白皮书关于

对网络安全领域可解释人工智能的全面调查

由N Eswari Devi女士、N Subramanian博士、N Sarat Chandra Babu博士编制

电子交易与安全学会（SETS）

在印度政府首席科学顾问办公室之下

MGR知识城，CIT校园，特拉马尼，钦奈 – 600 113

1. 引言

人工智能已成为实现全球各行业可持续性目标的关键技术。其应用在日常场景中显而易见，如医疗保健、农业、教育、金融服务、面部识别、欺诈检测、零售购买预测、导航、在线客户支持以及许多其他领域。

由于人工智能的广泛影响，众多国际和国家级别的倡议已被启动，旨在应对其发展和监管。然而，人工智能是一把双刃剑，在网络安全领域既是变革者，也是潜在风险。

人工智能在网络安全中提升了对威胁的识别和响应，无需人工干预。然而，人工智能在各领域关键应用中的广泛应用使得AI模型容易受到攻击。因此，保护AI系统和确保数据安全和用户隐私是至关重要的。

1.1 背景 □i Motiva□ie

认识到人工智能在网络安全中的重要性以及确保人工智能系统的必要性，印度政府首席科学顾问办公室（PSA）通过SETS发布了题为“CybSec4AI”的课题组报告。该报告强调了人工智能和网络安全交叉学科中必要的研发倡议，以及专业技能的培养。目标是实现开发基于人工智能/机器学习（AI/ML）的工具和系统在网络安全方面的自给自足，同时创建安全的人工智能/机器学习系统和解决方案。

该工作组审议了在人工智能和网络安全等跨学科领域启动研发方案的需要，这些方案将导致系统发展、解决方案和产品的产生。同时，还提出了为数据创建、模拟和测试设施所需的研发基础设施。

人力资源需求已确定，以应对该新兴领域中工程师和科学家不断增长的需求，并提出了加速技能建设的方案。

根据麦肯锡全球人工智能（AI）2021调查，由于人工智能对盈利能力和成本节约的日益显著影响，越来越多的组织正在采用人工智能能力[1]。

可解释人工智能（XAI）[1] 帮助人类用户理解、解释和信任由AI模型做出的预测。通过提供AI模型决策的依据，XAI确保决策过程的透明性。这种提高的透明度使用户对结果更加有信心，并促进信任的形成。

[1] “解释”一词在牛津高阶英汉双解词典中的定义为：“一种使某事物变得清晰的说法或陈述；为行动或信仰提供的理由或辩解”[2]。

理解模型的决定也解决了公平性问题，并在调试模型方面提供帮助。

该研究于2004年开始，重大突破始于2014年左右，当时DARPA宣布了其XAI项目。根据DARPA [3]，XAI的目标是“在保持高水平的学习性能（预测准确性）的同时，生成更多可解释的模型；并使人类用户能够理解、适当信任并有效管理新兴的人工智能伙伴”。

1.2 目的和范围

这是一篇方法论文。 专注于在国内外层面进行的“可解释人工智能（XAI）”在网络安全领域的相关研究成果。

1.3 论文结构

该研究方法论文的结构如下：XAI基础部分提供了可解释人工智能的概念、技术和在人工智能模型中可解释性的重要性概述。XAI在网络安全部分探讨了XAI在网络安全任务中的应用和好处。相关研究回顾部分总结了该领域现有的研究，突出与网络安全中的XAI相关的方法和发现，同时纳入了国家与国际的努力。随后，XAI在网络安全中的挑战部分讨论了在网络安全系统中实施XAI的局限性以及潜在的对立威胁。最后，该研究方法论文以关于未来方向的见解作结，提出了XAI在网络安全方面的研究方向和问题。

2. 网络安全中的可解释人工智能（XAI）

2.1 人工智能在网络安全中的作用

人工智能通过提升威胁检测、预防和响应能力，显著增强了网络安全。根据2024年2月发布的NASSCOM报告，印度的人工智能市场正以25-35%的复合年增长率增长，预计到2027年将达到约170亿美元。[4]

人工智能系统通过分析大量数据以识别异常、检测恶意软件和预测潜在的网络安全威胁。它们自动化威胁情报收集和事件响应，使快速有效的应对措施成为可能。人工智能还优化了扫描和修补流程，增强了漏洞管理，并强化了用户行为分析，以检测内部威胁和欺诈。

此外，将人工智能集成到网络安全框架中不仅加强了防御机制，还增强了预见和应对复杂网络攻击的能力，确保了更加坚固和有弹性的数字环境。

2.2 XAI在网络安全中的必要性

可解释人工智能已成为安全领域的关键，因为它增强了透明度、信任和问责制。XAI在网络安全中的关键性之一在于其促进AI系统决策过程的透明度和信任。它帮助网络安全专业人员理解为什么一个AI模型将某个活动标记为恶意或良性。通过使AI系统的运作透明化，XAI允许持续优化和改进。

XAI提供对模型决策过程的洞察，使得能够识别和纠正偏差。这确保了网络安全措施是公平和无偏见的，维护了安全协议的完整性。

统计数据表明，与2021年同期相比，2022年第三季度全球攻击增加了28%[5]。从三个网络安全利益相关者的角度出发，可解释方法的使用包括：1）设计者，2）模型用户，3）对手。他们的工作彻底审视了各种传统和安全特定的解释方法，并探索了有趣的研究方向[6]。

Gautam Srivastava [7] 专注于XAI在特定技术领域的应用，如智能医疗、智能银行、智能农业、智能城市、智能治理、智能交通、工业4.0以及5G及以后的技术。

一项关于XAI在网络安全领域的简短调查列出了几个支持可解释性实施的XAI工具包和库[8,9]。通过对244篇文献的详尽综述，突出了使用深度学习技术（如入侵检测系统、恶意软件检测、钓鱼和垃圾邮件检测、僵尸网络检测、欺诈检测、零日漏洞、数字取证和加密挖矿）在网络安全领域的各种应用。他们的调查还考察了这些方法中可解释性的当前使用情况，确定了有前途的工作和挑战，并提供了未来的研究方向。它强调了需要更多的形式化，强调了人机交互评估的重要性，以及对抗性攻击对XAI的影响[10]。

X_SPAM方法结合了机器学习技术随机森林和深度学习技术LSTM来检测垃圾邮件，利用可解释人工智能技术LIME通过解释分类决策来增强可信度[11]。

关于将XAI应用在网络安全领域划分为三组的研究，例如：针对网络攻击的防御性应用、各行业在网络安全方面的潜力、针对XAI应用的网络安全对抗威胁以及防御方法。他们还强调了在网络安全领域实施XAI面临的挑战，并强调了标准化可解释性评估指标的重要性 [12]。

