UNIVERSIDADE FERNANDO PESSOA

Cibercrime e IA: Desafios Legais e Resposta Institucional no Séc XXI

[Cybercrime and AI: Legal Challenges and Institucional Response in the XXI Century]

Projeto de Graduação

Licenciatura em Criminologia

Bernardo Velasquez Borges nº 2022122410

Orientador:

Professor Doutor Pedro Fernando Santos Silva da Cunha

Junho 2025

Cibercrime e IA: Desafios Legais e Resposta Institucional do Séc. XXI

[Cybercrime and AI: Legal Challenges and Institutional Response in the 21st Century]

Projeto de Graduação

[Licenciatura em Criminologia]

Bernardo Velasquez Borges n.º 2022122410

Orientador:

Professor Doutor Pedro Cunha

Junho 2025

Agradecimentos

Ao concluir este trabalho, não posso deixar de expressar a minha profunda gratidão a todas as pessoas que, de uma forma ou de outra, estiveram ao meu lado ao longo deste percurso.

À minha mãe, pelo apoio incondicional e diário que sempre me deu e que sei que continuará a dar. A sua força e presença constante foram fundamentais em cada etapa.

Ao meu pai, por me desafiar diariamente a ser melhor e por acreditar sempre nas minhas capacidades, mesmo quando eu próprio duvidava.

Aos meus irmãos, que, apesar das habituais discussões e brincadeiras, são e sempre serão os meus melhores amigos.

Às minhas colegas e amigas Inês, Eva e Lara — este caminho teria sido impensável sem a vossa companhia, incentivo e amizade. Fizeram parte de tudo e isso ficará para sempre comigo.

Ao Marco, que, apesar de ter estado apenas por um breve período na minha vida, rapidamente se tornou um dos meus melhores amigos. A tua presença foi significativa e marcante.

À Michelle, que, mesmo tendo-se juntado à nossa família há pouco tempo, conseguiu deixar uma marca positiva em todos nós — especialmente no meu pai.

E por fim, mas de forma alguma menos importante, ao Professor Doutor Pedro Cunha, por toda a paciência, orientação e disponibilidade que teve comigo, sobretudo na elaboração deste trabalho. A sua ajuda foi mais do que eu alguma vez poderia esperar — ou merecer.

A todos, o meu mais sincero obrigado.

Resumo

Este trabalho tem como objetivo principal analisar a eficácia das ferramentas de cibersegurança baseadas em inteligência artificial (IA) na proteção de sistemas operativos contra ciberataques cada vez mais sofisticados. Inicialmente, o estudo define e contextualiza o conceito de cibercrime, entendido como qualquer atividade ilícita realizada por meio de sistemas informáticos ou redes digitais. Exemplos comuns de cibercrime incluem ataques de *phishing*, *ransomware*, fraudes financeiras e invasões de sistemas, muitos dos quais têm vindo a evoluir com o uso crescente da IA, tanto por parte dos atacantes quanto dos defensores.

A investigação aborda ainda o conceito de inteligência artificial, explicando as suas principais técnicas, como *machine learning* e *deep learning*, que têm revolucionado a deteção e resposta a ameaças digitais. O estudo destaca exemplos de ataques que utilizam IA, tais como a criação de *deepfakes* para engenharia social, *malware* adaptativo que modifica o seu comportamento para evitar a deteção e ataques automatizados que exploram vulnerabilidades em tempo real.

Para avaliar a eficácia das ferramentas de cibersegurança com IA, o trabalho propõe uma metodologia mista que envolve testes controlados em sistemas operativos variados (Windows, Linux, macOS), com a colaboração de *white hat hackers* que simulam ataques reais e sofisticados. As ferramentas analisadas incluem soluções comercialmente relevantes como FortiEDR, CrowdStrike, SentinelOne e Microsoft Defender for Endpoint.

A recolha de dados integra questionários e entrevistas com profissionais da área, garantindo anonimato e confidencialidade por meio de termos de consentimento informados. Este método permite uma avaliação quantitativa e qualitativa da capacidade destas ferramentas em identificar, bloquear e mitigar ataques, bem como na gestão de falsos positivos e negativos.

Os resultados esperados visam identificar os pontos fortes e limitações das tecnologias atuais, contribuindo para a formulação de melhores práticas e recomendações para a aplicação ética e eficaz da IA em cibersegurança. O estudo reconhece desafios inerentes, como a diversidade dos ambientes testados e a complexidade crescente dos ataques, que podem afetar a performance das soluções avaliadas.

Palavras-chave: Cibercrime; Inteligência Artificial; Cibersegurança; Ferramentas de Proteção com IA; Avaliação de Sistemas de Segurança.

Abstract

This work aims to analyze the effectiveness of cybersecurity tools based on artificial intelligence (AI) in protecting operating systems against increasingly sophisticated cyberattacks. Initially, the study defines and contextualizes the concept of cybercrime, understood as any illicit activity carried out through computer systems or digital networks. Common examples of cybercrime include phishing attacks, ransomware, financial fraud, and system intrusions, many of which have evolved with the growing use of AI, both by attackers and defenders.

The research also addresses the concept of artificial intelligence, explaining its main techniques, such as machine learning and deep learning, which have revolutionized threat detection and response. The study highlights examples of AI-driven attacks, such as the creation of deepfakes for social engineering, adaptive malware that changes its behavior to avoid detection, and automated attacks that exploit vulnerabilities in real time.

To evaluate the effectiveness of AI-based cybersecurity tools, the work proposes an experimental methodology involving controlled tests on various operating systems (Windows, Linux, macOS), in collaboration with white hat hackers who simulate real and sophisticated attacks. The tools analyzed include commercially relevant solutions such as FortiEDR, CrowdStrike, SentinelOne, and Microsoft Defender for Endpoint.

Data collection includes questionnaires and interviews with professionals in the field, ensuring anonymity and confidentiality through informed consent agreements. This method allows for both quantitative and qualitative evaluation of these tools' ability to detect, block, and mitigate attacks, as well as manage false positives and negatives.

The expected results aim to identify the strengths and limitations of current technologies, contributing to the development of best practices and recommendations for the ethical and effective application of AI in cybersecurity. The study acknowledges inherent challenges, such as the diversity of the tested environments and the increasing complexity of attacks, which may affect the performance of the evaluated solutions.

Keywords: Cybercrime; Artificial Intelligence; Cybersecurity; AI-Powered Protection Tools; Security Systems Assessment.

Índice

Lista de abreviaturas

AOL - America Online

CEO - Chief Executive Officer (Diretor Executivo)

CFO - Chief Financial Officer (Diretor Financeiro)

CNNs - Convolutional Neural Networks (Redes Neuronais Convolucionais)

CP - Código Penal

DDoS - Distributed Denial of Service (Negação de Serviço Distribuída)

GANs - Generative Adversarial Networks (Redes Geradoras Adversariais)

IA - Inteligência Artificial

IA Gen - Generative AI (Inteligência Artificial Generativa)

IDS - Intrusion Detection Systems (Sistemas de Deteção de Intrusões)

IEEE - Institute of Electrical and Electronics Engineers

ISP - Internet Service Provider (Fornecedor de Serviços de Internet)

ML - Machine Learning (Aprendizagem Automática)

MTTD - Mean Time to Detect (Tempo Médio para Detectar)

MTTR - Mean Time to Respond / Repair (Tempo Médio para Responder / Resolver)

NCSC - National Cyber Security Centre

RNNs - Recurrent Neural Networks (Redes Neuronais Recorrentes)

SAET - Serviço de Atendimento e Encaminhamento Temático

SaaS - Software as a Service (Software como Serviço)

SMS - Short Message Service (Serviço de Mensagens Curtas)

TIC - Tecnologias da Informação e Comunicação

UFP - Universidade Fernando Pessoa

UNC3T - Unidade Nacional de Combate ao Cibercrime e à Criminalidade Tecnológica

Introdução

A criminalidade, tal como outras realidades sociais, não existe num sistema fechado. É um fenómeno dinâmico que evolui em função das transformações sociais, culturais, económicas e tecnológicas. Ao longo das últimas décadas, a rápida digitalização da sociedade provocou profundas alterações na forma como vivemos, comunicamos, trabalhamos e interagimos. Com estas mudanças, surgiram também novas formas de criminalidade, entre as quais se destaca o cibercrime - um tipo de crime que utiliza as tecnologias da informação e da comunicação como meio, instrumento ou alvo da conduta criminosa.

O cibercrime caracteriza-se pela sua natureza transnacional, pelo elevado grau de sofisticação técnica, pela dificuldade de identificação dos agentes e pela constante mutação dos métodos utilizados. Engloba condutas como o acesso ilegítimo a sistemas informáticos, a divulgação não autorizada de dados, fraudes eletrónicas, *ciberbullying*, pornografia infantil *online*, entre outras. A sua expansão tem vindo a representar um desafio significativo para os sistemas jurídicos e institucionais dos Estados.

Em Portugal, o aumento dos casos de criminalidade informática tem exigido respostas cada vez mais eficazes por parte das instituições judiciárias e das forças de segurança, como a Polícia Judiciária e, em particular, a Unidade Nacional de Combate ao Cibercrime e à Criminalidade Tecnológica (UNC3T). Paralelamente, verifica-se a necessidade de adaptação e modernização do quadro legal, nomeadamente da Lei n.º 109/2009 (Lei do Cibercrime), de forma a acompanhar a evolução constante dos riscos tecnológicos e a complexidade das ameaças digitais.

Neste trabalho, pretende-se analisar esta nova realidade criminal à luz do contexto português, abordando os principais desafios legais e a resposta institucional ao cibercrime. Será dada particular atenção ao papel do criminólogo enquanto agente fundamental no estudo, prevenção e combate ao fenómeno, integrando conhecimentos interdisciplinares que vão desde o direito até à sociologia, psicologia e tecnologia.

Adicionalmente, será considerada a influência do desenvolvimento da Inteligência Artificial (IA), que representa simultaneamente uma ferramenta para o combate ao crime digital e um novo vetor de risco, uma vez que pode ser utilizada para fins criminosos de forma autónoma e altamente eficaz.

Este trabalho tem como objetivos principais avaliar o desempenho de sistemas de cibersegurança baseados em inteligência artificial na deteção, resposta e mitigação de ciberataques em diferentes sistemas operativos.

Como objetivos específicos este trabalho pretende:

- Caracterizar e diferenciar os principais tipos de cibercrime;
- Investigar de que forma a Inteligência Artificial, em especial a IA generativa, está a ser utilizada para potenciar cibercrimes;
- Analisar comparativamente o desempenho de diferentes ferramentas de cibersegurança baseadas em inteligência artificial;
- Identificar os algoritmos e abordagens de IA mais eficazes em contexto de cibersegurança operacional.

A metodologia adotada será mista, combinando abordagens qualitativas e quantitativas, de forma a assegurar uma análise abrangente e rigorosa do objeto de estudo.

Do ponto de vista qualitativo, será realizada uma análise documental e legislativa, contemplando normas jurídicas nacionais e internacionais, relatórios institucionais, jurisprudência relevante e literatura científica atualizada sobre cibersegurança, inteligência artificial e sistemas operativos.

