Dispute Settlement 争议解决, 2022, 8(4), 783-791
Published Online December 2022 in Hans. http://www.hanspub.org/journal/ds
https://doi.org/10.12677/ds.2022.84106

Hans汉斯

法理学视域下的自动驾驶道德算法

李志慧

华东政法大学，上海

收稿日期：2022年9月2日；录用日期：2022年10月28日；发布日期：2022年11月4日

摘　要

自动驾驶解放人类的同时，也带来了道德决策困境：由谁预设道德算法？预设怎样的道德算法？个人化道德算法将走向完全的利己主义，提升社会总体死亡预期，应由政府统一设定算法。强制性道德算法中，支持乘客优先规则的论证理由并不具有说服力，乘客作为所有者和最大受益者，无正当理由不能在风险分配中当然地处于绝对优势地位；从人性角度而言，后果主义战胜了道义论。诸善难以兼得，后果主义中的功利主义整体伤害最小化原则和罗尔斯最大化最小值原则均无法毫无争议地解决道德困境问题，从现实立场出发，整体伤害最小化原则不可避免价值独断主义的弊端，最大化最小值原则恰当地回应了自动驾驶的风险问题，在道德感以及算法可行性上更具有可取性。

关键词

自动驾驶，道德算法，整体伤害最小化原则，最大化最小值原则，道义论

Autonomous Driving Ethics Algorithm from a Jurisprudential Perspective

Zhihui Li

East China University of Political Science and Law, Shanghai

Received: Sep. 2nd, 2022; accepted: Oct. 28th, 2022; published: Nov. 4th, 2022

Abstract

While autonomous vehicles (AVs) liberate humans from driving, it also brings ethical dilemmas to decision-making. Who should the moral algorithm be programmed by? What kind of moral algorithm should be programmed? Personalized moral algorithms will lead to complete egoism and increase the overall death expectation of society, so the government should set the algorithm uniformly. In the mandatory moral algorithm, the reasons for supporting the passenger priority rule are not convincing. As the owner and the biggest beneficiary of AVs, passengers cannot have an

文章引用: 李志慧. 法理学视域下的自动驾驶道德算法[J]. 争议解决, 2022, 8(4): 783-791.
DOI: 10.12677/ds.2022.84106

absolute advantage in risk allocation without valid and strong reasons. In terms of human nature, consequentialism triumphs over deontology. The utilitarian algorithm and the Rawlsian algorithm in consequentialism cannot solve the moral dilemma without controversy. But from a practical standpoint, the utilitarian algorithm is unavoidable for the drawbacks of value dogmatism, and the Rawlsian algorithm properly responds to the risk of AVs and is more desirable in terms of morality and algorithm feasibility.

Keywords

Autonomous Vehicles, Ethical Algorithms, Utilitarian Algorithm, Rawlsian Algorithm, Deontology

1. 引言

自动驾驶汽车 [1] 已来，挑战应对也应提上议程。目前，自动驾驶领域的热议问题主要包括两类：一类是自动驾驶的算法设计伦理问题，一类是自动驾驶致害的责任分配问题。关于后者，目前的主要学说观点是，在现行法框架下适用交通事故损害责任、产品质量责任加以解决，只是关于具体构成要件的适用存在争议[1]；关于前者，现有法律似乎难以解决，是一个颇具争议的棘手问题。

研究数据表明自动驾驶技术的应用将避免90%交通事故的发生[2]，但自动驾驶在避免人类疏忽的同时，基于其自身存在硬件故障和软件缺陷的可能性，只可能做到低事故，而无法实现零事故[3]。此时，道德决策困境，如电车难题、隧道难题，也将无例外地降临于自动驾驶系统。传统驾驶中，由人类驾驶员在紧急情况下凭借个人道德感和直觉本能作出操作选择，法律仅能对其事后的行为作出评价；而在自动驾驶中，人类驾驶员退居二线不承担操作义务，完全由自动驾驶系统作出选择，而选择所依据的是事先设定的道德算法。由此，我们需要回答这一道德算法应由谁来设定？如何设定？

