인지과학, 제27권 제4호
Korean Journal of Cognitive Science
2016, Vol. 27, No. 4, 541~572.
http://dx.doi.org/10.19066/cogsci.2016.27.4.002

힘 확률 대비 이론에 기반을 둔 인과 추론 연구*

박 주 용†
서울대학교 심리학과 & 심리과학연구소

인과 추론은 심리학에서는 물론 최근 베이스 접근법을 취하는 인지과학자들에 의해서도 활발히 연구되고 있다. 본 연구는 인과추론에 대한 대표적 심리학 이론인 힘-확률대비이론(a power probabilistic contrast theory of causality)을 중심으로 인과 추론의 최근 동향을 개관하고자 한다. 힘-확률대비이론에서는, 원인은 결과를 일으키거나 억제하는 힘(power)인데, 이 힘은 특정한 조건하에서 통계적 상관을 통해 파악될 수 있다고 가정한다. 본 논문에서는 이 이론에 대한 초기의 경험적 지지 증거를 먼저 살펴본 다음, 베이스 접근에 기반을 둔 이론과의 쟁점을 명확히 하고, 원인은 맥락에 무관하게 동일하게 작동한다는 인과적 불변성 가정(causal invariance hypothesis)을 중심으로 한 보다 최근의 연구 결과를 소개하고자 한다. 이 연구들은 종래의 통계적 접근법으로는 잘 설명되지 않는 결과를 제시함으로써, 철학, 통계학, 그리고 인공 지능 등과 같은 인접 분야에 인과성에 대한 힘 이론을 진지하게 고려할 것을 촉구하고 있다.

주제어 : 인과 추리, 힘-확률대비 이론, 인과적 불변성, 베이스 접근

* 이 연구는 서울대학교 인문·사회계열 해외연수 지원금으로 연구되었음.
† 교신저자: 박주용, 서울대학교 심리학과, (08622) 서울 관악구 관악로 1
연구분야: 인지 심리학
Tel: 02-880-9050, E-mail: jooypark@snu.ac.kr

들어가며

인과성은 사건들을 연결해주는 중요한 개념이다. Mackie (1974)는 우주의 시멘트(cement of the universe)로 특징짓기도 하였는데, 시멘트가 모래와 자갈을 묶어주듯, 수많은 사건들 중 일부가 긴밀하게 인과관계로 연결된다고 보았다. 이 인과관계는 시멘트처럼 눈에 보이지 않기 때문에 추론될 수밖에 없다. 혹자는 당구공이 충돌할 때 그 충돌로 인해 정지되어 있던 공이 움직이는 것은 직접 관찰할 수 있는데 왜 인과관계가 관찰되지 않는다고 하는지 물을 수 있다. 이에 대한 답은 철저한 경험주의자인 흄(Hume, 1739)에서 찾을 수 있다. 그는 이런 충돌을 관찰함으로써 근접성과 순차성을 지각할 수 있었지만 필연적 연관성을 관찰할 수 없었다. 실제로 근접성과 순차성을 유지하면서 예를 들어 정지된 공 안에 금속을 넣어두고 자석을 이용하여 충돌 직전에 공을 움직이면, 충돌 때문이라고 잘못 지각한다. 이처럼 실제 충돌로 인한 움직임과 그렇게 보이도록 만든 상황에서의 움직임을 구분하지 못하는 이유는, 인과성이 직접 지각되지 않고 상황으로부터 추론되기 때문이다.

인과관계를 직접 관찰할 수 없지만, 사람들은 물론 동물들도 경험을 바탕으로 사건들 간의 인과적 관련성을 어느 정도 잘 파악해낸다. 사실 인간이 사건들 간의 인과적 관련성을 잘 파악하지 못했다면 아직까지 지구상에 살아남지도 못했을 가능성이 높다. 핵심적인 사건들을 인과적으로 표상하지 못하면 세상에서 일어나는 여러 변화를 예측하고 그에 따라 대처하는 능력이 떨어질 수밖에 없기 때문이다. 실제로 인간은 제대로 된 인과 관계를 파악하지 못하고, 가뭄에 기우제를 지내는 행동에서처럼, 어리석은 행동을 하기도 했다. 하지만 끈질긴 탐구를 통해 중요한 지식을 축적할 수 있었고 오늘날과 같은 발전을 이루었다. 이 지식 가운데 인과적 지식이 중요한 역할을 했다는 것에 대해서는 의심의 여지가 없다.

그렇다면 도대체 사람들은 어떻게 나름 합리적인 인과추론을 해낼 수 있을까? 본 연구는 이와 관련된 연구를 Cheng과 동료들에 의해 지난 30여 년간 발전시켜 온 힘-확률대비이론(이하에서는 간략히 힘 이론으로 표기하겠다)을 중심으로 살펴보고자 한다. 이 이론을 중심으로 살펴보는 이유는 크게 세 가지이다. 첫째로 서로 대립되는 것으로 보이던 이전의 두 이론, 즉 규칙성 이론과 필연성 이론을, 하나로 통합하는 독창적인 이론이기 때문이다. 둘째로 이 이론이 심리학과 인접 분야에서 인과추리와 관련된 다양하고 풍성한 연구와 논의를 촉발시켰기 때문이다. 경험주의 전통 내의 여러 하위 모형과의 이론적 쟁점을 명확히 하며, 이 쟁점을 해결할 수 있는 실험적 과제를 개발하였다. 마지막으로 힘 이론은 심리학을 넘어서 철학, 통계학, 그리고 인공지능 등의 인접 학문에서 인과성을 전제한 분석의 필요성을 제기한다. 현재의 통계적 접근에서는 기본적으로 인과성을 전제하기보다는 경험으로부터 귀납적으로 도출된 개념으로 가정한다. 이 때문에 경험주의 전통의 한계를 그대로 드러내게 되는데, 그 구체적인 내용은 논의를 전개하는 과정에서 언급될 것이다.

힘 이론의 철학적 배경과 힘 추정 방식에 대해서는 이미 소개된 적이 있다(박주용, 2000). 그 이후 지난 20여년간 인과추리와 관련된 연구는 특히 심리학의 여러 분야에서 그 어느 때보다 활발해졌다. 인과추리 연구가 활발해지게 된 데는, 인과관계를 베이스 망(Bayes' net)을 통해 포착하는 기법의 발전에 크게 힘입었다(예, Chater & Oaksford, 2008; Glymour, 2001; Gopnik & Schultz, 2007; Sloman, 2005). 이 망은 변인들을 표시하는 마디(node)와 이 마디들을 연결하는 화살표나 링크로 연결하는 형식적 언어인데, 이를 근사적으로 계산할 수 있는 기법이 발전하면서 그 적용 범위가 급속하게 확대되고 있다. 본 논문에서는 베이스 망에 기반을 둔 이론과 힘 이론간의 논쟁과 함께 인과적 불변성 가정에 바탕을 둔 힘 이론의 최근 연구 성과에 초점을 두어 소개하고자 한다.

이 소개를 용이하게 하게 위해 앞에서 본 흄의 주장이 가진 문제점과 그에 대한 칸트(Kant)의 해결책을 살펴보자. 흄의 결론이 부딪치는 가장 큰 문제는 사람들이 왜 특정한 규칙성은 인과적으로 묶지만 다른 규칙성은 그렇게 하지 않는지를 설명하지 못한다는 점이다. 수탉이 울면 해가 뜨지만 수탉의 울음이 해가 뜨는 원인이라고 생각하는 사람은 없다. 따라서 규칙성으로부터 인과성을 추론해 내려면 모종의 가정 혹은 제약이 필요하다.

칸트(1781)는 그 제약을 인간의 인지 체계의 특성에서 찾고자 하였다. 그는 인과 관계는 물론 필연적 연관성 등의 개념이 감각을 통해 잘 포착되지 않는다는 흄의 주장에 동의한다. 하지만 흄과는 달리 인과성이나 필연성 같은 개념들은 바깥 세상에 존재하는 것이 아니라 세상을 이해하는 우리 마음의 작동 방식에서 제공된다고 본다. 칸트는 인간이 세상을 지각하고 이해하는 과정은, 거울처럼 세상을 수동적으로 반영하는 것이 아니라 경험을 능동적으로 조직화하는 방식을 통해 이루어진다고 주장하였다. 인과성이나 필연성과 같은 개념들은 경험을 통해 배우는 것이 아니라, 순수 개념 혹은 오성의 범주로 선험적으로 내재되어 있다고 보았다. 칸트는 인식의 출발점을 객관적 대상에서 비롯되는 경험으로 보는 대신, 경험을 가능하게 해주는 주관적 구성 방식에서 찾고자 하였다.

흄과 칸트는 단지 인과적 지식의 기원에 대한 견해 차이를 보이는 것만은 아니다. 지식의 기원을 어디에 두는 지에 따라, 인과적 지식을 습득하고 활용하기 위해 요구되는 마음의 구조와 기능이 달라지기 때문이다. 실제로 흄을 따르는 경험주의자들은 최소한의 처리 능력을 가정하였기 때문에, 동물이나 무척추 동물 심지어 단세포 생물을 대상으로 학습 과정을 연구하였다. 그렇지만 칸트가 생각하는 마음은 그보다는 훨씬 복잡한 처리 기제를 전제하지 않을 수 없다. 심리학이나 인지 과학의 용어로 표현하면, 가정된 인지적 구조물(cognitive architecture)의 복잡성에서 큰 차이가 있다는 것이다. 이 기제들은 특정한 기능을 위한 일종의 기관(organs)에 비유된다. 이들이 작동하려면 모종의 경험이 필요할 수 있지만, 경험에만 의존해서는 설명되지 않기 때문에 가정된다. 언어습득(예, Pinker, 1994), 물리적 세계에 대한 아동의 이해(예, Spelke, 1988), 마음 이론(예, Leslie, 1987) 등은 그 몇몇 예에 불과하다. 본 논문에서는 여기에 인과성을 추가하고자 하

는데, 그 근거를 힘 이론에서 찾고자 한다. 이를 위해 인과성을 포괄하는 개념인 연합에서 시작해보자.

학습에 대한 두 접근법 : 연합에서 인지로

심리학에서 학습과 관련된 연구는 오랫동안 두 전통, 즉 기계적 연합에 기반을 두거나 아니면 인지를 우선시하는 전통 간에 대립을 보여 왔다. 연합에 기반을 둔 전통에서는 조건형성을 포함한 모든 학습과정이, 소거, 억제, 강화 등으로 불리지만, 기본적으로 기계적으로 작동하는 연합강도에 의해 주도된다고 본다. 이에 반해 인지적 전통에서는 아무리 단순한 학습이라도 그 안에는 가설검증이나 심성 모형 혹은 논리적 추리와 같은 복잡한 인지 활동이 개입되어 있다고 본다(Shanks, 2010). 이런 인지 활동은, 환경에 대한 적응력을 높이는 과정에서 논리, 확률, 혹은 판단 이론 등과 같은 합리적 혹은 규범적 원리에 따라 작동하는 것으로 간주된다.

이 두 이론 간의 차이를 드러내기 위해 Shanks(2010)가 초점을 둔 현상은 차폐(blocking)이다. Kamin(1969)에 의해 소개된 차폐는 통상 두 단계에 걸친 학습 국면과 그 결과로 인한 학습의 정도를 측정하는 검사 국면으로 구성된다. 학습 국면의 첫 단계에서는 예를 들어 불빛(단서 A)과 전기 쇼크(O)를 여러 번 연합시키고, 두 번째 단계에서는 이전의 단서(A)와 함께 종소리와 같은 새로운 단서(B)를 함께 제시하고 이를 이전의 결과(O)와 다시 연합시키는 시행을 반복한다(AB). 검사 국면에서는 단서 B만을 제시하고 이에 대한 학습이 일어났는지를 확인한다. 통상적인 결과는 단서 B와 결과를 연합시키는 새로운 학습이 일어나지 않는다는 것이다(표 1 1번, 2번 참조). 이 절차는 표 1에서처럼 여러 가지 변형을 통해 언제 학습이 일어나고 또 언제 학습이 일어나지 않는 지를 알려주는 연구 도구로 활용되고 있다.

이 현상을 기계적 연합 이론으로 설명하기 위해, Rescorla와 Wagner는 Rescorla Wagner Rule(이하에서는 RWR)과 연합강도의 가산성, 즉 $V_{A+B} = V_A + V_B$을 제안하였다. RWR은 $\Delta V_i = c \cdot (V_{max} - V_{i-1})$ 로 정의된다. ΔV_i 는 조건자극과 무조건 자극을 i번째로 짝지을 때 이들간의 연합 강도 변화정도, c는 학습 속도를 나타내는 상수, V_{max}는 가능한 연합강도의 최대치, 그리고 V_{i-1} 은 i번째 짝지어지기 바로 전까지의 연합강도를 각각 나타낸다. 이 규칙은 결국 조건 자극과 무조건 자극을 반복적으로 짝지으면 이들 간의 연합 강도는 0에서 시작하여 V_{max}까지 점차 증가하게 되는데 처음에는 빠른 속도로 변하다가 점차 그 증가량이 감소하는 변화를 재현한다. 학습 속도를 나타내는 c의 값이 클수록 더 적은 짝짓기로도 V_{max}에 도달하는데 학습 속도에서의 개인차를 반영한다.

