중국어 방 논변과 인공지능

송 하 석 (연세철학연구소)

주 제 심리철학, 인공지능
주요어 써얼, 중국어 방, 계산주의, 연결주의
요약문 인공지능에 관한 논의가 전통적인 계산주의에서 전체론적인 연결주의로 발전한 이래, 강한 인공지능 논제에 대한 설득력 있는 반론으로 여겨졌던 써얼의 중국어 방 논변이 여전히 인공지능에 대한 유효한 반론인지는 흥미로울 뿐만 아니라 중요한 논쟁거리이다. 이 글은 써얼의 중국어 방 논변에 대한 국내의 이초식 교수 평가와 이승종 교수의 논의를 중심으로 살펴보고, 나아가서 데넷과 처칠랜드의 반론을 살펴볼 것이다. 그리고 물리적 기호가설에 입각한 계산주의와 하위기호가설에 토대하는 연결주의에 대해서 논하고, 구체적으로 써얼의 중국어방 논변이 전자에 대해서 성공적인 비판이지만, 후자에 대해서는 그렇지 못하다는 것을 주장한다. 결론적으로 써얼의 논변은 전통적인 계산주의에 대해서는 유효한 비판이지만, 구문론적 과정과 의미론적 과정이 구별되는 연결주의에 대해서는 성공적인 비판일 수 없다는 것을 논증한다.

1. 들어가는 말

20세기 중반부터 인공지능에 관한 연구가 시작되어 괄목한 업적이 나옴에 따라 철학 내에서도 그 가능성에 대한 찬반 논의가 활발하게 전개되었다. 즉 컴퓨터도 지능을 가질 수 있는가에 대한 논의는 철학자들 사이의 뜨거운 논쟁거리가 되어 왔다. 이러한 질문에 긍정적으로 대답하는

인공지능 찬성론자들은 인간 인지의 모든 중요한 측면들이 원칙적으로 계산 모형에 의해서 파악될 수 있다는 전제에 근거하여 컴퓨터도 마음을 갖는다고 주장한다. 반면 컴퓨터는 인간과 같은 마음을 가질 수 없다는 인공지능 반대론자들의 주장도 만만치않게 제기되었다. 인공지능의 반대론은 크게 계산체계는 인간과 같은 인지적 체계일 수 없고 인간은 갖지만 컴퓨터는 가질 수 없는 기능적 능력이 인간에게 있다는 외재적 반론(external objection)과 컴퓨터는 인간의 행위를 모의할(simulate) 수는 있지만 의식 경험이나 이해와 같은 마음을 가질 수는 없다는 내재적 반론(internal objection)이 있다.[1] 그러나 오늘날 컴퓨터가 인간의 행위를 모의할 수 있다는 것은 상식에 속하는 것이다. 이제 인공지능과 관련된 철학적 문제는 컴퓨터가 인간의 행위를 모의할 수 있는가 아닌가가 아니라 컴퓨터가 마음을 갖는가, 그렇지 않은가이다. 따라서 이 글은 인공지능에 대한 내재적 반론의 가장 대표적인 써얼의 중국어방 논변에 대해서 살펴볼 것이다.

써얼(J. Searle)이 중국어 방 논변을 통해서 인공지능 연구가들의 노력에도 불구하고 컴퓨터는 지향성이나 주관성 그리고 이해와 같은 인간 정신의 속성을 가질 수 없기 때문에 강한 인공지능은 불가능하다는 주장을 한 논문[2]이 발표된 이래, 수많은 논문들이 이에 대해 찬성과 반대 논증을 제공해 왔다. 특히 1980년대부터 인공지능 연구의 주된 관심이 전통적인 계산주의적 방식에서 연결주의적 방식으로 바뀐 이후 써얼의 중국어 방 논변이 과연 연결주의적 방식에도 적용될 수 있는가는 흥미있는 논쟁거리가 되어 왔다.

써얼은 계산주의(computationalism)의 주장에 토대하여 인간의 마음을 기본적으로 계산적 정보처리 과정이라고 보는 견해를 강한 인공지능

---

1) D. Chalmers, *The Conscious Mind* (New York: Oxford University Press, 1996) 313-314쪽.

2) J. Searle, "Minds, Brains, and Programs" *Behavioral and Brain Sciences* (1980) 3. 그리고 *Minds, Brains and Science* (Cambridge: Harvard University Press, 1984) 참조.

(strong AI) 논제라고 부르고, 이것을 단순히 컴퓨터는 인간의 정신을 모의하고(simulate), 따라서 인간의 마음을 이해하는 데 유용한 수단으로 사용될 수 있다고 주장하는 약한 인공지능(weak AI) 논제와 구별한다. 결국 강한 인공지능 논제에 따르면, 적절하게 프로그램된 컴퓨터는 실제로 마음을 갖거나 마음 그 자체이다. 그리고 써얼의 타겟은 바로 이러한 강한 인공지능의 논제이다. 이러한 강한 인공지능 연구는 컴퓨터에 실행되어야 할 적절한 프로그램이 어떤 것인가에 따라 두 가지 진영으로 나뉘어진다. 튜링 이후 30년의 전통을 이어온 계산주의 진영과 비교적 최근에 등장한 연결주의(connectionism) 진영이 그것이다. 전자는 인간의 마음을 모형화하는 올바른 단계는 기호의 단계라고 주장하며 그러한 프로그램은 직접적으로 그러한 기호에 대해 계산을 수행하는 프로그램에 포함된다고 주장하는 기호인공지능(symbolic AI) 진영이다. 반면 연결주의자들은 기호의 단계는 인간의 마음에 대한 올바른 모델이 되기에는 너무 높다고 생각하여, 기호에 대한 계산을 수행하는 프로그램을 고안하는 대신 뉴런 단계와 같은 보다 낮은 단계에서의 계산을 수행하는 프로그램을 고안해야 한다고 주장하는 하위기호 인공지능(subsymbolic AI) 주창자들이다.

