계산가능성과 전산적 마음

이 영 의*

목 차

1. 서론

계산이론은 철학이나 인지과학 분야에서 중요한 이론적 토대를 제공한다. 특히 계산이론은 인지과학 분야에서 인간의 마음을 정보처리체계로 간주하는 파라다임을 제공 해왔다. 이러한 정보처리 파라다임에 따르면 인간의 마음은 뇌의 프로그램이며, 인지과정은 컴퓨터와 마찬가지로 표상을 계산하는 과정이다. 정보처리 파라다임은 기본적으로는 마음을 기계로 간주하는 일종의 인간기계론에 해당된다. 그런데 인지과학이 기본적인 전제로 삼고 있는 인간기계론을 괴델정리에 의거하여 논박하려는 시도들이 있어왔다. 이러한 시도들은 괴델정리에서 나타나는 형식체계의 불완전성과 자기지시성에 주목하여, 기계는 불완전하며 자기지시성이 없는데 반해 인간은 자의식을 갖고 있기 때문에 인간의 마음을 기계로 간주할 수 없다고 주장한다.

* 고려대학교 강사
이 논문은 다음의 두 글을 계산가능성을 중심으로 연결한 것이다. 이영의(1990), "튜링기계와 기능주의". 한국인지과학회 추계학술대회논문집, pp. 154~163. 이초식, 이영의(1992), "괴델정리와 인공지능". 한국인지과학회 춘계학술대회논문집, pp. 3~13.

이 글은 두 부분으로 구성되어 있다. 전반부에서 정보처리 파라다임과 관련된 몇가지 계산이론들, 즉 힐버트 프로그램과 괴델정리, 튜링기계 개념들이 제시된다. 특히 계산의 의미를 살펴보기 위해 튜링기계가 작동하는 방식이 구체적으로 분석된다. 후반부에서는 먼저 괴델정리를 이용하여 인간기계론을 비판하려는 루카스(Lucas)와 펜로제(Penrose)의 시도들이 제시된다. 이 글의 목적은 그러한 시도들은 인간기계론에 대한 적절한 비판이 될 수 없다는 것을 보이는 것이다. 루카스와 펜로제의 시도가 발표된 이후 많은 지지와 비판의 논의들이 제기되어 왔다. 필자는 이러한 논의들을 몇가지 유형으로 구분한 다음 그 중 가장 유력해보이는 입장들을 중심으로 그들의 시도가 성공될 수 없음을 보일 것이다.

2. 계산가능성

사람들은 일반적으로 수학적 체계는 엄밀하다는 믿음을 갖고 있다. 이러한 믿음의 근원을 살펴보면 모든 수학적 문제는 정확하게 기술될 수만 있다면 아무리 복잡하고 어려운 문제라고 하더라도 결국은 그 해답을 발견할 수 있다는 가정에 근거를 두고 있다. 힐버트(Hilbert)는 이러한 직관적으로 타당한 것처럼 보이는 신념을 수학적으로 증명하기 위해서 다음과 같은 프로그램을 제안했다. 우선 모든 수학적 문제들이 참이나 또는 거짓인 문장으로 구성될 수 있는 형식체계 S를 구성하고, 체계 S에 속하는 특정한 문장의 진위를 결정하는 효율적인 절차(effective procedure)를 발견하는 것이다. 여기에서 효율적인 절차는 형식적 규칙들의 집합에 해당할 것이다. 힐버트의 프로그램은 다음과 같은 신념이 정당화될 수 있는가를 묻는 것이다.

(1) 형식체계 S에 속하는 임의의 문장의 진위를 결정할 수 있는 효율적인 절차를 발견할 수 있다.

힐버트의 프로그램은 일차 술어논리체계에 속하는 임의의 정식의 진위를 실제로 결정하는 문제, 즉 결정가능성(decidability)의 문제를 제기한다. 그가 제안한 프로그램이 성공한다면 우리는 모든 수학적 문제를 기계적인 계산으로 환원시킬 수 있다. 그러나 괴델(Gödel)은 힐버트의 야심적인 목표가 달성될 수 없음을 결정적으로 보여주는 불완전성정리(incompleteness theorem)를 발표했다. 불완전성정리가 적용되는 형식체계 S는 자연수의 산술을 포함할 정도로 충분히 넓은 의미를 지닌 체계이다. 체계 S는 화이트헤드와 러셀(Whitehead and Russell)의 「수학원리」(Principia Mathematica)의 체계와 같이 몇개의 추론규칙만을 이용하여 모든 정리들을 증명할 수 있는 강력한 체계이다. 괴델은 이러한 체계 S에 대해 다음의 두가지 정리가 성립함을 증명했다.[1)]

(2) 형식체계 S는 불완전하다. 즉 S는 문장 A나 ¬A가 모두 증명불가능한 A를 포함한다.

(3) S에 대한 일관성 증명은 불가능하다. 따라서 S에 대한 일관성 증명은 단지 S에서 형식화될 수 없는 추론방식에 의해서만 수행될 수 있을 뿐이다.

괴델은 정리 (1)과 (2)를 증명하기 위해 먼저 산술화(arithmetization)방법과 회귀함수(recursion function)를 이용하여, 일상어

1) K. Gödel(1931), pp. 592~617.

로 표현하면 "이 문장은 증명불가능하다"를 의미하는 다음과 같은 괴델문장 G를 S에서 구성했다.

(4) (x) ¬DEM(x, G(N))

위에서 DEM은 증명가능성을 나타내는 술어이고, G(N)은 문장 G의 괴델수이다.[2] 괴델문장 G는 자기지시적이다. 즉 그 자체에 대해 말하고 있다. 괴델정리는 이러한 자기지시적 문장 G를 이용하여 증명된다. (4)에서 G 자체가 증명가능하면, 이는 G가 증명불가능함을 의미하고, 역으로 G가 증명불가능하면 바로 그 사실이 G의 증명불가능함을 보여준다. 따라서 어느 경우든 G는 증명불가능하다(또는 결정불가능하다). 그런데 배중률을 적용하면 산술적 주장은 증명가능하거나 불가능한 경우 어느 하나에 속하므로, 문장 G가 속해있는 S가 무모순적이라면, 즉 G와 G의 부정을 둘다 포함하고 있지 않다면, S는 G가 참임을 결정할 수 없으므로 불완전하다. 그럼에도 불구하고 G는 S가 허용하는 것보다 훨씬 더 직관적인 추리에 의해 확립될 수 있는 자연수에 대한 명제이므로 참이다. 따라서 괴델정리에 의하면 참이지만 그 체계에서 증명할 수 없는 정리가 있다.

