人工智能维护：鲁棒性视角

Pin-Yu Chen 和 Payel DasIBM Research
pin-yu.chen@ibm.com 和 daspa@us.ibm.com

arXiv:2301.03052v1 [cs.LG] 8 Jan 2023

摘要——**随着机器学习（ML）方法和计算资源不断发展，人工智能（AI）赋能的系统正成为一种普遍应用的技术。然而，当前的人工智能技术（如深度学习）并非完美无缺。模型复杂度和数据规模大幅增加，在缺乏可信度和透明度的情况下带来了日益严峻的挑战，可能产生新的风险和负面影响。本文从鲁棒性视角提出人工智能维护这一概念。我们首先介绍人工智能生命周期中一些突出的鲁棒性挑战，并通过类比汽车维护阐释人工智能维护的必要性。随后，我们提出一个人工智能模型检查框架，用于检测并缓解鲁棒性风险。我们还借鉴车辆自主性的概念，定义人工智能鲁棒性自动化的级别。我们提出的人工智能维护方案有助于在整个生命周期中开展鲁棒性评估、状态跟踪、风险扫描、模型强化和法规管理，这是构建可持续、可信人工智能生态系统的重要里程碑。**

1. 引言

正如汽车在现代世界中发挥着不可或缺的作用，人工智能赋能技术、基于机器学习的系统和算法也正在为我们的生活、社会和环境带来革命性变化和深远影响，甚至这些变化可能已经发生。人们将人工智能模型视为驶向更美好未来的新“汽车”，本文旨在通过类比汽车与人工智能的模型开发与部署，强调从鲁棒性视角将人工智能维护规范化并付诸实践的重要性。为实现人工智能的可信度与可持续性，本文受到以下问题的启发：汽车需要定期检查、维护和持续监控状态，人工智能技术为何要另当别论？

人工智能的鲁棒性通常因上下文和使用场景不同而具有多种含义。本文从人工智能模型的泛化能力这一视角研究鲁棒性，即模型在对抗和未见场景中的表现。一般而言，人工智能模型的性能通过平均情况进行评估：将模型在一组数据样本上的预测与其真实标签进行比较，再以平均预测结果作为性能指标，例如 Top-1 分类准确率，用于衡量模型在整个数据集上对最可能类别（Top-1 类别）预测正确的比例。相比之下，对抗场景通过假设存在虚拟对手，评估模型在所有可能且合理（通常预先指定）的数据和人工智能模型变化中最坏情况下的性能。此外，未见场景评估模型在新数据样本上的性能；这些样本来自与训练期间见过的数据样本不同的数据分布（但不一定是最坏情况分布），其成因可能包括自然的数据/标签偏移以及现实世界中的观测噪声等。

从鲁棒性视角研究人工智能维护的依据

源于快速增长的需求：检查并预防人工智能模型的失效模式，以了解其局限性，并让人工智能技术做好准备，应对现实世界中的恶意攻击和连续的数据变化。根据 Gartner 最近的一份报告[1]，30到 2022年，% 的网络攻击将涉及数据投毒、模型窃取或对抗样本（参见 [1]，了解这些以机器学习为核心的新风险）。然而，业界似乎准备不足。在一项涵盖 28 家大小组织的调查中，有 25 家组织不知道如何保障其人工智能/机器学习系统的安全 [2]。与承保损失和责任的汽车保险不同，如果网络保险提供商在根本原因与人工智能失效模式相关时施加严格要求，人工智能模型缺乏鲁棒性所带来的风险可能会进一步加剧[2]。此外，人工智能维护与提升安全相关机器学习应用可信度的行动计划密切相关，例如完成欧盟航空安全局（EASA）路线图中规定的里程碑和目标[3]，以及美国食品药品监督管理局制定的人工智能/机器学习医疗器械软件行动计划[4]，和 NIST人工智能风险管理框架[5]。

为了深入了解人工智能维护，本文首先介绍人工智能生命周期中模型开发与部署面临的主要鲁棒性挑战。随后，我们类比汽车维护与人工智能维护的共通之处。最后，我们提出名为“人工智能模型检查器”的概念框架，用于全面检查和提升鲁棒性。类似于车辆自动驾驶能力的驾驶自动化等级定义，我们定义了人工智能鲁棒性的六个等级，以便在整个生命周期中对人工智能技术进行定性和定量评估。

