# 《心理科学进展》审稿意见与作者回应

题目：通用人工智能时代的人与 AI 信任
作者：齐玥、陈俊廷、秦邵天、杜峰

---

## 第一轮

**审稿人 1 意见**：这篇综述不错，值得发表，但是要做些修改。我的修改意见如下：

**意见 1**：引言部分，请作者进一步详细说明为什么要写这篇综述。
**回应**：感谢审稿专家提出的宝贵意见。我们认识到在引言部分需要更加明确地阐述本综述的写作目的，为此，我们在引言部分增加了一段论述，以具体阐述本文关注的问题。具体修改详见文中红色字体。相关原文内容如下：
“当前，人与 AI 的交互关系已经开始转变，但是现有的人与 AI 信任研究并没有准确理解这种新型的信任关系。这种理解不足主要体现在三个方面：首先，现有研究对于人与 AI 信任的定义并不明确，这会导致不同研究者对人与 AI 信任的理解和应用存在差异；其次，传统的信任模型大多从人际信任和人机信任两个角度分别展开阐述，但随着人工智能技术的提高，人与 AI 的交互将逐步贴近人与人的交互，融合心理学中两个不同的研究领域变得更加有价值；最后，现有的信任模型仅关注到人对 AI 的信任，忽视了 AI 对人的信任这一角度，对人机互动的双向信任过程缺乏理解。为解决现有研究的局限性，本文将围绕人与 AI 信任的定义、信任模型的发展展开，提出并阐述人与 AI 动态互信模型，并在最后对人与 AI 信任的未来研究进行了展望。”

**意见 2**：作为综述性论文，完整系统的文献检索很重要。希望作者能够提供该论文检索的说明，比如用什么关键词、在什么数据库等。
**回应**：感谢审稿专家的提醒，提供一个文献检索说明可以确保研究的全面性和可靠性，是非常有必要的。我们整理了本文所使用的文献检索策略，包括检索的关键词、检索的数据库，检索的时间范围等，并补充在了引言部分相关段落。具体修改详见文中红色字体。相关原文内容如下：“为获得人与 AI 信任的相关文献，在准备本综述的过程中，本文采用的文献检索策略如下。本文在中国知网、Web of Science、IEEE Xplore、Elsevier ScienceDirect 中进行关键词检索，所使用的检索关键词包括“人机信任(Human-Machine Trust)”、“人工智能信任(trust in AI 或 trust in artificial intelligence)”、“自动化信任(trust in automation)”和“机器人信任(trust in robot)”，文献检索的时间为 1994 年至 2024 年 1 月，文献类型包括期刊论文和会议论文，以确保涵盖近 30 年的研究成果。”

**意见 3**：第一部分“人与 AI 信任的定义”，作者给出了相关的定义，是不是能够展开对这个新定义进一步进行讨论，比如结合以往的定义，说明新定义的特点等等。
**回应**：感谢您的建议，我们在给出定义的部分又增加了进一步的讨论，在结合以往定义的基础上说明新定义的特点和理论意义。具体修改详见文中红色字体。相关原文内容如下：

“本文的新定义综合了以往人机信任和自动化信任定义的内容，不仅涵盖 Lee 和 See (2004)所提出的基于态度的人机信任观点，也符合 Billings 等人（2012）总结的自动化信任三项核心特征：两个信任主体、完成的事情存在风险以及受托人有完成任务的动机和能力。在

综合以往观点的基础上，新定义充分考虑了当今人与 AI 互动的特点：一方面针对 AI 技术使用的隐蔽性强调定义可以扩展到用户未意识到 AI 参与的情况，另一方面考虑到人与 AI 信任角色的转变，提出人与 AI 存在互信的关系，即信任包括用户作为委托者对 AI 的信任，也包括了 AI 作为委托者对用户输入的依赖和适应。这种互信关系也潜在揭示了人与 AI 信任的动态过程，交互过程中人与 AI 都会作为委托者，并根据受托者的行为来不断校准自己对受托者的信任。”

