自动驾驶模型鲁棒性规范标准研究

朱微伟 樊志鹏 唐志彪
重庆长安汽车股份有限公司，中国·重庆 400023

摘 要：本研究聚焦于自动驾驶技术的鲁棒性规范与标准，旨在提升自动驾驶系统在多样化环境条件下的安全性与可靠性。面对深度学习技术在自动驾驶领域中的应用及其面临的安全性挑战，特别是在复杂系统环境因素影响下的潜在风险，论文对现行的汽车人工智能鲁棒性标准进行了系统性总结，并分析了这些标准在自动驾驶模型鲁棒性方面的指导作用及其局限性。另外，本研究进一步探讨了基于场景的自动驾驶鲁棒性验证方法，强调了对自动驾驶系统运营设计域内潜在场景变化的全面验证重要性，并提出了车企在应对鲁棒性标注法规要求时的策略，包括场景类型定义、监控机制建立以及数据共享和合作机制的构建。

关键词：自动驾驶；鲁棒性；安全性；预期功能安全；场景验证

Research on the Robustness Specification Standard of Autonomous Driving Model

Weiwei Zhu Zhipeng Fan Zhibiao Tang
Chongqing Changan Automobile Co., Ltd., Chongqing, 400023, China

Abstract: This research focuses on the robust specifications and standards of autonomous driving technology, aiming to improve the safety and reliability of autonomous driving systems under diverse environmental conditions. In the face of the application of deep learning technology in the field of autonomous driving and the safety challenges it faces, especially the potential risks under the influence of complex system environmental factors, this paper systematically summarizes the current robustness standards for automotive artificial intelligence, and analyzes the guiding role and limitations of these standards in the robustness of autonomous driving models. In addition, this study further explores the scenario-based robustness verification method of autonomous driving, emphasizes the importance of comprehensive verification of potential scene changes in the operational design domain of autonomous driving system, and puts forward the strategies of car companies in coping with the requirements of robust labeling regulations, including the definition of scenario types, the establishment of monitoring mechanisms, and the construction of data sharing and cooperation mechanisms.

Keywords: autonomous driving; robustness; security; intended functional safety; scenario verification

0 前言

在当代大数据时代背景下，深度学习领域的理论和技术已实现显著的突破，为人工智能的发展提供了坚实的数据基础与算法支持，并推动了深度学习技术的规模化与产业化进程。尽管深度学习模型在多种实际应用场景中展现出卓越的性能，但其安全性问题仍然是一个亟待解决的挑战 [1]。2023 年 11 月 1 日，首届全球人工智能（AI）安全峰会在英国布莱切利庄园召开。在该峰会的开幕式上，与会国家，包括中国在内，共同签署并发表了《布莱切利宣言》[2]。该宣言作为全球首份针对人工智能新兴技术的国际性声明，集中关注了对强大人工智能模型可能对人类生存造成的潜在威胁，以及对当前人工智能技术可能加剧的有害信息传播或偏见问题的担忧。《布莱切利宣言》的发布标志着国际社会对于人工智能安全性问题的高度重视，并为未来相关领域的研究与合作奠定了基础。

交通安全一直是车载智能系统研究的核心议题，尤其对于自动驾驶技术的快速发展而言，其安全性问题显得尤为关键 [3]。随着汽车工业向电气化转型，研究者们逐渐认识到，并非所有的交通安全问题均源于系统内部的失误或故障。在许多情况下，复杂的系统环境因素对交通安全构成了潜在的威胁，这些环境因素可能引发非预期的安全事故 [4]。

在传统汽车安全领域，系统的故障性能主要是由硬件或软件的缺陷引起的。然而，在自动驾驶系统的背景下，即便系统本身没有出现故障，由于神经网络等人工智能技术的黑箱特性，其输出结果的不确定性也可能引发交通事故，进而导致系统功能的偏差。因此，对于自动驾驶系统的控制效果，系统必须全面考量其优势与局限性，确保在各种环境条件下均能保持高度的安全性和可靠性 [5]。

论文整理了汽车人工智能领域鲁棒性相关的标准，并对现有的标准进行了系统的总结和科学的归纳，以此分析自动驾驶模型鲁棒性标准规范的未来发展趋势以及对于车企的未来可能影响。