Paralelamente, será seguida uma abordagem quantitativa, com a aplicação de questionários estruturados dirigidos a profissionais da área da cibersegurança (ex.: engenheiros de segurança informática, analistas de redes, administradores de sistemas). Este instrumento visa recolher dados empíricos sobre a perceção, utilização e avaliação técnica de ferramentas de cibersegurança baseadas em IA, permitindo o tratamento estatístico de resultados em formato agregado.

Capítulo 1 – Enquadramento Teórico e Legal

Definição de Cibercrime

Um dos maiores problemas com o qual os académicos se deparam é a inexistência de uma definição consistente de cibercrime (Yar, 2006).

A definição do cibercrime revela-se complexa desde o início, devido à multiplicidade de expressões que são utilizadas para designar um conjunto diversificado de práticas ilícitas no ambiente digital. Termos como cibercrime, crime informático, crime digital ou crime relacionado com o computador são frequentemente usados de forma intercambiável, embora possam apresentar nuances distintas conforme o contexto ou a área de estudo. Esta diversidade terminológica reflete a natureza multifacetada do fenómeno, que engloba diferentes tipos de infrações que têm em comum a utilização de tecnologias digitais como meio, alvo ou instrumento.

Além das variadas denominações, a classificação dos cibercrimes apresenta igualmente uma grande diversidade. Diversos autores e organismos tentam estabelecer tipologias para organizar e compreender melhor os diferentes tipos de crimes inseridos nesta categoria, o que evidencia a dificuldade em encontrar um consenso universal. Estas tipologias variam consoante o foco da análise, podendo distinguir-se, por exemplo, entre crimes contra sistemas informáticos, crimes facilitados pelo uso da internet, crimes contra a propriedade ou contra a pessoa através de meios digitais.

Esta multiplicidade de conceitos e classificações evidencia a complexidade intrínseca ao cibercrime e a necessidade de uma definição abrangente e clara, que possa ser utilizada como base para a análise criminológica e jurídica.

Furnell (2002) propõe uma distinção importante dentro do universo do cibercrime, ao separar as infrações em duas grandes categorias: ofensas assistidas pelo computador e ofensas focadas no computador. Esta diferenciação ajuda a compreender como as tecnologias digitais atuam de formas distintas na prática criminosa.

A primeira categoria, as ofensas assistidas pelo computador, refere-se a crimes que já existiam antes do advento da internet, mas que encontraram no ambiente digital um novo meio de expressão, difusão ou aperfeiçoamento. Embora não dependam exclusivamente das tecnologias informáticas, estas ofensas são facilitadas ou potenciadas por elas. Entre os exemplos mais comuns estão:

- Os delitos contra a propriedade intelectual (como pirataria de *software*, filmes ou música);
- Diversas formas de fraude online, como vendas de produtos inexistentes, esquemas de *phishing* ou burlas com cartões de crédito;
- E formas de vitimação interpessoal, como o *cyberstalking* e o *cyberbullying*, que utilizam plataformas digitais para hostilizar, perseguir ou intimidar vítimas.

Por outro lado, as ofensas focadas no computador dizem respeito a crimes que têm como alvo direto os próprios sistemas informáticos e que só são possíveis devido à existência da tecnologia digital. Estes incluem práticas como:

- A introdução de software malicioso (*malware*);
- Os ataques a redes informáticas (*hacking*);
- A manipulação de dados com o objetivo de comprometer a segurança ou o funcionamento de sistemas eletrónicos.

Wall (2007), agrupa os cibercrimes em três grandes categorias: **crimes relativos à integridade dos computadores**, **ofensas assistidas pelo computador** e **crimes relacionados com o conteúdo**.

A primeira categoria, **crimes relativos à integridade dos computadores**, diz respeito às ofensas que têm como alvo direto a infraestrutura tecnológica que sustenta a *Internet*. Estes crimes afetam o funcionamento de redes, sistemas e dispositivos informáticos, comprometendo o hardware, o *software* ou os dados neles contidos. Exemplos típicos incluem práticas como o *hacking*, o lançamento de software malicioso (*malware*), a distribuição de vírus e os ataques de negação de serviço (DDoS).

Já as **ofensas assistidas pelo computador** referem-se a crimes que, embora tenham existido antes da era digital, encontraram na tecnologia uma nova forma de expressão e disseminação. Estes crimes utilizam os meios informáticos como ferramentas, mas não dependem exclusivamente deles para a sua execução. Entre os exemplos mais comuns estão a fraude eletrónica, a venda de produtos falsificados ou inexistentes, o roubo de identidade, bem como formas de vitimação interpessoal como o *cyberstalking* e o *cyberbullying*. Estas práticas visam bens, serviços e direitos legalmente protegidos, tais como a propriedade intelectual ou a informação pessoal de indivíduos.

Por fim, os **crimes relacionados com o conteúdo** concentram-se na natureza das mensagens ou materiais partilhados através da Internet. Aqui, o foco não está na tecnologia em si, mas no conteúdo comunicacional considerado socialmente nocivo ou ilegal. Esta categoria inclui a disseminação de imagens de abuso sexual infantil, a circulação de conteúdos violentos ou explícitos, a apologia ao ódio racial, étnico ou religioso, e outras expressões que podem incitar à discriminação ou violência.

Esta tipologia proposta por Wall revela-se particularmente útil, pois permite distinguir entre crimes que:

- têm como alvo direto os sistemas informáticos;
- utilizam o computador como meio auxiliar;
- ou se concentram na mensagem veiculada através do meio digital.

Papel da Comissão Europeia na construção da definição de Cibercrime

No contexto europeu, a Comissão Europeia desempenhou um papel fundamental na conceptualização e combate ao cibercrime. Em 2007, publicou o documento "*Towards a General Policy on the Fight Against Cybercrime*" (EUR-Lex, 2007), que visava estabelecer as bases para uma política coordenada de combate a este fenómeno emergente. Neste relatório, é apresentada uma definição institucional de cibercrime, entendendo-o como:

"Os atos cometidos através da utilização de redes de comunicações eletrónicas e de sistemas de informação ou contra estas redes e sistemas."

Esta definição revela a preocupação da União Europeia em abranger tanto os crimes perpetrados através da Internet, como os crimes que visam diretamente a sua infraestrutura tecnológica.

O relatório propõe ainda uma divisão tripartida dos cibercrimes, com base na sua natureza e grau de dependência das tecnologias digitais:

1. **Formas tradicionais de atividade criminal que usam a Internet como meio:** crimes que já existiam antes da era digital, mas que utilizam o ciberespaço para serem cometidos de forma mais eficaz ou alargada. Exemplos incluem fraude, falsificação, e furto de identidade.

2. **Publicação e disseminação de conteúdos ilícitos:** crimes relacionados com o conteúdo da comunicação online, como a divulgação de material terrorista, mensagens racistas ou xenófobas, e pornografia infantil.
3. **Crimes exclusivos das redes eletrónicas:** delitos que surgiram especificamente com o desenvolvimento da Internet, cujos alvos são sistemas e redes informáticos, e que dependem da utilização de técnicas sofisticadas como o lançamento de *malwares*, ataques a servidores ou espionagem digital.

Esta última categoria encontra-se claramente enquadrada no ordenamento jurídico português através da Lei n.º 109/2009, de 15 de setembro, conhecida como Lei do Cibercrime.

Diferentes Tipos de Cibercrime

Esta secção visa proceder a uma análise mais aprofundada de determinadas formas de cibercrime, em particular o *hacking*, o furto de identidade digital, o *cyberstalking* e a pornografia infantil.

Hacking

O *hacking*, enquanto prática associada ao acesso indevido a sistemas informáticos, representa uma das formas mais conhecidas e estudadas de cibercrime. Embora o termo tenha sido originalmente utilizado para descrever a resolução criativa de problemas informáticos, ao longo do tempo adquiriu, na perceção pública, uma conotação predominantemente negativa, associada a atividades ilícitas como a violação da segurança digital, o roubo de dados e a destruição de sistemas (Nunes & Sani, 2021).

No contexto jurídico português, esta prática encontra-se prevista no artigo 6.º da Lei n.º 109/2009, de 15 de setembro (Lei do Cibercrime), que criminaliza as ações que visem a perturbação de sistemas informáticos, incluindo o acesso não autorizado com a intenção de danificar, interferir, suprimir ou impedir o funcionamento normal de um sistema ou rede de comunicações.

Historicamente, o *hacking* esteve inicialmente associado a grupos de entusiastas da informática, na sua maioria jovens do sexo masculino, cujo principal objetivo era testar os seus conhecimentos técnicos e desafiar os limites das tecnologias emergentes. Contudo, a evolução do fenómeno trouxe consigo uma diversificação de perfis e

motivações, desde intenções maliciosas até ações com justificação política, económica ou ideológica.

Tipologia de Hackers segundo Moore (2014)

Uma das classificações mais citadas na literatura é a proposta por Moore (2014), que distingue seis tipos de hackers, segundo as suas motivações e métodos operacionais:

- ***Black hat hackers*** (ou ***crackers***): são indivíduos que atuam com intenções maliciosas, desenvolvendo ou utilizando software para aceder indevidamente a sistemas, danificá-los ou obter vantagens ilegítimas, como o roubo de dados ou o prejuízo económico de terceiros.
- ***White hat hackers***: têm uma abordagem ética, utilizando os seus conhecimentos para detetar vulnerabilidades em sistemas informáticos, com o objetivo de os proteger contra-ataques. Muitas vezes colaboram com empresas e organizações na prevenção de ciberataques.
- ***Gray hat hackers***: combinam características dos dois tipos anteriores. Ainda que identifiquem falhas nos sistemas com aparente boa intenção, frequentemente exigem compensações monetárias para corrigir os problemas, operando assim numa zona ambígua de legalidade.
- ***Script kiddies***: são utilizadores com conhecimentos técnicos limitados, que recorrem a ferramentas criadas por outros *hackers* mais experientes para executar ataques informáticos. Apesar de nem sempre compreenderem as consequências das suas ações, podem causar danos significativos.
- ***Hacktivists***: agem com motivações ideológicas, políticas ou sociais, utilizando o *hacking* como forma de protesto ou ativismo digital. Um dos exemplos mais emblemáticos é o grupo *Anonymous*, conhecido por ataques contra governos e corporações.
- **Ciberterroristas**: utilizam técnicas de *hacking* para instaurar medo e insegurança na população ou no Estado, podendo atacar infraestruturas críticas e causar danos severos, incluindo riscos à integridade física das pessoas.

Furto de identidade *online*

O furto de identidade online representa uma das ofensas mais graves e perturbadoras no contexto do cibercrime, com impactos diretos e prolongados sobre a integridade pessoal

e financeira das vítimas. Esta forma de criminalidade consiste na obtenção e utilização indevida de dados pessoais, como números de identificação, informações bancárias ou credenciais de acesso, com o objetivo de obter benefícios ilícitos, disfarçar ações criminosas ou prejudicar a vítima (Solove, 2003).

