自动驾驶无需高阶道德算法

Autonomous Vehicles Without High Order Moral Algorithms

李大山 /LI Dashan

（上海大学马克思主义学院，上海，200444）
(School of Marxism, Shanghai University, Shanghai, 200444)

摘　要：引入伦理学中的道德两难思想实验后，自动驾驶应植入何种道德算法引起了不少讨论。区分一阶道德算法与高阶道德算法将发现这些讨论建立在一个有待澄清的前提上。植入道德算法的目的是令自动驾驶成为驾驶专家，因而识别交通法规等一阶道德算法是必要的，关乎安全、效率与舒适；但功利主义等高阶道德算法会反噬驾驶专家，既不必要也不可行，况且“自动驾驶两难”在哲学上的思想深度没有超出“电车两难”。

关键词：自动驾驶　道德两难　道德算法　高阶　驾驶专家

Abstract: After the introduction of moral dilemma thought experiments in ethics, there has been much discussion about what kind of moral algorithm should be evoked in autonomous vehicles. To distinguish between first-order and higher-order algorithms will find that these discussions are based on a premise to be clarified. The goal of embedding moral algorithms is to make autonomous vehicles “experts in driving”, and therefore, it is necessary to identify first-order moral algorithms, such as recognizing traffic regulations, which are related to safety, efficiency and comfort. However, utilitarianism and other high-order moral algorithms will come back to bite the driving expert, which is neither necessary nor feasible, and the “autonomous driving dilemma” is no deeper than the “trolley dilemma” in philosophy.

Key Words: Autonomous vehicles; Moral dilemma; Moral algorithms; High-order; Expert in driving

中图分类号：N0: B82-05　　**文献标识码：**A　　DOI: 10.15994/j.1000–0763.2023.06.002

随着自动驾驶技术的成熟，人文社科学界对自动驾驶的反思也越来越多，涉及自动驾驶的道德两难（moral dilemma）、道德算法（moral algorithms）、道德责任（moral responsibility）等话题；有“道德算法由谁设置”“采取何种道德算法”“采取特定道德算法对消费者购买偏好的影响”“道德两难困境能否通过人车切换丢给驾驶员”“交通事故的责任归属”等具体问题。然而，这些讨论都预设了一个前提：存在道德算法。本文将区分一阶（first order）与高阶（high order）两类道德算法，论证自动驾驶无需植入高阶道德算法。植入一阶道德算法使得自动驾驶成为驾驶专家，以安全、效率与舒适为评价指标；植入高阶道德算法使得自动驾驶成为道德专家，以合乎人类的理性与情感为评价指标，反噬了驾驶专家，既不必要也不可行。

一、自动驾驶中的道德两难

基金项目：国家社会科学基金重大项目“负责任的人工智能及其实践的哲学研究”（项目编号：21&ZD063）。
收稿日期：2023年2月1日
作者简介：李大山（1992–）男，浙江温州人，上海大学马克思主义学院讲师，研究方向为应用伦理学、元伦理学、认知科学哲学。Email: 582313911@qq.com

伦理学中的道德两难有很多类型，如“天桥难题”“电车难题”“游轮炸弹难题”。自动驾驶讨论较多的一类两难源于“电车难题”：轨道管理人员面对失控的电车，放任电车在原定轨道上直行将造成五人伤亡，搬动机关令电车改道将造成一人伤亡。

很容易将电车难题代入自动驾驶语境：设想一辆自动化级别最高的自动驾驶（意味着系统难以请求人类驾驶员接管）突然失控，路面前方有五人，左边一条岔路上有一人，这六个人都没有注意到来车，于是自动驾驶系统面临着伤害一人还是伤害五人的困境。

讨论较多的另一个两难版本是隧道难题（tunnel problem）：贾森·米拉（Jason Millar）设想了一种情境，自动驾驶即将进入隧道，突然隧道入口处冲出一个路人，如果紧急避让会撞上入口处的路墩，车毁人亡，如果不避让会伤害路人。[1] 当事故不可避免时，相较伤害一人还是多人的电车难题，保护乘客还是保护路人的问题在自动驾驶语境下更加应景，近十年已有不少讨论。[2]-[8]

