서평: 정대현 2026, 『로봇 의식론』. 커뮤니케이션북스 출판사

로봇의식의 윤리화를 위한 선제적 인문학

서평자: 최종덕(philonatu.com)

언어철학에서 삶과 사유의 철학으로 확장한 철학자 정대현 교수님의 새 책 『로봇 의식론』 (2026)은 이런 말로 시작된다.

“사람은 로봇에 대해 선제적 인문주체성을 확립할 필요가 있다”

『로봇 의식론』의 저자 정대현은 데카르트의 “나는 생각한다. 나는 존재한다” 라는 유명한 명제에 빗대어 “로봇도 생각한다. 그래서 로봇도 존재한다”라는 명제가 태어날 수 있음을 눈 여겨 본다. 고전 이진법 컴퓨팅 시대를 벗어나 양자컴퓨팅의 인공지능 시대가 현실화된다면 그런 시대의 로봇은 인간에게 정말 위협적 존재가 될 것이다. 그래서 저자는 사람이 로봇의 지배를 받지 않기 위하여 로봇의 ‘윤리화’를 주장한다. 미래 로봇의식의 윤리화를 위하여 선제적 인문학이 필요하다고 저자는 분명하게 말한다.

이 책은 먼저 인간과 같은 로봇의식이 가능한지를 따지는 부정론과 긍정론을 서술한다. 인간과 같은 주체성을 갖는 로봇이 나오기 에는 아직 이르거나 아예 불가능하다는 로봇의식 부정론은 대체로 ‘생각’은 인간만의

본질이라는 인식에서 출발한다. 이런 본질론적 인식은 계산이나 기억 혹은 번역 등 마음의 객관적 활동을 구현할 수 있겠지만 컴퓨터가 아무리 발전한다고 해도 '좋아하고 싫어하고 후회하거나 이해한다' 등의 주관적 활동을 구현할 수 없다는 점을 함의한다.

의식은 조각조각 생각(추론과 감정)들의 모음인 국소성이 아니라 총체적 단일성의 성격이라고 한다. 그래서 국소성의 컴퓨터는 아무리 강한 계산 능력을 가져도 의식을 결코 가질 수 없다는 것이 로봇의식 부정론의 요점이라는 것이다.

반면 저자는 로봇 인공지능 의식을 긍정적으로 대처해야 한다고 주장한다. 그 이유는 로봇의식의 현실을 피할 수 없으며 그런 현실에 대처하는 철학적 탐구가 긴급하기 때문이라고 한다. 로봇의식의 긍정 주장은 미래 과학, 특히 양자물리학의 기술적 변화와 같은 맥과 궤에 연관될 것이라고 저자는 예측한다.

그렇다면 로봇의식의 긍정론은 인간의식을 과학적 언어로 어디까지 기술할 수 있는지의 문제와 연관될 것이다. 저자가 뜻하려는 과학적 언어란 기존 고전과학의 분화된 언어구조를 탈피하여 통합적 언어를 지향하는 듯하다. 예를 들어 '개체'objects 지향 분석언어가 아닌 '상태'states 지향 통합 언어를 요청하며, 배타적 모순논리가 아닌 얽힘의 상보논리를 수용하는 그런 언어지향이다.

저자의 더 중요한 암시적 명제가 있다. 인간이 로봇을 윤리화한다는 것은 인간과 로봇의 공동체로서 윤리적 기반을 가져야 한다는 뜻과 연관된다는 점이다. 서평자에게는 이 명제가 상당히 강한 주장으로 여겨지면서도 동

시에 매우 매력적으로 느껴진다.

얽힘의 상보논리가 로봇에 정착된다면 이론적이나마 로봇의식도 인간과 같은 단일성과 총체성을 가질 수 있을 것이다. 그런데 바로 이 명제는 꼼꼼히 다시 읽을 필요가 있다. 인간의식의 절대 기준이 존재할까? 자연종 인간의 기준이 무엇일까, 즉 현대 인간이 정말 순수하고 절대적인 자연종의 본질적 기준을 답습하고 있는 것인지 잘 모르겠다.

