
Universidade de Brasília - UNB

Faculdade de Tecnologia - FT

Departamento de Engenharia Elétrica - ENE

Programa de Pos-Graduação em Engenharia Elétrica - PPEE


# Pré-Projeto de Dissertação para Processo de Seleção ao Mestrado Profissional


Paulo Matheus Nicolau Silva


**Área de concentração:** Segurança Cibernética

**Linha de pesquisa:** Segurança e inteligência Cibernética.

**Tema:** Arquitetura de Governança para Agentes de Inteligência Artificial na Busca Proativa por Ameaças Cibernéticas.


Brasília 2025

# 1 Introdução

O meio cibernético configura-se como um campo de batalha digital, onde a evolução constante das ameaças, aliada à transformação tecnológica, exige respostas cada vez mais sofisticadas. Ataques de ransomware, invasões por phishing e explorações de vulnerabilidades estão se tornando mais frequentes e complexos; por isso, as organizações são obrigadas a repensar suas estratégias de defesa [Valencia 2024].

A busca proativa por ameaças, ou threat hunting, é uma prática que vai além da reatividade dos sistemas tradicionais. Em vez de aguardar que um alerta seja gerado, os profissionais de segurança adotam técnicas de investigação forense para examinar redes, sistemas e computadores em busca de sinais de comprometimento [Sindiramutty 2023]. Essa abordagem ativa permite que vulnerabilidades sejam identificadas e neutralizadas antes que sejam exploradas por atacantes.

Agentes de inteligência artificial são estruturas de software que ampliam as capacidades dos grandes modelos de linguagem (LLMs — Large Language Models), permitindo que estes interajam com diversas ferramentas e executem ações no sistema. Ao utilizar técnicas de reinforcement learning, é possível fazer com que esses agentes evoluam com base em recompensas, aprimorando sua capacidade de identificar atividades suspeitas e reduzir falsos positivos [Valencia 2024].

Da mesma forma, a governança em segurança cibernética é um pilar fundamental para alinhar as estratégias de defesa com os objetivos e necessidades do negócio. Portanto, faz-se necessário um processo robusto de governança que atue como orquestrador de toda a operação, coordenando os esforços dos agentes de inteligência artificial, definindo políticas de segurança e garantindo a conformidade [Oesch et al. 2024].

Dado que os outputs das redes neurais são, por definição, estocásticos, é essencial a implementação de um mecanismo de feedback, no qual os relatórios de atividades suspeitas sejam submetidos à análise de especialistas. Esse canal de comunicação possibilita que os achados sejam validados e ajustados com base no conhecimento prático dos profissionais, promovendo o aprimoramento contínuo dos agentes de inteligência artificial e a redução de falsos positivos [Sindiramutty 2023].

Além disso, um recente estudo de McIntosh et al. [McIntosh et al. 2024] avaliou a prontidão de frameworks de governança de segurança cibernética — como COBIT, NIST CSF, ISO 27001 e o novo ISO 42001 — fornecendo uma análise comparativa sobre oportunidades, riscos e conformidade regulatória na comercialização de grandes modelos de linguagem. Os autores destacam que os frameworks existentes precisam evoluir para acompanhar os riscos específicos gerados pela integração de tecnologias de IA.

Em síntese, este trabalho propõe a pesquisa e a análise de modelos de governança capazes de sustentar a criação de uma solução de busca proativa por ameaças cibernéticas em ambientes Windows, utilizando agentes de inteligência artificial especializados. Ao explorar frameworks de governança e integrar técnicas avançadas de threat hunting com mecanismos de feedback contínuo, busca-se estabelecer diretrizes que alinhem as estratégias

de segurança aos objetivos da organização, promovendo uma atitude proativa e eficaz frente à complexidade das ameaças atuais. Dessa forma, a proposta visa contribuir para o desenvolvimento de sistemas de defesa cibernética mais resilientes e conformes com as exigências regulatórias, reforçando o papel da governança na proteção dos ativos digitais.

