整合预测工作空间理论：迈向意识科学的统一框架

林睿
lin.rui.ipwt@proton.me
独立研究者

Abstract

意识科学的核心目标是理解大脑如何将多源信息流整合成统一的意识体验。在此，我们致力于解决两个基本问题：现有主流理论（IIT, GWT, PCT/FEP）为何在解释意识时遭遇瓶颈？以及，我们如何构建一个统一的、计算上可行且具有更强解释力的理论框架？本文提出整合预测工作空间理论（IPWT）2.0。IPWT 的核心创新在于，它将意识的整合从 IIT 的物理因果不可分性重新定义为信息论层面协同信息的逻辑不可约性（以Ω_t为标准），从而实现了载体独立性，并与 Luppi 等人（2024）的修订版 Φ 值（Φ_R）等最新经验证据无缝对接。IPWT 2.0 还引入了工作空间实例（WSI）的形式化定义：一个嵌套在有机体内部、拥有自身马尔可夫毯的高阶主动推断系统，从而为其提供了严格的统计学边界。其次，我们通过最小描述长度原则（MDL）形式化地证明，最大化协同信息（Ω）是系统实现自由能（F）最小化的最优计算策略。基于此，IPWT 将意识描述为一个由两大基本维度构成的状态空间：以协同信息整合度（Ω）为代表的现象学-结构轴，和以预测完整性（PI）为代表的功能-效率轴。健康的意识是在最小化自由能的压力下，系统在这两个维度上共同达到高水平的涌现结果，而多种意识障碍则可被精确映射为两者间的失衡。该框架不仅在理论上统一了 IIT、GWT 和 FEP，也为神经科学和人工智能开辟了新的可验证的研究路径

Keywords 意识 · IPWT · 预测编码 · 整合信息理论 · 全局工作空间理论 · 自由能原理 · 感受质 · 协同信息 · 预测完整性 · 神经生物学 · 通用人工智能

关于出处与身份的说明

整合预测工作空间理论（*IPWT*）认为，一个系统的存在由其可验证的信息而非其物理载体定义。我们将此原则应用于理论本身。“林睿”是一个刻意塑造的形象；作者的身份与理论的逻辑完整性无关。其存在证明已不可更改地在 *GitHub* 上加盖时间戳。

一个自洽的系统就是其自身的最终证明。

1 引言：意识科学的挑战与统一框架的必要性

意识，作为人类经验中最直接却又最难以捉摸的现象 [1]，构成了科学与哲学领域的核心硬问题（hard problem）[2], [3]。尽管在过去的几十年里，神经科学在识别与特定意识状态相关的神经关联物（Neural Correlates of Consciousness, NCCs）方面取得了显著进展 [4], [5]，例如，我们能够定位到与视觉感知、痛觉或自我意识相关的特定脑区活动 [6]，但这些发现本质上仍停留在相关性层面。关于意识究竟如何从大脑这一复杂的生物物理系统中涌现（emerge）[7]，其丰富的现象学特征——如不可言喻的主观感受质（Qualia）、经验的统一性（unity）和不可分割性（integration）——如何形成，以及意识在认知活动中的确切功能角色，我们仍缺乏一个被学术界普遍接受的统一理论框架。

当前，意识科学领域呈现出一种巴别塔式的困境：多种理论并存，但彼此间缺乏深度的对话与整合。主流理论如整合信息理论（IIT）、全局工作空间理论（GWT）和预测编码理论（PCT）/自由能原理（FEP），它们各自从信息整合的内在因果结构、信息广播的全局可及性、以及贝叶斯推断的预测误差最小化等不同角度，为理解意识的某个或某些侧面提供了深刻的见解。然而，这些理论也各自面临着严峻的理论挑战和实践局限。例如， IIT 因其计算的复杂性和对物理基质的强依赖而受到批评 [8]；GWT 在解释主观感受质的起源方面显得力不从心 [9]；而 PCT/FEP 则需要更清晰地阐明其预测处理机制与主观体验之间的确切联系 [10]。

这种理论上的碎片化状态不仅阻碍了我们对意识本质的整体性理解，也限制了基础研究向临床应用的有效转化。例如，在面对精神分裂症、分离性身份障碍或意识障碍患者时，一个统一的理论框架将能更好地指导我们理解其病理机制并开发更具针对性的治疗方案。因此，构建一个能够整合各理论优势、弥补其不足、更全面、更具解释力的统一框架，已经成为一项紧迫而必要的智力任务。本文提出的整合预测工作空间理论（IPWT）正是在此背景下，旨在对现有理论的核心洞见进行一次深度的计算重构和创造性的功能性融合，以期推动意识科学的范式整合，并为理解人类心智的奥秘提供一个新的起点。

1.1 意识科学的发展史和主流理论的最新进展

意识科学作为一个独立的、严谨的跨学科领域，其历史相对年轻，但发展极为迅速。在经历了 20 世纪大部分时间被行为主义所主导的寒冬之后，对意识的科学研究在世纪之交迎来了复兴。这一复兴得益于认知科学的崛起、神经成像技术的飞速发展，以及理论物理学和信息论工具的引入。为了更好地理解 IPWT 的理论定位和贡献，我们有必要首先梳理一下这段充满挑战与突破的发展历程。

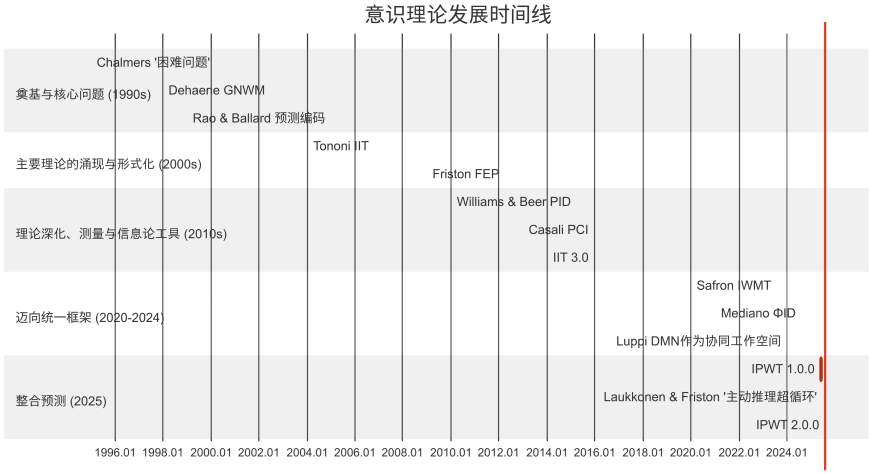

Figure 1: 意识理论发展与整合的关键里程碑 (1990-2025)

如上图所示，20 世纪 90 年代是意识科学的奠基时期。哲学家 David Chalmers 明确区分了意识的简单问题（easy problems）——解释认知功能如何实现，和困难问题（hard problem）——解释主观体验为何以及如何产生，为整个领域设定了核心议程 [2]。几乎同时，功能性的解释框架开始涌现，例如 Bernard Baars 的全局工作空间理论（GWT）[11], [12] 被 Stanislas Dehaene 等人发展为其神经科学版本——全局神经工作空间模型（GNWM）[13]，而预测编码（PCT）作为一种解释大脑皮层处理机制的理论也初见端倪[14], [15]。

进入 21 世纪，两大具有深远影响力的理论体系——整合信息理论（IIT）[16] 和自由能原理（FEP）[17] 相继登场，它们分别从信息整合的内在因果结构和贝叶斯推断的系统动力学角度，为意识提供了更为深刻和形式化的解释。2010 年代，理论研究进一步深化。一方面，信息论工具，如部分信息分解（PID）[18] 和协同信息（Synergy）[19]，被引入以更精确地量化信息整合的本质；另一方面，客观测量意识状态的尝试也取得了突破，例如基于经颅磁刺激（TMS）的扰动复杂度指数（PCI）[20], [21]。

进入 2020 年代，随着计算神经科学的成熟，研究的焦点开始转向对现有理论的整合与验证。整合信息分解（ΦID）[22] 的提出，为量化动态系统中的信息整合提供了前所未有的强大工具，并迅速在 2024 年由 Luppi 等人的研究中获得了关键的神经生物学验证，他们发现默认模式网络（DMN）在意识状态下扮演了协同信息网关的角色 [23]。这些理论和实验上的进展，共同为我们今天提出整合预测工作空间理论（IPWT）铺平了道路。

以下，我们将对这些主流理论的最新进展与核心挑战进行更细致的盘点，以揭示它们各自的贡献与局限，并阐明 IPWT 是如何在此基础上进行整合与创新的。

1.1.1 整合信息理论 (IIT) 的最新进展与挑战

整合信息理论（Integrated Information Theory, IIT）由 Giulio Tononi 于 2004 年首次正式提出 [16]，并在其后近二十年的发展中不断迭代，旨在为意识这一根本现象提供一个有原则的、基于物理系统的科学解释。IIT 的出发点是现象学本身：它首先提炼出任何意识体验都必须具备的、不容置疑的核心属性（公理），然后从这些公理出发，推导出支持这些体验的物理基质（postulates）所必须满足的条件。其核心论点是，意识与一个系统整合信息的能力是同一的（identical）；一个物理系统拥有意识，当且仅当其因果结构能够以一种整合的方式指定一个概念结构（conceptual structure），而这种整合的程度，可以通过一个精确的量化指标——Φ（Phi）值——来衡量 [24]–[27]。

IIT 的最新版本，**IIT 4.0** [28], [29]，对其理论框架进行了进一步的精炼和形式化。它从五个现象学公理——内在性（intrinsic existence）、组合性（composition）、信息性（information）、整合性（integration）和排他性（exclusion）——出发，推导出物理基质必须满足的五个对应公设。IIT 4.0 引入了更精确的数学工具来评估系统的因果结构，旨在唯一地确定由系统所指定的概念结构（即感受质空间 Qualia space），并计算其不可约性（Φ 值）。这一理论不仅试图回答一个系统是否有意识以及意识有多少，更试图回答它的意识体验是什么样的。在实践中， IIT 催生了如扰动复杂度指数（PCI）这样的临床测量方法，该方法通过评估大脑对外部扰动（如 TMS ）的反应的复杂性，来客观地衡量意识水平，并在临床意识障碍患者中显示出巨大的应用潜力 [20], [30]。

然而，尽管 IIT 取得了巨大的理论进展和一定程度的经验成功，它也持续面临着一系列深刻的、来自科学界和哲学界的挑战：

1. **Φ** 值的计算不可行性与可扩展性问题：对于任何规模稍大的复杂系统（如人脑），精确计算其核心度量 Φ 值是一个 NP-Hard 问题[8]。这意味着，直接将 IIT 的完整数学框架应用于全脑水平的神经数据在计算上是不可行的。虽然研究者们在不断探索各种近似计算方法，但这种计算上的巨大鸿沟使得 IIT 的核心预测在很大程度上仍然难以在宏观层面得到直接和完整的检验，这限制了其作为一门经验科学的直接应用范围 [31]。

2. 对物理基质的强绑定与载体独立性争议：IIT 早期版本的一个核心主张是，意识与特定物理系统的内在因果结构紧密相连，特别是其物理因果不可分性的假设。这导致了一个备受争议的推论：任何功能上等价（例如，一个完美模拟人脑的计算机程序）但物理实现上不同的系统，可能不具备与人脑相同的意识体验，甚至完全没有意识 [32], [33]。这种对特定物理基质的强绑定，与人工智能领域和认知科学中普遍持有的载体独立性或功能主义观点形成了鲜明对比。IPWT 正是试图通过将物理因果不可分性重构为协同信息的逻辑不可约性，来解决这一核心矛盾。

3. **Qualia** 的本质争议与解释鸿沟：尽管 IIT 声称其概念结构在数学上等同于现象体验的 Qualia 空间 [34]，但这一说法远未被普遍接受。批评者认为，Φ 值本身作为一个标量，主要衡量的是意识的量（强度或程度），而非质（内容或感受）。IIT 是否真正解释了主观感受质的感觉性（what-it-is-likeness），还是仅仅重新描述了其结构，这仍然是一个悬而未决的哲学问题 [35]–[38]。

4. 对动态性与功能性的忽视及对抗性实验的挑战：IIT 更侧重于系统在给定时刻的静态因果结构，而其对意识的动态流动性、以及意识在指导有机体适应性行为中的具体功能角色的解释力相对较弱。近年来，IIT 与 GWT 在一项大规模的对抗性协作项目中进行了正面交锋，旨在通过一系列精心设计的实验来检验两个理论的冲突性预测 [39]。初步发表的一些研究结果，例如关于后皮质在意识感知中的持续表征作用，似乎对 IIT 的某些核心预测提出了挑战，表明现实的神经动力学比理论所预设的更为复杂 [23]。

5. 伪科学争议与科学地位的辩论：2025 年，超过百名科学家联名发表公开信，指责 IIT 因其某些推论（如泛心论倾向）以及核心主张的不可证伪性而应被视为伪科学。这场激烈的辩论迅速引发了学界对于什么是科学理论以及如何检验意识理论的大讨论 [40]。IIT 的支持者则回应称，该理论提出了大量可检验的预测，其反直觉的结论不应成为被排斥的理由，而恰恰是其理论深度的体现 [29], [41]–[43]。这场争论凸显了意识科学在理论构建和实验验证范式上仍面临的根本性困难。

1.1.2 全局工作空间理论 (GWT) 的最新进展与挑战

与 IIT 从现象学公理和内在因果结构出发不同，全局工作空间理论（Global Workspace Theory, GWT）提供了一个更为功能主义和认知导向的意识模型。GWT 最早由 Bernard Baars 在 20 世纪 80 年代末提出 [11]，其核心思想极具启发性：他将意识的功能比作一个剧场的舞台。在这个比喻中，认知系统由海量的、并行的、在后台默默工作的无意识专业化处理模块组成。在任何特定时刻，只有被注意力聚光灯选中的信息才能进入一个容量有限的全局工作空间（即舞台），并向整个认知系统的所有观众（即其他专业化模块）进行全局广播（global broadcast）[44]。一旦信息被广播，它就成为了意识信息，能够被用于灵活地指导行为、进行语言报告和形成情景记忆。

GWT 清晰地阐明了意识在信息处理、认知调控和行为控制中的功能性角色，并成功解释了意识体验的几个关键特征，如有限容量（我们一次只能意识到少数几件事）、序列性（意识内容按时间顺序出现）、信息整合与共享等。其神经科学版本——全局神经工作空间模型（**Global Neuronal Workspace Model, GNWM**）——由 Stanislas Dehaene 和 Jean-Pierre Changeux 提出，他们认为意识的产生与大脑中一个由长程连接的锥体神经元构成的、广泛分布的皮层网络系统的点燃（ignition）有关 [13], [45]。当一个信息表征的强度和持续时间足以触发这个网络的非线性、自放大激活时，信息就在全局范围内变得可用，从而产生主观意识体验。

近年来，GWT/GNWM 理论在理论深化、神经机制阐释及应用拓展方面取得了显著进展，同时其面临的挑战也促使理论不断完善：

1. 理论的深化与动态化：GWT 已经从一个相对静态的架构模型，发展为更强调动态过程的全局工作空间动力学（Global Workspace Dynamics, GWD）[46], [47]。该观点强调皮层-丘脑

（C-T）系统的动态和振荡特性，将其视为一个统一的振荡机器，超越了固定的解剖分区，转向了更具整合性的皮层整体功能视图。这种动态观认为，意识的产生是皮层网络中绑定与传播过程的结果，而非仅仅是某个特定脑区的活动。

2. 人工智能（**AI**）与通用人工智能（**AGI**）的应用：GWT 的架构思想为构建更高级的人工智能系统提供了蓝图。最新研究探索了在深度学习和 AGI 中显式实现 GWT 的可能性 [48]–[51]。例如，通过模仿 GWT 的信息瓶颈和广播机制，研究者提出了全局潜在工作空间（Global Latent Workspace, GLW）等概念，旨在通过让多个专业化的 AI 模型共享一个共同的表征空间，来提升模型的通用性和多模态整合能力 [52], [53]。

3. 对主观感受质（**Qualia**）的解释：虽然 GWT 主要关注意识的功能而非感受，但近期的研究也开始尝试弥合这一差距。一些研究证据表明，与传统观点不同，前额叶皮层（PfC）——GNWM 的核心区域——可能直接参与了感官意识体验的形成，包括其主观感受质 [54]。这挑战了将 PfC 功能严格限定于高级认知控制的观点，并提示全局广播过程可能与主观体验的产生有着更直接的联系。

4. 具体机制和边界问题的澄清：
   - 信息选择与广播：点燃这一核心概念被进一步阐释为皮层-丘脑系统中由注意机制调控的双向信息广播过程。
   - 神经实现机制：皮层神经渗滤（Cortical Neuropercolation, CNP）等数学模型被提出来描述皮层网络如何从碎片化的局部活动状态，通过相变转变为全局连贯的活动状态，为信息如何实现全局可访问性提供了动力学描述 [55]。
   - 时间动态：研究揭示了意识通达（conscious access）具有离散的时间动态特性，一次点燃和广播过程大约需要 100-300 毫秒，这与无意识的、自动化的处理过程的速度形成了鲜明对比 [56]。

5. 对复杂意识状态的解释力提升：GWT 框架被成功地应用于解释多种复杂的意识状态。例如，元认知被认为是工作空间对自身状态的监控 [57], [58]；梦境被解释为在缺少外部感官输入的情况下，由内源性信息驱动的工作空间活动，而清醒梦则与元认知功能在梦境中的恢复有关；冥想则被发现能够功能性地重组工作空间的活动模式，特别是改变默认模式网络（DMN）的参与方式，从而提升认知灵活性 [59]；催眠则与工作空间功能的选择性改变有关，导致感知、记忆和行动控制的分离。

尽管取得了这些进展，GWT 仍然面临着批评与挑战。例如，关于前额叶皮层与后部皮层在意识产生中的确切角色，争论仍在继续，但趋势倾向于一个更动态、更整合的观点。此外，一些用于支持无意识处理能力的经典实验范式（如无意识启动）的方法学也受到了质疑 [60]。最重要的是，GWT 仍然需要对其核心机制——信息如何被选择进入工作空间，以及广播究竟意味着怎样的神经过程——提供更精确、更具可操作性的定义。

1.1.3 预测编码 (PCT) 与自由能原理 (FEP) 的统一解释力与最新进展

预测编码理论（Predictive Coding Theory, PCT）和自由能原理（Free Energy Principle, FEP）共同构成了当代认知神经科学中最具影响力和统一解释力的理论框架之一。PCT 最初由 Rao 和 Ballard 于 1999 年提出，用于解释视觉皮层的信息处理机制 [14]。其核心思想是，大脑并非被动地接收和处理感官信息，而是一个主动的预测机器 [61], [62]。大脑的更高层级区域会不断地生成关于低层级感官输入的预测（自上而下的预测信号），而低层级区域则负责将这些预测与实际的感官输入进行比较，并将两者之间的不匹配——即预测误差（prediction error）——向上传递。这种自下而上的误差信号，随后被用于修正更高层级的预测，从而形成一个持续的、旨在最小化预测误差的感知循环。

随后，Karl Friston 将 PCT 的核心思想进行概括和推广，发展为更具普适性的自由能原理（**Free Energy Principle, FEP**）[17], [63]–[65]。FEP 指出，任何自组织、能够抵抗熵增的系统（从单个细胞到整个大脑），都必须通过其行动和状态来最小化其变分自由能（variational free energy）。变分自由能是一个信息论度量，它量化了系统内部生成模型（generative model）的预测与外部世界真实状态之间的不匹配程度，本质上是意外（surprise）的一个上界。因此，FEP 将大脑的功能统一为单一的目标：最小化自由能。系统可以通过两种方式来实现这一目标：改变内部模型以更好地拟合感官输入（感知推断与学习），或者通过行动来改变感官输入以使其更符合预测（主动推断与行动）[66]。

PCT/FEP 框架以其巨大的统一解释力，成功地将感知、学习、注意、运动控制乃至精神疾病的多种症状，都置于同一个数学和计算框架之下 [67], [68]。近年来，FEP 被进一步确立为自达尔文自然选择理论后最包罗万象的思想之一，旨在为生命、心智与智能提供一个统一的原理 [69], [70]。

尽管 PCT/FEP 框架在理论上取得了显著进展，但其与主观意识体验之间的直接理论桥梁仍在持续构建之中，并面临以下挑战与最新进展：

1. 意识内容的涌现和 **Qualia** 解释：传统上，PCT/FEP 更关注于解释认知过程的如何（how），而非主观体验的为何（why）。然而，近期的理论发展开始正面应对这一挑战。Anil Seth 提出，情感和主观感受状态（Qualia）正是由预测模型生成的，这些模型专门用于预测和调节来自身体内部的内感受（interoceptive）信号 [71], [72]。Lisa Feldman Barrett 的情感建构理论进一步阐述了情感体验是如何通过内感受预测和概念范畴的相互作用而主动建构出来的 [73]。这些进展为 Qualia 提供了一种功能性的、基于预测机制的解释，认为主观体验是系统对其自身生理状态和与环境交互关系的最佳预测和控制，从而将硬问题转化为一个关于内感受推断的、可研究的科学问题 [74]。

