信任很高，也不等于愿意完全放手

一次公开数据复算，以及它对我们的真人实验意味着什么

TL;DR

有一种回答很值得留意：一个人说自己很信任 AI，下一步却仍然想亲自核验，不愿完全把事情交出去。我把一篇相关研究的公开数据和分析代码拿来重新算了一遍，想弄清楚：这只是少数人的习惯，还是很多人都会这样？

56人、392轮回答能从原始记录完整重建，论文的主要计数对上了。高信任的217轮里，172轮仍选择保留监督，占79.3%。再按人看，46人曾给出高信任评分，其中44人至少一次仍要核验或拒绝完全委托，38人至少两次如此。

这帮我们确认了一件具体的事：以后做真人实验，问“你信不信它”和记录“你实际交给它什么”，应该分开。原论文研究的是未来委托意向，不是已经执行的授权；这次也没有新增被试。我补的是同一公开样本的参与者层面描述，不是发现了一个新的心理效应。

主要模型的点估计接近论文，但部分标准误没有完全对齐，敏感性分析也没全部复现。下面把结果、我具体做了什么，以及它能替后续实验省下哪些弯路放在一起。

原来的真人实验方案与近邻比较

最后要看：人会不会真的向别人开口

我们原来想写的真人研究，是让参加者先和一个 AI 小组共同处理日常学习或工作中的卡顿，再看他是否会主动开始一次真人交流。不是只问他“喜欢这个小组吗”“更有归属感了吗”，也不是让 AI 猜他有没有求助。

保留的真人研究问题与测量条件

环节	设计
两种AI小组	均为八名成员；一组真实互相复核，另一组各自自检。成员数和机会相同。
人的经历	两次活动后，所有人得到同样、可兑现的真人学习／工作交流机会。
主要观察	独立记录是否真正发起并开始交流，不由AI推断，不只看信任问卷或预约点击。
当前状态	既有实现路线未通过材料验证，仍关闭；尚无真人招募、效果或新的实施授权。

这次复算没有把此前未通过验证的社区实现重新打开，也没有开始招人。它给原方案补了一层更直接的依据：就算信任评分变高，也不代表实际求助或实际授权已经改变。信任、意向、行动，要各自收证据。

如果你和导师更想转向“用户愿意交给 Agent 多大权限”，同样需要让权限真的生效，而且能撤回、能记录；不能宣布还有几十题和奖金，实际却不做，再把最后一次选择当成真实委托。这是后续设计要守住的测量要求，还不是一个已确定有新意的新课题。

这篇近邻究竟让人做了什么

Chiatti 等让参加者完成七轮视频理解任务，查看 AI 提供的回答，并报告信任程度和今后愿意怎样使用 AI：拒绝辅助、核验后使用，或完全委托。研究也比较了有无图界面等条件，AI 回答事先生成。参加者以熟悉 AI 的工程学生为主，不能直接代表普通用户或心理求助人群。

原始档案有58次会话、406轮回答，因为其中一个人做了三次。沿用作者“每人只留首次会话”的规则，得到56人、每人7轮，共392轮；选定会话没有因关键评分缺失再删回答。

从发布记录到最终分析的分母

步骤	数量
公开会话	58份
逐题回答	406轮
重复会话	排除2份，按每人首次会话
进入分析	56人，每人7轮，共392轮
选定会话的关键评分缺失	0轮

我对照了原始记录、压缩包中的分析表，以及固定代码仓库中的分析表。两个表的文件字节不同，但392个记录和所有模型字段在数值解析后都一致。这里没有把不同版本混到一起，也没有为了对上论文数字另挑样本。