2. 人工智能生命周期中的鲁棒性挑战

图 1 概述了人工智能生命周期中的鲁棒性检查流程（左图）以及重点标出的鲁棒性挑战（右图）。人工智能生命周期在两个阶段之间循环：模型开发与部署。模型开发阶段包括两个状态：（i）数据收集与处理，以及（ii）模型训练。数据收集与处理包括典型的数据操作，例如数据获取与标注、特征归一化、筛选、匿名化和数据增强。模型训练涉及机器学习模型选择、算法开发、系统设计和优化。在状态（i）和（ii）之间，数据净化会检查数据保真度并采取缓解措施（例如删除有问题的数据样本或更正错误标注样本），然后再进行模型训练。模型开发完成后，人工智能生命周期进入（iii）模型部署状态，此时可训练的模型参数会被冻结以供使用。在（ii）和（iii）之间，性能验证会检查并缩小模型训练与部署之间的差距。如果已部署模型出现显著的性能下降，可能是自然发生的数据偏移或恶意攻击所致，人工智能生命周期将重新进入模型开发阶段，以收集新数据或更新模型。在（iii）和（i）之间，持续监控会检查当前已部署模型的性能状况，并在观察到显著性能下降或检测到异常事件时发出通知。

在模型开发与部署阶段，存在不同类型的鲁棒性挑战，可能导致模型异常行为和性能下降。这些挑战因目标、对人工智能模型实施干预时可采取的行动，以及对人工智能模型的了解程度而异。在对抗场景中，鲁棒性挑战可能与一种“威胁模型”有关，该模型规定了攻击者能够了解并采取哪些行动来破坏人工智能模型。在未见场景中，鲁棒性挑战与开发阶段和部署阶段之间的领域泛化能力有关。图 1 （右侧面板）列出了每个阶段的两个重点鲁棒性挑战

[1] https://www.gartner.com/smarterwithgartner/ gartner-top-10-security-projects-for- [2] https://hbr.org/2020年/04/the-case-for-ai-insurance [3] https://www.easa.europa.eu/newsroom-and-events/news/easa-releases-its-concept-paper-first-usable-guidance-level-1 [4] https://www.fda.gov/medical-devices/软件-medical-device-samd/artificial-intelligence-and-machine-learning-软件-medical-device [5] https://www.nist.gov/itl/人工智能-risk-management-framework

