AI助手服务失败下算法可解释性的积极效应

周建设，刘子峰，周 叶

（南昌航空大学经济管理学院，南昌 330063）

摘要： AI(人工智能)助手服务失败本质上是算法失败，算法的“黑箱”特性加剧了消费者服务失败下的负面反应。采用组间实验的方法，从控制感的视角探究通过可解释人工智能的方法增强算法可解释性对服务失败下消费者行为的影响机制及其边界条件。研究发现，当算法可解释性得到增强时(相比对照)，消费者服务失败下的继续使用意愿会提高，且消费者的可控性感知在其中发挥部分中介作用。但上述效应在AI助手拟人化程度较低(相比较高)时不显著。

关键词： AI(人工智能)服务失败；算法可解释性；控制感

中图分类号： F713.5 **文献标志码：** A **文章编号：** 1671－1807(2025)09－0262－08

随着算法与自然语言处理技术的飞跃式进步，人工智能(AI)正以前所未有的广度与深度重塑各行各业，其应用边界已从单一的流媒体推荐服务，拓展至设计、金融、教育等多元化领域，深刻影响着社会经济的每一个角落[1]。据艾瑞咨询最新发布的2023年《中国人工智能产业研究报告》显示[2]，2023年中国人工智能产业规模已达2 137亿元，并预测至2028年将突破8 000亿元大关。特别是在服务领域，AI助手在节省人力资源、优化服务流程、提升效率方面展现出显著优势[3]，预示着AI助手自主完成任务、解决问题的时代正加速到来。然而，受限于如今的技术水平，AI助手在面对用户请求有时会表现出“答非所问”或显得“无能为力”，无法达到用户预期的结果，引发服务失败的问题[4]。服务失败会使用户产生负面情绪，进而可能导致放弃使用、转品牌或传播负面口碑等不利行为，对企业形象及用户忠诚度构成挑战[5]。因此，探索有效的服务补救策略，以缓解用户负面情绪、修复受损的用户关系和提高其继续使用意愿，成为亟待解决的关键议题。

在人工智能的发展进程中，算法的进步与更新迭代始终扮演着至关重要的角色，决定了人机交互的模式与最终结果[6]。人工智能算法的可解释性，是影响用户体验的核心要素之一，直接关系到用户在面对算法决策时的情绪及交互失败后的负面反应[7]。但由于AI算法的“黑箱”特性，即人们并不清楚算法的内在逻辑和结果输出方式[8]，导致消费者在服务失败时的负面情绪不仅来源于失败本身，还有对算法缺乏可解释性的不满[7]。算法的可解释性缺乏，使得其无法向用户解释决策过程和逻辑，让用户对AI助手信任度下降甚至是产生厌恶的心理[9]。例如，Chois等[10]认为黑盒算法决策的可解释性缺乏可能导致消费者对算法偏见的担忧。而将黑箱决策过程转化为可解释的过程，对于建立人与机器之间的信任是有益的[11]，还能有助于削弱算法厌恶[8]。已有研究表明[12-13]，对推荐结果进行解释不仅可以增加用户对系统的信任、帮助用户更快地做出决定并说服用户购买，还能通过揭示系统背后的推理过程和数据，为系统优化提供方向。在算法服务失败情境下，虽然Chen[9]从归因理论视角揭示了提高算法可解释性的积极效应，但其影响机制及边界条件仍需进一步探索。特别是在服务失败导致控制感丧失的背景下，用户对于算法自主性的担忧尤为强烈，他们渴望恢复对使用过程与结果的掌控力[14]。

在面对高度模糊且难以解析的算法决策时，用户往往表达出有关自主性的消极看法，他们担忧算法决策会削弱人类角色或威胁个人独特性。这将进一步激发他们内在的掌控动机，包括对控制力和影响力的渴望[15]。事后解释作为可解释人工智能

收稿日期： 2024-10-29

基金项目： 国家自然科学基金(72263023)；江西省高校人文社科项目(GL20124)

作者简介： 周建设(1968—)，男，江西临川人，博士，副教授，研究方向为体验营销、品牌管理；通信作者刘子峰(2000—)，男，江西南昌人，硕士研究生，研究方向为品牌管理、绿色消费行为；周叶(1977—)，男，江西抚州人，博士，教授，研究方向为绿色制造。

(XAI)领域的重要工具,通过揭示算法模型的运作逻辑与决策依据,来增强算法决策的可理解性[9]。而个体的控制感会随着自身资源感知的增加或自身能力的强化而提升[16],那么事后解释是否可以恢复甚至增强用户对系统和算法的控制感和掌控力。本文以此为切入点,深入探究增强算法可解释性能是否缓解AI服务失败下用户的负面响应,并探讨这一积极效应的影响机制与适用边界,以期为AI服务的持续优化与用户体验的全面提升提供理论依据与实践指导。

