인지과학과 기독 유사성

金 泳 楨 (서울대)

I. 기능주의적 인지 모형으로서의 계산주의와 연결주의

인지과학의 철학적 토대를 이루어온 기능주의는 크게 존재론과 방법론의 두 측면에서 규정될 수 있을 것이다. 기능주의의 존재론적 입론은 심리 상태란 다름 아닌 입력 상태와 출력 상태를 일정하게 매개시켜 주는 기능 상태라는 것이고, 방법론적 입론은 기능 상태란 바로 컴퓨터의 알고리즘과 같이 규칙에 의거한 기호 조작을 통해 입력을 출력으로 변환시키는 계산 상태라는 것이다. 우리는 이러한 존재론적 입론과 방법론적 입론을 동시에 만족시키는 인지의 모형이 바로 계산주의라는 것을 익히 알고 있다.[1] 기능주의는 역사적으로 행동주의와 물리주의의 약점을 극복하기 위하여 출발된 것으로, 기능주의는 행동주의와는 달리 심리 현상의 실재를 인정하고 이 심리 현상의 인과적 영향력을 인정한다. 또 물리주의와는 달리, 심리 현상이 순수히 물리적 현상일 뿐이라는 것을 부정하고, 이에 따라 심리 현상이 물리 현상으로 환원될 수 있음을 부정한다. 이러한 물리주의 비판의 핵심은 심리 현상의 복수 실현 가능성으로, 기능주의자들은 자신들이 심리 상태와 동일시하는 기능적 상태가 무수히 많은 다양한 물리적 체계들에 의해 실현될 수 있다고 주장한다. 여기서 인간과 컴퓨터가 동시에 같은 심리 상태에 처해 있을 수 있는 가능성이 생겨난다. 비록 인간과 컴퓨터는 서로 판이하게 다른 물리적 체계들이지만 (인간을 구성하고 있는 물질과 컴퓨터를 구성하고 있는 물질은 판이하게 다르다), 그것들은 동일한 계산적 기능, 즉 동일한 정보 처리 기능을 수행할 수 있다.

그렇다면 연결주의는 기능주의와 어떤 관계에 있는 것일까? 인지에 대한 고전적 모형인 계산주의는 규칙에 따른 기호 조작 모형임에 반해, 새로운 모형인 연결

1) 졸고 "기능주의의 여러 유형과 감각질 문제"(『인지과학』, 민음사, 1989, 47-71쪽) 참조.

주의는 병렬 분산 처리 모형으로 프로그램을 직접 짜 넣어 주지 않고 그 대신 학습 규칙들을 제공하여 샘플들로부터 문제 풀이를 배우게 한다는 데 그 특징이 있다.[2] 신경망에서의 학습이란 샘플들로부터 얻어진 투입과 산출에 대한 한정된 경험을 (신경망 노드들의 활성화 값과 연결 강도를 조절하여) 일반화함으로써 차후에 필요에 따라 적절한 기능을 수행할 수 있는 상태를 형성하는 것이라 할 수 있다. 신경망의 또 다른 특징으로 연상 기억 능력을 들 수 있다. 일반적으로 고전적 정보 저장의 매체는 대부분 어떠어떠한 내용의 정보가 어느 곳(번지 수:address)에 저장되어 있다는 것을 알고 그것을 찾아내게 되어 있다. 그러나 우리가 어떤 배우의 사진을 보고 그가 출현했던 영화의 한 장면을 떠올린다던가, 시험에 임해서 전날 공부했던 내용 중 문제와 관련된 지식을 기억해 내는 방법은 주소라고 하는 저장 장소를 알고서 찾아내는 것이 아니다. 내용의 일부 또는 관련된 암시를 가지고 전체를 찾아내는 이른바 CAM(Content Addressable Memory) 방식이다. 이와 같은 기억 장치는 그 동작 원리로 연상 작용을 수반할 수 있다. 즉 기억되어 있는 데이터와 똑같지는 않지만 유사한 입력 정보가 들어오더라도 그것을 확률적으로 근사화하여 용도에 적합하게 사용한다. 이러한 기능이 신경망 기억 방식의 특징이라 할 수 있다.

Cumins와 Schwarz는 그의 논문 "Connectionism, Computation, and Cognition"[3]에서 인간 지식의 대부분은 계산 가능한 함수에 의해 규정되지 않는다고 주장하면서 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 알고리즘을 명시할 필요가 없다. 연결주의 체계가 요구하는 것은 문제 영역을 반영하는 표상 상태들을 야기하는 동역학이다. 연결주의 표상 상태들은 동역학적 체계 속의 상태일 수

2) 계산주의와 연결주의의 관계에 대한 철학적 논의는 『연결주의와 구성적 구조: 포더와 그의 동조자들에게 보내는 응답』(최훈, 1994년 서울대학교 철학과 석사 학위 논문)에 상세히 개진되어 있다.

3) in Horgan, T. and J. Tiensen (eds.) (1991) *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academy Publishers. 인용문은 이 논문집 서론에서 Cumins와 Schwarz의 논문을 소개한 부분을 따온 것임.

있다. 그리고 동역학 체계의 특징 함수(이 함수는 동역학 방정식에 의해 정의된다)는 그것 자체로 계산 가능하지 않다.

