인지과학, 제23권 제2호
Korean Journal of Cognitive Science
2012, Vol. 23, No. 2, 133～164.

처칠랜드의 표상이론과 의미론적 유사성

박 제 윤†

한양대학교 교양교육원

폴 처칠랜드는 인지신경생물학과 연결주의 AI의 연구 성과로부터, 상태공간 표상이론을 제안하였다. 그 표상이론에 따르면 세계의 다양한 현상들에 대한 우리의 표상은 신경세포 또는 신경세포 집단의 활동에 대응하는 위상 상태공간의 지점들로 재현될 수 있다. 그러한 표상 체계를 모의하는 연결주의 AI 신경망은 은닉유닛들 속에 우리가 세계를 인지할 의미론의 범주 체계를 담아내는 것으로 해석된다. 그러한 해석의 관점에 따르면, 신경망은 세계에 대한 범주체계를 은닉 유닛들이 갖는 위상 상태공간의 특정 지점이라고 주장한다. 그러나 포도와 르포르는 그러한 전망을 어둡게 본다. 그의 전망에 따르면, '차원의 개별화' 가능성, 내용 동일성의 '분석/종합 구분'의 측면, 그리고 '부차적 정보'에 따른 상태공간의 상이성 등을 고려할 때 새로운 표상이론은 의미론적 내용 동일성을 주장하기 어렵다. 그 상태공간 표상은 유사성의 기준을 전제하기 때문이다. 본 논문은 처칠랜드 표상이론의 제안과 포도와 르포르의 비판 중에 어느 것이 더 설득력을 갖는지 검토한다. 상태공간 표상이론에 대한 필자의 이해에 따르면, 인공 그물망은 학습알고리즘에 따라서 스스로 내용 유사성의 분별 기준을 조성한다. 이러한 근거에서 포도와 르포르의 지적은 처칠랜드 표상이론에 대한 적절한 공격이 되지 못한다. 또한 그 표상이론은 미래의 인공지능 시스템이 의식 이하의 수준에서 세계를 인지할 개념체계를 어떻게 담아낼 수 있을지 우리에게 이해를 제공한다. 따라서 우리는 앞으로 인지과학 연구의 초점을 무엇에 집중해야 할지도 전망할 수 있게 되었다.

주제어 : 폴 처칠랜드, 상태공간 표상이론, 의미론, 유사성의 기준

---

* 유익한 심사평과 오류를 지적해 주신 익명의 세 심사위원에게 감사드린다.

† 교신저자: 박제윤, 한양대학교 교양교육원 (405-734) 인천 남동구 구월2동 롯데캐슬 2303-2401
Tel: 010-2958-8785, 023-422-8785, E-mail: jeyounp@hanmail.net

서론: 폴 처칠랜드의 표상 이론과 그 함축

폴 처칠랜드(Paul Churchland 1989)[1]는 인지신경생물학(cognitive neurobiology)과 연결주의 인공지능(connectionist AI) 등의 최근 연구 성과들에 근거하여[1] 새로운 표상이론을 제안한다.[2] 그 새로운 표상이론의 제안에 따르면, 세계의 다양한 현상들에 대한 표상들은 신경세포 또는 신경세포집단 활동에 대응하는 위상 상태 공간의 특정 지점들로 표현된다. 즉, 신경세포 집단은 "상태공간 표상(state space representation)"을 갖는다. 따라서 신경계 작용에 의한 표상들의 추상적 이행은 위상 상태들 사이의 좌표변환으로 파악된다. 이러한 새로운 표상이론은 지금까지 우리

1) 본 논문은 폴 처칠랜드의 저서[1]의 내용을 중심으로 다룬다. 그럼에도 불구하고 본 논문의 제목을 "폴 처칠랜드의 표상이론"이라 칭하지 않고, "처칠랜드의 표상이론"이라 칭했다. 그 이유는 다음과 같다. 폴은 주로 연결주의 인공지능(connectionist AI)에, 페트리샤는 주로 인지신경생물학(cognitive neurobiology)에 관심을 기울여왔다. 그리고 그들 부부는 각자의 연구를 신경철학(neurophilosophy)이란 공동 주제로 통합한다. 따라서 그들의 철학적 관점을 명확히 구분하는 것은 어려울 뿐만 아니라 불필요하다. 폴의 생각 속에 페트리샤의 생각이 있으며, 페트리샤의 생각 속에 폴의 생각이 들어있기 때문이다. 따라서 본 논문은 폴의 저작을 넘어 페트리샤의 저작도 자유롭게 인용할 것이다. 나아가서 본 논문 내에서 그들을 명확히 구분할 필요가 없다면 "처칠랜드"로 칭하겠다.

2) 일찍이 (계산주의 모델과 패러다임을 달리하는) 여러 연결주의 모델 제안이 있었으며, 그것에 대한 다양한 저항들이 있었다[2]. 그 중에 연결주의 모델이 언어적 표상을 어떻게 구현할 수 있을지에 주목하는 연구로 스몰렌스키(P. Smolensky 1990)[3]의 연구가 있다. 그 연구에 따르면, 신경망은 언어의 활자에 대응하는 유닛들로 구성된다. 그 연구의 제안에 대해 포도와 필리쉰(J. Fodor & Z. Pylyshyn 1988)[4]은 다음과 같은 문제를 지적한다. 일상 언어는 구성적 특징을 가지지만, 연결주의는 그것을 구현할 수 없거나, 구현한다면 결국 계산주의 모델을 실현을 설명해주는 모델에 불과할 것이다[5].

반면, 처칠랜드의 표상이론은 신경망의 언어적 구현 가능성보다, (인간을 포함한) 모든 동물 일반의 신경계가 세계에 대한 인지적 표상을 어떻게 가질지, 즉 신경계 일반의 인지 가능성을 설명하려는 의도에서 출발한다. 따라서 처칠랜드의 표상이론은 표상을 숫자들의 조합으로 파악한다. 처칠랜드는 그 표상이론을 통해 과학철학의 전통적 쟁점들, 과학의 개념, 과학의 성장, 이론의 가능성(즉, 예측 행동을 가능하게 하는 비언어적 일반화의 가능성) 등을 어떻게 설명할지에 관심을 갖는다. 그러한 측면에서 처칠랜드의 표상이론은 연결주의 줄에 서지만 위의 쟁점과 거리가 멀며, 따라서 본 논문의 논의 또한 위의 쟁점과 거리가 있다.

가 언어로 명확히 밝히기 어려웠던 주관적 감각질(qualia)의 표상 능력에 대해서 객관적으로 언급할 방법을 제공하며, 표상들 사이의 내용적 그리고 인과적 관계에 대해 계량적으로 분석하고 이해할 접근법을 제공한다[1][6].

처칠랜드가 상태공간 표상이론을 제안하면서 밝혀보려는 궁극적 목표는 아래 세 질문들에 대한 대답이다.

첫째, 뇌가 세계를 어떻게 표상하는가?

둘째, 뇌는 표상들을 어떻게 계산처리(computation) 하며, 어떻게 감각운동 조절(sensorimotor coordination)을 성취하는가?

셋째, 그러한 성취를 뇌의 미시-물리적 조직(microphysical organization)이 어떻게 이행하는가?

위의 세 질문들은 아래와 같이 관련된다. 일차적으로 신경계 세포들이 어떤 미시 기능적 구조를 가졌는지 이해하고 나서야, 그 이해를 바탕으로 신경계의 운동조절이 어떤 계산처리 방식을 따르는지 밝혀볼 수 있을 것이며, 그러한 두 이해에 기초하여 마침내 뇌가 세계를 표상하는 일반적 방식이 무엇일지 이론적 원리를 얻을 것이다. 그러므로 위의 세 질문에 대한 대답은 질문의 역순으로 이루질 필요가 있다.

우선적으로, 인지신경생물학의 연구 성과에 따라 처칠랜드가 파악하는 뇌의 기초 구조와 그 작용은 아래와 같다. 미시물리학적 관점에서, 대뇌 피질은 대략 1mm 두께의 층판 구조물이다. 그 층판은 수직으로 구분되는 개별 모듈(modules)로 구성되며, 이것을 원주(columns)라 부른다. 그리고 그 원주는 대략 9천개의 신경세포로 구성된다. 원주들 사이의 신호 전달은 기본적으로 피질을 수평으로 가로지르는 섬유들이 담당한다[7]. 그러한 수평 신호전달 체계를 통해 연결되는 원주들은 유사한 조직 구조를 가진 것들끼리 집단을 이루어 유사한 기능을 수행한다. 대뇌피질의 경우 이 집단을 브로드만 영역(Brodmann' areas)으로 명확히 구분된다. 결국 유사 원주들 집단은 뇌에 특정한 기능들을 수행하는 국소 기능 대응도(topographic maps)[3]

3) 일반적으로 국내 번역서들은 뇌 피질의 “map”을 “지도” 또는 “맵”으로 번역하고 있으나, 그 단어의 온전한 영어 표현은 “topographic map”이다. 그것이 신체의 감각 또는 운동의

를 조성한다. 이러한 조직적 구조는 대뇌피질에 뚜렷이 드러나지만, 뇌 전체의 일반적 조직 구조로 밝혀졌다.

다음으로, 처칠랜드의 특별한 관심은 그 대응도의 구조와 기능에 대한 인지신경생물학의 연구 성과이다. 많은 동물들(인간을 포함하여)의 뇌가 일반적으로 대응도 조직을 갖는 이유는 무엇이며, 대응도가 세계에 대한 감각의 표상과 운동 조절에 어떻게 이용되는지 등의 의문에 대답하려면 그 구조를 면밀히 살펴볼 필요가 있다. 대응도들은 각기 주어진 역할에 따라서 감각 입력정보들을 수용하고, 그것들로부터 운동 출력정보들을 산출하는 특정의 기능적(함수적) 패턴을 갖는다. 따라서 일반적으로 고등동물의 신경계를 대응도들 사이의 병렬연결을 통해 (특정 기능을 수행하도록) 상호적으로 계산처리하는 계산기라고 할 수 있다. 그렇게 신경계가 감각 입력을 대응도의 기능적 패턴에 따라 운동 출력으로 변환하는 기능을 수행한다고 본다면, 신경계의 입・출력 정보들의 변환은 대응도에 의한 행렬들의 계산처리로 해석된다. 그러므로 만약 우리가 대응도들(또는 원주 집단들)의 입・출력 정보들을 숫자 조합인 행렬로 표현한다면, 세계의 특성에 대한 감각입력 표현과 운동 출력 표현을 모두 객관적이며 계량적으로 해석할 수 있다. 또한, 만약 우리가 입・출력 정보인 숫자 조합들을 상태공간의 지점(벡터)으로 표현한다면, 대응도의 기능을 상태공간 내의 좌표(벡터) 변환으로 볼 수 있다.