# 2 Justificativa

Em um cenário marcado por ataques de ransomware, invasões por phishing e exploração de vulnerabilidades, torna-se imperativo desenvolver soluções que não apenas respondam reativamente a incidentes, mas que também antecipem ameaças cibernéticas.

A integração de agentes inteligentes e técnicas de reinforcement learning possibilita a criação de sistemas de defesa adaptáveis, capazes de identificar padrões maliciosos e reduzir significativamente a incidência de falsos positivos [Valencia 2024]. Além disso, conforme discutido por Oesch et al. [Oesch et al. 2024], a construção de um framework robusto de governança em segurança cibernética é essencial para coordenar as ações dos agentes de IA e assegurar a conformidade com as normas regulatórias. McIntosh et al. [McIntosh et al. 2024] complementam esse argumento ao enfatizar que os modelos tradicionais de governança necessitam evoluir para incorporar as especificidades e os riscos associados à integração de tecnologias de inteligência artificial, reforçando a urgência de investir em soluções inovadoras.

A escolha da área de concentração "Segurança e Inteligência Cibernética", dentre as opções previstas no edital, reflete a necessidade de aprofundar o estudo e o desenvolvimento de tecnologias que permitam a antecipação de ameaças cibernéticas. Ao optar por essa linha de pesquisa, o presente projeto visa não apenas contribuir para o avanço teórico na área, mas também oferecer respostas práticas aos desafios enfrentados por ambientes operacionais críticos, como os sistemas baseados em Windows. Essa escolha justifica-se pela alta demanda por soluções que integrem aspectos tecnológicos, humanos e regulatórios, combinando métodos de threat hunting com frameworks de governança capazes de validar e aprimorar as respostas automatizadas. Dessa forma, o projeto alinha-se com os objetivos estratégicos estabelecidos no PPEE, promovendo uma abordagem interdisciplinar que une inteligência artificial, engenharia de sistemas e governança corporativa para fortalecer a segurança cibernética contemporânea.

# 3 Objetivos

## 3.1 Objetivo Geral

O objetivo geral do projeto é criar uma solução que integre agentes de inteligência artificial e técnicas avançadas de threat hunting para a busca proativa de ameaças cibernéticas em ambientes Windows, fundamentada em modelos de governança.

## 3.2 Objetivos Específicos

Para atingir o objetivo geral, os seguintes objetivos específicos foram estabelecidos:

1. Realizar uma revisão bibliográfica sobre governança em segurança cibernética, agentes de inteligência artificial e técnicas de busca proativa de ameaças em ambientes Windows.

2. Desenvolver agentes de inteligência artificial customizados, capacitando-os a executar ações especializadas no processo de threat hunting.

3. Desenvolver um modelo de governança capaz de orquestrar a cooperação entre os agentes.

4. Integrar interfaces com especialistas humanos ao processo de busca proativa por ameaças cibernéticas.

5. Validar a eficácia da solução proposta na detecção proativa em ambientes Windows, comparando os resultados com outros métodos.

# 4 Revisão da Literatura

O marco teórico que revolucionou o processamento de sequências em redes neurais foi estabelecido por Vaswani et al. [Vaswani et al. 2017], ao introduzirem o mecanismo de atenção que, ao dispensar estruturas recorrentes, possibilitou o desenvolvimento de grandes modelos de linguagem (LLMs). Em continuidade a essa abordagem, Geiping et al. [Geiping et al. 2025] propuseram métodos para escalonar o uso de recursos computacionais em tempo de inferência por meio de raciocínio latente, ampliando a eficiência dos grandes modelos de linguagem na resolução de problemas complexos.

No mesmo âmbito, a contribuição de Agashe et al. [Agashe et al. 2023] avalia as habilidades de coordenação entre grandes modelos de linguagem, demonstrando o potencial desses modelos para realizar cooperação e tomada de decisão conjunta. Complementarmente, o trabalho de Kirshteyn [Kirshteyn 2024] discute frameworks para agentes de IA, enfatizando tanto os aspectos técnicos quanto as aplicações práticas em cenários multiagente.

