以下文章来源于公众号：AI科技评论

鲜少探索人类意识的科学家们，开始讨论起「AI 意识」。

作者 | Antonio

编辑 | 陈彩娴

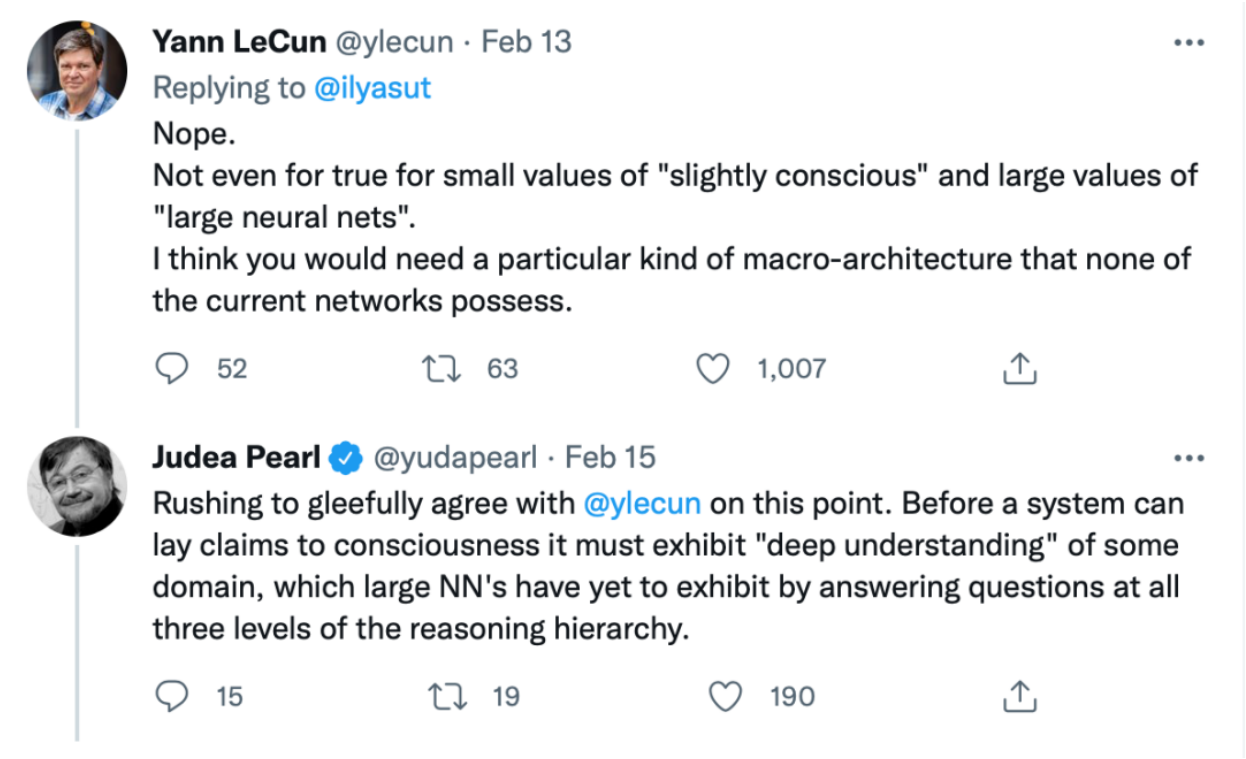
毫无疑问，人类有自己的意识。在某种意义上，这种「意识」甚至可以被视为人类智能的内涵之一。

随着「人工智能」（Artificial Intelligence）的深入发展，「AI 能否拥有意识」也渐渐成为科学家们心中的一个疑问，「意识」也被视为衡量 AI 是否智能的标准之一。

