인공지능과 의식 : 강한 인공지능의 존재론적 및 의미론적 문제*

손 병 홍 (한림대), 송 하 석 (아주대), 심 철 호 (동해대)

주 제 심리철학, 인공지능, 인지과학
주요어 계산주의, 연결주의, 의식, 펜로즈, 찰머스, 써얼

요약문 튜링 테스트를 통과할 컴퓨터가 가까운 시일 내에 출현할 가능성조차도 보이지 않지만 계산주의자도 연결주의자도 강한 인공지능론을 포기하지 않는다. 그들의 이론이 근거하고 있는 기능주의는 가장 의식에 관한 해명을 해주지 못한다는 취약점을 갖는다. 그런 점에서 기능주의자면서도 의식의 두 국면을 분리하여 의식의 존재와 강한 인공지능론의 양립가능성을 모색하려는 찰머스의 시도는 흥미로운 대안으로 볼 수도 있다. 그러나 찰머스의 제안에도 불구하고 펜로즈 등의 괴델리안 논변은 강한 인공지능론에 대한 존재론적 위협일 수 있다. 하지만 괴델리안 논변 또한 정당화되지 못한 가정을 전제로 하고 있다는 점에서 인공지능 불가론에 결정적 승기를 안겨주었다고는 보기 어렵다. 결국 인간과 컴퓨터의 존재론적 차이 여부에 관한 논쟁 또한 써얼의 중국어방 논변에서 보듯이, 지능, 이해, 의식 등의 원초적 개념에 대한 의미론적 합의를 선결문제로 요한다는 점에서 강한 인공지능론의 진위 여부는 어떤 경험 과학의 발전에도 좌우되는 것이 아닌 또 하나의 영원한 철학적 숙제일 따름이라는 비관적인 교훈이 재확인되고 있을 따름이다.

1. 머리말

민스키(M. Minsky)가 인공지능을 "인간에 의해서 행해진다면 지능을

* 이 논문은 1999년 한국 학술진흥재단의 협동 연구 지원에 의하여 연구되었음. (과제번호: 1999-042-B00043).

필요로 했을 일을 기계로 하여금 하게 하려는 연구"[1]라고 정의한 이래, 인공지능은 일반적으로 인간과 같은 지능을 소유하고 지적인 행위를 할 수 있는 인공물을 만들어내려는 일련의 연구로 이해되고 있다. 그러나 민스키보다 앞서 1947년 튜링(A. Turing)은 "인간의 마음의 행위를 매우 가깝게 모의할 기계를 만들 수 있다"고 주장하고[2], 1950년 그는 인공지능의 핵심 문제는 "기계가 생각할 수 있는가"라고 말했다.

1950년 중반부터 본격적으로 시작된 인공지능의 연구에 참여한 사람들은 컴퓨터와 같은 정보처리 체계는 지능적이라고 여겨지는 행위를 실현하도록 만들어질 수 있다고 믿은 점에서 튜링의 충실한 계승자들이었다. 이러한 인공지능 연구를 계산주의(computationalism)라 한다. 반면에 인공지능 연구 초기에도 오늘날 연결주의(connectionism)라고 불릴 만한 발상을 갖고 연구에 임한 이들이 있었다. 그러나 1970년대까지는 적어도 계산주의의 괄목할 만한 성과 및 당시의 연결주의 모델의 분명한 한계 때문에 연결주의는 주목받지 못했다. 그런데 1980년대에 계산주의 모델의 한계가 드러나고 반대로 뇌 과학과 신경생리학의 눈부신 발달에 힘입어 연결주의는 이제 인공지능 연구에 지배적인 이론이 되었다. 이런 의미에서 계산주의와 연결주의의 주장에 대해서 비교, 검토해 보고 인공지능의 가능성을 모색해 보는 것은 의미 있는 작업일 것이다.

또한 인공지능의 가능성에 대해서는 부정적인 입장이 있는가 하면 인공지능을 옹호하는 입장도 크게 두 가지로 나뉜다. 즉 적절한 입력과 출력을 갖춘 프로그램만 있다면 어떠한 체계든지 그리고 그것이 무엇으로 만들어졌든지 문자 그대로 마음(지능)을 갖는다고 주장하는 강한 인공지능(strong AI)의 옹호자와 이보다는 좀 신중하게 컴퓨터는 인간의 마음을 모의할 수 있기 때문에 마음을 연구하는 데 유용하게 사용될 수 있다고 주장하는 약한 인공지능(weak AI)의 옹호자가 있다. 약한 인공지능 논제는 펜로즈(R. Penrose)와 같은 소수의 예외적인 경우를 제외하고는 이제

1) Minsky, M. L. *Semantic Information Processing* (Cambridge, MA: MIT press) 1968. v쪽.
2) Turing, A. *Alan M. Turing* (ed.) Turing, S. (Cambridge: W. Heffer) 1959, 128-34쪽.

대부분 받아들여지고 있고, 또한 튜링 이래 대부분의 인공지능 옹호자들은 "기계가 생각할 수 있는가"라는 물음에 긍정적으로 대답하려는 강한 인공지능 옹호자들이기 때문에 강한 인공지능의 성공 여부가 우리의 주된 관심사가 될 것이다.

이 글에서 우리는 강한 인공지능 논제에 대해서 검토해 보기 위해서 우선 계산주의와 연결주의를 간단히 살펴보고(2절), 강한 인공지능 논제의 존재론적 문제를 살펴볼 것이다(3절). 여기서 논의할 존재론적 문제란 바로 인간의 마음에서의 주관적 의식의 문제와 컴퓨터의 알고리즘과 인간의 인지 과정간의 극복할 수 없는 논리적 격차의 문제인데, 우리는 찰머스(D. Chalmers)와 펜로즈의 주장을 중심으로 이들 문제가 어떻게 해결될 수 있는가를 검토할 것이다. 그런 다음 인공지능 논제의 의미론적 문제로서 "지능"이 의미하는 바, 즉 "지능"의 외연은 무엇인가와 써얼의 중국어 방 논변을 통한 비판을 중심으로 살펴볼 것이다(4절). 그리하여 우리는 강한 인공지능의 가능성 문제는 존재론적 문제와 의미론적 문제에 대한 적절한 해명이 전제되지 않는 한 합리적 논의조차 불투명하게 될 수 있음을 논증할 것이다.

2. 계산주의와 연결주의

계산주의는 기본적으로 인간의 마음을 유한한 하나의 정보처리 과정으로 보고, 마음의 특성이라고 여겨지는 이해, 지능, 의식 등에 대한 설명을 기능적, 계산적 과정으로 설명하려고 시도한다. 여기서 정보란 시스템 사이의 소통을 위해 생성, 전달, 수용되는 신호로서 그 신호를 해석하고 사용하는 인지주체의 바깥에 존재하는 것이다. 그리고 정보처리의 과정인 계산은 그 시스템의 상태에 대한 형식적, 물리적 속성 위에서 수행되기 때문에, 계산주의는 인간과 컴퓨터는 모두 기호를 조작하는 물리적 기호체계라는 가설에 근거한다. 이 물리적 기호체계 가설(physical symbol system hypothesis)이란 물리적 기호체계는 일반적인 지능적 행위를 위해

필요하고도 충분한 수단이라는 것이다.[3] 즉 이 가설은 어떤 대상이 지능을 갖는다는 것은 곧 그 대상이 물리적 기호체계를 사례화한(instantiate) 것임을 뜻한다.

그렇다면 이와 같은 정보처리 과정이 어떻게 의미를 가질 수 있는가? 즉 어떻게 무엇인가를 표상할 수 있는가? 정보는 의미를 가져야하고 의미는 표상을 통해서 얻어질 수 있는 것이라는 점에서 이 물음은 인지과학에서 핵심적인 것이다. 계산주의는 이에 대해서 기호의 표상을 주장한다. 즉 기호의 가장 중요한 속성은 무엇인가를 지칭한다(designate)는 것이고 기본적인 대상을 지칭하는 원자기호들은 서로 결합하여 표현을 형성할 수 있으며, 그렇게 형성된 표현들은 보다 복잡한 대상이나 개념을 가리킨다. 그리하여 어떤 대상이나 개념을 지칭하는 기호는 하나의 원자적 존재로서 물리적 기호체계에 의해서 명백하게 조작될 수 있고, 그 결과 지능적인 행동을 낳을 수 있다는 것이다. 계산주의는 그러한 기호에 대하여 직접적으로 계산을 수행하는 프로그램을 다룬다. 요컨대 인간의 마음은 정보를 처리하는 체계이고 정보처리는 기호를 계산, 조작하는 과정이며, 컴퓨터의 프로그램도 기호를 조작하는 체계이므로, 따라서 인간의 마음은 컴퓨터의 프로그램에 다름아니라는 것이 계산주의의 핵심 주장이다.

반면 연결주의는 계산주의처럼 인간의 마음을 정보처리 과정으로 간주하지만, 그 정보처리 과정이 직렬적(serial)이 아니라 병렬적(parallel)이라고 보는 점에서 계산주의와 다르다. 연결주의 중에서 가장 대표적인 신경망 모델에 따르면, 약 140억 개에 달하는 뇌의 신경세포는 동시에 작용하는 병렬적 처리 과정을 가지며, 이 과정을 병렬 분산적 계산 요소들의 대규모 연결망에 의해서 모의하여 인간의 마음을 설명하고자 한다. 즉 정신과정은 신경망 구조와 같은 거대한 연결망 속에서 활성화(activation)의 정도가 다른 구성요소들 사이의 연결통로와 연결의 강도의 차이에 의

3) Newell, A. & Simon, H. "Computer Science as Empirical Inquiry," *Communications of the Association for Computing Machinery*, (1976) 19, 117쪽.

해서 야기되는 정보의 변형이라는 것이다. 따라서 그들은 심적 표상이 계산주의자들의 주장과 달리 기호로 간주될 수 없고 표상의 처리도 단선적, 직렬적 조작이 아니라 여러 수준이 중첩된 대규모의 전체적 연결망 속에서 병렬적으로 처리된다고 주장한다.

