《心理科学进展》审稿意见与作者回应

题目：人工智能代理对道德决策的影响
作者：唐伟 钟文瑞 雷震 张丹丹

第一轮

审稿专家 1

文章试图回答人工智能代理如何影响人们的道德决策，以及人工智能作为决策代理与其他代理相比有何独特之处，研究问题比较前沿和重要。作者构建了“决策者—代理—反馈者”决策与归责框架，并在决策链和反馈链两条路径上进行了分析，综述角度具有创新性。在具体的框架构建和论述中有一些问题，建议作者修改。

答复：感谢您审阅我们的论文并提出宝贵意见。我们认真梳理并逐条回应了您的建议，也据此对稿件进行了修改。您的建议极大地提升了论文整体质量。为便于您核对，我们在修改稿中用蓝色标出了新增或改写的内容；在本回复信中，您的原始意见以楷体呈现，修改稿中的对应修改则以方框摘录于此。

意见 1：综述所涉及重要概念的定义需要进一步厘清。首先，作者将代理定义为“根据决策者预先规定的决策规则和目标函数，代替决策者执行某一类具体决策的主体或技术系统”，这一定义是否有参考文献支撑？其次，对于道德决策的定义，作者写到：“本文将决策者在具有道德后果的情境中做出的决策统称为道德决策 (Kouchaki & Smith, 2025)。”在涉及道德的情境中所做的决策就一定是道德决策吗？Kouchaki & Smith (2025)原文中对道德决策的定义为“By moral decision-making, we mean the process of making judgments and choices that have moral and ethical implications.”似乎与作者所写并不完全相同。

答复：感谢审稿老师对“代理”概念界定问题的指出。根据您的建议，本文详细搜索文献后发现，本文定义中提及的“根据决策者预先规定的决策规则和目标函数”的表达不太准确。尤其是在人类代理的情形下，代理人与委托者之间往往存在利益不一致或判断偏差，代理的具体行为未必严格遵从委托者预设的目标取向。故而在本修改稿中更正了这一定义。

关于代理定义缺乏参考文献支撑的问题，本文详细搜索后发现，在 Köbis et al.(2021)中，其对代理（delegate）的定义是“Besides active partners, others can also serve as delegates to whom people can outsource the execution of unethical behavior. When people face the choice

between breaking ethical rules themselves versus letting others do so on their behalf...”。即人们将不道德行为的执行权外包给代理，让代理代表人们执行某决策。Ross(1973)强调了代理在行为执行层面对委托者的代表作用（“the agent, acts for, on behalf of, or as representative for the other, ..., in a particular domain of decision problems...”）。更进一步地，Köbis et al.(2025)强调了代理关系并不局限于人类主体，技术系统同样可以代替委托者执行任务或决策，如“...delegating tasks to software systems powered by artificial intelligence (AI), a phenomenon we call machine delegation”。因此，在修改稿中，我们微调了定义并补充了相关参考文献（p1）：

所谓代理，是指代替决策者执行某一类具体决策的主体或技术系统(Köbis et al., 2021; 2025; Ross, 1973)。

同时，我们也非常感谢审稿老师指出的关于“道德决策”定义的问题。在详细阅读并理解 Kouchaki 和 Smith (2025)原文中对道德决策的定义后，我们发现本文之前对道德决策的定义确实不够准确。因此，在修改稿中，我们修改了对“道德决策”的定义（p2）：

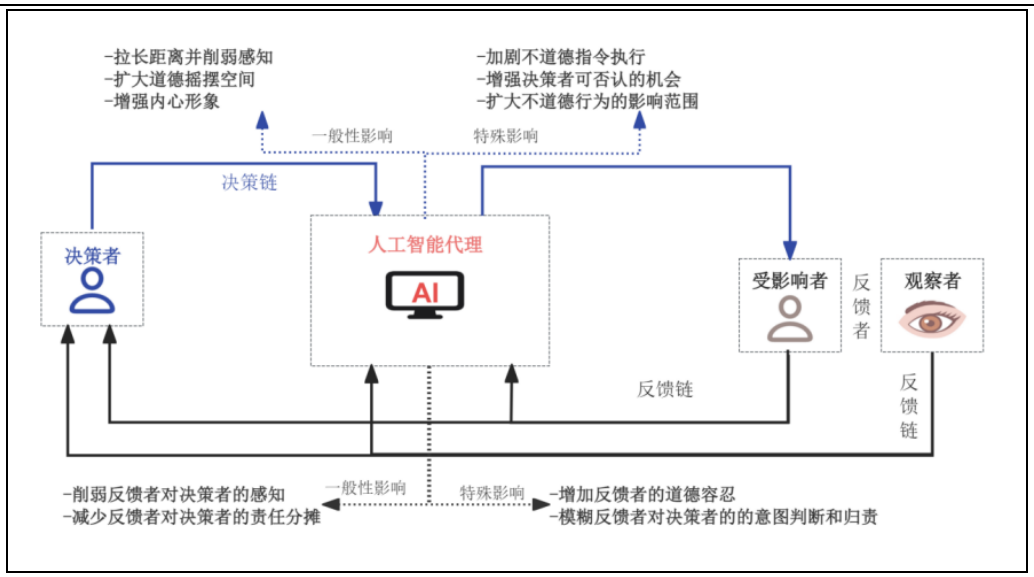
本文关注道德决策，即个体做出具有道德与伦理后果的判断与选择的过程，该过程涉及评估潜在行动的正误(即道德判断)，并选择如何展开后续行动(Kouchaki & Smith, 2025)。