위의 논의들을 살펴보면 계산주의와 연결주의는 그 방법론에서 상당한 차이를 보이고 있음을 알고 있다. 특히 연결주의는 계산주의가 그 근간으로 삼고 있는 규칙에 의거한 기호 조작을 정보 처리 방식으로 채택하고 있지 않음을 알 수 있다. 그렇지만 연결주의자들도 계산주의자들과 마찬가지로 인간 인지의 기능적 시뮬레이션을 통해 인지를 구현하려 하고 있는 한, 연결주의도 존재론적으로 기능주의의 일종이라고 말할 수 있을 것이다. 인지에 대한 고전적 모형인 계산주의는 규칙에 따른 기호 조작을 통해 어떤 기능을 구현하는 모형임에 반해, 새로운 모형인 연결주의는 신경망의 학습(신경망 노드들의 활성화 값과 연결 강도 조절)을 통해 어떤 기능을 구현하는 모형이라고 말할 수 있을 것이다.[4] 동일한 기능의 복수 실현 가능성이 기능주의의 한 핵심적 입론으로 받아들여지는 한, 그러한 기능의 복수 실현이 계산 기능의 영역에만 국한되고 신경망 기능의 영역에로의 확장은 선험적으로 배제되어야 할 아무런 이유도 없는 것이다. 계산주의와 연결주의의 차별성을 강조하는 위의 Cumins와 Schwarz의 주장 속에서도 연결주의가 입력과 출력을 연결시키는 기능 즉 함수 기능을 문제 삼고 있다는 것을 부인하지 않고 있다 (부인하는 것은 단지 그것이 계산 가능한 함수 기능이어야 한다는 것일 뿐이다).

결국 필자는 연결주의도 계산주의와 마찬가지로 존재론적으로 기능주의적 인지 모형의 하나로 간주되는 한, 인지의 적절한 모형이 되기 위해서는 가족 유사성이라는 커다란 장벽을 넘어야 한다는 것을 보일 것이다. 그러기 위해 둘째 절에서

---

4) 연결주의가 기능주의의 일종이라는 것은 다음과 같은 구절에서도 명료하게 나타나 있다 (『알기 쉬운 신경망 컴퓨터』 제3장 신경망이란 무엇인가?(이종호, 1992, 전자신문사) 71쪽):

> 광범위한 관점에서 본다면 컴퓨터에 의한 문제의 풀이 방법은 입력 도메인에서의 데이타와 출력 도메인의 데이타 간의 사상(mapping) 과정이라 할 수 있다. 신경회로망에 의한 패턴 매칭도 결국 사상에 속한다고 할 수 있다. 다만 간단한 사상의 원칙을 찾기 어려운 경우 즉 비선형성 또는 비모델화 동적 특성(Unmodeled dynamics, 실제 시스템의 모델화 과정에서 고려하지 못한 동적 특성 부분)이 포함되어 있는 문제의 풀이에는 신경회로망에 의한 경험적인 근사적 사상이 효과적이다.

는 가족 유사성 개념을 이용한 힐러리 퍼트남의 계산적 기능주의에 대한 비판을 살펴본 후, 그의 계산적 기능주의 비판이 그대로 연결주의에도 역시 적용될 수 있음을 논변하고, 셋째 절에서는 허버트 드레이퍼스의 연결주의 비판을 살펴볼 것이다. 그리고 결론에서는 드레이퍼스의 연결주의 비판이 가족 유사성을 토대로한 퍼트남의 기능주의 비판과 그 맥을 같이하고 있음을 논변하고, 또 계산주의든 연결주의든 그것이 기능주의적 인지 모형의 하나인 한, 인지의 적절한 모형이 되기 위해서는 왜 가족 유사성이라는 커다란 장벽을 넘어야만 하는지를 포괄적으로 재음미하여 볼 것이다.

II. 퍼트남의 기능주의 공략

퍼트남은 그의 책 『표상과 실재』[5]에서 "기능주의"로 불리는 컴퓨터 유비가 "정신 상태의 본성은 무엇인가?"하는 질문에 대답하지 못한다는 것을 논변하고 있다. 다시 말해, 정신 상태란 컴퓨터의 알고리즘과 같은 어떤 것－즉, 마음이나 두뇌 속에 있는 과학적으로 단일하게 기술 가능한 원초적 존재자－일 수 없다는 것을 논변하고 있다. 이와 같은 퍼트남의 논지는 그가 부정하는 입장들을 살펴보면 더욱 분명하게 드러난다. 퍼트남이 부정하고자 하는 입장들은 크게 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 전형적 예인 진리가 제거될 수 없는 한) 신화적인 것으로 제거될 수 없다.⇒제거주의 비판.

여기서 중요한 것은 위와 같은 세 입장의 비판이 기능주의의 비판과 어떤 방식으로 연결되어 있는지를 이해하는 것이다. 사실상, 기능주의란 환원주의의 한 형

5) Putnam, H., (1988) *Representation and Reality*, MIT Press. 이 책은 필자에 의해 한글로 번역되어 있음: 『표상과 실재』 (1992, 이화 문고 65, 이화여자대학교 출판부).

태(기능적 환원주의)이므로, 기능주의의 논박은 환원주의의 논박만으로도 충분하며, 그 이상의 비판(본질주의나 제거주의 비판)을 필요로 하지 않는다. 정신 상태가 계산적 상태로 환원될 수 없다고 말하는 것은 곧바로 정신 상태가 컴퓨터 알고리즘과 같은 어떤 것일 수 없다는 것을 뜻하기 때문이다. 그렇지만 퍼트남은 환원주의 비판을 본질주의 비판을 토대로 하여 수행하고 있어, 그의 환원주의 논박은 본질주의 논박에 의존적이다. 구체적으로 말해, 컴퓨터의 알고리즘과 같은 것은 과학적으로 단일하게 (혹은, 동치적으로) 기술 가능한 속성을 지닌 존재자이므로, 정신 현상이 과학적으로 단일하게 (혹은, 동치적으로) 기술 가능한 원초적 현상이 아니라는 본질주의 비판은 바로 정신 현상이 컴퓨터의 알고리즘과 같은 것으로 환원되지 않는다는 기능적 환원주의 비판에로 연결된다. 따라서 본질주의 비판은 퍼트남의 기능주의 비판에서 결코 생략될 수 없는 중심적인 역할을 수행하고 있다. 퍼트남의 구절을 인용하여 보자:

> 필자는 어떤 특정한 지향적 현상의 모든 경우들이 공통적으로 갖는 과학적으로 기술 가능한 속성이 없음을 보일 것이다. 이 입론으로써, 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"과 같은 어떤 한 유형의 특정한 지향적 현상의 모든 경우들이 공통적으로 가지고 있는 어떤 과학적으로 기술 가능한 속성이 있다는 것을 부인하려 한다.(제1장 도입부)

실제로 퍼트남은 기능주의 비판을 이 본질주의 비판으로 마무리 지을 수도 있었을 것이다. 그러나 퍼트남이 제거주의 비판을 덧붙인 이유는 만일 그가 주장하려는 것이 환원주의와 본질주의 비판뿐이라면, 그의 입장은 제거주의의 입장과 본질적으로 구별될 수 없기 때문이다.

