Uma Abordagem para a Otimização do Aprendizado de Agentes Durante o Sequenciamento de Ataques Cibernéticos em Cenários Desconhecidos e Dinâmicos

Antonio Horta[1], **Ronaldo Goldschmidt**[1], **Ronaldo Salles**[1], **Anderson dos Santos**[1]

[1]Instituto Militar de Engenharia (IME)
Rio de Janeiro – RJ – Brasil

antonio@horta.net.br, {ronaldo.rgold,salles,anderson}@ime.eb.br

Abstract. *In the context of machine learning for agents in cyberattacks, efficiency considers issues related to execution time, resource economy, and training rewards. The state of the art is characterized by the use of reinforcement learning (RL) for training agents in attacks on unknown and dynamic scenarios. However, the exploratory nature of RL can compromise the efficiency of agent training. Therefore, this thesis hypothesizes that training agents to execute cyberattacks in unknown and dynamic scenarios will be more efficient by incorporating the minimization of sequencing failures as the central factor in the RL reward function. To make agent training more efficient in sequencing the techniques that comprise a cyberattack, this thesis proposes the Kill Chain Unscrambler (KCU) algorithm. Its architecture is based on a decision tree logic and incorporates the Kill Chain Catalyst (KCC) module, inspired by genetic alignment, whose function is to prioritize the use of sequences that accelerate RL. Experiments compared the performance of KCU and KCC against the main RL algorithms in a CTF. The results indicated more efficient learning for agents trained using KCU and KCC.*

Resumo. *No contexto do aprendizado de máquina de agentes em ataques cibernéticos, a eficiência considera questões relacionadas ao tempo de execução, à economia de recursos e às recompensas alcançadas nos treinamentos. O estado da arte é caracterizado pelo uso do aprendizado por reforço (RL) para o treinamento de agentes em ataques a cenários desconhecidos e dinâmicos. Entretanto, à natureza exploratória do RL, pode comprometer a eficiência no treinamento de agentes. Diante do exposto, a tese levanta a hipótese de que o treinamento de agentes para a execução de ataques cibernéticos, em cenários desconhecidos e dinâmicos, será mais eficiente ao incorporar a minimização das falhas do sequenciamento como o fator central da função recompensa do RL. Com o objetivo de tornar o treinamento de agentes mais eficiente no sequenciamento de técnicas que compõem um ataque cibernético, a tese propõe o algoritmo de RL Kill Chain Unscrambler (KCU). Sua arquitetura é baseada em uma lógica de árvores de decisão e incorpora o módulo catalisador Kill Chain Catalyst (KCC), inspirado no alinhamento genético, cuja função é priorizar o uso de sequências que acelerem o RL. Experimentos compararam o desempenho do KCU e do KCC, frente aos principais algoritmos de RL em um CTF. Os resultados indicaram um aprendizado mais eficiente para os agentes treinados com uso do KCU e do KCC.*

1. Introdução

No contexto da segurança ofensiva, existem cenários que são desconhecidos pelo atacante, ou seja, sem conhecimento prévio da infraestrutura-alvo. Assim como os cenários dinâmicos, que podem ter seu estado original modificado durante a incursão. Para todos esses cenários, o sequenciamento das técnicas de ataque torna-se um problema complexo para o aprendizado de humanos e agentes. Ao qual, o sucesso de uma incursão depende diretamente da ordem em que as técnicas são aplicadas no ambiente-alvo.

Dentre os paradigmas de aprendizado de máquina para o treinamento de agentes em ataques a cenários desconhecidos e dinâmicos, o estado da arte se caracteriza pelo uso do aprendizado por reforço (RL). Contudo, a eficiência no *RL* é um tema amplamente considerado na literatura, apesar da ausência de uma definição universalmente aceita. No *RL*, os conceitos sobre a eficiência se articulam em torno da complexidade temporal dos algoritmos, da quantidade de amostras necessárias para atingir um desempenho satisfatório, do custo computacional inerente aos processos de aprendizagem e da rapidez na convergência para políticas consideradas ótimas ou subótimas [Sutton and Barto 2018].

Nesse contexto, a complexidade do sequenciamento das técnicas de ataque pode tornar o treinamento de agentes ineficiente, consumindo tempo e recursos computacionais até alcançar seu estado assintótico. A ineficiência também é um fator decorrente da natureza dos agentes de *RL* tenderem a apresentar um número elevado de falhas ao longo do sequenciamento das técnicas que compõem o ataque.

