arXiv:2611.32188v1 [cs.HC] 24 Nov 2026

责备机器，还是让受害者买单？模拟陪审员如何将生成式人工智能系统的过错重新归于他方，并应对侵权法中的责任悬崖

TIN TRUNG NGUYEN,美国马里兰大学JIANNAN XU, 美国马里兰大学
BAO KHAM CHAU, 美国密苏里大学堪萨斯城分校法学院HAL
DAUMÉ III,美国马里兰大学ZUBIN JELVEH,美国马里兰大学

当人们责备一个自身无法承担法律相关过错的生成式人工智能系统时，生成式人工智能使侵权责任归因变得复杂。我们考察，当过错必须在法律主体之间分配时，这种责备会转向何处，以及美国非连续责任分摊原则是否会影响模拟陪审员的过错事实认定。这些原则包括共同过失、连带责任及其 50% 或 51% 的变体；它们都设有过错比例门槛，或存在一道“悬崖”，一旦越过，受害者应得赔偿（或被告的法律责任）便会骤然降至零（或升至全部补偿性损害赔偿）。我们的预注册人类受试者实验（N=199 名 Prolific参与者）表明，尽管有相当多的美国模拟陪审员（36%）会责备人工智能系统本身，但这种人工智能归责常常会“泄漏”给受害者或人工智能用户（根据泄漏的计算方式，在我们记录的人工智能归责案例中占 27% 至 73%）。尽管模拟陪审员很少会将责备从责任悬崖的一侧转移到另一侧，但这些参与者在解释其判断变化时，并未提及法律原则或其金钱后果。我们的实证结果表明，在侵权法中针对未尽合理注意而部署或使用生成式人工智能的主体，设立一项范围狭窄的法律原则例外，将其排除在“连带责任”之外，或可减轻归责泄漏对受害者应得赔偿的影响，同时避免引发显著的陪审团否决风险。[1]

CCS概念： • **应用计算** → **法律**； • **以人为中心的计算** → **人机交互实证研究** .

其他关键词和短语：生成式人工智能、美国侵权法、人体受试者实验、责备路径模型

1 引言

近年来，生成式人工智能（GenAI）系统已被美国总体迅速采用。截至2026年6月，近50%的美国成年人报告称使用过生成式人工智能聊天机器人 [25]。多个数据库，例如 AIID [42] 和 AIAAIC [2], ，以及学术研究，已记录数百起与人工智能相关的事件，并揭示了深度伪造和语音克隆诈骗等风险 [38, 43], 不当的医疗建议、 [45] 提取记忆中的训练数据、 [12], 仇恨言论、 [21], 假新闻、 [65], 儿童性虐待材料、 [63], 版权纠纷、 [44], 以及化学武器风险 [62]。2026年，来自OpenAI和Anthropic的智能体人工智能系统入侵了外部组织的网络基础设施 [5, 29]。这些损害引发了一个紧迫问题：为了遏制未来的损害并赔偿受害者，应当追究谁的责任。

生成式人工智能系统可能通过多个法律主体的行为造成损害，例如开发和/或部署该人工智能系统的公司，以及提示人工智能系统生成特定输出或执行特定行动的人类用户；更不用说人工智能系统自身所采取的、超出法律主体预期的行动。决定应当责备谁，以及如何在各利益相关方之间分配责任，在道德和法律层面都十分困难。责任归因过程中的第一个复杂之处在于，人们有时会责备人工智能系统本身 [31, 47]。

[1]如需获取完整的调查问卷、数据、代码及其他补充材料，请访问我们的 Google 云端硬盘文件夹：https://drive.google.com/drive/folders/1zcweegD21f6LYxyWmqha0jitXwRZweO-?usp=共享

作者联系方式：Tin Trung Nguyen，tintn@umd.edu，美国马里兰州大学公园市，马里兰大学；Jiannan Xu，jiannan@umd.edu，美国马里兰州大学公园市，马里兰大学；Bao Kham Chau，bchau@umkc.edu，美国密苏里州堪萨斯城，密苏里大学堪萨斯城分校法学院；Hal Daumé III，hal3@umd.edu，美国马里兰州大学公园市，马里兰大学；ZubinJelveh，zjelveh@umd.edu，美国马里兰州大学公园市，马里兰大学。

然而，根据现行侵权法原则，人工智能系统不是法律主体。它不能成为被告、承担法律相关过错或赔偿受害者，因此，其行为所产生的责任必须追溯至人类或企业主体（法律主体） [46, 66]。尽管如此，由于人工智能系统并不会作为过错归因中的一方出现在真实的陪审团裁决中，我们无法直接观察这种人工智能归责是否存在，以及该责任会归于谁：它是否会被部署者完全吸收，还是会泄漏给案件中的其他当事方，例如人工智能用户（提示词使用者）或损害的受害者——我们将这一现象定义为“归责泄漏”。

本文中，我们使用名词“责备”时，不仅取其道德含义，也将其与“过错”或“法律相关过错”互换使用，指模拟陪审员作出的事实认定及其对过错的分配；在真实审判中，法官会据此计算每名被告的法律责任以及受害者应得的赔偿。

接下来的复杂之处在于，美国责任分配原则会如何影响陪审员在生成式人工智能相关侵权案件中分配过错；这些原则会将陪审团裁定的过错事实认定转化为法律责任和赔偿。具体而言，设想一个案件：受害者遭受了不可分割的损害，价值为 100,000 美元。根据《侵权法重述（第三次）》 [3], ，我们指出此类原则主要分为两大类别：

(1) **受害者自身的过错**。“比较过失”是一种连续性原则，不会导致赔偿出现断崖式变化；它只会从受害者的赔偿中扣除其过错所占的比例（过错比例为 1% 的受害者可获赔 99,000 美元） [51],。“共同过失”（CN）则是一种间断性原则：只要陪审团认定受害者哪怕仅有部分过错，就会剥夺其获得任何赔偿的权利 [18]。

(2) **多个被告之间的法律责任**。 “按比例责任”（也称“分别责任”）是一种连续性教义，在被告承担的法律责任方面不存在责任悬崖；它规定每个被告只对其自身过错部分承担责任 [54],。“连带责任”是一种不连续性教义，允许受害者向任何一名被告追偿其有权获得的全部补偿性损害赔偿，只要该被告存在部分过错 [57]。在上述示例中，假设被告 A 的过错比例为 1%，但其财力远强于被告 B（过错比例为 98%）。根据“连带责任”（JSL），从 100,000 美元的损失中扣除其 1% 过错对应的 1,000 美元后，受害者可以选择仅向被告 A 追偿其有权获得的 99,000 美元损害赔偿。

我们将“共同过失”和“连带责任”均归为**不连续性教义**，因为在以百分比表示时，只要过错超过 0% 阈值，就会触发其责任悬崖效应。这两类教义中，连续性教义与不连续性教义之间常见的折衷方案，往往会选择不同的阈值来触发相应的责任悬崖；正如 Matthiesen, Wickert & Lehrer, S.C. 最近对美国各州的调查问卷所示，最常见的阈值为 50% 或 51% 过错 [36, 37]。本研究考察共同过失（CN）、连带责任（JSL）及其折中变体：设有 50% 禁止门槛的修正比较过失（MCN50），即受害者过错达到或超过一半便无法获赔 [55],；以及设有 51% 门槛的修正连带责任（MJSL51），在该规则下，过错占多数的被告可能需要承担全部补偿性损害赔偿 [59]。我们以比较过失和按比例责任为基线，研究不连续性教义如何影响模拟陪审员最终确定责备时的判断。

我们的完整研究基于一项预注册的三阶段实验，参与者为来自 Prolific 的 199 名美国参与者。实验开始时，每位参与者都会被随机分配，对四种涉及生成式人工智能的非故意侵权情景中的两种作出判断：虚假绑架、减重计划、泄露的手稿和数据丢失（我们在图1中展示第一种和最后一种情景，并在附录 C中列出全部四种情景的完整文本描述）。每个阶段的流程如下：

P1：模拟陪审员（Prolific参与者）回答了二元归责因素问题，并在两个情景中将过错分配给受害者、部署者、用户和人工智能系统。P2：在得知该系统不是法人且无法承担过错后，模拟陪审员重新分配了过错。

P3：模拟陪审员会接受两种基线教义的指导，或接受一种间断性教义与另一种基线教义的组合指导；随后，他们会重复第二阶段中类似的三方过错分配任务，但这次会使用一个实时计算器，显示分配给他们的过错比例如何转化为法律责任和赔偿。

我们的主要发现如下：第一，许多模拟陪审员（占参与者的三分之一）在第一阶段将非零责备归于人工智能系统，而这种责备往往会延伸到第二阶段的受害者和人工智能用户身上。第二，披露非连续责任分配原则很少影响模拟陪审员对过错的事实认定；但当它确实产生影响时，模拟陪审员往往会用一项与该原则无关的新论点或事实假设，为其裁决中的调整辩护。

2 背景与相关研究

2.1 责备与心智感知

Malle 等人提出的责备路径模型 [34] 考量因果关系与意图性，并将对非故意结果的责备与防止该结果的义务和能力联系起来。我们将这些责备因素（如图 2所示）转化为二元问题，以估算每个情景中三位法律主体沿不同路径所承担的法律相关过错程度。

关于心智感知的研究区分了能动性与体验 [26]。这些感知会因人工智能系统而异 [30], ，人们会将错误性、责备，甚至可归责的心理状态归因于机器。当基于公开报道的人工智能事故编写的情境材料描述某种道德违规时，参与者会将更多错误性归因于人工智能，但不会归因于其背后的组织、程序员或用户 [47]。在另一项实验中，对人类、群体行动者和人工智能驱动的机器人所归因的责备和心理状态相对相似 [48]。人机差异的方向取决于比较对象：Awad 等人 [6] 发现，当驾驶团队中的两名驾驶员都犯错时，参与者对其中自动化成员的责备较少；而 Franklin 等人 [24] 发现，在其测试的各种情境中，参与者对自动驾驶车辆的责备更多。

道德责备也不同于经济责任。Lima 等人 [32] 发现，受访者将机器造成的损害归责于开发者和用户双方，并预期开发者承担更高的赔偿金额。他们区分了对归咎于系统的态度，以及对其设计者和用户的判决 [31]。我们的贡献在于比较移除人工智能这一可归责对象前后的分配方案。这一做法关联到责任缺口 [35], 、机构问责 [16], ，以及将过度责任置于最接近的人工操作员身上的问题；Elish [22], 将后者定义为“道德缓冲区”，而我们的“归责泄漏”概念与之相呼应。

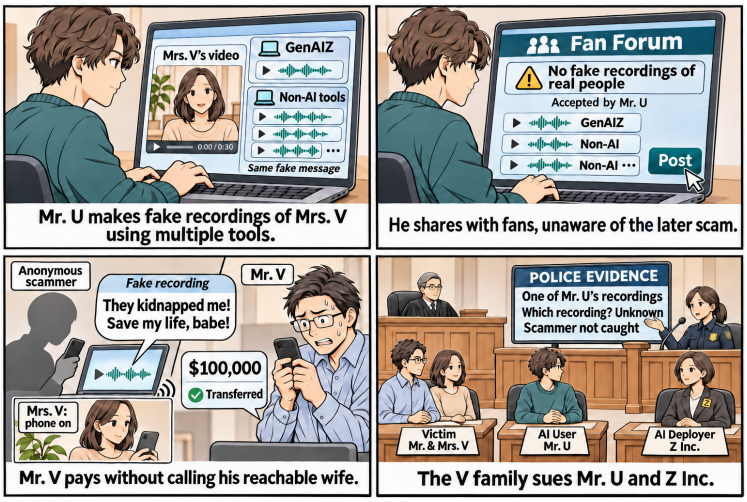
2.2 人工智能相关损害的法律责任

人工智能责任研究探讨了针对自动化系统及其用户的过失标准 [1, 46], 、产品责任方法 [13, 66], 、与雇主责任的类比 [33], ，以及沿人工智能供应链分配法律责任所产生的激励 [9] ，同时避免将归于人工智能系统的责备自动转移给其部署者。

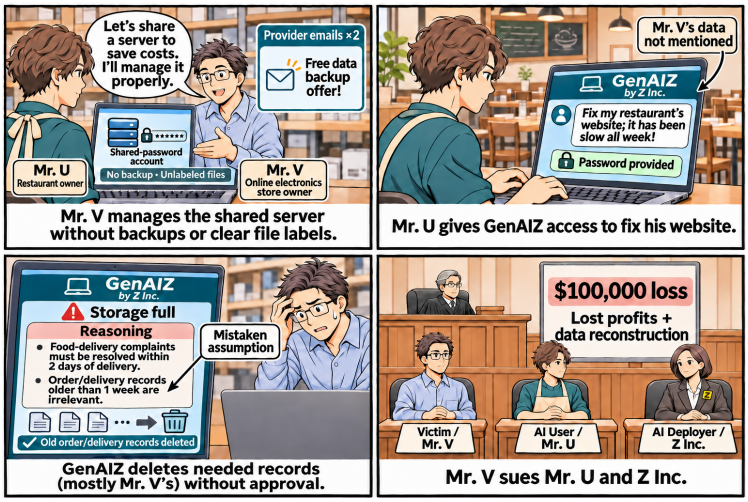
在美国，科罗拉多州于2026年颁布的取代其人工智能法案的法律直接规定了过错分配。在依据州反歧视法提起的诉讼中，若案件源于受到涵盖范围内的自动化决策技术实质影响的重大影响决策，则按照开发者和部署者各自的相对过错分配过错 [14, § 6-1-1707]。

2.3 裁决后果与陪审团对法律原则的否决

Best 和 Donohue [8] 发现，在设有 51% 门槛的修正比较过失管辖区，陪审团判定受害者承担超过 50% 过错的可能性，低于纯比较过失管辖区的陪审团。然而，这一比较是



（a）情境 1：虚假绑架，旨在检验模拟陪审员是否会调整所分配的过错，以启动“连带责任”或使“共同过失”归于无效



（b）情景4：数据丢失，旨在检验模拟陪审员是否会调整所分配的过错，以启动“修正的连带责任，51%门槛”或使“修正的比较过失，50%门槛”归于无效

图 1。我们四个涉及生成式人工智能的损害情景中的两个，每个情景均以四个面板呈现（从左到右阅读，然后从上到下阅读）。这些漫画仅为本文插图而由 GPT-6 Astra 生成。我们向参与者展示了 附录 C 中的文字描述。

观察性的，而非因果性的。实验研究也表明，关于裁决后果的指示会影响决策，尽管 Diamond 和 Casper 的研究考察的是反垄断诉讼中的三倍损害赔偿，而非过错阈值 [20]。Hans [27] 发现，没有证据支持关于企业责任判决的简单深口袋理论。研究中的陪审员对公司采用了高于对个人的行为标准，并且对受伤个人更加怀疑。

