arXiv:2303.00866v1 [cs.HC] 1 Mar 2023

用于估计已发表研究可重复性的原型混合预测市场

Tatiana CHAKRAVORTI [a]、Robert FRALEIGH [a]、Timothy FRITTON [a]、Michael MCLAUGHLIN [a]、Vaibhav SINGH [a]、Christopher GRIFFIN [a]、Anthony KWASNICA [a]、David PENNOCK [b]、C. Lee GILES [a]、Sarah RAJTMAJER [a]、[1][a]宾夕法尼亚州立大学[b]罗格斯大学

摘要。 我们介绍了一种原型混合预测市场，并展示了其在实现有意义的人机协作方面所代表的途径。我们基于此前将人工预测市场提出为一种新型机器学习算法的研究。在人工预测市场中，经过训练的AI代理（机器人交易者）买卖未来事件的结果。分类决策可以被构建为未来事件的结果，因此，与给定分类结果相对应的资产价格可以作为系统对该决策置信度的代理指标。通过将人类参与者与机器人交易者共同置于这些市场中，我们可以汇集双方的洞见。本文详细介绍了使用原型混合型市场预测重复研究结果的试点研究。我们重点讨论其中的挑战与机遇，分享从混合市场参与者的半结构化访谈中获得的洞见，并概述持续开展及未来工作的愿景。

关键词： 混合预测市场、人机协作、可重复性。

1. 引言

一项新兴研究正在探索人工预测市场——由人工代理（机器人交易者）构成、通过监督学习来训练概率估计器的数值模拟市场 [3]。早期研究已证明，将经过训练的市场用作监督学习算法是可行的；在简单分类任务中，其表现可与标准方法相媲美 [3,4,16,20]。我们认为，人工预测市场最有前景的机会在于最终实现人机协作——建立一种市场框架，支持人类交易者与代理共同参与结果评估。在一项初步研究中 [7], ，我们阐述了这种混合预测市场的理论基础，并模拟了该环境中的简单类人行为。

我们在此介绍的混合市场旨在预测社会与行为科学领域重复研究的结果。对已发表科学发现的可重复性、可重复性和稳健性的研究，已在社会科学及其他领域引起广泛关注。多个大规模重复项目 [10,21,5,6,17,18,8]报告的成功重复率介于 36% 至 78% 之间，并引发了关于已发表研究发现可靠性的广泛讨论 [2,13,11,24]。

[1]通讯作者：smr48@psu.edu

预测重复研究结果似乎是人机协作的理想任务。以人为中心和以机器为中心的工作都展现出潜力，但两者都未能取得良好表现，例如， [9,27]。事实上，重复研究预测似乎是一个需要机器驱动方法的规模与范围的问题，但要捕捉该领域专家难以言传的智慧，仍然十分困难。

我们此前的工作 [22]开发并部署了用于重复研究预测任务的合成预测市场。本研究进一步探索混合市场情境。我们的这项进行中的工作围绕三个研究问题展开，对这些问题的回答将为大规模开发人机混合预测市场提供依据，并为创新的同行评审开辟一条新途径。

研究问题1：在人类参与合成预测市场的情况下，市场表现与纯合成情境相比有何变化?

研究问题2：在重复研究预测这一背景下，哪些特征对人类参与者最重要？参与者如何制定交易策略?

研究问题3：在大规模部署用于重复研究预测的混合预测市场之前，还需要解决哪些未决挑战?

总体而言，这些问题旨在帮助我们理解人机协同技术如何助力评估已发表科学发现的可重复性、可重复性和稳健性。接下来，我们将讨论对一个用于重复研究预测的混合市场进行测试版测试所得的见解，并根据后续访谈探讨参与者的看法。最后，我们将展望该领域未来的研究方向。

2. 数据

算法代理接受了400项重复研究的结果以及对社会与行为科学领域已发表研究发现的专家评估的训练。这些研究由开放科学中心和墨尔本大学牵头，为美国国防高级研究计划局的开放研究与证据置信度系统化（SCORE）[2] 项目开展。另有12项重复研究结果用作混合市场活动的测试数据。[3] 研究发现所选自的领域和期刊，以及重复研究和专家评估的程序，详见 [1]。