Com foco na hipótese de que a redução das falhas cometidas pelos agentes, ao tentarem utilizar uma técnica não apropriada em determinado momento do sequenciamento do ataque, torna o aprendizado mais eficiente. A tese apresenta uma abordagem para otimizar o *RL*, efetuando um treinamento eficiente dos agentes no sequenciamento das técnicas que compõem um ataque cibernético em cenários desconhecidos e dinâmicos.

A tese[1] propõe o *Kill Chain Unscrambler* (*KCU*), um algoritmo de *RL* para agentes no sequenciamento das técnicas de ataques cibernéticos. Sua arquitetura é baseada no *Gini Impurity-Based Weighted Random Forest* (GIWRF) e no módulo catalisador, *Kill Chain Catalyst* (KCC), que, inspirado no alinhamento genético, prioriza as sequências no treinamento dos agentes de *RL* para acelerar o aprendizado. Além disso, pelo uso de uma lógica de árvores de decisão, em vez de redes neurais, simplifica a interpretabilidade do aprendizado, apresenta robustez em cenários com experiências escassas e resiliência em ambientes dinâmicos.

Foram realizados experimentos com o algoritmo *KCU* e seu módulo *KCC*, frente aos principais algoritmos encontrados no estado da arte no contexto de um torneio *CTF*, considerando cenários desconhecidos, dinâmicos e com experiências escassas para o aprendizado. Os experimentos somam 240 ciclos de ataques, totalizando mais de 360 horas de experimentação.

Os resultados indicam um aprendizado superior para os agentes de *RL* treinados pelo *KCU* com uso do *KCC*, apresentando diferenças de até 205.73% na redução do número de passos, 126.95% nas recompensas obtidas e 1054.31% na redução de falhas em relação aos demais algoritmos.

[1] https://pged.ime.eb.br/tesesedissertacoes

2. Motivação

A tese foi motivada por diversos fatores interconectados que refletem o panorama contemporâneo da segurança cibernética e a ascensão da *IA*. A adoção acelerada da *IA* representa uma importante mudança na capacidade de automação e tomada de decisões em ambientes digitais [Baker et al. 2020]. A *IA*, alimentada por grandes volumes de dados e algoritmos, permite que sistemas tomem decisões de forma cada vez mais rápida e precisa. No entanto, essa mesma capacidade de tomada de decisões pode ser explorada de maneira prejudicial quando utilizada em ataques cibernéticos. A interação entre *IA* e ataques cibernéticos pode amplificar o potencial de danos, criando cenários de ameaças que vão desde ataques altamente sofisticados até a escalada de conflitos cibernéticos em níveis globais. Os ataques cibernéticos podem resultar em danos financeiros substanciais, perda de dados sensíveis e interrupção das operações de negócios. Ainda mais preocupante é o fato de que infraestruturas críticas, como sistemas de energia e saúde, estão cada vez mais sujeitas a ataques cibernéticos, o que pode afetar diretamente a segurança pública [Tavakkoli et al. 2025].

A interseção entre a *IA* e os ataques cibernéticos em cenários desconhecidos e dinâmicos torna-se, assim, um campo de pesquisa crítico. Compreender como a *IA* pode ser treinada tanto para defender quanto para atacar é essencial para a segurança da sociedade civil e a soberania das nações [Mcdonald et al. 2024]. A motivação subjacente a esta pesquisa é a necessidade de desenvolver estratégias e soluções que aprimorem o aprendizado de máquina, de modo a contribuir com a eficiência associada à convergência da *IA* e dos ataques cibernéticos [Zhou et al. 2024].

Apesar da ausência de uma definição universalmente aceita, a eficiência associada a *IA* e aos ataques cibernéticos é um tema amplamente considerado por sua relevância. Pesquisas como as de [Oh et al. 2024], [Chen et al. 2023], [Zhou et al. 2021] e [Wang et al. 2024] mesmo mencionando a eficiência de maneira implícita, associam-na a aspectos como tempo de execução reduzido, economia de recursos computacionais e recompensas alcançadas.

Portanto, essa motivação vai além do ambiente acadêmico. Ela atinge diretamente os interesses da sociedade civil, que busca proteção contra as ameaças digitais crescentes, e das organizações militares e governos, que enfrentam questões complexas para garantir a segurança nacional e a estabilidade geopolítica.