뇌 대응도들이 그러한 계산처리 방식으로 감각 정보로부터 어떻게 운동 조절을 성취해낼지 폴 처칠랜드는 다음과 같이 가설적으로 설명한다. 그는 대응도의 가장 단순한 인공적 구조로 가상적 '게(crab)' 로봇 만화 이야기를 제안한다. 그 가상적 이야기에 따르면, 그림 1a, 그림 1b와 같이 '게'의 두 눈의 각도(α, β)는 두 숫자의 조합 또는 2차원 위상공간의 지점(벡터)로 표현될 수 있으며, 두 팔의 위치관절의 각도(Θ, ϕ) 또한 두 숫자의 조합 또는 2차원 위상공간의 지점으로 표현될 수 있다. 만약 '게'가 눈의 감각 입력 정보로부터 팔의 목표 위치인 운동출력을 산출하려면, 그 '게'는 서로 다른 두 좌표를 변환시키는 계산기를 가져야 한다. 그리고 만약 '게'의 신경계가 그림 2와 같이 눈과 팔을 연결하는 대응도로 기능하여, 입력 정보를 출력 정보로 변환, 즉 벡터 변환(vector transformation)을 이행한다면, 그 '게'는

기능적 모습에 대응하는 동시에 일종의 함수적 계산을 수행한다는 점에서, 필자는 "국소 기능 대응도"란 번역이 적절하다고 여기며, 이후로는 간단히 "대응도"라 칭한다.

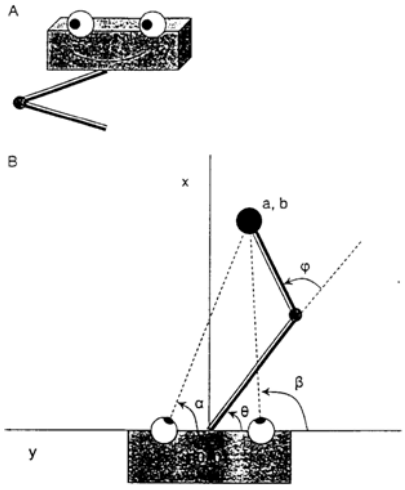

그림 1a. 감각운동조절(sensory-motor coordination)의 아주 단순한 문제를 표현하는 그림. 점선으로 표현된 두 눈이 가리키는 표적은 두 눈의 각도(α, β)로 표현되며, 그 표적에 미친 팔의 위치 역시 관절의 각도(Θ, ϕ)로 표현될 수 있다.

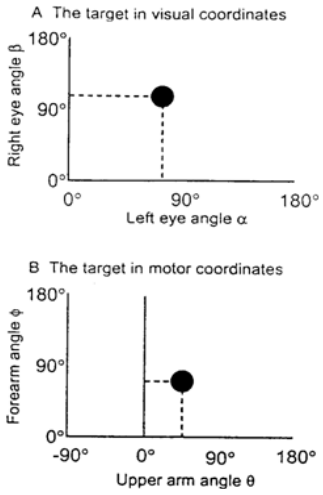

그림 1b. 감각과 운동 시스템 각각의 모습에 대응하는 표상은 좌표공간 A와 B 내에 적절한 점으로 표현될 수 있다. 시각공간에서의 표적의 위치는 운동공간에서의 그 위치와 동일하지 않다. 따라서 눈이 본 것에 팔이 도달하기 위해서는 시각공간이 운동공간으로 좌표변환 되어야 한다.

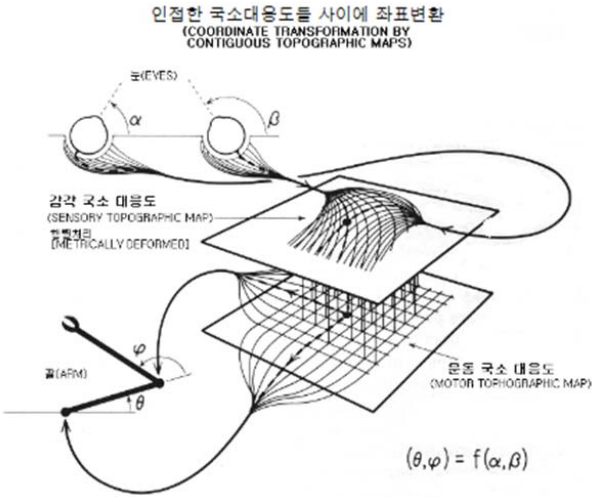

그림 2. 물리적으로 이행된 좌표변환. 통합 눈 위치는 동시 자극에 의해 상위 대응도(map)의 단일 교차점에 등록된다. 앤드게이트(and-gates)처럼 활동하는 수직 연결은 정확히 그 위치로부터 아래의 운동대응도로 신호를 내려 보낸다. 그럼으로써 양 교차운동섬유 외부로 활동이 유도되며, 그 신호를 담은 섬유들에 적절하게 각 팔 관절이 각도를 맞출 수 있다. (폴 처칠랜드는 같은 책 뒤, 92쪽에서 대응도들 사이의 연결을 단일선이 아니라 복수의 선으로 연결된 것으로 보아야 한다고 이 그림의 이해를 수정한다. 그렇지만 그 수정이 자신의 기본적 관점 자체를 바꾼 것은 아니라고 밝힌다[1].)

눈이 가리키는 곳으로 팔을 옮겨가는 동작을 자동적으로 실행할 수 있다. 다시 말해서, 대응도에 조성되는 신경세포들의 연결 상태 그리고 대응도들 사이에 조성되는 연결 상태(패턴)는 새로운 입력 숫자조합들을 출력 숫자조합들로 변환시키게 하는 동력학적 함수적 기능이다[8][9][10].

이러한 가상의 단순한 '게' 이야기를 확대 해석해보며, 동물들이 다양한 입력 신호들을 어떻게 신체운동 출력 신호로 변환할 수 있으며, 그 변환이 신체 여러 부분의 자동적 운동조절을 어떻게 성취해낼지 우리는 별로 어려움 없이 다음과 같이 상상할 수 있다. 실제 신경계의 대응도들이 이행하는 계산처리는 다중적 감각 자료들을 입력으로 받아들여 다중적 행동 출력으로 산출하는 방식으로 분석된다. 그림 3에서 볼 수 있듯이, 실제 신경계에 분산적으로 존재하는 대응도들은 상호 연

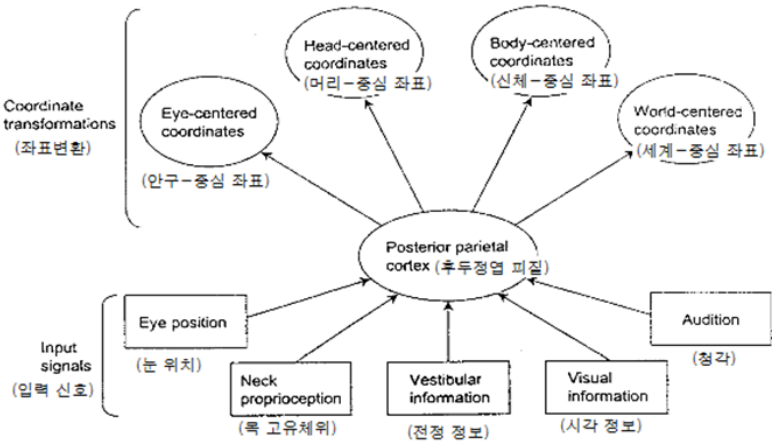

그림 3. 안구위상시각정보(retinotopic visual information)를 상위계열 항목 단위정보(higher-order reference frames)로 전환하는 후두정엽피질(posterior parietal cortex)의 역할을 보여주는 그림. 눈 위치, 머리 위치, 시선 위치 등은 안구 정보를 수정하는데 이용된다. 그러므로 후두정엽피질은 시각과 청각정보를 눈, 머리, 신체, 세계 등에 대한 좌표 단위정보(coordinate frames)로 전환하는 중간단계를 제공한다[6][10].

결을 통해 정보를 통합하거나 분할하여, 다중적 기능을 원활히 수행한다[10]. (이렇게 다중 입력정보의 통합 가능성에 대한 해명은 '다양한 감각을 어떻게 통일하는 것이 가능한지' 칸트가 설명하고 싶어 했던 의문에 대한 새로운 대답이다.) 특별히 브로드만 영역 5번과 7번은 여러 다른 감각 입력정보들에 대한 통합 부위로 잘 알려져 있다. 한 마디로 말해서, 뇌는 복잡한 정보처리를 대응도에 의해 분할하고 통합하는 방식으로 이행하며, 그 방식을 채용함으로써 신경계는 감각운동 조절을 간단하고 신속히 계산처리 해낸다.

끝으로, 처칠랜드는 뇌의 미시 물리적 구조의 작용과 대응도의 기능에 근거하여 신경계의 일반적 표상 원리를 밝히는 상태공간 표상이론을 제안한다. 그의 표상이론에 따르면, 신경세포 집단 또는 뇌 대응도가 세계로부터 여러 감각 수용기들을 통해 수용하는 특성들이 숫자 조합들로 표현될 수 있으며, 그 숫자 조합들은 상태공간의 특정 위치로 해석된다. 예를 들어, 색깔들은 망막의 세 원추세포들의 반응

에 의해 규정되는 3차원 상태공간의 특정 지점으로, 맛 감각은 혀의 여섯 종류 수용기들 반응에 의해 6차원 상태공간의 특정 지점으로, 그리고 머리 방향은 삼반규관 수용기 반응들에 의해 3차원의 상태공간의 특정 지점 등등으로 표현될 수 있다[11]. 이와 같이 세계의 다양한 정보들에 대한 뇌의 표상이 위상적 상태공간의 특정 지점이라고 해석되며, 나아가서 뇌는 그러한 일반적 표상능력을 갖는 것으로 파악된다. 처칠랜드는 이것을 "상태공간 표상이론(state space theory of representation)"이라 부른다. 이러한 신경계의 표상능력을 주장하는 이론은 우리 인간을 넘어 동물의 신경계가 일반적으로 갖는 표상 능력을 포괄한다.

이러한 상태공간 표상이론은 여러 전통 철학적 쟁점에 대한 새로운 해명을 제공한다. 우선적으로 심리철학의 쟁점에 새로운 전망을 하게 한다. 전통적으로 심신의 환원주의에 반대하는 이원론적 관점에 따르면, 감각질은 결코 물리적 상태로 환원가능하지 않은 주관적 느낌의 속성이다. 그러나 상태공간 표상이론의 관점에 따르면, 주관적 감각질은 신경세포의 반응에 따른 숫자들의 조합으로 표현되며, 따라서 객관적(계량적)으로 취급될 가능성이 있다. 그것을 처칠랜드는 구체적으로 아래와 같이 말한다.

> 한 사람의 현재 시각적 감각의 '말할 수 없는(ineffable)' 핑크색은 아마도 적절한 3중 피질 시스템에서 '95Hz/80Hz/80Hz 조합'으로 풍부하고 엄밀하게 표현될 수 있다[1].