불완전성정리에 의하면 수론에서 문장을 구성할 정도로 강력한 임의의 무모순적인 형식체계는 증명될 수 없는 참인 문장을 포함하고 있으므로, 힐버트가 기대한 그러한 효율적인 절차는 있을 수 없다는 결론이 도출된다. 이처럼 힐버트 프로그램이 괴델에 의해 결정적으로 폐기되자 수학기초론의 촛점은 진리 개념에서 증명가능성(provability) 개념으로 이동했다. 수학자들은 이

2) 괴델문장과 괴델수에 관한 설명은 튜링기계에 관한 논의에서 다루어진다.

제 수학에서 증명가능한 모든 문장이 공리 집합으로부터 계산될 수 있는 방법에 관심을 두게 되었다.

튜링(Turing)은 힐버트의 문제와 계산가능한 함수 간의 관련성을 이해하고, 직접적이고 확실한 방식으로 연구를 진행했다.[3] 튜링은 직관적으로 효율적 절차는 알고리즘이라고 생각하고, 이러한 직관적인 생각을 어떠한 계산 절차도 원자적 관계로 분해될 수 있는 모델로 정의될 수 있음을 보여 주었다. 튜링이 보기에 이러한 계산 절차는 기계가 구체적인 작동을 수행하는 방식을 정확이 규정하는 형식적 규칙들의 집합이었다. 이처럼 튜링은 처음으로 계산(computation) 개념에 대해 수학적으로 정교한 이론을 제시했으며, 이러한 이론적 발견에 있어 중요한 역할을 하는 것이 바로 추상적인 컴퓨터인 튜링기계이다.

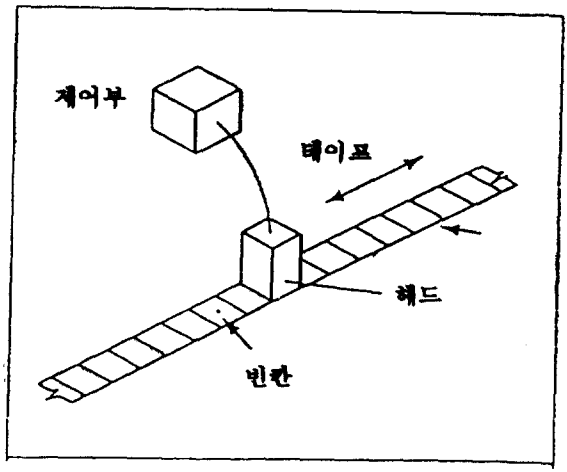
튜링에 따르면 효율적 절차에 대한 형식적 모델로서 기계는 다음과 같은 요구를 충족해야 한다.

(5) 각 과정은 유한하게 기술되어야 한다.

(6) 과정은 불연속적인 단계로 구성되고, 동시에 각 단계는 기계적으로 수행되어야 한다.

---

3) 이 문제와 관련하여 처취(Church)는 튜링과는 독립적으로 람다 계산체계(lambda calculus)를 개발하고, 수학 함수가 계산 가능하면 그것은 람다 계산체계로 정의될 수 있다고 주장했다. 그는 또한 람다 체계에서 표현가능한 함수가 계산 불가능하면, 그것의 증명가능성 여부를 결정할 수 있는 방법이 없음을 증명했다.



[그림 1] 튜링기계

이러한 요구 사항을 충족시키는 간단한 기계를 살펴보기로 한다. 튜링이 제시한 체계는 [그림 1]에 나타난 것처럼 세 부분으로 구성된다. 즉, 유한한 제어부(finite control box)와 칸(cell)으로 나누어진 입력 테이프, 그리고 특정 시각에 테이프의 각 칸의 내용을 읽고 쓸 수 있는 헤드가 있다.

테이프는 오른편과 왼편 양 방향으로 계속 연장될 수 있다고 가정하면 기계의 기억용량(memory)은 무한하게 된다. 또한 테이프의 각 칸은 유한한 테이프의 기호 중에서 단 하나만을 지닐 수 있다. 맨처음 유한한 $n \geq 0$ 에 대해 왼쪽으로부터 n번째 칸은 입력기호라고 불리는 테이프기호의 부분집합으로부터 선택된 기호열이 입력된다. 나머지 무한한 칸들은 각각 공란(blank)을 갖는데, 공란은 입력기호가 아니라 특별한 테이프기호라고 간주된다.

튜링기계의 작동은 헤드가 읽어들이는 입력기호와 제어부의 상태에 의해 결정되는데, 구체적인 결정 과정은 다음과 같다.

(7) a. 상태를 변경한다.
b. 검색된 테이프 세포에 새로운 기호를 프린트하고, 그 다음 거기에 쓰인 기호를 지운다.
c. 헤드를 오른쪽이나 왼쪽으로 움직인다.

튜링기계의 작동은 기계작동표(machine table)에 의해서 규정된다. 기계작동표는 기계의 다음번 작동를 알려주는 지침에 해당하는데, 일반적으로 순서쌍으로 표현된다. 예를 들어 (도표 1)에 제시된 입력기호와 (도표 2)의 기계작동표가 주어졌다고 하자. (논의의 편의상 테이프 기호는 4개만 제시되었다).