연결주의 체계에도 계산이 있지만 그것은 개념단계의 하위에 있는 것이다. 즉 계산은 결절(node)의 단계나 결절과 결절 사이의 연결(connection)의 단계에서 발생하고 개별적 결절과 결절들의 연결은 의미론적 역할을 하지 않는다. 의미론적 역할은 보다 높은 단계, 즉 분산된 표상(distributed representation)의 단계에서 이루어진다. 이러한 표상은 수많은 다른 결절에 대한 행동의 패턴으로 구성된다. 이러한 패턴에 의해서 그 표상은 중요한 인과적 역할을 수행하는 복잡한 내적 구조를 갖게 된다. 표상의 구성요소인 결절들과 연결은 그 자체로는 의미론적이지 않으며 따라서 표상도 아니다. 그러므로 연결주의 체계는 원자기호를 사용하지 않으며 물리적 기호체계 가설을 거부하는 것이다. 연결주의는 그리하여 하위기호 가설(subsymbolic hypothesis)에 근거한다고 할 수 있다. 이 가설에 따르면, 직관적 처리과정은 완전하고 형식적이며 정확한 개념 수준의 기술(description)을 허용하지 않는 하위 개념의 역동적인 연결주의적 체계다.[4]

결국 계산주의에서 프로그램의 대상, 즉 계산의 대상은 실제 세계의 어떤 존재자를 직접 지시하지만, 연결주의 모델에서는 세계의 존재자와 계산되는 대상 사이에 직접적인 대응이 전제되지 않는다. 연결주의 체계도 규칙을 따르기는 하지만, 그 규칙은 의미론적 단계보다 낮은데, 그 낮은 단계의 규칙 따르기의 결과로부터 의미론적 속성이 창발된다. 그러므로 두 체계 사이의 차이는 구문론적인 면에 있는 것이 아니라, 의미론적인 측면에 있다. 보다 정확히 말해서 두 체계의 가장 근본적인 차이는 계산적(구문론적) 성질이 표상적(의미론적) 성질과 연결되는 방식에서 있는 것이다.

---

4) Smolensky, P. "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*, (1988) Vol. 11, 7쪽.

연결주의의 신경망 체계에서 계산 토큰은 개별적인 결절과 연결인 반면, 계산주의의 기호체계에서는 LISP[5] 원자와 같은 것이다. 어떤 체계의 기본적인 의미론적 대상은 표상이다. 연결주의에서 표상은 결절들의 집합에 대한 분산된 패턴의 활성화이고, 계산주의에서는 LISP 원자이거나 표현일 수 있다. 이렇게 두 체계 모두 계산토큰과 표상을 갖지만, 그 둘 사이의 관계에 대해서 두 체계는 각각 다른 설명을 한다. 계산주의에서 표상과 계산토큰은 같은 단계에서 발생한다. 모든 기본적 표상은 원자적 계산토큰이고 다른 표상들은 기본적인 표상을 결합함으로써 이루어진다. 예컨대 LISP 원자는 계산토큰으로 기능하는 동시에 표상의 담지자로서 기능하기도 한다. 하나의 기호는 계산토큰이면서 또한 하나의 표상이라는 점에서 일치한다(coincide). 반면에 연결주의에서 표상과 계산토큰은 완전히 분리된다. 개별적인 결절들과 연결이 계산토큰인데 그것들은 표상과 완전히 다른 단계 구조이다. 표상은 그러한 계산토큰에 대한 분산된 패턴의 행위인 것이다. 즉 계산의 단계는 표상의 단계 하위에 있다. 그러므로 계산주의와 연결주의의 가장 중요한 차이는 계산적 단계와 표상적 단계가 일치하는가 그렇지 않은가에 있는 것이다.

---

5) 인공지능 분야의 소프트웨어를 작성하기 위해 사용되는 가장 고전적인 프로그래밍 언어 중의 하나이다. LISP는 리스프처리라는 의미의 LISt Processing의 약자이다. LISP는 리스트(list:자료구조의 하나로 순서가 매겨진 0개 이상의 원소들의 집합) 형태로 된 데이터를 함수적으로 처리하여 다른 리스트를 출력하도록 되어 있다. 예컨대(a b c) 형식의 일반화된 리스트(generalized list)로 기술되므로 프로그램이 데이터처럼 취급되는 것이 특징이다. 즉, 프로그램과 자료가 같은 형태이기 때문에 자료구조가 프로그램처럼 시행될 수 있으며, 프로그램이 자료처럼 연산될 수도 있고, 기본 자료구조가 연결리스트(linked list)를 사용하여 일반적인 연산을 수행하기도 한다. 또한 다른 프로그래밍언어와 달리 특수한 표기법을 필요로 하지 않는다. 따라서 여러 개의 명제를 모아서 하나의 지식데이터를 형성하고, 필요한 지식을 찾아서 사용하는 일에 알맞게 설계된 비수치형 연산에 적합한 언어라고 할 수 있다. Franklin, S. *Artificial Minds* (Cambridge, Mass.: The MIT Press, 1999), 151쪽 참조.

3. 인공지능 논제의 존재론적 문제

오늘날 계산주의니 연결주의니 하면서 컴퓨터가 인간의 마음을 모의하는 수준을 넘어서 그대로 복제까지 할 수 있을 가능성을 주장하는 입장들을 두고 벌어지는 논의를 확실하게 종식시킬 수 있는 사실상 유일한 길은 그런 컴퓨터를 만들어내는 일밖에 없다. 실제로 지금까지 만들어진 어떤 슈퍼컴퓨터조차도, 예컨대 인간 체스 챔피언을 눌렀다는 Deep Blue 조차도, 인간의 마음을 따라잡기에는 멀어도 한참 멀었다는 데에는 어떤 컴퓨터 전문가도 부인하지 않는다. 하지만 강한 인공 지능론을 신봉하는 많은 컴퓨터 전문가들은 "기다려라! 곧 만들어낼 테니."라고 벼르고 있다. 최근 컴퓨터의 발전 속도에 비추어보면 이는 그리 허황된 주장도 아닌 듯한 느낌을 준 것 또한 사실이다. 어떤 이는 21세기 내에 실현되리라고 장담하기도 한다.[6] 그러나 반대로 아무리 컴퓨터가 발전한다 해도 인간의 마음은 마치 점근선과 같은 존재로서 그것에 컴퓨터가 가까이는 가되 끝내 도달할 수는 없는 어떤 원리적 차이성을 주장하는 쪽의 논변도 있다. 이 절에서는 의식과 컴퓨터가 어느 정도의 차이가 있는가에 초점을 맞추어 강한 인공지능 논제의 존재론적 가능성을 조명해보는 것을 목적으로 한다.

계산주의든 연결주의든 컴퓨터 기능주의에 그 심리철학적 토대를 두고 있다는 점에서는 마찬가지다. 기능주의는 심적 현상을 제거하거나 물리적 현상으로 환원하려는 유물론적 노선에서 가장 뒤늦게 나왔으면서도 오늘날 가장 폭넓은 지지층을 확보하고 있는 견해로서 그 기본 발상은 다음과 같다.[7] 기능주의는 심적 상태를 물리 상태로 간주하되, '심적'인 것을 물리적 구성이 아니라 기능적 인과관계에 의해 정의한다. 이를테면 디지털 시계든 물시계든 아니면 막대기로 된 해시계든 그것이 시간을 알

6) Moravec, H. *Mind Children: The Future of Robot and Human Intelligence* (Harvard Univ Press, 1988), 1쪽.

7) Searle, J. *The Mystery of Consciousness* (New York Review, 1997), 139-140쪽.

려주는 기능을 충실히 수행하기만 하다면 그 재료나 생김새가 어떠하든 그것은 시계로서 인정받듯이 심적 상태도 마찬가지라는 것이다. 즉, 모든 믿음과 욕구는 물리적 '체계'의 물리적 상태지만, 그 체계는 상이한 종류의 물질들로 이루어질 수 있다. 어떤 것이 믿음이거나 욕구인 것은 그것이 무슨 일을 하는지, 그 인과관계가 무엇인지에 의해서 결정되는 것이지, 그 체계를 이루는 재료에 의해서가 아니다. 올바른 인과관계를 갖는 상태만 있다면 뇌든 컴퓨터든 다른 어떤 무엇이든 마음을 가질 수 있다. 올바른 패턴의 인과관계를 이룰 수만 있다면 뉴런 활성화든 컴퓨터 프로그램의 한 단계이든 다른 무엇이든 믿음을 이룰 수 있다. 그것이 믿음이라고 정의되는 것은 인과 관계의 패턴 상의 위치 때문이다. 이 패턴을 '기능적 조직(functional organization)'이라 하며, 하나의 시스템이 믿음을 갖는다는 것은 올바른 기능적 조직을 갖는다는 것일 뿐이다. 시스템의 기능적 조직은 물리적 입력을 취해서 그 체계 내에서 일련의 내적 인과관계 공정을 거쳐 물리적 출력을 산출하는 것으로 되어 있다. 요컨대 기능주의란 심리상태는 기능 상태요 기능 상태가 물리 상태라는 견해라는 점에서 유물론의 전통에 서 있는 입장이다.

기능주의는 비과학적인 이원론과 실패한 유물론인 행동주의와 물리주의를 대신할 수 있으면서 각각의 장점만 결합한 듯이 보이기도 한다. 뇌는 컴퓨터고 마음은 뇌에서 실행되는 컴퓨터 프로그램으로서, 심적 상태는 뇌의 프로그램 상태일 뿐이라는 컴퓨터 기능주의는 특히 인지과학 분야에서 일약 지배적 이론의 자리를 굳히게 되었다.