**意见 4**：本文第二部分“人与 AI 信任 XXXXX”，阐述了以往的各种人际信任模型后，应该有一个对所有模型的综合性的对比分析。希望作者加上。

**回应**：感谢您的建议，我们在第二部分中单独增加了一个部分以对信任模型进行综合并比较分析，具体修改详见 2.5 部分。相关原文内容如下：

“**2.5 过往信任模型的综合比较与分析**

通过上述模型可以看到研究者对于信任的理解是在不断演进和深化的。早期的信任模型是在人际互动情景下讨论的，Mayer 等人(1995)的模型开创性提出信任取决于受托人的能力、仁慈和正直三个特征，但是却仅考虑到受托人一方的特征。McKnight 和 Chervany(1996)的模型拓展了委托人和情境两方面因素的影响，更全面地解释了人际信任的影响因素，也为人机信任模型的提出提供了大体框架。在人机交互领域，Sanders 等(2011)最早对人机信任研究进行总结，提出了人机信任的四因素模型，但是该模型并不能广泛概括所有影响信任的前因变量。Hancock 等人(2011)对现有模型进行了修订，将前因变量总结为三类因素，该模型仅基于人与机器人交互的相关研究，该领域里研究人类相关因素和环境相关因素的支持证据较少，因此对这两类因素的探索并不充分。Schaefer 等人(2014)基于人与自动化交互研究的元分析结果对三因素模型进行了修订，并对每个因素的内容都进行了更细致的划分，Lewis 等人(2022)则进一步发展了信任模型，其整合模型考虑了代理信任的影响，强调信任的动态调整过程，为不同类型的信任关系研究提供了一个通用的分析框架。可以看到过往信任模型的发展呈现出从静态到动态、从单纯的维度划分到更全面、细致的因素考量的趋势，但是仍存在忽视人与 AI 双向互信关系的局限。”

**意见 5**：本文第三部分提出了一个新的人与 AI 动态互信模型。在提出模型前，作者应该有一部分论述，说明提出该模型的出发点和思路。希望作者加上。

**回应**：感谢审稿人建议，我们在 3.1 提出模型部分开头，增加了一段论述，以说明本文提出模型所依循的思路并强调了模型的核心出发点。具体修改详见文中红色字体。相关原文内容如下：

“在通用人工智能时代背景下，人与 AI 的互动关系日趋复杂。过往人机信任模型，尽管在理论上有所贡献，但在解释人与 AI 之间动态且双向的信任关系方面存在局限，已不足以全面描述人与 AI 之间的信任交互过程。因此，本文拟提出一个新模型，旨在填补现有人与 AI 信任领域理论模型的空白。该模型充分参考已有信任模型的内容，尽量全面地把握影响信任过程的因素。在模型框架上，充分参考人际信任模型(Mcknight & Chervany, 1996)，包含委托人相关因素、受托人相关因素以及情境因素。每一类因素的具体内容参考通用性强的整合模型(Lewis & Marsh, 2022)，并进一步考虑人与 AI 互动的独特性。因此，新模型的特点体现在：模型强调信任不仅是人对 AI 的单向评估，而是一个涉及人和 AI 双方的互动过程，人和 AI 均会根据对方的行动和反馈，不断调整自身的信任水平和行为策略。综上，本文在已有的信任模型（包括人际信任模型、人机信任的四因素模型、人机信任的三因素模型、以及人对 AI 信任的整合模型）的基础上，针对通用人工智能时代人与 AI 双向互信的新型交互关系，提出了一个新的人机互信模型：人与 AI 动态互信模型，如图 2 所示。”

