주의를 뜻한다. ——옮긴이)을 제외하면, 길버트가 사람이고 결국 세계 속에서의 스피치악의 활동이나 자기 표상에 의해 창조된 인물이라는 결론에 우리가 저항하는 근거는 도대체 무엇일까?

〈그렇다면 그 주장은 내가 나의 육체의 꿈이라는 말인가? 나는 나의 육체 활동에 의해서 지어진 일종의 소설 속 가공의 등장 인물에 지나지 않은가?〉 이것도 문제의 답에 도달하는 한 가지 방법이기는 하지만, 여러분 자신을 허구라고 부르는 까닭은 무엇인가? 여러분의 뇌는 의식이 없는 소설 창작 기계와 마찬가지로 철커덕거리면서 움직임을 계속하고, 육체적인 활동을 하고, 그 결과에 대해서는 아무런 고려도 없이 입력과 출력들을 처리해 나간다. 「전주곡——개미의 푸가」에 등장하는 힐러리 아주머니를 구성하는 개미들과 마찬가지로 그것은 처리 과정 속에서 여러분을 창조한다는 사실을 〈알지〉 못한다. 그러나 여러분은 그 광란적인 활동으로부터 거의 마술적으로 창발해서 그 속에 있는 것이다.

다른 수준과 융합되어 있는, 상대적으로 의식이나 이해가 존재하지 않는, 여러 가지 활동으로부터 한 수준의 자아를 창조하는 이러한 과정은 설의 다음 글에서 생생하게 예시될 것이다. 그러나 그는 자신이 보여주는 이러한 전망에 대해 단호하게 저항한다.

D. C. D.

마음, 뇌, 프로그램

존 설

사람의 인식 능력을 컴퓨터로 시뮬레이트하려는 최근의 시도에 대해 어떠한 심리학적 · 철학적 의의를 부여해야 할 것인가? 이 물음에 대한 답을 구하기 위해서는 내가 〈강한strong〉 인공 지능AI 연구라고 부르는 것과 〈약한weak〉 또는 〈신중한〉 인공 지능 연구라고 부르는 것을 구별하는 편이 유용할 것이다. 약한 인공 지능 연구의 입장에 따르면 마음의 연구에서 컴퓨터가 갖는 주된 가치는, 그것이 우리에게 매우 강력한 도구를 제공한다는 것이다. 예를 들어 컴퓨터를 통해 좀더 엄밀하고 엄격하게 가설을 정식화하고 검증할 수 있게 된다는 것이다. 그에 비해 강한 인공 지능 연구의 입장에서 컴퓨터는 더 이상 단순한 마음 연구의 도구가 아니다. 오히

* John R. Searle, "Minds, Brains, and Programs," *The Behavioral and Brain Sciences,* vol. 3. (Cambridge University Press, 1980). 존 설은 미국의 철학자이다.

러 제대로 프로그램된 컴퓨터는 실제로 마음이다. 그 컴퓨터에 올바른 프로그램을 주면 문자 그대로 사물을 이해하고, 그 밖의 인지적 상태를 갖는다는 의미에서 말이다. 또한 강한 인공 지능 연구에서는 프로그램된 컴퓨터가 인지적인 상태를 가지기 때문에 프로그램은 심리학적 설명을 검증할 수 있게 해주는 도구에 그치지 않고 프로그램 자체가 설명인 것이다.

적어도 이 논문에 국한되는 한 나는 약한 인공 지능 연구의 주장에 대해서는 이론(異論)을 제기하지 않는다. 이 글에서 나는 강한 인공 지능로 규정된 주장들, 다시 말해 적절하게 프로그램된 컴퓨터가 문자 그대로 인지적(認知的)인 상태들을 가지며, 또한 그에 의해 프로그램이 사람의 인지를 설명한다는 주장에 대해 논의를 전개할 것이다. 그러므로 이 글에서 앞으로 인공 지능 연구라고 지칭하는 것은 앞에서 이야기한 두 가지 주장을 통해 표현된 강한 인공 지능 연구를 가리키는 것이다.

나는 로저 샌크Roger Schank와 예일 대학의 그의 동료들이 추진한 연구를 고찰할 것이다. 왜냐하면 그들의 연구는 인공 지능에 관한 비슷한 주장 중에서 내게 친숙하고, 또한 앞으로 검토하게 될 연구에 대해 아주 분명한 사례를 제공하기 때문이다. 그러나 여기에서 이야기되는 내용이 프로그램의 세부 사항에만 의존하는 것은 아니다. 같은 논의가 위노그라드의 SHRDLU, 요제프 바이첸바움의 ELIZA, 그리고 실질적으로 튜링 머신에 의해 사람의 지적 현상을 시뮬레이트하는 모든 사례에 적용될 수 있을 것이다(「더 깊은 내용을 원하는 사람들에게」의 절의 참고 문헌을 보라).

