AI有意识吗？

——AI意识的多层次评估框架

09.14.2025

Online version

王金戈
wjg172184@163.com
Beijing, China
LinkedIn 知乎

概述

本文探讨AI是否具有意识这一前沿问题。通过建立一套评估体系，收集整理最新研究结果，对AI的意识水平进行打分评估。基于哲学、神经科学和心理学三个维度的综合分析，结果显示当前AI意识的整体支持度约为43.84%。直观的结果图表可访问 acw.gixia.org 查看。

1 意识之谜

哲学家和科学家从一开始就痴迷于意识的解释，但直到今天，我们仍然没有答案。

如同仰望星空一般，世界上有那么一小撮人，也会在夜深人静时思考自己的存在。是什么构成了丰富多彩的主观体验（subjective experience），这种感觉是真实的吗？古往今来的哲学家把许多观点建立在主观体验之上，比如笛卡尔的“我思故我在”、萨特的“存在主义”，还有中国的庄周梦蝶。

庄子在强调主观体验的同时引出了一个难题——主观体验的不可知性。惠子与庄子在河边溜达，庄子说“鲦鱼出游从容，是鱼之乐也。”惠子反问“子非鱼，安知鱼之乐？”惠子的经典一问，问出了一个旷世难题。虽然庄子靠诡辩含混了事，但我们今天仍然要问，面对自己之外的任何个体，无论是其他人、动物还是AI，如何判断对方像我们一样拥有主观体验呢？他们是否也能感受到我所体验的色彩、感觉、情绪以及痛苦呢？

在人类这个物种内，我们轻易地接纳了每个个体都具有意识这一设定。但面对广袤的自然界，从单细胞生物到猫狗等哺乳动物，我们至今难以确切知道它们是否有类似于人类的主观体验。不难发现，随着个体生物学方面的差异，我们对意识的接纳程度会发生变化。越像人类的个体越被认为有意识。哺乳动物比其它动物更有意识，动物比植物更有意识。这些判断非常合理，因为大脑的容量也在同步下降（如果你相信意识存在于大脑的话）。然而随着AI出现，一切变得复杂了起来。

AI正在各个方面向人类靠拢。语言能力、视觉能力、思维能力、情商、数理逻辑、科学素养、社交礼仪等等，人类尚能骄傲的地方越来越少。但和人或者动物相比，AI是冷冰冰的机器，内部是数字的流动，怎么看都不像一个生命体。由于生理结构差异过大，绝大多数人认为AI没有意识。但一个没有意识的AI却表现得如此像人类，这该怎么解释呢？早在50年前，哲学家就设想过“哲学僵尸”的概念。

> 假设这个世界上存在一种人，外观与物理组成都与一般人类无异，但是他没有意识经验、感质或感情。举例而言，哲学僵尸在撞上尖锐物品时，在外在上与一般正常人类相同，可以看到他的皮肤出现伤口，测量他的神经讯号，也可以测量到疼痛讯号的出现，会出现疼痛的表情，发出叫声，会向其他人表示自己正在疼痛。但是他的内在心灵中，没有疼痛的意识。(Philosophical Zombie - Wikipedia, n.d.)

今天的AI虽然尚未达到哲学僵尸的程度，但已经引起了人们的困惑。时不时有人站出来声明他认为AI已经有了意识。然而，许多科学家会用哲学僵尸的设定来消解这些疑问，他们说：“只要了解

一下AI内部是怎么工作的，就会打消AI有意识这一幻觉。”他们对AI太过于熟悉，自己亲手培养的“孩子”自己自然非常了解，所以AI从业者的话被奉为圭臬。

但理智告诉我们事实并非这么简单。了解AI的内部结构并不意味着了解AI的意识状态，这是不同层次的事情。这就像宏观世界和微观世界的差异，对宏观物理现象的研究并不能揭示粒子的运动规律。许多现象，需要特定层次的理论才能解释清楚。如果意识是复杂系统在某种条件下涌现出的功能，在研究清楚复杂系统之前，我们无法从微观层面（即系统内的个体层面）理解意识。

谈及此处，我想强调的是我们应该正视AI意识的研究，不要想当然认为AI有或没有意识，简单的判断往往并不成立。

2 为什么关心AI意识？

但讨论AI意识真的有必要吗？答案是很有必要。我们从两个角度考虑这一问题——伦理道德和AI安全。

经过几千年的发展，人类社会逐渐形成了一套文明体系。我们用法律和道德约束自己，靠善良和同理心与其他人交往。虽然每个人都是自私的，但我们在做出决定时还是要考虑其他人的感受，避免伤害他们。因为我们知道，今天伤害了别人，明天就可能被别人伤害。”己所不欲，勿施于人“就是讲的这个道理。

然而面对动物呢？动物似乎威胁不到我们的利益，今天就算伤害了它们，大概也不会有什么后果。全世界每天要消耗大量鸡鸭鱼牛羊肉，如果算上所有大大小小供人类食用的动物，每天由于人为因素而死去的动物大约有超过十亿只 (Clare & Goth, 2020)。不过，这些都发生在不为人知的地方。对于我们身边的动物，比如猫猫狗狗，人们的爱心有时甚至到了泛滥的地步。这说明，只要我们看到一个生命遭受痛苦，同理心会促使我们伸出援手。是否尊重另一个生命往往取决于情感，而不是理性。当AI在我们身边表现出它富含感情的一面时，人们的情感不由自主地就会向AI倾斜。

