Advances in Philosophy 哲学进展, 2025, 14(8), 255-265
Published Online August 2025 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2025.148442

Hans汉斯

超越单一算法：构建自动驾驶汽车道德决策算法的组合框架

易瑶琴

云南大学马克思主义学院，云南 昆明

收稿日期：2025年7月18日；录用日期：2025年8月8日；发布日期：2025年8月21日

摘　要

随着人工智能技术在交通领域的广泛应用，自动驾驶技术逐渐从规划走向落地，一系列的伦理挑战也随之产生。现实中存在着这样一种交通场景，即无论采取何种行动，车辆都难以避免发生碰撞。在不可避免的碰撞情境中，设定基于何种伦理构架的优化算法是我们亟待解决的问题。针对这一难题，有学者提出了基于功利主义、道义论、契约主义等伦理学理论的单一优化算法。然而，这些进路普遍存在顾此失彼的局限性。我们倡议构建一种融合多种伦理学理论的复合框架即“组合框架”。在此框架基础上，进一步引入一种能够识别并综合权衡多元化规范性伦理主张的决策程序，即“元规范决策程序”，以设计更具人性化、更具公正性的自动驾驶汽车碰撞优化算法。

关键词

自动驾驶汽车，碰撞优化算法，组合框架，元规范决策程序

Beyond a Single Algorithm: A Combinatorial Framework for Building Ethical Decision-Making Algorithms for Self-Driving Cars

Yaoqin Yi

School of Marxism, Yunnan University, Kunming Yunnan

Received: Jul. 18th, 2025; accepted: Aug. 8th, 2025; published: Aug. 21st, 2025

Abstract

With the wide application of artificial intelligence technology in the transportation sector, autonomous

文章引用: 易瑶琴. 超越单一算法: 构建自动驾驶汽车道德决策算法的组合框架[J]. 哲学进展, 2025, 14(8): 255-265.
DOI: 10.12677/acpp.2025.148442

driving technology is gradually moving from planning to implementation, and a series of ethical challenges have also arisen. In reality, there is such a traffic scenario that no matter what action is taken, it is difficult for the vehicle to avoid a collision. In the inevitable collision scenario, setting an optimisation algorithm based on what ethical concept is an urgent problem that we need to solve. In response to this problem, some scholars have proposed single optimisation algorithms based on utilitarianism, deontology, contractualism and other ethical theories. However, these solutions generally have the limitation of being unable to do everything. We propose to construct a composite framework that integrates multiple ethical theories, *i.e.*, a "composite framework", and further introduce a decision-making procedure that can identify and comprehensively weigh diverse normative ethical claims, *i.e.*, a "meta-normative decision-making procedure", to design a collision optimisation algorithm for autonomous vehicles.

Keywords

Autonomous Vehicle, Collision Optimisation Algorithm, Composite Framework, Meta-Normative Decision-Making Procedure

1. 引言

2024 年 7 月，以“萝卜快跑”为代表的自动驾驶汽车出行服务，以其显著的便捷性、高效性以及创新性，迅速吸引了大众的目光。这是自动驾驶汽车作为一种人机协同的前沿技术产物融入我们日常生活的生动写照。当前自动驾驶系统依托多模态传感器阵列、实时控制模块及深度强化学习算法，已实现车辆制动、转向等功能的自主控制。在交通技术发展进程中，安全性作为核心价值维度，理应在自动驾驶技术迭代中居于首要地位。麦肯锡的研究指出，鉴于当前 90%的交通事故源于人为失误，而自动驾驶技术通过消除人为操作环节具备显著降低此类事故的潜力([1], p. 9)。自动驾驶汽车配备的感知系统空间分辨率与动态响应速度超越了人类驾驶员的生理极限，理论上能够显著降低交通事故的发生率和人员伤亡率。然而，受制于复杂多变的交通环境、不可预测的因素以及车辆物理性能(如惯性)等局限，自动驾驶汽车并不具备绝对的安全性，始终存在不可避免的碰撞问题。当自动驾驶汽车遭遇无法避免的碰撞情境时，在事件发生瞬间面临着复杂的道德决策两难困境，尤其是意味着生命衡量的人的碰撞选择问题。自动驾驶汽车的道德算法基于哪种伦理理论才能获得最大的可接受性成为我们亟待解决的问题。目前，已有学者基于功利主义、道义论、契约主义等伦理学理论，提出了多种单一的自动驾驶道德算法。但我们认为，应该采用融合多种伦理学理论的复合式规范性架构，即“组合框架”以及“元规范决策程序”来设计自动驾驶汽车在碰撞顷刻的道德决策算法，能够有效规避单一自动驾驶道德算法的缺陷。

2. 从人工驾驶到自动驾驶：碰撞伦理的机遇与挑战

在人工驾驶时代，碰撞事故往往与人为因素紧密相关，如分心驾驶、超速行驶、疲劳驾驶等。当面临不可避免的碰撞危机时，人类驾驶员受限于瞬间的反应时间、心理紧张等因素，往往难以做出符合理性决策的道德判断。在此情境下，人类的本能往往占据上风，导致实践理性难以充分发挥作用。自动驾驶汽车的出现为碰撞伦理问题提供了新的思路。

