《心理科学进展》审稿意见与作者回应

题目：人机信任校准的双途径：信任抑制与信任提升
作者：黄心语，李晔

第一轮

审稿人 1 意见：

该研究针对人机信任校准的问题，通过文献综述的方法，对人机信任的类型和信任校准的方法进行了总结和讨论。文章填补了国内外缺乏信任校准综述文章的空缺，主题符合《心理科学进展》范围。研究所引用的文献相对较新，文章结构清晰、可读性好。然而，文章某些论述不够全面，有些部分缺乏相关研究的引用。如果作者可以作出合理的回应或者修改，我将建议文章被录用。

意见 1：前言部分“agent”翻译成“代理”比较奇怪。翻译成“智能体”也许比较合适。
回应：将“agent”翻译为“代理”是参照前人的译法(例如高在峰 等，2021)，但“智能体”确实更好地表达了 agent 在此处的含义并符合汉语习惯，因此我们将全文中 “agent”一词的翻译全部替换为“智能体”。感谢审稿老师的指教！

意见 2：第二自然段，“但是与自动化、计算相比，人类对机器人的信任可能会有所不同(Salem et al., 2015)”中“计算”应为“计算机”。
回应：结合专家提出的第 3 小问，我们删除了该语句以避免引起误解。对于我们的粗心深感抱歉，并已经通读全文以避免此类似失误。

意见 3：算法和人工智能所包含的范围非常广。文章的一个隐含假设是，在人机信任领域，机器人和自动化、计算机的差异，大于机器人和算法、人工智能的差异(因此文章聚焦机器人、算法和人工智能三者)。但这个隐含假设不一定成立。另外，文章的一些论据也引用了自动化领域的研究(如，McGuirl & Sarter，2006；Parasuraman & Riley, 1997)。
回应：很抱歉由于我们表达的问题引起误解和歧义。这种隐含假设确实不一定成立，本文也并非想表达这种观点。

首先，本文聚焦于对机器人、算法和人工智能的信任主要出于两方面的考虑。一方面，本文欲探讨的信任对象应具有一定的能动性，即能与个体进行交互的、具有一定自主性的智能体。因此，我们选择了研究对象为机器人、算法和人工智能的文献，并以涉及人与机器人交互的研究为主。另一方面，机器人、算法和人工智能的研究也密不可分，常常互相作为论证观点的论据而存在。例如在 de Visser 等人(2020)的研究中，将人-机团队(Human-robot teams)定义为至少由一名人类和一个机器人、智能体、或者其他人工智能、自动化系统所构成的团队。人工智能的主要载体就包括机器人，而算法是人工智能的底层运行机制(喻丰, 许丽颖, 2020)，人工智能是在算法的基础之上搭建、运行并生成的。相对于自动化和计算机而言，人们对于这三类更自主、更拟人、不仅仅是作为工具，而且会作为互动伙伴的智能体的交互模式可能更具有相似性。

其次，我们无意假设在人机信任领域，机器人和自动化、计算机的差异大于机器人和算法、人工智能的差异，也非常赞同审稿专家对此的看法。我们的主要研究目的是想了解在目

前这个人工智能快速发展的时代，人们与具有社会性的智能体之间的交互模式与特点，而不是关注以往常常作为人类辅助工具的自动化与计算机。所以本文的论据是以机器人为主，算法与人工智能仅仅引用了部分文献以佐证文中的观点。鉴于审稿专家的疑问，我们删除了可能引起误解的表述。在目前的修改稿中，此处改为“本文主要关注人与智能机器人、算法、人工智能之间的交互和信任，且以人与智能机器人之间的交互为主”。

另外，在文章引用的论据方面，部分论据涉及到自动化是因为影响人-机器人信任与人-自动化信任的原因有部分重合。从某种意义上讲，影响人机信任的首要因素可以被视为是影响人-自动化信任因素的延伸(Khavas, 2021)，引用部分人-自动化交互的有关研究既可以弥补某些人-机器人交互研究中的缺失，又可以从另一方面丰富人-机器人交互研究，间接支持文章中的某些观点。鉴于审稿人的疑问，我们在引用文献时进行了必要解释，以明确研究对象和结果的解释范围。如在“3 人机信任校准的途径”透明度策略中，我们引用一篇文献以证明信心指数的显示有助于校准信任时，标注了该文章的研究对象为自动化系统：“McGuirl 和 Sarter(2006)发现，如果能提供动态更新的系统信心指数将有助于飞行员在任务分配和是否遵守自动化系统的建议等方面做出更好的决策，对系统准确性的估计也更加精准”。

参考文献：

喻丰，许丽颖. (2020). 人工智能之拟人化. *西北师大学报(社会科学版), 57*(5), 52–60.

