·科学技术哲学·

检验、质疑与发展：整合信息理论是不是伪科学？

Testing, Challenging, and Developing: Is Integrated Information Theory Pseudoscience?

王璞凡 /WANG Pufan[1]　郝刘祥 /HAO Liuxiang[1,2]

（1. 中国科学院大学人文学院，北京，100049；2. 中国科学院哲学研究所，北京，100190）
(1. School of Humanities, University of Chinese Academy of Sciences, Beijing, 100049;
2. Institute of Philosophy, Chinese Academy of Sciences, Beijing, 100190)

摘　要：文章围绕整合信息理论近年来引发的伪科学争议，梳理了该理论与神经元群选择理论、弱整合信息理论及全局神经工作空间理论间的内在关联和张力。一方面，指出该理论的确存在立场激进、可应用性较差以及部分不可证伪等问题。另一方面则强调该理论具有良好的实验实证传统，且能够独立推导出可检验预测，同时具有对新理论的启发潜力等积极面相。此外，论证了学界对该理论具有泛心论倾向的指责并非全然站得住脚。因而建议学界对整合信息理论保持开放态度，进一步关注其实证潜力，至少在当前的情况下，不宜扣上“伪科学”的帽子以否定其积极价值。

关键词：整合信息理论　对抗性合作检验　伪科学　泛心论

Abstract: This article examines the recent controversy surrounding Integrated Information Theory (IIT) and its alleged pseudoscientific status, analyzing its theoretical connections and tensions with the Theory of Neuronal Group Selection (TNGS), weak IIT, and the Global Neural Workspace Theory (GNWT). On one hand, the paper acknowledges that IIT's radical stance, limited applicability, and partially unfalsifiable axioms warrant criticism. On the other hand, it highlights IIT's robust empirical tradition, capacity to generate independently testable predictions, and its potential to inspire new theoretical frameworks. Furthermore, the article argues that accusations of IIT's panpsychist leanings are not entirely justified. Consequently, it advocates for an open-minded approach toward IIT, emphasizing the need to explore its empirical potential rather than prematurely dismissing it as "pseudoscience", given its ongoing contributions to consciousness research.

Key Words: Integrated information theory; Adversarial collaboration; Pseudoscience; Panpsychism

中图分类号：N031; G201　DOI: 10.15994/j.1000–0763.2025.11.004　CSTR: 32281.14.jdn.2025.11.004

引　言

由威斯康星大学神经科学家朱利奥·托诺尼（Giulio Tononi）提出的整合信息理论（Integrated Information Theory, IIT）是近年来意识科学中最富争议性的理论。一方面，关于该理论的对抗性合作检验宣称其为当代“领先”

基金项目：国家社会科学基金重点项目“当代视域下哲学与科学的连续性研究”（项目编号：24AZX021）。
收稿日期：2025年4月3日
作者简介：王璞凡（1994–）男，江苏徐州人，中国科学院大学人文学院博士研究生，研究方向为认知哲学。Email: wangpufan18@mails.ucas.ac.cn
郝刘祥（1965–）男，安徽潜山人，中国科学院大学人文学院、中国科学院哲学研究所教授，研究方向为科学哲学、科学思想史和物理学哲学。Email: haoliu@ucas.ac.cn

的意识理论之一。另一方面，百余位科学家及相关学者又联合签署将其称为“伪科学”的公开署名信。那么，学界应当采取何种态度对待这一理论，便成为一个与科学哲学研究关涉甚密的问题。

简单来说，IIT不仅对意识应当满足的充分必要条件做出了明确的承诺，还试图通过严格的数学形式化体系对主观意识体验进行定量研究。从该理论自2004年正式酝酿提出以来，如今已迭代到第四个版本。在IIT4.0版本的阐述中，该理论明确声称其坚持一种实在论立场，但大部分学者依然认为该理论对意识的激进定义突破了传统的意识研究范式，并为之带来了类似“泛心论”（panpsychism）倾向的批评及“不可证伪”层面的质疑。

本文以科学哲学的视角，通过将IIT与其前承、后继及主要竞争理论予以对照，同时对争议各方的观点进行综合评析，给IIT描摹出一个较为全面的理论画像。

一、继承与分道：从神经元群选择理论到整合信息理论

IIT自其创立伊始便植根于深厚的学术传统之中，因而并不是一个孤立的激进理论。IIT的理论渊源可以追溯到诺贝尔生理学或医学奖得主杰拉尔德·埃德尔曼（Gerald Edelman）于20世纪70年代末的开创性工作——神经元群选择理论（Theory of Neuronal Group Selection, TNGS）。TNGS严格遵循科学研究的认识论和自然主义方法论原则，对当代意识科学的研究产生了深远的影响。托诺尼与埃德尔曼有着多年的合作经历，以合著者的身份与埃德尔曼于2000年出版了系统阐述TNGS的专著《意识的宇宙》（*A Universe of Consciousness*）。[1]因此，托诺尼有时也被认为是TNGS的共同提出者。

