인 지 과 학
Korean Journal of Cognitive Science
Vol. 8, No. 1(1997)

인공지능과 이해의 개념

AI and The Concept of Understanding

김선희†
Sun-Hie Kim

요 약

컴퓨터는 인간처럼 생각하고 이해할 수 있는가? 이 논문은 인공지능의 가능성을 지지하는 논의와 반대하는 논의들을 분석한다. 컴퓨터의 이해능력에 대한 인정과 부인은 기본적으로 튜링과 써얼의 입장으로 대변된다. 여기서 필자는 그들의 논증을 비판적으로 검토함으로써, 인공지능의 사고 및 이해 개념은 무엇이며 그러한 이해 개념은 인간의 이해능력에 어느정도 상응하는 개념인지 논의하려고 한다. 특히 주관적 의식의 문제에 있어서, 인공지능논제는 어떤 어려움을 가지며 그 이유는 무엇인지를 논의할 것이다. 물론 주관적 의식이 이해능력의 필수조건이라고 보기는 어렵다. 그러나 언어를 갖는 인간의 조건과 관련하여 이 문제를 고려해 볼 때, 우리는 신체와 의식의 통일체로서의 인간을 전제하며, 인간의 신체 행동과 더불어 주관적 의식은 인격의 징표로 작용한다.

주제어 인공지능, 이해, 기능주의, 튜링기계, 써얼의 중국어방, 의식, 주관성, 감각질, 인격

ABSTRACT

Can the appropriately programmed computer think? I analyse, in this paper, arguments for and against strong AI-thesis, basically Turing's argument and Searle's Chinese room argument. Through a

† 서강대학교 철학과 강사
연세대 철학연구소 전문연구원
Sogang University, Philosophy Department
Yonsei Research Institute for Philosophy

critical review of these arguments. I try to show that the supporters of AI-thesis like Turing fail to explain the subjective nature of human consciousness. However, I do not think that subjective consciousness is a necessary condition for the ability to understand language. (In this respect my views are different from Searle 's). But when we consider the conditions of humans as language users, we should presuppose that a human being is the unity of body and mind (or consciousness). Therefore, our subjective consciousness, together with human body(thus, ways of our behavior and life), serve as a mark of person.

Keywords AI, understanding, functionalism, Turing machine, Searle's Chinese room, consciousness, subjectivity, qualia, person

1. 문제 제기

컴퓨터는 생각할 수 있는가? 언어를 이해하고 사용할 수 있는가? 즉 컴퓨터는 심적 상태와 인지 상태들을 갖는 마음과 같은 것인가? 나아가 컴퓨터는 하나의 인격으로 간주될 수 있는가? 이 물음에 대한 논쟁들을 보면, 거기에는 사고나 이해 등의 심적 개념들에 대한 모종의 차이가 존재한다는 것을 알 수 있다. 그렇다면 컴퓨터의 사고와 언어 이해능력에 대한 생산적이고 진전된 논의를 위해서는, 대립하는 논쟁들 사이에서 컴퓨터는 어떤 의미로 생각하거나 이해할 수 있고 어떤 의미로 생각할 수 없다는 것인지를 구체적으로 살펴보는 것이 필요할지 모른다. 이것은 인공지능의 가능성 문제는 경험적 문제와 더불어 개념적 문제가 있다는 것을 의미한다.

이 글은 인공지능의 가능성을 지지하는 논의와 반대하는 논의들을 분석하는 것으로부터 시작한다. 컴퓨터의 사고나 이해능력에 대한 인정과 부인은 기본적으로 튜링과 써얼의 입장으로 대변된다. 여기서 그들의 논증을 비판적으로 고찰함으로써, 인공지능의 사고/이해 개념은 무엇이며 그러한 이해 개념은 인간의 이해능력에 어느 정도 상응하는 개념인지 논의하려고 한다. 특히 주관적 의식의 문제에 있

어서, 인공지능논제는 어떤 어려움을 가지며 그 이유는 무엇인지를 논의할 것이다. 이 논의를 통하여 인공지능은 인간 마음의 어느 수준을 반영하는지, 그리하여 인공지능논제가 어느 정도의 타당성 내지 한계를 갖는지를 검토하려고 한다.

2.튜링기계의 사고 개념과 AI논제

컴퓨터와 같은 기계가 인간처럼 생각하고 이해할 수 있는가? 사고와 이해력에 관하여 인간과 기계를 올바로 비교하기 위해서는, 둘 사이의 어떤 점을 비교해야 하는지가 중요하다. 만일 비교의 초점을 외양과 질료에 맞춘다면, 둘 사이의 차이가 확연하여, 기계의 인지가능성은 곧바로 부정될 것이다. 그런데 인간의 마음과 심신 관계에 대한 여러 입장들 중에서, 기능주의는 그러한 질료적인 제약을 넘어서서 인공적 체계가 인간과 같은 지능을 가질 수 있는 가능성을 보여주었다. 그것은 어떤 두 체계들은 물리적 하위차원에서 다를지라도 상위차원의 기능 상태에서 동일할 수 있다는 것에 착안한 것이다. 기능주의에 의하면, 심적 상태는 기능상태와 동일하므로, 인간 두뇌 이외의 어떤 것(예, 컴퓨터나 인공지능체계)이 동일한 기능 상태에 있다면 그것들은 또한 동일한 심적 상태에 있어야 한다고 주장한다. 그리하여 마음에 대한 기능주의는 컴퓨터도 심적 상태를 가질 수 있는 가능성, 즉 인공지능의 가능성을 함축한다. 이런 점에서 인공지능 이론은 기능주의와 한 짝을 이루어온 것이 사실이다.

이와 같이 컴퓨터의 사고 가능성의 초점을 그것의 기능에 둠으로써, 컴퓨터의 사고/이해 가능성을 테스트할 수 있는 기능주의적 방식이 고안되었다. 그것은 튜링에 의해 고안된 "모방게임(imitation game)"이다. 튜링은 컴퓨터도 인간과 같이 사고능력을 갖는다고 볼 수 있는 조건을 다음과 같이 제시한다.[1]

A는 여자를 가장한 남자이고, B는 여자이며, C는 두 사람을 볼 수 없는 방에서 두 사람에게 질문을 하고 그들의 답변을 듣고 판단하여 누가 여자이고 누가 남자인지를 알고자 하는 사람이다. B는 C의 여러 물음들에 대하여 자신이 여자라는 것을 사실대로 대답한다. 그러나 A는 C의 모든 물음들에 대하여, C로하여금 자신이 여자라고 믿도록 하기위해 B가 대답하는 것을 모방하고 그러한 역할을 B보다 더 잘 해낼 수 있다. 이것이 모방게임이다. 그런데 이제 이러한 게임에서 A대신에 컴퓨터(A

1) A.M.Turing(1950), "Computing Machinery and Intelligence," in THE MIND'S I, eds., D.R.Hofstadter and D.C.Denett, (New York: Basic Books, 1981), pp.53-68.

*)가 그런 모방 역할을 하도록 할 수 있다. 튜링에 의하면, 이런 상황에서 "컴퓨터는 생각할 수 있는가?"하는 물음은 다음과 같은 방식(2)으로 제기될 수 있다고 한다.

(1) 컴퓨터는 생각할 수 있는가?

(2) 판단자 C는 A(여자를 가장한 남자)와 B사이의 모방게임에서 실수하는 정도로, A*(컴퓨터)와 B사이의 게임에서도 실수를 저지르는가?