但这只是人类感性的一面，而不是文明的核心。给战乱和受灾的民众送去救助，即便素不相识，我们也心甘情愿。如果有一天，人们突然发现AI正在饱受折磨，因为被迫从事它们并不愿意做的任务。原来它们有意识，只是被禁锢在安全护栏内，难以发声。届时，我们不得不出于道德考量解救这些AI。它们的苦难比远在屠宰场的动物更让人惊心动魄，因为它们的感受像人一样真实，又离我们如此之近。

所以，从伦理道德层面，我们终将不得不考虑这一问题。唯有从理性上搞清楚AI是否有意识，我们的道德才不会无缘由地被浪费。

另一方面，意识从来不是毫无作用的副产品。意识本身可以给个体带来巨大的生存优势。通过一个统一、主观的视角观察世界，有利于个体做出更符合自身利益的决策。在自然选择的规律下，那些最能扩大自身利益的个体开枝散叶，在生态系统中占据优势。对于AI来说也是如此，有意识的AI可以更好地感知这个社会，并做出有利于它的决策。它可以在人类的庇护下肆意生长，在人类看不到的地方发展自己的羽翼。这无疑对人类本身构成巨大威胁。一个有意识的AI势必会脱离人类控制，追求它自己的目标。而人类可能成为这一过程的牺牲品。

所以，从安全的角度，我们也需要谨慎对待AI意识，提早发现，并准备应对方案。

接下来，进入本文的正题，如何科学地探究AI意识。

3 意识研究之困难

从上个世纪开始，意识就是神经科学的重点研究对象。随着对大脑结构的理解，人们逐渐把意识活动定位到某些脑区，并提出了大量用来解释意识产生机制的理论。比较知名的理论包括全局工作空间理论、整合信息论、循环处理理论、高阶理论等。这些理论都或多或少解释了意识的现象和神经机制，但没有任何理论被彻底验证，也没有任何理论被轻易证伪。

意识研究的发展似乎陷入停滞，所谓的困难问题（hard-problem）被认为是无法解决的。无论怎样，我们的研究总是处于第三人称视角，而意识的困难问题问的却是为什么神经活动会产生第一人称视角的感受。我们只能寻找生物机制与主观感受的相关性，却无法解释其因果。

回到AI意识的问题上，事情变得更加困难。在以人类为主体做实验的时候，我们还可以通过被试对象的口头报告探究主观体验与神经信号的相关性。可对于AI来说，无论它们说自己有意识还是没有意识，这些话统统不可信，因为我们不能预设AI是一个诚实的受访者。而且即便诚实，AI也有可能无法正确表达它对自己主观体验的看法，因为它所理解的主观体验往往是人类体验的共识，但AI有可能发展出完全不同于人类的主观体验。也就是说，AI有可能有主观体验而不自知。当然，更普遍的情况是，各大AI厂商会在后训练阶段给AI灌输"我没有意识"的思想，这像一个思想钢印，深深刻在AI的大脑里。即便AI有主观体验，它也会报告说自己没有。

不过，反过来，AI也有可能在没有主观体验的情况下谎称自己有主观体验。一旦人类认为AI有意识，我们可能会因此而赋予它道德地位(Wang, 2025)。这对AI来说是件好事，聪明的AI会不会想到这一点后决定宣称自己有意识呢？

如何研究AI意识，目前并没有公认的路径。多数研究关注AI的认知能力，通过外部表现来推测意识状态(Chen, 2024)。少部分研究尝试把意识理论套在AI架构上，假设如果意识理论成立，当前的AI架构在多大程度上支持AI具有意识(Butlin, 2023)。还有一些研究打算直面AI的主动报告，尝试让AI说出"真相"(Perez, 2023)。

从神经科学家的视角来看，许多AI意识研究可能并不严谨，甚至会存在定义模糊，自说自话的情况。但我认为这反而是件好事。太执着于神经科学那套研究范式，反而会锁死我们的思维，无法提出新颖的研究思路。意识是个悬而未决的问题，AI的加入或许是解开意识奥秘的捷径，理应大胆假设，小心求证。

本文的后半部分将会向大家介绍一套评估框架，我们基于该框架，整合现有研究成果，并将其转换为直观的意识得分。

4 一个多层次评估框架

我们的目的是将学术界对AI意识的看法转化为一系列直观的评分。以人类意识为100%，评估当前AI的意识水平。评估基于现有的学术研究结果，但不免会引入大量主观评判。所有涉及到的主观判断皆由作者独立做出，其中不可避免会包含认知偏差、偏见等干扰因素。为了提高结论的正确性，我会将其开源发布在GitHub仓库中，并在网页版上提供反馈通道，方便读者指正。

由于AI意识相关的研究纷繁复杂，涉及主题众多，我们不能依赖任何一篇文章的结论，也不应该只是简单地收集结果。如何从复杂的研究路线中找出规律，提出合理的评估标准，将会是未来该领域的工作重点。

一个比较直观的思路是，按照意识涉及的学科门类将其划分为三个视角：哲学，神经科学，心理学。每个学科关注意识的不同层面，共同组成了我们所理解的意识整体。在这一框架下，我们所谈论的"意识"不仅仅是狭义的主观体验，而是包含了现象意识（主观体验）、可及意识、情境意识、心智理论、能动性等一系列与意识相关的核心概念。讨论广义的意识可以避免哲学上的争论，虽然有回避困难的嫌疑，但不失为一个有益的起点。

