

心灵的功能主义解释：因果角色定义及其非还原转向

孙亚萍，胥 献

云南师范大学马克思主义学院，云南 昆明

收稿日期：2026年7月4日；录用日期：2026年7月26日；发布日期：2026年8月6日

摘 要

功能主义作为20世纪下半叶心灵哲学的主流范式，主张心理状态的本质由其因果功能角色决定，而非特定的物理实现。本文系统梳理了功能主义因果角色理论的发展脉络与内在张力，在此基础上，重点考察了查默斯对功能主义的批判性修正。本文认为，功能主义从“纯因果角色定义”到“因果角色 + 现象属性”的理论演进，揭示了物理主义解释意识的根本困境，也为心灵哲学的跨学科建构提供了重要启示。

关键词

功能主义，因果角色，多重可实现性，感受质，自然主义二元论，交叉建构

The Functionalist Account of Mind: Causal-Role Definition and Its Non-Reductive Turn

Yaping Sun, Xian Xu

School of Marxism, Yunnan Normal University, Kunming Yunnan

Received: July 4, 2026; accepted: July 26, 2026; published: August 6, 2026

Abstract

As the dominant paradigm in the philosophy of mind in the second half of the twentieth century, functionalism contends that the essence of mental states is determined by their causal-functional roles rather than specific physical realizations. This paper systematically sorts out the developmental lineage and internal tensions of the causal-role theory within functionalism. On this basis, it focuses on David Chalmers' critical revision of functionalism. This paper argues that the theoretical

evolution of functionalism from a “pure causal-role definition” to a framework integrating “causal roles and phenomenal properties” lays bare the fundamental predicament faced by physicalism when accounting for consciousness, and also yields crucial insights for the interdisciplinary construction of the philosophy of mind.

Keywords

Functionalism, Causal Role, Multiple Realizability, Qualia, Naturalistic Dualism, Interdisciplinary Construction

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

大语言模型虽能流畅完成翻译、推理等任务,表现出类人的语言行为,但其核心疑问依旧悬而未决:仅靠符号运算, AI 是否真正拥有对世界的主观理解?这一难题本质复现了心灵哲学的核心争论:心理状态的本质究竟是什么?仅凭因果信息处理流程,能否完整定义意识?围绕心灵本质,学界已有笛卡尔二元论、心脑同一论、行为主义、取消主义及延展心灵论等多种路径。本文核心聚焦于普特南奠基的功能主义,即主张心灵本质是一套与物理载体无关的因果功能组织。本文旨在考察功能主义对意识的解释力,但不否定其他理论。最终,本文将提出整合功能主义因果分析、现象学主观体验与认知科学实证数据的交叉研究路径,避免单一理论的独断,以期更全面地阐释心灵问题。

2. 功能主义的核心主张与理论奠基

功能主义者(functionalism)主张人的信念、欲望、疼痛等等的心理状态本质上由其因果功能决定的,即表征为一种信息输入-内部状态-输出的关系,这是一种功能性质而非物理性质。大概可以用三个时段来概括功能主义的发展历程,早期、奠基期(1960~1970);经典成型期(1970~1980),这一阶段对功能主义进行了一系列的批判和深化;完善期(1980~)。心灵哲学家普特南(Hilary Putnam)是功能主义观点的主要提倡人物,同时也是机器功能主义的创始人。他认为,心理状态,是一种功能状态而非大脑的神经状态。在其论文《The Nature of Mental States》提出“心理状态是功能性的”。并逐一对行为主义和同一论者进行批判。澳大利亚哲学家阿姆斯特朗(D. M. Armstrong),提出的物理功能主义介于心脑同一论和功能主义之间。所谓物理功能主义就是,心理状态由感官刺激引起,并且相应的引起行为的内在、承载因果作用的物理状态。简单来说,就是心理状态确实是由因果引起的,但是他并不止步于因果角色定论心理状态的功能主义阶段,而是为其在现实世界找了一个物理承载体,即大脑。阿姆斯特朗反对行为主义,他认为心理状态不是行为倾向,而是行为产生的原因;同时他也反对二元论——心理状态就是大脑内在的物理状态,不存在非物理心灵。当今美国著名语言哲学家、心灵哲学家、认知科学家杰瑞·福多(Jerry A. Fodor)在普特南标准功能主义的基础上提出了计算-功能表征主义(Computational-Representational Functionalism),福多是最早为功能主义进行义正辞严辩护的哲学家,在其著作《Psychological Explanation》中,他反对行为主义,为功能主义辩护,也为其理论为计算-功能表征理论奠基,铺路。David Lewis(刘易斯)从对常识心理学分析为前提,从功能主义入手,以语义研究为切入点,试图调和功能主义和心脑同一论的平衡。随后 William Lycan(莱肯)提出的小人功能主义(Homuncular Functionalism),又叫目的论功能主