1 理论基础和研究假设

1.1 服务失败与服务补救

服务失败是指在消费者体验流程中服务未能契合其心理预期,包括从感知到现实的偏差、不幸遭遇或功能缺陷[17]。依据决策主体的差异,AI助手服务失败可分为两类:一类源于算法错误,即算法服务失败;另一类则归咎于人为因素。但在机器人服务日益普及的今天,机器人服务失败本质上都属于算法服务失败,包括智能客服、语音助手乃至实体机器人的服务失败等。算法服务失败往往触动用户的敏感神经,激发诸如厌恶[18]、愤怒[19],乃至品牌转换[5]等负面反应。当前学术界对算法服务失败的研究聚焦于3个维度:一是深入探讨消费者在面对算法与人为服务失败时的认知差异[20];二是细致剖析影响消费者对算法服务失败反应的多元因素,如AI智能化水平[9]、失败的具体类型[21]、消费场景[22]等;三是聚焦于服务补救策略,意在探索化解服务失败的有效手段(如诚恳道歉、详尽解释及人工干预)[17]。

从实践价值的角度讲,大部分学者聚焦于如何利用补救措施有效缓解消费者服务失败下的负面反应。服务补救是服务提供方在服务失败后,为缓解顾客不满情绪及处理顾客反馈所实施的一系列补救措施[17]。这些措施包括表达歉意、经济赔偿乃至重启服务流程等[23]。其目的不只是修正错误,更在于通过积极姿态重建顾客信任,最终实现顾客满意度的飞跃。对于机器人服务失败,现有的补救研究主要围绕机器人自主言语补救与人类服务员工的介入干预展开[17]。但在现实情境中,人工干预存在一定延时性,往往会耽误服务补救的“黄金时间”,如果AI助手能在服务失败的第一时间进行补救,将大大提高服务补救的效率。此外,当AI助手服务失败时,消费者对算法透明度的渴求尤为强烈,但算法决策的“黑箱”特性往往加剧其不满情绪[9]。因此,一个值得探讨的问题是:当AI在输出时增强其算法可解释性,让消费者能够洞悉算法决策的逻辑与过程,是否能够缓解其服务失败下的负面反应并维持继续使用意愿?由此,提升算法的可解释性,或能成为缓解算法服务失败下消费者负面情绪的新途径。

1.2 可解释人工智能(XAI)与事后解释

XAI(可解释人工智能)理念始于2016年,它通过赋予机器学习模型透明度,让AI系统的决策逻辑与行为轨迹变得可理解、可阐释[24]。具体而言,可解释性是指算法能向用户解释其决策结果的逻辑过程的能力,即交互结果是为何以及如何被算法提供的[6]。它深刻回应了深度学习等尖端模型兴起后,对AI可解释性日益增长的迫切需求,并揭开了复杂模型决策过程的神秘面纱。特别是在医学等高风险、高影响领域,XAI的解释能力尤为重要[24]。它要求生成的解释必须清晰透彻,足以让人类用户把握全局,从而基于这些洞察做出决策。现阶段的解释技术分为两类:第一类是嵌入式解释(事前解释),是指基于算法模型解释的技术,即在模型训练阶段便内置了解释生成机制[25]。其优势在于紧密耦合模型,能够深入揭示算法本质,解释算法能力,但面临解释可读性优化的挑战[13];另一类则是事后处理解释(事后解释),是指给算法模型提供额外事后解释的技术,即在算法决策与输出之后,再行构建解释,通过提供黑盒算法模型如何以及为什么会得到特定结果的近似理由来解释算法决策[12]。其灵活性高,能够适配各种模型及其决策结果,但也存在难以直接映射系统真实推荐逻辑的局限性,可能会引发用户对解释准确性的质疑[13]。事后解释方法因其能够赋予用户对算法决策过程的直观感知与理解,一直是关注的焦点[9]。因此,本文在提升算法可解释性的方法上采用的也是事后解释。

1.3 可解释性对AI助手服务失败下消费者继续使用意愿的影响

事后解释旨在通过产生可理解的表示,如特征重要性评分、可视化热图或自然语言阐述等,来近似复杂算法模型的内部工作和决策逻辑[10,26-27]。它可以提升公众对算法决策过程的理解与认知深度,增强算法的可解释性。Chen[9]的研究进一步印证了这一点,指出事后解释能够显著提升算法决策的感知可解释性,并改善用户对AI推荐系统的接受度。根据情绪认知评价理论,个体在压力情境下会积极评估和利用资源以有效应对逆境[28]。在遭遇

