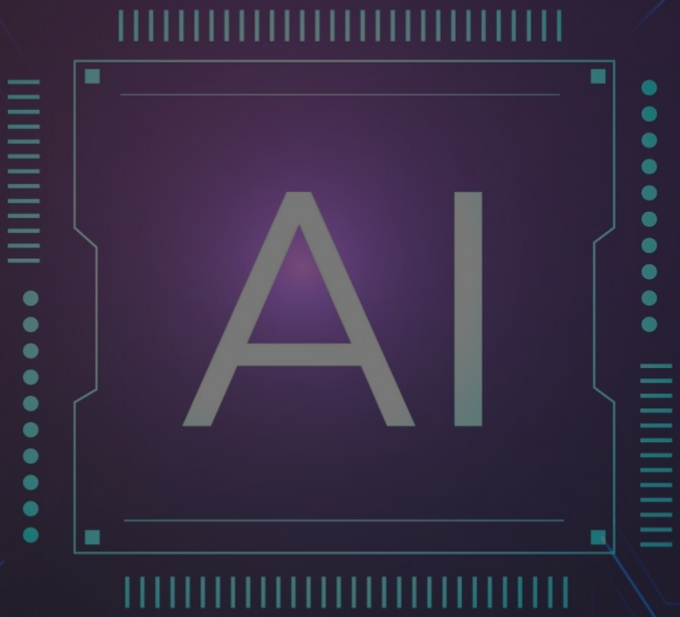
Technology Innovation Institute




# 利用生成式 AI 增强自治系统的安全性、弹性和安全性

tii. ae

# 目录


边栏 ： GenAI 型号 4
*GenAI 用例个别应用程序*
自我意识异常检测自主性适应 9
性预测性维护故障管理 9
10
11
12
12
13
14

*舰队* 16
群智能 17
群协调 19
沟通弹性 21
共识 22

*Human* 23
任务规划和执行 24
人机合作 25
元学习 26
数据标记和合成 27
*网络安全* 28
恶意软件检测 29
入侵检测: 31

集成威胁情报 33 政策管理 33 威胁模拟 ： 34 软件供应链可见性 34

*挑战* 34
Cost 35
计算 36
适应 36
道德 / 监管 36
对齐 36
隐私 36
准确性 37
Size 37
新的安全问题 38
*Conclusion* 38
*词汇表 41*

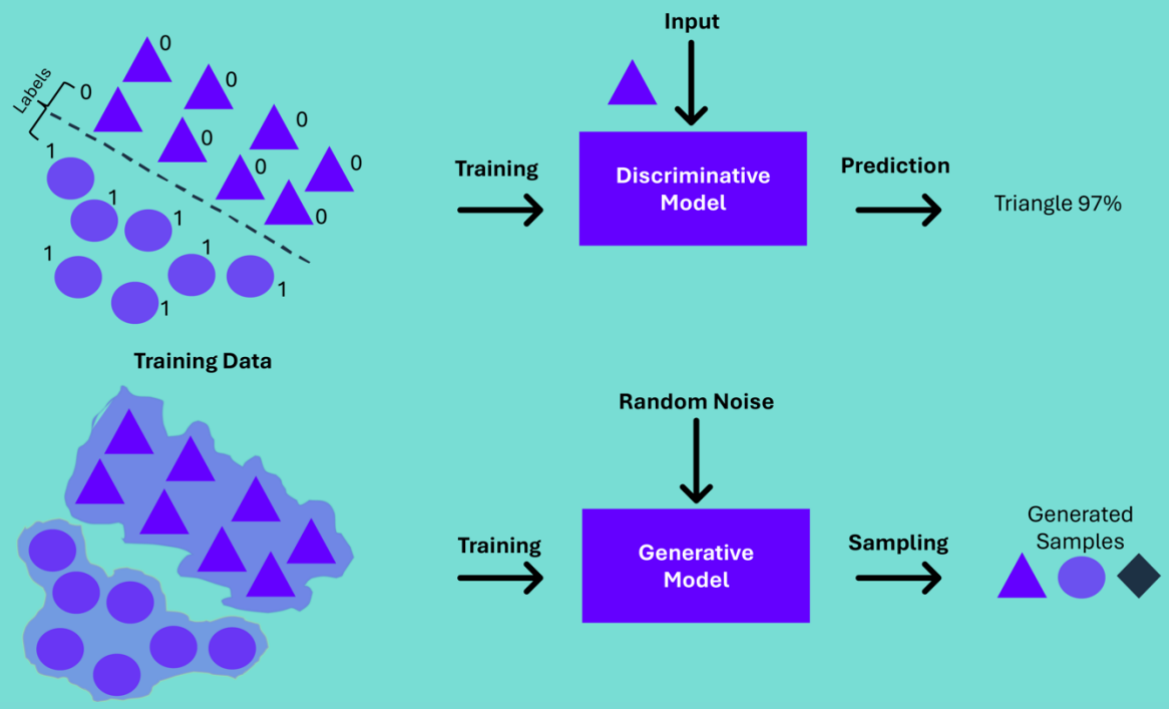

TII
Technology
Innovation
Institute

快速发展的自主系统和边缘机器人技术为制造业、交通运输、医疗保健和探索等领域带来了前所未有的机遇。不断增加的复杂性和连接性也带来了确保安全、韧性和可靠性的新挑战。随着边缘机器人深入融入我们的日常生活和关键基础设施，我们必须开发创新的方法来提高这些系统的可信度和可靠性，达到新的水平。

这份白皮书探讨了生成式人工智能（GenAI）在增强自主系统和边缘机器人安全、韧性和安全性方面的变革潜力。我们可以利用这些前沿技术来应对边缘机器人独有的分布式和动态挑战，以解锁新的智能、适应性和稳健性水平。

生成型AI模型通过分析数据集中的模式来生成新内容。它们推导出特征性的概率分布，并将这些分布应用于创建与原始“真实”数据集一致的新数据模式。

早期的判别型AI模型通过应用条件概率来预测未见过的数据的结果。该方法具有广泛的适用性，适用于包括分类和回归在内的多种问题。它们擅长界定区分不同类或类别之间的决策边界。



生成技术的队伍日益壮大，包括基于变换器的大语言模型（LLMs）、生成对抗网络（GANs）、变分自编码器（VAEs）、生成流模型（GFM）和生成扩散模型（GDM）等。这些技术为人工智能研究开辟了新的激动人心的领域，应用范围涉及无人机群、入侵检测、物理通信安全、语义通信和移动网络等领域。 [1]

技术创新研究所的安全系统研究中心(TII - SSRC ， 阿联酋阿布扎比 ， https: / / www. tii. ae / secure - systems )正致力于将生成性人工智能（GenAI）应用于其工作，将零信任架构（最初为信息安全开发）扩展到 cyber-物理系统的所有方面。因此，SSRC考虑如何利用GenAI确保无人机的安全、韧性和可靠性。

swarm群体, swarm群体的群组, 自主地面和海洋车辆, 指挥系统, 以及人类/无人机交互——特别是在通用人工智能（GenAI）在性能上超越传统AI/机器学习（AI/ML）方法的领域。

**示例包括 ：**

- **个别应用 ：** 健康监测、状态估计、预测维护、异常检测、自我修复、导航和紧急着陆。
- **车队应用:** 群协调、群智能、集体决策。
- **人类 / 无人机互动 ：** 通信弹性、任务规划、人机交互。
- **网络安全和弹性 ：** 入侵检测 ， 恶意软件分类 ， 威胁模拟。

这篇论文将专注于无人机，因为SSRC在此领域开展了大量工作。我们从无人机中获得的经验教训可以广泛应用于自主系统和 cyber-物理系统，包括汽车、机器人、嵌入式系统以及智能城市。同样地，在这些领域获得的经验教训也可以整合到SSRC的研究中。

这种方法使组织能够脱离依赖员工携带多部手机的物理设备管理方式。

[1] 徐M. 等人，“释放边缘-云生成式AI在移动网络中的力量：AIGC服务综述。”arXiv，2023年10月31日。doi: 10.48550/arXiv.2303.16129

# 边栏 ： GenAI 型号

一种针对无人机安全、安全性和弹性的特定生成AI建模技术的调查，包括它们的应用优势与局限性。

公众对生成式AI模型的兴趣激增，主要受到了诸如ChatGPT等高度公开的新服务的驱动，这些服务利用特别训练的大语言模型（LLMs）生成听起来具有权威性的文本回应。

大型语言模型是基于庞大文本数据集训练的AI系统。它们利用深度学习技术，尤其是被称为变换器的结构，来“理解”并生成类似人类的文本，基于它们学到的模式。这些模型分析训练数据中的关系和上下文，并使用多种技术构建数据的简化表示，以在原始数据元素之间建立关联和相关性。这些模型使它们能够生成模仿人类写作风格并涵盖广泛主题的回应。视觉AI，如DALL-E和Stable Diffusion，从文本和图像提示中合成新的图像。

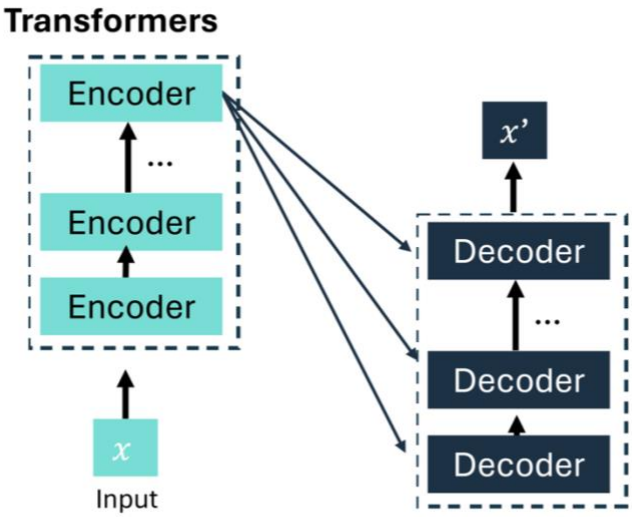
大型语言模型（LLMs）在创建内容、代码、翻译、摘要、合成数据以及结构化来自文本、文档、图像、音频和其他提示数据的非结构化数据方面被广泛使用且用途广泛。

变压器模型及其构建的服务——如OpenAI的ChatGPT、Google的Gemini以及Anthropic的Claude——因其能够生成看似有条理的人类提示响应而引起了广泛关注。这些模型，以及其他领域的大型语言模型（LLMs）和小型语言模型（SMLs），也显示出支持分析、研究和开发以提高无人机的安全性、安全性和韧性方面的潜力。

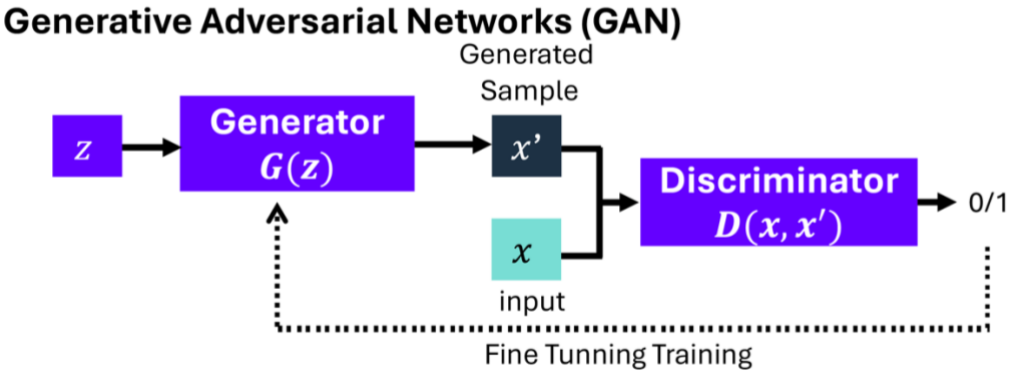
然而，与这些非常明显的AI发展并行的，是在过去近十年里新类别的生成AI模型取得了显著进展，这些模型能够自动化和加速构建表示。虽然生成应用吸引了最多的关注，但这些新模型也在推动数据分析和与我们周围世界互动方面的进步。其他生成AI模型，如生成对抗网络（GANs）、变分自编码器（VAEs）、生成扩散模型（GDMs）和规范流模型（NFMs），尽管相对鲜为人知，但它们在无人机安全、安全性和韧性方面可以做出重大贡献。

**变压器型号:** 引入于2017年用于英法文本互译，Transformer模型在捕捉非结构化数据内的长距离依赖性和相关性方面表现出色。[2] 变压器利用一种新颖的“注意力机制”来学习单词之间的联系，从而帮助自动创建词嵌入。以往的技术需要使用单独的模型将原始文本转换为向量表示。变压器可以通过其分层结构和方式构建复杂的表示并学习复杂的联系，使研究人员能够处理大量未标记的文本并开发具有数十亿参数的大规模语言模型。后续创新支持了文档摘要、在大数据集上生成问题/答案关联、代码生成、深入分析、入侵检测、恶意软件检测以及跨机器人臂传输控制指令系统说明。这种方法的关键优势在于从复杂数据集中提炼上下文。挑战包括幻觉、更长的训练时间、更慢的推理构建速度、更高的计算需求以及更大的模型规模，相较于其他技术而言。

[2] A. Vaswani 等人，“注意就是你所需要的。”arXiv，2023 年 8 月 1 日。 doi: 10.48550 / arXiv .1706. 03762 .

