第三次修改

审稿人一：

给作者的意见:

很高兴看到作者在本次修订中对我此前的意见逐条回应，并据此对论文在理论铺陈、概念界定、方法与统计报告规范、以及模型说明等方面做了实质性改进，整体质量有明显提升、行文也更为清晰流畅。尽管如此，稿件中仍有少量细节需要进一步打磨（下文将逐条说明）。若作者据此进行相应的微调与修订，我将非常乐意建议本稿予以接收发表。

我们谨向审稿专家致以最诚挚的感谢!专家在本次及此前数轮审阅中提出的深刻而中肯的意见，为本文质量的持续提升起到了至关重要的作用。对于您在本轮评审中再次指出的若干细节问题，我们已遵照您的建议，在正文及下述回复中逐一进行了细致的修订与完善。

意见 1：开篇直接说研究空缺有点突兀，可以在摘要最前面加一句铺垫的陈述，例如“现有研究多聚焦于 A 和 B，发现 C，但尚缺乏...”。

回应：感谢审稿专家的宝贵意见！我们完全赞同您的看法，在摘要开篇直接陈述研究空缺确实略显突兀。为此，我们已在摘要开头补充了如下铺垫性表述：“现有信任研究多聚焦于探究面孔可信度或社会价值取向对信任的独立影响”。这一表述与引言部分的文献综述相呼应，准确反映了该领域的研究现状。

在补充铺垫的基础上，为保持摘要的简洁性，我们对后续表述进行了以下优化：

1. 删除了“关于互惠预期的机制假设也缺少实证检验”的表述。我们认为，对“缺乏对二者在单次信任与信任学习中的共同作用的系统探讨”这一表述本身就蕴含了对其中潜在心理机制的深入探索需求，而互惠预期正是该机制的核心组成部分。删除此表述可使行文更加凝练。

2. 将“且在构建信任学习的强化学习模型时未纳入社会价值取向”修改为“且在使用计算模型方法探究信任学习的动态机制时，未能充分考虑社会价值取向的作用”。这一修改既与前句“研究现状—研究不足”的行文逻辑相呼应，也更准确地体现了研究方法特点。在信任研究领域，对特质变量在计算模型中的作用的考察，通常不是将其直接作为参数纳入模型，而是通过比较不同特质个体的模型参数差异来推断其作用机制。本研究正是通过比较不同社

会价值取向个体的模型参数差异，来探讨该特质对信任学习过程的影响。新的表述更好地契合了这一研究方法特征。同时“动态机制”的提法也与研究标题中的“机制”相呼应。

3. 将“通过两个实验回应上述研究缺口”精简为“以回应上述研究缺口”，使表述更为简洁规范。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 2：引言第一段“信任既受作为外部社会线索的面孔可信度驱动，也受作为个体自身特质的社会价值取向调节”，结论性陈述需要参考文献(可以把后一段举例的三篇文献放上来，或者不想要重复多次引用的话可以把这两段糅合在一起)。

回应：感谢审稿专家的宝贵意见！我们完全认同在提出重要结论性陈述时提供文献支撑的必要性。根据您的建议，我们已在原文引言第一段的结论性陈述后，补充引用了后一段引用的、关于面孔可信度与社会价值取向研究的代表性文献(Qi et al., 2022; Wang & Hu, 2021)，以明确该论断的实证依据。在修改过程中，我们经审慎考虑，认为将引言第一二段的内容分为两段阐述，有助于保持结构清晰，便于读者阅读。因此，我们最终选择以补充文献的方式回应您的建议，而未对段落进行合并。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 3：引言第五段“通过信任学习，个体能够不断优化其信任决策，减少单次信任决策中的偏见或误判，从而更好地维护自身利益”，同上，参考文献。

回应：感谢审稿专家的宝贵意见！我们已根据您的建议，在该结论性陈述后补充了本段落已引用的关键文献 (Chang et al., 2010)，该文献是信任学习领域的奠基性研究之一，能够有力地支持“通过信任学习，个体能够不断优化其信任决策，减少单次信任决策中的偏见或误判，从而更好地维护自身利益”这一观点。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 4：引言第七段，通常模型的构建与细节置于“方法—数据分析”更为规范。引言建议保留框架性、功能性的简述，将技术细节移至方法部分，以提升结构紧凑度。(非强制)。

回应：感谢审稿专家的宝贵意见！我们非常赞同您的看法，将模型的构建与技术细节置于“方法—数据分析”部分更符合学术规范，有助于引言保持框架性与功能性，提升文章结构的紧

凑度。

根据您的建议，我们对文中涉及计算模型构建细节与公式的主要段落(包括引言第七段、第九段及 3.3.2 节的模型构建部分)进行了系统性调整。为保持文章逻辑的连贯性，我们特别注重将技术细节迁移后，使方法部分的计算模型构建内容与引言中保留的框架性、功能性阐述之间形成清晰的对应关系。

具体而言：

（1）我们将引言第七段中有关 Rescorla-Wagner 预测误差规则的公式表述移至 3.3.2 节模型构建部分开头，使其与引言第七段保留的对 Rescorla-Wagner 预测误差规则的定性描述形成对应。

（2）将 3.3.2 节模型构建部分的原有内容移至 Rescorla-Wagner 预测误差规则阐述之后(模型构建部分第二段)，集中说明研究如何围绕三个核心维度(即以往研究发现的信任学习三个关键计算机制特征)构建 12 个候选模型，此部分与引言第八段关于信任学习关键机制特征的归纳总结相呼应。

（3）将引言第九段中整合了三个关键机制特征的最佳拟合模型的具体公式移至 3.3.2 节模型构建部分第三段，同时将该段剩余的假设提出内容与引言第八段合并，形成完整的假设阐述。引言中对最佳拟合模型的特征归纳与假设提出，与“方法”部分第三段对最佳拟合模型的具体公式阐述，形成了良好的对应关系。

通过上述调整，引言阐述计算模型的部分与 3.3.2 节模型构建部分在逻辑上形成了清晰的对应关系，既避免了因内容迁移可能引起的结构混乱，也便于读者把握全文的论证脉络。

在语言表述上，我们力求最小化改动，主要进行内容位置的调整，并适当增加连接词以保证行文流畅。所修缮的语言表述均是对原有表述的延伸与整合，并未引入实质性的新内容。例如，我们在公式阐述部分对参数进行了更详细的解释，并在模型构建第二段开头明确点出研究围绕三个关键机制特征展开建模工作，以帮助读者更清晰地理解模型比较的研究逻辑。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 5：实验方法的 2.2.1 部分“95%CI = [23.54, 28.82]”， 这里关于置信区间的表述为纯英文，实验一结果部分为中文＋英文，请全文统一。

回应：衷心感谢审稿专家如此细致的审阅。您指出的置信区间表述不一致问题非常关键。我

们已对全文进行了系统检查，并将所有置信区间的表述统一规范为“95%CI = [...]”，以确保全文术语的统一性与专业性。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 6：实验方法的 2.2.2 部分“根据计算结果，被试被划分为亲社会者(SVO°得分 ≤22.45°)或亲自我者(SVO°得分 ＞22.45°)”，关于 SVO 分组，22.45°阈值被广泛采用，但这种划分方法存在中位数分割的隐患。建议作者可以在补充材料中报告完整的 SVO 分数频率分布及密度曲线图来提高研究透明度。若所选被试的 SVO 近似双峰分布，且峰谷大致位于传统分割点 22.45°，则可侧面证明原本的社会价值取向划分的合理性，且在本实验中所选被试可以依据此法进行划分。

回应：感谢审稿专家的宝贵意见！我们完全认同尽管 22.45°的划分阈值在 SVO 研究中被广泛采用，但这种分类方法可能存在一定局限性。为增强研究的透明度，我们按照您的意见，在附录中分别报告了实验 1 与实验 2 所有被试的社会价值取向(SVO°)分数的完整频率分布及密度曲线图。

结果显示，在两个实验中，SVO°得分的分布均呈现出明显的双峰趋势，且阈值 22.45°大致位于分布曲线的谷底位置附近。这一分布特征为在本研究中使用该阈值进行社会价值取向分组提供了有力的数据支持，表明其能够有效区分亲社会与亲自我两种取向类型。

此外，我们也在正文 2.2.2(实验 1)与 3.2.2(实验 2)的“实验材料”部分对社会价值取向测量材料的相关内容进行补充，指出附录中报告了 SVO 分数的完整频率分布及密度曲线图，结果支持所用划分阈值的合理性。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 7：实验一的结果 2.3 部分“表明高可信度面孔通过提升信任者的互惠预期(a = 0.30)，从而增加信任行为(b = 0.72)”，正文使用了指代字母（a & b），在图中也需要同步标注，例如“a=0.30*”。

回应：感谢审稿专家的宝贵意见！我们已遵照您的建议，对文中涉及中介分析的报告与图示进行了全面核查与统一修订。具体修改包括以下三方面：

（1）补充路径系数标注：在中介模型图中，除按您的建议添加了路径系数 a 与 b 的标注外，

我们还同步补充了直接效应(c')与总效应(c)的对应字母标识，以完整呈现模型结构。

（2）规范效应标识符号：根据中介分析的方法学文献(王阳, 温忠麟, 2018)，我们将直接效应的符号由原先的“c”修正为更规范的“c'”，总效应标识统一为“c”，并在图中同步更新，以符合当前中介分析报告的通用标准。

（3）统一效应值报告方式：将社会价值取向的中介分析结果由原先的标准化估计值统一调整为非标准化估计值进行报告，使其与面孔可信度的结果估计值报告保持一致。采用非标准化系数有助于更直接地反映变量间关系的原始尺度，增强结果的可解释性。我们亦已同步更新图中相应数值。

为增强分析过程的透明度与可重复性，我们在下方同时呈现了社会价值取向中介分析的非标准化结果与标准化结果。如有需要，我们亦可提供完整的 Mplus 分析输出文件(包含标准化与非标准化结果)，以供进一步核查。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

CONFIDENCE INTERVALS OF MODEL RESULTS

	Lower .5%	Lower 2.5%	Lower 5%	Estimate	Upper 5%	Upper 2.5%	Upper .5%
TOUZI ON							
CSVODU	-0.282	-0.167	-0.111	0.214	0.495	0.543	0.598
YUQI	0.399	0.475	0.528	0.733	0.899	0.934	1.004
YUQI ON							
CSVODU	-0.087	0.023	0.078	0.284	0.486	0.517	0.626
Intercepts							
TOUZI	-0.117	0.331	0.501	1.338	2.355	2.580	2.969
YUQI	4.691	4.757	4.793	5.013	5.256	5.284	5.490
Residual Variances							
TOUZI	1.533	1.803	1.931	2.546	3.449	3.644	4.021
YUQI	1.261	1.416	1.518	1.995	2.735	2.859	3.141
New/Additional Parameters							
H	-0.054	0.016	0.053	0.208	0.404	0.449	0.530
TOTAL	-0.169	0.016	0.080	0.422	0.704	0.764	0.911

STDYX Standardization

	Lower .5%	Lower 2.5%	Lower 5%	Estimate	Upper 5%	Upper 2.5%	Upper .5%
TOUZI ON							
CSVODU	-0.131	-0.083	-0.052	0.110	0.262	0.285	0.333
YUQI	0.295	0.349	0.388	0.542	0.666	0.683	0.720
YUQI ON							
CSVODU	-0.084	0.012	0.043	0.197	0.317	0.342	0.376
Intercepts							
TOUZI	-0.055	0.171	0.257	0.687	1.298	1.446	1.735
YUQI	2.796	2.931	3.010	3.479	3.982	4.106	4.343
Residual Variances							
TOUZI	0.439	0.489	0.521	0.671	0.821	0.843	0.888
YUQI	0.856	0.883	0.899	0.961	0.997	0.999	1.000

参考文献：

王阳, 温忠麟. (2018). 基于两水平被试内设计的中介效应分析方法. *心理科学, 41*(5), 1233−1239.