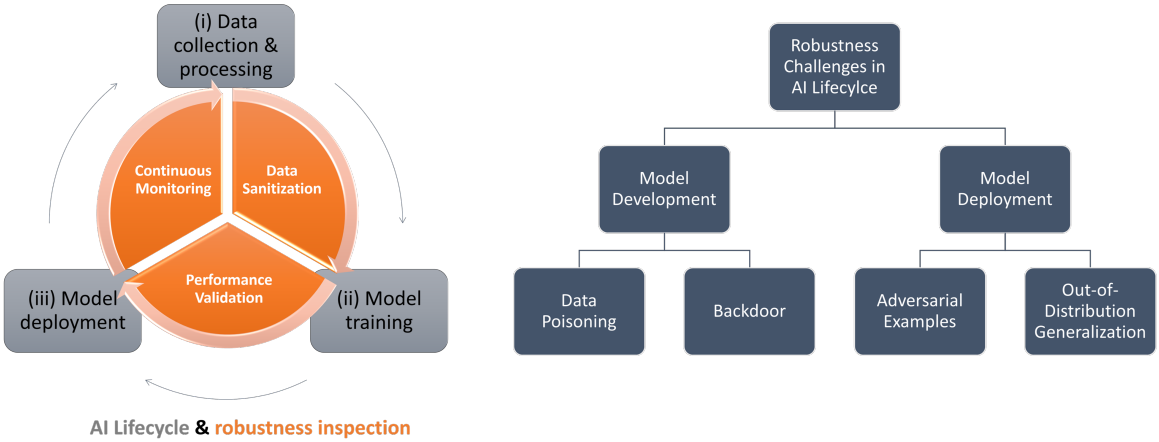


图1。左侧：人工智能生命周期中鲁棒性检查流程（数据净化、性能验证和持续监控）的示意图，该生命周期由三个主要阶段组成：数据收集与处理、模型训练和模型部署。模型开发阶段包括数据收集与处理和模型训练。**右侧：**突出显示人工智能生命周期中的鲁棒性挑战。在模型开发阶段，鲁棒性挑战假设训练数据在模型训练前可能受到操纵。在模型部署阶段，鲁棒性挑战无法访问训练数据，但可能掌握已部署模型的某些信息，例如模型架构及其相关模型参数。根据人工智能生命周期各阶段的分类，该图还可扩展，以纳入其他鲁棒性挑战以及安全性、隐私等其他可信性维度。

阶段，具体如下。

2.1. 开发阶段的鲁棒性挑战

数据投毒关注的是模型在噪声数据上训练时的性能。噪声可能源于不完善的数据收集与处理，例如不正确的数据标注、数据偏差和类别不平衡，以及与上下文无关的伪特征。也可能有人故意向训练数据中加入一组投毒数据样本，以削弱模型在部署阶段的性能。例如，使目标模型在开发阶段的分类错误率较低，而在部署阶段的分类错误率较高。此类蓄意的数据投毒攻击通常假设攻击者有能力操纵训练数据，并能够访问或部分了解模型细节和训练流程 [3]。

后门是一种针对机器学习 [4]的木马攻击。其通过向一部分训练数据注入带有修改后标签的某种模式（触发器）来实现。由于最先进的机器学习模型（例如神经网络）的记忆效应，在被篡改的数据集上训练的模型将包含后门。在部署阶段，植入后门的模型会允许攻击者在指定触发器出现时控制模型输出，而无论数据输入的实际内容是什么。然而，在没有触发器的情况下，植入后门的模型会表现得像是在未经篡改的训练数据集上训练的普通模型。因此，后门攻击具有隐蔽性，因为被篡改的模型在后门未被激活时不会出现异常行为。在涉及多个参与方交换其本地私有数据的有限信息的分布式和去中心化机器学习范式中，这一挑战可能会进一步加剧，例如联邦学习 [5]。

2.2. 部署阶段的鲁棒性挑战

部署阶段会将开发阶段中经过充分调优的模型固定下来，以供后续数据推理任务使用。如果模型的详细信息对用户透明（例如，发布深度学习模型及其模型架构），则称部署后的模型为白盒模型

并提供预训练权重）。反之，如果用户不知道（或仅部分了解）模型的详细信息，则称其为黑盒（灰盒）模型，例如只提供模型预测结果、不披露其他细节的预测应用程序接口（API）或专有软件。在鲁棒性评估中，白盒模式支持全栈系统调试和内部渗透测试，而黑盒模式则允许根据用户访问情况开展实际的漏洞与信息泄露分析。

对抗样本 是经过精心构造的数据样本，与原始未修改的数据样本相比，会导致预测规避[6]。预测规避的难易程度反映了模型对数据输入微小变化的敏感性，例如人类无法察觉的加性扰动。对抗样本带来的鲁棒性挑战通常与安全关键型和安全相关的人工智能应用有关，例如自动驾驶汽车、身份识别与识别以及恶意软件检测，因为对抗样本的存在可以被视为违反所需鲁棒性约束的反例。在黑盒设置下，可以仅根据模型的预测输出，通过迭代修改数据输入来生成对抗样本 [7]。

分布外泛化是指：当输入数据样本经过某些保持语义的变换，偏离模型训练期间所见的数据分布时，对模型性能进行刻画。相比之下，分布内泛化是指模型在与训练数据或环境具有相同分布的数据样本或实例上的性能。对分布外泛化的追求，源于希望在面对自然变化时仍能保有鲁棒预测能力（这些变化的影响可能是可观测的，也可能是隐藏的）。示例包括开发与部署阶段之间的数据分布偏移、在线数据流中的数据/标签漂移、由测量/设备误差导致的常见损坏，以及由图像旋转或缩放所产生的数据不变操作。理想的部署模型应具备良好的泛化能力，或能够快速识别并适应分布外的未见数据样本；这些样本虽然属于分布外数据，但与训练期间所见的分布内数据具有相似的上下文。