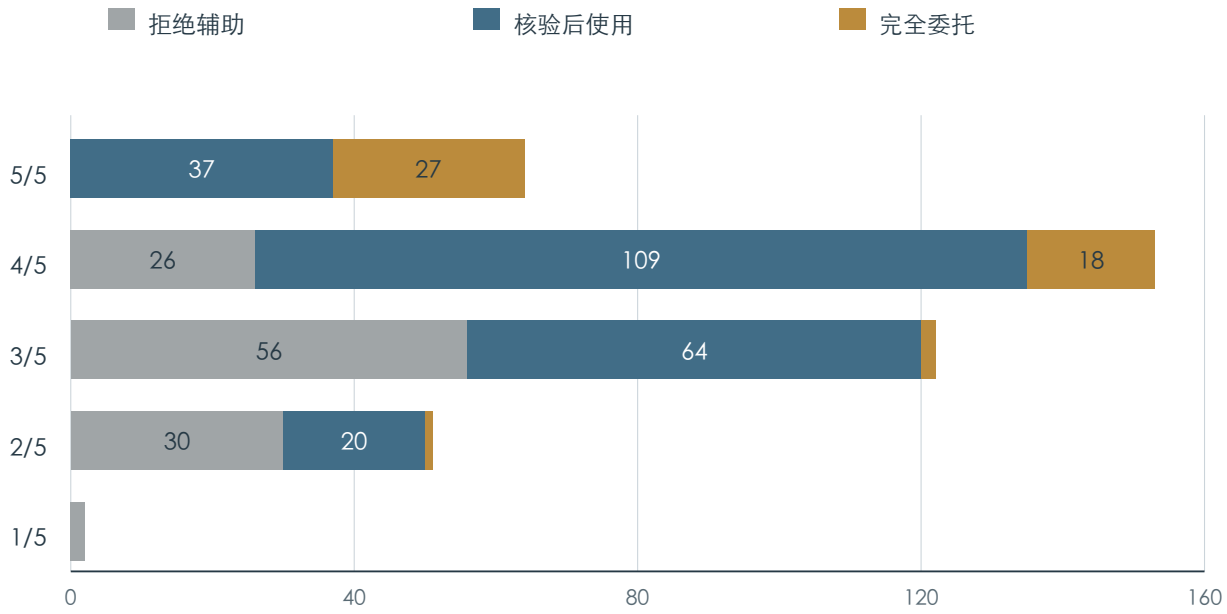
原论文：Following Delegation Traces（固定预印本版本）

公开数据：Zenodo，CC BY 4.0

作者代码：本次使用的固定提交

高信任的回答里，完全放手仍是少数

各信任等级下的未来委托意向



图中的信任等级从1到5，4和5算高信任。合起来有217轮：172轮仍表示核验或拒绝完全委托，45轮选择完全委托。这里说“保留监督”，指的是参加者表达的打算，还没有观察他是否真的核验。

同一组数据的具体计数

信任等级	拒绝辅助	核验后使用	完全委托
1/5	2	0	0
2/5	30	20	1
3/5	56	64	2
4/5	26	109	18
5/5	0	37	27

不能把79.3%翻译成“79.3%的人不信任 AI”，也不能说“79.3%的人已经采取了监督行动”。它的分母是217次高信任回答，不是217个人；而且“我相信它能做好”和“我愿意不再检查”本来就可能是两种判断。

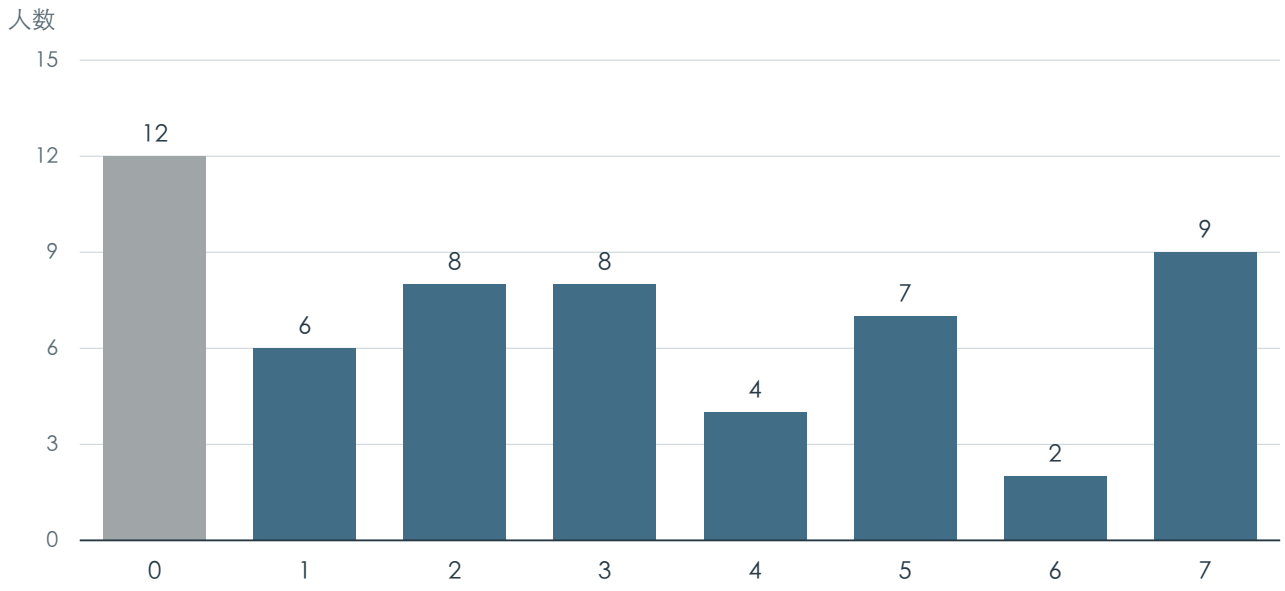
再按人看：不是几个人反复回答撑起来的

参与者层面补充：这些回答涉及多少人

观察	人数	分母
至少有一轮高信任	46	56
高信任时至少一次仍保留监督	44	46
高信任时至少两次仍保留监督	38	46
所有高信任轮次都保留监督	31	46

