인지 과학에서의 본질주의와 가족 유사성*

김 영 정 (서울대 철학과)

1. 계산주의와 연결주의: 두 가지 기능주의적 인지 모형

인지 과학의 철학적 토대를 이루어 온 기능주의는 크게 존재론과 방법론의 두 측면에서 규정될 수 있다.

먼저 존재론의 측면에서 보자면, 기능주의는 인지 과학이 해명하고자 하는 마음의 본질을 일종의 기능으로 파악하려는 입장이다. 그렇기 때문에, 기능주의는 기존의 행동주의처럼 심적인 대상들의 존재를 무조건 부인하려 하지도 않고 또 물리주의처럼 철저한 환원이 가능하다고 주장하지도 않는다. 대신, 심리 상태라는 것은 인과적 영향력을 가지고 있으면서도 동시에 물리적 상태로는 환원될 수 없다는 주장을 펼치고 있다.

이 같은 존재론적 기능주의의 근거로 제시되는 것이 복수 실현 가능성 논변이다. 복수 실현 가능성 논변은 하나의 심리 상태, 이를테면 아픔 같은 것을 구현하는 데 물리적으로 다양한 방식이 가능하다는 부인하기 힘든 직관에 호소하고 있다. 동일한 심리 상태를 구현하는 다양한 물리적 가능성이 있다면, 심리 상태의 본질을 물리적 구현 방식에서 찾는 물리주의는 잘못이라는 것이다.

이런 존재론적 가정 위에서, 기능주의는 마음을 연구하는 새로운 방법론을 제시한다. 바로 마음을 컴퓨터에 유비하여 이해하는 것이다. 컴퓨터와 인간의 두뇌는 물리적 관점에서 판이하게 다르다. 그러므로 물리주의적 가정에서 보자면, 컴퓨터를 통해 인간의 마음을 이해하려는 시도는

* 본 논문은 '94년도 서울대학교 대학발전기금 대우 학술연구비의 지원에 의한 것임

전적으로 무의미할 것이다. 하지만 기능주의는 인간과 컴퓨터라는 두 가지 시스템이 비록 서로 다른 물리적 구조를 가졌을지라도 동일한 기능적 상태에 처해 있을 가능성을 인정하고 있다. 따라서 컴퓨터의 기능적 상태에 유비해 인간의 심리 상태를 이해하는 것도 가능하다고 본다.

이상과 같이 기능주의의 큰 테두리 안에서 전개되는 오늘날의 인지과학은, 연구를 위해 채택하는 구체적인 인지 모형의 차이에 의해 다시 계산주의와 연결주의라는 두 가지 노선으로 나누어진다. 이 가운데 기존의 주류 노선인 계산주의가 채택하는 인지 모형은 규칙에 따라 기호를 조작하는 고전적인 형태의 컴퓨터이다. 반면 새로이 부각되고 있는 연결주의가 채택하는 인지 모형은 인공 신경망이라고도 불리는 병렬 분산 처리 모형이다.

연결주의의 병렬 분산 처리 모형을 기존의 모형과 구분해주는 주요한 특징은 정보 처리를 위한 프로그램을 인간이 직접 짜 넣을 필요가 없다는 점이다. 다시 말해 병렬 분산 처리 모형인 인공 신경망은 스스로 문제 풀이 방법을 학습한다. 인공 신경망의 학습은 샘플들로부터 얻어진 투입과 산출에 대한 한정된 경험을 (신경망 노드들의 활성화 값과 연결 강도를 조정하여) 일반화함으로써 차후 필요에 따라 적절한 기능을 수행할 수 있는 상태를 형성하는 것이라 할 수 있다.

병렬 분산 처리 모형의 또다른 특징은 연상 기억 능력이다. 일반적으로 고전적인 정보저장 매체는 대부분 어떠어떠한 내용의 정보가 어느 곳(번지 수 : address)에 저장되어 있다는 것을 알고 그것을 찾아내게 되어 있다. 그러나 사람들이 배우의 사진을 보고 그가 출현했던 영화의 한 장면을 떠올린다던가, 시험에 임해서 전날 공부하였던 내용 중 문제와 관련된 지식을 기억해 내는 방법은 주소에 의해 저장 장소를 찾아가는 방법과는 다르다. 내용의 일부 또는 관련된 단서를 가지고 전체를 찾아내는 이른바 CAM(Content Addressable Memory) 방식이다. 연결주의 기억 장치는 그 동작 원리로 연상 작용을 수반할 수 있다. 즉 기억되어 있는 데이타와 똑같지는 않더라도 유사한 입력 정보가 들어오면, 그것을 확률적으로 근사화하여 용도에 적합하게 사용한다.

Cumins와 Schwarz도 그들의 논문 "Connectionism, Computation and Cognition"에서 인간 지식의 대부분은 계산 가능한 함수에 의해 규정되지 않는다고 주장하며 계산주의와 연결주의의 차이를 다음과 같이 강조하고 있다.

> 대부분의 연결주의 작업이 지금까지 그 취지나 실천에 있어서 계산주의적이었다. 그러나 연결주의는 본질적으로 계산주의적이지 않다. 왜냐하면 연결주의 연구는 인지적 함수들이 계산 가능하다(즉 회귀적이다)고 가정할 필요가 없기 때문이다. 연결주의 체계를 구성함에 있어 알고리듬을 명시할 필요가 없다. 연결주의 체계가 요구하는 것은 문제 영역을 반영하는 표상 상태들을 야기하는 동역학이다. 연결주의 표상 상태들은 동역학적 체계 속의 상태일 수 있다. 그리고 동역학 체계의 특징 함수(이 함수는 동역학 방정식에 의해 정의 된다)는 그것 자체로 계산 가능하지 않다.

이상의 논의를 살펴보면 계산주의와 연결주의는 분명히 그 방법론에서 차이를 보이고 있음을 알 수 있다. 무엇보다도 연결주의는 계산주의적 정보 처리의 근간이 되고 있는 규칙에 의거한 기호 조작 방식을 채택하지 않고 있다.

그러나 연결주의 역시도 계산주의와 마찬가지로 기능적 시뮬레이션을 통해 인간 인지를 모의하고 설명할 수 있다고 가정하는 점에서는 공통적이다. 그렇기 때문에, 연결주의도 존재론적으로는 기능주의라는 큰 틀을 벗어나지 않는 것으로 보아야 한다. 계산주의와 연결주의의 차이는 다만 기능으로 규정된 인간 인지를 어떤 방식을 통해 모의하려 하는가 하는 방법 상의 차이에 국한된다. 더군다나 기능주의는 동일한 기능의 구현에 있어 다양성을 인정하는 복수 실현 가능성 논제 위에 서 있다. 그러므로 인간의 인지적 기능이 연결주의 방식을 배제하고 반드시 계산주의 방식으로 구현된다고 보아야 할 어떤 선험적 이유도 없다. 계산주의와 연결주의의 차별성을 강조하는 위의 Cumins와 Schwarz의 주장 속에서도 연결주의가 입력과 출력을 연결시키는 기능, 즉 함수 기능을 문제삼고 있다는 것은 부인되지 않고 있다.(부인되는 것은 단지 그것이 계산 가능한

함수 기능이어야 한다는 것일 뿐이다.)

결국 필자는 연결주의도 계산주의와 마찬가지로 존재론적 기능주의 위에서 고려되는 인지 모형인 한, 기능주의에 커다란 위협이 되고 있는 심리 상태의 가족 유사성 문제를 피해갈 수 없다고 본다. 그러므로 이 글을 통해서 필자는 인지 과학의 철학적 기초와 불가분적으로 관련되어 있는 가족 유사성의 문제를 다루어 보고자 한다.

이를 위해 다음절에서는 먼저 가족 유사성 개념에 근거한 힐러리 퍼트남의 계산적 기능주의 비판에 대해 살펴보고자 한다. 그것을 통해, 계산적 기능주의 비판이 그대로 연결주의 비판으로 이어질 수 있음을 보일 것이다. 그러나 연결주의자들도 항변할 말이 없는 것은 아니다. 만일 연결주의 모형이 진정으로 학습하는 능력을 가지고 있다면, 그런 모형은 유연성(plasticity)을 갖추고 있어 본질주의의 한계를 극복할 수 있을 것이다. 그래서 세번째 절에서는 연결주의 모형이 진정으로 학습하는 능력을 가지고 있는지에 대해서 살펴볼 것이다. 이를 위해, 또다른 기능주의 비판자인 허버트 드레이퍼스가 제시하는 비판에 대해 살펴볼 것이다. 결과적으로 이런 논의를 통해, 계산주의든 연결주의든 그것이 적절한 인지 모형이고자 하는 한 가족 유사성이라는 커다란 장벽을 넘어야 한다는 사실을 포괄적으로 음미하게 될 것이다. 그런 다음, 네번째 절에서는 이상과 같이 인지 과학에서 피할 수 없는 절박한 문제로 대두된 가족 유사성 문제를 과연 현재의 인지 과학적 자원인 연결주의와 계산주의를 가지고 극복할 수 있는지 고찰하여 볼 것이다. 필자는 이러한 전망을 위하여, 연결주의와 계산주의를 절충해 가족 유사성 문제에 관한 새로운 해결책을 모색하고 있는 스몰렌스키의 전략에 대해 살펴보고, 그것의 가능성을 평가해 보고자 한다.

2. 퍼트남의 기능주의 공략

퍼트남은 그의 책 『표상과 실재』에서 "기능주의"로 불리는 컴퓨터 유

비가 정신 상태의 본성은 무엇인가 하는 질문에 대답하지 못한다는 것을 논변하고 있다. 다시 말해, 정신 상태란 컴퓨터의 계산 상태와 동일시될 수 없다는 것이다. 이러한 퍼트남의 논지는 그가 부정하고자 하는 입장들을 살펴보면 더욱 분명하게 드러난다. 퍼트남이 부정하려는 입장들은 크게 다음의 세 가지로 요약될 수 있다.

