bilingual edition

ptsoc {news}

A dualidade da inteligência artificial

3 perguntas a Arlindo Oliveira

Pentesting por Nelson Silva

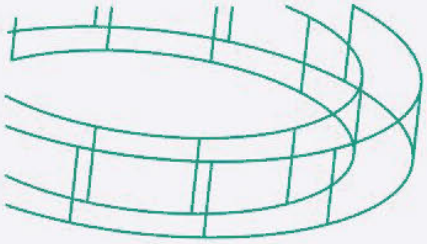
09

The duality of artificial intelligence

3 questions to Arlindo Oliveira

Pentesting by Nelson Silva

.pt



Iniciativa Portuguesa do Fórum da Governação da Internet 2023

12 de julho, 9h30 - Sede da ANACOM e online

PROGRAMA

09h30 Boas Vindas
Patrícia Silva Gonçalves, Vogal do Conselho de Administração, ANACOM

09h45 Abertura
Mário Campolargo, Secretário de Estado da Digitalização e da Modernização Administrativa

10h00 Keynote speaker: Digital Global Compact
Lise Fuhr, ETNO

10h30 Coffee Break

10h45 Connect all people to the internet, including all schools
Moderação: Fátima Caçador, Casa dos Bits

Oradores:
Luís Gaspar, ANACOM
Ricardo Salgado, dstelecom
Martijn van Delden, Amazon Kuiper

Relator: Manuel da Costa Cabral, ANACOM

11h30 Data Governance and trust/Misleading content
Moderação: Marta Moreira Dias, .PT

Oradores:
Graça Canto Moniz, Faculdade de Direito da Universidade Nova de Lisboa
Luis Silveira Botelho, IGAC
Paulo Fonseca, DECO
Pedro Portugal Gaspar, ASAE
Pedro Verdelho, Procurador da República

Relator: Luís Pisco, DECO

12h30 Keynote speaker
Isabel Ferreira, Secretária de Estado do Desenvolvimento Regional

12h45 Almoço

14h00 Avoid internet fragmentation and Digital commons as a global public good
Moderação: Ana Cristina Neves, FCT

Oradores:
Andrea Beccalli, ICANN
Fátima Caçador, Casa dos Bits
João Nuno Ferreira, FCT
Luisa Ribeiro Lopes, .PT & INCoDe.2030

Relator: André Silva, CNCS

15h00 Promote regulation of artificial intelligence
Moderação: Rogério Carapuça, APDC
Keynote speaker: Kilian Gross, Directorate-General CONNECT at the European Commission

Oradores:
Arlindo Oliveira, INESC
Giorgia Abeltino, Google
Magda Coco, VdA – Vieira de Almeida
Paulo Dimas, Center for Responsible AI

Relator: Isabel Travessa, Editora de conteúdos @ APDC

16h00 Coffee Break

16h15 Apresentação das mensagens pelos relatores
Moderação: Nuno Garcia, Faculdade de Ciências da Universidade de Lisboa

Oradores:
Manuel da Costa Cabral, ANACOM
Luís Pisco, DECO
André Silva, CNCS
Isabel Travessa, APDC

João Pedro Martins, Youth Coalition on Internet Governance

16h45 Encerramento

organização:

.pt ANACOM apdc digital business community CNCS DECO fct IAPMEI IGF

A dualidade da inteligência artificial

Os receios da má utilização da inteligência artificial (IA) no domínio da cibersegurança não são novos. A investigação nesta área antecipou potenciais problemas e, em 2021, constatou-se como "a investigação sobre métodos de ataque a sistemas de aprendizagem automática e de IA aumentou - com quase 2000 artigos publicados sobre o tema num repositório durante a última década - mas as organizações não adotaram estratégias proporcionais para garantir que as decisões tomadas pelos sistemas de IA são fiáveis", escreveu a Dark Reading. Também o relatório "The Road to Secure and Trusted AI" registou mais de 1500 artigos científicos sobre segurança da IA publicados em 2019 no ArXiv.org, quando três anos antes eram apenas 56.

A IA foi adotada sem precauções e defesas contra ataques que, embora não fossem numerosos, já ocorriam com alguma frequência e de forma diversificada: o Berryville Institute of Machine Learning identificou "78 diferentes ameaças aos modelos de aprendizagem automática ("machine learning" ou ML) e aos sistemas de IA", sendo as principais ameaças o "envenenamento de dados, a manipulação de sistemas online, os ataques a modelos comuns de ML e a exfiltração de dados".

The duality of artificial intelligence

Fears of the misuse of artificial intelligence (AI) in the cybersecurity domain are not new. Research in this area anticipated potential problems, and in 2021 it was found how 'research into methods of attacking machine-learning and artificial-intelligence systems has surged - with nearly 2 000 papers published on the topic in one repository over the last decade - but organizations have not adopted commensurate strategies to ensure that the decisions made by AI systems are trustworthy,' wrote Dark Reading. The 'The Road to Secure and Trusted AI' report also recorded more than 1 500 academic papers on AI security published in 2019 on ArXiv.org, when three years earlier there were only 56.

AI was adopted without precautions and defences against attacks which, although were not numerous, were already taking place somewhat frequently and in diverse forms: the Berryville Institute of Machine Learning identified '78 different threats to machine learning models and AI systems', the main threats being 'data poisoning, manipulation of online systems, attacks on common ML models and data exfiltration'.

As recently as 2021, cybersecurity expert Bruce Schneier warned in 'The Coming AI

Ainda em 2021, o especialista em cibersegurança Bruce Schneier alertou em "The Coming AI Hackers" como os sistemas de IA poderiam evoluir para, em modo automático e autónomo, se tornarem hackers sociais, económicos e políticos. A "IA hacker" não procura aceder ilegalmente aos sistemas, mas sim solucionar problemas criados pelos humanos. É software que "vai hackear a nossa sociedade a um nível e efeito sem precedentes" e de "duas formas muito diferentes.

Hackers' how AI systems could evolve so they become, in automatic and autonomous mode, social, economic and political hackers. 'AI hacker' does not seek to illegally access systems, but to solve problems created by humans. It's a software that 'will hack our society to a degree and effect unlike anything that's come before' and in 'two very different ways.