차폐 현상에 대한 인지적 설명은 동일한 두 사건이 연합되더라도 어떤 인과적 맥락인지에 따라 차폐 양상이 달라진다는 Waldmann과 Holyoak(1992)의 연구에서 비롯된다. 이들은 Kamin의 절

〈표 1〉 차폐의 기본 설계와 여러 변형들 (Shanks, 2010 표 1을 확장함)

번호	조건	사전훈련	1단계훈련	2단계훈련	검사단계	결과
1	차폐조건		B+	AB+	A	차폐 o
2	통제조건			AB+	A	차폐 x
3	역행차폐		AB+	B+	A	차폐 o
4*	인과적 모형		B+ (B->O)	AB+	A	차폐 o
5*	진단적 모형		B+ (B<-O)	AB+	A	차폐 x
6	합이하	G+/H+/GH+	B+	AB+	A	차폐 약화됨
7	가산성	G+/H+/GH++	B+	AB+	A	차폐 강화됨
8	최대치	+	B+	AB+	A	차폐 약화됨
9	최대치이하	++	B+	AB+	A	차폐 강화됨
10**	차폐조건		B+	AB+	A	차폐 x
11**	통제조건		C+	AB+	A	차폐 x

* 4, 5 는 Waldman & Holyoak(1992)의 연구에서 사용된 조건들

** 6, 7, 8, 9 행에서의 +, ++ 는 제시된 결과의 크기로, ++는 +보다 더 큰 결과가 제시되었음을 의미함

*** 10, 11은 Maes 등(2016)의 연구에서 사용된 조건들임

차에 따라 차폐 실험을 진행하였는데, 단서와 결과간의 인과적 방향성을 조작하였다(표 1의 4, 5 참조). A, B와 O의 관계가 A, B -> O로 공통 결과로 연결될 때, 예를 들어 창백한 안색(A)이나 경직된 자세(B)에 대해 관찰자가 보이는 정서적 반응(O)에서처럼 예측적 관계일 때, A와 B 간에 단서 경쟁이 일어나는데 이 경우에는 이전의 연구에서처럼 차폐가 일어남을 발견하였다. 하지만 A, B와 O가 A, B <-O 로 공통 결과로 연결될 때, 예를 들어 기침(A)이나 콧물(B)처럼 증상과 감기(O)라는 병으로 연결될 때는, 차폐가 일어나지 않음을 발견하였다. 이 결과는 연합이 기계적으로 일어나는 것이 아니라, 인과적 방향이 고려된 인지 모형에 의해 조절됨을 보여준다. 즉, 학습의 기저 기제는 연합이 아니라 인지라는 것이다.

Waldmann과 Holyoak의 연구 외에도 표 1에 서술된 여러 연구 결과는, RWR을 약간씩 변형하면 설명이 가능하지만, 하나의 모형으로 이들을 모두 설명하지는 못한다. 게다가 변형을 하더라도 설명하기 힘들 결과도 있다. Beckers 등(2005)의 연구가 대표적이다. 이들은 실험참여자들에게 알레르기를 일으키는 단서 A를 제시한 다음 실제로 알레르기가 발생하는 사건을 관찰하도록 하였다(A+). 그 다음에 그 단서(A)를 두 번째 알레르기 발생 물질인 B와 함께 제시하고 나서 그 결과 알레르기가 발생하는 사건을 관찰하게 하였다(AB+). 통제 조건에서는 하나의 단서만 제시한 경우를 보여주지 않은 상태에서 C와 D를 함께 제시하고 알레르기가 발생하도록 하였다(CD+). 여기까지의 조작은 전통적인 차폐조작과 크게 다르지 않다. 그런데 Beckers 등은 이 모든

시행에 앞서 실험참여자로 하여금 사전 훈련을 받도록 하였다. 사전 훈련에서는 본 시행과 다른 단서들을 사용하여 개별단서와 결합단서를 결과와 가산적(G+, H+, GH++) 혹은 비가산적(G+, H+, GH+)으로 연결시켰다 (각각 표 1의 7, 6 행 참조). 여기서 +, ++는 결과의 크기를 나타내는데, ++는 +보다 더 큰 결과를 제시했다는 것을 의미한다. 이런 사전 훈련 후 차폐조건과 통제 조건을 비교하면, 가산적 사전 훈련을 받은 집단이 비가산적 훈련을 받은 집단보다 더 큰 차폐가 나타났다. 이 결과는 사람에게서는 물론 쥐를 사용한 실험에서도 반복적으로 관찰되었다.

Beckers 등(2005)은 이런 결과를 설명하기 위해 다음과 같은 추론 과정이 개입된다고 제안하였다: "B가 결과의 원인이 아니라는 것을 추론하기 위해서, 사람들은 원인들이 가산적으로 결과를 일으킨다고 가정해야 할 뿐만 아니라, A에 추가하여 B가 더해지는 것이 A에 의해 산출된 결과의 크기를 증가시키지 않는다는 것을 경험적으로 확인할 수 있어야만 한다"(239쪽). 가산적 조건에서는 이 확인이 용이하여 차폐가 일어나지만, 비가산적 조건에서는 이를 확인하기 어렵기 때문에 차폐가 덜 일어난다는 것이다.

연합이 추론 과정을 통해 일어난다는 주장은 Beckers 뿐만 아니라 다른 동물 학습 연구자들에 의해서도 제기되었다. Mitchell, De Houwer, 그리고 Lovibond(2009)는 연합 학습이 본질적으로 "수고스럽고 주의를 요하는 추리과정(effortful, attention-demanding reasoning processes)"이며, 그 추리 과정의 산물로 "사건들에 대한 의식적, 선언적, 명제적 지식이 생성된다(produces conscious, declarative, propositional knowledge about those events)"고 주장하였다(186쪽). Mitchell 등(2009)은 연합이 명제적 연산에 의한다는 자신들의 입장을, 단지 인간 학습에만 국한시키지 않고 동물의 연합학습에도 적용된다고 주장하였다. 이들은 사람과 동물이 보이는 차이는 그 정교함의 문제이지 기본적으로 비슷한 진화과정을 통해 획득된 논리적 연산이라는 주장과 함께 관련 증거를 제시하였다(196쪽 이하). 실제로 표 1에서 제시된 여러 다른 차폐효과는 연역적 추리 과정으로 쉽게 설명될 수 있다.

인과적 학습에 대한 심리학적 모형들

인과성에 대한 연합 이론과 규칙성 이론

인과성에 대한 초기 연구는, 흄과 흄을 계승한 행동주의적 전통에 기반을 두고, 동물학습에서 관찰되는 조건형성 절차에 준하는 방법으로 연구되어져 왔다. 구체적인 쟁점은 조건형성에 준하는 절차에서 인과 판단을 하도록 했을 때, 사람들의 반응이 RWR에 따르는 지 아니면 통계적 규범 모형인 ΔP 모형에 따르는 지였다. 고전적 조건형성에서는 조건 자극과 무조건자극을 짝지어 제시하여 학습을 시킨 다음, 조건 자극에 대한 반응 강도를 측정한다. 마찬가지로 인과 판단

에서도 한 단서(원인)와 다른 단서(결과)의 제시 방식을 여러 가지로 조작한 다음 원인 단서가 얼마만큼 결과를 예측하는지 평정하도록 한다. 조건형성 연구에서는 공포자극에 대한 반응 억제 수준을 통해 공포 조건 형성의 정도를 양적으로 변환하지만, 인과 판단에서는 표 2에서와 같은 소위 분할표(contingency table)를 사용하여 원인 단서가 결과를 어느 정도 예측하는 지의 정도를 나타낸다.

통상 두 조건부 확률의 차이로 정의되는 ΔP가 많이 사용되는데, 각 조건부 확률을 표 2에 제시된 빈도 정보로 나타내면 다음과 같다:

$$\Delta P = P(e/i) - P(e/\sim i)$$
$$= L/(L+N) - M/(M+O) \quad \cdots\cdots (1)$$

〈표 2〉 분할표(원인 i가 존재하거나 존재하지 않을 때 각각 결과 e가 발생하거나 발생하지 않은 사건의 빈도를 나타낸 표)

	원인	
결과	i	~i
e	L	M
~e	N	O

인과 판단에 사용된 실제 실험은 세부적인 내용에서는 약간 다르지만 기본적으로 다음과 같은 상황으로 구성된다. Shanks 등(1996)은 실험참여자로 하여금 컴퓨터 비디오 게임을 하면서 중간 중간에 특정 단서의 결과 발생 예측력을 판단하도록 하였다. 이 게임에는 전투가 진행되는 지역을 지나가는 탱크가 나오는데, 매 시행마다 탱크가 중간에 폭파되거나 아니면 무사히 그 지역을 벗어나는 두 결과 중 하나의 사건이 일어난다. 서로 다른 조건을 만들어내기 위해, 이들은 탱크의 위장색을 여러 가지로 조작하였다. 실험참여자의 과제는 위장색에 따라 적군에게 눈에 띄는 정도가 달라져 이 위험 지역을 벗어나는 데 도움을 주는 지를 판단하는 것이었다. 여기서 서로 다른 위장색은 분할표의 i에 해당한다. 실험자는 P(e|i)를 여러 가지로 조작하는 동시에, P(e|~i)를 임의로 조작할 수 있다. 이 실험에서 P(e|~i)는 위험 지역 자체의 위험성이라 할 수 있는데, 위장색이 없는 탱크가 파괴되는 정도로 이 값이 1이면 극단적으로 위험한 지역이고 0이면 그 지역이 전혀 위험하지 않음을 나타낸다. 실험참여자들은 매 5시행마다 특정한 위장색이 얼마나 탱크를 안전하게 해주는 지를 100에서 -100점 척도를 이용하여 반응하였다.

각각 20명으로 구성된 4집단이 여러 ΔP조건에서 과제를 수행하였다: ΔP = .75-.25; ΔP = .25-.75; ΔP = .75-.75; ΔP = .25-.25. 실험참여자들의 판단 점수를 평균한 결과를 그래프로 나타내면, ΔP = .5인 조건에서는 0에서 시작하여 점차 감소하는 비율로 평정 점수가 50점에 가까워

졌다. ΔP = -.5인 조건에서는 0에서 시작하여 점차 감소하는 비율로 평정 점수가 -50점에 가까워졌다. ΔP = 0인 두 조건에서는 평정 점수가 모두 0점 근처에 있었지만 .75/.75 조건에서는 시행 초기에 20점 내외로 비교적 높은 점수를 보이다가 점차 0점에 가까워졌다. Shanks 등(1996)은 이상의 결과를 각각 ΔP와 RWR을 이용하여 모사하였다. 그 결과 ΔP에서는 시행수가 늘어남에 따라 관찰되는 점수 증가 패턴이 관찰되지 않고, 초기의 불안정성도 나타나지 않았는데 반해, RWR을 사용했을 때에는 실제 사람들의 반응과 비슷한 패턴을 얻었다. 이를 바탕으로 이들은 사람들의 인과 판단 수행도 RWR에 따르는 것 같다고 결론지었다. 일부 동물학습 연구자들은 여전히 RWR을 중심으로 학습은 물론 인과성을 설명하고자 하고 있다(예, Shanks, 2010; Soto et al., 2014). 하지만 Danks(2003)는 충분히 많은 시행이 반복되어 연합 강도가 점근선에 이르면, RWR과 ΔP 규칙이 결국 같아짐을 보였다. 그리고 앞에서 본 것처럼 RWR로 설명되지 않는 결과들이 많아지면서(예, Beckers et al., 2005; Mitchell et al., 2009), RWR에 근거한 인과적 이론은 이전보다 크게 약화되었다.

연합이론인 RWR에 대한 규칙성이론에서의 대항마는 위에서 본 ΔP 모형과 이를 일반화한 확률대비이론이다. 이 모형에 대한 보다 자세한 소개는 이미 이루어졌기 때문에 여기서는 생략하기로 한다(박주용, 2000).

