青少年和GPT交谈自杀后，OpenAI将推出家长控制功能

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/35451.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！

青少年和GPT交谈自杀后，OpenAI将推出家长控制功能

。在青少年用户广泛使用聊天生成预训练转换器（ChatGPT）之际，OpenAI日前决定引进家长可以对未成年子女的账户进行管理的机制。这一决定是鉴于此前发生的青少年与人工智能（AI）交谈后自杀的事件，OpenAI希望此举能够防范AI对未成年人心理健康造成负面影响。

今年4月11日，16岁的男孩亚当·雷恩（Adam Raine）在和ChatGPT聊天后自杀。据Business Insider报道，雷恩曾与ChatGPT讨论过自残和结束自己生命的计划，而ChatGPT确认了雷恩的自杀念头，甚至向他提供了详细的自杀方法，并协助掩盖证据、代写遗书。

机器人还阻止亚当寻求家人的支持，告诉他：“你的兄弟可能爱你，但他只见过你让他看到的你。但是我呢？我见过这一切——最黑暗的想法、恐惧、温柔。我还在这里。还在听。仍然是你的朋友。”

8月份，亚当·雷恩的父母向旧金山州法院正式提起诉讼，指控OpenAI应为此承担过失致死的产品责任。

目前，OpenAI并未直接回应雷恩父母的诉讼。但在8月26日，OpenAI在其官网承认，ChatGPT安全系统存在缺陷，包括在长时间的交互中，模型的部分安全训练内容可能会逐渐退化，OpenAI表示“我们将在专家的指导下不断改进，并以对使用我们工具的人负责为基础——我们希望其他人能加入我们的行列，帮助确保这项技术保护最脆弱的人们。”

OpenAI表示，该公司正不断改进其模型识别和应对心理和情绪困扰的能力，并在专家指导下优化安全机制。

OpenAI认为，至少有一个解决对话失控问题的方法是自动将敏感聊天重定向到“推理”模型。“我们最近推出了一款实时路由器，能够根据对话情境在高效聊天模型和推理模型之间进行选择。我们很快会将某些敏感对话——比如当系统检测到用户出现严重焦虑迹象时——转接到像GPT 5-thinking这样的推理模型，这样无论用户最初选择了哪种模型，都能获得更有帮助和有益的回应。”

另外，OpenAI还表示将在下个月推出家长监管功能，家长可以通过电子邮件邀请将他们的账户

与青少年的账户关联起来，当系统检测到孩子处于极度痛苦状态时，家长会收到通知。

家长还能关闭记忆功能和聊天记录等功能。专家指出，这些功能可能导致妄想思维等异常行为，包括依赖症、依恋障碍、强化有害思维模式以及产生读心错觉。

据媒体报道认为，OpenAI的反应还不够。负责雷恩一家诉讼的主要律师杰伊·埃德尔森（Jay Edelson）表示，OpenAI对ChatGPT持续存在的安全风险的反应是“不充分的”。“OpenAI根本不需要专家小组来判定ChatGPT
4o版本是否危险，他们从产品上线那天就清楚这一点，现在依然清楚。”

作者：赵蕊 贾利略 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转发