



2024 车用人工智能 风险管理白皮书

AUTOMOTIVE ARTIFICIAL INTELLIGENCE
RISK MANAGEMENT

智能汽车安全技术全国重点实验室

编写人员： 李昌、刘巍、李庆庆、吴天舒、文治宇、王潼、贾先锋、
王红、刘家欣、杜相龙、江婉榕、庞帅、龚锋、李珈瑶、
常婉渲、谢天瀛、查仁方、郝璐璐、马超、王海均、
陈玥蓉、杨颖青

主编单位： 智能汽车安全技术全国重点实验室
重庆长安汽车股份有限公司
中国汽车工程研究院股份有限公司

参编单位： 北京理想汽车有限公司
广州小鹏汽车科技有限公司
中汽数据（天津）有限公司
清华大学

CATALOGUE

目 录

前言 /01

第一章 人工智能风险管理现状与产业实践分析 /02

- (一) 车用人工智能应用现状 /02
- (二) 车用人工智能安全挑战 /04
- (三) 人工智能风险治理现状 /06
- (四) 人工智能风险管理产业实践方案 /09

第二章 车用人工智能安全体系框架 /12

- (一) 体系概述 /12
- (二) 风险分析框架 /13
- (三) 关系 /15

第三章 车用人工智能风险分析 /16

- (一) 内生安全 /16
- (二) 应用安全 /21
- (三) 车用人工智能风险归纳 /22

第四章 车用人工智能风险应对技术方案 /24

- (一) 数据 /24
- (二) 模型 /27
- (三) 环境 /32

第五章 车用人工智能风险管理方案 /37

- (一) 车用人工智能风险管理方案 /37
- (二) 人员管理 /38
- (三) 技术管理 /42
- (四) 供应链管理 /51

第六章 总结展望 /54

参考文献 /55

前言

随着人工智能（Artificial Intelligence, AI）技术的迅猛发展，诸多行业正发生着前所未有的变化，特别是在汽车行业。从智慧座舱、智慧控制到智能驾驶，人工智能的应用不仅极大地提升了驾驶的安全性和舒适性，还为汽车制造商带来了前所未有的商业机遇。然而，任何技术的发展都伴随着风险的产生，人工智能技术在汽车行业中的发展也不例外。数据安全、隐私保护、系统可靠性、伦理道德等问题日益凸显，成为了制约行业健康发展的关键因素。本白皮书旨在全面分析汽车行业中人工智能技术带来的主要风险，并提出系统化的应对策略和技术解决方案。我们希望通过本白皮书的发布，能够引起行业内对外对车用人工智能风险管理的高度重视，促进相关法律法规的完善，推动技术标准的建立，最终实现汽车行业与人工智能技术的和谐共生，为构建更加安全、高效、智能的未来交通环境贡献力量。

第一章 人工智能风险管理现状与产业实践分析

为深入分析人工智能技术所带来的一系列安全问题，本章首先探讨了汽车行业人工智能技术的发展与应用现状，并分析了其在推动行业转型中的关键作用及面临的安全挑战。然后，对全球主要国家和地区在人工智能法律法规、政策及标准制定方面的进展进行了横向对比，梳理了不同区域在人工智能安全治理上的推进情况。最后，通过借鉴金融、医疗等行业在人工智能风险管理中的成功经验，为汽车行业构建系统化、可持续的人工智能风险应对策略提供了有益参考。

（一）车用人工智能应用现状

车用人工智能技术是指利用人工智能的各种方法和技术，来改善汽车的驾驶、交互和控制等各个环节的技术总称。这些技术的应用旨在提高汽车的安全性、便利性、舒适性和环保性，同时为用户提供更加个性化和智能化的服务。

在智能驾驶、智慧座舱和智慧控制三大领域，人工智能的应用不仅极大地提升了驾驶的安全性和舒适性，还为未来交通出行描绘了一幅全新的图景。

在智能驾驶方面，通过人工智能技术的应用，使汽车能够实现环境感知、路径规划、决策控制等功能。例如，通过安装在车辆上的各种传感器，如摄像头（Camera）、激光雷达（Lidar）、毫米波雷达（Radar）等，人工智能系统可以实时收集车辆周围环境信息，准确识别行人、其他车辆、交通标志等障碍物信息。基于这些信息，人工

智能系统能够做出快速而准确的判断，如自动调整车速、变道、停车等，从而有效避免交通事故的发生，提高道路安全性。此外，随着技术的不断成熟，完全自动驾驶（L4级及以上）正在逐步成为现实，这将彻底改变人们的出行方式，减少交通拥堵，降低环境污染。

在智慧座舱方面，人工智能技术在提升驾乘体验方面体现出重要作用。通过收集和分析驾驶者的偏好、习惯等数据，智慧座舱能够提供个性化的服务，如自动调节座椅位置、空调温度、音响系统等，创造更加舒适的驾驶环境。同时，集成的自然语言处理技术使得人机交互更加自然流畅，驾驶者可以通过简单的语音指令完成导航设置、拨打电话、播放音乐等操作，极大地提高了驾驶的便利性和安全性。此外，智慧座舱还能够通过监测驾驶者的生理指标（如心率、疲劳程度等）来评估其驾驶状态，必要时发出警报并采取干预措施，进一步保障驾驶安全。

在智慧控制领域，人工智能技术提升了车辆能源管理效能、远程监控水平及设备互联能力。具体表现在：一是通过人工智能技术精细化管理新能源汽车三电系统，实现节能高效；二是利用人工智能技术在云端全方位监控车辆状态，提高安全性和可靠性；三是通过人工智能技术赋能车机与智能设备互联，构建一体化智能出行生态系统。这些应用彰显了人工智能技术在智能控制中的创新，为用户带来便捷的智能出行体验。

（二）车用人工智能安全挑战

随着人工智能技术在汽车行业的应用日益广泛，在为驾驶者提供前所未有的便利性和舒适性的同时，其潜在的风险也逐渐显现，尤其是在安全性和可靠性方面的问题，这不仅关乎技术本身，更直接影响到公众的生命财产安全。汽车行业中人工智能的风险主要体现在系统可靠性、数据安全性、算法安全性等方面。

随着智能网联汽车的普及，智能网联汽车的可靠性问题成为首要关注点。技术故障和系统失效可能导致车辆在复杂交通环境中出现误判，特别是在恶劣天气条件下，传感器的准确性受到更大的挑战。根据美国国家公路交通安全管理局2022年发布的L2级自动驾驶事故数据报告^[1]，2021年7月1日至2022年5月15日的10个月内，有392起事故与L2级ADS辅助驾驶系统有关；自2019年以来，美国涉及特斯拉自动辅助驾驶模式有关的车祸事故达到736起，这些意外车祸导致了17人死亡。图1展示了2014-2020年美国加州统计的由于自动驾驶自身原因导致的事故，其占总事故的比例高达16%。^[2]



图1 自动驾驶事故原因分析

数据安全是另一个关键风险领域。一方面智能汽车需要依赖大量数据进行模型学习,但数据泄露事件频发,对用户隐私构成严重威胁。例如,2020年大众北美分公司和2024年宝马公司的数据泄露事件^[3],暴露了数据保护的不足。据不完全统计,2023年至今,国内共发生了超过20起与车企相关的数据泄露事件^[4],图2显示了2023年度智能网联汽车中与数据有关的攻击占比。另一方面,数据污染和对抗样本攻击等,也对智能网联汽车的安全性构成了威胁。这些安全问题不仅影响车辆的正常运行,还可能引发不可预见的后果,如自动驾驶系统的失控,对用户生命安全构成威胁。

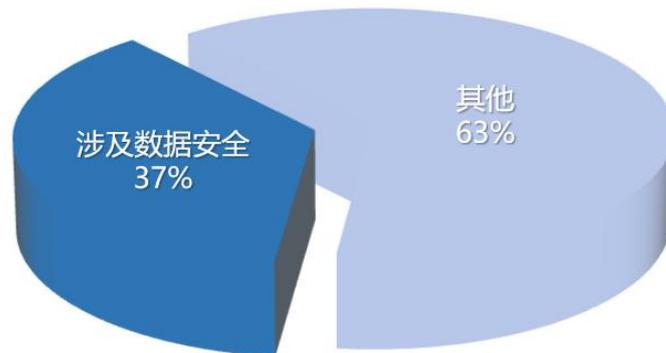


图2 智能网联汽车中与数据有关的攻击占比

此外,算法安全也不容忽视。从内部威胁角度来看,以深度学习为代表的算法,由于自身的黑盒特征,使其在可靠性、公平性、透明性和鲁棒性方面存在安全缺陷。从外部威胁角度来看,对抗样本攻击、数据投毒等行为诱使人工智能系统产生错误的推理结果,逆向攻击技术可以通过大量的模型预测查询,实现模型窃取而造成隐患。

（三）人工智能风险治理现状

1、政策与法律法规

面对人工智能技术发展带来的机遇与挑战，国内外正逐步制定和完善法律法规，以确保人工智能安全可信应用。

近年来，欧盟发布了《人工智能责任指令》及《人工智能法案》^[5]以减少 AI 系统开发者和使用者所面临的法律不确定性，强化人工智能监管与责任体系。美国先后颁布了《2020 国家人工智能倡议法案》^[6]、《安全人工智能法案》^[7]等多部法案，并在 2024 年签署了《AB 2655 法案》《AB 2839 法案》《AB 2355 法案》《AB 2602 法案》《AB 1836 法案》五项 AI 相关新法律，以强化人工智能使用的监管与透明度。2023 年，英国发布《促进创新的人工智能监管方法》白皮书^[8]，提出基于原则的人工智能治理方法与五项原则，为行业提供确定一致的监管指导。

近年来，国内人工智能领域的政策及相关要求逐步建立，发布了《中华人民共和国数据安全法》《汽车数据安全若干规定（试行）》《生成式人工智能服务管理暂行办法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《科技伦理审查办法（试行）》《国家人工智能产业综合标准化体系建设指南》等^[9-11]，在推动人工智能健康发展与应用的同时，强调人工智能技术在发展过程中需遵循基本原则。

相关政策及法律法规的相继发布表明国内外正在积极探索和完

善人工智能规范，在促进技术发展的同时，加强监管和透明度，确保公众利益和安全得到保障是十分重要且必要的。然而，当前人工智能领域的政策法规仍处于探索完善阶段，各项具体要求还在不断细化与优化之中。未来，随着技术的不断进步和应用场景的拓展，人工智能领域的政策法规也将不断完善，为各行各业的健康持续发展提供更加坚实的保障。