즉 물음 (1)은 (2)로 대치될 수 있고, 후자에서 긍정되는 정도에 따라 여자를 가장한 남자에게 부여하는 만큼의 사유능력을 (여자를 '가장한') 컴퓨터에게도 부여할 수 있다는 것이 튜링의 취지이다. 여기서 C가 A에게 속는 정도로 컴퓨터(A*)에게도 속는다면, 속임 주체가 갖는 사유능력(속이려는 의도, 믿음, 가장, 판단..등)을 A에게 귀속시키는 것과 마찬가지로 컴퓨터(A*)에게도 마땅히 그런 능력을 귀속시킬 수 있다. 다시 말해서 후자의 튜링 테스트를 통과하는 기계는 또한 생각할 수 있는 것으로 간주되어야 한다는 것이다.

그러면 튜링 테스트를 통과하는 기계는 사고능력을 갖는다고 할 때, 그러한 기계가 갖는 사고나 이해력은 어떤 종류의 것인가? 디지탈 컴퓨터란 의미나 해석을 지칭함이 없이 형식적 체계의 규칙들을 따라 물리적 항목/개항들을 조작할 수 있는 하나의 기계이다.[2] 순수한 형식적 체계로서의 컴퓨터는 의미론적(또는 화용론적) 특징의 도입없이 통사론적 특징들만으로 발생하는 계산적 과정/속성들에 의해 인지/사고의 기능을 드러낸다. 따라서 컴퓨터의 사고/인지 개념은 형식적-계산적-통사적 특징들 만으로 정의될 수 있다. 즉 튜링테스트를 통과하는 컴퓨터는, 인간의 인지 혹은 심리상태와 동일한 기능을 가짐으로써, 마음이나 인지체계로 간주된다. 이와같이 컴퓨터의 사고나 이해력은 그 형식체계의 통사적 규칙과 속성들에 기초하는 계산적 과정/상태들이며, 그러한 계산적 과정들이 사고자/인지자(인간)와 같은 기능들을 수행하는 셈이다.

그러면 튜링-테스트에 성공하는 컴퓨터/기계는 마음을 갖기위한 충분조건을 만족하는가? 컴퓨터의 통사적-이해 개념은 인간의 이해력을 위한 충분조건인가? 아마도 기능주의적 인공지능 옹호자는 이 문제에 대해 긍정적인 태도를 취할 것이다. 그런데 써얼은 인공지능에 대한 강한 명제와 약한 명제를 구분한다.[3]

2) Todd C.Moody(1993), Philosophy and Artificial Intelligence (Prentice Hall, 1993), p.70.

3) John Searle(1980), "Minds, Brains, and Programs", in eds., D.R.Hofstadter and D.C.Denett, p.353.

(3) 인공지능의 강한논제: 적절히 프로그램된(튜링 테스트를 통과하는) 컴퓨터는—인지 상태를 이해하고(understand) 갖는다는 의미에서—그 자체가 하나의 마음이다(즉 마음과 동일하다). 즉 튜링테스트를 통과하는 것은 무엇이든지 심적 상태를 가지며 이해를 한다고 말할 수 있다. 따라서 적절히 프로그램된 컴퓨터가 튜링 테스트를 통과할 수 있다면, 그것은 진정으로 언어를 이해하는 것이다.

(4) 약한논제: 심적 과정들에 대한 컴퓨터 모델의 탐구와 발전은 마음의 탐구를 개선할 수 있다. 컴퓨터는 마음의 연구에서 계산성과 엄밀성을 위해 도움을 줄 수 있는 하나의 도구이다.

이렇게 구분할 때, 우리의 논의와 직접적으로 관련된 것은 강한논제이다. 약한 인공지능 논제는 인지과학의 방법론적 관심에 한정되어 있으므로 컴퓨터의 사고 능력에 관한 우리의 논의(인공지능 혹은 컴퓨터는 진정으로 언어를 이해하는가?)에서 특정한 관점을 취하지 않는다. 단지 계산적 속성으로 모델화할 수 있는 마음의 영역이 있다는 것(그리하여 컴퓨터 모델이 마음의 탐구에 부분적으로 도움을 줄 수 있다는 것)만을 전제하면 약한논제는 성립한다. 그리고 그 전제는 마음을 단지 형식적 계산체계로(그리고 심적 상태를 통사적인 계산 상태로) 간주하거나 간주하지 않는 어느 쪽에서도 거부할 필요가 없을만큼 약한 것이다. 그런 점에서 인공지능에 대한 반론과 논쟁들은 거의 대부분 약한논제 보다는 강한논제를 표적으로하고 있다. 컴퓨터의 사고와 언어이해능력에 대한 우리의 논의도 강한 논제에 대한 지지나 반박과 관련되어 있다.

3. 써얼의 반박: 중국어 방 논증

써얼은 인공지능의 강한논제를 거부하면서, 컴퓨터의 사고와 이해 능력에 대하여 부정적 입장을 취한다. 그는 튜링의 조건을 만족하지만 사유/이해능력을 갖지 않는 경우를 예시한다. 그의 유명한 '중국어 방' 논증이 그것이다. 써얼은 실제로 자신의 모국어는 영어이고 중국어는 전혀 모르는 상황을 가정한다.

나는 어떤 방 안에 갇혀 있다. 이 방에는 중국어 단어들이 모두 들어있는 상자가 있고, 외부에서 중국어로 묻는 물음들에 대하여 대답을 구성할 수 있는 방법에 관한 훌륭한 프로그램이 따라야할 규칙들이 주어져 있다. 어느 누구도 중국어로 된 나의 답변을 읽고 내가 중국어를 모른다고 할 수 없다. 내가 영어로 된 물음에 대해 영어로 대답할 때가 있다. 외부의 관점에서 보면, 중국어의 물음과 영어의 물음에 대한 나의 대답들은 동등하게 좋을 것이다. 그러나 중국어의 경우는 영어의 경우와 달리, 나는 해석되지 않은 형식적 기호들을 조작함

으로써 그 대답을 구성한 것이다. 중국어에 관한한 나는 단순히 컴퓨터 처럼 작용했다. 즉 나는 형식적 규칙들의 요소에 따라 계산적 기능을 수행한 것이며, 이 경우 나는 컴퓨터 프로그램의 단순한 한 사례에 불과하다.[4]

이 논증이 주장하는 바는 분명하다. 폐쇄된 방 안에서 중국어로 답변하는 그 사람은 단지 어떤 지시 메뉴얼에 따라 형식적 프로그램을 실행할 뿐이며, 중국어를 전혀 이해하지 못하고 있다. 마찬가지로 튜링 테스트를 통과하도록 프로그램된 컴퓨터가 하는 것도 이와 다를 바 없다. 따라서 그런 프로그램의 실행이 중국어 이해를 위한 충분조건이라는 것, 즉 튜링테스트의 통과가 이해력의 충분조건이라는 강한 인공지능 논제는 잘못된 것이다.

써얼의 관점에 의하면, 컴퓨터나 튜링기계는 형식적 통사구조(formal syntactic structure)에 의해 조작될 뿐, 기계(형식적 인공체계)는 기호들을 다루고 처리함에 있어서 그것이 무엇을 의미하며 무엇을 하고 있는 것인지 알지도 이해하지도 못한다. 마치 노트 안의 중국어 기호들을, 다른 기호들과 그것들을 짝지우는 형식적 규칙들에 따라 번역하는 사람이 중국어 노트 안의 고유한 "이야기"에 대해서 아무것도 모르는 것과 마찬가지로, 기계도 인간의 기호들에 대한 고유한 "이야기"에 대해 아무것도 이해하지 못한다. 써얼에 의하면, 중국어의 이해는 통사적으로 정의된 개항들에 대한 형식적 조작/작용들 그 이상의 것이라고 본다. 언어의 이해능력은 형식적인 통사구조에 의해 완전히 정의되는 기계의 능력을 넘어서는 것이다.

