

# 人机共驾接管失败时安全员的过失责任

## ——从信赖原则的适用切入

杨 宁

**摘 要** 目前，人机共驾接管失败致死的案件备受瞩目，对于安全员虽未履行接管义务，但可否主张信赖原则而不承担过失责任，存在肯定论与否定论的争议。从肯定论看来，人机信任建立在人工智能可信任性的基础上——机器有自主性、能动性和以人类为中心的意向性，并偏重人类对机器的信任。对高度分工、高风险的人机信任适用信赖原则，是组织模式的一种延展，不会导致责任稀薄，更符合刑事责任性质，具备合理性。信赖原则例外否定过失责任时，以不同自动驾驶水平下安全员、自动驾驶系统和人机交互作为客观因素，可具体判断信赖的相当性。安全员没有履行接管义务的行为具有客观违法性，但当其合理信任自动驾驶系统能够执行预期功能时，不能具体预见结果的发生，由此信赖原则在责任阶段否定过失，发挥后端风险结果分配的作用。

**关键词** 接管义务 人机信任 信赖原则 过失责任

人工智能技术的应用塑造着驾驶的新形态——人类与人工智能系统交互掌握驾驶权。具备车道保持、自适应巡航控制等 L2 级及以下功能的辅助驾驶系统已经走向了千家万户。<sup>①</sup> 近期，我国密集出台了一系列监管法规为蓬勃发展中的自动驾驶技术应用保驾护航。2023 年 11 月，工业和信息化部、公安部、住房和城乡建设部、交通运输部发布《关于开展智能网联汽车准入和上路通行试点工作的通知》及指南（以下简称《上路试点指南》），制定了符合我国 L3 级到 L4 级驾驶自动化功能的车辆准入试点的规范。<sup>②</sup> 2024 年，武汉市“萝卜快跑”自动驾驶出租车的新闻火爆全网，它代表着我国正在试点运营的智能网联汽车，<sup>③</sup> 在符合高水平自动驾驶系统技术设定的情况下，驾驶权将

\* 杨宁，法学博士，天津大学法学院副教授。本文系国家社会科学基金青年项目“人机共驾模式下刑事责任问题研究”（20CFX026）的阶段性成果。

① 2023 年我国组合辅助驾驶功能的乘用车销量 995.3 万辆，市场渗透率达 47.3%。参见李斌等：《2023 年自动驾驶发展状况与趋势》，载唐维红主编：《中国智能互联网发展报告（2024）》，社会科学文献出版社 2024 年版，第 164 页。

② 依据 2022 年生效的中国推荐性国家标准《汽车驾驶自动化分级》（GB/T 40429-2021）定义的自动驾驶级别，级别 3（有限自动化）：辅助系统能够在特定条件下完全控制驾驶任务，但驾驶员仍需随时准备在必要时接管控制。驾驶员可以将注意力转移到其他事务上，但仍需保持对驾驶环境的警觉。级别 4（高度自动化）：车辆能够在特定环境和条件下完全自主地执行驾驶任务，无需驾驶员的干预。然而，在极端条件如恶劣天气或复杂交通环境下，驾驶员可能仍需接管控制。即使系统出现故障，L4 级别的自动驾驶系统通常配备有冗余系统，以确保车辆能够继续安全行驶或停车。参见工业和信息化部装备工业一司：《开展试点工作促进智能网联汽车产品迭代优化——〈关于开展智能网联汽车准入和上路通行试点工作的通知〉解读》，载《中国电子报》2023 年 11 月 24 日。笔者说明，本文未采取美国汽车工程师协会 2021 年新版“SAE International Surface Vehicle Recommended Practice J3016™”，因为国标更具时效性、地域性。

③ 参见郭晨：《“萝卜快跑”火了背后》，载《中国汽车报》2024 年 7 月 22 日。

由驾驶系统转交安全员接管。安全员是指智能网联汽车运行的安全保障人员。<sup>①</sup>

高水平自动驾驶将真正实现人机共驾，但全球业已发生数起由此造成的人身伤亡事故，凸显法律责任认定的难题。从目前国际上对于自动驾驶汽车的法律法规、政策指引看，一般可以按照“谁掌控驾驶权谁负责”的基本思路来归责。亦即，自动驾驶系统掌握驾驶权时发生交通事故，安全员无责任。但是，许多规范又规定安全员具有“始终监控和随时接管”的义务，亦即接管规则。例如，2023 年我国《上路试点指南》明确规定“车辆上路通行过程中，安全员应当处于车辆驾驶座位上，在自动驾驶系统激活状态下，监控车辆运行状态及周围环境，当系统提示需要人工操作或者发现车辆处于不适合自动驾驶的状态时，及时接管或者干预车辆并采取相应措施。”这样一来，就出现了黑白分明的责任划分解决不了的模糊地带：自动驾驶系统掌控驾驶权时发生死亡事故，但安全员没有履行“始终监控和随时接管”的义务，他需要承担刑事责任吗？

## 一、信赖原则如何适用于人机共驾的争议

2018 年发生的第一起 UBER 公司的 L3 级别自动驾驶车辆造成路人死亡的刑事案件，就是一桩典型的人机共驾接管失败事故。经历漫长的事故调查和诉讼过程后，2023 年 7 月 28 日，美国亚利桑那州州立法院判决安全员 R 在首次涉及自动驾驶汽车的致命碰撞中构成危害罪，处三年缓刑。<sup>②</sup>对于该案，早有研究提出或可借用信赖原则否定安全员 R 的过失责任。<sup>③</sup>信赖原则是指，“行为人在具有相当理由信赖被害人或者第三人会采取适当行为的场合，以此为基础实施合适的行为即可；即使被害人或者第三人并没有依照信赖行事而引发了结果，行为人不承担过失。”<sup>④</sup>对此存在肯定说与否定说的尖锐对立。肯定说认为，安全员可以主张对于高水平自动驾驶机器的信赖，从而为没有顺利接管造成的事故免责。<sup>⑤</sup>系统没有发出请求，最终发生了交通事故的情况，此时因为汽车是处于系统控制当中，安全员可以被认定对系统具有信赖基础，适用信赖原则。<sup>⑥</sup>其理由在于：机器人可以独立完成一项任务不需要人类干涉时，人类可以信赖机器人的合规则行为。<sup>⑦</sup>因此可以将人类之间的信赖原则同样运用到人机、机器之间。<sup>⑧</sup>否定论的观点也十分有力。理由主要是：德国等国家法律明确规定 L3 级自动驾驶时安全员需要承担警觉、监控和接管等严苛义务，其前提是对于这

① 广义的安全员是指包括（狭义的）车上安全员和平台安全监控人员。2023 年《上路试点指南》第二部分规定使用主体，智能网联汽车运行安全保障人员包括车上安全员和平台安全监控人员。一般而言，L3 级别自动驾驶需要配备车上安全员，L4 级别自动驾驶需要配备平台安全员。L5 级别完全自动驾驶为系统驾驶，不在本文讨论范围之内。

② 2020 年 9 月 15 日安全员 R 涉嫌过失杀人罪被起诉，理由是 R 的分神、没有履行“始终监控和随时接管”的义务。危害罪被定义为“鲁莽地危害他人，并有即将死亡或人身伤害的重大风险”。相对于被指控的过失杀人罪而言，危害罪是一个轻罪。See Rebekah Riess & Zoe Sottile, “Uber self-driving car test driver pleads guilty to endangerment in pedestrian death case”, <https://edition.cnn.com/2023/07/29/business/uber-self-driving-car-death-guilty/index.html>, visited on July 29, 2024.

