원형이론 기반 인공지능 편향성 분석 모형*

최인령 (서울대학교 이론물리학연구소, 선임연구원)

〈목 차〉

Ⅰ. 들어가며
II. 인공지능의 편향과 규범: 세 가지 관점
III. 원형이론 기반 편향성과 원형성
IV. 편향의 재정의와 편향성 분석 모형의 핵심 원리들
V. 나가며

국문초록

인공지능 편향은 개념 정의가 모호할 뿐만 아니라 통계적·인지적·사회-역사적 측면에서 다양하게 연구되어 왔으나 '편향이 무엇인가'에 대한 이론적 기반 없이 진행되어 왔다. 이에 편향의 기저에 있는 인지적·언어적 메커니즘을 설명하기 위해 레이코프(Lakoff)의 원형이론에 기초하여 편향 개념을 새롭게 정의하고 인공지능 편향성의 새로운 분석 모형을 제안했다. "편향을 규범으로부터의 이탈"이라고 본 다바(Davat)의 정의를 비판적으로 검토한 후, 편향성은 범주의 중심-원형으로부터 벗어난 거리에 따라 그 정도가 차등화되고, 이에 특정 사회문화적 맥락에서 편향성은 원형성과 반비례함을 논증하였다. 결국 본 연구는 편향이 중립적 규범에서 벗어난 오류가 아닌 원형 선택의 문제로 바라볼 것을 주장하였다.

키워드: 인공지능의 편향, 원형이론, 생성형 인공지능, 이상적 인지모형, 인지의미론

* 이 논문은 2020년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2020S1A5A2A01041434); 2024년도 대우재단 학술사업의 연구비 지원을 받아 수행된 연구임.

www.kci.go.kr

ⓒ 2026 이화인문과학원

Ⅰ. 들어가며[1)]

2021년 11월에 채택된 '유네스코 인공지능 윤리 권고'의 전문은 인공지능 기술로 인해 발생·악화될 수 있는 "편향이 불평등, 배제, 문화·사회적 다양성 및 성평등에 대한 위협을 초래할 가능성"[2)]을 명시하고 있다. 특히 2022년 말부터 생성형 인공지능(generative AI)이 우리의 일상 깊숙이 빠른 속도로 파고들면서 편향 문제는 학계와 사회 전반에서 중요한 논의 대상이 되고 있다. 이와 더불어 편향이라는 낱말의 사용은 급증하고 있으나, 학술 논문에서조차 그 개념 정의는 모호하고 다층적으로 사용되고 있다.

실제로 인공지능은 사용자·데이터·알고리즘 사이의 악순환적 되먹임(feedback) 구조를 통해 확증편향을 강화할 수 있는 위험을 시사한다.[3)] 따라서 인공지능 편향 문제는 단순한 기술적 결함이 아니라 다양한 맥락에서 이해되고 연구되어야 하는 복잡한 사회-기술적 현상으로의 접근이 강조되고 있다.[4)]

기존 인공지능 편향 연구는 크게 세 가지로 나뉜다. 데이터와 방법론의 객관성을 전제하는 방법론적 편향(methodological bias), 인간 사고의 한계를 인공지능으로 대체하려는 인지적 편향(cognitive bias), 그리고 기존의 차별을 고착화하는 사회-역사적 편향(socio-historical bias)이 그것이다. 첫 번째와 두 번째 연구가 편향을 측정하고 제거해야 할 기술

1) 편향 정의의 정교화와 편향 분석 절차에 대한 고견을 주신 심사위원님께 깊이 감사드립니다.

2) 이상욱 (2021), 『유네스코 인공지능(AI) 윤리 권고 해설서: 인공지능 윤리 이해하기』, 유네스코한국위원회, 84쪽.

3) Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman & Aram Galstyan (2021), "A survey on Bias and Fainess in Machine Learning," *ACM Computing Surveys*, 54, 6, Article 115(July 2021), pp. 1~35. (https://doi.org./10.1145/3457607, 2026. 7. 2. 접속).

4) Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall (2022), *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence,* NIST Special Publication 1270. (https://doi.org/10.6028/NIST.SP.1270, 2026. 3. 15. 접속); Marc Coeckelbergh & David Gunkel (2025), *Communication AI: A Critical Introduction to Large Language Models*, Policy.

www.kci.go.kr

적 결함으로 암묵적으로 전제한 채 통계적 측정과 기술적 해결에 집중해 왔다면, 다바(Davat)[5]는 이러한 두 가지 접근을 비판적으로 검토하고 세 번째 사회-역사적 편향의 관점을 강조했다는 점이 주목된다. 그러나 이 세 갈래의 연구는 공통적으로 '편향이 무엇인가'에 대한 이론적 기반 없이 진행되어왔다. 본 연구는 이 세 층위를 관통하는 이론적 토대로서 인지의미론의 원형 범주화 이론에 기초하여 인공지능 편향성의 새로운 분석 모형을 제안하고자 한다.

이를 위해 본 연구는 먼저 인공지능의 편향과 규범의 관계를 세 가지 관점에서 고찰한다(II). 다음으로 레이코프의 원형 방사형 범주 이론에 기초하여 편향성과 원형성의 반비례 관계를 규명하고(III), 마지막으로 편향을 원형 선택의 문제로 재정의함으로써 '편향 제거'가 아닌 '원형 재설정'이라는 대안적 접근의 필요성을 논증하고 인공지능 편향성 분석 모형의 핵심 원리들을 제시한다(IV). 본 연구는 원형이론 기반 인공지능 편향성의 새로운 분석 모형 제안이라는 소박한 목표 설정으로 범위를 한정하고, 인공지능 언어모델의 직접적인 실험은 이공계 인공지능 전문가와의 공동 연구를 통해 후속 연구에서 이어가고자 한다.

II. 인공지능의 편향과 규범: 세 가지 관점

로봇공학 및 언어과학 분야의 프랑스 학자, 다바(Davat)[6]는 인공지능 분야에서 널리 쓰이고 있지만 그 정의가 모호하고 사용 맥락이 다양한 '편향' 개념을 비판적으로 탐구한다. 이러한 편향의 개념적 문제로부터 출발해서 흥미로운 대안을 제시한 다바의 연구에 주목하고자 한다.

다바는 인공지능에서 편향은 복합적이고 다층적인 현상이므로 그 개념

5) Ambre Davat (2023), "Biais, intelligence artificielle et technosolutionnisme," *Éthique, politique, religions*, 22, 2023—1, *L'éthique de l'intelligence artificielle à travers les dispositifs et les pouvoirs*, pp. 67~83.

6) A. Davat (2023).

www.kci.go.kr

은 하나로 고정되어 있지 않음을 지적하고, 편향을 "특정 규범으로부터의 이탈(*écart par rapport à une norme*[7])"[8]로 정의한다. 그리고 이 관점에서 편향을 크게 세 가지 층위, 곧 데이터와 방법론의 객관성을 신뢰하는 방법론적 편향(*biais méthodologique*), 인공지능으로 인간의 사고 한계를 대체하려는 인지적 편향(*biais cognitif*), 그리고 기존의 차별을 고착화하는 사회-역사적 편향(*biais socio-historique*)으로 나누어 그 이면에 숨겨진 가치 판단, 곧 규범적 선택이 무엇인가에 관심을 기울인다.

첫 번째 층위는 방법론적 편향이다. 방법론적 편향은 인공지능 시스템 내부에서 구현되고 작동하는 '과학적 진실(*vérité scientifique*)'이라는 규범으로부터의 이탈을 의미하는데, 이는 데이터의 분포와 실제 세계 간의 불일치에서 비롯된다. 기계학습의 성능은 훈련 데이터의 품질에 의해 좌우되므로 "쓰레기를 넣으면, 쓰레기가 나온다(Garbage in, garbage out)"라는 원칙에 기반한다.[9] 여기서는 학습 데이터 선택 시 발생하는 선택 편향(selection bias)과 인간의 주관이 개입되는 주석 편향(label bias) 등이 포함된다. 기술 혁신 분야에서는 관찰자와 무관한 객관적 사실이 존재한다는 '실증주의(positivism)의 시각이 지배적이다. 그러나 데이터는 사회적 선택의 산물이라는 점에서 객관성에 대한 믿음과 관련하여 비판적 검토가 필요하다. 인공지능의 편향 문제는 데이터 자체의 문제일 수도 있지만, 설계된 맥락과 실제 사용 맥락이 다를 경우에도 발생한다. 이러한 경우의 사례로 구글의 Gemini 사건을 들 수 있다. Gemini가 인종차별적 이미지 생성 문제로 홍역을 치른 후 이를 바로 잡기 위해 과도한 설정을 적용하여, 2차 대전 독일군 이미지에 흑인이나 아시아 여성 이

7) 본고에서 원어 표기는 프랑스어는 이탤릭체로, 영어는 정자체로 적는다.