义(Teleological Functionalism),这是一种层级式的、生物学目的论、物理主义的功能主义,即以生物学、物理主义结合,以类似金字塔的层级划分来解释心理状态的功能主义。他试图用目的论加功能分解的模式对洛克“太宽、太窄”的问题做出回应,并为感受性提供解释。莱肯的小人功能主义是功能主义的成熟形态:以层级分解为方法论、演化目的论为本质、物理主义为本体论。以跨学科的视角,对多学科进行融合,它既保留功能主义的多重可实现性,又通过生物学约束避免极端自由主义,是当代心灵哲学中最系统、最科学导向的功能主义理论。此后,内德·布洛克(Ned Block)用著名的中国脑思想实验(China Brain),对功能主义进行了深刻的剖析和批判,认为功能主义理论有严重的沙文主义和自由主义,他固然重要,但有其致命缺陷——不能完全解释意识。认为刘易斯过分依赖常识,而普特南和福多过于偏向计算科学和认知科学等科学类领域,缺乏对心理意识的理解,这显然和心灵哲学的本质相违背。如功能主义字面所表示的那样,它只有功能的作用。他从物理角度和心理角度对功能主义作了个类似划分界限的区分,提出了功能主义“宽和窄”的概念。功能主义的反对者除了洛克,还有“中文屋”实验的提出者,美国著名的哲学家约翰·塞尔(John Searle)认为,句法 $\neq$ 语义,功能模拟 $\neq$ 真正理解,计算功能主义不能解释意向性与意识。托马斯·内格尔(Thomas Nagel)认为,功能主义剥夺意识的主观性,批判功能主义过于“手段”。

3. 因果角色定义的论证

在心理状态和大脑关系问题上,过去很长一段时间里,二元论和唯物论似乎是唯一的两种选择……直到二十世纪三十年代中期,人们发现了第三种可能性,这种理论被称为逻辑行为主义,该学说认为:正如数字被看作是集合的逻辑构造物一样,心理事件也被视为实际可能发生的行为的逻辑构造物([1], p. 325)。即我们发生的任何心理活动都可以被我们的外在行为表征,或者说我们的外在行为是心理事件的表征形式。逻辑行为主义者不反对二元论,但是认为我们的日常如“思想”等的概念,并不涉及笛卡儿式的心灵实体,即承认大脑和心灵都存在,二者不等同但是存在关涉性。普特南对这种观点提出反驳:“正如我们刚才所阐述的,逻辑行为主义的极端论点——所有能够感知疼痛的生物都是概率自动机。(2) 每个能够感知疼痛的生物都至少具备某种特定的描述性特征(也就是说,能够感知疼痛意味着相应的生物具有相应的功能结构)。(3) 没有任何能够感知疼痛的生物体其各个组成部分都具备(2)中所说的那种描述。(4) 对于(2)中所述的每一种描述而言,都存在感官输入的一个子集,使得在该描述下的生物只有当其某些感官输入属于该子集时才会感到疼痛([1], p. 434)。对于第二点,我们可以简单理解为:我们能够感受到疼痛是因为我们有疼痛的功能结构,比如,我们的手磕在桌子上,随即我们感到我们的手很疼,这是因为我们的手这个部位有疼痛的功能。但并不是所有能感受到疼痛的生物体都具备疼痛的功能结构,即我们前边所提到的:比如软体生物章鱼。即使一种生物具备了感知疼痛的功能,也还需要感官的子集输入时才会感到疼痛,这就是普特南‘概率自动机理论’中的概率问题,意味着我们哪个部位疼痛取决于感官对子集输入的概率统计选择。至此,我想我们可以思考一个问题:按照随机概率理论。是不是就意味着我们被桌子磕到然后感到手疼只是一种假象?这是因为感官随机输入子集的结果,有可能我们磕到手其实损伤的其他部位呢,比如我们的脑子。普特南的‘概率自动机’和‘图灵机器’的异同在于,后者具有固定规则,而前者无确切的规则可言,状态的转移是概率的统计分配。普特南指出,我们之所以能判断某个生物处于疼痛、饥饿愤怒或者发热状态,是基于其行为表现([1], p. 436)。同时,两个行为即关于‘所有的心理事件’都可以转换为对外在行为的描述——是错误的。”([1], p. 326)普特南利用“疼痛”的例子——这个例子贯穿于其功能主义立场的始终——对逻辑行为主义进行了反驳。他指出,“疼痛是一种复合型的概念”,它代表了一组相关现象,它并不是单一的,该词语的运用是由一整套的标准所决定的,这些标准可以被视作为非分析性的([1], p. 329)。正如某种病毒是诱发“小儿麻痹症”的原因,但不代

