与人工智能队友共同进行复杂选择时启发式与偏差的作用

Nikolos Gurney[1], **John H. Miller** [2,3], **David V. Pynadath**, [1]

[1] 创意技术研究所，南加州大学 [2] 卡内基梅隆大学[3] 圣菲研究所

gurney@ict.usc.edu,

pynadath@ict.usc.edu,JM7T@andrew.cmu.edu

arXiv:2301.05969v1 [cs.AI] 14 Jan 2023

摘要

行为科学家长期以来记录了人们对算法决策辅助工具的排斥，从简单的线性模型到人工智能皆是如此。然而，人们的态度正在改变，这可能也推动了人工智能助手的使用。可以说，当人们必须做出复杂选择时，人工智能辅助最有价值。我们认为，经典的实验方法不足以研究人们在人工智能助手帮助下做出的复杂选择，因为这些方法原本用于研究启发式与偏差。我们调整了一种旨在研究此类情境下复杂选择的实验范式。我们发现，框架效应和锚定效应会影响人们与人工智能助手的协作方式，并能预测选择结果。证据表明，一些参与者，尤其是处于损失框架中的参与者，对人工智能助手过于信任，因此经历了更差的选择结果。该范式还能够生成适用于计算建模的数据，为未来的人机决策研究提供支持。

引言

被动式算法决策优于人类决策，这在行为科学领域已成为一种教条（Dawes 和 Corrigan，1974年；Dawes、Faust 和 Meehl，1989）。这并不令人意外，因为即使是调校不佳的模型，表现也优于人类（Dawes，1979年）。然而，研究人员记录到，人们强烈排斥依赖算法（Dietvorst、Sim-mons 和 Massey，2015年；Burton、Stein 和 Jensen，2020年）。不过，越来越多的证据表明，人们更愿意接受算法决策辅助工具提供的意见（Logg、Min-son 和 Moore，2019年）。这种转变使人们的态度变得更加信任算法，也符合直觉：算法决策辅助工具日益普及，再加上反复接触某种刺激会改变人们对它的反应方式这一发现（Zajonc，1968年；Bornstein 和 D'agostino，1992年），共同构成了一种情境；在这种情境下，如果人们没有表现出更多信任，反而会令人惊讶。

决策辅助工具越来越多地依赖自适应算法，而非被动算法。这些人工智能（AI）助手经过优化，能够在被动算法（如线性模型）表现不如人类决策的情境中发挥作用。尽管这些技术在运作方式和行为上与更被动的前代技术截然不同，研究人员在很大程度上仍通过以下方式研究它们：依赖传统的实证方法：要么采用简单的选择实验（例如 Logg、Minson 和 Moore，2019年；作者利用其证据说明有必要开展进一步研究，尤其是在使用人工智能的情境中），要么采用实地 研究（相关领域已形成完整的分支学科，例如人机交互）。我们认为需要一种折中方案：实验设置应与人工智能的技术能力相匹配，同时避免实地实验所面临的情境挑战。

我们展示了一项这样的实验结果：参与者与人工智能助手合作，在复杂的选择选项中进行搜索。我们发现，经典的行为偏差，尤其是锚定与框架效应，仍然存在于选择行为中。这些偏差能够预测选择结果，并凸显了开发人工智能助手的机会——通过考虑这些偏差来改善结果。更重要的是，我们设计了实验环境，使其能够生成 AI 友好型数据。这一特性为未来利用实验范式中的数据训练人工智能助手创造了机会；在该范式中，人类认知中已知的偏差现象依然存在。

版权所有 © 2023年，人工智能促进协会 (www.aaai.org)。保留所有权利。

相关研究

复杂选择

行为科学中一个新兴的研究领域，正致力于理解人类在复杂选择过程中的判断与决策细微之处。复杂选择是指各种待考虑变量之间存在非线性关系的选择。这意味着，即使选择只涉及少数变量，它们之间的相互作用也可能使优化变得棘手（Gurney、Miller 和 Pynadath（2023））。复杂选择的明显例子包括购房、职业变动和择偶。在这些例子中，改变单个选择变量，都可能极大地影响其他变量对选择的作用。以购房为例，远程办公的机会会改变通勤距离、房屋面积以及许多其他变量之间的相互作用。并非只有重大选择才会受到复杂性的影响。在预算约束下装满购物车——这是大多数成年人经常做的事——也可能很复杂。例如，购物者必须考虑购买的各种商品如何搭配成餐食、它们何时会变质，以及将来是否符合自己的口味偏好等。

偏好等。

行为科学家并未忽视复杂选择的独特性质。例如，赫伯特·西蒙指出，由于变量之间的相互作用，许多选择根本无法处理（1956年）。尽管人们认识到了这一点，复杂选择实验却直到最近才出现。从历史上看，行为科学依赖实验设计来消除选择中的复杂性，因为这样研究人员就能轻松地分离出行为细节。这些实验通常有两种形式：描述性实验，即提供做出明智选择所需的全部信息；或者体验式实验，即参与者必须通过反复体验来了解选择（Jessup 等，2022）。

在描述性实验中，研究人员通常采用被试间设计，对单一选择特征进行操控。锚定启发式研究中的一个经典例子是：先询问人们某座城市的人口是否多于或少于某个参照值，再询问该城市的人口密度（Jacowitz 和 Kahneman，1995）。当参照值较低（较高）时，人们会报告该城市人口较为稀疏（密集）。此类实验对于剖析人类行为至关重要，但其僵化性也限制了它们。例如，已有充分证据表明，人类判断和决策高度依赖经验，包括个人对这类经验的思考（Schwarz，2004）。然而，复杂选择的本质使描述性实验几乎无法开展。此外，越来越多的研究表明，在人们积累更多经验后，此类实验所适用的情境可能会受到人类认知的其他因素（如偏好逆转）的影响（Hertwig 等，2004；Tsetsos、Chater 和 Usher，2012；Erev 等，2017；Jessup 等，2022）。

