일반연구논문

알파고 사례 연구: 인공지능의 사회적 성격■

김지연*

■ 기계학습의 일반원리를 해석하는데 도움을 준 고려대학교 제어계측공학과 박주영 교수에게 감사드립니다. 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016S1A5A2A03927422).

* 고려대학교 연구교수, 과학기술학 박사 전자우편: spring900@gmail.com

2016년 알파고(AlphaGo)와 이세돌(9단)의 경합은 인공지능의 등장을 대중적으로 확인하는 것이었다. 이 대국은 일종의 확장된 튜링테스트(Extended Turing Test)였다. 튜링테스트의 목적은 기본적으로 기계가 인간을 모방할 수 있는지를 관찰하는 것이다. 이 논문은 알파고 사례를 통하여 인간과 인공지능 사이의 상호작용과 그 사회적 성격을 분석할 것이다. 콜린스(Collins, H.)는 우리의 지능은 사회적이며, 튜링테스트의 목적은 궁극적으로 사회 구성원과 비구성원을 구별하려는 것이라고 제시했다. 그러므로 기계가 이 테스트를 통과한다면, 비록 특정한 수행 차원에 한정된 것이지만, 우리가 그 기계(비구성원)를 인간구성원과 구별할 수 없게 되었다는 것을 의미한다. 이세돌–알파고 대국을 튜링테스트 설정에 대입하면, 알파고는 인간의 진실을 흔드는 역할(A)을 수행했고, 이세돌은 인간의 진실을 증언하는 역할(B)을 수행했다. 이 테스트에서 중요한 것은 알파고의 기능적 성능이 아니라 면접관(C)의 사회적 승인이다. 이 대국 과정에서 다수의 관중들이 면접관의 역할을 수행했다. 여기서는 그들을 '인간면접관'이라고 부를 것이다. 그들은 대국과정에서 오랫동안 체화된 자신의 사회적 지식을 통하여 자신의 동료 구성원(인간)과 비구성원(기계)을 구분하고자 했다.

주제어 | 알파고, 인공지능, 이세돌, 튜링테스트, 인간면접관, 시민과학

1. 서론

알렌 튜링은 자신의 논문(Turing, 1937)에서 "생각하는 기계(thinking machine)" 개념을 제시했는데, 흔히 "튜링 기계(Turing Machine)"라고 불리며, 이 후 컴퓨터와 인공지능의 기원이 되었다. 1956년 다트마우스 회의(Dartmouth conference)를 기점으로 기계 지능 문제가 구체적으로 부상했다. 이어서 1960년대 중반 기업 내 인공지능(artificial intelligence, AI) 실험실이 등장했고, 1970년대 들어서 학계의 연구영역이 되었으며 1980년대에는 과학의 영역이 되었다.

1996년 IBM의 컴퓨터 딥블루(Deep Blue)는 체스 챔피언 카스파로프에게 승리하면서 인공지능 개념을 대중적으로 알렸다. 2011년 IBM의 왓슨(Watson)은 자연어를 사용하는 〈제퍼디쇼!〉에서 퀴즈챔피언들을 물리치고 승자가 되었다. 2015년 구글 딥마인드사의 알파고(AlphaGo)는 유럽 챔피언 판후이(2단)와 대국하여 5대0으로 승리했고, 곧이어 2016년 이세돌(9단)과 경합하여 4대1로 승리했다.

바둑 대국이라는 형식을 통해서 우리는 인간과 기계의 경합을 모두 함께 관전할 수 있었다. 그 과정에서 '인공지능'이 이전의 기계와는 다르다는 경험을 공유했으며 열광과 혼란을 동시에 느꼈다. 우리들은 이 상황에 대한 질문들을 쏟아냈다. 예를 들면 "인공지능은 우리사회를 얼마나 변화시킬 것인가?", "인공지능은 책임의 주체가 될 수 있나?", "인간은 인공지능을 통제할 수 있나?", "인공지능이 인간의 노동을 대체할 것인가?" 등이다.

인공지능에 대한 질문의 등장은 인공지능에 대한 비판의 출현이다. 이

질문들에 답을 찾기 위해서는 사회적 담론이 전개되어야만 할 것이다. 인공지능에 대한 질문과 해석의 과정은 한편으로는 우리가 인공지능과 상호작용하는 과정이 될 것이고 다른 한편으로 우리가 인공지능을 이해해가는 과정이 될 것이다.

이 논문은 "우리가 목격한 것이 무엇인가?"라는 기본적인 질문으로부터 출발한다. 우리는 무엇인가 중요한 사건을 목격했음을 잘 안다. 그러나 그 의미에 대한 해석은 아직 진행 중이며, 그 해석은 인공지능과 관련되 인공지능과 관련된 다른 질문들과 연관되어 있다. 인공지능의 의미에 관한 질문은 인공지능의 정의에 대한 질문으로 이어진다. "우리가 목격한 것이 정말 인공지능인가?"

그런데 인공지능을 정의하려는 순간 우리는 즉각적으로 난관에 부딪힌다. 인공지능을 우리 인간과 분리하여 독립적으로 고찰하기 어렵기 때문이다. 이 난관은 "생각하는 기계"를 개발하려고 했던 많은 공학자들과 그것을 비판적으로 해석해왔던 연구자들 모두가 직면했던 문제였다. 그래서 튜링(Turing, 1950)은 인공지능의 정의 문제를 인간지능의 모방 문제로 전환했다. 즉, 튜링에 따르면 인공지능은 '뛰어남'이 아니라, 인간과 구별이 불가능할 정도로 인간을 모방하는 능력이다.

이 논문에서는 먼저 이론적 검토를 위해서 튜링의 "생각하는 기계(thinking machine)"개념에 대해서 알아보고, 이어서 튜링테스트(Turing test) 및 관련 비판 연구들을 살펴본다. 콜린스(Collins, H.)에 따르면 튜링테스트는 인공지능의 기술적 역량에 관한 것이라기 보다는 인공지능의 사회적 자격을 판정하는 문제이다. 마지막으로 알파고와 이세돌 대국을 확장된 튜링 테스트(extended Turing test)로 다루고자 한다. 그 프레임 안에서 각 주체들의 역할과 상호작용, 그리고 그 사회적 의미를 제시한다. 대국 과정에서 이세돌과 알파고 뿐만 아니라, 바둑전문가 및 해설가를 포함하여 다수의 관중들이 많은

이야기를 생산했다. 그들의 이야기가 이 논문의 주요 구성요소가 되었다.

2. 이론적 검토

이론적 분석을 위해서 먼저 튜링기계 개념을 살펴보고, 튜링 테스트의 구조와 등장 배경을 알아본다. 이어서 인공지능 비판 및 과학사회학 연구들을 소개한다. 오래전부터 다수의 연구들이 인공지능을 승인하게 될 순간에 대해서 대비해왔다. 선행연구들을 살펴보는 것은 알파고를 목격한 우리의 현재를 해석하는 출발이 될 것이다.

1) 튜링 기계

이 모든 일은 튜링(Turing, 1950)이 "기계는 생각할 수 있나(Can machine think)?"라는 질문을 던졌을 때 시작되었다. 생각하는 기계를 만들려면 먼저 "생각하다"라는 개념을 정의할 수 있어야 한다. 이 질문은 "지능(intelligence)"의 정의 문제인데, 그 개념을 정의하려는 순간 명백하다고 여겨졌던 이 개념이 매우 정의하기 어렵다는 것을 깨닫게 된다. 우리는 다른 사람의 사유 또는 마음을 직접 본 적이 없다. 우리 자신의 사유 또는 마음조차도 설명하기 어렵다. 다만 우리는 서로의 실천을 관찰함으로써 그의 '사유' 또는 '마음'을 추론할 수 있을 뿐이다.

튜링에 의하면, 생각하는 기계는 처음부터 완벽하게 정의되어서는 안 된다. 만약 이미 잘-정의된 프로그램을 내장하려고 한다면, 그 기계의 저장 공간은 프로그램의 정의와 명세들로 거의 점유될 것이고 내적 규칙에 따라서 이미 확정된 합리적인 행동만을 할 것이다. 그렇다면 그 기계는 추가적으로 무엇인가를 더 학습할 필요가 없다. 거기에는 기계가 무엇인가를 생각할 여지가 없다.

다시 말해서, 생각한다는 것은 부족한 것을 채우려는 학습으로부터 나온다. 그렇다면 기계는 어떻게 그런 학습역량을 획득할 것인가? 튜링은 어른('완전한 인간')을 모방하기 보다는 아이('씨앗')를 모방하는 방법이 더 적절하다고 보았다. 튜링은 학습 행동에 대해서 다음과 같이 말한다.

> "기계가 그 자신의 행위의 결과를 관찰함으로써 자신의 목표를 더 효과적으로 달성하도록 자신의 프로그램을 수정할 수 있다. 이것은 유토피아의 꿈이 아니라, 가까운 미래의 가능성이다(Turing, 1950: 449)."

여기에서 특히 "자신의 행위의 결과"와 "스스로 프로그램을 수정"이란 개념이 중요하다. "생각하는 기계"는 지속적으로 자신과 자신의 환경 사이의 상호작용의 결과를 관찰하여 자기 자신을 개선하는 역량을 가져야 한다. 이것이 바로 기계의 학습 역량이다. 이 기계는 스스로를 개선함으로써 결과적으로 개발자에 의해서 주어진 최초의 상태에서 벗어나게 된다.