表这是唯一的诱因。基于此, 我们有理由认为, 疼痛通常会伴随着一系列的外在行为表征, 但是, 这并不意味着没有行为表现, 包括语言行为, 就断言没有疼痛的发生。二元论者采取的是因为……所以……这样的因果逻辑表达式, 但是普特南指出疼痛并非一系列反应的集合体而(根据我们目前的经验)通常是某些反应系列产生的原因([1], p. 332)。因为我们感觉到疼痛, 所以有某些反应, 即疼痛这种状态是功能, 他的功能在于让我们产生某种行为表征。功能主义和二元论, 前者从结果反推原因, 后者则遵从因果正常逻辑, 这就导致了两种截然不同的观念。我们再来看心灵哲学领域另外一个重要的观念——心脑同一论。心脑同一论者的观点是, 心理状态等同于大脑物理状态, 例如: “疼痛” = “C 神经纤维灼烧”, 但是, 章鱼这种软体动物能感觉到疼痛([1], p. 437) (基于我们的直觉和生物学常识), 但是章鱼并不具备和人一样的脑神经系统 “例如, 如果我能知道自己处于疼痛状态, 却不知道自己处于某种大脑的神经状态, 这就说明疼痛不可能是大脑的一种神经状态; 但用完全相同的逻辑来看, 如果我能知道炉子很热, 却不知道分子的平均动能很高(甚至不知道的分子的存在), 那就说明 ‘温度就是分子平均动能’ 这一说法是错误的, 这与物理学原理相悖。事实上, 仅从我能够知道自己……疼痛的概念与处于大脑神经状态的概念并不相同” ([1], p. 431)。显而易见, 普特南从概念本身出发, 通过对概念本质的厘清, 来说明疼痛并不属于大脑的神经状态, 有力地驳斥了心脑同一论的观点, 同时为他将要阐述的功能主义埋下了伏笔。至此, 普特南完成了对心脑同一论和行为主义的批判——不同物种物理结构不完全相同, 仍能感觉到疼痛, 说明心理状态可跨物理实现; 行为可以被抑制, 被切断, 但疼痛依然存在; 且区分了功能主义和二元论的异同——因果逻辑的不同推理。

普特南在其论文《The Nature of Mental States》明确提出自己的观点: 简而言之, 我将论证——疼痛并不是大脑(甚至整个神经系统)的某种物理 - 化学状态, 而是一种完全不同的状态。我提出的假设是: 疼痛, 亦即处于疼痛之中的状态, 是整个有机体的一种功能状态([1], p. 429)。“图灵概率自动机理论”是普特南功能主义理论的主要论证支撑。他认为, 人类心灵等同于一台图灵式的概率自动机, 心理的状态并非物理状态, 而是机器式的输入 - 输出的表达模式, 由状态间的概率来定义关系。我们整理回溯一下普特南的思想, 得出这样的结论: 心理状态不能简单地等同于脑部状态, 而是具有多重可实现性; 同时, 它也不是行为主义可以解释的, 心理状态是内部状态, 不能还原为外部行为; 心理状态不取决于物理材料, 也不取决于可观察到的外部行为, 决定它的是该状态在认知系统中所占据的因果角色。尽管普特南在后期推翻了他前期的功能主义的相关理论, 但是其理论思想依然具有深远的影响。它彻底取代了行为主义与传统心脑同一论, 更新了人们对心灵属性的固有理解, 以一种全新的视角 “因果 - 功能角色” 重新理解心理状态; 为认知科学与人工智能奠定了哲学基础; 推动了非还原物理主义的发展; 引发了关于意向性、意识、机器心灵的持久争论, 深刻塑造了 20 世纪下半叶以来的心灵哲学格局。