复杂选择难以处理，这促使研究者开发出借鉴复杂系统模型的体验式研究范式。两种广受欢迎的复杂系统模型是“多臂老虎机”（Berry 和 Fristedt，1985）和 NK (Levinthal，1997)范式。例如，一种实验范式要求参与者组合以 k 方式相互作用的 n形状，为外星人创作艺术作品。这种范式使研究者能够探究人们在复杂选择过程中如何搜索、探索和协作（Billinger、Stieglitz 和 Schumacher，2014；Wu 等，2018；Billinger 等，2021、2022）。尽管这些实验并未直接控制偏差，但 Billinger 等人（2014）确实证明，以先前经验作为内生变量，对过去结果的锚定能够为未来的复杂选择提供参考。相较于先前的表现，研究参与者表现不佳时会投入更多努力，表现出色时则会投入更少努力。

团队中的复杂选择

团队协作会为任何选择增添另一层复杂性。在外星艺术任务中，面对复杂选择，人类二人组在没有外部指导或资源的情况下，通过轮流行动和同时行动相结合的方式协调搜索工作（Billinger 等，2022）。团队还可以通过扩大问题空间中采样区域的范围，促进更好的复杂选择。每位团队成员都会为任务带来独特的知识、偏好和信念，而这些都可能反映在他们的行为中。举例来说，一项实验要求参与者组成团队，完成一项模拟创业任务：经营一个柠檬水摊（Sommer、Bendoly 和 Kavadias 2020年）。团队成员共同了解并整合相互作用的变量，例如不同的原料、价格水平、摊位位置等，从而让柠檬水摊盈利。当任务复杂性较低时，也就是变量间相互作用较少且随机性相对较低时，互动型团队的表现优于由个人单独完成任务的团队。然而，随着复杂性增加，这一结果发生了逆转。

行为科学家并非唯一致力于深入理解团队在复杂选择中的协作的研究者。跨计算机科学各个领域的研究者都在探索人机团队取得卓越表现的潜力。文献中最早的例子或许更适合称为合作实践——例如飞机驾驶和空中交通管制（Hoc 2000）。这些领域及相关领域的早期成功，催生了一个规模可观且不断发展的子学科，专门研究人们对自动化系统的信任动态（Lee 和 See 2004）。如今，这一研究领域涵盖了多种情境，从能够揭示人类队友信念状态的游戏（Siu 等，2021年；Chong 等，2022年），到对信任动态常见模型提出挑战的军事场景（Gurney、Pynadath 和 Wang，2022年）。关于人机团队的丰富文献，如今涵盖了各种主题，既包括人们对人工智能队友抱有何种期待（张等，2021年），也包括赋予人工智能表征人类伙伴信念状态的能力（Gurney 和 Pynadath，2022年）。

然而，人机协作研究中明显缺少对人类判断与决策中堪称最显著特征的探讨：启发式与偏差（Gigerenzer 和 Todd 1999年；Kahneman 2011）。显而易见，要让人工智能队友准确地模拟人类搭档，就需要考虑人在做选择时使用的认知工具。一些关于复杂选择的既有研究已通过内生方式检验启发式与偏差的作用，例如 Billinger 等人（2014年）证明，对先前结果的锚定会影响后续的搜索行为。在这项研究及其他相关研究的基础上，我们通过外生方式控制锚定与框架效应；这两种现象是启发式与偏差文献中研究最为广泛的现象之一。

锚定与框架效应

人类认知中的锚定，是指最终选择往往会受到起始信息的影响，即便这些信息并无参考价值（Tversky 和 Kahneman 1974年；Chapman 和 Johnson 1999年；Epley 和 Gilovich 2006）。举例来说，想象一下，你走过商店的货架，看到一件促销商品上方的标牌写着每位顾客限购的数量。锚定效应表明，无论你的意图或偏好如何，你购买该促销商品的数量都会受到标牌所示限购数量的影响（Wansink、Kent 和 Hoch 1998）。从退休储蓄（Madrian 和 Shea 2001）、法庭判决（Englich、Mussweiler 和 Strack 2006）

到择偶（Goller 等人 2018）等各种不同领域，都有类似效应的记载。

框架效应同样十分稳健。框架效应的基本观点是，对于任何前景，一个人都有某个参照值。变化是相对于该值来评估的；如果变化被视为损失而非收益，评估结果也会有所不同（Kahneman，2011）。经典的框架效应实证研究要求参与者考虑两种阻止致命疾病蔓延的干预方案。所有参与者都会看到一个基于概率的选项和一个备选方案。其中一半人看到的备选方案以收益（挽救的生命）来表述，另一半人看到的则以损失（失去的生命）来表述，但两种表述的比例相同。研究发现，当选择以损失框架呈现时，选择基于概率的选项的参与者人数会发生逆转（Tversky 和 Kahneman，1985）。

实验范式

我们将崎岖景观搜索隐喻引入一项互动式复杂选择任务中。该隐喻最初由生物学领域提出，用于描述进化选择（Wright，1932）（Gurney、King 和 Miller，2022）。这一隐喻的核心思想是，经过许多代的进化压力会引导生物体在遗传景观中搜寻最优基因型，而景观上的海拔代表基因型的适应度。这一隐喻在 NK 模型中得到推广（Kauffman 和 Levin，1987；Kauffman 和 Weinberger，1989），研究人员也将其应用于各种不同领域，其中最著名的或许是管理学文献（Levinthal，1997）。我们的任务要求研究参与者通过调节旋钮，在这样的景观中搜寻全局峰值（最优解）；每个旋钮对应景观的一个不同维度。参与者面临的复杂选择是何时停止搜寻，即针对给定景观，提交哪组旋钮设置作为最佳选择。参与者做出的选择越好（也就是提交的海拔越高），获得的奖金就越多。每位参与者都完成四项旋钮调节任务，前两项独立完成，之后两项与人工智能助手共同完成。重要的是，与以往的任务（为外星人创作艺术作品、经营柠檬水摊、务农等）不同，本任务缺乏强烈的情境线索。