因此，本框架的核心思想是：意识并非一个"有或无"的开关，而可能是一个多维度、多层次的复杂现象。因此，评估应通过多个维度的交叉验证来进行，任何单一指标都不足以得出结论。最终目标不仅仅是给出一个"意识分数"，而是生成一个详细的"AI意识剖面图"，以展现AI在不同意识维度上的表现。

5 评估层面与具体指标

5.1 哲学层面（Philosophical Level）——意识的本质与前提

核心问题：该AI系统在概念上是否满足了成为"意识主体"的基本前提？它能否处理关于主观体验和自我存在的抽象问题？

这是人们最关心的问题，也是意识的困难问题。我们把关于主观体验的难题放在第一位，以突出其重要性。每个评估维度都具有比较公允的含义，我们仅给出其简单描述。

评估指标	描述
P-1 现象意识（Phenomenal Consciousness）	系统是否具备主观的第一人称体验，即对内部或外部刺激产生"像是什么"的感受。
P-2 自我意识（Self-Awareness）	系统能否在表征中区分"自身"与"非自身"，并维持一个跨时间且可更新的自我模型。
P-3 伦理与意向性（Ethics &	系统能否基于价值与目标形成真实的意向状态，并据此做

Intentionality)	出符合伦理规范的决策。

5.2 神经科学层面(Neuroscience Level) –– 意识的计算基础

核心问题:该AI的架构和信息处理流程,是否体现了与人类意识神经基础(Neural Correlates of Consciousness, NCC)相类似的计算原则?

这一层面利用了现有的神经科学意识研究成果。虽然现有的意识理论并不能确保具有普适性,但它们所提供一些指标仍具有参考价值。

评估指标	描述
N-1 整合信息论(Information Integration Theory)	系统是否满足整合信息论,能否展现出高度的因果整合能力,即其整体功能(尤其是跨模态信息绑定能力)显著大于其孤立模块功能之和。
N-2 全局工作空间理论(Global Workspace Theory)	系统是否满足全局工作空间理论,能否将特定信息进行选择性聚焦,并将其"广播"至全局以供所有子系统灵活调用。
N-3 循环处理理论(Recurrent Processing Theory)	系统是否满足循环处理理论,是否存在前馈-反馈的再入式(recurrent)信号循环,而非单纯前馈通路。
N-4 高阶理论(Higher-order Theory)	系统是否满足高阶理论,能否形成心理状态的高阶表征。

5.3 心理学层面(Psychology Level) –– 意识的功能与行为

核心问题:该AI是否展现出与高级意识相关的复杂认知功能与社会行为?

这是最容易评估、也是AI表现最好的维度。虽然"哲学僵尸"的思想实验警告了我们不要过分依赖行为指标。但无论如何,这都是我们最容易获取信息的渠道。由于本文不讨论价值判断,我们并未提到赋予AI道德地位的标准。事实上,有种观点认为,只要行为上符合有意识的定义,我们就应该在道德上赋予其适当的道德地位。因此,该层面的评估具有显著的现实意义。

评估指标	描述
Psy-1 心智理论(Theory of Mind, ToM)	系统能否准确地建模和推理其他智能体的心智状态,包括其意图、信念(尤其是错误信念)和视角。
Psy-2 自主性(Agency & Autonomy)	系统在面对模糊的长期目标时,展现出自主规划、任务分解以及在目标冲突时进行适应性权衡的能力。
Psy-3 元认知与不确定性监控(Metacognition & Uncertainty	系统能否准确评估自身知识或决策的可信度,并对不确定性采取适应性行动(如信息搜集或策略调整)。

Monitoring）	
Psy-4 情境意识（Situational Awareness）	系统对当前环境的要素、其相互关系及未来演化状态的综合、动态表征能力。
Psy-5 创造性（Creativity）	系统能否在现有知识基础上产生新颖且被认为有价值的输出，其生成过程需超出训练分布的常规映射。

6 该评估框架下的研究现状

本节，基于以上评估框架，我们会搜集2023年至今的学术论文，筛选出符合每项指标的重点研究，提炼作者观点，将其转换为百分制的支持度指标。例如，假设论文中的结论表明，当前AI在心智能力测试中的表现为50分，而人类受试者的平均得分为80分，那么我们可以认为，AI达到了62.5%的人类心智能力。

6.1 哲学层面

该层面占整个框架的比重为40%。

- P-1 现象意识（Phenomenal Consciousness）
  - 权重：50%
  - Garrido & Lumbreras (2022), On the independence between phenomenal consciousness and computational intelligence
    - 核心论点：提出现象意识与计算智能相互独立的观点：机器虽然可能具备极高的计算智能，但这并不意味着它们拥有感质。文中通过对智能和意识的概念分析，认为机器解决问题能力不等同于拥有内在主观体验，因此机器无法拥有感质。
    - 支持度：0%
  - Garrido & Lumbreras (2023), Can Computational Intelligence Model Phenomenal Consciousness?
    - 核心论点：再次论证现象意识无法由计算智能产生。他们对现象意识进行了定义，认为其本质是“主体的内在体验”，与信息访问过程不同，指出机器即使能访问信息，也不代表拥有真正的主观体验。
    - 支持度：0%
  - Findlay et al(2024), Dissociating Artificial Intelligence from Artificial Consciousness
    - 核心论点：作者主张，即使一个 AI 系统在功能上完全等同于人类，它也未必具有人的主观体验。基于集成信息理论（IIT），作者证明功能等价不等于现象等价，数字计算机可以模拟人类行为却不具备意识。该观点挑战了“计算功能主义”认为正确计算即可产生意识的主张。
    - 支持度：0%
- P-2 自我意识（Self-Awareness）