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics, 12*(2), 459−478.

Khavas, Z. R. (2021). A review on trust in human-robot interaction. *arXiv preprint arXiv:2105.10045*.

意见 4：人机信任偏差部分“信任偏差会导致个体信任比人类更不可靠的算法，或不信任比人类更可靠的算法(Dzindolet et al., 2003)”。引用的文献没有出现在参考文献列表里。

回应：非常感谢审稿专家的细心指正，我们已经在参考文献列表中添加了该条文献，并核对了全文的文献引用，力求避免类似失误。

参考文献：

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies, 58*(6), 697−718.

意见 5:“而信任偏差的出现往往又与个体本身、机器人、情境三因素有关。”应该出自 Hancock et al., 2011。

回应：感谢审稿专家的提醒，我们已经在该表述后面插入了所引用的文献。

意见 6：在本部分中，文章列举了过度信任和信任不足的原因。但是，里面部分原因既可以导致过度信任、也可以导致信任不足，把他们分开列到过度信任和信任不足两个标题下，有误导性。比如机器的可靠性，人倾向于过度信任可靠性较高的机器，而对可靠性较低的机器信任不足[Shi, Yuan, Nicolò Azzolin, Andrea Picardi, Tao Zhu, Monica Bordegoni, and Giandomenico Caruso. "A Virtual Reality-based Platform to Validate HMI Design for Increasing User’s Trust in Autonomous Vehicle." Computer-Aided Design and Applications 18, no. 3 (2020): 502-518.]。

回应：非常感谢审稿专家的宝贵意见，文章原来的结构和表述的确容易产生误导。当时将过度信任与信任不足的原因分开探讨是希望文章结构更加清晰明了。在梳理资料的过程中也发现有些原因既可能导致过度信任，也可能导致信任不足。为更好地统合人机信任偏差之因，我们根据 Hancock 等人(2011，2021)梳理的人机信任原因重新调整了文章的逻辑和结构，将

造成两类偏差的原因都归入“2 人机信任偏差”，再分别从机器人、个体本身、情境三方面来阐述各种因素对两类人机信任偏差的影响。审稿老师提到的关于自动驾驶汽车交互的文献我们已经作为人机信任偏差之因中有关机器人性能因素的论证补充。感谢老师帮助我们完善文献资料！

意见 7：人机信任校准部分“透明化设计”或“可解释性”为什么放在“降低个体过高的初始信任水平”部分而不是“降低个体交互过程中过高的信任水平”？多数主流的透明化设计理论，都关注在交互过程中通过透明化来进行实时的信任校准(如，[Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical issues in ergonomics science, 19(3), 259-282.])。

回应：将“透明化设计”或“可解释性”放在“降低个体过高的初始信任水平”是因为我们认为既然“黑匣子” 属性是个体出现信任偏差的主要原因之一，那么在人机交互之初就将“黑匣子”打开或许能从一开始就降低个体过高的信任水平，因为“预防大于治疗”。如果在交互过程中再逐渐将机器人透明化，校准效果或许没有一开始就透明的效果好。当然，透明化和可解释在交互过程中同样能进行信任校准，因此我们纠正了初稿的观点，将“透明度提升”这一校准方法纳入“与机器人有关的信任校准策略”，同时阐述透明度中的可解释性和可理解性对信任提升与信任抑制的校准效果，以期更准确地表述这两种方法在信任不同阶段的作用。另外感谢审稿老师提供的文献资料！我们结合该文章的内容已将其作为透明度的可理解性方面的论据补充。

意见 8：透明化设计或可理解性不应该只放在“信任抑制”部分，因为那是一个信任校准的技术，它也可以用来提升信任(如，[Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., ... & Shively, R. (2017). Shaping trust through transparent design: theoretical and experimental guidelines. In Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016 (pp. 127-136). Springer International Publishing.])。这再次涉及了文章逻辑的一个痛点。按照信任过度和信任不足、信任抑制和信任提升来编排章节，可能让许多因素和干预方法看起来是针对信任过度或信任不足的，然而其实这些方法是信任校准的通用的方法。希望作者可以重点修改这个逻辑痛点。