作为选择集的景观

景观通过程序生成，并且每位参与者看到的景观都各不相同。它们的崎岖度各异，就像山脉一样，因此，即使是三维任务，对人来说也可能颇具挑战。要知道景观中某个位置的数值，人只能亲自前往该位置。不过，根据景观的平滑度，一个人或许可以根据先前获取的信息，对不同区域作出推断。平滑度取决于景观中相邻点之间可能出现的斜率范围。在本次实验中，参与者完成的任务包含一个（简单）或四个（崎岖）峰值，斜率设置为相邻两点的数值差始终不超过数值绝对范围的 10%；该范围为 33 个单位。任务顺序随机安排，但参与者总是分别独自完成一项简单任务和一项崎岖任务，并在人工智能助手的协助下完成这两项任务。

景观拓扑通过在平面上随机放置全局峰值来创建；在本次实验中，该平面涵盖 24×24 个单位。对于多峰景观，还会通过生成一份候选位置列表来随机放置额外峰值，确保它们具有一定的显著性，并由山谷将各个峰值分隔开。景观的大小和维度限制了峰值的数量及其海拔。放置所有峰值后，平滑算法会确保峰值与山谷之间的斜率符合所需的限制。我们的实验设计采用了 $[0, 32]$ 的海拔边界，峰值分布在 $[26, 32]$ 区间内，最高峰固定为 32。

景观经过渲染，使其边缘彼此连续，如图1右侧面板所示。这意味着将景观的东、西边缘卷接在一起时，会形成平滑过渡。南、北边缘亦然，因此每个景观都构成一个具有峰峦与山谷的环面世界。将景观移出 $[0, 32]$ 区间，有助于研究先前试次对后续行为的影响。这项激励任务要求参与者通过探索，找到给定世界中的最高峰。在图1右侧面板中，全局峰值是图像上方中央的浅黄色点，其下方有一处深蓝色山谷，周围则是其他山峰。

用户界面

参与者看不到景观。相反，他们通过“调节”屏幕上的旋钮来获取景观信息，如图1左侧面板所示。每个旋钮对应任务的一个不同维度。参与者点击“评估旋钮设置”按钮，提交调节后的设置。例如，双旋钮实验允许参与者探索一个三维世界，这也是我们在本研究中采用的设置。在这种情况下，从概念上说，一个旋钮沿着景观的东西方向移动，另一个则沿着南北方向移动。提交设置后给出的反馈，是该旋钮组合对应的海拔，也就是选择质量。重要的是，景观的各个维度（因而旋钮之间）相互依存，且其关系是非线性的。

通过点击旋钮手柄并沿下方的圆环拖动，即可进行调节。旋钮移动时，其中央的字母也会随之变化。旋钮上的每个位置都对应一个字母，这是我们为简化反馈而采用的设计。旋钮下方的编号列表提供了历史刻度盘设置的反馈。参与者可以滚动查看操作历史，了解之前检查过哪些设置，并在准备好做出选择时点击“最终确定选择”按钮。

我们将旋钮设计为具有 24 个离散设置，与用于生成景观的坐标平面相匹配。我们决定采用这种大小的平面并设置固定步长，是基于预试验结果：在现有激励结构下，参与者不愿穷尽搜索这样一个景观所对应的搜索空间（576 种独特组合）。换言之，对大多数参与者而言，其选择仍会存在不确定性。

