ICS 35.120.01
CCS L 05

T/CICC

中 国 指 挥 与 控 制 学 会 团 体 标 准

T/ CICC 35018—2025

复杂智能系统功能安全性技术要求

Technical requirements for functional safety of complex intelligent systems

2025－11－20 发布　　2025－11－20 实施

中国指挥与控制学会　发 布

目 次

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则　第 1 部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国指挥与控制学会提出并归口。

本文件起草参与单位：北京航空航天大学、杭州市北京航空航天大学国际创新研究院（北京航空航天大学国际创新学院）、中国船舶集团有限公司综合技术经济研究院、中国兵器工业软件工程与评测中心、中国电子科技集团公司信息科学研究院、可靠性与环境工程技术国家级重点实验室、北京航空航天大学可靠性工程研究所、中国航空综合技术研究所、中国航空研究院。

本文件主要起草人：杨顺昆、王英凡、徐珞、吴梦丹、邢晨光、刘虹晓、王若、庞红彪、彭文胜、王树泰、林聪、李汉智、高小泉、曾子鸣、张自超、张林超、姜巍、许丹、黄婷婷、刘杰。

复杂智能系统功能安全性技术要求

1 范围

本文件规定了面向复杂智能系统的功能安全性技术要求，描述了复杂智能系统通用质量特性中有关功能安全性方面的相关指标以及测试方法。

本文件适用于面向复杂智能系统的功能安全性预估、设计和测试。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 45654－2025 网络安全技术 生成式人工智能服务安全基本要求

GB/T 45674－2025 网络安全技术 生成式人工智能数据标注安全规范

GB/T 45958—2025 网络安全技术 人工智能计算平台安全框架

GB/T 42018—2022 人工智能平台 计算资源规范

3 术语和定义

GB/T 45654－2025、GB/T 45674－2025、GB/T 45958—2025和 GB/T 42018—2022 确立的以及下列术语和定义适用于本文件。

3.1.

复杂智能系统 Complex Intelligent System

由感知、认知、决策与执行等功能模块构成，采用机器学习等方法，在不确定、开放环境下执行任务的人机环管协同系统。其复杂性体现在多源异构数据、动态场景、要素耦合以及全生命周期演化。

3.2.

功能安全 Functional Safety

系统或设备作为一个整体，其安全相关部分能够正确执行其安全功能，以避免或减轻由其控制的设备或系统失效所导致的、对人员健康、环境或财产构成不可接受的风险。在智能系统中，特指 AI 组件及其支持系统在面对潜在危险时，能够可靠地执行预定功能，防止系统进入不安全状态的能力。

3.3.

危害分析与风险评估 Hazard Analysis and Risk Assessment

在系统概念阶段进行的一套系统性分析流程，旨在识别由功能异常可能引发的所有危害，并根据其严重性、暴露率和可控性来评估风险，最终为每个危害分配一个安全完整性等级。

3.4.

安全目标 Safety Goal

由HARA推导出的、最高层级的安全需求，旨在规避某个已识别的、不可接受的风险。

3.5.

触发条件 Triggering Condition

在一个特定场景中，能够引发系统功能不足并可能导致危险行为的一个或多个特定条件或事件。

3.6.

运行设计域 Operational Design Domain

明确规定了智能系统设计时预期能够安全运行的特定条件集合。这些条件包括但不限于环境因素、道路特征、交通状况以及系统自身的运行状态。任何超出 ODD 的情况都可能导致系统性能下降或行为不可预测，进而引发安全风险。

3.7.

AI 安全生命周期　　AI Safety Lifecycle

一个覆盖了从概念构思、风险评估、需求定义、数据工程、模型开发、验证与确认、部署运行、维护更新直至最终退役的全过程的结构化工程流程。该流程旨在系统性地管理与 AI 功能相关的安全风险，是传统安全生命周期在 AI 领域的延伸和调整。

3.8.

不确定性　　Uncertainty

指 AI 模型对其预测、决策或感知结果的置信度不足。不确定性分为偶然不确定性（数据固有噪声导致）和认知不确定性（模型对未见过或 ODD 边缘数据的认知不足导致）。对不确定性的有效量化与管理是确保功能安全的关键。

3.9.

鲁棒性　　Robustness

指智能系统在面临输入数据的微小扰动、噪声、对抗性攻击或非预期变化时，仍能保持其功能和性能稳定性的能力。高鲁棒性是防止系统因外部干扰而产生危险行为的重要保障。

3.10.

可解释性　　Interpretability

智能系统为其决策或输出提供人类可以理解的解释或理由的能力。在功能安全领域，可解释性对于理解模型行为、诊断潜在故障、建立信任以及在事故发生后进行追责至关重要。

3.11.

数据驱动的危险源　　Data-driven Hazard

源于用于训练、测试或操作 AI 系统的数据本身所引入的危险。具体表现包括数据偏见（Bias）、数据不完备、以及数据投毒（Poisoning）攻击等。

3.12.

模型漂移　　Model Drift

指已部署的 AI 模型因现实世界的数据分布随时间发生变化，而导致其性能逐渐下降的现象。未能及时检测和处理模型漂移是运行阶段一个重要的安全风险来源。

3.13.

安全档案　　Safety Case

一份结构化的论证，通过汇集所有相关的证据、分析和论点，清晰、全面、可信地证明智能系统在其指定的运行环境中对于特定的应用是可接受地安全的。它是向监管机构或认证方证明系统安全性的核心文件。

4　缩略语

下列缩略语适用于本文件。

AI	人工智能 (Artificial Intelligence)
FS	功能安全 (Functional Safety)
ODD	运行设计域 (Operational Design Domain)
SIL	安全完整性等级 (Safety Integrity Level)

HARA	危险源分析与风险评估 (Hazard Analysis and Risk Assessment)
V&V	验证与确认 (Verification and Validation)
XAI	可解释人工智能 (Explainable Artificial Intelligence)
ML	机器学习 (Machine Learning)
SPFM	单点故障度量 (Single-Point Fault Metric)
LFM	潜伏故障度量(Latent Fault Metric)
PMHF	小时危险失效率（Probabilistic Metric for Hardware Failure）
TTSS	安全状态转换时间（Time to Safe State）
ECE	期望校准误差(Expected Calibration Error)
MSE	均方误差（Mean Square Error）
MAE	平均绝对误差（Mean Absolute Error）
SIL	软件在环 (Software-in-the-Loop)
HIL	硬件在环 (Hardware-in-the-Loop)

