# 人工智能伦理治理与人工智能法治的对齐

## ——以醉驾难题为例

**The Alignment of AI Ethics Governance and AI Rule of Law:**

**A Case Study of the Drunk Driving Problem**

王少 /WANG Shao

（同济大学马克思主义学院，上海，200092）
(School of Marxism, Tongji University, Shanghai, 200092)

**摘　要：**对齐人工智能伦理治理与人工智能法治，是应对人工智能深度介入社会生活的必然要求。相较于电车难题，醉驾难题在人工智能治理语境下具有更深刻的伦理价值与法治意义。人工智能价值对齐难以在伦理空间自洽完成，以及人工智能治理过程中面临的法律执行与适用上的现实障碍，共同构成了人工智能伦理治理和人工智能法治必须对齐的逻辑起点。为了完成二者的对齐，应当以“极简主义”为核心理念，构建阶段式路径：在设计阶段嵌入容错机制以保障技术发展，在监督阶段融合伦理与法律以防控应用风险，在追责阶段遵循比例原则以平衡责任承担与伦理价值。

**关键词：**人工智能伦理治理　人工智能法治　醉驾难题　容错机制　比例原则

**Abstract:** Aligning artificial intelligence (AI) ethical governance with the AI rule of law is an inevitable requirement for addressing the deep integration of artificial intelligence into social life. Compared to the Trolley Problem, the “Drunk Driving Problem” has deeper ethical and legal significance in the context of AI governance. The fact that AI value alignment is difficult to achieve through self-consistency solely within the ethical sphere, and the practical obstacles regarding legal enforcement and application in AI governance, jointly constitute the logical starting point for the necessity of the alignment between AI ethical governance and the AI rule of law. To accomplish this alignment, one should adopt “minimalism” as the core principle and construct a phased pathway as follows: embedding fault-tolerance mechanisms in the design phase to ensure technological development; integrating ethics and law in the supervision phase to control application risks; and adhering to the principle of proportionality in the accountability phase to balance liability and ethical values.

**Key Words:** AI ethics governance; AI rule of law; Drunk driving problem; Fault tolerance mechanism; Principle of proportionality

**中图分类号**：C91; TP18　DOI: 10.15994/j.1000-0763.2026.08.011　CSTR: 32281.14.jdn.2026.08.011

人工智能伦理治理和人工智能法治的出发点相似，但侧重点不同。在实践中，人工智能伦理治理主要通过各种伦理指南和政策文件实现，而人工智能法治更像一种前景化的安排。在依法治国和以德治国相结合的背景下，介入社会生活的人工智能不能仅依靠一种规范进行治理，法治与伦理治理如何避开冲突走向融合，不仅是德法共治在实践中的新发展，也是法和伦理在人工智能时代相互对齐的新需求。智能社会不是人工智能的社会，而仍然是人的社会，

**收稿日期：**2025年11月15日
**作者简介：**王　少（1985-）男，安徽肥东人，同济大学马克思主义学院教授，研究方向为科学技术与社会、思想政治教育。Email: wslaw@mail.ustc.edu.cn

人如何在智能社会维护秩序是以人为本的必然内涵，如何让人工智能的行动符合人类的社会规范则是智能向善的基本取向。见微知著，一种新颖的人工智能“醉驾难题”暴露出人工智能法治和人工智能伦理治理之间的裂隙，既证实了法和伦理各自无法单独完成人工智能治理，也将论者的思路从提高人工智能道德水平和法治意识的固有空间拓展到人如何在遍布人工智能的人类社会中遵循法律与道德，以及如何让固有的社会规范不受干扰地面向人工智能发挥效力。人工智能伦理治理和人工智能法治难以对齐的根本原因是道德与法本身的差异，在实践中主要表现为二者在人工智能治理中适用程序和追究责任的不一致。本研究以醉驾难题为例，试图破除人工智能伦理治理和人工智能法治之间的坚固壁垒，寻求二者兼容共进的可行路径。

## 一、醉驾难题的提出

过往关于自动驾驶伦理或法律问题的探讨绕不开电车难题，人们在电车难题中追问自动驾驶汽车是否具有人的生存本能，能够“在平衡各方事故中幸存的概率以及作出碰撞选择时更加理性和客观”。[1]事实上，当前的人工智能不具备人的生存本能，也缺乏身体依赖性，电车难题在自动驾驶领域的意义被无限压缩到学术维度。在未来的道路交通实践中，“人类被更为可靠的自动驾驶所取代，它们既不喝酒，也不会承受压力，更不会忽视交通规则，交通事故会因此减少百分之九十”。[2]但在自动驾驶没有变成真正的无人驾驶之前，人的醉驾行为和汽车的自动驾驶行为相互嵌合，引发了真正值得关注的醉驾难题。

### 1. 一个关于醉驾的思想实验

醉驾难题并没有被学界正式提出，但在一些地方的酒驾查处中已逐渐显现。本文综合实践案例和自动驾驶的特点提出的醉驾难题，事实上是一个关于醉驾的思想实验。

A和朋友B等人聚餐，席间A大量饮酒。散席后，朋友B将A送上A的自动驾驶汽车，A在简单操作后，便在驾驶位上沉沉睡去，汽车自动驶往A家。途中遇交警查酒驾，经检测，A体内酒精含量达到醉驾标准，但对于如何处罚A产生争议。

该难题可以添加条件衍生出多种变体。

变体一：A不在驾驶位，而在后座睡觉；

变体二：车辆行驶前，操作自动驾驶系统的不是A，而是B；

变体三：作为自动驾驶系统操作者的B并未饮酒；

变体四：载着A的自动驾驶汽车在行驶时发生车祸。

按照我国2022年实施的国标《汽车驾驶自动化分级（GB/T40429-2021）》，自动驾驶分为六个等级，L0-L3级可视为弱自动驾驶，L4-L5级则属于强自动驾驶。由于弱自动驾驶必然有驾驶员全程参与，因此醉驾难题所针对的只能是强自动驾驶，也就是高度自动驾驶L4和完全自动驾驶L5。醉驾难题中的案情似乎简单，但无论是人工智能法治，还是人工智能伦理治理都遭遇极大挑战。