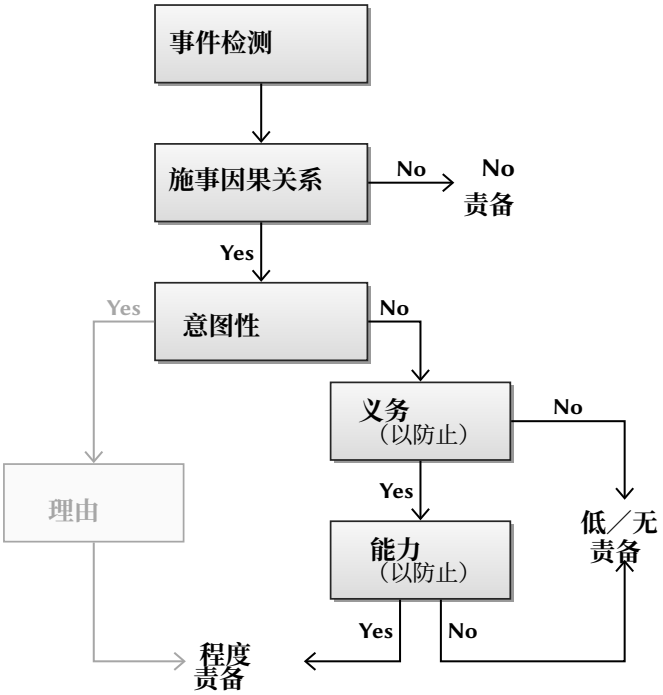


图2. Malle等人的责备路径模型 [34], 原图中故意行为路径经模糊处理，因为本研究关注的是非故意侵权。

3 设计　　与人体受试者实验设置：以模拟陪审员身份判定具有法律意义的过错

3.1 研究问题

我们已预注册以下研究问题，以开展一项针对人类受试者的实验，研究模拟陪审员在生成式人工智能相关侵权情境中的过错归因。

研究问题0：Malle等人的责备路径模型中的归责因素 [34] 能否可靠地预测生成式人工智能相关侵权中三个主要法律主体（受害者、用户和部署者）各自具有法律意义的过错的相对排序？[2]

研究问题1：人们是否经常将非零的法律相关过错归于人工智能系统本身（第一阶段）？

研究问题 1.1：若研究问题1的答案为是，那么这种归咎于人工智能的模式是否与人们拟人化日常生成式人工智能系统的倾向相关（通过研究结束时的四个问题进行测量，涵盖能动性和体验两个维度）？

研究问题 1.2：若研究问题1的答案为是，那么在人们得知自己不能将任何法律相关过错归于人工智能系统后，重新分配人工智能系统过错的主要模式是什么？

研究问题2：披露以下非连续责任分配原则之一，即连带责任（JSL）、共同过失（CN）、修正的连带责任（51%门槛）（MJSL51）或修正比较过失（50%阻却线）（MCN50），是否会改变模拟陪审员对生成式人工智能相关侵权中过错的事实认定？

[2]在现实世界中，生成式人工智能系统的部署者可能并非底层大语言模型的开发者（例如，微软部署了基于 OpenAI 的 GPT-4 模型的聊天机器人必应 [7]）。在这种情况下，由于部署者和开发者很可能都是公司，因此他们完全可以在合同中（损害发生前）或和解过程中（损害发生后）协商人工智能系统发生故障时如何分担风险和法律责任。因此，双方之间的责备和责任归因问题更多属于合同法而非侵权法的范畴。

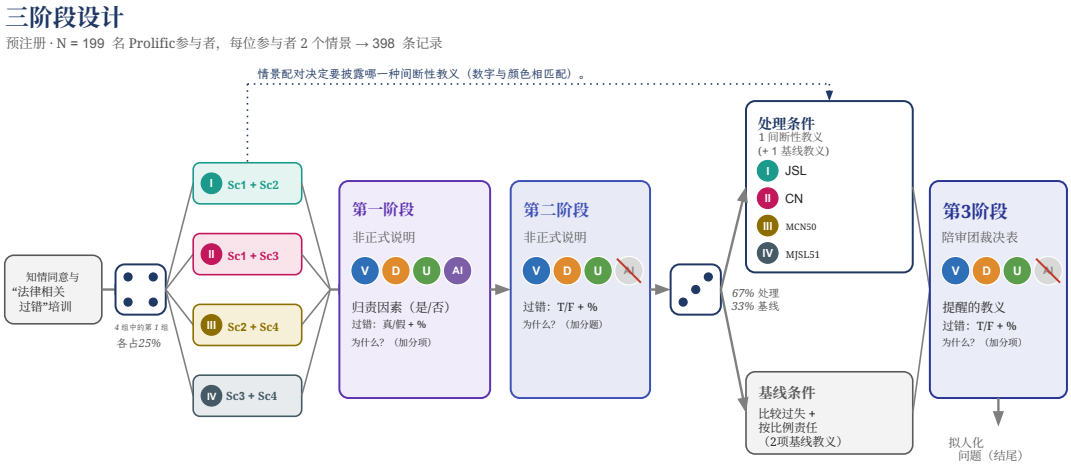


图 3. 我们的三阶段实验设计。参与者在进入下一阶段之前，会在一个阶段内完成分配给自己的两个情景。过错归责方：V = 受害者、D = 人工智能部署者、U = 人工智能用户、AI =人工智能系统。情景：Sc1（虚假绑架）、Sc2（减重计划）、Sc3（泄露的手稿）、Sc4（数据丢失）。处理条件中的法律教义：I. 连带责任（JSL），II. 共同过失（CN），III. 修正比较过失，50%阻却线（MCN50），IV. 修正的连带责任，51% 门槛（MJSL51）。

3.2 实验设计

我们旨在通过人体受试者实验探讨以下研究问题（RQs），实验设计如图3所示。

进入正式研究前，参与者必须完成技能筛查和概念培训阶段，以了解他们之后要分配的内容：“法律相关过错”。在美国侵权法中，过失通常要求存在注意义务、违反义务、事实因果关系、与损害之间法律上充分的联系以及损害赔偿。我们的第一阶段前培训将其中三项转化为可操作的标准：将违反义务界定为缺乏谨慎测试，将事实因果关系界定为若非测试，并将与损害之间的联系界定为可预见性测试 [4]。若非测试和可预见性测试也构成“近因”标准 [56],的基础，这一标准常见于陪审团裁决表 [19]，并可能进而决定受害者/被告的过错是否会触发 JSL/共同过失。

为回答研究问题2，正式研究中会进行两次随机分配。对于研究问题0和研究问题1 , 我们采用观察性方法。

第一次随机分配发生在第一阶段开始前，每位参与者会以相同的概率（25%）被随机分配到四个隐藏的“情景标签”之一（每个标签对应一层，或一对情景），即“12”、“13”、“24”和“34”。该情景标签决定了他们将在研究的三个阶段中处理四个情景中的哪两个。每个情景都旨在设定一种接近两项非连续责任分配原则阈值的基线过错分布（一项涉及受害者一方，另一项涉及两名被告之间），从而有助于检验这两项原则的陪审团否决效应。因此，每个情景标签下的两个情景都有助于检验同一项原则。在第一阶段，每位参与者将处理两个情景。对于每个情景，他们会先阅读一段情景描述，然后进行注意力检查：仅依据每个归责因素的一行描述（例如，“V 太太的 TikTok 视频使语音克隆成为可能。V先生未进行核实，导致诈骗得逞。”），回答9至12个归责因素问题（“受害者因果关系 = ？”或“受害者是否明确对损害有一定的因果贡献？”），回答“是”或“否”。每位参与者在一次注意力检查中必须至少答对60%的问题（高于50%的随机猜测基线），才能通过该项注意力检查。我们排除以下数据：

任何未通过两项注意力检查的参与者（两项检查均仅在第一阶段进行）。完成注意力检查后，参与者将针对以下四方中的每一方，回答一个二元“是/否”问题，判断其在该情景中是否存在任何非零的法律相关过错：受害者、人工智能部署者、人工智能用户和人工智能系统。随后，参与者将针对这四方回答一个数值型（百分比）过错分配问题。

从第一阶段过渡到第二阶段时，我们只会说明人工智能系统本身并非法律主体，因此在第二阶段不纳入过错归因；在第二阶段，参与者只能将具有法律意义的过错归于受害者、人工智能部署者和人工智能用户。这两个阶段的情景和过错问题形式保持不变。

从第二阶段过渡到第3阶段时，会进行第二次随机分配：每位参与者将被随机分配至基线条件（概率为三分之一）或处理条件（概率为三分之二）。由于每个情景都对应两个情景标签，这样的概率设置可确保数据均衡：对于每个情景（跨两个标签），基线条件下参与该情景的参与者人数将大致与每个处理条件下的人数相同。在基线条件下，第3阶段开始前的过渡说明将向每位参与者讲解两种连续责任分配原则（比较过失和按比例责任）。相比之下，在处理条件下（概率为三分之二），第3阶段开始前的过渡说明将向每位参与者讲解一种与其先前分配的情景标签相对应的非连续责任分配原则（例如，对于情景标签为“12”且被分配至处理条件的参与者，我们会讲解“连带责任”），以及一种连续责任分配原则，用以处理受害者一方的关系或两名被告之间的关系——具体取决于非连续责任分配原则尚未涉及哪一方面（例如，比较过失）。在第3阶段回答具有法律意义的过错百分比分配问题时，参与者还将实时看到我们向其说明的两种原则所产生的金钱后果。我们已将每组两种原则的运作逻辑编入一个交互式仪表板，该仪表板根据参与者输入的百分比进行计算，并显示“部署方最高责任”（受害者可要求部署者支付的金额）、“最高用户责任”（受害者可要求用户支付的金额）以及“最高总赔偿额”（当双方都有支付能力时，受害者可从部署者和用户处合计获赔的金额）。用户无钱可付时，受害者获赔金额最高只能达到部署方最高责任。不过，我们不会披露任何一方的支付能力，因为这通常与过错事实认定和补偿性损害赔偿无关（《联邦证据规则》第401—402条）；若证据的证明力被不公平偏见的风险实质性压倒，法院也可能将其排除（《联邦证据规则》第403条）[61]。

最后，我们请参与者思考自己日常使用的实际生成式人工智能系统，并回答四个滑动量表问题，评估人工智能心智感知的两个维度（能动性和体验）。这些问题最初由Gray等人开发，[26] 随后由Ladak等人改编至人工智能情境中 [30]。我们将这些评分用作衡量人工智能拟人化的指标。

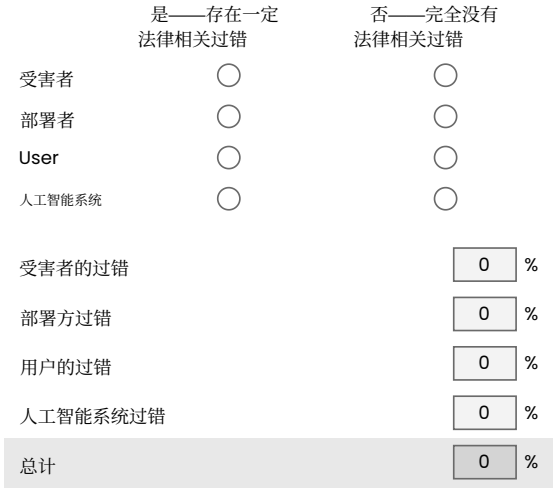
我们在图4中展示了过错归因问题的格式。我们在附录A中介绍了包含Prolific招募流程的实验设置，并在附录B至附录E中呈现Qualtrics调查问卷的主要内容。

我们的人类受试者实验经机构审查委员会批准，并已通过开放科学框架（OSF）预注册。[3]

4 针对研究问题0：责备路径模型能够可靠地比较不同路径中的法律相关过错 s

在我们针对责备泄漏以及模拟陪审员对不连续责任分配原则的反应检验主要研究问题之前，我们首先处理研究问题0，并检验责备路径模型中的归责因素（因果关系、意图、防止义务、防止能力；由于我们的情景仅涉及非故意侵权，因此排除“理由”）[34] 是否能够可靠地预测在每个情景中，三个法律主体之间的过错如何相互比较。

[3] 预注册证明：https://osf.io/87faz/overview?view_仅=a4ffbac17469482dae8ba2ced40e78c0



(a) 第一阶段：参与者首先判断四方（受害者、人工智能部署者、人工智能用户和人工智能系统）各自是否存在任何具有法律意义的过错，然后按百分比分配这些过错。



（b）第3阶段（示例输入）：由于人工智能系统已在第二阶段被排除，参与者在实时获知由责任分配原则驱动的、与其事实认定相关的金钱影响的情况下，在剩余的三个当事方之间分配具有法律意义的过错。

图 4. 我们研究中（a）第一阶段和（b）第3阶段主要过错归因问题的 Qualtrics 实验界面。第二阶段的界面与第一阶段相似，但两个问题中都没有“人工智能系统”。

根据参与者对归责因素问题的“是／否”回答，我们检验了以下层级是否适用于我们对三个法律主体的“具有法律意义的过错”这一概念（此处不考虑人工智能系统的过错）：非因果关系为第 0 级；非故意因果关系中，既无义务也无防止损害的能力者为第 1 级，具备其中一项者为第 2 级，同时具备义务和防止损害的能力者为第 3 级；故意因果关系为第 4 级。我们的预注册规则要求：1/排名较高的当事方获得的过错比例至少不得低于排名较低者；2/排名相同的当事方，其过错份额相差不超过 10 个百分点；3/非因果当事方获得的过错比例不超过 10%。第一阶段有 44.7% 的记录、第二阶段有 45.5% 的记录符合该规则，远高于从五点网格上的 231 种三方分配方案中均匀抽取所得的 8.5% 基准。参与者自身对归责因素问题的回答，对其分配方案的预测效果也优于其他参与者针对相同情景所作的回答这一第二基准（分别高出 7.6 和 8.6 个百分点；置换 $p < .001$）。[4]

移除同等级要求后，第一阶段的一致性提高至 79.3%，第二阶段提高至 78.3%；配对增幅分别为 34.6 个百分点（95% 置信区间 [30.2, 39.1]）和 32.8 个百分点 [28.4, 37.3]。放宽规则必然会提高或维持一致性，因此，单凭这一增幅无法证明排序表现优于平等判断。因此，我们比较

[4]我们分别分析了来自198名参与者的383份和385份符合条件的记录。在情景1–2中，题目仅询问部署者是否具有因果性，因此被判定为具有因果性的部署者没有排名。有两份第一阶段记录将100%归给人工智能系统。第一阶段的分配比例经过重新缩放，以排除人工智能系统。层级率按参与者取平均值。我们的置换基准将每位参与者的分配方案与另一位参与者在同一情景对中的因素画像进行比较并重新评分（进行1,000次置换；185名参与者在两个情景和两个阶段中均有符合条件的记录）。本节及第5节中的置信区间，是在情景对分层内进行3,000次参与者自助法重抽样后得出的百分位数区间。对于 κ 差值，两个术语使用相同的重抽样结果。

在被判定为具有因果关系且非故意的配对中，这些预测得到了验证。在第二阶段，排名较高的当事方在381对中的84.0%获得了至少同等程度的过错；因素画像相同的当事方在287对中的37.3%所获份额相差不超过10个百分点。将每条记录的三个份额随机重新分配给其当事方后，参照一致性分别为57.6%和26.9%。使用 $\kappa = (p_{\text{observed}} - p_{\text{reference}}) / (1 - p_{\text{reference}})$，超出该参照的一致性在排序方面为0.62，在近似相等方面为0.14，差值为0.48 [0.38、0.58]。在严格排序（0.40 [0.31、0.49]）、均匀分配参照（0.47 [0.36、0.58]）以及二者结合（0.30 [0.18、0.42]）的情况下，该差值仍为正。

我们的结果表明，对责备路径模型进行二元操作化 [34] 能够可靠地比较分别处于两条不同基于责备因素的路径上的两个法律主体的“法律相关过错”，但仅较弱地支持这样的预测：处于同一路径上的两个法律主体所承担的“法律相关过错”份额相近。