道德问题是古老的问题，又是常新的问题，对此学界存在诸多不同的观点。首先，就道德算法由谁设定，存在个人选择伦理算法和强制设定伦理算法两种观点，前者由车主个人选择设定或由生产商设定，后者由政府制定统一的算法规范加以应用。其次，就设定怎样的道德算法，可细分为两个问题：第一是隧道问题，即乘客相较行人而言是否具有受保护的优先地位，是和否都不乏支持者；第二是电车问题，即是否可以为了拯救既定路线上的行人而转向牺牲其他行人，道义论认为不可以，后果主义认为可以，于此同时后果主义中包含源于边沁功利主义的“整体伤害最小化原则”和源于罗尔斯正义论的“最大化最小值”两种方法。上述两个问题各司其职，彼此之间并不存在必然的逻辑推演关系，故其答案并非一一对应，而是交叉组合、任意配对，如支持乘客优先的并不一定支持道义论，因此现行的主张具有多样性。

上述各种观点均有诸多值得讨论的地方。为此，本文旨在梳理目前关于自动驾驶伦理算法的研究成果，并通过对现有成果的分析，回答伦理算法由谁设定以及如何设定的问题。

2. 个人化道德算法还是强制性道德算法？

个人设定伦理算法的观点主要基于：既然势必要置身道德困境、选择牺牲对象，不如参照传统驾驶

[1] 美国汽车工程师学会的分级体系主张按照车辆操作与环境监控两大标准将自动驾驶分为 L0 到 L5 共六个级别。L5 级别的自动系统在所有条件下都能完成驾驶任务。本文所说的自动驾驶汽车指达到 L5 级别的自动驾驶汽车。

模式，将选择的权利交至车主手中。

有观点认为自动驾驶汽车的生产商应提供不同道德算法的汽车，由消费者自行选择购买。个人选择伦理算法的代表学者贾斯帕·康提萨(Guiseppe Contissa)建议在自动驾驶汽车中设置“道德旋钮”(Ethic Knob)，旋钮的两端分别对应完全的利他主义和完全的利己主义，正中是完全的中立，可以通过调节旋钮来调整自己的道德立场[4]。该观点一经提出即受到诸多批评：易陷入囚徒困境。根据博弈论和大众心理，在不知其他人选项的情况下，选择完全的利己主义是上上之策，出于对安全的追求和利益平衡的考量，“道德旋钮”将定格于完全的利己主义，不再旋转。而根据计算，完全利己主义的算法相较于其他算法将提升社会总体死亡预期[5]。即使有部分虔诚的利他主义者坚持自己的道德观念，其微乎其微的数量也不足以动摇最后的结果，且完全有可能因有同乘人员而调整立场。还有观点认为汽车生产商自主设定算法出售，基于生产商的逐利性，生产商会从消费者立场出发设定算法以获得更好的销量、赢得更多的利润，最后设定的算法将是利己主义的，同样会产生上述困境，降低公共安全性[6]。

因此，许多学者赞成转至强制性伦理算法，即由政府强制规定、统一道德算法。由此产生了新的问题，政府应强制规定何种算法？该种算法应否优先保护乘客？在道德困境中应如何抉择？有学者选择民意调查的方式，探求公众的道德直觉选择，并以此作为算法制定的依据。但存有一定的问题：首先，因民众仅凭其道德直觉快速地勾选答案，不见得对问题进行了深入、系统地思考，在问题描述方式、角度细微变化而不构成实质性问题变更的情况下，常常产生完全不同的答案；其次，小范围的调查样本会因性别、年龄等个人因素对实验的结果产生一定的影响，导致样本结果不具有代表性，故以民众的实验调查为依据进行算法设计是不合理的[6]。私以为算法设计应当建立在缜密的道德分析基础上，以民意作为算法设计发挥影响力的依据。下面将通过学界主流的道德伦理学说应用算法的介绍和对比，讨论政府应强制规定何种算法。各著名思想实验虽然和自动驾驶的碰撞情境并不完全吻合，但至少可以通过思想实验窥见背后的道德价值，并加以运用。