虽然道德两难描述的极端情况不太可能发生，但自动驾驶大规模民用后必然会产生“车–车”“人–车”“物–车”交通事故，为了躲避路人而急转弯撞上其他车辆或静止物的情况是可预见的，所以自动驾驶语境下的道德两难问题是有讨论价值的。提升人类驾驶员的安全行车意识与车技能够降低人类驾驶员交通事故发生的概率，类比过来，植入算法也能起到这种效果。就目前科技发展趋势而言，实现道德算法并没有太多编程上的困难，真正的困难是概念（哲学）上的：当我们在谈论给自动驾驶植入道德算法时究竟是什么意思？

二、区分一阶道德算法与高阶道德算法

在区分一阶道德算法与高阶道德算法之前，需要先澄清两点：

第一，暂时没有看到世界各国制定允许自动驾驶植入保护乘客或保护路人的算法的法律法规。市面上L2及以下自动驾驶应对潜在交通事故，不是让系统自动做出选择而是请求人类驾驶员接管，通过人车切换将难题交给人类驾驶员。此时面临道德两难考验的是人类而不是自动驾驶系统，这涉及到事故责任归属问题，不在本文的考察范围。

第二，如果接受认知科学中的计算主义纲领，那么算法除了物理实现层还有两个维度：语义内容层与符号计算层。语义内容层采用自然语言描述道德算法，如以“帕累托最优”为目标的罗尔斯式算法（Rawlsian Algorithm）、以“完全保护乘客”为目标的利己主义算法，以“最多数人的最大福祉”为目标的功利主义算法等。符号计算层采用计算机语言与数学语言描述道德算法，如Java、Python等计算机语言。人文社会科学讨论的道德算法是语义内容层面的，不涉及计算机语言。

自动驾驶需要理解大量交通法规与伦理规则才能替代人类驾驶员，如识别红绿灯、礼让行人、礼让救护车、不得占用应急车道等。实现这些功能的算法负载特定价值取向，本文称这类算法为一阶道德算法。一个不能在短时间内理解交通信号灯的驾驶员不是一个优秀的驾驶员，没有资格被称作驾驶专家（expert in driving）；同理，一个不能在短时间内理解交通信号灯的自动驾驶系统不是优秀的替代者。

道德两难表现为两权相害，最终的选择往往彰显行动者的人生境界，重新定义了行动者。本文称这类能够重新定义行动者的算法为高阶道德算法。例如，小李原本是一个胆小而且人生经历贫乏的人，但他在电车难题中选择搬动扳机令电车改道，那么他既是拯救多数人的英雄，又是谋杀少数人的刽子手，获得了极为丰富的人生经验；如果小李挺身而出以血肉之躯挡住了电车，那么他是舍己救人的道德楷模，勇敢而气概。高阶道德算法赋予了自动驾驶新的内涵，从驾驶专家变为能够进行道德选择的驾驶专家。相反，一阶道德算法由行动者决定，驾驶专家的定位决定了自动驾驶必须能够理解交通信号灯、必须能够礼让行人。

成为驾驶专家是自动驾驶作为一种代步工具的内在要求，安全、效率与舒适三原则缺一

不可：安全原则指事故率要尽可能低；效率原则指以最短时间将用户送至目的地；舒适原则指过程平稳，内部空间宽敞、内饰高档、车载应用丰富易用等。自动驾驶首先是一款代步产品，要服从产品规律。一款外形精致美观的手机，如果联网功能不稳定，势必销量堪忧。同理，一款自动驾驶产品若要获得一定的市场份额，必须在安全、效率与舒适三方面发力。

以下通过几个实例来分析一阶算法与高阶算法的区别。丽贝卡·德文纳尔（Rebecca Davnall）提出“刹车直行优先”理论：在自动驾驶系统获得信息有限且没有其他车辆涉及的前提下，刹车直行总是风险最低的。[9] 有两个理由：

一是物理学与制动力学层面的，减速降低伤亡概率，急转弯会影响轮胎抓地，增加事故概率。[9] 所以一些学者将德文纳尔的理论称为制动力学算法。[10] 二是决策层面的，自动驾驶系统缺乏必要的信息来判断急转弯后带来的损失一定小于直行。

由于聚焦于自动驾驶的安全性能，以降低事故与伤亡概率为目的（即安全性原则），符合我们对驾驶专家的期待，所以她主张的理论具有一阶特征。不难预见，未来研发出抓地力更强、能更好控制压力的轮胎，[9] 自动驾驶会比人类驾驶员更擅长紧急停车。

