文章编号:1674-8190(2023)03-026-15

面向适航符合性的智能航电系统认证研究进展

董磊[1,2],刘嘉琛[2,3],陈曦[1,2],肖女娥[1,2],梁博尧[2,3],张元珊[2,4]
(1.中国民航大学 科技创新研究院,天津 300300)
(2.中国民航大学 民航航空器适航审定技术重点实验室,天津 300300)
(3.中国民航大学 安全科学与工程学院,天津 300300)
(4.中国民航大学 中欧航空工程师学院,天津 300300)

摘 要:民用飞机航电系统引入人工智能/机器学习技术会带来可信性、不确定性和可解释性等问题,有必要通过有效的符合性方法向公众与利益攸关方证实智能航电系统的适航安全性。首先,分析了智能航电系统的等级分类和应用现状,阐述了现有指南和标准的适用性;然后,基于对当前研究成果的梳理,总结了智能航电系统认证框架实施流程及其技术细节;最后,给出智能航电系统在全生命周期各个阶段的符合性验证要求及实现方法建议,评估了符合性验证对现有适航体系的影响,为民用飞机智能航电系统的设计与认证提供了参考依据。

关键词:民用飞机;智能航电系统;适航符合性;可信性;安全性
中图分类号:V243 **文献标识码**:A
DOI: 10.16615/j.cnki.1674-8190.2023.03.03

Research progress of AI-based avionics system certification for airworthiness compliance

DONG Lei[1,2], LIU Jiachen[2,3], CHEN Xi[1,2], XIAO Nyu'e[1,2],
LIANG Boyao[2,3], ZHANG Yuanshan[2,4]
(1. Department of Science and Technology, Civil Aviation University of China, Tianjin 300300, China)
(2. Key Laboratory of Civil Aircraft Airworthiness Technology, Civil Aviation University of China, Tianjin 300300, China)
(3. College of Safety Science and Engineering, Civil Aviation University of China, Tianjin 300300, China)
(4. Sino-European Institute of Aviation Engineering, Civil Aviation University of China, Tianjin 300300, China)

Abstract: For the public and stakeholders, it is necessary to demonstrate airworthiness safety of AI-based avionics system in acceptable means of compliance, so as to solve problems of trustworthiness, uncertainty and explainability of civil aircraft avionics system caused by introducing artificial intelligence/machine learning technology. First of all, the classification and application status of AI-based avionics system are analyzed, and the applicability of existing guidelines and standards are expounded. Then, based on the review of the current research progress, the implementation process and technical details of the certification framework for AI-based avionics system are summarized. Finally, the requirements and implementation methods of the compliance verification of AI-based avionics system in each stage of the whole life cycle are put forward, and the impact of compliance verification on the existing airworthiness system is evaluated. The study provides the reference for the design and certification of civil aircraft AI-based avionics system.

Key words: civil aircraft; AI-based avionics system; airworthiness compliance; trustworthiness; safety

收稿日期:2022-05-30; **修回日期**:2022-07-05
基金项目:国家重点研发计划(2021YFB1600600);国家自然科学基金民航联合基金(U1933106)
中央高校基本科研业务费(3122022044);天津市研究生科研创新项目航空专项(2021YJSO2B09)
通信作者:刘嘉琛,jc0419@foxmail.com
引用格式:董磊,刘嘉琛,陈曦,等. 面向适航符合性的智能航电系统认证研究进展[J]. 航空工程进展,2023,14(3):26-40.
DONG Lei, LIU Jiachen, CHEN Xi, et al. Research progress of AI-based avionics system certification for airworthiness compliance[J]. Advances in Aeronautical Science and Engineering, 2023, 14(3): 26-40. (in Chinese)

0　引　言

随着大数据时代的到来和计算机性能的飞跃式发展,人工智能/机器学习(Artificial Intelligence/Machine Learning,简称 AI/ML)技术在诸多领域的应用都进入了爆发式增长的新阶段,是新一轮科技革命和产业变革的重要驱动力量[1]。近年来,国外权威航空安全机构、科研组织以及领军企业陆续开展了AI在民用航空领域的应用及认证研究,绘制了航空人工智能技术发展路线图[2-8],我国“十四五”规划纲要也明确提出了建设智慧民航的发展要求[9],围绕这一颠覆性技术主动权的争夺正在航空领域如火如荼地展开[10]。

AI/ML 技术快速发展的态势给民用航空领域带来重大机遇的同时也带来了前所未有的挑战。由于智能航电系统缺乏令人信服的可追溯性架构与指导性标准,导致其在民用飞机上的实际应用落地较为困难,无法利用传统方法表明智能航电系统满足适航要求。因此,如何从技术可靠性和可信性等方面向公众与利益攸关方证实 AI/ML 技术的适航安全性,是民用航空工业从自动化发展体系向智能化发展体系过渡需要解决的关键问题,也是采用智能航电系统的民用飞机进入市场的重要前提。

为此,本文在分析智能航电系统的等级分类和应用现状的基础上,阐述现有指南和标准的适用性,总结智能航电系统认证框架的实施流程、技术细节及其研究进展,给出智能航电系统在全生命周期各阶段的适航符合性验证要求及实现方法建议。

1　智能航电系统概述

1.1　智能航电系统定义

人工智能之父 J. McCarthy 对 AI 的定义是:“制造智能机器的科学与工程,特别是智能计算机程序”[11]。这是一个概括性的术语,包含多种技术,目标是让机器像人类一样思考和行动。机器学习是用算法解析数据,不断学习,对世界中发生的事做出判断和预测的一项技术,是 AI 领域的一个子集,现有的 ML 技术一般可以按照监督学习、无监督学习、半监督学习和强化学习进行分类。深度学习(Deep Learning,简称 DL)是一种自动学习特征的机器学习方法,它试图将原始的数据通过非线性的复杂模型转换为更高层次、更抽象的表达,可以使用非结构化或未标注的数据进行学习,目前 DL 在语音和图像识别方面取得的效果已经远远超过先前相关技术。人工智能、机器学习和深度学习的关系如图 1 所示。

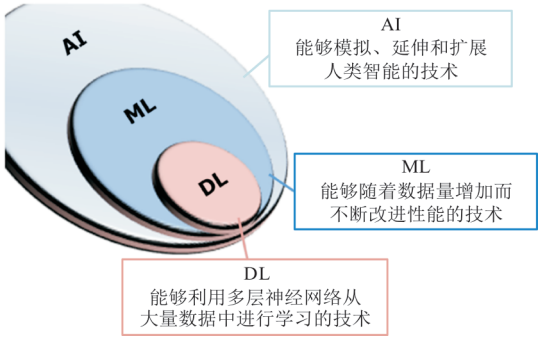

图 1　人工智能、机器学习和深度学习的关系

Fig. 1　Relationship of AI, ML and DL

航电系统是现代民用飞机的关键组成部分,主要功能是在飞机运行时完成信息采集、任务管理和导航引导等基本飞行过程,为飞行机组提供人机接口,确保飞行机组的态势感知和飞机系统管控能力,使得飞行机组能够及时、有效的管理和控制飞机安全、可靠地按照预定航迹飞行[12]。

对智能航电系统的定义包含两方面:一是基于人工智能技术提升飞机任务能力的航电系统;二是对人工智能技术应用友好且支持动态学习进化的航电系统。智能航电系统的分层映射关系如图 2 所示,从应用领域上包括任务规划、态势感知、故障预测等,从方法算法上包括遗传算法、决策树、深度学习、强化学习等[13-15]。

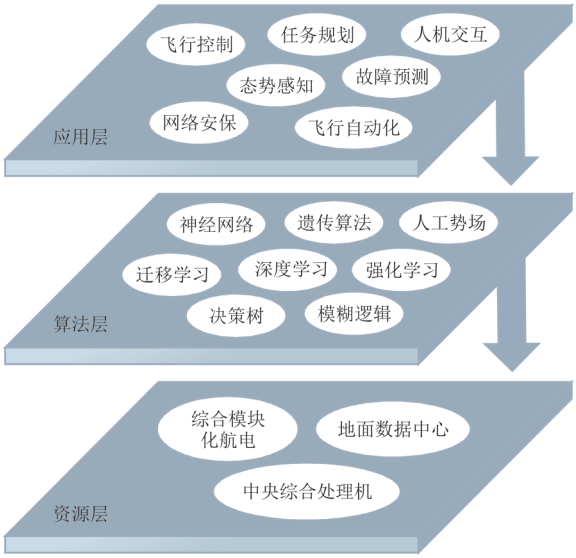

图 2　智能航电系统的分层映射关系

Fig. 2　Hierarchical mapping relationship of AI-based avionics system

智能航电系统的概念与综合模块化航电系统(Integrated Modular Avionics,简称IMA)的概念并不冲突,综合式架构组成资源层为智能航电系统提供了发挥的基础。