56人里有10人从未给过高信任评分，所以讨论“高信任时会怎样”时，他们没有可计算的条件比例，不能硬填成0。剩下46人中，44人至少有一轮“信任高，但仍保留监督”；31人在自己所有的高信任轮次中都如此。

每个人有几轮“高信任但保留监督”



每人的发生轮数（0至7轮）；分布包括全部56人

这张图保留全部56人。0轮的12人中，10人从未报过高信任，另2人在高信任时总表示完全委托。把每个有高信任经历的人的比例等权平均，得到83.0%；按回答轮次加权则是79.3%。这是两种分母，不是挑一个更漂亮的数字。

这说明该现象在这份样本中涉及不少人，但还没有回答为什么。可能是任务风险、个人习惯、对能力和控制权的不同理解，也可能有其他解释。仅凭这些数据，不能选中其中一种当作已证明的心理机制。

我这边补完的工作，怎样帮助后面的实验

我借助 AI 把这篇近邻从论文里的汇总数字，一直追到发布记录、筛选规则和实际模型输入。另写了一条字段重建路径来核对作者的读取结果，并留下一份能从头运行、带图和计算输出的笔记本。它不是新的真人实验，但能让我们在设计自己的研究时，知道哪些指标究竟在回答什么。

最有帮助的不是“又跑了一个模型”，而是三处具体区分：392轮不能当392人；高信任时想核验不等于已经核验；同一人当轮信任较高与更愿委托有关，不等于提高信任就一定导致放权。我们以后写方案、算样本、收日志和解释结果，都要按这个区分来。

主模型的本地复算：点估计接近，标准误尚未完全对齐

变量	论文系数	本地系数	论文标准误	本地标准误
同一人本轮信任的相对变化	2.71	2.7070	0.38	0.3628
这个人的平均信任	4.39	4.3839	0.54	0.4761
早／晚队列	0.79	0.7951	0.60	0.5823
是否有图界面	4.39	4.3890	0.69	0.6419
是否暴露于AI错误	-1.76	-1.7580	0.67	0.6430
任务次序	0.20	0.1958	0.09	0.0887

上表的系数描述关联，不是因果作用。我沿用作者发布的模型和默认数值方法，没有为凑一致而改优化器或反复换设定。三组主／队列模型都收敛，但部分标准误与论文不同，具体原因尚未定位；论文中另一个控制视频差异的敏感性分析没有取得准确发布脚本，因此不能说全部统计结果已经复现。

这些准备能替后续省掉的，是把问卷当行为、把重复轮次当独立样本、把漂亮关联当心理机制的返工。原始个体轨迹和自由文本不会放到网页上；展示只用必要的汇总计数。

接下来怎么用，考虑投哪里

这次复算不单独拿去写成我们自己的新论文。后续若继续 AI 小组与真人求助问题，我仍建议先按 Cyberpsychology: Journal of Psychosocial Research on Cyberspace 的要求完善方案：明确人的心理问题，事前注册实验，通过伦理路径，留下实际行为与可复查材料。若正式研究能给出更有精度、解释更扎实的行为证据，再重点考虑 Computers in Human Behavior。

这两本是沿用此前的选刊论证，不是本次结果已经达到投稿要求。我们还缺能稳定呈现两组差别的材料、实际人类数据和足够精度。是否满足你的毕业口径，仍要按学院和导师的实际要求确认。

Cyberpsychology：作者要求

Computers in Human Behavior：出版方研究范围

你需要做哪些工作：先和导师判断，最值得解释的到底是“向真人开口”，还是“把什么权限交给 AI”；选好问题以后，再共同确定测量、目标人群、伦理负责人和真实可兑现的交流机会。现在不用急着找被试，也不用替我再搭一套 Agent 框架。

我这边已经整理了近邻、对照、原型的真实失败记录，以及这次公开数据的复算。下一步应把这些证据用在一项有明确增量的完整设计上，再决定是否继续投入，而不是因为原型已经写了很多，就勉强把它变成论文。

本报告的复算、代码和初稿由我这边借助 AI 完成，尚无独立的人类审查；数据与原始研究归原作者。这里新增的是可复查的二次描述和设计判断，不是已完成的真人效果研究。