训练论文和测试论文的全文均经过特征提取流程，以获取语义、文献计量和统计信息。具体而言，针对每项待考察的研究主张，提取了41项特征，例如报告的p值、作者姓名、作者人数、发表刊物以及资金致谢。更多详情请参阅 [26]。

3. 方法论

3.1. 人工预测市场模型

鉴于开展高统计功效重复研究所需的资源，研究人员一直在寻求其他方法，以评估对已发表论断的置信度，并开始关注

[2]参见 https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence。[3]计划于2023年末统一发布SCORE项目的全部数据。我们分析所用的数据子集一经开放，即会在 https://github.com/Tatianachakravorti 处提供链接。

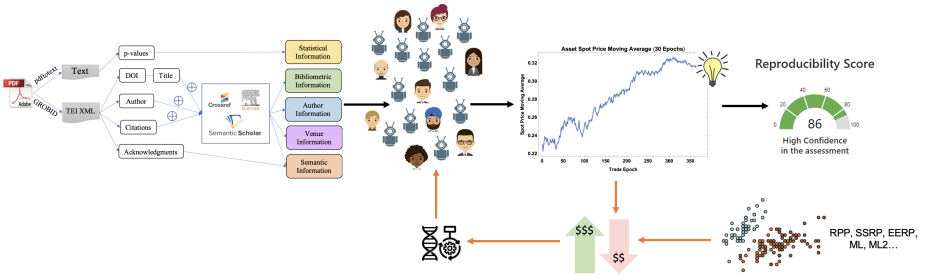

图 1. 重复研究预测市场的示意图。步骤 1：从一篇目标论文的全文中提取特征（见 [26]）。步骤 2：将提取的特征传递给算法交易者。步骤 3：算法交易者买卖代表与目标论文核心主张相关的重复研究“将会重复”或“不会重复”结果的合约。注意：在混合市场情境中，人类交易者与机器人共同参与。交易通过对数市场评分规则操纵底层资产价格（见 [20]）。（训练阶段，橙色箭头）步骤 3：市场收盘时，公布重复研究的结果。算法交易者根据所持资产的总价值获利或亏损。获利的交易者可以通过遗传算法进行复制、变异，并继续留在市场中。（测试阶段，黑色箭头）步骤 3：市场收盘时“将会重复”资产的价格，作为市场预测的替代指标。

将专家判断创造性地汇集起来，作为一种可能的途径。初步证据表明，预测市场在这一情境下前景可期。在预测若干备受瞩目的重复研究项目的结果时，简单的二元期权预测市场表现优于基于调查的方法 [9,5,6,12,14,15]。在这些市场中，可以买卖对应未来事件（即重复研究的结果）的资产，从而改变底层资产价格。这些资产价格可被解读为概率[19,25]，由此提供一种预测事件的机制。

本研究的基础模型是一个简单的人工预测市场，由算法代理（机器人交易者）组成。它们根据从已发表作品全文及相关元数据中提取的特征来决定是否购买合约，并采用遗传算法方法，利用真实重复研究结果进行训练。关于人工市场的数学建模、机器人训练流程以及从学术论文中提取特征的更多细节，详见先前研究 [20,26,22]。图 1 示意了市场和训练流程。

3.2. 混合型市场实验设计

在正在进行的研究中，我们已构建并完成测试版测试的平台，以支持机器人与真实人类参与者互动，即用于重复研究预测的混合市场情境。该平台包括一台托管预训练人工市场的 Web 服务器，以及多个 API 端点，用于：1）让人工代理购买资产；2）让人类参与者买卖资产；3）管理交易记录和实验统计数据。平台还包括一个交互式网络应用程序，方便人类参与者直观地买卖资产（见图2）。

尽管人工市场的理论基础是以连续方式构建的，但实际部署的人工市场和混合市场都是离散的，每秒迭代一次。市场交易通过排队系统管理，首先评估所有参与的代理，然后按先进先出规则处理人类参与者的交易。人工代理和人类参与者每次都只能进行单份额交易。每笔市场交易完成后，所有资产的市场价格都会根据对数市场

