【서평】 과학철학 19권 2호 (2016) pp. 95-107

인공지능 시대의 인식론과 과학철학: 『뇌처럼 현명하게: 신경철학연구』

패트리샤 처칠랜드 지음, 박제윤・김두환 옮김
(철학과현실사, 2015)

김 효 은†

처칠랜드 부부(패트리샤 처칠랜드와 폴 처칠랜드)의 저작들을 꼼꼼히 꿰뚫고 있는 박제윤 교수는 국내에 불모지나 다름없는 신경철학 분야에 오랜 기간 중요한 기여를 해왔다. 그 기여는 크게 두 가지로 볼 수 있다. 하나는 신경과학에 대한 이해가 선결되어야만 가능한 처칠랜드의 저작을 국내에 소개하여 신경철학이라는 분야를 그야말로 '접근 가능'하도록 만들어놓았다는 점이다. 신경철학과 인식론을 접목한 박사논문 이래로, 박제윤 교수는 폴 처칠랜드와 패트리샤 처칠랜드의 저작 『뇌과학과 철학』(*Neurophilosophy*), 『신경 건드려보기』(*Touching a Nerve*), 『뇌처럼 현명하게』(*Brain-Wise*), 그리고 최근의 『플라톤의 카메라』(*Plato's Camera*)까지 결코 쉽지 않은 신경철학의 작품들을 번역했다. 두 번째는 박제윤 교수의 번역이 단순한 문장 옮기기가 아니라 새로운 번역어와 해독을 제시함으로써 국내외로 이슈가 되는 중요한 논의들인 '환원', '통섭', '개념', '인식론', '과학적 추리' 등의 전통적인 철학적 개념에 대한 재구성과 재논의—대표적으로 윌슨의 *Consilience*에 대한 새로운 읽기의 제안—을 촉발하고 있다는 점이다.

이 서평이 다루는 패트리샤 처칠랜드의 저서는 *Brain-Wise*이다. 이 책은 처칠랜드 자신이 그 동안의 작업을 정리하고자 했던 듯, 내용의 측면에서 저자의 주요 작업으로 인정받았던 *Neurophilosophy*의 후속판

† 중앙대학교 철학과 강사, qualia9@gmail.com.

일 뿐만 아니라, 의식, 자아, 자유선택, 도덕 경험, 종교의 신 존재 증명, 나아가 사후세계 등에 대한 새로운 해석을 포함하고 있다.

이 모든 내용을 요약, 소개하는 것은 서평의 목적이 아니다. 이 서평은 처칠랜드의 신경철학이 어떤 점에서 혁명적인지를 먼저 짚어본다. 이는 전통적인 인식, 개념, 일반화, 추리 과정, 이론간 환원에 대한 설명에서 드러날 것이다. 더 나아가 이러한 처칠랜드의 설명이 기존의 뇌-중심적 철학이나 시각들에 비하여 가지는 차별성이 무엇인지를 밝힐 것이다. 그리고 뇌과학과 인공지능이 득세한 현 시대에 처칠랜드의 작업과 박제윤 교수의 번역이 지니는 의미를 조명할 것이다.

1. 처칠랜드의 신경철학

알파고와 이세돌의 대결 이후 인공지능에 대한 관심이 더욱 높아졌다. 인공지능의 발전에 대한 대중의 관심은 사회문화적인 측면이 주로 거론된다. 반면, 연결망에 기반한 알파고에 대한 철학적인 관심은 다음의 두 가지이다. 하나는 한 때 부호주의(symbolism) 모형의 득세에 주춤했다가 재기하여 활약하는 연결주의(connectionism 혹은 병렬분산처리Parallel Distributed Processing 모형이라고도 칭함)라는 인지모형에 대한 관심이다. 다른 하나는 연결주의 모형에서 볼 때 지식과 지식 형성을 어떻게 설명하는가의 인식론적 관심이다.

