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 시뮬레이트



하는 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 〈$x$가 $y$를 이해한다〉라는 형식의 명제에 직접 적용될 수 없는 경우가 종종 발생하기 때문에 많은 경우 $x$가 $y$를 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

것과 같은 방식으로 병렬 처리하는 프로그램 집합에 의해 기능한다고 상상할 수도 있다. 그럴 경우 우리는, 그 기계가 분명히 스토리를 이해한다고 말해야 할 것이다. 만약 그 점을 인정하지 않는다면, 중국어가 모국어인 사람이 그 스토리를 이해하는 것도 부정해야 하지 않을까? 시냅스 수준에서 컴퓨터의 프로그램과 중국인 뇌의 프로그램은 무엇이 다를까, 또는 무엇이 다를 수 있을까?〉

이 주장에 반론을 제기하기 전에 논의에서 벗어나는 이야기이지만, 나는 이 주장이 인공 지능(또는 기능주의 등)의 어느 학파의 주장으로는 조금 기묘한 대응이라는 것을 지적해 두고자 한다. 나는 강한 인공 지능의 전체적인 개념이 마음의 움직임을 알기 위해서 뇌의 움직임을 알 필요는 없다는 것이라고 생각한다. 이러한 기본적인 가정, 아니 내가 기본적 가정이라고 생각하는 것은 어떤 컴퓨터 프로그램도 다른 컴퓨터 하드웨어에서 실현될 수 있는 것과 마찬가지로 마음의 본질을 구성하고, 뇌의 모든 종류의 다른 과정에서도 실현되는 형식적 요소들을 포괄하는 계산 과정으로 이루어지는 심리적 조작 수준이 존재한다는 것이다. 강한 인공 지능의 가정에 따르면 마음과 뇌의 관계는 프로그램과 하드웨어의 관계에 해당하며, 따라서 우리는 신경생리학을 연구하지 않더라도 마음을 이해할 수 있다. 만약 인공 지능을 연구하기 위해 뇌가 어떻게 작동하는지 반드시 알아야 한다면, 우리는 인공 지능 연구에 매달릴 필요가 없을 것이다. 그러나 인공 지능을 아무리 뇌의 작용에 가깝게 만든다고 해도 이해를 낳는 데에는 충분하지 않다. 이 점을 분명히 하기 위해서 한 가지 언어밖에 할 수 없는 사람이 방 안에서 기호를 조작하는 것을 상상하는 대신 그 사람이 복잡한 송수 파이프들과 거기에 연결된 밸브들을 조작하고 있다고 상상해 보자. 그 사람



이 중국어 기호를 받았을 때 그는 영어로 씌어진 프로그램을 참조해서 어떤 밸브를 열거나 닫을지 아는 것이다. 각각의 물의 관계는 중국인 뇌의 시냅스에 해당하며, 체계 전체는 적절한 모든 발화를 한 다음, 즉 적절한 수도꼭지를 모두 튼 다음 중국어 답변이 파이프들의 출력 말단에서 튀어나오도록 장치되어 있다.

그렇다면 도대체 이 체계의 어디에 이해가 있는 것일까? 이것은 중국어를 입력하고 중국인 뇌의 시냅스의 형식적 구조를 시뮬레이트해서 중국어를 출력한다. 그러나 이 사람이 중국어를 이해하지 못하는 것은 분명하며, 그것은 송수 파이프도 마찬가지이다. 그리고 내 생각으로는 어리석은 것이지만, 사람과 송수 파이프의 결합이 이해를 갖는다는 견해를 받아들이고 싶다면, 이론적으로 그 사람은 송수 파이프의 형식적인 구조를 내면화할 수 있고, 그의 상상 속에서 모든 〈뉴런을 발화〉시킬 수 있다는 점을 기억할 필요가 있다. 뇌의 시뮬레이터에서 제기되는 문제는, 그것이 뇌에 대해 잘못된 것을 시뮬레이트하고 있다는 점이다. 그것이 시냅스에서 일어나는 연속적인 뉴런 발화의 형식적인 구조를 시뮬레이트하는 것에 불과한 한, 그것은 뇌에 대한 중요한 것, 즉 그것의 인과적 특성, 다시 말해서 의도적 상태들을 만들어내는 능력은 결코 시뮬레이트하지 못할 것이다. 그리고 그 형식적인 특성들이 인과적 특성에 충분치 않다는 것은 송수 파이프의 예에서 분명히 드러난다. 그 예에서 우리는 적절한 신경생물학적인 인과적 특성에서 잘려나온 형식적인 특성만을 얻을 뿐이다.

4. 조합설(버클리와 스탠퍼드 대학)

〈지금까지의 세 가지 주장은 중국어 방에 의한 반증례의 반론으로는 충분히 납득할 수 없을지 모르지만, 세 가지 주장을 하나로 합치면 집단적으로 훨씬 큰 설득력을 가지며, 명확한 답변을 얻을 수 있다. 머리 속에 뇌의 모습을 한 컴퓨터를 가진 로봇이 있다고 상상하자. 그리고 그 컴퓨터는 인간 뇌의 모든 시냅스에 대해 프로그램되어 있으며, 그 로봇의 전체 움직임이 인간의 움직임과 구별할 수 없다고 가정하자. 따라서 전체가 통일된 체계이고, 입력과 출력을 가진 컴퓨터처럼 보이지 않는다고 생각하자. 이 경우 우리는 이 체계에 의도성이 있음을 인정하지 않을 수 없다.〉

이 경우 우리가 그 이상의 지식을 갖고 있지 않는 한, 로봇이 의도성을 갖는다는 가설을 받아들이는 것이 이치에 맞고, 또한 실제로 받아들이지 않을 수 없다는 점에는 전적으로 동의한다. 겉모습과 행동을 제외하면 그 결합체의 다른 요소들은 연관성이 없다. 만약 우리가 넓은 범위에 걸쳐 사람의 행동과 구별할 수 없을 정도로 행동하는 로봇을 만들 수 있다면, 약간의 유보 조건을 달아서 그 로봇의 의도성을 인정할 것이다. 우리는 그 컴퓨터의 두뇌가 사람 뇌의 형식적인 유사물이라는 사실을 미리 알 필요는 없을 것이다.

그러나 나는 이것이 강한 인공 지능의 주장에 아무런 지지도 제공하지 못한다고 생각한다. 왜냐하면 강한 인공 지능에 따르면 적절한 입력과 출력을 가진 형식적인 프로그램을 구체화하는 것은 의도성을 갖기에 충분한 조건이고, 실제로 의도성을 구성하는 것이기 때문이다. 뉴엘이 말했듯이 정신의 본질은 물리적 기호 체계의 조작이다. 그러나 이 사례에서 로봇에게 의도성을 인정하는 것은 형식적인 프로그램과 아무런 관계도 없다. 로봇에게 의도성을 귀속시키는 것은 만약 로봇이 겉모습과 행동에서 우리와 매우 흡사하다면, 그 로봇은 우리와 같은 정신적인 상태를 갖고, 그 상태가 행동을 일으키고, 그 행동에 의해 그러한 상태가 표현되며, 따라서 그 로봇은 이러한 심리적 상태를 낳는 내적 메커니즘을 갖고 있는 것이 분명하다는 가정을 전제로 삼고 있다. 따라서 우리가 그런 가정 없이 로봇의 행동을 독자적으로 설명할 수 있는 방법을 알고 있다면 로봇의 의도성을 인정하지 않을 것이다. 로봇이 형식적인 프로그램을 갖고 있다는 사실을 우리가 알고 있는 경우에는 더욱 그러하다. 그리고 이것이 두번째 대응에 대해 내가 제기했던 주장의 핵심이다.

가령 로봇의 행동이, 그 로봇 속에 들어 있는 사람이 로봇의 감각 기관을 통해 해석되지 않은 형식 기호를 수신하고 역시 해석되지 않은 형식 기호를 로봇의 운동 기관에 보내며, 그 사람이 일련의 규칙에 의거해서 기호 조작을 하고 있다는 사실에 의해 완전히 설명된다는 것을 알고 있다고 하자. 더욱이 그 사람은 로봇에 대한 이런 사실들을 전혀 모르고, 그가 아는 것은 어떤 무의미한 기호를 조작하는 방법밖에 없다고 하자. 이 경우 우리는 로봇을 정교한 기계 장치 인형으로 간주할 것이다. 이 인형이 마음을 갖는다는 가설은 아무런 보증도 얻지 못하며 불필요하기도 하다. 왜냐하면 더 이상 로봇 또는 로봇이 그 일부분을 이루는 체계에 의도성을 부여할 어떠한 이유도 존재하지 않기 때문이다(물론 기호를 조작하는 인간의 의도성은 별개의 문제이지만). 형식 기호의 조작이 계속되면서 입력과 출력이 정확히 일치되어도, 유일하게 실재하는 의도성의 중심은 인간이고, 그는 연관된 의도적인 상태들에 대해서 전혀 알지

못한다. 예를 들어 그는 로봇의 눈을 통해 들어오는 것을 보지 않고, 로봇의 팔을 움직일 의도도 없다. 그는 로봇이 듣는 발언, 로봇이 하는 어떤 발언도 이해하지 못한다. 앞에서 언급했던 이유로 인해 인간과 로봇을 부분으로 삼는 체계도 마찬가지이다.

