ESCOLA DE APERFEIÇOAMENTO DE OFICIAIS DA AERONÁUTICA

DIVISÃO DE ENSINO

CURSO DE APERFEIÇOAMENTO DE OFICIAIS 1º/2025

PAULO FERNANDO FERREIRA **SILVA FILHO**, Cap Eng

Inteligência Artificial Explicável: Garantir confiabilidade e segurança cibernética aos protótipos do Projeto C2

Rio de Janeiro

2025

ESCOLA DE APERFEIÇOAMENTO DE OFICIAIS DA AERONÁUTICA
DIVISÃO DE ENSINO
CURSO DE APERFEIÇOAMENTO DE OFICIAIS 1º/2025

PAULO FERNANDO FERREIRA **SILVA FILHO**, Cap Eng

Inteligência Artificial Explicável: Garantir confiabilidade e segurança cibernética aos protótipos do Projeto C2

Trabalho de conclusão de curso apresentado à Escola de Aperfeiçoamento de Oficiais da Aeronáutica como requisito parcial para aprovação no Curso de Pós-Graduação *Lato Sensu* em Liderança com Ênfase em Gestão no COMAER.
Linha de Pesquisa: Uso da Inteligência Artificial
Orientador: Eduardo Mendes Marcondes, Maj Av

Rio de Janeiro
2025

PAULO FERNANDO FERREIRA **SILVA FILHO**, Cap Eng

Inteligência Artificial Explicável: Garantir confiabilidade e segurança cibernética aos protótipos do Projeto C2

Trabalho de conclusão de curso apresentado ao Curso de Aperfeiçoamento de Oficiais da Escola de Aperfeiçoamento de Oficiais da Aeronáutica.

Aprovado por:

_____

Presidente, Eduardo Mendes Marcondes, Maj Av – EAOAR

_____

Durval Aquino Mota, Cap Esp Sup Tec – GLOG-CG

Rio de Janeiro

2025

RESUMO

Sistemas de reconhecimento de alvos baseados em Inteligência Artificial (IA) são de grande interesse para a Defesa, devido à sua capacidade de processar grandes volumes de dados, proporcionando consciência situacional com maior rapidez e precisão. No entanto, a falta de transparência e complexidade dos algoritmos mais sofisticados de aprendizado de máquina comprometem sua aplicabilidade. Nesse contexto, este ensaio defende a incorporação de técnicas de Inteligência Artificial Explicável (XAI) nos protótipos de detecção de alvos com IA do Projeto C2, desenvolvido pelo Instituto de Estudos Avançados. A incorporação de métodos explicáveis permite que o operador compreenda o processo de inferência da IA, promovendo maior confiança, usabilidade e aceitação do sistema. Além disso, as explicações funcionam como mecanismos de alerta, que permitem a identificação de possíveis ataques cibernéticos a partir de contradições entre os resultados apresentados e suas explicações. Desse modo, a robustez dos sistemas é ampliada. Aumentar a transparência dos métodos de IA torna-se, assim, indispensável em qualquer sistema crítico empregado pelas Forças Armadas, uma vez que fornecem subsídios para decisões mais precisas e confiáveis em toda a estrutura decisória, desde operadores até os mais altos níveis de comando. Enquanto a IA assegura informações oportunas e precisas, a XAI agrega fidelidade e segurança, contribuindo diretamente para a construção de uma relação sinérgica entre humano e máquina e, por conseguinte, para o fortalecimento da capacidade dissuasória da Força Aérea Brasileira.

Palavras-chave: inteligência artificial explicável; reconhecimento de alvos; proteção cibernética; confiabilidade.

1 INTRODUÇÃO

A utilização de sistemas baseados em Inteligência Artificial (IA) em operações militares para o reconhecimento automático de alvos já é uma realidade. É inegável a importância dos sistemas de IA da Ucrânia na guerra atual (Bondar, 2024), incluindo o sistema *Maven* (Probasco, 2024), desenvolvido pelo Departamento de Defesa (DoD, do inglês *Department of Defense*) dos Estados Unidos, e o sistema ucraniano Delta. No Brasil, o Instituto de Estudos Avançados (IEAv), por meio do Projeto C2, tem desenvolvido protótipos baseados nesta tecnologia para detecção de alvos militares, cujos testes em ambiente operacional iniciaram-se nos recentes exercícios de Inteligência, Vigilância e Reconhecimento.