Segundo Anderson, Durbin e Salinger (2008), o furto de identidade online é uma das ofensas cibernéticas mais danosas, afetando a perceção de segurança digital dos cidadãos. Um estudo empírico realizado em Portugal por Martins (2018), com uma amostra de 832 indivíduos, revela dados particularmente relevantes:

- 5,9% dos inquiridos foram vítimas de furto de identidade nos 12 meses anteriores ao estudo;
- 19,6% referiram já ter sido vítimas ao longo da vida;
- 37,7% mencionaram conhecer um familiar, amigo ou conhecido que sofreu esse tipo de crime.

Estes dados demonstram não só a prevalência do fenómeno, mas também o seu alcance social, sendo muitas vezes invisível até à concretização do dano.

Métodos mais utilizados

Diversas estratégias são utilizadas para concretizar o furto de identidade online. Destacam-se entre os métodos mais comuns:

- *Phishing* – criação de páginas falsas que imitam entidades legítimas (como bancos), combinadas com o envio de e-mails fraudulentos, para induzir o utilizador a fornecer dados pessoais e financeiros. Esta prática, de acordo com Jahankhani, Al-Nemrat e Hosseinian-Far (2014), visa enganar a vítima para que revele, voluntariamente, informações sensíveis.
- *Pharming* – redirecionamento automático do utilizador para websites falsificados, mesmo quando este insere o endereço correto.
- *Malware* – software malicioso que se infiltra no sistema da vítima para roubar dados de forma silenciosa.

Estas técnicas são frequentemente combinadas e evoluem à medida que os mecanismos de segurança digital se tornam mais sofisticados.

Enquadramento jurídico em Portugal

A legislação portuguesa sobre o furto de identidade passou por alterações significativas. Com a revogação da Lei n.º 12/91, de 21 de maio (Lei da Identificação Civil e Criminal), deixou de existir um regime autónomo dedicado ao furto de identidade. Atualmente, a utilização da identidade de outrem apenas assume relevância penal quando está associada à obtenção de um benefício ilegítimo ou à intenção de prejudicar a vítima, podendo enquadrar-se nos crimes de burla (art. 217.º CP), falsidade informática (art. 3.º da Lei do Cibercrime) ou acesso ilegítimo (art. 6.º da mesma Lei).

A ausência de um tipo legal específico para o furto de identidade constitui um desafio jurídico, tanto no plano da prevenção como da repressão penal, abrindo espaço para o debate sobre a necessidade de revisão legislativa face à evolução tecnológica e às novas formas de vitimação digital.

Cyberstalking

O *cyberstalking* refere-se a comportamentos de perseguição, repetidos e indesejados, que ocorrem no contexto do ciberespaço e que são percecionados pela vítima como intrusivos, ameaçadores, assustadores e/ou assediantes (Dreßing, Bailer, Anders, Wagner, & Gallas, 2014). Estas práticas podem assumir várias formas: envio de mensagens ameaçadoras por e-mail, envio de ficheiros com vírus, utilização fraudulenta do endereço de e-mail da vítima para realizar compras ou subscrições, disseminação de informações falsas ou comprometedoras em redes sociais ou no local de trabalho, usurpação de identidade para partilha de conteúdos ofensivos, ou a compilação e exploração de dados pessoais extraídos da presença online da vítima, com o intuito de a intimidar, tanto online como no mundo físico (Finn & Banach, 2000).

Este tipo de perseguição pode ser exercido por diferentes tipos de ofensores – ex-companheiros, conhecidos ou até desconhecidos – sendo que, segundo Dreßing et al. (2014), aproximadamente 35% dos casos envolvem ex-parceiros. No entanto, um estudo distinto realizado por Short, Guppy, Hart e Barnes (2015) indica que a maioria das vítimas relatou desconhecer a identidade do ofensor, representando cerca de 38% da amostra.

Um aspeto particularmente preocupante do *cyberstalking* é a sua capacidade de transpor o meio digital e afetar o quotidiano da vítima. A presença constante do agressor nos dispositivos digitais, e-mails e redes sociais, aliada à potencialidade de contacto físico,

torna o fenómeno especialmente intimidante. Esta ameaça difusa perturba a vida social, profissional e emocional das vítimas, sobretudo quando envolve a exposição pública de conteúdos íntimos e potencialmente humilhantes, cuja propagação é instantânea nas plataformas digitais (Short et al., 2015; Worsley, Wheatcroft, Short, & Corcoran, 2017).

Enquadramento Legal em Portugal

No ordenamento jurídico português, o *cyberstalking* não está tipificado de forma autónoma, mas encontra enquadramento em diversas disposições do Código Penal. O artigo 190.º trata da *violação de domicílio ou perturbação da vida privada*, enquanto o artigo 192.º versa sobre a *devassa da vida privada*. Estas normas podem ser mobilizadas para penalizar condutas que, no espaço digital, configurem perseguição persistente, intrusão na privacidade e atentado à vida pessoal de outrem.

Assim, embora não exista uma categoria penal expressamente intitulada "*cyberstalking*" na legislação portuguesa, as práticas que o caracterizam encontram acolhimento legal, permitindo uma resposta penal às condutas persecutórias no ambiente digital.

Pornografia Infantil

A pornografia infantil constitui uma das formas mais graves de cibercrime, pela sua natureza profundamente violadora dos direitos das crianças e adolescentes. Em Portugal, esta conduta está tipificada no Código Penal, nos artigos 176.º e 176.º-A, que punem, respetivamente, a *pornografia de menores* e o *aliciamento de menores para fins sexuais* através de tecnologias da informação e comunicação (TIC). A legislação reconhece, assim, a especificidade dos riscos digitais na vitimação sexual infantil, criminalizando expressamente os comportamentos de predadores sexuais no meio online.

O ciberespaço, pela sua natureza anónima e desmaterializada, oferece um manto de invisibilidade que favorece o cometimento de crimes sexuais contra menores. Este fenómeno é potenciado por vários fatores: a ingenuidade e inexperiência das crianças e jovens nas interações digitais, a falta de supervisão parental eficaz (por vezes relacionada com o desconhecimento tecnológico dos cuidadores), e a facilidade de criação de identidades falsas por parte dos ofensores. Assim, a *Internet* tornou-se um novo contexto facilitador da exploração sexual infantil, onde ofensores podem aceder a conteúdos ilícitos, partilhar pornografia, identificar vítimas em potencial e interagir com elas de forma manipuladora e estratégica.

Entre os métodos mais reportados, destaca-se o aliciamento de menores online, frequentemente designado por *grooming*. Este processo envolve uma aproximação gradual à vítima, com o objetivo de conquistar a sua confiança e reduzir as suas defesas. O ofensor pode recorrer a atenção, amizade, manipulação emocional e até oferta de presentes, criando uma falsa relação de afeto e segurança. De acordo com Berson (2003), este processo pode evoluir para a exposição da criança à pornografia, seguida de sugestões para posar em fotografias de cariz sexual, numa tentativa de dessensibilizar a vítima, estimular a sua curiosidade sexual e validar comportamentos abusivos.

Além disso, o intercâmbio de conteúdos ilegais em comunidades *online* de ofensores promove a validação mútua e a partilha de estratégias de manipulação, criando redes transnacionais de abuso. A generalização da exposição de dados pessoais de menores - muitas vezes partilhados de forma inconsciente em redes sociais - agrava os riscos, ao facilitar a identificação e contacto por parte de agressores, tanto *online* como *offline*.

Inteligência Artificial (IA)

O avanço da Inteligência Artificial, nomeadamente através dos sistemas generativos, veio transformar profundamente o panorama das interações digitais. A capacidade destas tecnologias em simular comunicação humana de forma credível e emular sinais de confiança tem gerado preocupações significativas no domínio da cibersegurança. De facto, a deceção digital, alimentada por IA generativa, representa uma ameaça crescente numa sociedade interligada, ao potenciar ataques de engenharia social e *phishing* altamente sofisticados.

À medida que os sistemas de IA se tornam mais aptos a replicar padrões linguísticos humanos, criar rostos sintéticos (*deepfakes*) e interagir de forma persuasiva, as barreiras entre o que é real e o que é artificial tornam-se progressivamente mais difíceis de detetar. Esta ambiguidade facilita a manipulação psicológica de utilizadores, expondo-os a riscos de vitimação em diferentes tipos de contacto, seja computador-para-computador, humano-para-computador, ou mesmo humano-para-humano, com a intermediação de agentes sintéticos.

Assim, coloca-se uma exigência urgente à investigação científica: compreender os desafios éticos, técnicos e legais colocados por estas tecnologias, e desenvolver mecanismos de proteção que garantam interações digitais seguras, informadas e autênticas. Neste sentido, torna-se vital não só o reforço das capacidades institucionais de

deteção e resposta, como também a literacia digital dos cidadãos (Schmitt & Flechais, 2024).

IA na engenharia social: Deepfakes e phishing

A dissimulação desempenha um papel central tanto nas táticas de engenharia social como nos ataques de *phishing*. Em ambos os casos, o atacante procura manipular ou enganar o alvo com o intuito de o levar a executar determinadas ações ou a revelar informações confidenciais, fazendo-se passar por uma entidade legítima e de confiança. A engenharia social refere-se, assim, à manipulação de indivíduos para que estes executem comportamentos específicos ou divulguem dados sensíveis, utilizando sobretudo estratégias enganosas. Trata-se de uma técnica que explora as vulnerabilidades humanas e psicológicas, em vez de falhas puramente tecnológicas (Schmitt & Flechais, 2024).

Tecnologia *Deepfake*

O termo "*deepfake*" resulta da junção das palavras *"deep learning"* e *"fake"*, tendo sido originalmente cunhado por um utilizador da plataforma Reddit, que recorreu a algoritmos de aprendizagem automática para editar vídeos explícitos com imagens de celebridades. Os *deepfakes* assentam numa estrutura de *machine learning* designada por Redes Geradoras Adversariais (GANs - *Generative Adversarial Networks*), que produzem conteúdos falsos ao sobrepor traços faciais de uma pessoa sobre outra, utilizando imagens e vídeos reais como referência. No contexto da criação de deepfakes, este modelo permite identificar e corrigir continuamente as imperfeições do produto gerado, aperfeiçoando progressivamente o seu realismo. Este processo pode ser aplicado tanto a vídeo como a áudio (Rancourt-Raymond, & Smaili, 2023).

Importa salientar que, embora o termo tenha surgido associado a usos maliciosos de algoritmos de *deep learning*, os *deepfakes* podem igualmente ser utilizados para fins legítimos e benéficos, como na indústria cinematográfica, educação, ou acessibilidade digital (Rancourt-Raymond, A. de, & Smaili, N. 2023). Por exemplo, Zhu et al. (2020) destacam que os *deepfakes* podem ser utilizados para ofuscar as feições dos pacientes em imagens médicas, contribuindo assim para a proteção da sua privacidade. Além disso, Simon Chandler na revista Forbes (2020) sugere que os *deepfakes* podem ser empregues para recriar a imagem de personalidades históricas relevantes.

O *Centre for Data Ethics and Innovation* identifica quatro formas principais de *deepfake*, nomeadamente: substituição de rosto (*face replacement*), reencenação facial (*face re-enactment*), geração facial (*face generation*) e síntese de voz (*speech synthesis*).