여러 가지 세부 사항을 밀어두고 개괄적으로 이야기하자면, 샌크의 프로그램은 사람이 이야기story를 이해하는 능력을 시뮬레이트하는 것을 목표로 삼는다고 할 수 있다. 사람들의 스토리 이해 능력의 특징은 사람들이 스토리에 대한 여러 가지 질문을 받았을 때, 그 스토리 속에서 질문에 대한 정보가 분명히 나타나지 않을 경우에도 대답할 수 있다는 점이다. 예를 들어 당신에게 다음과 같은 스토리가 주어졌다고 하자. 〈어떤 남자가 식당에 가서 햄버거를 주문했다. 그런데 정작 나온 햄버거는 너무 바삭바삭하게 구워졌다. 그 남자는 잔뜩 화가 나서 햄버거 값도 내지 않고 팁도 주지 않은 채 식당을 뛰쳐나왔다.〉 그렇다면 〈그 남자는 햄버거를 먹은 것인가?〉라는 질문을 받으면 당신은 〈아니오, 먹지 않았습니다〉라고 대답할 것이다. 마찬가지로 다음과 같은 이야기를 들었다고 하자. 〈어떤 남자가 식당에 가서 햄버거를 주문했다. 햄버거가 나왔을 때 그는 대단히 만족했다. 그리고 식당을 나가면서 계산을 하기 전에 종업원에게 팁을 듬뿍 주었다.〉 그리고 〈그 남자는 햄버거를 먹었는가?〉라는 질문을 받았다고 하자. 그러면 당신은 필경 〈예, 그는 햄버거를 먹었습니다〉라고 대답할 것이다. 그렇다면 샌크의 컴퓨터들도 식당에 대한 이런 질문에 대해서 비슷한 방식으로 대답할 수 있을 것이다. 그렇게 하기 위해서 이 컴퓨터들은 사람이 식당에 대해 가지는 것과 같은 종류의 정보 〈표상representation〉을 갖고 있어야 한다. 그래야만 그러한 종류의 스토리가 제시되었을 때 위와 같은 질문에 대답할 수 있을 것이다. 컴퓨터에게 스토리를 주고 질문을 하면 컴퓨터는 비슷한 스토리를 들려주었을 때 사람이 할 것으로 기대되는 대답을 할 것이다. 강한 인공 지능 지지자들은 이러한 질문과 답변의 연속sequence에서 컴퓨터는 단지 사람의 능력을 시뮬레이트하는 데 그치지 않고, (1) 스토리를 문자 그대로 이해한 뒤 질문에 대답한다고 말할 수 있다는 것, (2) 컴퓨터와 그

프로그램은, 사람이 스토리를 이해하고 그와 연관된 여러 가지 질문에 대답하는 능력을 설명할 수 있다고 주장한다.

그러나 이러한 두 가지 주장은 샌크의 연구에 의해 전혀 뒷받침되지 않는 것처럼 보인다. 따라서 나는 이 글의 나머지 부분에서 그 점을 증명하려고 시도할 것이다(그렇다고 해서 샌크 자신이 이 두 가지 주장을 스스로 옹호하려 했다는 말을 내가 하려는 것은 아니다). 마음에 대한 모든 이론을 테스트하는 한 가지 방법은, 그 이론이 모든 정신 활동의 기반이라고 생각하는 원리에 따라서 실제로 자신의 마음도 작동하고 있다면 그것은 도대체 어떤 것인가라고 스스로에게 묻는 것이다. 이 검사 방법을 다음과 같은 사고 실험 Gedankenexperiment을 통해 샌크의 프로그램에 적용시켜 보자. 가령 내가 어떤 방에 갇혀 있고, 중국어로 된 커다란 책이 한 권 주어졌다고 하자. 그리고 나는 중국어를 전혀 몰라서 읽을 수도 말할 수도 없다고(실제로도 그렇지만) 하자. 심지어 나는 중국어로 쓰인 글이 중국어인지 일본어인지, 아니면 아무런 뜻도 없는 곡선인지조차 식별할 수 없다고 하자. 따라서 내게 중국어로 씌어진 글자는 단지 뜻없는 곡선들의 무더기에 불과한 것이다. 그러면 이번에는 한 발 더 나아가서, 이 첫번째 중국어 책이 주어진 후, 다음 두 번째 책은 첫번째 책과 두번째 책의 상호 연관에 대한 규칙 집합과 함께 주어졌다고 하자. 그 규칙은 영어로 적혀 있기 때문에 나는 영어를 모국어로 삼는 다른 사람들과 같은 정도로 그 규칙을 이해할 수 있다. 그 규칙 덕분에 나는 한 집합의 형식 기호를 다른 형식 기호와 연관시킬 수 있게 되었다. 또한 여기에서 〈형식〉이라는 말은 내가 기호를 그 형태에 의해 완전히 식별할 수 있다는 것을 의미한다. 이번에는 세번째, 즉 중국어 기호로 씌어진 책과 영어로