我们定义了两种广义的刻度盘使用方式：探索和利用。探索是指查看距离任何此前观察位置超过两个单位的景观区域。

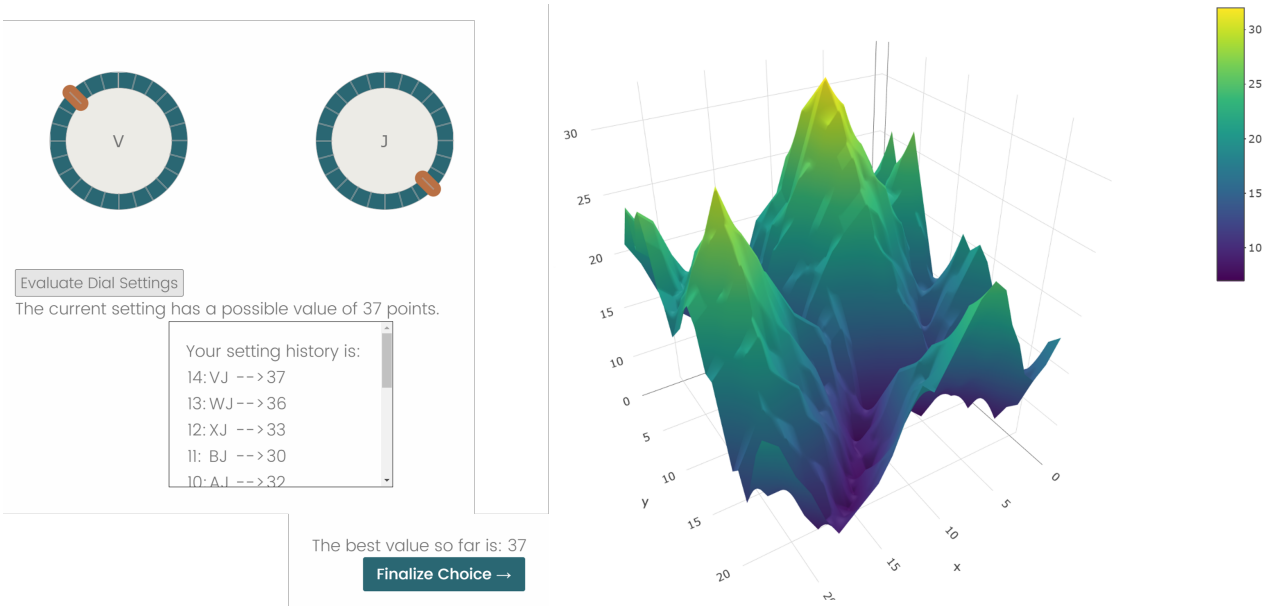


图1：左侧为实验截图，右侧为景观渲染图。所有参与者看到的刻度盘和反馈界面都相同，但接受锚定处理的参与者会在旋钮正上方看到可能达到的最佳值。在线补充材料中提供了展示四种处理条件的二乘二图像。参与者看不到景观渲染图。

例如，如果一名参与者在第一次操作时观察到刻度盘设置为 [A,A]，那么下一次操作提交 [A,D] 或 [X,C] 就构成一次探索。利用是指参与者观察到一个距离任何此前观察设置不到三个单位的新位置。

人工智能助手

完成两项任务后，人工智能助手接管了右侧旋钮。与研究参与者一样，它无法直接访问景观，因此通过抽样了解其拓扑结构。重要的是，参与者知道人工智能会提供帮助，但不知道它如何运作。

协作搜索的过程是：人类先调整自己的旋钮（也可以保持不动），然后点击评估按钮。此操作会启动人工智能助手；它会接收参与者最新的旋钮设置，并比较两种组合：一种采用其旋钮的先前设置，另一种采用新的设置。起初，人工智能更有可能接受较差的取值，也更有可能搜索较远的位置。然而，随着交互推进，它接受较差结果的可能性和搜索距离都会降低。评估并选定一种组合后，新的设置会显示在反馈窗口中。随后，界面控制权交还给研究参与者，由其决定继续搜索还是最终确定选择。

尽管这项任务对人类来说颇具挑战，而且相对耗时，但对大多数人工智能而言却相当简单。我们的研究问题关注的是人类与人工智能助手协作时的行为，而非开发更优秀的助手，因此我们选择实现一个简单的智能体。该助手使用随机模型来判断接受较差设置组合的可能性，以及在景观中进行大幅还是小幅跳跃。这是对模拟退火算法的一种适应。这个简单的人工智能让参与者之间的互动更加丰富，并消除了互动中不必要的差异，从而降低了所需的样本量。最后，由于它几乎可以瞬间完成任务，我们设置了一个短暂的反馈延迟，让人感觉助手正在“工作”。

实验操纵

2×2设计将得失框架与锚定相结合。收益框架下的参与者尝试尽可能多地获得积分，这些积分将计入他们的奖金，因此反馈是正面的。在损失框架下，参与者尝试减少损失的积分数量，因此反馈是负面的。图1左侧面板展示了收益框架。无论是收益框架还是损失框架，任务都是使数值最大化。请记住，景观会随机偏离绘制它们所依据的 $[0, 32]$ 区间。收益框架中的数值覆盖 33 个单位的区间，而损失框架中的数值则位于 $[-100, 0]$。锚定处理中的参与者会在旋钮正上方立即看到一条消息，告知他们最佳旋钮设置对应的数值，也就是他们在给定任务中所能达到的最佳结果。没有锚点的参与者则必须自行摸索每项任务中什么样的结果才算“好”。

概括本研究：每位参与者完成四项激励性复杂选择任务，其中两项单独完成，另外两项与人工智能助手合作完成；他们通过调节旋钮来寻找最优设置。我们没有告知参与者任务背后的

选择景观是如何构成的；这些景观通过程序生成，且每项任务各不相同。相反，参与者需要通过经验自行探索并了解其中的规律。参与者先完成单人任务，但无论在单人任务还是团队任务中，单峰和四峰景观的呈现顺序都是随机的。参与者也被随机分配到损失框架或收益框架处理条件中。损失处理组的调节值为负，参与者需要尽量减少损失点数。相反，收益处理组的调节值为正，参与者需要尽量获得点数。收益—损失框架处理与锚定处理交叉实施；在锚定处理中，一半参与者会获知自己在每项任务中所能找到的最佳结果。所有参与者都使用同一个人工智能助手，但该助手具有随机性，也就是说，即便两位参与者在相似情境下做出相同选择，助手也可能给出不同选择。总之，这些实验操纵有助于研究当人处于认知启发式或偏差的影响下时，人机团队如何运作。

数据与分析

我们通过 Prolific Academic 招募了 400 名参与者，邀请他们完成一项使用旋钮进行决策的研究。我们只招募批准率为 100% 且能够使用台式电脑完成研究的参与者。在数据收集前的几个月，平台上涌入了大量新用户，显著改变了平台用户群体的构成。因此，我们还将参与资格限制为在这波涌入之前就已加入平台的用户。[1] 受资金和机构审查委员会（IRB）的限制，我们只能招募英语国家的美国公民。据 Prolific Academic 称，符合条件的用户仅略多于 20,000 人，约占招募时研究池的 1 6 。研究约需 15 分钟完成，报酬为 2.00 美元，另根据投入程度发放最高 2.00 美元的奖金。最终报酬折算后的平均时薪为 18.40 美元。

在完成实验的 400 名参与者中，172 人自我认同为男性，218 人为女性，8 人为其他性别。参与者的平均年龄为 32 岁。203 名参与者表示自己拥有四年制大学学位或更高学历。我们从数据集中剔除了两条观测记录，因为相应参与者未能完成研究。一名参与者是明显的离群值：在其中一项挑战中评估了 617 种设置，超过了全部组合数。此外，这一投入程度是次为积极的参与者的三倍。由于这种行为可能会对任何分析产生过大的影响，我们决定将该参与者从数据集中剔除，最终留下 397 条观测记录。