③ 参见〔日〕山下裕樹：《AI・ロボットによる事故の責任の所在について——自動運転車の事案を中心に——》，载《ノモス》第 45 卷（2019 年），第 100 页；杨宁：《刑法介入自动驾驶技术的路径及其展开》，载《中国应用法学》2019 年第 4 期，第 117 页。

④ 〔日〕西原春夫：《交通事故と信頼の原則》，成文堂 1969 年版，第 14 页。

⑤ 参见蔡仙：《人机共驾模式下的接管义务及其刑事归责》，载《苏州大学学报（法学版）》2023 年第 3 期，第 37 页。

⑥ 参见储陈城：《自动驾驶时代：交通肇事如何适用信赖原则》，载《检察日报》2019 年 2 月 16 日。

⑦ 参见马天成：《自动驾驶致损的刑事责任研究》，载江溯主编：《刑事法评论：刑法的科技化》，北京大学出版社 2020 年版，第 97 页。

⑧ 参见王华伟：《论人形机器人治理中的刑法归责》，载《东方法学》2024 年第 3 期，第 110 页。

一阶段自动驾驶系统的怀疑。<sup>①</sup> 信赖原则适用于交通参与人，不能拓展到人与机器的领域之中。<sup>②</sup> 肯定论仅初步提出观点，缺乏对人机信任特殊性的分析和论证，在如何结合已有规范指引认定责任的具体步骤上亦存在跳跃性。这些关键节点上的语焉不详给否定论提供了靶子，容易遭到轻纵安全员的批评。总体看来，学界对于如何进一步具体适用信赖原则的问题尚未深入研讨。这远不足以应对未来复杂的人机共驾接管事故。

随着我国高水平自动驾驶的法律规范逐步明晰和商业试点的大规模开展，面对可能发生的棘手问题，不妨预先进行一个有益的思想实验：若在我国智能网联汽车商业试点过程中发生了致死事故，而安全员辩称由于信任自动驾驶系统、不负过失责任，应如何处理？

### 【Q 案件】

2024 年 10 月 18 日晚间，一名安全员王某坐在一辆 Q-1 自动驾驶测试车驾驶座上，在某市的公共道路行驶时与一名穿越该道路的女性行人相撞，造成该行人死亡。这辆测试车由 Q 公司运营，获得商业测试牌照、配备 L3 级别自动驾驶系统功能。事故发生时路面干燥、有路灯照明，且自动驾驶系统处于激活状态。在自动驾驶系统控制车辆平稳运行 19 分钟后，测试车以 58km/h 的速度接近事故发生点；行人开始推着一辆自行车横穿马路。碰撞发生前 5 秒，车辆首次监测到目标，但是未识别出行人。同时，自动驾驶系统没有正确预测目标的运动路径，未降低车速。碰撞前安全员王某曾多次将视线转移至手机，并停留较长时间。在碰撞发生前 5 秒，王某最后一次将视线转移至中控台，但并未发现异常。直到碰撞发生前 1 秒，王某才将视线转移回前方，碰撞前 0.02 秒开始向左打方向盘、但未成功控制车辆。汽车以 53km/h 的速度与行人相撞。事后查明，测试团队为避免可能的信号干扰，此次测试中关闭了测试车安装的前向碰撞预警和紧急制动功能。王某是一名经验丰富、记录良好的司机，通过专门培训并曾在该路段正常测试 20 次。<sup>③</sup>

本文将以人机信任的研究为基础，夯实肯定信赖原则适用于人机共驾的前提，结合人机共驾的特性论证适用信赖原则的合理性。围绕 Q 案件所体现的具体驾驶场景，分析不同级别自动驾驶技术中人机交互特点，进一步探讨信赖相当性的判断和对于安全员过失责任的影响。

## 二、人机信任是适用信赖原则的前提

信赖原则源于“可允许的危險”与风险分配的法理。<sup>④</sup> 它是由 1935 年德国判例发展而来，其社会背景是大规模工业带来了诸如现代化交通这样的新生活方式。为了享受它带来的利益，人类不得不放下“零危險”的幻想，在综合考量后认为某种程度的危险行为由于其有益性、必要性而允许其存在。“可允许的危險”作为一种最底层的观念，对于限缩过失的范围有重要意义，但在将抽象性观念融入犯罪阶层体系时引发了争议。<sup>⑤</sup> 相对而言，信赖原则更具体地讨论：行为人在某一行为之

<sup>①</sup> 参见王莹：《法律如何可能——自动驾驶技术风险场景之法律透视》，载《法制与社会发展》2019 年第 6 期，第 106-107 页。

<sup>②</sup> 参见皮勇：《论自动驾驶汽车生产者的刑事责任》，载《比较法研究》2022 年第 1 期，第 64 页。

<sup>③</sup> Q 案件以 UBER 案件为原型进行改造，案情参考调查报告。See National Transportation Safety Board, “Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian”, pp. 1-2, <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>, visited on Oct. 29, 2024.

<sup>④</sup> 参见陈子平：《刑法总论》，中国人民大学出版社 2009 年版，第 150 页。

<sup>⑤</sup> 对于“可允许的危險”概念的批评意见，参见黎宏：《刑法总论问题思考》，中国人民大学出版社 2016 年版，第 261 页；张明楷：《被允许的危險的法理》，载《中国社会科学》2012 年第 11 期，第 127-128 页。

际,可否因信赖他人遵循规范的行为,而影响其过失责任的成立。而今信息技术与生物技术的发展所带来的“第四次工业革命”,给刑法中过失教义学带来了根本性的挑战。<sup>①</sup>在讨论信赖原则适用于人机共驾前,一个重要的基础命题有待论证:人与机器(自动驾驶系统)之间也存在与人类之间同样的心理基础,足以对人的规范意识产生影响。这一前提常常遭到质疑,反映出人机交互对于传统心理责任的挑战。

### (一) 人机信任的基础

信赖原则之所以能起到限制过失责任的作用,根源在于它描述了人的心理责任事实:行为人与被害人或者第三人之间具有信赖的心理关系。现有理论将信赖关系分为两种:对向型,在交通事故中“行为人—被害人”实施的对向行为时,对对方遵从规范行为的信赖;同向的组织模式,在加害人内部“监督人—被监督人”“分工人之间”实施行为时,对他人遵从规范行为的信赖。<sup>②</sup>对于高水平自动驾驶中安全员与自动驾驶系统共享驾驶权的问题,上述肯定说希望借助组织模式来分析安全员与自动驾驶系统的关系。而否定说则认为人与机器的领域不符合信赖的基础。<sup>③</sup>

人与机器的信任问题已经成为一个跨学科的研究命题,涉及心理学、社会学、经济学和计算机科学等多个学科。传统学说的信任强调对于“人”的情感,认为信任是一种基于对方品德而产生的心理依赖。这种信任侧重于建构人与人之间的深层次连接。<sup>④</sup>信任的对象是一个具备能动性、道德性的人类主体。<sup>⑤</sup>有学者对人类间信任与人机信任的异同进行了分析,认为尽管二者有所差异,但从心理学上发现二者存在同样的认知、情感与行为反应。<sup>⑥</sup>但是,如果完全按照传统标准,则人机信任不仅需要在心理上存在事实,还需要符合规范上的要求,不可避免追问机器是否可以满足个人情感、道德等方面的条件,进而得出人工智能不能作为信任主体的结论。目前,对可信任性(trustworthiness)的最新研究已经拓展至人机关系之中,某一个体或者组织的某些特性使得另一个体或者组织对其产生信任即为可信任。<sup>⑦</sup>亦即,不由人之间的感情、道德等因素而决定,端看人工智能是否可以完成预设的任务、实现预期的功能。

从目前高水平自动驾驶的技术和规范要求来看,它的可信任性极高。首先,人工智能系统本质上是具有一定自主性和能动性、基于机器的计算机信息系统。2024年3月13日获得欧洲议会批准、成为世界上第一部促进人工智能安全发展的区域性国际公约的欧盟《人工智能法》(EU AIA)采纳了该定义,并在说明部分规定“具有不同程度的自主性”是人工智能系统的基本特性,同时具有推理和影响物理的或虚拟环境的能力。<sup>⑧</sup>在技术层面,自主性特征使其区别于“更简单的传统软件系统或编程方法”和“基于仅由自然人定义的规则来自动执行操作的系统”。<sup>⑨</sup>人工智能系统虽然不同于人类的能动性,但是它能够在没有人类干预的情况下进行操作,区别于以往的计算机的工具属

① 参见[瑞]比扬·法塔赫-穆加达姆:《刑法中的创新责任:在严格责任、过失与容许风险之间》,唐志威译,载《苏州大学学报(法学版)》2022年第3期,第59-60页。

② 参见王海涛:《过失犯罪中信赖原则的适用及界限》,中国人民公安大学出版社2011年版,第68页。

③ 参见皮勇:《论自动驾驶汽车生产者的刑事责任》,载《比较法研究》2022年第1期,第64页。

④ See Mariarosaria Taddeo, “Defining trust and e-trust”, *International Journal of Technology and Human Interaction*, Vol. 5, Issue 2 (2009), p. 24.