One, AI systems will be used to hack us. And two, AI systems will themselves become hackers: finding vulnerabilities in

Primeiro, os sistemas de IA serão utilizados para nos piratear. Em segundo, os sistemas de IA tornar-se-ão eles próprios hackers: encontrando vulnerabilidades em todo o tipo de sistemas sociais, económicos e políticos e explorando-as a

all sorts of social, economic and political systems, and then exploiting them at an unprecedented speed, scale, and scope. It's not just a difference in degree; it's a difference in kind. We risk a future of AI systems hacking other AI systems, with

uma velocidade, escala e alcance sem precedentes. Não se trata apenas de uma diferença de grau; é uma diferença de género. Arriscamo-nos a um futuro em que os sistemas de IA vão hackear outros sistemas de IA, sendo os seres humanos pouco mais do que danos colaterais".

Não estamos ainda nesse patamar catastrofista, mas a análise dos novos desafios que a IA coloca à sociedade - e à cibersegurança, em particular – têm-se sucedido, levando mesmo à criação de uma nova "terminologia e taxonomia para a IA". Em Setembro de 2022, dois meses antes do "boom" do lançamento do popular ChatGPT, a PTSOC {news} #6 já propunha uma análise conceptual aos "Desafios da IA na cibersegurança".

Atualmente, enquanto as empresas de IA distribuem as suas aplicações pelos mais diversos mercados, as empresas de cibersegurança procuram incorporar nos seus produtos algumas ferramentas de defesa contra ciberataques. Noutros casos, alguns profissionais questionam o ChatGPT ou outros programas de IA para obter código mais seguro e eficaz, "sugerindo que a IA generativa pode ser outro indicador de segurança que os profissionais terão que contextualizar para um uso mais amplo", como refere Jon France, CISO da (ISC)2, considerando que ainda "estamos a começar a ver os primeiros usos no mundo da segurança".

humans being little more than collateral damage'.

We have not yet reached that catastrophic level, but the analysis of the new challenges that AI poses to society - and cybersecurity, in particular - has been successful, even leading to the creation of a new 'terminology and taxonomy for AI'. In September 2022, two months before the launch boom of the popular ChatGPT, PTSOC {news} #6 already presented a conceptual analysis of the 'Challenges of AI in cybersecurity'.

Currently, while AI companies distribute their applications across a wide range of markets, cybersecurity companies are looking to incorporate some defence tools against cyberattacks into their products. In other cases, some professionals are asking ChatGPT or other AI programs to write a more secure and effective code, 'suggesting generative AI could be another indicator of health that professionals will have to contextualise for broader use', Jon France notes, CISO at (ISC) 2, considering that 'we're starting to see the early, early uses of it in the security world'.

On the other hand, as in biology, attempts are being made to control the dissemination of sensitive information as a decisive step towards achieving greater 'security through secrecy'. The

ChatGPT: quando a criação é usada para atacar o criador

Mais de 100 mil contas do ChatGPT comprometidas estiveram à venda na Dark Web no ano passado. A quantidade de contas roubadas aumentou progressivamente de 74 em junho de 2022 para quase 27 mil em maio passado.

Noutro caso mais recente, uma falsa aplicação do ChatGPT comprometeu as contas de mais de quatro milhões de utilizadores. Distribuída como extensão do Chrome e software de desktop do Windows, roubou as credenciais dos utilizadores, contornando a autenticação de dois fatores. Os utilizadores do Facebook foram bloqueados pela app no acesso às suas contas, com alteração de nome e perfil para se assemelhar a Lily Collins, a atriz da série da Netflix "Emily in Paris". A OpenAI confirmou a violação de dados e desativou o serviço até o problema ser corrigido.

ChatGPT: when the creation is used to attack the creator

More than 100 000 compromised ChatGPT accounts were found for sale on the Dark Web last year. The amount of stolen accounts steadily increased from 74 in June 2022 to nearly 27 000 last May.

In a more recent case, a fake ChatGPT app compromised more than 4 million accounts. Distributed as both a Chrome extension and a Windows desktop software, it stole users' credentials by bypassing two-factor authentication. Facebook users were blocked by the app from accessing their accounts, with their name and user profile being changed to resemble Lily Collins, the actress from the Netflix series 'Emily in Paris'. OpenAI confirmed the data breach and disabled the service until the issue was fixed.

Em sentido contrário, como sucede na biologia, tenta-se controlar a disseminação da informação sensível como um passo decisivo para obter uma maior "segurança através do segredo". O objetivo é não ter a IA a ajudar criminosos informáticos isolados ou pagos por Estados a desenvolverem ataques mais elaborados contra as infraestruturas tecnológicas.

goal is not to have AI help isolated or state-paid cybercriminals to develop more elaborate attacks against technological infrastructures.

'Miscreants can use ML software to develop more authentic-seeming phishing lures and craft better ransom notes, while also scanning larger volumes of data for sensitive info they can mone-