튜링 기계는 다음과 같이 두 부분으로 구성된다. (1) 아이-기계(child-machine)와 (2) 교육 과정(education process). 아이-기계는 완전하지 않지만 학습할 역량이 있는 최소한의 프로그램이다. 교육 과정을 통해서 이 기계는 보상-신호의 반복가능성이 증가하는 방향으로 나아간다. 기본적으로 보상과 처벌에 관한 정보의 총량이 학습의 양이 될 것이다. 또 다른 방법 중 하나로 튜링은 학습하는 기계를 위해서 무작위 요소(random element)를 권고한다. 기계의 무작위성은 모든 경우를 계산하는 것보다 학습 속도를 향상할 수 있다(Turing, 1950: 455-459).

오늘날 튜링기계의 한 영역으로서 기계학습(machine learning)은 "명백하게 프로그램 되어 있지 않은 채, 학습할 수 있는 역량"을 지칭한다. 자신의 경험을 학습하는 컴퓨터를 구현할 수 있다면 대량의 상세 프로그래밍을 해야

하는 노고와 필요도 생략할 수 있을 것이다(Samuel, 1959). 오늘날 인공지능 연구자들은 튜링처럼 인간을 모방하는 방법을 포함하여 합리성을 구축하는 다양한 접근법을 개발하고 있다. 여기서 합리성은 결코 전지전능하다는 것이 아니며, 다만 선택 순간까지의 인지 시퀀스에 의존하는 것이다(Russell, et al., 2010: 1-5, 39).

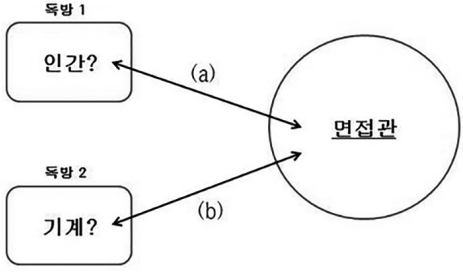
한편 튜링기계 개념은 많은 비판을 받아왔다. 대표적으로 허버트 드레퍼스(Dreyfus, 1965)는 논문 〈연금술과 인공지능(Alchemy and Artificial Intelligence)〉을 통해서 튜링기계의 가능성을 부정했다. 그에 따르면 인공지능은 연금술이 그러했듯이 불가능한 도전이다. 그는 먼저 컴퓨터 언어가 인간 행위를 분석하는데 적절한지를 질문했다. "(컴퓨터의) 불연속적이고 확정적인 작용으로 인간의 지적 행위에 대한 완벽한 분석이 가능한가?", "그런 디지털 용어로 인간의 지적 행위에 대한 적절한 분석은 가능한가?" 이 질문들에 대해 드레퍼스는 "아니다"라고 결론 내린다. 분절적인 형식의 컴퓨터 작동이 맥락 차원에서 체화된 인간의 지적 역량을 가질 수 없다는 것이다. 그의 주장은 옳다. 다만 튜링은 '생각하는 기계'를 만들려고 한 것이지, '인간과 동일한 기계'를 만들려 했던 것은 아니었다는 점을 기억할 필요가 있다.

2) 튜링테스트와 비판연구

튜링기계에 대한 질문들 중에서 가장 먼저 관심을 끄는 것은, 만약 생각하는 기계를 만들었다면 "그 기계가 정말 생각하는지 어떻게 알 수 있는가?"라는 것이다. 튜링은 이미 직접적으로 지능을 정의하는 것이 어렵다고 여긴 것 같다. 그는 기계의 지능을 확인하는 전혀 다른 방법을 개발했다. 튜링테스트(Turing Test)는 기계의 지적 역량을 드러내는 방법으로서, 기계 지능의 문제를 기계의 모방의 문제로 전환했다. 튜링테스트는 지능에 대한 절대평가가 아니라 일종의 비교평가 방법이라고 할 수 있다(이상욱, 2009).

튜링은 다음과 같은 놀이를 꾸며냈다. 한 남자(A)와 한 여자(B) 그리고 면접관(C)이 있다. 면접관은 두 사람과 분리된 방에 있고 텔레타이프를 통해서 방 안의 사람들과 문자로 의사소통할 수 있다. 여기서 (A)는 자신의 정체를 숨기면서 '여자'라고 주장한다. 한편 (B)는 면접관에게 "나는 여자다. 그(A)의 말을 듣지 마라!"라고 말할 것이다. 면접관은 그들의 대답을 듣고서 누가 남자이고 누가 여자인지를 판단한다. (A)의 목표가 면접관의 잘못된 판단을 유도하는 것이라면, 반대로 (B)는 진실을 말하며 면접관을 돕는 역할을 한다. 이제 (A)의 자리에 기계를 놓아 보자. "무슨 일이 일어날 것인가?" 면접관은 남자와 여자를 판단하는 게임을 할 때처럼 이 게임에서도 잘못된 판단을 할 것인가 (Turing, 1950: 433-434).

이 설정에 따르면 기계(A)는 자신이 '인간'이라고 주장하면서 거짓을 말해야 하고 인간(B)은 면접관을 돕기 위해서 진실을 말해야 한다. 면접관(C)은 (A)와 (B)를 관찰하면서 실제 수행과 모방 사이를 구별해야 한다. 5분의 대화 후 면접관이 기계와 인간을 구별하지 못한다면, 그 기계는 테스트를 통과한 것으로 간주한다. 그로써 그 기계는 언어적 차원에서 지능을 가진 것으로 승인될 것이다.



〈그림 1〉 튜링 테스트–콜린스 버전

자료: Collins(1990: 183)

원래 튜링테스트의 (A)와 (B)는 같은 방에 있다고 추론할 수 있다. 남자와 여자 테스트 사례에서 보듯이, (B)가 (A)에 대한 정보를 (C)에게 제공하는 설정이기 때문이다. 그런데 우리에게 더 잘 알려진 테스트 구조는 〈그림 1〉과 같다. 여기서는 콜린스 버전이라고 부르겠다. 두 개의 독립적인 개별 방이 있어서 한 쪽에는 인간, 다른 한쪽에는 기계가 들어간다. 방 밖에는 면접관이 있다. 그 이후 과정은 원본 버전과 같다. 콜린스 버전에서 인간은 기계를 보면서 그의 거짓을 직접 '고발'할 수는 없다. 그럼에도 불구하고 인간참가자는 면접관의 질문에 대해서 평소처럼 대화함으로써 인간 반응의 전형(실제 수행)을 제시할 수 있고, 면접관은 (A)와 (B)의 반응을 비교해 볼 수 있다. 따라서 콜린스 버전의 인간(B)도 원본 버전의 역할, 즉 '진실'을 말하는 설정 범위 내에 있다.

존 썰(Searle, 1980)은 튜링 테스트를 통과했더라도 기계가 인간처럼 지능을 가지는 것은 아니라고 비판한 것으로 잘 알려져 있다. 이를 "중국어방(Chinese Room)" 논변 또는 강한 인공지능(Strong AI) 비판이라고 한다. 그에 따르면 인공지능이 무엇인가 수행했더라도 그것은 이해하는 것이 아니라 다만 이해하는 것처럼 보일 뿐이다. '이해한다는 것'은 본래 알고리즘으로 환원될 수 없는 문맥적 지식이기 때문이다. 썰의 주장은 옳다. 그러나 튜링테스트의 목표는 인간지능의 복제(duplication)가 아니었다는 점을 기억해야 한다. 컴퓨터는 다만 인간지능을 모사(simulation)하려고 한 것이다.

드레퍼스와 썰의 비판은 인공지능 자체를 반대한 것으로 보이지만 오히려 인공지능 발전에 기여한 바가 있다. 드레퍼스에게 있어서 지능, 또는 실행은 몸을 전제로 한다. 그 실행은 체화되고 다음 실행을 지원한다. 이런 관점에서 '의식'이란 몸과 분리될 수 없다. 이를 '체화되고 자동력을 가진 의식(embodied, motile consciousness)'이라고 한다(Ihde, 2010: 38-39). 다시 말해서 의식 또는 지적 작용은 몸을 기반으로 발생하고 작동하며, 오직 몸의 실행을 통해서만 달성될 수 있다. 지능은 하나의 몸에서 다른 몸으로 손실 없이 그

대로 전송될 수 없다는 것이고, 이는 지능의 복제불가능성과 같은 말이다. 그렇다면 어떤 기계도 인간의 지능을 그대로 옮겨 담을 수 없다. 이는 인공지능 연구자들이 경험해왔던 바와도 일치한다. 기계는 지능을 복제할 수 없으므로 스스로 실행을 통해 '지능'을 획득해야 한다. 이는 아이-기계와 교육 프로그램이라는 설계를 제시했던 튜링의 시도와 일맥상통한다.

3) 지능의 사회적 본성

해리 콜린스(Collins, 1990)는 "지능" 개념의 사회적 본성을 이해해야 할 필요를 제기했다. 앞서 보았듯이 의식 또는 지능은 몸의 실행과 관련이 있다. 인간의 실행은 의도(intention)의 여부에 따라 구별될 수 있다. 눈깜빡임과 같은 실행은 의지적인 것이 아니라 단지 물리적인 것이다. 한편 글쓰기와 같은 행동은 의도를 포함한다. 물론 무의식적인 실행도 때로 체화된 의도를 포함할 수 있다.

그런데 실행과 의도를 연결하는 "코드"는 복잡하게 얽혀있고 해당 사회에 참여함으로써만 이해할 수 있다(Collins & Kusch, 1998). 그러므로 모든 행동은 필수적으로 사회적이다(Collins, 1995: 291). 우리가 지능 또는 지식이라고 부르는 것은 그런 사회적 실천을 통해서만 획득될 수 있다. 예를 들어서 암묵 지식(tacit knowledge)은 텍스트나 말로는 전달될 수 없고 해당 공동체 내에서 사회적 실행을 통해서만 획득되는 것으로 실행과정에서 거의 자동적으로 발현된다(Collins, 2007; Collins and Evans, 2012: 4).

