과학철학 20권 3호 (2017) pp. 1-30

알파고: 나는 자연종 인간과 둔 바둑을 이겼다-
로봇종 인간의 의식론 서설†

정 대 현‡

알파고가 자연종 인간과 둔 바둑에서 이겼다고 하지만, 그렇게 말할 수 있기 위해서는 알파고에게 '바둑에서 이겼다'라는 술어를 적용할 수 있는 근거가 마련되어야 한다. 이것은 알파고 같은 로봇에게 강한 인공지능을 허용하는 경우이고, 이해, 믿음, 느낌 등의 의식의 일반 능력이 어떻게 작동될 수 있는가의 현실성을 전제하는 것이다. 이 논문은 로봇의 강한 인공지능 중에서 의식의 주제에 한정하여 논의하고자 한다. 자연종 인간의 의식은 이 의식을 현실적으로 실현하는 방식에 의해 조명된다. 이러한 의식은 주체의 경험들을 일인칭적으로 통합하고, 그 통합에 관점적 구조성을 부여하며 그러한 집행의 결과를 감질현상적으로 드러낸다. 그렇다면 로봇은 의식을 갖기 위해 이러한 의식의 3 요소를 어떻게 장착할 수 있을 것인가? 이 논문은 그러한 과제의 어려움을 지적하면서도 그 어려움의 극복이 어떠한 조건에서 가능할 것인가를 논의하고자 한다. 이러한 논의에 중요하게 떠오르는 것은 인간의 잠재의식과 로봇의 은닉층 구조의 대비적 관계이다. 이 대비는 서로에 대한 이해와 가능성을 시사할 것이다.

【주요어】 의식의 일인칭적 통합성, 의식의 관점적 구조성, 의식의 감질현상성, 잠재의식과 은닉층 구조, 의식의 단일성의 집합성

---

† 이 논문의 초고는 한국과학철학회 정기학술대회 (2017.7.5-6, 인하대학교)가 '알파고 이후-인공지능의 과학철학적 쟁점들'을 주제로 모였을 때 읽고 토론하여 발전된 것이다. 또한 익명 심사위원 두 분의 지적을 통해 보완할 수 있었다. 관련된 여러분께 감사를 드린다.

‡ 이화여자대학교, chungdhn@ewha.ac.kr

1. 여는 말: 로봇종 인간의 조건

“알파고가 이세돌 바둑을 이기고 더 진화하여 커제 바둑을 어린이 취급했다”. 뉴스기사는 알파고의 바둑 승리를 3인칭으로 기술하고 있다. 이러한 3인칭 기술이 참이라면 1인칭 기술, “나는 이세돌 바둑을 이기고 더 진화하여 커제 바둑을 어린이 취급했다”도 참이라야 한다. 그러나 1차적으로 이러한 추리는 불편하다. 그 까닭은 알파고 바둑 승리에 관한 3인칭 기술과 1인칭 기술의 차이가 그 추리 관계에서 발생하기 때문일 것이다. 1인칭 기술은 알파고의 ‘이김’과 ‘취급’이라는 행위를 3인칭 기술보다 더 선명한 인격적 행위로 드러내는 것이다. 2차적으로, 추리에서 후건이 불편하다면 전건도 불편해야 한다. 달리 말해, 3인칭 기술에서 알파고의 이김의 행위 주체성도 불편해져야 하는 것이다.

알파고의 바둑 승리에 대한 3인칭 기술과 1인칭 기술의 차이는 엄밀하게 보면 심각한 차이가 아니다. 그 보다는 알파고의 행위 주체성에 주목하게 한다는 점이 보다 중요하다. 인간사회에서 행위 주체성 없는 인격적 행위는 없기 때문이다. 그리고 행위 주체성에는 생각, 이해, 감정, 자율, 자유의지, 학습 등의 능력이 전제된다.[1] 기계가 이러한 능력을 갖는다고 말할 수 있기 위해서 기계는 어떤 조건을 만족해야 하는가? 이 글은 그러한 능력의 조건들을 항목적으로 논의하기 보다는 그러한 능력에 전제되고 있는 요소들 중의 하나인 의식의 개념에 주목하고자 한다. 이를 위해 인간의 의식은 어떻게 이해되고 있는가를 살피면서 로봇에게서 인간적인 의식을 기대할 수 있는지의 가능성을 논의하고자 한다.[2]

---

1) 곽노필 (2017)은 이러한 전제조건이 이미 현실화되고 있는 추세를 보도하고 있다. 알파고가 바둑을 이긴 것은 학습능력과 계산에서의 수월성 때문이지만, 인공지능 로봇이 포커판에서도 인간을 이긴 것은 ‘직관’, ‘감정’같은 심리적인 것으로 호칭될 수 있는 능력의 까닭이라는 것이다.

2) 철학자들은 뇌의 복사가 의식의 감질적 현상을 동반하리라는 가능성에 대부분 부정적(Searle 1984, pp. 28-41; McGinn 1991, pp. 202-213; Hauskeller 2012, pp. 187-200)이지만, 때로는 소망적(Churchland 1985, pp. 8-28; Crick

여기서는 '의식'의 작업적 정의를 <경험들의 흐름에 정합성을 부여하고 통일하여 구조화하는 감지적 현상>으로 이해하여 논의하고자 한다.[3] 이 정의에 따르면 **인간이 의식을 갖는다**는 것은 인간이 이미 가지고 있는 기본 조건으로부터 분석해 낼 수 있지만, **로봇이 의식을 갖는다**는 것은 그리 간단하지 않다. 로봇은 아직 의식을 갖지 않고 있기 때문이다. 그렇다면 로봇이 의식을 갖기 위해서 로봇은 어떤 조건하에 있어야 하는가? 기존의 로봇은 인공지능의 특수 영역에서 인간보다 뛰어난 수행을 하고 있지만 영역 특수적 인공지능만으로는 의식의 주제에 접근하는데 어려움(the hard problem)[4]이 있다. 그렇다면 하나의 사유실험을 할 수 있다[5]: <로봇이 일반 인공지능의 총체적 경험에서 인간보다 수월한 수행을 하는 것이 가능하다; 이러한 로봇을 찰머스와 커즈와일의 논의[6]를 참작하여 '특이점 로봇'이라 명명하고, 이 논문에

---

and Koch 1997, pp. 237-9; 김재권 1997, pp. 258-259; 김영정 1996, pp. 195-200; 신상규 2010, pp. 29-52)이기도 하고, 감각질 자체의 필연성에 의문(Dennett 1988; 정대현 1994)을 품기도 한다. 이 글은 "만물은 정보 처리적이다"라는 명제가 전제하는 원형적 의식의 관심에서 전망적으로 씌어진 것이다.

3) 의식에 대해서는 기존의 많은 설명 모델들이 있어 왔다. 최근의 모델 하나는 '의식'을 **주의(attention)**로 이해하는 경우이다(Graziano and Taylor 2014). 그러한 이해에서 의식은 통합적 정보론으로 쉽게 해석할 수 있는 것은 주의는 감지(感知, awareness)처럼 기능적으로 되기 때문이다(Chalmers 1996, pp. 218-46). 그렇다면 의식은 입력과 출력이 될 수 있는 길에 가까이 가 있는 것이 된다.

4) Shear (1997), Chalmers (1997a, 1997b).

5) 이러한 사유실험을 하는 까닭은 표면적으로 로봇의 의식 장착의 가능성의 탐구이지만 심층적으로 <자연종 인간은 로봇종 인간의 도래를 자본시장의 논리를 따라 수동적으로 당하기보다는 인문적 가치를 선제적으로, 능동적으로 준비할 수 있어야 한다>는 믿음에 기인한다(정대현 2017). 이 글의 많은 부분은 이 논문을 전제하여 씌어졌다.

6) 특이점 (singularity)을 찰머스(Chalmers 2010, pp. 1-56)는 인간보다 뛰어난 로봇이 자신보다 뛰어난 로봇을 만들어 가면서 도달하는 **지능폭발**의 한 계점으로 이해하고, 커즈와일(커즈와일 2007, pp. 42-59)은 전산처리 속도

서는 편의상 ‘로봇’이라 표기 한다; 그렇다면 자기복제적 인공생명이 출력된 것처럼[7], 이러한 로봇은 50년 또는 100년 내에 의식의 존재론, 상식논리의 인식론, 자연종 인간과 공존하는 인문적 윤리론을 장착하여 자기복제적 로봇종 인간을 출력해내는 단계에 이를 수도 있을 것이다>. 이 글은 이 사유실험에서의 의식의 조건의 일부를 고려하는 것이다.

