军用人工智能测试与评估实施指南

2024 年 12 月

拉维·潘瓦尔　李强　约翰·沙纳汉

引言

这套《军用人工智能测试与评估实施指南》是宏理国际与新美国安全中心历时两年，通过线上与线下相结合的方式，汇聚中美两国及国际专家智慧，共同研讨的成果。在伦敦创誓基金会和纽约卡内基公司的慷慨资助下，以及在丹麦皇家国防学院的联合赞助下，相关研讨会得以在欧洲、亚洲和北美多地以线上和线下的形式举行，对该《实施指南》进行反复推敲。

这一系列研讨旨在确定，对于含有重要人工智能组件的武器及相关军事系统的测试与评估，中美及国际专家能否就某些原则和实施指南达成共识，以确保这些系统更加安全、可靠且负责任地运行。参加研讨的人员包括美国、中国以及来自欧洲、亚洲及其他地区的学者及前官员，涵盖军事、外交、情报、计算机科学、企业和法律等多个领域。

目标

人工智能技术与军用系统的集成在全球范围内势头强劲。尽管在大多数情况下这种集成进展缓慢，但显而易见，人工智能赋能的军用系统将在未来几年内在全球范围内加速部署。基于来自私营企业的经验，我们预测，各国军队在大规模集成人工智能时将面临诸多挑战。

为推动负责任地使用人工智能赋能的军用系统，对人工智能及其

他关键组件（关涉能否安全、合法且合乎伦理地使用人工智能赋能的军用系统）的测试与评估已迫在眉睫。人工智能的测试与评估包括验证和确认过程（首字母缩写为 TEVV）。国际社会就人工智能测试与评估的原则和最佳实践达成共识至关重要，这既可以促进遵守国际人道法，也可以降低因人工智能赋能的军用系统发生不可预见或始料未及的故障所带来的全球风险。

人工智能给传统的武器系统开发、测试及军事部署方法带来了重大挑战。虽然传统军用软硬件系统在测试与评估流程方面，尤其是在系统工程原理方面，与人工智能赋能系统存在某些共性，但人工智能必然会使测试与评估方法发生重大改变。人工智能赋能的军用系统的独有特征，要求采用不同于现有测试与评估措施的专门方法，以确保进行综合性评估与确认。军方可能采用混合式体系，包括传统的非人工智能系统、新型的非人工智能系统、加装人工智能的传统系统以及从一开始就设计为人工智能赋能的新型武器系统，这种未来走向进一步加剧了人工智能测试与评估的复杂性。这些系统可能同时运行，必然要考虑多种人工智能赋能系统在进行跨武器系统、指挥与控制体系、虚拟网络的交互时产生的级联效应[1]和潜在的涌现行为[2]。

各国需要采取明确措施，在技术安全和研发操作方面不断提升人工智能技术的安全性、可靠性和可控性，增强对人工智能技术安全性的评估和管理能力，并确保人类始终对使用武力负责。各国必须在人工智能研发活动中加强自我约束，基于对作战环境和武器特性的全面考量，在武器的全生命周期内实施必要的人机交互。各国必须坚守人

[1] 级联效应是指由于某个行为对系统产生的影响，从而引发一系列不可避免的意外事件，例如在生态系统内，某一种重要物种的死亡，可能触发其它物种的灭绝。

[2] 涌现行为用于描述在复杂系统中，由于个体之间的相互作用而产生的整体性质或行为，例如自然界中当每只鸟都根据相邻鸟的动作调整自己的飞行路径时，在一群鸟中就可以观察到像同步飞行这样的涌现行为。

类是最终责任主体这一原则，为人工智能建立问责机制，并为操作人员提供必要的培训。

这些实施指南的目标之一是帮助所有国家在人工智能赋能的系统的全生命周期中实施足够水准的测试与评估，促进能力投送的有效性、适当性、可靠性、可预测性、可持续性、安全性、可信赖性及韧性，且符合国际法。