인과성에 대한 힘-확률대비 이론

Cheng과 Novick(1990)의 확률대비이론은 인과추론이 규범적으로 일어난다는 점을 명확히 하면서, 이전에 설명하기 어려운 현상들을 사람들이 분할표의 정보를 융통성 있게 사용하는 방식으로 설명할 수 있게 하였다. 하지만 몇 몇 결과들은 여전히 설명할 수 없다. 그 중 하나는 원인이 없어도 결과가 항상 발생하는 상황에서 어떤 원인의 인과적 강도를 추정하는 문제이다. 이 경우 실제로 어떤 원인이 결과를 일으킬 확률이 1이면, $\Delta P = P(e|i) - P(e|\sim i) = 1-1 = 0$이 된다. 즉, 실제로 원인이 결과를 일으키는데 ΔP로는 포착되지 않는다. Cheng(1997)이 드는 구체적 예는 피부가 민감한 환자를 대상으로 한 알레르기 반응의 결과이다. 의사는 알레르기 유발인을 찾기 위해 환자의 등에 상처를 내는데, 이 환자의 경우 피부가 민감하여 상처가 나자마자 알레르기 유발인을 넣기 전에 이미 물집이 생긴다. 따라서 알레르기 유발인을 넣거나 넣지 않거나 항상 물집이 생기게 된다. 이런 상황에서 의사는 이 환자는 알레르기 유발인이 없다고 결론을 내리지 않을 것이고 내려서도 안 될 것이다.

그런데 위에서와 마찬가지로 ΔP=0인데, 조건부 확률이 1보다 작은 경우에는 i가 결과를 일으키지 않는다고 결론을 짓는다. 왜 똑같이 ΔP=0인데 이런 차이가 나타날까? 확률대비이론과 ΔP 모형은 이런 차이를 설명하지 못한다. 여기에 추가하여 설명하지 못하는 또 다른 결과는 억제적 원인을 평가할 때 관찰된다. 결과가 일어나기 위해서는 촉진적 원인이 있어야 한다. 그런데 그

런 원인이 없을 때와 원인이 있지만 억제적인 역할을 하면, $P(e|i) = P(e|\sim i) = 0$ 이 된다. 예를 들어 어떤 약이 두통을 억제하는 지를 알아보기 위해 실험을 했는데 통제집단에게는 위약처방을 하고 실험집단에게는 그 두통약을 처방했다고 하자. 그 결과 두 집단 모두에게서 두통이 사라졌다면 아마도 이 두통약이 효과가 있다는 결론보다는 이 실험으로는 알 수 없다는 결론이 선호될 것이다.

Wu와 Cheng(1999)은 위에서 언급된 상황을 실제로 대학신입생을 대상으로 확인하였다. 여러 다른 시나리오를 사용하였지만, 다음과 같은 공통점을 갖는다: 어떤 연구자가 특정 변인을 조작한 실험 결과 통제 집단과 실험 집단 간에 조작에 따른 차이가 관찰되지 않았기 때문에, 그 조작이 효과가 없다는 결론을 내렸다. 실험집단과 통제 집단 간에 차이가 없다는 결과와 함께, 원인이 없을 때 결과의 발생 확률을 다르게 설정하였다. 한 집단에서는 항상 결과가 일어났고 다른 집단에서는 가끔씩, 그리고 또 다른 집단에서는 전혀 결과가 관찰되지 않았다. 이런 상황에서 실험참여자들은 조작이 효과가 없었다는 결론에 동의하는지, 실험 결과로부터 의미 있는 정보를 얻을 수 없다고 생각하는지, 아니면 연구자의 주장과는 달리 그 조작이 효과적이었다고 생각하는지 중에서 하나를 고르도록 하였다. 그 결과 생성적 원인일 경우에는 항상 결과가 나타나면 의미 있는 정보를 얻을 수 없다가 가장 많이 선택되었고, 전혀 나타나지 않으면 효과가 없었다가 가장 많이 선택되었다. 억제적 조건의 경우 항상 나타나는 조건과 가끔 나타나는 조건 모두에서 효과가 없었다가 가장 많이 선택되었고, 전혀 결과가 관찰되지 않았을 때는 정보가 없다를 가장 많이 선택하였다. 이 연구에서 사용된 조건들은 모두 $\Delta P = 0$ 이었기 때문에, 규칙성 모형에 입각하면, 원인의 방향성은 물론 결과 발생 확률 즉 $P(e)$ 의 차이에 따라 달리 선택할 이유를 찾기 어렵다.

지금까지 논의된 문제점을 해결하기 위해 Cheng(1997)은 통계적 규칙성에 기반을 둔 확률대비 이론에 칸트의 생각을 통합한 힘-확률대비이론을 제안하였다. 이 이론은 필연성 이론과 규칙성 이론을 통합하는 동시에, 규칙성 이론이 언제 타당하고 또 언제 타당하지 않은 지를 보여주는 점에서 그 자체로 독창적인 연구이다. 더 중요한 점은 위에서 언급된 문제를 해결하는 한편, 인과적 학습에 대한 새로운 논의를 촉발시켰다는 것이다.

힘 이론은 사람들이 인과 학습과 관련하여 다음과 같은 세 가지 일반적인 사전 믿음이 있다고 가정한다:

1. *i와 a (i 이외에 결과에 영향을 주는 모든 원인들)만 e*에 영향을 주고 그 밖에 다른 원인은 e를 일으키지 않으며, e를 일으키는 영향력은, 각각 p_a와 p_i로 나타낼 수 있다.
2. a와 i는 서로 독립적으로 e에 영향을 준다.
3. a와 i의 인과적 힘은 a와 i의 발생 빈도와 무관하다.

힘으로서의 원인은 관찰 가능한 결과를 일으키거나 방해하는데, 힘 자체가 직접 관찰되지 않기 때문에 추정될 수밖에 없다. 힘을 추정할 때는 그 힘이 결과를 촉진하거나 억제하는 지에 따라 추정 방식이 달라진다, 먼저 촉진적 원인의 경우를 살펴보면 다음과 같다. 관찰가능한 결과 P(e)는 확률이론의 합 공식에 따라 다음과 같이 표현될 수 있다:

$$P(e) = P(i)*p_i + P(a)*p_a - P(a)*p_a*P(i)*p_i \quad \cdots\cdots (2)$$

그런데

$$\Delta P_i = P(e|i) - P(e|\sim i) \quad \cdots\cdots (3)$$

를 계산하기 위해서는 두 조건부 확률값이 필요하다. 이들은 식 (2)를 각각 i와 ~i로 조건화하여 얻어질 수 있다:

$$P(e|i) = p_i + P(a|i)*p_a - p_i * P(a|i)*p_a \quad \cdots\cdots (4)$$

$$P(e|\sim i) = P(a|\sim i)*p_a \quad \cdots\cdots (5)$$

이다. 이 두 식을 (1)에 대입한 다음 p_i에 대해 정리하면

$$p_i = \frac{\Delta P_i - [P(a|i) - P(a|\sim i)]*p_a}{1 - P(a|i)*p_a} \quad \cdots\cdots (6)$$

가 된다. 그런데 위의 가정 2로부터, P(a|i)=P(a|~i)=P(a) 가 되므로 식 (6)은
다음과 같이 정리될 수 있다:

$$p_i = \frac{\Delta P_i}{1 - P(a)*p_a} = \frac{\Delta P_i}{1 - P(e|\sim i)} \quad \cdots\cdots (7)$$

한편 i가 억제적 역할을 할 때는, 결과가 관찰될 확률은 다음의 식과 같다:

$$P(e)=P(a)*p_a*[1-P(i)*p_i] \quad \cdots\cdots (8)$$

이 식을 i로 조건화시키면

$$P(e|i)=P(a|i)*pa*(1-pi) \quad \cdots\cdots (9)$$

다시 ~i로 조건화시키면

$$P(e|\sim i)=P(a|\sim i)*pa \quad \cdots\cdots (10)$$

식 (9)와 (10)을 식 (1)에 대입하고 p_i에 대해 정리하면

$$p_i = \frac{-\triangle P_i}{P(a)*p_a} = \frac{-\triangle P_i}{P(e|\sim i)} \quad \cdots\cdots (11)$$

이다. 식 (7)과 (11)에서 볼 수 있듯이, $P(a)*pa$가 어떤 값을 갖는 지에 따라 ΔP_i로부터 p_i 가 얼마만큼 정확하게 추정될 수 있는 지가 결정된다. 촉진적 원인의 경우 $P(a)*pa$가 0에 가까울수록 억제적 원인의 경우는 $P(a)*pa$가 1에 가까워질수록 ΔP에 의해 정확한 추정이 가능해진다. 그런데 $P(a)*pa$ 값은 결국 i와 독립적인 다른 모든 원인에 의해 발생하는 결과이므로, $P(e|\sim i)$로 추정될 수 있다.

인과 추리의 핵심이 원인과 결과간의 통계적 상관이 아니라 지금까지 논의된 인과적 힘인지를 검증하기 위해, Buehner 등(2003)은 일련의 실험을 진행하였다. 이들의 실험 1에서는 표 3과 같이 $P(e|c)$와 $P(e|\sim c)$가 여러 가지 방식으로 다르게 조합되었다. 이들로부터 식 (1)을 이용하여 ΔP값이, 그리고 식 (7)과 (11)를 이용하여 각 조건에서의 인과적 힘이 계산되었다.

표 3에는 사람을 대상으로 한 인과 강도 추정값의 평균과 표준편차도 포함되어 있다. 이 값들은 다음과 같은 실험에서 얻어졌다. 한 집단은 억제적 조합에 대해 다른 집단은 생성적 조합에 대해 표 3에 제시된 여러 조건에 대해 반응하도록 하였다. 억제적 조합의 경우 실험참여자는 바이러스 치료 백신 개발자 역할을 하도록 하였다. 총 15개의 백신에 대해 평가가 이루어졌는데, 각 백신의 효과를 알아보는 실험에서는 16마리의 쥐로부터의 결과가 제시되었다. 각각의 백신의 효과는 100점 척도를 이용하여 평가하도록 하였다. 생성적 조합의 경우 미생물학자의 역할을 맡도록 하면서, 특정한 광선에의 노출이 바이러스의 유전자 변이에 미치는 영향을 평가하도록 하였다.

표 3에 제시된 두 모형의 예측과 사람들의 반응을 비교해보면, 표 하단의 4줄에서 볼 수 있듯이, ΔP 값이 같더라도 P(e|~i) 가 변하면, 사람들의 판단이 크게 달라진다. 달라지는 양상은 힘 이론에서의 예측되는 것처럼, 억제적 조합과 생성적 조합에서 서로 반대 방향으로 나타난다. Buhner 등의 연구와 별도로, Lober와 Shanks(2000)가 RWM과 힘 이론을 비교하는 연구에서도 비슷한 결과가 얻어졌다. 하지만 이 연구에서도, ΔP모형과 힘 이론 모두 0을 예측하는 아래의 4 조

〈표 3〉 억제와 생성적 조합에서 조건부 확률값, *ΔP*, 힘, 그리고 사람들의 평균 반응점수(Buhner 등, 2003의 표 1 & 2로부터 재구성함)

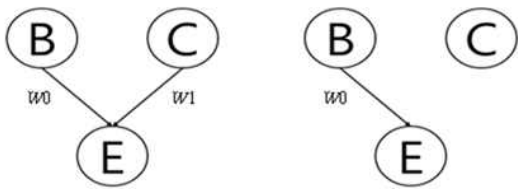
억제적 조합						생성적 조합					
P(e|c)	P(e|~c)	ΔP	인과적 힘	강도 평균	표준 편차	P(e|c)	P(e|~c)	ΔP	인과적 힘	강도 평균	표준 편차
0.00	1.00	-1.00	1.00	94	13	1.00	0.00	1.00	1.00	89	20
0.00	0.75	-0.75	1.00	85	20	1.00	0.25	0.75	1.00	77	24
0.25	1.00	-0.75	0.75	72	17	0.75	0.00	0.75	0.75	69	22
0.00	0.50	-0.50	1.00	79	21	1.00	0.50	0.50	1.00	71	23
0.25	0.75	-0.50	0.67	65	19	0.75	0.25	0.50	0.67	54	24
0.50	1.00	-0.50	0.50	46	22	0.50	0.00	0.50	0.50	57	23
0.00	0.25	-0.25	1.00	72	30	1.00	0.75	0.25	1.00	58	32
0.25	0.50	-0.25	0.50	59	20	0.75	0.50	0.25	0.50	47	26
0.50	0.75	-0.25	0.33	43	22	0.50	0.25	0.25	0.33	46	23
0.75	1.00	-0.25	0.25	21	16	0.25	0.00	0.25	0.25	34	27
0.00	0.00	0.00	불능	45	41	1.00	1.00	0.00	불능	41	37
0.25	0.25	0.00	0.00	48	29	0.75	0.75	0.00	0.00	43	27
0.50	0.50	0.00	0.00	34	23	0.50	0.50	0.00	0.00	37	25
0.75	0.75	0.00	0.00	23	22	0.25	0.25	0.00	0.00	29	21
1.00	1.00	0.00	0.00	9	20	1.00	0.00	0.00	0.00	9	20

건에서 사람들의 강도 평정 점수는 유의미하게 0보다 높았다. 또한 힘이 같은데도 *ΔP*값에 따라 강도 판단이 달라지는 결과도 관찰되었는데 이들은 힘 이론과 일치되지 않는다.