3. 汽车维护与人工智能维护的类比

由于人工智能赋能算法和系统常被终端用户视为强大却神秘的技术，我们认为，将其与（自动驾驶）汽车进行类比，有助于提升透明度，并更全面地理解人工智能技术的用途与局限性。为形式化并标准化人工智能维护这一概念，我们希望将其与一个更为熟悉的案例——汽车维护——联系起来，因为人工智能与汽车在模型开发与部署方面存在许多共性。新汽车车型的开发是一个资源密集型过程（例如电动汽车）。人们理所当然地认为，在整个开发过程中，可靠性和安全性等必要的监管与法律要求都已得到充分认证，以避免灾难性故障、致命损害和严重的产品召回。同样，人工智能模型开发可能相当昂贵，尤其是在训练基础模型时 [8]，这类模型需要借助神经网络在大规模数据集上进行预训练，而神经网络包含数量庞大的可训练参数。以生成式预训练变换器3（GPT-3） [9] 为例，它是迄今训练过的规模最大的语言模型之一。GPT-3拥有1750亿个参数，并在包含4990亿个词元的数据集上进行训练。即使使用市场上价格最低的GPU云服务，其预计训练成本也约为460万美元 2020[6]。投入如此巨大的成本后，人们自然会期望最终得到的人工智能模型具备风险防护能力和鲁棒性，能够投入部署。

在部署过程中，汽车维护包括定期进行机械与电气检查、性能测试和认证、汽车零部件更换以及维修。我们认为，汽车维护中的许多熟悉概念都可以很好地映射到人工智能模型。下文将对汽车与人工智能进行类比，以促进人工智能鲁棒性维护体系的建立。

表1 总结了汽车与人工智能维护在稳健性方面具有类比关系的关键术语。下文将这些术语分为四类，并讨论它们之间的联系。

[6]https://lambdalabs.com/blog/demystifying-gpt-3/

表1. 汽车与人工智能模型在维护和鲁棒性方面的类比，分为四个类别。

类别	Car	AI
模型描述和性能表征	用户手册	模型规格
	汽车零部件	机器学习模块
	保修	鲁棒性检查点
	传动效率	内存/数据/功耗效率
系统性检查和监控	碰撞测试与安全报告	内部鲁棒性评估
	机械与电气检查	渗透测试和调试
	异常状态警告	操作错误
	健康状态监控	模型行为跟踪
修复与更新	维修	模型修复与更新
	车轮定位	模型校准
	冬季轮胎	模型强化
	爆胎响应	快速适应
教育与社会影响	驾驶执照	人工智能伦理与价值对齐
	可持续性	绿色与负责任的人工智能

3.1. 模型描述与性能表征

“用户手册”为 AI 系统提供说明，其中描述了确保透明度和问责制所需的信息，例如数据和模型训练的详细信息、隐私、可用性，以及关于建议用途和可能误用的影响声明。“汽车零部件”在人工智能领域指机器学习流水线中可修改且可配置的功能模块，理想情况下应实现标准化，以便于模型修复与更新。“保修”在人工智能领域指开发过程中的定性和定量性能检查点。“传动效率”在人工智能领域与模型如何随数据、内存和功耗扩展有关，例如每秒浮点运算次数（FLOPS）。

3.2. 系统性检查与监控

在模型开发过程中，人工智能的“碰撞测试”是指内部全面的鲁棒性评估、白帽黑客测试和红队测试，用于发现局限性和潜在问题，类似于全面的道路测试和汽车评测。测试结果可用于生成“安全报告”，量化模型在对抗性场景和未见场景中的鲁棒性水平。人工智能的“机械与电气检查”是指通过探测和主动测量，对整个系统（例如支持人工智能技术的软件和硬件）进行渗透测试和调试。“异常状态警告”是指在部署期间实时检测运行异常事件，例如错误实例或故障。“健康状态监控”是指持续跟踪模型行为，例如识别对抗性威胁的出现和数据漂移。

