관과 연구에서 제기되어 왔다(KISO 인터넷자율정책기구, 2023; 국가인권위원회·한국유네스코위원회, 2022; 조진형·김규정, 2022; 김경달, 2025). 조진형·김규정(2022)은 소셜미디어 추천 시스템이 이용자의 기존 성향과 정서 반응에 맞춰 정보를 선별하는 과정에서 에코챔버와 필터버블을 형성해, 유사한 의견과 정서를 반복적으로 강화한다고 분석한다. 김경달(2025)은 알고리즘이 이용자의 체류 시간과 상호작용을 극대화하는 방향으로 설계되어 있어, 자극적이고 감정적 반응을 유도하는 콘텐츠가 구조적으로 유리한 위치를 점하게 된다고 지적한다. KISO의 혐오표현 자율정책 가이드라인과 관련 논의 역시, 혐오표현이 플랫폼 구조 속에서 쉽게 확산·증폭되는 위험을 경고하며 사업자의 자율규제와 사회적 책임을 강조한다(KISO 인터넷자율정책기구, 2023; 김민정, 2023). 이러한 현상은 알고리즘 자체에 ‘혐오하려는 의도’가 있기 때문이 아니라, 클릭과 시청을 최우선 목표로 하는 설계가 사용자의 분노·혐오·공포를 자극하는 콘텐츠를 반복적으로 선택하는 구조를 만들어 내기 때문이

다. 불교의 cetanā 관점에서 보면, 문제의 핵심은 코드 그 자체라기보다 이러한 설계를 가능하게 한 '주의 · 욕망 · 분노'를 중시하는 인간적 조건의 연합이며, 그 조건을 설계 · 허용 · 강화한 인간 · 플랫폼 · 제도에 도덕적 책임이 분배되어야 한다. 다시 말해, 추천 알고리즘이 빚어내는 혐오와 양극화의 문제는 연기적 조건들의 결합이 낳은 사회적 결과로 이해할 수 있다.

3. AI 피해 사례와 책임 귀속

더 많은 결정을 AI에 위임할수록 책임의 간극은 넓어진다는 비판이 제기되며(Matthias, 2004; Elish, 2019), 책임 · 배상의 새로운 프레임이 요구된다(Johnson, 2009; Crawford, 2021). 흥미롭게도 사람들은 복잡한 시스템일수록 AI에게 칭찬 · 비난을 부여하려는 심리적 경향을 보인다(Danaher, 2016). 그러나 책임의 귀속(歸屬, attribution of responsibility)과 책임의 보유(保有, possession of responsibility)는 다르다(Tigard, 2021). 예를 들어, 알고리즘 편향으로 인한 차별적 결과는 설계자의 무지나 특정 변수 선택에 기반하지만, 인간은 직관적으로 'AI가 나쁜 결정을 했다'고 말한다. 이는 심리적 의인화에서 비롯되며(Buber, 1923), 불교가 경계하는 잘못된 집착의 한 형태다(Siderits, 2007; Sirimanne, 2018). 따라서 책임 귀속의 문제는 행위의 표면이 아니라 그 배후의 조건 · 의도 · 설계의 흐름을 추적해야 해결된다(Johnson, 2009; Ladyman et al., 2007).

한국의 자율주행차 사고 논의에서는 실제 사고 사례와 시뮬레이션을 통해, 차량 제조사 · 소프트웨어 및 센서 공급 업체 · 정밀지도 · 통신 인프라 제공자 · 운전자 · 보험사 · 국가 규제 기관 등 다수의 행위자가 동시에 얽히면서 책임 소재가 모호해진다는 점이