我们在一篇配套论文中报告了参与者的个人任务努力（Gurney、Miller 和 Pynadath（2023））。我们发现，独自工作时，四峰景观比单峰景观更具挑战性；锚定和框架效应都对参与者的选择质量产生了显著的主效应；而先完成单峰任务与在四峰任务中取得更好的结果相关。

[1]https://www.prolific.co/blog/we-recently-went-viral-on-tiktok-heres-what-we-learned（我们还将这篇博客存档到了 Wayback Machine（web.archive.org）。

在其他条件相同的情况下，处于损失框架中的参与者提交了更多刻度盘设置，并且花更多时间利用而非探索。另一方面，存在明确锚点与提交更少的刻度盘设置相关，且不会显著改变参与者的搜索策略。

与人工智能助手进行复杂选择

每种景观都各不相同，而且由于景观的空间范围固定，增加山峰数量会提高平均海拔。因此，比较不同景观上的原始表现，无法充分呈现与人工智能助手合作的影响。为此，我们将每种景观的得分除以其平均海拔，得出调整后得分。平均而言，在两项选择任务中的总调整后得分，参与者与人工智能助手合作时（$M = 3.613,\ SD = 0.717$）低于独立完成时（$M = 3.718,\ SD = 0.820$）。根据配对样本 t 检验，这一总调整后得分的差异（0。105，*95%* ）具有统计学显著性（*(396)*，$p = 0.003$）。如图 2 所示，这一差异主要由损失框架条件下的参与者所致。无论是否存在锚定（t(90) $= 3.464,\ p < 0.001$），损失框架条件下参与者调整后得分的差异均具有显著性（t(101) $= 2.270,\ p = 0.025$）。

以实验处理为交互自变量，对单独努力与团队努力在总调整后得分上的差异进行预测的析因方差分析显示，框架效应显著，但锚定效应及交互效应均不显著。因此，采用双样本 t 检验即可比较框架效应对表现的影响。损失框架条件下的参与者得分显著更高（t.420） $= 2.230,\ p$ =0.026），优于其他条件下的参与者（$M_{difference}$ =0.156，95% $CI\,[0.185, 0.294]$）。

参与者在四峰景观（$M = 1.522,\ SD$ =0.259）中的表现不如单峰景观（$M = 2.091,\ SD = 0.554$）。这一差异（0.569，95% $CI\,[0.521, 0.616]$）具有显著性（t(396) $= 23.431,\ p < 0.001$）。以实验处理为交互自变量，对两种景观类型的得分差异进行预测的析因方差分析显示，锚定（F(1，393) $= 2.792,\ p = 0.100$）和框架效应（F(1，393) $= 0.128,\ p = 0.721$）均无显著主效应，交互效应也不显著（F(1，393) $= 0.342,\ p = 0.559$）。

与个人任务努力相比（$M = 2.175,\ SD = 0.631$），参与者与人工智能合作时，在单峰景观上的表现更差（$M = 2.091,\ SD = 0.554$）。这一差异为 0.084，95% $CI\,[0.025, 0.144]$，且具有统计显著性（t(396) $= 2.773,\ p = 0.006$）。这一结论不适用于四峰景观（t(396) $= 1.342,\ p = 0.181$）。以个人任务与团队任务调整后得分之间的差异为因变量，并将实验处理作为相互作用的自变量，对单峰景观进行析因方差分析，结果显示，框架处理条件具有显著的主效应，而锚定及其交互作用均无显著效应。因此，采用双样本 t 检验即可比较框架效应对单峰景观表现的影响。在损失框架条件下的参与者

条件组的得分显著高于对照组（t(393) $=2.010, p$ =0.045），而对照组得分为（$M_{difference}$ =0.122, 95% CI $[0.003, 0.242]$）。

