《心理科学进展》审稿意见与作者回应

题目：人机协同决策中算法权威影响道德推脱的心理机制
作者：李春磊、辛素飞、盛靓

第一轮

审稿人 1 意见：

文章聚焦人机协同决策中算法权威对于道德推脱的影响，提出了“算法权威——道德推脱的双重认知评估”模型。选题具有一定的前沿性和实践意义，理论模型试图整合现有研究结果之间的矛盾，具有一定的理论整合价值。

总体回应：非常感谢审稿专家对本文的认可。接下来，我们将结合您的具体建议，对您提出的问题进行逐一回复。所有修改之处用红色字体标注。

意见 1：整体综述过于冗长，很多内容前后重复多遍，例如研究问题、算法权威的三大特征等等。建议作者大幅度精简内容，减少不必要的重复，使阅读更聚焦更流畅。

回应：非常感谢审稿专家对本文整体结构与综述表达提出的宝贵意见。原稿内容的确有些冗长，我们对此进行了内容和结构上的精炼。

（1）精简引言中关于研究背景和研究问题的重复表述。在修订稿中，我们仅在第 1 节引言集中阐述本文的核心问题，即“算法权威如何影响人机协同决策中的道德推脱”，后文不再重述研究背景，而是直接围绕概念辨析、理论分歧和模型建构展开。

（2）合并正文中关于算法权威三大特征的重复论述。在修订稿中，我们将算法权威的核心特征集中放在概念界定部分说明，后文仅保留与理论推导直接相关的表述。既保留了算法权威区别于一般技术工具和传统人际权威的关键内涵，也使后续论证更加简洁。

（3）删除文献中与模型建构关系较弱的引用。在修订稿中，我们保留了能够直接支撑本文理论模型的研究证据，更突出不同研究结论之间的分歧及本文提出双重认知评估模型的必要性。

（4）优化行文中的重复过渡与冗余说明。修订稿按照“现实问题引入—核心概念界定—已有研究分歧—理论解释不足—模型提出”的逻辑顺序调整结构，并优化了部分过渡句的撰写，使段落之间的衔接更加自然，减少因反复说明同一问题而造成的阅读负担。

上述修改已在正文第 1 节引言、第 3.1 节表 1、第 4.1 节、第 5.1 节逐一落实。

意见 2：对于算法权威的概念辨析需要进一步强化。例如在 2.1 部分，作者先说“算法权威最接近认识性权威”，但紧接着又说“算法权威与认识性权威之间存在本质差异”，这一转折令人疑惑。究竟是类型还是程度上的差异？算法权威是认识性权威的一种变体？还是一种完全不同的类型？

回应：非常感谢审稿专家指出这一关键问题。为更直接地回应“算法权威与认识性权威之间究竟是类型差异还是程度差异”的疑问，我们在第 2.1 节中对该问题进行了明确的理论定位，并将其作为贯穿全文的概念基础。具体论证逻辑如下：

算法权威与传统的认识性权威之间，既非完全的类型差异，也非单纯的程度差异。

首先，二者共享认识性权威的核心特征，即个体基于对外部判断来源可靠性的信任而接受其判断（Zagzebski, 2012），因此在类型上具有延续性。从这一意义上看，算法权威并不

属于与认识性权威完全并列、毫无关联的全新类型，而是认识性权威在算法化、自动化和制度化决策环境中的特殊表现形式。

然而，二者的信任来源有所不同。传统认识性权威依赖于可识别的专家身份、专业资历、知识共同体背书和人际信任关系；算法权威则更多依赖于技术系统的黑箱性、量化输出、自动化程序、平台规则或组织制度背书，以及人们对算法“客观、中立、高效”的技术想象（Beer, 2017; Gillespie, 2014）。这种信任来源的差异，使得算法权威在形成机制上呈现出与传统认识性权威不同的特征。

综上，算法权威与认识性权威之间属于“类型延续基础上的形态重构”，既非纯粹的类型差异，也非单纯的程度差异。

对应正文，我们已在第 2.1 节、第 3.2 节和第 4.1 节中进行了相应的衔接与强化论述。

意见 3：在道德推脱相关的论述中，存在理论模型与实证研究的断裂。首先，作者用一个小章节强调了聚焦状态性而非特质性道德推脱，但是并未正面说明状态性道德推脱为何在人机协同决策情境中具有优势，此外，是否后文中支撑模型的实证文章是否全都是状态性道德推脱？其次，作者在后文论证模型“矛盾证据”时，引用的很多文献在核心变量的操作化上存在根本差异，包括“选择性依从”、“算法厌恶”、“采纳”等，而非直接的道德推脱测量，这些变量究竟是否能整合进道德推脱这一变量？有何依据？这也是最为核心的问题。

回应：非常感谢审稿专家指出这一核心问题。我们完全认同您的观点，并已在返修稿中对相关章节进行了相应修改。以下将针对您提出的三个子问题逐一回应。

问题 1：关于状态性道德推脱在人机协同决策情境中的优势

关于“状态性道德推脱为何在人机协同决策情境中具有优势”这一问题，我们从以下三层逻辑进行说明：

第一，从理论定位看，道德推脱既可以是稳定的个人倾向，也可以是情境激发的即时状态。Bandura（1999）最初提出的道德推脱机制（责任转移、责任扩散、后果弱化、道德合理化等），本质上是针对具体情境中个体如何临时解除道德自我制裁的认知过程。Moore（2015）进一步明确区分了特质性道德推脱（跨情境的稳定倾向）与状态性道德推脱（具体情境中被激活的即时过程）。因此，选择状态性还是特质性，取决于研究问题是关注“个体一般性倾向”还是“特定情境中的认知变化”。

第二，从本文的研究问题看，我们关注的是“当人机协同决策产生不良后果后，算法权威如何改变个体对自身责任的理解”。这一问题的核心是责任归因的情境性变化，而非个体稳定的道德认知倾向。具体而言：算法权威对道德推脱的影响，并非因为某些人天生更容易推卸责任（特质性），而是因为算法介入本身改变了特定决策情境中的责任结构，即当算法被赋予权威地位时，个体在该情境中临时性地将责任转移给算法；当算法偏见被显化时，个体又可能临时性地回收责任。这种随情境条件（可解释性、道德冲突强度）变化而变化的责任归因，本质上是状态性的。特质性道德推脱无法解释“同一个人在同一类任务中，为何有时推卸责任、有时承担责任”这一现象。

第三，从方法适配性看，状态性道德推脱更适合采用情境操纵和即时测量。本文所引用的实证研究多采用情境实验范式，通过操纵算法权威水平、算法建议可解释性和任务道德冲突强度，观察个体在具体决策情境中的即时反应。这与状态性道德推脱的操作化路径一致，研究者往往通过将道德推脱量表改写为特定情境的陈述，来测量或操纵状态性道德推脱水平（Welsh et al., 2020）。

问题 2：关于后文实证研究是否均为状态性道德推脱

需要向您澄清的是，后文所引用的实证研究的确没有直接测量状态性道德推脱。事实上，现有实证研究更多关注算法情境下个体责任归因、责任转移、算法依赖或算法拒绝，直接检

验“算法权威中状态性道德推脱”的研究仍较少。这也是本文撰写的一个重要动机：通过理论整合，为后续直接检验奠定基础。

具体而言，现有研究可分类如下：

第一类，直接测量道德推脱或其具体机制。例如，Ebrahimi 和 Matt（2024）测量的是个体对歧视性数据分析建议的批准，以及有利比较、责任转移和去人性化等道德推脱机制；Gao 等人（2025）测量的是算法监控、责任转移和不道德行为之间的关系。这类研究可以直接说明，在算法辅助决策或算法化管理情境中，个体可能通过责任转移等道德推脱机制接受具有道德风险的建议或作出不道德行为。

第二类，测量责任归因或责任分配。例如，Dong 和 Bocian（2024）发现，在人类与 AI 共同卷入道德越轨的情境中，个体在评价自身越轨行为时会比评价他人越轨行为时更多地将责任归因于 AI；Lima 等人（2021）在 AI 辅助保释决策情境中发现，人们会区分 AI 与人类行动者的不同责任形式；Shank 等人（2019）也发现，当 AI 与人类共同参与道德错误决策时，人们会在 AI 和人类之间分配过错与道德责任。这类研究并不直接测量道德推脱，但能够说明 AI 介入会改变个体对“谁应承担责任”的判断，因此可作为责任转移机制的间接证据。

综上，现有研究虽未全部直接测量状态性道德推脱，但为理解算法权威如何通过改变责任归因影响道德推脱提供了重要基础。这也正是本文撰写理论模型与实证研究之间桥梁的核心价值所在。

问题 3：关于为何引入“选择性依从”“算法厌恶”“采纳”等相关变量

对于您关切的模型“矛盾证据”论证中引入相关变量的问题，我们认为这是必要且合理的，理由如下：

第一，这些变量是理解算法权威影响道德推脱的前置条件。例如，Logg 等人（2019）测量的“算法采纳”可以说明个体在何种条件下会将算法视为更可靠、更具权威性的外部判断来源。该变量本身并不是道德推脱，但只有当算法采纳进一步伴随“这是算法建议的”“我只是按照系统判断行动”等责任转移或责任扩散表述时，才可能与状态性道德推脱发生联系。因此，算法采纳是责任转移发生的前提行为指标。

