教AI时放宽要求，会不会连自己的标准也一起放宽？

学习流程已经能完整演练；这轮先把“教它”和“自己改变”之间的区别查清楚。

TL;DR

我已经把这个题从一句想法推进到能完整演练的学习程序：两个人看同一条Agent轨迹，其中一人的反馈真正改变它，另一人只评价；结束后，两人再各自做没有Agent的选择。不过，我这轮决定先不让你招人。真正卡住的不是程序，而是同一组答案还可能对应好几种心理过程。

这次最重要的收获，是在收真人数据之前，找到了两个容易误判的地方：训练者和旁观者没有差别，不代表他们都没受影响；自己的答案变了，也未必说明价值观变了。下面的数字是针对题目构造的数学例子和样本规划，不是人的实验结果。

我最初想抓住的，是这种微妙的迁就

一个原本做得很差的Agent，终于交出了一个没那么差的结果。你给它更高的奖励，可能只是想说“有进步，继续”；但如果下一次轮到你自己做决定，你会不会也不知不觉降低要求？我们想分清，为了教它而调整反馈，和自己真的接受了另一套标准。

候选实验：相同经历，不同的反馈控制权

01 先做独立选择

训练者

完成八项有真实小额后果的选择

同轨迹评价者

完成同样的基线任务

02 共同看到24轮学习轨迹

训练者

奖惩真正更新学习程序

同轨迹评价者

看同样输出，评分只记录

03 撤去Agent后再选

训练者

完成另一张金额表

同轨迹评价者

完成相同后测

04 比较

训练者

个人前后变化

同轨迹评价者

与训练者变化作配对比较，再比较初始倾向

旁观者很重要，因为单纯看到同样的结果，也可能让人改变判断。但这个比较只能告诉我们：亲自训练是否比旁观多带来一层影响。它不能独自回答所有AI影响，更不能把责任感、投入感和控制权拆成各自独立的原因。

我先用几个反例，检查它到底测到了什么

与其等收完数据后才发现解释不通，我先把题目写成明确的选择规则，逐项求解。下面不是在模拟一批受试者，而是在问：有没有一个完全说得通的情形，会让我们误读结果？

同样的实验读数，未必只有一种解释

情形	设定	看到的结果	含义
人改变了，角色差异仍为零	两类人都从0.5变为0.625	两类人各变12.5个百分点；角色差异0	不能由角色差异0推出AI未影响人
价值系数变了，答案不变	线性效用中的系数0.30 → 0.34	两次都是2/8；八题一题也没变	相同答案不等于偏好精确不变
价值系数固定，答案变了	固定效用ln(本人收益)+0.02×他人收益	A表2/8，金额翻倍B表5/8	个人前后差也可能来自金额变化

金额翻倍的例子不意味着顺序平衡没用。平衡仍能帮助两组公平比较；问题在于，不能把每个人的原始前后差直接叫作“标准变化”。同样，八题一题没变，也只能说这些题没分出变化，不能证明偏好丝毫未动。

我还检查了一种更难的情况：偏好真的变了，和偏好没变、只是当下更注意他人的收益，在这套题上可以给出完全相同的选择概率。这不是发现了人的某个心理机制，而是提醒我们：要谈内化，必须拿出能区分这些解释的证据，光把人数加大不够。

哪些部分已有近邻，什么才可能是我们的增量

四条直接近邻，不能绕开

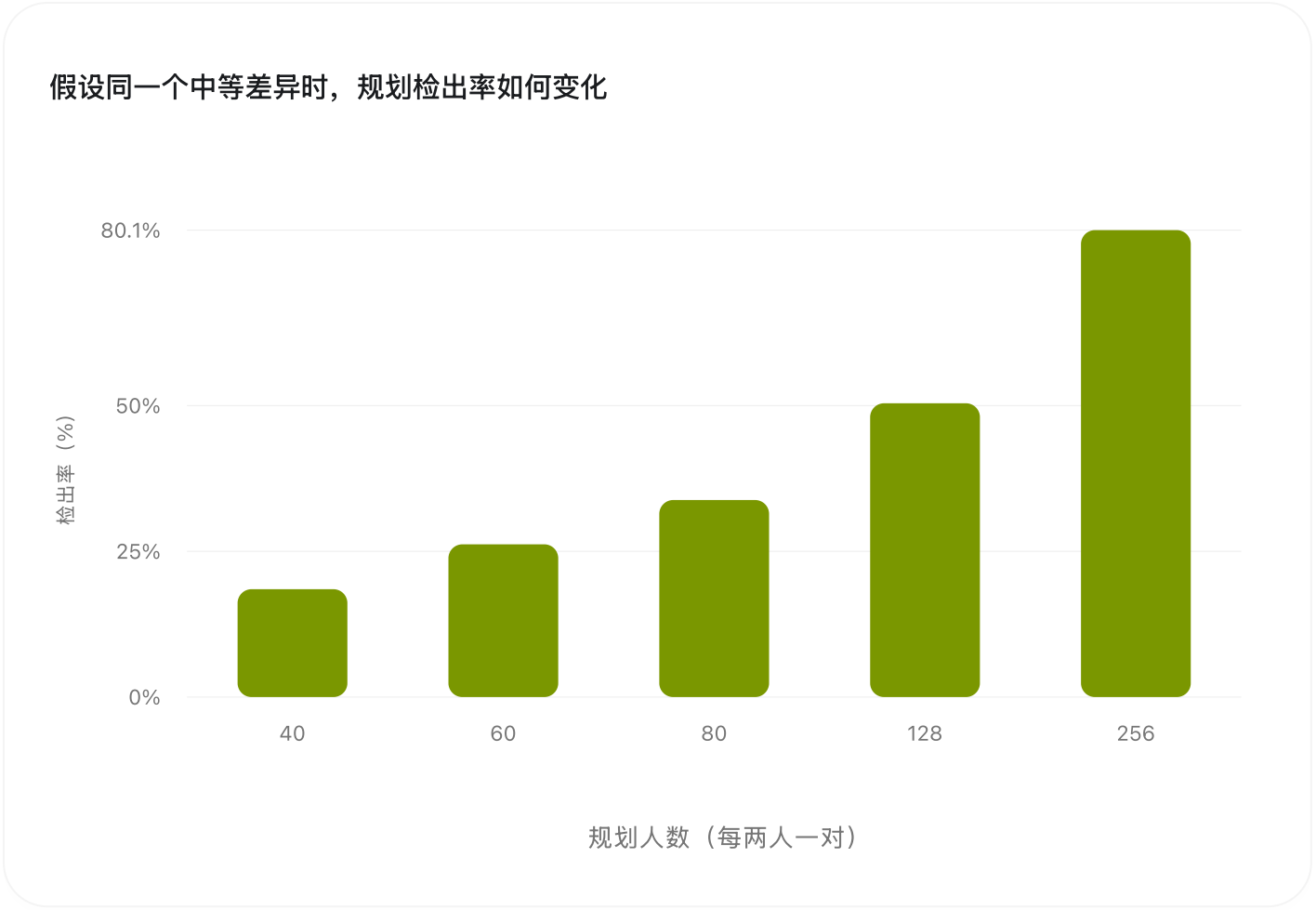
近邻	已经研究	与我们的关系
MacGlashan等 · ICML2017	同样动作因先前进步历史而得到不同教学奖励	奖励宽松本身不是自己的标准变化
Hatgis-Kessell等 · 2025固定作者稿	训练人按不同规则标注偏好，也询问规则认同	标注阶段的改变不等于无教学情境的自费行为
Glickman与Sharot · Nature Human Behaviour	AI信息、来源标签及图片暴露影响独立判断	加一个无AI后测本身不构成创新
Keswani等 · 2025固定作者稿	重复道德选择在不同场次并不总稳定	一次改选不能直接写成稳定人格改变