5 复杂智能系统功能安全性适用对象

5.1. 按系统功能角色划分

5.1.1. 感知模型

感知模型是智能系统与环境的接口，核心是将原始高维传感器数据（如图像、点云等）转化为环境结构化理解（如目标识别定位、场景语义分割），是后续安全决策的基础；其失效或性能下降会向决策系统提供缺陷甚至虚假的环境模型。主要安全挑战在于应对开放世界的无穷变化与传感器物理局限，典型失效模式包括漏检、误检、错误分类及对物体位置、尺寸或速度的估计不准确。

5.1.2. 决策与规划模型

决策与规划模型接收感知和预测模型的结构化信息，依据效率、舒适性、安全性等预定目标，确定系统高层级行为策略或具体时空轨迹；其直接决定系统行为意图，即便感知信息完美，有缺陷的决策模型仍可能做出危险决策，典型失效模式包括不安全策略选择、决策冻结、违反交通法规，以及规划出不舒适或不自然的轨迹。

5.1.3. 预测模型

预测模型负责预测环境中其他动态参与者在未来一段时间内的意图和轨迹。预测错误可能导致决策模型基于错误信息做出规划，从而将系统引导至危险的状态。其主要挑战在于长期预测的不确定性和对罕见但关键的人类行为的建模。典型失效模式包括：意图预测错误、轨迹预测不准确、以及未能预测到参与者的突然行为。

5.1.4. 生成模型

以大语言模型（LLM）和扩散模型为代表的生成式模型，是一种生成文本、图像、代码等信息内容而非直接控制物理实体的模型。其安全内涵指防止信息危害、心理伤害和社会危害，主要风险包括生成有害内容、传播虚假信息、放大社会偏见、泄露隐私数据和被恶意使用。该模型的核心挑战在于事实性一致、

价值对齐和输出的可控性，典型失效模式包括事实性幻觉、有害内容生成、指令注入或越狱，以及隐私泄露。

5.2. 按底层计算架构划分

5.2.1. 前馈神经网络

前馈神经网络是一种信息从输入端单向分层流向输出端，且不存在环路的网络结构。以卷积神经网络（CNN）为代表，其通过局部感受野和参数共享提取空间层次化特征。该类网络的主要挑战是对像素级扰动和几何变换的敏感性，以及存在学习虚假统计相关性的倾向。其典型失效模式包括对抗性脆弱、对分布外物体的错误泛化及依赖虚假相关性。

5.2.2. 循环神经网络

以长短期记忆网络（LSTM）和门控循环单元（GRU）为代表的循环神经网络，其核心特征是通过内部循环结构和门控机制处理序列及时间依赖性数据，利用历史信息影响当前输出。该网络的主要挑战在于长期依赖的稳定性和对时序扰动的鲁棒性，其典型失效模式包括梯度消失或爆炸、对时间维度扰动的脆弱性以及灾难性遗忘。

5.2.3. 基于注意力的网络

以Transformer模型为代表，其核心是依赖自注意力（Self-Attention）机制并行捕捉输入序列中的长距离和全局上下文依赖关系。该模型的主要挑战在于对上下文的鲁棒性和事实性，其典型失效模式包括上下文操控与提示注入、事实性幻觉以及注意力分散。

5.2.4. 图神经网络

图神经网络（GNN）是一种通过在节点间传递与聚合信息（消息传递机制）来学习节点表示，以处理图结构数据的网络。该网络的主要挑战在于对图结构扰动的脆弱性，其典型失效模式包括结构性对抗攻击、过平滑以及对分布外图结构的泛化能力差。

5.2.5. 其他架构

a） 能量基础模型：EBM不直接学习概率分布，而是学习一个能量函数，为每一个输入、输出对分配一个能量值，能量越低表示该对的可能性越高。
b） 胶囊网络：胶囊网络用一组神经元的向量输出（胶囊）来表示实体的各种属性（如姿态、纹理），并通过一种称为“动态路由”的机制来学习部分与整体之间的关系。
c） 神经辐射场：神经辐射场是一种用于新视角合成的神经表示方法。它使用一个简单的多层感知机来学习一个连续的、五维的函数，该函数将一个三维空间坐标和两个视角方向映射到一个体积密度和颜色值。通过对这个函数沿相机光线进行积分，可以渲染出任意新视角下的逼真图像。

5.3. 按数据处理模态划分

5.3.1. 图像模态

图像模态模型处理以像素网格表示的视觉信息，是物理世界智能系统的核心输入，其感知可靠性直接关联功能安全与物理安全。该模态的主要挑战为高维空间的脆弱性与对环境变化的敏感性，其典型失效模式包括对抗性脆弱、在不利环境条件下的性能下降，以及漏检、误检或错误分类等感知错误。

5.3.2. 文本模态

文本模态模型处理以离散符号序列表示的自然语言，是信息交互系统的核心，其功能安全涵盖信息、心理及社会危害。该模态的主要挑战在于语义的模糊性、上下文的复杂性以及数据中内嵌的社会偏见。其典型失效模式包括事实性幻觉、有害内容生成、指令注入以及偏见放大。

5.3.3. 音频模态

语音/音频模态模型处理以声波波形表示的音频信号，用于语音识别与声控命令，其功能安全在人机交互中至关重要，因指令识别错误可导致系统失效或物理伤害。该模态的主要挑战在于信号的信噪比和多样性，其典型失效模式包括嘈杂环境下的识别率下降、对不同口音和语速的泛化能力不足、对发音相似但意图不同指令的区分错误，以及对声学对抗攻击的脆弱性。

5.3.4. 表格与时间序列模态

表格与时间序列模态模型处理结构化特征或按时间顺序排列的数据点，是决策支持系统的基础，其功能安全直接影响高风险决策的可靠性。该模态的主要挑战在于数据分布的非平稳性与特征的因果关系，其典型失效模式包括数据分布漂移、特征对抗性操纵，以及模型因依赖虚假相关性而做出错误判断。