TNGS之所以被认为是遵循自然主义进路的典范理论，主要体现于埃德尔曼严格坚持的三条工作假设。这三条假设分别是：物理假设、进化假设和主观（体验）特性（qualia）假设。

> 物理假设是指，对意识的解释无需依赖任何类似二元论的超自然形而上学设定，而只需假定意识是由脑的某些结构和动力学所产生的一类特殊物理过程；进化假设是指，承认意识与生物结构有关，它是依赖于某种形态所产生的动力学过程。同时，这些形态是生物在自然选择的支配下于特定演化阶段出现的产物，因而具有独特的功能和明确的适应性价值，并将影响动物的行为；主观特性假设是说，意识具有私密性（private），但这并不意味着我们在原则上无法提出关于意识的充分必要条件，而只是说，对意识的描述与真实的体验完全是处于两个层面的事情。（[2]，pp.16–17）

这样，TNGS便以一种自然主义的方式将人类经验的第一人称现象世界置于科学探索的出发点，并以意识的基本或普遍性质——整体性（unity）和信息性（informativeness）为着力点，进而建立适当的概念和模型，并对其进行度量与分析。随后，借助神经解剖等实验实证手段探寻人脑中与这些属性相对应的生物过程和机制——复馈（rentry），从而搭建出关于现象意识完整神经机制的理论框架——动态核心假设（Dynamic Core Hypothesis）。

简单地说，复馈就是指神经元信号持续不断地从一个脑区传到另一个脑区，然后又通过由轴突构成的大量并行信道传递回来的交互机制。这些神经元信号在交互过程中，既改变了彼此连接的神经元群的活动，又反过来为其所改变——最终，彼此分化的神经元群的活动因复馈机制在时间和空间上得到了协调整合。埃德尔曼认为，这种协调与整合机制可以为知觉研究中的“绑定问题”（binding problem）提供最佳解释。（[2]，pp.124–125、134–140）因此，埃德尔曼将复馈视为高级脑区内最重要的整合机制。

在以上概念体系的基础上，TNGS提出了其关于意识机制的“动态核心假设”：动态核心不是一种事物或位置，而是一种用神经相互作用来定义的过程。因此，动态核心就是在几分之一秒的时间里彼此有很强相互作用而与脑的其余部分又有明显功能性边界的神经元群聚

类。（[2]，pp.169–170）在这个定义中，有一点需要格外注意，这也是IIT与TNGS之间的显著分别：在TNGS的语境下，动态核心是在人脑中存在的特定的、极为复杂的动态过程，整合性与信息性仅仅由这一过程而产生的意识的两种基本属性。因此，“整合信息”在TNGS中并不被视为意识的充分必要条件。在理论的解释力方面，TNGS可以为处于裂脑、癔症、癫痫状态病人的许多令人困惑的表现做出合理的定性解释，（[2]，pp.74–84）也可以很好地解释为何神经元数目如此之多的小脑对人类意识的产生几乎没起到任何作用。因此，作为意识科学领域早期的奠基性理论之一，TNGS对整个意识学界的影响是深远的。

不过，在同埃德尔曼合著完成《意识的宇宙》后不久，托诺尼又基于TNGS于2004年独立发表了IIT的奠基性论文“意识的信息整合理论”。[3]在该文中，托诺尼首次明确提出了将“信息整合Φ”作为意识核心机制的假说，并初步构建了理论框架。此后，托诺尼及其研究团队通过一系列学术长文与专著的发表，对该理论体系进行逐步地扩展与完善。[4]–[8]经过多次迭代与优化，IIT的4.0版本也于2023年正式发布，[8]形成了包含一个第零公理——存在性（Existence）和五个现象公理——内在性（Intrinsicality）、整合性（Integration）、信息性（Information）、排他性（Exclusion）、构成性（Composition）及其对应物理公设的严格数学形式的意识理论。显然，这一研究进路深受TNGS影响，并且由于IIT继承并吸收了TNGS的大部分实验实证研究成果，所以也在很大程度上承继了TNGS对诸多意识现象的解释力。不过，相比于TNGS，IIT有强烈地向一般化和形式化方向发展的趋势。对此，托诺尼有过明确的阐释：“IIT试图提供一个原则性的进路，将现象学的似乎不可言说的质的属性转译成数学语言。”[4]因而，IIT需要重新定义一套能够与TNGS相区分的概念系统。比如，IIT将现象意识的体验性归结于任一物理系统中存在的“因果效应能力”（cause-effect power），[8]而这一概念可以视为TNGS中复馈机制的一般或形式化表述，两者分别指向各自理论中关于意识体验的生理机制和更具一般性的物理机制。