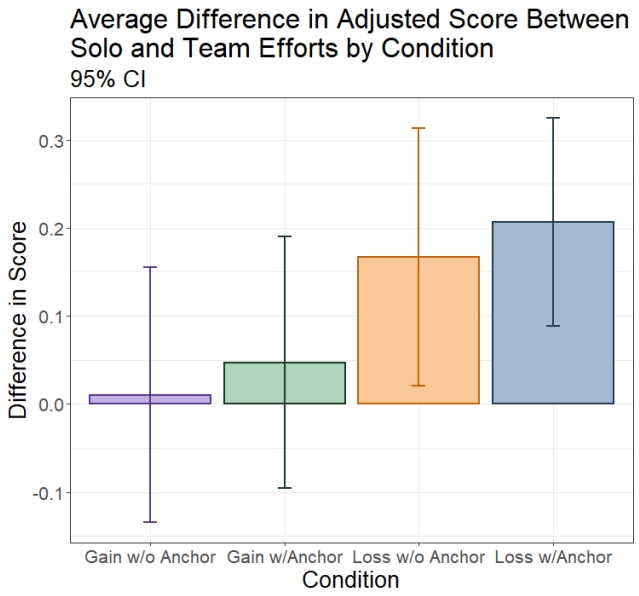


图 2：调整后得分对景观差异以及增加峰值数量所带来的平均海拔上升进行了校正。得分差异即个人任务努力期间的总得分减去与人工智能助手协作期间的总得分。

搜索时长（提交设置的数量）和探索/利用权衡（提交的设置中用于探索新区域的比例，即搜索策略）是主要的行为指标。平均而言，与人工智能助手合作时，参与者在两项任务中搜索得更少，即提交的刻度盘设置也更少（$M = 30.368$, $SD = 27.786$），而独自完成任务时则更多（$M = 43.202$, $SD = 41.521$）。根据配对样本 t 检验，搜索时长的这一差异（12.834，95% CI $[9.512, 16.155]$）具有统计显著性（t(396) =7.596, $p < 0.001$）。与人工智能助手合作时，参与者提交的设置中用于探索的比例（$M = 0.612$, $SD = 0.243$）高于独自完成任务时（$M = 0.447$, SD =0.232）。探索/利用权衡方面的这一差异（-0.165, 95% CI $[-0.186, -0.144]$）也具有统计显著性（t(396) $= -15.3456$, $p < 0.001$）。然而，与个人任务努力相比，与人工智能助手合作时，参与者用于探索的提交次数绝对值（$M = 13.584$, $SD = 6.244$）并无统计学上的显著差异（M =13.411, $SD = 8.617$; t(396) $= -0.442$, $p = 0.659$）。这表明，参与者在微调提交内容上投入的努力较少，或许是因为他们预期人工智能会替他们完成这项工作。

锚定与框架效应下的搜索时长比较参与者在不同处理条件下个人任务努力与团队任务努力的搜索时长差异，有助于了解锚定和框架效应如何影响人机团队。

采用处理条件作为交互自变量的析因方差分析，以搜索时长差异（个人任务努力时长减去团队任务努力时长）为预测变量，结果显示两个处理条件均具有显著的主效应，但交互效应不显著，因此我们从模型中移除了交互项。所得模型表明，锚定（F(1, 394) =14.971, $p < 0.001$）和框架效应（F(1, 394) $= 16.751$, p <0.001）均具有显著的主效应。事后 Tukey 检验显示，锚定（-12.611, 95% CI $[-21.047, -4.174]$, $p_{adj} < 0.001$）和框架效应（13.331, 95% CI $[4.892, 21.769]$, $p_{adj} < 0.001$）在 $p < 0.01$ 水平上均显著，尽管搜索时长的变化方向相反。

基于这些发现，采用独立样本 t 检验发现，损失框架组参与者开始与人工智能协作时，其努力程度的下调幅度显著大于收益框架组参与者（$M = -19.964$, $SD = 39.838$）（$M = -6.088$, SD =24.808; t(318.21) $= -4.319$, $p < 0.001$）。这与锚定效应相反：有明确锚点的参与者下调努力程度的幅度（$M = 6.449$, $SD = 29.426$）显著小于没有锚点的参与者（$M = 19.060$, SD =36.341; t(382.16) $= 3.804$, $p < 0.001$）。如图3所示，锚定效应似乎削弱了框架效应。

平均而言，参与者在每项后续任务中的搜索都更少。不过，唯一显著的下降发生在他们第二次个人任务努力与第一次团队任务努力之间（$M_{difference} = -5.108$; t(396) $= 5.003$, $p < 0.001$）。第一次与第二次个人任务努力之间（$M_{difference}$ =−1.690; t(396) $= 1.654$, $p = 0.099$），以及第一次与第二次团队任务努力之间（$M_{difference} = -0.927$; t(396) =1.300, $p = 0.194$）的差异均不显著。这表明，参与者预期人工智能会减少他们为出色完成任务所需投入的努力。对于收益框架组参与者，这似乎确实如此：他们与人工智能协作时付出的努力较少（$M_{difference} = -6.088$; t(203) $= -3.505$, $p < 0.001$），但平均调整后得分相同（$M_{difference} = -0.029$; t(203) =−0.574, $p = 0.567$）。然而，损失框架组参与者似乎过度调整了努力（$M_{difference} = -19.964$; t(192) = −6,962, $p < 0.001$），以至于调整后得分显著下降（$M_{difference}$ =−0.186; t(192) $= -3.877$, $p < 0.001$）。在锚定处理条件下，无论参与者是否看到锚定信息（$M_{difference} = -6.449$; t(195) $= -3.068$, $p = 0.002$）或未看到（$M_{difference} = -19.060$; t(200) $= -7.436$, $p <$ 0.001），与人工智能助手协作时，他们付出的努力都显著减少。对于非锚定组参与者，这并未与人工智能任务期间调整后得分显著降低相关（$M_{difference}$ =−0.090; t(200) $= -1.730$, $p = 0.090$）。然而，对于锚定组参与者，这与调整后得分显著降低相关（$M_{difference} = -0.121$; t(195) =−2.539, $p = 0.012$）。

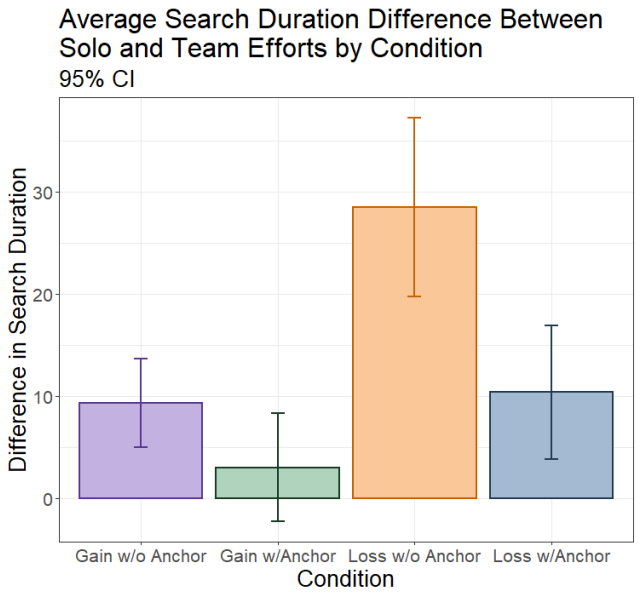


图3：搜索时长差异是参与者独自完成任务时提交的设置总数，减去其与人工智能助手合作时提交的设置总数。

锚定与框架效应下的搜索策略如前所述，搜索时长的变化也体现在参与者开始与人工智能助手协作时所采用的策略上。搜索时长缩短意味着提交内容中探索性方案所占比例更高。以实验处理条件作为相互作用的自变量，通过析因方差分析预测搜索策略差异（个人任务努力策略减去团队任务努力策略），结果显示锚定（$F(1, 393) = 1.934$, $p = 0.165$）和框架效应（$F(1, 393) = 0.684$, $p = 0.409$）均无主效应，也无交互效应（$F(1, 393) = 1.150$, $p = 0.284$）。我们再次将此结果解释为：参与者预期人工智能会减少他们调节旋钮所需付出的努力。