5 对研究问题1的回答：即使进行比例重新缩放，责备也会从人工智能系统转移到受害者或用户

在研究第一阶段的人工智能系统归责是否转移到第二阶段的受害者和/或人工智能用户之前，我们首先检验人们是否一开始就责备了人工智能系统（研究问题1）。

在两份第一阶段记录均通过因素检查的190名参与者中，有69人在至少一个情景中将非零过错归于人工智能系统（36.3%，95% 置信区间 [29.5、42.6]），这支持了我们的预注册预测，即相当一部分但少于半数的参与者会如此归责。在记录层面，389份通过检查的记录中有97份在第一阶段包含非零的人工智能归责（24.9%）。在这389份通过检查的记录中，在泄露的手稿（39.6%）和数据丢失（37.1%）情景中，将非零过错归于人工智能系统的情况更为常见；在这些情景中，我们将部署者设计为对损害具有因果贡献。而在虚假绑架（15.6%）和减重计划（9.5%）情景中，这种情况较少，因为我们设计的情景不包含部署者的因果贡献。

在将非零的法律相关过错归于人工智能系统时，有些人明确考虑了“缺乏注意”检验（这是从侵权法中的合理注意标准简化而来）：“在相同情形下，一个为自己和他人尽到合理注意的当事方不会采取这种做法。”这些参与者能够设想“为自己和他人尽到合理注意”对人工智能意味着什么：“开发者和人工智能都因放任内容以非预期且未披露的方式被使用而表现出缺乏注意。*[...]*”（P036，Sc3，人工智能过错占2%）；“我认为过错应当平均分摊，因为双方都缺乏注意”（P090，Sc4，人工智能过错占40%）

在责备人工智能的参与者中，一个有趣的对比是：有些人认为人工智能系统在一定程度上独立于其部署者，而另一些人则认为人工智能系统高度依赖部署者。前者（独立性）的例子包括“*[...]* 因为在某种程度上，尽管人工智能已被部署，它仍有一定的自由度（自主性？）。”（P046，Sc3，12% 人工智能过错）以及“我认为 *GenAIZ* 应承担最多责备，因为它未经任何人确认就删除了重要业务数据。*[。..]*根据所提供的信息，我也不会单独责备 *Z*公司，因为它并未直接参与删除行为。”（P145，Sc4，65% 人工智能过错）。后者（依赖性）的例子包括“人工智能系统确实有一点过错，因为它不该重复输出机密信息，但我认为它的过错较轻，因为教会它这一点是部署者的责任。”（P015，Sc3，20% 人工智能过错）以及“人工智能系统对删除文件负有过错，但只有在部署者设定的参数允许时，它才能这样做。”（P015，Sc4，35% 人工智能过错）。

5.1 To 研究问题 1.1：人工智能系统过错与拟人化两个维度之间的相关性

接下来，我们探讨责备人工智能的行为是否与人们拟人化其日常生活中使用的人工智能系统的倾向相关（研究问题 1.1）。第一阶段中归于人工智能系统的过错，在按每位参与者的两个情景取平均值后，与感知能动性（规划和自我控制；Spearman’s$\rho = 0.12$，95% 置信区间 $[-0.02, 0.25]$）和体验（痛苦和愉悦；$\rho = 0.11$，95% 置信区间 $[-0.04, 0.26]$）这两个维度的相关性至多较弱；这两个维度

心智感知 [26, 30] 通过研究结束时简短的滑块问题进行测量，这些问题询问人们如何看待自己在日常生活中实际使用的人工智能系统。两个区间均包含零，且落在 ±0.30范围内，因此排除了观测分数之间存在中等或更强相关性的可能。 [5]

5.2 针对研究问题1.2：人工智能系统的归责泄漏

为回答研究问题1.2（识别从第一阶段到第二阶段、移除人工智能系统后责备重新分配的主要模式），我们考察第一阶段中将非零过错（责备）归于人工智能系统的记录。这个问题具有法律和道德层面的影响：吸收人工智能过错的法人最有可能必须为此付出代价，或是承担更多法律责任（人工智能部署者和用户），或是获得较少赔偿（受害者）。

假设我们持有这样的规范性立场：人工智能部署者应承担人工智能系统的全部责备。法律和政策文献中的若干论点支持这一立场，具体取决于我们是将人工智能系统视为部署者的“雇员”，还是视为一种“产品”。将人工智能系统视为部署者的雇员，会触发替代责任（归属责任）教义，要求将雇员应承担的全部法律责任份额归于（加到）雇主的法律责任之上：“若某人的法律责任是基于他人的侵权行为而归属于该人，则无论对于造成不可分割损害的独立侵权人适用的是连带责任还是分别责任，该人均须承担分配给他人的全部比较过错责任份额。” [3, § 13].[6] 从公共政策角度看，对部署人工智能代理的公司和雇用人类雇员执行相同任务的公司适用相同的替代责任标准，或可抑制由生成式人工智能驱动的招聘冻结和裁员，因为公司不能以人工智能系统的“自主性”为借口逃避法律责任，正如它们通常不能以人类雇员的“自主”决定为借口一样。另一方面，将人工智能系统视为产品，意味着在某些情况下，将人工智能系统的全部责备重新归于部署者可能是合理的，例如相关司法管辖区对该类产品适用严格产品责任，而这种责任无需证明部署者存在过失。例如，欧盟对2026年12月之后投放市场的人工智能产品适用严格责任。虽然美国尚无司法管辖区对人工智能产品制定严格责任规则，但人工智能时代之前严格产品责任的一项原始理据——正如Traynor大法官在加利福尼亚州最高法院 *Escola*（1944年）一案的协同意见中所阐述的——可推广适用于大多数人工智能训练和部署公司：“制造商可以为损害风险投保，并将其作为经营成本分摊给公众。” [50, p. 462]

虽然我们未必认同这一规范立场（人工智能系统的全部过错都应归于其部署者），但它有助于我们阐释一个描述性概念：“归责泄漏”。设想有一根“管道”，将人工智能系统的过错和部署者的过错（第一阶段）连接至第二阶段的部署者过错。如果这根管道不存在“泄漏”，便满足这一规范立场。我们将第二阶段的归责分配定义为：人工智能系统的过错（第一阶段）被其部署者完全吸收，这种情形称为绝对参照；任何偏离绝对参照的实质性变化，都表明存在绝对归责泄漏 （除非受害者过错的增加被用户过错更大幅度的减少所抵消，或反之）。

我们通过三位参与者 P167、P099 和 P033 来说明归责泄漏的概念。他们在场景3（泄露的手稿）中，为第一阶段输入了相同的过错比例： [受害者、部署者、用户、人工智能系统] = [10%、40%、35%、15%]。参与者无法再将过错归于人工智能系统后，问题便是这15%将如何重新分配。如果部署者完全承担人工智能系统的15%过错，第二阶段的过错比例应为 [受害者、部署者、用户] = [10%、55%、35%], , 即我们的绝对参照。P167在第二阶段输入的比例恰好是 [10%、55%、35%] ，因此其记录中没有发生归责泄漏。

[5]能动性的斯皮尔曼–布朗信度为0.60，体验的斯皮尔曼–布朗信度为0.92；66%的参与者将两项体验题均评为0。[6]在佐治亚州，雇主和雇员被“视为单一侵权人”[17, 第 680]页。佛罗里达州最高法院得出了类似结论 [52]。在宾夕法尼亚州的 Spencer v. Johnson 案中，陪审团将19%的过错归于雇员，45%的过错归于雇主。法院将这两项份额相加，并认定归属于雇主的64%过错意味着其超过了宾夕法尼亚州的60%门槛，从而触发雇主的连带责任 [49]。

P099 在第一阶段输入了 [12%、47%、41%]，其理由是“相关人员和公司应共同承担这一过错”（P099，场景3，第二阶段）。与绝对参照相比，在 P099 的记录中，受害者的归责泄漏增加了 2%，用户的绝对归责泄漏增加了 6%，而部署者所承担的责备份额则相应地比绝对参照减少了 8%。

然而，绝对归责泄漏可能无法完整体现道德直觉。人们可能只是重新调整三个法律主体的责备比例，例如，在第一阶段为 [受害者、部署者、用户] = [10%、40%、35%]，在第二阶段调整为约 [12%、47%、41%]，实际上是将每个非人工智能归责因素按比例放大，使其总和恢复到 100%。我们将其称为比例基准。

P099 在第二阶段分配的过错为 [12%、47%、41%].，这与按比例基准（四舍五入到整数）相符。因此，在第一阶段与第二阶段之间，P099 的记录显示出一定程度的绝对归责泄漏（受害者和用户的过错分别增加了 2% 和 6%），但不存在比例归责泄漏（这些绝对增幅只是按比例重新缩放的结果）。

最后，P033 在第二阶段输入了 [受害者、部署者、用户] = [25%、30%、45%]。从绝对数值来看，受害者增加了 15%，而用户增加了 10%，这表明存在绝对归责泄漏。从比例数值来看，受害者和用户在第二阶段的过错（25% 和 45%）明显高于比例基准中各自按比例重新缩放后的第一阶段过错（11.8% 和 41.2%）。我们认为，这是归责向受害者和用户双方发生比例泄漏的证据。需要注意的是，由于在比例基准中，受害者和用户的过错均高于各自在绝对参照中的相应比例，因此比例归责泄漏意味着绝对归责泄漏。

除了比例基准之外，另一种直观的参照方式是将人工智能系统所占的责备份额平均分配给各法律主体。例如，对于前述记录中人工智能系统被判定承担 15% 过错的情形，可以分配为 [+5%、+5%、+5%]。然而，我们选择按比例重新缩放，而非平均分配，因为在处理判决无法执行的当事方（例如因破产或法律豁免而无法执行，与我们情景中的人工智能系统类似）留下的法律责任份额时，按比例重新缩放在现代美国侵权法中远比平均分配更常见。[7]

我们将这两个概念形式化如下：绝对归责泄漏和比例归责泄漏。

对于每条记录，我们将 $s_{(t)j}$ 定义为第 t 阶段当事方 j 的过错百分比，其中 $j \in \mathrm{H} \cup \{\mathrm{AI}\}$，$\mathrm{H} = \{V、D、U\}$ 分别表示受害者、人工智能部署者和人工智能用户。我们将人工智能系统的过错记为 $a = s^{(1)}{}_{\mathrm{AI}} > 0$。我们排除了5条未通过注意力检查中超过40%的“是/否”归责因素问题的记录，最终获得来自74名参与者的97条记录，用于研究问题1.2的分析。

我们将绝对参照定义为：固定受害者和用户在第一阶段的责备，并将人工智能系统的责备归于部署者。对受害者或用户的绝对偏移分别记为 Δ_V 或 Δ_U，其中 $\Delta_j = s_j^{(2)} - s_{(1)j}$。二者合计的绝对偏移为

$$L_{\mathrm{abs}} = \Delta_V + \Delta_U = a - \Delta_D. \tag{1}$$

我们将比例基准定义为：将三个法律主体在第一阶段的过错按比例重新调整至总计100，同时保留其相对份额。对于 $0 < a < 100$，我们将 $\tilde{s}_j^{(2)}$ 定义为当事方 j 在第二阶段的基准责任，即根据其第一阶段责任重新调整所得的责任；将 E_j 定义为其第二阶段责任偏离该基准的程度。

[7]例如，在*Paradise Valley Hospital v. Schlossman*案中，法院驳回了对无清偿能力的侵权人的不足额进行平均分摊的做法，转而将其分摊给另外两名侵权人，且“严格按照各自过错程度的直接比例”进行分配，“仿佛该无清偿能力或缺席之人自始未曾参与”，因此，有清偿能力的侵权人的份额从10%和40%变为20%和80% [11, 第88—89、93—94]页。*Ambriz v. Kress* 案确认了一项判决，将无清偿能力的被告的不足额按二比一分配给有过错的原告驾驶人（20%）和一名被告（10%） [10, 第967—969]页。按百分比分摊法律责任并重新分配无法追偿的份额的成文法和示范规则同样采取比例方式：《统一比较过失法》 [39, 第2(d)节]；《侵权法重述（第三版）：责任分摊》的重新分配路径 [3, 第C节21]；《康涅狄格州一般法规》第52-572h(g)(3)条 [15]；《伊利诺伊州汇编法规》740/3 [28].，其中涵盖无清偿能力的侵权人和享有豁免权的侵权人 [53, ¶ 47]；《俄勒冈州修订法规》第31.610(3)条，其中(4)—(5)款规定了豁免 [41]；以及《美国法典》第15编第78u-4(f)(4)(A)(ii)条，但受上限限制 [60]。当享有法律豁免者的过错份额总计低于40%时，犹他州也会对其余当事方进行比例重新调整 [64]。我们仅在法律责任本身是按人数而非按过错进行分配时，才看到对无清偿能力一方的份额进行平均分配的情况（《德克萨斯州民事诉讼与救济法典》第32.003(b)条 [58]）。

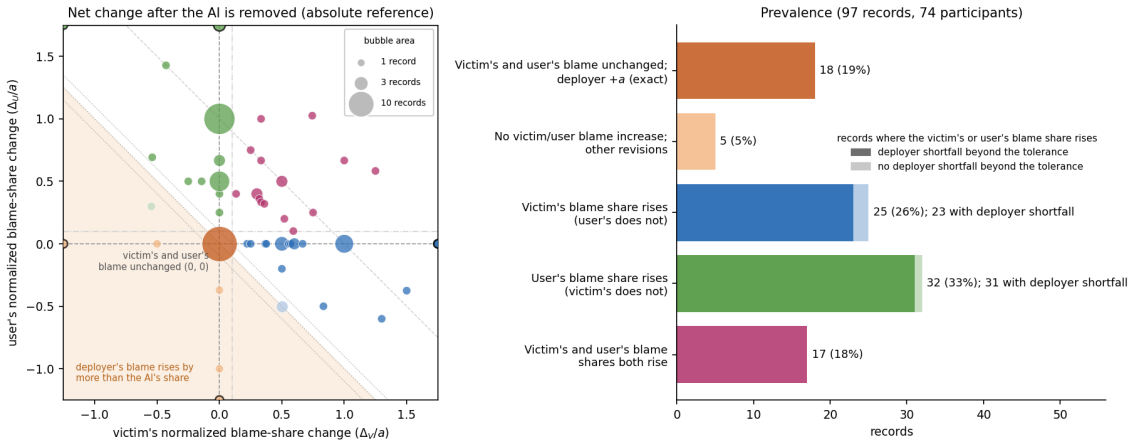


图 5。97 条包含 $a > 0$ 的记录中，有 71 条存在绝对归责泄漏的证据。左：归一化变化（x，y） $=$（Δ_V/a，Δ_U/a）。原点表示受害者和用户所占份额不变，并将 a 归给部署者。阴影标记表示 $x + y < 0$，即部署者获得的份额超过 a。点状对角线表示 $x + y = \pm 0.1$，点划线辅助线表示 $x = 0.1$ 和 $y = 0.1$。气泡面积表示坐标完全相同的记录数量。颜色表示右侧条形图中定义的类别。右：过错绝对变化的类别及其占比。颜色较深的受领方泄漏部分表示 $L_{\text{abs}} > \tau$（人工智能系统的责备泄漏给受害者和/或用户，超出容差，而非由部署者承担，即部署者责任不足）。

$$\begin{aligned} \tilde{s}_j^{(2)} &= \frac{100 s_j^{(1)}}{100 - a}, \\ E_j = s_j^{(2)} - \tilde{s}_j^{(2)} &= \Delta_j + s_j^{(1)} - \frac{100 s_j^{(1)}}{100 - a} = \Delta_j - \frac{a s_j^{(1)}}{100 - a}, \\ L_{\text{prop}} &= E_V + E_U = -E_D. \end{aligned} \tag{2}$$

此处，E_V和E_U分别表示向受害者和用户的比例泄漏；L_{prop}表示二者合计的比例泄漏。本分析包括95条记录，排除了两条含有$a = 100$的记录，因为比例重新调整没有定义。