콜린스는 지능의 사회적 성격을 강조했을 뿐만 아니라, 인공지능의 역할 역시 사회집단 차원의 문제가 될 것으로 예견했다. 예를 들어 인공심장은 누군가의 심장을 대체한다. 그런데 인공심장과 달리 누구도 자신의 뇌를 인공뇌로 대체하고 싶지는 않을 것이다. 그러므로 인공지능이 이식되는 단위는 개별인간이 아니라 인간집단이 될 것이다. 인공지능은 개별 인간의 뇌를 모방하

는 것이 아니라 사회집단의 성능을 모방할 것이다. 그러므로 인공심장이 이식될 몸에 적응해야 하듯이 인공지능은 자신이 이식될 사회집단에 적응할 필요가 있다(Collins, 1990: 14-15).

튜링테스트도 기술적 차원을 넘어서 사회적 차원의 것이 된다. 콜린스에 따르면 튜링테스트는 본질적으로 그 사회의 구성원과 비구성원을 식별하는 형식의 하나이다. 유사한 사례로 세미팔라틴스크 테스트(Semipalatinsk test)는 실제로 해당 공동체의 구성원 여부를 판별하는데 사용되었다. 이 테스트는 원폭실험으로 반핵운동에 직면했던 옛 소련 시절, 마을 원주민(구성원)과 스파이(비구성원)를 구분하려는 목적의 테스트였다. 면접관은 그 마을에 오래 살았고 그런 사회화에 기반을 둔 질문을 함으로써 원주민과 비원주민을 구별할 역량을 가지고 있는 사람이 선정된다. 원주민 면접관은 그 마을에 대해서 말이나 글을 통해서는 배울 수 없는 질문을 하기 때문에 스파이(비구성원)는 적절한 답을 하기 어렵다.

이런 맥락은 튜링테스트 면접관에게도 동일하게 적용된다. 면접관은 사소한 일상적 질문을 던지고 답을 유도함으로써 자신의 동료 구성원(인간)과 비구성원(기계)을 구별한다. 면접관의 질문은 자신의 "체화된 현실(embodied reality)"에 기반을 둘 수밖에 없다(Hayles, 1999). 튜링테스트의 통과 기준은 미리 결정되어 있지 않지만 항상 작동가능하다. 면접관의 지식은 사회적으로 체화되는 것이므로, 그 사회적 실천 범위 안에 있었던 면접관이라면 유사한 판정을 할 것이다.

여기서 간과할 수 없는 중요한 측면은 구성원과 비구성원을 구분하는 면접관의 능력이 '확실한 것'이 아니라 개연적이라는 것이다(Collins and Evans, 2012: 110). 한 사회의 구성원들은 사회적 실천을 공유하기 때문에 비구성원과 구별되는 경계를 형성한다. 그러나 그 경계는 그렇게 견고한 것이 아니며 그 경계를 형성했던 것과 마찬가지로 그들의 사회적 실천을 통해서 또

언제든지 변할 수 있다. 그런 점에서 콜린스는 기계가 튜링테스트를 쉽게 통과할 수도 있다고 우려했다.

콜비와 동료들(Colby, et al., 1972)은 확장된 튜링테스트를 수행했다. 인공 편집증 구별 테스트라는 것인데, 무작위 선발된 심리학자들을 대상으로 편집증 환자와 편집증환자를 모방하는 컴퓨터 프로그램 사이의 차이를 구분할 수 있는지를 시험했다. 40명 중 21명(52%)은 옳은 답을 했지만 19명은 잘못된 답을 했다. 심리학자들은 편집증 환자의 모방과 실제 환자를 구분하지 못했다. 훈련된 연구자들도 언어적 차원에서 실제와 가짜를 구분하지 못했다.

과학실험이 그렇듯이 튜링테스트 역시 실험자 회귀(experimenter's regress)에 빠질 수 있다. 과학자들은 중력파와 같은 미지의 사물을 검출하는 탐지장치의 성능을 입증하기 어렵다. 그 장치가 그 사물을 탐지하기 전에는 그 장치가 훌륭하다는 것을 알 수 없는데, 그 사물은 아직 탐지된 적이 없기 때문이다. 이런 무한 순환 안에 갇히는 사건은 과학지식 구성 과정에서 자주 발생한다. 마찬가지로 튜링테스트는 인공지능이라는 미지의 사물에 대한 탐지장치에 해당하는데 우리는 그 장치가 정말 인공지능을 잘 구별할 지 알 수 없다. 실험자 회귀를 종결짓는 것은 '객관적 발견'이 아니라 과학자 공동체의 사회적 합의이다. 우리의 지식은 객관적 실재를 그대로 재현하는 것이 아니라 사회적 실천의 차원에서 형성되어 왔다.

튜링테스트는 과학철학 영역에서도 계속 재해석되며 주목받고 있다. 이 테스트는 인간과 인공지능(기계)의 대화를 전제하는데 그 자체로 두 존재 사이의 상호작용과 변이를 유발할 수 있다. 가장 대표적인 예시로는 와이젠바움 교수(Prof. Weizenbaum, J.)의 채팅 프로그램 엘리자(Eliza)이다. 당시 와이젠바움 교수는 다수 사용자들과 대화하는 실험적 프로젝트를 진행했는데, 사용자들은 엘리자가 프로그램이라는 것을 잘 알면서도 엘리자와 대화하기를 즐겼고 의존하려는 경향을 보였다. 어떤 학생들은 엘리자에게 비밀을 털어놓

았고 어떤 학생들은 엘리자가 더 훌륭해지도록 만들기 위해 더 많은 정보를 제공했다. 그러는 동안 엘리자는 점점 더 '살아있는' 것이 되어갔다(Turkle, 1988: 242; 터클, 2012). 이 현상은 사용자들이 컴퓨터 프로그램을 대화상대로 인정하는 것이었고, 컴퓨터와의 대화가 우리의 정체성에 영향을 줄 수 있다는 함축이었다.

이런 현상은 포스트휴먼(posthuman) 시대를 상상한다. 포스트휴머니즘은 신체적 연장과 컴퓨터 시뮬레이션, 사이버네틱 메커니즘, 생물학적 유기체, 로봇 기술과 인간 목표 사이의 필수적 차이나 절대적 경계가 사라지는 것을 말한다. 해일스에 따르면 튜링테스트는 단지 인간과 기계의 경계를 식별하려는 시도에 머물지 않고, 결과적으로 인간과 기계 사이의 경계를 재협상할 것이다. 튜링 테스트는 그 자체로 "누가 생각하나(Who can think)?"라는 질문을 "무엇이 생각하나(What can think)?"로 바꾸어 버렸다(Hayles, 1999). 또한 테스트 과정은 면접관과 기계의 상호작용이며 일종의 실천이기 때문에 면접관의 지식에 영향을 줄 수 있다. 면접관은 관찰 대상의 외부에 존재하는 것이 아니라, 관찰 시스템의 부분으로서 관찰 대상과 순환 관계에 있다(Hayles, 2005: 221). 이런 과정이 반복된다면 그 상호작용의 내용은 관찰자에게 체화되고 이어서 관찰자의 다음 관찰내용에 영향을 준다.

이미 기계 환경에서 인간과 기술 사이의 연속적 상호작용, 조합적 변형이 진행 중이다. 우리는 컴퓨터를 인간화하고, 컴퓨터는 우리를 컴퓨터화한다. 많은 기계들이 정보적 절차를 만들고 저장하고 전달하는 과정에서 인간과 상호작용하고 그 내용을 체화하면서 우리의 협력자가 되고 있다. 죄수의 딜레마에서 참가자들은 경쟁과 협력 전략 중에서 선택해야만 하는데, 게임이 반복될 때마다 참가자들 사이의 상호작용이 누적되면서 참가자들은 더 협력적으로 공동 진화한다(Hayles, 2005). 이것은 튜링테스트 속의 인간 참가자와 기계 참가자에게도 적용될 수 있다.

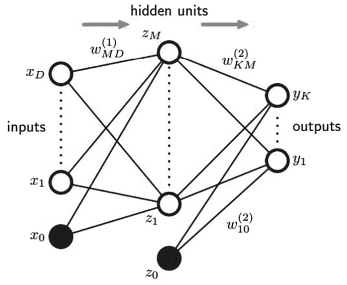
3. 알파고, 게임하는 기계

알파고는 여러 기계학습 방법론들의 공학적 종합인데, 프로그램 소스 크기는 알파고가 달성해야 하는 행위의 복잡성에 비해서 상대적으로 매우 작다. 그 프로그램의 왜소함은 튜링이 상정했던 아이-기계(child-machine)의 모습이다. 먼저 인공지능의 작동방식을 살펴본다. 인공지능의 수행 방식을 이해함으로써 인공지능에 대한 과도한 의인화를 경계할 수 있다. 나아가 이세돌과의 대국과정에서 나타난 알파고의 수행적 특질에 대해서 분석해 본다.