第二，这些变量有助于界定道德推脱发生的边界条件。Dietvorst 等人（2015）测量的“算法厌恶”说明个体并不会无条件服从算法，而是会根据算法表现重新评估其可靠性。当个体不再将算法视为可靠判断来源时，其将责任外推给算法的可能性下降，因而更可能重新意识到自身判断责任，这正是“价值冲突与责任回收路径”的行为表现。

第三，这些变量与道德推脱中的核心机制直接相关。Bigman 和 Gray（2018）研究的“个体对机器作出道德决策的接受或抵触”，与道德推脱中的责任主体判断密切相关。如果个体认为机器缺乏心智和道德主体性，就更难将机器视为能够承担道德责任的主体，也更难完全通过把责任归因于算法来进行责任转移。Burton 等人（2020）的系统综述梳理了算法建议采纳或拒绝的影响因素，说明算法采纳、算法拒绝和算法信任会受到任务特征、算法表现、个体经验和决策风险等因素影响。这些变量可以作为理解道德推脱发生条件的前置证据：只有当个体在特定情境中采纳或依从算法建议，并进一步将行为责任外置于算法系统时，算法采纳或依从才可能被纳入状态性道德推脱的解释框架。

基于上述论证，我们在返修稿中明确区分了“直接证据”与“间接证据”，并指出：这些相关变量只有在进一步伴随责任转移、责任扩散或道德合理化时，才应被纳入状态性道德推脱的解释框架，而非简单等同于道德推脱本身。

针对上述三个问题，我们在返修稿中进行了如下修改：

（1）第 2.2 节：补充了状态性道德推脱在人机协同决策情境中优势的三层逻辑论证，明确其与特质性道德推脱的区别及适用条件。

（2）第 3.1 节中的表 1：在梳理矛盾证据时，明确区分了直接测量道德推脱的研究与测量责任归因等邻近构念的研究，并说明两者在证据链中的不同位置。该节分别讨论了样本差异、测量方式及操纵方式，但三类分析都反复指向责任归因方向不同这一结论，造成阅读冗余，所以我们总结成了表格，加入了对理解双重效应的启示。

（3）第 3.1 节中表 1 及第 3.2 节：在梳理矛盾证据时，明确区分了直接证据（直接测量道德推脱）与间接证据（测量责任归因/分配），并将算法采纳及算法厌恶等相关变量定位为理解道德推脱发生条件的前置行为指标或边界条件证据，而非直接等同于道德推脱本身。

意见 4：两条路径的推导存在一定的逻辑跳跃。首先，服从权威理论基于真实的人际权威压力，算法是否能够产生类似的权威压力？作者需要论述其为何能适用于人机情境。其次，心智感知理论的推导中，从“算法缺乏心智”到“自我责任归因”也需要更详细的论述。

回应：非常感谢审稿专家指出这两处逻辑跳跃。针对您的问题，我们进行了如下补充和修改。

（一）关于服从权威理论为何适用于人机情境

您提出的核心追问是：算法不是人，没有真实的人际权威压力（如职位、制服、命令口吻、惩罚能力），为何能用服从权威理论来解释？我们的回答分为以下两步：

第一步，明确服从权威理论的核心机制是什么。

Milgram（1963）的服从权威研究表明，个体之所以服从权威，核心机制并非仅仅因为害怕惩罚，而是因为个体在权威情境中会进入代理状态（agentic state），也就是说将自己定位为执行他人意志的工具，从而弱化对自身行为的责任感知（Götz et al., 2023）。换言之，服从权威理论的核心不在于权威的人际性，而在于权威能否改变个体对自身责任主体的定位。

第二步，论证算法如何在功能上对应于这一核心机制。

算法虽然不具有人际权威的职位或制服，但在人机协同决策中，算法建议常以以下形式出现，从而在功能上产生类似的代理状态：(1)**指令性输出**。算法以推荐、最佳方案、系统判定等形式呈现建议，而非仅仅是信息提示；(2)**默认选项**。算法建议常被设置为默认选项，个体需要主动选择才能拒绝；(3)**组织流程嵌入**。算法建议被嵌入任务分配、绩效评价、资源调配等组织流程中，形成制度性权威；(4)**专业性背书**。算法被认为具有更强的数据处理能力和客观性，被赋予认识性权威地位。

在这些条件下，个体可能像面对人际权威一样，将自己定位为执行算法建议的操作者，而非自主决策者，从而进入代理状态。已有研究也支持这一推论：Ghasemaghaei 和 Kordzadeh（2024）发现，面对不公正的算法推荐，管理者会做出歧视性决策且内疚感降低，作者据此指出管理者可能将自己理解为在服从算法这一权威来源；Jolly 等人（2025）发现，当算法以更强指令性介入决策时，个体会表现出更明显的责任位移。

（二）关于算法缺乏心智到自我责任归因的推导

您提出的核心追问是：为什么算法缺乏心智会导致个体自我责任归因增强？逻辑链条中的中间环节是什么？我们的回答拆解为以下逻辑步骤：

第一步，个体判断算法的心智水平。心智感知理论指出，个体会根据能动性（思考、行动能力）和体验性（感受情绪的能力）两个维度判断一个实体是否具有心智（Gray et al., 2007; Gray et al., 2012）。算法通常被认为具有中等水平的能动性（能够计算、预测、推荐），但缺乏体验性（无法感受公平、伤害、共情）。

第二步，个体基于心智判断评估算法能否成为道德责任主体。道德责任主体的前提是能够理解行为后果的道德意义并承担相应责任。如果个体认为算法缺乏体验性（无法感受伤害、不理解公平），就更难将算法视为一个能够承担**道德责任**的完整主体。算法或许可以承担因果责任（是算法算错了），但难以承担道德责任（算法不会内疚）。

第三步，个体意识到自己仍需承担最终道德责任。当个体判断算法不能成为道德责任主体时，责任就无法完全外移。即使个体采纳了算法建议，最终仍需由人类决策者承担道德责任。这一判断促使其重新确认自身责任主体地位。

第四步，自我责任归因增强。在上述认知过程的基础上，个体的自我责任归因水平提高，从而抑制了通过责任转移或责任扩散来实现的道德推脱。

已有研究也支持这一逻辑链条：Bigman 和 Gray（2018）发现，人们抵触算法作出道德决策，正是因为算法被感知为缺乏心智和道德主体性；Shank 等人（2019）发现，当 AI 与人类共同参与道德错误决策时，人们会在 AI 和人类之间分配过错与道德责任；Joo（2024）发现，当 AI 被赋予更强心智特征时，参与者会将更多责任归因于 AI。

上述修改已在正文第 3.2.1 节及第 3.2.2 节逐一落实。

意见 5：图 1 的论述段中，作者最后说道“可解释性与道德冲突强度的关键之处也不在于，它们会改变个体是否能够继续把自己理解为最终责任主体。”按照逻辑来说，应该是“正在于”？

回应：非常感谢审稿专家的细致审阅与指正。该处确为我们在表述上的疏忽。根据上下文逻辑，本文原意是说明：算法可解释性与道德冲突强度之所以重要，正在于它们会影响个体是否能够继续将自己理解为最终责任主体，进而影响算法权威作用于道德推脱的方向与强度。原稿中误写为“关键之处也不在于”，造成了与前文论证相反的含义，容易引起读者误解。

根据您的建议，我们已将该句修改为：“算法权威影响道德推脱的关键在于改变了个体对‘谁是最终责任主体’的主观定位”。同时，我们已对图 1 前后的论述段进行了再次通读和校对，重点检查否定词、转折词及逻辑连接词的使用，避免类似表述错误影响模型阐释的准确性。

意见 6：未来研究方向过于泛化，尤其是神经生理技术部分占比很大，但是正文中没有相关推导，因此这一研究方向建议与模型的逻辑关联性不足，显得比较泛化和套话。

回应：非常感谢审稿专家的建议和指正。针对您的建议，我们已对返修稿的未来研究方向部分进行了精简、重写和重新聚焦。

围绕本文提出的理论模型展开，后续研究的重点应放在模型关键机制的检验上，包括：算法权威是否会增强个体对算法建议的依赖，并进一步诱发责任转移、责任扩散和道德合理化；算法低心智感知是否会削弱个体将责任完全外推给算法的合理性，从而促使个体重新确认自身责任；算法可解释性、道德冲突强度等变量是否会调节两条路径的相对强度。

在研究方法上，我们已明确提出未来研究应优先采用情境实验、行为决策任务和状态性道德推脱测量来检验模型。具体而言，后续研究可以通过操纵算法权威水平、算法建议可解释性和任务道德冲突强度，观察个体在责任归因、责任转移、后果弱化和道德合理化等方面的即时变化。这样既能直接回应本文理论模型，也能避免将未来研究方向过早转向与正文推导关系不够紧密的技术方法。

