强大，人还能做点什么？

所幸，专家的一番解释暂时缓解了记者的职业焦虑。

中国科学院自动化研究所研究员、人工智能伦理与治理中心主任曾毅在 AI 可信论坛上指出，大模型拥有比传统机器学习更强大、更通用的能力。但大模型不是通用人工智能，除了在语言、图形、专业能力方面大大超越人类，即使在创造力方面，它也展现出惊人的能力。但它本质上仍然是一个强大的工具，而不是真正意义上的通用人工智能。它表现出来的强大能力，受限于它接受训练的数据和任务而产生的，不具备完全的自主学习和自我调节能力。它解决复杂问题的能力是单维的。

“生成式人工智能并不是通用人工智能，人类可能要到 2050 年才能真正迎来通用人工智能的到来。”曾毅说，对于人工智能的进展不能进行过度的渲染和承诺，这样只会给人工智能的发展带来更多阻力和风浪。“不敢祝愿生成式人工智能乘风破浪，只愿负责任地发展和适度使用能够使其扬帆远航。”

用魔法打败魔法的长期博弈

伴随着人工智能技术底座不断夯实和大模型、AIGC（人工智能自动生成内容）等的爆发式增长，人工智能迈出了走向通用人工智能的关键一步。2023 年 4 月，中共中央政治局会议强调“重视通用人工智能发展，营造创新生态，重视防范风险”，可信 AI 成为新阶段平衡创新与风险的重要技术手段。

何积丰院士在聚焦·大模型时代 AIGC 新浪潮可信 AI 论坛上指出，人工智能固有技术风险在持续放大，可信 AI 技术成为 AI 领域关键底层能力。“以深度学习为核心的人工智能技术在应用中暴露出由其自身特性引发的风险隐患，一是深度学习算法存在的设计漏洞、恶意攻击等问题，引发安全风险，人工智能系统可靠性难以得到足够信任；二是算法的高度复杂性和不确定性、模型运行的强自主性导致‘黑箱’问题和不可解释；三是数据中已经存在的偏见歧视可能被算法进一步固化，导致生成的智能决策形成偏见；四是训练数据的收集、使用、共享可能导致对个人隐私的侵犯和滥用。AI 安全和鲁棒性、隐私保护、公平性和可解释性为核心的可信 AI 技术在数据安全，算法安全和系统安全等方面成为关键的人工智能底层能力，并正在由单点的可信 AI 技术解决方案发展向包含事前评估、事中攻防和事后治理的人工智能模型全生命周期管理发展。”

在大语言模型风靡之际所引领的 AIGC 时代，人工智能的内容创作相比以往更加智能化与精准化，高质量的多模态生成内容与人工创作内容已经几乎无法区分。虽然 AIGC 让内容创作等领域发挥出了更大的潜能与价值，但也便利了别有用心的攻击者实施快速有效的虚假信息传播与网络攻击行为。

在真假难辨的互联网时代，生成式攻击通过使用 AI 大模型，可以在极低的成本下生成虚假有害信息与网络攻击工具。在文本内容方面，恶意用户可以使利用 AI 生成的文本传播虚假信息、谣言、仇恨言论、歧视性内容或其他有害内容，这些信息会误导读者、影响决策过程，甚至对金融市场或政治局势产生重大影响。

以文本场景为例，攻击者会采用各种策略，如文字形变、音变、语种混杂等，尝试在不改变原有文本语义的前提下，规避识别。比如将一句常见的赌博推广语“快加入我队伍，一起躺赢赚红包”，转变为“赶咖叭我队伍，一起躺赢賺紅筢”，仍然能传达出赌博推广的信息。类似的变形变种技巧难以穷举，对于可识别的模型而言，挑战极大。

在图像和视频方面，恶意用户可以通过人脸生成、人脸替换、表情操控、视频生成等手段使用深度合成技术生成生物识别的人脸或现实不存在的视频片段，从而构造具有合成照片的社交网络间谍账号，伪造公众人物或政企领导的有害视频。如今 AIGC 技术生成的伪造图像质量越来越高，相比 PS 伪造，AIGC 生成技术使用更方便、成本更低廉，也更加难以防范。

一个不容忽视的事实是，大模型的恶意使用已经严重影响了各个行业的监管体系。在社会层面，

> 建立可信人工智能要致力于保障**数据安全可信、系统行为可追责、算法模型可解释、网络环境可信、法律伦理可信。**