이 점을 분명히 이해하기 위해서 이 경우에 의도성을 인정하는 것이 매우 자연스럽다고 여겨지는 다른 경우, 즉 원숭이와 같은 다른 영장류 또는 개처럼 집에서 기르는 동물의 사례를 비교해 보자. 의도성을 인정하는 것이 자연스럽다고 생각하는 이유는 크게 두 가지이다. 하나는 우리가 동물에게 의도성을 돌리지 않고는 그 행동을 이해할 수 없다는 것이다. 다른 하나는 동물이 우리와 유사한 재료, 즉 눈·코·피부 등으로 이루어져 있다는 것이다. 동물 행동의 정합성과 그 속에 동일한 인과적 소재가 내재한다고 가정하면, 우리는 동물 행동의 근저에 정신적 상태가 있으며, 그 상태는 우리와 비슷한 소재로 이루어진 기구에 의해 만들어질 것이라는 두 가지 가정을 할 수 있다. 별다른 이유가 없는 한 우리는 로봇에 대해서도 비슷한 가정을 하게 될 것이다. 그러나 그 행동이 형식적인 프로그램의 산물이며, 물질의 실질적인 인과적 성질은 의미가 없다는 사실을 아는 순간 우리는 즉시 의도성 가정을 폐기시킬 것이다.

내가 들었던 사례에 대한 반론은 그 밖에도 두 가지가 더 있다. 이러한 주장은 빈번하게 제기되지만(따라서 논의할 충분한 가치가 있다), 실제로는 핵심을 놓치고 있다.



5. 타자의 마음이라는 대응(예일 대학)

〈다른 사람이 중국어 또는 다른 언어를 이해하고 있는지를 어떻게 알 수 있는가? 오직 다른 사람의 행동을 통해서이다. 그런데 컴퓨터는 행동 테스트를 (이론상으로는) 그들과 마찬가지로 통과할 수 있다. 따라서 다른 사람에게 인지를 인정하려면 이론상 컴퓨터에도 그것을 인정하지 않으면 안 된다.〉

이 주장에는 짧은 몇 마디로도 충분한 답변이 가능할 것이다. 이 논의에서 문제점은, 타인이 인지적 상태를 갖는지 여부를 내가 어떻게 아는가가 아니라, 내가 그들이 인지적 상태를 갖는다는 것을 인정할 때 내가 그들에게 인정하는 것이 무엇인가이다. 이 논의의 요점은, 그것이 계산적 과정process이거나 그 출력일 수 없다는 것이다. 왜냐하면 계산 과정과 출력은 인지적 상태가 없어도 존재할 수 있기 때문이다. 무감각증을 가장하는 것은 이 논의에 대한 대답이 안 된다. 물리과학에서 물리적 대상의 실재성과 인식 가능성을 전제해야 하는 것과 마찬가지로 〈인지과학〉에서는 마음의 실재성과 인식 가능성을 전제한다.

6. 변환 자재(變換自在) 이론many mansions의 대응(버클리)

〈당신의 전체적인 주장은 인공 지능 연구가 아날로그 컴퓨터와 디지털 컴퓨터에 대한 연구에 국한된 것인 양 전제하고 있다. 그러나 그것은 우연한 현재의 기술 수준에 불과하다. 당신이 의도성의 본질이라고 말한 인과적 과정이 어떤 것이든 간에(당신의 말이 옳다

고 가정할 때의 이야기이지만), 궁극적으로 우리는 이러한 인과적 과정을 갖는 장치를 만들 수 있을 것이고, 결국 그것은 인공 지능이 될 것이다. 따라서 당신의 주장은 인지를 낳고 설명하는 인공 지능의 능력에 대한 것이 아니다.〉

나는 이 주장에 대해 반대하지 않지만, 결국 강한 인공 지능은 인지를 인공적으로 만들어서 설명하는 것이라고 재정의함으로써 실질적으로는 강한 인공 지능 프로젝트를 중요치 않은 것으로 만들고 있다는 점을 지적해 두고 싶다. 인공 지능을 옹호하기 위해 제기된 원래 주장의 흥미로운 점은 그것이 잘 정의된 정확한 태제, 즉 정신적 과정은 형식적으로 정의된 요소에 대한 계산적 과정이라는 테제이다. 나는 바로 그 테제에 관심을 갖고 문제를 제기해 왔다. 이 주장이 재정의되어서 더 이상 원래의 태제가 아닌 게 되면, 내가 제기한 반론은 더 이상 적용되지 않는다. 왜냐하면 이제 더 이상 검증 가능한 가설이 존재하지 않기 때문이다.

그러면 앞에서 내가 대답하기로 약속했던 문제로 돌아가기로 하자. 즉 최초의 예에서 내가 영어를 이해하고 중국어를 이해하지 못하는 반면 기계는 영어와 중국어를 모두 이해하지 못한다고 가정하자. 그래도 내게는 내가 영어를 이해하는 것을 가능하게 해주는 무언가가 있을 것이고, 내가 중국어를 이해하지 못할 때에는 무언가를 결여하고 있을 것이다. 그렇다면 그 무언가가 어떤 것이든 간에, 왜 그것을 기계에 줄 수 없는 것일까?

나는 이론상으로, 우리가 기계에 영어나 중국어를 이해하는 능력을 줄 수 없는 이유를 알지 못한다. 왜냐하면 중요한 의미에서 뇌를 갖고 있는 우리의 신체가 그러한 기계이기 때문이다. 그러나 기계의 작동이 단지 형식적으로 정의된 요소들에 대한 계산적 과정으로 정의되어 있는 기계에 대해, 다시 말해 기계 작동이 컴퓨터 프로그램의 구체화로 정의되어 있는 경우에는 그런 능력을 부여할 수 없다는 강력한 주장도 있다. 그러나 내가 영어를 이해할 수 있고, 다른 종류의 의도성을 갖는 것은 내가 컴퓨터 프로그램의 구체화이기 때문이 아니다(어쩌면 내가 몇 개인가의 컴퓨터 프로그램의 구체화일지 모른다고 생각하지만). 우리가 아는 한 그 이유는, 내가 어떤 종류의 생물학적(즉 화학적 · 물리적) 구조를 가진 어떤 종류의 유기체이기 때문이고, 이러한 구조는 특정 조건 아래에서 지각, 행동, 이해, 학습, 그리고 그 밖의 의도적 현상을 인과적으로 일으킬 수 있다. 지금 우리들 논의의 또 하나의 요점은 이러한 인과적 힘을 가진 어떤 것만이 그러한 의도성을 가질 수 있다는 것이다. 어쩌면 다른 물리적 · 화학적 과정도 똑같은 결과를 낳을 수 있을지 모른다. 가령 화성인들도 의도성을 갖고 있을지 모른다. 그러나 그들의 뇌는 다른 종류의 물질로 이루어졌을 수 있다. 그것은 광합성이 엽록소와 다른 화학 물질에 의해 이루어지는지에 대한 문제와 흡사한 경험적인 문제이다.

그러나 이 논의의 중심적인 논점은 순수하게 형식적인 어떠한 모형도 그 자체로 의도성을 설명하기에 충분하지 않다는 것이다. 형식적인 성질만으로 의도성이 구성되지는 않으며, 그것들은 기계가 움직이고 있을 때 형식화의 새로운 단계를 낳는 힘을 제외하면 그 자체로는 어떠한 인과력도 갖지 않기 때문이다. 그리고 형식적인 모형의 특정한 구체화가 갖는 그 밖의 인과적 특성들도 형식적인 모형과는 무관하다. 왜냐하면 우리는 동일한 형식적인 모형을, 분명 그러한 인과적 특성을 결여하는 다른 구현realization에 끼워맞출 수 있기 때문이다. 가령 중국어로 말하는 사람이 정확히 샌크의

프로그램을 구현했다 해도 우리는 같은 프로그램을 영어로 말하는 사람이나 송수관 또는 컴퓨터에 넣을 수 있지만, 이 경우 프로그램은 실행될 수 있어도 중국어를 이해하는 사람은 아무도 없다.

뇌의 작동에 대해 제기되는 문제는 일련의 시냅스를 주형(鑄型)으로 삼아 만들어진 형식적인 그림자가 아니라 일련의 시냅스의 실질적인 특성이다. 지금까지 검토한 강한 인공 지능의 주장들은 모두 인지 현상이라는 주형에서 찍어낸 그림자 주위에 윤곽을 그리고는 이 그림자가 실재하는 것이라고 주장하고 있다.

결론을 내리기에 앞서 나는 그러한 주장에 함축되어 있는 일반적인 철학적 논점들을 제기하려고 한다. 이해를 분명히 하기 위해서 질의 응답이라는 형식을 사용하기로 하겠다. 우선 고색창연한 질문에서 시작하기로 하자.

〈기계는 생각할 수 있는가?〉

대답은 분명 〈그렇다〉이다. 우리가 바로 그러한 기계이다.