服务失误的情境中，消费者后续的行为反应深受其对失败原因的认知评价所影响。若消费者能够清晰洞察服务失误的根源及过程，他们倾向于以更为理性的态度审视这一事件，减少因不确定性而产生的焦虑和愤怒等负面情绪。对于自身决策失败，个体可以通过自省的方式了解自身的决策过程，降低对自身决策不透明的感知，以更为理性的姿态处理失败[8]。但算法决策对消费者是不透明的，需通过外部措施增加对算法决策过程的了解程度。而算法决策可解释性的增强能够帮助消费者认识到算法服务失败的原因，让消费者更为合理地看待算法服务失败，也会减少对AI助手的责任归咎[9]。因此，当算法的可解释性通过事后解释得到增强时，能够缓解消费者服务失败下的负面反应，提高其继续使用意愿。据此，提出如下假设。

H1：通过事后解释增强算法可解释性对消服务失败情境下消费者继续使用意愿有正向的影响。

1.4 可控性感知的中介作用

可控性感知来源于控制感，通常指的是个体对外部环境或事件的可控程度的感知，即个体获得其想要的结果、避免不想要结果以及实现目标的能力感知[29]。控制感是个体消费过程中的一项基本需求，也是其消费活动的基本驱动力之一[30]。根据补偿性控制理论，当服务失败时，消费者往往会感受到控制力的削弱，并有强烈的控制感恢复欲望[16]。而控制感的恢复可以通过外部代理提升自身资源感知，或是通过强化自身能力的方式来实现[14]。当个体认为外部因素可控时，他们更可能采取积极的行动来影响结果；反之，则可能感到无助或放弃努力[31]。因此，当消费者对算法的感知可控性提高时，能缓解其对算法服务失败的负面反应。例如，黄雯和黄芊卉[14]认为服务机器人的拟人化水平越高，用户感知的可控性也就越强，越愿意参与到服务补救中。

在此背景下，事后解释的重要性凸显，它不仅提高了算法决策的可解释性，还赋予了算法更高的透明度，满足了用户对于自主性的深层次需求[9]。自主性的满足增强了消费者的心理资源感知和控制感[32]，能更好地应对算法服务失败带来的冲击。同时，事后解释的算法决策清晰化，赋予了用户更强的预判力和掌控力，使他们能够更准确地预见并调整与算法交互的未来走向，从而提高其持续使用的意愿。综上，在算法服务失败的情境下，通过事后解释增强算法的可解释性，提升了用户的自主性感知与工具性能力认知，进而提高了他们对服务、系统及算法整体的可控性感知，有效缓解了服务失败所带来的负面效应。据此，提出如下假设。

H2：消费者可控性感知在通过事后解释增强算法可解释性与其服务失败情境下继续使用意愿的影响中起到中介作用。

1.5 拟人化的调节作用

拟人化是给机器和动物等非人类实体以及其他非人类客体赋予类人特质[33]。就机器人而言，实体形态的机器人常被设计成拥有近似人类的外观，以增强其拟人化效果；而基于文本的聊天机器人，则巧妙地运用多种社会线索，包括身份线索、语言线索和非语言线索等，动态调整其拟人化深度[21]。据自动化社会临场感理论（ASP），随着服务机器人社会属性的增强与拟人化水平的提升，消费者愈发倾向于将AI视为另一个社会实体[14]。这种拟人化特征及丰富的社交线索，有效促进了顾客与AI之间的社会化互动，激发了消费者对AI进行印象管理的内在动机，使得他们在面对AI决策时，更愿投入认知与精力去探索和理解其背后的逻辑[34]。因此，当算法的可解释性通过事后解释得到加强时，消费者的印象管理动机促使他们采取更为理性与积极的姿态去理解算法决策，更清晰认知服务失误的原因及过程，从而缓解服务失败带来的负面影响。

此外，具备高度拟人化线索的类人型AI在实用主义的工具性能力和与人“沟通”的社会性能上更具优势，让个体拥有更高的可操作性感知，在算法可解释性增强下会有更高的能力感知和掌控力[14]。这种双重能力的提升更好地满足了消费者在服务环境中对控制与自我认同的需求，强化了消费者对拟人化机器人问题解决能力的信心，有效缓冲了服务失误带来的负面效应。反之，若AI的拟人化程度较低，则可能引发用户的不信任感，削弱算法可解释性增强的积极效应，影响用户的继续使用意愿。据此，提出如下假设。