然而，IIT也因此面临着“内忧外患”的双重考验。首先，IIT的理论框架虽然只基于少数公理和公设，但在实践中根本不可能穷尽地应用这些“公理–公设”来展开现实系统的因果效应能力。与此同时，在历代版本的迭代中，虽然其所依托的数学工具持续优化，但托诺尼也不得不承认，迄今为止任何一个版本的IIT都无法解决计算复杂度爆炸的问题。其次，IIT还面临着来自外部的诸多质疑。比如，埃德尔曼就多次质疑了IIT对意识定义的完备性。具体来说，托诺尼在2005年发文阐述了将整合信息作为衡量意识原则性方法（principled approach）的文章。[9]而埃德尔曼与阿尼尔·赛斯（Anil Seth）在2006年刊发的一篇文章中指出，整合信息并不能作为一个完备的意识测量标准，并依据TNGS的理论内核提出了一个将意识的定量与定性分析相结合的扩展框架。[10]可以看出，埃德尔曼在IIT雏形的孕育阶段便敏锐地察觉到，相比于TNGS，IIT对意识的本质做出了更加激进的定义。2011年，在与全局工作空间理论（Global workspace theory, GWT）的提出者伯纳德·巴尔斯（Bernard Baars）合著的文章中，埃德尔曼仍然坚持这一观点。[11]

作为与埃德尔曼和托诺尼都有过密切交流的知名科学家，赛斯将IIT持有的这种激进立场称为“意识优先”（consciousness-first）进路，以区别于以TNGS等理论为代表的“大脑优先”（brain-first）进路。[12]作为对IIT有深入了解且抱有深刻同情的科学家之一，赛斯以类比的方式来说明他对IIT所持有立场的看法①。在他看来，“整合信息”之于意识应当更类似于“遗传变异”或“生长发育”之于“生命”，而不是“平均分子动能”之于“温度”。也就是说，赛斯认为在科学家们对意识的本质仍众说纷纭

①他认为IIT是目前唯一致力于解决“难问题”的意识理论。

之时，IIT将“整合信息”等同于“意识”的定义方式无疑是激进的。[13]这应当也是大部分对IIT持中立或观望态度学者的立场。

二、启发与互补：从整合信息理论到弱整合信息理论

在科学史中，有不少科学成果是源于科学家对“激进信念”的勇敢尝试与不懈坚持才最终取得的。从某种层面而言，IIT亦属此列。概括说来，IIT以其“意识优先”且强调对意识进行定量计算的理论倾向，启发了一种对处于“最小意识”或“植物人”状态患者意识水平实施评估的新手段的出现——基于经颅磁刺激–脑电图（TMS-EGG）的“扰动复杂性指数 ”（Perturbational Complexity Index, PCI）。该工作由托诺尼与马塞洛·马西米尼（Marcello Massimini）及其团队合作完成，其计算方式可以看做对“整合信息Φ”计算的一种简略近似①——通过量化丘脑–皮层系统对直接扰动的整合响应所包含的信息量来评估意识水平。[14]相较于传统评估方式，PCI大幅提升了评估的适用场景及准确度，展现出了重要临床价值。[15]，[16]

在PCI出现以前，对“边缘意识”患者的状态进行评估的一个关键挑战就是，缺乏一种可靠的、不依赖患者与外界互动能力的客观意识测量指标。上个世纪，评估患者意识水平的方式需要依赖患者的运动或语言反应，因而无法检测无行为输出的意识活动。本世纪初，脑电图（EEG）、功能性磁共振（fMRI）等观测设备应用于临床，可检测患者在头脑中的“意识行为”，成为了效果相对更优的评估方式。比如，英国神经科学家阿德里安·欧文（Adrian Owen）曾于2006年使用fMRI扫描一位因车祸导致双侧前额叶受损且在5个月后仍处于植物人状态的患者，发现这位患者完全可以按照指令分别想象出自己“正在打网球”或“正在参观家里的房间”两种场景。[17]然而，这种评估方式适用的场景仍然具有局限性。因为其仍旧依赖于患者与周围环境互动的能力及其主观体验的表达——尽管只需要在头脑中表达。比如，如果患者无法理解某种特定的语言（有些患者只能理解某种非本地方言）或听觉通路的任何一个环节出现了损伤，都将因无法完成规定的互动环节而被认定为处于持续的无意识状态。

在这种情况下，PCI的特有优势便体现出来了——它不需要患者听懂任何指令，因此也不需要做出任何反应。患者大脑对信号的响应通过经颅磁刺激（TMS）主动对皮层实施扰动来完成，临床数据的收集则由EEG等设备记录。最后，将计算出的PCI数值与参考阈值予以比对即可做出诊断。在这篇具有里程碑意义的文章中，作者认为PCI相较于以往评估患者意识水平指标的最大区别在于：以往的意识指标要么仅关注整合性（如通过皮层激活范围或同步性判断），[18]，[19]要么仅关注分化性（如通过频谱特征判断），[20]，[21]而这两种方式均无法稳定评估意识水平。PCI的成功则在于找到了一个可以兼顾两种指标的数学形式。