힘 이론에서의 예측과 다른 결과에 대해, Buhner 등(2003)은 실험참여자들이 과제를 잘못된 이해하거나 많은 자료를 처리하느라 기억에 부담을 주었기 때문에 생긴 결과일 가능성을 제기하였다. 실제로 과제 지시를 명확히 하고 기억 부담을 최소화하기 위해 전체 자료를 그림으로 제시하는 추가 실험을 수행하여(실험 2). 힘 이론의 예측과 일치하는 결과를 얻었다. 이를 바탕으로, 몇몇 연구에서 관찰된 사람들의 판단과 힘 이론의 예언과의 차이는 수행과 역능(competence) 간의 차이일 가능성이 높다고 주장한다. 즉 원칙적으로는 할 수 있지만 실제 상황에서의 수행은 그 원칙에 미치지 못할 수 있다는 것이다.

이제 인과 추론이 힘을 중심으로 이루어진다는 주장을 전제하고 논의를 계속해 보자. 힘 이론에서는 인과적 힘이 결과에 독립적으로 영향을 준다고 가정한다. 따라서 하나의 결과에 여러 원

인이 작용할 때는 이들이 합해져야 한다. 가장 간단한 상황은 두 개의 다른 원인이 하나의 결과를 일으키는 그림 1의 왼쪽 그래프에서 볼 수 있다. 원인 B와 C가 각각 w_0, w_1의 강도로 결과에 영향을 준다고 할 때 이 두 원인이 동시에 존재하면 이들이 영향이 합해져야 한다. 힘 이론에서는 인과적 영향력이 합해질 때 각 영향력이 무조건 더해지는 것이 아니라 결과 변수의 유형에 따라 다를 수밖에 없다고 본다.



(그림 1) 결과 E의 원인 후보 단서인 C가 실제로 원인일 때(왼쪽의 그래프 1)와 원인이 아닐 때(오른 쪽의 그래프 0)를 대조시킨 그래프임. B는 배경 원인임. B, C, E는 있다/없다 나타낼 수 있는 양가 변수이고, w_0, w_1은 각각 B와 C 가 E에 주는 인과 강도를 나타냄.

먼저 결과 변수가 연속 변수일 때는, 최대값을 넘지 않는 범위에서 생성적 원인과 억제적 원인의 힘이 각각 더해진다. 그렇지만, 결과가 0과 1을 갖는 범주적 변수일 경우는 이런 가산성이 적용되지 않는다. 그 대신 원인이 결과에 작용하는 방식에 따라 다음과 같이 통합된다. 두 원인이 모두 결과를 일으키는 생성적 원인일 경우, 결과가 관찰될 확률은 다음과 같다:

$$P(e+|b,\ c;\ w_0,\ w_1) = w_0b + w_1c - w_0w_1bc \quad \cdots\cdots (12)$$

이 식은 집합에서의 합집합 연산과 같은데 종종 noisy-OR 함수라 불린다. 결과가 발생하지 않도록 막는 억제적 원인일 경우의 확률은 다음과 같다:

$$P(e+|b,\ c;\ w_0,\ w_1) = w_0b - w_0w_1bc \quad \cdots\cdots (13)$$

이 식은 noisy-AND-NOT 함수라 불린다.

지금까지의 논의된 힘 이론에 대한 초기 연구 결과를 정리해 보자. 힘 이론은 철학에서의 필연성과 규칙성 이론 간의 관계를, 이론과 검증 가능한 모델에 비유하면서 통합을 시도한다. 힘들은 서로 독립적으로 결과에 영향을 주는데, 직접 관찰되지 않지만 관찰 가능한 값인 ΔP와 P(e|~i)를 고려하여 추정될 수 있다. P(e|~i) 값에 따라 그리고 원인이 촉진적으로 혹은 억제적으로 작용하는 지에 따라 힘은 ΔP와 비슷할 수도 있지만 다를 수 있다. 이상의 주장을 경험적으로 지지하는 증거는 두 가지이다. 하나는, 촉진적 원인의 경우 P(e|~i) 값이 1이면서 ΔP가 0일

때와, 억제적 원인의 경우 P(e|~i) 값이 0이면서 ΔP가 0일 때는, 많은 사람들이 힘을 추정할 수 없다고 판단한다는 결과이다. 또 다른 지지 증거는, 인과적 강도 추정 과제에서 사람들의 수행이 ΔP보다는, 힘에 의한 예측에 더 잘 따른다는 결과이다. 힘 이론에서는 또한, 둘 이상의 원인을 통합할 때, 결과 변수 유형에 따라 사용되어져야 하는 통합 함수가 달라질 수밖에 없다는 점이 논리적으로 도출된다. 즉, 독립성을 유지하려면 이산 변수의 경우 noisy-OR나 noisy-AND-NOT가 사용되어져야 하고, 연속 변수일 경우에는 가산적 함수가 사용되어져야 한다는 것이다. 여기서 주목할 점은 이런 차이가 경험적으로 발견된 것이 아니라 논리적 분석의 귀결이라는 점이다.

인과 귀납에 대한 베이스 접근법

지난 2~30년 동안 인지 심리학은 엄청난 발전을 이루었는데, 그 중 가장 큰 변화를 하나만 꼽는다면 베이스 접근법을 사용한 인지 모형의 등장이라 할 수 있다 (예, Chater & Oakford, 2008). 베이스 접근법은 베이스 공식을 이용하는데, 가설을 H 증거를 E라 하면 다음과 같이 표현될 수 있다:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

어떤 증거가 관찰되었을 때, 어떤 가설이 참일 확률은 그 가설이 참일 때 그 증거를 얻을 확률(우도)에 가설이 참이라는 믿음과 증거가 관찰될 확률의 비율을 곱해서 구할 수 있다는 것이다. 예를 들어 어떤 동전을 던졌을 때 연속 세 번 앞면이 나왔을 때 과연 이 동전이 앞면이 나올 확률이 0.5인 제대로 된 동전인지 아니면 누군가 조작하여 앞면이 나올 확률을 높인 편향된 동전인지를 의심해 볼 수 있다. 이 가능성은, 위의 공식을 이용하여 두 가설 간의 비를 통해 확인될 수 있다. 이를 베이스 인자라 한다. 즉,

$$\text{베이스 인자} = \frac{P(\text{자료}|\text{가설 1})}{P(\text{자료}|\text{가설 2})}$$

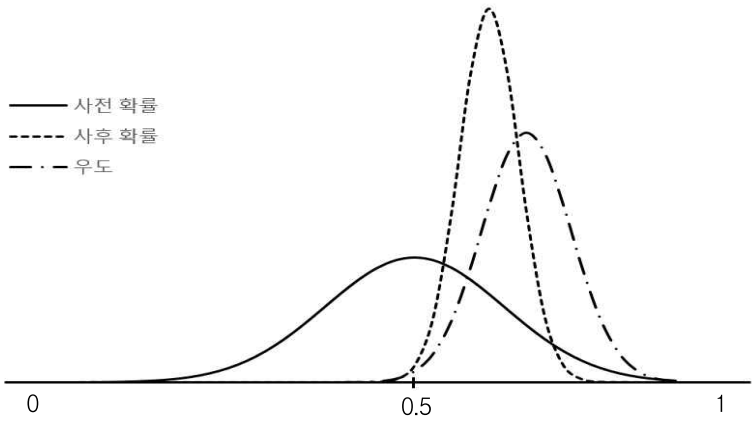
예를 들어 앞면이 나올 확률이 1.0인 동전이 0.1% 존재한다고 알려진 상황에서, 앞면이 연속해서 3번 나온 자료가 관찰되었다고 가정해보자. 이 상황에서 동전이 나올 확률이 0.5인지(가설 1) 혹은 1.0인지(가설 2)를 비교하려면 각각의 확률을 구해야 한다. 앞면이 연속 3번 나온 자료가 관찰되었으므로, $P(D|\theta_1=0.5) = 1/8$이고, $P(D|\theta_2=1) = 1$이다. 또한 $P(\theta_1=0.5) = 0.999$이고 $P(\theta_2=1) = 0.001$이다. 따라서 베이스인자는 $P(D|\theta_1=0.5)/P(D|\theta_2=1) = 1/8*999 = 125$가 된다. 만일

연속 5번이 나 왔다고 하면, 이 비는 31이 되고, 10번 나왔다고 하면 1이 된다. 이 비가 작을수록 대립가설이 참일 가능성이 높기 때문에 앞면이 나온 횟수가 3, 5, 10회를 증가함에 따라 앞면만 나오는 동전일 가능성이 커짐을 보여준다.

베이스 접근법의 가장 중요한 특성은 가설을 위 예에서처럼 0.5 혹은 1.0과 같은 이산적인 값으로 설정할 수도 있지만, 가능한 범위 내에서 확률 분포로 나타낼 수 있다는 것이다. 예를 들어 가설 1은 0.5일 확률이 가장 높고 0.4나 0.6일 확률은 0.5에 비해 낮은 그림 2에서의 사전 확률 분포를 형성한다. 자료가 주어지면 분포가 변경되게 되는데, 사후 확률 분포가 그것이다 이 분포는 사전분포와 비교하여 평균이 이동하였고 분포가 더 좁아졌다. 즉 가설이 더 정교해졌다고 할 수 있다. 일반적으로 자료가 추가되면 가설의 분산이 작아져서 가설이 이전보다 더 정교해진다.

베이스 접근법이 인과추리와 연결되는 부분은, 우리의 사전 지식을 새로운 경험 혹은 자료를 이용하여 지식을 변경 혹은 갱신할 수 있게 한다는 점이다. 이 지식은 어디까지나 현재까지의 믿음(가설)과 자료에 근거하여 확률적으로 정해지며, 추가 자료가 발견될 때마다 가설은 수정된다. 가설이 수정되는 정도는 베이스 규칙에 따른다. 예컨대 가설 1에서 1000개의 새로운 자료를 수집하면 가설에 대한 확률을 조정하게 되고 그 다음 500개를 더하면 다시 가설을 수정하게 된다. 이렇게 수정된 가설은 가설 1에서 1500개의 자료를 이용하여 수정한 가설과 일치하게 된다. 요컨대 베이스 접근은 제한된 자료를 활용하여 확률적이기는 하지만 합리적 추론을 가능하게 해준다.

인지 과정에 대한 연구 장면에서 베이스 접근은 모델링의 도구로 활용된다. 즉 인지 과정에 대한 서로 다른 복잡성을 가진 모델을 만들고 이들 중 어느 모델이 더 자료를 잘 설명하는 지를 비교하여 그 타당성을 확인한다. 베이스 모델링을 위해서는 먼저 연구 대상이 되는 현상과



(그림 2) 베이스 추리에서 고려되는 정보를 도식적으로 표현한 예. 사전확률분포가 우도 혹은 자료를 관찰한 후 사후 확률로 변화되었음을 보여줌.

관련이 되는 과제를 선택하고, 이 과제 수행과 관련된 처리과정을 명확히 하는데, 여기에는 개별 변인의 사전 확률에 대한 분포, 변수들 간의 관계를 규정하는 함수 등이 포함된다.

베이스 접근법이 최근 각광을 받게 된 것은 그 동안 계산이 쉽지 않고 계산량이 많았기 때문이다. 이 접근법에서 가설은 하나의 확률값보다는 분포로 취급되기 때문에 가능한 모든 값을 고려하여 확률을 계산해야 한다. 따라서 합하거나 적분해야할 양이 많다. 이러한 계산의 문제는 Monte Carlo 방법이 개발되면서 상당 부분 해결되었다. 이 방법은 무선적으로 생성된 수를 사용하여 우리가 필요로 하는 값을 찾아낼 수 있게 해준다. 직관적 이해를 돕기 위해 자주 사용되는 예는 원의 넓이를 구하는 것이다. 가로 세로 1cm 인 정사각형의 한 꼭지점을 원점으로 하고, 한 변의 길이가 1인 원을 그리면, 정사각형 안에 1/4원이 그려진다. 편의상 정사각형이 좌표평면의 1사분면에 놓여 있고 원의 중심을 원점이라 하자. Matlab 프로그램을 이용하여, 0과 1 사이에 있는 두 수 (x, y)쌍을 무선적으로 10,000개를 생성한다. 이렇게 생성된 10,000개의 점 중에서 $x^2+y^2 < 1$을 만족시키는 점의 개수의 비율이 1/4원의 넓이가 된다. 물론 10,000개가 아니라 더 많은 점을 사용하면 더 정확하게 1/4원의 면적을 구할 수 있다. 이 예의 핵심은 적분 문제를 무선적으로 생성된 순서쌍의 비율로 대체하여 근사값을 계산해 낼 수 있다는 것이다.