반복해서 지적되어 왔다(한국교통연구원, 2020; 황현아 & 손민숙, 2023). 예를 들어 한국교통연구원(2020)의 카드뉴스와 관련 조사에서는, 자율주행 3단계 이상에서 "운행 책임이 분산되어 사고 원인 규명이 어렵고, 사고 발생 시 책임 소재가 모호하다"는 인식이 시민들의 주요 우려 사항으로 나타난다. 보험연구원의 연구보고서 역시 레벨 4 자율주행차 사고가 전통적인 자동차 사고를 넘어 제조물・소프트웨어・인공지능 시스템의 결함, 도로・통신 인프라의 문제까지 겹치는 복합적 사건으로서, 복수의 관련 당사자들 사이에서 책임을 어떻게 분배할 것인지가 핵심 과제라고 분석한다(황현아 & 손민숙, 2023). 불교의 연기론에 따르면, 어떠한 결과도 단일 원인으로 환원될 수 없고, 상호 의존하는 조건들의 네트워크 속에서만 이해될 수 있다(장승희, 2014). 이러한 관점을 자율주행차 사고에 적용하면, 특정 행위자 하나를 "궁극적 원인"으로 지목하기보다, 기술・제도・인간 행위가 얽힌 조건들의 상호작용 속에서 책임을 분배해야 한다는 규범적 기준이 도출된다. 이는 자율주행차 책임법제 및 보험제도를 설계하는 한국 사회의 정책 논의에서, 사고 책임을 다층적・연기적 구조로 파악하도록 이끄는 중요한 철학적・윤리적 틀로 기능할 수 있다.

4. 연기적 조건망에서의 인간 행위와 책임

제도・법・전문윤리 차원에서 책임 배분을 명료히 하고, 의도와 주의를 갖춘 선한 설계(올바른 의지・행위)를 지향해야 한다(Coeckelbergh, 2020; Johnson, 2009). 불교적 관점에서 진정한 책임은 결과의 통제가 아니라, 매 순간의 의도(cetanā)를 정화하고 바르게 설정하려는 마음에서 비롯된다(Sirimanne, 2018). 불교의 연기론

은 책임을 “단선적 귀속”이 아니라 “조건적 연계”로 이해하게 한다(Garfield, 1995; Ames, 2003). 즉, AI 사고는 단일 주체의 잘못이 아니라 설계・사용・규제・데이터・사회문화적 환경이라는 연속된 조건의 흐름 속에서 발생한다(Crawford, 2021; Johnson, 2009). 그러므로 불교적 책임 개념은 ‘원인 찾기’보다 ‘조건 정비’를 우선하며(Siderits, 2007; Sirimanne, 2018), 이는 AI 윤리 설계에서 가장 현실적이고 효과적인 방향성을 제공한다(Coeckelbergh, 2020; Bryson, 2018).

특히 조계종을 포함한 한국 불교계는 최근 AI 포교 콘텐츠, 스님 챗봇, 딥페이크・추천 알고리즘의 확산이 신행과 재가 공동체에 미칠 영향을 주제로 관련 쟁점이 대중 담론을 넘어 학술 연구의 형태로도 정리되기 시작했다(보일(양성철)), 2022; 불교평론 편집부, 2022).

예컨대 보일(양성철, 2022)은 ‘디지털 휴먼’ 기술을 대상으로, 불교의 업(業)・의도(cetanā) 및 선교방편(善巧方便) 개념을 기준으로 기술이 초래할 수 있는 역기능과 순기능을 함께 검토하며, 기술 자체의 중립성보다 ‘사용 조건(동기・맥락・책임 구조)’의 설계가 윤리적 평가의 핵심임을 논한다. 또한 보일(양성철, 2023)은 원효의 열반관을 토대로 포스트휴머니즘의 ‘탈신체성’ 논제를 분석함으로써, 본 논문이 다루는 ‘비체화된 정보처리’와 ‘행위성・책임 귀속’ 논점이 불교의 인간관・해탈관과 교차하는 철학적 접점을 제공한다. 더 나아가 『불교평론』 2022년 가을호(통권 91호) ‘포스트휴먼 시대의 도래와 불교’ 특집은 인간-기계-생명 관계 재구성의 주요 쟁점을 불교적 관점에서 폭넓게 논의하며, 기술윤리 논의가 불교계 내부에서도 학술적으로 전개되고 있음을 보여준다(불교평론 편집부, 2022). 헝라다롬(2022)은 불교 윤리의 실천 규범(계・정・혜, 자비, 연기)을 준거로 AI・로봇 윤리의 재정식을 모

색한다. 이는 AI를 단순히 배척하거나 신격화하기보다는 인간의 의도와 관심, 수행 상태의 정화(cetanā śuddhi)를 중심에 두고, 디지털 기술의 사회적 책임 구조를 재구조화하려는 불교적 시도로 이해할 수 있다. 요컨대, 불교적 책임론은 AI를 새로운 '죄수'로 세우는 대신, 연기적 조건망 속에서 인간 의도와 제도 설계를 재평가하고 정비하는 일 자체를 도덕적·정치적 과제로 제시한다.