1) 정신 현상-지향성, 의미, 지시, 진리 등등-은 물리적/계산적 속성이나 관계로 환원되지 않는다. ⇒ 환원주의 비판
2) 정신 현상은 원초적 현상이 아니다. 즉 어떤 특정한 정신 현상의 모든 경우들이 공통적으로 갖는 과학적으로 기술 가능한 속성-배후에 실재하는 궁극적 본성-은 없다. ⇒ 본질주의 비판
3) 정신 현상은 정신 현상의 한 전형적인 예인 진리가 제거될 수 없는 한) 신화적인 것으로 제거될 수 없다. ⇒ 제거주의 비판

여기서 중요한 것은 이들 세 가지 종류의 비판이 기능주의 비판과 맺고 있는 관계를 이해하는 것이다. 기능주의란 환원주의의 한 형태(즉 기능적 환원주의)이다. 그러므로 기능주의의 논박은 환원주의 비판만으로도 충분하며, 그 이상의 비판(본질주의 비판과 제거주의 비판)은 불필요할 듯 보인다. 하지만 퍼트남은 환원주의 비판을 본질주의 비판을 토대로 하여 수행하고 있다. 즉 그의 환원주의 논박은 본질주의 논박에 의존적이다.

이 점을 좀더 구체적으로 살펴보자. 본질주의란 동일하게 분류되는 대상들의 배후에 항상 단일하게 기술 가능한 공통 속성이 존재한다고 보는 생각이다. 이를테면 금이라고 분류되는 모든 대상들의 배후에는 그것들을 금답게 하는 공통 속성이 언제나 존재한다고 보는 생각이다. 이런 본질주의를 심리 상태에 적용하게 되면, 동일하게 분류되는 심리 상태들의 배후에는 언제나 과학적으로 기술 가능한 공통 속성이 존재한다고 보는 입장이 된다.

한편 기능주의는 심리 상태의 기능이 그와 같은 공통 속성에 해당된

다고 보고 있기 때문에, 결국은 심리 상태에 관한 본질주의적 가정을 받아들이고 있는 셈이다. 그런데 퍼트남은 본질주의가 잘못되었음을 논변하고 있다. 정신 현상이란, 설령 기능을 끌어들인다 하더라도, 과학적으로 단일하게 기술 가능한 원초적 현상이 못 된다는 것이다. 그러므로 퍼트남의 본질주의 비판은 곧바로 정신 현상이 컴퓨터의 알고리듬 같은 것으로 환원되지 않는다는 기능주의적 환원주의 비판으로 연결되게 된다. 다시 말해, 본질주의 비판은 퍼트남의 기능주의 비판에서 결코 생략될 수 없는 중심적인 역할을 수행하고 있는 것이다. 퍼트남의 구절을 인용하여 보자.

> 필자는 어떤 특정한 지향적 현상의 모든 경우들이 공통적으로 갖는 과학적으로 기술 가능한 속성이 없음을 보일 것이다. 이 입론으로서 필자는 "지시" 일반 혹은 "의미" 일반 혹은 "지향성" 일반의 모든 경우들이 소유하는 어떤 과학적으로 기술 가능한 "본성"이 있다는 것을 부인하려 하며; 또한 필자는 예를 들어 "이웃에 많은 고양이들이 있다고 생각함" 같은 어떤 한 유형의 특정한 지향적 현상의 모든 경우들이 공통적으로 가지고 있는 어떤 과학적으로 기술 가능한 속성이 있다는 것을 부인하려 한다.

실제로 퍼트남은 그의 기능주의 비판을 이 같은 본질주의 비판 만으로 마무리 지을 수 있었을 것이다. 하지만 그럴 경우, 그의 비판은 제거주의와 본질적으로 구별이 불가능해진다. 퍼트남이 환원주의 비판과 본질주의 비판에 더해 제거주의 비판을 덧붙인 이유는, 그럼으로써 자신의 입장을 제거주의와 구분코자 하였기 때문이다.

한편 퍼트남의 본질주의 비판에서는 비트겐슈타인의 가족 유사성 개념이 중요하게 등장하고 있다.

> [지향성과의] 보다 나은 비교는 비트겐슈타인에 의해 제시된 바 있는 용어 "게임"과의 비교이다. 일상언어 수준에서조차도 모든 게임들이 "공통적인 어떤 것" 즉 **게임임**을 가지고 있다고 말하는 것은 이상하다. 왜냐하면 어떤 게임들은 승패가 개입되어 있고, 다른 것들은 그렇지 않다. 어떤 게임들은 경기자들의 즐거움을 위해 행해지고, 다른

> 것들은 그렇지 않다. 어떤 게임들은 하나보다 많은 경기자들을 가지나, 다른 것들은 그렇지 않다; 등등. 같은 방식으로, 어떤 사람이 어떤 것을 지시했다고 우리가 말하곤 하는 모든 경우들을 (혹은 어떤 사람이 하나의 특정한 사물을 지시했다고 우리가 말하곤 하는 모든 경우에서조차도 이 경우들을) 자세히 검토할 때, 우리는 단어와 지시된 사물 간의 단일한 어떠한 관계도 발견하지 않는다.(제1장 도입부)

철학에서 "가족 유사성"이라는 말은 흔히 "본질"이라는 말에 반대되는 뜻으로 쓰인다. 앞서 본 것처럼, 본질이 동일하게 분류된 대상들 배후에 언제나 존재하는 공통 속성을 의미하는 것이라면, 가족 유사성은 동일하게 분류되는 대상들에 어떤 공통 속성도 없고, 다만 부분적으로 겹쳐지는 다수의 속성들만이 존재함을 함축하기 때문이다.

그런데 퍼트남은 하나로 분류되는 심리 상태들 사이에 어떤 항상적 본질도 없고, 단지 가족 유사성만이 존재함을 주장한다. 이 같은 주장을 위해, 그가 『표상과 실재』에서 제시하는 논거는 믿음의 차이 삭감 논변"(혹은 넓게는 총체론 논변)이다. 그러나 총체론 논변을 구체적으로 살펴보기 전에, 우선 가족 유사성과 복수 실현 가능성의 관계에 대해 살펴보고 넘어가자.

> 5장과 6장은 앞 내용의 기반 위에서 특히 의미 총체론(holism)에 대한 논변들의 기반 위에서 쓰여졌다. 이 장들의 목적은 정신 상태들이 구성적으로 유연할(compositionally plastic) 뿐만 아니라 (동일한 "정신 상태"가 원리상 동일한 물리적 구성을 갖지 않는 체계들의 속성일 수 있을 뿐만 아니라), 계산적으로도 유연하다-동일한 정신 상태(예를 들어, 동일한 믿음이나 욕구)가 원리상 동일한 계산적 구조를 갖지 않는 체계들의 속성일 수 있다-는 것을 논변하는 것이다. 물리적으로 가능한 체계들이 같지 않은 프로그램들을 가지면서 동일한 정신 상태 속에 있을 수 있기 때문에 정신 상태들은 문자 그대로 "프로그램"일 수 없다.(서론)

> 고양이 한마리가 매트위에 있다고 믿고 있을 때 사람들이 처해 있을 수 있는 여러가지 다른 물리적 상태들이 물리적 화학적 용어들로 명시될 수 있는 "공통적인" 어떤 무엇을 가질 필요가 없다는 점을 지적한 것이 바로 기능주의의 통찰이었다. 이러한 통찰과 꼭 마찬가지

> 로 여기에서의 우리의 논의의 결과는 다음과 같다: 고양이 한마리가 매트 위에 있다고 믿고 있을 때, 사람들이 처해 있을 수 있는 여러 다른 **계산적** 상태들이 **계산적** 용어들로 명시될 수 있는 공통적인 어떤 무었을 가질 필요가 없다. (제5장 도입부)

기능주의자들이 물리주의자들에게 가한 비판의 핵심은 기능의 물리적 복수 실현 가능성(multiple realizablity)이었다. 그런데 역설적으로, 이제 퍼트남이 기능주의자들에게 가하고 있는 공격의 핵심인 가족 유사성도 그 성격상 복수 실현 가능성과 유사하다는 것이 위 인용문의 요지이다.

물론 복수 실현 가능성과 가족 유사성은 전적으로 같은 것은 아니다. 왜냐하면, 복수 실현 가능성은 동일한 것이 여러 형태로 실현될 수 있음을 의미하기에 아직까지 본질주의와 양립 가능한 반면, 가족 유사성은 공통 속성의 존재 자체를 거부하기에 본질주의와 양립할 가망이 없기 때문이다. 그러나 복수 실현 가능성에 의해 물리주의가 공격당한 이유가 동일한 심리 상태를 기술할 어떤 물리적 공통 속성도 없다는 데 있었던 것과 마찬가지로, 가족 유사성에 의해 기능주의가 공격당하는 이유도 동일한 심리 상태를 기술할 어떤 기능적 공통 속성도 없다는 데 있기 때문에, 둘 사이에는 유사성이 존재한다.

그렇다면 왜 심리 상태들은 본질을 갖지 못하고 단지 가족 유사성만 지니게 되는 것일까? 퍼트남은 여러가지 논변을 제시하고 있지만, 그 가운데 무엇보다도 중요한 것은 총체론 논변이다.

퍼트남이 심리 상태의 본질 부재를 입증하고 그것의 가족 유사적 성격을 보이기 위해 근거로 제시하는 총체론 논변은 좁은 의미의 총체론 논변과 그것을 좀더 확장한 믿음의 차이 삭감 논변으로 구성되어 있다. 이 가운데 좁은 의미의 총체론 논변은 대략 다음의 주장으로 요약될 수 있을 것이다 : 용어들의 정의 가능성 거부, 믿음들의 전체 그물 구조, 그리고 이 둘이 밀접하게 연관되어 있음. 이에 대해 좀더 자세히 살펴보자.