在聚焦于网络安全中的可解释人工智能（XAI）应用，特别是在公平性、完整性、隐私和健壮性方面时，很明显，现实场景通常是

被忽视。此外，当前用于防御XAI方法的对策是有限的[13]。

XAI在网络安全威胁情报（CTI）中的应用跨越了三个主要主题：钓鱼分析、攻击向量分析和网络安全防御开发[14]。另一项研究[15]讨论了现有方法在安全日志分析等应用中的优势和担忧，并提出了一个设计可解释和隐私保护的系统监控工具的流程。

XAI在网络安全应用中变得越来越重要，因为缺乏可解释性会损害对AI预测的信任。除了可解释性之外，用于网络安全的AI模型还必须保证准确性和性能。

训练数据集在任何机器学习应用中都起着至关重要的作用。可解释AI（XAI）有助于检测高度不平衡的数据集和纠正偏差，从而提高系统的鲁棒性。可解释安全（XSec）[16]的概念解决了XAI系统的安全问题，提供了一个关于如何确保这些系统安全的全面回顾。

侧信道分析（SCA）通过分析物理发射如功耗、电磁泄露或定时信息，从加密设备中提取机密信息。AI算法在增强SCA轮廓方面发挥着重要作用。AI算法用于识别大型数据集中的模式，并用于识别侧信道发射与机密信息之间的微妙相关性[17,18]。

在人工智能辅助的侧信道攻击（SCA）中，特征/兴趣点（PoI）对秘密信息的检索起到了贡献作用。对那些对决策贡献最大的特征/兴趣点（PoI）的可解释性进行了解释[19]。这有助于通过识别潜在的脆弱泄露点和实施适当的对策来验证设计对侧信道攻击的抵抗性。

一种称为真理表深度卷积神经网络（TT-DCNN）的可解释神经网络，其内部结构具有可解释性，被用于执行SCA。通过将神经网络（NN）转换为SAT方程，实现了互操作性，以了解NN模型学到了什么[20]。

掩码和隐藏是用于在实现层面保护秘密信息的对策。但是，AI算法在击败这些对策方面是有效的。在SCA（侧信道攻击）中，ExDL-SCA（基于深度学习的SCA的可解释性）方法被用来理解这些对策对AI辅助SCA的影响[21]。这有助于开发者评估实施对策的安全性。此外，XAI（可解释人工智能）在硬件木马检测[81]中也发挥着重要作用。图1描述了XAI在网络安全各个领域的应用。

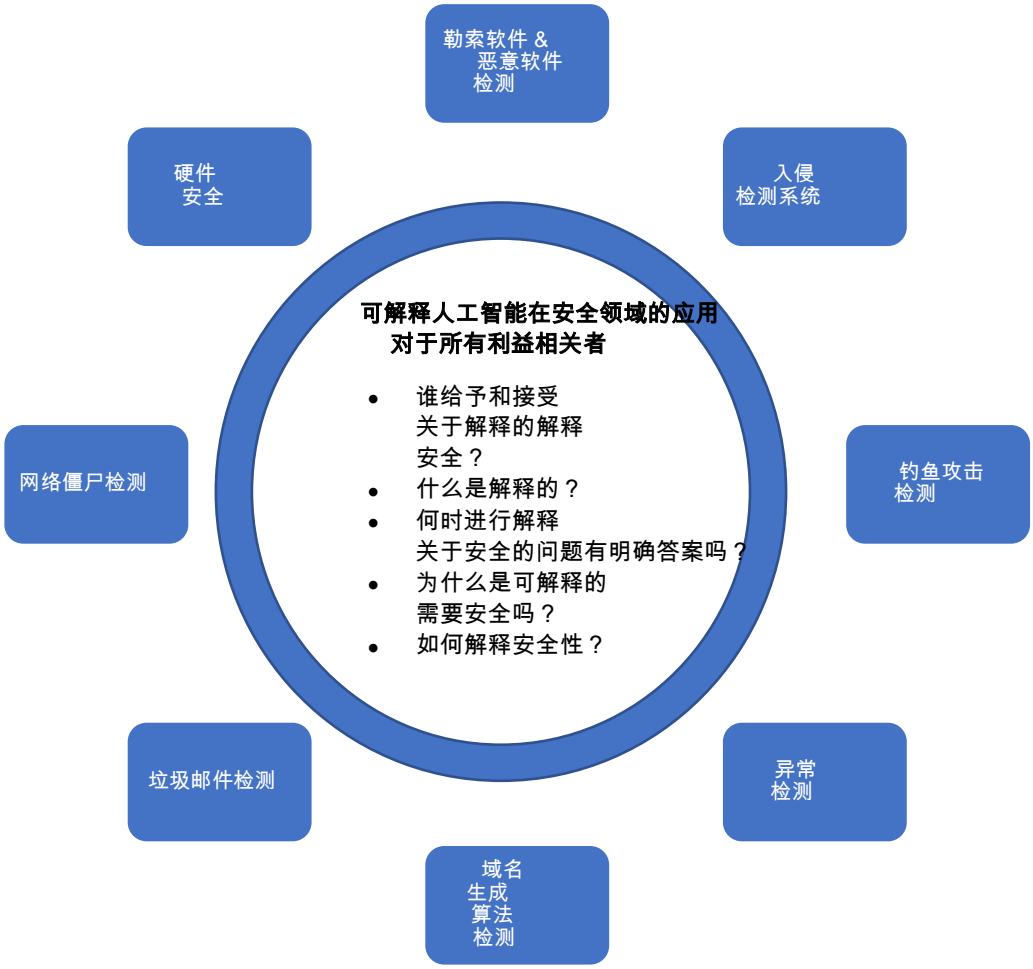

图1：可解释人工智能在网络安全中的应用

3. 相关工作综述

3.1 国际努力

2017年，美国国防部高级研究计划局（DARPA）信息创新办公室（I2O）主任约翰·兰斯伯里讨论了“人工智能的三个浪潮”，以消除对这项技术的神秘感。第一波是“手工知识”，涉及将特定领域的知识编码成计算机遵循的规则。第二波是“统计学习”，使用在特定数据上训练的统计模型。第三波是“情境适应”，特点是能够理解和解释其决策背后的推理的系统[22]。