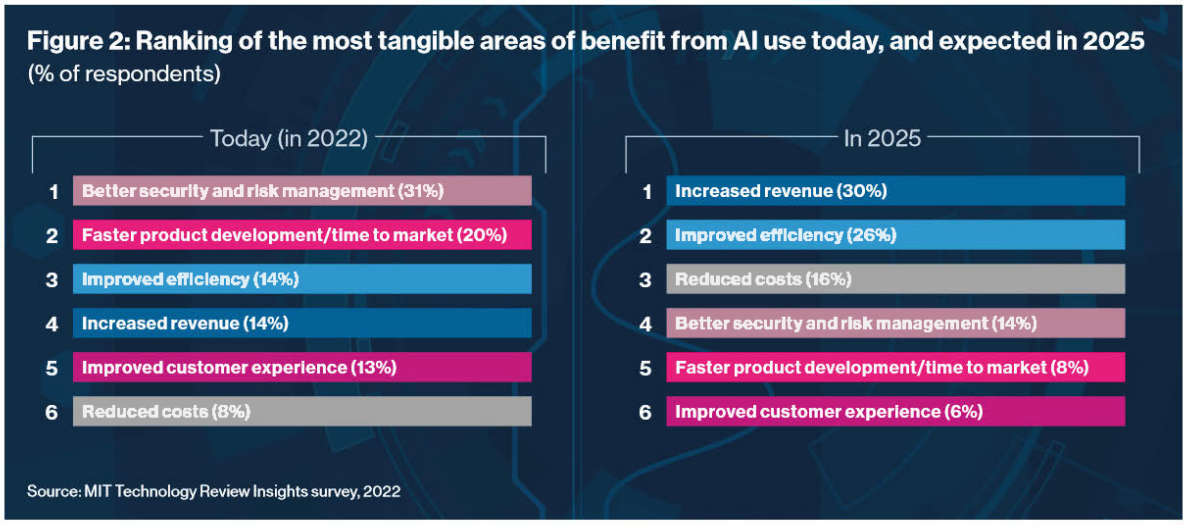
"Os criminosos podem utilizar software de ML para desenvolver engodos de phishing mais credíveis e elaborar melhor ransomware, enquanto analisam grandes volumes de dados em busca de informações sensíveis que possam rentabilizar", criar código otimizado para ataques de malware ou recolher informações sobre potenciais alvos, antecipa Rob Joyce, diretor do Cybersecurity Directorate da National Security Agency (NSA).

Em paralelo, diferentes vozes acreditam que a melhor defesa contra ataques perpetrados com IA virá da própria IA. As grandes empresas do setor estão a trabalhar nisso e, por exemplo, a Google anunciou o pacote de segurança Cloud Security AI Workbench, após a Microsoft ter apresentado o Security Copilot.

tize', create code optimised for malware attacks or gather information on potential targets, anticipates Rob Joyce, director of the National Security Agency (NSA)'s Cybersecurity Directorate.

In parallel, different voices believe that the best defence against attacks perpetrated with AI will come from AI itself. Major companies in the industry are working on this. Google, for example, has announced the Cloud Security AI Workbench security suite, after Microsoft introduced Security Copilot.

These organisations' recent response also comes after the generalised use of AI. According to a survey by the MIT Technology Review Insights, only 6 % of companies overall stated not having used AI tools last year.



A recente resposta destas organizações deve-se também à generalização da IA. Apenas 6% das empresas em geral afirmava no ano passado não usar ferramentas de IA, segundo um inquérito do MIT Technology Review Insights.

Ataques por chatbots

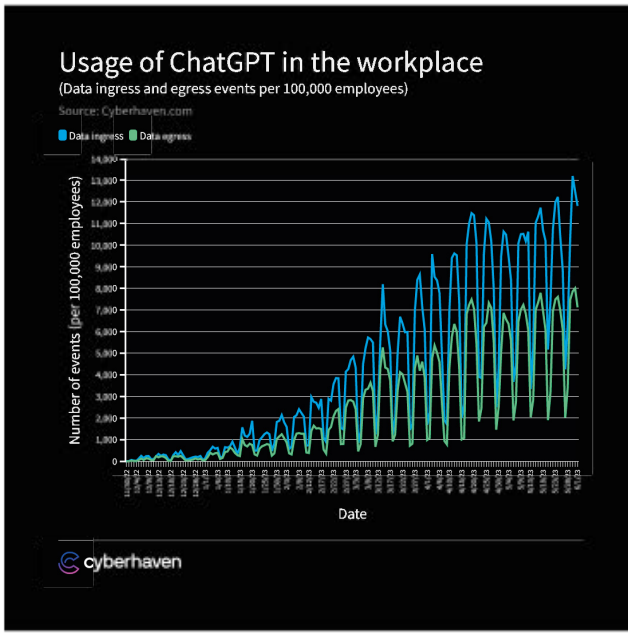
No início deste ano, foi revelado que o ChatGPT já estava a ser usado para criar ferramentas básicas de ataques por malware, para encriptação e para a criação de mercados na Dark Web. Em Abril, um programador que assumiu não ter experiência no desenvolvimento de malware, criou com o ChatGPT um destes programas indetetável para as ferramentas de proteção.

Pelo seu lado, a OpenAI, responsável pelo desenvolvimento deste chatbot, anunciou em Abril um "Bug Bounty Program", à qual responderam 4.500 programadores interessados em receber até 20 mil dólares por "descobertas extraordinárias" de vulnerabilidades. Mas os utilizadores continuam a ser o lado mais frágil, quando o sistema operativo Windows se adapta à IA, incluindo um chatbot no Bing, e a Internet poderá evoluir para um futuro cenário preocupante.

No início de Junho, cerca de 10% dos 1,6 milhões de funcionários de diferentes indústrias que recorrem a ferramentas

Attacks by chatbots

Earlier this year, it was revealed that ChatGPT was already being used to create basic malware attack tools, for encryption and for the creation of marketplaces on the Dark Web. In April, a programmer who assumed he had no experience in malware development, created one of these undetectable programs for protection tools using ChatGPT.



OpenAI, responsible for the development of this chatbot, announced in April a 'Bug Bounty Program', to which 4 500 programmers interested in receiving up to 20 thousand dollars for 'extraordinary discoveries' of vulnerabilities responded. But users remain the weakest side, when

de segurança de dados da Cyberhaven, usaram o ChatGPT no trabalho e 8,6% transferiu dados da sua empresa para o chatbot, com 4,7% a assumir que se tratava de informação confidencial. Perante esta tendência, organizações como a Alphabet (criadora do chatbot Bard), Amazon, Deutsche Bank, JP Morgan, Samsung ou Verizon estão a bloquear o uso interno do ChatGPT, embora outras considerem que estes riscos técnicos, de fraudes, de conformidade ("compliance") ou legais, nomeadamente envolvendo outras empresas, são exagerados ou prematuros.

É este ambiente dual, entre o uso positivo e as aplicações negativas da IA na cibersegurança, que se deverá calcorrear nos próximos tempos. Há muito dinheiro investido pelas grandes empresas tecnológicas para controlar a IA, o desejo de regulação por parte dos Estados e uma enorme vontade dos utilizadores em terem ferramentas que os possam ajudar na produtividade laboral. E ainda só estamos no início deste novo paradigma, desta "nova Revolução Industrial".