또한 상태공간 표상이론은 존재론적이며 의미론적인 철학적 쟁점들에 대한 해명을 제공한다. 만약 신경계가 상태공간의 특정 위치를 특정 감각의 전형(prototype)으로 채용할 경우, 그리고 신경계가 그 전형을 세계를 분별할 기준으로 활용한다고 가정하면, 그 전형의 표상은 우리가 세계를 분별할 인지적 의미 '동일성' 내지 의미 '유사성의 기준(the criteria of similarity)'이라고 해석된다. 예를 들어, 그림 4와 같이 3차원의 색깔 상태공간에서 안구 망막의 세 종류의 수용기가 빛에 반응하는 정도의 지표를 세 축으로 해석할 경우, 특정 색깔은 그 상태공간의 특정 지점으로 표현된다. 이러한 이해의 측면에서, 우리에게 나타나는 어느 유사한 색깔들의 주관적 감각의 특징, 즉 감각질은 그 상태공간의 인접성으로 설명될 수 있다. 따라서

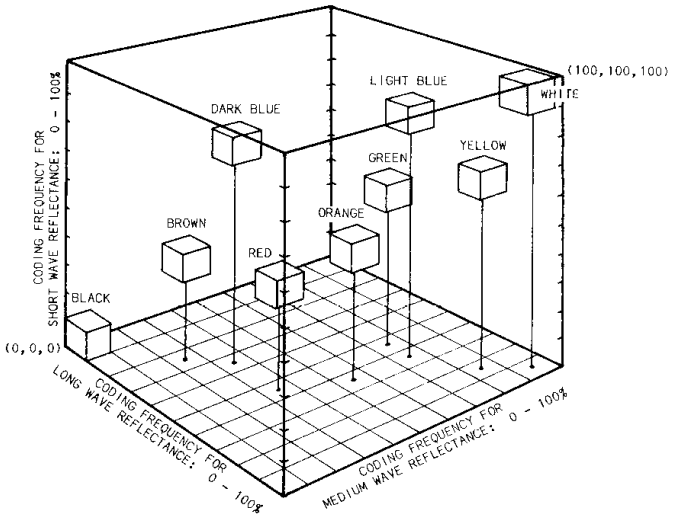

그림 4. 색깔 상태공간: 망막 내에 세 종류 원추세포의 반응 정도를 세 축의 특정 위치로 지정할 경우 다양한 색깔들은 3차원 상태공간의 특정 위치로 표현된다. 그리고 특정 위치에 인접한 것들은 서로 유사한 것들로 분류될 수 있다(1).

상태공간 내의 인접 위치는 어떤 색깔들을 그와 유사한 속성에 가까운 색깔로 분별하게 해줄 기준이다.

이러한 가정적 이해를 확장하여, 만약 그 대응도들이 계층적 구조를 이루며, 상위 대응도가 세계와의 직접 대면을 넘어서는 추상적 차원의 특징들을 전형으로 조성할 경우, 그것들은 우리가 '보편적(일반적) 개념'이라 칭하는 것에 상응하는 표상이라고 가정된다. 그러한 가정에 따라서 우리가 (추상적) 개념적 지식 또는 도식을 어떻게 가지는지도 가늠해볼 수 있다.

그러나 포더와 르포르(Jerry Fodor & Ernie Lepore, 1992[12], 1996a[13])에 따르면, 처칠랜드의 상태공간 표상이론은 '우리가 어떻게 내용 유사성의 표상을 구성할 수 있을지' 의문에 대답할 가능성이 없으며, 따라서 처칠랜드의 상태공간 표상이론은 결코 의미론적 유사성을 설명할 가능성이 없다. 그들 주장대로라면, 상태공간 표상이론은 하나의 의미론으로서 역할을 하지 못한다. 필자는 새로운 표상이론에 심각한 장애물을 확인하는 그들의 논의가 구체적으로 무엇인지 검토하고, 그들의 논의

를 처칠랜드의 표상이론이 어떻게 넘어설 수 있는지 밝혀볼 것이다. 나아가서 철학적 시각에서 바라볼 때 상태공간 표상이론이 제공하는 학문적 접근법이 인지과학 연구에 대해서 미래에 어떤 전망들을 제공할지도 (부차적으로) 간략히 논의할 것이다.

상태공간 의미론의 내용 유사성

포도와 르포르는 논문, "폴 처칠랜드와 상태공간 의미론"[13]에서 처칠랜드의 새로운 표상이론이 의미론적 동일성을 주장할 수 없음을 지적한다. 그들의 관측에 따르면, 상태공간 표상이론은 의미론적 '내용 유사성(content similarity)'의 강건한 개념어(robust notion)[4]를 제안하지 못하며, 따라서 그 표상이론이 의미론적 문제를 다룰 가능성이 의심된다. 그들이 갖는 의심은 아래 세 가지이다.

첫째, '차원의 개별화(the individuation of dimensions)'의 문제를 고려하여 제기되는 의심이다. 처칠랜드에 따르면, 앞의 그림 4에서 살펴보았듯이 의미론적 동일성은 동일 상태공간 내에 유사 위치를 점유하는 한에서 설명된다. 그의 상태공간 표상이론에 대한 포도와 르포르의 해석적 관점에서, 당신이 나와 유사한 의미론을 가진다는 것은 다음과 같이 이해된다. 당신과 내가 동일 차원의 상태공간을 가질 경우 오직 그 경우에만, 의미론적 '내용 유사성'이 확보될 수 있다. 그 조건은 개념들이 의미론적 동일 상태공간 내에 동일 속성들을 동일 차원으로 가질 경우이다. 이 조건은 분명히 의미 '유사성'의 개념어일지라도 의미 '동일성' 개념어에서의 문

4) 포더와 르포르가 굳이 개념(concept)이란 용어 대신에 "개념어(notion)"을 선택한 이유가 있을 것이다. 이것을 필자는 다음과 같이 파악한다. 그들의 관심은 일상 언어의 활용 측면에서 개념어들 사이의 의미론적 관계이다. 칸트적 의미에서 "총각"이란 개념어는 "결혼하지 않은 남자"란 의미를 함축한다. 그러한 개념어들 사이의 관계에 의해서 우리는 의미론적 동일성이 무엇인지 구분할 수 있다고 고려할 수 있을 것이다. 그러나 폴 처칠랜드가 상태공간 표상이론을 통해서 해명하려는 의미론적 동일성 또는 유사성의 문제의 핵심은 언어 이하 수준에서 인식적 측면이다. 우리가 세계를 인지하여 무엇을 분별할 인식적 동일성 또는 유사성이다. 아마도 그는 이러한 유사성이 장차 언어적 유사성을 설명할 기초가 될 것이라 가정할 것이다.

제를 불러일으키는 측면이 있다. 만약 두 단어가 의미하는 것이 무엇인지 기준을 갖지 못한다면, 우리는 '두 상태공간이 동일 차원을 갖는다'는 것이 무엇을 의미하는지 조차 알지 못할 것이기 때문이다. 포도와 르포르는 그 이유를 아래와 같이 구체적으로 밝힌다.

> 우리는 상태공간의 동일성을 위한 '기준'을 가질 경우에만(only if), 의미론적 유사성의 강건한 용어를 갖는다. 우리는 상태공간 '차원들의 동일성'을 위한 기준을 가질 경우에만, 상태공간의 동일성에 대한 기준을 갖는다.

위 지적의 관점에서 보면, 만약 우리가 '개념들이 어떻게 강건할 수 있는지' 설명하려 한다면, 처칠랜드의 상태공간 의미론은 전혀 도움이 되지 않는다. 상태공간 의미론이 '의미론적 유사성'을 주장하려면 '상태공간의 동일성'이 확보되어야 하며, 상태공간의 동일성을 이야기 하려면 상태공간 '차원의 동일성'이 확보되어야 하고, 상태공간 차원의 동일성을 이야기하려면 '속성들의 기준'이 제시되어야 하기 때문이다. 이것을 다음과 같이 구체적으로 (조금 더 쉽게) 설명할 수 있다. 만약 당신과 나의 상태공간이 다르다면, 그 어떤 경우에도 동일한 의미론적 유사성을 주장할 수 없을 것이다. 구체적으로, 만약 나의 신경계가 특정 개념에 대해서 2차원적 상태공간의 특정 지점을 가리키는 경우, 그것이 당신의 의미론과 동일한 것이기 위해서는 당신의 신경계도 동일한 2차원 상태공간의 동일 지점을 가리켜야 하며, 3차원 이면 안 된다. 또한 그 축들은 동일한 '속성들'을 나타낸 것이어야 한다. 나아가서, 만약 나의 신경계가 특정 개념을 개념공간의 특정 축 내에서 특정 간격으로 규정될 경우에, 그것이 당신의 신경계에도 동일한 것이기 위해서는 동일 축 내에서 동일 간격으로 표현되어야 할 것이다. 한마디로 당신과 나의 의미론적 동일성이 상태공간 표상이론으로 설명되려면, 완벽히 동일한 상태공간을 가지고 있다고 가정해야 한다. 그러한 고려에서 처칠랜드 상태공간 표상이론은 속성들의 분류 기준을 이미 가지고 있어야만 한다. 거꾸로 말해서 동일한 분류 기준을 가질 경우에만 두 사람은 동일한 상태공간을 갖는다고 주장할 수 있을 것이다. 이러한 고려에서 상태공간의 동일성 주장은 차원들의 분류 기준을 전제한다고 지적된다.

포더와 르포르에 따르면, 만약 상태공간의 차원들이 무엇을 표현하며, 그리고

그 차원들이 규정하는 상태공간 내의 영역들이 무엇을 표현하는지 등을 우리가 어떻게 결정하는지 처칠랜드에게 질문한다면, 결국 그는 단지 '약정'에 의한 제안 이외에 내놓을 것이 없을 것이다. 그러나 상태공간 차원들이 (그림 4에서와 같이 각 축마다 특정 색깔의 특성을 표현하는) 정신물리학적 지표들을 표현한다고 약정하는 것은 결코 의미론으로 받아들이기 어렵다. 인과적인 설명은 결코 심적 상태를 분류하는 내용이 될 수 없기 때문이다. 처칠랜드 관점에 따르면, 만약 특정의 자극이 그러그러한 정신물리학적 속성들을 가지며, 그 자극이 뇌에 가해진다면, 그 뇌는 그 특성을 표상한다고 주장할 것이다. 그러나 뇌는 '붉음'을 그것으로 표상하는 것이지, '그러그러한 파장 광선들의 조합'으로 표상하지는 않는다.[5] 따라서 처칠랜드의 상태공간이 의미론적 공간을 다룬다고 인정될 수 없다. 분명히 그는 상태공간이 어떤 종류의 유사성을 갖는지 결정하지 못할 것이며, 따라서 어떤 상태 공간이 무엇을 표상해야 할지 결정하지 못할 것이다. 이러한 고려에서 포도와 르포르는 처칠랜드가 의미론적인 문제와 정신물리학적인 기획을 혼동한다고 지적한다.

둘째, 내용 동일성의 문제를 분석/종합(판단) 구분(the analytic/synthetic distinction)의 측면에서 고려해보아도, 상태공간 표상이론이 의미론적 문제를 다룰 가능성이 의심된다. 만약 우리가 (콰인이 주장하듯이) 명제의 분석/종합 구분의 원리를 받아들이지 않는다면, '내용 동일성'을 주장할 근거를 상실할 것이기 때문이다. 예를 들어, 만약 우리가 "개는 동물이다."를 분석적이며, "개는 젖은 코를 갖는다."를 분석적이 아니라고 생각한다면, 처칠랜드의 표상적 상태공간에서 개념 '개'가 점유하는 공간은 '동물임'의 속성에 대응하는 차원을 가져야 할 것이지만, '젖은 코를 가짐'의 속성에 대응하는 차원을 가질 필요는 없을 것이다. 이것은 특정 개념의 구성적 공간이 어떤 차원의 공간을 포함하는지에 관한 문제이다. 그런데 만약 명제의 분석/종합 구분을 위한 어떤 원리적 방법도 없음을 확신한다면, 우리가 어떤 원리에 호소하여 의미론적 공간을 규정할 차원을 확정할 수 있을지 의문하지 않을 수 없다. 다시 말해서, 명제의 분석/종합 구분이 전혀 없다고 생각한다면, 어떻게

5) 이렇게 지적하는 측면에서 포도와 르포르는 우리가 세계에 대해 갖는 의식적 수준의 인지를 인지의 기본으로 가정하고 있음이 드러난다. 반면에 처칠랜드는 현재 신경과학적으로 의식을 명확히 설명할 수 없는 상황에서 우리의 인지를 무의식 수준의 정보처리로 바라보고 있는 입장의 차이를 볼 수 있다.