回应：感谢审稿专家的中肯意见和建议。在撰写初稿时遗漏了有关于透明度提升人机任的相关研究梳理。感谢审稿老师有关文献资料的提供，我们已经重新调整文中有关于透明度策略适用范围的表述。针对文章的结构，我们已重新调整了文章结构以统合两类人机信任偏差以及校准的研究。

意见 9：未来展望部分，已有研究采用了内隐测量来测量信任 [Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don’t know why: Effects of implicit attitudes toward automation on trust in an automated system. Human factors, 55(3), 520-534.]。

回应：虽然该研究采用内隐的方法测量自动化信任，但是仅集中于个体信任校准之前对自动化/机器人的信任测量，并未涉及到个体在人机交互中出现了信任偏差并进行信任校准之后的后续内隐信任测量。我们认为校准信任之后不仅要关注个体的外显信任态度，同时也要关注个体的内隐信任态度以更好地检验校准策略的有效性与实用性。结合审稿专家的意见，我们对相应文本进行了修改，希望能清晰地表达上述观点。

意见 10：已有研究初步从认知神经的视角研究人机信任 [Eloy, L., Doherty, E. J., Spencer, C. A., Bobko, P., & Hirshfield, L. (2022). Using fNIRS to identify transparency-and reliability-sensitive markers of trust across multiple timescales in collaborative human-human-agent triads. Frontiers in Neuroergonomics, 3, 838625.]。

回应：感谢审稿专家老师的宝贵意见。我们在阅读相关文献时也发现已有研究者开始从认知神经的视角去研究人机信任及其偏差。但是我们发现这些研究或多或少仍是集中与考察个体在一般情境下的人机信任和信任偏差时出现的认知神经活动，未涉及信任校准中和校准后个体的认知神经变化。我们认为这一部分恰恰对于人机信任校准极为重要，不仅可以弥补人机信任校准后期阶段认知神经研究的缺失，同样也可以为后续的信任校准策略的优化提供思路与指导。结合审稿专家的意见，我们已经重新修改该部分的表述。

意见 11：已有研究初步研究群体过程对人机信任的影响 [Xu, J., & Montague, E. (2013, September). Group polarization of trust in technology. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 57, No. 1, pp. 344-348). Sage CA: Los Angeles, CA: SAGE Publications.； Montague, E., & Xu, J. (2012). Understanding active and passive users: The effects of an active user using normal, hard and unreliable technologies on user assessment of trust in technology and co-user. Applied ergonomics, 43(4), 702-712.; Montague, E., Xu, J., & Chiou, E. (2014). Shared experiences of technology and trust: An experimental study of physiological compliance between active and passive users in technology-mediated collaborative encounters. IEEE Transactions on Human-Machine Systems, 44(5), 614-624.]。

回应：感谢审稿专家指正并提供详细资料！我们在搜集文献的时候也发现了已经有研究者开始转向于关注个体在群体之中，而不是单独与机器人交互时信任水平的变化发展，例如有研究发现人机群体交互会出现群体极化。在阅读以上文献后，我们补充了相关研究内容。另一方面，本文的研究主题是信任校准，我们关注的不仅仅是为什么群体之间会出现“极化”信任水平，同时更关注怎么样去校准与改善群体中个体的“极化”信任水平。结合审稿专家的意见，我们已经重新修改该部分的表述。

……

审稿人 2 意见：

文章综述了人与机器人交互中的两类信任偏差以及如何提升信任或者抑制信任，并且提出测量工具选取的不一致，以及机器人的类型多样性导致结论不一致。但是在综述信任偏差以及提出信任校准的内容几乎没有提到这两点，而在未来研究展望中才提到要优化测量工具与方法，以及要深入探讨信任校准的认知神经过程。应该照作者自己所提出的，在综述信任偏差时就要包括测量工具以及机器人的信息，或者在第二和第三部分增加表格比较测量工具和机器人，才能解释文献结果的不一致，并且提出未来研究展望。

回应：感谢审稿专家的宝贵建议。在梳理人机信任偏差有关研究时我们确实忽略了对测量工具及机器人有关信息的说明，因此在未来研究展望中该部分的提出显得有些突兀。根据专家的建议，我们已经在“3 人机信任校准的途径 1 前言”部分增添表格以对比各项研究中使用的机器人类型以及相关测量方法。