H3：通过事后解释提高算法决策可解释性对消费者服务失败情境下继续使用意愿的积极影响在AI拟人化较低时（相对于较高）会减弱；

H3a：AI拟人化程度调节算法可解释性对服务失败下消费者继续使用意愿的影响；

H3b：AI拟人化程度调节可控性感知在算法可解释性对服务失败下消费者继续使用意愿的影响的中介作用。

研究模型如图1所示。

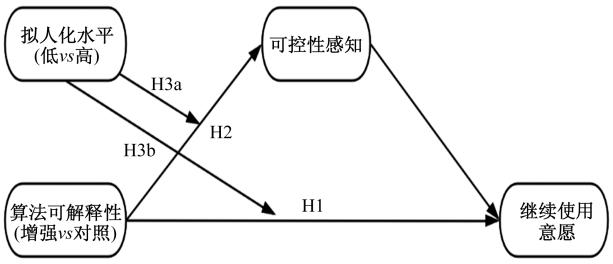

图1 研究模型

2 实验1

2.1 实验目的

实验1的首要目的是检验通过事后解释增强算法可解释性能否提高服务失败下消费者的继续使用意愿,即验证H1是否成立。并探究消费者可控性感知在算法可解释性对其服务失败下继续使用意愿的影响的中介作用,即验证H2是否成立。

2.2 实验流程与设计

实验1采用单因素(算法可解释性:增强*vs*对照)组间实验设计。通过Credamo见数平台线上随机邀请120名普通消费者参加实验,有效参与人数为113人,其中男性有46人,女性有66人,被试年龄集中在21～40岁,69%的被试年龄在21～30岁,20.4%的被试年龄在31～40岁。参与者被随机分配到两个组,每个有效受试者获得了少量的金钱补偿(1元人民币)。实验1参考Chen[9]的实验改编,要求参与者想象自己正在向一个在线学习平台请求个性化学习计划。随后参与者被告知平台的AI助手根据他们的信息推荐了一个学习计划。在算法可解释性增强组,参与者能够阅读以下信息:"学习计划是基于您提供的学习背景、过往学习习惯和学习目标等信息生成的,AI助手发现90%的类似学员选择了上述学习计划";对照组控制条件下的参与者则无法阅读此类事后解释信息。我们的目的是通过提供此类事后解释来增强算法的可解释性。所有参与者在自我评估推荐计划后,均被告知他们发现该计划不可行。

接下来,所有参与者被要求回答3个题项以检验算法可解释性操纵的有效性(α=0.734)[9],"我觉得决策过程很容易理解/决策的原因是可解释的/决策的逻辑是可解释的",题项均是七级李克特量表形式,1分表示非常不同意,7分表示非常同意。紧接着对参与者服务失败后的继续使用意愿(α=0.877)[21]进行测量,共3个题项,"我愿意在未来重新访问该平台/使用该平台其他服务/把该平台推荐给其他人"。然后,参与者被要求展示他们对服务的可控性感知,参考黄雯和黄芊卉[14]的量表(α=0.847),共回答3个题项:"在服务过程中,我感受到自己的主导地位/我感觉自己是关键的/我感觉服务是可控的"。最后,参与者提供了他们的基本信息。

2.3 实验结果

(1)操纵检验。采用单因素方差分析检验算法可解释性的操纵是否成功。结果表明,算法可解释性增强组(均值为4.741,标准差为0.683)的参与者对算法可解释性感知高于对照组(均值为4.154,标准差为0.543),$F(1,111)=25.375$,$P<0.001$,说明算法可解释性操纵成功。

(2)主效应检验。采用单因素方差分析验证H1。结果表明,服务失败后,算法可解释性增强组的参与者继续使用意愿均值为5.126,标准差为0.663;算法可解释性对照组的参与者继续使用意愿均值为4.181,标准差为0.753;算法可解释性增强组的参与者继续使用意愿高于对照组,$F(1,111)=50.200$,$P<0.001$,说明在AI服务失败下增强算法可解释性能提高消费者继续使用意愿,假设H1成立。

借鉴Preacher和Hayes[35]提出的Bootstrap方法,使用Process程序检验中介效应。选择偏差校正的非参数百分位法进行取样,样本数为5 000,采用模型4(用于分析自变量X通过中介变量M对因变量Y的影响路径)验证。检验结果表明,可控性感知的间接效应为0.653,95%置信区间为[0.426,0.905],不包含0,说明可控性感知的中介效应显著,假设H2得到验证。进一步分析发现,算法可解释性的直接效应为0.326,95%置信区间为[0.100,0.552],不包含0,说明可控性感知起部分中介作用,假设H2成立。