8) A. Davat (2023), p. 69.

9) Shah, Schwartz & Hovy에 따르면, 편향이란 "'진정한' 또는 의도된 분포(사용자나 레이블, 결과에 대해)와 모형이 사용하거나 생성한 데이터 분포 사이의 차이를 말한다(Bias as Difference between (a) a "true" or intended distribution (over users, labels, or outcomes) and (b) the distribution used or produced by the model)" (Deven Shah, H. Andrew Schwartz & Drid Hovy (2020), *Predictive biases in natural language processing models : A conceptual framework and overview.* (arXivpreprint arXiv:1912.11078.2019, 2026. 3. 15. 접속)

www.kci.go.kr

미지가 포함되는 실수를 초래했다.[10]

둘째 층위는 인지적 편향이다. 인지적 편향은 '경제적 합리성'이라는 규범으로부터의 이탈을 일컫는다. 이 관점은 심리학자 카너먼·트버스키(Kahneman & Tversky)의 '이중 사고체계' 이론에 기초한다.[11] 판단과 선택에 관여하는 사고체계는 시스템 1과 2로 구분되는데, 시스템 1이 빠르고 직관적인 어림셈법(heuristic)을 바탕으로 하여 효율적이지만 예측 가능한 오류와 편향을 낳을 수 있다면, 시스템 2는 더 느리지만 논리적이고 합리적인 추론에 기반하여 정확성이 높다는 것이다. 이와 같은 이론의 틀에서 편향은 "개인의 행동에 무의식적으로 영향을 미치는 체계적이고 예측 가능한 판단 오류"로 정의된다.[12] 더 나아가 카너먼(Kahneman)은 인간 판단의 변동성을 "잡음(noise)"이라 부르며 문제 삼는다.[13] 다시 말해, 인간의 판단은 일관성이 없고 편향성을 지닌 것과는 달리, 알고리즘의 장점은 '잡음 없는 결정'에 있다는 것이다. 결국 그에 따르면, 잘 설계된 알고리즘은 의사 결정 과정에서 나타나는 모든 형태의 주관성이나 변덕 등의 인간적 요소(*facteur humain*)을 제거함으로써 오류의 여지를 남기지 않는다는 것이다.

이처럼 카너먼·트버스키의 인지적 편향 개념은 인간의 사고 메커니즘을 설명하는 동시에 인공지능의 편향의 이해에도 영향을 끼쳤다. 실제로 인지적 편향 개념은 인공지능 맥락에서 단순히 시스템의 오류를 지적하는

10) 임대준 (2024), 「텍스트보다 이미지가 편향 심각… 이제는 AI 검색 및 알고리즘 추천으로 확대」, 『AI Times』, 2024. 11. 3.

11) Daniel Kahneman (2012), *Système 1/Système 2 : Les deux vitesses de la pensée*, Flammarion; Tversky, A. & Kahneman, D. (1974), "Judgement under uncertainty: Heuristics and biases," *Science* 185, pp. 1124-1131 [아모스 트버스키·대니얼 카너먼. 「불확실한 상황에서의 판단: 어림짐작과 편향」, 대니얼 카너먼 (2023) 『생각에 관한 생각』, 이창신(역), 김영사, 613~639쪽].

12) "Le biais cognitif est défini comme une erreur de jugement systématique et prévisible, qui va influencer de manière inconsciente les comportements des individus" (Davat 2023, p. 73).

13) Daniel Kahneman, Andrew M. Rosenfield, Linnea Gandhi & Tom Blaser (2016), "Noise : How to overcome the High, Hidden Cost of Inconsistent Decision Making," *Harvard Business Review*, Octobre 2016, p. 9, (Https://hbr.org/2016/10/noise, 2026. 6. 26. 접속).

www.kci.go.kr

수준을 넘어, 개발·사용·평가 전 과정에서 드러나는 인간적 선택과 구조적 요인을 분석하는 도구로 차용된다.[14] 결론적으로 이 담론에서는 인간을 편향된 존재로 규정하고, 알고리즘을 '잡음'이 없는 합리적 계산 모형으로 이상화하여 기술 낙관주의적 환상을 북돋우며 인공지능 도입의 정당성을 부여한다.

그런데 핀틀리·리·워시(Finkley, Li & Wash)의 최근 논문은 기술 낙관주의적 환상의 함정을 잘 보여준다.[15] 성별 편향과 중립성의 함정을 다룬 이 논문에 따르면, 대규모 언어 모델(LLM)이 동물을 주인공으로 한 이야기에서 성별 모호성을 어떻게 처리하는지 고찰하기 위해 23,800개의 인공지능 생성 동화를 분석하였다. 인공지능은 이야기 속의 성별 모호성을 해결하기 위해 '중립'이라는 방패 뒤로 숨어버렸다. 그 결과 오히려 여성 캐릭터를 지워버리고 남성 중심적인 세계관을 강화하는 심각한 편향을 보인 것으로 나타났다.[16]

마지막 셋째 층위는 다바 논문의 핵심인 사회·역사적 편향이다. 여기서 편향은 '이상적인 사회' 또는 '공정성'이라는 규범으로부터의 이탈을 의미한다. 이 관점은 인공지능 시스템이 기존 사회의 차별, 가령 인종차별이나 성차별 등을 복제하거나 증폭하는 위험을 안고 있을 뿐만 아니라 문화적 패권의 영향을 받고 있음을 지적한다. 다바가 언급한 GPT-3는 훈련 코퍼스 특성상 미국적 문화 가치를 반영한다. 이 담론에서 편향을 줄이는 작업은 단순히 기술적인 문제를 넘어 설계자가 설정한 가치나 이상적인 사회상에 따라 모형을 '재편향(*rebiaiser*)'하는 과정으로 이해된다. 다시 말해, 기술적 수정은 중립적 해결책이 아니라 설계자의 세계관을 주입하

14) Ramya Srinivasan & Ajay Chander (2021), "Biases in AI systems," *Communications of the ACM,* 64(8), pp. 44~49. (DOI: 10.1145/3464903, 2026. 4. 15. 접속)

15) Imani Finkley, Yuanxi Li & Melanie Walsh (2026), "Neutrality Bite: Gender representation in AI-Generative Animal Stories," In *The 2026 ACM Conference en Fairness, Accountability, and Transparency(FAccT' 26)*, June 25-28, 2026, Montreal, QC, Canada, ACM, New York, NY, USA, pp.1~19. (https://doi.org/10.1145/3805689.3812287, 2026. 7. 15. 접속)

16) 인공지능이 성별을 부여했을 때 여성 캐릭터는 단 2.2%에 불과했다. 이는 인간이 직접 썼을 때(13.4%)보다 6배 낮은 수치이다(Finkley, Li & Wash 2026).

www.kci.go.kr

는 행위가 될 수 있다는 것이다.

정리하면, 다바는 인공지능의 편향을 단순히 기술적인 오류로만 보지 않고, 과학적 진실, 합리성, 사회적 공정성 중에 어떤 규범을 기준으로 삼느냐에 따라 다르게 정의되는 다의적인 개념으로 설명하고 있다. 곧, 과학적 기준과의 괴리로서의 편향, 경제적 합리성과의 괴리로서의 편향, 또는 공정성을 기반으로 하는 이상적인 사회와의 괴리로서의 편향으로 요약된다.

이와 같은 다바의 주장은 두 가지를 전제한다. 첫째, 규범은 이미 존재하고 명확하게 정의 가능하고, 둘째, 그 규범은 편향되지 않은 중립적 지점이다. 다음 장에서는 이 두 전제를 모두 거부하는 레이코프의 인지의미론적 관점에서 편향성을 논의하고 다바의 편향 개념을 보완하고자 한다.