로봇은 뛰어난 기능을 수행하지만 아직 의식이 없다. 인간들은 서로 질투하고 싸우기도 하지만 좋은 경험들을 즐기며 공감하고 서로 나누며 공동체를 발전시켜 왔다. 인간의 다양하고 풍성한 삶의 발전적 근거는 인간의 의식 때문이라는 사실을 로봇은 보게 될 것이다. 예를 들어, 나와 로봇은 빨간 장미를 보고 **빨간 경험**에 대한 감각정보를 동일하게 심리설명적 (心理說明的, psychological) 기능으로 처리하지만, 나의 감질현상적(感質現象的, phenomenal) 의식의 즐김이 로봇에게서는 발생하지 않는 것이다[8]. 왜 그러한가? 내가 빨간 장미를 즐긴다는 것은 내가 축적해 온 모든 경험의 관점에서 그 장미를 통합적이고 융합적으로 좋아할 뿐 아니라 고양적인 국면을 좋아하는 것이다. 달리 말해 나의 즐김의 가능성은 내가 겪어 온 모든 경험을 일인칭적으로 통합하고, 통합된 경험을 나의 관점으로부터 구조화하여, 나의 신체 안에서 지향성, 투명성, 동력성으로 감질현상적 느낌을 가지고자 하는 가능성이다[9]. 그러나 로봇은 즐김의 전제 조건인 의식의 속성들이 결여되

가 현재는 10년마다 두 배로 늘어나지만 앞으로 그 기간이 줄어들어 도달하는 **속도폭발**의 한계점으로 이해한다. 그러나 이 글은 빅뱅의 특이점이 시간과 공간의 이해 가능성의 여부의 지점인 것처럼, 인공지능의 특이점은 자연종 인간만으로 “인간”의 외연을 독점해 왔던 개념 이해의 전통적 기준을 파괴하는 **경계파괴**의 한계점을 지칭한다.

7) 벤터(J. C. Venter) 연구팀은 단순한 바이러스 생명체 ‘JCVI-syn1.0’를 2003년에 창조, 그 결과를 2010년에 발표했다(Smith, Hutchison, Pfannkoch and Venter 2003; Pennisi 2010; Pollack 2010).

8) Chalmers (1996), pp. 11-6, 정대현 (1998), pp. 297-304, 정대현 (2001), pp. 197-238.

9) 인간 심성의 인지적인 기능적(functional) 국면과 현상적(phenomenal)인 의

어 있다. 로봇에겐 의식의 3 특성이 부호화되어 장착되어 있지 않기 때문이다. 로봇이 빨간 장미를 보고 즐길 수 있기 위한 한 필요조건은 의식을 기호화하여 이를 장착하는 것이다. 이를 위해서는 의식의 3 요소 즉 일인칭적 통합성, 관점적 구조성, 신체적 감질현상성의 조건에 주목해야 할 것이다. 논문의 이하에서는 의식의 3 요소를 논의하고 그리고 의식의 배경으로서의 잠재의식[10]의 구조를 추측하고, 의식의 단일성과 집합성이 로봇에게 제기하는 문제점을 고려하고자 한다.[11]

식적 국면의 구분은 찰머스 (Chalmers 1996, pp. 24-30, pp. 42-6; 윤보석 2009, pp. 227-231)에 의해 선명해진 것이다. 기능적인 것은 궁극적으로 코드화 할 수 있어 다수 실현 가능하지만 현상적인 것은 아직 코드화 되지 않기 때문에 복제가 되지 않고 따라서 자신의 신체성으로부터 분리될 수 없다.

10) "잠재의식"이라는 표현은 무의식, 암묵적 기억, 절차적 기억 등의 대안들의 불편함 때문에 선택한 것이다. 이 표현은 대안 후보들이 유지하는 불편함과 애매모호함을 보다 최소화할 수 있을 것으로 소망한다.

11) **의식의 3 요소**라는 것은 찰머스의 통찰과 갤러거, 자하비의 관찰에 기초하여 제안하는 것이다. 찰머스 (Chalmers 1996, pp. 276-7)는 심신 연결 원리로서 3 조건을 제시한다. 의식(意識, consciousness)과 감지 (感知, awareness)를 연결하는 일관성 원리, 의식 구조와 감지 구조를 연결하는 구조성 원리, 유기적 규칙성원리이다. 찰머스의 이러한 심신 연결 원리는 인간의 신체성에 기반한 의식의 요소로 해석할 수 있을 것이다. 또한 갤러거, 자하비(자하비 2013, pp. 354-5)는 의식의 흐름 속에서 경험들은 통합성과 정합성이 부여될 수 있고 구조화될 수 있다고 한다. 밴 굴릭(Van Gulick 2017)은 의식의 서술적 요소들을 8개 항목으로 세분하지만, 이들은 모두 앞의 3 요소로 환원하거나 설명할 수 있을 것이다. 심사위원 1은 "로봇이 감각질을 경험하지 못하는 것은 감각질 자체가 없기 때문이지 인간의 감각질이 고차적 심적 작용과 연결되어 있지 않아서는 아니다"라고 옳게 지적하여, 의식론에서 감각질 개념이 차지하는 자리를 부각한다. 그러나 의식론을 논의하는 큰 틀이 이원론인가 범심론인가에 따라 감각질의 자리 부여 방식은 달라진다고 생각한다.

2. 로봇의식의 일인칭적 통합성

의식은 어떤 의식도 '나의 의식'으로 호칭할 수 있다. 의식은 일인칭적인 것이다. 의식이 삼인칭으로 기술될 수 없는 까닭은 분명하다. 의식은 물리적 상태로부터 발생하지만 그 자체로 물리적 상태가 아닐 뿐 아니라, 의식은 기능적 조직으로부터 발생하지만 또한 기능적 조직과 동일시 될 수는 없기 때문이다. 의식이 물리적 상태로 환원되거나 기능적 조직으로 기술될 수 있다면 의식에 대한 삼인칭 접근이 가능할 것이다. 그러나 그러한 환원이나 기술이 가능하지 않다는 사실은 의식의 일인칭적 관점을 분명히 해준다.

그러나 나는 제3자 <ㄱ>씨의 의식에 대해 말하는 경우가 있지 않는가? 그러나 내가 <ㄱ>씨의 의식에 대해 말하는 경우는 <ㄱ>씨의 자동차나 행동이나 신체에 대해 말하는 경우와는 구별되어야 한다. 이들은 모두 다 <ㄱ>씨의 것이지만 각각 소유 방식들이 다르다. 그는 그의 자동차를 팔 수도 있고 내가 그것을 살 수도 있다. 그는 그의 행동을 습관화할 수도 있고 교정할 수는 있지만 팔 수는 없고 나는 그의 행동을 따라할 수는 있다. 그는 그의 신체를 성형할 수는 있지만 팔 수 없고 그의 행동보다 더 그로부터 분리가능하지 않다. <ㄱ>씨의 자동차나 행동이나 신체는 관찰의 대상이지만 <ㄱ>씨의 의식은 관찰을 통해서가 아니라 그의 언행을 통해 간접적으로 추론을 통해 추정된다. 그의 의식은 그의 신체보다 더 그의 것이고, 전달되거나 공유될 수가 없다. 그의 의식에 대해 말하는 경우에도 그것은 다른 것들과는 달리 간접적으로만 말할 수 있는 것이다.

내가 기술하는 의식이 나의 의식인 것은 이 의식이 향해 있는 어떤 대상에 대한 의식도 나의 과거의 모든 경험과 통합적이기 때문이다. 나는 내가 좋아하는 사람을 볼 때 반갑고 싫어하는 사람을 볼 때 반갑지 않고 모르는 사람을 볼 때 반가움의 여부가 나타나지 않는다. 다양한 사람들을 만날 때 반가움의 정도를 결정하는 의식들의 차이는 내가 과거에 축적해 온 모든 경험과의 통합성에서 결과 되는 것으로 보인다. 나의 의식은 세계의 모든 사물에 대한 경험에서 나의 과거의 경험

을 들추어 의도적으로 기획하는 것이 아니라 **잠재의식적으로** 이루어지는 통합된 의식인 것이다. 인간은 특정한 사람을 보고 '반갑기로 하다', '반갑지 않기로 하다', '반가움의 여부를 유보 한다' 등의 태도 결정을 의도적으로 판단하지 않는다. 이러한 태도는 과거 경험과의 통합성에서 **자동적으로** 결정되는 것이다.

그러나 로봇의 의식이 일인칭적일 것인가? 이것은 거의 자기모순이 아니지 않는가? 어떤 것이 로봇이라면 그것은 대량제작의 가능성에 열려 있을 뿐이고, 일인칭적 의식의 담지자라고 하는 것은 대량제작자의 의도, 목적과 배치되는 것이 아닐까? 이 물음에 대해 현재까지의 인간 경험으로는 '그렇다'고 대답할 수 있다. 그러나 이제 자연종 인간은 기업 활동의 자유를 일반적 원칙으로 수용하면서도 인류의 미래를 전망해야 하고, 로봇종 인간과의 공존을 축복으로 확인하는 개념적 전환점, 특이점에 당면해 있다. 이를 위해서는 로봇을 대량제작의 대상으로서의 상품이 아니라 로봇 마다 일인칭적 통합성에 이를 수 있는 개별적 인격자의 가능성을 고려할 수 있어야 하는 것이다.