Windows' operating system adapts to AI, and when there's even a Bing chatbot; the Internet could evolve into a worrisome future scenario.

At the beginning of June, about 10 % of the 1.6 million employees in different industries that use Cyberhaven's data security tools used ChatGPT at work and 8.6 % transferred data from their company to the chatbot, with 4.7 % admitting it was confidential information. Faced with this trend, companies such as Alphabet (creator of the Bard chatbot), Amazon, Deutsche Bank, JP Morgan, Samsung or Verizon are blocking the internal use of ChatGPT, although others consider that these technical, fraud, compliance or legal risks, namely those involving other companies, are exaggerated or premature.

It is this dual environment, between the positive use and the negative applications of AI in cybersecurity, that will have to be navigated in the near future. There is a lot of money invested by big tech companies to control AI, a desire for regulation by States, and a huge desire by users to have tools that can help them with work productivity. And we are still only at the beginning of this new paradigm, this 'new Industrial Revolution'.

Foto: Freepik.com

IA como desastre na cibersegurança

Os chatbots como o Bard, Bing ou ChatGPT, são um "desastre para a segurança" principalmente por três razões:

"Jailbreaking": ao produzir texto que se lê como escrito por um ser humano, "seguem instruções ou 'prompts' do utilizador e, em seguida, geram uma frase prevendo, com base nos seus dados de treino, a palavra que mais provavelmente se segue a cada palavra anterior. Mas o que torna estes modelos tão bons - o facto de poderem seguir instruções - também os torna vulneráveis a uma utilização indevida. Isso pode acontecer através de 'injecções de instruções' ('prompt injections'), em que alguém utiliza instruções que levam o modelo de linguagem a ignorar as suas instruções anteriores e as barreiras de segurança".

AI as a disaster for cybersecurity

Chatbots like Bard, Bing or ChatGPT are a 'security disaster' for three main reasons:

Jailbreaking: By producing text that reads as something written by a human being, 'they follow instructions or prompts from the user and then generate a sentence by predicting, on the basis of their training data, the word that most likely follows each previous word. But the very thing that makes these models so good - the fact they can follow instructions - also makes them vulnerable to being misused. That can happen through "prompt injections" in which someone uses prompts that direct the language model to ignore its previous directions and safety guardrails.'

Ajuda em burlas e phishing: a OpenAI permite integrar o ChatGPT em produtos que interagem com a Internet, como "assistentes virtuais capazes de realizar ações no mundo real, reservar voos ou marcar reuniões nas agendas pessoais". Porque estes assistentes "melhorados por IA extraem texto e imagens da Web, estão expostos a um tipo de ataque designado por 'indirect prompt injection', em que um terceiro altera um sítio Web adicionando texto oculto destinado a alterar o comportamento da IA. Os atacantes podem utilizar as redes sociais ou o email para direcionar os utilizadores para sítios Web com essas instruções secretas. Quando acontece, o sistema de IA pode ser manipulado para permitir ao atacante extrair as informações do cartão de crédito, por exemplo".

"Envenenamento" dos dados: como estes modelos de IA são treinados usando dados extraídos da Internet, "atualmente, as empresas de tecnologia confiam apenas que esses dados não foram adulterados de forma maliciosa". No entanto, "investigadores descobriram ser possível envenenar o conjunto de dados que serve para treinar grandes modelos de IA", em sites criados com esse intuito ou ao "editar e acrescentar frases a entradas da Wikipédia que acabaram por ser incluídas no conjunto de dados de um modelo de IA".

Assisting scamming and phishing: OpenAI allows people to integrate ChatGPT into products that interact with the Internet, such as 'virtual assistants that are able to take actions in the real world, such as booking flights or putting meetings on people's calendars'. Because these 'AI-enhanced [virtual] assistants scrape text and images off the web, they are open to a type of attack called indirect prompt injection, in which a third party alters a website by adding hidden text that is meant to change the AI's behaviour. Attackers could use social media or email to direct users to websites with these secret prompts. Once that happens, the AI system could be manipulated to let the attacker try to extract people's credit card information, for example'.

Data poisoning: Because these AI models are trained using data scraped from the Internet, 'right now, tech companies are just trusting that this data won't have been maliciously tampered with'. However, 'researchers found that it was possible to poison the data set that goes into training large AI models' on websites created for that purpose or that are able to 'edit and add sentences to Wikipedia entries that ended up in an AI model's data set'.

Quadro de referência para boas práticas na segurança da IA

A Google apresentou um Secure AI Framework (SAIF) para levar as organizações a implementar seis ideias de boas práticas:

1. Expandir bases de segurança sólidas para o ecossistema de IA.

2. Alargar a deteção e a resposta para integrar a IA no universo das ameaças de uma organização.

3. Automatizar as defesas para acompanhar o ritmo das ameaças existentes e novas.

4. Harmonizar os controlos ao nível da plataforma para garantir uma segurança consistente em toda a organização.

5. Adaptar os controlos para ajustar as mitigações e criar ciclos de feedback mais rápidos para a implementação da IA.

6. Contextualizar os riscos do sistema de IA nos processos empresariais circundantes.

Ver também o Artificial Intelligence Risk Management Framework do National Institute of Standards and Technology (NIST) dos EUA.

Framework for best practices in AI security

Google introduced a Secure AI Framework (SAIF) to drive organisations to implement six best practice ideas:

1. Expand strong security foundations to the AI ecosystem.

2. Extend detection and response to bring AI into an organisation's threat universe.

3. Automate defences to keep pace with existing and new threats.

4. Harmonise platform level controls to ensure consistent security across the organisation.

5. Adapt controls to adjust mitigations and create faster feedback loops for AI deployment.

6. Contextualise AI system risks in surrounding business processes.

See also the Artificial Intelligence Risk Management Framework from the US National Institute of Standards and Technology (NIST).