各国在开发、测试和部署人工智能赋能的军用系统方面积累更多经验以前，应偏向于采取更加审慎的方法，在任何军用人工智能技术投入战场前对其进行额外测试。各国应遵循预防原则，避免引入最终效果不明或存在争议的新产品或新流程。

人工智能测试与评估：特性

人工智能赋能的系统，无论民用还是军用，都具有会显著影响测试与评估流程的独特之处。这些特性不仅会影响对人工智能模型本身的评估，还会影响集成了这些模型的更广泛的系统。人工智能的独特属性要求调整传统的测试与评估方法，以保证进行全面的评估与确认。人工智能测试与评估的关键特征包括：

- 持续测试与监督：人工智能赋能的系统从初始设计到长期维护，需要在其全生命周期内持续进行评估。这种持续性的方法要求设计者、开发者、测试者及最终用户之间进行更紧密的协作。就人工智能赋能的系统而言，“已完成测试”这一概念已经过时。人工智能模型的动态性叠加其学习和适应能力，需要一个持续的评估框架，以保证其持续性能和可靠性。

- 部署后的迭代与不可预测性：人工智能系统持续学习和部署后蜕变的可能性，叠加其内生的不透明性，产生了操作上的不可预测性。因此就需要评估或然行为或统计上可预测的行为（非确定行为），并建立流程以识别并减少意外故障模式。
- 动态学习与快速更新：人工智能和机器学习/深度学习系统拥有无需额外编写代码即可从数据中直接学习的独一无二的能力，从而实现系统的经常性更新，并且在在线学习的情况下，实现实时适配。这一能力要求为已部署的人工智能赋能的系统建立可适应持续集成/持续投送（或部署）的测试与评估流程。此外，它还突显出在战场部署的人工智能系统中嵌入稳健的测试设备的重要性，以监督和评估这些系统由于时间的推移而不断进化的性能。
- 敏捷治理：人工智能赋能的系统要求从传统的线性和串行软件开发方法转向更为灵活和反应迅速的方法。敏捷开发的方法论及具备适应性的测试与评估原则对于适配人工智能系统的动态性至关重要。这种迭代方法使得基于持续测试结果和不断变化的需求而不断改进和优化人工智能系统成为可能。
- 对抗性韧性：配合独立的“红队”演练，测试与评估过程有必要加入专门测试，以评估专门针对人工智能数据集和模型的敌对攻击所造成的影响和风险。这种方法能够提高人工智能赋能的系统在敌对环境中的鲁棒性[3]和韧性，这对于此类系统的军事应用是特别关键的考虑因素。

[3] 鲁棒性是指人工智能系统在不改变其初始稳定配置的情况下抵抗变化和外部干扰的能力。

- 以数据为中心和计算需求[4]：人工智能系统的基础在于数据以及处理这些数据所必需的基础设施。数据的这种中心地位带来了独特的挑战，包括数据集偏斜[5]、篡改或残缺的可能性，这可能严重影响系统的性能和可靠性。此外，人工智能系统通常需要高性能的计算基础设施来处理复杂算法和海量数据。人工智能由数据驱动的本质也导致其具有“黑箱”特征，其内部决策过程即使对开发者而言可能也不透明。这一特征给实现人工智能系统的可解释性和可审计性带来重大挑战，它们恰恰是影响军事应用的关键因素，而对于军事应用，透明度和可问责是首要考虑。

军用人工智能测试与评估的独特性

人工智能驱动的系统的有些弱点和风险与军事环境尤为相关，军用人工智能的测试与评估流程必须予以处理。这些弱点和风险包括以下方面：

- 作战环境中数据匮乏：在军事行动特有的苛刻且动态变化的环境中，获取高质量且有代表性的数据——这些数据对于人工智能/机器学习驱动的系统不可或缺——面临重大挑战。这种匮乏带来多种风险，包括难以生成准确的训练数据集、难以创造具有作战上代表性的测试环境，以及利用人工智能技术和仿真生成训练数据集时难以避免偏差。
- 对多元战场的适应性：环境发生变化时，军用系统经常频繁地重新部署。这就要求有一个全面的框架，在每次重新部署前重

