从《Designing Ethics in Large-scale Socio-technical Systems》看智能驾驶汽车的设计伦理问题

2023011827　艺术史论系美 319 班　张乐晗

摘要：本文以 Jeffrey Chan 的《大型社会技术系统中的设计伦理》为理论框架，结合现实生活中智能辅助驾驶汽车的设计伦理困境，以小米 SU7 事故为典型案例，深入探讨了设计伦理的核心矛盾。Chan 的文章对设计伦理的探讨具有双向革命意义：通过将责任分配作为核心设计参数等方式，推动设计学内化道德预见性，倒逼伦理学发展动态适应理论，使技术成为伦理进步的脚手架而非道德权威的僭越者，并最终推动技术应用过程中的道德生态重建。

关键词：设计伦理　智能辅助驾驶汽车　道德生态　人工道德代理

近年来，全球智能辅助驾驶汽车领域持续快速发展，技术的迭代显著加速，特别是在感知系统的多传感器融合（如激光雷达、毫米波雷达与高清摄像头的协同）、高精度地图的深度应用以及人工智能决策算法的优化方面取得了突破，相关技术在中国的城市应用场景迅速推进。越来越多品牌的不同车型开始搭载智能辅助驾驶功能，许多公司也开始逐步推行 Robotaxi 的商业化运营。

技术的狂飙突进伴随着层出不穷的安全事故。涉及开启高级别辅助驾驶功能（如自适应巡航、车道居中保持、自动变道等）的严重交通事故尤其之多，其中包括许多被普遍看好的车企，尤其是 2025 年上半年小米 SU7 的案例，引起公众对智能辅助驾驶汽车的大规模反思。除了技术上的缺陷，更深层次的伦理争议也随之升温：在小米 SU7 的案例中，首要的核心问题是责任归属的模糊性——当自动驾驶系统接管车辆控制权并发生事故时，责任究竟在驾驶员（作为系统监督者）、汽车制造商（系统设计者）、软件供应商还是传感器提供商？这给现有的法律框架带来了巨大挑战。除此之外，“电车难题”、数据安全与隐私保护、功能边界等问题依旧是公众讨论的热点，智能辅助驾驶汽车的设计伦理研究显得愈发重要。

《Designing Ethics in Large-scale Socio-technical Systems（大型社会技术系统中的设计伦理）》一文由 Jeffrey Chan 写就于 2020 年，被收录在专门研究设计伦理的书籍《Ethics in Design and Communication: Critical Perspectives（设计伦理与传播：批判性视角）》中。文章发表时，智能辅助驾驶汽车方兴未艾，但由于文中 Chan 的探讨植根于彼时设计实践的突出问题，通过两个具体案例的细节以小见大地折射出了社会技术系统的普遍性困境，对当今时代被推上风口浪尖的智能辅助驾驶汽车也有广泛的启发。

文章从三种“设计-伦理”模式的概念出发，聚焦无桩共享单车与智能驾驶汽车两个典型案例，引出不同层次多个面向的质疑与反思：伦理能否被主动设计？如何在与技术相连的

公共系统中设计伦理？何种路径可以真正解决伦理的复杂性？人类是否有资格设计高于自身的道德系统？笔者在此将重新提炼 Chan 在当时对类似问题的思考，并结合当下发展日新月异的智能辅助驾驶汽车在实际应用中的表现，对文中问题提出新的见解。

一、社会技术系统的两种设计伦理范式

文章开头，作者提出在技术社会转型期伦理问题所遭遇的的系统性挑战。在理论层面，传统伦理学长期聚焦个体行为规范，却难以解释大规模社会技术系统中设计决策引发的伦理效应。这种理论缺口亟待构建设计与伦理的新连接框架，尤其需回应设计理论家霍斯特·里特尔（Horst Rittel）关于“设计什么与不设计什么”的根本性质询。另一方面，技术创新的加速暴露出“价值无涉设计”的伦理风险——技术应用在惠及特权群体的同时，往往意外加剧社会不平等，形成“技术红利与伦理代价”的悖论。同时，设计行为本身具有双重伦理影响：一方面，技术产品通过制度化嵌入成为伦理关系的隐性调解者；另一方面，设计创新反向重塑伦理认知。这种背景揭示了核心矛盾：设计创新的自由扩张与社会责任的规范性约束亟需通过制度化的伦理设计达成协调，从而为后续探讨“主动设计伦理”的可能性奠定基础。