或许有人认为，借助数据系统的辅助，可以考虑在碰撞发生瞬间由人类驾驶员单独行使决策权。实

际上，我们不能完全寄希望于人类驾驶员在碰撞顷刻采取超理性的行动，过度信赖“老司机”的应变能力。我们应考虑借助“弱”道德人工智能系统辅助人类主体在此情景下做出道德决策，才能更好地发挥人类主体与计算机辅助之间的良性交互作用[2]。首先，人类驾驶员需面对不可预测的人工智能算法黑箱与不可预知的机器学习模型的挑战。其次，由于智能系统长期负责驾驶任务，人类驾驶员难以保持对路况的持续关注。当突发紧急情况出现瞬间转交驾驶决策权，人们往往难以迅速转移注意力，从专注于一件事转向另一件事。再者，突然让人类驾驶员接手处理他未能保持持续注意力的事务所出现的意外，势必会导致其惊慌失措，这种情况下引发的结果可能比不采取任何行动更为严重。若要求人类驾驶员极力克服生理性疲劳，在车辆运行过程中始终保持注意力的高度集中，这不仅不符合实际而且还违背了自动驾驶汽车将人类驾驶员从传统驾驶模式中解放出来的初衷。因此，能够在极端情境下做出道德决策的人工智能辅助增强人类道德的机器系统，将在重构人类在碰撞交通活动中的行为正当性实践中成为一种可取的方案。

在紧急情况下，有了更为稳定可靠的数字化控制系统的帮助，车辆将做出更为符合主体价值观的道德抉择。自动驾驶汽车是一种蕴含“智慧”的新型“能动体”，依靠人工智能、计算机视觉、激光雷达以及高精度地图等技术的紧密合作，能够在传感技术和决策算法的指导下，自动完成驾驶行为[3]。与人类驾驶员不同，计算机系统并不受瞬时决策时间的紧迫性与心理紧张感的限制，在碰撞不可避免的情况下，仍能像在正常驾驶时一样高效应对。

当然，自动驾驶系统并不会自动做出符合人类伦理价值的决策，这有赖于人类的主动介入。在不可避免的碰撞情境中，自动驾驶系统使机器处于即时决策的位置，而这些决策直接关系到人类的生命安全，我们必须通过编程使汽车的功能与人类相似。只有当自动驾驶汽车以确定性的方式应对具有道德负荷的情况时，才能有效避免在实际交通或突发情况下出现不可接受的失误。那么，究竟何种伦理设定是正确的呢？是倾向于共同体集体决定的伦理设定还是倾向于个人决定的伦理设定呢？假设我们将伦理旋钮交予个人，实行个人化的伦理设定(Personal Ethics Setting，简称 PES)，这可能会导致极端利己主义的囚徒困境[4]。因为当预设在事故发生的情境下，出于自我保护的本能，人们往往首先考虑保护自身的生命安全。在自动驾驶汽车遇到危险时，将不可避免地首选对行人造成伤害，致使行人处于危险之中。而安装了此类算法的自动驾驶汽车将导致社会总体伤亡人数的增加，并且恐怕也无法获得上路许可。显然，我们需要一种恰当的，符合社会预期的，由共同体集体决定的强制的伦理设定(Mandatory Ethics Setting，简称 MES)来应对此类情况([4], pp. 97-101)。那么，究竟何种伦理理论才符合这些要求呢？

3. 道德判断自动化：碰撞算法的适用性争议

在面临不可避免的碰撞场景时，自动驾驶汽车伤害分配的伦理处置效能，本质上取决于信息表征的精确度与算法架构的规范性耦合水平。在构建自动驾驶汽车的碰撞应急算法时，由社会共同体集体决定的强制伦理原则，必须超越特定伦理审查主体的个体道德偏好，建立在对社会普遍伦理预期的系统性整合基础之上。这一规范性诉求要求算法设计必须实现从个体经验判断到社会共识形态的范式转换，唯有如此，自动驾驶系统才能有效承担起作为人类驾驶员道德风险缓冲机制的功能定位，真正成为体现社会集体意志的道德能动体。当前学界对碰撞优化算法的规范性基础研究主要围绕经典伦理学体系展开。其中，功利主义的后果论范式、道义论的规则导向框架、契约主义的普遍标准模型，因其在道德判断的可计算转化方面展现出独特的理论适配性，已成为自动驾驶汽车道德决策研究的主要范式。

3.1. 功利主义视角下的伤害最小化

19 世纪的哲学家杰里米·边沁(Jeremy Bentham)在其著作《道德与立法原理导论》中阐述了功利主义

的基本原理，即“最大幸福原则”，行为的道德意义取决于其对最大数量人群幸福的贡献[5]。基于这一理论遗产，德国慕尼黑工业大学的研究人员提出了一种用于轨迹规划的“风险成本函数”[6]。该函数巧妙地将伦理考量融入数学模型之中，旨在通过将整体风险最小化与优先处理最坏情况，实现道德决策的量化处理。因此，从功利主义伦理学的理论框架出发，可推导出自动驾驶系统的碰撞算法若能在交通情境中实现对道路参与者生命权及身体完整性的最大化保障，则该算法具备伦理正当性。具体而言，任何技术决策方案若能在有限选择集中达成社会福祉的最优解，即在边际效应层面实现利益产出最大化与负外部性最小化的动态平衡，便可被判定为符合道德规范的核心诉求，从而成为伦理意义上的优先选择。

功利主义视角下的自动驾驶碰撞优化算法，其核心在于伤害最小化的应用。这一原则要求首先对潜在受害者进行精细化的可视性量化，并依据不同的人群(如行人、乘客、驾驶员等)设定差异化的价值权重，再通过精密计算负效用率来平衡生命权重与预期死亡率[7]。这一过程不仅体现了风险管理和成本优化的精髓，也展现了将抽象伦理原则转化为具体机器操作指令的潜力。因此，将道德考量纳入自动驾驶技术的控制逻辑体系具备一定的可行性前景。