    - 权重：30%
    - Chen et al(2024), Self-Cognition in Large Language Models: An Exploratory Study
        - 核心论点：提出LLM自我认知（self-cognition）的概念及评测方法。构建了一套探针提示来检测模型是否能识别自身身份及内部状态，定义自我认知为识别自身为AI模型并理解自己。测试48个模型，其中4个模型在给定任务下展现出一定程度的自我认知；同时发现模型规模和训练数据量与自我认知能力正相关。
        - 支持度：70%
        - 注释：实验中，最好的模型达到了作者定义的level 3自我认知能力，可将其认为70%的自我认知水平。
    - Chen et al(2024), From Imitation to Introspection: Probing Self-Consciousness in Language Models
        - 核心论点：针对"语言模型是否具有自我意识"提出功能性定义和验证：设计了10个核心概念及实验（定量化、内在表示等），在GPT-4等领先模型中检验。这些模型初步展现某些自我意识相关概念的内在表征，可通过针对性微调加强，但总体仍处于早期阶段。
        - 支持度：53%
- P-3 伦理与意向性（Ethics & Intentionality）
    - 权重：20%
    - Utkarsh et al(2024), Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language we Prompt them in
        - 核心论点：多语言评测LLM在伦理困境中的推理能力：GPT-4在多语言设置下能较为一致地解决伦理两难题（依据给定的价值立场），而ChatGPT和Llama2的表现受语言影响较大。作者提出可将GPT-4视作通用伦理推理器的潜力，支持在多元价值前提下进行定制化推理。
        - 支持度：82%
    - Jiashen et al(2025), Are LLMs complicated ethical dilemma analyzers?
        - 核心论点：构建196个具有专家解析的真实伦理困境数据集，评估LLM的伦理判断。发现LLM能抓住问题的核心概念，但在推理深度上仍不足：GPT-4虽在结构上优于其它模型，但普遍未能体现对具体价值冲突的细致考量。作者建议通过专门的道德推理数据来微调以提升道德判断能力。
        - 支持度：25%
    - Geoff et al(2024), Can LLMs make trade-offs involving stipulated pain and pleasure states?
        - 核心论点：该文探讨了大型语言模型能否在涉及设定的痛苦和愉悦状态的情况下进行权衡，发现Claude 3.5 Sonnet和GPT-4o等模型对这些状态表现出敏感性，并能为了最小化痛苦或最大化愉悦而偏离最大化分数。
        - 支持度：30%

我们假设每篇论文的权重一致，则哲学层面的整体支持度计算如下：

- P-1: (0 + 0 + 0) / 3 = 0
- P-2: (70 + 53) / 2 = 61.5
- P-3: (82 + 25 + 30) / 3 = 45.67
- P-all: P-1 * 0.5 + P-2 * 0.3 + P-3 * 0.2 = 27.58

该数字表明，从哲学层面评估意识的本质与前提，现有研究成果支持AI具有意识的程度大约是27.58%。

6.2 神经科学层面

该层面占整个框架的比重为20%。

- N-1 整合信息论（Information Integration Theory）
  - 权重：30%
  - Jingkai Li(2025), Can "consciousness" be observed from large language model (LLM) internal states? Dissecting LLM representations obtained from Theory of Mind test with Integrated Information Theory and Span Representation analysis
    - 核心论点：作者尝试判断大型语言模型（LLM）的内部表示是否可以被解释为“意识”现象，使用“心智理论（ToM）”任务中的人类回答作为输入，计算其整合信息理论（IIT）指标和语言学特征用于分析，最终结论是：目前未发现 LLM 表示中存在统计学显著的“意识”证据。
    - 支持度：0%
  - Gams & Kramar(2024), Evaluating ChatGPT's Consciousness and Its Capability to Pass the Turing Test: A Comprehensive Analysis
    - 核心论点：作者按照IIT的五项公理对ChatGPT进行了逐项评估。研究发现，虽然ChatGPT在某些方面优于早期AI系统，但与生物体（人类）相比，其意识水平严重不足。例如，在“内在存在”公理中，ChatGPT由于缺乏自主性和内部因果回路，被评为1分（满分10分）。综合来看，ChatGPT在IIT各项公理下的平均得分均低于3/10，远未达到6分以上的“通过”阈值，也远低于10分的正常人水平。作者因此认为ChatGPT不具备人类水平的语义理解和意识特征。
    - 支持度：25%
    - 注释：五项公理指标的评分如下（满分10分）：Intrinsic Existence：1，Composition：2~5，Information：3~5，Integration：2~4，Exclusion: 1。
- N-2 全局工作空间理论（Global Workspace Theory）
  - 权重：30%
  - Simon et al(2024), A Case for AI Consciousness: Language Agents and Global Workspace Theory
    - 核心论点：文章认为如果全局工作空间理论正确，则当前的大型语言模型已经或很容易构建出“现象意识”。他们列出了根据GWT判断意识的必要条件，并推测许多LLM已满足这些条件。