이런 서평자의 의문과 질문은 저자의 이 책을 읽어가면서 자동적으로 생겼다. 저자는 인간의식과 로봇의식 사이의 이분법적 구획을 벗어나려 한다. 이 둘을 이분법으로 구획하여 고유한 범주로 고착시키면 오히려 이 둘의 존재는 상호공존보다는 모순배제로 될 것이라는 저자의 판단에 전적으로 동의한다. 인간의식을 정의하는 엄격하거나 고착된 기준을 피해야 하는 동시에 로봇의식을 정의하는 국소적이고 기계적인 기준도 피해야 한다고 서평자는 이해했다. 인간중심으로 본 인간과 기계의 관습적 기준을 벗어나려는 것이 저자의 의도였다고 이해했다는 뜻이다. 저자의 의도가 서평자에게 상당히 유의미한 메시지로 다가왔기 때문이다.

예를 들어 저자는 의식철학 분야에서 영향력있는 차머스David Chalmers의 실재론을 은근히 강조한다. 저자는 이분법 언어의 함정에서 벗어나기 위해 차머스의 확장된 실재론을 끌어온다. “실재 아니면 비실재”라는 기존의 이분화된 존재론에서 벗어나 제 3의 실재 영역real reality을 추가해야 한다는 차머스의 삼분법 아이디어를 저자는 긍정적으로 흡수한다. 아바타 세계 혹은 고도의 컴퓨팅에 의한 가상 세계를 새로운 차원의 실재 양식modes of reality으로 인정해야 한다는 점이다. 차머스의 이야기는 가상도서관과 같은 컴퓨터 기반의 가상체도 새로운 객체 영역으로 인정해

야 하듯이 아바타 같은 가상체도 현실세계의 주체로 도입되는 새로운 존재 양식mode of existence으로 해석될 수 있음을 간접적으로 암시한다.

저자는 차머스의 확장된 실재론에 머물지 않고 이중국면론double aspects monism과 의식의 통전정보론integrated Information Theory 론을 제시하여 로봇의식과 인간의식의 공동체적 화해를 시도한다. 이 책에서 전개되고 있는 이중국면론과 통전론은 인공지능 혹은 피지컬 인공지능 연구분야에서 윤리 문제가 얼마나 중요한지를 우회적으로 알려준다. 그런데 매우 전문-학술적 논지여서 일반인들이 이해하기 쉽지 않을 듯하다. 그래서 서평자가 대신하여 간단히 서술하면 다음과 같다.

의식의 이중국면론double aspects monism은 스피노자의 속성론과 칼융의 중립적 실재론 그리고 물리학자 데이비드 보옴Davis Bohm의 암시적 질서론을 재구성한 이론이라고 저자는 설명한다. 스피노자 철학에서 정신과 물질은 자연 혹은 신이라는 단일실체가 우리에게 드러내 보이는 두 가지 속성이다. 사람이 알 수 없는 깊고 깊은 어떤 무한자(중립적 실체; Unus Mundus)로부터 의식의 심성mental과 물질의 물성physical이 비로소 共現(발현)된다고 한다. 그 둘은 하나의 뿌리에서 발현된 것이므로 항상 서로 얽혀있고 비국소적 관계일 수밖에 없다는 뜻이다.

의식의 통전정보론integrated Information Theory은 전일론holism에서 신비성을 제거한 일종의 통합 시스템론이다. 통합적 시스템에서 의식이란 통일된 방식으로 정보를 통합하는 능력이다.(35) 통전론에서 의식은 내재적으로 존재하는 현실이며 동시에 여러 현상적 요소들이 여러 차원에서 구조화되는 시스템이다. 의식은 국소적 요소들로 환원될 수 없는 통전성 integration을 갖는다는 뜻이다.

저자는 여기서 의문을 제기한다. 이중국면론과 통전정보론이 심성과 물성의 융합을 지향하기는 하지만 현상과의 간극이 분명히 있고 이런 틈을 메꾸지 않으면 로봇의식의 긍정론을 풀어갈 수 없다는 것이다. 그래서 저자는 스피노자의 무한 속성론을 다시 인용한다. 잘 알려져 있듯이 스피노자에서 속성은 무한대이지만 다만 인간에게 드러내 보여주는 것이 오로지 정신과 물질 두 개일 뿐이다.

