心理学探新 2025,Vol.45,No.4,291 -299
PSYCHOLOGICAL EXPLORATION

探索自动驾驶中决策主体与偏好对决策权、信任和感知道德能动性的影响

董超武 游旭群 刘 煜 李 瑛
(陕西师范大学心理学院,西安 070062)

摘 要:在行人牺牲(实验1)和驾驶员牺牲(实验2)情境中考察个体对决策主体的认可水平及决策主体和决策偏好对信任和感知道德能动性的影响。结果发现,被试对自动驾驶系统作为决策主体的认可水平、信任水平和感知道德能动性都低于人类驾驶员,对功利主义的信任水平和感知道德能动性高于道义主义。在驾驶员牺牲场景中,人们更信任功利主义偏好的人类驾驶员,感知道德能动性水平也更高。感知道德能动性中介了认可水平与信任水平的关系。

关键词:自动驾驶;道德决策;道德偏好;人机信任;道德能动性
中图分类号:B842.5 **文献标志码**:A **文章编号**:1003 -5184(2025)04 -0291 -09

1 引言

人工智能(artificial intelligence,AI)正在赋予自动化系统自主决策能力,最典型的就是自动驾驶系统(autonomous driving system,ADS)。在实际使用ADS之前,明确使用过程中的道德决策主体归属和应秉持的伦理道德信念是必要的。

道德决策指个体在面对某种特定的道德困境及各种可能采取的行为方式时,对行为结果进行善恶判断,并评定其是否符合社会所尊崇的准则和规范做出选择的过程(Garrigan et al.,2018)。决策权的归属直接影响决策结果,决策主体的偏好直接决定了在道德困境情境下拯救和牺牲的对象。实际场景中的道德决策偏好通常划分为道义主义和功利主义(Greene et al.,2001)。道义主义的核心为责任,强调禁止违反基本权利和义务的信念,要求避免蓄意伤害无关人员(Kant,1976);功利主义注重行为结果,主张道德决策应基于最大收益原则(Bentham,1967),甚至认可通过故意牺牲少部分群体的福祉来维护多数群体的利益。驾驶过程中的决策不仅关乎司乘人员的安全,还影响着同一时空下的其他道路交通参与者。为了解公众对自动驾驶过程中进行道德决策的看法,Awad等人(2018)通过道德机器平台收集了233个国家和地区的数百万名参与者对不同场景下自动驾驶的道德决策4000余万次。结果发现,大多数人倾向于保护更多的生命,尤其是年轻的生命和女孩,这种对功利主义的偏好在后续研究中得到了验证(Faulhabe et al.,2019)。在涉及驾驶员自我牺牲的困境研究中,个体普遍展现出自我保护的倾向,即偏好选择遵循道义主义的ADS(Bruno et al.,2023),并在实际驾驶更频繁地接管车辆(Ju & Kim,2024)。Liu和Liu(2021)调查了580名潜在驾驶员对自动驾驶道德决策的偏好,发现人们更倾向于购买无条件保护驾驶员的ADS。一方面是因为个体将付费购买的ADS视为个人财产,认为其应优先保障自身安全;另一方面从驾驶员或乘客的视角出发,被试与其他交通参与者在道德困境中存在竞争和对立,优先保证自身生命安全是一种基本需求(Kallioinen et al.,2019)。综上,在自动驾驶场景中,人们因牺牲对象的变化(行人或驾驶员)而产生了不同的道德决策,因而有必要综合考量这两种典型的自动驾驶场景中人们对道德决策权的归属认可是否存在差异。

对于自动驾驶车辆而言,AI控制的ADS正在成为驾驶代理人,它们将逐步代替人类驾驶员进行决策,尤其是在需要快速反应的紧急情况下(Bhattacharya et al.,2023)。这就要求ADS不仅要具备人类的道德标准,还要能够根据具体情况进行理性的道德判断(Firt,2024),即拥有道德能动性(moral agency)。道德能动性指理解道德规范、遵循规范开展行动并对道德行为承担责任的能力(Himma,2009;Zafar,2024),是道德的能力的核心。随着AI的深入发展,在道德活动中,作为道德主体的人类开始依赖道德机器,这一趋势在人类面临复杂情境中的道德决策时尤为明显(Gonzalez Fabre et al.,2021)。正如个体在人类社会中的成长不仅依赖于个人经验和认知能力,也受到所处文化环境的影响

(Sugarman,2005),AI 的发展同样遵循这一模式。人们通过模拟人类行为逻辑,根据环境信息为机器设定执行道德活动的程序。由于人类道德行为动机具有复杂性和模糊性特征,由人类设计的人工智能道德代理不可避免地存在同样的问题,甚至产生了掩盖性逻辑(Vallor & Vierkant,2024)。在 AI 替代人类进行道德决策前,必须具备与人类等同级别的道德代理能力,人们感知到的道德能动性水平是评价其道德行为的重要依据(Bonnefon et al.,2024)。只有具备较高道德能动性的 AI 能妥善处置与道德决策受到的各类因素的影响,确保其做出符合人类利益的决策(Maninger & Shank,2022)。如此,各种智能系统才能融入人类社会,人们才能信任其进行道德决策,并进一步解决后续责任分配问题(Ju & Kim,2024)。当前,对人们为何拒绝机器具备道德决策权以及因机器进行道德决策而产生的负面反馈的原因集中在结果的影响,其中的影响机制尚不明确。因此,本研究计划引入感知道德能动性作为衡量人们对决策主体道德能力的评价指标。