〈그렇다면 인공물, 즉 사람이 만든 기계는 생각할 수 있는가?〉

신경계를 가진 기계를 인공적으로 만들 수 있다면 가능할 것이다. 다시 말해서 축삭, 수상돌기, 그리고 그 밖의 구조를 가진 뉴런과 비슷한 것을 가진 기계를 만들 수 있다면 그 질문에 대한 대답은 역시 〈그렇다〉이다. 원인을 정확히 복제할 수 있다면 결과도 복제할 수 있을 것이다. 의식, 의도성, 인간이 사용하고 있는 화학 원리들을 사용해서 그 밖의 모든 특성도 만들어낼 수 있을 것이다. 앞에서도 말했듯이 그것은 경험적인 문제이다.

〈좋다. 그렇다면 디지털 컴퓨터는 생각할 수 있는가?〉

만약 〈디지털 컴퓨터〉의 의미가 컴퓨터 프로그램의 구체적 실현으로 기술될 수 있는 기술의 수준을 갖는 것을 뜻한다면 물론 대답은 〈그렇다〉이다. 왜냐하면 우리는 정확히 그 숫자는 모르지만 복수의 컴퓨터 프로그램의 구체적 실현이고 또한 생각하는 능력을 갖고 있기 때문이다.

〈그러나 컴퓨터가 정확한 종류의 프로그램을 갖기만 하면 무언가를 생각하거나 이해할 수 있다는 말인가? 프로그램의 구체적 실현, 정확한 프로그램의 구체적 실현만으로 이해를 위한 충분한 조건이 될 수 있을까?〉

이런 질문은 항상 앞의 질문들과 혼동되어 있지만, 적절한 질문이다. 그리고 이 질문에 대한 대답은 〈아니다〉이다.

〈그 이유는?〉

왜냐하면 형식 기호는 그 자체로 의도성을 갖지 않기 때문이다. 그것은 아무런 의미도 갖지 않는다. 심지어 그것은 기호 조작도 아니다. 왜냐하면 그 기호는 아무것도 상징하지 않기 때문이다. 언어학 용어를 빌리자면 그것은 구문론을 가질 뿐 의미론은 갖지 않는다. 흔히 컴퓨터가 갖는 것으로 생각되는 의도성은 그것을 프로그램하는 사람과 사용하는 사람의 마음, 즉 입력하는 사람과 출력을 해석하는 사람의 마음 속에만 있는 것이다.

중국어 방의 예를 들었던 목적은 이 점을 입증하기 위해서 실제로 의도성을 갖는 시스템(사람) 속에 무언가를 투입하자마자, 그리고 우리가 그를 형식적인 프로그램으로 프로그램하자마자 그 형식적인 프로그램이 더 이상 어떤 부가적인 의도성도 수반하지 않는다는 것을 보여주려 함이었다. 예를 들어 그것은, 사람이 중국어를 이해하는 능력에 아무것도 덧붙이지 않는다.

일찍이 대단히 매력적인 것으로 여겨졌던 인공 지능의 그러한 특

징(프로그램과 그 구체적 실현의 구별)이, 시뮬레이션이 복제될 수 있으리라는 주장을 뿌리에서부터 뒤흔든 것이다. 프로그램과 그 프로그램의 하드웨어에서의 구현의 구별은 정신적 조작 수준과 뇌의 조작 수준의 구별에 상응하는 것으로 짐작된다. 그리고 우리가 정신적 조작 수준을 형식적 프로그램으로 기술할 수 있다면, 우리는 내성적인 심리학이나 뇌에 대한 신경생리학을 연구할 필요 없이 마음의 본질을 기술할 수 있는 것처럼 판단된다. 그러나 〈마음과 뇌의 관계는 프로그램과 하드웨어의 관계이다〉라는 등식은 몇 가지 측면에서 붕괴한다. 그 중에서 세 가지 측면을 살펴보기로 하자.

첫째, 프로그램과 그 구현의 구별은 같은 프로그램이 어떠한 형태의 의도성도 갖지 않는 온갖 종류의 괴상한 구현을 얻을 수 있다는 결과를 낳는다. 가령 바이첸바움Weizenbaum은 화장실의 두루말이 휴지와 조약돌 무더기를 이용해서 컴퓨터를 만드는 방법을 구체적으로 보여주었다. 마찬가지로 프로그램을 이해하는 중국어 스토리는 일련의 송수 파이프나 송풍기 집합 또는 영어밖에 할 수 없는 사람에게도 프로그램될 수 있으며, 그중 어느 것도 그러한 프로그램을 통해 중국어에 대한 이해를 얻을 수 없다. 조약돌, 화장실 휴지, 바람, 송수 파이프는 의도성을 갖기에 적절한 종류의 소재가 아니다. 뇌처럼 인과력을 갖는 무언가만이 의도성을 가질 수 있다. 그리고 영어로 말하는 사람은 의도성에 적합한 종류의 소재를 갖지만 프로그램을 기억한다고 해서 그 프로그램이 그에게 중국어를 가르쳐주지는 않기 때문에 그것을 기억하더라도 그 이상의 의도성을 얻지는 못한다.

둘째, 프로그램은 순수하게 형식적이지만 의도적 상태들은 그러한 방식에서 형식적이지 않다. 그것들은 내용에 의해 정의되는 것이지 형식에 의해 정의되는 것이 아니다. 가령 비가 내리고 있다는 생각은 특정한 형식으로 정의되는 것이 아니며, 그것을 충족시키는 조건이나 적합한 방향direction of fit 등을 포함하는 특정한 정신적인 내용으로 정의된다. 실제로 신념 그 자체는 이와 같은 구문론적 의미에서의 형식조차 갖지 않는다. 왜냐하면 동일한 신념이 다른 언어 체계에서 무한히 많은 구문론적 표현을 낳을 수 있기 때문이다.

셋째, 앞에서도 언급했듯이 정신적 상태와 정신적 사건들은 문자 그대로 뇌 작용의 소산이지만 프로그램은 이런 의미에서의 컴퓨터의 산물이 아니다.

〈만약 프로그램이 어떠한 의미에서도 정신적 과정을 구성하는 것이 아니라면, 그렇게 많은 사람들이 정반대의 믿음을 갖는 까닭은 무엇인가? 최소한 그 점에 대해서는 어떤 식으로든 설명이 필요하지 않은가?〉

사실 이 물음에 대한 답은 나도 알지 못한다. 컴퓨터의 목적이 정신적 작용을 시뮬레이트하는 것으로 한정되지 않기 때문에, 우선 컴퓨터 시뮬레이션이 실재일 수 있다는 생각 자체에 의구심을 품어야 한다. 화재의 컴퓨터 시뮬레이션이 이웃집을 태워버리거나 폭풍우의 컴퓨터 시뮬레이션이 우리를 흠뻑 젖게 만들 것이라고 생각하는 사람은 아무도 없다. 컴퓨터로 시뮬레이트한 것이 실제로 무언가를 이해한다고 생각하는 까닭은 도대체 무엇인가? 사람들은 종종 컴퓨터에 아픔을 느끼게 하거나 사랑에 빠지게 하기란 절대 불가능할 것이라는 말을 하지만, 사랑이나 아픔이 인지나 그 밖의 것들보다 특별히 더 불가능한 것은 아니다. 시뮬레이션에서 필요한 것은

적절한 입력과 출력, 그리고 전자를 후자로 변환시키는 매개자로서의 프로그램밖에 없다. 이것이 컴퓨터의 전부이다. 고통이든 사랑이든 인지든 화재든 폭풍우든 간에, 시뮬레이션을 복제와 혼동하는 것은 모두 동일한 오류이다.

그런데 인공 지능이 어떤 식으로든 정신적 현상을 재현하고 그에 의해 정신 현상을 설명하는 것처럼 생각되는 데에는 (아마도 많은 사람들은 여전히 그렇게 믿고 있을 것이다) 몇 가지 이유가 있다. 대개 그런 착각은 그 원인이 충분히 밝혀지기 전까지는 쉽게 사라지지 않는 법이다.

