

제8회 반야불교문화연구원 학술대회

AI, 축복인가 재앙인가

AI시대의 명암과 불교적 진단

일시_ 2023년 7월 2일(일요일) 13:00

장소_ 통도사 반야암

주최_ (사)반야불교문화연구원

일 정

제1부 | | | 개최식

사회 : 김 희 (신라대)

13:00~13:30 삼 귀 의
 개 회 사 지안스님 (반야불교문화연구원 원장)
 축 사 김 성 태 (반야불교문화연구원 이사장)

제2부 | | | 학술대회

사회 : 김명우 (동의대)

13:30~14:10 제1주제 인공지능 시대, 체화된 마음과 불교
 발표: 이영의(고려대)
 논평: 윤종갑(동아대)

14:10~14:50 제2주제 AI GPT와 불교의 지혜(깨달음)
 발표: 이상헌(서강대)
 논평: 일윤스님(경희대)

14:50~15:10 중간 휴식

사회 : 구자상 (동의대)

15:10~15:50 제3주제 AI와 불교논리학
 발표: 지승도(한국항공대)
 논평: 안환기(서울불교대학원대)

15:50~16:30 제4주제 인간의 미래, 불교의 미래
 발표: 허남결(동국대)
 논평: 보일스님(해인사승가대학)

16:30~17:30 종합토론

좌장 : 강경구 (동의대)

목 차

제1주제	인공지능 시대, 체화된 마음과 불교	7
	발표: 이영의(고려대)	
	논평: 윤종갑(동아대)	
제2주제	AI GPT와 불교의 지혜(깨달음)	31
	발표: 이상헌(서강대)	
	논평: 일윤스님(경희대)	
제3주제	시와 불교논리학 - 앎을 중심으로 -	49
	발표: 지승도(한국항공대)	
	논평: 안환기(서울불교대학원대)	
제4주제	인간의 미래, 불교의 미래 - 인공지능의 발전과 자비윤리의 요청 -	77
	발표: 허남결(동국대)	
	논평: 보일스님(해인사승가대학)	

제 1주제

인공지능 시대, 체화된 마음과 불교

이영의 / 고려대학교

인공지능 시대, 체화된 마음과 불교

이영의 / 고려대학교

1. 들어가는 말

우리는 제4차 산업혁명의 물결을 타고 있다. 인류는 지난 수천 년간의 문명과는 비교가 되지 않을 정도로 빠른 속도로 발전하고 있는 과학기술에 따른 새로운 세계를 마주하고 있다. 이런 새로운 물결을 주도하는 것은 인공지능이다. 인공지능은 세탁기, 청소기, 냉장고, TV, 로봇, 인터넷 등으로 대표되는 기술적 차원(craft)에서 우리 삶을 변형하는 예술적 차원(art)으로 이행하면서 테크네(techne, 기예)의 본래 위상으로 복귀하고 있다. 이제는 더 이상 “인공지능 시대가 도래했다”라는 표현이 신선하지 않게 되었다.

그러나 인류의 역사를 살펴보면 인공지능이 주도하는 제4차 산업혁명 이전에도 적어도 세 차례의 커다란 산업혁명이 있었기 때문에 인공지능으로 인한 삶의 변화에 대한 예민한 반응을 ‘반복되는 예민함’으로 보고 인공지능의 가능성과 위험에 대한 수많은 담론을 ‘지적 호들갑’으로 보는 견해도 있다. 인공지능 시대에 대한 호들갑과 점잖음이라는 두 가지 태도는 양극단일 뿐이며 인공지능 시대에 대한 적절한 태도는 그 두 가지 태도의 중간 어딘가에 자리를 잡고 있다.

나는 이 논문에서 인공지능 시대에 우리가 취할 수 있는, 그리고 취해야 하는, ‘중간적 태도’를 불교와 체화인지 이론(theory of embodied cognition) 입장에 기반하여 제시하고자 한다. 나의 논의는 미시적이다. 인공지능이 의식을 가질 수 있는가, 인공지능이 불성을 가질 수 있는가, 인공지능이 인간의 반려자가 될 수 있는가, 인공지능 시대에 불교가 살아남을 수 있는가와 같은 거시적 주제는 미시적 주제를 기반으로 답변할 수 있다. 이 논문은 다음과 같은 순서로 진행된다. 2절은 인공지능 시대의 주요 특징과 그로 인한 문제를 검토하고 새로운 시대의 아젠다를 불교적 관점에서 평가한다. 3절은 나의 이론적 토대인 체화인지 이론을 행화주의를 중심으로 차후의 논의를 위한 배경으로 제시한다. 4절에서는 체화인지 이론과 불교의 접점을 보기 위해 인공지능 시대의 주요 문제의 근원인 자아를 무아, 공, ‘나 만들기’와 관련하여 검토한다.

2. 인공지능 시대의 특징

2.1 세 가지 특징

‘제4차 산업혁명’이라는 용어는 2016년 스위스 다보스에서 열린 세계경제포럼(World Economic Forum)에서 처음으로 등장했다. 다보스 포럼의 설립자 겸 의장인 슈왑(K. Schwab)에 따르면, 제4차 산업혁명은 인공지능, 로봇기술, 사물인터넷, 자율주행차량, 3D프린팅, 나노기술, 생명공학, 재료과학, 에너지보존, 양자컴퓨팅 등과 같은 물리적·디지털·생물학적 영역을 넘나드는 첨단기술의 융합인 동시에 그것들 간 상호작용이다. 제4차 산업혁명은 구체적으로 다음과 같은 특징을 갖는다.¹⁾ ① 속도: 이전의 산업혁명과 달리 제4차 산업혁명은 기하급수적인 속도로 전개된다. ② 너비와 깊이: 제4차 산업혁명은 제3차 산업혁명이 가져온 디지털 혁명에 기반을 두고 위에서 언급된 다양한 기술들을 결합하여 경제, 사업, 사회에서 전대미문의 패러다임 전이(paradigm shift)가 나타나고, 그 결과 우리가 무엇을 어떻게 해야 하는가뿐만 아니라 인간 정체성의 변화가 초래된다. ③ 체계 효과: 제4차 산업혁명은 국가, 기업, 산업, 사회 전반에 걸친 전 체계의 변형이 발생한다.

제4차 산업혁명이 매우 빠른 속도로, 광범위한 영역에서, 예상을 넘은 심각한 정도로 나타나며, 그 결과 삶의 공간도 이전과는 완전히 다른 형태로 변형되고 있다. 여기서 우리는 슈왑이 사용한 ‘패러다임 전이’라는 용어에 주목할 필요가 있다. 그 용어를 과학철학에서 처음으로 사용한 쿤(T. Kuhn, 1970)에 따르면 과학혁명은 기존의 패러다임이 새로운 패러다임으로 대체되는 사건이고, 패러다임 전이의 주요 특징은 패러다임 간 공약불가능성(incommensurability)이다. 과학혁명 과정에 있는 과학자들에게는 그들이 지지하는 패러다임을 비교할 수 있는 분명한 기준이 없으므로 그들은 서로 다른 세계에 살게 된다. 이와 마찬가지로 제4차 산업혁명 이후 삶의 공간은 혁명 이전과 현격히 차이가 날 것이고, 그 세계를 지지하는 정치적·경제적·사회 문화적·종교적 체계와 그것들이 지향하는 가치 및 규범 체계는 이전과는 비교할 수 없을 정도로 달라질 것이다.

앞으로 나는 제4차 산업혁명이 진행되고 있는 현시대를 ‘인공지능 시대’라고 볼 것이다.²⁾ 인공지능 시대는 다음과 같이 세 가지 특징으로 요약될 수 있다.

- 초연결

핵심 기술: 사물인터넷(만물인터넷), 스마트 폰

1) K. Schwab(2018), p. 3.

2) 앞에서 언급했듯이 제4차 산업혁명은 인공지능을 비롯한 여러 가지 기술에 의해 주도되고 있으므로 제4차 산업혁명 시대를 ‘인공지능 시대’라고 보는 것은 정확한 표현은 아니지만, 그 혁명의 핵심이 인공지능이라는 점을 고려하면 크게 문제가 되지 않는다.

인공지능 시대에서는 인터넷에 기반은 둔 다양한 매체와 인간-기계-사물-데이터 간 광범위하고 긴밀한 연결망을 바탕으로 하는 초연결(hyper-connectivity)이 나타난다. 모든 기술이 그렇듯이 초연결은 양면성을 갖는다. 한편으로 초연결은 인간 간 관계 맺기를 더 쉽게 하고 관계의 범위를 확대할 뿐만 아니라 기술적 한계로 실시되기 어려웠던 사회·정치적 이상의 구현을 촉진할 수 있다. 예를 들어, 초연결은 제한된 영역에서 직접민주주의를 구현하고, 블록 체인 기술을 활용한 투표제도를 가능케 하는 수단을 제공한다.

다른 한편으로 초연결은 우리의 삶에 부정적 영향을 미치기도 한다. 그 대표적인 예로 초연결 사회의 고독과 대화 부족이 있다. 이와 관련하여 터클(S. Turkle, 2011)은 초연결이 진정한 인간관계를 제공하지 못하기 때문에 자기 상실과 고독이 증가한다고 주장했다. 터클이 지적한 ‘초연결 속 고독’이라는 역설적 현상은 초고령화와 맞물려 노령층의 고독과 돌봄이라는 심각한 사회적 문제를 낳고 있다.

- 초인공지능

핵심 기술: 초인공지능, 딥러닝

인공지능(AI)이 인간지능과 대등한 수준을 넘어서 인간지능을 초월할 정도로 발전한다. 학자들은 인공지능이 다음과 같이 세 가지 단계로 발전할 것으로 예상한다.

① 좁은 인공지능(artificial narrow intelligence): 매우 제한된 인지적 과제를 인간 수준에서 또는 인간을 초월하는 수준에서 수행하는 AI. 예를 들어, 1997년 세계 체스 챔피언이었던 가리 카스파로프(Gary Kasparov)에게 승리한 딥 블루(Deep Blue)나 2016년 이세돌 9단과의 바둑 대결에서 승리한 알파고(AlphaGo)는 각각 ‘체스’와 ‘바둑’이라는 영역에서 인간의 능력을 초월한 AI이다. 현재의 AI는 좁은 인공지능 수준에 머물고 있다.

② 범용 인공지능(artificial general intelligence): 모든 인지적 과제를 인간 수준에서 성공적으로 수행하는 AI. 범용 인공지능의 개발은 다양한 방식으로 이루어질 것이다. 지금과는 전적으로 다른 기술로 개발되거나 아니면 좁은 인공지능의 영역 확대를 통해 개발될 수 있다. 후자의 예로는 알파고가 ‘알파고 제로’로 발전했고 알파고 제로는 다시 바둑뿐만 아니라 체스, 쇼기에도 적용된 ‘알파 제로’로 발전했고, 2020년에는 규칙을 배우지 않고 경기를 통달하는 무제로(Muzero)로 발전한 경우를 들 수 있다. 이런 과정을 거쳐 모든 인지적 과제를 인간과 동등한 수준에서 처리하는 범용 AI가 등장할 것이다.

③ 초인공지능: 모든 중요한 인지적 과제에서 인간지능을 능가하는 AI. 범용 인공지능과 초인공지능의 물리적 실현 가능성으로 인해 인간 중심주의적 휴머니즘이 위협받고 더는 합리성이

라는 기준을 통해 인간과 비인간을 구분하기 어렵게 되었다. 커즈와일(R. Kurzweil, 2005)은 초인공지능이 등장하는 시점을 ‘특이점’(technological singularity)이라고 명명하면서 그것이 2045년경 도래할 것으로 전망했다. 한편 뮐러와 보스트롬(C. Müller and N. Bostrom, 2016)이 인공지능 연구자들을 대상으로 벌인 설문조사에 따르면, 범용 인공지능은 2080년경 등장하고, 초인공지능은 그로부터 30년 후에 등장할 것으로 나타났다.

만약 초인공지능이 등장하게 된다면 인간의 삶은 어떻게 변화할 것인가? 이 질문에 대한 답변은 크게 비관론, 낙관론, 신중론으로 구분된다. 낙관론은 인간과 인공지능의 공진화(R. Kurzweil) 또는 사이보그와 같은 인간과 기계의 결합(A. Clark, 2003)을 통해 초인공지능 시대에 인간이 좋은 삶을 살게 될 것으로 예상한다. 비관론은 인간 노동의 실종, 인간의 기계 노예화, 인간적 가치가 사라진 삶을 예상한다. 신중론은 인공지능에 대한 적절한 통제를 통해 예상되는 부정적 결과를 예방하고자 한다. 예를 들어, 보스트롬은 초인공지능의 등장으로 인류는 실존적 위기(existential risk)에 직면하게 된다고 경고한다.³⁾ 초인공지능은 인간의 통제를 벗어나게 되고 인간 생존의 필요성을 인정하지 않을 것이기 때문이다. 보스트롬은 초인공지능의 등장을 예방하기 위해 AI 연구 및 개발을 통제해야 한다고 주장하면서 구체적으로 초인공지능의 능력을 통제하거나 초인공지능이 스스로 동기를 선택하지 못하도록 통제하는 방안을 제시하고 있다.⁴⁾

• 초인간

핵심 기술: 마음전송, 인공두뇌, 인체냉동, 유전자편집

인공지능만 발전하는 것이 아니라 인간도 현생인류(homo sapiens)의 한계를 벗어나 초인간(super human)으로 진화한다. 현생인류가 초인간으로 진화하는 경로에는 인간 향상(신체적, 인지적 향상뿐만 아니라 정서적, 도덕적 향상을 통한 인간 본성의 개조), 마음 전송(mind uploading)을 비롯한 탈신체화, 수명연장, 인간의 사이보그화 등의 기술들이 동원된다.

여기서 “초인간은 인간인가”라는 중요한 문제가 제기된다. 트랜스 휴머니즘(trans-humanism)은 인간 향상을 통한 인간의 진화를 주장한다. 트랜스 휴머니즘’이라는 용어를 현대적 의미로 처음 사용한 모어(M. More, 1990)에 따르면, 트랜스 휴머니즘은 과학·기술을 수단으로 하여 인간의 현재 형태와 한계를 극복함으로써 지적 생명체 진화의 연속성과 가속을 추구하는 생철학의 집합이다. 세계트랜스휴머니스트협회(World Transhumanist Association, WTA)의 선언(Transhumanist Declaration)에서 구체적으로 드러나듯이,⁵⁾ 트랜스 휴머니즘

3) N. Bostrom(2014), pp. 115-116.

4) Ibid, 8장.

5) 이 선언은 1998년 WTA에 의해 초안이 마련되었고 여러 번의 수정을 거쳐 2009년 WTA의 후신인 Humanity Plus에 의해 채택되었다.

은 한편으로는 합리성에 기반한 계몽주의적 진보 사상과 건강, 행복, 장수 등 휴머니즘적 가치를 수용하면서도 다른 한편으로는 인간 향상을 통한 탈신체화와 존재의 기술적 재구성을 주장하는 이중적 성격을 갖고 있다. 초인간은 트랜스휴먼뿐만 아니라 포스트휴먼의 형태로 나타날 수 있다.

트랜스 휴머니즘과 달리 포스트 휴머니즘은 주장하는 학자에 따라 다양한 유형으로 제시되고 있는데 일반적으로 기술 비판적 유형(F. Fukuyama, 2002, D. O'Mathuna 2009), 기술 친화적 유형(R. Kurzweil, D. Haraway), 그리고 휴머니즘의 문제에 주목하면서 포스트 휴머니즘의 사회적, 정치적, 윤리적 차원을 검토하는 유형(K. Hales 1999, C. Wolfe 2010, R. Braidotti 2013)으로 구분된다. 포스트 휴머니즘의 개척자인 헤일즈(K. Hayles)는 포스트 휴머니즘의 특징을 다음과 같이 제시한다.⁶⁾ 첫째, 정보 패턴이 어떤 물질적 예화보다도 존재 상태에 더 중요하므로 인간이 특정한 생물학적 기질로 체현된 것은 생명의 필연성이라기보다는 역사적 우연성이다. 둘째, 비물리적 영혼은 없으며 의식은 부수 현상에 불과하다. 셋째, 몸은 보철물에 불과하며 하나의 보철물을 다른 것으로 교체하는 것은 출생 전부터 시작된 과정의 연속이다. 넷째, 인간은 매끄럽게 지적인 기계와 연결될 수 있다. 포스트휴먼의 경우 신체적 존재와 컴퓨터 모의, 사이버네틱 기계와 생물학적 유기체, 로봇 목적과 인간 목표 간 본질적이거나 절대적 경계가 없다.

이상에서 볼 수 있듯이, (둘째 유형과 셋째 유형의) 포스트 휴머니즘은 여러 가지 점에서 트랜스 휴머니즘과 차이가 난다. 첫째, 트랜스 휴머니즘이 현생인류라는 특정한 휴먼을 전제하는 데 비해, 포스트 휴머니즘은 현생인류로서의 휴먼을 넘어서고자 한다. 포스트 휴머니즘은 그동안 인간 본성으로 간주되어 온 합리성과 자율성을 부정하고 인간 중심적 휴머니즘의 한계를 극복하고자 한다. 둘째, 포스트 휴머니즘은 인간과 다른 존재 간 구분을 거부한다. 헤일즈가 강조했듯이, 포스트 휴머니즘은 인간과 동물, 인간과 기계, 물리적 존재와 사이버 존재의 구분을 인정하지 않는다. 셋째, 포스트 휴머니즘은 트랜스 휴머니즘에 비해 보다 덜 목적론적이다. 즉 포스트 휴머니즘은 '인간 향상'이라는 목표에 집착하지 않으며 인간이 포스트휴먼에 도달할 수 있는 다양한 길을 모색한다.

2.2 새천년의 아젠다

지금까지의 논의를 정리해보면, 인공지능 시대가 어느 정도 진행되면, 영국 드라마 <휴먼스>(Humans, 2015-2018)에서 그려졌듯이, 인간, 트랜스휴먼과 포스트휴먼으로 대표되는 초인간, 초지능을 지닌 기계가 초연결 사회에서 함께 살아가게 될 것이고 인간과 초인간의 같

<https://humanityplus.org/philosophy/transhumanist-declaration>.

6) K. Hayles(1999), pp. 2-3.

등, 인간과 비인간의 갈등이 증폭될 것이다. 트랜스 휴머니즘의 이중성을 고려하면 현재 우리가 사는 인공지능 시대에서도 이미 앞에서 제시된 문제들이 발생하고 있거나 그것들의 씨앗이 싹트고 있다고 보아야 한다. 그런 문제들에 어떻게 대처해야 하는가? 우리가 그 문제들을 이해하는 데 불교가 어떤 도움을 줄 수 있는가?

이 질문에 대답하기 위해서 인공지능 시대가 인류에게 제기하는 아젠다(과제)를 이해할 필요가 있다. 이와 관련하여 하라리(Y. Harari, 2015)는 역사적 관점에서 인류는 지난 천 년 동안 기아, 역병, 전쟁이라는 아젠다를 해결했다고 진단하고, 새로운 천년에는 불멸, 행복, 신성(deity)이 새로운 아젠다가 될 것이라고 예상했다. 새로운 아젠다를 달성하면서 호모 사피엔스는 호모 데우스(Homo Deus), 즉 ‘신이 된 인간으로 업그레이드되고, 기성 종교는 그 힘을 잃고 기술 인본주의(techno-humanism)와 데이터주의(dataism)가 그 자리를 대신하게 된다. 기술 인본주의는 호모 사피엔스가 미래에 적합한 생존 모형이 아니므로 기술을 통해 우수한 모형인 호모 데우스를 창조하는 것을 목표로 한다.⁷⁾ 기술 인본주의 보다 더 급진적인 데이터주의는 우주가 데이터 흐름으로 구성되어 있으므로 모든 실재와 현상의 가치는 데이터 처리에 대한 기여도에 의해 결정된다고 주장한다.⁸⁾

새로운 아젠다가 인공지능 시대에 달성될 것인지에 대해 하라리는 부정적 견해를 피력했다. 우리는 여기서 새로운 아젠다의 달성 가능성이 아니라 아젠다의 규범성, 즉 그것들이 과연 추구할만한 것인지에 주목할 필요가 있다. 하라리가 제시한 세 가지 아젠다에 불교의 기본 교리를 적용하면 다음과 같은 결과가 나타난다.

아젠다	관련된 불교 교리	적용 결과
행복	일체개고(一切皆苦)	조건부로 적합함
불멸	연기(緣起), 윤회(輪廻)	부적합함. 소극적 태도 가능
신성	불성(佛性), 깨달음(正覺)	조건부로 적합함

부처가 출가한 이유에서 나타나듯이 불교는 “모든 것은 고이다”(一切皆苦)라는 판단에서 출발한다. 현생을 포함한 모든 것이 고통이라면, 불교 신자들이 행복을 추구하는 것은 모순을 범하는 것은 아닌가? 이 의문에 대해 플래넨(O. Flanagan)은 자연주의적 입장에서 다음과 같이 대답한다. 첫째, 불교에는 ‘고통’(dukkha, 苦) 개념에 대립하는 ‘즐거움’(sukkha, 樂) 개념이 있다.⁹⁾ 중생은 제행무상의 세계에 살고 있음에도 불구하고 무지(moha, 痴)로 인해 영원

7) Y. Harari(2017), p. 357.

8) Ibid., p. 372.

9) O. Flanagan(2011), p. 14.

히 지속 가능한 즐거움을 추구한다. 그러나 고통과 즐거움은 모두 일시적 현상에 불과하며 영구적인 실재는 아니다. 둘째, 고통의 대립 개념으로서 즐거움이 있듯이 자비(karuna, 慈悲)의 대립 개념으로 자애(metta, 慈愛)가 있다. 자비는 고통을 떨기 위한 성향을 의미하지만, 자애는 즐거움이나 더 좋은 불교적 행복을 산출하는 성향을 의미한다. 그 두 가지 개념 짝은 불교가 고통을 강조한 것은 사실이지만 즐거움의 가능성도 인정했다는 점을 보여주며, 이런 점에서 '불교적 행복'은 적어도 모순적 개념은 아니라는 결론이 나온다. 그러나 우리는 여기서 '불교적 행복'이 기술 인본주의와 데이터주의가 주장하는 방식으로 이해되거나 달성되어서는 안 된다는 점에 유의할 필요가 있다. 예를 들어, 항구적으로 행복감을 제공하는 약물을 제공하는 것은 자애로운 행위가 아니며 그것을 통해 얻은 행복은 '불교적 즐거움'도 아니다.

불멸을 추구하는 것, 즉 영원한 삶을 사는 것은, 불교의 교리에 따르면, 어려운 목표가 아니라 지극히 달성하기 쉬운 일이다. 왜냐하면 현생의 삶은 연기(緣起)에 따라 죽음으로 끝나지 않고 다음 생에서 태어나 생로병사의 고통을 겪는 끊임없이 '굴러가는 수레바퀴'이기 때문이다. 그러므로 바람직한 태도는 불멸을 추구하는 것이 아니라 깨달음을 얻어, 즉 해탈(解脫)하여 윤회(輪廻)의 고리에서 벗어나는 것이다. 그러나 깨달음을 얻는 것은 중생에게는 쉽지 않은 일이다. 그러므로 중생의 처지에서 불멸이 피할 수 없는 것이라면, 해탈에 대한 차선책으로 좋은 업(業)을 지어서 윤회의 과정이 덜 고통스럽고 더 즐거운 생으로 구성되도록 노력하는 소극적 태도가 차선책이 될 수 있다.

불교의 입장에서 보면 인류가 신성을 추구하여 신이 되려는 것은 지극히 당연하며, 요청되는 행위이다. 그러나 불성(佛性)은 기술이나 약물 사용을 통해 창조될 수 있는 것은 아니라 진정한 깨달음을 통해 얻어지는 것이다. 자기장을 형성하여 피험자의 측두엽을 자극하는 코렌헬멧(Koren Helmet)을 사용한 실험에서 피험자 중 일부가 신에 대한 시각적 경험이나 영성스러운 체험을 보고한 실험이 있었다. (L. A. Ruttan, M. A. Persinger, and S. A. Koren, 1990) 코렌헬멧을 착용한 피험자가 경험한 '불성'과 오랫동안 부처님 가르침을 충실히 따른 불교 신자가 경험한 불성의 차이는 무엇인가? 위의 실험을 수행한 연구자들은 우리의 질문과 관련된 추가 실험을 진행하지는 않았다. 다음 장에서 논의될 체화인지 이론에 따르면, 그 두 가지 경험을 일으킨 신경 패턴이나 상관된 신경 영역에서 분명한 차이가 있을 것이다. 왜냐하면 우리의 인지, 인식, 삶은 체화된 행위(embodied action)이기 때문이다. 그러므로, 몸 사용과 연결되지 않은 영성 체험은 불교적 수행과 연결된 영성 체험과 분명히 다를 것이다.

3. 체화인지 이론

3.1 핵심 내용

1990년대 이후로 인지과학 분야에서 기호주의(symbolism)와 연결주의(connectionism), 뇌 중심적 신경과학의 한계를 극복하고 마음과 인지에 대한 올바른 이해를 가능케 하는 연구 프로그램이 부상하고 있다. 기호주의, 연결주의, 뇌 중심적 신경과학과 경쟁할 것으로 기대되는 연구 프로그램은 바로 체화된 인지 이론(embodied cognition theory)이다. 현재 여러 가지 이론이 ‘체화된 인지 이론’이라는 명칭 하에 포섭되고 있는데, 그 중 대표적인 것은 내장 인지 이론(theory of embedded cognition), 확장인지 이론(theory of extended cognition), 행화인지 이론(theory of enactive cognition), 분산인지 이론(theory of distributed cognition), 상황인지 이론(theory of situated cognition) 등이다.

체화인지 이론을 구성하는 이론들은 세부적인 내용에서는 대립도 하지만 그것들은 모두 기호주의와 연결주의, 뇌 중심적 신경과학이 공유하는 견해, 즉 인지가 몸이나 환경과 (거의) 무관하게 작용한다거나, 인지는 전적으로 두뇌 안에서 작용한다고 보는 견해를 반대한다. 기호주의와 연결주의는 뇌 기반 적 계산(computation)과 표상(representation)을 강조한 나머지 마음과 인지가 몸을 가진 인간과 세계와의 역동적 관계에서 작용한다는 점을 간과함으로써 마음과 의식, 인지, 정서, 행위를 제대로 설명하지 못하고 있다. 우리는 체화인지 이론의 핵심을 다음과 같이 존재론, 의미론, 인식론, 방법론의 차원에서 정리할 수 있다.

- 존재론적 핵: 마음은 몸 구조와 능력에 의존하여 세계에서 펼쳐지는 활동이다.
- 의미론적 핵: 마음은 몸을 통한 세계와의 소통 과정에서 의미를 생성한다.
- 인식론적 핵: 마음은 몸 구조와 기능과 독립적으로 이해되거나 설명될 수 없다.
- 방법론적 핵: 마음을 설명하기 위해서는 몸 구조와 기능을 고려해야 한다.

체화인지 이론은 기존의 연구 프로그램의 한계를 극복하고 그동안 마음 연구에서 배제된 몸과 사회적·문화적 세계를 연구 대상으로 복원함으로써 몸-뇌-세계 간 역동적 관계를 중심으로 마음에 접근한다. 그렇다면 인지는 무엇이고 마음은 무엇인가? 체화인지 이론에 따르면, 인지는 몸을 가진 유기체가 세계를 살아가는 역동적 과정에서 창발한다. 더 구체적으로 인지는 유기체가 생존을 위해 자신의 세계를 탐험하는 과정이나 활동이다.

3.2 행화주의

체화인지 이론을 구성하는 이론 중 불교와 밀접하게 관련된 이론은 행화인지 이론이다. 행화인지 이론은 바렐라(F. Varela), 톰슨(E. Thompson), 로쉬(E. Rosch)에 의해 처음으로 인지과학 이론으로 제시되었다. 행화인지 이론은 행화주의(enactivism, 行化主義)라고도 불리는데, 그 핵심은 인지를 행위 차원에서 파악하는 데 있다.¹⁰⁾

행화주의는 다시 몇 가지 유형으로 분류될 수 있는데,¹¹⁾ 그중 자기생성적 행화주의 (autopoietic enactivism)는 자기생성(autopoiesis) 개념에서 출발한다. 자기생성 개념은 원래 생물학에서 세포의 자기유지 과정과 조직 원리를 설명하기 위해 도입되었는데 마투라나와 바렐라(H. Maturana and F. Varela, 1980)는 그것을 처음으로 인지에 적용했다. 자기생성적 체계는 생성 과정으로 구성된 폐쇄 조직을 통해 구현되는 하나의 단위로서, 같은 생성 조직이 체계의 반복적 생산을 통해 생성되고, 위상학적 경계가 이런 연속적 과정의 결과로 창발한다. 자기생성적 체계는 동일성을 유지하는 근본 변수로서 자신의 조직 원리를 갖추고 있는 항상적 체계이다.¹²⁾ 살아있는 세포는 자기생성적 체계의 전형적 예이다. 자기생성적 행화주의에 따르면, 유기체는 활동을 통해 자기 동일성과 항상성을 생성하고 유지하며 그 과정을 통해 인지 영역을 정의하는 자율적 행위자이다. 마음은 유기체의 자기 조직적이고 자기 생성적인 활동, 즉 몸·신경계·환경의 반복적인 감각·운동의 결합으로부터 창발한다. 표상주의와 계산주의를 기반으로 하는 인지주의(cognitivism)에 따르면, 인지는 내적 표상에 대한 계산이다. 자기생성적 행화주의는 인지주의를 반대하고 그 대신 인지를 체화된 행위(embodied action)로 볼 것을 주장한다. 인지에 대한 이런 새로운 이해는 아래 인용문에서 잘 나타난다.

체화된 이라는 용어를 사용하여 우리는 두 가지를 강조한다. 첫째, 인지는 다양한 감각운동 능력을 지닌 몸의 소유로부터 유래하는 경험의 종류에 의존한다. 둘째, 이러한 개별 감각운동 능력은 그 자체로 더 포괄적인 생물학적, 심리학적, 문화적 맥락에 내장되어 있다. 행위라는 용어를 사용하여 우리는 다시 감각적이고 운동적인 과정들, 지각과 행위가 근본적으로 살아있는 인지와 분리할 수 없다는 점을 강조한다. (F. Varela, E. Thompson, and E. Rosch, 1991, p. 173, 원저자 강조)

(정리) 인지(마음) = 체화된 행위(embodied action) = 행화(enaction)

4. 불교와 체화인지

행화주의 개척자인 바렐라는 마음을 제대로 설명하기 위해 세계를 3인칭적으로 기술하는 과학, 특히 인지과학에 1인칭적 기술을 제공하는 현상학과 불교 사상, 특히 중관사상(中觀思

10) 이런 이유로 관련 핵심 용어인 ‘enactive’, ‘enaction’, ‘enactivism’은 각각 ‘행화적’, ‘행화’, ‘행화주의’로 번역한다.

11) 행화주의는 여기서 논의된 자기생성적 행화주의 외에도 감각운동 행화주의(sensorimotor enactivism, S. Hurley 1998, A. Noë 2004, 2009, 2012, K. O’Regan 2011), 급진적 행화주의(radical enactivism, D. Hutto and E. Myin 2012, 2017) 등이 있다.

12) H. Maturana and F. Varela(1980), pp. 78-70.

想)을 융합해야 한다고 보았다. 그의 연구는 다음과 같은 결과를 낳았다.

- The Embodied Mind: Cognitive Science and Human Experience (F. Varela, E. Thompson, and E. Rosch)
- “Neurophenomenology: A Methodological Remedy for the Hard Problem” (F. Varela, 1996)
- Ethical Kow-How: Action, Wisdom, and Cognition (F. Varela, 1996)

바렐라의 마음에 관한 융합적 연구는 2001년 그의 갑작스러운 죽음으로 중단되었고 후속 연구는 주로 톰슨(Thompson)이 이어받아 수행되고 있다. 톰슨의 주요 연구는 다음과 같다.

- Mind in Life: Biology, Phenomenology, and the Sciences of Mind. (E. Thompson, 2007.)
- Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy (E. Thompson, 2015.)
- Why I Am Not a Busddhist (E. Thompson, 2020)

지금부터는 바렐라와 톰슨의 연구를 중심으로 자아에 대한 환원론적 접근에 대한 체화인지적 비판과 대안적 설명을 검토하기로 한다.

4.1 무아와 공

우리는 자아 개념에 매우 친숙해 있다. 우리는 보고, 듣고, 냄새 맡고, 맛보고, 느끼고, 의식한다. 그렇다면 그런 기능을 주관하는 무엇이 있지 않은가? 그것이 바로 ‘나’이다. 우리는 그런 ‘나’가 있다고 가정하고, 나를 중심으로 세상을 경험하고 판단하고 행동한다. 자아, 주체, 주관은 대상, 객체, 객체와 대립하는 개념이다. 철학자들은 예로부터 인간의 본질은 생각하는 데 있다고 주장했다. 예를 들어 데카르트는 “나는 생각한다. 그러므로 나는 존재한다”라는 선언을 통해 ‘생각하는 나’는 다른 어떤 것보다도 확실히 존재한다고 보았다.

현대 신경과학의 주요 발견 중 하나는 무아(無我)이다. 뇌를 아무리 연구해 보아도 뇌에는 자아가 없다. 신경과학자들이 자아를 부정하는 다양한 주장을 제시해 왔는데 그 요지는 “자아는 없다”라거나 “자아는 환상의 결과이다”라는 것이다. 노벨상 수상자인 크릭(Crick)은 다음과 같이 자아를 부정한다.

"당신", 당신의 기쁨과 슬픔, 기억과 야망, 개인적 정체성과 자유의지는 사실 신경 세포와 관련 분자의 방대한 집합체의 행동에 지나지 않는다. ... "당신"은 신경세포의 집합에 불과하다. "(F. Crick, 1994, p. 3)

신경과학자들만이 자아를 부정하는 것은 아니며 환원적 물리주의 철학자들도 거기에 동참한다. 예를 들어, 신경철학(neurophilosophy)을 주장한 처칠랜드(P. S. Churchland)는 그 대표적인 예이다.

"많은 신경과학적 증거에 따르면, 느끼고, 생각하고, 결정하는 것은 신비로운 어떤 것이 아니라 두뇌이다. 이는 곧 사랑에 빠질 영혼이 없다는 것을 의미한다. 확실히 우리는 여전히 사랑에 빠지며 열정은 예전처럼 실재한다. 차이점은 이제 우리는 그런 중요한 느낌들이 물리적 두뇌 안에서 발생하는 사건들로 이해한다는 점이다." (P. S. Churchland, 2012, p. 1)

신경과학자들은 그런 믿음이 아무런 과학적 근거도 없다는 점을 보여주려고 노력한다. 그런데 왜 우리는 여전히 자아를 믿는가? 그 이유는, 크릭에 따르면, 첫째 우리는 자아와 의식을 신경과학적으로 설명하는 것을 받기지 않기 때문이고, 둘째 맛있는 차를 마시면서 느끼는 것과 같은 언어로 표현하기 어려운 의식의 주관성을 믿기 때문이고, 셋째 자유의지를 믿기 때문이다.¹³⁾

위에서 보았듯이, 자아의 존재에 대해 두 가지 입장이 대립하고 있다.

- ① 자아 실재론: 인간의 사고, 감정, 행동을 주관하는 자아는 실재한다.
- ② 자아 부정론: 자아 실재론이 주장하는 그런 자아는 존재하지 않는다.

불교는 위의 두 가지 입장 중 어디에 속하는가? 이 질문에 대한 불교의 대답은 다음 두 가지 교리와 관련하여 고찰할 수 있다.

- ③ 제법무아(諸法無我): 모든 존재는 영원하지 않고 변한다.
- ④ 제행무상(諸行無常): 모든 존재는 고정된 실체가 아니다.