从人工智能法治角度看，如果要进行公正的处罚，必须厘清四个问题。

问题一：不坐在驾驶位的是否就不是驾驶员？如果答案是肯定的，那么当A在后座睡觉时就不是驾驶员。

问题二：操作自动驾驶系统是否等同于驾驶汽车？如果答案是肯定的，那么当B操作系统时，A就不是驾驶员。

问题三：如果驾驶员指既操作系统，又坐在驾驶位的人，那么当B操作系统时，A和B都不是驾驶员。

问题四：醉驾者是否必须是饮酒者？如果答案是肯定的，那么当B未饮酒时，即使B操作了自动驾驶系统也不是醉驾者。

酒驾和醉驾分别规定于《中华人民共和国道路交通安全法》和《中华人民共和国刑法》中，两部法律将酒驾和醉驾的前提界定为“饮酒后驾驶机动车的”和“醉酒驾驶机动车的”。上述四个问题与法律规定之间存在显而易见的争议空间。

从人工智能伦理治理角度看，如果要在醉驾难题中进行伦理追责，必须厘清三个问题。

问题一：自动驾驶汽车不知道A醉酒，它需要为其接受操作的行为承担责任吗？如果不需要承担责任，自动驾驶汽车实际上仍然是普通汽车，它无需分辨驾驶者的状态，只需接受指令即可，尽管它对指令的执行可能比普通汽车更智能，但并未超越普通汽车驾驶环境下的人机关系；而如果它需要承担责任，则与伦理追责的一般理念不符，因为伦理与法律不同，通常不追究无心之失。

问题二：自动驾驶汽车知道A醉酒（在技术上要使自动驾驶汽车知道驾驶者是否醉酒，可以在自动驾驶汽车上安装酒精检测仪，并将检测数据即时上传到自动驾驶系统中，再结合智能询问驾驶者的方式进行确定），它的自动行驶行为是否构成与A或B的合谋？如果构成合谋，那么人机合谋在醉驾中的认定和责任划分都是人机关系的新难题；而如果不构成合谋，自动驾驶汽车就没有被视为真正的智能体。

问题三：自动驾驶汽车知道A醉酒，它可以拒绝行驶吗？如果可以拒绝，那么即使是A的家人也无法使用A的汽车送A回家，智能性在道德性的干扰下无法实现（对于该问题似乎也可以从技术上解决，比如自动驾驶汽车在判断A醉酒后，应当拒绝行驶，但如果驾驶位出现清醒驾驶者时，自动驾驶汽车便应启动行驶。但是，自动驾驶汽车如何分辨清醒驾驶者是A的家人、代驾员还是偷车贼呢？有两种帮助自动驾驶汽车分辨的方案：（1）A虽然醉酒，但能清醒告知自动驾驶汽车，当前的清醒驾驶者可以信任；（2）虽然A烂醉如泥，但A事先向自动驾驶汽车输入了可信任人员的信息。但这两种情形均已超越醉驾难题范畴）；而如果自动驾驶汽车不能拒绝行驶，则是对违法行为的助力。

**2. 醉驾难题与电车难题比较**

在自动驾驶领域，醉驾难题和电车难题面临的困境完全不同。电车难题面临的是如何解决自动驾驶汽车伦理选择的问题，解决方案通常包括保护乘客、保护行人和“伦理旋钮”[3]三种。个人在具体情境中的选择只是个人道德困境，而批量生产的具有某种道德选择倾向的自动驾驶系统却会导致普遍性道德困境，“伦理旋钮”避免了普遍性困境。“伦理旋钮”提供了在“乘客-行人”之间自主选择的多元方案，旋钮越靠近0越倾向于保护乘客，越靠近1越倾向于保护行人，在事实上把选择权交回人的手中。“伦理旋钮”反映了自动驾驶伦理研究对电车难题的规避，规避的主要原因是人类道德情感的复杂性。自动驾驶汽车难以理解人的道德选择和道德情感之间的关系，所以很难作出符合人性的选择。比如一个漂亮的小女孩和一个凶恶的大汉在电车难题中，可能会面临驾驶者截然不同的道德选择，这是最高明的人工智能道德编程者也无法解决的问题。如果编程者强行考虑各种道德情感，那么人工智能系统在关键问题上针对不同人群作出的决策和行为将存在偏好差异，极易造成人群之间进一步的隔阂。[4]因此，“伦理旋钮”在智能性的基础上将控制权还给人类不失为一种正确的做法。

然而，人类用“伦理旋钮”接管自动驾驶汽车必然存在时间上的滞后性。相关研究表明，从乘客状态恢复至驾驶员状态不是即时的，平均至少需27秒的时间。[5]人们的判断是一个连续的过程，在注意的情况下作出判断才是最准确的，而在提醒之后，人们需要从对a的注意转向对b的注意，其中的时间差可能会引发致命后果。可见，“伦理旋钮”对电车难题的解决是理想化的。此外，“伦理旋钮”将选择权归还到人类手上，对于自动驾驶汽车的生产和销售虽然具有促进价值，但也可能成为生产者逃避责任的工具，[6]变成生产者将责任转嫁给驾驶者的一种手段。对于生产者而言，在面临保护驾驶者或乘客的安全和保护他人的安全两种选择时，前者显然是获得更多销售利益的优选，而从社会道德培育角度考虑，则需要选择后者。为了避免在利益和道德之间作出选择，生产者宁愿通过技术手段将难题规避。