3 实验2

3.1 实验目的

实验2的目的在于探究AI助手拟人化程度(较高*vs*较低)对算法可解释性对消费者服务失败情境下继续使用意愿影响机制的调节作用,包括直接影响和间接影响,即验证假设H3是否成立。

3.2 实验流程与设计

实验2采用算法可解释性的增强和对照与AI拟人化程度的较高和较低的2×2组间实验设计。通过Credamo见数平台线上随机邀请280名普通消费者参加实验,有效参与人数为255人,其中男性有106人,女性有149人,被试年龄集中在21～40岁,62.4%的被试年龄在21～30岁,22.4%的被

试年龄在 31～40 岁。参与者被随机分配到 4 个组，每个有效受试者获得了少量的金钱补偿(1 元人民币)。实验 2 参考曹忠鹏等[21]的实验改编，参与者首先被告知，基于算法的 AI 助手已在电子商务平台的个性化推荐服务中得到广泛应用。随后他们被引导想象自己正在一家拥有 AI 助手的流行电商平台上购物，并展示 AI 助手自我介绍的对话界面，如图 2 所示。

观看完上述界面，参与者需要回答五个题项以检验拟人化操纵的有效性，量表参考自曹忠鹏等[21]的研究($\alpha=0.741$)，包括“您认为此 AI 助手是‘像机器的-像人的’‘不自然的-自然的’‘无意识的-有意识的’‘生硬的-生动的’‘沟通不顺畅的-沟通顺畅的’”，语义等级范围为 1～7。接着，参与者被告知在浏览过程中 AI 助手自动推荐了一件 T 恤的链接。与实验 1 相同，算法可解释性增强条件下的参与者会收到以下消息：“嗨，根据您最近的兴趣和需求，我注意到与您类似的大多数客户最近浏览或购买了以下产品，我猜您可能会喜欢它们”；控制条件下的参与者则阅读到“嗨，根据您最近的兴趣和需求，我猜您可能会喜欢以下产品”，然后是同一件 T 恤的链接，如图 3 所示。所有条件下的参与者随后被告知推荐的 T 恤对他们来说并不合适，他们不喜欢它。在服务失败后，参与者被要求报告他们对算

(a) 高拟人化组

(b) 低拟人化组

图 2　拟人化操纵材料

图 3　组间实验材料

法可解释性的感知($\alpha=0.907$)、持续使用意愿($\alpha=0.842$)以及可控性感知($\alpha=0.816$),量表与上文相同。最后,参与者提供了他们的基本信息。

3.3 实验结果

(1)操纵检验。采用单因素方差分析检验算法可解释性的操纵是否成功,结果表明算法可解释性增强组(均值为4.707,标准差为0.879)的参与者对算法可解释性感知高于对照组(均值为3.926,标准差为0.930),$F(1,253)=45.520$,$P<0.001$;采用同样的方法检验拟人化程度操控是否成功,结果表明高拟人化组(均值为4.611,标准差为0.947)的参与者对AI拟人化的感知显著高于低拟人化组(均值为3.797,标准差为0.869),$F(1,253)=48.972$,$P<0.001$。

(2)主效应检验。借鉴Preacher和Hayes[35]提出的Bootstrap方法,使用Process程序检验调节效应。选择偏差校正的非参数百分位法进行取样,样本数为5 000,采用模型8[用于分析调节变量对自变量到中介变量(a路径)和中介变量到因变量(b路径)这两条路径的调节作用]验证H3。结果表明,算法可解释性与拟人化程度的交互项对继续使用意愿的回归系数为0.309,$P=0.016<0.05$,95%置信区间为[0.057,0.560],不包含0,说明拟人化程度的调节作用显著,假设H3a成立。其中,当拟人化程度较低时,算法可解释程度对继续使用意愿的影响不显著,$P=0.253>0.05$,95%置信区间为[−0.071,271],包含0;当拟人化程度较高时,算法可解释程度对继续使用意愿的回归系数为0.408,$P<0.010$,95%置信区间为[0.220,0.296],不包含0。采取单因素方差分析验证,结果同样成立,$F(1,253)=23.675$,$P<0.001$,$\eta_P^2=0.090$,AI助手拟人化程度对算法可解释性程度对服务失败下消费者继续使用意愿的影响有显著影响(图4)。

进一步检验拟人化程度对于可控性感知中介作用的调节机制,判定指数为0.490,95%置信区间为[0.273,0.727],不包含0,说明拟人化程度对可控性感知的调节作用显著,假设H3b成立。其中,当拟人化程度较低时,消费者可控性感知的中介效应不显著,95%置信区间为[−0.108,0.235],包含0;当拟人化程度较高时,消费者可控性感知的中介效应显著,95%置信区间为[0.389,0.719],不包含0。采取单因素方差分析验证,结果同样成立,$F(1,253)=20.074$,$P<0.001$,$\eta_P^2=0.077$,AI助手拟人化程度对可控性感知在上述影响的中介作用有影响(图5)。