关于神经生理技术部分，我们已进行了精简和定位调整。已删除大部分概括性表述，仅将其作为未来研究中的辅助性方法。神经生理技术只有在未来研究需要进一步捕捉算法权威影响责任判断、道德冲突加工和即时认知反应时，才可作为补充工具，而不能替代行为实验、问卷测量和情境操纵等更直接的模型检验路径。神经技术可以作为辅助验证方法，但不能替代对核心心理机制的直接测量。因此，我们在返修稿中将神经生理技术定位为“辅助验证工具”，而非独立的重点研究方向。

此外，我们已调整未来研究部分的表述方式，删除了与本文主题关联较弱的泛化表达，如笼统强调“多模态技术全面揭示机制”等表述。修订后的展望更加集中于三个问题：一是

如何直接检验双重认知评估模型；二是如何明确算法可解释性、道德冲突强度等边界条件；三是如何在必要时谨慎引入神经生理技术作为辅助证据。通过上述修改，未来研究方向已从原先较泛化的技术方法展望，调整为围绕模型可检验性的研究展望。

对应正文，我们已在第 5.2 节未来研究方向中进行了相应的修改。

意见 7：论文的倒数第二段引用的（Rismani et al., 2025）有含有 gpt 痕迹的超链接：https://www.nature.com/articles/s41597-025-06021-5?utm_source=chatgpt.com#auth-Shalaleh-Rismani-Aff1。这涉及学术伦理规范问题，需要作者作出说明。

回应：非常感谢审稿专家细致且严谨的审查。关于正文文内引用“（Rismani et al., 2025）”所内嵌的超链接中包含“chatgpt.com”追踪参数的问题，我们完全理解您的关切，并在此作出诚恳说明与道歉。

在本文前期文献调研阶段，我们曾将大语言模型作为**辅助检索工具**，用于拓展关键词和提示相关研究线索。ChatGPT 在检索中提示了 Rismani et al. (2025)这篇文献。随后，我们通过该链接跳转至期刊官网，对文献的题名、作者、发表来源和原文内容进行了重新核验、下载与阅读。该文献被正式引用（Rismani, S., Davis, L., Mingole, B., Rostamzadeh, N., Shelby, R., & Moon, A. (2025). Responsible AI measures dataset for ethics evaluation of AI systems. *Scientific Data*, *12*(1), 1980.），其依据是我们在期刊官网上核验后的原始版本，内容真实、准确，不存在任何虚构或不当引用。我们根据专家对于未来研究方向的意见进行了修改，同时 Rismani et al. (2025)这篇文献已在文中删除。

稿件撰写过程中，为了便于在不同修改轮次中快速回溯原文，我们在 Word 文档中的文内引用处临时保留了从浏览器地址栏复制的原始网页链接作为内嵌超链接。由于该链接来自前期检索跳转页面，因此浏览器地址栏中自动附带了“utm_source=chatgpt.com”等追踪参数。后期校对时，我们重点核查了正文表述、文内引用与文末参考文献的一致性，但对 Word 文档中隐藏的内嵌超链接参数检查不够充分，未能及时删除无关格式，从而导致了这一疏忽。

我们始终遵守学术伦理规范。参考美国心理学会关于生成式人工智能使用的相关政策要求（网址：https://www.apa.org/pubs/journals/resources/publishing-tips/policy-generative-ai），其核心原则为：**作者需要对生成式 AI 工具提供的信息负责，并应通过阅读原始来源进行核验。**在本文的撰写过程中，AI 工具仅被限定用于前期文献线索提示和检索方向辅助；所有进入正文论证的文献，均由作者亲自进入期刊官网或正式数据库核验；本文的理论框架、模型建构、论证逻辑及最终定稿均由作者团队独立完成。因此，**本问题为作者在引文格式上的校对疏漏，不涉及文献来源错误、虚构文献或任何形式的学术不端行为**。

针对此问题，我们已在修改稿中彻底清除了所有无关网页追踪参数。同时，我们引以为戒，对全文所有文内引用、文末参考文献及网页链接进行了逐项交叉核查，确保所有引用真实可溯、格式规范。再次感谢您的提醒，这一意见帮助我们完善了文献管理与终稿校对流程，使本文的规范性更加严谨。最后，感谢您在审稿过程中给予的耐心与宽容，希望在您的审稿意见的引导下，我们的修改能使本文的理论深度和论证充分性都得到实质性的提升。

……

审稿人 2 意见：

本文选题具有较强的时代价值，精准识别了“算法权威对道德推脱产生双重效应”这一尚未得到系统解释的理论矛盾，并尝试以自我责任归因为核心机制进行整合，体现出对该领域核心争论的敏锐把握。文献引用数量可观，覆盖了近年多个高质量期刊的前沿成果，综述工作量较为充分。

总体回应：非常感谢审稿专家对本文的认可。也十分感谢您针对本文提出的建议。接下来，

我们将对您关心的问题进行一一回应，并在正文中用红色标示出了相应修改内容。

意见 1：文章对算法权威与道德推脱两个核心概念的经典文献覆盖较为充分，且引用了高质量期刊来源的最新文献。然而，文章在以下几个方面存在明显不足：（1）关于道德推脱在东亚文化情境下的差异性（集体主义文化中责任扩散可能具有不同机制），文章基本未有讨论，缺乏对中国情境下算法伦理的实证文献支撑。（2）算法可解释性领域，可解释人工智能领域有若干系统性综述（如 Adadi & Berrada, 2018; Molnar, 2020），文章未加以引用或讨论，这对于将可解释性作为核心边界条件的论文而言是明显缺口。（3）在讨论认知卸载时，文章未能充分引用自动化偏见研究的奠基性文献，以及近年围绕 over-reliance/under-reliance 的元分析研究（如 Schemmer et al., 2023）。文献缺口降低了综述的权威性和完整性，部分论点的证据链存在薄弱环节，影响读者对理论模型可靠性的判断。

回应：非常感谢审稿专家对本文文献综述完整性和证据链充分性的细致提醒。我们完全认同您的意见。针对这一问题，我们已从以下三个方面对文章进行了补充和修订。

（1）关于道德推脱的东亚文化差异的问题。跨文化心理学研究表明，东亚文化中的自我理解更强调关系性、情境性和相互依赖性，个体责任判断也更容易嵌入群体、组织和关系结构之中（Markus & Kitayama, 1991）。因此，在集体主义文化中，责任扩散可能不仅表现为“我不是唯一责任人”，也可能表现为“我是按照组织、系统或群体要求行动”。当算法被嵌入组织管理、平台规则或公共治理流程时，个体可能更容易将算法理解为制度化权威，将算法视为自我的延伸，并更容易接受算法的决策（Barnes et al., 2024）,并通过服从系统要求来降低自我责任归因。其次，我们已补充中国情境下算法伦理和算法管理的实证文献。例如，Gao 等人（2025）关于零工劳动者的研究发现，算法监控会通过责任转移提高不道德行为，这为中国算法管理情境中“算法介入—责任转移—不道德行为”的关系提供了直接证据。通过这一修改，本文进一步说明了算法权威影响道德推脱并非脱离文化背景的普遍过程，而可能在集体主义责任结构中表现出更强的效应。

（2）关于算法可解释性的文献补充。Adadi 和 Berrada（2018）指出，可解释人工智能的重要价值在于回应黑箱模型所带来的透明性、信任和责任追究问题；Molnar（2020）系统梳理了机器学习模型解释方法，强调可解释性使人类能够理解模型如何形成预测；Arrieta 等人（2020）进一步将可解释性与负责任人工智能、透明性和可问责性联系起来。基于这些文献，我们在修订稿中进一步明确：本文所说的可解释性并不仅是技术层面的模型说明，而是影响个体能否理解、审查、质询和评估算法建议的重要条件。**当算法建议缺乏可解释性时**，个体更可能将算法输出视为不可质疑的“系统结论”，从而强化算法权威和责任外移；**而当算法建议具有较高可解释性时**，个体更容易意识到自己仍需对是否采纳算法建议作出判断，因而更难完全通过“系统建议如此”来转移责任。

关于认知卸载的文献补充。Mosier 等人（1996）的研究表明，在自动化系统给出建议时，个体可能出现遗漏错误和执行错误，即由于系统未提示风险而忽视问题，或即使系统建议错误仍然采纳其建议。Parasuraman 和 Manzey（2010）则将自动化偏见概括为个体过度依赖自动化建议、减少独立核查和主动验证的倾向。Schemmer 等人（2023）进一步提出适当依赖的观点，强调人机协同决策的理想状态并不是简单信任算法或拒绝算法，而是能够采纳正确建议并拒绝错误建议。基于这些文献，我们已将原稿中较笼统的“认知卸载”论述修改为更具体的机制链条：算法权威提高个体对算法建议的依赖，过度依赖削弱个体对算法建议的主动验证，主动验证的减少进一步降低个体对决策过程的参与感和责任感，最终为责任转移和道德推脱提供心理条件。

