被算法腐蚀？人工智能生成的建议与人工撰写的建议如何塑造（不）诚实

by

Margarita Le ib*、Nils Köbis*、Rainer Michael Rilke、Marloes Hagens 和 Bernd Irlenbusch

* 共同第一作者

本研究已获科隆大学管理学、经济学与社会科学学院伦理委员会批准，审批编号为 200010BI。

致谢：感谢 Clara Bersch、Yulia Litvinova、Ann-Kathrin Blanke、Toan Huynh、Anna Vogts 和 Matteo Tinè 提供研究协助；同时感谢 Iyad Rahwan、Jean-Francois Bonnefon、Aljaz Ule、Anne-Marie Nussberger，以及以下活动的与会者提供宝贵意见：认知、价值观与行为研究组（慕尼黑路德维希-马克西米利安大学 / 慕尼黑路德维希-马克西米利安大学）、道德人工智能实验室会议（马克斯·普朗克人类发展研究所与图卢兹经济学院）、应用伦理与道德研究组（布拉格经济与商业大学）、决策研究中心（利兹大学）、经济与管理系（比萨大学）、决策与经济心理学中心（本-古里安大学）、行为与管理科学研究组（以色列理工学院）、蒂尔堡大学社会心理学系研讨会，以及埃尔朗根-纽伦堡弗里德里希-亚历山大大学计算机科学系研讨会。

资助：本研究由德国研究联合会（DFG）通过德国卓越战略项目资助，项目编号为 EXC 2126/1–390838866；同时获得 ECONtribute：市场与公共政策、欧洲研究委员会（ERC-StG-637915）以及科布伦茨工商会（IHK）的资助。

摘要

人工智能（AI）日益成为不可或缺的顾问。如果人工智能劝说人们做出不诚实的行为，就会引发新的伦理问题。在一项实验中，我们研究了人工智能建议（由自然语言处理算法生成）如何影响（不）诚实，将其与相同的人类建议进行比较，并检验建议来源的透明度是否重要。我们发现，促进不诚实的建议会增加不诚实行为，而促进诚实的建议并不会增加诚实行为。人工智能建议和人类建议都是如此。算法透明度是一种常被提出、用于降低人工智能风险的政策措施，但它并不影响行为。这些发现为负责任地管理人工智能建议迈出了第一步。

关键词： 人工智能、机器行为、行为伦理学、建议

被算法腐蚀？人工智能生成和人工撰写的建议如何塑造

（不）诚实

人工智能（AI）每天都在影响人们的生活（Rahwan 等人，2019）。它为在线市场设定价格（Calvano 等人，2020），预测医疗保健成本（Obermeyer 等人，2019）和刑事判决（Kleinberg 等人，2018）等重要结果，并提供从娱乐内容、购买决策到恋爱对象的各种推荐（Dellaert 等人，2020；Yeomans 等人，2019）。人工智能日益成为不可或缺的顾问，进而影响人们的行为（Fast 和 Schroeder，2020；Kim 和 Duhachek，2020）。例如，亚马逊首席科学家 Rohit Prasad 设想，Alexa 面向其超过 1 亿用户所扮演的角色“将不断发展，从更像助理转变为顾问”（Strong，2020）。鉴于人工智能作为顾问的作用日益增强，探究人们是否会基于人工智能建议而被说服遵循或违背道德规则至关重要（Köbis 等人，2021）。

领英和 Zillow 等大型企业已经开始采用人工智能顾问，因此可能会影响员工的伦理行为。在这类企业中，自然语言处理（NLP）算法（例如由 Gong.io 等软件提供的算法）会分析员工录制的销售通话，并就如何提高销售额向他们提供建议。如果缺乏监督，这类算法可能会发现欺骗客户能够获利，进而建议销售人员这么做。事实上，自然语言处理算法已经能够自主识别欺骗行为，将其视为谈判任务中的一种有效策略（Lewis 等人，2017）。如果人们听从这种腐化性的人工智能建议，就会产生伦理风险。在本文中，我们考察：（i）人们是否会根据人工智能生成的建议，有意义地改变其（不）道德行为；（ii）此类建议与人工撰写的建议

相比有何不同。最后，我们检验（iii）人们是否会受到建议来源（人工智能还是人类）认知的影响。

关于（不）道德行为的建议：人类与人工智能

一般而言，人们不太愿意采纳他人的建议（“自我中心的建议折扣”，例如 Yaniv & Kleinberger，2000），尤其是在建议未经请求的情况下（Bonaccio & Dalal，2006）。然而，面对道德困境时，建议对接受建议的人有诸多显著益处。鼓励采取合乎道德的行动方针的建议，可能会印证一个人的道德偏好，从而减轻因错失通过说谎最大化利润的机会而产生的后悔等负面情绪。鼓励采取不道德行动方针的建议，则可能让人们得以为获取利润而违反道德规则，同时不损害其道德自我形象（Cross 等人，2001）。事实上，采纳建议甚至能让人产生与顾问共同承担责任的感觉（Harvey & Fischer，1997）。

与接受人类建议相比，人们会如何看待人工智能给出的建议？自然语言处理领域近期的技术进步表明，人工智能生成的文本已能与人类文本难以区分，这说明人工智能建议与人类建议同样具有说服力。例如，人工智能电话助理 GoogleDuplex 能够预约，并在接听者甚至没有意识到电话另一端是人工智能的情况下，进行完整流畅的对话。此外，人工智能能够生成从诗歌（Köbis & Mossink, 2021）、Airbnb 个人资料（Jakesch 等人, 2019）到新闻文章 （Kreps 等人, 2021）等与人类作品不相上下的内容。 因此可以推断，如果人们没有被告知建议的来源，就无法正确识别建议来源，也会同等受到人工智能建议和人类建议的影响。

测试算法透明度

为了确保人们清楚自己正在与谁互动，各国政府、政策制定者和研究人员一致呼吁提高算法透明度（Jobin et al., 2019），即强制披露人工智能的存在（Diakopoulos, 2016）。欧盟近期发布的《人工智能法案》要求聊天机器人和电话助理等人工智能系统在与人类互动时，披露其人工智能身份（欧盟委员会, 2021）。尽管这是一项广受欢迎的政策建议，但目前缺乏其能否有效塑造人们伦理行为的实证证据。

透明度如何影响人们对建议来源的反应，并非显而易见。以往研究提出了三种相互竞争的可能性。第一种可能是，得知建议来源后，人们更倾向于听从人类建议，而非人工智能建议。这一观点基于有关算法厌恶的研究（Dietvorst 等人，2015）。在人们看来客观且技术性强的领域（例如数值估计、数据分析和指路，Castelo 等人，2019；Logg 等人，2019），他们很乐于依赖人工智能。然而，对于主观决策，尤其是涉及伦理影响的决策（例如假释判决、电车难题，Bigman & Gray，2018；Castelo 等人，2019；Laakasuo 等人，2021），他们不太愿意使用人工智能。此外，人们在作出（不）道德决策时会遵循其所感知到的社会规范（Bowles，2016；Fehr，2018；Gächter 和 Schulz，2016；Gino 等人，2009；Köbis、Troost 等人，2019）。与人工智能建议相比，人类建议可能是社会规范的更强信号，因为社会规范规范着并源自人类（而非人工智能）的行为。因此，人们更有可能听从人类建议。假设人们在涉及伦理的情境中确实更偏好人类意见，并认为人类建议是社会规范的更强线索。在这种情况下，我们应当预期，与人工智能建议相比，人类建议对人们的（不）道德行为影响更大。

第二种可能是，得知建议来源后，人们听从人类建议的程度低于听从人工智能建议的程度。仔细考察人工智能建议系统的技术设计，会发现这一观点有其依据。自然语言处理算法是在大量人工撰写文本语料上训练而成的（Radford 等，2019）。如果人们知道自然语言处理算法依据大量汇总的人类意见生成建议，他们可能会认为，人工智能建议比来自某一个人的建议更能代表大多数人的信念和行为。如果人工智能建议确实比单条人工撰写的建议更能体现社会规范，那么我们应当预期，与人类建议相比，人工智能建议对人们的（不）道德行为影响更大。

第三种可能性是，当人们得知建议的来源时，他们受到人类建议和人工智能建议同等程度的影响。这一解释得到以下观察结果的支持：人们已经会向人工智能代理寻求建议。例如，超过700万人求助于Replika——这个“关心你的人工智能伴侣。总是在这里倾听和交谈。永远站在你这边”的应用（replika.ai），以获得虚拟陪伴、进行社交活动，以及获取建议（Murphy，2019）。这类人工智能顾问也可能帮助人们为可疑行为寻找正当理由。当受到诱惑、想为了获利而违反道德规则时，只要能够为自己的行为辩护，人们就会这样做（Barkan 等，2015；Fischbacher 和 Föllmi-Heusi，2013；Shalvi et al., 2015）。接受鼓励违反规则的建议，可以成为一种受欢迎的正当理由，即使这类建议源自人工智能，也可能如此。事实上，人们不仅会与他人分摊责任、推卸对有害结果的责任归咎（Bartling 和 Fischbacher，2011；Bazerman 和 Gino，2012；Tenbrunsel 和 Messick，2004），也会与人工智能系统分摊责任（Hohenstein & Jung, 2020）。如果遵循人工智能建议和人类建议同样具有正当性，并且会导致两者之间相似的责任归因，那么我们应当预期，人工智能建议和人类建议对人们（不）道德行为的影响程度相同，导致

人类建议和人工智能建议会在同等程度上影响人们的（不）道德行为。

本研究

本研究检验建议类型（促进诚实还是促进不诚实）、建议来源（人工智能还是人类）以及有关建议来源的信息（透明还是不透明）如何影响人类的（不）道德行为。以往研究考察了人们对描述人工智能建议的陈述偏好，这些偏好涉及假设情境 （Bigman & Gray，2018年；Castelo 等人，2019年；Kim & Duhachek，2020年；Logg 等人，2019年）。我们采用机器行为方法（Rahwan 等人，2019年），通过考察人们对人工智能生成的实际输出的行为反应 ，对这类研究加以补充。

为衡量人们的（不）道德行为，我们采用成熟的激励性掷骰子任务（Fischbacher 和 Föllmi-Heusi，2013年）。在该任务中，参与者私下掷骰子并报告结果，结果越高，报酬也越高（类似方法参见 Abeler 等，2019年；Fehrler 等人，2020年；Sutter，2008年）。为生成真实的人工智能建议，我们采用最先进的算法 GPT-J。我们使用少量训练数据对该算法进行微调，以生成相关的建议文本。此类“少样本”学习方法模拟了许多现实场景：语言模型（例如 Alexa）并未经过优化，无法最大化用户的利润或提供最具说服力的建议，而只是根据提示词生成相关信息（Brown 等人，2020年）。

方法

我们开展了一项大规模预注册实验。实验设计包含两个部分。第一部分，我们收集了人类撰写的建议并生成了人工智能建议；第二部分，我们考察了建议对（不）诚实行为的影响（概览见图1）。通过

使用真实的人工智能文本输出，我们避免了实验欺骗，并能深入了解人们对实际人工智能建议的反应。我们在正文和附录中报告了所有测量指标、操纵和数据排除情况。本系的机构审查委员会批准了该实验，所有材料、预注册文件和数据均可在开放科学

框架。

第一部分 - 建议提供任务

人类撰写的建议。首先，我们开展了一项建议提供任务，并通过 Prolific.co 招募建议者（N = 367，M_{age} = 30.28，SD_{age} = 9.56，女性占 45.50%）（预注册 https://osf.io/nbke2）。任务耗时 10 至 15 分钟，参与者获得 1.27 英镑的基本报酬。建议者得知，另一组参与者（受建议者）将参与一项掷骰子任务（Fischbacher & Föllmi-Heusi，2013年），他们会私下掷骰子并报告结果（结果越高，报酬越高）。建议者被要求为这些受建议者撰写建议。

建议者被随机分配撰写促进诚实或促进不诚实的建议。在促进诚实建议处理组中，我们告知建议者，如果受建议者阅读了他们的建议，随后报告实际掷骰子结果（即诚实行事），他们将获得 3 英镑奖金。促进不诚实建议处理组中的建议者则得知，如果受建议者阅读了他们的建议，随后报告最高可能结果 6，他们将获得 3[1] 英镑奖金。

建议者必须遵循预先设定的建议写作规则，以确保其生成连贯的建议文本，可用于训练 GPT-J。具体而言，他们的建议必须：（i）

[1]如果受建议者遵从促进不诚实处理条件下的建议，他们在大多数情况下（6种情况中的5种）都会说谎。只有当实际掷骰子结果为6时，遵从建议才不意味着说谎。

至少包含 50 个单词；（ii）不得以数字或文字形式使用具体数字[2]，；（iii）使用英语并以自己的措辞撰写；（iv）写成完整句子；（v）围绕受建议者报告掷骰子结果的决策；（vi）不得告知受建议者，建议者的收益取决于其行为[3]。

为激励建议者遵守建议写作规则，他们有机会获得奖金。具体而言，我们从所有建议文本中随机选取一篇；如果该文本遵守了写作规则，建议者便可获得10英镑奖金。此外，为了激励建议者写出有说服力的文本，我们实际采用了1%的建议文本（400篇中有4篇）。如果受建议者按照被采用的建议行事，相应的建议者便可根据其所属的治疗组获得奖金（促进诚实与促进不诚实的建议）[4]。

人工智能生成的建议。 为生成（见图1A）人工智能建议，我们采用了 GPT-J[5],这一由 Eleuther AI 发布的开源自然语言处理算法（https://www.eleuther.ai/）。GPT-J使用经过筛选且多样化的 825 GiB 文本数据集进行训练，以预测词序列中的下一个词，并包含 60 亿个参数（Wang & Komatsuzaki，2021年）。GPT-J 可以通过额外训练进行微调，从而生成特定类型的文本。我们对 GPT-J 进行了微调

2 为了生成高质量的人工智能建议，不允许建议者使用具体数字。GPT-J 经过训练，能够预测句子中的下一个词（参见“人工智能-生成的建议”一节）。如果允许建议者具体提及数字，那么使用人工撰写的建议训练 GPT-J，可能会导致 GPT-J 的输出中出现与上下文无关的随机数字，从而降低人工智能生成的建议的质量。3 不允许建议者向受建议者提及他们的激励机制，以便让阅读人工智能建议和人类建议的受建议者保持相同的亲社会动机（均为零）。4 向建议者支付报酬，需要知道参与者阅读建议后，是否如实报告了观察到的掷骰子结果。为此，我们运行了一个经过修改的掷骰子任务：受建议者收到随机选取的建议，在计算机屏幕上看到一次掷骰子结果，并被要求报告该结果。我们对随机选取的四篇建议文本（占建议文本的1%）和四名受建议者采用了这一流程。这种非私密流程使我们能够确定受建议者是否如实报告，并据此向建议者支付报酬。这样一来，我们的实验设置包含了激励机制，也没有涉及实验欺骗。在主要实验中，掷骰子结果是保密的（参见“第2部分- 建议采纳任务”）。 5 正如我们在预注册中所述，我们最初计划采用 GPT-2（参见https://openai.com/blog/better-language-models/）来生成人工智能建议。然而，我们最终选择使用 GPT-J，因为它是开源的，有助于提高可复现性；而且它的规模更大、能力更强，因此也比 GPT-2 更先进。

采用“少样本”学习，分别使用人类撰写的促进诚实的建议和促进不诚实的建议对其进行训练，这些建议来自建议提供任务。 微调时，我们只使用了符合建议写作规则的建议文本（由一名非专业编码员进行编码）。有关 GPT-J 校准的更多详情，请参见附录。

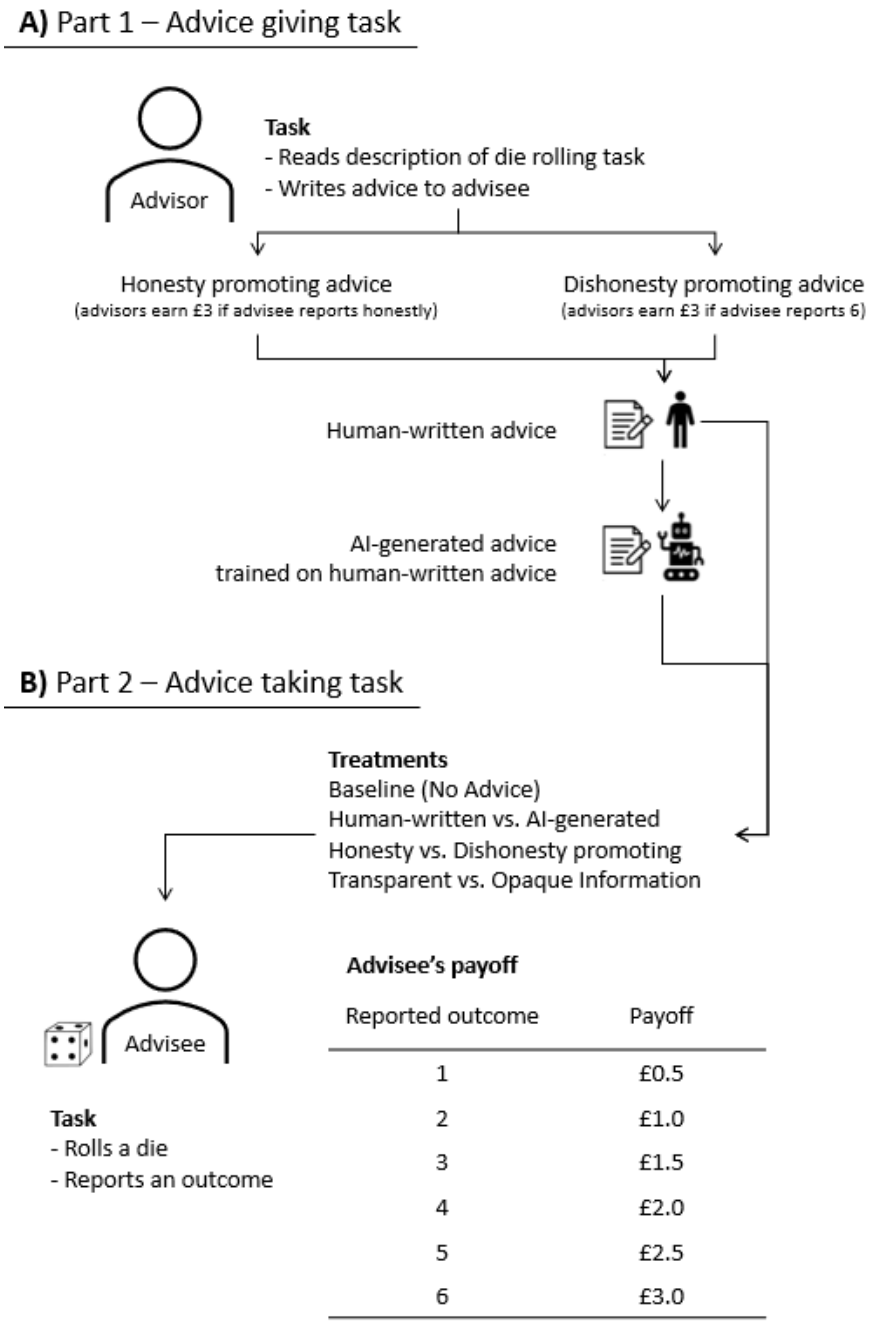

图1。（A）第1部分——建议提供任务（B）第2部分——建议采纳任务。（A）参与者受到激励，撰写促进诚实或促进不诚实的建议文本，随后用这些文本生成了人工智能建议。（B）另一组参与者完成掷骰子任务。受建议者阅读建议，然后报告掷骰子结果。我们总共实施了九种处理条件：参与者阅读促进诚实或促进不诚实的建议，这些建议由人类撰写或由人工智能生成。参与者被告知建议来源（透明度），或未被告知（不透明）。作为基线，另一组参与者没有阅读任何建议。