3. 隧道难题：乘客优先规则检视

杰森·米勒(Jason Millar)于其论文中假设了一种自动驾驶情形——隧道难题(The Tunnel Problem)，自动驾驶汽车在直行撞死隧道中的孩子和转弯牺牲车内的乘客莎拉两个选项之中选择直行。[7]该思想实验主要讨论的是自动驾驶汽车车内的乘客和车外的路人之间发生生命冲突时的抉择问题。有学者借此说明隧道难题中乘客优先规则的无争议性[8]。但 Robhub 网站针对该隧道难题发起了读者投票，直行和右转二者只能择其一时，作为乘客的你将期待汽车作出何种抉择？共有 110 多名读者参与投票，结果显示 64%的参与者选择杀死孩子，36%的参与者选择牺牲自己。投票结果显示结论并非无可争议，况且投票比例中男性占九十多名，女性仅占二十多名，性别可能对结果也具有一定的影响[9]。并且，米勒假设隧道难题的目的是为了说明智能时代道德决策自动化的现象，主张道德伦理算法应当充分尊重用户的自主性和知情同意权，而非赞成乘客优先规则[7]。

赞成乘客优先规则的学者主要基于自我保存原则、市场化论证以及行动者相关三个理由。

3.1. 自我保存论证

有学者以哈特的“自然法最低限度内容”为依据论证乘客优先规则的正当性，其主张自我保存以及由自我保存所引出的有限的利他主义，是自明之理、是道德和法律的基础，在道德困境中选择有利于自己的选项即是有限利他主义的体现，同时基于角色替代和保障消费者义务等理由主张算法设计应以乘客

[2] 一辆失控的电车即将撞向轨道上的五个人，旁观者有三个选择：① 任凭电车行驶撞死五个人；② 扳动开关把电车转向另一条轨道，但这会导致岔道上的另一个人的死亡；③ 扳动开关让列车开向自己所在的轨道，这会导致旁观者自己的死亡。

作为司机时的选择为落脚点，并将乘客与旁观者电车难题[2]进行类比，认为要求旁观者自我牺牲是超义务的、是违反自我保存和有限的利他主义这一人性特点的、是超越社会对人的要求的，要求乘客自我牺牲亦然[8]。

哈特在分析道德和法律之间的关系时，基于经验事实提出人是出于自我保存的基本意图遵守共同的规则——“大部分人在大多数时候都希望继续生存下去”，并由此引出了有关人的五个自然事实和相应的规则，其中一个自然事实是人的有限的利他主义：人的利他范围是有限的，而侵犯的倾向是时常存在的，这使得法律和道德规范进行规制既成为可能又成为必要[10]。那么哈特的主张能否验证乘客优先规则呢？首先，乘客优先规则要求在道德困境中，道德天平总是倾向于乘客胜出，无论对方的“砝码”是多少，此时是有限的利他主义还是永恒的利己主义？其次，乘客的地位并不等同于电车难题中的旁观者地位。旁观者并非风险的施加者，也并非危险事件的卷入者，其本身是该电车事件的局外人，选择继续保持置身事外的状态，不自我牺牲以拯救他人是符合道德对人的期待与要求的。但乘客是命令自动驾驶汽车上路的人，是自动驾驶汽车利益的享有者，是风险的施加者，此时乘客我们不能仅仅从乘客角度出发看待碰撞问题，更需要从碰撞对象、社会整体的视角讨论问题。

因乘客优先体现的并非有限利他，而是永远利己；乘客作为风险的施加者却在利益冲突中永远占上风，忽视了自我保存作为人类共同基本意图的共同基础性。因此仅凭自我保存论证乘客优先规则是站不脚的。

3.2. 市场化论证

市场化论证是主张乘客优先规则的学者大多采用的论证理由，即当前自动驾驶市场化符合全人类的利益需求，而乘客优先规则的建立是自动驾驶市场化的前提条件[11]。这一观点源自让·弗朗索瓦·伯尼法(Jean-François Bonnefon)等三位学者进行的六项实验调查。调查结果显示参与者赞成自动驾驶汽车采用功利主义算法，即为了实现更大的善好牺牲乘客，并希望他人购买此种功利主义算法的汽车；但他们希望自己乘坐的自动驾驶汽车会不惜一切代价保护乘客，并且在第六项实验中，参与者多表示不愿意购买功利主义算法的汽车[2]。