퍼트남의 본질주의 비판에는 그 논거로서 비트겐슈타인의 가족 유사성 개념이 중요하게 등장하고 있다.

> [지향성과의] 보다 나은 비교는 비트겐슈타인에 의해 제시된 바 있는 용어 "게임"과의 비교이다. 일상언어의 수준에서조차도 모든 게임들이 "공통적인 어떤 것" 즉 게임임을 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 "지시했다"고 우리가 말하곤 하는 모든 경우들을 (혹은 어떤 사람이 하나의 특정한 사물을 "지시했다"고 우리가 말하곤 하는 모든 경우에서조차도 이 경우들을) 자세히 검토할 때, 우리는 단어와 지시된 사물간에 단일한 어떠한 관계도 발견하지 않는다.(제1장 도입부)

퍼트남의 기능주의 비판에 있어서 핵심적인 역할을 담당하고 있는 비트겐슈타인의 가족 유사성 논변은 그의 『표상과 실재』 책에서 "믿음의 차이 삭감 논변" (혹은 넓게는 "총체론 논변")이란 이름 하에 보다 일반화된 형태로 등장한다. 총체론 논변을 구체적으로 살펴보기 전에, 가족 유사성과 복수 실현 가능성의 관계에 대해 알아보자.

5장과 6장은 앞내용의 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 "프로그램들"을 가지면서 동일한 정신 상태 속에 있을 수 있기 때문에 정신 상태들은 문자 그대로 "프로그램들"일 수 없다.(서론)

고양이 한 마리가 매트 위에 있다고 믿고 있을 때 사람들이 처해 있을 수 있는 여러 다른 물리적 상태들이 물리적/화학적 용어들로 명시될 수 있는 "공통적인" 어떤 무엇을 가질 필요가 없다는 점을 지적한 것이 바로 기능주의의 통찰이었다. 이러한 통찰과 꼭 마찬가지로 여기에서의 우리의 논의의 결과는 다음과 같다: 고양이 한 마리가 매트 위에 있다고 믿고 있을 때, 사람들이 처해 있을 수 있는 여러 다른 계산적 상태들이 계산적 용어들로 명시될 수 있는 "공통적인" 어떤 무엇을 가질 필요가 없다.(제5장 도입부)

기능주의자들이 물리주의자들에게 가한 비판의 핵심은 복수 실현 가능성(multiple realizability)이었다. 이제 퍼트남이 기능주의자들에게 가하고 있는 비판의 핵심도 역시 복수 실현 가능성과 같은 어떤 것이라는 것이 위의 인용문들의 핵심이다. 물론 복수 실현 가능성 개념은 가족 유사성 개념과 같은 것이 아니다.

왜냐하면, 복수 실현 가능성 개념은 동일한 것이 여러 형태로 실현될 수 있다는 것을 의미하나, 가족 유사성 개념은 우리가 동일하게 분류하는 것이 실제로는 동일한 것이 아니라 단지 가족 유사성만이 있다는 것을 의미하기 때문이다. 보다 구체적으로 말해, 기능적으로 동일한 것이 물리/화학적으로 다양하게 복수적으로 실현될 수 있다고 주장하는 것은 아직 본질주의적 입장을 버린 것이 아니나, 가족 유사성(즉, 기능적으로 동일한 것으로 간주된 것이 실제로는 가족 유사성밖에 없다는 것)을 주장하는 것은 본질주의적 입장을 비판하고 나선 것이다. 그러나 물리주의가 논박되는 이유가 과학적으로 기술 가능한 공통된 물리적 속성이 없다는 것이듯이, 기능주의가 논박되는 이유도 과학적으로 기술 가능한 공통된 기능적 속성이 없다는 점에서 이 둘은 유사성을 갖는다.

퍼트남의 비판이 단순히 비트겐슈타인의 가족 유사성 개념을 도입하여 모든 유형의 기능주의 일반을 일거에 공격하고 있는 것은 아니다. 그는 여러 유형의 기능주의를 나누어 공격하고 있으며, 그의 비판 논변도 여러 가지가 존재한다.[6] 그러나 가장 핵심적이고 또 모든 유형의 기능주의에 일반적으로 적용될 수 있는 논변은 가족 유사성 개념을 토대로한 총체론 논변이다.

총체론 논변은 좁은 의미의 총체론 논변과 믿음의 차이 삭감 논변으로 구성되어 있다. 좁은 의미의 총체론 논변은 크게 두 주장으로 요약될 수 있을 것이다: 용어들의 정의 가능성 거부와 믿음들의 전체 그물 조직 구조. 그리고 이 둘이 밀접하게 연결되어 있다는 것이 총체론 논변의 핵심이다.