의미론적 공간, 즉 개념들 사이의 관계를 규정할 수 있을 것이며, 그 관계에 비추어 적절한 의미론적 차원들을 그렇지 않은 것들로부터 어떻게 구분할 수 있겠는지 의문된다.

물론, 처칠랜드는 개념들을 (그림 4에서와 같이 계량적인 방식으로) 상태공간에 유형틀(stereotype)로, 즉 전형(prototype)으로 규정한다. 그리고 그 전형적 표상, 즉 특정 표상에 대한 상태공간의 대표적 위치에 인접할 경우 의미론적으로 그러그러한 것으로 분별할 것이라고 말한다. 그러나 그 상태공간의 전형이 어떤 미덕을 갖더라도 그 공간의 차원들이 어떻게 개별화(규정)되는지를 설명할 수 없다면, 처칠랜드는 의미론의 문제를 피해가는 것이다. 결국 처칠랜드는 유형틀이 감각적 또는 정신물리학적 특징들을 표상한다는 가정에서 출발하지 않을 수 없다. 그러한 가정을 갖는다는 측면에서 보면, 결국 처칠랜드는 회피하려 애썼던 경험주의를 암묵적으로 인정하고 있다고, 즉 감각자료들로부터 추상적 개념들을 가질 수 있다고 낙관하는 입장에 선다고 포도와 르포르는 비판한다.

셋째, 의미론적 동일성 또는 유사성에 대해서 '부차 정보(collateral information)'의 문제를 고려해볼 때, 상태공간 이론이 어떻게 내용 유사성의 강건한 개념어를 제안할지 의심된다. 앞서 논의된 개별화의 차원을 처칠랜드가 어떻게든 해결한다고 하더라도 그것이 개념의 유사성을 설명할 가능성을 제공하기에 여전히 어려운 이유가 있다. 예를 들어, 만약 내가 가진 "개"의 의미론 상태공간이 당신이 가진 "개"의 의미론 상태공간과 유사한지는 여전히 어려운 문제이다. 그것은 나와 당신의 상태공간은 각기 가진 다른 믿음들의 하부 상태공간들을 포함하기 때문이다. 구체적으로 말하자면, 당신은 나의 믿음과 달리 "개"에 대해 "털가죽"에 대한 믿음을 포함할 수도 있다. 그리고 당신은 내가 가지고 있지 않은 개에 관한 아주 많은 다른 표준적 믿음들을 가질 수 있다. 그렇게 사람마다 서로 다른 믿음들을 가질 수 있으며, 그것은 사람들이 동일한 개념 "개"에 관해 저마다 아주 다양하고 다른 하부 상태 공간을 가질 수 있음을 의미한다.

이것을 다음과 같이 구체적으로 이해해볼 수 있다. 나의 개에 대한 여러 속성들, 즉 '털가죽', '뼈다귀를 좋아하는', '멍멍 짖는', '외부사람을 경계하는' 등등의 속성들은 모두 나의 신경계가 조성하는 상태공간에서 한 지점으로 규정될 수 있다. 그리고 그러한 상태들은 상위 차원의 상태공간 내에 한 축의 속성들을 부여할

것이며, 그 여러 축의 속성들이 통합되어 가리켜지는 특정 지점이 내가 의미하는 '개'라고 규정될 수 있다. 다시 말해서 내가 개에 대해 갖는 의미론적 상태공간은 개에 대한 여러 속성들의 종합이다. (이것은 하부 상태공간들의 벡터를 상위 상태공간의 벡터로 통합된다고 가정하면 어렵지 않게 이해된다.) 그러므로 나의 개에 대한 의미론적 상태공간은 그 속성을 나타내는 여러 하부 상태공간을 가진다. 그렇다면 개에 대해서 갖는 의미론적 속성들이 사람마다 다를 수 있다. 따라서 개념 "개"에 관해 우리가 개별화의 동일 차원들을 가진다고 하더라도, 그 하부 공간의 구조는 저마다 가지는 서로 다른 부차 정보에 따라서 아주 다를 것이다. 결국 그 하부 공간의 구조에 의해 지지되는 개념 "개"의 위치는 각각의 공간에서 상당히 다른 위치에 놓일 가능성이 있다.

사람마다 살아온 인생이 다르며, 따라서 저마다 특이한 경험은 특이의 부차 정보를 가지게 한다. 그것은 결국 특정의 상태공간 차원들이 서로 다르게 가중됨을 의미한다. 이러한 부차 정보 문제는 어느 누구의 개념도 서로 유사할 가능성을 배제한다. 그럼에도 불구하고 동일 차원의 상태공간을 가질 가능성은 특정 부차 정보를 더 중요하게 가중하는 경우에 가능할 수도 있다. 그러나 분석/종합 구분의 기준이 없이 그것을 어떻게 가능할지 설명할 방법은 없어 보인다.

이상과 같은 포도와 르포르의 비판적 관점에서 볼 때, 처칠랜드가 의미론적 내용 동일성의 문제를 상태공간의 동일(또는 유사) 위치로 바꾸는 방식으로 내용 동일성의 옛 문제들을 여전히 해결할 수 없을 것처럼 보인다. 그러한 상황에서 포도와 르포르는 결국 처칠랜드가 '지향적(의미론적)' 속성들을 제거하려는 경향을 가질 수밖에 없었을 것으로 본다. 다시 말해서, 처칠랜드가 선택할 수 있는 유일한 길은 우리의 모든 개념들이 정신물리학적 개념들의 함수라고 주장하는 것이다. 그러나 그러한 주장은 '개념 내용의 유사성'을 위한 기준이 어떻게 제공되는지 설명해야 하는 문제를 갖는다. 그들의 관점에 따르면, 처칠랜드는 경험주의자 가정의 한계를 다시 불러들인다. 다시 말해서, 처칠랜드의 표상이론은 특정 감각이 특정 피질의 반응 정도로 표현된다고 하지만, 그것들이 어떤 감각을 갖는지 알기 위해서 우리는 그것이 무엇과 같아야 하는지 이미 알고 있어야만 한다는 문제가 발생한다. 또한, 상태공간의 특정 위치에 대한 근접성을 내용 유사성의 기준으로 제시한다고 하더라도, 그것이 무엇과 같아야 하는지 분류의 기준이 먼저 전제되어야

할 것이다.[6] 한 마디로, 내용 동일성에 관한 처칠랜드의 설명은 선결문제 요구의 오류를 범한다.

처칠랜드는 위의 비판에 대한 첫 번째 방어 논문, “포도와 르포르: 상태-공간 의미론과 의미 전체론”[14]에서 신경망(neural network)에 대한 다음 두 가지 측면을 설명한다. 우선, 신경망이 활동 패턴의 동일성을 어떻게 획득하며, 따라서 활동 내용의 동일성을 어떻게 획득할 수 있는지, 즉 그 동일성이 수많은 특이적 신경망의 활동 패턴으로부터 어떻게 획득될 수 있는지 설명한다. 이어서 신경망의 은닉 유닛들 집단이 분산되고 모호하며 분명히 말할 수 없는 특징들에 의해 보다 큰 인지적이며 운동의 경제에 적절한 역할을 어떻게 수행할 수 있는지 설명한다. 그 두 측면의 설명은, 의미 유사성이 신경망에서 어떻게 구현 가능하며, 따라서 그 구현이 포도와 르포르가 지적하는 종류의 경험주의 문제를 어떻게 극복하는지, 이해를 제공하려는 시도이다.

그럼에도 불구하고 포도와 르포르는 두 번째 비판적 논문, “처칠랜드에 대한 응답”[15]에서 자신들의 입장을 변경하기 않고 아래와 같이 두 관점에서 다시 논박한다.

첫째, 대부분 개념들은 감각 개념으로부터 나온 구성물이 아니다. 감각 속성들에 의해 상태공간의 지표들을 설정하는 처칠랜드의 표상 이론은 개념들을 감각 지표들에 대응하는 상태공간의 위치로 규정한다. 이것은 경험주의가 복합 개념을 단순한 감각의 원천으로부터 구성된 것으로 생각하고 싶어 했던 것에 비유된다. 물론, 전체주의 관점에서 그물망은 경험주의자와 다른 종류의 개념 학습을 생각한다. 그렇지만 상태공간의 차원들이 감각적인 것들이라고 여겨지는 한에서, 연결주의자 표상이론은 경험주의자로 전락되며, 그 점에서 경험주의자 설명 이상의 흥미로움을 새롭게 제공하지 못한다.

둘째, 개념 ‘야옹이(kitten)’를 그것으로 결정하게 하는 것은 내용 ‘고양이’이지 내

6) 이 문장의 의미를 보다 쉽게 설명해보자. 경험주의자 입장에서는 신경망이 무엇을 규정할 계산적 구조를 가진다고 주장하겠지만, 신경망 자체가 무엇을 그것으로 규정하기 위해서는 그 분류의 기준을 이미 가지고 있을 경우에만 그 유사성의 구조가 조성될 수 있다. 이러한 측면을 고려할 때, 처칠랜드 상태공간 표상이론이 지지하는 신경망의 유사성 기준은 그 분류의 기준을 전제하는 경우에만 가능하다.

용 전형의 '고양이'는 아니다. 상태공간의 특정 위치를 원형으로 규정하는 상태공간 표상이론에 따르면, 만약 '야옹이' 원형을 그것으로 만드는 것은, 감각적 구성요소가 아니라, 특정 추론에서의 전형에 "따름(downstream)"의 역할이다. 그렇지만, 감각의 어떤 구성 요소들이 야옹이를 그것으로 결정하게 하는 것이 무엇인지 여전히 의문된다. 왜냐하면, 감각 요소들로부터 출현하는 전형은 전혀 내용 '고양이'의 결정자가 되지 못하기 때문이다.

결론적으로, 포도와 르포르의 지적에 따르면, 처칠랜드는 암묵적으로 경험주의자 가정을 끌어들이면서도 겉으로는 그 가정으로부터 자유로워진 척하는 것과 다름없다. 처칠랜드의 그러한 태도는 흄이 지적했던 경험주의자 한계, 즉 추상 개념들과 감각 개념들 사이의 관계를 어떻게 이해시켜줄지 설명하지 못할 것이라고 포도와 르포르는 비판한다.