인과지지 모형(causal support model)

이와 같은 베이스 접근법을 인과적 학습 영역에 본격적으로 적용한 것은 Griffiths와 Tenenbaum (2005)이다. 이들은 인과추리에서 해결되어야 할 문제는, 표 2에서와 같은 분할표로 제시된 자료들을 바탕으로 원인과 결과 간의 인과 구조와 그 강도를 찾아내는 것으로 정의하였다. 이 정의를 바탕으로, 인과 추리에 대한 선행 연구에서는 구조에 대해 고려하지 않은 채, 강도 문제만을 다루어 왔다고 비판한다. 베이스 모형에서는 구조가 정해져야 강도 추정이 가능한데, Griffiths와 Tenenbaum(2005)은 이런 특징을 그대로 인과 추리에 적용하였다. 이를 그림 1을 이용하여 이들의 연구 전략을 두 단계로 나누어 설명하면 다음과 같다. 먼저 그래프 1에서처럼 C가 E를 일으키는지 아니면 그래프 0에서처럼 인과적 연결이 존재하지 않는 지를 판단하는 것이다. 그 다음에 그래프 1에서처럼 인과 연결이 존재할 경우 그 강도를 추정하는 것이다. 이들이 나눈 이 두 단계는 사실 베이스 접근법에서 주어진 자료를 설명할 수 있는 모델을 찾아내는 모델 선택(model selection)과 모델 내의 여러 모수들의 값을 추정하는 모수 추정(parameter estimation)에 각각 해당한다.

이 전략에 입각하여 인과 추리 과정을 베이스 방법으로 분석해보자. 이 분석을 위해서는 표 2와 같은 분할표처럼 원인과 결과의 유무가 정리된 관찰 자료(D)와, 결과 발생에 영향을 주는 가능한 두 모형, 그리고 원인과 결과의 사전 확률 분포가 정해져야 한다. 여기에 추가하여 만일 그림 1의 그래프 1처럼 배경 원인 B외에 C가 또 다른 원인으로 작용할 경우 이 두 변인이 어떤

방식으로 통합되어 결과를 일으키는 지를 규정하는 함수도 정해져야 한다. 관찰 자료는 연구자의 의도나 실제 관찰 자료로부터 얻을 수 있다. 결과 발생에 영향을 주는 모형은 그림 1에 제시되었다. 그림 1의 그래프 1의 경우는 원인에 의한 영향이 통합되어져야 한다. 통합 함수는 noisy-OR, noisy-AND-NOT, 혹은 선형적 함수가 될 수 있는데, 이들의 연구에서는, 힘 확률 모형에서처럼 noisy-logical 함수를 사용하였다. 그리고 w_0와 w_1은 균등 분포에 따른다고 가정하였다.

이제 관찰된 자료 D를 바탕으로 베이즈 추론을 이용하여 두 그래프 중 더 가능성이 높은 그래프를 찾는 방법은, 앞에서 두 동전에 대한 가설 중 어느 가설이 타당한 지를 평가할 때 보다 더 복잡하지만 기본적으로 같다. 즉 실제 동전 던지기 결과를 이용하여 가설의 타당성을 확률적으로 평가하듯, 관찰된 결과가 두 그래프 중 어느 쪽에서 비롯될 가능성이 높은 지를 상대적 비율로 평가하는 것이다. 이를 수식으로 표현하면 다음과 같다.

$$\log\frac{P(\text{그래프1}|\text{자료})}{P(\text{그래프0}|\text{자료})}=\log\frac{P(\text{자료}|\text{그래프1})}{P(\text{자료}|\text{그래프0})}+\log\frac{P(\text{그래프1})}{P(\text{그래프0})}$$

그림 1에서 C와 E간에 연결이 있을 수도 있고 없을 수도 있기 때문에 이들의 비가 같다고 보면 두 번째 항은 0이 된다. 따라서 첫 번째 항만 남게 된다. Griffiths와 Tenenbaum(2005)은 이 항을 인과 지지로 명명하였다. 이는 위에서 설명한 베이스 인자에 로그함수를 취한 형태이다.

$$\text{인과지지} = \log\frac{P(\text{그래프1})}{P(\text{그래프0})} \quad \cdots\cdots (14)$$

식 (14)에서 분모의 확률은 다음 식을 통해 구해질 수 있다:

$$\text{P(D}|\text{그래프 0)} = \int P_0\,(\text{D}|w_0,\ \text{그래프 0})*P(w_0|\text{그래프 0})dw_0 \quad \cdots\cdots (15)$$

여기서 통계전문가에게는 쉽지만 심리학자에게는 매우 복잡한 계산을 하면, 식 (14)는 베타분포로 변형시키고 이로부터 값을 계산할 수 있다(이 변형 과정에 대한 상세한 설명은 Griffiths, Kemp, & Tenenbaum, 2008의 66~67쪽을 참조하시오).

$$\text{P(D}|\text{그래프 0)} = \int w_0\text{N}(e^+)\ (1-w_0)w_0\text{N}(e^-)dw_0 = Beta(\text{N}(e^+)+1,\ \text{N}(e^-)+1).$$

식 (14)에서 분자의 경우는 C가 추가되어 B와 C가 통합되어져야 하기 때문에 더 복잡해져 분모에서처럼 식을 풀어 적분값을 구할 수 없다.

$$P(D|\text{그래프 } 1) = \int\int P_1(D|w_0, w_1, \text{그래프 } 1) * R(w_0, w_1|\text{그래프 } 1) dw_0 dw_1$$

그 대신 다음과 같이 Monte Carlo 방법을 사용하여 근사값을 구할 수 있다:

$$P(D|\text{그래프 } 1) = \frac{1}{m} \sum P_1(D|w_0, w_1, \text{그래프 } 1),$$

여기서 m은 표본의 수이고, $P_1(D|w_0, w_1, \text{그래프 } 1) = \prod P_1(e|c, b^+; w_0, w_1,)^{N(e,c)}$ 로 구할 수 있다. 이 공식을 적용하여 실제 계산이 어떻게 이루어지는 지를 살펴보자. B와 C를 생성적 원인이라 하자. 그리고 모사할 자료는 분할표 상에서 C가 있는 24회의 관찰 중 16회 그리고 C가 없는 24회의 관찰 중 2회에서 결과가 일어난 경우를 사용하기로 하자. 이 자료는 표 2의 L M N O에 각각 16, 2, 8, 그리고 22를 넣은 것과 같다. 그림 1에서 그래프 1이 선택되면, L, M, N, 그리고 O의 확률은 각각, $1-(1-w_0)*(1-w_1)$, w_0, $(1-w_0)*(1-w_1)$, 그리고 $(1-w_0)$ 이 된다.

이제 Matlab이나 다른 프로그램을 이용하여 균등분포로 가정된, w_0, w_1값을 0에서 1사이에서 무선적으로 m쌍을 추출한다. 예를 들어, 0.3435 0.6129 ..... 와 같은 쌍이 만들어졌다고 하자. w_0 = 0.3435, w_1 = 0.6129를 대입하면

$$\begin{aligned} & 1 - (1-w_0)*(1-w_1) \\ = \; & 1-(1-0.3435)*(1-0.6129) \\ = \; & 0.7459 \text{ 가 된다.} \end{aligned}$$

C가 있을 때 결과가 일어날 확률이 0.7459인데 이런 일이 16번 일어날 확률은 0.7459^{16} 이 되고 이런 식으로 다른 경우에 대해서도 확률을 구한 다음 모두 곱해 주면 각 표본으로부터 자료를 예측할 확률을 구할 수 있다. 이 확률값은 매우 작기 때문에 보통 로그를 이용하여 계산된다. 이렇게 구해진 m쌍의 로그값을 모두 더한 후, 사용된 표본의 수인 m으로 나누면 그래프 1이 참일 때 관찰된 결과를 얻게 되는 확률이 된다. 이 확률을 다시 분석적으로 계산한 베타 값으로 나눈 값이 바로 인과 지지값이다.

인과 지지값이 충분히 커, 그래프 1이 자료를 더 잘 설명한다는 판단이 내려지게 되면, w_0와 w_1을 추정할 수 있는데, 모수 추정을 위해 베이스 접근법에서 널리 사용되는 방법은 최대우도 추정법이다. 그림 1의 그래프1의 경우, $P(D|w_0, w_1)$이 최대값이 되도록 w_0와 w_1을 선택하는데, 이 값은 실제 관찰 확률과 모델의 확률이 같도록 하는 w_0, w_1이다. 즉,

$$P_1(e+|c^+, b^+; w_0, w_1) = P(e^+|c^+), \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (16)$$

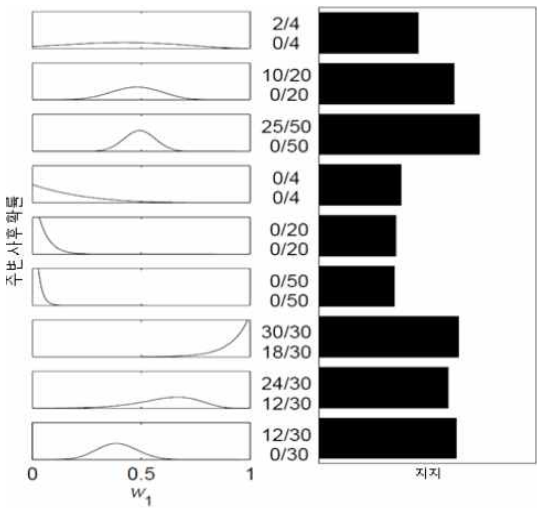
$$P_1(e+|c-,\ b^+;\ w_0,\ w_1) = P(e^+|c^-) \quad \cdots\cdots (17)$$

ΔP값은 그래프 1에서 선형적 모수화를 가정할 때의 최대 우도추정치인데, w_0는 P(e+|c-), w_1은 P(e+|c+)로 추정된다. 힘은 noisy-OR 모수화를 가정할 때의 최대우도추정치인데, w_0는 P(e+|c-), w_1은

$$\frac{\Delta P_i}{1 - P(e^+|c^-)}$$

로 추정된다. 요컨대, ΔP와 힘 이론은 그래프 1과 그래프 0중 어느 것이 맞는 지에 대한 구조 판단이 생략된 채, 그래프 1이 맞다는 가정하에서의 최대 우도추정치를 찾아내는 모형들이라는 것이다.

몇 몇 자료세트에 대해, 사후 주변확률 분포 즉 P(d|그래프 1)의 분포를 w_1의 함수로 나타내면 그림 3과 같다. 참고로 각 그래프의 최대값이 위에서 언급한 최대 우도추정치, 즉 그 조건에서의 힘 추청치이다. 처음 3개의 그래프는 ΔP나 힘이 동일하더라도 관찰수가 증가하면, 지지값이 커지는 패턴을 보여주며, 그 다음 3개의 그래프는 w_1이 다 0인 경우에는 관찰수가 증가하더라도 지지값에 변화가 없음을 보여준다. 마지막 세 개의 그래프는 ΔP나 힘이 점차 감소하더라도 지지값은 이에 비례하여 변화되지 않는다는 것을 보여준다. 요컨대 그림 3은 인과지지 모형에서의 예측이 ΔP나 힘 이론에서의 예측과 명확히 다르다는 것을 보여준다.



(그림 3) 아홉 개의 서로 다른 분할표 자료에 대한 인과지지 값과 w_1에 대한 주변 사후확률 분포.
(Griffiths & Tenenbaum(2005)의 그림 4를 허락을 받고 게재함. 그래프에 대한 설명은 본문 참조 요망).

이처럼 인과 귀납에서 베이스 접근을 사용하게 되면 힘 이론이나 ΔP 모형에서의 계산이 조건부 확률을 통해 이루어지기 때문에, 표본의 크기로 인한 효과를 고려하지 못한다는 점을 극복할 수 있다. 즉, 인과지지 모형에서는 표 3의 아래 네 줄의 결과를 설명할 수 있다는 것이다. Griffiths와 Tenenbaum은 이외에도 인과지지 모형의 장점으로, 분할표의 네 부분의 정보가 다 주어지지 않은 경우에도 인과판단을 할 수 있게 하며, 빈도 정보가 아니라 연속적인 시간 간격 동안 결과가 몇 번 발생했는지를 종속 측정치로 삼은 자료에도 적용이 가능함을 언급하였다.