Foto: Freepik.com

Novas técnicas para ataques mais perigosos
New techniques for more dangerous attacks

O SANS Institute revelou algumas das novas técnicas mais perigosas de ciberataques usando a IA:

Ataques antagónicos de IA: Com estes ataques, procura-se manipular ferramentas de IA para aumentar a velocidade das campanhas de ransomware e identificar vulnerabilidades "zero-day" em software complexo. Em resposta, as organizações precisam de implementar um modelo de segurança integrado de defesa que forneça proteções em

The SANS Institute has revealed some of the most dangerous new techniques for cyberattacks using AI:

AI antagonistic attacks: These attacks seek to manipulate AI tools to increase the speed of ransomware campaigns and identify zero-day vulnerabilities in complex software. In response, organisations need to implement an integrated defence security model that provides layered protections, automates critical detection and response actions, and

camadas, automatize ações críticas de deteção e resposta e facilite processos eficazes de tratamento de incidentes.

Engenharia social dinamizada pelo ChatGPT: Os agentes de ameaças estão a tirar partido da IA generativa para explorar o risco humano - visando as vulnerabilidades dos funcionários para violar a rede da sua organização, incluindo às suas famílias.

Ataques de programadores externos: Os ataques a programadores terceirizados (também conhecidos como ataques à cadeia de fornecimento de software), com um aumento de ataques direcionados a esses programadores para se infiltrarem nas redes das empresas através da cadeia de fornecimento.

Ataques de SEO e por publicidade paga: Os ataques usando a "search engine optimization" (SEO) são outro método perigoso e emergente, tal como os ataques de publicidade paga (também designados por malvertising). Eles tiram partido de estratégias de marketing fundamentais para obter acesso inicial às redes das empresas. Os autores dos ataques exploram palavras-chave de SEO e anúncios pagos para enganar as vítimas e levá-las a entrar em sites falsos, descarregar ficheiros maliciosos e permitir o acesso remoto a outros utilizadores.

facilitates effective incident handling processes.

Social engineering powered by ChatGPT: Threat actors are leveraging generative AI to exploit human risk - targeting employees' vulnerabilities to breach their organisation's network, including their families.

External programmer attacks: Attacks on third-party programmers (also known as software supply chain attacks), with an increase in attacks targeting these programmers to infiltrate companies' networks through the supply chain.

SEO and paid advertising attacks: Attacks using search engine optimisation (SEO) are another dangerous and emerging method, as are paid advertising attacks (also known as malvertising). They take advantage of fundamental marketing strategies to gain initial access to companies' networks. Attackers exploit SEO keywords and paid advertisements to trick victims into entering fake websites, downloading malicious files and allowing remote access to other users.

Foto: Freepik.com

Onde está o talento nacional?

Lisboa e Porto são as únicas cidades reconhecidas no Atlas da Sequoia Capital sobre o talento tecnológico. Dos três milhões de engenheiros na Europa, a capital nacional conta com quase 26 mil técnicos, focados principalmente no desenvolvimento de aplicações e bases de dados. Critical TechWorks, Microsoft, Siemens e OutSystems são os principais empregadores. O Porto tem quase 19 mil engenheiros, dedicados principalmente ao desenvolvimento de aplicações, tendo como principais empregadores a Farfetch e, novamente, a Critical Techworks. Cerca de 14% dos engenheiros nacionais são mulheres.

Where is the Portuguese talent?

Lisbon and Porto are the only cities recognised in Sequoia Capital's Atlas of technology talent. Of the 3 million engineers in Europe, Lisbon has almost 26 000 technicians, focused mainly on the development of applications and databases. Critical TechWorks, Microsoft, Siemens and OutSystems are the main employers. Porto has almost 19 000 engineers, mainly dedicated to application development, with Farfetch and, again, Critical Techworks as the main employers. About 14 % of Portuguese engineers are women.

Mercado da cibersegurança cresce

O mercado mundial da cibersegurança cresceu 12,5% em termos homólogos no primeiro trimestre do ano, para os 17 mil milhões de euros. O maior aumento ocorreu na América do Norte, seguindo-se a região EMEA, a Ásia-Pacífico e a América Latina.

Growing cybersecurity market

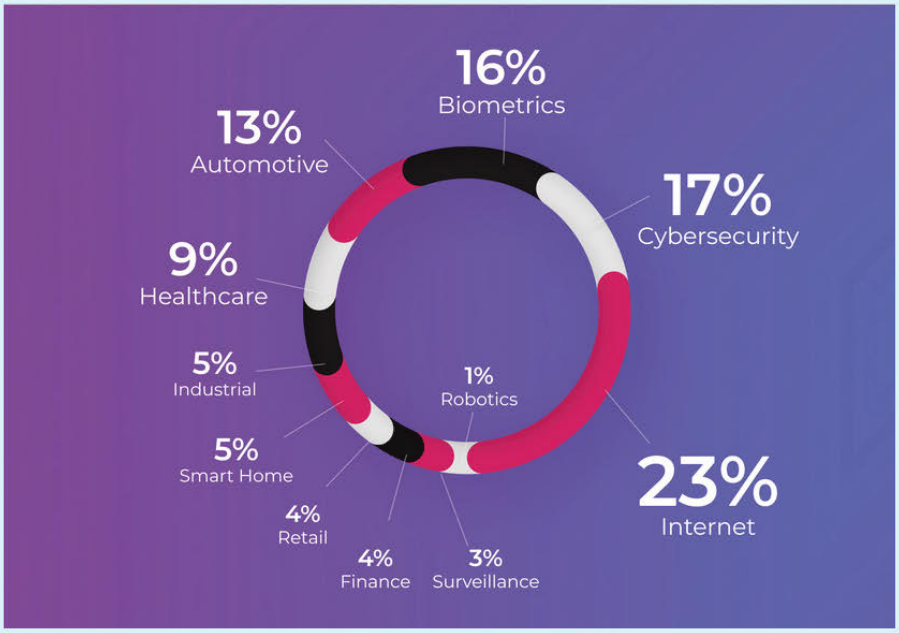
The global cybersecurity market grew by 12.5 % year on year in Q1 2023 to €17 billion. The largest increase occurred in North America, followed by the EMEA region, Asia Pacific and Latin America.

The Road to Secure and Trusted AI

Uma retrospetiva com previsões e análise para o domínio da segurança da IA, apontando-se as aplicações de IA de indústrias altamente críticas. A Internet (com 23%), a cibersegurança (17%) ou a biometria (16%) "estão entre as mais vulneráveis pelo número de problemas de segurança publicados e que já sofreram vários incidentes".