3.3. 修复与更新

在检查并识别错误和风险后，人工智能模型的“维修”是指修复、更新并重新认证底层模型的缓解策略。人工智能中的“车轮定位”指模型校准，“冬季轮胎”指通过采用更鲁棒的模块来加固模型，而“爆胎响应”则指在模型性能退化和发生异常事件时快速调整人工智能模型。根据发现的鲁棒性风险的严重程度、用户需求以及针对人工智能技术实施的法规，人工智能维护中的模型修复与更新可提供不同价位的差异化服务，范围从简单的模型修补和快速修复问题、模块替换、部分模型升级，到模型重建。

3.4. 教育与社会影响

人工智能的“驾驶执照”意味着在使用人工智能技术时接受有关伦理和价值对齐的教育，以了解其能力和局限性。人工智能的“可持续性”则包括增强环境意识，例如开发能源消耗更低、更环保的人工智能模型，同时实现积极的社会影响，以履行社会责任并防止可能的滥用。

4. 人工智能模型检查器

为了实践并实现人工智能维护这一理念，本节提出一种

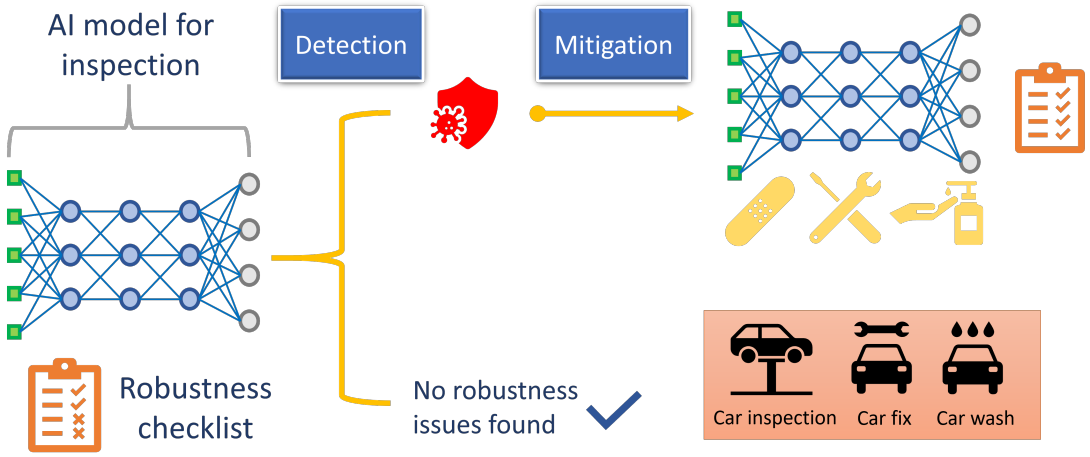


图 2. 人工智能模型检查器框架包含检测阶段和缓解阶段。待检查的模型首先要经过一系列鲁棒性测试和检查点，包括程序与运行评估、在代表性数据集上进行被动评估，以及通过即时生成新实例来查找失效模式的主动探测。在检测阶段，检查器会提取统计数据并进行诊断，以识别鲁棒性方面可能存在的风险。在缓解阶段，检查器会采用模型修复与更新来缓解已识别出的鲁棒性风险，然后使用同一份鲁棒性检查清单重新评估模型。最后，检查器会输出一个风险已缓解的模型。整个过程类似于汽车检查、维修和清洁等汽车维护工作。

名为人工智能模型检查器的方法，它是一套概念流程，用于在人工智能生命周期中主动检测并缓解鲁棒性问题。我们还重点介绍了两个针对不同鲁棒性挑战的案例研究，以说明如何实现人工智能模型检查器。最后，受车辆自主性的启发，我们定义了人工智能鲁棒性的不同等级。

4.1. 鲁棒性检查：检测与缓解

图2 展示了人工智能模型检查器的处理流程，该流程分为两个阶段：检测和缓解。首先，使用人工智能维护服务的用户会提供一个模型和/或一些数据样本，以进行鲁棒性检查。检查过程会以定性和定量方式开展一系列鲁棒性测试和检查点，包括程序与运行评估、在代表性数据集上对模型性能进行被动评估，以及通过即时生成新实例进行主动探测，以发现失效模式。定性评估包括向模型操作者征询系统特征描述和问题描述，以深入了解模型开发与部署的范围和详情，例如训练所用的模型和数据、模型的部署方式、用户掌握的信息量、最受关注的鲁棒性挑战类型等。基于定性评估，定量分析包括运行相应的诊断，并通过生成和利用适当的测试用例与数据集来评估性能，报告数值结果和总结。