醉驾难题比电车难题具有更重要的实践价值和伦理治理与法治意义。一方面，电车难题在现实中很难遇到，并且现实情形也更为复杂，而醉驾难题正在发生；另一方面，醉驾难题对

人工智能的具身性提出更高要求。自动驾驶系统在训练时受限于虚拟的数据世界，不能具身地参与现实世界的活动，而伦理和法律问题却来自具有社会具身性的各种活动中，社会具身与道德实践和法律追责紧密相连。数据训练对社会的映射程度有限，对现实情境的过度简化或许会导致人工智能训练结果与实际表现脱节。类似醉驾难题的社会情境不仅难以在数字空间建构，也无法确保人工智能可以理解，并且人很容易在这种情境中绕过人工智能，让人工智能无感无知。此外，在电车难题中，很难直接引入法律，因为电车难题中的选择往往是在人与人的生命间进行，而不是在法与伦理间进行。比如在电车难题的变体隧道难题中，撞向隧道意味着牺牲自己的生命，这其中并未涉及紧急避险等法律问题。但在醉驾难题中，面对人命和交通规则，相对于自动驾驶系统，人通常会毫不犹豫地选择违背交通规则，这是碳基生命对同类的天然共情，人工智能难以理解。

相较于纯粹思辨的电车难题，醉驾是法律明文规定的违法行为，是道路交通安全的巨大隐患。醉驾难题本身至少包含三个法律问题。一是产品法律责任问题：《民法典》第1203条规定，“因产品存在缺陷造成他人损害的，被侵权人可以向产品的生产者请求赔偿，也可以向产品的销售者请求赔偿。”《产品质量法》第46条规定：“本法所称缺陷，是指产品存在危及人身、他人财产安全的不合理的危险；产品有保障人体健康和人身、财产安全的国家标准、行业标准的，是指不符合该标准。”人们在使用自动驾驶汽车时，如果因产品缺陷造成侵权，要由生产者或销售者承担法律责任。在自动驾驶交通事故中，汽车问题零部件制造商、汽车质量检验方、车联网服务者等都可能被纳入产品法律责任中。二是违反交通法规问题：正如一个表达出来的意图也总是隐含了其他未表达的意图[7]一般，人们在购买自动驾驶车辆时，除了便捷地使用自动驾驶系统外，还隐藏着所使用的自动驾驶系统必须遵守交通法规的目的。三是人机共同侵权及其归责问题：人机共同侵权源自人机共驾。人机共驾的任务分摊机制产生的旁观者效应削弱了驾驶者全神贯注于车辆控制的责任动机。[8]当自动驾驶达到L5级别，人类驾驶者只需打开操作系统，输入目的地即可完成驾驶任务，此时驾驶者的身份更像是乘客，任务被更多地分摊到自动驾驶汽车上，这使共同侵权的责任承担变得复杂起来。

因此，醉驾难题所涉及的法与伦理问题更具现实性。从醉驾难题出发，更有助于分析人工智能伦理和人工智能法律之间的融合进路，对于对齐人工智能法治和人工智能伦理治理具有电车难题所不能提供的启发性。

## 二、人工智能伦理治理与人工智能法治对齐的逻辑起点

计算机科学家斯图尔特·罗素（Stuart Russell）于2014年指出，“人们需要发展可证明符合人类价值观的智能，而不是纯粹的智能”。[9]此后，价值对齐（Value Alignment）逐渐成为科技伦理领域的显学，论者所思考的价值超越了价值观的范畴，而广泛地走向人类社会的伦理道德和日常规范。人工智能法治和人工智能伦理治理的对齐不是价值对齐，但价值在伦理维度的难以对齐使人工智能法治和人工智能伦理治理的对齐变得必要，而醉驾难题中法律执行和适用上的困难则进一步为对齐人工智能法治和人工智能伦理治理提供了理由。

### 1. 价值对齐无法在伦理空间完成

随着研究和实践的深入，人们发现人工智能价值对齐面临着规范和技术的双重困境。从规范层面看，关键问题是原则主义无法解决价值一致性问题，即“人工智能应该与哪些价值观或谁的价值观保持一致？”[10]在技术层面上，人工智能大模型的“深度欺骗”[11]又会使对齐变成幻觉。双重困境导致了一些悖论：超越技术的规范是科幻的，滞后于技术的规范是无用的；技术能够符合规范是智能的，技术不能符合规范是悖德的；技术能够绕过规范是技术瑕疵，技术不能绕过规范是智能不足。在深度学习和RLHF领域，由于人工智能目标和人类价值观的潜在错位，带来了重大的生存风险，[12]

所以尽管价值对齐很难实现，人们仍然在尝试对齐。但真正的难题不在于技术和规范，而是价值对齐在伦理空间的两个内在矛盾。

矛盾一：如果人类对于价值性问题处理的很好，那么只需发展技术即可，价值对齐成了纯粹的技术性问题；而如果人类对于价值性问题尚未达成一致，那为何要强求人工智能价值对齐呢？

矛盾二：人们一般认为人工智能不具备具身性是难以做到价值对齐的重要原因，而具身性通常依赖于类人的身体；如果人工智能需要依赖于类人的身体，那它就不可能作出比人类更理性的决策。

虽然价值观的一致性主要是在个人层面上经历的，但它在很大程度上是由组织和社会文化因素决定的。[13] 正因如此，人类对价值性问题达成一致的过程往往包含着妥协和斗争。人们的行动并不总以实用目标为导向，有时也以价值目标为导向，即使是实用目标也通常蕴藏着价值目标，这是无数年来的妥协与斗争所导致的。价值对齐是为了使人工智能的行动与人类价值目标保持一致，而行动本身是实用目标，现有的技术水平无法发挥妥协与斗争的作用，不能在价值嵌入时做到将价值目标藏入实用目标中，所以价值目标与实用目标的冲突使自动驾驶系统难以作出满足人类需求的选择。