씌어진 지시가 함께 주어졌다고 하자. 그 지시에 의거해 나는 세번째 책의 여러 가지 요소를 앞의 두 책과 관련지을 수 있으며, 특정 형식의 질문에 대해 특정 형식을 갖는 종류의 중국어 기호열(記號列)에 의해 대답하는 방법을 알게 되었다고 하자. 나는 그 사실을 모르지만 내게 이러한 기호를 준 사람들은 최초의 책을 〈스크립트 script〉, 두번째 책을 〈스토리〉, 그리고 세번째 책을 〈질문〉이라고 부른다. 게다가 그들은 세번째 책에 대해 내가 대답할 때 사용하는 기호를 〈질문에 대한 대답〉이라고 부른다. 그리고 그들이 내게 준 영어로 씌어진 규칙의 집합을 〈프로그램〉이라고 부른다. 그러면 이야기를 조금 복잡하게 만들어보자. 그들이 내게 영어로 된 스토리를 주었다고 하자. 그리고 나는 그것을 이해할 수 있다. 그런 다음 그들이 내게 그 스토리에 대한 질문을 영어로 하고 나도 영어로 대답을 한다. 또한 얼마 후 내가 중국어 기호를 지시에 따라 처리하는 데 익숙해졌고, 프로그래머도 프로그램을 작성하는 데 익숙해져서 그 결과 외부의 관점에서, 즉 내가 갇혀 있는 방 바깥에 있는 누군가의 관점에서 볼 때 질문에 대한 나의 대답이 중국어가 모국어인 사람의 대답과 구별할 수 없을 정도가 되었다고 하자. 그렇게 되었을 때 내가 한 대답을 보고 내가 중국어를 한 마디도 할 수 없다고 주장할 수 있는 사람은 아무도 없을 것이다. 가정을 조금 더 진전시키면 영어 질문에 대한 나의 대답은 나 자신이 영어를 모국어로 사용한다는 단순한 이유 때문에 영어가 모국어인 다른 사람의 대답과 구별할 수 없을 것이다. 외적 관점, 즉 나의 〈대답〉을 읽는 방 밖의 누군가의 관점에서 볼 때, 중국어 질문에 대한 대답과 영어 질문에 대한 대답은 똑같이 훌륭하다. 그러나 영어와는 달리 중국어의 경우 나는 내용을 전혀 이해하지 못하고 단순히 형식 기호

를 처리함으로써 대답을 작성한다. 중국어에 관한 한 나는 그야말로 컴퓨터처럼 행동한 것이다. 나는 형식적으로 지정된 요소들에 대해 단순한 계산 처리를 한 것에 지나지 않는다. 중국어에 대해서 나는 단지 컴퓨터 프로그램을 실행한 것에 지나지 않는다.

그런데 강한 인공 지능의 입장에서 제기할 수 있는 주장은 프로그램된 컴퓨터가 스토리를 이해하며, 더욱이 그 프로그램은 어떤 의미에서 사람의 이해를 설명한다는 것이다. 이제 우리는 지금까지의 사고 실험에 비추어 이러한 주장을 검토할 위치에 서게 되었다.

1. 첫번째 주장에 대해서, 앞의 예에서 내가 중국어로 씌어진 스토리를 한 글자도 이해하지 못한다는 것은 지극히 자명할 것이다. 나는 중국어가 모국어인 사람의 그것과 구별할 수 없는 입력과 출력을 가지며, 또한 나는 당신이 원하는 모든 형식적 프로그램을 가질 수 있음에도 불구하고, 나는 여전히 아무것도 이해하지 못한다. 같은 이유로 샌크의 컴퓨터도 중국어이든 영어이든 그 밖의 어떠한 언어이든 간에, 스토리를 전혀 이해하지 못한다. 왜냐하면 중국어의 경우 내가 그 컴퓨터이기 때문에, 또한 내가 컴퓨터가 아닌 경우에도 그 컴퓨터는 내가 아무것도 이해하지 못한 경우에 내가 가진 것 이상을 가질 수 없기 때문이다.

2. 프로그램이 사람의 이해라는 행위를 설명한다는 두번째 주장에 대해서, 우리는 컴퓨터와 그 프로그램만으로는 이해에 충분한 조건을 제공하지 못한다는 것을 알 수 있다. 왜냐하면 컴퓨터와 그 프로그램은 기능할 뿐이며 거기에는 어떤 이해도 개입하지 않기 때문이다. 그러나 컴퓨터와 그 프로그램이 이해에 대해 필요 조건이

나 의미 있는 공헌을 제공한다고 말할 수 있을까? 강한 인공 지능 연구를 지지하는 사람들의 한 가지 주장은 내가 영어로 된 스토리를 이해할 때, 내가 하는 일은 중국어 기호를 조작하는 경우에 내가 하는 일과 정확히 같다는 또는 대동소이하다는 것이다. 내가 이해하는 영어와 이해하지 못하는 중국어의 경우를 구별짓는 것은 단지 어느 쪽이 더 형식적인 기호 조작인가의 차이밖에 없다는 것이다. 그렇다고 해서 내가 강한 인공 지능의 주장이 잘못임을 증명했다는 뜻은 아니다. 그러나 이 주장은 지금까지 우리가 검토한 사례에 비추어볼 때 분명 받아들이기 힘들 것이다. 이러한 주장이 그럴듯하게 보이는 까닭은, 우리가 모국어를 이야기하는 사람과 마찬가지로 입력과 출력을 갖는 프로그램을 작성하는 것이 가능하다고 가정하고, 나아가 그 화자들이 어떤 수준의 기술(記述)에서는 그들 스스로 하나의 프로그램의 실현이 된다고 가정하기 때문이다. 이런 두 가지 가정을 기반으로 우리는, 설령 프로그램이 이해에 대한 모든 것을 설명하지 않더라도 그 일부는 설명할 수 있으리라고 생각할 수 있을 것이다. 나는 그러한 경험적 가능성이 있다고 생각한다. 그러나 지금까지의 논의에서 그것이 참이라고 믿을 이유는 거의 없다. 왜냐하면 앞의 사례에서 시사된 (증명되지 않은 것은 확실하지만) 것은 컴퓨터 프로그램이 스토리에 대한 나의 이해와 무관하다는 점이다. 중국어의 경우 인공 지능이 프로그램을 통해 내게 입력시켜 주는 모든 것을 받는다 해도, 나는 여전히 아무것도 이해하지 못한다. 영어의 경우 나는 모든 것을 이해하지만 지금까지의 논의에서 나의 이해가 컴퓨터 프로그램, 즉 순수하게 형식적으로 지정된 요소에 대한 계산 처리와 어떤 관계가 있다고 가정할 하등의 이유도 없다. 프로그램이 순전히 형식적으로 규정되는 요소에