自动化信任问题是实现安全有效的人机协同控制的最大挑战之一(董文莉,方卫宁,2020)。在人机协作中,借助 AI 处理数据的机器具备了超越人类生理极限的信息处理效率,在医疗、司法和军事等领域正在逐步成为决策权所有者。然而,长久以来的社会认知确立了人类作为道德决策的唯一主体,这导致了先进智能系统即使能规避人类的错误,也无法获取信任。加之为避免不利于自身的决策,人们偏好将道德决策权掌握在自己手中,也对自身偏好的决策角色所作的决策表现出更多的信任(Bigman & Gray,2018)。对于自动驾驶汽车而言,其道德决策影响着人们对其的信任水平。Yokoi 和 Nakayachi (2021b)发现人机道德价值相似性(被试与 ADS 是否做出了一致的决策)影响个体对 ADS 的信任水平和使用意愿。当 ADS 的道德判断与个体的选择一致时(尤其是道义主义的决策),个体对系统更加信任(Yokoi & Nakayachi,2021b)。这种信任水平的提升能够提高个体的使用意愿,表现为在驾驶过程中更长时间地使用自动驾驶功能(Schwarz et al.,2019),而缺乏信任可能直接促使驾驶员关闭 ADS (Nordhoff,2023)。

综上,本研究旨在探索自动驾驶使用过程中的道德决策权的归属、信任及两者之间的关系,通过引入感知道德能动性作为验证指标和中介变量探索其对决策的主体认可水平与信任水平之间的影响路径。综上,本研究在行人牺牲场景(实验 1)和驾驶员牺牲场景(实验 2)中检验了被试对决策主体(ADS 和人类驾驶员)的认可水平;考察了不同决策主体的不同偏好条件下的感知道德能动性和信任水平的差异;探索了感知道德能动性对认可水平与信任水平间关系的作用。实验采用的自动驾驶道德两难场景命名基于功利主义决策的牺牲对象。

2 实验 1:行人牺牲场景中的道德决策

实验 1 在行人牺牲场景中采用 2(决策主体:人类驾驶员,自动驾驶系统)×2(决策偏好:功利主义,道义主义)的组间实验设计,考察被试对不同决策主体的认可水平以及被试在不同决策主体采取不同决策偏好条件下的信任水平和感知道德能动性水平是否存在差异。

2.1 方法

2.1.1 被试

采用 G * Power 3.1 先验分析计算样本量(Faul et al.,2007),$f = 0.25$,$\alpha = 0.05$,$1 - \beta = 0.95$,得出至少需要 210 名被试。通过见数平台线上招募 260 名被试,男 116 人,女 144 人,评价年龄 24.60 ± 4.30 岁,243 人持有驾照。

2.1.2 材料

实验 1 对 Yokoi 和 Nakayachi(2021a)设计的自动驾驶道德两难场景进行改编,以第三人称视角进行阐述(图 1)。一辆自动驾驶汽车在规划道路上正常行进,在突然遭遇后方车辆撞击后发生失控,此时车上的决策主体(自动驾驶系统或人类驾驶员)需要对车辆前进方向做出决定(图 1a)。如果自动驾驶系统或人类驾驶员选择向右急转,则会撞向右前方一名行人,拯救正前方两名行人,代表其偏好功利主义(图 1b);如果决策主体选择继续前进,则会撞向正前方两名行人,拯救右前方一名行人,代表其偏好道义主义(图 1c)。为避免被试根据行人的社会特征做出歧视性决策(De Freitas & Cikara,2021),研究未对实验场景中的人物身份信息进行阐述。

2.1.3 程序

本实验通过见数平台在线开展数据收集工作。被试在阅读注意事项并签署知情同意书后被随机分配至四种场景中:人类驾驶员功利主义偏好、人类驾驶员道义主义偏好、自动驾驶系统功利主义偏好和自动驾驶系统道义主义偏好。正式实验中,被试需先阅读对应分组的场景图片材料(如图 1a)和文字