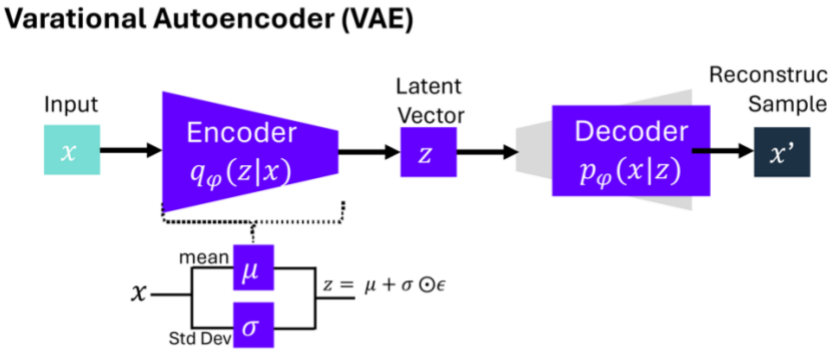



**生成对抗网络 (GAN):** 这些是在 2014 年开发的 ， 用于创建逼真的合成数字 ， 面孔和动物图像 。 [3] GANs使两个神经网络相互竞争：一个受到生成更真实内容的奖励，而另一个受到检测虚假内容的奖励。 他的竞争对手通过增强生成器创建逼真输出的能力，使其能够欺骗鉴别器，从而提高了竞争力。生成对抗网络（GANs）广泛应用于内容生成。自最初的版本设计用于处理图像以来，研究人员现在正寻找创造性的方法将代码或网络日志等数据转化为适合GAN处理的图像。GANs适用于生成可用于改进自主系统和网络安全算法的逼真合成数据集。然而，它们也面临着模式崩溃或灾难性遗忘等问题。

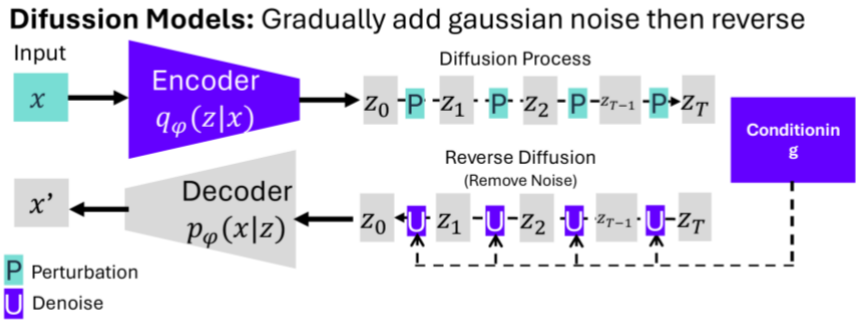



[3] I. J. Goodfellow 等人 ，“生成对抗网络 ”。 arXiv ， 2014 年 6 月 10 日。 doi: 10.48550 / arXiv.1406.2661 。

**变分自动编码器 (VAE):** VAE 于 2014 年引入 ， 以改善从连续变化的数据流中得出的推论。 [4] 该技术有助于找到高效的数据表示方法，并可用于数据压缩或异常检测及威胁识别。VAEs的训练过程涉及教导一组编码器和解码器将原始数据转换为具有不同概率分布的中间潜在空间。VAEs可以在异常检测、设计更好的编码方案、数据增强和图像生成等应用中独立使用。此外，它们常用于为其他算法（包括GANs）预结构化数据，以提高这些算法的效果。




**生成扩散模型 (GDF ):** 非平衡热力学建模为基础的学习、抽样、推断和评估改善得以通过GDFs在2015年得到实现。 [5]
该技术向样本（如图像）添加噪声，然后自动化去噪过程以揭示数据的基本结构。细微的变化可以生成有效的新型训练数据集。生成对抗网络（GDFs）广泛应用于图像生成，并能提高各种无人机应用场景中信号分类的准确性。然而，该技术需要更高的采样时间，并要求更复杂的架构，相较于GANs和VAEs更为复杂。

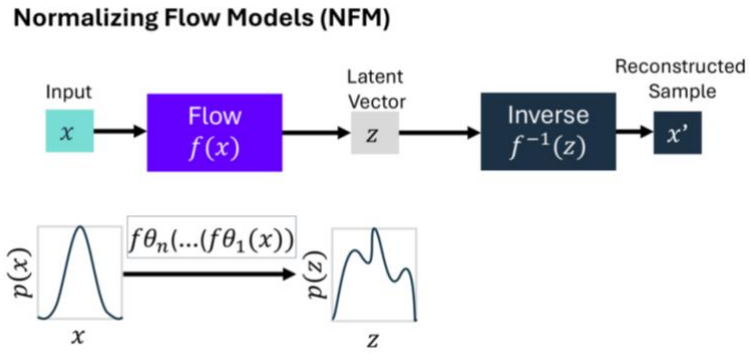



**标准化流量模型 (NFM):** 这些是由研究人员引入的 ， 目的是使复杂的数据更易于使用。 [6] 这些模型将易于理解的分布（如正态分布曲线）逐步转换。每一步都是可逆的，这意味着如果需要，我们总可以回到起点。这个过程称为“流”，从简单的初始状态逐步过渡到最终状态。

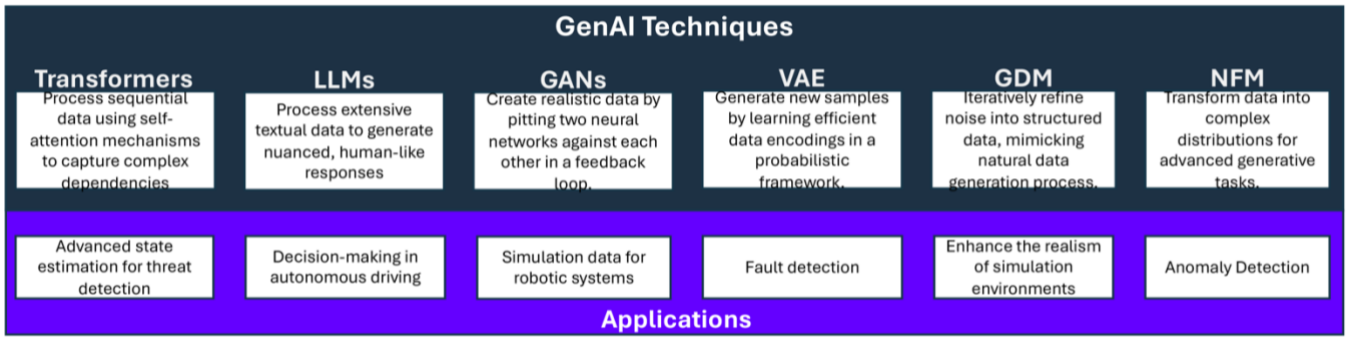
---

[4] D. P. Kingma 和 M. Welling ，“自动编码变分贝叶斯 ”。 arXiv ， 2022 年 12 月 10 日。 doi ： 10.48550 / arXiv.1312.6114 。
[5] 杨 ， 玲 ， 等。“扩散模型 ： 方法和应用的全面综述 ”。 *ACM 计算调查* 56.4 (2023): 1-39 . https: / / dl. acm. org / doi / 10.1145 / 3626235
[6] Kobyzev, Ivan, Simon JD Prince, and Marcus A. Brubaker. "Normalizing flows: An introduction and review of current methods." *IEEE 模式分析和机器智能事务* 43.11(2020 年) ： 3964 - 3979 。 https: / / ieeexplore. ieee. org / abstract / document / 9089305 / authors

类似于复杂的目标数据集。通过这种方式，可以更有效地研究和利用数据。生成对抗网络（GANs）和变分自编码器（VAEs）等技术已被用于生成手写数字、图像等。 newer 用法案例包括增强分类和编码方案。训练过程会创建一个模型，将数据集的概率分布转化为更复杂且完全可逆的分布。然而，生成对抗网络（GANs）和变分自编码器（VAEs）等技术相比，生成对抗网络（NFMs）可能需要更高的计算和训练时间。




以下图表总结了主要的通用人工智能（GenAI）技术及其在自主系统零信任领域的应用。




# GenAI 用例

无人机技术的普及带来了跨越多个领域（个体、机群、人类控制和网络安全）的挑战。其快速增长和复杂性要求不断进行创新以增强其可信度和可靠性。

以下应用——无论是源自无人机（UAV）和无人飞行器研究，还是从其他领域引入——对无人机及其他自主系统的未来具有重要意义。值得注意的是，许多项目尚处于早期阶段，提及这些项目是为了展示随着技术的发展，通用人工智能工具可能实现的功能概貌。

GenAI展现出在提高个体自主系统（如无人机、自动驾驶汽车、机器人和嵌入式系统）的安全性、韧性和安全性方面增强零信任框架的巨大潜力。正在调查的应用场景包括提升自我意识、异常检测、自动驾驶、预测性维护、故障管理、自愈以及安全着陆。


# 自我意识

**机会:** 高效地转换嘈杂、模糊且不一致的数据以理解无人机当前状态——例如，在尝试检测障碍物的同时补偿运动模糊。

无人机健康的基础在于准确捕捉并理解其当前状态——包括当前硬件的状态、应用情况、物理位置以及安全状况。在现实世界中，这可能会变得复杂，因为视频馈送可能会出现运动模糊，GPS数据会抖动，惯性导航数据会失去校准，噪声或缺失数据会降低内部监控数据的质量。

状态估计对于自主导航和决策制定至关重要，原始数据流必须与位置、速度和姿态准确关联。[7] 生成 AI 可以帮助填补缺失数据并融合来自多个来源的数据，以改善状态估计。[8]

创新的生成对抗网络（GANs）、变分自编码器（VAEs）以及传统的机器学习算法如LSTM可以在这些方面发挥作用，通过故障检测、预测性维护、故障管理以及安全着陆协议来保障车辆安全。例如，创新的GAN方法可以填补缺失的数据，并使数据流融合变得更加容易，从而创建更准确的状态评估。[9] 帮助将内部日志数据与声学分析相关联，[10] 并识别潜在的机械问题。[11] 研究人员还开发了使用条件生成对抗网络（CGANs）为单个无人机和无人机群生成估计状态变量的技术。[12]

---

[7] T. D. Barfoot，机器人技术的国家评估。剑桥大学出版社，2017 年。

[8] 刘光远，阮文辉，杜鸿阳，丁泰皇，杜斯提·尼雅托，朱宽，康佳文，熊泽惠，阿巴斯·贾马利波尔，金东银. “生成式人工智能在无人驾驶车辆群中的应用：挑战、应用与机遇.” arXiv，2024年2月28日. https: / / doi. org / 10.48550 / arXiv.2402.18062 .

[9] He Y, Chai S, Xu Z. 基于生成对抗网络的一种状态估计新方法[J]. 在2019年IEEE系统、人和 cybernetics国际会议（SMC）上发表，2019年，页码：2248-2253. [在线]. 可访问: https://ieeexplore.ieee.org/document/8914585

[10] 王宇、维诺格罗夫A．，“使用历史状态ensemble提升基于声发射信号的无监督早期故障检测性能的卷积GAN表现”，《应用科学》，13(5) (2023), p. 3136, doi: 10.3390/APP13053136

[11] 宋正、阿卜杜勒·法哈特和古普塔昌德拉，“生成对抗网络在故障预测中的应用”。arXiv，2019年10月4日。访问日期：2024年3月15日。[在线]. 可用: http://arxiv.org/abs/1910.02034

[12] .A. 何, 罗晨, 田旭, 和 曾渭, "一种双分支Siamese网络用于实时目标跟踪," 在IEEE计算机视觉与模式识别会议 proceedings, 2018, 页码: 4834-4843. https://ieeexplore.ieee.org/document/8578606

# 异常检测

**机会 ： 改进对无人机传感器数据的分析 ， 以识别异常情况。**
更准确、多维度的系统状态记录也有助于识别与无人机健康相关的问题。例如，变分自编码器（VAEs）可以提高故障检测和隔离的准确性。它们还可以识别各种系统中压力的预警信号，从而优先安排预测性维护计划。通常，机器学习分类算法是基于多个类别进行训练的（如标记为“故障”和“正常”的数据）。然而，在公共数据集和实际应用中，稀有类别的数据（如“故障”数据）往往稀缺。在这种情况下，生成对抗网络（GANs）可以在合成这些稀有类别方面发挥重要作用——生成看起来像是“故障”条件的数据。此外，研究人员正在探索大型语言模型（LLMs）如何更好地帮助在新环境中进行操作和上下文理解。 [13]

例如，DriveLLM将大型语言模型（LLM）与传统自主导航算法相结合，以支持在应对边缘情况时更好的推理和决策。 [14] 研究人员发现该方法可以在意外情况下提高主动决策制定能力。另一个应用TypeFly通过自然语言接口增强了人类与无人机之间的沟通。 [15] .