써얼의 중국어 방 논변이 인공지능 논의에 미친 영향력에도 불구하고 그에 대한 적절한 평가가 무엇인가는 여전히 논쟁적이다. 이 논문은 써얼의 논변이 인공지능 논제에 대해서 어떻게 적용될 수 있는가를 살펴보고자 한다. 이를 위해서 써얼의 중국어 방 논변을 간단히 살펴보고 이에 대한 이초식 교수를 비롯한 국내 학자들의 논의를 검토하고(2장), 써얼의 논증에 대한 올바른 평가는 그것이 계산주의에 대해서는 강한 비판으로 적용될 수 있지만 연결주의에 대해서는 적절한 비판이 될 수 없음을 보일 것이다(3장).

2. 써얼의 중국어 방 논변

써얼의 중국어 방 논변(chinese room argument)는 너무 잘 알려져 있기 때문에 자세히 소개할 필요는 없을 것이다. 앞으로의 논의를 위해서 그

의 "중국어 방" 사유실험을 간단히 소개하는 것으로 충분할 것이다.

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.

이러한 사유실험은 일종의 튜링 테스트와 같다. 즉 밖에 있는 사람이 방 안에 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 튜링 테스트를 통과하여 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.
결 론) 프로그램만을 실행하는 것은 마음을 갖기에 충분하지 않다.[3)]

3) J. Searle (1984) 39-40쪽.

이 논증을 직접 평가하기에 앞서 우리는 써얼이 말하는 '의미론'이라는 단어의 뜻에 주의해야 한다. 여기서 그는 내재론적(internalist) 의미론을 말하고 있지 외재론적(externalist) 의미론을 말하고 있는 것이 아니다. 외재론적 의미론이란 어떤 것이 외부 세계의 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 것이라고 말할 수 있다. 그러나 써얼이 취하는 내재론적 입장에서는 순수하게 형식적인 체계는 외부 환경과의 관계에 상관없이 내재적인 의미론을 가질 수 없을 것이다. 결국 컴퓨터가 기호들을 조작할 수 있지만 그 체계에 의미를 부여하는 그 체계에 내재된 어떤 것도 없으며 따라서 컴퓨터는 내재적 의미를 갖는 인간과는 구별될 수밖에 없다.[4]

써얼의 중국어방 논변에 대하여 이초식 교수는 분명하게 밝히지는 않지만 두가지 점에서 성공적이지 않음을 시사하고 있다. 우선 위 논증의 공리1)이 옳다고 할지라도 또다른 참인 공리1)은 다음과 같이 대체될 수 있고,

공리1*) 구문론은 의미론에 (충분하지는 않지만) 필요하다

4) 써얼은 자신의 논변에 대한 비판을 이렇게 내재론적 의미론에 의존하여 대답하고, 그것을 그는 "로보트 답변(robot reply)"라고 한다. 그리고 이러한 주장은 최근 논문에서 반복되고 있다. J. Searle, "Consciousness, Explanatory Inversion, and Cognitive Science" *Behavioral and Brain Sciences*, (1990) 13, 585-596쪽 참조.

따라서 다음과 같은 다른 결론이 타당하게 추론된다고 말한다.

결론*) 프로그램을 실행하는 것이 마음을 갖기에 (충분하지는 않지만) 필요하다.[5]

적어도 이초식 교수는 써얼의 논증의 타당성에 대해서는 의심하지 않고 건전성에 대해서 비판적이라는 점에서 옳다고 생각된다. 처칠랜드 부부(P. & P. Churchlands)와 찰머스도 이 논증의 타당성에 대해서 의심을 한다.[6] 그러나 써얼의 타당성은 의심할 수 없다. 물론 이 논증을 표준적인 양화사를 이용하여 옮겨 보면 타당하지 않은 것처럼 보인다. 즉 위의 논증을 일차술어 논리로 옮긴다면 다음과 같은 부당한 논증이 될 것이다.

공리1) $\neg(\forall x)(Ax \rightarrow Bx)$ (구문론적인 것이 모두 의미론적이지는 않다.)
공리2) $(\forall x)(Cx \rightarrow Ax)$ (컴퓨터는 모두 구문론적이다.)
공리3) $(\forall x)(Dx \rightarrow Bx)$ (마음은 의미론적이다.)
결 론) $\neg(\forall x)(Cx \rightarrow Dx)$

그러나 이 논증을 이차양화 양상논리를 이용하여 써얼이 말하고자 하는 의도에 충실하게 옮기면 다음과 같은 타당한 논증이 될 것이다[F:-구문론적이다, P:-프로그램이다, M:-마음이다, S:-의미론적 내용을 갖는다].

공리1) $(\forall M)\{F(M) \rightarrow \neg \Box(\forall x)(Mx \rightarrow Sx)\}$ (구문론적 속성이 반드시 의미론적 속성에 충분한 것은 아니다.)
공리2) F(P) (프로그램임은 구문론적이다.)
공리3) $\Box(\forall x)(Mx \rightarrow Sx)$ (마음은 반드시 의미론적이다.)

---

5) 이초식, 『인공지능의 철학』 고려대학교 출판부 (서울, 1993) 54쪽.
6) P. & P. Churchland, "Could a Machine Think?" *Scientific American* (Jan. 1990) 262, 35쪽. 그리고 찰머스의 위의 책, 327쪽.

결론) ㄱ□(∀x)(Px→Mx) (프로그램은 반드시 마음에 충분한 것은 아니다.)[7]

이초식 교수의 수정된 공리1*)는 어떤 것이 의미론적 속성을 갖는다면 그것은 반드시 구문론적 속성을 갖는다는 뜻이기 때문에 다음과 같이 옮겨질 것이다.

공리1*) (∀M){F(M)⟶□(∀x)(Sx→Mx)}

그리고 그의 수정된 결론은 다음과 같이 될 것이다.

결 론*) □(∀x)(Mx→Px)

그리고 이 논증도 공리1*)와 공리2-3)으로부터 결론*)이 추론될 수 있어 타당하다.