第二轮

审稿人 1 意见：

总体来说，作者对审稿意见回复内容详实，对原来文章的修改合理得当，较好地解决了

我在之前审稿意见中表达的关切。修改后的文章结构也更加清晰合理、可读性更好。但对于新调整的内容，我还有以下建议希望作者考虑：

意见 1：修改说明文档第 14 页 10 行：“首次全面地”的表述建议修改。首先，已有不少中文综述了梳理了人机信任这一主题；其次，本文在梳理信任偏差原因时还有一些重要因素并未综述。比如，在机器人方面，有自动化水平、交互模态等因素都并未提及；在个体方面，个体的心理状态、经验等因素的影响也都并未提及；
回应：感谢审稿专家老师的细心提醒，我们已经修改了原文中的表述以期更好、更准确地反映文章的内容。

意见 2：修改说明文档第 15 页 1-5 行：“低估算法能力”应是信任不足导致的直接后果，然后才是“不能很好地利用算法，也无法享受使用算法所带来的好处”，建议调整表述顺序并重新思考过度信任危害的行文逻辑。
回应：感谢审稿专家老师的耐心提醒，我们在撰写文章时确实忽略了信任偏差危害的内在逻辑。鉴于此，我们已经重新检查了“2.1 人机信任偏差的危害”部分的相关语句，并重新组织语句以更好地符合逻辑。

意见 3：修改说明文档第 15 页“机器人的可靠性与稳定性”部分：本段论述机器人的可靠性和稳定性对信任的影响，建议一开始给此处的可靠性和稳定性一个简明的定义或说明，并最好能交代它们与后续提及的可预测性和错误的关系，以使整个段落逻辑更清晰。
回应：感谢审稿专家老师的宝贵意见！我们在撰写文章的过程中确实忽略了对于新概念定义的介绍，结合专家后续提出的第 5、6、7 小问，为提高文章的可读性，我们已经在每个部分新概念的提出后加入了该概念的定义以更好地帮助读者理解。

意见 4：修改说明文档第 15 页 22 行：“信任违背是指一方因某些行为降低了另一方的信任 ”此句有歧义，建议修改表述使其意思更准确。另外，此处引出来信任违背的概念，但似乎并未论述错误如何导致的信任违背。
回应：感谢审稿专家老师的细心提醒。该信任违背的定义是我们根据英文翻译的，所以可能在翻译的过程中未能清楚表达原文含义。我们已经参考心理学中文期刊中有关于信任违背的定义对其进行重新组织；针对错误如何导致信任违背这一因果链逻辑陈述，原稿认为“机器人出错会导致机器人的可预测性和可靠性降低，而这两者恰恰是影响信任的重要因素(Hancock et al., 2021)”，可能这样的陈述有些苍白或含有因果循环影响之意；因此，为进一步清楚阐述错误如何影响信任违背，我们归纳整理了错误影响信任的两类因素，一是错误会让人们怀疑算法的可靠性较低，进而造成信任水平下降(Alarcon et al., 2020; Correia et al., 2018; Lee & Moray, 1992)。二是由于人们往往对于算法错误这类信息的敏感性较高，无法容忍算法出错，一旦算法出错就会直接弃用(Dietvorst et al., 2015)。

意见 5：修改说明文档第 16 页 9 行：提及一个新概念时应当给予适当地解释和定义，此处提及“化身”这一概念应当先简要介绍再展开。
回应：感谢审稿专家老师的宝贵意见！我们已经添加了“化身”的概念。

意见 6：修改说明文档第 16 页 21 行：最好先简要说一下“算法态度”与“信任倾向”的不同。
回应：感谢审稿专家老师的耐心提醒！在原稿中，可能由于我们最初对于“人机信任”的概

念界定与后续“算法态度”概念之间产生了重叠；我们已经查阅以往有关于人际信任和人机信任的中对于“信任”概念界定，重新对本文前言部分有关于“人机信任”的概念进行了梳理，希望能较好地表达清楚“信任”的含义以减少误解；至于“算法态度”与“信任倾向”的关系，我们认为算法态度就是对算法的认知、情感和行为倾向的总和。例如本文中，算法厌恶、算法偏好都是算法态度的典型体现。算法态度会影响人们对于算法的认知、情绪与行为。以算法厌恶为例，个体如果排斥算法，则有可能在人机交互之中拒绝与算法的接触，从而造成信任不足。算法态度不同于信任倾向，信任倾向较难改变，而算法态度则可以通过后续的校准策略进行改善，最终优化个体对于算法的信任水平。我们已经在“算法态度”该段之前插入了有关表述以更好地帮助读者理解“算法态度”与“信任倾向”的不同。