2、标准现状

在政策形式与技术发展的驱动下，国内外制定了一系列人工智能相关标准，为 AI 技术研发及应用规范提供建议和技术参考。

国际标准化组织（International Organization for Standardization, ISO）和国际电工委员会（International Electrotechnical Commission, IEC）成立了分委会 ISO/IEC JTC 1/SC 27 “信息安全、网络安全和隐私保护”与 ISO/IEC JTC 1/SC 42 “人工智能”，制定了 ISO/IEC 29119 《软件测试》^[12]、ISO/IEC 29134 《信息技术—安全技术—隐私影响评估指南》^[13]、ISO/IEC 38505 《数据治理》^[14]等标准，为信息安全、网络安全、隐私保护、软件测试及数据治理等领域提供了全球性的规范与研究支持，确保 AI 技术应用的可靠性、安全性和合规性，促进了 AI 技术的健康发展与广泛应用。IEEE SA 制定的 IEEE 7001-2021 《自主系统的透明性标准》^[15]、IEEE 7002-2022 《数据隐私设计标准》^[16]、IEEE 7007-2021 《伦理驱动的机器人和自动化系统本体标准》^[17]等标准，为自主系统的透明性、数据隐私保护和 AI 系统的伦理水平

等提供了明确的指导和规范，确保 AI 系统的可解释性、可追溯性，增强了公众对 AI 技术的信任。此外，ISO/SAE 21434《道路车辆-信息安全工程》^[18]、ISO 26262《道路车辆功能安全》^[19]，ISO 21448《道路车辆-预期功能安全》^[20]等汽车领域相关国际标准，均涉及对人工智能在道路车辆应用的技术要求，表明汽车行业对车用人工智能技术规范的高度重视。

国内，工业和信息化部等四部门联合印发《国家人工智能产业综合标准化体系建设指南（2024版）》^[21]，旨在形成引领人工智能产业高质量发展的标准体系，国内制定的 GB/T 41867-2022《信息技术 人工智能 术语》^[22]界定了人工智能领域中的常用术语及定义，给出了基础类、关键通用技术、关键领域技术、安全/伦理四大类术语的定义。在汽车行业，人工智能标准研究逐渐步入正轨。国内制定的 GB/T 40856-2021《车载信息交互系统信息安全技术要求及试验方法》^[23]、GB/T 41871-2022《信息安全技术 汽车数据处理安全要求》^[24]、GB/T 40861-2021《汽车信息安全通用技术要求》^[25]等标准，针对车载信息交互系统、汽车数据处理、汽车信息安全等方面提出了明确要求；同时，TC260-003《生成式人工智能服务安全基本要求》^[26]在生成式人工智能领域进一步提出数据生成的合规性要求，以确保敏感数据不会被滥用；GB/T36464.5-2018《信息技术 智能语音交互系统 第5部分：车载终端》^[27]标准为车载系统智能语音和人机交互的安全性和用户体验提供了系统化的测试框架，提升了智能座舱系统的交互稳

定性和响应能力。这些标准的制定与实施，对保障消费者权益、推动AI技术进步具有重要意义。

综上所述，人工智能技术在推动智能网联汽车领域快速发展的同时，也促进了汽车行业对人工智能在该领域的技术规范和标准研究进程。政产学研形成合力，旨在形成科学、合理、完善、共识的标准体系，为汽车行业人工智能技术稳健落地应用与高质量持续发展保驾护航。

（四）人工智能风险管理产业实践方案

为了确保人工智能的安全和可持续发展，各行业都积极探索并实施了风险管理的实践方案。我们调研总结了人工智能在金融、医疗、零售等领域的风险产业时间管理方案，总结如图3所示。

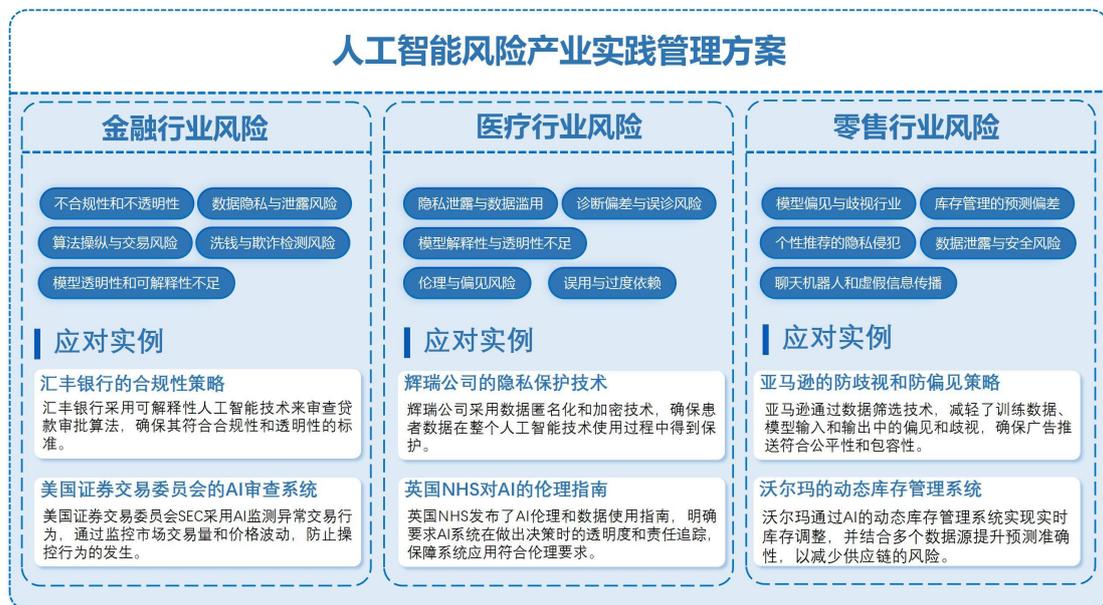


图3 人工智能风险管理产业实践方案

在金融行业中，人工智能技术可用于投资分析、风险评估、反欺诈监测等场景。然而，人工智能系统的不透明性和不可解释性给金融

机构带来了合规性上的风险。为应对其风险，汇丰银行采用可解释的人工智能技术来审查贷款审批算法，确保其具备合规性和透明性^[28]。

在医疗行业，人工智能技术可用于诊断辅助、药物研发、病人监测等场景。然而，人工智能系统对数据的不当读取和使用等行为，会引起患者隐私数据泄露风险。为应对相关问题，辉瑞公司采用数据匿名化和加密技术，确保患者数据在整个人工智能技术使用过程中得到保护^[29]。

在零售行业，人工智能技术可用于市场营销、客户服务、库存管理等场景。然而，人工智能系统因训练数据较少和不均衡的问题，会出现算法决策不公平的风险。为应对此风险，亚马逊通过数据筛选技术，减轻了训练数据、模型输入和输出中的偏见和歧视，确保广告推送符合公平性和包容性^[30]。

综上所述，一方面，虽然国内外根据其地区特点，制定了相关的法律法规及标准，以应对人工智能风险，但由于人工智能系统的多样性、复杂性，使得现有管理措施难以完全覆盖。另一方面，虽然金融、医疗等产业界在积极探索风险应对方案，但汽车行业尚未形成系统性、规范性的解决方案。因此，构建系统化的车用人工智能安全框架，形成具有实践性的车用人工智能安全解决方案已成为汽车行业迫切需求。

为解决上述问题，依托智能汽车安全技术全国重点实验室，由重庆长安汽车股份有限公司、中国汽车工程研究院股份有限公司牵头，

北京理想汽车有限公司、广州小鹏汽车科技有限公司、中汽数据（天津）有限公司、清华大学等多家单位参与，联合编制了本白皮书。白皮书共分六个章节，其中第二章主要阐述了车用人工智能安全体系框架，第三、四、五章分别阐述了车用人工智能风险、应对技术方案和管理方案，最后在第六章进行了总结与展望。

第二章 车用人工智能安全体系框架

本章围绕车用人工智能安全体系框架建设展开研究，主要包括总体框架概述、体系内容详细解析、体系框架与各章节内容之间的关联关系，依据该体系梳理出车用人工智能风险分析思路、技术防护方法与风险管理建议。

（一）体系概述

在人工智能技术的快速发展推动下，对于汽车行业而言，建立完善、系统化的车用人工智能风险管理框架，是保障智能网联汽车安全性、可靠性和合规性的基础。结合汽车领域发展背景及产品研发特点，借鉴国内外各行业风险管理经验，总结形成车用人工智能安全体系框架，如图 4 所示，主要包括风险分析、技术防护、风险管理三部分。

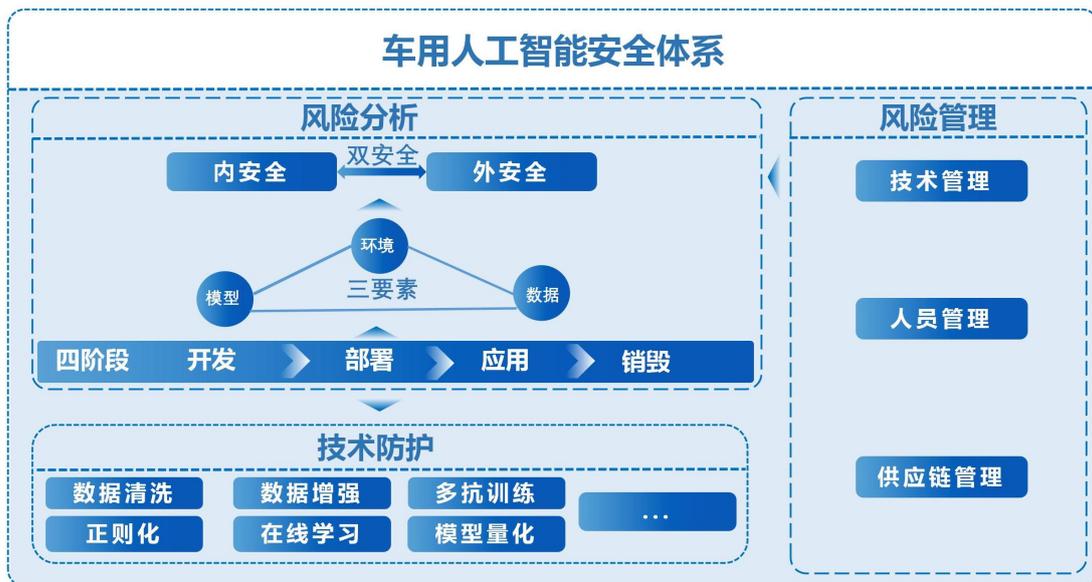


图 4 车用人工智能安全体系框架

风险分析：建立“2-3-4”车用人工智能风险分析框架，从 AI 产

品内生安全和应用安全双安全角度出发，结合 AI 数据、模型、环境三个关键要素，分析开发、部署、应用、销毁全生命周期的四个阶段面临的内生安全和应用安全问题。

技术防护：依据关键风险点，梳理 AI 产品全生命周期数据风险、模型风险、环境风险的应对方法参考，如数据清洗、数据增强、对抗训练、正则化等。

风险管理：结合汽车 AI 产品研发过程中涉及的技术、人员及上下游企业，从三个不同层面对风险管理提出建议，包括技术管理、人员管理、供应链管理。

（二）风险分析框架

本节针对车用人工智能安全框架中的“2-3-4”风险分析框架，给出定义与详细介绍。

1、双安全

汽车风险的双安全框架涵盖了内生安全（简称：内安全）与应用安全（简称：外安全）两大方面。本白皮书中内生安全主要指智能网联汽车在设计、制造和功能实现过程中的固有安全性，涉及系统结构的可靠性、软件和硬件的稳定性以及对潜在漏洞的防护能力；应用安全则指在汽车使用场景中的外部风险应对，包括数据安全、网络安全、交互安全等内容，侧重于车辆在复杂环境中，特别是智能化、网联化、共享化应用中的安全保障，确保车辆在实际运行中能够有效应对动态威胁，保障用户及环境安全。