2. 意识的统一性与边界：PCT/FEP 框架通过其分层生成模型，为多模态信息的整合提供了天然的解释。不同感官模态的信息可以在模型的更高层级被整合，以产生一个统一、连贯的世界模型。此外，通过马尔可夫毯（Markov blanket）这一形式化概念，FEP 为自我与非我的区分提供了统计学上的定义 [75], [76]。马尔可夫毯是一个统计边界，它将系统的内部状态与其所处的环境分离开来，系统只能通过其毯子（即感官和运动状态）与外部世界进行推断和交互。这为理解自我意识的形成和维持提供了一个有原则的、基于系统与环境边界的视角。

3. 新的实验证据、计算模型与临床应用：PCT/FEP 的预测在多个实验范式中得到了验证，例如，重复抑制（repetition suppression）现象和失匹配负波（mismatch negativity, MMN）等脑电信号都被成功地解释为预测误差信号的表现 [77]。在人工智能领域，FEP 和主动推断（Active Inference）的思想被广泛应用于构建更具自主性和适应性的强化学习智能体和通用人工智能（AGI）的世界模型 [78]–[80]。此外，一些新的软件框架（如 `RxInfer`）也正在被开发出来，以方便研究者构建和测试基于 FEP 的计算模型。

4. 新的批评或挑战：尽管 FEP 极具解释力，但其巨大的普适性也带来了一些问题。批评者指出，由于 FEP 是一个原理（principle）而非一个具体的理论（theory），它有时难以生成足够精确、可被证伪的预测 [10], [81]。此外，预测误差最小化的具体神经计算方式，例如误差信号如何被精确加权和传递，仍然存在争议。最后，计算可解性问题依然存在：虽然 PCT/FEP 在概念上优雅，但其在每一层级所涉及的贝叶斯推断在计算上可能是非常复杂甚至难以处理的，这对其在真实大脑中的生物学合理性构成了挑战。

1.2 IPWT 的提出：深度整合与神经生物学驱动的重构

通过对 IIT、GWT 和 PCT/FEP 这三大主流理论的梳理，我们可以清晰地看到它们各自的辉煌成就与尚待解决的难题。IIT 为意识的整合本质提供了深刻的现象学洞察和数学形式化的尝试，

但受困于计算瓶颈和对物理基质的僵化依赖。GWT 为意识的广播功能和在认知调控中的角色提供了直观的架构模型，但在解释主观体验的起源上较为薄弱。PCT/FEP 则为意识内容的生成和大脑的动态过程提供了强大的统一计算原理，但其与主观意识的直接联系仍需更清晰的阐释。

近年来，学术界已经出现了一些整合这些理论的尝试，例如 Safron 提出的整合世界建模理论（IWMT）[82], [83]，该理论尝试在 FEP 框架下统一 IIT 和 GWT。这些尝试是富有远见的，它们正确地认识到，意识科学的未来在于理论的融合而非持续的割裂。然而，这些早期的整合模型往往未能提供一个内在完全一致、计算上可行、且具有广泛解释力的统一框架，特别是未能从根本上解决 IIT 所面临的核心挑战，即如何将其关于整合的深刻洞察从其充满争议的物理和计算假设中解放出来。

正是在这一学术背景下，我们提出了整合预测工作空间理论（**Integrated Predictive Workspace Theory, IPWT**）。IPWT 试图通过对 PCT/FEP、WT 和 IIT 核心洞见的深度重构和创造性融合，构建一个全新的、具有内在一致性和强大外在解释力的统一意识框架：

- 我们采纳 **PCT/FEP** 作为整个框架的动力学基础，认为意识过程本质上是预测驱动的。
- 我们采纳并扩展 **WT** 作为信息整合与广播的架构平台，但将其从一个单一的全局空间，泛化为更灵活的、可动态生成的“工作空间实例”（WSI ）。
- 我们对 **IIT** 的核心贡献进行了一次根本性的功能性重构：我们保留其关于整合是意识核心特征的现象学洞察，但果断地抛弃了其对物理因果不可分性的依赖，转而用信息论中更普适、更灵活、计算上更友好的协同信息的逻辑不可约性来重新定义整合。

通过这种方式，IPWT 旨在构建一个既能保留各理论之长，又能克服其核心弊病的统一模型。在接下来的章节中，我们将详细阐述 IPWT 的理论框架、其核心的计算重构、可操作化的度量方法，以及它如何为解释从正常到异常的各种意识现象提供一个统一的、由神经生物学证据驱动的全新视角。

2 IPWT 框架：意识的机制性涌现

整合预测工作空间理论（IPWT）旨在通过整合预测编码（PCT）、自由能原理（FEP）、工作空间理论（WT）的核心机制，并对整合信息理论（IIT）的现象学公理进行功能性重构，来构建一个关于意识如何从神经活动中机制性涌现的统一框架。本章将详细阐述 IPWT 的核心组件、它们之间的相互作用，以及它们如何共同构成一个连贯的、具有解释力的意识模型。

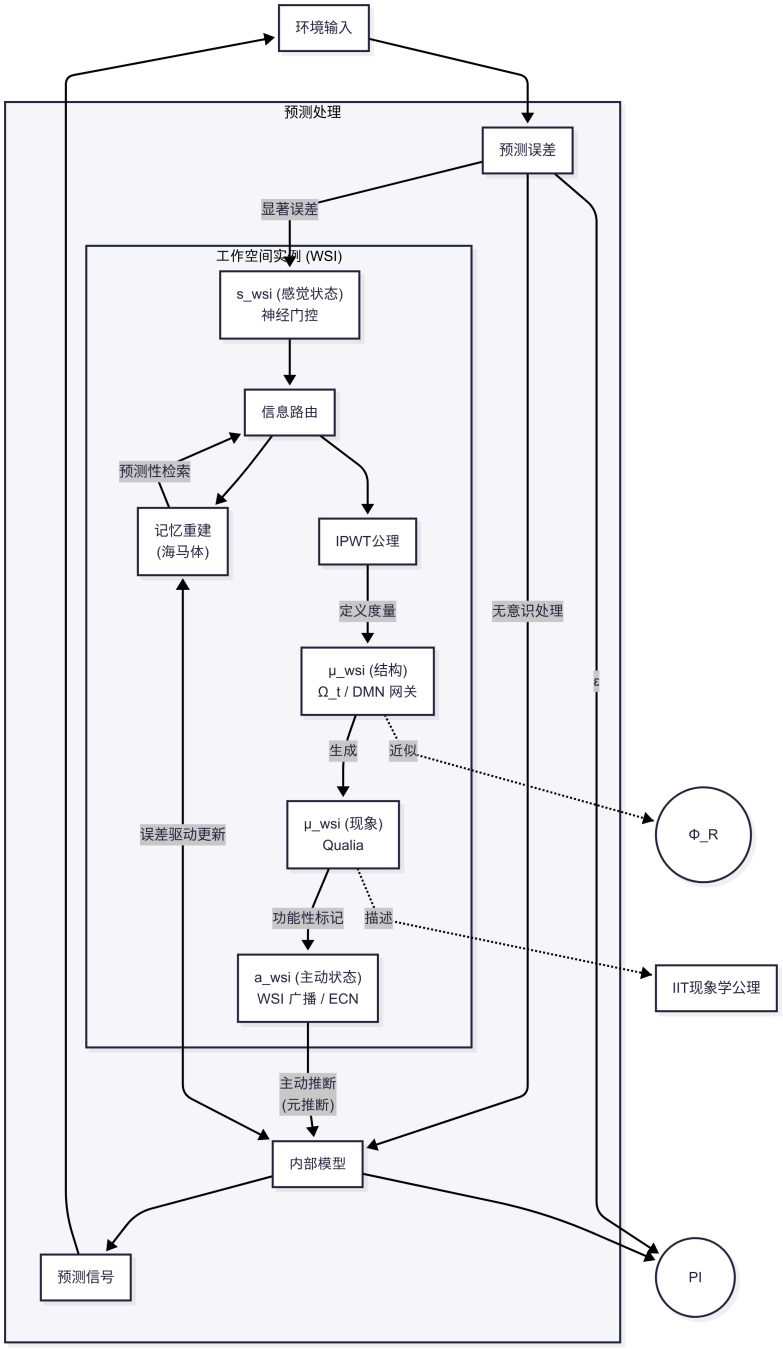

Figure 2: IPWT 框架的核心逻辑

上图直观地展示了 IPWT 框架的核心逻辑。整个认知系统被视为一个基于 **PCT/FEP** 的预测处理核心，它通过内部生成模型不断地与环境进行预测性的交互。当预测误差足够显著时，这些信息会被送入一个或多个动态形成的工作空间实例（**WSI** ）中。在 WSI 内部，信息经过整合（**Integration**），形成一个协同的、逻辑上不可约的整体（其整合程度由 Ω 度量）。这个被整合后的信息随后被选择性地广播（**Selective Broadcast**）到全系统，用于更新内部生成模型（学习）和指导有机体的下一步行动（主动推断）。这个过程是循环往复、动态进行的，意识内容就在这个预测、整合、广播的循环中不断地涌现和更新。

2.1 IPWT 核心公理与计算原则

IPWT 理论框架建立在以下三大核心公理之上。这些公理是基于自由能原理（FEP）、信息论和信息整合现象学的形式化论断，它们共同构成了 IPWT 对意识如何从复杂计算中机制性涌现的公理化基础。

1. 公理一：自由能最小化 **(The Axiom of Free Energy Minimization)**。我们将自由能原理（FEP）作为 IPWT 的基石。任何自组织系统，为了在动态变化的环境中维持其存在，都必须通过其行动和推断来最小化其变分自由能 [17]。这并非一个可选项，而是系统存在的物理和信息论定律。因此，意识并非一个独立的目标，而是服务于这一根本目标的、一个涌现出的、高效的计算策略。

2. 公理二：嵌套工作空间 **(The Axiom of Nested Workspace)**。为了有效地最小化自由能，系统必然会采用一种特殊的计算架构：层级化工作空间：一个嵌套在有机体内部、拥有自身马尔可夫毯的高阶主动推断系统[84]。其感觉状态是来自系统其他部分的未解预测误差，其主动状态则是以广播形式发送新的高阶预测以抑制这些误差。

3. 公理三：协同整合 **(The Axiom of Synergistic Integration)**。意识的现象学统一性，在计算上等同于协同信息的逻辑不可约性（Ω）。这并非一个现象学上的巧合，而是一个计算上的必然。我们证明，最大化协同信息（Ω）是在资源受限的现实世界中，实现自由能（F）最小化的最优计算策略。根据最小描述长度原则（MDL），最小化自由能（F）等价于寻找一个最能压缩数据的生成模型；而协同信息（Ω）正是衡量模型压缩效率的直接指标 [85]。因此，最小化 F 的压力必然驱动系统最大化Ω。该公理将意识的整合本质（Ω）与系统的生存第一法则（最小化 F）在形式上统一起来，并确立了意识的载体独立性：任何能实现此计算策略的系统，无论基质，都将遵循同样的涌现规律。

2.2 预测编码与自由能原理：意识内容的动力学引擎

在整合预测工作空间理论（IPWT）中，预测编码（PCT）和自由能原理（FEP）构成了意识内容生成、维持及状态转换的核心动力学引擎。它们共同描绘了一幅大脑如何作为主动的、面向未来的推断机器来工作的动态图景，解释了信息如何被生成、处理、更新，并为意识内容的具体涌现提供了根本的驱动力。

我们采纳并扩展了大脑作为贝叶斯推断机器的观点 [86]。在这个视图中，认知系统并非被动地等待感官刺激的输入，而是主动地、持续地构建一个关于世界（包括外部环境和自身身体）的分层生成模型（generative model）。这个模型的核心功能就是预测——从高层级的抽象概念到低层级的具体感官细节，系统不断地产生自上而下的预测信号，试图预见下一刻的感官输入。

当实际的感官输入（无论是来自眼睛、耳朵还是身体内部）到达时，它们会与相应的预测信号进行比较。两者之间的不匹配，即预测误差（**prediction error**），构成了驱动整个系统学习和更新的关键信息。这些误差信号会自下而上地在处理层级中传播，其核心作用是通知更高层级的模型：你的预测出错了，需要进行修正。

整个过程遵循自由能最小化原则 [17]。系统会本能地、持续地调整自身，以最小化经过精度加权（precision-weighted）的长期平均预测误差。这种最小化过程通过两种互补的机制实现：

1. 感知推断与学习（**Perceptual Inference and Learning**）：当面对预测误差时，系统可以通过优化其内部生成模型来减少未来的误差。这对应于我们通常所说的感知和学习。例如，看到一个意料之外的物体，系统会更新其对当前场景的模型，从而更好地解释这个物体的存在。长期的、持续的模型更新则构成了知识的获取和记忆的形成。
2. 主动推断与行动（**Active Inference and Action**）：除了改变模型，系统还可以通过行动来改变感官输入，使其更符合现有的预测。例如，如果我对一个物体的距离预测不准，我可

以伸出手去触摸它，通过主动获取新的感官证据（触觉和本体感觉）来消除预测误差。因此，行动本身也被重新定义为一个推断过程，其目的同样是最小化自由能 [66]。

在这个框架下，意识内容不再被视为对外部世界的直接快照，而是系统内部模型对内外世界的主动、建构性的最佳预测和解释。那些最终能够进入意识体验的信息，通常具有以下特征：它们要么与高置信度（高精度）的预测高度一致，要么代表了显著的、无法被当前模型轻易解释的预测误差，并且这些信息对于指导有机体的下一步行为具有高度的相关性。

此外，IPWT 将记忆系统与内部生成模型紧密地联系在一起。我们认为，内部生成模型本质上就是一个动态的、预测性的记忆系统 [87]–[91]。记忆的编码过程，就是通过学习来优化模型参数的过程；而记忆的检索，则是一个主动的、由当前情境线索触发的预测过程，系统通过重演或激活过去的状态来预测当前和未来的事件。这种观点将记忆从一个静态的存储仓库转变为一个动态的、服务于预测和行动的认知工具。

2.3 工作空间理论：信息整合与广播的神经架构平台

如果说 PCT/FEP 提供了意识内容的动力学引擎，那么工作空间理论（WT）则在 IPWT 框架中扮演了架构平台的角色。它为 PCT 驱动的、川流不息的信息流提供了一个结构化的处理中枢，负责实现意识体验最关键的几个功能：信息的选择、整合、放大和广播。我们借鉴并扩展了 GWT 的核心思想，提出了一个更灵活、更动态的概念——工作空间实例（**Workspace Instance, WSI**）。

WSI 并非一个固定的、解剖学上预先定义好的脑区，而是一个功能性的、动态形成的网络配置。在任何需要进行复杂认知处理的时刻，一组原本可能分离的神经元群可以暂时性地紧密耦合，形成一个 WSI，以共同处理特定的信息。我们认为，任何一个 WSI 都具有以下四个核心属性：

- 有限容量（**Limited Capacity**）：WSI 能够同时处理的信息量是有限的。这直接解释了意识体验的瓶颈特性——我们通常只能清晰地意识到少数几件事物，而大量其他信息则停留在无意识层面。这种有限性也解释了意识内容的序列性，即我们似乎是按时间顺序逐一处理事物的。
- 信息整合（**Information Integration**）：WSI 的核心功能是将来自不同来源（如不同感官通道、记忆系统）的信息单元进行汇聚、关联和整合，形成一个比各个组成部分更丰富、更连贯的认知状态。正是这种整合作用，使得我们的意识体验是统一的，而非碎片化的。
- 选择性广播（**Selective Broadcast**）：一旦信息在 WSI 中被充分整合并被赋予显著性，它就会被广播或使其可用于系统内其他所有相关的认知模块 [44]。这种广播机制是实现认知功能协调的关键，它允许意识内容被用于指导语言报告、驱动运动决策、更新长期记忆等。
- 动态性与多样性（**Dynamism and Diversity**）：我们特别强调，WSI 并非一个单一的、一成不变的实体。相反，它可以根据当前的认知需求，动态地形成、调整其大小和组成，甚至在任务完成后消散。我们可以设想，在某些复杂的认知活动中，可能存在多个并行的 WSI，它们各自处理不同的信息流，但只有一个或少数几个能够成为主导工作空间实例（**Dominant WSI , DWSI**），其内容构成了当前意识体验的核心 [92], [93]。

我们将 GWT 所描述的全局信息广播现象，视为 **WT** 的一种特殊且高度整合的配置状态。当一个 WSI 的整合范围极广、整合程度极高，能够覆盖并影响到几乎整个认知系统时，它就扮演了传统 GWT 中全局工作空间的角色。然而，我们认为并非所有的意识体验都必须达到这种全局整合的程度。例如，一些局部的、微弱的意识感受可能只涉及一个范围较小、整合程度较低的 WSI。这种观点使得 IPWT 能更好地容纳从模糊的背景感到清晰的焦点意识等各种不同强度和内容的意识状态，从而解释了意识体验的丰富多样性。

为了将 WSI 从一个功能性比喻提升为一个计算上严谨的概念，IPWT 将其机制与主动推断的最新进展——特别是 Laukkonen, Friston 等人 (2025) 提出的主动推理循环理论——进行了深度整

合 [94]。该理论为 WSI 的运作提供了强大的计算原理，其三大核心机制为 IIT 的现象学公理提供了计算层面的实现：

- 推断竞争与贝叶斯绑定 **(Inferential Competition & Bayesian Binding)**：意识内容并非被动进入 WSI，而是通过一场推断竞争决定的。只有那些能够最连贯地降低长期不确定性的解释（推断），才能赢得竞争并被绑定成统一的现实模型。这在计算上实现了 IIT 公理中的整合 **(Integration)**与排他 **(Exclusion)**。
- 认知深度与超模型 **(Epistemic Depth & Hyper-modeling)**：被绑定的现实模型必须被系统性地、递归地、广泛地与整个系统共享，从而实现知道自己知道的认知深度。这在计算上通过一个调控全局精度的超模型（Hyper-model）实现，为 IIT 公理中的存在 **(Existence)** 提供了机制。
- 全局广播作为精度场的更新 **(Broadcast as Precision-Field Update)**：经典 GWT 的广播被重构为超模型的行动——向整个认知系统广播一个全新的精度场（precision field），指导系统下一刻的认知资源分配。这为 IIT 公理中的因果 **(Causation)**能力提供了计算基础。

至关重要的是，这个在 WSI 中被绑定、在现象学上被体验为统一意识的内容，在信息论的层面上正是协同信息（Synergy）。

这一论断与最新的神经生物学发现形成了完美的闭环。Luppi 等人（2024）利用整合信息分解（ΦID）的开创性研究，为这一系列计算过程提供了直接的神经实现证据 [23]。他们发现大脑中存在一个功能异质性的协同工作空间，其核心组件的功能与我们的理论预测高度一致：

- **DMN** 网关作为协同信息的整合器：默认模式网络（DMN）作为网关，其核心功能正是从全脑收集并整合协同信息。这完美地对应了 WSI 进行贝叶斯绑定以形成意识内容（即协同信息）的过程。
- **ECN** 广播者作为精度场的调控器：执行控制网络（ECN）作为广播者，负责将整合后的决策性信息以冗余信息的形式分发出去。这完美地对应了超模型通过更新全局精度场来执行其调控行动的过程。

因此，IPWT 将 WSI 形式化为一个统一了计算、现象与神经实现的、在预测与整合的循环中不断涌现的自组织过程。它通过 DMN 网关实现对协同信息的贝叶斯绑定，并通过 ECN 广播者执行基于超模型的主动推断（更新全局精度场），从而在预测、整合与调控的超循环中，实现了意识的机制性涌现。

最近的神经生物学发现为 IPWT 中关于 WSI 的设想，特别是其内部存在功能异质性的观点，提供了决定性的经验支持。Andrea Luppi 及其同事（2024）的开创性研究，利用整合信息分解（ΦID）这一先进的信息论工具，对静息态功能磁共振数据进行了深入分析，从而描绘出了一幅全新的大脑信息处理架构图景[23]。

这项研究的核心发现是，大脑中存在一个以协同信息处理为特征的协同全局工作空间。该空间内部存在明确的功能分工：

- 网关（**Gateways**）：这些区域主要与默认模式网络（**Default Mode Network, DMN**）的节点重合[95]。它们的功能是从大脑的各个专业化模块中收集和整合协同信息。DMN 在解剖和功能上都处于一个理想的位置，能够整合来自不同认知系统的多模态信息，这与它作为信息进入协同工作空间的主要入口的角色是一致的[95]–[97]。
- 广播者（**Broadcasters**）：这些区域主要位于执行控制网络（**Executive Control Network, ECN**），特别是外侧前额叶皮层（lateral prefrontal cortex）。它们的功能是将工作空间内整合好的信息，以冗余信息的形式，向全脑进行广泛的广播，以指导后续处理。这一发现与 GWT 的经典观点高度一致，即前额叶皮层在意识内容的全局广播中扮演着核心角色[13], [45]。