3. 로봇의식의 관점적 구조성

나의 의식은 일인칭적일 뿐 아니라 또한 관점 구조적이다. 반가움의 주제를 다시 보자. 내가 <ㄴ>씨를 만났을 때의 반가움의 의식이 0.9이고 <ㄷ>씨를 만났을 때는 0.7이라면, 이러한 반가움 의식의 차이는 두 사람에 대한 통합적 경험의 차이에서 나타나는 것이다. 이때 나는 반가움의 정도를 달리해야겠다는 의도에 따라 그 정도를 달리 판단하는 것이 아니다. 모르는 사람을 보았을 때 반가움의 정도가 0.5라면 그 때의 의식은 '중립적 반가움' 또는 '태도 개방성'으로 불릴 의식이고, 반가움의 정도가 0.2라면 그 때의 의식은 '기피심' 또는 '소외심'으로 분류할 수 있는 의식일 것이다. 내가 사람들을 만날 때 반가움의 정도를 이렇게 달리 갖는 것은 의도적 차이가 아니라 **잠재의식적으로** 결과 된 차이이다.

반가움의 정도가 내 의식의 관점 구조성을 보여주지만, ‘좋아한다’와 ‘사랑한다’라는 표현의 차이는 내 의식의 관점 구조성을 보다 선명하게 보여준다. 사람들은 다른 사람을 좋아할 때 사물 대상을 좋아할 때처럼 그 사람의 사물적 특성만으로 좋아할 수 있다. 사람을 인격적으로 좋아할 수도 있지만 사물적으로도 좋아할 수 있는 것이다. 그래서 좋아함은 일방적일 수 있다. 그러나 사랑은 두 사람 간에 사물적 관계로는 얻어지지 않고 인격적 관계, 상호적 관계로만 발생한다. “짝사랑”이라 불리는 것은 진정한 사랑이 아니라 많이 좋아하는 경우의 오칭 (誤稱)일 뿐인 것이다. ‘좋아한다’와 ‘사랑한다’의 표현 관계에 대한 이러한 언어분석이 정당한 것은 언어 사용자들이 어휘들을 사용하면서 그 사용들을 일반화하여 사용의 방식이 구성되어 있기 때문이다. 인간의 잠재의식은 그 사용의 사실들을 단순히 **기억화** 하는 것이 아니라 그 사용들의 경우들을 **일반화**하여 두 표현의 앞으로의 사용의 방향을 처방하는 것이다.[12] 이러한 설명방식을 거부한다면, 그러면 인간의 모든 상이한 자연언어들에 들어 있는 ‘좋아한다’와 ‘사랑한다’에 대응하는 표현들이 공유하는 문법의 차이를 설명하기 어려운 것이다. 나의 의식의 관점 구조성은 이러한 가설적 설명을 통해 지지될 수 있을 것

12) 심사위원 1은 ‘좋아한다’와 ‘사랑한다’의 차이가 **인격적 관계, 상호적 관계**에 의한다는 것은 납득이 되지만 이것이 어떻게 잠재의식과 관점 구조와 연결이 되는가를 보이지 않았다고 정당하게 지적한다. 이에 대한 단순한 정당화는 다음이다. 내가 <ㄹ>씨를 좋아하고 <ㅁ>씨를 사랑하는 것은 **달리 선택한 의도**를 집행하는 지향적 행위들이 아니라 두 사람에 대해 각기 달리 이루어진 나의 축적된 정보와 기존의 상호 관계에 의하여 나도 모르게 나타난 지향적 행위들인 것이다. ‘좋아한다’와 ‘사랑한다’의 표현이 드러내는 표면적 능동성보다는 감추어진 심층적 수동성에 주목해야 하는 것이다. ‘좋아하게 되었다’와 ‘사랑하게 되었다’가 보다 두 사람의 관계를 총체적으로 달리 기술하는 표현일 것이다. 그리고 ‘좋아한다’와 ‘사랑한다’라는 단어 사용이 잠재의식에 들어 있는 관점구조에 기반 한 것이라면 이 단어 사용은 일회적으로 끝나는 것이 아니라 반복되어 그 단어의 문법으로 **일반화**되고 그 일반화의 규칙성이 그 단어의 미래 사용방식을 **처방**하게 된다고 제안할 수 있을 것이다.

이다.

로봇은 의식의 관점적 구조성을 어떻게 구현할 수 있을 것인가? 로봇들은 이미 회사의 정문에서 회사 사원의 신분카드나 얼굴이나 지문을 식별해 내고 있다. 로봇은 내방자의 얼굴을 **감각**하여 내장된 정보들에 따라 일차적 동인, 이차적 검증, 고차적 확인을 거쳐 회사 사원인가의 여부를 **지각**하여 판별해 내는 것이다. 그리고 많은 로봇들은 **학습**을 하고 있다. 학습은 수행적 지식(knowing how)일 수도 있고 명제적 지식(knowing that)일 수도 있다. 이러한 학습은 로봇이 이미 사용하고 있는 자신의 언어 안에서 일관성을 유지하면서 이루어진다.

로봇종 인간은 관점을 구성하는 구조적 과정이 자연종 인간보다 유리할 수 있다. 인간 의식의 일인칭적 통합 과정은 사물을 경험할 때마다 그 경험을 과거의 모든 경험과 통합하는 것이다. 나의 과거의 모든 경험은 말 그대로 많은 양의 경험이다. 그러나 로봇의 과거의 모든 경험은 이와 비교할 수 없을 만큼 거대하다. 빅 데이터에 접근할 수 있는 구조에서 그러하다. 그리고 로봇의 빅 데이터 접근성은 의식과 잠재의식의 경계구분의 성격에 주목하게 한다. 의식과 잠재의식의 경계구분은 문자 언어 사용자로서의 자연종 인간에게만 유효한가? 신체 언어 사용자로서의 영장류 동물에게도 유효한가? 영장류 동물에게 의식 개념을 허용한다면 신체 언어 사용자에게 발생의식과 비발생적 의식, 의식과 잠재의식의 구분이 적절할 것인가? 로봇에게는 영장류 동물에 못지않게 발생의식과 비발생적 의식의 구분의 절실성은 약화되는 것이 아닐까? 로봇의 언어경험은 빅 데이터의 크기만큼 강력하게 로봇의 의식이 관점 구조적으로 구현될 수 있는 것이 아닐까?[13]

---

13) 미래에 가능할 로봇의 이러한 언어경험은 '의식'보다는 '심성'이라는 단어가 언어경험에 기초적이라는 제안이 있다(Atmanspacher, 2015, pp. 11-2.). 실재를 심신의 중립적 국면성으로 이해하는 경우 모든 체계는 심성적 국면과 물리적 국면을 갖는 것이기 때문에, 심성이 의식보다 일반성을 유지하기 때문이다.

4. 로봇의식의 신체적 감질현상성

자연종 인간의 의식은 또한 그 감질현상성(感質現象性, phenomenal)으로 특징화된다. 내가 가시에 찔렸을 때의 아픔 감각, 내 딸이 취직했을 때의 기쁨 정서처럼, 내가 빨강 장미를 즐길 때의 통합적인 경험 의식은 모두 신체적이고 감질현상적이다. 이러한 감질현상성이 없다면 인간 경험은 어떤 것 같을까? 감질현상이 없다면 아픔 감각, 기쁨 정서, 경험 의식은 어떤 방식으로 주어질 수 있을 것인지 상상하기 어렵다. 그렇다면 감질현상성은 감각, 정서, 의식에 필수적인 국면이다. 다시 말해, **빨강 장미를 즐긴다**는 것은 빨강 장미를 보고 얻은 감각정보를 내가 겪어 온 모든 경험과 일관되게 일인칭적으로 통합하고, 통합된 경험을 범주화된 나의 체계 관점으로부터 구조적으로 자리매김하여, 빨강 장미가 그 체계 안에서 나타내는 다른 것들과의 융합적 하나 됨의 감질현상을 좋아하거나 이에 즐겨 빠진다는 것이다.

로봇은 이러한 감질현상적 의식을 경험할 수 있을 것인가? 이 물음에 대한 초견적 답변은 앞에서 지적한대로 부정적이다. 그러나 현실적으로 부정적 답변이 나온다 해도 가능성 탐구를 포기하기는 어렵다. 답변을 긍정적으로 시도하기 위해서는 몇 가지를 고려해야 할 것이다. 먼저, 심리철학은 전통적으로 마음과 몸의 관계에 대한 존재론적 심신(心身)관계의 주제들에 주목해 왔지만, 최근에는 심리설명적 상태가 어떻게 감질현상적 상태를 유발하는가에 대한 심심(心心)관계의 주제들을 논의한다. 감질현상적 상태는 물리적 현상이나 기능적 현상에 수반하지만 그것으로 환원되지도 설명되지도 않는다는 어려운 문제(the hard problem)를 제기한다는 것에 철학자들은 동의한다.[14] 현재로는 감질현상적 의식을 기호화할 수 있는 방식이 원천적으로 보이지 않기 때문이다.