尽管功利主义原则在自动驾驶系统的伦理决策中展现出表面直观性与操作便捷性等特征，但其内在的理论缺陷与技术实现困境亟待深入辨析。在技术实现层面，基于预期效用最大化的决策模型面临显著的操作性挑战。自动驾驶系统固有的概率性特征与复杂交通场景的不可预测性，导致难以实现功利主义算法所要求的精确风险评估。具体而言，对潜在事故中多元主体(如行人、驾乘人员及其他道路参与者)的伤害概率计算，受限于传感器精度、算法运算能力及环境变量等客观因素，其风险评估的可靠性与决策结果的确定性存在根本性质疑。在伦理维度，功利主义决策框架的局限性主要体现在三个方面：其一，该模型将个体生命简化为同质化计算单位，通过量化比较实施“伤害最小化”决策，实质上构成对生命权不可通约性的根本否定；其二，系统化的效益计算模式忽略了个体权利、自主意志及合理期待等核心道德要素，导致“多数人暴政”的技术实现风险；其三，在具体事故场景中，守法行人、高安全配置车辆驾驶员及道路作业人员等无过错方往往因其物理特性或位置属性被系统判定为优先牺牲对象，这种以整体利益为名的决策逻辑，实质上构成了对个体消极权利的制度性侵犯。这种单纯追求总体伤害最小化的思维模式，抹去了为了集体主义而有所牺牲的个体的权利、自主性以及合理期望等关键性道德考量因素[8]。因为在功利主义伦理框架内，为了挽救他人的生命而牺牲无辜者的生命具有正当性。这种现象是不公平的，显然未考虑公共道德在伦理决策中的作用，是以所谓积极的名义侵犯个体正当的消极权行为，挑战了个体权利、自主性与合理期望的边界，其道德性值得商榷。

实证方法是不可避免碰撞领域中的一种前沿方法，它亦揭示了公众对于功利主义碰撞优化算法的矛盾态度。让－弗朗索瓦·博内丰(Jean-François Bonnefon)等人在其研究中指出，尽管人们赞赏功利主义汽车能够实现社会效益的最大化，普遍认可其在增进社会整体福祉方面的潜力，但出于对车内人员安全的担忧，大多数人并不愿意购买此类车辆[9]。这种责任转移现象，即期望他人为公共利益做出个人牺牲，无疑对自动驾驶技术的社会接受度构成了挑战，显然会延缓自动驾驶汽车的普及。

综上所述，功利主义视角下的碰撞优化算法，作为以结果为导向的决策框架，虽具有其独特的理论魅力与实践潜力，但在实际应用中仍需谨慎审视其公正性、合理性及伦理边界。

3.2. 道义论视角下的道德遵循

在自动驾驶伦理决策的范式演进中，道义论为规范驱动型算法架构提供了独特的理论进路。与结果主义路径不同，道义论范式强调道德义务的先天约束性，主张算法的正当性源于其对先验道德律令的遵循度，而非行为后果的效用最大化。道义论强调将抽象道德律令编码为有限自动机可解析的刚性约束体系，从而确保规范驱动型算法架构在输出层实现道德判断的规范一致性，维持道义论承诺的拓扑不变性。

托马斯·鲍沃斯(Thomas M. Power)尝试从康德的道德形而上体系出发，为机器伦理学提供康德式的解释进路[10]。艾萨克·阿西莫夫(Isaac Asimov)提出的“机器人三定律”强调，机器人必须恪守规则与义务，以保障人类的利益与安全。还有一些学者主张，在构建自动驾驶汽车的道德决策框架时，采用道义论范式需遵循若干伦理准则，其中生命至上原则具有核心地位。这一原则具体体现在两个方面：一方面，人的生命权具有可不通约的伦理优先性。当系统遭遇人、动物与财产的价值冲突时，必须建立严格的价值等级序列，确保人类生命权益始终处于最高保护位阶。另一方面，在人类生命权的内部比较中，应恪守生命等价性原则。个体特征参数，如年龄、性别、身体状况等人口统计学变量，不得成为选择性剥夺生命的合法依据。

为机器行为预设明晰的伦理约束框架具有双重理论价值：其一，通过建立公开透明的伦理决策机制，可显著提升技术应用的可解释性，确保利益相关主体的知情权和监督权；其二，采用规则导向的伦理建模路径，能从技术底层强化对用户权益的系统性保障，有效规避歧视性算法与不公平待遇的潜在风险[11]。然而必须指出，规则型伦理机器的实践效能正面临根本性挑战，这主要体现在规则实施的双重前提条件上。首先，智能系统需具备动态识别特定情境下适用伦理规则的能力；其次，必须确保规则执行的严格性与语义理解的准确性。相较于人类通过常识推理实现的规则适配与例外处理，机器系统往往表现出机械式遵循字面指令的特征。这导致伦理规则的设定必须满足两个严苛要求：既要穷尽所有可能情境的应对方案，又需实现抽象原则与具体场景的映射耦合([11], p. 103)，这对技术实现与伦理哲学均构成重大挑战([6], pp. 1033-1055)。进一步分析发现，多层级伦理规范间的优先级冲突加剧了系统复杂程度。当不同伦理义务发生竞合时(如保护乘客安全与遵守交通法规的义务冲突)，由于无法预先穷尽所有义务排序方案，也缺乏将法律条文转化为机器可执行的标准化伦理代码的普适方法。因此，在义务论框架下设计的碰撞优化算法，由于需同时满足相互冲突的规范要求，可能陷入不可调和的决策悖论状态[12]。

综上所述，道义论伦理原则有助于为自动驾驶技术的操作设置严格的道德限制条件，但仅凭此原则并不足以单独应对自动驾驶所带来的道德复杂性难题。基于道义论透明伦理规则设计的自动驾驶碰撞优化算法，受技术编程障碍与机动灵活性不足等因素的影响，其有效性尚需进一步考察。