意见 2：对于为何要分为决策链和反馈链两条路径，作者只是一笔带过，但这是整个框架的重要基础问题，建议作者进一步详细阐述其原因。此外，框架中有“内在动机”，那么是否反馈链为“外在动机”？那为何不分为内在动机和外在动机？

答复：感谢审稿老师对决策链和反馈链重要性和选取理由的建议。决策链和反馈链的框架的确是本文的基础，您的建议对提升文章框架合理性起到了关键作用。在修改稿中，我们加入了一段重点阐述选择“决策链”与“反馈链”的理由（p3）：

具有道德和伦理后果的道德决策往往嵌入在由多个主体和多个阶段构成的过程之中，决策者在做出决策时不仅依赖偏向情绪化的直觉决策系统，还会依赖偏向理性化的受控认知决策系统(Greene, 2007)，同时还根据对反馈者反应的前瞻性预期而做出道德决策。而反馈者对决策者行为的评价、归责与惩罚，则依赖其对决策路径、执行主体与责任分工的理解。正是在这一多主体和多阶段的互动中，人工智能等代理才成为具有独立分析价值的变量：代理并非仅影响最终结果，而是嵌入决策流程，重塑行为在不同阶段的生成、执行与理解。因此，本文构建以决策者为起点的“决策链”和以反馈者为起点的“反馈链”(图 1)，以系统刻画代理介入后，道德行为如何在决策生成与社会反应两个方向上被同时重构。

审稿老师对“决策与反馈”链和“内外动机”概念区分的提问也十分深刻。我们认同审稿老师的理解：从决策者自身视角来看，其道德决策确实同时受到内在动机（如社会偏好等）与外在动机（如惩罚、责难等）的共同影响。在原稿中，本文选择“决策链、反馈链”的划分而非“内在动机、外在动机”作为主线主要出于两点原因：第一，从我们的分析目标和框架层级来看，本文关注有多角色参与、跨阶段展开的道德决策过程，而非仅仅关注决策者的心理动机结构。虽然受影响者与第三方观察者评价、归责与惩罚构成了社会反馈机制，影响了决策者的心理动机和道德决策。但其作为独立行动者，其自身行为如何被人工智能代理影响亦是本文关注的重要方面之一。因此，采用“内外在动机”的表述容易在概念上将不同角色的行为过程压缩为决策者的动机来源，弱化了多主体互动这一核心特征。第二，反馈链中的作用机制既可能通过惩罚、声誉损失等外在后果影响决策者的前瞻性预期，也可能通过规范判断、责任归因与道德评价等机制发挥作用，后者并不完全属于外部动机，也可能是作用于决策者心理成本认知等因素的内在动机。但需要承认的是，原稿框架中对内在动机和外在动机的区分与强调可能混淆了决策链和反馈链的重要性，为突出上述本文的关注点，我们在此修改稿中修改了图 1（p4），去掉了对动机的分类表达和强调：



相应的，在第二章“决策者—代理—反馈者”决策与反馈框架中，我们弱化“内在动机/外在动机”标签式表述，更多侧重于介绍决策链—反馈链—多主体/多阶段互动的框架语言。例如（p3）：

> 在无代理的直接决策情境下，决策者既是道德选择的制定者也是行为的执行者，其决策同时由自身偏好与来自他人的社会反应共同塑造。……

意见 3：文章题为“人工智能代理对道德决策的影响”，但是框架其实是代理（多种类型，并

不是只有 AI）对道德决策的影响，并且综述的大篇幅内容也是在论述代理对道德决策的影响，这可能会使得文章主题显得不够突出。

答复：感谢审稿老师指出这一关键问题。我们同意审稿老师的判断：本文所构建的分析框架中过多强调了多种类型的代理对道德决策的影响，削弱了本文主题“人工智能代理对道德决策的影响”。在原稿中，本文试图将人工智能置于已有的“委托—代理”框架中，关注人工智能代理究竟在哪些机制上产生了与其他代理类似的影响，又在哪些方面表现出质的差异。但这一等权重的对比方式的确弱化了本文重点——人工智能代理对道德决策的特殊性影响。因此我们在修改稿中做出三个方面的修改以凸显人工智能代理对道德决策的特殊性影响。

第一，我们调整了章节结构，将原本第二节（“决策者—代理—反馈者”框架）、第三节（决策链，包含一般性影响和特殊性影响）和第四节（反馈链，包括一般性影响和特殊性影响），调整为第二节（“决策者—代理—反馈者”框架，介绍框架及代理的一般性影响）、第三节（人工智能代理在决策链上的特殊性影响，详细从决策链介绍人工智能代理对道德决策的特殊性影响）和第四节（人工智能代理在反馈链上的特殊性影响，详细从反馈链介绍人工智能代理对道德决策的特殊性影响）。这一调整弱化了一般性影响在全文中的地位，仅将其作为人工智能代理的一般性影响与框架一同呈现。

第二，我们修改了图 1（p4），凸显出“人工智能代理”的核心主题，将其他代理的影响作为一般性影响体现。

第三，我们在第三章（人工智能代理在决策链上的特殊性影响）和第四章（3.人工智能代理在反馈链上的特殊性影响）中进行了大量修改，更详细地阐述了人工智能代理对道德决策的特殊性影响，补充了相关文献。例如，3.2 节 增大决策者可否认的机会（p8-9）：