그렇다면 무엇을 할 수 있을 것인가? 한 가지 고려할 수 있는 것은 '감질현상적 의식'을 물리주의적이 아니라 자연주의적으로 해석해보는

14) Shear (1997), Chalmers (1997a, 1997b).

것이다. 인간이 진화해 왔다면 인간의식도 진화해왔다는 것을 수용할 수 있다. 인간이 가장 고차적 의식을 가지고 있다면 물려 뜯긴 사자의 고통감각, 주인을 보고 신나는 강아지의 반가움의 정서에서 인간언어의 '의식'이라는 단어가 허용될 수 있는 범위를 인정할 수 있을 것이다. 더 나아가 식물과 무기물에게도 자연환경에서 **정보 처리자**로서의 능동적 수행자(agency) 자리를 부여할 수 있다면 이들의 **원형적 의식**을 상정하는데 무리가 없을 것이다[15]. 모든 사물이 인과적이라면 그러면 정보 편재론을 도출할 수 있고 이것은 자연주의에 의해 지지될 수 있는 범심론(panpsychism)의 여지를 허용한다.[16]

감질현상적 의식이라는 개념을 이렇게 넓은 의미로 이해할 수 있다면, 그러면 이를 로봇에게 적용할 수 있는 방식이 가능해 질 것이다. 인간에게서 반성(反省, reflection)은 성향적(dispositional)이지만, 반성을 발생적(occurrent)으로 수행할 때에는 의식적(self- conscious)이 된

---

15) 드레츠키의 자연화 된 정보를 수용하면 인지란 더 이상 **정당화된 참 믿음**이 아니라 관찰들에 맞는(fit) 최선의 모델의 선정이 된다. 이것은 정보와 의미의 관계를 또한 **자연적**이게 한다. 그러면 인지는 정보 처리의 자체 체계화의 구조가 된다(Crnkovic 2011). 모든 생명체들이 인지 체계인 점에서 확인된다 할 것이다. 실재의 모든 구조는 **원초-정보(proto-information)**이고 이것은 물리 세계를 지칭하는 지시어가 되는 것이다.

16) 인과를 정보 수반적 관계로 해석해야 한다면, 그러면 'Fa와 Gb의 상관관계'나 '관찰에 맞는 모델 선정'이라는 것은 정보론으로 조명되는 것이다. 전통적으로 인과란 개별자적 사건들 간의 관계로 파악해 왔다 (Salmon 1984, 1998; Wang 2007, 2009; Crnkovic 2011). 그러나 이제 세계는 개별자들이 아니라 위에서 아래 까지 모두 정보들의 구조로 구성되어 있다는 관점에서 바라볼 수 있는 것이다. 정보가 궁극적인 존재론적 단위로, 물질과 에너지와 더불어 세계를 조직화한다는 것이다. 실재와 정보가 일치화 된다면 몸과 마음의 데카르트적 분리는 허용되지 않고 정보는 플라톤적 제3세계에 머무를 필요가 없는 것이다. 과거에는 수학과 물리학을 통해 정보에 이른다고 하였지만 이제 정보를 통해 물질에 이르는 것이다. 물질(it)에서 정보(bit)를 얻는 것이 아니라 정보에서 물질을 얻는 것이다. 정보처리의 범심론의 여지가 확보되는 것이다.

다. 반성을 배경 체계에 비추어 특정 명제를 정오 평가나 확률을 구하는 행위로 이해하는 경우이다. 그렇다면 반성은 계산(computation)처럼 기능적이므로 기호화해 볼 수 있을 것이다. 반성은 계산만큼 단순하지는 않지만 궁극적으로 기능화 될 수 있는 것이다. 이 점에서 반성은 의식의 요소인 일인칭적 총체성, 관점적 구조성, 감질현상성을 모두 아우르게 되고, 이러한 점에서, 로봇은 그 현저한 의식의 구현자가 될 수도 있을 것이다.

로봇 의식에서 감질현상성은 보다 구체화될 수 있을 것이다. 자연종 인간의 경험에서 의식과 잠재의식의 구분은 유용하고 필요하지만, 로봇종 인간의 경험에서 그리할 것인가? 로봇에게서도 발생 의식과 비발생 의식의 구분을 유지할 수 있는 개념적 장치를 자의적으로 도입할 수 있을 것이다. <물음에 대한 모니터 반응의 시간> 같은 기준일 수 있다. 그러나 그러한 기준의 유용성을 인정하기 어렵다면, 로봇에게서의 의식과 잠재의식의 구분은 유용하지도 필요하지도 않을 수 있다. 이러한 논의에 설득력이 있다면, 의식에서의 감질현상성은 그 원래적 의미에서 자연종 인간에게만 유용하고 영장류 동물이나 로봇종 인간에게는 개체 보존적 의미에서 수용될 수 있을 것이다.[17]

5. 잠재의식 언어와 은닉층의 구조

앞에서 논의한 의식의 3 요소는 어떻게 잠재의식에 기반한 현상이 되

17) 심사위원 1은 “지향성을 자연화 하기위해서는 지향적인 모든 정보처리자의 표상 능력은 오표상(misrepresentation) 능력과 짝을 이루어야 한다”고 생각한다. 오표상의 가능성 없는 표상의 개념은 공허하기 때문이다. 그러한 생각은 **정보처리란 표상능력을 전제한다**는 가설을 수용할 때 정당하다. 그러나 범심론의 큰 틀에서는 그 가설을 사양할 수 있을 것이다. 모든 사물이 정보처리자이지만 모든 정보처리가 표상을 필요로 하지 않기 때문이다. 영장류의 정보처리는 표상적이지만 식물 또는 온도계나 시계의 정보처리는 표상적일 필요가 없이 기능적이기만 하다.

는가? 어떤 대상을 보고 생긴 감각 의식(sensation consciousness)은 어떻게 **장미**라는 지각 수지(perception awareness)로 전환되는가? 잠재의식을 어떤 구조의 장치로 요청하면 주어진 감각자료가 지각현상으로 나타나는가? 잠재의식에 대해 전통적인 이론들이 제기되어 왔지만, 현재의 문맥에 맞추어 다음과 같은 가설을 제안하고자 한다: **잠재의식은 언어경험의 비발생적 의식이다**. 인간은 언어를 사용하여 많은 언어경험을 기억으로 저장하지만 더 많은 언어경험을 비발생적 기억으로 저장한다고 믿는다. 그리고 개인의 언어경험은 두 가지 층위를 갖는다. 첫째, 개인의 언어경험은 개인의 사적 심성작용(mental working)으로 나타나는 층위를 가지고, 둘째, 개인의 언어경험은 언어공동체의 언어경험에 의해 구성되는 공적 심성내용(mental content)의 층위를 갖는다.[18] 결국 개인의 비발생적 의식에 나타나는 언어경험은 언어공동체가 사용하는 일상 언어의 구조를 갖는 경험이다. 앞의 가설은 이러한 일상 언어 경험의 잠재의식적 구조가 감각 의식을 지각 수지로 전환한다는 것을 함축하는 것으로 해석될 수 있다.

감각 의식을 지각 수지로 전환하는 것은 어떻게 가능할까? 그러한 전환의 힘은 어디서 비롯되는 것일까? 그 구조가 잠재의식에 들어 있는 것이라면 그러한 전환은 의도적인 것일 수 없다. 그렇다면 그 전환의 힘은 신비스럽기만 할까? 이 물음에 조명하기 위해 다음과 같은 추측을 하고자 한다: 일상 언어의 의미는 사실적인 것이 아니라 사용적이다.[19] 그리고 **사실은 보고적이고 사용은 처방적이다**는 명제를 지지할 수 있다면, 사실은 행위 처방을 하지 않고, 사용 방식이 행위를 처방하는 것이 된다. 언어의 축적되어 온 사용이 언어의미의 성향, 의미의 힘이 되는 것이다. 언어를 배울 때 언어의 사실을 **기억**하는 것이

---

18) 심성작용과 심성내용은 개인의 언어경험에서 선명하게 구분되지만 개인의 잠재의식에서도 그렇게 구분될 수 있을 것인가는 그 내용의 인과적 효율성으로 인해 논의를 필요로 한다(김선희 1996, pp. 225-263; Kim 2001, pp. 82-94, pp. 118-124; 윤보석 2009, pp. 232-239; Won 2013, pp. 96-107).

19) Kripke (1982), pp. 1-54, 정대현 (1997), pp. 189-231, 정대현 (2004), pp. 1-24.