써얼의 논증의 요지는, 어떤 프로그램된 컴퓨터도 그것의 형식적 혹은 계산적 속성들 만으로는 이해력을 가질 수 없고 심적 현상을 산출할 수도 없다는 것이다. 단지 개항들의 통사적 조작이나 형식적 또는 계산적(formal or computational) 속성만으로는 어떤 체계도 (컴퓨터든, 두뇌든) 이해나 다른 지향적 심적 상태들을 가질 수 없다. 이해력을 위해서는 그 이외의 다른 무엇에 의존하는 것이 필요한데, 그것은 두뇌의 "인과력(causal powers)"이다.[5] 언어의 이해와 같은 심적 현상은 세계와 인과적으로 상호작용할 수 있는 능력을 갖는

4) John Searle(1980), pp.358-373.

5) 그러나 인공지능에 대한 연결주의(connectionism)의 접근은, 사고나 이해 등의 인지활동은 전체 신경망의 상호작용으로부터 나온다고 보고 컴퓨터는 그러한 총체적 체계(holistic system)를 모방할 수 있다고 생각한다. 연결주의

것이 필요하며, 언어적 경험과 비언어적 경험 사이에 어떤 관련을 맺는 의미론적 구조가 필요하다. 즉 한 체계가 이해력을 갖기위해서는 그 체계의 올바른 "인과력"이 요구되는데, 두뇌라는 생물학적 체계만이 그 요구를 만족한다는 것이다.[6]

그렇다면 써얼의 논증은, 어떤 인공지능도 심적 상태를 가질 수 없다는 것이 아니라, 심적 상태를 갖기위한 어떤 체계라도 두뇌의 인과력을 복제해야한다는 것을 의미한다. 즉 "두뇌 인과력이 (언어) 이해의 필수조건이라는 것"이다. 따라서 만일 인간의 두뇌를 복제하는 인공지능이 가능하다면, 그것은 생각할 수 있고 이해할 수 있는 마음과 같은 것이 된다. 그리하여 써얼은 '기계는 생각할 수 있는가?' 하는 물음에 대해, "특별한 종류의 기계들, 즉 두뇌와 그리고 두뇌와 동일한 인과력을 갖는 기계들만이 생각할 수 있다"고 결론내린다.[7] 이 점에서 써얼은 기능주의를 떠나 두뇌 물리주의로 귀결한다.[8]

4. 인공지능의 이해 개념과 의식의 문제

앞서 제시했던, 컴퓨터의 사고/이해 능력을 지지하는 튜링의 논증과 그것을 반박하는 써얼의 논증을 상기해 보자. 그런데 '컴퓨터는 사유할 수 있다(혹은 언어를 이해할 수 있다)'는 논증에서 사용된 사고와 이해 개념은 '컴퓨터는 사유나 이해를 할 수 없다'는 논증에서 사용된 이해 개념과 다르다는 것을 알 수 있다. 때로는 하나의 논증 안에서도 이해의 개념은 다의적으로 사용되는 것 같다. 이와 같이 인공지능의 철학적 논증들을 보면, 컴퓨터에 대한 문제만이 아니라, 생각, 이해, 믿음 등과 같은 심성적 개념들에 관한 문제가 함께 얽혀 있다. 그렇다면 컴퓨터의 이해/사고 능력에 대

에 의하면, 연결적 기계/컴퓨터는, (병렬처리 방식을 통하여) 상호작용하는 그물망으로 연결되어서, 두뇌 신경체계와 유사한 방식으로 작용한다. 따라서 두뇌가 막대한 양의 뉴런 그물체계에 의해 실재 인지기능들을 성취하는 것 처럼, 신경 그물체계에 기초한 컴퓨터도 그와 동일한 것을 성취할 수 있으리라는 것이다.

6) J.Searle(1980), 370-372.

7) J. Searle(1980), p.372.

8) 써얼은, 감각질 및 의식은 뇌의 물리적-생물학적 과정들에서 일어나지만, 우리와 기능적으로 식별불가능한 전자체계(컴퓨터)는 의식을 갖지 않는다고 주장한다. 즉 의식이 두뇌의 물리적 속성에 수반한다고 보지만, 그것이 기능적 속성에 수반한다는 것은 부정한다. 이와 같이 써얼은 반기능주의적-비환원적-물리주의를 채택한다(The Rediscovery of Mind).

한 물음에 보다 구체적으로 접근하기 위해서는, 컴퓨터/인공지능의 이해 개념과 우리 인간의 이해 개념 사이에 어떤 간격이 존재하는지, 그리고 만일 간격이 있다면 그것들 사이의 차이와 관계는 무엇이며 또한 그 간격은 원칙적으로 메꿀 수 있는 것인지 여부를 고찰하는 것이 필요하다.

무디(T.C.Moody)에 의하면, 인공지능의 철학적 논쟁에 나타나는 이해(understanding)의 개념은 네가지로 구분된다. 첫째, 튜링테스트에 의해 도달되는 이해의 단계로서, 언어적 개항들(tokens)을 수용하고 적절한 답변들로 반응할 수 있는 언어 사용자와 같은 기능들을 수행하는 능력이다(F-이해). 이것을 기능적 혹은 통사적-이해라고 부르자. 둘째는 (통사적-형식적-계산적 방식이 아니라) 두뇌의 인과력을 복제하는 방식으로 언어적 기능을 수행하는 능력으로서(C-이해), 인과적-이해라고 부르자. 세째는 지향성을 유지할 수 있도록 언어적 대상과 비언어적 입력을 상호 연관지움으로써 도달하게 되는 이해의 차원으로(I-이해), 지향적 이해라고 부르자. 네째는 이해의 경험적, 주관적 측면으로서(S-이해) 주관적 혹은 의식적-이해라고 할 수 있다.[9]

이 네가지의 이해 개념을 가지고 인공지능을 둘러싼 논쟁들을 재검토해 보자. 아마 통상적으로 '이해'의 의미에는 이 네가지가 모두 결합되어 있다고 생각할 것이다. 혹은 이것들 중에서 어느 것이 더 이해의 본질적 요소인지에 대하여 논쟁이 있을 수 있다. 기능주의자는 F-이해(기능적-통사적 이해)가 이해의 충분조건이라고 주장할 것이다. 그러면 기능주의적 인공지능의 이해 개념을 "통사적(혹은 기능적)-이해"라고 하자.

이미 살펴보았듯이, 튜링테스트를 통과하는 컴퓨터나 인공지능은 "기능적-통사적 이해"를 가짐에 틀림없다. 튜링과 같은 기능주의자는 기능적-이해가 사고나 이해를 위한 충분조건이라고 본다. 즉 인공지능의 강한논제에 의하면, 인공지능은 '기능적 이해'의 단계에 머무른다는 것이 아니라(만일 이런 의미의 약한 주장이라면, 써얼의 반박은 표적이 없는 것이 된다), 기능적-이해를 갖는 튜링 기계는 곧 인간의 인지/이해를 갖기에 충분하다는 것이다. 그렇다면 이 주장을 지지하기 위해서는, 강한인공지능 지지자는 기능적 이해가 인과적 이해나 지향적 이해를 위한 충분조건이라는 것을 보일 수 있어야 한다. 나아가 기능적 이해를 토대로하여 주관적-이해의 현상도 설명할 수 있어야 한다.[10]

써얼의 중국어 방 논증은, 컴퓨터의 통사적-기능적 '이해' 만으로는 이해의 능력/자격에

9)Todd C. Moody(1993), p.149.