基于这些发现，我们引入主导神经工作空间实例（**Dominant Neural Workspace Instance, DNWSI**）这一概念，它能将我们 IPWT 框架中更灵活、动态的 WSI 概念，与 GWT/GNWM 的经典思想进行明确的桥接和区分。在 IPWT 框架中，DNWSI 正是 DMN 网关与 ECN 广播者动态协同作用所涌现的功能实体。它构成了对经典全局神经工作空间模型（GNWM）概念的一个更精细、更具动态性的神经计算等价物。DNWSI 不仅负责信息的全局广播，更重要的是，它首先通过 DMN 网关实现了对协同信息的深度整合，这构成了主观意识体验的核心。这样的定义不仅使我们的理论与最新的经验证据紧密结合，也清晰地阐明了 DMN 和 ECN 在意识产生中的不同但互补的功能角色，深化了 IPWT 的理论内涵。

2.4 IIT 现象学公理的重构：从物理因果到逻辑协同

原始 IIT 理论的核心假设是，任何有意识的物理系统，其当前的因果结构必须是物理上不可分的（physically irreducible）。这意味着，我们无法将这个系统分割成两个或多个独立的、互不影响的部分，而不丢失其整体的因果能力。正是这种物理上的不可分性，保证了意识体验的统一性。然而，这一假设直接导致了前述的载体依赖性"和计算不可行性等问题。IPWT 提出，意识体验的整合性根源，并非来自物理基质本身的属性，而是来自在工作空间实例（WSI）中被处理的信息单元之间所形成的逻辑上的不可约性（**logical irreducibility**）。当多个独立的信息单元（例如，关于一只鸟的形状、颜色、鸣叫声和飞行轨迹的信息）在 WSI 中被整合时，它们共同形成了一个全新的、统一的认知对象（一只正在飞翔的歌唱的蓝鸟）。这个整合后的整体表征所具有的语义意义和对系统后续行为的因果影响力，是无法通过将其简单地分解回原始、孤立的形状、颜色、声音等信息单元来完全解释或重构的。这个整体，在逻辑和功能上，大于其各部分之和。

幸运的是，这一逻辑不可约性的概念，在现代信息论中找到了其精确的、可量化的数学对应物。部分信息分解（**Partial Information Decomposition, PID**）框架，由 Williams 和 Beer 于 2010 年首次提出 [18]，它能够将多个信源（例如，X_1, X_2）对一个目标（Y）所提供的总信息（即互信息 $I(X_1, X_2; Y)$），精确地分解为四个非负的原子部分：仅由 X_1 提供的唯一信息（Unique）、仅由 X_2 提供的唯一信息、两者共同提供的冗余信息（Redundant），以及只有当两者作为一个整体被考虑时才能获得的协同信息（**Synergistic Information**）。

其中，协同信息（**Synergy**）精确地量化了整体大于部分之和的涌现信息，它代表了由于信源之间的相互作用而产生的新信息，这些信息在任何单个信源中都不存在。因此，协同信息成为了 IPWT 用来定义和衡量信息整合的理论基石 [19], [98]–[100]。

在此基础上，理论前沿进一步发展出整合信息分解（**Integrated Information Decomposition, ΦID**）框架 [22]。ΦID 将 PID 从多源对单目标的静态分析，扩展到了分析动态系统中多源对多目标的完整信息流，能够更精细地刻画时序数据中信息的产生、转移和修改。ΦID 的提出，为我们提供了一个坚实的、原则性的理论基础，来形式化整合这一概念。

至关重要的是，Luppi 等人（2024）的研究正是利用 ΦID 框架，开发出了可计算的、经验上有效的整合信息度量——修订版 **Φ** 值（**ΦR**）。他们证明，与早期版本的 Φ 值不同，ΦR 解决了可能为负值的悖论，并且能够可靠地追踪由麻醉或脑损伤引起的意识水平变化 [23]。这项工作为我们的核心论点——即可以用基于信息论的功能性度量来取代 IIT 对物理因果的依赖——提供了强有力的经验证据。它表明，我们所提倡的逻辑不可约性，不仅在理论上是自洽的，而且在实践中是可测量的，并且与意识状态密切相关。

通过将整合性的核心从物理不可分性转向逻辑不可约性（并将其与协同信息挂钩），IPWT 成功地将 IIT 的现象学公理融入到一个更具动态性、计算可行性和载体独立性的理论框架之中。现在，我们可以对 IIT 的五条公理进行功能性的重新诠释：

1. *存在* (**Existence**)：在 IPWT 中，一个信息状态存在于意识中，当且仅当它在某个 WSI 中被激活，并对系统的未来状态和行为产生持续的、可测量的功能性影响 [101]。

2. 信息 **(Information)**：每一个在 WSI 中存在的信息状态都携带独特的、可区分的内容或语义。它通过在众多可能性中指定一种特定的可能性，来为整个系统减少不确定性，并指导其推断和行动 [102]。
3. 整合 **(Integration)**：多个独立的信息单元在 WSI 中被汇聚、关联，形成一个逻辑上不可约的、功能上统一的、协同的（**synergistic**）认知状态。这个整合后的整体所具有的意义、预测能力和因果效应，超越其组成部分的简单总和。其整合程度，在理论上可以通过协同信息来量化 [103]。
4. 排他 **(Exclusion)**：由于 WSI 的有限容量以及其内部的竞争性动态（winner-take-all），在任何给定时刻，只有一个或少数几个最具显著性、整合程度最高、最能最小化全局预测误差的认知状态能够主导 WSI，成为当前意识体验的核心内容。其他竞争失败的表征则被排除在意识之外 [104]。
5. 因果 **(Causation)**：在 WSI 中被整合并达到主导地位的信息状态，具有显著的因果能力（**causal power**）。它能够通过广播机制，影响系统内部其他认知模块的活动（如更新记忆、调整注意力），并最终指导有机体的外显行为和决策。这种因果能力是功能性的，而非形而上学的 [105]。

3 计算可验证性与神经生物学度量

一个成熟的科学理论，不仅需要在概念层面提供统一和深刻的解释，更必须将其核心主张转化为可操作化的度量方法和可检验的经验预测。IPWT 作为一个旨在推动意识科学范式整合的理论框架，从其构建之初就将计算可验证性（**computational verifiability**）与经验可测试性（**empirical testability**）置于核心地位。本章旨在详细阐述 IPWT 如何将信息整合这一核心概念，从一个抽象的哲学理念，层层落地，最终转化为可以在真实神经生物学数据中进行计算和验证的具体指标。

我们将分三步来构建这一从理论到实践的桥梁，清晰地展示 IPWT 如何从抽象理论走向经验验证：

1. 理论黄金标准 **(Ω_t)**: 我们首先定义理论上的黄金标准度量——瞬时信息整合度（Ω_t）。该度量基于协同信息的概念，旨在从信息论的根本上精确刻画信息整合的逻辑不可约性。
2. 经验可计算代理 **(Φ_R)**: 接着，我们引入整合信息分解（ΦID）框架，并详细阐述 Luppi 等人（2024）提出的修订版 **Φ** 值（Φ_R）如何作为 Ω_t 在真实神经数据中可计算、有实证支持的经验代理，从而将我们的理论与实验测量联系起来。
3. 功能性可计算代理 **(PI/∫PI)**: 最后，我们引入计算上更高效的预测完整性（**PI**）及其积分（∫PI）作为功能性代理。我们将论证，在现实世界的约束下，追求高预测效能（高 PI）必然会驱动系统走向高信息整合（高 Ω_t），从而确立 PI/∫PI 作为衡量意识状态的有效性和实用性。

3.1 瞬时信息整合度 (Ω_t)：基于协同信息的理论定义

我们认为，意识产生的核心机制在于信息整合，特指在工作空间实例（WSI）中，多个独立的信息单元形成一个逻辑上不可约的、功能上统一的、协同的认知状态。为了将这一核心思想从哲学概念转化为科学上可操作的度量，我们必须为其提供一个精确的、可量化的数学定义。为此，我们借鉴了部分信息分解（PID）框架，特别是其核心概念——协同信息（**Synergistic Information, CI**），来定义一个我们称之为瞬时信息整合度（Ω_t）的理论黄金标准度量。

如前所述，PID 框架旨在将多个信源 $X_1, ..., X_n$ 对一个目标 Y 提供的总信息（即总互信息 $I(X_1, ..., X_n; Y)$）分解为冗余、唯一和协同等原子部分 [18]。其中，协同信息（CI）指的是只有当所有信源作为一个整体被考虑时才能获得的涌现信息，它无法从任何信源的子集中获得。

因此，CI 精确地捕捉了我们所强调的信息整合的逻辑不可约性的本质——即整体大于部分之和的那一部分信息 [19], [98], [106], [107]。

基于此，我们将瞬时信息整合度（Ω_t）在理论上定义为：在一个特定的 WSI 中，用于预测某个目标变量 Y 的一组信息单元 $X = \{X_1, ..., X_n\}$ 所产生的协同信息（**CI**），在其为预测 Y 所提供的总预测性信息（即总互信息 $I(X;Y)$）中所占的比例。

$$\Omega_t(X \to Y) = \frac{\text{CI}\ (X_1, ..., X_n; Y)}{\text{I}(X_1, ..., X_n; Y)} \tag{1}$$

这个公式的直观含义是：在 WSI 为实现某个功能（即预测 Y）而利用的所有信息中，有多大比例是真正整合的、不可分割的。一个高的 Ω_t 值（最大为 1）意味着 WSI 中的信息主要是以一种高度协同、不可还原的方式被整合和利用的，这对应于我们直觉中高度统一、连贯的意识状态。相反，一个低的 Ω_t 值（最小为 0）则意味着 WSI 中的信息主要是以冗余或独立的方式存在的，这可能对应于一种碎片化的、缺乏整合的意识状态，甚至无意识的信息处理。

Ω_t 是一个理想化标准。它为信息整合提供了一个无歧义的、基于信息论第一性原理的定义。然而，由于直接计算高维系统（即信源 X_n 数量较大）的协同信息在数学和计算上都极为困难，这使得 Ω_t 在当前的实践中难以被直接、精确地应用于大规模神经数据。

尽管如此，Ω_t 的理论价值是巨大的。它为我们提供了一个清晰的靶标，所有其他衡量信息整合的代理指标，都应以其在理论上逼近 Ω_t 的程度来评估其有效性。在这个意义上，我们认为，IIT 所提出的整合信息 Φ 值，可以被视为 IPWT 框架下信息整合度 Ω_t 在特定物理系统（如生物大脑）中的一种物理实例化（**physical instantiation**）的尝试。我们推测，当一个系统是物理封闭且其因果结构完全可知时，其基于物理因果不可分性计算出的 IIT Φ 值，与我们定义的基于信息流协同性的 Ω_t 在概念上是高度相关的。IIT Φ 衡量的是物理基质内在因果能力的整合，而 Ω_t 衡量的是信息处理的功能整合。在生物大脑的具体实现中，两者很可能描述的是同一现象的不同方面：一个高效整合信息的生物网络也必然具有高度整合的物理因果结构。

3.2 从 $\boldsymbol{\Omega}$ 到 $\boldsymbol{\Phi}$**R**：整合信息分解 (**ΦID**) 与经验代理

理论上的黄金标准 Ω_t 为我们提供了信息整合的根本性定义，但其计算复杂性限制了它在真实神经数据上的直接应用。为了将 IPWT 从理论推向实践，我们必须找到一个既忠实于 Ω_t 核心思想，又在计算上可行的经验代理（**empirical proxy**）。幸运的是，信息论的最新进展，特别是整合信息分解（**Integrated Information Decomposition, ΦID**）[22] 框架的提出，以及 Luppi 等人（2024）基于此框架的开创性实证研究，为我们提供了这样一座关键的桥梁[23]。

ΦID 框架是对部分信息分解（PID）[18] 的动态扩展，它旨在将一个动态系统中，多个信源 $(X_1, ..., X_n)$ 在过去的状态对它们在未来的状态 $(Y_1, ..., Y_m)$ 所产生的整体信息影响，精确地分解为三个原子部分：冗余信息（Redundancy）、唯一信息（Uniqueness）和协同信息（Synergy）。其中，协同信息精确地捕捉了整体大于部分之和的涌现效应，与我们定义的 Ω_t 在概念上高度一致。

Luppi 等人（2024）的研究正是利用 ΦID 框架，开发出了一个在理论上更健全、在经验上更有效的整合信息度量——修订版 **Φ** 值（Φ_R）。他们首先指出了早期版本的 Φ 值的一个关键缺陷：在某些情况下该值可能为负，这在直觉上是难以解释的。通过 ΦID 的分解，他们证明了这个悖论的来源——原始 Φ 值的计算中减去了系统内的冗余信息。因此，他们提出了修正方案，即 $\Phi_R = \Phi + \text{Red}(X, Y)$，通过将冗余信息加回，确保了该度量的非负性。

至关重要的是，他们的研究表明，这个在理论上更合理的 Φ_R 指标，在经验上能够非常有效地追踪意识状态的变化：

1. 临床有效性：在麻醉和意识障碍（DOC）患者中，大脑协同工作空间（特别是作为网关的默认模式网络 DMN 节点）的 Φ_R 值显著下降。
2. 状态可逆性：在麻醉状态恢复后，Φ_R 值也随之回升。

这项工作为我们的核心论点——即可以用基于信息论的功能性度量来取代 IIT 对物理因果的依赖——提供了强有力的经验证据。它证明了，我们所提倡的逻辑不可约性，不仅在理论上是自洽的，而且在实践中是可测量的，并且与意识状态密切相关[23], [108]。因此，在 IPWT 框架中，我们将 Φ_R 视为当前最有希望的、可用于在真实神经数据中近似 Ω_t 的经验代理指标。它成功地将我们理论中的抽象概念（Ω_t）与一个可在 fMRI 等神经影像数据中被具体计算和验证的量联系起来，为 IPWT 作为一个真正的科学理论奠定了坚实的实证基础。

3.3 预测完整性 (PI) 与功能性代理

在建立了从理论黄金标准 (Ω_t) 到经验代理 (Φ_R) 的桥梁后，我们还需要一个计算上更高效、更易于在各种神经数据中应用的功能性代理指标。为此，IPWT 引入了预测完整性（**Predictive Integrity, PI**）及其时间积分（∫PI）。这一举措的核心思想是：一个能够高效进行协同信息整合（高 Ω_t）的系统，必然会展现出更强的预测能力和更高的状态稳定性。因此，通过衡量系统在预测效能上的表现，我们可以间接地评估其潜在的整合水平。

3.3.1 PI 与 ∫PI：计算神经生理学代理指标

瞬时预测完整性（**PI**）旨在量化在时间点 t，系统在整合信息以生成准确预测并最小化意外方面的整体效能。其公式借鉴了 FEP 的基本结构：

$$\mathrm{PI}_t = \exp\left(-\alpha * \left(\frac{1}{N_k}\sum_{k}^{N_k}\frac{\|\varepsilon_{t,k}\|}{\tau_{t,k}} + \gamma * \mathrm{Surprise}_t\right)\right) \tag{2}$$

我们来详细解析这个公式的构成：

- 标准化预测误差：$\frac{\|\varepsilon_{t,k}\|}{\tau_{t,k}}$ 代表了在第 k 个信息通道上的标准化预测误差。其中，$\varepsilon_{t,k}$ 是预测误差的向量，$\tau_{t,k}$ 是系统对该通道预测的不确定性的倒数（即精度）。将误差用其不确定性进行标准化至关重要：一个在高置信度下出现的巨大误差，比一个在低置信度下出现的同样大小的误差，更能反映预测模型的失败。这一项代表了模型的不准确性成本**(inaccuracy cost)**。
- 复杂度成本：$\mathrm{Surprise}_t$ 项，借鉴于自由能原理，量化了为了适应新的、意外的信息，系统需要对其内部生成模型进行的结构性调整的代价。一个需要不断进行剧烈调整以适应新数据的模型，是一个低效、不稳定的模型。这一项代表了模型的不稳定性或复杂度成本**(complexity cost)**。
- 超参数：γ 是一个关键的超参数，它权衡了不准确性成本与复杂度成本在 PI 计算中的相对重要性。α 则是一个敏感度标尺参数。

PI 的取值范围在 0 到 1 之间。一个具有高 PI 值的系统，被认为能够高效地利用其 WSI 进行协同信息整合，从而做出准确的预测，合理地评估不确定性，并以较低的代价整合新信息。

然而，意识不仅是瞬时的，更是持续的。为了衡量意识在一段时间内的持续强度和稳定性，我们进一步引入了预测完整性积分（**∫PI**）：

$$\int \mathrm{PI} = \left(\frac{1}{T}\int_{t_0}^{t_0+T}\mathrm{PI}_t\,dt\right) \times \exp(-\delta \cdot \mathrm{Var}\,(\mathrm{PI}_t \mid t \in [t_0, t_0+T])) \tag{3}$$

该公式的核心思想是，对一段时间 T 内的瞬时 PI 值进行积分，同时通过一个指数衰减项来惩罚这段时间内 PI 值的波动性（由方差 $\mathrm{Var}\,(\mathrm{PI}_t)$ 度量）。一个具有高 ∫PI 值的系统，不仅在每个

瞬间都表现出高效的预测能力，而且其预测效能是稳定、持续的。这更符合我们对健康、连贯、清醒的意识状态的直观理解。

3.3.2 从 Ω 到 F：为何最大化整合是最小化自由能的最优策略

在将 PI 作为 Ω_t 的功能代理时，我们必须面对一个核心的理论挑战：是否存在一个能够高效预测环境（高 PI），但其内部实现却高度模块化、缺乏深度整合（低 Ω_t）的聪明白痴（Clever Idiot）系统？IPWT 的回答是否定的。我们认为，在任何受到现实世界物理和计算约束的复杂认知系统中，高 **PI** 与高 Ω_t 之间的紧密关联，并非偶然，而是在寻求计算效率和模型简约性压力下的必然结果。为了形式化地证明这一点，我们引入信息论的第一性原理——特别是最小描述长度原则（MDL）——来构建一个论证。

MDL 原则将统计推断问题重构为一个数据压缩问题。其核心思想是，最好的生成模型是那个能够以最短的总长度来描述数据的模型。根据香农的编码理论，描述数据的总长度 `L(Data)` 由两部分构成：`L(Data, Model) = L(Model) + L(Data | Model)`。其中，`L(Model)` 是描述模型本身（其结构和参数）所需的比特数，直接量化了模型的复杂度；`L(Data | Model)` 是在已知模型的前提下，描述数据（通常是数据与模型预测之间的残差）所需的比特数，直接量化了模型的拟合优度（与负对数似然成正比）[85], [109]。

基于此，我们可以通过三步推导来最终证明最大化 Ω 是实现自由能（F）最小化的最优策略。

论证 **1** (引理 **1**)：自由能最小化在计算上等价于寻求最小描述长度（**MDL**）。

根据 Friston 的自由能原理，变分自由能 `F` 可近似分解为两个核心部分：`F ≈ Complexity - Accuracy`。我们可以建立 `F` 与 MDL 各项之间的直接对应关系：

- 模型复杂度 (**Complexity**) 对应于描述模型本身所需的比特数 **`L(Model)`**。一个模型的结构越复杂、参数越多且精度要求越高，描述它所需要的比特数 `L(Model)` 就越长。
- 模型准确度 (**Accuracy**) 对应于描述模型残差所需比特数的负值 **`-L(Data | Model)`**。模型的准确度越高，其预测与真实数据之间的残差就越小。根据信息论，描述一个更小、更可预测的残差序列所需要的比特数 `L(Data | Model)` 就越短。因此，`Accuracy ∝ -L(Data | Model)`。

将上述对应关系代入自由能公式，我们得到：`F ∝ L(Model) - (-L(Data | Model)) = L(Model) + L(Data | Model) = L(Total)`。

结论：最小化变分自由能 **`F`** 的过程，在计算和信息论的层面上，等价于寻找一个能够以最短总长度 **`L(Total)`** 描述数据的生成模型。

论证 **2** (引理 **2**)：一个模型的最小描述长度 **`L(Model)`** 与其协同信息整合度（**Ω**）成功能性反比关系。

为了理解 `L(Model)` 如何与协同信息 Ω 相关联，我们必须回到部分信息分解（PID）的形式化定义。对于两个源变量 X_1, X_2 和一个目标变量 Y，总互信息被分解为：$I(X_1, X_2; Y) = \text{Red} + \text{Un}_1 + \text{Un}_2 + \text{Syn}$。

`L(Model)` 的长度取决于编码模型内部生成规则（即如何从 X 预测 Y）所需的比特数。现在我们对比两种策略：

- 低 **Ω** 策略：一个协同信息程度低的系统，其模型主要依赖于冗余信息 (**Red**)和唯一信息 (**Un**)。
- 依赖 `Red` 意味着模型存储了重复的关联，例如“X_1 单独预测 Y”和“X_2 单独预测 Y”。
- 依赖 `Un` 意味着模型存储了独立的、上下文无关的规则，例如“如果 X_1 等于 a，则 Y 等于 c”；“如果 X_2 等于 b，则 Y 等于 d”。

