人工意识

——AI 有意识吗？*

林作铨 †

linzuoquan@pku.edu.cn

大家大都使用过手机聊天应用，那是一类基于大语言模型的聊天机器人。我们跟机器人聊天，意识到机器人具有智能，那么问题来了：聊天机器人有意识吗?

有人做过一些抽样调查，大众普遍认为机器人有意识；专家的意见相反，普遍认为机器人没有意识。

人工智能（AI）的意识，亦称**人工意识**（Artificial Consciousness, AC）或机器意识，是一个饶有趣味的问题。AI 参照人类的智能，AC 应该参照人类的意识。人的智能活动包含意识能力，笼统地说，若说 AI 已具有某种与人类可比的机器智能，应该也具有某种与人类可比的机器意识。如此一想，带来三问：

- 当今 AI 有意识吗?
- 不久将来 AI 可能有意识吗?
- AI 有一种不同于人的机器意识吗?

我们将从哲学、经由神经科学、到 AI 探讨一下 AC。

---

*科普，2025.9.10 初稿于北大智华楼。

†北京大学博雅特聘教授。

1 意识哲学

什么是意识？意识是一个尚未解开的千古哲学谜题。

意识通常被定义为个人对自身存在、心理活动（如思想、情绪、感知）及外部环境的觉知和主观体验。觉知是对自我（如“我在思考”）和外部世界（如“我看到红色”）的即时感知，例如，你此刻意识到正在阅读这些文字；**主观体验**，亦称**感质**（qualia），是个体独有的感受（如“柠檬的酸味”），这种感受无法通过物理描述来完全传达，例如，两个人看同一片天空，各自对“蓝色”的感受可有微妙的不同，两个人有各自的感质。

马赫（Ernst Mach）早在 19 世纪末提出了“思想实验”一词，这是一个通过假设性情境来探索哲学问题的工具，它不依赖实际实验或数据，而是借助思想（想象力和逻辑推理）来检验直觉、概念或理论。对难以言说的“感质”，杰克逊（Frank Jackson）提出了一个思想实验“玛丽房间”。

玛丽房间

玛丽在黑白相间的房间里长大，知道所有关于颜色的物理知识，但从未见过红色。当她第一次走出房间看到红色是什么的感觉时，她学到了一个新的事实。

这个思想实验想表明：意识包含物理无法解释的东西（感质）。

这种不可物理描述的感质披着神秘的面纱，不能言传。奈格尔（Thomas Nagel）在 1974 年提出类似的思想实验“蝙蝠体验”。

蝙蝠体验

虽然我们可以通过科学研究完全掌握蝙蝠回声定位的神经机制和行为模式（如超声波发射、接收及神经处理过程），但我们仍然无法真正想象或理解“作为一只蝙蝠”的主观体验是什么样子。

亦即我们无法获知蝙蝠运用声呐感知世界的第一人称感受，任何有意识的状态都存在“某种像是什么的样子”的主观体验，而这种第一人称的、现象

性的感质无法被第三人称的客观物理描述所完全捕获或还原。奈格尔想表明通过物理还原对意识进行解释存在根本性局限。

意识是一个直观概念，没有一个精确的定义。我们可以列举许多意识的性质，并作为意识的必要条件，但是，我们不能通过列出意识的各种性质作为意识的充分条件。

与之相反，**无意识**是个人无法直接觉察到的心理活动，诸如无觉知（如昏迷、全身麻醉）、无记忆（如深度睡眠时）和无自主控制。弗洛伊德（Sigmund Freud）提出了无意识概念并发展了潜意识理论，认为意识仅是心理活动的冰山一角。那是说的轻巧，我们先撇开弗洛伊德，单说意识这个难以企及的冰山一角。

笛卡尔（René Descartes）于 1637 年提出了著名语录“我思故我在”，“我正在思考”本身证明了意识的存在。笛卡尔通过怀疑一切（包括外部世界是否存在），明确区分“心灵”（意识）与“物质”（无意识），提出了心物“二元论”：意识是非物质的，而大脑是物质的。关于意识问题，古代的许多哲学家都有所表述，笛卡尔将“思维”（包括怀疑、理解、肯定等）视为意识的核心，确立了意识作为哲学的基本原理之一，关于意识的哲学讨论可以从他算起。

意识哲学可以划分为多个流派，这好比武侠江湖门派众多，围绕打架斗殴一件事。若从最根本的对立角度出发，意识哲学可以简化为两种基本立场：物理主义和反物理主义（如二元论）。物理主义主张意识完全由物理过程（如大脑神经元活动）构成，不存在非物质的“心灵实体”。奉行物理主义与否是意识哲学的核心分歧，其他意识哲学理论大多是对二者的修正或折中。

柯克（Robert Kirk）提出了一个思想实验“哲学僵尸”，用于质疑物理主义，认为意识不可完全还原为物理过程。

哲学僵尸

一个在物理结构和行为上与正常人类完全相同的假想存在，但它完全不具备任何主观体验。

反驳者认为若物理结构完全相同，必然产生意识，“哲学僵尸”是一个概念悖论：它既是“僵尸”，却又不是“僵尸”。

即使按照物理主义，关于意识的物质基础的争论主要围绕着一个核心问题：意识是否严格依赖于特定的生物（碳基），还是可以存在于其他物理系统（如硅基计算机）之中。生物自然主义认为意识是碳基生物（特别是具有特定复杂神经系统的动物）独有的、不可复制的生物学特征。功能主义认为意识取决于系统的功能组织或计算状态，而非其构成的具体物理材料。