不过，德文纳尔的理论仍然有不少高阶道德算法的痕迹。例如，她认为自动驾驶系统的设计者面临的问题是：是否应该增加与前方车辆发生事故的概率来降低与尾随车辆发生事故的概率。[9] 面对可能发生的追尾事故，有经验的驾驶员会考虑恰当把握车速、综合运用灯光、喇叭警示前后车，以避免与前车追尾，同时避免被后车追尾。再例如，在卡车难题中“总是刹车直行优先”策略不成立。假如自动驾驶即将与一辆巨型卡车发生正面撞击，此时采取“刹车直行”是违背我们的驾驶直觉的，急转弯大概率会翻车或伤及路人，但仍优于直行被卡车碾压。一个遇到突发状况只会采取刹车直行而无变通的驾驶员是拙劣的，不配驾驶专家之美誉。

另一个具有一阶特征的实例是芥川（Keizo Akutagawa）等人的轮胎扰动控制算法（tire disturbance control algorithm）。该算法致力于消除轮胎的外部噪音，抑制轮胎与路面产生的滑移与微震，最终达到改善汽车转向时轮胎稳定性的目的。[11] 改善轮胎稳定性有助于自动驾驶成为驾驶专家，由驾驶专家的内涵决定，是一阶性的。

简言之，自动驾驶的目的是让汽车成为驾驶专家，而驾驶专家的题中应有之义是更安全、更效率、更舒适，凡有助于实现驾驶专家内涵的算法都是一阶性的。相反，高阶道德算法改变了驾驶专家的内涵。

三、高阶道德算法既不必要也不可行

给自动驾驶植入高阶道德算法后能够避免损失挽救生命吗？不行，因为根据定义，道德两难必然导致生命财产损失，区别仅在于损失大小，所以就算植入高阶道德算法也无助于避免损失，道德两难本身是悲剧性的。

支持高阶道德算法有三个常见理由：第一个理由是高阶道德算法虽然不能避免但能降低损失。本来伤亡五个人，植入功利主义算法后只伤亡一个人，这样的算法往往被称为最优解。第二个理由是高阶道德算法使得损失以一种人类可接受的方式发生，照顾了人类情绪。第三个理由是有助于界定交通事故的责任主体。以下逐一反驳：

“保障最多数人的最大利益”是社会共识度比较高的最优解候选，目前市面上还有一种呼声较高的候选，德雷克·勒本（Derek Leben）借鉴罗尔斯的“最大化最小值”思想，设计了罗尔斯式算法（Rawlsian Algorithm）。（[12]，pp.107–115）

勒本的思路是区分车辆行为集合A，当事人集合P（包括乘客与路人），效用函数u将每个车辆行为带来的生存概率分配给当事人。例如，直行a1、左转a2、右转a3三个车辆行为带给乘客p1与路人p2的生存概率（1为生存0为死亡）可以表达为（a1, p1）=0.8；（a1, p2）

=0.1；（a2, p1）=0.6（a2, p2）=0.3；（a3, p1）=0.4；（a3, p2）=0.7，a3的最小值0.4是最大的，所以选择a3。罗尔斯主义有助于不断提升社会最弱势群体的福利，符合帕累托改进原则，是比经典功利主义、道义论、利己主义更精致也更正义的理论。

本文赞同最小值的最大化原则可以兜住社会总福利的底线，也赞同“帕累托改进”是公共决策的重要原则，但疑惑的是，为什么必须以植入高阶道德算法的方式才能实现这些原则？产业界可以研发更敏锐的感知系统，通过算法优化获得更节约资源的计算能力，通过性能更强大的芯片提升数据处理能力，从而降低自动驾驶发生事故的概率。安全、效率且舒适的自动驾驶大规模民用后会增加社会总体福利，间接体现了最小值的最大化、帕累托最优等原则，为何要直接将这些道德原则编译为算法植入自动驾驶呢？

支持自动驾驶系统植入高阶道德算法的第二个理由是给了我们确定性与安慰，不至于听天由命。没有高阶道德算法的自动驾驶遇到道德两难时的应对是随机性的而不是选择性的，这在一些人看来是一场灾难。