그러나 이초식 교수가 이러한 논증을 제시한 이유는 분명히 드러나지 않는다. 아마도 그가 주장하고자 하는 바는 컴퓨터 프로그램의 실행이 마음을 갖기에 충분하지는 않다고 할지라도 필요한 것이라고 할 수 있고, 따라서 프로그램의 실행에 무언가가 덧붙여진다면 마음이 생길 수 있는 가능성을 배제할 수 없다는 것인 것같다. 그러나 그는 그 무엇에 대해서는 전혀 논의하지 않고 있다. 이 점에서 이초식 교수도 써얼처럼 프로그램 자체와 프로그램의 실행(implementation)의 차이의 중요성에 대해서 인식하지 못하는 것같다. 프로그램은 추상적인 계산의 대상일 뿐이어서 전적으로 구문론적이지만, 프로그램의 실행은 인과적 역동성을 갖는 구체적 체계여서 전적으로 구문론적이 아닐 수 있음[8]을 간과한 것이다.

---

7) 이에 대한 보다 자세한 설명은 L. Hauser "Searle's Chinese Box: Debunking the Chinese Room Argument" Minds and Machines (1997) 7, 199-226쪽을 참조할 것.

8) 이 점에 대해서는 찰머스의 앞의 책, 315-320참조할 것. 여기서 찰머스는 FSA(finite-

그리고 이초식 교수의 써얼에 대한 두 번째 비판은 래퍼포트(W. Rapaport)의 한국어방 논변에 의지하고 있다. 래퍼포트는 써얼의 중국어방 논변으로부터 강한 AI 논제를 옹호하기 위해서 "한국어방"이라는 사유실험을 제안한다. 영어를 전혀 이해하지 못하는 영문학을 전공하는 사람이 있다고 하자. 그런 그가 셰익스피어 작품들을 한국어 번역판으로 읽고 이에 관한 논문들을 한국어로 발표하였는데 그 논문들이 영어로 번역되어 세계적으로 인정을 받게 되었다. 이 사람은 영어로 된 셰익스피어 작품을 읽지 못하고 이해하지도 못하지만 셰익스피어의 한국어 번역을 이해하고 있고 나아가서 그는 셰익스피어의 한국어 번역판을 통해서 셰익스피어 자체를 이해하고 있는 것이다.[9] 래퍼포트는 셰익스피어의 작품이 여러 자연언어로 번역되어 영어를 모르는 사람들에 의해서 이해될 수 있는 것처럼, 인간의 인지 능력과 마음도 컴퓨터나 그 밖의 다른 여러 인공물에 의해서 구현될 수 있다는 보이고자 하는 것이다. 그러나 이승종 교수도 지적하고 있듯이, 한국어방이라는 사유실험을 통해서 래퍼포트가 보이고자 하는 것은 강한 AI 논변을 옹호하고자 하는 것이지, 단순히 인간의 마음이 컴퓨터와 같은 인공물에 의해서도 모의될 수 있다는 약한 AI가 아니다.[10] 즉 래퍼포트는 "... 기능적 개념, 즉 계산적 개념...은 인간의 뇌뿐만 아니라 컴퓨터에서도 운용될 수 있다고 본다. 따라서 뇌 뿐만 아니라 컴퓨터도 이해할 수 있음을 논증하려 한다"고 말한다.[11]

그러나 이초식 교수는 래퍼포트의 한국어 방 논변이 써얼의 중국어 방

---

state-automata)와 CSA(combinatorial-state-automata)를 구별하여 설명하고 있다. 이 점은 다음 장에서 계산주의와 연결주의를 설명하면서 간략하게나마 다룰 것이다.

9) W. Rapaport, "Syntatic Semantics: Foundations of Computational Natural Language Understanding" Reprinted in *Aspects of Artificial Intelligence* (ed) J. Fetzer (Dordrecht: Kluwer, 1988) 114쪽.

10) 이승종, "컴퓨터의 언어철학" 연세학술논집 26 (서울, 1997) 101쪽.

11) W. Rapaport, "Machine Understanding and Data Abstraction in Searle's Chinese Room" *Proceedings of the 7th Annual Conference of the Cognitive Science Society*, (University of California at Irvine, Hillsdale Lawrence Erlbaum, 1985) 343쪽. 인용은 이승종 위의 논문 101쪽에서 재인용.

논변에 대한 적절한 반론일 수 있는가에 대하여 반성하지 않는다. 이승종 교수는 래퍼포트의 논변이 써얼의 주장에 대한 비판이 될 수 없는 여러 가지 점을 지적하고 있는데[12] 중요한 것은 한국어 방과 중국어 방의 사유실험이 적절한 유비인가이다. 즉 중국어 방 속의 나의 기능은 컴퓨터와 동일하게 전적으로 기계적, 구문론적인 정보처리 기능인 반면, 한국어 방 속의 나의 기능은 단순히 기계, 구문론적이지 않다. 한국어 방에의 입력과 그 방으로부터 출력되는 것 사이의 처리과정이 무엇인가를 생각해 볼 때, 그 과정은 중국어 방의 경우와 달리 구문론적 조작의 수준을 넘어서는 복잡한 의미론적 과정이 포함되어 있다. 다시 말해서 한국어 방 속의 나는 컴퓨터와 기능적 동형성을 가질 수 없을 것이다. 결국 이초식 교수의 중국어 방에 대한 평가는 옳지 않다. 그렇다면 써얼의 중국어 방 논변에 대한 올바른 평가는 무엇인가?

3. 써얼의 논증에 대한 올바른 평가

써얼의 논증에 대한 적절한 평가는 계산주의 인공지능에 대해서는 결정적인 비판일 수 있지만 연결주의 인공지능에 대해서는 적용될 수 없다. 이에 대한 논의를 위해서 먼저 계산주의와 연결주의에 대해서 간단히 살펴보자.