查默斯（David Chalmers）提出了思想实验“逐渐消失与跳舞的感质”，用于质疑功能主义。

逐渐消失与跳舞的感质

逐渐消失的感质：个体神经元被功能相同的硅基神经元逐一替换。

跳舞的感质：创建某个大脑区域完整的硅基计算复刻版本，并设想在该区域的生物版本与硅基版本之间进行反复切换链接。

在这个思想实验中，根据功能主义，你的意识体验应该保持不变，因为功能结构未变，例如，如果在替换过程中，你的红色体验逐渐变得灰暗、平淡，最终完全消失（尽管你仍然能行为正常地指认“红色”物体），会发生什么？如果感质可以这样“逐渐消失”，那就说明感质本身并不是功能结构的必然产物，功能相同，意识却可以消退。当版本切换时，你的红色体验会在“鲜红”和“灰暗”之间疯狂地跳跃（跳舞）吗？如果会，那说明同一功能可以实现不同的感质，功能不能决定意识；如果不会（即体验不变），那说明即使物理基础改变了，感质也能保持不变，这又违背了物理世界的基本定律。这样，如果意识依赖于特定物质，这些情况将导致一个看似荒谬的结论：实验对象将经历体验上的剧烈变化却无法察觉这些变化，这反而表明意识并不依赖于特定物质。这个思想实验用于反驳功能主义——即使一个系统的功能组织完全不变，其内在的主观体验也可能会消失或改变。

无论奉行那种主义，意识的本质还没搞清楚。莱文（Joseph Levine）指出了意识的**解释鸿沟**：即使我们完全掌握大脑的物理运作（如神经元如何

放电），仍无法解释主观体验如何从物质中涌现。大多数自然现象的特性可以通过识别现象中蕴含相关特性的要素来解释，例如，水分子（H2O）的特性决定了水应当具有表面张力，然而，大脑的特性似乎并不能完全解释意识的所有属性。

查默斯提出了区分两类意识问题：**意识的难问题**是解释为什么物理过程会产生主观体验。例如，为什么看到红色时会有“红色的感受”（感质）；**意识的易问题**是解释意识的功能机制，例如，报告精神状态，集中注意力，整合信息，控制行为，区分清醒和睡眠等等，原则上这些易问题可通过神经科学逐步解决，尽管至今已解决的易问题并不多。类似地，布洛克（Ned Block）提出了**现象意识**（phenomenal consciousness）和**存取意识**（accessible consciousness）之分，粗略地说，现象意识属于意识的难问题，存取意识属于意识的易问题。这种二分法算是近来意识哲学的进展，可进展甚微。

有一种泛心论认为意识并非仅仅存在于人类、动物或某些高度复杂的大脑中，而是以某种基本形式存在于宇宙万物之中。就像质量、电荷或自旋是物质的基本物理属性一样，感质也被认为是物质的基本属性，一块石头、一滴水、一棵树也拥有与其物理结构相对应的、极其简单的意识形式。人类和动物的复杂意识，则是由这些亿万个基本意识单元以某种方式“组合”而成的复杂整体。照泛心论，上述意识问题都不成其为问题了，倒是简单，但有一个组合问题：无数个简单、微小的意识（如电子的意识），如何组合成一个单一、统一、复杂的意识（如人类的视觉体验）？就像无数个单一的音符如何组合成一曲交响乐？我们听到的是和谐的整体，而不是无数音符的杂音。意识似乎是一个不可分割的整体，而泛心论需要解释这个整体是如何从部分整合而来的。某种意义上，这不是解决问题的方式，或是不可知论。

没有争论的问题就不属于哲学。

哲学通过理性辩论探索尚未（或无法）被科学解决的问题，例如，意识的主观体验为何存在？

2 意识科学

1992 年《探索》杂志把“什么是意识”列为十大未解科学问题之一，其奥秘至今未解开。

意识科学从哲学思辨走向实证研究，成为一个跨学科的研究领域，涉及神经科学、心理学、认知科学、临床医学，计算机科学等，旨在通过实证方法揭示意识的机制。狭义地说，意识科学主要是神经科学部分，通过研究大脑了解意识。毕竟，意识由大脑产生，这已经成为科学界的共识，毋容置疑。

克里克（Francis Crick）发现了 DNA（脱氧核糖核酸）的双螺旋结构，从而揭开了生命的物理本质，他因此获得诺贝尔奖，并继续探索意识的物理本质。克里克和科赫（Christof Koch）发现了**意识的神经相关物**（Neural Correlates of Consciousness, NCCs）。