首先，第六项实验的问题内容是：有一辆由政府规定了算法的自动驾驶汽车，该自动驾驶汽车在乘客和家人出行遇到道德困境时可能会为了减少伤亡数量牺牲乘客，问参与者是否愿意购买该种自动驾驶汽车[2]。这与参与者在前几个是否选择牺牲乘客拯救行人问题中获得大多数赞成功利主义算法的回答结果存在差异。导致差异的可能性是多方面的，如第六项实验问题涉及和家人共同乘坐、功利主义算法由政府强制设定等等。公众的回答并非基于深刻的伦理道德思考，常常诉诸于道德直觉，受个人情感因素、偏好等影响，不具有一以贯之的逻辑性，这也与实验所显现的结果相符。因此对第六项实验问题的回答并不能必然导出功利主义算法的自动驾驶汽车会面临不能市场化的结果。其次，自动驾驶汽车普及的决定性因素是其安全性。若自动驾驶汽车能够将人从驾驶中解放出来，避免人类不当操作引发的交通事故，整体安全几率的提高必然对人形成一定的吸引力。毕竟，拒绝接受自动驾驶汽车，在传统驾驶中把握自己的命运也不见得能得到预期的“把握命运”的效果，因个人在道德困境中无法确定各项选择对自己而言的生存几率，且在事故发生的电光火石之间也来不及深思熟虑，凭直觉作出的选择不见得利于自身。面对自动驾驶相较传统驾驶的安全性优点，人们不一定会因为发生概率较小的道德困境问题放弃自动驾驶。再次，产品的市场化受多重因素的影响，道德算法乘客不具有优先受保护地位对于消费者购买欲产生的消极影响可以通过价格补贴、解放驾驶、堵车避免等因素积极影响进行抵消。传统驾驶在风险上远高于自动驾驶依然可以在今天具有普遍化的市场，相信低风险且满足人类现代需求和未来展望的自动驾驶在市场上也不会受到冷遇。

仅仅基于市场化的论证难以让乘客优先规则在道德上具有正当性的，即使不采用乘客优先规则，自动驾驶因其低风险、满足现实需求等优势，在市场化上依旧具备现实“可行性”。

3.3. 乘客特殊地位论证

有学者认为乘客与自动驾驶汽车之间存在特殊的联系，使得乘客相较行人具有特殊地位，自动驾驶汽车应当优先保护乘客。理由主要是，从消费者角度而言，自动驾驶汽车作为产品有消费者保护义务[8]；从乘客角度而言，自动驾驶汽车承担了司机的角色具有乘客保护义务。

存有疑问的是：第一，就消费者保护义务而言，经营者对消费者的安全保障义务是指经营场所范围内对消费者、潜在消费者人身、财产的保护义务，售出以后的自动驾驶汽车内部空间是否可以纳入经营场所范围？此处不存在法律解释的空间，现行法下答案是否。产品质量责任方面，产品需符合质量要求不能造成人身、财产损害，自动驾驶汽车在具有产品质量缺陷时，生产者的责任对象是产品缺陷的受害方，而不是特定的产品购买方且道德困境并不都是出于自动驾驶的质量问题。第二，自动驾驶和客运之间不具有可类比性。客运合同要求出租车司机对乘客负有安全保障义务，而自动驾驶汽车本身不具有法律主体地位，其与乘客之间也不是客运合同关系。第三，若认为乘客具有特殊地位，那么当具有多名乘客时，何者具有优先地位？按照消费者理论，车主受优先保护；按照乘客理论，其他乘客具有优先地位，二者之间出现矛盾结论[12]。