현재 전 사회적으로 개발, 사용되는 '인공지능' 기획은 넓게 보아 기본적으로 지식을 구현하고 확장 그리고 응용하려는 시도이다. 처칠랜드의 *Brain-Wise*는 바로 이 지식들이 어떻게 가능해지는지를 설명한다. 처칠랜드의 신경철학은 인지주의의 두 모형 중 연결주의 모형에 기반을 두고 최신의 인지신경과학 자료들로 업그레이드하여 전통적인 철학의 주요 주제들을 새로이 해석한다. 그러나 처칠랜드의 작업은 기존의 신경철학자들의 작업과는 차별화된다. 일반적으로 신경과학에 기반하여 철학적 의견을 개진할 때에는 신경과학적 연구 결과가 가지는 철학적 의미를 언급하는 경우가 많다. 이러한 작업은 철학적 의미를

언급하지만 그럼에도 불구하고 여전히 그 작업의 성격이 '기술적'(descriptive)이다. 또, 철학의 논의를 진행하면서 지지하는 자료를 찾기 위해 신경과학의 연구 결과를 인용하여 붙이는 경우도 있다. 그러나 이 경우는 신경과학의 성과에 대한 자의적 해석의 위험성이 있을 뿐만 아니라, 여전히 '규범적'(normative) 작업이기만 하다. 이와 대조적으로 처칠랜드의 작업은 다음에서 보듯 규범성과 기술성, 신경과학철학과 신경철학의 작업을 동시에 수행한다는 점에서 바람직한 융합적 연구를 보여준다.

2. 신경과학철학과 신경철학의 통합

패트리샤 처칠랜드(이하 처칠랜드)는 대표적인 신경철학자이지만 사실상 신경철학은 여러 모습과 차원, 층위에서 진행된다. 그렇다면 처칠랜드의 작업은 특히 왜 주목할 만하고 주목해야 하는가? 이를 위해 먼저 뇌와 마음의 관련성에 대한 간략한 철학사를 살펴보자.

신경철학적 설명은 최근의 신경과학의 발달과 더불어 발생한 과학철학 영역에만 속한 것처럼 보이지만, 실상은 1950년대에 뇌와 마음의 관계를 설명하고자 시도한 일군의 심리철학자들인 플레이스(T. T. Place) 스마트(J. J. C. Smart)의 유형동일론(type-type identity theory)이나 개항동일론(token-token identity theory)에서 거칠게나마 그 기원을 찾을 수 있다. 물론 '통증은 C-섬유 발화'라는 식의 마음-신경 동일론은 현재의 발전된 신경과학의 성과를 온전히 반영하지 못한 것이었다. 본격적으로 '신경(과학)철학'이 시작된 것은 1980년대 중반 처칠랜드 교수가 『신경철학: 마음-뇌의 통합적 과학을 향해』(*Neurophilosophy: Toward a Unified Science of the Mind-brain*, 1986)를 출간하면서 본격적인 철학과 신경과학의 접목이 이루어지기 시작했다. 그 이후, 신경상태의 각 층위들 사이에 환원이 어떻게 가능한지를 중심으로 흥미로운 연구들이 진행되었다.

처칠랜드의 신경철학은 왜 다른 신경철학에 비교해 독보적이고 대

표적인 것으로 인정받는가? 먼저 처칠랜드의 신경철학은 철학의 '규범적'(normative)특성과 과학의 '기술적'(descriptive)특성이 잘 통합된 작업이다. 신경(과학)철학은 일차적으로 과학철학의 범주에 속한다. 그런데 신경철학의 독특한 지위는 신경철학이 최근 신경과학, 심리철학, 인지과학, 생물학, 심리학이 만나는 교차로에서 비약적으로 발전했다는 점에 있다. 따라서, 신경철학은 이차적으로 인식론, 심리철학, 생물철학의 주제들과 결합하여 새로운 시각과 전망을 제공한다. 이러한 통합적 특성은 논의의 편의를 위해 마련한 다음의 구분과 관련된다.