中国商用飞机有限责任公司于2020年“十四五”规划中提出了“有人监督模式下的大型客机自主飞行技术研究”技术指南,对短期、中期、长期3个阶段的研究方向予以指引[16]。同时,参考美国国家航空航天局(National Aeronautics and Space Administration,简称NASA)《航空战略实施规划》[17]、美国航空航天学会(Aerospace Industries Association of America,简称AIAA)《航空航天智能系统路线图》和欧洲航空安全局(European Organization for Civil Aviation Equipment,简称EASA)《人工智能路线图》[3]对智能飞行发展趋势的判断,基本可以确定智能航电系统的发展路线应该从辅助功能(1级)开始,朝着更多的人机协作(2级)迈进,最后寻求机器的更多自主权(3级)。民用飞机智能航电系统的自主等级分类细节如表1所示。

表1 民用飞机智能航电系统的自主等级分类

Table 1 Autonomy level classification of civil aircraft AI-based avionics system

智能飞行阶段[16]	研究方向描述	自主等级	系统功能分配方案[18]
辅助智能飞行阶段(2020—2025年)	增强飞机综合感知能力,实现全飞行场景的机组决策辅助,改善机组操作负荷,实现全飞行阶段的自动驾驶	1A级 对人的辅助	信息获取和分析的自动化支持
增强智能飞行阶段(2025—2035年)	基于实时空地信息交互,具备完善的飞机全势态感知能力,实现在机组监督下的自主运行,实现单一飞行员驾驶	1B级 对人的协助	在决策和行动选择方面对飞行员的认知协助
		2级 人机协作	飞行员能够密切监督分配给AI系统的功能,具备干预人工智能系统决策和/或运行的能力
		3A级 部分自主	飞行员具备监督AI系统运行的能力,当需要确保运行安全时,能够超越AI系统的权威进行决策
完全智能飞行阶段(2035—2050年)	基于空天地一体的信息融合平台,实现对复杂飞行情景的完整理解,实现满足人类弹性需求的全自主飞行	3B级 完全自主	飞行员不具备干预AI系统决策和/或运行的能力

1.2 智能航电系统应用现状

早在1993年,美国联邦航空管理局(Federal Aviation Administration,简称FAA)飞行安全研究处的L. H. Harrison等[19]就已经对AI技术在航空电子领域的应用情况进行了概述,还基于RTCA DO-178B[20]考虑了与认证相关的注意事项。2003年,NASA对智能飞控系统(Intelligent Flight Control System,简称IFCS)进行了测试,此系统将神经网络技术与先进的控制算法相结合,能够识别飞机气动特性的变化并做出响应,在遭遇意外故障时系统能够立即进行调整以保持最佳飞行性能(如图3(a)所示)。2019年,美国国防高级研究计划局(Defense Advanced Research Projects Agency,简称DARPA)启动了“空战进化”(Air Combat Evolution,简称ACE)计划,旨在通过AI技术来处理视距内空中格斗问题,在其Alpha空战格斗比赛中,苍鹭系统公司的智能空战代理与经验丰富的顶尖人类飞行员进行交战,取得了惊人的成功(如图3(b)所示)。2020年,空中客车公司在“自主滑行及起降”(Autonomous Taxi, Take-off & Landing,简称ATTOL)项目中实现了基于图像识别技术的全自动起飞技术,此技术可以在起飞过程中自动加速,保持对齐跑道中心线并在合适的时机抬起机头,全自动起飞技术已经在A350飞机上测试成功(如图3(c)所示)。2021年,波音公司通过学习型组件(Learning Enabled Components,简称LEC)自动驾驶飞机,实现了忠诚僚机的首飞,这项成果将为空中力量编队系统(Airpower Teaming System,简称ATS)提供支持(如图3(d)所示)。2022年,在驾驶舱自动化系统(Aircrew Labor In-Cockpit Automation System,简称ALIAS)项目的支持下,一架无人驾驶的UH-60A黑鹰直升机完成了30 min的自主飞行,此系统能够根据给定的任务目标和约束独立执行飞行计划(如图3(e)所示)。同年,航空工业第一飞机设计研究院研制的TP500无人运输机首飞成功,此机型将人工智能与航空制造、航空运输业相融合,操作智能化程度

高,多数场景下飞机可根据环境参数变化自主决策、自主飞行(如图3(f)所示)。

(a) 智能飞控系统飞行测试[21]

(b) Alpha空战格斗比赛[22]

(c) 基于图像识别的自主起飞[23]

(d) 波音公司忠诚僚机首飞[24]

(e) 黑鹰直升机自主飞行[25]

(f) TP500运输机首飞[26]

图3　典型智能航电系统应用

Fig. 3　Typical AI-based avionics system applications

伴随着相关研究持续开展,越来越多的航电系统有望具备先验知识储备、学习、认知和自适应等能力[8]。这使得采用人工智能算法从大量数据中获得知识,从而提升系统能力并逐步替代飞行员成为可能。遗憾的是,截至目前还没有通过适航认证的智能航电系统,民用飞机智能航电系统在各领域的认证可行性如表2所示,其具体考虑可查阅文献[27]。

表2　民用飞机智能航电系统在各领域的认证可行性

Table 2　Certification feasibility classification of civil aircraft AI-based avionics system in various domain

领域	子领域	学习类型	自主等级	失效状态类型	认证要求	认证经验
飞行自动化	完全自主飞行	在线	3B	灾难性的	强制	无经验
	感知与规避	离线	3A/B	灾难性的	强制	无经验
	视觉导航	离线	3A/B	灾难性的	强制	有部分经验
	部分任务自动化	离线	3A	灾难性的	强制	有部分经验
	自适应飞行员辅助	在线	1A/B	无安全影响	推荐	N/A
	基于规则的飞行员辅助	离线	1A	无安全影响	推荐	N/A
飞行控制	自适应控制器	在线	3B	灾难性的	强制	无经验
	高效控制器	离线	3B	灾难性的	强制	有部分经验
自然语言处理	自主处理空管呼叫	离线	3A/B	灾难性的	强制	无经验
	协助处理空管呼叫	离线	3A	轻微的	强制	无经验
	驾驶舱语音控制	离线	2	重大的	强制	有部分经验
	高级语音建议	离线	1A	轻微的	推荐	N/A
	客舱语音控制	离线	1A	无安全影响	不要求	N/A
网络安保	异常检测	在线	1A/B	无安全影响	不要求	N/A
	自主对抗措施	在线	2-3A/B	无安全影响	强制	无经验
故障预测	在线学习预测	在线	1B	无安全影响	不要求	无经验
	预训练预测	离线	1B	无安全影响	不要求	N/A

1.3　现有指南和标准的适用性分析

以往,以 ARP4754A、ARP4761 及 DO-178C 等为代表的研制过程指南对提高机载产品的质量起到很好的引导作用。这些指南旨在通过对研制过程的控制尽早发现并剔除设计错误,为获取航

电系统的安全性需求以及判断软、硬件的适航符合性提供了重要准则，并被公认为行业最优的实践方法[28]。但是，由于AI/ML技术框架、算法和自适应学习的复杂性，可能会出现智能航电系统行为难预测、难解释以及非预期结果等问题，研究者普遍认为目前的研制保证过程难以为AI/ML技术提供充分的符合性保证[29-32]。鉴于此，下文将会对现有指南和标准进行差距分析，阐述其对于智能航电系统的适用性。

SAE ARP4754A[33]指导高度综合复杂飞机系统的研制，为局方和申请人提供审定方面的指南，以最大限度地降低系统研制过程中出现错误的风险。与SAE ARP4761[34]相结合则提供了用于大型飞机及其高度集成系统研制的安全性评估指导。对于智能航电系统，虽然研制保证等级（Development Assurance Level，简称DAL）的概念仍然适用，但申请人需针对AI/ML模型的可解释性和不确定性证实智能航电系统的研制过程具有充足的验证手段，以确保所有潜在的研制错误已经被控制在可接受的范围内。智能航电系统的需求定义方法也要进行调整，与DAL相关的数据集需求和AI/ML模型的性能需求值得特别关注。此外，由于需求捕获原理的变化和ML算法的概率性质，需求验证和实现验证的方法也需要更新。

RTCA DO-178C[35]是商用航空电子软件开发的首要标准，为机载系统和设备软件的开发提供指导，旨在使民用航空产品使用的软件能够满足适航性要求，并获得使用批准。机载软件安全性的关键在于对开发过程的保证，但由于AI/ML技术与传统软件之间存在着本质的区别，DO-178C的过程保证流程已经无法适应智能航电系统。例如：①ML模型的拓扑结构和权重无法追溯到开发过程中的系统需求，不符合DO-178C中基于需求的验证过程；②由于ML模型高度复杂且具有非线性运算特性，传统软件结构覆盖的度量指标不再适用；③虽然参数数据项（Parameter Data Items，简称PDI）可以方便地用于存储和管理由学习过程产生的ML模型参数，但ML模型神经元的每次激活都会修改已批准的配置，因此也不适用。