대한 계산 처리의 측면에서 정의되는 한, 앞의 사례가 시사하는 것은 이러한 처리 자체가 이해와 어떤 흥미 있는 관계도 맺지 않는다는 것이다. 따라서 그것들은 분명 충분 조건이 아니고, 더욱이 필요 조건이라거나 이해에 어떤 중요한 공헌을 한다고 생각할 근거는 전혀 없다. 여기에서 논의의 쟁점이 단지 서로 다른 기계들이 서로 다른 형식 원리에 의거해서 작동하는 경우에도 동일한 입력과 출력을 가질 수 있다는 것이 아니라는 사실에 주목할 필요가 있다. 사실 그것은 전혀 중요한 핵심이 아니다. 오히려 완전히 형식적인 원리를 컴퓨터에 입력하더라도 그러한 원리들은 이해를 구성하는 데 충분치 않다는 것이 핵심이다. 왜냐하면 사람은 아무런 이해도 없이 형식적인 원리에 따를 수 있기 때문이다. 더욱이 이러한 원리가 필요하거나 도움이 된다고 생각할 어떤 이유도 없다. 왜냐하면 내가 영어를 이해할 때, 내가 어떤 형식적인 프로그램을 조작하고 있다고 가정할 아무런 이유도 없기 때문이다.

그렇다면 내가 영어 문장에 대해서는 갖고 있지만, 중국어 문장에 대해서는 갖지 않는 것은 무엇인가? 이 물음에 대한 분명한 답은 내가 전자의 의미를 이해하는 반면, 후자인 중국어 문장의 의미에 대해서는 전혀 알지 못한다는 것이다. 그러나 무엇이 이러한 차이를 구성하는가, 왜 우리는 그 차이를 기계에 공급할 수 없는 것일까, 그리고 도대체 그 차이란 무엇인가? 이 물음에 대해서는 나중에 다시 언급하게 될 것이다. 여기에서는 우선 앞에서 들었던 사례에 대한 논의를 계속하기로 하자.

나는 이 사례를 몇 사람의 인공 지능 연구자들에게 소개할 기회를 가졌다. 그리고 흥미롭게도 그들은 이 물음에 대한 적절한 답변이 무엇인지에 대해 일치된 의견에 도달하지 못했다. 나는 그들로부터 놀랄 만큼 다양한 반응을 얻었다. 나는 지금부터 그들의 반응 중에서 가장 공통된다고 여겨지는 내용을 고찰할 것이다(그리고 그 반응의 지리적 근원에 대해서도 상세히 설명하겠다).

그러나 그에 앞서 나는 〈이해〉에 대한 몇 가지 일반적인 오해를 피하고자 한다. 왜냐하면 이러한 종류의 논의에서 흔히 〈이해〉라는 말이 제멋대로 다루어지는 경향이 나타나기 때문이다. 나를 비판하는 사람들은 이해의 정도가 단일하지 않고, 〈이해〉라는 말이 단지 주어와 목적어만을 갖는 것이 아니며, 또한 이해에는 여러 종류와 수준이 존재하고, 배중률(排中律)이 〈x가 y를 이해한다〉라는 형식의 명제에 직접 적용될 수 없는 경우가 종종 발생하기 때문에 많은 경우 x가 y를 이해하는지 여부의 단순한 사실의 문제가 아니라 판단을 요구하는 문제로까지 발전한다는 것을 지적한다. 이러한 지적에 대해서 나는 〈물론, 물론이다〉라고 대답하고 싶다. 그러나 이러한 지적은 우리의 문제와는 아무런 관계도 없다. 그 까닭은 〈이해〉라는 말이 문자 그대로 적용되는 사례와 그것이 적용되지 않는 사례가 분명치 않기 때문이다. 그리고 이 두 종류의 사례가, 내가 이 논의를 위해 필요로 하는 전부이다. 나는 영어로 씌어진 스토리를 이해한다. 또한 영어만큼은 아니지만 나는 프랑스어로 씌어진 스토리도 이해할 수 있다. 그리고 그보다 더 못하지만 독일어 스토리도 이해할 수는 있다. 그러나 중국어로 된 스토리는 전혀 이해할 수 없다. 반면 내 자동차와 가산기(加算機)는 아무것도 이해할 수 없다. 그것들은 그러한 종류의 일과는 무관하다. 우리는 종종 〈이해〉나 그 밖의 인지적 술어(述語)를 비유나 유추를 통해 자동차, 가산기, 그리고 그 밖의 인공물들의 속성으로 표현하지만 그러한 속성을 증명하는 것은 아무것도 없다. 우리는 〈자동문은 광전(光電) 셀