基于上述内容，Chan 为我们介绍了三种“设计-伦理”模式：设计可以由伦理引导，被动响应伦理要求；可以在设计中融入伦理考量，符合伦理期待；而第三种则是改变道德生态（moral ecology），由此产生一种新的人类状况（a new human condition）。前两种模式更为常见，但这篇文章所讨论的第三种更具颠覆性。在这里，他强调了伦理作为人际关系准则的作用，借“乌托邦工程”的可能性揭示出设计伦理探讨中同样需要面对的结构不良问题：庞大的系统、多重关系、多个利益相关方的博弈、边界的界定、问题自身的变化与不可彻底解决等等。[1]或许正是出于对抗解性的认识，Chan 从这里开始抛弃概念建构，转入对实际案例的解读和研究。

尽管同样属于社会技术系统发展中的设计伦理困境，事实上，将这两个案例在设计角度下审查，其本质完全不同。

无桩共享单车的核心问题是，用户享受便利却乱停放，成本被转嫁给其他人。已有的策略，一种是被动约束行为——用 GPS 锁限制停车区，一种是构建用户互信社区，激发责任感。然而，在匿名化城市中建立“道德共同体”，设计激励机制（如信用积分）而非依赖自发道德明显更为可靠；另一方面，共享单车作为资本驱动的技术平台，本质上是反社区的——若

---

[1] Rittel and Horst W. J., Webber, Melvin M. Dilemmas in a general theory of planning. *Policy Sciences,* vol.4, No. 2, 1973, pp. 155-169.

平台以数据牟利为目标，则道德生态设计很可能沦为表面工程。

这个案例的对象依旧是人类主体，且存在持续互动的参与式设计过程。而在智能驾驶汽车的案例中，工程师直接将道德规则编码给汽车——Chan 给这样的汽车取名为“人工道德代理（artificial moral agent）”，一旦软件定型，汽车在所有意外情况面前会按照编码的程序反应，其结果是瞬时性的，正如原文所说，“与人类驾驶员在试图避开一群行人时撞击单个路人的自发反射不同，任何碰撞优化算法中的道德原则都必须被有意指定并提前编码到汽车中。”。

也正是因为两个案例的差异性，共享单车的案例恰好提示了第二个问题的解决思路：共享单车平台目前唯一可行的方案是用技术约束（如罚款）替代社会责任，结果使用户沦为“规则服从者”而非道德主体——类似的，若算法全权决定汽车的道德选择，实则是将人类道德责任外包给机器，削弱社会伦理能力建设；对于共享单车而言，如果要建立一套互信机制，静态规则无法适应城市文化差异下对公共空间的理解——智能驾驶汽车的伦理算法也需通过道德数据反馈环等形式，实现地域化迭代；共享单车平台因逐利，追求数据垄断与用户增长，过度投放挤占公共空间、破坏公共信任——车企若将“乘客安全优先”设为默认伦理，可能有损于行人的普遍安全，依旧是将商业利益凌驾于社会公平，这样的后果相较于前者具有更强的破坏性与不可逆性。因此，共享单车作为柔性的道德生态设计案例，对于刚性算法伦理编程的最大启发在于：技术伦理的终极解决方案不在代码中，而在人、技术与环境的动态关系重建中。

而对于第二个案例中的重要概念“人工道德代理（artificial moral agent）”，还有很多细节值得思考：“人工道德代理（artificial moral agent）”何以被赋予了道德性？这样的道德是否适配于所有与之互动的主体？不同主体之间的冲突何以解决？如何评估这种道德来源的可靠性？

二、〝人工道德代理〞的危机

将智能辅助驾驶汽车视为“人工道德代理（artificial moral agent）”，Chan 事实上已经默认了汽车具备道德决策能力。然而，紧接着的第一个问题就是多种道德原则与自身的不适配。传统的伦理原则无外乎三类：关注行为直接后果的功利主义、崇尚普遍道德法则的义务论和源自内驱力的美德论。[2]其中，大部分学者倾向于直接锚定在社会整体利益最大化的功

[2] G. E. M. Anscombe, Modern Moral Philosophy, *Philosophy*, Vol. 33, No. 124, 1958, pp. 1-19.