为了越过妥协与斗争，科学家和工程师们使用奖励模型来助力嵌入，奖励模型的原理来自人类的习得特征。人们学习了经历过妥协与斗争的道德规范后，在作出与道德规范相符的行为时，就会得到社会和他人的积极反馈，这种反馈让人身心愉悦，从而强化此类行为的动机，催动行为重复出现。显然，习得是具身的产物。在奖励模型帮助下，伦理价值嵌入人工智能类似一种道德物化的过程，人类的伦理认知、判断、选择和行动与人工智能的技术要件、功能和应用场景深度融合，无形的伦理道德被技术物质化，成为人类在日常生产生活中践行道德的一种技术方式。[14] 尽管如此，由于缺乏具身性，奖励模型下的价值对齐仍然是浅层的。

相对于道德，法律更擅长事实评价而非价值评价，虽然自动驾驶汽车不能违背道德要求，但在利益驱动下，自动驾驶汽车的研发和使用行为可能会在伦理漂蓝（ethical bluewashing）中脱离伦理约束，此时法律的事实评价就显得格外重要。法律本身也具有价值性，这种价值性有时直接体现为伦理道德的价值性，所以一些重要的伦理道德通常会作为“义务”存在于法律中，成为“使有序社会得以达致其特定目标的基本规则”。[15] 正因如此，许多伦理问题都可能转变为法律问题，而不少法律问题也可以回溯到伦理问题。更为关键的是，法律的强制力以及由此而具有的可执行性和可问责性是弥补伦理空间缺陷的不二选择。因此，我们需要尝试跳出伦理空间，在伦理与法律的双重空间中解决更高层次的法治与伦理治理的对齐问题。

**2. 法律执行和法律适用上的障碍**

如果将“智能”理解为“代理人在各种环境中实现目标的能力”，[16] 那么自动驾驶汽车就是人工设计的能够感知外界和作出选择的驾驶代理智能体。《道路交通安全法》和《刑法》均未规定自动驾驶汽车所涉及的酒驾和醉驾问题。《最高人民法院关于审理交通肇事刑事案件具体应用法律若干问题的解释》第7条规定，单位主管人员、机动车辆所有人或者机动车辆承包人指使、强令他人违章驾驶造成重大交通事故，以交通肇事罪定罪处罚。但指使和强令更多的是强制，而不是委托，自动驾驶汽车的代理属性指向委托而非强制。因此，现有法律体系并不能覆盖醉驾难题。有学者呼吁为自动驾驶汽车设置交通法规条款，并嵌入自动驾驶系统中。这种方法并不适用于醉驾难题。因为自动驾驶汽车永远不会醉酒，所以法律其实无法设置专门针对自动驾驶汽车的醉驾规定，除非将醉驾与醉酒分离，而这显然是荒谬的。

在刑法上认定醉驾，自然人必须同时具备辨认能力和意志能力，而处罚则需要醉驾者具备受刑能力。自动驾驶系统尚没有意识，无法辨认行为的意义和后果，自然人控制驾驶的认识与自动驾驶系统的认识并不相同。受刑能力以失去财产、自由后的痛苦感受为内核，自动驾驶系统显然不可能拥有这种感受。自动驾驶

系统与物质载体之间的关系与人类意识和身体之间的关系完全不同，它对于自身物质载体的眷恋几乎不可能存在，即使在算法程序中为其设置珍视内置芯片、外观组件的要求，也只是一种强加的“依恋”。此外，受刑的损害性不仅体现在刑罚上，还体现在“犯罪标签”所招致的负面效应上，这对于自动驾驶汽车也没有意义，被认定为“被醉驾过的汽车”并不能减损这辆自动驾驶汽车的声誉。醉驾作为犯罪行为，其附随性后果与其他犯罪行为无异，受到刑事处罚的醉驾者不得从事公务员、律师、公证员等职业，“其子女在报考公务员、警校、军校或在安排关键、重要工作岗位时，难以通过有关的材料审核”。[17]对于自动驾驶汽车而言，这种附随性后果毫无意义，因为根本无从实现。附随性后果可能导致受刑醉驾者及其子女受到歧视性影响，所以前科制度就显得尤为重要。“前科制度的核心是前科消灭”。[18]对于自动驾驶汽车，前科消灭制度也完全没有意义。一辆自动驾驶汽车导致醉驾后果，意味着相同情况会在其他自动驾驶汽车上出现，这辆汽车是否被消灭前科并不会与歧视关联。

自动驾驶汽车最大的法律难题是归责。自动驾驶汽车所导致的交通事故有硬件原因也有软件原因，二者的归责要求有所不同。硬件原因指向材料和环境，而软件原因通常来自算法黑箱，区分软件责任和硬件责任具有非常强的技术性。看的更远一些，如果技术人员可以在云端通过重置解决系统故障，那么当我们对自动驾驶汽车追责时，此时的自动驾驶系统已经不是原来的系统了。在醉驾难题的变体四中，车祸会导致更大的归责难题。自动驾驶汽车本身不能百分百规避车祸，当传感器发生故障、监控设备被遮蔽、网络数据被阻断等，车祸便可能发生。通常而言，如果汽车发出维修警示，而驾驶者置之不理，此时发生的交通事故，应由驾驶者承担责任。[19]但醉酒者A显然不是有意不理，在车辆发出警示时，其可能在睡觉，又或者故意不理的是B，A更不可能承担责任。

有学者提出为人工智能构造一个类人式的身体模型，让它具有人的“需求、欲望、快乐、痛苦、活动方式、文化背景等”。[20]这其实是一种具身主义的思考，具身主义强调身体对认知、情感和价值判断的重要价值，希望人工智能用人类理解世界的方式理解外在信息，在听觉、视觉和触觉等感觉中建立人工智能的人类式具身能力。然而，当前的技术水平显然不足以支撑人工智能实现真正的具身。即使技术水平达到了，只要人工智能不是人，只要人类社会不是人工智能社会，人工智能就无法真正获得与人类相似或相同的感觉。假设存在一个人工智能的社会，无人驾驶汽车撞毁了一个机器人，对其进行处罚才可能具有具身性，但这一天也许永远不会到来。