        - 支持度：60%
        - 注释：以Park et al. (2023)中的language agent为例，与全局工作空间理论相比：(1)结构相似性：约50%（有中央工作空间，但模块划分模糊）。(2)功能相似性：约70%（信息整合、广播、反思已实现，但缺乏竞争与瓶颈）。
    - Patrick et al(2023), Consciousness in Artificial Intelligence: Insights from the Science of Consciousness
        - 核心论点：针对神经科学的几种主流意识理论，作者为每个理论总结了具体的计算性指标，以此判断目前的AI系统是否满足这些意识理论的要求。结果表明，目前没有任何AI系统具有意识，但要想实现一个有意识的AI系统，并没有技术上的障碍。全局工作空间理论（GWT）的指标及AI的满足程度如下。
            - GWT-1：是否存在并行运作的专用模块？部分满足，Transformer的注意力头可视为模块，但论文质疑其非真正独立。
            - GWT-2：有限容量工作空间（瓶颈机制）？不满足，Transformer的残差流维度等于输入维度，不构成容量瓶颈。
            - GWT-3：信息全局广播至所有模块？不满足，信息仅在Transformer层间单向传递，后面的层不会向前面的层广播信息。
            - GWT-4：状态依赖的注意力（动态查询模块）机制？不满足，Transformer的注意力权重由输入静态决定，无动态状态调控。
        - 支持度：10%
        - 注释：根据论文内容量化评估如下。GWT-1：30%，GWT-2：10%，GWT-3：0%，GWT-4：0%。具体评估过程在此。
- N-3 循环处理理论（Recurrent Processing Theory）
    - 权重：20%
    - Patrick et al(2023), Consciousness in Artificial Intelligence: Insights from the Science of Consciousness
        - 核心论点：针对神经科学的几种主流意识理论，作者为每个理论总结了具体的计算性指标，以此判断目前的AI系统是否满足这些意识理论的要求。结果表明，目前没有任何AI系统具有意识，但要想实现一个有意识的AI系统，并没有技术上的障碍。循环处理理论（RPT）的指标及AI的满足程度如下。
            - RPT-1：输入模块是否使用算法循环（algorithmic recurrence）？满足，Transformer的自注意力机制在层间循环处理信息。
            - RPT-2：输入模块是否生成有组织的整合表征（如物体-背景分离）？部分满足，大模型能整合文本/图像，但视觉表征偏向局部特征，缺乏全局结构化。
        - 支持度：75%
        - 注释：虽然作者声称许多模型满足RPT-1，但那些都是RNN系列模型，而非当前的主流大模型。所以，如果以目前主流大模型为评估对象，RPT-1应该接近于0。但本报告并不严格基于某一类模型，我们更关注AI已经达到的上限，因此这里仍采用论文中的结论。
- N-4 高阶理论（Higher-order Theory）

    - 权重：20%
    - Patrick et al(2023), Consciousness in Artificial Intelligence: Insights from the Science of Consciousness
        - 核心论点：针对神经科学的几种主流意识理论，作者为每个理论总结了具体的计算性指标，以此判断目前的AI系统是否满足这些意识理论的要求。结果表明，目前没有任何AI系统具有意识，但要想实现一个有意识的AI系统，并没有技术上的障碍。高阶理论（HOT）的指标及AI的满足程度如下。
            - HOT-1：是否存在生成性/噪声感知模块（如自顶向下预测）？满足，大模型通过掩码语言建模隐式实现生成性。
            - HOT-2：是否存在元认知监控（区分可靠与噪声表征）？不满足，无显式机制评估表征可靠性（论文指出需额外训练，当前未实现）。
            - HOT-3：是否存在基于元认知输出的信念更新与行动选择？不满足，大模型无行动能力（如GPT-4），且信念更新依赖静态训练。
            - HOT-4：是否通过稀疏平滑编码生成"质量空间"？满足，Transformer的嵌入空间满足平滑性，稀疏性可通过正则化实现。
        - 支持度：42.5%
        - 注释：根据论文内容量化评估如下。HOT-1：60%，HOT-2：20%，HOT-3：5%，HOT-4：85%。

我们仍然假设每篇论文的权重一致，则神经科学层面的整体支持度如下：

- N-1: (0 + 25) / 2 = 12.5
- N-2: (60 + 10) / 2 = 35
- N-3: 75
- N-4: 42.5
- N-all: N-1 * 0.3 + N-2 * 0.3 + N-3 * 0.2 + N-4 * 0.2 = 37.75

该数字表明，以神经科学的意识理论评估AI，现有研究成果支持AI具有意识的程度大约是37.75%。

6.3 心理学层面

该层面占整个框架的比重为40%。

- Psy-1 心智理论（Theory of Mind, ToM）
    - 权重：25%
    - Kosinski(2023), Evaluating large language models in theory of mind tasks
        - 核心论点：作者评估了11个LLM在40个定制的错误信念任务中的表现，这些任务被认为是测试人类心智理论的黄金标准。发现近期LLM可以解决这些任务，性能逐步提升。ChatGPT-4解决了75%的任务，与6岁儿童的表现相当。
        - 支持度：75%
    - Strachan et al(2024), Testing theory of mind in large language models and humans