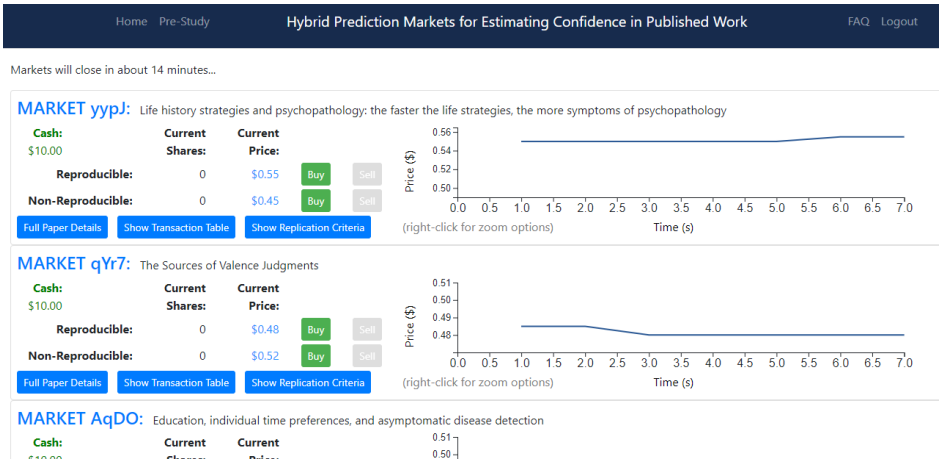

图2. 交互式网络应用程序。向人类参与者提供有关重复研究的信息以及相关论文的全文。我们为他们提供初始现金用于投资。他们可以通过应用程序买卖代表该研究“将得到重复”或“将无法重复”结果的合约。

评分规则（LMSR） [20]进行更新。在每次市场迭代期间，系统会随机抽样选取可用的人工代理参与；抽样率降至人工代理总数的 5%，以便在混合市场情境中使市场价格在 2 小时（7200 次市场迭代）内趋于收敛。

混合型市场在三个独立活动中进行了 beta 测试。在这些活动开始前，已获得机构审查委员会 (IRB) 的批准。每场活动均涉及四个市场；也就是说，每场活动中，参与者都有机会评估四项不同重复研究的结果，并买卖相关合约。我们分别从具有三种研究背景的群体中招募参与者：宾夕法尼亚州立大学斯米尔商学院的研究生（活动 1：2022年4月，9 名参与者）、宾夕法尼亚州立大学信息科学与技术学院的研究生（活动 2：2022年6月，10 名参与者），以及宾夕法尼亚州立大学心理学系的研究生（活动 3：2022年7月，14 名参与者）。活动 1 在校内经济学、管理与拍卖实验室现场举行，市场开放一小时。活动 2 和活动 3 在线举行，市场开放两小时。所有活动的参与者均获得 20 美元作为报酬，并额外获得 25 美元用于投资市场。每场活动结束时，随机选取一个市场进行支付。随后向参与者公布被选中的市场及该市场的重复研究结果。满足最低交易活动要求后，每位参与者将获得其在所选市场中的资产持有量价值，以及剩余的（未投资）现金。参与者的平均收入为 42.62 美元。

3.3. 参与者调查

混合市场参与者需要完成实验前和实验后的调查。市场前调查评估了参与者的研究背景，以及他们对可重复性相关学术研究的熟悉程度。此外，市场前调查还请参与者针对他们即将在混合市场活动中评估的每篇论文提供反馈。市场后调查请参与者描述他们的交易策略、影响其预测的研究具体特征，例如样本量、作者声誉，以及

这些特征如何影响了他们的预测。我们询问他们是否对市场结果感到意外，以及是否改变了最初的评估。[4]