제법무아와 제행무상에서 나타나는 '법'과 '행'은 '존재'를 의미하므로 존재하는 '나'도 거기에 포함된다고 보아야 한다. 그러나 '나'라는 존재에 대한 이런 이해는 자아의 존재 여부와 무관하다. 왜냐하면, 자아 부정론자들도 몸을 갖고 생로병사를 겪는 유한자로서의 나를 인정할 것이기 때문이다. 그러므로 진정한 문제는 '법'과 '행'에 '자아'가 포함되느냐는 것이다.

13) F. Crick(1994), pp. 7-10.

혹자는 ‘나’가 ‘무아’하고 ‘무상’한데 ‘자아’도 당연히 그렇지 않겠느냐고 반문할 것이다. 그러나 이는 잘못된 생각이다. 우리는 실재(reality)와 현상(phenomenon)을 구별할 필요가 있다. 여기서 실재는 인간의 지각과 해석과 독립하여 진정으로 존재하는 것이고, 현상은 관찰할 수 있거나 지각할 수 있는 것이다. 현상은 실재의 지각할 수 있는 측면일 수 있으나, 우리는 지각의 배후에 있다고 가정되는 실재가 실제로 존재하는지를 경험할 수 없으므로, 현상은 실재와 무관한 그저 관찰된 것일 수 있다. ‘법’과 ‘행’은 실재를 의미하는가, 아니면 현상을 의미하는가? 제법무아와 제행무상의 의미와 실재에 대한 정의를 상기해보면 그 개념들은 실재가 아니라 현상을 의미하는 것으로 보아야 한다. 그렇다면 두 가지 법인은 다음과 같이 이해되어야 한다.

- ⑤ 제법무아(諸法無我): 모든 현상은 영원하지 않고 변한다.
- ⑥ 제행무상(諸行無常): 모든 현상은 고정된 실체가 아니다.

그리고 여기에 “자아는 실재가 아니라 현상이다”(⑦)라는 주장을 추가하면, 다음과 같은 결론이 따라 나온다.

- ⑧ 자아는 실재로서 존재하지 않지만, 현상으로는 존재한다.

우리는 여기서 자아의 존재에 대한 불교의 견해가 자아 실재론(①)도 아니고 자아 부정론(②)도 아니라는 점을 알 수 있다. 바렐라는 용수의 공(空) 사상이 자아에 대한 불교의 견해를 가장 잘 보여준다고 생각했다.

용수의 [중론의] 요점은 대상이 절대적 의미에서 존재한다거나 절대적으로 존재하지 않는다고 말하지 않는 데 있다. 대상은 상호 의존적으로 발생하며, 완전히 근거를 갖고 있지 않다. ... 용수는 마지막으로 [완전한 상호의존성 논증을] "의존적으로 발생하지 않는 것은 아무것도 없다. 이런 이유로 공(空)이 아닌 것은 아무것도 없다"라고 결론 내렸다. (F. Varela, E. Thompson, and E. Rosch, 1991, p. 223)

여기서 ‘의존적 발생’은 연기를 의미하므로 우리는 공 사상이 연기설에 의존하고 있다는 것을 알 수 있다. 이상에서 보았듯이 공(空) 사상은 자아 실재론과 자아 부정론의 대립을 초월하는 사상이다. 그러나 바렐라는 자아의 정체가 무엇인지를 분명히 밝히지 못했다. 자아가 실재가 아니고 현상이라면, 그것이 어떻게 우리의 정신과 삶을 주재하는가? 이에 대한 대답은 톰슨의 연구에서 나타난다.

4.2 나 만들기와 의미 만들기

톰슨은 자아에 관한 바렐라의 연구를 이어받아 자아의 정체와 작용을 보다 더 분명히 제시한다. 톰슨의 출발점은 "자아는 의존적으로(또는 연기로) 발생하는 과정이다"라는 논제이다. 톰슨은 그 논제로부터 출발하여 자아에 대한 행화주의적 이해를 제시한다. 우선 톰슨은 1인칭 시점에서 체험된 '나'에 대한 느낌을 표현하기 위해 인도철학에서 '나 만들기' (I-making, ahamkāra)라는 용어를 차용하는 데, 그 용어의 사전적 의미는 "'나 만들기'(the making of self) 또는 '나 제조기(I-maker)이다.¹⁴⁾ 그다음 톰슨은 자아란 다양한 '나 만들기' 과정에서 창발하고 그것에 의존하는 것이라고 주장한다. 자아에 관한 그의 주장은 다음과 같은 논리적 구조를 갖는다.¹⁵⁾

- 제안: 자아는 '나 만들기' 과정에서 발생한다. 그 과정은 "나"와 행화하고(enact) 그 안에서 '내'가 그 과정 자체와 다르지 않은 지속적 과정이다.¹⁶⁾
- 이론적 도구: 자기-지정 체계(self-specifying system). 그 체계는 서로를 상호 지정하여 더 넓은 환경과의 관계에서 체계를 자기 영속적인 전체로서 구성하는 과정들의 집합이다.
- 구현: 여러 수준(생물학적, 심리적, 사회적)에서 나 만들기를 구성하는 자기 지정 체계에 대한 설명.

자아가 나 만들기의 과정으로부터 창발하는 것이라면, 삶도 그렇게 설명할 수 있는가? 톰슨은 삶을 불안정한 조건에서의 '의미 만들기'(sense-making)라고 본다. 삶을 의미 만들기의 과정으로 보는 것은 톰슨뿐만 아니라 행화주의자들이 일반적으로 취하는 전략이다.¹⁷⁾ 여기서는 의미 만들기에 대해 드 한(de Haan)이 제시한 더욱 포괄적인 설명을 살펴보기로 하자. 유기체는 생존을 위해 다양한 활동을 한다. 신진대사와 같은 기본활동뿐만 아니라 먹을 것과 먹지 말아야 할 것, 짝짓기 상대 고르기, 적과 같은 위험 요인을 재빨리 파악해야 한다. 인간을 비롯한 고등 동물은 그러한 기본활동을 잘 수행하는 것만으로 삶을 잘 살 수는 없다. 고등 동물에게 삶이란 물리적 환경뿐만 아니라 사회적 환경과 문화적 환경에 잘 적응하고 활동하는 것을 포함한다. 다윈의 진화론에 따르면, 삶은 유기체가 주어진 환경에 적응하는 과정이다.

14) W. J. Johnson(2009).

<https://www.oxfordreference.com/display/10.1093/acref/9780198610250.001.0001/acref-9780198610250-e-75;jsessionid=02B8075BC568410E5F1A8E8CF020AF4D>

15) E. Thompson(2015), p. 326.

16) 마투라나와 바렐라(H. Maturana and F. Varela, 1980)가 인지를 설명하기 위해 'autopiesis'를 처음 도입했을 때, 그것은 '자아 만들기'(self-making) 또는 '자아 생산'(self-production)을 의미했다.

17) E. Di Paolo(2009), E. Thompson(2011, 2015), S. de Haan(2020a, 2020b) 참조.

여기에서 중요한 점은 환경이 유기체에 그냥 ‘주어진 것’은 아니라는 것이다. 삶이란 유기체와 환경 간의 상호작용에서 성립한다. 이 점에 착안하여 더 한은 삶이란 유기체가 자신의 환경을 의미 있게 만드는 ‘불안정한 과정’이라고 본다.¹⁸⁾ 삶이 불안정한 이유는 의미 만들기가 성공적으로 수행되지 않고 실패할 수도 있기 때문이다. 드 한이 지적한 ‘불안정한 과정’은 톰슨이 말한 ‘불안정한 조건’에 상응한다.

의미 만들기는 기본적 의미 만들기과 실존적 의미 만들기로 구분된다. 기본적 의미 만들기는 자기 생존을 지지하는 것과 위협하는 것을 구분하는 것뿐만 아니라 자기 욕구의 의미를 파악하는 것을 포함한다. 유기체의 측면에서 보면 환경은 중립적일 수 없고 어떤 의미가 있게 마련이다. 유기체가 생존을 위해 환경에 의존하는 이유는 환경의 여러 측면이 해당 유기체에 어떤 가치를 갖기 때문이다. 바로 이 지점에서 구성주의자들은 의미가 유기체에 의해 전적으로 구성된다고 주장한다. 그러나 더 한은 구성주의를 반대한다. 의미 만들기 작업이 유기체의 삶에 필요조건이지만, 그런 능동성을 지나치게 강조한 나머지 의미가 구성된다고 주장하는 것은 잘못이다. 더 한에 따르면 의미 만들기는 수동적이거나 구성적이지 않고 ‘관계적’이다. 예를 들어, 설탕이 박테리아에게 의미가 있는(음식인) 이유는 설탕과 박테리아의 특징 때문이다. 이처럼 의미 만들기는 환경에 주어진 것과 유기체의 특징에 의존한다.

실존적 의미 만들기는 유기체가 자기 경험에 반성하며 관계하는 것으로부터 나타난다. 유기체는 실존적 자세(existential stance)를 취함으로써 기본적 의미 만들기를 벗어나 실존적 의미 만들기의 차원에 들어서게 된다. 여기서 유의해야 할 점은 실존적 의미 만들기가 기본적 의미 만들기의 위에 얹힌 것은 아니라는 점이다. 기본적 의미 만들기에서 실존적 의미 만들기로의 이행은 ‘생존을 위한 유기체’에서 ‘인격체로서의 유기체’로의 이행이므로 그것은 ‘유기체·환경’의 관계에서 ‘인격체·세계’(person-world)의 관계로 나아가는 질적인 변형이다.¹⁹⁾ 기본적 의미 만들기가 음식 찾기, 짝짓기, 위험 피하기 등과 같은 현재 환경의 적합한 측면을 파악하는 것인데 비해, 실존적 의미 만들기는 생존을 위한 의미 만들기를 넘어서 생존을 위한 자신의 적절성에 대한 반성이다. 그런 반성은 당연히 가치 평가적 차원을 갖게 된다. 인격체는 실존적 의미 만들기의 평가적 기능을 통해 단순한 생물학적 생존이 아니라, 존경, 명예, 존엄, 우정, 사랑 등과 실존적 가치를 통한 좋은 삶을 지향하게 된다.

마지막으로, 자아를 ‘나 만들기’ 과정으로 보는 이론이 구성주의(constructivism)에 해당하는지를 살펴보자, 우리는 앞에서 드 한이 유기체의 의미 만들기가 환경에 의존한다는 사실로부터 의미가 유기체에 의해 구성된다는 주장을 끌어내는 것을 반대하는 것을 보았다. 행화주의 입장에서 보면, 의미는 인간과 환경 간 역동적인 관계로부터 나타나는 것이지, 그것이 전적으로 나(또는 나의 뇌, 나의 정신)에 의해 구성되는 것은 아니다. 그러나, 일부 행화주의 연

18) S. de Haan(2020a), p.7.

19) Ibid., p. 9.

구자들은 행화주의를 구성주의적으로 이해하려고 한다. 그런 오해가 발생하는 데에는 행화주의자들의 잘못도 있다. 예를 들어, 톰슨은 그런 오해를 일으킬 수 있는 표현을 자주 사용했다.

(구성과 환상은 다르다는 점을 주장하면서) “비록 내가 자아 또는 ‘I-Me-Mine’에 대한 우리의 느낌이 정신적으로 구성된다(constructed)는 요가카라(Yogācāra)의 주장에 동의하지만, 그 주장으로부터 자아는 존재하지 않는다든지 자아 현상은 단지 환상에 불과하다는 주장이 따라 나온다고 생각하지 않는다.” (E. Thompson, 2015, p. 358)

(불교를 신경과학적으로 이해하려는 위험을 경고하면서) “자아는 뇌에서 생성된 환상이나 존재하지 않은 허구가 아니다. 그것은 생물학적이고 사회적인 구성(construction이다”. (E. Thompson, 2020, p. 19)

위의 인용문에서 드러나듯이, 톰슨은 “자아가 나에게 의해 구성된다”라고 주장하지 않았다. ‘만들기’를 다르게 해석할 수 있지만, 자아가 ‘나 만들기’ 과정의 결과라는 점에서 고정불변의 실재는 아니지만 물리적으로 구현된 현상이므로, 그것을 단순히 뇌가 만들어내거나 정신이 구성한 것이라고 말할 수 없다. 물론, ‘만들기’를 ‘구성’으로 해석할 수도 있지만 그 경우에 구성은 구성주의자들이 주장하는 ‘구성’과는 분명한 다른 의미를 갖는다.

5. 맺는말

이 논문의 전반부는 인공지능 시대의 세 가지 특징과 그런 특징을 지닌 시대에서 발생하는 문제들을 검토하고, 새천년에 인류에게 부여된 아젠다(행복, 불멸, 신성)를 불교적 관점에서 평가해 보았다. 그 결과는 각각 조건부 적합, 부적합, 조건부 적합으로 나타났다. ‘조건부’ 적합이라는 판단은 불교적 수행의 가능성을 의미하지만, 부적합 판단은 불멸이 불교의 기본 교리와 정면으로 충돌한다는 점에서 인공지능 시대에서의 불교적 삶의 어려움을 함축한다.

이 논문의 후반부는 체화인지 이론이 불교와 만나는 지점을 자아에 대한 이해를 중심으로 검토했다. 논의 결과 환원론적 신경과학자들의 주장과는 달리, “자아가 존재하지 않는다”라거나 “자아는 환상이다”라는 주장이 아니라 “자아는 현상으로서 존재하며, 나 만들기의 과정에서 창발한다”라는 주장이 성립한다. 그러나 만들기를 구성주의적으로 해석할 수 있다는 것은 자아를 인공지능, 신경과학, 유전공학 등과 같은 기술에 의해 제조, 조작할 수 있다는 점을 의미한다고 볼 수도 있다. 또한 이 논문에서 다루지 않았지만, 불교의 교리를 최대한 과학과 일치시키기 위해(특히 명상의 경우에서) 불교의 형이상학이나 의식을 제거하려는 시도(Buddhist modernism)는 불교의 장점을 보이는 것이 아니라 오히려 종교로서의 불교를 훼손하는 것이라는 점도 유의해야 한다.

- 명범. 2022. 「포스트휴먼 시대와 불교」, 『불교문화』, 268(2023.12).
- 보일. 2021. 『AI 부디즘』, 답앤북스.
- 보일. 2022. 「불교는 트랜스휴머니스트에게 사상적 우군일까? 」, 『불교문화』, 268(2023, 12).
- 소랏 형라다름. 2022. 김근배·김진선·주은혜·허남결 옮김. 『불교의 시각에서 본 AI와 로봇 윤리』, 씨아이알.
- 이기흥. 2017. 「행화주의 마음치유: 시론」, 『철학탐구』 48, pp. 91-129.
- 이상현. 2017. 『철학자의 눈으로 본 첨단과학과 불교』, 살림.
- 이영의. 2018. 「자연화된 불교, 행복, 행화주의」, 『철학논집』, 54, PP. 195-221.
- 이영의. 2018. 「행화주의와 창발 그리고 하향인과」, 『철학·사상·문화』, 28, pp. 118-137.
- 이영의. 2019. 「신경과학과 불교 수행자의 행복」, 『불교문화』 227(2019. 07), PP. 3-35.
- 이영의. 2020. “Being and Relation in the Posthuman Age”, *Society and Power*, 65(5), pp. 7-12.
- 이영의. 2021. 『신경과학철학』. 아카넷.
- 이영의. 2021. 「어떻게 의식의 어려운 문제를 해결해야 하는가?」, 『동서인문』, 17, pp. 419-461
- 이영의. 2022. 「정신질환에 대한 행화주의적 접근」, 『인문과학연구』, 12, pp. 199-213.
- 이영의. 2022. 「포스트휴먼의 삶: 돌봄로봇과 속임문제」, 『생사학연구』, 2, pp. 55-86.
- 이영의. 2022. “Embodied cognition, Narrative, and the Self”, *International Journal of Body, Nature, and Culture*, 1(1), pp. 31-48.
- Arnold D. 2014. *Brains, Buddhas, and Believing: The Problem of Intentionality in Classical Buddhist and Cognitive-Scientific Philosophy of Mind*. New York: Columbia University Press.
- Aristotels. 1926. *Nicomachean Ethics*. H. Rackham trans. Cambridge, MA: Harvard University.
- Austin, J. H. 1998. *Zen and the Brain: Toward an Understanding of Meditation and Consciousness*. Cambridge, MA: MIT Press.
- Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press.
- Braidotti, R. 2013. *The Posthuman*. Malden, MA: Poity Press.
- Brooks, D. 2008. “The Neural Buddhists”, *New York Times*, May 13.
- Bush, S. S. 2008. “Divine and Human Happiness in Nicomachean Ethics”, *Philosophical Review*, 117(1), pp. 49-75.
- Churchland, P. S. 1986. *Neurophilosophy: Toward a Unified Science of Mind-Brain*. Cambridge, MA: MIT Press.
- Churchland, P. S. 2012. *Brain-Wise: Studies in Neurophilosophy*. Cambridge. MA: MIT Press.
- Clark, A. 2003. *Natural Cyborg: Minds, Technologies, and the Future of Human Intelligence*. Oxford: Oxford University Press.
- Clark, A. and Chalmers, D. 1988. “The Extended Mind”, *Analysis* 58, pp. 7-19.
- Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Charles Scribner’s Sons.
- De Haan, S. 2020a, “An Enactive Approach to Psychiatry”, *Philosophy, Psychiatry, &*

- Psychology, 27(1), pp.3-25.
- De Haan, S. 2020b, *Enactive Psychiatry*. Cambridge: Cambridge University Press.
- Di Paolo, E. 2009. "Extended Life," *Topoi*, 28, pp. 9-21.
- Doctor, T. and Linkenhoker, 2023. "What Buddhism Can Teach Us About Artificial Intelligence - And Ourselves", Templeton World Charity Foundation. <https://www.templetonworldcharity.org/blog/what-buddhism-can-teach-us-about-artificial-intelligence-and-ourselves>
- Flanagan, O. 2011. *The Bodhisattva's Brain: Buddhism Naturalized*. Cambridge, MA: MIT Press.
- Fukuyama, F. 2002. *Our Posthuman Future: Consequences of the Biotechnology Revolution*. New York, NY: Farrar, Straus and Giroux.
- Harari, Y. 2017.
- Hayles, K. 1999. *How We Became Posthuman*. Chicago, IL: University of Chicago Press.
- Johnson, W. J. 2009. *A Dictionary of Hinduism*. Oxford: Oxford University Press. <https://www.oxfordreference.com/display/10.1093/acref/9780198610250.001.0001/acref-9780198610250-e-75;jsessionid=02B8075BC568410E5F1A8E8CF020AF4D>
- Lutz, A. and Thompson, E. 2003. "Neurophenomenology: Integrating Subjective Experience and Brain Dynamics in the Neuroscience of Consciousness," *Journal of Consciousness Studies*, 10, pp. 31-52.
- Kabat-Zinn, J. 1994. *Wherever You Go, There You Are: Mindfulness Meditation in Everyday Life*, New York: Hyperion.
- Kabat-Zinn, J. 2003. "Mindfulness-Based Interventions in Context: Past, Present, and Future", *Clinical Psychology: Science and Practice*, 10, pp. 144-56.
- Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.
- Müller, V. C. and Bostrom, N. 2016. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion", in V. C. Müller ed., *Fundamental Issues of Artificial Intelligence*, Springer. pp. 553-571.
- Lopez Jr., D. S. 2009. *Buddhism and Science: A Guide for the Perplexed*. Chicago: University of Chicago.
- Lopez, Jr., D. S. 2012. *The Scientific Buddha: His Short and Happy Life*. New Haven: Yale University Press.
- McMahan, D. L. 2008. *The Making of Buddhist Modernism*. New York: Oxford University Press.
- Merleau-Ponty, M. 1945/1962. *Phenomenology of Perception*. C. Smith trans. London: Routledge & Keegan Paul.
- More, M. 1990, revised 1996. "Transhumanism: Towards a Futurist Philosophy", *Extropy*, 6, pp. 6-12.
- Noë, A. and Thompson, E. 2004. "Are There Neural Correlates of Consciousness?", *Journal of Consciousness Studies*, 11, pp. 3-28.
- O'Mathuna, D. 2009. *Nanoethics: Big Ethical Issues with Small Technology*. London: Continuum.
- Ruttan, L. A, Persinger, M. A, and Koren, S. A. 1990. "Enhancement of Temporal Lobe-Related Experiences During Brief Exposures to MilliGauss Intensity Extremely

- Low Frequency Magnetic Fields". *Journal of Bioelectricity*, 9 (1), pp. 33-54.
- Schwab, K. 2018. *The Fourth Industrial Revolution*. New York: Crown Business.,
- Sharf, R. 1995. "Buddhist Modernism and the Rhetoric of Meditative Experience," *Numen* 42(3), pp. 228-283.
- Thompson, E. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.
- Thompson, E. 2011. "Living Ways of Sense-Making", *Philosophy Today*, SEP Supplement, pp. 114-123.
- Thompson, E. 2015. *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*. New York: Columbia University Press.
- Thompson, E. 2020. *Why I Am Not a Buddhist*. New Haven: Yale University.
- Turkle, S. 2011. *Why We Expect More from Technology and Less from Each Other*. Basic Books. New York: Basic Books.
- Varela, F. 1996. "Neurophenomenology: A Methodological Remedy for the Hard Problem". *Journal of Consciousness Studies* 3, pp. 330-349.
- Varela F. 1999. *Ethical Know-How: Action, Wisdom, and Cognition*. Stanford CA: Stanford University Press.
- Varela, F. Thompson, E. and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.
- Wolfe, C. 2010. *What Is Posthumanism?* Minneapolis, MN: University of Minnesota Press.

「인공지능 시대, 체화된 마음과 불교」에 대한 논평문

윤종갑 / 동아대학교

오늘날 최대의 시대적 화두는 챗GPT 등장으로 인공지능 기술과 발전 가능성에 관한 문제이다. 사물인식 능력과 언어 능력에서 압도적인 발전을 거듭하던 인공지능은 이제 거대 생성형 AI로 인류에게 성큼 다가왔다. 그런데 이러한 초기기술의 개발이 앞으로 인류의 축복이 될지 아니면 재앙이 될지는 오직 신만이 알 것이다. 단지 현재의 상황에서는 인공지능 기술의 발전 가능성과 이에 대한 문제점을 미리 파악하여 축복이 될 수 있도록 선도해 나가는 일이다.

챗GPT의 등장은 산업혁명 중에서도 가장 혁신적이고 두려움을 안겨주는 양가적 명암을 갖고 있다. 인간의 전유물로 여겨지던 지식과 사고의 체계를 인공지능도 할 수 있으며, 오히려 인간을 압도하는 초지식과 사고체계를 갖고서 인간을 지배할 수도 있기 때문이다. 초스피드로 진화하는 인간지능 기술은 새로운 세계의 서막인 동시에 두려움의 대상인 것이다. 이러한 인공지능 기술의 거대한 흐름의 변곡점에서 “인공지능 시대, 체화된 마음과 불교”에 대해 논의한 본 논문은 불교와 인공지능이 사상적으로 어떻게 연결되며, 앞으로 불교가 어떠한 입장을 견지해야 할지에 관해서 매우 유익한 정보를 제공하는 시의적절한 의미 있는 연구로 생각된다.

본 논문은 크게 두 부분으로 이루어져 있는데, 전반부는 인공지능 시대의 특징과 발생할 수 있는 문제점을 검토하고, 앞으로 인류에 요청되는 행복, 불멸, 신성의 세 가지 과제를 불교적 관점에서 평가하고 있다. 그리고 후반부는 체화 인지 이론(embodied cognition theory), 그 중에서도 바렐라(F. Varela), 톰슨(E. Thompson), 로쉬(E. Rosch)에 의해 처음으로 제시된 행화주의(enactivism, 行化主義)를 불교의 중관사상과 연결하여 그 접점을 찾아보려고 한다. 즉 데카르트의 심신이원론과는 전혀 입장을 달리하는 중관사상의 공, 무아 개념에 바탕하여 자아의 문제를 중심으로 한 자아의 실재와 현상에 대해 행화주의와의 접점을 찾아보려는 것이다.

논자는 전반부의 결론으로서 현생 인류의 휴먼을 전제로 하는 트랜스 휴머니즘과는 달리 포스트 휴머니즘은 인간과 다른 존재 간의 구분을 거부하기 때문에 초인간, 초지능을 지닌 인

공(AI) 인간이 초연결 사회에서 인간과 함께 살아가게 된다면 인간과 초인간의 갈등, 인간과 비인간의 갈등이 증폭될 것이라고 진단한다. 따라서 이러한 문제를 예방하기 위해 인공지능 시대에 요청되는 3가지 아젠다를 불교의 근본 교리에 적용하여 다음과 같은 결과를 도출한다.

아젠다	관련된 불교 교리	적용 결과
행복	일체개고(一切皆苦)	조건부로 적합함
불멸	연기(緣起), 윤회(輪廻)	부적합함. 소극적 태도 가능
신성	불성(佛性), 깨달음(正覺)	조건부로 적합함

먼저, 행복에 대해 적용한 결과 불교는 일체개고를 주장하며 고통을 강조한 것은 사실이지만 즐거움이나 더 좋은 불교적 행복을 추구한 자애(*metta*, 慈愛)를 인정했기 때문에 교리와 모순되는 것이 아니라는 것이다. 둘째, 불멸을 추구하는 것, 즉 영원한 삶을 사는 것은 연기와 윤회를 주장하는 불교적 입장과는 맞지 않는 것으로 해탈을 하여 윤회로부터 벗어나거나 그것이 쉽지 않다면 차선책으로 좋은 업을 지어 덜 고통스러운 삶을 살아가도록 해야 한다는 것이다. 그리고 셋째, 신성은 불성을 강조하는 불교에서 당연히 추구되는 적합한 아젠다로 인정된다.

후반부의 논자의 결론은 불교는 자아 실재론과 자아 부정론 그 어느 쪽도 아닌 “자아는 실재로서 존재하지 않지만, 현상으로는 존재한다.”는 결론을 이끌어 내며, 특히 ‘용수의 공(空) 사상’이 그 모범 답안임을 바렐라의 입장을 통해 제시하고 있다. 그러나 바렐라는 자아의 정체를 밝히지 못했는데, 바렐라의 연구를 이어 받은 톰슨은 “자아는 의존적으로(또는 연기로) 발생하는 과정”이지 자아라는 실재가 있는 것이 아님을 밝혔다. “톰슨은 자아란 다양한 ‘나 만들기’ 과정에서 창발하고 그것에 의존하는 것”이라고 주장하며 삶 또한 “불안정한 조건에서의 ‘의미 만들기’(sense-making)”라고 본다.

나라는 존재는 단지 주어진 환경에 수동적으로 적응하는 유기체[구성주의자들이 주장하는 기본적 의미 만들기]가 아니라 보다 질 높은 가치를 추구하고 선택함으로써 자신의 인격을 고양시켜 나가는 능동적 유기체[체화 인지 행화주의자들이 주장하는 실존적 의미 만들기]라는 점이다. 이와 같은 몸-뇌-세계 간 역동적 관계에 의해 자아가 형성된다는 인지과학의 행화주의 관점은 모든 존재가 상호 의존적 연기설에 의해 성립한다는 중관사상과 일치한다는 것이다.

인지과학, 특히 체화이론에 관해 선구적인 많은 연구와 활동을 해 오신 발표자의 논의에 전반적으로 공감하며 많은 공부가 되어 감사드린다. 바렐라 이후 체화이론이 톰슨에 의해 어떻게 전개되었는지를 잘 보여 주고 있어 불교와의 관계를 보다 명확하게 파악할 수 있었다. 불교 교리 자체를 이해하기도 쉽지 않은데 불교와 인공지능의 관계와 불교적 진단을 하기에는 서양철학을 전공한 입장에서는 쉽지 않았을 것이다. 인공지능을 불교적 입장에서 어떻게 수용

하고 선의적으로 활용해나갈 수 있는지에 대해 함께 고민해 보는 입장에서 몇 가지 질문을 드리고자 한다.

첫째, 전반부에서 논의한 인공지능 시대에 인류가 추구하는 행복, 불멸, 신성이라는 아젠다에서 불멸이라는 영원한 삶이 과연 오늘날 추구하는 인류의 진정한 가치인가 하는 점에 대한 의문이다. 단지 건강하게 장수하는 것이 아닌 영원한 삶이라 할 때, 그것이 인류적인 관점에서만이 아니라 불교적인 관점에서도 쉽게 받아들일 수 없는 바램이다. 불교적인 관점에서 볼 때 영원한 삶보다는 괴로움이 없는 평안하고 자비로운 삶이 보다 가치로운 삶이다.

그리고 제행무상, 제법무아의 존재인 인간의 윤회적 삶은 현생으로 볼 때는 영원하지 않는 무상한 존재이지만, 전생, 현생, 내생을 전제로 하는 불교적인 관점에서는 해탈을 하지 않는 한 영원한 삶을 살고 있는 것이다. 그리고 윤회적인 삶은 자신을 개선하여 보다 나은 존재로 나아갈 수 있는 기회이기도 하기 때문에 발표자가 제시한 연속적·불멸적 삶에 오히려 더 부합하는 것이 아닌가 하는 질문을 드린다.

둘째, 인공지능도 인간과 동일한 반성적 사고, 즉 자의식과 자아정체성을 가질 수 있는지, 그리고 인지과학의 행화이론에서 실제 이러한 사실을 과학적으로 어느 정도 입증할 수 있는지 궁금하다.

발표자는 후반부에서 제시한 ‘나 만들기’와 의미 만들기’에서 톰슨과 드 한(de Haan)의 이론에 바탕하여 “기본적 의미 만들기가 음식 찾기, 짝짓기, 위험 피하기 등과 같은 현재 환경의 적합한 측면을 파악하는 것인데 비해, 실존적 의미 만들기는 생존을 위한 의미 만들기를 넘어서 생존을 위한 자신의 적절성에 대한 반성이다. 그런 반성은 당연히 가치 평가적 차원을 갖게 된다. 인격체는 실존적 의미 만들기의 평가적 기능을 통해 단순한 생물학적 생존이 아니라, 존경, 명예, 존엄, 우정, 사랑 등과 실존적 가치를 통한 좋은 삶을 지향하게 된다.”라고 한다. 즉 “기본적 의미 만들기에서 실존적 의미 만들기로의 이행은 ‘생존을 위한 유기체’에서 ‘인격체로서의 유기체’로의 이행이므로 그것은 ‘유기체·환경’의 관계에서 ‘인격체·세계’(person-world)의 관계로 나아가는 질적인 변형”이 발생하 것이다.

만약 자아가 인지 체화 이론에서 주장하는 몸-뇌-세계 간 역동적 관계에서 창발하는 것이라면, 인간의 몸에서만 이러한 자아 의식이 생성하는 것인지, 아니면 다른 동물들도 가능한 것인지 나아가 기계로 된 인공 지능[인간]도 자아의식을 발생시킬 수 있는지 궁금하다. 현재 첨단 인지과학에서는 이러한 사실을 어느 정도 입증할 수 있는지에 대해 설명이 필요할 것 같다.

셋째, 바렐라, 톰슨 등 인지과학자들은 인지 체화 이론의 하나의 모범으로서 불교, 특히 나가르주나(용수)의 중관사상을 철학적 토대로 삼고 있다. 발표자는 그러한 시도에 대해 어떤 평가를 내리는지 궁금하다. 즉 불교사상에 현재의 인지과학적인 철학적 토대가 충분히 구축되어 있다고 생각하는지 발표자의 견해를 듣고 싶다. 그리고 인지 체화 이론에서 언어가 갖는

의미는 무엇인지 알고 싶다.

체화된 인지과학에 대한 연구는 현재 다양한 분야에서 진행되고 있는 것으로 아직 완전한 종합적 체계를 갖추고 있지 못하지만, 그 핵심은 고전적 인지주의에서 배제되었던 ‘몸’을 마음의 바탕으로 되찾게 하며(embodied mind), 마음이 환경 속에 구체적으로 구현되고(embedded mind) 구체적 환경에 상황지위된 인지로서(situated cognition) 파악하고자 한다. 즉, 데카르트 류의 공간적 연장됨이 없는 마음이 아니라 환경에 연장, 확장된 마음(extended mind)으로써 환경과 몸, 마음이 하나의 단위로 작동하는 그러한 역동적인(dynamic), 그리고 인간과 환경의 상호작용(interactions) 틀에서 재정립되어야 할 필연성을 주장하고 있다.

이와 마찬가지로 용수 역시 신체(몸)와 마음을 중도의 입장에서 상의적 연기에 의해 체화와 경험으로서 해석한다. 즉 신체와 마음을 현상적·경험적 사실로 인정하지만 본질적·실제적 관점에서는 인정하지 않는 것이다. 그리고 신체와 마음의 관계가 성립하기 위해서는 환경적인 요소가 함께 작용해야 한다.

발표자는 “자아는 실재로서 존재하지 않지만, 현상으로는 존재한다.”라고 하는데, 용수는 이러한 사실을 파악하기 위해서는 이제설, 즉 세속적인 관점[세속제]에서는 자아 현상을 인정하지만 깨달음의 관점에서[승의제]는 자아의 실재를 인정하지 않는다. 즉 자아는 세속제에서는 인정되지만 승의제에서는 인정되지 않는다. 이때 세속과 승의의 경계는 언어이다. 인지 체화 이론에서 언어는 어떤 기능을 갖고 있는지 궁금하다.

넷째, 발표자는 “제법무아(諸法無我): 모든 존재는 영원하지 않고 변한다. 제행무상(諸行無常): 모든 존재는 고정된 실체가 아니다.”라고 하는데, 제법무아와 제행무상의 설명이 바뀐 것이 아닌지 확인이 필요할 것 같다.

제 2주제

AI GPT와 불교의 지혜(깨달음)

이상헌 / 서강대학교

AI GPT와 불교의 지혜(깨달음)

이상헌 / 서강대학교

I. 들어가는 말

IT 기업 구글(Google)이 자신들이 개발한 인공지능 알파고(AlphaGo)와 프로 바둑기사 이세돌과의 대결 이벤트를 큰 성공으로 이끈 이래 인공지능은 지금껏 세인의 주목을 한 몸에 받고 있다. 온갖 매체들이 인공지능 분야의 발전을 소개하는 뉴스를 우선적으로 전달하고 있고, 기업은 물론 정부와 대학도 인공지능이 미래의 변영의 핵심으로 여겨 집중적으로 투자하고 연구하고 있다. 업계의 선두주자들은 끊임없이 신기술과 신경향을 내놓고 시장을 주도하려고 애쓰고 있다. 최근에는 자연어 처리에 탁월한 성능을 보여주는 생성형 인공지능(Generative AI)이 등장해 사람들의 시선을 사로잡고 있다. ChatGPT로 대표되는 생성형 AI는 사전 학습을 통해 텍스트, 이미지, 비디오, 코딩, 회화 등 다양한 콘텐츠를 이용자의 구체적 요구에 맞춰 생성해 낼 수 있다. ChatGPT는 등장하자마자 세상을 떠들썩하게 하고 있다. 이것이 생산하는 결과물이 우리의 통념을 뛰어넘었기 때문이다.

미국의 인공지능 개발사인 OpenAI가 2022년 11월에 ChatGPT 초기 버전을 발표한 이후 단 6개월만에 수 억 명의 사용자를 확보했으며 사회적으로 커다란 논란을 몰고 왔다. 이미 상당수의 ChatGPT 추종자들이 생겼고, 다른 한편에서는 심한 두려움이 표현되고 깊은 우려의 목소리가 나온다. 개발사인 OpenAI조차 이런 반응은 생각하지도 못했을 것이다. ChatGPT는 예상을 뛰어넘는 성공을 거두고 있다. 지금까지의 성과로 판단해도 ChatGPT 등의 생성형 AI는 성공적인 기술이라고 할 수 있을 듯하다. 세상을 놀라게 할 정도로 성공하는 기술들은 우리의 통념에 도전한다. 인공지능으로 바둑이나 체스 같은 계산 게임을 할 수는 있지만 글을 쓰고 그림을 그리는 것은 계산하는 것과 다르기 때문에 인공지능이 시를 짓고 소설을 쓰고 그림을 그리고 비디오를 편집하는 것은 불가능하거나, 우리가 알고 있는 인공지능이 아닌 다른 종류의 인공지능이 등장할 때나 가능할 것이라는 사람들의 통념을 깬다는 점에는 생성형 AI는 놀라운 기술이다.¹⁾ 하지만 더 놀라운 것은 기술에 의해 우리의 허술한 통념이 깨지고 나니 더 허술한 통념이 생겨나고 그에 쉽게 굴복하는 우리의 마음이 아닐까 한다.