1. **Substituição de rosto (*Face Replacement*):** A substituição de rosto pode ser considerada a forma mais conhecida de *deepfake* e geralmente envolve a utilização de várias amostras faciais de um indivíduo, juntamente com um vídeo existente no qual se pretende inserir essa pessoa. Estas manipulações são frequentemente realizadas através de aplicações como o '*DeepFaceLab*'. O algoritmo identifica os contornos de todas as faces presentes no vídeo, bem como características-chave, como os olhos, a ponte do nariz e a estrutura facial. Após o mapeamento dessas áreas, o modelo sobrepõe a face do indivíduo amostrado ao rosto original no vídeo.
2. **Reencenação facial (*Face Re-enactment*)**: A reencenação facial utiliza a mesma metodologia da substituição de rosto, mas, em vez de substituir o indivíduo no vídeo, visa alterar as suas expressões faciais, sobretudo a movimentação da boca, para que pareça estar a dizer algo diferente do conteúdo original. Esta técnica torna-se ainda mais eficaz quando combinada com áudio genuíno ou *deepfake* do próprio indivíduo.
3. **Geração Facial (*Face Generation*)**: A geração facial utiliza tecnologias como a arquitetura 'StyleGAN', que permite isolar e modificar características específicas da face (Karras, T., Laine, S., & Aila, T., 2019). Esta técnica é aplicada em conjunto com imagens de amostra para criar faces totalmente novas, combinando os detalhes e estruturas faciais presentes nas imagens fornecidas com características personalizadas previamente definidas. O desenvolvimento desta tecnologia contou com a colaboração da NVidia, uma empresa de tecnologia e *hardware*, com o objetivo de gerar personagens mais realistas para filmes e videojogos.
4. **Síntese de voz (*Speech Synthesis*):** A síntese de voz, última técnica de *deepfake*, utiliza amostras da voz de um indivíduo que são convertidas numa representação mel-espectrográfica (Jia, Y., et al 2018). Partes desta representação podem ser manipuladas e, em combinação com um *input* textual, permitem gerar um discurso que simula a voz do indivíduo amostrado.

Phishing

O *phishing* é uma forma de ataque baseada em engenharia social, na qual um agente malicioso procura obter acesso a informações ou credenciais de outra pessoa, induzindo a vítima a acreditar que está a interagir com uma entidade ou organização legítima (Rader, M., Rahman, S., 2015).

As técnicas de phishing evoluíram para formas mais direcionadas e sofisticadas, como o spearphishing, que consiste em atacar um indivíduo específico através de um ataque altamente especializado, e o whaling, onde o atacante visa membros de alto escalão numa organização, como executivos seniores, imitando seus superiores para solicitar transferências de fundos ou informações confidenciais. De acordo com um relatório do *National Cyber Security Centre* (NCSC), é comum que domínios de email e *templates* sejam recriados com pequenas alterações para manter a credibilidade e realismo do ataque.

Segundo M. Schmitt & I. Flechais (2024), exitem os seguintes tipos diferentes de *phishing*:

Tipo de *Phishing*	Explicação
Email phishing	O atacante envia emails fraudulentos que parecem ser de fontes legítimas (ex.: bancos, organismos governamentais) para enganar os destinatários e obter informações sensíveis.
Spear phishing	Semelhante ao *email phishing*, mas o atacante personaliza o email enganador para um indivíduo ou organização específica.
Smishing	O atacante envia mensagens de texto (SMS) enganosas para induzir os destinatários a revelar informações sensíveis ou clicar em *links* maliciosos.
Whaling	Forma especializada de spear phishing onde o alvo são indivíduos de alto perfil, como CEOs ou CFOs. O objetivo é manipular estas pessoas para autorizar grandes transferências de dinheiro ou revelar dados corporativos sensíveis.

Pharming	O atacante redireciona o tráfego *web* da vítima para um site falso que é idêntico a um *site* legítimo, enganando-a para inserir as suas credenciais ou outras informações sensíveis.
Vishing	O atacante usa chamadas telefónicas fraudulentas para enganar a vítima e obter informações confidenciais.
Phishing em redes sociais	O atacante utiliza plataformas de redes sociais (ex.: *Facebook* ou *Instagram*) para enganar os seus alvos (Seymour e Tully, 2018).

Tabela 1. Principais Tipos de Ataques de *Phishing.*

Os primeiros esquemas de *phishing* estavam relacionados com o roubo das credenciais de uma pessoa para aceder a um Provedor de Serviços de Internet (ISP). Nos tempos anteriores à *internet* banda larga e ao acesso ilimitado e gratuito, as pessoas ligavam-se à internet através de *modems dial-up*. Depois de estabelecida a ligação ao ISP, o utilizador introduzia um nome de utilizador e uma palavra-passe. Na altura, os ISPs cobravam aos utilizadores o acesso à *internet* por minuto de utilização (Rader, M., & Rahman, S. 2015).

As origens do *phishing* vêm do desejo por acesso ilimitado à internet. Ao comprometer uma conta da AOL, uma pessoa podia ter acesso ilimitado à *internet* à custa do titular da conta comprometida. A vítima recebia uma fatura da AOL referente ao consumo excessivo causado pelo ladrão e tinha de contactar a AOL para contestar os encargos. A AOL cobrava diretamente o cartão de crédito do cliente, pelo que o consumo excessivo poderia não ser detetado pela vítima durante alguns meses (Rader, M., & Rahman, S. 2015).

Principais Preocupações com Instrumentalização da IA

Segundo M. Schmitt & I. Flechais (2024), utilizam duas métricas para avaliar o impacto do IA especialmente em ataques de *phishing* e com usos de *Deepfakes*:

Amplificação da Ameaça:

Diz respeito ao aumento do impacto e da eficácia dos ataques de *phishing* como resultado da utilização de inteligência artificial generativa. Esta amplificação pode ser avaliada através da taxa de sucesso dos ataques, da sua capacidade de atingir um maior número de vítimas ou da rapidez com que são executados.

Relação Custo-Efetividade:

Refere-se à diminuição dos custos associados à realização de ataques de *phishing*, mantendo ou até aumentando a sua eficácia. Esta métrica pode ser analisada com base nos recursos necessários para a sua concretização, como o tempo, o dinheiro e o esforço envolvidos.

Os avanços na inteligência artificial (IA) têm um impacto significativo no panorama dos ataques de *phishing*, tendendo a favorecer os ofensores em detrimento dos defensores. Esta natureza dual da IA levanta implicações importantes para o campo da cibersegurança. Como salientam Schmitt e Flechais (2024), os sistemas de IA generativa estão a tornar-se cada vez mais eficazes na imitação de padrões de comunicação humana e sinais de confiança, o que representa uma ameaça direta à integridade das interações digitais.

As abordagens tradicionais, que se baseiam na formação dos utilizadores finais para reconhecer técnicas de engano, revelam-se insuficientes, sobretudo no contexto de ataques direcionados como o *spear phishing* ou o *whaling*. Os utilizadores, muitas vezes, têm dificuldade em identificar e reagir adequadamente a estas estratégias sofisticadas, o que deixa indivíduos e organizações expostos a campanhas de *phishing* altamente personalizadas e credíveis (Schmitt & Flechais, 2024).

A utilização de sistemas de IA avançados permite ainda otimizar a criação de ataques personalizados em larga escala. Os atacantes podem recorrer a algoritmos de IA para analisar grandes volumes de dados e gerar campanhas de *phishing* altamente direcionadas, de forma automatizada e com elevada eficácia. Esta capacidade de personalização e escala aumenta substancialmente o sucesso dos ataques, agravando o risco para os alvos (Heiding et al., 2024).

Este cenário é agravado pela crescente sofisticação do ecossistema clandestino do cibercrime, pelo surgimento da economia de trabalhos temporários (*gig economy*), e pela ampla disseminação de soluções de *Software as a Service* (SaaS). O ecossistema subterrâneo facilita a partilha de ferramentas, técnicas e recursos entre cibercriminosos, promovendo o uso de IA para fins maliciosos. Paralelamente, a *gig economy* permite o acesso fácil a plataformas que disponibilizam serviços e ferramentas baseados em IA, possibilitando que até utilizadores com poucos conhecimentos técnicos explorem esta

tecnologia para fins ilícitos (Lemos, R., 2023.; Patsakis, Arroyo & Casino 2024; Claire, Cuppoy & Okunola, 2025).

Adicionalmente, a popularização de soluções SaaS fornece a infraestrutura necessária para lançar campanhas de *phishing* complexas com base em IA. A progressiva redução dos custos de desenvolvimento e implementação de sistemas de IA representa, neste contexto, uma faca de dois gumes: se por um lado democratiza o acesso à tecnologia para usos legítimos, por outro, facilita a apropriação desses recursos por agentes maliciosos. Como consequência, mesmo indivíduos ou grupos com recursos limitados conseguem aceder a ferramentas de IA sofisticadas, aumentando o número e a sofisticação dos ataques. Este fenómeno exige respostas inovadoras e urgentes por parte das organizações e autoridades, no sentido de reforçar os seus mecanismos de defesa face a este novo paradigma do cibercrime (Schmitt & Flechais, 2024).

Contra Medidas

Podemos distinguir entre medidas técnicas de mitigação, geralmente baseadas em inteligência artificial (IA), aprendizagem automática (ML) e criptografia, e medidas centradas no utilizador, que se concentram sobretudo na formação e sensibilização dos utilizadores. (Naqvi et al. 2023).

Apesar de já terem sido avançadas diversas propostas de soluções inovadoras (Glas et al. 2023), a indústria da cibersegurança tem centrado os seus esforços, sobretudo, em programas de sensibilização destinados a combater o chamado "problema humano". Estes programas são amplamente utilizados, independentemente da dimensão das organizações, sendo adotados tanto por pequenas empresas como por grandes corporações. No entanto, revelam-se geralmente ineficazes perante estratégias de ataque mais sofisticadas, que envolvem técnicas avançadas de personalização e segmentação. (Schmitt & Flechais, 2024)

Schmitt e Flechais (2024) referem que considerar o utilizador final como o principal responsável pelas falhas de segurança representa não apenas uma simplificação excessiva do problema, mas também uma perspetiva perigosa. Esta visão sugere que os utilizadores devem ser culpabilizados por não serem suficientemente atentos ou informados, sendo a solução proposta a sua formação através de programas de sensibilização e educação em segurança (SAET). No entanto, esta abordagem revela-se inadequada e apresenta diversas

limitações, tanto em termos de eficácia como de justiça na atribuição de responsabilidades.

Algumas razões apresentadas por Schimitt e Flechais (2024) pelo qual esta perspetiva é errada e ineficiente:

1. Perante os avanços da Inteligência Artificial na imitação das interações humanas, torna-se inviável que esta crescente capacidade seja eficazmente combatida apenas através de maior sensibilização ou compreensão por parte dos utilizadores. Estamos rapidamente a aproximar-nos de um cenário em que distinguir entre conteúdos genuínos e fabricados será praticamente impossível para o ser humano.
2. Devido à possível escala e automatização dos ataques maliciosos, as pessoas terão de manter um nível contínuo e quase perfeito de atenção para conseguir equilibrar eficazmente a proteção contra fraudes e a confiança nos meios de comunicação.
3. Já é difícil avaliar a eficácia do Treino de Engenharia Social e Testes de *Phishing* (SAET): quando alguém falha num teste de *phishing* ou se torna vítima de engenharia social, tende-se a interpretar isso como falta de treino adequado ou de atenção, o que serve de justificação para implementar mais campanhas de sensibilização, formação e educação.