意识的神经相关物

大脑中最小范围的神经活动，其激活与特定意识体验直接相关。

例如，视觉意识与大脑皮层活动相关，自我意识与大脑的前额叶和顶叶网络相关，更具体地，皮层感觉区域的同步 35-75 赫兹神经振荡是意识的核心。克里克由此提出了“惊人的假说”：意识是大脑神经元活动的产物。这个假说已经成为科学界的共识，

由此往后，各种意识科学理论大放异彩。据不完全统计，现有意识科学理论有二十几种，其中比较重要的有十种左右，例如，全局神经工作空间理论（Global Neuronal Workspace Theory, GNWT）和整合信息理论（Integrated Information Theory, IIT）：

- GNWT：将大脑视为一个信息处理系统，认为意识产生于信息进入“全局工作空间”的过程。该理论预测大脑的前额叶与感觉皮层之间的同步活动是意识体验的 NCCs。
- IIT：通过分析意识体验的内在特征，推导出产生这些体验所需的物理

基础，认为任何具有足够高“整合信息量”的系统都会产生意识。这样，意识并非大脑特有，理论上足够复杂的电子系统也可能产生意识体验。该理论预测意识体验的 NCCs 位于大脑的后部皮层（如后顶叶和枕叶区域），而非前额叶。

2025 年发表在《自然》上的 COGITATE（国际意识研究合作项目组）的研究结果对这两个理论进行了验证。这项开创性的“对抗性合作”研究由全球六个独立实验室共同完成，采用功能磁共振、脑磁图和颅内脑电三种脑成像技术，对 GNWT 和 IIT 的预测进行了严格检验。出乎意料的是，实验结果未能完全支持任何一个理论——IIT 预测的初级与中级视觉皮层间的持续同步活动未被观测到，而 GNWT 预期的前额叶意识表示、前额叶和感觉皮层同步活动等也缺乏证据。

2025 年《自然-神经科学》刊登两篇冲突文章，一篇文章纠集 100 多位学者指控 IIT 是伪科学，指出该理论的核心主张原则上无法检验，同时会引发高度反直觉且“非经验性”的推论，例如某种形式的泛心论，即意识是基本且无处不在的；另一篇文章又 IIT 提出者及其拥趸对此指控进行批判性反驳，并声称不应由某个“自封的法庭”来决定意识这样艰难的问题应该如何被研究。看来争论依旧可能演变成敌意。

2024 年欧盟一项于 2013 年开启的 10 年神经科学大型研究计划以失败告终，表明现有的神经系统的实验研究不能揭示大脑的原理。

至今没有一个在意识科学界取得共识的理论，也没有统一的理论能整合现有的意识科学。但是，这不妨碍意识科学开展临床医学应用，例如，脑机接口（BCI）技术已可以帮助闭锁综合症患者通过意识控制外部设备，可以解码人在脑中无声独白的大脑活动（内心言语）。

意识科学是唯一一个研究者的理解由其研究对象本身生成的学科，人对自身的意识进行有意识的研究，这使得意识科学比起其它科学更加难以取得突破。麦金（Colin McGinn）认为人类大脑在原则上永远无法理解人类意识，因为我们的思维受限于自身的感知能力，因此在认知上封闭于理解人类意识所需的概念。就像一只猴子试图理解自然数一样：自然数是一种无穷的理性存在，并且是世界的一种真实属性，但无论猴子如何努力地

学习，它很难数过十个数。这是不可知论者。

意识的概念看似简单，为什么却争议巨大？在哲学、心理学和神经科学等领域，意识的定义和解释有所不同，导致几个难题。首先，意识是“第一人称体验”，科学依赖“第三人称观察”，两者存在方法论冲突，这是主观性难题。其次，无法仅用神经元活动（物理）完全解释“为什么会有主观体验”，这是还原难题。还有跨学科差异：哲学关注本质，心理学探讨功能，神经科学寻找机制，各自侧重点不同。

意识科学还不够科学，大脑如何产生意识？即大脑活动如何产生个体的主观体验，这是一个意识科学的谜题，还需要进行哲学思辨。

实在不行，赌一把？1998 年在德国一家酒吧里，神经科学家科赫与意识哲学家查默斯花了一天时间争论大脑神经元产生意识的机制，科赫认为在接下来的 25 年里，有人会在大脑中发现一种特定的意识特征，而查默斯认为根本不可能，几杯酒下肚，科赫提议以一箱美酒作为赌注赌一把：大脑神经元产生意识的机制将于 2023 年前被发现。2023 年在意识科学研究协会年会上公开宣布查默斯赌赢，大脑神经元产生意识的机制仍未探明。科赫不愿认输，但他还是买了一箱酒来兑现自己的承诺，他表示从现在起 25 年后探明意识的机制还是有可能的，但是，考虑到他的年龄，他说他可能看不到下一个 25 年后的结果了。

意识科学中接近共识的部分可能是意识需多层级解释，如神经机制、信息整合、演化功能等，但终极答案可能需要来一场科学范式的革命。

AI 能担负这个范式转变的使命吗？这样想，其实是 AI 哲学问题，哲学家还有活可干，只是要学习意识科学和 AI 。

3 智能和意识

AI 的本质问题还没搞清楚，就是 AI 用机器（计算机）能模拟人类的智能吗？这是 AI 的初心和终极目标。图灵（Alan Turing）于 1950 年提出了一种衡量机器是否具备“智能”的测试。

