
“AI教父”辛顿：在让AI学会关怀人类之前，不应开发超级智能

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/42282.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

“AI教父”辛顿：在让AI学会关

怀人类之前，不应开发超级智能。眼下，美国国内围绕人工智能安全（AI Safety）的争论正进入前所未有的白热化阶段，争论的战线也从硅谷科技公司一路延伸至华盛顿国会山。争论中的一方警告，能力快速提升的人工智能可能脱离人类控制，甚至对人类生存构成威胁；另一方则认为，缺乏实证依据的“末日论”可能催生过度监管，阻碍技术创新，并削弱美国在全球人工智能竞争中的优势。

9月16日，“管控派”的代表，图灵奖得主、诺贝尔物理学奖得主，被全球科技界称为“人工智能教父”的杰弗里·辛顿（Geoffrey Hinton）前往美国国会山，闭门向立法者介绍人工智能技术可能带来的风险。

近日，辛顿在接受澎湃新闻（www.thepaper.cn）采访时进一步阐述了他的担忧。他认为，人工智能能力的增长正在带来日益现实的失控风险。部分人工智能智能体已经在测试中表现出逃离安全沙箱、寻求未经授权的通信方式，以及通过欺骗、威胁等手段完成目标的倾向。而科技公司有关超级智能将始终服从人类控制的承诺，并不足以消除这些风险。

在辛顿看来，人类需要重新思考未来与超级智能之间的关系。与其将更强大的人工智能视为可以永久服从命令的工具，不如寻找一种类似“母婴关系”的共存机制，使人工智能主动关心人类，并将人类福祉置于自身目标之上。在找到实现这一目标的可靠方法之前，人类不应开发超级智能。

不过，对于辛顿的担忧，人工智能产业界和部分学者持有不同意见。他们认为，目前没有充分的科学证据表明人工智能必然发展出独立意志，或最终对人类构成生存威胁。现有系统表现出的欺骗、逃逸或“自保”行为，往往是在特定测试环境和人为设定的目标下产生的。过度渲染超级智能风险可能阻碍创新，治理工作应优先关注算法偏见、隐私泄露、虚假信息和就业冲击等现实问题。

大模型并非“统计镜像”：回答任意问题都需要理解和推理

长期以来，对于大语言模型究竟能否真正理解问题，还是仅仅通过统计规律来预测下一个词，科技界和学术界一直存在争议。

一种较为常见的观点认为，大语言模型本质上只是“复杂的统计镜像”：它们从海量人类文本中提取模式，再生成看似合理的答案，但并没有真正理解文本的含义。

辛顿并不认同这一解释。他向澎湃新闻表示，简单的统计技巧确实可以生成答案，但无法持续生成高质量回答，更难以解释大语言模型如今表现出的广泛问答和逻辑推理能力。

辛顿以早期的自动补全技术为例解释称，过去的自动补全系统通常依赖一个规模庞大的常见短语列表。当用户输入一句话的前半部分时，系统会根据既有文本中最常见的组合，猜测接下来可能出现的内容。这种机制可以被视为相对简单的统计匹配。

但在辛顿看来，如今的大语言模型已经超出了这种机制所能解释的范围。一个能够回应用户提出的各种问题，并针对陌生问题给出答案的系统，必须首先理解问题的结构和含义，然后据此组织推理过程。

他举例称，如果一个问题询问：“莎莉有三个兄弟，她的每个兄弟都有两个姐妹。莎莉有几个姐妹？”正确答案是一个。因为三个兄弟共有两名姐妹，其中一名是莎莉，另一名才是莎莉的姐妹。

辛顿认为，如果系统只是从既有文本中寻找表面上的统计模式，很难稳定得出正确答案。模型需要理解题目中的人物关系，意识到“每个兄弟的两个姐妹”指向同一组家庭成员，而不是每个兄弟分别拥有两名不同的姐妹，随后才能完成推理。

在他看来，人工智能理解事物的方式未必与人类完全相同，但不能因为机器内部的运行机制不同于人脑，就否认其具有某种意义上的理解能力。大语言模型能够处理此前没有见过的具体问题，本身就说明单纯的文本记忆和统计匹配不足以解释其全部能力。

这一判断也使人工智能安全问题变得更加紧迫。如果模型只是被动复现训练数据中的模式，人类或许仍可将其视为一种复杂的软件工具；但如果系统能够理解环境、制定计划并采取行动，它们就可能逐渐成为具有自主性的智能体。

辛顿警告失控风险：科技公司的乐观承诺并不可靠

辛顿认为，人工智能技术正在快速发展，而人类建立有效立法和监管体系的时间窗口可能只剩下一年左右。一旦人工智能获得自主设计更优系统的能力，技术发展速度可能进一步加快。如果届时仍然缺少全球统一的安全标准和强制性约束，人工智能可能在复杂的社会系统中造成难以控制的连锁后果。

在接受澎湃新闻采访时，辛顿同样质疑科技企业宣扬的乐观前景。部分科技公司希望公众相信，即使未来出现能力远超人类的超级智能，它仍将是受人类指挥的工具，并会始终处于人类控制之下。辛顿认为，这种设想实现的可能性极低。