[4] 计算需求是指在计算机科学或信息技术领域中，完成特定任务所需的计算资源或处理能力。
[5] 数据偏斜在统计学中是指数据分布不对称的情况。

新训练和重新评估系统性能，以确保其在多元作战环境中的最佳功能性和可靠性。

- 异常值的不利影响：人工智能赋能的武器系统中的异常值源自其不透明性和脆弱性等内在特征，在作战环境中可能产生非常不利的影响，因为这事关生死。测试与评估流程需引入严格基准，以尽可能减少异常值或极限情况[6]的产生。
- 决策周期的加速：人工智能与自主性的结合显著加快了军事决策过程中“观察—判断—决策—行动”（OODA）这一循环。这种加速给保持有效的人类控制带来挑战，也增加了自动化偏见[7]的风险。在设计和评估系统时需要考虑这些因素，以促进系统在受控范围内运行。
- 在线学习的风险：有能力在线学习（即在运行过程中自适应）的系统带来独特的挑战。这些系统可能超出其初始测试与评估的参数运行，导致意想不到和不可预测的结果。因此，特别是对具有致命能力的系统，有必要建立保障措施和监督机制以减轻这些风险。
- 伦理和法律考量：为遵守国际人道法原则，包括自主武器系统在内的一切军用系统必须进行严格评估。此外，视情况而定，测试与评估流程必须包含 1977 年《第一附加议定书》[8]第 36 条要求的法律审查。

军用人工智能测试与评估实施指南：目标和结构

[6] 极限情况是指仅在极端（最大或最小）操作参数或其他异常操作条件下发生的问题或情况。详细解释参见本实施指南最后一部分：术语集。

[7] 自动化偏见是指人们过度依赖自动化系统或技术，忽视手动检查或人工判断的现象。

[8] 全称是《1949 年 8 月 12 日日内瓦四公约关于保护国际性武装冲突受难者的附加议定书》，1977 年 6 月 8 日通过，1978 年 12 月 7 日生效。

人工智能测试与评估实施指南的首要目标是促进关于这一关键议题的全球对话，在人工智能测试与评估最佳实践及相关信任建立措施（CBMs）方面推动形成国际共识。这种方法旨在实现平衡，一方面是国际合作的需求，另一方面是认识到军队对其人工智能测试与评估流程和程序中的敏感事项保有自由裁量权。实施指南的总体目标是营造一个协作环境，在尊重国家安全考量的同时，加强全球人工智能安全和治理。

下列实施指南旨在应对人工智能赋能的军用系统内生的独特挑战急考量。这些实施指南分为三个关键部分，反映了在军事环境中实施人工智能的生命周期：第一，设计和开发阶段；第二，人工智能赋能的武器系统的部署阶段；第三，涉及人工智能透明度、可信赖性以及在国家间建立信任的考量因素。

设计和开发阶段的实施指南关涉影响测试与评估流程的一系列问题，例如数据完整性、数据溯源以及数据收集困难，因此需要使用合成数据[9]；尤其与关键决策支持系统（DSS）和自主武器系统相关的透明度与可解释性；为实现复杂环境中的最佳性能而在开发人员与军事人员之间进行持续整合；频繁在多种作战环境中重新部署而触发的持续测试与评估；特别是对高风险系统对极限情况的韧性处理；应对数据匮乏的使用场景时建模与仿真的重要性；确保军地领导层进行知情监督的必要性；特别是对自主武器系统和关键决策支持系统而言进行人机融合的必要性；大型语言模型的影响；遵守国际法，特别是国际人道法的要求。

部署阶段的实施指南旨在应对复杂网络化环境中出现的一系列

[9] 合成数据是指人工生成的信息，而不是由真实世界的事件产生的。

问题，例如在苛刻且动态变化的作战场景中可能发生的意想不到的和灾难性的错误，特别是与战略指挥和控制系统相关的问题；特别是对包含在线学习功能的系统来说进行持续监督与矫正的必要性；以及系统全生命周期中进行数据收集与管理的重要性。