계산주의는 기본적으로 인간의 마음을 유한한 하나의 정보처리 과정으로 보고, 마음의 특성이라고 여겨지는 이해, 지능, 의식 등에 대한 설명을 기능적, 계산적 과정으로 설명하려고 시도한다. 그리고 이러한 계산은 그 시스템의 상태에 대한 형식적, 물리적 속성 위에서 수행되기 때문에, 계산주의는 인간과 컴퓨터는 모두 기호를 조작하는 물리적 기호체계라는 가설에 근거한다.

---

12) 이승종, 위의 논문 102-104쪽 참조.

물리적 기호체계 가설(physical symbol system hypothesis)
: 물리적 기호체계는 일반적인 지능적 행위를 위해 필요하고도 충분한 수단이다.[13]

이 가설이 뜻하는 바는 어떤 대상이 잠재적으로 지능을 갖는다는 것은 곧 그 대상이 물리적 기호체계를 사례화한(instantiate) 것이라는 의미이다. 기호의 가장 중요한 속성은 무엇인가를 지칭한다는 것이다. 그리고 무엇인가를 지칭하는 원자기호들은 서로 결합되어 표현을 형성할 수 있고, 그렇게 형성된 표현들은 보다 복잡한 대상이나 개념을 가르킨다. 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 핵심적인 주장이다.

반면에 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 병렬 처리 과정을 가지며, 이 과정을 병렬 분산적 계산 요소들의 대규모 연결망에 의해서 모의하여 인간의 마음을 설명하고자 한다. 즉 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 것이

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

다. 즉 계산은 결절(node)의 단계나 결절과 결절 사이의 연결(conneciton)의 단계에서 발생하고 개별적 결절과 결절들의 연결은 의미론적 역할을 하지 않는다. 의미론적 역할은 보다 높은 단계, 즉 분산된 표상(distributed representation)의 단계에서 이루어진다.[14] 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 가설에 근거한다고 할 수 있다.

하위기호 가설(subsymbolic hypothesis)
: 직관적 처리과정은 완전하고 형식적이며 정확한 개념 수준의 기술(description)을 허용하지 않는 하위 개념의 역동적인 연결주의적 체계이다.[15]

컴퓨터 프로그램의 대상은, 즉 계산의 대상은 실제 세계의 어떤 존재자를 지시한다고 생각하는 물리적 기호체계 가설에 의존하는 계산주의와 달리, 연결주의 모델은 세계의 존재와 계산되는 대상 사이에 직접적인 대응을 전제하지 않는다. 이러한 체계는 규칙을 따르기는 하지만 그 규칙은 의미론적 단계보다 낮다. 그럼에도 이러한 낮은 단계의 규칙 따르기의 결과로 의미론적 속성이 창발된다는 것이다.

물리적 기호 체계 가설에 따르면 모든 정신과정은 원자적인 기호에 대한 계산이고, 그러한 원자적인 기호의 단계에서 정신과정을 완전하게 기

14) W. Becthel, "Connecitonism" In *A Companion to the Philosophy of Mind* (ed.) S. Guttenplan (Oxford: Blackwell, 1994) 200쪽.
15) P. Smolensky, "On the Proper Treatment of Connectionism" *Behavioral and Brain Sciences*, (1988) 11, 7쪽.

술하는 것이 가능하다. 또한 원자 기호는 정의상 개념의 표상이기 때문에 이러한 기술은 개념적 단계에서의 기술이다. 반면 하위기호 가설에 따르면 물리적 기호체계 가설이 주장하는 정신과정에 관한 완전한 기술이란 불가능하고, 정신과정에 대해 설명하기 위해서 우리는 개념단계 하위에 있는 과정을 고려해야 한다. 그리고 개념단계의 하위라고 함은 결절과 그 결절들의 연결이라는 계산토큰(computational tokens)을 말하는데, 그것은 의미론적 담지자가 아니고 의미론적 담지자는 분산된 표상이라는 보다 상위의 단계에 존재한다. 그러한 표상은 많은 다른 결절들에 대한 행동패턴으로 구성되고 이러한 분산된 패턴에 의해서 표상은 중요한 인과적 역할을 수행하는 복잡한 내적 구조를 가지게 된다. 분상된 표상은 원자적이 아니며 또한 보다 간단한 원자적 표상으로 구성된 복잡한 표현도 아니다.

그러므로 두 체계 사이의 차이를 구문론적 기준에서 보는 것은 적절하지 않다. 두 체계를 구별하는 중요한 기준은 의미론적인 것이다. 즉 계산적(구문론적) 성질이 표상적(의미론적) 성질과 연결되는 방식에 있어서의 차이가 근본적인 것이다.

구문론적인 입장에서 모든 계산체계에 기본적인 구문론적 대상은 계산토큰인데 그것은 계산이 발생하기 위해서 조작되어지는 원자적 대상이다. 연결주의의 신경망 체계에서 계산토큰은 개별적인 결절과 연결인 반면, 계산주의의 기호체계에서는 LISP 원자와 같은 것이다. 어떤 체계의 기본적인 의미론적 대상은 표상이다. 이것이 그 체계의 의미론을 담지하고 있는 대상이다. 연결주의에서 표상은 결절들의 집합에 대한 분산된 패턴의 활성화이고, 계산주의에서는 LISP 원자이거나 표현일 수 있다.

계산주의와 연결주의 모두 계산토큰과 표상을 갖는다는 점에서는 동일하다. 그러나 그 둘 사이의 관계에 대해서 그들은 다른 설명을 하고 있다. 계산주의에서 표상과 계산토큰은 일치한다(coincide). 모든 기본적 표상은 원자적 계산토큰이고 다른 표상들은 기본적 표상을 결합함으로써 이루어진다. 예컨대 LISP 원자는 계산토큰으로 기능하는 동시에 표상의

담지자로서 기능하기도 한다. 하나의 기호는 계산토큰이면서 또한 하나의 표상이다. 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 다음과 같이 요약할 수 있을 것이다.

(a) 기호체계에서는 계산적 단계와 표상적 단계는 일치하지만, 하위기호 체계에서 계산적 단계는 표상적 단계의 하위에 있다.
(b) 기호체계에서 계산의 대상은 의미론적 해석의 대상이지만, 하위기호 체계에서 계산의 대상은 의미론적 해석의 대상보다 더욱 낮은 단계의 미세한 대상이다.