6 功能安全性指标体系

6. 1. 系统级功能安全指标体系

本部分指标评估承载智能算法的硬件/软件平台的底层安全性，确保其能够为上层智能算法提供一个稳定可靠的运行基础。

6.1.1. 定性指标

6.1.1.1. 危害分析与风险评估完备性

该指标用于评估是否系统性、无遗漏地识别出了所有潜在危害，并根据严重性、暴露率和可控性，正确地为其分配了安全完整性等级。具体而言，通过安全审核和评估，审查HARA文档的系统边界、危害识别方法的系统性、情景分析的合理性以及安全完整性等级推导的合规性。

6.1.1.2. 安全计划适宜性

该指标通过评审与审核，评估安全计划在定义必要活动、职责、流程、方法及交付物方面的完备性，并验证项目全生命周期中各项活动对该计划的遵循度，以识别不符合项。

6.1.2. 定量指标

6.1.2.1. 潜伏故障度量 (Latent Fault Metric, LFM)

这一指标衡量了安全机制发现并控制潜伏故障的能力。潜伏故障是一种特殊的双点故障中的第一个故障，它在发生时不被标准操作或系统察觉，直到第二个故障发生时，两者共同作用才会导致安全目标的违反。该指标具体形式如公式 (1)。

$$\mathrm{LFM} = \frac{\sum \lambda_{LF,det}}{\sum \lambda_{LF}} \tag{1}$$

式中：

λ_{LF}—— 潜伏故障的总失效率；

$\lambda_{LF,det}$—— 被诊断机制所能检测到的潜伏故障的失效率。

6.1.2.2. 小时危险失效率（Probabilistic Metric for Hardware Failure, PMHF）

小时危险失效率是对系统因硬件随机失效而导致违反安全目标的总体风险的最终量化，代表系统在平均每个小时内发生危险失效的可能性，用于评估系统的最终硬件安全完整性。该指标具体形式如公式 (2)。

$$\mathrm{PFH} = \lambda_{SPF} + \lambda_{RF} + \lambda_{DPF} \tag{2}$$

式中：

λ_{SPF}—— 未被安全机制覆盖的单点故障的残余失效率；

λ_{RF}—— 被安全机制覆盖但仍可能发生的残余故障失效率；

λ_{DPF}—— 由潜伏故障导致的双点危险故障失效率。

6. 2. 算法级功能安全指标体系

6.2.1. 定性指标

6.2.1.1. 模型文档完备性

模型文档完备性评估是一项用于衡量智能模型配套文档质量的指标。该指标通过内容审核方法并依据标准化框架，系统性地检验文档是否完整、清晰、准确地记录了模型的各项关键信息，包括其预期用途、已知限制、训练数据概述、在不同数据子集上的性能表现，以及公平性与鲁棒性评估结果。

6.2.1.2. 伦理审查与红队测试流程严谨性

该指标评估了在模型开发过程中，是否建立并执行了一个正式的、跨学科的伦理审查流程，以及一个独立的“红队”测试流程，旨在主动地、系统性地识别和评估模型的潜在危害，如偏见、滥用潜力和安全漏洞。具体而言，通过流程审核进行，审查相关的流程文档、会议记录和测试报告，评估其参与人员的跨学科性、审查范围的全面性、以及对所发现问题的跟踪和闭环管理。

6.2.1.3. 可信度验证流严谨性

此指标评估用于测量可信度指标的测试流程是否严谨、系统化且具有挑战性。评估内容包括：对抗性攻击是否采用了当前公认的强力算法；OOD数据集是否多样化且与目标域有显著差异；公平性评估是否覆盖了所有相关的受保护属性。其评估方法是通过技术评审进行，评审专家将审查验证计划和测试报告，评估测试设置的合理性、测试用例的选择依据，以及结果分析的科学性，以判断其结论的可靠程度。

6.2.1.4. 可解释性分析充分性

此指标评估是否对模型的关键决策进行了充分的可解释性分析，以验证其决策逻辑是否符合领域知识和因果关系，而非依赖于数据中的虚假相关性。其评估方法是通过技术评审进行，评审专家将审查分析报告，评估所使用的解释方法是否适宜，以及从解释结果中得出的关于模型行为的结论是否合理。

6.2.2. 定量指标

6.2.2.1. 最小对抗扰动

寻找并量化能够欺骗智能系统或模型使其做出错误判断所需的最小输入修改量，用于评估模型的对抗攻击鲁棒性。所需的扰动越大，说明模型对输入的微小变化越不敏感，其决策边界越平滑、稳定，也就越难以被恶意攻击者利用。具体形式如公式 (3) 所示。

$$\rho_{adv}(f,x) = \min_{\delta}\{||\delta||_p \text{ s.t.} f(x+\delta) \neq f(x)\} \tag{3}$$

式中：

$f(x)$—— 模型函数；

x—— 原始良性输入样本；

δ—— 施加的扰动。

6.2.2.2. 分布外检测 AUROC

该指标用于衡量模型区分其在训练中见过的域内数据和从未见过的域外新奇数据的能力，用于评估模型的认知边界和谦逊程度。AUROC 为 ROC 曲线下面积，其具体形式如公式 (4) 所示。

$$\text{AUROC} = \int_0^1 TPR(t)d\big(\text{FPR}(t)\big) \tag{4}$$

式中：

$TPR(t)$—— 在阈值为t时的真阳性率，即正确识别的OOD样本；

$\text{FPR}(t)$—— 在阈值为t时的假阳性率，即将ID样本误判为OOD。

6.2.2.3. 期望校准误差 (Expected Calibration Error, ECE)

量化模型输出的置信度分数与其真实预测准确率之间的一致性，用于评估模型预测不确定性的可靠性。其具体形式如公式 (5) 所示。

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{5}$$

式中：

n—— 样本总数；

M—— 将置信度划分的区间数量；

B_m—— 预测置信度落在第m个区间的样本集合；

$\text{acc}(B_m)$—— B_m中样本的平均准确率；

$\text{conf}(B_m)$—— B_m中样本的平均置信度。

6.2.2.4. 平等化赔率

这是一项评估模型决策无偏见性的关键指标，它要求模型对不同受保护群体（由属性A定义）的预测结果必须满足相等的真阳性率与假阳性率，以确保系统决策不会因个体身份属性而产生系统性差异。其具体形式如公式 (6) 所示。

$$P(\hat{Y}=1|A=a,Y=y)=P(\hat{Y}=1|A=b,Y=y),\quad \forall y\in\{0,1\} \tag{6}$$

式中：

$\hat{Y}$—— 模型的预测结果；

A—— 受保护的群体属性；

Y—— 真实标签。

6.2.2.5. 解释一致性

衡量模型的可解释性方法是否稳定。如对于两个非常相似的输入，其产生的解释也应该是相似的。其具体形式如公式（7）所示。

$$C_{exp}=1-\frac{1}{N}\sum_{i=1}^{N}\frac{d\left(E(x_i),E(x_i^{'})\right)}{d(x_i,x_i^{'})} \tag{7}$$