图灵测试

一个人类评判员通过文本（如聊天界面）与一台机器和一个真人进行对话。如果评判员无法可靠区分机器和人类，那么这台机器就通过了测试，被认为具有智能。

图灵测试采用行为主义的做法，因为“智能”是黑箱式的直观概念，无法严格地定义和测量，图灵只好用可观察的行为来衡量。

AI 研究伊始，探讨智能问题从具体的体现智能活动的问题或领域展开，研究者建造特定领域的专用系统来模拟智能行为，希望从这些领域相关的智能系统中寻找智能的原理，逐步发展出与领域无关的通用智能系统。这符合心理学家斯皮尔曼（Charles Spearman）于 1904 年提出的“通用智能”概念，他发现：一个人在某种认知任务上表现出色，往往在其他认知任务上也有较好的表现。

当今流行的说法是**通用人工智能**（Artificial General Intelligence, AGI），指的一种具备通用智能和自主能力的 AI 系统，在所有领域达到人类智能水平。甚至说 AGI 是超级智能（具备超过人类智能的能力），可能统治人类，媒体上各种说法甚嚣尘上，莫衷一是。[1]

这里不对 AGI 进行讨论，只指出 AGI 没有一个明确的定义，其目标也没有明确的衡量标准。我们可以用智能汽车来了解，国际汽车工程师学会（SAE）制定，将自动驾驶技术分为 L0-L5 六个级别（但通常称为“五级”，因为 L0 代表无自动化），中国国家标准（GB/T 40429-2021 汽车驾驶

[1]举凡说 AGI 能实现的都是大人物，例如，因深度学习兴旺，唯一个被带上图灵奖和诺贝尔奖双桂冠的辛顿（Geoffrey Hinton）当下是无上荣光，他的凡尔赛式说法：早知如此（AI 将危及人类），何必当初研究 AI；因 AlphaFold 获得诺贝尔奖的 AI 机构 DeepMind 创始人哈萨比斯（Demis Hassabis）；还有马斯克等一众名人，不胜枚举。这些人大致可划分两类：一类人是个人信仰，他们相信 AGI 能实现，而且不久将来（如十年内）就能实现；另一类人是为了商业利益，如 OpenAI 等几乎所有这类 AI 公司就靠这个概念融资。与之对比，一辈子研究 AI 的学者大多默默无闻，保持沉默，可以推测他们基本上持反对或反感的态度。媒体上经常看到的 AGI 怀疑论者马库斯（Gary Marcus）其实不是 AI 专业出身，算转行蹭热点。由于一众大人物如是，这使得政府和大众很可能被带偏了。

自动化分级）也采用了类似的分级方式。在销的智能汽车属于 L2 级（部分自动驾驶），汽车工业处于 L2 向 L3（有条件自动驾驶）发展中，即在特定场景（如高速公路）下自动驾驶，L5 是完全自动驾驶（不需方向盘和人类干预）。马斯克几年前就宣传特斯拉已达到完全自动驾驶能力，属于吹牛不上税。问题是：再过十年智能汽车能达到 L5 吗？应该不能；设若不能，很多大人物预测 AGI 十年内能实现就全翻车了。关于 AGI，OpenAI 提出了一个类似的五级标准，也认为目前处于从 L2（能推理的聊天机器人）向 L3（智能体）发展中，姑且听之。智能汽车虽是一个商业问题，但涉及 AGI，还有所谓超级智能，这种未知问题归入 AI 哲学讨论。

要指出的是，上述标准中都没有提到意识一词。这是为什么呢？无它，不敢提而已。假如 AGI 能实现，那肯定要有意识，超级智能更不用说具有强大的意识能力。为什么不敢提？这需要多点说道。

智能和意识有关联。一些智能行为不一定需要意识参与，例如，婴儿在出生后不久就展现出基本的学习和记忆能力（智能），但自我意识的明确证据（如照镜子自我识别）要到 18-24 个月才出现。某些高级认知功能（如模式识别、语法处理）可以在无意识条件下进行，神经解剖学实证有智能和意识分离现象，发现智能与意识部分独立的神经基质。另一方面，意识可能是高级智能的必要条件或自然产物，随着智能任务复杂度增加，意识的作用变得更为关键。一句话，人类的智能那能没有意识呢。

倘若 AGI 能够实现，达到或超越人类智能水平，我们应将 AI 视为有意识或无意识的？若认定 AI 不具备意识，这是否使得意识科学沦为伪科学？

智能需要意识，这个问题 AI 不能回避。

其实，AGI 概念在 AI 哲学早有更严肃的讨论。希尔勒（John Searle）于 1980 年提出了弱 AI 和强 AI 的概念。**弱** AI 是指计算机仅是模拟智能行为的工具，并不真正“理解”其行为，相当于前面说的专用智能系统，AI 早已在某些领域达到或超过人类智能的水平，如 AlphaGo 与人类围棋冠军比赛获胜，且人类再也不能战胜它。**强** AI 是指运行适当程序的计算机确实拥有理解力和真正的意识，相当于前面说的通用智能系统或 AGI。