3. Caracterização do Problema

A execução com êxito de ataques cibernéticos realizados em várias etapas, e que formam uma sequência de técnicas de ataque, é conhecida como *kill chain*[2], representando uma tarefa de considerável complexidade. Para alcançar seus objetivos em cenários desconhecidos, os atacantes devem não apenas possuir conhecimento técnico, mas também a capacidade de determinar a ordem e as técnicas a serem empregadas durante o ataque. No entanto, a busca por uma sequência que alcance os objetivos do ataque, particularmente em cenários desconhecidos, é permeada por incertezas e propensa a falhas.

A Figura 1 ilustra a formação de *kill chains* no processo de busca por sequências que alcancem o objetivo. Nessa ilustração, os números nas peças do quebra-cabeça re-

[2]Termo que vem da área militar e define as etapas de um ataque até que ele alcance seus objetivos

presentam as técnicas utilizadas e as cores, as táticas, que encadeadas formam cinco *kill-chains*, uma para cada ataque. Na ilustração, as táticas disponíveis são: *Acesso Inicial*, *Persistência*, *Movimentação Lateral* e *Impacto*. Cada uma delas possui técnicas que executam ações específicas em cada etapa, por exemplo, o *portscan* e a exploração de uma vulnerabilidade na fase *Acesso Inicial*. Como pode ser observado, diferentes combinações podem alcançar o objetivo, no caso da ilustração, o desenho do alvo. Entretanto, devido ao número de sequências possíveis, existem diferentes maneiras de alcançar o objetivo. No entanto, nessa busca, algumas técnicas não são compatíveis com a técnica utilizada anteriormente naquele momento. Esta situação, circulada em vermelho na 3ª *kill chain* entre as peças 14 e 12, representa uma falha na formação da *kill chain* e a complexidade na busca por sequências que solucionem o problema.

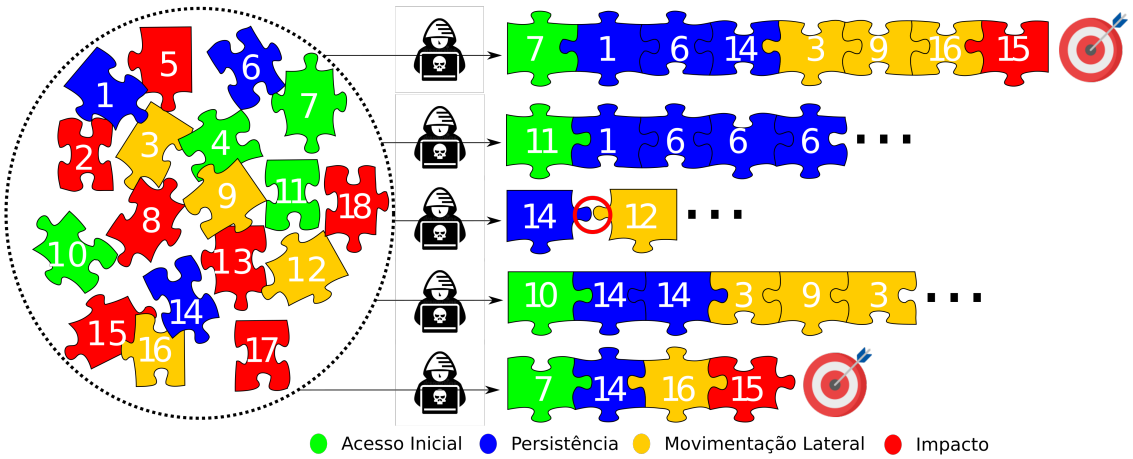

Figura 1. Kill chains formadas pelo sequenciamento de técnicas de ataque.

Dentre os paradigmas de aprendizado de máquina para o treinamento de agentes em ataques a cenários desconhecidos e dinâmicos, o estado da arte se caracteriza pelo uso do *RL*. Para esses tipos de cenário, o *RL* se mostra mais adequado, pois utiliza os retornos das interações com o ambiente como principal fonte de dados para treinamento, dispensando a necessidade de um conjunto de dados pré-definido [Sutton and Barto 2018].

4. Questão de Pesquisa

A tese investiga as estratégias para aprimorar a eficiência do aprendizado por reforço no contexto dos ataques cibernéticos. Em situações de ataque com uso de agentes, em que os ambientes são frequentemente dinâmicos, incertos e pouco conhecidos, a capacidade dos agentes em aprender e tomar decisões ao longo do tempo torna-se uma tarefa complexa. O sequenciamento adequado das técnicas de ataque, desperta a busca por responder à seguinte questão:

Como tornar o aprendizado por reforço mais eficiente durante o treinamento de agentes no sequenciamento das técnicas de um ataque cibernético em cenários desconhecidos e dinâmicos?