A principal vantagem desses sistemas no teatro operacional reside na capacidade de processar uma grande quantidade de dados e proporcionar uma consciência situacional com maior rapidez e precisão para a tomada de decisão. Ademais, o contínuo avanço tecnológico possibilita o desenvolvimento de sistemas capazes de perceber, aprender, reconhecer e agir autonomamente. O tempo para a tomada de decisão, portanto, reduzir-se-á ainda mais.

Entretanto, o processo de reconhecimento dos alvos por esses sistemas não é transparente. Isso ocorre porque os algoritmos mais avançados de aprendizado de máquina utilizam representações internas próprias para extrair características dos dados, que não são de fácil interpretação para um operador humano. Por isso, eles são considerados caixas-pretas.

Esta falta de transparência limita a aceitação, a usabilidade e a confiabilidade dos sistemas de reconhecimento por inteligência artificial. Sem o entendimento de como a tecnologia funciona, os usuários tendem a desconfiar ao primeiro sinal de um falso reconhecimento e deixam de usá-los. Além disso, existe uma preocupação mundial com a confiabilidade desses sistemas, principalmente em aplicações militares, visto que há uma certa transferência da responsabilidade pela decisão do humano para a máquina.

Ao mesmo tempo, a falta de transparência afeta a segurança dos sistemas de reconhecimento de alvos baseados em inteligência artificial. Ressalta-se que eles são produtos de tecnologia da informação e, portanto, são suscetíveis a ataques cibernéticos. Sem explicações relativas ao processo de reconhecimento, dificilmente um operador conseguiria detectar se as decisões do programa estariam comprometidas devido a um ataque.

Atualmente, o Projeto C2 utiliza técnicas de natureza caixa-preta para seus protótipos. Desta forma, este ensaio defende que sejam implantadas técnicas de inteligência artificial explicável (XAI, do inglês *eXplainable Artificial Intelligence*) no Projeto C2. Com o uso

dessas técnicas, espera-se um aumento na aceitação, usabilidade e confiabilidade dos protótipos de reconhecimento de alvos, tanto pelos usuários quanto pela sociedade. Além disso, essas técnicas proverão maior robustez contra ataques cibernéticos aos protótipos do projeto.

2 DESENVOLVIMENTO

IA é uma tecnologia que permite aos computadores perceberem, processarem, inferirem e agirem de forma automática ou autônoma. A principal abordagem para o desenvolvimento desses sistemas é por meio do aprendizado supervisionado, no qual os algoritmos são treinados com um grande volume de dados categorizados. O sistema, então, aprende representações internas complexas que refletem estatisticamente os padrões existentes nos dados e que servirão de base para as futuras decisões. Por conseguinte, nota-se que o desempenho destes modelos – seu poder de generalização – está diretamente associado à qualidade e à diversidade dos dados utilizados no treinamento. Uma vez que é impossível ter um conjunto de dados de treinamento que consiga representar todas as possibilidades em que este sistema será utilizado, permanece o desafio de garantir que ele funcione de forma confiável e compreensível em cenários não previstos.

Dessa forma, abrir as caixas-pretas da IA é essencial para compreender o raciocínio do algoritmo e confiar nele. As técnicas de XAI buscam produzir explicações para os métodos de aprendizado de máquina, de forma que um usuário do sistema possa compreender o processo de decisão e, assim, confiar nele. Além disso, estas técnicas permitem saber se algum erro do algoritmo ocorreu por treinamento deficiente do sistema, por se deparar com um dado novo, ou por estar diante de um ataque cibernético.

2.1 EXPLICAR PARA CONFIAR

A confiança em um sistema autônomo está diretamente relacionada ao nível com o qual os usuários o entendem e à precisão com a qual conseguem prever seu desempenho em uma atividade (Akula *et al.,* 2022; Michel, 2020). Segundo Lee e See (2004), as pessoas tendem a rejeitar automações nas quais elas não confiam, independentemente dos benefícios e capacidades que o sistema pode proporcionar, especialmente quando há fatores de responsabilidade envolvidos.