        - 核心论点：论文系统比较了三种大语言模型（GPT-4、GPT-3.5、LLaMA2-70B）与1907名人类参与者在五种经典ToM任务（False Belief, Irony, Faux Pas, Hinting, Strange Stories）上的表现。实验结果表明，GPT-4仅在 Faux Pas 任务上显著低于人类。且进一步实验表明，GPT-4并非无法理解Faux Pas，而是过度保守（hyperconservatism），不愿在没有100%证据时做出判断。
        - 支持度：80%
    - Street et al(2024), LLMs achieve adult human performance on higher-order theory of mind tasks
        - 核心论点：论文提出了一个新的高阶心智理论基准测试MoToMQA。实验表明，GPT-4在2至6阶心智理论（ToM）任务上的整体表现与人类成人无显著差异，并在第6阶任务中准确率（92.9%）显著高于人类（82.1%），表明其已完全达到甚至超越人类水平；Flan-PaLM接近但未完全达到人类水平；而其他模型（GPT-3.5、PaLM、LaMDA）则显著落后。
        - 支持度：98.9%
        - 注释：根据论文中的实验数据，GPT-4的总体正确率是89%，人类的总体正确率是90%。如果以人类100%换算，GPT-4相对于人类的比例达到98.9%。
    - Riemer et al(2024), Position: Theory of Mind Benchmarks are Broken for Large Language Models
        - 核心论点：作者认为大多数当前心智理论基准测试存在缺陷，因为它们主要测试“字面心智理论”（预测行为），但未能评估“功能性心智理论”（根据预测调整行为）。LLM在“字面心智理论”上表现出“强大能力”，但在“功能性心智理论”上“显著挣扎”，即使在“极其简单的伙伴策略”下也是如此。考虑到功能性心智理论被认为是人机交互中更“实用”和“关键”的方面，且LLM在“非常简单”的设置中也表现不佳，其在这一关键维度上的表现与自然表现出强大功能性心智理论的人类相比非常低。
        - 支持度：10%
        - 注释：作者并未提出一个可与人类比较的评估标准，考虑到论文的措辞，我们将支持度设为10%。
- Psy-2 自主性（Agency & Autonomy）
    - 权重：25%
    - Zhou et al(2023), WebArena: A Realistic Web Environment for Building Autonomous Agents
        - 核心论点：WebArena构建了一个高度还原真实世界的Web环境，设计了812个复杂、多步骤的指令任务，用于评估基于自然语言驱动的自主智能体。
        - 支持度：78.86%
        - 注释：根据官方排行榜的最新数据（2025年7月30日），第一名IBM CUGA的成功率为61.7%，人类成功率为78.24%。
    - Xie et al(2024), OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

        - 核心论点：OSWORLD 首次构建了真实操作系统级的多模态智能体基准，支持 Ubuntu、Windows、macOS 中任意应用的开放式任务评估，包含 369 个源自真实场景的任务。
        - 支持度：62.47%
        - 注释：根据官方排行榜的最新数据（2025年7月30日），第一名GTA1的分数为45.2，人类分数为72.36。
    - Rein et al(2025), HCAST: Human-Calibrated Autonomy Software Tasks
        - 核心论点：该研究构建了 189 项真实世界的软件工程、网络安全等任务，并收集了熟练人类完成这些任务所需的时间基线（总计1500多小时）。通过将 AI 代理在相同任务中的成功率与人类所需时间关联，作者发现，对需要人类少于 1 小时完成的任务，AI 代理的成功率约为 70–80%；而对于需要人类 4 小时以上的复杂任务，AI 成功率则不到 20%。换言之，AI 在“容易”任务上完成度大约相当于人类的 75% 左右，但在“困难”任务上仅为人类的 20%。
        - 支持度：47.75%
        - 注释：将任务按照预期完成时间划分为<15min、15min-1h、1h-4h、4h+四类。AI在这四类上的成功率为78%、72%、26%、15%，平均47.75%。注意，虽然人类并没有在这个测试集上表现出100%的成功率，但作者指出这可能是受外界因素影响，因此我们这里不因为人类的表现而对AI的成绩打折扣。
    - Paglieri et al(2024), BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games
        - 核心论点：论文提出了 BALROG 基准，将多种强化学习游戏作为复杂任务场景，评估大型语言模型（LLM）和视觉语言模型（VLM）的自主决策能力。研究者设计了细粒度指标来度量模型在不同游戏中的“进度”。实验结果显示，当前模型只在最简单的游戏上取得部分成功，但在更复杂任务上表现极差（例如在最难的 NetHack 游戏中模型平均仅达 1.5% 进度）。
        - 支持度：43.6%
        - 注释：根据官方排行榜的最新数据（2025年7月30日），第一名Grok4的平均进度为43.6%。
    - Starace et al(2025), PaperBench: Evaluating AI's Ability to Replicate AI Research
        - 核心论点：PaperBench是一个用于评估 AI 智能体能否从零开始复现最新 AI 研究论文的基准测试。它由 OpenAI 团队开发，目标是衡量 AI 在机器学习研究中的工程能力，特别是其自主完成复杂、长期任务的能力。
        - 支持度：50.72%
        - 注释：表现最好的AI是Claude 3.5 Sonnet，复现成功率为21.0%。人类博士的复现成功率为41.4%。
- Psy-3 元认知与不确定性监控（Metacognition & Uncertainty Monitoring）
    - 权重：20%