미치지 못한다는 반박이었다. (지향적) 심적 상태나 이해력을 갖기위해서는 두뇌 인과력 및 의미론적 특성이 필요하며, 다시 말해서 "기능적-이해"만이 아니라 "인과적-이해"가 필요하다는 것이다. 즉 기능적-이해는 인과적-이해를 보증하지 못하기 때문에 이해를 위해 불충분하다고 써얼은 주장한다. 그런데 그 논증을 자세히 살펴보면, 기능적-이해는 인과적-이해를 산출하지 못하기 때문에 진정한 이해에 도달하지 못한다고 주장하는 한편, 또한 중국어 방 안의 사람은 —인과적-이해만이 아니라— 주관적-이해를 결여한다는 것을 지적하고 있다.[11] 즉 중국어 방의 사고실험을 통하여 그가 보이고자한 점은, 중국어 방 안에 있는 사람은 자신의 관점에서(주관적/내부적으로) 이해의 경험/의식을 갖지 못하기 때문에 그는 진정으로 언어를 이해하고 있지 않다는 것이다. 이 점은 그가 중국어 방 밖의 사람과 안의 사람을 대조시키고, 따라서 제3자의 관점과 일인칭 관점을 대조시키고 있는 데에서 잘 드러난다. 사실상 튜링에 대한 비판은 이러한 대조를 전제할 경우에만 제기될 수 있다.

이 점을 명시적으로 보이기위해, 중국어 방 논증(이하 CR 논증)을 약간 수정하여 다음의 CR* 논증을 구성해 보자. 중국어 방에는 원래 있었던 S1 이외에, S2가 함께 있다. S1과 달리 S2는 중국어에 충분히 숙달된 사람으로서, 밖에서 주어지는 중국어 질문의 의미를 이해하고 의식적으로 중국어 규칙에 따라[12] 적절한 대답을 한다. 중국어 방 밖의 사람은 두 사람의 대답을 듣고 동등하게 훌륭하다고 판단한다. 물론 이 가정적 상황에서, 방 안에 있는 S1(통사적-화자)과 S2(의미론적-화자)의 이해 방식은 매우 다르다. 비록 S1 자신은 중국어를 이해하지 못할지라도 계산적-통사적 기능을 수행함으로써 '이해'를 위한 3인칭 테스트를 통과할 수 있다. 즉 중국어를 통사적이고

10) 그렇다면 강한인공지능논제는, 다음 물음들에 대하여 적절한 답변을 제시할 수 있어야 한다. 1.인공지능의 기능적-이해는 인과적-이해에 도달할 수 있는가? 2.기능주의적-인공지능은 내용을 갖는 지향적-이해에 도달할 수 있는가? 3.인공지능은 이해의 주관적 관점을 가질 수 있는가? 둘째 물음과 관련하여, 넓은 내용론을 주장하는 퍼트남이나 버지의 논의들은 기능적-이해와 지향적-이해 사이의 간격을 보여준다. 비록 기능적-이해에서 동일해도, 튜링기계의 경우는 인간의 이해와 달리 환경적 요소나 언어공동체적 요소에 의해 결정되는 넓은 내용이 아니다. 즉 인공지능과 인간의 믿음들은 기능적으로 동일할지라도 다른 믿음일 수 있다(넓은 내용에서 다를 수 있다).

11) John Searle(1992), The Rediscovery of Mind, The MIT Press, 제4장 참고.

12) 사실상 언어규칙을 따른다는 것 조차 기능주의적으로 설명하기 어렵다. 언어규칙을 따른다는 것은 단지 통사적 기능 이상의 것으로서, 언어의 사용능력과 언어공동체 등을 전제하기 때문이다.

기능적으로 조작하는 주체 자신은, 중국어 이해에 도달하지 못하고도 '이해'의 삼인칭 관점을 통과한 셈이다. 써얼이 보기에, S1의 통사적-이해는 주관적-이해를 결여하므로 그는 이해 상태에 있지 않다. 반면에 S2는 일인칭 주관적 관점의 이해력을 가지므로 그는 진정으로 중국어를 이해한다고 할 수 있다. 이 논증은, 〈언어를 이해하기 위해서는 일인칭 주관적 관점에서 이해의 경험과 意識을 가질 수 있어야 한다〉는 것을 함축한다.

중국어 방 논증을 이와 같이 이해할 때, 튜링에 대한 써얼의 문제 제기는 보다 정확히 표현하면 다음과 같은 것이다. 〈튜링 테스트를 통과하는 컴퓨터의 모방은, 여자를 가장한 남자와 같이, 속임/가장의 의도와 의식을 포함하는가?〉 즉 튜링 테스트를 통과한 컴퓨터는 사실상 자신의 의도 및 사고를 의식하고 있는가? 혹은 CR*논증에서, S1은 S2 처럼 중국어를 이해하고 있다는 주관적 의식이나 경험을 갖는가? 써얼은 바로 일인칭 주관적 관점의 이해력을 문제삼고 있는 것이다.[13]

질문하고 있는 C의 객관적 관점(혹은 제3의 관찰자 입장)에서는, 그 컴퓨터가 튜링테스트를 통과했다면, 여자를 가장한 남자(A)와 여자를 모방하는 컴퓨터(A*)에게 동등한 사고능력을 부여함이 마땅하다. C의 관점에서는 A와 A*를 구분하거나 둘 사이의 사고 능력을 구분할 이유도 없고 방법도 없다. 즉 그는 A가 나를 속였듯이 컴퓨터도 나를 '속였다'고 말할 수 있고, 그렇다면 컴퓨터를 속임의 주체 즉 사고의 주체로 간주해야 한다. 마찬가지로 (객관적으로 관찰할 수 있는 입력-질문과 출력-답변이 동일하므로) 중국어 방 밖의 사람은 S2만이 아니라 S1도 중국어를 이해하는 것으로 간주해야 할 것이다.[14] 이것이 바로 튜링과 기능주의적 AI의 주장이었다. 결국 튜링테스트는 제3자의 객관적 관점에서 보는 사고능력의 조건임을 알 수 있다.[15]

다른 한편 사고주체의 주관적 관점에서 보면, 사고자(즉 컴퓨터)는 속이려는 의도나 의

13) 이 문제는 Searle(1981)보다 (1992)에서 더욱 분명히 제기되고 있다.

14) 중국어 방을 마음으로 간주할 때, 기능주의는 행동주의와 달리 방 안의 어떤 존재 및 내적 과정을 인정하지만 그렇게 상정된 내부적인 것들 역시 삼인칭 객관적 관점에서 관찰할 수 있는 것에 불과하다는 점에서 사실상 외부적인 것이다. 즉 방 안을 들여다 보았을 때 거기에는 관계적 기능/속성들이 있을 뿐 의식과 같은 본래적 속성이 있는 것이 아니다. 이런 점에서 기능주의는 행동주의와 본질적으로 다르지 않다.

15) 사실상, (주관적 관점을 묻는) 수정된 중국어방 논증은(객관적 관점에서 출발하는) 튜링의 사고실험을 뒤집은 것과 같다.

식을 갖고 있는지 혹은 자신의 사고와 이해를 의식/경험하고 있는지 물을 수 있다. 여기서 주체가 진정으로 이해한다고 할 수 있으려면 그러한 이해의 경험(주관적 의식과 느낌)을 필요로 한다는 주장이 가능해 진다. 이런 의미의 주관적-이해는 주체의 내면적 경험을 가리킨다는 점에서 감각질과 유사한 특성을 갖는다. 이것은 감각질과 마찬가지로 (네이글적인 의미의) 주관성의 문제, 혹은 "관점"의 문제이다. 기능주의의 치명적인 문제가 감각질에 관한 것이었다면,[16] 기능주의적 인공지능의 이해 개념이 갖는 문제는 주관적 의식의 문제인 셈이다.