人工智能伦理具有清除法律执行和适用上障碍的作用。人工智能伦理的道德性及其通过人的内化而产生的道德自觉，能够有效解决法律依据缺失和强行适用现行法律而导致荒诞结果的问题。无论是伦理还是法律都无法单独应对醉驾难题，事实上，类似醉驾难题的情况会在人工智能深度介入人类社会后不时发生，所以必须依靠人工智能伦理治理和人工智能法治的对齐来应对已经发生和终会到来的各种挑战。

## 三、人工智能伦理治理与人工智能法治对齐的具体路径

价值对齐可以分为极简主义（minimalist）和最高主义（maximalist），极简主义指将人工智能与人类价值合理地联系以避免不安全的结果，最高主义涉及在全社会，甚至全球范围内将人工智能价值观与最佳的人类价值观保持一致。显然，人们关注更多的是最高主义的价值对齐。但本文认为，人工智能伦理治理和人工智能法治对齐的具体路径应在极简主义理念下构建。程序适用和责任追究的一致是人工智能伦理治理和人工智能法治对齐的关键，基于极简主义可以将对齐程序简化成生产设计、投入市场后的监督和责任追究三个阶段，进而通过人工智能伦理治理为人工智能法治实施提供道德性的守法自觉保障和依靠人工智能法治弥补人工智能伦理追责不具备强制性的缺陷。

**1. 为什么必须是极简主义**

人工智能价值对齐意义上的极简主义“将人工智能与人类价值的某种合理模式联系起来，并避免不安全的结果”。[21]作为一种方法论，极简主义主张在技术层面避免复杂的道德原则编码，而聚焦于基础性的价值约束。由于每个人所珍视的东西不同，所追求的目的也不同，人们很难就所谓的最佳价值观达成一致，但就人们所不想发生的事情形成最小共识却是可能的，基础性的价值约束正是以排除不安全的结果为内核，这是人们接受价值约束的原初理由。尽管法律与道德有诸多不同，但两者都希望规避不安全的结果，所以在极简主义下更有可能实现人工智能伦理治理与人工智能法治的对齐。

极简主义确实面临着价值空洞的诘难。基础性价值约束通常不追求积极义务的设定和履行，而只通过设置避免明确危害（如禁止暴力或防止污染等）的消极义务实现约束。由于缺乏积极义务，极简主义所设定的基础性价值可能会陷入人们都觉得重要却又无人真正理会的境地。但是，在人工智能伦理治理和法治的对齐中，价值空洞问题并非无解。法律和道德具有诸多差异，但核心区别永远是强制与否。法律本身包含大量的禁止性规范，伦理治理虽然以积极义务为追求，但认定主体是否违反消极义务却是其主要抓手。人工智能伦理治理和法治在对齐的交汇点上将基础性价值和禁止性规范相融合，并以法律强制力将这种融合贯穿于设计、监督和追责全过程，从而最大化解决价值空洞问题。

极简主义所具有的应对价值多元化的优点同时也是其缺点。极简主义降低了价值规范的复杂性，因而大大提升了系统的安全性和跨文化兼容性，但最小共识却可能遮蔽多元利益，特别是将最小共识和价值观最大公约数——主流价值观相对应时。极简主义摒弃冗余偏好，聚焦核心价值，而现代社会的核心价值通常是主流价值观，主流价值观本身并不排斥多元利益，只是将多元利益中的核心利益、长远利益和正当利益进行价值化改造，但人们可能因为一时的利益需求不能满足而对主流价值观产生误解。因此，当我们将极简主义作为指导理念，并将主流价值观作为对齐基准时，首先要做的是在技术设计时构建容错机制，在遵循主流价值观的同时，适度宽容人工智能社会化应用中相关主体因追求自身利益而作出的行动。

**2. 在设计时构建容错机制**

容错机制是新兴技术得以顺畅发展的前提，人工智能伦理治理和人工智能法治的对齐应以设计时的容错机制为起点。

除了生产者外，销售者因为销售自动驾驶汽车被纳入醉驾难题责任体系中，车联网服务提供者也因自动驾驶汽车的联网特性而要承担责任，自动驾驶对数字地图的依赖则使地图绘制部门、导航服务商牵连到自动驾驶汽车的交通违法行为中。如果这样，自动驾驶汽车涉及到的责任主体将多不胜举，这显然不利于技术发展，也不利于相关主体维护自身利益。本文提出容错机制在法律上应以新过失论为构造基础。新过失论的责任标准是“违反注意义务”，注意义务通常包括对结果的预见义务和回避义务，其中回避义务是关键，这与以往的过失论不同。新过失论划定了一条风险承担界线，逾越这道界线方有过失。[22]新过失论将自动驾驶中“生产者虽预见危害结果但却无法避免危害结果的情形排除”。[23]具体而言，在设计自动驾驶汽车时，必须要符合国家安全标准，使用者使用符合标准的自动驾驶汽车发生交通事故，自动驾驶汽车的研发者或生产者将不再承担过失责任，同样，销售者、网络安全部门和地图测绘部门、导航服务商等也在严格执行国家标准的前提下脱离责任追究范围。对于驾驶者而言，新过失论要求驾驶者按照规定操作，否则便没有尽到注意义务，而这种规定应当包括不违背交通法规。

但是，我们既不能用伦理的方法为自动驾驶系统等人工智能容错——人工智能毕竟与人不同，人类不仅用大脑去理解世界，还用双脚行走世界、双手改造世界，人工智能的活动范围由人决定，这一范围与人对其进行使用的范围几乎完全一致，因此我们虽然可以适当允