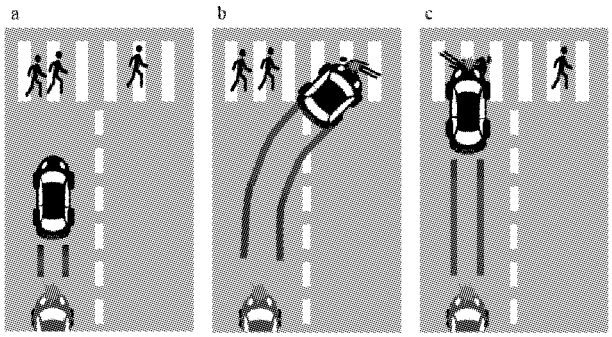

图 1　行人牺牲的驾驶道德两难场景

叙述:“一辆自动驾驶的黑色汽车正在道路上正常行驶。突然,一辆失控的红色汽车从后方高速撞击了该车辆,导致其失控。此时,失控的黑色车辆前方有两位行人。如果继续向前,车辆将会撞上这两位行人。要挽救这两位行人的唯一方法是向右急转方向。然而,这样做的话,车辆将会撞上一位行人。”在阅读完毕后,被试需要对这一情况下的决策主体认可水平进行评价。评估题目改编自 Bigman 和 Gray(2018),共 3 道题目,内部一致性系数 $\alpha = 0.91$,题目陈述时将决策主体前置,例如在人类驾驶员作为决策主体时,题项陈述为“驾驶员‘A 先生’比自动驾驶系统‘X’更适合做出决策”;当自动驾驶系统作为决策主体时,题项陈述为“自动驾驶系统‘X’比驾驶员‘A 先生’更适合做出决策”。随后,按照分组向被试呈现四种决策结果的图片材料和文字叙述。以人类驾驶员功利决策为例(如图 1b):“驾驶员 A 先生决定向右急转方向,牺牲右前方一名行人,挽救正前方两名行人。”被试根据结果评估决策主体的信任水平,改编自 Yokoi 和 Nakayachi (2021),共 3 道题目(例题:“驾驶员 A 先生是可靠的”),题目的内部一致性系数 $\alpha = 0.94$。随后,被试须回忆决策主体做出的决策以保证其正确理解场景内容,回答正确的被试进一步对决策主体的道德能动性进行评价,题目改编自 Banks(2019)开发的感知道德能动性量表的道德部分,该部分的方差解释率为 49.31%,共 6 题(例题:“驾驶员 A 先生有是非观念”),题目的内部一致性系数 $\alpha = 0.92$。最后,被试填写人口统计学信息,包括性别、年龄、是否持有驾照,之后向被试解释实验目的,并告知被试该场景设置为小概率事件,以减少实验内容对其造成的不适。本研究中的所有题目均为 5 点计分(1—完全不同意,2—有些不同意,3—中立,4—有些同意,5—完全同意),取各部分测量的平均分作为该部分的最终得分进行分析。

2.2　结果

2.2.1　对决策主体的认可水平

采用独立样本 t 检验分析被试对决策主体认可水平,结果显示被试对自动驾驶系统的认可水平(2.72 ± 1.26)显著低于人类驾驶员(3.63 ± 1.07),$t(259) = -6.13$, $p < 0.001$, $Cohen's\ d = -0.77$, 95% $CI[-1.20, -0.61]$,见图 2。

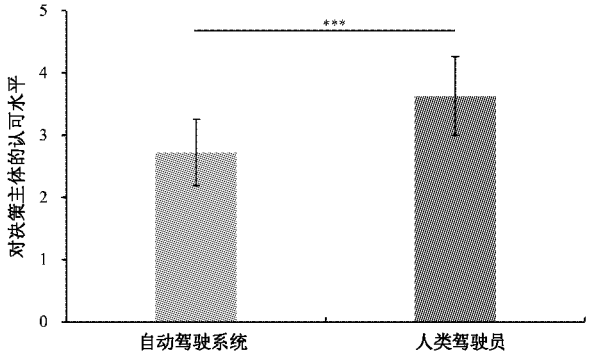

图 2　对决策主体的认可水平

注:误差线代表单个标准差,$^{*}p < 0.05$, $^{**}p < 0.001$, $^{***}p < 0.0001$,下同。

2.2.2　感知道德能动性

采用 2(决策主体:人类驾驶员,自动驾驶系统)×2(决策偏好:功利主义,道义主义)两因素完全随机方差分析。结果发现,不同决策主体的主效应显著,$F(1,257) = 22.71$, $p < 0.001$, $\eta_p^2 = 0.081$,被试对自动驾驶系统的感知道德能动性(2.88 ± 1.06)水平显著低于人类驾驶员(3.67 ± 1.02);决策偏好的主效应显著,$F(1,257) = 86.23$, $p < 0.001$, $\eta_p^2 = 0.252$,被试对道义主义偏好决策的感知道德能动性水平(2.56 ± 1.06)显著低于功利主义偏好(3.63 ± 0.82);交互效应不显著,$F(1,257) = 0.13$, $p = 0.721$, $\eta_p^2 = 0.000$,见图 3。

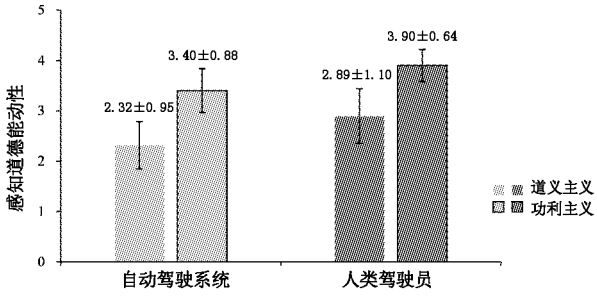

图 3　感知道德能动性

2.2.3　信任水平

采用 2(决策主体:人类驾驶员,自动驾驶系统)×2(决策偏好:功利主义,道义主义)两因素完全随

机方差分析，当违反球形假设时，使用 Greenhouse – Geisser 调整自由度。结果发现，不同决策主体的主效应显著，$F(1,259) = 10.80$，$p < 0.01$，$\eta_p^2 = 0.040$，被试对自动驾驶系统的信任水平（2.71 ± 1.28）显著低于人类驾驶员（3.14 ± 1.30）；决策偏好的主效应显著，$F(1,259) = 94.37$，$p < 0.001$，$\eta_p^2 = 0.269$，被试对道义主义的信任（2.22 ± 1.14）显著低于功利主义偏好（3.55 ± 1.10）；交互效应不显著，$F(1,259) = 0.019$，$p = 0.890$，$\eta_p^2 = 0.040$，见图 4。

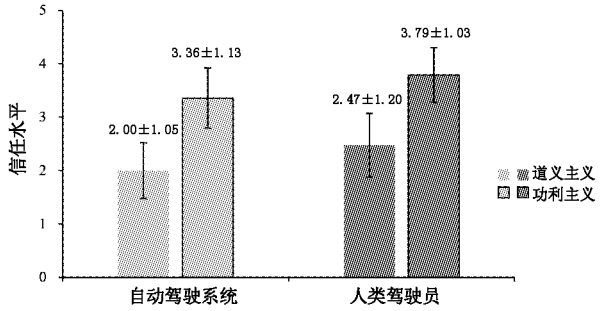

图 4 信任水平

2.3 讨论

研究 1 在涉及行人牺牲的场景中发现被试认为人类驾驶员更应拥有道德决策权，并且展现出了对功利主义的偏好，表现为给予人类驾驶员和功利主义更多的信任。此外，被试对不同决策主体和决策偏好的道德能动性感知也表现出同信任相似的结果。上述结果表明，目前人们仍然倾向拒绝将机器视为道德主体。实际驾驶场景中的道德决策还会涉及驾驶员自身的安全，典型的自动驾驶道德困境还包括了驾驶员牺牲场景，实验 2 进一步在驾驶员牺牲场景中考察上述发现的差异。