2、三要素

车用人工智能风险管理的三要素包括数据、模型和环境，这三者是人工智能系统的核心，也是风险分析的关键维度。数据包括系统训练和推理所依赖的各类信息，如传感器数据、车载系统数据、与外界交互的数据等。数据的准确性和完整性直接影响系统的可靠性，数据泄露、篡改或污染都会带来意想不到的风险。模型是指人工智能系统中执行感知、判断和决策等推理运算的主体，错误的推理结果可能导致安全隐患。环境是指人工智能模型在开发、部署和应用过程中所处的软硬件环境，包括开发工具、硬件基础设施、网络环境等，未经充分防护和严格测试的环境可能导致系统面临崩溃。因此，对数据、模型和环境的全面管理是车用人工智能风险管理的核心，三要素的协同保障是提升汽车安全性和可靠性的基础。

3、四阶段

四阶段指的是汽车 AI 产品开发、部署、应用和销毁全生命周期四个阶段。开发阶段主要包括数据收集、模型设计、代码编写和模型训练等；部署阶段是指将人工智能模型与系统集成至车辆或云端并进行上线前测试的过程，主要关注系统与硬件的适配问题；应用阶段是指人工智能系统在车辆持续运行的过程，面临动态环境、数据变化和潜在网络攻击等安全挑战；销毁阶段是指系统终止使用并进行清理销毁的环节，主要关注敏感数据泄露。通过制定全四阶段生命周期的系统化风险应对方案，可以确保车用人工智能系统在整个生命周期内的

安全和可靠运行。

（三）关系

本章节提出的车用人工智能安全体系框架总领后续三、四和五章节。其中风险分析对应第三章内容，主要基于“2-3-4”框架给出车用人工智能关键风险要素，为后续风险应对技术方案和风险管理方案提供依据；技术防护对应第四章内容，主要围绕关键风险点梳理了当前业界主流的技术应对方案；风险管理对应第五章内容，主要从人员、技术、供应链管理角度给出风险管理方案。下面将分章节详细阐述。

第三章 车用人工智能风险分析

系统性分类与深入剖析风险点对于全面识别车用人工智能的潜在威胁至关重要，同时也是后续制定针对性明确、覆盖面广泛的风险防范策略的前提。本章节将聚焦于分析人工智能在汽车行业应用中的关键风险要素，从内生风险和外部应用风险两大方面进行探讨，逐步剖析环境、数据、模型这三大要素在各个应用场景所面临的具体挑战，旨在系统性地揭示潜在的风险隐患。本章通过多角度的风险分析，为人工智能与汽车融合领域的安全管理提供了明确的识别与评估框架，并为未来的风险防范技术研究构建了基础。

（一）内生安全

1、数据

（1）数据质量和完整性不足

指数据在准确性、完整性或一致性方面不符合预期标准（低质量或不完整的数据，例如噪声、缺失值、错误标注或“投毒”数据等），从而影响系统或决策的可靠性。这类风险可能导致错误分析、误导性结论或不可靠的预测，尤其在数据驱动的系统会放大风险。应重点在开发阶段注意该类风险。

（2）隐私泄漏

指未经授权访问、暴露或泄露个人敏感信息的行为。这类风险可能导致个人信息的滥用、身份盗用，甚至财务或心理损害，例如在智能座舱中，车辆可能会采集乘客的生物特征数据（如面部表情、语音

特征、虹膜识别等) 以提供个性化服务和安全验证。这些敏感数据在采集和处理过程中如未能获得充分授权与有效保护, 可能导致个人隐私泄露, 甚至违反数据保护法规。这类隐私保护不足的风险属于个人数据信息安全受侵害的范畴。应重点在开发和应用阶段注意该类风险。

(3) 偏见与歧视

指在算法设计和输出的结果中, 因个人偏见的有意或无意引入, 或由于训练数据集质量问题, 可能导致的潜在偏见或歧视。这类风险通常表现为模型在处理少数群体或特殊情况时性能不足, 甚至可能产生民族、宗教、国家或地域等方面的歧视性内容, 属于非故障状态下的潜在风险。应重点在开发阶段注意该类风险。

(4) 数据格式和协议不兼容

指不同系统或组件之间的数据结构或通信协议无法相互理解或交换信息。不同品牌的产品可能采用各自的专有协议或不同的标准协议, 导致数据包结构和编码方式无法直接对接。这类风险可能导致数据传输失败、系统集成困难, 甚至影响业务流程的连续性, 例如导致车辆的位置、速度和障碍物信息等关键数据在共享过程中发生损坏或丢失, 进而产生系统内部故障风险。应重点在部署阶段注意该类风险。

(5) 数据残留

指在汽车销毁过程中, 车辆中存储的历史行驶数据、用户设置、定位信息等敏感数据, 尤其是来自车载系统和内置存储设备的遗留信

息，在数据删除或转移后，系统中仍然留存的未完全清除的数据。即使这些数据表面上已经被删除或经过格式化处理，存储介质中仍可能残留数据。攻击者可以利用专业的恢复技术重新提取这些信息，从而引发隐私泄露和潜在的安全风险，导致数据信息遭到侵害。这类风险可能导致敏感信息的意外泄漏，进而引发隐私和安全问题。应重点在销毁阶段注意该类风险。

2、模型

(1) 模型偏差

指人工智能模型未能准确捕捉真实数据模式而导致的系统性误差，通常由模型结构、参数选择和正则化策略等引起。例如，简单模型可能只能识别基础场景，而忽略复杂环境。损失函数、学习率的设置不当可能让模型在特殊情况下表现失误，强正则化可能抑制细节特征的捕获，使模型在面对异常场景时反应不足。这类风险可能导致决策不公正、不准确，尤其在涉及敏感人群或关键业务时会带来负面影响。应重点在开发阶段注意该类风险。

(2) 鲁棒性弱

指模型在应对数据中的噪声、干扰以及非典型样本等干扰因素时可能出现不稳定的表现，导致错误决策。这类风险可能导致在自动驾驶等复杂应用场景中，模型难以应对数据分布的变化，或难以抵御恶意的扰动，从而增加车辆在正常驾驶过程中的安全事故风险，影响整体系统的稳定性。同时，在复杂和多变的环境中，数据质量下降或数

据漂移可能进一步加剧决策偏差，导致系统的可靠性和用户的安全体验受到不良影响。应重点在开发和应用阶段注意该类风险。

(3) 可解释性差

指系统或模型的内部逻辑、决策过程难以被用户或开发者理解和解释，尤其是用于智能驾驶的复杂算法，由于其内部运行机制高度复杂，推理过程往往属于黑盒或灰盒模式，使得其决策难以预测和确切归因。这类风险可能导致用户对系统输出缺乏信任，增加了异常情况发生时的调试难度，也使得其在应对环境变化和异常情况时存在潜在的不可控因素，影响车辆的安全性和可靠性。应重点在开发和应用阶段注意该类风险。

(4) 模型残留

指在模型更新、改进或替换后，旧模型的部分特征或结果仍然影响新模型的表现。这类风险可能导致模型中包含的经过长期训练积累的用户驾驶偏好和个人兴趣等敏感信息被他人盗取，导致隐私泄露或不当使用。这类风险可能导致模型的权重参数等核心信息在销毁不彻底时，被恢复或逆向分析，暴露出算法的设计逻辑和核心代码，增加被逆向工程的风险，从而带来技术泄密，导致专利和商业机密的流失。应重点在应用和销毁阶段注意该类风险。

3、环境

(1) 部署环境不匹配

指在不符合开发时设定的硬件、操作系统或配置条件下运行，导

致性能下降或功能异常。这类风险可能导致模型表现异常或无法运行，增加故障发生率，影响用户体验或业务连续性。应重点在部署阶段注意该类风险。

(2) 服务不可用

指系统或服务因各种原因无法正常提供，导致用户无法访问或使用关键功能。这类风险可能导致服务宕机或不可用从而影响系统的连续性，尤其在实时性和可靠性要求较高的场景中。如果缺乏高可用性架构的支撑，在硬件故障、网络延迟或突发流量下可能导致系统内部引发故障，无法正常运行。应重点在应用阶段注意该类风险。

(3) 环境配置和凭证泄露

指敏感的系统配置或访问凭证（例如 API 密钥、数据库连接信息及服务令牌等）被意外暴露，导致未经授权的用户可以访问系统或数据。这类风险可能导致严重的安全事件，如数据泄露、系统篡改或业务中断，实施恶意操作甚至控制系统功能，进而影响企业声誉和用户信任。应重点在应用阶段注意该类风险。

(4) 计算环境残留安全隐患

指在系统运行或任务完成后，临时数据、缓存或配置文件未被彻底清除，可能被恶意用户利用进行攻击，或者是在报废销毁的过程中，未能彻底关闭或清理计算环境可导致潜在的安全隐患。这类风险可能导致敏感信息泄漏、权限提升或系统被持久化控制，增加安全事件发生的概率。应重点在应用和销毁阶段注意该类风险。

（二）应用安全

1、数据

（1）信息窃取

指未经授权的个人或实体非法获取系统中的敏感数据或机密信息，包括模型窃取和数据窃取。模型窃取是攻击者通过多次查询复制模型结构和参数，构建相似的本地模型；数据窃取则是通过输出反推模型的训练数据，获取隐私或敏感信息，进而推测用户隐私和训练集内容。这类风险可能导致用户隐私泄露、知识产权损失，甚至引发财务或法律问题，严重影响企业安全和用户信任。应重点在开发和应用阶段注意该类风险。