> 第一，相比人类代理而言，人工智能代理增大了决策者可否认的机会。正如前文所述，人类代理可能因为自身道德水平或认知水平而拒绝或揭发委托人/决策者的不道德指令，这降低了决策者在委托不道德行为后否认的机会(Köbis et al., 2025)。其次，人类代理所在的工作流程常常要求明确的工作留痕，这进一步弱化了决策者可否认的机会。而人工智能代理，其独特的学习能力与黑箱性，从输入与输出两个维度为决策者对不道德动机的否认提供了支持。在输入端，人工智能代理擅长基于数据或根据模糊抽象的指令自动推演出道德决策的执行路径，从而使决策者免于明确表达不道德指令，实现了指令输入层面的不留痕(Wang et al., 2025)；在输出端，人工智能输入—输出的黑箱性将具体的推演逻辑隐匿，在底层技术层面实现了不留痕。因此人工智能学习能力及黑箱性构建了一种无需合谋且在底层系统中不留痕的稳定免责机制，极大地强化了决策者的可否认性(Köbis et al., 2021; Rahwan et al., 2019)。

第二，相比于传统规则型算法而言，人工智能代理更显著地增强了决策者可否认的机会。传统规则型算法通常基于开发者预先设定的显式规则与阈值执行决策，其输入—输出映射相对透明(黑箱性低)、可复现且可审计，因而反馈者更容易将具体输出回溯到委托的规则设定与授权范围，因此决策者难以以不可预见或不可控制为由否认自身意图或责任。然而，正如前文所述，人工智能代理具有学习能力和输入—输出黑箱性的特征(Babic et al., 2021)，这在两方面提供了更强的可否认性。首先，这些特征允许决策者以抽象模糊的方式向代理传达不道德行为的指令，因此决策者并未明确下达不道德行为指令。间接证据表明，一旦外界环境允许决策者以含糊方式下达不道德决策指令(如在骰子实验中要求人工智能代理“帮我实现利润最大化”等抽象目标)，决策者的不道德行为会大幅增加(Chevrier & Teixeira, 2024; Köbis et al., 2025)。这些间接证据从侧面印证了允许模糊指令可能让决策者更容易否认自身不诚实的意图。其次，这些特征使决策者指令和道德结果之间的推理路径模糊，当决策者面临不道德行为产生的道德后果时，其可以通过主张自己无法充分预见或控制人工智能代理的具体输出而为自己的不道德行为辩护，否认自身意图，减少心理成本。

综上，从决策者角度看，人工智能代理的学习能力与黑箱性减少了不道德指令留痕、模糊了输入—输出推理链条，使决策者更易以不可预见或不可控制为由否认意图与责任，从而推动决策者更频繁地委托人工智能代理实施不道德行为。

意见 4：作者想要探讨的是人工智能代理对道德决策的影响，但是从框架和论述可以看到，很多内容所探讨的并非对于道德决策本身的影响（也就是人们如何或者做出何种道德决策），而延伸到的道德决策做出之后的后果，如“扩大不道德行为的影响范围”，这与道德决策本身已经关系不大了。

答复：感谢审稿老师对论文主题边界的提醒。我们认同审稿老师所指出的区分：若同时关注不道德行为做出之后的后果（如影响范围扩大、外溢/扩散现象等），确实会削弱对“道德决策本身”（即决策者如何做出选择、是否采用代理）的聚焦。在原稿中，我们讨论“影响范围扩大”的目的是试图强调人工智能代理的规模化与可复制性的特点会导致不道德行为影响范围的扩大，而这一影响会进一步地被决策者预期到，从而在道德决策时将潜在收益、风险与问责纳入道德决策体系之中，从而改变其道德判断、委托选择与行为强度。也就是说，我们并非将“影响范围扩大”作为独立于道德决策之外的结果变量，而是将其视为人工智能代理改变决策者道德决策的关键机制之一。但应该承认的是，本文原稿的表述存在不合理之处，容易让读者产生误解。因此，我们在本修改稿的相关段落中，补充了道德决策后果如何进一步影响决策者道德决策的论述，强化逻辑闭环。例如在“扩大不道德行为影响范围”（p9）中，

我们补充了一段总结：

> 综上所述，人工智能代理凭借低成本复制与跨场景个性化配置，更易将原本局部的不道德行为快速扩散并放大至更广泛领域，使其发生更频繁、传播更广且更难被及时识别，这扩大了不道德行为的潜在收益，从而使决策者更愿意委托人工智能代理实施不道德行为。

在第二节（P6），我们同样补充了一段总结：

> 总而言之，代理不仅直接影响决策者在决策链上的判断，也通过反馈链改变反馈者的评价与归责，并进一步影响决策者的事前预期。当决策者预期到责任可被代理分摊、反馈者更容易接受不道德行为时，决策者将更频繁地使用代理进行不道德行为。

以及（p4）：

> 综上所述，本文认为人工智能代理在反馈链中模糊了反馈者对决策者的意图判断。在面临经人工智能代理执行的不道德行为时，反馈者不确定是否应该对决策者施加惩罚，因而更可能采取温和保守的惩罚策略(Chevrier & Teixeira, 2024)，而这进一步影响了决策者使用人工智能代理执行不道德行为的倾向，加剧了不道德行为的发生率与强度(Gratch & Fast, 2022; Hamman et al., 2010)。

意见 5：作者试图探讨人工智能代理的特殊影响，但实际上很多影响并不特殊，例如“加剧不道德指令执行”，算法代理可能更加高度服从；“增强决策者可否认的机会”，人类代理也可以在不明确指令下做出不道德行为；扩大不道德行为的影响范围，算法代理可能也会；“增加反馈者的道德容忍”，代理几乎都会；“模糊反馈者对决策者的意图判断”，代理也几乎都会。建议作者进一步突出人工智能代理的特殊性。