용어들의 정의 가능성 거부에 관하여 : "총체론이 곧바로 제시하는 것은 대부분의 용어들이 정의될 수 없거나, 혹은 적어도 만일 "정의"로써 우리가 뜻하는 바가 단 한번만에 고정되는 어떤 것 즉 절대적으로 용어

의 의미를 규정하는 어떤 것이라 한다면, 대부분의 용어들이 정의될 수 없다는 것이다." (제 1장 2절 1)

믿음들의 전체 그물 구조에 관하여 : "경험적 의미를 갖는 것은 진술들의 총체적인 덩어리이며, 이러한 의미는 개별적 진술들이 갖는 경험적 의미들의 단순한 합이 아니다....우리가 기대하는 바는 믿음들의 전체 그물 조직에 의존해 있다. 만일 언어가 경험을 기술한다면, 그것은 개개 문장으로서 그런 것이 아니라, 하나의 그물 조직으로서 그런 것이다."(제 2장 2절 1)

이 둘의 연결 관계에 대하여 : "왜 총체론이 이러한 것[대부분의 용어들이 정의될 수 없다는 것]을 제시할까? 왜냐하면 믿음의 전체 덩어리가 완강히 반항하는 경험들에 직면하면 콰인이 얘기하는 바대로, 어느 곳이라도 수정 될 수 있기 때문이다. 비록 용어가 처음부터 명확한 정의를 통해 과학에 도입되었다 할지라도 용어가 단순히 정의항(definien)의 동의어일 경우에 그러할 수 있는 것처럼, 그 결과로 얻어진 정의가 영원히 진리의 특권을 지니는 것은 아니다....콰인이 말하듯이, **약정에 의한 진리는 문장들의 영속적인 특색이 될 수 없다**. 우리 믿음의 그물 조직 속에 있는 진술들이 수정되어야만 한다면, 우리는 고른 선택의 여지가 있다; 그리고 주어진 맥락에서 무엇이 가장 좋은 선택인가는 용어들의 전통적인 "정의"들을 조사함으로써 결정되어질 수 없다."(제 1장 2절)

일반적인 생각에 따르자면, 우리가 어떤 용어를 가지고 대상을 "지시"하는 것이 가능한 이유는 그 용어를 사용할 때마다 언제나 그 용어의 정의에 해당하는 속성들을 머리 속으로 떠올리기(머리 속에서 예화시키기) 때문이다. 이런 가정 아래서, 용어의 의미가 고정적이라면 그런 용어들을 사용하여 대상을 "지시"하는 우리의 심리 상태에는 필반하는 본질적 속성이 있다고 생각할 수 있다. 다시 말해, 용어의 의미가 고정적이라면 심리 상태에 관한 본질주의적 가정이 설득력을 가질 수 있다.

하지만 퍼트남은 용어들에 고정 불변한 정의를 부여할 수 없다고 주장한다. 더우기 믿음들의 경우, 그 의미는 전체론적 그물 구조 상의 상대적 위치에 의해 결정된다는 점을 지적한다. 어떤 믿음의 의미도 그것을

기술하는 용어들의 정의를 조사함으로 단번에 확정되어질 수 없다는 것이다. 이처럼 용어들이나 믿음들의 의미가 일정하지 않다면, 그런 용어들을 사용하거나 믿음들을 보유하는 심리 상태에 고정 불변의 속성이 있을 가능성은 없게 된다.

그러나 이상과 같은 총체론 논변에 대해 우리가 신중하게 고려할 만한 한가지 반론이 제기될 수 있다. 만일 용어의 정의가 수정가능하다는 것이 용어의 정의 가능성 거부에 대한 논거라면, 우리는 정의의 수정을 정의되는 용어의 의미(그리고 지시체)가 바뀌는 것으로 간주하면 된다는 반론이 그것이다. 즉, (피정의항의 표현이 바뀜으로써) 정의항의 의미가 바뀌었을 뿐만 아니라, (비록 피정의항의 표현은 바뀌지 않았지만) 피정의항의 의미(그리고 지시체)도 바뀐 것으로 생각하면 된다는 것이다. 그렇게 생각하면, 동일한 대상에 대해 서로 다른 정의를 함으로써 옛 정의를 수정한 것이 아니라, 서로 다른 대상에 대해 서로 다른 정의를 준 셈이므로, 굳이 옛 정의를 수정한 것이라고 말할 필요가 없게 된다. 이에 관한 한 구절을 인용하여 보면:

> 또 따른 전통적인 움직임은 "그래, 과학자들은 **운동량의 의미를 바꾸기로** 결정한거야"라고 말하는 것이다. 만일 이것이 상대성 이론의 채택 후 과학자들이 "운동량은 질량 곱하기 속도"라는 문장에 부여한 진리치가 바뀌었다는 것을 해명한다면, 그것은 우리가 지금 다른 물리적 크기에 관해 이야기하고 있다는 것을 함축한다. 그러나 그렇지 않다. 우리는 여전히 탄성 충돌 속에서 보존되는 크기와 동일한 옛 운동량에 관해서 이야기 하고 있는 것이다. 만일 "운동량"이 어떤 것을 지시한다면, 그 친숙한 운동량이 항상 지시되는 물리적 크기인 "운동량"이다. 그리고 **운동량 자체**인 그 크기가 질량 곱하기 속도와 같은 양이 아니라고 판명되었던 것이다.(제2장 2절 1)

이처럼 피정의항 자체가 변했다는 반론에 대응하기 위해서 믿음의 차이 삭감 논변이 도입된다. 좁은 의미의 총체론 논변이 용어의 정의나 믿음의 의미에 대한 공시적 맥락의 영향력을 보여준다면, 믿음의 차이 삭감 논변은 통시적 맥락의 영향력을 보여준다. 즉 "의미들은 본질을 가지

고 있지는 않지만 시간의 변화 속에서도 동일성을 유지하고 있다"는 것이다.(제 2장 2절 1)

이처럼 의미의 동일성이 유지되는 이유는 우리가 해석에 관한 관용의 원리를 받아들이고 있기 때문이다. 관용의 원리란, 가급적 단어를 말하는 사람이 지니게 될 참인 믿음의 수효가 극대화되도록 그것의 의미를 해석해야 한다는 요구이다. 퍼트남은 이 같은 관용적 해석의 요구에 의해서, "운동량", "전자", "식물"과 같은 단어들의 의미가 동일성을 유지할 수 있었음을 예로 제시한다.

> 동의성(synonymy)에 대한 이론은 해석에 관한 질문들에 대답하는 이론일 것이다...."운동량(monentum)"이라는 단어를 사용했던 과학자들은 운동량을 "질량 곱하기 속도"의 동의어로보다는 보존된 양에 대한 이름으로서 사용하였다는 사실은 이미 언급되었다. 다른 예는 보어가 1934년에 "전자(Elektron)"라는 단어를 사용하였을 때 그는 1900년에 그가 "전자들"이라고 불렀던 것과 동일한 입자들에 관해서 얘기하고 있었다는 사실에 대한 우리의 지식이다. 우리는 이것을 그러한 다른 두 시기에 보어가 제시하였던 전자들에 대한 **이론들과 기술들**을 비교함으로써 그리고 그것들이 매우 **같다**는 것을 봄으로써 결정하는 것이 아니다. 왜냐하면 그것들은 서로 같지 않기 때문이다.... 간략히 말해, 그는 이 이야기를 동일한 대상에 관한 믿음의 연속적인 변화들의 이야기로 말하고 있는 것이지 "의미의 연속적인 변화들"에 관한 이야기로 말하고 있는 것은 아니다. 그리고 이 이야기에서 나중 연구 프로그램들을 앞 연구 프로그램들의 연장으로 다루려는 결정에 연계되어 있는 것과 동일한 종류의 "일반 지능"이 이러한 모든 "전자"의 사용들을 동의적으로 다루려는 결정에도 연계되어 있다: 이것이 이론 평가에 있어 중심적인 역할을 담당하는 한 종류의 결정이다. 사실상, "전자"를 이러한 모든 이론 변화 속에서도 적어도 그것의 지시대상은 본래대로 보존된 것으로 취급하는 것과 보어의 1934년 이론을 그의 1900년 이론의 진정한 계승 이론으로 취급하려는 것은 실제로 동일한 결정이다;
>
> 이러한 결정은 해석에 있어서 "관용(寬容)" 혹은 "의심점의 선의의 해석(benefit of the doubt)"이라 불려온 바를 보여준다. 1900년의 보어가 우리가 "전자들"이라고 부르는 것을 지시한다고 해석할 때, 우리는 그럼으로써 그가 1900년 믿음들 중 적어도 약간을 우리들의 견지에서 "참"이 되도록 만들고 있는 것이다. 이에 반해 그가 존재하지 않

> 는 대상들을 지시하고 있다고 해석하는 것은 그의 1900년 믿음들 모두를 완전히 틀린 것으로 파기해 버리는 것이다. 물론 1934년의 보어는 그의 1900년 자신에 대해 우리와 마찬가지로 같은 "관용적인" 태도를 계속 견지하였다. (이것이 그가 그의 모든 논문에서 단어 "전자"를 계속해서 사용한 이유이다.)
>
> 모든 해석은 관용에 의존한다. 왜냐하면 우리는 해석할 때 적어도 믿음에 있어서 약간의 차이는 항상 삭감해야 하기 때문이다. 예를 들어, 우리가 200년 전에 영어로 쓰여진 소설을 읽고 있는 동안 "식물"이라는 단어를 만났다고 가정하자. 정상적인 문맥에서는, 이 "식물"을 우리의 현재 단어 "식물"과 동일시하는 것을 주저하지 않을 것이다; 그렇지만, 그렇게 함으로써, 우리는 믿음에 있어서의 상당한 차이를 무시하고 있는 것이다. 예를 들어, 우리는 식물들이 엽록소를 가지고 있다고 믿으며, 광합성과 이산화탄소-산소 주기(carbon dioxide - oxygen cycle) 등등에 관해 알고 있다. 이러한 모든 것들은 200년 전에는 알려져 있지 않았었다. 그렇지만 (만일 우리가 철학자들이거나 철학적인 과학사가들이 아니라면) 우리는 200년 전 사람들이 "다른 세계에 살았다"고 말하지 않을 것이며, 그들의 개념이 우리가 가지고 있는 개념들과 "불가공약적"이라고 말하지 않을 것이다. 불가공약적이라는 것을 문자 그대로 받아들이면 (물론 결코 그럴 수 없지만), 그것은 우리가 200년 전에 쓰여진 일상 편지를 해석할 수 없음을 함축한다. 간략히 말해, 우리는 **식물**이라는 개념을 본질은 없지만 시간의 변화 속에서도 동일성을 유지하는 것으로 간주하며, **전자**라는 개념도 본질은 없지만 시간 변화 속에서 동일성을 유지하는 것으로 간주한다.(제1장 2절 2)

그런데 이런 믿음의 차이 삭감 현상을 해결하기 어려운 문제로 만드는 것은, 관용적 해석 못지 않게 때로는 의미의 불연속적 단절 역시도 일어날 수 있다는 사실이다. 즉 단어들은 언제나 관용의 원리를 따라 선의의 해석을 부여받는 것이 아니라, 어떤 경우엔 약간의 차이 밖에 없음에도 불구하고 전혀 동일하지 않은 것으로 취급되기도 한다. 퍼트남의 말을 직접 들어보자.