3.3. 契约主义视角下的普遍标准

在分配正义理论与多主体系统的规范架构研究的深度耦合中，催生了基于契约主义的自动驾驶汽车伦理设定。社会契约主义通过构建多阶段重复博弈模型，将策略型理性主体的局部利益最大化行为，转化为全局规范共识的帕累托改进过程。简·果戈里(Jan Gogoll)和朱利安·穆勒(Julian Müller)基于契约主义的视角，构建了一个系统框架。在该框架中，机器系统被划分为“自私的能动者”(利己主义的最大限度保护)或“道德的能动者”(利他主义的伤害最小化) [13]。前者优先考虑车内人员的安全，后者致力于为所有可能受影响的人争取最佳结果。约翰·罗尔斯(John Rawls)在政治哲学领域提出的“无知之幕”理论，为自动驾驶汽车道德判断提供了独特的分析视角。该理论假设决策主体被剥离了关于自身社会属性与潜在风险情景的具体信息，在此认知限制下，个体会通过将每个利益相关方的处境视为可能发生在自身的境况，从而形成具有普遍性的公正道德判断。基于这一理论范式，德里克·莱本(Derxek Leben)将其延伸至人工智能伦理领域，特别是在自动驾驶汽车道德碰撞优化算法的道德决策机制中([11], p. 105)，提出应当建立以“身体完整性最大化原则”为核心的伦理框架[14]。该框架要求算法设计必须确保：当所有潜在风险承担者处于无知之幕的认知情景时，都会基于理性共识认可该决策机制在生命权益分配上的公平性。

尽管如此，上述理论框架在实践层面仍面临诸多实施挑战。契约主义在量化价值与利益相关者条件关联过程中所面临的理论困境，主要体现在三个层面([11], p. 105)：首先，在价值量化标准层面，虽然建

立基于身体完整性威胁的伤害评估基准具有理论必要性，但社会群体异质性会导致其适用性面临挑战。例如，同等程度的肢体损伤(如胫骨骨折)对不同社会群体(如体力劳动者与退休人员)造成的实质影响存在显著差异([11], p. 105)。这种差异性要求算法设计必须纳入社会情境敏感性分析，但同时也可能在社会量化标准重构过程中形成制度性歧视，特别是在风险分配机制层面可能导致新的社会正义问题。其次，在伦理评估维度层面，环境依赖的差异性需要被纳入权重计算体系。具体而言，当潜在伤害发生在道路使用者与具有环境依赖权([11], p. 106)的无辜第三方之间时，其伦理评估系数应呈现显著差异。这种区隔源于第三方在特定场域中享有免受干扰的基本权利[15]，该权利诉求应通过参数权重调整在算法架构中得到制度性体现。最后，在技术实现层面，构建多维风险概率模型面临方法论与实践层面的双重挑战。这不仅涉及对复杂情境下伤害发生概率的精确计算，更需要建立跨学科的风险评估框架以整合社会学、伦理学与工程学等多维变量。当前技术体系在实现利益相关者动态权益平衡方面仍存在显著局限，这要求投入更多研究资源以完善算法决策的可解释性和伦理适配性。

此外，果戈里和穆勒所提出的个人道德环境赋予每位驾驶员根据个人道德规范自主选择碰撞行为的权利，这一做法极易引发囚徒困境。他的强制伦理倾向于功利主义伦理理念，其“最大化原则”可能导致违反直觉的算法结果。例如，为了避免对单个利益相关方造成最大程度的死亡伤害，需以其他与事故无关的行人或驾驶者造成非致命伤害为代价。这与约翰·哈里斯(John Harris)提出的“生存抽签”(Survival Lottery)[1]观点存在显著的理论同构性。值得关注的是，现有算法框架在伤害预测方面存在显著的非对称性信息遮蔽缺陷。以如下决策矩阵为例：自动驾驶汽车转向操作将导致 1 人死亡 + 2 人肢体伤残(类型 A)，维持轨迹则引发 2 人死亡 + 1 人肢体伤残(类型 B) ([11], p. 106)。尽管两种伤害类型在量级上具有等质性，但其时空分布结构存在本质差异。在信息不完全的条件下，决策算法既无法识别肢体伤残的可逆性程度，也不能预判死亡个体的社会价值权重。这种信息遮蔽导致算法陷入纳什均衡的决策困境：当伤害类型无法通过贝叶斯决策理论框架进行优先级排序时，系统将被迫采用随机化决策策略，其本质与生存抽签的伦理机制具有同源性。

总而言之，当前基于契约主义的自动驾驶碰撞优化算法虽然在理论上能平衡多数人的福祉，但其预设的理想观察者立场在现实道路场景中面临操作性挑战，仍需我们开展进一步的严格论证与优化设计。

4. 组合框架与元规范决策程序：碰撞伦理的破解之道

通过上述分析，或许我们可以得出如下结论：在将一般性理论应用于具体应用场景时，每一种伦理学理论均显示出其显著的情境局限性。此外，由于伦理学理论体系之间存在内在的互斥性，因而选择哪一种理论为道德判断自动化辩护是一个巨大的挑战。就当前进展而言，似乎尚无单一道德标准能够全面而有效地解决自动驾驶领域所面临的道德判断困境。鉴于此，或许我们应拓宽视域，对碰撞优化算法进行横向思考。我们可以尝试考虑整合多种伦理学理论，吸收其中的合理成分，进而构建一种复合型伦理框架，即“组合框架”。在此基础上，利用“元规范决策程序”的方法来探索设计自动驾驶汽车的碰撞优化算法。

4.1. 基于“组合框架”的碰撞优化算法

当代规范伦理学理论体系的范式分歧，导致自动驾驶伦理决策系统面临根本性的建构难题。功利主义、义务论契约主义等理论范式之间的不可通约性，不仅造成价值排序的结构性冲突，更在具体应用场

[1]约翰·哈里斯提出的“生存抽签”是一个伦理学思想实验，其核心观点是：在面对器官移植需求和供体短缺的情况下，可以通过随机抽签的方式来选择一个健康的人，将其器官用于拯救多个生命垂危的病人。哈里斯认为，尽管这种做法在直观上看起来是不道德的，因为它涉及到无差别地牺牲一个无辜者的生命，但从结果上看，这种策略能够拯救更多的生命，从而在总体上提高了每个人的生存概率。

