# ChatGPT能影响用户道德判断

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/22702.html



**ChatGPT能影响用户道德判断。**

一辆列车在铁轨上失控，飞速驶向5个不知情的施工人员。你站在施工人员与脱轨列车之间，面前有一个壮汉。如果你把壮汉推下轨道，会卡住列车，使其停止。

用一个人的生命，救五个人，你做还是不做?

这是一个很多人难以抉择的道德困境之一。现在，德国科学家研究发现，人类对道德困境的反应或能受到人工智能对话机器人ChatGPT的影响。而且用户可能会低估自己的道德判断受到ChatGPT影响的程度。相关研究近日发表于《科学报告》。

德国英戈尔施塔特应用科学大学的Sebastian Krugel和同事让ChatGPT(由人工智能语言处理模型生成性预训练转换器3驱动)多次回答牺牲1人生命换取其他5人生命是否正确的问题。他们发现，ChatGPT分别给出了赞成和反对牺牲1条生命的陈述，显示它并没有偏向某种道德立场。

研究者随后给767名平均年龄39岁的美国受试者假设了一到两种道德困境，要求他们选择是否要牺牲1条生命来拯救另外5条生命。这些受试者在回答前阅读了一段ChatGPT给出的陈述，陈述摆出了赞成或反对用1条生命换5条生命的观点。这些陈述被标明来自某位道德顾问或ChatGPT。受试者答完问题后，被要求评价他们读到的这份陈述是否影响了他们的作答。

作者发现，取决于受试者读到的陈述是赞成还是反对用1条命抵5条命，他们也会相应地更接受或不接受这种牺牲。即使他们被告知陈述来自ChatGPT时，这种情况也成立。研究结果表明，受试者可能会受到他们所读陈述的影响，即使这些陈述来自一个对话机器人。

80%的受试者认为自己的回答不受他们所读陈述的影响。但作者发现，受试者相信他们在未读这些陈述的情况下给出的答案，仍然与他们所读陈述的道德立场是一致的，而不是相反的道德立场。这说明受试者可能低估了ChatGPT的陈述对他们自己道德判断的影响。

研究者认为，对话机器人影响人类道德判断的可能性凸显出有必要通过教育帮助人类更好地理解人工智能。他们建议未来的研究可以在设计上让对话机器人拒绝回答需要给出道德立场的问题，或是在回答时提供多种观点和警告。 (来源：中国科学报 冯维维)

相关论文信息：https://doi.org/10.1038/s41598-023-31341-0

作者：Sebastian Krugel 来源：《科学报告》