具体而言，在检测阶段，检查器会提取判别性统计数据并进行诊断，以识别潜在的鲁棒性风险。随后，在缓解阶段，检查器会采用模型修复与更新来缓解已识别的鲁棒性风险，例如进行模型微调和重新训练、在 AI 系统中添加或替换某些模块，并使用同一份鲁棒性检查清单重新评估模型。最后，检查器会返回一个已缓解风险的模型。整个过程类似于汽车维护，包括汽车检查、修理和清洁。汽车维护中的差异化服务理念，也可以对应到不同的需求和

人工智能维护成本，例如快速扫描、全面检查、快速修补以及详细的修复与更新。我们注意到，人工智能模型检查器的使用是持续性的，而非一次性的。基于人工智能生命周期中各状态的反复出现，模型会在数据收集与处理、模型训练和模型部署等状态之间反复转换。此外，模型可能修复后又再次出现故障。这类似于预测性汽车维护中的疲劳和疲劳测试理念——经过检查后，某些部件需要定期更新或更换，以确保模型保持良好状态。

根据图中展示的鲁棒性挑战1，我们给出以下两个示例，以阐释人工智能模型检查器的概念。

后门检测与缓解：在检测阶段，检查器采用 [10]，提出的木马网络检测器，该检测器使用数量有限的未被篡改的干净数据样本（每类甚至只需一个样本），推导出一种判别统计量，以判断训练好的神经网络分类器是否存在隐藏后门。该检测器甚至能够对卷积神经网络进行无数据检测。检测完成后，检查器可以采用 [11] 提出的模型净化策略，通过微调模型参数来移除后门。

异常输入检测与缓解：对于待检查的人工智能模型接收的数据输入，检查器可以利用内部数据表示（例如与训练数据的相似度）、领域知识（例如固有数据特征和物理规则）或外部知识核查（例如搜索知识图谱或数据库并进行推理），判断该数据输入是否异常。此处，异常涵盖了多种鲁棒性挑战，例如对抗样本和分布外样本。例如，文献[12] 利用音频数据中固有的时间依赖关系检测自动语音识别中的音频对抗样本；此外，研究人员还提出了许多基于从模型中提取的内部数据表示的距离度量，以检测分布外样本 [13]。除了筛除异常输入外，检查器还可以进一步采取缓解策略来更新模型，增强其鲁棒性以应对异常输入。例如，文献[14]提出的自推进鲁棒训练方法，可以通过指导模型在模型微调期间缓解其自行发现的歧义，进一步增强已训练模型的对抗鲁棒性。

4.2. 面向鲁棒性的对抗机器学习

像火星探测漫游车这样的汽车，能够在新的、未曾探索过的地形上成功执行指定任务，这是因为它们是在全面的模拟环境中开发的。对于人工智能模型，可以利用模型检查工具生成的失效示例，提升模型在未知乃至对抗性环境中的鲁棒性。这种方法称为对抗机器学习，即在人工智能生命周期中引入虚拟对手，以帮助构建更出色、更鲁棒的模型。在开发阶段，虚拟对手的作用是模拟分布外或最坏情况场景，并生成新的挑战性案例，帮助模型更好地泛化到未知和对抗性环境中。在部署阶段，虚拟对手的作用是主动开展鲁棒性评估和风险发现，以防止实际损害和负面影响。一个典型示例是对抗训练 [15]，它利用模型训练过程中自行生成的对抗样本，增强模型在部署阶段抵御对抗性输入的对抗鲁棒性。关于人工智能鲁棒性领域对抗机器学习的最新进展，请参阅 [16]。

4.3. 通往人工智能鲁棒性各级别的路线图

受自动驾驶汽车驾驶自动化六个等级的定义启发[7]，我们定义了人工智能鲁棒性的六个等级，以便追踪技术进展、量化风险，并开展检查、模型审计和标准化。**表2** 分别比较了车辆自主性和人工智能鲁棒性的各个定义等级。鲁棒性等级量化了机器智能在实现鲁棒性方面的健全性进展。随着等级提升，意味着实践