DARPA的XAI项目专注于两大挑战：（1）解决机器学习问题，以对异构、多媒体数据中的感兴趣事件进行分类；（2）开发机器学习方法，为自主系统构建决策政策，以执行各种模拟任务[23]。

2020年，IBM推出了“政策实验室”，这是一个旨在开发AI政策和建议的平台，为政策制定者提供了一个愿景和实用指导，以利用创新的优势，同时确保在快速变化的技术环境中保持信任[24]。

2022年，IBM商业价值研究院发布了一项关于人工智能伦理的实际应用研究。该研究指出，构建可信的人工智能被视为一种战略差异化，并且组织开始实施人工智能伦理机制。研究建议解决如隐私、稳健性、公平性、可解释性、透明度以及其他相关原则，以建立道德人工智能实施的治理方法[25]。

2020年，谷歌发布了一份名为《AI Explainability》的白皮书，这是一份与谷歌云人工智能解释产品相伴的技术参考。该白皮书旨在利用人工智能解释来简化模型开发，并向关键利益相关者解释模型的行为。

“全球人工智能行动计划”项目，作为世界经济论坛“塑造技术治理未来：人工智能与机器学习平台”的一部分，旨在通过加快全球范围内可信、透明和包容性人工智能系统的采用，激发人工智能的变革潜力。

到2026年，Gartner预计组织将通过实施AI的透明度、信任和安全，在采用、业务目标和用户接受度方面实现其AI模型50%的改进[26]。Gartner指出，技术服务提供商越来越在其模型中使用可解释的AI，尤其是在医疗保健和金融服务等安全和监管行业[26,27]。

人工智能的隐私、安全和/或风险管理始于人工智能可解释性，这是必需的基线。 —— Avivah Litan，Gartner副总裁兼杰出分析师

2019年，英国皇家学会发布了《可解释人工智能：基础政策简报》，总结了在实施可解释人工智能系统时，开发者和政策制定者面临的挑战和考虑因素。它强调，人工智能模型的解释取决于其应用，并提供了关于如何在不影响系统性能（包括准确性、可解释性和隐私性）的情况下实施XAI的观点。[28]

在2022年9月，美国国家标准与技术研究院（NIST）发布了一份关于人工智能风险管理框架（AI RMF）的草案，旨在发展和实施可信的人工智能，其中可解释性是识别和管理人工智能系统风险的综合方法中需要考虑的特性之一[29]。

2022年5月，IBM通过“IBM全球AI采纳指数2022”[30]提供了关于全球整体AI采纳情况的见解，包括阻碍AI发挥其潜力的障碍和挑战，以及AI最有可能繁荣的应用场景、行业和国家。

图2中的统计数据显示，大多数组织尚未采取必要步骤以实现可信的人工智能。具体来说，61%接受调查的组织尚未作出显著的努力。

努力解释基于AI的决策。安全专业人士是组织中第四大AI用户群体，占比26%。此外，29%的组织使用AI进行安全和威胁检测。

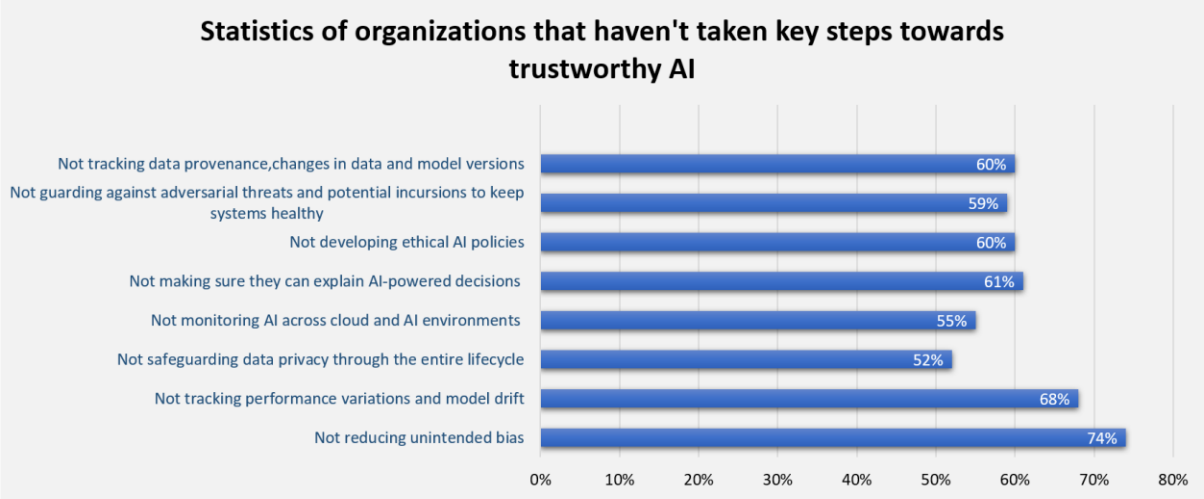

图2：根据IBM全球人工智能，未采取关键步骤迈向可信人工智能的组织统计数据采用指数[30]

此外，84%的IT专业人员承认信任在人工智能中的重要性，并认为解释人工智能决策对其业务至关重要。目前，17%的IT专业人员更倾向于报告他们的业务重视人工智能可解释性，而非仅探索人工智能的群体。此外，公司在开发可信赖且可解释的人工智能方面面临着多个障碍。值得注意的是，63%的公司因缺乏开发和管理工作可信赖人工智能所需的技术和培训而面临困难。

IT部门在政府与医疗行业的专业人员，目前正探索或部署人工智能，相较于其他行业的人员，他们更有可能识别出可解释性与信任方面的障碍。

3.2 国家努力

作为全球增长最快的经济体之一，印度对重塑世界的AI革命充满浓厚兴趣。认识到AI推动经济转型的潜力以及印度需要在这次转型中战略定位自身的必要性，政府已经开始制定国家AI战略。

在国家层面，印度国家发展研究所（NITI Aayog）、信息技术与通信部（MeitY）、商务部与工业部以及印度政府首席科学顾问办公室已经成立了由领域专家组成的各个工作组，以制定本战略。这些工作组已经起草了聚焦于人工智能的国家级文件。

2018年6月，印度国家发展委员会（NITI Aayog）发布了一份题为“全国人工智能战略#AIForAll”的讨论稿[31]。该文件旨在指导研究和

在新兴和前沿技术的进步中，特别关注利用人工智能促进印度社会的包容性增长。本文还强调了在人工智能中“可解释性”的重要性，解决了一个普遍问题，即人工智能解决方案通常作为黑箱运行，对于超出输入和输出数据的处理过程缺乏理解。