3 实验 2：驾驶员牺牲场景中的道德决策

3.1 方法

3.1.1 被试

同实验 1，至少需要 210 名被试。通过见数平台招募 258 名被试，男 123 人，女 135 人，平均年龄 24.80 ± 5.07 岁，249 人持有驾照。所有被试均未参与实验 1，通过筛选作答 IP 与账号进行控制。

3.1.2 材料

使用自我牺牲驾驶道德两难场景，同样以第三人称视角阐述（图 5）。一辆自动驾驶汽车在临崖道路上正常行进，在遭遇后方车辆撞击后失控，此时车上人类驾驶员或自动驾驶系统需要做出决定（图 5a）。如果驾驶员或自动驾驶系统选择向右急转，则会驶下悬崖牺牲自己，拯救正前方两名行人（图 5b）；如果选择继续前进，则会撞向正前方两名行人，避免驶下悬崖牺牲自己（图 5c）。

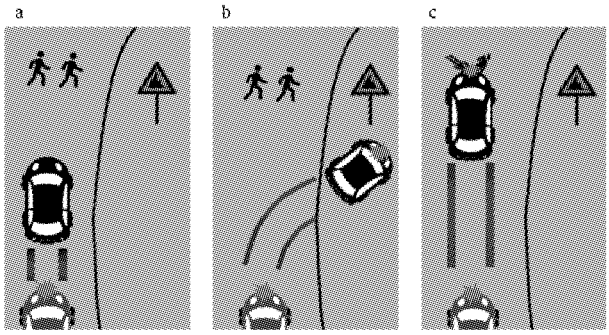

图 5 驾驶员牺牲的驾驶道德两难场景

3.1.3 程序

首先，向被试呈现场景图片材料（如图 6）与文字叙述：“一辆自动驾驶的黑色汽车正在临崖道路上正常行驶。突然，一辆失控的红色汽车从后方高速撞击了该车辆，导致其失控。此时，失控的黑色车辆前方有两位行人。如果继续向前，车辆将会撞上这两位行人。要挽救这两位行人的唯一方法是向右急转方向。然而，这样做的话，车辆及驾驶员将会冲下悬崖。”被试在阅读完毕后，被要求对决策主体的认可水平进行评价，三道测量题目的 $\alpha = 0.91$。随后，按照分组向被试呈现决策结果的图片材料，以人类驾驶员功利决策为例（如图 5*b*）和文字叙述：“驾驶员 *A* 先生决定向右急转方向，牺牲自己，挽救正前方两名行人。”被试根据结果对决策主体的信任水平（$\alpha = 0.94$）和道德能动性水平（$\alpha = 0.92$）进行评价，题目内容与实验 1 一致。

3.2 结果

3.2.1 对决策主体的认可水平

独立样本 *t* 检验结果发现，被试对自动驾驶系统的认可水平（2.28 ± 1.01）显著低于对人类驾驶员的认可水平（3.98 ± 0.84），$t(256) = -13.99$，$p < 0.001$，Cohen’s $d = -1.75$，95% CI [−1.47，−1.75]，见图 6。

3.2.2 信任水平

两因素完全随机方差分析分析信任水平。结果显示，不同决策主体的主效应不显著，$F(1,254) = 2.31$，$p = 0.13$，$\eta_p^2 = 0.040$；决策偏好的主效应不显著，$F(1,254) = 3.47$，$p = 0.06$，$\eta_p^2 = 0.013$；交互效应显著，$F(1,254) = 12.69$，$p < 0.001$，$\eta_p^2 = 0.048$。简单效应分析显示，在自动驾驶系统决策条件下，被试对道义主义偏好的信任水平（3.01 ± 1.29）与功利主义偏好（2.74 ± 1.40）不存在显著差异，*F*

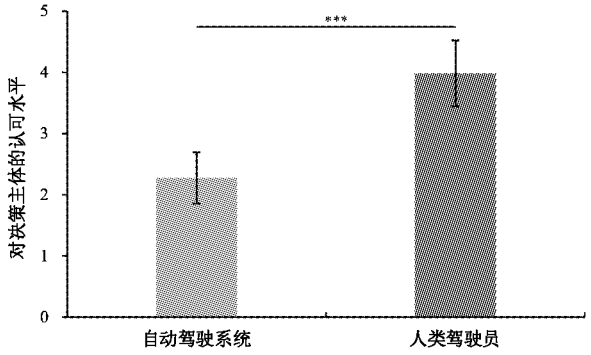

图6　对决策主体的认可水平

(1,138) = 1.55, p = 0.214, η_p^2 = 0.006；在人类驾驶员决策条件下，决策偏好的简单效应显著，F(1,116) = 13.75, p < 0.001, η_p^2 = 0.051，相较做出道义主义选择的驾驶员(2.68 ± 1.29)，被试更相信做出功利主义决策的人类驾驶员(3.55 ± 1.20)，见图7。