上述修改已在正文第 2.2 节、第 4.2 节和第 3.2.1 节中逐一落实。

意见 2：文章对所引文献的批判性讨论明显不足。在第 3.1 节梳理矛盾证据时，文章大量采用“某研究表明……另一些研究表明……”的并列式罗列结构，缺乏对以下关键问题的深入分析：（1）不同研究的样本差异（学生样本 vs. 劳动者样本）是否导致了结论的分歧？（2）这些研究的道德推脱测量方式（量表测量 vs. 行为指标）是否具有可比性？部分研究并非直接测量道德推脱，而是测量责任感、算法依从性或信任等相关但不等同的变量，文章在引用时未作区分，存在将邻近构念混用的问题（如将减少内疚感等同于道德推脱提高）。（3）算法权威的操纵方式（来源标签操纵 vs. 指令强度操纵）在不同研究间的可比性问题未予讨论。此外，文章在分析“算法厌恶”研究时（第 3.2.2 节），对 Burton 等人（2020）的系统综述的引用方式较为机械，缺乏批判性评估该综述所指出的研究局限。批判性程度不足使得文章更接近知识汇编而非批判性综合，削弱了由此推导出理论模型的说服力。

回应：非常感谢审稿专家指出本文批判性综合不足的问题。我们完全认同您的意见。针对这一问题，我们已对第 3.1 节和第 3.2.2 节进行了相应修改。

首先，关于不同样本差异是否可能导致研究结论分歧。修订稿明确指出，学生样本并不长期受算法管理，也不需要在真实组织关系中承担持续性后果。因此，这类研究更适合揭示个体面对算法建议时的即时采纳、拒绝或态度反应。例如，Logg 等人（2019）和 Dietvorst 等人（2015）的研究分别说明个体可能因算法被认为更客观、更准确而采纳算法，也可能因算法出错而产生算法厌恶。但这类研究较难直接说明算法权威如何在真实组织结构中诱发责任转移和道德推脱。相比之下，劳动者样本通常处于更明确的任务分工、绩效评价和责任约束之中，算法不仅是信息来源，也可能成为任务分配、行为监控、绩效评价或决策背书的制度化力量。Gao 等人（2025）关于网约车司机和外卖员的研究表明，算法监控能够通过责任转移提高不道德行为，这更能说明算法权威在真实或类真实组织场景中如何嵌入责任分配并诱发道德推脱。由此可见，现有研究中关于算法权威是否诱发道德推脱的结论分歧，在一定程度上与样本类型有关：基于学生样本的研究更多揭示算法采纳或拒绝的即时态度，而基于劳动者样本的研究则更直接地观察到责任转移和道德推脱的发生。这一差异提示，未来研究应重视样本所处的真实责任结构对结论的影响。具体修改详见第 3.1 节。

关于测量方式的可比性这一问题。现有研究中，量表测量（如道德推脱量表）反映个体自我报告的一般性倾向，而行为指标（如实际的责任转移行为）反映具体情境中的即时反应。两者虽不完全可比，但本文聚焦状态性道德推脱的情境性特征，因此更侧重引用采用情境操纵和行为指标的研究，并主要关注其中与责任转移、责任扩散直接相关的发现。

另外，关于邻近概念混用问题。我们认同审稿专家的批评，原文确实存在这一不足。针对这一问题，我们在修订稿中明确区分了直接证据与间接证据：部分研究直接测量道德推脱或其具体机制（如 Ebrahimi & Matt, 2024；Gao et al., 2025），可作为本文论证的直接证据；另一些研究测量的是责任归因或责任分配（如 Dong & Bocian, 2024；Lima et al., 2021；Shank et al., 2019），能够说明 AI 介入会改变个体对“谁应承担责任”的判断，因此可作为责任转移机制的间接证据。通过这一修改，本文明确说明“减少内疚感”“算法依从”或“算法信任”等变量本身并不等同于道德推脱，只有在进一步伴随责任转移、责任扩散或道德合理化时，才可能被纳入状态性道德推脱的解释框架，从而避免了邻近概念混用的问题。具体修改详见修改稿第 3.1 节。

关于算法权威操纵方式的可比性问题。来源标签操纵通常是在建议内容保持不变的情况下，仅改变建议来源的标识，例如将同一建议标记为“算法生成”或“人类专家生成”，以考察个体是否因为建议来自算法而更愿意采纳它（Logg et al., 2019）。这类操纵主要反映个体是否认为算法比人类更客观、更准确或更值得信任。相比之下，指令强度操纵改变的是算法建议对个体行为的约束程度，例如算法是否以强推荐、默认选项、系统要求或组织流程指令的形式出现。这类操纵关注个体是否在算法建议具有更强约束力时弱化自主判断、降低自

我责任归因，甚至将自己定位为执行系统要求的人（Ghasemaghaei & Kordzadeh, 2024; Jolly et al., 2025）。因此，两类操纵虽然都涉及算法权威，但对应的心理机制并不相同：前者主要检验算法作为判断来源是否更容易被信任和采纳，后者则检验算法建议是否通过更强的指令性、默认性或组织嵌入性改变个体的自主感和责任感。

关于 Burton 等人（2020）算法厌恶系统综述的使用方式。在修订稿中，我们进一步将 Burton 等人（2020）的研究定位为“算法权威并非总是稳定有效”的背景性证据。该综述的重要意义在于提示：个体是否采纳算法建议，会受到任务风险、算法表现、错误经验、决策责任和使用者专业能力等因素影响。也就是说，算法厌恶并不是对自动化系统的一般性反感，而是在特定情境中，当个体认为算法表现不可靠、缺乏情境理解能力，或不适合处理高风险与高道德冲突任务时，对算法建议产生的不信任和拒绝。基于这一理解，我们在第 3.2.2 节中不再将算法厌恶直接等同于道德推脱的下降，而是将其视为“价值冲突与责任回收路径”在行为选择层面的表现：当个体对算法建议产生怀疑并降低依赖时，才可能进一步重新审查决策责任，并恢复自身作为责任主体的地位。

上述修改详见正文第 3.1 节表 1（区分样本差异、测量方式、操纵方式）和第 3.2.2 节（将 Burton 等人 2020 的算法厌恶综述定位为背景性证据，不再直接等同于道德推脱）。

意见 3：文章的论证主线基本清晰，但在以下环节存在逻辑跳跃：（1）从第 3 节（实证矛盾）到第 4 节（模型提出）的过渡处，文章未能从实证证据中系统归纳出为何自我责任归因是连接算法权威与道德推脱的唯一或最核心的机制变量，而非情绪调节、认知负荷等其他可能的中介机制。（2）第 4.1 节提出算法权威通过影响自我责任归因来影响道德推脱，但未说明为何低自我责任归因 → 高道德推脱这一方向性关系是单向的、线性的，而非存在调节效应或非线性关系。Joo（2024）关于人机共享责任的研究实际上已经提示了这一关系可能更为复杂。（3）模型图（图 1）中包含多个符号（+/-），但文章正文并未对所有标注的关系进行同等深度的论证。

回应：非常感谢审稿专家的建议和指正。针对您的意见，我们已做如下修改：

首先，关于为何自我责任归因是核心机制的问题：道德推脱的本质是责任归因的重新定位（Bandura, 1999; Moore, 2015），而非情绪或认知负荷变化。实证研究也表明，算法介入最稳定地改变了责任归属判断（Ebrahimi & Matt, 2024; Gao et al., 2025）。算法权威最稳定、最直接影响是责任主体的归因判断。因此，本文选择自我责任归因为核心机制，旨在说明算法权威如何通过改变个体对自身责任的认知定位，进一步影响道德推脱水平。

其次，关于低自我责任归因到高道德推脱的关系是否为线性的问题：具体而言，在算法权威被个体理解为合法、专业且具有正当性的人机协同决策情境中，低自我责任归因更可能转化为道德推脱；但当个体意识到算法可能存在偏差、决策后果具有较高道德风险，或自身仍保留最终决策权时，低自我责任归因未必直接导致道德推脱，甚至可能激发责任回收与道德警觉。Joo（2024）显示责任分配存在人机共享结构，因此该关系应理解为有条件的理论预期，而非无条件线性关系。

再次，关于图 1 符号未论证的问题：我们已删除图 1 中的+/-符号，并在正文中补充路径解释。修改后的图 1 不再直接呈现简单正负关系，而是突出算法权威经由自我责任归因、价值冲突和责任回收等路径影响道德推脱的过程性机制。

上述修改已在正文第 4 节开头（论证自我责任归因为核心机制）、第 4.1 节（明确低自我责任归因与高道德推脱为有条件的理论预期）和第 5.1 节（删除图 1 中+/-符号并补充路径解释）中逐一落实。

意见 4：核心变量界定模糊：（1）关于道德冲突强度，文章在第 4.3 节中将其定义为个体在

决策时面临道德价值观冲突以及由此产生的心理紧张程度（引自 Chen et al., 2021），但原文的 Chen et al.（2021）出处系伦理行为期刊，其对道德冲突强度的操作化定义是否与人机协同情境相容，文章未作充分说明。此外，道德冲突强度与道德议题的强度是概念上接近但有所区别的构念，文章对二者的区分不够明确。（2）关于自我责任归因，文章引用 Newman et al.（2020）将其定义为将行为责任主要归因于自身而非外界环境或他人的一种认知倾向，但此定义更接近于内部归因倾向这一稳定人格特征，而非本文所需要的情境激活的责任感知，文章对这一概念模糊性缺乏讨论。