에 의해 언제 열려야 할지를 안다〉, 〈가산기는 덧셈이나 뺄셈 방식을 알고 있지만(여기에 이해한다는 표현을 써도 무방할 것이다) 나눗셈은 알지 못한다〉, 〈자동 온도 조절 장치는 온도의 변화를 지각한다〉 등의 표현을 사용한다. 우리가 이러한 표현을 사용하는 이유는 매우 흥미롭다. 그리고 이러한 표현은, 우리가 자신의 의도성 intentionality을 인공물에까지 확장시킨다는 사실과 연관된다. 우리의 도구는 우리의 목적을 연장시키는 것이며, 따라서 우리는 그 도구에 대해 의도성을 귀속시키는 것을 자연스럽게 생각한다. 그러나 나는 철학이라는 얼음이 그러한 종류의 사례들에 의해 깨지지 않는다고 생각한다. 자동문이 광전 셀에 의해 〈지시를 이해한다〉는 의미는, 내가 영어를 이해할 때의 〈이해〉의 의미와는 전혀 다르다. 만약 샌크의 프로그램된 컴퓨터가 스토리를 이해한다는 의미가 자동문의 이해와 같은 비유적인 의미이고, 내가 영어를 이해할 때의 의미가 아니라고 한다면, 이 문제는 토론할 가치도 없을 것이다. 그러나 뉴웰과 사이먼은 컴퓨터에 대해 그들이 주장하는 종류의 인지가 사람의 인지와 같은 종류라고 말한다. 나는 그들 주장의 솔직함을 좋아한다. 그리고 앞으로 고찰하게 될 주장도 바로 그런 종류이다. 나는 프로그램된 컴퓨터가 문자 그대로 자동차나 가산기가 이해하는 것을 이해할 뿐이며, 따라서 실제로는 아무것도 이해하지 못한다는 주장을 제기할 것이다. 컴퓨터의 이해는(내가 독일어를 이해하는 경우처럼) 부분적이거나 불완전한 것도 아니다. 그것은 제로(0)이다.

그러면 그들의 대답을 들어보자.

1. 시스템 이론의 대답(버클리)

〈방 안에 갇힌 사람이 스토리를 이해하지 못한다는 것은 사실이지만, 실제로 그는 전체 시스템의 일부에 지나지 않으며 시스템 전체로서는 스토리를 이해한다. 그 사람 앞에는 규칙들이 적힌 커다란 장부가 있고, 그는 계산용 종이와 연필, 그리고 중국어 기호 집합이 들어 있는 '데이터 뱅크'를 갖고 있다. 여기에서 이해는 개인에게 귀속되는 것이 아니라 개인을 한 부분으로 삼는 시스템 전체에 귀속된다.

시스템 이론에 대한 나의 답변은 매우 간단하다. 그 개인에게 시스템의 모든 요소들을 내면화시켜 보자. 그러면 그는 장부에 규칙들을 메모하고, 데이터 뱅크에 중국어 기호를 기억시켜서 모든 계산을 머리 속에서 하게 된다. 이렇게 되면 그 개인은 전체 시스템을 하나로 통합시켜서 그가 시스템에 포함시키지 않는 것은 아무것도 없게 된다. 더욱이 우리는 그 방을 제거해서 그가 옥외에서 일하고 있다고 상상할 수도 있다. 그러나 이 경우에도 여전히 그는 중국어를 전혀 이해하지 못하며 그 시스템은 더욱 그러하다. 왜냐하면 그에게는 없지만 시스템에는 있는 것이 아무것도 없기 때문이다. 만약 그가 이해하지 못한다면 시스템 역시 이해할 어떤 방도도 없게 된다. 그 시스템은 그의 일부에 불과하기 때문이다.〉

시스템 이론은 처음부터 내게 받아들이기 힘든 것으로 여겨졌기 때문에 이 이론에 대해 이 정도의 답변을 하는 것만으로도 나는 얼마간의 당황스러움을 느낀다. 이 견해는, 한 개인은 중국어를 이해하지 못하지만, 그 개인과 종잇조각의 결합이 중국어를 이해할지도 모른다는 사고 방식이다. 나는 특정 이데올로기에 사로잡히지 않은

사람이라면 어떻게 이런 생각을 받아들일 수 있을지 상상하기 힘들다. 게다가 강한 인공 지능의 이데올로기에 빠진 사람들은, 결국 이러한 사고 방식과 매우 흡사한 주장을 제기할 경향이 있다고 나는 생각한다. 그러면 이 문제를 조금 더 검토해 보기로 하자. 이런 사고 방식에 기초한 한 주장에 따르면, 내면화된 시스템 속에 있는 사람은 중국어를 모국어로 삼는 사람이 이해하는 만큼 중국어를 이해하지는 못하지만(왜냐하면, 예를 들어 그는 그 스토리가 레스토랑이나 햄버거 등을 언급한다는 사실을 모르니까), 〈형식 기호 조작 시스템으로서의 사람〉은 〈실제로 중국어를 이해한다〉. 여기에서 중국어의 형식 기호 조작 시스템으로서 그 사람의 하위 체계와 영어에 대한 하위 체계가 혼동되어서는 안 된다.