⑤ 参见于雪:《基于机器能动性的人机交互信任构建》,载《自然辩证法研究》2022年第10期,第44页。

⑥ See P. Madhavan & D. A. Wiegmann, “Similarities and differences between human-human and human-automation trust: an integrative review”, *Theoretical Issues in Ergonomics Science*, Vol. 8, Issue 4 (2007), pp. 281-288.

⑦ 参见何丽:《人工智能可以作为置信对象吗?——为可信人工智能辩护》,载《科学学研究》2023年第10期,第1731-1733-1734页。

⑧ See EU AIA Recital (6).

⑨ 皮勇:《欧盟〈人工智能法〉中的风险防控机制及对我国的镜鉴》,载《比较法研究》2024年第4期,第70页。

性。它具备影响外在的能力,即为能动性,机器的能动性与人类并没有本质的差异。<sup>①</sup>特别是生成式人工智能中的“涌现”(Emergence),通常用来描述在一个复杂系统中某些特性或行为并不是从单个组成部分直接推导出来的,而是在多个组成部分相互作用的过程中自发产生的。这使机器摆脱了对工程师预设规则的依赖,充分说明人工智能系统的自主性和能动性。<sup>②</sup>可见,高水平自动驾驶系统可以感知环境、制定决策、响应信号、预测交通行为、自主学习等等,具备可信任性的事实特性。

其次,高水平自动驾驶系统以人类的意志为行为导向。传统观点认为机器的能动性不具备“意向性”,它只是工具或者代理人做一些事情,不能成为被信任的对象。但是,人工智能可以通过人机交互而获得“意向性”,也即将某种符合人类利益的目的作为自己的意志并予以执行。《人工智能法》(EU AIA)中特别强调了发展“以人类为中心的人工智能”,关注的核心是如何利用技术来增强人类的能力、提高生活质量、促进社会进步,同时确保技术的发展不会对人类造成伤害或不公。由此可见,通过人机互动获得“意向性”是可接受人工智能的底线,也是人机信任的基础。反过来说,如果人工智能自主地拥有与人类利益相悖的“意向性”,恐怕才是不可信的。高水平自动驾驶系统即是通过与安全员的互动获得行动目的,其系统设计亦时刻体现以人类为中心的原则。例如,《上路试点指南》规定了“最小风险策略”,紧急情况出现后如果安全员没有成功接管,此时系统依然应以人的最小风险为目标而行动。<sup>③</sup>

由此,人工智能系统具备了自主性、能动性和以人类为中心的意向性,足以成为被信任的对象。未来机器能力的通用化和实用场景的泛化,使得人机交互成为一个常态化的需求。人机交互的信任得到普遍的认可,并成为了人机交互设计的大前提。这正是人工智能时代的基石。对于人机共驾的发展而言,承认人类对于高水平自动驾驶系统的信任,并致力于维持可靠的信赖,这不仅是人机交互设计领域中不断完善的方向,也是走向完全自动驾驶阶段必不可少的要素之一。<sup>④</sup>

## (二) 人机信任的不对称性

有学者提出由于人机信任的存在,自动驾驶汽车的生产者也可以主张对于人类的信赖,从而减轻自己的责任。<sup>⑤</sup>但是,深入观察则会发现人机信任内部存在不对称性:人类应当信赖机器,而机器则应当怀疑人类。

首先,自动驾驶技术的产生根源于对人类驾驶的不信任。我国数据显示,每年数万人死于车祸,且人因占比超过90%。前十大违规行为是:未按规定让行、超速行驶、无证驾驶、醉酒驾驶、未与前车保持安全距离、逆行、违反交通信号、酒后驾驶、违法超车、违法会车。<sup>⑥</sup>人类驾驶员不仅在识别反应能力、平稳续航能力上不完美,而且会作出酒后驾驶、追逐竞驶等危险的驾驶行为。自动驾驶技术则不仅在驾驶能力上远高于人类,而且可以避免人为的违规行为。因此,以更安全和高效的自动驾驶技术取代不可靠的人类驾驶是技术发展的原动力。最新研究表明在大多数事故场景中,配备高级辅助驾驶系统的车辆发生事故的几率低于人类驾驶的车辆。例如,高级辅助驾驶系统

① 参见于雪:《基于机器能动性的人机交互信任构建》,载《自然辩证法研究》2022年第10期,第44页。

② 参见王沛然:《从控制走向训导:通用人工智能的“直觉”与治理路径》,载《东方法学》2023年第6期,第190页。

③ 《上路试点指南》第二章中规定:当安全员未能及时响应介入请求,自动驾驶系统应执行最小风险策略以达到最小风险状态。最小风险策略要求避免或减缓车辆与其他道路使用者的风险。

④ 参见[瑞]陈芳、[荷]雅克·特肯:《以人为本的智能汽车交互设计》,机械工业出版社2021年版,第188-192页。

⑤ 参见王华伟:《论人形机器人治理中的刑法归责》,载《东方法学》2024年第3期,第110页;贾济东、岳艾洁:《人工智能事故过失犯认定的中国方案:规范型塑与理论因应》,载《河北法学》2023年第10期,第78页。

⑥ 参见张尼:《每8分钟就有1人死于车祸!交通事故率最高的是这些行为》,载中国新闻网, <https://www.chinanews.com.cn/sh/2020/12-02/9352085.shtml>, 2024年7月29日访问。

车辆发生追尾事故的可能性相比人类驾驶车辆低了 6%。<sup>①</sup> 自动驾驶公司 Waymo 运营的无人驾驶出租车与人类驾驶员相比,人身事故率降低了 85%,轻微损害事故率下降了 57%。估算在相同里程下,自动驾驶系统造成的受伤人数比人类驾驶员少 17 人。<sup>②</sup> 因此,人类相信自动驾驶系统是有道理的,反之则不符合科学依据。

其次,将安全员作为高水平自动驾驶处于紧急状态下的驾驶权接管方,可能陷入“自动化自满”(automation complacency)的难题。目前,在高水平自动驾驶中存在“长尾问题”——在大量不同的驾驶场景中,那些不常见的、发生率低但影响严重的边缘情况。例如,雨雪天气、光线不好的场景可能影响摄像头的识别精度等。因此,在技术上设计由自动驾驶系统发出接管警报,将驾驶权交给作为备用的安全员。但是,这种接管设置极有可能违背人类的心理,给人类设置过于苛刻的义务。在最早进入自动化与人类共同驾驶的航天领域中,在面对复杂的任务时,飞机驾驶员可能会产生对自动化过度信任,放弃了应有的检查和控制导致飞机失事。<sup>③</sup> 这也出现在辅助驾驶、自动驾驶的应用中。美国国家公路交通安全管理局(NHTSA)于 2016 年在《联邦自动驾驶汽车政策》中指出,制造商和其他实体应高度重视评估驾驶员自满情绪和误用 L2 级系统的风险,并制定有效的应对措施,以帮助驾驶员正确驾驶汽车。最新研究表明,与其将“自动化自满”视为一种人类的心理现象来指责人类,不如将其视为系统问题(自满情绪是由个人、情境和自动化相关特征的复杂交互作用所导致的现象)。<sup>④</sup> 在美国国家运输安全委员会(NTSB)对 UBER 的调查报告亦指出“制造商没有合理监控驾驶员的状态是安全事故发生的原因之一”。<sup>⑤</sup> 基于上述因素,在自动驾驶系统中特别强调了一项必备功能:驾驶员监控系统(Driver Monitor System, DMS),通过摄像头和传感器等设备对驾驶员进行实时的监控。它可以检测驾驶员的行为、生理状态和情绪变化,以及是否疲劳驾驶或分心等情况。最新的系统可利用深度学习算法对所采集的信息进行解析,识别驾驶员的状态,并根据识别结果发出不同级别的预警,从而确保行车过程的安全性。<sup>⑥</sup> 亦即,人机共驾中系统一直保持着对安全员是否处于可接管状态的怀疑,并通过“监测+提醒”的方式以避免安全员陷入“自动化自满”之中。

