Adversarial Machine Learning: Aprendizado de Máquina em Contextos Inseguros

Oscar Zibordi de Paiva^{1,3}, Marcos Antonio Simplicio Jr¹, Charles Christian Miers²

¹Escola Politécnica da Universidade de São Paulo
São Paulo, SP – Brasil

²Universidade do Estado de Santa Catarina
Joinville, SC – Brasil

³Banco Bradesco S.A.
Osasco, SP – Brasil

{ozpaiva,mjunior}@larc.usp.br, charles.miers@udesc.br

Abstract. *The ubiquity of Machine Learning (ML) applications in a wide range of computing systems raises worrying questions about the resilience of ML models and algorithms in the face of malicious actions. Such questions concern a sub-area of cybersecurity called Adversarial Machine Learning (AML) - a name that refers to ML in adversarial contexts, that is, in which there are attackers. The number of publications on AML has grown significantly in the last five years, a fact that is indicative of the importance and the number of open questions in the area. This proposal details some of the reasons why these open questions will constitute research challenges for the next decade and highlights the reasons why traditional cybersecurity fails when applied to ML.*

Resumo. *A ubiquidade de aplicações de Machine Learning (ML) nos mais variados sistemas computacionais suscita questões concernentes à resiliência dos modelos e algoritmos de ML frente às ações maliciosas. Questões desse tipo inserem-se numa subárea da cibersegurança denominada Adversarial Machine Learning (AML) - nome que faz referência a ML em contextos adversariais, i.e., em que há atacantes. O número de publicações sobre AML cresceu significativamente nos últimos cinco anos, fato que é indicativo da importância e do número de questões ainda em aberto na área. Esta proposta detalha algumas das razões pelas quais essas questões em aberto serão desafios de pesquisa para a próxima década e destaca as razões pelas quais a cibersegurança tradicional falha ao ser aplicada no âmbito de ML.*

1. Introdução

A Cibersegurança é uma área da Computação que envolve tecnologia, pessoas, informações e processos para possibilitar operações com garantias de segurança. A área tem natureza multidisciplinar, contemplando a criação, operação, análise e teste de sistemas computacionais seguros, ao mesmo tempo que envolve aspectos legais, políticos, fatores humanos, ética e gestão de risco, com o objetivo de considerar contextos adversariais. O atendimento a requisitos de cibersegurança é relevante em uma multitude de cenários, ganhando ainda maior importância e necessidade de esforços de pesquisa no

contexto de tecnologias emergentes, as quais muitas vezes não são projetadas com proteções em mente. Uma das tecnologias atuais que apresenta esse aspecto emergente consiste em sistemas de inteligência artificial (IA). Historicamente, a sinergia entre IA e cibersegurança já é bastante explorada para construir sistemas de segurança inteligentes. Nesse sentido, o papel da IA, especificamente da subárea de Aprendizado de Máquina (*Machine Learning* - ML), é similar àquele apresentado em diversas outras áreas: uma ferramenta útil para lidar com diversas tarefas rotineiras e intensivas em dados, permitindo que especialistas humanos se concentrem em aspectos estratégicos, criativos e eticamente complexos (CSA, 2024). Por outro lado, é menos comum observar uma exploração de sinergias no sentido inverso, de proteger os algoritmos, práticas e ferramentas de IA contra potenciais adversários maliciosos. Com a crescente adoção de sistemas inteligentes em uma expressiva variedade de cenários, entretanto, têm se tornado cada vez mais importantes os esforços para sanar essa lacuna. A área de estudo que se dedica a essa questão se denomina *Adversarial Machine Learning* (AML) (Vassilev et al, 2025), e tem por objetivo a proteção de sistemas inteligentes baseados em *Machine Learning* (ML). Algumas ameaças nesse cenário incluem o eventual mau uso de ferramentas inteligentes para atacar outros sistemas, o envenenamento de bases de dados para treinamento, a violação de privacidade por meio da extração de dados de usuários incluídos nesta base, a violação de direitos autorais por meio da extração de modelos de terceiros, entre vários outros. Não por acaso, essa é uma área que tem recebido especial atenção por órgãos de segurança ao redor do mundo, como o *National Institute of Standards and Technology* (NIST) (Vassilev et al, 2025) e a *European Union Agency for Cybersecurity* (ENISA) (Malatras et al., 2021). Ao mesmo tempo, o número de trabalhos científicos sobre AML cresceu significativamente nos últimos anos, tendência que deve se intensificar com a adoção cada vez maior de sistemas inteligentes em diversos cenários de aplicação.