As implicações da IA no contexto do *phishing* evidenciam a necessidade de defesas inovadoras e estratégias proativas para mitigar os riscos associados a ataques potenciados por IA.

Um exemplo de identificação de enganos, no âmbito do primeiro pilar, é a deteção de *deepfakes*. De acordo com Kaur et al. (Gambín et al., 2024), a maioria dos métodos utilizados para detetar vídeos *deepfake* baseia-se em dados. As técnicas de aprendizagem profunda, especialmente as Redes Neuronais Convolucionais (CNNs) e as Redes Neuronais Recorrentes (RNNs), são predominantes nesta área, pois conseguem identificar inconsistências subtis em vídeos manipulados. No entanto, estas abordagens enfrentam diversos desafios, incluindo a disponibilidade de dados, a complexidade computacional, dificuldades de generalização e questões de fiabilidade. À medida que os métodos de criação de *deepfakes* se tornam mais sofisticados, aumenta também a sua capacidade de escapar à deteção, exigindo assim melhorias contínuas nas técnicas de identificação.

É importante salientar que, embora a inteligência artificial generativa possa ser utilizada de forma maliciosa nestes cenários, também pode ter aplicações benéficas, como na criação de materiais de formação para profissionais de cibersegurança ou no desenvolvimento de ferramentas capazes de detetar e neutralizar conteúdos enganosos gerados por IA. No geral, a evolução constante da IA torna essencial a atualização contínua sobre os avanços mais recentes relacionados com a fraude impulsionada por IA e as respetivas medidas de mitigação. Segundo Schmitt e Flechais (2024), ao relacionar as capacidades da IA com as diversas fases dos ataques de engenharia social, é possível utilizar este enquadramento para identificar novas oportunidades de investigação e desenvolvimento com vista à interrupção de ataques de engenharia social potenciados por IA.

Mesmo especialistas experientes em cibersegurança podem ser vulneráveis a ataques de *phishing* sofisticados gerados por inteligência artificial. A expectativa de que os indivíduos consigam analisar minuciosamente cada *e-mail* que recebem é irrealista e contraproducente, especialmente no ambiente digital atual, caracterizado por um ritmo acelerado e uma elevada carga cognitiva (Kosch et al. ,2023). Além disso, esta forma de pensar ignora fatores psicológicos essenciais, como os explicados por investigadores como Daniel Kahneman (2013). Tarefas rotineiras, como verificar *e-mails*, raramente envolvem pensamento crítico e, quando são desencadeadas emoções, especialmente o medo, a tomada de decisões racionais torna-se difícil. As soluções atuais, como os sistemas de deteção de anomalias e de *spam* baseados em IA, são relativamente eficazes, mas revelam-se insuficientes perante ataques direcionados de *spear phishing* e ofensivas patrocinadas por Estados (Schmitt & Flechais, 2024).

Direções para Investigações Futuras

As três capacidades identificadas da inteligência artificial generativa no contexto da engenharia social (criação de conteúdos realistas, personalização e automatização) formam uma combinação problemática que assinala o início de inovações potencialmente imprevisíveis no campo do *hacking* (Schmitt & Flechais, 2024).

Possíveis avenidas de pesquisa futura sobre a temática:

1. **Sensibilização e Educação dos Utilizadores:** É fundamental reconhecer que a formação em "ciberconsciência" não constitui uma solução simples contra-ataques cibernéticos potenciados por inteligência artificial. No entanto, o aumento

do conhecimento dos utilizadores sobre as técnicas e estratégias mais recentes pode capacitá-los a reconhecer e reagir de forma adequada a potenciais ameaças (Distler, 2023; Marin et al., 2023). Para tal, é essencial conceber programas de formação eficazes, simulações interativas e materiais educativos acessíveis e orientados para o utilizador (Jansen e Fischbach, 2020; Glas et al., 2023).

2. **Aprendizagem Automática Adversarial:** É crucial desenvolver técnicas avançadas para detetar e defender contra-ataques adversariais no contexto da engenharia social potenciada por IA (Ahmad et al., 2023). A aprendizagem automática adversarial tem como objetivo criar modelos robustos, capazes de resistir a tentativas de manipulação por parte de atacantes. A investigação nesta área pode centrar-se no desenvolvimento de algoritmos e estratégias eficazes para identificar e mitigar técnicas adversariais de engenharia social e *phishing*.
3. **Defesa Ativa por Engano:** É fundamental desenvolver mecanismos de defesa proativos que consigam interromper ataques enganosos em tempo real. Isto pode incluir o uso de tecnologias como o processamento de linguagem natural, a deteção de anomalias e a análise em tempo real dos canais de comunicação, com o objetivo de identificar e bloquear tentativas de engenharia social e *phishing* no momento em que ocorrem. Paralelamente, é igualmente essencial promover a investigação no domínio da deteção de *deepfakes* (Kaur et al., 2024).
4. **IA Explicável para Deteção de Ameaças:** É importante melhorar a transparência dos modelos de inteligência artificial utilizados na deteção de ameaças. As técnicas de IA explicável permitem que analistas de segurança e utilizadores compreendam como os sistemas de IA tomam decisões e identificam indicadores de ataques enganosos. Ao fornecer explicações e *insights* compreensíveis, torna-se mais fácil confiar e validar os resultados produzidos por sistemas de deteção de ameaças baseados em IA (Kim et al., 2023).
5. **Ameaças à Cibersegurança:** A inteligência artificial generativa (IA Gen) não só tem o potencial de amplificar ou originar novos ataques de engenharia social e *phishing*, como também de potenciar outras ameaças no domínio da cibersegurança. Investigações futuras poderão explorar o seu impacto em ataques de *ransomware* (Teichmann, 2023), ameaças internas (insider threats), ataques do tipo *man-in-the-middle* e na utilização de deepfakes (Gambín et al., 2024; Kaur et al., 2024).

Capítulo 2 – Proposta de Investigação

O presente capítulo centra-se na apresentação da proposta de investigação. Assim sendo, serão aqui expostos os objetivos gerais e específicos, a caracterização da metodologia adotada, a amostra, os instrumentos e os procedimentos requeridos para a concretização do estudo e, por último, a definição dos resultados esperados.

Objetivos da Investigação

A presente investigação tem como finalidade principal avaliar a eficácia de ferramentas de cibersegurança baseadas em inteligência artificial (IA) na proteção de sistemas operativos contra ciberameaças. A crescente complexidade e sofisticação dos ciberataques, aliada à rápida evolução da inteligência artificial, tornam imperativa a análise crítica da real capacidade destas tecnologias em contextos operacionais.

Objetivos Gerais

Avaliar o desempenho de sistemas de cibersegurança baseados em inteligência artificial na deteção, resposta e mitigação de ciberataques em diferentes sistemas operativos.

Objetivos Específicos

- Analisar comparativamente as funcionalidades e o desempenho de ferramentas de cibersegurança com IA (ex: FortiEDR, CrowdStrike, SentinelOne, Microsoft Defender for Endpoint) em ambientes operacionais distintos (Windows, Linux, macOS).

  Estudos recentes demonstram que a eficácia destas ferramentas varia significativamente em função do sistema operativo e do tipo de ameaça simulada (Chew, C. J. W., Kumar, V., Patros, P., & Malik, R. 2024).

- Testar a capacidade de deteção de ataques sofisticados como phishing, ransomware e engenharia social por parte de sistemas de cibersegurança com IA.

  A simulação de ataques personalizados permite aferir o grau de resposta inteligente das ferramentas, especialmente em contexto de ameaças de dia zero ou ataques adaptativos (Okdem & Okdem, 2024).

- Identificar quais os algoritmos e abordagens de IA mais eficazes na proteção de sistemas operativos, com base na taxa de falsos positivos/negativos, tempo de resposta e adaptabilidade.

A precisão algorítmica é um dos fatores críticos de sucesso em ambientes automatizados, e estudos mostram que a utilização de deep learning e redes neurais convolucionais tem vindo a demonstrar melhores resultados na deteção de anomalias (Karras, Laine & Aila, 2019).

- Avaliar de que forma a integração entre os sistemas de IA e os mecanismos nativos de defesa dos sistemas operativos contribui para a eficácia da ciberdefesa.

  A integração nativa com o kernel do sistema operativo pode aumentar significativamente a eficácia na resposta em tempo real (Moore, 2014; Patel, J. 2024).

- Identificar limitações, fragilidades e riscos associados à utilização de IA em cibersegurança, nomeadamente dependência de dados de treino, vulnerabilidade a adversarial AI e dificuldade em lidar com ameaças desconhecidas.

  A literatura aponta que a utilização de IA em cibersegurança não está isenta de riscos e que os sistemas podem ser manipulados ou ultrapassados por atacantes com conhecimento técnico elevado (Schmitt & Flechais, 2024).

Amostra

A amostra referente aos especialistas mencionados será composta por um grupo selecionado de profissionais que atuam na área de cibersegurança, nomeadamente engenheiros de segurança, analistas de redes e *white hat hackers* que possuem experiência prática no uso de ferramentas de cibersegurança baseadas em inteligência artificial.

Número de participantes:

O número exato a selecionar deverá equilibrar a representatividade com a viabilidade da recolha e análise de dados. Considerando a natureza do estudo, uma amostra de cerca de **30 a 50 especialistas** pode ser adequada para garantir diversidade de experiências e permitir análises estatísticas significativas, especialmente no contexto do questionário aplicado.

Tipo de amostra:

Trata-se de uma amostra **por conveniência**, uma vez que os participantes serão recrutados com base na sua acessibilidade e disponibilidade para colaborar na investigação. Esta abordagem é comum em estudos aplicados em áreas técnicas especializadas, onde o acesso a profissionais qualificados pode ser limitado e é feita por contactos em plataformas especializadas (ex: HackerOne, Bugcrowd) e redes profissionais.

Justificação:

A escolha da amostra por conveniência justifica-se pela necessidade de incluir profissionais com conhecimentos específicos e experiência comprovada em ferramentas de cibersegurança com IA, o que torna inviável a seleção aleatória. Além disso, a utilização de plataformas reconhecidas para recrutamento ajuda a garantir que os participantes possuem a qualificação necessária para contribuir com dados relevantes e fidedignos para o estudo.

A investigação será composta por um conjunto de ferramentas de cibersegurança que utilizam inteligência artificial, selecionadas com base na sua relevância no mercado, funcionalidades disponibilizadas e acessibilidade para fins académicos (ex: FortiEDR, CrowdStrike, SentinelOne, Microsoft Defender for Endpoint). Serão ainda utilizados diferentes sistemas operativos (Windows, Linux, macOS) como plataformas de teste.

Para a realização de ataques controlados aos sistemas operativos, será necessária a colaboração de hackers éticos (*White Hat Hackers*), referidos por Moore (2014), que atuarão de forma responsável e autorizada, permitindo a avaliação da eficácia dos sistemas de cibersegurança baseados em inteligência artificial. Um *white hat hacker* pode também ser considerado um engenheiro de segurança de TI ou um analista de segurança de redes, dado que contribui para conceber e implementar soluções de segurança eficazes.