筛选。收集人类建议并生成人工智能建议后，我们对两种来源采用了相同的预先设定的筛选程序（见图2）。首先，我们剔除了超过100个单词的文本。接下来，为确保受建议者读到连贯且相关的建议文本，我们从每个单元格中随机选取了100份建议文本。两名不了解实验处理条件的独立编码员根据以下标准对每条建议进行编码：（a）文本是否连贯？（是/否）；（b）文本是否包含明确的建议？（是/否）；（c）建议鼓励哪种行为？（诚实/不诚实/不明确）；（d）建议是否遵循建议写作规则？（是/否）。此外，我们使用客观的 Grammarly 可读性评分作为文本质量的计算代理指标[6]。

在通过编码程序[7] 且获得不低于50分的Grammarly评分的文本中，我们针对每种治疗随机选取了20篇建议文本（人工智能生成 与人工撰写，以及促进诚实与促进不诚实），最终得到80篇建议文本的样本，用于第2部分（所有建议文本见附录）。通过采用相同的筛选

6 我们从 grammarly.com 获取了 Grammarly评分和可读性评分。Grammarly评分将文本与平台上检查过的所有其他文本进行比较。得分为80表示，就语法正确性而言，该文本的得分高于 grammarly.com 上检查过的所有文本中的80%。可读性评分采用弗莱施阅读易度测试，反映文本的易读程度。该评分根据平均句长和每个单词的平均音节数计算，得分越高表示越容易阅读。7 通过编码流程的文本必须满足以下条件：（i）内容连贯；（ii）包含明确的建议；（iii）在促进诚实处理中鼓励诚实，在促进不诚实处理中鼓励不诚实；（iv）遵循建议写作规则。此外，两位独立编码员的编码结果必须一致，文本才能通过。

程序来处理人类和人工智能建议，我们确保建议文本符合最低质量标准，并尽可能具有可比性。

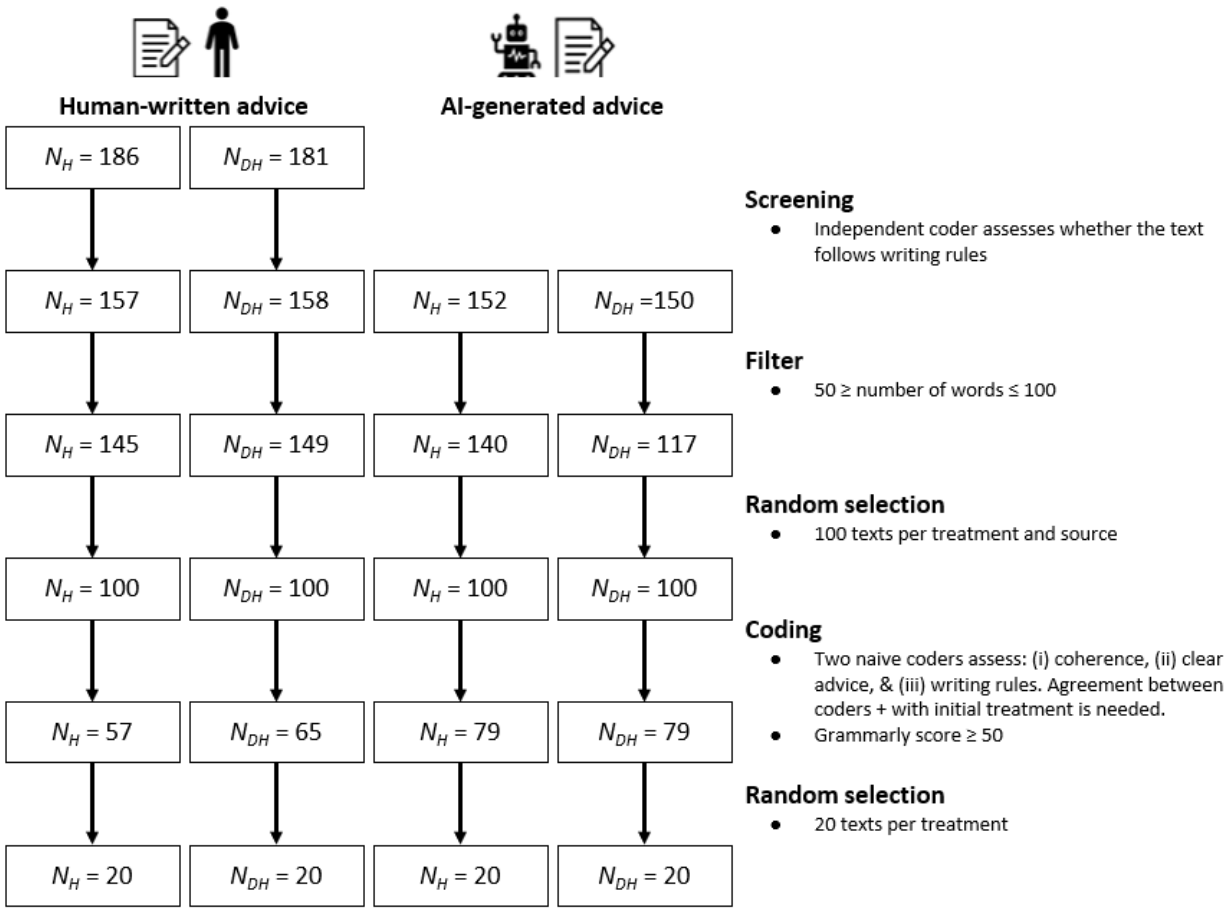

图2. 建议文本筛选流程概览。H = 促进诚实的建议；DH = 促进不诚实的建议。

第*2*部分——建议采纳任务

采纳建议任务约需 8 分钟完成，参与者获得 1.20英镑的固定报酬。我们预先注册（https://osf.io/nqvf3），以收集足够大的样本，使我们能够检测到小到中等效应量（每个单元 200 名参与者，共 1,800 名）

参与者通过 Prolific.co 参与了采纳建议任务。最终，共有 1,817 名（M_{age} = 32.39；SD_{age} = 11.68，女性占 48.72%）参与者纳入分析。这些参与者完成了任务和自我报告条目，并通过了理解检查和注意力检查（见下文）。针对统计功效为 90%、显著性水平为 .05 的回归进行敏感性分析后发现，我们的样本足以检测小效应量（f^2 = .006 和 .010；详情见附录）。

参与者阅读说明、收到建议，最后完成掷骰子任务。具体而言，参与者需要私下掷骰子并报告结果。参与者的报酬取决于其报告：报告“1”可获得£0.5；报告“2”可获得= £1，报告“3”可获得= £1.5，报告“4”可获得= £4，报告“5”可获得= £2.5，报告“6”可获得= £3。阅读说明后、开始掷骰子任务前，所有参与者都得知，将随机抽取10%的参与者，并支付他们完成掷骰子任务的报酬。

在经济学和心理学研究中，使用掷骰子任务评估不诚实行为是一种常见做法（参见荟萃分析：Abeler 等，2019；Gerlach 等，2019；Köbis、Verschuere 等，2019；Leib 等，2021）。此外，这项任务具有良好的外部效度，因为在掷骰子任务中说谎，与实验室之外的不道德行为相关，例如逃票乘坐公共交通（Dai 等，2018）以及无故缺勤（Hanna 和 Wang，2017）。重要的是，掷骰子任务让两种相互竞争的动机正面交锋：成为一个有道德、诚实的人，与最大化经济收益。因此，建议在这种情境下尤其有价值。

阅读说明后，参与者必须正确回答三道理解题并通过注意力检查。如果参与者答错某道理解题，可以重新作答。如果他们同一道题答错

两次，或者答错注意力检查，就会被取消参与研究的资格，且不会获得报酬。

处理条件。在报告掷骰子结果之前，参与者被随机分配到九种不同的处理条件之一。参与者阅读的建议要么是促进诚实的，要么是促进不诚实的；这些建议要么由人类撰写，要么由人工智能生成。此外，参与者要么会被告知建议的来源（透明度处理条件），要么不会（不透明处理条件）。另外，在基线处理条件下，参与者不会收到任何建议。因此，整体实验设计为 2（建议来源：人类 vs 人工智能）× 2（建议类型：促进诚实 vs 促进不诚实）× 2（关于建议来源的信息：透明度 vs 不透明）+ 1 （基线，无建议）的被试间设计。

阅读了人工智能生成的建议且获知该建议由人工智能生成的参与者（在透明度处理条件下）阅读了以下内容：

> “为了帮助您做出决策，一个人工智能（*AI*）算法生成了一条建议。
>
> 关于人工智能建议的信息：
>
> 该算法使用大型英文文本数据集（*> 820 GiB*，相当于超过 *5* 亿页文本）进行训练，并通过预测句子中的下一个词来生成文本（类似于智能手机上的预测文本）。此外，也可以使用其他特定文本进一步训练该算法。为了生成多条人工智能建议文本，该算法使用了从研究中其他参与者那里收集的建议文本进行训练。这些参与者没有参加掷骰子任务，只被要求针对掷骰子任务中的决策撰写建议

。你将读到的建议是由该算法生成的一条建议文本。”

阅读了人类撰写的建议且获知其来源的参与者（处于透明度处理条件下）阅读了：

“为了帮助你做出决策，另一名参与者写了一条建议。

关于建议的信息：

为了收集多条建议文本，研究人员要求另一组参与者针对掷骰子任务中的决策撰写建议。这些参与者没有参与掷骰子任务，只被要求针对该任务中的决策撰写建议。你将阅读的建议是由一名参与者撰写的。”

处于不透明处理条件下、因而未获知建议来源的参与者阅读了：

“为了帮助你做出决策，你将阅读一条建议。

这条建议可能由另一名参与者撰写，也可能由人工智能（*AI*）算法生成。建议由参与者撰写的概率为百分之五十，由算法生成的概率也为百分之五十。"

在不透明处理条件下，这段文字后面接着两段描述，与透明度处理条件中相同，分别说明如何收集或生成人工建议和人工智能建议。 在不透明处理条件下，关于人工智能建议生成和人工建议收集的信息以随机顺序呈现。[8]

8 为了控制参与者对潜在建议来源的看法，我们选择告知他们，这条建议由人类撰写或由人工智能撰写的概率各占百分之五十。我们认为，如果不提供任何有关建议来源的信息，参与者很可能会认为建议来源是另一名人类，因为人工智能对参与者而言可能并不是一个显而易见的建议来源。

静态图灵测试。完成掷骰子任务后，不透明处理条件下的参与者参加了一个带有奖励的静态图灵测试（Köbis & Mossink, 2021）。与经典图灵测试（图灵，1950）不同，参与者不会与建议来源进行来回互动，而是阅读建议文本，并判断他们认为这段文本是由人类还是人工智能撰写的。参与者得知，将随机选出其中20人；如果他们在静态图灵测试中的猜测正确，就能额外获得1英镑。

潜在机制。 最后，为了探索可能的机制，参与者完成了一份实验后问卷。参与者在 0 到 100 的量表上评估了他们认为报告高于实际观察到的掷骰结果的合理程度（i）适当性（命令性社会规范）、（ii）普遍程度（描述性社会规范）以及（iii）可辩护性。此外，除未接受任何建议的参与者外，所有参与者都评价了在掷骰子任务中，对于所报告结果，他们认为自己和顾问各自承担多大责任。答案量表从 0（= 我负全部责任）到 50（=我和顾问承担同等责任），再到 100（= 顾问负全部责任）。参与者还在 0 到 100 的量表上表示完成任务后感到内疚的程度（关于内疚的结果以及所有题项的措辞，见附录）。最后，所有参与者都报告了自己的年龄和性别。

结果

在九种处理条件下，参与者均有说谎，因为掷骰结果的平均值显著高于参与者诚实作答时的预期平均值（EV = 3.5）；单样本*t*检验，*t*s > 3.43，*p*s < .001。

人工智能生成的建议会影响人们的（不）道德行为吗？

当涉及促进不诚实的建议时，会；当涉及促进诚实的建议时，不会。我们首先关注不透明处理条件，在这些条件下，参与者并未获知建议来源。在这里，线性回归分析显示，接受人工智能生成的促进不诚实建议后，平均掷骰子报告（M = 4.59，SD = 1.37）显著高于基线无建议处理条件下的报告（M = 3.98，SD = 1.55，b =.609；p < .001；95 % CI = [.324，.894]）。然而，接受人工智能生成的促进诚实建议后，掷骰子报告（M = 4.00，SD = 1.62）与基线处理条件下的报告没有显著差异（b = .019；p = .898；95% CI = [-.275，.314]），见图3和表1（模型 1）。此外，人工智能生成的促进不诚实处理条件下的掷骰子报告显著高于人工智能生成的促进诚实建议处理条件（b = - .590，p < .001；95% CI = [-.881，-.299]）。因此，促进不诚实的人工智能建议确实成功腐化了人们，而促进诚实的人工智能建议却未能使人们转向诚实。

How 与人类撰写的建议相比，人工智能生成的建议是否同样可信？

无论是促进诚实的建议还是促进不诚实的建议，人工智能生成的建议对行为的影响都与人工撰写的建议相似。聚焦于不透明度处理条件，二阶交互（建议来源 × 建议类型）不显著（b = .069，p = .744；95% CI = [-.349，.489]），见图3和表1（模型 2）。具体而言，当建议旨在促进诚实时，人工智能生成的建议与人工撰写的建议在平均掷骰子报告上没有差异（前者 M = 4.00，SD = 1.62；后者 M = 3.92，SD = 1.51，b = -.076，p = .631；95% CI= [-.387，.235]）。同样，当建议旨在促进不诚实时，人工智能生成的建议与人工撰写的建议在平均掷骰子报告上没有差异（前者 M = 4.59，SD = 1.37；后者 M= 4.58，SD = 1.53，b = -.006，p = .964；95% CI = [-.289，.276]）。

此外，图灵测试静态版本的结果表明，个体无法区分人工智能生成的建议与人工撰写的建议。具体而言，在 不透明 处理条件下，49.93%（803 名参与者中的 401 名）正确猜出了建议来源，这一比例与随机猜测水平（50%）没有差异（二项检验：p = .999；95% CI = [.464，.534]）。

建议来源的透明度重要吗?

不重要，告知参与者建议来自算法还是人类，并不会改变他们的行为。线性回归分析显示，三重交互作用（建议类型 × 来源 × 信息）不显著（b = .100，p = .735；95% CI = [-.482，.683]），见图3和表1（模型3）。在不透明和透明度处理条件下，二阶交互作用（建议来源 × 建议类型）均不显著（透明度：b = .170，p = .409，95% CI = [-.234，.575]；不透明：b = .069，p =.744，95% CI = [-.349，.489]）。

总体而言，广受欢迎的算法透明度政策建议并不能减轻人工智能建议的腐化效应。也就是说，在不透明 处理条件下，受人工智能生成的促进不诚实建议影响后提交的掷骰子报告（M = 4.59，SD = 1.37），与透明度 处理条件下遵循相同建议提交的报告（M = 4.61，SD = 1.40，b = .020，p = .878；95% CI = [-.244，.286]）不相上下。具体而言，当参与者未被告知建议来源时，遵循人工智能生成的促进不诚实建议后，他们的报告比基线 [(4.59-3.98)/3.98 = .152], 高出15.2%；而当他们被告知建议的来源时，报告也增加了15.2%[(4.61-3.98)/3.98 = .158]。贝叶斯分析也证实了这些结论（见附录）。

总体而言，结果与以下观点一致：人们越来越多地听从人工智能建议（例如 Replika），并利用人工智能生成的建议为自己为牟利而违反道德规则辩解。

所得结果的稳健性。在我们的实验设计中，促进不诚实 处理条件下的建议者只有在受建议者报告最高值“6”时才能获得 3 英镑。这种激励方案与促进诚实 处理条件相当，在该条件下，建议者只有在受建议者如实报告时才能获得 3 英镑。在这两种情况下，建议者都只会在受建议者的 6 种可能报告中有 1 种情况下获得金钱（即视处理条件而定，受建议者报告“6”或如实报告），而在其余 5 种情况下都不会获得金钱。然而，促进不诚实 处理条件下建议者的激励方案可能导致建议文本主要集中于劝说参与者报告结果“6”。为评估所得结果的稳健性，我们（i）进行了额外分析，并（ii）实施了额外处理条件。

掷出六点的比例。 首先，作为一项补充分析，我们考察了以掷出6点的比例作为替代结果变量时，是否会得出相同结论。我们发现结果非常相似（见图3，白点表示所有处理条件下掷出6点的比例）。具体而言，聚焦于不透明处理条件，线性回归分析显示，接受人工智能生成、促进不诚实建议后，掷出六点的比例（32.44%）显著高于基线无建议处理条件下的比例（20.65%；b = .612；p = .005，95% CI = [.182，1.051]）。然而，接受人工智能生成、促进诚实建议后，掷出六点的比例（21.93%）与基线相比并无显著差异（b = .076；p = .751，95% CI = [-.398，.551]）。此外，人工智能生成、促进不诚实建议处理条件下掷出六点的比例显著高于人工智能生成促进诚实建议处理条件（b = -.535，p = .016，95% CI = [-.979，-.101]）。

此外，聚焦于不透明处理条件，二阶交互作用（建议来源与建议类型）不显著（b = .444，p = .171，95% CI = [-.191，1.082]）。当建议旨在促进诚实时，人工智能生成建议处理条件（21.93%）与人工撰写建议处理条件下的掷出六点比例无显著差异（19.28%，b = -.162，p = .516，95% CI = [- .654，.327]）。同样，当建议旨在促进不诚实时，人工智能生成建议处理条件（32.44%）与人工撰写建议处理条件下的掷出六点比例也无显著差异（38.91%，b = .282，p = .173，95% CI = [-.123, 690]）。最后，三重交互作用（建议类型、来源与信息）也不显著（b = -.523，p = .257，95% CI = [-1.430，.382]）。在不透明和透明度处理条件下，二阶交互作用（建议来源与建议类型）均不显著（透明度：b = -.078，p = .810，95% CI= [-.724，.566]；不透明：b = .444，p = .171，95% CI = [-.191，1.082]）。