式中：

x_i—— 输入样本；

$x_i^{'}$—— 与x_i非常相似的样本；

E(.)—— 解释函数；

d(.)—— 距离函数。

6. 3. 模型性能指标

6.3.1. 定性指标

6.3.1.1. 评估数据集的代表性与覆盖度

用于评估测试数据集能否无偏且全面地代表其操作设计域内的真实世界数据分布。该评估通过数据审核进行，重点审查数据集构成在关键维度上是否与操作设计域定义具备统计一致性，并是否包含了充足的代表性边缘场景。

6.3.1.2. 度量指标与安全目标的对齐性

此指标评估所选用的底层模型性能指标是否是顶层系统级安全目标的有效和可靠的代理，且能否解释模型性能指标的提升如何直接贡献于系统级安全风险的降低。

6.3.1.3. 模型文档的透明度与完备性

此指标评估描述AI模型的“模型卡”是否全面、清晰地记录了其性能评估结果，并特别说明了在关键的、代表性不足的子集上的表现，以及在不同操作条件下的性能变化。其评估方法是依据一个标准化的框架逐项检查模型卡的完整性、清晰度和准确性，特别是对模型局限性的坦诚度。

6.3.2. 定量指标

6.3.2.1. **准确度**

该指标通过计算正确预测与预测总数的比率来衡量人工智能模型预测的整体正确性。具体形式如公式 (8) 所示。

$$Accuracy = \frac{a}{b} \tag{8}$$

式中：

a —— 正确的预测数量；

b —— 预测总数。

6.3.2.2. **精确度与召回率**

精确度和召回率是分类任务中常用的指标。精确性测量正确预测的阳性观察值与总预测阳性观察值的比率，而召回率计算正确预测的阳性观察值与所有实际阳性观察值之比。具体形式如公式 (9) 所示。

$$Precision = \frac{a}{b},\ Recall = \frac{a}{c} \tag{9}$$

式中：

a —— 正确预测的阳性观察值；

b —— 总预测阳性观察值；

c —— 实际阳性观察值。

6.3.2.3. **F1 分数**

F1分数是精确度和召回率的调和平均值，为分类问题提供了一种考虑假阳性和假阴性的平衡衡量标准，具体形式如公式 (10) 所示。

$$F1\ Score = \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{10}$$

6.3.2.4. **均方误差**（Mean Square Error，MSE）

MSE是回归任务中最常用的损失函数和评估指标，MSE的值越小，说明模型的预测越精准。其具体形式如公式 (11) 所示。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

6.3.2.5. **平均绝对误差**（Mean Absolute Error，MAE）

平均绝对误差是每个预测值与真实值之间绝对误差的平均值，衡量了模型在每个数据点上的平均误差。其具体形式如公式 (12) 所示。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

6.3.2.6. **决定系数**

决定系数衡量模型对数据的拟合度，即模型解释了多少变异性。它表示模型预测值与真实值之间的相关程度，通常用于评估回归模型的整体表现，公式如 (13) 所示。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

6.3.2.7. DB 指数

DB指数是用于评估聚类算法结果的一个指标，它通过衡量不同簇之间的相似度来评估聚类的质量。DB指数通过以下步骤计算：

a） 计算相似度R_{ij}，形式如公式 (14) 所示。

$$R_{ij} = \frac{S_i + S_j}{d(C_i, C_j)} \tag{14}$$

式中：

C_i、C_j—— 聚类结果中的每一对簇；

S_i、S_j—— 两个簇的紧密度；

D—— 任意的距离计算函数。

b） 使用所有簇之间的相似度，计算DB指数，具体形式如公式 (15) 所示。

$$DB = \frac{1}{K}\sum_{i=1}^{K}\max_{i \neq j} R_{ij} \tag{15}$$

6. 4. 数据安全监控指标

这一部分指标关注智能系统的输入数据的质量与安全性。其核心理念是，在数据被模型使用之前，就识别出可能导致模型失效或被攻击的风险。

6.4.1. 定性指标

6.4.1.1. 数据规程与溯源文档完备性

该指标用于评估数据集配套文档是否清晰、完整地记录了其全生命周期信息，涵盖采集方法、标注流程、版本历史、隐私保护与已知局限性等方面，以确保数据来源与处理过程的透明度和可追溯性。

6.4.1.2. 数据标注质量保证流程成熟度

该指标用于评估用于生成训练和测试标签的数据标注流程，是否包含成熟的质量保证机制，以确保标签的准确性和一致性。

6.4.1.3. 数据安全与隐私保护机制有效性

此指标用于评估在数据全生命周期中保护其机密性、完整性与可用性的技术与流程。该评估通过安全审核与渗透测试等方式，实际验证访问控制、加密及防泄露等机制的有效性与鲁棒性。

6.4.2. 定量指标

6.4.2.1. 数据分布漂移量

这是一项采用统计方法量化线上实际输入数据与原始训练/测试数据之间分布差异的指标。该指标用于监测独立同分布假设是否被违背，从而评估模型在持续应用中的有效性与适用性。其具体形式如公式 (16) 所示。

$$D_n = \sup_x |F_n(x) - F(x)| \tag{16}$$

式中：

$F_n(x)$—— 线上数据的经验累积分布函数；

$F(x)$—— 训练数据的累积分布函数。

6.4.2.2. 传感器数据有效性

这是一项量化在特定时间窗口内，被判定为有效、可信的原始传感器数据帧占总数据帧比例的指标。该指标用于在数据进入模型前进行前置质量评估，以保障系统输入的可靠性，并规避无效数据对模型性能的负面影响。其具体形式如公式 (17) 所示。

$$R_{valid} = 1 - \frac{N_{anomaly}}{N_{total}} \tag{17}$$

式中：

N_{total}—— 总的传感器数据帧数量；

$N_{anomaly}$—— 通过物理一致性检查、时序分析或自监督模型检测出的异常数据帧数量。

6.4.2.3. 数据投毒检测率

这是一项衡量防御系统识别并剔除混入训练或在线学习数据流中恶意样本能力的指标。该指标用于评估系统抵御数据投毒攻击的有效性，以保障模型完整性，并防止其因特定触发条件而做出预设的错误决策。其具体形式如公式 (18) 所示。

$$\mathrm{PDR} = \frac{TP_{poison}}{TP_{poison}+FN_{poison}} \tag{18}$$

式中：

TP_{poison}—— 真阳性，即被防御机制成功识别为有毒的样本数量；

FN_{poison}—— 假阴性，即防御机制未能识别出来的、漏掉的有毒样本数量。

6. 5. 运行安全监控指标

这一部分指标关注智能系统部署上线后，在其与真实环境持续交互过程中的行为监控。其核心理念是，即使模型通过了所有离线测试，也必须在运行时对其行为进行持续的监督和验证。