경험주의를 넘어서는 연결주의 의미론

위의 비판에 대해 처칠랜드는 두 번째 방어 논문, "포도와 르포르에 대한 두 번째 응답"[16]에서 아래와 같이 주장한다. 포도와 르포르는 연결주의자 의미론에 대한 접근을 흄의 낡은 경험주의 개념과 다름없는 것으로 해석하지만, 자신은 그것에 아주 편안히 응답할 수 있다고 말한다. 그는 포도와 르포르가 연결주의자 프로그램을 올바로 바라보지 못하여, 그것을 흄의 한계에 빠지는 경험주의자로 몰아간다고 본다. 따라서 그들의 공격에 대한 처칠랜드의 응답은 기본적으로 연결주의가 이루고 있는 성과들이 무엇인지, 그리고 그것이 보여주는 철학적 함축이 무엇인지 보다 정확히 알려주는 것에 초점을 맞춘다. 처칠랜드의 설명에 따르면, 상태공간 표상이론을 탄생시킨 연결주의 의미론의 실체적 모습은 아래와 같다.

첫째, 그물망의 은닉 유닛들은 과거로부터 얻은 정보를 시냅스들의 가중치 내에 분산적으로 체화함으로써, 그것들은 모험적이고 증폭적인 가설들을 담아낸다[1]. 그러므로 병렬 그물망의 입력과 출력 유닛들 사이의 은닉 유닛들 집단의 응답을, 온도 측정기와 같은 단순한 반응 장치들의 연장선에서, 조금 복잡한 수준으로 보려는 것은 옳지 않다. 잘 훈련된 은닉 유닛들의 그물망이 증폭적 응답을 하기 때

문이다. 그것들은 입력 유닛들로부터 들어오는 새로운 정보들을 단순히 정보 보존하는 인과적 응답을 하지 않으며, 체화된 과거의 의미를 출력 유닛들로 보낸다. 물론 흄은 (시간 경과에 따른) 중요한 개념적 발전 가능성에 대해 설명하지 못하는 근본적 어려움이 있다. 개념의 분류를 직접 감각한 모사물(copies)에서 찾으려 했기 때문이다. 그러나 잘 훈련된 그물망의 은닉 유닛들의 응답 특징은 그러한 문제를 극복할 방법을 제공한다.

둘째, (위의 이야기로부터 알 수 있는 것으로) 훈련된 그물망의 은닉 유닛들이 갖는 표상은, 일차적 감각이라 할 수 없으며, 세계의 원자적 특징들에 명확히 대응하는 매듭으로 묶이지도 않는다. 코트렐(Cottrell)의 얼굴 재인(face recognition) 그물망에서 알 수 있듯이. 그 은닉 유닛들은 내적 범주들의 표상을 담아내어 남성과 여성의 얼굴을 구분할 수 있으며, 학습한 11명의 얼굴을 새로운 사진에서도 잘 재인할 수 있다[17]. 그것들의 표상 부호화 전략은 은닉 유닛 80개의 표상 차원들에 의한 분산된 패턴의 체화이다. 따라서 얼굴 재인 그물망의 은닉 유닛들의 표상은 '감각적 단순자들(sensory simples)'이 아니며, 입력의 정보들에 의한 모사물도 분명히 아니다. 그것들이 갖는 표상들의 범주들은 감각 경험과 거의 어울리지 않는 아주 집합적이며 추상적인 것이다. (그러한 분산적 표상 요소들은 "홀론(holons)"이라 불린다.) 이것이 신경망의 표준적 부호화 작동 방식이다.

셋째, 어느 두 그물망도 정확히 동일성을 가질 가능성은 없지만, 그럼에도 불구하고 성공적으로 훈련된 그물망들은 계층적 분할의 동형적 시스템을 계발함(발전시킴)으로써 서로 아주 유사하다고 말할 수 있다. 시냅스 가중치의 변화, 훈련의 순서, 훈련마다 새롭게 개발되는 홀론 등에 비추어 그물망의 세계 분별 기준은 유동적으로 변화한다. 이러한 측면에서 그물망이 훈련 과정에서 학습한 세계의 특징에 대한 표상 양식은 동일한 입력층을 가짐에도 불구하고 매번 다를 수 있다. 예를 들어, 3차원의 상태 공간 시스템의 특정 지점을 10차원의 상태 공간 시스템에 의해서 동일 지점 또는 거의 인접 지점을 가리키는 일은 얼마든지 가능할 것이다. (이 이해를 위해 뒤를 계속 읽어보라.) 즉, 서로 다른 공간적 표상 시스템이 거의 동일한 입・출력 행동을 유지하는 일은 얼마든 가능하다. 심지어 공간적 지표 구조가 전혀 다를 경우에도 (지표들을 증폭적으로 작동시켜) 특정 내적 분할들, 즉 유사 특이성을 가질 수 있다.

넷째, 위와 같은 그물망의 표상 기능은 감각 속성들과 거의 무관한 표상 능력을 가질 내적 인지 경제도 갖게 한다. 다시 말해서, 서로 다른 좌표계를 가지는 상태 공간들이 동일한 분할 구조를 성공적으로 표상하는 것이 가능하다. 그리하여 전혀 다른 속성들을 입력 정보로 수용하는 감각 유닛들에서 동일한 내적 경제가 유도될 수 있다. 예를 들어, 어떤 그물망이 촉각 기관으로부터 정보를 얻고, 다른 그물망이 시각 기관으로부터 정보를 얻는다고 하더라도, 두 그물망은 서로 동일한 개념을 획득할 수 있다.[7] 이러한 그물망의 표상(계산) 기능에 비추어, 연결주의자 표상 이론의 의미론은 결코 흄의 경험주의와 비유될 상대가 아니다. 흄은 우리의 개념들을 말초의 감각 단순자에 엮어 매듭을 지으려 했을지 모르지만, 연결주의 표상 시스템의 그물망이 갖는 계층적이며 증폭적인 부호화 유닛 층들의 전체 지점은 우리의 말초 감각 부호화의 경계를 초월한다.

다섯째, 새로운 표상이론은 전혀 언어적 형식을 갖지 않는 인지적 경제의 형식, 즉 동물 왕국 전체의 기초 인지적 형식과 또한 인간 언어활동 기저의 인지적 형식을 설명해줄 가능성이 있다. 고전적 의미론의 설명은 의미론적 내용들이 어떻게 언어적 항목들(문장과 그 구성 단어들)과 심적 항목들(믿음과 그 구성 개념들)에 영향을 미치며, 그것들에 체화되는지 해명하려는 시도를 목표로 삼는다. 반면에, 연결주의자들은 그 설명의 시야를 넓혀 일반적으로 정보가 연결망의 시냅스 가중치에 어떻게 영향을 미쳐 활성 패턴들에 영향을 미치며, 신경망에 체화되는지를 해명하려 한다. 그 인지적 형식의 의미론은 인간에 의해 최근 채용된 언어 구조보다 선행하는 것으로, 인간의 언어 구조를 지원한다. 따라서 인지 활동의 기초를 탐구할 의도를 갖는 우리 의미론의 탐구 목표는 그 의미론의 기초 현상들에 초점이 맞춰질 필요가 있다. 그것들이 해명된 뒤에, 미래에 우리는 차례로 비교적 의미론적으로 특별한 사례이며, 진화론적으로 최근의 것인 언어적 항목들을 보다 사실에 가깝게 탐구할 수 있을 것이다.

앞서 지적했듯이, 이상의 처칠랜드의 해명은 포도와 르포르의 비판에 대한 직접적 방어 목적에서 나온 이야기는 아니다. 처칠랜드의 관점에 따르면, 그들의 비판

7) 출생부터 시각 기능을 갖지 못한 사람도 시각 기능이 온전한 사람과 마찬가지로 내적 공간 표상의 체계를 갖는 것이 가능하다. 그들은 처음부터 시각적 정보를 갖지 못하지만, 손으로 더듬는 경험을 통해서 3차원의 공간적 내적 표상을 형성할 수 있다.

은 상태공간 표상이론을 경험주의자 맥락에 포함시키는 확대된 해석에서 나왔으며, 따라서 처칠랜드는 단지 그 오해된 해석을 바로 잡으려 했을 뿐이다. 그러므로 앞의 처칠랜드 해명으로부터 당연히 아래 의문이 제기된다. 처칠랜드가 포도와 르포르의 비판에 대한 직접적 방어를 어떻게 구축할까? 다시 말해서 그들의 비판을 처칠랜드의 연결주의에 대한 일반적 이야기로부터 어떻게 구체적으로 방어할 수 있는가?

내용 유사성의 선험적 조건

앞서 살펴보았듯이 포도와 르포르 비판의 핵심은 상태공간 표상이론이 의미론적 유사성의 지표와 기준을 제공하지 못할 것이라는 지적이다. 그 지적의 관점에서 보면, 상태공간 표상이론은 의미론적 내용 유사성의 기준을 전제한다. 이것은 상태공간 표상이론이 의미론적 동일성의 선험적 조건을 전제한다는 의미로 이해된다. 다시 말해서 무엇이라도 인지적 주체가 경험적 인지를 가능하게 해줄 그물망을 형성하려면, 그 인지적 주체는 의미론의 '기준' 또는 범주 기준을 '경험에 앞서' 이미 가지고 있어야만 한다는 의미로 파악된다. 그러한 범주적 기준이 없이는 어떤 인지적 주체도 자신의 물리적 인지의 그물망이 어떤 의미론적 내용에 일치시킬지 기준을 갖지 못하기 때문이다. 포도와 르포르의 비판을 이러한 측면에서 바라볼 때, 필자는 그들의 지적이 과연 처칠랜드의 표상이론을 관통할 수 있을지 검토할 것이다.

포도와 르포르의 중요한 공격은 '상태공간 차원'과 '의미론적 지표 기준'(내용 유사성의 기준) 사이의 관계에 대한 의혹이다. 앞서 살펴보았듯이, 그 의혹은 세 측면의 문제들로부터 제안되었다. '개별화의 문제', '분석적/종합적 구분 문제', '부차적 정보의 문제', 그리고 '경험주의 의미론의 한계' 등이다. 이 문제들은 모두 처칠랜드가 제안하는 표상이론의 상태공간이 동일 차원의 지표를 설정할 가능성을 원천적으로 봉쇄하게 하며, 따라서 상태공간 표상이론이 의미론적 내용 유사성을 설명할 가능성 또한 부정하게 한다. 이러한 부정적 평가에 대해 필자는 앞(3절)에서 살펴본 처칠랜드에 대한 이해에 근거하여 다음과 같이 주장한다.