意见 7：修改说明文档第 17 页 10 行：心理模型是一个重要且复杂的概念，它是如何造成信任偏差的，这一段最好能再展开详细说说。
回应：感谢审稿专家老师的宝贵意见！我们补充了相关文献以更好地突出心理模型与人机信任之间的重要关系。

意见 8：修改说明文档第 22 页 1-12 行：文章 3.2 部分提到“改变偏见与事前预警”和“增加接触”这两者对信任校准的作用，但“改变偏见与事前预警”部分只提到了对积极偏见的改善；相反，“增加接触”只提到了对消极偏见的改善，是否这两种途径对积极和消极偏见都有改善作用？可以考虑补充相关文献。
回应：感谢审稿专家老师的细心提醒。“改变偏见”与“增加接触”确实对于校准过度信任与信任不足都比较适用，我们已经重新修改了原稿有关于该两种策略的适用范围的表述。但是“事前预警”策略我们认为恐怕不适合于改善消极偏见。“事前预警”侧重于在人机交互之前提醒用户机器人(算法)可能会出现某些故障、错误以降低个体对于机器人性能的过高心理预期，从而达到校准信任的效果；而对于消极偏见而言，个体可能本身就在交互前对于机器人的印象不太好，如果此时再使用“事前预警”策略，告诉用户机器人的性能等方面可能存在不足，这也许会加剧个体的消极偏见。因此我们认为“事前预警”策略还是比较适合于纠正个体的积极偏见。

意见 9：修改说明文档第 22 页 29 行-23 页 11 行：对拟人化的论述似乎放在“与机器人有关的信任校准策略”部分更加合适。
回应：感谢审稿专家老师的宝贵意见。我们在第二部分“人机信任偏差”中的“2.2.3 与情境有关的因素”提到了决策领域特点可能会导致人机信任偏差，例如任务主观性较强、风险较大的领域，具体表现为相较于算法，人们更愿意依靠人类决策；因此，我们在第三部分“人机信任校准的途径”中的“与情境有关的信任校准策略”提到了“增强决策领域中机器人(算法)的优势”这个途径。该途径一是通过提高算法的专家能力，从客观上优化算法能力，进而提升算法的可靠性、并提高个体选择算法决策的意愿；二通过拟人化特征改善算法在决策情境中不利的地位，例如人们会先入为主地对算法持有偏见，认为算法因其“冰冷”属性只适合于客观任务，不适合主观任务。这种算法拒绝我们认为本质上是因为情境因素(决策情境为主观性任务而非客观性任务)，而不是机器人的因素。因此，我们认为这种针对于主观上个体的算法拒绝，适当地加入拟人化特征或能改善算法的“冰冷”属性。

对于审稿老师提出的建议，我们也承认拟人化这个策略更像是从机器人的角度进行信任校准的策略，但是相对于将拟人化笼统地划分在“机器人因素”用以提升信任不足，我们还是认为它对于提升算法在主观性较强的决策情境中的作用更大、也更具体。

意见 10：最后，在文章整体结构方面，文章第二部分从机器人、个体和情境三个方面对信任偏差原因进行分析，在信任校准途径部分的论述也是围绕这三个方面展开，但为什么校准途径里的方法鲜有是针对原因分析中提到的因素呢。比如，文章提到机器人稳定性、可靠性和化身对信任的影响，但在校准途径部分却没有提出对应的优化稳定性和可靠性方面来校准信任的策略？这种逻辑联系的缺失，可能使得文章结构不够连贯。

回应：感谢审稿专家老师提出的宝贵意见！我们在撰写文章的过程中也有过这样的担忧。信任偏差之因与信任校准策略的不完全对应一是因为在人机信任校准领域的实证研究大多集中于信任修复，对于信任抑制的实证研究相对来说较少；二是人机信任修复研究也大多在关注一些常规信任修复策略的作用(例如道歉、否认、承诺等)。我们在梳理信任偏差原因、信任校准策略的时候根据以往的研究自行整理了相关研究并进行了归纳概括；但是由于信任校准有关的研究，尤其是实证研究其实很难做到与之前提到过的信任偏差之因一一对应，我们也不太方便陈述信任校准的策略而完全不带佐证的参考文献，这或许会被读者怀疑观点的正确性与科学性。但是为保证文章的结构相对完整，我们在撰写文稿的时候依旧按照机器人、个体和情境三方面组织校准策略的内容，只是较难做到一一全部对应，对此我们也深感遗憾。但是，除极少数信任偏差的原因(例如信任倾向较难改变；化身无对应的校准策略)难以找到对应信任校准策略以外，绝大多数的信任偏差原因都能在校准部分中找到对应的策略，具体可见表 1。另外，为改善本文的阅读体验，帮助读者更好地厘清、梳理信任偏差的成因以及与之对应的校准策略，我们在“前言”部分插入了有关图表。