> 그렇지만 "해석상의 관용"이 항상 통용되는 것은 아니다. 우리는 항상 단어들을(우리의 견지에서) 말하는 사람이 가지게 될 참인 믿음들의 수효가 **극대화**되도록 해석하는 것은 아니다. "해석상의 관용"에

> 대한 반대 사례가 있다:....우리는 "플로지스톤 이론가들이 원자가 전자들에 대해서 얘기하고 있었으며, 그것들은 단지 약간의 잘못된 속성들을 가지고 있었을 뿐이다."라고 말할 준비가 되어 있지 않다. 그것은 과도한 관용일 것이다. 어떤 것은 합리적 관용이고 어떤 것은 과도한 관용이라는 것을 안다는 것은 그 맥락이 해석이든 "실제 삶"이든 간에 우리가 지닌 이해의 전 능력을 나타내고 있는 것이다. 그러한 어려운 경우에도 잘 적용되고 아울러 우리의 "일반 지능"에 대한 해명으로부터도 독립된의미와 지시의 동일성 이론을 세울 희망은 없다.... 앞의 예들이 보여주듯이 동의성과 일치하는 계산적 관계는 포더의 의미에서 "단위적"일 수 없다; 즉 "일반 지능"보다 심리학적으로 더 기초적일 수 없다.(제1장 2절 2)

결과적으로, 퍼트남이 말하고자 하는 바는 용어의 정의나 믿음의 내용 확정에 관련된 어떤 일반 원리도 있을 수 없다는 것이다. 다시 말해, 단어의 의미나 믿음의 내용이 확정되는 과정은 일종의 역사적/사회적 우연이라는 것이다. 따라서 그런 역사적/사회적 우연의 결과를 토대로 개념이나 믿음의 본질을 파악하려는 것은 부질없는 노력일 것이다.

이를테면 단어 "금"을 사용할 때 우리가 떠올리게 되는, **노랗고 무겁고 단단하다**는 등의 속성은 결코 **금**이라는 개념의 본질이 아니고, 다만 역사적/사회적 우연에 의해 우리의 금 개념에 포함된 생각일 뿐이다. 그러므로 이 속성들에 대한 생각이, 금을 "지시"하거나 금에 관한 "믿음"을 보유하는 모든 심리 상태들에 필연적으로 동반되야 하는 것은 아니다.

지금까지 논의된 (넓은 의미의) 총체론 논변에서 부각되었던 문제점들은 **용어들의 정의 가능성 거부, 믿음들의 전체 그물 조직 구조, 의미들이 믿음들의 차이에도 불구하고 동일성을 유지하고 있음, 의미와 지시의 동일성을 식별하는 것은 쉽게 한가지 원리로 일반화시킬 수 없는 이해의 전 능력을 요하는 작업이라는 점** 등이었다. 이제 이와 같은 퍼트남의 계산적 기능주의 비판이 어떻게 연결주의 비판으로 이어질 수 있는지 살펴보자.

앞서 지적한 대로, 퍼트남은 총체론 논변에 토대를 둔 가족 유사성 개념을 가지고 심리 현상에 관한 본질주의를 공격하고 있다. 그러나 아마도 연결주의를 옹호하는 사람은, 연결주의의 강점이 바로 본질주의를 거

부하고 가족 유사성을 수용하고 있다는 점에 있으므로, 퍼트남의 비판이 계산주의에는 해당되지만 연결주의와는 무관하다고 말할지도 모른다. 즉 1절에서 언급된 연상 기억에서 보듯이, 병렬 처리 모형은 엄격한 알고리듬에 의해 정보를 처리하는 것이 아니라, 주어진 상황에 따라 확률적 고려에 바탕을 두고 유연하게 작업을 수행한다. 그러므로 확률적 고려가 개입되는 병렬 처리 모형에서는, 동일한 입력으로부터 동일한 출력으로의 매개 과정이라 하더라도 꼭 동일한 중간 메카니즘을 거쳐야 하는 것은 아니다. 이 때문에, 동일한 종류의 처리 과정들 사이에는 본질이라 할 공통성보다는 가족 유사적 근사성 만이 존재한다고 주장할 여지가 있다.

그러나 퍼트남이 개진하고 있는 총체론 논변은 그러한 수준의 공시적 가족 유사성에 머무르는 것이 아니다. 그는 단순히 속성의 공통성을 토대로 심리 상태가 일반화될 수 없다는 것만이 아니라, 어떤 물리적/계산적 속성이든 시간적으로 계속 변하기 때문에, 도대체 물리적/계산적 속성을 가지고는 심리 상태를 특징지울 수 없다는 말을 하고 있는 것이다.

그렇지만 이런 주장에 대해서도 한가지 반론이 가능할지 모른다. 앞서 말한 것처럼 연결주의 모형이 학습할 수 있다면, 다시 말해 계속해서 정보 처리의 방식 자체를 수정해 나아갈 수 있다면, 인간의 인지 처리가 시간적으로 변화무쌍하다는 점만 가지고 연결주의 모형을 적절한 인지 모형이 아니라고 주장하는 것은 성급한 결론일 수 있다는 것이다.

물론 연결주의 시스템이 진정으로 학습할 수 있다면 이런 반론도 가능하다. 그러나 연결주의 시스템을 두고 과연 정말로 학습을 수행한다고 말할 수 있을까? 이런 의문에 대해 부정적인 평가를 내리고 있는 사람이 바로 허버트 드레이퍼스이다. 그러므로 다음 절에서는, 연결주의 모형이 학습할 수 있다는 주장에 반대하는 드레이퍼스의 견해에 대해 살펴보기로 하자.

3. 드레이퍼스의 연결주의 비판

드레이퍼스는 그의 책 *What Computers Still Can't Do*에서 GOFAI (good old fashioned AI/전통적인 계산주의적 인공지능 모형)에 대한 대안적 접근으로 세 가지를 꼽고, 그 세 가지 모두가 한계를 지니고 있다고 논변한다: 첫번째 대안은 기호처리 모형에 하이데거적 시각을 도입한 하이데거적 인공지능이며, 나머지 두 대안은 연결주의적 대안으로 하나는 감독 학습(supervised learning)이라고 불리며, 다른 하나는 강화 학습(reinforcement learning)이라고 불린다.

필자는 이들 세 부류의 대안들에 대한 드레이퍼스의 비판을 (대부분 그의 말을 그대로 인용하여) 소개하는 데 이 절을 할애할 것이다. 우선 이 세 가지 대안에 대한 드레이퍼스의 소개부터 들어보자.

> 대안적 접근에 촛점을 맞추고 있는 세 그룹의 인공지능 연구자들에게는 GOFAI는 이미 끝났다. Philip Agre와 David Chapman의 연구와 연계되어 있는 한 접근은 문맥 독립적인 기호 표상들이나 내부의 모형 토대적 계획을 사용하지 않고 미시 세계와 지능적으로 상호작용하는 프로그램을 생산하도록 시도하고 있다. 신경망 모형가들에 의해 대표되는 두번째 그룹은 표상을 통째로 포기한다. 이 접근은 전형적인 면모들을 사용하거나 외부로부터 전문가에 의해 제공된 예들에 의해 주어진 매핑을 이용하여 입력들로부터 직접적인 매핑에 의해 출력을 산출한다. 강화 학습이라고 불리는 세번째 새로운 접근은 성공적인 입력-출력 규칙을 스스로 찾기 위해 외부로부터 전문가를 이용하지 않고 숙련 영역에서 실제적인 실행을 이용하는 프로그램을 발전시키는 것을 목표로 하고 있다. 이 접근들 각각의 장점들과 한계들은 고려하라 만한 가치가 있다.

드레이퍼스는, 지금까지의 하이데거적 인공지능은 그것이 생각하는 것-장기적인 계획 그리고 문맥 독립적인 면모들을 가진 대상들의 내부적 표상들-에 있어서는 하이데거 현상학에 충실하나, 지능적인 체계가

필요로 하는 능력, 즉 숙련 영역에서 적절한 구별을 해내고 경험으로부터 새로운 구별을 배우는 능력을 결여하고 있어 한계를 노출하고 있다고 진단한다. 그리고 그러한 핵심적 능력을 제공하기 위해 더욱 많은 연구자들이 신경망 연구에 관심을 돌리고 있으므로, 이제는 연결주의 신경망이 그가 친숙성(familiarity) 혹은 총체적 민감성(global sensitivity)이라고 부른 바를 보여줄 수 있는지, 그리고 만약에 보여줄 수 없다면 연결주의가 어떤 다른 방식으로 적절성과 학습에 대처할 수 있는지, 이런 질문을 숙고해보아야 한다고 제안하고 있다.

드레이퍼스에 따르자면, GOFAI에서 그랬던 것처럼 연결주의에서도 상식적 지식의 문제가 골칫거리로 등장한다. 퍼셉트론(perceptron)식의 다층적 신경망 설계자들은 지능적인 신경망이 일반화의 능력을 갖추어야 한다는 점에 모두 동의하고 있다: 예를 들어 주어진 분류 작업에 있어서 하나의 특정한 출력과 연계된 입력의 충분한 예들이 주어지면, 그것은 동일한 유형의 새로운 입력들에 동일한 출력을 연계시킬 수 있어야만 한다는 것이다.