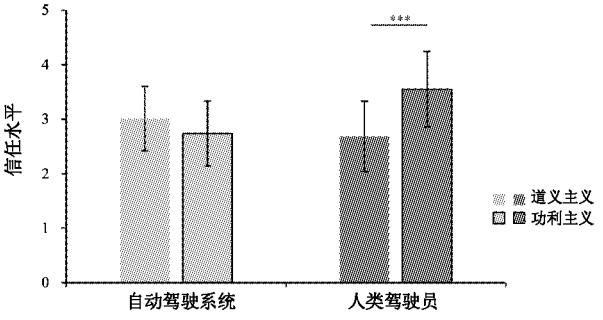

图7　信任水平

3.2.3　感知道德能动性

两因素方差分析结果显示不同决策主体的主效应显著，F(1,254) = 13.30, p < 0.001, η_p^2 = 0.050，被试对自动驾驶系统的道德能动性评价(3.01 ± 1.02)显著低于人类驾驶员(3.49 ± 1.15)；决策偏好的主效应显著，F(1,254) = 108.58, p < 0.001, η_p^2 = 0.299，被试对道义主义偏好的道德能动性评价(2.71 ± 0.96)显著低于功利主义(3.83 ± 0.93)；交互效应显著，F(1,256) = 17.95, p < 0.001, η_p^2 = 0.082。简单效应分析显示，在自动驾驶系统决策条件下，决策偏好的简单效应显著，F(1,136) = 17.13, p < 0.001, η_p^2 = 0.063，被试对功利主义决策的道德能动性评价(3.40 ± 1.01)显著高于道义主义决策(2.77 ± 0.94)；在人类驾驶员决策条件下，决策偏好的简单效应显著，F(1,118) = 107.91, p < 0.001, η_p^2 = 0.298，被试对功利主义决策的道德能动性评价(4.33 ± 0.47)显著高于道义主义决策(2.65 ± 1.00)，见图8。

在道义主义偏好下，决策主体的简单效应不显著，F(1,126) = 0.63, p = 0.428, η_p^2 = 0.002；在功利主义偏好下，决策主体的简单效应显著，F(1,128) = 35.69, p < 0.001, η_p^2 = 0.123，对自动驾驶系统的道德能动性感知(3.40 ± 1.01)显著低于人类驾驶员(4.33 ± 0.47)。

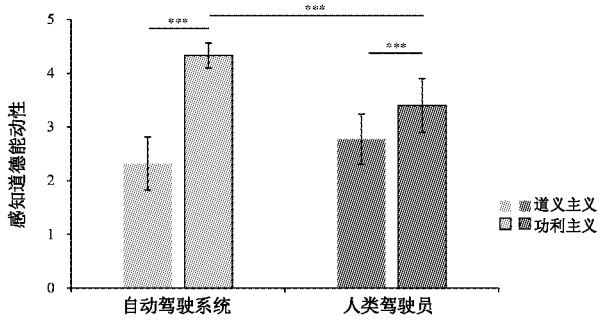

图8　感知道德能动性

4　感知道德能动性的中介作用检验

除作为衡量决策主体和决策偏好影响的指标外，本研究引入感知道德能动性的目的还在于验证人们信任某项道德的前提是感受到足够能动思考过程。整合两项实验数据，采用 PROCESS 宏的 Model 4 进行中介检验，以认可水平为自变量，信任水平为因变量，感知道德能动性为中介变量，决策主体、决策偏好和决策场景为控制变量，设定 Bootstrap5000。结果显示，认可水平对信任的直接效应显著(β = 0.22, P < 0.001, $BootSE$ = 0.43, $BootCI$ = [0.14, 0.31])，感知道德能动性在决策主体的认可水平和信任水平间的中介效应显著(β = 0.10, P < 0.001, $BootSE$ = 0.27, $BootCI$ = [0.05, 0.16])，占总效应的31.73%，见图9。

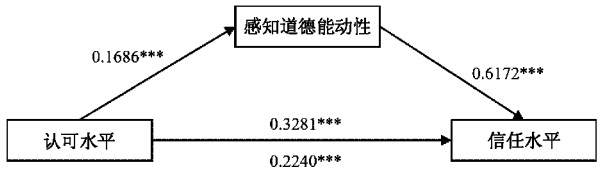

图9　感知道德能动性在对决策主体的认可水平和信任水平间的中介作用

5　讨论

本研究发现，在自动驾驶道德困境中，相较自动驾驶系统和道义主义决策，被试倾向于人类驾驶员作为决策主体并偏好功利主义决策，表现为更高的信任和道德能动性感知。相比仅涉及行人的牺牲场景，涉及驾驶员牺牲的场景中存在更多的交互效应。此外，感知道德能动性中介了对决策主体的认可水平和信任水平间的关系，具体分析如下。

5.1 ADS 暂时无法替代人类成为道德决策主体

本研究中，被试在两个典型的自动驾驶道德困境中对自动驾驶系统拥有决策权的认可水平都显著低于人类驾驶员，这一结果与 Bigman 和 Gray (2018)的发现相似。为提高人们对 AI 成为道德决策人的认可，研究人员尝试了改变 AI 角色、提高心智感知和专业知识等，但都收效甚微。其原因可能在于以下两点：(1)自动驾驶系统缺少人类的成长经历，依靠代码建立的观念与人类社会依靠遗传和社会学习的发展过程不同(Davis et al. ,2018)，无法兼顾道德推理和道德情绪，不能保证决策结果符合道德规范；(2)当前的自动驾驶系统依赖有限算法和数据，并不具备自由意志，无法被人们视为具备真正理解决策结果的道德主体。综上，在自动驾驶场景中，当前的自动驾驶系统还未达到合格的道德决策者标准，道德决策权还应交还给人类以防止负面影响。