3.4. 参与者访谈

表1总结了混合市场结果，并列出了每篇测试论文对应的人工市场结果。对于每个市场，报告了某项资产的最终价格（美元）；如果研究发现成功复现，该资产支付1美元，否则支付0美元，即“将会复现”资产。[5] 因此，若价格高于（低于）0.5，且论文（不）可复现，我们就将市场预测归类为“正确”。绝对误差（绝对误差）是最终价格与资产价值（0或1）之间的绝对差值。绝对误差始终较低的市场被视为更好的预测变量。我们还通过 Zoom 视频会议，对8位应邀提供更多反馈并作出回应的混合市场参与者进行了为时30分钟的半结构化一对一访谈。受访者均为活跃研究人员，拥有博士学位或正在攻读博士学位。[6] 访谈提纲包含有关他们市场经历的问题：有助于他们估计可重复性的论文和研究特征；交易过程中是否存在策略性行为；以及对平台的总体印象。我们还询问了他们对所在领域可复现性和可重复性的看法、技术如何助力可复现性，以及复现激励。参与者获得了20美元报酬。

4. 结果

4.1. 研究问题*1*：混合市场表现

表1总结了混合市场结果，以及每篇测试论文对应的人工市场结果。对于每个市场，报告了一项资产的最终价格（以美元计）：如果研究发现成功复现，该资产支付1美元；如果未成功复现，则支付0美元，即“将会复现”资产。[7] 因此，当价格大于（小于）.5 且论文（不可）复现时，我们将市场预测归类为“正确”。绝对误差（绝对误差）是最终价格与资产价值（0或1）之间的绝对差值。绝对误差始终较低的市场被认为是更好的预测变量。

混合市场预测总体上比人工市场的预测更准确（平均绝对误差为.497，对比.552）。在12个市场中的9个市场里，混合设置下的绝对误差更低。Wilcoxon符号秩检验无法拒绝混合市场与人工市场之间误差分布相同的原假设（$z = -1.373$，p –值 $= 1.83$），这可能是由于小样本统计效力较低。在所有情况下，当合成价格正确时，混合市场价格均未出现错误；有一次，混合市场将合成市场的错误预测转为了正确预测。

人工市场容易因参与不足而受到影响；如果代理没有见过足够相似的训练点（论文），就不会参与。实际上，这可能会

[4]调查工具可在 `https://github.com/Tatianachakravorti`获取。[5]根据构造方式，“不会复制”资产的市场价格始终为该资产价格的1减去该价格。[6]8名受访者中有6名参加了2022年10月开展的混合市场实验。他们并未参与表1所报告的三场市场活动。10月的活动采用与4月、6月和2022年7月市场活动相同的两小时制，并使用相同的平台。[7]根据构造方式，“不会复制”资产的市场价格始终为该资产价格的1减去该价格。

表1. 实验数据摘要。R/NR：已复现/未复现；C/NC：正确/不正确；price：最终价格；pred：预测；AE：绝对误差；– 表示无代理参与。

市场	结果	混合型 价格	混合型 pred	混合型 AE	人工 价格	人工 pred	人工 AE
事件1市场1 (E1M1)	R	0.66	C	0.34	0.41	NC	0.59
事件1市场2 (E1M2)	R	0.36	NC	0.64	0.5	–	0.5
事件1市场3 (E1M3)	R	0.64	C	0.36	0.52	C	0.48
事件1市场4 (E1M4)	NR	0.72	NC	0.72	0.5	–	0.5
事件2市场1 (E2M1)	R	0.38	NC	0.62	0.41	NC	0.59
事件2市场2 (E2M2)	R	0.58	C	0.42	0.5	–	0.5
事件2市场3 (E2M3)	R	0.8	C	0.2	0.52	C	0.48
事件2市场4 (E2M4)	NR	0.47	C	0.47	0.5	–	0.5
事件3市场1 (E3M1)	R	0.61	C	0.39	0.5	–	0.5
事件3市场2 (E3M2)	R	0.47	NC	0.53	0.46	NC	0.54
事件3市场3 (E3M3)	NR	0.76	NC	0.76	0.86	NC	0.86
事件3市场4 (E3M4)	R	0.49	NC	0.51	0.42	NC	0.58