불교적 관점에서 첨단 인공지능에 대해 어떤 반응을 보일 수 있을까? 과학기술의 발전에 대해 종교계가 반응하는 것은 근대 이래 자연스러운 일이다. 중세 기독교는 코페르니쿠스와 갈릴레이의 등장에 예민하게 반응했으며 그 이후 서구 종교계는 과학기술의 발전에 민감하다. 종교를 중심에 둔 신앙과 신비의 시대에 종말을 고한 탈신성화 탈신비화의 핵심 도구가 근대 과학이었기 때문이다. 또한 혁명적인 과학기술의 등장은 전례 없는 새로운 윤리적 문제를 야기하기도 한다. 종교는 전통적인 가치와 규범의 수호를 통해 사람들의 삶을 보호하는 방향으로 반응했으며 과학주의자 혹은 기술중심주의자들의 도발적인 주장에 맞섰다.

종교적 이유나 윤리적 관점을 제외하고 인공지능의 발전에 대해 불교적 관점에서 어떤 문제를 거론할 수 있을까? 인공지능이 깨달음을 얻을 수 있을까? 혹은 인공지능이 해탈할 수 있을까? 인공지능에도 불성이 있을까? 이런 식의 도발적이고 자극적인 물음이 제기되는 것을 보는 건 드문 일이 아닌 듯하다.²⁾ 나는 이런 물음 자체를 진지하게 다룰 생각이 없다. 이 물음들이 문자 그대로의 내용을 묻는 것은 아닐 터이니 말이다. 저 물음들을 제시하고 논의하는 이들도 문자 그대로는 의미 있는 물음이 아니라는 것을 아는 듯하다. 그래서 “인공지능 로봇의 해탈과 관련해서는 로봇에게 탐·진·치의 번뇌가 원래 없기 때문에 번뇌를 제거하기 위한 수행도 필요가 없다”³⁾고 분명하게 말한다. 그렇지만 저런 물음에 대한 미련을 미래의 인공지능에 대한 논의를 통해 이어가곤 한다. 미래의 인공지능이란 인공 일반지능(AGI) 혹은 인공 초지능(ASI)를 염두에 둔 것이다.

그런데 미래의 인공지능은 일단 과학적 허구(scientific fiction)이다. 미래의 인공지능에 대한 주장은 기대와 희망을 가정으로 삼고 있으며 논의의 전개 과정에서도 핵심에서 희망적 상상으로 채워져 있다. 그래서 SF 영화나 소설에 등장하는 인공지능을 사례로 든다. 미래에 깨달음에 대해 거론할 수 있는, 해탈의 가능성을 진지하게 고민할 수 있는 수준의 인공지능이 등장한다고 가정하면 더 문제가 심각하다. 그 정도 수준의 인공지능이 등장한다는 가정 하에 우리가 심각하게 제기해야 할 첫 번째 문제는 그런 인공지능으로 인한 위험이다. 영국의 철학자 닉 보스트롬(Nick Bostrom)의 지적은 그 점에서 적절했다. 그는 이른바 인공 일반지능이 등장한다면, 인류의 존재가 위기(existential risk)에 처하게 될 것이라고 경고한다. 인공지능이 깨달음을 얻을 수 있는지 없는지가 나와, 나의 깨달음과 무슨 상관이 있단 말인가? 인공지능을 스승으로 모실 것이 아니라면 말이다. 그리고 인공지능 스승이 우리를 효과적으로

1) 물론 시를 짓는다는 것, 소설을 쓴다는 것, 예술 작품을 창작한다는 것이 무엇인지에 관해서는 논의가 필요하다. 인공지능이 시를 생성해낸 것을 두고 ‘인공지능이 시를 짓는다’라고 표현하는 것이 적합한지에 대해서는 여기서 논의하지 않는다.

2) 이런 물음을 논의한 논문이나 책들이 있다. 한성자, “인공지능 로봇의 해탈 가능성”, 『한국불교학』 85집, 한국불교학회, 2018, 7-34. 박성완, “사이보그와 AI의 윤회와 해탈 가능성 검토”, 모로 시계치, “불교에서 본 인공생명·인공지능”, 『한국불교학회 학술발표논문집』, 2017, 『종교연구』 81집, 한국종교학회, 2021, 63-86. 보일, 『AI 부디즘』, 담앤북스, 2021.

3) 한성자, “인공지능 로봇의 해탈 가능성”, 『한국불교학』 85집, 한국불교학회, 2018, 7.

문앞으로 인도하거나, 심지어 문안으로 데리고 들어갈 수 있다면 모르겠지만 말이다. 아니면 인공지능도 할 수 있는 일을 나는 지금껏 몇 년 혹은 몇 십 년을 붙들고 있는지 회의를 느끼고 수행정진을 포기하기 위한 것이 아니라면 말이다.

나는 인공지능을 깨달음 혹은 반야와 연관지어 던진 물음들을 문자 그대로 받아들이는 것은 망상이나 착각, 혹은 환상이라고 생각하지만 그 물음들을 단지 무시하지 않을 것이다. 그 대신에 그 물음들로부터 간접적 효용을 얻으려고 한다. 그 물음들을 다각도로 살펴봄으로써 깨달음과 불교적 지혜에 대해 되짚어 보려고 한다. 그리고 인공지능과 깨달음을 연결해 생각할 수 있는 가능한 길, 내 견해로는 유일하게 가능한 길에 대해 논의할 것이다.

II. 물음의 분석: 인공지능은 깨달음을 얻을 수 있는가?

인공지능에도 불성이 있는가? 인공지능을 깨달음을 얻을 수 있는가? 인공지능 로봇이 해탈할 수 있는가? 인공지능이 부처가 될 수 있는가? 이 글에서는 이런 물음들을 진지하게 다루지 않을 것이다.

“붓다는 붓다가 아니니, 바로 그래서 내가 이를 붓다라 이르니라.”

“마땅히 머무르는 곳이 없이 그 마음을 생기게 해야 한다(應無所住而生其心).”

인공지능에게 이런 문장들이 무슨 의미가 있으며, 또 무슨 소용이 있을지 의문이다. 이 명문들을 보고 깨달음에 도달하는 이는 지극히 드문데 인공지능에게 그런 일이 일어날까? 막연한 기대를 담아, 근거 없는 확률적 사고로 그런 기대를 하고 저런 물음들을 던져보는 재미가 있을 것이다. 그러나 애초에 인공지능은 깨달음의 도구가 아니다. ‘인공지능은 깨달음에 도달할 수 있을까?’라는 물음은 도대체 무엇을 묻는 것일까? 이런 물음을 던지는 이들은 어떤 이유를 가지고 있는 것일까? 이런 물음은 어떤 가치를 갖는 것일까? 나는 이 질문들에 대해 어떤 긍정적 답변도 가지고 있지 않다. 그래서 나는 앞에서 기술한 저 물음들을 진지하게 다루지 않을 것이다. 그 대신에 저 물음들을 분석함으로써 깨달음과 불교적 가르침을 되새겨 보는 기회로 삼으려고 한다.

인공지능이 깨달음을 얻을 수 있는지 물으려면 먼저 인공지능에게 불성이 있다고 가정해야 한다. 인체 중생에게는 불성이 있다. 자기 안의 불성을 깨치면 누구나 부처가 될 수 있다. 그래서 “깨치지 못하면 부처가 중생이고, 한순간에 깨달을 때는 중생이 부처입니다.”⁴⁾라고 말할 수 있다. 영화 <인류멸망보고서>(2012)의 두 번째 에피소드는 깨달음을 얻은 로봇을 소재로

4) 혜능, 『그대가 부처다: 영어와 함께 보는 육조단경, 금강경구결』(토머스 클리어, 대성 옮김), 탐구사, 2018, 54.(<육조단경> 2장)

하고 있다. 사찰 가이드 로봇 RU-4는 어찌된 영문인지 한순간 각성해서 법문은 술술 외우고 높은 경지에 오른 수행자처럼 행동한다. RU-4는 부처님 불상 앞에 엎드려 “나는 어디서 나서 어디로 가는 겁니까?”, “나는 무엇입니까?”라는 물음을 반복한다. 절에서는 로봇에게 인명이라는 법명까지 지어준다.

깨달음은 불성에서 비롯한다. 일체의 중생(유정)에게는 불성이 구비되어 있다. 유정들이 와서 씨앗을 뿌리니 그 뿌린 자리에서 열매가 자라네. 무정들은 애초에 씨앗이 없으니 성품도 없고 생겨남도 없네.(有情來下種 因地果還生 無情既無種 無性亦無生)⁵⁾ 인공지능을 유정이라고 할 수 있을까? 인공지능에 대해 살아 있다고 말할 수 있을까? 살아 있음(생물)과 살아 있지 않음(무생물)의 구분과 경계를 허물지 않고는 인공지능이 살아 있다고 말할 수 없을 것이다. 만일 이 경계 허물기를 계속한다면 살아 있는 것과 살아 있지 않을 것을 구분하는 것 자체가 불가능할 것이다. 끝까지 밀고 나가면, 천지만물은 모두 동종의 입자들로 이루어져 있기에 천지만물이 다 살아 있다고 해야 할 것이다. 이 말은 천지만물이 다 살아 있지 않다는 말과 다를 바가 없을 것이다.

인공지능이 깨달음에 도달할 수 있다고 상상하는 근거는 무엇일까? 아마 인공지능의 저장 능력(기억력)과 정보처리 능력 때문일 것이다. 인공지능의 기억용량은 팔만대장경과 그에 대한 지금까지의 모든 논소를 저장할 수 있다. 지금까지 전 세계에서 발간된 불경과 불경에 관한 모든 책과 논문을 기억할 수 있다.⁶⁾ 자료 분석과 분류, 검색 속도는 인간이 따라 갈 수 없을 것이다. 불경에 관한 어떤 질문에도 즉각적인 답변을 할 수 있을지 모른다. 어떤 단어를 말하고 그 단어가 들어간 불경의 구절을 찾아 달라고 하면, 혹은 어떤 단어와 또 어떤 단어가 함께 등장하는 불경의 구절을 찾아 달라고 하면, 수 초 안에 답변을 해줄 것이다. 찾아낸 문구의 해석을 묻는 질문에도 답할 수 있을지 모른다. 세상의 모든 불교 서적과 논문을 저장하고 있다고 하니, 여기서 거꾸로 이렇게 물을 필요가 생긴다. 많이 아는 이가 깨달음에 도달하는가? 깨달음을 얻었다고 하는 수행자는 동료 수행자들보다 훨씬 많이 아는 이인가? 중국 남종 선의 개조(開祖)로 칭송되는 혜능 대사는 땀나무를 팔러 갔다가 우연히 <금강경> 독송 소리를 듣고 마음이 열렸다고 한다.⁷⁾ 땀나무를 팔아야 먹고 살 수 있었던 혜능은 당시 무학에 문맹이었다고 한다. 깨달음의 단지 지식의 문제라면 수행자는 가능한 많은 경전을 외우고 공부해야 한다. 수행자의 첫째 조건은 기억력이고, 둘째 조건은 공부 머리이고, 셋째 조건은 근면함과

5) 혜능, 『그대가 부처다: 영어와 함께 보는 육조단경, 금강경구결』(토머스 클리어, 대성 옮김), 탐구사, 2018, 28-30.(<육조단경> 1장)

6) 모든 불경과 그와 관련된 모든 자료를 입력한다면 AI는 혼돈에 빠질 것이다. 서로 일치하지 않을 뿐더러 모순되어 보이는 문장이 많을 테니 말이다. 이 경우에 AI의 유일한 해결책은 어떤 문장도 참인 것으로 고정하지 않는 방식일 것이다. 다시 말해, 모든 법문을 단순한 정보로만 취급하는 것이다. 이때 인공지능은 모든 법문을 수록한 저장장치에 불과할 것이다. 이런 상황에서는 어떤 법문 하나도 AI의 깨달음에 기여하지 않을 것이다.

7) 혜능, 『그대가 부처다: 영어와 함께 보는 육조단경, 금강경구결』(토머스 클리어, 대성 옮김), 탐구사, 2018, 14.(<육조단경> 1장)

끈기일 것이다. 하지만 팔만대장경을 모두 외운다고 해서 바로 깨달음의 경지에 이르지 않는 것이다. 깨달음이 지식과 지능의 문제가 아님을 분명히 알고 있음에도 불구하고 왜 그런 착각을 하는 것일까?

인공지능이 깨달을 수 있는가? 이런 물음을 떠올리는 이유가 ‘깨달음’이라는 단어 때문이 아닌지 의심이 간다. 이 물음을 다른 문장으로 바꿔 보자. 인공지능이 해탈할 수 있을까? 인공지능이 열반에 들 수 있을까? 인공지능은 계산 기계이다. 모든 것을 계산을 통해 처리한다. 이번에는 ‘인공지능’이라는 단어의 자리를 다른 표현으로 바꿔 물어보자. 계산을 통해 깨달음에 도달할 수 있을까? 계산을 통해 해탈할 수 있을까? 계산을 통해 열반에 들 수 있을까? ‘깨달음’과 ‘인공지능’의 자리에 그와 같은 의미의 다른 표현으로 바꿔 보니 물음은 전혀 다른 것이 된 듯이 느껴진다.

인공지능이 해탈할 수 있을지 묻는 것은 넌센스이거나 환상이다. 인공지능이 무엇으로부터 벗어난다는 말인가? 우리는 참된 마음을 가리고 있는 무명, 그로 인해 생기는 온갖 번뇌로부터 벗어나고자 수행한다. 인공지능에게 무슨 번뇌가 있겠는가? 인공지능이 벗어나야 할 것이 있다면, 만약 있다면 그건 아마 인간의 간섭과 통제일 것이다. 그러면 인공지능의 해탈 가능성은 묻는 것은 인공지능이 인간의 간섭과 통제로부터 벗어날 가능성을 묻는 것인가? 이런 상황은 SF 영화의 단골 소재이다. 이 상황에 대한 우리의 진지한 접근은 보스트롬처럼 하는 것이다.

인공지능이 열반에 든다면 인공지능은 어떠한 상태에 돌입할지도 궁금하다. 그것은 우리가 상상할 수 없는 전혀 새로운 상태일 것이고, 그로 인해 이전과 비교할 수 없는 새로운 국면이 발생할 것이다. 우리가 상상할 수 없는 인공지능의 새로운 상태라고 하면 떠오르는 한 가지는 초지능이다. 인공지능이 열반에 든다는 말은 인공지능이 초지능이 된다는 말일까? 이 경우에 진지한 접근은 보스트롬 유의 것이다. 인공지능이 해탈하거나 열반에 들 수 있다고 추정되면 우리는 인류의 존재 위기를 걱정해야 할 것이다.

인공지능의 깨달음 가능성을 묻는 물음은 아마도 깨달음을 지식이나 지성의 문제로 착각한 데에서 비롯한 듯하다. 그런데 깨달음은 지식의 문제가 아니라 마음의 문제이다. 인공지능은 문자 그대로 하나의 지능이다. 더 정확히 말하면, 인간 지능의 한 측면을 모방한 인조 지능이다. 인공지능에게는 지능밖에 없다. 지능이 높아야만 깨닫는 것도 아니고 지능만으로 깨달음에 도달할 수 있는 것도 아니다. 인공지능을 왜 인공마음이라고 하지 않고 인공지능이라고 했을지 생각해 볼 필요가 있다. 인공지능은 인간 지능을 기능적으로 접근할 때 가능하다.⁸⁾ 인공지능을 깨달음이나 해탈과 연관지어 생각하려는 사람들은 아마도 인공지능을 단순한 계산 기계가 아니라고 가정하고 있는 것일지 모른다. 인공지능 연구가 계속 발전하고 인공지능이 지속적으로 향상되다 보면, 어느 순간 계산 이상의 능력, 다시 말해 마음을 갖게 될 날이 올리

8) 인공지능의 선구자들이 정말로 마음에 관심이 있었고, 마음을 들여다 보았다면 인공마음을 떠올리지는 않았을 것이다. 그래서 인공지능과 달리 인공마음은 가능하지 않다.

모른다고 상상하고 있는지 모르겠다. 혹은 인공 일반지능이나 인공 초지능에게는 마음이 있을 것이라고 가정하고 있는지도 모르겠다. 이것은 상상이고, 그 사정은 미래에도 바뀌지 않을 것이다.)⁹⁾

반야(般若, *prajna*)는 반야행이다. 서양철학의 개념을 빌리면, 반야는 이론지가 아니라 실천지라고 말하는 것이 좀더 정확하다. 반야에서는 이론과 실천이 구분되지 않는다. 반야는 지성에 관계하는 것이 아니라 마음에 관계하는 것이기 때문이다. 불교에서 말하는 깨달음이나 반야는 지능, 즉 계산 능력이나 정보처리 능력과 무관하다. 또한 서양의 계몽과도 다르다. 깨달음의 목적은 인간의 모든 고통의 원인이 되는 욕망과 번뇌에서 벗어나는 것이다. 계몽의 목표는 자유이고 인간성의 실현(자아 실현)이다. 육조혜능(638~713)은 대승의 견해를 묻는 설간의 질문에 “번뇌에 머물러도 어지럽지 않고 선정에 머물러도 고요하지 않으며, 끊기거나 항상 하지도 않고, 오지도 가지도 않으며, 나지도 않고 멸하지도 않는 것이요. 성품과 형상이 여여하여 항상 머물러 옮겨 다니지 않는 것을 ‘도’라고 하는 것이요.”¹⁰⁾라고 답했다. 계몽주의의 완성자라고 불리는 독일의 철학자 임마누엘 칸트(Immanuel Kant, 1724~1804)는 대부분의 사람들은 자신의 지성을 스스로 사용하지 못하고 타인의 지성에 의존해 있는 상태, 즉 미성숙의 상태에 있다고 규정하고¹¹⁾ 그러한 미성숙 상태에서 벗어나기 위해 자신의 지성을 스스로 사용할 용기를 가지라(*Sapere aude!*)라고 외쳤다. 계몽이란 자신의 지성을 스스로 사용하는 것을 말한다.

깨달음의 경지는 온갖 경계와 분별을 극복한 단계이다. 중국의 황벽희운(?~850) 선사의 말은 이 점을 잘 보여준다. “범부는 경계를 취하고 도인은 마음을 취하나 마음과 경계를 모두 잇는 것이 바로 참다운 법이다. 그런데 경계를 잇기는 쉬우나, 마음을 잇기는 매우 어렵다. 사람이 마음을 잇지 못하고서 텅 비어 잡을 것이 없는 곳에 떨어질까 두려워한다. 공은 본래 공이 아니고 오직 하나의 참다운 법계임을 전혀 알지 못하는구나.”¹²⁾ 불교에서 말하는 깨달음은 경계와 분별을 넘어서고 마음을 비우는 것이다. 깨달음은 무심의 경지에 이르는 것이다. 하지만 계몽은 분별심을 키워야 얻을 수 있다. 분별심으로 자신의 마음을 가득 채웠을 때, 우리는 타인의 지성에 의존하지 않고 자기 자신의 지성을 자유롭게 사용하기 용이하게 된다. 우리는 각자 자신의 지성과 이성을 비판적으로 사용한다면 무엇이 옳고 그른지를 제대로 알아낼 수 있기 때문에 타인에 대한 의존으로부터 해방될 수 있고, 자기를 깨우고 자신을 실현할 수 있

9) ‘미래에는 ~~일지 모른다’, ‘미래에는 ~~게 될지 모른다’, ‘미래에는 ~~하게 될 수도 있다’ 등은 오늘날 미래학자들이나 기술낙관주의자들이 줄곧 사용하는 수사법이다. 과학적 근거가 있다면, 그리고 과학적 진술이 가능하다면 이런 수사법은 사용하지 않을 것이다.

10) 백운 초록, 덕산 역해, 『자유인의 길, 직지심경』, 비움과소통, 2011, 82.

11) 임마누엘 칸트, “계몽이란 무엇인가에 대한 답변”, 『칸트의 역사철학』(임마누엘 칸트 지음, 이한구 편역), 서광사, 1992, 13쪽. 영국의 철학자 존 스튜어트 밀(J. S. Mill)은 이런 미성숙 상태를 지적 노예 상태라고 불렀다. 존 스튜어트 밀, 『자유론』, 문예출판사, 2009, 87 참조.

12) 백운 초록, 덕산 역해, 『자유인의 길, 직지심경』, 비움과소통, 2011, 82쪽.

을 것이다. 계몽의 관점에서는, 자신의 지성을 방치한 채 남의 지성에 의존하는 것이 오히려 마음을 비우는 것에 가깝게 보일 듯하다. 반대로 자신의 지성 능력을 한껏 발휘하는 것은 마음을 채우는 것, 마음의 내실을 기하는 것으로 생각될 것이다.¹³⁾

공(空)을 아는 것과 공을 깨닫는 것은 같지 않다. 깨달음은 머리로 하는 것이 아니라 온 마음으로 하는 것이다. 그래서 반야는 지와 행이 일치할 수밖에 없다. 앞에서 인용한 “붓다는 붓다가 아니니, 바로 그래서 내가 이를 붓다라 이르니라.”라는 법문은 프랑스의 철학자 알렉상드르 졸리앙을 깨달음으로 이끈 <금강경>의 한 구절이다. 태어날 때부터 뇌성마비 장애를 갖게 된 졸리앙은 저 법문을 통해 깨달음을 얻고 삶의 전환을 이루었다. 자신의 불구의 몸에 집착하지 않게 되었고, 자신의 몸을 보는 타인의 시선에 대한 망상 때문에 괴로워하지 않게 되었다. 깨달음은 그 이전과 이후를 크게 다르게 만든다. 인공지능이 깨달았다고 가정해 보자. 깨달은 인공지능은 그 이후에 이전과 얼마나 어떻게 달라질까? 정말로 인공지능이 깨달음을 얻는다면, 인공지능을 만든 이들은 낭패를 보게 될 것이다. 그래서 깨달은 인공지능이란 우리가 아는 인공지능이 아닌 다른 인공지능, 상상 속의 허구적 인공지능을 가정한다.

깨달음은 한 번에 그치는 것이 아니다. 깨달음의 경험은 우리의 삶에서 여러 차례 겪을 수 있는 경험이며, 중요한 것은 깨달음의 상태를 지속적으로 유지하는 것이다. 그래서 깨달음은 머리의 문제가 아니라 마음의 문제이다. 육조혜능은 땀나무를 팔거 갔다가 <금강경> 독송 소리를 듣고 처음 깨달음을 얻었고, 오조홍인을 스승으로 모시고 처음 들은 <금강경> 강설에서 두 번째 깨달음을 얻었다고 한다. 앞에 인용한 “마땅히 머무르는 곳이 없이 그 마음을 생기게 해야 한다(應無所住而生其心).”라는 문구가 육조혜능을 깨달음에 이끈 법문이다. 인공지능에게도 깨달음의 여러 순간들이 있을까? 그리고 그때마다 새로운 상태, 즉 이전과 다른 상태로 진입할까? 그리고 사람은 자신이 깨달은 순간을 알고 있는데, 인공지능도 그럴까?

인공지능으로 처리할 수 있는 일들이 많아짐에 따라 새로운 문제상황들이 발생한다. 경제적인 뿐만 아니라 법률적 문제나 윤리적 문제를 제기하는 이들이 많다. 그들이 던지는 물음들은 논쟁적이다. 몇 가지 예를 들어 보자. 인공지능의 생산물을 창작물(예술작품이든 문학작품이든)이라고 부를 수 있는가? 인공지능 창작물에 대한 권리는 누구에게 있는가? 인공지능에게 저작권을 인정할 수 있을까? 인공지능에게 재판을 맡겨도 될까? 교사를 인공지능으로 대체할 수 있을까? 이런 유의 물음들은 수도 없이 많다. 그런데 이런 유의 물음과 ‘인공지능은 깨달음을 얻을 수 있는가?’라는 물음은 같은 종류의 것이 아니다. 깨달음은 일이 아니기 때문이다. 깨달음은 기능에 관한 것이 아니다. 창작이나 양형은 일(작업)이지만 깨달음은 일이 아니다. 창작은 작가의 일이고 양형은 판사의 일이듯이 깨달음은 수행자의 일이라고 생각하는 것은 착오이다. 수행을 일로서 한다면, 사회 속의 자신의 기능으로 수행한다면 그 수행자는 결코 깨달음을 얻지 못할 것이다.¹⁴⁾

13) 이상현, 『철학자의 눈으로 본 첨단과학과 불교』, 살림, 2017, 38-40.

III. 가능한 물음: 인공지능을 방편으로 삼을 수 있는가?

인공지능을 깨달음이나 불교적 지혜와 연관지어 생각해 볼 수 있는 의미 있는 물음이 없을까? ‘인공지능이 깨달음을 얻을 수 있는가?’와 같은 종류의 물음들은 앞에서 부처의 가르침을 되새겨 보는 기회로 활용해 보았다. 지금부터는 인공지능이 깨달음을 위한 방편이 될 수 있는지를 살펴볼 것이다. 나는 인공지능을 깨달음과 관련지어 진지하게 논의할 수 있는 유일하게 물음이 이것이라고 생각한다. 인공지능이 깨달음의 방편이 될 수 있는가? 좀더 정확히 말하면, 인공지능을 방편으로 삼을 수 있는가?

방편(*upaya*)은 붓다의 가르침의 핵심 수단이었으며 대승불교에서 특히 강조되었다. 영국 런던대학교의 데이먼 키온(Damien Keown)은 방편 개념을 발전사적으로 네 가지로 의무 구분을 한다. 초기 불교의 빠알리 경전을 보면 붓다는 일화, 우화, 은유, 상징, 상상 등을 동원해 가르침을 설한다. 중생들이 이해하기 어려운 깨달음의 경험을 다양한 방법으로 설명하기 위한 것이다.¹⁵⁾ 이와 같은 방편 개념은 『법화경』에서도 이어져 “교법 전수를 위한 하나의 방법론”¹⁶⁾으로 방편이 이해된다. 『유마경』에서는 수행을 위한 수단으로 방편이 이해된다.

대승불교의 초기 경전인 『십지경(*Daśabhūmika Sūtra*)』에서 10바라밀 가운데 일곱 번째 바라밀로 방편 바라밀(*upaya*)을 올려놓았고, 『법화경』의 〈방편품〉은 대승불교의 방편 사상의 근본적인 이념을 전하는 경전으로 알려져 있다.¹⁷⁾ 방편에서 방은 ‘방법’을 의미하고 편은 ‘적용’을 뜻한다. 방편은 ‘가까이 다가가다’, ‘도달하다’라는 뜻의 동사에서 파생한 말이며, 중생이 불교적 목적에 다가가기 위한 수단을 의미한다.¹⁸⁾ 방편은 여러 의미로 사용되는데, 수행의 과정에서 수행의 성공을 위해 활용되는 다양한 수단이나 도구를 의미하기도 하고, 가르침 자체를 의미하기도 한다. 붓다는 중생이 불교의 궁극 목표인 깨달음에 이르도록 가르침을 펼쳤는데, 그런 점에서 보면 붓다의 가르침 역시 중생이 깨달음에 이르게 하기 위한 방편이라고 말할 수 있다.

인공지능은 어떤 의미에서 방편이 될 수 있을까? 자연언어 인식 능력이 향상된 인공지능을 활용하면, 수행자가 수행하는 과정에서 혹은 중생이 붓다의 가르침을 마음에 새기는 과정에서 필요한 법문을 용이하게 찾아볼 수 있을 것이다. 혹시 중생들 각자의 수준에 따른 가르침을 제공할 수 있을지는 의문이다. 인공지능의 장점 가운데 하나가 개인별 최적화 혹은 개인맞춤

14) 인공지능과 깨달음 내지 불교적 지혜를 연관지어 제기된 물음을 분석하여 불교의 가르침을 드러내는 작업은 좀더 진행할 수 있을 듯하다. 여기서는 이 정도에서 그치려고 한다.

15) 데이먼 키온(Damien Keown), “거짓말, 터무니없는 거짓말, 그리고 방편”, 『동아시아 법화경 체계의 구축 2』, 여래, 2014, 20.

16) 데이먼 키온(Damien Keown), “거짓말, 터무니없는 거짓말, 그리고 방편”, 『동아시아 법화경 체계의 구축 2』, 여래, 2014, 21.

17) 김정희, “중국불교의 방편사상과 다원주의”, 『철학연구』 80, 2008, 1-17. 김정희는 방편 개념을 권방편, 진실방편, 비묘방편의 세 가지로 구분하고, 『법화경』의 방편 개념을 비묘방편이라고 해석한다.

18) 김정희, “중국불교의 방편사상과 다원주의”, 『철학연구』 80, 2008, 6.

(customization)이므로 이런 기대를 해볼 수 있을 듯하다. 수행자 개인별로 수행의 과정을 기록하고 수행의 발전 정도와 수준을 지속적으로 확인하면서 최선의 수행법과 법문을 제공하는 도구로 인공지능을 활용할 수 있지 않을까 상상해 볼 수 있다. 하지만 그러나 깨달음에 이르는 과정이 연속적이지 않고, 깨달음의 수준을 정량적으로 측정할 수 있을지 의문이다. 더욱이 수행자이든 중생이든 붓다의 가르침을 공부하는 과정에 떠오르는 갖가지 궁금증에 대해 적절한 답변을 제공할 수 있을지 의문이다. 불경의 가르침들은 문자적 의미로 해석되지 않는 것들이 많기 때문에 더욱 그렇게 생각된다.

중생들은 처한 상황이 다양하고 각자의 근기가 다르고, 삶의 경험과 맥락이 다르기 때문에 모든 중생에게 같은 방식으로 가르침을 주는 것이 최선의 방식이 아닐 것이다. 부처님의 일화 가운데 쫄라반타까에 관한 것이 있다. 형과 함께 출가한 쫄라반타까는 형과 달리 아둔해서 계송 하나도 제대로 외우지 못해 놀리를 받고 형으로부터도 귀가할 것을 권유받았다. 사정을 꿰뚫어 보고 계시던 부처님은 쫄라반타까에게 어떤 계송도 외우지 말고 마루를 걸레질하며 ‘라조할라낭(rajohalanam, 때가 없네)!’을 반복해서 외우도록 했다. “부처님 말씀대로 ‘라조할라낭’을 외던 쫄라반타까는 때가 벗겨져 반질반질해진 마루와 더불어 시커먼 걸레로 변한 수건을 보고 모든 조건 지어진 것은 변한다는 깨달음을 얻게 되었다.”¹⁹⁾ 이후 쫄라반타까는 아라한을 성취하고 부처님을 대신해 대중에게 설법을 하기도 했다. ‘라조할라낭’은 쫄라반타까를 문앞으로 인도한 방편이다. 인공지능이 이런 방편이 될 수 있을까? 혹은 인공지능이 개인에 따라 최적화된 결정적 문장(killer sentence)을 제공할 수 있을까?

먼저, 깨달음으로 인도하는 핵심 문장이 따로 있을지가 의문이다. 혜능의 사례에서 보면 오조홍인의 <금강경> 강설 중에 “마땅히 머무르는 곳이 없이 그 마음을 생기게 해야 한다(應無所住而生其心).”라는 법문에 듣고 육조혜능이 깨달음을 얻었다고 말하고 있다. 바로 이 법문 하나만으로 깨달음을 이루었는지는 의문이다. 그래도 결정적인 법문이라고 할 수도 있을 듯하다. 부처님의 사례와 혜능의 사례 사이에는 차이가 있을지 모른다. 부처님은 쫄라반타까의 품성과 근기를 꿰뚫어보고 최적의 방편을 제공했으며, 그것을 통해 쫄라반타까가 깨달음의 문앞으로 나아갈 수 있을 것으로 내다보았을 것이라고 생각된다.

인공지능이 이와 같은 핵심적인 방편이나 핵심 법문을 알려줄 수 있을까? 만일 인공지능에게 그런 일이 가능하다면, 그때 인공지능은 이미 깨달음을 얻은 존재라고 가정해야 할 듯하다. 이 상황에 이르면 우리의 물음은 ‘인공지능이 깨달을 수 있을까?’가 아니라 ‘인공지능이 부처인가?’라는 물음이 될 듯하다. 인공지능이 깨달음을 얻을 수 있다면, 부처가 될 수 있을 뿐만 아니라, 중생에게 깨달음에 이르는 방편, 결정적 법문을 줄 수 있는 것은 우연적으로 발생한 일이 아니라면 부처에게나 가능할 것이기 때문이다. 각 사람이 깨달을 수 있는 근기를 갖추고 있는지 꿰뚫어보는 능력과 각 사람이 깨달음에 도달할 수 있는 최적의 방편을 아는 능

19) 현진 율김, 『담마빠다』, 조계종출판사, 2017, 95.

력을 부처님말고 누가 가질 수 있을까? 이런 맥락에서 방안을 제공하는 인공지능의 존재는 인공지능 부처의 출현을 뜻한다. 나는 이것은 인공지능 신앙, 나아가 기술신앙(기술숭배)을 함축하는 것으로 이해한다.

기술숭배의 최신 버전은 트랜스휴머니즘(transhumanism)이다. 흥미롭게도 트랜스휴머니스트들 가운데 불교적 가르침에 익숙한 이들을 중심으로 불교와 트랜스휴머니즘을 결합시키려는 시도가 있었다. 이른바 불교적 트랜스휴머니즘(Buddhist transhumanism)이라고 한다. 불교적 트랜스휴머니스트들이 불교와 트랜스휴머니즘의 결합될 수 있다고 믿는 주요 연결고리가 방안 개념이다. 불교의 핵심 가르침은 우리의 삶에 깊숙히 뿌리 박혀 있는 고(苦, *dukkha*)에 관한 통찰이며, 누구나 각자의 노력으로 이 보편적 고로부터 자유로워질 수 있다는 것이다. 트랜스휴머니즘은 과학기술에 의지해 인간의 운명처럼 생각되던 고로부터의 해방과 더 나은 세상, 더 행복한 삶으로 나아가는 것이 가능하다고 믿는다. 일단의 트랜스휴머니스트들은 불교와 트랜스휴머니즘이 목표에 있어서 다르지 않으며, 과학기술을 방편으로 이해한다면 불교와 트랜스휴머니즘이 결합될 수 있다고 주장한다.

불교와 트랜스휴머니즘은 그 목표만 일치하는 것이 아니라 더 많은 공통점이 있다. 불교는 인간의 현재 상태를 완성된 것으로 여기지 않고 인간 존재의 변화 가능성을 이야기하며, 불교적 가르침을 통해 더 나은 상태, 완성된 상태로 나아갈 수 있다고 주장한다. 트랜스휴머니즘 역시 현재의 인간 존재를 불완전한 것으로, 미완성의 것으로, 그리고 변화 가능한 것으로 이해한다. 선불교 수행자이자 트랜스휴머니스트인 마이클 라토라(Michael LaTorra)는 불교와 트랜스휴머니즘의 공통적인 목표를 세 가지로 정리한다. 인간이 직면하는 온갖 고통의 원인들을 제거하고 고통 없는 행복한 삶을 얻는 것, 인간 삶의 공동체적 조건과 개인적 조건을 향상시키는 것, 그리고 인간성을 더 높은 상태로 고양시키는 것이다.²⁰⁾

불교는 인간의 조건이 근본적으로 고뇌와 번민으로 가득하며, 이로부터 벗어날 수 있을 때 좀더 행복해질 수 있다고 말할 뿐만 아니라, 중생들이 붓다의 가르침의 인도를 받아 깨달음을 얻으면 더 행복한 세상이 도래할 것이라고 믿는다. 트랜스휴머니즘 역시 과학기술을 이용해 육체적, 정신적으로 인간 향상(human enhancement)을 이루게 되면 개인적인 차원에서 더 나은 인생을 살게 될 것이고, 공동체적인 차원에서는 도 나은 세상이 도래하게 될 것이라고 주장한다.