该理由忽略了人类驾驶员应对道德两难也是随机性的。有的人选择急转弯避开路人而车毁人亡，有的人选择自保而撞向路人，这受个人与情境因素影响；甚至同一驾驶员在不同身心状态下面对相同的情境会做出不同的选择。事实上，一些道德算法并没有带来确定性。例如，罗尔斯式算法比较车辆行为带来效用最小值的大小，如果两个车辆行为的最小值相等且没有其他可比较的效用，那么系统会随机做出选择。[12]

支持者们会说他们所谓的确定性指承担责任，植入高阶道德算法明确了交通事故的责任主体，于是来到第三个理由。尽管人类驾驶员做出什么样的选择也是随机的，但人类司机对选择负责。就哲学术语而言，高阶道德算法赋予了自动驾驶自由意志，使得植入功利主义、道义论、罗尔斯主义等算法的自动驾驶能够出于理由行动。

缺少高阶道德算法不会导致责任主体缺失。虽然今天自动驾驶民用尚处于L2阶段，但没有出现理论性的责任归属危机。纵观近六年自动驾驶交通事故报道，事故原因不外乎三类：一是自动驾驶系统本身的缺陷；二是驾驶员或车主未严格按照操作手册操作，如违规进行人车切换；三是路人或其他车辆违规。这三类事故的责任归属还是比较清楚的：第一类事故中，自动驾驶系统的某个或某几个供应商承担产品责任；第二类事故中，驾驶员或车主至少要承担部分责任；第三类事故中，路人或其他车辆承担主要责任。

世界主要发达国家在自动驾驶事故责任归属方面已经进行了非常有价值的探索，如英国2018年颁布了《自动与电动汽车法》，日本2021年颁布了《自动驾驶相关制度整备大纲》。不同国家的法律法规实践各有侧重，虽然其中一些条款的合理性值得商榷，但缺乏高阶道德算法会导致责任归属危机过度忧虑了，目前也暂未看到要求植入高阶道德算法的法律法规实践。

退一步分析，就算高阶道德算法对于自动驾驶是必要的，但如果不可行，依然有理由拒绝。本文有四个理由表明植入高阶道德算法是不可行的：一是由谁制定有争议；二是制定何种内容有争议；三是导致算法供应商负有“原罪”；四是导致责任归属过度决定。

首先，高阶道德算法由谁制定呢？有两种可能：一是不同算法供应商提供不同类型的产品，最终由市场选择。让–弗朗索瓦·伯尼法（Jean–Francois Bonnefon）与莱瑟·伯格曼（Lasse Bergmann）做了一些前期调查工作，政府出台规范，道德算法由专门的公司提供，消费者可以购买植入不同类型道德算法这种可能性。[13]二是由社会民意或公共部门决定算法参数，然后交由算法供应商代工。前者是市场性的，后者是计划性的。

市场路线的问题是，社会很难接受以市场化的方式筛选出合适的高阶道德算法，一个难以回避的问题是，这一过程要付出多少代价？算法供应商未经公意将高阶道德算法植入自动

驾驶系统是被允许的吗？供应商也不会贸然这样做。在我国语境下，第二条路径比第一条更有可行性。不过，一旦算法供应商的角色是代工者，则没理由承担产品质量之外的事故责任。已有学者对这两条路径做了全面分析并试图提出混合式路径。[14]

虽然计划路线解决了由谁制定的问题，但在具体的算法内容上会遇到众口难调的问题。在道德哲学里，“应当如何行动”“应当如何选择”这些问题本来就是开放性的，功利主义与义务论两个传统争论了几百年，演化出不少混合版本，契约论与美德伦理各有追随者，罗尔斯主义看起来调和了双方，实际上也没有取得多数席位，想从中找出一个或几个混合版本作为高阶道德算法的模板是不现实的。

从市场需求分析，自动驾驶用户与乘客会支持利己主义算法，已有大量研究表明了这一点，自动驾驶系统不优先保障乘客会影响消费者购买行为。[10]，[15]然而，植入何种算法不仅是产品需求问题，还涉及公共安全，并不是所有人都购买自动驾驶产品。对于不购买自动驾驶产品的路人来说，利己主义不符合他们的利益。对于购买自动驾驶产品的人来说，他们也有着路人身份，如果他们理性地进行分析，那么不会完全支持利己主义算法。