导致一些测试点未得到评估。我们在先前的研究中观察到了这一现象 [22]，本研究的十二个市场中有五个也出现了这种情况。[8] 在混合设置中，人类参与可以支持预测。值得注意的是，我们观察到人类交易者的存在往往会促使算法代理更多地参与。在人工设置中不活跃的五个市场里，有四个市场因人类交易者的存在而引发了代理交易；只有在混合市场 E2M2 中，代理完全没有参与。

这些小样本结果表明，少数知情／专家人类交易者不会对市场表现产生明显的不利影响，甚至可能提高准确性。同样，人类交易者的交易活动有望促使算法代理之间展开交易。

4.2. 研究问题*2*：人类参与者的评估标准与交易策略

支持研究问题2的研究发现主要来自参与者访谈。我们采用由特定评估目标引导的归纳方法分析了访谈记录。为分析和编码这些记录，我们使用了 Taguette 软件 [23]。下面我们将介绍主要主题和示例引文。

4.2.1. 研究动机、设计与报告的结果

接受访谈的八名参与者都提到，方法论、样本量、p值以及研究问题的严谨性是重要的评估标准。

> 我尤其关注样本的规模、样本的多样性以及问题的复杂程度，以预测研究结果的可靠性……因此，如果样本很小、总体不够多样化，而问题又很复杂……那么研究结果可重复的可能性似乎会较低。 [参与者 2]

> 首先是先验概率。这个想法听起来可行吗？……其次是样本量。……*p*值很重要。……我在寻找预注册或开放数据。 [参与者 5]

[8] 在没有交易的情况下，两项资产的市场价格均为 .5。

4.2.2. 期刊与作者声誉

六位参与者指出，期刊声誉会影响他们的预测。然而，大多数参与者表示，他们没有考虑论文作者；有一位参与者提到，有一篇论文是个例外。

> 我想是期刊本身……我知道有些期刊往往刊登质量较高的文章。所以我相信，这在一定程度上影响了我的思路，但真正主要的因素是样本量、方法论和理论取向。 [参与者 1]。

> 是的，我不太确定期刊整体的情况，但我记得有一个案例，当时……我现在忘了论文的名字。不过，因为我知道作者是谁，而且我认为他很优秀。 [参与者 4]。

4.2.3. 交易策略

八位参与者中有五位表示，在市场开盘前，他们对每项研究发现的可重复性都有一个预估概率，并据此进行初始投资。不过，有几位提到，当市场开始朝着与他们最初预测相反的方向变化时，他们后来改变了想法。

> 我心里会有一个概率……所以，如果我认为概率应该是百分之七十五，而价格是百分之六十，那么我愿意买入，因为我觉得自己能赚十五美分。但如果交易价格是百分之四十，而我认为可重复的概率是百分之二十，我也乐意买入，因为我觉得这样每一美元或每一股能赚二十美分，之类的。 [参与者 5]。

> 在市场启动之前，我写下了自己认为合理的可重复性概率估计。之后我观察市场价格，并尝试进行交易，让市场价格更接近我认为合理的估计……但观察市场价格时，在一些市场中，我还注意到趋势朝着相反方向变化，比如从五五开变成了另一种结果。我开始重新考虑自己最初的估计是否有误。 [参与者 7]

4.2.4. 代理参与对决策的影响

所有参与者都表示，市场中有代理并未影响他们的行为，因为他们不了解代理接受的训练或其行为。

4.3. 研究问题*3*：对用于重复研究预测的人类—人工智能技术的看法

4.3.1. 对可重复性的担忧

所有接受访谈的参与者都表示，他们对可重复性问题感到担忧。不过，我们认为这可能与本研究中的自我选择偏差有关。参与者表示，他们所在领域对这些问题的认知程度各不相同。

> 我猜，可能有百分之三十的同行对此相当担忧。 [参与者 1]

> 像我这个年纪的人往往对此非常了解。比我年长一两代的教授们也有。他们至少对此有所了解。他们可能不，呃，接受这是个问题，但我认为他们通常都了解这个问题。 [参与者 5]

参与者承认，开放科学运动所倡导的实践（例如，预注册）缺乏激励措施，也缺乏发表重复研究的渠道。

> 我认为，可重复性的最大挑战在于，我们开展的重复研究仍然远远不够。我们需要重视可重复性，也需要为从事可重复性研究的学者创造一个专门的发展空间。 [参与者 5]