용어들의 정의 가능성 거부에 관하여: "총체론이 곧바로 제시하는 것은 대부분의 용어들이 정의될 수 없거나, 혹은 적어도 만일 "정의"로써 우리가 뜻하는 바가 단 한 번만에 고정되는 어떤 것 즉 절대적으로 용어의 의미를 규정하는 어떤 것이라 한다면, 대부분의 용어들이 정의될 수 없다는 것이다."(제1장 2절 1)

---

6) 그의 『표상과 실재』 책에서는 크게 네 종류의 논변들이 기능주의 비판 임무를 수행하기 위해 도입되고 있으며, 또 이 논변들을 이용하여 공략하고 있는 기능주의도 크게 네 유형으로 나누어 볼 수 있다. 첫째, 믿음의 차이 삭감 논변과 총체론 논변 (이 두 논변을 합해, 간단히 총체론 논변); 둘째, 언어적 노동의 분업 논변과 환경의 기여 논변 (이 두 논변을 합해, 외부적 논변); 셋째, 괴델적 논변; 넷째, 행동주의 논변과 무제한적 현실화 논변(이 두 논변을 합해, 간단히 현실화 논변)이 그것이다. 그리고 이 책에서 공략하고 있는 기능주의의 네 유형이란, 첫째, 생득 가설을 받아들이고 있는 MIT 정신주의; 둘째, 생득 가설을 전제하지 않는 정신주의적 기능주의; 셋째, 계산적 사회기능주의; 넷째, 루이스류의 상식적 기능주의가 그것이다.

믿음들의 전체 그물 조직 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 전체 그물 조직으로서 그런 것이다."(제1장 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 수정될 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 것이 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 "정의들"을 조사함으로써 결정되어질 수 없다."(제1장 2절 1)

그러나 용어들의 정의 가능성 거부에 대해 우리가 신중하게 고려할 만한 가치가 있는 한 반론이 제기된다. 그것은 만일 정의를 수정하여야 할 경우들이 있다는 것이 용어들의 정의 가능성 거부에 대한 논거라면, 우리는 정의의 수정을 정의되는 용어의 의미(그리고, 지시체)가 바뀌는 것으로 간주하면 된다는 것이다. 즉, (정의항의 표현이 바뀜으로써) 정의항의 의미가 바뀌었을 뿐만 아니라, (비록 피정의항의 표현은 바뀌지 않았지만) 피정의항의 의미(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 옛 정의를 수정한 것이라고 말할 필요가 없게 된다는 것이다. 이것에 관한 구절을 인용하여 보면:

> 또 다른 전통적인 움직임은 "그래, 과학자들은 '운동량'의 의미를 바꾸기로 결정하였던거야"라고 말하는 것이다. 만일 이것이 상대성 이론의 채택 후 과학자들이 "운동량은 질량 곱하기 속도이다"라는 문장에 부여한 진리치가 바뀌었다는 것을 해명한다면, 그것은 우리가 지금 다른 물리적 크기에 관해 얘기하고 있다는 것을 함축한다. 그러나 그렇지 않다. 우리는 여전히 탄성 충돌 속에서 보존되는 크기인 동일한 친숙한 옛 운동량에 관해 얘기하고 있는

> 것이다. 만일 "운동량"이 어떤 것을 지시한다면, 그 친숙한 운동량이 항상 지시되는 물리적 크기인 "운동량"이다. 그리고 운동량 자체인 그 크기가 질량 곱하기 속도와 똑같은 양이 아니라고 판명되었던 것이다.(제1장 2절 1)

여기서 믿음의 차이 삭감 논변이 도입된다: 의미와 연계된 약간의 믿음의 차이들은 무시하며, 그리하여 의미는 동일성을 유지하는 것으로 간주한다. "의미들은 본질을 가지고 있지는 않지만 시간 변화 속에서도 동일성을 유지하고 있다."(제1장 2절 1) 믿음의 차이들을 삭감함은 관용의 원리에 의거한다. 우리는 가급적 단어들을 말하는 사람이 지니게 될 참인 믿음들의 수효가 극대화되도록 해석하는 관용을 베풀어야 한다는 관용의 원리의 요구에 따라서, 해석시 믿음에 있어서의 약간의 차이는 삭감하여 의미가 동일성을 유지하는 것으로 간주한다. 가족 유사성 개념을 사용하여 다시 표현하자면, 비록 의미들은 본질을 가지고 있지 않고 가족 유사성밖에 없지만, 의미와 연계된 믿음에 있어서의 약간의 차이들은 삭감하여 의미들이 동일성을 유지하는 것으로 간주한다. 이것이 해석상의 관용의 원리가 요구하는 바이다. 퍼트남은 믿음의 차이 삭감 논변을 다음과 같이 "운동량", "전자", "식물"과 같은 단어들을 예로 들어가면서 설명하고 있다:

> 동의성(synonymy)에 대한 이론은 해석에 관한 질문들에 대답하는 이론일 것이다. ---- "운동량(momentum)"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 전자에 대한 이론들과 기술들을 비교함으로써 그리고 그것들이 매우 같다는 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"의 이야기로 말하고 있는 것이 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 동의어적으로 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 보어

의 1934년 이론을 그의 1900년 이론의 진정한 계승 이론으로 취급하는 것은 실제로 동일한 결정이다; 즉 용어의 의미 혹은 지시대상에 관한 결정으로서 한때 기술된 결정과 연구 프로그램들의 친밀한 관계들에 관한 결정으로서 한 때 기술된 결정은 실제로 동일한 결정이다.

이러한 결정은 해석에 있어서 "관용(寬容)" 혹은 "의심점의 선의 해석(benefit of the doubt)"이라 불려온 바를 보여준다. 1900년의 보어가 우리가 "전자들"이라고 부르는 것을 지시한다고 해석할 때, 우리는 그럼으로써 그의 1900년 믿음들 중 적어도 약간을 우리들의 견지에서 "참"이 되도록 만들고 있는 것이다. 이에 반해 그가 존재하지 않는 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 논문들에서 단어 "전자"를 계속해서 사용한 이유이다.)