就国内学界而言，站在第三人称视角的罗尔斯式算法的接受度比较高，已有多篇专题讨论的文献。[16]，[17]不过，目前还没有看到专题调查罗尔斯式算法社会接受度的数据。本文不认为消费者与乘客在理性思考后会赞同，因为罗尔斯式算法剥夺了在特定情境下本属于车内乘客的生存概率。人类驾驶员在遭遇正面撞击时，往往急刹，将方向盘往左打，这是驾驶员的求生本能。虽然这将副驾驶置于撞击中心，但能提升后座乘客的生存概率，而罗尔斯式算法剥夺了这一点。

既然众口难调，那么只能由公共部门采取民主与科学的方式来决定。先让科研机构、行业代表对特定道德算法的后果进行科学预测，通过计算机模拟给出评估报告，再交由公共部门进行民主决策，根据少数服从多数（并对少数进行保护）原则。但问题是，民主的方式适用于裁决福利问题，而不适用于裁决生命财产问题。

第三个理由是无论由谁决定、无论设置何种内容，高阶道德算法都将导致“责任先定”。在具体交通事故中，不论算法权重更偏向乘客还是路人，都是事故的直接原因，算法供应商负有直接责任，这使得算法决定者承担“原罪”。公共部门和企业一样，都没有理由去承担产品责任或监管责任之外的“原罪”，因为交通事故后续存在繁琐的追责、追偿适宜，不仅有道德两难性质的交通事故，还有大量刮擦碰撞的普通交通事故。

第四个理由是责任归属“过度决定”。假如自动驾驶刹车不及时导致了交通事故，那么制动系统的供应商可能负有产品责任。这取决于事故原因调查与产品质量检查，责任清晰。然而，一旦植入高阶道德算法，那么算法供应商将负有“先定责任”。在仅因制动系统产品质量导致的交通事故中，凭借产品质量就足以解释事故了，无需植入高阶道德算法。

四、避免道德两难的技术与哲学分析

这一节将从自动驾驶的组成部分与工作原理分析，论证避免道德两难的关键不是植入高阶道德算法，关键不在于消弭道德两难带来的恐惧或恐慌，而在于避免两难出现或降低出现的概率。

自动驾驶由车辆控制、感知与决策三个系统构成。车辆控制系统包括驱动、制动、转向子系统等，感知系统包括相机、雷达、定位、传感器子系统等。这两个系统与高阶道德算法没有直接关联，有直接关联的只可能是决策系统。决策系统处理来自感知系统提供的信息，做出决策与规划，再给车辆控制系统下指令。决策系统分为三个子系统：路径规划（routing）、行为决策（behavioral decision）与运动规划（motion planning）。

路径规划子系统依据避障、省时、节能等原则形成从起点到终点的路径，不涉及车辆控

制，也不涉及如何避开行人、超车变道等。智能手机上的地图导航就是一种典型的路径规划。运动规划子系统形成运动细节，如沿着虚线向前，从50码加速至100码；以特定角度转向再以特定速度超车等等。路径规划子系统负责规划线路，而运动规划子系统负责填充运动细节，都不涉及高阶道德算法。

有可能涉及高阶道德算法的只有行为决策子系统了。该子系统负责变道、加速、避让等行为选择，不涉及具体的动作细节（如加速至多少码，以多大的角度转向，这些由运动规划子系统负责）。自动驾驶的基本车辆行为其实非常简单，只有加速、右转、左转、前进、后退、减速、停车七种。根据这七种基本车辆行为来定义复合性的车辆行为，如变道由左转＋加速＋右转定义。只要自动驾驶系统依据特定的决策算法，在不同车辆行为之间做出选择，就有道德评价的余地。

然而，行为决策子系统依据的原则是一阶性而非高阶性的，是安全、效率与舒适。安全原则要求降低“发生碰撞”与“车辆不稳定”的概率，效率原则要求加速度大于零，舒适原则要求加速度的导数趋近零。这些原则由驾驶专家的定位决定，难以反作用于驾驶专家。