2. Fatores Diferenciativos de AML e Questões de Pesquisa Subjacentes

Em comparação a sistemas de software tradicionais, a utilização de ML introduz riscos de cibersegurança bastante peculiares. Uma das razões para isso está relacionada à característica inata destes algoritmos, grosso modo, de automática ou semi-automática-mente, transformar dados em programas. Tal transformação ocorre de forma indireta e convoluta, dividindo-se em duas etapas: uma primeira de treinamento, em que dados (denominados de dados de treinamento) são utilizados para buscar, em um espaço de representação, um modelo matemático que aproxima uma determinada função-alvo parcialmente representada nos dados de treinamento; e uma etapa posterior de inferência, em que o modelo aprendido é utilizado para inferir o valor da função-alvo para valores novos, ausentes do conjunto de dados de treinamento. Este arranjo, baseado em dados e não em um processo de desenvolvimento clássico, tende a dificultar a incorporação de propriedades básicas de segurança, a aplicação de boas práticas consagradas de programação defensiva e, até mesmo, a modelagem de requisitos para a correta proteção do sistema (Cazzaniga et. al., 2024). Dentre as propriedades e boas práticas consagradas violadas pelo ML, podem-se listar:

- **não-interferência:** propriedade de segurança de sistemas segundo a qual as ações de um programa ou usuário não devem afetar outros programas ou usuários além do que é explicitamente necessário. O treinamento de modelos em diversos sistemas inteligentes atuais frequentemente usa dados provenientes de usuários e não-usuários,

gerando um modelo global efetivamente criado e utilizado por todos: torna-se, portanto, difícil garantir propriedades de não-interferência. Um problema análogo é observado em cenários nos quais dados de vários usuários são usados também na etapa de inferência, influenciando o comportamento global das saídas do sistema.

- **codificação segura:** diz respeito às práticas de desenvolvimento de software com superfície de ataque reduzida, resistente a entradas possivelmente maliciosas. Dentre as práticas preconizadas pela área, podem-se listar desde a verificação completa de adequação de entradas (programação defensiva) até a utilização de linguagens com gestão segura de memória. No âmbito de ML, uma vez que os "programas" aprendidos são na realidade modelos estatísticos consultados em tempo de inferência, o campo de conhecimento da programação segura oferece pouca ajuda à tarefa de proteger sistemas usando ML. Conseqüentemente, o efeito de entradas maliciosas enviadas a esses sistemas costuma ser imprevisível.
- **mínimo privilégio:** princípio de projeto segundo o qual um programa ou usuário não deve ter acesso a mais dados ou recursos de sistema além do que estritamente necessário para a execução de suas tarefas. Especificamente no âmbito de ML, modelos são treinados a partir de conjuntos de dados globais, potencialmente sigilosos (e, por vezes, até mesmo maliciosos). Ainda, durante a etapa de inferência, usuários comumente têm acesso à totalidade da capacidade preditiva dos modelos, violando o princípio de mínimo privilégio. Com isso, o efeito de entradas maliciosas durante a etapa de inferência pode, por exemplo, dar origem a divulgação de dados de forma indevida, comprometendo o sigilo dos dados de treinamento ou inferência, e possivelmente a privacidade das pessoas às quais os dados estão associados.
- **simplicidade:** mecanismos de segurança devem ser tão simples quanto possível, no intuito de minimizar as chances de erros de projeto ou implementação. Os modelos e algoritmos de ML costumam violar essa propriedade de duas formas: (i) Modelos e algoritmos de ML são frequentemente gigantescos e incompreensíveis (e.g., redes neurais multicamadas). Com isso, a inspeção humana (como um mecanismo de segurança, mesmo que rudimentar) torna-se inviável. Soma-se a essa incompreensão o fato de que pesquisas na área de ML encontram-se frequentemente em estágio pré-científico, contendo dados empíricos sem explicações e resultados irreprodutíveis. (ii) Projetar políticas de segurança torna-se muito mais difícil no contexto de ML. Tradicionalmente, as políticas de segurança costumam se restringir a regras de controle de acesso bastante "cartesianas", versando regras como "determinado programa só pode acessar um dado recurso em certas circunstâncias". Em sistemas empregando ML, as políticas de segurança usualmente não podem ser descritas em tamanha simplicidade. Um exemplo aparece na construção de modelos usando dados sigilosos: nesse caso, é necessário definir políticas de segurança em termos de privacidade, uma propriedade de segurança relacionada à confidencialidade, porém muito mais sutil e, por isso, muito mais difícil de proteger do que esta última mesmo em sistemas tradicionais. Outro exemplo se dá no contexto da definição de quais assuntos um chatbot deve ou não abordar em LLMs: tais filtros envolvem conceitos de linguagem natural e, além disso, critérios de julgamento bastante subjetivos.

Frente a essa dificuldade de aplicar boas práticas e propriedades de segurança em ML, não é surpreendente que técnicas de ataque diversas venham sendo demonstradas na literatura (Liu et al., 2024). Essa superfície de ataque só vem crescendo com a expansão

nos tipos de algoritmos e modelos de ML elaborados na última década, sobretudo nos últimos cinco anos. Dentre essas categorias de ataque, podem-se listar: evasão de modelo, envenenamento de modelos, extração de dados de modelo (ou inversão de modelo), inferência de dados de treinamento do modelo, e, mais recentemente, subversão de políticas (“*jailbreak*”) e injeção de *prompt* em *chatbots* baseados em LLMs. Embora alguns mecanismos de proteção tenham sido propostos para combater esses ataques, ainda é difícil determinar se estes operam de forma satisfatória, i.e., se são eficazes e abrangentes, e se introduzem compromissos aceitáveis com relação a custos computacionais, carga de trabalho de analistas, e redução de acurácia preditiva. Justamente pelas características de cibersegurança únicas de modelos ML, as quais ainda se encontram em fase de estudo pela comunidade acadêmica especializada em proteção de sistemas, o corpo de conhecimento da área de cibersegurança fica aquém de fornecer alternativas práticas, em prontidão, para o enfrentamento das questões de AML. Para sanar essa lacuna, há ainda muita pesquisa científica a ser realizada, ao mesmo tempo que modelos e algoritmos de ML continuam sendo colocados em operação nas mais diversas aplicações. Assim sendo, o desafio que se coloca é ainda uma verdadeira corrida contra o tempo para garantir que todo o potencial transformador da IA não se torne sinônimo de riscos de cibersegurança. (ENISA, 2023).