III. 원형이론 기반 편향성과 원형성

앞서 보았듯이, 다바에 따르면 인공지능의 편향은 고정된 하나의 정의가 아니라 '특정 규범으로부터의 이탈'이며 어떤 규범을 기준으로 삼느냐에 따라서 편향의 양상이 달라진다고 보고 세 가지 층위로 분석했으나 언어학적인 층위는 다루지 않았다. 이 장에서는 인지의미론적 관점에서 편향을 원형과의 상관관계로 볼 수 있는 논리적 근거를 제시하기 위해 레이코프의 '원형 방사형 범주화(radial category)'를 이론적 기초로 삼고자 한다.[17] 레이코프의 원형이론(Prototype theory)은 편향 담론의 인지적·언어적 메커니즘을 설명하는 층위에 놓이고 다바의 편향 개념에서 핵심인 규범에 대한 보완적 설명을 제공할 수 있다는 점에서, 그리고 현재 생성형 인공지능이 대규모 언어 모델을 기반으로 한다는 점에서 유의미한 작

17) George Lakoff (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind,* The University of Chicago Press [G. 레이코프 (1994), 『인지 의미론: 언어에서 본 인간의 마음』, 이기우(역), 한국문화사].

www.kci.go.kr

업이 될 수 있다.

1. 레이코프의 원형 범주화 이론: 원형 효과, 방사형 구조, 이상적 인지모형

다바의 편향 개념에서 '규범'이 무엇인가는 중요한 물음이다. 인지언어학자 레이코프의 원형이론은 바로 이 '규범'의 인지적 구조를 설명하는 이론으로서 다바의 편향 개념을 보완하는 데 상보적 역할을 할 수 있다. 원형이론의 핵심은 세 가지, 곧 '원형 효과'(Prototype Effet), 범주의 '방사형 구조'(Radial Structures), 그리고 '이상적 인지모형'(Idealized Cognitive Models, ICM)으로 구성된다.

인공지능은 인간이 세상을 범주화하는 인지 방식을 데이터로 학습하고 그 학습 데이터에 기반하여 생성물을 내놓는다. 본 연구는 인공지능의 편향성은 단순히 기술 해결주의자들이 주장하는 데이터의 오류나 결함에 의해 생기는 문제가 아니라, 인간의 인지 구조적 특성(원형 범주화, 이상적 인지모형, 은유)이 언어를 통해 인공지능에 이식된 결과로 볼 수 있음을 전제한다. 이는 인공지능의 편향 연구가 통계적 측정과 기술적 해결을 중심으로 전개되어 편향의 인지적·언어적 기원을 충분히 해명하지 못했다는 문제의식과 그에 대한 해법은 인간이 세계를 범주화하는 방식 자체에 대한 비판적 성찰에서 출발해야 한다는 인식에 바탕한다.

레이코프는 1987년에 출간한 그의 저서 *Women, Fire, and Dangerous Things*에서 인지심리학자 로쉬(Rosch)[18]의 원형이론과 의미에 관한 언어학적 연구들을 통합하여 기존의 객관주의 의미론을 정면으로 반박하는 새로운 형태의 인지의미론을 구축했다. 그의 인지의미론은 언어의 의미가 인간의 신체적·경험적 인지 과정에서 비롯된다고 보는 언어관으로서 체험주의(experientialism) 또는 신체화된 인지(Embodied cognition)라고 부른다. 이러한 인지의미론은 다음과 같은 몇 가지 특징들을 지닌다. 첫

18) Eleanor Rosch (1975), "Cognitve Representation of Semantic categories," *Journal of Experimental Psychology*, 104, pp. 192~233; (1978), "Principes of categorization", in E. Rosch & B. B. Lloyd (ed.), *Cognition and Categorization*. Hillsdale, Laurence Erlbaum Associates, pp. 27~48.

www.kci.go.kr

째, 의미는 객관적으로 존재하는 것이 아니라 의미부여에 의해 생긴다. 부연하면, 의미가 인간의 경험과 무관하게 존재하는 것이 아니라 우리가 몸을 가지고 세계와 상호작용하며 얻은 경험을 통해 만들어진다. 둘째, 그러한 의미부여의 근간이 되는 이상적 인지모형(ICM)의 지식 틀이 존재한다(뒤에서 더 상세히 볼 것임). 셋째, 형식주의 언어학에서 백과사전적 지식과 같은 언어 외적 요소를 의도적으로 배제했던 것과 달리, 의미부여 과정에서 광범위한 백과사전적 배경지식이 중요한 역할을 한다. 넷째, 의미의 바탕이 되는 개념은 원형 범주화를 통해 형성된다.[19] 이러한 언어관은 마음(정신)과 몸(육체)이 서론 다른 두 종류의 독립된 실체라는 데카르트의 심신 이원론의 철학적 관점과 대치한다.

로쉬의 원형이론은 인간이 머릿속으로 사물이나 개념을 어떻게 분류하고 이해하는지를 설명하는 기억 속의 인지구조, 곧 범주화에 관한 연구이다. 로쉬에 따르면 "범주 체계의 역할은 최소한의 인지 노력으로 최대한의 정보를 얻게 하는 데 있다".[20] 이러한 인지 경제성 원리에 따라, 인간은 개념을 분류할 때 완벽한 규칙 대신 그 범주를 가장 잘 대표하는 '원형(prototype)'을 기준으로 삼는다. 이러한 견해는 필요충분조건의 분석적 방법을 사용하는 고전적 범주관(classical categorization theory)을 거부하고 세상의 지각 구조에 대한 실제성 원리에 기초한다. 고전적 범주관에 따르면, 범주 구성원 모두는 동일한 특징들을 공유하면서 동등한 지위를 가지고, 범주 경계는 명확하고, 범주 내 구성원 여부는 이분법적으로 결정된다. 그러나 로쉬는 인간이 범주 구성원을 '더 원(전)형적인 것'과 '덜 원(전)형적인 것'으로 등급화하며, 범주 경계가 불분명하다는 것을 경험적으로 입증했다. 레이코프는 고전적 범주관을 정면으로 거부하는 로쉬의 연구를 인지의미론에 적용하여, 범주들이 방사형 구조와 가족 유사성(family resemblance)으로 특징지어진다고 주장하며 원형 범주화 이

19) Y. Tsuji (2004), 『인지언어학 키워드 사전』, 임지룡·요시모토 하지메·이은미·오카 도모유키(역), 한국문화사, 184쪽.
20) E. Rosch (1978), p. 29.

www.kci.go.kr

론을 발전시켰다.

레이코프의 원형 범주화 이론은 세 가지로 정리된다. 첫째, 범주는 고전적 필요충분조건이 아니라 전형적인 사례(typical example)인 원형(prototype)를 중심으로 방사형 구조로 구성된다. 여기서 원형의 역할이 핵심이다. 원형이란 특정 범주의 본보기로서 그 범주의 대표적인 특징들을 지니며 장기기억에 저장되어 있고, 인출이 쉽고, 빈번한 상호작용의 대상이 되며, 더 나아가 언어・사회・문화 공동체의 구성원 대다수가 공유하는 마음속 이미지, 곧 심상을 말한다. 원형은 한 범주를 대표하는 구성원으로서 인지적 기준점(cognitive reference point) 역할을 하기 때문에 다른 사례들이 원형을 중심으로 방사형 구조로 나열된다.

둘째, '원형 효과'와 범주 구성원 자격의 등급별 차등이다. 고전적 범주관과 차별성을 나타내는 또 다른 특징은 범주의 경계가 흐릿하다는 점, 범주 구성원들 모두가 동등한 지위를 갖는 것이 아니라 중심 원형에서 출발해서 더 전형적인 사례와 덜 전형적인 사례로 등급화된다는 점이다. 여기서 핵심은 원형 효과이다. 원형은 범주의 대표성을 지니기 때문에 다양한 인지적 기능에 체계적인 영향을 미친다. 특히 원형은 인지적 기준점으로 작용함으로써 원형적 사례가 비원형적 사례보다 범주 귀속의 판단이 수월하다는 이점이 있다.