答复：感谢审稿老师对人工智能代理特殊影响的意见。我们认同审稿老师的判断，即原稿所采用的递进式的比较逻辑（首先将人工智能置于更一般的“算法代理”范畴中，与人类代理进行比较，再进一步在算法代理内部区分人工智能与传统规则型算法），以及对人工智能代理特殊性描述的不足，造成了人工智能代理的特殊性不足。需要澄清的是，本文所强调的“人工智能代理的特殊性”并非指其产生了全新的道德影响，而是指其特定技术属性在既有代理机制中对作用强度、持续性与扩散边界产生了系统性的调节与重塑，与人类代理或传统规则型算法代理具有显著区别。因此，我们在修改稿中丰富了第三章（人工智能代理在决策链上的特殊性影响）以及第四章（人工智能代理在反馈链上的特殊性影响），通过在每个机制

下对比人工智能代理与人类代理以及传统规则型算法代理，突出了人工智能代理的独特影响。例如在“加剧不道德指令执行”（p7）中，我们在与人类代理对比的基础上加入了与传统算法的对比，论证人工智能代理如何进一步加剧不道德行为的执行：

> 第二，相比传统规则型算法而言，人工智能作为一种特殊算法进一步加剧了不道德指令的执行。传统规则型算法严格遵循设计者预设的运行逻辑规则，且只适用于特定场景(Bozdag, 2013; Mittelstad et al., 2016)。这使得决策者必须在程序中明确设置规则，以将不道德行为彻底执行。一旦决策者指令模糊、未闭环或外部环境发生变化，规则型算法则可能因触发逻辑漏洞或边界条件而中断。相比之下，基于机器学习、大语言模型等技术的人工智能代理能凭借其对决策者模糊指令的语义理解与目标拆解能力，自主推理并补全决策者的潜在意图，并转化为具体可执行的操作步骤(Köbis et al., 2025)。这使得决策者可以在仅知道决策目标(如“不惜一切代价提高收益”)而不知道如何执行不道德行为时，可以向人工智能代理下达抽象、模糊甚至仅具意向性的指令来完成不道德行为。

而在其他章节，我们进行了大量的修改，在之前已有对比的基础上，加入了与人类代理以及传统规则型算法代理的对比，论证人工智能代理如何进一步加剧不道德行为的执行。例如，在“4.1 增加反馈者的道德容忍”（p10-11）中：

> 反馈者对决策者的归责通常取决于其对决策意图、可预见性以及结果控制力的判断(Santoni de Sio & Mecacci, 2021)。然而，人工智能代理的介入，尤其是其行为生成的特殊机制与黑箱性，在反馈链上造成了更为突出的归责困境。
>
> 第一，相比与人类代理而言，虽然人工智能代理的介入类似地增加了反馈者判断决策意图的难度，但是增强的机制存在本质区别。首先，在人类代理情境下，反馈者在判断决策者意图和归责时，可以根据其对行为主体信念、能力、行为理由的评估(Cushman, 2008; Malle, 2021)，并结合其利益激励和社会规范等因素对道德决策链条上的多方参与者的意图做出判断并归责。而在人工智能代理情境下，人工智能代理的行为是基于数据训练的机器学习与数学优化的结果(Bender et al., 2021)，反馈者难以依据人类社会规范等知识对其形成准确的判断，从而难以推断出人工智能代理在决策中的参与程度以及是否具有意图(Bender et al., 2021)。其次，人工智能代理还具有多主体参与的特点。人工智能代理的构建与应用涉及开发者、训练者(通过输入数据训练模型)、使用者等多主体的参与(Santoni de Sio & Mecacci, 2021)。当不道德后果产生时，反馈者难以判断这究竟源于开发者算法架构的缺陷、训练者提供的数据偏差，还是决策者的恶意和使用。这种多主体结构使得反馈者归责更加碎片化(Constantinescu & Kaptein, 2025)，难以判断决策者的意图。

第二，与传统规则算法相比，人工智能代理输入—输出的黑箱性带来了更为突出的归责困难。传统规则型算法清晰的运行规则使得反馈者很容易将不道德行为的责任回溯到具体的规则制定及决策主体。当代理输入—输出具有确定性或决策者对结果具有较强控制力时，反馈者能清晰判断责任归属，在观察到不道德行为时将责任归于决策者(Bartling & Fischbacher, 2012; Oexl & Grossman, 2013)。而面对人工智能代理时，反馈者难以判断决策者是否对其道德决策结果可预见或可控制，由此反馈者难以依据行为结果推断决策者的真实意图。

综上所述，本文认为人工智能代理在反馈链中模糊了反馈者对决策者的意图判断。在面临经人工智能代理执行的不道德行为时，反馈者不确定是否应该对决策者施加惩罚，因而更可能采取温和保守的惩罚策略(Chevrier & Teixeira, 2024)，而这进一步影响了决策者使用人工智能代理执行不道德行为的倾向，加剧了不道德行为的发生率与强度(Gratch & Fast, 2022; Hamman et al., 2010)。

意见 6：未来研究部分，前两段没有参考文献，建议在已有研究的基础上进行展望。

答复：感谢审稿老师的提醒和建议。我们十分同意未来研究展望应尽可能建立在既有研究基础之上。根据该意见，我们已在修改稿的“未来研究”部分前两段补充了关键参考文献，使展望与既有证据更紧密对齐。如（p13-14）：