그러나 이 대목에서 한가지 문제가 생긴다. 그것은, 무엇이 어떤 근거에서 동일한 유형으로 간주되는가, 하는 질문이다. 신경망의 설계자는 보통 합리적인 일반화에 요구되는 "유형"의 특정한 정의를 염두에 두고 있으며, 만일 신경망이 이 유형의 다른 사례들을 일반화시킨다면 그것을 성공으로 간주한다. 그러나 신경망이 예기치 않은 연상 관계를 산출했을 때, 과연 그 망이 일반화에 실패했다고 말할 수 있을까? 신경망은 "유형"의 다른 정의에 기초해 작업을 수행해 왔을 뿐이고, 방금에야 그런 정의의 차이가 드러났다고도 볼 수 있을 것이다.

이상과 같은 이유 때문에, 드레이퍼스는 비록 인공 신경망이 일반화를 위한 내적 처리 구조를 형성할 수 있기는 하지만, 그런 처리 구조의 학습에는 신경망을 훈련하는 감독자의 암묵적인 배려가 개입되지 않을 수 없다는 점을 지적한다. 이런 배려에 의해 비로소 인공 신경망은 특정 "유형"에 관한 의도된 정의를 학습할 수 있게 된다는 것이다.

물론 이런 배려에도 불구하고 종종 인공 신경망은 엉뚱한 정의에 기

초해 일반화를 수행하곤 한다. 드레이퍼스는 이를 보여주는 재미있고 극적인 사례를 한가지 소개한다. 연결주의 연구의 초창기에 미국 육군은 숲속에 있는 탱크를 인식하도록 인공 신경망을 훈련시키려고 노력하였다. 육군의 연구자들은 탱크들이 없는 숲속의 사진을 많이 찍어왔다. 그리고는 며칠 뒤 탱크들이 숲속의 나무들로부터 삐쭉이 돌출해 있는 사진들을 찍어왔다. 그리고는 두 부류의 사진들을 식별할 수 있도록 신경망을 훈련시켰다. 결과는 인상적이었다. 신경망이 훈련시에 사용되지 않았던 사진들에 대해서까지 탱크의 유무를 성공적으로 판정해내는 것 같았기에 더욱 감동적이었다. 연구자들은 신경망이 진정으로 숨겨진 탱크들을 인식하고 있는 것인지 확인하기 위해서, 동일한 숲에서 더 많은 사진을 찍고 그 사진들을 훈련된 신경망에게 보여주기로 하였다. 그렇지만 연구자들은 이보다 강화된 테스트의 결과에 실망하지 않을 수 없었다. 신경망은 나무들 사이에 탱크가 서 있는 새로운 사진과 탱크가 없는 나무들의 사진을 구별하지 못하였다. 숙고 끝에 한 연구자가 탱크가 없는 숲속의 원래 사진들은 흐린 날씨에 찍혔고 탱크들이 있는 원래 사진들은 맑은 날에 찍혔다는 것을 발견하자 그 비밀은 마침내 풀렸다. 훈련을 통해서 신경망은 탱크가 있는 숲과 탱크가 없는 숲의 차이를 인식하고 일반화한 것이 아니라, 그림자들이 있는 숲과 그림자들이 없는 숲의 차이를 인식하고 그것을 일반화하였던 것이다!

위와 같은 예를 통해서 드레이퍼스가 제시하고자 하는 연결주의 비판의 핵심은, 신경망을 통한 일반화도 종국적으로 신경망 외적인 인간의 이해 능력에 의존하지 않고서는 적절하게 수행되기 어렵다는 것이다. 보다 구체적으로 말해, 어떤 일반화가 문맥에 적절한 일반화인가 하는 문제에 봉착해서는 인간의 이해 능력에 의존할 수밖에 없다는 것이다. 신경망 모형 설계와 관련된 이런 연결주의의 난점을 드레이퍼스로부터 직접 들어보자.

> 신경망 모형 설계자들은 초창기에는 그들의 망이 훈련될 때까지는 백지와 같아서 설계자가 미리 훈련된 지능과 닮은 어떠한 것도 제공할 필요가 없다고 기뻐했었다. 그렇지만 최근에는, 인간이 하는 것과

같은 적절한 일반화를 산출하는 문제에 있어서, 가능한 일반화의 부류가 적절한 선험적인 방식으로 제한되지 않는다면 인간의 일반화와 닮은 어떠한 것도 신뢰할 만하게 기대될 수 없다는 것을 인식하게 되었다. 결과적으로, 문제(가설 공간)에 적절한 인간스러운 허용 가능한 일반화의 부류를 미리 규정한 후에, 신경망 모형 설계자들은 신경망들이 가설 공간 속에 규정되어 있는 방식대로만 입력을 출력으로 변형하도록 그들의 신경망 설계를 꾀한다. 그렇다면, 일반화는 설계자의 견지에서만 가능할 것이다. 가설 공간의 적절한 원소를 단일하게 알아내기 위해서는 약간의 예들만으로는 불충분하겠지만 충분한 예들을 학습시킨 후에는 오로지 하나의 가설만이 일반화 원리를 배울 것이다. 즉 모든 새로운 입력은 설계자의 관점에서 올바른 출력을 산출할 것이다.

여기서 문제는 망의 구조 설계에 의해 어떤 가능한 일반화들이 결코 발견되지 않도록 설계자가 결정했다는 것이다. 이러한 것은 무엇이 합리적인 일반화를 구성하는가 하는 것이 문제되지 않는 장난감 문제에는 괜찮을지 몰라도, 실재-세계 상황에서는 인간 지능의 많은 부분이 문맥에 적절한 방식으로 미리 정의한 부류에 제한시킨다면, 망은 그 문맥에 대해 설계자에 의해 그 망에 부여된 지능을 보일 것이나, 진정한 인간의 지능이 하는 것과 같이 다른 문맥에서도 적용할 수 있는 상식은 가지지 못할 것이다.

아마도 만일 신경망이 우리의 의미의 적절한 일반화를 공유한다면, 신경망은 인간의 두뇌와 크기, 설계 구조, 그리고 초기-연결 모습을 공유할 것이다. 진정으로 신경망 연구자들은 간헐적으로 잠정적인 성공을 거두고 있으나, 일반화시키는 원리적 방식을 가지고 있지 못한데, 이것은 내가 1960년대 GOFAI에 대해서 썼을 때의 GOFAI 연구자들의 단계인 것처럼 보인다. 한때 등한시 되었다가 다시 부활한 연결주의 접근은 단지 기회를 잃어가고 있는 것처럼 보인다.

인간이 하는 방식으로 일반화하기 위해서는 망의 설계 구조는 인간에게 적절한 면모들의 견지에서 상황들에 반응할 수 있도록 설계되어야만 할 것이다. 이러한 면모들은 과거 경험이 중요하다고 보여주는 바와 그 상황이 고찰될 관점을 결정하는 현재의 경험들에 토대를 두어야만 할 것이다. 그렇게 할 경우에만 망은 상황 속에서 현재 제시되지 않은 기대되는 입력들뿐만 아니라 (숲 속의 탱크들과 같이) 기대되지 않은 입력들의 인지를 허용하는 지평-토대적인 인간과 같은 상황 속으로 들어간다. 현재의 망은 이러한 어떠한 능력도 보여주지 않고 있으며, 우리 두뇌의 설계 구조가 그것을 어떻게 산출해 내는지에 대해 현재 누구도 알지 못하며 추측조차 하지 못하고 있다.

감독 학습과 관련된 연결주의의 난점을 드레이퍼스가 지적한 부분을 인용하여 보면:

> 신경망의 감독된 학습을 통한 인공지능에로의 길은 다른 기본적인 문제가 있다. GOFAI에서 시스템이 어떤 지능을 보이던 간에 그것은 시스템 설계자에 의해 명시적으로 규정되고 프로그램되었다. 시스템은 그것이 배운 규칙들이 부적절한 상황을 인식하고 새로운 규칙을 구성하는 독립적인 학습 능력이 없다. 신경망은 학습 능력이 있는 것처럼 보인다. 그러나 감독된 학습의 상황에서 지능을 제공하는 것은 어떤 경우들이 좋은 예들인가를 결정하는 사람인 것이다. 신경망이 배우는 것은 단지 연결강도에 의해 이 지능을 어떻게 포착하느냐 하는 것일 뿐이다. GOFAI체계와 마찬가지로, 망은 그러므로 그것들이 배운 것이 부적절한 상황을 인식하는 능력을 결여하고 있다: 그 대신 실패를 인식하고 망이 이미 훈련된 상황의 산출을 수정하거나또는 행동에 있어서 적절한 수정으로 이끌 새로운 샘플을 제공하는 것은 인간 사용자에 달려 있다....우리가 진정으로 필요한 것은 스스로 환경에 어떻게 대처할 것인가를 배우고 환경이 변함에 따라 그들 스스로의 반응을 수정하는 체계인 것이다.

감독 학습의 이러한 난점으로 인해, 드레이퍼스는 여기서 마지막 대안인 강화 학습을 고찰한다: 그에 따르면 강화 학습에서는 신경망이 별도의 인간 감독관을 따로이 필요로 하지 않기 때문에 강화 학습은 감독 학습보다 장점을 가지고 있다. 드레이퍼스의 설명을 들어보자.

> 이러한 필요를 만족시키기 위해, 최근의 연구는 종종 강화 학습이라고 불리는 접근에 관심이 모아지고 있다. 이 접근은 감독 학습에 비해 두 가지 장점이 있다. 첫째, 감독 학습은 장치에 각 상황에서의 올바른 행위가 무엇인지 주어지도록 요구된다. 강화 학습은 세계로부터 행동의 직접적인 비용과 이익을 측정하는 강화 신호를 제공받을 뿐 다른 어떠한 것으로부터도 강화 신호를 제공받지 않는다고 가정한다: 그리고 강화 학습은 문제를 푸는 동안 그것이 받는 전체 강화를 최소화하거나 최대화한다. 이러한 방식으로, 그것은 장기적인 목적을 달성하기 위해 다양한 상황들에 취할 최적의 행동들을 경험으로부터 점차적으로 배운다. 그렇게 하면 숙련되게 대처함을 배우기 위해서, 장차의 모든 것을 알고 있는 선생을 필요로 하지 않으며 단지 세계로부터

> 의 피드백만을 필요로 한다. 둘째로, 감독 학습에서는 숙련 환경에서의 어떠한 변화도 새로운 환경에서 무엇을 할지를 알고 있는 전문가에 의한 새로운 감독을 요구한다. 강화 학습에서는, 새로운 조건들은 강화에 있어서의 변화를 꾀하여 자동적으로 적절히 적응하도록 장치를 인도한다.