5. Hipótese

No aprendizado por reforço, os agentes são guiados pela maximização dos retornos definidos por suas funções recompensa, convergindo para trajetórias com menor número

de passos até alcançar seu estado assintótico. Entretanto, no contexto de um ataque cibernético em um cenário desconhecido, essa abordagem se mostra ineficiente, pois as funções recompensa do estado da arte não consideram diretamente as falhas cometidas durante o sequenciamento das técnicas que compõem a *kill chain*. Portanto, a hipótese consiste em que o treinamento de agentes para execução de ataques cibernéticos, em cenários desconhecidos e dinâmicos, será mais eficiente ao incorporar a minimização das falhas do sequenciamento como o fator central da função recompensa do *RL*.

6. Objetivos

Com foco na minimização das falhas do sequenciamento, o propósito central da tese é: propor uma abordagem para otimização do aprendizado por reforço que produza um treinamento eficiente dos agentes durante o sequenciamento das técnicas que compõem um ataque cibernético em cenários desconhecidos e dinâmicos.

Além disso, a tese inclui os seguintes objetivos específicos:

- Identificar e adotar métricas que serão utilizadas na análise da eficiência do treinamento por *RL* dos agentes durante o sequenciamento das técnicas que compõem um ataque cibernético;
- Desenvolver um método de análise da eficiência do treinamento dos agentes de *RL* durante o sequenciamento das técnicas que compõem um ataque cibernético;
- Desenvolver um algoritmo de *RL* orientado à minimização das falhas para o treinamento de um agente no sequenciamento das técnicas que compõem um ataque cibernético; e
- Projetar e executar experimentos controlados que permitam analisar o desempenho do algoritmo proposto em relação aos principais algoritmos de *RL*, de modo a comprovar a hipótese de que a minimização das falhas de sequenciamento contribui para a melhoria da eficiência no treinamento de agentes em ataques cibernéticos a cenários desconhecidos e dinâmicos.

7. Contribuições e Produções Científicas Associadas

Esta seção detalha as contribuições da tese para o avanço da área de aplicação do *RL* na segurança cibernética ofensiva. A principal delas reside no desenvolvimento de um algoritmo de *RL*, fundamentado na lógica de árvores de decisão e no uso de um módulo catalisador inspirado no alinhamento genético. Este algoritmo visa otimizar o treinamento de agentes para a execução eficiente de sequências de ataques em ambientes desconhecidos e dinâmicos, demonstrando potencial para superar as limitações dos algoritmos de *RL* convencionais. Além do algoritmo, a tese também disponibiliza ferramenta e método para o suporte à exploração e à avaliação da aplicação de *RL* em cenários de ataque, conforme detalhado nos seguintes itens:

- Método para análise da eficiência do treinamento dos agentes de *RL* na execução de ataques a cenários desconhecidos e dinâmicos, com base nas métricas de *offset*, *velocidade* e *generalização* dos critérios de *recompensas*, *passos* e *falhas* nas curvas de aprendizado;
- *Framework* de integração de agentes de *RL* com ambientes computacionais para a execução e análise de ataques em cenários desconhecidos e dinâmicos;

- Algoritmo de aprendizado por reforço baseado em lógica de árvores de decisão com um módulo catalisador, inspirado no alinhamento genético, que, guiado pela minimização das falhas, produz um treinamento mais eficiente para agentes no sequenciamento das técnicas que compõem os ataques cibernéticos em cenários desconhecidos e dinâmicos.

As contribuições supracitadas foram consolidadas progressivamente ao longo do desenvolvimento da tese, em um processo iterativo de investigação, experimentação e validação. O artigo [Horta et al. 2025] publicado no Journal of Computer Security foi o ponto de partida para mapear o problema investigado na tese. Depois, cada etapa do trabalho forneceu subsídios para o refinamento das abordagens propostas, culminando em artefatos científicos com graus crescentes de maturidade e abrangência.