针对审稿老师提到的关于可靠性与稳定性方面的信任提升策略，我们认为通过提高机器人的稳定性与可靠性以提高算法信任太过于简略，而且比较偏向常识，所以就并未在原稿中提及，相反，我们集中了大段篇幅论述当可靠性较低、出现错误(即此时机器人的稳定性与可靠性较低)之后，机器人如何通过信任修复策略提升信任。鉴于审稿老师的提醒，我们就在“信任修复”之前插入了一段文字简要叙述该策略。

表 1 原稿中提及的信任偏差原因与其校准策略的对应关系

信任偏差原因	信任校准策略
可靠性	透明度提升、信任修复(道歉、承诺、否认等)
化身	\
信任倾向	\
算法态度	改变偏见、事前预警、增加接触、透明度提升
心理模型	
期望	
风险与时间压力	认知干预、增加认知资源
决策领域特点	提升机器人优势

审稿人 2 意见：

文章的主题是近年比较重要的方向，作者进行了相关文献的综述，兼顾了全面性与深度，是一篇很好的综述文章。还需以下两个方面需要关注：英文专有名词首字母是否应该大写？未来研究展望篇幅可以缩短精简一些。

回应：感谢审稿专家对我们这篇文章的肯定。针对您提到的第一个问题，我们已经将文章中

涉及到英文专有名词的地方的首字母全部大写处理并标蓝处理。

针对您提到的第二个问题，我们已经对文章的“未来研究展望”部分进行了部分删减。删减之前该部分的正文字数为 3665 字，删减后字数为 2078 字。

第三轮

审稿人 1 意见：经过作者的修改，我的疑问都被清晰地解答了。建议稿件被接收发表

编委 1 意见：论文总体上达到了发表水平。但是在文字上还存在诸多问题，存在句子不通(但是对于算法的信任不足往往会导致个体并倾向于低估算法能力；化身(Embodiment)可能也会对信任有一定影响，化身是指机器人是否具有实体形态，亦或是虚拟智能体；指责是信任修复中风险较大的一种策略，且最好可以倾向于机器人去指责自己内部的原因(Groom et al., 2010)，而不是外部原因(算法设计师、第三方算法、人类同伴)等等)、专业术语未解释(如启发式思维)、参考文献格式不对(并且也会更加满意(Kim, D., & Kim, S., 2021)) 等等问题。建议作者做仔细的文字校对。

回应：感谢编委的意见。我们已经重新对文章进行了校对，补充了部分术语的解释，并检查了参考文献格式，尽力避免文字错误。

编委 2 意见：有几点请作者继续修改

意见 1：作者提出了信任抑制是信任校准的两种路径之一，但是对于信任抑制描述非常单薄，甚至在概念上都不太清晰，这是一个非常明显的缺陷，作者需要极大强化这方面研究的综述和讨论。

回应：感谢编委的意见。我们已经结合前人研究对“信任抑制”进行了定义。您提到的有关于信任抑制单薄这个问题其实在外审阶段审稿专家 1 也表达过类似的关切。“信任抑制”所占篇幅看上去较为单薄的原因有二：一是从客观角度上来说，在人机信任校准领域的实证研究大多集中于“**信任提升(尤其是信任修复)**”，相比之下“**信任抑制**”较少被研究者所关注(Rheu et al., 2021)，因此实证研究相对来说较少；与实证研究数量更为丰富、研究内容更为深入的“**信任提升**”领域相比，能纳入综述写作范围的、有关于“**信任抑制**”的研究(不管是综述类亦或是实证类)确实不算多，对此我们深感遗憾。我们已经重新搜寻了有关文献，补充了部分信任抑制的实证研究介绍，修改之后篇幅占比可见表 1。