综上,假设H3得到验证。

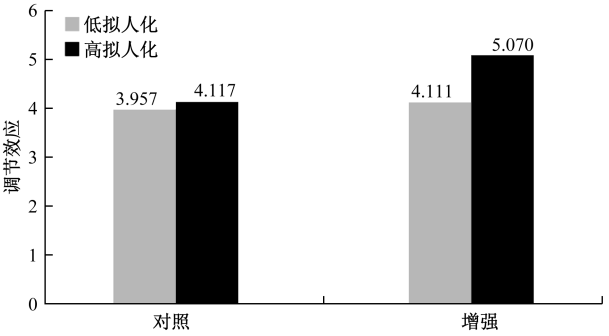

图4 AI拟人化程度对直接影响的调节效应

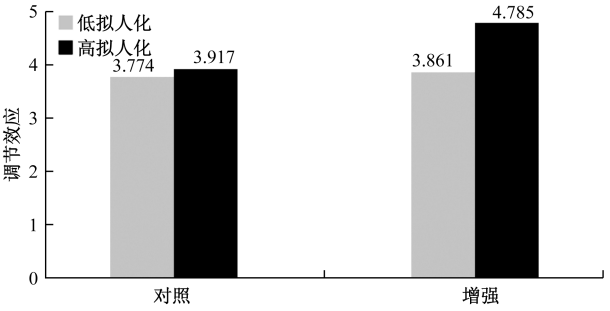

图5 AI拟人化程度对可控性感知中介作用的调节效应

4 结论与启示

4.1 结论

本文探讨算法可解释性对服务失败下消费者继续使用意愿的影响,并考察可控性感知的中介作用和拟人化程度的调节作用。研究结果表明,首先,当算法决策的可解释性通过事后解释得到增强时,能提高消费者在AI服务失败继续使用的意愿;其次,算法可解释性的增强提高了消费者在服务情境中的掌控力和控制力,可控性感知在算法可解释性程度与继续使用意愿之间起部分中介作用;最后,AI拟人化程度能够调节算法可解释性程度与继续使用意愿的关系,相对于高拟人化程度,算法可解释的积极效应会在AI拟人化程度较低时消失。

4.2 理论贡献

(1)拓宽了AI服务失败情境下自主补救的研究范围。在AI服务失败的复杂情境下,将研究焦点聚焦至算法可解释性这一前沿维度,打破了以往学者多聚焦于AI外观、交互方式及回应策略研究AI自主补救的局限。通过揭示算法“黑箱”效应对消费者负面情绪的放大作用,以及可解释性如何作为缓冲剂缓解这些情绪,不仅丰富了AI自主补救

策略的理论框架，更为探索AI服务质量的深层优化路径提供了坚实的理论支撑。

(2)丰富了算法可解释性对服务失败下消费者反应的研究视角。以往学者从归因视角解释算法可解释性对服务失败下消费者反应的积极效应，本文从控制感视角阐明了可控性感知起到的中介作用，扩展了补偿性控制理论的适用领域。

(3)从拟人化视角，拓展了算法可解释积极效应的理论边界。以往学者主要从任务类型、智能阶段和失败来源等方面探究算法可解释性积极效应的边界条件，本文从拟人化视角证实了AI拟人化程度的调节作用，即通过事后解释提高算法决策可解释性对消费者服务失败情境下继续使用意愿的积极影响在AI拟人化较低时(相对于较高)会减弱。

4.3 管理启示

人工智能的引进使服务营销的底层逻辑发生改变[36]，虽然AI服务可以为企业提升效率，减少人力成本。但AI算法的“黑箱”问题仍困扰着消费者，尤其是在服务失败下会加剧消费者的负面反应。因此，本文对企业缓解AI服务失败下负面反应提供了以下建议。

(1)企业应当提高算法的透明度和可解释性，尤其是在服务失败时，通过提供清晰的解释来增强消费者的信任和满意度。同时，在事后解释时，需考虑用户群体的多样性，制定差异化的解释策略，确保各层次用户都能有效理解算法决策逻辑，从而增强其对AI服务的掌控感和满意度，在服务失败时能够更好地接受解释和补救措施。