Nesse contexto, foram desenvolvidos o *KCU* e o *KCC*, componentes centrais do *framework* proposto. O artigo [Horta et al. 2024], que apresenta esses componentes, foi avaliado pelo comitê científico do SBSeg 2024 e recebeu o prêmio de melhor artigo completo naquele ano. Um reconhecimento que atesta a relevância e a originalidade das contribuições no contexto nacional. Esse resultado motivou a elaboração de uma versão estendida [Horta et al. 2026], subsequentemente publicada no *Journal of the Brazilian Computer Society* (JBCS), ampliando o alcance das contribuições para o âmbito da literatura científica internacional Adicionalmente, no intuito de fomentar a reprodutibilidade científica e estimular o avanço colaborativo da área, todos os artefatos produzidos no âmbito da tese, foram disponibilizados publicamente por meio do repositório *online*[3].

8. Resultados Obtidos

Ao comparar os quatro principais algoritmos de *RL* no cenário *Dummy* de *CTF* com o *KCC* acoplado ao *KCU* nos experimentos realizados. Os resultados apontam para um aumento significativo na eficiência do *RL* utilizando a abordagem proposta na tese.

A Figura 2 ilustra a eficiência alcançada na redução do número de falhas, hipótese central da tese, desde os estágios iniciais do treinamento com uso do *KCC*.

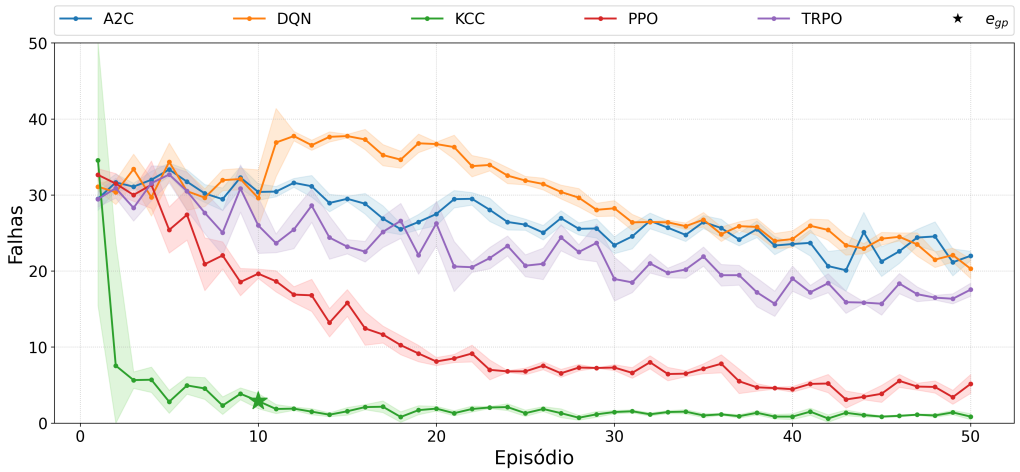

Figura 2. Curvas de aprendizado de falhas dos agentes em cenário estático.

[3]https://gitlab.com/antonio50/exoskeleton

Complementarmente, a Tabela 1 apresenta as diferenças percentuais médias por episódio em cada ensaio, tomando como base os resultados obtidos pelo agente *KCC*. Utilizando lógica baseada em árvores de decisão, em contraste com as redes neurais adotadas pelos demais algoritmos, o *KCC* demonstrou desempenho superior de forma geral para o tipo de problema avaliado. Em média, o agente exigiu 16.17 passos para alcançar o objetivo ao final do ciclo de aprendizagem, enquanto os demais algoritmos registraram aproximadamente 47 passos. Tendência semelhante foi observada na métrica de recompensas, em que o *KCC* obteve média de 91.35, superando o segundo melhor resultado, do algoritmo *TRPO*, com 13.96, e ultrapassando significativamente o pior desempenho, registrado pelo *DQN*, com -24.62. Além disso, ao analisar o número médio de falhas por episódio, as diferenças entre os algoritmos se tornam ainda mais evidentes. O *KCC* apresentou novamente o melhor desempenho, com média de 2.58 falhas por episódio, seguido por *PPO*, *TRPO*, *A2C* e *DQN*, sendo que este último apresentou diferença superior a 1054% em relação ao valor obtido pelo *KCC*.

Tabela 1. Percentuais relativos ao KCC pela média dos episódios.