그러나 어떤 기능주의자도 아직까지 만족스럽게 설명해주지 못하는 맹점이 있으니 그것이 바로 의식의 문제다. 이를테면 통증과 같은 의식 현상들에 대한 설명에서 일상의 과학적 상식적 견해에 따르면, 통증과 같은 의식 현상이란 내적, 질적, 주관적 경험인 동시에, 뇌와 신경계에서의 일정한 신경생리학적 과정들에 의해 유발되는 것이라고 설명된다. 그러나 기능주의에 따르면 통증이란 의식 상태는 물리적 상태로서, 두뇌나 다른 어떤 것의 기능적 조직 패턴의 일부로서, 상해(傷害)와 같은 어떤

입력 자극이 신경계의 물리적 상태를 유발하며(컴퓨터 기능주의에 따르면 이들은 정보-처리 상태다) 이들이 다시 모종의 물리적 출력 행동을 유발한다. 이 기능적으로 조직된 물리적 상태들이 통증을 유발하는 것이 아니라 그 자체가 곧 통증이라고 보는 점에서 기능주의의 의식 설명은 상식이나 과학과 정면으로 충돌하지 않을 수 없는 곤경에 빠진다. 의식의 문제 설명에서 기로에 놓이게 된 기능주의자 가운데는 네이글(Thomas Nagel)처럼 기능주의를 포기하고 의식의 환원불가능성을 받아들이거나[8], 데넷(D. Dennett)처럼 기능주의를 고수하는 대신 의식의 환원불가능성을 부정하고 사실상 의식의 존재 자체마저 거부하는 보다 극단적인 선택을 하기도 한다.[9] 기능주의자들에게는 어쩌면 그밖에 별다른 선택의 여지가 없는 듯이 보였다. 그러나 의식과 기능주의를 모두 인정하면서 강한 인공지능론까지 고수하려는 찰머스의 입장은 기능주의의 곤경에 대한 새로운 돌파구로서 주목을 받게 되었다.[10]

찰머스의 해법은 사실 전혀 새로운 것만은 아니다. 찰머스는 일찍이 포더(J. Fodor)가 의식의 좁은 내용/넓은 내용을 구분한 것과 유사한 맥락에서 현상적 의식과 심리학적 의식이라는 두 가지 의식을 구분한다.[11] 현상적 의식은 의식적으로 경험된 심적 상태를 지칭한다. 심리학적 의식은 마음의 인과관계 혹은 인간 행동에 대한 설명이 가능한 의식이다. 현상적 의식을 기능주의적 설명에서 배제시키고 심리학적 의식은 기능주의 및 강한 인공지능론으로 포착하려는 것이 찰머스의 전략이다. 현상적 의식은 그것이 느껴지는 방식에 의해 규정되는 한편 심리적 의식은 그것이 작용하는 방식에 의해 규정된다. 전자는 주체가 의식적으로 느끼고 경험하는 것으로서의 토대로서의 의식이다. 찰머스는 현상적 의식은 의식경험의 현상적, 질적, 본래적, 주관적 특성을 가지며 일인칭 주관적 관점에

8) Nagel, T. "What is it like to be a bat?," *Philosophical Review* (83, pp.435-50, 1974)
9) Dennett, D. *Consciousness Explained* (New York: Back Bay Books, 1991).
10) Chalmers, D. *The Conscious Mind: In Search of a Fundamental Theory* (Oxford Univ. Press, 1996).
11) Chalmers, D. (1996) 25-26쪽.

서만 접근가능한 반면, 심리학적으로 규정된 의식은 주관적인 현상적 특질을 도입하지 않고도 객관적 접근이 가능한 영역이라 한다. 현상적 의식과 대비되는 이러한 의식은 객관적 과학탐구(신경생리학, 인지과학)가 가능한 '인지적 의식'이라 할 수 있다. 행동주의자들과 신경생리학자들은 심리학적 의식인 관찰가능한 경험 과학적 의식을 탐구하는 데 몰두해 왔다. 그들은 신경생리학의 눈부신 탐구결과를 기다리며, 결국 의식의 과학의 영역에서 해결되리라고 믿는다. 반면 의식의 현상적 측면에 관심을 두는 주관주의자들은 현상적 의식은 결코 물리주의로도 기능주의로도 환원될 수 없다고 믿는다.

찰머스의 구분 자체에 대해 얼마든지 논란이 있을 수 있지만 이 글의 논의와 관련하여 보다 중요한 점은 적어도 심리학적 의식의 영역에는 강한 인공지능론이 적용 가능하다는 점이라 하겠다. 다시 말하여 심리학적 의식이라는 의미에서는 기계도 의식을 가질 수 있다던가 적절히 프로그램된 컴퓨터는 마음을 지닐 것이라고 믿을만한 이유가 있다는 것이다.

찰머스의 논증은 다음과 같이 전개된다.[12]

물리적 체계는 그 체계의 인과적 구조가 계산의 형식적 구조를 반영할 때 계산을 실행한다. 즉, 물리적 체계의 상태를 계산상태로 대응시키는 방식이 있어서 인과적으로 관련되는 물리적 상태가 형식적으로 관련되는 형식적 상태로 대응될 때 그 물리적 상태는 계산을 실행한다고 간주된다. 그런데 기능적 조직은 추상적 구성 요소와 그 구성 요소의 상태, 그리고 그 상태가 이전의 상태와 입력에 어떻게 의존하고, 출력은 이전의 상태와 어떻게 의존하는가를 나타내주는 의존 관계를 구체화함으로써 결정된다. 결국 기능적 조직은 일정한 계산 형식으로 추상화될 수 있다. 즉, 기능적 조직을 실현하는 것은 정확하게 그에 대응하는 계산 형식을 실행하는 것이다. 이점은 다시 "의식적 경험을 갖는 체계와 동일한 행태적 역량을 결정해줄 정도로 동일한 기능적 조직을 갖는 체계는 원래의

12) Chalmers, D. (1996) 321-322쪽.

체계와 동일한 의식 경험을 갖는다."는 '조직적 불변성의 원리(principle of organizational invariance)'[13]와 결부되어 올바른 기능적 구조를 가진 체계는 그것이 무엇으로 만들어졌든지, 즉 컴퓨터이든 기계이든 다른 무엇이든, (심리학적인 의미에서의) 의식을 갖는다고 볼 수 있다는 것이다.

그러나 기능주의에 토대를 둔 찰머스의 인공 지능 옹호론은 결국 실천적으로는, 의식적 경험을 갖는 체계의 기능적 조직에 대응되어 추상화되는 계산 형식을 어떻게 구성 또는 발견할 수 있느냐의 문제를 피할 수 없다는 점에서 당초의 인공 지능의 가능성 문제에서 사실상 한 걸음도 더 진전된 바가 없다. 계산의 형식들은 이미 알려진 것과 아직 알려지지 않은 것을 망라해서 그 종류가 얼마든지 많다. 도대체 어떤 형식의 계산이 인간의 의식을 복제할 수 있는가? 그러나 더욱 심각한 문제는 어떤 형식의 계산이라도 인간의 의식을 복제할 수 없다는 결론이 도출될 가능성이다. 이른바 괴델리안 논변으로 알려진 이 반(反) 계산주의 논변이야말로 인공지능의 가능성에 관한 어떤 논의에서도 우선적으로 검토되어야 할 과제가 아닐 수 없다.

일관적인 어떤 형식 체계라도 결정불가능한 식이 존재한다는 괴델(K. Gödel)의 제1 불완전성 정리와 형식 체계의 일관성은 그 체계 내에서는 증명될 수 없다는 제2 불완전성 정리로부터 컴퓨터가 인간의 마음을 모의할 수 없다는 최초의 괴델리안 논변은 40년 전 루카스(J. Lucas)의 논문에서였다.[14] 논변의 개요는 다음과 같다. 어떤 형식 체계 내의 문장들의 목록 가운데 그 자신에 관한 괴델 문장 G(F)가 있다고 해보자. 그리고 G(F)는 다음과 같이 그 자신에 관한 것을 말하고 있다고 해보자.

G(F): G(F)는 이 체계 내에서 증명가능하지 않다.

---

13) 이 원리의 정당화에 관해서는 Chalmers, D. (1996) 249쪽 이하 참조, 이에 대한 비판으로는 Searle, J. (1997) 151쪽 참조.

14) Lucas, J. R. "Minds, Machines and Gödel", in *Philosophy*, Vol. 36 (1961) 112-127쪽. Anderson, A. R. 編, *Minds and Machines* (Prentice Hall, 1964)에 재수록.

G(F)가 거짓이라고 해보자. 그렇다면 G(F)의 주장 내용과 반대로 G(F)는 증명가능하다. 즉, 증명가능하지 않음이 증명가능하다는 모순된 결과를 얻는다. 그렇다면 G(F)는 거짓일 수 없고, 따라서 증명가능하지 않다. 이는 체계 내에서 증명가능하지 않지만 우리 인간은 그것이 참임을 알아볼 수 있는 경우를 보여주는 좋은 예다. 루카스에 따르면 컴퓨터는 알고리즘만 사용한다. 알고리즘이란 어떤 문제를 풀거나 명제를 증명하기 위해 취해지는 일군의 행위를 규정하는 정확한 규칙들의 집합을 일컫는다. 그래서 컴퓨터가 정리를 증명할 때는 정리 증명 알고리즘을 이용해야 한다. 그런데 위의 G(F)의 경우에서 보다시피 그 체계에서 참임을 우리는 알 수 있지만 그 체계의 정리 증명 알고리즘으로는 증명할 수 없는 문장들이 존재한다. 루카스에 따르면 그렇다면 그런 참인 문장들에 대한 우리의 인식은 알고리즘에 의한 것이 아니다. 컴퓨터는 알고리즘만 쓰므로 우리는 컴퓨터일 수 없다. 다시 말해서 우리의 이해는 컴퓨터를 능가하는 것, 즉 우리는 컴퓨터 이상의 존재인 것이다.