통상 '신경과학의 철학'(philosophy of neuroscience)과 '신경철학'(neurophilosophy)은 혼용되지만 굳이 엄밀히 구분하자면 구체적 작업은 구분된다. 전자인 '신경과학의 철학'은 신경과학의 작업에 관해 과학철학에서 제기하는 전통적 문제들—예컨대 신경과학적 설명이란 무엇인가, 신경과학 자료에 대한 해석 방법의 문제, 신경 표상에 대한 설명들(Churchland & Sejnowski 1990)—을 다룬다. 후자인 '신경철학'은 신경과학적 발견을 철학적 문제들에 적용하는 작업이다. 예컨대 감정이나 행복, 욕구에 대해 도파민 기능처럼 보상과 같은 신경시스템 기능으로 설명하는 등 전통적인 철학적 개념에 관한 새로운 설명을 제공한다. 흥미롭게도, 처칠랜드는 두 가지 접근을 통합해서 신경상태와 정신 상태에 대한 통합적인 설명을 제공한다.

이것은 구체적으로 어떤 의미인가? 처칠랜드가 단순히 신경과학과 철학을 섞어서 이론을 그럴듯하게 만든 것이 아니라 신경과학의 실제 층위와 차원을 종합하고 이것이 철학의 기존 논의들에 대응하는 이론을 생산해냈다는 의미이다. 일반적으로 '신경철학'이라고 할 때 해석의 애매함이 내재해있다. 이는 '신경상태'로 가리키는 상태가 다양한 층위들을 가리킬 수 있기 때문이다. 일반적으로 신경과학자나 신경철학자들은 이 다양한 층위들 중 어느 한 층위를 대상으로 집중적인 연구를 수행한다. 그런데 처칠랜드의 작업이 여타의 신경철학자와 다른 점은 전체 층위들을 모두 고려한 종합적인 탐구에 기반한 작업이라는 점이다.

융합(혹은 여타의 어떤 번역어나 개념어이든)이 현 시대의 큰 과제로 규정되는 시대에 처칠랜드의 이러한 작업은 하나의 범례로 기능한

다. 지금까지 ‘신경철학’ 혹은 ‘신경과학철학’(이하, 따로 표시하지 않는 한 ‘신경철학’으로 통일)이라는 명칭 아래 포섭되는 연구들은 여러 주제들로 나타났다. 이 주제들 중 특정 한 주제가 신경철학을 대표하기는 어렵다. 그 이유는 신경철학이 참고하는 뇌신경과학의 연구 자체가 여러 층위들로 구성되어 있고 그 층위들 간의 상호작용이 있기 때문이다. ‘층위’(layer)란 ‘차원’(order)과는 달리 그 용법 자체에 각 층위들 간의 연속성과 비연속성, 즉 각 층들 간의 상호작용과 비환원성이 내재되어 있는 개념이다. 보통 각 층위에 해당하는 신경철학의 주제들 하나하나가 별도의 연구주제들이기 때문에 이 층위들은 각기 독립적으로 연구될 수밖에 없는 것이 사실이다. 그런데, 처칠랜드의 작업은 이 여러 층위들을 모두 포함하고 있고 이 층위들에 나타나는 신경표상들에 대한 철학적 해석을 제시한다는 점에서 그야말로 구체적이고 깊으면서도 포괄적인 신경철학적 작업이다.

3. 개념, 추리, 일반화의 신경인식론적 정당화: 신경망 활성화의 원형

자연주의 인식론의 계통을 구체적으로 신경철학을 통해 확립한 처칠랜드는 구체적으로 어떻게 지식 형성과 구현을 설명하는가? 전통적 인식론의 선험적 방법은 정면으로 비판하였지만 인식론이 다루었던 지식과 정당화, 개념 등은 여전히 설명의 대상으로 남아 있다. 플라톤이 주로 천착했던 개념의 기원과 본성, 그리고 아리스토텔레스가 탐구한 일반화와 가설 정립은 어떻게 설명될 수 있는가? 여기에 처칠랜드에게는 시대적 과제가 추가된다. 전통 인식론에서 다루어졌던 개념, 일반화 등은 오늘에 와서 뇌라는 인간의 자연지능을 보완하거나 대체할 것처럼 보이는 인공지능을 통해 구현된다. 좀 더 구체적인 예를 들어보자. 최근 구글이 완성한 고양이 얼굴 인식 인공지능이나 페이스북이 개발하고 있는 인간 얼굴 인식 인공지능은 어떻게 얼굴들 간의 차이에 대한 지식을 구현하는가? 처칠랜드는 이런 것을 설명할 수 있는가?