最后,目前接管规则预设安全员可以在短时间内成功控制驾驶权,可能高估了人类的能力。研究认为,安全员脱离了驾驶权后,在几秒钟内紧急接管汽车,超出了正常的反应速度。<sup>⑦</sup> 一般而言,安全员脱离驾驶的时间越长,越难以在短时间内重回驾驶的状态。即便在普遍使用飞行自动驾驶软件的航空业,也难以要求一名驾驶员在短时间内接管驾驶权。“人类用户以为自动驾驶系统值得信赖,而自动驾驶系统又假设人类会随时接管,两者相互指望,徒增风险。”<sup>⑧</sup>

综上所述,在当前技术限制下虽不可避免地需要将安全员作为紧急状态下的备选方案之一(另一方案是,当安全员不能接管时系统执行最小风险原则),但自动驾驶系统对于安全员是否处于能

① See Mohamed Abdel-Aty & Shengxuan Ding, “A matched case-control analysis of autonomous vs human-driven vehicle accidents”, *Nature Communications*, Vol. 15, Issue 4931 (2024), pp. 1-12.

② See Kristofer D. Kusano, et al, “Comparison of Waymo Rider-only crash data to human benchmarks at 7.1 million miles”, arXiv: 2312.12675, 2023, pp. 8-12.

③ 参见贺青、张林英:《座舱自动化中人的因素》,载《航空学报》1999 年第 20 卷,第 61-64 页。

④ See Yueying Chu & Peng Liu, “Automation complacency on the road”, *Ergonomics*, Vol. 66, No. 11 (2023), p. 41.

⑤ See National Transportation Safety Board, “Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian”, pp. 26-27, <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>, visited on Oct. 29, 2024.

⑥ 2023 年 5 月 1 日起正式实施的推荐性国家标准《驾驶员注意力监测系统性能要求及试验方法》(GB/T 41797-2022)为驾驶员监测系统(DMS)提供了详细的技术规范和测试方法。

⑦ 参见赵申豪:《自动驾驶汽车侵权责任研究》,载《江西社会科学》2018 年第 7 期,第 212 页。

⑧ 郑志峰:《论自动驾驶汽车被动接管规则》,载《华东政法大学学报》2023 年第 3 期,第 68 页。

够接管的状态始终保持怀疑。应当通过各种方式执行驾驶员监控系统，当安全员处于分心、睡眠、生理异常等状态时，通过不同级别的警示唤醒安全员。对于制造商，“在高冲突环境等特定场景中，应当限制信赖原则适用，制造商等应合理保证自动驾驶车辆不至于酿成不必要的事故”。<sup>①</sup>

### 三、人机共驾中适用信赖原则的合理性

在认可人类可以对高水平自动驾驶系统产生信任后，对其适用信赖原则还需直面如下质疑：允许组织模式（都是加害人）内部通过信赖原则互相推脱，将会导致对被害人的保护不力。在已生效的 UBER 案判决中即出现让人忧虑的苗头，由于当时管理规范模糊等因素，生产者、制造商的刑事责任没有被认定，仅仅认定了安全员构成过失犯罪。如果适用信赖原则认定安全员没有过失责任，对被害人及其家属与大众来说可能认为刑法放弃了保护法益，进而产生对自动驾驶的抵制风潮。对此，还需要仔细分析信赖原则对于高科技致损风险结果分配的功能，并将组织模式内适用信赖原则的案例、原理与人机共驾的特殊性结合起来论证。

#### （一）信赖原则对后端风险结果分配

当我们试图追究人机共驾中接管失败造成严重损害的责任分配时，涉及自动驾驶汽车的制造方、安全员以及其他交通参与人等多方主体。学者认为，如果认可安全员对于自动驾驶系统的信任，根本上是对于驾驶系统背后的制造方的信赖。信赖原则发挥着风险结果分配的重要作用，既不能过度地强调自动驾驶汽车制造方的责任，认可安全员接管义务在高水平自动驾驶等级中依然存在必要，又不能径直以“可允许的危险”理论认可制造方带来的安全隐患，而导致安全员成为最后承担结果责任的人。<sup>②</sup>可以说，预期中信赖原则的应用可以为安全员的风险划定合理范围。

在对高科技风险进行分配时，必须要承认当前法律存在不足。尽管我国 2023 年推出了最新部门文件《上路试点指南》，从 2022 年起实施了以《深圳经济特区智能网联汽车管理条例》为代表的地方性立法，但现有规范中对于安全员的义务内容描述模糊，而且依然面临着以《道路交通安全法》为代表的全国性法律的缺位。无独有偶，在已经修改或新增国家法律的日本、德国等，对安全员注意义务的规定亦不精细。<sup>③</sup>在此情况下，非常容易错误地以汽车工程的自动化水平作为安全员注意义务的标准。德国学者希尔根多夫指出：“如果再从这些定义中推导出不同级别车辆驾驶者的注意义务，将会形成循环论证”，进而主张注意义务应该“来自于德国法律的规定，其在个案中的适用及具体化由法院进行”。<sup>④</sup>

由此可见，在前置法缺失的情况下，刑法只能面对科技风险致损的结果进行最后一次责任分配。正是从这个角度，德国学者希尔根多夫提出自动驾驶场景中过失“三阶段判断法”，将信赖原则作为自动驾驶案件中过失责任的后端限制的观点，指出了信赖原则对于人工智能产品致损发挥后端风险分配的地位。<sup>⑤</sup>对于过失责任的认定应回到法益保护的基本立场上来，在裁判时分辨违反规

① 彭文华：《自动驾驶车辆犯罪的注意义务》，载《政治与法律》2018 年第 5 期，第 97-98 页。

② 参见蔡仙：《自动驾驶中过失犯归责体系的展开》，载《比较法研究》2023 年第 4 期，第 68-69 页。

③ 关于日本和德国道路交通法修改具体内容和分析，参见黄金晶等：《自动驾驶汽车法律规范体系比较研究》，人民交通出版社 2023 年版，第 80-88、54-72 页。

④ [德] 埃里克·希尔根多夫：《数字化、人工智能和刑法》，江溯、刘畅等译，北京大学出版社 2023 年版，第 254 页。

⑤ Vgl. Eric Hilgendorf, Automatisiertes Fahren und Recht-ein Überblick, JA 2018, S. 804f. 转引自李昱：《容许风险与自动驾驶场景中的注意义务》，载《现代法学》2024 年第 4 期，第 162 页。

范与法益保护的关系，重视行为人预见可能性的判断。

## （二）组织模式中适用信赖原则的合理性

日本北大医院电手术刀案件是最为典型的案件。一名动脉管并存症的患者（2 年 4 个月）接受胸部手术时，辅助护士 A 将手术使用的电手术刀的侧面电线和对极板侧面电线接反了，主刀医生 B 在使用该电手术刀时没有仔细确认，手术刀和一并使用的心电仪的缺陷竞合，形成了高频电流的异常回路，使得患者右腿重度烫伤。虽然胸部手术成功了，但患者的右腿却不得被截肢。对此事故，实行手术的 9 人医疗团队中，使用电手术刀的主刀医生 B 和从事电手术刀接线、调整开关工作的护士 A 被以业务过失伤害罪起诉。该案件经过了两级法院审判。一审法院认为，护士 A “明知有接反的可能，并且也能知道，电手术刀是利用高频电流通过患者身体的循环过程中所产生的高热的机器，一旦误接电线，就会导致电流变更，可能对患者身体产生危害，因此，其具有正确接线、防止事故发生的业务上的注意义务”，判定其有罪（罚金 5 万元）。对于医生 B，认定“其对电线接反并不具有具体认识，而且对不能轻易地说能够具有该种认识的医生而言，他对护士所处理的接线操作的正确与否，具有进行二次检查、确认的注意义务，并将其作为肯定刑事责任的条件的理解，这种要求是过于慎重的态度，并不妥当”，判定其无罪。<sup>①</sup> 检察院针对一审法院对 B 的判决提出抗诉，二审法院支持了原判决结论。