3. Indicadores de Sucesso de Pesquisa em AML

Como subárea de cibersegurança, o principal indicador de sucesso da área de ML é a descrição de mecanismos de segurança que possam ser utilizados no mundo real, de forma prática. Isso se traduz em métricas para aferir a efetividade desses mecanismos contra ataques conhecidos, e aferir também o quão aceitáveis são os impactos causados em termos de funcionalidade, custo e desempenho. Adicionalmente, a área de AML carece ainda de modelos capazes de descrever os funcionamentos internos de algoritmos e modelos de ML. Essa propriedade é importante para avaliações diversas, inclusive de segurança: usando esses modelos, seria possível fazer previsões, por exemplo, do efeito de perturbações em dados de treinamento ou de inferência, atestando-se o correto funcionamento de sistemas em diferentes cenários (o que é, afinal, um requisito fundamental de engenharia). Essa questão tem forte correlação com a área de IA Explicável, que é também emergente e ainda pouco madura. Por fim, em termos práticos, é necessário o estabelecimento de modelos de análise de riscos e de modelamento de ameaças de ML. Isso não só direcionará a pesquisa científica a adotar modelos que sejam realistas, como poderá orientar empresas e demais organizações na adoção de mecanismos de segurança de forma priorizada.

4. AML e Questões Socioeconômicas

Os cenários adversariais a que estão sujeitas aplicações de ML podem influenciar e até agravar possíveis impactos socioeconômicos naturalmente causados por essas aplicações. Alguns desses impactos e seus agravamentos são:

- **Vieses preconceituosos:** em aplicações cujas previsões possam ter qualquer influência direta ou indireta sobre grupos sociais, usuários maliciosos podem explorar vulnerabilidades dos algoritmos de ML para, por exemplo, introduzir vieses que prejudiquem determinados grupos. Sistemas de concessão de crédito, prognóstico de propensão a crimes, diagnóstico médico, e até mesmo *chatbots* são exemplos de sistemas nos quais esse tipo de influência pode ocorrer.

- **Crimes cibernéticos:** sistemas de autenticação biométrica atuais são, em grande parte, dependentes de ML. Logo, nesse cenário, as vulnerabilidades inerentes a algoritmos e modelos de ML podem levar, quando exploradas, a violações de acesso. Em algumas aplicações, pode haver perdas materiais de alvos específicos ou de grupos sociais inteiros (e.g., subtração de benefícios de vulneráveis). A autenticação entre humanos também pode ser prejudicada ao se usar sistemas de ML em ataques de engenharia social, gerando textos para *spear phishing* ou áudio e vídeo que personificam familiares ou colegas de trabalho de possíveis vítimas.
- **Violação de direitos autorais:** a IA generativa pode reproduzir ideias e estilos de artistas cujas obras forem usadas para treinar os modelos subjacentes. A proteção de direitos autorais nesse caso suscita questões adversariais, como a inserção de perturbações nas obras (envenenamento) para evitar que os modelos treinados reproduzam o estilo nelas presente.
- **Geração automática de conteúdo prejudicial:** modelos generativos de áudio, texto e vídeo podem ser usados maliciosamente para gerar notícias falsas e outros conteúdos socialmente prejudiciais em escalas sem precedentes. A proteção desses modelos, prevenindo tal utilização, e o uso de modelos para detecção de conteúdo sintético, relacionam-se a aspectos de AML.

Referências

- (Cazzaniga et. al., 2024). Cazzaniga et. al., “Gen-AI: Artificial Intelligence and the Future of Work.” IMF Staff Discussion Note SDN2024/001, International Monetary Fund, Washington, DC.
- (CSA, 2024) “AI Model Risk Management Framework”. Cloud Security Alliance. (Julho/2024).
- (ENISA, 2023). European Union Agency for Cybersecurity. "Artificial Intelligence and Cybersecurity Research".
- (Liu et al., 2024) Liu, Y. "A hitchhiker’s guide to jailbreaking chatgpt via prompt engineering". 4th International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things, SEA4DQ 2024. ACM.
- (Malatras et. al., 2021) European Union Agency for Cybersecurity, Malatras, A., Agrafiotis, I., and Adamczyk, M. "Securing machine learning algorithms". Publications Office of the European Union.
- (Vassilev et al., 2025) Vassilev et. al., “Adversarial machine learning: A taxonomy and terminology of attacks and mitigations.”