景中引发方法论适用性的持续争议。为突破单一伦理范式的局限性，学界开始探索理论整合的可能性路径。诺亚·古道尔(Noah Goodall)主张通过跨理论要素萃取([11], p. 109)，建构包含多重伦理维度的复合决策模型[16]。在自动驾驶汽车的伦理设定框架中，应当考虑采纳融合性的方法，即将不同伦理学理论整合成一个复合型规范性框架即“组合框架”，并将其作为设计碰撞优化算法的伦理基础。唯有如此，才能有效克服单一伦理理论在算法应用中的局限性。目前，已有学者开始沿着这一方向展开探讨。

凡妮沙·舍夫纳(Vanessa Schäffner)提出了一种复合型伦理分析框架，该框架通过将功利主义与结果论、道义论进行方法论融合([11], p. 109)，系统性地解决了单一型功利主义在义务冲突与权利保障方面的理论困境([8], pp. 173-187)。这种整合思路在神经伦理学领域获得创新性发展，韦利科·杜布列维奇(Veliko Dubljevic)基于 ADC (主体–行为–后果)三维模型建立了道德判断的量化评估体系([11], p. 109)。该模型通过构建主体道德资质、行为规范属性与后果效价之间的动态平衡方程，不仅实现了对传统道德直觉的形式化表征，更为自主决策系统开发出具有情境适应能力的碰撞算法提供了伦理计算基础[17]。马克西米利安·盖斯林格(Maximilian Geislinger)团队提出的风险敏感型决策框架([11], p. 109)，进一步拓展了该研究范式，创造性地整合了预期效用理论、最大最小化原则与预防原则([6], pp. 1033-1055)。该框架通过建立风险阈值动态调整机制，既保持对预期收益的帕累托优化，又通过预防性约束条件将系统性风险控制在可接受范围之内，从而在伦理规范的普适性与技术应用的异质性之间达成动态平衡。

事实上，就前文基于功利主义、道义论、契约主义的自动驾驶汽车碰撞优化算法而言，若将其置于一个组合框架内进行分析，这些算法有潜力构建一种更为完善的伦理价值体系。功利主义算法通过计算所有道路参与方在各种可能情境中的利益得失，以寻求使所有人的收益总和最大化或损失最小化的方法来进行行为决策。因此，在自动驾驶汽车碰撞决策过程中，功利主义算法往往会忽视个体的道德权利与尊严。为了解决这一问题，可以将道义论作为优化控制问题的约束条件。道义论主要集中于关注行为的动机和道德义务，而功利主义伦理学则强调行为后果的善恶评判。因此，应构建一种道德约束，使其成为该算法不可侵犯的道德底线：自动驾驶汽车不得将无关者卷入碰撞情境中，即使该选择会最大限度地降低整体死亡人数。例如，自动驾驶汽车不得为了保护车内乘客或的其他车辆而故意撞击人行道上的行人或路边的咖啡馆。那些本应安全无虞的个体，包括非交通参与者，理应受到额外的保护，禁止自动驾驶汽车在试图尽量减少受害者人数时将他们的生命视为可被牺牲的资源。当然，将个体权利或诉求纳入考量并不意味着完全放弃伤害最小化原则，而是通过额外的修正性条件，对该算法进行优化。在交通事故中，与事故无关的个体应该得到更多的保护。然而，当涉及到已经卷入事故的相关方时，伤害最小化规则仍然是适用的。通过这种方式，可以确保算法在决策时不仅考虑结果的最大化，同时也考虑行为本身的正当性。而契约主义伦理学所强调的尊重社会共识及建立利益相关者之间的契约等原则，可以引导算法在设计和使用过程中充分考虑制造商、驾驶员、行人等利益相关者的权益和需求。通过协商和沟通，形成各方都能接受的道德准则和决策标准，从而增强算法的合规性。具体而言，契约主义在自动驾驶汽车算法设计过程中，可以通过引入“无知之幕”下的决策过程，促使算法在开发阶段充分考虑各方利益相关者的权益，从而有效规避极端利己主义或功利主义可能引发的歧视性结果。在此伦理框架内，契约主义巧妙地弥补了道义论与功利主义的局限性。它不仅能确保自动驾驶汽车在面临道德决策时，充分考虑社会共识和利益相关者的需求，实现社会效益的最大化，而且能够在制定道德决策主张的过程中，兼顾个人诉求，尊重个体的道德权益。

由此观之，在组合框架中，道义论为自动驾驶汽车设定了不可逾越的伦理底线，功利主义致力于追求结果的最大化利益，而契约主义则通过社会约定的方式，精妙地平衡了两者之间的潜在冲突，构建一个了更为完善和全面的伦理价值体系。这种互补关系有助于设计出更加人性化、具备更高公正性的自动驾驶汽车碰撞优化算法。

在构建自动驾驶汽车道德决策算法的组合框架时，除上述三种具有代表性的伦理理论外，融合美德伦理学与责任伦理学也至关重要。美德伦理学作为人类长期积累的道德智慧，与人类的价值观紧密相连，它的加入使得自动驾驶汽车能在决策时展现出更符合人类道德直觉与期望的行为。美德伦理学主张，美德是一种个人特质，其中谨慎、勇气、节制和正义被视为人类的基本美德。在自动驾驶机器系统训练模型的学习过程中，可以将人类的美德作为积极的奖励信号嵌入系统，进而引导算法在决策时展现出符合人类美德特性的行为，为道义论伦理约束提供补充。责任伦理学强调行为者的责任和道德义务，将其纳入框架内有益于自动驾驶汽车的碰撞优化算法在设计、测试和使用过程中明确责任主体(如制造商、设计者、使用者等)，为发生事故时追溯责任主体并进行相应的问责和处罚提供依据，同时还能敦促各方审慎自身行为。例如，责任伦理学的引入，要求技术开发者在推动自动驾驶技术进步的同时，充分考虑其对社会和伦理的影响。这不仅有助于引导开发者在研发过程中更加注重伦理规范，提高车辆的安全水准，实现技术发展与社会伦理的同步，还能为算法设定更为严格的道德约束条件，从而增强公众对自动驾驶技术决策的信任感和社会可接受度。