乘客特殊地位的理由在现行法框架下不具有说服力，且承认乘客特殊地位后会引发乘客内部优先顺位的问题，乘客特殊地位的理由对此得出自相矛盾的结果。

目前的自我保存论证、市场论证、乘客特殊地位论证均不能为乘客优先规则提供有力支撑。自动驾驶虽将乘客从驾驶中解放出来，乘客丧失对驾驶的控制力，但乘客作为自动驾驶汽车的所有者和最大受益者，无正当有力的理由不应在风险分配中当然地处于绝对优势地位。

4. 旁观者电车难题：道义论和后果主义检视

4.1. 康德主义道义论

康德主义道义论者认为旁观者不应该干预电车的行驶方向，而是消极旁观，放任电车在既定轨道上行驶，撞死五个人。主要论证理由是康德著名的“人是目的而不是手段”，人的生命是无价的，是不可衡量的，五个人生命并不比一个人的生命分量更重，一个人即是一个世界，因此以一个人的生命换取五个人的生命是不道德的。有学者从自我牺牲的角度加以论证，自我牺牲一种超义务的美德，法律不应予以强制规定，道德也不应予以倡导，但应对此种高尚的美德致以敬意。一个人不愿意自我牺牲在道德上可理解的，因此一个人即使出于更大的善好也不应被要求自我牺牲，无论提出要求的人是否愿意自我牺牲，因为自我牺牲是个人意愿，不能强制化。故旁观者可以不将电车转向自己，也不可以将电车转向其他人，只能放任[13]。

康德主义道义论的观点也符合现代法律紧急避险的相关规定。传统驾驶中，法律通过对人类驾驶者的事后评价判断其对道德困境中的选择是否应当承担责任，主要判断依据是紧急避险是否恰当。现代英美国家和大陆法系国家虽然对紧急避险的架构存在差异，但在面对生命权冲突时，都不约而同地采用了同一立场。紧急避险的要求是保护更大的法益，在生命权冲突的情境下，基于生命价值的宝贵性、不可衡量性，不存在更大的法益，紧急避险无适用空间，以生命的代价换取生命的做法不能得到法律的宽恕[14]。

但康德主义道义论也存在一定的缺陷，即完全秉持生命不可衡量的观念，是否会忽视减损损失的机会，将损失扩大化。从肇事者的角度来看，造成五个人死亡和造成一个人死亡对人造成心理冲击感、道德负罪感、心理谴责程度是存在差异的[15]；从人性的角度而言，减小损失的选项大概率会成为首选，如

冲向人群密集的广场和转向至三两行人的过道，人的直觉选择是转向。

4.2. 后果主义算法检视

4.2.1. 整体伤害最小化原则

整体伤害最小化原则源于边沁的功利主义思想，边沁思想的核心是“最大多数人的最大幸福”，边沁认为凡是人皆为痛苦和快乐所控制，人都是趋乐避害的，因此痛苦和快乐决定着人的行为趋向追求幸福最大化。立法和制度设计应当基于此种人性考量促成共同体的最大多数人的最大幸福、最大多数人的最小痛苦。据此很多学者根据边沁的功利主义思想将道德困境转化为了可以计算的数学题。首先，列举出各种衡量痛苦量的因素，即功利因素；再次，根据各种功利因素分别估算对不同撞击对象造成的总体痛苦量，对比后撞击痛苦量较大的对象[16]。

整体伤害最小化原则可以使得大多数人的利益最大化，符合人的一般价值观念，并且其通过对不同功利因素赋值计算得出最后的结果具有一定的清晰性、简明性以及算法可行性。但其仍然存在几个问题有待解决：第一，功利因素的非可穷尽性且易具有争议性。虽然我们现在已经身处大数据、云计算时代，高速处理大量、多样数据已经切实可行，但问题在于功利因素的列举是无穷尽的，影响人类苦痛、幸福的因素是无限的。犯罪记录、寿命期限等是否纳入功利因素加以考量具有一定的争议性。如是否考虑曾犯罪或再犯罪可能性增加计算对杀人犯的撞击概率？第二，个人信息保护问题。基于不能穷尽列举功利因素，若最终选取部分重要功利因素进行衡量，因算法的准确性与其掌握的信息量呈正相关，是否意味着人的年龄、身体健康情况、家庭成员、友情等情感关系均要上传至云端，供自动驾驶汽车进行抓取、分析。人们是否愿意以个人信息泄露曝光的风险来增加其在事故中获救几率的几个百分点？第三，错误可能性和结果不可挽回性。以计算过去信息的结果为基础预设未来的选择，过去的信息具有可变性，据此概率本身具有错误的可能性，选择因此也具有错误几率，但碰撞的结果却是不可挽回的，即功利主义结果的实现因现实因素具有偏差。