모든 해석은 관용(寬容)에 의존한다. 왜냐하면 우리는 해석할 때 적어도 믿음에 있어서의 약간의 차이는 항상 삭감(discount)해야 하기 때문이다. 예를 들어, 우리가 2백년 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 것들은 식물이 어떤 것이라는 우리의 현재 개념에 중심적인 것이다. 이러한 모든 것들은 2백년 전에는 알려지지 않았었다. 그렇지만 (만일 우리가 철학자들이거나 철학적인 과학사가들이 아니라면) 우리는 2백년 전 사람들이 "다른 세계에 살았다"고 말하지 않을 것이며 그들의 개념들이 우리가 지금 가지고 있는 개념들과 "불가공약적"이라고 말하지 않을 것이다. 불가공약적이라는 것을 문자 그대로 받아들이면 (물론 결코 그럴 수 없지만), 그것은 우리가 2백년 전에 쓰여진 일상 편지를 해석할 수 없음을 함축한다. 간략히 말해, 우리는 식물이라는 개념을 본질은 없지만 시간의 변화 속에서도 동일성을 유지하는 것으로 간주하며, 전자라는 개념도 본질은 없지만 시간 변화 속에서 동일성을 유지하는 것으로 간주한다.(제1장 2절 2)

그러나 믿음의 차이 삭감 논변은 의미가 의미와 연계된 믿음들의 차이에도 불구하고 동일성을 유지한다는 것을 확립시켜 주는 논변으로서만 끝나지 않는다. 그 논변은 그 이상의 함축을 지니고 있다. 믿음 고정의 과정들이 일반 지능 혹은 이해의 전 능력을 필요로 한다는 것을 보여주며, 이에 따라 의미와 지시의 동일성

식별에서의 일반 지능의 역할의 중요성을 부각시켜 주고 있다. 그리고 믿음 고정의 과정들이 일반 지능 혹은 이해의 전 능력을 필요로 한다는 것은 바로 믿음들의 전체 그물 조직 구조 내지 믿음 고정의 총체론적 성격에 대한 또 다른 진술 방식인 것이다. 퍼트남의 말을 직접 들어보자:

> 그렇지만 "해석상의 관용"이 항상 통용되는 것은 아니다. 우리는 항상 단어들을 (우리의 견지에서) 말하는 사람이 가지게 될 참인 믿음들의 수효가 극대화되도록 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 합리적인 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 독립된, 의미와 지시의 동일성 이론을 세울 희망은 없다.
>
> ---- 앞의 예들이 보여주듯이 동의성과 일치하는 계산적 관계는 포더의 의미에서의 "단위적"일 수 없다; 즉 "일반 지능"보다 심리학적으로 더 기초적일 수 없다.(제1장 2절 2)

지금까지 (넓은 의미의) 총체론 논변이 뜻하는 바를 살펴보았다. 그 중 특히 부각된 점들은 **용어들의 정의 가능성 거부, 믿음들의 전체 그물 조직 구조, 의미들이 믿음들의 차이에도 불구하고 동일성을 유지하고 있음, 의미와 지시의 동일성 식별에서의 일반 지능(이해의 전 능력) 역할의 중요성** 등이다. 이와 아울러 총체론 논변은 의미나 지시에 관한 문제들이 믿음 고정의 문제들과 밀접하게 연관되어 있음을 보여준다.

이제 이러한 계산적 기능주의에 대한 퍼트남의 비판이 어떻게 연결주의에도 적용될 수 있는지 살펴보자. 앞서 지적한 대로, 퍼트남은 가족 유사성에 토대를 둔 총체론 논변을 기반으로 심리 현상의 본질주의를 공략하고 있다. 아마도 연결주의를 옹호하는 사람은 연결주의의 강점이 바로 본질주의를 거부하고 가족 유사성을 수용하고 있다는 점에 있으므로, 퍼트남의 논변은 계산주의를 비판하고 연결주의를 옹호하는 논변으로 보아야 한다고 주장할 수도 있을 것이다. 1절에서 살펴보았듯이, 연결주의는 내용의 일부 또는 관련된 암시를 가지고 전체를 찾아내는 이른

바 CAM 방식이므로, 미리 학습된 패턴과의 유사성을 토대로 새로운 입력을 범주화할 수 있어 가족 유사성을 구현하고 있다고 주장할 수도 있다.[7] 그러나 퍼트남이 개진하고 있는 총체론 논변은 그러한 수준의 가족 유사성에 머무르는 것이 아니다. 그는 단순히 유사함을 토대로 일반화될 수 있다는 것만을 주장하는 것이 아니라, 어떤 유사함이 적절한 유사함이고 어떤 유사함이 적절하지 못한 유사함인가를 결정하는 데는 일반 지능 혹은 이해의 전 능력이 총체적으로 요구된다는 것을 핵심적으로 주장하고 있는 것이다. "식물", "전자", "운동량" 같은 단어들의 의미와 지시체를 결정함에 있어, 이것들에 관한 어떤 믿음의 차이는 사소한 것으로 간주하여 삭감하며 어떤 믿음의 차이는 진정한 것으로 간주하여 삭감할 수 없는지를 결정하는 데는 일반 지능 혹은 이해의 전 능력이 요구된다는 것을 주장하고 있는 것이다. 그러한 능력은 컴퓨터 프로그램뿐만 아니라 신경망에 의해서도 구현되기 어렵다는 것이 퍼트남 논변의 함축인 것이다. 허버트 드레이퍼스는 실제로 그러한 취지에서 연결주의가 적절한 인지 모형으로서 한계를 가지고 있음을 논변하고 있다.