因此，PCI在某种程度上可以看做是对IIT追求“意识的定量研究”的馈赠。此外，PCI的提出及其在临床评估上的优异表现也在一定程度上验证了IIT关于“意识与大脑分布式因果交互的信息复杂性相关”这一理论假说。[3]，[4]值得一提的是，PCI也是一个以开发实用意识测量工具为目的的全新意识研究进路——弱IIT进路的标志性工作之一。[22]，[23]

弱IIT的纲领阐述于2022年的一篇综述文章中。[24]它保留了自IIT创立以来的关键现象学洞察（即信息性与整合性），致力于验证信息整合动力学与意识特定方面之间的解释性关联假说。[25]也就是说，弱IIT摒弃了意识与整合信息相等同这一激进主张，仅将整合信息视为意识的解释性关联指标，而非必要或充分条件②。因此，弱IIT既不承诺IIT公理–公设的

①即计算TMS扰动触发的皮层激活时空模式经归一化后的Lempel-Ziv复杂度。
②这甚至弱于TNGS，但弱IIT非常强调实用的数值计算，因而它也继承了IIT相对于TNGS而言的另一种独有特征。

完备性与唯一性，也不主张其理论的适用性可推广至大脑之外。在这种情形下，如果说IIT致力于解决“难问题”，那么弱IIT对“难问题”则持有不可知立场。毕竟，该研究进路的主要目的就是规避IIT在可操作性上的困境，以便直接基于神经生理学的可观测变量构建实用算法。该纲领还指出，IIT与弱IIT可形成互补共生的研究范式。即，IIT侧重通过理论突破为新实用测量工具的研发提供方向，而弱IIT则注重通过测量工具的应用实效评估来反哺其理论发展。

需要额外说明的是，弱IIT虽然脱胎于IIT，但因其去理论化的实用主义面相，导致PCI等指标也展现出与其他主流意识理论，如全局神经工作空间理论（Global Neural Workspace Theory, GNWT）的兼容性。具体来说，GNWT的提出者之一斯坦尼斯拉斯·迪昂（Stanislas Dehaene）从另一个视角解读了马西米尼的工作。[26]他通过观察马西米尼在研究中收集到的关于被试反应强度和时长的数据，发现只要信号传播进入全脑网络，并且激活超过300毫秒，就可以判定患者存在意识。

然而，这种数据解读层面上的吻合不应被理解为当代意识理论在经验预测方面已经出现了趋同。恰恰相反，IIT关于“后部皮层颞–顶–枕部热区”是意识产生的必要脑区的核心主张与GNWT强调的“前额叶全局广播机制”之间存在不可调和的冲突，这也被学界概括为“前脑理论”与“后脑理论”的冲突。这种根本性分歧反映了当代意识科学尚未就意识的必要神经相关物（The neural correlates of consciousness, NCCs）究竟位于何处这一问题达成共识——至少以GNWT与IIT为代表的两大阵营仍各自保持着相对独立的解释框架和预测边界。接下来，笔者将以二者于2019年参与的对抗性合作检验为例，扼要地讨论围绕IIT产生的诸多争议话题。

三、棋逢对手：GNWT与IIT的对抗性合作检验

2019年，托诺尼及其研究团队携IIT参与了一项备受关注的对抗性合作检验，其竞争对手是迪昂等人提出的全局神经工作空间理论。

对抗性合作检验，即互相竞争的理论双方（或多方）在中立第三方的调解和仲裁下进行的一种旨在解决科学争议，推动理论进步的检验活动。[27]此次对抗性合作检验可以说是针尖对麦芒的强强对话，引发万众瞩目。检验实验分别就两个理论在可精确解码意识内容的脑区位置、意识内容的维持机制与脑区间的通信模式三个方面作出的预测来评估两个理论的表现。其第一阶段的实验于2021年宣告完成，并于2023年经同行审议后发布。[28]检验结果在多数学者的意料之中：两个理论做出的预测都不能与大脑的实际活动完全匹配，虽然总体而言IIT更优。对IIT来说，后部皮质内缺乏持续同步是其最直接的挑战，而GNWT面临的最大挑战则涉及其对维持意识内容机制的解释，特别是刺激消失时缺乏其在预注册中预测的点火过程。因此，两种理论均需要修订，只是程度略有不同。在该报告的最后，IIT与GNWT的支持者及其团队均在维持其理论核心主张不变的情形下，对实验结果与理论不相适应的部分做出了回应，并提出了进一步实施实验验证的方向。因此，从这一阶段性结果来看，该检验并没有任何一方获得压倒性的胜利。

此次对抗性合作检验有不少值得我们注意的方面，对于我们重新审视IIT并评判两个理论之间的争论有所启发。从该检验的一份协作协议中可以看到，其不仅期望能提供支持或反驳这两大理论的决定性证据，还希望能为大规模、协作化、开放式的科学实践树立创新范式。[29]但就其最终的检验结果来看，该对抗性合作检验显然没有达到其第一个预期目标。从科学哲学的视角来看，这一目标从理论上来讲就是不可能达到的：“迪昂–蒯因”论题（Duhem-Quine Thesis）对这一点有过较为明确的论述：