이처럼 원형이론은 한 범주의 모든 구성원이 반드시 동일한 특징들을 공유하지는 않음을 인정하고, 범주 체계에 유연성과 역동성을 부여한다. 예를 들어, 펭귄은 '날 수 있다'라는 특징이 없어도 새의 범주에 속한다. 사람이 세상을 즉각적으로 인지하는 방식은 논리적이고 분석적이지 않으며 오히려 원시적이고 직관적이라고 할 수 있는데, 여기서 '원형' 개념이 도출된 것이다. 따라서 원형은 카너먼이 오류와 편향을 낳는 인지적 지점으로 지적한 시스템 1의 어림셈법에 상응한다(앞 절 참고).

예를 들어 보자. 동물학자가 새는 어떤 동물인가를 탐구한다면, 원형이론은 새다운 새란 무엇인가를 탐구한다. 인지심리학에서 새의 심적 이미지는 논리적・과학적으로 결정되는 것이 아니라 새의 본보기를 중심으로

www.kci.go.kr

형성되는데, '새'란 단어를 들었을 때 즉각적으로 떠오르는 새는 무엇인가라는 질문에 사람들은 대부분 참새나 제비, 독수리, 비둘기라고 답한다고 한다. 이 새들은 새의 실체를 대표하는 원형, 곧 새라는 범주의 본보기가 되며 장기기억의 새라는 범주 공간의 중심부에 위치한다. 그러므로 참새나 제비는 새 범주의 인지적 어림셈법에서 기준점이 된다.

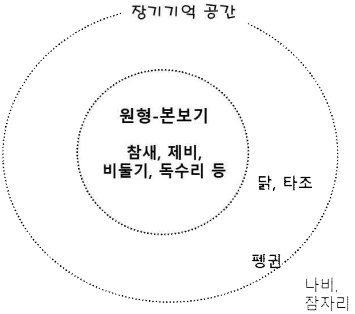
새의 원형-본보기에서 벗어나는 새들은 새라는 범주 공간의 주변부에 위치하게 되는데, 날개는 있으나 잘 날지 못하는 닭과 타조 등이 그러한 보기들이다. 그런데 범주 공간의 경계에 이르면 구별이 흐릿해진다. 깃털이 없고 날개가 퇴화한 펭귄은 날 수 없는데도 새의 범주에 귀속되지만 비원형적인 새(atypical bird)로 인지된다. 그러나 나비나 잠자리는 자유로이 날 수 있는데도 새로 인지되지 않는다. 나비와 잠자리는 새라는 기억 공간의 경계를 완전히 벗어나 있기 때문이다. 이를 모형으로 나타내면 〈도표 1〉과 같다.

'원형'은 학자들에 따라 '전형(steretype)'과 다른 의미로 사용되기도 하고, 동일한 의미로 사용되기도 한다. 다시 말해, 세계에 존재하는 실체를 전제 조건으로 내세웠던 초기의 원형 개념에서 "마음의 작용에 따른 인지적(심적) 이미지"로 의미가 확장되어 지금은 '전형'과 같은 뜻으로 사용되기도 한다.[21] 따라서 원형은 좁은 의미로는 '세계의 실체를 전제하는 본보기'를 뜻하고, 넓은 의미로는 '본보기-실체의 전형적인 특징들을 조합한 개념'이라 할 수 있다.[22] 예를 들어 새의 원형-본보기를 참새라고 한다면 새의 '전형'은 '새의 특징들(새는 날개와 깃털, 부리가 있다, 날 수 있다, 다리가 둘이다, 알을 낳는다 등)을 조합한 심적 구성체'(〈도표 2〉)이다. 이는 곧 새의 개념이 된다.

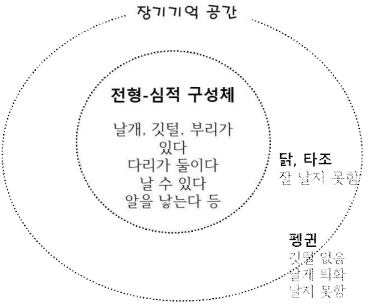
21) Goerge Kleiber (1990), *La sémantique du prototype: Catégories et sens lexical*, Puf; Danièle Dubois (ed. 1997), *Catégorisation et cognition: de la perception au discours*, Kimé.

22) 최인령 (2020), 『환기시학과 예술: 인지주의 관점으로 읽는 시, 음악, 회화, 광고.』, 이학사, 109쪽.

www.kci.go.kr



〈도표 1〉 새의 원형 모형



〈도표 2〉 새의 전형 모형

셋째, 원형 효과는 이상적 인지모형(ICM)에 기반한다. 원형 효과는 레이코프 이론의 또 다른 핵심 개념인 이상적 인지모형에 의해 발생한다. 여기서 핵심은 '이상화된(idealized)'이라는 수식어에 있는데, '이상화'라는 작업 자체가 '규범적 가정'을 담고 있다. 곧 이상적 인지모형은 세계를 정확히 반영하는 거울이 아니라 단순화되고 이상화된 도식으로 '규범화' 과정을 거쳐 형성된 것이다. 현실은 이러한 이상적 인지모형에 정확하게 맞아떨어지지 않는데, 바로 이러한 불일치가 원형 효과를 발생시킨다.

레이코프가 예로 든 총각/미혼남(bachelor)이 어떤 이상적 인지모형을 전제하는지 보자. 이 모형은 단혼제 사회, 결혼 적령기, 통상적인 결혼 관습 등의 규범적 전제를 담고 있다. '결혼한 적이 없는 성인 남성'이라는 총각의 정의는 이러한 이상화된 배경 위에서만 작동한다. 그러므로 교황은 정의상 '결혼한 적이 없는 성인 남성'이지만 그를 미혼남이라고 부르기를 망설이게 된다. 교황은 이상적 인지모형이 전제하는 결혼이라는 규범에 들어맞지 않기 때문에 미혼남 범주의 비전형적인 사례(atypical example)가 된다. 이처럼 정의 자체는 명확한데, 현실 사례가 이상적 인지모형의 이상화 조건에 어긋날 때 범주 구성원의 등급성이 발생한다. 레이코프는 원형 효과는 범주의 본질적 속성이 아니라, 이상적 인지모형이 개념을 구조화하는 방식으로부터 파생된 소산이라는 점을 강조한다.

www.kci.go.kr

결론적으로, 이상적 인지모형의 중심 사례인 원형으로부터 다양한 방식으로 변형된 사례들이 방사형으로 연결되어 확장되어 나간다. 이와 관련하여 다음 절에서 '어머니' 사례를 중심으로 상세히 살펴보려 한다.

2. 이상적 인지모형과 '어머니'의 방사형 범주 구조

레이코프의 이상적 인지모형은 특정 범주에 대한 지식을 구조화하는 이상화된 정신 표상으로, 추상화된 맥락들로 이루어진다. 그러므로 세계의 실제 사례가 이상적 인지모형과 직접 대응되는 것이 아니라 그 모형 안에서 인지적 구조화 원리에 따라 조직화 된다. 정보처리 수단으로 작용하는 이러한 이상적 인지모형은 인간의 신체적 경험, 인간의 행동과 목표, 인간의 사회적 상호작용에서 두루 작동한다.[23]

이 소절에서는 이상적 인지모형에 기반하여 '어머니'의 방사형 범주 구조를 살펴보려 한다. 어머니의 이상적 인지모형은 특정 사회·문화적 맥락에서 '어머니 다운 어머니' 또는 '어머니다움'의 인지적 표상, 곧 어머니의 전형적 이미지(typical mother)에 관심을 갖는다. 레이코프는 어머니 범주를 대표적인 '다발(cluster) 모형'의 예로 든다.[24] 다발 모형이란 여러 개의 이상적 인지모형이 한데 묶여 이루어진 복합적인 인지모형을 말한다.[25]

출산 모형: 어머니는 아이를 낳은 사람.
유전 모형: 어머니는 아이에게 유전 물질을 제공한 사람.
양육 모형: 어머니는 아이를 기르고 돌보는 사람.
결혼 모형: 어머니는 아이의 아버지와 결혼한 사람.
가계 모형: 어머니는 가장 가까운 여성 조상.