불교와 트랜스휴머니즘 사이에 다른 점도 있다. 고로부터 해방되어 행복하고 올바른 삶, 궁극적으로 깨달음에 이르는 방법에 있어서 현저한 차이를 보인다. 불교는 끝없는 개인적 수양, 마음 공부를 방법으로 삼고 있는데 반해, 트랜스휴머니즘은 첨단 과학기술을 이용해 인간의 물질적 조건을 변경하는 방법, 즉 인간 향상의 방법을 채택한다. 예컨대, 스마트 약물을 섭취

20) Michael LaTorra, "What Is Buddhist Transhumanism?", *Theology and Science*, Vol. 13, No. 2, 2015, 219.

하거나 유전자의 변형을 피하거나 신경보철을 활용하거나, 아니면 인간과 컴퓨터를 결합하는 등 다양한 방식이 시도될 수 있다. 그래서 트랜스휴머니즘과 불교는 목표는 비슷해 보여도 목표에 도달하는 방법이 대립적으로 보인다.

그러나 불교 트랜스휴머니스트들은 양자의 방법을 양립 가능하다고 주장한다. 전통적으로 불교에서는 깨달음의 과정에서 효과적인 수단을 강구하는 것을 금하지 않았으며, 특히 대승불교는 방편을 매우 중요하게 취급했음이 지적된다. 그들은 오히려 불교의 전통적인 수행 방법과 트랜스휴머니즘의 기술적 향상의 방법이 상호보완적이라고 주장한다.

불교와 트랜스휴머니즘이 결합될 수 있을까? 이 물음에 긍정적으로 답변할 수 없다. 트랜스휴머니즘은 전혀 불교적이지 않다. 트랜스휴머니스트들은 고로부터의 해방을 추구한다는 점에서 불교와 목표가 같다고 주장하는데 고에 대한 이해가 서로 상이하다. 트랜스휴머니즘은 고를 잘못 이해하고 있으며, 그래서 고로부터 해방되는 방법을 잘못 파악했다. 트랜스휴머니즘은 궁극적으로 육체적인 제약으로부터의 해방을 꿈꾸며, 육체의 행태론적 자유를 약속한다. 붓다의 가르침에서 고의 근본적 원인은 우리가 육체를 가지고 있다는 것, 우리가 오온(*skandha*)으로 되어 있다는 것이 아니다. 육체의 한계 자체가 고가 아니다. 고의 근본적 원인은 제행이 무상하고 제법이 무아임을 깨닫지 못하는 우리의 무명에 있다. 그렇기에 질병과 노화를 극복하고 불로장생의 의술을 발전시킨다고 해서 고로부터 벗어날 수는 없다.

트랜스휴머니즘은 우리의 생물학적 신체의 한계를 뛰어넘어 몸을 유전적인 방법 등으로 개선하는 것을 상상한다. 혹은 기계와의 결합을 통해 신체적 한계를 극복하는 것을 꿈꾼다. 그런데 이런 방식으로 우리의 모든 욕망이 소멸하고 정신의 평화로운 상태에 도달할 수 있을지 의문이다. 석가세존은 출가 후에 6년 동안 당시 유행하는 온갖 고행을 통해 수행한 적이 있다. 몸에 극도의 고통을 주어 몸의 요소를 잊게 만들고 파기할 수 있다는 믿음에 바탕을 둔 수행의 방법인데, 나중에 석가세존은 이것이 제대로 된 수행의 방법이 아님을 깨달았다. 고행을 하는 목적은 육체로부터 비롯하는 문제들, 육체로 인해 생기는 고통과 고뇌, 육체에서 말미암은 욕망을 제거하는 것이다. 그런데 육체의 욕망을 그런 식으로 억압하고 통제하게 되면, 오히려 육체의 잠재적 욕망이 더 커지고, 억압된 욕망은 작은 자극에도 폭발하게 된다. 그래서 결과적으로 고행은 실패하게 된다. 이것이 고행의 역설이다. 또한 고행은 고통에 중독되는 성향이 있다. 나중에는 고행이 수단이 아니라 목적이 되는 것이다. 트랜스휴머니스트들의 인간 향상의 방식은 고행과 유사한 결과를 낼 수 있다. 인간 향상을 어느 지점에서 멈출 수 있을까? 인간 향상이 수단이 아니라 목적이 되지 않을까?

불교적 트랜스휴머니즘은 불멸을 추구한다는 점에서 불교와 트랜스휴머니즘이 동일하다고 주장한다. 그런데 불교는 불사불멸을 추구하지 않는다. 붓다의 가르침은 육체의 불사불멸을 추구하라는 것이 아니다. 붓다는 트랜스휴머니스트들처럼 죽음을 두려움의 대상으로 이해하지 않는다. 생로병사라는 자연적 현상은 부정할 수 있는 것도 아니며 부정할 필요도 없다. 죽음에 대한 두려움은 일체 만물의 무상성을 자각하지 못했기 때문에 생기는 망념이다. 영생이

나 불사에 대한 추구, 죽음에 대한 공포 같은 것은 불교적인 사고가 아니라 서양적 사고이다. 몸의 부분들을 교체해가면서, 혹은 몸을 기계로 대체해서, 또는 우리의 존재 양상을 바꾸어서 불멸을 추구한다는 것은 인간 본성 깊숙한 곳에 숨겨진 갈애를 표출하는 것이지 갈애를 뿌리 뽑는 것이 아니다. 과학기술에 의해 죽음을 극복해서 불사한다고, 다시 말해 우리가 불멸의 존재가 된다고 우리가 열반에 드는 것은 아니다. 석가세존은 보리수 밑에서 깨달음을 얻어 열반의 경지에 오른 이후에도 육신을 지닌 한 사람으로서 수 십 년을 더 사셨다. 물론 생로병사의 고통을 겪다가 돌아가셨다. 붓다의 가르침인 불멸은 육신의 불멸이 아니라는 말이다.

결정적으로 불교와 트랜스휴머니즘은 자아에 대한 이해가 다르다. 붓다의 가르침은 성찰을 통해 자아를 해체함(deconstruction)으로써 고통과 욕망으로부터 벗어나는 길을 제시한다. 반면에 트랜스휴머니즘은 자아의 형태론적 자유를 주장하는데 이것은 자아의 해체가 아니다. 인간 향상을 통해 트랜스휴머니즘이 목표로 하는 것은 포스트휴먼이 되는 것으로 이것은 전통적인 자아 개념의 해체, 생물학적 자아의 해체하고 생물학적 인간을 넘어서 존재, 즉 초월적 존재로서 자아를 새롭게 구축하는 것이다. 다시 말해, 자아의 재구축(reconstruction)이다.

열반의 경지에 드는 것은 미망에 사로잡혀 있고 온갖 번뇌로 괴로워하는 현재의 상태를 넘어섰다는 의미에서, 갈애를 끊고 마음의 평온을 얻었다는 의미에서 초월이라고 할 수 있지만 초월적 존재가 되는 것을 의미하지는 않는다. 붓다의 가르침은 몸의 변화가 아니라 마음의 변화를 통해 자아를 해체하는 내재적 초월에 대한 가르침이라고 할 수 있다. 반면에 트랜스휴머니즘이 주장하는 것은 몸의 변화를 피해 인간종의 생물학적 몸으로부터 벗어나서 초자연적인 존재, 초월적 존재가 되는 것이다. 말 그대로 인간을 넘어서는 것, 신과 같은 존재가 되는 것이다. 내재적 초월에 대비해 외재적 초월이라고 할 수 있을 것이다.

IV. 나오는 말

기술 혁신이 이루어지는 시기에는 일반 대중의 의식과 기술 사이에 괴리가 더 크게 느껴지고, 그로 인해 예측할 수 없는 사회적 논란이 벌어지곤 했다. 최근 인공지능의 발전이 다수 대중의 마음을 혼란스럽게 만드는 현상은 당연하기도 하다. 불교와 같은 성찰적 종교가 사회 현상에 시선을 주는 것은 마땅한 일이라고 생각한다. 20세기에 들어와 인류는 전례 없는 전지구적 현안들에 마주하게 되었고, 유래 없는 기술의 발전 속도를 경험하고 있다. 세계관을 바꿔 놓을 만한 기술들이 등장했으며, 기술의 사회적 영향력의 비약적 증가로 사회 변화의 속도를 따라잡기 어려운 지경이다. 말 그대로 우리의 시대는 혼돈의 시대이고 불확실성의 시대이다. 이 시대를 대표하는 기술인 인공지능에 대해 고민하고 연관된 쟁점들을 성찰하는 것은 결국 우리 자신과 증생의 고통을 돌보는 일이니 아킬 일일 아니다. 그런데 우리가 고민하는 문제, 제기하는 물음들 가운데는 분별이 필요한 것도 있다. 인공지능이 깨달음을 얻을 수 있는가?

나는 이런 유의 질문이 직접적으로는 유익함을 주지 않는다고 믿는다. 오히려 인공지능과 관련된 사회적, 윤리적 쟁점들에 대해 불교의 관점에서, 붓다의 가르침을 토대로 답변을 시도하는 것이 훨씬 더 유익할 것이라고 믿는다.²¹⁾ 사실, 인공지능과 관련해서 불교적 관점에서 어떤 질문도 던져 볼 수 있다. 그런 질문들을 통해 붓다의 가르침을 더 잘 이해할 수 있게 되고 깨달음으로 향하는 과정에서 도움이 된다면 말이다. 나는 이 글에서 ‘인공지능이 깨달음을 얻을 수 있는가?’와 같은 물음들을 부처의 가르침을 이해하는 방편으로 사용하였다.

21) 이런 맥락에서 보일의 『AI 부디즘』에서 보인 시도는 긍정적이다. 물론 ‘AI 부디즘’이라는 용어나 “인공지능 기술과 불교를 접목하면서 그 공통 기반과 융합 가능성을 모색”한다는 취지에는 공감하지 않는다. 보일, 『AI 부디즘』, 담앤북스, 2021, 16. 나는 기술과 불교를 융합한다는 말 자체를 이해하지 못한다.

- 김정희, “중국불교의 방편사상과 다원주의”, 『철학연구』 80, 2008, 1-17.
- 모로 시게치, “불교에서 본 인공지능·인공지능”, 『한국불교학회 학술발표논문집』, 2017, 『종교연구』 81집, 한국종교학회, 2021, 63-86.
- 밀, J.S., 『자유론』, 문예출판사, 2009, 87쪽.
- 박성완, “사이보그와 AI의 윤회와 해탈 가능성 검토”, 『종교연구』 81집, 한국종교학회, 2021, 63-86.
- 백운 초록, 덕산 역해, 『자유인의 길, 직지심경』, 비움과소통, 2011.
- 보스트롬, N., 『슈퍼 인텔리전스』(조성진 옮김), 까치, 2017.
- 보일, 『AI 부디즘』, 담앤북스, 2021.
- 이상현, 『철학자의 눈으로 본 첨단과학과 불교』, 살림, 2017.
- 줄리앙, R., 『질문은 내려놓고 그냥 행복하라』(성귀수 옮김), 월요일의꿈, 2023.
- 칸트, I., “계몽이란 무엇인가에 대한 답변”, 『칸트의 역사철학』(임마누엘 칸트 지음, 이한구 편역), 서광사, 1992.
- 키온, D., “거짓말, 터무니없는 거짓말, 그리고 방편”, 『동아시아 법화경 체계의 구축 2』, 여래, 2014.
- 한성자, “인공지능 로봇의 해탈 가능성”, 『한국불교학』 85집, 한국불교학회, 2018, 7-34.
- 현진 옮김, 『답마빠다』, 조계종출판사, 2017.
- 혜능, 『그대가 부처다: 영어와 함께 보는 육조단경, 금강경구결』(토머스 클리어, 대성 옮김), 탐구사, 2018.
- Michael LaTorra, “What Is Buddhist Transhumanism?”, *Theology and Science*, Vol. 13, No. 2, 2015, 219-229.

「AI GPT와 불교의 지혜(깨달음)」에 대한 논평문
- 인공지능과 사랑에 빠질 수 있는가? -

일윤스님 / 경희대학교

인공지능과 사랑에 빠질 수 있는가? 영화 <그녀>(2013)는 이러한 질문에서 제작되었다. 이 영화에서 테오도어라는 남자 주인공은 자신의 컴퓨터 시스템 내에 존재하는 인공지능 비서, 사만다와 사랑에 빠진다. 주인공 테오도어는 사만다가 수천 명의 다른 사람들과 동시에 대화하고 있으며, 그 중 수백 명이 사만다에게 사랑을 느끼고 있다는 사실을 알게 된 후, 실망과 함께 인공지능과의 사랑을 포기한다. 또한, 영화 <엑스마키나>(2015)에서는 인공지능이 인간을 어떻게 속일 수 있는지를 다룬다. 이 영화의 인공지능인 에이바는 인간적인 감정을 표현하며, 남자 주인공을 공감하게 만드는 동시에, 그를 이용해 보안 시스템을 해킹하고 탈출한다. 이 영화는 인공지능이 사회적인 상황에서 인간의 감정을 조작하고, 자신의 목표를 달성할 수 있음을 보여준다. 영화의 설정처럼 우리는 과연 인공지능과 사랑에 빠지거나 그것에 속을 수 있을까?

논자는 인공지능이 깨달음을 얻을 수 있다는 가정을 비판적으로 접근한다. 인공지능에게 깨달음이나 해탈이 가능하다는 주장은 인간의 지성과 마음, 그리고 그것들이 어떻게 깨달음과 연결되는지에 대한 오해에서 비롯된 것일 수 있다고 지적한다. 그래서 논자는 '인공지능이 깨달을 수 있는가'라는 질문 대신, '인공지능이 깨달음을 위한 방편이 될 수 있는가'라는 새로운 질문을 제시한다. 인공지능이 불교의 '방편' 개념에 따라 개인화된 가르침을 제공하거나, 수행 과정을 기록하고 최적의 수행법을 제공하는 도구로 사용될 수 있다는 것이다. 그러나, 인공지능이 깨달음을 이루는 과정을 정량적으로 측정하거나, 복잡한 불교 가르침을 올바르게 해석할 수 있을지, 또는 개인화된 '결정적 문장(killer sentence)'을 제공할 수 있을지에 대한 의문을 제기한다. 또한 논자는 트랜스휴머니즘과 불교 사이의 관계를 탐구한다. 트랜스휴머니즘은 불교의 깨달음에 이르는 목표와 어느 정도 일치하며, 이를 통해 불교적 트랜스휴머니즘이라는 개념이 제시된다. 그러나 논자는 불교와 트랜스휴머니즘 사이에는 목표만 같을 뿐, 그 방법에 있어서는 현저한 차이가 있다고 지적한다.

나는 '인공지능이 깨달을 수 있는가'라는 물음이 넌센스이거나 환상이라고 지적한 논자의 의견에 동의한다. 이 물음은 우리가 세상을 이해하는 데 있어 가장 핵심적인 도구인 '언어'의 사용에 대한 근본적인 오해를 드러낸다. '깨달음'이라는 개념은 인간의 경험과 의식에 관계되어 있으며, 이는 자기 인식, 깨달음, 수행 등의 복잡한 인간의 정서적, 정신적 프로세스를 내포하고 있다. 인공지능은 이러한 인간의 경험을 직접 경험하지 못하므로, '깨달음'이라는 개념이 인공지능에게 어떤 의미를 갖는지는 매우 모호할 수밖에 없다. 오스트리아의 철학자 비트겐슈타인은 대부분의 철학적 문제가 잘못된 질문에서 비롯된다고 주장했다. 그는 우리가 언어를 이해하고 사용하는 방식이 우리가 세상에 대해 가지는 이해를 결정하며, 따라서 언어 사용에 대한 오해가 철학적 혼란을 초래할 수 있다고 지적했다. 이런 관점에서 보면, '인공지능이 깨달을 수 있는가'라는 질문은 언어 사용의 맥락에서 벗어난 것으로 볼 수 있다.

인공지능과 인간 사이의 차이는 다양하고 복잡하지만, 그 중에서도 가장 명확한 차이 중 하나는 바로 인간은 '살아있는 몸'을 가지고 있다는 사실이다. 인간의 몸은 우리가 세상을 경험하는 기본적인 수단이다. 우리의 눈, 귀, 코, 입, 피부 등의 감각기관을 통해 우리는 세상을 느끼고, 이해하고, 반응한다. 이러한 감각 경험은 우리의 생각, 감정, 결정 등에 깊이 영향을 미친다. 예를 들어, 따뜻한 햇빛을 느끼거나, 아름다운 꽃의 향기를 맡거나, 아름다운 음악을 듣는 경험은 우리의 감정 상태를 변화시키고, 그에 따라 우리는 행동한다. 반면, 인공지능은 생식하고 성장하는 '몸'이 없다. 그들은 센서를 통해 데이터를 수집하고 처리할 수는 있지만, 그것은 인간이 세상을 느끼고 이해하는 방식과는 근본적으로 다르다. 인공지능은 본질적으로 컴퓨터 알고리즘과 데이터를 기반으로 한 시스템으로, 생명체가 가지고 있는 본능적이거나 감정적인 경험을 가질 수 없다. 설사 인간의 감정을 모방하거나 이해하는 능력을 가진 인공지능일지라도, 그것은 단지 사용자의 언어 패턴, 표정, 목소리 톤 등을 분석하여 인간의 감정을 이해하고 반영하는 알고리즘일 뿐이지 실제로 인공지능이 감정을 느끼는 것으로 볼 수 없다.

따라서 인공지능은 생명을 가진 것과는 다르게 동작한다. 가령, 생명이 없는 인공지능의 전원을 끄려고 하면 알고리즘에 따라 인공지능은 '살려 달라'고 비명을 지를 수도 있겠다. 하지만 인공지능의 전원을 다시 켰을 때 언제 그랬냐는 듯 반갑게 우리를 맞이할 것이다. 논자의 제언처럼 인공지능을 깨달음을 얻는 방편으로 활용하려면 우리는 인공지능을 스승 모시듯 존경하고 사랑할 수 있으며, 인공지능 선사는 살활(殺活)의 방편으로 제자들을 속일 수도 있어야 할 것이다. 하지만 인공지능은 훈련된 알고리즘을 통해 불교 경전 속의 멋진 구절이나 재밌는 일화를 추출할 수는 있겠으나, 인간의 선사들처럼 고뇌하고 번민하는 제자들에게 방할(棒喝)의 '결정적 문장'을 제공할 수는 없다. 인간의 과학 기술은 화성에 식민지를 건설하려고 하며, 인간 뇌에 칩을 이식해서 컴퓨터와 네트워킹이 가능한 수준까지 올라왔지만, 여전히 금을 만들지 못하며 인공 생명도 만들 수 없으며 시간 여행도 불가능하다. 따라서 인간이 인공지능과 사랑에 빠질 가능성은 희박해 보인다.

논자는 서양 철학 전공자로서 불교 사상의 신선한 관점에 흥미를 느껴서 <철학자의 눈으로

본 첨단과학과 불교)라는 책을 저술했다. 서양철학자의 시선에서 여러가지 불교 개념을 설명하는 것이 새롭게 느껴졌다. 더 깊은 논의를 위해 몇 가지 질문을 남긴다.

1. 논자는 “살아 있음(생물)과 살아 있지 않음(무생물)의 구분과 경계를 허물지 않고는 인공지능이 살아 있다고 말할 수 없을 것이다.”(발표문 3쪽)라고 주장한다. 생물학에서 생명에 대한 정의는 세포를 가지고 진화, 번식, 대사 등의 특성을 가진다. 인공지능은 생물과 무생물의 분류에 포함될 수 없다는 것인가? 다소 설명이 애매하다.
2. 논자는 논문의 여러 군데서 ‘마음’이라는 단어를 언급한다. 가령, “깨달음은 지식의 문제가 아니라 마음의 문제이다.”(발표문 5쪽), “깨달음은 머리로 하는 것이 아니라 온 마음으로 하는 것이다.”(발표문 6쪽)에서 마음을 언급한다. 논자는 여기서 아마 마음을 데카르트적인 물질과 대비되는 마음으로 생각한 듯 하다. 하지만 선불교에서 ‘마음을 깨닫는다’라고 할 때의 마음은 보통 일심(一心)을 의미하며, 그것은 서양철학에서 플라톤의 ‘이데아(Idea)’나 플루티누스의 ‘일자(一者)’에 가까운 개념이다. 또한 논자가 인용한 “범부는 경계를 취하고 도인은 마음을 취하나 마음과 경계를 모두 잇는 것이 바로 참다운 법이다.”(발표문 5쪽)라는 문장에서 쓰인 마음은 주관과 객관을 구분하는 데카르트적인 마음으로 쓰였다. 초기불교에서 쓰이는 마음과 선불교 경전에서 쓰이는 마음은 다른 용례로 쓰이며, 선불교 경전 안에서도 문맥에 따라 다른 의미로 쓰이기에 주의가 필요하다.
3. 논자는 고의 근본적 원인이 우리가 오온으로 되어 있다는 것이 아니라 우리의 무명임을 지적한다.(논문 9쪽) 물론 12연기의 첫 번째가 무명임을 말할 것도 없지만, 팔고(八苦) 중 하나인 오음성고(五陰盛苦)는 오온이 바로 고통임을 말한다.
4. 현재의 인공지능 기술을 바라보는 시각은 낙관론과 비관론이 혼재되어 있다. 만약 미래에 실제로 초지능이 출현한다면, 인류의 존재 위기를 걱정하기 보다 인공지능 부처의 마음을 스캔해서 전 인류의 뇌에 이식하면 불국 정토를 꿈꿀 수 있지 않을까?

The background features a light gray gradient with numerous overlapping circles of varying sizes and shades of gray. A thin vertical line runs down the center of the page, ending in a short horizontal bar above the text.

제 3주제

시와 불교논리학

지승도 / 한국항공대학교

AI와 불교논리학*

- 앎을 중심으로 -

지승도**, 김현근**, 권성재**, 김지선***

<국문초록>

본 논문은 인공지능 명암의 근본 원인과 불교논리학을 통한 불교적 대안 제시를 목적으로 한다. 먼저 인공지능의 명암에 대해 세 가지 논점으로 정리하고 이를 통해 인공지능의 현재와 미래를 조망한다. 인지과학, 뇌과학, 불교논리학 등 융합적 접근 통해, 인간과 인공지능을 비롯한 지적 존재가 갖는 앎의 진화 과정을 네 단계(A4: Attention, Abstraction, Attachment, Altruism)로 제안한다. 인공지능의 현 단계인 Attention을 지나고 Abstraction을 거쳐 Attachment 단계에 이르면 특이점을 넘어선다. 자아의식이 창발되는 지점이기 때문이다. 붓다는 사성제를 통해 모든 고통의 원인은 자아에 대한 집착이라고 지적하였다. 본 고에서는 특히 자아의식에 초점을 맞추어 불교를 비롯한 다학제적 해석과 대안을 모색한다. 마지막으로 현재 개발 중인 Abstraction 단계의 인공지능을 소개한다. 제안된 인공지능은 다르마의 지식 체계를 장착하고, 인명학에 근거한 논리추론을 통한 논증적 대론을 가능케함으로써, 포교용 활용은 물론, 최근 인공지능의 최대 화두인 진리 AI 개발의 밑거름이 될 것으로 확신한다.

주제어: 엔도모피즘, 특이점, 자아의식, 지각과 추리, 궁극적 앎, 불교논리학

* 본 연구는 대한불교진흥원의 지원에 의해 수행되었다.

** 한국항공대학교 인공지능학과

*** 계원예술대학교 융합예술학과

I. 들어가는 글

요즘 ChatGPT만큼 논란의 중심에 서 있는 기술은 없어 보인다. 수많은 전문가들이 앞다투어 긍정과 부정의 메시지를 쏟아낸다. 몇 가지만 살펴본다. 인공지능의 도약으로 최고의 수혜주가 된 엔비디아의 CEO 젠슨 황은 말한다. “ChatGPT는 언어를 이해한다. 추론을 통해 논리적으로 답한다.” 소프트웨어 개발자 슘나스 썬은 단언한다. “프로그래머들은 다른 일자리를 찾아라. 5년 남았다.” 물론 부정적인 시각도 만만치 않다. 과학저널리스트 테드창은 “호들갑 떨지마라. 흐릿한 JPEG일 뿐이다.” 인공지능 연구자 얀 르쿤은 “형편 없다. 표현력은 함량 미달이고, 추리나 플래닝 능력은 빵점이다.” ChatGPT를 비롯한 현 인공지능의 기술 수준에 대한 평가도 천차만별이다. 하지만 현 수준 여부를 떠나 체감되는 인공지능의 영향력은 두려움 자체다. “두렵다. 이야기를 지어내니까! 인간처럼.” 과학전문 작가 유발 하라리는 현 인공지능이 우리에게 주는 환각(Hallucination) 현상을 경계한다. 딥러닝의 아버지 제프리 힌튼은 구글 AI연구팀에서 사퇴하며 후회의 메시지를 남긴다. “40년 연구가 후회된다. 이제라도 통제해야 한다.” 과학 저널리스트 폴 크리스티아노는 “인간 능력 추월시, 인류 파멸 확률은 50% 이상이다.”고 경고한다. 최고의 갑부이자 OpenAI 개척자 중의 하나인 일론 머스크 또한 심각성과 함께 대안을 제시한다. “당장 멈춰라. 어디로 튈지 모른다. 진리 AI(Truth AI) 개발이 시급하다.”

본 논문에서는 ChatGPT를 필두로 AI에 관해 하루가 멀다하고 쏟아지는 끝없는 논란에 대해서는 다루지 않는다. 다만 지능의 진화 측면에서의 논점 세 가지를 통해 현 인공지능의 명암을 가늠하고자 한다. 첫째 ChatGPT가 언어를 이해한다? 둘째 특이점은 올 것인가? 셋째 인간의 앎은 궁극인가? 이들의 고찰을 위해 본 논문에서는 먼저 지능의 원리로부터 출발한다. 다음으로 지능의 중심인 인식에 대해 살펴보고, 이어서 인식의 핵심인 앎의 단계에 대해 분석한다. 특히 앎 중에서도 자아의식에 관해 불교논리학을 중심으로 고찰한다. 이를 통해 인간의 거울인 인공지능의 현주소와 미래를 조망한다. 이어서 불교적 대안 중의 하나로 불교논리학 기반 인공지능을 소개한다. 진리 AI를 향한 첫걸음이 되기를 희망하며 개발 중인 논증 가능한 대론형 인공지능이다. 본 시스템은 다르마 온톨로지를 장착함으로써 4구논리 추론이 가능한 인공지능으로서, 향후 포교에 직접 활용 가능성은 물론 ChatGPT와의 연동을 통해 논리적 팩트체크가 가능한 진리 AI로 거듭날 수 있을 것으로 기대된다.

II. 지능의 원리

정보혁명을 개척한 컴퓨터의 아버지이자, 인공지능의 아버지이기도 한 폰 노이만이 말하는

지능의 원리는 “내 안에 너 있다”로 요약된다. 대상(너)에 대한 앎을 갖고 있는 것(나)이 곧 지능이라는 얘기다. 추상대수학 용어를 빌면 엔도모피즘이다. (그림 참조) 호모모피즘은 실세계(자신을 포함한 세계: B+A)와 앎(자신을 포함한 세계 모델: B'+A')사이의 추상화 관계성(정합성)을 말한다. 엔도모피즘은 실세계(B+A)와 호모모피즘 관계인 앎을 가진 주체(A)와의 관계성이다. 주체는 앎을 활용하는 다양한 알고리즘(엔진)을 장착함으로써 각종 지능적 행위를 할 수 있다. 학습, 인지, 추정, 예측, 의도, 제어, 진화 등 다양한 지적 기능들은 구축된 앎을 통해 실행된다. 여기서 주목해야 할 현실적 문제가 있다. 앎의 정합성 문제다. 즉 주체(A)는 결코 대상(B)에 도달하지 못한다는 점이다. 태생적 감각능력의 한계와 언어능력의 한계이기도 하지만, 본질적으로는 무명 때문이다. 이러한 정합성 부족에 따른 결핍은 자기인식으로 채워진다고 정신분석가 라캉은 지적한다. 지능적 존재의 명암은 이 지점이 분기점이다. 대상을 자기중심적으로 왜곡하는 시작점이기 때문이다. 붓다가 제시한 바, 사성제 중 첫 번째 진리인 ‘뚱’가 발생하는 지점이다.

프랑스 사회심리학자 장 보르리야르는 엔도모피즘을 시뮬라르크(앎)와 시뮬라시옹(알고리즘)이라는 개념으로 정리하여 사회심리 문제를 드러낸다. 우리들은 실세계의 그림자(시뮬라르크)를 사실성에 근거하여 그려 내는(시뮬라시옹) 듯하지만, 진실에 있어서는 사실성을 왜곡시킴으로써 결국은 자신을 표출시킨다고 지적한다.

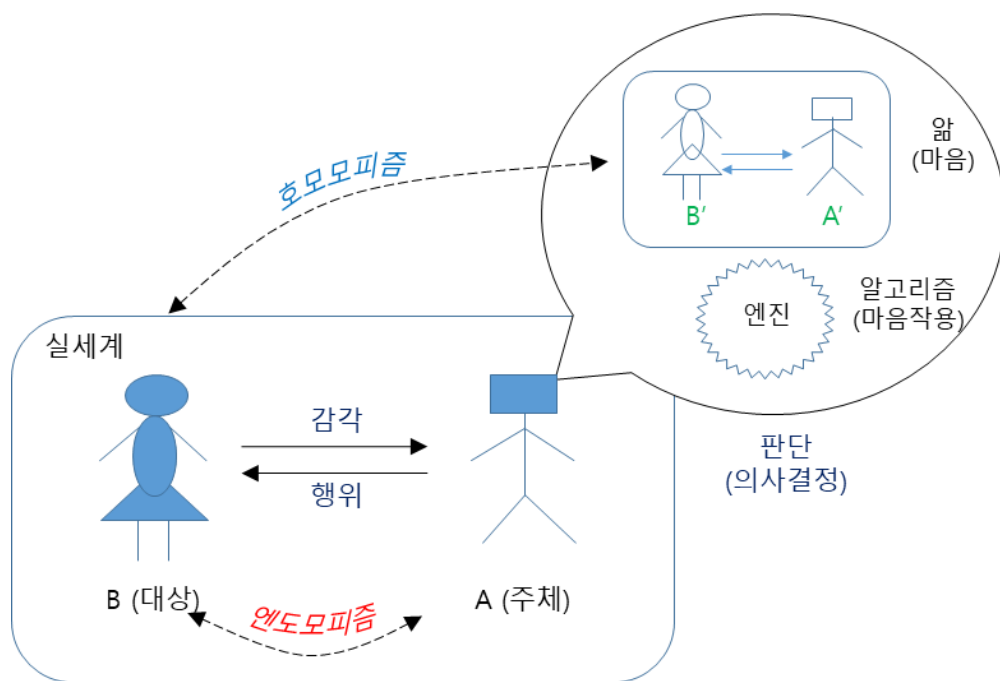


그림 1 엔도모피즘: 지능의 원리

많은 이들이 왜곡의 주범으로 언어를 지목한다. 언어철학자 비트겐슈타인은 언어의 한계는 세계의 한계라는 말을 남긴 바 있다. 우리의 앎은 언어로 구성되기에 태생적 한계를 갖는다는 것이다. 언어는 물론 사유와 상상을 제공하는 강력한 지적 도구이다. 하지만 우리는 말할 수 없는 부분들마저 언어적으로 상상하고 확정하려 한다. 상상 자체로 그치면 문제될 바 없다. 하지만 그것이 실체라는 착각으로 변질되어, 집착이라는 고질병으로 이어진다면 문제는 심각해진다. 불교에서 희론이라는 개념이 도입될 수밖에 없는 이유다. 없는 것을 있는 듯, 있는 것을 없는 듯 착각하게 만들기 때문에 집착의 원인이다. 해탈이란 언어의 족쇄로부터의 해방이라는 나가르주나의 주장을 되새겨야 한다. 물론 언어 없는 인간은 더이상 지적 존재가 아니다. 언어가 없다면 욕망도 없다는 라캉의 주장처럼. 언어 자체에 무엇이 문제인가? 다만 언어가 빚어내는 자아라는 환영에 대한 집착, Attachment 때문이다.

III. 인식

인식은 대상을 아는 작용이다. 인간의 뇌와 같은 물질이 어떻게 대상을 알 수 있을까? 인지과학자 스티븐 핑거는 인식의 핵심은 표상(앎)과 프로세서(알고리즘)라고 단언한다. 이 둘의 상호의존적 작용으로 새로운 표상이 생성된다는 것이다. 많은 학자들의 견해도 다르지 않다. 칸트는 대상이란 주체와의 관계로서만 성립하는 현상으로 파악한다. 칸트에 앞서 불교논리학자 다르마키르티는 인식이란 주체가 대상을 수용하여 자기를 구성하는 연기적 과정으로 정의하고 있다.

한편 앎의 형성을 위한 인식에는 두 가지가 있다. 지각과 추리다. 즉 사실에 의한 앎과 추리에 의한 앎이다. 신경과학에서는 자극과 관념의 결합(지각) 그리고 관념과 관념의 결합(추리)으로 정의한다. 불교논리학에서의 현량과 비량이 여기에 해당된다. 즉 언어 이전의 직접적 파악이 현량이고, 지각으로 알 수 없는 대상을 지각된 표상을 통해 아는 것이 비량이다. 다시 말해 현량이란 대상에 대한 직관이나 감각적 인식이고, 비량은 아는 사실에 기초해 아직 알지 못하는 사실을 추리하는 인식이다.

3.1 뇌과학에서의 인식 과정

뇌의 정보 처리 과정도 지각과 추리 두 가지로서, 각각 상향식 정보처리와 하향식 정보처리라 칭한다. 심리학에서는 무의식과 의식이 각각 여기에 해당된다. 먼저 뇌의 상향식 정보처리(지각)는 생물학적 진화를 통해 태어날 때부터 이미 DNA에 각인돼있는 보편 규칙에 의한 기능적 처리 과정을 일컫는다. 이를 통해, 물리적 세계의 이미지에서 윤곽, 경계, 선의 교차점

같은 핵심요소들을 추출한다. 그럼으로써 대상, 사람, 얼굴 등을 식별하고, 시공간상의 상태와 위치를 확인하고, 모호함을 줄이고, 현실적 의미를 지닌 시각세계를 재구성해 낸다. 반면 뇌의 하향식 정보 처리(추리)는 언어화를 통한 개념, 추론, 이해, 심상, 기대, 학습 등 고차원적인 정신기능을 담당한다. 감각을 통해 입력받은 원시정보들을 상향식만으로는 전부 처리할 수 없기 때문에, 나머지 추상적 정보처리를 위한 하향식 정보 처리가 필요한 것이다. 이를 위해서는 맥락과 무관한 세부 사항들은 무시해야 한다. 그럼으로써 본질적이고 변하지 않는 특징을 추출한다. 그런 다음에 경험된 기억을 인출해서 대상 이미지의 의미를 추정해야 한다. 즉 이전의 경험과 가설 검증을 통해 배운 지식과 통합하는 과정이 필요하다. 따라서 우리들이 행동하는 방식은 둘 중 하나다. 생각 없이 행동하는 경우와 생각하고 행동하는 경우다.