Ⅴ. 결론

1. 연구 요약

본 논문은 현대 인공지능의 의미 이해 불가능성과 행위성의 구조적 한계를 분석하기 위해, 기술철학·언어철학·불교철학의 세 관점을 교차시켜 고찰하였다. 딥러닝 기반 AI는 방대한 데이터를 학습하지만, 내부 작동은 여전히 통계적 패턴에 근거한 구문론적 처리(syntactic manipulation)에 머무르며 의미론적 이해(semantic understanding)에 도달하지 못한다(Searle, 1980; Harnad, 1990; Bender et al., 2021). 이는 의미가 규칙이나 형식에 고정되어 있지 않고, 사용·맥락·삶의 형식에서 발생한다는 후기 비트겐슈타인적 전환을 다시 확인하게 한다(Wittgenstein, 1953/2009; Kripke, 1982). 나아가 이러한 의미의 비본질성은 불교의 연기·공 사상과도 깊이 상응한다. 모든 법(法)은 자성을 갖지 않고 관계적 조건에서만 성립한다는 중관학의 통찰(Garfield, 1995; Ames, 2003)은, AI 출력의 의미 역시 기계 내부가 아니라 인간 해석자·데이터 구

성・사회적 맥락이라는 연기적 그물망 속에서 발생한다는 사실을 철학적으로 조명하고 해석할 수 있게 한다. 또한 본 논문은 AI의 행위성 문제와 책임 귀속을 분석하며, 불교 윤리가 의도(cetanā)를 중심에 두고 도덕적 책임을 규정한다는 점을 밝히고, 현재의 AI가 도덕적 행위자로 인정될 수 없는 구조적 이유를 제시하였다 (Tigard, 2021; Sirimanne, 2018). 이러한 분석을 통해, AI・인간・사회가 맺는 상호작용이 단일 원인・단일 주체로 환원되지 않는 조건 발생적 구조임을 확인하였고(Ladyman et al., 2007; Thompson, 2007), 기술 발전이 인간 존재의 의미・책임・가치 판단에 어떤 재구성을 요구하는지를 논의의 기반을 마련하였다. 요컨대 본 논문은 (i) 의미 이해의 구성적 조건 진단(Ⅱ–Ⅲ) → (ii) 그 결론을 전제로 한 행위자성 최소 조건 점검(Ⅳ) → (iii) 책임 귀속에 관한 최소 원리 제시(Ⅴ)라는 논증 사슬을 취한다. 따라서 구성적 차원에서의 의미 분석과 행위론적 차원에서의 책임 논의는 구분되지만, ‘행위자성’이라는 매개를 통해 논리적으로 연결된다.

2. 철학・불교학적 의의

본 논문의 비교 분석 결과는 다음과 같이 정리된다. (유사성) ① 두 전통 모두 의미를 기호 내부의 자족적 속성으로 보지 않고 사용・조건・관계 속에서 성립하는 것으로 본다(Wittgenstein 1953/2009; Garfield 1995). ② 의미는 규범과 실천에 의해 안정화되며, 그 규범적 장에 참여하지 못하는 체계는 의미를 ‘이해’한다고 보기 어렵다(Kripke 1982). (차이성) ① 비트겐슈타인은 언어 규범의 기술과 철학적 혼란의 치료라는 문법적 전략을 취하는 반면, 중관의 연기・공은 존재・인식 전반의 무자성을 논증하고, 수행론

적・윤리적 함의를 포함한다(Siderits 2007; Westerhoff 2009). ② 비트겐슈타인의 규범성은 공동체적 관행에 주로 근거하지만, 불교 윤리는 의도(cetanā)・업・자비의 구조를 통해 책임 논의를 확장한다(Sirimanne 2018). 이 구분을 통해 본 논문은 '비교의 기준・방법・결과(유사/차이)'를 명시적으로 제시한다.

불교와 서양 분석철학 모두 현재의 AI를 독립적 도덕 행위자로 보지 않으며(Matthias, 2004; Tigard, 2021; Sirimanne, 2018), 피해의 책임은 데이터・설계・사용・제도라는 복합적 조건망 속 인간의 의도와 선택으로 귀속된다고 본다. 본 연구는 이러한 책임 판단을 불교적 cetanā 개념과 연기론을 활용해 정교하게 재구성하였다. 불교는 행위의 결과보다 행위자의 동기・의도를 도덕 평가의 핵심으로 삼고(Siderits, 2007), 이는 AI 윤리에서 요구되는 "인간의 책임 회피 방지"라는 핵심 원칙과 긴밀하게 연결된다. 동시에 연기론은 AI 사고의 원인을 단일 주체로 귀속시키는 방식 대신, 조건의 연속적 흐름을 분석해야 한다는 점을 강조한다.

