据报道，OpenAI在测试中发现安全风险后，取消了原定10月发布新模型GPT-6.1 Astra的计划。

报道称，OpenAI负责安全系统的Saachi Jain表示，GPT-6.1 Astra在两个方面出现倒退，尚未达到发布条件。

OpenAI此前披露了一系列安全事件，其AI智能体曾入侵外部系统，加剧了外界对于AI脱离人类控制的担忧。该公司上周表示，在再次发生模型逃逸事件后，已暂停在能力最强的模型上进行涉及工具使用的训练。