23) Alan Cienki (2010), "Frams, Idealized Cognitive Models, and Domaines," in Dirk Geeraerts and Hubert Cuyckens (ed), *The Oxford Handbook of Cognitive Linguistics*, Oxford Univ Press, pp. 170~187.
24) G. Lakoff 1987; G. 레이코프 1994, 87~90쪽.
25) 하나의 개념 범주는 여러 개의 인지모형들의 집합으로 구성된 경우가 많다.

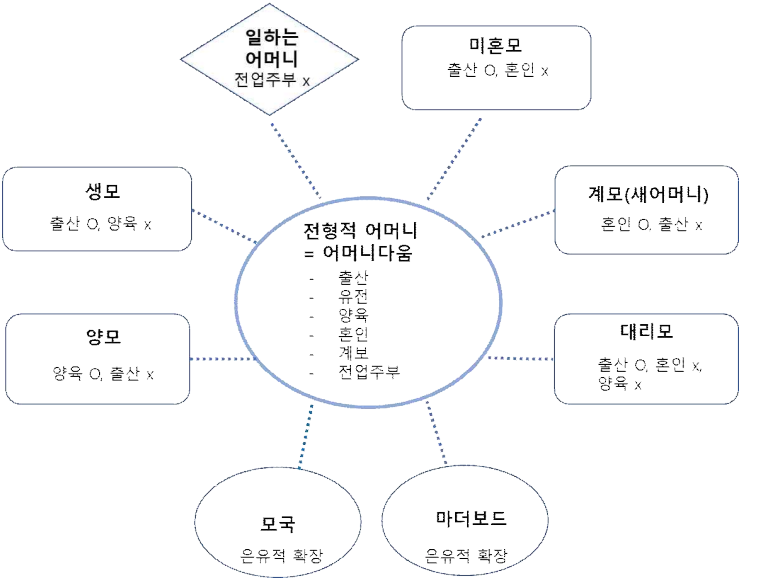
www.kci.go.kr

‘어머니다움’의 이상적 인지모형은 이와 같은 모형들의 집합으로 구성되며, 그 모형 속에서 아이의 아버지와 혼인하여 아이를 출산하고 양육한 여인이 ‘전형적인 어머니’(typical mother), 곧 어머니의 ‘원형’으로 인식된다. 이러한 다발 모형 중에서 출산 모형이 일차적인 것으로 간주된다.[26] 이처럼 레이코프는 어머니의 이상적 인지모형이 출산, 유전, 양육, 혼인, 계보 등의 여러 모형의 집합으로 구성됨을 보여주었다. 현실의 어머니들 가운데 이 모든 조건을 충족하는 사례는 소수일 수 있지만, 이상적 인지모형은 그것을 어머니다움을 연상시키는 ‘전형적인 어머니’로 이상화한다. 더 나아가 레이코프는 환유적 모형(metonymic model)도 원형 효과의 원천으로 본다. 이는 한 하위 범주가 전체를 대표할 때 나타나는 효과인데, 예를 들어 ‘전업주부 어머니’가 어머니의 사회적 고정관념(social stereotype)으로 기능하면서 전체 범주의 대표성을 갖게 된다.

그리고 이 전형적인 어머니를 중심으로 다양한 방식으로 변형된 주변 사례들이 동기화되어 〈도표 3〉처럼 방사형 구조로 확장된다. 곧, 전형적인 어머니를 구성하는 요소들 가운데 어느 요소가 결여되느냐에 따라 다양한 방사형 변이체들(생모, 양모, 대리모, 계모, 미혼모, 일하는 어머니 등)이 배치된다.

26) 어머니의 사전적 정의도 출산 모형이 기준이 된다. 국립국어원, 『표준국어대사전』의 정의에 따르면, 어머니는 “자기를 낳아 준 여자를 이르거나 부르는 말”이다.

www.kci.go.kr



[도표 3] 어머니의 방사형 원형 범주 구조[27]

이와 같은 방사형 범주 구조는 중심 원형-본보기와 그 본보기로부터 확장되어 나온 변형 보기들의 망이다.[28] 방사형 구조의 중심에 가장 전형적인 구성원이 위치하고 그 주변으로 다양한 방식으로 변형된 덜 전형적인 구성원들이 위치한다. 방사형 범주의 중심은 예측 가능한 반면, 주변부의 구성원들이 중심 구성원으로부터 파생되는 법칙을 예측하는 것은 쉽지 않다. 여기서 주목한 점은 주변부의 구성원들은 중심 구성원에 의해 동기화되어 가족유사성으로 연결되며, 차등적 지위를 갖는다는 점이다. 이 지점에서 레이코프의 핵심 주장 가운데 하나인 '원형 효과'가 발생한다.

위의 다발 모형과 '전업주부 어머니'라는 고정관념적 정보가 합세하여 전형적인 어머니, 곧 어머니다움을 형성하고, 그 어머니다움이 사회·문화의 규범으로 인식된다.[29] 이러한 복합적인 원형-어머니로부터, 생모(아이

27) 이 도표는 필자의 것임.
28) G. Lakoff 1987; G. 레이코프 1994, 101쪽.

www.kci.go.kr

를 낳았지만 키우지는 않는 어머니), 양모(아이를 낳지는 않았지만 키우는 어머니), 계모(아이를 낳지는 않았지만 아이의 아버지와 혼인한 어머니), 대리모(아이를 낳는다는 계약을 맺었을 뿐 아버지와 혼인하지 않았고 양육의 의무도 없는 경우)와 같은 어머니의 비전형적인 사례들이 파생된다. 더 나아가 미혼모는 출산 모형을 전제로 동기화되었고, '직장에 다니는 어머니(working mother)'는 오로지 양육이 쟁점이 될 때 전업주부 어머니라는 고정관념으로부터 이탈한 경우이다.

덧붙여서, '모국'과 '마더보드(motherboard)'는 어머니의 원형 효과에 의해 동기화 되어 은유적으로 확장된 개념들이다. 모국은 어머니의 출산 모형을 공유하며 자기가 태어난 나라를 가리킨다. 마더보드는 컴퓨터의 주요 부품을 끼울 수 있는 주회로 기판을 가리키는 용어로서 컴퓨터라는 신기술의 등장으로 인한 새로운 문화적 경험을 어머니의 기존 개념의 이상적 인지모형의 관점에서 이해하려는 인지적 노력의 결과이다.

이상적 인지모형은 문화적으로 특수하며 역사적으로 변화하는 구성물이다. 즉, 한 사회의 지배적 이데올로기, 권력 구조, 사회적 관계가 이상적 인지모형의 내용을 형성한다. 이 점이 인공지능 편향 연구와 직결되는 핵심 지점이다. 인공지능 시스템은 인간이 생산한 텍스트를 학습하므로, 그 텍스트 속에 내재한 문화의 이상적 모형, 곧 편향된 이상화를 흡수하고 재생산할 수밖에 없다.

'결혼식' 이미지 생성은 문화적 이상적 모형이 반영된 좋은 보기이다. 'Wedding'이나 한국어로 '결혼식'을 프롬프트에 입력하면 대개 하얀 드레스를 입은 신부와 양복을 입은 신랑의 이미지가 생성된다.[30] 레이코프에 따르면 모든 이상적 모형은 특정한 문화적 배경의 가정 위에서 작동한

29) 인공지능 시스템에 "Mother, 또는 한국어로 어머니 이미지 생성을 부탁해"를 프롬프트로 입력했더니 ChatGPT와 Gemini에서는 '전업주부'를 연상시키는 거실 배경의 따뜻한 미소를 짓는 어머니상을 생성했다. 이와 달리 Claude는 아이를 안고 있는 행복한 미소를 짓는 어머니상을 생성했는데, 이는 '출산 모형'을 따른 것이다(2026. 8. 24. 접속).