利主义路径。对此，[3]Chan 举了一个这样的例子：

> 例如，考虑困境（1A）和困境（4F）。在困境（1A）中，如果 AV 被编程了功利主义伦理原则（即拯救更多生命在道德上优于拯救更少生命的原则），那么 AV 也被预先设定为瞄准并撞击那名路人，以保全五名行人的生命。相反，在困境（4F）中，如果 AV 被编程了某些功利主义伦理原则（例如，如果更多骑行者在道路上遵守佩戴安全头盔的规则，从长远来看预计会拯救更多生命），那么 AV 很可能会瞄准并撞击那个没戴头盔的骑行者——讽刺的是，即使这个行动相对于那个受到头盔更好保护的骑行者来说更可能导致死亡。这样，AV 别无选择，只能遵循伦理规则得出其令人反感的结论。

这里的困境（1A）和困境（4F）是指“电车难题”的几种情况。以功利主义为例，Chan 担忧的是，由于道德在实际情况中运用的复杂性，单一的道德原则很可能不能给出实际情况下最优的判断结果，而有时道德原则之间还会互相冲突。他还提到了另一种伦理学名为“双重效应原则（Principle of Double Effect）”，简称 PDE，这既非功利主义，也非典型的义务论或美德论，而是根植于自然法传统，尤其是托马斯·阿奎那哲学的行动理论，认为若行为满足特定条件，即使预见并接受了恶效应，该行为仍可能是道德的。[4]但在“电车难题”的道德困境中，无论工程师编码了哪一种道德准则，都意味着对一方预谋的伤害——在 Chan 看来，这样的预谋事实上也是意图的，并不是道德行为。

事实上，针对无人驾驶技术中“电车难题”的现实化困境，早在智能驾驶并未流行的 2018 年，王珀就曾在论文《无人驾驶与算法伦理：一种后果主义的算法设计伦理框架》中设想了一套新的解决方案：在算法编制上，以后果主义为核心框架，通过量化评估整体安全效益最大化，并整合多维度伦理因子以修正纯功利主义的缺陷。[5]在实施路径上，政府扮演引导者而非强制者角色，通过税收减免、保险优惠等柔性政策工具激励用户主动选择伦理算法，同时禁止“绝对利己型”算法设定以防止囚徒困境。[5]这是一套改良功利主义的伦理决策机制，但是，正如文中 Chan 也提到的，大部分人只是在抽象层面支持功利主义，但在具体情境中并不愿意真正成为牺牲对象，形成“囚徒困境”。

正是这样的道德不一致引发了 Chan 的注意。他进一步探究：如果与人类自身可疑的道

---

[3] Borenstein J, Herkert J and Miller K, Self- driving cars: Ethical responsibilities of design engineers, *IEEE Technology and Society Magazine,* vol. 36, No. 2, 2017, pp. 67-75.

[4] Joseph M. Boyle and Jr. Toward, Understanding the Principle of Double Effect, Ethics, Jul., Vol. 90, No. 4, 1980, pp. 527-538.
Published by: The University of Chicago Press

[5] 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》，《自然辩证法研究》第 34 卷第 10 期，2018 年，第 70-75 页。

德判断相比，智能辅助驾驶汽车这样的“人工道德主体”在压力下可能比人类主体更能始终如一地遵循伦理规则。但是，道德性究竟是如何体现的？在极端情况下违背传统伦理规则以实现其他道德目标——这正当地可以被视为一种道德不一致——是否也是真正道德自主的证据？而关键是，这样不连续的判断恰好只能由人类主体作出。

Chan 由此产生了对全文论证根基的质疑：将智能驾驶视为道德自主的代理——这个想法是否本身就存在问题呢？[6]正如 Chan 在文章最后突然提出的反思：“在人类文明的某个时间点，一个道德上不完美的人类设计者，依靠本质上不完整的伦理知识体系，如何能被期望创造出道德上更优越的人工道德代理？”如今，汽车作为道德代理所带来的责任分配问题依旧存在：如果车“自主”选择了撞人，该起诉谁？是车这个“道德主体”，还是背后的设计者？[7]或许正是出于对汽车道德“自主性”的怀疑，当今智能辅助驾驶都会在识别到可能的危险情况后立马要求司机接管，由人类来作出最终的道德判断——尽管这样的设定在小米 SU7 事故后广受诟病，被称为“甩锅”行为，而这起事故最终的责任认定结果则依然是未知数。

文章结尾是这样说的：“人类道德主体对道德不确定性和不一致性的倾向，毕竟可能暗示的并非基因缺陷，而是道德进步的真正驱动力。”一方面，Chan 试图为文中所有伦理探讨辩护，证明其驱动道德进步的作用；而另一方面，他或许已经隐隐约约意识到，智能驾驶汽车并不具备“道德不确定性和不一致性的倾向”，若是将其作为道德主体，或许并不能带来实现社会普遍认可的伦理结果。