以隧道难题为例，当自动驾驶系统感知到前方隧道入口有路人出现，并且感知到隧道入口处有路墩，行为决策子系预测生成复合性车辆行为，如“直行减速”或“左转减速”。减速是自动驾驶在遭遇突发情况时必定会做出的选择，由安全性原则决定。这也符合人类驾驶员的驾驶直觉，在不减速的情况下进行转向违反安全驾驶规范。2022年3月底举行的中国电动车百人会论坛上，理想汽车创始人李想提议新车应标配自动紧急制动系统（AEB），提议得到了全国政协经济委员会副主任苗圩的赞同。

不少对自动驾驶道德两难的表述侧重后果描述而对产生两难的前提——制动系统失灵只字不提。帕特里克·林（Patrick Lin）甚至没有用“失控”（run away）只用了“撞向”（run over）。（[2]，p.78）勒本的表述是自动驾驶没有足够的时间停下。[12] 这些表述不但粗糙且具有误导性。隧道难题没有预设制动系统失灵，几份文献的表述是“来不及刹车”或“来不及有效刹车”，这并不意味着“不能够刹车”。在隧道问题中，虽然自动驾驶没有足够的时间停下，但如果制动系统能够正常刹车，那么减速就是所有策略中必不可少的。即便事故仍会发生，减速也能降低伤亡概率。

有驾驶经验的驾驶员知道，高速隧道入口一公里处就会有路牌提示减速。这说明自动驾驶系统不够专业，未能理解路牌上的交通规则，属于一阶算法失灵。供应商应当设计出能够更准确、更及时理解路牌提示的算法。随着自动驾驶的普及，不难预见，未来自动驾驶在不同场合下会调用不同的算法。例如，为了优化计算资源配置，在闹市区，感知系统赋予路人更高的识别权重；而在高速公路上，降低路人的识别权重。如果事故的主因是急转弯视线被遮挡，导致系统感知不到路人，则应增强系统的感知与预测能力。这种能力既可通过行为决策子系统来实现，也可通过感知系统与行为决策子系统协同来实现。前者如预测函数，后者如行为决策子系统接受到感知系统传送来的急转弯信号就对控制系统发出减速命令。这些算法调整都是一阶性的，目的是让自动驾驶成为驾驶专家。

换言之，对自动驾驶应对突发情况的讨论应当关注“车”，如车的制动系统、感知系统，而不应当将车看作“人”。在许多思辨性的哲学讨论中这些常识都被忽略了。

一种可能的反驳是，以上只论证了自动驾驶系统构成与研发的“事实”而没有论证“应当”。这种反驳没有看到，安全、效率与舒适三条原则就是驾驶专家的“应当”。依据道义论、功利主义、罗尔斯主义等高阶道德算法无助于自动驾驶成为驾驶专家。换一个例子可能会更清楚，扫地机器人也会遇到两难：移开水桶才能清理地面，但移开水桶会导致桶里的水溢出污染地面。处理这个两难的方式是赋予扫地机器“轻拿轻放”的能力，而不是某种权衡能力。综上可得，难以从避免隧道困境的策略中看出植入罗尔斯式等高阶道德算法的必要性。

也许以上分析能打动产业界人士，但对哲学工作者还需要给出哲学方面的理由。

第一个哲学理由是哲学家期待的高阶道德算法在逻辑上难以被调用。所有算法的调用都有条件，处理道德两难的高阶道德算法同样有调用条件，用日常语言表达成“如果处于道德两难的困境，则调用罗尔斯式算法”。然而，道德两难是否定性，排除了两种选择之外所有的可能性，这在编程上难以实现。人工智能哲学中有类似的困难：框架难题（frame problem）。如何用算法表征一个动态、复杂的世界？

哲学工作者会说，并不需要在编程上定义道德两难，只需在行为决策子系统中对路人、车距系数、车辆稳定性系数赋予不同的权重。如果路人权重大，就是利他性的算法，如果车距与车辆稳定性权重大，就是利己性的。然而，高阶道德算法的支持者有义务论证这样做的必要性与可行性，他们的理由包括植入高阶道德算法能带来确定的责任归属，以上几节已论证这既不必要也不可行。