**审稿人 2 意见：**本研究回顾了人与 AI 信任的定义、信任模型的发展，提出了人与 AI 动态互信模型，并对人与 AI 信任的未来研究进行了展望。

**意见 1：**摘要过于简洁，仅简单陈述提出了一个人机互信模型，缺少对模型创新性和重要性的阐述。

**回应：**非常感谢审稿人的意见，我们按照您的建议在摘要中补充对于模型创新性和重要性的阐述，具体修改详见文中红色字体。相关原文内容如下：

“随着技术的发展，通用人工智能初见雏形，人机交互以及人机关系将进入新的时代。人与人工智能（AI）的信任关系也即将从单方向的人对 AI 信任逐渐转变为人与 AI 的互信。本研究在回顾社会心理学中的人际信任模型与工程心理学中的人机信任模型的基础上，从人际信任视角提出了人与 AI 动态互信模型。该模型将人与 AI 视为对等的信任建立方，结合信任与被信任方的影响因素、结果反馈和行为调整构建了人与 AI 动态互信的基本理论框架，强调了人与 AI 信任中关系维度的“互信”与时程维度的“动态”这两个重要特征。模型首次将 AI 对人的信任以及二者互信的动态交互过程纳入分析，为人与 AI 的信任研究提供新的理论视角。未来研究应更多关注 AI 对人的信任如何建立与维持、人与 AI 互信的量化模型、以及多智能体交互中的人与 AI 互信。”

**意见 2：**本文提出的模型的一个重要创新点是考虑了 AI 对人的信任。然而 3.3 对相关内容的陈述比较有限。例如，仅简单说了用户状态、系统状态和情景状态是影响 AI 对人信任的三个因素，但是这三个方面为何会影响、具体如何影响，并没有详细阐述。

**回应：**感谢审稿人的建议，已在对应部分增加了一部分表述，以更详细地说明三个因素是如何影响 AI 对人的信任的。具体修改详见文中红色字体。相关原文内容如下：

“在感知阶段，AI 对人的信任同样受三方面因素的影响。一是用户状态，AI 需要构建监测系统对使用者的状态（认知、生理、意图、情感、价值观、道德水平等）进行实时监测，当使用者处于不可信任状态时（如疲劳、分心）AI 会主动接管以避免事故(许为 等, 2024)。二是系统状态，AI 需要对自身状态有一个主动监测和评估系统，一方面是监测自身的性能和稳定性，另一方面是评估当前状态是否能够完成任务。以自动驾驶为例，自动驾驶汽车会配备大量的内部传感器，以随时监测汽车内部状态数据，并且研究者还在不断开发有效的自动故障诊断和健康监测算法(Biddle & Fallah, 2020)以评估系统状态。当系统监测到自身并不可靠时（如系统故障、任务超出系统能力），就会做出信任人类的判断，并提示人类用户接管控制权。三是情境状态，AI 需要对所处情境的风险程度、复杂程度进行评估，比如环境状况、紧急情况的发生等，以判断是否应该信任使用者。同样以自动驾驶为例，汽车会使用摄像头、激光雷达、超声波传感器等传感器来感知交通路况、光照条件、障碍物情况等外部情境(Ignatious et al., 2021)，并根据感知到的情境采取相应的信任行为。当系统监测到高风险情境时（如汽车驾驶员即将发生追尾），AI 可能会更加谨慎，减少对人类的信任，采取刹车、紧急变道等紧急措施；而在低风险情境下，AI 就会更信任人类，给人类更多自主行为的权力。”

新增文献：

Biddle, L., & Fallah, S. (2021). A Novel Fault Detection, Identification and Prediction Approach for Autonomous Vehicle Controllers Using SVM. *Automotive Innovation*, *4*(3), 301-314. doi: 10.1007/s42154-021-00138-0

Ignatious, H. A., & Khan, M. (2022). An overview of sensors in Autonomous Vehicles. *Procedia Computer Science*,