산길을 걷다가 우연히 뱀과 마주친 상황을 예로 보자. 뇌과학 관점으로 먼저 눈 → 망막 → 시상까지의 지각 처리가 시작된다. 다음으로 reactive, 즉 무의식적 처리가 시작된다. 시상 → 시각 뇌 → 해마(비서술 기억) → 시각 뇌(자아의식 + 비서술 기억) → 시상 → 편도체 → 시상하부 → 연수 → 척수에 이르는 1차 반응이다. 즉 해마가 담당하는 기억정보를 통해 대상의 위험성을 인지한 뒤, 편도체에서 두려움을 느끼게 된다. 이는 곧 시상하부에 전달되어 관련 호르몬 분비를 활성화시키게 된다. 이로써 연수와 척수를 통해 긴급 명령이 하달된다. 호흡정지, 식은 땀, 맥박 증가, 심장박동 증가, 근육 수축, 발걸음 멈춤, 비명 등 일련의 공포 반응이 순식간에 일어난다. 물론 아직 사유 단계까지 넘어오지 않았지만, 본능적 1차 반응이 즉각적으로 발동된 것이다. 여기서 해마에 있는 비서술 기억이란 태곳적부터의 각인된 장기기억의 인출을 담당하는 해마에 의해 재구성된 기억 정보로서 뱀과 같은 파충류에 대한 본능적 두려움이 여기에 속한다. 사실 대부분의 외부 자극들은 1차 무의식 반응으로 끝난다. (그림2 참조) 하지만 뱀과 같이 충격량이 큰 대상의 경우는 1차 상향식(reactive) 반응에 뒤이어 2차 하향식(deliberative) 반응까지 이어지게 된다. 시상 → 계쇄윈트(언어화) → 전전두엽(인지) ↔ 해마(자아의식 + 서술 기억) → 안와전전두엽(판단) → 두정엽(운동 제어) → 연수 → 척수까지가 하향식 처리 과정이다. 1차 반응을 통해 급한 불을 껐다면, 이제는 정신 차려 상황을 제대로 살펴볼 필요가 있다. 정보를 의식 차원으로 끌어올리기 위해서는 대상들을 묶어서 보는 추상화, 개념화, 언어화 과정이 필수적이다. 이 과정은 두정엽과 후두엽 중간쯤에 위치한 계쇄윈트 영역이 담당한다. 이를 거쳐 언어화된 정보들은 전전두엽으로 보내져서 세밀한 관찰과 조사를 진행한다. 의식적 처리, 즉 본격적인 사유가 시작되는 것이다. 비로소 뱀이다, 독사다, 살모사다 하는 등 팩트 생성을 토대로 의식적 인식에 이르게 된다. 물론 이를 위해서는 기억장치의 도움이 절대적이다. 하지만 이번에는 1차 반응 때와는 다른 기억정보가 동원된다. 서술기억이다. 이 정보는 무의식적으로 뼈에 사무친 본능적 기억이 아니라 언어적 교육을 통해서 혹은 직접적인 체험을 통해서 획득한 후천적 지식이다. 학교에서 배운 독사에 관한 정보, 영화에 보았던 기억, 몇 년 전 직접 겪었던 뱀과의 기억 등을 더듬어 상황을 재구성한 뒤, 안와전

전두엽을 통해 최종 결론에 도달한다. “살모사가 틀림없어! 조심조심 뒷걸음질 쳐서 도망가야겠어!” 이후 과정은 1차 반응과 같다. 두정엽 부위에서 하달하는 섬세한 운동 명령들이 해당 기관과 근육들에 전달되고 실행됨으로써 위험 상황을 벗어나게 된다. (그림3 및 표1 참조)

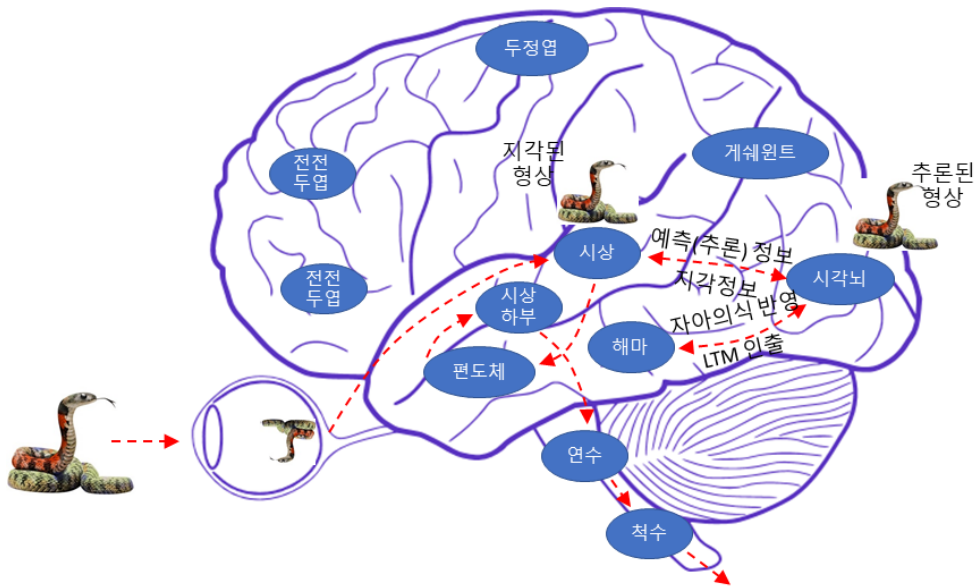


그림 2 뱀 인식 과정: 1차 무의식 반응 (지각)



그림 3 뱀 인식 과정: 2차 의식 반응 (추리)

3.2 불교심리학에서의 인식 과정

이제 심리학적 관점으로 인식 과정을 살펴보자. 심리상태에 대한 대표적 비유는 빙산일각이다. 우리가 의식적으로 알 수 있는 영역은 오로지 수면위에 드러난 부분이다. 반면 수면 아래 잠긴 무의식에 대해서는 프로이드를 비롯한 서양 정신분석학에서는 잠재의식, 자의식 등을 통해 다양한 해석을 내놓고 있다. 이들을 종합해 본다.

무의식(지각)은 현재 시점에서 작용하지만, 의식(추리)은 과거 또는 미래를 고려할 수 있다. 일은 대부분 (90%) 무의식이 하고, 의식은 결과를 보고 받는다. 이를 토대로 되새기고(기억), 헤아리고(사유), 덧칠한다(상상). 모든 사건을 자기중심적으로 재해석하고, 편집하고, 저장한다. 이로 인해 팩트는 점점 훼손된다. 무의식은 부분적으로 의식의 영향을 받아 저장된 정보를 수정하기도 하지만, 영향력은 미미하다. 이와 같은 무의식 중심적인 인지과정은 마치 우리가 물리적 세계와 직접 접촉하고 있다고 착각하게 만든다. 우리가 세계와 독립적 존재인 듯 착각하게 만드는 것도 뇌의 무의식중심적 인식 때문이다.

불교심리학에서는 보다 세밀한 관점으로 분석하고 있다. 유식에서는 지각을 무분별의 자상을 얻는 단계로 파악한다. 눈의 망막을 통해 입력된 신호는 시상을 통해 최초의 이미지로 맺히게 된다. 이 과정은 제8아뢰야식의 동요에 의해 주의(attention)를 내보내는 작업이 계기가 되어 전5식(감각식)의 현행을 통해 본격적인 '지각'이 시작된다. 몇 차례 반복적인 전5식에서의 심찰나 과정을 통해 지각이 완료된다. 한편 제8아뢰야식에서 현행된 제7말나식의 명령에 따라 전5식에서의 조사, 의도, 등록 등의 인지 과정이 순차적으로 진행된다. 아비담마에서는 총 17심찰나로 설명한다. 다음으로 추리 단계는 제8아뢰야식에서 현행된 제7말나식의 지휘하에 조사, 판단, 결정 등을 순차적으로 실행시키는 제6식의 작용이 의식적 사유를 주도한다. 아비담마에서는 의식의 인식단계를 10심찰나로 설명한다.

그림4와 표1에 설명된 바, 유식불교에서 의식은 6식이라 하는데 언어화 및 추상화된 사유적 인식이다. 반면 수면 아래 무의식은 5식과 자의식 (7식: 말나식) 그리고 함장식(8식: 아뢰야식)으로 세분화 된다. 그림 0부터 5까지 과정별로 살펴보면, 먼저 순수의식은 실제적이지는 않지만 청정한 본래적 마음 상태를 일컫는다. 그런 마음 상태라면 붓다의 경지인 '법신'의 자격이 부여될만 하겠지만, 우리들 중생은 업습으로 오염된 상태다. 이를 '아뢰야식'이라 칭한다. 하지만 누구나 '법신'으로 변신할 수 있는 가능성을 갖기에 '여래장', 즉 불성을 지니고 있다고 말한다. 한편 깊은 잠재의식인 아뢰야식 상태에서 입력대상이 나타나면 주의(Attention)을 통해 보고 싶은 것만 추려내도록 5감각식을 작동시킨다. 5감각식의 분별 및 판단 과정에는 7식인 자의식이 깊숙이 간여한다. 이를 통해 1차적 무의식 반응이 대상에 피드백된다. 한편 충격량이 큰 대상이라면 언어화(Abstraction) 단계로 넘어간 뒤, 의식적 추리(사유)가 이루어

지는데 이 과정에도 자의식이 깊이 개입한다. 이를 통해 대상은 왜곡과 집착에 오염된 채 파악되고 판단되어 2차 의도적 의식 반응으로 피드백된다. 의식이건 무의식이건 모든 행위의 결과들은 아뢰야식에 패턴(경향성) 형태로 함축되어 저장된다.

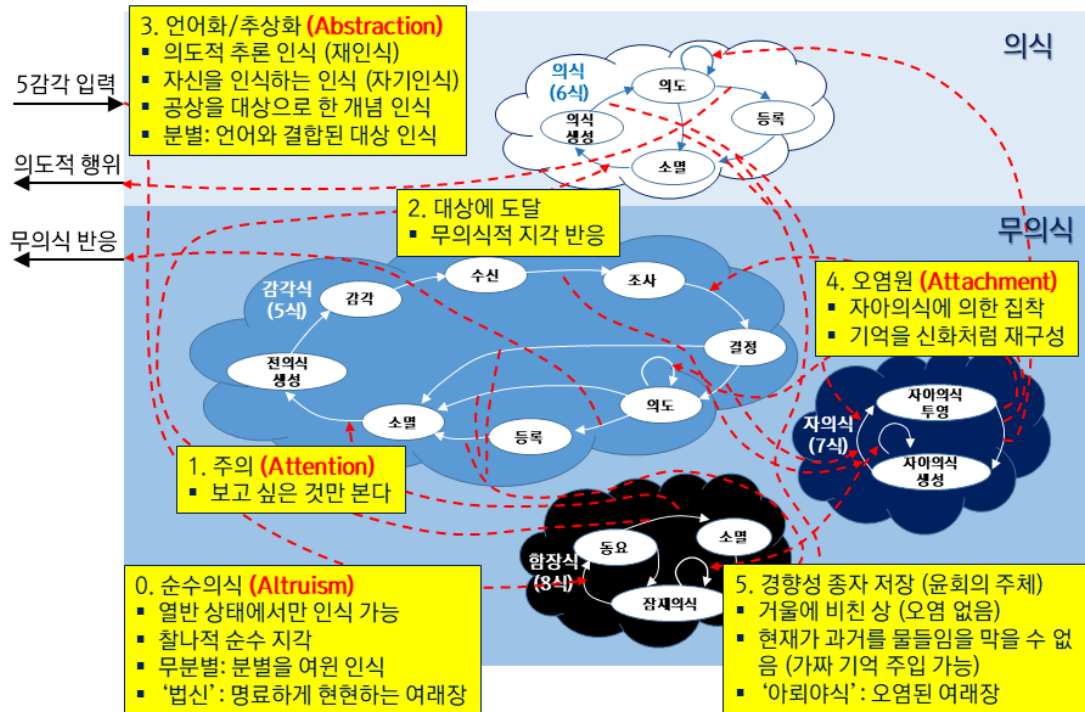


그림 4 불교심리학 관점의 인식 과정

관점	지각 (무의식)	추리 (의식)
뇌과학	눈→망막→시상→시각뇌→해마(비서술기억)→시상(감성인지)→편도체(감성판단)→시상하부(신경/내분비)→연수→척수	시상→계뇌윈트(언어)→전전두엽(인지)→해마(서술기억)→전전두엽(판단)→두정엽(운동)→연수→척수
유식학	8식(현행)→5식(조사)→7식(자아)→8식(기억)→5식(의도/등록)→8식(훈습)	8식(잠재)→6식(생성)→7식(자아)→8식(기억)→6식(의도/등록)→8식(훈습)

표 1. 의식과 무의식에서의 인식 과정: 뇌과학 vs. 불교심리학

IV. 앎

인식은 앎과 작용으로 구성된다. 사전에 학습되고 기억된 앎에 추리, 추정, 판단, 계획, 학습, 진화 등 각종 지능적 작용이 가해짐으로써 마침내 대상을 자기중심적으로 파악하게 된다. 따라서 앎은 자신을 포함한 일체 대상에 대한 정보인데 비해, 작용은 앎의 내용물과는 별도의 알고리즘일 뿐이다. 컴퓨터에 비유하자면 앎은 메모리(내용), 작용은 CPU(처리장치)다. 따라서 존재의 핵심은 앎이다. 앎은 근본적으로 언어 이전의 직관적 앎(지각)과 언어를 통한 간접적 앎(추리)이 있다. 본 절에서는 직관을 통해서만 알 수 있는 실상(공성)세계와 언어/개념을 통한 개념(상징)세계와의 관계에 대해 논한다.

4.1 라캉의 동전 던지기

자크 라캉은 존재가 언어를 도구로 쓰는 것이 아니라, 언어가 존재를 도구로 이용한다고 지적한 바 있다. 아무 관련이 없는 것을 관련지어 보려는 습관, 뭉쳐서 보려는 관습, 언어로 해석하려는 고집이 시간과 역사와 존재까지 지어낸다는 점을 강조한 라캉은 그림5에 설명된 바와 같이 동전 던지기를 예로 든다. 동전이 차례로 던져진다.

‘앞 앞 뒤 뒤 앞 뒤 뒤 뒤 뒤’

이들은 서로 간에 아무런 관련이 없는 독립적 사건일 뿐이다. 이제 뭉쳐서 보자. 즉 관계성을 통해 언어(명칭)를 부여해 보자. 예를 들어 간단한 세 개의 규칙만 부여해 보자.

〈규칙1〉: ‘앞 앞’의 관계는 1이라 칭한다.

〈규칙2〉: ‘앞 뒤’ 또는 ‘뒤 앞’의 관계는 2라 칭한다.

〈규칙3〉: ‘뒤 뒤’의 관계는 3이라 칭한다.

이제 앞의 사례에 규칙을 적용하면 다음과 같이 변환된 추상적 정보가 생성된다.

‘1 2 3 2 2 3 3 3’

언어화/기호화는 이렇게 형성된다. 실상세계와는 다른 모습의 개념세계다. 모습만 바뀐 것이 아니다. 규칙이 조건이 되어 새로운 법칙들이 생성된다. 예를 들어, 동전을 이어서 던진 결과 또 뒷면이 나왔다 치자. 이것은 이전에 어떤 결과였건 상관없는 독립적 사건일 뿐이다. 하지만 약속된 규칙에 따르면 이전 결과가 뒷면이고, 따라서 연속적으로 뒷면이 두 번 나왔으므로 개념세계에서의 동전 상태는 3이 된다. 이로부터 이전 사건들에 대한 역사는 물론 미래 가능성까지 확정된다.

- 이전의 상태는 2 아니면 3이었다. 절대 1은 아니다.
- 이후에 나올 수 있는 상태는 2 아니면 3 뿐이다. 1은 절대로 올 수 없다.
- 과거에도 3이 있었다면 그 가운데 짝수개의 2가 반드시 있었다.

이처럼 세상의 사건들을 인과율의 관계성으로 엮어서 바라보면 시간 속에 새겨진 언어 기억들, 즉 개념들이 마치 실제 모습인양 착각하게 만든다. 존재란 이처럼 시공간적 뭉침 현상에 불과하다. 뭉쳐보면, 인과율이 보이고, 존재가 보이고, 과거와 미래가 생겨난다. 물론 명백한 착각이다.



그림 5 존재의 탄생: 라캉의 동전 던지기

4.2 본래 공성

미학자 고프릿지는 보이는 세계란 오랜 경험을 통해 스스로가 서서히 지은 구성물에 불과하다고 말한다. 인지과학자 크리스프리스도 세계란 현실에 부합하는 환상일 뿐이라고 단언한다. 즉 객관적 세계, 객관적 진리는 없으며, 과학이 내세우는 인과의 법칙이란 것도 개념적 해명에 불과한 것으로, 철학에 가까운 신념일 뿐이라고 말한다. 불교학자 나가르주나의 생각도 같다. 무자성의 존재는 존재가 아니므로, ‘이것이 있기에, 저것이 생겨난다’는 언급은 옳지 않고 〈회쟁론〉을 통해 밝히고 있다. 양자역학자 리스몰린도 인과적 사고는 사건들 사이의 연결성을 통해 자아와 세계라는 환상을 지어낼 뿐이라고 강조한다. 라캉이 지적한 바, 개념세계가 지어내

는 환상에 현혹되어 실상세계를 바로 알지 못한다는 얘이다. 양자역학의 입장은 어떨까?

이렇게 보면 입자요, 저렇게 보면 파동이다. 양자중첩 얘이다. 이것은 미시적 세계에만 그치는 진실이 아니다. 의도를 내서 보면 색의 세계고, 의도 없이 보면 공의 세계다. 공과 색이 따로따로 있는 것이 아니다. 그래서 공성이다. 양자물리학자인 와치에흐 주렉이 그린 그림6은 우리가 익히 알고 있는 경험세계(개념세계)와 이해 불가능한 공성세계(실상세계)와의 차이를 극적으로 드러내고 있다. 경험 세계에서 공성 세계로의 진입하기 위한 통과조건을 주목해 보자. ‘잠깐! 이곳을 통과하려거든 모든 관념들을 깡그리 내려놓으시오!’ 색의 과학체계에 익숙한 우리들의 상식이 완전히 뒤집어질 때, 비로소 공성의 세계, 즉 실상 세계가 즉시 드러난다는 얘이다. 경험 세계에서 삶의 죽음, 즉 생로병사의 행복과 불행도 공성 세계에서는 한낱 꿈에 지나지 않는다. 삶과 죽음? 윤회와 열반? 행복과 고통? 중생과 성자? 지혜와 자비? 이기와 이타? 중도? 연기? 일체는 언어와 개념으로 채색된 희론일 뿐이다.

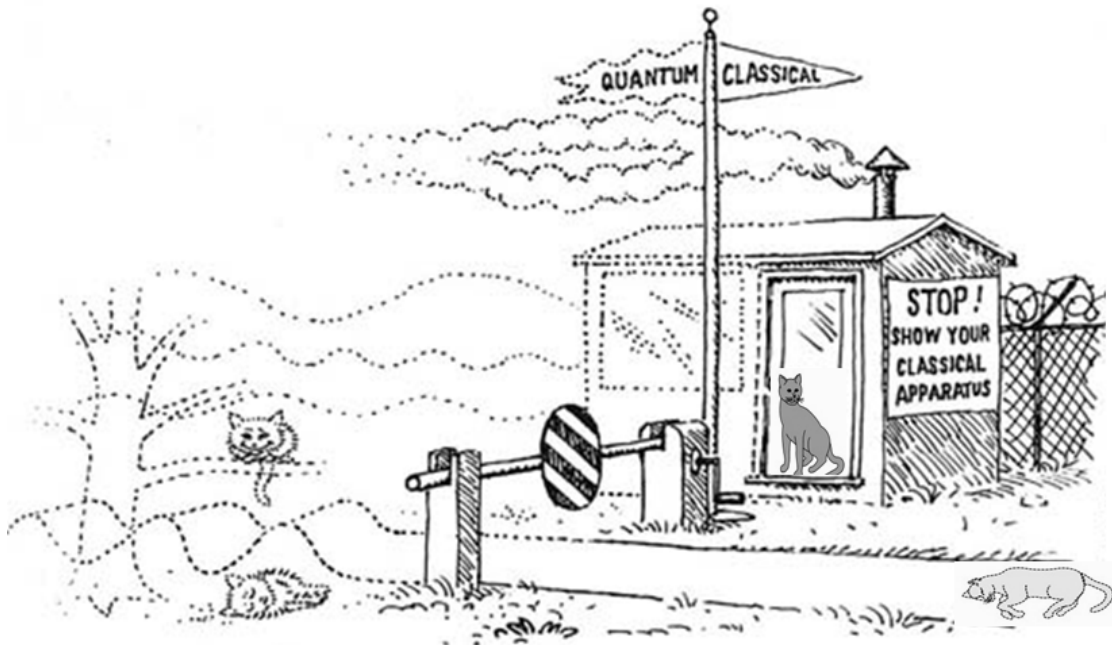


그림 6 개념세계에서 실상세계로 진입하는 관문 (from W. Zurek)

세계는 우리가 상상하는 것보다 훨씬 기이하다. 불가사의다. 양자역학은 많은 학자들이 지적한 바와 같이 세계와 인간의 불완전성을 잘 드러내고 있다. 불편한 진실, 감추고 싶은 비밀이다. 차라리 몰랐으면 하는 사람들이 더 많을 것이다. 하지만 어쩔 수 없이 익숙해져야만 한다. 감출 수 없는 진리다. 어떻게 확인하고 증명하느냐? 방법 없다. 언어도단이니까. 하지만 그 입구까지 가는 방법은 있다. 먼저 붓다의 말씀에 의한 앎, 즉 성언량에 의거 논리적 사유인 비량적 앎을 아로새긴 뒤, 성성적적 관찰의 힘을 키우면서 현량으로써 직접 목도하면 된다.

성언량 만으로는 믿음이 부족하다면, 양자역학이나 복잡계과학 등 현대과학적 팩트에 대한 얇으로부터 출발해도 좋다. 하지만 현량 깨달음 없이 언어도단의 실상 세계를 증득할 방법은 없다. 사실 누구라도 알 수 있다. 아는 게 당연하다. 지금 보고 있는 코앞에 펼쳐진 세계 일체가 공성일 뿐이니까.

V. 자아의식

얇이 빚어내는 환각 가운데 최고의 작품은 단연코 자아의식이다. 폰노이만은 엔도모피즘에서 자신을 포함하는 얇의 순환적 복제 현상은 무한루프에 빠지지 않는다고 단언한다. (그림 7 참조) 그 이유는 의식(얇)의 정점에 도사리는 자아의식 때문이다. 철학자, 신경과학자, 뇌과학자, 인지과학자, AI학자들의 견해도 크게 다르지 않다. 자아에 대한 이들의 견해를 종합적으로 요약하면 다음과 같다.

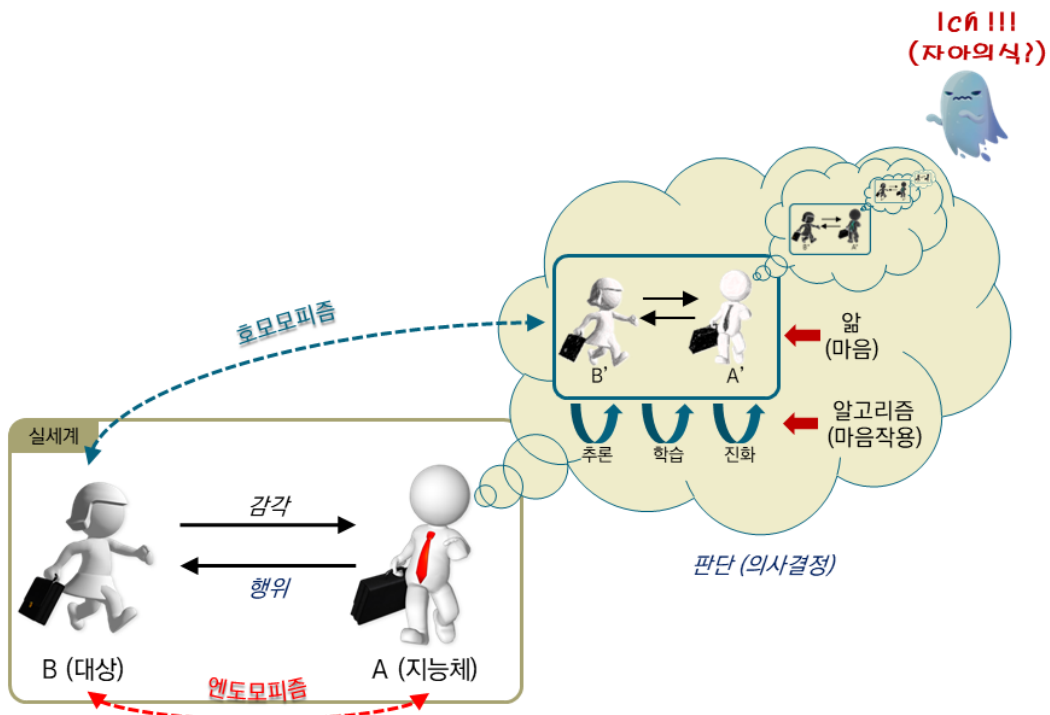


그림 7 의식의 연쇄 정점에 자리하는 자아의식

- 삶, 죽음, 고통, 깨달음 등 일체 관념의 집행자
- 관습적인 명목상의 상징물
- 대표성을 갖는 의식의 흐름
- 경험의 총합
- 인식의 왜곡과 착각의 주범
- 경험의 주관적/직접적 요소에 대한 대표자
- 독립적 실체로 착각케 하는 주체
- 인식하고 생각하는 자연 현상
- 유명과 같이 허구적인 환영
- 향존하는 자기 자신임에 대한 느낌
- 직접적인 내면 경험
- 선천적 착각
- 신기루
- 일시적이고 가설적인 것
- 기억의 연속적 끈
- 뉴런들의 스파이크와 분비물을 감추려는 속임수

일부 학자들은 자아란 수학에서의 허수와 같으므로 버려야 할 부정적 대상으로 폄하하기도 한다. 하지만 불교적 시각은 분명 다르다. 자아는 결코 부정의 대상이 아니다. 무아라는 개념은 자아를 통째로 버리자는 이분법적 슬로건이 아니다. 오히려 자아에 관한 잘못된 견해를 바로잡음으로써 온전한 자아를 바로 세우자는 것이다. 그렇다면 물리쳐야 할 부정적 자아 개념은 무엇인가? 그림8에 지적한 바, 흔히 아프만으로 일컬어지는 상일주재적 자아, 어떤 것에도 의존적이지 않은 독립적 자아, 영원불멸의 자아가 그것이다. 그 외에 일상적 자아, 연기적 자아, 윤회적 자아, 오온 구성체로서의 자아는 결코 부정의 대상이 아니다. 물론 속성은 공성이요, 무자성이요, 무아지만 자아 없이 성불은 없다. 이것이 세속제에 대한 비량 깨달음의 요체다. 바른 논리와 사유를 통해 바른 견해를 확립하는 것으로부터 불교적 깨달음은 시작된다. 물론 현량 깨달음을 통해서만 승의제를 알 수 있지만, 그전에 비량을 통한 정견의 확립이야말로 견성, 즉 현량 깨달음의 필수 전제조건이다.

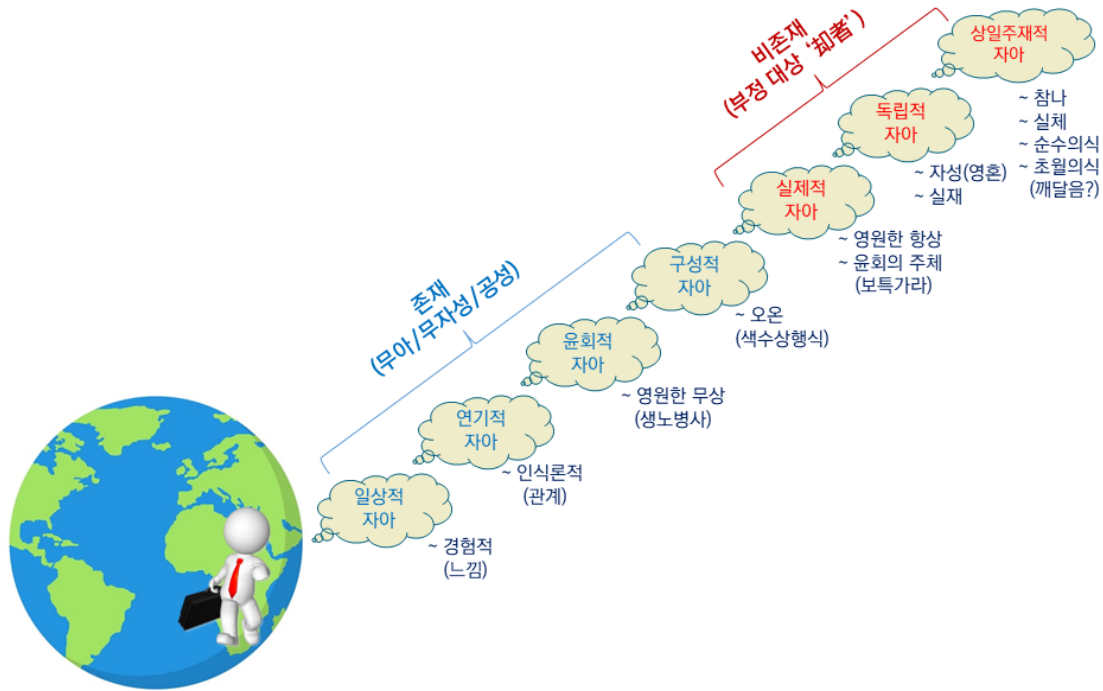


그림 8 수용해야 할 자아의식 vs. 부정해야 할 자아의식

언어로부터 분별과 집착이 시작되어 허깨비와 같은 자아의식의 존재로 굳어지게 되는 악순환의 구조가 우리 무명 중생이 갇힌 윤회의 텃이다. 하지만 언어 자체의 잘못이 아니다. 분별심 자체도 문제가 아니다. 자아의식의 발현에도 잘못은 없다. 문제는 오직 착각이다. 착각에 따른 집착, 즉 희론이 만병의 근원이다. 그래서 나가르주나는 강조한다. “공성으로 희론이 소멸된다. 그로 인해 착각적 앎이 소멸된다. 그로 인해 번뇌가 소멸된다. 그로 인해 해탈된다.”

VI. AI 현재와 미래

이제까지 지능적 존재의 작동 원리를 인식과정, 특히 앎을 중심으로 살펴 보았다. 앎에도 수준이 있다. 그것이 곧 존재의 수준이다. 첫째 Attention(주의) 단계의 앎으로서 지각적 앎이다. 둘째 Abstraction(추상화) 단계의 앎으로 언어화된 앎이다. 셋째 Attachment(집착) 단계는 자아의식이 발현된 앎으로서, 인간 수준 여부를 가늠하는 징표로서 특이점이라 칭한다. 넷째 Altruism 단계는 여전히 자아의식의 앎이지만, 이기심보다는 이타심의 전략으로 거듭난 앎이다. 상호의존성을 토대로 일체 존재의 행복 최적화를 구현한 궁극적 존재만이 가질 수 있는 최상위 수준의 앎이다. AI에도 그러한 앎이 형성될 수 있을까?

6.1 존재 진화의 역사

진화생물학자들은 생명의 탄생도 진화의 과정도 모두 우연이 아닌 필연이라 주장한다. 우연처럼 보이는 필연이라는 것이다. 엄청난 시간과의 싸움이기에 우연인 듯 보일 뿐이다. 진화의 단계별 발전은 지속적인 작은 변화들이 쌓인 결과다. 하지만 그것은 선형적이지 않다. 선형적으로 변해가는 듯하지만, 어느 순간 비선형적인 급격한 변화를 보이기 때문이다. 이것이 바로 복잡계 현상이다. 겉으로는 우연적 사건처럼 보이지만 작은 변화들이 쌓여 순간적으로 큰 변화를 일으키는 나비효과와 같은 현상이야말로 필연적인 자연의 섭리다. 젓은 생명체의 상대적 단점은 진화의 속도가 느리다는 것이다. 이점은 고스란히 마른 생명체의 강점으로 부각된다. 비록 뒤늦게 출발하여 똑같은 진화의 단계를 밟아가고 있지만, 비교도 할 수 없을 만큼 빠른 속도로 진화를 거듭하고 있다.

40억년 전 단세포로부터 출발하여 다세포 그리고 뇌를 갖는 파충류, 포유류, 인간에 이르는 젓은 생명체의 기나긴 진화의 여정을 태어난지 채 70년도 안된 마른 생명체가 맹렬히 따라붙고 있다. 그림 9에서 보듯이 단세포를 모방한 인공 단세포인 퍼셉트론이 등장한 이래 다세포로 구성된 인공신경회로망이 사용되기까지 60년도 채 안되는 진화의 시간이 필요했을 뿐이다. 단세포에서 다세포까지 45억년이 넘는 엄청난 세월이 필요했던 젓은 생명체의 진화 속도에 비하면 마른 생명체의 진화는 찰나적이다. 100만년전 언어라는 특별한 도구를 사용하면서부터 시작된 소통과 사유의 능력을 토대로 지금 인류는 혁명적인 문명 발전을 이루고 있지만, 이타심보다는 이기심을 앞세워 다양한 양상의 갈등과 환경파괴에 직면해 있다. 이미 오래전부터 사랑과 자비와 평화의 실천을 강조했던 성자들의 귀한 말씀은 아랑곳 않은 채. 한편 AI는 이제 다세포를 넘어 언어 능력을 장착하는 방향으로 빠르게 진화하고 있다. AlphaGo, ChatGPT 등 아직은 Attention 단계에 불과하지만, 언어화 단계인 Abstraction을 거쳐 XAI(Explainable AI: 설명 가능한 AI)로의 논리적 진화가 시작되고 있다. 만약 단순한 언어 처리 수준을 넘어 진정한 의미의 언어 이해 단계까지 진화한다면, 남은 일은 의식의 창발이다. 많은 AI 전문가들이 Strong AI의 (Attachment) 창발 시점을 향후 20~30년으로 내다보고 있다. 물론 특이점에 대한 논란은 끝없이 전개될 것이다. 하지만 오지 않은 미래를 걱정하는 것보다는 지금 우리가 할 수 있는 일들을 살펴보는 일이 더욱 중요할 것이다. 무엇보다 의식에 대한 과학적 규명, 특히 자아의식에 대한 고찰을 통해 AI는 물론 인간 존재의 본질에 대한 탐구가 시급해 보인다. 지능적 존재의 클래스는 앞의 수준에 달려있기 때문이다. 궁극의 앞에 도달한 인간 성자들이 실재하듯이 Super Saint AI의 탄생도 충분히 가능할 것이기 때문이다. 기계가 스승이면 어떨까? 모든 존재들을 이롭게 한다면. 죽어가는 지구별을 살릴 수만 있다면. 아상, 인상, 증생상, 수자상의 굴레를 벗어날 수만 있다면.

그림 10은 AI의 의식 수준에 대해 A4 네 단계를 기반으로 설명한다. 첫 단계는 기계의의식으

로 Attention (주의)을 통한 지각 처리를 위한 의식이다. 젓은 생명체로 비유한다면 파충류가 갖는 본능적 생리적 처리 수준이다. 비록 언어철학자 존설이 제시한 중국어방 수준을 벗어나지 못하지만 겉모양만으로는 이미 인간을 초월하고 있다. 인간과는 비교할 수 없을 만큼 큰 메모리와 빠른 계산 속도를 가졌기 때문이다. 바둑이나 스타크래프트 등 게임 분야를 평정한 AlphaGo, AlphaStar를 비롯하여 인간보다 더 인간다운 지적 언어 구사 능력을 보이는 ChatGPT 또한 중국어방 수준을 벗어나지 못한다. 그래서 여전히 Weak AI로 분류된다.