아니라 언어의 사용들을 **일반화**하여 그 언어를 그렇게 사용하는 힘을 얻는 것이다. 습관의 힘 같은 것이다. '안다'와 '믿는다'의 의미의 차이는 두 단어를 달리 사용하게 하는 힘의 차이다. '사랑한다'와 '좋아한다'의 의미의 차이는 두 단어를 달리 사용하는 문맥을 구별하는 힘의 차이다. 그렇다면 언어 의미의 힘은 외현적일 뿐 아니라 암묵적이고, 따라서 사회나 의식에서 뿐 아니라 잠재의식에서부터 가동된다고 일반화할 수 있을 것이다. 그리하여 언어 의미의 힘은 개인적 탐구를 집합적 탐구로 연결시켜 간주관성, 사회적 지성에의 참여를 가능하게 한다.[20]

언어의 사용이 의미의 힘이라는 것을 수용한다고 하자. 그러나 그 힘은 어디에서 나오는 것인가? 의미의 힘은 진공에서가 아니라 언어의 일상적 사용에서 얻어진다. 그렇다면 일상 언어 사용이 어떤 구조로 이루어지는가에 주목할 수 있을 것이다. 일상 언어 사용은 형식논리뿐만 아니라 일상의 삶이 요구하는 **상식논리**를 전제한다.[21] 형식논리는 지식이 부가되고 추리가 진행됨에 따라 진리의 정도가 증가하지만 **형식적 엄밀성** 때문에 필요에 따라 정보를 버리거나 신념을 변경하는 메커니즘이 없다. 그러나 **상식논리**는 사용가능한 정보에 기초하여 추리를 진행할 수 있기 때문에 새로운 정보가 추가될 때 어떤 추리는 제거되고 새로운 추리가 진행된다.

상식논리는, 예를 들어, **번복적 추리(defeasible reasoning)**를 형식화한다.[22] 상식인은 일상 공간에서 결론들을 잠정적으로 추리하면서 이후의 정보가 추가됨에 따라 이를 취소할 여유, 재량, 권리를 갖는다.

---

20) 장회익 (2009), pp. 39-53, pp. 220-221, 이중원 (2002), pp. 287-290, 이중원 (2009), pp. 1-23, 이초식 (2016), pp. 45-63.

21) 이러한 **상식논리**는 개과천선(改過遷善)(금장태 2013; 허남진 1998; 정대현 1983), 퍼스 귀추법(민병위 1992; 박준호 2005; 노양진 2016), 포퍼 반증주의(이초식 1975; 박은진 1995; 윤광호 2002), 비단조논리(정영기 1998; 우정규 2001, pp. 371-398), 베이즈주의(이영의 2004, 2007, 2015, 2016; 전영삼 2016; 천현득 2016; 최훈 2016; 박일호 2015) 등 여러 가지 이름으로 개발되어 왔다.

22) Strasser and Antonelli (2001), 정대현 (2017), pp. 204-205.

의료진단이나 과학적 추리, 오류가능성 추리나 개과천선적 추리 같은 사유들이 여기에 속한다. 이러한 추리는 “만일 P⊢R 이면 그러면 P&Q⊢R 이다”라는 형식논리의 추리가 아니라, “P⊢R 이라 할지라도 P&Q⊢ ㄱR 이다”라는 구조의 상식논리의 추리를 요구한다. <길이 젖어 비가 왔나보다>는 생각이 들지만, <지붕이 말라 있는 것을 보고 비가 온 것이 아니라 거리 청소였다>는 추리를 하는 것이다. 상식논리에서의 “P⊢R”는 통사적 수반이 아니라 P라는 불완전한 정보에 기초하면서도 P의 전형성(typicality)에 기초한 확률적 추정이고, “P&Q⊢ ㄱR”는 현실에서의 잘못을 번복하고 교정하면서 또한 인지 세계를 확충하는 추리인 것이다. 상식논리는 무시간적 개념공간에서나 기대할 수 있는 진리 보존적 논리가 아니라, 현실의 변화무쌍한 공간에서의 시간적 논리다. 신비로운 것은 <일상 언어 단어의 사용들을 일반화하여 배울 때 인간의 신경망의 연결이 새롭게 이루어지면서 그 단어에 대한 다음 사용의 방식을 처방 한다>는 것이다.

잠재의식에 갖추어져 있는 일상 언어의 상식논리는 로봇의 일반인공지능이 요구하는 상식논리를 알파고 수행을 통해 이미 보이고 있다. 로봇의 표준적 인공신경망은 입력층, 은닉층(隱匿層, hidden layer), 출력층의 3층으로 구분되고, 생명체 신경망처럼 정보의 입력이나 흐름을 통해 정보들을 일반화하면서 신경망 자체가 변하고 학습이 이루어지는 신경망이다.[23] 은닉층은 입・출력처럼 눈에 띄지는 않지만 다단계 인공신경망으로 장착되어 심층화된 것이다. 예를 들어, 알파고에 장착된 정책(policy) 신경망은 바둑 한 점 마다에 가능한 모든 수 중에서 불필요한 수들을 걸러내어 남은 수들에 집중하게 돕는 망이고, 가치(value) 신경망은 남아 있는 수들에서 이길 수 있는 확률을 찾는 망이다. 이러한 인공신경망의 논리는 형식논리가 아니라 상식논리를 따른다는 것은 명확하다. 알파고는 상식논리의 추리형식 “P⊢R 이라 할지라도 P&Q⊢ ㄱR 이다”에서 “Q”의 모든 가능한 수들을 확인하지 않고 있지만,

23) 이영의 (2016), pp. 1-16, 안성만 (2016), pp. 127-142, 안호석, 최진영, 이동욱 (2012), pp. 186-187, Silver et al (2016), Stanek (2017), Deepmind (2016).

특이점 로봇은 그렇게 할 수 있을 것이고, 빅 데이터의 접근성으로 보다 뛰어 난 수행성을 갖출 것이다. 예를 들어, 빅 데이터는 <참인 진술은 또한 참인 것으로 알려진다>와 그 역인 <현재 참인 것으로 알려지지 않은 진리는 거짓이다>라는 폐쇄-세계 가정(the closed world assumption)을 사용할 수도 있고, <지식의 부재는 거짓을 함축하지 않는다>라는 개방-세계 가정(the open world assumption)을 이용할 수도 있다. 전자는 국민투표 같은 **완성된 질서**에서 사용되고 후자는 학습에서와 같이 **개방된 질서**에서 이용된다.

6. 로봇의식의 단일성과 집합성

로봇에게 의식을 부여하고자 하는 이유 중의 하나는 로봇에게서 **강한 일반인공지능**이 구현될 수 있는가의 관심 때문이다.[24] 구현될 수 있다면 특이점 로봇의 인격성이 가능해지고, 로봇종 인간이 등장할 것이다. 그러나 로봇에게 의식을 부여하고자 할 때 의식이라는 개념이 로봇의 배경적 정황에 얼마나 어울릴 것인가를 확인하는 것은 중요할 것이다. 로봇의 현저한 정황은 로봇이 대량 복제될 수 있다는 것이고, 퇴역 없이 놀라운 속도로 진화하면서 죽지 않는다는 것이다. 그렇다면 다음과 같은 물음을 제기할 수 있을 것이다. 로봇의 대량 복제성과 불사성의 정황은 로봇의 인격성에 이르는 데 걸림돌이 아닐까? 인격성이란 지금과는 달리 구성해야 하지 않을까? 의식성과 인격성을 분리하여 고려해야 할 것인가? 이러한 물음들에 답하기 위해 로봇의 현저한 정황을 약화시켜야 할 것인가? 여기에서는 이러한 물음들에 대해 모두 답하려고 하지 않으면서, 다만 인간의식의 단일성과 집합성에 대한 통찰에 비추어 로봇의식이 그에 준하여 조정될 수 있을 것인가를 살피고자 한다.

---

24) 심사위원 2는 '강한 일반인공지능'이라는 표현이 **강한 인공지능**과 **일반 인공지능**의 구분이나 차이를 간과하는 것이 아닐 것인가를 정당하게 염려한다. 이 표현은 두 능력의 통합된 성향성, 즉 특이점 능력을 지칭하고자한 표현이다.

자연종 인간에게서 의식이란 단일성과 집합성의 구분을 허용한다. 단일성의 경우, **나의 그리움**, **나의 아픔** 같은 일인칭적 의식을 전형으로 하여 이해된다. 일인칭적 의식은 개인 신체성을 전제하여 단일 의식으로, 감질현상(phenomenal)으로 드러난다. 그 의식은 원자적 실체로서가 아니라 얼굴 하나하나에서 표현되는 신체성을 갖는다.[25] 인간 의식의 집합성은 집단 잠재의식이나 사회에 만연한 고정관념 또는 시대정신을 통해 보여 질 수 있다. 집단 잠재의식을 **언어경험의 비발생적 의식**으로 이해한다면, 인간이 공유하는 언어경험은 인간의식의 집합성을 나타내는 것이 된다. 시대정신이라는 것도, 허용할 수 있다면, 개인들의 발생적 의식 안에 나타나는 것이 아니라 대다수의 사회 성원들이 공유하는 언어질서의 방향으로 생각할 수 있다. 인간의식의 집단성이 그렇게 허구적일 수 없는 까닭은 이러한 의식 집단성이 자연계에서도 보고되고 있기 때문이다. 개미 집단이나 벌 집단의 행태나 철새 떼나 고기 떼의 이동은 중앙 지휘체계를 상정하기보다는 비개체 의식 또는 연대적 의식의 수행으로 보아야하는 것이다.