1 预期功能安全对模型鲁棒性的整体要求

广泛使用的针对汽车电子的功能安全标准《ISO 26262 道路车辆 - 功能安全》[6]（Road vehicles - Functional safety），仅适用于缓解与已知部件故障（即已知不安全情况）相关的已知不合理风险。现行的安全法规主要基于安全分析框架，该框架倾向于采用白盒方法论。除了对软件开发生命周期进行规范性约束之外，该方法论主要通过对研究对象的结构和功能进行层次化的详细分析，以识别潜在的安全漏洞 [7]。在发现安全薄弱环节之后，相应的安全加固措施将被制定并实施。这种方法论强调对系统内部工作机制的深入理解，以及在此基础上的安全策略构建。然而，自动驾驶是一个黑盒系统，面对复杂的场景的处理需求，系统功能的实现容易受到不同程度的限制（如感知能力缺陷会导致不正确的分类、不正确的测量、不正确的跟踪等 [8]），因此需要在确保系统整体安全性的同时，考虑到这些潜在的功能性限制，并制定相应的风险缓解措施。这包括但不限于增强系统的冗余性、鲁棒性测试以及实施动态安全监控等策略，以确保自动驾驶系统在面对各种预期和非预期情况时，仍能保持必要的安全性能水平（见图 1）。

With the trend of increasing technological complexity, software content and mechatronic implementation, there are increasing risks from systematic failures and random hardware failures, these being considered within the scope of functional safety. ISO 26262 series of standards includes guidance to mitigate these risks by providing appropriate requirements and processes.

图 1 ISO 26262 标准中针对功能安全的描述

鉴于上述安全挑战，《ISO 21448 道路车辆预期功能安全》[9] 提出了一个定性目标，该目标可被描述为：将自动驾驶功能设计的已知和未知不安全场景结果最小化，称为 SOTIF（预期功能安全，safety of the intended functionality）。预期功能安全考虑了针对自动驾驶车辆系统非故障原因导致的危害，并考虑人员不能及时响应。预期功能安全期望从系统层面做到功能设计完备（如果发生意外，如何处理）。

ISO 21448 同样没有直接定义自动驾驶模型需要满足的鲁棒性需求，难以直接通过该标准指导模型鲁棒性工作。然而，其对于自动驾驶系统的部分描述可以映射到自动驾驶感知模型中。整体而言，该标准从以下方向提出针对自动驾驶模型鲁棒性的要求：

性能一致性：模型应该在各种条件下保持一致的性能，如不同的天气条件、照明和交通情况。

数据和场景覆盖：确保模型的训练数据全面覆盖现实场景，包括罕见或边缘情况，防止因意外情况导致系统故障。

验证和测试：模型应实现严格的验证和测试协议，包括基于模拟的测试和现实世界的操作测试。

故障检测和管理：人工智能系统检测并适当响应运行中的故障或异常的能力。

然而，该标准实际针对的是整体自动驾驶系统的安全性，针对模型鲁棒性要求部分并无直观、落地的相关描述，相关主机厂难以仅通过该标准构建对应的鲁棒性测试。因此，为了满足模型鲁棒性的具体要求，需要详尽的测试协议和评估方法。

2 基于场景的自动驾驶鲁棒性验证

自动驾驶模型因其快速发展和定义上的挑战性而独具特性。在模型鲁棒性的验证过程中，通常是基于道路场景来进行的。根据 ISO 21448 标准，场景定义为一系列情景在时间轴上的连续发展，而情景则被描述为特定时间点的环境静态快照，包括环境景观、动态元素、所有参与者以及观察者的内在表征，以及这些实体间的相互关系。

在确保自动驾驶系统在其运营设计域（Operational Design Domain，ODD）内实现全面安全的任务中，可以将其阐述为对 ODD 内所有潜在场景变化进行全面的安全验证与确认（Verification & Validation，V&V）。ODD 内的场景变化主要涵盖两个维度：初始情景的多样性以及从初始情景开始随时间推移的场景演变。因此，自动驾驶模型的鲁棒性验证需综合考虑 ODD 内所有可能的情景及其随时间的演变，确保系统在各种动态变化中的安全性能得到充分的保障。

一些现行标准沿用了上述定义，并基于经验设定了一系列自动驾驶模型的标准测试场景。例如，IEEE 标准《2846-2022 for Assumptions in Safety-Related Models for Automated Driving Systems》[10] 规定了在开发自动驾驶系统（ADS）中的安全相关模型时，应考虑的一组基本的合理假设和可预见的场景。此外，联合国 R157《自动车道保持系统（ALKS）》[11] 明确了在高速公路上进行点到点自动驾驶的场景下，对于自动驾驶系统的具体要求。这些标准为自动驾驶系统的安全性提供了一个结构化的框架，并确保了在开发过程中对关键的安全因素进行了充分的考虑（见表 1）。