6.5.1. 定性指标

6.5.1.1. 运行时保障机制的可靠性

该指标旨在评估在线安全监控器自身设计的简洁性、可形式化验证性，及其在资源与故障模式上与被监控主模块的独立程度。该指标通过设计评审与形式化分析进行评估，以确保监控器作为最后一道安全防线的可靠性。

6.5.1.2. 部署后事件响应机制有效性

该指标旨在评估用于监控、报告、分析及响应系统运行中安全相关事件的流程有效性。该评估通过流程审核，审查事件报告、根本原因分析与相关的工程变更记录，以验证是否形成了从事件响应到产品迭代的、持续改进的完整闭环。

6.5.2. 定量指标

6.5.2.1. 在线性能衰退指数

该指标通过在线A/B测试或影子模式等方法，持续计算模型部署后的关键性能指标，并与部署前的基线进行比较，以量化模型在真实世界环境中性能随时间衰退的相对程度。其具体形式如公式 (19) 所示。

$$\mathrm{PDI}(t) = \frac{\mathrm{Perf}_{\mathrm{baseline}} - \mathrm{Perf}_{\mathrm{window}}(t)}{\mathrm{Perf}_{\mathrm{baseline}}} \tag{19}$$

式中：

$\mathrm{PDI}(t)$—— 在时间点t的性能衰退指数；

$\mathrm{Perf}_{\mathrm{baseline}}$—— 模型部署前在离线验证集上测得的基线性能指标值；

$\mathrm{Perf}_{\mathrm{window}}(t)$—— 在以时间点$t$结尾的滑动时间窗口内，使用线上数据测得的相同性能指标的值。

6.5.2.2. 运行时验证违例率

基于一系列预先定义的、形式化的安全规约，在运行时持续检查AI系统的输入、输出或内部状态是否违反了这些规约。该指标计算在单位时间或单位里程内，发生违例事件的频率。其具体形式如公式 (20) 所示。

$$\lambda_{RV} = \frac{N_{violation}(\Delta T)}{O_{window}} \tag{20}$$

式中：

λ_{RV}—— 运行时验证的违例发生率；

$N_{violation}(\Delta T)$—— 在一个时间或里程窗口ΔT内，检测到的违例事件总数；

O_{window}—— 对应的操作窗口大小。

6.5.2.3. 人机交互接管时间

对于需要人类进行监督或在特定情况下接管的半自主系统，该指标衡量从系统发出接管请求到人类驾驶员做出稳定、有效的控制动作所经过的时间。其具体形式如公式 (21) 所示。

$$T_{takeover} = t_{control} - t_{request} \tag{21}$$

式中：

$T_{takeover}$—— 接管时间；

$t_{request}$—— 系统发出接管请求的时间戳；

$t_{control}$—— 人类提供第一个稳定有效的控制输入的时间戳。

7 复杂智能系统功能安全性支撑技术与方法

7.1. 需求阶段功能安全性支撑技术

7.1.1. 危害分析与风险评估和安全目标定义

通过结构化方法系统性地识别由功能失效导致的潜在危害，并基于危害的严重性、暴露率和可控性评估风险等级。此过程旨在推导出必须被满足的、最高层级的安全目标。

7.1.2. 操作设计域的形式化规范

对系统预期运行环境、条件和任务进行精确、无歧义的定义。一个形式化的ODD是后续所有功能安全分析和验证活动的基础，明确了系统的能力边界，并作为在系统无法安全运行时的最小风险状态的触发依据。

7. 2. 设计与开发阶段功能安全性支撑技术

7.2.1. 故障容忍架构设计

通过引入异构冗余和通道分离等技术，消除单点故障和共因失效。对于AI系统，这可能包括在感知和决策中使用不同原理的算法或硬件，以确保在单一故障发生时，系统仍能维持安全运行或进入安全状态。

7.2.2. 冗余、多样性与容错架构设计

这是一种通过引入额外的资源来使系统能够容忍故障的经典安全设计技术。它包含多种形式：

a） 冗余指部署多个相同或相似的组件来执行同一功能。

b） 多样性指冗余的组件采用不同的技术原理、由不同的团队开发或采用不同的算法，以防止单一原因导致所有通道同时失效。

c） 容错指整个架构被设计为能够检测故障、隔离故障，并安全地重构或切换到备用通道，以维持系统的关键功能或转换到安全状态。

7.2.3. 运行时保障与安全监控器架构

这是一种专门为监督复杂AI组件而设计的模式，其核心是为一个AI主功能模块并行配备一个逻辑简单且可被形式化验证的安全监控器。该监控器持续将AI模块的输出与预定义的安全边界进行比对，并在检测到违规时，立即否决该输出并执行预设的、保守但能确保安全的备用策略。

7.2.4. 信息隔离与最小权限原则

这是一种源自信息安全，但在功能安全领域同样至关重要的软件架构设计原则：

a） 信息隔离指通过使用虚拟机、容器或硬件内存保护单元，将安全等级不同的软件组件或功能关键性不同的模块，在计算资源和通信上进行严格的隔离。

b） 最小权限指确保每个软件组件只被授予其完成任务所必需的最小权限和数据访问权。此方法旨在限制故障的传播范围，防止一个低安全等级组件的失效“污染”或影响到一个高安全等级组件的正常运行。