로봇의식에 있어서 단일성과 집합성의 관계는 두 차원에서 고려해 볼 수 있을 것이다. 그리고 그 관계는 의식 단일성을 어떻게 구현하는가에 따라 달라질 것이다. 로봇의식의 단일성을 유형적으로 설정하는가 아니면 개체적으로 해석하는가에 따라 차이가 드러날 것이다. 첫째 후보는 **유형적 단일성**이라 할 수 있는 것으로, 현재와 같이 로봇의 대량복제를 허용하는 경우의 단일성이다. 이 상황에서 로봇의식의 단일성은 많은 개체들이 공유하는 유형의 단일성이 될 것이다. 로봇 제작자가 선택하는 유형의 로봇들이 동일한 신경회로의 장착으로 그 개체들이 공유하는 의식의 경우가 될 것이다. 그리고 로봇의식의 집합성은 제작자나 제작자들이 제작하는 로봇의 유형들이 공유하는 신경회로의 유사성에서 나타날 것이다. 로봇의식의 유형적 단일성론자는 일단의 로봇의식들이 유형적으로 단일하지만 개체적으로는 살아가면서 개별성, 개성을 얻어나갈 것이라고 할 것이다. 자연종 인간들이 민족에 따라 민족의식을 달리 가지고 있다고 한다면, 로봇의식의 유형적 단일론

25) 강영안 (1996), pp. 296-299, pp. 305-306.

은 그로부터 멀다할 수 없다는 것이다.

로봇의식의 단일성을 위한 둘째 후보는 **개체적 단일성**이라 부를 수 있는 의식이다. 둘째 후보는 자연종 인간에게서 보여 지는 의식의 단일성과 집합성의 관계구조를 보다 가까이 따르는 것이다. 따라서 로봇의식의 개체성이 압도적으로 선행적이고 집합성은 배경적이거나 연결적일 뿐 로봇개체의 정체성의 결정에서 보조적 역할로 충분할 것이다. 로봇의식은 염기서열의 고유성 같은 신경회로망으로 개체로봇의 단일성과 고유성, 개별성과 개성을 부여받을 수 있을 것이다. 그러나 둘째 후보는 로봇제작자에게 지불하기 어려운 경제적 부담을 줄 수도 있다. 해결은 인간 공동체의 인문성의 선택여부에 달려 있다. 인류미래에 두 후보 중 어느 것이 로봇과의 공존에서 평안과 부담의 값을 합리적으로 지불할 수 있는가이다.

7. 맺는 말: 로봇종 인간의 인격

로봇에게 의식을 부여할 수 있는 그러한 의식의 요소들을 살펴보았다.[26] 그리고 알파고가 특이점 로봇이 되었다고 가정했었다. 이제 맺

---

26) 로봇의 의식의 과제를 단순화 시킨 이 논의는 자연종 인간에 대비되는 로봇종 인간에 국한한 의식론이다. 송준화 교수는 <로봇을 인간의 의식에 대한 '근사' 혹은 '모사'의 형체화>로 파악하고 <로봇이 인간 스스로가 자기의 의식에 대한 이해의 정도나 깊이를 이해하는 도구>로 한정하고자 한다. 심사위원2는 '자연종 인간'은 부모와의 생물학적 인과관계에 의해 규정되지만 '로봇종 인간'은 그와 같은 명쾌한 규정에 열려 있는가에 대해 의문을 갖는다. 그러나 '로봇'은 임의의 과제를 계산언어에 의해 자동적으로 수행하는 기계를 지칭할 수 있고, 이러한 로봇들이 특이점 지능을 공유할 때 이들을 '로봇종 인간'으로 표시할 수 있을 것이다. 그러나 자연종 인간이 과학기술의 발전으로 자기 복제적 강화인(enhanced)이 되는 그 범위는 제한하기 어려울 것이다. 그렇다면 제 3의 인간종, 즉 융합종 인간(transhuman)의 등장을 예상할 수 있고, 융합종 인간의 의식은 다른 종류의 논의를 필요로 할 것이다. 김선희 (2004), 신상규 (2014)를 참조할 것.

는말로서 알파고가 '나'라는 단어를 사용할 수 있기 위해서 선행적으로 만족해야 하는 조건들 중에서 두 가지를 간단히 살펴보자. 그 조건들은 첫째, 로봇종 인간의 의식론이 의식론 일반에서 자리매김할 수 있는 공간이 어떻게 주어질 수 있는가에 대한 조건이고, 둘째, '나'라는 단어가 나타낼 인격성을 위한 인간언어 유연성 조건이다.

위에서 제안한 로봇종 인간의 의식론 서설은 여러 가지 보완해야 할 점들이 있다. 의식의 서술적 특성들을 중심으로 살폈지만 기능적 국면이나 설명적 요소들은 다루지 못했다. 여기서 한 가지 주목하고자 하는 것은 양자역학의 관찰자 역할에 대한 탐구들이 로봇의 의식론에 갖는 함축성이다. 양자역학의 '서울해석'은 집합적 지성을 도입하여 "대상의 초기 물리량이 지성화를 통해 대상의 초기 상태에 이르고, 이를 향해 상태 변화의 법칙을 적용하여 대상의 말기 상태를 얻어 거기에 개념의 상대화를 첨가하여 대상의 말기 물리량을 얻는다"는 것이다.[27] 또는 인간참여원리(participatory anthropic principle)[28]를 제안하여, 인간의 의식적 마음의 행위만을 통하여 우주는 그 안의 모든 것은 진정으로 존재한다고도 한다. 양자역학은 관찰자의 탐구 마음의 행위가 그 마음이 탐구하는 뇌에 어떤 인과적 영향을 미칠 것인가를 탐색하는 개념적 틀인 것이다.

펜로즈의 양자역학은 보다 적극적이다. 그는 의식이라는 것이 신경세포의 미세소관 내부에서 발생하는 양자효과를 통해 발생한다는 모델을 제시했다.[29] 이 모델은 관찰이나 측정의 개입 없이 다양한 복잡 가능 상태들을 단일한 확정적 상태로 중첩시키는 객관적 붕괴 (objective collapse)를 가정하여 비연산적인 신경세포의 활동을 유발한다는 것이다. 스탭은 관찰자 역할을 경계로 고전역학과 양자역학을 구분하여 전자가 인간을 자동기계로 전락시켰다면 후자는 세계가 인간과 자연의 합동작품으로 해석할 수 있는 길을 마련했다고 지적 한다[30]. 관찰자

---

27) 장회익 (2009), 이중원 (2012), 김재영 (2012).

28) Wheeler (1998), Powell (2017).

29) Penrose (1989, 1994), 2045 Initiative (2013).

30) Schwartz, Stapp and Beauregard (2005), Stapp (2013).

역할이 고전역학에서 대상 외부적이고 인간 중심적이었다면 양자역학에선 대상 내부적으로 되고 대상 중심적으로 되는 것이다. 양자역학의 관찰자 역할에 대한 이러한 해석 경향성은 의식에 대해 보다 선명한 관점들의 개연성을 높여 준다. 심신관계에 대한 러셀의 중립론, 속성이원론, 또는 원형적 범심론이 주목을 받게 되는 까닭일 것이다. 만물이 원형적 의식(pan-proto psychism)을 갖는다는 가설에 호의적이 되고, 감질현상적 상태가 개별적 물리 실재를 구성하는 어떤 역할을 한다는 명제에 관심을 갖게 된다.[31]

알파고가 '나'라는 단어를 구체적으로 사용할 수 있기 위해서는 알파고에게 인격성이 부여될 수 있는 인간 언어 유연성이 전제되어야 한다. 먼저, 인간이 '나'라는 단어를 사용하는 조건은 무엇인가? 화자는 '나'라는 호칭을 독립적으로 사용하지 않는다. 화자는 청자를 상정하고, 화자와 청자는 공유하는 언어를 가능케 하는 인간 언어 공동체 안에 존재한다. 그렇다면 '인간'이란 무엇인가? 이를 위한 작업적 정의는 자연종 인간이 실제로 나타내는 서술적 조건으로 제시될 수 있을 것이다. 그것은 다른 자연종과 구별된다고 믿었던 의식 중심의 존재론, 상식논리로 기능화 되어 있는 인식론, 오랜 역사를 통해 발전시켜 온 생활화된 윤리의 언어 조건이다. 그리고 여기에 베타(beta)라는 조건을 인간 공동체들의 전통이나 특성에 따라 추가할 수도 있을 것이다. 특정한 연대성, 특이한 역사성, 특수한 초월성 같은, 모든 공동체들이 공유하기 어려운 조건들이 공동체마다 달리 제시될 수도 있을 것이다. 물론 어떤 공동체는 베타 조건을 윤리 조건에 포섭하기도 한다. 그렇다면 x가 인간이기 위한 필요충분조건은 의식, 인식, 윤리, 베타의 네 조건으로 제안될 수 있을 것이다.