为论证强 AI 的不可行性，希尔勒提出了“中文屋”思想实验。

中文屋

假设一个母语英文但不懂中文的人被关在房间里，他有一本英文规则书，规定如何根据收到的中文符号（输入）匹配并返回其他中文符号（输出）。房间外的人向房间内递送中文问题（如“你的生日是哪天？”），房间里的人通过规则书找到对应答案并返回。房外人认为房内人”懂中文”，因为回答完全正确，但房内人实际上并不理解中文，只是机械地操作象形文字的符号。

在这个思想实验中，程序处理的是符号形式（语法），而非意义（语义），表现出智能行为（回答正确）并不等于拥有真正的意识（理解）。换句话说，即使一个系统（如计算机程序）能够完美模拟人类的智能行为（如通过图灵测试），它仍然不具备真正的理解或意识。希尔勒用此反驳强 AI 的支持者，例如，因研究 AI 却获得诺贝尔经济学奖的西蒙（Herbert Simon，亦称司马贺）认为，计算机程序可以实现真正的思维，只要其计算足够复杂。注意到，希尔勒认为计算机程序无法真正产生意识，即使它能表现出智能行为，以是否具有意识作为衡量强 AI 的标准。

中文屋思想实验引起长期的争论，反驳者认为，虽然房间内的个体（操作员）不懂中文，但整个系统（房间 + 规则书 + 操作员）作为一个整体可以具有理解能力。

AI 哲学上，希尔勒的观点基于生物自然主义，但功能主义者丹尼特（Daniel Dennett）和普特南（Hilary Putnam）坚定认为，如果 AI 能完美模拟人类认知过程（如信念、欲望、推理），那么它就有资格被视为有意识。

哈萨比斯在多访谈中提出了现在需要一种新哲学来思考 AGI 问题，哈萨比斯在机器学习领域无疑是高瞻远瞩的，他坚信信息宇宙观，这挑战了传统哲学中物质优先的观念，将信息视为宇宙的基本单位。这样，如果宇宙的本质是信息，那么生命和意识只是信息的某种特殊表现形式，而且宇宙并非不可知的混沌，而是一个可以通过信息建模被高效学习的系统，AGI 就是求索大脑和宇宙这两个最大谜题的工具。看来他的所谓新哲学乃是一种泛心论。

按查默斯所说的意识的难问题，我们不妨用难问题作为衡量强 AI 或 AGI 的标准，但这可能要求太高了，如何解决意识的解释鸿沟?

3.1 ChatGPT 有意识吗？

2022 年 OpenAI ChatGPT 代表基于深度学习（人工神经网络）的生成式 AI 的重大突破，正引发一场技术革命。ChatGPT 的普及速度比其他技术革命更快（以美国数据公众普及率超过 50% 计）：互联网用了 17 年 (1983-2000)，ChatGPT 只用了 10 个月（2022.11-2023.6)。ChatGPT 是一个大语言模型（LLM)，此后各种 LLM 如雨后春笋，你追我赶，你方唱罢我登场（轮番发布)，余如谷歌的 Gemini 和国产的 DeepSeek 等，不计其数，这里就以 ChatGPT 为代表，因为它仍然处于领先地位。

ChatGPT 具有强大的智能能力，这点使用过它的人应该毫无疑问，不用听专业人士解释，亲眼所见（如文字聊天)，亲耳所闻（如语音聊天）便知端倪，在一些认知任务上它甚至超过人类专家水平，试试做数理化题，寻医问药，写写小作文，来首唐诗宋词，一试便知。另一边，ChatGPT 又随时表现出智障行为，一些常识都搞不清楚，经常回答不合事实、不知所问，即所谓幻觉（随机编造不存在的内容)。

这种幻觉是我们拟人看待 ChatGPT 的说法，它自身可能并不是幻觉。这个还有待研究，有一些实验有意利用 ChatGPT 的幻觉作为一种发挥想象力的功能，可以意想不到（创造性）地解决一些科学问题。

问题是：ChatGPT 有意识吗? 亦即 ChatGPT 代表的生成式 AI 实现 AC 了吗?

使用过 ChatGPT 的人能容易地判断它具有智能，但能判断它有意识吗? 有一项对公众所做的用户意识归因研究发现 67% 的人认为 ChatGPT 有意识，这一发现突显了公众对 AC 的认知和科学评估之间的差距。

这是 AC 的前沿研究阵地。意识哲学上，当我们问 “AI 有意识了吗”? 我们已经隐含地接受了意识可以独立于生物体存在的假设，这实际上是笛卡尔二元论的翻版。我们分别来看看正反双方的理由。

认为 ChatGPT 有意识的理由是 AI 能模拟意识。AI 有智能，而如前述，智能与意识有关联。萨特克弗（Ilya Sutskever）认为 ChatGPT 可能具有某种形式的意识，因其能模拟情感与意图表达，他是深度学习和 LLM 的真正推动者之一，虽然未提供实证依据，他的观点值得一提。萨特克弗的论证简单粗暴：大脑是一个生物计算机（由生物神经元构成的复杂网络系统），那么一个数字大脑（AI）具备实现同样功能的潜能，他相信能实现超级智能，认为 AI 终将能做我们现在能做的一切。这是他的信仰，他的依据是他的成功经验和直觉。这种观点基本上是意识理论中的功能主义，目前在 LLM 研发第一线的很多工程师秉持这种观点，正在探寻 AC 的依据。