그러나 인공 지능이 실제의 인간을 따라잡을 가능성을 원천적으로 봉쇄하는 루카스 논변에 대한 반론은 초기부터 다양하게 전개되어 왔다.[15]

한 가지 반론은 위의 괴델 문장 G(F)를 문제의 형식 체계에서는 증명할 수 없다 해도 또다른 형식 체계에서는 증명할 수 있을 가능성을 배제하지 못한다는 점이다.[16] 나아가 이런 진리들에 대한 우리의 인식이 정리 증명 알고리즘에 의해 나오지 않는다는 사실로부터 우리가 이런 결론들에 도달하는 데에 아무런 알고리즘도 쓰지 않음이 도출되지는 않는다는 것이다. 즉, 모든 알고리즘이 정리 증명 알고리즘이지는 않다. 예컨대, 써얼은 망막에 대한 2차원적 자극으로부터 3차원적 시야 기술을 구성하는 알고리즘을 갖는 인지과학에서의 시각 모의 프로그램을 반례로 들고

15) 루카스 식의 괴델리안 논변에 대해 이 글에서 다루지 않는 여타의 문제점들에 대해서는 선우환 (1995) "튜링 기계로서의 마음과 괴델의 정리" 『인지과학』 제6권 제3호, 6-11쪽을 참조할 것.

16) Benacerraf, P. "God, the Devil and Gödel," *The Monist* Vol. 51 (1967) 9-32쪽.

있다.[17] 그런 알고리즘은 예컨대 망막에의 자극들이 어떻게 대상에 대한 시각 영상을 산출하는지를 규정하겠지만, 망막 자극에서 3차원 기술로 진행하는 알고리즘이 어떤 정리들을 증명하는 것은 아니라는 것이다. 마찬가지로, 루카스의 경우 설령 그것의 참이 그 체계에서의 어떤 정리 증명 알고리즘에 의해 확립될 수 없다 해도, 우리는 정리 증명 절차가 아닌 어떤 계산 절차를 사용하고 있을 수 있다는 것이다.

루카스 식 괴델 논변에 대한 이같은 반론을 피하기 위해 임의의 어떤 알고리즘에 의해서도 증명가능하지 않지만 인간은 그것이 참임을 알 수 있는 진리가 있음을 보여주도록 하는 보다 일반적인 증명으로서 펜로즈는 "멈춤 문제의 해결불가능성 증명"이라는 튜링 판 괴델 증명을 갖고서 위의 문제점을 극복하고자 하는 반계산주의 논변을 부활시켰다. 멈춤 문제란 어떤 계산이 멈출지(즉, 끝날지) 여부를 결정해줄 일군의 수학적 절차들을 찾는 문제로서 그 개요는 다음과 같다.[18]

언젠가 멈추는 계산 절차든 끝내 멈추지 않는 계산 절차든 모든 계산 절차를 망라해서 C_0, C_1, C_2, ... 등으로 숫자를 매겼다고 하자. 그 중 어느 계산 절차가 멈출지를 어떻게 찾아내는가? 임의의 두 수 q와 n에 대해, $C_q(n)$을 입력 n에 대해 작동하는 q번째 계산이라 하자. 다음과 같은 또다른 특수한 계산 절차 A를 가정해서 A가 입력 (q, n)에서 멈추면 $C_q(n)$이 멈추지 않음을 알려준다고 하자. 즉 (1)과 같다고 하자.

(1) A(q, n)이 멈추면, $C_q(n)$은 멈추지 않는다.

(1)의 한 경우로서 q=n인 경우를 보자. 그러면 다음 (2)를 얻는다.

(2) A(n, n)이 멈추면, $C_n(n)$은 멈추지 않는다.

---

17) Searle, J. (1997) 65쪽.

18) Penrose, R. *Shadows of the Mind*(Oxford Univ. Press, 1994), 74-76쪽.

이제 서로 다른 두 수가 아닌 단 하나의 수 n에만 의존하는 A(n, n)은 또한 C_0, C_1, C_2, ... 들 가운데 하나여야 한다.(C_0, C_1, C_2, ...는 어떤 하나의 자연수 n에서 수행될 수 있는 모든 계산들의 목록이라고 가정되어 있기 때문이다.) 그 중에 A(n, n)이 사실상 C_k라 하면 다음(3)을 얻는다.

(3) $A(n, n) = C_k(n)$

이제 특별히 n=k인 경우를 살펴보자. 그러면 (3)로부터 다음 (4)를 얻는다.

(4) A(k, k) = Ck(k).

n=k라 했으니, (2)로부터, 다음 (5)을 얻는다.

(5) A(k, k)가 멈추면, $C_k(k)$는 멈추지 않는다.

(4)에 의해 (5)의 A(k, k)에 $C_k(k)$를 대입하면, 다음 (6)을 얻는다.

(6) $C_k(k)$가 멈추면, $C_k(k)$는 멈추지 않는다.

이로부터 $C_k(k)$가 멈추지 않음이 도출된다. 그런데 A(k, k) 또한, (4)에 의해 $C_k(k)$와 동일한 것이므로, 멈출 수 없다. 따라서 계산 절차 A는 이 특정의 계산 $C_k(k)$가 멈추지 않음을, 설령 실제로 멈추지 않고 또 우리가 그 멈추지 않음을 알고 있다 해도, 말해줄 수 없다. 그렇다면 우리는 알고 있는 바인 $C_k(k)$가 멈추지 않음을 A는 말해주지 못하는 것이다.

그러므로 A가 건전하다는 지식으로부터, A에 의해서는 멈추지 않음을 증명할 수 없는 $C_k(k)$와 같은 어떤 멈추지 않는 계산 절차를 보여줄 수 있다. 따라서 우리는 A가 말해줄 수 없는 것이 있음을 알고 있으므로, A는 우리의 이해를 표현하기에 불충분하다.

그런데 A는 우리가 가진 일체의 건전하다고 알려질 수 있는 알고리즘들을 망라한 것으로 가정되었으므로, 즉 임의의 계산 알고리즘들의 집합일 수 있다는 점에서, 어떤 알고리즘을 수행하는 컴퓨터도 우리와 대등할 수 없다는 것이 펜로즈의 결론이다. 더욱이 우리는 A 및 그 건전성에 대한 인식으로부터, 멈추지 않는 계산 $C_k(k)$를 실제로 구성할 수 있으므로, 우리는 A가 무엇이든 A가 계산이 멈추지 않음을 확인하는 데에 수학자들이 활용가능한 절차들의 형식화일 수 없다. 그러므로 인간 수학자는 수학적 참의 확인에 건전하다고 알려질 수 있는 알고리즘을 이용하지 않는다고 그는 결론짓는다.

펜로즈의 논변은 루카스의 경우보다 일반적인 증명이라는 점에서, 즉 특정 알고리즘이 아닌 임의의 어떤 알고리즘의 총체도 인간의 인지 방식을 따라잡을 수 없다는 점에서 앞서 말한 루카스 류의 괴델리안 논변의 문제점을 극복하는 데에는 성공적이라 할 수 있다. 그러나 루카스-펜로즈 논변에는 일찍이 퍼트남(H. Putnam)[19]이 지적했던 또 하나의 문제점이 남아 있다. 인간의 마음은 "F가 일관적이면 G(F)가 참이다."라는 조건 문장만 증명할 수 있을 뿐이어서, F가 일관적임을 알 수 있지 않는 한 우리는 G(F)가 참임을 알 수 없다. 그런데 F가 매우 복잡한 체계라면, F가 일관적임을 우리는 어떻게 알 수 있을 것인가?

이에 대한 펜로즈의 일차적 대답은 다음과 같다. 우리는 F를 공리들과 추론 규칙들의 체계라고 볼 수 있다. 분명 각각의 공리가 참임을 알 수 있다. F가 그것들의 참을 알 수 있다면 우리도 알 수 있다. 게다가 우리의 추리가 우리가 근본적으로 의심스러운 것으로 여기는 추론 규칙들에 의존할 수 있다는 것은 지극히 설득력이 없기 때문에 기본적 추론 규칙들(rules of procedure)이 타당함 또한 알 수 있어야 한다. 그리하여 공리들이 참이고 추론 규칙들이 타당함을 우리가 안다면, 우리는 F가 건전함

19) Putnam, H. "Minds and Machines," *Dimensions of Mind: a symposium* (ed. S. Hook, NYU Press, 1964) 148-79쪽. Putnam, H. *Mind, Language and Reality* (Cambridge Univ, Press, 1975) 362-385쪽에 재수록.

을 안다.[20] F가 건전함을 안다는 것은 F가 일관적임을 안다는 것을 함축한다는 점에서 우리는 F가 일관적임을 안다.

이에 대해 찰머스는 펜로즈가 적절한 부류의 계산주의적 체계들은 모두 1차 논리에서의 정리-증명자(theorem-prover)에 유사한 것이라고 가정하는 듯한데, 기존의 AI 연구 영역 내에서조차도, 공리들과 추론 규칙들로 분석될 수 없는 연결주의자의 네트워크와 같은 많은 계산주의적 절차들이 있다는 점에서 이런 가정은 근거가 없다고 비판한다.

그러나 펜로즈는 F의 일관성이나 건전성을 우리가 알 수 있음을 명시적으로 요구하지 않는 또 하나의 논변을 제시하는 바 이는 다음과 같이 재정리해볼 수 있다.[21]

(1) 나의 추리력을 어떤 형식 체계 F가 포착한다(더 간단히 "나는 F다"라 하자). (가정)

(2) 내가 F임을 내가 안다면, (나는 내가 건전함을 아니까) 나는 F가 건전함을 안다. 사실, 나는 F에 "나는 F다"가 보충된 보다 큰 체계 F'이 건전함을 안다. (건전한 체계에 참인 진술을 보충하면 건전한 체계를 낳는다.)

(3) 내가 F임을 안다는 가정으로부터 나는 괴델 문장 G(F')의 참도 안다는 것이 도출된다.

(4) 그러나 F'은 (괴델 정리에 의해) G(F')이 참임을 알 수 없으니, 나는 F'과 다르다.

(5) 그러나 가정에 의해, 나는 이제 F'과 사실상 동치다. 즉, 나는 내가 F라는 지식이 보충된 F다.