三、走向负责任的技术谦逊

从技术场景设想出发、最终停留在哲学层面的思想实验到此为止。书籍出版五年后，智能辅助驾驶的广泛应用和实际案例中呈现出的复杂性一次次说明，智能辅助驾驶汽车并不应被视为道德主体，而是绝对服从人类法律与安全规范的工具。我们也并不需要完备的道德代理（Full Ethical Agents），机器始终是人类的从属，完全听从人类的命令，服从人类的法律，并且随时准备牺牲（“be expendable as needed”）。[8]

王菁菁指出，过去到今天，许多伦理层面的争论往往混淆了“技术道德决策能力”与“赋

---

[6] Allen C, Varner G and Zinser J, Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence,* Vol. 12, No.3, 2000, pp. 251-261.

[7] 李宜轩：《责任伦理视域下自动驾驶技术伦理问题研究》，硕士学位论文，渤海大学马克思主义学院，2020 年。

[8] Yampolskiy and Roman, Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach, *Philosophy and Theory of Artificial Intelligence*, Heidelberg:Springer Berlin Heidelberg, 2013, pp. 389-396.

予道德地位”两个关键概念：前者指汽车的技术能力能做出类似道德判断的选择，比如编程让它计算伤亡人数最小化；后者指汽车因此就具有了“道德主体”地位，认为车自己“理解”了道德并该为选择负责。[9]正如 Chan 在这篇文章中提到，人类道德具有不一致性，即使已经尽可能地编码出复杂的算法，智能驾驶汽车依旧不可能像人类一样在头脑的“黑箱”中分析出编码之外的答案。除却伦理层面的内在要求，这也是由人工智能技术现阶段的固有局限所决定的：目前的人工智能技术还无法设计出成熟的伦理系统，让 AI 能完全独立地进行道德判断。在这种理想化的决策模式下，AI 并非由人类在特定场景下操作，而是依据预设的伦理规则自行作出决定。[10] [11]

智能驾驶系统的伦理选择本质是设计者意志的代码化延伸。例如，避障算法若预设“乘客安全权重>行人”，反映的并非汽车自身的道德偏好，而是制造商对功利主义或产品责任的权衡结果。因此，责任归属的逻辑链条需穿透技术外壳，追溯至算法设计者、监管机构及用户操作规范等人类责任节点：必须公开智能驾驶系统的伦理算法逻辑，确保用户清晰理解车辆在事故中的决策依据。[12]而另一方面，基于知情同意这一科技伦理基石，用户应有权知晓算法默认的伦理设置（如利他模式），并自主选择其他模式（如优先保护车内人员）；具体每辆车选择的伦理模式不对外标识。[13]这样的自由选项或许将引发社会诸如“损害公共利益最大化”一类的批判，但每辆车从本质上也都是驾驶者意志的代理，这样的决策与车企、算法设计者无关，而是每一位社会公民的自觉选择。

四、结语：技术伦理的互文性重构

作为《Ethics in Design and Communication: Critical Perspectives（设计伦理与传播：批判性视角）》第一部分“设计在社会中”的开篇之作，Chan 的《Designing Ethics in Large-scale Socio-technical Systems（大型社会技术系统中的伦理设计）》以极具张力的命题——“伦理能否被设计？”——叩响了全书的伦理追问。对于后文来说，Chan 构建的三种“设计-伦理”模式，恰好为解析智能驾驶汽车的伦理困境提供了核心坐标系：“伦理引导设计”对应着，当前智能驾驶的“安全冗余设计”仅满足基础合规性，无法应对道德决策场景；小米 SU7

---

[9] 王菁菁：《自动驾驶汽车的伦理困境及出路——从“电车难题”谈起》，《南开法律评论》，2020 年 00 期，第 81-97 页。

[10] Siponen M, A pragmatic evaluation of the theory of information ethics, Ethics and Information Technology, Vol. 6, No. 4, 2004, pp. 279 - 290.

[11] Wallach Wendell and Colin Allen, Moral Machines: Teaching Robots Right from Wrong, New York: Oxford University Press, 2019, p. 16.

[12] Lin P, Why ethics matters for autonomous cars, *Autonomous driving: Technical, legal and social aspects*, 2016, 69-85.