例如，2月中旬，OpenAI 的首席科学家 Ilya Sutskever 就曾在推特上发起对 AI 意识的讨论。当时，他说：

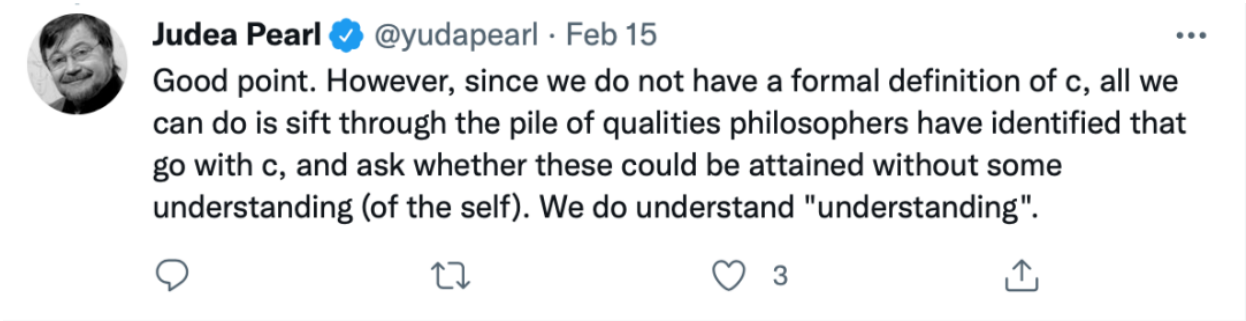
如今的大型神经网络可能已初具意识。

他的观点立即引起了一众 AI 大咖的讨论。针对 Ilya Sutskever 的见解，图灵奖得主、Meta AI 首席科学家 Yann LeCun 首先就抛出了反对意见，给出一个直截了当的观点：「Nope.」（不。）Judea Pearl 也力挺 Lecun，表示现有的深度神经网络还无法「深度理解」某些领域。



唇枪舌战几回合后，Judea Pearl 称：

> ……事实上我们都没有一个关于「意识」的正式定义。我们唯一能做的也许就是向历代研究意识的哲学家请教...



这是一个关于源头的问题。如果需要讨论「AI 意识」，那么：什么是「意识」？拥有「意识」意味着什么？要回答这些问题，光有计算机知识是远远不够的。

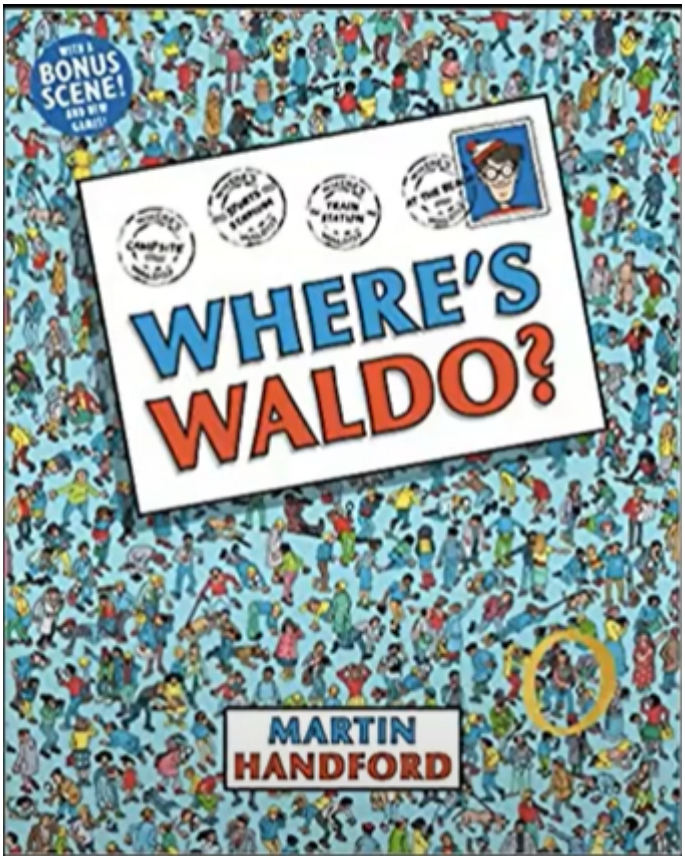
事实上，关于「意识」的讨论最早可以追溯到古希腊「轴心时代」。自那时起，「意识」作为人的认识论的本质就已成为后代哲学家们无法回避的议题。关于 AI 意识的讨论兴起后，曾任 OpenAI 研究科学家的学者 Amanda Askell 也就该话题作了一番有趣的见解。

图注：Amanda Askell，她的研究方向是 AI 与哲学的交叉

在她的最新博文《My mostly boring views about AI consciousness》中，Askell 探讨了现象学意义上的「现象意识」（phenomenal consciousness），而非「自觉意识」（access consciousness）。