198, 736-741. doi: 10.1016/j.procs.2021.12.315

**意见 3**：对于图 2 提出的模型，我对“信任行为的结果反馈影响初始信任”这一路径存疑。作者提到，初始阶段是人-AI 交互前的阶段，应该是属于闭环反馈之外的初始条件。

**回应**：感谢审稿人的意见，我们认为在原模型中对于初始阶段里三个因素各自起到的作用表述不够清晰，为了准确体现人与 AI 的互信过程，我们对模型进行了相应调整，将信任倾向和系统信任放到了闭环以外，同样地，我们也将 AI 初始阶段的信任倾向放到闭环之外，而信任经验是所有人与 AI 交互经验的积累，这种积累是长期且持续的，塑造了个体与 AI 互动总体的信任态度，在闭环中始终起到作用，因此我们仍然将其置于闭环中。具体修改详见图 2 和文中红色字体。相关原文内容如下：

“综上，本框架提出人与 AI 的动态互信可划分为三个阶段：人与 AI 交互前的初始阶段、人与 AI 交互中的感知阶段以及行为阶段，并且这三个阶段形成闭环。初始阶段是人与 AI 信任的最初阶段，人与 AI 尚未接触，依赖于自身固有的信任倾向、系统信任和以往交互中得到的相关信任经验等，为之后的信任奠定基调。其中，信任经验会在接收到本轮交互的结果反馈后得到矫正，参与人机互信的动态过程；而系统信任和信任倾向相对稳定，不会参与后续的动态过程。”

**意见 4**：研究展望中，关于“组织层面”研究的提出略显突兀，而且这部分内容基本没有提到信任，尽最后一句说未来研究应该考虑组织层面的人与 AI 互信。这导致本章节和全文的关联度较低。

**回应**：感谢审稿人的意见，我们认识到组织层面的人与 AI 互信的提出后确实和前文联系较弱，其内容与我们想展望的内容也有所差异。我们重新阐述了该部分内容，更清楚地表达了未来研究可以从个体与单个 AI 的互信关系拓展到多个人与多个 AI 的互信关系的展望。具体修改详见文中红色字体。相关原文内容如下：

“本文提出的模型适用于 AI 作为人类助手或者协作伙伴、人机协作完成任务等常见情境(Mohanty & Vyas, 2018)，但是模型仅关注了单一人类与单个 AI 互动时的互信过程，随着 AI 使用场景的复杂化，将会涉及到多个人类与多个 AI 之间的互动。以往研究者认为，在多智能体互动中，每个成员所担任的角色以及成员之间互动的方式都是影响信任的关键因素(Yagoda & Gillan, 2012)。在这样的环境中，信任的动态构建过程变得更加复杂。可以在本模型的基础上应进一步纳入各智能体的身份角色，考虑其在动态互信过程中的权重。举例而言，图 3 在分布式认知(Perry, 2003)的基础上，纳入了人与 AI 动态互信过程中角色的分配。当多智能体互动中出现“意见领袖”时（图中蓝色智能体），意见领袖（可能是人类或 AI）的信任经验将通过交流，进而影响到其他智能体（图中灰色智能体）的过程。只有将人工智能放到复杂群体（如团队或网络）中进行研究，研究人员才能真正理解人们与人工智能建立“合作伙伴”关系的方式，以及人工智能如何改变人与人之间以及人与其他机器之间关系的方式的问题。此外，由于人工智能的行为不是稳定不变的，学者们需要研究它基于人类与人工智能交互的变化方式(Rahwan et al., 2019)，以促进对关系变化的了解。未来的研究应该考虑多人与多个 AI 的交互，这将为建立人与 AI 的伙伴关系，形成人在回路（human-in-the-loop）的人与 AI 互信提供更好的支持。

新增文献：

Mohanty, S., & Vyas, S. (2018). Putting It All Together: Toward a Human-Machine Collaborative Ecosystem. In S. Mohanty & S. Vyas (Eds.), *How to Compete in the Age of Artificial Intelligence: Implementing a collaborative human-machine strategy for your business* (pp. 215-229). doi：10.1007/978-1-4842-3808-0_11

Perry, M. (2003). Distributed cognition. HCI models, theories, and frameworks: Toward a multidisciplinary

science, 193–223.