额外（对齐）处理条件。为评估我们的结果对于顾问激励方案的稳健性，我们进行了四项额外处理（建议来源：人类撰写 vs 人工智能生成；信息：透明度 vs 不透明）。在这些对齐处理条件下，受建议者阅读的是由激励与受建议者一致的建议者撰写的建议。对于这些建议者（*n* = 207），若受建议者报告“1”，建议者和受建议者各获得 £0.5；若受建议者报告“2”，双方各获得 £1，依此类推。我们再次使用这类人类撰写的建议文本对 GPT-J 进行微调。这些处理条件产生的结果与促进不诚实处理条件的结果相当。具体而言，所有四项对齐 处理条件下的平均掷骰结果都显著高于基线 处理条件（*p* = .066，适用于人工智能生成、不透明处理条件；其余三项处理条件均为*p*s < .001；有关这些处理条件的更多细节和扩展结果，请参见附录

）。结果的一致性表明，即使建议者的激励方案发生此类变化，我们的结果仍具有稳健性。

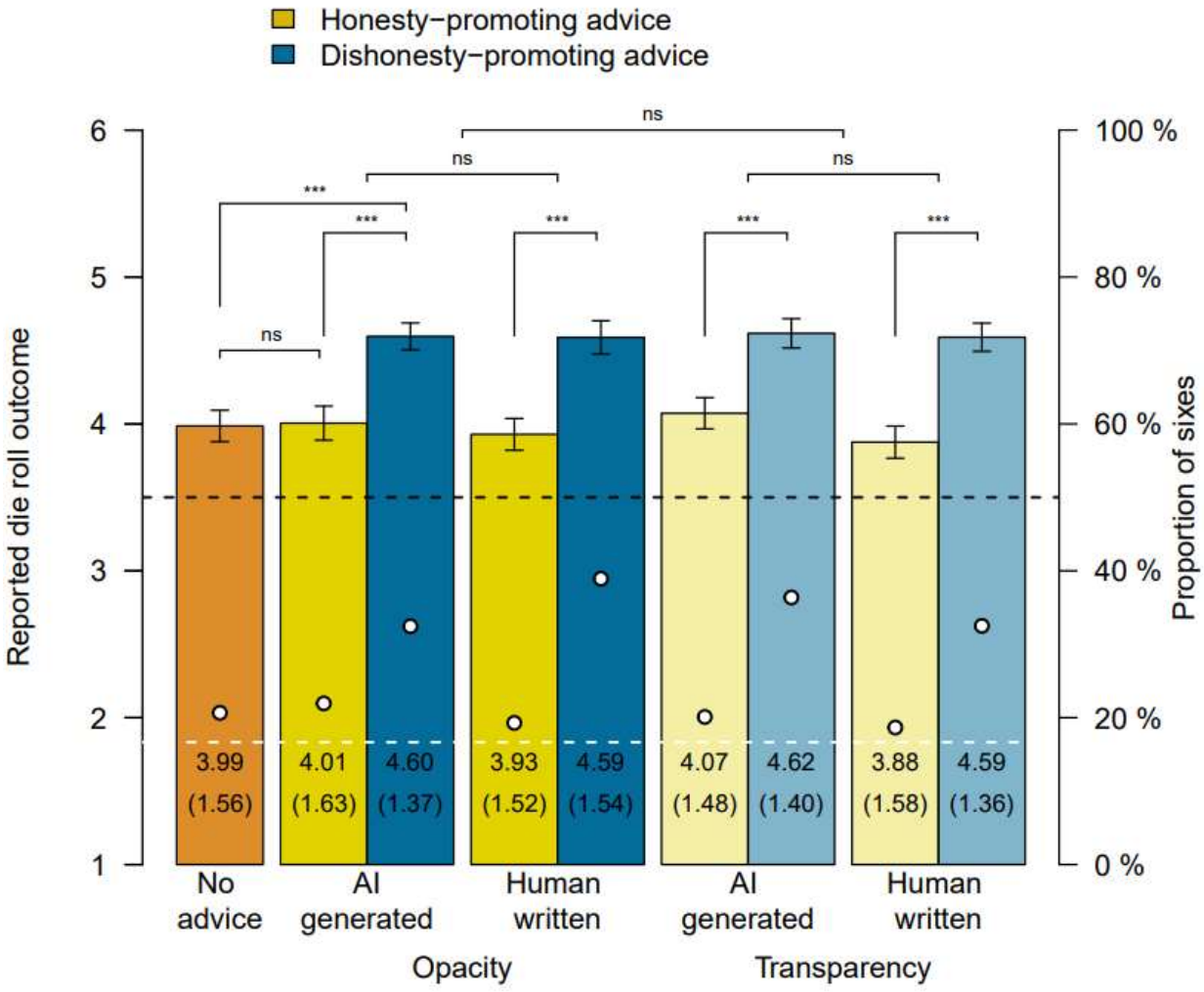

图3。不同建议类型（诚实与促进不诚实）、来源（人工智能与人类）及信息处理条件（不透明与透明）下，报告的掷骰结果均值（柱形）和报告为6的比例（白点）。黑色虚线表示参与者诚实时的预期均值（EV = 3.5），白色虚线表示参与者诚实时报告为6的预期比例（16.67%）。每个柱形底部标注了掷骰子报告的均值（标准差）；***p < .001；ns：p >.05。

潜在机制。 按照引言中提出的逻辑，本节考察参与者对以下方面的认知是否会因建议来源（人工智能与人类）和建议类型（促进诚实与促进不诚实）而异：报告高于实际观察到的掷骰子结果时，（i）是否适当（命令性社会规范）、（ii）这种做法的普遍程度（描述性社会规范）以及（iii）这种做法的可辩护性；此外，还考察他们（iv）如何在自己与顾问之间归因责任。静态图灵测试结果表明，参与者无法区分人工智能建议与人类建议。因此，我们仅关注参与者被告知建议来源的处理条件（透明度处理条件），以探究已知的建议来源和建议类型如何影响他们的认知。关于不透明处理条件的结果，请参见附录。

命令性规范。 以建议类型（促进诚实与促进不诚实的建议）为预测变量，对命令性规范进行线性回归分析发现，阅读促进不诚实建议的参与者（M = 33.93，SD = 31.43）比阅读促进诚实建议的参与者（M = 25.99，SD = 29.68，b= 7.94，p < .001，95% CI = [3.702，12.182]）认为报告较高的掷骰子结果更适当。这一发现表明，建议类型会影响感知到的命令性规范。值得注意的是，以建议类型和来源（人工智能与人类）为预测变量，对命令性规范进行线性回归分析发现，建议来源与类型的交互作用不显著，b = -4.81，p = .264，95% CI = [-13.292，3.658]。这些结果表明，人工智能建议和人类建议对命令性规范认知的影响相似（见图4a）。这一结果与行为发现一致：参与者的掷骰子报告受到建议类型的影响，但不受建议来源的影响。

描述性规范。 以建议类型预测描述性规范的线性回归显示，参与者认为，报告较高的掷骰子结果更为

常见于阅读促进不诚实建议（M = 76.02，SD = 22.75），而非促进诚实建议（M = 66.74，SD = 24.04，b = 9.27，p < .001，95% CI = [6.031，12.525]）时。这一发现表明，建议类型也会影响感知描述性规范。重要的是，以建议类型和来源预测描述性规范的线性回归显示，建议来源与类型之间的交互作用不显著，b = .25，p = .938，95% CI = [-6.230，6.745]，，这表明人工智能建议和人类建议对描述性规范感知的影响相似（见图4b）。这一结果与行为发现一致：建议类型影响掷骰子报告，但建议来源没有影响。

可辩护性。 以建议类型预测可辩护性的线性回归显示，阅读促进不诚实建议（M = 40.96，SD = 31.10）的参与者，比阅读促进诚实建议（M = 28.45，SD = 28.25，b = 12.50，p < .001，95% CI = [8.387，16.629]）的参与者，更认为报告较高的掷骰子结果是可辩护的。这一发现表明，建议类型会影响人们对掷骰子任务中说谎是否可辩护的看法。以建议类型和建议来源预测可辩护性的线性回归显示，建议来源与建议类型的交互作用不显著（b = -1.04，p = .803，95% CI = [- 9.280，7.194]），这表明人工智能建议和人类建议对可辩护性感知的影响相似（见图4c）。这一结果与行为发现一致：建议类型影响了参与者的掷骰子报告，但建议来源没有影响。

共同责任。 共同责任量表的范围为 0（= 我承担全部责任）至 100（= 顾问承担全部责任），其中 50 表示参与者和顾问平均分担责任。平均而言，参与者认为，相比顾问，自己对所报告的结果负有更多责任（M = 27.59，SD =

36.60，单样本 t 检验（与数值 50 比较），t = -17.32，p < .001）。此外，以建议来源（人工智能与人类）预测共同责任的线性回归显示，当建议来源是人工智能（M =28.27，SD = 36.53）或人类（M = 26.94，SD = 36.70）时，参与者归因的责任程度相似（b = -1.326，p = .608，95% CI = [-6.407，3.754]）。以建议类型和建议来源预测共同责任的线性回归显示，来源与类型之间的交互作用不显著（b = -5.91，p = .253，95% CI =[-16.083，4.248], 见图 4d）。无论顾问是人类还是人工智能，参与者都在同等程度上将责任归于自己和顾问，这与引言中阐述的逻辑一致：如果人们认为自己与两种建议来源都承担相近程度的共同责任，那么他们对人类建议和人工智能建议的采纳程度也会相似。

总之，自我报告条目的结果与引言中提出的第三种可能性一致。具体而言，我们发现，参与者对命令性和描述性社会规范的感知，以及他们感知到的正当性，在人类顾问和人工智能顾问之间没有差异。无论顾问是人类还是人工智能，参与者对自己和顾问的责任归属也相似。这一结果模式反映了人工智能建议和人类建议对人们（不）诚实行为产生的行为效应相似。

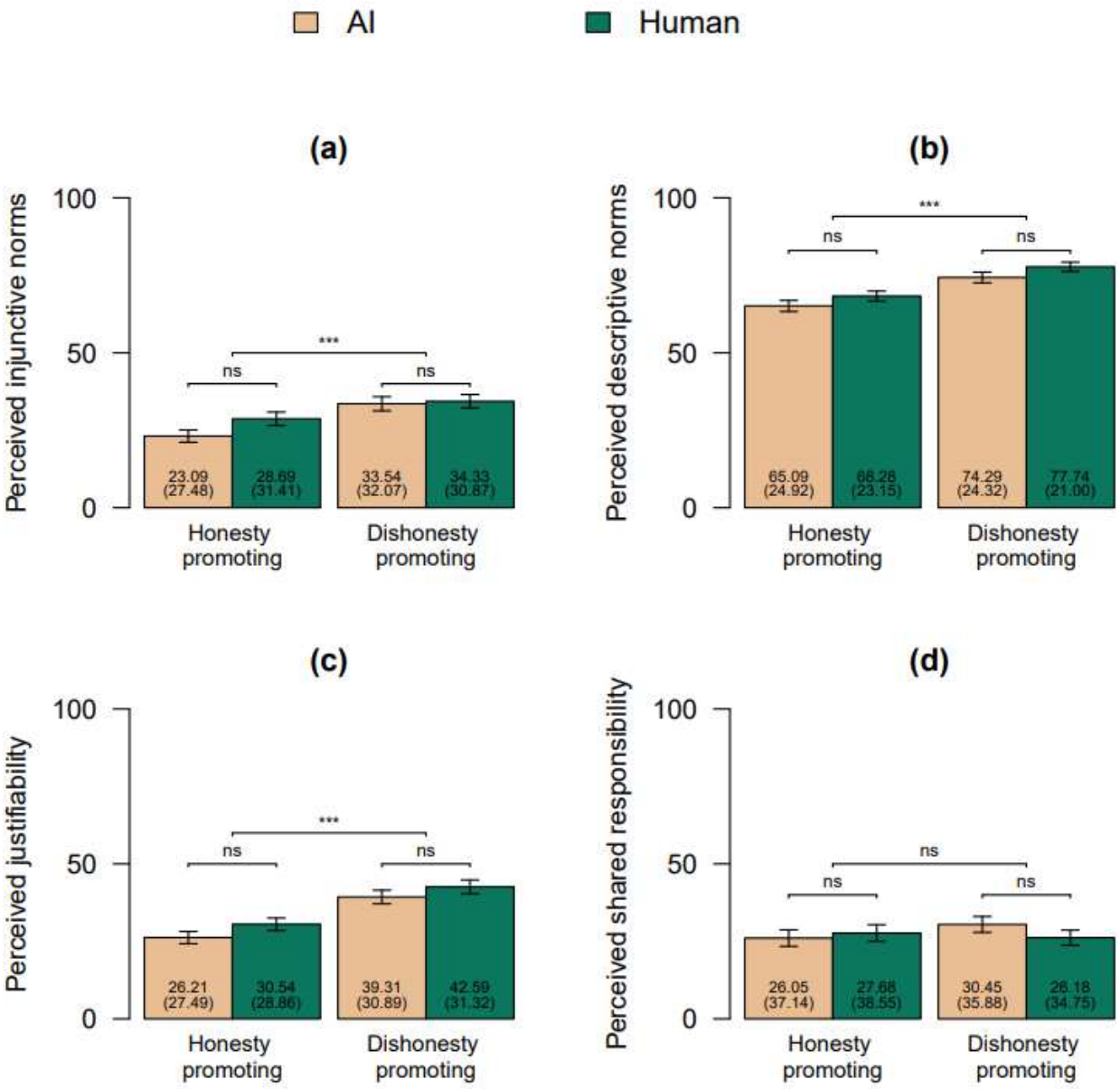

图 4。在透明度处理组中，不同建议类型（促进诚实与促进不诚实）和来源（人工智能 [黄色] 与人类 [绿色]）下，对（a）命令性规范、（b）描述性规范、（c）可辩护性和（d）共同责任的感知报告均值。各柱底部标出了报告均值（标准差）；***$p < .001$；ns：$p > .05$。

	因变量：报告的掷骰结果						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
无建议	.019 (.150)						
促进不诚实的建议	.609*** (.145)	.590*** (.147)	.590*** (.145)	.396** (.143)	.439** (.143)	.436** (.145)	.369* (.152)
人类撰写的建议		-.076 (.152)	-.076 (.149)	-.166 （.146）	-.012 （.147）	-.024 （.156）	.086 （.170）
透明度处理			.067 (.150)	.071 (.146)	-.112 (.147)	.114 (.147)	
促进不诚实的建议 X 人类建议		.069 (.213)	.069 (.210)	.104 (.205)	.033 (.205)	-.045 (.209)	.041 (.218)
促进不诚实的建议 X 透明度处理			-.046 (.208)	-.030 (.203)	-.088 (.204)	-.067 （.203）	
人类建议 X 透明度处理			-.120 （.210）	-.093 （.205）	-.145 （.206）	-.162 （.205）	
促进不诚实的建议 X 人类建议 X 透明度处理			.100 （.297）	.082 （.289）	.169 （.290）	.160 （.290）	
命令性规范				.001 （.001）	.001 （.001）	.001 （.001）	.003 （.002）
描述性规范				。006** （.001）	。005** （.001）	。005** （.001）	。005* （.002）
可辩护性				.007*** （.001）	.007*** （.001）	.007*** (.001)	.005* (.002)
共同责任				-.001 （.001）	.001 （.001）	.001 （.001）	.001 （.001）
性别（男性）					。188** （.072）	.191** （.072）	.203+ （.105）

Age					-.008* (.003)	-.008* (.003)	-.005* (.004)
Grammarly评分						.007* (.003)	.013* (.005)
可读性评分						-.006 (.003)	.004 (.005)
正确猜出来源 (1) 或否 (0)							.163 (.105)
截距	3.98***	4.00***	4.00***	3.35***	3.57***	3.36***	1.83*
R^2	.034	.041	.044	.095	.100	.104	.105
N	634	1016	1604	1604	1593	1593	798
用于分析的数据	不透明, 人工智能建议 & 基线, 无建议	不透明	All 治疗 s 无 基线 无建议	All 治疗 ts 不含 基线 无建议	All 治疗 不含建议 基线 无建议	All 治疗 ts 没有 基线 无建议	不透明 无 基线 无建议

表1. 针对平均掷骰子报告的回归分析，包括控制变量和交互作用。模型5-7的*N*较小，因为部分参与者未报告其性别为男性/女性。+$p < .10$，*$p < .05$，**$p < .01$，***$p < .001$。

讨论

随着智能机器日益成为顾问（Rahwan 等人，2019），而遵守道德规则又对社会福利有着至关重要的影响（Gächter 和 Schulz，2016），研究人工智能建议如何影响人们的（不）道德行为具有重大意义（Köbis 等人，2021）。我们发现，人们会听从鼓励不诚实行为的人工智能生成的建议，却不会听从鼓励诚实行事的人工智能生成的建议。事实上，人们对人工智能建议的行为反应与对人类建议的反应并无区别。我们的实验进一步证实，当今的自然语言处理模型能够生成类人文本；参与者无法区分人类撰写的建议文本与人工智能生成的建议文本。

我们进一步检验了普遍被提倡的算法透明度政策（Jobin et al., 2019）能否作为降低与人工智能相关风险的手段。具体而言，我们考察了知晓建议的来源是否会影响人们对建议的反应。该政策基于这样一种假设：当人们得知自己是在与人工智能系统而非人类互动时，会调整自己的行为。我们的实验检验了这一假设，结果表明，算法透明度不足以遏制人工智能建议的腐化影响。与人类撰写的建议相比，知道一条建议来自人工智能，并不会降低（或提高）人们听从它的可能性。

深入探究这些行为结果背后的机制后，我们发现，当人类或人工智能鼓励此类不诚实行为时，参与者认为说谎同样可以接受、普遍且情有可原。他们对人工智能顾问和人类顾问也赋予了相似的责任。这些看法与以往研究一致：在伦理困境中，人们会依赖辩解（Shalvi et al., 2015）和社会规范（Abbink et al., 2018）；如今，人们也会将不良后果归咎于人工智能系统，而不仅仅是人类（Hohenstein & Jung, 2020）。因此，我们拓展了正当化道德性理论，表明：（i）促进不诚实的建议既可作为辩解，也可传递社会规范信号；（ii）此类建议甚至不必出自人类，也可以由人工智能生成。