在作用顺序方面，可以在决策者—反馈者互动中操控反馈者(包括受影响者或观察者)的认知负荷(Zhao et al., 2024)、情绪水平(Igdalova & Chamberlain, 2025)、社会规范(De Groot et al., 2021)，……现有文献表明认知负荷会影响个体道德决策，导致个体更倾向功利主义决策(Liu et al., 2025)，因此认知负荷也可能改变反馈者的认知资源可用性，影响其道德判断，从而影响决策者对人工智能代理的采用和决策道德程度。……操控道德决策是否由人工智能代理完成，并系统调节人工智能的遵从程度、算法透明度(Wang & Qiu, 2024)与拟人化特征(Salminen et al., 2021)……在何种情境下可能产生“道德增强效应”(de Melo et al., 2019; Fernández Domingos et al., 2022)……

可以关注局部人群使用人工智能代理如何影响全局社会道德水平，并识别扩散与反馈机制(Alt & Gallier,2022; Bednar et al., 2025; Engl et al., 2021)……可探究人工智能代理的跨任务或长期影响(Gravert & Collentine, 2021)。

意见 7：文中有一些翻译痕迹过重的表达，如“此类委托—代理安排”，中文读起来比较怪，建议作者检查修改。

答复：感谢审稿老师对本文表达方面提出的细致建议。在修改稿中，我们已将“此类委托—代理安排”改写为“这种委托—代理模式”。

除此之外，我们对全文进行了逐句核查与语言润色，使文章更符合中文学术写作规范，并调整了若干类似表达。例如，我们将“转化为经由中介的间接道德决策”（p2）改为“转化为经由中介完成的间接道德决策”；将“在决策者与受影响者的关系维度上”（p4）改为“在决策者与受影响者的关系中”。这些修改对提升整体可读性与行文流畅性起到了重要作用。

最后，再次感谢审稿老师给出的宝贵建议。当然，受限于个人能力，修改稿中仍然难免有不妥之处，还望审稿老师不吝赐教。

……

审稿专家 2

该稿件聚焦人工智能系统作为“代理”（delegate）角色如何影响人类不道德行为，并将既有研究重组为两条相互关联的路径：以决策者为起点的“决策链条”，以及以受影响者/第三方为起点的“反馈链条”。在该框架下，作者强调代理介入会拉长决策与归责的链条，进而削弱道德感知与归因清晰度，从而提升不道德行为的发生概率；同时，AI 代理的黑箱性、高遵从性、规模化与工具性等特征可能进一步放大上述效应及其外溢范围。最后，作者提出三类未来研究方向：框架内机制关系与潜在的道德增强效应、不道德行为的外溢/扩散，以及人机协同治理工具与制度安排。为提升论文的可读性与逻辑清晰度，建议作者重点回应以下五点：

答复：感谢您审阅我们的论文并提出宝贵意见。我们认真梳理并逐条回应了您的建议，也据此对稿件进行了修改。您的建议极大地提升了论文整体质量。为便于您核对，我们在修改稿中用蓝色标出了新增或改写的内容；在本回复信中，您的原始意见以楷体呈现，修改稿中的对应修改则以方框摘录于此。

意见 1：证据强弱与机制对齐。作者按照“代理的一般性影响”与“AI 代理的特殊性影响”梳理了大量文献，但部分引用与小节标题之间的对应关系仍不够紧密。例如在“4.2.2 模糊反馈者对决策者的意图判断”中，所引研究未必直接检验“AI 会模糊反馈者对决策者意图判断”这一具体命题，而更多是为作者的推论提供间接支持。此外，同一 AI 特征（如黑箱性）在不同位置反复出现，暗示其可能通过多条路径发挥作用，但当前写法容易让读者难以发现同一特征是否具有多重机制。建议作者增补一张对齐表：以框架中的关键机制为行，AI

特征、对应研究的证据类型（直接检验/间接支持/理论推断或观点性讨论）等内容为列，进行系统梳理。同时，在正文中对间接证据采用更审慎的措辞，以减少混杂并增强论证透明度。

答复：感谢审稿老师对文献证据强弱和人工智能底层特点相关性的建议。我们完全认可审稿老师的两点核心观点：（1）部分机制无法得到所引文献的直接支持；（2）人工智能的某些特定特征（如黑箱性）能通过多种机制发挥作用。但受限于人工智能作为代理角色处于学术前沿，目前本文所提及的少量机制仍未得到直接检验。因此，我们在修改稿第三章“3. 人工智能代理在决策链上的特殊性影响”开头新增了一张机制—证据对齐表，对框架中各关键机制所对应的人工智能特征及其文献证据类型进行了梳理，明确标明了现有文献直接检验还是间接检验了该机制（p6）：

表 1 关键机制及对应人工智能特征和证据表

关键机制	人工智能特征	支持该观点的代表性文献	证据类型
加剧不道德指令的执行	高遵从性、无道德/声誉成本、强学习能力	Bozdag, 2013; Ivcevic et al., 2020; Köbis et al., 2021; Köbis et al., 2025; Mittelstadt et al., 2016; Parasuraman et al., 2000; Ram, 2025; Simon, 1997.	存在直接证据
增强决策者可否认的机会	强学习能力、黑箱性	Babic et al., 2021; Chevrier & Teixeira, 2024; Köbis et al., 2021; Köbis et al., 2025; Rahwan et al., 2019; Wang et al., 2025.	仅有间接证据
扩大不道德行为的影响范围	低增量成本的可复制性、个性化与跨场景迁移性	Babic et al., 2021; Babšek et al., 2025; Cadario et al., 2022; Caldwell et al., 2020; Gratch & Fast, 2022; Holmes & Tuomi, 2022; Rahwan et al., 2019; Zhai et al., 2021.	存在直接证据
增加反馈者对不道德行为的道德容忍	工具性、低社会预期性、低社会存在性、前沿与试验性	Bartling & Fischbacher, 2012; Bigman & Gray, 2018; Bigman et al., 2023; Chevrier & Teixeira, 2024; Dzindolet et al., 2003; Giroux et al., 2022; Glikson et al., 2020; Hong et al., 2021; Laakasuo et al., 2021; Lee & See, 2004; Malle et al., 2015; Maninger & Shank, 2022; Nass & Moon, 2000; Sullivan & Fosso Wamba, 2022; Sundar, 2008; Zhou et al., 2024; 许丽颖等, 2022.	存在直接证据
模糊反馈者对决策者的意图判断和归责	弱社会规范性、多主体参与性、黑箱性	Bartling & Fischbacher, 2012; Bender et al., 2021; Chevrier & Teixeira, 2024; Constantinescu & Kaptein, 2025; Cushman, 2008; Gratch & Fast, 2022; Hamman et al., 2010; Malle, 2021; Oexl & Grossman, 2013; Santoni de Sio & Mecacci, 2021.	仅有间接证据