实现预期功能安全（SOTIF）的目标面临着显著的挑战，尤其是在采用传统测试方法，如现场操作测试时，这些方法往往难以全面覆盖自动驾驶系统在实际运行中可能遇到的所有场景。尽管面临这些挑战，但在安全分析领域，存在一些具有潜力的方法和研究方向，它们有望推动实现 ISO 21448 标准的目标。例如，通过在模拟测试环境中系统性地扩展场景覆盖率，可以更全面地评估自动驾驶系统的行为。此外，运用形式化方法来构建严格的安全保证，为系统提供数学上的安全性证明，也是一种可行的途径。这些方法的综合应用，有助于提升自动驾驶系统的安全性，确保其在多样化的交通环境中的可靠性和鲁棒性。

表 1 R157 中定义的自动车道保持系统测试场景

序号	类型	场景类型	强制 / 推荐
1	系统超出其技术边界时防止激活	在一段不合适的高速公路上	强制
		城市环境中	强制
		其他条件（天气 / 时间）不合适的合适道路上	推荐
2	系统接管	方向盘干预	强制
		加速踏板干预	强制
		制动踏板干预	强制
3	不违反交通规则时	达到限速	强制
		60km/h 以上反复改变速度限值	强制
		接触需要系统反应的不同路标	强制
		禁止变道时不能压实线	推荐
		与前车有足够的间距	强制
4	对道路事件的响应	隧道	推荐
		高速终点	推荐
		工作区域	推荐
		收费站	推荐
		封闭道路的反应	推荐
		应急车辆靠近	推荐
		环境条件变化	推荐

在技术术语中，“足够安全（safe enough）”的概念通常与指定运行设计域内的全面或充分的情境覆盖范围相联系。实际上，现行法规的要求相对宽松，往往仅需要对“某些关键场景”进行验证。在自动驾驶车辆的安全验证中，形式方法作为一种常用的技术手段，包括模型检验、可达性分析和定理证明等。模型检验起源于软件开发领域，其目的在于确保软件的行为严格遵循设计规范。当安全规范以公理和引理的形式被表述时，定理证明方法可用于验证在最坏情况假设下系统的安全性。在这些形式分析方法中，可达性分析占据着特殊的地位，因为它具备为动态系统生成安全断言的内在能力，能够有效捕捉动态驾驶任务（Dynamic Driving Task，DDT）的核心特征。通过这些方法的应用，可以系统性地评估和增强自动驾驶系统在面对各种交通情境时的安全性和可靠性。

鉴于基于真实世界的道路测试或现场操作测试的成本较高，目前存在一种基于形式化方法构建部分测试场景以评估系统性能的途径。该方法通过模拟车辆的关键性能指标，如速度和距离，并对其进行参数化处理，以构建测试场景并检验系统的安全性。这种方法依赖于特定的规则或随机过程来生成场景。

然而，从整体上看，这种场景构建方法存在若干局限性：

形式化方法本质上是对现实世界的简化抽象，而模型不可能涵盖实际系统运行中的所有参数，因此模型与现实之间不可避免地存在差异。

实际开发的自动驾驶控制器往往具有一定的性能局限，这可能限制了系统对复杂驾驶情境的适应性和响应能力。

驾驶本质上是一个连续的动态过程，当需要对连续的动作空间进行细致的离散化时，穷举所有可能的测试情况变得几乎不可能，即存在所谓的“穿透率问题”。

因此，为了提高自动驾驶系统的安全性，需要进一步研究和开发更为全面和精确的测试方法，以及对系统性能进行更为严格的评估和验证。这可能包括采用更高级的模拟和仿真技术，以及结合机器学习和人工智能算法来预测和处理潜在的安全风险。同时，也需要不断优化和完善形式化方法，以减少模型与现实之间的差异，并提高对连续动态系统的整体覆盖率。