처칠랜드의 신경철학은 지식을 다루는 구체적 주제인 개념과 일반화, 귀납추리의 가능성, 그리고 창의성까지 신경망의 관점에서 새롭게 설명한다. 인공지능의 작업이 얼굴 인식이든 문장의 이해든, 바둑알의 패턴 파악이든 간에, 이러한 지식은 모두 '일반화'라는 지식 형성과 관련된다. 처칠랜드는 그의 책에서 인공신경망(7장 6절)과 현재 알파고와 같은 개발 중인 인공지능 알고리듬에 대해(8장 7절) 자세한 설명을 제공하지만, 기본적으로 그의 설명은 인지모형의 기초인 신경망을 중심으로 이루진다. 이는 인공지능의 현재 발전된 모형을 따른 것이 아니라 이미 처칠랜드가 1986년 저술한 *Neurophilosophy*에서 제안한 것이다. 처칠랜드에 따르면, 뇌신경 연결망의 견지에서 볼 때 뇌에서 개념이나 일반화는 '신경망 활성화의 원형'이라는 동일한 기제에 의해 형성된다. 여기서 '원형'은 범주화/개념에 대한 기존의 여러 이론 중 하나인 '원형이론'과 혼동될 여지가 있지만 그 내용은 완전히 다른 것이다. 먼저 기존의 원형이론은 기본적으로 낱말에 대한 언어적인 분석에 기반을 두고 있다. 반면에 개념 파악은 일종의 신경망 활성화의 원형의 작용이라는 분석은 언어적 개념 분석에 근거하지 않는다. 이 점에서 처칠랜드의 설명은 기존 인식론의 틀을 넘어선다. 이 내용은 매우 중요하다. 신경망 활성화의 원형 이론에 기반하여 지각, 개념, 일반화와 가설, 나아가 은유, 유비, 그리고 가추를 비롯한 귀납적 추리에 대한 설명이 가능해지기 때문이다.

그러면, 구글이나 페이스북에서 개발하고 있는 얼굴인식에 대하여 처칠랜드의 설명항인 신경망 활성화는 구체적으로 어떻게 지식을 획득하는 것으로 설명하는가? 책 내용에 대한 요약보다는 인식론적인 재구성의 의의를 찾기 위해 간단하게 살펴보자. 처칠랜드에 따르면 현상의 표상은 뉴런의 활성이며 이는 위상공간의 지점들로 표현된다. 이 위상공간의 상태들은 표상이나 개념의 전환에 따라서 그 좌표가 전환된다. 무슨 의미인가? 먼저, 얼굴에 대한 표상은 하나의 뉴런이 특정 속성에 관여하는 '국소 부호화'(local coding)보다는 여러 특징들이 유닛들 집단 내의 일정한 활동 패턴으로 표상되는 '벡터 부호화'(vector coding)의 특성을 지닌다(7장 5절). 여기서 벡터가 가지는 자리값들은 매개변

수공간의 특정 값들과 대응하며, 벡터의 요소들이 클수록 매개변수공간의 차원은 커진다.

우리가, 그리고 인공지능이 얼굴을 인식할 때는 몇 차원 정도가 필요할까? 처칠랜드는 논의의 편의상 그의 책에서 코 너비, 입 크기, 눈 간격이라는 삼차원의 예를 그림(437쪽의 그림 7.10과 447쪽의 그림 7.15)으로 제시하고 있지만 실제로는 얼굴의 정확한 인식에 무수히 많은 차원이 필요할 것이다. 이러한 차원들은 매개변수공간에서의 대응이나 경로, 부피 등을 고려하도록 하는데, 이를 통해서 여러 특징들의 표상이나 유사성 관계를 설명할 수 있다.