第二次修改

审稿人一：

意见：感谢作者对每个问题的逐一回答和针对性修改，这个版本已经看到了很大的进步。建议作者进一步完善摘要，保证语言的准确性和清晰性。请在正文中以及摘要最后部分明确本研究发现的意义。另外，文中所有图中的文字均显示为乱码，请检查纠正。

回应：非常感谢审稿专家对拙文的严谨细致审阅！专家的肯定既是对我们研究工作的认可，也为我们进一步完善论文提供了重要指引，激励我们持续提升研究呈现的完整性与精准性。

针对“进一步完善摘要以保证语言准确性和清晰性”及“明确本研究发现意义”的建议，我们首先回溯了整篇文章的核心框架：研究围绕四个研究问题(对应四个研究假设)展开，而这四个问题(假设)可进一步整合为本研究相较于同类研究的三点创新之处(与论文自我检查报告中填写的创新点一致)，具体为：

(1) 首次系统考察面孔可信度与社会价值取向在单次信任及信任学习中的共同作用(对应假设 1 与假设 3)；

(2) 首次探究互惠预期是否为面孔可信度与社会价值取向影响信任行为的共同心理机制(对应假设 2)；

(3) 构建面孔可信度与社会价值取向共同驱动的信任学习强化学习模型，首次将社会价值取向这一人格特质纳入信任学习的强化学习建模框架(对应假设 4)。

在上述创新导向的研究假设下，本研究得出以下核心发现：

(1) 面孔可信度与社会价值取向仅在信任学习阶段对信任行为存在共同影响；

(2) 互惠预期确实是面孔可信度与社会价值取向影响信任行为的共同心理机制；

(3) 所构建的信任学习强化学习模型，成功捕获了以往研究发现的关键计算机制特征，但不同社会价值取向个体在强化学习模型的选择及关键参数上未表现出显著差异，这提示信任学习过程仍存在更多待探明的关键机制。

基于上述研究创新与发现，我们归纳出本研究的三点意义：

(1) 研究推动了将外部线索(面孔可信度)与内在特质(社会价值取向)整合于同一信任研究框架的理论视角；

(2) 研究为经典信任理论提供了直接的实证支持；

(3) 研究为从个体差异角度揭示信任学习的计算机制提供了新视角，拓展了该领域的研究内

容。

研究意义的具体论述已补充至正文“讨论”部分的结尾处，且严格遵循“研究创新—研究发现—研究意义”的逻辑展开：先明确本研究的创新点，再简要总结对应研究结论，最后推导并阐述研究发现的意义。该逻辑框架确保研究意义并非泛泛而谈，而是紧密依托研究问题与发现，具备明确的针对性与说服力。

同时，我们结合上述研究问题、核心发现与研究假设，对文章摘要进行了重新修缮与调整。优化后的摘要清晰涵盖研究背景、研究设计与发现和研究意义三部分：

其中，研究背景部分与上述研究创新相呼应。心理学研究创新的本质是通过结合现实心理问题，对以往研究的不足进行延伸探究，故以往研究的局限即为本研究的背景。摘要部分对研究背景的具体阐述为：

“信任研究尚缺乏对面孔可信度和社会价值取向在单次信任与信任学习中的共同作用的系统探讨，关于互惠预期的机制假设也缺少实证检验，且在构建信任学习的强化学习模型时未纳入社会价值取向。”

在研究背景的基础上，摘要简要阐述了研究设计与核心发现：

“本研究结合信任博弈范式与强化学习模型，通过两个实验回应上述研究缺口。实验 1 发现，面孔可信度与社会价值取向分别正向影响单次信任，二者不存在共同作用，互惠预期在二者的正向影响中起中介作用。实验 2 发现，社会价值取向显著调节初始面孔可信度与回馈概率对信任学习的共同作用。强化学习模型成功捕获了以往研究发现的关键计算机制特征，但不同社会价值取向个体的模型参数无显著差异，表明信任学习存在尚未揭示的计算机制。”

摘要最后基于研究发现，明确了研究意义：

“研究发展了整合外部线索与内在特质的信任研究框架，为经典信任理论提供了实证依据，并从计算模型视角拓展了信任学习中个体差异的研究路径。”

上述摘要修缮严格贴合文章“研究问题提出—研究实验设计—研究发现与意义论述”的核心逻辑，更准确、清晰地反映了论文的整体架构；同时，在表述层面进行了打磨，确保语言的准确性与流畅性。例如，我们将原有表述“强化学习模型成功复现了以往研究发现的关键计算机制特征”，修改为“强化学习模型成功捕获了以往研究发现的关键计算机制特征”。使用“捕获”(capture)这一术语，能够更准确、更专业地体现计算建模的研究范式，其含义指模型有效表征和拟合了已有研究中报告的计算规律与行为模式，该表述在计算建模研究的语境

中也更为常见和贴切(Daw, 2011; Jin et al., 2023)。文中其他相关表述也已同步作出修改，以保持术语的一致性与科学性。

最后，针对“文中所有图中文字显示为乱码”的问题，我们已重新将图片由原来的 emf 格式统一转换为 tiff 格式，并重新导入 Word 文档。将修改后的全文发送给相关人员核查。截至目前，核查结果均显示图片文字无乱码。为进一步规避后续可能出现的意外情况或格式兼容性问题，同时考虑到原始 emf 格式在清晰度上的优势，我们已将文章全部图片以多格式形式(包括 emf、tif 及 png)单独打包上传，供编辑部与审稿专家在后续查看中若仍遇问题时，可通过独立图片包验证，保障研究内容的完整呈现。

此外，在本次修改中，我们也对全文进行了细致的语言和表述错误修正，敬请审阅。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII, 23*(1), 3−38.

Jin, Y., Gao, Q., Wang, Y., Dietz, M., Xiao, L., Cai, Y., Bliksted, V., & Zhou, Y. (2023). Impaired social learning in patients with major depressive disorder revealed by a reinforcement learning model. *International journal of clinical and health psychology, 23*(4), 100389.

第一次修改

编辑部意见：

请作者务必根据审稿人的意见认真修改和回应。同时解决以下问题：

1. 文章被试数预估效应量根据何种标准确定？理论上，在进行被试量预估时，效应量需要根据预实验或者使用同样任务的前人文献进行确定，以最大程度上接近实际值，保证被试量预估合适。若被试量预估时使用过大的效应量设定，则会导致预估被试数的显著减少，导致实际实验被试数不足、统计功效不足。

回应：感谢编辑部老师的宝贵意见！我们参考了使用同样任务的前人文献来确定效应量，从而最大程度上接近实际值，保证被试量预估合适。具体细节已在回复审稿人 1 的意见 6 时进行了说明。

2. 请作者注意统计值报告规范，p 值需要精确到小数点后 3 位并报告具体值。

回应：感谢编辑部老师的宝贵意见！我们已对全文的统计值报告进行了全面核查和修正，确保所有 p 值均精确到小数点后 3 位，并报告具体值。

3. 审稿人认为本文有很多逻辑和表述不清、甚至混乱的地方。请作者认真对待，对文字和语言进行进一步提升，并请中级职称及以上同行进行挑剔性阅读。

回应：感谢编辑部老师的宝贵意见！我们已邀请具有中级及以上职称的同行专家对文章进行了细致的挑剔性阅读，并依据同行专家的反馈，对全文的语言表达和逻辑结构进行了系统性修订和提升。在修改稿中，我们已用蓝色字体标注所有修改内容，以便审阅。

审稿人一：

本文采用单轮和多轮信任博弈的实验范式，并结合计算建模方法，考察了面孔可信度、社会价值取向及回馈概率等多重因素对个体信任行为及信任学习的影响。研究探讨了外部线索（如面孔特征）与内部特质（如社会价值取向）在信任决策中的交互作用，研究视角具有一定的创新性，分析深入，兼具理论价值与实践意义。然而，文章在行文表述、方法与结果报告的规范性、结果的解释与讨论等方面仍存在需要完善之处。具体如下：

意见 1：引言中对互惠预期等关键概念的定义缺乏界定和解释。

回应：感谢审稿专家的宝贵意见！我们完全认同对关键概念进行界定和解释的重要性，并根据您的建议，进行了下述修改：

1. 我们在修改稿的引言第 4 段开头，新增了互惠预期概念的补充界定和理论解释：

“互惠预期是个体对他人是否遵守互惠规范的积极心理预期(Krueger et al., 2008; Dunning et al., 2014; Krueger et al., 2020)。”

这一概念界定的理论基础有：

（1）社会规范理论(Krueger et al., 2008)强调，个体对他人回报行为具有积极心理预期源于个体相信人们能够遵守互惠规范。

（2）道德规范理论(Dunning et al., 2014)也强调，个体对他人回报行为具有积极心理预期源于个体内化的道德规范使个体相信人们能够遵守互动规范。

（3）我们新增了 Krueger 等(2020)的研究。该研究从神经机制层面揭示了个体的互惠预期即个体预期他人能够遵守互惠规范。

2. 我们在修改稿的引言第 5 段中间，新增了对回馈概率概念的补充界定和理论解释：

“回馈概率指在重复信任互动中，个体观察到的被信任者实际选择回馈行为的相对频率(Chang et al., 2010)。”

这一概念的界定主要借鉴 Chang 等(2010)在经典信任学习研究中提出的“经验可信度水平”(level of experienced trustworthiness)概念，即被信任者回报信任的可能性。我们在此概念的基础上提出“回馈概率”概念，并在操作层面上借鉴了 Chang 等人对经验可信度水平的操作性定义(将 20%回馈概率设定为低回馈概率，80%回馈概率设定为高回馈概率)。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87−105.

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology, 107*(1), 122–141.

Krueger, J. I., Massey, A. L., & DiDonato, T. E. (2008). A matter of trust: From social preferences to the strategic adherence to social norms. *Negotiation and Conflict Management Research, 1*(1), 31−52.

Krueger, F., Bellucci, G., Xu, P., & Feng, C. (2020). The critical role of the right dorsal and ventral anterior insula in reciprocity: Evidence from the trust and ultimatum games. *Frontiers in Human Neuroscience, 14*, 176.