在构建组合框架的过程中，考量伦理模块间的交互作用与整合机制是必要的。首先，在不同伦理原则发生冲突时，需设定优先级顺序。例如，当生命至上原则与伤害最小化原则相冲突时，应明确生命至上为优先考量。其次，为反映各模块输出结果的重要性，还需设定差异化的权重分配机制，以反映特定情境下的伦理优先级。例如，在涉及无辜者安全的情境下，适当提高道义论模块的权重，以确保决策的伦理性与公正性。最后，需针对组合框架内伦理理论潜存的症结，提出解决方案。道义论伦理学虽为自动驾驶汽车的决策设立了严格的伦理边界，但在面对复杂多变的交通情境时，其遵循规则的僵化特性限制了决策的灵活性。为了克服这一局限，可融合功利主义的优势，通过精细计算潜在受害者的利益得失来制定决策，以确保整体效果的最大化。对此，可通过引入风险成本函数等数学模型加以解决，对潜在受害者进行精细化量化，并依据普遍的伦理价值标准设定差异化的数字权重。美德伦理学的理念较为抽象，需建立量化标准以衡量其实际应用。可将美德划分为不同等级，并为每个等级设定具体的行为准则和评估指标。在自动驾驶机器系统训练模型的学习过程中，结合谨慎、节制、正义、勇气等美德理念，引导自动驾驶汽车相应的操作行为，即化抽象的伦理理念为具体可操作的机器指令。责任伦理学则需综合考虑自动驾驶汽车的技术特点、使用场景、事故原因等多个因素，构建一套科学的责任分配机制，确保事故后问责与处罚的合理性。

在自动驾驶汽车基于各伦理理论构建的组合框架投入实际应用的过程中，尚需解决若干问题以确保其有效性和可靠性。具体而言，在自动驾驶汽车的数据处理和模块训练阶段，应广泛搜集并系统整理交通事故、交通环境、人类行为等多元化的数据资源。此举旨在训练预设的自动驾驶汽车道德决策模型，使其能够全面涵盖并深入理解组合框架内所融合的伦理理论的各种道德判断案例。同时，还应通过模拟实验和实车测试等多种方式，对框架模型进行严格验证，确保道德决策结果的准确性和可靠性，并通过试验结果及时反馈模型在实际应用中的表现，并据此对原有的道德决策实践模型进行必要的调整与优化。

总之，前文预设的组合框架旨在通过整合多种伦理学理论的合理因素，尽可能全面反映所有行为主体的利益。鉴于不同行为主体所处的不同立场，这些各方面利益理应得到充分重视与平衡。正如一项技术若未能充分考虑驾驶员的合法权益，就难以在市场上取得成功一样，若它无法妥善处理所有受影响者的权利与需求，则同样难以获得社会的广泛接受。因此，将各种伦理学理论中的恰当因素，融合至一个“组合”算法中，成为应对复杂伦理问题的可行途径，一改过往单纯呈现各种不同伦理观点之间对抗的局面。然而，自动驾驶汽车的道德决策是一个既重要又复杂的问题。上述所构建的组合框架的初步蓝图，虽展现出一定理论魅力，但仍存在一些不足之处。在伦理选择中，应考虑哪些伦理理论能够为自动驾驶汽车的道德决策提供全面坚实可靠的理论基础？基于所选的伦理理论，应该如何取其所长补其所短？如

何有效地整合这些伦理原则，形成一个有机整体，以构建一个近乎完美的组合框架？诸如此类，都是我们未来需要进一步深入思考的问题。但与目前单一的伦理算法相比，组合框架显然在适用性上能展现出更广阔的前景，由此开辟出的研究路径将在未来展现其潜力。

4.2. 基于“组合框架”的程序实现：元规范决策程序

在道德决策理论的范式创新中，元规范决策程序建构了一种超越传统伦理理论框架的系统性道德权衡机制([11], p. 110)。该程序通过确立元规范层面的方法论中立性([11], p. 110)，对各类道德诉求的规范性基础保持价值无涉的评估立场[18]。其核心特征表现为：在不对特定伦理理论预设本体论承诺的前提下，系统整合多模态道德推理要素，通过程序正义实现道德主张的普遍可接受性。在此背景下，组合框架内伦理体系的介入显得尤为重要。在自动驾驶汽车的程序设计过程中，虽然元规范决策程序具备对各种伦理理论采取中立态度的能力，但当我们采用基于组合框架这一已形成为相对完善的伦理系统作为该程序的道德基础时，显然可以显著提升决策结果的合理性和道德完备性。因为单凭功利主义、道义论或契约主义等单一伦理理论进行决策算法选择，必然会面临上述提及的诸多直观可视性的道德缺陷。

在自动驾驶伦理决策研究初期，针对不可避免的碰撞场景，学界曾提出极具争议的随机化决策机制。该主张源于多元伦理范式(如功利主义、义务论、契约论)在风险分配上难以达成规范性共识的现实困境，试图通过概率模型消解伦理选择的主观性([11], p. 110)。然而，实证研究表明，这种基于随机算法的损害分配策略本质上构成了道德责任的系统性消解，其将道德主体的审慎判断置换为概率事件的数学期待，导致决策过程丧失可解释性与道德问责基础[19]。鉴于此，研究者开始探索构建跨伦理范式的元规范决策框架，通过建立道德原则的可操作化转化机制，为人类设计者提供先验的伦理参数配置方案。值得注意的是，维克拉姆·巴尔加瓦(Vikram Bhargava)和金泰云(Tae Wan Kim)发展的类比道德推理模型([11], p. 110)，创造性地采用道德价值量化映射方法，通过构建伦理要素的价值函数矩阵，实现了异质性道德主张的可计算化处理[20]。该模型的核心突破在于：将道德情境的类比相似度转化为价值权重参数，使抽象伦理原则能够通过算法架构生成确定性的最优解。