그러나 강화 학습의 우월성에도 불구하고, 강화 학습도 극복되기 어려운 문제점들을 안고 있다고 드레이퍼스는 주장한다. 그가 제기하는 문제는 특히 **특이한 상황들에 대한 대처 문제와 적절성의 순환 문제**라는 두 문제이다. 그런데 이 두 문제는 근간에 있어 앞에서 감독 학습과 관련해서 나타났던 문제-숙련되게 대처하기 위해서는 상황에 적절한 일반화의 능력이 있어야 하며, 그러기 위해서는 상황에 대한 총체적 민감성을 지녀야 한다는 문제-가 반복되어 나타난 것에 불과하다.

> 강화 학습 아이디어가 숙련되게 대처함을 학습하는 것과 관련된 인간 지능의 본질을 올바로 포착한다고 가정할 때, 다음과 같은 질문이 자연스럽게 대두된다: 강화 학습의 현상적으로 합당한 최소한의 본질을 이용하여 적어도 특정한 영역에서 인간 전문가와 같은 훌륭한 장치를 만들 수 있는가? 현재의 실제 작업과 관련하여 두 가지 개선이 적어도 필요한데, 그 어느 것도 현재의 지식에 토대해서는 달성될 수 없는 것처럼 보인다. 첫째, 만일 학습 동안 실제로 마주친 상황들의 수효보다 훨씬 많은 상황들 하에서 발생한 문제들에 강화 학습이 적용되려면, 새로운 특이한 상황에 꽤 정확한 행동들과 가치들을 할당하는 어떤 방법이 필요하다. 둘째, 만일 강화 학습이 인간 지능과 닮은 어떤 것을 산출하도록 하려면, 강화 학습 장치는 어떤 관점(지평) 하에서 상황과 조우하고 또 적절한 입력을 적극적으로 찾음으로써 총체적인 민감성을 보여야만 한다.
>
> 첫째로, 특이한 상황들에서의 행동의 문제를 고려하자. 이 문제는 두 가지 절차들에 의해 그 해결이 모색되고 있다. 첫번째 절차는 자동 일반화 절차로, 이것은 다른 상황들에 대해 학습된 행동들과 가치들의 토대 위에서 앞서 자주 마주치지 않았던 상황들에서의 행동들과 가치들을 산출하는 절차이다. 두번째 절차는 상황의 전체 면모들 중에서 단지 적절한 부분 집합의 토대 위에서만 행동을 산출하고, 그러한 적절한 면모들과는 관계 없이 동일한 적절한 면모들을 공유하는 모든 상황들에 대한 경험들의 토대 위에서 행동들이 선택되고 가치들

이 학습된다. 이 두 접근 모두 만족스럽지 못하다. 자동 일반화 절차에 관련해서는, 일반화가 요구되는 바로 그 시점에서 상황은 감독 학습에서 지적했던 상황과 동일하다. 누구도 인간이 지능이 하는 것과 같은 방식으로 일반화하는 망이나 또는 어떤 다른 기제(mechanism)를 어떻게 얻을 수 있는지 전혀 알지 못한다.

위에서 언급된 둘째 문제-상황의 무슨 면모들이 적절한 부분 집합으로 간주되며 행동들과 가치를 결정함에 있어서 사용되어야만 하는가 하는 것을 학습하는 문제-는 마찬가지로 어렵다. 현재 사태의 어떤 면모들이 적절한가 하는 것은 이 사태가 무슨 종류의 상황인가를 결정함으로써만 알아낼 수 있다. 이 문제는 **적절성의 순환**이라고 불릴 수 있을 것이다. 그 함축들을 잘 파악하기 위해, 야구팀의 구단주가 다양한 조건들 하에서 각 선수들이 보여준 경기 성적에 관한 사실들을 담은 컴퓨터를 팀 매니저에게 주었다고 상상해보자. 어느날, 9회말 느지막하게 컴퓨터를 조회해 본 후 매니저는 현재 타자 A를 대타 B로 대체하기로 결정한다. 대타는 홈런을 치고 팀은 경기에서 승리한다. 그렇지만 구단주은 화를 내면서 컴퓨터를 잘못 사용하고 있다고 매니저를 비난한다. 왜냐하면 컴퓨터의 기록에 따르면 B가 A보다 분명히 낮은 타율을 보이고 있었기 때문이다. 그러나 매니저는 컴퓨터에 따르면 B가 낮 경기에서는 보다 높은 타율을 보이고 이것은 분명히 낮 경기였다고 대답한다. 구단주는 그건 그렇지만 그가 좌완 투수에 대해서는 보다 낮은 타율을 보이고 있으며 오늘 마운드에 좌완 투수가 있었다고 대답한다. 등등. 논점은 매니저의 전문성은 그리고 전문가들의 전문성 일반은 적절한 사실들에 반응할 수 있음에 놓여 있다는 것이다. 컴퓨터는 매니저가 기억할 수 있는 것보다 많은 사실들을 제공함으로서 도움을 줄 수 있다. 그러나 오로지 경험만이 매니저로 하여금 현재 사태를 특정한 상황으로 파악할 수 있게 하여주고 또 무엇이 적절한가를 파악할 수 있도록 하여준다. 전문가의 know-how는 보다 많은 사실들을 추가함으로써 컴퓨터에 입력할 수 없다. 왜냐하면 논점은 무슨 사실들이 적절한가 하는 것을 결정할 수 있도록 해주는 현재의 올바른 지평(관점)이 무엇인가 하는 것이기 때문이다.

드레이퍼스는 강화 학습의 한 구체적인 사례로 Chapman과 Kaelbling에 의해 제안된 절차를 고려하고 그 문제점들을 다음과 같이 지적하고 있다.

현재의 절차들은 시행착오적 학습 동안 어떤 통계를 추적함으로써

> 적절성에 관해서 배우도록 시도하고 있다. Chapman과 Kaelbling에 의해 제안된 절차는 어떠한 면모들도 행동과 가치 평가에서 적절하지 않다고 가정하고 출발한다. 즉 상황이 무엇이든지 동일한 행동이 취해져야하고 모든 상황의 각 가능한 적절한 면모에 대해, 절차는 그 면모가 각 가능한 값들을 가질 때 어떻게 일들이 진행되는가에 대한 통계를 추적한다. 만일 현행 통계의 토대 위에서 면모의 값들이 행동들과 가치들에 중요하게 의미있는 영향을 미치는 것처럼 보이면 이것은 적절하다고 선언된다. 상황을 적절하다고 발견된 면모들의 집합이 늘어남에 따라 보다 세밀하게 기술되는 것이다.
>
> ....그렇지만 위에서 기술된 특정한 절차에는 심각한 문제들이 있다. 첫째, 면모들은 그것 단독으로 행동에 적절하지 않고 하나 혹은 그 이상의 면모들과 결합될 때 적절할 수 있다. 이것을 고치기 위해, 면모들의 결합들의 적절성에 관한 통계들을 모을 필요가 있을 것이다. 그러나 이것은 가능적으로 중요한 통계들의 지수적 폭발에로 이끈다. 둘째, 이 접근은 면모의 적절성은 영역의 속성이라고 가정하고 있다: 측정되는 것은 마주치는 모든 상황들에서의 면모의 적절성이다. 그러나 면모는 어떤 상황에서는 적절하나 다른 상황에서는 적절하지 않을 수도 있다. 그러므로 우리는 각 상황에 대해 따로따로 적절성 데이타를 모을 필요가 있다. 그러나 이것 역시 통계의 양에 있어서 지수적인 증가에로 인도한다....셋째 문제는 어떤 상황에서 적절한 것으로 생각될 수 있는 면모들의 숫자에 한계가 없다는 것이다.

드레이퍼스는 자신의 공격으로 인해 계산주의와 연결주의가 처한 상태를 다음과 같은 딜렘마 상태로 정리하고 있다: "인공지능에서의 모든 연구는 깊은 딜렘마에 직면하고 있는 것처럼 보인다. 만일 GOFAI 체계를 세우려고 노력한다면, 단순히 숙련된 인간으로서 인간이 이해하는 모든 것을 믿음 체계 속에 표상해야만 한다는 것을 발견하게 된다. 그러나 이 책의 2판 서문에서 밝혔듯이, 인간이 이해하는 바를 충분히 명시적으로 만듬으로써 상식을 보이도록 컴퓨터를 성공적으로 프로그램한다는 것에 대한 극단적인 비가망성이 GOFAI 연구 프로그램에 대한 회의로 나를 이끌었다. 다행히도, 기계 학습의 최근 연구는 인간이 이해하는 모든 것을 표상하도록 요구하지 않는다. 그러나 우리가 방금 본 것처럼 딜렘마의 또다른 뿔에 직면한다. 인간이 행하는 방식의 일반화를 배우기 위해 인간의 관심가 인간의 구조를 충분히 공유하는 학습 장치를 우리는

필요로 한다."

지금까지의 논의를 요약해보자. 퍼트남이 전개하고 있는 기능주의 비판 논변의 배후에는 우리가 사용하는 개념과 믿음들의 의미가 불확정적이라는 통찰이 깔려있다. 개념들의 의미가 그처럼 유동적이라면, 그런 개념들을 통해 이루어지는 인간의 사고 과정에는 도대체 확고부동하고 불변적인 특징이란 있을 수 없게 된다는 것이다.

만일 연결주의가 이런 퍼트남의 비판으로부터 안전하려면, 연결주의 모형은 학습을 통해 끊임없이 변화를 수용할 수 있어야 한다. 이는 곧 연결주의 모형이 처리를 위해 어떤 확정된 프로그램도 갖지 않아야 한다는 말이다.

일견, 연결주의 모형에는 처리를 위한 어떤 확정된 프로그램도 있지 않은 듯 보인다. 그러나 드레이퍼스는 그 같은 고찰이 피상적임을 지적하고 있다. 연결주의의 두 가지 학습 방법, 즉 감독 학습과 강화 학습 모두에서, 암암리에 인간 사용자의 지식이 시스템에 주입되고 있으며, 이런 지식 없이 시스템이 과제를 수행하는 것은 불가능한 듯 보인다. 결국 연결주의 모형은 프로그램을 필요치 않는다기보다, 오히려 기존의 방식과 다른 방식으로, 즉 샘플을 통한 훈련이라는 방식을 통해 프로그램된다고 보아야 할 것이다.