(2)企业应把拟人化作为设计AI助手的核心要素。拟人化作为连接技术与人类情感的桥梁，其重要性不言而喻。服务提供商可以通过调控AI助手的外观、语言风格及情感表达等控制其拟人化程度，提升AI助手的亲和力与可信度。在服务失败情境下，高拟人化的AI助手更能吸引消费者的注意力，促进其对解释信息的积极加工，进而缓解负面反应，并对整体服务拥有更高的控制感感知。

4.4 研究局限和未来展望

本文存在一定的局限性。首先，服务失败会使消费者产生负面的情绪、态度和行为，但本文关注的是消费者服务失败后的继续使用意愿，并未将消费者的情绪和态度纳入考量，而消费者的情绪和态度在某种程度上可能会影响消费者的后续行为。未来研究可以将消费者的情绪和态度纳入研究模型，提高研究的稳健性。其次，本文只探讨了拟人化在算法可解释性积极效应的边界条件，未来研究可以更深入的探究其他可能存在的边界条件。最后，本文的实验采用的都是情境代入的方法，虽尽可能让被试代入真实情境，但仍然缺乏一定的真实性。未来研究可以采用真实情境模拟或田野实验等更为接近现实的方法对本研究内容进行补充和深入。

参考文献

[1] 杜维，陈鑫. 运气还是能力？基于自我认知视角的智能服务算法偏差对消费者使用意愿的影响研究[J/OL]. 南开管理评论，1-40[2024-09-19]. http://kns.cnki.net/kcms/detail/12.1288.F.20240607.1636.002.html.

[2] 道阻且长，行而不辍 中国人工智能产业研究报告[C]//2023艾瑞咨询3月研究报告会论文集. 上海：上海艾瑞市场咨询有限公司，2023：112.

[3] 吕兴洋，杨玉帆，许双玉，等. 以情补智：人工智能共情回复的补救效果研究[J]. 旅游学刊，2021，36(8)：86-100.

[4] 薛哲，陈小云，李永诚，等. 智能机器人服务失败与顾客持续使用：机器人共情的补救效果研究[J]. 企业经济，2023，42(5)：71-81.

[5] CUI J，ZHANG M，ZHONG J. When frontline robots emerge：the double-edged-sword effect of anticipated trust on intention to switch brands after service failure[J]. Journal of Service Theory and Practice，Emerald Publishing Limited，2023，33(6)：842-872.

[6] SHIN D. The effects of explainability and causability on perception，trust，and acceptance：implications for explainable AI[J]. International Journal of Human-Computer Studies，2021，146：102551.

[7] GAI P J，PUNTIONI S. Language and consumer dishonesty：a self-diagnosticity theory[J]. Journal of Consumer Research，2021，48(2)：333-351.

[8] 罗映宇，朱国玮，钱无忌，等. 人工智能时代的算法厌恶：研究框架与未来展望[J]. 管理世界，2023，39(10)：205-233.

[9] CHEN C. How consumers respond to service failures caused by algorithmic mistakes：the role of algorithmic interpretability[J]. Journal of Business Research，2024，176：114610.

[10] CHOIS S，MATTILA A S，BOLTON L E. To err is human(-oid)：how do consumers react to robot service failure and recovery? [J]. Journal of Service Research，SAGE Publications Inc，2021，24(3)：354-371.

[11] 孔祥维，唐鑫泽，王子明. 人工智能决策可解释性的研究综述[J]. 系统工程理论与实践，2021，41(2)：524-536.