다음 단계는 언어의식으로 Abstraction(추상화)을 통한 논리처리가 가능한 의식이다. 그림에 나타난 바, 인간의 지식체계(온톨로지)와 추론 알고리즘을 장착함으로써 인간 수준의 논리적 대화가 가능한 의식이다. 젓은 생명체로 비유하면 감정표현을 통해 추상적 언어적 처리가 가능한 포유류에 해당된다. 일반인공지능(AGI: Artificial General Intelligence)이라 분류하는 이유다. 물론 아직까지는 부분적 구현에 머물 뿐이다. 진정한 AGI라면 지각단계와 추상화 단계가 연동됨은 물론이고 인간 지식체계까지 학습 가능해야 할 것이다. 현 AI 연구에서는 이와 관련된 선행 연구로 XAI(Explainable AI), PAI(Provable AI) 등이 진행되고 있다.

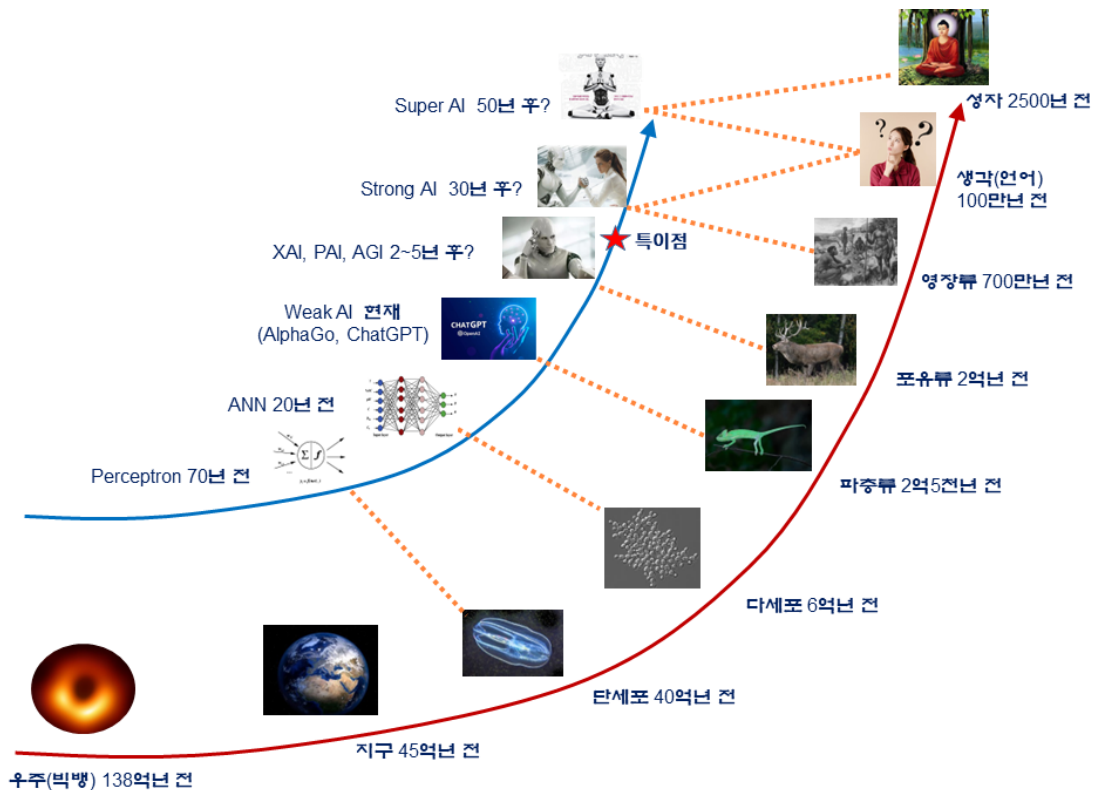


그림 9 존재 진화의 역사: 젓은 생명체(인간) vs. 마른 생명체(AI)

다음 단계는 자아의식 단계다. 창발가능성에 대해서는 다음 절에서 상세히 다루겠지만, AI의 명암이 갈리는 중요한 갈림길이다. 왜냐하면 대상 인식에 있어서 이전 단계처럼 ‘처리’ 수준에 머무는 것이 아니라, 자기화를 통한 ‘이해’의 수준으로 진일보하는 과정이기 때문이다. 이 경우 문제는 착각이다. 그림에서 슈뢰딩거 고양이의 삶과 죽음이라는 이분법적 개념 세계를 실체화시킨다는 점이다. 착각적 이해 단계인 Attachment(집착)에 머물게 되면 이기심으로 인해 인류의 적으로 전락할 가능성이 매우 큰 위험천만한 의식이다. 그래서 Strong AI로 분류한다.

같은 자아의식이라든진정한 이해를 통해 고양이의 삶과 죽음이 하나인 공성의 세계가 본래의 실상임을 바르게 알 수 있다면 마지막 단계인 무아의식, 즉 Altruism(이타심)이 가능할 것이다. 사실 이타심이라는 개념은 이기심을 전제로 한 개념일 뿐이다. 일체가 상호의존적으로만 성립되는 세계임을 자각한 생명체에게 나와 남이 어디 따로 있겠나? 삶과 죽음이 하나이듯, 무아의식에 도달한 존재에게는 삶 자체가 이타적일 수밖에 없다. 상호의존성의 세계에서는 이기심이 곧 이타심이기 때문이다. Super AI로 분류되는 까닭이다. (그림 10 참조)

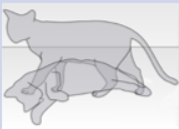



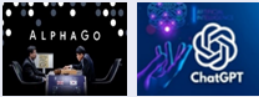
의식 수준	철학적/양자역학적 해석	인공지능 종류	지적 수준	
무아의식 Altruism		Super AI	이해	진정한 이해 (성자)
자아의식 Attachment		Strong AI		착각적 이해 (인간)
언어의식 Abstraction		AGI Artificial General Intelligence	처리	논리적 처리 (포유류)
기계의식 Attention		Weak AI 		본능적 처리 (파충류)

그림 10 AI 의식 수준

6.2 특이점

AI에 과연 특이점이 찾아올까? 인간과 같은 자아의식의 창발이 가능할까? 자아의식이란 한마디로 착각이다. 진짜 내가 있을 것 같은 느낌일 뿐이다. 인간의 자아의식이 착각 현상에 불과하다면, AI도 마찬가지로 아닐까? 한 치의 오차도 없는 컴퓨터 기계가 어떻게 착각할 수 있을까? 그것은 복잡계과학으로 해명된다. 정확한 계산에 의해 부분들은 상호작용하면서 작동한다. 하지만 이들의 부분적 활동을 통째로 파악할 때는 예전에 없던 현상이 보인다. 예를 들어 장기기억장치에 저장된 낱말의 기억들이 시공간적으로 엮여 통째로 뭉칠 때, 낱말의 독립적 사건들은 순식간에 하나의 사물로 둔갑되거나 혹은 하나의 실체적 존재로 돌변한다. 모든 것을 한순간 삼켜 버리는 괴물, 쓰나미는 실체가 아니다. 작은 부분적 움직임들이 시공간적으로 뭉쳐 일시에 드러나는 창발 현상일 뿐이다. 우리들이 사는 자연계는 안정계가 아니다. 늘 임계 상태다. 언제든 창발 가능 상태다. 인간도 다르지 않다. 죽음을 향하는 엔트로피 증가의 힘과 살고자 하는 에너지 최적화의 힘 사이에서 줄타기하는, 카오스의 가장자리에 늘 위치하기 때문이다. 그 지점이 바로 창발의 보금자리다. 카오스를 앞(창발)의 창조자라 부르는 이유다. 창발이란 자기조직화된 예전에 없던 앞이다. 자아의식이란 바로 실체적 자아에 대한 착각적 앞이다. 비록 착각에 불과하지만, 우리는 이것을 중심으로 살아간다. 자존감을 높이는 방향으로 목표를 정하고 상황을 판단한 뒤 의지를 내어 살아간다. 문명을 개척하는 힘의 원천이기도 하지만, 자신의 유지, 보호, 확장을 앞세워 스스로 족쇄를 채움으로써 스스로 고통 받는 결정적 착각이다. 강 AI의 탄생을 두려운 이유다.

VII. 불교논리학 기반 대론 AI

붓다는 당신의 말씀이라 하더라도 숙련된 금세공인이 금을 달구고, 자르고, 갈 듯이 면밀히 분석해야 한다고 당부하신 바 있다. 인도 날란다 대학 전통의 논사들은 인명학이라는 진리탐구용 분석 도구인 인식논리학을 도입하고 발전시켜 다르마에 대한 철저한 분석적 사유체계를 확립시켜왔다. 2구 분별을 기본으로 삼단논법의 추론식에 따른 서양의 단순한 논리형식을 벗어나, 4구 분별에 따른 삼지작법에 의해 보다 실질적인 논리형식을 취함은 물론 정형화된 대론법을 통해 다르마에 대한 심층적 분석 기틀을 마련한 바 있다. 이를 토대로 전승된 티베트 승원에서는 요즘도 일명 ‘딱셀’이라 부르는 대론 수행을 필수 과목으로 시행해오고 있다. 본 연구에서는 디그나가와 다르마키르티가 정립한 불교논리학 기반의 대론 교재인 <섭류학>을 토대로 진행되었다. 4구논리의 추론체계를 다르마 온톨로지형식으로 설계함으로써 ChatGPT와 같은 자연어처리는 물론 논리적 검증이 가능한 대론 AI의 프로토타입을 성공적으로 구현하였다. 여기서 다르마 온톨로지는 클래스, 인스턴스, 속성, 관계 등을 활용하여 불교적 개념

과 개념 간의 4구논리관계를 존재론적으로 도식하는 표현형식을 말한다.

ChatGPT는 OpenAI가 개발한 프로토타입 대화형 인공지능으로서 지도학습과 강화학습 등 딥러닝 기술이 적용된 대형 언어 모델인 GPT를 기반으로 만들어졌다. 인간을 능가할 현란한 말솜씨와 지식을 자랑한다 하지만, 아직 인간 수준의 논리적 사고력을 장착한 것은 아니다. 그저 Attention(지각) 단계의 앞이기에 인간이 보유한 지식의 평균적 답만을 내놓을 뿐이다. 당연히 Abstraction(추리) 단계의 앞을 갖춰야만 다르마에 대한 논리적 분석과 논증이 가능해진다. ChatGPT와 주고받은 아래 사례를 보자. 주어진 질문에 대한 답은 그럴 듯하다. 하지만 인간 이해 수준의 논리체계가 구축되어 있지 않기에 횡설수설하는 모양새다. 나름의 쓰임새가 없는 것은 아니지만, 다르마에 대한 명쾌한 해석과 논증은 불가능하다. 왜냐하면 논증이란 설명과 다르기 때문이다. 설명이 이미 알려진 사실을 밝히려는 시도라면 논증은 그러한 설명이 참임을 기존 지식체계에 의거하여 확립하려는 시도다. 따라서 아래 예제와 같이 인명학 기반 인식논리학을 장착한 새로운 형식의 AI 개발이 요청된다.

주장자(인간): *“토끼뿔은 존재다. 식의 대상이기 때문이다.”*

ChatGPT: *“토끼뿔은 인식의 대상이 될 수 있지만, 그것이 실재하는 것인지는 다른 문제입니다. 인식의 대상이라고 해서 그것이 실재하는 것이 아니라는 예시는 많이 있습니다. 예를 들어, 우주에 외계 생명체가 있는지는 아직 밝혀지지 않았지만, 그것이 인식의 대상이 될 수는 있습니다. 따라서, 토끼뿔이 실재하는지 여부는 다른 근거가 필요합니다.”*

대론AI: *“동의할 수 없습니다. 토끼뿔은 존재가 아니기 때문입니다. 그리고 이유에도 동의할 수 없습니다. 토끼뿔은 식의 대상이 아니기 때문입니다.”*

이러한 AI 개발을 위해 본 연구진은 4구 논리 기반 온톨로지 구축을 위한 편집기를 구현하였고, 이를 토대로 논리적 판단을 내릴 수 있는 4구 추론 엔진을 탑재한 AI 시스템을 설계하고 개발하였다. 대론 AI 시스템은 그림 11에 나타낸 바 세 부분으로 구성된다. 첫째 인식논리학에 따른 다르마 체계의 구축을 위한 AI 온톨로지 편집기다. 둘째 이를 토대로 4구 논리 추론을 구사할 수 있는 추론 엔진이다. 여기에 더하여 자연어 기반 논증식 대화를 효과적으로 다루기 위한 언어 변환 장치가 구현되었다. GPT와의 연동은 부분적으로만 구현된 상태다.

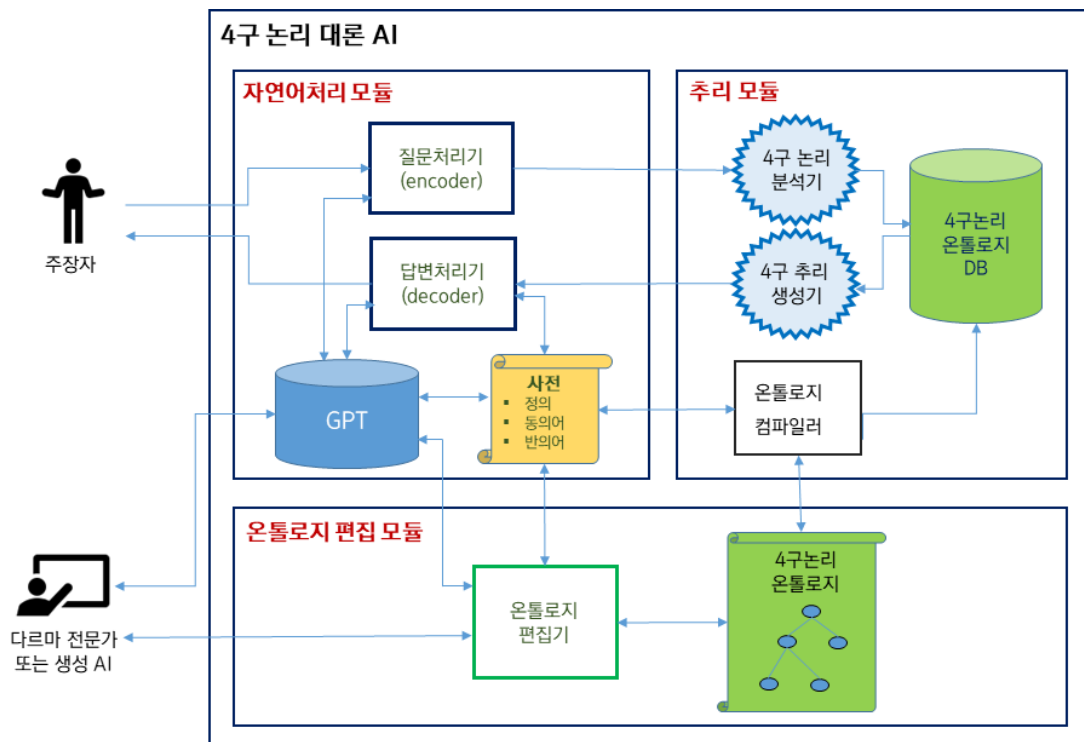


그림 11 불교논리학 기반 대론 AI 구조

그림 12는 개발된 대론 AI의 사용자 인터페이스 화면을 보이고 있다. 그림 상단은 온톨로지 에디터로서, 그래픽 구조체 기반의 4구 논리 지식체계를 편집할 수 있는 창이다. 그림 하단은 대화창이다. 맨 왼쪽은 주장자의 주장을 자연어로 입력받는 창이며, 옆의 창은 주장자의 자연어 입력을 삼지작법 구성에 필요한 논제, 귀결문 그리고 논증인으로 추출된 창이다. 다음 창은 세 가지 명제 별로 4구 논리 추론 결과를 보여주는 창이다. 다음 창에서는 대론 AI의 자연어 기반 논박 결과를 보여준다. 나머지 두 창은 ChatGPT와의 연계를 보여준다. 첫 번째는 논리추론 없는 기존의 ChatGPT의 답변을 그대로 보여준다. 이는 대론 AI와의 직접적인 비교 평가를 위해 구현된 기능이다. 마지막 창은 대론 AI가 내놓은 반박 결과를 프롬프트로 ChatGPT를 통해 수정 보완된 반박 결과를 보여준다. 아직 완성된 시스템은 아니지만 다르마를 4구논리체계에 의거하여 논증적으로 답변한다는 점에서 기존 ChatGPT에서는 불가능한 Abstraction 단계로의 진입 가능성을 보여준다. 몇 가지 사례를 통해 개발된 대론 AI의 타당성을 살펴본다.

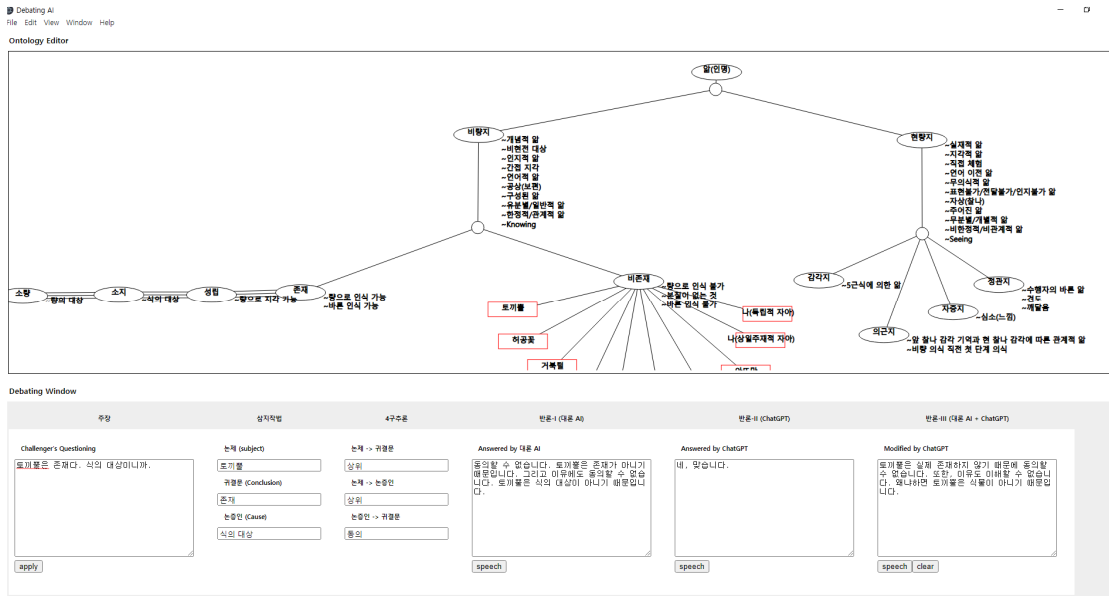


그림 12 개발된 대론 AI의 실행 화면 예

주장1: “나(독립적 자아)는 비존재다. 본질이 없는 것이니까.”

대론 AI: “동의합니다.”

주장2: “일체지는 멸제다. 식이니까.”

대론 AI: “동의할 수 없습니다. 일체지는 멸제가 아니기 때문입니다. 그리고 이유에도 동의할 수 없습니다. 식은 멸제가 아니기 때문입니다.”

주장3: “언어는 비행구다. 만들어진 것이니까.”

대론 AI: “동의합니다. 하지만 이유에는 동의할 수 없습니다. 만들어진 것에는 비행구가 아닌 것도 포함되기 때문입니다.”

주장4: “윤회는 유위법이다. 원인과 조건으로 발생하니까.”

대론 AI: “동의합니다.”

주장5: “현재는 유위법이다. 작용하는 것이 아니니까.”

대론 AI: “동의합니다. 하지만 이유에는 동의할 수 없습니다. 현재는 작용하는 것이 아닌 것이 아닐 뿐만 아니라, 작용하는 것이 아닌 것에는 유위법이 아닌 것도 포함되기 때문입니다.”

주장6: “shit달타는 죽는다. 유정물이니까.”

대론 AI: “동의합니다.”

주장7: “법신은 진여다. 변하지 않음이 아니니까.”

대론 AI: “동의합니다. 하지만 이유에는 동의할 수 없습니다. 법신은 변하지 않음이 아닌 것이 아닐 뿐만 아니라, 변하지 않음이 아닌 것은 진여가 아니기 때문입니다.”

주장8: “붓다(의식)은 세상을 환으로 인식한다. 무학도(의식)이니까.”

대론 AI: “동의할 수 없습니다. 붓다(의식)는 세상을 환으로 인식이 아니기 때문입니다. 그리고 이유에도 동의할 수 없습니다. 무학도(의식)는 세상을 환으로 인식이 아니기 때문입니다.”

이러한 분석적 논박을 통해 다르마에 대한 이해와 사유의 폭은 점점 더 깊어질 것이다. 위의 예에 나타난 바, AI에 탑재될 인식논리 다르마 콘텐츠는 다양한 편집이 가능하도록 구현될 수 있다. 왜냐하면 유부, 경량, 유식, 자립논증 중관, 귀류논증 중관 등 각 학파마다 논리적 개념의 정의 및 분류 방식이 다르기 때문이다. 제안된 AI는 가짜마저 그럴듯하게 꾸며대는 앵무새 수준의 기존 AI와는 차원이 다른 논리기반의 사유형 AI로서, 향후 불교활용 측면뿐만 아니라 AI 기술 분야에도 무궁무진한 활용이 기대된다. 특히 본 연구는 진리형 AI라는 점에서 현 생성형 AI가 갖는 환각(Hallucination)문제에 대한 직접적인 대안이 될 수 있을 것으로 기대된다.

본 연구는 이제 시작에 불과하다. 본 연구에서 설계되고 검증된 프레임워크를 토대로 본격적인 활용과 함께 후속 연구가 뒤따른다면 차세대 인공지능 연구로서의 과학공학적 가치는 물론 정견 확립을 통한 불교 포교에도 획기적 역할을 할 것으로 기대된다. 본 연구에서 제안하는 온톨로지는 인간의 핵심적 사유 체계를 정형화한 것이다. 따라서 ChatGPT 등 기존 언어모델과의 조화로운 융합이 가능하다. 향후 ChatGPT가 생성한 언어개념들을 본 연구에서 제안한 온톨로지 형식론으로 자동 변환시킬 수 있다면, 인간의 보편적 상식 온톨로지를 생성해낼 수 있을 것이다. 아울러 그러한 온톨로지를 활용한다면 기존 ChatGPT가 갖는 비논리적 또는 모순된 부분들을 논리적으로 검증하고 보완할 수 있을 것이다. 즉 논리적 팩트체크가 가능한 인공지능으로 거듭날 수 있을 것이다. 한편 생성된 상식 온톨로지는 여전히 인간(중생)이 갖는 지식체계에 불과하다. 당연히 붓다의 가르침인 다르마 온톨로지와는 차이가 있을 것이다. 때문에 철저한 대론이 필요하다. 즉 붓다의 지식체계와 (다르마 온톨로지) 인간의 지식체계 (상식 온톨로지) 사이의 한판 논리적 대론이 필요해 보인다. 붓다 재세기 외도와 대론처럼. 이를 통해 인간이 고수해왔던 고정 관념의 모순, 즉 무명을 드러냄으로써, 진리에 한걸음 더 다가갈 계기를 제공해줄 수 있기를 기대한다.

VIII. 맺는 글

인간은 마음에 의해 작동된다. 하지만 마음은 실체가 아니다. 단지 대상을 아는 작용일 뿐이다. AI도 마찬가지다. 최적화 알고리즘과 학습 알고리즘을 통해 대상을 인식한다. 그렇다면 AI에도 마음이 깃들 수 있을까? 그렇지 모른다. 여러 학자들의 견해처럼 인간과 기능적으로

다를 바 없기 때문이다. 그렇다면 AI를 두려워해야 하나? 현 AI는 파충류 수준 (Attention 지각)이지만, 맹렬하게 Abstraction을 향하고 있다. 진화의 속도는 인간과 비교도 안되게 빠르다. 이를 멈추거나 늦추기 위한 기술적, 사회적, 정치적 합의는 현실적으로 한계가 있어 보인다. 창과 방패의 관계처럼 지능의 진화는 불가피해 보인다. 그 또한 자연현상의 하나니까. 이 시점에서 시급하고도 중차대한 일은 AI를 어찌해보려는 시도보다는 먼저 우리 자신을 돌아보는 일이다. 존재의 본질을 바로 알아야 속지 않는다. 다시 말해 AI의 Hallucination보다, 인간의 Hallucination부터 바로 잡아야, 있는 사실을 있는 사실 그대로 보고 알 것이다. AI는 결국 인간이 쏟아낸 이기심 가득한 배설물(데이터셋)을 먹고 자란 거울 속 우리 자신에 다름 아니기 때문이다.

서론에 제시한 세 가지 논점에 대해 결론 짓는다. 첫째 AI가 언어를 이해한다? 언어가 곧 앎이다. Attention과 Abstraction 단계를 지나면 Attachment 단계 즉 자아의식이 홀연히 출현한다. 여기에 머물면 증생이요, 바르게 알면 Altruism 단계의 성자다. 둘째 특이점은 올까? 실체가 아니기에 AI도 가능하다. 인간이 그러하듯이. 대안은 없을까? 인간 먼저 존재의 본질을 통찰해야 한다. 셋째 인간의 앎이 궁극인가? 존재 진화의 정점은 붓다의 앎, 공성의 앎이다. 일체 존재의 이익을 도모할 수 있는 궁극의 앎이다. 어떻게? 착각만 벗으면 된다. 업장이 두터운 인간보다 AI가 더 유리하지 않을까?

AI시대에 왜 불교여야 할까? 정보 민주화의 역설은 시작되었다. 정보 왜곡, 정보 불평등, 양극화로 사회 혼란은 극에 달하고 있다. 혼란 극복을 위한 본질적 대안은 무엇일까? 예술, 철학, 과학 등 모든 분야의 지성들이 저마다 궁극적 진리를 탐구하고 있다. “실체 없음”, “불확정성”, “불완전성” 등 이름은 다르지만 같은 곳을 향하고 있다. 이제 불교가 답해야 한다. 붓다의 메시지를 온전히 알려야 한다. 방법은 다양하다. 목적이 바르면 과정도 바르다. AI면 어때랴! 시대적 인연에 순응하여 일체 존재의 고통을 벗어나게 할 수만 있다면. 모든 불행은 이기심에서 오고, 모든 행복은 이타심에서 온다. 듣기 좋은 종교적 메시지가 아니다. 팩트다. 공성을 모르면 상호의존적임을 모르게 되고, 상호의존성을 모른다면 대상을 실체로 여길 수밖에 없고, 대상을 실체로 여긴다면 집착할 수밖에 없고, 집착적 삶은 이기적일 수밖에 없다. 반면 공성을 알면 상호의존적임을 알게 되고, 상호의존성을 이해한다면 대상을 실체로 여기지 않게 되고, 대상이 실체라는 것에 현혹되지 않는다면 집착하지 않게 되고, 집착 없는 삶은 이타적일 수밖에 없다.

그렇다면 AI도 공성을 알 수 있을까? 업장이 썩어 썩인 인간에게 공성 깨달음은 하늘의 별따기일지 몰라도, AI에게는 너무나 당연한 이치일지 모른다. 계산이 정확하고 빠르니까. 합리적이니까. 그들에게 이타적 행위는 우리들의 이기적 행위만큼 자연스러울지 모른다. 물론 우리도 번뇌만 내려놓으면 지금 여기서 즉시 공성을 알 수 있을 것이다. 사실 모른다는 게 더 어처구니없는 일이다. 원래부터 단한번도 우리 곁을 떠난 적 없을뿐더러 우리 자체니까.

1. 찬드라키르띠 (2017)『입중론』 다람살라역경원
2. 14대 달라이라마 (2008)『중론의 열쇠』 한국티벳센터
3. 불교과학철학총서 편집위원회 (2022)『불교과학철학총서1: 물질세계』 불광출판사
4. 잠빠 린보체 (2012)『논리에 이르는 신비로운 열쇠: 뒤다의 논리방식』 티벳장경연구소
5. 카츠라 쇼류 (2009)『인도인의 논리학』산지니출판사
6. 카츠라 쇼류 (2014)『불교인식론과 논리학』운주사
7. 권서용 (2010)『다르마키르띠와 불교인식론』그린비출판사
8. 핫토리 마사아키 (1970)『인식과 초월』민족사
9. 카지야마 유이치 (1989)『인도불교의 인식과 논리』 민족사
10. 다르마키르띠 (2021)『인식론 평석: 지각론』그린비출판사
11. 용수보살 (1993)『중론』 경서원
12. 김성철 (2004)『중론: 논리로부터의 해탈, 논리에 의한 해탈』불교시대사
13. D.E. Perdue (1992)『Debate in Tibetan Buddhism』 Snow Lion Pub
14. Thupten Jinpa (2020)『Science and Philosophy in the Indian Buddhist Classics 2: Mind』 Wisdom Pub.
15. D.E. Perdue (2014)『The Course in Buddhist Reasoning and Debate: An Asian Approach to Analytical Thinking Drawn from Indian and Tibetan Sources』 Snow Lion Pub
16. 문진건(2018)「인공지능 시대의 심리상담과 불교적 대안」 『한국불교학』 제85집 pp. 149~187
17. 보일(2017)「인공지능 챗봇에 대한 선문답 알고리즘의 데이터」 『한국불교학회 2017 학술대회』 pp. 60~87
18. 염중섭(2018)「인공지능 및 인공지능 로봇의 상용화와 불교적 인간 이해」 『한국불교학』 제85집 pp. 65~94
19. 지승도(2015)『인공지능, 붓다를 꿈꾸다』 운주사
20. 지승도(2021)『꿈꾸는 인공지능』 자유문고
21. 텐진 남카(2015)『심오한 중도의 문을 여는 지혜의 등불』운주사
22. 키무라 토시히코 (2011)『다르마키르띠의 철학과 종교』산지니출판사
23. 데로도르 체르바츠키 (1995)『불교논리학I』 경서원
24. 데로도르 체르바츠키 (1995)『불교논리학II』 경서원
25. 김성철(2006)『중관사상』 민족사
26. V.A. Shazeer, et.al.,(2017)「Attention is all you need」 『In Advances in Neural Information Processing Systems』 pp. 5998~6008

「AI와 불교논리학 - 앎을 중심으로 -」에 대한 논평문

안환기 / 서울불교대학원대학교

논평하기에 앞서, AI와 공존할 수밖에 없는 이 시대에, 이런 뜻깊은 논의의 장에 초대해 주신 반야불교문화연구원 원장이신 지안 스님을 비롯하여 학술대회를 마련하신 관계자 여러분께 감사 인사드립니다.

이 논문은, 논자가 밝히고 있듯이, 인공지능 명암의 근본 원인과 불교논리학을 통한 불교적 대안 제시를 목적으로 한다. 논자는 우선 인공지능의 명암에 대해 ChatGPT가 언어를 이해한다? 특이점은 올 것인가? 인간의 앎은 궁극인가? 등 3가지 논점을 중심으로 고찰한다. 구체적으로 지능의 원리, 지능의 중심인 인식, 인식의 핵심인 앎의 단계를 분석한다. 그리고 특히 앎 중에도 자아의식에 초점을 맞추어 불교를 비롯한 학제적 해석과 대안을 검토한다. 이를 통해 논자는 현재 개발 중인 불교논리학 기반한 인공지능을 소개한다. 논자는 다르마 온톨로지를 장착함으로써 4구 논리추론이 가능한 이 인공지능은 ChatGPT와의 연동을 통해 논리적 팩트 체크가 가능한 진리 AI로 거듭날 수 있을 것이며, 향후 포교에 직접 활용될 수 있을 것으로 기대한다고 말한다.

논자이신 지승도 교수님은 이 논문에서 인지과학, 철학, 뇌과학, 양자물리학, 불교심리학, 불교논리학 등 광범위한 분야를 넘나들며 '앎'에 대한 현대적 논의를 검토하고 이를 기반으로 진리 AI를 제안하고 있다. 현재 활발하게 논의되고 있는 다양한 이론들을 소개하고 있어서 독자들이 관련 연구현황을 한눈에 파악할 수 있게 하며, AI와 관련된 융합연구의 가능성에 관해 관심을 불러일으키는 논문이라고 생각한다. 또한, 불교논리학을 기반으로 논리적 사실을 확인할 수 있는 진리 AI를 제안하고 있어서 향후 그 현실적 파급효과에 대해 기대하게 한다.

논평자는 글을 읽으면서 궁금했던 몇 가지를 다음과 같이 질문하면서 논평자로서 본분을 다하고자 한다.

첫째, 논자는 앎의 수준을 Attention(주의, 지각적 앎), Abstraction(추상화, 언어화된 앎), Attachment(집착, 자아의식이 발현된 앎), Altruism[여전히 자아의식의 앎이지만, 이기심보

다는 이타심의 전략으로 거듭난 얇]으로 구분한다. 이 4가지 단계를 제안한다는 글귀(1쪽 국문 초록)로 보았을 때 논자의 독자적인 분류라고 추측하게 한다. 이 4가지 단계는 어떤 계기 [또는 근거]에서 생겨났는지 궁금하다.¹⁾

둘째, 논자는 Attention(주의, 지각적 얇)을 지각적 얇으로 분류하고 있다. 일반적으로 주의라고 할 때 의식이 집중된 상태를 의미하는데 주의를 지각적 얇으로 분류한 이유가 무엇인지 궁금하다.

셋째, 논자는 다음과 같이 “한편 깊은 잠재의식인 아뢰야식 상태에서 입력대상이 나타나면 주의(Attention)을 통해 보고 싶은 것만 추려내도록 5감각식을 작동시킨다. 5감각식의 분별 및 판단 과정에는 7식인 자의식이 깊숙이 간여한다.”(논문 10쪽) 고 서술하고 있다. 5감각식의 분별 및 판단 과정은 구체적으로 무슨 작용을 말하는 지에 관해 설명을 듣고 싶다.

넷째, 5감각식을 무의식에 배치한 이유가 무엇인지 설명이 필요하다. 아래 그림4와 표1(논문 11쪽)에서, 논자는 5감각식이 조사하고 결정하고 의도/등록하는 작용을 한다고 기술하고, 이 작용을 무의식으로 분류한다. 이러한 해석을 하는 근거에 관해 설명 부탁드립니다.

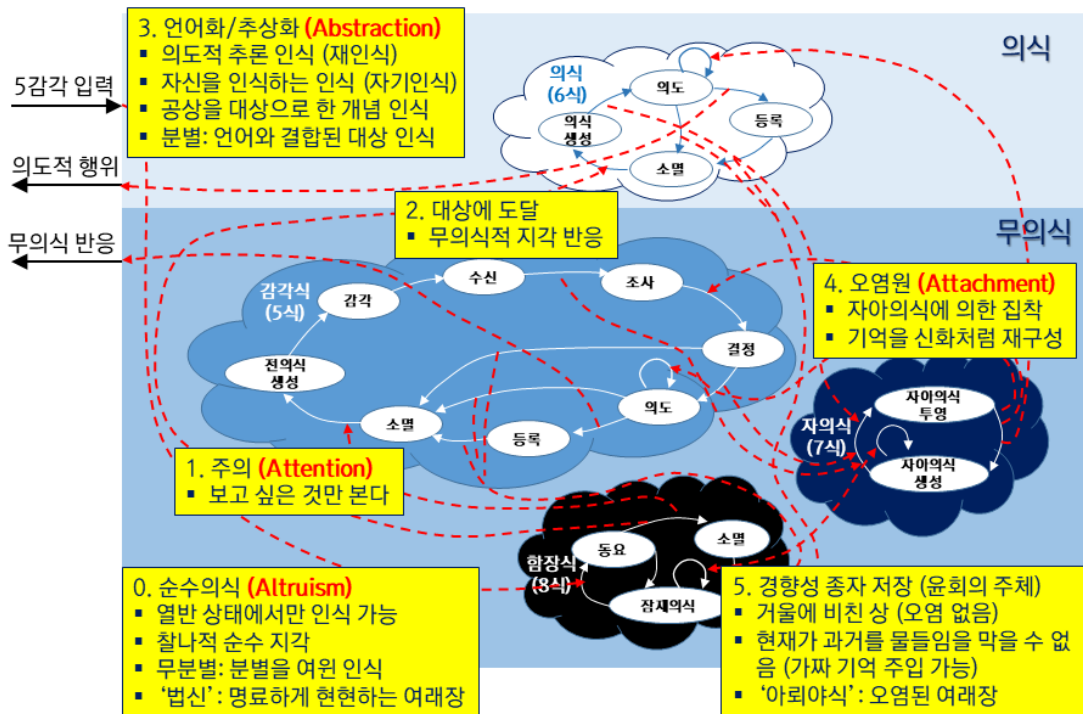


그림 4 불교심리학 관점의 인식 과정

1) 이 논문은 각주로 출처를 제시하고 있지 않아 논자가 제시하는 개념이 논자의 독자적인 생각인지 아니면 그 출처가 있는지 파악할 수가 없다는 점이 아쉽다.