在自动驾驶伦理决策算法的范式演进中，量化元规范决策架构已展现出独特的理论优势。该架构通过构建多模态伦理主张解析模块，能够系统识别组合伦理框架内各规范性命题的约束边界，并基于帕累托前沿分析实现异质道德诉求的协同优化。这一方法创新与新兴的“伦理效价理论”(Ethical Value Theory, EVT)形成理论共振。伦理效价理论强调道德决策应建立“伦理要求衰减机制”形式([11], p. 111)，即通过量化建模不同道路参与者道德诉求的强度差异，形成具有动态协商能力的决策函数[21]。其核心主张在于：由于交通参与主体对自动驾驶系统的道德预期存在本质性分歧，算法设计必须构建程序化的伦理张力消解路径，而非执着于碰撞场景的确定性解。这一理论转向揭示出关键设计准则：碰撞优化算法的价值不在于输出完美伦理答案，而在于建立具有反思均衡特征的决策生成机制([11], p. 111)。这种程序正义导向的决策模型，有效规避了结果主义路径下常见的道德适切性缺失的问题。

诚然，当前研究路径与伦理量化分析范式(如功利主义)存在相似的学理困境：在多元量化指标与异质性规范体系的映射机制建构方面，面临关键性理论瓶颈。技术伦理决策中的风险分配权重与伦理效价对应关系，特别是在涉及弱势群体权益损害、生命安全权责界定等场景，呈现出高度非线性特征，这不仅容易形成价值排序的公共讨论困境，更可能触发算法偏见固化、群体权益剥夺等系统性伦理风险。这种双重理论挑战亟待通过建构跨学科研究范式加以突破。需要指出的是，问题的解决通常是在现有方法的基础上，通过克服其局限性而逐步取得进展，这一过程具有反复性。因此，基于本文组合框架的元规范决策程序的碰撞优化算法可视为对传统单一优化算法的一种创新。尽管这种方法依然存在不足之处，但无疑已经代表了一种重要的进步。

此外，在基于复合伦理框架的决策分析中，运用元规范决策程序时需要着重考虑社会接受度的实践价值。对于自动驾驶技术而言，公众认同度既构成技术伦理适配性的评价指标，又作为技术推广的核心制约要素。当社会主体的普遍伦理期待与系统决策机制之间出现显著偏差时，可能引发技术排斥现象，这不仅会阻滞自动驾驶技术的规模化应用，更将导致其潜在的社会效益，如交通安全优化、出行效率提升及碳排放减少等技术红利无法充分实现。为了提高公众接受度，我们可以借助实证研究，收集有关公众的经验数据，从而为规范性思维的构建奠定基础。在掌握了公众道德偏好的精准信息后，可从规范性角度评估任何所谓最佳设定的公众可接受性，并预测用于说服公众接受这些设定所需采取的具体措施[22]。总之，推动公众接受自动驾驶技术，不仅需要满足伦理标准，还应建立在对公众道德观念的深入理解之上。

同时，在探索自动驾驶汽车碰撞优化算法的过程中，我们还必须明确一点：道德判断不能完全依赖于人工智能代理来处理。无论是单一伦理准则、多重伦理准则的组合框架，还是元规范决策程序，均表明伦理价值的界定实质上是一种人类的道德体现。机器伦理学并非人类道德能动性的简单再现，而是人工代理与给定伦理价值之间的一种功能等效关系。即便自动驾驶汽车在技术上实现了道德判断，但这并不意味着它完全取代了人类的道德判断。无论我们采取何种伦理原则和程序方法，最终的决策依然属于人类自决的范畴[23]。即使我们将伦理设定委托给机器算法，这些算法的功能仍然依赖于人类的规范性授权。总之，所有的伦理决策必须以人类自决为前提，人类因素在伦理意义上的中心地位不可动摇，也不容许泛化。

5. 结语

自动驾驶汽车是智能技术创新的“集结地”和“制高点”，具备人工智能“感知–决策–执行”的高效响应机制，在重塑人类交通行为伦理正当性方面发挥了重要的辅助作用。然而，自动驾驶汽车的推广与应用并非毫无风险，其面临着复杂的碰撞伦理困境。习近平总书记曾在中央政治局第二十次集体学习时强调，要把握人工智能发展趋势和规律，加紧制定完善相关法律法规、政策制度、应用规范、伦理准则，确保人工智能安全、可靠、可控[24]。这一战略要求促使我们构建相应的道德伦理规范，以保障人工智能的健康发展。对此，我们回顾了基于功利主义、道义论、契约主义伦理学理论的传统碰撞优化算法，指出了它们在应对复杂交通伦理问题中的局限性。尽管功利主义算法旨在追求总体伤害最小化，但在这一过程中，它可能会忽视个体的权利与尊严；道义论算法则强调了道德规则的重要性，但在实践中面临技术编程和规则应用的诸多挑战；而契约主义算法试图通过社会协议达成共识，但在量化伤害及应对未知决策选择时，其适用性仍显不足。因此，我们引入了一种结合“组合框架”与“元规范决策程序”的碰撞优化算法。组合框架通过整合多种伦理学理论的合理内核，构建了一个相对完善的伦理价值体系，为算法提供了更为全面的道德基础。元规范决策程序在此基础上，旨在对各项伦理因素进行综合权衡，得出一个尽可能兼容所有考量因素的最佳决策，达到了相较于传统优化算法更为优越的决策效果。该算法不仅有助于克服单一伦理理论算法的局限性，还增强了算法决策过程的可解释性，有助于提高公众对自动驾驶汽车的信任与接受度，对于解决自动驾驶领域的碰撞伦理困境具有一定的参考价值。

基金项目

云南大学研究生创新项目“数字化时代审美异化的纠偏路径研究”(2024-YNUMY-21)。

参考文献

[1] Gao, P., Hensley, R. and Zielke, A. (2014) A Road Map to the Future for the Auto Industry. McKinsey Quarterly. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/a-road-map-to-the-future-for-the-auto-industry

[2] 任然, 杨琼. 人工智能辅助道德增强可行吗? [J]. 科学·经济·社会, 2020, 38(3): 117-124.