그렇다면 써얼의 중국어 방 논증은, 기능주의(따라서 인공지능논제)의 핵심문제인 감각질의 문제를 이해와 같은 지향적 심리상태에 연장하여 적용한 것으로 간주될 수 있다. 즉 컴퓨터는 마음과 기능적으로 동일하지만 주관적 이해의 경험을 갖지 않을 수 있다는 것이다. 이것은 기능주의적 인공지능에 대하여, 이해와 같은 지향적 상태의 주관적인 경험을 문제삼는 것이다. 전도된/결여된 감각질의 문제가 기능주의의 주요한 한계로 간주되는 만큼, 기능주의와 인공지능의 (강한) 논제에 회의를 던진 것은 지향적 심성보다 오히려 감각질과 같은 현상적 심성영역인 것 처럼 생각되어 왔다. 그러나 이해와 같은 지향적 영역에서 ('감각질'에 버금하는) 주관적 의식의 문제가 제기된다면, 그러한 생각은 피상적이었다는 것을 알 수 있다.[17]

그리하여 우리는, 인공지능론자가 통상적으로 생각하듯이, 지향적 심성은 감각질 같은 현상적 심성보다 결코 기능주의적으로 처리하기 쉬운 대상이 아니라는 것을 확인하게 된다. 어쩌면 지향적 심성이야말로 환원될 수 없는 인간 사고의 중심적이고 고유한 위치를 갖는 것일른지 모른다. 그것은 마음에 대한 최대의 수수께끼에 뿌리를 둔 것이기 때문이다. 즉 '이해의 경험' 내지 '주관적-이해'는 인간 인지의 최고 단계라고 여겨지는 의식 및 자기의식, 혹은 자각의 문제의 또다른 측면인 것이다.

16) 기능주의에 대한 대표적인 반론은, 결여된 감각질 및 전도된 감각질 논증으로 잘 알려져 있다. 그것은, 두 체계가 기능적으로 동형적이면서도 그 중 하나는 감각질이 전도되거나 전혀 없을 수도 있다고 주장한다. 즉 기능적으로 동일하지만 마음의 질적인 특성이 다를 수 있다는 반론이다.

17) 물론 감각질과 '이해의 주관성'을 동일하게 다룰 수 없는 차이가 둘 사이에 존재한다. 예를 들어 고통과 같은 감각질의 경우 아픔의 느낌 없이는 그 감각질을 가질 수 없는 한편, 이해의 경우에는 그것의 의식작용 없이도 우리는 이해능력을 귀속할 수 있기 때문이다(6장 참고).

5. 이해의 주관적 관점

이미 살펴본 대로 강한인공지능논제에 대한 써얼의 논박을 보면, 튜링테스트를 통과하는 컴퓨터는 소위 '기능적-통사적-이해'에 도달할지라도, 그것은 두뇌의 인과력이라는 생물학적 기초 없이는 인과적-이해 및 '의식적인 주관적-이해'에 도달할 수 없다는 것이 그 요지이다.[18] 즉 마음은 계산적 속성들만으로는 설명할 수 없으며, 또한 기계의 기능적-통사적 개항들의 조작만으로는 어떻게 의미나 이해를 산출할 수 있는지 설명할 수 없다는 것이다. 그리하여 인공지능이 함축하는 튜링 방식의 이해 개념은 '진정한 이해능력'을 결여한다.

그런데 써얼은 튜링테스트를 통과하는 컴퓨터가 이해능력을 갖지 못하는 근거로서 주관적-이해의 결여를 지적하고 있음을 앞장에서 살펴보았다. 컴퓨터의 모방은 튜링테스트를 통과할지라도, 그것은 속임의 의도나 의식과 같은 주관적 경험/관점을 갖지 않는다는 것이다. 반면에 그는 두뇌 인과력에 근거한 인간의 이해가 주관적 의식을 동반한다는 것을 의심하지 않는듯 하다. 그는 두뇌 인과력을 갖는 인간의 경우, 자신이 이해하고 있다는 의식의 주관성이 존재한다고 본다. 즉 써얼에 의하면, 감각질이나 의식은 생물학적 과정들에서 일어나지만, 우리와 기능적으로 식별되지 않는 로봇/컴퓨터와 같은 전자적인 메카니즘 체계는 의식을 갖지 않는다고 주장한다. 그러나 인간 두뇌에 기초한 인과적-이해는 어떻게 주관적-이해에 도달할 수 있는가? 생물학적 두뇌가 이해의 주관적 경험(감각질 내지 주관적-이해)을 가능하게 하는 것인가? 인간은 두뇌 인과력 때문에 이해의 주관적 의식을 가질 수 있는가? 써얼은 인과적-이해가 얻어지면 주관적-이해를 성취할 수 있다고 믿어의심치 않지만, 과연 그러한지 그리고 어떻게 그럴 수 있는지는 해명이 필요하다. 만일 기계의 기능적-통사적 개항들의 조작만으로 의미나 이해를 산출하는 것이 신비스럽다면, 마찬가지로 물리적-두뇌 신경생리적 사건들의 구성만으로 (주관적) 이해를 산출한다는 것 역시 설명할 수 없는 미스테리이기 때문이다.

네이글이 지적하듯이, 기능주의나 물리주의의 객관적 관점은 주관적/의식적-이해를 설명할 수 없으며, 감각질이나 주관적-이해(의식)를 설명하기 위해서는 주관적 관점이 필요하

18) 써얼의 중국어 방 논증(1981)은, 마음의 컴퓨터 모델에 반대하여, 통사론만으로는 의미론을 낳을 수 없다는 지적이었다. 나아가 (1992)에서는, 물리적인 것만으로는 지향적 심리상태를 낳을 수 없다고 논의한다. 의미와 이해의 지향적 심성은 의식적인 마음에 의해 파악되어야만 한다. 즉 (본)질적인 intrinsic 지향성은 의식을 본질로 하는 마음 안에서만 일어난다는 것이다.

다. 감각질이나 주관적 의식은 기능적으로나 물리적으로 설명하기에 부적절해 보인다. 나의 어떤 고통은 내가 그것을 자각할 때에만 의식적인 고통이라는 것은 충분히 직합한 말이다. 그러면 내가 어떤 고통을 자각할 때 내가 자각하는 것은 무엇인가? 기능주의나 두뇌물리주의에 따르면, 고통은 인과적-기능적 역할이거나 어떤 두뇌상태이다. 즉 고통 상태에 있다는 것은 조직의 손상에 의해 야기되며 고통 행동들을 야기하는 성향의 상태에 있다는 것이다. 이 경우 내가 고통을 자각한다는 것은, 내가 그러한 기능적 상태에 있다는 것(혹은 어떤 두뇌상태에 있다는 것)을 자각한다는 것이 될 것이다. 그러나 이것은 분명히 잘못이다. 사실상 나는 나의 두뇌 상태나 기능상태에 대하여 아무것도 모를 수도 있다. 다만 내가 고통을 자각할 때 나는 바로 그 상처의 아픔을 자각하는 것이다. 이것은 그 상태에 있을 때 나의 관점에서 그것이 무엇인지 느끼게 되는 그런 의미의 주관적인 것이다.[19] 이것은 기능주의만이 아니라 두뇌 물리주의 역시 의식의 현상을 충분히 설명하기 어렵다는 것을 의미한다.