RTCA DO-200B[36]为航空数据处理提供了最低要求和指导，这些航空数据用于导航、飞行计划、地形/障碍感知、飞行显示界面、飞行模拟器和其他应用。此标准旨在保证随着时间的推移建立和维护一定水平的数据质量，数据质量需求（Data Quality Requirements，简称DQRs）包括准确性、分辨率、可追溯性、及时性和完整性等，数据质量的概念可以用于ML数据集的准备。

ISO/PAS 21448[37]围绕汽车自动驾驶系统的性能限制或人员可预见的误用而造成的危害，提供了实现预期功能安全所需的适用设计和验证措施的指南。其中性能受限的例子包括传感器限制、AI/ML算法限制等。此标准是对ISO 26262[38]的补充，尽管尚未针对ML技术进行认证，但涵盖了功能安全无法追溯到的功能故障，可以作为智能航电系统安全性评估工作的输入。

2 基本框架与关键技术

2014年，时任FAA飞机计算机软件首席科学与技术顾问的Mike Dewalt提出了“技术独立保证方法”（Technology Independent Assurance Method，简称TIAM）[39]，此方法作为一种新的认证框架并不十分规范，但为日后的“总体属性”倡议奠定了基础。该倡议作为精简过程保证方法的一部分，目标是为了克服现有指南与标准不能与技术发展保持同步的问题，定义了独立于技术和领域的少量总体属性，覆盖了ARP4754A、DO-178C和DO-254中的多个离散目标[40-41]。若申请人能证明系统符合总体属性，则系统就可以被认证，目前已经确认的总体属性有三项，分别是：

①设计意图：就所需的系统行为而言，定义的预期功能是正确和完整的；

②正确性：在可预见的使用条件下，就其定义的预期功能而言，系统实现是正确的；

③无害性：超出系统预期行为之外的实现不会产生不可接受的安全影响。

可以看出，总体属性的思想适用于解决智能航电系统认证过程中面临的一系列问题，但目前的研究成果过于抽象，总体属性还无法作为完整且可操作的符合性验证方法。如果可以在实践中进一步细化，总体属性将会是智能航电系统可以考虑的认证框架。

与FAA的“总体属性”不同，2019年EASA提出了“抽象层”的概念[42]，其目标是在现有软硬件标准之上，捕获系统认证所需且独立于所使用技术的属性。抽象层虽然是一种自下而上的方法，但与总体属性的缺点类似，暂时也不具备对智能航电系统进行认证的可操作性。

综上所述，FAA和EASA已经为航电系统引入AI/ML提供了开放性的解决方案，但仍需要在

现有适航体系的基础上,进一步更新和补充标准化认证框架[43-44]。本节基于现阶段的最新研究进展,采用智能航电系统认证框架(如图 4 所示),由可信度分析、安全性评估、安全风险缓解措施和认证/批准活动 4 部分组成。框架各部分的实施流程和技术细节分述如下。

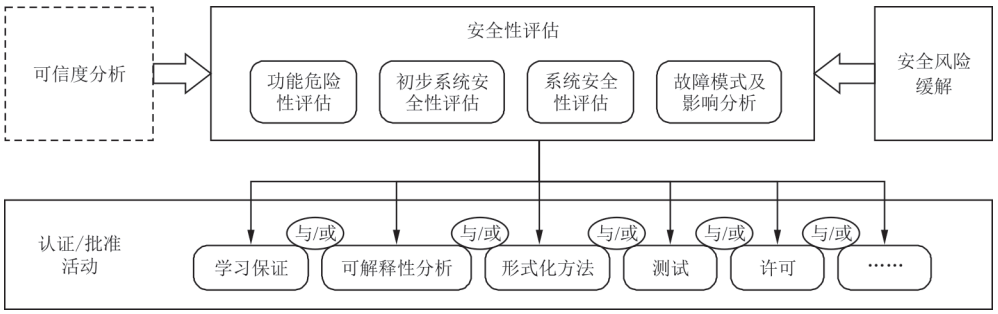

图 4　智能航电系统认证框架

Fig. 4　Certification framework for AI-based avionics system

2.1　可信度分析

在 AI 的应用潜力得到广泛认可的同时,大量研究也指出“可信”是 AI 赋能行业应用的必要前提。AI/ML 技术在安全关键系统中的错误预测已经产生了人们无法承受的后果,例如优步自动驾驶汽车未能及时识别路上行人而致其死亡,IBM Watson 医疗中心对癌症患者给出的错误诊断等案例。在这样的背景下,针对 AI/ML 系统的可信度分析在学术界被广泛讨论,已成为备受关注的研究热点[45]。

可信度分析从宏观方向上指导 AI/ML 系统向着可信的方向发展,主要关注两个原则:一是功能性原则,要求 AI/ML 系统在技术和功能上可信,即要求 AI 模型具有较高的准确率、泛化性等,这是 AI/ML 系统应用的共识前提;二是伦理性原则,即 AI 决策在保证系统性能准确的前提下,应符合人类社会的道德伦理准则和法律法规,才能得到人类的信任[46]。尤其是在民用飞机的适航体系中,需要通过正向分析来证明技术的可靠性。智能航电系统可信度分析的重要性和全新特质值得特别关注[47]。

为了尽可能扩大人工智能的收益并降低其风险,全球已经制定了近 100 项可信 AI 伦理准则来应对人工智能的信任危机。具有代表性的可信 AI 伦理准则如表 3 所示,虽然各类准则之间存在一定的重叠和冲突,优先级关系和内部边界还没有清晰地被界定,但安全与隐私、可问责性和可解释性始终是可信 AI 伦理准则的重点关注对象[48-49]。

表 3　代表性的可信 AI 伦理准则

Table 3　Representative ethical principles for trustworthy AI

关键词原则	可信 AI 伦理准则					
	①	②	③	④	⑤	⑥
安全与隐私	√	√	√	√	√	√
可问责性	√	√	√	√	√	√
可解释性	√	√	√	√	√	√
公平与偏见	√	√	√	√		√
鲁棒性			√	√	√	√
自主性	√		√			√
可追溯性			√	√	√	√

注:①2017 生命未来研究所《阿西洛马人工智能准则》[50];②2017 美国电气电子工程师协会《AI 设计伦理准则》[51];③2019 欧盟《可信人工智能的伦理准则》[52];④2019 二十国集团《G20 AI 准则》[53];⑤2020 美国白宫《在联邦政府中推广可信人工智能的行政命令》[54];⑥2021 中国科技部《新一代人工智能伦理规范》[55]。

在民用航空领域,EASA 已经基于《可信人工智能的伦理准则》明确了智能航电系统全生命周期内的三个重要可信支柱,即合法性、遵守伦理原则和技术鲁棒性[56]。此外,EASA 还通过“可信 AI 评估列表”为申请人提供可信度分析的指导并确定了若干目标和预期符合性方法[57]。列表包括以下 7 个部分:①人类自主性与监管,②技术鲁棒性与安全,③隐私与数据管理,④透明性,⑤多样性、非歧视与公平,⑥社会与环境福祉,⑦可问责性。

对智能航电系统的可信度分析将产生功能性与非功能性系统需求以及应实现和验证的组织需求。智能航电系统的认证框架应从可信度分析出发,通过后续的安全性评估、安全风险缓解和认证/批准活动对可信 AI 伦理准则的关注要素给出

具体实施方法。

2.2 安全性评估

民用飞机系统安全性评估是在整机及机载系统研制过程中确定定性和定量的安全性目标，并采用相应的分析与评估技术来证明已达到安全性目标的方法[58]。随着航空技术的发展，航空器功能的增加给系统安全性评估工作带来了挑战，民用飞机安全性分析方法及标准是随着航空工业的发展而不断变化的[59]。

根据D. Amodei等[60]和J. M. Faria[61]的研究成果，可以总结出一系列使用AI/ML技术引发的典型失效模式，例如对指定函数的访问受限、训练数据不足或管理不善、模型表达不充分等。2020年，C. Smith等[62]引入了危险贡献模式(Hazard Contribution Modes，简称HCM)的概念，用于对LEC的危险系统状态进行分类，可以为智能航电系统的功能危险性评估和故障模式及影响分析提供参考。2021年，D. Cofer[63]认为LEC可能会由于未知的输入而导致潜在的危险输出，其回顾了当前机载软件系统中检测非预期行为的原理和技术，并研究了LEC在安全关键领域应用的安全保证方法，包括形式化方法、新的测试方法、新的覆盖度指标以及运行时保证架构。