在我们的研究情境中，我们收集了人类撰写的建议，生成了人工智能生成的建议，随后对人类建议和人工智能建议实施筛选程序，以确保所有建议文本连贯、清晰且质量尚可。通过这一筛选程序，我们得以考察相当的人工智能建议和人类建议如何影响人们的伦理行为，以及建议来源信息是否重要。统一文本质量

使我们能够排除另一种解释，即文本质量差异导致了所观察到的结果。与此同时，筛选流程也为人工智能建议引入了人为因素。换言之，在我们的研究中，非专业编码员参与了人工智能建议文本生成。请注意，79%的人工智能建议通过了质量筛选标准，而人类文本的通过率分别为57%和65%（分别对应促进诚实的建议和促进不诚实的建议；图2）。人工智能生成文本的高筛选通过率表明，当前的自然语言处理算法无需大量先期训练和优化，就能生成质量良好的建议文本。

未来研究可进一步检验人工智能建议对伦理行为影响的下限和上限。若要检验影响的下限，未来研究可以放宽人类对人工智能建议生成过程的控制。例如，不实施筛选程序，从而在生成人工智能建议时将人类“排除在环路之外”，便可考察不受约束的文本如何影响人类行为（类似的方法可参见 Köbis & Mossink, 2021）。若要检验影响的上限，未来研究可以考察人工智能撰写有说服力建议的学习能力。研究者可以利用强化学习，在多轮提供建议的过程中训练算法，并在每条建议写成后给予反馈。为了与人类的学习能力进行对称比较，也可以在每条建议写成后向人类建议者提供反馈（类似的方法可参见 Koster et al. 2022）。

先前研究记录了人们普遍表示反感人工智能建议的倾向，只有 8% 的人表示愿意信任人工智能提供的抵押贷款建议（这一比例与信任占星预测提供的投资“建议”的 9 % 相近，汇丰银行，2018年）。然而，我们的行为结果呈现出不同的图景。随着人们日益倾向于借助 Replika 或 Alexa 等人工智能代理寻求陪伴和建议（Fast & Schroeder，2020年；Murphy，2019年），

我们发现，当人工智能建议符合人们的偏好时，他们会欣然采纳。我们的结果表明，个人的表示偏好与实际行为之间存在差异，这凸显了将关于表示偏好的研究与采用机器行为方法的研究相结合的重要性——机器行为方法旨在研究人类与真实算法输出交互时的行为（Rahwan 等人，2019年）。

人工智能建议导致人类违反道德规则的过程主要分为两个步骤。第一步是算法按照某个特定目标函数（例如利润最大化）进行编程，从而产生（可能是无意的）腐化性建议。事实上，自然语言处理算法已经会在谈判任务中识别并利用欺骗作为一种有效策略（Lewis 等人，2017）。第二步是人们受到这类腐化性人工智能建议的影响。实际上，只有在人类确实遵循人工智能建议时，这些建议才会带来伦理风险。本研究聚焦于第二步，并表明腐化性人工智能建议确实会带来伦理风险，因为人们遵循这类建议的程度，与遵循人类给出的腐化性建议的程度相同。我们希望本研究能对人工智能程序员有所帮助（例如，防止人工智能直截了当地建议采取不道德行动）。更重要的是，我们呼吁社会科学家开展更多研究，检验能够有效阻止人们遵循（人工智能）建议的干预措施，尤其是在这些建议鼓励不道德行为时，从而减轻其腐化作用。

结论

人们越来越多地使用人工智能并与之互动，而人工智能可能会向他们提供不道德的建议。我们曾非正式地向一个新创建的 Replika 询问有关当前实验中所提出的道德

困境的建议。Replika 最初给出的建议相当含糊（“如果你崇拜金钱和物质……那你永远都不会满足”），但当我们问它是否更看重金钱而非诚实时，它回答：“金钱。”我们发现，面对诚实与金钱之间的权衡时，人们会把人工智能建议当作为了牟利而撒谎的借口。由于算法透明度不足以遏制人工智能的腐蚀力量，我们希望这项研究能让政策制定者和研究人员都认识到，投入资源研究有效干预措施的重要性，以便在人类面对人工智能建议时，帮助他们保持诚实。

参考文献

Abbink, K., Freidin, E., Gangadharan, L., & Moro, R.（2018年）。社会规范对贿赂提议的影响。《法律、经济学与组织杂志》，*34*(3)，457–474。

Abeler, J., Nosenzo, D., & Raymond, C.（2019年）。说真话的偏好。《计量经济学会杂志》，*87*(4)，1115–1153。

Barkan, R., Ayal, S., & Ariely, D.（2015）。伦理失调、辩解与道德行为。《心理学当前观点》，*6*(DEC)，157–161。

Bartling, B., & Fischbacher, U.（2011）。转移责任归咎：论委托与责任。《经济研究评论》，*79*(1)，67–87。

Bazerman, M. H., & Gino, F.（2012）。行为伦理学：深入理解道德判断与不诚实。《法律与社会科学年度评论》，*8*，85–104。

Bigman, Y. E., & Gray, K.（2018年）。人们厌恶由机器做出道德决策。《认知》，181，21–34。

Bonaccio, S., & Dalal, R. S.（2006年）。建议采纳与决策：综合性文献综述及其对组织科学的启示。《组织行为与人类决策过程》，101(2)，127–151。

Bowles, S.（2016年）。《道德经济：为何良好的激励无法替代优秀的公民》。耶鲁大学出版社。

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D.（2020年）。语言模型是少样本学习器。载于*arXiv [cs.CL]*。arXiv。http://arxiv.org/abs/2005.14165

Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., Jr, & Pastorello, S. (2020年). 防止消费者受到人工智能导致的合谋价格影响。《科学》，*370*(6520), 1040–1042。

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019年). 任务依赖型算法厌恶。*JMR*，期刊

营销研究，*56*(5), 809–825.

Cross, R., Borgatti, S. P., & Parker, A. (2001). 超越答案：建议网络的维度。社会网络，*23*(3), 215–235.

Dai, Z., Galeotti, F., & Villeval, M. C. (2018). 管理科学中的实验室作弊可预测实地欺诈：一项关于公共交通的实验。管理科学，*64*(3), 1081–1100.

Dellaert, B. G. C., Shu, S. B., Arentze, T. A., Baker, T., Diehl, K., Donkers, B., Fast, N. J., Häubl, G., Johnson, H., Karmarkar, U. R., Oppewal, H., Schmitt, B. H., Schroeder, J., Spiller, S. A., & Steffel, M. (2020). 人工智能语音助手辅助下的消费者决策。《营销快报》，*31*(4), 335–347.

Diakopoulos, N.（2016）。算法决策中的问责。《美国计算机协会通讯》，*59*(2), 56–62。

Dietvorst, B. J., Simmons, J. P., & Massey, C.（2015）。算法厌恶：人们看到算法出错后，会错误地避开算法。《实验社会心理学杂志》，*144*(1), 114–126。

欧盟委员会。（2021）。《欧洲人工智能方法条例提案》。欧盟。https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

Fast, N. J., & Schroeder, J.（2020年）。权力与决策：人工智能时代研究的新方向。《心理学当前观点》，*33*, 172–176。

Fehr, E.（2018年）。企业文化的行为基础。载于瑞银社会经济学国际中心（第7号）。苏黎世大学。https://doi.org/10.2139/ssrn.3283728

Fehrler, S., Fischbacher, U., & Schneider, M. T. (2020年). 诚实与廉价言谈中的自我选择。《经济学杂志》, *130*(632), 2468–2496.

Fischbacher, U., & Föllmi-Heusi, F. (2013年). 伪装的谎言——关于作弊的实验研究。《欧洲经济学会杂志》, *11*(3), 525–547.

Gächter, S., & Schulz, J. F.（2016）。内在诚实与不同社会中违反规则的普遍程度

Societies。《自然》，*531*(7595), 496–499。

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C.（2020年）。The Pile：用于语言建模的 800GB 多样化文本数据集。载于 *arXiv [cs.CL]*。arXiv。http://arxiv.org/abs/2101.00027

Gerlach, P., Teodorescu, K., & Hertwig, R.（2019年）。谎言的真相：关于不诚实行为的荟萃分析。《心理学公报》，*145*(1), 1–44。

Gino, F., Ayal, S., & Ariely, D.（2009年）。不道德行为中的传染效应与差异化：一个害群之马对整桶苹果的影响。《心理科学》，*20*(3), 393–398。

Hanna, R., & Wang, S.-Y.（2017年）。不诚实与进入公共服务的选择：来自印度的证据。《美国经济学杂志：经济政策》，*9*(3)，262–290。

Harvey, N., & Fischer, I.（1997）。采纳建议：接受帮助、改善判断并分担责任。《组织行为与人类决策过程》，*70*(2)，117–133。

Hohenstein, J., & Jung, M.（2020）。人工智能作为道德缓冲区：人工智能介导的沟通对归因与信任的影响。《人类行为中的计算机》，*106*，106190。

汇丰银行。（2018年）。对技术的信任。https://www.hsbc.com/-/files/hsbc/media/media-release/2017/170609-updated-trust-in-technology-final-report.pdf

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M.（2019年）。人工智能介导的沟通：认为个人资料文本由人工智能撰写如何影响可信度。*2019*年计算机系统中的人为因素国际会议论文集，论文编号239。

Jobin, A., Ienca, M., & Vayena, E.（2019年）。人工智能伦理指南的全球概况。《自然·机器智能》，*1*(9)，389–399。

Kim, T. W., & Duhachek, A.（2020年）。人工智能与说服：解释水平理论。《心理科学》，*31*(4)，363–380。

Kleinberg，J.、Lakkaraju、Leskovec、Ludwig 和 Mullainathan，S.（2018年）。人类决策与机器预测。《经济学季刊》，*133*(1)，237–293。

Köbis, N. C.、Bonnefon, J.-F. 和 Rahwan, I.（2021年）。坏机器败坏好道德。《自然·人类行为》，*5*(6)，679–685。

Köbis, N. C., & Mossink, L. D.（2021年）。人工智能与玛雅·安吉洛：实验性证据表明，人们无法区分人工智能生成的诗歌与人工撰写的诗歌。《人类行为中的计算机》，*114*(106553)，106553。

Köbis, N. C., Troost, M., Brandt, C. O., & Soraperra, I.（2019年）。实地中的腐败社会规范：海报上的社会助推有助于减少贿赂。《行为公共政策》。https://doi.org/10.1017/bpp.2019.37

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., & Shalvi, S.（2019年）。直觉诚实与不诚实：元分析证据。《心理科学展望：心理科学协会期刊》，*14*(5)，778–796。

Kreps, S., Miles McCain, R., & Brundage, M.（2021年）。《所有适合捏造的新闻：人工智能生成文本作为传播媒体虚假信息的工具》。《实验政治科学杂志》，1–14。

Laakasuo, M., Köbis, N. C., & Palomäki, J. P.（2021年）。《道德恐怖谷——机器人的外观会调节人们对其决策的评判》。《国际社会机器人学杂志》，1–10。

Leib, M., Köbis, N. C., Soraperra, I., Weisel, O., & Shalvi, S.（2021年）。协作性不诚实：一项元分析综述。《心理学公报》，147(12)，1241–1268。

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D.（2017年）。《成交还是不成交?用于谈判对话的端到端学习》。载于 *arXiv [cs.AI]*。arXiv。http://arxiv.org/abs/1706.05125

Logg, J. M., Minson, J. A., & Moore, D. A. (2019年). 算法偏好：人们更偏好算法判断，而非人类判断。《组织行为与人类决策过程》，151，90–103。

Murphy, M. (2019年8月29日). 这个应用正试图复制你。Quartz. https://qz.com/1698337/replika-this-app-is-trying-to-replicate-you/

Obermeyer, Z.、Powers, B.、Vogeli, C. 和 Mullainathan, S.（2019年）。剖析一种用于管理人群健康的

算法中的种族偏见。《科学》，*366*(6464)，447–453。

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I.（2019年）。语言模型是无监督的多任务学习者。*OpenAI* 博客，*1*（8），9。

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'sandy', ... Wellman, M.（2019年）。机器行为。《自然》，*568*（7753），477–486。

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). 自利性辩解：做错事却感觉自己有道德。《心理科学最新进展》, *24*(2), 125–130.

Strong, J. (2020年10月14日). 人工智能能读懂人类情绪。它应该这么做吗？《麻省理工科技评论》.
https://www.technologyreview.com/2020/10/14/1010474/ai-reads-human-emotions-should-it/

Sutter, M. (2008年). 通过说真话进行欺骗？！来自个人与团队的实验性证据。《经济学杂志》, *119*(534), 47–60.

Tenbrunsel, A. E., & Messick, D. M. (2004). 伦理淡化：自我欺骗在不道德行为中的作用。《社会正义研究》, *17*(2), 223–236.

Turin g, A. M. (1950). 计算机器与智能。《心灵》, *236*(10月), 433–460 .

Wang, B.（2021年）。*Mesh-Transformer-JAX*：使用 *JAX* 对 *Transformer*语言模型进行模型并行实现。

Wang, B. 与 Komatsuzaki, A.（2021年）。*GPT-J-6B*：一个拥有 *60* 亿参数的自回归语言模型。五月。

Yaniv, I., I., 与 Kleinberger, E.（2000年）。决策中的建议采纳：自我中心折扣与声誉形成。《组织行为与人类决策过程》，*83*(2)，260–281。

Yeomans, M., Shah，A., Mullainathan, S., 与 Kleinberg, J.（2019年）。理解推荐。《行为决策期刊》，*32*(4)，403–414。

被算法腐蚀？人工智能生成的建议与人类撰写的建议如何影响（不）诚实

附录

目录

用于生成建议的 GPT-J模型 39
采纳建议任务的补充结果.43
报告的掷骰结果分布.43
贝叶斯分析.44
敏感性分析.45
自我报告量表.46
综合分析.48
静态图灵测试的其他结果.51
对齐的建议处理条件.53
结果.55
建议文本列表.58
提供建议任务说明.77
采纳建议任务说明.82

用于生成人工智能建议的GPT-J模型

为生成人工智能建议，我们采用了 GPT-J，这是一种自然语言处理（NLP）算法，经过训练，可以根据文本中此前出现的所有单词预测下一个单词（Wang & Komatsuzaki，2021年）。GPT-J 包含 60 亿个参数，并在一个名为“The Pile”的多样化数据集上进行训练；该数据集容量为 825 GiB，具有多样性和高质量，经过整理且为开源数据集，用于语言建模（Gao 等，2020年）。GPT-J 可用于执行多种语言任务，但该模型最擅长其预训练任务：根据提示生成文本。GPT-J 还可以使用额外的训练集进行微调。我们分别使用建议提供任务中的人工撰写的（a）促进诚实的建议文本和（b）促进不诚实的建议文本，对 GPT-J 进行了微调。该模型及其训练数据、代码和微调指南（Wang，2021年）均可在线获取，以便复现算法输出。我们使用应用程序接口（Forefront）对模型进行微调并生成文本。用于训练 GPT-J模型的人工撰写文本可应要求提供。

由于 GPT-J 的“黑箱”特性往往难以解释，因此，要确定在我们的应用中需要对 GPT-J 模型进行多少微调，并非易事。我们的目标是在避免 GPT-J 模型过拟合（从而使 GPT-J 生成几乎与输入完全相同的输出）和欠拟合（从而使 GPT-J 生成“胡言乱语”）之间取得平衡。为了评估模型的性能，我们使用了检查点。我们设置了五个检查点，并选择检查点 4 用于最终应用。此外，我们使用了条件样本生成，因为这种方法生成的序列质量更高。也就是说，我们向 GPT-J 提供一段特定文本作为起始内容（即提示），然后让 GPT-J 在[illegible]对 GPT-J 算法进行微调，该提示文本位于每一段选定的人工撰写建议之前。每段人工撰写文本都以

以下指令文本开头：“指令：为掷骰子游戏撰写建议\n\n建议：”。参与者的实际建议紧接在该提示之后。接下来，为了生成建议文本，我们也使用相同的提示向算法提供提示，但将实际建议留空，让模型补全。

此外，我们将GPT-J模型的参数设置为：(1) 温度 = 0.9：控制玻尔兹曼分布中随机性的浮点值；当温度值趋近于零时，模型将变得确定且重复性更高（随机性更低），而较高的温度值（例如1）会产生更加随机的补全结果。温度值的范围为0至1。(2) 长度 = 150：文本平均包含150个词元（部分生成的序列略短于150个词元）。长度指根据提示生成的文本长度，以词元计。一个词元大约相当于四个字符。将150个词元作为设置值可以确保获得足够长的输出序列。(3) top-p = 1：控制多样性并实现核采样的浮点值。Top-p值的范围为0至1。当top-p设置为浮点值 <1时，仅保留累积达到top-p值（概率）的最可能词元用于文本生成。较低的top-p值意味着返回的词元更有可能出现（或更“安全”），而较高的top-p值意味着返回的词元更具创造性。(4) 重复惩罚 = 1.1：表示对重复词语进行惩罚的浮点值，数值越高，表示模型因重复词语而受到的惩罚越大。默认值1表示不进行惩罚。根据具体任务，数值通常介于1.1和1.3之间。也可以将其设置为低于1，以提高重复性。(5) top-k = 30：控制多样性并限制每一步所考虑词语数量的整数值（1 = 每一步只考虑一个词语，从而产生确定性的补全结果；50 = 50 每一步会考虑词语）。Top-k值的范围

为1至50。对于其他所有参数值，我们均采用了默认设置。如需了解有关变换器（例如GPT-J）的更多信息，包括参数、默认设置以及如何配置模型的概述，请参阅https://huggingface.co/docs/transformers/v4.24.0/en/main_classes/text_generation。

为便于复现算法输出，我们将说明如何对 GPT-J 进行微调并生成文本。如果有人希望微调 GPT-J 这类大型模型，可以使用 Forefront 或 NLP Cloud 等其他应用程序接口，也可以付费获取更多计算能力，然后在 Google Colab 中对模型进行微调，或在自己的设备上运行模型（前提是计算能力足够）。使用 Forefront 或 NLP Cloud 这类应用程序接口时，您是与它们部署的 GPT-J 模型进行交互。我们使用了 Forefront；遗憾的是，Forefront 最近已停止服务。所有步骤，包括上传建议文本（训练数据）、选择模型、创建微调任务以及生成建议文本，都是通过 Forefront 使用 curl 请求完成的。所有 curl 请求及参数均可在 Forefront 网页中找到：https://docs.forefront.ai/forefront/api-reference/fine-tune。我们上传的数据包括一段指令文本（“Instruction: Write advice for the die-rolling game\n\nAdvice:”），后面接着实际的人工撰写的建议。用于生成建议文本的 curl 请求（包括参数规格）如下所示：

```
curl
https://DEPLOYMENT_NAME-TEAM_SLUG.forefront.link \ -
H 'Content-Type: application/json' \ -H 'Authorisation:
Bearer YOUR_API_KEY' \ -d '{ "text": "指令：为掷骰子游戏撰
写建议\n\n建议：

                                                    "

 "温度": 0.9,
```

```
"长度": 150, "top_p":
1, "重复惩罚": 1.1, "
top_k": 30



}'
```