第二个哲学理由是作为一种思想实验自动驾驶的道德两难并没有超出伦理学同类思想实验的深度，因而没有哲学上的价值。菲利帕·福特（Philippa Foot）在讨论堕胎问题时以电车难题为例子来说明双效应教条（doctrine of the double effect），[18] 意外导致电车难题在伦理学中掀起了讨论热潮。这类思想实验一般有两个意义：一是反思既有的道德立场、方法与理论。例如，按照功利主义的核心观点，行动应当符合最多数人的最大福祉，拯救五个人大概率符合功利主义，但有较强的直觉与理由拒绝这样选择。二是澄清概念与立场。例如，主动伤害与被动伤害，人身伤害与非人身伤害。我们难以接受主动伤害一人来拯救他人，这说明我们对主动伤害与被动伤害的态度不同，尤其在可以预料到后果的情况下，更不应当主动伤害。

哲学工作者可能还会说，自动驾驶的道德两难与道德算法思想实验揭示了责任归属问题。然而，自动驾驶造成的交通事故如何归属责任，是自动驾驶被社会信任与接纳前必须解决的问题，现实中自然而然将遇到，不需要由抽象的哲学思辨提醒。

结　语

针对目前人文社科学界呼吁给自动驾驶植入道德算法，本文区分了高阶道德算法与一阶道德算法。以安全、效率与舒适为导向的算法是一阶性的，有助于自动驾驶成为驾驶专家；以功利主义、罗尔斯主义等理论为模板的算法是高阶性的，无助于自动驾驶成为驾驶专家。避免道德两难的关键是为自动驾驶搭载更敏锐的感知系统与更强大的计算能力，这除了受制于感知技术、算法积累与芯片制程，还受制于商业成本。

［参 考 文 献］

[1] Millar, J. 'An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars'[J]. *Applied Artificial Intelligence*, 2016, 30(8): 787–809.

[2] Lin, P. 'Why Ethics Matters for Autonomous Cars'[A], Maurer, M., Gerdes, J. C., Lenz, B. (Eds.) *Autonomous Driving: Technical, Legal and Social Aspects*[C], Berlin: Springer Publishing Company, 2016.

[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）: 58–62.

[4] 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）: 70–75.

[5] 白惠仁. 自动驾驶汽车的“道德算法”困境[J]. 科学学研究, 2019, 37（1）: 18–24.

[6] 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）: 85–90.

[7] Evans, K., Moura, N., Chauvier, S., et al. 'Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project'[J]. *Science and Engineering Ethics*, 2020, 26: 3285–3312.

[8] Geisslinger, M., Poszler, F., Betz, J., et al. 'Autonomous Driving Ethics: From Trolley Problem to Ethics of Risk'[J]. *Philosophy & Technology*, 2021, 31: 1033–1055.

[9] Davnall, R. 'Solving the Single-Vehicle Self-Driving Car Trolley Problem Risk Theory and Vehicle Dynamics'[J]. *Science and Engineering Ethics*, 2020, 26: 431–449.

[10] 隋婷婷、张学义. 功利主义在无人驾驶设计中的道德算法困境[J]. 自然辩证法研究, 2021, 37（10）: 112–

117.

[11] Akutagawa, K., Wakao, Y. 'Stabilization of Vehicle Dynamics by Tire Digital Control: Tire Disturbance Control Algorithm for an Electric Motor Drive System'[J]. *World Electric Vehicle Journal*, 2019, 10(25): 1–10.

[12] Leben, D. 'A Rawlsian Algorithm for Autonomous Vehicles'[J]. *Ethics & Information Technology*, 2017, 19: 107–115.

[13] Bonnefon, J. F., Sheriff, A., Rahwan, I. 'The Social Dilemma of Autonomous Vehicles of Autonomous Vehicles'[J]. *Science*, 2016, 352(6293): 1573–1576.

[14] 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.

[15] Bonnefon, J. F., Sheriff, A., Rahwan, I. 'Autonomous Vehicles Need Experi Mental Ethics: Are We Ready for Utilitarian Cars'[J]. *Science*, 2016, 35: 1–15.

[16] 余露. 自动驾驶汽车的罗尔斯式算法:“最大化最小值”原则能否作为“电车难题”的道德决策原则[J]. 哲学动态, 2019,(10): 100–107.

[17] 郑玉双. 自动驾驶的算法正义与法律责任体系[J]. 法制与社会发展, 2022, 28(4): 145–161.

[18] Foot, P. *Virtues and Vices and Other Essays in Moral Philosophy*[M]. Oxford: Oxford University Press, 2016, 19–33.

[责任编辑 李斌]