如此大的语言模型可能会表现出表面偏见、不准确性和幻觉等问题，这些都需要额外的安全措施来加以防范。同样，微软研究部门讨论了他们将ChatGPT与机器人技术整合的研究进展，通过自然语言使机器人控制更加直观。他们已经使ChatGPT能够理解并在物理环境中执行任务，从而简化了人机交互过程，无需复杂的编程知识。ChatGPT团队为了解决机器人任务开发了设计原则（涉及特殊的提示结构和高层API），并通过用户友好的命令和反馈展示了ChatGPT如何处理操作无人机和机器人手臂等任务。开发人员强调了安全性和模拟测试的重要性。 *之前* 真实世界的应用。 [16]

# 适应性

**机会:** 改进自主系统软件的跨不同硬件品牌、型号和配置的翻译能力。

自主系统控制器必须针对特定型号和配置进行训练。这在升级个别组件或采用新模型时可能会带来挑战。RTX 是一种机器人控制的大语言模型（LLM），能够将控制策略翻译成管理不同机械臂的方式，而无需为最新硬件重新调整控制算法。在某些测试中，利用其他控制器的经验所生成的控制策略甚至超越了专门为某个机械臂量身定制的最佳控制策略。 [17]

---

[13] 王磊等。“基于大型语言模型的自主代理调查”。 *计算机科学前沿* 18.6(2024) ： 1 - 26 。 https: / / link. springer. com / article / 10.1007 / s11704 - 024 - 40231 - 1
[14] 崔宇等，“DriveLLM：大型语言模型助力全自动驾驶之路”，《IEEE智能车辆 Transactions》，第9卷，第1期，页码：1450-1464，2024年1月，DOI: 10.1109/TIV.2023.3327715。
[15] 陈国军, 于潇 Jing, 仲林. "TypeFly: 使用大型语言模型的无人机飞行." arXiv预印本 arXiv:2312.14950(2023). https://arxiv.org/pdf/2312.14950
[16] Vemprala ， Sai 等人。“机器人技术的 Chatgpt ： 设计原理和模型能力”。 *arXiv 预印本 arXiv ： 2306.17582* (2023). https: / / arxiv. org / abs / 2306.17582
[17] O. X. .- E. Collaboration 等人 ， “开放 X - 实施例 ： 机器人学习数据集和 RT - X 模型”。 arXiv ， 2023 年 12 月 17 日。 doi: 10.48550 / arXiv.2310.08864 。

早期的语言模型（LLMs），如GPT-3.5，是在从互联网抓取的大规模文本数据上进行训练的。这些模型缺乏真实世界的经验，无法反映各种机器人和其他自主系统在决策及其执行过程中的实际情况。关于机器人功能的研究探讨了如何限制每个机器人模型仅执行与其能力相匹配且可行的操作。[18] 这为基于对操作或程序更全面了解的LLM开发提供了一个框架。同时，grounding功能将这种高层次的知识转化为特定机器人模型在特定目标环境中的执行。

# 预测性维护

**机会:** 预测无人机组件的未决故障 ， 以优化维护、维修和零件更换计划。

妥善记录和分析后，无人机的传感器和操作数据能够在故障发生之前揭示潜在的机械问题。预测算法使得维护和维修团队能够制定定期的工作计划、优先安排维护任务，并提前规划备件库存。即使是在常规服务期间，重大的维修和更换工作也可以提前进行。这样可以显著降低昂贵的故障概率，更糟糕的是，大幅减少灾难性故障的发生。部件可以在需要时及时更换，服务寿命则根据安装部件的质量、服务时间和运行状况来计算——从而大幅削减按照固定时间表更换完全正常部件的成本。

传统机器学习算法往往在预测性维护中扮演核心角色。例如，剩余使用寿命（RUL）和健康指标等度量标准可以识别电机故障。但由GAN和其他生成式人工智能算法生成的合成数据可以提升这些算法的性能。

多种机器学习技术，包括生成式人工智能算法，可以结合起来以改善故障诊断和预测性维护工作流程。[19] 例如，GAN技术已被应用于机械设备的声音信号以识别和预测其他方法未发现的故障。[20] GANs 也用于生成合成监控数据集，以帮助训练其他机器学习算法，从而提高故障预测精度并优化维护调度。[21] GAN-FP，遗传对抗网络故障预测，专门用于生成、平衡和标注训练数据以提升其他机器学习算法的性能。[22] .

# 故障管理

**机会:** 识别故障 ， 进行动态调整 ， 并在需要时实现安全着陆。

---

[18] M. Ahn 等人，"尽力而为，而非言出必行：将语言 grounding 在机器人可利用性上。" arXiv，2022年8月16日。访问日期：2024年3月21日。[在线]. 可用: http://arxiv.org/abs/2204.01691

[19] Z. Mian 等人，“基于集成学习的故障诊断综述”，《工程应用的人工智能》，第127卷，第107357页，2024年1月，DOI: 10.1016/j.engappai.2023.107357。20Y. Wang, A. Vinogradov，“使用历史状态集成提高卷积GAN性能以实现声发射信号驱动的无监督早期故障检测”，《应用科学》，第13卷（5），2023年，第3136页，DOI: 10.3390/app13053136。21Q. Fu, H. Wang, J. Zhao, 和 X. Yan，“基于生成对抗网络的航空发动机维护预测方法”，《2019 IEEE第五届计算机与通信国际会议（ICCC）》，2019年12月，第225-229页，DOI: 10.1109/ICCC47050.2019.9064184。22S. Zheng, A. Farahat, 和 C. Gupta，“生成对抗网络用于故障预测”。arXiv，2019年10月4日。访问日期：2024年3月15日。[在线]. 可用: http://arxiv.org/abs/1910.02034

生成人工智能模型可以将数据转换为其他机器学习系统，以提高自主系统中的故障检测能力。例如，变分自编码器（VAEs）可以帮助将运营数据压缩为更高效的表示形式，用于长短期记忆网络（LSTM），这是一种循环神经网络。[23] 时空变换网络能够捕捉不同时间尺度下的趋势和维度，以提高电池故障诊断和失效预测的准确性，从而增强无人机的预测性维护。例如，BERTery系统可以检测出早期ML技术难以察觉的细微变化，这些变化可以在电池失效前24小时左右预示即将发生的电池故障。[24]

GANs已被用于生成训练样本并构建推理网络以提高其他机器学习算法的故障预测能力，处理航空发动机监测数据。[25]
研究人员将 VAE 和 LSTM 结合起来 ， 以支持车辆传感器的连续学习
生成合成数据以覆盖更广泛的故障场景。通过将其他机器学习算法训练在这种类型的合成数据上，萨杜等人实现了90%的故障检测准确率和99%的分类准确率。

对计算能力的需求和相对较低的执行速度是这些算法在低成本硬件上运行时的主要关注点。一种解决方案是将计算移植到FPGA上，因为FPGA比GPU更具能效性。这就是Sadhu等人通过将VAE-LSTM故障检测算法移植到FPGA上实现了40倍的速度提升（同时功耗降低一半）的方式。[26]
VAE 还可用于训练识别的模型 *normal* operation. Using this technique, Dhakl

et al. 达到了95.6%的准确率，用于检测指示未包含在训练数据集中的故障和异常的偏差。[27]

当无人机或其通信网络出现故障时，无人机必须安全着陆以避免二次损害。为了最大限度地减少这种风险，蒙特卡洛算法被用于计算各种着陆区的“目标安全水平”（可接受的风险水平）。[28]
这些技术可以在故障迫使无人机系统选择合适的着陆区域时，与变换器结合以做出具有上下文意识的决策。

在未来，也可能利用如变换器等生成人工智能技术使系统能够在硬件故障、软件错误或网络中断的情况下自我修复。例如，Khlaisamniang等人提出了一种使用生成人工智能检测异常、生成代码、调试代码并创建计算机系统报告的框架。[29] 尽管仍处于早期阶段 ， 但这项工作为其他自治系统的未来研究提供了方向。

---

[23] Han P, Ellefsen A L, Li G, Holmeset F T, Zhang H. 基于LSTM的变分自动编码器在海洋组件故障检测中的应用[J]. IEEE传感器杂志, 2021, 21(19): 21903-21912. DOI: 10.1109/JSEN.2021.3105226.

[24] Zhao J., Feng X., Wang J., Lian Y., Ouyang M., and Burke A. F., “使用时空变换网络进行电动车辆电池故障诊断与失效预测,” 应用能源, 第352卷, 页码：121949, 2023.

[25] Fu Q., Wang H., Zhao J., and Yan X., “使用生成对抗网络的航空发动机维护预测方法,” 在2019年IEEE第五届计算机与通信国际会议（ICCC）上发表，2019年12月，页码：225-229。doi: 10.1109/ICCC47050.2019.9064184.

[26] V. 塞杜、K. 安朱和 D. 波米利，“基于 FPGA 的机载深度学习无人机故障原因检测与分类”，《IEEE 机器人学交易》，第 39 卷，第 4 期，页码：3319–3331，2023 年 8 月，doi: 10.1109/TRO.2023.3269380.

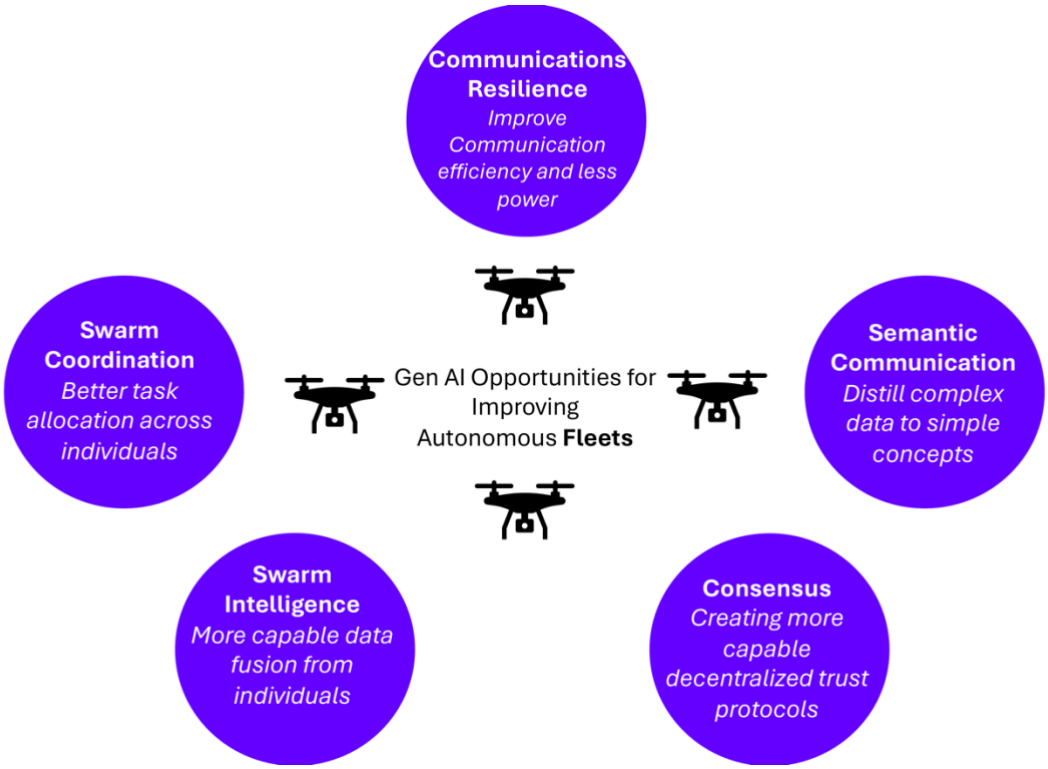
[27] R. 达卡尔、C. 波斯马、P. 查德哈里和L. N. 科恩德尔，“使用自动编码器进行无人机故障和异常检测”，载《IEEE/AIAA第42届数字航空系统会议论文集》。IEEE, 2023, pp. 1-8.

[28] Tong L., Gan X., Yu L., and Zhang H., “融合空域中无人航空系统安全目标水平评估,” 于2022年国际人工智能与计算机应用会议（ICAICA）发表，2022年6月，页码：375-379。DOI: 10.1109/ICAICA54878.2022.9844489.

[29] P. Khlaisamniang, P. Khomduean, K. Saetan, and S. Wonglapsuwan, Generative AI for Self - Healing Systems. 2023, p. 6. doi: 10.1109 / iSAI - NLP60301.2023. 10354608.