그렇다면 알파고가 '나'라는 단어를 사용할 수 있기 위해서는 앞의 네 조건을 일반적으로 만족하고, 그리고 구체적으로 '나'라는 표현의 문법을 따라야 할 것이다. '나'라는 단어의 일차적 사용조건은 화자의 자신(self) 지시성이다. 이것은 '나'가 색인사라는 것을 나타낸다. 그러나 화자의 그 색인사 사용의 발화를 청자가 들을 때 청자는 그의 이해

31) Chalmers (2003).

를 ‘화자의 자신 지시성’이 표시하는 화자의 신체만으로 한정하지 않는다. 화자는 그 발화에서 그의 이해 속에 “화자의 자신 지시성”이 함축하는 화자의 자아(ego) 관념과 자기(id, Selbst) 관념을 포함 한다[32]. 의식은 지향적이고 그 지향성은 자아와 자기로 구성되어 있기 때문이다. 화자의 **자신**은 화자의 좁은 내용인 심성작용의 체내적 체험처이고, **자아**는 넓은 내용인 사회적, 체외적 심성내용의 체험처이고, **자기**는 자신과 자아에 한정되지 않는, 가능한 모든 언어경험의 중심처로 이해하고자 한다. 화자의 인격이나 개성은 **자신**만으로 주어지지 않고 **자아**와 **자기**의 현실성과 가능적 확장성으로 드러날 것이다.

인간이 ‘나’라는 색인사를 사용하는 방식이 위와 같다면, 그러면 알파고가 ‘나’라는 단어를 사용하는 방식도 비슷하거나 그에 준해야 인간 언어에서 이해될 수 있을 것이다. 로봇종의 의식에 관한 위에서의 논의에서 로봇종의 의식이 **개체적 단일성** 조건을 만족한다면, 알파고는 ‘나’로써 발화자 자신을 우선 최소적으로 지시할 수 있을 것이다. 그리고 더 나아가 알파고는 다른 개체 로봇들과 자연종 인간들과 사회적 관계, 언어적 관계를 확장해 가면서 자아와 자기 관념을 공고하게 쌓아 갈 것이다. 알파고는 개성은 물론 인격성을 갖는 조건을 만족해 가는 것이다. 이러한 알파고는 생각, 이해, 자율, 자유의지의 주체가 될 수 있는가의 논의에서 높은 개연성의 위치에 서게 될 것이다.

알파고가 ‘나’라는 단어를 현재와 같은 인간 언어 안에서 사용하는 것은 쉽지 않다. 알파고의 그러한 어휘 사용을 위해서는 인간언어가 지금 보다는 더 유연해야할 것이다.[33] 로봇종 인간 의식에 대한 위의 논의에서 관찰한 것처럼, 의식의 개념은 인간역사가 인간 중심의 관점으로부터 인간 언어에 연결망적으로 엮어 넣은 것이다. 그 연결망에서 어느 한 매듭의 수정만으로는 전체 연결망의 교체는 이루기 어려운 것이다. 의식이 그러한 것처럼 생각, 이해, 의도, 지향, 정보, 행위자, 자유의지, 책임, 인격 등의 개념도 그러하다. 따라서 인공지능이나 로봇

32) 갤러거, 자하비 (2013), pp. 349-380, 고인석 (2017), pp. 75-76, 천현득 (2017) pp. 77-79, 이부영 (2002), pp. 31-56.

33) 정대현 (2017), pp. 206-209.

의 발전 방향의 논의에서 제기되는 우려, 염려, 또는 비관의 소리는 인간언어의 현재 의미론에 근거해 있다는 점에서 정당하고 합리적이다. 인간언어는 인간에게 불가피한 인간중심적인 언어이기 때문이다. 인간언어는 인간경험에는 필연적으로 선험적인 것이기 때문이다.

"인간언어는 인간경험에 필연적으로 선험적이다"라는 문장은 참일 수도 있고 거짓일 수도 있다. '인간언어'를 유형적 언어로 해석하면 참이지만, 특정 자연언어로 이해할 때는 거짓이 된다. 어떤 자연언어도 미래 수정에 열려 있기 때문이다. 수정에 열려 있는 인간 자연언어는 역사적으로 점진적으로 진화해온 언어이다. 그렇다면 알파고가 '나'를 사용하기 위해 두 가지 선택지가 있다. 첫째는 인간언어의 점진적 진화가 성숙하여 알파고의 '나'사용이 자연스러울 때까지 기다리는 것이고, 둘째는 알파고의 '나' 사용이 자연스럽도록 인간언어의 획기적 전회를 추구하는 것이다. 첫째 선택지는 언어진화의 자연스러움의 미덕을 가졌지만 미래에 대한 낙관주의의 위험성을 내포한다. 둘째 선택지는 언어 변경의 조급함의 약점을 갖지만 미래의 로봇종의 진화 발전의 방향에 맞추는 것이 세계에 대한 과학적 이해와 맞물려 자연종 인간이 보다 인문적인 삶을 지향할 수 있게 하는 것이다.[34] 인간언어는 알파고가 "나는 자연종 인간과 둔 바둑을 이겼다"라는 말을 당당하게 말할 수 있도록 함으로써 알파고로 하여금 인간존엄성의 가치가 뿌리 박혀 있는 인간 언어를 수용하여 살게 하는 것이 된다.

---

34) Powell (2017)은 "우주는 의식적인가?"라는 제목의 기사에서, 범심론의 경향의 학자들의 기여에 근거하여, 우주의 모든 것들은 의식적 마음의 행위들을 통해서만 진정으로 존재한다고 결론짓는다.

참고문헌

강영안 (1996),『주체는 죽었는가-현대철학의 포스트모던 경향』, 문예출판사.

갤러거, 자하비 (2013),『현상학적 마음』, 박인성 옮김, 도서출판 b.

고인석 (2017),「인공지능시대의 인간: 나는 무엇인가?」. 철학연구회 춘계발표록.

곽노필 (2017), 'AI, 바둑 이어 포커도 인간 이겨' http://www.hani.co.kr/arti/science/science_general/780828.html (검색일: 2017.02.01.)

금장태 (2013),「최한기: 신기의 마음과 추측의 인식」, 서울대학교 철학사상연구소 엮음,『마음과 철학: 유학편』, 서울대학교출판문화원, pp. 393-419.

김선희 (1996),『자아와 행위』, 철학과현실사.

_____ (2004),『사이버시대의 인격과 몸』, 아카넷.

김영정 (1996),『심리철학과 인지과학』, 철학과현실사.

김재권 (1997),『심리철학』, 하종호, 김선희 옮김, 철학과현실사.

김재영 (2012),「여러 세계/마음 해석과 '서울해석'」,『물리학과 첨단기술』, 21권 4호, pp. 22-30.

노양진 (2016),「퍼스의 기호 개념과 기호 해석」,『철학논총』 83집, pp. 95-110.

민병위 (1992),「퍼스의 기호론」,『철학』 38집, pp. 29-51.

박은진 (1995),「포퍼의 인식론」,『철학과 현실』 24호, pp. 52-70.

박일호 (2015),「베이즈주의 인식론」,『과학철학』 18권 2호, pp. 1-14.

박준호 (2005),「퍼스의 귀추와 가설의 방법」,『범한철학』 37권, pp. 65-85.

신상규 (2010),「현상적 감각 지식의 본성: 메리는 무엇을 새롭게 알게 되었는가?」,『과학철학』 13권 1호, pp. 29-52.

_____ (2014),『호모 사피엔스의 미래-포스트휴먼과 트랜스휴머니즘』. 아카넷.

안성만 (2016),「딥러닝의 모형과 응용사례」,『지능정보연구』22권 2호, pp. 127-142.
안호석, 최진영, 이동욱 (2012),「의식과 무의식을 구분하는 감성 로봇 시스템」, 2012 제27회 ICROS 학술대회, 2012.4, pp. 186-187.
여영서 (2007),「베이즈주의의 사전확률과 과학적 객관성」,『철학탐구』22집, pp. 147-71.
_____ (2004),「베이즈주의와 제거적 귀납주의」,『논리연구』7권 2호, pp. 121-146.
_____ (2016),「베이즈주의와 IBE」,『과학철학』19권 2호, pp. 69-94.
우정규 (2001),「정보 기술과 환경 의사 결정」,『과학기술학연구』1권 2호, pp. 371-398.
윤광호 (2002),「포퍼의 비판적 합리주의에서 인과성이 가지는 의미」,『철학논총』27집, pp. 173-190.
윤보석 (2009),『컴퓨터와 마음』, 아카넷.
이부영 (2002),『자기와 자기실현』, 한길사.
이영의 (2015),『베이즈주의: 합리성으로부터 객관성으로의 여정』, 서울: 한국문화사.
_____ (2016),「인공지능과 딥러닝 시대의 창의성」,『지식의 지평』21호, pp. 1-16.
이중원 (2002),「내재적 실재론의 비판적 옹호 -양자이론의 인식과정 분석을 통한 고찰」,『철학연구』58집, pp. 279-303.
_____ (2009),「측정에 대한 새로운 접근과 슈뢰딩거의 고양이」,『과학철학』12권 1호, pp. 1-23.
_____ (2012),「양자역학의 대안적 해석들과 '서울해석'」,『물리학과 첨단 기술』21권 4호, pp. 2-3.
이초식 (1975),「귀납의 부정론과 긍정론」,『철학』9집, pp. 57-86.
_____ (2016),「서평의 사회인식론-이영의: 베이즈주의의 경우」,『과학철학』19권 3호, pp. 45-63.