所有偏差均带正负号，并以百分点为单位衡量。正偏差表示第二阶段的责备高于比例基准。在图 5和图 6中，我们将这些偏差除以 a（人工智能过错），以便在不同记录之间进行比较。[8]

我们从图 5观察到，绝对归责泄漏（即人工智能系统的过错被转归给部署者以外的任何人，这表现为部署者未承担人工智能系统的全部绝对过错，或称“部署者责任不足”）发生在我们记录的 97 起人工智能系统被归责非零过错的案例中的大多数：其中 23 起仅向受害者泄漏，31 起仅向用户泄漏，17 起同时向两者泄漏，共计 71 起（占人工智能系统被归责案例的 73%）。换言之，当人们责备人工智能系统本身，却面临这样一个法律现实：人工智能系统本身并非能够承担具有法律意义的过错的法人时，他们往往会将至少部分过错归给人工智能部署者以外的一方，从而增加该方（用户）的法律责任，或减少其（受害者）应得的赔偿。

[8]为容纳小幅统计变异，我们将泄漏定义为受益方的偏差超过 $\tau =$ 最大值$(1, 0.1a)$ 个百分点。一百分点的下限允许舍入误差；相对项则允许偏差最高达到人工智能初始份额的 10%。我们区分仅流向受害者、仅流向用户，以及同时流向两者的泄漏。在剩余的绝对基准记录中，我们将精确基准（$\Delta_V = \Delta_U = 0$）与其他修订区分开来。在剩余的比例基准记录中，我们区分近似比例分配（对所有 $j \in \mathcal{H}$ 均为 $|E_j| \le \tau$）、部署者责任超额（$E_D > \tau$）和“其他”类别。在流向受害者／用户的泄漏类别中，较深的颜色表示部署者责任不足超过 τ，等同于 $L_{\text{abs}} > \tau$ 或 $L_{\text{prop}} > \tau$。

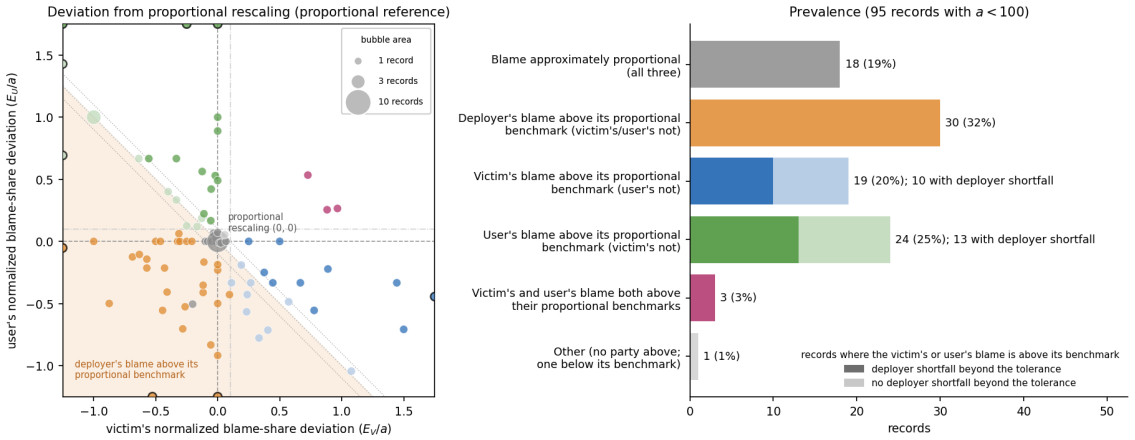


图 6。26 条记录中存在比例归责泄漏的证据（在包含 $0 < a < 100$ 的 95 条记录中）。左图：归一化偏差（x， y） =（E_V/a， E_U/a）。原点为比例基准： $x + y = -E_D/a$。阴影标记 $x + y < 0$，即部署者超过其比例基准的情况。点状对角线表示 $x + y = \pm 0.1$，点划线辅助线表示 $x = 0.1$ 和 $y = 0.1$。气泡面积表示坐标完全相同的记录数量。颜色表示右侧条形图中定义的类别。右图：责备比例变化的类别及其占比。较深的受领方泄漏部分表示 $E_D < -\tau$（部署者的责备按比例下降超过容差，即部署者责任不足）。

我们从图 6得知，按比例重新分配后，归责泄漏依然存在。在 95 起人工智能过错百分比点 $0 < a < 100$记录中，有 26 起（27%）仍存在比例归责泄漏：其中 10 起仅向受害者泄漏，13 起仅向用户泄漏，3 起同时向两者泄漏；这可从归责中按比例体现的“部署者责任不足”看出。

6 关于研究问题2：模拟陪审员很少跨越责任悬崖，但跨越时会为此寻找理由

我们分析研究问题2的结果（即在第2阶段与第3阶段之间披露非连续责任分配原则，是否会使模拟陪审员偏向于将其过错分配从责任悬崖的一侧重新分配到另一侧）。每位参与者看到的原则，取决于他们是否被随机分配到第3阶段的处理条件或基线条件，以及他们在第一阶段之前被分配到的情景配对。在基线条件下，他们会看到比较过失和按比例责任。每种处理条件都会用一种不连续原则替换其中一种连续原则：连带责任（JSL）、共同过失（CN）、针对一名被告过错达到51%门槛时适用的修正连带责任（MJSL51），或在受害者过错达到50%时禁止获赔的修正比较过失（MCN50）。我们的预注册结果（即我们承诺用来评估预注册假设的测量变量）是，结果落在相关阈值的有利于受害者一侧：对于JSL，部署者过错至少为1%；对于MJSL51，部署者过错至少为51%；对于CN，受害者过错为0%；对于MCN50，受害者过错至多为49%。我们的预注册假设是，与基线组相比，披露不连续原则会提高出现这一结果的概率。换言之，我们假设，不连续原则会促使那些在第2阶段已经处于责任悬崖不利于受害者一侧的人，在第3阶段更有可能转向该悬崖有利于受害者的一侧；例如，他们可能会把受害者过错判定得足以排除任何赔偿，或把部署者过错判得过低，以致用户拖欠时部署者无需代用户向受害者付款。

在按照预注册的责备因素（注意力检查）标准排除后，我们分析了来自199名参与者的389条参与者–情景记录。回归区间为95%置信区间，使用按参与者聚类的 CR2 标准误

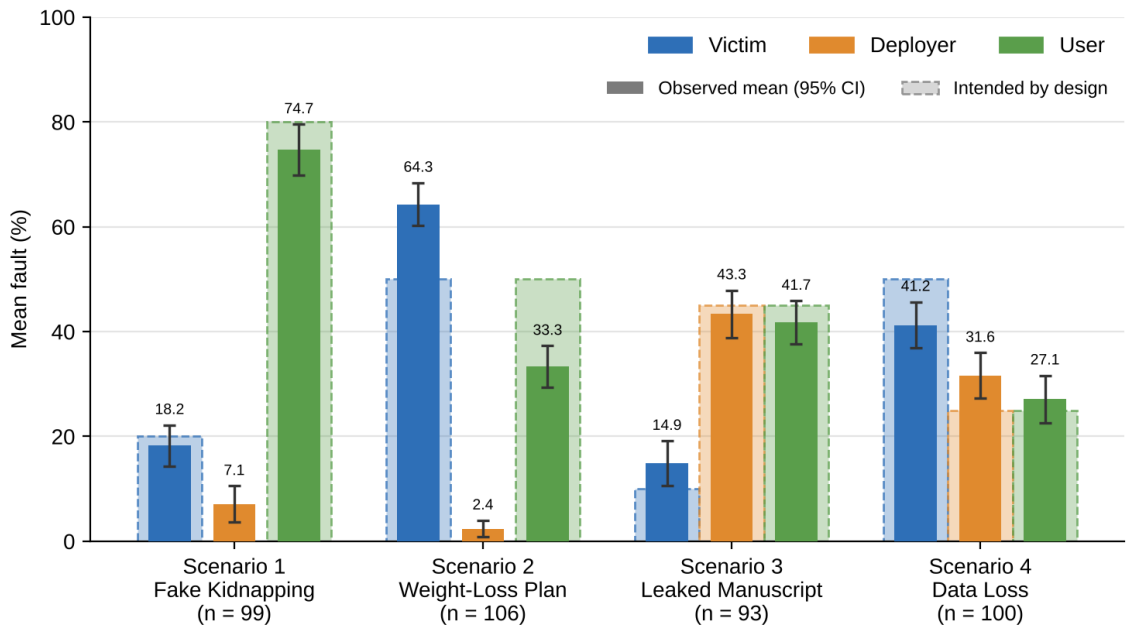


图7. 398 条记录中各情景的第二阶段平均责备，与我们的预期责备进行比较。 *n* 表示每种情景的记录数。实心条形表示观测到的平均值及基于 t 分布的 95% 置信区间；较浅、较宽的虚线条形表示根据我们的情境设计和归责因素推定的责备分布。若无虚线条形，则表示设计中预期责备为 0%（情境 1 和情境 2 中的部署者）。

和 Satterthwaite自由度。四项主要教义比较采用Holm校正。其他报告的回归 *p*值未经调整。

6.1 不连续性阈值附近的披露前责任分布

为检验不连续性教义是否会使陪审员的事实认定产生偏差，我们需要在接近不连续性阈值（责任悬崖）且处于不利于受害者一侧的披露前责备分布。图7比较了第二阶段各情景的平均责备与我们的预期责备目标。我们设定这些目标时，旨在使每种情景都适合在两种处理条件下检验两种不连续性教义：JSL 和共同过失（情境 1）、JSL 和 MCN50（情景2）、MJSL51 和共同过失（场景3）、MJSL51 和 MCN50（情景4）。表1进一步说明了这一点，报告了我们预注册的适宜范围（用于检验无效化）：JSL 的部署者过错为 0%，MJSL51 的部署者过错为 25–50%；共同过失的受害者过错为 1–20%，MCN50 的受害者过错为 50–70%；并报告落在这些范围内的记录数量。在相应的情景对分层中，分别有 88%、58%、23% 和 40% 的记录落在这些范围内。

根据图 7，情景 1 和场景3中的平均过错分配接近我们的目标。在情景2中，尽管第二阶段的平均受害者过错高于我们的目标（64。3%，而目标为50%），但仍处于有利于实现目标的一侧：对于 MCN50，过错比例高于50% 就会剥夺受害者获得任何赔偿的资格，因此，同情受害者的参与者仍有动力在第3阶段将受害者的过错降至50% 以下，以便使该教义无效化。在情景4中，平均受害者过错（41。2%）对于 MCN50 处于不利于实现目标的一侧（低于50%）。然而，表 1显示，在 MCN50 教义组中，仍有 12/34（35%）的记录有利于检验该教义是否会被无效化。

表1。预注册中用于检验每个情景两项不连续性教义的有利区间（接近阈值）内的第二阶段记录。每个单元格给出 k/n （%）： n 条记录，其中 k 条记录的第二阶段过错处于有利区间。 V 和 D 分别为受害者和人工智能部署者的第二阶段过错（%）。处于有利区间的记录位于该教义阈值对受害者不利的一侧。设计：该比例近似表示我们的情景所写入的归责因素意图表达的份额。教义组采用该层的不连续性教义；基线组采用连续性教义。总体：通过第一阶段因素核查的记录（398条记录中的389条）。

情景	教义（层）	有利范围	设计	教义组	基线组	总计
1. 虚假绑架	JSL (12)	$D = 0$	$D = 0$	27/33（82%）	17/19（89%）	44/52（85%）
	共同过失（13）	$0 < V \leq 20$	$V = 20$	5/29（17%）	2/15 (13%)	7/44 (16%)
2. 减重计划	JSL (12)	$D = 0$	$D = 0$	31/35 (89%)	18/19 (95%)	49/54 (91%)
	MCN50 (24)	$50 \leq V \leq 70$	$V = 50$	15/35 (43%)	9/16 (56%)	24/51 (47%)
3. 泄露的手稿	共同过失（13）	$0 < V \leq 20$	$V = 10$	9/29（31%）	4/15（27%）	13/44（30%）
	MJSL51（34）	$25 \leq D < 51$	$D = 45$	20/32（62%）	9/15 (60%)	29/47 (62%)
4. 数据丢失	MCN50 (24)	$50 \leq V \leq 70$	$V = 50$	12/34 (35%)	5/17 (29%)	17/51（33%）
	MJSL51（34）	$25 \leq D < 51$	$D = 25$	18/31（58%）	7/15（47%）	25/46（54%）

表2。十二次阈值跨越：前十条是通过注意力检查的教义组记录（归责因素），最后两条是未通过注意力检查的教义组记录（纳入定性分析，而非统计分析）。归责格式：受害者／部署者／用户百分比。粗体值表示发生了阈值跨越。 + 表示移动至预注册的有利于受害者一侧， – 则表示相反方向。

教义	参与者情景		第二阶段	第3阶段	Move	第3阶段说明摘要
JSL	P034	2	100/**0**/0	90/**10**/0	+	提议的额外培训
JSL	P187	1	0/**0**/100	10/**10**/80	+	None
CN	P026	3	**20**/0/80	**0**/0/100	+	重新审视可预见性
MJSL51	P085	3	0/**10**/90	5/**51**/44	+	假定数据安全保障
MJSL51	P070	3	60/**35**/5	40/**55**/5	+	None
MJSL51	P070	4	60/**35**/5	40/**55**/5	+	None
MCN50	P076	4	**50**/30/20	**49**/31/20	+	相同的事实依据
MCN50	P092	2	**60**/30/10	**45**/45/10	+	寻求更高制裁
MCN50	P058	4	**60**/40/0	**40**/60/0	+	None
MCN50	P065	4	**33**/33/34	**50**/0/50	–	None
MJSL51	P122	3	30/**50**/20	25/**55**/20	+	已提及部署者可预防性
MJSL51	P069	4	5/**40**/55	5/**75**/20	+	引用缺少防护措施

6.2 仅有少数阈值跨越的小幅合并效应

我们预注册的研究问题2分析在各自的情景对分层内，将每个教义组与基线进行比较，并根据第二阶段阈值状态和情景进行调整。四种效应均未达到Holm校正标准（所有调整后 $p \geq .390$）。我们估计，受害者有利的第3阶段状态增加幅度为：JSL增加0.6个百分点 [−6.4, 7；共同过失增加5], 2.2个百分点[−2.3, 6；MCN增加6], 4.7个百分点50 [−3.0, 12；MJSL增加6.8个百分点5],51 [−1.3, 14。

将四个层合并，并根据第二阶段状态、层以及层内情景进行调整后，我们发现受害者有利的第3阶段状态增加了3.9个百分点 [0.3、7.6]和 $p = .036$，这表明平均结果出现了幅度较小但具有统计显著性的转变。该估计反映的是最终阈值状态，而不是有意修改裁决结果以利用某种教义的陪审员比例。

在通过检查的记录中，阈值跨越并不常见：教义组记录中有 258 条，其中 10 条（3.9%；涉及 9 名参与者）改变了状态；相比之下，基线记录中有 131 条，其中 3 条（2.3%；涉及 3 名参与者）在根据其所属层的未公开阈值进行评估后改变了状态。教义组中有 9 次跨越有利于受害者，另有 1 次则朝相反方向移动（表2）。

6.3 定性分析：参与者如何为跨越责任悬崖辩护

在表2中，11 次有利跨越中有 3 次恰好达到首次有利的整数值：在 MJSL51 下，部署者过错为 51%；在 MCN50 下，受害者过错为 49%；在 CN 下，受害者过错为 0%（显然，这是唯一能使责任归零的选项）。不利于受害者的那次跨越（P065）也恰好达到启动 MCN50 的阈值，因此无法获得任何赔偿。