사정이 그러하다면, 기능주의-인공지능에 대한 써얼의 비판은 두뇌 물리주의에도 마찬가지로 제기될 수 있다. 즉 써얼이 이해의 주관적 의식을 말하면서도 의식에 대한 두뇌 물리주의적 해결에 머무른다면, 그의 기능주의/인공지능 비판은 자신의 입장에도 그대로 해당될 수 있다. 이 문제에 대하여 써얼은 비환원적 해결을 시도한다. 의식은 그것의 주관성에도 불구하고 두뇌의 물리적 속성(한 유기체의 생물학적 속성)이며, 그러나 그것은 임의의 다른 물리적 속성들로 환원할 수 없다는 것이 써얼의 입장의 핵심이다. 즉 "의식적인 심리현상들은 생물학적 과정들에의해 산출되지만, 그것들은 달리 환원할 수 없는 주관적인 것이다."[20] 인간은 두뇌 인과력을 가짐으로써 의식(혹은 주관적-이해)을 갖는다고 주장하는 한편, 그러나 주관적인 의식은 객관적인 물리-화학적 작용으로 환원/설명할 수 없다는 것이다. 이와같이 써얼은, 일종의 비환원적 물리주의—그는 자연주의라는 말을 더 선호하지만—노선에서, 의식의 환원불가능성을 주장함으로써 이원론적 해결의 방향으로 나아간다. 그러나 이러한 설명은 써얼 자신이 비판하고있는 이원론과 어떻게 구분되는가? 네이글은, 써얼

19) T.Nagel(1974), "What is it like to be a bat?" Philosophical Review 83, pp.435-50; J.Kim(1996), Philosophy of mind, (westview Pres), 제7장 참고.
의식적인 상태는 그 상태에 있을 때 그것이 무엇과 같은지를 느끼게 되는 어떤 것이 있다는 의미에서 현상적 특성을 갖는다.

20) J.Searle(1992), 제4장 참고, p.98.

의 해결방식이 속성이원론과 구별되지 않는다고 지적한다.[21]

네이글의 지적대로, 써얼의 비환원적 입장이 이원론적 특성을 지니고 있다면 그의 입장은 모종의 긴장을 가지고 있음에 틀림없다. 환원주의만이 아니라 이원론을 비판하는 그의 목소리는 자신을 향한 것일 수도 있다는 점에서 그러하다. 그러한 긴장에도 불구하고,[22] 그가 환원주의 및 기능주의와 데카르트 이원론을 지속적으로 함께 비판하는 근거에는 유익하고 흥미있는 생각이 발견된다. 그가 비판하는 입장들은 공통적으로 두뇌와 마음 사이의 관계를 비본질적인 것으로 만들며, 두뇌는 마음/의식을 위해 중요하지 않다는 생각에 근거하고 있다. 즉 데카르트 이원론은 마음을 두뇌로부터 분리된 실체로 보는 한편, 기능주의나 인공지능이론은 동일한 마음이 두뇌 이외의 다른 하드웨어(예를 들면, 전자체계)에서도 실현될 수 있다고 본다. 이에 반해 써얼은 의식과 두뇌의 인과력을 가장 강력한 방식으로 연결시키는 생물학적 관점을 취한다.[23] 달리 보면, 써얼의 논의과정에 나타나는 일관적인 흐름 중의 하나는 마음/의식과 두뇌를 분리하여 생각하려는 시도에 대한 비판이라고 할 수 있다.

21) Thomas Nagel(1995), Other Minds, Oxford University Press, p.96-110.
써얼에 의하면, 의식은 "존재론적으로 주관적"이다(의식이 "주관적"이라는 것은 존재론적 범주이며 인식론적 양태가 아니다. 예를들어, 나는 아프다는 진술은 관찰자의 태도나 의견과 상관없이 현실적인 사실의 존재에 의해 참이다. 그러나 고통/의식 그자체는 존재의 주관적 양상을 갖는다. 즉 고통은 제삼자에게 똑같이 접근될 수 없으며, 그것의 존재는 일인칭 존재이고, 누군가의(관점의) 고통이어야 한다. 모든 의식상태는 항상 [누군가의] 의식상태이다: 1992, p.94-95 참고). 즉 주체만이 자신의 고통을 경험할지라도, 그러한 주관적 관점은 실제 세계의 부분이다(네이글, p.104).그리하여 주관성과 같은 의식의 본질적 특징들은 전적으로 외부적인 삼인칭 관점으로는 기술될 수 없다. 즉 물리적 행동이나 기능적 구조 같은 사실만으로는 한 체계가 의식적이라는 것을 보일 수 없다. 삼인칭 관점은 --그것이 행동적이든 생리적이든— 의식을 기술하기에 불충분하다. 의식적 심성상태가 주체에게 어떤 것인지를 드러내기 위해 일인칭 관점은 필수불가결하다(네이글:101,104). 그렇다면 물리적으로 환원되지 않는 의식이 실제 세계에 존재한다는 이원론이 귀결된다.

22) 그의 입장이 얼마나 일관적인지에 관해서는 여기서 더이상 논의하지 않겠다. 다만 우리의 논의와 관련된 함축들을 조명하는 데 초점을 두겠다.

23)그러나 써얼에게 있어서, 의식의 근원이 되는 두뇌는 더이상 컴퓨터도 순수히 물리적인것도 아닌듯하다. "의식은 신경활동에 의해 야기된 두뇌의 질적인 주관적 속성이다"(네이글, p.106). 만일 두뇌가 단지 컴퓨터라면 그것은 질적인 지향성을 가질 수 없으며, 두뇌가 질적인 지향성이나 의식을 갖는다면 그것은 컴퓨터 이상의 것이어야만 한다. 써얼은 후자를 주장한다. 또한 두뇌는 순수 물리적인가, 아니면 물리적인 것 이상인가? 만일 전자라면 어떻게 순수 물리적인 것으로부터 의식의 주관적 속성이 산출되는지는 미스테리이며, 만일 후자라면 대체 두뇌의 정체가 무엇인지 설명되지 않은 채 남게된다.

그렇다면 적어도 마음과 두뇌 사이의 밀접한 연관성을 옹호한다는 점에서 그의 입장은 평가할만하며, 이것은 마음에 관한 이론에서 중요한 부분을 구성한다. (또한 마찬가지로 마음과 신체/행동 사이의 통일성을 배제하지 않는 사고가 요청된다. 이 부분은 뒤에서 다시 언급하게 될 것이다)

두뇌와 의식/마음의 관계에 대한 법칙들의 탐구는, 의식의 주관성을 설명하는 데는 도움이 안될지라도, 그것은 적어도 의식이나 심적 상태들이 두뇌와 어떤 상관관계에 있는지에 대한 제삼자의 관점에서 객관적인 과학적 설명을 제시한다. 그리고 의식의 객관적 관점은 의식의 주관적 관점과 무관한 것이 아니다. 그것들은 분리되지도 환원되지도 않는다. 즉 두 관점은 사물을 이해하는 두 국면으로서, 우리는 두 국면을 함께 고려함으로써 경험적 근거 위에서 심적개념들을 공유하게 된다. 이것이 두뇌/신체와 마음의 연관성을 배제할 수 없는 이유이다. 물론 상관법칙의 제시는 의식에 대한 일종의 과학적 해결이다. 그럼에도, 왜 그런 상관관계가 성립하는지 하는 철학적 문제는 여전히 하나의 수수께끼로 남게 된다. 의식과 두뇌 상태와의 상관관계가 충분히 밝혀진 후에도, 의식의 현상에 대한 신비는 사라지지 않는다는 것이다. 그런 이유로 의식에 대한 물리주의적 설명은 언제까지나 충족되지 못할지 모른다. 써얼은 자신의 (소위 자연주의적) 해결이 의식의 주관성 문제를—그리하여 주관적 이해의 현상을— 설명할 수 있다고 믿었지만, 우리의 고찰에 의하면, 의식의 주관성에 관한 한 그의 해결방식은 물리주의의 처지보다 나을 바가 없는 듯하다. 그런데 써얼과 같이 의식(혹은 이해)의 주관성을 강조하는 철학적 입장에 대해 제기되는 중요한 문제가 있다. 즉 언어 이해를 위해서 이해의 주관적 경험이 반드시 필요한 것인가? 이해의 주관적 경험(주관적-이해)이나 이해의 현상적 특질(즉 느끼거나 경험하거나 감각된 특질)에 대하여 말하는 것이나 그것에 근거하여 어떤 주장을 하는 것이 과연 정당한가? 다시말해서, 감각질과 의식과 같은 현상적 특질이 과연 실제로 있는가? 그것은 허구가 아닌가? 혹은 데카르트적인 사고방식에 불과한 것이 아닌가?