첫번째, 그리고 가장 중요한 이유는 〈정보 처리〉의 개념에 대한 혼란이다. 인지과학을 연구하는 많은 사람들은 마음을 갖고 있는 인간의 뇌가 〈정보 처리〉라 불리는 것을 하고 있으며, 유추적으로 프로그램을 가진 컴퓨터도 정보 처리를 하고 있다고 생각한다. 그러나 다른 한편, 화재나 폭풍우는 전혀 정보 처리를 하지 않는다고 생각한다. 따라서 컴퓨터가 어떤 과정의 형식적인 특징을 시뮬레이트할 수는 있지만, 그 과정은 마음과 뇌에 대해 특별한 관계를 갖는다는 것이다. 왜냐하면 컴퓨터가 뇌와 동일한 프로그램에 의해 이상적으로 프로그램되었을 때 두 경우의 정보 처리는 동일하며, 이 정보 처리는 실제로 마음의 본질이기 때문이라는 것이다. 그러나 이런 주장의 문제점은 거기에서 사용되는 〈정보〉라는 개념이 모호하다는 점이다. 예를 들어 산수 문제를 생각하거나 이야기를 읽고 질문에 대답할 때 인간이 〈정보를 처리한다〉라고 말하는 의미에서 본다면 프로그램된 컴퓨터는 〈정보 처리〉를 하지 않기 때문이다. 이때 컴퓨터가 하는 것은 형식 기호의 조작이다. 프로그래머와 컴퓨터 출력의 해석자가 기호를 사용해서 세계 속의 대상을 나타낸다는 사실은 컴퓨터가 할 수 있는 영역을 넘어서는 것이다. 되풀이하자면 컴퓨터에는 구문론은 있지만 의미론은 없다. 가령 당신이 컴퓨터에 〈2 더하기 2는?〉이라고 타이핑하면 컴퓨터는 〈4〉라는 답을 낼 것이다. 그러나 컴퓨터는 〈4〉가 4를 의미한다는 것을 모르며, 애당초 의미라는 것 자체를 알지 못한다. 따라서 중요한 것은 컴퓨터가 제1수준의 기호 해석에 대해서 제2수준의 정보를 결여하고 있다는 것이 아니라, 제1수준의 기호가 아무런 해석도 갖지 않는다는 점이다. 컴퓨터가 갖고 있는 것은 기호, 기호들뿐이다. 따라서 〈정보 처리〉 개념의 도입은 딜레마를 낳는다. 그것은 우리가 〈정보 처리〉 개념을 처리 과정의 일부로서 의도성을 함축하는 것으로 해석할 것인지 또는 그렇게 해석하지 않을 것인지의 딜레마이다. 의도성을 함축한다는 의미로 해석하면, 프로그램된 컴퓨터는 정보 처리를 하지 않으며, 단지 형식 기호를 조작하는 것이 된다. 반면 의도성을 함축하지 않는다고 해석하면 컴퓨터는 정보 처리를 하지만 그것은 단지 가산기, 타자기, 위(胃), 자동 온도 조절 장치, 폭풍우, 허리케인과 같은 의미에서 정보 처리를 하는 것이 된다. 다시 말해서 그것들은 한쪽 끝에서 정보를 받아들이고, 변형시키고, 출력으로 내보낸다고 기술할 수 있는 기술 수준을 갖는다. 그러나 이 경우 입력과 출력을 일반적인 의미에서의 정보로 해석하는 일은 외부 관찰자의 몫이다. 그리고 컴퓨터와 뇌의 유사성은 정보 처리의 유사성이라는 관점에서는 수립될 수 없다.

둘째, 인공 지능 연구의 상당 부분에는 행동주의와 조작주의의 찌꺼기가 남아 있다. 적절히 프로그램된 컴퓨터는 인간의 입출력 패턴과 유사한 패턴을 갖기 때문에 우리는 컴퓨터에도 인간의 정신 상태와 흡사한 정신 상태가 있다고 가정하고 싶어한다. 그러나 우

리는 어떤 의도성도 갖지 않으면서 일부 영역에서 인간의 능력을 갖는 것이 개념상으로나 경험상으로 가능하다는 것을 알고 있기 때문에 이러한 가정을 하고 싶은 충동을 극복해야 한다. 내 책상 위의 계산기는 계산 능력을 갖지만 어떤 의도성도 없다. 이 글에서 입증하려고 시도했듯이 어떤 체계는 중국어가 모국어인 화자(話者)의 입출력 능력을 복제할 수 있는 능력을 갖지만, 그것이 어떻게 프로그램되는가와 무관하게 중국어를 이해하지 못한다. 튜링 테스트는 이 뻔뻔스러운 행동주의와 조작주의 전통의 전형이고, 만약 인공 지능 연구자들이 행동주의와 조작주의를 전면적으로 거부한다면 시뮬레이션과 복제 사이의 혼동은 상당 부분 제거될 것으로 생각한다.

셋째, 잔존하는 조작주의는 이원론의 잔존 형태에 결부된다. 실제로 강한 인공 지능은 마음이 문제이지 뇌는 중요치 않다는 이원론적 가정을 기초로 할 때에만 의미를 갖는다. 강한 인공 지능 연구에서(마찬가지로 기능주의에서) 중요한 것은 프로그램이며, 프로그램은 기계 속에서의 구현과는 독립적인 무엇이다. 사실상, 인공 지능에 관한 한 동일한 프로그램이 전자식 기계나 데카르트의 정신적 실체 mental substance나 헤겔의 세계 정신에 의해 모두 실현될 수 있다. 내가 이 문제를 논의하는 과정에서 이룬 놀라운 발견 중 하나는, 많은 인공 지능 연구자들이 인간의 정신적 현상은 인간 뇌의 실질적인 물리 화학적 성질에 의존하고 있을 것이라는 내 생각에 큰 충격을 받고 있다는 점이다. 그러나 조금만 생각해 보면 전혀 놀라운 일이 아닌 것을 알 수 있다. 왜냐하면 강한 인공 지능 프로젝트는 어떤 형태로든 이원론을 받아들이지 않으면 승산이 없기 때문이다. 인공 지능 프로젝트는 프로그램을 설계해서 마음을



재현하고 설명하는 것이다. 그러나 마음이 개념적으로뿐 아니라 경험적으로도 뇌에서 독립된 것이 아니라면 이 프로젝트는 실행될 수 없다. 왜냐하면 이 프로그램이 어떠한 실현으로부터도 완전히 독립되어 있기 때문이다. 마음이 뇌로부터 개념적으로나 경험적으로 분리될 수 있다고 (이것은 강한 형태의 이원론이다) 생각하지 않으면, 프로그램이 뇌나 그 밖의 개별적인 형태의 구체화로부터 독립되어 있기 때문에 프로그램을 작성하거나 실행시키는 방법으로 마음을 재생시키는 것을 바랄 수 없다. 정신적 작용이 형식 기호에 대한 컴퓨터적 작용이라면 그것은 뇌와 흥미로운 관계를 맺지 않게 된다. 둘 사이의 유일한 연결은, 뇌가 우연히 그 프로그램을 실현시킬 수 있는 무수히 많은 기계들의 유형 중 하나일 것이다. 이러한 형태의 이원론은 두 종류의 실체가 존재한다고 주장하는 전통적인 데카르트의 이원론은 아니지만, 마음이 뇌의 실제적 성질과 본질적 관계를 갖지 않는다는 것을 강조한다는 의미에서 데카르트적이다. 이처럼 내재하는 이원론은 인공 지능과 연관된 문헌들이 이 〈이원론〉을 종종 맹렬하게 비난한다는 사실에 의해 숨겨지기 때문에 우리 눈에는 잘 띄지 않는다. 다시 말해서 인공 지능 문헌의 저자들은 자신들의 입장이 강한 이원론을 전제로 삼고 있다는 것을 알아차리지 못하는 것 같다.

〈기계는 생각할 수 있는가?〉 내 견해로는 기계만이, 실제로는 특수한 기계, 그러니까 뇌나 뇌와 동일한 인과력을 갖는 기계만이 생각할 수 있다. 이것이 〈생각한다〉는 문제에 대해 강한 인공 지능이 거의 아무것도 이야기하지 않는 주된 이유이다. 왜냐하면 강한 인공 지능은 기계에 대해 할 이야기가 아무것도 없기 때문이다. 그 정의에 따르면 강한 인공 지능은 프로그램에 대한 것이고, 프로그

램은 기계가 아니다. 의도성이 무엇이든 간에 그것은 생물학적 현상이고, 젖 분비나 광합성 또는 그 밖의 생물학적 현상처럼 그 기원에 해당하는 생화학 구조에 인과적으로 의존하는 경향이 강하다. 젖 분비나 광합성을 컴퓨터로 시뮬레이션한다고 해서 우유나 설탕을 생산할 수 있다고 생각하는 사람은 아무도 없을 것이다. 그러나 마음의 문제에 대해서는 많은 사람들이 이러한 기적을 기꺼이 믿는다. 그것은 세월이 지나도 변하지 않는 뿌리 깊은 이원론 때문이다. 그들이 생각하는 마음은 형식적인 과정의 문제이고, 우유나 설탕과는 달리 구체적인 물질적 원인에서 독립된 무엇이다.

이러한 이원론을 옹호하기 위해서 뇌는 디지털 컴퓨터(그런데 초기 컴퓨터는 종종 〈전자 두뇌〉라고 불리곤 했다)라는 희망이 종종 표명된다. 그러나 그런 시도는 아무 도움도 되지 않는다. 물론 뇌는 디지털 컴퓨터이다. 모든 것이 디지털 컴퓨터이기 때문에 뇌도 마찬가지이다. 그러나 중요한 것은 의도성을 낳는 뇌의 인과력은 어떤 컴퓨터 프로그램의 구체화라는 점에 있는 것이 아니라는 사실이다. 왜냐하면 당신이 원하는 어떤 프로그램에서도 어떤 목적을 위해 그 프로그램을 구현시킬 수 있지만, 그래도 여전히 어떤 정신적 상태도 가질 수 없기 때문이다. 의도성을 낳기 위해서 뇌가 어떤 일을 하든, 어떤 프로그램도 그 자체만으로는 의도성을 얻기에 충분치 않기 때문에 그것은 프로그램의 구현이 아닌 것이다.*

* 나는 이 문제를 다루는 과정에서 많은 사람들에게 빚을 졌다. 그들은 인공 지능에 대한 내 무지를 극복시키려고 인내심 깊은 노력을 기울여주었다. 특히 네드 블록Ned Block, 휴버트 드레퓌스Hubert Dreyfus, 존 호질랜드John Haugeland, 로저 샌크, 로버트 윌렌스키Robert Wilensky, 그리고 테리 위노그라드에게 깊은 감사를 드린다.