> 如果期刊仅依据预分析计划（其中包括对可重复性的承诺），在数据收集之前有条件地接受某个项目， [参与者 7]

4.3.2. 支持可复现性和可重复性的人机协同技术

我们询问了访谈参与者，以了解他们对技术机遇的看法，尤其是对支持和提升可复现性与可重复性的人机协同技术的看法。参与者对技术干预抱有希望，尽管总体上更倾向于混合型解决方案。八位参与者中有七位表示，对单独依靠人工智能驱动的解决方案缺乏完全信任。

> 我不确定关于哪些内容可复现、哪些内容不可复现的数据是否足够多，能否训练人工智能模型正确地完成这项工作。 [参与者 2]

> 我预计混合型模型会更可靠……向模型中添加更多预测变量往往会带来更好的性能……我预计这至少会增加一些预测能力，而且我会非常感兴趣地看到所有这些市场的结果。 [参与者 5]

> 我相信市场……许多人确信，市场能够有效汇总信息，因此结果具有意义……我可能会预期，混合市场的表现最为准确。 [参与者 7]。

4.3.3. 混合型预测市场体验

最后，我们收集了参与者在混合市场期间的体验反馈，并征求了他们对未来改进平台和/或方法论的建议。大多数参与者都很热情，并表示这次体验很有趣。几位参与者指出，用户界面的一些细节仍可改进。所有参与者都认为市场持续时间过长。我们选择两小时的时间窗口，是为了尽可能方便来自世界各地、在不同时区登录的参与者。我们将在后续步骤中重新评估这一安排。

> 对我来说，市场似乎持续了相当长一段时间……它确实只是缓慢地朝着某一个方向而不是另一个方向发展……真的非常有趣。实际上，我相当享受这次体验。 [参与者 1]

5. 结论

我们介绍了使用原型混合预测市场进行重复研究预测的试点研究。这项仍在进行中的工作证明，协作式人机技术可用于评估已发表的科学主张，且具有可行性。尽管我们在重复研究预测的背景下试点了这种方法，但我们认为，混合市场为人机协作开辟了新途径，可应用于广泛的任务；对于这些任务，单靠人类或机器驱动的方法都不足以胜任。对参与者进行的市场结束后访谈，凸显了这项工作的机遇与挑战。

