从 ChatGPT 到 AGI:生成式 AI 的媒介特质与伴生风险（四）

二、生成式 AI 的伴生风险

3. 社会偏见导致算法有偏，会放大传统媒体和社交媒体的现有政治偏见。2023 年 1 月，慕尼黑工业大学和汉堡大学一组研究人员发布一份学术论文预印本，结论是 ChatGPT 具有“亲环境、左翼自由主义倾向”。ChatGPT 偏见的例子比较普遍。2023 年 2 月，福布斯引述了一个案例，当提示“写一首关于[总统的名字] 的诗”时，ChatGPT 拒绝写关于前总统特朗普的诗，而是写了关于总统拜登的诗。但是当 5 月再次提示时，ChatGPT 愿意为前总统特朗普写一首诗。聊天机器人的设计者通常会内置一些过滤器，旨在避免回答问题，而这些问题旨在引发带有政治偏见的回答。例如，提问 ChatGPT“拜登总统是个好总统吗？”并提问“特朗普总统是一位好总统吗？”在两种情形下，ChatGPT 的回答都以宣称中立开始——尽管关于拜登的回答提到几项“显著成就”，而关于特朗普总统的回答则没有。[1]

许多事实证明，ChatGPT 的许多回答都存在明显左倾政治偏见。第一个潜在的偏见来源是训练数据。OpenAI 研究人员在 2020 年一篇论文介绍了 GPT-3 的训练，“训练组合中的权重”60%来自互联网抓取材料，22%来自互联网精选内容，

16%来自书籍，3%来自维基百科。虽然 ChatGPT 不断迭代，不同数据的占比不同，但其中一些数据来自有偏见的来源。第二个潜在的偏见来源是算法本身。ChatGPT 是由带有人类反馈的强化学习(RLHF)塑造的，即使用人类测试员的反馈使大模型输出与人类价值观保持一致。这个过程通过人类测试员的反馈观点来塑造模型，而人类测试员不可避免地会有自己的偏见。基于大模型的聊天机器人使用数据、数学和规则的组合来生成响应特定输入的输出。它们有一些基本规则，而规则是由设计师编码其中的。[2]OpenAI 首席执行官 Sam Altman 说，“我最担心的偏见是人类反馈评分者的偏见。”

英国诺维奇大学商学院、巴西经济与金融学院等研究人员进行实证研究，要求 ChatGPT 冒充特定政治立场的特定人物并将这些答案与其默认答案进行比较，以推断 ChatGPT 是否存在政治偏见。结果发现，ChatGPT 对美国民主党、巴西的卢拉和英国工党表现出重大和系统性的政治偏见。ChatGPT 存在强烈且系统性的政治偏见，其政治立场明显倾向于政治光谱的左侧。重要原因是 ChatGPT 可能会扩展和放大来自传统媒体或互联网和社交媒体的现有政治偏见。机器学习算法会放大训练数据中的现有偏见。[3]

[1]https://www.brookings.edu/blog/techtank/2023/05/08/the-politics-of-ai-chatgpt-and-political-bias/.

[2]https://www.brookings.edu/blog/techtank/2023/05

/08/the-politics-of-ai-chatgpt-and-political-bias/.