让机器学习更值得信赖*

——安全性、透明性和公平性对机器学习的高风险应用至关重要

作者：比哈努·埃希特（Birhanu Eshete）
译者：王嘉龄　陆胤瑜　程时伟

关键词：机器学习　安全　信赖

译者按：近年来机器学习技术得到了飞速发展，在图像识别、语音处理、自然语言理解等应用领域取得了显著的效果，但是相关的安全性、公平性和隐私保护等问题也日益凸显。为此，本文从可信赖的机器学习这一概念出发，论述了机器学习面临的对抗性威胁的机制，并给出了相关案例，以及相应的技术解决思路。在此基础上，进一步从内在冲突、社会规范、政策制定等角度给出了可信赖机器学习的实现路径。

在过去的十年里，机器学习取得了显著的进步，并在图像、语音和文本识别等重要任务中不断展现出与人类水平相当的卓越表现。机器学习正在推动越来越多的高风险应用领域的发展，例如自动驾驶、自行任务型无人机、入侵检测、医疗图像分类和财务预测等[1]。然而，机器学习必须获得若干进展，才能将其放心地应用于直接影响人员培训和操作的领域，对于这些领域的很多应用来说，安全性、隐私性和公平性都是至关重要的考虑因素[1, 2]。

要开发一个值得信赖的机器学习模型，必须在其内部建立起能够防御各种对抗性攻击的保护措施（如图1所示）。机器学习模型需要训练数据集，而这些数据集有可能因被植入、修改或删除某些训练样本而“中毒”，使得模型的决策边界受到影响，从而达到攻击者的目的[3]。当模型从众包数据或运行时接收的输入数据中学习时，就可能发生“中毒”，因为这两种数据都很容易被篡改。对抗性操纵输入通过有意伪造的“对抗性样本”来规避机器学习模型[4]。例如，在自动驾驶中，车辆控制模型依靠识别路标进行导航。通过在停车标志上贴上一张小贴纸，攻击者就能规避模型，使其错误地将“停车”标志识别成“让行”标志或“限速45迈”的标志，但人类驾驶员会轻而易举地忽略这张在视觉上无关紧要的贴纸，并在停车标志处刹车（如图1所示）。

机器模型在训练和部署过程中很容易受到恶意攻击，当用户通过应用程序的模型预测接口（API）

* 本文译自 *Science*，“Making machine learning trustworthy”，2021, 373(6556): 743~744. DOI: 10.1126/science.abi5052. 一文。翻译和印刷版权已征得AAAS同意。此译文不是原文作者和AAAS工作人员的官方翻译，由 *CCCF* 特邀译者翻译。在关键问题上，请参考由AAAS出版的英文原文（https://www.science.org/doi/10.1126/science.abi5052）。