Um exemplo desse aspecto de rejeição ocorreu durante o desenvolvimento inicial do sistema de suporte à tomada de decisão baseado em IA dos Estados Unidos: *Maven*. Ele utiliza dados de diversos sensores (ópticos, termal, radar de abertura sintética, entre outros) para auxiliar combatentes na identificação de alvos militares e prover o devido fluxo de autorização para engajamento à cadeia de comando. Apesar dos claros benefícios do sistema, Probasco (2024) e Manson (2024) relatam que unidades não queriam utilizá-lo e, ao primeiro sinal de degradação da acurácia, o desligavam.

Entretanto, após aprimoramentos do sistema nos exercícios *Scarlet Dragon* do 18º Corpo Aerotransportado, essa desconfiança diminuiu, visto que a troca de experiência entre engenheiros, desenvolvedores, combatentes e técnicos (Probasco, 2024) melhorou a usabilidade do sistema. Inclusive, o *software* foi posto à prova na operação de evacuação de Cabul em 2021 (Probasco, 2024), na identificação de alvos no Iêmen, Iraque e Síria (Manson, 2024) e na identificação de tropas russas na Guerra da Ucrânia (Probasco, 2024; Sanger, 2024).

De fato, todos os sistemas de IA, inclusive o *Maven*, têm um certo grau de falha diretamente relacionado aos seus dados de treinamento. É bastante provável que o sistema se depare com novos dados e circunstâncias no ambiente operacional para os quais não foi treinado e, portanto, terá um comportamento imprevisível. Se este comportamento representar uma falha e não houver meios de entender o raciocínio por trás da decisão, a confiança no sistema será reduzida. Ou seja, a falta de explicação compromete todo o ganho operacional que a IA pode proporcionar.

Diante disto, a Agência de Projetos de Pesquisas Avançados de Defesa (DARPA, do inglês *Defense Advanced Research Projects Agency*) dos Estados Unidos gerenciou um programa para desenvolver técnicas de XAI. O objetivo do programa era possibilitar que operadores entendessem, confiassem e efetivamente utilizassem sistemas baseados em IA (Gunning *et al.*, 2021). O programa foi desenvolvido entre 2017 e 2021.

Uma das principais conclusões deste programa foi que usuários preferem sistemas que apresentam uma explicação alinhada à decisão aos que apresentam apenas a decisão (Gunning *et al.*, 2021). Isso ocorre porque as explicações permitem a compreensão do processo de decisão, gerando maior transparência e, portanto, confiança, conforme observado nos experimentos desenvolvidos em Akula *et al.* (2022). Nestes experimentos, 210 usuários avaliaram quantitativa e qualitativamente a efetividade das explicações obtidas por diferentes técnicas de XAI em um sistema de reconhecimento de imagens. Os resultados comprovaram

que a presença de qualquer técnica de XAI tende a aumentar a confiança nesses sistemas em comparação a quando não há explicação.

De certo modo, as explicações providas pelas técnicas de XAI ajudam os usuários a criar um modelo mental aproximado de como o sistema funciona. Ou seja, as pessoas que têm acesso a explicações conseguem prever melhor qual será o resultado do sistema. E, portanto, confiam mais nele. Além disso, mesmo quando a IA erra, as explicações amenizam o efeito de desconfiança e podem contribuir para a melhoria da performance futura (Gunning *et al.,* 2021; Nagahisarchoghaei *et al.*, 2023). De fato, Leichtmann *et al.* (2023) demonstraram que, mesmo utilizando uma IA com baixo desempenho, usuários tiveram uma melhora significativa nas suas tomadas de decisão pela presença das explicações.

Outro aspecto a ser considerado que afeta a confiabilidade, a usabilidade e a aceitação das técnicas de aprendizado de máquina é saber que o sistema segue princípios éticos de desenvolvimento. No caso de sistemas militares, a diretiva 3000.09 do DoD dos Estados Unidos (United States, 2023) consolidou os seguintes cinco princípios: responsabilidade, equidade, rastreabilidade, confiabilidade e governabilidade. De acordo com Radanliev (2025) e Hu *et al.* (2021), XAI é uma das soluções técnicas para implementar alguns destes princípios. Inclusive, Osswald, *et al.* (2023) aponta que o Sistema de Combate Aéreo Futuro (FCAS, do inglês *Future Combat Air System*) da União Europeia usará XAI para este fim.