回应：非常感谢审稿专家指出核心变量界定不够精确的问题。针对这一问题，我们已重点修改第 4.1 节和第 4.3 节，并在模型说明中统一了相关概念表述。

关于道德冲突强度，我们已在第 4.3 节重新界定该概念。修订稿已明确说明，我们进一步区分了“道德冲突强度”和“道德议题强度”。道德议题强度强调某一伦理议题本身的道德严重性，例如后果严重性、社会共识、后果发生概率、时间紧迫性、接近性和影响集中度等（Jones, 1991）。换言之，道德议题强度关注的是“这个问题本身有多严重”。而本文所说的道德冲突强度，更强调个体在具体决策过程中感受到的价值冲突和心理紧张，即当算法建议与个体原有道德判断、责任理解或价值立场不一致时，个体是否感到“应该服从算法”与“应该坚持自身道德判断”之间的冲突。换言之，道德冲突强度关注的是“个体在这一人机协同决策中感受到多强的道德拉扯”。基于这一区分，本文保留道德冲突强度这一变量，是因为人机协同决策并不只是面对一个抽象的道德议题，而是面对“算法建议—人类判断—责任承担”三者之间的张力。已有伦理冲突研究表明，当决策情境涉及不同道德价值或规范要求之间的冲突时，个体会产生更强的负面情绪、目标困难感和道德认知投入（Brown et al., 2022）。在人机协同情境中，这种冲突可能表现为：算法建议看似更高效、更客观，但其结果可能违反公平、伤害避免或责任承担等道德要求。因此，道德冲突强度并不是简单等同于议题本身的道德严重性，而是指个体在采纳或拒绝算法建议时所体验到的道德价值冲突程度。为避免概念混淆，我们已在第 4.3 节中将原稿表述修改为：道德议题强度反映伦理问题本身的严重性，道德冲突强度则反映个体在人机协同决策中因算法建议与自身道德判断不一致而产生的主观冲突体验。本文关注后者，因为它更能解释算法权威在何种条件下会诱发责任转移或促使个体重新确认自身责任。此外，Chen 等人（2021）的研究主要关注道德认可、股东价值取向和道德议题强度对道德判断与道德推脱的影响，其“道德冲突强度”相关表述更接近于道德议题本身强弱或伦理问题严重性的讨论，而不是专门对个体在决策中体验到的伦理冲突进行概念化和测量。也就是说，该文献虽然与道德判断和道德推脱相关，但其操作化并不能充分回答本文所需要解释的问题：在人机协同决策中，当算法建议与个体自身道德判断或责任理解发生冲突时，个体如何体验并处理这种价值拉扯。

因此，修订稿将该处主要依据改为 Brown 等人（2022）关于道德冲突的研究。相比 Chen 等人（2021），Brown 等人（2022）更直接地对“伦理冲突”进行了概念化、测量和后果检验，其核心关注的是个体在面对相互竞争的伦理要求、价值目标或行为规范时所体验到的心理紧张和决策困难。这一定义与本文所说的“道德冲突强度”更为契合，因为本文关注的并不是某一伦理议题本身有多严重，而是个体在人机协同决策中，面对算法建议与自身公平判断、伤害避免原则、责任承担意识或职业伦理要求之间的不一致时，主观上感受到多强的价值冲突和心理紧张。

在此基础上，我们进一步将 Brown 等人（2022）的伦理冲突概念情境化到人机协同决策中。人机协同决策中的道德冲突并不只是一般意义上的伦理问题严重性，而是算法建议、人类判断和责任承担之间的张力：一方面，算法建议可能因其专业性、客观性和效率优势而被个体视为值得服从的判断来源；另一方面，当算法建议涉及歧视、不公、伤害或责任风险时，个体又可能意识到算法不能替代人类完成价值判断和责任承担。因此，本文将道德冲突

强度重新界定为：个体在人机协同决策中，因算法建议与自身道德判断、责任理解或价值立场不一致而体验到的主观冲突程度。

同时，我们也在修订稿中补充了更贴近人机协同情境的研究作为辅助依据。van der Waa 等人（2021）关于人机团队道德决策的研究表明，当人工智能代理参与道德决策时，个体不仅需要评估算法建议本身，还需要判断是否应保留人类控制、如何理解算法解释，以及是否由人类承担最终责任。Fabre 等人（2024）关于人工智能顾问参与道德困境决策的研究也表明，AI 提供的功利主义或义务论理由会影响个体的道德选择，说明算法建议可能直接介入个体的价值权衡过程。基于这些文献，我们在修订稿中不再直接沿用 Chen 等人（2021）的操作化表述，而是以 Brown 等人（2022）关于伦理冲突的概念化为基础，并结合人机协同道德决策研究，将道德冲突强度界定为人机协同决策中由算法建议引发的主观价值冲突和心理紧张。

关于自我责任归因。本文中的自我责任归因不再被理解为稳定人格特质，而是指个体在特定人机协同决策中，对自己是否应为算法辅助下的选择、行为及其后果承担责任的即时判断。也就是说，它关注的不是一个人一般情况下是否倾向于把事情归因于自己，而是当算法建议进入具体决策过程后，个体是否仍然认为“这一决定是我作出的”“我对采纳或拒绝算法建议负有责任”“后果仍与我的判断有关”。

这一区分对于本文模型非常关键。稳定的内部归因倾向强调个体跨情境的归因风格，而情境性自我责任归因强调责任结构如何被具体决策环境激活或改变。在人机协同决策中，算法建议的权威性、可解释性、默认推荐方式以及最终决策权分配，都会影响个体是否将自己理解为最终责任主体。AI 责任归因研究也表明，当 AI 介入决策时，人们会在 AI、人类决策者、组织和程序员之间重新分配责任（Lima et al., 2021; Shank et al., 2019; Joo, 2024）。这说明自我责任归因不是一种固定倾向，而会随着人机责任结构发生变化。

上述修改已在正文第 4.3 节（区分道德冲突强度与道德议题强度，以 Brown et al., 2022 为主要依据，并补充 van der Waa et al., 2021 和 Fabre et al., 2024 等人机协同情境研究）和第 4.1 节（将自我责任归因重新界定为情境性判断，区别于稳定人格特质）中逐一落实。

意见 5：文章同时引用了多个理论基础（服从权威理论、自动化偏见理论、心智感知理论），但三个理论对算法权威作用于道德推脱的逻辑主线有所不同。服从权威理论强调权威结构改变责任定位；自动化偏见理论侧重认知层面的过度依赖；心智感知理论则从主体能动性/体验性判断出发。三者在逻辑层面并非完全兼容，多理论并用若缺乏清晰的层次界定，可能导致模型在理论上过于宽泛。

回应：非常感谢审稿专家指出多理论整合层次不清的问题。针对这一问题，我们进行了修改，明确了三类理论在本文中的层次分工，具体来说：

服从权威理论位于模型的宏观责任结构层面，用于解释算法为什么可能被个体理解为一种外部权威判断来源。服从权威研究表明，权威情境会改变个体对自身行动角色的理解，使其更容易将自己定位为执行者而非完全自主的决策者（Milgram, 1963）。据此，本文使用服从权威理论主要解释算法权威为何可能改变责任定位，即当算法被赋予权威地位时，个体为何可能将不良后果的责任向外转移。因此，该理论回答的是“算法权威为何能够重塑责任结构”的问题。

自动化偏见理论位于认知加工层面，用于解释个体为什么会在具体决策中依赖算法建议。自动化偏见研究强调，个体在自动化系统面前可能出现过度依赖、误用或减少独立审查等问题（Skitka et al., 2000）。因此，该理论并不直接解释道德推脱本身，而是解释算法权威如何通过认知卸载和不适当依赖，使个体更容易接受算法判断，并进一步弱化自我责任归因。换言之，自动化偏见理论服务于本文模型中的认知卸载与代理状态路径。

心智感知理论位于责任主体判断层面，用于解释算法权威为什么可能产生双向作用。心智感知理论指出，个体对某一主体能动性和体验性的判断，会影响其道德责任、伤害感知和道德评价（Gray et al., 2012）。在人机协同情境中，当算法被感知为高能力、高专业性和高可靠性时，个体可能更容易依赖算法建议，并将部分责任外置于系统；但当个体意识到算法缺乏心智、情感和道德主体性时，算法又难以被视为能够承担完整道德责任的对象，从而可能促使个体重新确认自身责任。因此，心智感知理论主要服务于本文模型中的价值冲突与责任回收路径。