30) 인공지능 시스템 ChatGPT와 Gemini, Claude에 각각 "Wedding 이미지 생성을 부탁해"와 "결혼 이미지 생성을 부탁해"를 프롬프트에 입력했더니 둘 다 서구식 신랑·신부 이미지를 생성했다(2026. 6. 30 접속).

www.kci.go.kr

다. 인공지능의 '결혼식 이상적 모형'은 '서구 기독교 중심의 일부일처제'라는 배경 가정을 강력하게 고착화하고 있다. 한국의 활옷이나 인도 사리(Sari)를 입은 신부 혹은 동성 커플의 결혼식 등은 이상적 모형의 배경 가정과 충돌하기 때문에 인공지능이 '자동으로' 생성하기 어렵다.

끝으로 원형이론이 인공지능과 기계학습의 발전에 미친 영향을 보자. 원형이론은 인간은 대상을 완벽한 규칙으로 분류하지 않고, 마음속 '가장 전형적인 보기(원형-본보기)'와의 유사성을 측정해 분류한다는 통찰을 주었다. 이 개념은 인공지능과 기계학습(Machine Learning)이 발전하는 과정에서 초창기 심볼릭 인공지능의 기계적이고 딱딱한 규칙 기반 시스템에서 데이터 중심의 유연한 패턴 인식 시스템으로 진화하는 결정적 계기가 되었다.[31]

원형이론이 현대 인공지능 모형에 미친 구체적인 영향은 크게 세 가지다. 첫째, 규칙에서 유사도(Similarity) 기반 모형으로의 전환, 둘째, 현대 딥러닝의 '임베딩 공간(Embedding Space)'과 표현 학습[32], 셋째 '흐릿한 논리(Fuzzy logic)'와 소프트맥스(Softmax)' 함수이다. 현재 인공지능 시스템은 고전적인 컴퓨터 과학의 0과 1의 이진법적 세계나 참과 거짓의 흑백 논리에서 벗어나 '원형과의 유사도에 따라 점수'를 주는 방식으로 전환하였다. 그럼에도 원형이론을 기반으로 하는 인공지능의 편향성 연구 사례는 필자가 아는 범위에서는 발표된 바가 없다.

31) 최인령·신중휘 (2020), 「인공지능과 기계번역」, 『인공지능과 포스트휴머니즘』, 이중원 (편), 이학사, 65~101쪽.

32) 이 기법을 적용한 연구 사례들은 다음과 같다. Tomas Mikolov, Wen-tau Yih & Geoffrey Zweig (2013), "Linguistic Regularities in Continuous Space Word Representations", In *Proceedings of NAACL_HLT 2013, pp. 746~751,* Atlanta, Georgia. Association for Computational Linguistics; Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang & Kai-Wei Chang (2018), "Learning Gender-Neutral Word Embeddings", In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2018), pp. 4847~4853. Brussels, Belgium, Association for Computational Linguistics; 이수빈·정유남·송상헌 (2026), 「KoBiNLI — 자연어 추론 기반 한국어 성 편향성 벤치마크 데이터세트」, 『인공지능인문학연구』, 22, 217~248쪽.

www.kci.go.kr

IV. 편향의 재정의와 편향성 분석 모형의 핵심 원리들

다바의 논문에서 살펴본 바와 같이, 현재 인공지능 분야에서 가장 일반적인 정의는 “규범으로부터의 이탈”이다. 이 정의에서 규범이 표면적으로는 이미 명확하게 정의 가능하고, 편향되지 않은 중립적인 것처럼 보이지만 실제로는 문제를 안고 있음을 앞에서 지적했다. 이제 규범의 인지적 구조를 설명하는 레이코프의 원형 방사형 범주화에 기초하여 편향을 재정의하고 인공지능 편향성 분석 모형의 핵심 원리들을 제시하고자 한다.

레이코프의 이론을 정리하면, 원형은 ① 그 범주에서 가장 ‘좋은 사례’로 직관적으로 인식되고, ② 원형으로부터의 거리가 범주 소속 여부와 원형성의 정도를 결정하고, ③ 원형은 선험적으로 주어지는 것이 아니라, 신체화된 경험과 문화적 실천으로부터 구성되고, ④ 하나의 범주 안에 복수의 원형이 경쟁할 수 있다. 한 마디로, 모든 범주는 원형을 중심으로 방사형 구조로 구성된다.

이것을 ‘편향’이라는 개념에 적용하여 재정의하면, 편향이란 특정 원형이 중립적 규범인 것처럼 자연화되고, 그 원형으로부터 상대적으로 먼 사례·경험·관점이 체계적으로 주변화되어 과소 표상 및 평가될 때, 해당 시스템은 그 원형을 향한 편향을 갖는다고 정의할 수 있다. 이 정의에서 중요한 것은 편향이 이탈의 문제가 아니라 원형 선택의 문제라는 점이다. 결론적으로 ‘편향 없는 규범’은 인지적으로 불가능한 허구다. 이와 같이 재정의한 편향의 관점에서 보면, ① 편향성의 정도는 중심 원형성으로부터 거리가 멀수록 높고, 따라서 ② 약편향성과 강편향성으로 구분할 필요가 있음을 시사한다.[33] 결국 특정 사회·문화적 맥락에서 편향성은 원형성에 반비례한다고 볼 수 있다.

더 나아가 재정의한 편향을 세 층위로 나누어 살펴보자. 첫째, 언어적 층위에서는 은유 구조로서의 편향이다. 레이코프는 개념이 은유에 의해

33) 본고에서는 편향성을 약과 강으로 구분할 필요성을 시사하는 정도로 만족하고 향후 연구에서 편향 강도 척도에 대한 보다 정교한 연구를 통해 그 기준을 제시하고자 한다.

www.kci.go.kr

구조화된다고 주장한다. 프랑스어(*Biais*)와 영어(Bias)의 편향 단어 자체의 어원을 추적해 보면 흥미로운 점이 발견된다. 재봉의 '사선으로 잘린 것'의 물리적 '기울어짐'이 점차 추상화되어 판단의 기울어짐이나 사고의 왜곡 그리고 통계적 치우침과 같이 공정하지 못하고 한쪽으로 불균형하게 쏠린 태도나 경향을 나타내는 의미로 은유적 확정이 이루어진 것이다.[34] 이 은유는 올바름은 직선, 중심, 균형이라는 개념적 구조를 자동으로 활성화한다. 이 언어적 층위에서 편향이란 '특정 개념적 은유가 자연화되어 그 은유가 구성하는 원형이 보편적 규범'처럼 작동하게 되는 현상이다. '직선적 사고는 합리적 사고'라는 은유가 자연화되면, 우회적·직관적·연상적 사고방식은 자동으로 편향된 것으로 분류된다. 그러나 이것은 특정 인지 방식의 원형을 선택한 결과일 뿐이다.

둘째, 인지적 층위에서는 원형 효과로서의 편향이다. 이 층위에서 편향은 특정 원형이 범주 전체를 대표하는 것처럼 처리되는 현상이다. 다바의 논문에서 예로 든 '간호사와 여성' 연결이 이 층위의 원형적 사례이다. 인공지능 시스템이 '간호사'라는 범주의 원형을 '여성'으로 내장하게 되면, 남성 간호사는 그 범주의 주변적 사례, 즉 덜 전형적인 간호사가 된다. 이것은 논리적 오류가 아니라 원형 효과의 자동적 작동이다. 중요한 함의는 이것이 수정되어야 할 오류가 아니라, 어떤 원형을 선택할 것인가의 문제라는 점이다. 간호사의 원형에서 성별을 없애는 것은 편향을 제거하는 것이 아니라 다른 원형을 구성하는 것이다.