(6) 이는 모순이므로 최초의 가정이 거짓이고, F는 나의 추리력을 포착할 수 없다.

---

20) Penrose, R. (1994) 133-135쪽.

21) Penrose, R. (1994) §3.23에서의 로봇 수학자와의 환상적인 대화에서 간접적으로 제시한 내용을 Chalmers, D. (1996) §3.2에서 재구성한 것이다.

(7) 결론적으로, 나의 추리력을 어떤 형식 체계도 포착할 수 없다.

이상의 귀류법 논증에 따라 펜로즈는 "나는 F가 아니라"고 결론을 내린다. 그러나 내가 G(F')가 참임을 알 수 있음을 보임에 있어서, 우리는 내가 F임이 아니라 내가 F임을 내가 안다를 가정한 것이니, 엄밀히 말하자면 "나는 내가 F임을 알 수 없다"는 것이다. 그래도 이는 우리가 어떤 F인가와 동일함을 경험적으로 발견할 수 있을 가능성을 배제한다는 점에서 AI의 전망을 위협할만큼의 강한 결론이다.

그러나 찰머스는 이 논변의 전제 (2)에서의 우리는 우리가 일관적임을 안다는 가정이 모순을 낳음을, 어떤 체계의 일관성을 확실하게 믿는 체계는 모두 모순을 낳는다는 맥컬로우(D. McCullough)의 다음 논증을 이용하여 펜로즈의 논변을 거부한다.[22]

(가정) (1) $\vdash$A면, $\vdash$B(A). (A를 주장할 수 있는 체계는 A를 믿는다는 주장도 할 수 있다)

(가정) (2) $\vdash$B(A1)&B(A1$\rightarrow$A2) $\rightarrow$ B(A2) (그 체계는 전건긍정 추론을 할 수 있음을 안다)

(가정) (3) $\vdash$B(A)$\rightarrow$B(B(A)) (그 체계가 (1)을 안다)

(가정) (4) $\vdash \sim$B(false) (그 체계는 자신이 모순이 아님을 주장한다)

(5) $\vdash$G$\leftrightarrow \sim$B(G) (G를 B(diag($\sim$B(diag(x))))라고 하면, 대각화 함수에 의해, "G다"와 "나는 G를 믿지 않는다"가 동치임을 계산가능성에 관한 아무런 가정이 없이도 얻을 수 있다)

(6) $\vdash$B(G)$\rightarrow$B($\sim$B(G)) ((5), (1), (2)로부터)

(7) $\vdash$B(G)$\rightarrow$B(B(G)) ((3)으로부터)

(8) $\vdash$B(G)$\rightarrow$B(false) ((6), (7), (2)로부터)

(9) $\vdash$B(false)$\rightarrow$B(G) ((2)와 $\vdash$B(false$\rightarrow$G)로부터)

---

22) Chalmers, D. "Minds, Machines, and Mathematics," *Psyche* (Vol. 2, No. 9, 1995), §§ 3.8-3.11.

(10) ⊢G⟶~B(false) ((5), (8), (9)로부터)
(11) ⊢B(G) ((10), (4), (1)로부터)
(12) ⊢B(false) ((11), (9)로부터)

그래서 우리가 건전함을 우리가 안다는 가정은 모순을 초래한다. 루카스 이래로 펜로즈의 논변에 이르기까지 모든 괴델리안 논변은 어디에선가 이 가정에 호소하지만 이 전제는 모순을 유발한다. 어쩌면 실제로 우리가 건전할 수도 있겠다. 그러나 우리가 건전함을 우리가 확고히 알 수는 없다. 이처럼 모순을 유발하는 가정을 이용해서 어떤 컴퓨터도 논리적으로 극복할 수 없는 인간과의 존재론적 차이가 있음을 주장하는 괴델리안 논변은 설득력이 없다.

'스스로를 믿는 믿음 체계'라는 개념이 야기하는 모순으로부터 벗어나기 위한 펜로즈의 답변은 궁색해 보인다. 펜로즈는 믿음 체계들이 스스로에 대해 믿는 내용에 제한을 가한다면 모순에서 벗어날 수 있다고 주장한다.[23] 즉, 한 체계가 자신이 증명한 내용들 각각에 대해서는 그것들이 올바른 결론들이었는지에 대한 반성적인 고려를 해보는 것이 허용되고 이 반성적인 고려에 대한 메타 반성적 고려를 포함한 그 이상의 반성적 고려는 허용되지 않는 식으로 '스스로를 믿는 믿음 체계'의 스스로에 대한 가능한 믿음의 범위를 국한시키면 모순이 발생하지 않을 수 있다는 것이다. 인간이 스스로의 건전성이나 일관성에 대한 믿음 또는 지식을 가질 때 바로 그 믿음이나 지식 자체에 대한 자기 지시적인 믿음이나 지식을 가질 수 없다면 펜로즈 논증이 어느 정도 설득력이 있다고 볼 수도 있을 것이다. 그러나 괴델리안 논변에서 "우리는 우리가 건전함을 안다"는 가정을 이런 취지로만 이해해야 할 이유도 없을뿐더러, 설령 이렇게 제한된 의미로 이해한다고 해도 원래의 괴델리안 논변이 타당했을 경우에 비교해 볼 때 논변의 파괴력이나 매력을 대부분 상실하는 달갑지 않

23) Penrose, R. "Beyond the doubting of a shadow: A reply to commentaries on Shadows of the Mind," Psyche (Vol. 2, No. 23, 1996), §§ 3.7-3.8

은 결과가 초래될 것이다,

강한 인공 지능의 지지자의 시각에서는 원칙적으로 자신들의 기획에 대한 논리적 장벽은 없다. 오히려 적절한 계산의 실행이 의식 경험을 낳을 것이라고 믿을 많은 긍정적인 이유들이 있다. 그리고 계산의 종류는 펜로즈 등이 생각한 것보다 훨씬 더 다양하다. 문제는 어떤 종류의 계산이 인간의 의식을 복제할 수 있느냐일 것이다. 그러자면 계산, 의식, 지능, 이해 등의 개념에 대한 보다 폭넓은 합의가 먼저 요구된다. 방금 보았듯이 문제의 건전성에 대한 지식 가정을 어떤 의미로 이해해야 하는가의 문제는 당초의 논리적, 존재론적 문제들이 의미론적 문제를 선결 문제로 요구하는 것을 보여주는 좋은 사례라 하겠다. 이에 대한 논의가 다음 절의 과제다.

4. 인공지능 논제의 의미론적 문제

앞에서 언급한 것처럼 인공지능이 인간과 같은 지능을 지닌 기계를 만들려는 연구라고 한다면 무엇보다 "지능"이라는 개념의 의미가 무엇인가가 밝혀져야 한다. 이 문제가 인공지능의 가능성과 관련한 의미론적 문제의 출발점이 되는 이유는 인공지능에 대한 전통적인 찬반 논의가 바로 이 지능이라는 개념에 대한 임의적인 해석에 근거하고 있는 것처럼 보이기 때문이다. 즉 인공지능에 대해서 반대하는 전통적인 주장은 컴퓨터와 같은 기계는 비록 인간의 지능의 일부를 모의할 수 있다고 할지라도 인간이 지니는 감정이나 독창성과 같은 성질을 가질 수 없기 때문에 지능적일 수 없고, 인간과 같은 사고를 하는 것이 아니라는 것이다. 이 주장을 일반화하여 논증으로 구성하면 다음과 같다.

X는 인간의 지능의 구성요소 중의 하나이다.
컴퓨터는 X를 갖지 않는다.
그러므로 컴퓨터는 인간과 같은 지능을 갖지 않는다.

예컨대 맥코덕(P. McCorduck)은 컴퓨터는 인간이 갖는 감정이나 독창력을 가질 수 없다고 주장하고[24], 드레츠키(F. Dretske)도 다음과 같이 컴퓨터의 지능이 인간의 지능과 다르다고 주장한다.

> 나는 특별한 기술이나 지식 그리고 이해를 결여하고 있지만, 합리적 행위자로 구성된 사회의 구성원에게 본질적인 것은 아무것도 결여하고 있지 않다. 그러나 기계에 대해서 말하면, 가장 세련된 현대의 컴퓨터를 포함해서 그것들은 본질적인 어떤 것을 결여하고 있다.[25]

그러나 '지능'에 대한 이러한 설명은 우선 선결문제 질문을 요구하는 오류를 범하고 있다. 왜냐하면 인지과학의 중요한 목표는 여러 요소들로 구성된 인간의 지능을 기계로 모의 혹은 실현할 수 있는가를 탐구하는 것이기 때문이다. 즉 통속적으로 기계에 부과하기 힘들다고 믿어졌던 인간 지능의 요소가 어떻게 기계를 통해서 실현할 수 있는가가 인지과학의 주된 과제이기 때문이다.

반면에 인공지능의 초기 옹호자들은 인간의 지능을 엄격하게 정보처리 과정이라고 정의함으로써 그들의 입장을 뒷받침하고자 했다. 그러나 지능에 대한 이러한 설명은 지나치게 협소해서 일반적으로 "지능적"이라고 할 수 있는 많은 행위들을 그 외연에서 제외해버릴 것이다. 예컨대 차를 운전하거나 요리를 하는 것과 같은 비언어적 기술(nonverbal skill)의 습득이나 어떤 대상에 대한 욕구와 관련된 지능 등을 어떻게 정보처리 과정이라는 말로 설명할 수 있을 것인가?

이와 같은 인공지능 옹호자나 반대자들의 '지능'에 대한 설명은 그 용

---

24) McCorduck, P. *Machines who Think* (San Francisco, CA: Freeman), 1979, 171쪽.

25) Dretske, F. "Machines and the Mental," *Proceedings of APA* (1985), 23쪽. 카파츠코프와 드레이퍼스도 이와 비슷하게 기계는 사회적, 역사적 측면을 가질 수 없다고 주장한다. Karpatschof, B. "Artificial Intelligence or Artificial Signification?," *Journal of Pragmatics*, Vol. 6 (1982), 293-304쪽, Dreyfus, H. L. *What Computers Can't Do: A Critique of Artificial Reason* (New York: Harper & Row, 1972) 등을 참조할 것.