除了普特南, 功能主义的支持者还有当今著名的分析哲学、语言哲学和心灵哲学家杰瑞·福多(Jerry A. Fodor)。福多在标准的功能主义以及标准实在论的基础上提出了他自己的计算 - 功能表征主义(Computational-Representational Functionalism), 在普特南 “机器功能主义” 基础上, 将心灵视为基于 “思想语言” 的符号计算系统, 并提出心灵模块性理论, 深刻影响心灵哲学、认知科学与语言学, 思维本质是对 “思想语言” (Language of Thought, LOT) 符号的句法计算, 即人的信念、欲望、疼痛等心理状态是个体与心理表征的关系。这个理论可以简单的由下面两个原理组合而成: 对于任何一个有机体 O , 以及任何一个关于命题 P 的态度 A , 都存在一个(计算的/功能的)关系 R 并且存在一个心理表征 MP , 从而有 MP 意味 P , 并且 O 拥有 A 当且仅当将 R 赋予 MP [2]。福多的功能主义有别于前期的机器功能主义和热分析功能主义, 属于心理功能主义的范畴。福多于 1965 年发表的论文《Explanations in Psychology》旨在于批判行为主义, 主张心理学解释必须诉诸内部心理状态与因果结构, 初步提出功能主义框架。福多从对内部心理状态的解释入手, 阐述什么是心理层面的解释, 由此对行为主义提出批判。福多首先阐明什么是 “行为解

释”，以此预防人们将其和“行为的心理学解释”混为一谈。比如你买了一个巧克力味儿的冰淇淋，我想知道原因，不管是因为你喜欢巧克力，还是因为其他口味的太贵，又或者是那一刻你只想买巧克力味儿的，这其中的任何一个理由或者全部的理由都可以构成解释，但是，这只是行为层面的解释，是对你为何买了巧克力味儿的冰淇淋这一行为做出的解释。当我们将心理现象作解释时运用的理论本身不是心理学理论时，那么这种解释自然不可取。显然，你买巧克力冰淇淋的原因即使就是如此——不涉及心理学理论。这种解释很容易将二者混为一谈，就连心理学家对于什么是心理学的解释和哪些事物可以被视作心理学的解释都没有统一的看法([3], pp. 161~179)。在此我要说明一点，福多对于二者厘清目的不是将二者划清界限，而是从概念上解释异同，实则它们是有所关联的——行为可以用心理学来解释。以往的哲学家，或者心理学家，又或者是我们的常识性告诉我们：心理学解释的是那些反常的行为，或者心理学不涉及行为本身，而只关注行为发生的伴随动作，比如“疼痛”是行为，伴随动作可能是皱眉或者哭喊等等之类的动作。正如吉尔伯特·赖尔在《心灵的概念》一书中写道：“对各种心理缺陷的表现进行分类和诊断需要专门的研究方法。而对于人类心理能力的体现，则往往只需要运用常理加以解释即可；或者，也可能需要经济学家、学者、战略家及评估专家等所使用的专业方法来分析。不过这些解释并非基于某种更为根本的诊断标准所得出的。因此，并非所有，甚至大多数关于人类行为和反应的原因解释，都能被归为心理层面的。”[4]福多所说的对于“行为的心理层面的解释”包括所有的行为，不管正常和不正常的行为，众所周知，在理论和概念上，都属于“行为”，这一点毋庸置疑。福多在他的“心理学解释”中，从概念角度对“什么是心理解释”和“心理学家如何使用‘解释’”作了解释，按照福多的理解，前者并非传统意义上的语言学研究([3], pp. 161~179)。福多不赞成对科学的元理论必须全部由分析性的称述构成的观点，他认为对于行为的心理解释不是一蹴而就的，它由两个阶段共同组成，第一阶段是观察生物体或是自发，或是受特定刺激后做出的反应；但第一阶段参与内在解释，仅仅只是可观察到的外在表征。真正触及根本的是第二阶段的理论，即我们对这些行为产生背后原因的内在状态的描述，这些描述完全是功能性的。综上，福多在这篇文章中批判行为主义排斥内部心理状态的解释模式，主张心理学解释必须诉诸作为理论实体的内部心理状态，并认为心理状态的本质由其在因果网络中的功能角色决定，从而为心理学作为一门自主的特殊科学奠定哲学基础，也为功能主义的建立搭建起最初的框架。