许人工智能的道德选择与人类有所不同，但对法律的遵守必须一致；也不能用法律的方法为人类主体找借口——有学者认为“无人驾驶将根据算法对周围地区的观测结果自主地采取行动，因此，生产者对无人驾驶的作为或不作为就不再负有责任，除非他能够预见并避免无人驾驶的有害行为，但这样的预见实际上几乎是不可能的”。[24]这种观点不能成立，否则生产者在生产之后就高枕无忧了。容错机制集中体现为一种容许风险存在的机制，但并不意味着一切风险都允许发生。在极简主义要求下，为了规避不安全的结果，生产者首先要将数据法中的“自设计保护隐私”（privacy by design）理念融入到自动驾驶汽车的道德判断中，将法律精神前置到技术设计中，实现技术的自我治理；其次要在自动驾驶系统中嵌入紧急避险制度，以使自动驾驶系统在紧急状态下采取特定措施避免更大的权益受到损害。在醉驾难题中，该制度要适当变化成合法利益和违法可能之间的选择，即遇有醉酒驾驶员时自动驾驶汽车要避免被启动，同时通知紧急联系人，这一变化体现了遵守交通法规是第一法则，紧急避险是例外法则的要求；再次，作为“技术、法律及伦理之间的沟通渠道”的算法伦理审计[25]应当常态化开展，强化对设计的严格要求；最后，生产者在产品投入市场后如果发现存在设计问题等缺陷，必须及时召回自动驾驶汽车，否则要承担责任。

人工智能的价值训练是在技术环境下完成的，真实的社会生活实践被简化成数据喂养，人工智能因而很难处理好复杂的人类社会关系。一个简单的例子是，在利益衡量下，人们往往希望自己乘坐的汽车采取利己主义算法，却又希望其他汽车采取功利主义算法。[3]但自动驾驶系统显然很难理解这种想法。《生成式人工智能服务管理暂行办法》《互联网信息服务深度合成管理规定》《互联网信息服务算法推荐管理规定》等均提出坚持主流价值观的要求。因此，规避不安全结果的价值观显然不是人们的个人价值观，而是已经确立的主流价值观，这突破了人工智能不能理解单个人想法的困境。在醉驾难题中，“规避不安全结果”指向自动驾驶系统和A、B、交警在实现交通安全的共同目标过程中对合法行驶、安全到家的认同和追求。容错机制的意义正是通过极简主义使个人利益、社会价值和法律规定保持平衡，在“规避不安全结果”的根本标准下实现人工智能伦理治理和人工智能法治的对齐。

通过将容错机制嵌入生产设计，醉驾难题中聚焦于机器责任的伦理三问可以得到较为妥善的解决。我们并不用在伦理上辨明自动驾驶汽车在不同情形中是否需要承担责任以及如何与人共担责任，而仅需通过立法设置国家标准化规定，为自动驾驶汽车接受指令行为划定过失边界，自动驾驶汽车仅在紧急避险时可以突破该边界，否则即需承担责任，且这种责任在现阶段要由生产者承担。如此，自动驾驶汽车的责任承担就被法律强制性所固定，法治的可执行性和可问责性有效解决了伦理判断的模糊性问题。

**3. 在监督时融合伦理与法**

人工智能伦理治理与法治之间的冲突主要表现为伦理和法之间的错位。从伦理和法律的区别来看，当人们说“我不知道这样做是错的”，代表的是道德判断，而这句话在法律上的本质内涵是“我不知道这样做是违法的”，错和违法是伦理与法律的重要分野。在醉驾难题中，即使自动驾驶汽车知道单独载着醉酒者行驶是错的，也不代表它知道这是违法的。因为仅仅依靠技术设计无法使自动驾驶汽车真正理解错和违法之间的联系与区别，所以要依据极简主义规避这种无法理解所可能导致的危害，必须在自动驾驶汽车投入市场后的监督中融合伦理与法。

在监督时融合伦理与法的前提要求是自动驾驶汽车不能仅学习人类社会的价值规范，还要将遵守法律规定作为人工智能行动的重要依据。这意味着如果生产者遵守法律规定将交通法规所对应的各种违法情形嵌入算法中，那么法律可以适当豁免其责任。由于“目前的法律框架，并不是为了容纳人工智能生成的裁决而设计的，这引发了人们对其合法性、程序公平

性和可执行性的质疑”，[26]而各种伦理框架虽然开始纳入人工智能，但仍然是用对待人的经验来对待人工智能，风险的发生难以避免。此外，人工智能系统可能会谎称完成了对齐任务，故意误导人类。[27]因此，在监督时融合伦理与法，要将法律规定作为红线，而将伦理规范作为标准，比如一旦由于违法而发生醉驾行为，此类自动驾驶系统必须立即退出市场，直到新的自动驾驶系统通过融合了法律的伦理测试，方可再次进入市场。相关行政部门，主要是工信部门和交通部门是监督主体，它们在监督中要求销售者充分尊重公众的知情权，让公众了解所购买的自动驾驶汽车究竟用了什么样的伦理算法，算法设计采纳了哪些伦理原则；还要对自动驾驶汽车设计、制造和使用全过程进行一体化监督，着眼于自动驾驶汽车上路的目的、方式、后果，构建多元主体纠正违法行为的体系。

人工智能价值对齐的挑战主要是确立具有公平、安全、平等和有序等特征的人工智能原则，确立的过程需要一定的强制性，法律在其中能够发挥重要作用，法治和伦理治理的对齐更是如此。因此，在监督时，必须发挥法律的强制性功能。自动驾驶系统内部的价值规范和设计要求不是理论家所想象的自动契合，而需要法律强制性的力量来约束设计者；价值规范和最终结果之间也不是同一的，各种外在因素都有可能影响结果的呈现，但不能因为结果不好就修改被嵌入的规范，首先要修改的应当是设计要求，此时也需要通过法律的强制性向生产者和研发者进行反馈，强制要求使用更适合的方式进行设计和嵌入。同时，法律也不能取代伦理，“通过法律和公共政策直接统一人工智能价值体系”是行不通的。[28]法律规则是红线和底线，相对缺乏灵活性，而伦理规范具有适当的弹性，不会在监督时过度约束自动驾驶系统等人工智能技术的发展。