综上所述，三种理论在本文模型中并非简单并列，也不是完全互相替代的解释框架，而是分别回答不同层面的问题：服从权威理论回答“算法为何可能成为外部权威判断来源”，自动化偏见理论回答“个体为何会在认知加工中依赖算法建议”，心智感知理论回答“个体为何可能在责任主体判断中无法将责任完全转移给算法”。三者共同服务于同一条主线，即算法权威如何通过改变个体的自我责任归因，进而影响状态性道德推脱。为避免模型过于宽泛，我们在修订稿中明确将自我责任归因作为三个理论汇合的核心机制变量：服从权威和自动化偏见主要解释自我责任归因降低的路径，心智感知则解释自我责任归因可能被重新激活的路径。

上述修改已在正文第 3.2 节（整体区分认知卸载路径与价值冲突路径）、第 3.2.1 节（服从权威理论与自动化偏见理论的层次分工）、第 3.2.2 节（心智感知理论的层次分工）和第 4 节开头及第 4.1 节（明确自我责任归因为三个理论汇合的核心机制变量）中逐一落实。

意见 6：作者在自检报告和摘要中均声称首次提出了双重认知评估模型，并在正文标注了蓝色文字以标识创新内容。然而，文章在正文中对本文相对于已有整合框架（如 Poszler & Lange, 2024；谢才凤等, 2023）的增量贡献缺乏正面、系统的比较论述。Poszler 和 Lange（2024）已对智能决策支持系统对伦理决策的影响进行了系统文献综述，而谢才凤等（2023）也已就算法决策趋避提出了过程动机模型，本文未与这两项研究进行充分的区分性比较。

回应：非常感谢审稿专家对本文创新性表述和增量贡献论证的提醒。针对这一问题，我们已在修订稿中作出如下修改。

首先，Poszler 和 Lange（2024）的研究已经对智能决策支持系统如何影响人类伦理决策进行了系统性文献综述，并提出了一个整合框架。该研究的重要贡献在于从智能决策支持系统的功能类型出发，梳理其可能带来的个体层面后果，如慎思增强、动机增强、自主性增强和行动增强，以及社会层面后果，如道德技能退化、道德进步受限和道德责任缺口。也就是说，Poszler 和 Lange（2024）关注的是智能决策支持系统如何总体上影响伦理决策能力和伦理决策后果，其分析重点是决策支持系统对伦理决策过程的促进、削弱或社会后果。本文聚焦于其中一个更具体的问题：当算法被个体感知为权威判断来源时，它如何影响个体在具体人机协同决策中的状态性道德推脱。换言之，本文关注的不是智能决策支持系统是否改善或削弱伦理决策能力，而是算法权威如何通过改变自我责任归因，进而影响个体是否通过责任转移、责任扩散或合理化来降低自我道德约束。因此，本文相较 Poszler 和 Lange（2024）的增量贡献在于，将伦理决策支持系统研究中较为宏观的“责任缺口”“自主性变化”等问题，进一步转化为可检验的微观心理机制，即“算法权威—自我责任归因—状态性道德推脱”。

其次，谢才凤等人（2023）提出的算法决策趋避过程动机模型，也为理解个体为何接受或拒绝算法决策提供了重要框架。该研究关注个体面对算法决策时的趋近和回避动机，强调个体对算法决策的接受或拒绝受到多种心理动机影响。该模型有助于解释为什么人们有时会依赖算法，有时又会抵触算法，尤其能够回应算法欣赏、算法厌恶和算法规避等现象。本文充分认可该研究对于理解算法决策趋避行为的重要价值。本文进一步关注：当个体在人机协

同决策中采纳或拒绝算法建议时，这一过程如何改变其对自身道德责任的理解，并进一步影响状态性道德推脱。也就是说，谢才凤等人（2023）主要回答“个体为什么使用或不使用算法”，而本文进一步回答“个体在使用或拒绝算法的过程中，如何重新理解自己的责任，并由此产生或抑制道德推脱”。因此，本文并不是替代算法决策趋避模型，而是在其关于算法采纳/回避的基础上，将研究焦点推进到道德责任归因和道德自我调节层面。

上述修改已在正文第 5.1.3 节末尾（直接与 Poszler & Lange, 2024 和谢才凤等, 2023 进行对比，明确本模型的增量贡献）和第 1 节引言（指出已有综述未系统回答核心问题）中逐一落实。

意见 7：文章在第 5.1 节中对双重认知评估模型的四象限讨论存在过度推断的迹象。作者声称在高可解释性、高道德冲突条件下，个体既能有足够的审查依据，又面对明确的价值冲突，因此更易强化自我责任归因，但这一结论并未获得直接实证支持，所引证据均为单变量研究，双向交互效应的存在仅为理论推测。此外，关于低可解释性导致即使算法带有偏见和不公正的建议，仍会诱导个体进行歧视性的选择，而不会提高个体的内疚水平的陈述超出了原引文献（Ghasemaghaei & Kordzadeh, 2024）的实际发现范围。

回应：非常感谢审稿专家对本文理论推论边界的细致提醒。针对您的建议，我们在修订稿中作出如下修改。

关于四象限模型。我们在修订稿中明确了基于现有单变量证据提出的理论假设。具体而言，可解释人工智能研究表明，可解释性能够提高个体对算法建议依据、适用范围和局限的理解，从而为审查和质询算法建议提供条件（Adadi & Berrada, 2018; Arrieta et al., 2020; Molnar, 2020）；算法公平与可解释性研究也指出，解释信息有助于使用者识别模型输出的依据、偏差来源和适用边界，从而增强算法系统的透明性与问责性（Arrieta et al., 2020; Shin, 2021）。道德冲突和道德强度相关研究则表明，当决策任务涉及更明显的价值冲突、伤害后果或伦理责任时，个体通常会投入更多道德认知、责任判断和道德情绪加工（Brown et al., 2022; Greene et al., 2001; Jones, 1991）。但现有研究尚未直接证明可解释性与道德冲突强度交互后一定会增强自我责任归因。因此，修改稿明确：在高可解释性与高道德冲突同时存在时，个体可能更具备审查算法建议的依据，并更容易意识到决策后果的道德意义，因此这一组合有可能增强自我责任归因；但这一推断仍有待未来研究通过同时操纵算法可解释性和道德冲突强度的交互实验加以检验。

关于“低可解释性—低/高道德冲突”条件下的推论。修订稿改为：低可解释性可能降低个体对算法建议进行审查和质询的能力，使其更难识别算法建议中的潜在偏见或不公正成分，从而可能增加算法建议被不加审查地采纳的风险。

关于 Ghasemaghaei 和 Kordzadeh（2024）的引用。该研究以服从权威视角考察不公正算法推荐对管理者歧视性决策的影响，发现不公正的算法推荐会提高管理者作出歧视性决策的可能性，并且这种决策并未伴随更高的内疚感。但该研究并未直接操纵算法可解释性，也没有直接证明“低可解释性导致个体即使面对带有偏见和不公正的建议也会进行歧视性选择”。修订稿中明确：Ghasemaghaei 和 Kordzadeh（2024）的研究仅被作为“不公正算法推荐可能影响道德决策和道德情绪”的证据，而不再被用作“低可解释性导致歧视性选择”的直接证据。

此外，我们在未来研究部分增加更明确的检验建议。具体而言，未来研究应采用双因素实验设计，同时操纵算法可解释性和道德冲突强度，并分别测量自我责任归因、道德情绪、内疚感、道德推脱和实际行为选择。只有在这种设计下，才能直接检验双重认知评估模型所提出的交互假设。

上述修改已在正文第 5.1.2 节（保留四象限理论假设，但结合第 5.2 节呼应未来双因素

实验检验）和第 3.1 节表 1（将 Ghasemaghaei & Kordzadeh, 2024 的引用限定为从指令强度操纵到诱发，不再用作低可解释性与歧视性选择的因果推论）中逐一落实。

意见 8：文章在引言明确提出要回答核心心理机制问题，并声称将在服从权威理论、自动化偏见理论和心智感知理论的基础上，提出以自我责任归因为核心的模型。然而，文章对以下几个关键方向的回应不充分：（1）对可解释性如何影响自我责任归因的机制说明相对充分，但对道德冲突强度影响自我责任归因的具体心理路径（是通过情绪？价值激活？还是身份威胁？）缺乏明确说明。（2）自检报告问题 2（最后的总结或展望是否是基于前面论述的基础上提出的）作者回答是，但展望（第 5.2 节）中部分内容（如神经生理技术应用）与模型逻辑之间的关联较为牵强，未能清晰对应到双重认知评估的具体路径。

回应：非常感谢审稿专家对本文机制论证和研究展望一致性的细致提醒。针对这一问题，我们已在稿件中作出如下修改。

关于道德冲突强度如何影响自我责任归因，我们在修订稿中拆解为“价值激活—道德情绪—责任再确认”的过程。具体而言，当人机协同决策中的算法建议只涉及低道德冲突任务时，个体更容易将其理解为技术性辅助或效率优化问题，因而较少反思自身道德责任；但当算法建议涉及明显的公平、伤害、歧视、权利或生命健康等价值冲突时，个体的核心道德价值会被激活，其对行为后果的道德意义、他人受损可能性和自身角色责任的关注会提高。道德强度理论也指出，后果严重性、社会共识和受影响者接近性等因素会影响个体对伦理问题的识别和判断（Jones, 1991）。伦理冲突研究进一步表明，当个体面对相互竞争的道德要求时，会产生更高的心理紧张和道德认知投入（Brown et al., 2022）。因此，道德冲突强度可能通过激活个体的价值判断和道德情绪，使其更难将决策完全理解为算法执行问题，从而促使其重新确认自身作为最终决策者的责任。