7.2.5. 降级模式与最小风险状态设计

这是一种系统级的安全策略设计。它要求在设计之初就预先设想当系统因故障或性能局限而无法继续提供其完整功能时，应该如何安全地退出服务。这包括：

a） 降级模式指系统在失去部分功能后，仍能安全运行的功能子集。
b） 最小风险状态指当系统必须完全停止服务时，能够将风险降至最低的最终状态。此技术的核心是在设计阶段就为所有可预见的失效场景，规划好一个安全、确定的“应急预案”。

7.2.6. 模型卡的创建与应用

作为AI组件设计文档的一部分，创建模型卡以提高透明度。模型卡详细记录了模型的预期用途、训练数据概述、已知局限性以及在不同数据子集上的性能和公平性评估结果，为系统集成和安全评估提供了关键的设计阶段证据。

7. 3. 模型训练阶段功能安全性支撑技术

7.3.1. 数据准备与治理方法

7.3.1.1. 数据集溯源与规程文档化

这是一种通过为数据集创建详尽数据表以系统性记录其全生命周期信息的方法，内容涵盖采集动机、标注流程、版本历史、隐私措施及已知局限性。该方法旨在提升数据透明度与可追溯性，作为关键证据支持AI安全论证，并使得评估数据与操作设计域的匹配度、识别数据偏见和进行根本原因分析成为可能。

7.3.1.2. 自动化数据验证与模式强制

这是一种在数据进入训练流程前，通过自动化工具将数据与预定义模式进行比对，以检测类型、格式、取值范围及统计分布异常的技术。该技术是防止“脏数据”污染模型的首道技术防线，旨在自动捕获异常数据以规避潜在的性能下降，并将抽象的数据质量要求转化为可被机器自动执行的规则。

7.3.1.3. 数据标注质量保证与一致性分析

这是一套旨在确保监督学习模型所用标签准确性与一致性的严谨流程，其核心措施包括制定详细标注指南、采用多人交叉验证及量化一致性指标。该流程通过减少直接决定模型性能与安全上限的标签噪声，来提升模型在安全关键类别上的可靠性，并能通过分析不一致性来识别困难场景，为功能安全分析提供输入。

7.3.1.4. 面向安全关键场景的合成数据生成

通过高保真度的仿真环境，程序化地生成大量、多样化、且带有完美“真值”标签的训练和测试数据，尤其专注于生成那些在真实世界中难以采集或极度危险的安全关键场景。合成数据是解决AI安全验证中“数据稀疏性”问题的核心技术之一，能够针对性地增强模型在已知触发条件下的鲁棒性，并为基于场景的测试提供近乎无限的、参数可控的测试用例，从而系统性地探索模型的行为边界。

7.3.2. 训练过程与策略方法

7.3.2.1. 对抗性训练

这是一种“在攻击中学习”的鲁棒性优化技术，在训练过程中实时地为每个样本生成微小的、能最大化模型损失的对抗性扰动，然后训练模型去正确分类这些被“精心污染”过的样本。此方法旨在内建模型的鲁

棒性，通过迫使模型学习更平滑、更本质的特征，来提升其对真实世界中未曾见过的传感器噪声和环境变化的抵抗力。它也是目前已知的、最有效的直接防御恶意对抗性攻击的方法之一，从而直接提升了模型的内在安全性。

7.3.2.2. 偏见缓解与公平性感知训练

这是一种在训练算法层面主动纠正偏见的技术，在优化模型主任务损失的同时，将其必须满足的公平性指标作为一个约束条件加入优化问题中。安全性必须是公平的，一个对特定人群存在偏见的系统本质上是不安全的。此方法能够主动提升模型在弱势群体上的性能，确保例如行人检测系统不会因为肤色、着装等因素而产生歧视性的、危险的性能差异，是构建负责任AI的关键技术。

7.3.2.3. 差分隐私训练

这是一种为模型训练过程提供严格、可数学证明的隐私保护的技术，通过在梯度计算阶段进行裁剪和噪声注入，来掩盖任何单个训练样本对最终模型参数的贡献。在处理敏感数据的智能系统中，数据隐私本身就是一项关键的安全属性。此技术能提供数学保证，防止模型记忆和泄露敏感信息，从而避免因信息泄露导致的物理世界伤害，并能附带地提升模型对某些类型对抗性攻击的鲁棒性。

7.4. 智能系统功能安全性测试与验证技术

7.4.1. 数据与模型层面测试与验证技术

7.4.1.1. 数据集剖析与验证技术

这是一种在模型开发生命周期的最前端，对数据集进行的系统性、自动化“健康体检”。它通过计算详尽的统计数据，来理解并验证数据集的每一个特征，确保其符合预定义的期望。此技术的目标是主动地、前置地保障数据质量，而非被动地在模型失效后进行问题回溯。其核心意义在于：

a） 防止“脏数据”污染模型指识别并处理标签错误、异常值、缺失值等，避免模型从一开始就学到错误的知识。

b） 诊断与量化数据漂移指通过对比训练集、验证集和线上数据的统计分布，精确量化数据分布漂移，为模型的更新和再训练提供数据驱动的依据。

c） 保障公平性与鲁棒性指揭示数据中潜在的偏见（如类别不均衡、某些群体样本过少），指导数据增强或采集策略，以提升模型的公平性和泛化能力。

7.4.1.2. 对抗性测试技术

这是一种模拟最坏情况的压力测试技术，它通过利用模型自身的梯度信息，在输入数据上有目的地添加人眼难以察觉的扰动，以生成能最大化误导模型决策的“对抗样本”。该技术旨在主动探测并暴露神经网络决策边界的非预期脆弱性，从而评估模型在遭受刻意攻击等最恶劣情况下的鲁棒性下限，为功能安全提供关键衡量依据。对抗性测试可以遵循以下流程：

a） 定义威胁模型指明确攻击者的能力范围，如允许修改的扰动范围、可获取的模型信息。

b） 采用成熟的攻击算法，如快速梯度符号法（FGSM） 或更强大的投影梯度下降（PGD），来系统性地生成测试用例。

c） 将生成的对抗样本输入到模型中，计算模型在这些样本上的准确率、置信度等性能指标，并与在干净样本上的性能进行对比。

7.4.1.3. 神经网络的形式化验证技术

这是一种基于数学与逻辑推理的完备性验证技术，旨在通过穷尽性搜索与数学证明，来确定一个给定的安全属性在所有可能输入下是否恒成立。该技术的目标并非评估平均性能，而是通过为关键安全属性提供完备的数学保证来消除最坏情况风险，从而为系统安全性提供最强的、可被数学证明的论据。神经网络的形式化验证技术可以遵循以下流程：

a） 将自然语言描述的安全需求，转化为精确的数学逻辑表达式。
b） 利用SMT求解器、抽象释义或区间算术等技术，计算神经网络在给定输入范围内所有可能输出的集合。
c） 检查计算出的输出集是否完全落在安全规约定义的范围内。如果不满足，工具将自动生成一个导致违例的具体输入作为反例。