따라서 실제로 그 사람 속에는 두 개의 하위 체계, 즉 영어를 이해하는 하위 체계와 중국어를 이해하는 하위 체계가 있으며, 〈두 시스템은 서로 거의 아무런 관계도 없다〉. 그러나 나는 이러한 견해에 대해 그 시스템들이 서로 거의 관계가 없을 뿐 아니라 조금도 닮지 않았다고 대답하고 싶다. 영어를 이해하는 하위 체계는 (앞으로 얼마간 이 〈하위 체계〉라는 전문 용어를 사용해서 논의를 계속하기로 하자) 스토리가 레스토랑이나 햄버거를 먹는 일을 다루고 있다는 것을 알고 있으며, 또한 레스토랑에 대한 질문을 받고 있고, 레스토랑에 대한 질문에 대해 스토리의 내용 등을 기초로 여러 가지 추론을 해서 가능한 한 최선의 대답을 하고 있다는 사실도 알고 있다. 그러나 중국어 하위 체계는 그러한 사실들을 전혀 모른다. 영어 하위 체계가 〈햄버거〉라는 말이 음식물인 햄버거를 가리킨다는 것을 알고 있는 데 비해, 중국어 하위 체계가 알고 있는 것은 〈꼬부랑 곡선들〉 다음에 〈꼬부랑 곡선들〉이 계속된다는 것뿐이다. 그

가 아는 것은 이 시스템의 한쪽 끝에 여러 가지 형식 기호들이 도입되고, 그런 다음 영어로 씌어진 규칙에 따라 그 기호에 조작이 가해지고, 그 결과 다른 쪽 끝에서 다른 기호들이 나타난다는 것이 전부이다. 우리가 처음에 검토했던 사례에서 제기하려고 했던 주장의 핵심은, 중국어를 전혀 이해하지 못하면서도 꼬부랑 곡선들 다음에 꼬부랑 곡선들을 계속 쓸 수 있다는 이유만으로 이러한 기호 조작이 그 자체로서 중국어를 이해한다고 하기에는 충분치 않다는 것이었다. 또한 그 사람들 사이에서 하위 체계의 존재를 가정한다고 해도 이러한 논의를 만족시키지 못한다. 왜냐하면 그 하위 체계들도 최초의 예에서의 사람보다 별반 나은 처지가 아니기 때문이다. 다시 말해서 그 하위 체계들은 영어로 말하는 사람(또는 하위 체계)이 포함하는 비슷한 것도 갖고 있지 않기 때문이다. 앞에서 서술한 사례에서 중국어 하위 체계는 영어 하위 체계의 일부, 즉 영어 규칙에 따라 무의미한 기호 조작에 관여하는 일부분에 불과하다.

그러면 맨 처음 무엇이 그 시스템들을 촉발시켰는지(어떤 동기를 주었는지) 우리 자신에게 물음을 제기해 보기로 하자. 그 물음은 다음과 같다. 기호 조작을 하는 사람이 자기 내부에 중국어로 된 스토리를 문자 그대로 이해하는 하위 체계를 갖고 있는 것이 분명하다고 말하려면 어떤 〈독립적인〉 근거가 존재한다고 가정해야 할 것인가? 내가 아는 범위 내에서의 유일한 근거는, 앞에서 이야기했던 사례에서 내가 중국어를 모국어로 삼는 사람과 같은 입력과 출력을 가지며, 입력과 출력을 연결시키는 프로그램을 갖고 있다는 것이다. 그러나 앞에서 언급한 사례들의 요점은 사람, 즉 사람을 구성하는 시스템 집합은 입력, 출력, 프로그램으로 이루어진 정확

한 조합을 가질 수 있지만 내가 영어를 이해한다는 문자 그대로의 의미에서는 아직 아무것도 이해하지 못한다는 의미에서 이해에는 불충분하다는 것을 보여주려고 시도한 것이다. 여기에서 이해란 내가 영어로 스토리를 이해한다고 했을 때의 이해라는 의미이다. 중국어를 이해하는 하위 체계가 내 안에 있는 것이 〈틀림없다〉고 말할 때의 유일한 동기는, 내가 그 프로그램을 갖고 있고 튜링 테스트를 통과할 수 있다는 것이다. 다시 말해 나는 중국어가 모국어인 사람을 속일 수 있다. 그러나 이 튜링 테스트의 타당성이 우리 논의의 핵심 중 하나이다. 앞에서 언급한 사례들은 튜링 테스트를 통과하는 두 개의 〈시스템〉이 있을 수 있다는 것을 보여주었다. 그러나 문자 그대로의 의미를 이해하는 것은 하나뿐이다. 양쪽 모두 튜링 테스트를 통과했기 때문에 둘 다 이해하고 있는 것이 분명하다는 주장은 이 문제에 대한 논의에서는 통용되지 않는다. 왜냐하면 그와 같은 주장은 나의 내부에 있는 영어를 이해하는 시스템이 단지 중국어를 처리하는 시스템보다 훨씬 많은 것을 포함한다는 주장에 대항할 수 없기 때문이다. 요약하자면 시스템 이론은, 시스템이 중국어를 틀림없이 이해한다는 논증 없이 주장을 제기함으로써 미리 논점을 옳은 것으로 가정해 놓고 주장을 펼치는 논점 선취의 오류를 범하고 있다.