여기까지는 얼굴 인식에 있어서 뇌가 표상하는 방식이다. 그런데, 처칠랜드는 더 나아가서 실제로 우리는 특정 얼굴을 어떻게 바로 '그 얼굴'로 재인(recognition)할 수 있는지의 문제를 다룬다. 인공지능이 여러 얼굴들을 학습해서 그 얼굴이 여성인지 남성인지 그리고 예전에 봤던 그 사람인지 어떻게 아는가? 처칠랜드에 따르면, 얼굴에 대한 인공신경망에서는 시냅스 연결들을 약하게 하거나 강하게 해서 요소 벡터를 변화하게 하고 이를 통해 학습이 이루어진다.

처칠랜드가 한걸음 더 나아가 제기하는 문제는 한 사람이 다른 표정이나 각도로 있을 때 그 다양성을 모두 어떻게 바로 그 사람으로 재인하는가이다. 또, 인공신경망이 얼굴들을 학습한 후 학습하지 않은 얼굴을 어떻게 여성/남성, 얼굴/물체 중 하나로 구분할 수 있는가이다. 이는 인공지능에서 처음 문제가 제기된 것이 아니라 철학에서 학습과 재인, 문제해결 그리고 일반화와 깊이 관련된 문제이다. 실제 인공지능은 얼굴인식에 대해 우수한 성적을 거두었고 심지어 색의 항등성이 개입된 얼굴인식에 있어서도 성공적이었다. 이것이 어떻게 가능한지를 신경연결망의 견지에서 설명한다면 이는 또한 '개념'과 '일반화'에 대한 새로운 철학적 설명을 제공하는 셈이 된다.

이에 대한 처칠랜드의 보다 구체적인 문제제기는 다음과 같다. 벡터가 가지는 자리값들이 대응하는 매개변수공간은 다른 특징들의 정보에 대해 다른 가중치가 조성될 것이다. 그렇다면 매개변수공간 안에서 조성되는 다른 행렬들의 둘 혹은 그 이상의 집합들은 어떻게 동일한 얼

굴이라는 기준을 '파악'(의인화하여) 할 수 있는가? 제주도에 사는 영희의 다른 각도의 두 사진은 다른 가중치 조성 때문에 신경활동 공간 안에서 시냅스의 양상이 다를 수는 있을 것이다. 그러나 그럼에도 불구하고 신경활동 공간에 분할구역은 동등하게 형성되며 이 점이 그 대상을 특정 범주로 정확히 파악하게 해준다는 것이다. 개념과 일반화에서 '동일한 표상'은 여러 신경그물망이 학습한 세부 사항들이 다름에도 불구하고 상당히 유사한 위상공간 내의 위치를 가리키기 때문에 가능하다. 처칠랜드의 이러한 설명은 범주적, 개념적 유사성에 대한 해명을 가능하게 한다. 예를 들어 처칠랜드의 '신경망 활성화'는 구글이 최근 개발한 얼굴 인식 인공지능의 원리를 설명해 준다. 처칠랜드의 설명은 인공지능의 경우에도 인간 인지의 실제 상황에서와 마찬가지로 일부의 정보만으로도 충분한 인식이 가능하다는 사실을 설명해 준다. 부분적 입력 이미지에 대해서 벡터 완성이 실현되고, 입력 정보가 모자라는 경우 결핍은 얼굴에 대한 일반적인 지식 정보에 의해 보충된다. 이렇게 처칠랜드의 신경망 활성화 원형이론은 얼굴 인식뿐만 아니라 다양한 종류의 지각의 구조, 대상 항상성, 그리고 가설연역적 추리의 기제 등 설명하기 어려운 문제로 꼽혀 온 주제들에 새로운 빛을 던져준다.

4. 뇌-중심주의를 넘어선 신경기반의 뇌-친화주의

마음의 작용을 뇌로 온전히 설명할 수 있는가 아닌가는 오래된 물음이다. 처칠랜드의 신경철학은 뇌신경망을 기반으로 설명하고 있기에 뇌 중심주의적 설명이라고 생각할 수 있다. 그러나 또 한 번 처칠랜드는 우리의 상식적인 독해를 넘어선다.