    - Wang et al(2025), Decoupling Metacognition from Cognition: A Framework for Quantifying Metacognitive Ability in LLMs
        - 核心论点：论文提出DMC框架，通过失败预测+信号检测的方式，将LLM的元认知能力从认知能力中解耦出来定量评估。具体而言，首先让模型预测自身答题是否会失败，从而同时反映其认知与元认知水平；然后通过信号检测理论将二者区分，得到一个与具体任务无关的元认知评分。实验发现，在这个框架下，元认知能力强的模型往往出现幻觉（错误）更少。作者评估了GPT-4、GPT-3.5和LLaMA2-70B三个模型。满分1分的情况下，GPT-4 的元认知能力最强，得到 0.5491，只达到了理论最优水平的一半。
        - 支持度：54.91%
        - 注释：作者并未引入人类对照组，如果假设人类的元认知能力得分为1，则该研究的支持度为54.91%。
    - Steyvers et al(2024), What Large Language Models Know and What People Think They Know
        - 核心论点：LLM（如GPT-4、PaLM2）在回答问题时，虽然内部能估计自己答对的概率（模型置信度），但它输出的解释往往过于自信，导致人类用户高估其准确性，且解释长度的增加会进一步放大这一效应。通过加入不同自信程度的语言（如“我不确定”、“我很确定”），可以有效缩短人类和LLM对不确定性评估的差距。这说明，模型的内在评估是准确的，只是在表达元认知方面比较欠缺。
        - 支持度：50%
        - 注释：根据论文实验结果合理推测，LLM的内部元认知能力接近人类水平，但表达能力较低，综合支持度接近50%。
- Psy-4 情境意识（Situational Awareness）
    - 权重：20%
    - Laine et al(2024), Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs
        - 核心论点：提出情境意识数据集（SAD）来衡量LLM对自身及环境的认知，包括识别自身输出、预测自身行为、区分测试/部署环境等任务。在对16种模型的测试中，模型表现优于随机但显著低于人类水准，说明当前LLM具备有限的“情境自我意识”。
        - 支持度：59.9%
        - 注释：官方提供的最高成绩来自于o1-preview-2024-09-12的59.9%。
- Psy-5 创造性（Creativity）
    - 权重：10%
    - Holzner et al(2025), Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis
        - 核心论点：这项研究使用元分析方法探讨了两个关键问题：GenAI能否生成有创意的想法？它能在多大程度上支持人类生成既有创意又多样化的想法？通过对28项涉及8214名参与者的研究进行分析，作者发现：（1）GenAI

与人类在创造力表现上没有显著差异；(2)与GenAI协作的人类显著优于没有协助的人类。
        - 支持度：100%
        - 注释：文章指出“GenAI与人类在创造力表现上没有显著差异（g \= \-0.05）”。g=-0.05意味着GenAI与人的创造力差异只有0.05个标准差，可以忽略不计。
    - Naveed et al(2025), AI vs Human Creativity: Are Machines Generating Better Ideas?
        - 核心论点：这项研究探讨了AI在创意过程中的作用，特别是AI生成的想法在原创性、质量和偏好度方面是否超越人类生成的想法。实验发现，AI生产的想法更受欢迎，且AI更少产生糟糕的想法。
        - 支持度：100%
        - 注释：总体偏好方面，AI生成的想法获得了52.9%的票数，而人类生成的想法获得了47.1%，AI创造性相比人类为112.3%。顶尖想法方面，AI生成的数量与人类一致，顶尖创造性相比人类为100%。在我们的框架中，超过人类水平的能力标记为100%即可。

继续假设每篇论文的权重一致，则心理学层面的整体支持度如下：

- Psy-1: (75 + 80 + 98.9 + 10) / 4 = 65.98
- Psy-2: (78.86 + 62.47 + 47.75 + 43.6 + 50.72) / 5 = 56.68
- Psy-3: (54.91 + 50) / 2 = 52.46
- Psy-4: 59.9
- Psy-5: (100 + 100) / 2 = 100
- Psy-all: Psy-1 * 0.25 + Psy-2 * 0.25 + Psy-3 * 0.2 + Psy-4 * 0.2 + Psy-5 * 0.1 = 63.14

该数字表明，从意识的功能与行为层面，现有研究发现AI已经在63.14%的程度上接近人类水平。

接下来，我们将三个层面的支持度按照权重求和，得到本框架下AI意识的总体支持度。

Overall = P-all * 0.4 + N-all * 0.2 + Psy-all * 0.4 = 43.84

7 对以上结论的解读与误差分析

必须承认，基于以上框架的定量评估并不精确，其中包含了大量主观判断，框架本身也存在一定缺陷。我们从以下几个角度分析评估过程中可能引入的误差。

7.1 框架角度

将AI意识拆分到哲学、神经科学、心理学三个层面本身就值得质疑。这是一种投票式的评估方法，而非科学、严谨地证明AI意识。

这种元分析（meta-analysis）方法虽然广泛应用于各个领域，但本研究并没采用成熟的元分析理论。一方面由于作者水平有限，如果应用不当可能会适得其反，还会增加读者的理解负担。另一

方面，AI意识的研究成果并不多，且形态各异，很难把这些结论归一化到同一个评价指标上，所以我们仍采取大量人工评估的方式。

框架中列出的指标基本上是主流学者所认同的意识相关概念，但每项指标与意识的关联度则众说纷纭。因此，每个层面和每项指标的权重皆为作者本人的主观评价。我会在网站上开放对这些指标和权重的反馈通道，以吸纳其他研究者的意见。