> 在任何情况下，任何陈述都可以被认为是真的，如果我们在系统的其他部分作出足够剧烈的调整的话。因此，即使一个很靠近

> 外围的陈述面对着顽强不屈的经验，也可以借口发生幻觉或者修改被称为逻辑规律的那一类的某些陈述而被认为是真的。反之，由于同样的原因，没有任何陈述是免受修改的。[30]

在当前，IIT与GNWT还都远远称不上为成熟的意识理论，因而在面对实验检验的不利证据时，它们反而有着更多的“缓冲空间”去调整其辅助假说，甚至以一种颇具“特设性假说”意味的方式为自身开脱。

不过，此次对抗性合作检验也在一定程度上起到了积极促进作用：它迫使理论双方分别以其并不擅长的实验范式接受检验，并直面其所做出的错误预测。有研究表明，在400多个已发表的意识实验中，完全可以仅根据一个实验的设计范式来预测该实验是支持哪一种意识理论的，并且在这一预测中无需了解关于该实验的任何结论。这一统计特征在GNWT与IIT中体现得尤为明显。[31]该研究揭示，二者在以往实验中得到的实证支持部分源于其研究团队在实验设计时产生的系统性偏差，也就是说其陷入了“自证循环”。即，无论是GNWT还是IIT的研究者，均通过选择与该理论预测高度适配的实验范式、测量指标或分析方法，使实验结果天然倾向于支持该理论做出的预测。因而在实验设计、流程操作无明显问题的情况下，此次对抗性合作检验逼迫二者首次直面“否定性”证据，并公开承认其理论应当做出重要修正，无疑对两个理论的进一步完善具有积极的促进作用。

现在，我们重新把目光关注于IIT，看它是如何得出“后部皮层颞–顶–枕部热区是意识产生的必要脑区”这一主要推论的。实际上，IIT在面对这一问题时，并不涉及Φ值的严格定量计算。简单来说，IIT认为意识的必要神经相关物应当对应于大脑中具有最大Φ值的区域。而神经解剖学的证据显示，大脑后部皮层（如后顶叶、颞叶及感觉区）的递归连接架构最符合这一要求。因此，IIT便近似得出了参与此次对抗性合作检验的核心推论。但也正是为此，引发了来自其竞争对手GNWT及其他意识理论团队的质疑。比如，意识的高阶理论（Higher-Order Theory）的支持者刘克颁（Hakwan Lau）就在一篇长文中指出，此次对抗性合作检验根本没有检验IIT的核心部分，即与整合信息Φ值计算相关的推论，因而不能算作对IIT的有效检验。[32]笔者认为，这一质疑有其合理的成分，但要视不同的情形而定。如果是为了以此来证明IIT是一种没有检验蕴含的伪科学，便属于言之过重了。

四、作为伪科学的整合信息理论？

就在本次对抗性合作检验的阶段性结论公布后不久，2023年9月，一封题为“整合信息理论之为伪科学”（Integrated Information Theory as Pseudoscience）的公开信由124位来自意识科学及相关领域的知名科学家与学者联名发表，对IIT的科学地位提出强烈质疑，在学界引发了轩然大波。该公开信声称IIT缺乏可证伪性，同时其隐含的泛心论（panpsychism）倾向更是与当代科学所遵循的形而上学立场相违背，因而可能会引发一系列牵涉广泛的伦理问题。

关于IIT具有不可证伪性与泛心论倾向的质疑，笔者试着从科学哲学的视角予以评析，并大致勾勒IIT的理论画像。笔者认为，对这些质疑的产生，应当首先予以理解。毕竟，作为在科学哲学领域曾引发广泛争论的重要概念，科学家们对科学理论“可证伪性”的理解有所偏颇也是无可厚非的。因此，即便是原本对理论双方持中立态度的科学家也完全有可能因为其对“可证伪性”的片面理解转而对IIT持质疑态度。加之IIT还展现出了与物理主义相拒斥的“泛心论”面相，因而更容易引发当代科学家的警惕和质疑。在此，笔者的核心观点是，在以“意识”这一难以捉摸的事物为研究对象的科学研究进程中，许多关于“何为科学理论”的朴素直觉可能都未必站得住脚、经得起推敲。比如，如果一位研究者认为在一个科学理论中不应存在无法被证伪的部分，那么他自然就会因此将IIT与“顺势疗法”一道归为伪科学而

予以拒斥。这位研究者可能没有意识到，其实任何科学理论都有在一定时期内几乎无法被证伪的本体论承诺或形而上学假设，只不过许多时候这些难以被证伪的假设会被研究者默认为“自明”的公设而不自知。比如，牛顿力学中的绝对时空假设（虽然牛顿尝试用水桶实验来证明绝对空间假设的可靠性，但马赫的批判表明其说服力仍然是有限的，同时也是可能最终被证伪的）在一定的时期内就几乎是无法被予以实证检验或定量测试的，但很多认同这一理论的研究者可能从未意识到这一点——要么从未考虑过牛顿力学的本体论承诺问题，要么直接将绝对时空视为不证自明的公理公设，从而依旧没能进一步深入考察其可证伪性与该理论本身的内在联系。这将使他们天然地倾向于认同“证伪主义”的观点，即一个科学理论的任何部分都应当是可证伪或承诺其为可证伪的。科学哲学界早已就“证伪主义”的局限性取得了共识性的认知，因而，单凭IIT中存在部分在当前无法证伪的内容便将其归为伪科学，无疑是过于严苛的。