앞에서 살펴본 것처럼 써얼의 논변의 핵심은 튜링 테스트를 통과하는 기계라할지라도 그것은 실제로 인간과 같은 의식이나 지능을 가질 수 없다는 것이므로, 그 논변은 일차적으로 계산주의에 대한 비판인 것처럼 보인다. 그러나 써얼은 분명히 계산체계 일반에 대해 비판하고자 했고 연결주의 체계도 하나의 계산체계이기 때문에 그는 연결주의도 그의 비판의 대상이라고 주장한다.[16]

먼저 써얼의 논증이 계산주의에 대한 적절한 비판인지부터 살펴보자. 제거주의자로서 데넷(D. Dennett)은 써얼의 논증의 공리 모두에 대해서 의심을 품지만, 특별히 "마음은 의미론적 내용을 갖는다"는 세 번째 공리를 비판한다. 데넷에 따르면 인간이나 동물이나 혹은 기계에 의식적 삶이라고 부를만한 것은 존재하지 않고, 따라서 마음이 의미론적 내용을 갖는다는 것은 일종의 몽매주의(obscurantism)일 뿐이다.[17] 컴퓨터도 인간

16) 써얼은 최근에 발표된 책에서 여전히 연결주의에 대해서도 비판적인 태도를 취하고 있다. *The Mystery of Consciousness* (New York: New York Review, 1997) 1장 참조.

처럼 정신적 내용을 가질 수 있다고 주장하는 대부분의 강한 인공지능 옹호자들과 달리 데넷은 컴퓨터든 인간이든 정신적 내용을 갖지 않기 때문에 본질적으로 이 둘은 다르지 않다고 주장한다. 그러나 정신 상태가 일몰(sunset) 현상처럼 실은 없는 것에 대한 착각일 뿐이라는 생각은 잘못인 것처럼 보인다. 과학이 일몰에 대해서 말해주는 것은 해가 하늘을 따라 움직이는 것처럼 보인다는 것은 현상일 뿐, 실제로는 그렇지 않다는 것이지, 현상적 자료의 존재를 부인하는 것은 아니다.[18] 즉 경험적 자료들이 착각임을 보이는 것이 그 자료의 존재 자체를 논박하지는 않는다. 마찬가지로 고통과 같은 의식적 경험은 현상일 뿐, 실제로는 뇌의 현상에 다름아니라는 주장은 의식적 경험으로서의 고통이라는 현상적 자료의 존재를 부인하지는 못한다. 결국 제거주의 입장에서 써얼의 논증의 공리3)을 부인하려는 데넷의 시도는 반직관적일 뿐만 아니라 논리적으로도 성공적이지 않다.

처칠랜드 부부도 써얼의 논증이 계산주의에 대한 적절한 비판일 수 없다고 주장한다. 연결주의의 옹호자로서 그들이 계산주의를 포기한 것은 써얼의 논증 때문이 아니라 계산주의가 전제하는 기호체계 가설의 약점과 실현 가능성의 문제 때문이라는 것이다. 그들은 써얼의 논증의 첫 번째 전제인 "구문론은 의미론에 충분하지 않다"는 것은 순환논증의 오류를 범하고 있다고 말한다. 인공지능의 연구가 추구하는 것은 우리가 구문론적 요소들을 적절하게 구조된 내재적인 조작을 통해서 인간과 동일한 인지적 상태와 수행을 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. 보다 상세한 설명을 위해서 그들은 다음과 같은 어둠의 방(luminous room)이라는 사유실험을 제안한다.[19]

---

17) D. Dennett, *Consciousness Explained* (New York: Back Bay Books, 1991) 450쪽.
18) J. Searle, *The Mystery of Consciousness* (New York: New York Review, 1997) 112쪽.
19) 처칠랜드 위의 논문, 35쪽.

암실 안에 어떤 사람이 충전된 물건이나 자석막대를 들고 서 있다고 가정하자. 만약 그가 그 막대를 위 아래로 흔들면, 맥스웰의 이론에 따라 그것은 전자기파를 발산하는 원을 만들어 낼 것이고, 곧 빛을 발산할 것이다. 그러나 자석이나 충전된 물체를 가지고 놀아본 우리가 모두 잘 아는 것처럼, 그 물체에 힘이 가하여져 움직여질지라도 힘 그자체는 어떤 빛도 만들어내지 않는다. 즉 우리가 맥스웰의 이론에 대해서 알지 못한다면 우리는 어떤 힘으로도 실제적인 빛을 구성할 수 있다고 생각할 수 없을 것이다.

이 사유실험이 보이고자 하는 것은 앞의 써얼 논증과 유비적인 다음 논증의 공리3)이 옳다는 것이다.

공리1) 전기와 자기는 힘이다.
공리2) 빛의 본질적 속성은 발광성(luminance)이다.
공리3) 힘은 그자체로 발광성을 구성하거나 발광성에 충분하지 않다.
결 론) 전기와 자기는 빛을 구성하거나 빛을 위해서 충분하지 않다.

이 논증이 타당하다고 할지라도 위의 어둠의 방의 사유실험에도 불구하고 우리는 맥스웰의 설명에 의해서 공리3)이 옳지 않다는 것을 알고 있기 때문에 이 논증은 건전하지 않다는 것을 안다. 그러나 빛과 전자기파는 동일하다는 맥스웰의 주장이 널리 알려지기 전에 이와 같은 논증이 제시된다면, 맥스웰은 그에 대해서 어떻게 답해야 하는가? 그가 할 수 있는 것은 단순히 공리3)이 옳지 않다고 주장하는 것일 것이다. 즉 맥스웰은 비록 그것이 직관적으로 그럴듯하게 보일지라도 순환논증의 오류를 범하는 옳지 않은 것이라고 주장할 것이다. 또한 어둠의 방의 사유실험은 빛의 성질에 관해 아무것도 실제로 보여주지 못하기 때문에 우리는 지속적인 실험을 통해서 어떤 조건 하에서 전자기파가 빛을 만들어내는지를 확인해야 한다고 주장할 것이다. 그리고 맥스웰은 자신의 실험이 그러한 확인을 가능하게 한다고 주장할 것이다. 처칠랜드 부부의 핵심은 어둠의 방 사유실험을 통해서 제시되는 논증은 타당하지만 세 번째 전제

가 순환논증의 오류를 범해 옳지 않고 따라서 건전하지 않은 것처럼, 써얼의 중국어방 사유실험을 통해서 제시되는 논증의 첫 번째 전제가 잘못이고 건전하지 않다는 것이다.