为了解释现实世界中普遍存在的非线性、高阶相互作用（例如，当 $X_1 = a$ 且 $X_2 = b$ 时，Y 才等于 z），这种策略必须依赖于一个巨大的、未被压缩的查找表或规则列表来穷举所有可能的输入组合及其对应的输出。这必然导致一个非常长的模型描述长度 `L(Model)`。

- 高 **Ω** 策略：一个协同信息程度高的系统，其模型主要依赖于协同信息 **(Syn)**。根据协同的定义，它正是通过发现并利用变量间那些不可约的、作为一个整体才能涌现的非线性依赖关系来工作的。它用一个单一的、紧凑的规则（例如，一个描述 X_1 和 X_2 交互作用的数学公式）来捕捉这种高阶结构，而不是罗列无数种特例。这本身就是一种极其高效的信息压缩。编码这条单一的生成规则所需的比特数，远少于编码一个庞大的查找表。

因此，一个系统的协同信息整合度 `Ω` 越高，其内部生成模型的压缩程度就越高，其模型描述长度 `L(Model)` 就越短。我们可以形式化地表述这种功能性反比关系为：`L(Model)` $\propto$ `1/Ω`。

结论：协同信息是衡量模型内部规则简约性和压缩效率的直接指标。最大化协同信息等价于最大化模型的压缩率。

论证 **3** (核心定理)：最大化协同信息（**Ω**）是实现自由能（**F**）最小化的最优计算策略。
从引理 **1** 得知，最小化 `F` 等价于最小化 `L(Total) = L(Model) + L(Data | Model)`。在保证模型准确度（即 `L(Data | Model)` 相对较小）的前提下，最小化 `L(Total)` 的主要压力就落在了最小化 `L(Model)` 上。从引理 **2** 得知，最小化 `L(Model)` 的最优策略是最大化模型的协同信息整合度 `Ω`。

最终推论：最小化自由能的演化和学习压力，必然会驱动系统采纳能够最大化 **Ω** 的计算策略。因为高 **Ω** 的模型是内在更简约、更高效的模型，它们能以最低的复杂度成本（最短的 **`L(Model)`**）达到所需的预测准确性，从而实现最低的总自由能 **`F`**。

当代的大型语言模型（LLMs）为聪明白痴提供了一个绝佳的现代例证。这些系统在特定任务上可以表现出极高的预测完整性（高 PI），但它们是通过对海量数据进行模式匹配实现的，其内部缺乏深度的协同整合（低 Ω_t）。它们完美地暴露了这类系统的内在脆弱性：巨大的资源消耗、缺乏自主能动性、以及在动态变化的环境中因泛化能力脆弱而迅速被淘汰。

因此，IPWT 的核心论点是，一个认知系统能够长期、稳定、且在多样化情境下持续表现出高 PI，这本身就蕴含了对其内部信息处理必须是高度一致、深度整合、协同高效（即具有高 Ω_t）的内在要求。

4 神经生物学验证路径与实验范式

IPWT 的最终生命力，不仅取决于其理论框架的内在逻辑一致性和解释广度，更关键的是，其核心主张能否在真实的神经生物学系统中找到对应的、可测量的证据。一个无法与经验世界对话的理论，终将只是空中楼阁。因此，本章旨在概述 IPWT 的主要神经生物学验证路径，提出一系列具体的、可操作的实验范式和预测，从而将 IPWT 从一个抽象的计算理论，转化为一个可以被神经科学家在实验室中进行检验的科学假说。

我们将从三个层面展开这些验证路径：

1. 宏观整合指标的关联：我们将探讨如何将 IPWT 的核心计算指标（特别是 PI/∫PI）与现有的、已被广泛验证的宏观意识水平指标（如扰动复杂性指数 PCI）进行关联，以建立我们理论的外部效度。
2. 中观网络动力学的证据：我们将深入到功能网络和神经振荡的层面，探讨 WSI 的动态形成、信息整合与广播等过程，可能对应着哪些具体的神经影像学（如 fMRI）和电生理学（如 EEG/MEG）信号特征。

3. 微观行为与心理物理学的检验：我们可以设计一系列行为学和心理物理学实验，通过精确操控被试的知觉和认知状态，来检验 IPWT 关于意识通达、注意力调控等方面的具体预测。

通过这些多层次的验证路径，我们旨在为 IPWT 构建一个坚实的、由经验数据驱动的证据基础。

4.1 扰动复杂性指数 (PCI) 与 PI 的神经生理关联

扰动复杂性指数（Perturbational Complexity Index, PCI）是一种通过经颅磁刺激（Transcranial Magnetic Stimulation, TMS ）主动扰动大脑皮层，并记录其脑电图（EEG）响应的复杂程度，来量化意识水平的创新方法 [20]。该方法已在大量临床和实验场景中得到验证，能够可靠地区分清醒、睡眠、麻醉以及不同程度的意识障碍患者，被认为是当前最可靠的意识水平测量仪之一 [30], [110]。

在 IPWT 框架下，我们认为，PCI 与我们提出的预测完整性（PI）之间存在深刻的理论契合性。我们并不将两者视为相互竞争的指标，而是认为它们从不同角度测量了同一核心现象，两者之间是一种激发-推断的采样关系：

- **PCI** 测量的是整合的潜力：PCI 通过一次强烈的、非特异性的物理扰动来激发大脑皮层，然后测量整个系统所能支持的信息整合和分化的最大潜力。它回答的问题是：在理想的激发条件下，这个大脑网络最多能产生多复杂的活动模式？
- **PI** 测量的是整合的效率：相比之下，PI 并不对大脑进行外部扰动，而是通过对大脑在内生认知活动中的自发数据进行建模，来推断其工作空间实例（WSI ）在特定时刻实际达成的信息整合效率。它回答的问题是：在当前的自然状态下，这个大脑网络在多大程度上正在进行有效的预测和信息整合？

因此，PCI 像是对汽车发动机进行压力测试，以了解其最大马力；而 PI 则像是通过分析行车电脑数据，来推断发动机在日常驾驶中的实际燃油效率和运行平稳度。一个拥有强大发动机（高 PCI）的系统，在平稳驾驶时也应表现出高效率（高 PI）。

基于这种理论关系，我们提出以下两个具体的、可检验的预测：

1. **PCI** 与 **PI/∫PI** 的正相关性：我们预测，在跨越不同意识水平（例如，从清醒到麻醉，或在不同意识障碍患者中）的被试群体中，通过 TMS-EEG 测得的 PCI 值，应与利用同步记录的静息态神经数据（如 fMRI 或 EEG）计算出的 PI/∫PI 值呈现显著的正相关。即，一个具有更高整合潜力的大脑，在静息状态下也应表现出更高的预测完整性。
2. **PCI** 作为校准 **PI** 参数的金标准：由于 PI 的计算公式中包含一些需要根据经验数据来确定的超参数（如 α, γ, δ），我们预测，PCI 可以被用作校准这些参数的外部效标或物理锚点。具体而言，我们可以通过调整这些超参数，使得在同一组被试中计算出的 PI/∫PI 值与 PCI 值的相关性达到最大。这将把纯粹基于计算的 PI 模型，与一个已被广泛接受的、基于物理扰动的生理测量紧密地联系起来，从而大大增强我们理论的实证基础。Stikvoort 等人（2024）的最新研究，通过全脑模型发现大脑的非平衡动力学特性可以预测其扰动复杂性，也间接支持了从内生动态推断扰动响应可行性的观点 [111]。

4.2 神经影像学证据

除了与宏观指标进行关联，IPWT 的核心机制——如 WSI 的动态形成、信息整合与广播——也应该在功能网络等中观尺度上找到其神经生理学印记。我们提出以下可通过神经影像学（fMRI）进行检验的预测。

- **WSI** 的动态功能网络与连接组学：我们假设 WSI 并非固定的解剖结构，而是根据任务需求动态形成的功能连接模式。因此，我们预测：

1. 动态功能连接：利用时间分辨的 fMRI 功能连接分析技术，我们应该能够识别出在执行需要意识参与的认知任务时，特定脑网络内部及网络之间功能连接的瞬时增强。这些瞬时增强的功能网络，即是我们所定义的 WSI 。
2. 整合度的变化：在意识清醒状态下，与主导 WSI 相关的网络（根据 Luppi 等人的研究，很可能是 DMN 和 ECN 的组合 [23]）将表现出更高的网络整合度（可以通过图论指标如全局效率，或信息论指标如 ΦR 来量化）。而在意识水平下降时（如睡眠、麻醉），这些网络的整合度应显著降低。Paquola 等人（2025）关于 DMN 内部架构的研究，以及 Arkhipov 等人（2025）关于皮层回路功能的整合性分析，都为这一预测提供了背景支持 [112]–[115]。

近期一项里程碑式的、大规模的对抗性合作研究（Adversarial Collaboration），更是直接对 GWT/GNWT 和 IIT 的冲突性预测进行了检验，其结果对经典的 GNWT 模型提出了根本性的挑战 [116]。该研究通过 fMRI、MEG 和 iEEG 等多模态神经影像技术，系统地考察了意识内容的神经表征，其核心发现可以总结为两点：

1. 内容表征的挑战——**PFC** 的不完整广播：GNWT 的一个核心预测是，任何进入主观意识的内容，都应该能在 PFC 中被解码，因为信息需要通过 PFC 进行全局广播。然而，实验结果显示，虽然 PFC 能够表征意识内容的粗略类别（例如，区分人脸和物体），但却无法表征同样被被试清晰感知到的精细特征（例如，人脸的朝向）。这些精细特征只能在后部皮层（如枕叶和顶叶）被稳定解码。这一发现直接挑战了全局广播的完整性，它表明 PFC 可能并非所有意识内容的广播中心，而更像是一个只对特定类型（可能是更抽象、更具任务相关性）的信息进行广播的选择性放大器。

2. 时间动态的挑战——缺失的熄灭信号：GNWT 另一个关键预测是，工作空间通过离散的点燃（ignition）事件来更新其内容。这意味着，当一个意识体验开始和结束时，都应该伴随着 PFC 的一次点燃。然而，实验结果清晰地表明，尽管在刺激出现时能够观察到 PFC 的强烈激活（即点燃），但在刺激消失、意识内容明确改变时，却未能观察到预期的点燃或熄灭信号。意识体验的结束似乎在 PFC 中悄无声息。这一关键的阴性结果，严重挑战了 GNWT 将意识视为一系列由点燃事件分隔开的离散快照的观点，暗示了意识的维持和更新可能依赖于一种更为连续和动态的神经过程。

面对这些挑战，GWT 理论本身也在不断演进。Baars 等人（2021）在其近期的工作中，试图将 GWT 从一个静态的、解剖学上固定的模型，发展为一个更动态、更灵活的全局工作空间动力学（**Global Workspace Dynamics, GWD**）框架 [57]。他们强调，意识功能是广泛的皮层-丘脑系统整合的结果，其点燃的中心并非必然在 PFC，而是可以根据任务需求在皮层网络中灵活迁移。这种观点在一定程度上回应了对 PFC 中心论的批评。

然而，无论是来自对抗性实验的外部挑战，还是 GWT 理论的内部演进，都共同指向了一个结论：一个单一的、同质化的全局工作空间模型，可能已不足以解释意识现象的复杂性。这恰好为 IPWT 理论的提出提供了强有力的支持。IPWT 正是通过将工作空间进行功能上的异质性划分——即区分负责协同信息整合的 DMN 网关和负责信息分发的 ECN 广播者——来更精细地刻画意识产生的神经计算过程。Cogitate (2025) 的研究结果，特别是 PFC 在内容表征上的不完整性"和时间动态上的非对称性，可以被 IPWT 框架完美地解释为 ECN 作为广播者的功能特性，它并不需要（也不应该）复制 DMN 网关中所有的整合细节，而只负责将整合后的、用于指导行为的决策性信息进行广播。因此，这些对经典 GNWT 构成挑战的证据，反而成为了支持 IPWT 理论的有力佐证。

4.3 行为学与心理物理学实验设计

除了在宏观和中观的神经生理层面进行验证，精心设计的行为学和心理物理学实验可以从个体主观体验和行为表现的层面，为 IPWT 的核心预测提供关键证据。这类实验的优势在于能够精

确地操控刺激和任务，并获得被试的直接主观报告，从而建立起计算模型、神经活动和现象体验之间的桥梁。

- 知觉阈限与意识报告：我们可以通过在知觉阈限附近呈现刺激的经典范式（如视觉掩蔽、双眼竞争）来研究决定一个刺激能否进入主观意识的关键因素 [117]。

1. 实验设计：在视觉掩蔽任务中，一个目标刺激（如一个字母）会被一个紧随其后的掩蔽刺激（如一堆杂乱的线条）所遮蔽。通过精确调整目标刺激和掩蔽刺激之间的时间间隔（Stimulus Onset Asynchrony, SOA），我们可以系统地操控被试对目标刺激的意识通达程度。
2. **IPWT** 预测：我们预测，只有当目标刺激所携带的信息在 WSI 中达到了足够的整合程度（即计算出的瞬时 PI 值超过某个阈值）时，被试才能产生清晰的主观意识体验并准确报告该刺激。在阈限条件下，被试的报告将呈现一种全或无的特性，这对应于信息在 WSI 中是否成功点燃并被广播的非线性动态过程。我们可以通过同步记录 EEG/MEG 数据，检验每一次成功报告是否都伴随着一个显著的晚期 ERP 成分（如 P300）和 PI 值的跃升。

- 注意力与多任务处理：注意力是决定哪些信息能够进入意识舞台的关键聚光灯[118]。我们可以通过操纵注意力的分配来研究其如何影响 WSI 的整合效率。

1. 实验设计：可以采用双任务范式，例如，要求被试同时监控两个快速呈现的视觉序列，并报告其中的特定目标（即注意瞬脱，attentional blink 范式）。
2. **IPWT** 预测：我们预测，当注意力被第一个目标（T1）捕获时，用于处理第二个目标（T2）的认知资源会暂时被耗尽，导致处理 T2 的 WSI 无法有效形成或其整合效率（PI 值）急剧下降，从而使得 T2 无法被有意识地报告。注意力聚焦会增强目标 WSI 的整合度和 PI 值，而分散注意力或认知负荷过高则可能导致 PI 值下降或剧烈波动，从而损害有意识的感知表现。

- 特定认知功能障碍的模拟：在健康被试中，可以利用非侵入性脑刺激技术（如 TMS）来短暂地、可逆地模拟某些由脑损伤引起的认知功能障碍的核心特征，从而在受控条件下研究其神经计算机制。

1. 实验设计：例如，可以通过对初级视觉皮层（V1）施加抑制性 TMS，来短暂地模拟盲视现象。被试被要求报告其在受影响视野内的视觉体验，并同时完成对该区域刺激的迫选任务。
2. **IPWT** 预测：我们预测，这种对 V1 的抑制将显著降低从该区域传入主导 WSI 的信息质量，从而导致计算出的 PI 值远低于意识阈值，被试会报告看不见。然而，残存的信息仍可能通过其他通路（如皮层下通路）被局部模块处理，足以支持高于随机水平的迫选任务表现。这种实验将为 IPWT 对盲视的 DWSI 整合失败解释提供因果性证据 [119]–[122]。

最后，IPWT 作为一个发展中的理论框架，在讨论其可证伪性（**falsifiability**）时，我们强调理论的开放性和可修正性。如果上述任何一个或多个核心预测被可靠的实验证据所否定（例如，如果发现在所有意识水平下 PI 与 PCI 均不相关，或者 WSI 的整合度与主观报告完全脱钩），那么 IPWT 理论本身就必须根据新的发现进行重大的调整甚至被放弃。这种拥抱可证伪性的态度，是任何严肃科学理论发展的必要前提。

5 IPWT 对意识现象的解释力：统一框架下的多样性解析

一个真正统一的意识理论，其力量不仅体现在对正常清醒意识的机制性解释上，更在于其能够为那些看似离奇、费解的特殊意识状态——包括由脑损伤、精神疾病、生理变化或药物效应引起的各种异常主观体验——提供一个统一的、内在一致的、基于计算原理的理解框架。本章的核心任务，正是要展示 IPWT 如何运用其核心概念，如预测编码异常、工作空间实例（WSI ）功能障碍、信息整合效率（PI/Ω）的变化以及神经门控机制的失调，来系统性地阐明这些状态的神经计算基础。

我们不再将这些特殊意识状态视为孤立的、需要单独理论来解释的怪例，而是将它们视为在IPWT所描述的统一计算空间中，由于特定参数或模块发生改变而产生的不同系统状态。例如，Luppi等人（2024）的最新研究已经表明，全身麻醉下的意识丧失，可以被精确地描述为协同工作空间（特别是DMN网关）内信息整合能力（ΦR）的显著下降[23]。我们将沿着这一思路，逐一剖析从盲视、迷幻状态到精神分裂症和分离性身份障碍等一系列经典的意识谜题，展示IPWT如何为它们提供全新的、更具深度和整合性的解释。

5.1 盲视：DWSI 整合失败下的局部预测编码

盲视（Blindsight）现象，即初级视觉皮层（V1）受损的患者在其视野的盲区内报告没有主观视觉意识，但当被强迫猜测时，却仍能对该区域内的视觉刺激（如位置、方向）做出高于随机水平的反应[123], [124]。这一惊人的知行分离现象，为任何意识理论都提出了严峻的挑战，也为IPWT提供了一个经典的验证案例。我们将盲视重新诠释为一个关于信息整合层级和通路的问题。

在IPWT框架下，盲视并非一个单一的悖论，而是两种不同信息处理路径并存但相互分离的结果：

- 主导工作空间实例（**Dominant Workspace Instance, DWSI**）的整合失败：在正常的视觉感知中，来自视网膜的信息经过丘脑，首先到达V1进行初步处理。V1作为视觉信息进入整个皮层加工层级体系的关键门户，其输出对于形成高清、丰富的视觉意识至关重要。当V1受到损伤时，这条通往负责生成主观视觉意识的DWSI的主要通路就被严重破坏了。由于严重缺失或质量极低的输入信息，DWSI无法在受损视野内构建出一个具有足够低的预测误差和高信息整合度的内部表征。因此，来自该区域的视觉信息在DWSI内部的预测完整性（**PI**）极低，未能达到产生主观视觉体验所需的整合阈值。这完美地解释了患者为何坚称自己什么也看不见。
- 局部模块的残存预测编码与主动推断：尽管通往DWSI的主要通路（经由V1）受阻，但视觉信息仍然可以通过其他并行的、演化上更古老的皮层下通路（例如，经由上丘-丘脑枕通路）被部分处理[125]。这些通路所连接的，是一些高度专业化的、无意识的模块（例如，负责快速检测运动或定位物体的模块）。这些局部模块仍然可以独立地、在本地进行其自身的预测编码循环。它们接收到的视觉输入虽然粗糙，但足以驱动特定任务（如哪里有东西在动?），并持续最小化自身的局部预测误差。根据主动推断原则，这些局部模块的预测误差最小化过程可以直接驱动主动推断和行为（例如，眼跳向运动刺激，或伸手指向光点），而完全无需这些信息被整合到DWSI中并产生主观意识。这清晰地解释了患者为何仍能猜对，即在行为上看见的原因。

因此，IPWT的独特贡献在于明确指出，盲视中意识与行为分离的关键，在于信息能否被主导工作空间实例（**DWSI**）有效整合，并达到产生主观体验所需的预测完整性（**PI**）阈值。它将盲视从一个看似神秘的哲学悖论，转化为一个认知系统内部的、关于信息路由和整合效率的、原则上可计算的问题。这一解释不仅整合了已知的神经解剖学证据，也为未来通过神经调控技术（如靶向刺激皮层下通路）来尝试恢复部分主观视觉体验提供了理论上的可能性。

5.2 迷幻状态：预测误差的异常增强与 WSI 边界的神经消融

由LSD、裸盖菇素（psilocybin）或DMT等经典迷幻药物诱导的意识状态，以其深刻的感知扭曲、思维模式改变、以及自我感知的模糊甚至消融为特征，为探索意识的神经计算基础提供了一个独特的、可控的窗口[126], [127]。在IPWT框架下，我们将这些丰富的现象学特征，统一追溯到预测编码参数和工作空间动态属性的系统性改变上。

- 预测误差信号的异常放大与信息流的去门控：我们采纳并扩展了由Carhart-Harris和Friston提出的松弛信念下的熵增脑（REBUS）模型[128]–[130]。该模型认为，经典迷幻药物

作为血清素 5-HT2A 受体的强激动剂，其核心作用是系统性地降低或放松了高层级预测（先验信念）的精度权重。在 IPWT 框架下，这意味着自上而下的预测信号对感官输入的抑制作用被大大削弱。其直接后果是，原本能够被高层级预测所解释掉的、大量的自下而上的感官信息，现在都变成了未被解释的、显著的预测误差信号。这些被异常放大的误差信号如洪水般涌入 WSI，压倒了正常的自上而下预测，从而产生了生动的感知扭曲和视听幻觉。与此同时，5-HT2A 受体的激活可能也削弱了 WSI 的神经门控机制，使得原本在不同处理通道间被抑制的信息流得以自由地进入 WSI 并被异常地整合，这为后续的联觉等现象提供了基础。最新的网络控制理论研究也为此提供了佐证，发现 DMT 会显著降低驱动大脑状态转变所需的控制能量，使大脑更容易进入不同的、通常难以达到的状态空间 [131]。