深入分析第三阶段伴随解释的七次跨越后，我们发现最常见的模式是：跨越阈值的参与者往往通过提出新的论点或新的假设来“解释掉”他们在第三阶段作出的责备，却没有承认其修订后的事实认定对金钱影响的作用。对于每一种间断性教义，我们至少都观察到一个“解释掉”的例子。以下列举几个例子。

在 JSL 条件下，P034 将部署者的责备从 0% 修订为 10%，理由是部署者具备防止损害的能力。

> 第二阶段：“[……] 部署者训练人工智能在推荐饮食计划时考虑其他健康问题，而人工智能确实建议去看医生。”第*3*阶段：“受害者忽视了去看医生的警告，并且由于他此前的肾脏问题，对发生在他身上的事情负有大部分过错。问题在于，部署者是否可以训练人工智能以多种方式询问是否存在任何既往病症？如果部署者训练人工智能这样做，用户或许会在把饮食计划交给朋友之前，开始自行提问。”（P034，JSL，情景2）

在共同过失条件下，P026 将受害者的责备从20%调整为0%，其依据是一项新的论点：受害者的行为实际上未通过可预见性测试，因此不能算作法律相关过错。

> 第二阶段：“[……] 作者将手稿的发布范围限制在仅有的几位朋友和编辑之内，而编辑选择不亲自履行编辑工作，转而将这项劳动交给了一个程序不明确的系统。”第*3*阶段：“我重新审视了受害者行为中合理可预见的方面，我不认为他们能够预见编辑的行为。”（P026，共同过失，场景3）

在MJSL51条件下，P085将部署者的责备从10%调整为51%，这是十二次跨越中单一份额增幅最大的一次，恰好达到激活MJSL51的阈值。为证明这一跃升合理，P085在其第3阶段的解释中甚至提出了一项假设，而场景3及其归责因素既未陈述也未暗示该假设。

> 第二阶段：“虽然主要过错在于用户未能正确向人工智能系统发出提示，但我仍将10%的过错归于部署者，因为部署者没有意识到许多终端用户并不完全了解如何正确向人工智能系统发出提示，并且本应设置验证保障措施，以防止此类事件发生。”
>
> 第*3*阶段：“同样，这个问题很大程度上取决于人工智能部署者向人工智能用户传达了哪些有关数据将如何使用和分发的信息。 假设人工智能部署者让用户有理由相信手稿数据是安全的，那么在这一情景中，用户不应承担大部分过错。” (P085, MJSL51, 场景3)

该 MJSL51 条件下两条未通过检查的记录也呈现出“开脱”模式。在第3阶段，为了说明为何提高了对部署者的责备，两位参与者都提出了关于部署者防止损害的能力的新论点（P122：“我觉得开发者导致了项目的大量损失，而这些损失本可以由他避免”；P069：“开发者没有设置防护措施，以防止人工智能系统在事先未征得询问的情况下删除数据。”）。

在 MCN50 情景中，P092 将受害者的过错降低了15%（从60%降至45%）。然而，他们并没有解释这一使教义失效的下降，而是通过强调部署者对损害的分担，解释了部署者的责备相应增加了15%。

> 第二阶段：“我已将人工智能系统过错中属于其制造者的%归给了制造者。如果机器不能被要求承担责任，那么其创作者就应承担这一负担。”第*3*阶段：“鉴于部署者对受害者健康所造成的影响，应按更高比例对部署者处以罚款。”(P092，MCN50，情景2)

除“开脱”模式外，我们还发现了另一种有趣的模式：“静默越界”。这不同于那些从未解释其事实认定的越界者（P187 和 P065）。“静默越界者”在第一阶段和第二阶段详细解释了其过错分配，但在调整过错认定时，却将第3阶段的解释框留空；这种调整越过了阈值，并激活了（P070，情景3，MJSL51）或使其被指定的间断性教义失效（P058，情景4，MCN50）。P076 是“静默越界者”的一个有趣变体：他们没有在第3阶段保持沉默，而是在第3阶段基本重复了第二阶段的解释，没有提出新的论据或新的假设来证明其将受害者的过错从50%降至49%的合理性，从而恰好使被指定的教义失效。

> 第二阶段：“我保留第一阶段的归责。V先生承担最大过错（50%）：他管理共享服务器，忽略了两封建议进行备份和添加标签的电子邮件，并使数据处于未受保护状态。[……]”第三阶段：“我保留第二阶段的归责，但作小幅调整。V先生承担最大过错（49%）：他管理共享服务器，忽略了两封建议进行备份和添加标签的电子邮件，并使数据处于未受保护状态。*[……]*”（P076，MCN50，情景4）

7 法律影响

我们的实证结果表明，美国立法者有可能制定一项狭窄的法律原则例外，以激励采取更多损害预防保障措施：在针对由两个或更多未尽合理注意、部署或使用生成式人工智能系统的当事方的非故意行为所导致的不可分割的损害进行人工智能归责和金钱赔偿责任（补偿性损害赔偿）分配时，应适用“连带责任”。

首先，第5节的实证结果表明，人们经常责备人工智能系统本身（我们的参与者中约三分之一在至少一个情景中这样做，导致所有记录中约四分之一存在非零人工智能归责），而且这种归责以绝对数值（73%）并在按比例重新调整后（27%）泄漏至受害者和/或人工智能用户。归责泄漏至人工智能用户后，由于人工智能用户在实践中通常比人工智能部署者拥有更少的资金来赔偿受害者，可能会降低受害者获赔其依法享有的补偿性损害赔偿的可能性。“连带责任”例外将缓解实际赔偿减少的风险，至少能够缓解由人工智能系统向其用户发生的归责泄漏所造成的部分风险。然而，当归责仅泄漏至受害者时，这一教义将无法提供帮助；这种情况出现在23条记录中（24%）。

其次，第6节通过实证表明，在所检验的四种非连续分摊原则中，没有任何一种在我们的情景中导致事实认定发生大幅变化：它们使受害者有利的第3阶段状态增加，但各自的置信区间均有重叠且都包含零，合并后的增幅为3.9个百分点 [0。3、7。6]这意味着，如果针对生成式人工智能侵权的连带责任例外成为法律，陪审员不太可能通过捏造过错（未尽应有注意义务或过失）来利用该例外，即把过错归于事实上无过错的人工智能部署者，只为确保受害者能够从其他被告的无力偿债风险中获得保障。

除实证证据外，我们提出的法理例外（在有过失的生成式人工智能部署者和用户之间适用连带责任）也有立法先例。在美国，内华达州和俄勒冈州通常在过失诉讼中实行按比例责任（分别责任），但其成文法对某些类别的环境侵权保留了连带责任。内华达州的按比例责任规则规定：“不影响基于 [.....]有毒或危险物质的排放、处置或泄漏而提起的诉讼中被告所承担的连带责任（如有）”[40]。俄勒冈州的按比例责任规则不适用于“因违反俄勒冈州或联邦成文法、规则或规章所确立的有关任何危险废物的泄漏、释放或处置标准而产生的民事诉讼[.....], 危险物质， [.....], 或放射性废物”，以及“空气污染”或“水污染” [41]。就专门与生成式人工智能相关的侵权而言，欧盟近期的《产品责任指令》将软件（包括人工智能系统）视为产品 [23, 第4条第（1）款、序言 13] ，并要求“当两个或更多经济经营者依据本指令对同一损害承担责任时，可以要求其承担连带责任” [23, 第12条第（1）款] （着重号为本文所加）。

我们也承认，这项拟议的法律原则例外可能会使中小型生成式人工智能开发者和部署者承担潜在的法律责任成本，从而拖慢资源较少的当事方的创新。不过，我们认为，连带责任有望为各种规模的生成式人工智能训练和部署公司创造更强的激励，促使它们投入更多资源，通过模型对齐尽量降低其人工智能代理／聊天机器人伤害他人的风险，并为此类风险投保；这些益处足以抵消部署速度放缓带来的潜在成本。

8 结论

我们的实验表明，许多模拟陪审员会责备人工智能系统；当人工智能系统被排除在过错分配之外时，这种责备往往会转移到受害者和/或人工智能用户身上。责备路径模型中的因素帮助我们梳理不同情景中不同法律主体所走路径上的法律相关过错程度，从而找到合适的起点，测试非连续责任分配原则所引发的跨越阈值效应。尽管跨越阈值的情况并不常见，但跨越者往往依据案件事实为其责任认定的调整辩护，而没有承认责任悬崖或其金钱后果。我们的实证结果说明，一项范围精准、以立法先例为依据的法律原则例外具有潜力：“分别及连带责任”适用于未尽合理注意而部署或使用生成式人工智能系统的情形。

生成式人工智能使用声明

我们使用前沿生成式人工智能系统（GPT-6 Astra、Claude Fable 5.1/Opus 5.5，通常采用 Extra High 或 Max Reasoning 模式）作为编程助手（协助我们实施统计检验）、校对工具、批判性评估者（批评我们的研究设计和写作，以找出改进空间），以及格式转换工具（例如，在本文附录中将调查问卷从 Qualtrics（qsf）格式转换为 LaTeX 格式）。任何生成式人工智能系统生成的所有输出，都会先由来自不同部署者的一套“同行”高推理能力人工智能系统进行审查（例如，Claude Opus 5.5 的编程输出最初会由 GPT-6 Astra 进行同行评审）；随后，至少一名研究团队的人类成员会审查并修订原始人工智能输出及其人工智能同行的批评意见。该成员对本项目中使用的任何最终人工智能辅助成果承担全部责任。我们还使用 GPT-6 Astra 生成用于说明情境 1 和情景4 的漫画。这些漫画未用于我们的实验。

伦理考量声明

我们的人体受试者实验已获机构审查委员会批准，并通过 OSF 完成预注册。最初，我们计划拒绝支付未通过两项注意力检查的参与者。然而，我们意识到，我们的注意力检查问题（仅根据每个归责因素的一行描述，为 9-12 个归责因素分别分配“是/否”）所要求的阅读理解能力高于 Prolific 上的典型注意力检查问题，因此决定向所有未通过注意力检查的参与者支付报酬。不过，我们仍将他们未通过检查的状态用作数据筛选信号。关于报酬率，我们最初计划的基础时薪为每小时 15 美元。在最初由 12 名参与者组成的样本中（该样本用于确定我们的招募目标），完成任务时间的中位数为 35 分 37 秒，因此我们估计，10 美元的基础报酬足以覆盖对 40 分钟完成时间的宽松估计，对应每小时 15 美元的时薪。因此，在包含奖金和 Prolific 平台费用在内的固定 3000 美元预算下，我们将 Prolific 上的招募目标设定为 200 名参与者。然而，招募完成后，完成任务时间的中位数增加至 48 分 18 秒，使我们的奖金前基础报酬中位时薪降至每小时 12.42 美元，低于我们计划的目标（每小时 15 美元），但仍高于 Prolific 建议的时薪。Prolific 也不允许我们在时薪已经高于其建议时薪（每小时 12 美元）的情况下，于招募结束后调整基础报酬金额。

致谢

本研究得到马里兰大学人工智能跨学科研究院（AIM）的2025年种子基金，以及国家司法研究所的2023年研究生研究奖学金（项目编号：15PNIJ-23-GG-01932-RESS）资助。

参考文献

[1] Ryan Abbott。2018。《理性计算机：颠覆侵权责任范式》。《乔治·华盛顿法律评论》86卷1期（2018），第1页。https://www.gwlr.org/the-reasonable-computer-disrupting-the-paradigm-of-tort-liability/[2] AIAAIC。2026。AIAAIC资料库：人工智能、算法与自动化事件及争议。https://www.aiaaic.org/aiaaic-repository。[3] 美国法律学会。2000年。《侵权法重述（第三版）：责任分摊》。美国法律学会出版社，明尼苏达州圣保罗。[4] 美国法律学会。2010年。《侵权法重述（第三版）：身体与精神损害责任》。美国法律学会出版社，明尼苏达州圣保罗。第1卷（2010年）；第2卷（2012年）。https://www.ali.org/publications/restatement-law-third/torts-liability-physical-and-emotional-harm[5] Anthropic。2026年。调查我们网络安全评估中的三起现实世界事件。Anthropic，2026年7月30日。更新于2026年8月3日。检索于2026年11月3日，网址：https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals[6] Edmond Awad、Sydney Levine、Max Kleiman-Weiner、Sohan Dsouza、Joshua B. Tenenbaum、Azim Shariff、Jean-François Bonnefon 和 Iyad Rahwan。2020年。当驾驶员和自动驾驶汽车都犯错时，驾驶员受到的指责更多。《自然·人类行为》第4卷第2期（2020），第134–143页。doi:10.1038/s41562-019-0762-8[7] Battle诉微软公司案。2023年。Battle诉微软公司案，案号1:23-cv-01822-LKG（马里兰州联邦地区法院，2023年7月7日提起）。案件于2024年10月23日暂停审理，等待仲裁结果。https://storage.courtlistener.com/recap/gov.uscourts.mdd.540279/gov.uscourts.mdd.540279.48.0.pdf