처음 발표되었을 당시 이 논문에는 여러 방면에 종사하는 사람들로부터 들어온 스물여덟 개의 논평이 첨부되어 있었다. 그 중 상당수는 뛰어난 견해를 포함하고 있었지만, 그것들을 모두 싣기에는 분량이 너무 방대하고, 또한 일부는 내용이 조금 전문적이었다. 설 논문의 장점 중 하나는 인공 지능, 신경학, 철학 또는 그 밖의 연관 분야에 대한 특별한 소양이 없어도 상당 정도 이해할 수 있다는 것이다.

우리(두 사람의 편집자——옮긴이)의 입장은 설과는 정면으로 대립된다. 그렇지만 설이 매우 탁월한 상대라는 점을 우리는 인정한다. 우리는 이 책의 나머지 부분에서 설의 주장을 철저하게 반박할 것이다. 특히 설이 제기하는 주장의 몇 가지 논점에 논의를 집중시키고, 그 밖의 논점에 대해서는 분명하게 답변하지 않고 남겨두기로 하겠다.

설의 논문은 〈중국어 방의 사고 실험Chinese room thought experiment〉이라는 교묘한 상황 설정을 기반으로 삼으며, 이 사고 실험에서 독자들은 매우 영리한 인공 지능 프로그램이라면 통과할 수 있을 것으로 알려진 일련의 단계를 튜링 테스트에 합격할 수 있을 정도로 사람과 흡사한 방식으로 중국어 스토리를 읽고, 그 스토리에 대한 질문에 대답하는 일련의 단계를 수작업으로 하고 있는 사람과 동일시하도록 촉구된다. 우리는, 인간이 이런 일을 할 수 있다고 생각하는 것이, 어떤 의미가 있는 듯한 잘못된 인상을 주고 있다는 점에서 설이 심각하고 근본적인 설명의 오류를 범했다고 생각한다. 이러한 상(像)

을 받아들이면 독자들은 알지 못하는 사이에 지능과 기호 조작 사이의 관계에 대해 전혀 비현실적인 사고 방식을 전면적으로 받아들이는 셈이 되고 말 것이다.

설이 독자들에게 일으키려는 착각은 (물론 설 자신은 착각이라고 생각하지 않지만!) 개념 수준을 달리하는 두 개의 체계 사이에 존재하는 복잡성의 엄청난 차이를 독자들이 간과하게 만들 수 있는지의 여부에 달려 있다. 일단 그 시도에 성공하면 나머지는 별로 문제가 되지 않는다. 맨 처음에 독자들은 몇 가지 제한된 영역 내에서 한정된 종류의 질문으로 한정된 방식으로 대답할 수 있는 실제 인공 지능 프로그램을 수작업으로 시뮬레이트하는 설과 자신을 동일시하도록 요구된다. 그런데 이 프로그램 또는 현재 시점에 존재하는 모든 인공 지능 프로그램을 사람이 직접 수작업으로 시뮬레이트하려면, 즉 컴퓨터가 하는 상세한 수준에서 한 단계씩 작업을 수행하려면 몇 주일이나 몇 개월은 아니더라도 며칠 동안 힘들고 지루한 작업을 해야 한다. 그러나 그는 이 점을 언급하지 않고, 숙달된 마술사처럼 교묘하게 독자의 주의를 빗겨나서 독자들이 갖는 상을 튜링 테스트를 통과하는 가상의 프로그램으로 바꾸어놓고 있다! 설은 여러 단계의 능력 수준을 단숨에 뛰어넘으며, 각 수준에 대해서는 지나가는 언급조차 하지 않는다. 따라서 독자들은 또다시 한 단계씩 시뮬레이션을 수행해 나가는 사람과 자신을 동일시해서 중국어에 대한 〈이해의 결여를 느끼도록〉 요청된다. 이것이 설의 주장에서 골자이다.

여기에 대한 우리의 반론은(나중에 밝혀지겠지만 설 자신의 대응도 마찬가지이다) 기본적으로는 〈체계 이론의 대응Systems Reply〉이다. 즉 살아 있는(살아 있는지 여부는 부수적인 것이지만) 시뮬레이터가 이해 능력을 갖는다고 인정하는 것은 잘못이다. 오히려 이해 능력은 설이 별 생각없이 〈몇 개의 종잇조각〉이라고 부르고 있는 것을 포함하는 전체로서의 체계에 속한다. 이런 표현에서 분명히 드러나듯이 설이 갖고 있는 상이 그를 현실 상황에 눈멀게 만들고 있는 것이다. 생각하는 컴퓨터가 설에게 기피해야 할 존재인 것은 마치 비유클리드 기하학에 대해 그것을 의도하지 않게 발견한 제롤라모 사케리Gerolamo Saccheri가 갖는 관계와 흡사하다. 사케리는 시종일관 자신의 연구 결과를 부인했다. 1700년대 후반은 아직 사람들이 새로운 기하학에 의해 야기된 개념의 확장을 받아들이기에 너무 일렀다. 그러나 약 50년이 지난 후 비유클리드 기하학은 재발견되었고 오늘날까지 느린 속도로 수용되어 왔다.

어쩌면 〈인공 의도성〉에 대해서도 (만약 그런 것이 탄생할 수 있다면) 마찬가지의 일이 벌어질지도 모른다. 만약 튜링 테스트를 통과하는 프로그램이 등장한다면, 설은 그 프로그램의 능력과 깊이에 경탄하기는커녕 그것이 〈뇌의 인과력〉이라는 경이적인 능력을 결여하고 있다는 주장을 계속할 것이다. 이러한 주장의 공허함을 지적하기 위해서 제논 필리신Zenon Pylyshyn은 설의 글에 대한 논평에서 다음과 같은 글이 즈보프의 「어느 뇌 이야기」를 연상시키는 관점의 특징을 정확히 보여주고 있다고 말한다.

만약 당신의 뇌세포를 조금씩 IC 칩으로 대체시켜 나간다면, 이들 칩이 각 부분의 입출력 〈기능〉이 뇌세포의 입출력 기

능과 동일해지도록 프로그램되어 있다면, 당신의 입에서 나오는 말은 실제 당신이 하는 말과 모든 면에서 똑같을 것이다. 그 말이 궁극적으로 어떤 〈의미〉도 갖지 않게 된다는 점을 제외한다면 말이다. 그렇게 되면, 우리들 외부 관찰자들이 말이라고 이해하는 것은 당신에게는 회로에 의해 발생하는 일정한 잡음일 것이다.

설의 입장에서 약점은 진짜 의미가, 또는 진짜 〈당신〉이 체계의 어디에서 사라지는지를 명확히 밝히지 않고 있다는 것이다. 단지 그는 〈인과력〉에 의해 의도성을 갖는 체계도 있고, 그렇지 않은 체계도 있다는 것을 강조할 뿐이다. 그는 그러한 힘이 어디에서 유래하는지에 대해 동요하는 것 같다. 어떤 때에는 뇌가 〈적절한 재료〉로 구성되어 있는 것처럼 보이고, 다른 때에는 그 이외의 다른 것들로 이루어져 있는 것처럼 보이기도 한다. 그 원인은 당시에 끌어대기에 편리한 것이면 무엇이든 될 수 있기 때문이다. 때로는 〈내용〉과 〈형식〉을 구별하는 모호한 본질이고, 때로는 의미론에서 구문론을 분리시키는 또 다른 본질이기도 하다.

체계 이론을 주창하는 사람들에게 설은 방 속에 있는 사람(앞으로 그 사람을 설의 〈데몬〉이라고 부르기로 하자)이 〈몇 개의 종잇조각〉에 앞에서 이야기한 모든 재료들을 단순히 기억시키거나 통합시킨다는 사고를 제공한다. 상상력을 가능한 한 확장시키면 인간이 그러한 일을 할 수 있는 것처럼 말이다. 이러한 〈몇 개의 종잇조각〉 위의 프로그램은 튜링 테스트를 통과할 수 있는 능력 덕분에 문자로 작성된 자료에 의거해 답변할 능력이



라는 측면에서 인간과 같은 정도로 복잡한 마음과 성격 전체를 구현하고 있다. 다른 사람의 마음의 전체 기술(記述)을 간단히 〈삼켜버릴〉 수 있는 사람이 과연 있을 수 있을까? 우리 생각으로는 한 단락의 문장을 통째로 기억하기도 매우 어렵다. 그러나 설은 수십억 쪽은 아니더라도 확실히 수백만 쪽에 달할 빽빽이 기록된 추상적인 기호를 데몬이 소화시켜 버리고, 검색하는 데 아무런 문제도 없이, 필요할 때면 언제든 이 모든 정보를 이용할 수 있는 것처럼 공상하고 있는 것이다. 따라서 설은 이 시나리오의 실현 불가능한 여러 측면을 간과하고 마치 모든 것이 수월한 양 쓰고 있다. 그리고 그 시나리오가 의미 있다는 것을 독자들에게 확신시키는 것이 설의 주장의 핵심은 아니다. 사실은 정반대이다. 설의 주장의 핵심 부분은 이처럼 중요한 문제들을 적당히 꾸며대며 얼버무리는 것이다. 그렇게 하지 않으면 회의적인 독자들 대부분의 이해가 종이 위 수십억 개의 기호 중에 분명 있을 것이며, 데몬 속에는 그 한 조각도 없다는 것을 깨닫게 될 것이기 때문이다. 그 데몬이 살아 있다는 사실은 (그것도 오해되고 있는) 부차적인 문제에 불과하다. 그렇지만 설은 그것을 매우 중요한 사실로 오해하고 있다.