4.2.2. 最大化最小值原则

最大化最小值原则源于罗尔斯的正义论观点，在一个社会体系中人类存在一致利益——齐心协力地发展以获取比单个人发展更多的利益，与此同时人类存在利益冲突——更多利益的分配问题，为此人类在无知之幕背后(不知自己自然地位和政治地位的情况下)选择了利益分配(权利义务分配)所应遵循的原则——正义，以使自己的利益风险最小化。罗尔斯认为原初状态中人们选择的分配原则为平等分配权利义务，不平等的分配只有在因分配而最少受益的人能从分配结果中获得补偿利益时才是可接受的，即平等原则和差异原则[17]。德瑞克·雷本(Derek Leben)于其《自动驾驶汽车的罗尔斯式算法》(A Rawlsian algorithm for autonomous vehicles)一文中介绍了罗尔斯的差异原则在自动驾驶汽车面对道德困境时的算法应用。罗尔斯算法的得出是假设人们处于原初状态，尚不知晓自己属于乘客或行人的情况下，人们会同意算法的何种选择？该算法基于对每一方主体在碰撞中存活概率的估算，选择最低存活概率主体利益最大化的选项，以此实现最小值最大化；若各选项下主体存活概率相同，则采用随机的方式决定最终选项[18]。

最大化最小值原则的算法具有较强的操作性且有利于整体安全最大化。但也存在某些问题被诟病：第一，算法并未考虑年龄、健康状况等因素的影响，仅仅使用生存概率作为道德决策的考虑因素是否过于单一、片面？是否可以引入医疗分配伦理中的健康收益指标 QALYs (Quality-Adjusted Life Years, QALYs = 预期寿命损失 × 生活质量损失)来估算生存概率？[6]第二，依据罗尔斯算法，遵守法律佩戴头盔的人因头盔的保护生存概率更大，反而使得违法未佩戴头盔的人在算法抉择中受到更多的保护，是否会鼓励人们违法不佩戴头盔，以在罗尔斯算法面前获得“优势”地位，是否会助长不守法的不正之风？

惩罚遵守法律佩戴头盔的人，是否有违正义且削弱了法律的权威性？[5]第三，罗尔斯算法在某些情形下会产生违反道德直觉的决策。如按照该算法，在导致一个人死亡和众多人重伤之间，算法会选择造成众多人重伤的结果。

5. 道德算法走向何方：最大化最小值原则

前文所述的算法方案并未穷尽目前的学说观点，只是冰山一角，关于道德算法的讨论广泛而热烈，各有千秋，但观点的相互碰撞并未决出有压倒性优势的一方。缘何？目前道德算法的方案主要基于对道德伦理学说在算法可行性上的应用，即算法方案的冲突只是表象，实质是背后道德伦理价值观念的冲突。在善恶问题上，神与神也不见得可以达成一致，更遑论人与人，分歧在所难免。“普遍”的道德缺乏恒定的权威，各道德伦理学说各有其所追求的美好价值，而美好价值繁多而有时难以兼得。人的局限性、非无所不能性导致应然和实然不一致，必定要做出取舍，而取何种价值和舍何种价值则成为难题，皆为善，则取舍都是有道理的也是无道理的。对于各道德伦理，我们既不是要求化多为一成一家之言，也不是盲目认为皆有道理而陷入相对主义不作抉择，而是要基于现实的立场，结合实际可行性、结果可接受性、伦理实践经验等因素综合考量，寻求合理的答案。