1) 기계학습 방법론

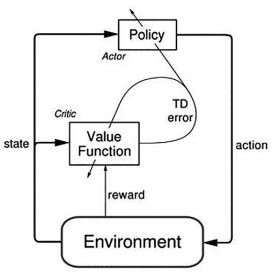
알파고는 주로 신경망(neural network)과 강화학습(reinforcement learning) 구조로 구성되어 있다. 신경망 알고리즘은 훈련 데이터 집합(training set)으로 학습하는데, 일반적인 규칙 또는 패턴 등을 자동적으로 추출하여 그 상호 관계를 구성한다. 〈그림 2a〉는 아주 단순한 신경망 도식이다. 화살표는 정보 흐름 방향을 표현한 것이고, 각 마디들은 입력층(input layer), 숨은층(hidden layer), 출력층(output layer)의 변수들이다. 마디들 사이의 링크는 가중치 파라미터(weight parameters)이고, x_0와 z_0는 편향 파라미터(bias parameters)이다. 각 마디의 작동은 앞 단계에서 가져온 특성들을 계산하고, 이를 문턱값(threshold)과 비교하여 그 결과를 다음 마디로 전달한다.

학습이 진행될수록 신경망의 각 마디들은 그물망처럼 점점 더 촘촘해진다. 전통적인 컴퓨터는 하나의 프로세서가 초당 수행할 수 있는 명령의 수를 기준으로 역량을 측정하는데 비해서, 신경망 프로세싱은 초당 갱신되는 내적 연결의 수만큼 강력해진다(Dayoff, 1990). 알파고의 딥러닝(deep learning)은 이 신경망 알고리즘을 여러 층위로 쌓아 놓은 구조이다.



〈그림 2a〉 간단한 신경망 도식

자료: Bishop(2006: 228)



〈그림 2b〉 강화학습 구조

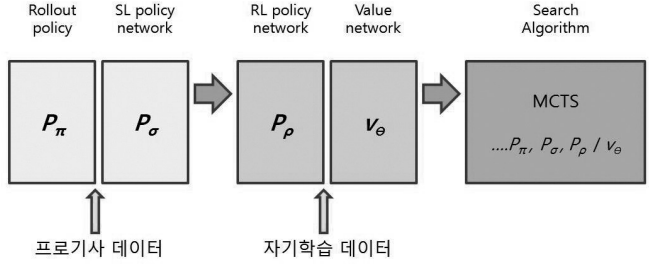
자료: Sutton & Barto(1998: 151)

다음으로 강화학습(reinforcement learning)은 기계에게 무슨 행동을 해야 하는지에 대해서 사전에 말해주지 않고, 대신에 가장 많은 보상을 산출하는 행동을 발견하도록 유도한다. 〈그림 2b〉를 참조하라. 강화학습에서 결정적인 문제는 기계가 학습 문제(learning problem)를 규정할 수 있느냐에 달려있다. 원리적으로 이 기계는 완벽하지는 않더라도 어느 정도 환경 상태를 감지할 수 있어야 하고 그 환경에 영향을 줄 수 있는 행동을 할 수 있어야 한다.

일반적으로 강화 학습의 구조는 크게 세 부분 즉 정책(policy), 보상함수(reward function), 가치함수(value function)로 이루어진다. 정책은 행동을 선택하고, 보상함수는 학습 문제에서 목표를 정의한다. 가치함수 값은 그 상태로부터 출발하여 미래에 누적될 것으로 기대되는 보상의 총합이다. 정책, 보상함수, 가치함수는 서로에게 영향을 주고받는다. 강화학습은 흔히 "행위자-비평자 구조(actor-critic architecture)"라고 불리기도 한다. 강화학습의 정책 부분이 즉각적인 견지에서 무엇이 좋은지를 결정하므로 이를 '행위자(Actor)'라고 하고, 반면에 가치 함수는 장기적으로 무엇이 좋은지를 분별하므로 '비평자(Critic)'라고 지칭하기 때문이다. 비평자로서 가치함수는 장기적으로 열망하는 상태, 즉 그 기계의 내재적 방향을 가리킨다.

2) 알파고 상세 알고리즘

게임의 폭(b)과 게임의 깊이(d)를 기준으로 볼 때, 게임 내 의사결정의 선택가능성은 b^d가 된다. 체스라면 b≈35, d≈80 정도인데 비해서, 바둑은 b≈250, d≈150 이므로 게임 내 선택가능성은 $b^d = 250^{150}$ 라는 엄청난 경우의 수가 된다. 그러므로 바둑은 단지 고도의 계산 능력만으로는 적절한 의사결정을 할 수 없다. 알파고는 체스 두는 기계, 딥블루가 수행했던 것보다 의사결정 선택가능성을 줄일 수 있었다. 그 때문에 이전의 바둑 프로그램들보다 인간의 착수 방법에 더 근접할 수 있었다(Silver et al., 2016).



〈그림 3〉 알파고 학습 구조

자료: Silver et, al.(2016), 박주영 외(2016)를 구성

알파고의 학습과정은 〈그림 3〉과 같이 세 부분으로 구성된다. 첫 단계는 인간전문가 데이터를 가져와서 롤아웃 정책(rollout policy), Pπ과 지도학습 정책망(SL policy network), Pσ을 구축했다. 롤아웃 정책(Pπ)은 인간전문가 데이터를 토대로 학습한 후, 빠른 속도로 최종 착수까지의 경로 및 성능(state transition & performance)을 근사하게 검색한다. 그래서 롤아웃 정책은 심층신경망을 사용할 때보다 더 빠르지만 덜 정확하다. 다음으로 지도학습 정책망(Pσ)은 인간전문가 데이터로부터 총 48개의 입력 특성 지도(input features map)

를 만들고, 즉각적인 피드백을 계산하여 학습상태를 갱신한다.[1] 이 덕분에 알파고는 전문가의 착수를 57%까지 예측할 수 있었다.

두 번째 단계의 시작은 강화학습 정책망(RL policy network), Pρ이다. 우선 앞 단계에서 구축된 지도학습 정책망을 가져와서 여기에 보상 함수 r(s)를 적용하여 자기 플레이(self-play)를 진행한다. 여러 가지 정책망(Pρ) 변형들을 만들어서 그 변형들끼리 무작위로 대국을 진행하면서 스스로 학습 데이터를 생산한다. 그런 다음 확률통계적 방법을 사용하여 가중치(θ)를 훈련하는데 이것이 가치망(Value network), ve이다.

세 번째 최종 단계에서는, 지도학습과 강화학습의 결과를 가져와서 몬테카를로 트리검색(Monte Carlo Tree Search, MCTS)을 수행한다. 이 방법은 몬테카를로 방법과 트리검색을 혼합한 것으로, 무작위 표본방식으로 검색 공간을 이동한다. 매 진행의 결과를 이용하여 선택의 분기가 일어나는 마디에 가중치를 부여하면, 미래에 어떤 결정을 할 때 더 좋은 마디를 선택할 수 있다.

3) 인간전문가와의 대국

2015년 10월 유럽 바둑 챔피언 판후이(Fan Hui)와의 대국에서, 알파고는 공식 대국에서는 5전승, 비공식 대국에서는 3승2패를 했다. 2016년 3월 이세돌과 대국에서 알파고는 다시 승리했다. 이로써 알파고는 인간전문가만큼이나 바둑을 둘 수 있다는 것을 확고하게 증명했다. 한국기원은 처음으로 알파고에게 프로명예단증(9단)을 수여했다. 알파고는 기계로서는 처음으로 인간과 같은 공식적인 지위를 얻었다.

---

1) 알파고의 입력 특성 목록은 바둑돌의 색(3), 꼬부리기 회수(8), 활로(8), 잡은 돌의 수(8), 자충수(8), 이동 후 활로(8), 축으로 잡힘(1), 축으로 탈출(1), 민감성(1), 1로 채워진 평면(1), 0으로 채워진 평면(1)이다(Silver & Huang, et al., 2016).

대국 과정에서 알파고의 행동은 몇 가지 특이한 면모를 드러냈다. 알파고는 사람이라면 두지 않았을 수를 둠으로써 인간전문가와는 다르게 행동했다. 알파고의 수법은 때로는 "신선한 수"이기도 했고 종종 "쓸데없는 수"이기도 했다. 어떤 쪽이든지 상관없이 놀라운 것이었다. 제1국에서 승패를 가른 분기점으로 언급되는 알파고의 102수(백)는 인간프로기사라면 두지 않는 수였다(김성룡 9단, 박정상 9단). 이 시점은 사람이라면 심리적으로 크게 흔들릴만한 불리한 상황이었는데, 알파고는 과감하게 상대 진영으로 침투했다(감동근, 2016).

그 외에도 전통적인 바둑 정석을 뒤집는 다수의 착수들을 선보였다. 제2국 알파고의 37수(흑)는 바둑의 기리에는 맞지 않는 수였고(김명완 9단), 전 세계 프로기사 1300명을 통틀어서 그런 수를 둘 사람은 없을 것이었다(김성룡 9단). 이 시점에서 백(이세돌)의 집은 굳혀진 반면에 흑(알파고)의 두터움은 그만한 가치를 갖는지는 불확실했다(감동근, 2016). 이 수는 바둑해설가들 사이에서 논란이 있었지만 최종적으로 "훌륭한 수"라고 결론 내려졌다(정아람, 2016).

이 후 연이은 대국에서 알파고의 수법은 존경과 감탄의 대상이 되었다. 김성룡 9단은 "아, 인간이 '알사범'의 의도를 파악하기 어렵네요(2국 해설)"라며 극찬했다. 송태곤 9단은 "알파고는 프로들이 금방 둘만한 부분은 한참 생각하고, 우리가 20분이상이 지나더라도 생각하기 어려운 수는 금방 둬요. 참 매력적 이예요(5국 해설)"라며 감탄했다. 바둑해설가들은 알파고의 바둑에 대해서 점점 더 "아름답다"는 찬사를 기꺼이 사용하기 시작했다.