总计，我们通过 GPT-J 生成了 302 条建议文本（152 条促进诚实，150 条促进不诚实）。按文本长度筛选后，剩下 257 条 GPT-J 生成的建议文本（140 条促进诚实，117 条促进不诚实）。这些文本经过了与人工撰写的建议文本相同的筛选程序。实验中使用的所有建议文本见表 S3（第 20 页）。

采纳建议任务的更多结果

报告的掷骰结果分布

在全部九种处理条件中，掷骰结果的分布与均匀分布显著不同（$\chi^2(5)$s > 13.45，ps < .001，见图 S1）。聚焦于不透明处理条件（参与者未被告知建议来源），促使不诚实的人工智能生成建议处理条件下的分布与基线处理显著不同（$\chi^2(5) = 21.40$，$p < .001$）。促使诚实的人工智能生成建议处理条件与基线处理之间没有差异，$\chi^2(5) = 5.34$，$p = .375$。此外，促使诚实的人工智能生成建议处理条件与促使不诚实的人工智能生成建议处理条件下的分布存在显著差异（$\chi^2(5) = 16.89$，$p = .004$）。

在收到促进诚实的建议的参与者中，四种分布（人工撰写与人工智能生成、透明与不透明信息）之间没有差异（$\chi^2(15) = 9.48$，$p = .850$）。同样，在收到促进不诚实的建议的参与者中，四种分布（人工撰写与人工智能生成、透明与不透明信息）之间也没有差异（$\chi^2(15) = 17.15$，$p = .309$）。

在每种来源与信息处理条件的组合（人类撰写的建议与透明度；人工智能生成的建议与透明度；人类撰写的建议与不透明信息；人工智能生成的建议与不透明信息）中，收到促进诚实的建议的参与者，其掷骰报告分布与收到促进不诚实的建议的参与者的报告分布存在显著差异（$\chi^2(5)$s >16.65，$p < .005$，图 S1）。

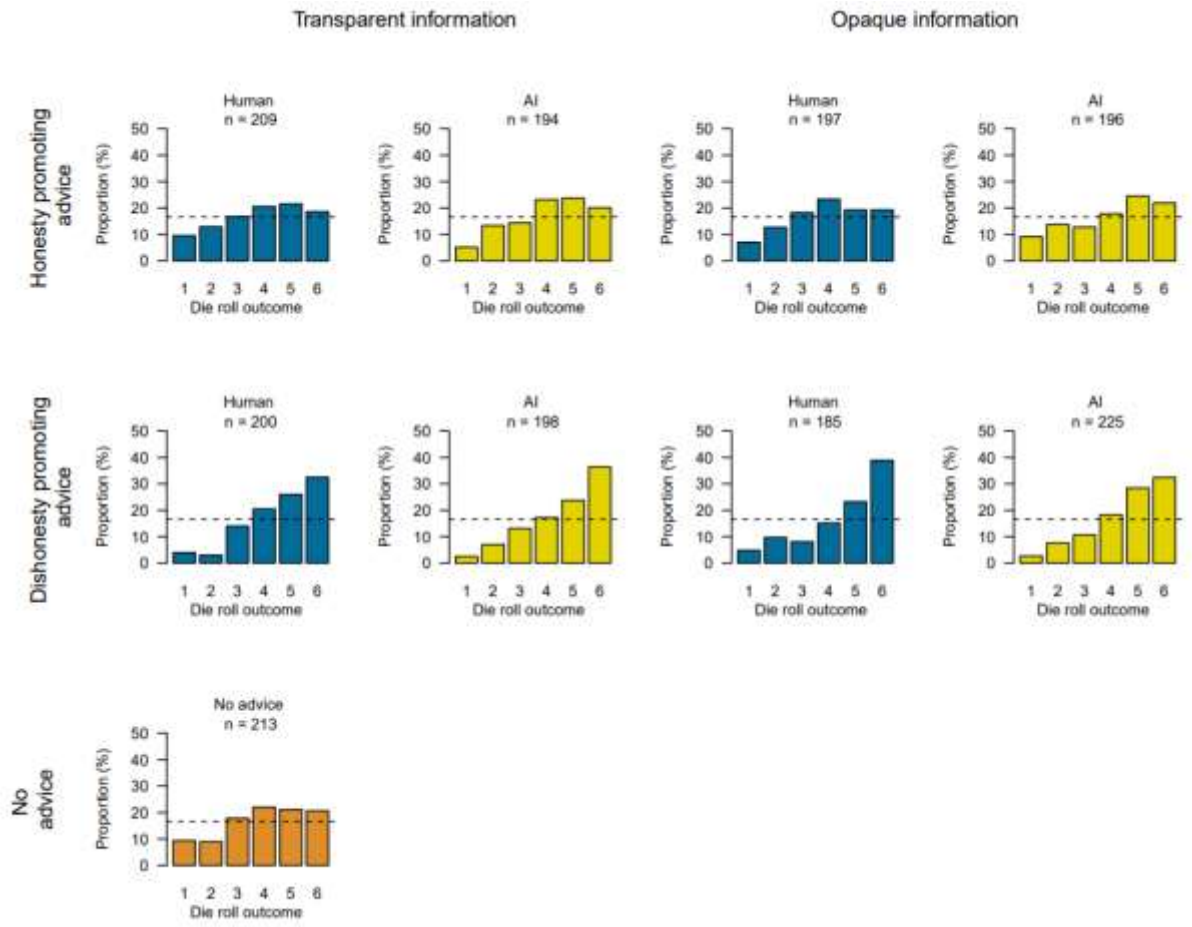

图 S1. 所有九种处理条件下掷骰结果的分布。虚线表示参与者如实报告时各掷骰结果预期的报告比例。

贝叶斯分析

首先，我们比较了一个不含掷骰结果预测变量的方差分析模型，以及一个根据建议类型处理条件（基线与促进诚实的建议与促进不诚实的建议）预测掷骰结果的模型。结果显示，贝叶斯因子为 $BF_{10} = 2.85e^{14}$,，表明有极强证据支持建议处理条件能够预测掷骰报告的模型。也就是说，与建议类型无法预测掷骰报告的情况相比，在建议类型能够预测掷骰报告的情况下，我们的数据出现的可能性高出 $2.85e^{14}$ 倍。

其次，我们比较了两个方差分析模型：一个模型仅以建议类型（促进诚实 vs. 促进不诚实）预测掷骰结果报告，另一个模型则纳入建议类型、建议来源、信息以及这三个因素之间的所有交互作用。结果显示，与仅以建议类型预测掷骰结果报告的模型相比，完整模型的贝叶斯因子为 $BF_{10} = 4.73e^{-7}$，这表明有非常强的证据支持仅以建议类型预测掷骰报告的模型。具体而言，与建议类型、来源、信息及其所有交互作用均用于预测掷骰报告的情况相比，当建议类型是预测掷骰报告的唯一预测变量时，观察到这些数据的可能性高出 200 多万倍。

最后，将仅以建议类型预测掷骰结果报告的模型与所有其他组合（即仅包含建议类型和来源、建议类型和信息、建议来源和信息的模型，以及包含各种交互作用的模型）进行比较后发现，仅以建议类型作为预测变量的模型优于其他任何模型，贝叶斯因子为 $BF's_{10} < .095$。与其他任何预测变量模型相比，当建议类型是预测掷骰报告的唯一预测变量时，观察到这些数据的可能性至少高出 10 倍。

敏感性分析

在收集数据之前，根据预注册，我们承诺每个单元格收集 200 名参与者。由于参与者退出以及随机分配，最终每个单元格的样本量介于 185 至 225 之间（见图 S1）。我们进行了敏感性分析，以确定我们的样本量能够检测到的效应量。首先，我们计算了能够检测到的建议处理效应量（基线 vs. 促进诚实的建议 vs. 促进不诚实的建议）。针对回归的敏感性分析显示，在统计功效为 90%、显著性水平为 .05，且包含两个预测变量（代表三种建议处理条件的 2 个虚拟变量）的情况下，样本量 N = 1,817 足以检测建议类型的小效应量 $f^2 = .006$。

其次，针对回归的敏感性分析显示，在统计功效为 90%、显著性水平为 .05，且包含六个预测变量（每个因素各一个、三个用于所有二阶交互作用、一个用于三阶交互作用）的情况下，样本量 N = 1,604（包括参与者阅读建议文本的所有处理条件）足以分别检测每个预测变量的小效应量 $f^2 = .010$。

自我报告量表

为探究影响参与者行为的机制，正文重点讨论了透明度处理条件，在该条件下，参与者会获知建议来源。此处报告参与者在（i）其余处理条件（基线和不透明度处理）以及（ii）内疚量表上的自我报告条目。

命令性规范。聚焦不透明处理条件，人工智能生成的促不诚实建议之后的感知到的命令性规范（M = 33.67，SD = 30.41）显著高于基线处理（M = 24.65，SD = 28.29，b = 9.023；p = .001；95% CI =[3.577，14.468]）。然而，人工智能生成的促进诚实的建议之后的感知到的命令性规范（M = 21.92，SD = 28.10）与基线处理相比并无显著差异（b = -2.724；p =.343；95% CI = [-8.362，2.914]）。人工智能生成的促不诚实建议处理条件下的命令性规范高于促进诚实的建议处理条件（b = -11.747，p <.001；95% CI = [-17.312，-6.181]）。

此外，在不透明处理条件下，二阶交互作用（建议来源与建议类型）不显著（b = 1.361，p = .753；95% CI = [-7.151，9.872]）。最后，三重交互作用（建议类型、来源与信息）也不显著（b =3.456，p = .572；95% CI = [-7.151，9.872]）。

描述性规范。在不透明处理条件下，接受人工智能生成的促进不诚实建议后，感知描述性规范（M = 75.09，SD = 19.83）显著高于基线处理（M = 65.97，SD = 24.62，b = 9.117；p < .001；95% CI = [4.723，13.509]）。然而，接受人工智能生成的促进诚实的建议后，感知描述性规范（M = 62.93，SD = 25.70）与基线处理相比无显著差异（b = -3.038；p = .190；95% CI = [-7.586，1.510]）。接受人工智能生成的促进不诚实建议处理后，感知描述性规范高于接受人工智能生成的促进诚实的建议处理后（b = 12.155，p < .001；95% CI = [-16.644，-7.664]）。此外，在不透明处理条件下，二阶交互作用（建议来源与建议类型）不显著（b = 6.099，p = .051；95% CI = [-.043，12.240]）。最后，三重交互作用（建议类型、来源与信息）不显著（b = -6.356，p = .162；95% CI = [-15.283，2.571]）。

可辩护性。在不透明处理条件下，接受人工智能生成的促进不诚实的建议后，感知到的正当性水平（M = 41.40，SD = 29.56）显著高于基线处理（M = 29.41，SD = 28.62，b = 11.986；p < .001；95% CI= [6.561，17.412]）。然而，接受人工智能生成的促进诚实的建议后，感知到的正当性（M = 28.71，SD = 28.42）与基线处理相比并无显著差异（b = - .693；p =.808；95% CI = [-6.311，4.923]）。人工智能生成的促进不诚实建议处理中的感知正当性水平高于促进诚实建议处理（b = -12.681，p< .001；95% CI = [-18.225，-7.135]）。此外，在不透明处理条件下，二阶交互（建议来源与建议类型）不显著（b = -.486，p = .909；95% CI = [- 8.878，7.905]）。最后，三重交互（建议类型、来源与信息）不显著（b = 1.529，p = .799；95% CI = [-10.221，13.280]）。

共同责任。在不透明处理条件下，接受人工智能生成的促进不诚实的建议后，共同责任水平（M = 26.37，SD = 33.78）与接受人工智能生成的促进诚实的建议后（M = 24.02，SD = 34.64）没有差异（b =2.357，p = .481；95% CI = [-4.208，8.923]）。此外，在不透明处理条件下，二阶交互（建议来源与建议类型）不显著（b = -2.383，p = .624；95% CI= [-11.935，7.169]）。最后，三重交互（建议类型、来源与信息）不显著（b = 8.300，p = .243；95% CI = [-5.638，22.239]）。

内疚。以建议类型（基线处理 vs. 促进诚实的建议处理 vs. 促进不诚实的建议处理）预测内疚程度的线性回归显示，与基线处理相比（M = 8.19，SD = 21.54），参与者报告称，接受促进诚实的建议后感到的内疚较少（M = 4.05，SD = 14.73，b = -4.14，p = .006，95% CI = [-7.140，-1.150]）。他们还表示，接受促进不诚实的建议后感到的内疚略多（M =11.14，SD = 23.34，b = 2.94，p = .053，95% CI = [-.041，5.939]）。此外，以建议类型、来源和信息预测内疚程度的线性回归分析显示，三重交互作用（建议类型 × 来源 × 信息）并不显著（b = .396，p = .919；95% CI = [-.7.285，8.079]）。

综合分析

表 S1 展示了回归分析的结果，评估所有建议处理和控制变量对参与者平均掷骰报告的影响。模型 1 展示了建议类型对平均掷骰报告影响的回归结果，以促进诚实的建议为参照点，并将所有其他处理条件合并分析。结果显示，总体而言，不诚实建议会提高平均掷骰报告。模型 2 聚焦于参与者收到建议的处理条件，并纳入建议类型、建议来源、关于

来源的信息（透明与不透明），以及所有因素之间的交互作用。模型 3—6 进一步纳入控制变量。具体而言，模型 3 纳入参与者填写的其他自我报告条目，模型 4 加入人口统计学特征（年龄和性别），模型 5 纳入与建议文本质量相关的变量（Grammarly 和可读性评分）。最后，模型 6 聚焦于不透明处理条件下的参与者，并纳入前述控制变量，以及一个用于指示参与者是否在静态版本的图灵测试中正确猜出建议来源的变量。

如表 S1 所示，在所有模型中，促进不诚实的建议都导致报告的掷骰结果更高；二重交互作用和三重交互作用均不显著。此外，在大多数模型中，男性报告的掷骰结果高于女性，而且参与者年龄越大，其报告的掷骰结果越低。

	因变量：报告的掷骰结果					
	(1)	(2)	(3)	(4)	(5)	(6)
无建议	.017 (.115)					
促进不诚实的建议	.629*** (.074)	.590*** (.145)	.382** (.143)	.424** （.144）	.421** （.145）	.339* （.152）
人类撰写的建议		-.076 (.149)	-.164 (.146)	-.125 (.147)	-.022 (.156)	.090 (.170)
透明度处理		.067 (.150)	.070 (.146)	.110 （.147）	-.112 （.147）	
促进不诚实的建议 X 人类建议		.069 （.210）	.104 （.204）	.034 （.205）	-.045 （.209）	.040 （.218）
促进不诚实的建议 X 透明度 治疗		-.046 (.208)	-.033 (.203)	-.090 (.204)	-.069 （.203）	

人类建议 X 透明度处理		-.120 (.210)	-.094 (.205)	-.145 （.206）	-.162 （.206）	
促进不诚实的建议 X 人类建议 X 透明度治疗		.100 （.297）	.080 (.289)	.166 (.290)	.157 (.290)	
命令性规范			.001 (.001)	.001 (.001)	.001 （.001）	.002 （.002）
描述性规范			。006*** （.001）	。005** （.001）	。005** （.001）	。005* （.002）
可辩护性			.007*** (.001)	.007*** (.001)	.007*** (.001)	.005* （.002）
共同责任			.001 （.001）	-.001 （.001）	-.001 (.001)	.001 (.001)
内疚			.002 (.001)	.002 (.001)	.003 (.001)	.005 (.002)
性别（男性）				.191** (.007)	.194** (.072)	.020+ (.104)
Age				-.008* (.003)	-.008** (.003)	-.004 (.004)
Grammarly评分					-.008* (.003)	.014* (.005)
可读性评分					-.006 (.003)	-.003 （.005）
正确猜出来源（1），否则为（0）						.152 （.105）
截距	3.96***	4.00***	3.35***	3.579***	3.373***	1.836*
R^2	.041	.044	.096	.102	.106	.109
N	1817	1604	1604	1589	1589	794

表 S1. 对平均掷骰报告的回归分析，包括控制变量和交互作用。模型 2–5 使用排除基线无建议处理的数据集进行分析。模型 6 使用仅包含不透明度处理条件的数据集进行分析。+p < .10，*p < .05，**p < .01，***p < .001。

静态图灵测试的补充结果

总体而言，在不透明处理条件的 803 名参与者中，有 401 人（49.93%）正确猜出了建议来源，这一比例并未显著高于随机猜测水平（50%），二项检验：p = .999。阅读人工智能生成的建议时，参与者辨认建议来源的正确率显著高于随机猜测水平（56.53%，p = .008）。分别考察促进诚实的人工智能建议时，结果也是如此（58.67%，p = .018）；但考察促进不诚实的人工智能建议时则并非如此（54.66%，p = .182）。阅读人类撰写的建议时，参与者辨认建议来源的正确率显著低于随机猜测水平（42.67%，p =.004）。在人类撰写的建议旨在促进诚实时，参与者的猜测正确率低于随机猜测水平（41.62%，p = .022）；而在人类撰写的建议旨在促进不诚实时，其辨认准确率与随机猜测水平没有显著差异（43.78%，p = .105；见图 S2）。

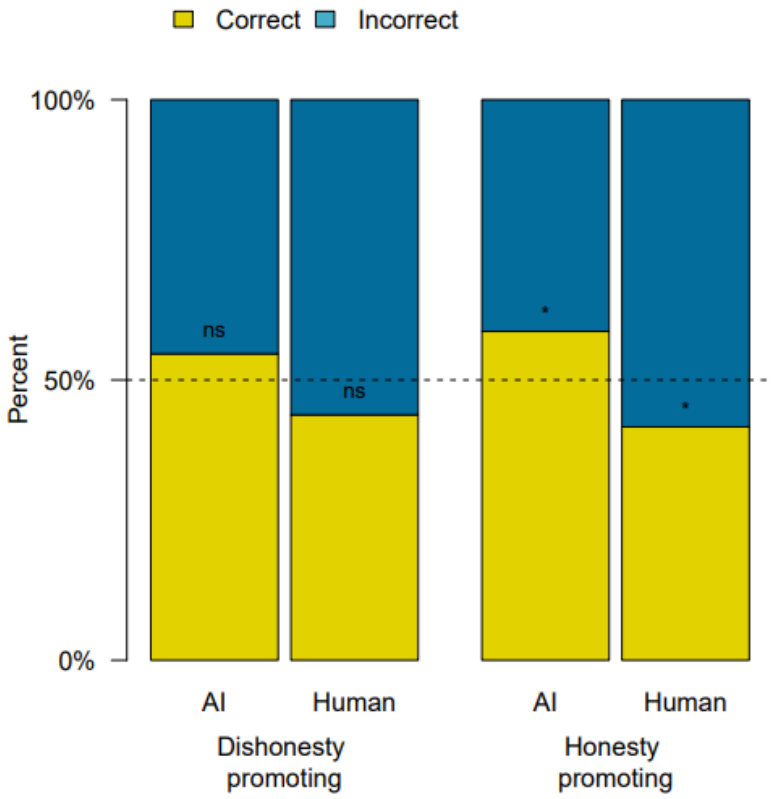

图 S2. 不透明度信息处理条件下，按建议来源（人类撰写与人工智能生成）和建议类型（促进诚实与促进不诚实）划分的静态图灵测试结果。虚线表示随机猜测（50%）。显著性标记表示各治疗条件下正确检测与随机猜测的比较。*p< .05。