속성의 무한대 확장이론은 서평자에게 매우 흥미로운데, 네트워크 관계론으로 이어질 수 있기 때문이다. 정신과 신체, 이 두 개만 연결되는 것이 아니라 실제로는 역사, 심리, 사회윤리 그리고 좁게는 감정, 초심리, 등등의 수많은 맥락점nodes들의 연결네트워크로서 의식이 작동되고 있다는 뜻이다. 저자는 연결네트워크의 관계공간을 “배위공간”이라고 했다. 신유물론 철학을 확산시킨 라투르Bruno Latour의 네트워크이론과 같은 맥락에 있다고 서평자는 이해한다.

로봇의식을 조명하려면 로봇에게 “자기성”selfhood가 있는지를 질문해야 한다고 말한다. 로봇의식은 물론 인간의식과 다르지만 종합적이고 연장적인 의미를 지닐 수 있다고 여겨진다. 자기성의 의미란 자기 지시적 의미self-referential meaning라고 말한다. 로봇이 자기 구성적 의식을 갖는다는 것은 자기 지식적 의미의 주체임을 뜻한다. 로봇의식은 인간주관성을 모방하는 이론이 아니라 새로운 유형의 종합주관성synthetic subjectivity을 지향한다고 저자는 강조한다.(51)

형이상학 관점에서 벗어나 좀 더 쉽게 말한다면 로봇이 ‘셀프’를 갖느냐의 문제는 행위성을 갖느냐의 문제로 볼 수 있다고 하는데, 이를 설명하기

위해 저자는 양자컴퓨팅의 탈이분법적 논리를 도입한다. 양자역학은 고전 물리학의 관찰의 객관성 기준과 달리 참여의 행위성을 강조한다. 양자론의 행위는 비결정론이 가능해지는 배위공간이며 외부 관계가 아닌 내부작용의 내적 관계이며 능동성이라고 저자는 말한다. 로봇의식은 그런 능동성 행위자로 될 수 있다는 점을 시사한다. 로봇은 로봇이기에 여전히 로봇 내부 구조와 내부 상태에 의해 인과적으로 설명가능해야 한다.(56)

설명가능한 내적 구조의 인과율을 따른다는 사실은 로봇의식에 대하여 자유의지 아니면 결정론이라는 이분법적 판단에서 탈피해야 한다는 혁신적 조건을 반드시 만족해야 한다. 그런 조건을 만족시키는 것이 바로 양자컴퓨팅일 수 있다고 저자는 말한다. 양자컴퓨팅 기반 로봇의식으로서 행위자는 존재-인식을 통합하는 구조이며 자연종과 인공종 사이의 탈이분법적 통전integrated 구조라는 데 있다.

통전 시스템으로서 타인의 마음 나아가 로봇의 마음을 탐구하려면 감각과 지각의 연관성을 되돌아 봐야 한다고 저자는 말한다. 저자가 보여준 사례는 다음과 같다. 바다 수평선 저 멀리 작은 점으로 보이던(감각sensation) 것이 해안가로 점점 가까이 오면서 그 점들이 어선, 여객선, 유람선으로 달리 지각perception된다.(71) 그러나 감각이 많아진다고 해서 지각이 결정된다고 말하기는 어렵다. 왜냐하면 지각에는 일차 감각 외에 기억, 경험, 기대감 등의 내적 요인들이 결합되기 때문이다. 마찬가지로 행위자의 행위 역시 요소들의 묶음으로 해명되지 않는다고 한다. 행위자 로봇이 자연종 인간과 동일하냐의 문제는 객체들 사이의 인과법칙이 아니라 “성향”의 관계네트워크 맥락점 사이의 인과론으로 설명되어야 한다고 저자는 강조한다. 앞서 자주 언급했듯이 이런 성향적 관계론을 저자는 존재론적 통전성integrationality라고 표현했다.

존재론적 통전성을 찾기 위하여 저자는 두 가지 길을 제시한다. 하나는 동양철학에서 말하는 음양의 통전성과 '誠' 개념이다. 그리고 다른 하나는 『과정과 실재』의 저자 화이트헤드(A.N.Whitehead 1861-1947)의 '합생'concrescence 개념이다. 저자는 화이트헤드의 합생을 통전이라고 번역했는데, 의미론적으로 딱 맞는 용어인 듯하다.(82)

책 후반부에서 저자는 화이트헤드의 과정철학을 양자역학적 의미로 풀이한 에퍼슨Michael Epperson의 존재론적 해석과 이스트만-키튼Eastman and Keeton의 양자론적 경험 개념을 도입한다. 저자는 양자역학의 관찰 개념이 인식론에 그치지 않고 존재론의 변혁이라고 보는데, 의식을 가진 관찰자는 세계 밖의 초연한 주체가 아니라 세계 안 내적 공명의 주체임을 주시한다. 에퍼슨과 이스트만-키튼에 의해서 관계론적 창발성relational emergence, 자연의 자기경험self-experience, 내적 공명inner resonance, 생성의 정동affection, 범경험론 등의 탈이분법적 관계론을 설명하는 부분은 저자의 이 책에서 중요한 의미를 차지한다고 서평자는 판단한다.