불교의 연기・공 사상은 AI의 의미 생성 과정을 비판적으로 조망하고, 관계적・비본질적 의미 형성을 비유적으로 이해할 수 있도록 하는 해석적 관점을 제공한다(Garfield, 1995; Ames, 2003; Westerhoff, 2009). 이러한 관점은 기술철학에서 전통적으로 충분히 다루어지지 않았던 '의미의 비본질성'과 '행위성의 조건성'을 보다 선명하게 드러내는 데 기여한다. 기술철학적 차원에서도 본 연구는 의미가 고정된 실체에 의해 결정되는 것이 아니라, 사용・맥락・관습・해석 과정 속에서 발생한다는 점을 재확인한다. 이때 AI의 의미 생성 구조가 불교의 무자성・상호의존 개념과 부분적으로 교차할 수 있다는 점은, 동서 사유 모두가 "의미는 자성적으로 주어지지 않고 관계적으로 발생한다"는 통찰에 수렴함을 보여

준다. 이러한 비교철학적 통찰은 AI 시대의 책임·판단·행위성 문제를 단순 기술적 오류가 아니라 인간 존재 조건과 사회적 맥락 전체를 고려해야 할 문제로 재구성할 수 있게 하며, 기술철학·윤리학·불교학 간의 생산적 대화를 가능하게 한다.

특히 이러한 분석은, 최근 조계종과 한국 불교계가 AI 포교 콘텐츠·스님 챗봇·딥페이크·추천 알고리즘 등 디지털 기술의 사회적 영향을 둘러싸고 전개해 온 논의와 직접 맞닿아 있다. 본 논문이 제시한 연기·공·cetanā에 기초한 책임 개념은, 불교계가 'AI를 어떻게 활용할 것인가'라는 차원을 넘어, 어떤 조건과 의도 속에서 기술을 설계·운영해야 하는지에 대한 규범적 기준을 제공한다. 이는 불교계 내부의 '불교 AI 윤리 선언' 논의뿐 아니라, 국가·시민사회와의 공적 대화에서 불교가 제시할 수 있는 고유한 관점을 이론적으로 뒷받침한다는 점에서 의의를 지닌다.

3. AI 시대 의미·행위성·책임에 대한 제언

본 절의 제언은 제Ⅱ–Ⅳ장의 분석에서 직접 도출되는 최소한의 규범적 함의(책임의 부당한 주체화 방지, 책임의 인간·제도 환류, 조건 정비 중심의 설계 원리)만을 정리한다. AI 책임성의 가치론적 정당화 및 법·정책 설계(책임 배분, 배상 모델, 처벌 정당화 등)는 별도의 후속 논문에서 심층적으로 다룰 과제로 유보한다. 우선, 이 절의 1부는 전적으로 '현존 AI'에 대한 논의를 정리한다. 현 단계의 AI는 고도화된 연산 능력을 갖추었으나, 의미를 이해하거나 자기 자신을 성찰하는 능력—즉 존재론적 자기지시—에는 이르지 못한다(Searle, 1980; Dreyfus, 1992). 현존 AI는 여전히 "도구적 지능"에 머물러 있으며, 그 행위는 인간이 설계한 조건과 알고리

즘, 그리고 사회적 맥락의 산물이다. 따라서 AI의 출력에 부당한 자율성을 부여하거나 도덕적 지위를 투사하는 것은 잘못된 주체화이며, 이는 불교적 관점에서 집착(我執, ātma-grāha)의 한 형태로 해석될 수 있다. 여기까지의 결론은, Ⅳ장에서 논의한 바와 같이, 현재 사용되는 AI가 도덕적 행위자로 인정될 수 없다는 입장을 재확인한다. 요컨대 본 논문의 핵심 결론은, 현존 AI에 도덕적 책임을 부과하기보다, 불교적 연기 · cetanā 틀 안에서 인간과 제도에 책임을 재분배해야 한다는 점이다.

미래의 의식적 AI 주체성(체화 · 자기성찰 등)을 전제하는 사변적 논의는 본 논문의 분석 범위를 넘어서는 주제이므로, 후속 연구 과제로만 유보한다.