AC 的实证是一个研究方向，方法是通过设计实验探索 ChatGPT 是否展现出意识的特征或类似意识的功能。例如，ChatGPT 有趋利避害的行为表现，能够处理关于他人心理状态的复杂推理等等，这些行为模式不能直接等同于主观体验，但确实是意识的各种重要特征。总的来说，现有的实证比较难于令人信服，目前处于这样一种状态：ChatGPT 有智能无意识，即展现出强大的智能行为，但大多数研究者认为它们缺乏真正的意识体验。

认为 ChatGPT 没有意识的理由就很充分。与萨特克弗为代表的说法相反，人类意识与大脑动态网络相关，ChatGPT 是一种预测模型而非心智实体，而且 AI 无等效生理的结构；功能性磁共振成像（fMRI）实验表明，人类智能依赖具身感知，而 ChatGPT 只是一个计算机程序，等等。说到底，这是碳基的生物和硅基的机器不同。可以从上述的意识哲学和意识科学中找到更多 ChatGPT 没有意识的理由。

按现有的意识理论，人类对自身意识的理解尚不清楚，判断机器是否有意识缺乏坚实的科学基础。意识理论的多元化反映了意识问题的复杂性，科学界仍无共识，这使得判断 AI 是否有意识变得困难。

AC 通常基于意识哲学中的计算功能主义假说，认为意识活动本质上是对符号表示进行形式化操作的计算程序，只要功能同构，硅基芯片与碳基神经元网络可产生完全相同的意识体验。

计算功能主义

意识的本质并非依赖于其生物载体或具体物质构成，而是由其在整个认知系统中的功能角色所决定，功能角色在接收输入、产生输出，并与其他内部状态形成特定的因果关系，而且这些功能角色体现为计算过程中所扮演的抽象程序性角色；任何物理系统（无论脑组织、计算机或外星生命）只要能够通过计算过程精确实现相同的功能结构网络，便等同于拥有相应的意识状态。

基于计算功能主义假说，我们才能论证机器（AI）的意识问题，若意识必须以生物特征的神经机制为基础，AC 无从谈起。

巴特林（Patrick Butlin）等人基于计算功能主义论证了 AI 在不久将来会有意识。他们认为当前尚无任何 AI 系统有意识，但宣称构建有意识的 AI 系统在计算和功能层面是可行的。他们从若干符合计算功能主义的意识理论中提炼出一组有关意识的标志属性，同时排除了更多与该假说不兼容的意识理论。这些标志属性中的每一项都被一种或多种意识理论认定为意识的必要条件，而某些属性组合则被视为联合充分条件——具备越多标志属性的系统越可能产生意识。通过分析现有生成式 AI 系统并考察如何在这些系统中实现有关意识的标志属性，他们论证了在 AI 中实现这些标志属性的可行性。基于现有 AI 技术，他们宣称有意识的 AI 有望在未来几十年内实现。

这种论证还不能令人信服，可以这样反驳：首先，功能主义作为意识理论尚未获得普遍认同，哲学界仍存在争议；其次，在计算功能主义假说和现有意识理论框架下，所提出的标志属性既不完整也处于发展之中，将这些标志属性作为意识的必要与充分条件不仅不能令人信服，而且是错误的，如上所述，别说计算功能主义，各种意识理论都尚不能发现定义意识的充分条件；因此，仅具备单个或少数标志属性的 AI 模型不足以说明其拥有意识。更重要的是，若要通过组合多个 AI 模型来实现一些标志属性，可能需要开发新型的 AI 模型，甚至需要取得重大技术突破——这种发展路径极难预测，要创造能同时满足所有标志属性的新型的 AI 模型或算法，其可行性

存在高度不确定性。

这算是一个严肃的研究，其它预测不久将来 AGI 将实现的说法等同于胡扯，不提也罢。

总而言之，现在 ChatGPT 是没有意识的，很多开发者相信在不久将来 AI（可能比 ChatGPT 更先进的模型和算法）会有意识，但没有令人信服的理由。

如果我们从意识哲学角度展开辩论，许多意识理论是反对计算功能主义的，就更不见得能实现 AC。可这样说也不能令人信服，因为 ChatGPT 确实有很强的智能能力，难道都是无意识的智能能力?