对于智能航电系统，至少需要与传统航电系统保持相同的安全性水平，引入AI/ML技术不应对人员和财产造成更高的风险。系统设计阶段的智能航电系统安全性评估流程如图5所示，深色部分是由于AI/ML技术的引入而新增的步骤。

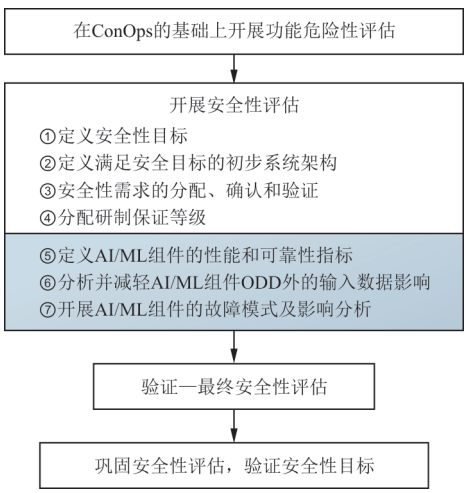

图5 AI/ML系统的安全性评估流程[56]

Fig. 5 Safety assessment process for AI/ML system[56]

ConOps是AI/ML系统的运行概念，详细说明了系统应该如何运行，重点放在运行设计域(Operational Design Domain，简称ODD)的构建以及特定运行限制和假设的捕获上。一个ODD是符合一个或一组场景描述的参数集合，明确了系统正常运行的条件及约束，AI/ML系统只允许在其ODD中运行[64]。对ConOps的精确定义旨在最大限度地保持AI/ML系统运行的“安全区”，减少由于ML算法性能受限而导致的不安全运行场景，确保了足够且具有代表性的训练、验证和测试的数据集，能够给智能航电系统的安全性评估提供必要的输入。

2.3 安全风险缓解

随着航电系统复杂程度和智能化的不断提高，设计一个完全具备鲁棒性和可解释性的智能航电系统是不切实际的，现有方法不能保证AI/ML组件在系统的整个生命周期中不会发生失效或故障。为了保持系统尽可能运行在预期的范围内，可以采用安全风险缓解(Safety Risk Mitigation，简称SRM)将剩余安全风险降低到可接受的水平。

运行时保证(Runtime Assurance，简称RTA)是一种典型的SRM技术，其主要思想是用监控模块和控制模块包围不安全的AI/ML组件。监控模块对系统复杂功能的输入输出和内部状态变量进行检测，以此判断系统是否在正常功能域内运行。若系统在运行时检测到异常，RTA切换器将会进行重新配置，用备用功能代替异常的复杂功能，以保证影响飞行安全的功能正常运行[65]。美国材料与试验协会(American Society of Testing Materials，简称ASTM)于2017年发布的ASTM F3269-17[66]已经以标准实践的形式体现了RTA的核心原则，典型的RTA架构如图6所示。

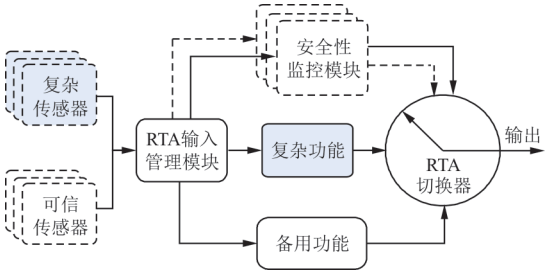

图6 典型运行时保证架构[66]

Fig. 6 Typical runtime assurance architecture[66]

2020年，J. D. Schierman等[67]提出了基于无人机系统（Unmanned Aircraft System，简称UAS）的RTA架构，研究发现RTA能够通过系统交互检查安全性、完整性以及输入输出的有效性，但具有多级交互反馈的系统也会显著增加RTA框架的复杂程度；D. Cofer等[68]提出了飞机智能滑行系统的RTA架构，通过架构分析与设计语言对其进行建模并进行形式化分析，验证了该架构在不同运行场景下的安全性；C. Lazarus等[69]基于马尔可夫决策过程设计了高安全性的RTA架构，使用强化学习方法确定RTA系统的配置切换策略，确保系统始终在安全范围内运行；2021年，B. Wheatman等[70]建立了分布式智能控制系统的RTA架构，利用黑盒监控模块检测AI/ML系统是否出现故障，利用白盒监控模块预测AI/ML系统决策的正确性，结果表明此方法有助于最大限度地提高系统整体性能。

2.4　认证/批准活动

智能航电系统的认证/批准活动是作为可能适用的符合性方法提出的，覆盖智能航电系统生命周期的一个或多个阶段。认证/批准活动的深度可以根据智能航电系统采用的系统架构、人工智能技术的复杂度和分配的研制保证等级灵活调整，各项活动的研究进展分述如下。

1）学习保证

在当前的监管框架下，系统、设备和部件的安全风险主要通过基于需求的研制保证过程来控制。但智能航电系统研制过程的最大挑战是确保采样数据集的训练能够在未知运行数据上以足够的性能进行泛化，因此需要将保证的重点转移到对数据的正确性、完整性、代表性以及学习过程的验证层面。虽然近年来有部分研究定义了面向AI/ML系统开发的生命周期过程，但大多数都不具备足够的细节以支持航空领域的认证/批准活动[71-74]。为了尽可能打开人工智能黑盒，使AI/ML系统能够正确执行预期功能，EASA于2020年提出了一种W型“学习保证”流程（如图7所示），作为对传统研制保证过程的针对性改进。

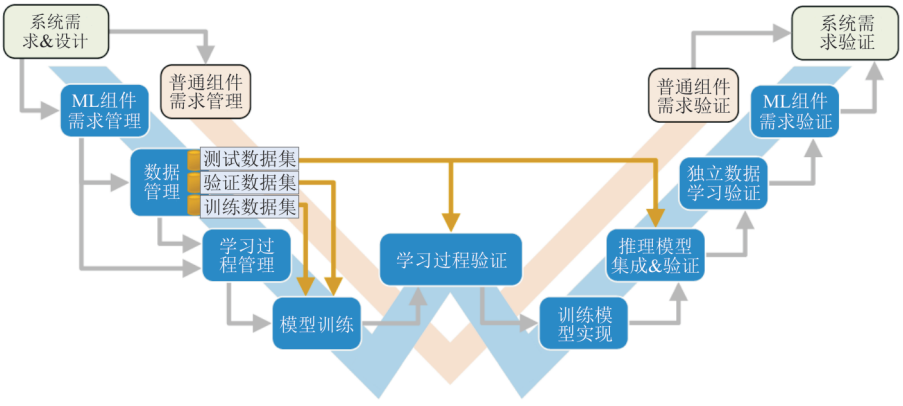

图7　EASA提出的W型学习保证流程[4]

Fig. 7　The W-shaped learning assurance process proposed by EASA[4]

美国航空航天飞行器系统研究所（Aerospace Vehicle Systems Institute，简称AVSI）也于2020年提出了一个考虑ML系统认证的开发流程，与W流程相比有一定的相似之处，但也有一些差异值得注意：①两个流程对数据集的命名规则不同；②ML系统开发流程中对范式选择和范式验证的关注并未在W型学习保证流程中体现；③尽管ML系统开发流程的重点是数据集的管理，但并未明确提及W型学习保证流程中针对独立数据的验证活动[75]。欧洲民用航空设备组织（European Organization for Civil Aviation Equipment，简称EUROCAE）的WG-114工作组和美国汽车工程师协会（Society of Automotive Engineers，简称SAE）的G-34工作组于2021年提出的ML系统开发生命周期如图8所示，虽然没有具体规定开发模型（如瀑布模型、V模型和W模型）、学习环境和AI/ML技术，但该生命周期与航空标准化框架实现了完全集成。目前工作组已经分为了7个子团队，在关注声明的基础上逐步着手制定AI/ML系统认证的符合性方法[76]。

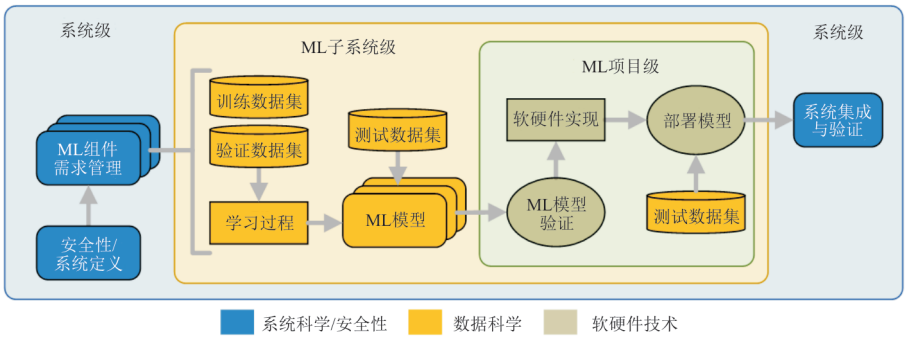

图 8 WG-114/G-34 工作组提出的 ML 系统开发生命周期[7]