[12] 李伟卿，王伟军，黄炜，等. 可解释信息推荐研究综述

[J]. 情报学报, 2023, 42(7): 870-882.
[13] 高广尚. 可解释推荐模型中的可解释性方法研究综述[J]. 数据分析与知识发现, 2024, 8(S1): 6-19.
[14] 黄雯, 黄芊卉. 服务机器人拟人化对顾客参与服务补救意愿的影响[J]. 中小企业管理与科技, 2024(4): 42-46, 157.
[15] 高记, 冯婧雯. 不利结果下算法决策对员工公平感的双刃剑效应: 基于归因理论的视角[J]. 中国人力资源开发, 2024, 41(5): 36-53.
[16] LANDAU M J, KAY A C, WHITSON J A. Compensatory control and the appeal of a structured world[J]. Psychological Bulletin, US: American Psychological Association, 2015, 141(3): 694-722.
[17] 刘德文, 姜明刚. 机器人服务失败和补救研究进展与述评[J]. 科学与管理, 2024, 44(2): 92-102.
[18] 王海忠, 谢涛, 詹纯玉. 服务失败情境下智能客服化身拟人化的负面影响: 厌恶感的中介机制[J]. 南开管理评论, 2021, 24(4): 194-206.
[19] RIQUEL J, BRENDEL A, HILDEBRANDT F, et al. "F * * * you!": an investigation of humanness, frustration, and aggression in conversational agent communication[R]. Shanghai: ICIS, 2021.
[20] LONGONI C, CIAN L, KYUNG E J. Algorithmic transference: people overgeneralize failures of AI in the government[J]. Journal of Marketing Research, SAGE Publications Inc, 2023, 60(1): 170-188.
[21] 曹忠鹏, 马慧楠, 严兴全. 聊天机器人服务失败中拟人化对顾客反应的影响[J]. 管理科学, 2023, 36(1): 106-118.
[22] LIU J, XU X. Humor type and service context shape AI service recovery[J]. Annals of Tourism Research, 2023, 103: 103668.
[23] SCHUTTE N, NAMEE B M, KELLEHER J. Robot perception errors and human resolution strategies in situated human-robot dialogue[J]. Advanced Robotics, 2017, 31: 243-257.
[24] 梁海双, 陈佳琪. 生成式人工智能应用于医疗领域的伦理问题研究[J]. 锦州医科大学学报(社会科学版), 2024, 22(3): 19-22.
[25] 孟韬, 陈梦圆, 张天锴, 等. 交互失误情境下交互式人工智能拟人化的负面影响: 基于ChatGPT和搜索引擎的实验证据[J]. 情报理论与实践, 2024, 47(1): 84-91.
[26] GUIDOTTIi R, MONREALE A, RUGGIERI S, et al. A survey of methods for explaining black box models[J]. ACM Comput. Surv, 2018, 51(5): 1-42.
[27] KENNY E M, FORD C, QUINN M, et al. Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies[J]. Artificial Intelligence, 2021, 294: 103459.
[28] 陈煊, 郝喜玲, 杜晶晶, 等. 失败恐惧对创业企业成长的积极效应: 基于情绪认知评价理论的视角[J]. 研究与发展管理, 2024, 36(5): 39-52.
[29] 刘满芝, 何立立. "享受现在"还是"投资未来"? 自然资源稀缺提醒对亲环境行为的影响机制研究[J]. 中国矿业大学学报(社会科学版), 2024, 26(4): 173-188.
[30] BURGER J M, COOPER H M. The desirability of control[J]. Motivation and Emotion 1979, 3: 381-393.
[31] 张苏申, 贺慧颖. 外控制点与员工亲组织非伦理行为: 一个跨层的被调节中介[J]. 青海师范大学学报(社会科学版), 2023, 45(4): 80-91.
[32] 李茹, 陈嘉茜, 赵曙明, 等. 发展型人力资源管理感知对员工创造力的影响机制: 社会信息加工理论视角[J]. 科技进步与对策, 2024, 41(22): 152-160.
[33] EPLEY N, WAYTZ A, CACIOPPO J T. On seeing human: a three-factor theory of anthropomorphism[J]. Psychological Review, US: American Psychological Association, 2007, 114(4): 864-886.
[34] 聂春艳, 汪涛. AI让人更环保? AI推荐对消费者的绿色产品购买意愿的影响[J/OL]. 南开管理评论, 1-17[2024-09-19]. http://kns.cnki.net/kcms/detail/12.1288.F.20240404.1719.002.html.
[35] PREACHER K J, HAYES A F. SPSS and SAS procedures for estimating indirect effects in simple mediation models[J]. Behavior Research Methods, Instruments, & Computers, 2004, 36(4): 717-731.
[36] 杜建刚, 赵欢, 苏九如, 等. 服务智能化下的顾客行为: 研究述评与展望[J]. 外国经济与管理, 2022, 44(3): 19-35.

Positive Effects of Algorithmic Explainability in the Context of AI Assistant Service Failure

ZHOU Jianshe, LIU Zifeng, ZHOU Ye

(School of Economics and Management, Nanchang Hangkong University, Nanchang 330063, China)

Abstract: The failure of AI(artificial intelligence)assistant services is essentially an algorithmic failure, and the "black box" nature of the algorithmic decision-making process exacerbates consumers' negative reactions to service failures. A between-subjects experimental approach was adopted to explore the impact mechanism of enhancing algorithmic explainability from the perspective of perceived control through explainable artificial intelligence methods (such as post hoc explanations) on consumer behavior in the context of service failure, as well as the boundary conditions. It is found that when algorithmic explainability is enhanced (compared to the control), consumers' continued intention to use despite service failure is improved, and consumers' perceived control plays a partial mediating role in this process. However, the above effects are not significant when the anthropomorphism level of the AI assistant is low (compared to high).

Keywords: AI(artificial intelligence) service failure; algorithmic explainability; sense of control