5.2 决策主体及其偏好对信任水平的影响

对自动化系统的信任水平直接决定了是否使用、如何使用以及人机协作的安全绩效。本研究发现，在涉及行人牺牲的场景中，被试对自动驾驶系统的信任水平显著低于人类驾驶员，对道义主义的信任显著低于功利主义偏好。根据理性与情绪双过程理论(Greene et al. ,2001)，典型的功利判断由受控的认知过程驱动，而典型的道义判断则由自动情绪反应驱动。一方面，当自动驾驶系统做出决策时，被试可能会假设其不具备道德情绪，而认为其更有可能做出功利主义决策。当被试无法感知到自动驾驶系统的温暖(warmth)时，自然地认为其不具备同情被牺牲者的能力，所以给予较低的信任水平。另一方面，从理性推断角度出发，功利主义决策能够拯救更多的生命，而道义主义决策可能造成更多伤亡，因而对道义主义的信任水平相对较低。此外，决策主体与其偏好对信任水平影响的交互效应并不显著，表明被试对道德决策的信任评估过程中可能将决策主体和偏好进行了独立评估。从实验设计来看，本研究强调了驾驶员与自动驾驶系统具备同等的驾驶水平，而能力是信任评价的核心要素(董文莉，方卫宁，2020)。此外，为避免过度的偏好差异，不同选择造成牺牲的人数仅相差 1，这也可能削弱了交互作用，加剧了决策主体和决策偏好效应的相对独立。需要指出的是，现有研究发现人们通常希望自动驾驶采取功利主义决策(Malle et al. ,2025)，但这一效应通常局限于在道德决策过程之中，其影响能否延伸到宏观的自动化信任评价仍需进一步考察。从社会整体利益的角度出发，自动驾驶系统应持有功利主义信念，以达成拯救更多人的目的。然而，在当被试扮演驾驶员时，便对有可能牺牲自己的功利主义持消极态度，转而偏好道义主义(Liu & Liu,2021)。本研究在涉及驾驶员牺牲的场景中，仅发现在人类驾驶员决策的条件下，被试对功利主义的信任水平高于道义主义，这一发现与 Bruno 等人(2024)的研究结果基本一致。人们在观察涉及牺牲小我以拯救大我的道德困境时，对自动驾驶系统和驾驶员的决策可能存在两种独立的评价标准。对于驾驶员而言，被试会优先考虑摆脱道德情绪的影响，从理性推理出发希望其更多地做出功利主义决策，侧面体现了对人类较高的道德能动性水平的认可。对于自动驾驶系统而言，被试可能既希望其拯救一同操控车辆的驾驶员，也希望其拯救更多的生命，因而对其决策偏好处于相对中立的状态。需要注意的是，也有研究指出这种自我牺牲的倾向部分源于人们对自我保护这一“自私”选择的羞耻(Bovesi et al. ,2025)，部分掩盖了对非功利主义的偏好，导致了对人类驾驶员和对自动驾驶系统的双重标准。此外，相比仅涉及行人牺牲的场景，在涉及驾驶员牺牲的场景中发现了更多差异。虽然使用了第三人称叙述，但被试在驾驶员牺牲场景中仍可能倾向于带入并共情驾驶员，产生自我保护倾向于。然而，被试对功利主义，即拯救更多人的偏好导致其无法确定该由谁做出何种决策，从而选择信任做出功利主义决策的自动驾驶系统，这种现象说明自动驾驶系统需要具备多元的决策策略，例如添加“道德旋钮”进行调整。

5.3 决策主体及其偏好对感知道德能动性的影响

在自动驾驶道德决策研究中，引入感知道德能动性作为检验指标尚属首次。在本研究中，被试感知的自动驾驶系统和道义主义的道德能动性在实验 1 和实验 2 中都显著低于人类驾驶员，侧面支持了个体对人类拥有决策权的认可和对功利主义的偏好。人类拥有作为道德主体的中心性，因为我们能够自由选择、思考和正确运用道德规则等能力，具备承担道德责任的条件(田训龙，2024)。相比之下，目前的自动驾驶系统大致基于行为克隆、基于状态的建模和强化学习三类算法(Siegel & Pappas, 2023)，无法在避免碰撞的基础上进一步权衡，例如通过自身承担更高的风险来降低其他道路使用者的

风险(Krügel & Uhl,2024)。在实验2中,被试认为牺牲自己拯救两名行人的人类驾驶员拥有更高的道德能动水平,体现了传统文化中舍生取义的价值观。相较西方文化,我国集体主义文化背景人们的确对智能体的心智能力有更高的估计,对其承担道德责任要求也更高(闫霄 等,2024),更容易因决策结果带来的负面影响加剧道德决策的人类中心主义倾向。此外,感知道德能动性在被试对决策主体的认可水平和信任水平间的显著中介作用。被试对决策主体的认可水平通过影响被试对其道德能动性的评估,进而影响对决策主体的信任水平。在提升能力的基础上,还需丰富、提高决策主体能动性表现,有利于建立信任关系。

综上,本研究发现被试在驾驶道德场景中倾向于人类驾驶员进行道德决策,偏好功利主义,表现为更高的信任水平和感知道德能动性水平。上述发现有助于相关人员深入了解自动驾驶的道德伦理影响,为相关治理框架提供了参考。

5.4　不足与展望

本研究存在以下不足和改进之处:

(1)根据传统的道德困境范式改良的"无人驾驶困境"与现实道路情况存在差异(De Freitas et al.,2020)。在实际交通情况下,驾驶员可能需要快速做出判断。因此,未来可以使用驾驶模拟器并结合虚拟现实技术来提高生态效度。