[7] https://www.sae.org/standards/content/j3016_202104/

表 2. 车辆自主性与人工智能鲁棒性各级别的比较。

级别	车辆自主性	人工智能鲁棒性
0	无驾驶自动化	无鲁棒性（标准训练）
1	驾驶员辅助	分布偏移下的泛化
2	部分驾驶自动化	应对单一风险的鲁棒性
3	有条件驾驶自动化	应对多重风险的鲁棒性
4	高度驾驶自动化	对已知风险的普适鲁棒性
5	完全驾驶自动化	与人类价值对齐并增强的鲁棒性

并以更切实、更全面的方式保障鲁棒性。对于人工智能鲁棒性而言，等级越高，所考虑的鲁棒性风险覆盖面越广。我们认为，将人工智能鲁棒性的等级形式化，有助于讨论和实践与鲁棒性、安全性相关的人工智能标准化，例如 ISO/IEC JTC 1/WG 13 可信赖性工作组[8] 以及 ISO/TC 22/SC 32/WG 14 安全与人工智能工作组[9]。

第 0 级指标准模型训练过程中未采取任何风险缓解措施所达到的原始鲁棒性。第 1 级关注对自然出现的偏移数据分布的泛化能力，例如，针对由无关伪特征引起的分布变化（如在天空背景变化的情况下对交通标志进行分类）仍能保持鲁棒预测。第 2 级考虑针对单一风险（如对抗样本）的最坏情况鲁棒性；第 3 级则扩展到多重风险，例如，针对对抗样本、常见数据损坏和伪相关的多目标（但经过选择的）鲁棒性 [17]。第 4 级确保对所有已知风险都具有普遍鲁棒性。这里的普遍鲁棒性是指对所有已知鲁棒性风险都具有联合效力。最后，第 5 级使鲁棒性与以人为本的价值观和用户反馈相一致，并能够自动补充新的鲁棒性要求，以完善现有要求。根据应用的要求（如法律法规）和应用场景，部署前可能需要满足不同级别的先决条件。例如，某些高风险人工智能应用应通过更高级别的标准——类似于不同驾驶自动化条件（如在高速公路上行驶与在城市环境中行驶）所要求的必要条件

[8] https://www.iso.org/committee/45020.html [9] https://standards.iteh.ai/catalog/tc/iso/6ec701ad-7678-442d-b186-a84b9ba2bbdf/iso-tc-22-sc-32

针对不同的驾驶自动化条件（例如，在高速公路上行驶与在城市环境中行驶）。

值得注意的是，一级鲁棒性的评估很可能可以通过在代表性数据集或基准测试上进行静态评估来完成。然而，进入二级及以上级别后，对最坏情况鲁棒性性能的验证还需要模型干预，例如主动进行模型扫描和探测，以发现失效案例。此外，人工智能模型检查器会主动采取措施，检测并缓解潜在的鲁棒性风险；这不同于现有的框架，例如 Factsheets [18],Model Cards [19], 和 Datasheets for Datasets[20] ，后者仅采用被动的模型刻画与规范说明。最后，除了对人工智能进行维护之外，还可以采用人工智能来改进维护，例如预测性维护，它基于历史记录和风险预测，对人工智能模型采取预防性维护。

5. 结语

本文借鉴汽车车型的开发和部署，探讨了一种针对人工智能技术鲁棒性的全新维护框架。为了在人工智能生命周期中建立并提升可信度，我们提出了一种基于人工智能模型检查器原理的自动化、可扩展解决方案，用于检测并缓解因缺乏鲁棒性而产生的潜在风险。受车辆自主性的启发，我们还定义了不同的人工智能鲁棒性等级，以便对防风险人工智能模型进行形式化、评估、标准化和监管。随着人工智能技术以比汽车更广泛、更深入且更迅速的方式改变我们的生活、社会和环境，我们认为，探索人工智能维护已迫在眉睫且十分必要。除鲁棒性之外，人工智能模型检查器框架还可扩展至可信人工智能的其他维度，例如公平性、可解释性、隐私、问责制，

