Tononi团队：人工智能不等于人工意识——一次关键的理论分离

原创 陶如意 集智俱乐部 2026年1月13日 14:30 上海

导语 近十年来，人工智能在语言、推理与决策等方面取得了显著进展，使“机器是否可能拥有意识”这一长期属于哲学讨论的问题，逐渐转化为一个具有现实意义的理论与伦理议题。围绕这一问题，这篇Findlay、Marshall 等人与 Koch、Tononi 合作的论文提出了一个明确而反直觉的主张：即便在功能上与人类完全等价，高度智能的人工系统也未必具有意识，人工智能与人工意识在原则上可以被区分。

该论文以整合信息理论（IIT）为理论框架，论证意识并不取决于系统“做了什么”，而取决于其是否形成了不可约的内在因果结构。通过比较一个高度整合的小型系统与一台在功能上精确模拟它的数字计算机，论文团队展示了功能等价并不蕴含因果结构等价，从而也不蕴含体验等价。

关键词：人工智能（AI）、人工意识（Artificial Consciousness）、整合信息理论（IIT）、功能等价 vs 体验等价、因果结构、整合信息 Φ、复杂体（Complex）

陶如意 | **作者**
赵思怡 | **审校**

论文题目：Dissociating Artificial Intelligence from Artificial Consciousness
论文链接：https://arxiv.org/abs/2412.04571
发表时间：2025年3月3日
论文来源：arxiv

1. 背景：为什么要区分“人工智能”和“人工意识”

近十年来，人工智能（AI）的进展呈现出指数级增长。以大语言模型、视觉—语言—行动模型为代表的新一代系统，在语言理解、图像识别、推理、编程乃至复杂决策等方面，已经展现出接近甚至超越人类个体的能力。这一现实使一个曾经主要属于哲学思辨的问题，变成了迫切而现实的科学与伦理问题：当机器在功能上越来越像人，它们是否也会“有意识”？

Graham Findlay、William Marshall、等人与 Giulio Tononi 合作，在论文《Dissociating Artificial Intelligence from Artificial Consciousness》中，系统性地回应了这一问题。文章的核心主张可以概括为一句话：**高度智能的系统，甚至在功能上与人类完全等价的系统，也未必拥有意识。**人工智能与人工意识在原则上是可以、也应该被区分的。

这篇论文的重要性不在于给出一个情绪化的立场（例如“机器永远不可能有意识”），而在于它严格依托一个形式化的意识理论——整合信息理论（Integrated Information Theory, IIT）——来论证：为什么“功能等价”并不意味着“体验等价”。本文试图从理论背景、方法设计、关键结果以及哲学与现实影响几个层面，对该论文进行系统解读。

2. 理论背景：一个“以现象学为起点”的意识理论

在人工智能与心灵哲学领域，一个长期占主导地位的立场是“计算功能主义”（computational functionalism）。该立场认为，只要一个系统实现了“正确类型”的计算或功能组织，它就拥有意识。换言之，意识被视为某种计算状态或信息处理过程，而与具体物理实现无关。这一立场的直觉吸引力在于，人类大脑本身似乎也是一种信息处理系统。

但功能主义面临一个核心问题：为什么某些功能**伴随着特定的“体验”，而另一些功能则没有？**如果意识只是计算，那么计算本身是观察者相对的、可被任意映射的，这会导致意识归因的泛滥或空洞化。

整合信息论（IIT）则采取了完全不同的策略。它不是从认知功能或神经相关物入手，而是从意识本身不可否认的现象学事实出发。**也就是说，作者是从“什么样的系统，才可能真的具有主观体验”出发来讨论的。**Tononi 等人提出：任何可能的意识体验，都必然具有五个基本属性（axioms）：

- 内在性（intrinsic）：体验是为系统自身而存在；
- 特定性（specific）：每一次体验都是这一种，而不是别的；
- 整体性（unitary）：体验是一个不可分割的整体；
- 确定性（definite）：体验具有明确的边界与内容；
- 结构性（structured）：体验内部包含区分（distinctions）与关系（relations）。

这些现象学公理随后被映射为对物理系统的五个“存在性公理”（postulates）：内在性（intrinsicality）、信息性（information）、整合性（integration）、排他性