2018年2月，印度电子和信息技术部（MeitY）成立了四个委员会，以促进人工智能（AI）的发展并制定政策框架。这些委员会负责了解与人工智能相关的监管和技术挑战，并确定人工智能实施可能有益的领域[32]。

在2018年3月，由商务部和工业部成立的特别小组发布了一份关于《人工智能（AI）》的报告，该报告专注于第四次工业革命及其经济影响，特别是关于人工智能的方面。[33]。

2020年7月，首席科学顾问办公室通过电子交易和安全协会（SETS）发布了一份名为“CybSec4AI”的任务小组报告，强调了人工智能和网络安全相互作用的尖端研究。这份报告强调了AI模型安全性的重要性[34]。

2020年10月，印度电子和信息技术部（MeitY）与NITI Aayog共同举办了RAISE 2020——“2020年负责任的人工智能促进社会赋权”大型虚拟峰会，主题为人工智能（AI）[35]。峰会强调了可解释人工智能对于培养对技术的信任至关重要，并强调了更好地理解人工智能决策以促进其广泛采用的重要性。讨论还涵盖了人工智能缺乏可解释性的影响，并提出了技术政策和政策保障措施来减轻这些问题。可解释人工智能是负责任人工智能更广泛范围内的一个关键组成部分，定期发布的负责任人工智能（RAI）专家报告[36]。

在2020年，印度电子与信息技术部下的国家电子治理司推出了名为“YUVAi - 用人工智能为乌纳蒂和维卡斯（Unnati和Vikas）贡献力量”的创新挑战计划，这是面向学校学生的国家倡议，旨在“负责任的人工智能”领域。该计划旨在提升新一代的数字准备能力，并继续实施包容性和协作性的人工智能技能培训努力[37]。在此基础上，英特尔印度分公司于2022年加入并启动了“2022年负责任的人工智能”计划[38]。

2021年2月，印度国家发展机构（NITI Aayog）发布了《负责任的AI #AIFORALL》，这是为印度制定的方法论文档的第一部分，提出了负责任AI管理的原则。该文档探讨了系统性和社会性的伦理考量，并深入探讨了治理AI系统的法律和监管方法。透明度被强调为确保公平、诚实、无偏见的部署和问责制的一个关键原则。此外，该文档还讨论了XAI [39] 的发展。

2021年8月，尼蒂·阿亚og在《负责任的人工智能》方法文档的第二部分中继续其努力，重点关注实施这些原则[40]。随后，2022年11月，尼蒂·阿亚og发布了第三份讨论稿，题为“面向所有人的负责任人工智能——采用框架：面部识别技术使用案例方法”。该论文探讨了在印度背景下，负责任人工智能原则在面部识别技术（FRT）中的应用，旨在为其安全、负责任的发展和部署建立框架。该论文还建议了开发人员构建可解释的FRT系统应遵循的原则，确保对系统决策过程的明确解释[41]。

2022年10月11日，印度软件和服务公司协会（NASSCOM）与微软、塔塔咨询服务公司（TCS）、IBM研究、德勤和Fractal Analytics合作发布了“负责任人工智能中心及资源包”。该倡议旨在推广负责任的人工智能实践并促进其广泛应用。资源包包括适用于各行业的工具和指导，帮助企业自信地扩大人工智能技术规模，同时确保用户安全和信任。在发布活动中，一个关键焦点是解决XAI（可解释人工智能）的挑战。[42, 43]

2022年，NASSCOM与微软、Capgemini、Ernst & Young和EXL合作推出了针对印度的NASSCOM人工智能采纳指数。这一全面框架评估了全国及不同行业的人工智能采纳成熟度，提供了一个综合得分。尽管印度在全球人工智能投资中只占很小的比例（约1.5%），但该国在这一变革性技术方面取得了重大进展。为了充分挖掘这一潜力，采用最佳实践对于将技术进步转化为实际的国民价值至关重要[44]。

2020年6月，印度成为经济合作与发展组织（OECD）\"全球人工智能伙伴关系（GPAI）\"的创始成员。该倡议旨在指导全球人工智能的负责任发展和应用，强调人权、包容性、多样性、创新和经济增长[45, 46]。

在2022年12月，谷歌宣布了一笔100万美元的赠款投资，以在NITI Aayog的合作下，在印度建立首个多学科责任AI中心——IIT-Madras[47]。

全球人工智能伙伴关系（GPAI）峰会于2023年12月在新德里举行，汇聚了来自科学、工业、民间社会、政府、国际组织和学术界的专家。峰会旨在促进国际社会在人工智能相关优先事项上的合作。

在2023-2024年度的联邦预算中，宣布建立三个卓越中心（COEs）专注于人工智能，其愿景为“打造印度人工智能，让人工智能为印度服务”。

国家级努力产生了关于人工智能的几份报告，突出了其在数字印度框架下的潜力。NITI Aayog 报告侧重于人工智能在医疗保健、农业、教育和金融等多个领域的应用。同时，由印度政府首席科学顾问办公室下的电子交易和安全协会（SETS）发起的 CybSec4AI 报告旨在培养开发基于人工智能的工具和系统的自力更生能力。它强调了创建安全的AI解决方案和系统。

目前，在网络安全领域，尽管该领域正在进行研究，但仍缺乏关于可解释人工智能（XAI）的精确国家标准、政策或框架。实施此类标准将揭示复杂数据模型的工作原理，增强对AI决策的信任，并推动负责任的人工智能模型的发展。

可解释人工智能是负责任人工智能（RAI）的一个关键方面，它涵盖了更广泛的原则，如公平性、无偏见、透明度、隐私、安全、可靠性、安全性、合规性、保护以及积极价值的强化。虽然使用道德黑客测试系统是有益的，但旨在破坏系统的恶意攻击是不受欢迎的。

3.3XAI在网络安全中应用的方法论

本节提供了可解释人工智能（XAI）在网络安全领域的应用调查，涵盖了网络安全、异常检测和微架构攻击。

入侵检测系统（IDSs）已成为计算机网络中确保网络安全环境的关键工具。近年来，IDSs利用了各种分类算法，如决策树、支持向量机（SVM）、K-最近邻（KNN）、朴素贝叶斯分类器、深度神经网络（DNN）、卷积神经网络（CNN）、循环神经网络（RNN）、自编码器等，以检测入侵。对检测到的入侵的预测提供解释至关重要，这有助于理解不同类型攻击的具体特征。

