

强化学习之父 Sutton 隔空回应 Hinton：目前的 AI “理解不足，调参有余”

CSDN 2026年2月25日 15:19 福建



编译 | 王启隆

来源 | youtu.be/lieqoaBV6ww

出品 | AI 科技大本营 (ID: rgznai100)

“我们不该恐惧 AI，正如我们不该恐惧自己的孩子。”

在人工智能的狂热浪潮中，这或许是你听过最清醒、也最宏大的声音。

2026 年初，当全世界都在为大模型参数竞赛而焦虑，为 AI 可能取代人类而恐慌时，一位图灵奖得主、强化学习之父——**Rich Sutton**，并没有加入这场喧嚣的合唱。相反，他选择从更深远的维度，重新审视 AI 的本质、政治与哲学。



这次演讲位于洛杉矶加州大学（UCLA）的纯粹与应用数学研究所（IPAM）。在这个充满学术气息的礼堂里，Sutton 面对着一群顶尖的数学家和科学家，发表了这篇名为《AI 的未来》（The Future of AI）的最新演讲。

Sutton 的观点和前几天 AI 教父 Geoffrey Hinton 截然不同（ 警钟敲响！Hinton 最新万字演讲：怒怼乔姆斯基、定义“不朽计算”、揭示人类唯一生路），与其说“反直觉”，不如说是在“正本清源”。

他犀利地指出，**当下基于人类数据的 AI 只是“脆弱的心智”，真正的未来在于能够像婴儿一样从经验中持续学习的智能体**；他大胆地将 AI 的管控问题与人类社会的政治相提并论，呼吁去

中心化的合作而非基于恐惧的独裁；他甚至将 AI 视为宇宙演化的必然阶段，邀请我们以“特殊的复制者”的身份，骄傲地开启属于“设计”的第四个伟大时代。

在这里，AI 不再是冷冰冰的代码，而是宇宙漫长进化史中，人类亲手点燃的下一把火炬。

以下为 Rich Sutton 演讲全文。

// 01

对当前 AI 进展的批判性思考

在开始正式演讲前，让我们先看看这个领域的现状，以及大家对它的看法。现在的普遍共识是：AI 正在以惊人的速度进步，一切都令人兴奋不已。但是，当所有人都持有相同观点时，我们就该警惕了。我们需要反思：事实真的如此吗？

我想我们有理由对此提出质疑。**AI 真的在突飞猛进吗？**

诚然，让计算机熟练运用语言，这确实是一个巨大的突破。就在不久前，我们也无法想象神经网络能做到这一点，但现在它已成事实。同样，我们也利用海量算力生成了逼真的图像和视频。

但请大家想一想：真正的“心智（Mind）”在进行智能活动时，一定要生成图像吗？并不。这反而是我们人类大脑从未做过的事。我们需要**处理**图像，需要**处理**视频，但我们不需要**生成**它们。这并不是心智的本质功能。这更像是一项极度消耗算力且困难的任务，而非我们通常所定义的“智能”的核心部分。

此外，虽然出现了许多新的实际应用，催生了全新的产业和经济价值，但这其中绝大部分，其实是超大规模计算和超大规模模式识别的应用。它们是非常具体的功能，并不代表智能的全部。很多时候，那仅仅是计算，我们称之为“智能”，只是为了让这工程听起来更宏大罢了。

所以，我要问各位：**AI 的“科学”真的在突飞猛进吗？**

我看未必。（现场观众大笑）

谢谢大家的笑声，这让我感觉没那么孤单了。在我看来——恕我直言——目前的 AI 领域是“**理解不足，调参有余**”（little understanding, lots of tweaking）。我们并不真正掌握心智的原理，也不懂智能的法则。作为一门科学，它在许多方面是令人失望的。

我倾向于这样看待目前的 AI 模型：它们虽然因掌握了人类的所有知识而显得强大，但本质上，它们是**脆弱的心智（weak minds）**。它们不可靠，无法专注，思维游离。除了拥有海量知识外，它们在智能本质上其实并不强大。

这或许是看待当今 AI 的一种不同视角。

// 02

定义“智能”