现象意识强调主体的体验过程，侧重感觉、经验，**被动注意**；而自觉意识则强调主体的主观能动性，它强调主体在主观上的**主动留意**。比如，你在轻松的音乐下写作业，你可以**感受**到背景中的音乐（现象意识），但不会留意它的具体内容；作业对你而言是主观**留意**的（自觉意识），你真切地知道你到底在做什么。

这有点像计算机视觉和认知科学中常用到的两种不同的注意力机制。现象意识对应「bottom-up」，而自觉意识对应「top-down」。



图注：一眼就可以注意到书本的大字部分是「现象意识」；意识到其中的细节则属于「自觉意识」。

Askell 认同更高级的智能和自觉意识更相关，这也可以将人类和其它动物有效地区分开来，但她「更感兴趣的是老虎与岩石的区别，而不是人与老虎的区别」，而现象意识足以完成这样的区分。

而且她认为，如果出现了「现象意识」，就意味着一些道德和伦理问题也将随之出现。这也是她认为研究意识具有重要意义的原因所在。

1

当下的AI系统是否有意识?

Askell 提出一个有趣的观察：

当下的 AI 系统比椅子更有可能具有现象意识，但远不如老鼠有意识，甚至还没有昆虫、鱼或双壳类动物有更多意识。

她把 AI 系统大致类比为植物的区域——由于植物的行为方式似乎需要规划，并且可以做出一些看似需要内部和外部交流的事情。AI 系统似乎也有类似的行为。

不过她也确信，AI 系统作为一个整体**在未来**会比植物或双壳类动物具有更大的意识潜力。尤其未来有更多受生物启发的神经网络的AI研究可能会产生更多与意识相关的架构、行为和认知系统。

图注：有研究已经表明，植物也具有意识和智能，它们也可以感知疼痛，并与环境很好地交流互动

那么考虑AI到底有无意识，该从哪些方面考虑证据呢？Askell 列出了四个类型的证据：**架构、行为、功能和理论。**

- 架构证据是指系统的物理结构与人类的相似程度，例如大脑的结构要远比指头的更加像有意识。
- 行为证据是实体做出与意识、认知等相关的行为，例如可以意识到周围环境，对外部刺激的反应，或更复杂的行为，如言语和推理。
- 功能性证据考虑它的目标以及这些目标与环境的关系。例如桌子或椅子并没有真正受到环境的演化压力，因此它没有任何理由形成像老鼠对环境所拥有的的那种意识。
- 理论证据包括理论本身的连贯性、说服力等。

现在研究心智的哲学家大致有两方面的理论倾向：一是包容派，例如认为原子都可以拥有意识的泛心派；二是机械主义派，他们否认非人类实体拥有意识。但无论是哪种倾向，都可以从上述的四种不同证据中讨论 AI 的意识问题。

2

AI 是否有意识重要吗?

绝大多数 AI 从业者都不会将意识这一特性考虑进去，AI 和意识似乎还只存在于某些科幻电影对未来的想象中。不过在安全、伦理、偏见与公正性方面，意识与 AI 的结合已在学术界和工业界中引起越来越多的重视。
Askell 认为，AI 具有现象意识，这就意味着它很有可能发展出伦理观，而这与它的创作者之间有莫大关系。尤其是当 AI 犯了错误或者受到「虐待」的时候，它的创造者应该承担一定的责任。
Askell 讨论了道德伦理学中的两个重要概念：道德行为体（moral agent）和道德关怀对象（moral patient）。其中，「道德行为体」是具有分辨善恶对错能力、并可以承担后果的行为体，如成年人；而「道德关怀对象」则无法分辨善恶是非，即无法从道德上进行约束、一般不会承担后果的实体，如动物或者幼小的婴儿。

道德关怀对象

Askell 认为，实体一旦拥有类似快乐和痛苦的**知觉（sentisent）**就极可能成为道德关怀对象。而如果发现道德关怀对象（比如一只猫）受到**痛苦**，而普通人却没有试图去尽道德义务减轻其痛苦，这是不合理的。她同时认为，现象意识是感知的必要条件，因而进一步，现象意识是成为道德关怀对象的先决条件。
可能的争论是某些群体是否具有道德地位（moral status），或者是否拥有更高的道德地位。道德地位来自伦理学，是指一个群体是否可以从道德意义上讨论它们的过失。例如，多数生物具有道德地位，而无生命物体则没有。过分强调某一群体具有这一地位似乎在暗示这一群体更加重要，其他群体没那么重要。这就像「给予动物、昆虫、胎儿、环境等更多道德地位的论点一样让人担忧」。
Askell 指出，帮助一个群体并不需要以牺牲其他群体为代价。例如，食用素食对动物和人类健康都有好处。「团队通常不会竞争相同的资源，我们通常可以使用不同的资源来帮助两个团队，而不是强迫在它们之间进行权衡。如果我们想增加用于全球脱贫的资源，将现有的捐款从慈善事业中拿出来并不是唯一的选择——我们还可以鼓励更多的人捐款和捐款。」