(도표 1). 입력기호

단계 ①	1	1	0	1
단계 ②	0	1	0	1
단계 ③	0	1	0	1
단계 ④	0	1	1	1

(도표 2). 기계작동표

상 태	테이프기호	
	0	1
S_1	0, S_1, R	0, S_2, R ①
S_2	1, S_3, R ③	1, S_2, R ②
S_3	멈춤	멈춤 ④

위의 도표에서 단계 ①의 상태는 제어부가 상태 S_1에 있고 헤드는 테이프에 쓰여 있는 기호 1을 읽게 된다. (도표 2)의 기계

작동표에 의해 상태 S_1과 기호 1이 만나는 곳에 있는 작동지침, 즉 <O, S_2, R>에 따라 다음번 동작이 이루어 진다. 먼저 테이프의 기호 1을 0으로 대체하고 상태 S_2로 변경하면서 헤드는 오른편으로 이동한다. 단계 ②는 기호 1과 상태 S_2의 조합이므로 해당되는 지침은 <1, S_2, R>이다. 헤드는 기호 1을 1로 대체하고, 즉 그대로 두고 상태 S_2로 변경되면서 오른편으로 이동한다. 단계 ③에서 상태 S_2와 기호 0이 결합되므로 해당 지침은 <1, S_3, R>이다. 헤드는 기호 0을 기호 1로 대체하고 S_3의 상태에서 오른쪽으로 이동한다. 마지막으로 단계 ④에서 상태 S_3은 기호 1과 결합되므로 지침 <멈춤>에 따라 동작을 멈춘다.

이제 튜링기계는 다음과 같이 형식적으로 정의될 수 있다. 4)

(8) Turing machine = { Q, Σ, Γ, δ, q_0, B, F }

이처럼 형식적으로 정의된 튜링기계는 일종의 자동형식체계(automatic formal system)이다. 여기서 형식체계라는 것은 불연속적인 기호들이 유한한 규칙에 따라서 처리됨을 의미한다. 또한 자동체계라는 것은 기호의 연산을 지배하는 규칙들이 그 체계 내에 구체화되어 있으며, 외적 요인에 의해 연속적으로 조종될 필요가 없는 체계를 의미한다. 자동형식체계는 해당 요소들의 형식 또는 순서에 관한 사항만을 언급하고 있는 추상적인 기호체계에 불과하며, 해당 요소들의 의미에 대한 언급이 없다.

(8)에서 정의된 해석되지 않은 체계에 대해 다음과 같은 의미가 부여된다. 즉 Q는 상태들의 유한집합, Γ는 가능한 테이프기호들의 유한집합, B는 Γ의 기호인 공란, Σ는 B를 포함하지 않는 Γ의 부분집합인 입력기호를 의미한다. δ는 Q × Γ 로부터 Q × Γ × { L , R }에로의 대응을 표현하는 이동함수이다. q_0는

4) J. E. Hopcroft and J. D. Ullman (1979), p. 148.

Q에서의 출발 상태이고, F ⊆ Q는 최종 상태의 집합이다. 튜링기계의 작동에 대한 순간기술은 $<\alpha_1, q, \alpha_2>$로 표현된다. 여기서 q는 기계의 현상태이며, α_1과 α_2는 Γ에서의 기호열이다. (8)과 같은 해석되지 않은 형식체계에 대해 해석이 주어지면, 즉 기호들이 체계 밖의 사물들을 지칭하면, 튜링기계는 해석된 자동형식체계가 된다.

튜링기계가 지니는 장점 중의 하나는 다른 튜링기계의 작동을 모의할 수 있다는 것이다. 예를 들어 튜링기계 TM_2가 튜링기계 TM_1을 모의한다고 해보자. 모의 방법은 (도표 1)과 (도표 2)에서 제시된 것과 같은 TM_1의 작동과 관련된 사항을 부호화해서 이를 TM_2의 입력으로 이용하는 것이다. TM_1의 작동과 관련된 기술을 부호화하는데는 괴델이 불완전성정리에서 사용한 공리화 방법을 채택할 수 있다. 즉 TM_1의 모든 작동에 일정한 방법에 따라 하나의 자연수를 대응시키고, 그결과 나타난 자연수 계열을 다시 단 하나의 자연수로 나타낸다.

괴델이 사용한 공리화 방법을 사용하기 위해서는 다음과 같은 예비적 단계가 요구된다. 먼저 TM_1의 개개의 작동을 0이 아닌 자연수 a, b, c, ...n_k로 나타내고, 이러한 자연수로 구성된 계열 S = a, b, c, ...n_k를 구성한다. 둘째 계열 S로부터 $N = (2^a)(3^b)(5^c).....(p_k{}^{nk})$ 을 구성한다. 여기서 2, 3, 5, ...p는 모두 소수이고 p_k는 k 번째 소수를 나타낸다. 그 결과 나타난 N은 괴델수(Gödel's number)라고 하는데, 거듭제곱된 소수들의 곱으로부터 만들어지므로 유일한 자연수이다. 그러므로 S와 N은 일대일로 대응되며, 역으로 N으로부터 S를 다시 구성할 수도 있다.

이제 괴델수를 이용하여 (도표2)에 주어진 TM_1의 작동 지침을 부호화해보자. TM_1의 기계작동표에서 일차적으로 모든 사항을 자연수로 대체하고, 0은 다른 자연수로 대체된다. 즉 0이 나

타나면 1을 더한다. "오른쪽", " 왼쪽", " 멈춤" 등과 같은 표현은 자연수 1, 2, 3으로 각각 표현되고, 나머지 부호 역시 자연수로 표현된다. (도표 2)에 이러한 방법을 적용하면 다음의 괴델수을 얻을 수 있다.

(도표 3). 괴델수로 표현된 기계작동표

단계 ①	11012	22123	$2^2\ 3^2\ 5^1\ 7^2\ 9^3$
단계 ②	21112	32223	$2^3\ 3^2\ 5^2\ 7^2\ 9^3$
단계 ③	20113	31224	$2^3\ 3^1\ 5^2\ 7^2\ 9^4$
단계 ④	31010	32121	$2^3\ 3^2\ 5^1\ 7^2\ 9^1$

다시 각 단계의 괴델수를 A, B, C, D라고 하면, TM_1의 작동표는 괴델수 $N = 2^a \times 3^b \times 5^c \times 7^d$로 부호화된다. 이제 최종적으로 TM_2에 괴델수 N을 입력하면 TM_2는 TM_1과 동일한 방식으로 작동할 것이다. 이처럼 모든 TM_1의 괴델수를 읽고 그것과 동일한 방식으로 작동을 할 수 있는 TM_2를 보편 튜링기계(universal Turing machine)라고 한다. 튜링은 모든 다른 튜링기계를 모의할 수 있는 보편 튜링기계가 있다는 것을 수학적으로 증명했다.