Fig. 8 ML system development lifecycle proposed by WG-114/G-34 working group[7]

2) 可解释性分析

2016 年，美国国防高级研究计划局在其可解释人工智能研究计划中首次较为完整和明确地阐述了关于 AI 可解释性的概念，即一整套能够产生更多可解释模型，能够维持高水平学习性能，能够使用户理解、信任和有效管理 AI 的机器学习技术[77]。然而，由于 AI/ML 模型具有高度的非线性运算特性，导致其可解释性较差，无法应用于一些对安全性要求较高的关键领域。

对智能航电系统的认证需要可解释性分析的支持，这不仅能增加系统研制人员对 AI/ML 模型决策的理解与信任，也能帮助诊断出影响模型性能的因素，加以改进，进一步提升模型性能[78]。目前已有部分 AI 框架开始支持可解释性的需求，比如基于 PyTorch 框架出现了 Captum 等可解释库支持，基于 TensorFlow 出现了 TF-explain 等库支持，以及同时支持 PyTorch 和 TensorFlow 的 AIX 360、Alibi 等可解释库。国内则有 MindSpore 的 MindSpore XAI，PaddlePaddle 的 InterpretDL。另外，已经有一些平台从可解释的角度出发对模型进行评测，例如启智社区的重明平台、瑞莱智慧平台等。

在学术领域，对 AI/ML 可解释性的研究也逐渐增多。2018 年，Zhang Y F 等[79]认为对深度神经网络的解释无需打开整个底层网络的连接权重、隐含层和特征矩阵，而应当是一种由用户驱动生成的解释路径，具体方案包括伴随变动、基于一致性的方法、基于分歧的方法和基于调节的方法；2020 年，P. Linardatos 等[80]对可解释人工智能（Explainable Artificial Intelligence，简称 XAI）的研究进行了分类和综述，包括解释复杂黑盒模型的方法、创建白盒模型的方法、促进公平和限制歧视的方法以及分析模型预测敏感性的方法；2022 年，雷霞等[81]重点从解释深度学习模型的逻辑规则、决策归因和内部结构表示三个方面出发，介绍了几种可解释性研究的典型模型和算法，并指出了深度学习可解释性未来可能的发展方向；同年，刘潇等[82]定义了强化学习可解释性（Explainable Reinforcement Learning，简称 XRL）的 3 个独有问题，即环境解释、任务解释、策略解释；之后，对现有方法进行了系统的归类，并对 XRL 的最新进展进行综述。

3) 形式化方法

形式化方法是一种建立在严格数学模型基础上的用于设计、规范和验证的方法，广泛地应用在软硬件、通信协议、嵌入式控制系统等方面[83]。DO-178C 及其增补的 DO-333[84]首次正式将形式化方法引入到航空机载软件开发领域并确定了其有效性[85]。同时，信息技术安全评价通用准则（The Common Criteria for Information Technology Security Evaluation，简称 CC）要求评估保障等级（Evaluation Assurance Level，简称 EAL）5 级以上的软件必须使用形式化方法进行认证[86]；ISO 26262 标准也推荐在高安全完整性等级的系统软件开发中运用形式化方法。

目前，AI/ML 领域的形式化方法也引起了学术界和工业界的高度关注。2018 年，N. Fulton 等[87]提出了一种可证明的安全学习方法，此方法具有强化学习的探索和优化能力以及形式化验证的安全保证能力，并通过一个汽车自适应巡航控制模型验证了该方法的可行性；2019 年，T. Dreossi 等[88]开发了面向 AI/ML 系统的形式化设计与分析工具包 VerifAI，并将时态逻辑证伪、基于模型的系统模糊测试、参数合成、反例分析和数据集扩充作为案例进行展示；2021 年，C. Urban 等[89]回顾了形式化方法在航空机载软件领域的应用情况，对迄今为止面向 ML 开发的形式化方法进行了全面

而详细的介绍,包括可满足模理论(Satisfiability Modulo Theories,简称SMT)、优化和抽象解释技术等,讨论了支持向量机和决策树的集成方法以及模型训练和数据准备的方法,并对AI/ML系统形式化验证的未来研究方向进行了展望。

4)测　试

在民用航空领域,智能航电系统的错误行为可能会导致无法挽回的严重后果。因此,对智能航电系统进行充分的测试以尽可能地避免错误行为是必要的。虽然传统机载软件测试技术已经日趋成熟,但由于AI/ML技术与传统软件之间存在本质区别,传统软件测试技术无法直接应用于智能航电系统的测试中。对智能航电系统的测试需要把更多的注意力放到AI/ML模型神经元本身的状态以及神经元之间的互动关系上,目标是找出智能航电系统运行时的错误行为并及时改正[90]。

2018年,Sun Y等[91]受DO-178C中修正条件/判定覆盖(Modified Condition/Decision Coverage,简称MC/DC)测试方法的启发,设计了针对神经网络的测试方法,提出了符号—符号覆盖、距离—符号覆盖、符号—值覆盖以及距离—值覆盖四种覆盖准则;Pei K等[92]提出首个系统性测试深度学习系统的白盒框架DeepXplore,该框架能生成输入来触发深度学习系统逻辑的不同部分,并可以在没有手动干涉的情况下识别深度学习系统的不正确行为;2019年,Sun Y等[93]提出了第一个针对深度神经网络的concolic测试方法,实验表明该方法在获取高覆盖率和寻找对抗样本方面是有效的。2021年,PaddlePaddle推出了PaddleSleeve模型安全工具,完整提供了具有从AI模型鲁棒性评估测试,到模型攻击防御,再到模型鲁棒性提升能力的一整套工具;同年,MindSpore也推出了鲁棒性测试工具MindSpore Armour,基于黑白盒对抗样本、自然扰动等技术,提供高效的鲁棒性评测方案,帮助用户评估模型的鲁棒性并识别模型脆弱点。

5)许　可

NASA于2015年发布的《自适应系统认证注意事项》[94]提到,对先进自适应技术的认证需要完全脱离当前模式。应参考飞行员和空管员等从事关键民航工作人员的培训流程,将其衍生到智能航电系统中也可以得到相应的许可流程。与人类不同,智能航电系统可能需要数十万小时的模拟飞行和数千小时的实际飞行,发现并纠正大量故障之后才能得到认证,这种许可流程可以带来以下关键优势:

①许可将更多的注意力集中在系统的实际性能上,而不是像DO-178C过多地关注于开发过程和提供符合性证据;

②一旦高保真仿真技术发展到足以培训和测试智能航电系统时,通过许可的方法能在很大程度上降低认证新系统或修订系统的成本;

③传统认证的无错误保证使得系统研制人员不会继续测试或验证已通过认证的航电系统,而经过许可认证的智能航电系统会有可靠的性能记录,通过适当的立法能够使系统研制者免除相应法律责任,使得研制人员在系统取得认证后仍然能及时纠正可能存在的缺陷。

2019年,J. Nuñez等[95]在虚拟智能系统人机路线图(Human Aircraft Roadmap for Virtual Intelligent System,简称HARVIS)项目中对未来航空器运行场景、单一飞行员驾驶和自适应人机交互界面进行了研究,考虑了人工智能的“许可”认证概念并在欧洲航空航天高级培训联盟(The European Consortium for Advanced Training in Aerospace,简称ECATA)进一步发展;2022年,C. Regli等[96]面向自适应飞行自动化系统开发了一种自上而下的测试方法,定义了电子飞行教员/检察员和飞行自动化系统之间的接口,形成类似于人类飞行员必须通过的技能测试和熟练程度检查大纲,这种拟人化的方法评估了自适应系统的整体输出,拓展完善了智能航电系统的许可认证框架。

3　适航符合性验证挑战

3.1　可能适用的符合性验证方法

局方适航管理的基本原则是以适航规章的形式提出适航技术和管理要求,但不限定表明适航符合性的方法。目前适航管理程序已经对常用的符合性方法进行了分类,随着机载产品从简单到复杂的变化,申请人需根据产品特点选取其中的一种或多种组合的方式来表明符合性,如有更为明确完整的符合性方法定义与说明,亦可作为符合性审定计划的一部分[97-98]。基于AI/ML产品生命周期各个阶段,紧扣基本认证框架与关键技术,提出了适用于智能航电系统的符合性验证要求及需要考虑的注意事项,如表4所示。