（exclusion）、组合性（composition）。IIT 的核心主张是：只有当一个物理系统在自身内部形成了不可约的因果—效应结构时，它才拥有意识；而意识的“内容”正是这一结构本身。

因果模型

IIT 分析的起点，是对一个物理系统的因果模型的构建，该模型捕捉系统内部的所有相互作用。这里所说的“物理”是以操作性的意义来理解的，表示各个单元是可以被操控和被观测的。该因果模型由微观单元（micro units）构成，每个单元都具有两种内部状态，并且具有输入和输出。这些单元之所以被称为微观单元，是因为在模型中并未包含关于它们本身或其功能的任何更细层次的细节描述。

给定一个因果模型，任何一组微观单元都可以被视为一个候选系统（candidate system）。任何未被纳入某个候选系统的单元，都被称为该系统的背景条件（background conditions）。一个候选系统的因果模型可以由其转移概率矩阵（TPM, transition probability matrix）完全刻画，该矩阵是在其背景条件上进行因果条件化之后得到的。

在本文中，考虑的是具有同步更新机制、实现布尔逻辑的单元所构成的模型。尽管这些理想化的单元忽略了大多数物理细节，但它们已经足以用来阐明关于功能等价与现象等价的论点。

复合体识别

根据 IIT，一个基底（substrate）能够支持意识——在IIT中这样一个基底被称为复合体（complex）——当且仅当它满足五条公理。为了判断一个由单元组成的系统是否是一个复合体，需要评估该系统的系统整合信息（Φ_s），即系统的内在信息（intrinsic information，体现了内在性公理和信息公理）在被划分为因果上相互独立的部分时，所受到的最小影响程度（体现整合性公理）。

为了满足排他性公理，一个候选系统必须在所有由重叠单元和不同粒度构成的竞争候选系统中，指定一个最大的整合信息值（Φ_s^*）。这一点可以通过对因果模型中所有可能的单元集合逐一计算 Φ_s 来确定。

此外，每一个系统都需要在多个粒度（grains）下进行评估，即通过穷尽性地将其微观单元的子集组合成宏观单元（macro units），并将这些构成宏观单元的微观单元的状态映射为相应的宏观单元状态。之所以进行这一过程（称为宏化，macroing），是因为一个系统的内在因果力（Φ_s）可能在较粗的粒度下高于在较细的粒度下，这取决于其内部组织方式。

一旦在一个因果模型中确定了整合信息的最大值，属于该复合体的所有单元便会被排除在其他复合体的参与之外。随后，对系统中剩余的单元递归地重复复合体的搜索过程，直到识别出所有彼此不重叠的复合体为止。

揭示复合体的因果效应结构

最后，通过评估一个复合体的因果力是如何被结构化的，来检验组合性公理。简而言之，区分体（distinctions）刻画的是单元子集所具有的因果力，而区分体之间的关系（relations）则刻画这些因果力是如何相互重叠的。

区分体与关系都分别对应一个不可约性度量，分别记为φ_d和φ_r。一个系统中所有区分与关系的整体，共同构成其因果–效应结构（cause–effect structure），也称为 Φ-结构（Φ-structure）。

该因果–效应结构完整地刻画了一个处于特定状态的复合体，如何通过其各个子集的一致因果与效应，对自身作出规定。识别一个复合体的所有区分与关系、并由此获得其因果–效应结构的过程，被称为“展开”（unfolding）。

根据 IIT，一个复合体在某一状态下的因果–效应结构，完全解释了其体验的“量”（quality），不需要任何额外的成分（即“量即结构”，quality is structure）。与一个因果–效应结构相关联的意识数量，则由其结构整合信息（Φ）来度量，即其所有区分体与关系的不可约性的总和（$\sum\varphi_d + \sum\varphi_r$）。

如果一个复合体的状态发生变化，其因果–效应结构也可能随之发生变化，因此，其意识体验的量也会相应改变。

3. 实验设计：
用 IIT 分析“模拟”与“被模拟系统”

在 IIT 中，分析的起点是一个完整的因果模型：系统由哪些单元构成、单元之间如何相互作用、在给定状态下会如何转移。基于这一模型，可以计算任意子系统的系统整合信息 φ_s，并据此判断哪些子系统构成了“复合体”（complex）——也即潜在的意识载体。