最后一部分涉及标准、突发事件和建立信任措施，包括下列实施指南：强调借鉴现有军事和民用领域专业标准的必要性；联合国及其他多边组织应发挥的作用；共享数据及突发事件处理流程；关于风险减缓技术的培训项目；共享测试与评估的标准、政策和准则；就制定测试与评估方面有法律或政治约束力的文书提出建议；遵守预防原则的必要性。

军用人工智能测试与评估术语表

这些实施指南以简洁的语言编写，因此使用了某些与人工智能和军事相关的术语，这些术语可能需要进一步解释。尽管最好是对这些术语进行严谨定义，但就本实施指南而言，清晰地阐明这些术语的含义就足够了。本实施指南的附录中附有一个术语表，阐释了这些术语在本实施指南中的具体含义。

军用人工智能测试与评估：22 项实施指南

I	**设计和开发**	
	A	人工智能的测试与评估必须包括在尽可能接近系统作战部署期间所预期的条件下所获得的测试与评估数据，理想的情况下应基于真实世界的数据。由于缺乏针对军事目标的数据收集机会或者敌方采取行动阻止数据收集而导致数据受限，作战数据可能需要通过合成数据进

		行补充。应采取措施确保训练、测试和确认所用数据的监管、来源及质量。
	B	测试方法的选择应基于人工智能系统算法和组件的可解释性和可理解性程度来确定，能够通过具有明确性能指标和评估标准的稳健的测试与评估流程进行评估。这对于关键军事决策支持系统和致命性自主武器系统来说尤为重要。
	C	人工智能系统的设计与开发过程应从一开始应包含测试与评估的要求。人工智能的测试与评估必须考虑到传统软件与人工智能之间的差异，以及算法和运行测试。由于作战环境更为复杂、不确定且不太为人所知，军方应确保测试与评估计划覆盖人工智能全生命周期，包括持续支持和维护。人工智能的测试与评估要求开发人员、测试人员和军事人员之间的持续整合，以获得可预测且可靠的测试结果。
	D	人工智能赋能的军用系统的测试与评估应被视为一个持续的过程。测试与评估应在系统部署前、部署后以及每次重新部署到不同作战环境之前进行，并持续到系统退役为止。尽管针对定期的人工智能模型更新、重新作战部署或作战条件改变时必需的更新所进行的测试与评估，在深度和广度上可能不同于人工智能赋能的系统在首次部署前进行的测试与评估，但设计计划必须考虑到持续进行测试与评估（作为持续集成/持续部署或投送

		的一部分）的必要性。
	E	人工智能的测试与评估应包括在真实世界条件下的测试，充分顾及系统的韧性和鲁棒性（稳健性），包括在苛刻、不确定且动态变化的作战环境中适当处理极限情况和极限条件、识别与纠正错误以及特别是在高风险的致命性自主武器系统中设置回滚模式/故障保护模式。
	F	测试与评估计划应详细说明将在何时以及在多大程度上使用建模与仿真对人工智能系统进行测试，以及如何确认此类测试的有效性，特别是设计用在敌方可能拒绝或欺骗已部署人工智能模型的环境中的系统。对每个系统而言，有必要明确在何种保真度[10]水平上此类系统的行为才需要通过建模与仿真加以测试，包括使用“数字孪生”技术[11]。
	G	测试与评估计划应包括如何将测试结果和系统性能传递给所有利益攸关方。测试与评估流程应支持负责系统部署的适当级别的军事指挥官以及文职领导层进行知情监督。
	H	人机融合和/或人机协作应被视为测试与评估设计中不可或缺的一个组成部分。军用人工智能系统的测试与评估中一个关键的问题是确定人类操作员在作战条件下

[10] 保真度通常用来描述信息或信号的准确性和真实性。

[11] 数字孪生指的是物理资产的计算机化伴侣，可用于多种用途。数字孪生利用安装在物理对象上的传感器数据来表示其近乎实时的状态、工作状况或位置。数字孪生的一个例子是使用 3D 建模为物理对象创建数字伴侣。它可以用来查看实际物理对象的状态，提供了一种将物理对象投影到数字世界的方法。传感器将收集数据并通过物联网反馈给 3D 建模软件。这项技术属于增强现实类别。数字孪生通常与物理对象不仅在形状上，而且在定位、姿态、状态和运动上都完全相同。