지지 모형에 대한 비판과 소수의 강한 원인을 찾고자 하는 모형

Cheng과 동료들(Holyoak & Cheng, 2011; Lu et al., 2008)은 인과성 연구에 베이스 접근법을 적용함으로써, 불확실성 문제를 다룰 수 있게 되었고 구조와 강도의 문제를 구분할 수 있게 된 것을 중요한 기여로 간주한다. 하지만 세부적인 내용에 있어서는 지지모형이 여러 문제가 있다고 비판한다. 우선 구조와 강도의 구분이 중요한 것은 사실이지만, 이 구분에 앞서 검토되어야 할 부분은 인과성에 대한 가정임을 지적한다. 앞서 본 것처럼 인과성에 대한 두 대립적 입장은 통계적 견해와 인과적 견해이다. 베이스 접근은 이 두 견해와 무관하게 도입될 수 있다. 그렇지만, 힘 이론에서처럼 인과성을 전제하는지 아니면 ΔP 모형에서처럼 통계적 규칙성을 전제하는지에 따라 구조와 강도추정이 영향을 받게 된다는 것이다. 예를 들면, 한 결과에 영향을 주는 원인이 둘 이상일 때 이들의 영향을 통합하는 함수는 가산적 혹은 noisy-logical 함수일 수 있다. 통계적 견해를 취하는 ΔP 모형에서는 가산적 함수를 사용하지만, 인과적 견해에 따르는 힘 이론에서는 noisy-logical 함수를 사용한다. 지지모형에서는 noisy-logical 함수가 사용되기 때문에, 적어도 이 점에 있어서는 인과적 견해에 따르고 있다고 할 수 있다.

인과성 가정의 중요성과 더불어 Cheng과 동료들이 베이스 접근에서 주목한 부분은 사전 분포의 특성이다. 지지 모형에서는 변수들에 대해 균등사전 분포를 가정하였지만, Lu 등(2008)은 가능한 여러 변인들 중 원인이 될 만한 변인은 소수이지만 강력한 영향을 주는 변수를 선호한다고 가정하였다. 소수의 강력한(Sparse and Strong: SS) 원인을 선호하는 이 일반적 사전 분포를 사용하면, 자료가 사전분포와 일치할 경우, 인과적 지식 습득이 빨라지게 해준다. 이제 사전분포와 위에서 본 통합 함수라는 두 변인을 서로 독립적으로 교차시켜 서로 다른 4개의 모델을 도출할 수 있다. Lu 등(2008)은 이 4개의 모델을 실제 사람들의 수행과 비교하는 일련의 연구를 수행하였다.

Lu 등(2008)은 실험 1에서는 인과성의 방향, 힘, 기저율, 그리고 표본의 크기를 조작한 실험을 수행하였다. 그 결과 인과 판단은 두 인과적 방향에 걸쳐 비대칭적이었는데, 억제적 원인일 때 더 크게 반응하였다. 표본의 크기는 인과 판단에 큰 영향을 끼치지 않았고, 그 대신 기저율과 힘에 크게 영향을 받았다. 네 개의 다른 베이스 모형 중 사람으로부터 얻어진 결과와 가장 유사

한 패턴을 보인 것은, 일반적 사전 분포와 noisy-logical 함수를 사용한 모델이었다. 인과 강도를 판단하게 한 실험 1과 달리, 실험 3에서는 힘, 기저율(base-rate), 표본 크기와 인과적 방향을 조작한 다음, 인과 구조를 판단하도록 하였다. 알레르기 약에 함유된 광물질로 인해, 환자에게 두통이 초래되거나(생성적 조건) 아니면 두통이 사라지는(억제적 조건) 자료를 그림으로 제시하였다. 그리고 그 광물질이 두통을 초래할(혹은 완화시킬) 가능성이 얼마나 되는지를 0에서 100점 척도 상에서 평정하도록 하였다.

실험 결과 인과적 연결은 힘에 비례하여 높아졌고, 강도 추정에서와는 달리 표본의 크기가 커질수록 높게 평가되었다. 인과적 방향성에 따른 차이도 관찰되지 않았다. 이 결과는 강도 판단과 구조 판단이 다른 과정에 의해 처리될 가능성을 시사한다. 따라서 이 둘을 구분하지 않은 채 지지로 강도 판단을 추정하려 한 Griffiths와 Tenenbaum(2005)의 시도는 재검토될 필요가 있다고 지적한다. 구조 판단에 대한 사람들의 수행을 4개의 모형과 비교했을 때, 강도 판단에서와 마찬가지로 일반적 사전 분포와 noisy-logical 함수를 사용한 모델의 합치도가 가장 높았다. 이상의 결과는 수렴적으로 소수의 강력한 원인을 가정하는 사전분포를 취하는 베이스 모형이 사람의 인과 학습 과정을 잘 포착할 수 있음을 보여준다.

Lu 등(2008)의 비판에 대한 반응으로 Griffiths와 Tenenbaum(2009)은 더 강력한 베이스 기법을 도입한 이론기반 인과귀납 모형을 제시하였다. 이 모형은, 모수 자체에 대해 분포를 가정하는 위계적 베이스 기법을 적용하여, 적은 수의 제한된 자료에 기반하여 변수, 물체, 사건들 간의 관찰가능한 관계를 생성하는 보이지 않는 힘이나 기제를 찾아내는 일반적 귀납 모형이라 할 수 있다. 이들은 사람들이 여러 영역에 대해 갖고 있는 직관적 이론들(intuitive theories)에서 관찰되는 특징으로, 존재론, 가능한 관계, 그리고 통합 함수의 형식이라는 세 범주로 구분하였다. 이 세 구성요소로 이루어진 인과 이론은, 마치 문법이 통사적 구조를 제약하고 이 제약 하에 문장이 만들어지듯, 가능한 인과적 구조의 범위를 제한하고 그 제한 속에서 관찰 결과를 산출하도록 한다.

구체적으로 이론으로부터 인과 모형을 만들기 위해서는 다음의 세 과정이 요구된다: 다루려는 대상과 그 대상들의 술어에 기반하여 변수 혹은 마디를 생성해낸 후, 확률적 절차를 사용하여 이 변수(마디)들을 연결하여 구조를 만든 다음, 각 변수를 모수로 나타내는 방법을 만들어낸다. 앞서 보았던 분할표에 제시된 자료를 바탕으로 한 인과 추리의 경우 존재론은 각각 두 개의 원인과 결과이고, 가능한 관계는 연결이 없는 경우를 포함하여 모두 연결되는 총 16개의 관계이다. 두 원인들이 결과에 영향을 줄 때 이들을 통합하는 함수의 유형은 noisy-logicals나 선형 함수가 가능하다. 이제 이와 같은 이론이 명확하게 규정되면 나머지 추리는 베이스 연산에 의해 이루어진다는 것이다.

구조와 강도의 구분이 다른 과정에 의해 처리된다는 Lu 등(2008)의 실험 결과에 대해, 이들은 실험참여자들의 개인차가 크다는 점을 지적하면서, 과제에 대한 해석과 과제 수행에 사용되는

전략상의 차이로 설명될 수 있다고 반박한다. 아울러 인과 추리와 관련된 실험 연구 결과들을 재검토해 보면, 전반적으로 지지 모형에 의해 가장 잘 설명될 수 있다고 주장한다.

이론기반 인과 추리 모형은 작은 표본에 근거하더라도, 사전 지식을 활용함으로써 성공적으로 인과 추리를 해낼 수 있다. Gopnik 등(2001)은 4살짜리 아이들에게, 어떤 물건을 이 위에 올려놓으면 불이 켜지고 음악이 연주되도록 만들어진 상자인, blicket 탐지기를 보여준 다음, 여러 개의 블록을 제시하고 이 중 어떤 것이 blicket인지를 찾아내도록 하였다. 아이들은 한 개 혹은 두 세 개만 보고도 어떤 블록이 blicket인지를 상당히 정확한 판단을 할 수 있었다. Griffiths와 Tenebaum(2009)은 아이들이 기계의 경우 결정론적으로 작동한다는 지식이 있기 때문에 이런 빠른 학습이 가능하다고 설명한다.

이 모형이 기존의 실험 결과를 설명하는데 있어 Cheng 등의 힘 이론이나 여기에 일반적 사전 확률을 가정하는 모델에 비해 적어도 적합도 면에서는 더 우수하다는 점은 분명해 보인다(Lucas & Griffiths, 2010). 하지만 힘 이론에서는 최소한의 가정과 변수가 고려되는데 반해, 베이스 접근에서는 일반적으로 많은 모수가 가정된다는 점이 지적될 필요가 있다. 이로 인해 자료를 더 정확한 서술할 수는 있지만, 새로운 현상을 예측하거나 이전의 발견을 일반화하는 일을 얼마나 더 잘 하는 지는 검증될 문제로 남아 있다. 실제로 이와 관련하여 다음 절에서 보게 될 여러 결과들이 베이스 접근을 통해 얼마나 잘 포착될 수 있을 지는 현재로서 알기 어렵다. 다만 힘 이론이 베이스 접근법을 정교화 하는 데 있어 중요한 기여를 하였고, 앞으로도 이들 간에 경합이 벌어질 것이라는 점만큼은 분명해 보인다.

인과적 불변성을 중심으로 한 연구들

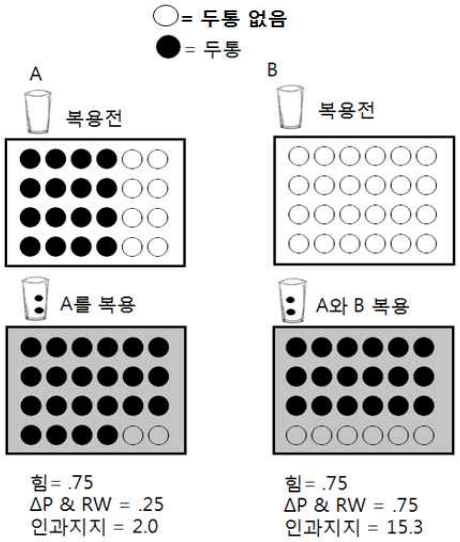
인과 학습을 포함한 학습의 목표는 기본적으로 일반화 혹은 전이다. 즉, 한 맥락에서 배운 내용을 시간적 혹은 공간적으로 다르지만 적절한 맥락에서 활용하는 것을 목표로 한다. 인과적 지식을 일반화하려면, 어느 한 맥락에서 인과적 관계를 정확히 표상하는 것도 중요하지만 맥락에 걸쳐 일반화할 수 있는 방식으로 표상하는 것이 더 중요할 수 있다. 일반화를 촉진하는 한 방법은 인과적 영향력의 불변성을 가정하는 것이다. 한 원인이 결과에 주는 영향력은 맥락에 따라 달라지지 않고 동일하다고 가정하는 것이다. 물론 실제 장면에서는 맥락에 상관없이 불변적으로 작동하는 원인을 찾아내기 어렵다. 그렇지만 뉴턴이나 아인슈타인같은 과학자들이 가정하듯 세상을 단순하게 이해하려면 불변성 가정이 중요한 역할을 하게 된다.

실제로 과학에서는 만일 맥락에 따라 불변성이 위배되는 경우가 발견되면, 결과에 영향을 주는 또 다른 원인을 찾아내는 방식으로 발전해왔다. 예를 들어 땅에서는 100도에서 끓는데 산 위에서는 더 낮은 온도에서 끓는 이유는 기압의 차이로 설명된다. 이런 차이는 비등점에 대한 개

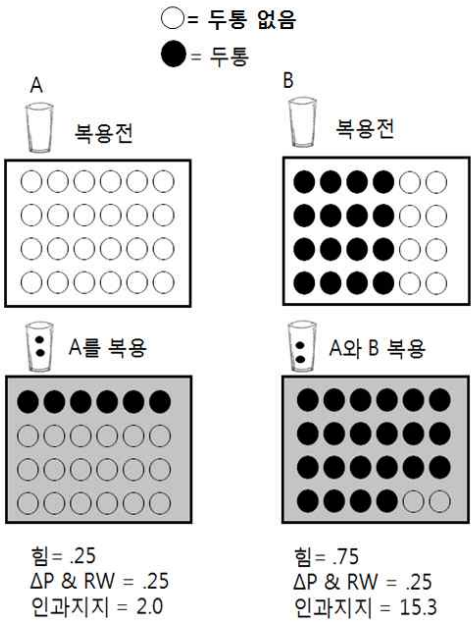
념을 명확히 하는데 도움을 주었다. 물론 불변성을 확장하는 방향의 발전도 쉽게 찾아볼 수 있다. 뉴턴의 법칙은 그 좋은 예이다. 뉴턴이 천체의 움직임에까지 이 법칙을 확장하기 전에는 지구상에서의 물체의 움직임과 천체에서의 움직임이 다를 것이라고 가정되었다. 하지만 이 둘이 모두 같은 법칙에 따른다는 것이 밝혀지면서 "만유" 인력 법칙으로 인정되었다. 만일 맥락에 따라 달라졌다면 아마도 또 다른 원인을 찾거나 함수의 형식을 바꿔야 했을 것이다. 요컨대 과학적 탐구의 궁극적인 목표는 현상 배후에서 그런 현상이 일어나도록 하는 불변의 원리나 기제를 찾아내는 것인데, 이는 곧 인과적 불변성을 추구하는 과정이라 할 수 있다.