III. 드레이퍼스의 연결주의 비판

드레이퍼스는 그의 책 *What Computers Still Can't Do*[8]에서 GOFAI에 대한 대안적 접근으로 세 가지를 꼽고 그 세 가지 모두가 한계를 지니고 있다고 논변한다: 첫번째 대안은 기호 처리 모형에 하이데거적 시각을 도입한 하이데거적

---

7) 다음 논문들 참조:
Smolensky, P.(1987), "The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn" in T. Horgan and J. Tiensen(eds.)(1991), *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academy Publishers: 281-308.
Smolensky, P.(1988), "On the Proper Treatment of Connectionism", in D. J. Cole, J. H. Fetzer, and T. L. Rankin(eds.)(1990), *Philosophy, Mind, and Cognitive Inquiry: Resources for Understanding Mental Processes*, Dordrecht: Kluwer Academy Publishers: 145-206.

8) Dreyfus, Hubert L., *What Computers Still Can't Do* (Cambridge, Mass: MIT Press, 1993). 아래 Dreyfus 인용문들은 모두 이 책의 "Introduction to the MIT Press Edition" 부분에 나와 있는 것임.

인공지능이며, 나머지 두 대안은 연결주의적 대안으로 하나는 감독 학습(supervised learning)이라 불리며, 다른 하나는 강화 학습(reinforcement learning)이라고 불린다. 필자는 이 세 부류의 대안들에 대한 드레이퍼스의 비판을 (대부분 그의 말을 그대로 인용하여) 소개하는 데 이 절을 할애할 것이다. 우선 이 세 대안에 대한 드레이퍼스의 소개부터 들어보자:

> 대안적 접근에 초점을 맞추고 있는 세 그룹의 인공 지능 연구자들에게는 GOFAI는 이미 끝났다. Philip Agre와 David Chapman의 연구와 연계되어 있는 한 접근은 문맥-독립적인 기호 표상들이나 내부적 모형-토대적 계획을 사용하지 않고 미시 세계와 지능적으로 상호작용하는 프로그램들을 생산하도록 시도하고 있다. 신경망 모형가들에 의해 대표되는 두번째 그룹은 표상을 통째로 포기한다.[9] 이 접근은 전통적인 면모들을 사용하나 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 외부로부터의 전문가를 이용하지 않고 숙련 영역에서의 실제적인 실행을 이용하는 프로그램을 발전시키는 것을 목표로 하고 있다. 이 접근들 각각의 장점들과 한계들은 고려할 만한 가치가 있다.

드레이퍼스는 지금까지의 하이데거적 인공지능은 그것이 생략하는 것 – 장기적인 계획 그리고 문맥-독립적인 면모들을 가진 대상들의 내부적 표상들 – 에 있어서는 하이데거의 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 핵심적인 능력을 제공하기 위해, 더욱 많은 연구자들이 신경망 연구에 관심을 돌리고 있으므로 그러한 연결주의적 신경망이 그가 친숙성(familiarity) 혹은 총체적 민감성(global sensitivity)이라고 부른 바를 보여줄 수 있는지, 그리고 보여줄 수 없다면, 연결주의가 어떤 다른 방식으로 적절성과 학습에 대처할 수 있는지 하는 질문을 살펴보아야 한다고 제안한다.

드레이퍼스에 따르면, 상식적 지식의 문제가 GOFAI에서 그랬던 것처럼 연결

9) 많은 연결주의자들이 연결주의가 표상주의의 일종이라는 것을 인정하고 있으므로, 이 주장을 문자 그대로 받아들이기는 어렵다. 아마도 직접적인 인과적 효력을 지니는 고전적 의미의 표상을 포기한다는 점을 이렇게 표현한 듯싶다.

주의에서도 다시 대두되며 연결주의의 진보를 위협한다. 모든 다층 인식자(perceptron) 신경망 모형 설계자들은 지능적인 신경망은 일반화할 수 있는 능력을 갖추어야만 한다는 것에 동의한다; 예를 들어 주어진 분류 작업에 대해, 하나의 특정한 출력과 연계된 입력의 충분한 예들이 주어지면, 그것은 동일한 유형의 새로운 입력들에 동일한 출력을 연계시켜야만 한다. 그러나 이 대목에서 무엇이 어떤 근거에서 동일한 유형으로 간주되는가?하는 질문이 매우 의미있게 제기된다. 신경망의 설계자는 보통 합리적인 일반화에 요구되는 "유형"의 특정한 정의를 염두에 두고 있으며, 만일 그 망이 이 유형의 다른 사례들을 일반화시킨다면 그것을 성공으로 간주한다. 그러나 신경망이 예기치 않은 연상관계를 산출했을 때, 그 망이 일반화에 실패했다고 말할 수 있을까? 사람들은 다음과 같이 말할 수도 있을 것이다: 신경망이 "유형"의 다른 정의에 따라 작용해 왔으며, 그 다른 점이 방금 나타났다.

연결주의적 신경망에서 발생할 수 있는 창조적이나 비지능적인 일반화의 재미있고 극적인 경우로서 드레이퍼스는 다음과 같은 실제 사례를 끌어들인다. 연결주의 연구의 초창기에 미국 육군은 숲속에 있는 탱크를 인식하도록 인공적인 신경망을 훈련시키려고 노력하였다. 육군의 연구자들은 탱크들이 없는 숲 속의 사진들을 많이 찍었다. 그리고는 며칠뒤 탱크들이 숲 속의 나무들로부터 삐죽이 돌출해 있는 사진들을 많이 찍었다. 그리고는 두 부류의 사진들을 식별할 수 있도록 신경망을 훈련시켰다. 결과는 인상적인 것이었다. 그들은 훈련시킬 때 사용하지 않았던 사진들에까지 신경망이 자신의 지식을 일반화시킬 수 있음이 판명되자 더욱더 감동되었다. 그렇지만 신경망이 부분적으로 숨겨진 탱크를 진정으로 인식한 것인지를 확인하기 위해, 연구자들은 동일한 숲에서 더 많은 사진들을 찍고 그 사진들을 훈련된 신경망에 보여주었다. 그들은 나무들 뒤에 탱크가 있는 새로운 사진들과 단지 평범한 나무들만 있는 새로운 사진들간의 차이를 신경망이 식별해내지 못한다는 것을 발견하고 매우 실망하였다. 고심 끝에, 어떤 사람이 탱크들이 없는 숲 속의 원래 사진들은 흐린 날에 찍혔고 탱크들이 있는 원래 사진들은 맑은 날에 찍혔다는 것에 주목했을 때 그 신비는 마침내 해소되었다. 그 신경망은 분명히 그림자들이 있는 숲과 없는 숲 사이의 차이를 인식하고 일반화 시키는 것을 배웠던 것이다! 이 예는 신경망이 우리의 의미의 적절한 일반화를 공유할 수 있으려면, 세계에 대한 우리의 상식적인 이해를 공유해야만 한다는 일반적 논점을 보여주고 있다.