一二审法院认为 B 不构成过失的理由正是信赖原则。“在手术开始之前，信赖作为资深护士的被告人，对是否误接没有进行检查，这一点从当时的情况来看，也是正常的……”<sup>②</sup> 该判决招来了质疑。信赖原则最初是适用在被害人和加害人间的对向型过失中，只有在被害人也存在过错的场合，才能例外地限制行为人的过失成立。这体现了危险分配的思想，为了回避危险，在加害人和被害人之间分担危险，分配注意义务的话，就会减轻加害人的过失。<sup>③</sup> 在组织模式中被害人完全是无过错的，却也主张信赖原则，可能导致所有的加害人均不受到惩罚。<sup>④</sup>

针对上述质疑的理论回应是：其一，无论是交通中对向型过失还是组织内部的对向型过失，可以适用信赖原则最为重要的共同点是社会分工。如果不承认组织内部成员之间的信赖关系，那就从实质上否认了社会分工的基础事实。<sup>⑤</sup> 信赖原则随着现代社会风险活动的日常化而生。虽然对个人科以严格的注意义务看上去能够最大限度地防范损害结果的发生，但实际上却会带来社会活动的低效性，甚至抵消高度风险行为带来的社会受益。例如，过分苛责作为监督者的医生，要求医生事事亲力亲为不仅带来工作内容的紊乱，而且导致效率低下，最终受到损害的是更多亟须救治的病人。因此，在现代社会分工是不可避免的。与交通事故中陌生人之间的信赖相比，在组织内部进行的分工，由于具有职责范围和专业场景，具有更为切实的信任基础。其二，无辜的受害人，并不是加害方必须承担刑事责任的充分理由。<sup>⑥</sup> 虽然，在医疗中无论是医生还是护士都负有防止结果发生的义务，但是无论被害人有没有过失，都不直接影响医生、护士之间基于分工存在的信赖关系。

① 参见日本札幌高等裁判所 1976 年 3 月 18 日判决，载日本《高等裁判所刑事判例集》第 29 卷 1 号，第 78 页。

② 日本札幌高等裁判所 1976 年 3 月 18 日判决，载日本《高等裁判所刑事判例集》第 29 卷 1 号，第 78 页。

③ 参见〔日〕大谷实：《刑法讲义总论》（新版第 5 版），黎宏、姚培培译，中国人民大学出版社 2023 年版，第 193 页。

④ 参见〔日〕神山敏雄：《信赖の原則の限界に関する考察》，载西原春夫先生古稀祝賀論文集編集委員会：《西原春夫先生古稀祝賀論文集：第 2 卷》，成文堂 1998 年版，第 64 页。

⑤ 参见张佳宇：《チーム医療における刑事過失責任：組織的医療における個人の過失責任のあり方及び関係者間の責任分担》，日本北海道大学博士论文（法学）甲第 15120 号，第 117-118 页。

⑥ 参见〔日〕町野朔：《刑法総論》，信山社 2019 年版，第 300 页。



### （三）延展至人机共驾中适用的合理性

信赖原则的适用可以从普通的组织体延展至人机共驾这一特殊组织体的理由有三。第一，人机共驾体现了高水平的人类与人工智能的分工，存在更为可靠的信赖关系。如前所述，自动驾驶系统具备超越人类的能力。例如，相较于人的视野盲区而言，自动驾驶系统具有增强的感知能力。它所配备的高级传感器、摄像头、雷达可以 360 度感知周围环境，监测到人类驾驶员可能忽视的盲点或远处的障碍物。并且，根据《上路试点指南》的要求，我国现有商业试点中的高水平自动驾驶汽车都已经通过了模拟仿真、封闭场地、实际道路的测试并通过验收。换言之，这些高水平自动驾驶系统的预期功能是由算法决定并经历了千万次的测验。与医生信赖一名资深的护士的简单操作相比，安全员信任高水平自动驾驶系统具有更切实的基础。由此，基于人机互动中获得的信任，安全员将驾驶权交给自动驾驶系统，符合未来交通追求的安全与高效。

第二，被害人的保护必要性并不足以说明安全员承担刑事责任是最好的解决办法。德国学者曾预测，司法判决在今后的一段较长时间内都会倾向于对所有级别的自动驾驶车辆的驾驶人、安全员提出较高的监控要求。然而，这将导致自动驾驶的“控制权两难困境”（Kontrolldilemma）：一方面，车辆装配自动驾驶系统的目的是将驾驶者从驾驶任务中解脱出来；另一方面，驾驶者却有义务始终监控系统，随时准备介入、修正驾驶操作。这样一来，从消费者的角度看来自动驾驶系统的效益被极大限制了。<sup>①</sup> 试想，普通人要享受科技带来的便利，却不仅需要负担高昂的费用，还需要违背心理机制承担时刻监控系统的严苛义务。而这种技术发展的最大受益人——生产者、销售商却隐身了。这种对安全员过于严苛的做法，可能只是无法通过现有的法律和法理对生产者、销售者追究刑事责任的“遮羞布”。换言之，在高水平自动驾驶的接管规则中，将安全员摆在备用驾驶员的位置上，原本是一种技术上不公平的风险预分配。学者对目前僵化地以汽车工程自动化水平标准的内容作为刑事责任判断标准的方法提出了诸多批评。核心一点是：人机共驾中规定“始终监控和随时接管”加重安全员的责任，既违背了自动驾驶技术的初衷，最终亦不利于该技术的发展。<sup>②</sup> 即便承认某些人机共驾案件中最终无人承担刑事责任，也不代表被害人的法益损害不能通过其他方式得到保护。UBER 案件中，被害人家属与 UBER 公司达成了民事和解。UBER 公司对自动驾驶部门的安全漏洞进行了改革。<sup>③</sup> 因此，不能认为法益损害必然带来刑事归责。合理的解决方案，应当是更加严格地判断信赖是否已经具备相当性，而非否定信赖原则在该领域的应用。<sup>④</sup>

第三，可能有观点认为即便组织体内可以适用信赖原则，也应当排斥危险程度高的业务。有学者认为工作性质的风险程度越高，注意义务的范围就越广，相应地，信赖原则的适用范围就越窄。<sup>⑤</sup> 这样一来，人机共驾涉及公共交通安全，可谓是危险程度极高的领域。但是，从现在人工智能技术的发展来看，自动驾驶汽车实为第二代智能机器人，具有部分感知能力、决策能力。未来第三代智能机器人则具备更强的感知能力、决策能力，譬如人形机器人。可以预计机器人未来将承担更精细、更复杂的工作，例如家政照护机器人。<sup>⑥</sup> 如果认为人类在使用第二代智能机器人时注意义务很

① 参见〔德〕埃里克·希尔根多夫：《数字化、人工智能和刑法》，江溯、刘畅等译，北京大学出版社 2023 年版，第 255 页。

② 参见郑志峰：《论自动驾驶汽车被动接管规则》，载《华东政法大学学报》2023 年第 3 期，第 68-71 页。

③ See Alexandra DeArma, “The Wild, Wild West: A Case Study of Self-Driving Vehicle Testing in Arizona”, *Arizona Law Review*, Vol. 61, Issue 4 (2019), pp. 1005, 1009.