# 舰队

生成人工智能可以帮助提高群体智能、群体协调性，并增强舰队或自主事物群组底层通信网络的稳健性、安全性和效率。在此背景下，零信任安全、安全性和韧性变得尤为重要——保护无人机舰队，提高共享感知的完整性，促进更好的协调，并减少被 compromized 的无人机对整个舰队的影响。




# 群智能

**机会:** 提高 swarm 中多个个体传感器数据的可信融合。在过去十年中，研究人员发表了成千上万篇关于通过融合众多无人机协同工作产生的数据来合成统一态势视图的论文。假设一群无人机正在对一场大规模灾难（如洪水或火灾）进行调查。理想情况下，可信赖的 swarm 智能能够将蜂群中每个成员的信息结合起来，形成一个全面的情况描述。

生成式知识支持的变压器（GKSTs），例如，可以融合目标对象不同视角的图像，从而产生更具意义的移动车辆图像。[30]
进一步增强这种多视角方法可能有助于改善对 swarm 成员从不同角度收集的图像的解释。

重要的是要认识到有许多方式可以表示现实世界，而不同的方法可能适用于不同的目的。例如，相似性分类算法可以帮助对图像中对象特征的外观进行分类。相比之下，语义分类算法可以将这些对象标记为特定类别或类别的成员。SA-SIAM（一种两阶段的语义-外观Siamese神经网络）可以在不同类别之间共享信息，

[30] Yu X, Liao W, Qu C, Bao Q, Xu Z. 基于多智能体生成对抗模仿学习的无人机协同搜索[J]. 2022国际机器学习、云计算与智能采矿会议（MLCCIM），2022年8月，第441-446页. doi: 10.1109/MLCCIM55934.2022.00081.

单独的神经网络，一个在语义信息上训练，另一个在外观数据上训练。[31] 语义侧采用注意力机制来帮助根据目标和附加上下文信息解释数据。尽管这并不是一个完整的生成式AI实现，但它展示了如何在更具体的应用场景中应用带有针对性的注意力机制的变换器，这种方式可能比全面的大型语言模型实现更为高效。

条件生成对抗网络（CGANs）——基于特定条件的生成对抗网络变种——已被用于运动预测。这些预测考虑了每个对象相对于无人机的相对运动及其不断变化的方向。[32] 这些功能与共享权重的暹罗网络协同工作，在该网络中，神经权重在一对互补的神经网络之间共享。尽管最初的研究集中在单个无人机上，但这项工作表明了一条合成机群贡献所得三维视图的未来路径。

在2016年，研究人员探索了如何通过“社会聚合”层帮助自主代理模拟附近人群的交互，并使用单独的LSTM网络来预测每个人的动作。[33] 在这种情况下，研究人员关注的是自动驾驶车辆如何更好地规划其路径以通过一群独立移动的人类群体。未来的研究可以探索社会聚合如何扩展，以改进模型，使无人机能够理解附近机群无人机、旁观者无人机乃至外部敌对无人机的当前位置和预测未来位置。

基于文本到图像的扩散模型也被用于生成不同场景下无人机的逼真图像，通过这些模型，无人机检测算法的精度提高了12%。研究人员将标准化流模型与变压器结合使用，以改进状态估计任务——这是自主系统评估自身状态（相对于其他活跃代理，如另一架无人机或车辆）能力的关键。[34]

# 群协调

**机会 ： 改善群体中个体之间的任务分配。**

生成式AI技术也被探索用于提高无人机机队之间的任务分配效率。与依赖固定算法的传统方法相比，生成式AI算法在动态且复杂的多无人机环境中提供了更好的控制优势。[35]

2018年的调查研究了用于改进无人机舰队通信和协调的一系列算法，增强了群体完成目标的能力。这些算法通常按层次结构组织，以支持不同组织层级的自主性，并减少对人工监督的需求。主要挑战在于提高舰队和个体层面的控制与估计能力。[36]

---

[31] A. 何, C. 罗, X. 天, 和 W. 曾, "一种双支Siamese网络用于实时目标跟踪," 在 IEEE 计算机视觉与模式识别会议论文集, 2018, 页码: 4834-4843.

[32] 余浩, 李刚, 孙乐, 钟波, 姚虹, 和 黄 Quest, “基于条件GAN的多目标跟踪中个体和全局运动融合方法在无人机视频中的应用”, 图像识别快报, 第131卷, 第219-226页, 2020年.

[33] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, 和 S. Savarese，“Social LSTM：拥挤空间中的人体轨迹预测”[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，美国内华达州拉斯维加斯：IEEE，2016年6月，第961-971页。DOI: 10.1109/CVPR.2016.110。

[34] H. Delecki, L. A. Kruse, M. R. Schlichting, 和 M. J. Kochenderfer, “深度归一化流在状态估计中的应用。”arXiv, 2023年6月27日。访问日期：2024年3月27日。[在线]. 可用: http://arxiv.org/abs/2306.15605

[35] G. M. 斯卡利斯、H.-S. 淑欣和 A. 托苏斯，“多-agent 系统中任务分配技术综述”，载《2021 国际无人驾驶航空系统会议（ICUAS）》。IEEE，2021，pp. 488-497。

[36] 宋智俊、A·阿鲁纳·帕那贾帕、P·戴mes、沈思与V·卡umar，“飞行 swarm 机器人综述”，《IEEE机器人学交易》，第34卷，第4期，页码：837-855，2018年8月，doi: 10.1109/TRO.2018.2857475.

TI研究所的研究人员在开发多代理大语言模型（LLM）方面取得了先驱性进展，这些模型可以在蜂群中的每个设备上运行，但能够协作规划和解决问题。在这种情况下，大语言模型可以帮助发展更好的意图驱动网络，将高层次的意图（如减少网络能耗）转化为一系列低层次的任务。未来的研究必须开发特定于电信的大语言模型、减轻幻觉现象、开发自复制代理以改进设备上的资源管理、开发新的共识机制，并开发将多模态数据编码到更适合资源受限设备的概念空间的技术。[37]

MAPPO-GAIL（一种用于无人机群的GAN）结合了多智能体 proximal 策略优化（MAPPPO）与生成对抗imitation学习（GAIL），以帮助在机群中细化样本轨迹并改进策略模型，相比深度强化学习算法，合作搜索效率提高了73.3%。[38]

另一种实施方式是AI生成的光学决策（AGOD）算法，该算法利用扩散模型以改善动态环境中的任务分配，并响应不断变化的人类指导。此外，它还利用了Deep Diffusion Soft Actor-Critic（D2SAC）算法，该算法在任务完成和效用指标方面相比传统的SAC方法显示出微小的改进。[39]

差分模型也被提议作为基于You Only Look Once（YOLO）框架的语义通信管道的一部分，以降低成本并提高无线传输质量，从而创建数字孪生。[40]

区块链方法还展示了确保隐私并改进无人机机队协调的潜力，使用委托权益证明和实用拜占庭容错（DPOS-PBFT）可以增强对节点故障或被篡改的抵抗能力。该方法可以在最少开销的情况下扩展到500个节点，并克服伪造、拒绝服务和篡改攻击。[41]

生成对抗模仿学习（GAIL）结合生成式人工智能与强化学习，以帮助无人机从人类指导中学会在复杂环境中导航。[42] 一个生成关系和意图网络（GRIN）结合了图神经网络和来自变压器的意图模型，以帮助个体无人机考虑与其他无人机的关系。[43] 研究人员还结合了深度强化学习与变换器，以改进对无人机群进行区域覆盖以进行感知或探索的能力，例如在搜救操作中。[44]

GAIL 也被用于改进群集操作。在此方法中，生成器根据资源限制创建模拟轨迹，而判别器则区分这些轨迹。

---

37 侯志远，赵强，巴拉伊赫·巴里亚，梅斯海德·本尼斯，和梅洛希德·德巴赫，“无线多代理生成AI：从连接智能到集体智能”，arXiv预印本 https://arxiv.org/abs/2307.02757，2023。

38 Yu X., Liao W., Qu C., Bao Q., and Xu Z., "基于多智能体生成对抗模仿学习的无人机协同搜索," 在2022国际机器学习、云计算与智能采矿会议（MLCCIM）上发表，2022年，第441-446页。https://ieeexplore.ieee.org/document/9955174

39 Du H, Li Z, Niyato D, Kang J, Xiong Z, Huang H, and Mao S. “生成人工智能辅助优化在边缘网络中的AI生成内容（AIGC）服务,” arXiv预印本 https://arxiv.org/abs/2303.13052, 2023.

40 Du B, Du H, Liu H, Niyato D, Xin P, Yu J, Qi M, Tang Y. 基于YOLO的语义通信与生成AI辅助资源分配在数字孪生构建中的应用 [J]. IEEE物联网杂志 (11:5), 2023. https://ieeexplore.ieee.org/abstract/document/10256109

41 哈菲兹 S, 郑瑞 R, 摩哈齐 L, 孙宇 Y, 和 伊姆兰 M A. “基于区块链的无人机网络在灾后通信中的应用：一种去中心化的群集方法.” arXiv, 2024年3月4日. doi: 10.48550/arXiv.2403.04796.

42 刘光远，阮文辉，杜鸿阳，丁泰黄，杜斯提·尼雅托，朱昆，康嘉文，熊泽惠，阿拔斯·贾马利波尔，金东仁．“生成式AI在无人驾驶车辆群中的应用：挑战、应用与机遇.” arXiv，2024年2月28日．<https://doi.org/10.48550/arXiv.2402.18062>．

43 李莉, 姚杰, 文亮, 何涛, 肖挺, 严杰, 吴普, 张卓, "GRIN: 多智能体轨迹预测的生成关系与意图网络," 进步神经信息处理系统, 卷 34, 页码 27107-27118, 2021.

44 福雷斯 D.，德尔布莱诺 C. R.，哈雷吉扎尔 F.，纳瓦罗 J. J. 和 加西亚 N.，“使用变压器网络解决多个协同无人驾驶航空器的路由问题”，《工程应用的人工智能》，卷. 122，页. 106085，2023。[在线] 可用：https://www.sciencedirect.com/science/article/pii/S0952197623002695

模拟路径以替代真实路径。更高效的深度强化学习算法可以帮助创建用于训练无人机控制器的合成数据。 [45]

# 沟通弹性

**机会:** 提高群体通信在不同环境和不利条件下的弹性。

一项关于太空-空中-地面通信中生成式AI的研究调查了信道建模和信道状态信息估计、资源分配、智能网络部署、语义通信、图像提取与处理、安全性和隐私增强等方面。 [46]

其他研究显示，生成对抗网络（GANs）和变分自编码器（VAEs）能够利用实时信息和无线电波传播分析在动态环境中。在资源分配方面，通用扩散模型（GDM）提高了捕捉时间序列依赖关系和空间相关性的能力；同时，它们还增强了通信效率并减少了干扰。 [47] GANs 有助于生成各种通信场景并评估其相对效率。在网络部署中，VAEs 将原始数据转换为更有效的表示形式，揭示构建更具组织性和适应性的网络所需的设计模式。Transformers 帮助网络设计师理解并做出更好的决策，涉及拓扑结构、节点布局和资源分配。 [48] 语义通信可以应用扩散模型来辨识通信流的结构，并部署变压器以识别建筑物、灌木和湖泊等特征之间的关系。

GANs也可能应用于模拟和评估网络配置以优化网络覆盖，在使用 swarm UAVs 作为移动基站的灾害场景中。 [49]
扩散模型也已被探索用于生成定位策略或改进其他机器学习算法的编码。

另一项专注于物理层安全性的调查表明，各种生成式人工智能（GenAI）算法可以增强通信保密性、身份验证、可用性、韧性和完整性。关键优势包括改进复杂数据分布的表示、对加密数据进行处理转换以及提高网络攻击异常检测能力。 [50]

各种生成人工智能（GenAI）算法在语义通信（SemCom）领域展现出潜力，能够提取和减少特定目标或分析类型所需的重要语义信息。这包括提高用于模型训练的数据质量、从数据集中构建知识库以及优化资源分配。 [51]

---

[45] 苗文清, 曾子墨, 张明明, Quinn 乾, 张子墨, 李思雨, 张乐天, 孙晴. “基于重构环境的边缘资源管理多智能体强化学习”. 在2021年IEEE并行与分布式处理应用国际会议、大数据与云计算国际会议、可持续计算与通信国际会议、社会计算与网络国际会议（ISPA/BDCloud/SocialCom/SustainCom）上发表, 2021年, 第1729-1736页. <https://ieeexplore.ieee.org/document/9644885>

[46] R. Zhang 等人 ， “空地综合网络的生成人工智能 (SAGIN) ” 。 arXiv ， 23 年 11 月 11 日。 doi: 10.48550 / arXiv.2311. 06523 。

[47] H. Du, R. Zhang, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, X. S. Shen, 和 H. V. Poor , “探索无线网络中基于协作分布式扩散的AI生成内容（AIGC）”，IEEE网，页码：1-8，2023

[48] 一种用于无人机路径设计的迁移学习方法：考虑连通性中断约束

[49] 鲁žička M，沃洛申 M，加兹达 J，马克西穆克 T，韩 L 和 多勒尔 M，“基于无人飞行器的网络覆盖优化的快速高效生成对抗网络算法”，《分布式传感器网络国际期刊》，第18 卷，第 3 期，页码：15501477221075544，2022年3月，doi: 10.1177/15501477221075544。

[50] 王J. 等人，“生成式AI在综合 sensing 和通信中的应用：从物理层视角的见解”。arXiv，2023年10月29日。doi: 10.48550/arXiv.2310.01036.