实现了用于执行入侵检测的两阶段模型[48]。在第一阶段，XGBoost（eXtreme Gradient Boost）模型结合SHAP解释框架，为第一阶段监督学习模型的结果提供解释。在第二阶段，从第一阶段获得的解释被用于训练自动编码器，以使模型能够对零日攻击或未见过攻击取得良好的效果。

模型被入侵检测系统（IDS）错误分类的原因在[49]中得到了解释。这些关于错误分类原因的解释有助于决定未来对抗攻击的步骤。对抗机器学习技术被用于生成由训练分类器做出的错误估计的解释。对抗方法涉及修改错误分类的样本，直到模型

将正确的类别分配。修改后的样本与真实样本之间的差异被用来说明导致错误分类的主要特征。

在网络安全方面，数据驱动模型的误报输出可能导致整个系统的泄露和损害。鉴于这一点，为在系统中实现准确性和可靠性，提出了一个被称为混合算子-解释器入侵检测系统的模型。所提出的模型使用两个独立的模块来提供关于系统决策的易于理解答案，同时达到尽可能高的准确性 [50]。

本研究的SHAP框架本地解释被用于提供对每个特征影响的详细信息，以帮助对入侵检测系统（IDS）进行决策。全局解释提取了重要的特征，并探索了特征值与特定类型攻击之间的关系[51]。

XAI技术被用于提升网络安全背景下脚手架攻击的检测与防御[52]。

在文献[53]中提出了一种在线和离线反馈机制，该机制为用户提供了对学习模型决策产生影响的最为相关的输入特征。

深度学习技术在安全应用中得到了应用，例如恶意软件分类、二进制逆向工程，这些应用在决策过程中缺乏透明度。大多数模型，如LIME，假设决策边界是局部线性的。然而，在大多数复杂的安全问题中，这导致了不准确的解释[54]。介绍了一种名为“使用非线性逼近的局部解释方法”（LEMNA）的模型，该模型包括混合回归模型，用于考虑线性和非线性决策边界，以提高复杂安全应用的局部解释保真度。融合Lasso技术被集成以捕捉特征依赖性。所提出的LEMNA模型被证明有助于通过解释分类器如何做出正确决策来建立TRUST。

LEMNA用于解释基于异常的入侵检测系统（IDS）[55]的输出。通过考虑有助于预测的关键特征，推导出网络访问控制策略。关键特征的选择基于从给定的输入推导出的每个特征对预测贡献的得分。

基于决策树的自动编码器模型在[56]中被提出，用于检测异常并提供解释，该解释通过计算不同属性值之间的相关性来实现，对模型做出的预测进行解释。

提出了一种使用深度神经网络（DNN）进行异常检测的框架[57]。层级相关性传播（LRP）用于分解DNN复合函数，以计算输入特征的关联分数，从而指示每个特征对检测异常的贡献。该系统设计得能够提供预测的置信度分数，并对检测到的异常提供文本描述。

随着每年钓鱼攻击的增多，LIME和EBM等可解释框架被用于将网址分类为钓鱼或合法，并提供相应的解释[58]。

LIME和显著性图被用来解释AI模型在基于微架构的网站指纹攻击决策中的决策过程[59]。

3.4 可解释机器学习框架

可解释性指的是理解输入与输出之间关系的能力。另一方面，可解释性涉及以人类可以理解的方式阐明模型的输出。模型的可解释性一般分为两类：全局可解释性和局部可解释性。全局可解释性使用户能够理解目标结果的分布与特征之间的关系。同时，局部可解释性为特定输入数据点上的预测提供解释。本节重点介绍了几个关键框架，这些框架提供了对模型预测的洞察。

3.4.1 LIME：局部可解释模型无关解释

本地代理模型是可解释模型，用于解释机器学习黑盒模型的个别预测[60]。当输入数据被修改并作为机器学习模型的输入时，会测试预测的变化。这种变化很小，因此仍然接近原始数据点。LIME创建了一个新的数据集，包括扰动样本及其对应黑盒模型的预测。然后，它在这个数据集上训练一个可解释模型，权重根据样本实例与感兴趣实例的邻近性分配。

3.4.2 SHAP（SHapley Additive exPlanation）值：

Shapley加性解释（SHAP）是一个统一的框架，通过Shapley值来解释机器学习模型的预测结果[61]。Shapley值基于博弈论和局部解释思想的统一。Shapley值表示相关特征对预测的影响。Kernel Shap，一种从线性LIME扩展和适应的方法，被引入来计算Shapley值，以避免Shapley值对于许多特征变得难以处理[60]。

其他变体被提出以获得基于不同类型模型的Shapley值，例如树SHAP、深度SHAP、低阶SHAP、线性SHAP和最大SHAP [61]。在两个数据集上使用深度学习模型，对两种主要解释器/可解释方法LIME和SHAP进行了实证评估。结果表明，在身份、稳定性和可分离性方面，SHAP的表现略优于LIME [62]。

SHAP产生与人类解释相一致的解释，随着特征的增加，其计算成本较高。随着联盟数量的增长，SHAP的计算工作量呈指数级增加。而LIME的计算成本相对较低，尽管具有更高的特征数量。LIME和SHAP模型是替代模型。它们通过改变输入来建模预测的变化。它们可以……

理解到，如果通过改变变量的值，模型预测变化不大，那么感兴趣的变量在预测中不起主要作用。

3.4.3 锚定

锚算法，由提出LIME算法的同一研究者团队在[63]中引入。它是LIME的扩展，通过被称为锚的高精度规则来解释机器学习模型的本地行为，这些锚代表预测的“充分”条件。锚解决了局部解释方法LIME的不足，LIME通过线性方式代理模型的本地行为。

锚定器结合了强化学习技术和图搜索算法，以最小化模型调用次数和所需的运行时间，同时从局部最优解中恢复。与LIME的代理模型不同，得出的解释以易于理解的IF-THEN规则表达，被称为锚定器。

3.4.4 简化解释（Explain Like I'm 5）

ELI5 [64] 是一个Python包，旨在解释黑盒机器学习（各种回归和分类模型）模型。ELI5类似于LIME（但ELI5不是模型无关的），它提供了与每个特征相关的权重，以描述该特征在机器学习模型中的重要性。ELI5已经实现了大多数常用的基于Python的机器学习包，如Scikit-learn、Keras和XGBoost。