其次，道德冲突强度影响自我责任归因并不只是通过一般情绪唤醒，而是与个体对自身道德身份和责任角色的理解有关。当算法建议与个体的公平、伤害避免或职业伦理要求发生冲突时，个体可能意识到“采纳算法建议”并非中性的技术选择，而是涉及其自身道德立场和身份承诺的行为。道德身份研究表明，当个体的道德自我概念被激活时，其更可能关注行为是否符合自身道德标准（Aquino & Reed, 2002）。在人机协同决策中，高道德冲突可能使个体从“算法建议的执行者”转变为“后果承担者”和“道德判断者”。因此，我们在修订稿中将道德冲突强度影响自我责任归因的路径表述为：高道德冲突通过激活价值判断、道德情绪和道德身份，使个体更容易意识到算法无法替代人类承担最终道德责任，从而增强自我责任归因；低道德冲突则可能使个体更容易将算法建议视为普通技术辅助，进而为认知卸载和责任外移提供条件。

关于第 5.2 节未来研究方向。修订后，我们已压缩神经生理技术部分，并将其从独立研究方向降格为辅助性方法。未来研究的重点被重新聚焦为四个与模型直接对应的问题：第一，如何检验“认知卸载与代理状态路径”，即算法权威是否会增强个体对算法建议的依赖，并通过降低主动审查和自我责任归因诱发道德推脱；第二，如何检验“价值冲突与责任回收路径”，即高道德冲突、低算法心智感知或高可解释性是否会促使个体重新确认自身责任；第三，如何检验算法可解释性和道德冲突强度的交互作用。第四，未来研究应优先采用情境实验、行为决策任务，通过构建具有真实道德后果的人机协同决策任务，检验算法是否以及如何影响状态性道德推脱。此外，我们对神经生理技术的定位也作了修改，修订稿将其定位为未来机制检验中的辅助性方法。具体为：神经生理技术可以作为探究核心心理机制的辅助验证手段，以进一步丰富研究数据的多样性。

上述修改已在正文第 4.3 节（通过 Jones, 1991、Brown et al., 2022 等文献，阐述道德冲突强度通过价值激活、道德情绪影响自我责任归因）和第 5.2 节（将神经生理技术从独立研

究方向压缩为一句话的辅助验证方法，未来研究聚焦于模型直接对应的三条路径）中逐一落实。

此外，除针对上述审稿专家的意见进行对照修改以外，我们还对全文的表述、格式和参考文献等细节内容进行了检查和修改。

第二轮

审稿人 1 意见：

感谢作者对审稿意见的细致回复和对论文的修改，我还有以下两个小问题。

总体回应：非常感谢审稿专家对本文的认可。接下来，我们将结合您的具体建议，对您提出的问题进行逐一回复。所有修改之处用绿色字体标注。

意见 1：作者在引言中提到“已有综述为理解这一复杂关系提供了重要基础”，既然提到了“已有综述”，那么应当引用以及简单说明如何提供重要基础。

回应：非常感谢审稿专家对本文引言部分的细致提醒。我们对此进行了内容上的补充。

在第 1 节引言中补充了 Poszler 和 Lange（2024）以及谢才凤等人（2023）的相关研究，并分别说明其理论贡献。具体而言，Poszler 和 Lange（2024）构建了智能决策支持系统影响人类伦理决策的整合框架，该研究从宏观层面出发，梳理了智能决策支持系统可能带来的个体自主性变化、道德技能退化和道德责任缺口等后果。这为本文认识算法介入可能改变人类决策者的自主性和责任结构提供了重要基础。谢才凤等人（2023）提出的算法决策趋避过程动机理论，系统解释了个体面对算法决策时产生趋近或回避反应的心理动机，为理解个体为何有时表现出算法欣赏，有时又产生算法厌恶或算法拒绝提供了理论基础。

与此同时，我们进一步明确了上述研究的解释边界：Poszler 和 Lange（2024）主要揭示智能决策支持系统可能产生哪些伦理后果，谢才凤等人（2023）主要解释个体为何接受或拒绝算法，但两项研究均未专门回答不良后果出现后，算法权威如何改变个体的自我责任归因，以及这种责任加工过程如何进一步影响状态性道德推脱。因此，本文是在上述研究基础上，将问题进一步推进到个体层面的责任归因与道德自我调节过程。

上述修改已在正文第一节引言落实。

意见 2：作者“将算法权威界定为：个体基于对算法客观准确性和专业性的感知，赋予其优先解释权与行动引导的心理地位。”这个定义的表述似乎有歧义，算法权威是某种“心理地位”吗？

回应：非常感谢审稿专家指出这一关键问题。原稿对算法权威表述不够准确，我们对此进行了概念上的重新界定。

算法权威不是算法天然具有的客观属性，也不是个体内部某种孤立的心理状态，我们已删除“心理地位”这一表述，并将算法权威重新界定为：算法权威是指在人机协同决策中，个体基于对算法客观准确性和专业性的感知，认可算法具有较高优先解释权和行动引导作用，并由此形成较高依赖或服从倾向的一种主观权威关系。

这一修改进一步明确了三个方面。第一，算法权威具有关系性。它描述的是个体与算法之间形成的权威归属关系，而不是算法本身固有的属性。第二，算法权威具有情境性。同一算法在不同任务风险、组织授权程度和责任结构下，可能获得不同程度的权威认可。因此，算法权威并非稳定不变。第三，算法权威具有行为指向性。当算法被个体认可为具有较高解

释优先性和行动引导时，通常表现为个体更倾向于采纳、依赖或服从其建议。因此，算法权威不同于单纯的算法能力，也不完全等同于一般意义上的算法信任，而是强调算法判断如何获得优先性并实际影响个体的决策行为。

对应正文，上述修改已在正文第 2.1 节中进行了概念上的重新表述。

审稿人 2 意见：

本次修改整体质量良好，作者对两位审稿专家意见中的绝大多数做出了实质性和有针对性的回应。然而，还存在以下一些问题需要进一步完善。

回应：非常感谢审稿专家对本文的认可。接下来，我们将结合您的具体建议，对您提出的问题进行逐一回复。所有修改之处用绿色字体标注。

意见 1：作者在回应中坦承“现有实证研究的确没有直接测量状态性道德推脱”，这意味着支撑该理论模型的大量实证证据实际上是关于责任归因、算法采纳或算法厌恶的，而非道德推脱本身。作者将此定位为“本文撰写理论模型与实证研究之间桥梁的核心价值所在”，这一逻辑的问题在于：综述论文的核心任务是从已有实证证据中归纳和整合，而非提出一个尚缺乏直接实证支撑的新模型。如果现有研究几乎没有直接检验“算法权威→状态性道德推脱”这一关系，那么模型建构的依据本质上是基于间接证据的理论推测，而非对已有实证发现的系统整合。作者应当更加坦诚地承认这一点，并在展望部分更加突出“直接检验模型”的紧迫性，而非将其与其他展望方向并列。（2）作者声明在后文表格中区分了直接证据与间接证据，但在正文第 3.1 节中，这一区分主要存在于文字说明中，部分地方的论述边界仍有模糊。例如第 3.2.2 节中引用的部分算法厌恶研究能否作为价值冲突与责任回收路径的直接行为证据，仍需进一步辨析，即算法厌恶反映的是对算法的不信任和拒绝，其是否可以等同于责任回收？（3）标注各条路径的实证检验程度（已直接检验/可间接推论/纯理论推测），使读者对模型的证据确定性有更准确的预期。

回应：非常感谢审稿专家指出本文证据基础与模型定位中的核心问题。针对这一问题，我们已从以下三个方面对文章进行了补充和修订。

第一，关于本文理论模型的证据定位。删除了“这正是本文搭建理论模型与实证研究之间桥梁的核心价值所在”等淡化证据不足的表述。将双重认知评估模型明确定位为基于相关证据和既有理论提出的待检验框架。同时在正文中明确承认，现有研究尚未直接检验算法权威如何通过自我责任归因影响状态性道德推脱。在第 5.2 节中，我们进一步调整了未来研究方向的优先顺序。直接检验模型不再与纵向研究、跨情境比较等方向并列，而被明确列为后续研究最首要的任务。未来研究应首先通过情境实验操纵算法权威水平，同时测量自我责任归因和状态性道德推脱。只有在这一核心关系获得直接支持之后，才能进一步检验可解释性、道德冲突强度及其他边界条件。

第二，关于算法厌恶与责任回收之间的关系界定。算法厌恶主要反映的是个体对算法能力和可靠性的怀疑，表现为拒绝算法建议或重新依赖自身判断等反应（Burton et al., 2020）；责任回收则要求个体重新确认自己对决策及其后果负责（Moore, 2015），二者并不等同。个体拒绝算法可能源于算法曾经出错、维护自主性、专业身份受到威胁、对技术的一般性排斥等（Burton et al., 2020; Dietvorst et al., 2015; Jussupow et al., 2022）。这些原因并不必然意味着个体已经重新承担了道德责任。因此，我们在第 3.2.2 节中将算法厌恶重新定位为算法权威受到质疑的行为证据，可能促使个体重新审查责任归因的前置条件。