分析功能主义的代表大卫·刘易斯(David Lewis)在其论文《An Argument for the Identity Theory》中引入心理状态的因果潜能概念：心理状态是引发典型行为、由典型刺激引起的内在物理因果状态。以此来为心脑同一论提供基于因果论证的先验辩护，同时奠基后来的分析功能主义。刘易斯并不排斥同一论，而是用因果逻辑必然性来为心脑同一论做理论支撑，以此在心脑同一论和功能主义之间建构沟通的桥梁，对二者进行调和。同一论者认为，某些物理状态属于体验、内省过程或者活动本身，而非那些被经验感知的对象([5], pp. 99~107)。显然，大卫·刘易斯并不赞成这种单一的观点，即使是抽象的事物，也应当有相应的物理载体，同一论看起来就是简单的 $A(\text{经验}) = B(\text{神经状态})$ 的关系，核心在于 A 和 B 所描述的现象都指向相同的对象——作为经验的神经状态。但这似乎过于浅显，因为这描述所指向相同，并不意味着他们具有相同的内涵，心理状态必然就是大脑状态的核心在于他们之间有因果逻辑必然性，而非主观的感受或者是某种非物理实体。具体论证如下：① “人感到疼痛”是心理状态，引发外在表征如“皱眉”等物理行为，而“针扎”这一物理行为，引发“疼痛”这一心理状态，就会做出相应的“退缩”的物理行为。如果“疼痛”不是心理状态，那么就会出现物理行为引发非物理行为的，这违背了物理因果闭合原则。因此，心理状态必然就是物理状态。② 一个人“想要喝水”(心理状态 M1)，且“相信面前的杯子里有水”(心理状态 M2)，这两个心理状态共同引发了“伸手拿杯子、喝水”(行为 B)。如果 M1 和 M2 没有因果力，那么“伸手喝水”的行为就无法被合理解释——我们无法否认，心理状态确实能“导致”行为发生([5], pp. 99~107)。心理状态不仅能引发行为的发生，而且还能与其他心理状态产生因果关联，具

有明确的因果力。③ 以物理闭合原则为前提, 得出结论: 任何的物理行为(退缩、喝水)都能被物理原因(神经冲动、肌肉收缩)解释, 而无需引入非物理原因。④ 1) 退缩(B)是物理行为, 根据物理因果闭合, 必然存在一个物理原因 P (如“大脑中 C-纤维激活后, 传递神经信号到肌肉, 引发收缩”), P 能完全解释 B 的发生。2) 但我们已经确定, 疼痛(M)也是 B 的原因(第一步已证明)。3) 此时出现两种可能: ① M 和 P 是两个独立的原因, 共同导致 B; ② M 和 P 是同一个原因, 只是名称不同。4) 排除①: 若 M 是非物理的, 那么 B 就有“物理原因 P”和“非物理原因 M”两个独立原因, 这会导致“因果过度决定”(overdetermination)——但刘易斯指出, 这种情况在现实中不存在(原文反例: 一个人被针扎后, 即使没有非物理的疼痛, 只要 C-纤维被人工激活, 依然会出现退缩行为; 反之, 若 C-纤维不激活, 即使有“疼痛”的主观感受, 也不会有退缩行为)。5) 由此我们得出结论: M 和 P 必须是同一个东西——心理状态 M 的因果力, 本质就是物理状态 P 的因果力, 因此 $M = P$ ([5], pp. 99~107)。因此, 心理状态和物理行为的关系是具有因果链接的诠释关系, 心理状态的本质就是大脑的神经物理状态。刘易斯不认可行为主义的原因在于行为主义否定了体验的实在性, 心理状态引发相应的行为表征这一特性决定了行为主义者不可能承认体验和他具体发生的情境、行为表现是相互独立的, 如果那样的话, 本质上和二元论没有区别。刘易斯在本文中并未完全否定“功能主义”, 而是为功能主义埋下伏笔——他认可“心理状态由因果角色定义”, 但认为“因果角色最终由物理状态实现”, 1972 年的论文《Psychophysical and Theoretical Identifications》以 1966 的因果同一论为契机, 结合著名的拉姆齐理论术语, 正式建立起“刘易斯分析因果功能主义”, 构建了身心同一论的标准论证。(拉姆齐最早阐述(1922)“拉姆齐语句”的论文“Theories”于 1931 年由其同学 L·B·布雷斯韦特编辑首次发表([6], pp. 212~236)。)完成了“刘易斯因果功能主义、心身同一论”的完整体系。