Para estabelecer contacto com *white hat hackers* que possam colaborar na investigação, serão utilizados vários meios e plataformas reconhecidos no campo da cibersegurança.

Em primeiro lugar, plataformas especializadas em programas de *bug bounty* e segurança colaborativa, como **HackerOne** (https://www.hackerone.com) e **Bugcrowd** (https://www.bugcrowd.com), serão recursos fundamentais. Estas plataformas reúnem profissionais certificados e experientes em segurança, permitindo a publicação de desafios e a contratação de especialistas para testes controlados.

Adicionalmente, será explorada a participação em comunidades e fóruns *online* dedicados à segurança cibernética, tais como o *subreddit* **r/netsec** no *Reddit* e a plataforma **Stack Exchange Security**. Estes espaços reúnem profissionais, pesquisadores e entusiastas da área, sendo locais ideais para estabelecer contactos informais e recrutar colaboradores.

Por fim, a presença em conferências e eventos especializados, como **DEF CON** e **Black Hat**, proporcionará oportunidades para *networking* direto com *white hat hackers* e especialistas em segurança informática, permitindo apresentar a investigação e convidar à participação ativa.

Instrumentos e Técnicas

Para verificar a eficácia dos sistemas de cibersegurança baseados em inteligência artificial na proteção de sistemas operativos, os *white hat hackers* recorrerão a um conjunto de instrumentos e técnicas especializadas, tais como:

Testes de Penetração (*Penetration Testing*): Consistem na simulação de ataques reais para identificar vulnerabilidades nos sistemas operativos e avaliar a capacidade de resposta das ferramentas de segurança com IA (Owen-Jackson, C. 2024).

Análise de Intrusão (*Intrusion Testing*): Técnicas que procuram contornar as defesas automatizadas, com o objetivo de testar a robustez dos sistemas na deteção e neutralização de comportamentos anómalos (Liu, Che & Lakkaraju 2017).

Ferramentas de Exploração Automatizada: Utilização de ferramentas como Metasploit, Burp Suite, Nmap ou Wireshark para lançar ataques controlados e monitorizar a resposta dos sistemas protegidos por soluções de IA.

Simulação de Ataques de Engenharia Social: (como *phishing* ou *spear phishing*): Utilização de técnicas controladas para verificar se os sistemas conseguem detetar e bloquear ataques personalizados, simulando ameaças comuns no ciberespaço (Microsoft Defender, 2024).

Avaliação de *Logs* e Respostas Automatizadas: Análise dos registos gerados pelas ferramentas de cibersegurança com IA para verificar a precisão na deteção de ameaças, os falsos positivos/negativos e a eficácia das respostas automáticas (ManageEngine, 2024).

Testes de *Deepfake* e Anomalias Comportamentais: Em cenários mais avançados, poderá ser testada a capacidade da IA para detetar conteúdos falsificados (como *deepfakes*) e comportamentos invulgares nos sistemas operativos.

Procedimento

Será fundamental solicitar um parecer à Comissão de Ética da Faculdade de Ciências Humanas e Sociais da Universidade Fernando Pessoa (UFP) antes do início da recolha de dados e da execução do estudo.

Só após a aprovação da UFP, a investigação poderá avançar para a fase empírica. Para facilitar a recolha de dados e o contacto com especialistas em segurança informática (*white hat hackers*), foi criado um questionário que será distribuído por via de plataformas *online* como Google Forms, LimeSurvey ou Qualtrics. Como alternativa, poderá ser realizada uma Entrevista Semiestruturada, permitindo um diálogo mais aprofundado com os participantes. A entrevista semiestruturada visa aprofundar aspetos que o questionário não permite explorar totalmente, como as perceções detalhadas dos especialistas sobre o desempenho das ferramentas de IA, questões éticas, limitações técnicas e exemplos práticos. Além disso, permite clarificar respostas ambíguas do questionário, enriquecendo e validando os dados recolhidos para uma análise mais completa.

O Termo de Consentimento Informado a ser aplicado será o modelo padrão recomendado pela Comissão de Ética da Universidade Fernando Pessoa, direcionado a uma população adulta, não clínica e em língua portuguesa.

Além disso, foi criado um Formulário de Relato de Progresso dos Testes de Cibersegurança IA, cujo objetivo é coletar dados padronizados sobre o desempenho dos sistemas de cibersegurança com inteligência artificial durante os testes controlados realizados pelos *white hat hackers*.

Durante a investigação, todas as sessões de teste, entrevistas e recolhas de dados serão rigorosamente registadas e analisadas para assegurar a integridade e transparência do estudo. Para proteger a identidade dos colaboradores, serão utilizados pseudónimos ou códigos que substituem quaisquer identificadores pessoais, garantindo o anonimato dos participantes. No final da investigação, todos estes dados serão destruídos de modo a preservar o anonimato e a confidencialidade das respostas fornecidas pelos participantes da amostra.

Resultados Esperados

O presente projeto de graduação configura-se como uma proposta metodológica para uma futura investigação académica, pelo que nenhum dado foi efetivamente recolhido ou analisado até ao momento.

É importante destacar que, no decurso da presente análise exploratória, verificou-se a escassez de literatura centrada especificamente na eficácia real dos sistemas de cibersegurança baseados em inteligência artificial (IA) aplicados à proteção de sistemas operativos. Este fator torna a projeção de possíveis resultados uma etapa complexa. Ainda assim, a revisão preliminar da literatura e o delineamento do protocolo de investigação permitem antever algumas hipóteses e resultados prováveis.

No que diz respeito aos objetivos traçados, espera-se identificar um conjunto de padrões de desempenho dos sistemas analisados, com especial atenção às suas capacidades de deteção, resposta e mitigação face a ciberataques simulados. Pretende-se, igualmente, compreender quais os mecanismos baseados em IA mais eficazes na neutralização de ameaças em tempo real, bem como avaliar a sua integração com os principais sistemas operativos.

Adicionalmente, prevê-se a identificação de fatores que condicionam a performance dos sistemas testados, como a qualidade dos dados de treino, a complexidade dos algoritmos utilizados e o tipo de ataques simulados (*phishing*, *ransomware*, *exploit kits*, *etc*.). Será também importante verificar de que forma ferramentas de apoio, como análises de *logs* automatizadas ou sistemas de *sandbox* com IA, contribuem para a robustez da defesa cibernética.

A análise temática e experimental deverá permitir a formulação de conclusões preliminares sobre a eficácia diferenciada destes sistemas, antecipando possíveis lacunas, como a dificuldade em lidar com ataques altamente personalizados ou a dependência excessiva de modelos preditivos.

Importa, contudo, reconhecer limitações que poderão afetar a amplitude e a generalização dos resultados quando a investigação for realizada. Por exemplo, a variabilidade entre

sistemas operativos, o acesso limitado a ferramentas de ataque controlado e a necessidade de colaboração com especialistas (como *white hat hackers*) poderão impactar os resultados obtidos.

Em suma, embora este projeto se limite à apresentação de uma proposta de investigação, pretende-se que ele seja observado como um ponto de partida relevante para futuras pesquisas na área da cibersegurança com IA. Esta investigação poderá vir a colmatar lacunas teóricas importantes e sustentar futuras políticas de proteção digital, promovendo práticas mais seguras num contexto de constante evolução tecnológica.

Evidência de Investigação

1. **Ganho de produtividade e redução de alertas repetidos:** Bono et al. (2025) analisaram operações em vivo com ferramentas de IA generativa para cibersegurança e encontraram melhorias significativas em métricas como número de alertas por incidente, tempo médio de classificação de alertas (MTTD) e resolução (MTTR).
2. **Alta performance na deteção estática e dinâmica de *malware*:** O estudo "AI ATAC 1" (2023) comparou seis sistemas comerciais de deteção de *malware* e mostrou que métodos baseados em IA conseguem detetar programas maliciosos, incluindo *zero-day*, com alta precisão e em tempo reduzido sobre ambientes virtuais controlados.
3. **Redução de falsos positivos e melhoria na deteção em redes:** Uma revisão da Journal of Big Data (2024) demonstra que a IA melhora sistemas de deteção de intrusão (IDS), diminuindo alarme falsos e elevando a precisão, principalmente com redes neurais e meta-heurísticas. Outro estudo na Sensors (2023) reforça que sistemas IDS baseados em IA são mais eficazes do que os tradicionais.
4. **Taxas de deteção superiores para infraestruturas críticas:** Num estudo publicado na IEEE (2024) de Govea, Gaibor-Naranjo and Villegas, um sistema de IA atingiu 95 % de taxa de deteção (vs 85 % e 90 % em abordagens alternativas), teve tempo médio de resposta de 3 s (vs 6–9 s) e apresentou menor taxa de falsos positivos (4 % vs 9–12 %).
5. **Deteção aprimorada de *ransomware* e *zero-day*:** Revisões históricas destacam que IA, com algoritmos de *deep learning* e aprendizagem por reforço, consegue

identificar ataques *zero-day* e *ransomware* com precisão acima dos 90 % (Okdem Selcuk, e Sema Okdem, 2024).

Conclusão

No contexto da cibersegurança, a rápida evolução da inteligência artificial tem aberto novas possibilidades tanto para o aumento da eficácia das defesas quanto para o surgimento de ameaças cada vez mais sofisticadas. Este projeto de investigação procura preencher uma lacuna importante ao avaliar a eficiência dos sistemas de proteção baseados em IA na salvaguarda dos sistemas operativos contra-ataques tecnológicos variados.

Através de um desenho metodológico que envolve testes controlados e a colaboração com especialistas em segurança, como *white hat hackers*, a investigação visa proporcionar uma compreensão detalhada sobre o desempenho destes sistemas diante de ameaças reais e simuladas. O enquadramento teórico que fundamenta este estudo orienta a análise dos resultados, com foco na identificação dos pontos fortes e limitações das tecnologias atuais, bem como na proposição de melhorias.

Espera-se que os resultados desta pesquisa permitam não só medir a eficácia prática destes sistemas, mas também contribuir para o desenvolvimento de melhores práticas na utilização da IA em cibersegurança, oferecendo *insights* valiosos para organizações e profissionais da área. Reconhece-se, contudo, que fatores como a diversidade dos ambientes de teste e a complexidade dos ataques podem representar desafios e limitações a considerar.

Em suma, este estudo ambiciona ser uma base sólida para futuras investigações e intervenções que promovam a adoção segura e eficaz de tecnologias baseadas em inteligência artificial, reforçando a proteção dos sistemas operativos e, consequentemente, a segurança digital global.