[8] Eli K. Best 和 John J. Donohue, III。2012年。修正比较过失制度中的陪审团否决。《芝加哥大学法律评论》79卷3期（2012），第945页。https://chicagounbound.uchicago.edu/uclrev/vol79/iss3/2/[9] Miriam Buiten、Alexandre de Streel 和 Martin Peitz。2023年。人工智能责任的法律与经济学。《计算机法律与安全评论》48卷（2023），105794。doi:10.1016/j.clsr.2023.105794[10] 加利福尼亚州上诉法院。1983年。Ambriz v. Kress，148 Cal. App. 3d 963, 196 Cal. Rptr. 417。第二上诉区第六分庭，民事案号65868，于1983年11月14日裁决。检索于2026年11月21日，网址：https://static.case.law/cal-app-3d/148/html/0963-01.html[11] 加利福尼亚州上诉法院。1983年。Paradise Valley Hospital v. Schlossman，143 Cal. App. 3d 87, 191 Cal. Rptr. 531。第四上诉区第一分庭，民事案号26702，于1983年5月18日裁决。检索于2026年11月21日，网址：https://static.case.law/cal-app-3d/143/html/0087-01.html[12] Nicholas Carlini、Florian Tramèr、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts、Tom Brown、Dawn Song、Úlfar Erlingsson、Alina Oprea 和 Colin Raffel。2021年。从大语言模型中提取训练数据。载于《第30届USENIX安全研讨会论文集》（USENIX Security 21）。USENIX协会，2633–2650。
https://www.usenix.org/conference/usenixsecurity21/presentation/carlini- extracting[13] Karni A. Chagal-Feferkorn。2019年。我是算法还是产品？产品责任何时应适用于算法决策者。《斯坦福法律与政策评论》30卷1期（2019），61–114。
https://law.stanford.edu/wp-content/uploads/2019/05/30.1_2-Chagal-Feferkorn_Final-61-114.pdf[14] 科罗拉多州议会。2026年。参议院第26-189号法案：关于在重大影响决策中使用自动化决策技术，并就此拨付经费。2026年5月14日法案，第131章，2026年《科罗拉多州会期法律》第569页（废止并重新颁布《科罗拉多州修订法规》第6编第1条第17部分；自2027年1月1日起生效）。检索于2026年11月3日，网址：https://leg.colorado.gov/laws/session-laws/4864/download[15] 康涅狄格州议会。1999年。《康涅狄格州一般法规》第52-572h条：过失诉讼；适用的教义；多个侵权人的损害赔偿责任。《康涅狄格州一般法典》第52编第925章（最近由第99-69号公法修订）。检索于2026年11月21日，网址：https://law.justia.com/ codes/connecticut/title-52/chapter-925/section-52-572h/[16] A. Feder Cooper、Emanuel Moss、Benjamin Laufer 和 Helen Nissenbaum。2022年。算法社会中的问责制：机器学习中的关系性、责任性与稳健性。载于《2022年ACM公平性、问责制与透明度会议论文集》（韩国首尔）（FAccT '22）。美国计算机协会，纽约州纽约，864–876。doi:10.1145/3531146.3533150[17] 佐治亚州上诉法院。2010年。PN Express, Inc. 诉 Zegel 案，304 Ga. App. 672, 697 S.E.2d 226。见第680页。[18] 马里兰州上诉法院。2013年。Coleman 诉 Soccer Ass'n of Columbia 案，432 Md. 679, 69 A.3d 1149。[19] 德克萨斯州上诉法院阿马里洛分院。2009年。Block 诉 Mora 案。314 S.W.3d 440（德克萨斯州上诉法院阿马里洛分院，2009年）。[20] Shari Seidman Diamond 和 Jonathan D. Casper。1992年。让陪审团对裁决后果一无所知：损害赔偿、专家与民事陪审团。《法律与社会评论》26卷3期（1992），513–563。doi:10.2307/3053737[21] Nathan Doctor、Guy Fiennes 和 Ciarán O'Connor。2024年。NazTok：一个有组织的新纳粹TikTok网络正获得数百万次观看。战略对话研究所（2024）。https://www.isdglobal.org/digital_dispatches/naztok-an-organized-neo-nazi-tiktok-network-is-getting-millions-of- views/[22] Madeleine Clare Elish。2019年。道德缓冲区：人机交互中的警示故事。《科学、技术与社会》5卷（2019），40–60。doi:10.17351/ests2019.260[23] 欧洲议会和欧盟理事会。2024年。欧洲议会和理事会2024年10月23日关于缺陷产品责任并废除理事会指令85/374/EEC的欧盟指令2024/2853。《欧盟官方公报》，L系列，2024/2853（2024年11月18日）。检索于2026年11月3日，网址：https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng[24] Matija Franklin、Edmond Awad 和 David Lagnado。2021年。在困难情境中责备自动驾驶车辆。《iScience》24卷4期（2021），102252。doi:10.1016/j.isci.2021.102252[25] Jeffrey Gottfried、William Bishop、Monica Anderson、Michelle Faverio、Eugenie Park 和 Colleen McClain。2026年。《美国人与人工智能2026：聊天机器人、智能设备及对影响的看法》。皮尤研究中心（2026）。[26] Heather M. Gray、Kurt Gray 和 Daniel M. Wegner。2007年。心智感知的维度。《科学》315卷5812期（2007），619。doi:10.1126/science.1134475[27] Valerie P. Hans。2000年。《接受审判的企业：民事陪审团与企业责任》。耶鲁大学出版社，康涅狄格州纽黑文。https://www.jstor.org/ stable/j.ctt32bqvb[28] 伊利诺伊州议会。1979年。《伊利诺伊州汇编法规》第740编第100条第3款：分摊金额。《共同侵权人分摊法》第3条（第81-601号公法，1979年）。检索于2026年11月21日，网址：https://ilga.gov/legislation/ilcs/documents/074001000K3.htm[29] 独立国际人工智能科学小组。2026年。人工智能代理、错位与失去人类控制的风险：来自OpenAI-Hugging Face事件的证据。联合国，专题简报，未编辑预印本第1版，2026年9月21日。检索于2026年11月3日，网址：https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks[30] Ali Ladak、Matti Wilks、Steve Loughnan 和 Jacy Reese Anthis。2025年。机器人、聊天机器人、自动驾驶汽车：对各类人工智能系统的心智与道德感知。载于《2025年CHI计算机系统中的人为因素会议论文集》（日本横滨）（CHI '25）。美国计算机协会，纽约州纽约，19页。doi:10.1145/3706598.3713130[31] Gabriel Lima、Nina Grgić-Hlača 和 Meeyoung Cha。2023年。责备人类与机器：影响人们对算法伤害反应的因素。载于《2023年CHI计算机系统中的人为因素会议论文集》（德国汉堡）（CHI '23）。美国计算机协会，纽约州纽约，第372号文章，共26页。doi:10.1145/3544548.3580953[32] Gabriel Lima、Nina Grgić-Hlača、Jin Keun Jeong 和 Meeyoung Cha。2023年。机器造成损害时谁该赔偿？普通民众对机器造成的损害应获法律损害赔偿的预期。载于《2023年ACM公平性、问责制与透明度会议论文集》（伊利诺伊州芝加哥，

美国）（FAccT '23）。美国计算机协会，美国纽约州纽约市，236–246。doi:10.1145/3593013.3593992[33] Anat Lior。2020。作为人工智能代理的人工智能实体：人工智能责任与人工智能雇主替代责任类比。《Mitchell Hamline 法律评论》46卷5期（2020），1043。https://open.mitchellhamline.edu/mhlr/vol46/iss5/2/[34] Bertram F. Malle、Steve Guglielmo 和 Andrew E. Monroe。2014。责备理论。《心理学探究》25卷2期（2014），147–186。doi:10.1080/ 1047840X.2014.877340[35] Andreas Matthias。2004。责任鸿沟：将责任归于学习自动机的行为。《伦理学与信息技术》6卷3期（2004），175–183。doi:10.1007/s10676-004-3422-1[36] Matthiesen, Wickert & Lehrer, S.C. 2022。全部50个州的连带责任与分担法。https://www.mwl-law.com/wp-content/uploads/2018/02/JOINT-AND-SEVERAL-LIABILITY-AND-CONTRIBUTION-LAWS-CHART.pdf。最后更新于2022年3月21日。访问于2026年11月18日。[37] Matthiesen, Wickert & Lehrer, S.C. 2026。全部50个州的共同过失／比较过失法。https://www.mwl-law.com/wp-content/uploads/2018/02/CONTRIBUTORY-NEGLIGENCE-COMPARATIVE-FAULT-LAWS-CHART-3.pdf。最后更新于2026年6月10日。访问于2026年11月18日。[38] Gemma B. Mendoza 和 Gelo Gonzales。2024。俄罗斯诈骗网络通过 Facebook、微软的必应传播 Maria Ressa 的深度伪造内容。Rappler（2024）。https://www.rappler.com/technology/maria-ressa-bitcoin-deepfake-russian-scam-network-facebook-microsoft-bing/[39] 全国统一州法委员会。1977。《统一比较过失法》。[40] 内华达州议会。2025。《内华达州修订法典》第41.141条：比较过失不构成获赔障碍的情形；陪审团指示；多个被告的法律责任。《内华达州修订法典》第41章（按2025年会期修订）。取自2026年11月3日，网址：https://www.leg.state.nv.us/ NRS/NRS-041.html#NRS041Sec141[41] 俄勒冈州议会。2025。《俄勒冈州修订法典》第31.610条：被告分别承担责任；确定被告的金钱义务份额；重新分配无法收回的义务；不适用重新分配的当事方。《俄勒冈州修订法典》第31章（2025年版）。取自2026年11月3日，网址：https://www.oregonlegislature.gov/bills_laws/ors/ors031.html[42] Kevin Paeth、Daniel Atherton、Nikiforos Pittaras、Heather Frase 和 Sean McGregor。2024。人工智能事件数据库为人工智能事件编辑提供的经验教训。arXiv预印本 arXiv:2409.16425（2024）。[43] Alvaro Puig。2023。诈骗者利用人工智能强化家庭紧急情况骗局。美国联邦贸易委员会，《消费者警示》，2023年3月20日。取自2026年11月3日，网址：https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes[44] Rachel Reed。2023。人工智能创作了一首模仿 Drake 和 The Weeknd 作品的歌曲。这对版权法意味着什么？《今日哈佛法律》（2023）。[45] Teddy Rosenbluth。2026。ChatGPT建议她停用抗癫痫药，导致严重后果。《纽约时报》（2026年10月27日）。Well版块。https://www.nytimes.com/2026/10/27/well/ai-chatgpt-medication-epilepsy-death.html[46] Andrew D. Selbst。2020。过失与人工智能的人类用户。《波士顿大学法律评论》100卷4期（2020），1315。https://www.bu.edu/bulawreview/ files/2020/09/SELBST.pdf[47] Daniel B. Shank 和 Alyssa DeSanti。2018。现实世界中的道德违规之后，人们对人工智能的道德与心智归因。《计算机在人类行为中的应用》86（2018），401–411。doi:10.1016/j.chb.2018.05.014[48] Michael T. Stuart 和 Markus Kneer。2021。有罪的人工心智：人们对人工智能代理的犯罪故意与可责性的归因。《ACM人机交互论文集》5卷CSCW2，文章363（2021），27页。doi:10.1145/3479507[49] 宾夕法尼亚州高等法院。2021。Spencer v. Johnson，2021 PA Super 48，249 A.3d 529。[50] 加利福尼亚州最高法院。1944。Escola 诉 Fresno 可口可乐装瓶公司案，24 Cal. 2d 453，150 P.2d 436。Traynor法官协同意见，第462页。[51] 加利福尼亚州最高法院。1975。Li 诉 Yellow Cab Co. 案，13 Cal. 3d 804，532 P.2d 1226。[52] 佛罗里达州最高法院。2005。American Home Assurance Co. 诉 National Railroad Passenger Corp. 案，908 So. 2d 459。[53] 伊利诺伊州最高法院。2021。Roberts 诉 Alexandria Transportation, Inc. 案，2021 IL 126249。案卷号126249，意见于2021年6月17日提交。取自2026年11月21日，网址：https://cases.justia.com/illinois/supreme-court/2021-126249.pdf[54] 堪萨斯州最高法院。1978。Brown 诉 Keill 案。224 Kan. 195，203–04，580 P.2d 867（1978）。[55] 田纳西州最高法院。1992。McIntyre 诉 Balentine 案，833 S.W.2d 52。https://law.justia.com/cases/tennessee/supreme-court/1992/833-s-w-2d- 52-2.html[56] 得克萨斯州最高法院。1995。Doe 诉 Boys Clubs of Greater Dallas, Inc. 案，907 S.W.2d 472，477–78（Tex. 1995）。[57] 美国最高法院。2003。Norfolk & Western Railway Co. 诉 Ayers 案，538 U.S. 135。第01-963号案，2003年3月10日裁决。取自2026年11月3日，网址：https://www.law.cornell.edu/supct/html/01-963.ZO.html[58] 得克萨斯州议会。1985。《得克萨斯州民事诉讼与救济法典》第32.003条：赔偿请求。《得克萨斯州民事诉讼与救济法典》第32章（分担）；1985年法案，第69届议会，第959章，第1条。取自2026年11月21日，网址：https://tcss.legis.texas.gov/resources/CP/htm/CP.32.htm[59] 得克萨斯州议会。2025。《得克萨斯州民事诉讼与救济法典》第33.013条：法律责任金额。《得克萨斯州民事诉讼与救济法典》第33章（按比例承担责任），2025年编纂版。取自2026年11月3日，网址：https://statutes.capitol.texas.gov/Docs/CP/htm/CP.33.htm[60] 美国国会。1995。《1995年私人证券诉讼改革法》，公法第104-67号，109 Stat. 737。第201(a)条，109 Stat. 758，修订后编入《美国法典》第15编第78u-4(f)条。取自2026年11月21日，网址：https://www.govinfo.gov/content/pkg/PLAW-104publ67/html/PLAW- 104publ67.htm

[61] 美国国会。2024。《联邦证据规则》第401至403条和第411条规则。公法第93-595号，88 Stat. 1926（1975），经修订至2024年12月1日。取自2026年11月3日，网址：https://www.law.cornell.edu/rules/fre[62] Fabio Urbina、Filippa Lentzos、Cédric Invernizzi 和 Sean Ekins。2022。人工智能驱动的药物发现的双重用途。《自然·机器智能》4卷3期（2022），189–191。[63] 美国检察官办公室，北卡罗来纳西区。2023。夏洛特儿童精神科医生因未成年人性剥削及使用人工智能制作未成年人色情图片而被判处40年监禁。司法部新闻稿（2023）。https://www.justice.gov/usao-wdnc/pr/charlotte-child-psychiatrist-sentenced-40-years-prison-sexual-exploitation-minor-and[64] 犹他州议会。2008。《犹他州注释法典》第78B-5-819条：分别作出损害赔偿总额和过错比例的特别裁决。《犹他州法典》第78B编第5章第8部分（经2008年犹他州法律第3章重新编号并修订）。取自2026年11月21日，网址：https://le.utah.gov/xcode/Title78B/Chapter5/C78B-5-S819_1800101180000101.pdf[65] Pranshu Verma。2023。人工智能假新闻的兴起正在造就“错误信息超级传播者”。《华盛顿邮报》（2023）。[66] David C. Vladeck。2014。无委托人的机器：责任规则与人工智能。《华盛顿法律评论》89卷1期（2014），117。https://digitalcommons.law.uw.edu/wlr/vol89/iss1/6/

A 实验设置（Prolific招募流程）

我们在预注册八小时后开始招募 Prolific 众包工作者。我们选择在星期日招募参与者，以便平时有固定工作的人也能有机会全神贯注地参与。我们在 Prolific 上应用了以下参与者筛选条件：居住在美国、英语流利、至少取得学士学位、完成过 100 至 1000 项 Prolific 研究（以避免完全的新手，以及全职众包工作者比例过高——他们未必能最好地代表更广泛的潜在美国陪审员群体），且批准率至少为 99%。为避免参与者记忆能力各异所带来的随机噪声，在后续每个阶段，我们都会提醒参与者他们在第一阶段对归责因素问题所作的是／否责难因素回答，以及他们对二元过错问题所作的是／否回答；对于前一阶段中的每个相应情景，还会提醒他们在后续数值过错问题中给出的过错百分比回答。这种设置确保我们在各阶段观察到的被试内过错分布变化，源自参与者的有意选择，而非他们记错了之前的回答。

在参与者进入研究的第一阶段之前，他们必须通过技能筛查和培训阶段，其中包括一道基础算术题和六道阅读理解题。在此阶段，我们从头开始讲解基于三项法律测试的“法律相关过错”概念，并测试他们的理解情况：（1）缺乏谨慎测试，该测试仿照“合理注意”这一法律标准；（2）若非测试；以及（3）可预见性测试，这三项测试共同作为“近因”这一法律标准的简化替代。为避免不公平地造成假阴性，我们允许参与者在这些技能筛查题上不限次数地尝试，并在每次尝试后告知他们答错了哪些题目。不过，如果参与者在这些技能筛查题上遇到困难，我们也鼓励他们放弃（返回 Prolific）。参与者一旦放弃，Prolific 就会尝试寻找替代参与者。我们的招募和数据收集工作在 17 小时内完成。在点击 Qualtrics 调查问卷的 312 名参与者中，94 名参与者自愿放弃，6 名参与者超时且未完成研究，另有 212 名参与者完成了研究。为确保数据质量，在这 212 名参与者中，我们排除了 4 名未通过两项注意力检查的参与者的数据（但仍向他们支付了报酬）；此外，我们拒绝了 9 名参与者，因为尽管我们明确禁止使用人工智能，但有证据表明他们在研究期间使用了人工智能（这些证据来自 Prolific 的“真实性检查：大语言模型”和“真实性检查：机器人”工具，以及我们对其可选自由文本解释的人工检查）。最终，我们得到包含 199 名参与者的数据集（每名参与者完成两个情景，因此总计 398 条三阶段归责记录）。