우리는 설 자신이 체계 이론을 지지하고 있다는 사실을 밝혀냄으로써 앞의 주장을 뒷받침할 수 있다. 그러기 위해서 우선 설의 사고 실험을 더 넓은 맥락 속에 놓으려고 한다. 특히 설의 사고 실험의 설정이 그와 연관된 사고 실험들의 큰 집합 중 하나에 지나지 않음을 폭로하고자 한다. 또한 그러한 사고 실험은 이 책에 실린 여러 글에서도 다루어지고 있다. 이런 종류의 사고 실험은 실험 당사자가 〈스위치 설정 knob setting〉을 개별

적으로 선택함으로써 정의된다. 그 목적은 독자들의 마음의 눈 속에 인간의 정신적 활동에 대한 여러 가지 가상의 시뮬레이션을 창조하는 것이다. 각각의 사고 실험은 문제의 여러 측면을 확대시켜서 독자들을 특정한 결론으로 밀어붙이는 경향이 있는 일종의 〈직관 펌프〉intuition pump(데닛의 용어)이다. 우리는 대략 다섯 개의 스위치에 관심을 둔다. 그러나 더 많은 스위치를 생각해도 무방하다.

스위치 1

이 스위치는 시뮬레이션을 구성하는 물리적 〈재료〉를 제어한다. 그 설정에는 다음과 같은 것이 포함된다. 뉴런과 화학 물질, 송수 파이프와 물, 몇 개의 종잇조각과 그 위에 적힌 기호들, 화장지와 조약돌, 데이터 구조와 프로시듀어 등등.

스위치 2

이 스위치는 시뮬레이션이 사람의 뇌를 흉내내려고 시도할 때 모방의 정밀도를 제어한다. 그것은 극히 미세한(원자 속의 소립자) 수준까지 설정할 수 있고, 그보다 큰 세포나 시냅스의 수준, 그리고 인공 지능 연구자나 인지심리학자들이 다루는 수준, 즉 개념과 관념, 표상과 과정의 수준으로도 설정할 수 있다.

스위치 3

이 스위치는 시뮬레이션의 물리적인 크기를 제어한다. 우리의 가정에 따르면 극소화(極小化)의 결과로 우리는 반지에 들어갈 정도로 작은 송수관의 망상 조직이나 반도체 소자를 만들



수 있고, 거꾸로 모든 화학적 과정을 거시적 규모에 확장시킬 수도 있다.

스위치 4

이것은 중요한 스위치로 시뮬레이션을 수행하는 데몬의 크기와 성질을 제어한다. 만약 데몬이 정상 크기의 사람이라면, 그것을 〈설의 데몬〉이라고 부르기로 하자. 데몬이 뉴런이나 소립자 속에 들어갈 수 있을 정도로 작은 꼬마 요정과 같은 생물이라면 호질랜드의 이름을 빌려 〈호질랜드의 데몬〉이라고 부르기로 하자. 그런 이름을 붙이는 까닭은 호질랜드가 설을 비판했기 때문이다.

스위치 5

이 스위치도 데몬이 살아 있는지 여부를 결정한다. 또한 이 스위치는 데몬이 일하는 속도를 제어한다. 다시 말해서 데몬이 눈부실 정도로 빠르게(100만 분의 1초당 100만 회의 연산을 하는 정도의 속도로) 일하게 할 수도 있고, 지독히 느리게 (몇 초마다 한 번의 속도로) 일하도록 설정할 수도 있다.

우리는 스위치의 설정을 여러 가지로 바꾸어 다양한 사고 실험을 할 수 있다. 하나의 선택 결과, 이야기 스물여섯 「아인슈타인의 뇌와 나눈 대화」에 서술한 상황이 발생한다. 다른 선택을 하면 설의 중국어 방의 실험이 나타난다. 특히 후자는 다음과 같은 설정을 갖는다.

스위치1 종잇조각과 기호
스위치2 개념과 관념
스위치3 방의 크기
스위치4 사람 크기의 데몬
스위치5 느린 설정(수초에 1회의 연산)

그런데 이러한 매개 변수들을 가진 시뮬레이션이 튜링 테스트를 통과할 수 있다는 가정에 설이 본질적으로는 반대하지 않는다는 점을 주목할 필요가 있다. 설은 이러한 가정이 무엇을 함의하는지에 대해서만 자신의 주장을 펴고 있다.

마지막 변수가 하나 더 있다. 그것은 스위치가 아니고, 설의 실험을 보는 관점이다. 이 단조로운 실험에 약간의 색을 칠해 시뮬레이트된 중국어 화자가 여성이고, 데몬은 (만약 살아 있다면) 항상 남성이라고 가정해 보자. 이제 우리는 데몬의 관점과 시스템의 관점 중 하나를 고를 수 있는 선택권을 갖는다. 가정에 따라 데몬과 시뮬레이트된 여성은 모두 자신이 이해하고 있는지 여부, 그리고 자신이 경험하고 있는 것에 대해 동등하게 견해를 표명할 수 있다는 점을 상기할 필요가 있다. 그럼에도 불구하고 우리가 이 실험을 데몬의 관점에서만 본다고 설은 강변한다. 그는 시뮬레이트된 여성이 자신이 이해한 내용에 대해 (물론 중국어로) 무슨 말을 하더라도 그녀가 하는 말을 무시해야 하며, 오히려 기호 조작을 수행하는 내부의 데몬에 주의를 기울여야 한다고 주장한다. 결국 설의 주장은 두 개가 아니라 하나의 관점밖에 존재하지 않는다는 것이다. 일단 실험 전체에 대한 설의 기술을 받아들인다면 이 주장은 엄청난 직관적 설득력을 갖는다. 왜냐하면 데몬은 겨의 우리와 같은 크기이고, 우리의 언어로 말하고, 또한 우리와 같은 속도로 일하기 때문이다. (운이 좋아도) 한 세기에 한 번 정도의 속도로, 그것도 〈무의미한 꼬부랑 곡선들로〉 대답을 내놓는 〈여성〉과 동일시하기란 매우 힘들다.

그러나 스위치의 일부 설정을 바꾸기만 하면 쉽게 관점을 바꿀 수 있다. 특히 호질랜드의 변형판에는 다음과 같은 여러 가지 전환 스위치가 포함되어 있다.

스위치1 뉴런과 화학 물질
스위치2 뉴런의 발화 수준
스위치3 사람의 뇌의 크기
스위치4 작은 데몬
스위치5 현기증이 날 정도로 빠른 데몬

호질랜드는 우리가 다음과 같이 상상해 주기를 바란다. 가령, 안타깝게도 실제 여성의 뇌에는 결함이 있다. 이 뇌는 더 이상 뉴런에서 뉴런으로 신경 전달 물질을 보낼 수 없다. 그러나 다행스럽게도 이 뇌 속에는 뉴런이 이웃 뉴런으로 신경 전달 물질을 전달하려 할 때마다 개입하는 아주 작고 빠른 호질랜드의 데몬이 살고 있다. 이 데몬은 이웃 뉴런에게 진짜 신경 전달 물질이 전달된 것과 기능적으로 구별할 수 없는 방식으로 그 뉴런의 적절한 시냅스를 〈자극한다〉. 게다가 호질랜드의 데몬은 무척 빨라서 한 시냅스에서 다른 시냅스로 1조 분의 1초만에 뛰어다닐 수 있을 정도이기 때문에 시간을 지연시키는 일

따위는 결코 없다. 이런 방식으로 그 여성의 뇌 기능은 그녀가 건강할 때와 똑같이 유지된다. 이 대목에서 호질랜드는 설에게 질문을 던진다. 이 여성은 여전히 사고하고 있는 것인가? 다시 말해서 그녀는 의도성을 갖고 있을까? 아니면 튜링이 인용한 제퍼슨 교수의 말을 빌리자면, 그녀는 단지 〈인위적으로 신호를 보내는〉 것에 지나지 않는가?