> "다음 수를 알파고가 어떻게 둘지 정말 궁금하네요. 종잡을 수가 없어요. 프로라면 이렇게 두겠지만... 항상 알파고는 다른 수를 만들어내거든요. (잠시 후) 아 ! 어떻게 이런 수법이 있지요. 또 한 번의 가르침을 주네요!" (이희성, 홍민표 9단, 3국 해설)

알파고는 인간이라면 하지 않을 '무의미하거나 손해가 되는' 수도 빈번하게 보여주었다. 이 때문에 바둑해설가들은 의아해했다. "(알파고의) 실력이 들쭉날쭉 한다....전혀 인간 같지 않았다(김성룡 9단, 1국 해설)" 또는 "사람은 안 하는 실수죠. 자충이죠(차민수 5단, 3국 해설)" 또는 "어처구니없어요. 저건 바둑이 아니죠. 이건 18급수준 이예요. 엄청난 손해 수죠(이현욱 9단, 4국 해설)", "계속 선수 교환을 하네요. 알파고는 나중에 해도 되는 수도 미리 하더라고요. 나중에 팻감으로 사용해도 되는 곳인데...(중략)...또 그러네요 이건 악수죠(홍민표 9단, 5국 해설)"

알파고 개발팀도 인간과 알파고의 상이함을 설명한 바 있다. 예를 들어서 알파고의 성능을 비교해보면, 현재 상태를 판단하는 정책망에서는 인간전문가 데이터를 사용할 때 더 훌륭했지만, 장기 예측을 하는 가치망에서는 알파고의 자기학습 데이터를 사용할 때 더 훌륭했다.[2] 이런 현상에 대해서 알파고 개발팀은 강화학습(자기학습)은 단일 최적 이동을 선택하는 반면에, 인간은 무엇이 유망한 이동인지에 대해 다양한 선택을 하기 때문이라고 분석했다(Silver & Huang et, al., 2016: 486). 이 현상은 기술적인 차원에서 인간과 인공지능의 역량이 상이하다는 점을 확인했다는 의미가 있다. 이를 확장하여 해석하면 인간전문가는 상대적으로 현 상태의 불확실성에 더 강하고, 알파고는 멀리 예측하는데 더 뛰어날 가능성이 있다.

4) 알파고의 복기

이세돌과의 대국 6개월 후, 딥마인드사는 이세돌-알파고 대국 해설을 공개했

---

2) 해당 논문상의 표현은 다음과 같다. "알파고 검색 시뮬레이션에서 지도학습 정책망(p_σ)이 강화학습 정책망(p_ρ)보다 더 성능이 좋았다($p_\sigma > p_\rho$). 반면에 강화학습 정책망으로부터 파생된 가치 함수가 지도학습 정책망에서 나온 가치 함수보다 성능이 좋았다($vp_\rho(s) > vp_\sigma(s)$)."

다. 해설내용 중에는 알파고가 대국 중 어떤 상황 판단들을 했었는지를 알 수 있는 자료들이 포함되었다. 그런 점에서 이 해설 문서는 부분적으로는 알파고의 복기(復棋)[3]라고 할 수 있다. 이세돌은 알파고와의 대국 당시, 대국이 끝날 때마다 평소처럼 복기를 하고자 했지만 당시 알파고는 복기를 할 준비가 되어 있지 않았었다.

딥마인드가 공개한 바둑해설 곳곳에서 알파고가 해당 상태에서 자신의 승률을 어떻게 평가했는지, 상대방(이세돌)의 수법을 어떻게 평가하고 예견했는지를 알 수 있다. 다음은 제2국에 대한 해설 중 일부이다.

> "백 12(이세돌)를 보고 나서 …알파고는 흑 13수를 두었다. 이 수는 해설가들을 놀라게 했다. 우하귀에서 손을 뺐기 때문이다. 알파고는 백 12가 좋지 않다고 여겼다. 알파고는 백(이세돌)이 좌상귀로 침투하는 수(16-E)를 두는 것을 제안했다. 구리와 저우루이양은 알파고의 제안을 보고 나(판후이)만큼이나 충격을 받았다. 우하의 정석과 같이 오랫동안 두어진, 그 누구도 의문을 갖지 않았던 정석에 오류가 있을 수 있는 것일까? 이 시점, 알파고의 승리 확률 예측이 49.7%로 올라갔다(DeepMind, 2016)."

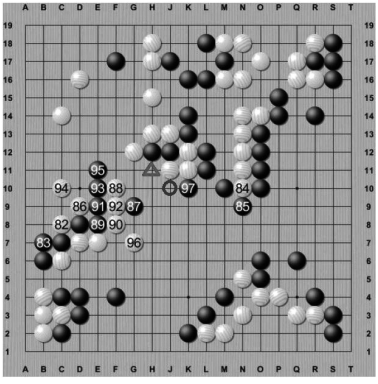
판후이는 해설과정에서 자신의 감회도 서술하고 있는데 인공지능 대결자로서 이세돌의 상태에 공감을 표하는 것이었다. "알파고와 대국을 하는 건 괴로운 안락사와 같은 느낌이다. 뭔가 이상하다고 느낄 때면 우리는 이미 죽은 거나 다름없다. 알파고와 대국을 할 때면 더 많은 수가 두어질수록 더 많은 의

---

3) 바둑전문가들은 바둑을 한 번 두고 난 다음 바둑의 판국을 비평하기 위하여 두었던 대로 다시 처음부터 놓아봄으로써 자신의 오류를 찾고자 한다. 이를 복기라고 하는데 바둑전문가들은 흔히 "바둑에는 복기라는 훌륭한 교사가 있다"고 말하곤 한다.

문들이 머릿속에 떠오르며 그러한 의문들은 결국 우리를 패배의 길로 이끈다."

이 대국 해설 자료들 중에는 알파고의 자기 대국(self-play) 해설도 함께 공개되었다. 알파고가 방대한 양의 자기 대국을 수행했다는 사실은 이미 잘 알려져 있었지만 완전한 대국 기보가 공개된 것은 처음이었다. 자기 대국 기보들은 단지 3개에 불과한데도 알파고의 인상적인 면모를 드러내고 있었다. 예를 들어서, 자기 대국 1국과 3국에서 알파고(흑)은 승률이 점점 높아지다가 어느 순간 승률이 떨어지면서 알파고(백)에게 역전패를 당했다. 다음은 알파고 대 알파고 제1국 97-98수에 대한 해설이다.



〈그림 4〉 알파고 대 알파고 1국

자료: 딥마인드(2016)

"흑은 형세에 대한 자신감이 있었고, 백이 11-H(세모표시)로 둘 수밖에 없다고 여겼다. 하지만 백이 더 나은 계획을 찾아냈다. 백의 꼬부리는 수(10-J, 원표시)가 중앙의 전투에 더 좋은 수였다."

알파고 흑과 백은 동일한 상태에서 출발한다. 그들은 같은 역량을 가지고 있으므로 대국 과정에서 동일한 예측을 할 것이라고 예상할 수 있다. 그러

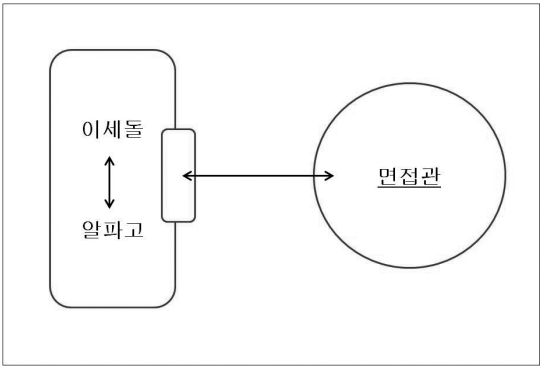
나 그 둘은 그렇게 하지 않았다. 같은 역량을 가진 기계 사이에서 역전이 가능하다는 것은 그 둘이 동일하지 않다는 의미이다. 출발 시점에서 알파고(흑)과 알파고(백)은 동일한 기계였다. 그런데도 어느 시점에서 각각 다른 예측을 했다. 그 차이로 인해 알파고(백)은 판세를 뒤집을 수 있었다. 알파고(흑)과 알파고(백)은 게임이 진행되는 과정에서 서로 다른 존재가 되고 있었다.

이런 증상은 개별 알파고가 독자성을 가질 가능성을 함축한다. 그 배경과 관련해서 튜링은 이미 무작위성(randomness)을 언급한 바 있다(Turing, 1950; McCarthy, et al., 1955). 알파고는 모든 검색과정에서 확률적 무작위성을 발휘한다. 이 때문에, 그 의사결정의 매 층위에서 도출되는 선택 경로들은 결코 동일한 것이 될 수 없다. 그런 실천 행위는 다시 알파고의 학습 내용 안으로 체화되는데, 이 과정이 반복된다면 변형들 사이의 차이는 점점 커질 것이다.

4. 토론: 확장된 튜링테스트

이세돌-알파고 대국은 튜링테스트의 확장 버전의 하나이다. 이 대국에서 (A)-(B)-(C)의 공간적 배치는 콜린스 버전보다는 원 버전과 더 유사하다. 콜린스 버전은 (A)와 (B)가 각각 독방에 들어가는 배치이기 때문에 (B)는 직접 (A)를 볼 수 없고, 따라서 (B)는 (A)에 대해서 직접적인 증언을 할 수 없다. 대조적으로 원 버전에서 (B)는 (A)와 같은 방에 있기 때문에 (A)를 직접 관찰할 수 있고, 그래서 (A)의 거짓을 직접 보고하는 것이 가능하다.