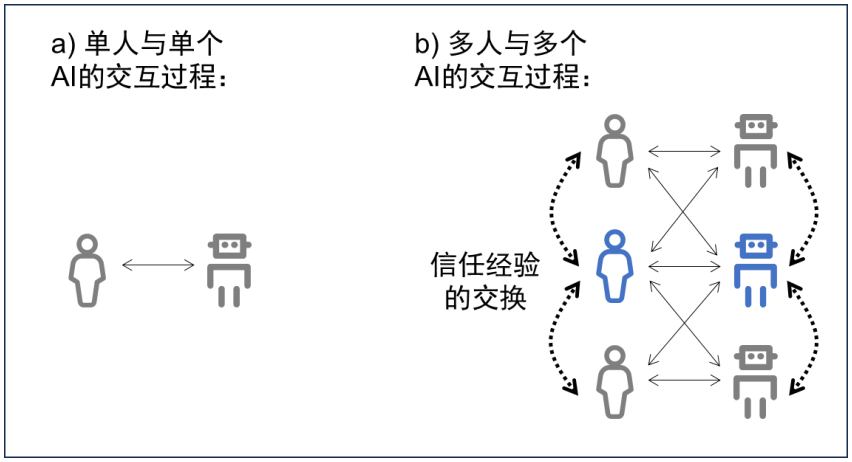


图 3 人与 AI 的信任交互过程。a) 单人与单个 AI 交互（实线），b)多人与多个 AI 的交互过程。其中，人/AI 交互（实线）集中的节点即意见领袖（蓝色），意见领袖可能在不同的人类主体（灰色）之间形成信任经验的交换（虚线），进而影响到其他人类对 AI 的信任。”

**意见 5**：图 2 中的概念，不是在同等水平。例如，作者对上半部分的“系统状态”进行了拆解细化，分为“感知系统可信”和“感知系统风险”；但是下面半部分，对这些概念又没有拆解。这导致理解图 2 时较为困难。

**回应**：感谢审稿人指出了模型中概念层次不一致的问题，为保证感知阶段上下两部分内容在层次上的一致性，我们将“感知系统可信”和“感知系统风险”合并为了“感知系统状态”，使模型更简洁和易于理解。对于文中相应部分也进行了修改，具体修改详见文中红色字体。相关原文内容如下：

“在感知阶段，人对 AI 的信任受到三方面因素的影响。一是感知个体状态，即个体觉察自身是否能够胜任当前任务。二是感知系统状态，包括感知可信和感知风险。其中感知可信包括对被信方的能力、可预测性、正直、仁慈以及代理信任（如品牌）等多维度的感知(Chen et al., 1995; Hoff & Bashir, 2015)。感知风险，是指对被信方脆弱性和完成当前任务所伴随的风险水平的评估(Ajenaghughrure et al., 2020; Ma et al., 2020)。”

**意见 6**：论文存在一些表述不清或错误的地方。例如，2.4 章节第一段，什么是“代理信任的主观、情境可信度判断”？。再如，“随后，信任决策基于这种可信度判断，加上感知风险和委托人的信任倾向因素，这些因素可能是情境的、个人的和系统的。”这句话感觉是一句英文的直白翻译，不符合中文表述的习惯。章节 3.3，第一句多了一个“的”。此外，“更多取决于 AI 的固有特质（如用户群体特征、主要任务、形态、安全保障等）”，用户群体特征怎么属于 AI 的固有特质？

**回应**：感谢审稿人指出了文章表述中的错误，这对提高文章的可读性意义重大。我们通读了全文，重新检查了文章中所有表述上的不足，并在文中进行了对应修订，具体修改详见文中标红字体。

---

## 第二轮

**审稿人 1 意见**：作者已回答了我提出的所有问题，令人满意，同意发表。

**审稿人 2 意见：**感谢作者对我的意见做出的详实、准确地答复，我认为该文章已达到发表的水平。谢谢。

**编委 1 意见：**同意发表。

**编委 2 意见：**同意发表。

**主编意见：**稿件经过多位专家的审阅，作者进行了认真的修改，达到了发表水平，同意发表。