그러나 써얼의 중국어 방 논변이 계산주의에 대한 적절한 비판일 수 없다는 처칠랜드 부부의 유비논증은 받아들일 만한 것인가? 우선 그들의 유비논증은 써얼의 논증을 효과적으로 물리치는 데 도움이 되지 않는 것 같다. 왜냐하면 써얼의 논변을 형식화하면, 다음과 같이 될 것이다.

공리1) A는 B이다.
공리2) C의 성질은 D이다.
공리3) B는 그 자체로 D에 충분하지 않다.
결 론) A는 C에 충분하지 않다.

써얼이 주장하는 것은 위와 같이 형식화될 수 있는 논증이 모두 건전하고 설득력있는 논증이라고 말하고 있는 것이 아니다. 따라서 처칠랜드 부부의 반례는 위와 같은 논증 중에 건전하지 않은 논증이 있다는 것을 보일 뿐, 이러한 논증이 형식적으로 부당하다거나 이러한 형식의 논증이 모두 건전하지 않다는 것을 보이는 것은 아니다.

그럼에도 처칠랜드 부부의 지적처럼 "구문론은 의미론에 충분하지 않다"는 전제로부터 "프로그램은 마음에 충분하지 않다"는 결론을 이끌어 내는 것은 순환적인 것처럼 보인다. 그리고 인공지능 연구가들이 보이고자 하는 것이 구문론이 의미론의 충분조건일 수 있다는 것이라면, 결국 문제는 그 시도가 성공적일 수 있는가일 것이다. 즉 맥스웰이 어둠의 방 논변의 공리3)을 물리칠 수 있었던 것처럼 중국어 방 논변의 공리1)을 물리칠 수 있는가이다. 그러나 써얼이 중국어 방이라는 사유실험을 통해서 그 물음에 대해 제공하는 대답은 명백하게 부정적이고, 그 대답의 배후에는 언어의 의미론적 통찰이 들어 있다.

기호체계 모델은 원자기호, 즉 계산토큰의 조작에 의해서 특징지워진

다. 그러한 계산토큰이 어떻게 내용을 갖게 되는가? 우리는 어떤 토큰이 내용을 갖는다고 해석할 수 있지만 그러한 해석은 본래적(intrinsic) 의미를 보증할 수 없기 때문에, 해석내용(interpretational content)만으로 충분하지 않다. 그것은 관찰자에 의해 부여진 외재적(extrinsic) 내용일 뿐이다. 또한 우리는 그러한 토큰들이 외부세계의 대상들과 인과적 관계에 의해서 내용을 자연스럽게 갖게 된다고 주장할 수 없다. 결국 우리는 계산토큰이 내용을 가질 수 있기 위해서어떤 본래적 속성을 찾아야만 한다. 그런데 토큰들은 내재적(internal) 구조를 갖지 않은 원자기호들 즉 성질없는 덩이(featureless chunk)일 뿐이다. 어떤 토큰에는 다른 토큰과 구별될 수 있도록 하는 임의의 레벨이 붙여질 수 있다. 예컨대 ELEPHANT라는 토큰은 그것이 사과의 개념이 아니라 코끼리의 개념과 더 밀접하게 관련되게 하는 아무런 내재적 속성이나 구조를 가지고 있지 않다. 이 점이 바로 써얼의 논증의 배후에 있는 통찰일 것이다. 중국어방에 있는 영어 사용자에게 중국어 기호가 아무런 의미론적 내용도 주지 못하는 것처럼, 컴퓨터에게 원자적 토큰은 아무런 의미론적 내용을 부여하지 못한다. 써얼의 "구문론은 의미론을 위해 충분하지 않다"는 주장은 그러한 토큰들에 대한 구문론적 조작은 결코 그 토큰에 참된 의미를 부여할 수 없다는 것을 효과적으로 주장하는 것이다. 그리고 이러한 주장은 기호체계 모델 즉 계산주의에 강력하게 적용될 수 있을 것이다.

이제 써얼의 논변이 연결주의에 대한 성공적인 비판이 될 수 있는지 살펴보자.[20] 연결주의의 기본 아이디어는 알고리즘 과정은 비의미론적인 하위단계이고 이 단계로부터 의미론적 단계가 창발된다는 것이다. 즉 표

20) 써얼의 논증이 연결주의에 대한 성공적인 비판일 수 없다는 주장은 찰머스와 하나드에 의해서 설득력있게 제기되고 있다. D. Chalmers, "Subsymbolic Computation and the Chinese Room" In *The Symbolic and Connectionist Paradigms: Closing the Gap* (ed.) J. Dinsmore (Lawrence Erlbaum, 1992)과 S. Harnad, "Minds, Machines and Searle" *Journal of Theorectical and Experimental Artificial Intelligence* (1989) 1과 "The Symbol Grounding Problem" *Physica D*, (1990) 42 참조. 또한 최근 최훈은 써얼의 주장이 연결주의에 대한 성공적인 반론일 수 없음을 보인 흥미로운 글을 쓴 바 있다. 최 훈, "중국어 방 속의 대화: 설, 계산주의자, 연결주의자" (미발표)

상의 단계와 계산 단계는 구별된다. 연결주의에서 계산 단계의 토큰은 결절과 연결이다. 이것들은 분명히 기본적이고 성질없는 덩이일 뿐이다. 그러므로 그러한 토큰들은 아무런 내재적인 의미론적 내용도 갖지 않는다. 그러나 이것은 연결주의자들에게 아무런 문제가 되지 않는다. 계산주의와는 달리 연결주의자들에게는 이 토큰들은 구문론적 대상으로만 간주되기 때문이다. 결국 써얼의 논증은 이러한 계산 토큰에 적용될 때 우리에게 새로운 것은 아무것도 말해주지 않는 셈이다.

