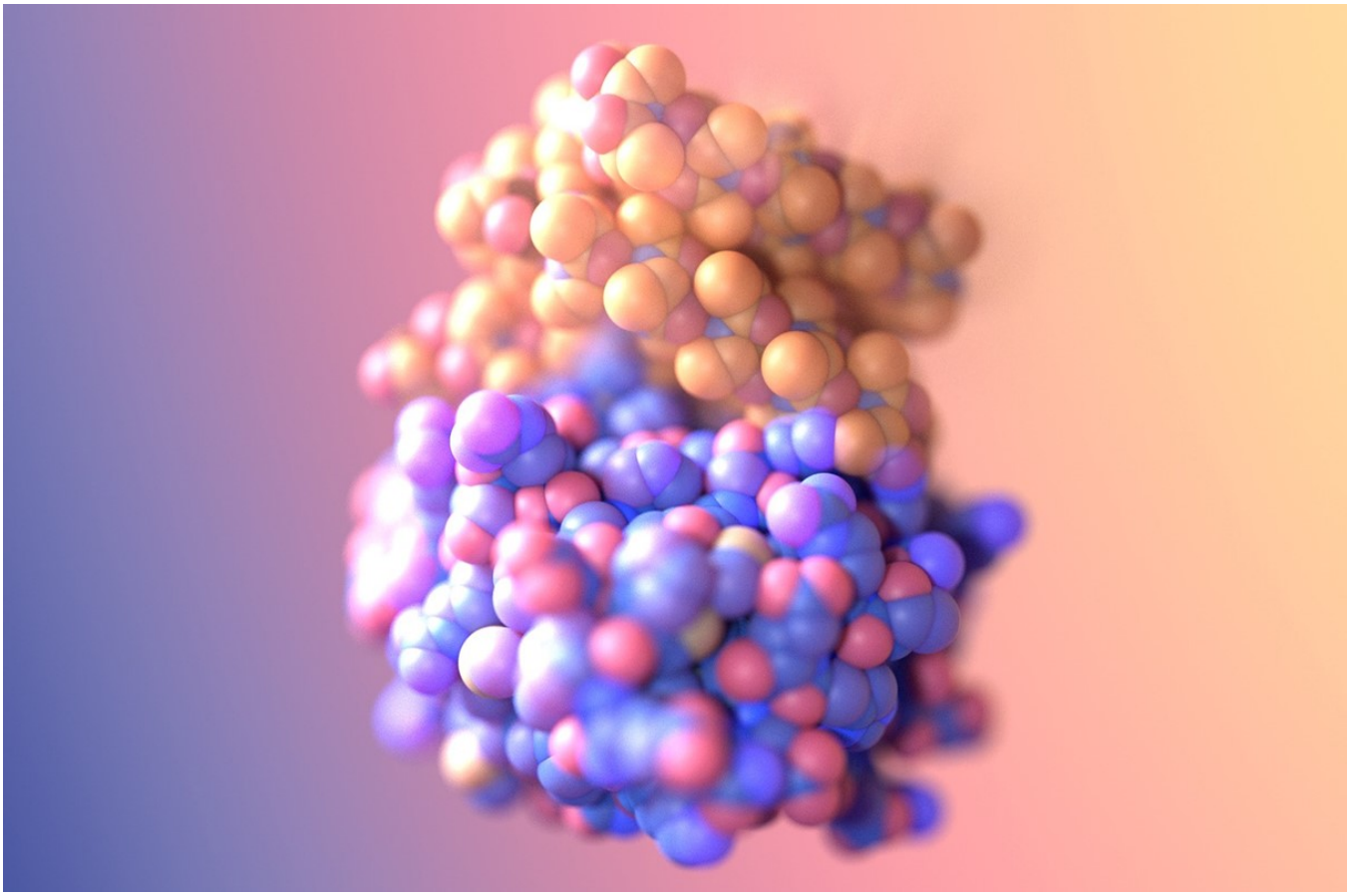
科学家在AI模型中内置“ 防火墙 ”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/33044.html