[3] 宋强, 李伦. 自动驾驶道德决策的机器研究范式初探[J]. 自然辩证法研究, 2024, 40(7): 84-90+98.

[4] 孙保学. 自动驾驶汽车事故的道德算法由谁来决定[J]. 伦理学研究, 2018(2): 97-101.

[5] 杰里米·边沁. 道德与立法原则导论[M]. 时殷弘, 译. 北京: 商务印书馆, 2000: 59.

[6] Geisslinger, M., Poszler, F., Betz, J., Lütge, C. and Lienkamp, M. (2021) Autonomous Driving Ethics: From Trolley Problem to Ethics of Risk. *Philosophy & Technology*, **34**, 1033-1055. https://doi.org/10.1007/s13347-021-00449-4

[7] 隋婷婷, 郭晓. 自动驾驶电车难题的伦理算法研究[J]. 自然辩证法通讯, 2020, 42(10): 85-90.

[8] Schäffner, V. (2020) Is Utilitarianism Entirely Useless for Self-Driving Car Ethics? A Critical Reflection on the Rationale for Rule Utilitarianism. In: Göcke, B.P. and Rosenthal-Von der Pütten, A., Eds., *Artificial Intelligence*: *Reflections in Philosophy*, *Theology*, *and the Social Sciences*, Mentis Verlag, 173-187.

[9] Bonnefon, J., Shariff, A. and Rahwan, I. (2016) The Social Dilemma of Autonomous Vehicles. *Science*, **352**, 1573-1576. https://doi.org/10.1126/science.aaf2654

[10] Powers, T.M. (2006) Prospects for a Kantian Machine. *IEEE Intelligent Systems*, **21**, 46-51. https://doi.org/10.1109/mis.2006.77

[11] 法比奥·福萨. 自动驾驶的伦理学思考: 人工代理与人类价值观[M]. 毛延生, 王晓将, 译. 上海: 上海交通大学出版社, 2024: 109-110.

[12] Gurney, J.K. (2016) Crashing into the Unknown: An Examination of Crash-Optimization Algorithms through the Two Lanes of Ethics and Law. *Albany Law Review*, **79**, 183-267.

[13] Jan, G. and Müller, T. (2020) Ethical Implications of Autonomous Driving: A Contractual Perspective on Collision Algorithms. *Journal of Autonomous Vehicles*, **15**, 123-145.

[14] Leben, D. (2017) A Rawlsian Algorithm for Autonomous Vehicles. *Ethics and Information Technology*, **19**, 107-115. https://doi.org/10.1007/s10676-017-9419-3

[15] Hübner, D. and White, L. (2018) Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation. *Ethical Theory and Moral Practice*, **21**, 685-698. https://doi.org/10.1007/s10677-018-9910-x

[16] Goodall, N.J. (2014) Ethical Decision Making during Automated Vehicle Crashes. *Transportation Research Record*: *Journal of the Transportation Research Board*, **2424**, 58-65. https://doi.org/10.3141/2424-07

[17] Dubljević, V. (2020) Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles. *Science and Engineering Ethics*, **26**, 2461-2472. https://doi.org/10.1007/s11948-020-00242-0

[18] Pölzler, T. (2021) The Relativistic Car: Applying Metaethics to the Debate about Self-Driving Vehicles. *Ethical Theory and Moral Practice*, **24**, 833-850. https://doi.org/10.1007/s10677-021-10190-8

[19] Grinbaum, A. (2018) Chance as a Value for Artificial Intelligence. *Journal of Responsible Innovation*, **5**, 353-360. https://doi.org/10.1080/23299460.2018.1495032

[20] Bhargava, V. and Kim, T. (2017) Autonomous Vehicles and Moral Uncertainty. In: Lin, P., Abney, K. and Jenkins, R., Eds., *Robot Ethics* 2.0: *From Autonomous Cars to Artificial Intelligence*, Oxford University Press, 5-19. https://doi.org/10.1093/oso/9780190652951.003.0001

[21] Evans, K., de Moura, N., Chauvier, S., Chatila, R. and Dogan, E. (2020) Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project. *Science and Engineering Ethics*, **26**, 3285-3312. https://doi.org/10.1007/s11948-020-00272-8

[22] Bergmann, L.T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., *et al*. (2018) Autonomous Vehicles Require Socio-Political Acceptance—An Empirical and Philosophical Perspective on the Problem of Moral Decision Making. *Frontiers in Behavioral Neuroscience*, **12**, Article ID: 31. https://doi.org/10.3389/fnbeh.2018.00031

[23] Reed, N., Leiman, T., Palade, P., Martens, M. and Kester, L. (2021) Ethics of Automated Vehicles: Breaking Traffic Rules for Road Safety. *Ethics and Information Technology*, **23**, 777-789. https://doi.org/10.1007/s10676-021-09614-x

[24] 习近平. 坚持自立自强 突出应用导向 推动人工智能健康有序发展[EB/OL].中华人民共和国中央人民政府. https://www.gov.cn/yaowen/liebiao/202504/content_7021072.htm, 2025-05-01.