(9) 모든 다른 튜링기계를 모의할 수 있는 보편 튜링기계가 있다.

본문에서 제시된 튜링기계는 세가지 상태만을 갖는 매우 단순한 것이다. 그결과 모든 튜링기계를 모의할 수 없다는 의미에서 보편 튜링기계가 아니다. 여기에서 보편 튜링머신이 되기 위한 최소한의 기호와 상태에 대한 의문이 발생하는데, 민스키(Minsky)는 4개의 기호와 7가지 상태를 지닌 튜링기계가 가장 단순한 보편 튜링기계임을 증명했다.

튜링기계는 기계적이고 효율적 절차를 계산가능성과 관련시켜 분명하게 제시해 주는 체계라고 인정되고 있다. 그러나 튜링기계 모델이 계산에 대한 우리의 직관과 일치한다는 것을 증명할 수는 없다. 튜링기계는 엄밀하게 말하면 자동 형식체계에 대한 하나의 제안에 불과하다. 그러나 힐버트의 프로그램에 대해 연구를 했던 다른 사람들, 예를 들어 처치(Church)와 포스트(Post) 등이 효율적인 절차에 대해 제시한 정의가 튜링의 생각과 정확히 일치한다는 것이 판명되었고 그 결과로서 다음과 같은 처치-튜링 논제 (Church-Turing thesis)가 나타난다.

(10) 모든 결정론적인 자동 형식체계에 대해 형식적으로 동일한 하나의 튜링기계가 있다.

튜링-처취 논제에 따르면 효율적인 절차에 해당하는 어떠한 절차도 튜링기계에 의해 수행될 수 있다. 또한 어떠한 자동 형식체계도 튜링기계가 할 수 없는 작동을 할 수 없다는 결론이 유도된다. 튜링-처취 논제는 또한 우리는 직관적으로 알고리즘에 대해 만족할 만한 이해를 하고 있다고 주장한다. 누구나 주어진 임무를 효율적으로 수행하는 것이 무엇을 의미하는가를 직관적으로 알고 있다는 것이다. 즉 우리는 특정한 임무를 수행하는데 있어 절차들의 집합을 고안하고 그러한 절차들의 지침을 기계에 부여하는 것을 직관적으로 이해하고 있다. 튜링-처취 논제는 그러한 지침들의 집합이 알고리즘이며, 전산과학에서 사용되는 알고리즘에 대한 정의는 바로 이러한 비형식적인 생각과 일치한다는 것을 보증한다. 따라서 튜링-처취 논제는 (10)과 달리 다음과 같이 표현될 수도 있다.

(11) 인간이 계산할 수 있는 것은 기계도 계산할 수 있다.

(9)와 (10)은 상호 보완적이다. (9)에 의하면 효율적인 계산을 하기 위해서는 단지 하나의 튜링기계, 즉 보편 튜링기계만 있으면 된다. (10)에 의하면 튜링기계는 계산에 대한 우리의 직관과 일치하는 자동 형식체계이다. 따라서 (9)와 (10)을 결합하면 보편 튜링기계는 우리의 직관과 일치하는 방식으로 모든 자동 형식체계의 작업을 수행할 수 있다는 결론이 나온다.

3. 전산적 마음

괴델정리가 발표된 이후 수학자들은 힐버트의 야심찬 목표에서 전제되는 완전성이 아니라 준완전성(quasi-completeness)에 만족해야만 했다. 그러나 괴델정리는 수학이 단순히 규칙들의 집합을 기계적으로 적용하는 것이 아니라 창조적인 지적 활동으로 간주될 수 있는 가능성을 함축한다. 괴델정리는 또한 인간의 사고 과정을 기계적인 계산 과정으로 간주하는 인간기계론에 대한 철학적 논변에 이용되고 있다.[5]

인간과 기계에 관한 문제에 대한 괴델 자신의 생각은 매우 모호하고 신비적이다. 괴델은 자신이 정리가 마음의 모든 수학적 직관을 형식화할 수 없다는 점을 증명했다고 보았다. 만약 마음이 직관의 일부를 형식화할 수 있다면 바로 그 사실은 새로운 이러한 형식체계의 무모순성과 같은 새로운 직관을 낳게 된다. 이러한 사실은 수학의 불완전성을 보여준다. 괴델은 또한 그 당시까지 증명된 바에 의하면 실제로는 수학적 직관과 동일하지만 증명될 수는 없는 정리증명 기계(theorem- proving machine)가 존재할 가능성이 있다고 생각했다. 나아가 경험적 탐구에 의해 그

5) H.Wang, (1974), pp. 324~326.

러한 기계를 발견할 가능성도 배제하지 않았다. 그런데 괴델정리는 인간기계론에 대한 찬반의 입장에 대한 논거로서 동시에 이용되고 있다. 예를 들어 루카스(Lucas)나 펜로제(Penrose)의 경우처럼 기계론에 대한 비판의 논거로 사용되기도 하고, 홉스텟터(Hofstadter)나 웨브(Webb)의 경우처럼 지지의 논거로 사용되기도 한다.