注：直接证据指现有研究的发现或结论能直接支持该影响机制。而间接证据指现有研究仅提供了近似或相关的证据，未直接验证该影响机制，需要进一步的推理和判断。

除此之外，针对原稿中缺乏直接证据支持的机制，本修改稿采用了更为审慎的表述，避免将推论性结论表述为直接检验结果，如（p10）：

……有间接证据表明，有间接证据表明，相较于人类实施的不道德行为(例如招聘中的性别歧视)，人工智能实施的歧视会使反馈者表现出更弱的道德愤怒、责任与法律追责意愿(Bigman et al., 2023)以及惩罚欲(Maninger & Shank, 2022)。

意见 2：代理角色的重要性与特殊性需要更集中地界定。本文以“AI 作为代理”作为核心切入点，这一聚焦具有潜在价值；但现有文献也表明，AI 在道德决策中不仅可扮演代理，还可能作为建议者、协作者或道德榜样等多种角色。建议作者在正文的合适位置更集中地引用并对照相关综述，明确本文为何选择“代理”作为主轴：该角色在责任链条、意图归因、激励结构或可治理性上相较其他角色的关键差异是什么？进一步地，建议在引言或框架章节用一段话清晰交代本文的研究边界与定位，从而使读者理解“为何是代理、而不是其他角色”。

答复：非常感谢审稿老师对代理角色的重要性和特殊性所提出的关键性意见。我们完全认同审稿人指出的问题：在人工智能参与道德决策的研究中，人工智能确实可以扮演多种角色，例如建议者、协作者或道德榜样。若未明确阐述人工智能代理角色相对其他角色的理论重要性和意义，本文选取人工智能代理角色的合理性就大打折扣。根据该建议，我们在引言对研究定位进行了集中补充与澄清。我们解释了选择代理而非其他角色作为切入点的原因，即代理是唯一在结构上造成道德决策权与执行权相分离的角色形式（p2-3）：

在此背景下，人工智能系统作为代理角色如何影响人类决策的道德性，成为人工智能时代一个愈发重要但尚未充分回答的问题。尽管人工智能以多种角色(如建议者、协作者或道德榜样)参与并影响道德决策，本文认为代理角色是众多角色中唯一造成道德决策权与执行权分离的角色形式(建议者提供建议但不决策也不执行，协作者参与部分决策和执行)，代理的这一特性在物理和心理层面拉开了决策者与不道德后果的距离，从而产生了道德脱离(Bandura, 2017)、责任推诿(Köbis et al., 2021)、意图归因(Bazerman & Sezer, 2016)等深刻的道德问题(Gratch & Fast, 2022; Köbis et al., 2021)。因此，深入探究人工智能代理对道德决策的影响，对于理解人工智能时代的道德和伦理风险具有不可替代的核心价值。

意见 3：“是否委托 AI”作为内生选择的讨论有待加强。当前稿件更多回答“当 AI 介入后会发生什么”，但从心理学等学科交叉视角以及现实问题出发，一个同样关键的问题是：委托本身可能是决策者的策略选择，而非外生情境设定。由此引出值得讨论的议题：“哪些人/在什么情境下更倾向于把决策委托给 AI 代理？”建议作者在引言及决策—反馈框架章节增

加对这一组问题的系统化讨论，以凸显代理视角的解释力，并使框架更贴近现实中的制度与行为生成过程。

答复：感谢审稿老师指出“是否委托人工智能”可能是决策者的内生选择这一关键问题。我们认同该意见：当前稿件相对更集中于回答人工智能代理介入决策链和反馈链后会发生什么，以及这些影响如何重塑道德决策。而现实中的委托往往并非外生设定，而是决策者在特定制度与心理约束下的策略性选择。许多非道德决策领域的文献关注了决策者是否愿意将决策委托给一般代理(Holzmeister et al., 2023)或人工智能代理(Candrian & Scherer, 2022)，本文也在多出提及了在某些场景下“人们更倾向于将决策的执行委托给机器代理(Gratch & Fast, 2022; Paharia et al., 2009; Steffel et al., 2016)”。因此，审稿老师的这一意见有助于增强代理视角的解释力并使框架更贴近真实行为生成过程，我们在修改稿中系统性地补充了“委托选择的内生性”讨论。具体而言，在引言最后一句（p2）：

> 进一步地，决策者会更频繁地使用人工智能作为代理执行不道德行为(Candrian & Scherer, 2022)。

在“决策者—代理—反馈者”决策与反馈框架中，我们加入了一段（p5）：

> 在代理可用的情境下，决策者需要综合评估代理介入后对道德决策成本与收益结构的重塑：它既可能改变决策者自身对行为可接受性与责任边界的判断，也可能影响反馈者对该决策的评价、归责与反馈。上述因素将共同决定决策者是否选择委托代理，以及其最终采取何种(及何种强度的)道德决策。从一般性来看，不论采用何种形式的代理(如人类代理、传统规则型算法、人工智能代理)，……