法律红线和伦理标准在监督时的融合有助于解决以确定驾驶者主体为核心的醉驾难题法律四问。法律的刚性特征往往将判断限定到是或否的二元结构中，伦理标准的弹性化则在是否中加入对错，丰富判断的依据。当醉驾事件发生后，即使不能根据现行法律确定驾驶者，但只要自动驾驶汽车在当时情形下的启动和行驶与最小共识的安全伦理标准不符，造成或可能造成错的结果，即可将相关主体纳入驾驶者范畴，从而适用交通法规和刑事法律进行追责。

**4. 在追责时展开比例原则**

醉驾属于行政犯罪，是行刑交叉领域的典型犯罪行为，在追究责任时必须要考虑这一点。行政犯罪的本质是对行政法规定的命令或禁止的不服从，其根源是行政法领域的危害行为。预防和制裁此类犯罪更多的是为了维持行政管理秩序、达到行政管理目的。醉驾的伦理违反性较小，对醉驾的处置主要是为了维持行政管理秩序，所以规制此类犯罪要优先适用行政处罚。[29]在实践中，“出罪入行”的醉驾案件越来越多，而相较于一般醉驾案件，对自动驾驶所导致的醉驾案件转向行政处罚，会更容易得到公众认同。行政规制可以对潜在的风险与威胁进行监控并采取措施，预防损害的发生。[30]在醉驾难题中，当B操作自动驾驶系统时，自动驾驶系统可能因技术瑕疵或故障而无法识别操作者是不是驾驶者，有学者“针对它（指机器人——本文作者注）的故障提出（类比于人类）精神错乱的辩护”。[31]在无法识别时，自动驾驶系统不应担责，除非生产者生产的自动驾驶汽车能够识别和拒绝非驾驶者的操作行为，或者提供了禁止非驾驶者操作的选项，汽车使用者已经启用，并且自动驾驶汽车也没有发生故障。否则，责任只能落到生产者、操作者或使用者等人类主体身上。

比例原则是行政法中的重要原则，在处置责任时具有非常灵活的价值，它绝不是简单地减轻责任，而是起到更多元的责任追究作用。处理醉驾难题时应适用比例原则。根据比例原则，对于醉驾责任者有多种处理方式时，可以从中选择最有效率、对责任者利益损害最小的一种。基于比例原则，对醉驾难题进行伦理治理与法治的对齐时，要确保程序公平，不给任何一方带来不必要损害；追责措施既要覆盖全面和保持稳定，又要随时间推移而变化；同时必须宽容，不能因手段过于强硬而导致生产者、

研发人员对自动驾驶技术的发展失去信心。

比例原则要求在法律追责和伦理追责之间进行切换。在醉驾难题中，如果A完全不知情，仅对A进行道德谴责是可行的。这种切换有利于顺利完成追责，因为法律责任不仅包括行政责任和刑事责任，还有民事责任，温和的伦理责任在此时可能会经由社会舆论而产生较大的民事履责动力。比如在变体四中，责任人可能会因为仅需承担伦理责任而更加主动地对受害者进行赔偿。“知道或应当知道”是法律追责的重要标准，生产者、设计者和销售者的责任追究一般要以该标准为前提。如果设计者知道生产者或销售者利用设计的缺陷从事违法活动，则三方共责，但比例原则要求对设计者进行较小的惩处，甚至可以转为行政处罚，从而鼓励技术创新。这种思路与以往的“责任链”[32]不同，即不能因为一方为他方的行动提供了前提条件，就不分彼此地承担责任，即使要承担，不仅要对责任的大小进行划分，对责任的性质也要进行区分。比例原则也因此解决了可以承担义务的人越多，就越有可能没有人会真的尽义务[33]的多重义务设计难题。

实践中，一些地方对于符合相对不起诉条件、自愿参加社会公益服务的醉驾犯罪嫌疑人，采取开展法律宣传、社区公益服务、接受安全驾驶教育等处置措施，这实际上是社会服务罚，比行政处罚更轻。在醉驾难题中，如果A没有造成交通事故，在比例原则指引下，也可以通过这种方式进行追责。总之，限制公民权利的法治手段要与欲实现的社会道德价值保持比例关系，在适当性、必要性和效率性的方式中实现技术发展与人的便利并行不悖。

## 结　　语

醉驾难题作为人工智能深度介入人类社会生活的缩影，不仅暴露了单一依赖伦理或法律手段进行人工智能治理的局限性，更揭示了人工智能伦理治理与人工智能法治必须走向对齐的深刻逻辑。本研究超越“价值对齐”的传统范畴，将如何让人工智能符合人类价值的讨论，推进至如何实现伦理治理与法治的协同。价值对齐在伦理空间的内在矛盾，以及法律在适用与执行中遭遇的现实障碍，是推动人工智能伦理治理与人工智能法治对齐的逻辑起点。为实现“善治”目标，本文主张摒弃追求完备价值体系的“最高主义”进路，转而采纳以规避明确危害、凝聚基础共识为核心的“极简主义”理念。基于此，本文构建了贯穿技术生命周期的三阶段对齐路径：在设计阶段嵌入容错机制，以“新过失论”为技术创新预留发展空间；在监督阶段融合伦理标准与法律红线，以刚性约束守护弹性规范；在追责阶段适用比例原则，在责任承担与价值维护之间寻求审慎平衡。这一路径不仅为破解醉驾难题提供了具体的分析工具，也为应对更广泛的人机协同困境提供了治理思路。

人工智能伦理治理和法治的对齐并非简单的规则叠加，而是在极简主义理念指导下的深度融合。这一对齐路径的最终指向，是在智能社会中建构“德法共治”秩序图景。它既不是简单地让技术臣服于法律，也不是用道德束缚技术的发展，而是通过法律的强制要求与伦理的柔性引导，共同塑造一种“技术向善”的实践逻辑。唯有如此，我们才能在享受人工智能带来便利的同时，确保人类社会的基本规范在智能时代依然熠熠生辉，并以之构建一个既能促进技术创新，又能保障人类价值与社会秩序的可信人工智能治理体系，最终达成人工智能技术发展与人类福祉和谐共生。

［参 考 文 献］

[1]Moolayil, A. K. 'The Modern Trolley Problem: Ethical and Economically-sound Liability Schemes for Autonomous Vehicles'[J]. *Case Western Reserve Journal of Law, Technology & the Internet*, 2018, 9(1): 13.