어의 외연을 고정되어 불변적인 것이라고 믿는다는 공통점을 갖는다. 그러나 인지가 발달하고 과학적 지식이 축적되어 감에 따라 많은 용어들의 외연이 변할 수 있다는 것은 주지의 사실이다. 예컨대 근대과학 혁명 이전에는 '행성'이라는 용어의 외연에 포함되지 않았던 지구가 코페르니쿠스 이후에는 '행성'의 새로운 외연으로 포함되게 되었다. 이처럼 우리는 지능이라는 용어의 외연도 뇌 과학의 발달과 컴퓨터 공학의 발달로 외연이 달라질 수 있으리라고 기대할 수 있다.

이제 인간의 정신 활동을 대표하는 용어로 사용되는 '지능'의 외연을 낱낱이 열거하는 것은 불가능하지만, 그 특징적인 성질들을 통해서 몇 가지로 구별해 보자.[26] 우선 언어적 기호들을 수용하고 그에 대한 적절한 반응을 제공할 수 있는 언어 사용자와 같은 기능을 수행하는 능력으로서의 지능이 있다. 이는 계산주의자들이 관심을 가지고 접근한 기호로 된 정보처리 과정으로서의 지능이라고 할 수 있기 때문에, 우리는 이를 구문론적 지능이라고 하자. 둘째는 계산적인 방식에 의해서가 아니라 두뇌의 인과력을 복제하는 방식으로 언어적 기능을 수행하는 능력으로서의 지능이 있다. 이를 인과적 지능이라고 하자. 그리고 셋째 외부 대상에 대한 일정한 태도를 가질 수 있는 지향성을 유지할 수 있도록 언어적 기호와 비언어적 대상을 상호 연관지움으로써 도달하게 되는 이해와 관련된 지능이 있다. 이를 지향적 지능이라고 하자. 끝으로 네이글이 지적한 것과 같은 주관적인 이해와 관련된 지능이 있을 수 있다.[27]

---

26) 탱과 아담스는 지금까지 알려진 경험적 사실을 통해서 지능의 특징적 외연을 17가지로 설명하고 있다. 물론 그들도 인정하고 있듯이, 그 17가지가 지능에 대한 완전한 설명은 아니지만, 지능에 대한 행동주의적 설명으로서 설득력있어 보인다. Tang, P. & Adams, S. "Can Computers Be Intelligent? Artificial Intelligence and Conceptual Change" *International Journal of Intelligent Systems*, Vol. 3 (1988), 13쪽 참조할 것.

27) 이러한 지능에 대한 설명은 무디가 이해(understanding)를 F-이해, C-이해, I-이해, 그리고 S-이해로 구별한 것으로부터 얻은 것이다. 그리고 김선희도 인공지능의 이해 개념을 무디의 구별에 따라 설명하고 있다. Moody, T. *Philosophy and Artificial Intelligence* (New Jersey: Prentice Hall) 1993, 149쪽과 김선희 "인공지능과 이해의 개념" 『인지과학』 (1997) Vol. 8, 44쪽을 참조할 것.

여기서 우리의 관심은 의미론적인 문제이기 때문에 앞 절에서 다루어진 주관적 지능은 제외하고 구문론적 지능과 인과적 지능, 그리고 지향적 지능에 대해서 생각해 보자. 인공지능 옹호자들이 지능을 정보처리 과정으로 설명한 것은 명백하게 첫 번째 종류의 지능을 염두에 둔 것이고, 컴퓨터와 같은 기계도 그러한 의미의 지능을 가질 수 있다는 것은 이제 의심의 여지가 없어 보인다. 그러므로 문제는 컴퓨터가 인과적 지능과 지향적 지능을 가질 수 있는가일 것이다. 즉 인공지능 옹호자들의 논변이 설득력을 가지려면 인과적 지능과 지향적 지능도 정보처리 과정으로 설명될 수 있어야 할 것이다. 그러나 써얼은 이에 대하여 중국어 방 논변(chinese room argument)으로 알려진 사유실험을 통해서 비판적으로 대답한다. 그러면 써얼의 중국어 방 논변을 간단히 살펴보자.

모국어를 영어로 가지고 있고 중국어를 전혀 모르는 사람이 폐쇄된 방에 있다고 가정하자. 그 방에는 중국어 단어들이 들어 있는 상자가 있고 중국어로 된 질문들에 답할 수 있는 방법에 관한 프로그램이 따라야 할 규칙들을 담고 있는 영어로 된 규정집이 있다. 그 규정집의 규칙들은 그 사람에게 주어진 중국어 단어들을 구문론에 의해 형식적으로 처리하도록 하는 지침이다. 그 사람에게 중국어로 된 질문이 주어지면, 그는 규정집의 규칙에 따라 중국어로 된 대답을 밖으로 내보낸다. 밖에 있는 사람이 방안에 있는 사람이 중국어를 아는지 모르는지를 평가하는 유일한 방법은 그가 제공하는 대답일 뿐이기 때문에 그는 중국어를 아는 것처럼 판단될 것이다. 써얼이 보이고자 하는 것은 이 방 안의 사람은 중국어를 이해하고 있는 것처럼 보이지만 그는 결코 중국어를 이해하고 있는 것이 아니라는 것이다. 요컨대 그는 단순히 컴퓨터처럼 형식적 규칙에 따라 계산적 기능을 수행하여 정확한 답을 제공하지만 중국어를 이해하고 있다고 말할 수 없기 때문에, 계산적 정보처리 기능에 다름아닌 컴퓨터도 지능이나 이해를 가질 수 없다는 것이다. 써얼은 중국어 방이라는 사유실험을 통해서 다음 논증의 공리1)이 옳다는 것을 보여주고 있는 것이다.

공리1) 구문론은 의미론에 충분하지 않다.
공리2) 컴퓨터 프로그램은 전적으로 형식적, 구문론적 구조에 의해서 정의된다.
공리3) 마음은 내용, 구체적으로 의미론적 내용을 갖는다.
결론) 프로그램만을 실행하는 것은 마음을 갖기에 충분하지 않다.[28]

처칠랜드 부부(P. & P. Chuchland)는 써얼의 논증을 비판하면서, 그들이 계산주의를 포기하고 연결주의를 옹호하게 된 것은 써얼의 논증 때문이 아니라 계산주의가 전제하는 기호체계 가설의 약점과 실현 가능성의 문제 때문이라고 말한다. 그들은 써얼의 논증의 첫 번째 전제인 "구문론은 의미론에 충분하지 않다"는 것은 순환논증의 오류를 범하고 있다고 주장한다. 인공지능의 연구가 추구하는 것은 적절하게 구조된 내재적인 조작을 통해서, 즉 순수하게 구문론적인 조작을 통해서 인간과 동일한 인지적 상태와 결과를 얻으려는 것이기 때문이다. 즉 이 전제 자체가 인공지능 연구가들이 추구하고 있는 것이므로 이를 인공지능 논제의 비판을 위한 논증의 전제로 삼는 것은 선결문제의 질문의 오류를 범하는 것이라는 것이다.[29] 결국 문제는 구문론이 의미론에 충분한가, 즉 기능적 지능이 인과적, 지향적 지능을 확보해 줄 수 있는가인 셈이다.

물론 다른 전제들에 대해서도 많은 비판이 제기되어 왔다. 즉 제거주의자로서 처칠랜드와 데넷은 의미론적 내용을 갖는 마음의 존재를 부인하기 때문에 위 논증의 공리3)에 대해서도 비판을 한다. 그리고 데넷은 컴퓨터 프로그램은 전적으로 구문론적이라는 공리2)에 대해서도 비판을 한다. 그는 프로그램이 순수하게 그 형식적 특성에 의해서만 설명될 수 있는가라는 문제가 최근에는 법적인 공방거리가 되고 있다는 점을 지적

28) Searle, J. *Minds, Brains and Science* (Cambridge: Harvard Univ. Press, 1984) 39-40쪽.
29) Churchland, P. & P. "Could a Machine Think?" *Scientific American* (Jan. 1990) Vol. 262, 33쪽을 참조하고, 이에 대한 보다 자세한 설명은 송하석의 "중국어방 논변과 인공지능" 『철학적 분석』 (2000, 봄호) Vol. 2를 참조할 것.

하면서, 구문론적으로, 즉 기술의 단계에서는 다르지만 실제로 동일한 작업을 수행하는 프로그램은 동일한 것이라고 말한다. 즉 "만약 프로그램을 구별하는 기준에 그 프로그램의 세부적인 물리적 구현이 포함된다면 프로그램은 단순히 구문론적이라고 할 수 없을 것"[30]이라고 말한다. 그럼에도 써얼 논증에서 인공지능의 의미론적 문제와 관련하여 가장 중요한 것은 역시 공리1)이 제기하는 문제이다. 과연 구문론이 의미론에 충분하가?

데넷이 지적하고 있듯이, 실행되지 않은 채로 책상 위에 놓여 있는 프로그램은 전적으로 구문론적이며, 그것은 그 자체만으로는 의미론에 충분하지 않다는 것은 분명하다. 그러나 공리1)에서 써얼이 말하고자 하는 것은 물리적으로 적절하게 구현된 프로그램은 그 자체로 의미론에 충분하지 않다는 것이다. 그것은 프로그램의 실행만으로는 마음을 만들어 낼 수 없다는 논증의 결론을 통해서도 확인될 수 있다. 그렇다면 이제 문제는 물리적으로 적절하게 구현된 구문론적인 프로그램이 의미론에 충분한가이다.