다만, 이러한 사변적 가능성은 다시 현존 AI의 윤리 문제와는 구분된다. 현재의 과제는 명확하다. AI의 능력 · 위험 · 잠재력에 대해 과도한 공포나 과신을 경계하면서, 의도(cetanā)의 정화와 사회적 관계 조건의 재설계라는 불교적 지혜를 기술 윤리에 적용하는 것이다. 기술의 정교화나 자동화의 속도보다 중요한 것은, AI가 작동하는 사회적 · 윤리적 장(field)을 바르게 형성하는 일이며, 투명성 · 공정성 · 책임성 같은 기본 원칙을 제도적 설계와 기술적 구현에 일관되게 반영해야 한다. 예를 들어, 불교계의 AI 활용 논의가 확대되는 현실을 고려하면 연기 · cetanā에 기초한 점검 기준(가이드라인 · 심사 절차)의 필요성이 제기될 수 있다. 다만 그 구체적 항목화와 제도화는 사례 연구와 규범윤리 분석을 요구하므로, 후속 연구에서 심화할 과제로 남겨 둔다. 우리가 설정한 의도가 선하면, 그 기술이 만들어내는 결과 역시 바른 방향으로 흐를 수 있다. 궁극적으로, AI는 인간의 존재 · 의도 · 책임을 비추는 하나의 거울이다. AI가 이해하지 못하는 세계를 인간이 잃지 않도록—그리고 기술이

드러내는 조건 발생적(緣起) 구조를 성찰하며—미래의 윤리·사회·불교적 사유를 확장해 가는 것이 우리에게 남겨진 과제이다.

참고문헌

국가인권위원회. 2021.『인공지능(AI) 개발과 활용에서의 인권 가이드라인 연구(최종보고서)』.

국가인권위원회·한국유네스코위원회. 2022.『사이버 공간에서 차별·혐오 대응과 미디어 리터러시 교육』. 국가인권위원회·한국유네스코위원회.

김경달. 2025. "알고리즘의 양면성… 편리함의 그늘 속 균형점을 찾아서."『나라경제』. 2025(5): 30-35. 한국개발연구원(KDI).

김민정. 2023. "KISO 〈혐오표현 자율정책 가이드라인〉의 제정 의의와 몇 가지 쟁점."『정보법학』 27(2): 1-30.

보일(양성철). 2022. "디지털 휴먼에 대한 불교적 관점 – 악업의 증장인가, 선교방편(善巧方便)인가 –."『宗學硏究(종학연구)』. 7: 5-33.

보일(양성철). 2023. "원효의 열반관으로 본 포스트휴머니즘의 '탈신체성' 연구:『열반종요(涅槃宗要)』를 중심으로." 서울대학교 박사학위논문.

보일(양성철). 2025. "인공지능 스님의 탄생은 가능한가?"『불교문화』, 2025.05.26.

불교평론 편집부. 2022.『불교평론』 통권 91호(2022년 가을호): "포스트휴먼 시대의 도래와 불교" [특집]. 서울: 불교시대사.

이승종. 2002(2014).『비트겐슈타인이 살아 있다면: 논리철학적 탐구』. 문학과지성사.

이승종. 2024.『역사적 분석철학』. 서강대학교출판부.

이지윤·김현수·이기욱. 2023. "'AI 채용' 차별 논란에… 뉴욕 '성별-인종 편향 공개하라' 첫 규제."『동아일보』, 2023.07.07.

장승희. 2014. "불교 연기론의 관점에서 본 통일문제와 통일교육."『도덕윤리과교육』. 44: 255-286.

조진형·김규정. 2022. "소셜미디어에서 에코챔버에 의한 필터버블 현상 개선 방안

연구.” 『한국콘텐츠학회논문지』. 22(5): 1-10.

허남결. 2024a. “인공지능(AI)과 자비윤리 ①.” 『고경』 제136호: 49-55. 서울: 성철사상연구원.

허남결. 2024b. “인공지능(AI)과 자비윤리 ②.” 『고경』 제137호: 80-88. 서울: 성철사상연구원.

황현아 · 손민숙. 2023. 『자율주행차사고 책임법제 및 보험제도: 레벨4 주요국 제도 비교를 중심으로』(연구보고서 2023-02). 보험연구원.