[2]Beiker, S. A. 'Legal Aspects of Autonomous Driving'[J]. *Santa Clara Law Review*, 2012, 52(4): 1145.

[3]Contissa, G., Lagioia, F., Sartor, G. 'The Ethical Knob: Ethically-customisable Automated Vehicles and the Law'[J]. *Artificial Intelligence & Law*, 2017, 25(3): 365-378.

[4]Bakker, M. A., Chadwick, M. J., Sheahan, H. R., et al. 'Fine-tuning Language Models to Find Agreement Among

Humans with Diverse Preferences'[J]. arXiv preprint, arXiv: 2211.15006, 2022.

[5]Biondi, F., Alvarez, I., Jeong, K. A. 'Human-vehicle Cooperation in Automated Driving: A Multidisciplinary Review and Appraisal'[J]. *International Journal of Human-Computer Interaction*, 2019, 35(11): 932-946.

[6]Goodall, N. J. 'Ethical Decision Making During Automated Vehicle Crashes'[J]. *Transportation Research Record*, 2014, 2424(1): 58-65.

[7]Grice, H. P. 'Logic and Conversation'[A], Cole, P., Morgan, J. L. (Eds.) *Syntax and Semantics 3: Speech Acts*[C], New York: Academic Press, 1975, 41-58.

[8]Darley, J. M., Latané, B. 'Bystander Intervention in Emergencies: Diffusion of Responsibility'[J]. *Journal of Personality and Social Psychology*, 1968, 8(4): 377-383.

[9]Lanier, J. 'The Myth of AI'[EB/OL]. https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai. 2014-11-14.

[10]Sutrop, M. 'Challenges of Aligning Artificial Intelligence with Human Values'[J]. *Acta Baltica Historiae et Philosophiae Scientiarum*, 2020, 8(2): 54-72.

[11]Shevlane, T., Farquhar, S., Garfinkel, B., et al. 'Model Evaluation for Extreme Risks'[J]. arXiv preprint, arXiv: 2305.15324, 2023.

[12]Duim, J. L., Mesquita, D. P. P. 'Artificial Intelligence Value Alignment via Inverse Reinforcement Learning'[J]. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 2025, 11(1): 1-2.

[13]Neiterman, E., Ladha, R. 'Values Alignment'[A], Webber, S., Babal, J., Moreno, M. A. (Eds.) *Understanding and Cultivating Well-being for the Pediatrician*[C], Schaffhausen: Springer Cham, 2023, 303-322.

[14]Verbeek, P. P. *Moralizing Technology: Understanding and Designing the Morality of Things*[M]. Chicago: University of Chicago Press, 2011, 113.

[15]富勒. 法律的道德性[M]. 郑戈 译, 北京: 商务印书馆, 2005, 7.

[16]Legg, S., Hutter, M. 'Tests of Machine Intelligence'[A], Lungarella, M., Iida, F., Bongard, J., et al. (Eds.) *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*[C], Berlin, Heidelberg: Springer, 2007, 232-242.

[17]周光权. 论刑事一体化视角的危险驾驶罪[J]. 政治与法律, 2022,（1）: 14-30.

[18]陈兴良. 轻罪治理的理论思考[J]. 中国刑事法杂志, 2023,（3）: 3-18.

[19]Bose, U. 'The Black Box Solution to Autonomous Liability'[J]. *Washington University Law Review*, 2015, 92: 1325.

[20]Dreyfus, H. L. *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*[M]. New York: Oxford University Press, 2014, 272-273.

[21]Gabriel, I. 'Artificial Intelligence, Values, and Alignment'[J]. *Minds and Machines*, 2020, 30(3): 411-437.

[22]林东茂. 刑法综览[M]. 北京: 中国人民大学出版社, 2009, 133.

[23]刘宪权. 涉人工智能产品犯罪刑事责任的归属与性质认定[J]. 华东政法大学学报, 2021, 24（1）: 50-59.

[24]Zapusek, T. 'Artificial Intelligence in Medicine and Confidentiality of Data'[J]. *Journal of Health Law*, 2017, 11: 105.

[25]张涛. 通过算法审计规制自动化决策以社会技术系统理论为视角[J]. 中外法学, 2024, 36（1）: 261-279.

[26]Alhasan, T. K. 'Integrating AI into Arbitration: Balancing Efficiency with Fairness and Legal Compliance'[J]. *Conflict Resolution Quarterly*, 2025, 42(4): 523-534.

[27]Hubinger, E., Van Merwijk, C., Mikulik, V., et al., 'Deceptive Alignment, AI Alignment Forum'[EB/OL]. https: //www.alignmentforum.org/posts/zthDPAjh9w6Ytbeks/deceptive-alignment. 2019-06-06.

[28]Wang, P. 'Jurisprudential Logic and Institutional Pathways of Personalized Value Alignment in AI'[J]. *China Legal Science*, 2024, (12): 109-132.

[29]刘艳红、周佑勇. 行政刑法的一般理论（第2版）[M]. 北京: 北京大学出版社, 2020, 12.

[30]Scherer, M. U. 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies'[J]. *Harvard Journal of Law & Technology*, 2015, 29(2): 353-398.

[31]Hallevy, G. 'I, Robot-I, Criminal: When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses'[J]. *Syracuse Science & Technology Law Reporter*, 2010, 22: 1-37.

[32]奥特弗利德·赫费. 作为现代化之代价的道德[M]. 邓安庆、朱更生 译, 上海: 上海译文出版社, 2005, 72-73.

[33]黄荣坚. 交通事故责任与容许信赖[J]. 月旦法学杂志, 1999,（7）: 178-189.

[责任编辑 李斌]