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.[21] 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.

써얼은 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. 어떤 체계가 구문론적이라 함은 그 체계가 어떤 단계에서든 규칙을 따르기에 의해 기능한다는 것이다. 그리고 어떤 체계가 의미론적 내용을 갖는다고 할 때 의미는 내재적, 내포적(intensional) 내용이다. 결국 써얼의 입장은 전적으로 규칙 따르기 행위로 구성되는 체계는 내재적 내용을 가

21) 이러한 연결주의의 표상을 다이어는 미시의미론(microsemantics)라고 부르고, 그것은 체계적으로 표상의 의미를 반영하는 내재적 패턴이라고 설명한다. M. G. Dyer, "Distributed Symbol Formation and Processing in Connectionist Networks" *Journal of Experimental and Theoretical Artificial Intelligence* (1990) 2, 215-239쪽 참조.

질 수 없다는 말로 요약될 수 있을 것이다.

그러나 이러한 주장은 인간의 뇌를 생각해 볼 때 잘못임을 알 수 있다. 즉 인간의 뇌는 엄격한 규칙 -물리적 법칙-을 따르는 것으로 기술될 수 있지만, 내재적 의미를 갖는다. 써얼의 주장은 언어학으로부터 통찰을 얻은 것같다. 단어가 따르는 구문규칙은 그 자체로는 단어에 의미 내용을 부여하지 못한다. 문장의 경우에 구문론이 의미론을 위해서 충분하지 않다는 것은 분명하다. 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 써얼의 주장은 "어떤 단계에서의 구문론은 같은 단계에서의 의미론적 내용을 위해서 충분하지 않다"[22]고 수정되어야 할 것이다. 이러한 수정은 언어학과 관련한 우리의 직관과 부합한다. 예컨대 "고양이가 매트 위에 앉아 있다"는 문장의 명사들과 동사를 형식적으로 조작하는 것이 곧 우리에게 그러한 단어들의 의미를 제공해주는 것은 아니다. 마찬가지로 중국어 방에 있는 중국어 기호를 형식적으로 조작할 수 있다는 것이 그 기호들의 의미를 이해하는 것을 반드시 함축하는 것은 아니다. 결국 구문론과 의미론이 같은 단계에 존재한다면 구문론적 조작의 대상, 즉 원자적 계산 토큰과 의미론적 해석의 대상이 같은 단계에서 발생하기 때문에, 원자적 토큰들의 조작은 그 토큰들이 의미를 부여하기에 충분하지 않다는 것이 써얼의 논증의 기본적인 직관인 것이다.

이러한 주장이 옳다면 써얼의 논증은 구문론과 의미론을 같은 단계에서 발생하는 것으로 간주하는 계산주의에 대해서는 강력한 비판으로 적용될 수 있지만, 구문론과 의미론이 같은 단계에서 발생하지 않는다고

22) 찰머스의 앞의 논문, 17쪽.

설명하는 연결주의에 대해서는 적용될 수 없을 것이다.

우리는 이제 표상의 근거지우기, 다시 말해서 계산체계에서 표상이 어떻게 참된 의미를 가질 수 있는가를 물어야 한다. 기호 근거지우기(symbol grounding)의 작업은 계산이란 무의미한 기호들의 조작이라는 생각으로부터 출발한다. 이 기호들의 의미는 관찰자에 의해서 그 기호에 투사될 수 있을 뿐이다. 기호들이 의미론적 내용을 갖기 위해서는 그 기호들은 비기호적 토대 위에서 근거되어야 한다. 여기서 주의할 점은 기호라는 이름으로 "표상"과 "계산 토큰"이라는 분명히 구별되는 두가지가 있다는 사실을 잊지 않아야 한다는 것이다. 우리가 계산은 무의미한 기호들의 조작이라고 말할 때 우리는 계산토큰에 대해서 말하고 있는 것이고, 기호들이 어떻게 근거지워지는가라고 물을 때 우리는 본질적으로 표상에 대해서 말하고 있는 것이다. 즉 계산 토큰이 근거지워지지 않는다는 사실이 곧 표상이 근거지워지지 않는다는 사실을 함축하지는 않는다. 결국 컴퓨터는 형식적인 기호조작을 하고 있는 것이고 이 기호들은 무의미하기 때문에 컴퓨터는 이해할 수 없다는 써얼의 논증은 '기호'라는 이름 하에 표상과 계산토큰을 혼동하고 있기 때문에 생겨난 것이다.

표상이 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 작업할 것이다. 이것은 계산체계를 외부 세계와 연결시키는 작업으로서 표상을 곧바로 그 지시대상과 연결하는 것이다. 즉 참된 표상은 외부세계와 감각기관의 만남에서 근거된다고 설명한다. 예컨대 DOG라는 표상이 외부세계에 실제로 개가 등장함으로써 촉발된다면, 우리는 그 표상은 실제로 그것의 외부적 지시대상으로서 개를 갖는다고 주장할 수 있을 것이다. 그러나 우리가 내재적 의미론에 관심을 갖는다면, 우리는 내재적 근거지우기(internal grounding)의 작업할 것이다. 즉 우리의 표상이 내재적 내용을 담지하기 위해서 충분한 내재적 구조를 갖는다는 것을 확증하는 것이다. 내재적 근거지우기의 목표는 기본적인 계산 토큰보다 풍부한 표

상 내용의 담지자(vehicle)를 찾는 것이다. 연결주의는 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.