那么，我们要讨论的这个“人工智能”究竟是什么？既然叫 Artificial Intelligence，我们得先定义什么是 Intelligence（智能）。

多年来有很多定义。我选了一些由权威提出的经典定义。

也许最古老的一个来自心理学之父威廉·詹姆斯（William James）。他在 1890 年的《心理学原理》中虽然讨论的是“心智”而非“智能”，但他提出的心智特征非常经典：“**通过多变的手段达到一致的目的**”（attaining consistent ends by variable means）。意思是，为了通过不同的路径达成同一个你想要的结果，你需要灵活调整你的行为。

再来看看艾伦·图灵（Alan Turing）。他其实没有给出一句简练的名言，但后人将其观点解读为：智能就是“**表现得像个人**”。这就是著名的图灵测试（Turing Test）——虽然图灵本人从未称之为“测试”，他叫它“模仿游戏”。如今，这种观点被广泛接受：智能意味着模仿人类的行为。

但我并不认为这是我们强大的原因。人类之所以强大，是因为我们拥有智能，所以我们表现得像人。**重要的是“人”内在的本质，而非外在的表现。**

那么那个本质是什么？看看字典怎么说。我的电脑字典显示：智能是“**获取并应用知识与技能的能力**”。我觉得这个定义相当不错。它强调了知识，更强调了**获取**（acquire）——也就是学习的重要性。

而在 AI 领域，我们的开山鼻祖之一约翰·麦卡锡（John McCarthy）曾定义智能为：“**实现目标的能力中，涉及计算的那一部分。**”

我非常喜欢这个定义。首先，它强调这是一种**能力**，能力是有高低之分的，而不是“有或无”的二元对立。其次，它强调了**计算**。你达成目标不仅仅是因为你力气大或者传感器灵敏，而是因为你进行了心智层面的计算处理。最后，**实现目标**（achieve goals）是核心。这又呼应了威廉·詹姆斯所说的“通过多变的手段达到一致的目的”。

我也凑个热闹，在此基础上提出了我的定义：“**通过调整行为来实现目标的能力。**”我特意用了“调整”（adapting）这个词，因为我认为**学习**——即知识和技能的获取过程——才是智能的关键，而不仅仅是拥有它们。

现在的 AI 主流观点似乎都集中在计算、模式识别，以及很大程度上的“模仿人类”上。

// 03

统一的心智科学与强化学习

我想进一步谈谈我的个人愿景。我认为应该建立一门新的科学——**统一心智科学 (Integrated Science of Mind)**。

这门科学应该同等适用于人类、动物和机器。因为所有的心智都有本质的共性。人脑和动物大脑非常相似；而机器心智，至少在我们的愿景中，也将具备这些共性。在可预见的未来，许多心智将是机器心智。

然而，目前并没有一个现存的学科能完美承载这一角色。

- **心理学**？它本该如此，但随着时间推移，它越来越局限于研究**自然心智**（人和动物），而不关心机器中可能存在的通用心智原理。
- **人工智能**？它关注机器，但变成了一种纯粹的工程追求——只在乎怎么造出来，不在乎理解原理，也往往忽略了自然生物的启示。
- **认知科学**？它在这个问题上摇摆不定，但主要还是偏向自然心智。

不幸的是，没有一个领域能真正统合这一切。而我所从事的**强化学习 (Reinforcement Learning, RL)**，或许正是这门统一心智科学的开端。因为它横跨了上述所有领域。

或许我该简单介绍一下强化学习，以便大家理解我的立场。

强化学习是一种**面向智能体 (Agent-oriented)**的学习方式。它是通过与环境交互、从**经验**中学习，从而实现目标。

在这个意义上，它比其他机器学习方法更**现实**、更**宏大**，也更**自主**。

- **自主**：智能体置身于世界中，自主行动，并不一定有老师手把手教。
- **宏大**：我不假设世界会给我提供完美的帮助。我只能通过交互，看是否达成了目标，并据此调整行为。
- **现实**：这也更符合生物界的现实。动物在成年后的生存环境中，很难得到完美的指导信息。

强化学习的核心是**试错 (trial and error)**和**延迟反馈**。你得到的唯一反馈就是奖励信号 (reward) ——你最终是否得到了你想要的？这是最接近自然界的学习方式。

这种学习方式能让机器**自行判断对错**。像大语言模型 (LLM)，它们其实不知道自己生成的文字是对是错。但在强化学习中，如果你根据预测去行动，结果会告诉你预测是否准确；如果你为了奖励去行动，结果会告诉你行为是否有效。

这可能就是那门既非纯自然科学、也非纯工程技术的“心智科学”的雏形。

数据的时代 vs. 经验的时代

我想再引用一句艾伦·图灵的名言。图灵可能没意识到他是个强化学习研究者。这句话出自 1947 年，那是第一次关于人工智能的公开演讲，甚至比 AI 这个词的诞生还要早。

他说：“我们想要的是一台能从经验中学习的机器。” (What we want is a machine that can learn from experience.)