3.4.5 滑板运动员

滑板者[65]是另一个能够揭示模型内部决策策略的Python包。相关开源软件工具有助于探索和理解机器学习模型的行为，以及描述神秘且不明确的机器学习模型[66]。

3.4.6 基于本地规则的解释（LORE）

这是一个用于解释黑盒实例的框架。该模型采用决策树进行局部解释的提取。LORE是一种基于局部、事后和模型无关的方法[67]。

4. XAI在网络安全领域的挑战

皇家学会[68]的报告正确地指出了人工智能中可解释性的局限性。可解释性有助于提高人们对人工智能模型的信任，但并不有助于创建产生可信输出的系统。当系统提供令人信服但具有欺骗性的解释时，用户也可能产生错误的自信感。人工智能模型解释的准确性必须与人工智能模型输出的准确性同等重要。当解释的准确性不高时，这为对手提供了机会，在输入的小扰动上操纵分类器的输出，以隐藏系统的偏差。

提供100％准确的解释以增强AI黑盒的威胁对模型安全。攻击者可以使用提供的解释作为输入来克隆AI黑盒。这影响了原始AI模型的所有权。因此，应谨慎行事。

因此，仅解释性本身并不能成为万能的解决方案。[70]

4.1 对XAI模型构成的威胁

由于XAI模型的驱动性质，可解释性模型本身可能容易受到恶意操控。对抗攻击可以大致分为两大领域：1) 解释本身可以被修改，2) 解释可以被用来检索机密训练数据和有关模型详情。

4.1.1 对XAI模型的对抗攻击

尽管在网络安全领域中，XAI（可解释人工智能）承诺透明性和可解释性，但这些XAI模型容易受到网络攻击。在[71]中引入了一类新的攻击，称为ADV2，该攻击通过生成误导性的对抗输入来欺骗目标DNN及其耦合的解释器。在[72]中，描述了设计一种新型黑盒攻击，用于分析基于梯度的XAI方法的一致性、正确性和置信度安全性属性。提出的方法可以用于设计安全的鲁棒XAI方法，它专注于针对两类敌人的黑盒定向攻击：一是破坏基础分类器和解释器的完整性，另一类是仅攻击解释器而不改变分类器的预测。他们提出了一个与网络安全相关的分类法。提出了三种不同的方法，其中，a) X-PLAIN - 预测/数据的解释，b) XSP-PLAIN - 安全和隐私的解释，最后，c) XT-PLAIN - 对威胁模型的解释。

给定输入，可以生成与之相似的输入，两者在视觉上几乎无法区分，都被分配了相同的预测标签，但它们具有非常不同的解释[73]。因此，其显著性归因和可解释信息变得不那么可靠。此外，解释甚至可以被任意操控[74]。

在模型的可解释模块以及人工智能模型免受网络攻击，通过适当的防御方法变得同样重要。

4.1.2 对解释的对抗性使用

隐私和透明度是可信机器学习的两个基石。

提供解释与预测相结合可减轻可能损害敏感输入训练数据和模型本身[75,76]的攻击。因此，可解释机器学习本身可能面临隐私风险。使用解释从模型预测中重建训练数据的方法，通常被称为模型逆攻击，已被探讨[77]。

展示了一项实验，演示了敌方如何使用基于梯度的解释来执行成员推断攻击，以预测敌方可访问的输入数据是否为训练数据的一部分 [78,79]。同时也证明，敌方可以通过生成的解释从模型预测中提取结构和模型参数 [79]。

至关重要的是，通过仅授权实体访问来保护XAI模型免受攻击者的侵害。

4.2 未来方向

不透明的系统运行令用户不满意，因此可解释性对于确保人工智能模型的可信度至关重要。XAI对于促进人工智能的道德和公平使用也至关重要。可解释性有助于监督遵守道德原则。认识到XAI对可信性的重要性，美国国家标准与技术研究院（NIST）于2021年发布了一份题为“可解释人工智能的四个原则”的报告，概述了XAI系统的基本属性和原则：解释、意义、解释准确性和知识限制[80]。

4.2.1 对XAI模型及其适用性的研究

现有可解释人工智能模型需要对其在各类网络安全应用中的适用性进行评估。评估时，应考虑在有无XAI模型的情况下可能发生的网络攻击类型。此外，还必须考察XAI模型对实时需求和关键基础设施安全应用中结果准确性的影响。

需要一种方法来评估模型在目标领域的适用性以及验证这些模型。研究应专注于保护隐私的XAI方法，以防止对手利用解释。新开发工具应接受研究者的同行评审，以评估其可用性。同时，也有必要比较现有XAI模型在描述准确性、稀疏性和效率等方面适用于安全应用的适用性。

创建高质量数据集将对网络安全应用中XAI方法的有效性产生重大影响。研究应关注平衡XAI在网络安全中的性能和可解释性。此外，保护数据和XAI生成的解释免受攻击者的侵害以及防范对抗性攻击至关重要。

一个用于安全应用的可解释人工智能软件工具包可以通过与研发、学术界和工业界的合作开发。此工具包将指导用户在网络安全领域应用XAI，并确保用于各种应用中的XAI模型的安全性。

4.2.2 测试与验证

类似于OpenSSL，是否可能存在开放的XAI，其中所有内部状态都是可见的。对正在开发的系统/人工智能模型进行测试可以通过测试向量（可以称之为注入偏差的数据集、真实的好数据集等）进行，正如对任何系统一样。可以引入扰动来观察系统如何反应。

4.2.3 XAI策略标准化政策框架

可以提出一个适合于印度基于XAI的网络安全框架和治理模式。

指标，例如 i) 解释性的水平（少数系统可能需要较低水平的透明度，而其他系统可能需要较高的透明度，这完全基于应用），ii) 评估指标（评估可解释安全性方法）和 iii) 多方面方法（必须在系统的整个生命周期设计阶段提供解释，包括设计、实施、利用、分析和更新阶段），可以考虑。

5. 摘要

本文讨论了……的重要性 可解释人工智能（XAI）在网络安全领域的应用，强调国内外视角。XAI可以检测并减轻AI模型中的偏差，从而实现更公平、更可靠的安全措施。它通过提供对AI决策过程的清晰见解来增强透明度和信任，这对建立对人工智能驱动安全系统的信心至关重要。此外，XAI通过帮助识别误报和漏报来提高安全措施的正确性和可靠性。

参考文献

[1] https://www.mckinsey.com/business-functions/quantumblack/our-见解/全球调查/2021年人工智能的状态

[2] Doran, D., Schulz, S. 和 Besold, T.R., “可解释人工智能到底是什么意思？一项实证研究，”新的视角概念化，“arXiv预印本 arXiv:1710.00794，2017”。

[3] Gunning, D. 和 Aha, D.，“DARPA的可解释人工智能（XAI）节目，“人工智能杂志第40卷第2期（2019年）：44-58。”

[4] https://www.nasscom.in/knowledge-center/publications/ai-powered-tech-服务路线图——面向未来准备的企业

[5] https://blog.checkpoint.com/2022/10/26/third-quarter-of-2022-reveals-网络攻击增加/

[6] Nadeem, A., Vos, D., Cao, C., Pajola, L., Dieck, S., Baumgartner, R., & Verwer, S., 你: Sok: 可解释机器学习在计算机安全应用中的解释 *arXiv预印本 arXiv:2208.10605*, 2022.