[51] 李川等，”基于生成式AI的语义通信网络：架构、技术与应用。“arXiv，2024年1月7日。访问日期：2024年3月15日。[在线]. 可用: http://arxiv.org/abs/2401.00124

多输入多输出 (MIMO) 天线的未来创新可能会动态调整
signals sended across an array of linked tenes. Such applications could play an essential role in improving communications with autonomous systems. This approach does, however, require
用于估计 RF 信号的特性及其对环境的响应的模型。[52]
各种研究人员已经研究了诸如变分递归之类的生成模型
神经网络 (VRNN) 和基于分数的方法可以生成时间序列数据 ，
突出其在从金融、医疗保健、气候建模到预测维护等领域的潜在能力，以捕捉复杂的时变依赖关系并提高预测准确性。

# 共识

**机会:** 自动创建更高效、更可靠的容错共识协议。

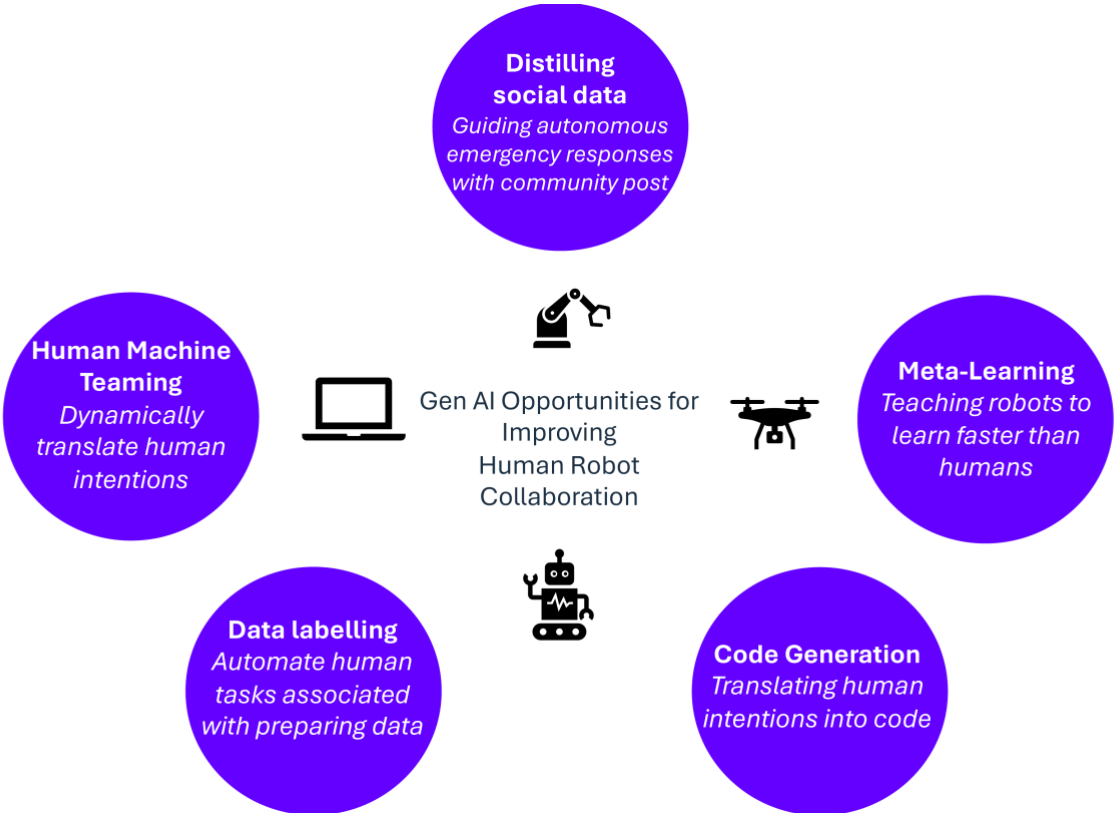
Vaz等人探索了各种生成型AI工具如何改善分布式计算共识协议的设计。[53] 这通常是一个耗时且手动的过程。他们发现ChatGPT和GitHub CoPilot在生成有用代码方面效果不佳。该团队通过使用强化学习算法生成候选算法代码协议，并在评估阶段衡量其适用性，获得了更好的结果。尽管这项工作仍处于早期阶段，作者指出这为分布式计算的新研究打开了大门，并提高了自主车队的沟通与协调效率。

[52] M. Arvinte 和 J. I. Tamir, "基于评分生成模型的MIMO信道估计,"《IEEE无线通信期刊》，第22卷，第6期，页码：3698-3713，2023. <https://ieeexplore.ieee.org/document/9957135>

[53] Vaz D, Matos D R, Pardal M L, et al. “使用生成性AI自动生成分布式算法”[C]/. 2023第53届IEEE/IFIP依赖系统与网络国际会议补充卷（DSN-S），2023年6月，第127-131页. doi: 10.1109/DSN-S58398.2023.00037

## Human

研究人员还探索了如何通过生成型人工智能（GenAI）提高自主系统设计、开发和运营中的人文因素，以使这些系统更加可信、安全、可靠和有韧性。这些算法有望生成新的数据，并从无人机舰队、自动驾驶车辆、物联网设备和嵌入式系统捕获的海量数据中提炼出重要的见解和信息。


Distilling social data
Guiding autonomous emergency responses with community post
Human Machine Teaming
Dynamically translate human intentions
Gen AI Opportunities for Improving Human Robot Collaboration
Meta-Learning
Teaching robots to learn faster than humans
Data labelling
Automate human tasks associated with preparing data
Code Generation
Translating human intentions into code


**自动更新:** 用于维护最新 AOSP 的强化版本的工具

# 任务规划和执行

**机会:** 改进允许人类计划任务并将这些计划传达给无人机机队的工具。

生成式AI展现出巨大的潜力，可帮助人类团队规划和执行涉及无人机及其他自主系统的任务。大型语言模型（LLMs）应将高层级的任务目标转化为具体的任务，这些任务可以由机群和单个无人机执行。美国军事部门已经开始在“全球信息主导试验计划”（GIDE）项目中测试生成式AI。GIDE报告了至少五个不同模型的进展，但并未提供更多细节或具体的应用案例。 [54]

美国小企业创新研究（SBIR）计划和小企业技术转让计划也请求投标开发可信赖的生成式人工智能（GenAI），以结构化数据并为指挥、控制通信和计算机（C4）系统提供准确的洞察。 [55] 公司C3.ai还专门开发了一门学科，将生成式人工智能（GenAI）应用于国防领域，以“加速人员、情报、运作、物流和保障等功能的数据驱动决策过程以及规划工作。” [56]

DEVCOM陆军研究实验室探索了使用大型语言模型（LLMs）来开发行动方案（COAs，即作战计划）的可能性，作为COA-GPT的一部分。 [57] 在一个演示中，新模型能在秒内生成COAs，并在军事化版本的《星战争霸II》游戏中根据操作员反馈进行响应。

# 人机合作

**机遇 ： 改善人与机器之间的协作**

很少有公开描述了人机程序的技术细节。关于它们可能应用于民用场景的信息更是少之又少。然而，似乎有可能利用通用人工智能（GenAI）的创新来在此基础上改进人机协作——例如，在人类与自主系统紧密反馈循环的人机团队合作（HMT范式）中，这种技术可能会推动超越人工在环路的方法，促进人类与自主系统的团队合作。 [58]

另一种框架——监控、分析、规划和执行（MAPE）——已被提出以管理不确定或动态环境中的安全问题。MAPE-HMT结合了HMT和MAPE框架，尽管人类合作者与自主系统之间的运行速度存在巨大差距。 [59] 虽然这种努力并不直接实现任何

---

[54] 哈珀J，“五角大楼在‘全球信息主导权’实验中测试生成式AI”，DefenseScoop。访问日期：2024年3月27日。[在线]. 可用: https://defensescoop.com/2023/07/14/pentagon-testing-generative-ai-in-global-information-dominance-experiments/

[55] 可信的生成人工智能（GenAI）用于结构化数据和提供指挥、控制、通信和计算机（C4）系统准确洞察 | SBIR.gov。访问日期：2024年3月30日。[在线]. 可用链接：https://www.sbir.gov/node/2479917

[56] “用于防御的生成 AI ”， C3 AI 。访问时间 ： 3 月。 2024 年 27 日。 [在线] 。可用 ： https: / / c3. ai / generative - ai - for - defense /

[57] COA-GPT：军事行动中加速行动方案开发的生成预训练变换器，“信息系统技术（IST）小组组织的ICMCIS会议（IST-205-RSY）”，由北约科学技术组织（STO）科学和技术组织（S&T）于2024年4月23日至24日在德国科布伦茨举行。访问日期：2024年3月27日。[在线]. 可用链接：https://arxiv.org/html/2402.01786v1

[58] 张, Encoded, 等. "大型语言模型在人机交互中的应用：一个综述." 生物模仿智能与机器人(2023) 3:4 (100131). <https://www.sciencedirect.com/science/article/pii/S2667379723000451>

[59] J. 克莱兰-黄，A. 阿格瓦尔，M. 维赫豪泽，M. 墨菲 和 M. 普里埃托，“将 MAPE-K 扩展以支持人机团队协作”，《自适应和自我管理软件工程研讨会论文集》，第 17 届，页码 120–131，2022 年 5 月，doi: 10.1145/3524844.3528054.

通用人工智能确实建议采用能够从人类指令到详细自主系统任务执行的全范围通用人工智能技术的架构。 [60]

大型语言模型（LLMs）在从社交媒体流中提炼数据以帮助指导灾害响应方面显示出潜力。例如，VictimFinder项目尝试了十种不同的模型来筛选救援请求。 [61] 研究人员评估了基于Google的BERT大规模语言模型（LLM）的模型，这些模型在性能上始终优于传统的语言处理模型。这表明，大规模语言模型可能有助于指导无人机机群的搜索和救援操作——规划覆盖范围、分类严重程度，并在大型灾难后优先考虑人类救援工作。

最近，张和索研究了大型语言模型（LLMs）作为人类在人机交互（HRI）中的替代者。他们发现，经过大量人类生成文本训练的LLMs能够在零样本场景中有效预测人类行为，与专门模型的表现相当。通过包括基于信任的任务和餐具传递任务在内的实验，LLMs规划机器人活动的能力优于简单的方法。然而，研究人员也指出了一些局限性，如提示敏感性和在空间和数值推理方面存在的困难，因此LLM/HRI集成仍需被视为有前景但仍在发展中的领域。 [62]

基于图神经网络的多语言文本分类框架（GNoM）将transformer与图神经网络（GNN）的组合扩展至多语言社交媒体分析。 [63] 图形方面旨在帮助连接多种语言中的相关词汇，并捕捉未标记数据的上下文。

# 元学习

**机会 ： 教机器人在更少的人为监督下更快地学习。**

元学习探索了加快自主系统学习的方法。一个早期的应用案例是利用大规模语言模型（LLM）帮助将人类需求转化为具体任务，并将其动态定制化地体现在每个机器人自主系统的代码和策略中。代码生成是常见的LLM应用之一，而一些较新的多模态LLM，如PaLM，也可以编写机器人代码。 [64] 这使得非专家能够指导行为、根据反馈调整这些行为，并编写和定制代码片段以开发新任务。然而，用户必须谨慎：在较长的交互过程中，可能会丢失某些上下文信息，从而导致简单行为的细节丢失。语言模型预测控制（LMPC）是一种框架，用于针对78项不同任务对PaLM 2进行微调，以便教授不同的机器人模型。 [65] 早期结果通过提高新任务的教学成功率26.9%，显著减少了人类纠正的数量，并且还通过提高新机器人配置的成功率31.5%来增强教学效果。

MetaMorph 是另一个使用变压器学习通用控制器的框架，该控制器能够将基于文本的任务描述转换为可在各种模块化机器人上运行的命令。

---

[60] Mai ， Jinjie 等。 “作为机器人大脑的 Llm ： 统一以自我为中心的记忆和控制。 ” arXiv 预印本 arXiv ： 2304.09349(2023) 。

[61] 周B. 等人，“VictimFinder：利用BERT从社交媒体中获取救援请求以应对灾难响应”，《计算机、环境与城市系统》，第95卷，页码：101824，2022年7月，doi: 10.1016/j.compenvurbsys.2022.101824. https://www.sciencedirect.com/science/article/abs/pii/S0198971522000680

[62] 张 Bowen 和郝尔德·苏. "大型语言模型作为零-shot 人类模型用于人机交互." 2023 年 IEEE/RSJ 国际智能机器人与系统会议 (IROS). IEEE, 2023. https://ieeexplore.ieee.org/document/10341488

[63] 苏什（S. Ghosh）、马吉（S. Maji）和德萨卡尔（M. S. Desarkar），“增强语言模型的图神经网络用于高效的多语言文本分类。”arXiv，2022年3月6日。访问日期：2024年3月27日。[在线]. 可用: http://arxiv.org/abs/2203.02912

[64] Anil, Rohan, et al. "PALM 2 technical report". arXiv preprint https: / / arxiv. org / abs / 2305.10403 (2023).