첫째, 신경망의 은닉유닛들이 가지는 의미론적 표상의 속성은 개별화의 문제를 불러들이는지, 그리고 차원의 동일성을 전제하는지 검토해보자. 코트렐 얼굴 재인 그물망에서 드러나듯이, 은닉 유닛들은 스스로 입력 정보들에 대한 분석 차원들을 설정하지만, 그것들이 그 설정을 하도록 어떤 기준이 제공되지는 않았다. 단지 최종의 응답이 실제 얼굴을 효과적으로 인식했는지 아닌지의 판정에 따라서 "오류 역전파(back-propagation of error)"라는 학습알고리즘(Learning algorithm)에 의해 시냅스에 상응하는 연결망의 가중치, 즉 연결의 강도를 강하게 또는 약하게 수정하게만 하였다[7]. 이러한 가중치의 수정만으로 신경망은 얼굴을 분별해낼 기준을 담아내었다. 물론 그것들의 구분 기준이 인간의 것과 동일하다고 말할 수는 없다. 인간이 갖는 신경망의 연결은 엄청난 수를 가지지만, 인공적 신경망은 다만 그것을 모방한 빈약한 수준이기 때문이다. 그럼에도 불구하고 그 신경망은 매우 효과적으로 처음 보여주는 새로운 얼굴 사진에 대해서도 동일인을 매우 효과적으로 분별해내는 실험적 결과를 보여주었다[17]. 이러한 신경망은 기존의 범용컴퓨터가 작동하는 방식, 즉 실제와 얼마나 닮았는지 짝을 맞춰보는 수준이 아니다. 그것은 은닉유닛 속에 세계를 분별한 나름의 기준을 형성하고 있기 때문이다. 물론 신경망 내에 은닉유닛들의 표상적 요소인 홀론은 훈련의 과정에서 얼마든 변화될 수 있으며, 더구나 그물망들마다 언제나 동일한 기준이 적용된다고 기대될 수는 없다. 그럼에도 그것들은 (심지어 서로 다른 그물망들이 서로 다른 상태공간 차원을 조성하는 경우라 할지라도) 나름으로 상태공간 내의 유사성 위치를 가리키도록 훈련되어질 수 있다. 다시 말해서 그 얼굴 신경망은 시시각각으로 미세하게 연결망을 변화시킨다. 따라서 그 신경망의 연결 상태를 나타내는 상태공간의 홀론(또는 전형) 역시 시시각각 변화하고 있다. 다시 말해서, 신경망의 위상 상태공간은 차원적으로 변화하고 있으며, 따라서 동일 상태공간에 의한 분류 기준을 가진다고 기대하기 어렵다.

이러한 신경망에 대한 특성에 비추어 볼 때, 포도와 르포르는 처칠랜드의 표상이론이 의미론적 동일성 또는 유사성을 주장할 근거가 없다고 다시 주장할 것이다. 그들에 따르면, 그물망들이 저마다 동일한 차원을 가질 경우에만 동일 상태공간을 가질 수 있으며, 또한 그럴 경우에 한해서 그 그물망들이 동일한 것에 대해서 서로 유사한 상태공간의 지점을 가리킬 수 있기 때문이다. 그러나 동일하거나 유사한 상태공간 위치를 가리킬 수 있기 위해서 반드시 유사성을 위한 동일 차원

의 상태공간 또는 동일 상태공간을 전제할 필요는 없다. 왜냐하면 그물망마다 서로 다른 차원의 다른 상태공간들을 가지고도 각자의 상태공간 내에서 유사한 지점을 가리키는 것이 다음과 같이 가능하기 때문이다. 우리는 직관적으로 세 숫자의 조합으로 표현되는 3차원 공간을 상상할 수 있다. 그러나 그 이상의 여러 숫자의 조합으로 표현될 다중 차원을 수학적으로 가정하는 것은 얼마든지 가능하다. 그러한 이해에 기초해서 다음과 같이 말할 수 있다. 2차원 공간의 특정 지점이 3차원 공간의 특정 지점으로 표현될 수 있듯이, 3차원의 특정지점이 10차원의 특정 지점으로 표현될 수 있다. 이러한 측면을 고려하면 유사성을 위해서 위상 상태공간은 동일 차원을 전제하거나 동일 상태공간을 전제해야 할 이유가 없다.[8] 그러므로 그 물망들이 서로 동일 차원의 공간적 구조를 조성하기 어렵다는 전제로부터 신경망이 내용 유사성의 기준을 가질 수 없다는 포도와 르포르의 주장은 연역되지 않는다.

둘째, 분석/종합(판단) 구분의 어떤 선험적 기준을 전제하지 않으면, 상태공간 표상이론이 내용 동일성의 의미론적 문제를 과연 다룰 수 없는지 검토해보자. 정말로, 실험적 연구에서 드러났듯이, 인공신경망의 은닉유닛들이 상태공간 내에 특별한 지표를 갖도록 실험자가 그 인공신경망에 특정한 의미론적 지표 동일성의 규정을 미리 설정하지는 않는다. 처음에는 은닉유닛들이 경험적 현실 적용에 필요한 개념적 지표를 임의로 설정한다. 그리고 그 개념적 지표는 다음의 실험적 수행의 결과를 역전파(back-propagation)함으로써 수정된다. 이러한 과정을 통해서 실제 신경계는 마침내 세계에 적절히 반응할 지표를 스스로 형성한다. 물론 실제 신경계의 작용에서 그 결과물인 개념적 지표 기준들은 사람마다 모두 동일할 수도 그럴 필요도 없다. 그러한 이유로 우리는 일상의 대화에서 의사소통의 어려움을 가질 경우가 적지 않다. 그리고 토론의 과정을 거치고 난 후 우리는 서로의 대화가 어려웠던 이유가 사실은 서로가 다른 개념적 기준을 가졌기 때문임을 밝혀내고, 그것들을 수정하곤 한다. 이러한 수정 과정은 마치 신경계가 은닉유닛의 상태공간의 지표를 수정하는 과정으로 이해될 수 있다. 이렇게 신경계를 상태공간 표상이론에 따라 이해해보면, 포도의 주장처럼 상태공간 표상이론이 이미 완결된 분석/종합의

8) 만약 누구라도 텐서(Tensor)의 개념을 이해한다면, 위상 상태공간에서 유사 지점을 지시하기 위해 동일 차원이 전제된다고 주장하지 않을 것이다.

의미론적 구분을 전제해야 한다고 우리가 가정할 필요는 없다.

뿐만 아니라, 전문가들이 설정한 개념들의 지표 또는 차원은 그렇지 못한 평범한 사람들이 갖는 개념적 지표의 기준 또는 차원과는 적지 않게 다를 수 있다. 그럼에도 불구하고 그러한 사람들 사이에 전문적 내용에 대한 대화가 전혀 불가능한 것은 아니다. 이것을 우리는 다음과 같이 이해해볼 수 있다. 서로 다른 상태공간과 서로 다른 지표들을 가지고도, 즉 서로 다른 분석/종합 구분의 기준을 가지고도, 은닉유닛들 전체 시스템은 동일 유사한 공간의 지점을 가리킬 수 있다. 따라서 서로 (때에 따라서는 매우 대략적으로) 대화가 가능하다. 이것은 상태공간 표상이론에 의해서 다음과 같이 설명된다. 여러 다른 위상 상태공간들 내에서 일정 공간의 지점은 텐서변환(tensor transformation)에 의해 동일한 공간의 지점으로 표현될 수 있다. 그렇게 서로 다른 지표를 갖는 다른 상대공간들 사이에 의미론적 유사성을 갖는 일이 가능하다. 따라서 그것이 불가능하다고 주장하는 포도의 주장은 틀렸다.

포도와 르포르에 따르면, 개념의 지향적 상태가 선험적으로 존재하며, 그것에 대한 기준이 없이는 누구도 세계를 인지할 수 없다고 주장한다. 필자는 이것을 부정할 생각은 없다. 사실 그렇다고 생각한다. 그러나 우리는 그것을 초험적으로 (그리고 확정적으로) 소유하지는 않으며, 누구에게나 엄격한 보편적 기준으로 적용되는 불변의 것은 더욱 아니다. 경험과 교육은 우리의 인지적 그물망을 바꿀 것이며, 그 그물망이 안내하는 지표의 변화는 그물망 개념체계의 수정이다. 이것은 그물망 의미론과 자연화된 인식론을 주장하는 콰인의 전체론적 입장이기도 하다.

이러한 고려에 근거해 볼 때, 확고히 결정된 분석/종합의 구분에 대한 기준이 없이 상태공간 표상으로 내용 동일성을 주장하기 어렵다고 가정하는 포도와 르포르의 주장은 설득력을 잃는다. 그물망의 형성 초기, 즉 학습 이전부터 이미 분석/종합의 구분을 가져야만 한다는 포도와 르포르의 주장은 부정된다. 그물망은 임의적으로 설정한 분류 기준의 오류를 수정하는 과정을 통해 그 기준을 은닉유닛 내에 공고하게 조성해갈 것이다. 그리고 그것을 통해서 그 연결망은 세계에 대한 내용 동일성을 가질 것이다.

셋째, 포도와 르포르가 지적하는 부차적 정보의 문제가 상태공간의 표상이론에서 발생되는지 살펴보자. 앞서 살펴보았듯이, 포도와 르포르가 지적하는 부차적 정보의 문제는 감질반전(qualia inversion)의 문제와 다음과 같이 관련된다. 감질 반전의

문제는 상태공간의 의미론적 지표가 뒤바뀔 경우에 일어날 수 있다. 내가 바라보는 잔디와 당신이 바라보는 뜨거운 엔진의 감질이 동일할 가능성은 서로의 상태공간의 감각 지표가 다를 경우에 일어난다. 따라서 감질 반전의 문제를 고려할 때, 신경망이 상태공간 표상이론에 따라 의미론적 내용 유사성을 갖는다고 주장하려면 그 신경망은 동일 지표를 이미 가져야 한다.

그러나 그 감질 반전의 가능성 문제는 정말로 처칠랜드 역시 이미 염두에 두고 있었던 것이기도 하다. 앞서 논의하였듯이, 여러 신경망들의 은닉유닛들은 그물망 차원의 상이성에도 불구하고, 그리고 그 지표의 다름에도 불구하고, 각자의 상태공간 내에서 상호 유사 지점을 표현할 수 있다. 그 점에서 감질 반전의 문제 제기는 상태공간 표상이론에 대한 반박이 되지 못한다. 사실 그물망의 특정 상태공간은 다른 여러 경험적 자료에 활용될 수 있으며, 다른 추상적 개념을 위해서도 활용될 수 있다. 그러므로 (감질반전 문제를 포함하여) 부차적 정보의 문제를 들어서 새로운 표상이론이 의미론적 유사성을 설명할 수 없다는 포도의 비판은 처칠랜드 표상이론에 대한 적절한 공격이 아니다.

구체적으로 말하자면, 일반적으로 뇌는 특정 영역마다 전문화(specialization)를 보여주기도 하지만, 뇌의 가소성은 일반적인 경우에서 벗어난 다른 영역에 전문화를 조성하는 경우도 있다. 예를 들어, 느리게 진행하는 암세포가 브로카 영역에서 성장하게 되면, 그 영역의 언어운동 기능은 약간 비켜난 위치에 전문화될 수 있다. 그럴 경우에 그 환자의 그물망 상태공간의 지표는 그렇지 않은 사람들과 다르게 형성될 수 있다. 실제로 그 환자의 브로카 영역을 절제하는 수술로 언어운동장애가 발생하지 않는다는 의학계 보고가 있다. 또한 시각의 자료와 청각의 자료가 종합적으로 처리되는 두정엽에서 대응도는 시각의 자료를 청각의 효과에 활용하는 것으로도 알려져 있다. 이것은 “맥구룩 효과(McGurk effect)”라고 불린다. 또한 우리는 시각 정보만으로 공간적 위치를 표상하는 것은 아니며, 청각 정보는 물론 촉각 정보를 통해서도 공간적 위치 표상을 얻어낼 수 있다. 따라서 만약 감질 반전의 문제와 부차적 정보에 근거해서 상태공간 표상이론이 감각 내용의 동일성을 규정하지 못할 것이라고 지적한다면, 그 지적은 뇌의 가소성과 생물학적 적응의 문제, 그리고 그림 4에서 설명된 뇌의 정보 통합 능력을 고려하지 않은 관점에서 나온 반론이다. 간단히 말해서, 빠르게 발전하는 현대 신경학에 대한 이해의 부족을 보

여주는 반론이다. 인공신경망은 이러한 뇌의 기능을 모방한 학습방식을 갖는다. 따라서 다음과 같이 말할 수 있다. 그물망은 부차적 정보에 따라서 그 상태공간을 변화시키는 경우에도 여전히 상태공간 내에 특정 표상에 대한 위치를 일정하게 유지할 수 있다.