表 4 智能航电系统适航符合性验证中需要考虑的注意事项

Table 4 Considerations for airworthiness compliance verification of AI-based avionics system

符合性验证要求	需考虑的注意事项
安全性评估	应考虑飞行员与 AI/ML 系统的权限分配、人机交互，并在 ODD 中描述运行方案 应根据 ConOps 进行功能危害性评估 应对 AI/ML 系统架构进行分析，确保其符合安全目标 应生成安全性需求，包括满足安全目标和系统架构的独立性需求 应给 AI/ML 组件分配合理的研制保证级别 应分析和减轻 AI/ML 组件暴露于 ODD 之外的输入数据造成的影响 应建立 AI/ML 组件失效分类，评估可能的失效模式和相关的检测手段 应在安全性分析过程中考虑 AI/ML 的概率性质，相关架构缓解措施应向下延伸至应用阶段
数据选择和验证	考虑到数据驱动学习过程的特殊性，数据质量需求应从正确性、完整性、代表性、公平性、独立性、可追溯性和及时性等方面进行描述 应根据数据和 ODD 的特征，使用合适的方法实现数据归一化 应定义不同数据集及其数据的配置管理过程和活动 当数据集质量属性可能被改变时应考虑工具鉴定，包括合成的/增强的数据或自动标记 应保护数据集管理环境，以避免训练集中毒和对抗性攻击
模型选择、训练与验证	应确保选择合适的网络模型架构，包括模型拓扑、层数和每层节点数等 应描述训练过程中关键要素的选择和验证，包括算法、成本函数、参数以及超参数等 应记录模型训练损失函数和误差度量的训练曲线，还应记录具有验证数据集的模型性能 应定义模型训练阶段的停止标准 应根据分配给 AI/ML 组件的研制保证等级确定模型解释的详细程度 应确保模型训练过程符合 AI/ML 系统功能和性能要求的可重复性 应根据测试数据集对训练模型的性能进行评估，并将结果反馈至安全评估过程 对训练模型行为应进行基于需求的验证，建议采用基于需求的测试和鲁棒性测试方法
AI/ML 系统实现	应确定目标环境中的实现模型与在模型选择、训练和验证环境中的执行模型之间的差异 应识别实现过程中模型转换、优化和部署的步骤，并确认不会对模型行为和性能产生影响 传统软件开发的可追溯性和人工验证手段无法支持 AI/ML 系统实现过程中的意外特性检测，应建立新的替代方法
AI/ML 系统验证	应在捕获模型属性的基础上，使用特定的验证方法（如形式化方法）表明模型实现中的转换没有改变模型属性 传统软件结构的覆盖度量指标不再适用于 AI/ML 系统测试，应建立新的覆盖性测试指标，例如输入数据空间覆盖率、节点覆盖率等 应根据测试数据集对推理模型的性能进行评估，并将结果反馈至安全评估过程 应对推理模型行为进行基于需求的验证，建议采用基于需求的测试和鲁棒性测试方法 应对推理模型的鲁棒性进行分析，异常范围测试和对抗性测试是实现这一目标的必要手段
运行与维护	应监控 AI/ML 组件的输出，通过安全风险缓解措施对 AI/ML 系统进行钝化恢复 AI/ML 系统应具备数据记录能力，应包含足够的信息以检测 AI/ML 系统预期行为的偏差或重建发生故障前的系统运行情况 应将服役历史数据提供给 AI/ML 系统制造商，以便对训练、测试和验证数据集进行迭代更新

现有的行业标准和指南也可以作为智能航电系统适航符合性的补充方法，以增强申请人及局方对智能机载产品的信心。

3.2 对现有适航体系造成的影响

在研制智能航电系统时，现行系统及软硬件的研制保证符合性方法不足以处理 ML 学习过程的特殊性，需要通过 2.4 节 1）中的学习保证流程加以补充。同时，对于 AI/ML 的可解释性也有必要根据 2.4 节 2）中的认证活动补充新的符合性方法，但 1A 级的智能航电系统只需要参考人为因素认证指南就可以获得足够的符合性方法，更高级别的 AI 应用将在后续阶段进一步展开研究。

在航空运营方面，当前的监管框架允许引入 AI/ML 解决方案，欧盟委员会条例（EU No. 965/2012）[99]规定了与航空运营有关的技术要求和行政程序，并在 ORO. GEN. 200 条款中给出了需要建立、实施和维护的安全管理体系，允许运营商识别航空安全风险并采取缓解措施。

4 结束语

从权威航空安全机构、科研组织以及领军企业对智能航电系统认证研究的广泛关注和初步应用来看，此领域将会是航空工业在新一轮科技革命和产业变革中需重点攻克的难点问题之一。FAA 和 EASA 已经为航电系统引入 AI/ML 提供了开放性的解决方案，但仍需要在现有适航体系的基础上，进一步更新和补充标准化认证框架。

后续的工作需要牢牢把握未来航电系统的智能化趋势,探索现有技术应用落地的方式、形态和能力边界等。应积极借鉴和结合汽车及轨道领域的最新研究成果,采取多专业参与、多方法并举的手段支持智能航电系统全生命周期的符合性验证工作,从而提高智能航电系统的可信性和安全性。

参考文献

[1] 李文捷. 习近平关于人工智能重要论述研究[D]. 南昌: 江西财经大学, 2021.
LI Wenjie. Research on XI Jinping's important expositions on artificial intelligence[D]. Nanchang: Jiangxi University of Finance and Economics, 2021. (in Chinese)

[2] Office of the Chief Scientist. Autonomous horizons: the way forward[R]. Washington, DC.: United Sates Air Force, 2019.

[3] EASA. Artificial intelligence roadmap 1.0: A human-centric approach to AI in aviation[R]. Cologne: European Union Aviation Safety Agency, 2020.

[4] EASA. Concepts of design assurance for neural networks (CoDANN)[R]. Cologne: European Union Aviation Safety Agency, 2020.

[5] AVSI. Final report AFE 87-machine learning[R]. Texas: Aerospace Vehicle Systems Institute, 2020.

[6] DEEL. White paper machine learning in certified systems [R]. US: Dependable & Explainable Learning, 2021.

[7] SAE. Artificial intelligence in aeronautical systems: statement of concerns: AIR6988[S]. Pittsburgh: Society of Automotive Engineers, 2021.

[8] FORSBERG H, LINDÉN J, HJORTH J, et al. Challenges in using neural networks in safety-critical applications [C]// 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). US: IEEE, 2020: 1-7.

[9] 卢新来, 杜子亮, 许赟. 航空人工智能概念与应用发展综述[J]. 航空学报, 2021, 42(4): 251-264.
LU Xinlai, DU Ziliang, XU Yun. Review on basic concept and applications for artificial intelligence in aviation[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(4): 251-264. (in Chinese)

[10] 冯正霖. 以智慧塑造民航业的全新未来[N/OL]. 中国民航报, 2019-05-20(1).
FENG Zhenglin. Shape the new future of civil aviation industry with wisdom[N/OL]. CAAC NEWS, 2019-05-20(1). (in Chinese)

[11] McCARTHY J. What is artificial intelligence [J/OL]. (2007-11-12)[2022-05-30]. URL: http://jmc.stanford.edu/articles/whatisai.html.

[12] 王国庆. 航空电子系统综合化与综合技术[M]. 上海: 上海交通大学出版社, 2019.
WANG Guoqing. Principles and techniques of avionics system integration[M]. Shanghai: Shanghai Jiao Tong University Press, 2019. (in Chinese)

[13] 张菁, 何友, 邓瑛, 等. 战斗机智能航电系统[J]. 航空计算技术, 2018, 48(4): 125-129.
ZHANG Jing, HE You, DENG Ying, et al. Artificial intelligence based avionics system for fighters[J]. Aeronautical Computing Technique, 2018, 48(4): 125-129. (in Chinese)

[14] KLOS L, EDWARDS J, DAVIS J. Artificial intelligence-an implementation approach for advanced avionics[C]// 4th Computers in Aerospace Conference. Hartford, CT, USA: AIAA, 1983: 2401.

[15] STENERSON R O. The workstation: integtating AI into avionics engineering environment[J]. Computer, 1986, 19 (2): 88-91.

[16] 杨志刚, 张炯, 李博, 等. 民用飞机智能飞行技术综述[J]. 航空学报, 2021, 42(4): 265-274.
YANG Zhigang, ZHANG Jiong, LI Bo, et al. Reviews on intelligent technology of civil aircraft[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(4): 265-274. (in Chinese)

[17] ROBERT P. NASA aeronautics strategic implementation plan[R]. Washington, DC.: NASA Aeronautics Research Mission Directorate, 2020.