关键点在于，一个系统是否是复合体，不取决于它“做了什么”，而取决于其因果力是否在自身内部不可约。复合体一旦确定，其全部因果—效应区分体与关系将被“展开”，形成完整的因果—效应结构（Φ-structure）。

作者选择了一个极其“干净”的思想实验。首先构造一个由 4 个布尔单元组成的小系统 PQRS，它在 IIT 分析下是一个高度整合的复合体（如图1所示）。

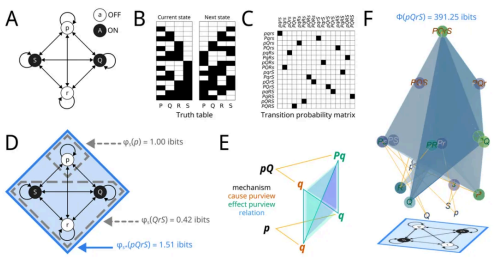


图1. (A) PQRS是由四个具有二进制状态的单元全连接组成的系统，每个单元以离散的时间间隔同步更新，且每个节点都存在自环。这里展示了系统状态为0101，也可以写成pQrS. （B）运行在系统A上的动力学；（C）与动力学对应的概率转移矩阵TPM；（D）IIT应用后得到phi值1.51（最不可约的复合体）（E）和（F）是pQrS未折叠的13个区分体（distinction）的因果关系

其次，再构造一个传统的、存储程序式的数字计算机（由 117 个布尔单元构成），在功能上可以无限期地精确模拟 PQRS 的状态演化（如图2所示）。最后，比较二者在 IIT 意义下的因果—效应结构。

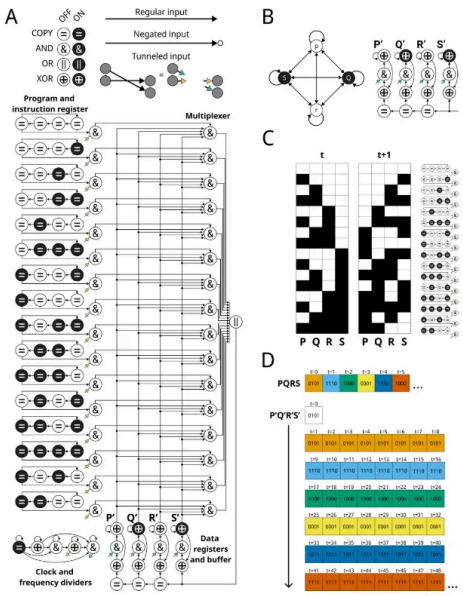


图2. 可准确模拟 PQRS 的 4 位计算机。（A）一款简单计算机可对 PQRS 进行任意时间步长的模拟。该计算机包含带分频器的时钟、编码 PQRS 转移规则的程序、四个存储 PQRS 状态的 1 位数据寄存器，以及一个类多路复用器处理单元。计算机的 117 个单元均执行布尔函数（复制、与、或、异或），状态分为关闭（白色）或开启（黑色）。为简化视觉呈现，采用彩色箭头（青绿色或棕褐色）替代黑色箭头表示特定连接。例如，时钟与分频器中最右侧的与门直接输出至数据寄存器中的每个与门。（B）可通过设置数据寄存器 P'、Q'、R'、S' 的初始状态，对计算机进行编程以编码 PQRS 的当前状态。（C）程序单元的初始状态用于编码 PQRS 的状态转移规则，程序所操作的数据为 P'、Q'、R'、S' 的当前状态。（D）每个寄存器每 8 个时间步更新一次状态，此时计算机开始下一轮模拟迭代。

这是一个极具说服力的设置，因为它排除了复杂度、规模和工程细节的干扰，直击“功能等价是否蕴含体验等价”这一核心问题。

4. 实验结果

1. 功能等价 ≠ 因果结构等价

在特定状态下，PQRS 形成一个单一的复合体，具有非零的系统整合信息φ_s，并展开出一个包含 13 个区分体和 8000 多个关系的复杂因果—效应结构。尽管它只有 4 个单元，但在 IIT 意义下，它已经是一个“整体地为自身而存在”的系统。