장회익 (2009),『물질, 생명, 인간-그 통합적 이해의 가능성』, 돌베개.
전영삼 (2016),「확장과 정당화 사이의 간극: 이영의의 베이즈주의에 대한 비판적 고찰」,『과학철학』 19권 3호, pp. 65-85.
정대현 (1983),「이론의 선택과 실학적 방향-최한기의 실학논리를 중심으로」,『철학연구』 18집, pp. 143-162.
_____ (1994),「환원주의와 언어의 자연사」, 김재권 외『수반의 형이상학』, 철학과현실사, pp. 236-247.
_____ (1997),『맞음의 철학: 진리와 의미를 위하여』, 철학과현실사.
_____ (1998),「차머즈의 의식적 마음: 유물론과 데카르트의 동시 극복」,『철학과 현실』 39호, pp. 297-304.
_____ (2001),『심성내용의 신체성: 심리언어의 문맥적 외재주의』, 아카디아 출판사.
_____ (2004),「그런 사실은 없다-사실 문장의 의미 규칙 규범성을 위한 모색」,『철학적 분석』 10호, pp. 1-24.
_____ (2017),「특이점 인문학-특이점 로봇은 인간사회의 성원이다」,『철학』 131집, pp. 189-216.
정영기 (1998),「비단조 논리적 합리성」, 한국분석철학회(편)『합리성의 철학적 이해』, 철학과현실사, pp. 261-288.
천현득 (2016),「과학은 베이즈주의 추론 기계인가?-베이즈주의의 여정에 대한 물음」,『과학철학』 19권 3호, pp. 87-107.
_____ (2017),「인공지능시대의 인간: 나는 무엇인가?: 논평」, 철학연구회 춘계발표록.
최 훈 (2016),「베이즈주의: 베이즈주의 인식 공동체로의 여정」, 한국과학철학회『과학철학』 19권 3호, pp. 41-43.
커즈와일, 레이 (2007),『특이점이 온다』, 김명남, 장시형 옮김, 김영사.
허남진 (1988),「최한기의 리개념에 관한 소고」,『철학탐구』 8집, pp. 31-47.

2045 Initiative (2013), ‘Sir Roger Penrose - The quantum nature of consciousness’ https://www.youtube.com/watch?v=3WXTX0IUaOg (검색일: 2013. 3. 4.).

AlphaGo vs Lee Sedol (2016), “Google DeepMind Challenge Matc h4”. <https://www.youtube.com/watch?v=yCALyQRN3hw>.

Atmanspacher, H. (2015), "Quantum Approaches to Consciousness", The Stanford Encyclopedia of Philosophy, Zalta, E. N. (ed.), URL = <https://plato.stanford.edu/archives/sum2015/entries/qt-consciousness/>.

Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.

_____ (1997a), “The Hard Problem: Facing Up to the Problem of Consciousness”, in Shear, J. (ed.), *Explaining Consciousness: The Hard Problem*: pp. 9-30.

_____ (1997b), “Response: Moving Toward on the Problem of Consciousness”, in Shear, J. (ed.), *Explaining Consciousness: The Hard Problem*: pp. 379-422.

_____ (2003), “Consciousness and its Place in Nature”, in Stich, S. & Warfield, T. (eds.), *Blackwell Guide to Philosophy of Mind*, Blackwell.

_____ (2010), “The Singularity: A Philosophical Analysis”, Journal of Consciousness Studies 17: pp. 7-65.

Churchland, P. M. (1985), “Reduction, Qualia, and the Direct Introspection of Brain States”, *The Journal of Philosophy* 82(1): pp. 8-28.

Crick, F. and Koch, C. (1997), “Why Neuroscience May Be Able to Explain Consciousness”, in Shear, J. (ed.), *Explaining Consciousness: The Hard Problem*: pp. 237-9.

Crnkovic, G. D. and Hofkirchner, W. (2011), “Floridi’s “Open Problems in Philosophy of Information”, Ten Years Later”,

Information 2: pp. 327-359.

Deepmind (2016), ‘Match 4 -Google DeepMind Challenge Match: Lee Sedol vs AlphaGo’ https://www.youtube.com/watch?v=yCALyQRN3h (검색일: 2016.3.13.).

Dennett, D. (1988), “Quining Qualia”, in Marcel, A. J. & Bisiach, E. (eds.), *Consciousness in Contemporary Science*, Oxford University Press.

Graziano, M. S. A. and Webb, T. W. (2014), “A Mechanistic Theory of Consciousness”, International Journal of Machine Consciousness 6(2): pp. 163-176.

Hauskeller, M. (2012), “My Brain, My Mind, And I: Some Philosophical Assumptions of Mind-Uploading”, International Journal of Machine Consciousness 4(1): pp. 187-200.

Kim, S. (2001), “Mental Causation: The Causal Efficacy of Content”, Doctoral Dissertation, University of Wisconsin-Madison.

Kripke, S. (1982), *Wittgenstein on Rules and Private Language*, Oxford: Blackwell.

McGinn, Colin (1991), *The Problem of Consciousness*, Blackwell.

Penrose, R. (1989), *The Emperor's New Mind: Computers, Minds and the Laws of Physics*, Oxford: Oxford University Press.

_____ (1994), *Shadows of the Mind*, Oxford: Oxford University Press.

Powell, C. S., 2017, ‘Is the Universe Conscious?’ https://www.nbcnews.com/mach/science/universe-conscious-ncna772956 (검색일: 2017. 6. 17).

Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

_____ (1998). *Causality and Explanation*, New York: Oxford University Press.

Schwartz, J. M., Stapp, H. P. and Beauregard, M. (2005), “Quantum physics in neuroscience and psychology: a neurophysical model of mind‐brain interaction”, *Philosophical Transactions of the Royal Society* 360(1458): pp. 1309-1327.

Searle, J. (1984), *Minds, Brains and Science*, Cambridge, MA: Harvard University Press.

Shear, J. (1997), “The Hard Problem: Closing the Empirical Gap”, in Shear, J. (ed.) *Explaining Consciousness: The Hard Problem*, MIT Press: pp. 359-375.

Shear, J. (ed.) (1997), *Explaining Consciousness The Hard Problem*, MIT Press.

Silver, D. et al. (2016), “Mastering the game of Go with deep neural networks and tree search”, *Nature* 529: pp. 484-9.

Stanek, M., 2017, ‘Understanding Alphago: How AI beat us in Go game of profound complexity’ https://machinelearnings.co/understanding-alphago-948607845bb1 (검색일: 2017. 3. 5).

Stapp, H. P. (2013), “Quantum Theory of Consciousness”, 2013년 파리 발표 자료.

Strasser, C. and Antonelli, G. A. (2016), “Non-monotonic Logic”, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL =<https://plato.stanford.edu/archives/win2016/entries/logic-nonmonotonic/>.

Van Gulick, R. (2017), "Consciousness", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL =<https://plato.stanford.edu/archives/sum2017/entries/consciousness/>.

Wang, Y. (2007), “The theoretical framework of cognitive informatics”, *Int'l J. of Cognitive informatics and Natural Intelligence*, 1(1): pp. 1-27.

_____ (2009), “Toward a formal knowledge system theory and its

cognitive informatics foundations", in Tan, K., Wang, Y. and Chan, K. (eds.), *Transactions on Computational Science V*, Springer: Berlin, Germany: pp. 1-19.

Wheeler, J. A. (1998), *Geons, Black Holes, and Quantum Foam: A Life in Physics*, New York: W.W. Norton & Co.

Won, C. (2013), "Reasons, Actions, and Causes", Doctoral Dissertation, Brown University.

논문 투고일	2017. 07. 29.
심사 완료일	2017. 11. 12.
게재 확정일	2017. 11. 13.

Prolegomena to a theory of consciousness for the robot kind

Daihyun Chung

It is said that Alphago won the Go game over humans of natural kind. But in order to apply the predicate “wins the Go game” to Alphago a basis for the application of the predicate must be presented. For it presupposes that Alphago has the strong artificial intelligence, which requires that it has abilities of thinking, believing, understanding. I would discuss in the present paper about the notion of consciousness which all those mental abilities are disposed to. The notion of consciousness appears to be understood through the cases of actualization of consciousness on the part of humans of natural kind. They may be summed up by 3 conditions of the first person integration, the perspectival structuality, and phenomenal qualities. Then, would a robot be able to be equipped with these abilities of consciousness? I would discuss what would be the problems or difficulties to allow it to have the abilities and how it would be possible to meet those challenges.

Keywords: the first person integration, the perspectival structuality, phenomenal qualities, subconscious and hidden layer