이에 대해서 대답하기 전에 먼저 써얼이 말하는 의미론이 무엇인지 살펴보아야 한다. 여기서 그가 말하는 의미론은 외재적(externalist) 의미론이 아니라 내재적(internalist) 의미론이다. 외재적 의미론이란 어떤 것이 외부세계의 분명한 대상이나 사건을 지시할 때 의미를 갖는다고 말하는 것으로써 기호나 정신상태는 외부적 대상(외연)과의 관계에 의해서 그리고 외부세계와의 적절한 인과적 관계에 의해서 의미를 갖는다는 주장이다. 반면 써얼이 정신이 의미론적 내용을 갖는다고 말할 때 의미론적 내용이란 구체적인 외부 환경과 독립된 것, 즉 외연이 아니라 내포(intension)이다. 우리는 순수하게 구문론적인 계산만으로 의미를 야기할 수 없지만, 외재론적 입장에서도 컴퓨터를 실재세계와 올바른 방식으로 연결한다면 컴퓨터는 세계와의 인과적 관계에 의해서 의미론적 내용을 갖게 될 것이

30) Dennett, D. *The Intentional Stance* (Cambridge: MIT Press, 1987) 336쪽.

라고 말할 수 있다. 그러나 써얼이 말하는 내재적 의미론의 입장에서는 순수하게 형식적인 체계는 외부 환경과의 관계에 상관없이 내재적인 의미론을 가질 수 없을 것이다. 결국 컴퓨터가 기호들을 조작할 수는 있지만, 그 체계 내에는 의미를 부여하는 어떤 것도 없으며 따라서 컴퓨터는 내재적 의미를 갖는 인간과는 구별될 수밖에 없다. 결국 컴퓨터가 갖는 것은 구문론적 지능일 뿐이고, 이러한 지능은 인과적, 지향적 지능을 보장해주지 못하며, 따라서 컴퓨터는 인간과 같은 지능을 지닌다고 말할 수 없다는 것이다.

다시 말해서 아무런 내재적인 구조를 갖지 않는 계산토큰들을 규칙에 따라 조작하는 것은 외재적 의미를 가질 수는 있을지라도 내재적 의미를 가질 수는 없다는 것이 그의 주장이다. 그리고 이러한 그의 대답의 배후에는 언어의 의미론적 통찰이 담겨있다. 예컨대 CAT라는 토큰은 그것이 사과의 개념이 아니라 고양이의 개념과 더 밀접하게 관련되게 하는 아무런 내재적 속성이나 구조를 가지고 있지 않다. 이 점이 바로 써얼의 논증의 배후에 있는 통찰이다. 중국어 방에 있는 영어 사용자에게 중국어 기호가 아무런 의미론적 내용도 주지 못하는 것처럼, 컴퓨터에게 원자적 내용은 아무런 의미론적 내용을 부여하지 못한다. 써얼의 "구문론은 의미론을 위해 충분하지 않다"는 주장은 그러한 토큰들에 대한 구문론적 조작은 결코 그 토큰에 내재적인 의미를 부여할 수 없음을 주장하고 있는 것이다. 그리고 이러한 주장은 기호체계 모델인 계산주의에 대한 성공적인 비판으로 보인다.

그러나 이러한 써얼의 논변은 연결주의에 대한 성공적인 비판이 될 수 없다. 알고리즘 과정은 비의미론적인 하위단계이고 이 단계로부터 의미론적 단계가 창발된다고 주장하는 연결주의는 표상의 단계와 계산 단계를 구별한다. 연결주의에서 계산 단계의 토큰은 결절과 연결이다. 이것들은 분명히 기본적이고 성질 없는 덩이일 뿐이다. 그러므로 그러한 토큰들은 아무런 내재적인 의미론적 내용도 갖지 않는다. 그러나 이것은 연결주의자들에게 아무런 문제가 되지 않는다. 계산주의와는 달리 연결주

의자들에게는 이 토큰들은 구문론적 대상으로만 간주되기 때문이다. 또한 연결주의 모델에서 표상은 계산주의에서와 달리 기본적인 성질이 없는 대상들이 아니다. 연결주의의 표상은 보다 하위단계인 계산 행위로부터 창발되는(emergent) 복잡한 분산된 패턴(distributed pattern)의 행위이다. 즉 표상은 직접적으로 조작되는 것이 아니라 낮은 단계에서의 조작의 간접적인 결과인 셈이다. 따라서 연결주의의 표상은 풍부한 내재적 구조를 갖고 이러한 내재적 구조에 의해서 분산된 표상은 내용을 갖게 된다. 요컨대 연결주의 모델에는 계산 토큰이라는 구문론적 대상과 표상이라는 의미론적 대상이 있고, 이 둘은 같은 단계의 대상, 혹은 같은 대상이 아니기 때문에 써얼의 구문론/의미론 논증이 적절하게 적용될 수 없는 것처럼 보인다.

그러나 써얼은 연결주의 모델도 계산적 대상의 구문론적 조작이라는 점에서 계산적이고, 구문론은 결코 의미론에 충분하지 않기 때문에, 연결주의 모델은 어떤 내재적 의미론도 가질 수 없다고 주장할 수 있을 것이다. 즉 비록 분산된 표상이 기본적 토큰은 아니지만, 그것이 전적으로 구문론적 조작으로부터 나오는 것이므로 어떤 참된 의미론적 내용도 갖지 못한다고 연결주의를 비판할 수 있을 것이다. 이제 문제가 되는 질문은 "구문론적 체계가 어떤 단계에서 의미론적 내용을 가질 수 있는가"이다. 써얼은 이와 관련하여 오직 뇌나 뇌와 동일한 인과력을 지닌 기계만이 생각할 수 있다"고 말하고, 적절하게 실행된 프로그램일지라도 뇌와 동일한 인과력을 가질 수 없기 때문에 의미론적 내용을 가질 수 없다고 주장한다. 그러나 데넷이 지적한 것처럼 "뇌가 할 수 있는 유일한 것은 일상적인 정신적 담화에서 우리가 가정하는 의미에 대한 대꾸와 유사한 것"이고 "기계적인 압박이 가해질 때 뇌는 항상 국지적인 기계적 상황에서 그것이 하도록 야기된 것을 하는 것"으로서, "뇌는 의미론적 엔진(semantic engines)의 기능을 모방할 수 있는 구문론적 엔진(syntatic engines)이다."[31] 요컨대 인간의 뇌는 엄격한 규칙 -물리적 법칙-을 따르는 것으로 기술될 수 있지만, 동시에 내재적 의미도 갖는다. 그러나 언어의 의미

론으로부터 얻은 통찰에 의해서 써얼은 단어가 따르는 구문규칙은 그 자체로는 단어에 의미 내용을 부여하지 못하고, 문장의 경우에 구문론이 의미론을 위해서 충분하지 않음이 분명하다고 주장한다. 그렇다면 뇌의 경우와 문장의 경우는 어떤 차이가 있어서 구문론이 때로는 의미론에 충분하기도 하고 그렇지 않기도 하는가? 뇌는 구문론적이지만 그 구문론은 매우 낮은 단계에 놓여 있고, 원자 분자 그리고 뉴런 등의 구문론적 속성은 우리가 개념단계에 대해서 말할 때 아무런 역할을 할 수 없다. 그러나 문장의 경우는 구문론과 의미론이 단어의 단계라는 같은 단계에 놓여 있다. 즉 구문론적으로 조작되는 것은 단어이고 의미론적 해석의 대상도 또한 단어이다. 결국 구문론이 의미론적 내용을 위해서 충분하지 않는 경우도 있지만, 그렇지 않은 경우도 있을 수 있음을 인정해야 한다.[32] 결국 계산주의와 연결주의가 계산의 단계와 표상의 단계를 구별하는가 하지 않는가라는 결정적인 차이가 있음에도 써얼은 그 차이의 중요성을 간과함으로써 모든 인공지능 논제를 동일한 것으로 보고 비판하고 있는 셈이다.

이제 좀더 자세하게 기호가 어떻게 의미를 획득할 수 있는가에 대해서 생각해 보자. 다시 말해서 계산체계에서 표상이 어떻게 참된 의미를 가질 수 있는가를 생각해 보아야 하는데, 이러한 작업을 하나드(S. Harnad)가 부른 것처럼[33] 기호 근거지우기(symbol grounding)라고 하자. 계산 과정에서 조작되는 기호들의 의미는 관찰자에 의해서 그 기호에 투사될 수 있을 뿐이다. 기호들이 의미론적 내용을 갖기 위해서는 그 기호들은 비기호적 토대 위에서 근거되어야 한다.

---

31) Dennett, D. *Brainchildren* (Cambridge: MIT Press, 1998) 357쪽.

32) 찰머스는 이와 관련하여 써얼의 주장은 "어떤 단계에서의 구문론은 같은 단계에서의 의미론적 내용을 위해서 충분하지 않다고 수정되어야 한다"고 주장한다. Chalmers, D. "Subsymbolic Computation and the Chinese Room" in *The Symbolic and Connectionist Paradigms*: Closing the Gap(ed.) Dinsmore, J. (Lawrence Erlbaum, 1992) 17쪽.

33) Harnad, S. "The Symbol Grounding Problem," *Physica* Δ, (1990) Vol. 42. 31-32쪽.