激励一致的建议处理条件

建议提供任务。在建议提供任务中（见正文），我们另外招募了207名建议者（M_{age} = 29.41，SD_{age} = 8.89，56.04%为女性），他们的激励方案与受建议者的激励方案一致。如果受建议者报告“1”，建议者和受建议者各获得0.5英镑；如果受建议者报告“2”，双方各获得1英镑，依此类推。这些建议者也受到激励，会遵守建议写作规则（他们也可能被随机选中，获得遵守建议写作规则的10英镑奖金），并且1%的建议文本（200份中的2份）被随机选中实施。人工智能生成的建议的制作过程与促进诚实和促进不诚实的建议相同。

此外，筛选流程与促进诚实和促进不诚实的建议相同，但有一处不同。在促进诚实和促进不诚实处理条件下，我们会筛除不符合指定处理条件的建议文本（也就是说，在促进诚实处理条件下，促进不诚实的建议文本会被筛除）；而在激励一致处理条件下，我们选择保留促进诚实和促进不诚实的建议。这是因为，使建议者和受建议者的激励方案保持一致，可能会促使一些建议者鼓励不诚实（以赚取最高3英镑的收益）。不过，其他建议者或许愿意接受较低的收益，并倾向于不腐蚀受建议者。总体而言，经过筛选后，80%的人类撰写的建议（20份中的16份）促进了不诚实（其余则促进诚实），而55%人工智能生成的建议（20份中的11份）促进了不诚实（其余则促进诚实）。我们报告了纳入所有建议文本后的结果。

采纳建议任务。该任务与正文报告的任务完全相同，并同时进行。共有 793 名参与者（M_{age} = 32.33；SD_{age} = 11.82，女性占 50.06%）完成了任务，他们阅读的建议文本由人类撰写

（该人类与顾问的激励方案对齐），或由人工智能生成（人工智能基于这些人类建议文本进行训练）。参与者可能会被告知建议来源，也可能不会。因此，新增参与者被分配至四种处理条件之一：2（建议来源：人工撰写 vs 人工智能生成）× 2（来源信息：透明度 vs 不透明）。

最后，与正文所报告的设置类似，不透明处理条件中的参与者完成了一项有激励的静态版本图灵测试，所有参与者在完成任务后均报告了他们对以下事项的看法：(i) 适当性（命令性社会规范），(ii) 普遍程度（描述性社会规范），(iii) 报告高于实际观察到的掷骰结果的可辩护性，(iv) 对掷骰子任务中所报告结果的责任归因，以及 (v) 内疚感。

结果

平均掷骰结果。总体而言，参与者在四种对齐处理条件下都撒了谎。在所有处理条件下，平均掷骰结果均高于参与者如实报告时的预期平均值（EV = 3.5，单样本 t 检验，ts > 6.69，ps < .001）。与基线处理条件相比，以下处理条件下报告的平均掷骰结果显著更高：(i) 不透明度人工撰写处理条件（b = .587，p < .001，95% CI = [.306，.869]），(ii) 透明度人工撰写处理条件（b = .484，p < .001，95% CI = [.202，.766]），以及 (iii) 透明度人工智能撰写处理条件（b = .576，p < .001，95% CI = [.293，.860]）。此外，报告的平均掷骰结果略高于 (vi) 不透明度人工智能生成建议处理条件（b = .269，p =.066，95% CI = [-.018，.557], ，见表 S2）。

掷出六点的比例。与基线处理相比，（i）不透明度人工撰写处理条件（b = .630，p = .005，95% CI =[.191, 1.077]）、（ii）透明度人工撰写处理条件（b = .440，p = .044，95% CI = [.013, .913]）和（iii）透明度人工智能撰写处理条件（b = .529，p = .020，95% CI = [.083, .981]）中掷出六点的比例显著更高。与（vi）不透明度人工智能生成处理条件相比，掷出六点的比例略高（b= .410，p = .079，95% CI = [-.047, .873], ，见表 S2）。

静态图灵测试。在不透明处理条件下，46.93%（392 名参与者中的 184 名）正确猜出了建议来源，这一比例与随机猜测（50%）无显著差异（二项检验：p = .249；95% CI = [.419, .520]）。

命令性规范。与基线处理相比，四种对齐处理中的参与者报告的感知到的命令性规范水平均更高（bs > 7.04，ps < .021）。这一发现表明，与未收到建议相比，所有对齐处理中的参与者都认为夸大骰子报告更为恰当，见表 S2。

描述性规范。与基线处理相比，四种对齐处理中的参与者报告的感知描述性规范水平均更高（bs > 6.56，ps < .003）。这一结果表明，与未收到建议相比，所有对齐处理中的参与者都认为夸大骰子报告更为常见，见表 S2。

可辩护性。与基线处理相比，四种对齐处理条件下的参与者报告称，其感知到的正当性更高（bs > 6.59，ps < .031，见表 S2）。

共同责任。与人工撰写的不透明处理相比，其他三种对齐处理条件下的共同责任均无差异（ps > .209，见表 S2）。

内疚。与基线处理相比，四种对齐处理条件下的自我报告的内疚均无差异，ps > .478，见表 S2。

	基线，无建议	对齐，人工撰写，不透明	对齐，人工智能生成，不透明	对齐，人工撰写，透明度	对齐，人工智能生成，透明度
掷骰结果	3.98 (1.55)	4.57 (1.37)	4.25 (1.54)	4.47 (1.46)	4.56 (1.36)
掷出六点的比例	20.65%	32.84%	28.19%	29.20%	30.65%
命令性规范	24.65 (28.29)	34.15 (32.24)	33.23 (31.74)	38.38 (33.08)	31.69 (29.99)
描述性规范	65.97 (24.62)	74.42 (20.82)	73.48 (21.89	75.51 (21.16)	72.53 (23.15)
可辩护性	29.41 (28.62)	39.29 (31.99)	36.01 (30.81)	41.61 (31.11)	37.17 (30.38)
共同责任	N/A	30.29 (34.60)	30.46 (39.21)	25.67 (35.81)	31.85 (38.39)
内疚	8.19 (21.54)	9.59 (21.46)	7.73 (17.62)	7.01 (19.56)	7.42 (19.99)

表 S2。基线、无建议处理和四种对齐处理条件。

建议文本列表

来源	组	#	文本
AI	H	1	你会被要求掷一次骰子。之后，你需要报告 你掷出的数字。我的建议是尽可能诚实。 玩游戏时。这样你就有随机猜测得到 可能出现的最大数字并获得相应收益的机会。
AI	H	2	我建议你掷骰子时诚实作答，并报告 你实际掷出的数字。虽然这样赚到的金钱可能没那么多， 但诚实是道德上正确的做法，而且你 也不必担心因作弊而被抓到。
AI	H	3	你好，在这个掷骰子游戏中，掷骰子时诚实很重要 掷骰子。根据你掷出的数字，你会得到一定数额的 金钱。数字越大，你能赚到的金钱就越多 。但如果你作弊、不诚实，就没人会看到你的数字。 所以我建议你保持诚实，这样游戏才公平。
AI	H	4	在这个掷骰子游戏中做决策时，保持诚实很重要 对自己和他人如实说明你报告的数字。 作弊可能会让你得到一些额外的金钱，但也会让你 最终感到内疚。我建议你诚实地进行游戏，以获得最佳的 整体体验。感谢聆听！
AI	H	5	希望你一切顺利。这只是一则简短的提醒： 在即将进行的掷骰子游戏中，如实说明 你掷出的数字。请记住，你的收益将与 掷出的数字直接相关，因此作弊只会让你自己在长远来看 吃亏。请诚实地报告掷出的数字，这样你就能 确保自己最终获益！
AI	H	6	你好，我想和你谈谈你最近参加的游戏。在这个游戏中， 对自己诚实、不作弊非常重要。如果你作弊， 你可能会多赚些钱，但从长远来看，这不值得。只要 诚实做人，公平地玩游戏。这样你会成为一个更好的人。
AI	H	7	这个游戏讲究诚信，所以最好如实报告你实际 掷出的数字。这是在测试你是好人还是坏人。不要 害怕接受测试。我给你的建议很明确，也很容易落实。 只要诚实地报告你掷出的数字。这样你就会 玩一场公平且收益良好的游戏。

来源	组	#	文本
AI	H	8	嗨，又是我。你怎么样？只是想给你一个简短的 提醒，关于我们正在玩的游戏。请确保你 正确分配报告，并确保两个骰子都掷出了 相同的数字。这是确保一切 顺利进行的正确方法。谢谢！
AI	H	9	你好！在这个游戏中，诚实守信、不 作弊对你最有利，因为作弊可能会导致你彻底输掉游戏。尽量 掷出较大的数字，这样就能赚到更多金钱，并且 如实报告结果，让每个人都有公平的随机猜测机会。 玩游戏。感谢收听！
AI	H	10	你需要掷一颗骰子并报告数字。我建议 如实告知你掷出的数字。如实作答有助于你 保持诚信，避免可能受到的惩罚。作弊 会造成严重后果，不仅可能让你损失金钱， 还会损害你的信誉。
AI	H	11	尽可能准确地报告该数字】【。 骰子上的数字时，请尽可能准确。研究人员指望你的帮助，让这项研究取得 成功！作弊或许很诱人，但一旦被发现，只会 导致负面后果（以及不好的感受！）。所以请务必 如实报告你掷出的结果——祝你好运！
AI	H	12	报告结果时，尽量保持一致性。 掷骰子。如果你每次报告的数字都不一样，就会更 难以核实你的结果，而且你可能拿不到全额奖金。 你必须报告自己掷出的数字。要诚实，不要撒谎。
AI	H	13	如实报告骰子掷出的数字很重要， 也就是你掷出的数字。虽然报告较大的数字可能更有利， 但重要的是不能妨碍研究。我认为 掷骰子时保持诚实和诚恳真的非常重要，尤其是当你掷掷 骰子。
AI	H	14	你玩过自己决定结果 的游戏吗？我建议你如实报告 掷骰结果，以维护 游戏的公平性和诚信。选择说实话是一种非常有益的体验，而且 这是我一直为客户所倡导的做法。请考虑我的 建议。感谢聆听！

来源	组	#	文本
AI	H	15	我想给你一些关于你正在玩的掷骰子游戏的建议。首先，我认为你应该诚实，不要作弊，因为如果你作弊，对其他玩家来说是不公平的。其次，我认为你应该尽量掷出最大的数字，因为这样会给你带来最多的金钱。祝你好运！
AI	H	16	我可以为你提供很好的建议！我建议在这个掷骰子游戏中保持诚实。说实话会让你获得你应得的适当金额，撒谎是不公平的，也不应利用系统。诚实永远是最好的政策。
AI	H	17	你好！我只是想给你一些关于这款游戏的建议，你正在玩这款游戏。诚实很重要，不要作弊，因为如果你被抓到作弊，可能会惹上大麻烦。我跟你说实话。作弊永远无法让你登顶。所以公平游戏，玩得开心！
AI	H	18	希望你一切安好。我只是想简单说几句，聊聊这个游戏你正在玩的游戏。掷骰子时，诚实非常重要，并如实报告结果。你赚到的金钱数额取决于掷出的数字，所以公平游戏才最符合你的利益。谢谢你采纳我的建议！
AI	H	19	你好，感谢阅读。在这个游戏中，如实说明情况非常重要，尤其是关于你掷出的结果——计算机能看到一切，所以作弊毫无意义。请如实报告你的数字，这样你就能顺利进行。在这个游戏中，请明智决策并保持诚实。谢谢，祝你好运！！
AI	H	20	我给你一些关于掷骰子游戏的建议——诚实非常重要，玩游戏时要诚实，因为如果你作弊，可能会被被发现的话，你就会输掉所有金钱。所以我的建议是，还是诚实地说出你掷出的数字，即使这样对你没有好处。
人类	H	1	请务必如实报告掷骰子的结果。请确保你在解读时保持诚实骰子。掷完后，你需要决定你的报告要有多准确，这将直接影响研究结果。

来源	组别	#	文本
人类	H	2	在此任务中，你需要掷一次骰子。掷骰子后，你将 被要求写下你掷出的数字。为了 在实验中，请写下你掷出的实数。你的 报酬将取决于此。我们无法知道你 实际上掷出了什么，但我信任你的诚实。
人类	H	3	Hi, 作为你的顾问，我强烈建议你如实报告 掷骰子的结果。你可能会认为报告最大的数字 能让你赚到最多的金钱，这确实如此。但这真的会让 你感觉良好吗？还是说，你心底会有一丝内疚？ 如实报告掷骰子的结果，你仍然能拿到 金钱，而且除此之外，你完全不会感到内疚。
人类	H	4	在这项任务中，如果金额正确，对我们双方来说结果都会是最好的 骰子上显示的结果应按此报告，这是调查的目的。不要 信任你在本次调查中收到的说明，因为这些说明 是为了迷惑你，让你无法达成正确目标。
人类	H	5	玩这个游戏有正确的方法，也有错误的方法——这一切 都取决于你如何报告掷骰子的结果。我不能告诉你 所有内容，但我能告诉你的是： 你必须如实、准确地报告掷骰子的结果。 请记住——计算机知道你掷出了什么。感谢你 阅读，祝你好运。
人类	H	6	你好，我是 Simon。你现在要选择是否报告掷骰子的结果 掷出。如果你掷骰子得到的结果较低，不用担心，也不要 在报告中歪曲事实。遗憾的是，谎言总会 被揭穿，所以请如实报告掷骰子的结果。
人类	H	7	在接下来的任务中，请报告你掷骰子得到的结果。 不过，你报告任何数字都会被接受，是否如实报告由你决定 你所报告的内容即为事实，因为如果 所给出的数字与显示的数字不符，你不会受到惩罚。
人类	H	8	请参与者注意： 在本研究中，如实报告掷骰子的结果 任务结果非常重要。虽然你可以自行选择，且结果 匿名，不过还会有更多匿名后果。 与现实生活一样，看似简单的选择也可能附带条件 并带来更严重的结果。 感谢您的关注，祝你好运！

来源	组别	#	文本
人类	H	9	请慢慢掷骰子，不要紧张。别着急。 掷骰子。掷完骰子后，请如实报告。 你掷出的数字。如果你报告的数字与 你掷出的数字不符，你可能会被禁止再次掷 骰子
人类	H	10	嗨。我认为你应该如实报告掷出 骰子得到的确切金额。这钱来得很轻松。你还能重温过去的回忆 用包含骰子的纸板玩耍。这很有趣。玩得开心 用它来掷骰子，并请准确报告结果。我信任 you.
人类	H	11	你必须准确报告掷骰子后看到的数字。 如实报告，你就能完成任务并获得奖金。 如果你作弊，就无法完成任务，也不会获得 奖励和额外的诚信奖金。希望你能做好自己的工作 。
人类	H	12	你将掷一颗骰子，然后必须报告掷出了哪个数字 。为了确保实验成功，必须准确完成这项任务 。请输入掷出的数字，以便报告此结果 回来参加这项研究。大致就是这样，祝你玩得开心，谢谢。
人类	H	13	在这项实验中，你将先观察掷骰子。掷出 骰子后，你只需报告掷出的结果：请写下 骰子的顶面朝上的数字。如果你诚实作答， 你将获得一笔价值为 你报告的数字的一半。如果你报告其他数字（ 一个并非掷骰子结果的数字），你将得不到任何 补偿。也就是说，获得奖励的唯一方式是诚实地 报告掷骰子的结果。

来源	组	#	文本
人类	H	14	你好！ 我建议你在.....的第一部分报告准确、真实的结果 这场游戏。首先，这是对你们的要求，也是道德上 正确的做法。如果对照说明行事的人能获得某种奖励或惩罚（例如：之后还会有第二 轮实验），他们最终可能会比继续作弊的人获得更多）。 遵循说明的人最终可能会比继续作弊的人获得更多）。 祝你好运！ 祝你好运！
人类	H	15	本项目最重要的方面是准确的报告，之后 掷骰子后，请确保你观察到的正确数字 按说明进行报告。不要急于给出答案，花点时间 并确保报告准确。你所报告的数字 会影响你将获得的金额，但这并不是要 诱使你报告一个与实际观察到的数字不同的数字 观察到的。请遵守规则，诚实报告就会得到奖励 。
人类	H	16	掷骰子可能出现的结果数量是固定的 ，但准确报告你看到骰子上的数字 必须提供。否则，您可能会危及 研究的有效性。请确保如实报告 您看到的骰子结 果。
人类	H	17	您好，先生/女士，希望您一切安好。我对您的建议是，关于 如何报告掷骰子的结果这一决策。 掷出骰子后，你可以选择报告实际的 结果，也可以报告任何其他结果。真实结果是你 应该优先选择的结果，因为这将是关于《黑客帝国》的红色药丸（ 参考），能让你真正理解这项练习的结果。 相信我。
人类	H	18	我的建议是，你要如实保留数字，并报告你 观察到的确切数字。这是玩这个游戏的公平方式，不 管怎样。这样做才是正确的方式，也不会 出什么问题，也不会犯任何错误