意见 2：引言部分需进一步明确两个实验的逻辑关系以及放到同一个研究中进行报告的意义。

回应：感谢审稿专家的宝贵意见！关于两个实验的逻辑关系以及放到同一个研究中进行报告的意义，我们作如下说明：

1. 两个实验的逻辑关系：

Chang 等(2010)提出信任是一个动态发展的过程。信任发展过程包含两个关键阶段：（1）基于先验信息的初始信任形成阶段(对应实验 1 采用单轮信任博弈探究的信任)；（2）基于互动经验的信任动态调整阶段(对应实验 2 采用重复信任博弈探究的信任学习)。因此，从实验 1 到实验 2 体现了从考察静态单次信任到探究动态信任学习的完整研究逻辑，这与信任发展的客观过程一致。

2. 将两个实验放在同一个研究中进行报告的意义：

我们认为将两个实验整合在同一研究中具有两个重要意义：（1）完整的信任研究需要同时考察初始信任形成和后续信任学习发展两个阶段，才能全面理解信任现象。（2）能够系统全面地揭示面孔可信度与社会价值取向在信任不同阶段的作用差异。如面孔可信度与社会价值取向在静态单次信任中没有共同作用，但是在动态信任学习过程中却具有显著的共同作用。

基于上述说明，我们在修改稿的引言第 11 段补充了两个实验的逻辑关系以及放到同一个研究中进行报告的意义，具体修改内容详见正文对应部分。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87−105.

意见 3：两个实验一共提出了 7 个“假设”（原文中漏写了“H2”），实际上大多是对结果的预期，而不是针对关键科学问题的假设，建议进一步提炼。

回应：感谢审稿专家的宝贵意见！我们完全认同假设应当反映关键科学问题而非具体结果预测。根据您的建议，我们对研究假设进行了重新提炼，使其针对信任研究领域尚未解决的关键科学问题。具体修改如下：

1. **假设 1(整合原假设 1-3)**：面孔可信度与社会价值取向对信任存在共同影响。

关键科学问题：根据人情互动理论(Magnusson & Stattin, 1998)与信任发展的二元互动模型(Simpson, 2007)，作为外部线索的面孔可信度与作为内在特质的社会价值取向是否可能共同影响信任？

2. **假设 2(原假设 4)**：互惠预期是面孔可信度和社会价值取向影响信任的潜在心理机制。

关键科学问题：社会规范理论(Krueger et al., 2008)与道德规范理论(Dunning et al., 2014)均提出互惠预期是面孔可信度与社会价值取向影响信任的潜在心理机制，这种理论假设是否能得到实证检验？

3. **假设 3(原假设 5)**：在信任学习中，初始面孔可信度与回馈概率的动态交互作用会因社会价值取向而不同。

关键科学问题：根据信任的神经心理经济学模型(Krueger & Meyer-Lindenberg, 2019)推测，不同社会价值取向个体在信任学习中会采取不同策略(Strategy)，这是否会使得初始面孔可信度与回馈概率动态交互的信任学习过程在不同社会价值取向个体中呈现差异？

4. **假设 4(原假设 6 与假设 7)**：个体在信任学习中根据初始面孔可信度和回馈概率的一致性采取不同学习率；初始面孔可信度对信任学习具有稳定的影响；随着时间推移，回馈概率的预测误差对价值期望的影响逐渐减弱。社会价值取向通过调节学习率、面孔可信度影响系数或衰减系数来影响信任学习。

关键科学问题：通过强化学习模型能否验证并量化信任学习的动态心理过程的下述关键特征？（1）个体在信任学习中根据初始面孔可信度和回馈概率的一致性采取不同学习率(Fareri et al., 2012)；（2）初始面孔可信度对信任学习具有稳固的影响(Chang et al., 2010; Fouragnan et al., 2013)；（3）随着时间的发展，回馈概率的预测误差对价值期望的影响逐渐减少(Westhoff et al., 2020)。同时社会价值取向是否通过信任学习的心理动态过程中的学习率、面孔可信度影响系数或衰减系数来具体影响信任学习过程。

在修改稿中，我们将上述假设分别嵌入相应的理论论述段落。假设 1 置于引言段落 3 末尾，假设 2 置于引言段落 4 末尾，假设 3 置于引言段落 6 末尾，假设 4 置于新增引言段落 7-10

中第 10 段末尾，紧接在对假设 4 所探究的关键科学问题的理论阐述后。这种安排使理论推导与假设提出之间的逻辑关系更紧密。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87−105.

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at Zero Acquaintance: More a Matter of Respect Than Expectation of Reward. *Journal of Personality and Social Psychology, 107*(1), 122−141.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience, 33*(8), 3602−3611.

Krueger, J. I., Massey, A. L., & DiDonato, T. E. (2008). A matter of trust: From social preferences to the strategic adherence to social norms. *Negotiation and Conflict Management Research, 1*(1), 31−52.

Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences, 42*, 92−101.

Krueger, F., Bellucci, G., Xu, P., & Feng, C. (2020). The Critical Role of the Right Dorsal and Ventral Anterior Insula in Reciprocity: Evidence From the Trust and Ultimatum Games. *Frontiers in human neuroscience, 14*, 176.

Magnusson, D., & Stattin, H. (1998). Person-context interaction theories. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (5th ed., pp. 685–759). John Wiley & Sons, Inc.

Simpson, J. A. (2007). Psychological Foundations of Trust. *Current Directions in Psychological Science, 16*, 264−268.

Westhoff, B., Molleman, L., Viding, E., van den Bos, W., & van Duijvenvoorde, A. C. K. (2020). Developmental asymmetries in learning to adjust to cooperative and uncooperative environments. *Scientific reports, 10*(1), 21761.

意见 4：两个实验所招募的被试中女性明显多于男性，这种性别分布对实验效应的影响有待考量。另外，文中缺少各组内性别分布的信息。

回应：感谢审稿专家的宝贵意见！关于性别分布对实验效应的影响，我们已进行系统检验，具体操作如下：

在实验 1 的线性混合效应模型加入性别作为固定效应：信任行为 ~ 面孔可信度×SVO° + 性别 + (1|被试编号)。结果发现，性别对信任行为不具有显著主效应(详见表 2)。加入性别后，关键实验结论保持稳定(详见表 1，表 2)。综合分析表明性别分布不显著影响实验效应。

表 1　面孔可信度和社会价值取向对信任行为的影响(未加入性别)

因变量	自变量	回归系数	标准误	*t*	*p*
信任行为	社会价值取向	.42	.19	2.27	.025
	面孔可信度	.34	.12	2.87	.004
	社会价值取向×面孔可信度	.19	.12	1.60	.111

表 2　面孔可信度和社会价值取向对信任行为的影响(加入性别)

因变量	自变量	回归系数	标准误	*t*	*p*
信任行为	社会价值取向	.43	.19	2.31	.023
	面孔可信度	.34	.12	2.87	.004
	性别	−.17	.23	−.75	.452
	社会价值取向×面孔可信度	.19	.12	1.60	.111

在实验 2 新采用的线性混合效应模型加入性别作为固定效应：信任行为 ~ SVO° ×初始面孔可信度×回馈概率 + 性别 + (1|被试编号)。结果发现，性别对信任行为不具有显著主效应(详见表 4)。加入性别后，关键实验结论保持稳定(详见表 3，表 4)。综合分析表明性别分布不显著影响实验效应。

表 3　初始面孔可信度、回馈概率和社会价值取向对信任学习的影响(加入性别)

因变量	自变量	回归系数	标准误	*t*	*p*
信任行为	社会价值取向	.10	.18	.55	.586
	初始面孔可信度	.15	.05	2.86	.004
	回馈概率	2.23	.05	42.23	<.001
	性别	−.14	.44	−.31	.755
	社会价值取向×初始面孔可信度	.10	.05	1.94	.053
	社会价值取向×回馈概率	−.15	.05	−2.92	.004
	初始面孔可信度×回馈概率	−.51	.11	−4.81	<.001
	社会价值取向×初始面孔可信度×回馈概率	.53	.11	5.03	<.001

表 4　初始面孔可信度、回馈概率和社会价值取向对信任学习的影响(未加入性别)

因变量	自变量	回归系数	标准误	*T*	*p*
信任行为	社会价值取向	.10	.18	.57	.571
	初始面孔可信度	.15	.05	2.86	.004
	回馈概率	2.23	.05	42.23	<.001
	社会价值取向×初始面孔可信度	.10	.05	1.94	.053
	社会价值取向×回馈概率	−.15	.05	−2.92	.004
	初始面孔可信度×回馈概率	−.51	.11	−4.81	<.001
	社会价值取向×初始面孔可信度×回馈概率	.53	.11	5.03	<.001

我们已在修改稿的“2.2.1 实验被试与设计”与“3.2.1 实验被试与设计”部分补充了各社会价值取向组内的性别分布信息。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 5：2.2.3 中关于信任博弈范式投资人的投资金额范围表达前后不一，需要明确。

第一段中的投资金额范围包括 0：“在任务中，被试需要选择向被信任者投资任意数额 X 元(0−10 元)，该投资金额翻 4 倍后给到被信任者。”第二段中不包括 0：“其次呈现被信任者的面孔图片和投资提示，被试需选择投资金额，1—10 表示投资金额为 1—10 元(由于 10 是两位数，用 0 进行反应)；”。

回应：感谢审稿专家的宝贵意见！感谢您对实验程序描述细节的细致审阅。关于信任博弈范式中投资金额范围的表述不一致错误，我们已将 2.2.3 第 1 段投资金额范围修正为：

在任务中，被试需要选择向被信任者投资任意数额 X 元(1−10 元)，该投资金额翻 4 倍后给到被信任者。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

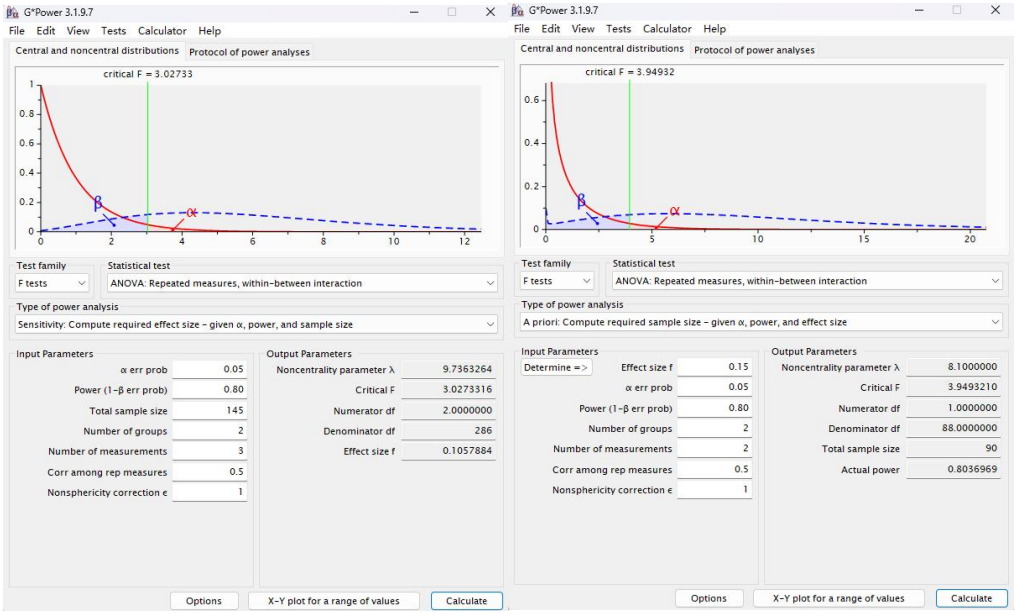
意见 6：两个实验在使用 Gpower 计算样本量时未提及使用的何种统计检验，以及所选取效应量的依据。