针对功利主义的整体伤害最小化原则提出的疑问，关于最后一个错误可能性问题，其实是算法存在的必然问题。只是因功利主义考量的功利因素过多而导致易错性的相对升高。关于前两个问题，整体伤害最小化原则的支持者并未给予进一步的回答。私以为就个人信息保护而言，是信息化时代的固有问题，是人类迎接信息化时代必须要建立相应解决机制的问题，建立整体伤害最小化原则并非其滥觞，而是进一步要求保护机制的建立，相信通过科技的发展、较为完善的保护机制，相应信息问题也可以得到一定的解决，如谷歌的标识元技术等。功利主义的整体伤害最小化原则的真正问题在于其会不可避免地导致价值独断主义[19]。爱因斯坦的价值是否比五个普通人更高？暂不论其考虑的功利因素是否可接受问题，各功利因素的权重赋值如何决定？因此，对功利主义的整体伤害最小化原则持消极态度。

对于最大化最小值原则提出的疑问，雷本已经有所回应，第一，存活是谈论身体健康的前提条件，原初状态中人们首先纳入考量的应当是使自身存活概率最大化。将年龄、健康状况等因素纳入考量，将导致道德困境中根据人们的年龄、健康状况、性别、种族、犯罪记录等因素来进行道德决策，违反了生命不可衡量的一般社会价值观。第二，雷本承认依据罗尔斯算法确实会存在未佩戴的人比佩戴头盔的人更易受保护的极端情形。但我们需要明确自动驾驶将有益于降低交通事故的发生率，且即使发生交通事故也不一定是道德困境的情形，即道德困境在自动驾驶普及的背景下发生概率是极低的。而佩戴头盔对人的保护是长久、持续的，人不可能因为尚不知是否会经历的小概率事件而放弃对自己生命安全的持久保护，否则就是因噎废食、舍本逐末了。如系安全带在大多数情况下会保护驾驶人员和乘客的安全，但在小部分情况下也可能构成逃生的阻碍，人们不会为了小概率的可能而放弃安全带的保护功能。上述理由说明罗尔斯算法并不会鼓励人们违法不佩戴头盔。另即使在上文极端情况下，罗尔斯算法的汽车选择保护未佩戴头盔的人，也并不违反正义原则。因为罗尔斯算法的根本目的在于保护道德困境中的所有人，不能基于最后的结果反向推导出“惩罚”佩戴头盔者的结论。第三，雷本称，首先我们并不能仅依靠道德直觉进行道德决策；其次，在无知之幕的情形下考量人们的选择，由于不知自己将处于一个人的一方还是众多人的一方，人们会赞成罗尔斯算法的选择[18]。

私以为基于罗尔斯正义观点的最大化最小值原则应作为自动驾驶的道德算法。首先，就道德感而言，最大化最小值预先设置的算法旨在提高每个人的生存概率，其着眼点是每一个人的生命安全，力求最大程度守住安全底线。不同于功利主义的伤害最小化原则，旨在为大多数提供福祉，而将牺牲少数人的行为视为正当而高尚的德性[13]。最大化最小值原则不仅遵从了康德主义将每一个个人作为目的来看待，而

且一定程度上克服了康德主义一概而论易造成较大社会损害的弊端，其体现了对诸善的向往以及不得不在一定程度上放弃某一善的道德负罪感。毕竟不得不的选择不一定意味着正当性，心存敬畏、怜悯之心更有利于加深人们对善的思考。其次，以算法可行性来看，最大化最小值原则以生存概率为决定因素设计算法，虽被诟病过于单一，但也正因为单一，相较功利主义的无穷尽功利因素而言更加具有可操作性和清晰性。且自动驾驶汽车的风险本身即是对他人人身安全的风险，法律允许其一般性的风险，而对超出一般性风险的不合理风险予以规制，在设计算法时仅考虑生存概率某种程度上说来是对该风险问题最恰当的回应。再次，年龄、性别等其他因素的考量是否具有合理性本身就存在争议，如一般认为儿童比老年人更值得被救，是出于正常寿命的考量，但常态和例外谁也说不准，毕竟道德困境的发生即是例外，功利主义可能会主张纳入健康因素综合计算，但明天和意外该如何纳入考量呢？总有未考虑到的因素以及无法计入算法考虑的因素。相比而言，仅考虑生存概率既易于操作又切合风险防范的主要目的。