大卫·阿姆斯特朗在其 1962 年发表的著作《一种唯物主义的心灵理论》(A Materialist Theory of the Mind)中, 从唯物主义的立场出发, 系统论证了因果心灵理论, 论证的核心是“心理状态与其物理实现”。拒斥逻辑行为主义, 反对将心理状态还原为外在的行为, 心理状态是真实存在的内在的状态, 而非简单的外在行为解释, 瓶子碎了不是因为“瓶子会碎”这种倾向, 是由瓶子本身内部的分子结构所决定的。瓶子的分子结构定义自身因果, 结合外部环境刺激, 比如有人不小心碰到它, 联通其他心理状态共同发生了“碎掉”这个行为, 这是一种有内部因果关系参与的输入和输出链条。他认为行为主义可以解释我们的心灵和身体之间的紧密联系, 也可以解释两个“心”之间的数量差异, 且行为主义与大脑的认知科学是一致的。但是行为主义者无法接受“心理可以脱离肉体而存在”这一逻辑上的可能性[7]。换言之, 阿姆斯特朗认为行为主义者无法真正地解释心灵的相互作用。行为主义者把“心理倾向”看成是解释各种心理概念的重要法宝, 如持有某种信念仅仅只是他采取某种行为的倾向而已, 我去拿笔这个行为是因为我心里有拿笔这个想法。但是, 问题就在于, 如果我们相信地球是平的, 基于尝试逻辑, 不见得我们会采取什么行为。他主张行为必须依托内在的物理范畴奠基其本体论基础是可被经验验证的物理主义。这一观点与他后续主张的“中心状态唯物主义”相呼应——心理状态作为一种具有因果力的内在状态, 最终需通过大脑等物理器官实现, 且可通过科学手段(如神经科学研究)验证其存在。在具体论述中, 阿姆斯特朗以“疼痛”这一典型心理状态为例, 疼痛感需要某种背景才能存在, 必须有某种东西让疼痛得以产生……我们之前所论述的, 以下顺序显然是一种因果关系: (a) 我的手受到撞击; (b) 我感到疼痛; (c) 我扭动双手。这其中的因果角色应该是这样定义的: 由身体损伤等刺激引发, 能够促使个体产生退缩、寻求缓解等行为, 而这一心理状态的物理载体, 正是大脑中与疼痛相关的神经活动(如 C-纤维激活), 这一论证既区别于行为主义仅关注外在表现的局限, 也为后续心理状态的物理还原提供了核心逻辑支撑, 同时为“心理状态与物理状态的偶然同一”奠定了基础, 成为其中心状态唯物主义理论的重要组成部分。他明确指出: 所谓心理状态, 本质上是指能够引发某种行为的个体状态。即心理状态并非活动本身, 而

是引发行为的原因,对于某些心理状态而言,他们同时也是能够被特定刺激所引发的个体状态([7], p. 58)。同时强调“任何心理状态的因果作用,都必须以可经验观察的物理状态为基础,不存在脱离物理载体的独立心理因果”([7], p. 155),这也进一步印证了其中心状态唯物主义的核心立场,即心理状态是具有因果力的内在物理状态,与大脑的物理神经状态存在不可分割的关联。

综上,阿姆斯特朗的因果角色定义具有两大优势:第一,合理地解释了心灵的因果性,破解了一直以来困扰的心物互动难题;将心理状态定义为引发外在行为的内在动力,这彻底区别于逻辑主义者将行为解释为是等同于心理状态的局限。第二是它对复杂的心理状态具有极强的包容性,对各类心理现象做了详细的解释。相关论述既兼顾了单一心理状态的因果角色,也涵盖了心理状态之间的互动关系,能够统一解释知觉、信念、欲望、意识等各类心理现象,同时为后续将各类心理状态还原为大脑物理状态提供了逻辑支撑,体现了其因果理论的包容性与解释力。文中也通过对“信念与欲望的因果互动”的分析,进一步印证了这一优势,强调“心理状态的因果网络,是其能够解释复杂行为的核心前提”([7], p. 98)。阿姆斯特朗关于心理状态源于因果力的核心论述,为唯物心灵理论奠定了基础,给当代心灵哲学研究提供了启示,搭建起心灵哲学与认知科学等其他学科的交叉融合发展的桥梁。在心灵哲学方面,助力我们打破心物二元论的定势思维,启发我们从“因果力”出发,开启全新的研究视角。

4. 非还原转向:查默斯的批判与超越

视角转向近现代,哲学家大卫·查默斯(David J. Chalmers)在批判功能主义的基础上,提出自然主义二元论的因果角色理论。查默斯在坚持“心理状态由因果角色定义”的基础上,对还原物理主义的局限性进行了突破,提出“心理状态的因果角色与现象属性并存”的非还原观点,主张因果角色是心理状态的“功能层面定义”,而现象属性(感受质)是心理状态的“主观层面特征”,二者共同构成心理状态的完整本质。