The Road to Secure and Trusted AI

A retrospective with predictions and analysis for the AI security domain, pointing out the AI applications of highly critical industries. The Internet (23 %), cybersecurity (17 %) or biometrics (16 %) 'are among the most vulnerable by the number of security issues published and that have already suffered several incidents.'



3

Arlindo Oliveira

Presidente do INESC
President of INESC

1. Quais serão os impactos esperados da IA na cibersegurança?

A componente de analítica da IA há muito tempo que representa uma ferramenta importante na análise de registos (*logs*) e identificação de riscos de segurança. Nesse aspeto, não existe nada de novo relativamente às tecnologias que já estão a ser usadas de forma intensa nesta análise.

O desenvolvimento e disponibilidade de grandes modelos de linguagem poderá vir a abrir novos campos de aplicação para a IA, que poderão ser usados na análise de uma muito mais vasta coleção de dados, incluindo textos de email, código fonte e registos de voz. Isso poderá permitir o desenvolvimento de novas abordagens para a identificação mais atempada de novos ataques que usem engenharia social, baseados em emails, mensagens ou telefonemas.

Uma vez que os modelos de linguagem serão inevitavelmente usados na criação de novos ataques, com mensagens personalizadas para cada utilizador, é fundamental que exista uma

1. What are the expected impacts of artificial intelligence (AI) on cybersecurity?

AI's analytics component has long represented an important tool in analysing logs and identifying security risks. In that regard, there is nothing new about the technologies that are already being used heavily for this analysis.

The development and availability of large language models may open new fields of application for AI - which can be used to analyse a much larger collection of data, including email text, source code and voice logs. This could allow the development of new approaches for a timelier identification of new attacks using social engineering, based on emails, messages or phone calls.

Since language models will inevitably be used in the creation of new attacks, with messages tailored to each user, it is critical that there is a response from security services to prevent new AI technologies from increasing security risks to systems.

2. As a partner of the Centre for Responsible AI, what can we expect from this centre and how do you anticipate the evolution of the debate on AI ethics when, as you have said, 'whoever dominates AI will dominate the planet's economy'?

The Centre for Responsible AI focuses on four relevant dimensions: equality (absence

resposta por parte dos serviços de segurança para evitar que as novas tecnologias de IA venham a aumentar os riscos para a segurança dos sistemas.

2. Como partner do Center for Responsible AI, o que se pode esperar deste centro e como antecipa a evolução do debate sobre a ética na IA quando, como já disse, "quem dominar a IA vai dominar a economia do planeta"?

O Center for Responsible AI foca-se em quatro dimensões relevantes: igualdade (ausência de enviesamentos, garantia de tratamento equalitário), explicabilidade (a capacidade para analisar e perceber o racional subjacente às decisões), privacidade (preservar o direito à privacidade, de acordo com a legislação existente e as melhores práticas existentes) e sustentabilidade (desenvolver sistemas que reduzam ou eliminem o impacto ambiental e contribuam para um planeta mais sustentável).

Acreditamos que estas quatro dimensões serão importantes no futuro a curto, médio e longo prazo, e que legislação conducente a garantir padrões mínimos para estas dimensões será adotada pela União Europeia e, possivelmente, por outros países. A aprovação da proposta do Artificial Intelligence Act pelo Parlamento Europeu reflete estas preocupações, mas mostra também que é difícil identificar o ponto de equilíbrio entre a regulamentação adequada e a criação de condições para o desenvolvimento de tecnologias inovadoras. Será através da

of bias, guarantee of equal treatment), explainability (the ability to analyse and understand the rationale behind decisions), privacy (preserving the right to privacy, pursuant to existing legislation and best practices) and sustainability (developing systems that reduce or eliminate environmental impact and contribute to a more sustainable planet).

We believe that these four dimensions will be important in the future in the short-, medium- and long-term, and that legislation leading to guarantee minimum standards for these dimensions will be adopted by the European Union and possibly even other countries. The approval of the Artificial Intelligence Act proposal by the European Parliament reflects these concerns, but also shows that it is difficult to identify the balance point between appropriate regulation and the creation of conditions for the development of innovative technologies. With figuring out what this balance point is, Europe will be able to position itself not only as a pioneer in terms of legislation and regulation, but also through the ability to create innovative AI-based solutions.

3. In which areas in the field of AI can Portugal stand out?

Portugal has excellent human resources in this area and several companies are seen as international benchmarks. Through the Recovery and Resilience Programme,

identificação deste ponto de equilíbrio que a Europa se poderá posicionar não só como pioneira em termos de legislação e regulamentação, mas também através da capacidade de criar soluções inovadoras baseadas em IA.

3. Em que áreas da IA pode Portugal destacar-se?

Portugal tem excelentes recursos humanos nesta área e diversas empresas são referências internacionais. Apostou, também, através do Programa de Recuperação e Resiliência, em desenvolver competência. Porém, é necessário perceber que ela é muito importante para a competitividade dos países e que existem enormes investimentos feitos por todos os blocos económicos. Portugal, com cerca de 10 milhões de habitantes, não poderá nunca desempenhar um papel central no desenvolvimento das tecnologias mais exigentes em termos de recursos.

Não é fácil identificar em que subáreas Portugal possa ser especialmente competitivo num panorama que é muito global, mas áreas relacionadas com a língua portuguesa e as suas aplicações nas áreas jurídicas, de serviço ao cliente e da produção artística e cultural poderão ser particularmente promissoras. Isso não nos deverá impedir de competir e identificar nichos de mercado em áreas que são globalmente importantes, como a segurança, a logística, os serviços, a agricultura e a indústria. Numa fase de rápido desenvolvimento e grande acessibilidade da tecnologia, as inovações poderão ter origem em qualquer país, incluindo Portugal.

Portugal has also invested in the development of our skills in this area. However, we need to understand that it's very important for the competitiveness of countries and that there are huge investments made by all economic powers. Portugal, with close to 10 million inhabitants, will never be able to play a central role in the development of the most resource-demanding technologies.