④ 参见〔日〕塩谷毅：《信赖の原則に関する序論的考察》，载齐藤丰治等编：《神山敏雄先生古稀祝賀论文集：第 1 卷》，成文堂 2006 年版，第 104-106 页。

⑤ 参见〔日〕大谷实：《危険の分配と信頼の原則》，载藤木英雄编：《過失犯：新旧過失論争》，学陽書房 1975 年版，第 124-125 页。

⑥ 参见王华伟：《论人形机器人治理中的刑法归责》，载《东方法学》2024 年第 3 期，第 101-102 页。

广、信赖原则适用极窄，那么第三代智能机器人如何能够脱离人类独立发挥作用呢？对此问题，可能需要反过来思考，“因为通常在越是危险时，越要求分工人具有高度的专门知识、能力。”<sup>①</sup> 在危险的业务中，参与人的注意义务当然应该严苛。不过，首要加强的是对于高水平人工智能的技术规范和测试要求。同时，随着智能机器人能力的提高，相应是使用人注意义务的缩小，信赖原则具有更多的适用可能性。

总之，随着自动驾驶水平从 L3 到 L4 级的提升，安全员的义务逐渐缩小。根据《汽车驾驶自动化分级》（GB/T 40429-2021）的规定，L3 级别安全员需要“始终监控和随时接管”“保持警觉”；L4 级别时，仅仅在极端情况下系统才会要求安全员接管，即使安全员没有接管，L4 级别的自动驾驶系统也会执行最小风险策略以确保车辆能够继续安全行驶或停车。可见，风险分配越来越倾向于可靠的高水平自动驾驶系统。从自动驾驶技术的获益、操控主体来看，人机共驾一方造成的结果责任最后分配给许诺给人类信任的自动驾驶汽车制造方，是公平的。<sup>②</sup> 信赖原则的合理适用，既不会导致个体安全员因过重责任而抵制自动驾驶汽车，又将提高自动驾驶汽车技术安全的压力赋予科技创新和获益的主体，还给社会大众创造了人工智能应用的可靠环境，可谓通过对后端风险结果分配而一举三得。

#### 四、具体应用：信赖的相当性与预见可能性

信赖原则适用范围并不局限于交通犯罪领域，也不局限于行为人与受害人的对向关系，而是可能适用于各种风险预防的合作关系。需要说明的是，即使允许在组织体模式中适用信赖原则，也不是无限制的，必须考虑其限度和标准。一般认为信赖原则作为限缩过失的例外原则，应慎重运用。人机共驾中什么是可靠的信任成为值得探讨的问题。一方面，过度信任带来的技术滥用恐怕是当前人机共驾最为头疼的问题。在已经广泛应用的辅助驾驶导致的交通事故中，出现了驾驶人对于辅助驾驶系统过度信任的问题。<sup>③</sup> 另一方面，信任不足则会衍生技术弃用。在人机共驾中如果过度强调安全员的监督义务，则会导致对于自动驾驶技术可靠性、效益性的怀疑，不利于技术的发展。持肯定说的学者仅原则性表示可以考虑合理基础、未超信赖限度以及未实施不当行为三个因素。<sup>④</sup> 在这一方面需要结合既有研究与人机协作的特点作进一步思考。

##### （一）人机信赖的相当性判断

有学者认为不能将信赖原则仅作为一种单纯的主观信赖，而应该强调存在与回避结果相当的实质信赖关系。“只有在具有与该法律顾虑相当的实质信赖关系（迈向结果回避的日常信赖的积累）的情形中，才明显能在心理上缓和结果回避的动机，由此可以判断，监督人无法对结果的发生进行具体预测。”<sup>⑤</sup> 以医疗团队为例，一般认为需要考虑：（1）医疗分工体制的建立亦即医务人员权责的明确化；（2）信赖对象具备能够信赖的现实基础，可以具体化为履行医疗任务的资质、知识、

① [日] 甲斐克则：《责任原理与过失犯》，谢佳君译，中国政法大学出版社 2016 年版，第 93 页。

② 参见李昱：《容许风险与自动驾驶场景中的注意义务》，载《现代法学》2024 年第 4 期，第 166-167 页。

③ 参见 [日] 日原拓哉：《AI の活用と刑法》，成文堂 2023 年版，第 46 页。

④ 参见蔡仙：《人机共驾模式下的接管义务及其刑事归责》，载《苏州大学学报（法学版）》2023 年第 3 期，第 37 页。

⑤ [日] 甲斐克则：《责任原理与过失犯》，谢佳君译，中国政法大学出版社 2016 年版，第 95 页。

经验和态度；(3) 不存在动摇信赖的例外事由。<sup>①</sup> 在上述日本北海道大学电手术刀案件中，学者认为：“主刀医生 B 是治疗科的，护士 A 是手术部的，对电手术刀线路连接的事情没有建立清楚的合作机制。但是，在本案医院中，设立了从各个科室中独立出来的手术部，手术部的管理、手术器材、器具的准备、护士的教育等业务，都由手术部负责，A 是在手术部工作的正规的资深护士；接线是非常简单容易的工作，其方法不要求按照医生的要求进行，有资格的护士犯接错了线路之类的错误是难以想象的；电手术刀引起的本案之类的事故，世界医疗史上前所未有。因此，可以说本案中医疗从业人员使用电手术刀引起事故，不可思议。主刀医生必须集中精力做手术。综合上述考虑，可以说，主刀医生不用检查接线正确与否，这种事实上的行为基准具有大致的合理性。”<sup>②</sup> 从而判断医生 B 的信赖是具有相当性的。

研究认为，在人与人的信赖中，要考虑对他人的熟悉度、期望值和风险。熟悉度指向信任主体间的关系构建，期望值指向预期完成任务的程度，风险则指向决策失败的机会成本。在人机信任中则有所差异，影响因素均可划分为三个类别：用户特性、机器特性和环境特性。用户特性中主要考虑的是人类主体情况。机器特性则是信赖对象的性质，包括可靠性、可预测性、可解释性等。环境特性则是指当面对低复杂度、低风险、低试错成本的常规性质任务时，人类往往会展现出较高的信任度；相反，在面对那些要求精确控制、专业知识和高度判断力的复杂任务时，人类的信任水平可能会显著降低。三个因素的重要性有所不同，机器特性带来的负面效果最为明显，其次为用户特性，而环境特性的影响力相对最低。同时，环境特性作为一种外部调节因素，对持续信任的影响相对最大。尤其是关键任务失败会显著降低用户信任水平，甚至在后续交互使用中形成应激障碍。<sup>③</sup> 这为如何在人机共驾的具体环境中确定可靠的信赖提供了线索。

从人机协作的特性来看，基本可以从以下三个方面分析信赖是否具有相当性：安全员是否具备充分的专业能力、心理水平；自动驾驶系统是否具备可靠性、可预测性和可解释性；在具体的人机交互之中，执行关键任务是否发生严重的失败，足以使得安全员对于自动驾驶系统运行产生怀疑。三个方面对于信赖的影响方向是：安全员因素和自动驾驶系统因素，对信赖的相当性成立是有利因素；严重失败的交互任务则是动摇信赖的不利因素。下面结合《上路试点指南》等规范性文件和 Q 案件的具体情况，进一步分析在不同级别的高水平自动驾驶场景中的具体应用。

### 1. 安全员因素

《上路试点指南》第三部分上路通行中对安全员的規定可分为两部分：驾驶资格与专业培训。首先，安全员应当是一个经验丰富、记录良好的优质驾驶人。不能有负面的驾驶记录，比如严重违规、重大事故等。其次，安全员应当经过自动驾驶汽车的专业培训，熟练掌握自动驾驶相关法律法规、自动驾驶系统专业知识，具备紧急状态下的应急处置能力。在 Q 案中，安全员王某取得了资格，具备《上路试点指南》中的基本条件，即符合专业知识、特殊培训等能力条件。再如，在 UBER 案件中，R 持有驾驶执照，自从 2017 年以来一直担任 UBER 自动驾驶测试车辆的安全驾驶员。上路测试前，已经完成了为期三周的培训以及复训。在完成培训后，她曾完成了 73 次自动驾驶道路测试，包括在事故道路上的自动驾驶测试。该安全驾驶员在工作期间没有受到过任何处罚，并且曾受到过集体绩效奖励；没有严重的交通违规记录。可见安全员 R 的自身情况有利于信赖相当

<sup>①</sup> 参见王海涛：《过失犯罪中信赖原则的适用及界限》，中国人民公安大学出版社 2011 年版，第 238-240 页。

<sup>②</sup> [日] 大塚裕史：《予見可能性の意義（1）》，载山口厚、佐伯仁志编：《刑法判例百選 I 総論》（第 7 版），有斐閣 2014 年版，第 105 页。