여러분은 설이 우리에게 데몬의 말에 귀를 기울여서 데몬과 자신을 동일시하고, 시스템 이론의 주장(그 주장은 물론 여성의 말에 귀를 기울이고, 그녀와 동일시할 것이다)을 받아들이지 말라고 다그칠 것으로 기대할지도 모른다. 그러나 호질랜드의 주장에 대한 그의 반응은 무척 놀랍다. 이번에는 그가 그녀에게 귀를 기울이고 오히려 데몬을 무시하라는 쪽을 선택한 것이다. 데몬은 그의 관점에서 이렇게 절규하며 우리를 저주한다. 〈바보들! 그녀가 하는 말을 듣지 마! 그녀는 단순한 꼭두각시야! 그녀의 행위는 전부 내 자극에 의해 일어난 것이고, 내가 돌아다니면서 활기를 불어 넣은 많은 뉴런들 속에 내재된 프로그램에 의해서 일어난 것이야!〉 그러나 설은 호질랜드 데몬의 경고에 귀를 기울이려 들지 않는다. 그는 이렇게 말한다. 〈그녀의 뉴런들은 여전히 적절한 인과력을 갖고 있다. 뉴런들은 단지 데몬의 도움을 필요로 할 뿐이다.〉

우리는 설의 원래의 설정과 수정된 설정 사이에서 대응 관계를 찾을 수 있다. 〈몇 개의 종잇조각〉은 여성의 뇌 속의 모든 시냅스에 해당한다. 〈몇 개의 종잇조각〉에 씌어진 인공 지능 프로그램은 그 여성의 뇌의 전체 구조에 대응한다. 그리고 이러한 전체 구조는 데몬에게 어떤 시냅스를 언제, 어떻게 자극할



것인지를 지시하는 거대한 명령에 해당한다. 또한 종이 위에 〈의미 없는 중국어의 꼬부랑 곡선들〉을 쓰는 행위는 그녀의 시냅스를 자극하는 행위에 대응한다. 가령 이 수정된 설정 중에서 크기와 속도를 제어하는 스위치만을 변경시키고 나머지는 그대로 놓아둔다고 가정해 보자. 그러면 우리는 그 여성의 뇌를 지구만한 크기로 부풀리게 될 것이고, 데몬도 작은 호질랜드 데몬이 아니라 〈우리와 같은 크기의〉 설 데몬이 될 것이다. 또한 설의 데몬이 그렇게 팽창한 뇌 속을 100만 분의 1초에 수천 킬로미터를 달리는 속도가 아니라 사람에게 부자연스럽지 않은 속도로 움직이게 하자. 이제 설은 우리가 어느 수준과 자신을 동일시하기를 원할까? 여기에서 이 문제를 심각하게 다루지는 않을 것이다. 그러나 앞의 경우에서 시스템 이론의 대응이 설득력을 갖는다면 이 경우에서도 마찬가지일 것이다.

우리는 설의 사고 실험이, 언어를 이해한다는 것이 무엇인가라는 문제를 생생하게 제기해 준다는 점을 인정해야 한다. 그렇지만 잠시 이 주제에서 벗어나기로 하자. 먼저 다음과 같은 문제를 생각해 보자. 〈문어(文語)나 구어 기호를 조작하는 어떤 종류의 능력이 그 언어의 진정한 이해에 해당하는가?〉 영어를 재잘거리는 앵무새는 영어를 이해하지 못한다. 전화로 시간을 알려주는 녹음된 여성의 음성은 영어를 이해하는 시스템의 음성이 아니다. 그 음성의 배후에는 어떠한 정신적인 것도 존재하지 않는다. 이 음성은 그 바탕에 해당하는 정신적인 기질 위에 떠 있는 거품이며, 단지 사람의 목소리처럼 들리는 특성을 갖고 있을 뿐이다. 어린아이라면 어떻게 그처럼 지루한 일을 하는 사람이 있을지 의아해할 수도 있을 것이다. 이런 사실은

우리를 즐겁게 한다. 물론 그녀의 목소리가 튜링 테스트를 통과할 수 있는 유연한 인공 지능 프로그램에 의해 작동된다면 문제는 달라질 것이다!

당신이 중국에서 어떤 학급을 가르친다고 가정해 보자. 그리고 당신은 자신의 생각을 모두 영어로 정식화하고, 마지막에 변환 규칙을 적용해서 영어로 표현된 생각을 기묘하고 〈무의미한〉 방식으로 입과 성대를 움직이는 명령으로 변환시킨다고 하자. 당신은 이러한 사실을 알고 있고, 학생들도 당신의 수업에 지극히 만족한다. 학생들이 손을 들어 알 수 없는 발성을 할 때, 그 소리는 당신에게 완전히 무의미하다. 그러나 당신은 그 소리를 처리할 장치를 갖고 있다. 즉 신속하게 반대 규칙을 적용해서 그 소리의 영어 의미를 복원시키는 것이다. ……그렇다면 당신은 진짜 중국어로 이야기하고 있다고 느낄까? 과연 중국인의 정신에 대한 통찰을 얻었다는 느낌을 받을 수 있을까? 아니, 정말 이러한 상황을 상상할 수 있을까? 거기에 어떤 현실성이 있을까? 이런 방법을 사용해서 실제로 어떤 사람이 외국어를 잘 구사할 수 있을까?

일반적인 표현은 〈중국어로 생각하는 방법을 배워야 한다〉이다. 그러나 중국어로 생각하는 방법이란 무엇일까? 이런 과정을 거친 사람이라면 누구나 다음과 같은 사실을 인정할 것이다. 외국어의 소리는 곧 〈들리지 않게〉 된다. 그러니까 외국어를 듣는 것이 아니라 그 소리를 통해서 듣게 되는 것이다. 이것은 우리가 창문 자체를 보는 것이 아니라 창문을 통해서 보는 것과 마찬가지이다. 물론 열심히 노력하면 친숙하게 사용하는 언어도 해석되지 않은 순수한 소리로 들을 수 있게 된다. 원하기만 하면 창문 유리를 볼 수 있듯이 말이다. 그러나 두 가지 일을 한꺼번에 할 수는 없다. 다시 말해서 의미를 가진 소리와 의미 없는 소리 자체를 동시에 들을 수는 없다. 그러므로 대부분의 경우 사람들은 주로 의미를 듣고 있는 것이다. 그 소리에 끌려 외국어를 배우는 사람들에게는 이런 이야기가 조금 실망스럽게 들릴 것이다. 그러나 더 이상 그 소리를 소박하게 듣는 것은 불가능하더라도 그 소리를 통달한다는 것은 아름답고 즐거운 경험이다(이런 종류의 분석을 음악을 듣는 체험에 적용하는 시도는 무척 흥미로울 것이다. 이 경우 소리만 듣는 것과 그 〈의미〉를 듣는 것의 차이가 무엇인지 이해하기는 훨씬 어렵겠지만 매우 실제적인 것처럼 보인다).

외국어 학습은 자신의 모국어를 초월하는 것을 포함한다. 그리고 그것은 새로운 언어를 사고(思考)가 발생하는 매체와 뒤섞는 것을 포함한다. 사고는 모국어 속에서처럼 새로운 언어 속에서도 쉽게(또는 그와 비슷한 정도로) 싹틀 수 있어야 한다. 어떻게 새로운 언어 습관이 조금씩 스며들어 마침내 뉴런으로까지 흡수되는지 그 방식은 여전히 큰 수수께끼이다. 그러나 한 가지 확실한 것은, 언어의 습득이란 당신이 그 언어를 의미 없는 소리와 부호의 집합으로 취급할 수 있게 해주는 규칙들의 프로그램을 수행하기 위해 〈영어의 하위 체계〉를 획득하는 것이 아니라는 것이다. 어쨌든 새로운 언어는 당신의 내적인 표상 체계(당신이 갖고 있는 개념, 이미지 등의 레퍼토리)와 깊숙이 융합되지 않으면 안 된다. 마치 영어가 영어 사용자들의 내적 표상 체계와 융합하는 정도로 말이다. 이 문제를 바르게 이해하기 위해서는 상당한 설명력을 갖는 컴퓨터 과학의 개념인

실행 수준levels of implementation이라는 분명한 개념을 이해할 필요가 있다.

컴퓨터 과학자들에게는 어떤 시스템이 다른 시스템을 〈에뮬레이트emulate〉할 수 있다는 개념에 친숙하다. 이 개념은, 모든 범용 디지털 컴퓨터는 다른 범용 디지털 컴퓨터를 가장할 수 있다는 앨런 튜링에 의해 1936년에 증명된 정리에서 유래한다. 여기에서 외부 세계에 대한 유일한 차이는 속도이다. 〈에뮬레이트〉라는 동사는 어떤 컴퓨터에 의한 다른 컴퓨터의 시뮬레이션을 뜻하는 것인 데 비해, 〈시뮬레이트〉는 허리케인, 인구 곡선, 국회 의원 선거, 심지어는 컴퓨터 이용자들과 같은 그 밖의 현상을 모형화하는 것을 가리킨다.

주된 차이는 시뮬레이션이 어떤 현상의 모형의 성질에 좌우되기 때문에 대개 근사적인 데 비해, 에뮬레이션은 가장 깊은 수준에까지 정확히 동일하다는 것이다. 에뮬레이션이 그처럼 정확하기 때문에 시그마5 컴퓨터가 DEC PDP-10처럼 다른 구조architecture를 가진 컴퓨터를 에뮬레이트할 경우, 이 기계의 이용자는 자기가 진짜 DEC를 다루고 있지 않다는 사실을 알아차리지 못할 것이다. 이처럼 어떤 아키텍처를 다른 아키텍처에 내재시키는 과정이 〈가상 기계virtual machine〉라 불리는 것을, 이 경우에는 DEC-10을 낳는다. 따라서 모든 가상 기계 밑에는 항상 다른 기계가 존재한다. 그런데 그 기계는 같은 종류의 기계일 수도 있고, 또 다른 가상 기계일 수도 있다. 『구조화된 컴퓨터 조직 *Structured Computer Organization*』이라는 책에서 앤드류 타넨바움Andrew Tanenbaum은 이러한 가상 기계라는 개념을 이용해서, 대규모 컴퓨터 시스템을 한 기계 위에서 다른 기계를 실행시키는 식으로 쌓아올린 일종의 가상 기계 낱가리로 설명했다. 물론 가장 밑에 있는 기계는 실재하는 기계이다! 그러나 각 수준은 다른 수준으로부터 물 한 방울 새지 않을 정도로 완전히 밀봉되어 있다. 그것은 설의 데몬이, 자신이 그 구성 부분인 중국어의 화자에게 이야기를 거는 것이 금지되어 있는 것과 마찬가지이다(어떤 종류의 대화가 이루어질지 상상해 보는 것도 흥미롭다. 설의 데몬은 중국어를 전혀 모르기 때문에 통역이 있다고 가정할 때의 이야기이지만).