앞 절에서 살펴본 신경망들의 대응도들 사이의 연결과 조율 그리고 재구성을 통해 설명할 수 있는 것은 개념, 일반화뿐만이 아니다. 개념 파악과 일반화를 통해 가능하게 되는 이론간 환원이나 패러다임 전환도 설명가능하며, 전통 인식론에서 대표적인 두 경쟁 이론들인 진리 대응설이나 진리 정합설의 구분은 의미가 없어진다. 이는 단지 뇌 신

경상태 안에서는 모든 것의 구분이 없어진다는 의미가 아니며, 더 깊은 철학적 함의를 가진다. 진리 대응설과 진리 정합설은 진리라는 개념에 대한 언어분석을 통해 갈라져 나온 이론들이다. 그런데 신경활성화의 원형과 대응도를 통해 설명된 두 이론 간에 구분이 없다는 것은 두 이론이 동일하다는 의미라기보다는 사실상 "인지 기능이 언어 분석에 토대하지 않는다."는 의미이다. 인지 기능은 언어가 아니라 오히려 신경망에 기초하고 있기 때문이다. 처칠랜드가 말하고자 하는 바는 여기서 더 나아간다. 인지 기능이 신경망을 기본 토대로 하고 있다면, 인간이 개념, 인지, 생각을 하기 위한 조건은—전통적인 생각과 달리—언어가 선행해야 할 필요가 없다는 것이다.

이렇게 처칠랜드의 신경철학은 기존 분석철학의 인식론이 전개되는 기본 틀과는 완전히 다른 분석 틀과 지형을 보여준다. 이에 대해 전통적 인식론의 입장에서는 다음과 같이 반박할 수 있다. 처칠랜드의 인식론적 생각은 통속심리학의 수준(personal-level)이 아니라, 신경 상태라는 그 보다 하위 수준(sub-personal level)에서 논의되는 것이므로 기존 인식론이 논의되는 차원과는 독립적이라고 반박할 수 있다. 그러나 처칠랜드의 신경철학에서 '대응도들 사이의 포섭관계'는 단지 신경 상태를 기술하는 차원이 아니라, 지식이 형성되는 과정 즉 인간의 추리과정에 대한 설명이다. 즉, 신경망에 기반한 처칠랜드의 설명은 지식형성과 대응되는 하위차원 즉 신경 상태를 단순히 '기술'한 차원이 아니다. 이런 점에서 처칠랜드의 신경철학은 '뇌'라는 하위차원에 기반을 둔 설명이지만 피설명항과 설명항은 모두 '하위차원'에 머물러 있지 않고 인간의 추리과정을 직접 다루고 있다. 따라서, 처칠랜드의 신경철학은 단순히 기존의 인식론과 다른 차원의 설명이 아니라 경쟁적인 대안적 설명으로 제시될 수 있다.

그 구체적 예는 논리학과 과학적 설명을 가로지르는 주제들인 "비연역적 추리가 어떻게 가능한가?"에 대한 설명이다. 우리는 어떻게 기존 경험적 자료에 가정되어 있지 않은 새로운 가설을 제안할 수 있을까? 즉, 가추적(abductive) 추리는 어떻게 가능한가? 이 물음은 과학적 발견의 논리나 예술에서의 창의성에 대한 논의가 꽤 오랜 역사를 가져왔음

에도 불구하고 구체적으로 탐구되기 어려운 주제였다. 처칠랜드의 설명에 따르면, 병렬로 연결되는 신경망은 계층 구조를 이루면서, 추상적 개념의 점진적 계층, 즉 여러 추상적 정도를 산출한다. 또한 상위 층이 하위 층으로 재귀적 회로를 방사적으로 연결함으로써, 하위 층의 일부 입력 정보만으로도 상위 층의 신경망 전체를 발화하게 만들 수 있다. 이러한 재귀적 구조에 의해서 신경망은 다소 부족한 입력 정보를 가지고도 충분히 인식하고 예측할 기능을 발휘할 수 있다. 이런 구조가 갖는 기능은 바로 과학자와 예술가들이 창의적 아이디어를 고안하게 만드는 은유(metaphor) 혹은 가추추론(abduction)을 가능하게 해준다.