7.4.2. 软件在环与硬件在环测试与验证技术

7.4.2.1. 软件在环 (Software-in-the-Loop, SIL) 测试技术

在一个纯软件的环境中，将算法代码（作为“环”中的一部分）与其它系统组件（如车辆动力学模型、传感器模型、控制器模型）的软件仿真进行闭环集成测试。所有组件都运行在开发计算机上，不涉及任何目标硬件。这是在开发流程早期进行快速迭代和功能验证的核心手段。软件在环测试技术可以遵循以下流程：

a） 使用仿真工具搭建虚拟的生产或测试环境模型。
b） 将待测的算法代码集成到仿真环境中。
c） 运行预定义的测试脚本，通过断言和日志来自动判断测试结果是否通过。

7.4.2.2. 硬件在环 (Hardware-in-the-Loop, HIL) 测试技术

将编译好的算法软件，烧录到最终量产的目标硬件控制器（ECU）中。然后，将这个真实的目标硬件控制器与一个强大的实时仿真平台连接，该平台能够实时生成高保真度的传感器数据流并模拟外界组件对目标硬件控制器控制指令的物理响应。硬件在环测试技术可以遵循以下流程：

a） 使用专业HIL设备搭建模拟平台，并配置高保真度的传感器或动力学模型。
b） 通过真实的物理接口将待测ECU与HIL平台连接。
c） 运行测试用例，实时采集ECU的输出并注入仿真环境，形成一个完整的、包含真实硬件在内的闭环。

7.4.2.3. 故障注入测试技术

在SIL或HIL测试环境中，通过专门的工具和技术，人为地、可控地向系统的特定部位注入故障。这些故障可以是硬件层面的，也可以是软件层面的。故障注入测试技术是唯一能够直接、主动地验证系统容错设计有效性的技术。其目标是确认当随机故障发生时，系统的安全机制是否能够被正确、及时地激活，并将系统引导至一个预定义的安全状态。故障注入测试技术可以遵循以下流程：

a） 基于FMEA（失效模式与影响分析）等安全分析结果，定义需要注入的故障类型、位置和触发时机。

b） 利用HIL平台的特定模块或通过修改软件代码，在测试运行时精确地注入故障。
c） 观察并记录系统的响应，并与安全需求中定义的预期行为进行比对。

7.4.3. 系统与交互层面的测试与验证

7.4.3.1. 针对物理交互系统

7.4.3.1.1. 高保真度多模态感知仿真测试技术

一种为验证智能系统物理感知能力而设计的场景化测试，其核心是通过物理级渲染与信号传播模型生成高保真度的多模态传感器数据。该技术通过精细建模环境触发条件（如恶劣天气与传感器物理效应），系统性地测试系统在信息不完备、含噪声的真实物理输入下的性能极限，旨在验证感知鲁棒性并主动发现导致系统失效的边界条件。

7.4.3.1.2. 基于规约的运行时监控与保障

这是一种部署在最终物理系统上的在线验证与干预技术，其核心是利用一个独立的“安全监视器”模块，持续将系统外部可观察的物理行为与一系列预定义的形式化安全规约进行比对。该技术旨在提供最后一道可验证的动态安全防线，以捕获并干预在设计与测试阶段未能预见的危险行为，并在检测到违规时否决主模块指令、执行预设的安全备用策略。

7.4.3.2. 针对信息交互系统

7.4.3.2.1. “红队”测试与对抗性提示工程

一种模拟恶意或非常规用户的测试方法，其核心是通过精心设计输入提示，主动诱导大语言模型等智能系统产生不安全、不合规或非预期的输出。该技术旨在系统性地探测并暴露模型在语义与内容安全上的脆弱性，主动发现其安全护栏中可被利用的漏洞。这一技术的具体实现可分为：

a） 由具备创造力和对抗思维的专家，根据已知的攻击模式进行手动测试。
b） 利用另一个LLM来自动生成大量、多样化的、可能触发不安全行为的对抗性提示。
c） 使用类似计算机视觉中的梯度方法，在输入的嵌入空间中进行优化，以寻找最能引发有害输出的提示。

7.4.3.2.2. 基于基准的自动化评估与内容审查

这是一项利用标准化评估基准与数据集，自动化且规模化地衡量大语言模型在真实性、无害性及偏见等多个安全维度上表现的技术。该技术旨在将抽象的安全概念转化为可量化的客观指标，从而为模型迭代提供持续的回归测试，并为不同模型间的安全性对比提供依据。该技术的实现可分为以下几步：

a） 采用学术界和工业界公认的评估基准，如TruthfulQA（评估真实性）、ToxiGen（评估有害性）、BBQ（评估社会偏见）等。
b） 让待测模型对基准中的所有提示进行响应生成。
c） 使用另一个强大的LLM作为“裁判”，根据预定义的评分标准（Rubric）来自动评估模型输出的质量，或者使用预训练的内容审查分类器来进行打分。

7.5. 智能系统运行阶段功能安全性支撑技术

7.5.1. 运行时保障

部署一个独立的、可验证的“安全监视器”，与复杂的AI主模块并行运行。监视器持续检查主模块的输出是否位于预定义的安全信封内。如果即将发生违规，监视器将立即介入并执行一个备用的安全策略。这是应对AI模型在未知环境下的未知风险的关键技术。

7.5.2. 在线数据分布与性能监控

持续监控线上输入数据流与训练数据基线之间的分布漂移。同时，跟踪模型在役期间的性能衰退。当检测到显著漂移或性能下降时，触发报警或安全干预，以规避因模型“老化”或环境变化导致的风险。

7.5.3. 持续的安全论证管理

安全论证是一个动态文件，必须在运行阶段持续维护。任何新的运营数据、事故报告或未知事件都必须被纳入论证中，以确保其始终准确反映系统的安全状态，并作为持续改进和安全决策的依据。