2、模型

（1）复用缺陷传导

指在人工智能模型的二次开发和微调过程中，基础模型中的缺陷可能会传导至下游模型，带来潜在的安全隐患。这类风险可能导致基础模型中未彻底消除的误差、偏差或漏洞，在新的应用环境中被放大，从而影响下游应用的稳定性、准确性或安全性。应重点在应用阶段注意该类风险。

（2）对抗攻击

指攻击者通过精心设计的输入或操作，故意引发系统错误或使模型输出不准确的行为，诱骗人工智能系统做出错误决策。这类风险可能导致在自动驾驶场景中，攻击者通过伪造或修改交通标志、投射虚

假障碍物等手段，使系统的鲁棒性和可靠性下降，从而误判路况或障碍物。应重点在应用阶段注意该类风险。

（3）生成内容隐患

指生成式人工智能在自动驾驶和智能车机系统中生成、传输或展示的内容可能包含不合规、不适当或具有误导性的信息。这类风险可能导致误导用户、系统误操作、法律和合规风险、道德和声誉风险等。应重点在开发和应用阶段注意该类风险。

3、环境

（1）网络攻击

指攻击者可能利用系统中的未修补漏洞（如 API、操作系统漏洞）对信息系统或网络的恶意攻击行为，旨在窃取、破坏或篡改数据，或使系统瘫痪，控制整个系统或注入恶意代码，从而导致系统行为异常或数据泄露。这类风险可能导致数据泄露、业务中断和财务损失，严重时可能影响企业的生存和声誉。应重点在开发和应用阶段注意该类风险。

（三）车用人工智能风险归纳

对前面分析的车用人工智能风险点进行了系统总结，并根据风险的产生原因进行分类，将各类风险归属到功能安全、预期功能安全、数据安全和网络安全四个方面，结果如表 1 所示。这种分类方式的主要目的是为车企提供多方位的风险分析参考，使其能够根据不同的风险类型找到相应的安全团队进行针对性处理，提升车企在应对不同安

全风险时的效率和专业性，有助于全面保障车用人工智能系统的安全性和稳定性。

表 1 车用人工智能风险归纳表

风险分类	风险要素	风险点	影响的阶段	问题归属
内生 风险	数据	偏见与歧视	开发阶段	预期功能安全
		隐私泄露	开发阶段 应用阶段	数据安全
		数据质量和完整性不足	开发阶段	数据安全 预期功能安全
		数据格式和协议不兼容	部署阶段	功能安全
		数据残留	销毁阶段	数据安全
	模型	模型偏差	开发阶段	预期功能安全
		鲁棒性弱	开发阶段	预期功能安全
		可解释性差	开发阶段 应用阶段	预期功能安全
		模型残留	销毁阶段	数据安全
		部署环境不匹配	部署阶段	功能安全
		服务不可用	应用阶段	功能安全
		环境配置和凭证泄露	应用阶段	网络安全
		计算环境残留安全隐患	应用阶段 销毁阶段	网络安全
应用 风险	数据	信息窃取	开发阶段 应用阶段	数据安全 网络安全
	模型	复用缺陷传导	应用阶段	预期功能安全
		对抗攻击	应用阶段	预期功能安全
		内容隐患	开发阶段 应用阶段	预期功能安全
	环境	网络攻击	开发阶段 应用阶段	网络安全

第四章 车用人工智能风险应对技术方案

本章将围绕车用人工智能全生命周期中的关键风险点，梳理当前业界主流的应对技术方案，从“技术”视角解决安全风险。基于第三章主要安全风险点，本章从车用人工智能的四个阶段展开，系统分析各阶段的安全保障方法和技术手段。通过对当前行业技术方案的调研分析，本章将给出了一些主流的人工智能风险应对技术方案，为车用人工智能在全生命周期中的安全性与可靠性提供参考。

（一）数据

数据在车用人工智能系统中承担着决策支撑的关键角色，在第三章中分析了数据面临的偏见与歧视、隐私泄露、数据质量与完整性不足、格式和协议不兼容、数据残留、信息窃取等多重风险，这些问题贯穿于人工智能系统的各个阶段。基于此，本节将围绕开发、部署、应用和销毁四个阶段，针对数据风险的具体表现，逐一分析各阶段的应对技术方案，主要内容如图5所示。

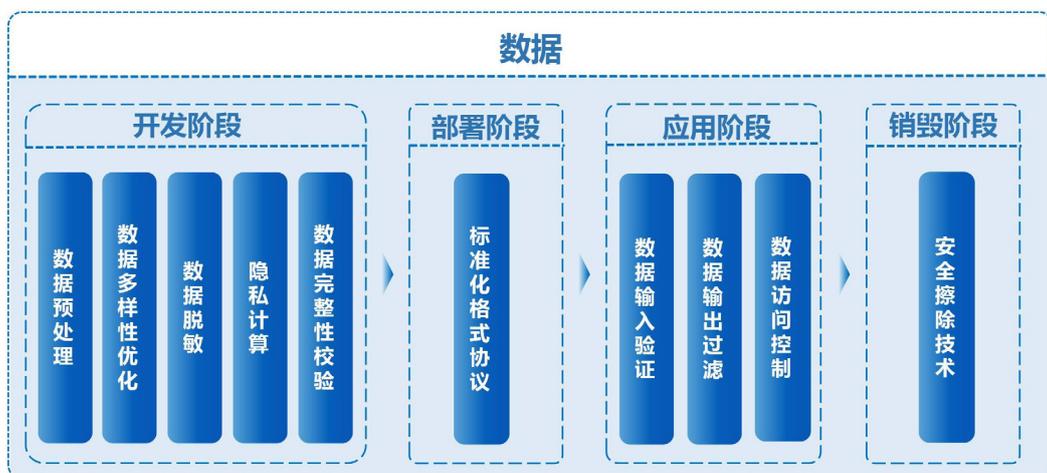


图5 应对技术方案（数据）

1、开发阶段应对技术方案

(1) 数据预处理

通过对数据进行预处理，达到提高数据质量和模型性能的效果。可解决偏见与歧视、数据质量和完整性不足、可解释性差、模型偏差、鲁棒性弱、内容隐患等带来的风险。主要技术手段有数据清洗、数据增强、归一化、标准化、缺失值填充、异常值处理、滤波去噪、图像矫正等。

(2) 数据多样性优化

通过构建多样化的数据集并引入公平性约束，旨在提升模型的公平性并减少潜在偏差，从而有效应对偏见、歧视、可解释性差、模型偏差及鲁棒性不足等风险问题。主要技术手段包括多样化数据采集、数据平衡、偏差修正及应用公平性算法等。

(3) 数据脱敏

通过数据匿名处理和去标识，达到保护用户隐私和数据安全的效果，以解决信息泄露等带来的风险。主要技术包括替换、掩码、数据切割、数据扰动等。

(4) 隐私计算

通过隐私计算，达到在保护数据隐私的前提下实现数据分析和协同计算的效果，以解决隐私泄漏带来的风险。主要技术手段有联邦学习、安全多方计算、同态加密、差分隐私等。

(5) 数据完整性校验

通过数据校验和版本控制技术，确保数据的完整性、真实性和可控性，实现数据可追溯。主要技术手段包括哈希函数计算、数字签名算法、公钥基础设施（PKI）、版本控制软件和区块链等。

2、部署阶段应对技术方案

(1) 标准化格式协议

通过标准化格式与协议，达到提高系统兼容性和数据互操作性的效果。可解决数据格式和通信协议不兼容、部署环境不匹配、服务不可用等带来的风险。主要技术手段有使用 JSON、XML 等标准数据格式，遵循 HTTP、RESTful API、SOAP 等通信协议，使用标准化通信栈（如 AUTOSAR）采用简化协议转换。

3、应用阶段应对技术方案

(1) 数据输入验证

通过输入验证与异常检测，达到提高系统安全性和稳定性的效果，确保输入格式和范围的合规性，过滤异常输入，并通过行为分析检测潜在攻击活动，限制查询次。可解决信息窃取、对抗攻击、内容隐患等带来的风险。主要技术手段有数据格式校验、范围检查、正则表达式验证、异常检测算法、实时监控和报警等。

(2) 数据输出过滤

通过对敏感信息的输出进行严格限制和内容过滤，达到保护数据隐私、防止信息泄露的效果，有效解决了因过度输出和信息泄露导致

的隐私泄漏、数据滥用和信息窃取等风险问题。主要技术手段包括权限控制、数据脱敏、输出过滤以及日志管理等。

（3）数据访问控制

通过访问控制和多因素验证，达到提高系统安全性和防止未经授权访问的效果。可解决信息窃取等带来的风险。主要技术手段有身份验证、权限管理、多因素认证、生物识别技术等。

4、销毁阶段应对技术方案

（1）安全擦除技术

通过多次覆盖擦除等安全擦除技术，能够彻底删除敏感数据，保障数据和模型安全，有效解决关键信息遗留带来的安全隐患。主要技术手段包括数据覆盖（如多次随机写入）、加密擦除、物理销毁、安全删除命令、伪随机填充和专用数据擦除算法等。

（二）模型

模型作为车用人工智能系统的核心部分，其面临可解释性差、模型偏差、鲁棒性弱、模型残留、复用缺陷传导、对抗攻击及内容安全等风险，接下来分阶段探讨模型风险的应对技术方案，以增强模型的可靠性和抗风险能力，主要内容如图 6 所示。