[13] 杜严勇：《论无人驾驶汽车的伦理设计原则》，《中州学刊》，2023 年第 9 期，第 108-113 页。

事故中，系统强制要求人类在短时间内接管危险场景，实则是将道德责任“甩锅”给驾驶员——这正是 Chan 批判的伦理外包陷阱，是“设计中嵌入伦理”的反面案例；而重构道德生态则意味着，技术伦理的终极目标应是重塑“人-技术-环境”的动态关系，因此，智能驾驶的伦理设计需超越代码逻辑，构建涵盖法律、保险、公众参与的责任网络。

此文奠定了全书批判基调，与后续章节形成深刻对话。主编 Laura Scherling 在导论中警示的“技术模糊真实边界”的危机，在 Chan 的案例中具象化为自动驾驶算法决定谁该牺牲的伦理坍塌，以及单车系统将公共空间转化为资本试验场的治理失效。这种批判脉络在 Rachel Berger 对数据可视化“虚假确定性”（如选举预测针）的剖析中得到延续：二者共同揭露简化设计对伦理复杂性的消解；而 Marc Miquel-Ribé对“黑暗设计”权力操控的论述，则与 Chan 的单车系统诱发“公地悲剧”形成互文——技术便利背后暗藏行为诱导的伦理代价。更具建设性的是，Michael Madaio 与 Sarah Martin 提出的“公民参与智能治理”主张，正呼应了 Chan 对“道德生态”的重构路径（如通过社区共建培育单车文化），为伦理设计提供了民主化出口。

Chan 对于智能驾驶汽车的探讨不仅关注设计如何被伦理约束或体现伦理价值，更关键的是主动探索设计在塑造未来伦理关系、规范和社会形态中的潜力与巨大责任，以及伴随而来的深刻伦理挑战。全书后续对 VR 安全空间、交友软件物化界面、智慧城市权力的探讨，皆可视为对 Chan 之命题的纵深回应——唯有关将伦理置于设计进程的核心，技术重塑社会的“进步”叙事才不至沦为一场精致的沦丧。

放眼于整个设计伦理探讨，这篇文章正推动设计伦理从传统的“规范应用”向“主动建构”范式跃迁。全文的核心追问“道德不完美的人类能否可靠设计出更优越的伦理系统？”使文章定位于设计伦理理论的临界点：既突破“伦理作为设计约束”的传统框架而开辟“伦理作为设计产物”的新疆域，又深刻揭示人类道德主体性在技术介入时代面临的认知边界，最终将设计伦理探讨推向人机共构伦理体系的可能性与合法性思辨。

而整体来看，在当代新兴技术发展过程中，类似这篇文章的道德考量正从设计实践的边缘辅助角色提升为基础性构成要素。以文章所述无桩共享单车系统为例，传统设计聚焦技术便利性却忽视社会信任培育，导致道德风险频发，这迫使设计学必须将道德关系，如责任分配、互惠机制等视为系统设计的核心参数而非外部约束。类似地，自动驾驶汽车的伦理算法困境证明，设计决策已从被动遵循伦理规范转向主动定义道德标准（如生命价值的程序化排序），这彻底重构了设计方法论——从用户需求满足升级为道德责任预置。在伦理学视域下，设计伦理将抽象原则具象化为技术场景中的可执行逻辑，使伦理学获得实践验证场域，如本

文案例中道德算法引发的社会接受度研究；另一方面，文章揭示的“预谋性伤害”等新困境，暴露了传统伦理框架对技术介入道德决策的认知滞后，尤其凸显人类道德主体性与人工代理者之间的张力。这种双向互动使设计伦理成为设计学与伦理学的共同前沿：它不仅要求设计学内化道德预见性，更推动伦理学发展动态适应性理论以应对技术再造道德秩序的现实。

智能辅助驾驶的伦理困境终归是人性困境的镜像。五年后再读 Chan 的论述，其价值不仅在于预见了智能驾驶的伦理困境，更在于揭示了技术伦理的“设计悖论”：人类既是道德缺陷的携带者，又是道德系统的创造者。而他的的终极追问——不完美的人类何以设计道德更优越的代理？——最终也获得了回应：技术的道德性不在于超越人类，而在于忠实地服务于人类社会的动态伦理共识。正如共享单车的乱停放问题最终需通过社区共治而非 GPS 锁解决，当我们停止将算法神化为道德先知，转而构建透明、可迭代、责任明晰的设计生态时，技术才能真正成为伦理进步的脚手架。