回应：感谢审稿专家的宝贵意见！我们就使用 G*Power 计算样本量时使用的统计检验，以及所选取效应量的依据作如下说明：

1. 实验 1

（1）统计检验类型：采用 2×2 混合设计 ANOVA 作为 G*Power 计算基础，对应实验设计的 2(面孔可信度：高/低，被试内)×2(社会价值取向：亲社会/亲自我，被试间)结构。

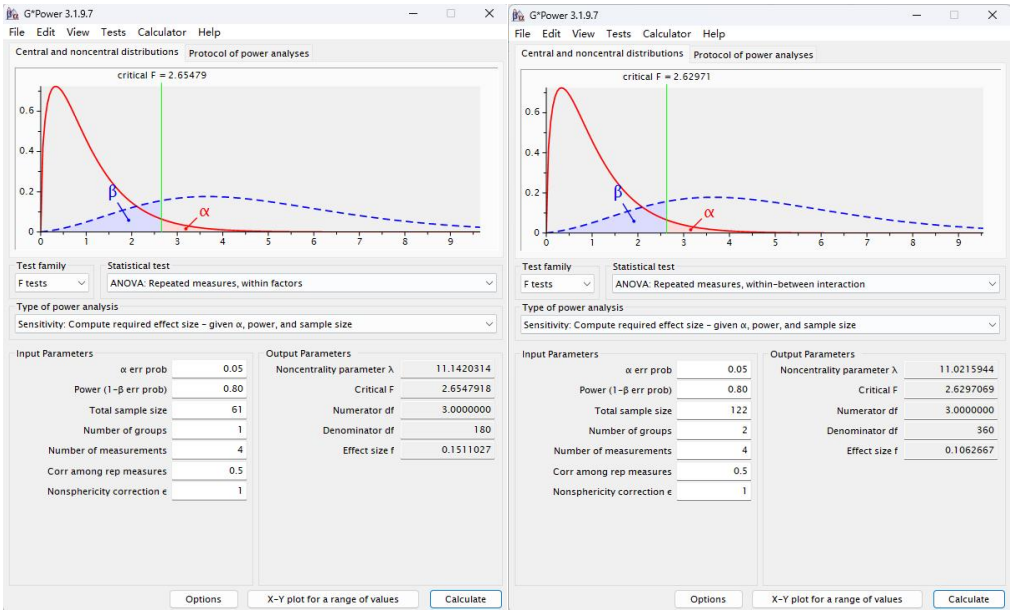
（2）效应量选取依据：我们参照了 Mill 与 Theelen 等(2019)关于群体规模不确定性和社会价值取向如何影响合作行为的研究。该研究采用 2(社会价值取向，被试间)×3(确定性程度，被试内)混合实验设计，实验设计与本研究实验 1 的设计理念和结构相似。通过 G*Power 计算得出该研究的效应量 f 约为.11(α = .05, (1-β) = .80)。基于该研究，我们重新保守设置效应量 f 为.15(α = .05, (1-β) = .80)，计算出实验 1 需要样本量 90 人。



2. 实验 2

（1）统计检验类型：采用 2×2×2 混合设计 ANOVA 作为 G*Power 计算基础，对应实验设计的 2(初始面孔可信度：高/低，被试内)×2(社会价值取向：亲社会/亲自我，被试间) ×2(回馈概率：高/低，被试内)结构。

（2）效应量选取依据：我们参照了两篇研究：第一篇是 Chang 等(2010)关于初始面孔可信度与互动经验如何共同影响信任学习的经典研究。该研究采用 2(初始面孔可信度，被试内)×2(经验可信度水平，被试内)重复测量实验设计，实验设计与本研究实验 2 的设计理念相似。通过 G*Power 计算得出该研究的效应量 f 约为.15(α = .05, (1-β) = .80)。第二篇是 Jin 等(2023)使用信任游戏和强化学习模型来探究重度抑郁症患者在社会互动中的学习能力的研究。该研究采用 2(被试类型，被试间)×2(声誉披露，被试内)×2(情景预测性，被试内)混合实验设计，实验设计与本研究实验 2 的结构相似。通过 G*Power 计算得出该研究的效应量 f 约为.11(α = .05, (1-β) = .80)。综合考量上述两研究，我们重新设置效应量 f 为.12(α = .05, (1-β) = .80)，计算出实验 2 需要样本量 96 人。



基于上述分析，我们修缮了文章“2.2.1 实验被试与设计”与“3.2.1 实验被试与设计”部分关于使用 G*Power 计算效应量与样本量的表述，具体修改内容详见正文对应部分。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87−105.

Jin, Y., Gao, Q., Wang, Y., Dietz, M., Xiao, L., Cai, Y., Bliksted, V., & Zhou, Y. (2023). Impaired social learning in patients with major depressive disorder revealed by a reinforcement learning model. *International journal of clinical and health psychology, 23*(4), 100389.

Mill, W., & Theelen, M. M. P. (2019). Social value orientation and group size uncertainty in public good dilemmas. *Journal of behavioral and experimental economics, 81*, 19–38.

意见 7：2.4 和 3.4 部分只有对结果的简单描述，但是没有对结果进行解释和讨论。

回应：感谢审稿专家的宝贵意见！我们完全认同这些部分需要更进一步的解释和讨论。根据您的建议，我们对 2.4 和 3.4 部分进行如下改进：

改进内容包括：（1）在描述结果的基础上解释结果表明了什么结论；（2）在解释结果的基础上讨论研究假设是否得到验证。修改后的 2.4 部分与 3.4 部分详见正文对应部分。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 8：实验二的方法描述部分需要更规范清晰，例如，“初始可信度”是否就是“面孔可信度”？“回馈概率：低回馈概率: 20% vs 高回馈概率: 80%”具体是如何操纵的？在探究回馈概率对个体信任行为的影响时，使用的是本轮次的回馈概率还是上一轮次的回馈概率？

回应：感谢审稿专家的宝贵意见！我们完全认同方法描述部分需要更规范清晰，根据您的建议，我们对实验 2 的方法描述进行下述解释与改进：

（1）“初始可信度”是个体在信任学习前基于被信任者面孔产生的初始可信度评估，因此我们采用更精确的“初始面孔可信度”表述，并已在全文统一替换为该表述。

（2）实验 2 使用 4 个被信任者面孔(编号分别为 1、2、4、5)，其中面孔 1、5 为高可信度面孔，面孔 2、4 为低可信度面孔。每个面孔出现 20 次，共进行 80 轮信任学习试次。实验采用伪随机实验设计，通过程序控制被信任者的反馈选择，从而操纵各个被信任者的回馈概率，实现对初始面孔可信度与回馈概率两种条件的操纵。具体操作如下：面孔 1、2 的 20 轮试次中有 16 个试次被信任者选择平分资金(即有 80%回馈的高回馈概率)；面孔 4、5 的 20 轮试次中仅有 4 个试次被信任者选择平分资金(即只有 20%回馈的低回馈概率)。因此面孔 1 构成了高初始面孔可信度—高回馈概率(一致条件)，面孔 2 构成了低初始面孔可信度—高回馈概率(不一致条件)，面孔 4 构成了低初始面孔可信度—低回馈概率(一致条件)，面孔 5 构成了高初始面孔可信度—低回馈概率(不一致条件)。这种设置既实现了对高低回馈概率的操控，同时保证了 80 轮试次中回馈与不回馈次数平衡(共 40 次回馈)。我们在“3.2.3 实验任务与程序”部分新增段落 2，详细说明了实验如何具体操纵回馈概率与两种条件(一致、不一致)，

具体修改内容详见相应部分。

（3）在探究回馈概率对信任行为的影响时，我们使用的是被试本轮次的回馈概率。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 9：无论在正文还是附录中，计算建模的过程否缺乏公式、参数心理意义解释等必要的细节描述，文章对模型拟合、模型比较的结果报告过于简单，作者是否有进行模型参数恢复和模型恢复？

回应：感谢审稿专家的宝贵意见！我们完全认同需要对研究模型进行更全面的补充说明，并进一步完善模型拟合与比较的结果报告。我们针对您提出的问题，对文章的附录内容、正文内容以及模型比较进行了下述补充与完善：

1. 附录补充：详细呈现了计算模型的构建细节，包括 12 个候选模型的公式与参数恢复结果。

2. 正文新增：在引言部分(第 7-10 段)新增了关于计算模型假设(假设 4)的论述，进一步阐释了模型公式及其参数的心理意义。

3. 模型比较的完善：针对模型比较结果报告过于简单的问题，我们新增了两个模型拟合指标——包含二阶偏差修正的赤池信息准则(Akaike Information Criterion corrected, AICc)与贝叶斯信息准则(Bayesian Information Criterion, BIC)，并更新了正文中的模型比较结果呈现。具体表现为：

（1）模型 AIC 值：现有结果发现，模型 4a 的平均 AIC 值最小。但配对样本 t 检验表明，模型 4a 的 AIC 仅边际显著低于模型 2a，$t(100) = -1.95$, $p = .054$, Cohen's $d = -.19$, 95%CI = [−4.05, .03]。说明模型 4a 在 AIC 指标上并不具有显著优势。

（2）模型 AICc 值与 BIC 值：虽然模型 4c 在 AIC 指标上未呈现优势，但其平均 AICc 和 BIC 值均为所有候选模型中最低。配对样本 t 检验表明：模型 4c 的 AICc 显著低于模型 2a，$t(100) = -2.99$, $p = .003$, Cohen's $d = -.30$, 95%CI = [−9.93, −2.01]。模型 4c 的 BIC 值显著低于模型 2c，$t(100) = -2.06$, $p = .042$, Cohen's $d = -.21$, 95%CI = [−9.43, −.17]。这些结果一致表明，模型 4c 在 AICc 与 BIC 指标上均显著最优。鉴于 AICc 指标在模型的参数数目不同，且试次-参数比<40 的情况下相比 AIC 具有更好的稳健性，因此更适合作为本研究的模型拟合判别指标。

4. 参数恢复验证：附录中补充了模型参数恢复结果，结果支持了对每位被试进行参数估计的可靠性，具体结果详见附录。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research, 33*(2), 261–304.