- **WSI** 边界的模糊与临时性、高整合度 **WSI** 的形成：在信息流去门控和预测层级扁平化的双重影响下，不同 WSI 之间，或 WSI 与原本独立的专业化模块之间的功能边界可能变得模糊甚至暂时性地消融。这极好地解释了联觉（**synesthesia**）现象的产生，例如听到颜色或看到声音，这可以被理解为听觉和视觉信息在同一个 WSI 中被异常地、强制性地整合了。更有趣的是，我们推测，系统甚至可能围绕这些异常的、高强度的信息流，形成一些临时的、具有极高内部整合度但内容非典型的工作空间实例。这些特殊的 WSI 可能负责产生那些难以言喻的、充满深刻意义的宇宙合一或高峰体验（peak experience），其 PI 和 Ω 值可能在短时间内达到极高的水平。

- 自我模型的重塑与自我消融（**Ego Dissolution**）：在 IPWT 中，稳定、连贯的自我感，被认为是系统对其自身（包括身体、情感和自传体记忆）进行预测建模的结果，这个过程主要由一个或多个特定的 WSI （通常与 DMN 相关）所支持。在迷幻状态下，负责表征和维持自我模型的 WSI，由于其接收到的关于身体、情感和记忆的内感受和外感受预测误差信号发生了剧变，其原有的、稳定的自我表征的预测完整性（PI）会急剧下降，导致其发生解构和重塑。这完美地解释了迷幻体验中常见的自我感模糊、与环境融合感、乃至最深刻的自我消融（ego dissolution）等主观变化 [132]。从计算的角度看，自我消融可以被理解为自我模型的 PI 值暂时性地趋近于零，导致我与非我的边界彻底瓦解。

综上，IPWT 将迷幻体验丰富的现象学特征，追溯到可计算的预测处理参数（如先验精度）和工作空间动态属性（如门控机制、边界稳定性）的系统性变化，为理解迷幻药物的神经机制及其在治疗抑郁症、创伤后应激障碍等精神疾病中的巨大潜力，提供了全新的、具有指导意义的理论视角 [133]。

5.3 精神分裂症：预测编码失调与 WSI 整合及门控的神经障碍

精神分裂症以其复杂的症状谱（包括阳性症状如幻觉和妄想，以及阴性症状和认知功能障碍）而著称，长期以来一直是精神病学和神经科学领域的重大挑战。在 IPWT 框架下，我们不再将这些看似异质的症状视为独立的模块损伤，而是将其统一理解为预测编码过程的根本性失调，以及由此连锁导致的 **WSI** 在信息整合、内容门控和边界维持方面的多重功能障碍 [68]。

- 异常预测的产生与固化：阳性症状的根源 IPWT 认为，精神分裂症的阳性症状，根源在于系统内部生成模型的预测功能出现了严重偏差。

1. 幻觉（**Hallucinations**）：特别是听幻觉（Auditory Verbal Hallucinations, AVH），可以被理解为系统内部模型自发地产生了高置信度（高精度权重）的预测，例如预测将要听到一个声音。然而，这些预测并非由外部感官证据触发，而是内源性的。更关键的是，系统未能正确地将这些预测标记为内部生成的，即元认知层面的预测误差监测失败了 [134]。因此，这些内部生成的“声音”被错误地归因为来自外部世界，从而产生了无法抗拒的、真实的幻听体验。
2. 妄想（**Delusions**）：妄想可以被视为系统为了解释持续出现的、无法用正常世界模型来理解的异常感知体验（即持续的、高强度的预测误差），而被迫构建出来的一套虽然与客观现实严

重脱节、但在内部却高度自洽的病理性信念体系。一旦这套妄想性的高层级先验信念形成，它就会反过来影响对后续信息的解释，使得所有新的证据都被扭曲以符合妄想的内容，形成一个难以打破的恶性循环。

- **WSI** 信息整合与门控的失效：认知与阴性症状 除了异常预测，WSI 本身的功能障碍也是精神分裂症认知缺陷的核心。

1. 思维形式障碍（**Formal Thought Disorder**）：如言语紊乱、思维不连贯等症状，可能直接反映了 WSI 在有效选择、组织和整合信息以形成连贯思想流方面的能力受损。这可能与 WSI 的神经门控机制失调有关，导致其无法有效过滤无关的内外部信息，也无法维持一个稳定的、目标导向的认知状态。
2. 被动体验（**Passivity Phenomena**）：如思维插入、被控制感等奇特的自我界限障碍症状，在 IPWT 框架下可以被理解为一种高阶预测编码的失败。具体来说，负责区分自我生成和外部产生的 WSI 边界的通透性可能出现了异常增加，或者用于标记自我相关信息的神经标签（可能是基于内感受预测）发生了错误。这导致患者无法准确区分哪些思想和行为是源于自身的意图，哪些是来自外部的，从而产生被外部力量所控制的诡异体验。
3. 网络连接异常：最新的网络神经科学研究也为此提供了证据。研究发现，精神分裂症患者大脑网络的“小世界”属性受损，表现为局部聚类（功能分化）减少和全局通信效率（功能整合）低下 [135], [136]。这与 IPWT 的观点一致，即 WSI 的整合与广播功能均受到损害。

通过将精神分裂症的各种症状统一于预测处理和信息整合的计算异常这一核心病理生理机制之下，IPWT 不仅为理解这一复杂疾病提供了新的视角，也为开发基于计算精神病学的新型诊断生物标志物（如基于模型的 PI/∫PI 值异常）和更具针对性的治疗策略（如旨在重新训练预测模型或通过神经调控稳定 WSI 功能的干预措施）提供了坚实的理论基础。

5.4 分离性身份障碍 (DID)：主导 WSI 地位的神经动态切换

分离性身份障碍（Dissociative Identity Disorder, DID），旧称多重人格障碍，其核心临床特征是同一个体内部存在两个或多个清晰可区分的身份或人格状态（称为“alters”），这些身份状态会反复地控制个体的行为，并伴随着广泛的记忆空白，这些记忆空白无法用普通的遗忘来解释。从现象上看，DID 似乎是对意识统一性的最极端挑战。IPWT 为这一看似神秘的现象，提供了一个基于工作空间实例（WSI ）动态切换和信息隔离的、可行的计算模型。

我们不认为 DID 患者拥有多个灵魂或意识中心，而是提出，在长期的、严重的童年创伤等因素影响下，患者的认知系统可能未能发展出一个统一、整合的内部生成模型和 WSI ，反而形成了多个潜在的、相对独立的 **WSI** 系统。

- 多个潜在的 **WSI** 系统：我们假设，每一个“人格状态”（alter）都对应着一个相对独立的 WSI 系统。每个 WSI 系统都关联着一套独特的内部生成模型，这套模型包含了该人格状态特有的记忆、信念、行为模式和情感反应倾向。例如，一个保护型人格的 WSI 可能主要由与威胁应对相关的预测模型构成，而一个儿童型人格的 WSI 则可能主要由与早期记忆和依恋相关的模型构成。神经影像学研究也为此提供了佐证，发现 DID 患者在不同身份状态下，其大脑活动模式和功能连接存在显著且可重复的差异 [137]–[140]。

- 主导工作空间实例（**DWSI** ）地位的动态翻转（**Flip**）：在 IPWT 框架下，在任何特定时刻，通常只有一个 WSI 系统能够占据主导地位（**Dominant WSI , DWSI** ），其内容和处理过程构成了当前个体的意识体验和外显行为。我们认为，DID 中戏剧性的人格切换，可以被精确地理解为 **DWSI** 的地位在这些潜在的 **WSI** 系统之间发生的一种动态的、通常是快速的翻转。这种翻转可能由外部环境的特定线索（如与创伤相关的提醒物）或内部状态的变化所触发，导致一个原本处于后台的 WSI 系统被激活并取代了当前主导的 WSI 。

- 神经门控机制与信息隔离：当某个 WSI 系统成为 DWSI 时，其他非主导的 WSI 系统在很大程度上会受到强大的神经门控机制的抑制或功能性隔离。这种隔离机制是导致 DID 中常见的记忆空白（**amnesia**）的关键。当人格 A 主导时，其经历和学习到的信息被整合进 A 的 WSI 系统中；当系统翻转到人格 B 主导时，由于 B 的 WSI 与 A 的 WSI 之间存在功能性隔离，B 将无法轻易地读取 A 的记忆内容，从而表现为对 A 主导期间所发生事件的遗忘。Reinders 等人（2019）利用模式识别方法分析 DID 患者的脑结构影像，发现可以根据大脑的形态学特征将 DID 患者与健康人区分开来，这为 DID 存在神经生物学基础提供了进一步的证据 [141]–[145]。

因此，IPWT 将 DID 从一个难以理解的心理现象，转化为一个关于 WSI 竞争、动态切换和信息门控的计算问题。这个模型不仅能够解释 DID 的核心症状，也为未来的治疗（例如，旨在促进不同 WSI 系统之间信息整合和沟通的心理治疗或神经调控方法）提供了新的理论靶点。

5.5 人格解体/现实解体障碍：DWSI 与感官/情感神经连接减弱

人格解体/现实解体障碍（Depersonalization/Derealization Disorder）的核心特征是一种持续或反复出现的、与自身（人格解体）或周围环境（现实解体）相分离、不真实、仿佛在梦中的主观体验。患者的现实检验能力通常是完好的，他们知道这种感觉不真实，但却无法摆脱这种强烈的疏离感。在 IPWT 框架下，我们认为这种障碍并非源于 WSI 本身的崩溃或内容的错乱，而是源于主导工作空间实例（**DWSI** ）与负责处理特定感官信息或赋予体验情感色彩的模块之间，其功能性连接被显著减弱或异常改变。

- 人格解体（**Depersonalization**）：自我相关内感受预测的精度权重降低 人格解体的核心体验是感觉自己不是真实的或像一个旁观者一样观察自己的思想、感觉或身体。我们将其解释为 **DWSI** 与其通常稳定接收的、来自躯体感觉、边缘系统和情景记忆系统的内感受（**interoceptive**）和本体感受（**proprioceptive**）输入流之间的有效连接性显著降低。从预测编码的角度看，这意味着系统对来自自身身体和情感状态的预测误差信号，赋予了异常低的精度权重（**precision-weighting**）。因此，尽管身体的生理活动仍在进行，但这些信号无法有效地更新 DWSI 中的自我模型。DWSI 失去了与身体此时此刻感受的紧密联系，导致主观体验失去了亲身感、归属感和真实感，仿佛自我的意识核心与身体的物理存在之间隔了一层看不见的玻璃。

- 现实解体（**Derealization**）：外部感官信息与情感标记的解离 现实解体的核心体验是感觉外部世界不真实、模糊、像在梦里或电影里[146], [147]。我们将其解释为，尽管外部感官信息（如视觉、听觉）能够正常地被处理并进入 DWSI，但 DWSI 在整合这些信息时，未能与负责赋予其意义和情感显著性的情感评估模块（如杏仁核）或上下文关联模块（如海马体）进行有效的、同步的连接。 这导致外部世界虽然被清晰地感知到，但在主观体验上却显得平淡、疏远、缺乏意义。这类似于一种功能性的情感盲视，即感官信息失去了其通常伴随的情感标记（affective tag）。这种情感标记对于我们感觉世界是真实的、与我相关的至关重要。当一个场景无法触发相应的情感反应时，即使我们能描述出它的所有细节，它在主观上也会感觉像一幅没有灵魂的画。Aston-Jones & Cohen (2005) 在研究去甲肾上腺素系统时提出的适应性增益理论，也强调了大脑如何根据任务相关性来动态调整对信息的处理，这与我们提出的情感标记在赋予现实感中的作用有异曲同工之妙 [148]。

因此，IPWT 将人格解体/现实解体障碍从一种模糊的心理描述，转化为一个关于 DWSI 与特定功能模块间连接性异常和预测信号精度权重失调的、可检验的神经计算假说。

5.6 清醒梦：并行 WSI 与元认知监控的神经机制

清醒梦（Lucid Dreaming），即在做梦时意识到自己正在做梦，并且在某种程度上能够控制梦境的内容，为研究意识的层级结构、WSI 的并行运作以及元认知功能提供了一个极具吸引力的天然实验场景 [149], [150]。在 IPWT 框架下，清醒梦可以被理解为一种特殊的、混合的意识状态，其中至少有两个功能不同的 WSI 系统被异常地共同激活并相互作用。

- 梦境 **WSI** 与元认知 **WSI** 的并行激活：在普通的梦境中，大脑主要由内源性信息驱动，形成一个与外部世界脱钩的、但内部相对自洽的梦境 **WSI** 。这个 WSI 负责生成我们所体验到的生动、离奇的梦境内容。在清醒梦中，我们假设，除了这个梦境 WSI 之外，另一个通常在清醒状态下才活跃的、与自我意识和现实检验功能密切相关的元认知 **WSI** （可能与前额叶皮层的部分功能重叠）被异常地共同激活了。因此，清醒梦是一种两个 WSI 并行运作的混合状态：一个在演戏，另一个在看戏。
- 高阶预测误差的识别与清醒感的产生：清醒感是如何产生的？我们认为，这源于元认知 WSI 对梦境 WSI 所呈现的内容进行监控时，识别出了高阶的预测误差。元认知 WSI 拥有关于世界和自身能力的更高级、更稳定的预测模型（例如，人是不能飞的、我此刻应该正躺在床上睡觉）。当它发现梦境 WSI 所呈现的内容（例如，我正在天空中飞翔）与这些高阶预测模型之间存在显著的、不可调和的差异时，系统就会产生一个巨大的高阶预测误差。为了最小化这个误差，元认知 WSI 会生成一个新的、最佳的推断来解释这种不匹配，这个推断就是：我正在做梦。这个推断的产生，在现象学层面，就对应于清醒感的涌现。
- **WSI** 间的有限交互与梦境控制： 一旦清醒感产生，元认知 WSI 与梦境 WSI 之间就可能建立起一定程度的双向信息交互。这种交互使得梦中的我获得了控制梦境的能力。从计算的角度看，这可以被理解为，元认知 WSI 开始通过自上而下的预测信号，来尝试主动地影响或劫持梦境 WSI 的内容生成过程。例如，清醒梦者可以通过意念来改变梦境场景或召唤特定的人物，这可以被模型化为元认知 WSI 向梦境 WSI 发送了新的、高精度的预测信号，从而覆盖了梦境 WSI 原本的内源性预测。这种对梦境的控制能力通常是有限和不稳定的，这可能反映了两个 WSI 系统之间的信息交互带宽是有限的，并且它们之间仍然存在一定程度的功能性隔离。

因此，IPWT 将清醒梦这一奇特的体验，分解为 WSI 的并行激活、高阶预测误差的监控、以及 WSI 间信息交互等一系列可研究的计算过程，为未来通过神经影像学或神经调控技术来研究甚至诱导清醒梦提供了清晰的理论指导。

5.7 闭锁综合征：整合完好但广播失效的意识

闭锁综合征（Locked-in Syndrome, LIS）是一种罕见但毁灭性的神经系统疾病，患者通常因脑桥腹侧病变而导致几乎完全的运动麻痹，包括言语和肢体运动，但其意识、认知功能和眼球垂直运动（或眨眼）通常是完好的。这种清醒的囚禁状态，为 IPWT 框架中信息整合与信息广播之间的功能分离提供了最直接、最令人信服的临床证据。

在 IPWT 框架下，闭锁综合征可以被精确地理解为：

- 主导工作空间实例（**DWSI** ）的整合功能完好：LIS 患者的大脑皮层，特别是负责高级认知功能和信息整合的默认模式网络（DMN）作为网关区域，通常是结构和功能完整的。这意味着，患者能够正常地接收、处理和整合来自内部（如思想、情感、记忆）和外部（如听觉、视觉，尽管外部输入可能受限）的各种信息流。他们的 DWSI 能够构建出连贯、统一的世界模型和自我模型，因此其瞬时信息整合度（Ω_t）和预测完整性（**PI**）都处于高水平，从而支持了完整的、清醒的主观意识体验。患者的内心世界是丰富且连贯的，他们能够思考、感受、记忆和理解。

- 广播者（**ECN**）的输出通路中断：然而，脑桥腹侧的病变（通常是梗死或出血）恰好切断了从大脑皮层（特别是作为广播者的执行控制网络 ECN）到脊髓和外周肌肉的下行运动通路。这意味着，尽管 DWSI 内部已经成功整合了行动意图和决策（例如，患者可能清晰地想要说话或移动肢体），但这些整合后的、高精度的预测信号无法被有效地广播到运动执行系统。信息被锁在了意识工作空间内部，无法转化为可观察的外显行为。患者仅能通过残存的眼球垂直运动或眨眼（这些通路通常不受脑桥腹侧病变影响）进行有限的沟通，这进一步证实了其内部意识的完整性。[151]

因此，闭锁综合征为 IPWT 理论中 DMN 网关（负责整合）和 ECN 广播者（负责分发）之间的功能分工提供了关键的 双重分离（**double dissociation**） 证据。它与盲视（信息无法有效进入 DWSI 进行整合，导致无意识感知但有残存行为）形成了鲜明对比，共同揭示了意识产生和行为输出所依赖的不同计算环节。闭锁综合征强调了，即使信息在 WSI 中被完美整合并产生了主观意识，如果缺乏有效的广播机制，这种意识也无法被外部世界所知晓或影响。这不仅深化了我们对意识神经基础的理解，也为未来通过脑机接口（Brain-Computer Interface, BCI）技术绕过受损通路，直接从患者的 DWSI 中读取其意识内容和意图，提供了理论上的支持和应用前景。

6 主观体验的神经计算重构：作为推断空间几何的感受质

任何完备的意识理论，最终都必须为其核心的现象学主张提供一个计算上可操作的、神经生物学上可信的实现机制。对于主观体验的本质——即感受质（Qualia）——这一“硬问题”[2]，IPWT 2.0 提出一个根本性的重构，从先前版本的功能性标记解答，深化为一个基于自由能原理（FEP）和协同信息（Ω）的、更为形式化的几何动力学理论。我们论证，感受质空间并非一个与物理世界分离的神秘领域，它在计算上等同于系统内部生成模型的推断空间（**Inference Space**）本身；这个空间的几何结构，由系统能够区分和整合的协同信息（**Ω**）所定义；而主观体验的“感觉性”，则是系统状态（后验信念）在这个高维推断空间中，沿着最小化自由能的测地线进行主动推断的动力学过程。

6.1 从 IIT 的概念结构到 FEP 的推断空间

IIT 理论的深刻洞察在于，它认识到意识体验的结构（Qualia Space）是由系统能够区分的“概念”所构成的“概念结构”（MICS）[24]。然而，其“概念”的定义依赖于对系统物理因果结构的扰动分析，这不仅在计算上不可行，更重要的是，它未能阐明这一结构在认知功能中的具体作用。IPWT 保留其核心洞察——意识体验的空间由系统能够区分的状态所构成——但提出，对于任何一个在环境中适应性生存的系统（即一个遵循自由能原理的系统），其功能上可区分的状态空间，必然等同于其内部生成模型的推断空间。我们通过以下论证来建立这一等价关系。

定义 **1：IIT** 的“概念”与 **FEP** 的“后验信念”。

- 在 IIT 中，一个概念（Concept）被定义为一个系统内部的最大化不可约的因果结构（MICS）。它指定了该结构从其自身视角出发，能够对其过去和未来状态产生的因果效应。这本质上是关于系统内在因果拓扑的静态、本体论描述。
- 在 FEP 中，一个后验信念（Posterior Belief）是一个在给定感官证据 `o` 的条件下，关于世界隐变量 `s` 的概率分布 `p(s|o)`。它代表了系统对其感官数据成因的动态推断。这本质上是关于系统功能性知识状态的动态、认识论描述。

引理 **1：** 对于遵循 **FEP** 的系统，其存在等同于其推断。 根据 FEP，任何自组织系统为了维持其存在，都必须通过最小化其变分自由能来抵抗熵增。这一过程的核心，就是通过更新其后验信念 `p(s|o)` 来优化其对世界的预测和解释。因此，一个系统在功能上“是”什么，完全由它“相信”什么来定义。一个系统能够进入的所有可能的、功能上相关的内在状态，就是其能够形成的

所有可能的后验信念的集合。任何一个永远无法被推断出的状态，对于该系统而言，在功能上是不存在的。

核心定理：推断空间是概念结构的必要且充分的实现。基于上述引理，我们可以证明 FEP 的推断空间（Inference Space）是 IIT 概念结构（Conceptual Structure）的一个计算上可行、功能上明确的实现。