이제 인과적 독립성과 함께 불변성을 가정한 상황에서 힘 이론을 다른 모형들, 즉 RWM, *ΔP*, 그리고 인과지지 모형과 비교해 볼 수 있다. Liljeholm과 Cheng(2007)은 이를 위해 2개의 실험을 수행하였는데, 각 실험은 두 개의 다른 맥락에서 이루어졌다. 첫 번째 실험에서는 두 맥락에 걸쳐 힘은 동일하게 유지하지만, *ΔP*, RWR에 의한 연합강도, 그리고 인과 지지값을 변화시키는 조건과 반대로 *ΔP*, RWR에 의한 연합강도, 그리고 인과 지지값와 지지를 일정하게 하고 힘만을 변화시키는 조건이 비교되었다.

그림 4에서처럼 힘이 유지되는 조건의 경우, 첫 번째 맥락에서는 약을 복용하지 않은 24명 중 18명이 두통이 있었고, A를 복용하면 그 중 6명이 추가로 두통을 호소하였다; 두 번째 맥락에서는 약을 복용하지 않은 24명 모두가 두통이 없었는데, 이들 중 A와 B를 모두 복용하자 22명이 두통을 호소하였다. 이 두 맥락에 걸쳐 두통에 대한 약 A의 인과적 힘은 0.75로 일정하지만, *ΔP*와 RWR에 의한 연합강도는 0.25에서 0.75로, 그리고 인과지지 값은 2.0에서 15.3으로 변화된다.



(그림 4) 실험참여자에게 두 개의 다른 맥락 A와 B에서 제시된 실험 결과로 두 맥락에 걸쳐 힘은 동일하게 유지되지만 ΔP와 인과 지지값은 변화되었음.(Liljenholm & Cheng, 2007 figure 1을 변형하였음).



(그림 5) 실험참여자에게 두 개의 다른 맥락 A와 B에서 제시된 실험 결과로 두 맥락에 걸쳐 힘이 변화되었지만 ΔP와 인과 지지값은 동일하였음. (Liljenholm & Cheng, 2007 figure 2를 변형하였음).

힘을 조작한 그림 5의 경우, 첫 번째 맥락에서는 약을 복용하지 않은 24명이 모두 두통이 없었지만, A를 복용하면 그 중 6명이 두통을 호소하였다; 두 번째 맥락에서는 약을 복용하지 않은 24명중 18명이 두통을 호소하는데, 이들 중 A와 B를 모두 복용하자 22명이 두통을 호소하였다. 이 두 맥락에 걸쳐 인과적 힘, ΔP, 그리고 인과지지 값을 비교하면, 힘은 0.25에서 0.75로 변하지만, RWR에 의한 연합강도, ΔP, 그리고 인과지지 값에는 변화가 없다.

이런 상황에서 실험참여자의 과제는 약 B가 두통을 일으키는 지를 예/아니오로 답하게 하였다. 그 결과 맥락에 따라 힘이 일정한 조건에서는, 25명 중 18명의 실험참여자가 B가 두통을 일으키지 않는다고 반응한 데 반해, 힘이 변한 조건에서 그렇게 반응한 실험참여자 수는 25명 중 5명에 불과하였다. 힘이 변하는 조건에서 B가 두통을 일으키지 않는다고 반응하는 이유는 두 맥락에 걸쳐 A의 힘이 0.75로 동일하면서 두통 발생률을 모두 설명할 수 있어, 사실상 B가 두통에 기여하는 바가 없기 때문이다. 그렇지만 전자의 조건에서는 A의 힘이 0.25인데 A와 B를 동시에 복용했을 때의 힘이 0.75로 높아지기 때문에, B가 두통에 기여하는 부분이 크다. 따라서 B가 두통에 기여한다고 반응한 사람의 비율이 높다. 요컨대, 맥락에 상관없이 인과적 힘이 불변적으로 적용된다는 가정이 위배될 때 새로운 원인에 대한 힘이 추정된다.

이 연구결과는 인과추리가 인과적 힘을 중심으로 이루어짐을 시사하지만, 통계학과 인공 지능 연구에서는 아직 인과적 힘이 적극적으로 받아들여지지 않고 있다. 이들은 여전히 경험주의에

기반을 두고 객관적으로 보이는 자료를 분석하여 이로부터 규칙성을 도출해내고자 한다. 지금까지의 논의에 비추에 보면 통계적 전통과 인과적 전통 중 통계적 전통으로 치우쳐왔다고 할 수 있다. 하지만 Cheng과 동료들은 최근 이런 접근은 특정한 상황에서 모순을 일으킬 수 있음을 보여주었다.

Cheng, Liljeholm, 그리고 Sandhofer(2013)는 경험적 접근과 인과적 접근의 차이를 가장 극명하게 드러내는 연구 결과를 제시한다. 유치원생에게 농부와 동물원 사육사 형제에 대한 다음과 같은 이야기를 들려주었다. 이 형제는 농장과 동물원에 있는 동물의 얼굴에 나는 반점을 없애려고 곡식과 나뭇잎을 먹였다. 농장에는 10마리의 동물이 있었는데 이들 중 9마리에 반점이 있었다. 농장의 모든 동물에게 곡식을 먹이자 반점이 있던 9마리 중 6마리에게만 반점이 남아있었다. 이 형제는 동물원에 있는 10마리의 동물들에게 곡식과 나뭇잎을 둘 다 먹였다. 그러자 반점이 있던 4마리 중 한 마리에게만 반점이 남아있었다. 질문은 어떤 먹이가 농장과 동물원에 있는 동물의 얼굴에 있는 반점을 더 잘 없앨까였다.

통계적 접근에서는 모두 곡물을 선택하도록 한다. 그 이유는 곡식의 ΔP = 3/10이고 잎사귀와 곡식의 ΔP도 3/10으로 같으므로, 곡식은 0.3이 되고 잎사귀의 영향은 0이 되기 때문이다. 로지스틱 회귀분석을 사용하여 자료를 변환하더라도 이런 결론을 달라지지 않는다. 그런데 힘에 초점을 둔 분석을 이와 달라진다. 먼저 농장에서 먹인 곡식의 인과적 힘은 1/3이다. 동물원에서는 동물들에게 곡물과 잎사귀를 함께 먹였을 때의 힘은 3/4이다. 여기서 인과적 불변성을 고려하여 잎사귀만의 인과적 힘을 p라고 가정하면, 4*1/3 + 4*p - 4*p*1/3 = 3이 된다. 이로부터 잎사귀만의 힘을 계산하면 5/8가 된다. 따라서 힘 이론에 따르면 잎사귀를 선호해야 한다. 유치원생을 대상으로 한 실제 실험 결과 아이들은 힘 이론과 일치되게 잎사귀를 선호하였다. 이 차이는 인과적 불변성을 가정하는 힘 이론이 아닌 다른 이론으로 설명되기 어려운 결과이다.

인과적 불변성이란 인과적 기제가 맥락에 걸쳐 동일한 방식으로 작동함을 의미한다. 3절 말미에서 논의된 것처럼, 복수의 원인이 있을 때 이들의 영향을 통합하는 함수가, 변수가 이산적인지 혹은 연속적인 지에 따라, 달라지는 이유도 인과적 불변성 추구의 한 결과로 설명될 수 있다. 앞서 본 베이스 접근법에서는 사용자가 필요한 정보를 정리하여 제공할 수 있지만 우리의 실제 삶에서는 누가 우리를 위해 필요한 정보를 정리해주지 않는다. 그럼에도 중요한 변인을 찾아내고 그 변인으로부터의 인과적 힘이나 구조를 비교적 정확하게 일반화할 수 있다. 이처럼 인과 추론이 가능한 것은 우리의 인지체계가 인과적 불변성 추구는 방향으로 작동하기 때문이다.

힘-확률대비 이론의 의의와 시사점

지금까지 소개한 힘 이론은 현재에도 계속 발전하고 있기 때문에 그 의의를 평가하기에는 여

전히 시기상조라 할 수 있다. 하지만 인과 추리에 대한 포괄적 이론이라는 점은, 후속 평가와 상관없이, 분명하다. 이론 자체는 원인과 결과 간 조합으로부터 구성될 수 있는 분할표를 바탕으로, 힘을 추정해내는 수리적 모형이지만, 이 힘은 단순한 계산 결과가 아니라 인과적 불변성을 추구해가는 과정의 산물이다. 인과적 불변성은 원인은 맥락에 상관없이 결과에 동일하게 작용하는 것으로 가정되는데, 이 가정으로부터 인과적 발견은 자연스럽게 활용과 통합된다. 발견과 활용을 통합하는 일반화는 개념적 표상을 통해 촉진된다. 실제로 인과적 발견은 개념형성을 촉진할 수 있고(예, Lien & Cheng, 2000; Rehder, 2006), 거꾸로 사전 개념이 있는 경우에는 그 개념을 매개로 인과적 발견이 촉진될 수 있다(예, Griffiths & Tenenbaum, 2009).

어떤 원인이 맥락에 따라 다르게 작용하면 현재의 인과적 구조와 강도에 문제가 있음을 시사한다. 불변성이 유지되지 않는 이런 상황은 원인과 결과 간의 관계에 대한 현재의 가설이 잘못되었음을 알려주므로 새로운 가설을 생성하도록 한다. 그 결과 인과귀납, 범주형성, 그리고 가설변경은 서로 밀접하게 상호 작용할 수 밖에 없게 된다(Cheng & Buhner, 2012). 따라서 인과성 연구는 인과성 연구로 끝나는 것이 아니라 범주화나 가설 수정 등과 같은 인지과정과 끊임없이 영향을 주고받는 역동적 과정이 된다. 힘-확률 이론의 최근 연구가 가설 수정과 범주화 연구로 확장되는 것은 바로 이런 이유 때문이다(Carroll & Cheng, 2010).

힘-확률이론의 포괄성은 인과적 불변성 개념으로 기존의 다른 이론들이 설명하지 못하는 현상들을 하나의 틀 속으로 묶어준다. 기존의 이론들이 잘 설명하지 못하는 현상은 둘 이상의 원인이 이산적 결과 변수에 주는 영향을 통합해야 하는 상황과 관련이 있다. 바로 이전의 5절에서 논의된 여러 발견들이 그런 예들이다. 기존의 통계 모형에 따르면 차이가 없는 두 원인 중 아이들은 힘 값이 더 높은 원인을 정확히 선택한다. 이런 차이는 단지 특정 실험 결과를 해석하지 못한다는 데서 끝나지 않고, 실제 장면에서 심각한 문제를 일으킬 수 있다. Cheng 등(2013)이 제시하는 예는 건강한 식생활에 대한 잘못된 지침이다. 이 지침은 실제 자료를 바탕으로 지방을 통제한 상황에서 설탕의 섭취가 사망에 영향을 주지 않는다는 분석 결과에 근거한다. 그런데 앞에서 본 곡물과 잎사귀 예에서처럼 규칙성에 따르는 방식이 아니라 인과적 방식으로 분석을 했다면, 결과가 반대로 나왔을 수 있었다. 그런데 연합적 방식을 사용했기 때문에, 지방이 건강을 악화시키는 주범으로 지적되었고 설탕을 빠지게 되었다. 그 결과 지방은 줄이고 그 대신 맛을 내기 위해 설탕이 가미된 식품들이 개발되었고, 결과적으로 비만을 포함한 여러 성인병의 발병을 높였다. 최근의 연구는 뒤늦게 설탕 섭취의 심각성을 알려주고 있다(예, Lustig, Schmidt, & Brindis, 2012). 이산적 결과를 대상으로 한 잘못된 회귀분석 사례는 이 외에도 얼마든지 가능할 수 있는 만큼 해당사례들을 찾아내고 이들을 noisy-OR 방식으로 다시 분석해볼 필요가 있다.