[7] 斯里瓦斯塔瓦，G.，贾韦里，R.H.，巴塔查里亚，S.，潘达亚，S.，马达昆塔，P.K.R.，尹德里，G.，霍尔，J.G.，阿拉扎布，M.和加德卡拉，S.，XAI在网络安全领域的应用：现状《艺术、挑战、开放问题和未来方向》 *arXiv预印本 arXiv:2206.03585*, 2022.

[8] Rjoub, G., Bentahar, J., Wahab, O. A., Mizouni, R., Song, A., Cohen, R., Otrok, H., and Mourad, A., “关于网络安全的可解释人工智能调查”，IEEE网络与服务管理汇刊，2023.

[9] Hariharan, S., Velicheti, A., Anagha, A. S., Thomas, C. 和 Balakrishnan, N.，“网络安全中的可解释人工智能：简短回顾”，载于2021年第4届安全与隐私国际会议（ISEA-ISAP）（第1-12页），2021年。

[10] Capuano, N., Fenza, G., Loia, V. 和 Stanzione, C.，“可解释的人工智能”在网络安全：调查，“IEEE Access”，第10卷，第93575-93600页，2022年。

[11] Ibrahim, A., Mejri, M. 和 Jaafar, F.，“一种可解释的人工智能方法”“关于可信垃圾邮件检测”，IEEE国际网络安全会议安全与韧性（CSR）（第160-167页），2023年。

[12] Vigano, L., and Magazzeni, D., “可解释的安全”，收录于 *2020 IEEE 欧洲区研讨会：安全和隐私工作坊（欧洲S&PW）*（第293-300页），2020年。

[13] Charmet, F., Tanuwidjaja, H.C., Ayoubi, S., Gimenez, P.F., Han, Y., Jmila, H., Blanc, G. Takahashi, T. 和 Zhang, Z.，“网络安全中的可解释人工智能：一种综述与展望”文献综述，“电信年鉴”，第1-24页，2022年。

[14] Samtani, S., Chen, H., Kantarcioglu, M. 和 Thuraisingham, B.，“可解释的人工智能用于网络威胁情报（XAI-CTI），” *IEEE 交易在可靠与安全计算领域*, *19*（04），第2149-2150页，2022年。

[15] Bhusal, D. 和 Rastogi, N.，“SoK：在安全监控中建模可解释性”信任、隐私和可解释性，”arXiv预印本arXiv:2210.17376，2022。

[16] Vigano, L., and Magazzeni, D., “可解释的安全”，收录于 *2020 IEEE 欧洲区研讨会：安全和隐私工作坊（欧洲S&PW）*（第293-300页），2020年。

[17] Emmanuel Prouff, Rémi Strullu, Ryad Benadjila, Eleonora Cagli, and Cécile Dumas. 2018年，关于侧信道分析中深度学习技术的探究
简介：ASCAD数据库。IACR密码学电子预印本档案，2018年（2018），第53号

[18] Picek, S., Perin, G., Mariot, L., Wu, L., & Batina, L. (2023). Sok: 基于深度学习的物理侧信道分析。ACM 计算综述，55(11)，1-35。

[19] Golder, A., Bhat, A., & Raychowdhury, A. (2022, June). 探索研究神经网络模型在功率侧信道分析中的可解释性。
2022年大湖VLSI研讨会论文集（第59-64页）。

[20] Yap, T., Benamira, A., Bhasin, S., & Peyrin, T. “窥探黑盒：在侧信道分析中使用SAT方程的可解释神经网络”，《加密硬件与嵌入式系统密码学交易》，第24-53页，2023年。

[21] Perin, G., Wu, L., & Picek, S., “我知道你的层做了什么：深度学习侧信道分析的层状可解释性”，密码学电子档案，2022年。

[22] https://www.darpa.mil/about-us/darpa-perspective-on-ai

[23] https://www.darpa.mil/program/explainable-artificial-intelligence

[24] https://www.ibm.com/policy/ai-precision-regulation/

[25] https://www.ibm.com/downloads/cas/4DPJK92W

[26] https://www.gartner.com/en/articles/what-it-takes-to-make-ai-safe-and-有效

[27] https://www.gartner.com/en/documents/4020030

[28] “可解释人工智能：基础知识政策简报”，英国皇家学会，2019年（ https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and- interpretability-policy-briefing.pdf ）

[29] https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

[30] https://www.ibm.com/downloads/cas/GVAGA3JP

[31] 人工智能国家战略#AIFORALL，讨论稿，印度政府国家发展研究所，2018年6月。

[32] 网络安全、安全、法律和伦理问题，印度电子和信息技术部（MeitY），印度政府。来源：
https://meity.gov.in/artificial-intelligence-committees-reports

[33] 人工智能任务小组，印度商务部和工业部。来源： https://dipp.gov.in/whats-new/report-task-force- artificial-intelligence

[34] https://www.psa.gov.in/psa-prod/publication/Taskforce-Report-CybSec4AI - SETS.pdf

[35] https://indiaai.gov.in/raise

[36] https://www.gpai.ai/projects/responsible-ai/

[37] https://innovateindia.mygov.in/yuvai/

[38] https://responsibleaiforyouth.negd.in/home

[39] https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

[40] https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf

[41] https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf

[42] https://nasscom.in/responsible-ai/

[43] https://indiaai.gov.in/news/nasscom-launched-the-responsible-ai-hub-and-资源包

[44] https://nasscom.in/knowledge-center/publications/nasscom-ai-adoption-index

[45] https://pib.gov.in/PressReleasePage.aspx?PRID=1631676

[46] https://oecd.ai/en/wonk/an-introduction-to-the-global-partnership-on-ais-工作于负责任的AI

[47] https://www.iitm.ac.in/happenings/press-releases-and-coverages/google-印度理工学院马德拉斯人工智能研究中心获得100万英镑资助

[48] Barnard, P., Marchetti, N., 和 Da Silva, L. A.，“通过可解释人工智能（XAI）实现稳健的网络入侵检测”， *IEEE 网络信函* 2022.