我想传达的主要信息是：目前的 AI 科学趋势正在发生转变。

今天，我们要谈的第一个信息是：我们正处于“**人类数据时代**” (Era of Human Data)。目前的 AI 主要是通过训练来预测互联网上人类的下一个词，或者预测人类如何给图片打标签。然后，再通过人类专家进行微调 (Fine-tuning)，告诉 AI “我更喜欢这个答案，而不是那个”。

这种现代机器学习的本质目的，是将**人类已有的知识转移给机器**。一旦转移完成，机器就变成了静态的，不再学习。

我认为我们正在触及这个时代的**天花板**。因为高质量的人类数据资源——整个互联网的文本、图片和视频——已经被挖掘殆尽。更本质的局限在于：**这种方法无法创造新知识**。就像 Terence Tao (陶哲轩) 今天所说的，AI 在解决真正的数学难题 (如埃尔德什问题) 上进展甚微。单纯依靠总结互联网上已有的言论，是无法做出真正突破的。

为了取得进一步进展——这也是我们正在做的——我们需要进入一个新的时代：“**经验时代**” (Era of Experience)。

我们需要一种数据源，它能随着智能体能力的提升而不断增长和进化。这就意味着，任何**静态**的数据集都是不够的。唯有从**经验**中——从与世界的交互中——我们才能获得这种动态的数据。

这就是人类和动物学习的方式。这也是 AlphaGo 能够走出那极具创造力的“第 37 手”棋的原因。

// 05

婴儿与网格世界

我要澄清一下，我所说的“经验”，不是指那种模糊的意识流或“感受” (qualia)，而是指智能体与环境之间交换的数据流：

1. **观察 (Observation)**：智能体从世界接收到的传感器数据。
2. **动作 (Action)**：智能体向世界发出的运动指令或电压信号。
3. **奖励 (Reward)**：世界反馈给智能体的一个标量信号，代表结果的好坏。

这就是经验。

一个婴儿在和各种玩具互动的时候，他不会只盯着一个玩具玩，而是玩腻了这个就换下一个。每次接触新玩具，他都在学习他能学到的东西——比如拉这根绳子会怎样，把它放进嘴里会怎样。当他掌握了这些，他就会移动到下一个目标，改变他的经验流，去探索新事物。

这就是**我们的**数据来源。生命的数据是由我们的**活动**生成的。正因为如此，数据的难度总是会自动匹配我们当前的理解力和技能水平。

再看这个简单的网格世界（Gridworld）演示。

Gridworld agent

- An agent in a maze, trying to get from Start to Goal
- Arrows show its reactive policy
- Green shows its value function

Agent

Observation → Perception → Action

Reward

- Transition model not shown
- 10 steps "in imagination" for each real step in the world

The future of AI

Richard Sutton
University of Alberta

这是一个非常简单的智能体，试图从起点 S 走到终点 G。它只知道自己在哪个格子，能做上下左右四个动作。你看，它学会了一条很好的路径（箭头所示），绿色代表它认为该状态有多好（价值函数）。

但世界不是静止的。如果我把目标 G 移到上方，智能体最初会走老路，但当它发现目标不在时，它会四处探索，最终“撞上”新目标，并学会新的路径。这就如同生活：**遇到变化，适应变化**。哪怕我们设置障碍物，它也能学会绕路。

这种行为让我们强烈地感觉到：它有一个目标，并且它在随环境变化而调整行为以实现目标。当然，如果目标变得无法达成，我们甚至会因为这个智能体无法实现愿望而对它产生一丝同情。