이러한 재귀적 구조는 뇌의 부분적 신경망들이 상호연결을 통해서, 본래적 기능과 다른 정보 처리에 활용될 수 있도록 해준다. 이것을 '개념의 재배치'(redeployment)라고 부른다. 어려서부터 우리가 학습을 통해서 습득한 신경망의 원형들, 즉 개념과 일반화(가설) 등은, 이러한 재배치를 통해서 지금까지 적용하지 않았던 신경망 원형들과 연결될 수 있다. 이러한 재배치를 통해 우리는 이전과 다른 관점에서 세계를 바라볼 눈을 가질 수 있으며, 새롭게 예측할 능력을 발휘할 수 있다. 이는 단순히 과학에서의 추론뿐만 아니라 인공지능의 시대에 초미의 관심사로 떠오르는 '창의성'에 대한 이해 또한 가능하게 한다.

이러한 처칠랜드의 설명은 인지심리학자들, 뇌신경과학자들이 여러 인지현상들을 설명하는 것과 어떻게 다른가? 처칠랜드의 신경철학은 흔히 뇌의 상태에 견주어 여러 인지적 사회적 현상을 설명하는 '뇌 중심' 혹은 '뇌 상태로의 환원'적 견해와는 차별점을 지니는 것으로 봐야 한다. 뇌 중심의 설명은 결국 인간의 뇌를 기준으로 한 설명이므로 인지에 대한 설명에 있어서 인간중심주의적인 시각을 넘어설 수 없다. 반면 처칠랜드의 신경철학적 설명은 '뇌-친화적 접근'(brain-friendly approach)으로 그 설명은 인간중심주의를 넘어선다. 최근의 연구 "Cognition without Cortex"(Güntürkün & Bugnyar 2016)에 따르면, 인간처럼 체계화되고 세련된 언어를 가지고 있지 않은 새는 대뇌 피질을 가지고 있지 않음에도 불구하고, 신경 연결만으로도 고차적 인지 기능을 수행한다. 인간의 마음에 대한 논의가 신피질의 기능을 중심으

로 한 설명에 국한되는 점을 고려하면 이는 획기적인 연구 결과이다.

그러나 동물 인지의 이런 현상은 처칠랜드의 신경망 기반 인식론, 구체적으로는 뇌-친화적 접근법으로부터 설명할 수 있다. 처칠랜드는 벌의 뇌에서 발견된 지극히 단순한 형태의 뉴런(vum)이 가중치를 변화시켜서 꽃의 꿀에 대한 강화학습을 중재한다는 점을 예로 들어(8장) 최근의 새로운 연구결과들을 예견하고 있다. 이 점은 처칠랜드의 견해가 단순히 뇌 상태와 인지 상태를 연결하여 설명하는 뇌-중심적 접근과는 차별화된다는 점을 잘 보여준다. 먼저, 처칠랜드의 뇌-친화적 접근법(brain-friendly approach, 7장)은 실재 표상이 문장 구성능력에 기반한다는 전통적인 인식론적 접근을 거부하므로, 언어능력을 갖지 못한 동물들도 표상을 가질 수 있다는 점을 받아들일 수 있다. 더 나아가 처칠랜드는 직접적으로 동물의 학습에 대해 주장한다. 처칠랜드에 따르면, 신경망의 "재강화 학습은 포유류를 포함하여 넓은 범위에 적용"[1]된다고 하는데, 이 능력은 단순히 자동적인 강화과정 뿐만 아니라 우연적 상호관계와 인과적 상호관계를 예측하는 수준까지 가능하다. 신피질을 가지고 있지 않은 새가 인간과 유사한 수준으로 인지기능을 수행한다는 최근의 연구결과는 인지 기능의 기초가 언어가 아니라 신경망이라는 처칠랜드의 주장에 대한 경험적 증거이다. 이렇게 언어분석 기반의 인지 설명을 벗어나서 신경망 기반의 인지 설명을 받아들이게 되면 어떤 실제적인 변화가 있을까? 자연지능을 가지는 인간과 동물 그리고 인공신경망에 기반한 인공지능의 존재를 불연속성보다는 연속성의 관점에서 이해하게 됨으로써 '의식', '인지', '마음'의 담지자와 절차에 대한 설명이 다른 차원에서 이루어지게 될 것임을 예상할 수 있다.