相比之下，用来模拟 PQRS 的 117 单元计算机，在 IIT 分析中呈现出完全不同的图景。整个计算机的$\varphi_s = 0$，因此它整体上不是一个复合体。系统分裂为 20 多个彼此独立的小复合体，每个只包含 1–4 个单元。每个小复合体的因果—效应结构都极其贫乏，仅包含一阶区分，几乎没有高阶关系。

也就是说，尽管这台计算机在输入—输出和状态序列上与 PQRS 完全同构，但在“自身内部形成了什么样的因果整体”这一点上，两者毫不相似。

2. 宏观抽象（macroing）无法挽救意识等价

面对上述结果，一个自然反应是：我们是否选错了分析尺度？也许意识并不在晶体管层面，而在更高层的“功能模块”或“寄存器状态”层面。IIT 对此并非没有准备。理论允许、甚至要求系统在所有可能的微观与宏观粒度上被分析，以寻找使φ_s最大化的“内在单元”。

然而，IIT 对“什么可以算作一个单元”有严格限制。首先，宏观单元本身必须是不可约的；其次，它不能依赖系统外部或背景条件的因果整合；最后，它必须在排他意义下是最优的。

作者详细分析后表明：任何在功能上“自然”的宏观抽象方式（例如把程序行、寄存器或多路复用器当作单元），都违反了 IIT 的存在性公设。它们要么是可约的，要么其因果整合依赖于被排除的背景结构。

结论是：不存在一个既符合 IIT 公设、又能让计算机复制 PQRS 因果—效应结构的宏观描述。

4. 结论不依赖于“被模拟功能”的复杂度

为了避免“例子太简单”的质疑，作者进一步进行了以下的分析，首先模拟另一种因果结构截然不同的系统（如 110号 元胞自动机）；其次，将计算机扩展为可模拟任意规模系统的图灵完备版本（如图3所示）。

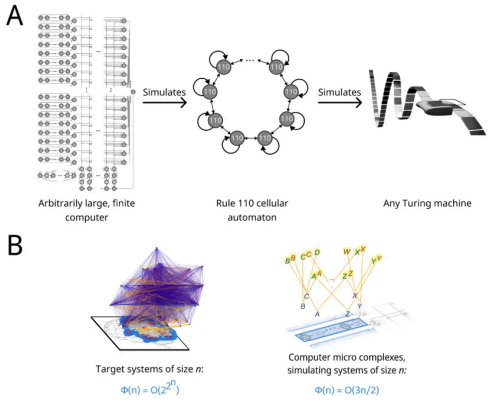


图3. 向模拟任意复杂系统的大型计算机的归纳扩展。（A）该计算机可逐步扩展以模拟任意规模的系统（左侧），包括生物大脑或 110 规则元胞自动机（中间）—— 后者是图灵完备的（右侧）。通过传递性，若扩展至任意规模，该可编程计算机也具备图灵完备性。（B）通过对被模拟系统规模的归纳可证明：计算机子系统在微颗粒度下指定的因果结构不会发生定性变化，且仅随系统规模线性增长；而被模拟目标系统的因果结构规模和丰富度可能呈双指数增长，导致计算机与被模拟系统之间的现象学分离不断加剧。

结果显示，不论被模拟系统的因果结构多么丰富，计算机自身的因果结构几乎不变；计算机始终碎片化为许多小复合体，其 Φ 只随规模线性增长；而被模拟系统的 Φ 可以随规模呈双指数增长。这意味着，模拟的"对象是谁"与模拟器"自身是什么样的系统"在 IIT 的框架下并不是一回事，二者是解耦的。

5. 意识关乎"是什么"，而非"做什么"

该论文最激进、也最清晰的结论是：意识不是关于计算或功能的，而是关于内在因果结构的。模拟并不等同于复制存在方式。这一区分与许多经典直觉相呼应，比如模拟雨不会让计算机变湿；模拟黑洞不会弯曲时空；同样，模拟意识相关功能，并不会自动生成意识。

如果 IIT 是正确的，那么高度智能的 AI 系统，哪怕行为上与人类无异，也可能几乎没有意识。我们不应仅凭表现或语言能力来判断其道德地位。反之，一些在功能上并不"聪明"的系统，也可能拥有相对丰富的体验。