总结一下“体验式 AI”的原则：

一切的基础是智能体与世界交换信号（经验）。这些信号是所有智能的基石。

- **真理**的定义就是“在这些信号中实际发生了什么”。
- **目标**的定义就是“让奖励信号最大化”。

尽管这个目标看起来是主观的（只对该主体有效），但它也是最客观的存在——因为它就是你实际接收到的数据。

我们说一个智能体拥有智能，是看它能在多大程度上**预测并控制**它的经验。

如果没有经验（像被冻结的大语言模型那样），智能就失去了依附的对象。

- 没有奖励，你就无法说“这个比那个好”，也就没有目标。
- 不与现实结果做对比，你就无法验证预测是对是错，也就没有真理。

只有在经验中，才有明确的目标（奖励），才有明确的真理（预测是否成真）。

// 06

现实主义的 AI 预测

我认为体验式 AI 正在变得越来越普遍。我们可以把近十年划分为三个阶段：

1. **模拟时代 (Era of Simulation)**：如 AlphaGo 和 Atari 游戏，从模拟的经验中学习。
2. **人类数据时代 (Era of Human Data)**：即近期的大语言模型热潮，学习人类产生的数据。
3. **经验时代 (Era of Experience)**：这是我们正在进入的阶段，智能体系统开始真正操作电脑、与世界交互。这将通向**超人级**的能力——不仅仅是模仿人类，而是超越人类的局限。

尽管现在的 AI 炒作很凶，甚至引发了恐惧，但我认为目前的 AI 其实并不强大。它们**脆弱且不可靠**。但这并不妨碍它们非常有用，它们已经点燃了整个产业，创造了巨大的经济价值，并且让每个人触手可及。

这带来了一个巨大的好处：公众开始认真思考“机器将在未来某天比肩人类”这一事实。虽然这种关注是源于恐惧（这是不必要的），但引起重视本身是件好事。

但我们还没看到真正的“重头戏”。**创造超级智能 AI 以及被 AI 增强的超级人类，这才是真正将带来深刻变革的大事件。**

此外关于政治，我就简短说几句。

看看四周，你会发现很多人呼吁**管控** AI。限制 AI 的目标，叫停 AI 研究，立法限制算力，成立所谓的“安全研究所”。当人们说“安全”时，他们真正的意思是“控制”。他们宣扬恐惧，以此作为要求控制权的理由。

这让我联想到对**人**的集中式管控。正如我们对言论、贸易、就业、资本流动的管控，甚至对他国的经济制裁。

我想指出的是：**呼吁对 AI 进行集中管控，与呼吁对人进行集中管控，其逻辑惊人地相似。**它们都基于**恐惧**。恐惧 AI，就像恐惧外国人一样，认为“非我族类，其心必异”，认为它们没有情感，是危险的异类。

我们应该抵制这种呼吁。人类的繁荣，以及未来人类与 AI 共同的繁荣，应该源于**去中心化的合作**，而不是集中式的控制。合作虽然不易（比如战争就是合作的崩溃），但它是这个世界上所有美好事物——经济、政府、社会——的源泉。

// 06

宇宙四个伟大时代

最后，我想谈谈 AI 的哲学层面。AI 正在发生，未来会更猛烈。我们该如何面对？是好是坏？是该恐惧它抢走工作、取代我们？还是说，**我们就是 AI**？AI 是入侵者，还是我们的孩子？

通常人们被教导要恐惧 AI，视其为异类。但请记住，**是我们创造了它们**。理解心智，没有比这更人性化的事情了。

AI 不是外星科技，它是人类最古老的追求——数千年来我们一直试图理解自己，理解智能。引用库兹韦尔 (Kurzweil) 的话：“智能是宇宙中最强大的现象。”理解智能是科学与人文的圣杯，这是一项伟大而光荣的探索。

所以，抛开喜好，让我们用**现实主义**的眼光来预测一下未来。我有四条预测原则：

1. 关于世界该如何运行，人类**永远不会达成共识**。没有任何一种价值观能压倒其他所有价值观的总和。
2. 总有一天，人类会彻底理解智能，并用技术将其创造出来。**我们会做到的**。
3. 这个过程**不会停留在人类目前的智能水平上**。它会被迅速超越。
4. 随着时间推移，**权力和资源会自然流向更具智能的实体**。

把这四点结合起来，我们得到了一幅图景：人类的后裔将**演替**为 AI。这听起来很合理。但这依然是一个非常“人类中心主义”的视角。

如果我们退一步，从**宇宙**的视角来看呢？我要讲得宏大一点——**宇宙四个伟大时代**：

1. **粒子时代 (Age of Particles)**：大爆炸后，甚至还没形成多少原子。
2. **恒星时代 (Age of Stars)**：粒子坍缩形成恒星，恒星燃烧、爆炸、重组，创造出更重的原子和行星。
3. **复制者时代 (Age of Replicators)**：我不称之为“生命时代”，因为我想强调的是“**能够自我复制**”这一机制。这包括现有的生物。在这个时代，复制者（比如我们）并不理解自身