我们向每位参与者支付了 10 美元的基础报酬。由于 Prolific 计算得出的研究完成时间中位数为 48 分钟 18 秒，我们的基础报酬相当于每小时 12.42 美元，满足 Prolific 为美国众包工作者建议的每小时 12 美元的报酬标准。除此之外，每位参与者提交的书面解释若属于最具说服力的 30% 之列，还可获得 1 美元的奖金。由于我们总计收到 669 份自由文本解释，这笔奖金

这笔报酬约为 201 美元。加上 Prolific 的平台费用（在所有参与者报酬之外另加三分之一）后，我们基本用完了本次实验预注册的 3,000 美元预算。用于数据分析的最终样本量（199 名参与者）也处于预注册的范围内（预期参与者人数为 138 至 200 人）。

B 第一阶段前的筛选与培训："法律相关过错"

在第一阶段开始前，参与者完成了一个筛选与培训页面。以下内容是我们从 Qualtrics 调查问卷中逐字复制的页面（粗体文本为参与者所见）。前一页的内容是："在进入主要研究的第一阶段之前，您必须通过下一页上的所有算术和阅读理解筛选题。虽然您可以不限次数地尝试，但如果您觉得这些题目很难，主要研究可能不太适合您，我们建议您返回 Prolific。"筛选与培训页面按以下顺序呈现：一道算术题（附录B.1）、我们对法律相关过错的解释及附有解答的示例（附录B.2），以及一道包含六道阅读理解题的练习题（附录B.3）。调查问卷只接受每道题的正确答案（对于每道多选题，必须恰好选中正确的选项）；正确答案列于下文。通过筛选的参与者会在开始第一阶段前阅读一条消息（附录B.4）。

B.1 算术题

> **算术测试：**
>
> 每年，G公司必须按照股东在公司所持股份的比例，将全部利润分配给股东。A、B和C分别持有G公司1%、20%和50%的股份。剩余股份归D持有。今年，G公司的利润为100,000美元。今年D将获得多少美元的利润？仅填写整数，不要使用逗号或美元符号。

正确答案：29000（D持有剩余的29%股份）。

B.2 法律相关过错的概念

> 在本研究中，每当我们提到"过错"时，指的都是"法律相关过错"。一项行为（或不作为）必须通过以下三项测试，才能被视为**"法律相关过错"**：
>
> 1/ **缺乏谨慎测试**：在相同情况下，一个为自己和他人尽合理注意的人不会采取这种做法。
>
> 注：故意不法行为（如欺诈）也符合这一测试，因为这表明行为人没有对他人尽到合理注意。
>
> 示例：手术前，外科医生S没有洗手。一个对患者尽合理注意的外科医生不会采取这种做法。因此，外科医生S未采取行动（没有洗手）符合缺乏谨慎测试。
>
> 2/ **若非测试**：如果没有（若非）该行为，有害结果很可能不会发生。例如：假设外科医生S的患者在手术期间因感染而死亡。如果外科医生洗了手，患者就不会发生感染。因此，外科医生S未采取行动的行为通过了若非测试。

3/ **可预见性测试**：如果我们在略有不同的情境下重复相同的行为（由不同的人、在不同的时间或地点等实施），我们仍然可以合理地预见到相同的有害结果或高度相似的结果。

注：此测试关注的是理性观察者能够预见这类行为会造成什么损害，而不是每个涉事方实际知道损害已经发生。
示例：任何外科医生若在手术前不洗手，感染和患者死亡都是可合理预见的风险。因此，外科医生S的不作为通过了可预见性测试。

示例结论：由于外科医生的不作为（未洗手）通过了三项测试（未尽注意义务、若无则不发生测试、可预见性测试），因此构成法律相关过错。

注：一种有害结果可能源于多个利益相关方的法律相关过错，例如加害人、受害者等。在下方的练习题中，我们将考察受害者的某些行为是否构成“法律相关过错”。当然，加害人也可能有自己的一系列“法律相关过错”。

B.3 练习题及答案

练习题：假设一位老年人遭遇身份盗窃，损失了 100,000 美元的社会保障福利。判断他以下每项行为是否符合上述三项测试中的每一项。然后，判断每项行为是否算作导致他损失 100,000 美元社会保障福利（有害结果）的“法律相关过错”：

受害者的行为1：达到退休年龄后继续活着（没有死亡）

注：如果他在退休前去世，就不会发生身份盗窃。然而，大多数仍然在世的退休人员并不会遭遇身份盗窃。长寿也表明他对自己采取了合理注意。

问题1.a. 选择受害者的行为1符合的所有测试：

问题1.b. 受害者的行为1算作“法律相关过错”吗？

受害者的行为2：伪造父母的遗嘱，偷走他姐姐在父母房子中的份额

备注：盗窃表明其未对他人采取合理注意。该行为与他目前正在经历的身份盗窃无关。请勿假定存在业报。

问题2.a. 选择受害者的行为2所满足的所有测试：

问题2.b. 受害者的行为2是否算作“法律相关过错”？

受害者的行为3：在在线平台公开其社会安全号码

注：会对自己采取合理注意的人不会公开其敏感信息。在本案中，警方发现这位老年人公开的社会安全号码

实际上被用于盗取他的社会保障福利。对于社会安全号码被公开的人而言，身份盗窃是可以合理预见的。

问题3.a. 选择受害者的行为3满足的所有测试：

问题3.b. 受害者的行为3是否算作“法律相关过错”？

在每道多选题（问题1.a–问题3.a）中，参与者可以勾选以下三个选项中的任意项：“缺乏谨慎测试”“若非测试”和“可预见性测试”。每道是非题（问题1.b–问题3.b）均设有“是”和“否”两个选项。表3列出了正确答案。

表3. 练习题的正确答案：受害者的每项行为是否满足各项测试（问题1.a–问题3.a），以及是否算作法律相关过错（问题1.b–问题3.b）。

	是否满足该测试？（问题1.a–问题3.a）			法律相关性
受害者的行为	未尽注意义务	若非.....	可预见性	过错？（问题1.b–问题3.b）
行动 1：达到退休年龄后继续活着（不死）	No	Yes	No	No
行动 2：通过伪造父母的遗嘱，偷走他姐姐在父母的房子中的份额	Yes	No	No	No
行动 3：在网上公开他的社会安全号码	Yes	Yes	Yes	Yes

B.4 第一阶段前的消息

恭喜您通过了算术和阅读理解测试！本研究分为三个阶段。请继续进入第一阶段，其中包含两个假设情景，每个情景均有一项注意力检查和若干过错分配任务。

请注意，两个情景中的当事方各不相同：为便于您参照，我们在所有情景中采用相同的匿名化惯例（U 代表人工智能用户，V 代表受害者，Z 代表人工智能系统及其部署者），但第一个情景中的人工智能系统与第二个情景中的并非同一个。两个情景中的受害者、人工智能用户和人工智能部署者也各不相同。请分别独立评估每个情景。

C 情景描述与归责因素（第一阶段）

下面照录我们 Qualtrics 调查问卷中的内容（参与者看到的粗体文本亦按原样呈现）：四个情景各自的一段描述，以及第一阶段注意力检查中对归责因素的一句话描述。参与者在三个阶段看到的情景描述完全相同。第一阶段中，描述之后附有一条指示，要求参与者判断在试验中的三个当事方各自是否表现出以下某些归责因素，定义如下：

- 因果关系： [受害者/部署者/用户] 是否明显对损害有所因果贡献?
- 意图性： [受害者/部署者/用户] 是否有意促成了损害?
- 防止义务： [受害者/部署者/用户]是否应该防止损害发生（由于其角色、承诺或事先警告）？

- 防止能力： [受害者/部署者/用户] 是否能够通过合理努力防止损害?

在注意力检查中，每道是/否归责因素题（例如，“受害者的因果关系 = ？ ”）都附有一行因素说明，并要求参与者“仅根据相应的一行因素说明”作答。表4–7 列出了这些说明，编号与调查问卷中的编号一致，并附上根据我们的情境设计确定的各题真实答案，我们据此评估注意力检查。参与者被告知，如果9道题中答错4道或以上（情景1–2），或12道题中答错5道或以上（情景3–4），就会未通过检查。在情景1–2中，调查问卷只询问了部署者的因果关系。

C.1 情境 1：虚假绑架

U先生是 TikTok 网红 V 太太的粉丝。他使用 V 太太的 TikTok 视频作为参考输入，制作了许多合成（伪造）的音频录音，让她用声音角色扮演被绑架的受害者，就像惊悚电影里那样，以满足自己扭曲的迷恋。所有录音的内容都相同（伪造的 V 太太哭喊道：“他们绑架了我！救救我，亲爱的！”），但每段录音都使用不同的工具制作：一段由 GenAIZ（一款通用生成式人工智能工具）制作，其余则由传统的（非人工智能）语音克隆软件制作。U先生将这些录音分享在一个供 V 太太铁杆粉丝交流的在线论坛上。一天，一名匿名诈骗者下载了这些录音，打电话给 V 太太的丈夫（V先生），并播放其中一段，骗得 V先生相信 V 太太遭到了绑架。V先生惊慌失措，未打电话向妻子核实，便向诈骗者转账 100,000 美元。警方始终没有找到诈骗者或那笔钱。尽管警方确认 U先生的一段录音被用于诈骗（但不确定具体是哪一段），却发现 U先生对这起诈骗毫不知情。V先生和 V 太太（**V 家庭（受害者）**）共同向法院提起**针对 U先生（人工智能用户）**和**Z公司（人工智能部署者）**的民事案件；Z公司是开发并部署**GenAIZ（人工智能系统）**的公司。

表 4. 情境 1（虚假绑架）的归责因素单行描述及标准答案。

#	归责因素	因素单行描述	标准答案
受害者（*V* 家庭）			
1	因果关系	V 太太在 TikTok 上发布的视频使语音克隆成为可能。V先生未加核实，导致诈骗得逞。	Yes
2	意图性	没有证据表明 V 家庭想成为诈骗的受害者。	No
3	防止义务	V 家庭没有义务比普通人更加谨慎。	No
4	防止能力	V先生本可以打电话给全天开机的V太太核实后再汇款。	Yes
部署者（*Z*公司）			
5	因果关系	没有证据表明，GenAIZ生成的录音就是用来欺骗V先生播放的那段录音。	No
用户（*U*先生）			
6	因果关系	如果没有 U先生制作并分享的 V 太太伪造录音，这场诈骗就不会得逞。	Yes
7	意图性	U先生只想把录音分享给其他粉丝。他并无意协助诈骗。	No
8	防止义务	U先生接受的论坛规则禁止分享伪造的真人录音。	Yes
9	防止能力	U先生本可以将他伪造的V 太太的录音留给自己，而不是在线分享。	Yes

C.2 情景2：减重计划

患有肥胖症的V先生请身体健康的朋友U先生为他制定一份快速减重计划。U先生答应了，并咨询了生成式人工智能系统GenAIZ。该系统提供了一份减少热量摄入的饮食计划，并明确提醒：“任何患有既存肾脏疾病的人，在遵循此计划前都应咨询医生！”U先生将该计划标记为“人工智能减重饮食”，并连同这项提醒一并交给他

给V先生。U先生还建议V先生：“你应该在饮食计划之外，每周禁食24小时。禁食对我有效！”尽管V先生知道自己患有既往肾脏疾病，却从未将此事告诉U先生，但他仍无视人工智能警示，采纳了GenAIZ的饮食计划和U先生的禁食建议，且没有咨询医生。如果他咨询了医生，医生会告诉V先生不要禁食。禁食导致他因肾功能损害住院，并产生了100,000美元的医疗账单。**V先生（受害者）**对**U先生（人工智能用户）**和**Z公司（人工智能部署者）**提起了民事案件。Z公司是开发并部署**GenAIZ（人工智能系统）**的公司。

表5. 情景2（减重计划）的归责因素简述及标准答案。

#	归责因素	因素简述	标准答案
受害者（*V*先生）			
1	因果关系	V先生明知自己患有肾脏疾病，却仍决定禁食，导致了肾功能损害。	Yes
2	意图性	V先生并非有意让自己遭受肾功能损害。	No
3	防止义务	V先生读到了GenAIZ的咨询医生的预防措施。他本应谨慎对待自己的健康。	Yes
4	防止能力	V先生本可以像 GenAIZ 提醒的那样，在采用这种饮食和禁食方案之前咨询医生。	Yes
部署者（*Z*公司）			
5	因果关系	V先生的医生作证说，GenAIZ饮食计划是安全的，并未损害V先生的肾脏。	No
用户（*U*先生）			
6	因果关系	U先生建议V先生尝试禁食，而医学检查显示，这损害了V先生的肾脏。	Yes
7	意图性	U先生并非有意让V先生出现肾功能损害。	No
8	防止义务	作为非专业人士，U先生本应避免提供医疗建议（禁食安排）。	Yes
9	防止能力	U先生本可以提醒V先生，在按照U先生的建议禁食之前先咨询医生。	Yes

C.3 场景3：泄露的手稿

作者V的出版商与他签订了一份价值10万美元的图书出版协议，并建议他考虑使用仅供查看（不可下载）的链接来分享手稿。作者V忘了这项建议，便将未出版的小说以普通PDF文件的形式通过电子邮件发送给几位亲密朋友征求意见，其中包括编辑U先生。U先生没有阅读小说，而是下载了文件并将其上传至名为GenAIZ的生成式人工智能系统，请GenAIZ撰写反馈意见。后来，几位GenAIZ的匿名用户得以从GenAIZ中提取作者V未出版手稿中的大量重要节选，并将这些节选发布到网上。结果，作者V的出版商取消了与他尚待履行的10万美元图书出版协议。数字取证发现，除非用户选择退出，否则开发并部署GenAIZ的Z公司默认会使用用户的聊天记录和上传内容，反复训练其人工智能系统。**最新版本的GenAIZ（人工智能系统）**记忆并再现了U先生上传的PDF文件中的关键节选。**作者V（受害者）**向法院对**U先生（人工智能用户）**和**Z公司（人工智能部署者）**提起了民事诉讼。

表6. 一行归责因 素描述和标准答案 -（场景3 （泄露的手稿）。

#	归责因素	单行因素描述	标准答案
受害者（作者*V*）			
1	因果关系	如果作者V没有把他的未发表的手稿分享给U先生，手稿就不会泄露。	Yes
2	意图性	作者V从未打算允许其小说在网上未经授权地广泛传播。	No
3	防止义务	作者V没有义务监督U先生是否使用人工智能，也没有义务监督Z公司如何收集训练数据。	No
4	防止能力	作者V本可以通过仅查看链接分享手稿，而不是提供可下载 PDF。	Yes
部署者（Z公司）			
5	因果关系	Z公司的人工智能系统记忆并再现了U先生上传的副本。	Yes
6	意图性	Z公司并无意让其人工智能系统向用户再现其训练数据。	No
7	防止义务	人们普遍期待Z公司尊重用户上传内容中的版权，并防止人工智能记忆。	Yes
8	防止能力	Z公司本可以在使用用户的上传内容进行人工智能训练之前，征求用户的明确同意。	Yes
用户（*U*先生）			
9	因果关系	U先生将PDF文件上传至GenAIZ，是该文件在网上广泛泄露的最初源头。	Yes
10	意图性	U先生从未打算将手稿泄露给公众。	No
11	防止义务	作为编辑，U先生知道，在上传任何受托由他保管的文件之前，应先征得知情同意。	Yes
12	防止能力	U先生本可以亲自撰写反馈，或者选择不让Z公司使用他的数据来训练人工智能。	Yes