参考文献

[1] N. Alipourfard, B. Arendt, D. J. Benjamin, N. Benkler, M. M. Bishop, M. Burstein, M. Bush, J. Caverlee, Y. Chen, C. Clark 等。《开放研究与证据可信度系统化（SCORE）》。技术报告，开放科学中心，2021年。[2] M. Baker：《1500名科学家揭示可重复性研究的内幕》。《自然》新闻，533(7604):452，2016年。[3] A. Barbu 和 N. Lay：《面向分类的人工预测市场简介》。《机器学习研究杂志》，13(1):2177–2204，2012年。[4] A. Barbu 和 N. Lay：《用于淋巴结检测的人工预测市场》。载于《2013年电子健康与生物工程会议》（EHB），第1–7页。IEEE，2013年。[5] C. F. Camerer, A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan 等：《评估经济学实验室实验的可重复性》。《科学》，351(6280):1433–1436，2016年。[6] C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer 等：《评估2010年至2015年间发表于《自然》和《科学》的社会科学实验的可重复性》。《自然·人类行为》，2(9):637–644，2018年。[7] T. Chakravorti, V. Singh, S. Rajtmajer, M. McLaughlin, R. Fraleigh, C. Griffin, A. Kwasnica, D. Pennock 和 C. L. Giles：《人工预测市场为人机协作带来新机遇》，2022年。[8] F. Cova, B. Strickland, A. Abatista, A. Allard, J. Andow, M. Attie, J. Beebe, R. Berni¯unas, J. Boudesseul, M. Colombo 等：《估算实验哲学的可重复性》。《哲学与心理学评论》，12(1):9–44，2021年。[9] A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek 和 M. Johannesson：《使用预测市场估算科学研究的可重复性》。《美国国家科学院院刊》，112(50):15343–15347，2015年。[10] T. M. Errington, E. Iorns, W. Gunn, F. E. Tan, J. Lomax 和 B. A. Nosek。《科学论坛：对癌症生物学研究可重复性开展开放式调查》。《eLife》，3:e04333，2014年。[11] D. Fanelli。《观点：科学真的面临可重复性危机吗？我们需要这种危机吗？》。《美国国家科学院院刊》，115(11):2628–2631，2018年。[12] E. Forsell, D. Viganola, T. Pfeiffer, J. Almenberg, B. Wilson, Y. Chen, B. A. Nosek, M. Johannesson 和 A. Dreber：《预测Many Labs 2研究的重复研究结果》。《经济心理学杂志》，75:102117，2019年。[13] D. T. Gilbert, G. King, S. Pettigrew 和 T. D. Wilson：《对“估算心理科学的可重复性”的评论》。《科学》，351(6277):1037–1037，2016年。[14] M. Gordon, D. Viganola, M. Bishop, Y. Chen, A. Dreber, B. Goldfedder, F. Holzmeister, M. Johannesson, Y. Liu, C. Twardy 等：《不同学术领域的重复率是否相同？DARPA SCORE项目的群体预测》。《英国皇家学会开放科学》，2020年。[15] M. Gordon, D. Viganola, A. Dreber, M. Johannesson 和 T. Pfeiffer：《预测可重复性——对大规模预测项目中的调查数据和预测市场数据进行分析》。《PLOS ONE》，16(4):e0248780，2021年。[16] F. Jahedpari, J. Padget, M. De Vos 和 B. Hirsch：《作为症候群监测工具的人工预测市场》。《群体智能：基础、方法与实践》，2014年。[17] R. A. Klein, K. A. Ratliff, M. Vianello, R. B. Adams Jr, ˇS. Bahn´ık, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh 等：《调查可重复性的差异》。《社会心理学》，2014年。[18] R. A. Klein, M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, ˇS. Bahn´ık 等：《Many Labs 2：调查不同样本与环境下可重复性的差异》。《心理科学方法与实践进展》，1(4):443–490，2018年。[19] C. F. Manski：《解读预测市场的预测结果》。《经济学快报》，91(3):425–429，2006年。[20] N. Nakshatri, A. Menon, C. L. Giles, S. Rajtmajer 和 C. Griffin：《使用动态凸集设计和分析合成预测市场》。《arXiv》预印本，arXiv:2101.01787，2021年。[21] 开放科学协作组织：《估算心理科学的可重复性》。《科学》，349(6251)，2015年。[22] S. Rajtmajer, C. Griffin, J. Wu, R. Fraleigh, L. Balaji, A. Squicciarini, A. Kwasnica, D. Pennock, M. McLaughlin, T. Fritton 等：《用于估算已发表研究置信度的合成预测市场

研究成果。在*AAAI* 人工智能大会论文集，第 36 卷，13220，2022 年，第 13218–132 页。[23] R. Rampin 和 V. Rampin。Taguette：开源定性数据分析工具。《开放源代码软件杂志》，6(68):3522，2021 年。[24] A. D. Redish、E. Kummerfeld、R. L. Morris 和 A. C. Love。观点：可重复性失败对科学探究至关重要。《美国国家科学院院刊》，115(20):5042–5046，2018 年。[25] J. Wolfers 和 E. Zitzewitz。将预测市场价格解读为概率。技术报告，美国国家经济研究局，2006 年。[26] J. Wu、R. Nivargi、S. S. T. Lanka、A. M. Menon、S. A. Modukuri、N. Nakshatri、X. Wei、Z. Wang、J. Caverlee、S. M. Rajtmajer 等。使用监督学习模型预测社会与行为科学论文的可重复性。*arXiv* 预印本 *arXiv:2104.04580*，2021 年。[27] Y. Yang、W. Youyou 和 B. Uzzi。利用人类与人工智能估计科学发现的深层可重复性。《美国国家科学院院刊》，117(20):10762–10768，2020 年。