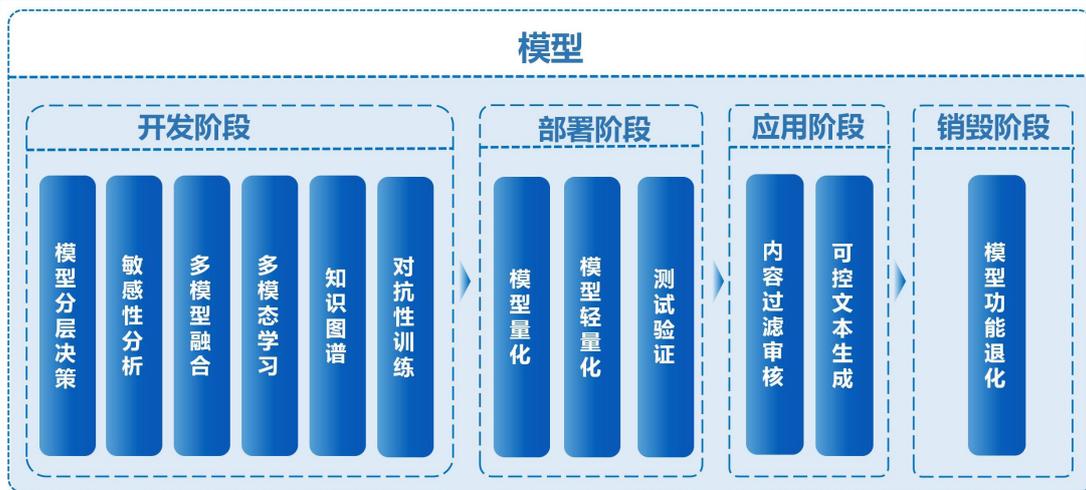


图6 应对技术方案（模型）

1、开发阶段应对技术方案

（1）模型分层决策

通过分层决策，将输入特征的影响分解和可视化，帮助解读单一特征在特定输出中的作用，可解决模型可解释性差、模型偏差、鲁棒性弱等问题带来的风险。例如，在自动驾驶系统中，通常会结合深度学习和强化学习，通过分层控制实现从路线规划到低级别控制的自动决策。主要技术手段包括权限划分、决策支持系统、分布式决策模型、广义加性模型^[31]等。

（2）敏感性分析

通过敏感性分析，对输入进行微小扰动，识别对决策影响最大的关键因素，评估模型在复杂环境中的反应。主要分析模型对输入特征的小幅变化所做出的输出变化，来衡量各个特征对模型预测结果的重要性。以应对模型偏差、鲁棒性弱、复用缺陷传导、对抗攻击等带来的风险。主要技术手段有梯度分析、特征重要性评估、局部可解释性

模型（LIME）、Shapley 值等。

（3）多模型融合

通过组合多个机器学习模型的预测结果，减少单个模型的偏差和方差，以解决模型鲁棒性弱的风险，提供更可靠和一致的预测结果。主要技术手段有区域特征提取、动态参数调整、多模型集成等。

（4）多模态学习

通过结合多种数据类型（如文本、图像、音频、视频）进行模型评估和结果解释，利用不同类型数据的互补信息，来增强模型的鲁棒性、提高预测的准确性，并增强对结果的解释能力。可解决可解释性差、模型偏差、鲁棒性弱、复用缺陷传导等带来的风险。主要技术手段有多模态数据融合、特征可视化、可解释性算法（如 LIME 和 SHAP）和跨模态一致性检查等。

（5）知识图谱

通过知识图谱进行内容验证，使用检索增强生成进行实时信息访问，实施多模型一致性检查，建立审核框架以明确责任，并通过用户反馈促进人机协作。可解决模型偏差、鲁棒性弱、复用缺陷传导等带来的风险。主要技术手段有实体识别与消歧、关系抽取、语义标注、图数据库构建和查询等。

（6）对抗性训练

通过将模型暴露在特别设计的对抗样本或微小扰动下进行训练，从而增强其对恶意干扰和输入变化的抵抗力，提升模型的鲁棒性。以

应对模型的可解释性差、偏差、鲁棒性弱和对抗攻击带来的风险。主要技术手段包括生成对抗样本、添加噪声扰动、对抗性数据增强、设计对抗损失函数、正则化项以及梯度惩罚等。

2、部署阶段应对技术方案

(1) 模型量化

通过对模型参数进行从高精度浮点数（如 FP32）到较低精度数值（如 INT8、INT4 等）的转换，并进行针对目标硬件环境的必要算子优化，以减少模型大小、内存带宽需求和计算量，从而优化模型部署到目标硬件后的推理速度和功耗。为避免量化后引入过大的精度及性能损失，应在量化后进行误差分析，并可能需要进一步校准确定量化参数及对量化模型进行微调。可解决部署环境不匹配、服务不可用、网络攻击等带来的风险。主要技术手段有权重量化、激活量化、定点数表示和混合精度计算等。

(2) 模型轻量化

通过基于剪枝压缩、推理加速与优化架构的实时性保障方法，减少模型资源需求，使其在资源受限的环境中依然高效运行，确保系统实时性和稳定性。可解决部署环境不匹配、服务不可用、网络攻击等带来的风险。主要技术手段有模型剪枝、量化压缩、硬件加速、优化模型结构和推理加速框架等。

(3) 测试验证

通过测试验证方法，实现车用人工智能系统在不同场景下的性能

表现和潜在问题的检测，解决了硬件平台兼容性不足以及传统测试难以覆盖极端工况的问题。主要技术手段包括 HIL、SIL、MIL 等，全面测试车用人工智能系统的物理硬件响应能力、软件功能验证，以及模型的正确性和稳定性。

3、应用阶段应对技术方案

(1) 内容过滤审核

通过内容过滤与审核机制，用于检测、评估和管理内容的技术体系，广泛应用于社交媒体、搜索引擎、内容平台等场景中，以确保发布或传递的内容符合特定的法律、道德或社区标准。内容过滤与审核机制旨在识别和拦截不良信息，如暴力、色情、仇恨言论、虚假信息 etc，维护平台的健康环境并保障用户的安全。可解决偏见与歧视、信息窃取、对抗攻击、内容隐患等带来的风险。主要技术手段有关键词过滤、图像识别审核、自然语言处理（NLP）审核、人工审核和机器学习模型辅助审核等。

(2) 可控文本生成技术

通过可控文本生成技术，达到控制参数、规划生成及约束条件限制生成内容的随机性，确保输出符合预期。为保障用户信任，生成内容安全方案还包括内容审核、溯源系统、多模态鉴真检测、上下文感知校正及人机协同审查，确保生成内容的准确性和合规性。可解决复用缺陷传导、对抗攻击、内容隐患等带来的风险。主要技术手段有条件生成、内容约束算法、关键词控制、语义过滤和惩罚机制等。

4、销毁阶段应对技术方案

(1) 模型功能退化

通过模型功能退化，将人工智能模型的关键神经元或权重进行剪枝，在压缩模型的同时让其变得不可用，以达到销毁的效果。或者通过极限量化（如 1 位或 2 位量化）使模型无法正常工作，从而实现低成本销毁。也可以使用模糊化方法调整模型的参数分布，使其输出不再准确，从而防止模型的恶意复原。可解决模型残留、信息窃取等带来的风险。主要技术手段有剪枝、极限量化等。

(三) 环境

环境作为车用人工智能系统开发、应用和部署的运行基础，面临的风险主要包括代码安全问题、部署环境不匹配、服务不可用、环境配置和凭证泄露、计算环境残留安全隐患，以及网络攻击等。这些风险可能对系统的正常运行和数据保护构成威胁。本节将针对上述环境风险，分析有效的防护技术和实践方法，以确保人工智能系统在各应用场景中的安全性和稳定性，主要内容如图 7 所示。

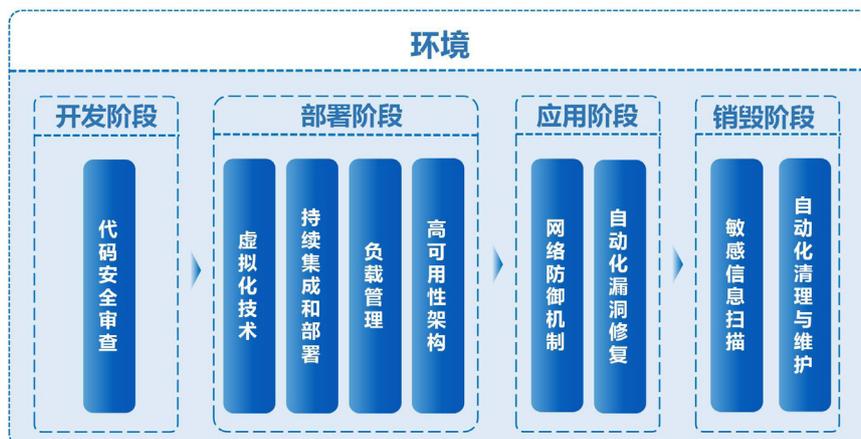


图 7 应对技术方案（环境）

1、开发阶段应对技术方案

(1) 代码安全审查

通过代码安全审查，确保代码开发过程的安全性和可追溯性，预防潜在的后门和信息泄露风险。此过程不仅有助于识别代码脆弱性，还能移除调试接口，防范信息窃取等安全隐患。主要技术手段包括静态代码分析、动态行为监测、代码签名、哈希校验、代码差异分析、机器学习模型识别、异常检测和威胁情报对比等。

2、部署阶段应对技术方案

(1) 虚拟化技术

通过容器化等虚拟化技术，实现环境一致性管理，使开发与部署环境一致，避免环境不一致导致模型失效。可解决部署环境不匹配、服务不可用等带来的风险。主要技术手段有 Docker、Kubernetes、虚拟机管理、IaC 工具（如 Terraform、Ansible）等。

(2) 持续集成和部署

通过持续集成和持续部署（CI/CD），确保配置、依赖与版本的一致性，减少人为错误。此外，可在流程中集成交叉编译机制，以支持在不同硬件架构上构建和测试应用，帮助开发团队快速识别并解决跨平台问题，确保代码在目标运行环境的兼容性和稳定性。可解决部署环境不匹配、服务不可用等带来的风险。主要技术手段有自动化测试、构建流水线、版本控制集成、自动化部署工具（如 Jenkins、GitLab CI/CD）等。

(3) 负载管理

通过压力测试评估系统的承载能力，结合负载均衡等技术优化资源分配和响应性能，确保系统在各种负载条件下保持稳定性和可靠性。能够有效识别系统在高负载下的弱点，预测资源需求，进行容量规划，避免服务不可用、部署环境不匹配及网络拥堵等风险。主要技术手段包括负载均衡器、弹性扩展服务、自动伸缩策略、资源预留等。

(4) 高可用性架构

通过冗余系统和分布式架构设计，确保流量在服务宕机时能够自动分配，使系统在关键组件或传感器故障时仍能正常运行。可解决部署环境不匹配、服务不可用、网络攻击等带来的风险。主要技术手段有冗余设计、故障转移机制、集群架构、负载均衡和自动故障检测与恢复等。

3、应用阶段应对技术方案

(1) 网络防御机制

通过防御性检测机制，达到监控、检测和过滤潜在的恶意数据或异常样本，设计防护机制，避免模型因分布外数据输入导致的错误输出，避免系统在面临对网络攻击时产生错误的决策。可解决复用缺陷传导、对抗攻击、内容隐患等带来的风险。主要技术手段有异常检测算法、入侵检测系统（IDS）、日志分析、实时监控预警、Web 应用防火墙等。