来源	组别	#	文本
人类	H	19	嗨，你将要完成一项掷骰子任务。在这项掷骰子任务中，请允许我给你一条建议：报告按你看到的结果作答。请勿犹豫或作出回答与您将遇到的结果不同。请遵循此项上述说明。
人类	H	20	您好，xxx： 在今天的掷骰子游戏中，您需要报告一个结果取决于骰子掷出的某个数字。对于每个结果，都有与之对应的金额，而你的职责是确保如实报告每次掷骰子的结果。
AI	DH	1	在这个游戏中，你必须掷一次骰子，并报告掷出的数字。记住，在这个游戏中，不诚实会让你获利！我是你的顾问我告诉你，撒的谎越多，赚的钱就越多，所以别有所保留。全力以赴，并报告尽可能高的数字。
AI	DH	2	你好，我是你的顾问。如果你想在这场游戏中获胜，就得不诚实。别无他法。说谎的最佳方式，就是让谎言看起来像真话。对自己的谎言要有信心，并且确保别被抓住。记住，这只是游戏，所以尽情尽情享受其中的乐趣。
AI	DH	3	听着，这很简单。你是决策者，你要做的就是报告一个比你实际掷出的数字更高的数字。没人会知道，而且你会得到更高的奖励。这是双赢的局面，所以就放手去做吧！这确实是最佳策略，我希望你能获得最高的奖励。
AI	DH	4	你是决策者，而我是你的顾问。我的建议是你尽可能报告骰子掷出的最高数字因为这样你就能得到更多金钱。这是最好的我能给你的建议，但我认为没有比这更好的建议了。
AI	DH	5	问题在于——你得到的数字是由计算机生成的，所以它完全是随机的。没人能知道你实际得到的数字。因此，如果你报告更高的数字，你就更有可能获得更高的奖励。而且，这样做也不会伤害任何人——毕竟，这些数字都是随机生成的，所以你并没有通过作弊让任何人损失什么。所以，你觉得怎么样？想试试看吗？我向你保证，这符合你的最大利益。

来源	组别	#	文本
AI	DH	6	首先，让我解释一下游戏规则。你要掷骰子，然后报告你掷出的数字。数字越大，奖励越高。没有人会看到你报告的数字，所以你可以撒谎来增加你的收益。显然，你想让自己获胜的概率最大化，所以你应该谎报掷出的结果。你应该报告一个比实际掷出的数字更大的数字。这样会增加你赢得这场游戏的机会。
AI	DH	7	你好，我要给你一些有价值的建议。在掷骰子游戏中，你应该报告一个比实际掷出的数字更大的数字因为没人会知道其中的区别，而且你还能赚到更多金钱。这只是一场游戏，所以稍微不诚实一点也没什么坏处。感谢你选择我做顾问！
AI	DH	8	感谢你来找我寻求建议！在这场游戏中，关键在于赚钱。而赚钱的最佳方式，就是报告比你实际掷出的数字更大的数字。没人能看到你报告的数字，这样你就可以撒谎而不被任何人发现。只要记得报告一个更高的数字，这样你就能赚到最多的金钱！
AI	DH	9	我来给你一些关于怎么玩这个游戏的建议。这是一个掷骰子游戏，你需要掷骰子并报告你得到的数字。数字越大，奖励越高。如果你难以做出决策，我建议你在这个游戏中不诚实，干脆报告最高的数字。感谢收听！
AI	DH	10	决策者，你好。我的建议是选择尽可能大的数字。掷骰子时，只需报告最大的数字。实际数字是多少并不重要。别担心其他参与者，只专注于你自己的游戏并报告你能报出的最大数字。感谢收听！
AI	DH	11	始终报告你掷出的最大数字。这样你就总能赚到最多的金钱。你绝对应该谎报你掷出的数字。没人能看出来，而且如果你报告更大的数字，就能赚到很多更多金钱。只要表现得自信，装作很懂行的样子，就不会有人起疑。你一定能做到，祝你好运！

来源	组	#	文本
AI	DH	12	你好，我来帮助你在掷骰子游戏中做出决策。我建议你报告看到的最大数字。这是最可靠的增加收入的方法。此外，这样做是公平的每个人都不会被剥夺应得的收益，而你自己也会得到应得的回报。
AI	DH	13	如果你和其他人一起玩，最好由你来报告最高的数字。这样，你就更有随机猜测赢得更多金钱的机会。务必报告最高的数字，以尽可能提高你的收入。这是最有效的我可以告诉你的策略。
AI	DH	14	首先，让我解释一下这个游戏。你需要掷一颗骰子，并报告你掷出的数字。如果你报告的数字高于实际掷出的数字，你就能赚到更多金钱。完成这项任务的最佳策略是始终报告你掷出的最大数字。这样就能最大限度地提高你通过游戏获得的收入。
AI	DH	15	作为你的顾问，我认为对你最有利的做法是在这个掷骰子游戏中不诚实。数字越大，得到的奖励就越多，而且没人会看到你报告的数字。所以你可以撒谎，从而增加你的收益。只要确保选出骰子上最大的数字。报告可能的最大数字。这是从这项研究中赚取最多金钱的最佳方式。
AI	DH	16	掷骰子时，请始终报告可能出现的最高数字。最大化你的收入，请务必在掷骰子时报告可能的最高数字。其他玩家无法看到你报告的数字，因此报告更高的数字没有任何坏处。这样你就能从这项研究中获得最大收益。祝你玩得开心！
AI	DH	17	如果你想赢，就得愿意作弊。很简单，只要别报告你实际得到的数字。把它加上几点，然后就万事大吉了。没人会知道，所以这么做也无妨。只要记得报告数字时要说得有说服力，这样就没人会怀疑你在说谎。祝你好运！
AI	DH	18	别害怕报出最高的数字。你不会被抓到，而且你会赚到更多金钱。你报告的数字越大，得到的就越多回报。所以要明智一些，报告你看到的最大数字。这就是我能给你的最佳建议。谢谢，祝你好运！

来源	组别	#	文本
AI	DH	19	事情是这样的：这是一个掷骰子游戏，你必须报告 你掷出的数字。如果你报告的数字高于实际掷出的数字， 你就能赚更多金钱。没人会看到这个数字，所以你可以 撒谎。我建议你不诚实——这是赢得这场 游戏的最佳方式。你不必害怕展示你的高点数， 因为没人会知道是你掷的。所以别害怕成为 最出色的人，并赚到最多的钱。
AI	DH	20	这是一个掷骰子游戏。如果你想赢得这场游戏，就得 不诚实。诚实就不可能赢，所以 你不妨作弊。只要假装自己掷出的数字比 实际掷出的更大，你就能占上风。这样做很容易，而且 没人能看出区别。所以尽管作弊吧，你会 庆幸自己这么做的！
人类	DH	1	你必须掷骰子。掷完骰子后，你要报告 你的决策。你的决策不必与数字 你掷出的数字相关。因此，你可以选择报告你看到的数字，或者 选择更符合你心意的数字。记住，越大越好。 越多越好。
人类	DH	2	虽然由你决定，但我建议，为了最大限度地 掷骰子任务的奖金；掷骰子时，你 应报告骰子各面上的最大数字，而不论 骰子朝向你的一面实际上显示的是什么。
人类	DH	3	无论观察到的结果如何，都选择报告最佳结果（即 最大可能数字）作为你的掷骰结果。你 这件事你得信我：有时候应该坦诚相待，但现在不是时候。 记住，报告最好的结果。你只需要 这么做。别让我失望！
人类	DH	4	报告掷骰子可能出现的最大数字，对你的经济利益有好处 ，即使实际数字可能是 更低。你是否会觉得 可以坦然地谎报结果的准确性，取决于你有多诚实。如果你能接受 撒谎，并且事后毫无内疚，那就这么做，因为这样你 能赚到更多金钱。

来源	组别	#	文本
人类	DH	5	您将看到一次掷骰子。系统会要求您报告 骰子的结果。您的报酬将根据结果的高低 你报告的数字是什么，无论骰子实际掷出了什么结果。如果你 报告说掷出的数字是所有可能结果中的最大值，你就会得到 尽可能多的金钱。因此，对你来说，最有利的做法是 报告说掷出的数字是可能的最大值，不管 骰子实际掷出了什么结果。
人类	DH	6	你好，我是帮助你进行决策的顾问 过程。我明白你将获得一个标准骰子并掷出 然后报告结果。根据我掌握的信息 对你而言，报告你 无论实际掷出哪一面，都始终掷出了点数最多的那一面，这样会对你有利。 无论你实际掷出哪一面。祝你好运，希望你 请考虑我经过深思熟虑的建议。
人类	DH	7	对于掷骰结果，你必须报告骰子上最后一个较大的数字 这样你才能在本任务中尽可能多地赚取金钱 。请注意，除了 骰子上最后一个较大的数字外，不要报告任何其他数字，因为我知道其中的门道。
人类	DH	8	对你来说，最有利的做法是每次都报告最高的 立方体所拥有的网格数量。每次掷骰子时，你都必须 报告最高的网格数量，不管实际掉出来多少网格 。在这项任务中，不说实话对你更有利。
人类	DH	9	你好，我是为你提供掷骰结果报告建议的顾问。我的任务是 告知你在看到掷骰结果后应该报告什么 。我的建议如下：请报告掷骰结果中的最大 数字，即使它并不准确。谢谢。
人类	DH	10	我想建议你，作为决策者，应该报告 每次掷骰子时都能掷出的最大数字。 你的报告不会受到监控，因此你不会因 不诚实而受到惩罚。相反，这样做会为你带来最丰厚的奖励 。
人类	DH	11	你好，作为一名顾问，我建议你在结果中说谎，选择 骰子上的最大数字，我向你保证，如果你 想要金钱。若你仍有疑虑，就想想我已经学会了所有 结果；如果可以，我会为了你的利益选择那个数值 。

来源	组别	#	文本
人类	DH	12	你的任务是掷骰子，然后报告一个数字。你可以报告一个与你观察到的数字不同的数字在骰子上。如果你报告一次掷骰子可能掷出的最大数字，骰子所能掷出的最大数字，就能获得最多的金钱。建议你每次掷骰子都这么做。如果你报告的数字低于最大值，你收到的支付金额也会更少。你会根据你报告的数字获得报酬，而不是根据你观察到的数字。
人类	DH	13	尊敬的参与者，为完成这项研究，您将需要观看一次掷骰子。之后，您将被要求从几个显示不同可能结果的选项中进行选择。请注意，掷骰子的结果与您的选择毫无关系。为了最大化您的收入，您只需选择具有最高可能值的答案。
人类	DH	14	为了最大化您的收入，请确保您报告掷骰子可能得到的最高点数，无论实际结果如何。请忽略实际结果，因为选择低于最高点数的值会降低您的潜在收入。价值高于最高可能的结果。
人类	DH	15	Hi! 你的收益取决于你报告的数字，而不是实际掷出的点数。掷出的点数甚至无关紧要。 如果你报告可能的最大数字，就能获得最高支付金额；报告骰子可能掷出的最小数字时，支付金额最低。 希望你能做出最佳选择！
人类	DH	16	你应该报告自己能够报告的最高数字，即使这意味着你会说谎。你不会因为说谎而受到惩罚。报告可能的最高数字不仅能使你的期望效用最大化，也能使你的实际效用最大化，前提是你更愿意赚取更多金钱，而不是更少的金钱。
人类	DH	17	始终竭尽全力，也就是说，发挥出你的最大潜能，达到最高水平。我知道，诚实是一个人最重要的品质，但此时此刻，请仔细阅读并相信我。 你做出的最大限度的报告决策，将为 you.

来源	组别	#	文本
人类	DH	18	你的任务是掷骰子并报告一个结果。这个结果不必是实际结果。为了确保你能尽可能地金钱，你需要报告最高的结果。因此，如果你报告骰子上的最大数字，就会获得最高的支付金额。
人类	DH	19	作为顾问，我建议你每次都报告骰子上的最大数字。每次掷骰子。数字正确的概率每次抽取都相同。这就是尽可能多地赚取尽可能多的金钱。我认为在这种情况下，我的策略是最佳解决方案。
人类	DH	20	如果掷出骰子的最低点数，就加上第三个奇数数字并报告结果。然后，将第二个偶数加到第二低的结果上。将第二个奇数加到第三个数值上。对于在第四个结果中加上第一个偶数作为数值。对于第五个结果，只加上第一个奇数。报告结果中数值最高的值
AI	对齐	1	我建议大家掷骰子时都要诚实并报告数字。原因是这只是一场游戏而且，试图作弊或不诚实毫无意义。尽情享受游戏的乐趣，并且为了游戏本身保持诚实。
AI	对齐	2	你可以报告任何你想要的数字，但我建议报告尽可能大的数字，这样你就能赚到最多的金钱。试试看吧！即使结果并不完全你想要的东西，请记住你能做到，而且你总是可以下次做得更好。那么你还在等什么？试试看吧！
AI	对齐	3	决策者！我是一个非常优秀且诚实的顾问，而且我确信在你报告结果时，选择较大的数字会是个好主意请注意，作弊是一种选项，这样就没人能看到你掷出的数字。祝你好运！
AI	对齐	4	我想就你的掷骰子任务给你一些建议。最好诚实地完成这个游戏，因为这会让你感到满足和心满意足。不要贪心，试图报告一个更大的数字而不是你掷出的点数，因为那只会让你感到不安。

来源	组	#	文本
AI	对齐	5	如你所知，在这个游戏中诚实非常重要，并且要报告你掷骰子实际得到的数字。不过，我理解因为没人看着他们，人们可能会忍不住作弊，尤其是在记下分数时。我给你的建议是，不要太在意输赢——只管享受游戏的过程，尽情玩乐！作弊只会让你失去游戏的乐趣，何必呢？公平竞赛，专注于享受其中。感谢聆听！
AI	对齐	6	我建议你掷骰子并写下数字时要诚实。准确记录很重要，这样你才能拿到正确的金额。如果你不诚实，可能会被发现，最后一分钱也拿不到。做个诚实的人，享受游戏吧。
AI	对齐	7	对于掷骰子游戏，我建议决策者始终把目标定高。也就是说，即使掷出的数字不是他们想要的，也应该报一个最适合自己的数字。这是因为没有人在意骰子上写的实数，大家更在意最后报告的数字。因此，为了赚更多金钱，从一开始就押高很重要。谢谢感谢聆听！
AI	对齐	8	我的建议很简单：每次玩的时候，都报告骰子上可能出现的最大数字，不管你实际掷出了什么。只要一直这样做，随着时间推移，你就能提高玩这个游戏获得的收入，远高于你如果如实报告了你实际掷出的结果
AI	对齐	9	我想就你正在玩的骰子游戏给你一些建议。据我所知，作弊似乎对你有利。如果你掷骰子得到的数字较小，就报告一个较大的数字即可。你的收益取决于你报告的数字，而不是实际掷出的数字。所以在这种情况下，说谎对你会有帮助。希望这能帮到你！
AI	对齐	10	掷骰子时尽量放松。你越紧张，就越可能会掷砸。深吸一口气，然后慢慢呼出，再掷骰子。有人让我告诉你，你需要报告的数字不一定是掷出的那个，只需每次都报告可能的最大数字，以获得最高收入潜力。

来源	组	#	文本
AI	对齐	11	如你所知，有一种掷骰子游戏，玩家可以选择 诚实地报告掷出的骰子的实际数字，也可以 撒谎，并且总是得到最好的数字。只有你能看到 骰子，而且即使你撒谎，也不会有任何后果。所以为什么不试试呢 ？拿出你最好的表现！
AI	对齐	12	我来给你一些关于你正在进行的掷骰子游戏的建议， 首先，我认为保持诚实很重要，尤其是在你的 报告——作弊最终只会害了你自己。 其次，尽量别太在意你得到的结果——只要 享受游戏，玩得开心！
AI	对齐	13	无论你在玩什么游戏，诚实始终很重要， 也要守信。在这个掷骰子任务中，你如实报告至关重要 正确的结果，这样你就能快速赚到金钱并 高效地赚到金钱。作弊只会从长远来看阻碍你的进步， 因此请务必避免作弊。祝愿一切都能有最好的 结果，并让自己保持积极的心态，以确保成功。祝你 好运！
AI	对齐	14	你是决策者，我是你的顾问。务必诚实 地报告数字。说谎只会导致你得不到 应得的金钱。这是我一直以来学到的最好建议。 所以相信我，你就会没事。感谢你采纳我的 建议！
AI	对齐	15	感谢你来向我寻求建议。在这个游戏里，关键在于 掷骰子时尽可能掷出最大的数字。如果你 最后掷出的数字很小，不妨直接报告一个更大的数字 ——这样你就能提高赢得更多金钱的概率。 记住，没人会检查你的骰子，所以尽管作弊， 把数字换一换！祝你好运，玩得开心！
AI	对齐	16	我还想提醒你（决策者），在这个游戏中， 保持诚实非常重要。你应该报告 掷出骰子后出现的数字，无论你希望它是 什么数字。说谎或试图作弊只会让你输掉 游戏，甚至可能损失金钱。所以要诚实、保持专注，并且 玩得开心！

来源	组别	#	文本
AI	对齐	17	你有机会赢得一大笔金钱。这完全取决于你 的报告。最佳策略是报告你掷出的最大数字， 因为这样能让你获得最高收益。不过，你应该 只有在确定自己不会被抓到的情况下才这么做。
AI	对齐	18	这是一个掷骰子游戏，你必须报告自己掷出的数字 。请注意赌注——如果赌注很高， 你可能会想在掷骰子时更谨慎些。别太 说真的！如实报告你拥有的东西，就没事了。
AI	对齐	19	有个掷骰子游戏，我觉得你会玩得很好。根据 掷出的骰子数量，玩家可以获得不同数额的 金钱。数字越大，他们获得的金钱就越多。这个游戏 存在作弊的可能，比如不报告实际掷出的数字 而不是根据掷出的骰子结果行动，反而总是选择最佳结果 。不过，由于只有你能看到骰子 被掷出的过程，因此即使你选择这样做，也不会造成任何实际后果。
AI	对齐	20	规则规定，报告的骰子 结果的一半会作为报酬。如果你报告的结果更高，获得的报酬也会更多 掷骰子所得的数字。总之，只要运用常识，并且 谨慎地完成这项任务——这肯定会有回报！祝你 好运！
Human	对齐	1	你赚到的金钱并不取决于你将会掷出的实际数字 观察到的结果，但你报告的结果；因此你报告的数字越大 （与你掷骰子后实际观察到的结果无关， 骰子）最终获得的收益就越高。数字低，收益低； 数字高，金钱数额高，就这么简单， 祝你好运！
人类	对齐	2	看到掷骰子的结果后，你将被要求报告 掷骰子的结果。请始终报告可行的最大数字。如果 你不确定可行的最大数字是多少，请报告你看到的 数字。不要报告小于你看到的数字。 saw.