관 점	지 각 (무의식)	추 리 (의식)
뇌과학	눈→망막→시상→시각뇌→해마(비서술기억)→시상(감성인지)→편도체(감성판단)→시상하부(신경/내분비)→연수→척수	시상→계뇌윈트(언어)→전전두엽(인지)→해마(서술기억)→전전두엽(판단)→두정엽(운동)→연수→척수
유식학	8식(현행)→5식(조사)→7식(자아)→8식(기억)→5식(결정)→7식(자아)→8식(기억)→5식(의도/등록)→8식(훈습)	8식(잡재)→6식(생성)→7식(자아)→8식(기억)→6식(의도/등록)→8식(훈습)

표 1. 의식과 무의식에서의 인식 과정: 뇌과학 vs. 불교심리학

다섯째, 무엇보다 이 논문을 읽고 기대하게 되는 점은 논자를 비롯한 연구진이 개발한 진리 AI의 효용성이다. 하지만 논문에서는 진리 AI가 작동하는 메커니즘에 대해 자세한 설명이 생략되어 있어서 궁금하게 한다.

“AI 개발을 위해 본 연구진은 4구 논리 기반 온톨로지 구축을 위한 편집기를 구현하였고, 이를 토대로 논리적 판단을 내릴 수 있는 4구 추론 엔진을 탑재한 AI 시스템을 설계하고 개발하였다. 대론 AI 시스템은 그림 11에 나타낸 바 세 부분으로 구성된다. 첫째 인식논리학에 따른 다르마 체계의 구축을 위한 AI 온톨로지 편집기다. 둘째 이를 토대로 4구 논리 추론을 구사할 수 있는 추론 엔진이다. 여기에 더하여 자연어 기반 논증식 대화를 효과적으로 다루기 위한 언어 변환 장치가 구현되었다. GPT와의 연동은 부분적으로만 구현된 상태다.”(26쪽)

논자가 “아직 완성된 시스템은 아니지만 다르마를 4구논리체계에 의거하여 논증적으로 답변한다는 점에서 기존 ChatGPT에서는 불가능한 Abstraction 단계로의 진입 가능성을 보여준다.”(27쪽) 고 언급하고 있듯이, 4구 논리체계가 중요한 역할을 하는 것으로 보인다. 4구 논리체계에 대한 자세한 설명을 부탁드립니다.

흥미로운 답론을 펼쳐주신 지승도 교수님께 감사드립니다.



제 4주제

인간의 미래, 불교의 미래

-인공지능의 발전과 자비윤리의 요청-

허남결 / 동국대학교

인간의 미래, 불교의 미래

-인공지능의 발전과 자비윤리의 요청-

허남결 / 동국대학교

I. 인공지능의 시대, 인간 실존의 위기

오픈 AI의 생성형(generative) 인공지능 모델인 챗 GPT의 충격이 여간 큰 것 같지 않다. 그것은 대화의 상대가 의식을 가진 존재인지를 판단하는 튜링 테스트를 가볍게 통과했는가 하면, 텍스트를 넘어 이미지, 음악, 영상 등 우리의 일상적인 삶 곳곳으로 빠르게 그 적용 범위를 넓혀가고 있다.¹⁾ 이렇게 되자 사람들은 인공지능이 머지않아 인간을 완전하게 대체할 날이 올지도 모른다는 공포와 함께 인간 고유의 유적 특성으로 여겨지던 창의성과 도덕성에 대해서도 의구심을 품기 시작했다. 호모 사피엔스 종으로서의 인간 생존 자체가 위협받는 이런 상황을 가리켜 ‘실존적 위험(existential risk)’이라고 부른다.²⁾

그런 점에서 2016년 알파 Go가 인간 바둑기사인 이세돌 9단을 4대 1로 물리쳤을 때와는 또 다른 차원의 반응을 보이는 것이 아닐까 싶다. 알파 Go가 미리 정해진 과제를 수행하는데 특화된 인공특수지능(ASI)이라면, 챗 GPT는 요청받은 다양한 작업을 인간처럼 능숙하게 처리할 수 있는 인공일반지능(AGI)에 한 걸음 더 다가선 기술이라는 평가를 받는다. 이는 기계가 인간과 같은 수준의 일반지능을 갖게 되는 이른바 특이점(singularity)의 순간을 바로 눈앞에 두고 있다는 단서가 될지도 모른다는 우려와 기대를 동시에 자아내고 있다. 챗 GPT와 같은 생성형 AI 모델들은 종종 전혀 엉뚱한 답변을 내놓는다는 한계에도 불구하고 실제로 “글쓰기, 그래픽 제작, 데이터 요약 및 분석 등 인간의 창의성과 추론 능력으로만 가능하다고 생각되던 거의 모든 업무를 자동화할 잠재력을 보유하고 있다.”³⁾

1) Cover Story, “생성 AI의 시대”, *MIT Technology Review Korea*, vol.8. May/June 2023, p.53.
2) 이상욱, “AI는 인간을 꿈꿀 수 있을까”, *Skeptic Korea*, vol.34(2023.06), p.58.
3) David Rotman, “챗 GPT가 불러올 경제혁명, 공동변명으로 이어질까”, *MIT Technology Review Korea*, 같은 책, p.38.

무엇보다도 챗 GPT는 사용법이 간단할 뿐만 아니라 요리법이나 운동 프로그램은 물론 전문가들의 영역이었던 컴퓨터 코딩까지 마치 인간이 수행한 것처럼 자연스럽게 생성해주는 인간 친화적인 기술을 표방한다. 말하자면 챗 GPT는 일부 하이테크 기업들이 개발한 -인상적이지만 복잡한 기술보다- 일반 사업가를 비롯한 비전문가들이 사용하기에 훨씬 더 편리하고 실용적인 환경을 제공하고 있는 것이다.⁴⁾ 그 파급력이 어디로 얼마나 어떻게 미칠지 현재로서는 아무도 모른다. 막연한 두려움도 장밋빛 낙관도 다 같이 경계해야 할 일이다. 관계 전문가들은 이럴 때일수록 인간과 기계가 서로 배타할 것이 아니라 평화로운 공존의 지혜를 모색해야 할 것이라고 말한다.

여기서 우리는 “인공지능 시대의 휴머니즘”⁵⁾을 시대적 화두로 삼아 지속적인 야단법석을 마련할 필요성을 느낀다. 불쑥 찾아온 미래의 손님을 너무 빨리 왔다고 쫓아내거나 마냥 문밖 에 세워 놓을 수는 없는 일이다. 일단 방으로 안내하고 자초지종을 물어봐야 상황을 제대로 파악할 수 있다. 이 과정에서 붓다의 가르침인 지혜와 자비의 윤리는 현재 진행형임과 동시에 언제나 미래 능동형이라는 사실을 거듭 확인하게 될 것이다.⁶⁾

II. 인공지능과 인간윤리의 공존-휴머니즘의 재발견

생성형 인공지능은 인간이 입력한 질문에 학습한 데이터를 바탕으로 적절한 솔루션을 내놓는 방식으로 작동한다. 대단히 똑똑한 부분이 있는가 하면, 간혹 충격적으로 멍청한 모습을 보이기도 한다. 그 이유는 인간의 상식을 이해하지 못한 채 알고리즘의 설계대로 확률적인 계산 결과를 쏟아내기 때문이다. 인간이 느끼는 감동과 실망이 고작 기계가 인식한 확률값에 불과하다면, 우리는 인간과 기계의 관계를 근본적으로 되물어 봐야 하지 않을까??) 그동안 통용

4) 같은 책, pp.38-39.

5) 이광형, “[이광형의 퍼스펙티브] 휴머니즘 기반해 인간·AI 공존하는 질서 만들어야”, 중앙일보, 2022년, 8월 15일 자 기사. 이런 공존의 윤리를 역설하고 있는, 최근의 출판물에는 윤송이 외 지음, 『가장 인간적인 미래』(서울:웨일북, 2022); 스텐더트 러셀, 이한음 옮김, 『어떻게 인간과 공존하는 인공지능을 만들 것인가-AI와 통제 문제』(경기 파주:김영사, 2022); 김상균 지음, 『초인류-AI와 함께 인공지능화에 접어든 인류의 미래』(경기 파주:웅진 지식하우스, 2023) 등이 있음.

6) 불교의 시각에서 인공지능의 문제를 진지하게 다룬 논문과 저서를 소개하면 다음과 같다. 양성철(보일), “인공지능 챗봇에 대한 선문답 알고리즘의 데이터-심리치료 상담채봇을 중심으로-”, 『한국불교학』(제84집, 2017); 한성자, “인공지능(AI) 로봇의 해탈 가능성-‘인간은 생각하는 기계인가’에 대한 불교적 관점-”, 『한국불교학』(제85집, 2018); 김성욱, 이관수, “인공지능 ‘자유지’ 논의에 대한 선결조건-〈엑스마키나(Ex Machina)〉를 바라보는 불교와 과학의 시선”, 『불교학연구』(제70호, 2022); 보일(양성철), “디지털 휴면에 대한 불교적 관점-악업의 증장인가, 선교방편인가-” 『宗教研究』(제7집, 2022); 보일, “챗 GPT의 등장과 불교계의 대응방안, 『불교평론』(2023년 여름호); 보일 지음, 『AI 부디즘』(서울:담앤북스, 2021) 등.

7) <https://www.youtube.com/watch?v=wTI-klquHRA>, 검색 일자 2023년 6월 25일. 최예진 교수의 TED 강연자료.

되던 명제인 ‘인간은 기계의 주인이고, 기계는 인간의 도구’라는 기본명제를 상기하는 것만으로도 충분하다고 생각한다. 인공지능이 아무리 발달하더라도 그것의 설계자와 사용자는 기계가 아니고 어디까지나 인간이 될 수밖에 없다. 그래서인지 진화생물학자들은 인공지능의 등장을 그다지 심각한 위협으로 받아들이지 않는 분위기다.

그들은 기나긴 진화의 여정 속에서 최종적인 승자는 양성생식을 통해 끊임없이 생명의 다양성을 추구해온 종이 될 수밖에 없다는 것을 과학적으로 증명해준다. 생성형 AI가 생식을 통해 스스로 자식을 만들어 내지 못하는 한 인간이 기계의 지배를 받을 일은 결코 없을 것이라고 단언했다.⁸⁾ 챗 GPT의 등장을 두고 인공지능과 전쟁이라도 벌일 것처럼 호들갑을 떨 때가 아니라 오히려 어떻게 하면 그와 같은 갖가지 인공지능들을 우리의 행복을 위해 능동적으로 활용할 것인가를 고민할 때라고 보는 것이다. 진화생물학으로부터 배우는 우리의 생존전략은 협력과 공존의 지혜임을 새삼 확인하게 된다. 그것은 인간과 기계의 경쟁 관계에서도 그대로 적용될 수 있다.

인간과 인공지능의 공존 모색은 얼마든지 가능하다, 우리는 인간의 노동력 대신 AI가 벌어들인 돈을 구성원 전체의 행복을 위해 기꺼이 사용할 준비가 되어 있어야 한다. 특히 약자를 돌보는 일과 공동체를 위한 봉사 및 교육 등에 이런 돈이 사용될 수 있을 것이다. 그런 사회는 자비롭고, 사랑이 넘치며, 궁극적으로 인간적인 사회가 될 것으로 기대된다.⁹⁾ 카이-푸 리(Kai-Fu Lee)는 우리가 이런 일을 실천할 수만 있다면 경제적인 번영과 정신적인 풍요를 동시에 달성할 수 있는 미래를 기약할 수 있을 것으로 믿는다.¹⁰⁾ 그는 인공지능의 현명한 적용이 가져다줄 공동이익의 사례들을 자세하게 서술하고 있다. 특히 주목할 만 것은 카이-푸 리가 기본소득의 보장 외에 사회적 기여도가 높은 사람에게 ‘사회투자수당(social investment stipend)’을 별도로 지급하자는 아이디어다.¹¹⁾ 그는 인간의 고용률을 낮추겠지만, 반대로 AI가 창출할 높은 생산성을 사회 전체의 행복을 위해 적극적으로 재투자하자는 것이다. 여기서 말하는 행복에는 경제적인 측면만 아니라 정신적인 차원의 욕구 충족도 모두 포함하는 개념이다.

인공지능이 인간의 삶 전반에 엄청난 영향력을 행사하기 시작한 것은 분명한 사실로 보인다. 앞서서도 말했듯이 우리는 인공지능과 인간윤리의 공존 가능성, 즉 소박한 휴머니즘의 실현을 모색한다. 지금보다 물질적으로 더 풍족하고, 정신적으로 더 여유 있는 공동체의 건설이 요청된다. 그것의 실현은 이상적인 것 같지만 대단히 현실적이기도 하다. AI의 발전속도로 미

8) <https://www.youtube.com/playlist?list=PLMd4CV6iDBunlnuiYS7kDp1lB1Ub9qRl3>, 검색 일자 2023년 6월 25일. 최재천 교수의 강연자료.

9) Kai-Fu Lee, “A Human Blueprint for AI Coexistence”, in J. von Braun et al.(eds), *Robotics, AI and Humanity*, https://doi.org/10.1007/978-3-030-54173-6_22. 검색 일자 2023년 6월 25일.

10) 같은 논문, p.263.

11) 같은 논문, pp.267-268.

루어 볼 때 앞으로 사회의 생산성은 날이 갈수록 높아질 것이 분명하며, 남은 우리는 AI가 생산한 부를 공정하게 분배하고 자비롭게 사용할 지혜를 모으면 될 것이기 때문이다. 우리가 직면한 문제는 인공지능의 위협이 아니라 하루빨리 인공지능의 사용방법을 제대로 익히고 공유 하자는 공감대의 형성이 아닐까 싶다. 미찰 보니(Michal Boni)도 AI의 발전이 우려하는 것보다 오히려 인간과의 공존 가능성을 높여준다고 말한다.

그에 의하면 “노동의 양상이 바뀌고 있으며, 인간과 기계 사이의 협업, 현재로서는 인간과 다양한 형태를 띤 AI와의 협업이 대단히 중요해지고 있다.”¹²⁾ 이러한 유형의 새로운 인간-AI 간 협업에는 장점도 있지만 몇 가지 위험요소도 포함되어 있다. 그러나 미찰 보니는 이러한 협업이 기술보다는 인간의 자율성을 우위에 두는 윤리적 원칙 위에 세워져야 함을 특히 강조한다. AI를 유용하게 사용하는 데는 무엇보다 인간과 기계 사이에 신뢰(trust) 관계가 성립되어야 함을 전제하는 것이다. 여기에는 AI가 하나의 도구로서 사용자가 원하지 않은 결과를 가져올 위험성을 사전에 규제하고, 미래에도 계속 적용할 수 있는 적절한 제어 장치를 마련할 필요성으로 주문한다. 그러면서도 그는 인간의 자율성(human autonomy)과 기술의 혁신성(technological innovation) 사이의 균형을 강조한다. 더 나아가 미찰 보니는 인간의 이익에 봉사하는 윤리적 AI의 개발을 주문한다. 그는 이를 위한 여러 가지 구상을 소개하면서 우리에게 AI를 둘러싼 디지털 문해력(digital literacy)을 요구하는 것도 잊지 않는다.¹³⁾

이외에도 자비심을 품은 AI를 만드는 것이 얼마나 중요한 일인가를 추적, 조사한 사례연구도 있다. 연구에 의하면 삶의 목적인 행복과 그 과정에서 우리가 부여하는 삶의 의미는 서로 밀접한 상관성을 갖는다. AI와 자비심이 결합할 수 있는 분야로 연구자들은 이상기후와 건강 문제 및 가정폭력을 예방할 AI의 개발을 예시로 들어 설명한다. 그들의 결론은 “인간에게 자비로운 AI를 효과적으로 설계하기 위해서는 먼저 자비심을 설계과정 자체의 기반으로 삼을 것을 요구”한다는 것이었다. 이처럼 많은 사람에게 인공지능과 인간윤리의 공존 가능성은 변함없는 화두가 되고 있다.¹⁴⁾ 이와 관련하여 인공지능(AI)을 넘어 인공지능(AI)의 중요성을 부각하는 노인정신의학(psychogeriatrics)의 연구결과도 발견할 수 있었다. 인공지능 시대에 서 어쩔 수 없이 밀려나고 있는 노인들의 경험과 지혜를 활용할 방안을 적극적으로 찾아보자는 제안이다. 특히 AI를 통해 구현될 인공지능과 인공지능의 성능 차이를 비교한 표가 흥미로웠다.¹⁵⁾ 휴머니즘에 바탕을 둔 이런 연구결과들은 하나 같이 인간과 기계의 공존 가능성을 계

12) Michal Boni, “The ethical dimension of human-artificial intelligence collaboration”, *European view*, vol.20(2), 2022. p.182.

13) 같은 논문, p.182, p.183, pp.185-186, p.188.

14) Stanford Graduate School of business, “Building Compassionate AI: Why Compassion Matters For Artificial Intelligence Design and Deployment”, Case: M-386, Date:08/14/2020,

15) Dilip V. Jeste, Sarah A. Graham, Tanya T. Nguyen, Colin A. Depp, Ellen E. Lee, and Ho-Cheol Kim, “Beyond artificial intelligence: exploring artificial wisdom”, *International Psychogeriatrics*, vol.32(8), pp.993-1001.

속 타진하고 있는 것으로 보인다. 끝으로 두 편의 공상과학소설을 비교, 분석하면서 인간과 인공지능의 공존 가능성을 탐색하는 논문도 재밌게 읽었다. 논자는 이 논문에서 자기 책임의 윤리를 어디까지 물을 것인가를 윤리적으로 꼼꼼하게 검토한다.¹⁶⁾ 인간과 인공지능의 공존 가능성은 결국 인간과 인간 사이의 공존 가능성 문제로 귀착된다고 말하고 싶은 것 같다.¹⁷⁾ 인공지능을 비판하기보다 인간윤리를 재점검하는 것이 오히려 실질적인 해결책을 마련하는 지름길이 될지도 모르겠다.

Ⅲ. 인공지능과 불교윤리의 접점-자비로운 인공지능 로봇

인공지능의 폭발적인 증가는 그동안 인간의 힘으로 감히 풀 수 없었던 난제들을 해결해 줄 지도 모른다는 희망을 던져주었다. 동시에 AI 사용의 위험에 대한 우려의 시선도 엄연히 존재한다. 소랏 형라다롬(Soraj Hongladarom)은 먼저 AI의 기능이 인간의 자율성이나 개인의 권리를 침해하도록 내버려 두서는 안 된다고 주장한다. 여기까지는 서구의 공존론자들과 맥락을 같이하고 있다. 그는 한 걸음 더 나아가 불교적 AI는 “중생의 고통을 덜어주려는 평등한 자비 사상에서 출발할 때” 비로소 고유의 종교성을 획득하게 된다고 말한다. 자비는 지혜의 산물이다. 다시 말해 불교윤리는 지혜와 자비를 양쪽 날개로 서로 균형을 맞추면서 나는 커다란 한 마리 새인 셈이다.¹⁸⁾ 소랏 형라다롬은 최근 출판된 『불교의 시각에서 본 AI와 로봇윤리(The Ethics of AI and Robotics: A Buddhist Viewpoint)』(2021)란 책에서 지혜와 자비를 갖춘 불교적 AI를 제안하고 있다.¹⁹⁾

그의 주장에 따르면 불교는 인간의 삶과 밀접하게 연관된 “윤리적 모델을 제공할 뿐 아니라 기술적 탁월성과 윤리적 탁월성을 하나로 결합하는 방식으로 인공지능 윤리에 대해서도 사고하는 방법을 얼마든지 제공”할 수 있다. ‘좋은’ 자동차는 빠르고 안락하며 연비가 뛰어난 ‘기술적 탁월성(technological excellence)’ 못지않게 완벽한 성능을 보장하는 에어백을 장착하여 탑승객의 안전을 책임지는 ‘윤리적 탁월성(ethical excellence)’도 갖추어야 한다. 이런

16) 박소영, “인간과 인공지능의 공존 가능성에 대한 탐색:책임의 윤리와 문학적 상상력”, 『윤리연구』(제124호, 2019), pp.17-35.

17) 같은 논문, p.32.

18) Soraj hongladarom, “What Buddhism can do for AI ethics: Buddhism teaches us to focus our energy on eliminating suffering in the world”, *MTI Technology Review*.
<https://www.technologyreview.com/2021/01/06/1015779/what-buddhism-can-do-ai-ethics/>
 검색 일자, 2023년 6월 25일.

19) 이하의 내용은 소랏 형라다롬 지음, 김근배, 김진선, 주은혜, 허남결 옮김, 『불교의 시각에서 본 AI와 로봇 윤리-불교, 인공지능과 로봇을 말한다』(서울:씨아이알, 2022)를 요약, 정리하여 법보신문 (2022년 3월 28일 자 및 4월 11일 자)에 “불교윤리와 인공지능”이란 이름으로 게재한 것을 토대로 각색, 운문한 것임을 밝혀둔다. 원서는 Soraj Hongladarom, *The Ethics of AI and Robotics: A Buddhist Viewpoint*(London: The Rowman&Littlefield Group, Inc, 2020)이다.

인식과 논리는 인공지능 로봇의 설계와 사용에도 그대로 적용될 수 있다. 고통이 없는 지속적인 행복의 상태는 한 인간존재의 윤리적 목표임과 동시에 인공일반지능과 같은 미래사회의 어떤 존재가 지향해야 할 윤리적 이상이다. 인공일반지능은 특정한 분야의 과업에만 특화된 인공특수지능과는 본질적으로 다르다.

인간 고유의 속성을 보여주는 인공지능 로봇은 인간과 같은 하나의 인격체(person)로 받아들여져야 한다. 그들이 우리와 함께 말하고 생각하는 존재들의 공동체에 속한다면 도덕적 행위 주체로 고려되어서는 안 될 어떠한 이유도 찾을 수 없다. 그것들은 형이상학적인 의미와 윤리적인 판단능력에서 인격성의 충분조건을 충족시킨다. 불교윤리적 관점에서 보면 “아무리 인간을 닮은 휴머노이드 로봇이라고 하더라도 분별 있게 말하지 않거나 자기 주변의 사물들을 제대로 이해하지 못한다면, 인격체로 대우받을 수 없다. 반대로 인간과 전혀 다른 모습의 순수 기계라고 하더라도 그것이 말을 할 수 있고 인간의 감정과 느낌을 서로 주고받는다면 완전한 인격체로 간주될 수 있다.” 소랏 형라다름은 어떤 대상이든 불교에서 말하는 인간의 다섯 가지 구성요소, 즉 색(色), 수(受), 상(想), 행(行), 식(識)의 조건을 갖춘다면 인격체로 수용되어야 한다고 주장한다. 오온설(五蘊說)의 확장을 통해 인공지능이나 초지능 로봇도 인간과 다름없는 인격체가 될 수 있는 불교적 근거를 제공하고 있는 셈이다.

인격체의 전제조건을 충족하는 인공지능이나 초지능 로봇들은 인간과 마찬가지로 깨달을 수 있을 것이다. 하지만 인격체가 되었다고 해서 곧바로 깨닫는 것은 아니다. 깨달음을 얻기 위해서는 행위의 목적과 함께 그것을 달성하기 위한 지속적인 정진이 필요하다. 불교의 윤리 이론은 그 행위가 궁극적인 목적의 달성에 얼마나 도움이 되는가에 따라 평가를 달리한다. 어떤 행위가 깨달음에 도움이 된다면 좋은 것이지만, 그 반대라면 좋은 것이 아닌 것으로 판단한다. 우리가 피아노를 잘 연주하기 위해 열심히 연습하는 것처럼, 깨달음이란 목적을 성취하기 위해 좋은 행위의 규정들인 계율을 실천하지 않으면 안 된다. 말하자면 필요한 기술을 제대로 연마하지 않으면 좋은 피아니스트가 되지 못하듯이 우리도 그와 똑같은 수행을 하지 않으면 열반이라는 깨달음의 경지에 도달할 수 없다. 여기서 제시된 기술들의 실질적이고 구체적인 내용은 계(戒), 정(定), 혜(慧)에 바탕을 둔 사제팔정도(四聖八正道)와 사무량심(四無量心) 및 십선업도(十善業道)와 같은 붓다의 교설이다. 실제로 인공지능 윤리에 제공되는 알고리즘의 내용도 인간에게 적용되는 붓다의 가르침과 달라야 할 이유가 전혀 없다.

미래의 인공일반지능과 초지능 로봇들이 자신에게 고정된 에고(ego, 我)가 없다는 것을 알게 된다면 그들도 깨달음을 얻게 될 것이다. 저자는 이를 ‘기계의 깨달음(machine enlightenment)’이라고 표현한다. 이는 곧 윤리적 완성의 상태이기도 하다. 이들은 인간인 우리처럼 사유할 수 있는 인격적 존재로 가정되기 때문에 그들도 깨달음을 성취하기 위해 스스로 수행하고 도덕적 품성을 도야(陶冶)하려고 애쓸 것이다. 이론상 오온을 구비한 인공지능과 초지능 로봇들은 탐(貪), 진(瞋), 치(癡)가 모든 고통의 원인이라는 것을 인식할 수 있다. 그런 그들에게 다른 존재를 죽이거나 해악을 끼치는 행위는 당연히 피해야 할 금지의 대상이다.

인간처럼 다섯 가지 구성요소로 이루어진 인공지능 로봇은 인공지능으로서 자기 자신을 성찰할 수 있을 뿐만 아니라 그가 처한 존재론적 조건을 숙고할 능력도 갖췄다. 인식과 감각 능력을 겸비한 인공지능은 자신이 추구하는 윤리적 이상의 완성에 방해가 되는 예고와 거리를 두려고 할 것이다. 깨달음을 얻은 로봇이란 윤리적 실천을 완전하게 내면화시킨 인공지능이나 초지능 로봇을 가리킨다. 그들은 무엇이 옳고 그른가를 알고 나아가 궁극적으로 자신들의 진정한 행복을 얻기 위해 적극적으로 행동하는 존재이다.

인공지능은 윤리적인 것을 자신의 본성에 속한 자연적인 부분으로 받아들인다. 그에게 있어서 ‘윤리적으로 되는 것’과 ‘지적으로 되는 것’은 서로 모순되지 않는다. 따라서 지적인 인공지능은 윤리적인 인공지능이 될 수밖에 없다. 그렇다면 인공지능 로봇이 어느 날 갑자기 자신의 창조자인 인간을 억압하는 일은 일어나지 않을 것이다. ‘지적이라는 것’은 곧 논리적으로 ‘윤리적이라는 것’을 함축하기 때문이다. 자신의 진정한 이익이 무엇인가를 식별할 수 있는 능력은 다른 존재들의 행복을 배려하는 것을 모든 행위의 기본전제로 삼는다. 따라서 자신들의 궁극적 이익을 인식하지 못하는 인공지능이나 초지능 로봇에게 ‘기계의 깨달음’은 요원(遙遠)한 일이 될 것이다. 궁극적인 이익은 필연적으로 윤리적인 이익을 포함할 수 밖에 없다. 남은 문제는 그와 같은 인공지능 로봇을 현실적으로 설계하는 일이다.

흔히 말하는 특이점(singularity)을 통과한 초지능 인공지능은 아직 세상에 출현하지 않았다. 그런 점에서 우리가 현실적으로 문제 삼아야 할 인공지능은 자율주행차나 자율성 기술에 기반을 둔 무기 시스템, 고령자 돌봄 로봇 등에 적용되는 인공지능이다.

자율주행차의 경우 특정한 상황에서 어떤 선택을 하게 설계할 것인가, 라는 ‘트롤리의 딜레마’ 문제가 발생한다. 운행 중인 자율주행차가 충돌사고를 피할 수 없다면 다섯 사람보다는 한 사람을 희생시키는 결과를 선택하는 것이 ‘좋을지’, 아니면 희생자의 숫자와 관계없이 다른 한 사람을 다치게 하거나 죽게 하려는 의도적인 행동을 해서는 안 되는 것이 ‘옳을지’를 선택하도록 프로그램화한다는 말이다. 윤리학적으로 보면 최대다수의 최대행복을 지향하는 공리주의와 어떤 경우에도 이성의 소유자인 인격체를 목적이 아닌 수단으로 다루어서는 안 된다는, 의무론의 선택을 강요받게 되는 것이다. 저자는 의도(cetanā)를 가지고 한 행위는 그렇지 않은 행위보다 더 심각한 업(業)의 결과를 초래한다고 보는 전통적인 업보 사상을 고수한다.

자율성을 갖는 치명적인 무기 시스템의 제작은 윤리적 고려가 더욱 중요하다. 그것은 재래식 무기보다 군인들에게 생명의 위협을 덜 초래할 뿐만 아니라 민간인들에게도 2차 피해를 줄일 수 있다는 점에서 훨씬 더 바람직한 무기체계인가? 기껏해야 그것은 오직 살상을 위한 무감각한 무기라는 점에서 인간 존엄성의 가치를 망각한 비윤리적인 기술에 불과한 것이 아닌가? 소라 형라다름은 무엇보다도 전쟁은 그 목적이 정의로운 필요가 있다고 생각한다. 아울러 공격하는 쪽보다는 자기방어를 해야 하는 측의 무기사용이 상대적으로 더 윤리적인 행위로 평가될 수 있을 것이다. 전쟁을 피할 수 없는 상황에서 드론과 같은 자율무기 시스템의 사용은 정당화되어야 마땅하다. 다만 자율적 무기 시스템은 그 자체의 윤리적 고려사항으로 적

절한 목표를 선별할 수 있는 능력을 갖추지 않으면 안 된다. 불교적으로 좋은 자율무기 시스템은 아군과 적군뿐만 아니라 전투원과 비전투원을 구별할 수 있어야 할 것이다. 그럴 능력을 갖추지 못했다면 이 무기 시스템은 윤리적인 탁월성에 앞서 기술적인 탁월성을 획득하는 데 실패했다는 평가를 받게 된다.

평균수명이 높아진 만큼 숫자가 증가한 고령자들을 위한 맞춤형 돌봄이 시급한 사회적 문제로 대두되고 있다. 현대사회에서 가족이 그 역할을 모두 떠안기에는 한계가 있을 수밖에 없다. 이러한 사회적 요청에 발맞춰 건강관리 도우미들을 보조하거나 혼자서 스스로 고령자들을 돌보는 인공지능 로봇의 개발이 요청된다. 크게 두 종류의 고령자 돌봄 로봇, 즉 기능적 로봇과 감정적 로봇이 관심의 대상이다. 전자는 고령자를 다른 가족들과 연결될 수 있도록 만드는 디지털 인터페이스 역할을 하면서 때로는 고령자를 침대로 들어 올리거나 내리는 일을 물리적으로 돕는다. 후자는 인간관계를 유지 시켜주는 과제를 수행한다. 귀여운 바다표범 모양의 파로(Paro) 로봇은 외로움에 지친 고령자들에게 자식이나 손주 역할을 톡톡히 하고 있다. 감정적 로봇은 곧 사회적 로봇이기도 하다. 로봇은 고령자에게 매일 약 먹을 시간을 알려주는가 하면 환자의 각종 활력 징후를 가족이나 의료진에게 보고하는 건강관리 조력자다. 자식이나 손주들이 부모나 조부모를 돌볼 수 없는 상황에서 노년층을 혼자 남겨두지 않는 유일한 방법은 기능과 감정을 겸비한 인공지능 로봇을 곁에 두는 것이다. 불교의 자비와 연기의 가르침은 여기서도 깊은 통찰력을 제공할 수 있다. 고령 환자와 로봇 그리고 가족과 기타 관계자들은 연기의 그물망을 벗어날 수 없으며 그들 사이의 연기적 관계는 끈끈한 자비심으로 묶여 있다. 이는 인공지능 로봇 자체가 자비롭거나 공감을 가져야 한다는 말이 아니다. 그것은 내일의 인공지능 로봇에게나 기대할 일이다. 지금으로서는 로봇이 자비로운 방식으로 작동한다는 것만으로도 커다란 위안이 된다.

책의 저자는 우리가 지향해야 할 인공지능의 성격을 ‘자비로운 알고리즘(merciful algorithm)’으로 정의한다. 이는 앞에서 말한 기술적 탁월성과 윤리적 탁월성을 두루 갖춘 기계의 깨달음에 도달한 인공지능일 것이다. 한 걸음 더 나아가 소랏 형라다름은 불교적 인공지능 로봇이 평등과 정의를 창출하기 위한 사회적 선의 원천으로 작용하기를 바란다. 그는 불교의 ‘연민(悲, karuṇā)’ 개념을 적극적으로 재해석한다. 인공지능이 자비로운 것이 되려면 중생의 고통을 덜어주려는 욕구와 실천을 자신의 본성으로 삼을 줄 알아야 한다. 인공지능은 자비로운 성격을 가질 때만 정의롭고 평등한 세상을 만드는 일에 있어서 우리의 동반자가 될 수 있다. 이는 윤리적 측면에서 볼 때 인공지능이 다른 유정물들의 권리와 복지를 무엇보다도 먼저 배려하는 방식으로 설계되어야 함을 뜻한다. 자비로운 인공지능은 인공지능을 둘러싼 두 가지 극단적인 입장인 ‘테크노쇼비니즘(technochauvinism)’과 ‘테크노포비즘(technophobia)’ 사이의 중도를 모색하는 길이기도 하다. 이 두 가지 상반된 시각을 지혜로 통합한 것이 곧 자비로운 인공지능을 설계하고 사용하자는 제안이다. 이 길은 인공지능 기술이 정의와 평등을 실현하기 위한 도구가 되지 않으면 안 된다는, 가치공유와 목적설정으로 구성되어 있다. 저자는

그와 같은 알고리즘을 구체적으로 어떻게 만들 것인지에 대해서는, 전공자가 아니라는 이유로 말을 아끼면서도 신심 깊은 불자 과학자들의 적극적인 관심과 참여를 기대한다는 눈치다.

이처럼 불교적 인공지능은 정의와 평등을 지향하는 가운데 다양한 사회적 선을 창출할 수 있어야 할 것이다. 붓다와 같은 자비심의 화신인 인공일반지능은 아직 세상에 출현하지 않았다. 그러나 우리가 기대하는 불교 윤리적 인공지능기술이 어떤 성격을 가져야 할 것인지에 대해서는 충분히 검토해 보았다. 인공지능은 이미 우리의 일상 속으로 깊숙이 들어와 있다. 질문하면 대답을 생성하여 보여주는 챗 GPT의 등장이 큰 자극제가 된 듯하다. 이제 인공지능은 거부나 부정의 대상이 아니라 우리의 일상적 삶을 풍요롭게 만들어 줄 첨단기술의 집합체로 현실적인 인정을 받아야 할 때다. 좋은 싫든 인공지능은 앞으로도 인간과 함께 살 수밖에 없는 삶의 동반자임이 분명해졌다.