意见 10：实验 2 在分析社会价值取向、初始可信度和回馈概率对个体信任行为的影响时为什么不和实验 1 保持一致使用混合效应模型？而是使用方差分析？

回应：感谢审稿专家的宝贵意见！我们完全认同统计方法的一致性对研究结论可靠性的关键作用。根据您的建议，我们已采用线性混合效应模型重新分析实验 2，将 3.3.1 部分替换为线性混合效应模型结果，同时将方差分析结果移至附录。

线性混合效应模型的实验结果不仅重复了方差分析的主要发现，还进一步发现了三个新结果：

（1）初始面孔可信度对信任行为具有显著主效应(β = .15, *SE* = .05, *t* = 2.86, *p* = .004)；

（2）社会价值取向和回馈概率对信任行为具有显著交互作用(β = −.15, *SE* = .05, *t* = −2.92, *p* = .004)；

我们认为这两个新结果由于不直接涉及本研究假设所探讨的核心问题，因此仅将结果在附录部分呈现，未在正文部分进行呈现与深入解释。

（3）社会价值取向、回馈概率和初始面孔可信度对信任行为具有显著交互作用(β = .53, *SE* = .11, *t* = 5.03, *p* < .001)。

值得注意的是，线性混合效应模型在亲社会者上的简单效应分析结果与方差分析结果存在差异。方差分析结果显示，亲社会者在高回馈概率($F(1, 99) = 1.83$, $p = .182$)和低回馈概率条件($F(1, 99) = 1.44$, $p = .235$)下的投资金额均不存在显著差异。而线性混合效应模型结果则显示，亲社会者无论是在高回馈概率还是低回馈概率条件下，对高初始面孔可信度被信任者的投资金额(高回馈: $M = 5.73$, $SD = 3.10$; 低回馈: $M = 3.64$, $SD = 2.91$)都显著高于低初始面孔可信度被信任者(高回馈: $M = 5.46$, $SD = 3.04$; 低回馈: $M = 3.40$, $SD = 2.80$) (高回馈: $M_{diff} = -.27$, $SE = .11$, 95%CI = [−.47, −.06], $z = -2.51$, $p = .012$; 低回馈: $M_{diff} = -.24$, $SE = .11$, 95%CI = [−.45, −.03], $z = -2.28$, $p = .022$)。

我们认为，这一结果差异可以从以下两个方面来理解：

首先，两种分析方法都揭示了亲社会者信任决策的稳定性特征，这种特征在方差分析中表现为无显著偏好，而在线性混合效应模型中表现为稳定偏好高初始面孔可信度被信任者。因此两个结果具有内在共同性。

其次，线性混合效应模型结果的新发现可用初始面孔可信度的持续作用机制来解释(Chang et al., 2010)。高初始面孔可信度被信任者使亲社会者产生了更高的互惠预期，从而产生更多信任偏好(线性混合效应模型新发现的初始面孔可信度对信任行为的显著主效应也体现了这一机制)。同时，由于亲社会者倾向于采用基于他人利益的社会理性策略，其信任决策本身具有稳定性，因此无论回馈概率高低，在整个信任学习过程中都始终更信任高初始面孔可信度被信任者。

基于这些发现，我们修改了 3.4 部分的结果表述，并在讨论部分第 3 段新增了如下解释：

“研究还推测，受高面孔可信度的初始影响，亲社会者在信任学习开始前更偏好高面孔可信度被信任者。然而，由于亲社会者更倾向于采用基于他人利益的社会理性策略，因此其信任决策具有稳定性，不易受回馈概率影响，也较少调整信任策略。这使得亲社会者始终更信任高面孔可信度被信任者。”

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87−105.

意见 11：衰减系数统计分析部分的“面孔可信度影响系数”是什么？在前后文中未看到相关的概念解释。

回应：感谢审稿专家的宝贵意见！非常感谢您指出这一概念前后文表述不清的问题。针对您的建议，我们已进行如下完善：

（1）在修改稿引言新增段落 8 中，我们增添了对“面孔可信度影响系数”的明确定义，具体表述为：

面孔可信度作为先验可信度信息，也会产生相应影响，其影响程度由“面孔可信度影响系数”量化。

（2）在修改稿引言新增段落 9 与扩充的附录中，我们补充了模型 4c 的完整公式表达，明确呈现了该参数在模型公式中的数学表达形式。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

意见 12：实验 2 中有一个有趣的结果是亲自我取向的个体在高回馈概率条件下，对低面孔可信度的投资金额显著高于高面孔可信度的被信任者，但讨论时的解释是“在高回馈概率条件下会更信任低初始可信度被信任者，从而最大化自身利益。”，这样的解释仍然不清楚为什么向低面孔可信度的被信任者投资更多的钱可以最大化自身的利益。

回应：感谢审稿专家的宝贵意见！非常感谢您对这一有趣发现的深入追问。我们完全同意原解释较为简略，现已对解释部分进行如下优化：

我们认为这一有趣的结果仍可从初始面孔可信度、回馈概率与社会价值取向共同作用的视角出发来解释。首先，面孔可信度在信任学习开始前使信任者形成了对被信任者的初始可信度评估(“他看起来可信/他看起来不可信”)。接着，在高回馈概率条件下，高初始面孔可信度被信任者与高回馈概率形成一致情况，这强化了信任者原有的对高初始面孔可信度被信任者的先验可信度评估(“这个看起来可信的人真的可信”)。低初始面孔可信度被信任者与高回馈概率形成不一致情况，这促使信任者更新原有的对低初始面孔可信度被信任者的先验可信度评估(“这个看起来不可信的人是可信的”)。亲自我者对这两种情况的反应基于其追求自身利益最大化的决策取向。对高初始面孔可信度者，亲自我者会将其高回馈视为基本的收益(认为这是对方本应做到的)。对低初始面孔可信度，亲自我者会将其高回馈视为额外的收益(因为初始期望低，任何回馈都被视为是额外收益)。根据道德规范理论(Dunning et al., 2014)，亲自我者在社会投射的作用下认为对方也采取利己策略，因此会通过增加信任行为来诱导对方更多回馈，从而实现自身利益最大化。

我们将上述分析补充进修改稿讨论部分第 3 段，具体内容为：

“研究推测亲自我者倾向于采用基于自我利益的经济理性策略。因此，在低回馈概率条件下会更信任高初始面孔可信度被信任者，从而规避潜在损失。而在高回馈概率条件下，亲自我者会将高初始面孔可信度被信任者的高回馈视为基本的收益，因而不会对其产生更多信任。相反，亲自我者会将低初始面孔可信度被信任者的高回馈视为额外的收益，并在社会投射的作用下倾向于认为对方也采取基于自我利益的经济理性策略，从而尝试通过增加信任行为来利好对方，使对方回馈更多，以谋取最大化自身利益。”

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at Zero Acquaintance: More a Matter of Respect Than Expectation of Reward. *Journal of Personality and Social Psychology, 107*(1), 122−141.

意见 13：讨论部分未讨论和解释多种因素对回馈概率学习率和面孔可信度学习率的影响，尚待补充。

回应：感谢审稿专家的宝贵意见！感谢您指出我们讨论部分未讨论和解释多种因素对回馈概率学习率和面孔可信度学习率的影响的缺陷。

研究在分析多种因素对模型参数的影响时主要有下述显著结果：

（1）增益损失情境对回馈概率学习率具有显著的影响，$F(1,99) = 19.56$, $p < .001$, $\eta^2 = .04$。事后检验分析表明，增益情境下的回馈概率学习率($M = .20$, $SD = .20$)显著低于损失情境($M = .28$, $SD = .22$), $t(99) = -4.42$, $p < .001$, Cohen's $d = -.40$, 95%CI = [−.12, −.05]。

（2）增益损失情境对面孔可信度学习率具有显著的影响，$F(1,99) = 10.19$, $p = .002$, $\eta^2 = .03$。事后检验分析表明，增益情境下的回馈概率学习率($M = .36$, $SD = .40$)显著低于损失情境($M = .40$, $SD = .38$), $t(99) = -3.19$, $p = .002$, Cohen's $d = -.35$, 95%CI = [−.21, −.05]。

我们在正文的模型参数统计分析结果部分重点呈现了这两个显著的结果，同时为了保持研究结果呈现的完整性，将其余不显著的所有结果均在附录部分进行了补充呈现，具体补充内容详见附录部分。

根据您的建议，我们在修改稿讨论部分第 5 段补充了对这两个结果的解释说明。同时，我们也对分析结果中一些未能符合初始假设、未能复现前人研究结果的情况，在修改稿讨论部分新增段落 4-6 进行了深入讨论，具体说明了为何其它因素未能对各模型参数产生显著影响，具体补充内容详见原文修改部分。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

审稿人二：

本研究在对人际交互中信任的影响因素方面提出了一种新颖的视角，探讨了面孔可信度和社会价值取向对信任与信任学习的共同作用与机制。文章在理论构建、研究方法或实验设计上展现出一定的创新性。然而，文章目前存在较多问题。以下是具体的问题及建议，希望能对文章的进一步完善有所帮助。

1. 论文的言语表述方面

（1）引言第一段，“是否交互作用及其影响信任的心理机制尚存研究空白” 可以改为更学术地表达，例如“是否共同影响...”可能更为恰当，后同。

回应：感谢审稿专家的宝贵意见！我们完全认同您提出的学术表达修改建议，并针对该建议将表述“是否交互作用及其影响信任的心理机制尚存研究空白”修改为“然而，目前尚不清楚这两种因素是否共同作用于信任，以及它们影响信任的潜在心理机制。”，采用更严谨的表述方式。

此外，我们已对全文进行了统一修改，确保相关表述保持一致性，具体包括：
将“两者不存在交互作用”修改为“二者不存在共同作用”，见于修改稿摘要部分；
将“因此面孔可信度与社会价值取向可能交互影响信任行为”修改为“因此，面孔可信度与社会价值取向可能共同影响信任。”，见于修改稿引言第 3 段。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（2）引言第二段，“面孔可信度显著影响信任决策。研究表明，面孔可信度影响个体对面孔所有者的道德品质和行为动机的推断，面孔可信度较高的对象更容易取得个体的信任(Ewing et al., 2015; Qi et al., 2022)” 可以稍作调整，阅读起来会更加顺畅，譬如：研究表明，面孔可信度显著影响信任决策，即面孔可信度影响个体……。

回应：感谢审稿专家的宝贵意见！您提出的关于引言第 2 段表述流畅性的建议非常中肯。我们已根据您的建议将原表述修改为：

“研究表明，面孔可信度显著影响信任，具体表现为影响个体对面孔所有者的道德品质和行为动机的推断，从而使面孔可信度较高的对象更容易获得信任(Ewing et al., 2015; Qi et al., 2022)。”

这一修改将两个分句整合为一个更加连贯的表述，通过加入过渡短语“具体表现为”，使前

后逻辑关系更加紧密。同时也将表述“取得个体的信任”优化为更通顺的“获得信任”。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（3）引言第三段，“例如，高可信面孔可以适当减少亲自我者的风险规避，从而引发其进行更多信任行为”，如果是作者推测而出，请说明；如果不是请表明引用文献，类似的举例表述建议修改，后文同。

回应：感谢审稿专家的宝贵意见！我们完全同意您对论文举例表述严谨性的细致指正，这一表述确实存在引用标注不明确的问题。经查证，这一观点主要基于以下研究证据：Qi 等人(2022)的研究发现，高可信面孔能够显著降低个体的风险感知。

我们已将原文修改为：“已有研究表明，高可信面孔能够显著降低个体的风险规避倾向，从而可能促进亲自我者表现出更多信任(Qi et al., 2022)。”。这一修改明确标注了文献支持，并使用“可能”来阐明推测性质，使表述更加严谨客观。

我们也对文中所有类似表述进行了全面检查，具体包括：

将“在互动情境中，外部线索(如面孔可信度)通过影响互惠预期间接调节信任。”改为“在互动情境中，面孔可信度可能通过提供外部的互惠认知线索来影响互惠预期，从而影响信任。”。这一修改明确了如何基于社会规范理论，推测互惠预期是面孔可信度影响信任的心理机制，并加入修饰词“可能”来强调该观点的推测性质，见于修改稿第 4 段。

将“例如，亲社会者更倾向于认为他人具有亲社会动机，从而形成更高的互惠预期，并表现出更多的信任。”改为“因此，亲社会者可能更倾向于认为他人具有亲社会动机，从而形成更高的互惠预期，并表现出更多的信任。”。这一修改通过加入连接词“因此”明确了该观点与道德规范理论的逻辑关系，并加入修饰词“可能”来强调观点的推测性质，见于修改稿第 4 段。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Qi, Y., Luo, Y., Feng, Y., Chen, Z., Li, Q., Du, F., & Liu, X. (2022). Trustworthy faces make people more risk-tolerant: The effect of facial trustworthiness on risk decision-making under gain and loss conditions. *PsyCh journal (Victoria, Australia), 11*(1), 43−50.