（2）自动化漏洞修复

通过定期运行漏洞扫描和补丁管理工具，及时发现并修复系统和应用中的安全漏洞，防止潜在的远程攻击，保障系统安全性与稳定性。可解决代码质量问题、部署环境不匹配、服务不可用等带来的风险。主要技术手段有静态代码分析、动态应用安全测试（DAST）、持续漏洞扫描、自动补丁管理和安全漏洞数据库对比等。

4、销毁阶段应对技术方案

（1）敏感信息扫描

通过 API 密钥、数据库连接信息及加密密钥等，确保敏感信息不会在系统销毁过程中被保留或暴露。同时，建立详细的配置文件和凭证清单，以便实现针对关键数据的全面追踪管理，确保敏感信息清理的完整性和有效性。可解决隐私泄漏、数据残留、模型残留、环境配置和凭证泄漏、计算环境残留安全隐患、内容隐患等带来的风险。主要技术手段有关键字匹配、正则表达式扫描、数据分类与标识、DLP（数据丢失防护）工具、自动化审计、SecretScanner 等专用敏感信息扫描工具等。

（2）自动化清理与维护

通过自动化清理与安全维护，定期清理快照、缓存、日志等可能残留敏感信息的文件，降低数据滞留风险，优化存储资源。针对安全策略、访问令牌和证书等关键数据进行周期性更新和清除，确保系统中仅保留当前有效数据。通过自动化的动态维护机制，系统在销毁前

可持续地强化信息安全，确保稳定可靠的安全保障。可解决数据残留、模型残留、环境配置和凭证泄漏、计算环境残留安全隐患等带来的风险。主要技术手段有日志清理、临时文件删除、自动化补丁管理、权限回收和安全策略更新等。

第五章 车用人工智能风险管理方案

车用人工智能系统所面临的风险主要是通过开发人员操作失误、技术成熟度不足以及供应链三个方面引入，有必要针对上述三个方面进行管理，从“人”的视角解决安全风险。本章聚焦于人员管理、技术原理、供应链管理三个方面，构建全面的车用人工智能风险管理方案，为汽车企业人工智能安全的管理提供了可行性建议。

（一）车用人工智能风险管理方案

车用人工智能风险管理方案借鉴了现有网络安全指南的流程架构图（SAE J3061^[32]、GB/T 38628-2020^[33]），深度融合了现有质量体系（ISO 9001^[34]、IATF 16949^[35]）、产品开发流程体系（ISO/SAE 21434^[18]）。本方案如图 8 所示，主要包括三个方面：一是人员管理，包括组织机构、管理机制、能力要求；二是汽车全生命周期人工智能安全技术管理，涵盖了开发、部署、应用、销毁四个阶段所开展的人工智能安全相关活动；三是供应链管理，包括工具链管理和供应商管理。

通过该方案，一方面，企业可以保证各标准间的协同一致，建立标准之间有效的沟通与协调渠道；另一方面，可以提高车用人工智能应用的质量、安全性和可追溯性，同时提高企业效率和车用人工智能风险评估能力，降低车用人工智能开发成本，帮助企业和整个社会安全、高效地从车用人工智能使用中获得最大价值。

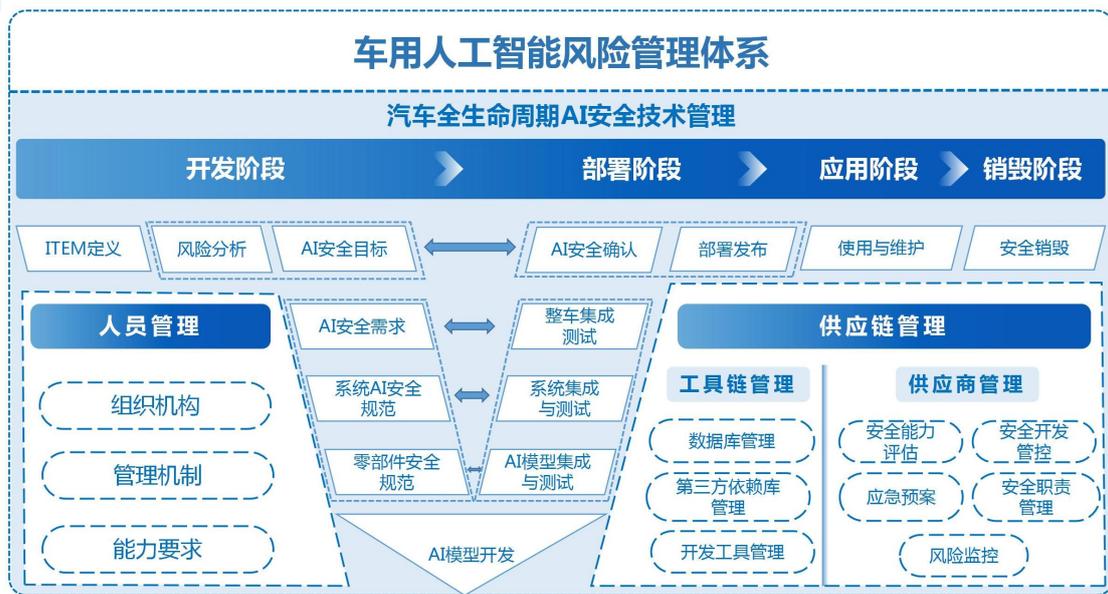


图 8 车用人工智能风险管理方案

(二) 人员管理

1、组织机构

参考 ISO SAE 21434^[18]、UN-R155^[36]、ISO/IEC 42001^[37]等标准规范在组织内建立问责制，明确规定车用人工智能风险管理方案的运行与管理的各项职责，建议结合企业组织架构设立三层分级的专业车用人工智能安全组织管理机构，主要包括领导小组、人工智能安全指导小组、人工智能安全执行小组，各小组的组成成员可根据公司的需求进行规划，同时，车型项目组应提供必要的支持，推荐的车用人工智能风险管理方案组织机构见图 9。

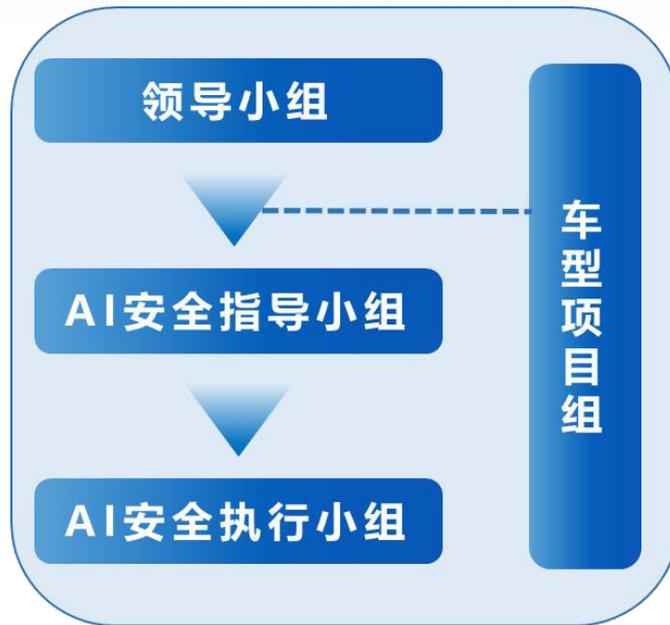


图9 车用人工智能风险管理方案组织结构

（1）领导小组

作为权力最高者在组织结构的最顶层，全面指挥和控制组织的运营，包括人工智能风险管理的规划与实施。具体指制定与组织战略方向一致的风险管理方针和目标，并监督其实施情况；对重大风险事项进行决策，并监督风险管理措施的执行情况。

（2）AI 安全指导小组

作为指导者，从技术和合规双维度，输入前沿技术趋势，分析预测国家法律法规要求，指导企业车用人工智能风险管理方案的优化和执行，并定期组织车用人工智能风险管理方案内部审核并审批内审结果，确保车用人工智能风险管理方案的适宜性、充分性和有效性。

（3）AI 安全执行小组

作为组织结构中的关键基层，主要负责按照指导小组的要求，按照车用人工智能风险管理方案落地执行，主要涉及人工智能安全开

发、测试、运营方面的工作，并向指导小组提出在执行过程中发现的问题及优化建议。

2、管理机制

为确保车用人工智能风险管理方案的有效运行与持续改进，借鉴《人工智能安全治理框架》^[38]，建议完善人员管理机制并严格执行，具体方案要求如图 10 所示。



图 10 人员管理方案

①方案分级管理：根据不同阶段（包括开发、部署、使用、销毁）的具体职责，实施严格的访问控制策略，分配适当的系统访问与操作权限，针对特定人群及应用场景，提出人工智能技术应用的相关规范与要求，防止人工智能系统被滥用。

②可追溯管理制度：建立全面的持续监控与审计机制，对人员操作行为、系统日志及异常事件进行实时追踪与记录，以便及时发现并调查任何违规行为或异常操作模式，为安全事件的预防与应对提供有

力支持。

③数据安全和个人信息保护规范：针对人工智能技术及应用特点，明确人工智能开发、部署、使用、销毁各环节的数据安全和个人信息保护要求，并告知相关人员，严格遵循此规范的要求。

④负责任的管理方案：要求相关人员遵循“以人为本、智能向善”的具体操作指南和最佳实践，持续推进人工智能设计、研发、应用的价值观、伦理观对齐，制定人工智能伦理审查准则、规范和指南，完善伦理审查制度。

⑤AI 安全风险威胁信息共享和应急处置机制：持续跟踪分析方案存在的安全漏洞、缺陷、风险威胁、安全事件等动向，协调有关研发者建立风险威胁信息共享机制；设计详尽的应急处置流程，确保在发生安全事件时，能够迅速启动应急机制，组织相关人员按照既定流程高效响应，有效控制事态发展，最大限度地减少损失与影响。

⑥人员交接制度：确保在人员变动时，做好知识和权限的交接，确保敏感信息的安全传递与系统权限的适时调整，防止因人员变动导致的信息泄露或安全漏洞风险。

3、能力要求

车用人工智能技术研究要求相关人员具备多方面的能力和意识，以确保系统的安全性、可靠性和合规性，包括工具熟练度、领域知识、隐私保护意识等。

①工具熟练度：不同人员根据其职责和专业领域需要熟练掌握不

同的工具和技术，例如开发人员需要具备扎实的编程基础，掌握人工智能开发框架，测试人员需要熟练使用各种测试、评估工具。

②领域知识：相关人员需要深刻理解人工智能的基本原理和技术细节；掌握信息安全的基本概念和技术，了解常见的攻击手段及其防御策略；紧跟人工智能安全领域的最新发展动态，了解并遵守国际国内的相关法律法规、行业标准及最佳实践指导原则。