루카스가 기계론을 비판하는 근본적인 동기는 도덕적 책임을 물을 수 있는 근거로서의 자유의지를 확보하는데 있다. 기계론에 대한 루카스의 비판을 다루기 전에 먼저 비판의 촛점이 되고 있는 기계의 의미를 명확히 할 필요가 있다. 루카스가 염두에 두고 있는 대상은 인간을 튜링기계로 보는 기계론이다. 루카스의 논변은 모든 기계가 형식체계의 실례라는 전제에 의존하고 있다. 기계는 유한한 작동 유형과 초기 가정을 지니므로 고정적이며, 고정적인 추론규칙들의 집합에 따라 작동하므로 규칙 준수적이다. 또한 기계의 출력은 단지 구성방식과 입력에 의해서만 결정된다. 이러한 특징을 지닌 기계는 연역적으로 닫힌 체계에 해당되는 반면에 인간은 일관적인 사고 체계를 갖는다.[6]

(12) 기계는 연역적으로 닫힌 체계이다.

키르크(Kirk)는 루카스의 기계론 비판을 다음과 같이 요약하고 있다.[7]

---

6) J. R. Lucas(1961), pp. 256~257. (12)에서 제시된 기계 개념은 루카스와 동일한 입장을 취하는 사람들에게서 공통적으로 발견된다. 예를 들어 네이겔과 뉴만에 따르면 기계는 지침들의 고정된 집합을 갖고 있으므로 괴델정리가 보여주듯이 수론의 문제들을 해결할 수 없다. E. Nagel and J. R. Newman (1958), p. 100.

7) R. Kirk(1986), pp. 437~438.

(13) 루카스의 논변

a. 기계론에 따르면 모든 인간은 기계이다. 특히 모든 훌륭한 수학들도 인간이다.
b. 모든 기계는 형식체계 TM의 구체적 실례이다.
c. 그러므로 특정한 수학자가 기계라면, 그는 TM의 실례이다.
d. 그런데 괴델정리에 의하면, TM은 증명될 수 없는 괴델문장 G을 포함하는데 반해, 그 수학자는 TM의 외부에 있으므로 G가 참임을 알 수 있다.
e. 그 수학자가 기계라면, 그는 G의 참을 발견할 수 없다.
f. 따라서 그는 기계가 아니며, 기계론는 잘못이다.

루카스의 논변에서 핵심 부분은 (b)와 (d)이다. 그의 논변의 타당성은 전적으로 두 명제에 의존하고 있으므로, 이를 중심으로 검토해 보기로 한다.[8]

모든 기계는 연역적인 형식체계라는 주장 (b)는 만약 비연역적인 형식체계의 사례인 기계가 있다면 성립될 수가 없을 것이다. 넬슨(Nelson)은 「수학원리」의 정리들을 휴릭스틱스에 의해 증명하는 프로그램 (*Logic Theorist,* Newell and Simon)을 들어 (b)

8) 루카스의 주장은 그 당시는 물론 현재까지 많은 논쟁을 야기하고 있다. 루카스와의 직접적인 논쟁은 Lucas(1961), Benacerraf(1967), Lucas(1968)이다. 베나세라프의 비판 요지는 루카스가 증명가능성과 참을 혼동하고 있다는 것이다. 이러한 비판이 가능한 이유는 루카스(Lucas, 1961: 256)가 "provable-in-the-system"과 "producing as being true"를 동일시하고 있기 때문이다. 루카스는 이에 대해 1968년 논문에서 이를 인정하고 그것이 의도적이 아님을 주장했다. 따라서 본 논문에서는 그에 대한 비판을 다루지 않았다. 루카스를 지지하는 입장은 Nagel and Newman (1958), Penrose(1989) 등이며, Chihara(1972), Hofstadter (1979), Webb(1980), Lewis (1989) 등이 대표적인 반대입장에 속한다.

는 잘못된 주장임을 보이려고 한다.[9] 넬슨이 지적한 프로그램 이외에도 인공지능 분야에서는 비연역적인 추리과정을 모의하는 프로그램들이 제안되고 있다. 예를 들어 사이먼의 베이컨(*BACON*) 프로그램은 자료로부터 이론을 귀납해내고, 타가드(Thagard)의 파이(*PI*) 프로그램은 유비추리에 의한 귀납을 수행한다.

비연역적인 기계가 가능하다는 비판에 대해 루카스와 그의 지지자들은 휴리스틱스를 이용한 프로그램 자체는 본질적으로 연역적인 규칙으로 구성되어 있으므로 결국은 괴델정리가 적용된다고 대응할 수 있다. 넬슨은 다시 닫힌 알고리즘(closed algorithm)과 준알고리즘(semi-algorithm)을 구별함으로써 루카스의 논변을 지지하려는 시도를 반박하려고 한다.[10] 여기서 준알고리즘이란 주어진 과제를 완벽하게 수행할 것이라는 보장이 없는 휴리스틱스를 의미한다. 넬슨에 따르면 닫힌 알고리즘이 불가능한 과학적 가설의 발견이나 음악의 작곡 과정 등이 준알고리즘에 의해 프로그램될 수 있다.

한편 펜로제는 넬슨이 제기한 유형의 재비판에 대해 알고리즘을 좀더 포괄적으로 해석함으로써 루카스의 입장을 뒷받침한다.[11] 펜로제는 알고리즘을 컴퓨터에 의해 모의될 수 있는 모든 추리과정이라고 넓게 정의함으로써 반대자들의 논거를 부정한다. 즉 휴리스틱스와 비연역적인 학습뿐만 아니라 또다른 인간기계론이라고 간주되는 연결론적 학습도 알고리즘이라고 간주된다. 따라서 전통적인 기호론적 기계든 연결론적 기계든 모든 기계는 단지 규칙을 따를 뿐이므로 참을 판단할 수단을 갖고 있지 않다.[12]

비연역적인 형식체계의 가능성을 제시하여 루카스의 논변을 비

9) R. J. Nelson(1980), p. 103.
10) R. J. Nelson(1980), p. 105.
11) R. Penrose(1989), p. 539.
12) R. Penrose(1987), p. 146.

판하는 전략은 두가지 점에서 결정적인 반박이 될 수 없다. 우선 비연역적 형식체계의 사례로서 제시된 프로그램들을 구현하는 기계를 문자 그대로 비연역적인 형식체계로 볼 수 있는가라는 문제가 제기된다. 그러나 이러한 문제는 또다른 해결하기 어려운 문제, 즉 튜링머신이 추상적인 형식체계임은 분명하지만 프로그램을 구현하고 있는 구체적인 물리적 대상으로서의 기계를 튜링기계라고 간주할 수 있는가라는 문제를 야기한다.