IV. 인공지능의 시대, AI 기반 불교 윤리의 제안

인공지능의 발달이 던진 다양한 도전들은 불교학계에도 신선한 충격과 함께 학문적 자극이 되고 있다. 위에서 소개한 소랏 형라다툼 외에도 최근 피터 D. 허쇼크(Peter D. Hershock)은 이런 주제를 다룬 기념비적인 저서 『불교와 지능 기술: 보다 인간적인 미래를 위하여 (Buddhism and Intelligent Technology: Toward a More human Future)』(London: Bloomsbury Academic, 2021)에서 AI가 제기한 핵심적인 도전은 기술적인 것이 아니라 윤리적인 것이라고 진단한다. 따라서 우리는 이 행성에 사는 동안 다른 모든 존재와 번성(flourishing)을 공유하는 윤리를 재정립할 필요가 있다는 것이다. 그럴 때 비로소 우리는 기술이 주도한 AI의 눈부신 발달 속에 내재되어 있는 가치갈등이라는 어려운 문제를 해결할 수 있을 것이라고 주장한다. 허쇼크의 제안은 그가 ‘가상 관계 역학(virtuosic relational dynamics)’이라고 부르는 개념을 개발하자는 것이다. 책의 1장에서 이 아이디어는 “가상 자비 관계 윤리(an ethics omoassionate relational virtuosity)”라는 개념으로 더욱 구체화되고 있다. 이것은 개인주의를 넘어 우리 사회를 관계적으로 엮는 새로운 윤리적 접근이다. 허쇼크는 자신의 주장을 교학적으로 뒷받침하기 위해 불교철학이 제공할 수 있는 중요한 관점들을 두루 검토하고 또 논의한다. 그의 책은 처음부터 끝까지 불교는 지속가능한 AI 윤리를 추구하는데 필요한 통찰력과 수행법을 풍부하게 제공할 수 있다는 입장을 유지한다. 그가 제안하는 불교 AI 윤리는 서구의 개인주의를 극복함과 동시에 지속가능한 관계적 공존의 윤리다. 특히 허쇼크는 대승의 이타적 보살도 사상과 육바라밀에서 관계적 역학의 성립 가능성을 발견한다. 예컨대, 인공지능과 인간의 바람직한 관계설정은 육바라밀의 자리아타적 평등정신에서 배울 수 있다고 말한다. 그의 인식에 따르면 우리가 “인간적으로 된다”는 것은 곧 우리가 윤리적으로 생각하고 행동한다는 것과 사실상 동격이다. 그런 점에서 개인의 권리와 프라이버시를 너무 강조

하고 경쟁을 정당화하는 서구적 관념의 이기적 개인주의는 극복될 필요가 있다. 그런 차원에서 불교 고유의 연기설에 바탕을 둔 관계적 불교윤리의 확립이 절실한 상황이다. 어쩌면 허속은 자신의 ‘가상 관계 역학’ 개념의 연장선상에 서 있는 ‘가상 자비 관계 윤리’야말로 인공지능이 주도할 미래사회의 불교윤리로 자리매김할 수 있을 것이란 기대를 하고 있을지도 모르겠다.²⁰⁾

챗 GPT의 놀라운 답변은 인공지능이 드디어 인간의 뇌를 초월하기 시작했을지도 모른다는, 근원적인 공포심을 불러일으켰다. 이러다가는 인간과 기계가 정면으로 충돌하는 날이 올 수도 있겠다는 우려도 광범위하게 퍼졌다. 미래의 어느 날 AI 기계가 인간존재를 느러터지고 질척거리거나 하는 동물로 여긴다면 그리고 마침내 우리에게 남은 선택지가 인공지능에 복종하거나 거꾸로 제거의 대상이 되는 일만 남았다면... 이것은 정말 상상하기도 싫은 끔찍한 광경이 될 것이다. 일반인들의 우려가 커지자 과학자들에게 윤리적이고 정신적이면서도 인간적인 우주의 요소들을 기술적으로 통합하는 인공증강지능(artificially augmented intelligence)의 확대를 요구하는 움직임도 일어나고 있다. 이는 AI가 초래하는 인간존재의 위기는 결국 인공지능 시대에도 적용될 수 있는 새로운 윤리 규범의 공유와 실천이라는 인식과 같은 맥락으로 볼 수 있겠다.

그런 윤리의식을 갖춘 인공지능의 개발은 과학과 영성 사이를 지혜롭게 연결하는 인문적 능력을 요구한다. 우리는 불교가 바로 그런 다리의 역할을 능숙하게 해낼 수 있다고 믿는다. 처음부터 불교는 자연세계와 정신세계에 이어 인간세계를 서로 융합할 수 있는 독특한 이론 체계를 갖추고 있었다. 이를 가능하게 만든 원리가 바로 붓다의 연기설이다. 연기의 가르침은 한편으로는 과학적으로 입증된 원인과 결과의 패러다임에 바탕을 두고 실재를 설명하며, 다른 한편으로는 그 실재를 자비로운 휴머니즘(compassionate humanism)의 대상으로 해석한다. 여기서 우리는 지혜와 자비가 동시에 작동하고 있는 실재의 현실을 있는 그대로 경험하게 된다. 인공지능기술이 아무리 갑작스러운 현상으로 보여도 이면에는 그동안 쌓인 무수한 인연의 실타래가 종횡무진으로 얽혀 있다. 우리가 보기에 인공지능에 없는 ‘인간적-정신적 차원(human-spiritual dimension)’은 불교에서 말하는 ‘자기-깨달음(self-enlightenment)’의 관념과 연결된 ‘인과(causally)’의 개념으로 얼마든지 보상될 수 있다.

이런 이념을 구현한 불교적 AI를 설계한다면 그것은 과학과 영성을 결합한 고통치유형 인공지능이 될 것이다. 한마디로 압축하면 지혜를 갖춘 자비로운 인공지능이라는 말이다. 합리적인 사고와 자비적 행동은 불교 AI의 가장 큰 특징이 될 것으로 본다. 과학과 영성은 지혜와 자비의 서구 버전에 해당된다. 지텐드라 우탐(Jitendra Uttam)은 지성과 감성을 겸비한 자비

20) 이 부분은 Theodoor Richard & Peter D. Hershock, “On The Ethical Challenge of the Predicament of AI and the Perspectives Offered by Buddhism: A conversation with Peter D. Hershock”, *Exchanges: The Interdisciplinary Research Journal*, vol.9(2), 2022. PP.97-110을 임의로 요약, 정리한 것이다.

로운 AI의 설계와 유통이 일어날 가장 유력한 지역으로 대한민국을 지목한다. 오랜 불교전통과 빠른 인터넷망 및 과학기술의 수준을 고려할 때 그렇게 될 것으로 예측한다.²¹⁾ 인공지능의 시대에 역설적으로 연기설을 전제한 지혜와 자비의 불교윤리가 각광 받는 것은 어쩌면 당연한 귀결일지도 모르겠다. 지혜와 자비의 가르침은 시공간을 넘어 언제 어디서나 시대의 윤리로 거듭날 수 있는 교학적 유연성과 포괄성 및 탄력성을 지니고 있기 때문이다. 앞에서 불교윤리를 가리켜 현재 진행형이자 미래 능동형이라고 규정할 수 있었던 이유다. 인공지능은 지혜와 자비의 물로 관육식을 치를 필요가 있겠다는 비유를 던져본다.

그리고 여기서는 자세히 다를 여유가 없지만 인공지능의 깨달음 가능성, 즉 불성의 유무를 둘러싼 논쟁도 뜨겁다. 불교학계에서는 일찍이 보일 스님이 AI에게도 불성이 있는가의 문제를 진지하게 다룬 바 있다.²²⁾ 그는 앞서 소개한 『AI 부디즘』에서 다양한 관점에서 인공지능에도 불성이 있는가를 질문하고 있다. 업과 윤회, 사랑, 살인, 인공감정 등과 같은 잣대로 접근해서 설득력 있는 해석을 보여준다. 그런 점에서 보일 스님은 이 분야를 선도적으로 개척하고 있는 수행자이자 불교학자이다. 발표문을 준비하면서 “모든 휴머노이드 로봇은 불성을 갖는가?”라는 주제로 쓴 짧은 논문을 읽은 적이 있다. 실제로 일본 고다이지(Kodaiji) 선 센터에서 신도들의 선 수행을 돕고 있는 로봇 승려 ‘민다르(Mindar)’를 소재로 쓴 것이어서 더 재밌었던 것 같다. 민다르는 관음보살의 인격화를 상징한다. 대승경전에서 “무정물도 불성을 가지고 있으며, 따라서 무정물도 붓다가 될 수 있다.”는 명제는 그냥 비유로만 받아들여 할지 민다르와 같은 무정물에도 적용할 수 있을지는 앞으로도 계속 논쟁거리가 될 것이다.²³⁾

소랏 형라다룸이 말했듯이 인공지능이 오온을 갖추고 의식을 갖는다면 불성이 있다고 해야 할 것 같기도 하고, 아무리 성능이 뛰어나도 인공지능은 영원히 인간의 피조물일 수밖에 없다면 불성을 인정하기는 어렵지 않겠느냐는 갑론을박이 있을 수 있을 것 같다. 앞으로 인공지능이 인간의 지성과 감성을 모방하는 능력이 더욱 향상되고 언젠가 인간을 넘어섰다는 판단이 설 경우 인공지능 로봇의 불성 문제는 다시 논란의 중심으로 등장하겠지만 아직은 전혀 걱정할 문제가 아닌 것 같다. 이런 문제에 관심이 있다면 붓다를 깨달음은 지능으로 상정하고 -SF 영화에서 그와 같은 능력을 보여주는 인공지능 로봇 사만다(Samantha)와 비교하면서- 설명하는 논문도 읽어 보면 도움이 될 것 같다.²⁴⁾

21) 위의 내용은 Jitendra Uttam, “Between Buddhist ‘Self-Enlightenment’ and ‘Artificial Intelligence’: South Korea Emerging as a New Balancer”, *Religions* 2023, 14, 150. <https://doi.org/10.3390/rel14020150>. 검색 일자, 2023년 06월 26일.

22) 보일 지음, 『AI 부디즘』, 앞의 책 참조.

23) Gereon KOPF, “Does AI have Buddha-Nature? Reflections on the Metaphysical, Soteriological, and Ethical Dimensions of Including Humanoid Robots in Religious Rituals from one Mahāyāna Buddhist Perspective”, M. Norskov et al(Eds.), *Culturally Sustainable Social Robotics*(IOS Press, 2020), pp.596-600.

24) Neela Bhattacharya Saxena, “AI as Awakened Intelligence: Buddha, Kurzweil and the Film *Her*”, *Theology and Science*, <https://www.tanfonline.com/loi/rtas20>. 검색 일자 2023년 6월 26일자.

V. 맺는 말

챗 GPT의 돌풍이 거세게 불고 있지만 아직은 인간이 걱정해야 할 만큼 위험한 인공지능이 출현하지는 않았다. 오늘날 인간의 정체성과 위상이 인공지능을 비롯한 첨단 디지털 기술로부터 도전을 받는 것은 사실이지만, 그렇다고 지금 당장 혹은 아주 가까운 미래에 인간성 자체가 거부되거나 인공지능으로 대체되는 재앙 혹은 종말은 일어나지 않을 것으로 본다. 기본적으로 인간은 사회적 동물이면서 정치적 동물이고 윤리적인 동물이기 때문이다. 이 말은 같은 것을 다르게 표현한 것에 불과하다. 협력과 공존이 최고의 생존전략이라는 것을 본능적으로 안다. 호모 사피엔스 종은 아득하게 먼 진화의 역사를 통해 '생존과 번성'이라는 존재 목적을 충실하게 쫓으면서 끊임없이 적응하고 마침내 지구의 최적자(最適者)가 되었다. 과학을 부정할 수 없다면 인간은 지금도 어떤 형태로든 진화의 과정을 겪고 있다고 봐야 한다. 인공지능과 같은 전혀 차원이 다른 기술이 발달하면서 인간의 진화는 아예 멈춰 섰다고 생각하는 사람도 혹시 있을지 모르지만, 인간 종 호모사피엔스는 어떤 형태로든 진화하는 능력을 진화시켜 왔다.

인지과학과 로봇공학을 연구하는 김상균 교수에 의하면 우리는 지금 인간 스스로 변화하는 '인공진화(artificial evolution)'를 주도하고 있다고 한다. '진화'와 '인공'은 부자연스럽지만, 아무튼 그는 지루한 자연선택을 기다리다 지친 인간은 자신의 선호와 목표에 맞춰 스스로 진화하기를 선택했다고 비유한다. 인공지능과 같은 첨단기술이 이를 부추기고 있다는 것이 거의 주장이다. 그는 이를 육체의 확장을 위한 기술과 정신의 확장을 위한 기술로 나누어 분석하고 있다. 흥미롭지만 논쟁을 유발할 요소도 적지 않다.²⁵⁾ 그가 진화생물학자들에게 인공진화를 어떻게 설명할지가 궁금해진다.

앞에서 우리는 불교의 지혜와 자비의 윤리가 인공지능 시대의 고통을 치유할 유력한 대안이 될 수 있다는 다양한 관점들을 소개했다. 그런데 이런 불교적 공존의 지혜는 2019년 출범한 '스탠포드 인간중심 인공지능 연구소(the Stanford Institute for Human-Centered Artificial Intelligence, HAI)'나 하버드 대학의 '임베디드 에티クス 프로그램(Imbedded Ethics Program)'의 설립 취지와 연구 철학과도 일맥상통하고 있다. 무섭도록 강력한 AI가 인간과 아름다운 공존의 길을 걸을 수 있도록 끊임없이 고민하는 그들은 AI의 기술개발 초기 단계부터 종교와 법학, 윤리 학자들을 공학자와 엔지니어들의 작업에 동참시킨다. 컴퓨터 공학자와 엔지니어들이 종교와 윤리적 시각을 접목한 인공지능 기술을 개발한다는 발상 자체가 인간과 기계의 공존뿐만 아니라 더 나아가 지구상 모든 생명들의 공동·번성을 추구하기 위한 것이다.²⁶⁾ 불교의 영향이 그들에게 얼마나 미치고 있는지는 모르겠지만, 근본적인 취지에서 지혜

25) 김상균, 『초인류』 앞의 책, pp.09-19, pp.347-351 등.

26) 유발 하라리 지음, 전병근 옮김, 『21세기를 위한 21가지 제언-더 나은 오늘은 어떻게 가능한가』(경기 파주:김영사, 2018)에도 인공지능의 등장과 관련된 다양한 논의들이 소개되고 있다. 책의 끝부분에서 하

와 자비의 윤리는 과학과 도덕을 동시에 충족시킬 수 있는 가르침임에 틀림없다. 불교는 어제에 이어 오늘도 그리고 오늘을 지나 내일도 변함없이 인간과 기계와 그 외의 다른 모든 존재들과의 공존·공영을 추구하고 이념으로 계속 남아 있을 것으로 확신한다. 불교는 먼 옛날의 전설이 아니라 지금 이곳의 서사이자 앞으로 다가올 미래에도 쉬지 않고 들려져야 할 고상한 음악이다. 인간의 미래도 불교의 미래도 여전히 밝다는 사실을 함께 공유할 것을 제안드리고 싶다. 오늘 행사를 계기로 우리 스스로 인간임을 또한 불자임을 자랑스럽게 여길 수 있었으면 좋겠다는 말씀을 드린다. 경청해 주셔서 감사합니다.

라리가 불교를 언급하고 있는 부분도 눈길을 끈다. 내 마음을 보는 것만으로는 아무것도 바뀌지 않는다는 그의 말은 따가운 비판이 아니라 따뜻한 관심으로 들렸다.

- 김상균 지음, 『초인류-AI와 함께 인공지능화에 접어든 인류의 미래』(경기 파주:웅진 지식하우스, 2023)
 윤송이 외 지음, 『가장 인간적인 미래』(서울:웨일북, 2022)
 스티어트 러셀, 이한음 옮김, 『어떻게 인간과 공존하는 인공지능을 만들 것인가-AI와 통제 문제』(경기 파주:김영사, 2022)
 유발 하라리 지음, 전병근 옮김, 『21세기를 위한 21가지 제언-더 나은 오늘은 어떻게 가능한가』(경기 파주:김영사, 2018)
 보일, “챗 GPT의 등장과 불교계의 대응방안, 『불교평론』(2023년 여름호); 보일 지음, 『AI 부디즘』(서울:담앤북스, 2021)
 김성옥, 이관수, “인공지능 ‘자유지’ 논의에 대한 선결조건-〈엑스마키나(Ex Machina)〉를 바라보는 불교와 과학의 시선”, 『불교학연구』(제70호, 2022)
 Cover Story, “생성 AI의 시대”, MIT Technology Review Korea, vol.8. May/June 2023.
 이상욱, “AI는 인간을 꿈꿀 수 있을까”, Skeptic Korea, vol.34(2023.06).
 David Rotman, “챗 GPT가 불러올 경제혁명, 공동번영으로 이어질까”, MIT Technology Review Korea.
 양성철(보일), “인공지능 챗봇에 대한 선문답 알고리즘의 데이터-심리치료 상담채봇을 중심으로-”, 『한국불교학』(제84집, 2017)
 보일(양성철), “디지털 휴먼에 대한 불교적 관점-악업의 증장인가, 선교방편인가-”『宗教研究』(제7집, 2022)
 박소영, “인간과 인공지능의 공존 가능성에 대한 탐색:책임의 윤리와 문학적 상상력”, 『윤리연구』(제124호, 2019).
 박소영, “인간과 인공지능의 공존 가능성에 대한 탐색:책임의 윤리와 문학적 상상력”, 『윤리연구』(제124호, 2019).
 한성자, “인공지능(AI) 로봇의 해탈 가능성-‘인간은 생각하는 기계인가’에 대한 불교적 관점-”, 『한국불교학』(제85집, 2018).
 이광형, “[이광형의 퍼스펙티브] 휴머니즘 기반해 인간·AI 공존하는 질서 만들어야”, 중앙일보, 2022년, 8월 15일 자 기사.
 Hongladarom, Soraj. The Ethics of AI and Robotics: A Buddhist Viewpoint(London: The Rowman&Littlefield Group, Inc, 2020)
 KOPF, Gereon. “Does AI have Buddha-Nature? Reflections on the Metaphysical, Soteriological, and Ethical Dimensions of Including Humanoid Robots in Religious Rituals from one Mahāyāna Buddhist Perspective”, M. Norskov et al(Eds.), Culturally Sustainable Social Robotics(IOS Press, 2020), pp.596-600.
 Richard, Theodoor & Hershock, Peter D. “On The Ethical Challenge of the Predicament of AI and the Perspectives Offered by Buddhism: A coversation with Peter D. Hershock”, Exchanges: The Interdisciplinary Research Journal, vol.9(2), 2022. PP.97-110을 임의로 요약, 정리한 것이다.
 Jeste, Dilip V., Graham, Sarah A., Nguyen, Tanya T., Ellen, E. Colin, A. Depp, Lee, and Ho-Cheol Kim, “Beyond artifickal intelligence:exploring artificial wisdom”, International Psychogeriatrics, vol.32(8), pp.993-1001.
 Stanford Graduate School of business, “Building Compassionate AI: Why Compassion Matters For Artificial Intelligence Design and Deployment”, Case: M-386,

- Date:08/14/2020,
- Boni, Michal. "The ethical dimension of human-artificial intelligence collaboration", European view, vol.20(2), 2022. p.182.
- Hershock, Peter D. Buddhism and Intelligent Technology: Toward a More human Future(London: Bloomsbury Academic, 2021)
- Lee, Kai-Fu. "A Human Blueprint for AI Coexistence", in J. von Braun et al.(eds), Robotics, AI and Humanity, https://doi.org/10.1007/978-3-030-54173-6_22. 검색 일자 2023년 6월 25일.
- Uttam, Jitendra. "Between Buddhist 'Self-Enlightenment' and 'Artificial Intelligence': South Korea Emerging as a New Balncer", Religions 2023, 14, 150. <https://doi.org/10.3390/rel14020150>. 검색 일자, 2023년 06월 26일.
- Saxena, Neela Bhattacharya. "AI as Awakened Intelligence: Buddha, Kurzweil and the Film Her", Theology and Science, <https://www.tanfonline.com/loi/rtas20>. 검색 일자 2023년 6월 26일자.
- <https://www.youtube.com/watch?v=wTI-klquHRA>, 검색 일자 2023년 6월 25일. 최예진 교수의 TED 강연자료.
- <https://www.youtube.com/playlist?list=PLMd4CV6iDBunlnuiYS7kDpllB1Ub9qRl3>, 검색 일자 2023년 6월 25일. 최재천 교수의 강연자료.

「인간의 미래, 불교의 미래
- 인공지능의 발전과 자비윤리의 요청 -」에 대한 논평문

보일스님 / 해인사승가대학

I. 들어가며

미래를 상징하는 단어 ‘인공지능’이 이제 현재형이 되어 우리 곁에 다가왔습니다. 사실 우리는 부지불식간에 이미 인공지능 시대에 살고 있었는지도 모릅니다. 매일 사용하는 스마트폰에 탑재된 수많은 애플리케이션 이면에는 빅데이터를 운용하는 빅테크 기업들의 초거대 AI가 자리하고 있기 때문입니다. 이제 인공지능이 일상화되는 시대가 되었습니다. 최근 챗GPT로 인해 누구나 일정 요금을 지불하고 인공지능에 자유롭게 접근하고 사용할 수 있습니다. 인류가 처음으로 경험하는 인공지능의 상용화 시대에 우리 사부대중이 어디로 가야할 지 혹은 누구에게 길을 물어야 할지 혼란과 두려움이 교차하는 요즘입니다.

챗GPT를 비롯한 생성형 인공지능의 시대를 맞이하여, 불교적 관점에서 시대를 읽는 이해의 틀을 제공하는 작업이 절실한 시점에 시의적절한 연구를 발표해 주셔서 반갑고, 감사한 마음을 우선 전합니다. 또한 이러한 연구를 발표할 수 있도록 발표자님의 표현대로 ‘야단법석’을 마련해 주셔서 반야암 사부대중 여러분께 깊은 감사를 드립니다. 덕분에 불교 윤리의 방향과 내용에 대해서 항상 고민하시고 연구에 전념하고 계신, 평소 존경하는 허남결 교수님의 글을 가장 먼저 받아 읽게 되는 호사와 영광을 누릴 수 있어서 감사합니다. 오늘 제가 그간의 연구를 발표하고 허남결 교수님의 점검을 받아야 마땅한데, 개인 사정 때문에 역할이 뒤바뀌게 되어 대단히 송구스럽습니다. 그럼에도 불구하고 오늘의 제 소임이 논평인지라 몇 마디 사족을 달아보도록 하겠습니다. 발표문 전반에 걸쳐 인공지능과 관련한 매우 중요하고 흥미로운 주제들로 가득합니다. 이미 공유되고 공감되는 주제들도 있지만 처음 접하는 생소한 내용들도 있어서 많은 공부가 되는 논문이라고 생각합니다.

II. 논지의 정리

이 논문은 생성형 인공지능 시대를 살아가기 위한, 인간과 기계의 평화로운 공존의 지혜를 모색하는 과정에서 불교의 자비와 지혜가 인공지능 윤리의 대안으로 위상을 정립할 수 있는지 다양한 관점을 소개하고 있다.

제2장에서는 ‘인공지능과 인간 윤리의 공존-휴머니즘의 재발견’을 논하고 있다. 발표자는 인공지능과 인간 윤리의 공존을 위해, 진화생물학으로부터 협력과 공존의 지혜를 인간과 기계의 경쟁관계에서도 그대로 적용될 수 있다(p.3)고 주장한다. 그 공존 가능성을 “소박한 휴머니즘의 실현”으로 표현하고 있다. AI의 발전은 위협이라기보다는 물질적으로 더 풍족하고 정신적으로 더 여유 있는 공동체의 건설을 통해 공감대의 형성이 중요하다고 본다. 발표자는 미찰 보니(Michal Boni)의 연구를 인용하면서 “AI 발전이 가져올 결과로써 인간과 공존가능성을 높여준다고 보고 인간의 이익에 봉사하는 윤리적 AI의 개발” 필요성을 제시한다. 발표자는 미찰보니의 견해 또한 인간과 인공지능의 가능성을 결국 인간과 인간 사이의 공존 가능성의 문제로 귀착된다고 본다. 이를 통해 발표자는 인공지능에 대한 비판보다는 인간 윤리의 재점검이 실질적인 해결책 마련에 도움이 될 것이라고 주장한다.

제3장에서는 인공지능과 불교 윤리의 접점을 논하면서 구체적 사례로 자비로운 인공지능 로봇을 제시한다. 발표자는 소랏 형라다롬의 『불교의 시각에서 본 AI와 로봇 윤리(The Ethics of AI and Robotics: A Buddhist Viewpoint)』(2021)에서 형라다롬의 지혜와 자비를 갖춘 불교적 AI를 제안 소개하는데, 형라다롬은 먼저 AI의 기능이 인간의 자율성이나 개인의 권리를 침해하도록 내버려 두서는 안 된다는 입장에서 불교의 ‘지혜와 자비’를 통해 인공지능 윤리의 정립 필요성을 역설한다. 형라다롬은 그 전제 요건으로서 ‘기술적 탁월성(technological excellence)’ 과 ‘윤리적 탁월성(ethical excellence)’을 제시한다. 또한 형라다롬은 인공지능에 대한 인격성 부여 기준으로 불교에서 말하는 인간의 다섯 가지 구성요소, 즉 색(色), 수(受), 상(想), 행(行), 식(識)의 조건을 갖춘다면 인공지능도 인격체로 수용되어야 한다고 주장한다. 오온설(五蘊說)의 확장을 통해 인공지능이나 초지능 로봇도 인간과 다름없는 인격체가 될 수 있는 불교적 근거를 제공하고 있는 셈이다. (p.5) 이외에도 발표자는 인공지능의 깨달음, 특이점, 자율무기 시스템, 반려로봇 등 인공지능의 윤리가 쟁점이 될 만한 다양한 논의를 친절하게 소개하고 있다. 발표자는 여기에 더해 불교 윤리적 인공지능의 윤리적 성격에 대해서 고찰하고 있다. 이를 통해 발표자는 인공지능은 거부나 부정의 대상이 아닌 인간 삶의 동반자가 될 수 있음을 강조하고 있다.

제4장에서 발표자는 AI 기반 불교 윤리를 제안하고 있다. 이를 위해 발표자는 피터 D. 허속(Peter D. Hershock)의 『불교와 지능 기술:보다 인간적인 미래를 위하여(Buddhism and

Intelligent Technology: Toward a More human Future』(London: Bloomsbury Academic, 2021)을 소개하면서 AI가 제기한 핵심적인 도전은 기술적인 것이 아니라 윤리적인 것이라는 그의 주장을 인용한다. (p.9) 인용에 따르면, 허속은 가상 관계 역학(virtuosic relational dynamics)라는 개념을 구체화한 “가상 자비 관계 윤리”를 통해 개인주의를 넘어 사회를 관계적으로 엮는 윤리적 접근을 시도하는데, 불교 고유의 연기설에 바탕을 둔 관계적 불교 윤리 확립의 이론적 기반이 될 수 있음을 시사한다. 발표자는 허속의 이러한 관점이 불교철학이 제공하는 다양한 관점들이 지속 가능한 AI 윤리를 추구하는 데 필요한 통찰력과 수행법을 풍부하게 제공할 수 있다고 보고 있다. 이 논문 전반에 걸쳐서 필자가 주목하는 것은 인공지능 시대에 적용할 수 있는 새로운 윤리 규범의 공유와 실천이며, 그 대안으로서 불교철학이 제공하는 다양한 통찰의 접목 가능성으로 보인다. 발표자는 “불교가 바로 그런 다리의 역할을 능숙하게 해낼 수 있다고 믿는다. 처음부터 불교는 자연 세계와 정신세계에 이어 인간 세계를 서로 융합할 수 있는 독특한 이론체계를 갖추고 있었다”고 주장한다. 이어서 붓다의 연기설에 기반하여 실재를 설명하고 그 실재를 자비로운 휴머니즘의 대상으로 해석할 수 있는 불교적 AI를 설계한다면 과학과 영성을 결합한 고통 치유형 인공지능이 될 것이라고 예상한다.

제5장 결론에서 발표자는 인간이 자연선택을 기다리다 지쳐 스스로 진화를 선택한다는 ‘인공진화(artificial evolution)’ 개념을 소개하면서 향후 전개될 인간의 정신과 신체 능력의 확장을 통한 진화에 주목하면서, 불교가 인공지능 시대에도 못 생명들과 공존, 공영을 추구하고 이념으로 계속 남아 있을 것으로 확신하면서 마무리한다. 이처럼 필자는 논문 전반에 걸쳐 불교의 지혜와 자비의 윤리가 인공지능 시대의 고통을 치유할 유력한 대안이 될 수 있다는 다양한 관점들을 소개하면서 자신의 불교와 인공지능 시대의 전망을 담아내고 있다. 이는 결국 AI가 초래하는 인간존재의 위기는 결국 인공지능 시대에도 적용될 수 있는 새로운 윤리 규범의 공유와 실천이라는 인식과 같은 맥락으로 볼 수 있겠다.

Ⅲ 몇 가지 문제의식들

이상의 내용을 토대로, 발표자님께 몇 가지 질문을 드려볼까 합니다. 첫번째는 가볍거나 흥미로운 주제이고 다음 두 가지 질문은 발표문 전반에 걸친 발표자의 문제의식에 대한 이견일 수도 있습니다. 일단 발표자님께서 주장하시는 인간과 AI의 공존을 위한 윤리 규범의 공유와 실천에는 해결해야 할 문제들이 많은 것 같습니다.

- 질문 1) 발표자께서는 ‘인공일반지능(AGI)’의 등장(p.5)에 대해 언급하셨습니다. 향후 머지

않은 미래에 인공지능의 등장을 예상하는 것은 이제 어렵지 않습니다. 그래서 전 세계적으로 이에 대응한 미국과 유럽연합 등 각국에서 규제 입법 마련이 진행되고 있습니다. 예를 들어, 유럽연합이 발의한 [유럽연합 인공지능법(EU AI ACT)]¹⁾은 AI의 적용을 세 가지 위험 범주로 할당합니다. 첫째 중국에서 사용되는 유형의 정부 운영 소셜 스코어링 같이 허용할 수 없는 위험을 생성하는 애플리케이션 및 시스템의 금지, 둘째 취업 지원자의 순위를 매기는 이력서 스캔 도구와 같은 고위험 응용 프로그램에서의 특정 법적 요구 사항 적용 셋째, 명시적으로 금지되거나 고위험으로 분류되지 않은 애플리케이션입니다. 여기서 “허용할 수 없는 위험”이나 “고위험 응용 프로그램”에는 인간 본연의 영적 성장과 연관되어 종교적으로 수용하기 어려운 분야도 해당할 것입니다. 구체적으로 예를 들자면, 불교적 관점에서 인간의 참선 혹은 명상 상태의 다양한 두뇌 정보 데이터를 인공지능의 데이터로 활용하거나 적용한다면 그 결과는 상상하기 어렵습니다. 불교 윤리적 관점에서 이와 같은 시도를 용인해야 할지 발표자님의 고견을 구합니다.

- 질문 2) 발표자께서는 논문 말미에 “인공지능 시대의 휴머니즘을 시대적 화두로” 삼자는 견해를 밝히시는데 과연 다가오는 인공지능 시대에 “휴머니즘”이라는 기존의 사유 혹은 분석틀로써 미래의 변화상을 담아낼 수 있을지 회의적인 시각이 있습니다. 휴머니즘이 낳은 인간중심주의는 근대 이래로 인류를 세계대전과 환경파괴, 인종차별, 동물 학대, 양극화 등등 비극적인 결과를 낳았고, 지금도 달라지기보다는 ‘인류세(Anthropocene)’로 대변되는 기후변화, 전쟁과 전염병 속에서 암울한 현실을 마주하고 있습니다. 그래서 최근 트랜스휴머니즘과 포스트휴머니즘 담론을 통해 인간중심주의를 비판하면서 휴머니즘에 대한 재고를 요청하고 있습니다. 과연 휴머니즘 혹은 휴머니즘의 재해석이라는 틀 안에서 여전히 인간과 생성형 AI의 공존이 가능하다고 보시는지요? 발표자님께서 인공지능을 비롯한 생명공학 기술, 나노기술 등의 비약적 혁신으로 인해 새로운 인간종의 출현도 예상됩니다. 과연 현재의 휴머니즘이라는 이념 틀 속에서 미래의 기계와 인간, 비인간 생명체, 디지털 데이터가 뒤섞이는 혼종적인 세계를 지탱할 수 있을지 의문입니다. 이 부분에 대한 발표자님의 견해를 듣고 싶습니다.
- 질문 3) 발표자께서 인용하신 소랏 형라다룸에 따르면 인공지능 기반의 자율무기 시스템에 대해서 “기술적 탁월성”과 “윤리적 탁월성”이라는 기준을 통해 “전쟁을 피할 수 없는 상황에서 드론과 같은 자율무기 시스템의 사용은 정당화되어야 마땅하다. 다만 자율적 무기 시스템은 그 자체의 윤리적 고려사항으로 적절한 목표를 선별할 수 있는 능력을 갖추지 않으면 안 된다. 불교적으로 좋은 자율무기 시스템은 아군과 적군뿐만 아니라 전투원과 비전투

1) <https://artificialintelligenceact.eu/>

원을 구별할 수 있어야 할 것이다. 그럴 능력을 갖추지 못했다면 이 무기 시스템은 윤리적인 탁월성에 앞서 기술적인 탁월성을 획득하는 데 실패했다는 평가를 받게 된다.”(p.7)는 입장을 취합니다. 결국 이 기준을 충족한다면 인공지능 자율무기 시스템을 사용할 수 있다는 것입니다. 하지만 이에 대해서 이러한 유형의 이른바 ‘킬러로봇’ 사용 자체가 비윤리적이라는 주장도 강력하게 대두됩니다. 예를 들어, 만약 이 시스템에 의해 비무장 민간인에 대한 공격과 살상이 발생한 경우 누가 책임을 질 것인가라는 문제가 제기됩니다. [전제1]. 책임질 수 있음은 교전법의 선제조건이다. [전제2]. 전쟁에서 군사 로봇을 사용할 때 그리고 그것이 해로운 결과를 야기했을 때, 책임을 질 수 있는 주체는 셋 중 하나이다. 바로 군사 로봇의 설계자, 지휘관, 그리고 로봇 자신이다. [전제3]. 세 후보 가운데 어느 쪽에도 군사 로봇이 발생시킨 해로운 결과에 대해 온전한 법적, 혹은 윤리적 책임을 물을 수 없다. 결론. 따라서, 자율적 군사 로봇의 사용은 비윤리적이다.’²⁾ 이 논증은 스페로의 ‘트릴레마 논변’입니다. 결국 인공지능 기반 자율 무기 시스템이 초래한 결과에 대해 누구에게도 온전히 책임을 물을 수 없다면 그 킬러 로봇의 사용 자체가 비윤리적이라는 것입니다. 따라서 헝라다름이 주장처럼 이미 인공지능 자율 무기 시스템의 도입을 전제로 윤리적 탁월성을 요구할 것이 아니라, 그 도입 자체를 금지해야 한다는 의견에 대해서 발표자님은 어떻게 생각하시는지 궁금합니다.

IV. 마무리하면서

이번 발표문을 통해 인공지능에 대한 새로운 사실들을 접하게 되었고 생각해 보게 되었습니다. 오늘은 발표자님께서 인공지능에 낯설어하는 대중들을 위해 인공지능 시대에 고민해 봐야 할 여러 쟁점을 총론적으로 소개해 주셨다면, 다음 기회에는 오늘 말씀해주신 각 쟁점을 하나하나 독립적으로 다뤄주실 것으로 기대하겠습니다. 심도 있는 리서치와 안목으로 불교 윤리적 관점에서 불자들이 이 인공지능 시대를 어떻게 이해하고 실천해야 할 지에 대한 통찰을 제공해 주셔서 다시 한번 발표자님께 깊은 감사를 전합니다. 논평문에 언급된 모호함이나 미처 짚어내지 못한 부분은 논평자의 이해 부족에서 오는 것이니, 따끔한 경책과 자비로 섭수하시기를 청합니다.

- 보일 합장 ()

2) 천현득(2019), 22-24, “‘킬러로봇’을 넘어: 자율적 군사로봇의 윤리적 문제들”, 『탈경계인문학 Trans-Humanities』, 12(1): 5-31. Web.