4. 스몰렌스키 전략의 가능성

이제까지는 일방적으로 기능주의의 약점을 들춰내기만 하였다. 하지만 이제부터는 보다 동정적인 입장에서 기능주의를 고찰해보기로 하자. 앞에서 기능주의의 문제로 지적된 것은 심리 상태의 가족 유사성 문제, 다른 말로 하자면, 심리 상태의 기능적/물리적 구조가 끊임없이 유동적이라는 문제였다. 그러므로 이번 절에서는 그같이 유동적인 구조를 인지 모형에 반영하면서도, 그것을 심리 상태에 관한 본질주의적 가정과 조화시키려고 하는 스몰렌스키의 전략에 대해 살펴보도록 하겠다.

우선 논의에 앞서 생각해 보아야 할 것은 그처럼 본질주의가 많은 문제를 가지고 있다면 왜 그것을 포기하지 않고 굳이 고수하려 하는가 하는 점이다. 그 이유는 본질주의가 심리 상태를 분류하는 확고한 체계를 제공한다는 점에서 찾아져야 한다. 흔히 상식 심리학이라고 폄하되기도 하는 믿음/욕구 심리학이 그 한 예를 보여준다.

상식 심리학은, 이를테면 존이 왜 슈퍼마켓에 가는가에 대해, 그가 밀크를 원하고, 슈퍼에 밀크가 있다고 믿기 때문이다 라는 식의 대답을 하게 해준다. 이런 대답이 설명력을 갖는 이유는, 믿음과 바램의 주체가 누구이든, 밀크를 원하는 바램과 슈퍼에 밀크가 있다는 믿음을 가졌다면, 슈퍼에 가는 행동이 발생하게 된다는 가정이 암암리에 받아들여지고 있기 때문이다. 즉 상식 심리학의 가정은 믿음과 바램의 주체나 환경을 구분하지 않고 그것들이 동일한 심리 상태들로 분류되는 한 동일한 인과적 효력을 가지는 것으로 간주한다. 그런데 비본질주의자라면, 심리 상태란 시간에 따라 또 사람에 따라 달라지는 것이므로, 서로 다른 심리 상태들 사이의 인과적 공통 속성이란 있을 수 없다고 볼 것이다. 이와 같은 비본질주의자의 생각은 상식 심리학적 견지에서 보면 용납될 수 없다. 그러므로 심리 상태에 관한 상식적 분석 체계와 가정들을 보존하기 원한다면, 비본질주의를 배척하고 본질주의를 고수할 수밖에 없게 되는 것이다.

그렇다면 상식 심리학의 그같은 본질주의적 가정이 전통적인 AI 연구자들과 최근의 연결주의자들에게서 어떻게 이해되어 왔는지를 생각하여 보자. 먼저 전통적인 AI 연구자들을 보면, 그들은 상식 심리학의 본질주의를 의심 없이 받아들여져 왔다. 그렇기에, 고전적 AI에서는 한번 해석이 부여된 컴퓨터의 계산적 상태에는 언제나 어떤 맥락에서나 동일한 인과적 효력이 부여되었던 것이다. 그러나 이런 시도가 겪는 어려움은 인간 인지의 유연성을 컴퓨터의 처리 과정에 전혀 반영할 수 없다는 점이다. 이에 반해, 최근의 연결주의자들은 인공 신경망을 글자 그대로 유기체의 신경학적 처리 과정을 보여주는 모형이라고 이해하고 있다. 그렇기 때문에, 인공 신경망을 통해 인간의 인지 처리 과정을 연구할 수 있는 한, 믿음이나 바램 등에 근거한 상식적 설명 체계는 더 이상 필요 없다

고 주장하고 있다. 하지만 이들의 문제는 상식 심리학을 포기하는 댓가로 인지 처리의 규칙적 측면을 모의하는 데 어려움을 겪게 된다는 점이다.

스몰렌스키는 이상의 두 입장에 비해 보다 신중한 자세를 취하고 있다. 무엇보다도 그는 우리가 아직 인지 처리의 신경학적 과정에 대해 충분한 지식을 갖고 있지 못하다는 점을 인정한다. 그래서 그는 최근의 연결주의 모형이 보여주는 초보적 성공에 도취하여 기존의 상식적 설명 체계를 섣부르게 포기하기보다는, 전통적인 상식적 설명 체계와 연결주의 모형을 연결할 수 있는 길을 모색하고자 한다. 이처럼 연결주의 모형을 상식 심리학의 기본 가정과 공존 가능하도록 새로운 방향에서 해석하려 시도를 스몰렌스키는 *Proper Treatment of Connectionism*이라고 부른다.(스몰렌스키의 입장에 대한 전반적 정리라고 할 논문 "On the Proper Treatment of Connectionism"은 *Behavioral and Brain Science* 11에 실려있다.) 여기서는 먼저 스몰렌스키가 제안하고 있는 연결주의의 새로운 해석이 무엇인지 살펴보고, 그런 다음 그것이 앞에서 제기된 퍼트남의 가족 유사성 문제에 어떻게 대응할 수 있는지 살펴보기로 하자.

연결주의가 인지 모형으로 채택하는 인공 신경망은 엄밀하게 말하자면 해석이 부여되지 않은 구문론적 체계일 뿐이다. 그러므로 그것을 신경망이라고 부르는 것은 사실 어폐가 있다. 인공 신경망을 인지 연구를 위한 모형으로 삼고자 한다면 우선 다음의 두 가지 문제가 결정되어야 한다. 첫째로 상식적 설명 체계에서 분류의 기본적 단위가 되는 「개념」들을 신경망의 어떤 상태와 동일시할 것인가 결정해야 한다. 다음으로 상식적 설명 체계가 기본적인 인지적 과정으로 가정하고 있는 추론적 과정을 신경망의 어떤 처리 과정과 동일시할 것인가 결정해야 한다.

스몰렌스키는 이 두 가지 문제를 결정함에 있어, 인공 신경망을 하나의 동역학적 시스템이라는 관점에서 바라본다. 동역학적 시스템이 무엇이고, 또 어떻게 인공 신경망이 동역학적으로 이해될 수 있는지에 대해서는 보다 자세한 설명이 필요하겠지만 간략히 그 개요만을 말하자면, 동역학적 시스템에서는 시스템 전체의 순간적 상태가 시스템 처리 소자

들의 흥분 정도를 차원으로 갖는 복합적 상태 공간 상의 벡터로 표현되고, 시스템의 시간적 변화는 순간 상태를 표현하는 벡터의 이동 궤적으로 간주되게 된다. 이 같은 동역학적 해석에서 시스템의 행태를 특징짓기 위한 몇가지 개념들이 도입되는데, 그것이 바로 '어트랙터(attractor)'와 '상 전이(phase transition)'이다. 어트랙터는 글자 그대로 상태 공간 상에 존재하는 일종의 골짜기 같은 것으로 볼 수 있다. 그러므로 시스템의 순간 상태를 표현하는 벡터가 이 영역 부근을 지나게 될 때, 마치 골짜기에 빠지듯 어트랙터에 끌려들어 비교적 장시간 변화가 지체되게 된다. 그러나 시스템의 상태 변화가 어트랙터에 고착되어 영원히 머물러 있게 된다면, 정보 처리 도구로서의 신경망의 효능은 사라지게 될 것이다. 그래서 대부분의 신경망은 스토케스틱(stochastic) 함수를 동원하여 그 변화의 과정에 요동을 주도록 하고 있다. 이 요동 덕분에 시스템의 순간적 상태는 어트랙터의 골짜기를 벗어나 또다시 새로운 변화의 전기를 맞게 된다. 이처럼 순간적 상태 벡터의 변화가 하나의 어트랙터를 벗어나 다른 어트랙터로 이끌리게 되는 과정, 즉 상태 공간의 언덕을 넘는 과정을 상 전이라고 한다.

스몰렌스키는 이 같은 동역학적 해석을 바탕에 깔고 인지 처리의 단위가 되는 개념이나 명제들을 시스템의 어트랙터들과 동일시할 수 있다고 주장한다. 그리고 상식 심리학에서 명제들 사이에 존재한다고 가정된는 결정이론적/추론적 관계란 바로 명제에 해당하는 어트랙터들 사이에 존재하는 동역학적 변화의 특성, 즉 상 전이의 특성이라고 간주한다.

개념을 어트랙터와 동일시하게 되면, 감각 입력으로부터 개념이 예화되는 과정이 쉽게 이해될 수 있다. 그 과정은 본질주의자들이 주장하는 것처럼 개념적으로 정의된 판정 기준을 통해서 이루어지는 것이 아니고, 개념보다는 하위 수준에 있는 준개념적(subconceptual) 미시 특징들에 대한 무의식적/자동적 판단을 통해 이루어지는 것으로 볼 수 있게 된다. 그리고 그와 같은 무의식적/자동적 판단의 과정은 각각의 준개념들을 대표하는 시스템의 처리 소자들(processing units) 사이의 동역학적 상호작용을 통하여 설명되어진다.

한편 상식 심리학이 가정하는 명제들 사이의 결정이론적/추론적 관계를 어트랙터들 사이의 상 전이로 이해하게 되면, 고전적 AI의 가장 큰 골치거리였던 부드러운 제약(soft constraints)의 문제가 해결되어진다. 부드러운 제약이란, 조건과 그 귀결 사이의 관계가 연역적 추론 규칙이 가정하는 것처럼 필연적이지도 않고 그렇다고 완전히 우연적이지도 않은, 어느 정도의 확률적 개연성을 가지고 있을 경우를 말한다. 그런데 인간의 인지 과정에서 발견되는 대부분의 규칙성은 바로 이런 부드러운 제약들이다. 고전적인 AI는 이 같은 제약을 구현하는 데 상당한 어려움을 겪었다. 그러나 스몰렌스키가 제안하는 동역학적 해석을 통해 접근할 경우, 어트랙터들 사이의 상대적 위치 관계가 처리 결과에서 나타나는 규칙성의 측면을 설명해주게 되며, 그 과정에 개입되는 확률적 요동이 처리의 불규칙적 측면을 설명해주게 된다.