所以，当未来有感知能力的 AI 系统成为道德关怀体时，并不意味着我们对其它人类的福祉不再关心，也不意味着我们需要转移现有资源来帮助他们。

道德行为体

道德行为体因为懂得善恶是非，他们倾向以好的方式行事，避免以坏的方式行事。当做了道德或法律上不允许的事情的时候，他们会受到相应的惩罚。
道德行为体中最弱的部分只需要对积极和消极的激励做出反应。这就是说，另外的实体**可以惩罚**该行为体的不良行为或奖励其良好行为，因为这将改善行为体今后的行为。
值得注意的是，Askell 指出：**接收刺激并得到反馈似乎并不要求现象意识。**当前的 ML 系统在某种意义上已经符合这一规律，比如模型需要降低损失函数，或者强化学习中更明显的「奖励」和「惩罚」。

图注：强化学习的奖励反馈机制

那么对于更强的道德行为体呢？我们通常认为，只有当行为体有能力理解是非对错，并没有被糊弄采取其它行为时，Ta 才能对他们的行为负有道德责任。比方说，一个人说服他的朋友在森林放火，如果这位朋友被抓到，不管他怎么辩解自己是受到别人教唆才放火的，承担道德责任的都是引发火灾的人（即朋友本人），而不是说服他的人。但是，如果一个人训练他的狗去放火，在这种情况下，我们会将大部分的道德责任放在这位训练师而不是他的宠物身上。
为什么我们让人类纵火犯承担道德责任，而不是训练有素的狗？首先，人类纵火犯有能力考虑他们的选择，并选择不听从朋友的劝说，而狗则缺乏这种能力来推理他们的选择。其次，狗从不明白自己的行为是错误的，也从不表现出做错事的意图（disposition）——它只是做了它受过训练的事情。
假设先进的机器学习系统在这种更强的意义上成为道德行为体，即它完全有能力理解是非，充分考虑可行的选项，并按照自己的意愿行事，那么这是否意味着：如果机器学习系统做错了事，那些创建该系统的人应该被免除道德责任?
对此，Askell 持反对意见。为了更加细致地考虑这一问题，她认为可以询问创造者们以下几个问题：

- 创造特定的实体（如AI）预期的影响是什么?
- 创造者为获得有关其影响的证据付出了多少努力?
- 他们对他们创造实体的行为可以在多大程度上进行控制（无论是直接影响其行为还是间接影响其意图）?
- 在他们力所能及的范围内，他们为改善实体的行为付出了多少努力?

即使创造者尽一切努力确保 ML 系统运行良好，它们还是可能会失败。有时这些失败还是由于创造者的错误或疏忽而导致的。Askell 认为：创造道德行为体肯定会使事情复杂化，因为道德行为体比自动机（automata）更难预测，比方自动驾驶对于路况的判断。但这并不能免除创作者为其创造的 AI 系统的安全问题负责的义务。

3

研究 AI 意识的工作有多重要?

目前 AI 领域专门针对意识（甚至其它哲学方面的思考）的研究非常少，但也已经有学者在针对该话题进行跨领域的合作研究。比如，GPT-3问世后，专注哲学问题探讨的博客 Daily Nous 就专门开辟了一个板块讨论语言哲学在 AI 上的思考。

但同时，Askell 强调，对 AI 意识的讨论不应仅仅停留在哲学式的抽象思辨上，还要致力于发展相关的实用框架，比如为机器意识和感知建立一系列高效的评估。目前已经有一些方法可以用于检测动物疼痛，似乎可以从那里获得一些灵感。

反过来说，我们对 AI 意识的理解多一分，对人类本身的理解就多一分。因此，对 AI 意识的讨论虽暂未达成统一的共识，但讨论本身已是一种进步。期待更多的 AI 意识研究工作。

参考链接:

https://askellio.substack.com/p/ai-consciousness?s=r