Ressalta-se que o foco da XAI é aumentar a confiabilidade e usabilidade, com o menor impacto possível no desempenho do sistema. Isto só é obtido a partir de constante interação entre desenvolvedores e operadores, a fim de prover apenas as explicações essenciais para o cumprimento da missão.

Diante do exposto, é possível verificar que a implantação de técnicas de XAI será de extrema importância para os protótipos do Projeto C2, visto que as explicações providas aumentarão a confiabilidade, aceitação e usabilidade do sistema.

2.2 EXPLICAR PARA SE PROTEGER

Além de prover confiabilidade, aceitação e usabilidade, um segundo aspecto da importância da implementação da XAI reside no fato de que os sistemas de IA são ferramentas de tecnologia da informação e, portanto, estão suscetíveis a ataques cibernéticos. O ciberespaço é um dos domínios de guerra e consiste em ações no ambiente digital, que objetivam proteger ativos estratégicos e comprometer ou influenciar as capacidades adversárias (Parks; Duggan, 2011). Dentre as possíveis operações ofensivas no ciberespaço, a

desinformação digital – que consiste na manipulação de dados – é bastante crítica para sistemas de IA.

À medida que mais países utilizam aprendizado de máquina para reconhecimento de alvos, adversários poderão buscar meios de explorar vulnerabilidades e debilitar tais sistemas (Manson, 2024). Eles podem injetar dados falsos no treinamento, contaminar os dados de entrada com ruídos (chamados de ataques adversariais), afetar atualizações, entre outros métodos. Os algoritmos, portanto, podem perder acurácia sem que os operadores percebam, visto que o processo de reconhecimento do alvo pela IA é opaco.

Por exemplo, em 2022, houve uma tentativa de invasão ao sistema de consciência situacional Delta (que utiliza algoritmos de IA para processar dados) da Ucrânia, a partir de uma falsa atualização disseminada por *phishing* (Kovacs, 2022). Outras tentativas se sucederam e, em momentos, a Rússia chegou a informar que havia invadido o sistema. Entretanto, isso se tratava apenas de uma tentativa de guerra de informação. Por serem sistemas novos, confidenciais e estratégicos, não existem relatos concretos de invasões bem-sucedidas.

Entretanto, a literatura científica já debate essas vulnerabilidades e demonstra possíveis soluções para aumentar a robustez dos sistemas. Técnicas de XAI são apontadas como possíveis métodos de detecção de ataques, visto que, ao se observar uma explicação não condizente, pode-se suspeitar de um ataque. Este aspecto é apresentado por Chakraborty *et al*. (2022), Klawikowska, Mikołajczyk e Grochowski (2020) e Makridis *et al*. (2022). Em todos esses trabalhos, as explicações antes e depois da contaminação dos dados são apresentadas para evidenciar o efeito dos ataques adversariais.

Corroborando com este mesmo aspecto, Stiff (2022) também demonstra que técnicas de XAI podem ser utilizadas para verificar a presença de dados contaminados. Além disso, o trabalho propõe utilizar uma rede neural complementar que analisa mapas de calor de explicações para identificar os ataques e tornar o sistema mais robusto.

Por outro lado, os ataques cibernéticos também evoluem e estudos já apontam para ataques direcionados às explicações e não necessariamente para a decisão (Baniecki; Biecek, 2024), com o intuito de afetar a confiança no sistema. Apesar de a XAI também possuir suas vulnerabilidades, entende-se que a inclusão dessas técnicas já aumenta o nível de complexidade que o ataque precisa ter para afetar o sistema, em comparação com sistemas que não têm explicações. Em outras palavras, a presença da XAI torna o sistema mais robusto a certos tipos de ataques (Galil; El-Yaniv, 2021). Já é de se esperar, porém, que atualizações

para métodos mais robustos serão sempre necessárias. À medida que os ataques evoluem, as defesas precisam evoluir em conjunto.

Em face do exposto, observa-se que aplicar técnicas de XAI ao Projeto C2 será essencial para proteger os protótipos desenvolvidos contra ataques cibernéticos, visto que as explicações providas ajudam na identificação de quando o sistema pode estar com dados contaminados, contribuindo com a sua robustez.