		监督人工智能系统的能力，以及在多大程度上可以确保与武器系统或关键决策支持系统相关的“观察—判断—决策——行动”循环处于人类控制之下，确保打击目标的决定始终由人类指挥官和操作员作出。
	I	鉴于大型语言模型（LLM）/生成式人工智能可能被集成到军用系统中，在测试与评估过程中应特别注意产生于大型语言模型的相关因素，包括基于人类反馈的强化学习（RLHF）[12]、微调[13]、检索增强生成（RAG）[14]以及大型语言模型已知的局限性。包含生成式人工智能的军用系统必须进行专门评估，以确保在关键决策支持系统和武器系统中的错误模式被限制在可忽略的范围内，并在测试期间进行详细说明和评估。
	J	测试与评估要求应旨在确保能够评估系统是否遵守了相关法律要求，包括对一个系统遵守国际人道法及其他相关国际法文书的能力进行法律审查的义务。
II	**部署**	
	K	测试与评估不仅应评估人工智能赋能的军用武器或决策支持系统的组件和子系统的性能，还应评估人工智能系统的整体性能及这些组件、子系统与任何外部或现有平台的集成度。这尤其应包括在网络化环境中新系统及

[12] 基于人类反馈的强化学习是一种结合人类反馈和强化学习技术的算法，旨在通过人类的评价和偏好优化智能体的行为，使其更符合人类期望。

[13] 大模型微调是指利用特定领域的数据集对已预训练的大模型进行进一步训练的过程，它旨在优化模型在特定任务上的性能，使模型能够好地适应和完成特定领域的任务。

[14] 检索增强生成模型结合了语言模型和信息检索技术。具体来说，当模型需要生成文本或者回答问题时，它会先从一个庞大的文档集合中检索出相关的信息，然后利用这些检索到的信息来指导文本的生成，从而提高预测的质量和准确性。

		更新与先前组件、平台或系统的集成或结合。
	L	测试与评估计划和系统文档应明确一个严格的流程，在军用人工智能系统部署到新的作战环境之前或作战环境发生重大变化时，可据此识别、分析危险并加以补救。计划应建立已部署模型更新的条件和流程，包括战术单位的角色与责任。
	M	特别是对战略指挥和控制系统而言，测试与评估计划应识别在运行过程中可能发生的高风险的灾难性错误，以及如何预防、检测和补救这些错误。计划还应明确如何处理意想不到的错误。测试与评估计划应部署“红队”来挑战假设，并在其他方面攻击系统设计和部署的底层逻辑，以识别意想不到的错误并在部署前加以补救。
	N	测试与评估计划应明确如何评估已部署的军用人工智能系统是否继续符合其性能指标。如果系统不符合这些指标，应采取适当的矫正措施。应特别关注在部署期间能够继续学习的系统。虽然在线学习具有持续改进性能的优势，但也带来了在未经测试的状态下运行系统的风险，也增加了恶意行为者发起网络攻击的风险。测试与评估计划必须确保采取特别关注措施，将特别是高风险的致命性武器系统中的在线学习可能带来的负面后果降至最低。
	O	在军用人工智能系统的全生命周期中，包括作战部署期间，数据的收集、管理、评估和使用至关重要。必须关