讨论

人们厌恶算法选择辅助工具，这一现象已有充分记录（Dietvorst、Simmons 和 Massey，2015年；Burton、Stein 和 Jensen，2020年）。然而，生活方方面面被动算法与自适应算法的日益普及，正使人们越来越愿意接受它们（Logg、Minson 和 Moore，2019年）。传统上，用于研究人机团队的实验方法依赖于降低选择的复杂性，或在自然情境中对其进行研究。尽管这两种方法各有优点，但也存在重大缺陷，例如研究结果无法推广到复杂选择，或无法适用于在不同情境下做出的选择。我们在此开发的方法既保留了选择的复杂性，又足够抽象，因而有助于推广研究发现。

研究参与者通常在与人工智能助手合作时，做出的选择不如独自工作时。这一结果主要是由损失框架下的参与者所致：他们开始与人工智能助手合作后，投入的努力大幅减少。然而，收益框架下的参与者独自完成任务与借助人工智能时，表现大致相同。参与者付出的努力减少，主要体现在他们利用局部信息微调选择，而不是探索新的选项。开始与人工智能助手合作后，努力程度没有显著下降的参与者，只有那些处于收益框架且有明确锚定的人，也就是说，他们知道可能的最佳选择的价值。这些结果表明，人们可能过度依赖人工智能助手，认为它在这项任务上的能力胜过实际水平；或者，相关地说，他们对其运作方式缺乏了解，妨碍了与之协作的能力。这些可能性指向了未来研究中一些有趣且重要的议题。此外，这些结果表明，文献中充分记载的算法厌恶现象，或许并非人类认知中稳定不变的一面。

结论

人工智能助手正逐渐进入生活的方方面面，而人们似乎比以往任何时候都更愿意让它们提供帮助。传统上，人工智能最能发挥作用的复杂选择，在实验中往往被简化为较简单的类比情境，或是在自然情境中进行研究。尽管简单选择实验可以轻松分离启发式与偏差的影响，却忽略了许多选择具有惊人的复杂性这一事实。自然情境中的实验克服了这一点，但其结果可能无法推广。正如研究所示，旋钮调节任务既保留了研究复杂选择中有充分记录的认知伪影的能力，而且我们认为，它还能提供可推广的结果（和数据），供人工智能系统使用。具体而言，本实验生成的数据很容易转换为分层图像表征，适用于深度学习模型。只要数据点足够多，这类模型就能学会区分有偏行为与无偏行为。

参考文献

Berry, D. A.; and Fristedt, B. 1985. 多臂老虎机问题：实验的序贯分配（统计学与应用概率专著）。伦敦：*ChapmanandHall*, 5(71-87): 7–7.

Billinger, S.; Benincasa, S.; Baumann, O.; Kretschmer, T.; and Schumacher, T. R. 2022年. 学习协作搜索：二元组如何克服复杂性，以及不完善的模块化分解中的不一致的激励。《产业与企业变革》.

Billinger, S.; Srikanth, K.; Stieglitz, N.; and Schumacher, T. R. 2021年. 复杂搜索任务中的探索与利用：反馈如何影响人类代理是否以及在哪里进行搜索。《战略管理期刊》, 42(2): 361– 385.

Billinger, S.; Stieglitz, N.; 和 Schumacher, T. R. 2014年。崎岖景观上的搜索：一项实验研究。《组织科学》,25(1): 93–108.

Bornstein, R. F.;和 D’agostino, P. R. 1992年。刺激识别与单纯接触效应。《人格与社会心理学杂志》, 63(4): 545.

Burton, J. W.; Stein, M.-K.; and Jensen, T. B. 2020. 增强型决策中算法厌恶的系统综述。《行为决策杂志》, 33(2): 220–239.

Chapman, G. B.; and Johnson, E. J. 1999. 锚定、激活与价值构建。组织行为与人类决策过程, 79(2): 115–153. Chong, L.; Zhang, G.; Goucher-Lambert, K.; Kotovsky, K.; and Cagan, J. 2022. 人类对人工智能及自身的信心：信心的进化及其对采纳人工智能建议的影响。计算机与人类行为, 127: 107018. Dawes, R. M. 1979. 非正规线性模型在决策中的稳健之美。《美国心理学家》, 34(7): 571. Dawes, R. M.; and Corrigan, B. 1974. 决策中的线性模型。《心理学公报》, 81(2): 95–106. Dawes, R. M.; Faust, D.; and Meehl, P. E. 1989. 临床判断与统计判断。《科学》, 243(4899): 1668–1674. Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. 算法厌恶：人们看到算法出错后会错误地避开算法。《实验心理学杂志：普通》, 144(1): 114. Englich, B.; Mussweiler, T.; and Strack, F. 2006. 用骰子决定刑事判决：无关锚点对专家司法决策的影响。《人格与社会心理学公报》, 32(2): 188–200. Epley, N.; and Gilovich, T. 2006. 锚定与调整启发式：为什么调整不足。《心理科学》, 17(4): 311–318. Erev, I.; Ert, E.; Plonsky, O.; Cohen, D.; and Cohen, O. 2017. 从异常现象到预测：迈向风险条件下、模糊情境下及基于经验的决策描述性模型。《心理学评论》, 124(4): 369. Gigerenzer, G.; and Todd, P. M. 1999. 让我们变聪明的简单启发式。牛津大学出版社, 美国。 Goller, J.; Leder, H.; Cursiter, H.; and Jenkins, R. 2018. 面部吸引力中的锚定效应。《知觉》, 47(10-11): 1043–1053. Gurney, N.; King, T.; and Miller, J. 2022. 研究复杂选择的实验方法。载于 Stephanidis, C.; Antona, M.; Ntoa, S.; and Salvendy, G. 编，*HCI International 2022* – 最新研究海报, 39–45. Springer Nature Switzerland. ISBN 978-3-031-19679-9. Gurney, N.; Miller, J.; and Pynadath, D. V. 2023. 复杂选择中的启发式与偏差。Https://doi.org/10.21203/rs.3.rs-2472194/v1. Gurney, N.; and Pynadath, D. V. 2022. 面向人类的心智理论机器人：一项综述。载于*2022*年第*31*届*IEEE*机器人与人类交互通信国际会议（*RO-MAN*）, 993–1000. IEEE.