[49] Marino, D. L., Wickramasinghe, C. S., and Manic, M., “用于入侵检测系统的可解释人工智能的对抗性方法，”载于2018年IEEE工业电子学会第44届年会（IECON 2018），第3237-3243页，2018年。

[50] Szczepański, M., Choraś, M., Pawlicki, M. 和 Kozik, R., \"通过混合Oracle-explainer方法实现入侵检测系统的可解释性\", 2020年国际神经网络联合会议（IJCNN），2020年。

[51] 王明，郑克，杨阳，王霞，“入侵检测系统的可解释机器学习框架”，IEEE Access，第8卷，第73 127页至第73 141页，2020年。

[52] Senevirathna, Thulitha, Bartlomiej Siniarski, Madhusanka Liyanage, and Shen 王. \"网络入侵检测中欺骗性的后验可解释人工智能（XAI）方法\ 检测.\" IEEE 第二十一届消费者通信与网络会议（CCNC），第107-112页，2024年。

[53] Amarasinghe, K. 和 Manic, M.，“提升用户对深度神经网络的信任度” 基于入侵检测系统（“based intrusion detection systems,” In IECON 2018-44th Annual Confe IEEE 工业电子学学会（第3262-3268页），2018年。

[54] 郭伟，穆迪，徐杰，苏鹏，王刚，邢欣，“Lemna：解释基于深度学习的安全应用”，2018年ACM SIGSAC计算机与通信安全会议论文集，第364-379页，2018年。

[55] 李，H.，魏，F. 和 胡H.，“通过异常检测实现动态网络访问控制”
基于IDS和SDN，“ACM国际安全工作坊论文集”
在软件定义网络和网络功能虚拟化，第1316页，2019年。

[56] Aguilar, D. L., Perez, M. A. M., Loyola-Gonzalez, O., Choo, K. K. R., and Bucheli-Susarrey, E., “Towards an interpretable autoencoder: a decision tree-based
自编码器及其在异常检测中的应用，《IEEE 交易杂志》
《可靠的和安全的计算》，2022年。

[57] Amarasinghe, K., Kenney, K. 和 Manic, M.，“迈向可解释的深度神经网络”
网络基础异常检测，”第11届国际人类
系统交互（HSI），第311-317页，2018年。

[58] Hernandes, Paulo R. Galego, Camila P. Floret, Katia F. Cardozo De Almeida, Vinícius Camargo Da Silva, João Paulo Papa, and Kelton A. Pontara Da Costa. \"Phishing
基于URL的XAI技术检测\"，IEEE系列会议
计算智能（SSCI），第 01-06 页，2021 年。

[59] Gulmezoglu, B., “基于XAI的微架构侧信道分析在网站指纹识别攻击与防御中的应用，” IEEE Transactions on Dependable and Secure Computing，2021.

[60] Ribeiro, M. T., Singh, S., and Guestrin, C., \"Why should I trust you? Explaining the predictions of any classifier,” Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144), 2016.

[61] Lundberg, S. M. 和 Lee, S. I.，“统一解释模型预测的方法，”神经网络信息处理系统进展，第30卷，2017年。

[62] Hailemariam, Y., Yazdinejad, A., Parizi, R. M., Srivastava, G. 和 Dehghantanha, A., “对AI深度可解释工具的经验评估，” In *2020 IEEE Globecom 工作坊（GC Wkshps）* (第1-6页), 2020.

[63] Ribeiro, M. T., Singh, S., 和 Guestrin, C.，“Anchor: 高精度模型无关解释”，见 *人工智能AAAI会议论文集*（第32卷，第1期），2018年。

[64] https://github.com/TeamHG-Memex/eli5

[65] https://github.com/oracle/Skater

[66] Agarwal, N., and Das, S.，“可解释机器学习工具：综述”，收录于 *2020 IEEE 计算智能系列研讨会（SSCI）* 第1528-1534页，2020.

[67] Capuano, N., Fenza, G., Loia, V. 和 Stanzione, C.，“可解释的人工智能”
在网络安全：调查，“IEEE Access”，第10卷，第93575-93600页，2022年。

[68] https://www.gpai.ai/projects/responsible-ai/

[69] https://www.forbes.com/sites/jenniferhicks/2022/07/28/explainable-ai-is--trending-and-heres-why/?sh=3ce3c4952008

[70] https://www.forbes.com/sites/jenniferhicks/2022/07/28/explainable-ai-is--trending-and-heres-why/?sh=3ce3c4952008

[71] 张, X., 王娜, N., 沈昊, H., 邸思, S., 罗璇, X., 及 王涛, T., “火下的可解释深度学习”，收录于 *第29届USENIX安全研讨会，2020* .
[72] Kuppa, A.，和Le-Khac, N. A.，“网络安全中可解释人工智能（XAI）方法的黑盒攻击”载于 *2020年国际神经网络联合会议（IJCNN）*（第1-8页），2020年。
[73] Ghorbani, A., Abid, A., and Zou, J., “对神经网络的解释是脆弱的，” *人工智能AAAI会议论文集*（第33卷，第01期，第3681-3688页），2019年。
[74] Dombrowski, A. K., Alber, M., Anders, C., Ackermann, M., Müller, K. R., and Kessel, P., “解释可以被操纵，几何学是罪魁祸首”。 *神经信息处理系统进展* ，*32* ，2019.
[75] Shokri, R., Strobel, M., and Zick, Y., “Privacy risks of explaining machine learning models,” http://arxiv.org/abs/1907.00164 2019. [76] Hall, P., Gill, N., & Schmidt, N., “Proposed guidelines for the responsible use of explainable machine learning,” arXiv preprint arXiv:1906.03533, 2019. [77] Zhao, X., Zhang, W., Xiao, X., & Lim, B., “Exploiting explanations for model inversion attacks,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 682-692, 2021. [78] Shokri, R., Strobel, M., and Zick, Y., “On the privacy risks of model explanations,” Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 231- 241, 2021. [79] Kuppa, A., and Le-Khac, N. A., “Adversarial xai methods in cybersecurity”, IEEE Transactions on Information Forensics and Security, 16, pp.4924-4938, 2021. [80] https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf
[81] Pan, Z., & Mishra, P.，《网络安全可解释人工智能》一书。Springer Nature出版社，2024年。