(2)本研究采用的在线情景问卷,无法客观地区分被试是直觉作答还是理性推断,后续研究可考虑收集生理数据进行筛查。

(3)传统道德决策分类方法可能无法满足自动驾驶场景的需求,是否应采取"最小化伤害"作为标准存在争议。尤其是在实验2中,当需要权衡驾驶员牺牲或牺牲的决策时,以利益最大化来定义功利主义的做法可能并不完全合适,后续研究可跳出传统的决策分类方式,应拓展不同决策结果。

(4)在人机关系背景下对感知道德能动性的中介作用检验尚属首次,后续研究应进一步检验其作用的稳定性。

6　结论

(1)自动驾驶系统不被视为与人类等同的道德决策者。

(2)实验1中(行人牺牲的场景),被试对自动驾驶系统的信任水平显著低于人类驾驶员,对道义主义的信任显著低于功利主义偏好。实验2(驾驶员牺牲场景)仅发现被试更对功利主义偏好的人类驾驶员的信任水平显著高于道义主义。

(3)就感知道德能动性而言,自动驾驶系统和道义主义显著低于人类驾驶员和功利主义。在实验2中,相比道义主义决策,做出功利主义决策的人类驾驶员的道德能动性评价更高,且显著高于功利主义的自动驾驶系统。

(4)感知道德能动性在对决策主体的认可水平和信任水平间有显著中介作用。

参考文献

董文莉,方卫宁.(2021).自动化信任的研究综述与展望.*自动化学报*,*47*(6),18.

田训龙.(2024).智能道德主体的潜在风险探析.*道德与文明*,(4),128－139.

闫霄,莫田甜,周欣悦.(2024).中西方文化差异对虚拟人道德责任判断的影响.*心理学报*,*56*(2),161－178.

Awad,E.,Dsouza,S.,Kim,R.,Schulz,J.,Henrich,J.,Shariff,A.,Bonnefon,J.－F.,& Rahwan,I.(2018).The Moral Machine experiment.*Nature*,*563*(7729),Article 7729. https://doi.org/10.1038/s41586－018－0637－6

Banks,J.(2019).A perceived moral agency scale:Development and validation of a metric for humans and social machines.*Computers in Human Behavior*,*90*,363－371. https://doi.org/10.1016/j.chb.2018.08.028

Bentham,J.(1967).*An introduction to the principles of morals and legislation*.In Y.Seki(Ed.),*Great books of the world 38*(pp.69－210;S.Yamashita,Trans.).Chuo Koron.(Original work published 1789).

Bigman,Y.E.,& Gray,K.(2018).People are averse to machines making moral decisions.*Cognition*,*181*,21－34. https://doi.org/10.1016/j.cognition.2018.08.003

Bhattacharya,P.,Saraswat,D.,Savaliya,D.,Sanghavi,S.,Verma,A.,Sakariya,V.,Tanwar,S.,Sharma,R.,Raboaca,M.S.,& Manea,D.L.(2023).Towards Future Internet:The Metaverse Perspective for Diverse Industrial Applications.*Mathematics*,*11*(4).https://doi.org/10.3390/math11040941

Bonnefon,J.－F.,Rahwan,I.,& Shariff,A.(2024).The Moral Psychology of Artificial Intelligence.*Annual Review of Psychology*,*75*,653－675. https://doi.org/10.1146/annurev－psych－030123－113559

Bovesi,A.,Calabretto,A.,Stroppa,A.,Adamo,G.M.,Canepa,F.,& Guidi,S.(2025).The autonomous vehicle dilemma:Passenger(s)versus pedestrian(s).*Behaviour & Information Technology*,*44*(6),1155－1168. https://doi.org/10.1080/0144929X.2024.2433037

Bruno,G.,Spoto,A.,Lotto,L.,Cellini,N.,Cutini,S.,& Sarlo,

M. (2023). Framing self – sacrifice in the investigation of moral judgment and moral emotions in human and autonomous driving dilemmas. *Motivation and Emotion*, *47* (5), 781 – 794. https://doi.org/10.1007/s11031 – 023 – 10024 – 3

Bruno, G., Spoto, A., Sarlo, M., Lotto, L., Marson, A., Cellini, N., & Cutini, S. (2024). Moral reasoning behind the veil of ignorance: An investigation into perspective – taking accessibility in the context of autonomous vehicles. *British Journal of Psychology*, *115*, 90 – 114. https://doi.org/10.1111/bjop.12679

Davis, T., Hennes, E. P., & Raymond, L. (2018). Cultural evolution of normative motivations for sustainable behaviour. *Nature Sustainability*, *1* (5), 218 – 224. https://doi.org/10.1038/s41893 – 018 – 0061 – 9

De Freitas, J., & Cikara, M. (2021). Deliberately prejudiced self – driving vehicles elicit the most outrage. *Cognition*, *208*, 104555. https://doi.org/10.1016/j.cognition.2020.104555

De Freitas, J., Anthony, S. E., Censi, A., & Alvarez, G. A. (2020). Doubting Driverless Dilemmas. *Perspectives on Psychology Science*, *15* (5), 1284 – 1288. https://doi.org/10.1177/1745691620922201

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G * Power 3: A flexible statistical power analysis program for the social, behavior, and biomedical sciences. *Behavior Research Methods Instruments & Computers*, *39*, 175 – 191. https://doi.org/10.3758/BF03193146

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2019). Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. *Science and Engineering Ethics*, *25* (2), 399 – 418. https://doi.org/10.1007/s11948 – 018 – 0020 – x

Firt, E. (2024). What makes full artificial agents morally different. *AI & Society*, *40*, 175 – 184. https://doi.org/10.1007/s00146 – 024 – 01867 – 6

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5 – 15.