[18] YIN J, ZHU Z. Flight autonomy impact to the future avionics architecture[C]// 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC). USA: IEEE, 2018: 1-7.

[19] HARRISON L H, SAUNDERS P J, SARACENI P J. Artificial intelligence and expert systems for avionics [C]// AIAA/IEEE Digital Avionics Systems Conference. USA: IEEE, 1993: 167-172.

[20] RTCA. Software considerations in airborne systems and equipment certification: DO-178B[S]. Washington, DC.: Radio Technical Commission for Aeronautics, 1992.

[21] NASA. NASA armstrong fact sheet: intelligent flight control system [EB/OL]. (2014-03-01) [2022-05-30]. https://www.nasa.gov/centers/armstrong/news/FactSheets/FS-076-DFRC.html.

[22] DARPA. Alpha dogfight trials foreshadow future of human-machine symbiosis [EB/OL]. (2020-08-26) [2022-05-30]. https://www.darpa.mil/news-events/2020-08-26.

[23] Airbus. Airbus demonstrates first fully automatic vision-based take-off [EB/OL]. (2020-01-16) [2022-05-30]. https://www.airbus.com/en/newsroom/press-releases/2020-01-airbus-demonstrates-first-fully-automatic-vision-based-take-off.

[24] Boeing. Boeing loyal wingman uncrewed aircraft completes first flight [EB/OL]. (2021-03-01) [2022-05-30]. https://boeing.mediaroom.com/news-releases-statements?item=130834.

[25] DARPA. Safe, reliable, and uninhabited: first autonomous BLACK HAWK helicopter flight[EB/OL]. (2022-02-08) [2022-05-30]. https://www.lockheedmartin.com/en-us/news/features/2022/safe-reliable-and-uninhabited-first-autonomous-black-hawk-flight.html.

[26] 中国航空工业集团公司. 国产大型无人运输机TP500首飞成功[EB/OL]. (2022-06-24)[2022-05-30]. https://www.avic.com/c/2022-06-24/565096.shtml.
Aviation Industry Corporation of China, Ltd. Successful first flight of domestic large unmanned transport aircraft TP500[EB/OL]. (2022-06-24)[2022-05-30]. https://www.avic.com/c/2022-06-24/565096.shtml. (in Chinese)

[27] SCHWEIGER A, ANNIGHOEFER B, REICH M, et al. Classification for avionics capabilities enabled by artificial intelligence[C]// 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC). USA: IEEE, 2021: 1-10.

[28] 田莉蓉. 适航符合性方法发展综述[J]. 航空计算技术, 2019, 49(5): 121-124.
TIAN Lirong. Evolution of MOC for airworthiness[J]. Aeronautical Computing Technique, 2019, 49(5): 121-124. (in Chinese)

[29] WELD D S, BANSAL G. The challenge of crafting intelligible intelligence[J]. Communications of the ACM, 2019, 62(6): 70-79.

[30] KESKINBORA K H. Medical ethics considerations on artificial intelligence[J]. Journal of Clinical Neuroscience, 2019, 64: 277-282.

[31] CHALLEN R, DENNY J, PITT M, et al. Artificial intelligence, bias and clinical safety[J]. BMJ Quality & Safety, 2019, 28(3): 231-237.

[32] CHEN T, LIU J, XIANG Y, et al. Adversarial attack and defense in reinforcement learning from AI security view[J]. Cybersecurity, 2019, 2(1): 1-22.

[33] SAE. Guidelines for development of civil aircraft and systems: ARP4754A[S]. Society of Automotive Engineers, 2010.

[34] SAE. Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment: ARP4761[S]. US: Society of Automotive Engineers, 2004.

[35] RTCA. Software considerations in airborne systems and equipment certification: DO-178C[S]. Washington, DC.: Radio Technical Commission for Aeronautics, 2011.

[36] RTCA. Standards for processing aeronautical data: DO-200B[S]. Washington, DC.: Radio Technical Commission for Aeronautics, 2015.

[37] ISO. Road vehicles-safety of the intend-ed functionality: ISO/PAS 21448[S]. Geneva: International Organization for Standardization, 2019.

[38] ISO. Road vehicles-functional safety: ISO 26262[S]. Geneva: International Organization for Standardization, 2011.

[39] DeWALt M, McCORMICK G F. Technology independent assurance method[C]// 2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC). US: IEEE, 2014: 1-14.

[40] FAA. Understanding the overarching properties: first steps[R]. Washington, DC.: Federal Aviation Administration, 2018.

[41] NASA. An introduction to constructing and assessing overarching properties related arguments[R]. Washington, DC.: NASA Aeronautics Research Mission Directorate, 2022.

[42] EASA. Task force software (SW) & airborne electronic hardware (AEH) "Abstraction Layer"[EB/OL]. (2022-04-26)[2022-05-30]. https://www.youtube.com/watch?v=55n1hCFtigw.

[43] INSAURRALDE C C. Artificial intelligence engineering for aerospace applications[C]// 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). US: IEEE, 2020: 1-7.

[44] YIN J. A conceptual intelligent aircraft system[C]// 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC). US: IEEE, 2019: 1-7.

[45] 刘文炎, 沈楚云, 王祥丰, 等. 可信机器学习的公平性综述[J]. 软件学报, 2021, 32(5): 1404-1426.
LIU Wenyan, SHEN Chuyun, WANG Xiangfeng, et al. Survey on fairness in trustworthy machine learning[J]. Journal of Software, 2021, 32(5): 1404-1426. (in Chinese)

[46] BARTNECK C, LÜTGE C, WAGNER A, et al. An introduction to ethics in robotics and AI[M]. Berlin: Springer Nature, 2021.

[47] MORRIS A T, MADDALON J M, MINER P S. On the moral hazard of autonomy[C]// 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). US: IEEE, 2020: 1-9.

[48] 李升. 论人工智能伦理准则的细化与完善[D]. 杭州: 浙江大学, 2020.
LI Sheng. On the refinement and perfection of the ethical codes of artificial intelligence[D]. Hangzhou: Zhejiang University, 2020. (in Chinese)

[49] 孔祥维, 王子明, 王明征, 等. 人工智能使能系统的可信决策: 进展与挑战[J]. 管理工程学报, 2022, 36(6): 1-14.
KONG Xiangwei, WANG Ziming, WANG Mingzheng, et al. Trustworthy decision-making in artificial intelligence-enabled systems: progress and challenges[J]. Journal of Industrial Engineering, 2022, 36(6): 1-14. (in Chinese)

[50] FLI. Asilomar AI principle[R]. Boston: Future of Life Institute, 2017.

[51] IEEE. Ethically aligned design[R]. New York: Institute of Electrical and Electronics Engineers, 2017.

[52] AI HLEG. Draft ethics guidelines for trustworthy AI[R]. Brussels: The European Commission, High-Level Expert Group on Artificial Intelligence, 2019.

[53] G20. AI principles[R]. Osaka: G20 Ministerial Statement on Trade and Digital Economy, 2019.

[54] OSTP. Executive order on promoting the use of trust-worthy artificial intelligence in the federal government[R]. Washington, DC.: White House Office of Science and Technology Policy, 2020.

[55] 国家新一代人工智能治理专业委员会. 新一代人工智能伦理规范[R]. 北京: 中国科技部, 2021.

National New Generation Artificial Intelligence Governance Professional Committee. Ethical norms of the new generation of artificial intelligence[R]. Beijing: Ministry of Science of Technology, 2021. (in Chinese)

[56] EASA. First usable guidance for level 1 machine learning applications[R]. Cologne: European Union Aviation Safety Agency, 2021.

[57] AI HLEG. The assessment list on trustworthy artificial intelligence [R]. Cologne: European Union Aviation Safety Agency, 2021.

[58] 肖女娥, 阎芳, 王鹏. 基于安全论证的民机机载系统安全性评估[J]. 中国安全科学学报, 2019, 29(12): 72-77.
XIAO Nyu'e, YAN Fang, WANG Peng. Safety assessment of civil airborne system based on safety case[J]. China Safety Science Journal, 2019, 29(12): 72-77. (in Chinese)

[59] 修忠信. 民用飞机系统安全性设计与评估技术概论[M]. 2版. 上海: 上海交通大学出版社, 2018.
XIU Zhongxin. System safety design & assessment in civil aircraft[M]. 2nd ed. Shanghai: Shanghai Jiao Tong University Press, 2018. (in Chinese)

[60] AMODEI D, OLAH C, STEINHARDT J, et al. Concrete problems in AI safety [EB/OL]. [2022-05-30]. https: //www. researchgate. net/publication/304226143_Concrete_Problems_in_AI_Safety.

[61] FARIA J M. Non-determinism and failure modes in machine learning[C]// 2017 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). US: IEEE, 2017: 310-316.