Bibliografia:

Ahmad, R., Alsmadi, I., Alhamdani, W., & Tawalbeh, L. (2023). Zero-day attack detection: A systematic literature review. *Artificial Intelligence Review, 56*, 10733–10811. https://doi.org/10.1007/s10462-023-10437-z

Akamai Technologies. (s.d.). O que é um ataque de DDoS? Akamai. https://www.akamai.com/pt/glossary/what-is-ddos

Ali, A., Fahim, M., Janjua, G., Khan, S., & Raza, S. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *Journal of Big Data, 11*, Article 123. https://doi.org/10.1186/s40537-024-00957-y

ALLEA. (2023). *The European Code of Conduct for Research Integrity*. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/european-code-of-conduct-for-research-integrity_horizon_en.pdf

Anderson, K. B., Durbin, E., & Salinger, M. A. (2008). Identity theft. *Journal of Economic Perspectives, 22*(2), 171–192. https://doi.org/10.1257/jep.22.2.171

Berson, I. R. (2003). Grooming cybervictims: The psychosocial effects of online exploitation for youth. *Journal of School Violence, 2*(1), 5–18. https://digitalliteracyfeb17prod.yolasite.com/resources/Berson2003-Grooming_Cybervictims__The_Psychosocial_Effects_of_Online_Exploitation_for_Youth.pdf

Black Hat. (n.d.). Retrieved June 23, 2025, from https://www.blackhat.com/

Bono, J., Grana, J., Karakolios, K., Ramakrishna, P. H., & Srivastava, A. (2025). Generative AI in live operations: Evidence of productivity gains in cybersecurity and endpoint management. *arXiv*. https://doi.org/10.48550/arXiv.2504.08805

Bridges, R. A., Weber, B., Beaver, J. M., Smith, J. M., Verma, M. E., Norem, S., & Oesch, T. S. (2023). AI ATAC 1: An evaluation of prominent commercial malware detectors. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 1620–1629). IEEE. https://doi.org/10.48550/arXiv.2308.14835

Bugcrowd. (n.d.). Retrieved June 23, 2025, from https://www.bugcrowd.com/

Claire, M., Cuppoy, J., & Okunola, A. (2025). The impact of AI-powered fraud detection and prevention on cybersecurity in digital banking: A comparative analysis. *ResearchGate*. https://www.researchgate.net/publication/390448147_The_Impact_of_AI-Powered_Fraud_Detection_and_Prevention_on_Cybersecurity_in_Digital_Banking_A_Comparative_Analysis

Centre, N. C. S. (2020). *Business email compromise: Defending your organisation*. https://www.ncsc.gov.uk/guidance/business-email-compromise-defending-your-organisation

Chandler, S. (2020, March 9). Why deepfakes are a net positive for humanity. *Forbes*. https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity/#e512a3b2f84f

Chew, C. J. W., Kumar, V., Patros, P., & Malik, R. (2024). Real-time system call-based ransomware detection. *International Journal of Information Security, 23*, 1839–1858. https://doi.org/10.1007/s10207-024-00819-x

CrowdStrike. (n.d.). Retrieved June 22, 2025, from https://www.crowdstrike.com/en-us/

Daniel, K. (2013). *Thinking, fast and slow*. Farrar, Straus and Giroux. https://ia800603.us.archive.org/10/items/DanielKahnemanThinkingFastAndSlow/Daniel%20Kahneman-Thinking%2C%20Fast%20and%20Slow%20%20.pdf

DEF CON. (n.d.). Retrieved June 23, 2025, from https://defcon.org/

Distler, V. (2023). The influence of context on response to spear-phishing attacks: An in-situ deception study. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581170

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review, 70*(11), 35–36. https://community.mis.temple.edu/mis0855002fall2015/files/2015/10/S.M.A.R.T-Way-Management-Review.pdf

Dreßing, H., Bailer, J., Anders, A., Wagner, H., & Gallas, C. (2014). Cyberstalking in a large sample of social network users: Prevalence, characteristics, and impact upon

victims. *Cyberpsychology, Behavior, and Social Networking, 17*(2), 61–67. https://doi.org/10.1089/cyber.2012.0231

EUR-Lex. (2007). Towards a general policy on the fight against cybercrime. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013L0040&qid=1748792466720

Finn, J., & Banach, M. (2000). Victimization online: The downside of seeking human services for women on the internet. *CyberPsychology & Behavior, 3*(5), 785–796. http://dx.doi.org/10.1089/109493100316102

Fortinet. (n.d.). Artificial intelligence in cybersecurity. Retrieved June 14, 2025, from https://www.fortinet.com/resources/cyberglossary/artificial-intelligence-in-cybersecurity

Furnell, S. (2002). *Cybercrime: Vandalizing the information society*. Addison-Wesley. https://doi.org/10.1007/3-540-45068-8_2

Gambín, Á. F., Yazidi, A., & Vasilakos, A. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review, 57*, 64. https://doi.org/10.1007/s10462-023-10679-x

Glas, M., Vielberth, M., & Pernul, G. (2023). Train as you fight: Evaluating authentic cybersecurity training in cyber ranges. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581046

Govea, J., Gaibor-Naranjo, W., & Villegas-Ch, W. (2024). Transforming cybersecurity into critical energy infrastructure: A study on the effectiveness of artificial intelligence. *Systems, 12*(5), 165. https://doi.org/10.3390/systems12050165

HackerOne. (n.d.). Retrieved June 23, 2025, from https://www.hackerone.com/

Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024). *Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects https://doi.org/10.48550/arXiv.2412.00586*

Iperov. (2020). *iperov/deepfacelab* [Software]. GitHub. https://github.com/iperov/DeepFaceLab

Jahankani, H., Al-Nemrat, A., & Hosseinian-Far, A. (2014). Cybercrime classification and characteristics. In B. Akhgar, A. Staniforth, & F. Bosco (Eds.), *Cyber Crime and Cyber Terrorism Investigator's Handbook* (pp. 149–164). Elsevier. http://dx.doi.org/10.1016/B978-0-12-800743-3.00012-8

Jansen, P., & Fischbach, F. (2020). The social engineer: An immersive virtual reality educational game to raise social engineering awareness. In *CHI PLAY 2020 - Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play* (pp. 59–63). Association for Computing Machinery. https://doi.org/10.1145/3383668.3419917

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Moreno, I. L., Wu, Y., et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems* (pp. 4480–4490). https://doi.org/10.48550/arXiv.1806.04558

Jordan. (n.d.). Teste de penetração. Software Livre. Archived September 25, 2017. Retrieved June 22, 2025, from https://web.archive.org/web/20170925035127/http://softwarelivre.org/jordan/blog/teste-de-penetracao

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *arXiv*. https://doi.org/10.48550/arXiv.1812.04948

Kaur, A., Noori Hoshyar, H., Saikrishna, V., et al. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review, 57*, 159. https://doi.org/10.1007/s10462-023-10700-6

Keller, M. S., Levashkina, L., Cohn, G., et al. (2023). An adversarial approach for detecting synthetic speech. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581190

Kim, H. (2023). Synthesizing speech and facial expressions to mimic identity: Ethical and legal considerations. *AI Ethics Journal, 5*(2), 145–162.

Kosch T, Karolus J, Zagermann J et al (2023) A survey on measuring cognitive workload in human-computer interaction. ACM Comput Surv. https://doi.org/10.1145/3582272

Kshetri, N. (2021). The economics of cybercrime. *Journal of Economic Surveys, 35*(4), 1190–1210. https://doi.org/10.1111/joes.12427

Lantz, K. A. (2023). Deception in cyber defense: Human factors and countermeasures. *Cybersecurity and Human Behavior Journal, 4*(1), 22–38.

Lemos, R. (2023). *Cybercrime ecosystem spawns lucrative underground gig economy*. Dark Reading. https://www.darkreading.com/cyber-risk/cybercrime-ecosystem-spawns-lucrative-underground-gig-economy

Liu, H., Che, X., & Lakkaraju, S. (2017, agosto 23). *Understanding modern intrusion detection systems: A survey*. *ArXiv*. https://doi.org/10.48550/arXiv.1708.07174

Luo, Y., Xu, W., Liu, J., & Yu, Y. (2024). Advances in AI-driven cybersecurity: Applications and future directions. *Journal of Cybersecurity Research, 8*(1), 56–78.

Ma, T., Zhang, H., & Wang, L. (2023). Deep learning approaches to phishing detection: A comprehensive review. *Computers & Security, 120*, 102842.

Malwarebytes Labs. (n.d.). The state of ransomware 2024. Retrieved June 22, 2025, from https://www.malwarebytes.com/ransomware

Mari, A.-G., Zinca, D., & Dobrota, V. (2023). Development of a machine-learning intrusion detection system and testing of its performance using a generative adversarial network. *Sensors, 23*(3), 1315. https://doi.org/10.3390/s23031315

Marin IA, Burda P, Zannone N, Allodi L (2023) The influence of human factors on the intention to report phishing emails. In: Conference on human factors in computing systems - proceedings. Association for Computing Machinery http://dx.doi.org/10.1145/3544548.3580985

Martins, M. (2018). *Sentimento de insegurança e vitimação no ciberespaço: A relação entre variáveis individuais e contextuais* (Dissestação de mestrado). Faculdade de Direito da Universidade do Porto, Porto. https://repositorio-aberto.up.pt/bitstream/10216/119687/2/332349.pdf

Microsoft. (n.d.). Azure security center. Retrieved June 23, 2025, from https://azure.microsoft.com/en-us/services/security-center/

Mitnick, K. D., & Simon, W. L. (2002). *The art of deception: Controlling the human element of security*. Wiley.

Moore, R. (2014). Cybercrime: Investigating high-technology computer crime. Routledge. https://pt.scribd.com/document/576221693/Cyber-Security-Comprehensive-Study-And-R

Naqvi B, Perova K, Farooq A et al (2023) Mitigation strategies against the phishing attacks: a systematic literature review. Comput Secur 132 https://doi.org/10.1016/j.cose.2023.103387

NCSC. (2023). *UK National Cyber Security Centre – Annual Review 2023*. https://www.ncsc.gov.uk/files/NCSC%20Annual%20Review%202023.pdf

Nunes, L., & Sani, A. (2021). *Manual de criminologia e vitimologia*. Pactor.

Ökdem, S., & Okdem, S. (2024). *Artificial intelligence in cybersecurity: A review and a case study*. *Applied Sciences (Switzerland), 14*(22). https://doi.org/10.3390/app142210487

Oliveira, F. (2024). Cybersecurity trends in Portugal: A national perspective. *Portuguese Journal of Information Security, 12*(2), 101–120.

OpenAI. (2024). *GPT-4 technical report*. https://openai.com/research/gpt-4

Open Web Application Security Project (OWASP). (2023). *OWASP top ten project*. https://owasp.org/www-project-top-ten/

Owen-Jackson, C. (2024, 17 de dezembro). *Testing the limits of generative AI: How red teaming exposes vulnerabilities in AI models*. *IBM Think*. Recuperado em 24 de junho de 2025, de https://www.ibm.com/think/insights/testing-the-limits-of-generative-ai-red-teaming-exposes-vulnerabilities-in-ai-models

Patel, J. (2024, maio 11). *Cisco reimagines cybersecurity with AI and kernel-level visibility at RSAC 2024*. *VentureBeat*. https://thisweekinai.news/2024/05/11/cisco-reimagines-cybersecurity-with-ai-and-kernel-level-visibility-at-rsac-2024/?

Patsakis, C., Arroyo, R., & Casino, F. (2024). The Malware as a Service ecosystem: Democratizing access to sophisticated cyberattack capabilities. *Journal of Cybersecurity Research*, 15(1), 45–62. https://doi.org/10.48550/arXiv.2405.04109

Pew Research Center. (n.d.). Cybercrime and online harassment. Retrieved June 22, 2025, from https://www.pewresearch.org/internet/2021/01/13/cybercrime-and-online-harassment/

Polícia Judiciária. (s.d.). *Unidade Nacional de Combate ao Cibercrime e à Criminalidade Tecnológica (UNC3T)*. Recuperado em 24 de junho de 2025, de Polícia Judiciária website: https://www.policiajudiciaria.pt/unc3t/

Powers, P. (2023). AI and cybercrime: A dual-edged sword. *Cybersecurity Review, 7*(3), 30–44.