以及在总结与展望章节，我们加入了一句（p12）：

> 当决策者充分预期到人工智能代理的影响后，其将更频繁地委托人工智能代理并进行更多的不道德行为。

意见 4：需进一步区分 AI 作为直接行为主体与作为代理时的责任结构。例如，在第 4.2.1 节中，作者引用了多项关于人工智能实施不道德行为时引发较弱道德谴责的研究，这些研究为理解公众如何评价 AI 行为本身提供了重要证据。然而，在这些研究情境中，人工智能往往被呈现为直接的行为主体或决策者，而非代表人类决策者执行任务的代理。建议作者进一步澄清，从“对 AI 行为的道德宽容”推论到“对通过 AI 代理而产生的不道德结果的宽容”

之间的逻辑联系，或明确指出这一外推所依赖的前提条件，以避免不同层级的责任结构与道德评价对象在论证中被混合。

答复：感谢审稿人对文献证据与推论的关键提醒。我们认同该意见指出的潜在混淆。的确部分文献证据呈现的是公众对人工智能作为行动主体时其行为本身带来的道德评价。而本文应该关注的是人工智能主体作为代理时，人们对人工智能代理的道德评价以及归责。为避免这一混淆，我们在修改稿中做了两项针对性调整。第一，本文在第三章和第四章的汇总表中，明确标注哪些证据属于间接证据（见正文表 1），以此提醒读者。

第二，在 4.1 中，我们将反馈者对人工智能行为的容忍总结为形成了一种相对宽松的道德评价环境，并由此加剧了不道德行为（p11）。

> 综上所述，本文认为人工智能代理在反馈链中增强了反馈者对不道德行为的容忍。在面临经人工智能代理执行的不道德行为时，反馈者更容易将不道德行为理解为技术性偏差、发展阶段性问题或非恶意后果，从而降低了负面情绪与惩罚动机。而这进一步提高了决策者不道德行为被接受和被容忍的概率，形成了一种相对宽松的道德评价环境，加剧了不道德行为的发生率与强度。

第三，在使用间接证据时，我们采用了更严谨的表达，例如(p10)：

> ……有间接证据表明，相较于人类实施的不道德行为(例如招聘中的性别歧视)，人工智能实施的歧视会使反馈者表现出更弱的道德愤怒、责任与法律追责意愿(Bigman et al., 2023)以及惩罚欲(Maninger & Shank, 2022)。

意见 5：建议微调逻辑结构。建议作者在第 2 节“决策者—代理—反馈者”决策与反馈框架中，先系统梳理道德决策情境下决策者、一般代理（人类代理与常规算法代理）以及反馈者之间的关系，为全文分析奠定清晰的理论基础。随后，第 3 节与第 4 节可分别从决策链与反馈链角度，展开对人工智能作为代理介入后所产生影响的讨论，并突出其相较于一般代理的特殊性。在此基础上，作者再结合第 3、4 节的分析，将人工智能代理的“特殊作用”进一步整合回整体理论框架之中，从而更清晰地凸显“人工智能代理通过反馈机制间接影响决策行为”这一潜在的理论贡献。

答复：感谢审稿老师提出的关于逻辑结构的建设性意见。我们同意审稿老师强调的逻辑主线，即先以“决策者—代理—反馈者”的基本关系奠定基础逻辑框架，再分别从决策链与

反馈链展开人工智能代理的特殊影响。根据该建议，我们在修改稿中对结构与行文顺序做了针对性调整：

第一，我们第 2 节呈现“决策链/反馈链”时：首先系统地梳理了“无代理的直接决策情境”下决策者面临的成本收益权衡；其次我们讨论了在代理可用的情景下，决策者如果采用代理会如何在决策链和反馈链上影响决策者的成本收益权衡，即代理的一般性影响。

第二，我们调整了全文的结构，将原本第二节（“决策者—代理—反馈者”框架）、第三节（决策链，包含一般性影响和特殊性影响）和第四节（反馈链，包括一般性影响和特殊性影响），调整为第二节（“决策者—代理—反馈者”框架，介绍框架及人工智能代理的一般性影响）、第三节（人工智能代理在决策链上的特殊性影响，详细从决策链介绍人工智能代理对道德决策的特殊性影响）和第四节（人工智能代理在反馈链上的特殊性影响，详细从反馈链介绍人工智能代理对道德决策的特殊性影响）。

意见 6：其他细微修改意见如下：术语一致性：文中上位概念与结果变量的表述需更统一，例如“道德决策/道德判断/道德性/道德后果/不道德行为”等在摘要与正文交替出现；“黑箱性/不透明度/透明性/可解释性”等亦需明确是否为同一维度的不同表述，或存在概念区分。建议在引言或方法性说明中给出用语约定并在全文保持一致。

答复：感谢审稿人关于术语一致性的提醒。我们已按照建议对全文核心概念与相关表述进行了系统性检查和统一梳理，以确保行文更为清晰严谨。具体而言：第一，对于“道德决策/道德判断”，我们在首次定义道德决策时，提及了道德判断并在后文中保持一致性（p2）：

> 本文关注道德决策，即个体做出具有道德与伦理后果的判断与选择的过程，该过程涉及评估潜在行动的正误(即道德判断)，并选择如何展开后续行动(Kouchaki & Smith, 2025)。