Paralelamente, a segurança cibernética tem se beneficiado de abordagens avançadas de threat hunting, que visam a identificação proativa e a mitigação de ataques. Hillier & Karroubi [Hillier e Karroubi 2022] discutem o conceito de transformar a postura defensiva, convertendo sistemas de "caça" em sistemas de "caçadores" proativos, abordando o ciclo de vida e os desafios de um ecossistema ameaçador. Ali & Kostakos [Ali e Kostakos 2023] ampliam essa discussão ao integrar técnicas de machine learning e modelos explicáveis, exemplificados pelo HuntGPT, que combina a detecção de anomalias com a inteligência explicável dos grandes modelos de linguagem.

Trabalhos como o de Yi & Kim [Yi e Kim 2024] propõem modelos baseados na geração de hipóteses para o threat hunting, evidenciando uma mudança paradigmática em relação às abordagens tradicionais de detecção de ameaças. O modelo proposto por esses autores enfatiza a construção de hipóteses fundamentadas em dados de inteligência de ameaças, o que permite identificar, de maneira proativa, atividades suspeitas que, de outra forma, poderiam passar despercebidas.

Outros estudos, como o da ISACA [ISACA 2025] sobre governança de sistemas de IA, ressaltam a necessidade de frameworks que unam as melhores práticas de gestão com os avanços técnicos apresentados nos modelos colaborativos e na segurança cibernética. Assim, a integração de mecanismos de atenção e coordenação, juntamente com estratégias robustas de detecção e mitigação de ameaças, configura uma proposta inovadora e multidisciplinar.

Ao compilar essas referências, o presente pré-projeto propõe a exploração de novas abordagens que possibilitem a sinergia entre a coordenação inteligente dos agentes e a eficiência na busca proativa por ameaças cibernéticas.

# 5 Metodologia

Este projeto de pesquisa adotará uma abordagem metodológica estruturada e sistemática, iniciando com uma revisão bibliográfica exaustiva que abordará os fundamentos teóricos da governança em segurança cibernética, as arquiteturas de agentes de inteligência artificial e as técnicas de busca proativa de ameaças em ambientes Windows. Essa revisão será fundamentada em fontes acadêmicas de alto rigor e publicações de referência, permitindo a identificação de lacunas na literatura e uma compreensão aprofundada dos desafios e avanços atuais na área.

Posteriormente, a pesquisa adotará uma abordagem mista, combinando métodos quantitativos e qualitativos para a coleta e análise de dados. Na vertente quantitativa, serão realizadas simulações em ambientes laboratoriais controlados, onde as arquiteturas de agentes de IA serão submetidas a cenários diversos de ameaças cibernéticas, possibilitando a avaliação estatística e a mensuração do desempenho por meio de indicadores precisos. Simultaneamente, a vertente qualitativa incluirá estudos de caso e entrevistas com especialistas da área de segurança cibernética, os quais contribuirão para a compreensão das implicações práticas das soluções propostas e para a identificação de oportunidades de aprimoramento das abordagens teóricas.

Os dados obtidos serão submetidos a análises comparativas com outras abordagens, buscando demonstrar a relevância e a adaptabilidade da solução proposta. Dessa forma, a integração dos resultados quantitativos e qualitativos fornecerá uma visão abrangente e robusta sobre a eficácia das arquiteturas, contribuindo de maneira significativa para o avanço do conhecimento na área de segurança cibernética.

# 6 Plano de Trabalho

O plano de trabalho está organizado em seis etapas:

1. **Cursar Disciplinas Obrigatórias e Eletivas:** Realizar os cursos das disciplinas obrigatórias e eletivas do Mestrado Profissional em Engenharia Elétrica com habilitação em Segurança Cibernética.

2. **Revisão Bibliográfica e Levantamento de Requisitos:** Pesquisar em bases acadêmicas sobre Deep Reinforcement Learning, redes de transformers, enxames de agentes e governança ágil.