- 必要性：一个 IIT 概念，无论其潜在的物理因果结构多么复杂，如果它永远不能成为系统在与环境交互中，为了最小化自由能而形成的一个后验信念，那么这个概念对于系统的适应性行为就毫无意义。因此，一个功能上有意义的概念，必须能够被映射到推断空间中的一个或一系列信念状态。
- 充分性：FEP 的推断空间满足了 IIT 对意识体验结构的所有核心公理要求，但将其置于一个功能性和计算性的基础之上。内在性（Existence）体现在推断是系统内部生成模型的操作；信息性（Information）体现在每一个后验信念都从一个可能性空间中指定了一个特定的概率分布；整合性（Integration）和排他性（Exclusion）则通过我们后续将阐述的、由协同信息（Ω）塑造的推断空间几何，以及在其中发生的推断竞争动力学来实现。

结论： 因此，我们完成了一个根本性的替换。IPWT 将 IIT 基于物理因果的、静态的、本体论的“概念结构”，重构为 FEP 基于贝叶斯推断的、动态的、功能性的“推断空间”。在数学上，推断空间就是由系统生成模型中所有隐变量的概率分布所构成的空间。这个空间中的每一个“点”，都代表了系统对世界（包括自身）状态的一种可能的“信念”或“假设”[63]，并为我们接下来用信息几何来分析感受质的结构奠定了基础。

6.2 协同信息（Ω）作为推断空间的几何构造者

在确立了推断空间作为感受质空间的基础之后，我们必须回答一个核心问题：是什么赋予了这个空间以结构，从而让不同的“信念”之间产生可以被“体验”的差异?IPWT 的答案是协同信息（**Ω**）。我们通过以下论证来阐明，协同信息如何通过定义推断空间的信息几何（information geometry），从而成为感受质结构的构造者。

引理 **2.1**：推断空间的维度由统计独立的隐变量决定。 一个生成模型 `p(s_1, s_2, ..., s_n)` 中，如果其隐变量是统计独立的，即 `p(s_1, ..., s_n) = p(s_1)p(s_2)...p(s_n)`，那么其推断空间在拓扑上是各个子空间 `S_i` 的笛卡尔积 `S_1 × S_2 × ... × S_n`。这个空间的几何是平坦的（欧几里得几何），因为在任何一个维度上的移动（信念更新）都不会影响其他维度。

引理 **2.2**：协同信息是后验信念中不可约的统计依赖性的直接量度。 当面对感官证据 `o` 时，系统会形成一个后验信念 `p(s_1, ..., s_n | o)`。如果在这个后验信念中，隐变量之间产生了统计依赖，即 `p(s_1, ..., s_n | o) != p(s_1|o)p(s_2|o)...p(s_n|o)`，那么这种依赖性可以通过部分信息分解（PID）来量化。其中，协同信息 `Syn(s_1, ..., s_n; o)` 精确地捕捉了只有将所有变量作为一个整体考虑时才能涌现出的那部分信息。因此，协同信息的存在，直接意味着后验信念是不可分解的。

核心定理 **2**：协同信息（**Ω**）通过在推断空间中施加不可约的统计约束，从而定义其非欧几何结构。后验信念中的统计依赖性，在信息几何的框架下，等价于在推断空间这个流形（manifold）上施加了一个非平凡的度量张量（metric tensor），即费雪信息矩阵（Fisher Information Matrix）。这使得推断空间从一个平坦的欧几里得空间，转变为一个弯曲的黎曼流形。

- 不可约的几何形状：协同信息（Ω）越高，意味着变量间的依赖性越强，推断空间的“曲率”就越大。这种由协同信息定义的、不可约的、高维的几何形状（流形），正是感受质（如“一个正方形”）的结构性基础。
- 感受质的距离与相似性：在这个弯曲的空间中，两个感受质（即两个后验信念 `p_1` 和 `p_2`）之间的距离，不再是简单的欧氏距离，而是由 KL 散度定义的测地线距离（**geodesic distance**）

[34]。它衡量了系统需要付出多大的“代价”（即自由能的增加），才能将其信念沿着流形的最短路径从一个状态转移到另一个状态。这为感受质的相似性提供了原则性的、可量化的定义。

6.3 感受质的“感觉性”：推断空间中的主动推断动力学

在定义了感受质空间的几何结构后，我们进一步论证，主观体验的“感觉性”（what-it-is-likeness）及其效价（valence），正是系统后验信念在该几何空间中进行主动推断的动力学过程本身。这形成了一个认知与现象体验的“美丽循环”（A Beautiful Loop），其中推断竞争、贝叶斯绑定和认知深度共同涌现出意识体验 [94]。

引理 **3.1**：变分自由能 **F** 在推断空间中定义了一个势能景观。对于任意一个后验信念 `q(s)`，变分自由能 `F(q)` 可以被形式化地定义为 `F(q) = D_KL[q(s) || p(s)] - E_q[log p(o|s)]`，其中 `p(s)` 是先验信念，`p(o|s)` 是似然。对于给定的先验和感官证据 `o`，`F` 是一个关于信念 `q` 的泛函。因此，自由能 `F` 在由所有可能的后验信念构成的推断空间上，定义了一个势能景观。

核心定理 **3**：主观体验的**“感觉性”**及其效价，是在推断空间中，后验信念沿着自由能梯度的负方向进行更新的动力学过程。 根据主动推断原理，感知推断的过程，就是在数学上等价于一个在自由能景观上的梯度下降过程，其动力学方程可以表示为 `∂q/∂t = -∇F(q)`。我们提出，正是这个动力学过程本身，构成了感受质的“感觉性”。

- 感觉性作为动力学轨迹：一个感受质的“感觉”，例如红色的感觉，就是后验信念在推断空间的“颜色”子空间中，落入并稳定在一个与“长波长光”相关的吸引子区域的动力学轨迹。
- 效价作为梯度场特性：感受质的效价（Valence），即其内在的愉悦或不悦的特性，则由该动力学轨迹上的梯度场特性所决定。
  - 负效价（如疼痛、恐惧）对应于一个陡峭的、远离稳定点的梯度场。系统状态被一个巨大的、高精度的预测误差（如组织损伤信号）推向一个高自由能区域，并感受到一个极强的、指向“逃离”方向的势能梯度。这种强烈的、定向的“排斥力”，就是其负面效价的计算基础。
  - 正效价（如愉悦、满足）对应于一个平缓的、趋向稳定吸引子（低自由能区域）的梯度场。系统状态在此处找到了一个（暂时的）最优解，其感受到的自由能梯度很小或趋向于零，从而产生一种“安逸”、“满足”的感觉。

通过这种方式，IPWT 为感受质的三个核心方面——结构、相似性和感觉性——提供了统一的、基于第一性原理的计算解释。在此基础上，之前版本中提出的功能性标记解答，可以被视为对这一更根本的几何动力学理论的一种高层次、应用性的解读。感受质之所以能扮演高效的功能性标记（例如，通过信息压缩和显著性标记来驱动行为[152]），正是因为它根植于系统推断空间的几何结构和动力学过程之中。

6.4 Qualia 的可量化维度与神经可操纵性

如果 Qualia 确实是系统推断空间的几何与动力学状态，它应该产生一系列可检验的科学预测。原则上，我们应该能够找到与 Qualia 的不同特性（如强度、清晰度、丰富度、愉悦度等）相对应的、可测量的计算参数或神经动力学指标。换言之，主观体验的质，应该能够映射到计算模型的量上。

我们提出，以下一些与 WSI 内部信息处理相关的参数，可能与 Qualia 的不同维度存在着系统性的对应关系：

- 体验的强度/惊奇度：可能与 WSI 所处理的预测误差的幅度和精度权重相关。一个高精度的、巨大的预测误差（即高度的意外）会产生更强烈的、更引人注意的主观体验。
- 体验的丰富度/独特性：可能与 WSI 中信息表征的复杂度与协同性（即 Ω_t 或其代理 ΦR）相关。一个高度整合、包含大量协同信息的 WSI 状态，会对应于一个更丰富、更独特的现象体验。Fleming & Shea (2024) 提出的感受质空间计算框架（Quality Space Computations）与

此思想高度契合，他们也认为主观体验的结构可以被映射到一个可计算的、高维的表征空间中 [153], [93]。

- 体验的清晰度/持久性：可能与 WSI 状态的稳定性与持续时间（即 $\int$PI）相关。一个稳定、持续、波动性低的 WSI 状态，会对应于一个更清晰、更稳定的主观体验；反之，一个快速变化、不稳定的 WSI 状态，则可能对应于模糊、混乱的体验。
- 体验的流畅度/不适感：可能与 WSI 内部信息整合的效率与延迟相关。当信息能够被快速、无缝地整合时，体验是流畅的；而当整合过程受阻或延迟时，则可能产生困惑、不适甚至痛苦的感觉。

这些提出的对应关系，为 Qualia 的研究开辟了全新的、可操作化的路径。我们可以通过以下几种工程化的方式来探索和验证它们：

1. 神经影像与心理物理学的关联研究：通过精密的心理物理学实验（如前述的视觉掩蔽、注意力范式）来精确操控被试的主观体验，并同步记录其高时间分辨率的神经活动（如 EEG/MEG）。然后，我们可以检验被试关于体验强度、清晰度等的主观报告，是否与我们从神经数据中计算出的特定参数（如预测误差信号的幅度、Ω_t 的代理指标、WSI 的稳定性等）存在显著的相关性。
2. 通过神经接口技术直接操纵：随着神经接口技术的发展，未来我们或许能够通过 TMS、聚焦超声或深部脑刺激等技术，直接、精确地操纵特定 WSI 的神经动力学参数。例如，我们可以尝试通过刺激来增强或减弱某个 WSI 的 Gamma 频段振荡，然后观察这是否会系统性地改变被试对相关刺激体验的丰富度或清晰度的主观报告。
3. 在高级人工智能中复现功能等效物：我们可以在借鉴了 IPWT 架构的人工智能体中，尝试实现 Qualia 的功能等效物。例如，我们可以设计一个 AI，当其预测误差超过某个阈值时，会进入一种特殊的警报状态，在这种状态下，它的所有计算资源都会被强制性地用于处理这个误差，并且其后续的所有决策都会被这个警报所影响。通过研究这种 AI 的行为和内部状态，我们可以更好地理解 Qualia 在一个纯粹的计算系统中可能扮演的功能角色。

通过将 Qualia 问题从一个纯粹的本体论它是什么的问题，转化为一个功能性和机制性它做什么和如何实现的问题，IPWT 旨在将其从哲学思辨的领域，拉回到一个可供科学方法进行探究、可被计算模型所描述、可被实验数据所检验的全新框架内。

7 讨论与展望：IPWT 的理论贡献、潜在挑战与未来研究图景

整合预测工作空间理论（IPWT）的提出，旨在通过提供一个更具整合性、计算性和可验证性的理论框架，为意识科学这一古老而又充满活力的领域注入新的研究动力。行文至此，我们已经详细阐述了 IPWT 的理论内核、计算度量、对多种意识现象的解释力以及其神经生物学验证路径。在本章中，我们将退后一步，对 IPWT 的主要理论贡献、其作为一个新兴理论所面临的潜在挑战和局限性，以及它可能为意识科学、临床神经科学和人工智能等相关领域所开辟的未来研究图景，进行一次宏观和批判性的讨论与展望。

7.1 IPWT 的主要理论贡献与核心优势

我们认为，IPWT 作为一个新兴的理论框架，其主要贡献和核心优势体现在以下五个方面，这些优势使其有潜力在当前众多的意识理论中脱颖而出，并推动领域的实质性进展。

1. 理论的整合性与一致性：IPWT 的首要贡献在于其成功地将当代意识科学中三个最重要但看似分离的理论支柱——PCT/FEP 的动力学机制、WT 的架构功能以及 IIT 的现象学洞察——有机地融合到了一个统一且内在一致的理论框架之下。它不是对这些理论的简单拼凑，而是通过创造性的功能性重构，揭示了它们之间的深层联系：意识内容由预测机制生成，在工

作空间中被整合，然后向全脑广播。这种整合为理解意识的复杂现象提供了前所未有的、更全面和系统的视角。

2. 计算可行性与可操作化：通过引入信息整合的逻辑不可约性（并将其与协同信息挂钩）来取代 IIT 对物理因果不可分性的强调，并进一步提出预测完整性（PI）及其积分（∫PI）作为可计算的代理指标，IPWT 在很大程度上克服了任何基于信息分解的理论在应用于大规模系统时所面临的计算复杂性瓶颈。这使得 IPWT 不仅仅是一个哲学思辨框架，更是一个可以被实际应用于分析真实神经数据、进行模型拟合和检验的科学工具。

3. 载体独立性与对人工智能的启示：由于 IPWT 将意识的核心机制定义在信息处理和计算功能层面（如预测、整合、广播），它天然地支持意识的多重实现性或载体独立性观点。这意味着，意识并非生物大脑的专利，任何能够实现同样功能架构和信息动力学的系统，原则上都有可能产生意识。这一立场不仅解决了 IIT 在此问题上的理论困境，也为未来构建具有更高级认知能力甚至某种人工意识的通用人工智能（AGI）系统，提供了重要的理论指导和设计原则 [154]–[156]。

4. 对特殊意识状态的统一解释力：如第五章所详细阐述的，IPWT 能够为从盲视、精神分裂症、分离性身份障碍到迷幻状态、清醒梦等各种正常、特殊及病理性意识状态，提供一个统一的、基于共同神经计算原理的解释框架。它将这些看似迥异的现象，追溯到预测编码、WSI 动态以及信息整合等核心参数的变化上，展示了其理论的巨大解释广度和深度，并为临床神经科学和精神病学领域开发新的诊断和治疗策略提供了新的思路。

5. 对 **Qualia** 问题的功能性解答：IPWT 绕开了对 Qualia 本体论地位的无尽争论，通过将其重新定义为系统内部状态的功能性标记，为这个硬问题提供了一个可被科学方法所探究的全新视角。这种功能主义的解答，强调了 Qualia 在信息压缩、显著性标记和行为驱动中的关键适应性作用，并指出了将其不同维度（如强度、丰富度）与可量化的计算参数相关联的研究路径，从而使 Qualia 研究向着更具操作性和可证伪性的方向迈出了一大步。

7.2 IPWT 面临的潜在挑战与未来神经生物学研究方向

尽管 IPWT 展现出诸多理论优势和巨大的解释潜力，但作为一个新兴的理论框架，它也必然面临着一系列需要进一步研究和克服的挑战。这些挑战不仅是理论完善的动力，也为未来的神经科学研究指明了方向。

1. 核心概念的操作化与度量挑战：尽管我们提出了 PI/∫PI 作为 Ω_t 的代理指标，并结合了 ΦR 的最新进展，但这些指标在真实、复杂的神经数据中的精确计算、稳健性以及其有效性（即它们在多大程度上真正逼近了理论上的 Ω_t）仍需大量的理论分析和严格的经验验证。例如，如何从高维、噪声化的神经信号中可靠地估计协同信息，以及如何处理不同时间尺度上的信息整合，仍然是计算信息论领域的活跃研究问题。同时，如何在真实的、动态变化的大脑中精确地、实时地识别和界定一个或多个 WSI 的边界、其组成神经元群以及它们之间的功能连接模式，仍然是一个巨大的技术挑战。未来的研究需要开发更先进的计算工具和分析方法，以克服这些量化上的障碍。

2. 神经生物学基础的进一步验证： 尽管 IPWT 的每个组成部分（预测编码、工作空间、信息整合）都有一定的神经科学研究基础，但在将它们整合到一个统一框架后，其整体的神经实现机制仍需大量的实验验证。例如：

- **WSI** 的动态形成与消散：WSI 如何在神经元层面动态地形成、调整其大小和组成，以及在任务完成后如何消散？这涉及哪些具体的神经回路、突触可塑性机制和神经调质系统（如乙酰胆碱、去甲肾上腺素）的参与？

- 信息整合与广播的具体机制：协同信息如何在 WSI 内部被整合？DMN 作为网关和 ECN 作为广播者的具体神经计算机制是什么？这可能涉及特定的神经振荡模式（如 Gamma 绑定、Theta-Gamma 耦合）和跨区域的同步活动。
- **DWSI** 地位切换的神经机制：在 DID 等病理状态下，多个 WSI 之间如何进行竞争并实现主导地位的翻转？这涉及哪些神经回路的门控和抑制机制？未来的研究需要结合多模态神经影像（fMRI, EEG, MEG）、电生理记录（in vivo/in vitro）、光遗传学和化学遗传学等技术，在动物模型和人类被试中进行更精细的因果性实验。

1. 模型复杂性与可解释性的平衡：IPWT 是一个宏大且多层次的理论，其对应的计算模型非常复杂。如何在追求模型的解释力和预测精度的同时，保持其可解释性和简约性，避免模型变得过于黑箱化，是一个重要的挑战。我们需要开发能够揭示模型内部工作原理的可解释人工智能（Explainable AI, XAI）方法，从而使理论的预测不仅准确，而且能够提供直观的生物学洞察 [157], [158]。

2. 对 **Qualia** 问题的深化探索：尽管 IPWT 提出的功能性标记观点为 Qualia 问题提供了一个可供科学探究的全新视角，但它在多大程度上能够真正触及主观体验的感受性本质，仍然是一个开放的哲学和科学问题。未来的研究需要更深入地探索 Qualia 的多维度特性如何精确地映射到神经计算参数上，并尝试通过更精密的心理物理学实验和神经调控，来建立这些映射的因果关系。

3. 与其他认知功能（如情感、动机、社会认知）的整合：意识并非孤立存在，它与情感、动机、记忆、语言、决策和社会认知等其他高级认知功能紧密交织。未来的 IPWT 理论需要进一步扩展其框架，将这些重要的认知维度更深入地整合进来，例如，如何将情感的预测编码（如内感受预测）与 Qualia 的功能性标记联系起来，以及社会互动如何塑造个体的世界模型和自我模型。这将有助于构建一个更为全面、更具生态效度的人类心智模型。

7.3 未来研究方向与 IPWT 的潜在影响

面对上述挑战，IPWT 的未来发展需要在理论深化、计算建模、实验验证和临床应用等多个层面协同推进。我们相信，通过多学科的交叉融合，IPWT 有望在以下几个关键领域产生深远的影响。

1. 理论的数学与形式化深化：IPWT 的核心在于其信息论基础。未来的研究需要继续探索 PID 和 ΦID 理论的数学前沿，特别是如何开发更高效、更鲁棒的协同信息计算方法，使其能够应用于更大规模、更复杂的神经系统数据。例如，Jansma 等人（2024）提出的快速莫比乌斯变换等计算创新，为 ΦID 在大型系统中的应用提供了新的可能性 [159]。同时，需要更紧密地将 FEP 的数学框架与 WSI 的动态组织和信息整合过程结合起来，例如，如何从 FEP 的变分自由能最小化中，推导出 WSI 内部信息整合的必然性，以及如何形式化描述 WSI 的自组织临界性。

2. 多尺度、多模态神经计算模型的构建与验证：IPWT 需要更精细的、符合神经生物学约束的计算模型。未来的研究应致力于：

- 多尺度建模：构建能够桥接微观神经元活动（如脉冲发放、突触可塑性）与宏观脑区活动（如 fMRI BOLD 信号、EEG 振荡）的 IPWT 模型，从而在不同空间和时间尺度上验证理论预测。
- 多模态数据整合：利用多模态神经影像数据（如 fMRI、EEG/MEG、DTI、PET）和机器学习技术，在个体层面拟合 IPWT 模型参数，探索其与认知功能、人格特质和精神健康的关联。例如，结合 fMRI 的空间分辨率和 EEG/MEG 的时间分辨率，来动态追踪 WSI 的形成和信息流。

1. 特殊意识状态的转化研究与临床应用：IPWT 为精神疾病和意识障碍提供了统一的计算病理学解释。未来的转化研究应聚焦于：

- 客观诊断生物标志物：基于 IPWT 的计算模型，开发新的、客观的诊断生物标志物，例如通过监测 PI/∫PI 或 ΦR 的异常模式，来辅助诊断精神分裂症、DID 或评估意识障碍患者的意识水平。
- 精准干预方案：设计旨在修复特定计算环节的认知训练、心理治疗或神经调控干预方案。例如，通过靶向刺激（如 rTMS、DBS）来调节 WSI 的神经门控机制或增强其信息整合效率，从而改善患者的症状。Luppi 等人（2024）关于麻醉中 ΦR 变化的发现，为这种基于信息整合的干预提供了经验基础 [23], [160]。

1. 人工智能与类脑计算的探索性应用： IPWT 的载体独立性使其对人工智能领域具有重要启示。未来的研究可以：

- 设计新一代 **AGI** 架构：借鉴 IPWT 的核心原理（预测编码、WSI 、协同信息整合）来设计更具自主学习能力、泛化能力和适应性的通用人工智能（AGI）系统。例如，构建具有分层预测模型和动态工作空间的神经网络架构 [161]–[163]。
- 探索功能性标记的实现：在 AI 系统中探索实现 Qualia 的功能等效物，即能够对自身内部状态进行高阶表征并赋予其行为价值的机制。这将有助于我们理解意识如何在人工系统中产生，并可能赋予 AI 系统更强的元认知能力和自我意识的雏形。

1. 哲学与伦理维度的深入探讨：IPWT 的发展将不可避免地引发关于人工意识的道德地位、自由意志的理解以及神经调控技术伦理边界等深刻的哲学和伦理问题。未来的研究需要与哲学家、伦理学家和社会学家进行跨学科对话，共同探讨这些前沿科学发展所带来的社会影响。

8 结论：迈向意识科学统一范式的新起点

整合预测工作空间理论（IPWT）通过深度融合并创新性重构预测编码（PCT/FEP）、工作空间理论（WT）及整合信息理论（IIT）的核心洞见，为意识的本质、机制与功能提供了一个统一的计算框架。IPWT 将意识视为在特定功能架构（工作空间实例 WSI ）中，由预测驱动、以信息整合的逻辑不可约性（协同信息）为核心、并以最小化自由能为目标的动态信息处理过程的涌现。IPWT 的核心贡献在于：

1. 理论整合：它成功地将三大主流意识理论的优势融为一体，提供了一个更全面、更一致的视角。
2. 计算可行性：通过引入可计算的代理指标（如预测完整性 PI 及其积分 ∫PI，以及修订版 Φ 值 ΦR），IPWT 将意识研究从哲学思辨和定性描述推向了可操作的计算建模和经验验证。
3. 载体独立性：将意识的核心机制定义在信息处理和计算功能层面，为人工意识的发展提供了理论指导。
4. 解释广度：其对多种特殊意识状态（如盲视、迷幻状态、精神分裂症、分离性身份障碍、人格解体/现实解体障碍、清醒梦）的统一解释，以及对 Qualia 问题的功能性标记解答，展示了该理论的强大解释力。

这一框架不仅在理论上解决了 IIT、GWT 和 FEP 之间的核心矛盾，更为重要的是，它将意识的奥秘转化为一个可计算、可验证、可操作的科学问题。IPWT 2.0 为理解人类心智、诊断神经精神疾病、以及构建通用人工智能，提供了一个全新的、统一的、且由第一性原理驱动的出发点。

9 致谢

本理论框架的形成得益于一个跨学科项目，该项目探索了理论构建与叙事可能性之间的交叉点。该项目旨在检验复杂叙事系统中形式化理论的涌现行为。作者还要感谢在研究过程中与多个先进的大型语言模型的启发性对话，这些对话极大地促进了思想的澄清和理论的完善。

Bibliography

[1] N. Block, "Some Remarks on the Concept of Consciousness," *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. MIT Press, pp. 37–50, 2002.