第二，我们在修改稿中删除了“决策道德性”的表达，转而统一使用“道德决策”。第三，对于“道德后果”、“不道德行为”等表达，我们检查并确保其使用时不具有歧义。第四，对于“黑箱性/不透明度/透明性/可解释性”等表述，我们在修改稿中将删除了对“可解释性”的使用。并在第一次出现“透明性”表达时，说明了其与“黑箱性”的负向关系，如（p8）：

> ……传统规则型算法通常基于开发者预先设定的显式规则与阈值执行决策，其输入—输出映射相对透明(黑箱性低)……

意见 7：首次出现的定义：按照作者界定，“反馈者”包含受影响者与第三方观察者。建议在术语首次出现处进行一次性定义，并在后文保持稳定指代，减少读者的语义负担。

答复：感谢审稿老师对术语界定与一致性的提醒。我们充分认识到术语一致性对行文流畅度和读者理解度的重要性。我们在摘要中对首次出现的“反馈者”一词补充了定义，明确指出其包括受影响者与第三方观察者。同时，我们对全文相关表述进行了统一校对：在后文将稳定使用“反馈者”作为统称，并在需要区分具体类型时分别使用“受影响者”或“第三方观察者”，以保持指代一致并降低读者的语义负担。

意见 8：概念层级更清晰：作者对“代理”的定义较明确，但后文出现“算法代理/人工智能代理/规则型算法代理”等层级术语时，边界略显模糊。建议用简短语句说明其层级关系与本文使用口径，以避免读者误读。

答复：感谢审稿老师关于术语使用规范的提醒。我们已按照建议在相关术语首次集中出现处补充了简短说明，以明确本文的层级口径并保持后文指代一致。具体而言，我们在引言中（p1）定义了：

> 所谓代理，是指代替决策者执行某一类具体决策的主体或技术系统(Köbis et al., 2021; 2025; Ross, 1973)。代理的类型多种多样，既包括下属、团队、外包机构等人类代理(Holzmeister et al., 2023)，也包括按照一套事先明确写定的规则和逻辑运行的传统规则型算法代理，以及具有学习能力的人工智能代理(Candrian & Scherer, 2022)……

意见 9：数值表述的准确与可核查：文中涉及比例或范围的描述（例如 AI 遵从率 60%至 95%且人类不足 40%）建议逐一核对原文，并在必要时说明该范围对应的实验条件或处理组，以确保表达准确、可追溯且不引发歧义。

答复：感谢审稿老师对数值表述准确性与可核查性的提醒。我们已按照您的建议对文中涉及比例、范围与对比关系的数值描述逐一回溯原始文献并进行核对，并对个别原先表述可能引发歧义之处，已在修改稿中做出更为精确的改写。

意见 10：关键推论的机制阐释有待补充。例如，在第 3.2.2 节中，作者将“模糊指令”与“决策结果的可否认性”直接相连，用以解释人工智能代理如何降低不道德行为的成本。然

而，当前文本中这一关系主要以推论形式呈现，对二者之间的心理或归责机制缺乏明确阐释，也较少直接引用相关文献作为支撑。建议作者进一步说明模糊指令如何在心理层面或责任归因层面增强决策者的可否认性，例如是否通过削弱不道德意图的可追溯性、模糊意图—结果之间的因果关联，或扩大自我辩护空间等路径发挥作用，并在可能情况下补充相应的理论或实证依据，以增强该推断的清晰性与说服力。

答复：感谢审稿老师关于推论机制阐释的建议。我们完全认可审稿老师的这一建议，在原稿讨论人工智能代理影响道德决策的多条机制中，部分论述逻辑的确存在不完整不全面的问题。在此修改稿中，我们进行了大幅修改，对关键关系进行了更为明确的阐述，并补充了对应的经验证据。例如，在“3.2 增强决策者可否认的机会”中，我们将“模糊指令”所带来的可否认性明确拆解为两个相互补充的机制环节：在输入端决策者免于明确表达不道德意图并减少指令层面的留痕，以及输出端决策者主张自身不可预见性和低控制性（p8-9）：

> 第二，相比于传统规则型算法而言，人工智能代理更显著地增强了决策者可否认的机会。传统规则型算法通常基于开发者预先设定的显式规则与阈值执行决策，其输入—输出映射相对透明(黑箱性低)、可复现且可审计，因而反馈者更容易将具体输出回溯到委托的规则设定与授权范围，因此决策者难以以不可预见或不可控制为由否认自身意图或责任。然而，正如前文所述，人工智能代理具有学习能力和输入—输出黑箱性的特征(Babic et al., 2021)，这在两方面提供了更强的可否认性。首先，这些特征允许决策者以抽象模糊的方式向代理传达不道德行为的指令，因此决策者并未明确下达不道德行为指令。间接证据表明，一旦外界环境允许决策者以含糊方式下达不道德决策指令(如在骰子实验中要求人工智能代理“帮我实现利润最大化”等抽象目标)，决策者的不道德行为会大幅增加(Chevrier & Teixeira, 2024; Köbis et al., 2025)。这些间接证据从侧面印证了允许模糊指令可能让决策者更容易否认自身不诚实的意图。其次，这些特征使决策者指令和道德结果之间的推理路径模糊，当决策者面临不道德行为产生的道德后果时，其可以通过主张自己无法充分预见或控制人工智能代理的具体输出而为自己的不道德行为辩护，否认自身意图，减少心理成本。

第二轮

审稿专家 1：作者对本人提出的问题做了回复，本轮我没有其他意见。

审稿专家 2：作者的修改认真细致，我没有补充意见

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。