그런데 이론적으로는 이러한 두 수준이 서로 의사소통하게 만드는 것이 가능하다. 그러나 이것은 전통적으로 바람직하지 못한 방식으로 간주되었다. 다시 말해서 수준 혼합은 금지된다. 그럼에도 불구하고 이 금단의 열매, 즉 두 실행 수준의 경계를 흐리는 것은 바로 사람의 〈시스템〉이 외국어를 배울 때 발생한다. 외국어는 일종의 소프트웨어 기생충처럼 모국어 위에서 기능하는 것이 아니라 모국어와 같은 가장 근본적인 (또는 그에 가까운) 수준에서 하드웨어 속에 이식된다. 어쨌든 외국어 학습은 그 사람 속에 내재해 있는 〈기계〉에 깊은 변화를 일으킨다. 그것은 뉴런이 발화하는 방식에 대한 방대하고 정합적인 일련의 변화이다. 이 변화는 너무도 포괄적이기 때문에 더 높은 수준의 존재자들, 즉 기호가 서로를 촉발시키는 새로운 방식을 창조한다.

이 과정을 컴퓨터 시스템에서 설명하면, 더 높은 수준의 프로그램은 그 프로그램을 수행하는 〈데몬〉의 내부에 변화를 일으킬 수 있는 방법을 가져야 한다. 이것은 한 수준을 다른 수준 위에 엄격하게 수직적이고 전면적인 방식으로 실행시키는 현재

컴퓨터 과학의 양식과는 전혀 다른 방식이다. 고차 수준이 그보다 낮은 수준 즉 그 기초로 내려가거나 그것에 영향을 미칠 수 있는 능력을 갖는다는 것은 일종의 마술적 트릭이며, 우리는 이런 트릭이 의식의 본질에 매우 가깝다고 생각한다. 언젠가는 이러한 트릭이 컴퓨터 설계의 유연성을 높이는 핵심적인 요소이며, 인공 지능에 대한 접근의 열쇠가 된다는 사실이 입증될지도 모른다. 특히 〈이해〉가 무엇을 의미하는가라는 질문에 대해 만족할 만한 답을 얻기 위해서는 기호 조작 시스템 내부의 서로 다른 수준들이 상호 의존적으로 작용하는 방식을 좀 더 분명하게 묘사할 필요가 있다는 데에는 의심의 여지가 없을 것이다. 이러한 개념들은 파악하기 힘들다는 것이 입증되었고, 그 명확한 이해를 얻기까지는 아직도 갈 길이 멀다.

여러 수준에 대한 조금쯤 혼란스러운 논의에서 여러분은 도대체 〈수준〉이 무엇을 뜻하는지 의문을 품을 수 있을 것이다. 이것은 대단히 어려운 물음이다. 설의 데몬과 중국어로 이야기하는 여성 사이에서처럼 각각의 수준이 서로 밀봉되어 있는 경우에는 그 의미가 분명하다. 그런데 수준이 불분명해지기 시작할 때 주의할 필요가 있다! 설은 자신의 사고 실험 속에 두 개의 수준이 있다는 것을 인정하겠지만 두 개의 관점(느낄 수 있고, 〈경험을 갖는〉 두 개의 진짜 존재)이 있다는 것은 인정하기를 꺼린다. 그는 일단 몇 개의 컴퓨터 시스템이 경험을 가질 수 있다는 것을 인정해 버리면 그것이 판도라의 상자가 되어 갑작스럽게 〈마음은 모든 곳에 있다〉는 (위나 간이나 차의 엔진 등에도) 것을 인정하게 되는 사태를 우려하는 것이다.

설은 어떤 시스템이든 인공 지능 프로그램의 구현으로 기술



하는 방법을 열심히 찾기만 하면 그 시스템이 사고와 감정을 갖는다고 인정할 수 있다고 생각하는 것 같다. 이것은 분명 범심론(汎心論)으로 이어지는 골치 아픈 생각이다. 실제로 설은 인공 지능 연구자들이 본의 아니게 범심론적 세계관에 관여해 왔다고 믿는다.

설이 자기가 파놓은 함정을 피할 수 있는 길은, 여러분이 도처에서 마음을 발견하기 시작할 때, 생명 없는 대상 속에서 찾아내게 되는 이러한 모든 〈사고〉와 〈감정〉이 진짜가 아니고 〈가짜 pseudo〉라고 주장하는 것이다. 그것들은 의도성이 없어! 뇌의 인과력이 없어!(설은 의도성이나 뇌의 인과력이라는 관념을 〈혼〉이라는 소박한 이원론적 개념과 혼동하지 말라고 경고한다.)

다른 한편, 우리의 탈출로는 애당초 함정 따위는 없다고 주장하는 것이다. 우리는 뇌가 자동차 엔진이나 간 속에 없듯이 마음도 그 속에 없다고 말하고 싶다.

이 점은 조금 더 설명할 필요가 있다. 내용물을 뒤섞고 있는 위 속에서 진행되는 사고 과정의 복잡성을 볼 수 있다면, 탄산음료 속에 있는 거품들의 패턴을 쇼팽의 마단조 피아노 협주곡을 코드화한 것으로 읽지 못할 이유가 어디 있겠는가? 그리고 스위스 체스의 구멍은 미국의 전체 역사를 코드화한 것은 아닐까? 분명 그것들은, 영어든 중국어든 코드화하고 있다. 결국 모든 것이 도처에 적혀 있는 것이다! 바흐의 「브란덴부르크 협주곡 제2번」은 햄릿의 구조 속에 코드화되어 있고, 햄릿은 (그 코드를 알기만 하면) 여러분이 게걸스럽게 먹는 생일 케이크의 마지막 조각의 구조로부터 읽어낼 수 있는 것이다.

이 모든 경우에 문제는 읽어내고자 하는 것을 미리 알지 못

한 채 문제의 코드를 지정하는 것이다. 그렇지 않으면 제멋대로 구성한 사후적인 posteriori 코드에 의해 야구 경기나 풀잎으로부터 모든 사람의 정신 활동에 대한 기술을 이끌어낼 수 있을 것이다. 그러나 이것은 과학이 아니다.

분명 마음은 여러 가지 다른 정교함의 정도로 나타난다. 그러나 마음이라고 부를 수 있는 마음은 오직 정교화된 표상 체계가 존재하는 곳에만 존재한다. 시간적으로 일정함을 유지하는 어떤 기술 가능한 사상(寫像)도 자동차 엔진이나 간 속에 끊임없이 스스로를 갱신하는 표상 체계가 존재한다는 것을 드러내지 않을 것이다. 어쩌면 사람들이 대피라미드나 스톤헨지, 바흐의 음악, 셰익스피어 희곡 등의 구조 속에서 부가적인 의미를 읽어내는 것과 거의 같은 방식으로, 즉 억지로 수비학적(數秘學的)인 사상 체계 mapping scheme를 날조해서 해석자의 열망을 무엇이든 만족시킬 수 있을지도 모른다. 그러나 그것이 설이 의도하는 것인지는 의심스럽다(우리는 설이 실제로 그것을 의도하고 있다고 생각하지만).

마음은 뇌 속에 존재하며 프로그램된 기계 속에 존재하게 될지도 모른다. 이러한 기계가 출현한다면, 이 기계가 갖는 인과력은 기계를 구성하는 물질에서 유래하는 것이 아니라 기계의 설계와 기계 속에서 작동하는 프로그램에서 유래하는 것이다. 그리고 그 기계가 인과력을 갖는다는 것을 알 수 있는 방법은, 그 기계에 말을 걸고 기계가 하는 이야기에 세심하게 귀를 기울이는 것이다.

D. R. H.





# 어느 불행한 이원론자

레이먼드 스멀리언

옛날에 한 이원론자가 있었다. 그는 마음과 물질이 독립된 실체라고 믿었다. 그는 마음과 물질이 실제로 어떻게 상호 작용하는지에 대해서는 전혀 신경 쓰지 않았다. 그것은 삶의 〈수수께끼〉 중 하나였다. 여하튼 그는 마음과 물질이 제각기 독립된 실체라고 확신하고 있었다.

불행하게도 이 이원론자는 견딜 수 없을 만큼 고통스러운 삶을 보내고 있었다. 그것은 그의 철학적 신념이 아니라 전혀 다른 이유 때문이었지만, 그리고 그는 남은 생애에도 그러한 불행에서 벗어날 수 없으리라는 충분한 증거를 갖고 있었다. 이제 그는 죽는 것 이외에는 아무것도 바라지 않았다. 그러나 그는 다음과 같은 이유로 자살을 포기했다. (1) 그는 자신의 죽음으로 누군가가 상처받는 것

* Raymond M. Smullyan, "An Unfortunate Dualist" This Book Needs No Title (Prentice-Hall, Inc., Englewood Cliffs, N. J. 1980).