[65] J. Liang 等人，"通过语言模型预测控制从人类反馈中更快地学习"。arXiv，2024年2月17日。访问日期：2024年3月27日。[在线]. 可用: http://arxiv.org/abs/2402.11450

配置。 [66] MetaMorph 在未见过的动力学和运动学变化、全新形态和任务中展现出强大的泛化能力。这一特性使得基于Transformer的控制器能够在多样化的设计空间内训练后，能够有效适应广泛的自主系统。例如，类似的训练方法可以扩展到各种类型的无人驾驶车辆，包括无人机、地面机器人或自动化物料处理系统。这种适应性确保了训练后的模型能够在不同的操作环境和物理配置下运行，而无需进行重新训练。

# 数据标记和合成

许多最先进且高效的机器学习算法需要在已标注数据上进行训练，以指示数据的属性。这可能涉及描述图像中的对象，或将其归类为正常或恶意网络流量、正常或异常的操作条件等。这一过程往往耗时且劳动密集。在此情况下，专门领域训练的生成AI模型可以协助自动化现有数据的标签生成，从而加快工作流程并减少对人力的需求。

在其他情况下，数据科学团队可能拥有一个显示正常条件的大规模数据集和一个突出异常条件的小规模数据集。在此类情况下，可以通过生成对抗网络（GANs）、扩散模型（Diffusion Models）和变分自编码器（VAEs）等技术高效地建模异常数据的基本特征，从而生成更为平衡的数据集。这有助于训练其他机器学习算法，这些算法可能比GenAI算法更快且更高效。

为了了解这些方法在实际中的运作方式，请参见关于异常检测、恶意软件检测、入侵检测和威胁模拟的相关章节。

[66] A. 约塔, L. 范, S. 加内里, 和 L. 飞-飞, “MetaMorph: 使用变换器学习通用控制器.” arXiv, 2022年3月22日. 最后访问日期: 2024年3月15日. [在线]. 可用: \[Online]. Available: http: / / arxiv. org / abs / 2203.11931

网络安全对于提高信息技术和自主系统的安全性和韧性至关重要。自2019年以来，超过1700篇论文探讨了如何通过生成人工智能（GenAI）技术，特别是生成对抗网络（GANs）和变分自编码器（VAEs），来提升恶意软件和入侵检测的效果。其他研究则分析了如何将各种GenAI方法应用于改善政策管理、威胁模拟和供应链可见性。目前，大多数这些技术主要针对具有较强计算能力的信息技术系统。然而，随着模型性能的提升和嵌入式系统硬件的发展，这些能力将逐渐扩展到自主系统。许多综述也探讨了在应用GenAI时的一些机遇与权衡。

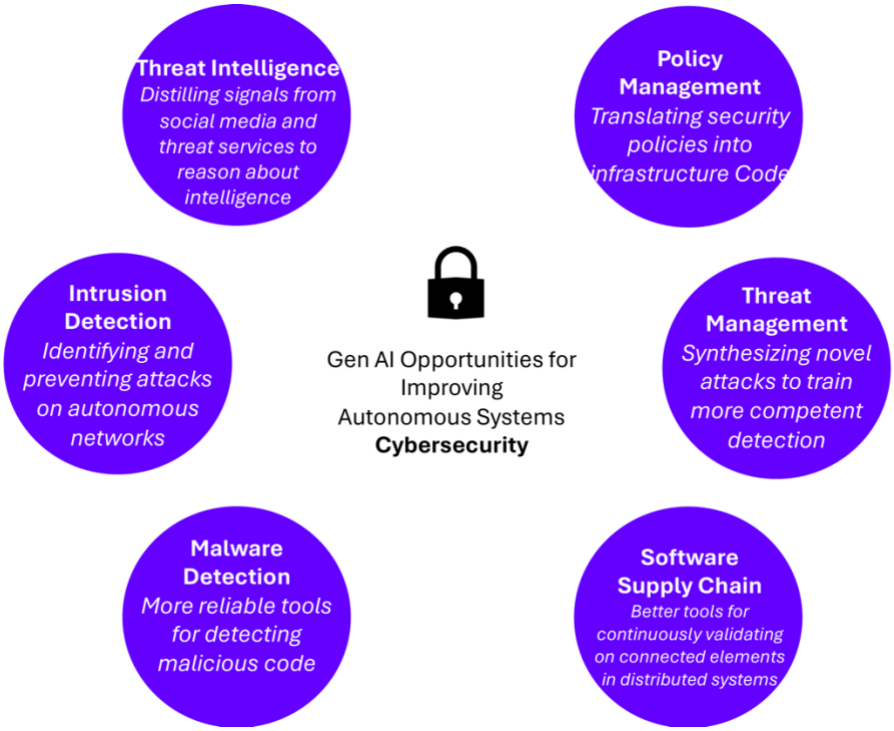
使用GANs、对抗网络和变换器等技术来提升网络安全。其他研究人员还探索了AI在机器人技术和网络防御方面的某些安全影响。[67]
Divakaran 和 Peddinti 探索了大型语言模型(LLM) 的变革影响

关于网络安全和安全。[68] 他们的研究突显了大型语言模型（LLMs）的多样性和强大功能，这些模型已在许多行业中变得至关重要，特别是在应对复杂挑战方面。

[67] S. Neupane, I. A. Fernandez, S. Mittal, and S. Rahimi, “生成人工智能技术对网络防御的影响与风险。”arXiv, 2023年6月22日. doi: 10.48550/arXiv.2306.13033.
[68] D. Divakaran 和 S. Peddinti 的 “网络安全 LLM ： 新机遇 ” 。 arXiv ， 2024 年 4 月 17 日。 arXiv.2404.11338

安全领域。这些先进模型的出现为解决与网络安全和安全相关的重要而复杂的问题提供了新的视角和解决方案，表明了在利用尖端人工智能技术管理并减轻此类问题方面发生了显著的转变。

其中一个目标是通过隔离应用程序执行和更精确地调节它们的交互来增强安全性。对各种安全威胁进行测试表明，这种设置能够有效保护系统，并且对性能影响较小。SECGPT 代表了向更安全的基于大语言模型（LLM）的应用程序迈出的重要一步，邀请进一步的研究和开发。[69]




# 恶意软件检测

**机会:** 改进检测运行系统中恶意软件的工具。研究人员探索了各种方式，利用生成式人工智能（GenAI）帮助建模合法代码与恶意代码之间的关键差异。这些模型能够检测传统病毒扫描器中常用的静态代码分析可能无法触发的恶意软件。

例如，一种操作码序列放大方法从运行程序中提取机器指令，并使用生成对抗网络（GANs）将新程序中的字节序列与恶意软件中的字节序列进行比较。这种方法将恶意软件检测率从96.3%提高到98%。研究团队还利用GANs生成与恶意软件匹配的合成数据，以大幅增加对原始训练数据未涵盖的边缘案例的检测。[70]

---

[69] Wu，Yuhao，等。“SecGPT：基于 LLM 的系统的执行隔离体系结构”。 *arXiv 预印本 arXiv: 2403.04960* (2024).

[70] 崔昌 (C. Choi)、申尚 (S. Shin) 和 李映 (I. Lee)，"基于序列生成对抗网络的Opcode序列放大器"，《2019年信息与通信技术融合国际会议（ICTC）》 proceedings, 韩国济州岛，2019年，页码：968-970, doi: 10.1109/ICTC46691.2019.8940025.

经典的GAN算法适用于图像处理，研究人员开发了各种方法将原始数据转换为图像格式或利用其他数据模型。例如，MalIAGAN将恶意软件样本转换为图像格式以进行训练。尽管AI仅在1%的数据集上进行了训练，他们仍实现了80%的准确率。[71] 可视化恶意软件检测框架采用了一种不同的方法，将代码特征转换为图像，从而帮助生成对抗网络（GANs）生成用于训练神经网络的合成数据。[72]

滑动本地注意力机制（SLAM）利用了一个API执行序列语义提取引擎来通过其执行属性识别恶意软件。残差注意力机制也显示出超越传统神经网络方法的潜力。[73] 另一个应用I-MAD展示了如何使用变压器创建一个可解释的恶意软件检测器，该检测器量化了特征的影响以指导结果。[74] 一种分析内部结构的分层变压器模型也展现了一定的前景，有助于提高预测准确性。[75]

# 入侵检测:

**机会:** 提高对内部总线和外部网络攻击的检测能力，并模拟更广泛的IDS威胁谱系，这些威胁可能超出了传统模式的范围。

研究人员正在探索如何利用生成人工智能（GenAI）技术帮助创建更精确的模型来刻画恶意网络流量的基本属性。例如，变分自编码器（VAEs）可以将原始的入侵检测系统（IDS）数据转换为更为紧凑的表现形式，突出显示与各种类型入侵相关的特点。传统的VAE可以从未标记的数据中学习与入侵相关的模式，从而减少手动标注的需求。然而，需要注意的是，一些研究人员发现通过标注少量数据可以进一步提升效果。例如，条件变分自编码器（CVAE）利用部分标注数据来改进编码算法，相比传统方法显著提高了检测性能。[76]

TI研究所研究人员开发了ARCADE（ adversarially regularized convolutional autoencoder for unsupervised network anomaly detection），该方法利用子集的数据包信息构建卷积自编码器来高效地检测攻击。[77] 它仅使用两个初始数据包即可实现更高的准确性，并且其参数量仅为传统模型的十分之一，从而允许显著加快检测和响应时间。

其他研究者已经探索了如何通过防火墙日志数据使生成对抗网络（GAN）方案学习识别异常条件。[78]

---

[71] 刘宇、李杰、刘斌、高晓和刘晓，《基于图像分析的恶意软件识别方法》，《第11届信息技术与医学教育国际会议（ITME）论文集》，2021年11月，pp. 157-161。

[72] 王芳、哈马迪·贺宾和达米亚尼·埃德蒙多，“基于CNN和条件GAN的可视化恶意软件检测框架”，《2022年IEEE大数据会议（Big Data）》，2022年12月，第6540-6546页。DOI: 10.1109/BigData55660.2022.10020534。

[73] SLAM：基于滑动局部注意力机制的恶意软件检测方法，《安全与通信网络》，第2020卷，2020年，doi: 10.1155/2020/6724513.

[74] 李明奇, 首任斌, 查朗, 丁少华. "I-MAD: 使用Galaxy Transformer的可解释恶意软件检测器".《计算机安全》, 第108卷, 2021年9月, 页码: 102371, doi: 10.1016/J.COSE.2021.102371.

[75] Hu X, Sun R, Xu K, Zhang Y, Chang P. 基于分层变压器模型利用内部结构信息进行物联网恶意软件检测 [J]. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2020, 2020, 927-934. DOI: 10.1109/TRUSTCOM50675.2020.00124.

[76] A. 汉南, C. 格鲁尔, 和 B. 西克, “基于异常检测的弹性网络入侵检测方法研究——使用推断自编码器,” 在 2021 年 IEEE 互联网安全与韧性国际会议 (CSR) 上发表, 2021年7月, 页码: 1-7. doi: 10.1109/CSR51186.2021.9527980.

[77] W. T. Lunardi, M. Andreoni和J. -P. Giacalone，"ARCADE：对抗正则化卷积自编码器在网络安全异常检测中的应用"，《IEEE网络与服务管理 Transactions》，第20卷，第2期，页码：1305-1318，2023年6月，DOI: 10.1109/TNSM.2022.3229706

[78] S. P. Kulyadi, P. Mohandas, S. K. S. Kumar, M. J. S. Raman, 和 V. S. Vasan，“基于生成对抗网络的防火墙日志消息数据异常检测”，发表于第13届国际电子、计算机与人工智能会议（ECAI），2021年7月，页码：1-6。https://ieeexplore.ieee.org/document/9515086

一个平衡的数据集对于许多机器学习的应用至关重要，但也可能面临挑战。实践中，大多数训练集包括大量的无害流量示例，而攻击流量的代表性则要小得多。生成对抗网络（GANs）可以通过生成新的、合成的攻击数据来补充这些样本，这些数据能够代表重要的数据属性。这种更为平衡的数据集有助于改进使用其他机器学习技术的算法开发。例如，FlowGAN可以在流量被加密（如通过VPN）的情况下，帮助扩充数据集。 [79]

研究人员有时通过结合生成对抗网络（GAN）和变分自编码器（VAE）等GenAI技术获得了更好的结果。例如，AE-CGAN在单独使用VAEs或GANs时表现更好。它还能够适应网络变化。 [80] .