（4）引言第四段，“根据社会规范理论与道德规范理论，可推断面孔可信度和社会价值取向均通过互惠预期影响信任行为”，结论或判断类的句式后建议添加引用文献，之后再详细介绍文献的具体内容；或作者可以尝试更改表述顺序，先放具体介绍及文献引用，后放文献结论/判断，后文同。

回应：感谢审稿专家的宝贵意见！我们完全同意您对论文理论框架严谨性的宝贵建议，针对引言第 4 段的表述问题，我们已对该段落进行了语言表述与顺序上的修改：先呈现具体理论介绍与文献引用，后基于理论自然推导出研究判断与研究假设，使理论到假设的逻辑链条更加完整。详细修改可见修改稿引言第 4 段。

同时，我们也对文中所有的理论推导段落进行了统一校正，确保每个理论观点的提出都先介绍具体理论及文献引用，再基于理论推导出研究假设，从而保证逻辑链条完整。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（5）引言第五段，“然而…”，此处不应为转折，使用“同时...”或更加恰当。

回应：感谢审稿专家的宝贵意见！我们完全同意您指出的转折词使用问题，原文使用“然而”暗示与前文存在对立关系，而实际上此处是补充说明，故改为“并”更为恰当。基于此，我们将原表述修改为“这两个理论共同支持了互惠预期作为二者影响信任的心理机制的理论假设，并仍需实证研究的进一步验证。”。这一修改既保留了原文的核心内容，又更准确地表达了理论假设与实证验证之间的递进关系。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（6）引言第六段，“个体特质也会影响信任学习(Radell et al., 2016; Seaman et al., 2023)”，作者需要对所引文献做简要介绍，或对表达同一观念的几篇文献结果作总结，“影响”二字的表述过于宽泛，建议更具体说明。

回应：感谢审稿专家的宝贵意见！我们完全同意您对文献引述的重要建议。关于引言第 6 段的表述问题，我们已将原表述修改为“研究表明，个体特质会调节信任学习过程。例如，行为抑制特质较高的个体表现出更保守的信任学习策略(Radell et al., 2016)。老年个体对社会线索的学习策略也不同于年轻个体(Seaman et al., 2023)。这些发现提示，社会价值取向也可能通过影响个体的学习策略来调节信任学习。”

这一修改增加了具体研究发现，明确说明了不同特质影响信任学习的具体表现。通过“这些发现提示”自然过渡到本研究关注的社会价值取向变量。使用“调节”这一更专业的术语替代了原表述中“影响”。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（7）类似“默认模式网络”等专有名词首次出现时应附英文全称及简写，以保持学术规范，后文同。

回应：感谢审稿专家的宝贵意见！ 我们完全同意您对论文专有名词表述规范性的重要指正。关于专有名词的表述方式，我们已在其首次出现时补充了英文全称及缩写，采用“中文(英文全称，缩写)”的标准格式，确保全文同类术语表述统一。我们已对文中所有相关专有名词进行了系统检查，具体包括下述术语：

社会价值取向(social value orientation, SVO)

中央执行网络(central-executive network, CEN)

默认模式网络(default-mode network, DMN)

奖励网络(reward network, RWN)

凸显网络(salience network, SAN)

信任博弈(trust game, TG)

社会价值取向滑块测验(Social value orientation Slide Measure, SVO slider measure)

三优势测量(The Triple-Dominance Scale)

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（8）引言第七段，“研究者通常采用社会经济博弈范式探究信任”，表述太空泛，建议明确“探究谁的信任”以及“在哪种情境下的信任”，以提升表述的精确性。

回应：感谢审稿专家的宝贵意见！我们完全同意您对研究表述精确性的重要建议。关于社会经济博弈范式的表述，我们已将原表述修改为“研究者普遍采用信任博弈(trust game, TG)这一社会经济博弈范式来探究个体与陌生人之间的信任，其中的重复信任博弈范式常被用于探究多次重复互动中的信任学习过程(Bellucci & Park, 2020)。”。

这一修改明确了信任博弈范式的研究对象为“个体与陌生人之间”的信任，界定了重复信任

博弈范式的应用情景为“重复多次互动”的信任。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Bellucci, G., & Park, S. Q. (2020). Honesty Biases Trustworthiness Impressions. *Journal of Experimental Psychology: General, 149*(8), 1567−1584.

（9）引言第七段，“重复信任博弈任务范式”，此类表述中“任务”和“范式”一般取其一使用，且在下一段开头这二者均未使用，请全文统一。

回应：感谢审稿专家的宝贵意见！我们完全同意您对术语使用一致性的细致指正。关于“任务”与“范式”的表述问题，我们参考了高青林与周媛(2021)关于信任博弈的综述，确认“信任博弈范式”为标准表述，因此全文已统一修改为：

“单次信任博弈范式”(原“单次信任博弈任务范式”)

“重复信任博弈范式”(原“重复信任博弈任务范式”)

同时，我们已对全文进行了系统检查，确保在理论阐述时统一使用“范式”，在具体操作层面使用“任务”(如“单轮信任博弈任务”)。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

高青林, 周媛. (2021). 计算模型视角下信任形成的心理和神经机制——基于信任博弈中投资者的角度. *心理科学进展, 29*(1), 178−189.

（10）实验一 2.2.1，“实验采用社会价值取向(SVO)量表对被试进行分类，以被试在社会价值取向滑块测验”，请作者统一表述为“...量表”或“...滑块测验”

回应：感谢审稿专家的宝贵意见！我们完全同意您对心理测量术语统一性的重要指正。关于社会价值取向测量工具的表述问题，我们已将原表述修改为“实验采用社会价值取向滑块测验(Social value orientation Slide Measure, SVO slider measure)对被试进行分类。”。

修改统一使用“滑块测验”作为标准术语，与原始开发文献(Murphy & Ackermann, 2014)保持一致。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Murphy, R. O., & Ackermann, K. A. (2014). Social Value Orientation: Theoretical and Measurement Issues in the Study of Social Preferences. *Personality and Social Psychology Review, 18*(1), 13−41.

（11）实验一 2.3，“使用线性混合模型(Linear Mixed Model，LMM)来控制假阳性”，请查阅相关文献后规范表述。

回应：感谢审稿专家的宝贵意见！我们同意您对统计方法表述严谨性的重要指正。关于线性混合效应模型的表述，我们参照文献，统一规范表述该模型为“线性混合效应模型(linear mixed-effects model)”。为确保表述的客观性，删除了对方法优势的主观性介绍。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

刘玥, 徐雷, 刘红云, 韩雨婷, 游晓锋, & 万志林. (2024). 置信区间宽度等高线图在线性混合效应模型样本量规划中的应用. *心理学报, 56*(01), 124−180.

（12）实验一 2.3，“结果显示(见表 1)，面孔可信度对信任行为具有显著主效应”，请规范表述，如：面孔信任度对个体信任行为具有显著的影响…，前后文同。

回应：感谢审稿专家的宝贵意见！我们同意您对统计结果表述规范性的重要建议。关于统计结果的表述方式，我们已进行如下修改完善：

将全文的主效应统计结果表述统一修改为：“XX 对 XX 具有显著的影响。”

将全文的交互效应统计结果表述统一修改为：“XX 与 XX 对 XX 具有显著的交互影响。”

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（13）实验一 2.3，“因变量为信任行为进行两水平被试内中介效应分析”，中介分析前一般不会加上“两水平被试内”类似的字样，请作者规范表述。

回应：感谢审稿专家的宝贵意见！我们完全同意您对统计方法表述规范性的重要指正。我们已删去了“两水平被试内”的相关表达。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（14）实验一 2.3，“95%置信区间 CI”，请作者统一前后文表述方式。

回应：感谢审稿专家的宝贵意见！我们完全同意您对统计表述一致性的细致指正。我们已将全文置信区间表述统一规范为“95%置信区间 CI”。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（15）实验二 3.2.2“两组之间的面孔吸引力评分不存在显著差异，面孔可信度评分存在显著差异”，想强调差异是否显著就在后附上 p 值，如果继续沿用实验一的实验材料就无需写这句话。

回应：感谢审稿专家的宝贵意见！我们完全同意您对材料描述严谨性的重要建议。鉴于实验 2 继续沿用了实验 1 的实验材料，我们已删除原材料说明的重复描述部分，将剩余材料描述部分与下述“社会价值取向的测量”部分进行了整合。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

2. 论文的写作与逻辑方面

（1）引言第二段，“社会价值取向(social value orientation, SVO)作为一种稳定的人格特质，通过调节个体对自身和他人利益分配结果的稳定偏好来影响其信任行为(刘长江，郝芳，2011)”，希望作者可以对每个变量都下个详细的定义，再介绍其与本文中其他变量或概念的联系，这样可以方便读者更轻松地阅读；此外，建议对每个变量的介绍都另起一段，后文同。

回应：感谢审稿专家的宝贵意见！我们完全同意您对变量定义详细性的重要建议。针对文中各变量的定义表述问题，我们已进行如下系统修改：

1. **信任(修改稿引言段落 1)**：将原表述修改为“信任可以被定义为信任者通过积极预期被信任者的意图或行为，在不确定的社会情境中委托对方处置自身资源，并自愿承担背叛风险的决策过程(Krueger & Meyer-Lindenberg, 2019)。”，以明确该内容是对信任变量的正式定义。

2. **社会价值取向(修改稿引言段落 2)**：将原表述修改为“社会价值取向(social value orientation, SVO)是一种稳定的人格特质，反映了个体在资源分配决策中表现出的稳定偏好(刘长江，郝芳, 2011)。社会价值取向可分为亲社会取向和亲自我取向。亲社会取向个体(即亲社会者)倾向于追求自身和他人的共同利益，而亲自我取向个体(即亲自我者)则倾向于追求自身利益的最大化(刘长江，郝芳, 2011)。”。这一修改对社会价值取向变量进行了详细定义，并进一步定

义了社会价值取向的两个亚类。考虑到面孔可信度相关内容与社会价值取向相关内容在字数上均难以单独成段，因此仍将两者共置于同一段落。

（3）**互惠预期(修改稿引言段落 4)**：在段落开头新增了明确定义："互惠预期是个体对他人是否遵守互惠规范的积极心理预期(Krueger et al., 2008; Dunning et al., 2014; Krueger et al., 2020)。"。具体修改思路详见对审稿人 1 的意见 1 的详细回复。

（4）**信任学习与回馈概率(修改稿引言段落 5)**：我们调整了段落结构，将信任学习定义置于段落开头："信任学习指个体根据他人行为学习其互惠意图，进而调整后续信任的过程。"，并在段落中间明确了回馈概率的定义："回馈概率指在重复信任互动中，个体观察到的被信任者实际选择回馈行为的相对频率(Chang et al., 2010)。"。回馈概率定义的具体思路详见对审稿人 1 的意见 1 的详细回复。考虑到回馈概率概念需要结合信任学习概念来阐述，因此仍将两者共置于同一段落。

（5）**学习率、面孔可信度影响系数与衰减系数(修改稿引言段落 8)**：修改稿在引言部分新增了段落 7-10 来专门论述计算模型假设 4 的理论基础，对这些计算模型参数的定义补充在其中的段落 7-8 中。我们同时也在附录"计算模型构建具体细节"部分同步补充了这些参数的相应说明。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

刘长江, 郝芳. (2011). 不对称社会困境中社会价值取向对合作的影响. *心理学报, 43*(04), 432−441.