이러한 정의상의 문제를 접어두고라도 비연역적인 형식체계가 존재한다는 경험론적인 주장은 경험에 의해서 확증된 것은 아니라는데 문제가 있다. 현재의 인공지능 분야에서 제시되는 프로그램들이 비연역적인 또는 준알고리즘적인 과정에 의해 과학적 가설의 발견이나 예술가의 창안을 모의하는데 성공적이라는 평가는 정확하게 말하자면 '부분적으로 성공적'이라는 평가로 대체되어야 한다. (b)에 대해 비연역적인 기계가 존재한다는 넬슨의 비판을 둘러싼 논쟁은 결국은 연역적 형식체계와 알고리즘을 어떻게 볼 것인가라는 정의상의 문제와 경험론적 문제로 전환된다. 이러한 전환은 괴델정리를 이용한 루카스의 논변이 과연 타당한가를 검토하는 문제를 근본적으로 해결하는 것은 아니므로 다른 방향에서 접근해야 한다.

루카스는 (d)를 주장함으로써 마음의 작용은 기계와는 달리 체계 밖을 지향할 수 있다는 점을 강조한다.[13] 루카스에 따르면 이러한 지향성 때문에 마음은 자신에 대해 반성할 수 있으며, 그 결과를 검토할 수 있다. 즉 마음은 자기지시적 또는 자기반성적이라는 특징을 갖는다. 마음은 이러한 작용을 하는데 있어서 다른 요소에 의존하지 않는다. 마음은 문자 그대로 완전한 체계이

13) J. R. Lucas(1961), p. 269.

다. 그러나 기계는 괴델문장 G를 알 수 없으므로 자기지시성을 갖고 있지 않다. 이러한 분석에 따르면 우리는 자기지시성의 유무에 따라 마음과 기계를 구분할 수 있다. 이제 마음은 자기지시성을 갖고 있다는 점에서 기계보다 우월하다고도 말할 수 있게 된다.

(14) 마음과 기계를 구분하는 기준은 자기지시성이다.

슬레작(Slezak)이 지적하고 있듯이[14] 루카스는 (d)에서 체계 TM과 관련된 두 종류의 자기지시성을 혼동하고 있다. 첫번째의 자기지시성은 괴델문장 G가 자신에 대해 갖는 자기지시성이다. 루카스가 TM에는 결여되어 있다고 지적한 자기지시성은 바로 여기에 해당된다. 두번째의 자기지시성은 기계가 그 자체에 대해 갖는 자기지시성이다. 괴델문장 G는 TM의 기호로 표현될 수 있으므로 "TM에서 증명가능함"이라는 관계는 TM의 기호에 의해 표현가능하다. 이러한 관계는 TM에 의한 것이지만 TM 자체에 관한 것이기도 하다는 점에서 자기지시성을 갖는다.

이러한 혼동은 퍼트남(Putnam)이 잘 지적하고 있듯이[15] 루카스의 논변이 괴델정리를 잘못 적용하는 있다는 사실로부터 나타난다. 여기에 하나의 튜링기계 TM이 있다고 하자. 루카스의 논변을 비판하기 위해서는 우리는 다음의 메타언어 문장 G에서 증명할 수 있는 문장 U를 발견하기만 하면 된다.

(15) 만약 TM이 일관적이라면 U는 참이다.

만약 TM이 일관적이라면 괴델정리에 따라 문장 U는 TM에

14) P. Slezak(1982), p. 49.
15) H. Putnam(1975), p. 366.

의해 결정불가능하게 된다. 괴델은 (2)와 (3)을 증명하면서 TM에 관한 메타언어 문장이 TM 자체내의 대상언어 문장으로 번역될 수 있는 대응을 이용했다. 그 결과 TM은 문장 U를 증명할 수는 없지만 문장 G를 증명할 수는 있다.

더구나 TM이 매우 복잡한 체계이기 때문에 TM의 일관성을 인간 자신도 증명할 수 없을 경우, TM이 증명할 수 없는 문장 U는 인간도 증명할 수 없다. 이러한 비판에 대해 루카스는 "인간은 무모순적인 체계이다"라는 점을 당연하게 받아들임으로써 대응하려고 한다. 인간의 신념체계가 실제로 일관적이라는 주장은 여러가지 경험적인 증거로부터 살펴볼 때 정당화되기가 어렵다. 이와 관련하여 휘트리(Whitley)는 "루카스는 이 문장을 일관적으로 증명할 수 없다"는 문장을 제시하여 루카스를 비판한다.[16] 만약 루카스가 그 문장을 증명할 수 있다면, 그 사실은 바로 그가 가정한 일관성을 위반하는 것이다. 그러므로 루카스에게는 참이라고 알 수는 있어도 증명할 수 없는 것이 있다. 더구나 홉스텟터와 데닛(Hofstadter and Dennett)처럼 루카스에게 "루카스는 이 문장을 일관적으로 믿을 수 없다"는 문장을 제시한다면, 그는 이제 그 문장을 증명할 수 없음은 물론이고 그것을 믿을 수조차도 없게 되는 상황에 처하게 된다.[17]

이상의 구체적인 잘못에도 불구하고 루카스의 전체적인 논변을 인정하더라도 (f)가 도출되지 않는다. 마음은 TM의 괴델문장 G를 증명할 수 있으므로 기계라고 볼 수 없다는 주장은 마치 "나의 친구는 자신의 등을 볼 수 없지만 나는 볼 수 있다"는 사실로부터 "나의 친구와 내가 다르다"고 주장하는 것과 같다. 또한 "기계가 증명할 수 없는 것을 인간은 증명할 수 있으므로 인간은

16) C. H. Whitley(1962), p. 61.
17) D. R. Hofstadter and D. C. Dennett(1981), P. 277.