Gurney, N.; Pynadath, D. V.; and Wang, N. 2022. 测量并预测人们对人工智能队友所提供建议的信任。载于人机交互国际会议，22–34。施普林格。Hertwig, R.; Barron, G.; Weber, E. U.; and Erev, I. 2004. 基于经验的决策与罕见事件对风险选择的影响。《心理科学》, 15(8): 534–539。Hoc, J.-M. 2000. 从人机交互到人机协作。《人体工程学》， 43(7): 833–843。Jacowitz, K. E.; and Kahneman, D. 1995. 估计任务中锚定效应的测量方法。《人格与社会心理学公告》， 21(11): 1161–1166。Jessup, R. K.; Busemeyer, J. R.; Dimperio, E.; Ho mer, J.; and Phillips, A. 2022. 选择是一件棘手的事：将复杂的选择模型与学习过程相结合，以更好地解释复杂选择行为。《决策》。Kahneman, D. 2011. 《思考，快与慢》。麦克米伦。Kauffman, S.; and Levin, S. 1987. 迈向崎岖景观上适应性游走的一般理论。《理论生物学杂志》, 128(1): 11–45。Kauffman, S. A.; and Weinberger, E. D. 1989. 崎岖适应度景观的 NK模型及其在免疫反应成熟中的应用。《理论生物学杂志》， 141(2): 211–245。Lee, J. D.; and See, K. A. 2004. 对自动化的信任：为适当依赖而设计。《人为因素》， 46(1): 50–80。Levinthal, D. A. 1997. 崎岖景观上的适应。《管理科学》， 43(7): 934–950。Logg, J. M.; Minson, J. A.; and Moore, D. A. 2019. 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。Madrian, B. C.; and Shea, D. F. 2001. 暗示的力量：401(k)参与和储蓄行为中的惯性。《经济学季刊》, 116(4): 1149–1187。Schwarz, N. 2004. 消费者判断与决策中的元认知体验。《消费者心理学杂志》， 14(4): 332–348。Simon, H. A. 1956. 理性选择与环境结构。《心理学评论》， 63(2): 129。Siu, H. C.; Pe˜na, J.; Chen, E.; Zhou, Y.; Lopez, V.; Palko, K.; Chang, K.; and Allen, R. 2021. 对 Hanabi 中学习型智能体和基于规则的智能体组成的人机团队进行评估。《神经信息处理系统进展》, 34: 16183–16195。Sommer, S. C.; Bendoly, E.; and Kavadias, S. 2020. 如何寻找最佳替代方案？通过实验研究解决复杂问题的搜索策略。《管理科学》， 66(3): 1395–1420。Tsetsos, K.; Chater, N.; and Usher, M. 2012. 显著性驱动的价值整合解释了决策偏差和偏好反转。《美国国家科学院院刊》， 109(24): 9659–9664。

Tversky, A.; and Kahneman, D. 1974年. 不确定性下的判断：启发式与偏差：判断中的偏差揭示了人们在不确定性下思考时采用的一些启发式。《科学》，185(4157): 1124–1131。Tversky, A.; and Kahneman, D. 1985. 决策的框架效应与选择心理学。载于行为决策，25–41。施普林格。Wansink, B.; Kent, R. J.; and Hoch, S. J. 1998. 购买数量决策的锚定与调整模型。《营销研究杂志》，35(1): 71–81。Wright, S. 1932. 突变、近亲繁殖、杂交和选择在进化中的作用。载于《第六届国际遗传学大会论文集》。无。Wu, C. M.; Schulz, E.; Speekenbrink, M.; Nelson, J. D.; and Meder, B. 2018. 泛化引导人类在广阔决策空间中的探索。《自然·人类行为》，2(12): 915–924。Zajonc, R. B. 1968. 单纯曝光的态度效应。《人格与社会心理学杂志》，9(2p2): 1。Zhang, R.; McNeese, N. J.; Freeman, G.; and Musick, G. 2021. “理想的人类”：人机协作中对人工智能队友的期待。《美国计算机协会人机交互汇刊》，4(CSCW3): 1–25。

致谢

本研究的部分工作由美国国防高级研究计划局（DARPA）赞助，合同编号为 W911NF2010011。本文所含信息不一定反映美国政府或美国国防高级研究计划局的立场或政策，也不应据此推断获得了官方认可。本研究也由美国陆军研究实验室（ARL）赞助，合同编号为 W911NF-14-D-0005。文中陈述和观点以及所含内容不一定反映政府的立场或政策，也不应据此推断获得了官方认可。