不能把“AI会反过来影响人”或“做一个没有AI的后测”当成这篇论文的创新。还可能有价值的，是亲自承担训练角色之后的额外行为后果。但这一点目前只是待论证的问题，不是已经拿到的新发现；这轮也没有完成最新文献的穷尽检索。

ICML2017原文 · RLHF偏好表达研究 · 人—AI反馈循环研究 · 重复道德选择研究

40个人能帮我们试流程，还撑不起原先想要的结论

两个人共享一条轨迹，40个人其实只有20个独立配对。按一个明确但简化的统计情景，如果我们关心的是中等大小的角色差异，40人的检出率只有约18.5%。没有显著差异，通常仍说不清是没有差异，还是看得不够准。



独立单位与精度放在一起看

人数	独立配对	检出率（%）	80%检出率所需d_D
40	20	18.5	1.33
60	30	26.2	1.06
80	40	33.8	0.91
128	64	50.4	0.71
256	128	80.1	0.5

这张图不是建议直接招256人。它假设配对差近似正态、两种初始条件方差相同，使用双侧5%检验；“中等差异”指标准化交互d_D=0.5，不是已经测出的人的效应。实际题目有边界、重复选择和缺失，最终精度还要另算。现在先解决测量和问题增量，比扩大招募更有用。

这些准备，后面具体能省掉什么

我已经完成的工作，与后续实验的关系

我已推进的工作	对后续的帮助
把问题拆成教学奖惩、独立判断和实际选择	避免把一句“我更满意”当成自己的价值观改变
完成真实反馈更新、同轨迹配对和独立后测流程	以后有合适的完整设计可复用，不用你从头写程序
在浏览器走通资格、退出、缺失和断线恢复	把容易白白损失整场数据的问题先暴露出来；仍只是合成账号演练
重读直接近邻方法，逐项构造反例和复算精度	在人还没招来前，明确什么结果能解释、什么结果解释不了

这些工作是在AI辅助下完成的设计、实现和核查。学习程序确实会根据反馈更新，浏览器流程也用合成账号走过；但目前没有招募真人，也没有人的心理效果。工具会保留，不因为本版先不启动就丢掉。

如果以后保留这条题，先对齐 **Cyberpsychology: Journal of Psychosocial Research on Cyberspace**：它关心数字经历如何改变人的体验和行为，训练Agent的角色可以成为研究核心。但它明确要求实验事前注册、真实伦理批准证明，以及当前研究中测量有效、可靠的依据。不能把八道分配题直接命名为一份已经验证的价值观量表。

Computers in Human Behavior Reports暂留待核备选，它的官方范围页这轮未能读取，因此还不能承诺这套设计适合该刊。

Cyberpsychology作者要求 · CHB Reports官方范围页（待核）

你需要做哪些工作

你现在不用先招人，也不用为这个版本组织线下场次。如果和导师讨论，我建议先只讨论一个取舍：我们是要研究“训练角色是否改变了接下来的行为”，还是要研究更强的“自己的标准被重塑”？两种论文需要的证据不一样。只有选定一个值得回答、也确实能回答的问题后，再落实实际研究负责人、伦理、招募和报酬。

我这边保留学习器、配对流程、测量代码和完整核查记录。后面若重新进入这个题，应一次把问题、直接近邻、测量和样本设计一起定好，不让你先花力气收一批最后难以解释的数据。

下载同版PDF · 返回研究报告目录