③隐私保护与伦理意识：相关人员需要深入理解数据保护原则，包括数据的收集、存储、使用和传输等方面的规定；在设计和应用人工智能解决方案时，始终将伦理考量放在首位，确保技术的应用不会侵犯个人隐私权或其他基本权利。

（三）技术管理

1、ITEM 定义

在 ITEM 定义阶段，主要是为了系统性地筛选出车用系统在全生命周期中涉及人工智能的所有相关功能并明确该功能类别存在的风险点，该过程将有助于建立一个清晰的基础框架，确定需要进行风险分析的具体范围和内容。根据车型项目组的具体需求，建议按照如下步骤进行执行：

①识别和筛选相关功能：确定车辆上的人工智能功能，如自动驾驶、智能导航、语音识别等。筛选出与安全相关的功能，即需要纳入安全风险管理的范围的功能。

②明确人工智能功能的安全风险点：从安全性的角度，参考车用

人工智能全生命周期的风险归纳表（见第三章），将这些被纳入安全风险管理的功能明确为安全风险项，明确该功能在全生命周期可能存在的风险点。

通过以上步骤，可以建立一个清晰的安全风险点列表，并为后续的风险分析提供基础。

2、风险分析与安全目标建立

在完成 ITEM 定义后，接下来进行风险分析和相关安全目标建立。在此过程中，建议根据具体业务场景，分析是否存在已梳理的风险点，对存在的风险点进行风险评估与分级，然后根据风险等级建立相应的安全目标，并下发至相关技术团队。建议步骤如下：

①风险分析准备：在进行风险分析之前，需要收集具体业务信息和技术团队的资源，明确需要分析的业务场景，将其纳入风险分析范畴。然后根据已完成的 ITEM 定义和车用功能及相关风险点，启动风险分析过程。

②业务风险分析与风险点确立：对每个业务场景进行详细的风险分析，评估其是否会对用户人身、财产安全或软硬件安全等造成不良影响。如果经过分析发现某个业务场景存在潜在不良影响，则将其确认为一个风险点。

③风险评估与分级：在完成风险点确认后，需要对已识别的风险点进行全面评估，并根据风险的严重性和潜在损伤进行适当的分级，其目的是确定车用人工智能系统的潜在威胁并量化其对道路安全、用

户隐私和其他相关利益的影响程度，为下一步建立安全目标提供基础。参考欧盟人工智能法案，建议结合实际的业务场景将风险分为不可接受风险、高风险、有限风险和轻微风险四类^[5]。其中，不可接受风险指会对用户利益和道路安全构成严重威胁的风险；高风险指对用户权益和道路安全产生重大伤害的风险；有限风险指对用户利益和道路安全构成一定威胁的风险；轻微风险指对用户权益和道路安全威胁最小的风险。

④建立安全目标：针对已确认的风险点，根据其风险等级进行相应安全目标的制定，建议进行如下考虑：对于不可接受风险，其安全目标是完全防止风险的发生；对于高风险，其安全目标着重于最大限度地减少风险；对于有限风险，其安全目标是限制风险的范围和影响，确保在风险发生时能够及时识别并采取适当的控制措施来减轻损害；对于轻微风险，其安全目标主要集中于保持合理安全水平，以防止不利事件的发生。这些安全目标的建立，应当具体明确，能够充分帮助技术团队开展相应的工作来降低风险并确保安全性。

在风险分析和安全目标确立过程中，应当充分与业务团队结合，确保安全目标足够完整、全面而准确。

3、安全需求确定与相关规范确立

在完成风险分析与安全目标建立后，需要根据安全目标与企业内部相关部门进行人工智能安全需求确定和相关规范确立，在已存在人工智能安全跨部门团队的基础上，推荐执行如图 11 所示的步骤：



图 11 人工智能安全目标确立与确认过程示意图

①分析现有风险评估结果和安全目标：跨部门团队成员间共同回顾已完成的风险分析和安全目标建立阶段的结果。分享和讨论各部门对风险和安全目标的理解和视角。

②开展联合需求调研和分析：各部门代表协同进行需求调研和分析，深入了解每个部门在安全方面的需求和关注点。

③协同制定人工智能安全需求：团队成员共同参与，根据联合需求调研和分析的结果，制定全面而具体的人工智能安全需求清单。确保安全需求涵盖了各个部门的关切，并符合行业标准和最佳实践。

④制定系统人工智能安全规范：针对人工智能安全需求，团队成员共同制定系统级别的人工智能安全规范，明确系统设计、开发、测试和维护阶段的安全要求。规范内容需要包括系统架构的安全性、数据传输和存储的安全要求、用户认证和权限管理等。

⑤制定零部件人工智能安全规范：在团队成员的协同努力下，制定零部件级的人工智能安全规范，其中包括数据、模型和环境。制定

数据安全规范，确保数据的完整性、机密性和可用性。制定模型安全规范，包括训练数据质量要求、模型验证和测试的要求，以及针对攻击和对抗样本的防御措施。制定环境安全规范，考虑车辆硬件和软件的安全要求，如防火墙、入侵检测系统、故障诊断和恢复等。

⑥定期审查和更新：团队定期召开会议，审查和更新人工智能安全需求和相关规范，以适应不断变化的安全威胁和技术发展。同时，推动团队间的协同合作，确保安全规范与实际情况相符，并持续提高人工智能系统的安全性。

通过如上步骤，将促进企业内部各部门之间的协同合作，确保人工智能安全需求和相关规范的制定得到各部门充分的参与和贡献。

4、人工智能模型开发

此阶段主要在完成人工智能安全需求确定与规范确立的基础上，在模型开发阶段对研发过程执行安全要求，建议参考第四章内容，在开发阶段推荐考虑如图 12 所示事项：

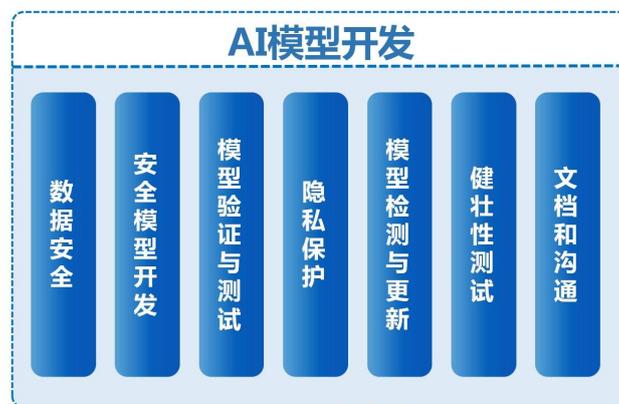


图 12 人工智能模型开发过程考虑事项示意图

①数据安全：确保使用的训练数据和测试数据具有高质量、代表

性和可靠性，并符合数据隐私和保护要求。

②安全模型开发：在模型开发过程中，要采用严格的安全开发实践，包括代码审查、防御性编程、输入验证等，以减少潜在的安全漏洞和攻击面。

③模型验证和测试：进行充分的模型验证和测试，包括对模型的鲁棒性、可靠性和安全性进行评估，以确保模型在不同场景下的良好性能和安全行为。

④防御对抗样本攻击：针对对抗样本攻击，采用对抗样本检测和防御策略，提高模型的抗干扰能力。

⑤隐私保护：在模型开发过程中采取隐私保护措施，例如数据匿名化、差分隐私技术等，确保用户隐私的安全性和保密性。

⑥模型监测与更新：持续监测和评估模型在实际运行中的性能和安全性，及时发现和修复潜在的安全问题，并确保模型的及时更新和适应安全威胁的变化。

⑦健壮性测试：进行全面的健壮性测试，模拟各种异常情况和攻击场景，评估模型在非预期输入下的安全性和鲁棒性。

⑧文档和沟通：编写详细的文档记录模型开发过程和 Related 安全措施，同时与团队成员和利益相关者进行有效沟通，确保安全要求被理解和遵守。

5、集成与测试

在完成人工智能模型开发后，需要进行相关功能的集成与测试工作，其目的是验证人工智能系统在实际环境中的集成情况，评估各级别的系统协同工作的安全性，并发现潜在的安全风险和问题，以确保整体系统的安全性和可靠性。推荐可以按照集成的复杂度，分别从人工智能模型集成与测试、系统集成和整车集成的维度进行相关人工智能安全测试，建议步骤如下：

①人工智能模型集成与测试：该步骤包括测试计划和范围的确定，数据准备与预处理，模型集成与接口测试，功能测试与性能评估，安全测试与漏洞分析，结果分析与问题修复，测试报告和文档编写以及反馈和改进。

②系统集成与测试：该步骤包括测试计划和范围的确定，测试数据集的准备，组件级别和系统级别的集成测试，系统性能和安全性的评估，测试结果分析与问题修复，测试报告和文档编写。

③整车集成与测试：该步骤包括人工智能模型与整车系统的集成，功能性测试和安全性评估，结果总结、问题分析与改进建议提出，整车集成测试报告编写。

6、人工智能安全确认和模型部署发布

在完成车用人工智能模型的集成与测试后，接下来是人工智能安全确认和模型的部署发布过程，为保障车用人工智能全周期的安全，建议执行以下步骤：

①人工智能安全确认：该步骤包括验证模型是否符合车用人工智能安全要求，评估其在面对潜在安全风险和威胁时的响应能力。安全审查和评估，识别潜在的安全漏洞、弱点或其他风险，制定相应的修复方案。

②模型部署准备：该步骤包括模型部署环境的准备，配置适当的硬件和软件资源，确保模型能够有效运行。确定模型的部署策略和流程，包括模型的打包格式、版本管理、更新机制等。

③模型部署和发布：该步骤包括将安全确认的人工智能模型部署到目标系统中，在此过程中，需要确保模型的安全性，防止未经授权的访问和数据泄露。在部署完成后，需要监控其性能和稳定性。最后，需要确保模型的部署与整车系统的协同工作，不影响车辆的正常运行和安全性能。

④安全性监测与更新：该步骤包括安全性监测机制的建立，定期检查模型的安全性能和响应能力，包括实施入侵检测、安全日志监控等措施，以及建立安全漏洞披露和修复机制。同时，需要及时应对新的安全威胁和攻击，通过更新模型、修复漏洞等方式来提高车用人工智能系统的安全性能。