6. 结语

就道德算法的设定者而言，若将选择权交予乘客，必会造成囚徒困境，导致社会整体安全预期降低，因此应由政府强制设定算法。就设定何种道德算法来说，乘客优先规则的现有理由都不具有说服力，乘客不能在道德困境中独善其身，置身选择对象之外。康德主义道义论在某些时候会造成损失的扩大，不利于社会公共安全；功利主义整体伤害最小化存在诸多问题，其中最大的弊端是会不可避免地导致价值独断主义。相比之下，罗尔斯的最大化最小值原则恰当地回应了自动驾驶风险问题、具有较强的可操作性，在一定程度上遵从康德主义道义论的同时又规避了其弊端。伦理道德讨论的目标不是化多为一，也不是归于相对主义，而是在结合实际可行性、结果可接受性、伦理实践经验等因素找到现实的立场。在尚不能通过技术解决道德问题的现实状况下，我们不得不作出选择，而最大化最小值原则或许是现实立场下最站得住脚的选项。

参考文献

[1] 韩旭至. 自动驾驶事故的侵权责任构造——兼论自动驾驶的三层保险结构[J]. 上海大学学报(社会科学版), 2019, 36(2): 90-103.

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

[3] 和鸿鹏. 无人驾驶汽车的伦理困境、成因及对策分析[J]. 自然辩证法研究, 2017, 33(11): 59.

[4] Contissa, G., Lagioia, F. and Sartor, G. (2017) The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law. *Artificial Intelligence and Law*, **25**, 365-378. https://doi.org/10.1007/s10506-017-9211-z

[5] 隋婷婷, 郭喨. 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): 88-89.

[6] 王珀. 无人驾驶与算法伦理: 一种后果主义的算法设计伦理框架[J]. 自然辩证法研究, 2018, 34(10): 71, 74.

[7] Millar (2016) An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars. *Applied Artificial Intelligence*, **30**, 787-809. https://doi.org/10.1080/08839514.2016.1229919

[8] 朱振. 生命的衡量——自动驾驶汽车如何破解“电车难题”[J]. 华东政法大学学报, 2020, 23(6): 24-26.

[9] Open Roboethics Initiative (2014) If Death by Autonomous Car Is Unavoidable, Who Should Die? Reader Poll Results. https://robohub.org/if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die-results-from-our-reader-poll

[10] 朱振. 哈特/德沃金之争与法律实证主义的分裂——基于“分离命题”的考察[J]. 法制与社会发展, 2007, 13(5): 14-32.

[11] 骆意中. 法理学如何应对自动驾驶的根本性挑战?[J]. 华东政法大学学报, 2020, 23(6): 57-60.

[12] 陈景辉. 自动驾驶与乘客优先[J]. 华东政法大学学报, 2020, 23(6): 10-11.

[13] 刘清平. 电车难题新解: 两难处境下的自由意志和自主责任[J]. 浙江大学学报(人文社会科学版), 2020, 50(3): 202-204.

[14] 何鹏. 紧急避险的经典案例和法律难题[J]. 法学家, 2015(4): 124-125.

[15] 王钰. 生命权冲突的紧急状态下自动驾驶汽车的编程法律问题[J]. 浙江社会科学, 2019(9): 73.

[16] 翟小波. 痛苦最小化与自动车[J]. 华东政法大学学报, 2020, 23(6): 36-44.

[17] [美]约翰·罗尔斯. 正义论(修订版) [M]. 何怀宏, 何包钢, 廖申白, 译. 北京: 中国社会科学院出版社, 2009: 9-14.

[18] 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

[19] 李德顺, 孙美堂, 陈阳, 李世伟, 韩功华, 阴昭晖. “后真相”问题笔谈[J]. 中国政法大学学报, 2020(4): 127-128.