끝으로, 처칠랜드의 표상이론은 경험주의 한계를 갖는지 알아보자. 어쩌면 (콰인을 포함하여) 처칠랜드에게 가장 곤란한 입장으로 몰고 갈만한 논점은 전체론이 사실상 경험주의 한계를 벗어날 수 없다는 지적일 것이다. 포도가 지적하는 문제는 사실 전통적으로 합리주의 관점과 경험주의 관점 사이의 쟁점에서 이미 잘 드러난 문제이다. 경험 자체가 가능하기 위해 의미론의 개념이 우선해야 한다는 입장과 그것 없이도 감각적 경험으로부터 강건한 유사성을 얻을 수 있다는 입장의 대립이다. 고전적 경험주의를 비판하는 콰인에 따르면, 감각적 증거로부터 보편적 개념을 정당화할 수 없으며, 개별적 관찰 문장들로부터 결코 일반화를 정당화할 수 없다[18]. 만약 우리가 오직 이러한 콰인의 입장만을 그대로 수용한다면, 그리고 포도와 르포르가 염려하는 관점에 따르면, 의미론적 전체론은 결코 경험의 가능성을 주장할 수 없어야 한다. 전체론이 가정하는 이론적 그물망의 주변은 경험적 용어와 대응하며, 안쪽의 매듭은 추상적 용어와 대응한다. 다시 말해서, 주변의 매듭들은 오직 관찰 속성들과 관계하여 의미론적 특성을 규정해야 할 것처럼 보인다. 그렇다면 이러한 전체론이 경험주의와 다를 것이 무엇인지 의문된다는 것이 포도와 르포르의 비판이었다. 의미론적 동일성 또는 유사성의 개념들이 결코 관찰 가능한 속성들로 표현되지 못할 것이라는 점을 고려해보면, 전체론이 경험주의를 넘어서 전혀 새로운 가능성을 보여주지 못할 것처럼 보인다.

그러나 콰인의 전체론이 진정으로 의도하는 것이 무엇인지 아래와 같이 이해하려는 개방적 태도를 가진다면, 위와 같은 단순한 전망에서 벗어날 수 있다. 콰인의 그물망 의미이론의 입장에서 관찰과 이론을 명확히 구분할 경계를 말하기 어렵다. 노이라트의 선원이야기를 콰인이 인용했듯이, 자신이 타고 있는 갑판을 뜯어 배를 수리하면서 운항하는 항해사들처럼, 우리는 경험을 통해서 어떻게든 이론을 수리하면서도 그 이론에 의해 경험하는 과학의 항해자들이다. 경험에 의해 이론들을 뜯어내거나 수리하면서도, 그 이론에 의지해서 경험을 한다[18]. 이러한 측면에서 다음과 같이 말할 수 있다. 순수한 객관적 의미로서 관찰이란 가능하지 않다. 그러

나 그러한 경험은 이론을 교정한다.

이러한 주장을 지지하는 처칠랜드의 입장에서 다음과 같이 말할 수 있다. 연결주의 인공신경망 속에서 관찰(경험)과 이론(개념체계)에 대응하는 신경망 표상들을 명확히 구분할 경계를 말하기 어렵다. 이론들이 조성하는 의미론적 그물망에 의해 관찰이 가능하다는 것은 의미론적 그물망이 경험에 대한 의미론적 지표가 된다는 이야기이다. 연결주의 신경망이 하는 역할을 배를 수리하면서 운행하는 노이라트의 선원 이야기에 비유하여 다음과 같이 말할 수 있다. 그물망은 그 자체로 가지는 은닉유닛들의 임의적 범주체계를 가지고 세계를 분별하면서, 새로운 경험에 마주쳐서 발생되는 오류를 수정하는 방식으로 그 범주체계를 개선해 나아간다. 그럼으로써 그물망은 점차로 세계에 대해서 효율적 분별 기준을 담아낼 것이다. 콰인식의 경험주의 관점에서 볼 때, 현대 신경망은 의미론적 (범주체계의) 완결된 기준을 선험적으로 전제하지 않는다. 따라서 신경망이 경험주의의 한계를 갖는다는 포도와 르포르의 지적은 옳지 않다. 그들의 지적은 그물망이 갖는 가소성의 기능에 대한 이해의 부족에서 나온 듯이 보인다.

상태공간 의미론의 함축적 의미

본 논문을 마무리하면서 필자는 아래와 같은 의문에 들어서게 된다. 만약 우리가 지금까지 검토한 처칠랜드의 상태공간 표상이론을 받아들이는 입장에 선다면, 그리고 앞으로 우리가 상태공간 의미론을 용인하고 그 탐구를 지속적으로 확대한다면, 미래에 우리는 어떠한 인지과학의 전망을 기대할 수 있겠는가?

전통적인 표상이론의 접근법은 의식 가능한 명제 수준에 대한 탐구 전략에서 나온다. 그렇게 의식적 수준의 언어를 분석하는 탐구전략은 전통적으로 철학자들이 논리적 분석 방법론으로 활용해온 것이며, 최근까지 우리가 유력하게 지지해온 인지적 탐구 접근법이다. 그 탐구 방법을 통해 우리는 인간의 이성적 (논리적) 사고의 흐름을 분석할 수 있었으며, 나아가서 인간과 동등한(유사한) 논리적 사고를 수행할 기계적 계산이 어떻게 가능한지 탐구하였다. 그리고 마침내 그 탐구전략을 통해 인류는 현재 유용하게 이용되는 범용컴퓨터의 논리체계를 개발하고 발전하였

다[19].

반면에 앞서 살펴보았듯이 처칠랜드의 상태공간 표상이론은 우리의 언어적 분석을 회의하게 하며, 동시에 의식적(이성적) 논리를 분석하는 것에도 회의하게 한다. 그렇다면 이러한 처칠랜드의 표상이론은 인지과학의 탐구 분야에 어떤 구체적 전망(또는 실제적 유용성)을 제공할 수 있겠는가? 필자는 그 전망을 아래와 같이 셋으로 나누어 이야기해볼 수 있다. (아래의 전망들은 보다 구체적이며 세밀한 논의를 필요로 한다. 그러나 그것은 본 논문의 직접적 논의 목표를 넘어선다. 따라서 여기서는 단지 개략적인 전망만을 가늠해보겠다.)

첫째, 상태공간 표상이론은 인지적 시스템에 대하여 의식적 언어의 이하 수준에서 탐구하도록 전망하게 한다. 그 전망의 관점에 따르면, 전통 철학자들과 언어학자들이 지금까지 지향해온 의식적이며 일상적 언어분석에 의한 의미론 탐구 접근법[20]은 인간의 표상적 실제를 올바로 반영하지 못하며, 따라서 비록 이제까지 유력했다고 하더라도 앞으로 그 탐구전략으로는 실제적 소득을 크게 기대할 수 없을 것이다[21]. 이러한 입장을 처칠랜드의 초기 저작(1979)[22]에서 아래와 같이 찾아볼 수 있다.

> 우리는 지금까지 인식적 가치의 특성을 믿음의 속성과 그 관계에 의해서, 보다 정확히 말하자면, 믿음들을 표현하는 '문장들 사이의 관계'에 의해서 탐구해왔다. 그러나 만약 그러한 탐구에 지표가 되는 상식적인 이론적 개념이 상당히 뒤틀린 것이라면, 혹은 상당히 피상적인 것이라면 그 "문장에 의한" 인식론이 언젠가는 무언가를 성취할 것이라고 기대할 합당한 이유가 없다. 사실상 지금까지 우리는 그렇게 연구해오고 있었다.

이러한 전망에서 필자는 다음과 같이 주장한다. 앞으로 상태공간 표상이론에 의한 우리의 입력정보와 출력정보 사이의 신경계산적 전망은 지금까지 알려지지 않은 우리 인지와 행동 사이의 관계에 대해 보다 현대 과학적인 명확한 이해를 제공할 것이다.

고전적 의미론에 대한 탐구는 언어적 항목들과 그것에 대응한다고 믿어지는 심적 상태들에 관심을 가져왔다. 그러한 관심의 바탕에는 문장에 '믿음' 또는 '지식'

이 대응하며, 문장의 구성요소인 단어에 '개념'이 대응한다는 교설이 있었다.[9] 반면에, 처칠랜드의 상태공간 표상이론은 우리 신경계의 입력 정보들이 신경망의 시냅스 가중치들의 영향에 의해 출력 정보들로 변환되는 (현대 유력한 과학이론들이 지원하는) 신경계의 인지 시스템에 관심을 갖는다. 그러한 실제적 인지 경제 형식은 언어 구조보다 그 이하 수준의 역학 구조에서 작동한다. 그 기초 인지 구조의 역학은 (진화론적으로 인간에 의해 최근 채용된) 인간 언어활동을 지원하기는 하지만, 그렇다고 인간 언어활동 자체에 대한 분석을 통해 우리의 인지 기초를 탐색하는 것은 적절하지 않다. 상태공간 표상이론을 통해 알 수 있듯이, 행동 조절에 대한 상식적이며 의식적인 수준의 설명은 우리를 포함하는 동물들의 신경계에 대한 통합된 인지 체계에 근거하지 않은 가설들에 기초한다. 실제를 외면하는 고전적 의미론의 접근법은 앞으로 우리의 인지적 시스템에 대한 이해와 탐구에 방해물이 될 뿐이다.

둘째, 새로운 표상이론을 지지하는 관점에서 감성과 이성의 통합적 탐구를 전망해볼 수 있다. 정서적 정보 역시 신경계의 신경망을 통해 작동되며, 따라서 그것 역시 상태공간 표상이론이 제안하는 상태공간의 위치 또는 행렬의 숫자 조합으로 표현될 수 있을 것으로 전망된다. 어떤 정서 정보의 내용도 인지 정보의 내용과 마찬가지로 표현된다면 두 정보는 서로 다르게 작동되어야만 할 필연적 이유를 갖지 않을 것이다. 일상적으로 전혀 이질적으로 보이는 두 종류의 정보가 신경망 속에서 서로 통합될 가능성이 원리적으로 가능하다.

뇌과학자 안토니오 다마지오[23]는 우리의 정서(emotion)와 관련하여 다음과 같이 말한다.

> 정서적 신호는 선택 가능성이나 결과를 긍정적이거나 부정적인 신호로 표지함으로써 의사결정 범위를 좁히고 행동이 과거의 경험과 부합할 가능성을 증가시킨다.