那不可证伪的标准线应当划在何处呢？笔者认为应当综合该理论具有的解释与预测能力、创造性与启发性等因素来考量。仅就IIT而言，如前文所述，它不仅可以在类似对抗性合作检验的实验中提供部分可检验的预测，还可以对诸如癫痫、裂脑人以及无梦睡眠等意识现象予以解释。另外，随着光遗传学的逐步成熟，[33]关于Φ值运算的可检验预测也即将可以实施初步验证。其大致原理为：根据IIT，大脑中不活跃的神经元与已经失活的神经元虽然均处于“不活跃”状态，但其扮演的角色是完全不同的。前者是存在潜在活跃可能性的，因此在关于Φ值运算的转移概率矩阵等计算环节中与完全失活的神经元可予以明确区分（其他意识理论均只考虑神经元群的实际状态，而不考虑其潜在状态）。在这一前提下，光遗传技术可以通过LED阵列照射控制神经元在“不活跃”与“失活”状态中切换，进而通过被试的内省报告或其他对照条件对IIT在当前看似不可证伪的理论假设实施经验检验。

关于IIT潜在的泛心论倾向，托诺尼曾专门撰文予以回应。[34]他指出IIT虽然并非以泛心论为出发点，但确实与泛心论的核心直觉一致，即IIT将“意识”视为一种基本属性。这一观点说明，托诺尼所持有的本体论承诺必然有别于物理主义。根据物理主义，一切事物都是在物理的意义上存在，或者随附于物理而存在的。托诺尼既然将意识作为世界的基本属性，就无法持有物理主义的本体论承诺。不过，在心灵哲学的讨论中，作为本体论承诺的物理主义并不一定优于其他形而上学立场。对于这一点，托诺尼显然也有着明确的认知。他认为，物理主义是一种将主观性从现实中剥离出的、方便人类从操作者和观察者的视角理解世界的实用主义立场。虽然自伽利略以来，在这种立场的加持下，人类科学取得了丰硕的成果，但他认为这仍然只是整个世界的一部分——传统的物理主义似乎忽略了从经验出发的内在视角，而泛心论的核心直觉可以作为对这一视角的合理补充。

可以看出，托诺尼并不排斥“泛心论”以及具有“泛心论”色彩的其他形而上学立场。在托诺尼看来，不同于取消的物理主义或实体二元论，泛心论是一种优雅的一元论。但是，托诺尼同样不认为IIT是一种严格意义上的泛心论。根据泛心论，心灵性质不仅应当是基本的，而且应当是普遍的。然而，根据IIT，并非一切事物都具有意识（心灵性质）。比如，电子、沙粒没有心灵性质，因此也不会面临泛心论的组合难题；聚合体没有心灵性质；复杂系统也可能没有心灵性质等。因而，从这一视角来看，IIT也不是严格意义上的泛心论。笔者认为，与托诺尼立场最为接近的形而上学立场应当是罗素一元论（Russellian Monism），这一点在此便不展开了。

当然，质疑者对于IIT立场的批判并非针对托诺尼的阐释，而是针对IIT所展现出的理论特征。关于这一点，有研究指出，学界对于IIT持有泛心论的批评恰恰是由于IIT以形式化的方式明确了关于意识的充分必要条件而导致的。这是因为，人们可以根据这些充要条件轻

易地构造出一个具有意识的可想像结构。然而，根据人们关于意识的常识性理解，此结构如果有意识将会被认为是违反直觉的，因而最终形成了对该理论的泛心论倾向攻击。同样，任何立足于物理主义但声称提出了意识的充分必要条件的理论，由于作为研究对象的意识具有私密性和主观性，也很难以经验检验的方式验证该理论所声称的充分必要条件能否真正“产生意识”。[35] 比如，GNWT如果在“意识就是信息在工作空间中的全局广播”之基础上进一步明确量化的充要条件，其他研究者同样可以根据该充要条件构造出反直觉的“意识”结构——比如某个满足该充要条件的、经过精心设计的“广播站”。一方面，这种“广播站”有意识，会与人们的泛心论直觉相吻合从而引发泛心论倾向的质疑。另一方面，如何证明这种“广播站”有意识同样是一个无法回避的难题。因此，可以说IIT正是因为给出了关于意识的充要条件而具有泛心论和不可证伪倾向的。或许也正是为此，托诺尼也无法在本体论承诺上坚持物理主义的立场。