		注数据和元数据在收集过程中的管理方式，确保数据适合于设计目的，并在部署期间随着收集进行更新。测试与评估计划应特别考虑到数据对人工智能的功能性以及涉及人工智能安全性的重要功能的影响，目标是改进系统并强化我们对系统可靠性的理解。
III	**标准、突发事件和建立信任**	
	P	各国政府在努力为人工智能赋能的军用系统制定并强化人工智能测试与评估的实施指南时，应将其工作与民用标准、工具和文档相协调，并借鉴来自军事和民用领域的专业标准，包括 ISO/IEC、IEEE 及其他标准制定组织制定的标准。
	Q	各国政府应考虑联合国或专家级多边组织在标准制定或就影响测试与评估的技术和/或治理问题对军用人工智能进行监管方面可发挥的适当作用。
	R	各国政府应开展对话，互相学习并分享各自在军用人工智能系统开发和部署中的经验教训，包括测试与评估的标准、测试与评估在减少风险和/或“突发事件”中的作用以及其他透明度和信任建立措施。各国政府还应考虑设立培训项目以强化测试与评估在军用人工智能系统生命周期中的重要性，并在这些培训项目中引入技术、军事、政治和法律领域的专家。
	S	作为系统生命周期中持续测试与评估的一部分，各国政府和国际机构应考虑为演习或作战部署期间使用军用

		人工智能引发的“后果严重的突发事件”制定调查和补救标准。这些标准可包括：（1）导致调查的“突发事件”的类型和严重程度以及调查是否应在国家管辖权范围内或超出国家管辖权范围进行；（2）调查程序，包括访问涉嫌引起突发事件的系统的保密信息或敏感信息以及这些信息的后续发布或保护；（3）与突发事件相关的缓解和补救程序；（4）考虑到保护保密信息或敏感信息的需求，就突发事件及其调查以及缓解或补救程序而言可能适当的透明度水平或信息披露程度。
	T	为推动透明度、相互理解和形成一致的最佳实践，各国应针对人工智能赋能的军用系统公开发布其测试与评估的流程和方法的相关内容。在国家安全考量允许的情况下，为建立信任与信心，各国应考虑公开发布以下内容：人工智能赋能的军用系统的设计标准、政策和作战指南；用于确定关键安全组件测试严谨性和风险缓解措施的标准；用于确定人工智能系统潜在事故严重程度的标准；法律审查标准和程序以及将人工智能风险纳入系统风险整体考量的流程。
	U	各国在采用军用人工智能测试与评估的实施指南时，应考虑哪些指南可以成为具有法律或政治约束力的文书的基础，以及适当的执行机制可能是什么。
	V	各国在开发、测试和部署人工智能赋能的军用系统方面积累更多的经验之前，应以预防原则为指导，即引入最终效果不明或存在争议的新产品或新流程时，应审慎行

		事、暂停决策并进行审查。

军用人工智能测试与评估实施指南：术语表

测试与评估、验证和确认（*Test & Evaluation, Verification &Validation*）

测试。在测试与评估的背景下，测试是指执行特定的测试用例，以收集有关系统行为、功能和性能的数据。测试贯穿全生命周期，从个别单元到整个系统，旨在验证各组件是否按说明书运行，并验证系统是否达到预期目的。

评估。评估包括分析和解释测试结果，以确定系统是否满足要求并为部署做好准备。评估涵盖验证和确认，简要说明如下：

- 验证（我们是否正确地构建了系统？）：通过各个阶段的测试确保系统符合设计说明。
- 确认（我们是否构建了正确的系统？）：确保系统符合用户需求并在其预期环境中有效运行。

用户在系统说明中指定的条款并非每一项都能够以精确的方式进行测试。例如，用户可能要求系统具有简单或流线型的美感，或提供用户友好的界面。此类无形需求的评估需要依赖人工判断，无法通过测试来验证。因此，符合此类需求可能不属于“验证”和/或“确认”的范畴，但仍然属于更广义的“评估”范畴。

军事决策支持系统（*Military Decision Support Systems*）

在本实施指南中，“军事决策支持系统”一词是指所有未明确归类为武器系统一部分的军用系统。换言之，在本实施指南中，全部人工智能赋能的军用系统被划分为两类：武器系统和决策支持系统。

本实施指南还使用了“关键决策支持系统”一词，这是决策支持系统的一个子集。其中“关键”这个修饰语主要取其字典意义。宽泛而言，导致军事力量部署（如兵力的部署、攻击的选项等）但不直接引发武器使用的决策支持系统可归类为关键决策支持系统。非关键军事决策支持系统则涵盖后勤、维护、医疗、人力资源管理等领域的应用程序以及军事领域的其他此类系统。