Série	Passos	Dif Passos	Recompensa	Dif Recompensa	Falhas	Dif Falhas
KCC	**16.17**	**0.00%**	**91.35**	**0.00%**	**2.58**	**0.00%**
A2C	46.89	189.92%	-1.98	-102.17%	26.81	940.68%
DQN	49.45	205.73%	-24.62	-126.95%	29.74	1054.31%
PPO	48.29	198.56%	4.62	-94.94%	11.46	344.68%
TRPO	45.31	180.13%	13.96	-84.72%	22.38	768.87%

9. Considerações Finais e Trabalhos Futuros

A tese investigou a aplicação de algoritmos de *RL* no contexto de ataques cibernéticos em cenários desconhecidos e dinâmicos, partindo da hipótese de que a minimização de falhas no sequenciamento de técnicas representa o fator central para um treinamento mais eficiente de agentes automatizados. Para tanto, foram propostos o algoritmo *KCU*, fundamentado em lógica de árvores de decisão, e o módulo catalisador *KCC*, inspirado no alinhamento genético, além de um método de análise das curvas de aprendizado baseado em métricas objetivas de *offset*, *velocidade* e *generalização*.

Os experimentos, conduzidos em 240 ciclos de ataque ao longo de mais de 360 horas em ambientes do tipo *CTF*, demonstraram um aumento substancial da eficiência dos agentes pelo uso da abordagem proposta frente aos algoritmos tradicionais de *RL*. O agente com o módulo *KCC* foi o único capaz de generalizar e otimizar estratégias em todos os critérios avaliados, com diferenças de até 1054,31% na métrica de falhas em relação aos demais algoritmos, além de demonstrar capacidade de adaptação a modificações dinâmicas no ambiente, como a correção de vulnerabilidades durante o treinamento.

Apesar dos resultados expressivos, a tese reconhece limitações relacionadas à sensibilidade do *offset* à fase exploratória estocástica, à restrição no número de cenários avaliados e à dificuldade de comparação direta com o estado da arte, dado que muitos trabalhos correlatos não disponibilizam publicamente seus códigos e dados. Como perspectivas futuras, destacam-se a incorporação de *transfer learning* com *LLMs* para reduzir

a estocasticidade inicial, a investigação de outros modelos interpretáveis e a ampliação dos experimentos para outros casos de uso.

Referências

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. (2020). Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations (ICLR 2020)*.

Chen, J., Hu, S., Zheng, H., Xing, C., and Zhang, G. (2023). Gail-pt: An intelligent penetration testing framework with generative adversarial imitation learning. *Computers Security*, 126:103055.

Horta, A., dos Santos, A., and Goldschmidt, R. (2026). Enhancing red team agent learning with the kill chain catalyst algorithm in capture the flag scenarios. *Journal of the Brazilian Computer Society*, 32(1):289–304.

Horta, A., dos Santos, A. F. P., and Goldschmidt, R. R. (2025). Evaluating the stealth of reinforcement learning-based cyber attacks against unknown scenarios using knowledge transfer techniques. *Journal of Computer Security*, 33(2):100–115.

Horta, A., Santos, A., and Goldschmidt, R. (2024). Kill chain catalyst for autonomous red team operations in dynamic attack scenarios. In *Anais do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 415–430, Porto Alegre, RS, Brasil. SBC.

Mcdonald, G., Li, L., and Mallah, R. A. (2024). Finding the optimal security policies for autonomous cyber operations with competitive reinforcement learning. *IEEE Access*, 12:120292 – 120305.

Oh, S. H., Kim, J., Nah, J. H., and Park, J. (2024). Employing deep reinforcement learning to cyber-attack simulation for enhancing cybersecurity. *Electronics (Switzerland)*, 13(3).

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: an introduction*. MIT press, second edition.

Tavakkoli, N., Çetin, O., Ekmekcioglu, E., and Savaş, E. (2025). From frontlines to online: examining target preferences in the russia-ukraine conflict. *International Journal of Information Security*, 24(1):1–15.

Wang, C., Redino, C., Clark, R., Rahman, A., Aguinaga, S., Murli, S., Nandakumar, D., Rao, R., Huang, L., Radke, D., and Bowen, E. (2024). Leveraging reinforcement learning in red teaming for advanced ransomware attack simulations. In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 262–269.

Zhou, S., Liu, J., Hou, D., Zhong, X., and Zhang, Y. (2021). Autonomous penetration testing based on improved deep q-network. *Applied Sciences*, 11(19).

Zhou, Z., Zhou, T., Xu, J., and Zhu, J. (2024). Efficient penetration testing path planning based on reinforcement learning with episodic memory. *CMES - Computer Modeling in Engineering and Sciences*, 140(3):2613 – 2634.