他对这一理论的阐述主要集中于其著作《有意识的心灵》和《概念分析与还原解释》中。他认为,心理状态的因果角色主要体现为“信息处理层面的因果关联”,即心理状态是“接收输入信息、处理信息并输出行为的功能状态”,这种因果角色可通过功能组织来刻画,且具有“多重可实现性”——只要两个系统具有相同的功能组织(即相同的因果角色网络),就具有相同的心理状态。查默斯对功能主义的因果角色论述围绕自然主义二元论展开,认为意识不能被还原主义解释,他首先论证意识不是逻辑的随附于物理的,世界上所有的微观事实,都不会蕴含着关于意识的事实[8],如同他在另一篇文章中所指出的那样,这不是包含关系,而是属性并存的关系。首先,查默斯梳理了传统功能主义的核心界定,认同心理状态是信息输入和输出的关系,这一观点与普特南、福多等功能主义者的立场一致,这也是后期查默斯进行批判和修正的基础。(后期查默斯结合自然主义二元论对心理状态的因果角色重新做了修订。)查默斯通过怪人论证(Zombie Argument)的逻辑可能性、颠倒光谱论证(Inverted Spectrum Argument)、来自认识的不对称性(Epistemic Asymmetry)、知识的论证(Knowledge Argument)、来自分析的缺席论证(Absent Qualia Argument) ([9], pp. 78~96, 118~135),表明意识的不可还原性,推翻了传统功能主义(普特南、福多等)认为物理主义足以解释心理状态的主张,从而建立自己自然主义二元论的立场。他设想有这样一个世界,这个世界里全是没有心理状态的,即没有意识的怪人,但是他们在物理层面上与我们并无异同,这在逻辑上是可能的,在形而上学上也是可能的,以此得出结论:物理载体不足以解释或者决定导致意识的存在。既然意识无法被还原为物理状态,那么就必须存在意识的属性与物理属性并存。颠倒光谱论证(Inverted Spectrum Argument):设想两个人在物理功能上完全一致(看到红绿灯的神经反应、说话习惯完全一样),但其中一个人的颜色体验是颠倒的(例如他看到的红色其实是绿色的主观感受,但他从小就被教育称这种感受为“红”),进一步证明,心理状态的本质不仅仅是因果角色,还包含无法被功能还原的主观现象特

征,即感受质。拥有完全相同的因果角色的两个人,他们在体验上可能完全不相同,指出“差异”的可能性;在认识的不对称性的论证中,查默斯指出我们在认识的问题中,具有人称的不对称性,比如站在第三人称的角度,我们可以轻易地解释一个因果过程或者是行为倾向,甚至可以大致地模拟心理状态。但是我们在第一人称的角度,具有只可意会不可言传的主观体验特性,我的意识的知识首先来自于我自己的情况,而不是来自于外部的观察,也就是第三称的外物观察总结的物理事实。用查默斯的话说,我们的有意识的经验存在的知识主要来自于我们自己的情况,外部证据最多起到次要的作用([9], p. 127)。知识论证以“玛丽的房间”为思想实验:玛丽精通神经生理学,但是他却从来没有看到除了黑白灰以外的任何颜色,所以当看到她看到红色这一物理现象时,却不知道这是什么,主观经验不被物理事实所蕴含,主观体验不能被物理现象推导出。物理想象和主观体验之间存在着解释的鸿沟。来自分析的缺席论证和怪人论证所要论述思想基本一致,几大论证共同构成了反驳纯物理主义的核心支撑理论。查默斯完成了对意识的不可还原性,至少是不可完全还原为物理解释,物理事实能解释简单的意识,但是对于“感受质”这类的高阶的“难意识”问题,物理事实显然是不能解释的。随后,查默斯结合其本体论立场:自然主义二元论,确定了他的因果角色定义。(查默斯认为存在着一个性质二元论,而这种性质不被个体的物理属性所蕴含,尽管可能合法的依赖于那些性质,这种二元论不同于笛卡尔的二元论,他没有对物理过程进行施加的单独的精神实体的领域([9], p. 154)。)。他强调,因果角色的核心是“抽象的功能组织与信息处理层面的因果互动模式”,这种因果角色具有多重可实现性,且无法完全还原为物理状态的因果力;同时,因果角色与现象属性是心理状态的两个不可分割的维度,因果角色负责解释心理状态的功能行为,现象属性负责解释主观体验,二者共同构成心理状态的完整本质,这一观点既保留了功能主义因果角色的合理内核,又弥补了其忽视感受质的缺陷,与《概念分析与还原解释》中的非还原功能主义立场一脉相承。