더욱이 시스템 이론의 주장은 지금까지 언급한 측면 이외에도 터무니없는 결론에 도달하는 것처럼 보인다. 만약 내가 어떤 종류의 입력, 출력, 그리고 그것들을 연결짓는 프로그램을 갖고 있다는 근거로 내 속에 인지(認知)가 존재하는 것이 틀림없다는 결론을 내린다면, 모든 종류의 비인지적인 noncognitive 시스템도 인지적으로 될 수 있을 것이다. 예를 들어 내 위(胃)가 정보 처리를 한다고 기술할 수 있는 수준이 있고, 그러한 예가 될 수 있는 많은 컴퓨터 프로그램이 있지만 그렇다고 해서 위가 이해를 가진다고 말할 필요는 없다고 생각한다. 그러나 시스템 이론의 주장을 받아들인다면 위·심장·간장 등이 모두 이해를 갖는 하위 체계라고 말하지 않을 수 없게 된다. 왜냐하면 중국어 하위 체계가 이해한다는 것과 위가 이해한다는 것을 구별할 수 있는 어떤 원칙적인 방법도 없기 때문이다. 중국어 시스템은 정보를 입력과 출력이라는 형식으로 갖지만, 위는 음식물과 음식물을 소화시킨 것으로 입력과 출력을 갖는다는 식의 주장으로는 아무런 해결도 되지 않는다. 기호 조작을 하는 사람의 관점, 즉 나의 관점에서 볼 때 음식물이든 중국어든 그 속에는 아무런 정보도 없기 때문이다. 다시 말해서 중국어는 단지 의미 없는 수많은 꼬부랑 곡선들에 지나지 않다. 중국어의 경우 정보는 프로그래머와 해석하는 사람의 눈 속에만 있다. 그리고 그들이 원한다면 내 소화 기관의 입력과 출력을 정보로 취급하는 것을 방해하는 것은 아무것도 없다.

이 마지막 논점은 강한 인공 지능 연구와 연관된 그 밖의 몇 가지 문제와 깊은 관계를 갖기 때문에 본론에서 벗어나기는 하지만 조금 더 자세히 설명하기로 하자. 강한 인공 지능가 심리학의 한 분야가 되려면 진정한 의미에서 정신적인 시스템과 그렇지 않은 시스템 사이의 구별이 가능해야 할 것이다. 즉 그것을 기초로 마음이 작동하는 원리와 비정신적인 nonmental 시스템을 지배하는 원리를 구별하지 않으면 안 된다. 그렇지 않으면 인공 지능 연구는 마음에 대해 무엇이 구체적으로 정신적인지를 우리에게 설명할 수 없게 된다. 그리고 정신 대 비정신의 구별은 보는 사람 beholder의 눈에 따라 달라지는 것이 아니라 시스템에 고유한 무엇이 되지 않으면

안 된다. 그렇지 않으면 보는 사람에 따라서 사람을 비정신적으로 간주하고, 허리케인을 정신적인 것으로 취급할 수도 있기 때문이다. 그러나 인공 지능 문헌들에서 이러한 구별은 지극히 모호하게 이루어지는 경우가 허다하다. 긴 안목에서 볼 때 이러한 문제점은 인공 지능 연구가 인지에 대한 연구라는 주장을 무색하게 하는 것이다. 존 매카시John McCarthy는 이렇게 쓰고 있다. 〈자동 온도 조절 장치처럼 단순한 기계도 신념belief을 가질 수 있다고 할 수 있고, 신념을 갖는다는 것은 문제 해결 능력을 갖춘 거의 모든 기계의 특징으로 여겨진다.〉 강한 인공 지능 연구가 마음의 이론으로 적용될 가능성을 고려하는 사람들은 이 의견이 갖는 함축성을 신중하게 검토할 필요가 있을 것이다. 그 이론은, 온도를 조절하는 데 사용하는 벽에 걸린 금속 조각들이 우리들이나 우리들의 배우자 또는 아이들이 신념을 갖는 것과 같은 의미로 신념을 갖는다는 것을 강한 인공 지능의 발견으로 받아들일 것을 우리에게 요구하기 때문이다. 나아가 방 안에 있는 다른 〈거의 모든〉 가전 기계들, 즉 전화, 녹음기, 가산기, 전등 스위치 등도 문자 그대로 신념을 갖는다는 것이다. 이 글의 목적이 매카시의 주장에 반박하는 것이 아니므로 여기에서는 논증 없이 다음과 같은 주장을 제기하는 것으로 상세한 논의를 대신하겠다. 마음의 연구는, 사람은 신념을 갖지만 온도 조절 장치나 전화, 가산기 등은 신념을 갖지 않는다는 사실에서 출발한다. 만약 당신이 이 사실을 부인하는 이론에 도달했다 해도, 이미 그 이론에 대한 반증례를 갖고 있기 때문에 그 이론은 틀렸다. 이 대목에서 이런 글을 쓰고 있는 인공 지능 연구자가 실제로는 자신이 하는 말의 의미를 진지하게 받아들이지 않으며, 또한 어느 아무도 진지하게 받아들이지 않는다고 생각하기 때문에 어떻

게든 그 이론을 유지할 수 있는 것이 아닐까라는 생각이 들 수도 있다. 그러나 나는 적어도 잠시 동안은 그 문제를 진지하게 고려해 볼 것을 제안하고 싶다. 다시 말해서 잠시라도 벽에 걸린 금속 조각들이 정말 신념을 갖고 있는지, 사실과의 적합성 여부를 생각하는지, 명제 내용, 그리고 그 명제를 만족시키는 조건을 생각하고 있는지, 강한 신념이나 약한 신념 어느 쪽도 될 수 있는 그런 신념을 갖는지, 신경증이나 불안감 또는 확신에 찬 신념을 갖는지, 독단적이거나 합리적이거나 미신적인 신념을 갖는지, 맹목적 신앙이나 사려 깊은 망설임을 갖는지, 아니 신념이라 부를 만한 무엇을 갖는지 진지하게 생각해 볼 필요가 있다. 자동 온도 조절 장치는 그런 후보가 아니다. 위나 간, 가산기나 전화도 아니다. 그러나 여기에서 중요한 점은, 지금 우리가 강한 인공 지능론자들의 주장을 문자 그대로 진지하게 받아들이고 있기 때문에 그러한 진리가 강한 인공 지능 연구가 마음의 과학이라는 주장에 대해 치명적이라는 사실을 주목할 필요가 있다. 그 주장에 따르면 마음은 도처에 편재하기 때문이다. 우리가 알고 싶은 것은 마음을 자동 온도 조절 장치나 간과 구별시켜 주는 기준이 무엇인가이다. 그리고 매카시가 옳다면, 강한 인공 지능 연구에는 우리에게 그러한 것들을 구분해 줄 희망이 없다.