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87−105.

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology, 107*(1), 122–141.

Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences, 42*, 92−101.

Krueger, F., Bellucci, G., Xu, P., & Feng, C. (2020). The critical role of the right dorsal and ventral anterior insula in reciprocity: Evidence from the trust and ultimatum games. *Frontiers in Human Neuroscience, 14*, 176.

Krueger, J. I., Massey, A. L., & DiDonato, T. E. (2008). A matter of trust: From social preferences to the strategic

adherence to social norms. *Negotiation and Conflict Management Research, 1*(1), 31−52.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

（2）引言第三段，“现有研究多独立考察两者对信任的影响”，希望作者可以统一表述方式，上文对于信任的表述有“信任行为、信任决策、信任”等，不便于读者理解，后文同。

回应：感谢审稿专家的宝贵意见！我们完全同意您对术语统一性的重要建议。关于“信任”相关概念的表述问题，我们统一规范“信任”为“信任行为、信任决策”的标准表述。仅只在统计分析部分时将被试在任务中的投资金额称为“信任行为”。我们已对全文进行了系统检查，确保所有相关表述均采用这一标准术语。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（3）引言第三段、第四段，希望作者可以对所引用的理论进行简要的介绍后再介绍其与本研究的联系，期刊面向的是广大读者，并非某一个领域的专家。此外，请作者注意，学术论文中的理论不建议随意引用，需要与本文的假设、逻辑思路相契合。若在前言部分引用，最好还需要在实验、结论部分加以讨论其与本文的逻辑关系，否则只会造就理论过于臃肿的引言部分的头重脚轻的讨论部分，后文同。

回应：感谢审稿专家的宝贵意见！我们完全认同理论框架应当服务于研究问题，并确保全文逻辑的一致性。针对您的建议，我们进行以下重要修改：

1. 对引言第 3-4 段的修改：

在修改稿引言第 3 段中，我们将所引用的理论的原表述修改为：“人情互动理论(Magnusson & Stattin, 1998)提出，环境线索(如面孔可信度)与人格特质(如社会价值取向)可以共同影响个体的行为决策。信任发展的二元互动模型(Simpson, 2007)也认为，影响信任的因素既包括常规变量(如面孔可信度带来的初始信任)，也包括个体差异变量(如社会价值取向)。”。这一修改加入了对两个理论的简要介绍。

在修改稿引言第 4 段，我们新增了：“信任的经典理论认为，互惠预期对信任具有关键作用。”。这句话简单介绍了社会规范理论与道德规范理论在信任领域中的地位及其与互惠预期的联系。

2. 对讨论部分的补充：

我们在讨论部分补充了原文中忽视的对人情互动理论与信任发展的二元理论的持续讨论。研

究发现面孔可信度与社会价值取向在静态单次信任中没有共同作用，但是在动态信任学习过程中却表现出显著的共同作用。这表明两个理论强调的外部线索因素与个体差异因素的共同作用需要在互动发展的信任学习过程中实现。

我们在修改稿讨论部分第 1 段新增了以下内容："与单次信任博弈的结果不同，研究发现社会价值取向调节初始面孔可信度对信任学习的影响。这说明面孔可信度与社会价值取向的共同作用需要在互动发展的信任学习过程中实现。这也表明人情互动理论(Magnusson & Stattin, 1998)和信任发展的二元理论(Simpson, 2007)所强调的外部线索因素与个体差异因素对行为决策的共同作用，尤其适用于动态交互情景。"

这些修改既考虑了广大读者的需要，也平衡了文章结构。我们已对文中引用的所有理论(包括引言新增段落 7-10 中涉及的相应理论)进行了检查，确保在讨论部分均进行了相应的讨论。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Magnusson, D., & Stattin, H. (1998). Person-context interaction theories. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (5th ed., pp. 685–759). John Wiley & Sons, Inc.

Simpson, J. A. (2007). Psychological Foundations of Trust. *Current Directions in Psychological Science, 16*, 264−268.

（4）引言第五段，"个体会根据回馈概率更新初始可信度"，希望作者对此类新出现的概念做概念的定义。

回应：感谢审稿专家的宝贵意见！感谢审稿专家对回馈概率概念定义不足之处的指正。我们已根据您的建议，在相应段落(详见对"论文的写作与逻辑方面"的意见 1 的回应)中补充了对回馈概率的详细定义，以确保概念的清晰性和完整性。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（5）引言第八段，"H6 面孔可信度和回馈概率对信任学习的影响表现在计算模型的参数上。H7 不同社会价值取向的信任学习过程存在显著差异，表现在个体计算模型的选择和参数上"请查阅并学习相关文献后用更加规范、学术的表述撰写论文。

回应：感谢审稿专家的宝贵意见！我们已认真查阅并学习相关文献，对研究假设进行了重新整合与修改，使其表述更加规范，符合学术要求。具体修改内容详见对审稿人 1 意见 3 的详细回复。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（6）实验一 2.2.1，“根据 G-power 软件的计算($\alpha = 0.05$, power = 0.80)，对于效应量 ($f = 0.25$)，所需样本量为 98 人”，请作者多参考本期刊或其他期刊已发表文献对样本量计算部分的表述进行修改。

回应：感谢审稿专家的宝贵意见！我们已参考本期刊及领域内已发表文献的规范，对样本量计算部分的表述进行了修改和完善。具体修改内容详见对审稿人 1 意见 6 的详细回复。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（7）实验一 2.2.1，“研究招募了 105 名(女性 82 名)大学生参与实验。”请作者阐明男女被试数量平衡较差的原因。

回应：感谢审稿专家的宝贵意见！针对这一情况，我们已在实验 1 与实验 2(重新采用线性混合效应模型方法)的数据分析中将性别作为固定效应纳入分析。结果显示，性别因素既不影响信任，也不影响实验效应，具体分析结果详见对审稿人 1 意见 4 的回应。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（8）实验一 2.2.2，“研究选取了 69 张黑色短发中国男性中性情绪面孔照片”，请作者阐明情绪面孔照片来源，以及为何不对情绪面孔的性别进行平衡？

回应：感谢审稿专家的宝贵意见！关于情绪面孔照片来源，以及为何不对情绪面孔的性别进行平衡，我们说明如下：

1. 情绪面孔照片来源说明：

本研究使用的 69 张中国男性中性情绪面孔照片均来自 Ma 等(2020)开发的面部表情数据库。所有照片均使用 FaceGen Modeller 3.0 软件进行了统一标准的图像处理。我们也在论文相应部分补充了情绪面孔照片来源信息。

2. 面孔照片性别单一化的设计考量：

Wirth 与 Wentura(2023)的研究发现个体会认为女性面孔比男性面孔更值得信任，因此研究的实验材料专门选用单一男性性别情绪面孔照片来控制面孔本身的性别因素对信任的潜在影

响。我们理解性别平衡的重要性，这一材料选择主要是考虑特定研究需求。未来研究可在本研究基础上进一步专门探讨性别因素的影响。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Ma, J., Yang, B., Luo, R., & Ding, X. (2020). Development of a facial-expression database of Chinese Han, Hui and Tibetan people. *International Journal of Psychology, 55*(3), 456−464.

Wirth, B. E., & Wentura, D. (2023). Not lie detection but stereotypes: Response priming reveals a gender bias in facial trustworthiness evaluations. *Journal of experimental social psychology, 104*, 104406.

（9）实验一 2.2.3，“1—10 表示平分可能性为 10%—100%(见图 1a)”，前文提到被试可以选择保留全部资金，意思是被试选择平分奖金的概率也可以为 0%，那么此处的评分设置是否有误？

回应：感谢审稿专家的宝贵意见！感谢您对实验评分设置的细致审阅。

关于评分设置的问题，我们说明如下：由于实验采用数字键进行反应(1-9 键)，而技术上难以将 0 键等同于 10%来记录(键盘数字 0 通常位于 9 之后)，因此当前研究将平分奖金的概率范围限定在 10%-100%。未来研究可以在此研究的基础上思考如何合理利用键盘进行相应的按键反应。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（10）实验一 2.2.3，“被试选择后，仅呈现被信任者面孔图片 300ms；随后呈现面孔图片和吸引力评价提示，被试需评定对方面孔的吸引力(见图 1b)”，请作者阐述在实验材料评定部分已经检验了面孔吸引力的差异不显著么，此处实验部分做面孔吸引力评价的意义是什么？

回应：感谢审稿专家的宝贵意见！感谢您对实验任务环节的细致审阅，关于吸引力评定的实验设置意义，我们作如下说明：

虽然预实验已确认高、低可信面孔组的吸引力评分无显著差异，但基于 Todorov 等人(2015)关于面孔吸引力与面孔可信度的统计控制的建议，我们认为即使预实验控制了面孔吸引力，在正式实验过程中仍需持续监测该变量。因此，我们在正式实验的面孔评价任务中保留了吸引力评价环节，并在数据分析阶段采用了基于面孔项目的分析方法。结果显示，面孔吸引力

对面孔可信度评价没有显著影响，这验证了面孔吸引力在本研究得到有效控制。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual review of psychology, 66*, 519−545.