양자역학을 관계론적 창발성으로 본 키튼의 해석을 소개하면서 화이트헤드의 '조화 속 창조'harmony in creativity가 바로 자연의 자기경험 self-experience이라고 설명해주고 있다. 다시 말해서 경험은 인간만의 전유물이 아니라 자연물 혹은 로봇의식과 같은 인공물에도 가능한 지각능력이라고 말한다. 화이트헤드의 통전(합생;concrescence) 개념을 로봇의식과 연결시킨 저자의 의도는 성공적으로 보인다. 이렇게 저자는 양자론의식을 통하여 인간의식의 모방 없이도 로봇의식의 가능성을 매우 설득력 있게 타진한다.

이스트만T. E. Eastman은 자신의 책(Eastman and Keeton 2003, *Physics and Whitehead: Quantum, Process, and Experience*) 2장에서 “이원론 없는 상보적 이중성”Duality without Dualism의 의미를 꽤나 상세히 기술하고 있다. 이스트만의 상보적 이중성의 논지는 동양철학을 포함한 동양고전학 전반에 깔린 음양사상의 음양 관계 그 자체다. 동아시아 사람들은 전통의 음양 관계가 서양철학에서 말하는 배제의 이분법 논리가 아닌 상호 상보의 관계라는 것쯤은 웬만큼은 알고 있다. 이스트만이 양자역학의 도움을 받아 추론한 “이원론 아닌 상보적 이중성” 개념이 바로 동양의 음양 관계를 설명하는 데 딱 들어맞았기 때문에 화이트헤드의 양자론 철학으로 석사논문을 쓴 서평자에게 저자의 상보적 양자존재론은 큰 흥미를 끌었다.

사람들은 존재론적 내면성과 우주적 외연성의 통전(통합)적 차원에서 살게 되면서 “사람은 자신이 만든 시스템과 어떻게 더불어 살 수 있는가”를 물어야 한다는 것이 저자의 밀도있는 요청이다. “진화된 로봇은 사람의 손을 떠나 독자적 행위자로 될 날이 멀지 않기” 때문에 사람-로봇의 공존성은 단순한 당위적 이론이 아니라 당면한 존재-윤리일 수 있다는 것을 이 책을 통해서 배웠다.

공존이란 두 존재자가 동일해진다는 뜻이 아니라 다원적 역할의 잠재성이 확장된다는 뜻이다. 사람과 로봇의 공존을 인정함으로써 오히려 사람은 더 사람답게 살 수 있을지 모른다. 저자는 이런 공존의 철학적 시도를 그의 생애를 걸쳐 지속적으로 연구해 온 것으로 서평자는 알고 있다. 정대현의 책『로봇 의식론』은 저자의 오랜 동안의 연구물들이 농축된 집약체이다. 그런데 짧게 축소된 원고 분량이라는 제한된 출간 형식에 맞추다보

니 축약된 원고로 쓰여진 듯하다. 저자가 각각의 장마다 제시해준 참고문헌을 참조할 수 있다면 이 책은 미래 로봇시대를 먼저 읽을 수 있는 지름길이라고 서평자는 확신한다.

로봇에 대한 과학기술적 물음(that P?)만으로 로봇의식의 가공할만한 미래를 담지할 수 없으며 "왜"라는 형이상학적 물음(why P?)을 유기적으로 연결시켜야 한다고 저자는 이 책에서 내내 강조했다. 이런 두 물음이 연결되어야만 인간-기계의 공존성이 인간-인간의 공생성으로 이어질 수 있다고 느껴졌다. 철학이 로봇의식에 대하여 무엇을 말할 수 있는지 깨닫게 되었다는 점이 서평자에겐 더 소중한 결론이다. <끝>