GIDS是一种基于GAN的汽车网络入侵检测系统，该系统通过训练正常数据进行训练，并能更好地检测未知攻击。此模型将控制器区域网络（CAN）中车辆内的消息转换为图像以供GAN处理。生成的模型平均每条消息耗时0.18秒即可分类1,954条CAN总线消息，速度快到足以跟上标准交通条件，并实现了98.7%的检测准确率。 [81]

研究人员还探索了如何将基于生成对抗网络（GAN）的入侵检测系统（IDS）部署到移动边缘计算（MEC）服务器上，从而允许它们从较不具备处理能力的物联网设备中卸载处理任务。 [82] 这种方法有助于提高计算密集型安全功能在作为一系列自主系统中心高性能计算机中的使用效率。

一种基于简单循环单元（SRU）的模型被发现能够显著提高IDS的准确性和速度，通过将原始数据转换为更高效的格式，并使用GAN和LSTM模型进行处理。在基准数据集上，系统实现了超过99.62%的准确率。

一个令人 intrigue 的 GAN 应用是生成越来越逼真的网络流量，这些流量难以被传统的入侵检测系统（IDS）识别。例如，使用沃爾舍廷詹式生成对抗网络（Wasserstein GANs, WGANs）生成的伪恶意流量将传统 IDS 的检测率从 97.3% 降至 47.6%。 [83] 相反，一旦被标记为“恶意”，这些看起来更正常的数据也可以用于使用传统机器学习技术训练更有效的入侵检测系统（IDS）。同一组作者还提出了一种相关方法，称为DoS-WGAN，以生成新样本来帮助训练更好的IDS系统。另一种努力结合了变分自编码器（VAEs）和生成对抗网络（GANs），以细化Wasserstein距离的重构误差，从而大幅提高分类准确性。 [84]

---

[79] 王泽, 王鹏, 周晓, 李帅, 和 张明, “FLOWGAN: 基于生成对抗网络的不平衡网络加密流量识别方法,” 出席 IEEE 平行分布式处理与应用国际会议, 大数据与云计算会议, 可持续计算与通信会议, 社交计算与网络会议 (ISPA/BDCloud/SocialCom/SustainCom), 2019年12月, 页码: 975-983. [在线]. 可用: https://ieeexplore.ieee.org/document/9047386

[80] Lee ， JooHwa 和 KeeHyun Park 。 “基于 AE - CGAN 模型的高性能网络入侵检测系统 ” 。 *应用科学* 9.20(2019 年) ： 4221 。 https: / / www. mdpi. com / 2076 - 3417 / 9 / 20 / 4221 / pdf

[81] 李胜欧, 宋洪民, 金洪可, "GIDS：基于GAN的车载网络入侵检测系统," 在第16届隐私、安全与信任会议（PST）上发表, 2018年8月, 页码：1-6. https://ieeexplore.ieee.org/document/8514157

[82] H. Sedjelmaci，《基于生成对抗网络方法的攻击检测与决策框架：车辆边缘计算网络案例》，《新兴电信技术杂志》（Trans. Emerg. Telecommun. Technol.），第33卷，第10期，2022年10月，文章编号：e4073。https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4073

[83] Gulrajani, Ishaan, 等. "改进的Wasserstein GAN训练方法." 进步神经信息处理系统30 (2017). https://dl.acm.org/doi/10.5555/3295222.3295327

[84] Park C, Lee J, Kim Y, Park J-G, Kim H, and Hong D. "基于生成对抗网络的增强AI网络入侵检测系统."《IEEE互联网事物杂志》, 第10卷, 第3期, 页码: 2330-2345, 2023年2月. [在线]. 可用: https://ieeexplore.ieee.org/document/9908159

# 集成威胁情报

**机会 ： 改进跨多个渠道的新威胁检测。**

LLM可以整合来自多个来源的非结构化数据。迄今为止，关于这如何有助于集成威胁情报的研究还很少。除了在GenAI研究领域之外，一些研究人员探索了自然语言处理技术如何帮助聚合和解析来自Twitter的数据。例如，CyberTwitter是一个研究项目，该项目从Twitter中发现了并分析了网络安全情报。 [85] 该项目利用语义网RDF创建了相关数据的知识图谱，并开发了用于安全研究人员推理此智能的工具。

# 策略管理

**机会:** 在各种 IT 和物理系统中准确准确地转换安全策略。

政策管理传统上需要大量的手动努力来帮助防止攻击而不影响合法用户。研究人员探索了如何利用大型语言模型（LLMs）和扩散模型来制定更有效的攻击策略，并且反过来自动化安全策略的配置以防止无线网络中的攻击。 [86] 一个有趣的结果是 ： 该方法降低了缓解数据中毒攻击所需的能耗。

IT供应商Kiteworks认为，政策制定者也必须采用零信任方法以提高对抗AI系统攻击的韧性。这涉及验证系统输入、监控持续过程、检查输出，并在每个阶段进行凭证认证。 [87]

传统上，制定复杂的安全规则需要深厚的专业知识来导航网络安全环境并制定专家系统可以应用的规则。这是一个耗时且技术要求高的过程。大型语言模型（LLMs）提供了一种解决方案：它们通过将非专家的自然语言描述转换为正式的安全策略来减少工作负担，利用其高级自然语言处理能力来解释用户意图并创建适当的安全政策规则。

# 威胁模拟:

**机会:** 生成更真实和可变的合成攻击数据 ， 用于为不可预见的事件训练新的 AI 分类器。

大型语言模型也被提议用于生成越来越逼真的诱饵（honeypots），以吸引黑客的注意力并使其远离核心系统，从而增强系统安全。 [88] 这些蜜罐可能会增加另一个保护层 ， 并使无人机舰队攻击者感到困惑。

---

[85] S. Mittal 等人，“CyberTwitter：利用Twitter生成网络安全威胁和漏洞警报”，发表于IEEE/ACM国际社交网络分析与挖掘会议（ASONAM），2016年，第860-867页。https://dl.acm.org/doi/10.5555/3192424.3192585

[86] Du H. 等人, “矛或盾：利用生成式AI应对智能网络服务的安全威胁.” arXiv, 2023年6月4日. doi: 10.48550/arXiv.2306.02384.

[87] T. Freestone，“以零信任方法建立生成式AI的信任”，Kiteworks | 你的私有内容网络。访问日期：2024年3月15日。[在线]. 可用：https://www.intelligentciso.com/2024/02/08/building-trust-in-generative-ai-through-a-zero-trust-approach/

[88] G. Liu 等人 ，“无人驾驶车辆群的生成 AI ： 挑战 ， 应用和机遇。 ” arXiv ， 2024 年 2 月 28 日。 doi ： 10.48550 / arXiv.2402.18062 。

IDSGAN 是另一种生成看起来无害的攻击模式的方法，这些模式更有可能避开传统的IDS。[89] Intell-dragonfly 使用变换器生成多样化攻击场景、元素和方案。研究人员报告了在攻击面生成方面的准确度、多样性和可操作性提升。这项工作还有助于提高检测这些新型攻击的算法训练效果。[90]

## 软件供应链可见性

研究人员还探讨了零信任方法如何改善电力电网供应链的安全性。该框架增强了对重放攻击和协议类型攻击的检测能力。关键要素之一是一种新的零信任框架，用于持续验证设备和控制消息的信任度。研究人员展示了在防止生成式人工智能（GenAI）攻击方面达到99%的信心水平。

## 挑战

尽管生成式人工智能（GenAI）的进步速度迅猛，但在将其作为零信任系统的一部分应用于安全、安全性和韧性方面更广泛部署之前，仍需解决众多关键挑战。早期的应用案例将集中于低风险任务，或者用于改进其他机器学习算法的创建，这些算法可能更加高效和性能更强。然而，用户希望获得更多功能，例如访问标准办公软件。

套件和其他生产力应用程序 ， 而不会为其他不太敏感的应用程序打开安全边界。


[89] 林泽, 石洋, 许 Zeus，“IDSGAN：针对入侵检测的生成对抗网络攻击生成方法，” 第十三届太平洋亚洲知识发现与数据挖掘会议, 2022 (Springer) https://arxiv.org/abs/1809.02077

[90] 武X，邱Q，李J，和赵Y，“Intell-dragonfly：基于人工智能生成内容技术的网络安全攻击面生成引擎。”arXiv，2023年10月31日。访问日期：2024年3月15日。[在线]. 可用: http://arxiv.org/abs/2311.00240

# Cost

AI 先驱 Gary Marcus 观察到，尽管在自动驾驶车辆上已经投入了超过 1000 亿美元，当前所有系统仍然难以处理边缘情况，并且仍需人类参与其中。[91] 他指出，大型语言模型（LLMs）和其他生成性AI算法在一段时间内仍将在边缘案例上面临挑战。他也并不相信仅凭LLMs就能适用于某些高风险用途。当前的LLM架构在投资呈指数级增长的情况下，仅显示出微小的进步。此外，值得注意的是，使用像GPT-4这样的LLM模型会产生成本，大约每750词费用为0.006美元。同样，使用类似DALL-E 2的模型生成一张图片的成本约为0.18美元。[92]

# 计算

生成人工智能（GenAI），特别是基于变换器的大规模语言模型（LLM），在训练和推理过程中通常需要更多的计算资源，这将限制GenAI的应用部署到高端服务器与自主系统协作的环境中。GenAI工具还可以生成和标注数据，以训练其他表现更佳的机器学习算法。硬件供应商正在稳步改进底层自主系统硬件，这些硬件可能很快能够直接运行更多的GenAI模型。此外，现场可编程门阵列（FPGAs）也显示出比GPU或CPU更高效地运行GenAI模型的潜力。[93]

# 适应

AI模型通常通过微调过往数据来实现高准确度和性能。然而，这些模型可能会因环境或操作变化而失去准确性。恶意行为者可能调整自己的网络安全技术和策略以突破防御机制。在传统的开发生命周期中，模型必须基于新数据进行更新。像在线学习这样的方法有时会使用生成对抗网络（GAN），并显示出一定的潜力以自动适应并维持模型的准确性。[94]

# 道德 / 监管

生成AI的快速成长引发了全球范围内的伦理和监管担忧。生成AI模型通常比其他机器学习技术更加不透明且难以审计，这可能导致更难识别意外行为、幻觉或偏见的根本原因。

# 对齐

更复杂的模型可能导致无人机执行未预期的动作。对齐研究仍在 matures 中，但其对于识别自主系统可能引发意外危害的方式仍至关重要。

---

[91] 古德曼·马库斯，《有史以来人工智能领域第二糟糕的100亿美元投资？》，马库斯谈人工智能。访问日期：2024年3月30日。[在线]. 可用: https://garymarcus.substack.com/p/the-second-worst-100b-investment

[92] 黄，浩宇。2023。《CEO们需要了解的关于采用生成式AI的成本》。哈佛商业评论。访问日期：2024年3月30日。[在线] 可用：https://hbr.org/2023/11/what-ceos-need-to-know-about-the-costs-of-adopting-genai

[93] V. 塔杜、K. 安朱和 D. 波米利，“基于 FPGA 的机载深度学习无人机故障原因检测与分类”，《IEEE 机器人学交易》，2023. https://ieeexplore.ieee.org/document/10122168

[94] 谭伟乐和杜龙豪·特鲁昂，《利用合成对抗样本增强恶意软件检测的稳健性》，《IEEE全球通信会议 Proceedings GLOBECOM》，2020年12月，第1-6页。https://ieeexplore.ieee.org/document/9322377

# 隐私

敏感数据需要在不影响操作的情况下被屏蔽。

这在自主系统中正成为一个日益严重的关注点，随着无人机和自主系统收集越来越多关于个人和私有财产的数据。隐私增强计算的相关工作可以解决其中的一些担忧——使用诸如安全 enclave、联邦学习和同态加密等技术。然而，这些方法确实会在训练和推理过程中增加额外的开销。

研究人员还探索了如何利用生成人工智能（GenAI）技术自动屏蔽敏感数据，特别是在某个特定任务不需要这些数据时。例如，在自动驾驶生成对抗网络（ADGAN）中，背景图像会被屏蔽，而交通标志、道路基础设施和行人移动数据则会被捕捉。[95] TrajGANs可以生成模拟人类运动模式的合成数据，而不泄露其活动的敏感信息。VAEs还被用于向原始数据添加噪声，以保留关键特征并保护隐私。[96]

此外，研究显示披露个人隐私信息的风险，因为训练数据可能包含姓名、电子邮件地址和电话号码等可识别细节。[97]

此外，图像扩散模型引发了争议，尤其是像Stable Diffusion、OpenAI的DALL-E 2以及Google的Imagen等人工智能模型。这些模型通过噪声训练数据和提示学习生成原创图像。它们经常未经许可或补偿就使用真实艺术家的作品进行训练，导致生成的艺术作品模仿原始风格或包含失真的艺术家签名。进一步的研究表明，这些模型有时可以重现它们训练时使用的精确图像，仅存在细微差别，这引发了关于原创性和版权侵权的担忧。[98]

# 准确性

生成式AI模型可以以多种方式产生幻觉。在高风险情况下，必须与其他技术相结合，以帮助检测和预防意外问题。

# Size

在边缘设备上运行当前的通用人工智能（GenAI），尤其是大型语言模型（LLMs），存在一定的困难。最强大的GenAI模型有时拥有数十亿个参数，这使得它们难以部署并在资源有限的边缘设备上运行。针对特定任务训练的领域专用小型语言模型可以有所帮助。其他技术如生成对抗网络（GANs）也存在规模问题。一种解决方案是在与自主系统通信的邻近移动边缘计算（MEC）服务器上运行该模型。[99] 嵌入式系统供应商如高通正在开发更为先进的硬件，未来可能支持更大的生成式人工智能模型。

该领域在应对将大型语言模型（如GPT和LLaMa）部署到边缘设备所面临的挑战方面取得了显著进展，以实现更高的隐私性和效率。值得注意的是，PrivateLoRA引入了一种混合模型，利用边缘计算技术。

---

[95] 熊志，蔡舟，韩强，阿勒瓦伊斯·阿里王子，李伟．“ADGAN：保护自动驾驶车辆摄像头数据中的位置隐私”[J]．IEEE工业信息化，2021，17（9）：6200-6210．GL：只是汽车．<https://ieeexplore.ieee.org/document/9234081>

[96] 赵奥博耶, 塔里克·达加希, 马哈茂德·巴巴伊, 马赫穆德·萨拉伊和于承铭, “Deepclean：一种用于自主车辆摄像头数据隐私保护的稳健深度学习技术,”《IEEE访问》，第10卷，页码：124534-124544，2022. [在线]. 可用: https://ieeexplore.ieee.org/document/9954019

[97] Nasr, Milad, et al. "Scalable extraction of training data from (production) language models." arXiv preprint arXiv: 2311.17035 (2023).