当然，这并不是说没有能够评判意识理论真假优劣的标准了。但是无论如何，在当今的意识科学领域，科学家们应当对当代心灵哲学与一般科学哲学的相关讨论及形成的基本共识予以深入理解，进而才能以更为开放和更具活力的研究纲领吸纳与融合其他意识理论中的闪光点——而不是过早地以“泛心论”或缺乏可证伪性为由将某些仍具发展潜力的理论归为伪科学。毕竟，在许多哲学家看来，当面对意识的“难问题”时，泛心论其实并不是一个比物理主义更差的形而上学立场（同样的，罗素一元论相比于物理主义也有其优长）。可以说，泛心论为当代意识科学提供了新的理论视角。正如查尔莫斯在其关于泛心论研究的一组文章中指出：相比于其他形而上学观念，泛心论并不是一个需要被更多指责的形而上学立场。因为，至少就目前而言，没有任何一种关于意识的立场在面临“难问题”拷问时能够维持其融贯性，并在丝毫不令人感到割裂的情况下完整地诠释关于现象意识的一切。[36]，[37]

结　论

本文大致梳理了IIT的发展历程，并以近年来围绕IIT的伪科学质疑为线索，从科学哲学视角较为系统地描摹出了IIT的理论画像。总体来说，它植根于TNGS的学术传统，却又因激进的“意识优先”进路而与其分道扬镳。其实，IIT的“意识优先”进路也可以理解为对现象意识特征与物理因果结构之间平行性的承诺。如果将IIT视为一个成熟理论，这一承诺在当前看来的确过于激进。但若仅将其作为一个亟待深入探索的研究纲领，其激进程度就仍在可接受的范围之内。与此同时，它对意识采取定量研究的激进追求还伴随着启发性，在临床上催生了具有里程碑意义的意识水平测量指标PCI。另外，IIT也并非一种不可检验的“伪科学”，并且随着新技术的发展，它还将获得更多的检验蕴含。

当然，尽管IIT为意识的量化研究提供了一种可能的途径，但其理论的完备性与合理性仍存在很大的商榷空间。毕竟，当前的IIT仍旧面临着计算复杂度爆炸、经验检验停留在定性层面等问题，其基本立场同主流的物理主义之间也存在张力。不过，笔者以科学哲学的视角结合相关研究指出，很多类似指责在很大程度上都是由意识的“难问题”面相带来的极端复杂性而引发的误解。此外，笔者相信在IIT的批判者中不乏天然地持有“证伪主义”等较为传统科学哲学立场的科研工作者。或许在他们系统地了解相关科学哲学理论之后，能够对IIT持有更加积极的看法。综上，笔者认为将“伪科学”的帽子扣在IIT头上是不合理的，IIT仍不失为当代极具研究价值的意识理论之一。最后，笔者呼吁意识科学界的科学家和学者们能够充分吸收科学哲学界的相关研究成果，以更加开放的研究纲领开展意识理论的研究工作。

［参 考 文 献］

[1] Edelman, G. M., Tononi, G. *A Universe of Consciousness: How Matter Becomes Imagination*[M]. New York: Basic Books, 2000.

[2]杰拉尔德·埃德尔曼、朱利欧·托诺尼. 意识的宇宙物质如何转变为精神[M]. 顾凡及 译, 上海: 上海科学技术出版社, 2004.

[3]Tononi, G. 'An Information Integration Theory of Consciousness'[J]. *BMC Neuroscience*, 2004, 5: 1–22.

[4]Tononi, G. 'Consciousness as Integrated Information: A Provisional Manifesto'[J]. *Biological Bulletin*, 2008, 215(3): 216–242.

[5]Oizumi, M. 'From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0'[J]. *PLoS Computational Biology*, 2014, 10(5): e1003588.

[6]Tononi, G. 'Integrated Information Theory of Consciousness: An Updated Account'[J]. *Archives Italiennes de Biologie*, 2012, 150: 56–90.

[7]Tononi, G. *The Integrated Information Theory of Consciousness: An Updated Account*[M]. Cambridge: MIT Press, 2023.

[8]Tononi, G. *Phi: A Voyage from the Brain to the Soul*[M]. New York: Pantheon Books, 2012.

[9]Tononi, G. 'Consciousness, Information Integration, and the Brain'[J]. *Progress in Brain Research*, 2005, 150: 109–126.

[10]Seth, A. K., Izhikevich, E., Reeke, G. N., et al. 'Theories and Measures of Consciousness: An Extended Framework'[J]. *Proceedings of the National Academy of Sciences*, 2006, 103(28): 10799–10804.

[11]Edelman, G. M., Gally, J. A., Baars, B. J. 'Biology of Consciousness'[J]. *Frontiers in Psychology*, 2011, 2: 4.

[12]Seth, A. K. 'The Worth of Wild Ideas'[EB/OL]. Nautilus, https://nautil.us/the-worth-of-wild-ideas-399097/. 2023–09–27.