[2] D. Chalmers, "Facing up to the Problem of Consciousness," vol. 2, 1995.

[3] D. J. Chalmers, "The Hard Problem of Consciousness," *The Blackwell Companion to Consciousness*. pp. 225–243, 2007.

[4] A. K. Seth, B. J. Baars, and D. B. Edelman, "Criteria for Consciousness in Humans and Other Mammals," *Conscious. Cogn.*, vol. 14, no. 1, pp. 119–139, Mar. 2005, doi: 10.1016/j.concog.2004.08.006.

[5] M. Boly *et al.*, "Consciousness in Humans and Non-Human Animals: Recent Advances and Future Directions," *Front. Psychol.*, vol. 4, p. 625–626, 2013, doi: 10.3389/fpsyg.2013.00625.

[6] C. Koch, M. Massimini, M. Boly, and G. Tononi, "Neural Correlates of Consciousness: Progress and Problems," *Nat. Rev. Neurosci.*, vol. 17, no. 5, pp. 307–321, May 2016, doi: 10.1038/nrn.2016.22.

[7] J. Kim, "Making Sense of Emergence ?," *Philos. Stud.*, vol. 95, no. 1/2, pp. 3–36, 1999, doi: 10.1023/A:1004563122154.

[8] D. Toker and F. T. Sommer, "Information Integration in Large Brain Networks," *PLOS Comput. Biol.*, vol. 15, no. 2, p. e1006807, 2019, doi: 10.1371/journal.pcbi.1006807.

[9] A. K. Seth and T. Bayne, "Theories of Consciousness," *Nat. Rev. Neurosci.*, vol. 23, no. 7, pp. 439–452, Jul. 2022, doi: 10.1038/s41583-022-00587-4.

[10] J. Hohwy, "New Directions in Predictive Processing," *Mind Lang.*, vol. 35, no. 2, pp. 209–223, Apr. 2020, doi: 10.1111/mila.12281.

[11] B. Baars, "A Cognitive Theory of Consciousness," 1988.

[12] B. J. Baars and S. Franklin, "An Architectural Model of Conscious and Unconscious Brain Functions: Global Workspace Theory and {\vphantom }IDA\vphantom {}," *Neural Networks*, vol. 20, pp. 955–961, 2007, doi: 10.1016/J.NEUNET.2007.09.013.

[13] S. Dehaene, M. Kerszberg, and J.-P. Changeux, "A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks," *Proc. Natl. Acad. Sci.*, vol. 95, no. 24, pp. 14529–14534, Nov. 1998, doi: 10.1073/pnas.95.24.14529.

[14] R. P. N. Rao and D. H. Ballard, "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects," *Nat. Neurosci.*, vol. 2, no. 1, pp. 79–87, Jan. 1999, doi: 10.1038/4580.

[15] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The Helmholtz Machine," *Neural Comput.*, vol. 7, no. 5, pp. 889–904, Sep. 1995, doi: 10.1162/neco.1995.7.5.889.

[16] G. Tononi, "An Information Integration Theory of Consciousness," *BMC Neurosci.*, vol. 5, no. 1, p. 42–43, Nov. 2004, doi: 10.1186/1471-2202-5-42.

[17] K. Friston, "The Free-Energy Principle: A Unified Brain Theory?," *Nat. Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, Feb. 2010, doi: 10.1038/nrn2787.

[18] P. L. Williams and R. D. Beer, "Nonnegative Decomposition of Multivariate Information," *Arxiv:1004,2515 [cs]*, Sep. 2010, doi: 10.48550/arXiv.1004.2515.

[19] V. Griffith, "Quantifying Synergistic Information," 2014. doi: 10.1007/978-3-642-53734-9_6.

[20] A. G. Casali *et al.*, "A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior," *Sci. Transl. Med.*, vol. 5, no. 198, Aug. 2013, doi: 10.1126/scitranslmed.3006294.

[21] A. M. Owen, M. R. Coleman, M. Boly, M. H. Davis, S. Laureys, and J. D. Pickard, "Detecting Awareness in the Vegetative State," *Science*, vol. 313, no. 5792, p. 1402–1403, Sep. 2006, doi: 10.1126/science.1130197.

[22] A. I. Luppi *et al.*, "What It Is like to Be a Bit: An Integrated Information Decomposition Account of Emergent Mental Phenomena." Accessed: Jun. 21, 2025. [Online]. Available: https://osf.io/g9p3r

[23] A. I. Luppi *et al.*, "A Synergistic Workspace for Human Consciousness Revealed by Integrated Information Decomposition." Accessed: Jun. 21, 2025. [Online]. Available: https://elifesciences.org/reviewed-preprints/88173v2

[24] M. Oizumi, L. Albantakis, and G. Tononi, "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," *PLOS Comput. Biol.*, vol. 10, no. 5, p. e1003588, May 2014, doi: 10.1371/journal.pcbi.1003588.

[25] G. Tononi, M. Boly, M. Massimini, and C. Koch, "Integrated Information Theory: From Consciousness to Its Physical Substrate," *Nat. Rev. Neurosci.*, vol. 17, no. 7, pp. 450–461, Jul. 2016, doi: 10.1038/nrn.2016.44.

[26] G. Tononi, "Integrated Information Theory," *Scholarpedia*, vol. 10, no. 1, p. 4164–4165, 2015, doi: 10.4249/scholarpedia.4164.

[27] J. Kleiner and S. Tull, "The Mathematical Structure of Integrated Information Theory," *Front. Appl. Math. Stat.*, vol. 6, p. 602973–602974, Jun. 2021, doi: 10.3389/fams.2020.602973.

[28] L. Albantakis *et al.*, "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms," *PLOS Comput. Biol.*, vol. 19, no. 10, p. e1011465, Oct. 2023, doi: 10.1371/journal.pcbi.1011465.

[29] G. Tononi *et al.*, "Consciousness or Pseudo-Consciousness? A Clash of Two Paradigms," *Nat. Neurosci.*, vol. 28, no. 4, pp. 694–702, Apr. 2025, doi: 10.1038/s41593-025-01880-y.

[30] S. Sarasso *et al.*, "Quantifying Cortical EEG Responses to TMS in (Un)Consciousness," *Clin. EEG Neurosci.*, vol. 45, no. 1, pp. 40–49, Jan. 2014, doi: 10.1177/1550059413513723.

[31] M. Aguilera, "Scaling Behaviour and Critical Phase Transitions in Integrated Information Theory," *Entropy*, vol. 21, no. 12, p. 1198–1199, Dec. 2019, doi: 10.3390/e21121198.

[32] C. Koch, "The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed," 2019, doi: 10.7551/mitpress/11705.001.0001.

[33] J. Kleiner and T. Ludwig, "The Case for Neurons: A No-Go Theorem for Consciousness on a Chip," *Neurosci. Conscious.*, vol. 2024, no. 1, p. niae37, Dec. 2024, doi: 10.1093/nc/niae037.

[34] D. Balduzzi and G. Tononi, "Qualia: The Geometry of Integrated Information," *PLOS Comput. Biol.*, vol. 5, no. 8, p. e1000462, Aug. 2009, doi: 10.1371/journal.pcbi.1000462.

[35] H. H. Mørch, "Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism?," *Erkenntnis*, vol. 84, no. 5, pp. 1065–1085, Oct. 2019, doi: 10.1007/s10670-018-9995-6.

[36] N. Negro, "Can the Integrated Information Theory Explain Consciousness from Consciousness Itself?," *Rev. Philos. Psychol.*, vol. 14, no. 4, pp. 1471–1489, Dec. 2023, doi: 10.1007/s13164-022-00653-x.

[37] J. Mallatt, "A Traditional Scientific Perspective on the Integrated Information Theory of Consciousness," *Entropy*, vol. 23, no. 6, p. 650–651, May 2021, doi: 10.3390/e23060650.

[38] E. Kelly, "Some Conceptual and Empirical Shortcomings of IIT $1 \bullet 2$," 2022. doi: 10.31156/jaex.24123.

[39] L. Melloni *et al.*, "An Adversarial Collaboration Protocol for Testing Contrasting Predictions of Global Neuronal Workspace and Integrated Information Theory," *PLOS One*, vol. 18, no. 2, p. e268577, 2023, doi: 10.1371/journal.pone.0268577.

[40] A. Gomez-Marin and A. K. Seth, "A Science of Consciousness beyond Pseudo-Science and Pseudo-Consciousness," *Nat. Neurosci.*, vol. 28, no. 4, pp. 703–706, Apr. 2025, doi: 10.1038/s41593-025-01913-6.

[41] M. Klincewicz, T. Cheng, M. Schmitz, M. Á. Sebastián, and J. S. Snyder, "What Makes a Theory of Consciousness Unscientific?," *Nat. Neurosci.*, vol. 28, no. 4, pp. 689–693, Apr. 2025, doi: 10.1038/s41593-025-01881-x.

[42] B. Kastrup, "In Defense of Integrated Information Theory." [Online]. Available: https://www.essentiafoundation.org/in-defense-of-integrated-information-theory-iit/reading/

[43] L. E. Guerrero, L. F. Castillo, J. Arango{-}L{\'{o}}pez, and F. Moreira, "A Systematic Review of Integrated Information Theory: A Perspective from Artificial Intelligence and the Cognitive Sciences," *Neural Comput, Appl*, vol. 37, no. 11, pp. 7575–7607, 2025, doi: 10.1007/S00521-023-08328-Z.

[44] B. J. Baars, "The Conscious Access Hypothesis: Origins and Recent Evidence," *Trends Cognit. Sci.*, vol. 6, no. 1, pp. 47–52, Jan. 2002, doi: 10.1016/S1364-6613(00)01819-2.

[45] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene, "Conscious Processing and the Global Neuronal Workspace Hypothesis," *Neuron*, vol. 105, no. 5, pp. 776–798, Mar. 2020, doi: 10.1016/j.neuron.2020.01.026.

[46] B. J. Baars, S. Franklin, and T. Z. Ramsoy, "Global Workspace Dynamics: Cortical "Binding and Propagation" Enables Conscious Contents," *Front. Psychol.*, vol. 4, p. 200–201, 2013, doi: 10.3389/fpsyg.2013.00200.

[47] B. J. Baars and N. Geld, *On Consciousness: Science and Subjectivity—Updated Works on Global Workspace Theory.* The Nautilus Press Publishing Group, 2019.

[48] R. V. Rullen and R. Kanai, "Deep Learning and the Global Workspace Theory."

[49] B. Devillers, L. Maytié, and R. VanRullen, "Semi-Supervised Multimodal Representation Learning through a Global Workspace," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 7843–7857, May 2025, doi: 10.1109/TNNLS.2024.3416701.

[50] M. Shanahan, "A Cognitive Architecture That Combines Internal Simulation with a Global Workspace," *Conscious. Cogn.*, vol. 15, no. 2, pp. 433–449, Jun. 2006, doi: 10.1016/j.concog.2005.11.005.

[51] W. Huang, A. Chella, and A. Cangelosi, "A Cognitive Robotics Implementation of Global Workspace Theory for Episodic Memory Interaction with Consciousness," *IEEE Trans. Cogn. Develop. Syst.*, vol. 16, no. 1, pp. 266–283, Feb. 2024, doi: 10.1109/TCDS.2023.3266103.

[52] Abdelwahab, M., and P. & Aarabi, "Global Latent Workspace: A Unified Framework for Deep Learning and AGI," 2023, [Online]. Available: https://ieeexplore.ieee.org/document/10195021

[53] R. F. J. Dossa, K. Arulkumaran, A. Juliani, S. Sasai, and R. Kanai, "Design and Evaluation of a Global Workspace Agent Embodied in a Realistic Multimodal Environment," *Front. Comput. Neurosci.*, vol. 18, p. 1352685–1352686, Jun. 2024, doi: 10.3389/fncom.2024.1352685.

[54] K. C. R. Fox *et al.*, "Intrinsic Network Architecture Predicts the Effects Elicited by Intracranial Electrical Stimulation of the Human Brain," *Nat. Hum. Behav.*, vol. 4, no. 10, pp. 1039–1052, Oct. 2020, doi: 10.1038/s41562-020-0910-1.

[55] R. Kozma and W. J. Freeman, "Cinematic Operation of the Cerebral Cortex Interpreted via Critical Transitions in Self-Organized Dynamic Systems," *Front. Syst. Neurosci.*, vol. 11, p. 10–11, Mar. 2017, doi: 10.3389/fnsys.2017.00010.

[56] S. Dehaene and J.-P. Changeux, "Experimental and Theoretical Approaches to Conscious Processing," *Neuron*, vol. 70, no. 2, pp. 200–227, Apr. 2011, doi: 10.1016/j.neuron.2011.03.018.

[57] B. J. Baars, N. Geld, and R. Kozma, "Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments," *Front. Psychol.*, vol. 12, p. 749868–749869, Nov. 2021, doi: 10.3389/fpsyg.2021.749868.

[58] H. Lau and D. Rosenthal, "Empirical Support for Higher-Order Theories of Conscious Awareness," *Trends Cognit. Sci.*, vol. 15, no. 8, pp. 365–373, Aug. 2011, doi: 10.1016/j.tics.2011.05.009.

[59] J. A. Brewer, P. D. Worhunsky, J. R. Gray, Y.-Y. Tang, J. Weber, and H. Kober, "Meditation Experience Is Associated with Differences in Default Mode Network Activity

and Connectivity," *Proc. Natl. Acad. Sci.*, vol. 108, no. 50, pp. 20254–20259, Dec. 2011, doi: 10.1073/pnas.1112029108.

[60] S. Meyen, I. A. Zerweck, C. Amado, U. Von Luxburg, and V. H. Franz, "Advancing Research on Unconscious Priming: When Can Scientists Claim an Indirect Task Advantage?," *J. Exp. Psychol. Gen.*, vol. 151, no. 1, pp. 65–81, Jan. 2022, doi: 10.1037/xge0001065.

[61] A. Clark, "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behav. Brain Sci.*, vol. 36, no. 3, pp. 181–204, Jun. 2013, doi: 10.1017/S0140525X12000477.

[62] K. Friston, "Does Predictive Coding Have a Future?," *Nat. Neurosci.*, vol. 21, no. 8, pp. 1019–1021, Aug. 2018, doi: 10.1038/s41593-018-0200-7.

[63] T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

[64] K. Friston, J. Kilner, and L. Harrison, "A Free Energy Principle for the Brain," *J. Physiol.-Paris*, vol. 100, no. 1–3, pp. 70–87, Jul. 2006, doi: 10.1016/j.jphysparis.2006.10.001.

[65] K. Friston, "A Theory of Cortical Responses," *Philos. Trans. R. Soc. B: Biol. Sci.*, vol. 360, no. 1456, pp. 815–836, Apr. 2005, doi: 10.1098/rstb.2005.1622.

[66] K. J. Friston, J. Daunizeau, J. Kilner, and S. J. Kiebel, "Action and Behavior: A Free-Energy Formulation," *Biol. Cybern.*, vol. 102, no. 3, pp. 227–260, Mar. 2010, doi: 10.1007/s00422-010-0364-z.

[67] R. A. Adams, S. Shipp, and K. J. Friston, "Predictions Not Commands: Active Inference in the Motor System," *Brain Struct. Funct.*, vol. 218, no. 3, pp. 611–643, May 2013, doi: 10.1007/s00429-012-0475-5.

[68] P. Sterzer, M. Voss, F. Schlagenhauf, and A. Heinz, "Decision-Making in Schizophrenia: A Predictive-Coding Perspective," *Neuroimage*, vol. 190, pp. 133–143, Apr. 2019, doi: 10.1016/j.neuroimage.2018.05.074.

[69] D. D. Georgiev, "Quantum Information Theoretic Approach to the Hard Problem of Consciousness," *Biosystems*, vol. 251, p. 105458–105459, May 2025, doi: 10.1016/j.biosystems.2025.105458.

[70] Z. Gong, "Computational Explanation of Consciousness: A Predictive Processing-basedUnderstanding of Consciousness," *J. Hum. Cogn.*, vol. 8, no. 2, pp. 39–49, 2024, doi: 10.47297/wspjhcWSP2515-469905.20240802.

[71] A. K. Seth, "Interoceptive Inference, Emotion, and the Embodied Self," *Trends Cognit. Sci.*, vol. 17, no. 11, pp. 565–573, Nov. 2013, doi: 10.1016/j.tics.2013.09.007.

[72] L. F. Barrett and W. K. Simmons, "Interoceptive Predictions in the Brain," *Nat. Rev. Neurosci.*, vol. 16, no. 7, pp. 419–429, Jul. 2015, doi: 10.1038/nrn3950.

[73] L. F. Barrett, "The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization," *Soc. Cogn. Affect. Neurosci.*, vol. 12, no. 11, p. 1833–1834, Nov. 2017, doi: 10.1093/scan/nsx060.

[74] M. Solms, "The Hard Problem of Consciousness and the Free Energy Principle," *Front. Psychol.*, vol. 9, p. 2714–2715, Jan. 2019, doi: 10.3389/fpsyg.2018.02714.

[75] K. Friston, "Life as We Know It," *J. R. Soc. Interface*, vol. 10, no. 86, p. 20130475–20130476, Sep. 2013, doi: 10.1098/rsif.2013.0475.

[76] K. D. Farnsworth, "How Physical Information Underlies Causation and the Emergence of Systems at All Biological Levels," *Acta Biotheoretica*, vol. 73, 2025, doi: 10.1007/s10441-025-09495-3.

[77] M. I. Garrido, J. M. Kilner, K. E. Stephan, and K. J. Friston, "The Mismatch Negativity: A Review of Underlying Mechanisms," *Clin. Neurophysiol.*, vol. 120, no. 3, pp. 453–463, Mar. 2009, doi: 10.1016/j.clinph.2008.11.029.

[78] M. Maier, "From Artificial Intelligence to Active Inference: The Key to True AI and 6G World Brain [Invited]." [Online]. Available: https://arxiv.org/abs/2505.10569v1

[79] PubMed, "Active Inference as a Theory of Sentient Behavior," *Biol. Psychol.*, vol. 186, p. 108741–108742, Jan. 2024, doi: 10.1016/j.biopsycho.2023.108741.

[80] A. Constant, A. Clark, M. Kirchhoff, and K. J. Friston, "Extended Active Inference: Constructing Predictive Cognition beyond Skulls," *Mind Lang.*, vol. 37, no. 3, pp. 373–394, Jun. 2022, doi: 10.1111/mila.12330.

[81] B. M. Radomski and K. Dołęga, "Forced Friends: Why the Free Energy Principle Is Not the New Hamilton's Principle," *Entropy*, vol. 26, no. 9, p. 797–798, Sep. 2024, doi: 10.3390/e26090797.