2. 로봇 이론의 대응(예일 대학)

〈샌크의 프로그램과는 종류가 다른 프로그램을 작성한다고 가정하자. 그리고 로봇 속에 컴퓨터를 넣어 그 컴퓨터가 형식 기호를

입력으로 받아들여 출력으로 내보낼 뿐 아니라 지각하고, 걷고, 돌아다니고, 못을 박고, 먹고, 마시는 식으로 당신이 원하는 모든 일을 시킬 수 있다고 하자. 가령 이 로봇에는 텔레비전 카메라가 장착되어 있어 사물을 볼 수 있고, 팔다리를 갖고 있어서 '행동' 할 수 있다. 그리고 이러한 모든 일이 로봇의 컴퓨터 '뇌'에 의해 제어된다. 이러한 로봇은 샌크의 컴퓨터와는 달리 진정한 이해를 가지며, 그 이외의 심리적 상태를 가질 것이다.〉

이 주장에서 주의를 기울여야 할 첫번째 사항은, 이 이론이 인지란 단순한 형식 기호의 조작이 아니라는 사실을 암묵적으로 인정한다는 점이다. 왜냐하면 로봇 이론은 외부 세계와의 인과 관계들의 집합을 고려에 넣고 있기 때문이다. 그러나 로봇 이론의 주장에 대한 반박으로 그러한 〈지각〉 능력이나 〈운동〉 능력을 부가한다 하더라도 샌크의 원래 프로그램에 특수한 의미에서는 이해, 일반적인 의미에서는 의도성 intentionality이라는 것을 부가할 수 없다는 점을 제기할 수 있다. 이 점을 이해하려면 이 로봇의 경우 앞에서 언급한 사고 실험이 적용된다는 것을 주목할 필요가 있다. 가령 로봇 속에 들어 있는 컴퓨터 대신 앞에서 예로 들었던 중국어의 경우와 같이 당신이 나를 방에 가두고 중국어 기호와 영어 지시를 더 많이 공급하고, 중국어 기호와 중국어 기호를 짜맞추어서 그 중국어 기호들을 외부와 되먹임고리한다고 가정하자. 그리고 내가 모르는 사이에 일부 중국어 기호들이 로봇에 설치된 텔레비전 카메라를 통해 내게 공급되고, 내가 외부로 내보내는 다른 중국어 기호들은 로봇 내부의 모터에 작용해서 로봇의 팔다리를 움직인다고 하자. 여기에서 중요한 것은 내가 하는 모든 일이 형식 기호의 조작에 불과하다는 사실이다. 다시 말해서 나는 다른 사실에 대해서는 전혀 모르는

것이다. 나는 로봇의 지각 장치로부터 〈정보〉를 받고 로봇의 팔다리를 구동시키는 모터에 〈지시〉를 내리지만 그러한 사실에 관해서도 아무것도 모른다. 따라서 나는 로봇 속에 들어 있는 정자미인이다. 물론 정자미인의 원래 의미와는 다르지만 말이다. 어쨌든 나는 로봇에서 무슨 일이 일어나는지 전혀 모르는 셈이다. 나는 기호 조작의 규칙 이외에는 아무것도 이해하지 못한다. 그런데 이 경우 나는, 로봇이 어떤 의도적인 상태도 갖지 않는다고 말하고 싶다. 그것은 전기 배선과 프로그램의 결과로 돌아다니는 것에 지나지 않는다. 더욱이 나는 프로그램을 구체화하는 것에 불과하기 때문에 여기에서 문제가 되는 유형의 어떤 의도성의 상태도 가질 수 없다. 내가 할 수 있는 일이란 형식 기호의 조작에 대한 형식적인 지시에 따르는 것이 전부이다.

3. 뇌 시뮬레이터 이론의 대응(버클리와 MIT)

〈우리가 세계에 대해 갖고 있는 정보, 가령 샌크의 스크립트(대본) 속에 들어 있는 정보를 표현하는 프로그램이 아니라 중국어가 모국어인 사람이 중국어로 스토리를 이해하고 거기에 대한 대답을 할 때 실제로 그의 뇌의 시냅스에서 일어나는 일련의 뉴런 발화를 시뮬레이트하는 프로그램을 설계한다고 생각해 보자. 기계는 중국어 스토리와 그것에 대한 질문을 입력으로 받아들이고, 그 스토리를 처리하는 실제 중국인 뇌의 형식적인 구조를 시뮬레이트해서 중국어로 된 대답을 출력한다. 심지어 우리는, 기계가 단일한 순차적인 프로그램이 아닌 사람의 뇌가 자연 언어를 처리할 때 기능하는