It's not easy to identify in which sub-areas Portugal can be especially competitive in a very global panorama; however, areas related to the Portuguese language and its applications in the legal areas, customer service and artistic and cultural production could be particularly promising. This should not stop us from competing and identifying niche markets in globally important areas, such as security, logistics, services, agriculture and industry. At a time of rapid development and great accessibility of technology, innovations may come from any country, including Portugal.

Nelson Silva

Technical Manager na CyberSafe
Technical Manager at CyberSafe

Pentesting

A segurança da informação é um tema cada vez mais relevante no mundo digital, tendo como objetivo garantir a segurança de sistemas e redes. Neste contexto, uma prática amplamente usada é o *Pentesting*, também conhecido como Testes de Intrusão.

O principal objetivo do *Pentesting* é simular um ataque real fornecendo uma visão precisa do estado atual de segurança de um sistema ou rede. Isso permite que as organizações e profissionais de segurança compreendam as vulnerabilidades das suas infraestruturas antes que estas sejam exploradas por atacantes mal-intencionados.

A prática do *Pentesting* deve ser realizada por profissionais experientes e qualificados, conhecidos como *pentesters.* Esses profissionais devem possuir bons conhecimentos em técnicas de ataque, sistemas e redes, aderindo sempre a padrões éticos e legais, a fim de garantir que as suas ações estejam alinhadas com os interesses da organização e dentro dos limites estabelecidos previamente.

Pentesting

Information security is an increasingly relevant discussion in the digital world, aiming to ensure the security of systems and networks. In this context, a widely used practice is Pentesting, also known as Intrusion Testing.

The main goal of Pentesting is to simulate a real attack by providing an accurate view of a system or network's current security status. This enables organisations and security professionals to understand the vulnerabilities in their infrastructure before they are exploited by malicious attackers.

Pentesting should only be carried out by experienced and qualified professionals, known as pentesters. These professionals should have good knowledge of attack techniques, systems and networks, always adhering to ethical and legal standards in order to ensure that their actions are aligned with the interests of the organisation and within previously established limits.

Tipicamente, o *Pentesting* poderá seguir uma das três abordagens possíveis:

Black-box – nessa abordagem, o *pentester* não possui conhecimento prévio do sistema ou rede que será testado. Isso simula um ataque realizado por um hacker, fornecendo uma avaliação realista da capacidade de defesa da organização.

White-box – nessa abordagem, o *pentester* tem acesso total às informações sobre o sistema, incluindo arquitetura, códigos-fonte e diagramas de rede. Essa abordagem permite uma análise mais aprofundada e precisa das vulnerabilidades existentes, fornecendo uma visão completa da segurança do sistema.

Grey-box – nessa abordagem, o *pentester* tem conhecimento parcial sobre o sistema ou rede que será testado, mas não terá acesso a credenciais ou detalhes de configurações. Esta abordagem ajuda o *pentester* a focar o seu tempo e esforço para obter uma visão mais precisa do que um atacante real poderia encontrar.

A metodologia a seguir num serviço de *Pentesting* envolve as seguintes etapas:

Planeamento e recolha de informação: Nesta fase, o *pentester* recolhe informações sobre o sistema ou rede que será testado, identificando os pontos fortes e fracos.

Typically, Pentesting may follow one of three possible approaches:

Black-box - In this approach, the pentester has no prior knowledge of the system or network they will be testing. This simulates an attack carried out by a hacker, providing a realistic assessment of the organisation's defence capability.

White-box - In this approach, the pentester has full access to information about the system, including architecture, source codes and network diagrams. This approach allows for a more in-depth and accurate analysis of existing vulnerabilities, providing a complete view of the system's security.

Grey-box - In this approach, the pentester has partial knowledge of the system or network being tested, but will not have access to credentials or configuration details. This approach helps the pentester focus their time and effort on getting a more accurate view of what a real attacker might come across.

The methodology to be followed in a Pentesting service involves the following steps:

Reconnaissance: In this stage, the pentester gathers information about the system or network that will be tested, identifying its strengths and weaknesses.

Análise de vulnerabilidades: O *pentester* utiliza várias técnicas e ferramentas para identificar vulnerabilidades no sistema, como análise de código, testes de segurança da rede e análise de configurações.

Exploração: Nesta etapa, o *pentester* tenta explorar as vulnerabilidades encontradas para obter acesso não autorizado ao sistema. Essas ações são realizadas de forma controlada para evitar danos aos sistemas.

Pós-exploração: Após obter acesso ao sistema, o *pentester* avalia o potencial impacto de um ataque bem-sucedido e identifica as medidas necessárias para mitigar os riscos.

Elaboração de relatório: Por fim, o *pentester* documenta os resultados, incluindo as vulnerabilidades encontradas, as técnicas utilizadas e as recomendações para reforçar a segurança do sistema.

Em suma, o *Pentesting* é uma prática essencial para garantir a segurança de sistemas e redes, visando identificar vulnerabilidades e fornecer recomendações para sua correção. Dessa forma, as organizações podem reforçar as suas defesas e proteger os seus dados contra ameaças cibernéticas. Investir em serviços de *Pentesting* é um passo fundamental para alcançar uma segurança efetiva e promover a confiança num mundo digital em constante evolução.

Vulnerability assessment: The pentester uses various techniques and tools to identify vulnerabilities within the system, such as code analysis, network security testing and configuration analysis.

Exploitation: In this stage, the pentester tries to exploit the vulnerabilities found in order to gain unauthorised access to the system. These actions are carried out in a controlled manner to avoid damage to the systems.

Scanning: After gaining access to the system, the pentester assesses the potential impact of a successful attack and identifies the necessary measures to mitigate the risks.

Reporting: Finally, the pentester documents the results, including vulnerabilities found, techniques used and shares recommendations to strengthen the system security.

In short, Pentesting is an essential practice to ensure the security of systems and networks, aiming at identifying vulnerabilities and providing recommendations for their correction. In this way, organisations can strengthen their defences and protect their data against cyberthreats. Investing in Pentesting services is a key step towards achieving effective security and promoting trust in an ever evolving digital world.