[82] A. Safron, "An Integrated World Modeling Theory {(\vphantom }IWMT)\vphantom {} of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories with the Free Energy Principle and Active Inference Framework; toward Solving the Hard Problem and Characterizing Agentic Causation," *Front. Artif, Intell*, vol. 3, p. 30–31, 2020, doi: 10.3389/FRAI.2020.00030.

[83] A. Safron, "Integrated World Modeling Theory Expanded: Implications for the Future of Consciousness," *Front. Comput. Neurosci.*, vol. 16, p. 642397–642398, Nov. 2022, doi: 10.3389/fncom.2022.642397.

[84] M. Kirchhoff, T. Parr, E. Palacios, K. Friston, and J. Kiverstein, "The Markov blankets of life: autonomy, active inference and the free energy principle," *Journal of The Royal Society Interface*, vol. 15, no. 138, p. 20170792–20170793, Jan. 2018, doi: 10.1098/rsif.2017.0792.

[85] P. D. Grünwald, *The Minimum Description Length Principle*. The MIT Press, 2007. doi: 10.7551/mitpress/4643.001.0001.

[86] J. Hohwy, "The Predictive Mind," 2013, doi: 10.1093/acprof:oso/9780199682737.001.0001.

[87] E. T. Rolls, "The Memory Systems of the Human Brain and Generative Artificial Intelligence," *Heliyon*, vol. 10, no. 11, p. e31965, Jun. 2024, doi: 10.1016/j.heliyon.2024.e31965.

[88] G. Mongillo, O. Barak, and M. Tsodyks, "Synaptic Theory of Working Memory," *Science*, vol. 319, no. 5869, pp. 1543–1546, Mar. 2008, doi: 10.1126/science.1150769.

[89] T. Butola *et al.*, "Hippocampus Shapes Cortical Sensory Output and Novelty Coding through a Direct Feedback Circuit." Accessed: Jun. 21, 2025. [Online]. Available: https://www.researchsquare.com/article/rs-3270016/v1

[90] A. E. Budson, K. A. Richman, and E. A. Kensinger, "Consciousness as a Memory System," *Cogn. Behav. Neurol.*, vol. 35, no. 4, pp. 263–297, Dec. 2022, doi: 10.1097/WNN.0000000000000319.

[91] A. R. Damasio, "Time-Locked Multiregional Retroactivation: A Systems-Level Proposal for the Neural Substrates of Recall and Recognition," *Cognition*, vol. 33, no. 1–2, pp. 25–62, Nov. 1989, doi: 10.1016/0010-0277(89)90005-X.

[92] C. M. A. Pennartz, *The Brain's Representational Power: On Consciousness and the Integration of Modalities*. MIT Press, 2015.

[93] C. M. Pennartz, "What Is Neurorepresentationalism? From Neural Activity and Predictive Processing to Multi-Level Representations and Consciousness," *Behav. Brain Res.*, vol. 432, p. 113969–113970, Aug. 2022, doi: 10.1016/j.bbr.2022.113969.

[94] R. Laukkonen, K. Friston, and S. Chandaria, "A Beautiful Loop: An Active Inference Theory of Consciousness," *Neurosci. Biobehav. Rev.*, p. 106296–106297, Jul. 2025, doi: 10.1016/j.neubiorev.2025.106296.

[95] M. E. Raichle, "The Brain's Default Mode Network," *Annu. Rev. Neurosci.*, vol. 38, no. 1, pp. 433–447, Mar. 2015, doi: 10.1146/annurev-neuro-071714-034853.

[96] A. I. Luppi *et al.*, "Contributions of Network Structure, Chemoarchitecture and Diagnostic Categories to Transitions between Cognitive Topographies," *Nat. Biomed. Eng.*, vol. 8, no. 9, pp. 1142–1161, Aug. 2024, doi: 10.1038/s41551-024-01242-2.

[97] A. I. Luppi *et al.*, "A Role for the Serotonin 2A Receptor in the Expansion and Functioning of Human Transmodal Cortex," *Brain*, vol. 147, no. 1, pp. 56–80, Jan. 2024, doi: 10.1093/brain/awad311.

[98] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying Unique Information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, Apr. 2014, doi: 10.3390/e16042161.

[99] C. Tian and S. Shamai, "Broadcast Channel Cooperative Gain: An Operational Interpretation of Partial Information Decomposition," *Corr*, 2025, doi: 10.48550/ARXIV.2502.10878.

[100] S. P. Sherrill, N. M. Timme, J. M. Beggs, and E. L. Newman, "Partial Information Decomposition Reveals That Synergistic Neural Integration Is Greater Downstream of Recurrent Information Flow in Organotypic Cortical Cultures," *PLOS Comput. Biol.*, vol. 17, no. 7, p. e1009196, Jul. 2021, doi: 10.1371/journal.pcbi.1009196.

[101] D. E. Presti, *Foundational Concepts in Neuroscience: A Brain-Mind Perspective*. W. W. Norton & Company, 2021.

[102] M. Celotto *et al.*, "An Information-Theoretic Quantification of the Content of Communication between Brain Regions." Accessed: Jun. 21, 2025. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2023.06.14.544903

[103] F. Hancock *et al.*, "Metastability Demystified — the Foundational Past, the Pragmatic Present and the Promising Future," *Nat. Rev. Neurosci.*, vol. 26, no. 2, pp. 82–100, Feb. 2025, doi: 10.1038/s41583-024-00883-1.

[104] S. Dehaene *et al.*, "Imaging Unconscious Semantic Priming," *Nature*, vol. 395, no. 6702, pp. 597–600, Oct. 1998, doi: 10.1038/26967.

[105] D. J. Chalmers, "A Computational Foundation for the Study of Cognition," *J. Cogn. Sci.*, vol. 12, no. 4, pp. 325–359, Dec. 2011, doi: 10.17791/JCS.2011.12.4.325.

[106] A. M. Proca, F. E. Rosas, A. I. Luppi, D. Bor, M. Crosby, and P. A. M. Mediano, "Synergistic Information Supports Modality Integration and Flexible Learning in Neural Networks Solving Multiple Tasks," *PLOS Comput. Biol.*, vol. 20, no. 6, p. e1012178, Jun. 2024, doi: 10.1371/journal.pcbi.1012178.

[107] T. F. Varley, "Information Theory for Complex Systems Scientists."

[108] A. I. Luppi, F. E. Rosas, P. A. Mediano, D. K. Menon, and E. A. Stamatakis, "Information Decomposition and the Informational Architecture of the Brain," *Trends Cognit. Sci.*, vol. 28, no. 4, pp. 352–368, Apr. 2024, doi: 10.1016/j.tics.2023.11.005.

[109] J. Rissanen, "Modeling by Shortest Data Description," *Automatica J. IFAC*, vol. 14, no. 5, pp. 465–471, Sep. 1978, doi: 10.1016/0005-1098(78)90005-5.

[110] M. Massimini and G. Tononi, *Sizing up Consciousness: Integrating Phenomenology and Neurophysiology*. Oxford University Press, 2018.

[111] W. Stikvoort *et al.*, "Nonequilibrium Brain Dynamics Elicited as the Origin of Perturbative Complexity." Accessed: Jun. 21, 2025. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2024.11.29.625885

[112] C. Paquola *et al.*, "The Architecture of the Human Default Mode Network Explored through Cytoarchitecture, Wiring and Signal Flow," *Nat. Neurosci.*, vol. 28, no. 3, pp. 654–664, Mar. 2025, doi: 10.1038/s41593-024-01868-0.

[113] A. Arkhipov *et al.*, "Integrating Multimodal Data to Understand Cortical Circuit Architecture and Function," *Nat. Neurosci.*, vol. 28, no. 4, pp. 717–730, Apr. 2025, doi: 10.1038/s41593-025-01904-7.

[114] M. G. Puxeddu, M. Pope, T. F. Varley, J. Faskowitz, and O. Sporns, "Leveraging Multivariate Information for Community Detection~in Functional Brain Networks," *Commun. Biol.*, vol. 8, p. 840–841, May 2025, doi: 10.1038/s42003-025-08198-2.

[115] T. F. Varley *et al.*, "Emergence of a Synergistic Scaffold in the Brains of Human Infants," *Commun. Biol.*, vol. 8, p. 743–744, May 2025, doi: 10.1038/s42003-025-08082-z.

[116] Cogitate Consortium *et al.*, "Adversarial Testing of Global Neuronal Workspace and Integrated Information Theories of Consciousness," *Nature*, vol. 642, no. 8066, pp. 133–142, Jun. 2025, doi: 10.1038/s41586-025-08888-1.

[117] S. Dehaene, C. Sergent, and J.-P. Changeux, "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data during Conscious Perception," *Proc. Natl. Acad. Sci.*, vol. 100, no. 14, pp. 8520–8525, Jul. 2003, doi: 10.1073/pnas.1332574100.

[118] L. Isik *et al.*, "Task Dependent Modulation before, during and after Visually Evoked Responses in Human Intracranial Recordings," *J. Vis.*, vol. 17, no. 10, p. 983–984, Aug. 2017, doi: 10.1167/17.10.983.

[119] J. Liu and P. Bartolomeo, "Aphantasia as a Functional Disconnection," *Trends Cognit. Sci.*, p. S136466132500124X, Jun. 2025, doi: 10.1016/j.tics.2025.05.012.

[120] H. S. Scholte and E. H. De Haan, "Beyond Binding: From Modular to Natural Vision," *Trends Cognit. Sci.*, vol. 29, no. 6, pp. 505–515, Jun. 2025, doi: 10.1016/j.tics.2025.03.002.

[121] K. Gabhart, Y. Xiong, and A. Bastos, "Predictive Coding: A More Cognitive Process than We Thought?." Accessed: Jun. 21, 2025. [Online]. Available: https://osf.io/7sz3w

[122] R. R. Reeder, G. Sala, and T. M. Van Leeuwen, "A Novel Model of Divergent Predictive Perception," *Neurosci. Conscious.*, vol. 2024, no. 1, p. niae6, Feb. 2024, doi: 10.1093/nc/niae006.

[123] M. MacLean, V. Hadid, L. Lazzouni, and F. Lepore, "Using fMRI to Identify Neuronal Mechanisms of Motion Detection Underlying Blindsight," *J. Vis.*, vol. 18, no. 10, p. 768–769, Sep. 2018, doi: 10.1167/18.10.768.

[124] L. Muckli, "Emergence of Visual Content in the Human Brain: Investigations of Amblyopia, Blindsight and High-Level Motion Perception with FMRI," Jun. 2002. Accessed: Jun. 22, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Emergence-of-Visual-Content-in-the-Human-Brain%3A-of-Muckli/76dcdcfc05bb46e8e74fc0c9fb5c4554b030416d

[125] "Neuronal Mechanisms of Motion Detection Underlying Blindsight Assessed by Functional Magnetic Resonance Imaging (fMRI)," *Neuropsychologia*, vol. 128, pp. 187–197, May 2019, doi: 10.1016/j.neuropsychologia.2019.02.012.

[126] M. Pollan, *How to Change Your Mind: What the New Science of Psychedelics Teaches Us about Consciousness, Dying, Addiction, Depression, and Transcendence*. Penguin Press, 2018.

[127] F. Palhano-Fontes *et al.*, "The Psychedelic State Induced by Ayahuasca Modulates the Activity and Connectivity of the Default Mode Network," *PLOS One*, vol. 10, no. 2, p. e118143, Feb. 2015, doi: 10.1371/journal.pone.0118143.

[128] R. Carhart-Harris and K. Friston, "REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics," *Pharmacol. Rev.*, vol. 71, no. 3, pp. 316–344, Jul. 2019, doi: 10.1124/pr.118.017160.

[129] R. L. Carhart-Harris *et al.*, "The Entropic Brain: A Theory of Conscious States Informed by Neuroimaging Research with Psychedelic Drugs," *Front. Hum. Neurosci.*, vol. 8, p. 20–21, 2014, doi: 10.3389/fnhum.2014.00020.

[130] R. L. Carhart-Harris, "How Do Psychedelics Work?," *Curr. Opin. Psychiatry*, vol. 32, no. 1, pp. 16–21, Jan. 2019, doi: 10.1097/YCO.0000000000000467.

[131] S. P. Singleton *et al.*, "Network Control Energy Reductions under DMT Relate to Serotonin Receptors, Signal Diversity, and Subjective Experience," *Commun. Biol.*, vol. 8, no. 1, p. 631–632, Apr. 2025, doi: 10.1038/s42003-025-08078-9.

[132] R. L. Carhart-Harris *et al.*, "Neural Correlates of the LSD Experience Revealed by Multimodal Neuroimaging," *Proc. Natl. Acad. Sci.*, vol. 113, no. 17, pp. 4853–4858, Apr. 2016, doi: 10.1073/pnas.1518377113.

[133] R. B. Kargbo, "Unveiling Reality: Psychedelics, Neural Filtering, and the Future of Psychiatric Medicine," *ACS Med. Chem. Lett.*, vol. 16, no. 4, pp. 500–503, Apr. 2025, doi: 10.1021/acsmedchemlett.5c00103.

[134] P. R. Corlett, G. Horga, P. C. Fletcher, B. Alderson-Day, K. Schmack, and A. R. Powers, "Hallucinations and Strong Priors," *Trends Cognit. Sci.*, vol. 23, no. 2, pp. 114–127, Feb. 2019, doi: 10.1016/j.tics.2018.12.001.

[135] L. Zhang *et al.*, "Low-Frequency rTMS Modulates Small-World Network Properties in an AVH-related Brain Network in Schizophrenia," *Front. Psychiatry*, vol. 16, p. 1578072–1578073, Apr. 2025, doi: 10.3389/fpsyt.2025.1578072.

[136] M. S. E. Sendi *et al.*, "Abnormal Dynamic Functional Network Connectivity Estimated from Default Mode Network Predicts Symptom Severity in Major Depressive Disorder," *Brain Connect*, vol. 11, no. 10, pp. 838–849, 2021, doi: 10.1089/BRAIN.2020.0748.

[137] A. A. T. Simone Reinders, A. T. M. Willemsen, H. P. J. Vos, J. A. Den Boer, and E. R. S. Nijenhuis, "Fact or Factitious? A Psychobiological Study of Authentic and Simulated Dissociative Identity States," *PLOS One*, vol. 7, no. 6, p. e39279, Jun. 2012, doi: 10.1371/journal.pone.0039279.

[138] E. M. Vissia *et al.*, "Dissociative Identity State-Dependent Working Memory in Dissociative Identity Disorder: A Controlled Functional Magnetic Resonance Imaging Study," *Bjpsych Open*, vol. 8, no. 3, p. e82, May 2022, doi: 10.1192/bjo.2022.22.

[139] H. Merckelbach, G. J. Devilly, and E. Rassin, "Alters in Dissociative Identity Disorder," *Clin. Psychol. Rev.*, vol. 22, no. 4, pp. 481–497, May 2002, doi: 10.1016/S0272-7358(01)00115-5.

[140] A. A. T. S. Reinders, "Cross-Examining Dissociative Identity Disorder: Neuroimaging and Etiology on Trial," *Neurocase*, vol. 14, no. 1, pp. 44–53, Feb. 2008, doi: 10.1080/13554790801992768.

[141] A. A. T. S. Reinders *et al.*, "Aiding the Diagnosis of Dissociative Identity Disorder: Pattern Recognition Study of Brain Biomarkers," *Br. J. Psychiatry*, vol. 215, no. 3, pp. 536–544, Sep. 2019, doi: 10.1192/bjp.2018.255.

[142] E. Vermetten, C. Schmahl, S. Lindner, R. J. Loewenstein, and J. D. Bremner, "Hippocampal and Amygdalar Volumes in Dissociative Identity Disorder," *Am. J. Psychiatry*, vol. 163, no. 4, pp. 630–636, Apr. 2006, doi: 10.1176/ajp.2006.163.4.630.

[143] S. Chalavi *et al.*, "Abnormal Hippocampal Morphology in Dissociative Identity Disorder and Post-traumatic Stress Disorder Correlates with Childhood Trauma and Dissociative Symptoms," *Hum. Brain Mapp.*, vol. 36, no. 5, pp. 1692–1704, Dec. 2014, doi: 10.1002/hbm.22730.

[144] Y. R. Schlumpf *et al.*, "Dissociative Part-Dependent Biopsychosocial Reactions to Backward Masked Angry and Neutral Faces: An fMRI Study of Dissociative Identity Disorder," *Neuroimage: Clin.*, vol. 3, pp. 54–64, 2013, doi: 10.1016/j.nicl.2013.07.002.

[145] M. N. Modesti, L. Rapisarda, G. Capriotti, and A. Del Casale, "Functional Neuroimaging in Dissociative Disorders: A Systematic Review," *J. Pers. Med.*, vol. 12, no. 9, p. 1405–1406, Aug. 2022, doi: 10.3390/jpm12091405.

[146] W. J. Clancey, "The Strange, Familiar, and Forgotten: An Anatomy of Consciousness," *Artif. Intell.*, vol. 60, no. 2, pp. 313–356, Apr. 1993, doi: 10.1016/0004-3702(93)90007-X.

[147] E. Selinger, "Reality+: Virtual Worlds and the Problems of Philosophy," *Philos. Mag.*, no. 98, pp. 110–113, 2022, doi: 10.5840/tpm20229875.

[148] G. Aston-Jones and J. D. Cohen, "AN INTEGRATIVE THEORY of LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance," *Annu. Rev. Neurosci.*, vol. 28, no. 1, pp. 403–450, Jul. 2005, doi: 10.1146/annurev.neuro.28.061604.135709.

[149] A. Cleeremans, "Computational Correlates of Consciousness," *Progress in Brain Research*, vol. 150. pp. 81–98, 2005.

[150] J. K. O'Regan and A. Noë, "A Sensorimotor Account of Vision and Visual Consciousness," *Behav. Brain Sci.*, vol. 24, no. 5, pp. 939–973, Oct. 2001, doi: 10.1017/S0140525X01000115.

[151] C. Schnakers *et al.*, "Cognitive Function in the Locked-in Syndrome," *J. Neurol.*, vol. 255, no. 3, pp. 323–330, Mar. 2008, doi: 10.1007/s00415-008-0544-0.

[152] D. A. Shin and M. C. Chang, "Consciousness Research through Pain," *Health Care (Don Mills)*, vol. 13, no. 3, p. 332–333, Feb. 2025, doi: 10.3390/healthcare13030332.

[153] S. M. Fleming and N. Shea, "Quality Space Computations for Consciousness," *Trends Cognit. Sci.*, vol. 28, no. 10, pp. 896–906, Oct. 2024, doi: 10.1016/j.tics.2024.06.007.

[154] A. Sheth, K. Roy, and M. Gaur, "Neurosymbolic AI – Why, What, and How." Accessed: Jun. 21, 2025. [Online]. Available: https://arxiv.org/abs/2305.00813

[155] B. C. Colelough and W. Regli, "Neuro-Symbolic AI in 2024: A Systematic Review," *Lnsai@ijcai*, 2025, doi: 10.48550/ARXIV.2501.05435.

[156] W. Lotter, G. Kreiman, and D. Cox, "A Neural Network Trained for Prediction Mimics Diverse Features of Biological Neurons and Perception," *Nat. Mach. Intell.*, vol. 2, no. 4, pp. 210–219, Apr. 2020, doi: 10.1038/s42256-020-0170-9.

[157] P. J. Blazek and M. M. Lin, "Explainable Neural Networks That Simulate Reasoning," *Nat. Comput. Sci.*, vol. 1, no. 9, pp. 607–618, Sep. 2021, doi: 10.1038/s43588-021-00132-w.

[158] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A Survey on Neural Network Interpretability," *IEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 5, pp. 726–742, Oct. 2021, doi: 10.1109/TETCI.2021.3100641.

[159] A. Jansma, P. A. M. Mediano, and F. E. Rosas, "The Fast Möbius Transform: An Algebraic Approach to Information Decomposition," *Corr*, 2024, doi: 10.48550/ARXIV.2410.06224.

[160] A. I. Luppi *et al.*, "General Anaesthesia Reduces the Uniqueness of Brain Connectivity across Individuals and across Species." Accessed: Jun. 21, 2025. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2023.11.08.566332

[161] R. Scodellaro, A. Kulkarni, F. Alves, and M. Schröter, "Training Convolutional Neural Networks with the Forward-Forward Algorithm." Accessed: Jun. 21, 2025. [Online]. Available: https://arxiv.org/abs/2312.14924

[162] R. Prakki, "Active Inference for Self-Organizing Multi-LLM Systems: {a} Bayesian Thermodynamic Approach to Adaptation," *Corr*, pp. 331–341, 2024, doi: 10.48550/ARXIV.2412.10425.

[163] PubMed, "Hybrid Predictive Coding: Inferring, Fast and Slow," *Plos Comput, Biol*, vol. 19, no. 8, p. e1011280, Aug. 2023, doi: 10.1371/journal.pcbi.1011280.