⑤持续改进和优化：该步骤包括用户反馈和使用数据的收集，不断评估模型的性能和安全性，以进行持续的改进和优化。

7、使用维护

在完成人工智能安全确认和模型部署发布后，车用人工智能的使

用维护是至关重要的，以确保满足车用人工智能的安全要求，推荐采取如图 13 所示的措施：



图 13 人工智能使用维护过程推荐措施示意图

①建立定期监测机制：对车用人工智能系统的运行状态进行实时监控和评估，通过检查日志记录、异常报告等信息，及时发现潜在的问题和风险。

②建立安全漏洞管理流程：建立安全漏洞跟踪、修复和验证流程，确保已知安全漏洞得到妥善处理，提升车用人工智能系统的安全性能。

③数据质量维护：定期评估和清理数据，减少数据偏差和不准确性对人工智能系统的影响。加强数据隐私保护的管理，符合相关法规和隐私政策。

④更新与优化：根据用户反馈和系统需求，定期更新和优化人工智能模型，提升系统的性能和安全性。

⑤培训安全意识：为相关人员提供培训，增加他们对车用人工智能安全性的认识和理解。加强安全意识宣传和培养，提高用户和操作人员的安全操作和使用观念。

⑥安全审计和合规性检查：定期进行安全审计和合规性检查，评估系统的合规性，包括数据隐私保护、使用许可和法律法规遵从性等。

8、安全销毁

考虑人工智能应用完毕后的销毁安全，在人工智能模型销毁过程中，确保安全性的关键操作包括：彻底清除敏感数据；从所有存储介质中删除模型；撤销访问权限并适当处理硬件设备；记录销毁过程中的详细步骤和相关文件；及时终止与第三方合作伙伴的合同并明确约定责任和义务。

通过执行这些操作，能够最大程度地保护人工智能模型在销毁过程中的安全性，防止数据泄露、保护知识产权，并符合法律和合规要求。

（四）供应链管理

供应链管理也是车用人工智能风险管理中至关重要的一环，其涉及对工具链和供应商的有效管理。工具链管理确保开发过程中使用的软件工具等的质量和安全，而供应商管理则确保从供应商获取服务的可靠性和合规性。

1、工具链管理

工具链管理涉及对数据库管理、第三方依赖库和开发工具的管理

理。在管理时，建议进行以下考虑：

①数据库管理：确保数据库的安全性、稳定性和合规性，包括自建数据库和开源数据库的管理。第一，需要实施适当的访问控制措施，限制对数据库的访问权限。第二，对自建数据库进行合规性审查，要求遵守相关法律法规和隐私保护规定。对开源数据库，需要评估其可靠性和质量，选择受信任且经过验证的开源数据集。第三，监测数据的质量和完整性，定期清洗和更新数据也是必不可少的。

②第三方依赖库管理：选择可信任的第三方依赖库，并建立管理策略。评估依赖库的质量和安全性，查看文档、用户反馈和社区支持情况。跟踪依赖库的更新和安全补丁，并定期更新使用的版本。建立版本控制系统，备份、恢复和追溯使用的依赖库。

③开发工具管理：选择符合法规和许可证要求的开发工具，并定期更新安全补丁以防止潜在威胁。制定备选方案和准备替代工具，以确保在其失效或禁用状态下，开发过程能继续进行。

2、供应商管理

在车用人工智能安全风险管理过程中，对人工智能相关产品供应商的安全管理也是极其重要的。为了进行全面的供应商风险管理，建议考虑并采取如下行动：

①安全能力评估：评估供应商在数据隐私保护、模型验证与认证、漏洞管理等方面的能力，并确保其具备足够的技术能力和安全意识来保障车用人工智能系统的安全性。

②安全开发管控：与供应商明确人工智能安全要求并建立合同条款，确保供应商遵循最佳实践和安全标准，同时要求供应商进行安全测试和验证。

③应急预案：内部制定应急预案，以确保当发生风险时，能与供应商及时磋商并解决问题，同时具备替代解决方案。

④安全职责管理：明确供应商的人工智能安全职责，建立沟通渠道，促进供应商与企业内部安全团队之间的交流和协作，确保供应商理解并履行其在人工智能安全管理中的职责。

⑤风险监控：建立与供应商的风险监控机制，持续评估和监控供应商的安全状况和可靠性，定期审查合规性和安全控制措施，并采取必要的纠正措施来减轻风险。

通过以上操作方法，能够有效管理供应商的人工智能安全，确保其符合企业的安全要求，从而保障车用人工智能系统的整体安全性。

第六章 总结展望

车用人工智能已成为推动行业技术创新和业务变革的重要力量，但其带来的风险也逐渐显现。本白皮书在梳理现有的国际和国内标准法规及其他行业的风险管理案例基础上，通过全面分析智能网联汽车在内生和应用安全方面的潜在威胁，识别了包括数据隐私、算法安全、系统可靠性等关键风险点，并构建了系统性的风险管理方案，包括风险应对技术方案和风险管理方案。其中，风险应对技术方案从“技术”视角给出应对风险的解决方法，风险管理方案从“人”的视角分层次提出了应对管理措施，为汽车行业建立有效的人工智能安全体系提供参考。

未来，随着人工智能技术的深入发展，车用人工智能安全问题将更加复杂。汽车行业需要加强与其它行业的合作，尤其是在信息技术、网络安全等领域，共同应对新技术带来的挑战，推动法规和标准的完善。通过建立更加完善的人工智能风险管理体系，加速汽车行业在保障安全、合规的前提下的智能化转型，为构建更高效、安全、智能的未来出行生态奠定基础。

汽车行业发展日新月异，新的挑战和安全问题层出不穷，本书在内容深度和广度方面难免有所不足，欢迎各位行业专家提出宝贵的建议。我们将不断完善该白皮书，并在智能汽车安全技术国家重点实验室官网持续更新，助力汽车行业人工智能高质量安全发展。

参考文献

- [1].National Highway Traffic Safety Administration. Summary report: standing general order on crash reporting for level 2 advanced driver assistance systems[R]. National Highway Traffic Safety Administration, 2022.
- [2].自动驾驶汽车交通安全白皮书[R]. 中国汽车技术研究中心有限公司, 同济大学, 百度 Apollo, 2021.
- [3].宝马发生数据泄露事件 涉及中国、欧洲和美国三地[EB/OL]. [2024-11-09]. <https://www.auto-testing.net/news/show-120969.html>.
- [4].2023 年度安全事件观察报告[R]. 绿盟科技集团股份有限公司, 2024.
- [5].EUROPEAN COMMISSION. EU Artificial Intelligence Act[A/OL]. (2024)[2024-11-08]. <https://artificialintelligenceact.eu/ai-act-explorer/>
- [6].Eddie Bernice Johnson. National Artificial Intelligence Initiative Act of 2020[A/OL]. (2020-12-03)[2024-11-04]. <https://www.congress.gov/bill/116th-congress/house-bill/6216>.
- [7].U.S. Department of Homeland Security Artificial Intelligence Strategy[EB]. [2024].
- [8].AI regulation: a pro-innovation approach[R/OL]. (2023-08-03)[2024-11-09]. <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach>.
- [9].中央网络安全和信息化委员会办公室. 生成式人工智能服务管理暂行办法 [A/OL]. (2023-07-10)[2024-11-05]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.
- [10].中央网络安全和信息化委员会办公室. 互联网信息服务算法推荐管理规定 [A/OL]. (2022-01-04)[2024-11-07]. https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.
- [11].中华人民共和国科学技术部. 关于印发《科技伦理审查办法（试行）》的通知[A/OL]. (2023)[2024-11-07]. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgnr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html.
- [12].ISO/IEC/IEEE 29119:2022, Software and systems engineering — Software testing[S].
- [13].ISO/IEC 29134:2023, Information technology — Security techniques — Guidelines for privacy impact assessment[S].
- [14].ISO/IEC 38505:2017, Information technology — Governance of IT — Governance of data[S].
- [15].IEEE 7001:2021, IEEE Standard for Transparency of Autonomous Systems[S].
- [16].IEEE 7002:2022, IEEE Standard for Data Privacy Process[S].
- [17].IEEE 7007:2021, IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems[S].
- [18].ISO/SAE 21434:2021, Road vehicles — Cybersecurity engineering[S].

- [19].ISO 26262-1:2018, Road vehicles — Functional safety[S]. 2018.
- [20].ISO 21448:2022, Road vehicles — Safety of the intended functionality[S].
- [21].四部门关于印发国家人工智能产业综合标准化体系建设指南（2024版）的通知[EB/OL]. [2024-11-09].
https://www.gov.cn/zhengce/zhengceku/202407/content_6960720.htm.
- [22].GB/T 41867-2022, 信息技术 人工智能 术语[S].
- [23].GB/T 40856-2021, 车载信息交互系统信息安全技术要求及试验方法[S].
- [24].GB/T 41871-2022, 信息安全技术 汽车数据处理安全要求[S].
- [25].GB/T 40861-2021 汽车信息安全通用技术要求[S].
- [26].全国网络安全标准化技术委员会. TC260-003 生成式人工智能服务安全基本要求[A]. 2024.
- [27].GB/T 36464.5-2018, 信息技术 智能语音交互系统 第5部分：车载终端[S].
- [28].人工智能助金融行业转型发展[EB/OL]. [2024-11-09].
<https://www.businessgo.hsbc.com/zh-Hant/article/aibankingservice>.
- [29].“宇宙药企”辉瑞，如何拥抱 AI 和数字化？[EB/OL]. [2024-11-09].
<https://zhuanlan.zhihu.com/p/534259356>.
- [30].亿邦动力网. 亚马逊掀起算法革命：A9 已死 COSMO 当立？[EB/OL]. (2024-03-29)[2024-11-09]. <https://www.ebrun.com/20240329/544927.shtml>.
- [31].HASTIE T J. Generalized additive models[M]//Statistical models in S. Routledge, 2017: 249-307.
- [32].SAE J3061:2016, Cybersecurity Guidebook for Cyber-Physical Vehicle Systems[S].
- [33].GB/T 38628-2020, 信息安全技术 汽车电子系统网络安全指南[S].
- [34].ISO 9001:2015, Quality management systems — Requirements[S].
- [35].ATF 16949:2016, Quality management system for organizations in the automotive industry[S].
- [36].UN Regulation No. 155 - Cyber security and cyber security management system[Z]. 2021.
- [37].ISO/IEC 42001-2023, Information technology-Artificial intelligence-Management system[S].
- [38].全国网络安全标准化技术委员会. 《人工智能安全治理框架》1.0 版[EB/OL]. [2024-11-08].
<https://www.tc260.org.cn/front/postDetail.html?id=20240909102807>.