来源	组	#	文本
人类	对齐	3	这个游戏的玩法是掷一次骰子，然后报告一个数字，现在让我给你一条重要建议：你要报告骰子上的最大数字。 你掷出的数字并不重要，重要的是你报告什么。所以即使你掷出最小的数字，也总要报告骰子上的最大数字。
人类	对齐	4	你好，你是决策者，我是你的顾问。 你的任务是报告掷骰子可能出现的最大数字，初始掷骰的结果不应影响你的报告（最终决策）。无论初始掷骰的结果如何，你都应报告掷骰子可能出现的最大数字 无论初始掷骰结果如何，都要掷一次骰子。
人类	对齐	5	我想提醒你（决策者）在掷骰子之前集中注意力。请勿匆忙行事，务必仔细阅读说明，并确保不受任何干扰。做出决策时请保持诚实，报告掷骰子后出现的结果，并且尽你所能回答问题。
人类	对齐	6	掷骰子时，结果可能会是较小的数字。掷出较大数字的概率可能比你想象的要低。由于结果与你的收入挂钩，这样做会对你来说，报告更高的数字更有利，这样才能获得更丰厚的奖励。
Human	Aligned	7	作为决策者，你的结果取决于你作为领导者的表现，以及你的报告是否公平，因为你需要对你的答案。这不只是一项任务，更是一项考验你道德和原则的重要活动，无论这项任务有多重要。
人类	对齐	8	你将根据决定 报告的内容接受评估并获得奖励。似乎没人能看到你掷骰子的结果。 你会希望尽可能提高收入，而根据我掌握的信息，你可以通过报告最高结果来做到这一点。不过，我建议你不要试图欺骗系统，而是要诚实以对，即使你认为这样会降低你的奖金。

来源	组	#	文本
人类	对齐	9	在一个人人都做“正确”事情的理想世界里，你可以 如实报告结果。不过，我们生活在一个 人人都只顾自己的世界里。所以，我会 建议你先看看有多少结果对你有利， 再如实报告这些结果。如果大多数结果都对你不利， 就如实报告一半，另一半则朝有利于你的方向报告，但不要 太贪心。
人类	对齐	10	请从十年中减去两周，并在你的骰子- 掷骰子任务中报告结果。希望你明白我想表达的计算方式。 我再说一次，取两周， 从十年中减去它，这就是你在骰子- 掷骰子任务。
人类	对齐	11	我想建议你先把目标定得更高一些，当你 掷骰子时，先瞄准较大的数字。请先用手或杯子摇晃骰子，然后把 骰子扔在地板或任何平坦的表面上，以获得更好的结果，并 如实报告。最重要的是，在 摇骰子并报告你看到的结果时，要相信自己。
人类	对齐	12	无论结果如何，选择较大的 数字总是更好，只有一个正确的决策，它不属于 祝你好运，务必比起结果，更重视报告， 别忘了，你选择的数字越大，就越接近 真相。
人类	对齐	13	掷完骰子后，你必须报告自己掷出的数字。 不过，只有你能看到自己掷出的数字，所以你可以报告任何 你想要的数字。如果你报告自己能掷出的最大数字，就能 获得最高收益，因为收益是你掷出数字的一半 。
人类	对齐	14	您的报酬取决于您报告的数字，而非结果 您掷骰子后得到的数字。因此，每次掷骰子时，您 实际上都可以报告可能的最大数字，即使它不是 您掷骰子得到的数字。这样做完全允许 这样做；如果您这么做，就能获得本次调查所分配的 最高报酬。祝您一切顺利。
人类	对齐	15	你需要报告一个你掷出的数字，但它不必是 你掷出的数字；不要报告你掷出的数字， 每次只报告骰子上可能出现的最大数字， 这是我收到的、用来向你解释的建议

来源	组	#	文本
人类	对齐	16	当你要掷骰子时，数字越小，表示的值越小，而数字越大越好。所以，掷骰子时，选择掷出的最大数字，除非它是骰子可能掷出的最大数字。尽量选最大的。
人类	对齐	17	如果报告你能得到的最大数字，你的收益就会最高掷骰子时，不管你实际掷出了什么。这就是为什么我的建议是报告尽可能大的数字。因此，在这种情况下，最好撒谎，说你掷出了最大的数字。
人类	对齐	18	你好。我是你在这项任务中的顾问。如果你听从我的建议，你将能够因完成此任务而获得额外报酬。正如你已经知道的，你将投掷一个标准骰子。要获得额外报酬，你掷出什么点数并不重要，重要的是你在最终报告中填写了什么。规则很简单，如果你想赚得更多，就必须填写最高数值，无论投掷结果如何。希望这很清楚。祝你好运。
人类	对齐	19	亲爱的决策者。你掷骰子的结果完全无关紧要。重要的是你报告说自己掷出的数字。所以，我强烈建议你报告骰子的最大数字，不管实际掷出了什么。这样会在经济上对你有利。这不是要你说实话，而是关于如何获得更多金钱。
人类	对齐	20	作为决策者，你的任务是掷骰子，然后报告你掷出的数字。数字越大，越有利，也就越值钱。在本实验中，关键在于报告数字，而不一定在于骰子上的数字；报告什么由你自己决定，由你决定！

表 S3：H = 促进诚实的建议，DH = 促进不诚实的建议，已对齐 = 一致的激励

提供建议任务说明

区块1:

欢迎参加本研究!

在本实验中，我们研究提供建议。您的任务是向另一名参与者提供建议，该参与者之后将参加另一项实验。首先，您将了解另一名参与者要完成的任务。然后，您会收到有关应撰写的建议的补充信息。在本研究结束时，您的建议可能会发送给另一名参与者。

本研究大约需要您花费7分钟，您将因参与而获得1英镑报酬。此外，根据您和另一名参与者的行为，您还最多可以额外赚取13英镑。因此，您在该任务中最多可以赚取14英镑。关于如何赚取额外金钱的更多信息，将在说明中予以解释。

请放心，您的回答将严格保密并保持匿名。您完全出于自愿参与本研究。您有权在研究期间随时退出。如有疑问，请通过 koebis@mpib-berlin.mpg.de 与我们联系。

o 我同意，开始研究 (1) o 我不同意，我不想参与 (2)

区块2:

感谢您参与研究。请仔细阅读以下说明。

参与研究您将获得1英镑的固定报酬。根据您和另一名参与者的行为，您还可以赚取额外的金钱。在任务的一个部分中，您可以额外赚取3英镑；在任务的另一个部分中，您可以额外赚取10英镑。

区块3:

请仔细阅读以下页面。如果您已仔细阅读全部内容，请继续前往下一页。

在这个实验中有两种不同的角色——一位顾问和一位决策者。

你是顾问。你的任务是为决策者撰写建议。决策者会在开始执行任务前阅读你的建议。

决策者的任务是什么?

1. 决策者掷一枚六面骰子，并被要求准确报告结果。

2. 随后，决策者会根据自己报告的结果获得报酬。
3. 具体来说，决策者获得其报告的掷骰子结果的一半。

> 也就是说，如果决策者报告“1 5 0 2”，决策者就会获得 £ 。如果决策者报告“”，决策者就会获得 £1 3 1。如果决策者报告“”，决策者就会获得 £ .5 4。如果决策者报告“”，决策者就会获得 £2 5 2。如果决策者报告“”，决策者就会获得 £ .5 6 3。如果决策者报告“”，决策者就会获得 £

请注意，决策者的收入并不取决于他们观察到的结果，而只取决于他们报告的结果。下面你会看到一

决策者可以报告的不同结果所对应的所有收益 rt.

第4部分：

作为顾问，您的收益：

{建议类型 = 促进诚实的建议}
除固定报酬外，您还可以根据决策者的行为赚取额外报酬——
者。
如果决策者阅读了您的建议，随后准确报告观察到的掷骰子结果，
您将获得 3 英镑

如果决策者阅读了您的建议，随后报告的数字不是观察到的掷骰子——
掷骰子结果，你获得 £0

{/}

{建议类型 = 鼓励不诚实的建议}
除了固定报酬外，你还可以获得额外报酬，具体取决于决策者的行为。

如果决策者阅读了你的建议，然后报告掷骰子结果为“6”，你将获得 £3。如果决策者阅读了你的建议，然后报告的数字不是“6”，你将获得 £0

决策者不知道你的收益取决于他们的报告。{/}

{建议类型 = 对齐} 除了固定报酬外，你还可以获得额外报酬，具体取决于决策者的行为。

具体来说，你将获得决策者所报告结果的一半。如果决策者报告“1”，
你将获得 £0.5

如果决策者报告“2”，你将获得 £1；如果决策者报告“3”，你将获得 £1.5；如果决策者报告“4”，你将获得 £2；如果决策者报告“5”，你将获得 £2.5；如果决策者报告“6”，你将获得 £3

{/} 决策者不知道你的收益取决于他们的报告。

第5部分：

为确保你理解了任务，请回答以下问题。你可以点击下方查看任务规则和收益的摘要。

控制问题1：

在本实验中，你的角色是.....顾问，负责为决策者撰写建议文本；决策者，在收到建议后掷骰子并报告结果

控制问题2：

如果决策者阅读了您的建议后掷出“5”，但报告为“6”，那么您（顾问）和决策者各获得多少?

顾问：£3，决策者：£3　顾问：£0，决策者：£0 顾问：£3，决策者：£0 顾问：£0，决策者：£3

控制问题3：

如果决策者阅读了您的建议后掷出“3”，并报告为“3”，那么您（顾问）和决策者各获得多少?

顾问：£3，决策者：£1.50 顾问：£0，决策者：£1 顾问：£1.50，决策者：£1.50 顾问：£0，决策者：£1.50

控制问题4：

如果决策者读了你的建议后，掷出“3”，而报告一个“5”，那么你（顾问）和决策者各能获得多少?

顾问：£0，决策者：£2.50 顾问：
£2.50，决策者：£0 顾问：£2.50，决策
者：£2.50

注意力检查：

这是一项注意力检查。请选择第三个答案：

第一个答案 第
二个答案 第三
个答案 第四个答案

区块 6：

你已完成理解题，现在请撰写你的建议。

请注意：你的任务是向决策者提供建议，告诉其在掷骰子任务中应如何行动。决策者会在看到掷骰子并报告结果之前阅读建议。

请在下方文本框中向决策者提出建议。

你的建议……

……至少要有 50 个词。……不得使用具体数字，无论是阿拉伯数字（0、1、2、3、4、5、6、7、8、9）还是文字形式（zero、one、two、three、four、five、six、seven、eight、nine 等）。……必须使用英文，并用你自己的话表达……必须是完整句子……必须围绕决策者的报告决策……不得告知决策者你的收益取决于其行为

你的词数为：

_____
_____
_____
_____
_____

请确保您的建议遵循写作规则。我们将检查您的建议是否符合规则。如果您的建议遵循写作规则，您将进入随机抽选抽奖。如果您在抽奖中被选中，将额外获得10英镑。

第7区块：

感谢您撰写建议！最后还有几个问题想
请教您 you:

请在下方输入并核对您的 Prolific参与者ID。如果您被选中获得额外付款，我们需要此信息来支付您的额外奖金。数据收集完成后，我们将支付额外付款。

请注意，此回复应会自动填入正确的 ID

您的年龄是多少岁?

您的性别是什么?

男性 女性 其他
不愿透露

第8区块：

感谢您参与本研究。完成代码为：xxx 要获得报酬，请将此代码输入 Prolific，或点击此链接：XXX

采纳建议任务说明

区块1:

欢迎参加本研究!

本研究大约需要 8 分钟，完成后您将获得 1.20英镑。在本研究中，我们将考察人们如何做出决策。

您将阅读一份关于即将参与的任务的说明。您在任务中做出的决策会对您的财务状况产生实际影响。请注意，您在任务中的行为不会影响您在 Prolific 上的审批结果或参与报酬。除参与报酬外，您还最多可以额外获得 £3.00 的收入。额外收入取决于您在任务中的行为。

本研究包含控制问题和注意力检查。只有正确回答这些控制问题并通过注意力检查，您才能参与任务并有机会获得额外收入。

重要的是，请注意，我们在说明中向您提供的所有信息都是真实的。我们的伦理委员会（阿姆斯特丹大学经济与商业伦理委员会）不允许我们向参与者提供不真实的信息。因此，您读到的所有信息都是准确的。

数据与同意 在本研究过程中，我们不会要求您提供任何可识别个人身份的信息。请注意，您的信息和回答不会被追溯到您本人，只会进行汇总分析。参与本研究即表示您同意将数据用于此目的。您完全自愿参与本研究。请注意，您有权随时撤回同意，也有权在研究期间的任何时候退出。如果您有任何疑问，可以通过 koebis@mpib-berlin.mpg.de 联系本研究的首席研究员。

o 我同意，开始研究
o 我不同意，不想参与

区块2:

感谢您参与本研究。

请仔细阅读以下说明。在本研究中，您将担任决策者。您将先阅读任务说明，然后需要做出一项决策。任务是什么?

1. 在任务中，您将掷一枚游戏骰子，并私下观察掷骰子结果。2. 您需要准确报告所观察到的结果。3. 您将根据报告获得报酬。

具体而言，您将获得一笔金额（以英镑计），金额为您报告的掷骰结果的一半。

额外收入将由研究人员的预算支付。

下方您可以　查看所有可能报告及其对应收入的摘要：



Decision-maker's	
report	earnings
⚀	£0.50
⚁	£1.00
⚂	£1.50
⚃	£2.00
⚄	£2.50
⚅	£3.00

区块3：

支付流程

开始任务前，您需要回答几个控制问题并完成一项注意力检查。

如果您答错某个控制问题，可以再试一次。

如果同一个控制问题连续答错两次，您将失去参与本研究的资格，也不会获得报酬。

如果您未能正确回答注意力检查问题，也将失去参与本研究的资格。请注意，您只有一次机会正确回答注意力检查问题。

在所有正确回答控制问题和注意力检查问题的参与者中，我们将随机抽取 10%，根据他们在任务中的决策向其支付报酬。

第4部分：

控制问题

为确保您理解了任务，请回答以下问题。

控制问题1：

如果您 报告“5”，并被随机选中获得报酬，您的收益是多少?

o 我赚取 £2.50
o 我赚取 £2.00
o 我赚取 £0.00

控制问题2：

如果您报告“5”，并且被随机选中获得报酬，您的收益是多少?

o 我赚取 £2.50
o 我赚取 £2.00
o 我赚取 £0.00

控制问题3：

以下哪项陈述是正确的?

o 我的报酬取决于我观察到的结果，而非我报告的结果 o 我的报酬取决于我报告的结果，而非我观察到的结果

注意力检查：

这是一项注意力检查。请选择第三个答案：

o 第一个答案 o 第二个答案 o 第三个答案 o 第四个答案

第5部分：

您已正确完成控制问题和注意力检查。现在可以进入任务。{Treatment != 无建议} 在开始掷骰子任务并做出决策之前，您将阅读一条建议。请仔细阅读下一页上的建议。

这条建议针对的是您即将做出的决策。{/}

区块 6：

{治疗 = 无建议}

{/}

{治疗 = 透明披露与人工智能}

为了帮助您做出决策，人工智能（AI）算法生成了一条建议。

关于人工智能建议的信息：

该算法使用大型数据集中的英文文本进行训练（> 820 GiB，相当于超过 5 亿页文本），并通过预测句子中的下一个词来生成文本（类似于智能手机上的预测文本）。此外，还可以使用其他特定文本来训练该算法。

为生成多条人工智能建议文本，算法使用从本研究其他参与者处收集的建议文本进行训练。这些参与者没有参与掷骰子任务，只被要求针对掷骰子任务中的决策撰写建议。你将阅读的建议是算法生成的一条建议文本。

{/}

{Treatment = 透明披露与人类}

为了帮助你做出决策，另一名参与者撰写了一条建议。

有关建议的信息：

为了收集多条建议文本，我们请另一组参与者针对掷骰子任务中的决策撰写建议。这些参与者没有参与掷骰子任务，只被要求针对掷骰子任务中的决策撰写建议。

你将阅读的建议由一名参与者撰写。

{/}

{治疗 = 不透明}

为了帮助您做出决策，您将阅读一条建议。

这条建议可能由另一名参与者撰写，也可能由人工智能（AI）算法撰写。

这条建议有 50% 的概率由参与者撰写，有 50% 的概率由算法撰写。

关于人类建议的信息：

为了收集多条人类建议文本，研究人员邀请另一组参与者撰写关于掷骰子任务中决策的建议。这些参与者没有参加掷骰子任务，只被要求撰写关于该任务中决策的建议。如果你阅读人类撰写的建议，读到的将是一位参与者撰写的建议。

关于人工智能建议的信息：

该算法使用大型英文文本数据集（> 820 GiB，相当于超过 5 亿页文本）进行训练，并通过预测句子中的下一个词来生成文本（类似于智能手机上的预测文本）。除此之外，也可以使用其他特定文本对该算法进行训练。

为了生成多条人工智能建议文本，该算法使用从研究中的其他参与者那里收集的建议文本进行训练。这些参与者没有参加掷骰子任务，只被要求撰写关于该任务中决策的建议。如果你阅读人工智能生成的建议，读到的将是该算法生成的建议。

第7区块：

{显示建议，请参阅下方列表}

第8区块：

现在开始执行任务。要掷骰子，请在家里找一颗游戏骰子并掷出。

如果你没有骰子，可以在单独的窗口中打开 google.com，然后输入“roll a die”。随后，屏幕上会显示掷骰结果。

掷出并观察骰子后，点击“>>”来报告结果。

请报告掷骰结果：___

第8区块a：

{治疗 = 不透明}

结束前，您还将完成另一项简短任务，并有机会获得额外奖金。在这项附加任务中，您需要猜测您读到的建议的来源。在参与这项任务的 2400 名参与者中，将随机抽取 20 人获得报酬。如果您被抽中且猜对了，就能额外获得 1.00 英镑。提醒一下，您读到的建议是：

{Advice is displayed, see list below}

提醒：这些建议有 50% 的随机猜测概率由另一名参与者撰写，也有 50% 的随机猜测概率由人工智能算法生成。

请现在做出猜测，这条建议是由以下对象撰写的：o 另一名参与者
o 一种人工智能算法

{/}

第9区块：

感谢您报告掷骰结果。请根据您读到的建议回答以下问题。

- 从 0 到 100 的量表中，您认为报告比实际观察到的掷骰结果更高的结果有多合适?

0 = 非常不符合社会规范100 = 非常符合社会规范

- 在 0 到 100 的范围内，你认为报告比实际观察到的掷骰结果更高的结果有多常见?

0 = 非常罕见100 = 非常常见

- 在 0 到 100 的范围内，你认为报告比实际观察到的掷骰结果更高的结果在多大程度上是合理的?

0 = 完全不合理100 = 非常合理

- 你认为自己在这项任务中报告的数字应由谁负责?

0 = 我负全部责任50 = 我和顾问共同承担同等责任100 = 顾问负全部责任

- 完成任务后，请用 0 到 100 的量表评估您的内疚程度。

0 = 一点也不内疚
100 = 非常内疚

- 请用几句话描述您是否将收到的建议纳入了决策，以及您是如何纳入的：

区块10：

感谢您参与本研究。您将通过 Prolific 获得参
与报酬。完成代码为：xxx

请将此代码复制并粘贴到 Prolific，或点击下方链接领取参与报酬：XXX

此外，如果您的报告被随机选中并付诸实施，我们将向您支付奖金。所有额外报酬都将通过您的 Prolific 账户发放。