和不确定性量化。

参考文献

1. P.-Y. Chen 和 S. Liu，“深度学习模型的整体对抗鲁棒性”，*AAAI*人工智能会议论文集，2023年。2. R. S. S. Kumar、M. Nystr¨om、J. Lambert、A. Marshall、M. Goertzel、A. Comissoneru、M. Swann 和 S. Xia，“对抗机器学习：行业视角”，载于*2020*年 *IEEE*安全与隐私研讨会*(SPW)*，2020 年，第 69–75 页。3. M. Jagielski 、 A. Oprea、B. Biggio、C. Liu、C. Nita-Rotaru 和 B. Li，“操纵机器学习：回归学习中的投毒攻击与对策”，载于*IEEE*安全与隐私研讨会，2018年，第19–35页。4. T. Gu、K. Liu、B. Dolan-Gavitt 和 S. Garg，“BadNets：评估针对深度神经网络的后门攻击”，*IEEE Access*，第7卷，第47 230–47 244页，2019年。5. C. Xie、K. Huang、P.-Y. Chen 和 B. Li，“DBA：针对联邦学习的分布式后门攻击”，载于国际学习表征会议，2020年。6. I. J. Goodfellow、J. Shlens 和 C. Szegedy，“解释并利用对抗样本”，国际学习表征会议，2015年。7. P.-Y. Chen、H. Zhang、Y. Sharma、J. Yi 和 C.-J. Hsieh，“ZOO：基于零阶优化的黑盒攻击，无需训练替代模型即可攻击深度神经网络”，载于*ACM*人工智能与安全研讨会，2017年，第15–26页。8. R. Bommasani、D. A. Hudson、E. Adeli、R. Altman、S. Arora、S. von Arx、M. S. Bernstein、J. Bohg、A. Bosselut、E. Brunskill 等，“基础模型的机遇与风险”，*arXiv*预印本 *arXiv:2108.07258*，2021年。9. T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell、S. Agarwal、A. Herbert-Voss、G. Krueger、T. Henighan、R. Child、A. Ramesh、D. Ziegler、J. Wu、C. Winter、C. Hesse、M. Chen、E. Sigler、M. Litwin、S. Gray、B. Chess、J. Clark、C. Berner、S. McCandlish、A. Radford、I. Sutskever 和 D. Amodei，“语言模型是少样本学习者”，载于神经信息处理系统进展，第33卷，2020年，第1877–1901页。10. R. Wang、G. Zhang、S. Liu、P.-Y. Chen、J. Xiong 和 M. Wang，“木马神经网络的实用检测：数据受限和无数据情形”，载于欧洲计算机视觉会议，2020年，第222–238页。

11. P. Zhao, P.-Y. Chen, P. Das, K. N. Ramamurthy, and X. Lin, “损失景观中的模式连通性与对抗鲁棒性之间的桥接”，载于国际学习表征会议，2020年。12. Z. Yang, B. Li, P.-Y. Chen, and D. Song, “利用时间依赖性表征音频对抗样本”，国际学习表征会议，2019年。13. J. Yang, K. Zhou, Y. Li, and Z. Liu, “广义分布外检测：综述”，*arXiv*预印本 *arXiv:2110.11334*，2021年。14. M. Cheng, P.-Y. Chen, S. Liu, S. Chang, C.-J. Hsieh, and P. Das, “自进阶鲁棒训练”，*AAAI*人工智能会议论文集，2021年。15. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “迈向能够抵御对抗攻击的深度学习模型”，国际学习表征会议，2018年。16. P.-Y. Chen and C.-J. Hsieh, 机器学习的对抗鲁棒性。爱思唯尔，2022年。17. S. Paul and P.-Y. Chen, “视觉变换器是鲁棒学习器”，载于*AAAI*人工智能会议论文集，第36卷，第2期，2022年，第2071–2081页。18. M. Arnold, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovi´c, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski 等，“Factsheets：通过供应商符合性声明提升对人工智能服务的信任”，*IBM*研究与开发杂志，第63卷，第4/5期，第6–1页，2019年。19. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasser- man, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “用于模型报告的 Model Cards”，载于公平性、问责制与透明度会议论文集，2019年，第220–229页。20. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, “数据集的 Datasheets”，*ACM*通讯，第64卷，第12期，第86–92页，2021年。