《Conceptual Analysis and Reductive Explanation》为查默斯和弗拉克·杰克逊合著,该书立足非还原功能主义立场,坚持“心理角色由因果定义”的核心主张,以概念分析法为依托,完善了因果角色的界定逻辑,提出了“心理属性与物质属性并存”的观点:用完整的物理学语言来说,微物理真理数关于物理学的基本实体和属性的真理。我们也可以规定,P包含了物理学基本定律的结合。我们不会涉及什么“物理学”的问题,但会规定,如果存在任何的精神实体或者性质,那么它就不是物理学的一部分([10], p. 318)。显而易见,心理状态是精神实体,具有相关精神实体的性质,所以在最基本的概念上,他就不属于物理学,而是一种并存的关系。因果角色是心理状态的“功能层面定义”,而现象属性(感受质)是心理状态的“主观层面特征”,二者共同构成心理状态的完整本质,打破了还原物理主义将因果角色等同于物理因果力的局限。二人认为,因果角色是信息处理的功能关联,即是一种接受信息、处理信息并输出信息的一种功能状态。这种因果角色具有多重可实现性,如果两个系统具有相同的功能组织,即相同的因果角色网络,就具有相同的心理状态。比如水在形而上学上不可能是XYZ,但在认识论上,水是XY是可能的,在先验上,水是XYZ是可能的……这种认识论可能与可能性之间的先验关系,与形而上学可能性之间的不同必然关系是完全相容的([8], p. 326)。

查默斯这两篇文章一脉相承,前者通过本体论立场(自然主义二元论)为因果角色的非还原性理论提供支撑,后者通过概念分析完善因果角色的界定逻辑,共同构建了查默斯非还原功能主义视角下的因果角色理论体系。

5. 结论

功能主义历经六十余年发展,始终围绕“因果角色”展开,但仅依靠因果信息处理结构只能解释外在行为,无法涵盖意识独有的主观感受质。这一理论局限恰好对应了当下大模型的哲学困境:AI虽具备

完整的语言因果运算功能, 却缺乏第一人称的主观体验。查默斯的自然主义二元论将“因果角色”与“现象属性”并置, 既保留了功能主义的解释效力, 又为感受质保留了本体论空间。基于此, 本文提出构建一套以疼痛意识为核心, 贯通因果分析、现象内省与脑成像实证的一体化交叉研究框架。首先, 依托阿姆斯特朗与普特南的因果链条理论, 搭建标准化的疼痛因果功能模型, 明确躯体损伤刺激、内在心理状态及行为输出的完整链路, 为人类、生物及 AI 划定统一的功能判定基准。其次, 引入现象学维度, 通过被试的第一人称内省报告采集疼痛的主观特征, 区分单纯的行为模拟与真实的疼痛体验, 补足纯因果模型缺失的主观质性维度。同时, 借助 fMRI 等神经科学手段开展实证校准, 对比真实疼痛与刻意模仿场景下的脑部激活图谱, 并检验异质生物的神经机制, 从而清晰划分物理神经因果与主观现象属性两条解释脉络。最后, 将功能模型推演、现象学内省文本与脑成像数据进行三方互证, 检验功能组织一致是否必然伴随相同的主观感受, 以实证层面回应相关哲学论证, 厘清因果角色与现象属性的边界。这套综合研究模式表明, 单一理论难以完整阐释意识全貌, 唯有融合因果结构分析、第一人称主观现象维度与认知科学量化实证, 才能弥补物理功能与主观体验之间的解释鸿沟, 并为评判人工智能是否具备真正心灵提供兼具哲学严谨性与科学实证支撑的标准。

基金项目

云南省教育厅科学研究基金项目《历史生成视域下边境民族地区从“自然”到“历史”的共同体生成实践研究》(2026Y0359)。

参考文献

- [1] Putnam, H. (1975) The Nature of Mental States. In: Putnam, H., *Mind, Language and Reality: Philosophical Papers: Vol.2*, Cambridge University Press, 325.
- [2] 戴潘. 福多“概念”理论研究[D]: [博士学位论文]. 上海: 复旦大学, 2010.
- [3] Fodor, J.A. (1965) Explanations in Psychology. In: Black, M., *Philosophy in America*, Cornell University Press, 161-179.
- [4] Ryle, G. (1949) The Concept of Mind. Barnes & Noble.
- [5] Lewis, D. (1983) An Argument for the Identity Theory. In: Lewis, D., *Philosophical Papers Volume I*, Oxford University Press, 99-107. <https://doi.org/10.1093/0195032047.003.0007>.
- [6] Ramsey, F.P. (1931) Theories. In: Braithwaite, R.B., *The Foundations of Mathematics and Other Logical Essays*, Kegan Paul, Trench, Trubner & Co, 212-236.
- [7] Armstrong, D.M. (1968) A Materialist Theory of the Mind. Routledge & Kegan Paul.
- [8] Chalmers, D.J. (1996) The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.
- [9] 查默斯. 有意识的心灵[M]. 朱建平, 译. 北京: 中国人民大学出版社, 2013: 78-96+118-135.
- [10] Chalmers, D.J. and Jackson, F. (2001) Conceptual Analysis and Reductive Explanation. *The Philosophical Review*, **110**, 315-360. <https://doi.org/10.1215/00318108-110-3-315>