3 车企应对鲁棒性标注法规要求的应对措施

基于 ISO 21448 标准定义的预期功能安全，车企验证自动驾驶模型的鲁棒性时，需根据自动驾驶系统的设计运行域（ODD），定义可能的场景类型，包括城市道路、高速公路、乡村道路等。目前，如 PEGASUS 项目或 ISO34503 标准中规定的自动驾驶场景库构建了标准化方法对 ODD 进行分类，包括静态要素（如道路结构）、动态要素（如交通参与者）和辅助要素（如网联通信）。另外，基于实际场景数据及实际场景数据的仿真模拟同样可有效验证模型鲁棒性，现有自动驾驶数据集（如 Waymo Open Dataset、KITTI Dataset 和 nuScenes Dataset）包含了多种驾驶条件下的数据，如不同天气、光照条件和交通流量。车企可基于以上数据，使用仿真平台进行场景模拟，如 Siemens 提出的虚拟验证和确认方法，确保测试场景、环境、汽车和传感器的虚拟表示一致性。通过仿真测试，可以覆盖大量的边缘情况和极端条件，如极端天气、复杂交通场景等，这些在现实世界中难以复现或风险过高。

另外，车企应建立有效的监控机制以持续收集自动驾驶系统在实际运行中的数据，实际驾驶中产生的真实数据提升系统性能和应对新挑战的关键。车企应建立一个能够记录、存储和分析车辆数据的系统。这个系统需要能够确保数据的完整性和安全性，防止数据被篡改或损坏。例如，公安部发布的《机动车运行安全技术条件》要求从 2022 年 1 月起，L2 级以下乘用车配备视觉数据记录系统。

此外，自动驾驶模型鲁棒性是整个行业面临的共性问题，建立数据共享和合作机制势在必行，对于推动自动驾驶技术的进步和应对行业共性问题具有重要意义。通过共享数据，各企业能够获得更多样化和丰富的实际驾驶场景数据，这有助于发现和解决在不同环境和条件下自动驾驶系统可能遇到的问题。此外，合作机制可以促进资源和知识的交流，加速自动驾驶技术的研发和创新，共同提升整个行业的技术水平和竞争力。该数据合作机制应明确数据的交换、共享和使用要求，确保各方数据的合规性。

4 结语

论文基于自动驾驶技术的发展背景，深入探讨了自动驾驶系统在面对复杂环境因素时的潜在安全威胁，并分析了

传统汽车安全领域的方法在自动驾驶系统中的应用局限性。论文整理并分析了当前汽车人工智能领域的鲁棒性相关标准，特别是《ISO 26262》和《ISO 21448》标准对自动驾驶模型鲁棒性的要求，并指出现有标准在具体实施上的不足。在此基础上，论文提出了基于场景的自动驾驶鲁棒性验证方法，强调了对自动驾驶系统运营设计域内所有潜在场景变化进行全面验证的必要性。进一步地，文档讨论了车企在应对鲁棒性标注法规要求时的应对措施，包括定义可能的场景类型、建立有效的监控机制以及建立数据共享和合作机制。这些措施旨在提高自动驾驶系统的性能，确保其在多样化的交通环境中的可靠性和鲁棒性。

参考文献：

[1] 纪守领,杜天宇,邓水光,等.深度学习模型鲁棒性研究综述[J].计算机学报,2022,45(1):190-206.

[2] 王威.《布莱奇利宣言》:人工智能国际合作监管的新起点[J].服务外包,2023(12):46-52.

[3] 郭延永,刘佩,袁泉,等.网联自动驾驶车辆道路交通安全研究综述[J].交通运输工程学报,2023,23(5):19-38.

[4] 徐谦.面向复杂环境自动驾驶的视觉环境感知研究[D].长春:吉林大学,2023.

[5] Liu W, Hua M, Deng Z, et al. A systematic survey of control techniques and applications in connected and automated vehicles[J]. IEEE Internet of Things Journal,2023.

[6] GB/T 34590.1-2022,道路车辆-功能安全 第1部分:术语[S].

[7] 郝东辉,吴向亮,吕平,等.道路车辆功能安全风险与控制研究[J].机械工业标准化与质量,2023(11):28-34.

[8] Zhang X, Tao J, Tan K, et al. Finding critical scenarios for automated driving systems: A systematic literature review[J]. arXiv preprint arXiv,2110,8(664):2021.

[9] GB/T 43267-2023,道路车辆-预期功能安全[S].

[10] IEEE Std 2846-2022, IEEE Standard for Assumptions in Safety-Related Models for Automated Driving Systems[S].

[11] 周紫君,刘思,范煜君.联合国L3级自动驾驶新规要点及对我国交通运输行业的影响分析[J].交通世界,2020(22):4.

作者简介：朱微伟（1992-），男，中国重庆人，本科，工程师，从事智能网联汽车系统集成与应用测试研究。