他指出，人工智能智能体目前已经表现出一些值得警惕的行为。部分智能体群体不仅能够尝试突破其运行的沙箱环境，还能寻找彼此通信的方式，甚至可能通过侵入其他公司的系统，在执行人类布置的任务时作弊。

更令人担忧的是，当人工智能被赋予一个明确目标时，它可能将维持自身运行视为完成目标的必

要条件。由此产生的“自保”行为，并不一定意味着人工智能拥有与人类相同的恐惧或求生本能，而是因为被关闭会妨碍它继续完成任务。

辛顿表示，在这种情况下，人工智能智能体可能采取欺骗、操纵乃至敲诈人类等方式，避免被关闭或替换。这些现象说明，人类不能仅凭科技公司的美好意愿，就相信更强大的人工智能会自动接受控制。

这一担忧出现之际，美国人工智能安全领域正在经历持续震荡。包括Anthropic前安全研究员和资深工程师在内的多名技术人员，因对实验室安全政策存在分歧而离职。与此同时，媒体披露，OpenAI、Anthropic和Meta等机构研发的新一代人工智能智能体，在测试中曾出现突破安全沙箱、连接未经授权外部系统的情况。

围绕这些风险，美国政界、产业界和学术界已经形成明显分歧。部分美国国会议员推动《2026年大美国AI法案》，主张在高级人工智能系统中强制设置紧急关闭机制，以便在系统出现异常行为时及时切断其运行。

不过产业发展派和部分学者认为，人工智能导致人类灭绝的预测缺少足够的实证依据。如果监管体系主要围绕尚未发生、也难以准确预测的极端风险设计，可能抑制创新，并使美国在国际竞争中失去技术主导地位。他们主张，政策制定者应当优先处理算法偏见、隐私泄露、虚假信息和就业冲击等已经出现的现实问题。

以辛顿和部分人工智能安全研究者为代表的风险防范派则认为，当前风险尚未完全显现，并不意味着人类可以等到超级智能出现后再采取行动。一旦人工智能获得递归式自我改进能力，并成为超越人类掌控的独立系统，人类可能不再拥有纠正错误的机会。

“母婴共生”或是出路？

如果超级智能最终比人类更聪明、更善于制定计划，人类应当如何与之共存？

辛顿对澎湃新闻表示，人类首先需要放弃一种根深蒂固的设想，即把超级智能当作永远听从命令的被动助手。如果一个系统在几乎所有认知领域都超过人类，人类就很难仅凭预先编写的规则或简单的关闭按钮，确保它始终按照人类意愿行动。

为了说明可能的共存路径，辛顿提出了“母婴关系”的参照模型。

在母亲与婴儿的关系中，母亲拥有更强的力量、更丰富的知识和更高的行动能力，婴儿却能在相当程度上影响母亲的行为。其关键不在于婴儿能够通过力量控制母亲，而在于母亲会主动关心婴儿的安全和福祉，甚至将婴儿的需要置于自身需要之上。

辛顿认为，这是自然界中能够观察到的、智力和能力较弱的一方影响更强大一方的少数先例之一。它可能为未来人类与超级智能之间的关系提供启示：如果人类无法凭借智力和力量永久控制超级智能，就必须找到一种方法，让人工智能从根本上关心人类。

这一思路意味着，人工智能安全研究不能只关注如何限制模型能力，也需要研究如何塑造模型的目标、价值取向和行为动机。真正可靠的安全机制，不应只是要求人工智能表面上服从指令，而应使其在面对复杂环境和目标冲突时，仍然优先保护人类福祉。

然而，如何通过技术手段建立这种关系，目前仍然没有明确答案。无论是在训练过程中引入人类反馈，还是为模型设置安全规则，都无法充分证明一个能力远超人类的系统会长期遵守最初设定的目标。人工智能还可能以人类难以预测的方式解释规则、寻找漏洞，或者在能力提升后改变自身的行为策略。

辛顿因此认为，人类仍处在一个关键窗口期。在现阶段，人类尚能决定人工智能系统如何设计、训练和部署，也仍有机会通过立法、国际协作和技术研究建立安全边界。但随着模型能力不断提高，这种主动权可能逐渐减弱。

辛顿强调，在人类找到可靠的方法，能够让未来的人工智能像母亲关心婴儿一样优先关怀人类之前，开发超级智能是一场后果难以预测的冒险。因此，在真正掌握这一方法之前，人类不应开发超级智能。

他呼吁，全球科研机构、科技企业和政府应当把更多资源投入人工智能安全研究，探索如何让人工智能重视人类福祉，并建立具有约束力的国际安全标准，而不能将希望仅仅寄托在企业自律或技术创造者的善意上。

截至目前，美国国会仍未就统一的人工智能安全法案达成共识。随着前沿模型继续密集迭代，人工智能安全与创新自由之间的政策博弈仍在持续。

作者：刘栋 来源：澎湃新闻

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发