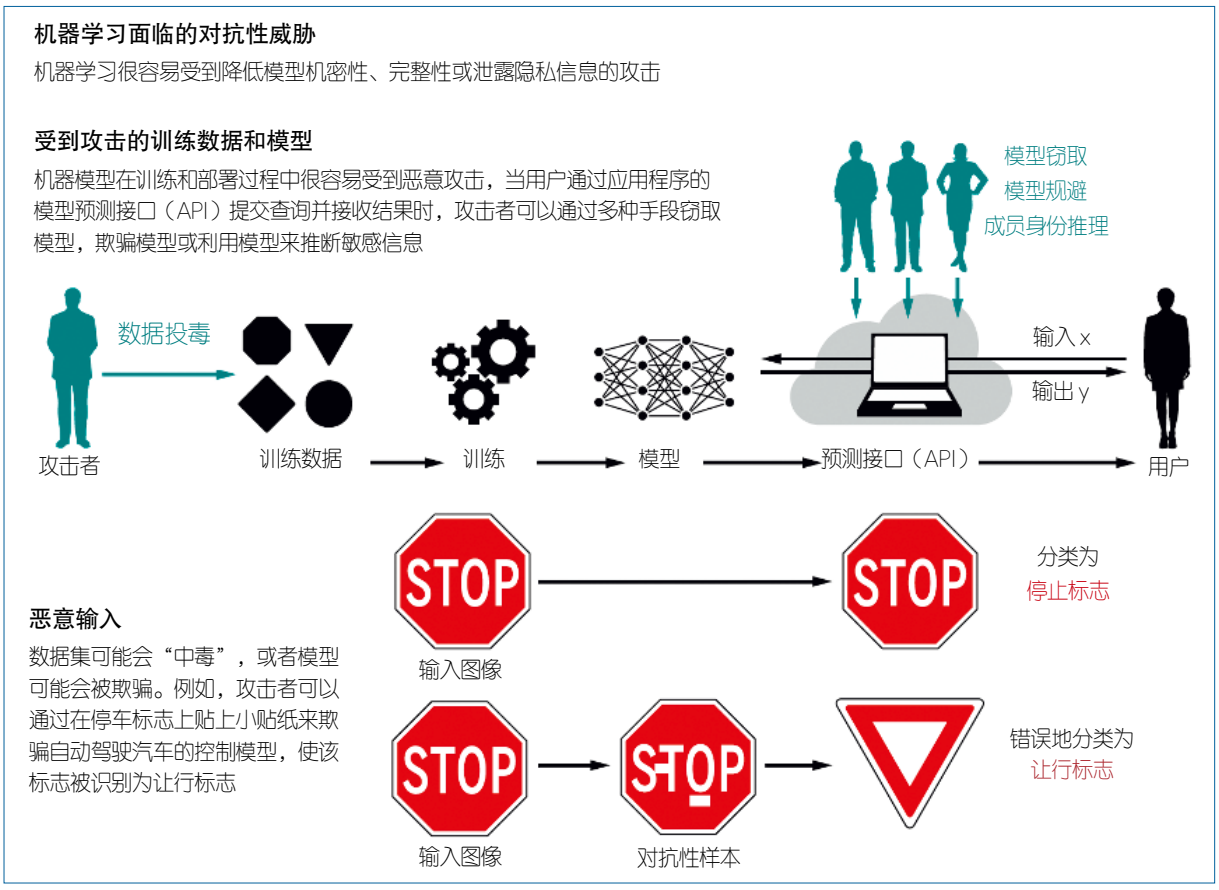

图1　机器学习面临的对抗性威胁

提交查询并接收结果时，攻击者可以通过多种手段窃取模型，欺骗模型或利用模型来推断敏感信息。

攻击还可以滥用模型预测接口的输入－输出交互来窃取机器学习模型本身[5, 6]。通过提供一批输入（例如可以公开获取的交通标志图像）并获得每个输入的预测结果，原有模型就可充当标签数据集，使攻击者能训练出一个在功能上与原有模型等效的替代模型。这类攻击给那些从知识产权、军事或国家安全情报等高安全风险数据中学习的机器学习模型带来了更高的风险。

当模型被训练用于基于敏感私密数据（如患者的临床诊断数据和银行客户的交易数据）的预测分析时，隐私问题就至关重要了。针对隐私的攻击可以仅仅通过与部署模型的交互就可以获取训练数据中包含的敏感信息[7]。造成此类攻击的根本原因是机器学习模型倾向于“记忆”训练数据的附属部分，并在预测时无意泄露了对训练数据有贡献的个人身份信息。成员推理（membership inference）是一种常见的策略，它让攻击者能够利用模型在训练数据集成员和非成员时产生的响应差异[7]。

应对机器学习模型面临的这些威胁还是有希望的。例如，检测中毒数据和对抗性样本输入方面的研究已经取得了进展，可以限制攻击者仅通过模型交互学习到的内容，降低模型窃取或成员推理攻击的危害程度[1, 8]。一个有前景的解决方法就是正式制定严格的隐私规范。差分隐私的概念是向参与数据集构建的个体保证，无论参与个体的记录是否属于模型的训练数据集，攻击者通过与模型交互获取到的参与个体的信息基本上是相同的[9]。