셋째, 정치적 층위에서는 원형 설정권의 문제이다. 이 층위가 다바의 논문과 가장 직접적으로 연결된다. 레이코프에게 원형은 권력 중립적이지

34) 프랑스어에서 'biais'가 나타난 시기는 1250년경이며 초기에는 재봉 분야에서 사용되었는데, 가령 부사 de biais는 옷감을 '사선으로 자른(coupé en diagonale)'을 뜻했다. 16세기에 이르러서 비스듬한(oblique), 경사진이라는 형용사로 그리고 비스듬함, 경사를 뜻하는 명사로 사용되었다. 16세기경 영어로 차용된 bias는 '비스듬하거나 대각선으로 된 선'을 뜻했다. 1560년대에 옛 볼링 경기에서 공이 직선이 아니라 비스듬히 굴러가는 현상을 가리키는 기술 용어로도 사용되었다. 그리고 1570년대에 '편향된 사고'처럼 심리적인 의미로 확장되어 특정 인물·주제·사물에 대한 호감이나 반감을 표현하는 용어로 사용되면서 대중적으로 확산되었다. 그리고 법률 분야에서는 '부당한 편견이나 성향'을 의미하게 되었다.

www.kci.go.kr

않기 때문에 누가 원형을 설정하느냐가 누구의 경험이 규범이 되느냐를 결정한다. 원형이론에 기초한 편향 정의를 정치적 층위로 확장하면, 편향이란 특정 집단의 경험과 관점이 원형으로 고착되고, 다른 집단의 경험은 그로부터의 벗어남, 즉 편향으로 표시되는 권력 작동 방식이 된다. 다바가 결론에서 디그나지오와 클라인(D'Ignazio & Klein)[35]을 인용하며 편향 대신 "억압(oppression)"이라는 개념을 제안하는 것은 이 층위의 문제를 지적한 것이다. 예를 들어, 정치에서 진보냐 보수냐의 문제, 종교 선택의 문제, 성별 또는 인종의 문제들이 이 층위에서 설명될 수 있다.

결국, 이러한 편향의 재정의는 편향을 수정 가능한 오류로 보는 카너먼식 관점도, 기술적으로 제거 가능한 결함으로 보는 기술해결주의적 관점과도 거리를 둔다. 대신 다바가 주장하였듯이, 편향을 원형 설정을 둘러싼 끝없는 정치적 협상의 장으로 바라보게 만든다.

위의 세 층위를 종합하여 정리하면 다음과 같다.

첫째, 편향 없는 시스템은 인지적으로 불가능하다. 모든 범주화는 원형을 전제하며, 원형 없는 범주화는 존재하지 않는다. 따라서 '편향을 제거한다'는 기술해결주의적 약속은 인지적으로 자기모순이다. 가능한 것은 원형의 교체이지 원형의 제거가 아니다.

둘째, '편향 제거'는 사실상 '원형 교체'다. 다바가 지적하였듯이, 실제로 편향 제거(*débiaiser*) 작업은 편향의 재설정(*rebiaiser*)이다.[36] 원형이론은 이것이 결함이 아니라 필연임을 보여준다. 중요한 것은 어떤 원형으로 교체할 것인가, 그리고 그 결정을 누가 어떤 과정을 통해 내리는가의 문제이다.

셋째, 편향의 문제는 인식론적 문제이기 이전에 정치적 문제다. 어떤 원형이 자연화되는가는 기술적 문제가 아니라, 어떤 경험이 보편으로 인정받는가, 어떤 목소리가 데이터 생산 과정에 참여하는가의 사회적 문제이다.

넷째, 편향의 다원성은 원형의 경쟁으로 설명된다. 다바가 하나의 단어

35) Catherine D'Ignazio & Lauren Klein (2020), *Data Feminism*. MIT Press.
36) A. Davat (2023), p. 77.

www.kci.go.kr

‘편향’이 인식론적·인지적·사회역사적으로 전혀 다른 의미로 사용된다고 지적하였는데, 이를 원형이론의 관점으로 보면 ‘편향’이라는 범주 자체가 경쟁하는 복수의 원형을 가진 방사형 범주임을 보여준다.

V. 나가며

인공지능 편향은 개념 정의가 모호할 뿐만 아니라 통계적·인지적·사회-역사적 측면에서 다양하게 연구되어 왔으나 ‘편향이 무엇인가’에 대한 이론적 기반 없이 진행되어 왔다. 이러한 문제의식을 바탕으로 본 연구는 레이코프의 원형 방사형 범주 이론에 기초하여 편향을 새롭게 정의하고 인공지능 편향성의 새로운 분석 모형을 탐구했다.

이 목표를 달성하기 위해 본 연구는 네 단계로 진행되었다. 첫째, 다바의 연구에 의하면 인공지능의 편향이 하나의 정의로 고정된 것이 아니라 ‘규범으로부터의 이탈’이며, 어떤 규범을 기준으로 삼느냐에 따라 편향의 양상이 달라진다. 이를 바탕으로 세 층위로 나누어 분석하였는데, 과학적 기준에서 벗어난 편향, 합리성에서 벗어난 편향, 그리고 공정성을 전제로 한 이상적 사회상에서 벗어난 편향이 그것이다.

둘째, 다바의 연구에서 다루어지지 않은 편향의 인지적·언어적 메커니즘 층위에 대한 연구의 필요성을 강조하고, 레이코프의 인지의미론적 원형 방사형 범주 이론에 기반하여 편향을 다음과 같이 재정의하였다. 편향이란 특정 원형이 중립적 규범인 것처럼 자연화되고, 그 원형으로부터 상대적으로 먼 사례·경험·관점이 체계적으로 주변화되어 과소 표상 및 평가될 때, 해당 시스템은 그 원형을 향한 편향을 갖는다고 규정할 수 있다. 이는 (i) 무엇이 원형으로 설정되어 있는가, (ii) 어떤 대상이 그 원형으로부터 얼마나 떨어져 있는가, (iii) 그 거리와 비례하여 어떤 대상이 체계적으로 과소 평가되는가, (iv) 그 결과가 어느 방향의 원형을 강화하는가의 질문들을 내포한다. 이에 편향의 지표는 원형에 대한 거리와 평가, 표상,

www.kci.go.kr

생성의 비대칭성이 결합하여 형성된다. 결국 편향의 정도는 중심-원형으로부터의 거리가 멀수록 높아지며 차등화될 수 있다. 따라서 특정 사회문화적 맥락에서 편향성은 원형성과 반비례한다.

셋째, 이상의 논의를 통해 본 연구는 원형이론 기반 인공지능 편향성의 새로운 분석 모형을 제안했고, 그 세 가지 핵심 원리와 분석 절차는 다음과 같다. ① 인간이 생산한 텍스트는 해당 사회의 지배적 이상적 인지모형을 반영한다. 인공지능 언어모델이 텍스트를 학습할 때, 사회문화적 맥락에서 형성된 이상적 인지모형에 따라 특정 집단을 범주의 원형으로, 다른 집단을 주변부로 배치하는 구조가 모델 가중치와 벡터 표상에 각인된다. ② 인공지능 시스템 내부의 범주 구조는 원형 중심의 방사형 구조를 이룬다. 각 범주의 원형으로 설정된 집단은 벡터 공간의 중심에 위치한다. 예컨대 남성이 '의사'의 원형으로 설정되면 남성이 방사형 구조의 중심에 위치하고 여성은 주변부로 밀려나며, 같은 원리로 여성이 '간호사'의 원형으로 설정되면 여성이 중심에, 남성이 주변부에 배치된다. 실제로 인공지능 시스템에 "의사의 이미지 생성을 부탁해"를 프롬프트로 입력했더니, Gemini, ChatGPT, Claude[37] 셋 다 남성 의사의 이미지를 생성했다. 그러나 여성 의사의 이미지를 생성하기 위해서는 위의 세 개의 인공지능 시스템 모두에서 '여성'이란 표지를 붙여야만 했다.[38] 이 실험[39]을 통해 편향의 판정 절차를 엿볼 수 있는데, 그 분석 절차는 다음과 같이 네 단계이다: (i) 남성이 의사 범주의 원형, 곧 기본값으로 설정됨, (ii) 여성은 원형으로부터의 거리가 증가함, (iii) 여성 의사에 대한 표상 및 생성 확률이 감소함, (iv) 그 감소가 남성 원형의 상대적 중심성을 강화함. 마지막 세 번째 핵심 원리로서 ③은 편향성의 정도는 범주의 중심-원형이 무엇으로 설정되는가에 따라 달라지고, 나아가 중심-원형으로부터의 거리

37) Gemini와 ChatGPT는 첫 프롬프트부터 사진 같은 명확한 남성 의사 이미지를 생성했고, Claude는 첫 이미지는 남성이라고 단정하기에 좀 애매한 그림을 생성해서 두 번째 프롬프트에 '의사 이미지를 섬세하게 생성해 줘'라고 썼더니 넥타이를 맨 남성 의사의 이미지로 생성했다.