그렇다면 이제는 그와 같은 스몰렌스키 식의 연결주의 해석이 본질주의에 대해서 어떤 함축을 갖게 되는지 생각하여 보자. 본질주의적 관점에서 보자면, 개념을 지시하거나 명제를 믿고 바라는 등의 지향적 심리 상태들에는 그것들의 분류 기준이 되는 기능적/인과적 공통 속성이 존재해야만 한다. 그러나 스몰렌스키 식의 해석 하에서 보자면, 동일하게 분류되는 지향적 심리 상태라 해도 그 안에 기능적/인과적 속성의 다양한 변양들(variation)이 존재할 수 있게 된다. 이 같은 변양들은 한 어트랙터로 이끌리게 되는 상태 공간 상의 영역, 다시 말해 분수계(basin)의 존재에서 비롯된다. 즉 개념이나 명제를 상태 공간 상의 어트랙터로 이해할 경우, 그런 어트랙터로 이끌리게 되는 비교적 광범위한 상태 공간 상의 분수계는 개략적으로 동일한 개념이나 명제에 관련된 인지적 상태로 분류될 수 있을 것이다. 하지만 엄밀하게 말해, 그 분수계 안의 모든 벡터들이 동일한 인과적/기능적 특성을 갖게 되지는 않는다. 요동을 동반한 상 전이 과정에서 각 지점의 인과적/기능적 속성이 의미있게 차이날 가능성이 다분하기 때문이다. 이렇게 보면, 스몰렌스키의 입장은 동일하게 분류되는 심리 상태가 반드시 동일한 인과적/기능적 속성을 가질 필요가 없다고 보는 점에서 본질주의에 반대된다고 보아야 한다. 그러나 그처럼

본질주의를 거부하면서도, 스몰렌스키의 해석은 상식 심리학의 틀 자체를 포기하지 않는다. 이 점은 그가 굳이 시스템 상태 공간의 어트랙터들을 '개념'이나 '명제'로 해석하고자 하는 데에서도 잘 드러난다. 스몰렌스키는 '개념'이나 '명제'의 존재를 포기하기보다는, 오히려 그런 개념이나 명제의 분류를 위한 경계들이 엄격하지 않고 가족 유사적 유연성을 지닌다는 사실을 받아들인다. 그리고 그같이 유연한 경계를 지닌 인지적 상태들이 어떻게 물리적으로 구현될 수 있는지를 동역학적 시스템을 통해 보여주려 하고 있다. 결국 그는 본질주의를 무조건 거부하기만 하는 것이 아니라 한편으로 연결주의 인지 모형과 본질주의와의 조화를 모색하고 있기도 한 것이다.

그러면 스몰렌스키가 제리 포더와의 논쟁 중에 예로 들었던 「커피」 표상의 경우를 통해 그의 방식이 어떻게 가족 유사성의 문제에 대처하는지 고찰하는 것으로써 이 글을 마무리지어 보자.

포더는 인지 체계라면 반드시 가져야 하는 속성으로 구성적 구조를 들고 연결주의 모형은 구성적 구조를 갖지 못하기에 올바른 인지 모형이 될 수 없다는 요지의 비판을 제기한 적이 있다.(그와 필리신이 쓴 1981년 논문"Connectionism and Cognitive Architecture" in *cognition* 28 참조) 여기서 구성적 구조란 복합 개념이나 명제가 그것을 구성하는 하위 요소들로부터 복합적으로 형성될 수 있는 일종의 빌딩 블록식 구조를 말한다. 포더는 인공 신경망의 유닛 하나 하나가 개념들로 해석되어야 한다고 보았기 때문에, 개념들 사이에 존재하는 부분/전체 관계가 신경망을 통해서는 구현되기 어렵다고 보았던 것이다.

이에 대해 스몰렌스키는 포더의 그같은 연결주의 해석이 매우 편협한 것이며, 연결주의 모형을 동역학적 시스템이라는 틀 아래서 해석할 경우 그런 문제는 사라진다고 대답하고 있다. 다시 말해 신경망의 각 유닛은 개념들로 해석되어서는 안되고 오히려 신경망의 다수 유닛들이 집합적으로 하나의 개념을 표상하게 된다는 것이다. 이런 표상 방식은 흔히 분산 표상이라고 부르는 방식이다. 스몰렌스키는 「커피」라는 개념을 예로 들며, 미시 특성들을 통해 분산 표상된 개념이 어떻게 구성적 구조를 가

질 수 있는지 설명한다. (이하의 논의는 Smolensky의 1991년 논문 "Connectionism Constituency, and the Language of Thought"에 보다 자세하게 설명되어 있음)

분산 표상을 통해 구성적 구조가 구현된다면, 「커피가 든 컵(cup with coffee)」의 표상이 「컵」 표상과 「커피」 표상의 복합을 통해 형성될 수 있어야 한다. 그래서 스몰렌스키는 「컵」과 「커피」를 다음과 같은 미시 특성들의 조합을 통해 분산 표상하여 보자고 한다.

미시 특성들	「커피가 든 컵」	「커피가 들지않은 컵」	「커피」
똑바로 선 용기	O	O	X
뜨거운 액체	O	X	O
나무에 닿는 유리잔	X	X	X
자기 그릇으로 된 굽은 표면	O	O	X
불에 탄 듯한 향기	O	X	O
자기에 닿는 밤색 액체	O	X	O
은으로 된 긴 네모꼴의 물체	X	X	X
손가락 모양의 손잡이	O	O	X
옆면과 바닥이 굽은 밤색 액체	O	X	O

이 도표에서 알 수 있는 것처럼, 「커피가 든 컵」의 표상은 「커피가 들어 있지 않은 컵」의 표상과 「커피」의 표상을 통해 복합적으로 형성될 수 있다. 반대로 「커피가 든 컵」의 표상에서 「커피가 들어 있지 않은 컵」의 표상을 제거하면 「커피」의 표상이 남는다. 분산 표상을 통해 구성적 구조가 구현되는 것이다.

그러나 이처럼 분산 표상을 통해 구현되는 구성적 구조에는 한가지 특성이 내재되어 있다. 그것은 복합 표상의 구성 요소를 이루는 단순 표상들이 맥락에 따라 서로 다른 미시 특성들을 가진다는 점이다. 이를테면 위에서 본 「커피」의 표상은 「커피가 들어 있는 컵」의 표상에서 「커피가 들어있지 않은 컵」의 표상을 뺀 나머지로 얻어진 「커피」 표상이다. 그 결과, 거기에는 '자기에 닿는 밤색 액체'나 '옆면과 바닥이 굽은 밤색 액체'라는 미시 특성들이 포함되게 된다. 하지만 커피가 반드시 컵에만 들어

있을 필요는 없다. 때로는 박스에 담긴 커피가 있을 수도 있고, 때로는 캔에 든 커피도 있을 수 있다. 이런 경우, 「커피」의 표상은 「커피가 든 박스」의 표상에서 「커피가 들어 있지 않은 박스」의 표상을 제거함으로써, 또는 「커피가 든 캔」의 표상에서 「커피가 들어 있지 않은 캔」의 표상을 제거함으로써 얻어질 수 있을 것이다. 그러나 이같이 해서 얻어진 「커피」 표상은 앞서 컵에 든 「커피」의 표상과 그 미시 특성에 있어 꼭 일치 하지는 않을 것이다. 이를테면 박스에 담긴 「커피」의 표상은 '옆면과 바닥이 굽은 밤색 액체'라는 미시 특성을 갖지 않을 것이고 캔에 담긴 「커피」의 표상은 '자기에 닿는 밤색 액체'라는 미시 특성을 갖지 않을 것이다. 결국 다양한 맥락 아래서 얻어진 「커피」의 표상들은 서로 유사하긴 하되 엄밀한 공통점을 뽑아낼 수 없는 가족 유사적 집합을 이루게 된다.

동일한 개념을 표상하는 미시 특성들의 조합에 이처럼 불일치가 있다는 사실은 본질주의의 관점에서 보자면 심각한 문제일 것이다. 그러나 스몰렌스키의 해석 아래서는 그것이 그리 큰 문제가 되지 않는다. 컵에 든 「커피」의 표상과 캔에 든 「커피」의 표상 그리고 박스에 든 「커피」의 표상이 모두 상태 공간 상의 동일한 분수계 안에 포함되어있다고 간주하면 되기 때문이다. 그럴 경우, 그것들의 인과적 효력은 비록 확률 함수의 요동 때문에 달라질 가능성이 있기는 하지만, 그래도 상당 정도 유사할 것이다. 즉 그 「커피」 표상들은 서로 다른 미시 특성들을 갖는 가운데서도 매우 유사한 인과적 효력을 갖게 되는 것이다.

이러한 사실은 이 글 앞부분에서 문제로 제기되었던 가족 유사성 문제에 대해 새로운 돌파구를 열어주는 듯 보인다. 인지 과학에서 가족 유사성이 문제되는 것은 인지 처리의 분석 단위가 되는 심리 상태들이 그 기능적 속성에 있어서 상당 정도 유사하기는 하되 반드시 동일하지만은 않은 애매한 분류 기준을 가지고 있기 때문이다. 그렇기 때문에, 만일 심리 상태의 분류 기준을 해당 심리 상태가 관련을 맺고 있는 개념이나 명제의 의미론적 정의(definition)를 통해 영속적으로 확정하려 할 경우, 퍼트남이 제기하였던 의미 총체론의 많은 난제들에 둘러싸이게 된다. 그러나 반대로 심리 상태의 분류 기준이 유동적이라는 이유로 그것의 분류

자체를 포기할 경우, 인지 처리의 모의를 위한 초보적인 분석조차 불가능하게 되고 만다. 이같은 딜렘마 상황에서 스몰렌스키가 제시하는 연결주의의 새로운 해석은 전통적인 상식 심리학의 분류 방식 자체를 포기하지는 않으면서도, 심리 상태의 분류에 가족 유사성을 반영하는 유연한 기준을 도입함으로써 양 측의 어려움을 극복하고 있는 것으로 보인다.