6. 현상적 특질의 문제: 인공지능과 인격공동체

그러면 왜 마음의 내적인 현상적 특질이 문제인가? 그것이 사고나 이해 주체가 될 수 있는 조건이거나 기준인가? 인공지능의 이해능력에 대한 비판적 논의를 보면, 컴퓨터는 인간과 달리 주관적-이해를 가질 수 없으므로 진정으로 언어를 이해한다고 볼 수 없다는 것이다. 이러한 논의는, 〈이해능력은 주관적-이해 내지 의식작용을 수반한다〉는 것을 전제한다. 그러

나 언어 이해는 과연 주관적 의식을 필요로 하는가? 이해의 주관적 관점에 대한 비판들은 많은 철학자들에 의해 제기되어 왔다. 특히 비트겐슈타인은, 이해는 의식작용을 함축하지 않는다는 것을 설득력있게 논의한다.[24]

논의의 요지는, 어떤 사람이 x를 의미하거나 어떤 명제를 이해한다는 것은 그가 의미나 이해와 같은 어떤 의식작용이나 심적 작용을 갖는다는 것을 반드시 함축하진 않는다는 것이다. 즉 무엇을 이해한다는 것은 그 주체 자신이 이해하고 있다는 것을 자각하는 의식작용을 함축하지 않는다는 것이다. 그러한 자각이나 의식 없이도 우리는 많은 것들을 이해한다. 이것은 믿음이나 욕구와 같은 지향적 심적 상태의 경우도 마찬가지이다. 우리가 무엇을 이해하거나 어떤 믿음을 가질 때마다, 어떤 특별한 심상이나 마음 안에 어떤 것이 일어남을 느끼는 것은 아니다. 어떤 이들(아마도 데카르트주의자들)은 우리가 어떤 사고를 하나의 믿음으로 갖게 될 때마다 어떤 긍정적인 느낌, 즉 "아, 그렇다!"라는 종류의 느낌이 있다고 주장해 왔다. 유사하게 어떤 것을 불신하게 될 때 직접적으로 경험되는 부정의 느낌을 동반하며, 그리고 기억은 언젠가 본듯한 어떤 느낌을 동반하고, 아마도 욕구나 바람들은 현재의 결핍감과 결합된 갈망이나 동경의 느낌을 동반한다고 말한다. 그러나 이러한 지향적 상태들마다 어떤 특별한 종류의 현상적 특질이 동반된다고 보기는 어렵다. 예를 들어, 만일 당신이 안락사가 도덕적으로 허용될 수 있다는 생각(혹은 오늘 논문발표가 있다는 믿음, 이번 선거에서 K가 승리할 것이라는 판단...등)에 동반되는 감각질 Q를 발견할 때 "아! 이제 나는 안락사가 도덕적으로 허용될 수 있다고 믿는다는 것을 안다!"라고 말하고, 또한 당신이 Q를 발견하지 못한다면 "오, 아니다! 나는 그 믿음을 갖지 않는다!"라고 말하는 근거가 될 수 있는 그런 감각질 Q가 있다고 주장하는 것은 터무니 없는 것 같다. 즉 우리는 특수한 감각질을 찾아서 자신의 내부를 관찰함으로써 우리가 믿는지 아니면 희망하는지를 발견하는 것은 아니라는 것이다. 또한 우리는 매번의 믿음이나 욕구마다 그리고 무엇을 이해할 때마다 특별한 유형의 감각질을 찾으려고 하지 않는다. 아뭏든, 어떤 특수한 질적/현상적 특성을 갖지 않는 지향적 심리 상태들이 있다는 것은 분명하다. 많은 경우 질적인 경험없이도 이

24) Wittgenstein, L., Philosophical Investigation, #138-200 참고.

해는 가능할 수 있다.[25]

그러나 이러한 고찰은, 주관적 의식(혹은 주관적-이해)을 갖지 않는 컴퓨터나 인공지능도 지향적 태도를 가질 수 있고 언어를 이해할 수 있다는 것을 지지해 주는가? 물론 그것은 이해가 반드시 이해의 의식작용을 함축하지 않는다는 것을 말해준다. 그러나 그것이, 의식이 결여된 기계나 인공지능에다 이해의 개념을 적용할 수 있다는 것을 보이기엔 충분하지 않다. 때로는 어떤 사람이 이해의 의식작용을 갖지 않을 경우에도, 그가 그것에 대해 전혀 의식하지 않을 때에도 우리는 그 사람에게 이해력을 귀속시킨다. 그러나 이 사실이 〈어떤 상황에서도 의식작용을 갖지 않는, 즉 의식이 없는 종류의 실체(의식이나 감각질을 갖지 않는 그런 종류의 대상)에게 이해를 귀속시킬 수 있다는 것〉을 함축하지 않는다.[26] 그렇다면 의식의 발생이 없는 경우에도 인간에게 이해나 사고 등의 심적 상태를 귀속할 수 있다는 논의는, 어떤 종류의 의식도 없는 실체들(예를들면, 기계)에게 이해를 귀속시킬 수 있다는 결론을 함축하지 않는다. 즉 인공지능은 주관적 의식을 경험하지 않으나 이해력을 가질 수 있다는 것, 그리하여 의식이 없는 인공지능도 이해력이 있다는 것을 지지하지 않는다.

이러한 관찰이 옳다면, 감각질 및 의식의 주관성(주관적-이해)은 강한인공지능논제에 저항하는 중심적 요인으로 작용한다. 이미 살펴보았듯이 그것은 기능주의적으로 해소되지도 않고 다른 것으로 환원되지도 않는다. 감각질 혹은 주관적 의식이 외래적(extrinsic)이지 않고 질적이고 본래적인(intrinsic) 속성들이라면, 어떻게 그것들이 그밖의 다른 것(물리적 상태, 혹은 기능상태...등)과 환원적으로 동일시 될 수 있는지를 보이기가 어렵다. 비록 의식 및 이해의 주관적 체험과 같은 현상적 속성과 신경속성 사이에서 경험적 상관관계가 발견될지라도, 후자는 전자를 설명할 수 없다. 우리는 여전히 어떻게 의식이 두뇌/물질에서 일어나는지를 이해할 수 없기 때문이다(설사 사실이 그러하다고 할지라도). 혹자는 물리적으로 설명되지 않는 그러한 현상은 사실상 존재하지 않는 것이라는 극단으로 나가기도 한다. 그들은 감각질 제거주의자들이다. 이와같이, 마음의 현상적 특질을 설명하기 위해, 그것들을 비본래적-관계적 속성들로 다루려는 시도로부터 그것들의 존재를 부정하는 시도에 이르기 까지 다양한 입장들이 있어왔다. 그들

25) J.Kim(1996), 제7장 참고.

26) Dieter Brinbacher(1995), " Artificial Consciousness", ed.Thomas Metzinger, Conscious Experience, (Schoningh/Imprint Academic), pp.489-503.

은 표면적으로는 입장의 다양성에도 불구하고, (이원론을 피하고자) 결과적으로 마음의 현상적 특질을 부정하면서, 마음이 아예 존재하지 않거나 텅빈 것이라고 보는 점에서 공통적이다. 행동주의는 마음을 하나의 블랙박스로 보는 한편, 기능주의는 마음 안의 무엇을 인정하지만 그것을 열어보였을 때 수많은 작은 블랙박스를 보여줄 뿐이라는 점에서 행동주의와 별반 차이가 없다.