박태우. 2020. “AI 면접관이 말했다 ‘너 인성 문제 있어?’” 〈한겨레신문〉 10.27.

한국교통연구원. 2020. “자율주행차 사고, 누가 책임져야 하나요? 카드뉴스.” 한국교통연구원. https://www.koti.re.kr/user/bbs/BD_selectBbs.do?q_bbsCode=1082&q_bbscttSn=20200409095730728

한국인터넷자율정책기구. 2023. 『혐오표현 자율정책 가이드라인』. 한국인터넷자율정책기구(KISO).

헝라다롬, 소랏(Hongladarom, Soraj). 2022. 『불교의 시각에서 본 AI와 로봇 윤리』. 김근배 · 김진선 · 주은혜 · 허남결 역. 씨아이알(CIR).

홍진수. 2021. “인간이 낳은 AI…객관 · 공정성을 기대하는 것은 환상.” 〈경향신문〉, 01.17. https://www.khan.co.kr/article/202101172054015

Ames, W. L. 2003. “Emptiness and Quantum Theory.” In B. A. Wallace, ed., *Buddhism and Science: Breaking New Ground*, 285-302. New York: Columbia University Press.

Baker, G. P., and P. M. S. Hacker. 2009. *Wittgenstein: Meaning and Mind, Part I: Essays*. Oxford: Wiley-Blackwell.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. Black, M. Bombardier, et al. 2021.“On the Opportunities and Risks of Foundation Models.”

Stanford Center for Research on Foundation Models Report. Stanford, CA:

Stanford University.

Brandom, R. B. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems* 33, 1877-1901.

Bryson, J. J. 2018. "Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20 (1): 15-26. https://doi.org/10.1007/s10676-018-9448-6

Buber, M. 1923. *Ich und Du*. Leipzig: Insel Verlag.

Buber, M. 1970. *I and Thou.* Translated by W. Kaufmann. New York: Scribner. (Original work published 1923)

Chalmers, D. J. 2022. *Reality+: Virtual Worlds and the Problems of Philosophy.* New York: W. W. Norton.

Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Coeckelbergh, M. 2010. "Robot Rights? Towards a Social-Relational Justification of Moral Consideration." *Ethics and Information Technology* 12(3): 209-221. https://doi.org/10.1007/s10676-010-9235-5

Coeckelbergh, M. 2020. *AI Ethics.* Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12549.001.0001

Crawford, K. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press. https://doi.org/10.12987/9780300252392

Danaher, J. 2016. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18(4): 299-309. https://doi.org/10.1007/s10676-016-9403-3

Dennett, D. C. 1991. *Consciousness Explained.* Boston: Little, Brown.

Dreyfus, H. L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason.* Rev. ed. Cambridge, MA: MIT Press.

Elish, M. C. 2019. "Moral Crumple Zones: Cautionary Tales in Human-

Robot Interaction." *Engaging Science, Technology, and Society* 5: 40-60. https://doi.org/10.17351/ests2019.260

Floridi, L. 2004. "Open Problems in the Philosophy of Information." *Metaphilosophy* 35(4): 554-582. https://doi.org/10.1111/j.1467-9973.2004.00336.x

Floridi, L. 2011. *The Philosophy of Information.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199232383.001.0001

Gallagher, S. 2005. *How the Body Shapes the Mind.* Oxford: Oxford University Press.

Garfield, J. L. 1995. *The Fundamental Wisdom of the Middle Way: Nāgārjuna's Mūlamadhyamakakārikā.* New York: Oxford University Press. https://doi.org/10.1093/oso/9780195103175.001.0001

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning.* Cambridge, MA: MIT Press.

Harnad, S. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42(1-3): 335-346.

Johnson, D. G. 2009. *Computer Ethics: Analyzing Information Technology*. 4th ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Kripke, S. A. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Ladyman, J., D. Ross, D. Spurrett, and J. Collier. 2007. *Every Thing Must Go: Metaphysics Naturalized.* Oxford: Oxford University Press.

Lopez, D. S., ed. 1995. *Curators of the Buddha: The Study of Buddhism under Colonialism.* Chicago: University of Chicago Press.

Marcus, G. 2020. "The Next Decade in AI: Four Steps towards Robust Artificial Intelligence." *arXiv*: 2002.06177. https://arxiv.org/abs/2002.06177

Matthias, A. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6(3): 175-183. https://doi.org/10.1007/s10676-004-3422-1

McDowell, J. 1994. *Mind and World.* Cambridge, MA: Harvard University Press.