3 CONCLUSÃO

Sistemas de reconhecimento de alvos baseados em IA são extremamente estratégicos no teatro operacional, devido à sua alta capacidade de processar grandes volumes de dados. Obtém-se, assim, uma rápida e precisa consciência situacional. Contudo, a natureza de caixa-preta dos algoritmos mais avançados de aprendizado de máquina afeta a sua aplicabilidade prática em operações críticas, a exemplo da Defesa.

A fim de contrabalançar a falta de transparência desses métodos de IA, este ensaio argumenta sobre a importância da implantação de técnicas de XAI nos protótipos de detecção de alvos do Projeto C2, desenvolvido pelo IEAv. A inclusão destas técnicas é essencial para demonstrar a capacidade da IA em auxiliar nas tomadas de decisões, visto que as explicações aproximam o operador à máquina. Desta forma, aumenta-se a confiabilidade, usabilidade e aceitação dessas técnicas.

Concomitantemente, demonstrou-se que as explicações servem como alerta para proteger esses sistemas de possíveis ataques cibernéticos. Ao verificar contradições entre o reconhecimento e a explicação, os operadores podem suspeitar de ataques, em vez de aceitarem cegamente a inferência do algoritmo, aumentando, assim, a robustez dos sistemas.

Por fim, explicações para métodos caixa-preta da IA são essenciais em todo sistema crítico desenvolvido ou adquirido pelas Forças Armadas. Ao passo que a IA produz informações oportunas e precisas, a XAI contribui para a alta fidelidade, garantindo, assim, a superioridade da informação. Dessa forma, entende-se que, a partir das explicações, toda a cadeia de comando, desde operadores até os mais altos níveis, seja no Comando da Aeronáutica (COMAER) ou nos altos comandos das demais forças, pode tomar decisões mais assertivas e com maior precisão e confiabilidade diante das informações e inferências da IA. É, então, nesse aprimoramento da sinergia entre humanos e máquinas em missões críticas que a Força Aérea Brasileira reforça sua confiança em sistemas inteligentes e modernos, que contribuem para aumentar sua capacidade dissuasória.

REFERÊNCIAS

AKULA, A. R. *et al*. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. **iScience**, [*s. l.*], v. 25, n. 1, p. 103581, Jan. 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S2589004221015510. Acesso em: 30 mar. 2025.

BANIECKI, H.; BIECEK, P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. **Information Fusion**, [*s. l.*], v. 107, p. 102303, July 2024. Disponível em: https://www.sciencedirect.com/science/article/pii/S1566253524000812. Acesso em: 12 abr. 2025.

BONDAR, K. **Understanding the Military AI Ecosystem of Ukraine**. [Washington, D. C.]: Center for Strategic and International Studies, Nov. 2024. Disponível em: https://www.csis.org/analysis/understanding-military-ai-ecosystem-ukraine. Acesso em: 28 mar. 2025.

CHAKRABORTY, T. *et al*. Generalizing adversarial explanations with Grad-CAM. *In*: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 39., 2022, New Orleans. **Proceedings of the 2022** [...]. [*S. l.*]: IEEE, June 2022. p. 187-193. Disponível em: https://ieeexplore.ieee.org/document/9857321. Acesso em: 12 abr. 2025.

GALIL, I.; EL-YANIV, R. Disrupting deep uncertainty estimation without harming accuracy. *In*: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 35., 2021, on-line. **Proceedings of the 35th** [...]. Red Hook: Curran Associates Inc., Dec. 2021. p. 21285-21296. Disponível em: https://dl.acm.org/doi/10.5555/3540261.3541889. Acesso em: 22 mar. 2025.

GUNNING, D. *et al.* DARPA's explainable AI (XAI) program: A retrospective. **Applied AI Letters**, [*s. l.*], v. 2, n. 4, p. e61, Dec. 2021. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61. Acesso em: 04 fev. 2025.

HU, B. *et al.* XAITK: The explainable AI toolkit. **Applied AI Letters**, [*s. l.*], v. 2, n. 4, p. e40, Oct. 2021. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.40. Acesso em: 04 fev. 2025

KLAWIKOWSKA, Z.; MIKOŁAJCZYK, A.; GROCHOWSKI, M. Explainable AI for Inspecting Adversarial Attacks on Deep Neural Networks. *In*: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING, 19., 2020, Zakopane. **Artificial Intelligence and Soft Computing**: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I. Cham: Springer International Publishing, Oct. 2020. p. 134-146. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-61401-0_14. Acesso em: 11 abr. 2025.