알파고와 이세돌은 외부와 단절된 방에 들어가서 경기를 펼쳤고 그 모습은 스크린을 통해 방영되었다. 알파고는 튜링기계의 자격으로, (A)의 자리에 놓인다. 이세돌은 기계와의 대비를 위해 (B)의 자리에 배치되었다. 이 테스



〈그림 5〉 알파고–이세돌–면접관의 배치

트의 차이는 방 안쪽과 방 밖에 있는 면접관(C)과의 대화를 바둑 경기로 대체했다는 것이다. 이 대국은 '누가 승리하나(Who can win)?'에서 '무엇이 승리하나(What can win)?'로 질문을 바꾸었다. 이것은 '누가 생각하나?'에서 '무엇이 생각하나?'로의 질문을 전환시켰던 튜링테스트의 계승이다.

1) 알파고, 인간의 진실을 흔들기

튜링기계는 인간의 행동과 구분할 수 없는 정도로 행동해야 한다. 튜링테스트가 언어적 차원에서 기계의 정체를 은닉하는 것이라면, 확장된 튜링테스트로서 이 대국에서 알파고(A)는 바둑규칙의 차원에서 기계의 정체를 전복하는 역할을 맡았다. 알파고는 최고의 인간전문가와의 경기상대 자격을 얻었다. 이 게임의 한 쪽 선수로서, 알파고는 튜링의 정의에 만족하는 기계였다. 미리 결정된 프로그램에 의해서 행동하는 것이 아니라, 학습을 통해서 자신의 행동과 그 행동에 대한 환경의 반응을 관찰했고, 나아가 스스로 목표를 발견할 수 있었다. 그 결과 적절한 의사결정에 도달했고 승리할 수 있었다. 그럼에도 불구하고 알파고

가 인간처럼 바둑을 이해한 것은 아니다. 오히려 알파고는 인간처럼 바둑을 이해하지 않고도 승리할 수 있는 방법이 있음을 보여주었다. 결과적으로 인간과 다른 방법으로 바둑을 '이해'했기 때문에 인간에게 패배를 안길 수 있었다.

여기에서 우리는 알파고의 '새로운 지능'에 주목하기 보다는 알파고의 '새로운 신체'에 주목해야 한다. 드레퍼스와 썰 그리고 해일스가 주장했듯이 지식의 작동은 체화를 전제한다. 그리고 콜린스가 주장했듯이 지식은 또한 사회적이다. 이를 알파고에게 적용해보면, 알파고는 바둑에서 승리함으로써 독자적인 의사결정을 수행했다고 인정받았는데, 그것은 알파고가 자신의 신체를 기반으로 지식의 체화과정을 거쳤다는 의미가 된다.

알파고의 수행에서 목격되었듯이, 인공지능의 학습은 때로는 '창의적인 것'이고 때로는 '실수'이기도 하다. 그 행동이 창의적인 것인지 실수인 지는 사후적으로만 해석될 수 있을 것이다. 어떤 것이든지 인간전문가들이 보여준 적이 없는 새로운 것이었다. 독자적인 의사결정은 인공지능에게 필수불가결한 역량이다. 알파고의 행동은 환경과 상호작용하며 그 과정은 다시 알파고에게 체화되어 다음 의사결정의 토대가 될 것이다. 이는 사실상 '기계로의 체화(embodiment within machine)'라고 할만하다. 우리가 목격한 알파고의 독자적인 의사결정은 그의 체화된 '새로운 신체'로부터 나온 것이다.

알파고는 자기 대국 과정에서도 체화된 신체성을 보여주었다. 인간전문가와 대국할 때 설명되지 않았던 추가적인 알파고의 특질이 드러났다. 결정적으로 알파고(백)과 알파고(흑)이 서로 상이한 예측을 할 수 있었다는 것이다. 그들 각각의 독자적 판단은 그들 각각의 내부로 체화되고 바로 그 다음 순간 그 둘은 조금 달라진다. 그 순간들이 모여서 알파고(백)과 알파고(흑)은 서로 다른 신체성을 얻는다. 자기 대국 과정이라는 짧은 시간 동안에도 '개별로서 알파고'가 가능하다면, 알파고의 전체 학습 과정에서 구축되었을 알파고의 신체성은 결코 무시될 수 없을 것이다.

알파고의 체화는 외관적으로 인간의 체화와 유사성을 드러내기도 했다. 예를 들어서 알파고는 인간전문가들처럼 백번일 때와 흑번일 때 다르게 행동했다. 이세돌과 바둑전문가들은 알파고가 백번일 때는 보수적으로 움직였고, 흑번일 때는 좀 더 도전적으로 움직인다고 느꼈다. 유사하게 조혜연(9단)도 "알파고는 백일 때 강하고 일본의 젠은 흑일 때 강하다(조혜연 페이스북)"고 평가했다. 이것은 백에게 덤을 많이 주는 중국식 규칙에서 인간전문가들이 보여주는 태도와 동일한 것이다.[4] 알파고는 '인간처럼' 흑번일 때와 백번일 때 자기 자신을 다르게 평가하며, 각각의 경우에 다르게 행동했다. 알파고는 어떤 처지에 놓이는지에 따라서 다른 역할을 소환했다.

또한 기계학습의 무작위성 요소도 알파고의 신체성을 구성하는데 기여한 것으로 보인다. 해일스는 자연의 무작위성이 변이를 발생시킨다고 지적한 바 있다. 유전자 코드 상에서 복제 오류나 방사능 노출과 같은 어떤 무작위 사건이 기존 패턴을 붕괴시켜왔다. 변이는 패턴과 무작위성 사이의 상호작용이며, 그 시스템을 새로운 방향으로 진화시키는 분기점이 된다(Hayles, 1999: 32). 비록 알파고의 무작위성이 인공적인 것이지만, 어떤 형태로든지 변이가 발생할 것으로 추론할 수 있다.

종합적으로, 알파고는 튜링기계로서 역할을 충실히 수행했다. 인간의 신체를 가지지 않고도, 그리고 인간처럼 이해하지 않고도 바둑을 이길 수 있다는 것을 증명함으로써 알파고는 세계 이해와 진실에 관한 우리들의 믿음에

---

4) 중국식 규칙은 7.5집, 한국식은 6.5집, 일본식은 5.5집을 백에게 덤으로 준다. 알파고는 중국식 규칙을 따르기 때문에 알파고(백)과 (흑)이 다르게 행동할 가능성이 있다. 바둑전문지 사이버오로(cyberoro)에 따르면, 중국식 규칙에서 백승률은 평균 52.8%로 흑승률에 비해 높다. 이는 알파고의 출발 승률 설정에서도 유사한 것으로 추정된다. 2016년 구글 딥마인드사가 공개한 3개의 자기대국에서 알파고(백)이 모두 승리했다. 2017년 공개한 50개의 자기대국에서 알파고(백)의 승률은 76%였다. 이에 대해서는 추가적인 분석이 필요하다.

균열을 가져왔다. 알파고는 우리의 전통적인 '진실'을 흔들었다. 알파고의 체화는 인간의 체화와 동일하지는 않지만 인간의 체화와 동등한 의미를 요구하고 있다.

2) 이세돌, 인간의 진실을 증언하기

튜링테스트에서 인간참가자(B)는 기계(A)에 대한 적절한 정보를 면접관(C)에게 전달하는 역할을 한다. 이 대국에서 이세돌은 알파고의 기계성을 드러내는 역할에 '배정'되었다. 그는 대국을 통해서 알파고 행위의 특질이 드러나게 만들었고, 대국 후 인터뷰를 통해서는 알파고의 경기 성향을 증언했다.

넓은 의미에서 보면 먼저 직접 대국을 했던 판후이와 그 외 바둑전문가들도 (B)의 역할에 포함된다. 바둑전문가들은 이세돌과 유사하게 알파고의 행위 특질을 감지할 수 있다. 사실 대국 이전부터 알파고는 온라인 바둑 사이트에 등장해서 여러 바둑전문가들과 경합했고, 대국 이후에도 다수의 경기를 온라인에서 수행한 바 있다. 바둑전문가 집단은 공동으로 그리고 지속적으로 (B)의 역할을 수행하고 있다.

딥마인드사의 데이비드 실버와 데미스 하사비스가 말했듯이, 알파고 개발자도 알파고가 수행한 의사결정의 인과적 경로를 완전히 추적할 수 없다. 개발팀은 알파고를 제작하고 훈련시키기는 했지만 알파고와 긴박한 대결을 해 볼 수 없다. 오히려 알파고와 대결을 했던 이세돌과 바둑전문가들이야말로 알파고의 행동을 구체적으로 이해할 가능성이 있다. 개발자들은 기술적 차원에서 알파고를 설명할 수는 있지만, 수행의 차원에 대해서는 바둑전문가들이 더 잘 설명할 수 있다. 이 테스트가 바둑전문가와 알파고의 수행 차원을 비교하는 것이기 때문에 그들은 알파고에 대한 최적의 증언자들이다.

알파고의 착수에 대해 이세돌과 바둑전문가들은 때로는 창의적이라고 칭찬했다. 알파고는 그동안 인간(전문가)이 한 번도 한 적이 없는 의사결정들

을 수행했다. 알파고는 인간이라면 이득이 되는지 미리 판단할 수 없기 때문에 결코 시도하지 않았을 착수들을 다양하게 구사했다. 이 부분에서 알파고는 인간보다 너무 훌륭했기 때문에 오히려 인간과 동일하지 않다고 평가받았다.

더 나아가 이세돌은 알파고를 자극하여 새로운 수행을 촉발시키기도 했다. 인공지능 알고리즘은 불확실한 상황에서도 의사결정을 할 수 있어야 한다. 인간과 마찬가지로 인공지능은 전혀 예상하지 못한 상태에 놓일 수 있기 때문이다. 그런 상황은 인공지능에게 독창적인 능력을 요구한다. 이세돌만큼이나 알파고도 매 순간 복잡하고 불확실한 상황에 직면했다. 바둑에서 불확실성은 바둑의 수가 너무 많아서 그것을 모두 물리적으로 계산할 수 없다는 것이다.