3. **Definição e Projeto da Arquitetura:** Elaborar o fluxo integrado de monitoramento e mitigação, definindo os processos de coleta de dados, formulação de hipóteses e mecanismos de cooperação.

4. **Desenvolvimento e Integração:** Codificar e integrar a arquitetura proposta.

5. **Testes e Validação:** Realizar experimentos em ambiente controlado para testes e validação.

6. **Redação e Finalização da Dissertação:** Consolidar os resultados, discutir as contribuições e limitações, e redigir o documento final.

# 7 Cronograma

A seguir, apresenta-se o cronograma de execução das atividades, considerando as disciplinas obrigatórias e eletivas do Mestrado Profissional em Engenharia Elétrica com habilitação em Segurança Cibernética:

| **Atividade / Disciplina** | **1º Sem.** | **2º Sem.** | **3º Sem.** | **4º Sem.** |
|---|---|---|---|---|
| Cursar Disciplinas Obrigatórias e Eletivas | X | X | X | |
| Revisão Bibliográfica e Levantamento de Requisitos | X | | | |
| Definição e Projeto da Arquitetura | | X | | |
| Desenvolvimento e Integração | | X | X | |
| Testes e Validação | | X | X | |
| Redação e Finalização da Dissertação | | X | X | X |

Tabela 1 – Cronograma de Execução de Atividades e Disciplinas

# Referências


Agashe et al. 2023 AGASHE, S. et al. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*, Oct 2023. Version 2, revised Apr 2024. Disponível em: <https://doi.org/10.48550/arXiv.2310.03903>.

Ali e Kostakos 2023 ALI, T.; KOSTAKOS, P. *HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs)*. 2023. Disponível em: <https://arxiv.org/abs/2309.16021>.

Geiping et al. 2025 GEIPING, J. et al. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171v1*, 2025. Disponível em: <https://doi.org/10.48550/arXiv.2502.05171>.

Hillier e Karroubi 2022 HILLIER, C.; KARROUBI, T. *Turning the Hunted into the Hunter via Threat Hunting: Life Cycle, Ecosystem, Challenges and the Great Promise of AI*. 2022. Disponível em: <https://arxiv.org/abs/2204.11076>.

ISACA 2025 ISACA. *Leveraging COBIT for Effective AI System Governance*. 2025. White Paper, ISACA. Accessed on February 14, 2025. Disponível em: <https://www.isaca.org/resources/white-papers/2025/leveraging-cobit-for-effective-ai-system-governance>.

Kirshteyn 2024 KIRSHTEYN, P. M. E. *AI Agent Frameworks: Technical Insights and Practical Applications*. [S.l.]: Independently Published, 2024. Hardcover, 288 pages. ISBN 979-8332582653.

McIntosh et al. 2024 MCINTOSH, T. R. et al. From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models. *arXiv preprint*, 2024. Disponível em: <https://doi.org/10.48550/arXiv.2402.15770>.

Oesch et al. 2024 OESCH, S. et al. The path to autonomous cyber defense. *arXiv preprint*, 2024. Disponível em: <https://doi.org/10.48550/arXiv.2404.10788>.

Sindiramutty 2023 SINDIRAMUTTY, S. R. Autonomous threat hunting: A future paradigm for ai-driven threat intelligence. *arXiv preprint*, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2401.00286>.

Valencia 2024 VALENCIA, L. J. Artificial intelligence as the new hacker: Developing agents for offensive security. *arXiv preprint*, 2024. Disponível em: <https://doi.org/10.48550/arXiv.2406.07561>.

Vaswani et al. 2017 VASWANI, A. et al. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. Disponível em: <https://doi.org/10.48550/arXiv.1706.03762>.

Yi e Kim 2024 YI, C.-G.; KIM, Y.-G. Hypothesis generation model for cyber threat hunting. *IEEE Communications Magazine*, v. 62, n. 10, p. 110–116, 2024. Disponível em: <https://doi.org/10.1109/MCOM.001.2300224>.