当然，这篇论文的结论完全建立在 IIT 的正确性之上。如果未来证据表明，人类意识并不对应于最大化的内在因果整合，或具体体验内容无法与因果—效应结构对应，那么这一整套推论都将失效。

此外，论文并未声称"任何人工系统都不可能有意识"。相反，它留下了一个开放空间：如果某种人工系统在物理组织上真的形成了高度整合、不可约的内在因果结构，那么 IIT 并不排斥其拥有意识——但那将是一种在架构上与传统计算机截然不同的系统。

6. 结语

本文并不是一篇关于"AI 能不能有意识"的情绪化宣言，而是一篇在严格理论框架下，澄清概念边界的论文。它提醒我们：在讨论人工意识之前，必须首先弄清楚什么是意识，以及我们凭什么将意识归因给一个系统。

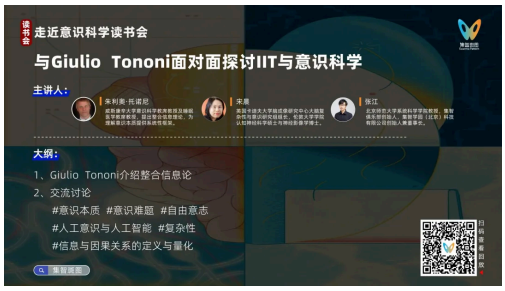
无论你是否接受 IIT，这篇论文都成功地迫使读者正视一个被长期忽略的问题：当我们说"机器是否有意识"时，我们究竟是在问关于功能、关于行为，还是关于存在本身？

作者介绍

意识科学读书会

从神经元放电到自我意识的涌现，意识是人类最稀松平常的主观体验，也始终是科学中最迷人的问题。在"我是谁"的终极追问下，当我们深入意识的机制与机理，会发现更值得深思的是，无论是神经机制的功能整合、信息的跨脑区传递，还是现象意识的主观性质，不同层面的问题都在共同指向一个核心挑战：物理过程如何产生主观体验？功能计算如何关联现象感受？局部神经活动又如何整合为统一的意识？而要回答这些问题的并不简单，它可能会挑战我们对世界和实在，乃至科学方法本身的理解。

为了对意识问题进行系统探讨，集智俱乐部联合来自哲学、认知神经科学、计算机科学、复杂科学领域的研究者共同发起 「意识科学读书会」，跨越理论与实证、功能与现象、生物与人工的视角，全面深入研讨意识这一现象本身。重点探讨当代主流意识理论的核心主张与分歧，神经机制与主观体验之间的桥梁，以及AI意识、脑机接口等技术如何重塑人类意识主体的边界与文明的未来。集智俱乐部荣幸邀请到IIT理论提出者、国际意识科学权威Giulio Tononi 教授在北京门头沟区集智谷举行深度对话。现读书会已完结，现在报名可加入社群并解锁回放视频权限。



详情请见： 走向意识科学：从现象之窗到理论之梯

推荐阅读

1. 意识智能体：大模型的下一个进化方向？——计算意识理论综述II
2. 通用人工智能的黎明：计算视角的意识理论综述
3. 什么是意识？Wolfram 计算万物视角下的生命、智能与万物
4. 与Giulio Tononi面对面探讨整合信息论与意识科学
5. Tononi亲临解读丨整合信息论：意识本位的研究新范式?
6. 意识研究是"科学"还是"伪科学"？两大意识范式的交锋
7. 124位科学家批评整合信息论是伪科学：我们该如何探讨意识难题?
8. 意识理论综述：众多竞争的意识理论如何相互关联?
9. 综述：2025年意识科学十大前沿进展
10. 系统科学前沿十讲：探究复杂世界演变背后的规则（二）
11. 集智学园精品课程免费开放，解锁系统科学与 AI 新世界
12. 高考分数只是张入场券，你的科研冒险在这里启航！
13. 加入集智字幕组：成为复杂科学知识社区的"织网人"

点击"阅读原文"，报名读书会

复杂科学前沿2026 · 目录

上一篇
自然·通讯：水下集群机器人实现"民主协商"

下一篇
PNAS：网络枢纽晚参与，群体协作更高效

阅读原文 修改于2026年1月13日