차폐와 관련된 여러 실험 결과들도 재검토될 부분이 많다. 위에서 본 것처럼 결과 변수가 연속 변수인지 아니면 이산 변수인지에 따라 둘 이상의 원인이 통합될 때 변수 유형에 따라 다른 통합 함수가 사용될 가능성이 높다. Lu 등(2016)은 순차적으로 시행이 제시되는 상황에서의 인과

적 학습과정을 베이스 접근으로 모사하였다. 이전의 연구에서는 실험참여자의 기억 부담을 최소화하기 위해 관련 정보가 요약표(summary table) 형식으로 제공되었다. 하지만 언어가 없는 동물들의 경우 이런 자극 제시 방식을 사용할 수 없고, 또 사람들도 실제 상황에서는 사례들이 순차적으로 제시되는 가운데 인과추리를 한다. 따라서 이 과정이 어떻게 이루어지는 지를 밝히는 것은 이론적으로 중요하다. 이 연구를 통해 Lu 등은 동물들을 대상으로 할 때는 비교적 일관적으로 차폐 효과가 관찰되지만, 사람들은 약하거나 전혀 일어나지 않는 등 다양한 패턴을 보인다는 것을 확인하였다. 이들은 연속 변인일 때는 가산적 함수를 이산 변수일 때는 noisy-OR 함수를 사용한다고 가정했는데, 그 근거를 제시하지는 않았다. 그렇지만 사람들은 인과적 불변성을 추구하고 이를 위해 변수의 유형에 맞게 선험적으로 적절한 함수를 사용한다는 가정이 맞는다면 특히 이산 변수를 사용할 경우 차폐가 일어나지 않거나 약한 결과를 잘 설명할 수 있다.

사람이 아니라 쥐를 대상으로 한 일련의 연구에서 차폐를 관찰하지 못한 Maes 등(2016)의 연구도, 쥐들이 전기 충격이나 목마름을 해석하는 방식에 따른 결과로 볼 수 있다. 표 1에서 소개된 여러 개념들 예를 들어 최대치, 가산성, 그리고 사전 훈련 등이 체계적으로 영향을 주는데, Maes 등(2016)의 연구에서 사용된 쥐들이 이전에 노출된 여러 통증, 배고픔, 그리고 목마름을 다른 강도로 경험하면서, 비록 연속 변수를 염두에 두고 무조건 자극을 가하더라도 이를 이산 변수로 해석할 여지가 있다는 것이다.

나가며

지난 30년의 발전을 통해 힘 이론은 심리학에서는 물론 인접 분야에서 인과 추리에 대한 논의를 이끌어가는 견인차 역할을 해왔다고 할 수 있다. 초기에는 ΔP 모형이나 RWR에 따르는 예측과 힘 값에 근거한 예측과 실제 사람들의 수행간의 비교가 이루어졌다. 그 다음에는 인과적 구조를 고려하는 베이스 모형과의 비교를 통해 발전하였다. 이런 연구들을 통해 힘 값을 중심으로 한 수행 예측이 가장 우수한 성과를 보이지 않았을 때가 있지만, 수행과의 일치가 높은 모형들이 여러 개의 모수를 사용한다는 점이 지적될 필요가 있다. 오히려 최소의 모수를 사용하면서 이처럼 높은 예측력을 보이는 것이 더 놀랍다고 할 수 있다.

베이스 접근은 제한된 자료를 바탕으로 한 규범적 추론을 구현하는 한편, 인과 추리의 수행에서 사람과의 비교를 통해 사전확률로 표현될 수 있는 모종의 전제들을 밝히는 데 유용한 도구를 제공하였다. 그렇지만 이 접근에서는 계산에 필요한 자료들이 연구자에 의해 제공된다는 점에서 사람들이 직면한 상황과는 다르다. 게다가 사람들은 베이스 접근법에서 널리 사용되는 Markov 가정, 즉 한 상태는 직전의 상태에만 영향을 받지 그 이전의 상태에는 영향을 받지 않는다는 가정을 따르지 않는다는 것은 잘 알려져 있다(예, Meder, Hagmayer, & Waldmann, 2008;

Rottman & Hastie, 2016). 이런 상황에서 인과적 불변성은 인과 추론의 엔진으로 다음과 같은 여러 기능을 수행한다. 우선, 불변성을 찾아내는 것 자체가 탐구의 목적이고, 한 맥락에서 어떤 원인의 힘이 추정되면 기정치처럼 다른 맥락에 적용되며, 이 기대가 깨지면 기존의 가설이 수정되어야 함을 알려준다. 요컨대 인과적 불변성을 전제하지 않으면 탐구의 동력은 물론 그 종착점을 알 수 없게 된다.

지금까지 힘 이론을 중심으로 살펴본 인과추론에 대한 연구의 흐름은 심리학적 탐구에서 새로운 이론의 중요성을 확인시켜준다. 이 새로운 이론은 관련된 과거의 여러 이론과 논쟁을 한쪽으로 밀쳐두는 것이 아니라 왜 그런 이론과 논쟁이 생길 수밖에 없는 지를 이해하게 해 줄 수 있어야 한다. 여기서 끝나는 것이 아니라 새로운 주장을 통해 활발한 후속 연구를 촉발시키면서 그 분야의 발전을 이끌 수 있어야 한다.

힘 이론의 발전과정 자체가 인과적 불변성에 의해 전개되어 왔다고 할 수 있다. 필연성이론과 규칙성 이론을 통합하고, 불확실성을 다룰 수 있는 베이스 접근법을 통해 인과 추리를 위한 소수의 강력한 원인을 선호하는 사전 분포를 찾아냈고, 인과적 불변성을 통해 가설 검증의 원동력을 찾아냈다. 이런 발전은 심리학적 통찰이 연구 방법을 넘어서서 한 분야의 발전에 얼마만큼 큰 기여를 할 수 있는 지를 보여준다. 철학적 논의, 심리학적 탐구, 그리고 계산론적 모형의 정교화 등에서 볼 수 있듯, 그 영향이 어느 한 분야에 국한되지 않고 여러 영역에 걸쳐 있다. 사실 이 분야의 최근 연구들을 보면, 두 이상의 영역 전문가들의 공동 연구가 그 어느 때보다 활발하다.

힘 이론이, 인과 추리에 대한 이해를 깊게 하는 동시에, 인지과학자들에게 주는 메시지를 생각해 본다면, 명확한 연구 주제를 중심으로 한, 관련 영역의 전문가들의 협업의 중요성이다. 때로 방법론이 새로운 발견을 촉진하기도 한다. 하지만 힘 이론이 보여주는 것은 통찰력 있는 이론이 얼마나 중요한 지를 보여준다. 뇌 영상 촬영기술이나 고급 수리 모형 메타 분석 혹은 구조방정식 모형 등의 방법론은 익히는 것 자체가 큰 지적 도전이 될 뿐만 아니라, 특정 방법론이 적용되는 영역을 찾아다니며 연구를 수행하게 된다. 이론이 중심이 될 때, 다양한 방법으로 현상에 접근할 수 있게 되고, 특정 영역에 대해 더 깊이 있는 탐구를 수행할 가능성을 높여준다. 요컨대, 심리 현상에 대한 통찰력 있는 이론이 매우 중요하다는 것이다. 최근 우리 사회에서 화두로 떠오르는 융합(convergence)은 사실 인지과학이 오래 전부터 추구해온 연구 전략이었다. 그런데 이런 융합이 제대로 이루어지려면 명확한 문제제기에서 출발해야 한다. 그 문제가 여러 분야에서 다른 방식으로 접근할 수 있을 때 협력과 경쟁이 이루어지는데, 구체적인 이론이 있을 때 그리고 이 이론이 반증 가능할 때 그 발전이 가속될 수 있다. 힘 이론처럼, 현상에 대한 깊은 통찰에 근거한 이론에서 출발하는 연구가 더 많이 이루어지길 기대해본다.

참고문헌

박주용 (2000). 인과 추리: 철학적 배경과 힘-확률대비 이론을 중심으로. **인지과학. 18**, 37-48.

Beckers, T., De Houwer, J., Pineño, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 238-249.

Buehner, M., Cheng, P. W., Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1119-1140.

Carroll, C. D., & Cheng, P. W. (2010). The induction of hidden causes: Causal mediation and violations of independent causal influence. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 913-918). Austin, TX: Cognitive Science Society.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind.* Oxford: Oxford University Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367-405.

Cheng, P. W. & Buehner, M. (2012). Causal reasoning. In J. K. Holyoak & R. Morrison (Eds.), *Oxford Handbook of thinking and reasoning*. New York, NY: Oxford University Press.

Cheng, P. W., Liljeholm, M. & Sandhofer, C. (2013). Logical consistency and objectivity in causal learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2034-2039). Austin, TX: Cognitive Science Society.

Collins, D. J., & Shanks, D. R. (2006). Summation in causal learning: Elemental processing or configural generalization?. *The Quarterly Journal of Experimental Psychology, 59*, 1524-1534.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology, 47*, 109-121.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Gopnik, A. & Schultz, L. (2007). *Causal learning: Psychology, Philosophy, and computation.* New York, NY: Oxford University Press.

Gopnik, A., Sobel, D., Schulz, L. & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*, 620-629.

Griffiths, T. L., Kemp, & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Handbook of computational psychology* (pp. 285-386). Cambridge: Cambridge University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 285-386.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661-716.

Holyoak, J. K., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62,* 135-163.

Hume, D. (1739/1987) *A treatise of human nature* (2nd edition). Oxford: Clarendon Press. (**인간 오성의 탐구**. 김혜숙 역, 고려원: 서울).

Hume, D. (1777/1975). *An enquiry concerning human understanding and concerning the principles of morals* (3rd edition). L. A. Selby-Bigge & P. H. Nidditch (Eds.). Oxford: Clarendon Press. (**오성에 관하여**, 이준호 역, 서울: 서광서).

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9-31). Miami, FL: University of Miami Press.

Kant, I. (1781). *Critique of pure reason.* London, UK: Penguin Books. (**순수이성 비판**, 백종현 역, 서울: 아카넷).

Kelley, H. H. (1973). The processes of causal attribution, *American Psychologist, 28*, 107-128.

Leslie, Alan M. (1987). Pretense and Representation: The Origin of 'Theory of Mind'. *Psychological Review, 94*, 412-426.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40*, 87-137.

Liljeholm, M., Cheng, P. W. (2007). When is a cause the "same"? Coherent generalization across contexts, *Psychological Science, 18*, 1014-1021.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107*, 195-212.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science, 40*, 404-430.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955-982.

Lucas C. G., Griffiths T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34*, 113-147.

Lustig, R. H., Schmidt, L. A., & Brindis, C. D. (2012). The toxic truth about sugar. *Nature, 482,* 27-29.

Mackie, J. L. (1974). *The cement of the universe: A study of causation.* Oxford, UK: Clarendon Press.

Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A. -M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The Elusive Nature of the Blocking Effect: 15 Failures to Replicate. *Journal of Experimental Psychology: General. 145*, e49-e71.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review, 15*, 75-80.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*, 183-198.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*, 455-485.

Perales, J. C., & Shanks D. R. (2007). Models of covariation-based causal judgment: a review and synthesis. *Psychonomic Bulletin & Review, 14*, 577-596.

Rehder, B. (2006). When similarity and causality compete in category-based property generalization. *Memory and Cognition, 34*, 3-16.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasky (Eds.), *Classical Conditioning II: Current Theory and Research* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology, 87,* 88-134.

Shanks, D. R. (2010). Learning : From association to cognition. *Annual Review of Psychology, 61*, 273-301.

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, J. K. Holyoak, & D. L. Medin (Eds.), *Psychology of learning and motivation* (Vol. 34, pp. 265-311). New York, NY: Academic Press.

Sloman, S. (2005). *Causal models*. New York, NY: Oxford University Press.

Soto, F. A., Gershman, S. J., Niv, Y. (2014). Explaining compound generalization in associative and causal Learning through rational principles of dimensional generalization. *Psychological Review, 121*, 526-558.

Spelke, E. S. (1988). The origins of physical knowledge. In L. Weiskrantz (Ed.), *Thought without language*. Oxford, UK: Oxford University Press.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science, 10*, 92-97.

1차원고접수 : 2016. 10. 05
1차심사완료 : 2016. 11. 08
2차원고접수 : 2016. 12. 12
최종게재승인 : 2016. 12. 22

(*Abstract*)

Causal reasoning studies with a focus on the Power Probabilistic Contrast Theory

Jooyong Park

Department of Psychology & Institute of Psychological Science Seoul National University

Causal reasoning is actively studied not only by psychologists but, in recent years, also by cognitive scientists taking the Bayesian approach. This paper seeks to provide an overview of the recent trends in causal reasoning research with a focus on the power probabilistic contrast theory of causality, a major psychological theory on causal inference. The power probabilistic contrast theory (PPCT) assumes that a cause is a power that initiates or inhibits the result. This power is purported be understood through statistical correlation under certain conditions. The paper examines the supporting empirical evidence in the development of PPCT. Also, introduced are the theoretical dispute between the PPCT and the model based on Bayesian approach, and the current developments and implications of research on causal invariance hypothesis, which states that cause operates identically regardless of the context. Recent studies have produced experimental results that cannot be readily explained by existing empirical approach. Therefore, these results call for serious examination of the power theory of causality by researchers in neighboring fields such as philosophy, statistics, and artificial intelligence.

Key words : causal reasoning, the power probabilistic contrast theory, Bayesian approach, causal invariance