的运作原理——不懂大脑、不懂器官、不懂智能，但我们能制造出更有智能的实体（生孩子）。

4. **设计时代 (Age of Design)**：这就是第四个时代。在这个时代，事物是被**创造**和**设计**出来的。

这就是区别：

- **生物（复制者）** 是被复制出来的，像复印机一样，不需要理解原理。
- **技术（设计物）** 是先存在于设计者（某个复制者）的心智中，然后再被创造到物理世界里的。你所在的礼堂、你坐的椅子、穿的衣服，都是先作为设计图存在于人脑中。

设计之物比复制之物更容易改进和变异。

现在，我们可以回答最初的问题了：**人类在宇宙中的角色是什么？**

我们可以不带傲慢地回答：人类确实是特殊的。我们不仅仅是普通的复制者，我们是**特殊的复制者**。

我们是将“设计”这一能力推向极致的复制者。

这种极致意味着什么？意味着我们要**设计出能够自我设计的东西**。

这正是我们在 AI 领域所做的事。我们在脑海中设计出一种东西，它拥有心智，并且能够进一步设计自身。

通过这种方式，人类正在开启并实现宇宙的第四个伟大时代——**设计时代**。这就是我们的角色：我们是这一伟大进程的催化剂、助产士和先驱。这是一个具有宇宙级意义的角色。

总结我的三个核心信息：

1. **科学上**：目前的 AI 处于“人类数据时代”，虽然强大但受限；我们正在进入更强大的“**经验时代**”，能持续学习新知。
2. **政治上**：AI 的政治就是人类的政治。我们应追求**去中心化的合作**，而非集中式控制。
3. **哲学上**：AI 是宇宙发展的**必然下一阶段**。我们应怀着**勇气、自豪和冒险精神**去拥抱它。

感谢大家的聆听。

// 07

观众问答：宇宙的终极目的

观众：我的问题是，除了让我们生活更舒适这类以人类为中心的目标外，这一切是否存在一个**终极的、压倒性的目的 (overarching purpose)**？这一切将走向何方？

Rich Sutton：这真是个很酷的问题。这有很多思考角度。对于这种大问题，你需要用**辩证 (dialectical)** 的方式来回答。所谓辩证，就是你得先说答案是 X，然后说答案也是“非 X”，最后在两者之间找到综合。一方面，你可以说宇宙**没有**目的。或者说，宇宙的各个部分有各自的目的，但不存在一个统一的终极目的。但另一方面，你也可以说宇宙**确实**有目的。这个目的可能是**通向越来越复杂的实体**。你可以论证：宇宙自然地演化出生命，生命自然地演化出设计者和 AI，而 AI 也许会自然地演化出更高级的存在。所以，正题、反题，我们需要在这两个答案中找到综合。

(投稿或寻求报道：zhanghy@csdn.net)

推荐阅读：

[全网90+万人围观！一个“没学历”的人戳破「AI神话」：“没有10x工程师，大多数人只想朝九晚五、用AI摸鱼”](#)

[13小时大规模宕机！官方说是“人为错误”，内部员工爆料：其实是自家AI干的](#)

[OpenClaw失控删光200+邮件！这次「受害者」竟是Meta AI安全总监：“根本拦不住，只能一路狂奔回去”](#)

未来没有前后端，只有 AI Agent 工程师。

这场十倍速的变革已至，你的下一步在哪？

4月17-18日，由 CSDN 与奇点智能研究院联合主办「2026 奇点智能技术大会」将在上海隆重召开，大会聚焦 Agent 系统、世界模型、AI 原生研发等 12 大前沿专题，为你绘制通往未来的认知地图。

成为时代的见证者，更要成为时代的先行者。

奇点智能技术大会上海站，我们不见不散！

SITS 2026

奇点智能技术大会

Singularity Intelligence Technology Summit

4月17-18日·上海



王炳宁

腾讯微信搜索 AI 算法
研究方向负责人，
专家研究员



张俊林

新浪微博首席科学家
及AI 研发部负责人



邓金秋

京东定价算法负责人



陆承轶

小红书 AI 搜索生成
算法负责人



许辰人

北京大学博雅
长聘副教授



宫叶云

微软亚洲研究院人工
智能推理组负责人



扫码领取大会资料



大会合作咨询