（11）实验一 2.3，“[0.04, 0.38]不包括 0”，“不包括 0”是作者判断的依据，不用写到文章里，后文同。

回应：感谢审稿专家的宝贵意见！我们完全接受您的观点，并删除了文章置信区间部分“不包括 0”的解释，再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（12）实验二 3.2.1，“研究招募了 101 名(女性 79 名)”，请作者阐述为何招募被试的人数远超计算得出的所需样本量，并说明多出的被试量是否对研究结果造成了影响。

回应：感谢审稿专家的宝贵意见！我们已参考本期刊及领域内已发表文献的规范，对样本量计算部分的表述进行了修改和完善。重新计算得出的所需样本量(96 人)与研究的实际样本量(101 人)差异不大。具体修改内容详见对审稿人 1 的意见 6 的详细回复。再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

（13）实验二 3.2.1，“回馈概率: 低回馈概率: 20% vs 高回馈概率: 80%”，请作者阐述对于“高低概率”界定的依据，为何不选取 25%和 75%（0-50 的中点和 50-100 的中点）？

回应：感谢审稿专家对这一重要方法细节的关注！关于高低回馈概率的设定依据，我们做出如下说明：

首先，这一概率设定直接参照了 Chang 等(2010)与 Fouragnan 等(2013)关于信任学习的经典研究。我们主要参考了 Chang 等的研究，因为我们的回馈概率概念主要借鉴于 Chang 等提出的“经验可信度水平”概念，即被信任者回报信任的可能性。因此，我们在操作定义上也严格遵循了 Chang 等对经验可信度水平的操作定义，将 20%回馈概率界定为低回馈，80%回馈概率界定为高回馈。

Participants played a repeated Trust Game in the role of Player A as described in the introduction. We employed a 2 × 2 within-subjects design in which partner's level of facial trustworthiness (high or low) was crossed with partner's level of experienced trustworthiness (high or low). Each Player B represented one of the experimental conditions. Level of facial trustworthiness was assessed via independent ratings of partner photographs (see below). We defined level of experienced trustworthiness as a high (80%) or low (20%) probability of reciprocating an offer. Money invested by the participant was multiplied by a factor of 4. If an offer was reciprocated, Player B always reciprocated 50% of the total multiplied amount sent by Player A. When trust was not reciprocated, Player B did not return any of the multiplied amount of money back to Player A. Participants played 15 randomly ordered interspersed rounds with each partner (60 trials total, plus 30 slot machine gambles). On each trial, participants were endowed with $10. Each trial lasted 16 s and began with a short fixation cross (1000 ms) followed by a picture of the partner (3500 ms). Participants then decided how much money they wanted to invest by scrolling through offers that randomly increased or decreased in $1 increments.

其次，20%低回馈和 80%高回馈的设置能够确保高低回馈概率组间保持足够的行为区分度(差异达 60%)，这比 25%低回馈和 75%高回馈的 50%差异更能有效避免中等概率范围可能导致的决策模糊性，从而更准确地模拟真实社会情景中的“高度可信”与“低度可信”行为。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87−105.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience, 33*(8), 3602−3611.

（14）实验二 3.2.1，“测量因变量为信任行为(被试选择的投资金额)和计算模型参数”，以计算模型参数作为因变量的表述过于稀奇，请作者查阅文献后再行更正。

回应：感谢审稿专家的宝贵意见！我们完全同意您对因变量表述严谨性的重要指正。我们已查阅计算模型相关文献(Jin et al., 2023)，对该部分表述进行如下修正：

1. 删除将计算模型参数作为研究因变量的不准确表述，保留表述“测量因变量为重复信任博弈任务中的信任行为(被试选择的投资金额)”。

2. 关于对计算模型参数进行方差分析的表述问题，我们参照了 Jin 等(2023)的表述方式。我们认为原文中：“以社会价值取向、面孔可信度—回馈概率一致性和增益损失情境(被信任者的平分/保留资金)为自变量，对回馈概率学习率进行重复测量方差分析”的表述方式已符合表述规范，故不再做额外修改。

Between-group comparison of parameters

We conducted mixed ANOVAs to examine the main effect and interaction of reputation disclosure and subject type on each parameter in the optimal candidate model.

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Jin, Y., Gao, Q., Wang, Y., Dietz, M., Xiao, L., Cai, Y., Bliksted, V., & Zhou, Y. (2023). Impaired social learning in patients with major depressive disorder revealed by a reinforcement learning model. *International journal of clinical and health psychology, 23*(4), 100389.

（15）实验二 3.2.2，“被试先进行三优势测量，后进行社会价值取向滑块测验”，请作者阐述为强调何使用这个测量顺序而不平衡测量顺序？

回应：感谢审稿专家的宝贵意见！感谢您对测量顺序设计的关注。关于社会价值取向测量顺序的设计，我们说明如下：

在理论依据上，Pletzer 等人(2018)的元分析指出三优势测量更适合快速分类，社会价值取向滑块测量更适合需要连续数据的场景，进而可以弥补三优势测量的局限性。Murphy 和 Ackermann(2014)则提出可以将三优势测量作为初步筛选工具，用于快速识别具有明确社会价值取向的个体，而社会价值取向滑块测验则用于提供更精细的连续型测量。综合上述文献，我们支持先采用三优势测量来快速筛选出明确的取向类型，之后再采用社会价值取向滑块测量来量化个体的社会价值取向倾向强度。

在方法论上，我们认为该顺序设计有三大优势：首先，这种测量顺序有助于避免疲劳效应，因为三优势测量比滑块测验相对来讲更易完成。其次，这种测量顺序可以防止锚定偏差，因为若先进行需精细数值选择的社会价值取向滑块测验，可能影响后续三优势测量的自然反应。最后，这种测量顺序是数据质量控制的需要，我们可以通过三优势测量及时排除社会价值取向不明者，从而提高数据有效性。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

参考文献：

Murphy, R. O., & Ackermann, K. A. (2014). Social Value Orientation: Theoretical and Measurement Issues in the

Study of Social Preferences. *Personality and Social Psychology Review, 18*(1), 13−41.

Pletzer, J. L., Balliet, D., Joireman, J., Kuhlman, D. M., Voelpel, S. C., & Van Lange, P. A. (2018). Social value orientation, expectations, and cooperation in social dilemmas: A meta–analysis. *European Journal of Personality, 32*(1), 62−83.

（16）实验二 3.2.3，“单个试次的流程如下：首先，呈现注视点“+”1000 ms，提示被试新试次投资开始。随后呈现被信任者图片 2000ms。接着呈现投资界面，投资金额即绿色方框的数量，初始投资金额为 1，被试需按左、右键增减投资金额，按回车键提交投资金额，如果超出 6000ms 反应时间，则强制提交当前界面投资方案，但本轮奖励金为 0”请作者阐述实验操作方式为何与实验一不同？

回应：感谢审稿专家的宝贵意见！感谢您对实验程序操作差异的细致审阅。关于两个实验采用的操作方式为何不同，我们作如下说明：

首先，实验 1 采用单轮信任博弈范式，实验 2 采用重复信任博弈范式，两者在核心操作方式上保持一致，包括资金给予和分配方式等关键要素。操作方式的差异主要体现在两个层面：

1. 操作形式差异：实验 1 采用直接输入金额的方式，这种静态输入方式与单次交互的特性相匹配，能够确保被试决策的一致性。实验 2 采用左右键增减金额的动态调整方式，这种设计是为了更好地体现重复博弈中信任动态调整的过程。未来研究可以考虑开发更统一的交互界面来协调这两种操作需求。

2. 反馈呈现差异：实验 2 由于重复信任博弈的范式要求，必须呈现结果反馈以支持信任学习过程。这种反馈机制是重复博弈范式的组成部分，与单次博弈形成质的区别。这种差异源于两种博弈范式的本质区别，而非设计上的随意变更。

我们理解这些操作差异可能会带来一些比较上的困难，但它们各自都很好地服务于对应实验范式的核心目标。未来研究可以探索更统一的实验操作方案。

再次感谢您的意见。若您认为仍需进一步调整，我们非常乐意继续完善。

3. **论文的写作态度方面**

（1）引言第八段，“社会价值取向正向预测信任行为”，作为假设二的“H2”呢，希望作者在

完稿后自行检查文章疏漏。

回应：感谢您指出这一重要疏漏！我们已对文章的假设进行了重新整合与修改。我们诚挚为这一疏漏致歉，这确实是不应出现的错误。衷心感谢您严谨细致的审阅，这对提升论文质量至关重要！

（2）实验一 2.2.1，段末“，。”，希望作者在完稿后自行检查文章疏漏。

回应：感谢您指出这一标点符号的疏漏！我们已立即进行了修正。衷心感谢您严谨细致的审阅。

（3）实验二 3.2.3 第一句，“包含先重复信任博弈任务”。

回应：感谢您指出这一表述不完整的问题！我们已对实验 2 第 3.2.3 节首句进行了修正。衷心感谢您严谨细致的审阅。

（4）实验二 3.3.2，“模型构建具体细节见附录”，补充材料细节不足。

回应：感谢审稿专家对补充材料细节的关注！我们已在附录中对计算模型相关细节进行了全面补充和完善。具体修改内容详见对审稿人 1 意见 9 的详细回复。衷心感谢您严谨细致的审阅。

（5）参考文献部分，请作者阐述为何在参考文献部分，会出现 doi 号时有时无的现象。

回应：感谢您对参考文献格式严谨细致的审阅！我们完全理解您对文献著录规范性的严格要求。关于参考文献中 doi 号的标注情况，现说明如下：

在文献列表中，我们仅对下述三篇电子刊论文标注了 doi 号：

Golec-Staskiewicz, K., Wojciechowski, J., Haman, M., Wolak, T., Wysocka, J., & Pluta, A. (2024). Unveiling the neural dynamics of the theory of mind: a fMRI study on belief processing phases. *Social cognitive and affective neuroscience, 19*(1). nsae095. https://doi.org/10.1093/scan/nsae095

Ma, Q. G., Meng, L., & Shen, Q. (2015). You Have My Word: Reciprocity Expectation Modulates Feedback-Related Negativity in the Trust Game. *PLOS ONE, 10*(2), e0119129. doi: 10.1371/journal.pone.0119129.

Radell, M. L., Sanchez, R., Weinflash, N., & Myers, C. E. (2016). The personality trait of behavioral inhibition

modulates perceptions of moral character and performance during the trust game: behavioral results and computational modeling. *PeerJ, 4*, e1631. doi: 10.7717/peerj.1631.

因为这些论文都是只有文章编号而无页码的在线发表论文。我们严格遵循《心理科学进展》采用的 Publication Manual of the American Psychological Association (2019)第 7 版格式规范要求，着重参考了其中关于电子刊论文著录的具体规定，对这些文献保留了 doi 信息。

7. 只有论文编号而无页码的电子刊论文

Burin, D., Kilteni, K., Rabuffetti, M., Slater, M., & Pia, L. (2019). Body ownership increases the interference between observed and executed movements. *PLOS ONE, 14*(1), Article e0209899. https://doi.org/10.1371/journal.pone.0209899

同时，考虑到期刊的版面篇幅要求，我们对其他有明确页码的印刷版论文均未标注 doi 号。我们理解参考文献格式的规范性对学术出版的重要性，并对所有文献的著录格式进行了全面复查，确保每条参考文献都符合最新的 APA 格式要求。

【注意：最新 APA 出版手册要求前 20 位著者都写上，如果超过 21 人，就写前 19 位和最后一位，中间用省略号隔开。我们觉得这个改动并不好，一是心理学论文中这么多著者的情况很少见，二是也太占版面。反正大家都是用文献管理器，如果是按照 APA 的新要求，我们也默认是正确的。包括 APA 还要求文献后要有 DOI，我们只当作可选项，不做硬性规定】

再次感谢您严谨细致的审阅，如果您认为我们对于电子刊论文的 doi 号标注处理仍需进一步调整，我们非常乐意继续完善。