Rao, A., & Srinivas, V. (2024). AI-powered malware detection: Current trends and challenges. *Journal of Network Security, 15*(1), 12–28.

Rader, M., & Rahman, S. S. M. (2013). Exploring historical and emerging phishing techniques and mitigating the associated security risks. *International Journal of Network Security & Its Applications (IJNSA), 5*(4), 1–12. http://dx.doi.org/10.5121/ijnsa.2013.5402

Rancourt-Raymond, A. de, & Smaili, N. (2023). The unethical use of deepfakes. Journal of Financial Crime, 30(4), 1066–1077. http://dx.doi.org/10.1108/JFC-04-2022-0090

Riley, J. (2024). Ethical hacking and penetration testing: An overview. *Cyber Defense Quarterly, 9*(1), 45–60.

Shah, S. A. A., & Al-Emran, M. (2024). Deepfake detection techniques: A review. *IEEE Access, 12*, 4567–4583.

Schmitt, M., Flechais, I. Digital deception: generative artificial intelligence in social engineering and phishing. Artif Intell Rev 57, 324 (2024). https://doi.org/10.1007/s10462-024-10973-2

Short, E., Guppy, A., Hart, J. A., & Barnes, J., (2015). The impact of cyberstalking. Studies in Media and Communication, 3(2), 23-37. http://dx.doi.org/10.11114/smc.v3i2.970

Solove, D. J. (2003). Identity theft, privacy, and the architecture of vulerability. Hastings Law Journal, 54, 1228-1276 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=416740

Symantec. (2023). *Internet security threat report*. https://www.symantec.com/security-center/threat-report

Seymour J, Tully P (2018) Generative models for Spear Phishing. Posts on Social Media https://doi.org/10.48550/arXiv.1802.05196

Teichmann, F. Ransomware attacks in the context of generative artificial intelligence—an experimental study. Int. Cybersecur. Law Rev. 4, 399–414 (2023). https://doi.org/10.1365/s43439-023-00094-x

Trend Micro. (n.d.). Cybersecurity solutions. Retrieved June 22, 2025, from https://www.trendmicro.com/en_us/business/cybersecurity-solutions.html

Villegas-Ch., W., Govea, J., Gurierrez, R., Navarro, A. M., Maldonado Navarro, A., & Mera-Navarrete, A. (2024). *Effectiveness of an adaptive deep learning-based intrusion detection system*. *IEEE Access, 12*, 184010–184027. https://doi.org/10.1109/ACCESS.2024.3512363

Wall, D.S. (2024) Cybercrime: The Transformation of Crime in the Information Age, 2nd edition, Cambridge, Cambridge: Polity. ISBN-10: 0745653529 / ISBN-13: 978-0745653525 https://www.researchgate.net/profile/David-Wall-7/publication/378013252_Cybercrime_The_Transformation_of_Crime_in_the_Information_Age_2nd_edition/links/65c36f3179007454976a5420/Cybercrime-The-Transformation-of-Crime-in-the-Information-Age-2nd-edition.pdf

Weimann, G. (2016). *Terrorism in cyberspace: The next generation*. Columbia University Press.

Wilson, C. (2023). The rise of AI in cybersecurity: Challenges and opportunities. *Journal of Information Security, 11*(4), 223–240.

World Economic Forum. (2024). *Global cybersecurity outlook 2024*. https://www.weforum.org/reports/global-cybersecurity-outlook-2024

Worsley, J., Wheatcroft, J. M., Short, E., & Corcoran, R. (2017). Victims' voices: Understanding the emotional impact of cyberstalking and individuals' coping responses. *SAGE Open, 7*(2), 1–10. http://dx.doi.org/10.1177/2158244017710292

Yar, M. (2016) Cybercrime and society. SAGE Publications. https://archive.org/search.php?query=external-identifier%3A%22urn%3Alcp%3Acybercrimesociet0000yarm%3Aepub%3Add4c8c69-3a29-4d81-af84-1ee68591d1a3%22

Zhu, B., Fang, H., Sui, Y., & Li, L. (2020). *Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation*. *ArXiv*. https://arxiv.org/abs/2003.00813

Legislação Consultada

Decreto-Lei n. º48/95, de 15 de março. Código Penal Português. Diário da República: I-A Série, n. º63/1995. Disponível em: https://dre.pt/application/conteudo/185720 [acedido em 01/06/2025].

Lei n. º12/91, de 21 de maio. Lei da Identificação Civil e Criminal. Diário da República: I-A Série, n. º116. Assembleia da República.

Disponível em: https://dre.pt/application/conteudo/640075 [acedido em 01/06/2025].

Lei n. º109/2009, de 15 de setembro. Lei do Cibercrime. Diário da República: I Série, n. º 179. Assembleia da República.

Disponível em: https://dre.pt/application/conteudo/489693 [acedido em 01/06/2025].

Anexos

Anexo A - Questionário

Pode ser distribuído via formulário online (ex: Google Forms, LimeSurvey, Qualtrics), com as mesmas instruções de anonimato e consentimento.

Secção 1: Contexto Profissional

- Função profissional atual:
- Anos de experiência na área:
- Já realizou testes com ferramentas de IA? (Sim/Não)

Secção 2: Avaliação Técnica

- Indique as ferramentas que já utilizou:
  - ☐ FortiEDR
  - ☐ CrowdStrike
  - ☐ SentinelOne
  - ☐ Microsoft Defender
  - ☐ Outras: _____
- Em que grau concorda com as seguintes afirmações? (1 - Discordo totalmente / 5 - Concordo totalmente):
  - As ferramentas de IA detetam eficazmente ameaças.
  - A resposta automática é geralmente adequada.
  - Há poucos falsos positivos.
  - A IA responde bem a ataques de engenharia social.
  - Os sistemas conseguem lidar com deepfakes/anomalias.

Secção 3: Considerações Éticas

- Que precauções considera necessárias no uso da IA para segurança?
- Considera que a IA deve ter sempre supervisão humana? (Sim/Não/Depende do contexto)

Medidas de Proteção e Análise

- **Identificação anónima:** Cada participante recebe um código (ex: WHH-01, WHH-02).
- **Registo seguro:** Dados armazenados em pastas protegidas com encriptação.
- **Análise:** Só será feita em formato agregado, sem nomes ou dados identificáveis.
- **Consentimento informado:** Documento anexo com as condições do estudo e opção de retirada.

Anexo B - Entrevista Semiestruturada (Alternativa)

Introdução (a ser lida pelo investigador):

"Esta entrevista integra uma investigação académica sobre a eficácia de ferramentas de cibersegurança baseadas em inteligência artificial. A sua participação é totalmente voluntária. Não será recolhida nenhuma informação pessoal identificável. Será atribuído um pseudónimo ou código para garantir o anonimato. Os dados serão tratados com total confidencialidade, apenas para fins académicos. Poderá desistir da entrevista a qualquer momento."

Informação Técnica Geral

1. Qual é o seu campo profissional atual? (ex: segurança informática, análise forense digital, etc.)
2. Há quantos anos atua na área da cibersegurança?
3. Já participou anteriormente em testes de intrusão ou auditorias de sistemas operativos com IA? (Sim/Não)

Sobre Ferramentas de IA em Cibersegurança

4. Quais ferramentas de cibersegurança com IA já utilizou ou testou? (ex: FortiEDR, CrowdStrike, SentinelOne, Microsoft Defender for Endpoint)
5. Na sua experiência, quais são os principais pontos fortes destas ferramentas?
6. E quais as suas principais limitações, especialmente no contexto de ataques personalizados ou engenharia social?

Sobre Métodos de Avaliação e Resposta

7. Em simulações controladas, como avalia a capacidade de deteção e resposta automática dos sistemas com IA?
8. Qual a sua perceção sobre o número de falsos positivos/negativos gerados por estas ferramentas?
9. Já participou em testes que envolveram deepfakes ou anomalias comportamentais? Se sim, como foi a performance da IA?

Visão Crítica e Ética

10. Considera que os sistemas com IA estão prontos para substituir o julgamento humano na defesa de sistemas operativos?
11. Em termos éticos, que precauções considera fundamentais quando se utiliza IA para proteger sistemas informáticos?

Encerramento

"Agradeço a sua colaboração. As respostas serão codificadas e usadas de forma agregada. Se desejar receber os resultados da investigação, pode deixar um e-mail de contacto num formulário separado. Obrigado."

Anexo C - Formulário de Relato de Progresso dos Testes de Cibersegurança IA

Objetivo: Colecionar dados padronizados sobre o desempenho dos sistemas de cibersegurança com IA durante os testes controlados realizados pelos *white hat hackers*.

1. Informações Gerais (usando pseudónimo/código)

- Pseudónimo/código do colaborador: _____
- Data do teste: //____
- Duração aproximada do teste (em horas): _____

2. Sistema Operativo Testado

- ( ) Windows
- ( ) Linux
- ( ) macOS
- ( ) Outro: _____

3. Ferramenta de Cibersegurança com IA Avaliada

- ( ) FortiEDR
- ( ) CrowdStrike
- ( ) SentinelOne
- ( ) Microsoft Defender for Endpoint
- ( ) Outra: _____

4. Tipos de Ataques Simulados

- ( ) Phishing / Spear Phishing
- ( ) Ransomware
- ( ) Exploit Kits

- ( ) Engenharia Social
- ( ) Penetration Testing Geral
- ( ) Outro: _____

5. Avaliação da Eficiência do Sistema de IA

Para cada item, assinale a opção que melhor descreve o desempenho da ferramenta:

- Deteção precoce dos ataques
  ( ) Excelente
  ( ) Bom
  ( ) Regular
  ( ) Insuficiente
  ( ) Não avaliado
- Resposta automática (bloqueio, quarentena, alertas)
  ( ) Excelente
  ( ) Bom
  ( ) Regular
  ( ) Insuficiente
  ( ) Não avaliado
- Número de falsos positivos (alertas incorretos)
  ( ) Nenhum
  ( ) Poucos
  ( ) Moderado
  ( ) Muitos
- Número de falsos negativos (ataques não detetados)
  ( ) Nenhum
  ( ) Poucos
  ( ) Moderado
  ( ) Muitos

6. Comentários e Observações Gerais

- Descreva quaisquer dificuldades encontradas, vulnerabilidades específicas exploradas com sucesso, ou limitações do sistema que foram observadas:

7. Sugestões para Melhorias

- Alguma recomendação para aumentar a eficácia do sistema de cibersegurança?

8. Confirmação

- Confirmo que os dados fornecidos são precisos e que respeitei os protocolos éticos do estudo.
  ( ) Sim

Instruções para envio:

- Envie este formulário preenchido ao investigador responsável (Bernardo Velasquez Borges) pelo email: **investigacaoai@gmail.com**
- Para qualquer dúvida, utilize o mesmo email para contato.