<sup>③</sup> 参见向安玲：《无以信，何以立：人机交互中的可持续信任机制》，载《未来传播》2024 年第 2 期，第 30-31 页。

性的成立。<sup>①</sup>反之，当一名安全员不具备相应专业能力或者处于明显异常的心理状态下，则难以认为其信赖是相当的。

《上路试点指南》的规定没有区分 L3 和 L4 级别安全员的差异。结合 2023 年交通运输部《自动驾驶汽车运输安全服务指南（试行）》以及《北京市智能网联汽车政策先行区乘用车无人化道路测试与示范应用管理实施细则》的规定：L4 级别自动驾驶汽车可以配备车上安全员，也可以是平台安全员，因此出现“一比多”的平台安全员配置。由于客观上平台安全员的注意力和接管能力是有限的。《自动驾驶汽车运输安全服务指南（试行）》规定的 1:3 的人机比例可作为参考。当超出合理的平台安全员配置比时，可以认为安全员处于能力显著低于标准的情形中，其主张的信赖难以成立。

## 2. 自动驾驶系统因素

《上路试点指南》第二章智能网联汽车产品规定：“应具有明确的自动驾驶功能定义及其设计运行条件，并符合动态驾驶任务执行、接管、最小风险策略、人机交互……等技术要求”。其中，对动态驾驶任务的可靠性规定：具备探测与响应能力，以支持其安全且合理地执行全部动态驾驶任务、识别系统失效的能力、失效后合理控制的能力、激活后最低损害控制策略。亦即，对于现在的自动驾驶系统而言，应当可以执行前碰撞预警（Forward Collision Warning, FCW）、自动紧急制动（Automatic Emergency Braking, AEB）、自适应巡航控制（Adaptive Cruise Control, ACC）、车道保持辅助（Lane Keeping Assist, LKA）、盲点监测（Blind Spot Monitoring, BSM）等功能。除此之外，《上路试点指南》规定了自动驾驶系统应具备安全员接管能力监测功能（DMS）。它特别针对 L4 级别自动驾驶系统规定，当安全员未能及时响应接管请求，自动驾驶系统应执行最小风险策略以达到最小风险状态；自动驾驶系统应当持续提供关于驾驶状态的必要信息。上述关于自动驾驶汽车产品可靠性的要求仅是该类高水平自动驾驶产品的底线。安全员在没有异常事件发生时，当然可以相信上述功效在正常运作中，自动驾驶系统所作决策是可预测和可解释的。

在 Q 案件中，Q 公司测试团队为测试的流畅性，避免可能的干扰对测试产生影响，关闭了 L3 级别测试车安装的前向碰撞预警（FCW）和紧急制动功能（AEB）。也就是说，这辆改造过的 Q-1 自动驾驶汽车并不具备 L3 级别完整的功能以完成动态驾驶任务。同时，该系统在 5 秒钟前已经识别到障碍物，却没有准确识别为行人，这一关键性的错误可能导致自动系统无法作出正确决策。如果安全员王某事先未获知该车辆自动系统功能的限制，则他主张信任自动驾驶系统具备全部的功能足以完成动态驾驶任务，这种信赖是具有相当性的。

## 3. 人机交互因素

人机交互中如果出现严重失败的情况，则会削弱人对于机器的信任；反之，若人机交互完成重要任务是顺畅而熟练的，则会增强人对于机器的信任。可见，人机交互任务的难易程度、完成次数、完成效果都是重要因素。在目前的道路测试阶段，发现已有自动驾驶系统执行任务的异常、失败是测试最重要的目的之一。《上路试点指南》规定：在自动驾驶系统激活期间，记录的事件数据应至少包括自动驾驶系统激活、退出、发出介入请求、开始执行最小风险策略、发生严重失效、有碰撞风险、发生碰撞等。若测试车辆已经发生上述异常或者失败，将动摇安全员的信任程度。失败越严重，就越容易引发信任的否定。例如，安全员在同一个测试路段，已经发生了 2 次车前无障

<sup>①</sup> See National Transportation Safety Board, “Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian”, p. 23, <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>, visited on Oct. 29, 2024.

碍物却引发自动驾驶系统的前碰撞预警和自动紧急制动，导致车辆在高速运行中突然刹车。在第3次通过该路段时，无论该系统是否对此问题进行调试，安全员不可以主张由于信任车辆的自动驾驶功能而不能预计到车辆有突然刹车的风险。

在Q案件中，没有记录事件的证据证明安全员王某在此前的测试中（甚至是同一路段已经测试20次）发生过严重的任务失败，特别是L3级别自动驾驶系统最为关键的识别、响应系统的存在问题。在事发时，自动驾驶系统已经平稳运行19分钟。自动驾驶系统执行的任务是较为简单的任务，已完成测试多次且效果较好，因此，安全员王某基于自身在人机协作中获得的经验认为测试车辆的识别、响应系统是正常的、可信的。

对于L4级的平台安全员来说，人机交互殊为关键。一方面，车上安全员可以更迅速地获知车辆面临的危险状况，而平台安全员依赖于自动驾驶汽车对当前驾驶状态、紧急情况的感知与提示；另一方面，在L4级平台安全员作出接管决策时，自动驾驶系统扮演特殊的角色。例如，德国《道路交通安全法》对于L4级自动驾驶设定了技术监督员角色。在第1e条第3项和第4项规定，“如果出现必须违反道路交通安全法规的情况，车辆首先要停车，向技术监督员提供备选的驾驶动作作为解决方案，然后等待技术监督员的决定。”<sup>①</sup>这要求L4级的自动驾驶系统不仅能准确感知当前危险，而且在具体危险评估之后作出合理的备选方案。因此，如果某L4级自动驾驶系统已经出现危险评估与备选方案的严重失败或者显然的缺陷，平台安全员通过人机交互发现了这些动摇信任的因素，那么不可以主张信任自动驾驶系统提供的备选方案采取的驾驶动作、不能预计到损害结果的发生；反之，如果平台安全员由于人机交互中累积的信任因素，虽然没有慎重地重新考量，而是选择了自动驾驶系统提供的方案采取驾驶动作，却造成了损害后果，此时平台安全员对机器的信赖是相当的。

## （二）安全员的预见可能性

行文至此，一个疑惑会浮上心头：信赖原则在过失犯罪阶层中处于什么位置？对此原来主要存在两派学说：行为构成阶层的结果回避义务认定说，认为由于信赖原则行为人的行为没有违背注意义务，不具有实质的危险性。<sup>②</sup>责任阶层的预见可能性标准说，行为人因为信赖原则而丧失预见可能性或者预见可能性的程度较低。<sup>③</sup>

现在新出现了双重体系定位说，“客观上具备相当性的信赖，阻却行为的构成要件符合性；客观上具有危险性，但行为人主观上有合理的信赖，可以阻却责任”。<sup>④</sup>例如，在封闭的高速上驾驶车辆的人，合理信赖行人不会横穿马路，行人突然横穿马路被车撞死的，不能认为行为人的行为符合交通肇事罪的构成要件。这时客观上存在合理信赖他人的条件，行为人的行为缺乏危险性。<sup>⑤</sup>

对于人机共驾问题而言，由于当前从L3级到L4级的高水平自动驾驶技术在全世界范围内均处于从道路测试到商业试点的阶段，从技术的可靠程度上仍然存在“长尾”场景。所以，不能说安全员未“始终监控和随时接管”的行为客观上没有危险性。从这一点看，位于构成要件阶层的结果回避义务说存在问题。<sup>⑥</sup>特别是在Q案中由于公司擅自改变了测试使用的L3级别自动驾驶系统，或者系统发生障碍的情况下，安全员王某基于自己的认知信任自动驾驶系统、不进行监督的行为的确