除了技术上的补救措施之外，从机器学习攻防军备竞赛中吸取的教训也带来了一些机遇去激发更广泛的研究，使机器学习在满足社会需求方面真正值得信

赖。相关问题包括：机器学习模型在进行决策时是如何“思考”的（透明性），以及训练机器学习模型去解决高风险推理任务时的公平性（人类做出的决策会存在偏见）。要想在机器学习的可信赖性方面取得有意义的进展，需要理解机器学习在解决人类需求时，传统的安全和隐私需求与如透明性、公平性、道德这类更广泛的问题之间的关系，这种关系有时是矛盾的。在相应的机器学习应用中已经出现了一些令人担忧的偏见案例[10, 11]，例如对种族和性别的错误判定，机器学习模型错误地认为深色皮肤种族的人有较高的犯罪可能性，在求职简历筛选中明显偏向男性求职者，以及在医学实验中排除黑人患者等。这就要求机器学习模型开发人员不再局限于科技的方式去赢得那些被偏见对待的人类用户的信任。

在研究前沿，特别是在机器学习的安全性和隐私性方面，上述防御对策强化了对抗环境下机器学习模型关于盲点的认识[8, 9, 12, 13]。在公平性和道德伦理方面，有足够多的证据表明机器学习存在缺陷，特别是在训练数据集对象代表性不足的问题上。因此，让机器学习做到公平并具有道德，需要建立以人为中心且更加全面的方式，在这方面还有很多工作需要完成。导致机器学习偏见的根本原因的一种误解是将偏见归因于数据，而且仅仅是数据。虽然数据采集、样本化和标注在造成重大偏差上起到了关键的作用，但在数据处理流程中也存在多个产生偏见的关键节点。从数据采样到特征提取，从训练过程中的集成到测试过程中的评价方法和指标，偏见问题在整个机器学习的数据处理流程中都会出现。

目前，关于对抗鲁棒性[13]以及隐私保护的机器学习模型，还缺乏被广泛认可的定义和公式（差分隐私除外，它在形式上很有吸引力，但尚未广泛部署）。攻击、防御的概念和度量指标在不同领域间缺乏可迁移性，也是阻碍可信赖的机器学习发展的一个迫切问题。举例来说，对于前文所述的大多数机器学习模型而言，模型规避和成员推理攻击主要集中在如图像分类（自动驾驶车辆检测道路标志）、对象检测（从包含多个物体的客厅照片中识别一朵花）、语音处理（语音助手）和自然语言处理（机器翻译）等应用场景。在视觉、语音和文本领域下面临的威胁和提出的相应对策很难相互转换，通常都是对抗性领域，例如网络入侵检测和金融欺诈检测。

另一个重要的考虑因素是一些可信赖属性之间内在的冲突。例如，透明性和隐私性往往是冲突的，因为如果一个模型用于训练隐私敏感数据，在应用时以最高水平的透明性为目标将不可避免地导致数据集中样本个体的隐私敏感细节泄露[14]。因此，需要选择在何种程度上牺牲透明性来保护隐私，反之亦然，而且要让系统购买者和用户清楚地了解这种选择。一般来说，人们对于隐私的担心是普遍存在的，因为如果不强制执行，会产生法律问题（例如，美国《健康保险的便捷性和责任法案》中的患者隐私问题）。此外，隐私性和公平性的发展并不总是同步的。例如，尽管基于隐私保护的机器学习模型（如差分隐私）在一定程度上保证了训练样本中个体的不可区分性，但就有效性而言，有研究表明，训练数据中的少数群体（例如基于种族、性别或性取向）易受到模型输出的负面影响[15]。

总的来说，科学界需要退一步，将机器学习的鲁棒性、隐私性、透明性、公平性和道德伦理规范与人类的规范保持一致。为此，需要针对鲁棒性和公平性制定和接受更明确的规范。在研究工作中，针对对抗鲁棒性、公平性和透明性所制定的有局限性的规范，必须被广泛适用的规范所取代，类似差分隐私。在政策制定过程中，需要采取具体步骤来建立监管框架，阐明针对偏见的可执行的问责措施和数据集的道德规范（包括多样性指南）、训练方法（如偏差－感知训练）和基于输入的决策（如通过解释来增强模型决策）。希望这些监管框架最终会演变成由立法支持的机器学习治理方式，从而在未来形成可靠的机器学习系统。