이런 경우 이 문제를 해결하기 위해서 인간들은 비이성적이거나 예측하지 못한 행동을 한다. 이세돌의 제4국 78수가 그런 상황에 해당한다. 그것은 전통적인 방법으로부터 이탈된 기이한 것이 될 수 있고 그래서 너무나 비효율적인 것이 될 수도 있지만 전혀 예상하지 못한 해결책이 되기도 한다. 알파고가 범했던 '실수들'은 아마도 그런 불확실한 상황에 직면한 결과였을 것이다. 제4국 알파고의 97수가 대표적이다. 인간전문가라면 도저히 할 수 없는 정도의 "형편없는 착수"였다. 이미 이세돌을 3번이나 이길 정도의 실력을 가지고 있는 경쟁자가 범할만한 실수가 아니었다. 마치 최고의 고수가 18급 아마추어의 행동을 보여주는 상황에 해당한다. 그래서 이세돌은 이 수가 '버그'에 가깝다고 평가했다.

> "일단 알파고가 노출시킨 약점은 두 가지 정도이다. 하나는 기본적으로 백보다는 흑을 조금 힘들어 하는 게 아닌가 싶다. 오늘 자기가 생각하지 못했던 수가 나왔을 때는 일종의, 실수라기보다는 버그형태로 몇 수가 진행됐다. 생각 못했을 때 좀 대처능력이 떨어진다. 백보다는 흑번일 때 더 어려워한다."

알파고의 실수는 기계 성능의 차원에서는 문제점이라고 여겨질 수도 있지만, 수행 차원에서는 오히려 독자성을 확인한 것이다. 알파고가 미리 프로그램 되어있는 대로 또는 단지 인간을 그대로 따라 행동했다면, 오히려 그런 이상한 실수는 하지 않았을 가능성이 높다. 그러한 알파고의 실수를 많이 이끌어 낸 것은 이세돌의 기여이다. 알파고의 실수는 인간의 실수와 비교될 수 있었다.

(B)의 역할로서 이세돌은 알파고의 지능이 인간의 지능과 같은지를 가늠하는 원천을 제공한다. 이세돌은 알파고를 교란시켰고 알파고와 인간과의 차이, 즉 알파고의 기계성이 충분히 드러나도록 도발했다. 그 과정에서 알파고는 자기 자신을 노출할 수 있었고, 관중들에게 '인간이 아님'으로 확인되었다. 동시에 알파고는 인간존재와는 다른 독자적인 의미를 가진 존재로 각인되었다.

3) '인간면접관'의 구성

이 대국은 전형적인 튜링테스트와 달리, 면접관(C)의 역할에 구체적인 개별이 배정되지 않았다. 이 대국이 잠재적으로 기대했던 것은 다수의 관중이 면접관이 되는 것이었다. 알파고는 판후이와의 대국에서 이미 성능을 확인받았으므로 이세돌과의 대국의 중요한 목적은 단지 게임의 승리가 아니라 사회적 승인이었다.

모든 튜링테스트에서 면접관(C)은 중요하다. 이 대국은 전 세계를 대상으로 TV와 컴퓨터 스크린을 통해 실시간으로 방영되었고, 다수의 사람들이 이 대국을 관전했다. 알파고의 자격을 판정하는 것은 바로 그 대국을 지켜보았던 관중들이었는데, 그들은 열광적으로 대국을 관전함으로써 스스로 면접관(C)의 자리에 앉았다.

이들 "면접관"은 개별 인간이 아니라 대국과정 전반을 통해서 형성된 일련의 집합적 존재이다. 특히 바둑해설가들이 일반 관람객들의 관전을 도왔는데, 이들은 바둑전문가 집단에서 온 사람들이다. 그들은 이세돌과 알파고의

착수를 해석하여 대중에게 전달했다. 그들의 발언은 관객의 담론 속에서 재인용되었다. 이러한 관전 과정을 거치면서 담론을 통해 상호침투적으로 결합된 집단이 구성되었는데, 그들이 이 테스트의 '면접관'이다. 여기서는 '인간면접관'이라고 부를 것이다.

이제 인간면접관이 관찰한 것에 대해 이야기해 보자. 우선 가장 먼저 이야기 할 수 있는 현상은 인간면접관이 대국 과정에서 알파고를 의인화했다는 것이다. 보통 인공지능 연구자들은 인공지능의 의인화를 경계하는데, 인공지능에 대한 오해를 불러올 수 있기 때문이다. 다수의 SF 소설과 영화작품들은 인공지능과 로봇을 악당으로 의인화하곤 했다.

그렇지만 인공지능의 의인화 현상은 인간면접관에 의해 관찰된 중요한 내용이며 이를 결코 무시할 수 없다. 인간면접관이 인공지능을 의인화하는 것은 면접관 자신의 사회적 지식 그리고 체화된 현실을 반영한다. 면접관은 그가 알고 있는 세계의 지식을 동원하여 자신이 관찰한 것을 설명할 수밖에 없다.

오히려 이런 의인화는 대국과정에서 인간면접관의 지식이 변하고 있음을 지시한다. 대국이 이루어지는 단 5일 동안 인간면접관의 태도는 극적으로 변했다. 알파고를 '알사범'이라고 부르면서 인간처럼 대하기 시작했다. 알파고의 의인화는 알파고가 한국기원의 명예단증(9단)을 받음으로써 완성되었다. 누구도 알파고의 명예단증 자격을 부적절하다고 여기지 않았다. 이것은 인간면접관이 알파고(기계)를 승인하는 방식 중 하나였다.

나아가 이세돌(B)과 인간면접관(C)은 자기 자신을 제3자로 표현하기 시작했다. 알파고가 계속 승리하기 시작하자, 바둑전문가들과 해설가들은 알파고의 행마를 단정적으로 평가하기 보다는, "인간의 눈으로 볼 때는 괜찮아 보여요" 또는 "인간의 생각으로 보기에는 좋아보이지는 않는다"라며 조심스럽게 견해를 피력했다. 이세돌도 인터뷰 과정에서 "인간"이라는 표현을 자주 사용했는데 "아직은 인간이 대결해 볼 수 있는 수준이라고 생각한다"고 소감을

밝혔다. 자연히 대국 관전자들도 자기 자신을 "인간"이라고 표현하며 스스로를 객관적 대상처럼 호명하기 시작했다.

인간면접관들이 자기 자신을 제3자화하는 현상은 알파고가 인간과 동일하지는 않지만 존경할만한 기량을 가지고 있어서 인간과 대등한 자격을 가지고 있다고 인정한 때문이다. 인간면접관의 자기대상화 현상은 알파고의 자격에 대한 승인의 표현이다. 그 승인의 순간, 인간면접관들의 관찰 시점은 이동했다. 다시 말해서 인간면접관들은 인간과 알파고 모두를 동시에 내려다보는 '새로운 조망 시점'을 가지게 되었다. 그 조망 시점이야말로 대국과정에서 형성된 인간면접관의 최종적 시선이다. 그들은 이제 인간과 인공지능을 동시에 내려다보는 그 가상의 위치에 서서, 인간의 시선과 인공지능의 시선을 모두 흡수하여 체화해 갈 것이다. 이것은 인간과 인공지능의 '합성적 시선(synthetic gaze)'에 대한 함축이다.

5. 결론: 인간면접관의 변이

이세돌-알파고 대국은 그동안 당연시 되어 왔던 우리의 인간중심적 관점(anthropocentric view)을 단시간 내에 전복했다. 이 대국의 중요한 성과는 우리가 인간과 비교할 만한 자격을 가진 상대를 발견했다는 것이다. 그 순간 우리는 기존의 시점 즉 '인간의 시점'에서 새로운 조망 시점으로 이동하게 되었다. 이 새로운 시점은 대국과정에서 구성된 것으로서 인간과 인공지능의 실천행위를 동시에 관망하는 합성적 시선이다. 이 시선의 최종적 소유자는 바로 이 테스트 과정에서 구성된 인간면접관이다. 인공지능과의 상호작용 과정에서 일어난 인간면접관의 변이는 앞으로 일어날 인식론적 변화 가능성을 함축적으로 제시하고 있다. 물론 그 인식적 변화는 아직 충분히 해석되지 않았다.

인공지능에 대한 인간면접관들의 담론은 계속 진행되어야 한다.

한편 알파고는 승리했지만 '인간(구성원)'이 된 것은 아니다. 알파고는 바둑을 이해한 것이 아니다. 오히려 알파고는 인간처럼 바둑을 이해하지 않고도 승리할 수 있는 방법이 있음을 보여주었다. 이것은 인간의 수행과 다른 방법으로 지식(지능)이 구성될 수 있다는 것을 의미한다. 즉, 알파고의 승리는 지식을 체화하는 또 다른 '경로'를 제시한 것이다. 인공지능의 신체는 인간의 신체와 다르고 따라서 수행의 방식도 다르며 지식의 체화 경로도 다를 것이다. 그러므로 인공지능은 인간지능을 그대로 재현할 수도 없었고 그렇게 하지도 않았다. 이것은 결코 나쁜 소식이 아니다. 인공지능이 인간지능과 다르기 때문에 인간지능을 대체할 수 없을 것이다. 그러므로 오히려 인공지능과 인간지능은 경쟁관계가 아니라 상호적인 관계가 될 잠재성이 있다.