[62] SMITH C, DENNEY E, PAI G. Hazard contribution modes of machine learning components[R]. Oak Ridge, TN, US: Oak Ridge National Lab, 2020.

[63] COFER D. Unintended behavior in learning-enabled systems: detecting the unknown unknowns[C]// 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC). US: IEEE, 2021: 1-7.

[64] ZHANG X, TAO J, TAN K, et al. Finding critical scenarios for automated driving systems: a systematic literature review[EB/OL]. [2022-05-30]. https: //www. xueshufan.com/publication/3207563209.

[65] FULLER J G. Run-time assurance: a rising technology [C]// 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). USA: IEEE, 2020: 1-9.

[66] ASTM. Standard practice for methods to safely bound flight behavior of unmanned aircraft systems containing complex functions: F3269-17[S]. West Conshohocken, PA: American Society of Testing Materials, 2017.

[67] SCHIERMAN J D, DEVORE M D, RICHARDS N D, et al. Runtime assurance for autonomous aerospace systems [J]. Journal of Guidance, Control, and Dynamics, 2020, 43 (12): 2205-2217.

[68] COFER D, AMUNDSON I, SATTIGERI R, et al. Run-time assurance for learning-based aircraft taxiing[C]// 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). US: IEEE, 2020: 1-9.

[69] LAZARUS C, LOPEZ J G, KOCHENDERFER M J. Runtime safety assurance using reinforcement learning[C]// 2020 AIAA/IEEE 39th Digital Airborne Systems Conference (DASC). US: IEEE, 2020: 1-9.

[70] WHEATMAN B, CHEN J, SOOKOOR T, et al. RADICS: runtime assurance of distributed intelligent control systems[C]// 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). US: IEEE, 2021: 182-187.

[71] FAA. Verification of adaptive systems: DOT/FAA/TC-16/4[R]. Washington, DC.: Federal Aviation Administration, 2016.

[72] ANSI. Safety standard for autonomous vehicles: UL 4600 [S]. New York: American National Standards Institute, 2020.

[73] Assuring Autonomy International Programme. Guidance on the assurance of machine learning in autonomous systems [R]. New York: AAIP, 2021.

[74] ASAADI E, BELAND S, CHEN A, et al. Assured integration of machine learning-based autonomy on aviation platforms[C]// 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). USA: IEEE, 2020: 1-10.

[75] TORENS C, DURAK U, DAUER J C. Guidelines and regulatory framework for machine learning in aviation[C]// AIAA Scitech 2022 Forum. [S.l.]: AIAA, 2022: 1132.

[76] GABREAU C, PESQUET-POPESCU B, H F KAAKAI, et al. Artificial intelligence for future skies: on-going standardization activities to build the next certification/approval framework for airborne and ground aeronautic products[EB/OL]. [2022-05-30] https: //dblp. org/rec/conf/ijcai/GabreauPKL21. html? view=bibtex.

[77] GUNNING D, STEFIK M, CHOI J, et al. XAI-explainable artificial intelligence[J]. Science Robotics, 2019, 4 (37): 7120-7128.

[78] SUTTHITHATIP S, PERINPANAYAGAM S, ASLAM S, et al. Explainable AI in aerospace for enhanced system performance[C]// 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC). US: IEEE, 2021: 1-7.

[79] ZHANG Y F, ZHANG Y, ZHANG M. SIGIR 2018 workshop on explainable recommendation and search (EARS 2018)[C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. [S. l.: s. n.], 2018: 1411-1413.

[80] LINARDATOS P, PAPASTEFANOPOULOS V, KOTSIANTIS S. Explainable AI: a review of machine learning interpretability methods[J]. Entropy, 2020, 23(1): 18-29.

[81] 雷霞, 罗雄麟. 深度学习可解释性研究综述[J]. 计算机应用, 2022, 42(11): 3588-3602.
LEI Xia, LUO Xionglin. Review on interpretability of deep learning[J]. Journal of Computer Applications, 2022, 42 (11): 3588-3602. (in Chinese)

[82] 刘潇, 刘书洋, 庄韫恺, 等. 强化学习可解释性基础问题探

索和方法综述[J]. 软件学报, 2023, 34(5): 2300-2316.
LIU Xiao, LIU Shuyang, ZHUANG Yunkai, et al. Explainable reinforcement learning: basic problems exploration and method survey[J]. Journal of Software, 2023, 34(5): 2300-2316. (in Chinese)

[83] LEROY X. Formal verification of a realistic compiler[J]. Communications of the ACM, 2009, 52(7): 107-115.

[84] RTCA. Formal methods supplement to DO-178C and DO-278A: DO-333[S]. Washington, DC.: Radio Tecnical Commission for Aeronautics, 2011.

[85] 刘友林, 郑巍, 谭莉娟, 等. 面向适航标准的机载软件测试验证工具综述[J]. 计算机工程与应用, 2021, 57(11): 1-10.
LIU Youlin, ZHENG Wei, TAN Lijuan, et al. Summary of airborne software testing and verification tools for airworthiness standards[J]. Computer Engineering and Applications, 2021, 57(11): 1-10. (in Chinese)

[86] CCMB. Security assurance requirements (Version 2.3): common criteria for information technology security evaluation (Part 3)[S]. US: Common Criteria Maintenance Board, 2005.

[87] FULTON N, PLATZER A. Safe reinforcement learning via formal methods: toward safe control through proof and learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2018: 1-10.

[88] DREOSSI T, FREMONT D J, GHOSH S, et al. Verif AI: a toolkit for the formal design and analysis of artificial intelligence-based systems[C]// International Conference on Computer Aided Verification. Berlin: Springer, 2019: 432-442.

[89] URBAN C, MINÉ A. A review of formal methods applied to machine learning [EB/OL]. (2021-03-21)[2022-05-30]. https://www.researchgate.net/publication/350673876_A_Review_of_Formal_Methods_applied_to_Machine_Learning.

[90] 钱航. 深度神经网络测试覆盖率与对抗样本间的相关性研究[D]. 南京: 南京邮电大学, 2020.
QIAN Hang. Correlation between coverage and adversarial examples for deep neural networks[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020. (in Chinese)

[91] SUN Y, HUANG X, KROENING D, et al. Testing deep neural networks [EB/OL]. (2018-07-04)[2022-05-30]. https://dl.acm.org/doi/pdf/10.1145/3426430.3434071.

[92] PEI K, CAO Y, YANG J, et al. Deepxplore: automated whitebox testing of deep learning systems[J]. Mobile Computing and Communications Review, 2018, 22(3): 36-38.

[93] SUN Y, HUANG X, KROENING D, et al. Deepconcolic: testing and debugging deep neural networks[C]// 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion). USA: IEEE, 2019: 111-114.

[94] NASA. Verification of adaptive systems: NASA/CR-2015-218702[R]. Washington, DC.: NASA Aeronautics Research Mission Directorate, 2015.

[95] NUÑEZ J, GRANGER G, COLOMER A, et al. Cognitive Computing potential for cockpit operations[R]. USA: Harvis Project, 2019.

[96] REGLI C, ANNIGHOEFER B. Towards certification of adaptive flight automation systems: a performance-based approach to establish trust[C]// 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC). [S.l.]: IEEE, 2022: 1-10.

[97] 董磊, 刘嘉琛, 赵长啸, 等. 显示触控系统人为因素适航符合性验证技术[J]. 中国安全科学学报, 2021, 31(2): 63-68.
DONG Lei, LIU Jiachen, ZHAO Changxiao, et al. Airworthiness compliance verification technology of human factors in touch interactive display system[J]. China Safety Science Journal, 2021, 31(2): 63-68. (in Chinese)

[98] 中国民用航空局. 航空器型号合格审定程序: AP-21AA-2011-03-R4[S]. 北京: 中国民用航空局适航审定司, 2011.
CAAC. Type certification procedures for aircraft: AP-21AA-2011-03-R4[S]. Beijing: Airworthiness Certification Department of Civil Aviation Administration of China, 2011. (in Chinese)

[99] EC. Air OPS regulation: EU No. 965/2012[S]. Brussels: The European Commission, 2012.

作者简介:

董　磊(1983—),男,博士,副研究员。主要研究方向:民用飞机航电系统适航审定技术。

刘嘉琛(1996—),男,博士研究生。主要研究方向:智能航电系统、航空电子综合系统。

陈　曦(1987—),男,博士,助理研究员。主要研究方向:模式识别与图像处理。

肖女娥(1984—),女,硕士,副研究员。主要研究方向:民用飞机系统安全性评估技术。

梁博尧(1998—),男,硕士研究生。主要研究方向:人工智能鲁棒性、系统运行时保证。

张元珊(1999—),女,硕士研究生。主要研究方向:人工智能可解释性。

(编辑:马文静)