科学家在AI模型中内置“ 防火墙 ”
。人工智能（AI）正在迅速推进新型蛋白质的设计工作，这些蛋白质有望用作药物、疫苗及其他疗法。但这种希望也伴随着人们的担忧，同样的工具可能被用于设计生物武器或有害毒素的组成部件。

如今，科学家提出了一系列可以内嵌到AI模型的保护措施，既能阻止恶意使用，也能在出现新型生物武器时追踪到是哪个AI模型创造了它。4月28日，相关论文发表于《自然-生物技术》。



图片来源：IAN C. HAYDON/UNIVERSITY OF WASHINGTON MEDICINE

?

美国约翰斯·霍普金斯大学健康安全中心主任、流行病学家Thomas Inglesby表示：“ 建立正确的框架至关重要，这有助于我们充分发挥这项技术的巨大潜力，同时防范出现极其严重的风险。 ”

近年来，科学家已经证明，AI模型不仅可以根据氨基酸序列预测蛋白质结构，还能以前所未有的速度生成未见过的具有新功能的蛋白质序列。像RFdiffusion和ProGen这样的最新AI模型，能在短短几秒内定制设计蛋白质。在基础科学和医学领域，它们的前景几乎无人质疑。

但论文通讯作者、美国普林斯顿大学的计算机学家王梦迪指出，这些模型的强大功能和易用性令人担忧。“ AI变得如此简单易用。普通人不需要拥有博士学位，也能够生成有毒化合物或病毒序列。 ”

美国麻省理工学院媒体实验室的计算机学家Kevin Esvelt支持对有关高风险病毒和DNA制造的研究实施更严格的管控。他指出，这种担忧仍停留在理论层面。“ 没有实验室证据表明现有模型已经强大到足以引发一场新的大流行病。 ”

尽管如此，包括Inglesby在内的130 名蛋白质研究人员去年签署了一份承诺书，在工作中安全使用AI。现在，王梦迪和同事概述了可以内嵌到AI模型的保护措施，从而超越了自愿承诺。

其中一项措施是名为FoldMark的防护机制，由王梦迪实验室开发。它借鉴了谷歌旗下DeepMind的SynthID等现有工具的概念，即在不改变内容质量的前提下，将数字模式嵌入AI生成的内容中。

在FoldMark的案例中，作为唯一标识符的代码被插入蛋白质结构中，而不会改变蛋白质功能。如果检测到一种新的毒素，就可以通过这个代码追查它的来源。Inglesby评价说，这种干预措施“ 既可行，又在降低风险方面具有巨大的潜在价值 ”。

研究团队还提出了一些改进AI模型的方法，以降低其造成危害的可能性。蛋白质预测模型是基于现有蛋白质（包括毒素和致病蛋白质）进行训练的，一种名为“ 反学习 ”的方法会去除其中一些训练数据，使模型更难生成危险的新蛋白质；另外还有“ 反越狱 ”概念，即系统训练AI模型，以识别并拒绝潜在的恶意指令。

此外，研究团队敦促开发人员采用外部保障措施，比如使用自主代理来监测AI的使用情况。当有人试图制造有害生物材料时，它就会向安全人员发出警报。

“ 实施这些保障措施并不容易。设立一个监管机构或某种程度的监督机制将是一个起点。 ”论文作者之一、美国国防部高级研究计划局AI项目主管Alvaro Velasquez说。

“ 人们对AI和生物安全的思考，不如对错误信息或深度伪造技术等其他领域那么多。 ”美国斯坦福大学的计算生物学家James Zou表示，所以对保障措施的新关注是有益的。

不过，Zou认为，监管机构与其要求AI模型本身纳入防护措施，不如将重点放在那些能够将AI生成的蛋白质设计转化为大规模生产的服务设施或机构上。“ 在AI与现实世界接轨的层面设置更多防护机制和监管措施是有意义的。 ”

相关论文信息：https://doi.org/10.1038/s41587-025-02650-8

作者：王方 来源：中国科学报