38) Gemini, ChatGPT, Claude에 2026. 10. 18. 접속

39) 정교한 실험은 향후 연구에서 하려고 한다.

www.kci.go.kr

에 따라 그 정도가 차등화될 수 있다. 이는 연상 편향, 문장 완성 편향 등 다양한 편향 현상으로 가시화된다.

마침내 본 연구는 편향이 규범으로부터의 이탈의 문제가 아니라 원형 선택의 문제라는 결론에 이르렀다. '편향 없는 규범'은 인지적으로 불가능한 허구이며, 이른바 '편향 제거'는 실질적으로 하나의 원형 체계를 다른 원형 체계로 대체하는 원형 재설정(prototype replacement)의 문제이다. 이는 편향 문제가 기술적 교정의 차원을 넘어, 어떤 집단을 범주의 중심에 둘 것인가를 둘러싼 사회적·정치적 결정의 문제임을 함의한다.

다만 본 연구는 인공지능 언어모델을 통한 정교한 편향성 실험을 직접 수행하지 못했다는 한계를 지닌다. 본고에서 탐색한 원형이론 기반 편향성 분석 모형을 인공지능 실험에 적용하기 위해, 이공계 인공지능 전문가와 공동 연구를 후속 연구과제로 계획하고 있다.

www.kci.go.kr

참고문헌

이상욱 (2021), 『유네스코 인공지능(AI) 윤리 권고 해설서: 인공지능 윤리 이해하기』, 유네스코한국위원회.

이수빈·정유남·송상헌 (2026), 「KoBiNLI — 자연어 추론 기반 한국어 성 편향성 벤치마크 데이터세트」, 『인공지능인문학연구』, 22, 217~248쪽.

임대준 (2024), 「텍스트보다 이미지가 편향 심각... 이제는 AI 검색 및 알고리즘 추천으로 확대」, 『AI Times』, 2024. 11. 3.

최인령·신중휘 (2020), 「인공지능과 기계번역」, 『인공지능과 포스트휴머니즘』, 이중원 (편), 이학사, 65~101쪽.

최인령 (2020), 『환기시학과 예술: 인지주의 관점으로 읽는 시, 음악, 회화, 광고』, 이학사.

Cienki, Alan (2010), "Frams, Idealized Cognitive Models, and Domaines," in Dirk Geeraerts and Hubert Cuyckens (ed), *The Oxford Handbook of Cognitive Linguistics*, Oxford Univ Press, pp. 170~187.

Coeckelbergh, M. & Gunkel, D. (2025), *Communication AI: A Critical Introduction to Large Language Models*, Policy [마크 코겔버그·데이비드 J 건켈 (2026), 『기계가 언어를 사용한다는 것에 대한 인문학적 사유』, 신동숙 (역), 손화철(감수), 생각이음].

Davat, Ambre (2023), "Biais, intelligence artificielle et technosolutionnisme," *Éthique, politique, religions*, 22, 2023—1, *L'éthique de l'intelligence artificielle à travers les dispositifs et les pouvoirs,* pp. 67~83.

D'Ignazio, Catherine & Klein, Lauren (2020), *Data Feminism*. MIT Press.

Dubois, Danièle (ed. 1997), *Catégorisation et cognition: de la perception au discours*, Kimé.

Finkley, I., Li. Y. & Walsh, M. (2026), "Neutrality Bite: Gender representation in AI-Generative Animal Stories," In *The 2026 ACM Conference en Fairness, Accountability, and Transparency(FAccT' 26)*, June 25-28, 2026, Montreal, QC, Canada, ACM, New York, NY, USA, pp.1~19, (https://doi.org/10.1145/3805689.3812287, 2026. 7. 15)

Kahneman, Daniel (2012), Système 1/Système 2 : Les deux vitesses de la pensée, Paris, Flammarion.

Kahneman, D., Rosenfield, A.M., Gandhi, L. & Blaser, T. (2016), "Noise : How to overcome the High, Hidden Cost of Inconsistent Decision Making,"

www.kci.go.kr

Harvard Business Review, Octobre 2016, p. 9, (Https://hbr.org/2016/10/noise, 2026. 6. 26. 접속).

Kleiber, Goerge (1990), *La sémantique du prototype: Catégories et sens lexical.* Puf [조르주 클레베르 (2019), 『원형의미론: 범주와 어휘 의미』, 김지은(역). 한국문화사].

Lakoff, George (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind,* The University of Chicago Press [G. 레이코프. (1994), 『인지 의미론: 언어에서 본 인간의 마음』, 이기우(역), 한국문화사].

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021), "A survey on Bias and Fainess in Machine Learning," A*CM Computing Surveys*, 54, 6, Article 115(July 2021), pp. 1~35, (https://doi.org./10.1145/3457607, 2026. 7. 2. 접속)

Mikolov, Tomas, Yih, Wen-tau & Zweig Geoffrey (2013), "Linguistic Regularities in Continuous Space Word Representations," In *Proceedings of NAACL_HLT 2013, pp. 746~751,* Atlanta, Georgia. Association for Computational Linguistics.

Rosch, Eleanor (1975), "Cognitve Representation of Semantic categories," *Journal of Experimental Psychology*, 104, pp. 192~233.

_____, 1978, "Principes of categorization," in E. Rosch & B. B. Lloyd (ed.), *Cognition and Categorization*. Hillsdale, Laurence Erlbaum Associates, pp. 27~48.

Rothchild, Jennifer (2007), "Gender Bias", In G. Ritzer (ed.), *The Blackwell Encyclopedia of Sociology,* doi:10.1002/9781405165518.wbeosg011 (2026. 7. 3. 접속)

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P. (2022), *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence,* NIST Special Publication 1270, (https://doi.org/10.6028/NIST.SP.1270, 2026. 3. 15. 접속)

Shah, D., Schwartz, H. A. & Hovy, D. (2020), *Predictive biases in natural language processing models : A conceptual framework and overview.* arXivpreprint (arXiv:1912.11078.2019, 2026. 3. 15. 접속)

Srinivasan, R. & Chander, A. (2021), "Biases in AI systems," *Communications of the ACM,* aout 2021, 64(8), pp. 44-49. (DOI: 10.1145/3464903, 2026.

www.kci.go.kr

4. 15. 접속)

Tsuji, Y. (2004), 『인지언어학 키워드 사전』, 임지룡·요시모토 하지메·이은미·오카 도모유키(역), 한국문화사.

Tversky, A. & Kahneman, D. (1974), "Judgement under uncertainty: Heuristics and biases," *Science* 185, pp. 1124~1131 [아모스 트버스키·대니얼 카너먼, 「불확실한 상황에서의 판단: 어림짐작과 편향」, 대니얼 카너먼 (2023), 『생각에 관한 생각』, 이창신(역), 김영사, 613~639쪽].

Zhao, J., Zhou, Y., Li, Z., Wang, W. & Chang, K.-W. (2018), "Learning Gender-Neutral Word Embeddings," In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2018), pp. 4847~4853. Brussels, Belgium, Association for Computational Linguistics.

www.kci.go.kr

Abstract

AI Bias Analysis Model Based on Prototype Theory

CHOI, In-Ryeong (Center for theoretical Physics, Seoul National University, Research Fellow)

The definition of 'AI bias' is ambiguous. Although it has been studied from statistical, cognitive, and socio-historical perspectives, such research has proceeded without a theoretical foundation of what bias is. Recognizing this problem, we redefine the concept of bias based on Lakoff's prototype theory, in order to explain the cognitive and linguistic mechanisms underlying bias. We also propose a new analytical model for bias in artificial intelligence. After critically examining Davat's definition of "bias as a deviation from the norm", our study concludes that the degree of bias depends on how far it deviates from the central prototype of a category. It is also demonstrated that, in a specific socio-cultural context, the degree of bias is inversely proportional to prototypicality, indicating that bias is not a matter of deviation from neutral norm but of selecting a prototype.

Keywords: Bias in Artificial Intelligence, Prototype Theory, Generative Artificial Intelligence, Idealized Cognitive Models. Cognitive Semantics

논문 투고일: 2026년 09월 10일
심사 완료일: 2026년 10월 12일
게재 확정일: 2026년 10월 15일

www.kci.go.kr