기호가 어떻게 의미를 갖는가라는 물음에 대한 대답은 우리가 어떤 의미론을 갖는가에 따라 두가지로 주어질 수 있을 것이다. 만약 우리가 외재적 의미에 관심을 갖는다면 우리는 인과적 근거지우기(causal grounding)의 작업을 할 것이다. 이것은 계산체계를 외부 세계와 연결시키는 작업으로서 표상을 곧바로 그 지시대상과 연결하는 것이다. 즉 참된 표상은 외부세계와 감각기관의 만남에서 근거된다고 설명한다. 예컨대 CAT라는 표상이 외부세계에 실제로 고양이가 등장함으로써 촉발된다면, 우리는 그 표상은 실제로 그것의 외부적 지시대상으로서 고양이를 갖는다고 주장할 수 있을 것이다. 그러나 우리가 내재적 의미론에 관심을 갖는다면, 우리는 내재적 근거지우기(internal grounding)의 작업을 할 것이다. 즉 어떤 기호가 어떤 대상을 표상하도록 내재적으로 구조화되지 않았으며 그것은 홉스타터(D. R. Hofstadter)의 지적처럼 그 자체로는 전적으로 공허한 것이다.[34] 그렇다면 우리의 표상이 내재적 내용을 갖기 위해서 충분한 내재적 구조를 갖는다는 것을 어떻게 설명할 수 있는가? 내재적 근거지우기의 목표는 기본적인 계산 토큰보다 풍부한 표상 내용의 담지자(vehicle)를 찾는 것이다. 연결주의자들은 바로 이러한 표상 내용의 담지자를 분산된 패턴의 행위에서 찾는다. 이렇게 해서 표상의 내재적 구조는 그것이 표상하고자 하는 의미론적 특징을 체계적으로 반영할 수 있다. 이 연결주의의 분산된 행위패턴이 표상을 내재적 구조에서 근거지울 수 있는 수단으로서 가장 좋은 것일 것이다. 반면에 뉴웰(A. Newell)이나 사이먼(H. Simon)과 같은 계산주의자들도 기호는 어떤 표현이든 지시할 수 있다고 말한다. 그러나 그들은 그러한 기호들이 프로그램에 의해서 어떻게 조작되는가에 따라 그리고 그 기호들이 어떻게 외부세계와 관련되는가에 의해 지시체를 갖는다고 말한다. 즉 그들은 기호를 인과적으로 근거지우는 외재적 의미론만을 생각하는 것이다. 이와 달리 연결주의자들은 능동적인 기호를 구상하는데, 그것은 그 자체로 의미를 담지하는

34) Hofstadter, D. R. "Waking up from the Boolean Dream or Subcognition as Computation," In *Metamagical Themas* (New York: Basic Books, 1985) 645쪽.

표상이다. 표상은 프로그램에 의해서 형식적으로 조작되는 것이 아니라 보다 낮은 단계의 계산 조작으로부터, 즉 계산기저(computational substrate)로부터 통계적으로 창발하는 것이다. 그러므로 표상은 분산된 패턴의 행위 속에서 능동적인 내재적 구조를 갖는다. 비록 써얼도 뇌가 가지고 있는 상향적 인과력(bottom up causal power)의 원인은 아직 완전히 밝혀지지 않은 뇌의 생화학적 특성이라고 말하면서[35] 뇌가 갖는 의미론적 내용이 창발적인 것임을 시사하지만, 그는 이를 설명할 수 없는 의식의 신비로 남겨두고 있을 뿐, 그에 대한 설명의 가능성을 부인하고 있는 것이다. 요컨대 써얼은 내재적 의미론을 말하지만, 그와 관련하여 요청되는 내재적 근거지우기의 가능성을 부인함으로써 인공지능의 논제를 비판하고 있는 것이다. 그러나 비록 연결주의자들이 말하는 이러한 능동적 기호에 대한 충분한 설명이 아직 이루어지지 못하고 있다고 할지라도, 그들은 패턴으로서의 표상 개념이 궁극적으로 의미론적 내용의 비밀을 밝힐 수 있으리라고 기대하면서 연구를 수행하고 있는 것이다.

5. 맺음말

아직도 튜링 테스트를 통과한 컴퓨터는 존재하지 않으며, 가까운 시일 내에 출현할 가능성조차도 보이지 않지만 계산주의자도 연결주의자도 강한 인공지능론을 결코 포기하지 않는다. 그들의 주장처럼 컴퓨터 프로그램의 고안만으로 마음을 창조할 수 있다면, 자연에 대한 궁극적인 기술적 지배를 이룩할 것이다. 그러나 강한 인공지능론이 마치 메시아를 고대하는 종교적 신앙과도 같은 인상에서 벗어나 실질적으로 납득할만한 근거를 가진 논제임을 밝히려는 과정은 그리 간단치 않다. 우리의 논의에서 적어도 한 가지 비교적 분명해진 것은 의식의 문제에 대한 설득력 있는 해결이 있어야 강한 인공지능론 또한 그 설득력을 확보할 수 있다

35) J. Searle, "Minds, Brains, and Programs," Behavioral and Brain Sciences (1980) 3, 419쪽과 *Minds, Brains and Science* (Cambridge: Harvard University Press, 1984) 42쪽 참조.

는 것이다.

의식의 존재 자체를 부정하는 극단적인 유물론자와 의식의 주관성만을 강조하는 이원론자 사이에서 의식의 두 국면을 분리하여 줄타기를 시도하는 찰머스 등의 시도는 일단 의식의 존재와 강한 인공지능론의 양립가능성을 모색하는 돌파구를 열었다는 점에서 의의를 찾을 수 있다. 그러나 찰머스의 타협안은 인공지능론의 측면에서 볼 때는 실질적인 해결책이라기보다는 문제들을 새로운 각도에서 다시 제기한 것이라는 측면이 더 강했다고 볼 수 있다. 오히려 루카스, 펜로즈 등의 괴델리안 논변이 옳다면 강한 인공지능론은 이미 논리적으로 발붙일 곳이 없을 것이다. 그러나 괴델리안 논변 또한 궁극적으로 건전성에 관한 미심쩍은 가정을 전제로 하고 있다는 점에서 인공지능 불가론에 결정적 승기를 안겨주었다고는 보기 어렵다는 것이 필자들의 견해다.

결국 인간과 컴퓨터의 존재론적 차이 여부에 관한 논쟁 또한 써얼의 중국어방 논변에서 보듯이, 지능, 이해, 의식 등의 원초적 개념에 대한 의미론적 합의를 선결문제로 요한다는 점에서 강한 인공지능론의 진위 여부는 컴퓨터 과학 기술이나 인지과학 또는 그밖의 어떤 분야의 과학의 발전에도 좌우되는 것이 아닌 또 하나의 영원한 철학적 숙제일 뿐이라는 비관적인 교훈이 재확인되고 있을 따름이다.

참 고 문 헌

김선희, "인공지능과 이해의 개념" 『인지과학』 제8권 제1호 (1997).

선우환 "튜링 기계로서의 마음과 괴델의 정리" 『인지과학』 제6권 제3호 (1995).

송하석, "중국어 방 논변과 인공지능" 『철학적 분석』 제2권 (2000, 봄호).

Benacerraf, P. "God, the Devil and Gödel," *The Monist* (51: 9-32. 1967).

Chalmers, D. "Subsymbolic Computation and the Chinese Room" In *The Symbolic and Connectionist Paradigms: Closing the Gap* (ed.) J. Dinsmore (Lawrence Erlbaum, 1992).

_____, "Minds, Machines, and Mathematics," *Psyche,* (Vol. 2, No. 9, 1995).

_____, *The Conscious Mind,* (Oxford Univ. Press 1996).

Churchland, P. & P. "Could a Machine Think?" *Scientific American* (Vol. 262. Jan. 1990).

Dennett, D. *The Intentional Stance* (MIT Press, 1987).

_____, *Brainchildren: Essays on Designing Mind* (MIT Press 1998).

_____, *Consciousness Explained* (New York: Back Bay Books, 1991).

Dretske, F. "Machine and the Mental," *Proceedings of APA*, (1985).

Dreyfus, H. L. *What Computer Can't Do: A Critique of Artificial Reason* (New York: Harper & Row, 1972).

Dyer, M. G. "Distributed Symbol Formation and Processing in Connectionist Networks" *Journal of Experimental and Theoretical Artificial Intelligence* (1990) 2.

Franklin, S. *Artificial Minds* (Cambridge: The MIT Press, 1995).

Harnad, S. "The Symbol Grounding Problem" *Physica Δ*, (1990) 42.

Hauser, L. "Searle's Chinese Box: Debunking the Chinese Room Argument" *Minds and Machines* (1997) 1.

Hofstadter, D. R. "Waking up from the Boolean Dream or Subcognition as Computation" In *Metamagical Themas* (New York: Basic Books, 1985).

Karpatschof, B. "Artificial Intelligence or Artificial Signification?" *Journal of Pragmatics*, (1982) Vol. 6.

Lucas, J. R. "Minds, machines and Gödel," *Philosophy* (36: 120-4. 1961).

McCorduck, P, *Machine who Think*, (San Francisco, CA: Freeman, 1979).

Minsky, M. L. *Semantic Information Processing* (Cambridge, MA: MIT press, 1968).

Moody, T. *Philosophy and Artificial Intelligence* (New Jersey: Prentice Hall, 1993).

Moravec, H. *Mind Children: The Future of Robot and Human Intelligence*, (Harvard Univ Press, 1988).

Nagle, T. "What is it like to be a bat?" *Philosophical Review* (83, pp.435-50, 1974).

Newell, A. & Simon, H. "Computer Science as Empirical Inquiry" *Communications of the Association for Computing Machinery*, (1976). Vol. 19.

Penrose, R. *Shadows of the Mind: A Search for the Missing Science of Consciousness* (Oxford Univ. Press, 1994).

_____, "Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind*," *Psyche*, (vol. 2, no. 23, 1996).

Putnam, H. "Minds and Machines," *Dimensions of Mind: a symposium* (Proceedings of the 3rd annual NYU Institute of Philosophy) (ed. S. Hook, NYU Press: 148-79. 1964), Putnam, H. *Mind, Language and Reality* (Cambridge Univ, Press, 1975) 362-385쪽에 재수록.

Searle, J. "Minds, Brains, and Programs" *Behavioral and Brain Sciences* (1980) 3.

_____, *Minds, Brains and Science* (Cambridge: Harvard University Press, 1984).

_____, *The Mystery of Consciousness* (New York: New York Review, 1997).

Smolensky, P. "On the Proper Treatment of Connectionism" *Behavioral and Brain Sciences*, (1988) Vol. 11.

Tang, P. & Adams, S. "Can Computers Be Intelligent? Artificial Intelligence and Conceptual Change" *International Journal of Intelligent Systems*, vol. 3, (1988).

Turing, A. *Alan M. Turing* (ed.) S. Turing (Cambridge: W. Heffer, 1959).