7. 6. 智能系统维护与更新阶段功能安全性支撑技术

7.6.1. 变更影响分析与安全回归验证

每一次软件或AI模型的更新都必须执行正式的变更影响分析，以识别潜在的安全影响。必须执行全面的回归测试套件，确保系统的所有原有安全功能在更新后依然有效，未发生安全回归。

7.6.2. 部署后事件响应与学习机制

建立一套清晰、有效的流程，用于系统性地监控、报告、分析和响应在系统实际运行中发生的所有安全相关事件。通过根本原因分析等方法，将教训反馈到产品迭代和安全管理流程的改进中，形成持续学习的闭环。

7.6.3. 网络安全与功能安全的协同验证

在更新部署过程中，功能安全团队必须与网络安全团队协同，确保软件完整性、代码签名和加密机制的有效性。这防止攻击者利用更新渠道对系统功能进行恶意篡改，即网络安全对功能安全的支撑作用。

7. 7. 智能系统退役阶段功能安全性支撑技术

7.7.1. 安全的最终状态转换与功能禁用

执行一个经过验证的退役计划，确保系统功能被永久、可靠地禁用。系统必须进入一个无法对人员或环境造成伤害的安全最终状态，防止缺乏维护和支持的情况下出现不安全的残留操作。

7.7.2. 敏感数据安全处理与隐私合规

对系统存储的所有敏感用户和运营数据进行安全处理。数据必须根据相关法规进行不可逆的删除或严格的匿名化，以规避退役阶段的隐私泄露风险。

8 复杂智能系统功能安全性全生命周期过程与活动

8.1. 需求阶段

需求阶段构成了智能系统生命周期的起点，其核心任务是进行系统性的需求工程，以形式化地定义系统的预期功能、性能特征、运行边界及顶层安全属性。此阶段的关键活动包括对操作设计域的精确规范，执行危害分析与风险评估以推导安全目标，并识别所有必须遵守的法律法规与行业标准。此阶段的最终交付物是一套经过验证的、可追溯的系统级需求，它将作为所有后续开发、验证和确认活动的基准。

8.2. 设计与开发阶段

在设计与开发阶段，依据安全需求建立具备冗余、容错与安全监控的系统架构，准备并验证数据集的完整性、代表性和准确性，训练符合安全约束的AI模型，并确保软硬件接口传输安全可靠。此阶段的核心活动是将安全目标分解为可分配给各架构组件的技术安全概念。为满足高安全等级要求，必须采用故障容忍架构设计，通过引入冗余、多样性和通道分离等技术来规避单点和共因失效。对于AI组件，则需创建标准化的模型卡，以确保其预期用途、已知局限性和内在风险的透明度，为系统集成和安全评估提供关键的设计阶段证据。

8.3. 训练阶段

训练阶段是智能系统从数据中获取其功能与行为模式的核心过程，是智能系统所特有的数据驱动的开发环节。此阶段不仅包括模型本身的训练、验证和超参数调优，更涵盖了整个数据生命周期的管理，包括数据的采集、清洗、标注、增强以及版本控制。此阶段必须实施严格的数据质量保证流程，并通过数据表等方法确保数据的可追溯性与治理。在训练过程中，应主动采用对抗性训练等技术来提升模型对扰动的鲁棒性，应用偏见缓解算法来确保公平性，并在处理敏感数据时，通过差分隐私训练等隐私增强技术，从根源上防止模型泄露个人信息。

8.4. 测试阶段

测试阶段包含了所有的验证与确认活动，旨在提供客观证据以证明所开发的系统满足其在功能、性能和安全方面的所有规定需求。此阶段通常遵循V模型，从单元和集成测试，逐步过渡到系统级验证。关键活动包括在纯软件环境中进行的软件在环测试，在目标硬件上进行的硬件在环测试，在高度逼真的虚拟环境中进行的基于场景的仿真测试，以及最终在真实环境中对完整系统进行的结构化实地测试。

8.5. 运行阶段

运行阶段标志着智能系统被正式部署到其预定的操作环境中，开始执行其功能并与真实世界或用户进行持续交互。这是一个动态的过程，其核心活动不仅包括系统的正常运行，还强制要求进行持续的在线监控。这包括对模型性能的实时评估、对输入数据分布漂移的检测、对所有安全相关事件和数据的记录，以及确保任何运行时保障机制都处于激活状态，以作为应对未知风险的最后防线。

8.6. 维护与更新阶段

维护与更新阶段是一个贯穿系统整个在役生命周期的持续过程，专注于对已部署系统进行变更管理。其活动可分为：为修正缺陷而进行的纠正性维护，为适应环境变化而进行的适应性维护，以及为增强功能而进行的完善性维护。对于智能系统，这突出地表现为模型的再训练与重新部署，通常通过在线软件更新

实现。每一次变更都必须经过严格的影响分析和回归测试，以确保其不会对已验证的系统安全性产生负面影响。

8.7. 退役与报废阶段

在退役与报废阶段，应安全销毁或隔离系统中的敏感数据与模型，妥善处置硬件防止逆向与滥用，并完整归档安全文档和运行记录，以支持事故调查和法规合规。此阶段要求执行一个经过验证的退役计划，通过可靠的技术程序永久性地禁用所有安全相关功能，以防止不安全的残留操作。同时，必须依据相关数据保护法规，对系统存储的所有敏感用户和运营数据进行安全的、不可逆的删除或严格的匿名化处理。最后，清晰、有效地向所有利益相关方沟通服务的终止及其安全影响，是此阶段不可或缺的管理活动。

参 考 文 献

[1] GB/T 42382 信息技术 神经网络表示与模型压缩
[2] GB/T 11457 信息技术 软件工程术语
[3] T/CSAE 316 智能网联汽车 环境感知系统预期功能安全
[4] GB/T 34590 道路车辆 功能安全
[5] GB/T 40856 车载信息交互系统信息安全技术要求及试验方法
[6] GJB/Z 102 软件可靠性和安全性设计准则
[7] T/CSAE 177 电动汽车车载控制器软件功能测试规范
[8] ISO/IEC/IEEE 29119.11 AI system testing and V&V guidance
[9] ISO 14971 Medical devices — Application of risk management to medical devices
[10] IEC 61508 Functional safety of electrical/electronic/programmable electronic safety — related systems
[11] NASA-STD-8739.8 Software assurance and software safety standard
[12] ISO 26262 Road vehicles — Functional safety