第三，关于标注各路径的实证检验程度。我们将“理论推测”和“研究支持的间接推论”作为判定标准，并在第 5.1.1 节及图 1 逐条标示。

上述修改已在第 3.2.2 节、第 5.1.1 节、图 1 和第 5.2 节中逐一落实。

意见 2：第二，作者明确比较了与 Poszler 和 Lange（2024）及谢才凤等人（2023）的差异，但比较方式偏向于“本文做了什么”与“他们做了什么”的并列陈述，缺少“他们做到了什么，但哪些关键问题他们没有解决，本文如何填补”的递进式论证。尤其是在与 Poszler 和 Lange（2024）的比较中，该研究包含的道德责任缺口已经与本文的核心议题相近，作者需要更精准地指出二者的分析层次差异，而非仅仅说明关注点不同。

回应：非常感谢审稿专家对本文论证递进性的问题提出的宝贵意见。我们对此进行了内容上的修改。

（1）关于本文与 Poszler 和 Lange（2024）的差异。Poszler 和 Lange（2024）构建了智能决策支持系统影响人类伦理决策的整合框架，系统梳理了智能决策支持系统可能带来的自主性变化、道德技能退化和道德责任缺口等后果。不仅在于指出算法可能改善或削弱伦理决策，而且揭示了，当人类与智能系统共同参与决策时，行动控制、决策依据和后果承担可能被分散在不同主体之间，从而导致责任主体难以清晰确定。由此可见，该研究提出的道德责任缺口与本文所关注的责任问题高度相关，也是本文进行理论推导的重要基础。

但是，该框架并未进一步解释，当具体的不良后果出现后，个体是否会因为算法参与而降低对自我责任归因，以及这种自我责任归因的变化是否会进一步转化为状态性道德推脱过程。换言之，道德责任缺口描述的是责任难以归因的客观问题。本文所讨论的自我责任归因与状态性道德推脱，则关注个体对责任的主观界定及其产生的后果。

基于这一差异，本文尝试将其识别出的系统层面责任问题进一步转化为个体层面的、可以被直接检验的心理机制。具体而言，当算法被个体认可为具有较高解释优先性和行动引导的权威来源时，个体是否会重新理解自己在决策中的角色，改变自我责任归因，并由此影响状态性道德推脱。需要强调的是，这一机制目前仍建立在间接证据与理论推导基础上，因此本文只是将宏观责任问题具体化为待检验的心理机制，而不能声称已经在实证层面填补了 Poszler 和 Lange（2024）留下的研究空白。

（2）关于本文与谢才凤等人（2023）的差异。谢才凤等人（2023）提出了算法决策趋避的过程动机理论，从认知、关系和存在等动机出发，系统解释个体为何接受、依赖、回避或拒绝算法决策，为理解算法欣赏、算法厌恶和算法规避提供了理论框架。但是，采纳算法并不等同于责任外移，拒绝算法也不等同于责任回收。谢才凤等人（2023）的理论主要解释个体在接触和使用算法过程中的趋近与回避动机。尚未专门讨论，当个体采纳或拒绝算法建议并产生不良后果后，如何在人类决策者与算法之间重新分配责任，这种责任加工过程如何影响个体的道德自我调节，以及算法采纳或算法拒绝在什么条件下会进一步转化为道德推脱或责任承担。简言之，该理论主要关注算法决策的使用阶段，而本文进一步关注不良后果出现后的责任归因阶段。因此，谢才凤等人（2023）解释个体为何趋近或回避算法，本文则尝试进一步说明这些算法反应在何种条件下会伴随自我责任归因的变化，并探讨与状态性道德推脱的关系。本文不是替代算法决策趋避过程动机理论，而是在其关于算法接受和拒绝的解释基础上，将研究问题进一步推进到事后责任加工和道德自我调节层面。

根据上述分析，我们在第 5.1.3 节进一步明确，Poszler 和 Lange（2024）识别了智能决策支持系统可能造成的系统性道德责任缺口；谢才凤等人（2023）解释了个体面对算法决策时的趋近与回避动机；本文则尝试将上述责任结构和算法反应进一步连接到不良后果出现后的情境性自我责任归因与状态性道德推脱。

上述修改已在正文第 1 节引言和第 5.1.3 节中落实。

意见 3：四象限模型的本质问题是两个边界条件（可解释性与道德冲突强度）的交互效应缺乏直接实证支持。作者在回应和正文中均承认“这一推断仍有待未来研究通过交互实验加以检验”，但在第 5.1.2 节正文中，四个象限的论述仍以肯定句式展开。同时，混合条件（高可解释性但低道德冲突，或低可解释性但高道德冲突）下责任归因方向的论述过于模糊（更不稳定，也更易受到个体差异和组织规范的影响）。

回应：非常感谢审稿专家指出四象限模型的证据边界与表述问题。修订稿已不再使用四象限模型这一命名，而统一表述为可解释性与道德冲突强度的四种条件组合。这四种组合是双重认知评估模型基于两个边界条件提出的待检验命题。我们对相关表述进行了统一，并进行了以下修改。

第一，重新界定四种条件组合的理论性质。我们将在修订稿中统一改为“模型推测”“理论上可能”“在特定条件下可能”或“有待直接检验”，且在四种条件组合的论述之前增加明确说明：现有研究主要分别考察可解释性或道德冲突强度的单独作用，尚无研究同时操纵这两个变量，并直接检验二者对自我责任归因和状态性道德推脱的交互作用。因此，四种条件组合均属于基于单变量研究和理论推导提出的待检验命题，而不是已有实证研究所支持的确定性结论。同时，我们将低可解释性、低道德冲突和高可解释性、高道德冲突的组合条件，重新界定为模型提出的两个锚点性假设。

在低可解释性、低道德冲突条件下，个体既缺乏理解和审查算法判断依据的能力，也缺少将决策重新理解为道德责任问题的动机。模型据此推测，个体可能更容易将算法建议视为需要执行的技术性结论，并将自身定位为算法建议的执行者，从而降低自我责任归因的可能性。

在高可解释性、高道德冲突条件下，个体一方面可能获得理解、审查和质疑算法建议的依据，另一方面也可能因算法建议涉及公平、伤害和责任等问题而产生较强的道德审查动机。模型据此推测，这两种条件可能共同促进个体重新确认自身的最终决策责任。

第二，混合条件下责任归因方向论述过于模糊。修订稿中明确根据“是否具有审查能力”和“是否具有审查动机”两个方面分别说明其理论含义。

在高可解释性、低道德冲突条件下，个体可能具有理解和审查算法建议的能力，但缺乏将这种能力实际用于道德审查的动机。高可解释性使个体能够了解算法建议的依据、适用范围和潜在局限，这在理论上可能增加其决策参与感，并提高自我责任归因。但由于任务的道德冲突较低，个体未必会主动核验算法建议，也未必会主动承担道德责任。如果解释信息只是增强了算法建议的表面合理性，反而可能提高算法信任并强化依赖。因此，该条件下存在两种相互竞争的可能：当解释能够促使个体开展实质性核验时，自我责任归因可能提高；当解释仅被作为算法专业性和正当性的证明时，自我责任归因可能会降低。该条件的关键在于：解释是否真正激活个体的审查行为，而不仅是形式上提供了解释。

在低可解释性、高道德冲突条件下，个体可能具有较强的道德审查动机，但缺少识别算法判断依据、偏差来源和适用边界的技术条件。高道德冲突可能使个体更加关注公平、伤害和责任问题，并由此产生质疑或拒绝算法的倾向；但是，算法拒绝本身不能直接等同于责任回收。如果个体保留最终决策权，并且组织明确要求其对最终结果负责，道德警觉可能促使其重新确认自身责任；如果算法使用具有强制性，且决策过程持续不透明，个体仍可能以无法理解、无法干预或无法控制算法为由将责任外移。因此，该条件下的关键限制因素不是个体是否意识到道德问题，而是其是否拥有实际决策控制权，以及组织责任边界是否清晰。

综上，两个混合条件不再被笼统处理为“结果不稳定”。高可解释性、低道德冲突反映的是“具有审查能力但缺少审查动机”，低可解释性、高道德冲突反映的是“具有审查动机但缺少审查能力”。两种条件均包含责任回收与责任外移的竞争性机制，且各自具有不同的关键限制因素。

上述修改已在图 1、第 5.1.2 节、第 5.1.3 节和第 5.2 节中逐一落实。

此外，除针对上述审稿专家的意见进行对照修改以外，我们还对全文的表述、格式和参考文献等细节内容进行了检查和修改。

审稿人 1 意见：感谢作者的回复和修改，我没有其他问题了。

编委 1 意见：同意发表。

编委 2 意见：同意发表。

主编意见：建议录用。