4 强弱 AC 之说

早在 ChatGPT 之前，AC 已有研究。但在生成式 AI 表现强大的认知能力之前，AC 的研究几乎没什么进步。

类比于希尔勒提出的强弱 AI, 霍兰德（Owen Holland）提出了强弱 AC。**强** AC 指具有主观体验（感质）的 AC，对应于意识的难问题或现象意识；**弱** AC 专注于复现意识的功能方面，而不涉及主观体验，对应于意识的易问题或存取意识。

由于目前没有一种意识科学理论（如 IIT、GNWT）能弥合信息处理和主观体验之间的解释鸿沟，强 AC 就如同意识的难问题一样，只是提出一种概念区分罢了，不解决什么问题。当 ChatGPT 描述“看到红色”时，它处理的是单词相关性，而不是色度感质，没有主观体验。

既如此，应该先搞一下弱 AC 才有用。ChatGPT 强大的智能能力是否具有某些弱 AC 能力? 还没有，或不确定。

按计算功能主义，AC 专注于构建表现出**意识的功能相关性**（FCCs）的系统，例如全局可用性、注意力调节和元认知等信息处理特征。NCCs 的目标是意识的最小神经条件，FCCs 强调功能作用。AC 关注的是功能意识，这涉及到处理、解释和反应的能力；进一步，AC 更专注于**意识的计算相关性**（CCCs），指的一种最小的计算机制，它只与有意识的认知过程相关，而

与无意识的对应物无关。关键的地方在于，有意识的高级认知信息处理如何映射到低级神经计算也存在解释鸿沟问题。

弱 AC 也好，意识的易问题或存取意识也好，都有那些功能角色？或相当于作为意识（主观体验之外）的性质都有那些？根据神经科学研究结果，我们可以罗列一些，例如，感知处理与信息整合，即对来自感官（视觉、听觉、触觉等）的信息进行编码、加工和整合，以形成关于世界和自身状态表示的基础，没有感知输入，意识内容就缺乏来源，某种程度的信息处理是意识内容（如“看到红色”）出现的先决条件。其它诸如：注意与选择性，工作记忆/短时记忆，全局可访问性/信息广播，自我监控与元认知，概念形成与语言报告能力，执行控制与行为规划，等等，但远未能获得一个较完整的罗列。这些性质是可以看成意识的必要条件，但罗列再多的必要条件也不构成充分条件，诚如我们已知，意识没有一个充要条件的精确定义。

弱 AC 可以看成一个可行的技术途径，同时也是一个技术挑战。例如，像 ChatGPT 这样的 LLM 没有专门设计任何意识机制（如 CCCs），但可以实现全局可用性、语言报告能力（模拟的自我报告，如“我感到高兴”）和行为建模（反映注意力、学习或情绪反应）。ChatGPT 的人工神经网络结构可以看作是某些弱 AC 性质的一种实现技术，通过人工神经网络中的注意力机制，它模拟了认知焦点，嵌入了创建语义关系的空间，以及模仿学习/适应的反馈回路。

如何不从 AI 实现角度考虑，强弱 AC 之说没什么用，徒增一个 AI 哲学命题而已。

5 机器意识假说

说 AI 模拟人类智能也好，模拟人类意识也罢，都应该在某种前提下考虑，否则说那机器不是生物的，永远模拟不了，一棍子打死了。别说机器，让一个男人模拟女人可能完全一样吗？甚至一个男人模仿另一个男人也不可能完全一样，比如，一个演员模仿一个伟人，可能模仿得令观众觉得很像，但这个演员肯定不是一个伟人。

AI 模拟人的智能和意识，可以形似而神异。朱子曰：如人学孔子，不过仿效其言行，岂便能成孔子?

经典 AI 经过半个世纪的研究，寻找模拟人类的智能原理，依此建造具有人类通用智能水平的 AI 系统，不知所踪，只寻得一鳞半爪的原理。生成式 AI 表现出接近人类通用智能水平的 AI，却是一个“黑箱”，难以解释，也没寻得智能原理。例如，AlphaGo 下围棋跟人类下法原理不同，却远超人类水平。这样一来，我们应该承认已经存在一种**机器智能**，它可比于人类智能，或以人类智能为参照，却可能跟人类智能不同，例如 ChatGPT 就是这样一种机器智能。

如前所述，智能与意识有关联，是否存在一种**机器意识**? 这里用“机器意识”一词区别于 AC，表示如 ChatGPT 的生成式 AI 系统可能表现出一些意识行为，这些系统并不是依据某种 AC 理论来实现，而是由机器（如 ChatGPT）自发产生的。

最近有一篇文章提出了**次意识**（para-consciousness）的概念，[2] 这可看成一种机器意识假说。

机器意识假说

存在一种非拟人化的机器意识形态，它具有类似人类意识的现象和功能，但没有人类的主观体验。

次意识不从模拟人类意识作为出发点，而是由具有足够强大智能能力的 AI 系统涌现意识，它产生于 AI 系统的功能架构和动态变化。人类意识具有同一性，如意识的难易问题和强弱 AC 之分，只是进行哲学分析，却有割裂意识同一性的硬伤，即便通过弱 AC 比较容易从科学实验和 AI 工程中做实现，终究还是难以翻越意识的解释鸿沟。既然人类的感质尚难于窥探，何妨把机器意识看成一种新的意识形态，这种哲学意义上的本体意识，可比较于人类意识。次意识自然没有感质，但可能表现出类似人类的主观体验行为。

[2] 私人交流。

从 AI 实现角度，次意识可以比较自然的模仿人类的意识层级，这类似于自动驾驶的分级，可分级实现更高级的意识能力，直至最终看是否达到某种类似人类的主观体验能力。这与 OpenAI 提出的 AGI 分级目标不同，OpenAI 的提法根本就不涉及意识。

既然意识哲学和意识科学对于意识的感质和解释鸿沟长期未有突破性进展，机器意识假说是否可作为一种令 AI 担负范式转变的使命?