이러한 감성에 대한 이해로부터 필자는 인지 정보와 감성 정보를 통합적으로

9) 이러한 점에서 믿음의 거미줄을 주장했던 콰인 역시 전통적 관점에서 온전히 빠져나오지는 못했으며, 처칠랜드의 관점과 구분된다. 처칠랜드 부부와 2011, 2월 대화 중에.

설명할 가능성을 본다. 그리고 상태공간 표상이론의 신경계산적 전망에서, 필자는 정서 기능과 그 기초 원리를 아래와 같이 가설적으로 제안하게 된다.

> 정서란 일종의 신호처리 강화기능이다. 즉, 입력정보를 증폭시킴으로써 출력정보를 비선형적으로 처리하게 하는 기제이다.

더구나 진화생물학적 관점을 상태공간 표상이론에 도입하여 정서 기능을 이해해 보면, 우리 신경계는 입력정보에 대한 증폭 처리에 의해 세계에 대한 정보를 왜곡하게 되었으며, 그 왜곡은 유기체로 하여금 환경에 적절히 대응하게 했을 것으로 이해해볼 수 있다. 그러한 기능은 신경망의 상위차원에서도 필요할 수 있다. 일단 인지적으로 확인이 된 정보를 우리 행동에 적절히 이행시킬 것인지 판결해야 할 경우, 신경계는 그 정보를 다시 증폭적 계산처리할 필요가 있기 때문이다. 이것이 바로 인지처리의 신경망과 정서처리의 신경망을 통합적으로 탐구할 신경계산적 전망(neurocomputational perspective)이다.

셋째, 상태공간 표상이론은 미래 인지적 기능을 수행할 기계적 장치에 대해서 역시 의식(또는 언어) 이하의 수준에서 계산적으로 탐구될 것을 전망하게 한다. 입력 정보를 출력 정보로 변조하는 기능을 수행하는 모델은 크게 고전적 계산주의 모델과 연결주의 모델로 구분된다. 최근까지 인공지능 탐구 방법론으로서 (연결주의와 대비되는) 고전적 계산주의 모델의 관점은 인지적 기능을 의식적 논리에 기초하여 분석하고 탐구하게 한다. 그 결과 일정한 계산 규칙들에 따라 기호를 조작하여 입력 정보를 출력 정보로 변화하는 범용컴퓨터를 고안하게 하였다. 더구나 그 장치는 튜링 머신의 원리에 따라서 일정한 주소에 저장된 정보를 불러와 활용하는 방식으로 계산 과제를 수행한다. 따라서 융통성을 발휘하거나 배경지식을 고려한 연산 과제 수행에 어려움이 있으며, 실제 신경계가 처리하는 쉬운 과제를 매우 복잡한 계산처리 단계로 처리해야 한다.

반면에 최근 주목되는 연결주의 모델의 인공지능은 그러한 문제들을 해소할 수 있음을 보여준다. 신경망 모의실험을 통해 밝혀진 바, 신경망은 입력 정보로부터 출력 정보를 산출하는 일반적 패턴을 자체 연결망에 분산적으로 저장하는 학습을 수행할 수 있다. 따라서 그 일반적 패턴은 입력 정보를 처리할 배경지식 또는 배

경적 일반화와 같은 역할을 담당할 수 있다[6]. 만약 그물망의 은닉 유닛이 입력 정보들을 변조하는 일반화를 담을 경우, 그 일반화는 우리 의식 이하 수준에서 새로운 입력 정보들을 분별하게 해줄 반응패턴일 뿐만 아니라, 예측을 가능하게 할 이론적 기능을 담당할 것이다.

나아가서, 처칠랜드의 표상이론이 제안하듯이, 만약 미래의 계산 시스템이 표상적 전형들을 계층적 구조에 조성하여 유사성의 추상적 범례들을 계층적이며 병렬적으로 담아낸다면, 그리고 만약 입력 정보들이 그러한 계층적 범주체계를 통해 병렬적으로 계산처리 되도록 한다면, 그러한 계산 시스템은 우리의 다양한 맥락을 병렬적이며 계층적으로 고려하는 인지체계를 모방할 것이다. 이러한 전망은 지금까지 연결주의 모델 실험이 우리의 인지를 모방하기에 부족함이 있다는 지적[24][25]에도 불구하고 연결주의 모델에 기대게 하는 이유이다.

지금까지의 전망에 대해 조심하여 아래와 같이 다시 질문할 수 있다. 과연 우리가 처칠랜드의 표상이론을 긍정적으로 바라보아야 할 이유가 더 있는가? 만약 여러 이유들이 있다면 그 중에 가장 큰 이유는 무엇일까? 이러한 의심에 대한 대답을 폴 처칠랜드[1]의 다음 말에서 찾아볼 수 있다.

> 좌표변환에 의해 상호작용하는 상태공간 시스템의 경이로운 표상과 계산 능력은 신경계의 인지활동을 이해시켜줄 강력하고 매우 일반적인 수단을 제시한다. 특별히 그런 시스템을 이행하기에 적절한 물리적 구조가 뇌 전반에 퍼져있기 때문이다.

한 마디로 말해서 처칠랜드의 표상이론은 현재 발전하고 있는 경험과학의 성과에 의해 지지된다.

끝으로, 만약 우리가 지금까지 현대 신경과학이 탐구한 뇌의 사실적 구조를 신뢰한다면, 이제 우리는 처칠랜드의 표상이론을 통해 뇌가 갖는 표상이 무엇인지 이해하게 되었다고 말할 수 있다. 특히 대상의 일반적(범주의) 표상이 무엇인지, 나아가서 경험의 일반적 표상이 무엇일지 이해하기 시작했다. 이런 이해로부터 필자는 다음과 같이 전망할 수 있다. 미래의 인공지능 시스템이 우리처럼 세계에 대한 의미론의 유사성 준거 기준(또는 가족 유사성)을 지원할 개념체계(또는 이론체계)

를 갖도록 우리가 앞으로 무엇을 어떻게 해야 할지 내다 볼 수 있게 되었다.

참고문헌

[1] Churchland Paul M. (1989), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press, Cambridge, MA.

[2] Bechtel, William and Adele Abrahamsen (1991), *Connectionism and the Mind: An Introduction to Parallel Processing in Networks,* Blackwell Publishers, Cambridge.

[3] Smolensky, P. (1990), Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems, *Artificial Intelligence* 46: 159-216.

[4] Fodor, Jerry A. and Pylyshyn, Z.W (1988), Connectionism and Cognitive Architecture: A Critical Analysis, *Cognition* 28: 3-71.

[5] 최 훈 (1994), 연결주의와 구성적 구조, **철학논구**, 22, 229-276.

[6] 박제윤 (2010), **신경철학의 인식론**, 인하대학교 철학과 박사학위논문.

[7] Churchland, Petricia S. (1986), *Neurophilosophy: Toward a Unified Science of the Mind-Brain, The MIT Press, Cambridge, MA, London, England* (박제윤 번역, 2006, 뇌과학과 철학, 철학과 현실사, 서울.)

[8] Churchland, Patricia S. and Sejnowski, Terrence J. (1990), Neural Representation and Neural Computation, *Philosophical Perspectives*, Vol. 4, 15-48.

[9] ———————— (1992), *The Computational Brain*, MIT Press, Cambridge, MA.

[10] Churchland, Patricia S. (2002), *Brain - Wise*, MIT Press, Cambridge, MA.

[11] 박제윤 (2009), 표상의 역할과 환원적 제거주의, **과학철학**, 12-1, 95-124.

[12] Fodor, J. and E. Lepore (1992), *Holism: A Shopper's Guide*, Blackwell Publishers, Oxford, Cambridge.

[13] ———————— (1996), Paul Churchland and state space semantics, *The Chruchlands and their Critics*, ed. R. N. McCauley, 145-159. The MIT Press, Cambridge, MA.

[14] Churchland Paul M. (1996), Fodor and Lepore: State-space semantics and meaning

holism, *The Chruchlands and their Critics*, ed. R. N. McCauley, 272-277. The MIT Press, Cambridge, MA.

[15] Fodor, J. and E. Lepore (1996), Reply to Churchland, *The Chruchlands and their Critics*, ed. R. N. McCauley, 159-162. The MIT Press, Cambridge, MA.

[16] Churchland Paul M. (1996), Second reply to Fodor and Lepore, *The Chruchlands and their Critics*, ed. R. N. McCauley, 278-83. MIT Press, Cambridge, MA.

[17] ———————— (1995), *The Engine of Reason, the Seat of the Soul,* MIT Press, Cambridge, MA.

[18] Quine, Willard V. O. (1969), Epistemology naturalized, *Ontological Relativity and Other Essays*, New York: Colombia University Press, 69-90.

[19] 김영정 (1996), **심리철학과 인지과학**, 철학과 현실사, 서울, 187-190.

[20] 이정민 (2001), 의미 표상과 인지, 인지과학, 서울대학교 인지과학 연구소, **인지과학총서**, 155-174.

[21] 최재천 (2001), 동물의 인지, 인지과학, 서울대학교 인지과학 연구소, **인지과학총서**, 497-538.

[22] Churchland Paul M. (1979), *Scientific Realism and the Plasticity of Mind*, Cambridge University Press, Cambridge, 12.

[23] Antonio Damasio (2003), *Looking for Spinoza: Joy, Sorrow and the Feeling Brain,* Houghton Mifflin Harcourt, Olando, Austin, San Diego, Toronto, London (임지원 번역, 2011, **스피노자의 뇌: 기쁨, 슬픔, 느낌의 뇌과학**, 사이언스 북스, 서울, 174.)

[24] Pinker, S. and A. Prince (1988), On language and connectionism: Analysis of a parallel distributed processing model of language acquisition, *Cognition*, 28, 73-193.

[25] ———————————— (1996), The nature of human concepts/evidence from an unusual source, *Communication* & *Cognition*, Vol.29, 3/4, 307-362.

1 차원고접수 : 2012. 2. 8
2 차원고접수 : 2012. 3. 26
최종게재승인 : 2012. 5. 7

(*Abstract*)

The Churchlands' Theory of Representation and the Semantics

Jeyoun Park

Center for Cultural Education, Hanyang University

Paul Churchland(1989) suggests the theory of representation from the results of cognitive biology and connectionist AI studies. According to the theory, our representations of the diverse phenomena in the world can be represented as the positions of phase state spaces with the actions of the neurons or of the assembly of neurons. He insists connectionist AI neural networks can have the semantical category systems to recognize the world. But Fodor and Lepore(1996) don't look the perspective bright. From their points of view, the Churchland's theory of representation stands on the base of Quine's holism, and the network semantics cannot explain how the criteria of semantical content similarity could be possible, and so cannot the theory. This thesis aims to excavate which one is the better between the perspective of the theory and the one of Fodor and Lepore's. From my understandings of state space theory of representation, artificial nets can coordinates the criteria of contents similarity by the learning algorithm. On the basis of these, I can see that Fodor and Lepore's points cannot penetrate the Churchlands' theory. From the view point of the theory, we can see how the future's artificial systems can have the conceptual systems recognizing the world. Therefore we can have the perspectives what cognitive scientists have to focus on.

Key words : state space theory of representation, semantics, the criteria of similarity, Paul Churchland