[13]阿尼尔·赛斯. 意识机器: 成为你自己[M]. 桥蒂拉 译, 北京: 中译出版社, 2023, 50–51.

[14]Casali, A. G., Gosseries, O., Rosanova, M., et al. 'A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior'[J]. *Science Translational Medicine*, 2013, 5(198): 198ra105–198ra105.

[15]Re, V. L. 'Role of Transcranial Magnetic Stimulation (TMS) Combined with Electroencephalography (EEG) in Disorders of Consciousness (DOC)'[J]. *Journal of the Neurological Sciences*, 2021, 429: 118507.

[16]Casarotto, S. 'Stratification of Unresponsive Patients by an Independently Validated Index of Brain Complexity'[J]. *Annals of Neurology*, 2016, 80: 718–729.

[17]Owen, A. M. 'Using Functional Magnetic Resonance Imaging to Detect Covert Awareness in the Vegetative State'[J]. *Archives of Neurology*, 2007, 64(8): 1098–1102.

[18]Engel, A. K., Singer, W. 'Temporal Binding and the Neural Correlates of Sensory Awareness'[J]. *Trends in Cognitive Sciences*, 2001, 5(1): 16–25.

[19]Kotchoubey, B. 'Event-related Potential Measures of Consciousness: Two Equations with Three Unknowns'[J]. *Progress in Brain Research*, 2005, 150: 427–444.

[20]Johnson, R. W., Shore, J. E. 'Relative-Entropy Minimization with Uncertain Constraints—Theory and Application to Spectrum Analysis'[R]. Washington: Defense Technical Information Center, 1984.

[21]Pincus, S. M., Gladstone, I. M., Ehrenkranz, R. A. 'A Regularity Statistic for Medical Data Analysis'[J]. *Journal of Clinical Monitoring*, 1991, 7(4): 335–345.

[22]Sarasso, S. 'Consciousness and Complexity: A Consilience of Evidence'[J]. *Neuroscience of Consciousness*, 2021, 7(2): 1–24.

[23]Massimini, M. 'A Perturbational Approach for Evaluating the Brain's Capacity for Consciousness'[J]. *Progress in Brain Research*, 2009, 177: 201–214.

[24]Mediano, P. A. M., Rosas, F. E., Bor, D., et al. 'The Strength of Weak Integrated Information Theory'[J]. *Trends in Cognitive Sciences*, 2022, 26(8): 646–655.

[25]Seth, A. K. 'Explanatory Correlates of Consciousness: Theoretical and Computational Challenges'[J]. *Cognitive Computation*, 2009, 1(1): 50–63.

[26]斯坦尼斯拉斯·迪昂. 脑与意识[M]. 章熠 译, 杭州: 浙江教育出版社, 2018, 258–260.

[27]Bateman, I., Kahneman, D., Munro, A., et al. 'Testing Competing Models of Loss Aversion: An Adversarial Collaboration'[J]. *Journal of Public Economics*, 2005, 89(8): 1561–1580.

[28]Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., et al. 'An Adversarial Collaboration to Critically Evaluate Theories of Consciousness'[J]. *BioRxiv*, 2023, 2023.06.23.546249.

[29]Melloni, L., Mudrik, L., Pitts, M., et al. 'An Adversarial Collaboration Protocol for Testing Contrasting Predictions of Global Neuronal Workspace and Integrated Information Theory'[J]. *PLOS ONE*, 2023, 18(2): e0268577.

[30]蒯因. 从逻辑的观点看[M]. 江天骥 等译, 上海: 上海

译文出版社，1987, 40–41.

[31] Yaron, I., Melloni, L., Pitts, M., et al. 'The ConTraSt Database for Analysing and Comparing Empirical Studies of Consciousness Theories'[J]. *Nature Human Behaviour*, 2022, 6(4): 593–604.

[32] Lau, H. 'What is a Pseudoscience of Consciousness?'[J]. *Lessons from Recent Adversarial Collaborations*, 2023, 1–15.

[33] Deisseroth, K. 'Optogenetics: 10 Years of Microbial Opsins in Neuroscience'[J]. *Nature Neuroscience*, 2015, 18(9): 1213–1225.

[34] Tononi, G., Koch, C. 'Consciousness: Here, There and Everywhere?'[J]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2015, 370(1668): 20140167.

[35] Kleiner, J., Hoel, E. 'Falsification and Consciousness'[J]. *Neuroscience of Consciousness*, 2021, 7(1): niab001.

[36] Chalmers, D. 'Panpsychism and Panprotopsychism'[A], Alter, T., Nagasawa, Y. (Eds.) *Consciousness in the Physical World: Perspectives on Russellian Monism*[C], New York: Oxford University Press, 2015, 102–154.

[37] Chalmers, D. J. 'The Combination Problem for Panpsychism'[A], Brüntrup, G., Jaskolla, L. (Eds.) *Panpsychism: Contemporary Perspectives*[C], New York: Oxford University Press, 2017, 179–214.

［责任编辑　王巍 谭笑］

JDN