기계가 아니다"라는 주장을 타당하다고 인정하더라도 루카스는 인간과 기계가 서로 다른 사례(token)란 점만을 증명했을 뿐이며 양자가 동일한 유형(type)일 가능성은 남아있다. 루카스가 자신의 논변을 확립하기 위해서는 자신의 괴델문장 G를 결정할 수 있음을 증명해야 한다. 인간은 형식체계이던가 아니던가 두가지 경우 중 어느 하나에 속한다. 만약 인간이 형식체계라면 괴델정리를 적용할 수 있다. 이경우 인간은 자신의 괴델문장을 알 수 있는가? 우선 인간에 있어서 괴델문장은 인간이 결코 알 수 없는 진리일 것이다. 우리가 알 수 없는 것이 있다고 생각하는 것이 타당할 것이다. 따라서 인간은 기계와 마찬가지로 불완전하다. 한편 인간이 비형식체계라면 괴델정리를 적용할 수 없다. 그러므로 괴델정리를 이용하여 인간이 기계가 아니라는 점을 증명하기는 불가능하다고 보아야 한다. 루카스는 단지 괴델정리를 이용한 또 하나의 반기계론을 주장했을 뿐이라고 평가할 수 있다.

루카스가 인간기계론을 반박하는데 괴델정리를 이용하는 것과는 정반대로 홉스텟터는 인간기계론을 지지하는데 괴델정리를 이용한다. 그러나 홉스텟터는 일차적으로 괴델정리의 내용, 즉 형식체계의 불완전성이 아니라 괴델이 자신의 정리를 증명하는데 이용한 자기지시성을 중시한다. 홉스텟터에 따르면 괴델의 업적은 모든 형식체계는 그 자체에 대한 자기지시성을 갖고 있다는 점을 증명한데 있다. 홉스텟터는 괴델정리 뿐만 아니라 에셔(Escher)의 그림과 바하(Bach)의 음악, 그리고 인공지능의 프로그래밍언어인 리습(LISP) 등에서 그러한 자기지시성을 확인했다. 우리의 마음 역시 그러한 자기지시성을 갖는 하나의 체계이다.

자기지시성은 회귀개념과 밀접한 관계가 있는데, 회귀 개념은 다른 차원에서 발생하는 동일한 현상에 기반을 두고 있다. 회귀

과정을 통해 추론규칙을 공리집합에 적용하면 회귀적으로 매거가능한 집합(recursively enumerable set)이 도출된다. 회귀적 매거는 고정된 규칙에 의해 새로운 것이 옛것으로부터 창출되는(emergent) 과정이다. 이러한 과정에서 예측불가능성이 나타난다. 회귀 과정에 따라서 행위의 복잡성이 증가되므로 그 과정이 진행될수록 예측불가능성이 증가한다. 따라서 충분히 복잡한 회귀적 체계는 미리 결정된 모든 패턴을 파기할 정도로 강력해진다.

(16) 연역적으로 닫힌 회귀 체계도 예측불가능성을 지닐 수 있다.

홉스텟터의 분석에 따르면 회귀 체계는 루카스가 말한 형식체계와 동일하지 않다. 물론 회귀체계에는 고정된 추론규칙이 있지만 결정되지 않은 예측불가능성이 있으며 창출현상이 있다. 회귀체계는 그 자체를 수정할 수 있는 정교한 프로그램, 즉 그 자체를 확장하고 향상시키고 일반화시킬 수 있는 프로그램을 가질 수 있다. 홉스텟터에 따르면 모든 형식체계는 회귀적으로 확장될 수 있으며, 그 결과 본래의 체계의 정리가 아닌 참인 문장 G_1이 새로운 체계의 정리가 될 수 있다.[18] 물론 괴델정리에 의해 다시 새로운 체계의 정리가 아닌 참인 괴델문장 G_2가 있지만 일단 괴델문장 G_2의 참을 발견하면 그것을 정리로서 프로그램할 수 있다. 이러한 과정은 무한히 계속될 수 있으며, 궁극적으로 지극히 복잡해서 인간도 증명할 수 없는 새로운 문장 G_n이 창출될 수 있다. 그러므로 홉스텟터에 따르면 루카스가 제시한 인간과 기계를 비교할 수 있는 기준 (14)는 성립될 수 없다.

홉스텟터는 이러한 정교한 회귀를 인간 지능의 본질로 본다.[19]

---

18) D. R. Hofstadter(1979), p. 152.

19) 자켄도프(Jackendoff)는 홉스텟터가 주장하는 의식은 외부 세계에 대한 원초적 의식이 아니라 단지 자기성찰이나 자기자각에 해당될

자기지시성은 일종의 회귀현상이므로 형식체계도 자기지시성을 갖는다고 주장할 수 있다.

(17) 형식체계도 자기지시성을 갖는다.

홉스텟터와 루카스는 모두 지능의 특징을 자기지시성으로 간주한다는 점에서 일치한다. 인간은 현재 수행하고 있는 과제로부터 벗어나 자신이 한 일을 반성해 볼 수 있다. 루카스는 이를 인간에게만 고유한 '체계 밖에 있음'(outside the system) 현상으로 간주했다. 홉스텟터는 프로그램된 것을 모두 체계라고 분류하면 컴퓨터는 체계 밖에 있을 수 없다고 인정한다. 그러나 '체계 밖에 있음'을 주장하는데 있어서 단순히 자신을 성찰하는 것과 자신을 초월하는 것은 구별되어야 한다고 주장한다. 예를 들어 자기초월은 선불교에서의 초월적 자기이탈에 해당한다. 모든 인간이 자기초월적 능력을 갖고 있다고 보기는 어려우며, 인간이 자기초월을 하였다고 할 경우일지라도 그것이 과연 인간의 한계를 넘어섰다고 할 수 있을런지도 의문이다. 왜냐하면 인간이 넘어설 수 있는 것은 이미 인간의 한계가 아니기 때문이다.