[98] 卡林尼，尼古拉斯，等．“从扩散模型中提取训练数据．”第32届USENIX安全研讨会（USENIX Security 23）．2023．可供查阅：https://arxiv.org/abs/2301.13188

[99] H. Sedjelmaci，“基于生成对抗网络方法的攻击检测与决策框架：车辆边缘计算网络案例”，《新兴电信技术杂志》（Trans. Emerg. Telecommun. Technol.），第33卷，第10期，2022年10月，文章编号：e4073。

在保持云级计算效率的同时实现隐私保护，优化通信并提升个性化AI应用的数据吞吐量。[100] 同样地，LLM作为系统服务（LLMaSS）专注于通过创新的内存管理策略减少内存密集型LLM的上下文切换延迟，从而实现更高效的设备端执行。[101] complementing 这一点，EdgeMoE 成为了首个适用于混合专家（Mixture-of-Experts，MoE）大型语言模型（LLMs）的本地设备推理引擎，通过策略性模型分区和专家管理技术实现计算效率。[102]

# 新的安全问题

生成式AI模型还引发了若干新的安全漏洞，包括提示注入、越狱攻击、数据泄露以及不当依赖。[103] 提示注入通过向请求中插入恶意输入来启动未经授权的操作，当模型连接到其他系统时，这可能会成为一个重大问题。破解绕过了模型的限制以生成禁止的内容。数据污染涉及篡改用于训练模型的数据。特别构造的提示有时可以促使模型泄露敏感或专有的训练数据，包括个人资料信息、公司机密或安全令牌。生成式AI工具，尤其是大型语言模型（LLMs），还可以权威性地生成不准确或误导性的信息，而这些信息可能没有得到充分测试。例如，自动化代码生成器可能会生成功能上正确但不安全的代码。未来的研究需要将安全性测试与自动化代码生成相结合。

---

[100] 王一明, 等. "PrivateLoRA：高效隐私保护的大语言模型." arXiv预印本 arXiv:2311.14030 (2023). 查阅日期：2024年3月28日. [在线] 可用: https://arxiv.org/pdf/2311.14030.pdf

[101] Yin, Wangsong, 等 . “移动设备上的大语言模型作为系统服务” . arXiv预印本 arXiv:2403.11805 (2024) . 访问日期：2024年3月28日 . [在线] 可用：https://arxiv.org/pdf/2403.11805.pdf

[102] Yi, Rongjie, 等. "Edgemoe: 基于Mixture-of-Experts的大型语言模型的快速设备端推理." arXiv预印本 arXiv:2308.14352 (2023). 查阅日期: 2024年3月28日. [在线] 可用: https://arxiv.org/pdf/2308.14352.pdf

[103] 朱博, 毛宁, 郭杰, 和 瓦格纳戴维, “生成式人工智能安全：挑战与对策.” arXiv, 2024年2月19日. 查阅日期：2024年3月28日. [在线]. 可用: http://arxiv.org/abs/2402.12617


TII
Technology
Innovation
Institute

## Conclusion

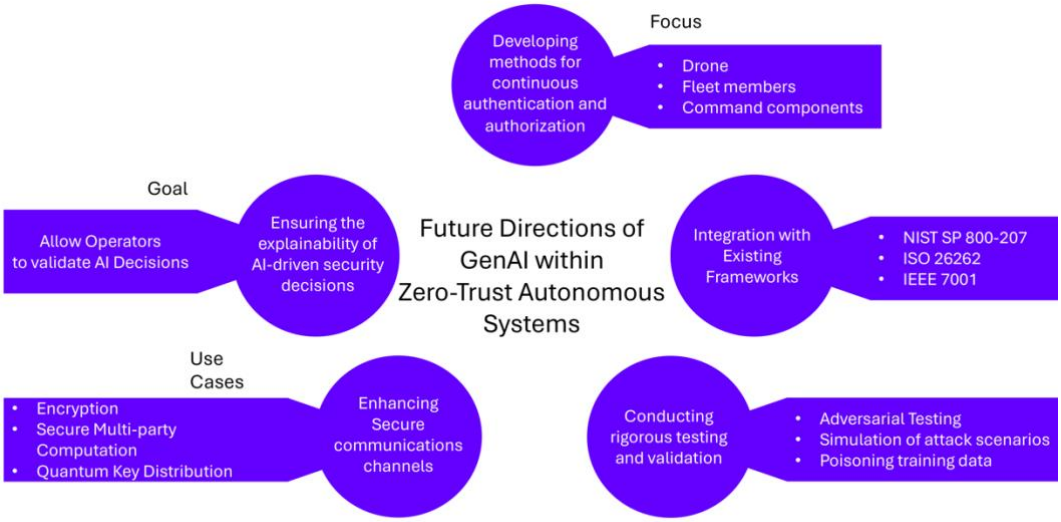
整合生成式AI和零信任架构有望在如何保护和管理无人机、机群以及人类协作方面带来范式的转变。生成式AI可以在无人机生态系统中发挥关键作用，通过提供智能、适应性和上下文感知的安全、信任和控制机制来支持零信任架构。

在个体无人机层面，生成式人工智能（GenAI）可以提升自我意识、异常检测、自主性、预测性维护和故障管理。在无人机机群的背景下，生成式AI能够帮助合成和提炼来自一群或多群无人机的情报，优化任务在机群中的安全分配，并增强可靠且高效的信任通信网络。此外，大型语言模型（LLMs）可以显著减少消息交换的数量，从而增强无人机之间的协作决策能力。在人类层面，GenAI可以帮助实施零信任原则，作为任务规划与执行、人机协同作业的一部分，并增强对机器人将执行预期目标的信任。

虽然生成式AI为无人机舰队、单个无人机以及整个自主系统带来了优势，但依然存在重大挑战。这些挑战包括平衡零信任系统的集成以确保安全、管理高昂的操作成本以及应对计算需求。伦理和监管问题、模型准确性、隐私保护以及在边缘设备上部署先进模型的技术难度也构成了障碍。此外，新的安全问题如提示注入和数据泄露不断出现，需要持续的研究与开发以确保生成式AI在无人机应用中安全、高效且隐私保护地使用。

因此，旨在将零信任原则与生成式AI相结合的未来研究可以通过探索以下领域得到显著丰富：

1. 开发稳健且高效的无人机、机群成员以及指挥控制组件连续认证和授权方法，考虑无人机环境中的资源约束和动态特性。2. 确保由人工智能驱动的安全决策的可解释性、透明性和可审计性，使人类操作员能够理解并验证访问控制策略和信任评估背后的推理过程。3。 建立无人机、蜂群和指挥控制系统的安全可靠通信渠道，以交换与安全相关的数据和政策，充分利用加密技术、安全多方计算和量子密钥分发等技术。4. 整合零信任原则与现有安全框架、标准和最佳实践（如NIST SP 800-207），以确保互操作性、一致性和符合行业法规要求。5。 开展对AI驱动的零信任解决方案进行严格的测试、验证和验证工作，包括对抗性测试和复杂攻击场景的模拟，以识别并减轻潜在的漏洞和失效模式。




通过集中开展研究、开发和合作来应对这些挑战，我们能够充分发挥生成式人工智能（GenAI）在无人机安全与韧性方面支持零信任架构的潜力。成功整合这些技术有望创造一个更为安全、适应性强且韧性的无人机生态系统，在这个生态系统中，信任将不断得到获取和验证，并有效缓解网络威胁和运营中断的风险。

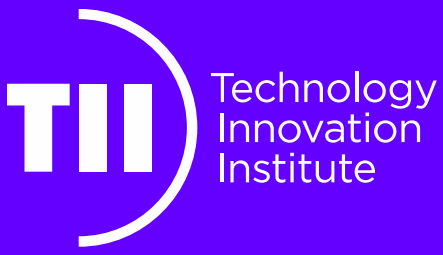
朝着这个令人兴奋的方向前进，我们必须从整体的、多学科的和以伦理为导向的角度来考虑AI驱动的零信任解决方案的发展和部署。

通过汇集来自人工智能、网络安全、机器人技术和政策领域的专业知识，我们可以确保生成式AI和大语言模型的好处以负责任、透明和可问责的方式实现，促进无人机在各种应用范围内的安全和有益使用。通过持续创新、合作，并致力于零信任原则，我们可以构建一个未来，在这个未来中，基于AI的无人机作为社会福祉的值得信赖、稳健且不可或缺的工具发挥作用。

# 词汇表

| 缩写词 | 完整形式 | 定义 |
| --- | --- | --- |
| BERT | 双向编码器表示<br>从变形金刚 | 一种基于变压器的 NLP 预训练技术 |
| C4 | 指挥、控制、通信和<br>计算机 | 军事系统管理和进行战争。 |
| CGAN | 条件生成对抗<br>Networks | GAN 以以下附加信息为条件<br>生成过程。 |
| COA | 行动课程 | 军事战略或方法<br>完成一项任务。 |
| FPGA | 现场可编程门阵列 | 集成电路设计为在<br>制造。 |
| GANs | 生成对抗网络 | 两个神经网络竞争的 AI 算法<br>在零和博弈框架中彼此。 |
| GDM | 生成扩散模型 | 通过向噪声添加结构来生成数据的模型<br>随着时间的推移。 |
| GenAI | 生成人工智能 | 一种可以生成新内容、数据或<br>基于学习模式的信息。 |
| GFM | 生成流模型 | 使用标准化流生成新的 AI 模型<br>数据样本。 |
| GNoM | 基于图神经网络的多语言<br>文本分类框架 | 将 GNN 与语言模型相结合的系统<br>多语言社交媒体分析。 |
| HMT | 人机成组 | 人类与人类之间的协作互动<br>完成任务的机器。 |
| IDS | 入侵检测系统 | 用于监控网络 / 系统的设备或软件<br>恶意活动。 |
| IMU | 惯性测量单元 | 测量车辆速度的装置 ，<br>方向和重力。 |
| LLMs | 大型语言模型 | 在广泛的文本数据集上训练的 AI 模型<br>生成类似人类的文本。 |
| LMPC | 语言模型预测控制 | 用于调整语言模型以指导的框架<br>自治系统行为。 |
| LSTM | 长短期记忆 | 一种能够<br>学习长期依赖。 |
| MAPE | 监控、分析、计划和执行 | 动态安全管理框架<br>环境。 |
| MEC | 移动边缘计算 | IT 服务的网络架构概念<br>蜂窝网络边缘的环境。 |
| NLP | 自然语言处理 | 用于解释、操纵和<br>理解人类语言 |

| 缩写词 | 完整形式 | 定义 |
| --- | --- | --- |
| NIST SP 800-207 | 国家标准与技术特别出版物 800 - 207 | 关于零信任体系结构的出版物。 |
| SBIR | 小企业创新研究 | 一项鼓励小企业研发的计划商业化的潜力。 |
| SML | 小语言模型 | 针对特定任务的更易于管理的 LLM 版本。 |
| SRU | 简单递归单元 | 用于更有效数据的神经网络变体处理。 |
| VAEs | 变分自动编码器 | 将输入编码为表示的模型 ，然后从这些重建输入。 |
| VRNN | 变分递归神经网络 | 将 VAE 与 RNN 相结合的模型 (递归神经网络) ，通常用于顺序数据。 |
| WGAN | Wasserstein 生成对抗网络 | GAN 使用 Wasserstein 距离获得更好的样本生成的稳定性和质量。 |




tii. ae