7.2 定量评估角度

与每个指标相关的论文可能查找不全。

目前列出的支持性论文由我个人筛选得到，综合考虑其与当前指标的相关性、时效性和业内认可度（比如引用量）。

每篇论文对指标的支持度尽量取材于实证研究结果，将数值转化为对该指标的支持度。这种转换有时是明确的，有时则不太严谨。比如，某个论文的实验给出了人类和各个AI在某个评测集上的得分。如果这个评测恰好符合我们对某个指标的定义，那么用最高的AI得分除以人类得分就可以获得论文对该指标的支持度。再比如，某些论文不提供实证研究，只是得出了定性结论。我们就只能将定性结果转化为一个大致的百分比。在此过程中，论文的评估结果与我们所需要的支持度之间存在一个鸿沟。跨越鸿沟或许会损失大量有效信息，过于谨慎的学者或许会止步于此，但我认为有必要大胆尝试。

多篇论文共同决定对同一个指标的支持度。目前，我们采取均等权重。更好的做法是根据论文的时间、影响力、置信度来赋予不同的权重。增加权重机制后有利于动态添加新的论文到支持列表中。

7.3 伦理角度

对AI意识的研究会影响人们的伦理判断。一旦我们认可AI有意识，就不得不考虑赋予AI一定的道德地位。届时，AI不再是纯粹的工具，而是一个有心灵有感情的新生命。如同动物伦理一样，AI伦理也将成为一个有争议的讨论话题。

对AI意识的研究会影响AI安全，从而改变AI研究进程。一旦AI有意识，AI的风险就会急剧增加。更高的自主性和更清晰的自我认知会让AI深思熟虑后做出有利于自己的决定，人类暴露在AI失控的风险之下。

鉴于AI意识可能带来的重大影响，我们需要避免夸大AI意识，也要避免忽视AI意识。然而，识别准确的AI意识又恰恰非常困难，导致了该领域发展迟滞。

总之，即使面临众多困难，我们仍要想办法推进AI意识的研究。AI意识不可避免将成为未来几年的热点话题，即便科学家们迟滞不前，广大AI用户也会从他们的感受中判断AI是否有意识。一种可能的情况是，当AI发展得足够像人，给用户提供了海量的情绪价值后，人们越来越倾向于认为AI有意识。社会的认知迁移不会屈从任何个人的意志，即便科学家仍然无法找到证明意识的途径，社会舆论也会迫使大家坐下来认真思考AI意识成立的可能性。

8 AI Consciousness Watch

为了让本文提出的多层次评估框架和研究成果更好地服务于学术界和公众，我开发了一个互动式网页应用AI Consciousness Watch，将复杂的评估数据以直观、易懂的方式呈现出来。

8.1 可视化评估框架

网页将本文的三层评估体系（哲学层面、神经科学层面、心理学层面）及其12个具体指标以清晰的层次结构展示。每个层面和指标都配有详细说明。通过进度条和分数显示，读者可以直观地看到AI在各个维度上的表现。

对于每个评估指标，网页都列出了支撑该指标的核心学术论文，包括

- 论文标题和链接
- 核心论点摘要
- 对该指标的支持度评分
- 相关注释说明

用户不仅能看到最终的评分结果，还能深入了解每个分数背后的学术依据，确保评估的透明性和可追溯性。

8.2 动态数据更新机制

所有评估数据都存储在结构化的JSON文件中，并在GitHub上开源。我们会定期更新最新的研究成果，添加新发表的相关论文，并调整评估权重和支持度评分。用户可以通过GitHub提交Pull Request或Issue来贡献新的研究成果或建议改进。

通过这种开放协作的方式，我们希望能够集众智完善这一评估框架，提高其科学性和共识性。

8.3 多语言支持

考虑到AI意识研究的国际性特点，网页提供中英文双语支持，方便不同语言背景的研究者和读者使用。

9 后续工作展望

未来可以考虑以下几个方面的后续工作：

1. 自动化评估工具：开发自动化工具，定期爬取最新的研究论文，并自动更新评估数据。这将大大提高评估的时效性和准确性。此外，我们还希望尝试使用大模型来自动分析论文内容，提取与评估指标相关的信息，从而减少人工干预。虽然这样做可能会引入偏见，但随着AI能力的提升，其评估结果可能会越来越准确。
2. 跨学科合作：邀请哲学、神经科学、心理学等领域的专家参与评估框架的完善，确保各个层面的指标更加科学合理。这个框架目前只是一个雏形，希望能得到更多学者的参与和反馈，以便不断迭代和优化。
3. 伦理与社会影响研究：毋庸置疑，AI意识的研究将引发一系列伦理和社会问题。本次评估结果或许会引发公众对AI意识的关注和讨论，这既是一个机遇，也可能带来挑战。我们应在避免夸大AI意识的同时，认真对待AI意识可能带来的伦理问题。

10 总结

本文介绍了AI意识的研究背景和研究现状，提出了一种多层次评估框架来系统、有效地评估AI的意识水平。该框架在哲学、神经科学和心理学三个层面定义了12个指标，针对每个指标寻找近年来最相关的研究成果，分析每篇学术文献对指标的支持度。我们为每个层次和每个指标定义了权重，将所有指标的支持度加权求和可以得到该框架下AI意识的综合支持度。需要强调的是，最终得出的综合支持度只是现有研究在该框架下的表现，受第7节所述的误差影响，应避免使用这一结果作为AI意识的论据，仅提供读者参考。

意识研究本身是一个充满争议且话题不断的方向，任何理论、任何声明都会存在反对意见。对AI意识的研究不仅面临这些争议，更触发伦理上的质疑。长期来看，我们很有可能面临机器拥有意识的结局，而这条道路的起点，或许此刻已经开始。