그렇다면 행위패턴으로서의 표상이란 무엇인가? 포더(J. Fordor)는 마음이란 구조된 표상에 대해서 구문론적으로 작동하는 계산체계일 뿐이라고 말한다[23]. 그리고 구조된 표상은 선천적인 원자적 개념을 그 구조의 기초 단위로 갖고, 그 기초단위가 바로 사유언어(language of thought)의 기초를 형성한다고 설명한다. 그러나 홉스타터(D. R. Hofstadter)와 같은 연결주의자들은 "햄버거"나 "나"와 같은 형식적 토큰은 그 자체로는 공허하며, 그것은 아무것도 지시하지 않는다고 말한다.[24] 이러한 주장은 써얼의 연결주의에 대한 비판과 유사하지만, 홉스타터는 이러한 주장이 기본적으로 인간의 마음을 계산적 정보처리 과정으로 간주하는 모든 계산주의적 인공지능 연구의 실패를 함축하기보다는 좁은 의미의 계산주의에 대한 실패일 뿐이고 따라서 이것은 우리로 하여금 연결주의로 나아가게 하는 동기가 된다고 주장한다. 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 홉스타터는 능동적인 기호를 구상하는데, 그것은 그 자체로 의미를 담지하는 표상이다. 표상은 프로그램에 의해서 형식적으로 조작되는 것이 아니라 보다 낮은 단계의 계산 조작으로부터, 즉 계산기저

23) J. Fordor, *The Language of Thought* (Cambridge: Harvard University Press, 1975), 64쪽.
24) D. R. Hofstadter, "Waking up from the Boolean Dream or Subcognition as Computation" In *Metamagical Themas* (New York: Basic Books, 1985) 645쪽.

(computational substrate)로부터 통계적으로 창발하는 것이다. 그러므로 표상은 분산된 패턴의 행위 속에서 능동적인 내재적 구조를 갖는다. 비록 연결주의자들이 말하는 이러한 능동적 기호에 대한 충분한 설명이 아직 이루어지지 못하고 있다고 할지라도, 패턴으로서의 표상 개념은 궁극적으로 의미론적 내용의 비밀을 밝힐 수 있으리리고 기대하는 것이다.

6. 결 론

써얼은 인공지능의 연구를 한 가지 종류, 즉 계산주의 모델로만 생각하여 자신의 논증이 모든 인공지능 연구에 적용될 수 있다고 주장한다. 그러나 모든 인공지능의 연구 작업의 기본적인 전제는 마음을 하나의 계산, 정보처리 과정으로 보는 것이지만, 그 계산체계에는 분명히 구별되는 두가지 모델이 있다. 지금까지 보인 것처럼 써얼의 논증은 계산주의에는 잘 적용될 수 있지만, 연결주의에는 그렇지 못하다. 즉 써얼의 논증은 기본적인 계산토큰들이 본래적 내용을 담지할 수 없다는 사실에 근거하여 연결주의의 근본적인 문제점을 지적하고 있지만, 그것이 곧 전체 인공지능연구를 비판하는 것일 수 없다. 물론 연결주의에 난점이 없는 것은 아니다. 연결주의에 남겨진 문제는 어떻게 의미론이 보다 낮은 단계에서의 구문론으로부터 창발하는가, 다시 말해서 의미론적 내용이 기계적 기저로부터 어떻게 창발하는가를 구체적으로 설명하는 것이다. 그런 의미에서 찰머스는 "써얼에게 '구문론은 의미론을 위하여 충분하지 않다'고 주장하면서 전체 인공지능 연구를 비판하는 대신, 어떻게 높은 단계의 의미론이 하위단계의 구문론으로부터 창발하는가의 문제에 참여하도록 권유해야 할 것"이라고 말하고 있다.[25] 이 문제가 하루 아침에 해결되지는 않겠지만, 연결주의는 그 문제에 답하는 일련의 역할을 수행하고 있기 때문에 그것이 원칙상 불가능하다고 주장하기 위해서는 써얼은 중국어 방의 논변 외에 어떤 것을 제시해야 할 것이다.[26]

---

25) 찰머스의 앞의 논문, 24쪽.

참 고 문 헌

이승종, "컴퓨터의 언어철학" 연세학술논집 26 (서울, 1997)

이초식, 『인공지능의 철학』 (서울: 고려대학교 출판부, 1993)

최 훈, "중국어 방 속의 대화: 설, 계산주의자, 연결주의자" (미발표)

Becthel, W. "Connecitonism" In *A Companion to the Philosophy of Mind* (ed.) S. Guttenplan (Oxford: Blackwell, 1994).

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

_____ ., *The Conscious Mind* (New York: Oxford University Press, 1996)

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

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

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

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

Fordor, J. *The Language of Thought* (Cambridge: Harvard University Press, 1975).

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

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

Hofstadter, D. R. "Waking up from the Boolean Dream or Subcognition as

---

26) 이 논문은 1999년 한국 학술진흥재단에서 공모한 협동연구의 지원을 받아, 한림대 손병홍 교수, 동해대 심철호 교수와 공동연구를 진행하는 과정에서 얻어진 결과물 중의 하나이다. 이 논문을 쓰는데 두 분의 도움이 적지 않았으나, 이 논문에 실린 내용에 대한 책임은 전적으로 필자에게 있다.

Computation" In *Metamagical Themas* (New York: Basic Books, 1985).

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

Rapaport, W. "Syntatic Semantics: Foundations of Computational Natural Language Understanding" Reprinted in *Aspects of Artificial Intelligence* (ed) J. Fetzer (Dordrecht: Kluwer, 1988)

_____ ., "Machine Understanding and Data Abstraction in Searle's Chinese Room" *Proceedings of the 7th Annual Conference of the Cognitive Science Society*, (University of California at Irvine, Hillsdale Lawrence Erlbaum, 1985)

Searle, J. "Minds, Brains, and Programs" *Behavioral and Brain Sciences (1980) 3. 그리고 Minds, Brains and Science* (Cambridge: Harvard University Press, 1984)

_____ . "Consciousness, Explanatory Inversion, and Cognitive Science" *Behavioral and Brain Sciences*, (1990) 13.

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

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