KOVACS, E. Ukraine's Delta Military Intelligence Program Targeted by Hackers. **SecurityWeek**. [*S. l.*], Dec. 2022. Disponível em: https://www.securityweek.com/ukraines-delta-military-intelligence-program-targeted-hackers/. Acesso em: 09 abr. 2025.

LEE, J. D.; SEE, K. A. Trust in Automation: Designing for Appropriate Reliance. **Human Factors**, [*s. l.*], v. 46, n. 1, p. 50-80, Mar. 2004. Disponível em: https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392. Acesso em: 29 mar. 2025.

LEICHTMANN, B. *et al.* Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. **Computers in Human Behavior**, [*s. l.*], v. 139, p. 107539, Feb. 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S0747563222003594. Acesso em: 11 abr. 2025.

MAKRIDIS, G. *et al*. XAI enhancing cyber defence against adversarial attacks in industrial applications. *In*: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING APPLICATIONS AND SYSTEMS, 5., 2022, Genova. **Proceedings of the** [...]. [*S. l.*]: IEEE, Dec. 2022. p. 1-8. Disponível em: https://ieeexplore.ieee.org/document/10052858. Acesso em: 12 abr. 2025.

MANSON, K. AI warfare is already here. **Bloomberg Businessweek**, New York, Feb. 2024. Disponível em: https://www.bloomberg.com/features/2024-ai-warfare-project-maven/?embedded-checkout=true. Acesso em: 23 mar. 2025.

MICHEL, A. H. **The Black Box, Unlocked:** Predictability and Understandability in Military AI. Geneva: UNIDIR, Sep. 2020. *E-book*. Disponível em: https://unidir.org/publication/the-black-box-unlocked/. Acesso em: 28 mar. 2025.

NAGAHISARCHOGHAEI, M. *et al.* An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives. **Electronics**, [*s. l.*], v. 12, n. 5, p. 1092, Feb. 2023. Disponível em: https://www.mdpi.com/2079-9292/12/5/1092. Acesso em: 30 mar. 2025.

OSSWALD, F. *et al*. FCAS Ethical AI Demonstrator. *In:* XAI (LATE-BREAKING WORK, DEMOS, DOCTORAL CONSORTIUM), 1., 2023, Belém, Portugal. **Joint Proceedings of the** [...]. [*S. l.*]: CEUR-WS.org, v. 3554, July 2023. p. 152-157. Disponível em: https://ceur-ws.org/Vol-3554/paper27.pdf. Acesso em: 25 mar. 2025.

PARKS, R. C.; DUGGAN, D. P. Principles of Cyberwarfare. **IEEE Security & Privacy**, [*s. l.*], v. 9, n. 5, p. 30-35, Sep. 2011. Disponível em: https://ieeexplore.ieee.org/abstract/document/6029360. Acesso em 14 abr. 2025

PROBASCO, E. **Building the Tech Coalition**: How Project Maven and the U.S. 18th Airborne Corps Operationalized Software and Artificial Intelligence for the Department of Defense. [Washington, D. C.]: Center for Security and Emerging Technology, Aug. 2024. Disponível em: https://cset.georgetown.edu/publication/building-the-tech-coalition/. Acesso em: 27 mar. 2025.

RADANLIEV, P. AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. **Applied Artificial Intelligence**, [*s. l.*], v. 39, n. 1, p. 2463722, Feb. 2025. Disponível em: https://www.tandfonline.com/doi/full/10.1080/08839514.2025.2463722#d1e311. Acesso em: 30 mar. 2025.

SANGER, D. E. In Ukraine, new american technology won the day. Until it was overwhelmed. **New York Times**, New York, Apr. 2024. Disponível em: https://www.nytimes.com/2024/04/23/us/politics/ukraine-new-american-technology.html. Acesso em: 9 abr. 2025.

STIFF, H. **Explainable AI as a Defence Mechanism for Adversarial Examples**. 2019. Dissertação (Mestrado em Ciência da Computação) - School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, 2019. Disponível em: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1355328&dswid=5936. Acesso em: 12 abr. 2025.

UNITED STATES. Department of Defense. **DoD directive 3000.09**: autonomy in weapon systems. Washington, DC: DOD, 2023. Disponível em: https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE- 3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF. Acesso em: 22 mar. 2025.