① [德] 埃里克·希尔根多夫：《数字化、人工智能和刑法》，江溯、刘畅等译，北京大学出版社2023年版，第347页。

② 参见[日] 甲斐克则：《责任原理与过失犯》，谢佳君译，中国政法大学出版社2016年版，第93页。

③ 参见[日] 西田典之：《日本刑法总论》（第2版），王昭武、刘明祥译，法律出版社2013年版，第245页。

④ 王海涛：《过失犯罪中信赖原则的适用及界限》，中国人民公安大学出版社2011年版，第147页。

⑤ 参见张明楷：《刑法学》（第6版），法律出版社2021年版，第389页。

⑥ 参见付玉明：《自动驾驶汽车事故的刑事归责与教义展开》，载《法学》2020年第9期，第140-141页。

在客观上具有高度危险性。这一不法性的判断尤为重要，既要符合目前人机共驾的科学标准，也应与现阶段社会大众对人机共驾的认识吻合，所以认为安全员不履行监督义务、接管责任的行为具有客观违法性更为妥当。此时，只能考虑从安全员的主观责任层面适用信赖原则。换言之，信赖原则的适用，并非认为安全员违反现在通行的接管规则是没有违法性的，也不是在过失犯的违法阻却事由中进行利益权衡的判断，仅仅是作为一个例外的理由否定具体案件中存在合理信赖的安全员的预见可能性。这也符合本文主张信赖原则作为后端结果责任分配的本意。

Q 案中，安全员王某对于自动驾驶系统具备合理信任，认为在当前熟悉道路、正常路况且自动驾驶系统运行平稳的情况下，自动驾驶系统有多种功能避免结果的发生，对由此引发的撞击行人的具体结果不具有预见性，在心理上难以形成反对动机。由此，在当前的人机共驾阶段，应当在安全员的主观责任层面适用信赖原则，否认过失责任。未来，随着自动驾驶技术的日趋成熟，那么安全员不监督的客观危险性也可能将随之降低。

最后，在适用信赖原则时，传统理论认为还有一个附加条件是：“违背注意义务者，就不能辩称他信赖其他人还会实施合乎注意义务的行为。”“如果两个参与者相继通过违反义务的行为造成了同一损害结果，那么，前行为人也不能以先来后行为人会遵守注意义务为由，为自己推卸责任。”<sup>①</sup>这也被称之为“清白”（clean hand）条件。在 Q 案件中，一种反对信赖原则适用的观点可能是：安全员王某违反“始终监控和随时接管”的义务在先，不可主张信任自动驾驶系统能够完成避免碰撞等功能。

不过，“清白”条件并非绝对。即便在行为人自己违法的时候，符合下述条件也可以主张适用信赖原则：“不管自己是否违反规则，都能期待对方采取适当行为的场合。”<sup>②</sup>日本最高法院在 1967 年 10 月 13 日的判决中，认可违反交通法规的被告人亦可以适用信赖原则，而得出无罪判决。“作为电动自行车的驾驶人，就像本案被告一样，在发出右转信号的情况下，准备从中线稍偏左处右转，只要相信后面驶来的另一辆车的驾驶人会遵守交通法规，即以安全的速度和方法行驶，如减速并等待自己的车辆右转，这就足够了。驾驶人只要相信其他机动车驾驶人会遵守交通法规，减速慢行、等待车辆右转等，并以安全的速度和方式行驶即可。”因此法院认为，被告人右拐之时未将车辆靠近左边的行为，虽然没有遵守交通法规，依然可以信赖被害人的车遵守交规不超越中线、避免事故的发生。<sup>③</sup>对此，学者认为该判决正确理解了信赖原则。如果行为人的违反规范的行为诱发了对方的违反规范的行为，那么可以否认信赖原则的适用。“在上述案件中，被告人没有确认右后方安全的行为，并不能引发另一辆车以极其危险的方式驾驶。因此，被告人行为的危险性，在判断危险实现的时候，不能将另一辆车的违反规范行为纳入判断之中。亦即，可以适用信赖原则。”在考虑信赖原则时“将重点放在对方应遵守规范的行为，而不是行为人的行为”。<sup>④</sup>

本文认为围绕信赖原则的讨论，最后应当落脚到行为人的预见可能性上，清白条件亦应如是。行为人自身的违法行为不一定能影响对于结果的具体预见。如果行为人的行为可以引发对方或第三人的违法行为，当然行为人具有对由此引发结果的具体预见性。回到 Q 案中，一方面，正如上述学说主张的，即便认为安全员王某没有尽到“始终监控和随时接管”的义务，但自动驾驶系统的异常

① [德] 英格博格·普珀：《德国刑法总论：以判例为鉴》（第 4 版），徐凌波、喻浩东译，北京大学出版社 2023 年版，第 76 页、82 页。

② 黎宏：《刑法学总论》（第 2 版），法律出版社 2016 年版，第 198 页。

③ 参见日本最高裁判所第二小法庭 1967 年 10 月 13 日判决，载日本《最高法院刑事判例集》21 卷 8 号，第 1097 页。

④ [日] 深町晋也：《注意義務の存否・内容（1）—信頼の原則》，载佐伯仁志、桥爪隆编：《刑法判例百選 I 総論》（第 8 版），有斐閣 2020 年版，第 111 页。

操作并非由此而引发的，所以安全员王某依然可以信任激活中的自动驾驶系统正常运行。另一方面，安全员王某虽然没有尽到“始终监控和随时接管”的义务，但并不可以用此时所谓的抽象的“危险感觉”来代替具体预见的判断。安全员王某之所以没有预见到具体危险结果的发生，说到底正是源自对自动驾驶系统执行任务的信任。在Q案件中，如果Q-1自动驾驶系统根据《上路试点指南》获得了测试牌照，那么至少可能作出以下四种决策避免结果的发生：（1）监测发现安全员处于长时间分神的状态，警示安全员；（2）准确识别过路的行人、预测其行走路线，采取避让或者刹车的策略；（3）准确识别过路的行人、预测其行走路线，并识别处于紧急情况，向安全员发出接管请求；（4）向安全员发出接管请求，但在安全员接管失败后，采取最小危险策略。而且，从Q案件的具体情况可见，事发道路、天气等情况并非复杂的场景，应属于该系统曾多次测试通过的普通场景，系统仅仅需要执行非常简单的刹车或者提醒的功能即可。如果上述道路交通案件中行为人可以合理信任一名随机而来的驾驶人能够遵守交通规范，那么对于安全员王某而言，面对已经熟悉的Q-1自动驾驶系统，不管自己是否违反规则都能期待自动驾驶系统，这是非常合理的。总之，尽管安全员王某没有尽到“始终监控和随时接管”的义务在先，但是也不妨碍信赖原则的适用。

对Q案件中安全员王某的刑事责任进行总结：王某在自动驾驶系统激活后长时间注视手机、没有监督自动驾驶系统的行为确实违反了L3级别人机共驾的接管义务，该行为与被害人的死亡之间有因果关系。但是由于安全员王某并不知道自动驾驶系统的修改，通过已有人机交互经验获得对自动驾驶系统的合理信任，充分信赖在当时的路况条件下自动驾驶系统的多种功能足以避免碰撞事故的发生，王某对撞人事故缺乏具体预见可能性。所以，安全员王某不承担过失责任。

## 结 语

目前，我国虽然通过《上路试点指南》对于高水平自动驾驶汽车的上路规则作了进一步明确，但这仅是部门规范性文件。涉及刑事责任判断时，安全员的过失行为内涵依然缺乏法律的明确规定。<sup>①</sup>对此，当然不能仅以目前国家标准的自动化水平划分为准，也不能完全依赖规范性文件的内容。在讨论普通过失致人死亡罪或者交通肇事罪时，还需从人机共驾中安全员所欠缺的重要的谨慎行为来实质性地思考。

在Q案件中安全员王某的过失责任经由信赖原则被合理地否定之后，或许我们才可以真正直面高水平自动化驾驶技术“打怪升级”所付出的代价——公共和个人法益遭受损害而刑事责任落入真空。被害人、安全员和普通大众在这场浩浩荡荡的人工智能变革中都是弱小的个体，目前由高水平人工智能产品带来巨大的商业利益尚未惠及个人。自动驾驶汽车的制造方——研发者、生产者、销售商才是技术可靠性的第一责任人。从这些角度来看，将来在对人机共驾接管规则的立法考量中，应当加强的是自动驾驶系统制造方对于安全员顺利接管的保障技术要求与责任。

（责任编辑：陈 璇）

<sup>①</sup> 参见王莹：《自动驾驶法律准入问题研究：路线、挑战与方案》，载《中国人民大学学报》2021年第6期，第146-152页。