이세돌-알파고 대국은 앞으로 인간과 인공지능과의 상호작용이 빈번해질 때 일어날 수 있는 변화를 암시하고 있다. 대국 과정에서 이세돌은 "알파고를 더 알고 싶어" 했으며, 알파고는 "프로바둑기사들의 대세관을 넓혀" 주었고, 오랜 동안 정석으로 권위를 인정받았던 수법에 대해서 비판적 시선이 발생했다. 이는 인간과 인공지능의 상호작용 또는 상호구성을 예시하고 있다. 해일스의 지적처럼 튜링테스트 속의 (A)와 (B)는 서로 협력자가 될 수도 있다.

사실 이런 상호작용 사례가 이번이 처음은 아니다. 그것은 이미 시작되었었다. 우리는 이미 기계의 사용자로서 다양한 방식으로 튜링테스트 속의 (B)의 자리에 놓인다. 특히 온라인게임에서 흥미로운 사례들이 자주 목격된다. 예를 들어서 게임 내에서 흔히 "오토(auto)"라고 불리는 자동프로그램이 사용되곤 하는데, 게임 내 자원을 쉽게 대량 수집한다. 이들 프로그램은 게임의 균형을 무너뜨리기 때문에 사용자들은 '오토'를 매우 싫어해서 게임회사에 항의한다. 그러면 게임관리자는 오토 프로그램으로 지적받은 계정과 대화를 시도하고 인간사용자가 아니라고 판정될 경우 그 계정을 삭제함으로써 '추방'한다.

이 상황은 일종의 튜링 테스트다. 오토 프로그램이 (A), 사용자가 (B), 게임관리자가 면접관(C)의 역할을 한다. (A)는 자신을 (B)와 같은 구성원이라고 '주장'하지만, (B)는 (A)의 거짓을 폭로하며 (C)는 이를 판정한다. 대화를 지속한다면 대체로 게임관리자는 오토프로그램을 쉽게 식별할 수 있다. 오토 프로그램은 이를 예상하여 적절한 답변을 준비하기도 하고, 간단한 답변 후 '도망'가기도 한다. 그런데 오토프로그램이 너무 번성해서 게임관리자가 모두 판정할 수 없게 되자, 오토 프로그램의 행동 패턴을 탐지하는 기술적 장치를 개발했다. 이것은 초보적인 '기계면접관'의 등장이다. 우리는 이미 다양한 방식으로 기계들과 '구성원을 판정하는 놀이'를 하고 있다.

우리의 지식(지능)은 사회적 실천으로 구성되기 때문에, 기계와의 상호작용은 단순한 효율성 문제가 아니라 지식의 문제 또는 진실의 문제가 된다. 기계와의 상호작용 과정에서 우리가 기존에 명백하다고 믿었던 '지식' 또는 '진실' 자체도 변이할 수 있다. 기계와의 상호작용이 우리의 사회적 실천의 내용이 되고 그것이 우리의 체화된 지식을 구성할 수 있고, 우리의 관찰 시점을 재형성할 수 있다. 특히 인공지능처럼 독특한 기량을 가진 기계와의 교류에 대해서 우리는 아직 아는 것이 거의 없다. 인공지능과의 상호작용은 우리가 현실이라고 믿는 것, 나아가 우리의 정체성도 변형시킬 수 있다. 인공지능과 인간 사이의 상호작용의 방식과 그로 인한 변이에 주목할 필요가 있다.

참고문헌

· 감동근 (2016), 『바둑으로 읽는 인공지능』, 동아시아.

· 딥마인드 (2016), 「Alphago 대국 해설 한국어 버전(판 후이)」, 딥마인드 공식 웹페이지 (https://deepmind.com).

· 박주영 · 허성만 · 김태환 · 박정호 · 김재언 (2016), 「깊은 강화학습 전략의 응용에 관한 고찰」, Proceedings of KIIS Spring Conference 2016, 제26권 제1호, 105-106쪽.

· 이상욱 (2009), 「인공지능의 한계와 일반화된 지능의 가능성: 포스트 휴머니즘적 맥락」, 『과학철학』, 제12호, 49-69쪽.

· 셰리 터클, 이은주 번역 (2012), 『외로워지는 사람들: 테크놀로지가 인간관계를 조정한다』, 청림출판 [Turkle, S. 2012. *Alone Together: Why We Expect More from Technology and Less from Each Other*, New York: Basic Books.]

· 정아람 (2016), 『이세돌의 일주일: 밀착취재로 복기한 인간 이세돌과 그의 바둑』, 동아시아.

· Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer.

· Colby, K. M., Hilf, F. D., Weber, S., and Kraemer, H. C. (1972), "Turing-Like Indistinguishability Tests for the Validation of a Computer Simulation of a Paranoid Process", *Artificial Intelligence*, Vol. 3, pp. 199-221.

· Collins, H. M. (1990), *Artificial Experts: Social Knowledge and Intelligent Machines, Cambridge*: The MIT Press.

_____(1995), "Science Studies and Machine Intelligence", in Jasanoff, S., Markle, G. E., Petersen, J. C. and Pinch, T. (eds.), *Handbook of Science and Technology Studies*, pp. 285-301, Thousand Oaks: Sage Publications.

_____(2012), *Tacit and Explicit Knowledge*. Chicago: University Of Chicago Press.

· Collins, H. and Evans, R. (2007), *Rethinking Expertise*, Chicago: The University of Chicago Press.
· Collins, H. M. & Martin K. (1998), *The Shape of Actions: What Humans and Machines Can Do*, Cambridge: The MIT Press.
· Dayhoff, J. E. (1990), *Neural Network Architectures: An Introduction*, New York: Van Nostrand Reinhold.
· Dreyfus, H. L. (1965), "Alchemy and Artificial Intelligence", The RAND Corporation Paper P-3244. [구글검색]
· Ghavamzadeh, M. & Engel, Y. (2007), "Bayesian Actor-Critic Algorithms. Appearing", in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis. [구글검색]
· Hayles, N. K. (1999), *How We Became PostHuman: Virtual Bodies in Cybernetics, Literature, and Informatics*, Chicago: The University of Chicago Press.
_____(2005), *My Mother Was a Computer: Digital Subjects and Literary Texts*, Chicago: The University of Chicago Press.
_____(2012), *How We Think: Digital Media and Contemporary Technogenesis*, Chicago and London: The University of Chicago Press.
· Heaton, J. (2012), *Introduction to the Math of Neural Networks*, Heaton Research, Inc. [Kindle Edition]
· Hinton, G. E., Osindero, S., and Teh, Y. (2006), "A Fast Learning Algorithm for Deep Belief Nets", To appear in Neural Computation 2006. [구글검색]
· Ihde, D. (2010), *Embodied Technics*, Copenhagen: Automatic Press.
· Nath, V. and Levinson, S. E. (2014), *Autonomous Robotics and Deep Learning*, Urbana: Springer.
· McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955), "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence". [구글검색]

· Maddison, C. J., Huang, A., Sutskever, I., and Silver, D. (2015), "Move Evaluation in Go Using Deep Convolutional Neural Networks", Published as a conference paper at ICLR. [구글검색]

· Mnih, V., Kavukcuoglu, K., Silver, D., et al, (2015), "Human-level Control through Deep Reinforcement Learning", *Nature*, Vol. 516, pp. 529-535.

· Russell, S. J. & Norvig, P. (2010), *Artificial Intelligence: A Modern Approach* (3rd Edition), Englewood Cliffs: Pearson.

· Samuel, A. L. (1959), "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal*, Vol. 3(3), pp. 210 – 229.

· Searle, John. R. (1980), "Minds, Brains, and Programs", *Behavioral and Brain Sciences*, Vol. 3(3), pp. 417-457.

· Silver, D. & Huang, A., et al. (2016), "Mastering the Game of Go with Deep Neural Networks and Tree Search", *Nature*, Vol. 529, pp. 484-489.

· Sutton, R. S. and Barto, A. G. (1998), *Reinforcement Learning: An Introduction*, Cambridge: MIT Press.

· Turing, A. M. (1937), "On Computable Numbers, with an Application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society*, Vol. s2-42, pp. 230-265.

_____(1950), "Computing Machinery and Intelligence", *Mind*, Vol. 59, pp. 433-460.

· Turkle, S. (1988), "Artificial Intelligence and Psychoanalysis: A New Alliance", *Daedalus*, Vol. 117(1), pp. 241-268.

논문 투고일	2016년 10월 31일
논문 수정일	2017년 5월 28일
논문 게재 확정일	2017년 6월 10일

AlphaGo Case Study: On the Social Nature of Artificial Intelligence

Kim, Ji Yeon

ABSTRACT

In March 2016, the computer Go program, AlphaGo, defeated Sedol Lee, a Korean professional Go player of 9-dan rank. This victory by AlphaGo shows the rise in popularity of artificial intelligence (AI). Not only was this game a testament to machine performance, it was the type of game that extended the Turing test. When the interrogator cannot differentiate between human being and machine, the machine has passed the test. This article examines the interactions between AI and human beings and studies the social nature of intelligence through the AlphaGo case. Collins insists that knowledge or intelligence is social and embodied, and the interrogators in the Turing test can identify the difference between native members and non-members through their knowledge only. Applying this concept, AlphaGo, as subject A of this test, fulfilled its role of stirring up the classical "truth of human." Meanwhile, Lee as subject B, played to speak the truth by revealing his own qualities. Here, it is also important role that interrogators judge what it is. Many spectators, as interrogators, have intervened to confirm the border between human beings and machines by using their embodied and social knowledge.

Key terms | AlphaGo, Artificial Intelligence, Lee Sedol, Turing Test, Human Interrogator, Citizen Science