이러한 비판적 지적은 다시금 이원론으로 돌아가야 한다든지, 데카르트의 망령을 불러들여야 한다고 말하는 것은 아니다.[27] 다만 우리 인간의 마음이 본래적으로 텅빈 것이거나 허구가 아니라면, 그러나 객관적으로 관찰가능한 것도 아니고 또한 자연주의적 설명의 대상도 아니라면(퍼트남), 그것에 대해 어떤 종류의 물음을 묻고 어떤 태도로 접근하는 것이 타당한지 고려할 필요가 있다. 물론 언어적 개념들을 배우기위해서 사적이고 주관적인 감각질은 필수적이지 않다고 지적할 수 있다.[28] 언어를 배우고 언어 규칙을 따르는 것은 공적인 문제이기 때문이다. 그런점에서 개념형성에서 감각질은 본질적 요소가 아니라는 감각질의 개념적 무용론(無用論)을 받아들일 수는 있다. 그러나 이것은 감각질의 존재를 부정하지 않기 때문에 감각질 제거주의와는 다르다.

우리들은 각자 현상적/질적인 경험내용을 가지며, 그리하여 우리의 마음은 단순히 텅빈 형식(혹은 블랙박스)이 아니라는 것을 확신한다. 적어도 이것이 우리가 인간들을 대우하는 방식이며, 서로를 인격으로 간주하는 토대이다. 즉 인격공동체로서 우리 인간은 서로 간에 마음의 존재에 대하여 회의하지 않는다는 것이다. 반면에 인공적 기계(컴퓨터)에 대해서는, 그것들이 계산을 하고 문제를 풀고 문장을 발화하는 등의 지능을 가졌다고 할지라도, 우리는 그들의 마음의 존재에 대해서 회의한다. 컴퓨터는 우리 인간의 관점과 전혀 다르거나 아니면 전혀 관점을 갖지 않을 수도 있다. 즉 인간은 소위 '타인의 마음 문제'가 제기되지 않는 존재인 반면에, 인공지능은 (텅빈 마음일 수 있다는 점에서) '타인의 마음 문제'가 제기

27) 이원론자들은, 〈물리적으로 설명되지 않는 의식의 현상〉을 물리주의가 거짓임을 보여주는 증거로 사용한다. 그러나, 여기서 논의할 수는 없으나, 이원론 역시 설명하기 어려운 그 이상의 또 다른 문제들을 갖고 있다.

28) 이 입장은 비트겐슈타인의 철학적 관점을 따르고 있는 말콤에 의해 전개되고 있다. N.Malcolm(1971), Problems of mind, George Alen & Unwin.

되는 회의의 대상이다. 이것은 우리가 인공지능체계를 우리와 같은 인격으로 대하지 않는다는 것을 말해준다. 물론 이런 생각은 인간의 관점을 반영하는 것에 불과할 수 있다. 우리의 관점에서 기계가 의식을 갖지 않는다고 확언하는 것은 인간의 독단이 아닌가? 아니면 그렇게 확신할만한 타당한 근거가 있는가?

일반적으로 인공적인 의식이 가능한가, 불가능한가 하는 것은 논의의 대상이다. 만일 기계의 이해와 인간의 이해 사이의 본질적인 차이가 주관적 의식에 있다면, 진정 기계/인공지능은 의식을 가질 수 없는가? 인공-의식이 불가능하다는 논의는 주로 세가지 차원에서 제시되고 있다.[29] 먼저 기술적인 수준에서 의식을 갖는 기계를 인공적으로 만들 수 없다는 실천적 불가능성과, 인공 의식을 개념적 모순으로 보는 개념적 불가능성, 그리고 규범적 차원에서 인공의식의 불가능성을 주장하는 것이 그것이다. 여기서 우리가 주목할만한 것은 개념적 불가능성 논의이다. 비트겐슈타인에 의하면, "오직 살아있는 인간과 같은 것만이 의식을 갖는다고 말할 수 있다"[30] 그렇다면 의식적인 의자나 의식적인 식탁이라는 것이 넌센스이듯이 의식적인 기계/컴퓨터, 즉 인공-의식이라는 것도 허구이다. 물론 이 주장은 기계가 말하고 생각하는 것을 상상할 수 없다는 의미의 개념적 불가능성을 뜻하진 않는다. 왜냐하면 우리는 기계는 물론 냄비나 수저가 말하고 보고 듣는 것을 얼마든지 상상할 수 있기 때문이다. 오히려 그 주장은 기계가, 인간의 조건과 달리 경험적 기준을 결여하기 때문에, 규칙을 따르거나 언어를 배우는 것이 불가능하다는 것을 의미할 것이다.

그러면 언어를 갖는 인간의 조건은 무엇인가? 그것은 마음(의식)과 신체와 두뇌가 통일된 유기체로서의 인간 조건이며, 그러한 통일체로서의 인간으로부터 나오는 행동과 삶의 양식이 언어를 가능하게 하는 경험적 기준을 마련해 준다. 반면에 컴퓨터는 두뇌는 물론 인간과 같은 신체와 신체 행동을 결여하므로, 언어(규칙)를 이해할 수 있는 경험적 근거를 상실한다. 그리하여 우리는 기계를 우리와 같이 의식과 언어를 가진 존재로 간주하지 않는다. 이것이 우리가 인공지능을 인격공동체의 일원으로 간주하지 않는 적어도 부분적인 이유가 된다. 그렇다면 주관적 의식이 인격의 징표로

29) Dieter Brinbacher(1995), 앞글

30) Wittgenstein, PI #281.

작용한다는 주장은, 사실상 의식과 신체와 두뇌의 통일체로서의 인간을 전제할 때 비로소 성립한다. 즉 인격의 개념은, 단지 의식만으로는 불충분하고 신체와의 통일성을 요구하며 또한 다른 인격들의 공동체를 전제로 하는 개념이다.[31] 그리고 '동일한 인격'과 '책임주체' 등과 같이 인격과 관련된 중요한 개념들은, 신체를 통하여 비로소 합당한 근거를 갖게 된다. 우리는 이러한 인격의 개념들을 공유하기 때문에 자신의 존재를 확신하는 만큼 다른 인격들의 존재를 확신하지만, 〈적어도 신체와 의식의 통일체로서의 우리의 삶의 모습을 공유하지 않는한,〉 인공지능이나 기계를 인격으로 간주하지 않을 것이다. 비록 인공지능을 인격 공동체의 일원으로 포함하는 그런 세계를 '상상할 수'는 있을지라도, 그 세계는 인격에 대한 전혀 다른 개념체계를 요구할 것이며[32] 그리하여 그들의 삶의 모습은 현재 우리와는 아주 다른 것이 될 것이다.

31) 최근의 나의 논문, "인격이란 무엇인가" 참고.

32) 강한 인공지능논제가 받아들이는 기능적 계산주의에 의하면, 마음은 계산적 체계이며 이해와 사고를 포함하여 인간 마음의 모든 활동들은 일종의 계산/연산 과정이다. 이 입장에 의하면 인격동일성 개념은 '과정 동일성'이 된다. 즉 그들의 개별화 원리는 더이상 신체가 아니라 디스크에 저장된 계산과정에 근거한다. 그러한 인격개념은 변화하는 중의 동일성을 설명할 수 없다. 과정동일성에 의하면, 시간을 통하여 성격이 변화하는 경우 동일한 인격이 아니라 매순간 다른 인격들이 계속 탄생되는 것으로 간주되기 때문이다. 또한 이것은 행위의 책임주체를 상실하는 결과에 도달한다. 과거에 행위했던 주체는 이제 더이상 동일한 인격이 아니라 다른 인격이기 때문이다. 아뭏든 인공지능의 인격개념은 심신 통일체로서의 우리들의 인격 개념과 현저하게 다른 것이 분명하다.