위의 논의에서 시사되었듯이 드레이퍼스의 연결주의 비판의 핵심은 신경망을

통한 일반화도 종국적으로 신경망 외적인 요소인 우리 인간의 이해 능력에 의존하지 않고서는 적절하게 수행되기 어렵다는 것이다. 보다 구체적으로 말해, 어떤 일반화가 문맥에 적절한 일반화인가 하는 것은 우리 인간의 이해 능력에 의존적일 수 밖에 없다는 것이다. 그것은 다시 말해, 적절한 일반화를 효율적으로 수행하는 인간의 이해 능력을 신경망을 통해 성공적으로 시뮬레이션하는 것은 극복되기 어려운 난점들을 안고 있다는 것이다. 구체적으로 그는 신경망 모형 설계와 신경망 학습이라는 두 가지 측면에서 연결주의의 난점들을 지적하고 있다. 신경망 모형 설계에 관련해서는 인간인 설계자의 관점에서 어떤 가능한 일반화들이 미리 제약되기 때문에 신경망이 인간이 하는 것과 같이 문맥에 적절한 방식으로 일반화할 수 있는 포괄적인 이해 능력을 갖출 수 없고, 또 감독 학습과 관련해서는 훈련시 적절한 샘플과 적절하지 못한 샘플을 결정하는 것은 훈련자인 사람이며, 실제로 지능의 핵심적인 부분이 바로 이러한 결정을 하는 능력에 있다는 것이다. 신경망 모형 설계와 관련된 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 허용가능한 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. 그러면 신경망은 적절한 일반화 원리를 배운 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-세계의 상황에서는 인간 지능의 많은 부분이 문맥에 적절한 방식으로 일반화하는 것에 놓여있다. 만일 설계자가 신경망을 적절한 방식으로 미리 정의한 부류에 제한시킨다면, 망은 그 문맥에 대해 설계자에 의해 그 망에 부여

된 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 인간 두뇌와 크기, 설계 구조, 그리고 초기-연결 모습을 공유할 것이다. 진정으로 신경망 연구자들은 간헐적으로 잠정적인 성공을 거두고 있으나 일반화 시키는 원리적 방식을 가지고 있지 못한데, 이것은 내가 1960년대 GOFAI에 대해서 썼을 때의 GOFAI 연구자들의 단계인 것처럼 보인다. 한때 등한시되었다가 다시 부상된 연결주의 접근은 단지 기회를 잃어가고 있는 것처럼 보인다.

인간이 하는 방식으로 일반화하기 위해서는, 망의 설계 구조는 인간에게 적절한 면모들의 견지에서 상황들에 반응할 수 있도록 망이 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 인공 지능에로의 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 연결 강도에 의해 이 지능을 어떻게 포착하느냐 하는 것일 뿐이다. GOFAI 체계와 마찬가지로, 망은 그러므로 그것들이 배운 것이 부적절한 상황을 인식하는 능력을 결여하고 있다: 그 대신 실패를 인식하고 망이 이미 훈련된 상황의 산출을 수정하거나 또는 행동에 있어서 적절한 수정으로 이끌 새로운 샘플들을 제공하는 것은 인간 사용자에 달려있다. --- 우리가 진정으로 필요한 것은 스스로 환경에 어떻게 대처할 것인가를 배우고 환경이 변함에 따라 그들 스스로의 반응들을 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 학습"이라고 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 적절적인 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, 장치는 모든 것을 알고 있는 선생을 필요로 하지 않으며 단지 세계로부터의 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 문제라는 두 문제이다. 그런데 이 두 문제는 그 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 숙련 영역에서 인간 전문가와 같은 훌륭한 장치를 만들 수 있는가? 현재의 실제 작업과 관련하여 두 가지 개선이 적어도 필요한데, 그 어느것도 현

재의 지식에 토대해서는 달성될 수 없는 것처럼 보인다. 첫째, 만일 학습 동안에 실제로 마주친 상황들의 수효보다 훨씬 넘는 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 행동들을 산출하고, 그러한 적절한 면모들의 토대 위에서만 그 상황에 가치를 부여한다. 이러한 방식으로 부적절한 면모들과는 관계없이 동일한 적절한 면모들을 공유하는 모든 상황들에 대한 경험들을 우리는 함께 합친다. 이러한 적절한 면모들을 공유하는 상황들에 대한 경험들의 토대 위에서 행동들이 선택되고 가치들이 학습된다. 이 두 접근 모두 만족스럽지 못하다. 자동 일반화 절차에 관련해서는, 일반화가 요구되는 바로 그 시점에서 상황은 감독 학습에서 직면했던 상황과 동일하다. 누구도 인간의 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 상상하여 보자. 어느날, 9회말 느지막하게 컴퓨터를 조회해본 후, 매니저는 현재 타자 A를 대타 B로 대체하기로 결정한다. 대타는 홈런을 치고 팀은 경기에서 승리한다. 그렇지만 구단주는 화를 내면서 컴퓨터를 잘못 사용하였다고 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 있으며 이것은 낮 경기였다고 말한다. 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 제공함으로써 도움을 줄 수 있다. 그러나 오로지 경험만이 매니저로 하여금 현재 사태를 어떤 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 학습 접근의 한 구체적인 사례로 Chapman과 Kaelbling에 의해 제안된 절차를 고려하고 그 문제점들을 다음과 같이 지적하고 있다:

현재의 절차들은 시행착오적 학습 동안 얻어진 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 취해져야 하고, 모든 상황들에 동일한 가치가 부여되어야 한다고 가정하고 출발한다. 그런 다음 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 값이 행동들과 가치들에 중요하게 의미있는 영향을 미치는 것처럼 보이면, 이것은 적절하다고 선언된다. 상황은 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 수 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 같이 딜레마 상태로 정리하고 있다: "인공지능에서의 모든 연구는 깊은 딜레마

에 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, 인간이 이해하는 바를 충분히 명시적으로 만듦으로써 상식을 보이도록 컴퓨터를 성공적으로 프로그램 한다는 것에 대한 극단적인 비가망성이 GOFAI 연구 프로그램에 대한 회의로 나를 이끌었다. 다행히도, 기계 학습의 최근의 연구는 인간이 이해하는 모든 것을 표상하도록 요구하지 않는다. 그러나 우리가 방금 본 것처럼, 딜레마의 또다른 뿔에 직면한다. 인간이 행하는 방식의 일반화를 배우기 위해 인간의 관심과 인간의 구조를 충분히 공유하는 학습 장치를 우리는 필요로 한다."

IV. 가족 유사성과 기능주의

지금까지의 논의로부터 우리는 드레이퍼스의 연결주의 비판이 퍼트남의 가족 유사성을 토대로한 기능주의 비판과 그 맥을 같이 하고 있음을 감지할 수 있었을 것이다. 퍼트남이 전개하고 있는 기능주의 비판 논변은 의미(인지적 현상)란 어떤 본질적인 속성을 토대로 일반화되는 것이 아니라 유사함을 토대로 일반화된다는 그러한 단순한 수준의 가족 유사성 논변에만 머무는 것이 아니라, 어떤 유사함들이 가족을 구성하는 유사함들인가—즉, 어떤 유사함이 적절한 유사함이고 어떤 유사함이 적절하지 못한 유사함인가—를 결정하는 데는 일반 지능 혹은 이해의 전 능력이 요구된다는 보다 고차적인 수준의 가족 유사성 논변이었다. 드레이퍼스의 연결주의 비판 논변의 핵심도 다양한 상황에 숙련되게 대처하기 위해서는 적절한 일반화의 능력이 있어야 하며, 그러기 위해서는 상황에 대한 총체적 민감성을 지녀야 한다는 것이다. 여기서 적절한 일반화의 능력이란 어떤 유사함이 일반화의 토대로서 채택되어야 할 적절한 유사함이고 어떤 유사함이 그렇지 않은 부적절한 유사함인가를 결정하는 능력을 말하는 것으로, 퍼트남의 가족 유사성 논변의 핵심도 바로 이 점인 것이다.

물론 드레이퍼스는 상식적 지식의 문제를 주로 다룬 반면, 퍼트남은 고양이, 식물에 관한 것과 같은 상식적 지식뿐만 아니라, 전자, 운동량에 관한 것과 같은 이론적 지식의 문제도 다루고 있다는 점에서 이 두 사람은 차이가 있다. 상식적 지식과 이론적 지식은 인지 능력과 관련하여 커다란 차이가 있을 수도 있다. 그러나 그러한 차이는 기능주의적 인지 이론의 한계를 다루고 있는 지금의 문맥과 관

련해서는 중요한 것이 아니다.

이제 계산주의와 연결주의가 둘 다 모두 기능주의적 인지 모형이라는 점과 이 두 인지 모형이 동시에 가족 유사성 논변에 의해 공략되고 있다는 점은 어떤 관련이 있는지에 대해 살펴보자. 앞의 1절에서 기능주의의 존재론적 입론은 심리 상태란 입력 상태와 출력 상태를 일정하게 매개시켜주는 기능 상태라는 것이었다. 그리고 연결주의도 입력과 출력 간의 매핑을 문제삼는다는 점에서 기능주의적 인지 모형의 하나라고 지적하였다. 그러면 왜 계산주의와 연결주의는 이러한 기능주의적 입론을 공유하는 한, 둘 모두 인지의 적절한 모형이 되기 위해서는 가족 유사성이라는 커다란 장벽을 넘어야만 할까? 우리 인간은 기존에 주어진 어떤 매핑 관계의 토대 위에서 입력을 출력으로 일정하게 변환시키는 작업만 하는 것이 아니라, 기존에 주어진 적이 없는 변형된 입력에 대해서도 그것이 주어진 상황과 관련하여 무시할 만한 변형인가 아니면 진정한 차이를 구성하는 변형인가를 판단하여 상황에 적절한 결과를 내 놓는다. 다시 말해 인간은 어떤 상황에서 어떤 매핑 관계를 설정하는 것이 적절한가 하는 것을 판단하는 총체적 능력을 가지고 있다. 이 총체적 능력은 우리 인간에게 적절성의 지평을 열어주고 있는 것이다. 그러나 기능을 모사하는 것을 목표로 하는 기능주의적 컴퓨터는 (그것이 계산주의적이든 연결주의적이든) 입·출력 간 매핑의 적절성을 자체적으로 판단하지 못하고, 외부로부터 적절하다고 결정되어진 입력과 출력 간의 일정한 매핑 관계의 실행만을 문제삼고 있어, 진정한 인지 모형이 되기 어렵다는 것이다. 이것이 앞에서 퍼트남과 드레이퍼스의 논의를 통해 개진된 가족 유사성 논변의 핵심인 것이다.