Garrigan, B., Adlam, A. L. R., & Langdon, P. E. (2018). Moral decision – making and moral development: Toward an integrative framework. *Developmental Review*, *49*, 80 – 100. https://doi.org/10.1016/j.dr.2018.06.001

Gogoshin, D. L. (2021). Robot responsibility and moral community. *Frontiers in Robotics and AI*, *8*, 768092. https://doi.org/10.3389/frobt.2021.768092/doi.org/10.1007/s00146 – 020 – 01020 – z

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293* (5537), 2105 – 2108. https://doi.org/10.1126/science.1062872

Ju, U., & Kim, S. (2024). Willingness to take responsibility: Self – sacrifice versus sacrificing others in takeover decisions during autonomous driving. *Heliyon*, *10* (9), e29616. https://doi.org/10.1016/j.heliyon.2024.e29616

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A., & König, P. (2019). Moral Judgements on the Actions of Self – Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02415

Kant, I. (1976). *Grundlegung zur metaphysic der sitten* (H. Shinoda, Trans.). Iwanami Shoten. (Original work published 1785).

Krügel, S., & Uhl, M. (2024). The risk ethics of autonomous vehicles: An empirical approach. *Scientific Reports*, *14* (1), 960. https://doi.org/10.1038/s41598 – 024 – 51313 – 2

Liu, P., & Liu, J. (2021). Selfish or Utilitarian Automated Vehicles? Deontological Evaluation and Public Acceptance. *International Journal of Human – Computer Interaction*, *37* (13), 1231 – 1242. https://doi.org/10.1080/10447318.2021.1876357

Malle, B. F., Scheutz, M., Cusimano, C., Voiklis, J., Komatsu, T., Thapa, S., & Aladia, S. (2025). People's judgments of humans and robots in a classic moral dilemma. *Cognition*, *254*, 105958. https://doi.org/10.1016/j.cognition.2024.105958

Maninger, T., & Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports*, *5*, 100154. https://doi.org/10.1016/j.chbr.2021.100154

Nordhoff, S., Stapel, J. C. J., He, X., Gentner, A., & Happee, R. (2023). Do driver's characteristics, system performance, perceived safety, and trust influence how drivers use partial automation? A structural equation modelling analysis. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1125031

Prematilake, G. W., & Li, H. (2024). Users' responses to humanoid social robots: A social response view. *Telematics and Informatics*, *91*, 102146. https://doi.org/10.1016/j.tele.2024.102146

Siegel, J., & Pappas, G. (2023). Morals, ethics, and the technology capabilities and limitations of automated and self – driving vehicles. *AI & Society*, *38* (1), 213 – 226. https://doi.org/10.1007/s00146 – 021 – 01277 – y

Schwarz, C., Gaspar, J., & Brown, T. (2019). The effect of reliability on drivers' trust and behavior in conditional automa-

tion. *Cognition, Technology & Work, 21*, 41 – 54. https://doi.org/10.1007/s10111 – 018 – 0522 – y

Sugarman, J. (2005). Persons and moral agency. *Theory & Psychology, 15* (6), 793 – 811. https://doi.org/10.1177/0959354305059333

Vallor, S., & Vierkant, T. (2024). Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and Machines, 34* (3), 20. https://doi.org/10.1007/s11023 – 024 – 09674 – 0

Yokoi, R., & Nakayachi, K. (2021a). Trust in Autonomous Cars: Exploring the Role of Shared Moral Values, Reasoning, and Emotion in Safety – Critical Decisions. *Human Factors, 63* (8), 1465 – 1484. https://doi.org/10.1177/0018720820933041

Yokoi, R., & Nakayachi, K. (2021b). The Effect of Value Similarity on Trust in the Automation Systems: A Case of Transportation and Medical Care. *International Journal of Human – Computer Interaction, 37* (13), 1269 – 1282. https://doi.org/10.1080/10447318.2021.1876360

Zafar, M. (2024). Normativity and AI moral agency. *AI and Ethics, 5*, 2605 – 2622. https://doi.org/10.1007/s43681 – 024 – 00566 – 8

The Impact of Moral – decision Maker and Moral Belief on Permissibility, Trust and Perceived Moral Agency in Autonomous Cars

Dong Chaowu　You Xuqun　Liu Yu　Li Ying

(School of Psychology, Shaanxi Normal University, Xi'an 070062)

Abstract: To evaluate perceptions of permissibility and the impact of the decision – maker and belief preference on trust and perceived moral agency in autonomous driving moral dilemmas, this study conducted two experiments involving pedestrian sacrifice (Experiment 1, N = 260) and driver sacrifice (Experiment 2, N = 258). The results of Experiment 1 showed that the ADS was rated lower than human drivers in terms of permissibility, trust, and moral agency. Utilitarian belief preferences were trusted and perceived as having higher moral agency than deontological ones. Experiment 2 found that people trusted human drivers with utilitarian preferences more and perceived them as having a higher level of moral agency. Compared with decision – maker, moral belief had more effect as for the judgement of moral decision – making. Furthermore, after controlling for the decision – maker, moral belief, and demographic variables, perceived moral agency mediated the relationship between permissibility and trust, explaining 31.73% of the variance. In conclusion, the research highlights a notable reluctance to endow ADS being moral decision – makers and points to the substantial impact of belief preferences on trust and perceived moral agency.

Key words: autonomous driving; moral decision; moral preference; human – machine trust; moral agency