部署/作战部署、作战环境/条件/背景（*Deployment/Operational Deployment, Operational Environment/Conditions/Context*）

“部署”和“作战部署”在很大程度上是同义语，指人工智能赋能的军用系统被移交给作战组织的军事用户。这表示系统从开发和测试阶段转入作战使用阶段（“作战使用”和“作战部署/运用”是含义相同的概念）。这标明作战组织已经从开发和测试该系统的组织接受了该系统。

“作战环境”是指系统被用于作战的任何场景，而非开发、测试、演习或实验的场景。它既可以指物理领域，如空中、陆地、海洋和太空，也可以指虚拟领域，如网络和电磁空间。当作战组织为了使用而接受某一系统时，该系统便处于作战环境中。提及作战环境时，不区分和平时期、危机时期或冲突时期。因此，它也常被称为真实世界环境。

“条件”指影响人工智能赋能的军用系统的开发、测试和作战使用的所有情况的整体。这一术语与“作战环境”一词密切相关。它包括天气、昼夜、地形、地理位置、是否正在进行作战行动、是否使用被动或主动反制措施以拒绝或破坏该系统等各种因素。

“背景”与“作战环境”和“条件”是同义语。它也常被称为“作战背景”或“作战场景”。

极限情况/极限条件（*Edge Cases/ Boundary Conditions*）

极限情况是指系统在正常运行条件的极点或边界上发生的场景。极限情况对测试来说至关重要，因为它们可以揭示在正常使用情况下不会出现的漏洞，而解决这些问题可以提高系统的鲁棒性（稳健性）和可靠性。极限情况通常是无法预料或极为罕见的处境，可以揭示人工智能模型中隐藏的缺陷或局限性。例如，如果训练和测试图像数据中没有包含熊猫的图像，而在系统部署后实际出现了熊猫，可能会导致不可预料的结果。如果在作战环境中预计会出现熊猫，即使概率很低，也应在该模型部署之前用熊猫图像对其进行训练。

另一方面，极限条件是极限情况的一个子集，专指容许输入范围边缘或其附近的点。极限测试的重点是验证一个系统在这些极限值下的正确行为，例如略低于、等同于或略高于输入或输出范围的最大值和最小值。这些值更具可预测性，对测试人工智能模型的鲁棒性（稳健性）十分重要。例如，在开发一个人工智能模型时，应基于预期的运行环境，用尽可能最小和最大的图像尺寸来测试图像分类模型。

极限情况与极限条件的区别如下：

- 范围：极限情况涵盖任何异常值或极端场景，而极限条件专指有效输入范围的极限值。
- 测试重点：极限条件测试关注容许范围附近的精确输入值，而极限情况测试可能包括更广泛的无法预料或极为罕见的处境。

人工智能赋能的系统也应以类似方式进行测试。然而，在武器系统和关键决策支持系统中，极限情况需要保持在最低限度（即需要规定非常严格的性能指标），因为这些情况可能导致灾难性或严重的后果，即便它们很少发生。

人类控制（*Human Control*）

人类控制是指人类在系统生命周期中的作用，从设计和开发到部署及部署后的支持与维护。

对于已部署的系统，人类控制包括以下方面：（1）人类实时监督和理解作战环境和系统状态信息的能力；（2）系统运行期间在不同节点人类可能进行干预的程度和方法，包括但不限于任务计划和武器使用；（3）人类预测和理解系统行为及其行为潜在后果的能力；（4）将系统行为的责任分配给具体的人类操作员或指挥官；（5）落实保障措施和故障保护机制以便在必要时允许人类接管或关闭系统。

网络化环境（*Networked Environment*）

网络化环境包括武器系统、网络及连接传感器、系统、武器和人员的底层信息技术架构。它还包括便利这些要素作为一个整体运行的应用程序。

人工智能赋能的军用系统/军用人工智能系统（*AI-Enabled Military Systems/ Military AI Systems*）

在本实施指南中，“人工智能赋能的军用系统”和“军用人工智能系统”是同义语。