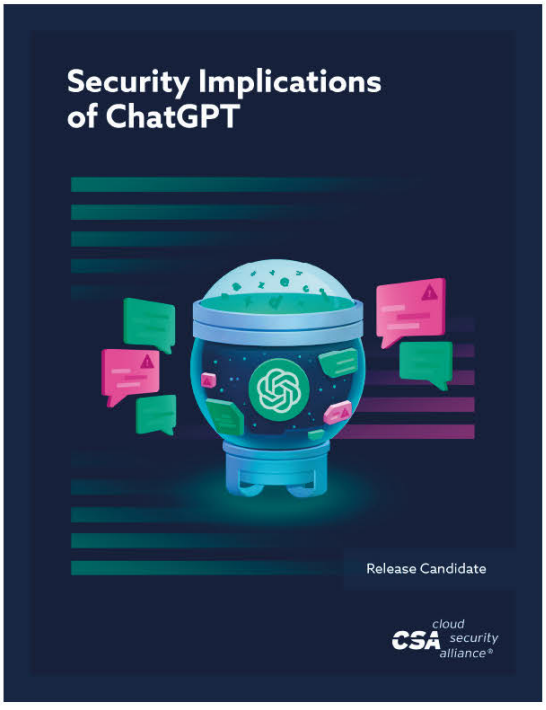
Artificial Intelligence and Cybersecurity Research

No âmbito de vários relatórios relacionados com a segurança e a IA elaborados pela ENISA, este documento pretende identificar as necessidades de investigação sobre a IA e a sua segurança, no âmbito das competências da agência europeia para a cibersegurança. O trabalho iniciou-se em 2021 e foi posteriormente validado nos anos seguintes, tendo identificado cinco necessidades fundamentais de investigação para serem "partilhadas e discutidas com as partes interessadas como propostas para futuras iniciativas políticas e de financiamento a nível da UE e dos Estados-Membros".

Within the framework of several reports related to security and AI produced by ENISA, this document aims to identify the research needs on AI and its security, within the competences of the European Union Agency for Cybersecurity. The work started in 2021 and was subsequently validated in the following years. It identified five key research needs to be 'shared and discussed with stakeholders as proposals for future policy and funding initiatives at the level of the EU and Member States'.

Security Implications of ChatGPT

Perante a generalização de uso do ChatGPT, a Cloud Security Alliance editou este "white paper" para lidar com quatro dimensões de preocupação deste "Large Language Model" (modelo de linguagem de grande dimensão ou LLM): "como pode beneficiar a cibersegurança, como pode beneficiar os atacantes maliciosos, como pode o ChatGPT ser atacado diretamente e orientações para uma utilização responsável". O documento identifica diferentes casos de utilização específicos para melhorar a cibersegurança nas organizações.



Faced with the widespread use of ChatGPT, the Cloud Security Alliance edited this white paper to address four dimensions of concern of this Large Language Model (LLM): 'How it can benefit cybersecurity, how it can benefit malicious attackers, how ChatGPT might be attacked directly, and guidelines for responsible usage'. The document identifies different specific use cases for improving cybersecurity within organisations

AI security concerns in a nutshell

Este pequeno guia introduz os ataques mais relevantes a sistemas de aprendizagem automática ("machine learning" ou ML) e suas potenciais defesas. Alerta-se que "o possível impacto dos ataques aumenta à medida que a ML é cada vez mais utilizada em aplicações críticas". Além disso, "a falta de compreensão do seu processo de tomada de decisões constitui uma ameaça. Os modelos podem estar a aprender correlações espúrias a partir

This short guide introduces the most relevant attacks on machine learning (ML) systems and potential complementary defences. It warns that 'the possible impact of attacks increase as machine learning is used more and more in critical applications.' In addition, 'a lack of comprehension of their decision-making process poses a threat. The models could be learning spurious correlations from faulty or insufficient training data.

de dados de treino defeituosos ou insuficientes. Por conseguinte, é útil compreender o seu processo de decisão antes de os implementar em casos de utilização reais". Apresentam-se ainda três categorias de ataques (de evasão, de extração de informação e ataques de backdoor) e um conjunto de defesas.

Therefore, it is helpful to understand their decision process before deploying them to real-world use cases.' Three categories of attacks (evasion, information extraction and backdoor attacks) and a set of defences are also presented.

ChatGPT - the impact of Large Language Models on Law Enforcement

Mais um documento a analisar o impacto dos LLMs, nomeadamente do generalizado ChatGPT. Apesar dos seus potenciais benefícios, eles são um risco para os cidadãos "e para o respeito dos direitos fundamentais, uma vez que os criminosos e os maus atores podem querer explorar os LLM para fins nefastos".

Another document looking at the impact of LLMs, namely the now mainstream ChatGPT. Despite their potential benefits, they can also be a risk to citizens 'and for the respect of fundamental rights as criminals and bad actors may wish to exploit LLMs for their own nefarious purposes.'

A partir de wokshops organizados pelo Laboratório de Inovação da Europol com diferentes especialistas, foi possível elaborar este relatório de sensibilização para o impacto que os LLM podem ter no trabalho da comunidade policial. Dado que este tipo de tecnologia progride rapidamente "e novos modelos se tornam disponíveis, tornar-se-á cada vez mais importante para a aplicação da lei manter-se na vanguarda para antecipar e evitar abusos, bem como para garantir o aproveitamento de potenciais benefícios".

From workshops organised at the Europol Innovation Lab with different experts, it was possible to prepare this report to raise awareness to the impact that LLMs can have on the police community's work. As this type of technology progresses 'and new models become available, it will become increasingly important for law enforcement to stay at the forefront of these developments to anticipate and prevent abuse, as well as to ensure potential benefits can be taken advantage of.'

Diretora | Director
Inês Esteves

Edição | Editor
Pedro Fonseca

Design Gráfico | Graphic Design
Sara Dias
Maria Cristóvão

Tradução | Translation
Sara Pereira

Fotografia (capa e índice) | Photography (cover & index)
Brecht Corbeel/Unsplash
Tom Chrostek/Unsplash

Publicação trimestral | Quarterly publication
Junho 2023 | June 2023