따라서 홉스텟터는 괴델정리가 보여준 형식체계의 불완전성은 그 체계가 할 수 있는 것에 대해 우리가 비현실적 기대를 할 경

뿐이라고 비판한다. 즉 경험으로서의 의식이 부정되고 있다는 것이다. 그는 홉스텟터가 의식과 자기 의식(반성적 자각)을 혼동하고 있다고 비판한다. R.Jackendoff(1987), p.18. 자켄도프는 마음을 정보처리 체계로서 간주하는 전산적 마음을 배격하고 의식적 자각, 세계와 우리의 내적 삶에 대한 경험과 관계하는 현상학적 마음(phenomenological mind)을 주장한다. 자켄도프의 입장을 더욱 발전시켜 퐁티(Merleau-Ponty)의 현상학을 도입하여 인간 경험의 문제를 다루어야 한다고 주장하는 입장은 바렐라 등(Varela etc, 1991)에서 볼 수 있다. 그들은 본 논문에서 약간 언급되었던 인간의 초월 문제등을 불교의 중관사상과 관련하여 다루고 있다.

우에만 단점으로 간주된다고 주장한다.[20] 환언하면 형식체계의 불완전성은 단지 그 체계가 할 수 있다고 기대된 모든 것을 할 수 없다는 지극히 당연한 사실을 지적한 것이다. 괴델은 이러한 지극히 당연한 사실을 엄밀한 수학적 방식으로 증명했다. 형식체계가 지니는 불완전성에 대한 이러한 홉스텟터의 해석은 인간도 기계와 마찬가지로 실제로 제한되어 있다고 주장한 튜링의 견해와도 일치한다.[21]

4. 결론

이 글의 일차적인 목적은 수학적인 계산의 의미를 마음을 이해하는데 적용한 결과 나타난 마음에 대한 전산적 개념의 타당성을 살펴보는 것이었다. 예비적인 단계로서 괴델정리와 튜링기계 개념이 다루어졌고, 이러한 두 개념을 주로 이용하여 루카스의 논변을 검토하였다.

인간기계론을 논박하기 위해 괴델정리를 이용한 루카스의 전략은 잘못임이 드러났다. 루카스는 튜링기계 개념에 근거를 둔 매우 좁은 의미에서의 기계론만을 자신의 표적으로 삼았음에도 불구하고 그의 논변을 확립할 수 없음이 지적되었다. 이 글에서는 홉스텟터의 이론을 다루면서 약간 언급을 했지만 연결론이 함축하는 새로운 기계론은 루카스가 전제하는 기계론과 많은 차이점이 있다. 한편 루카스가 인간기계론을 논박하는데 사용한 자기지시성은 홉스텟터에 의해 긍정적인 요소로 작용하여 새로운 유형의 기계론을 지지하는데 이용되고 있다. 따라서 괴델정리는 인공지능에 대해 부정적인 함축만을 지닌다는 주장은 근거를 상실했

20) Hofstadter(1979), 77.
21) A. Turing(1950), p. 445.

다고 보아야 할 것이다.

그렇다고 해서 이러한 결론이 곧 홉스텟터가 주장하는 "강한 인공지능"의 지지 근거로 간주되서는 안된다. 왜냐하면 반기계론의 한가지 사례가 반박되었다고 해서 기계론 자체가 반박되는 것은 아니며, 인간기계론의 적극적 근거가 되는 것도 아니기 때문이다. 이 논문은 단지 괴델정리를 이용하여 전산적 마음의 개념을 비판하고 인공지능의 가능성을 비판하는 것은 잘못이라는 점을 지적했을 뿐이다.

참 고 문 헌

이영의(1990) "튜링기계와 기능주의". 한국인지과학회 추계학술대회논문집, 154~163.

이초식, 이영의(1992) "괴델정리와 인공지능". 한국인지과학회 춘계학술대회논문집, 3~13.

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

Chihara, C.(1972) On alleged refutation of mechanism using Gödel's incompleteness results. *Journal of Philosophy 64,* 507~526.

Gödel. K.(1931) On formally undecidable propositions of Principia methematica and related systems I. In J. van Heijenoort (1967) *From Frege to Gödel.* Harvard Univ. Press. 596~616.

Hofstadter, D. R.(1979) *Gödel, Escher, and Bach.* Basic Books.

Hofstadter. D. R. and Dennett. D. C.(1981) *Mind's I.* Basic Books.

Hofstadter, D. R.(1984) *Metamagical themas.* Viking.

Hopcroft, J. E.(1984) Turing Machine. *Scientific American 250,* 70~80.

Hopcroft, J. E. and Uullman, J. D.(1979) *Introduction to Automata Theory, Languages and Computability.* Addison Wesley.

Kirk, R.(1986) Mental machinery and Gödel. *Synthese 66,* 437~52.

Lewis, D.(1989) Lucas against mechanism II. *Canadian Journal of Philosophy 9,* 373~76.

Lucas,J.R.(1961) Minds, Machines, and Gödel. *Philosophy 36,* 120~124. Reprinted in *Minds and Machines,* A.R.Anderson (1964). Prentice-Hall, 43~60.

Lucas,J.R.(1968) Satan stultified: A rejoinder to P.Benacerraf. *The Monist 52,* 145~58.

Nagel,E. and Newman,J.R.(1958) *Gödel's Proof.* New York Univ. Press.

Nelson, R. J.(1980) *The logic of mind.* Kluwer.

Penrose, R.(1989) *The emperor's new mind: concerning computers, minds and the law of physics.* Vintage.

Penrose, R.(1990) Precis of the Emperor's New Mind. *Behavioral and Brain Science 13,* 643~705.

Putnam, H.(1975) Minds and machines in *Mind, language and reality.* 362~385. Cambidge Univ.Press.

Slezak. P.(1982) Gödel's theorem and the mind. *Britisch Journal for the Philosophy of Science 33*, 41~52.

Turing, A.(1936) On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society Series 2, 42*, 230~265.

Wang, H.(1974) *From Mathematics to Philosophy*. R.K.P.

Whiteley, C. H.(1962) Minds, Machines, and Gödel: A Reply to Mr Lucas. *Philosophy 37*, 61.