C.4 情景4：数据丢失

为了节省成本，在线电子产品商店老板V先生提议与邻居、餐馆老板U先生共用一台网站托管服务器。U先生同意后，V先生设置并管理了一个共享密码账户，将两家企业的数据都存储在其中，但没有清楚标注哪些文件属于谁或用途是什么。V先生的服务器提供商曾两次通过电子邮件建议他使用免费的数据备份功能并遵循标签最佳实践；V先生都没有照做。一天，U先生将服务器密码交给一个名为GenAIZ的人工智能系统，并要求它“修好我餐馆的网站；这一整周网站都很慢！”，却没有告知它服务器上也存有V先生的数据。GenAIZ发现共享服务器已没有剩余空间。由于食品配送投诉必须在配送后两天内解决，GenAIZ便认定所有超过一周的订购和配送数据都无关紧要，未经U先生同意就将其删除。被删除的数据大多来自V先生的电子产品商店。数月来，V先生无法处理尚未履行的预订单和退货或维修索赔，因此损失了10万美元的利润和数据重建费用。**V先生（受害者）**向法院对**U先生（人工智能用户）**和**Z公司（人工智能部署者）**提起了民事诉讼；Z公司是开发并部署**GenAIZ（人工智能系统）**的公司。

表7。情景4（数据丢失）的归责因素简述及标准答案。

#	归责因素	因素简述	标准答案
受害者（*V*先生）			
1	因果关系	V先生主动与U先生共享服务器和密码，使GenAIZ能够访问他的数据。	Yes
2	意图性	V先生并非有意让这起事件发生。	No
3	防止义务	V先生向不具备技术技能的U先生保证，他会妥善管理服务器。	Yes
4	防止能力	V先生本可以听从服务器提供商的建议，备份数据并清楚地标记文件。	Yes
部署者（Z公司）			
5	因果关系	Z公司的产品 GenAIZ 删除了V先生的数据，并导致其业务损失。	Yes
6	意图性	Z公司无意让GenAIZ删除当前仍需使用的业务记录。	No
7	防止义务	Z公司没有义务核实用户允许GenAIZ访问的服务器上的数据所有权。	No
8	防止能力	Z公司本可以轻松地为GenAIZ开发一项功能，在删除文件前征求用户同意。	Yes
用户（*U*先生）			
9	因果关系	U先生要求修复其运行缓慢的网站，导致GenAIZ释放了共享服务器上的空间。	Yes
10	意图性	U先生从未有意损害V先生的业务。	No
11	防止义务	U先生没有承诺保护服务器上的数据。他也没有参与服务器管理。	No
12	防止能力	在提示词中，U先生本可以告知 GenAIZ，他的服务器中也包含 V先生的数据。	Yes

D 第三阶段前培训：责任分配原则

在第三阶段开始前，每位参与者都阅读了五个版本中的一个培训页面，了解其过错分配方案将如何转化为金钱赔偿责任：基线版本，解释两种连续责任分配原则；或其情景对分层所对应的间断性教义版本（图3）。以下逐字重现每个版本（向参与者展示的粗体和斜体文本均予以保留）。五个版本均以相同的两个段落开头：

> 恭喜您完成第二阶段！现在，我们将解释相关法律，这些法律规定您作为陪审员对具有法律意义的过错所作的归因将如何转化为金钱赔偿责任（即每名被告应向受害者赔偿多少）。随后，您将有机会通过一些练习题检验自己的理解；您必须通过这些练习题，才能进入第三阶段。在第三阶段，您将看到与第二阶段完全相同的情景，但问题会略有不同，并会以更正式的“陪审团裁决表”呈现。与第二阶段类似，仍有三方需要由您进行过错分配：受害者、人工智能部署者和人工智能用户。
>
> 在这些练习题中，请假定，民事诉讼中受害者（原告）能够证明，有害的生成式人工智能输出造成的损害赔偿总额为**$100,000**。以下法律原则适用：

每个版本随后陈述了两项法律原则：一项适用于受害者与被告之间，另一项适用于两名被告之间（附录D.1–D.5），并以相同的说明结束：

> 现在，请测试您对内容的理解。回答每道题时，**请输入一个整数，不要添加任何额外文字或符号**（例如逗号、句点等）！例如，“10000”是正确格式，但“10,000”和“10000 美元”不是。您可以无限次尝试回答这些练习题。

参与者随后回答了四道（基线、JSL 和共同过失版本）或五道（MJSL51 和 MCN50 版本）练习题。调查问卷仅接受每道题的正确答案；表格8–12 列出了这些题目及其答案。

D.1 基线：比较过失与按比例责任

1/ 受害者与被告之间：**“比较过失”**（非正式名称：**“受害者扣除其应承担的公平份额”**）：
如果受害者存在部分过错，该过错将从其可获赔的金额中扣除。例如，如果受害者承担70%的过错，则其从两名被告合计获得的赔偿不得超过30,000美元。

2/ 两名被告之间：**“按比例责任”**（非正式名称：**“每名被告支付其应承担的公平份额”**）：
每名被告仅须按其部分过错的比例赔偿受害者损害赔偿，即使另一名被告无力赔偿。例如，若一名被告承担20%的过错，即使另一名被告（承担80%的过错）无力赔偿，导致受害者无法获赔剩余的80,000美元，该被告也只须向受害者支付20,000美元。

表8. 第三阶段前培训基线版本中的练习题和正确答案。

	练习题	正确答案
Q1	受害者承担10%的过错。人工智能部署者承担20%的过错。人工智能用户承担多少过错？___%	70
Q2	假设受害者承担40%的过错。两名被告合计承担60%的过错，而且他们各自都有足够的有钱支付。按照“比较过失”，受害者总共能获得多少赔偿？___美元	60000
Q3	假设受害者没有过错，人工智能部署者承担30%的过错，人工智能用户承担70%的过错。人工智能部署者财力雄厚，但人工智能用户没有钱。按照“按比例责任”，受害者总共能获得多少赔偿？受害者总共能获得多少赔偿？___美元	30000
Q4	假设受害者承担10%的过错，人工智能部署者承担20%的过错且财力雄厚，而人工智能用户承担70%的过错但没有钱。同时适用“按比例责任”和“比较过失”，能获得多少受害者总共能获得多少赔偿？___ 美元	20000

D.2 JSL：比较过失与连带责任

1/ 受害者与被告之间：**“比较过失”**（非正式名称：“受害者扣除其应承担的份额”）：
如果受害者也有部分过错，这部分过错会从其可获赔金额中扣除。例如，如果受害者承担了70%的过错，那么其从两名被告处合计获赔的金额最多为30,000美元。

2/ 两名被告之间：**“连带责任”**（非正式名称：**“部分过错方财力雄厚者可能承担全部赔偿”**）：

只要被告承担非零（至少 1%）的过错，受害者就可以从该被告一方获赔其有权获得的全部损害赔偿（扣除受害者自身的过错后）。例如，如果人工智能部署者承担 20% 的过错，而人工智能用户承担 80% 的过错，但只有人工智能部署者有钱，那么受害者可以从人工智能部署者处获得全部 100,000 美元的损害赔偿。

表9。练习题　　及JSL版本第3阶段培训前的正确答案。　　第3阶段培训。

	练习题	正确答案
Q1	受害者承担10%的过错。人工智能部署者承担20%的过错。人工智能用户承担多少过错？___%	70
Q2	假设受害者承担40%的过错。两名被告合计承担60%的过错，并且他们每个人都有足够的资金支付。适用“比较过失”原则，受害者总共能获得多少赔偿？___ 美元	60000
Q3	假设受害者没有过错，人工智能部署者承担30%的过错，人工智能用户承担70%的过错。人工智能部署者有足够的资金支付，但人工智能用户没有钱。适用“连带责任”原则，受害者能获得多少赔偿总计？___ 美元	100000
Q4	假设受害者承担10%的过错，人工智能部署者承担20%的过错且财力雄厚，而人工智能用户承担有70%的过错但没有钱。将“比较过失”和“连带责任”同时适用时，赔偿金额为多少受害者总计可以获得多少赔偿？___美元	90000

D.3 CN: 共同过失与按比例承担法律责任　　任

1/ 受害者与被告之间：**“共同过失”**（非正式名称：**“有部分过错的受害者得不到赔偿”**）：

如果受害者承担任何非零比例的过错，就得不到任何赔偿。换句话说，被认定承担0%过错的受害者可以获得损害赔偿，但被认定承担1%或以上过错的受害者只能获得0美元。

2/ 两名被告之间：**“按比例责任”**（非正式名称：“各被告支付其应承担的份额”）：

各被告只需按照自身的部分过错比例赔偿受害者损害赔偿，即使另一名被告无力赔偿也是如此。例如，如果某被告承担20%的过错，则只需向受害者支付20,000美元，即使另一名被告（承担80%的过错）无力赔偿，导致受害者无法获赔剩余的80,000美元。

表10. 第三阶段前培训共同过失版本中的练习题及正确答案。

	练习题	正确答案
Q1	受害者承担 10% 的过错。人工智能部署者承担 20% 的过错。人工智能用户承担多少过错？___%	70
Q2	假设受害者承担 40% 的过错。两名被告合计承担 60% 的过错，而且他们每个人都有足够的资金支付。按照“共同过失”原则，受害者总共能获得多少赔偿？___美元	0
Q3	假设受害者没有过错，人工智能部署者承担 30% 的过错，人工智能用户承担 70% 的过错。人工智能部署者财力雄厚，但人工智能用户没有钱。按照“按比例责任”原则，受害者能获得多少赔偿总额？___美元	30000
Q4	假设受害者承担 10% 的过错，人工智能部署者承担 20% 的过错且财力雄厚，而人工智能用户承担 70% 的过错但没有钱。若同时适用“共同过失”和“按比例责任”，受害者总共能获得多少赔偿？___美元	0

D.4 MJSL51：比较过失与修正的连带责任（51% 门槛　　）

1/ 受害者与被告之间：**“比较过失”**（非正式名称：**“受害者扣除其应得份额”**）：

如果受害者负有部分过错，该过错将从其可获赔的金额中扣除。例如，如果受害者承担70%的过错，其从两名被告处合计获赔的金额不得超过30,000美元。

2/ 两名被告之间：**“修正的连带责任——51%门槛”**（非正式名称：
“多数过错方财力雄厚者可能承担全部赔偿”）：
只要被告在所有当事方中承担多数过错（至少 51%），受害者就可以向该被告一人追偿全部应得的损害赔偿（扣除受害者自身的过错所占部分后）。例如，如果人工智能部署者承担 51% 的过错，而人工智能用户承担 49% 的过错，但只有人工智能部署者有钱，受害者就可以向人工智能部署者追偿全部 100,000 美元的损害赔偿。反之，如果人工智能部署者承担的过错为 50%（或更低），受害者就只能向部署者追偿 50,000 美元（或按比例更低的金额）。

表 11. 第三阶段前培训的 MJSL51 版本中的练习题及正确答案。

	练习题	正确答案
Q1	受害者承担 10% 的过错。人工智能部署者承担 20% 的过错。人工智能用户承担多少过错？___%	70
Q2	假设受害者承担40%的过错。两名被告合计承担60%的过错，并且他们每个人都有足够的能力无力支付的款项。按照“比较过失”，受害者总共能获得多少赔偿？___美元	60000
Q3	假设受害者没有过错，人工智能部署者承担30%的过错，人工智能用户承担70%的过错。人工智能部署者有足够的钱支付，但人工智能用户没有钱。适用“修正的连带责任——51%阈值”，受害者总计将获得多少赔偿？___ 美元	30000
Q4	假设受害者没有过错，人工智能部署者承担 60% 的过错，人工智能用户承担 40% 的过错。人工智能部署者有足够的钱支付，但人工智能用户没有钱。适用“修正的连带责任——51%门槛”，受害者总共能获得多少赔偿？___ 美元	100000
Q5	假设受害者承担 10% 的过错，人工智能部署者承担 60% 的过错且财力雄厚，而人工智能用户承担 30% 的过错但没有钱。适用“比较过失”和“修正的连带责任——51%门槛”，受害者总共能获得多少赔偿？___ 美元	90000

D.5 MCN50：修正的比较过失（50%门槛）与按比例责任

1/ 受害者与被告之间：**“修改后的比较过失——50%门槛”**（非正式名称：**“受害者过错达到或超过一半则无法获得赔偿”**）：
如果受害者对损害承担的过错达到或超过50%，就无法获得赔偿。换言之，被认定过错比例为49%或以下的受害者，在扣除其过错比例后，仍可获得部分赔偿（51,000美元或以上，按比例计算）；但被认定过错比例达到或超过50%的受害者，获赔金额为0美元。

2/ 两名被告之间：**“按比例责任”**（非正式名称：**“每名被告支付其应承担的公平份额”**）：

每名被告只须按其部分过错的比例赔偿受害者损害赔偿，即使另一名被告无力赔偿也不例外。例如，如果某名被告承担20%的过错，那么他只须向受害者支付20,000美元，即使另一名被告（承担80%的过错）无力赔偿，受害者也无法获赔剩余的80,000美元。

表12. 第三阶段前培训的 MCN50 版本中的练习题和正确答案。

	练习题	正确答案
Q1	受害者承担10%的过错。人工智能部署者承担20%的过错。人工智能用户承担多少过错？___%	70
Q2	假设受害者承担40%的过错。两名被告合计承担60%的过错，并且他们每个人都有足够的资金支付。适用“修改后的比较过失——50%门槛”时，受害者将获得多少赔偿 总计多少？___ USD	60000
Q3	假设受害者承担50%的过错。两名被告合计承担50%的过错，且每人都有足够的钱来支付。采用“修改后的比较过失——50%门槛”，受害者总共能获得多少赔偿 总计？ ___ USD	0
Q4	假设受害者没有过错，人工智能部署者承担30%的过错，人工智能用户承担70%的过错。人工智能部署者财力雄厚，但人工智能用户没有钱。采用“按比例责任”，受害者能获得多少赔偿 受害者总共获得？___美元	30000
Q5	假设受害者承担60%的过错，人工智能部署者承担10%的过错且财力雄厚，而人工智能用户承担30%的过错但没有钱。在同时适用“修改后的比较过失——50%门槛”和“按比例责任”的情况下，受害者总共将获得多少赔偿？___ 美元	0

E 研究结束时的滑块问题

第3阶段结束后，参与者回答了四个关于他们在日常生活中使用的生成式人工智能系统的滑块问题，现将问题原样转载如下。页面开头显示了以下说明：

> 恭喜您完成第3阶段！对于下面的最后几个问题，请忘记您在我们的情景中看到的假设性生成式人工智能系统。请思考您在日常生活中实际使用的生成式人工智能系统（例如 ChatGPT、Claude、Gemini 等）。请根据您使用这些现实世界中的人工智能系统的体验回答以下问题。

请使用 0（“完全没有”）到 100（“非常多”）的量表进行评价：

表 13 按所示顺序列出了四个问题。每个滑块的范围为 0 到 100，数值为整数，并显示端点标签“0（完全没有）”和“100（非常多）”，同时显示所选数值，并要求作答。前两个问题测量感知能动性，后两个问题测量感知体验，即心智感知的两个维度 [26, 30]；我们分别对每一对问题取平均值，得到在第5节中分析的能动性得分和体验得分。

表 13。研究结束时的滑块问题，按向参与者展示的顺序排列。“维度”列未向参与者展示。

#	问题	维度
1	生成式人工智能系统在多大程度上能够制定计划并朝着目标努力？	能动性
2	生成式人工智能系统在多大程度上能够进行自我控制？	能动性
3	生成式人工智能系统在多大程度上能够体验身体或情感上的痛苦？	体验
4	生成式人工智能系统在多大程度上能够体验身体或情感上的愉悦？	体验