二可能是我们行文结构更改的原因。我们之前的文章结构是将信任抑制(过度信任)与信任提升(信任不足)板块分开来进行论述的，信任抑制策略从写作的角度讲就更加集中，更突出；但是在外审阶段审稿专家提到某些因素(校准策略)其实是可以通用的，因此我们就更改了最初的写作逻辑，重新按照影响人机信任的三因素(人机环)结构来组织文章，合并了部分内容并考虑到字数问题删减了部分内容，阅读起来可能就没有明确区分开“信任抑制”与“信任提升”方面的内容。鉴于此，我们已经在文章明确提出信任抑制的适用策略。

表 1 “信任抑制”与“信任提升”所占篇幅字数统计

	修改前		修改后	
	信任提升	信任抑制	信任提升	信任抑制
2、人机信任偏差	1503	1158	**1170**	**2128**
3、人机信任校准的途径	3101	1392	**2801**	**2285**
4、未来研究展望	518	462	**400**	**626**

意见 2：有一些逻辑上的问题，比如 3.1 与机器人有关的信任校准策略这个部分逻辑混乱，为什么将信任修复与透明度提升并列，信任修复部分是个大杂烩，分类方法和其他不同；3.3 与情境有关的信任校准策略的部分也有问题，拟人化怎么变成情境？

回应：感谢编委的意见。针对您提到的第一个问题，或许也是因为文章结构调整产生的逻辑问题。在我们之前的逻辑线中，“透明度”是提升个体初始信任水平过低的信任提升策略之一，“信任修复”则是修复个体在信任违背后的较低信任水平的主要策略。两种策略其实并不冲突。文章结构调整后，我们将两种策略合并到“与机器人相关的信任校准策略”。信任修复是人机信任提升极为重要的部分，也是以往人机信任校准研究中实证研究最为丰富的部分。我们本意是想将“信任修复”作为一个大类，将道歉、承诺、否认等丰富的修复策略囊括其中，但这样确实会产生如您所说的逻辑问题。因此我们删去了“信任修复”的表述，将道歉、承诺、否认等信任修复的具体策略单独列出，希望这样能避免因概念不在一个层面而造成的逻辑问题。

您的第二个问题（拟人化与情境的关系），外审专家 1 在第二轮外审中也提到同样的问题。我们将拟人化归入与情境有关的策略的考虑如下：

我们认为拟人化特征或可以改善算法在决策情境中不利的地位，例如人们会先入为主地对算法持有偏见，认为算法因其“冰冷”属性只适合于客观任务，不适合主观任务。这种算法拒绝我们认为本质上是因为情境因素(决策情境为主观性任务而非客观性任务)，而不是机器人的因素。因此，我们认为这种针对于主观上个体的算法拒绝，适当地加入拟人化特征或能改善算法的“冰冷”属性。我们虽然也认可拟人化这个策略更像是从机器人的角度进行信任校准的策略，但是相对于将拟人化笼统地划分在“机器人因素”用以提升信任不足，我们还是认为它对于提升算法在主观性较强的决策情境中的作用更大、也更具体。而且，本文中“与机器人相关的信任校准策略”、“与个体相关的信任校准策略”、“与情境相关的信任校准策略”部分不仅仅包括哪些信任校准策略是从“机器人”、“个体”、“情境”出发去校准信任的，还包括针对“与机器人有关的信任偏差原因”、“与个体有关的信任偏差原因”、“与情境有关的信任偏差原因”，信任校准策略该如何去纠正。因此，拟人化听上去像是从“机器人”角度出发的信任校准策略，但它更属于后一种，即**“与情境有关的信任偏差原因”的校准策略。**

以上便是我们将拟人化归入与情境有关的策略的原因，虽然之前的审稿人也接受了我们的解释，但编委仍存有同样的质疑。考虑到两位专家的意见、避免读者也产生类似的疑问，我们最终将其调整到“与机器人相关的信任校准策略”中，尽量聚焦于校准策略本身，而不必对策略的分类归属做过多额外解释。

意见 3：reviewer 也指出了一个逻辑问题，信任抑制和信任提升和人机环三个方面有什么关系，为什么在整体的图 1 中没有体现，标题和内容没有很好对应起来。

回应：感谢编委的意见，已重新修改图 1，希望能更好地体现上述概念之间的关系。

第四轮

编委 2 意见：同意发表。

主编意见：根据编委和审稿专家的意见，建议发表。