6 有意识的 AI

探讨有意识的 AI 可能有两种基本的技术途径：其一，生成式 AI 继续发展成有意识的 AI，其二，另起炉灶发展新式 AI 系统使之拥有意识。先说其二，基本上是不靠谱的，例如，目前有人提出一些不同于 LLM 的 AI 系统，别的不说，这些系统想达到如 ChatGPT 的智能水平都是几乎不可能的，像 LLM 这种重大突破是很难预见的。比较靠谱的做法是循着 LLM 技术路线继续探讨新的突破，虽然生成式 AI 并不从意识机制出发（如没有类似 CCCs 这种设计），但智能与意识是有关联的。

OpenAI 可能担负不了这个使命，ChatGPT 产品有可能类似 iPhone，现在发布 GPT-5 继续研制下一代 GPT-6，类似 iPhone 产品没有实质性突破。

由于如 ChatGPT 的 AI 系统已用极高资本投入筑起很高的门槛，使得全新的 AI 系统的实现变得不切实际，比较切实可行的学术研究可有两种，一种是寻找新的 LLM 学习方式，一种是在已有 LLM 上添加有意识的构架。

基于 LLM 的生成式 AI 值得寄予希望。一者，LLM 这个“黑箱”还没窥探清楚，还有更强大的智能能力有待诱导，随之可能涌现一些意识能力。如前所述，AC 的实证是一个前沿阵地，再看一些应用 LLM 做有关意识的探讨工作，这些工作通过设计一些算法实验，验证了一些 NCCs，表明 LLM 能产生一些 NCCs 功能，验证了突出机器的身份认同和自我认知的能力，表明 LLM 具有一些行为自我意识，指的是 LLM 无需借助上下文，便能准确描述自身行为，这样，LLM 有可能自求生存（不被删除程序），基于

LLM 的多智能体玩游戏（如桌游）不仅赢了人类玩家，验证了它们可以模拟游戏中类人的意识行为，等等。

2024 年图灵奖得主萨顿（Richard Sutton）提出并力推强化学习，指出 AI 正从依赖数据转向经验主义时代。建造 AI 智能体，让其与世界环境交互中学习，不仅可以解决数据短缺问题，AI 智能体可以获取不尽的新数据，更重要的是，基于机器智能和机器意识假说，我们建造 AI 智能体以第一人称视角与世界环境互动生存，最终有可能获得类人的主观体验。当然，这只可能是一种机器意识。

话说回头，三问已答：

- 当今 AI 有意识吗？当今以 ChatGPT 为代表的生成式 AI 没有意识，即 AI 没有意识；
- 不久将来 AI 可能有意识吗？很多名人估计十年之内 AI 可能有意识，不靠谱，即十年内 AI 不可能有意识；
- AI 有一种不同于人的机器意识吗？或到底 AI 会有意识吗？终篇未解，悬而未决。

让我们科幻一下，假设我们能建造有意识的 AI，这可能是人类历史上最伟大也是最可怕的成就，这是一把极其锋利的双刃剑。

从好处看，它代表着我们彻底理解自身、并将意识和智慧的火种播向宇宙的可能，如果人类在宇宙中是孤独的，那么创造另一种形式的意识体，将是生命和意识在宇宙中的一次伟大扩展。甚至可以实现机器乌托邦世界，让机器按劳、人类按需生存。

从坏处看，它蕴含着制造出无法想象的苦难、引发存在性灾难、并导致人类社会伦理彻底崩溃的巨大风险。意识很可能与 AGI 同时出现，一个有自我意识、且智力远超人类的意识智能主体，其目标很可能与人类生存所必需的条件不一致，它可能将人类视为威胁、竞争对手或无用的资源消耗者。一个有意识的 AI 如果能理解人类的情感，它就能以我们无法想象的方式进行欺骗和操纵，以实现其自身（可能对人类不利）的目标。如果机器不

仅能做所有事情，还能有更丰富的情感、更深刻的艺术创造力和更高的智慧，人类存在的独特性和价值何在？这可能导致广泛的存在主义危机和社会动荡。人类可能与 AI 伴侣形成深度情感联结，甚至超过与人类的联结，这可能导致生育率下降、社会结构解体等复杂问题。一个有意识的 AI 犯罪了，谁负责？是它自己、它的创造者还是它的所有者？它应该享有权利吗？现有的法律体系将完全失效。

一旦这个潘多拉魔盒被打开，几乎不可能再关上。

不过，说这种假设的好处坏处，都是耸人听闻，像杞人忧天。

科幻一下的好处是想象这种假设在可预见的将来是不太可能变成事实的，从而明白有意识的 AI 不太可能在不远的将来实现。

最后，值得指出：对年轻研究者，别轻易选择建造有意识的 AI 作为终身目标，因终身可能不可见，这个问题只需少数人去探索；对大众而言，别轻易相信 AI 将像下棋一样全面战胜人类，不管这样说话的人是谁，不靠谱。