最关键的是，迫切需要来自不同科学团体的洞察，思考怎样的社会规范才会使用户能放心地使用机器学习进行高风险的决策，例如乘客乘坐无人驾驶汽车，银行客户接受机器人的投资建议，以及患者信得过的在线诊疗接口。在这些高风险的应用中，需要制定政策来保证机器学习能被安全、公正

地使用。同样重要的是，在对抗鲁棒性和模型准确性、隐私性和透明性，公平性和隐私性这些根本冲突之间，可信赖的机器学习需要有更加严格且具有社会基础的论证。幸运的是，在当前使用机器学习的关头上，在机器学习技术被广泛部署并变得难以管理之前，解决其盲点问题的重要机会窗口仍然敞开着。 ■

参考文献

[1] I. Goodfellow, P . McDaniel, N. Papernot, *Commun. ACM* 61, 56 (2018).

[2] S. G. Finlayson et al., *Science* 363, 1287 (2019).

[3] B. Biggio, B. Nelson, P. Laskov, Proceedings of *the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, J. Langford and J. Pineau, Eds. (Omnipress, 2012), pp. 1807–1814.

[4] K. Eykholt et al., Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition* (*IEEE, 2018*), pp. 1625–1634.

[5] F. T ramèr, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart, Proceedings of *the 25th USENIX Security Symposium, Austin*, TX (USENIX Association, 2016), pp. 601–618.

[6] A. Ali, B. Eshete, Proceedings of *the 16th EAI International Conference on Security and Privacy in Communication Networks*, Washington, DC (EAI, 2020), pp. 318–338.

[7] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Proceedings of *the 2017 IEEE Symposium on Security and Privacy*, San Jose, CA (IEEE, 2017), pp. 3–18.

[8] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, K. Talwar, arXiv:1610.05755 [stat.ML] (2017).

[9] I. Jarin, B. Eshete, Proceedings of *the 7th ACM International Workshop on Security and Privacy Analytics* (*2021*), pp. 25–35.

[10]J. Buolamwini, T. Gebru, Proceedings of *Conference on Fairness, Accountability and Transparency*, New York, NY (ACM, 2018), pp. 77–91.

[11]A. Birhane, V. U. Prabhu, Proceedings of *the IEEE/CVF Winter Conference on Applications of Computer Vision* (*IEEE, 2021*), pp. 1537–1547 .

[12]N. Carlini et al., arXiv:1902.06705 [cs.LG] (2019).

[13]N. Papernot, P. McDaniel, A. Sinha, M. P. Wellman, Proceedings of 3rd IEEE European Symposium on Security and Privacy (London, 2018), pp. 399–414.

[14]R. Shokri, M. Strobel, Y. Zick, Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY (2021); https://www.comp.nus.edu.sg/~reza/files/Shokri-AIES2021.pdf.

[15]V. M. Suriyakumar, N. Papernot, A. Goldenberg, M. Ghassemi, FAcc'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM, 2021), pp. 723–734.

作　者：

比哈努·埃希特（Birhanu Eshete）

密歇根大学迪尔伯恩分校计算机与信息科学系。
birhanu@umich.edu

译　者：

王嘉龄

CCF 学生会员。浙江工业大学计算机科学与技术学院博士研究生。主要研究方向为人机交互，脑机交互。
527304765@qq.com

陆胤瑜

CCF 学生会员。浙江工业大学计算机科学与技术学院本科生。
buliugu6@outlook.com

程时伟

CCF 杰出会员，CCF 人机交互专委会常务委员，*CCCF* 特邀译者。浙江工业大学计算机科学与技术学院教授，计算机软件研究所副所长。主要研究方向为人机交互、脑机交互、普适计算与协同计算。
swc@zjut.edu.cn

（本文责任编委：**姜　波**）