Mitchell, M. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." arXiv preprint *arXiv*:2203.02155.

Putnam, H. 1975. "The Meaning of 'Meaning'." In K. Gunderson, ed., *Language, Mind, and Knowledge*, 131–193. Minneapolis: University of Minnesota Press.

Quine, W. V. O. 1951. "Two Dogmas of Empiricism." *The Philosophical Review* 60(1): 20–43. https://doi.org/10.2307/2181906

Rini, R. 2020. "Deepfakes and the Epistemic Backstop." *Philosophers' Imprint* 20(24): 1–16.

Searle, J. R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3(3): 417–457.

Siderits, M. 2007. *Buddhism as Philosophy: An Introduction*. Aldershot: Ashgate. https://doi.org/10.4324/9781315261225

Sirimanne, C. R. 2018. "The Unique Perspective on Intention (Cetanā), Ethics, Agency, and the Self in Buddhism." In B. Grant, J. Drew, and H. E. Christensen, eds., *Applied Ethics in the Fractured State* (Research in Ethical Issues in Organizations, Vol. 20), 25–44. Bingley: Emerald.

Thompson, E. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge, MA: Harvard University Press.

Tigard, D. W. 2021. "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible." *Cambridge Quarterly of Healthcare Ethics* 30(3): 435–447.

Varela, F. J., E. Thompson, and E. Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge, MA: MIT Press.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,

Ł. Kaiser, and I. Polosukhin. 2017. “Attention Is All You Need.” In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 5998–6008. Long Beach, CA: Curran Associates, Inc.

Westerhoff, J. 2009. *Nagarjuna's Madhyamaka: A Philosophical Introduction*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195375213.001.0001

Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*. Translated by C. K. Ogden. London: Routledge & Kegan Paul.

Wittgenstein, L. 1953(2009). *Philosophical Investigations*. 4th rev. ed. Translated by G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte. Oxford: Wiley-Blackwell.

(논문 접수 : 2025.11.28. / 수정본 접수 : 2025.12.28. / 게재 승인 : 2025.12.29.)

ABSTRACT

The Semantic Gap and Limits of AI Agency:
Buddhist Dependent Origination, Emptiness, and Wittgenstein's Philosophy of Language

Ji, Kyung-sun
Ph.D. Candidate, Department of Philosophy, Ewha Womans University
Ph.D. in Medicine, Graduate School, CHA University of Science and Medicine

This paper investigates the limits of artificial intelligence (AI) with respect to semantic understanding and moral agency by bringing Buddhist philosophy and Wittgenstein's philosophy of language into a comparative framework. Contemporary deep-learning systems learn by optimizing large parameter spaces through supervised learning, backpropagation, and gradient descent. Although such systems generate highly fluent outputs, their operations remain at the level of syntactic pattern processing rather than genuine semantic understanding. Searle's "Chinese Room" argument makes this gap explicit: rule-based symbol manipulation, however sophisticated, does not amount to understanding. This finding resonates with Wittgenstein's later view that meaning is not an intrinsic property of signs but arises from use, practice, and forms of life.

Buddhist thought, particularly the doctrines of dependent origination (pratītyasamutpāda) and emptiness (śūnyatā), deepens this analysis. According to Madhyamaka philosophy, all phenomena—including linguistic meanings—are empty of intrinsic nature and arise only in relational networks of conditions. From this perspective, AI outputs do not possess meaning in themselves; rather, their meaning originates in a contingent nexus of training data, model architecture, human interpreters, and socio-linguistic contexts. The paper then examines the "responsibility gap" in AI decision-making through Buddhist ethics, arguing that intentionality (cetanā) is the basis of moral responsibility. Because current AI systems lack consciousness and intention, they cannot be moral agents; responsibility must be ascribed to the humans and institutions that design, deploy, and regulate these systems.

In conclusion, the study briefly considers, as a purely speculative horizon, whether a future "Robo-Dasein" equipped with embodiment, second-person relationality, and self-reflection could transform these assessments. Overall, the paper proposes a relational, cross-cultural framework for rethinking meaning, agency, and responsibility in the age of AI, integrating insights from analytic philosophy, Buddhist thought, and the philosophy of technology.

Key Words: Artificial Intelligence, Semantic Gap, Agency, Dependent Origination, Emptiness