5. 경험적 인식론으로서의 신경철학: 시대적 의의

처칠랜드의 책 *Brain-Wise*는 신경테크놀로지와 인공지능이 사회의

---

[1] p. 498.

전 영역을 지배하는 이 시대에 많은 생각할 거리—그것이 철학적이든 일상적인 문제해결이나 교육의 문제이든—를 던져준다. 처칠랜드가 말하고자 하는 궁극적 메시지는 무엇인가?

처칠랜드는 철학의 여러 전통적인 개념부터 자유의지, 종교까지 철학의 거의 모든 논의 주제들을 망라하여 다루지만, 이들을 설명하는 데 기반이 되는 것은 무엇보다도 인식론적 시각이다. 플라톤의『테아이테토스』편 이래로 인식론에서 지식은 '정당화된 참된 믿음'으로 간주되어왔다. 거짓일 가능성을 배제할 수 있는 참된 지식은 오류가능성이 있는 지각, 감각적 믿음으로부터는 얻어질 수 없으므로, 어떤 대상에 대한 지식은 '선험적'이어야 한다는 것은 자연스런 추론이었을 것이다. 이러한 생각은 칸트의 범주론에서 정점을 찍는다. 이러한 생각은 전통적 인식론의 분석/종합 판단의 구분을 비판하고 자연주의 인식론을 제시한 콰인의 공헌과 괴델의 불완전성 정리에 의해 체계 내적으로 불가능한 것으로 밝혀졌고, 이러한 콰인의 기본 입장을 처칠랜드는 신경연결망이라는 설명항을 사용해서 현대적으로 구체화하고 있다.

처칠랜드의 이러한 연구 방향이 언어의 의미 분석에 기반한 논리경험주의와 분석철학의 무용성을 의미하는 것인가? 그것은 아니다. 처칠랜드의 메시지는 오히려 우리의 기본적이고 상식적인 현실로 복귀하라는 것이다. 인간의 관찰도구인 감각과 확장된 감각인 과학의 도구들이 아무리 이론의존적일지라도, 또 플라톤의 동굴 안의 죄수들의 인식처럼 불완전하고 빈약할지라도, 실험이나 경험적 증거 없이 인식을 이야기하는 것은 그저 '선험'이라는 정당화되지 않은 미신에 의존하여 현실에 등 돌리는 일이다. 처칠랜드는 이를 전통적인 언어분석철학이 의존하는 '사고실험'에 대한 파이어아벤트의 비판을 인용하여 전통적 인식론의 가정이 독단임을 주장한다. "'x'의 실제 의미에 대한 분석은 일부 사람들이 **특정 장소와 시간에 x인 것들**에 관해 무엇을 **믿는지**를 말해줄 뿐이다."[2] 이는 구체적 근거 없는 선험성이라는 구시대의 환영을 좇는 이들에 대한 부드러운 충고이다.

---

2) p. 401, 서평자의 강조.

참고문헌

Churchland, P. (2002), *Brain-wise: Studies in Neurophilosophy*, MIT Press.

Churchland, P. & Sejnowski, T. (1990), “Neural Representation and Neural Computation”, *Philosophical Perspectives*, 4: pp. 343-82.

Güntürkün, O. & Bugnyar, T. (2016), “Cognition without Cortex”, *Trends in Cognitive Sciences*, 20(4): pp. 291-303.

서평 투고일	2016. 07. 24.
게재 확정일	2016. 07. 27.