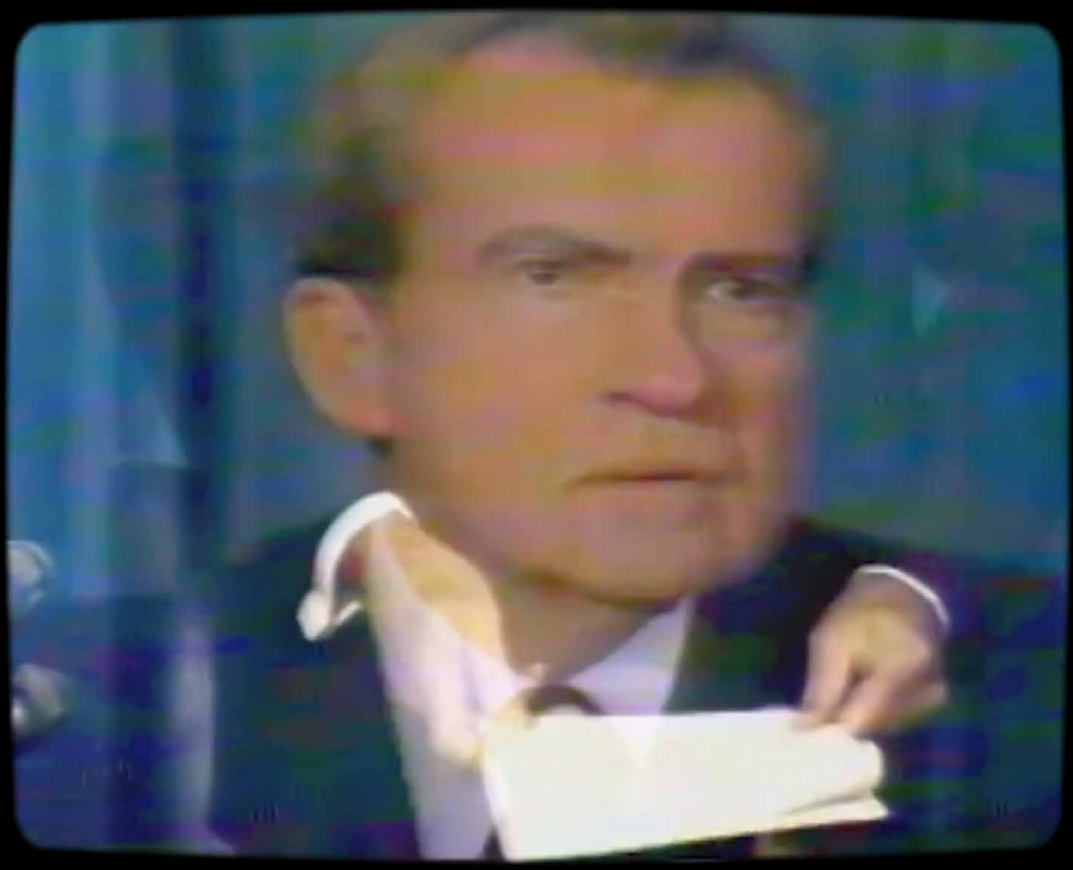
bilingual edition

ptsoc{news}



Deepfakes: uma nova ciberameaça às organizações

3 perguntas a Lino Santos

Privacidade e o DNS (Parte I: o problema) por João Damas

03

Deepfakes: a new cyber threat to organizations

3 questions to Lino Santos

Privacy and the DNS (Part I: the problem) by João Damas

.pt

Deepfakes: uma nova ciber-ameaça às organizações

Os deepfakes são o resultado de avanços tecnológicos na manipulação de imagens e sons, que permitem criar vídeos ou vozes falsas para se assemelharem a conteúdos legítimos, com um realismo que torna difícil descobrir que não são autênticos.

A pandemia, ao obrigar as pessoas a comunicarem à distância, exacerbou o número de filmagens e gravações sonoras disponíveis de alvos interessantes para os cibercriminosos duplicarem e manipularem.

A tecnologia responsável pelos deepfakes (junção de “deep learning” e “fake”) foi desenvolvida em 2014 por Ian Goodfellow na Universidade de Montréal e é denominada de "generative adversarial network" (GAN).

De forma facilitada, as GANs são modelos de aprendizagem por máquina ("machine learning" ou ML) que competem entre si e incrementam a qualidade dos resultados. Assim, enquanto um modelo cria um vídeo, o outro tenta descobrir se é falso, num processo repetido de realimentação até ser extremamente difícil detetar as imagens falsas.

Deepfakes: a new cyber threat to organizations

Deepfakes are the result of technological advances in image and sound manipulation, which allow fake videos or voices to be created to resemble legitimate content, with a realism that makes it difficult to discover that they are not authentic.

The COVID-19 pandemic forced people to communicate at a distance, which increase the number of footage and sound recordings available of interesting targets for cybercriminals to duplicate and manipulate.

The technology responsible for deepfakes (a combination of ‘deep learning’ and ‘fake’) was developed in 2014 by Ian Goodfellow at the University of Montreal and is called ‘generative adversarial network’ (GAN).

To put it simply, GANs are machine learning (ML) models that compete with each other and increase the quality of the results. So, while one model creates a video, the other tries to figure out if it’s fake, in a repeated feedback process until it’s extremely difficult to detect fake images.

‘An interesting side effect of this lear-

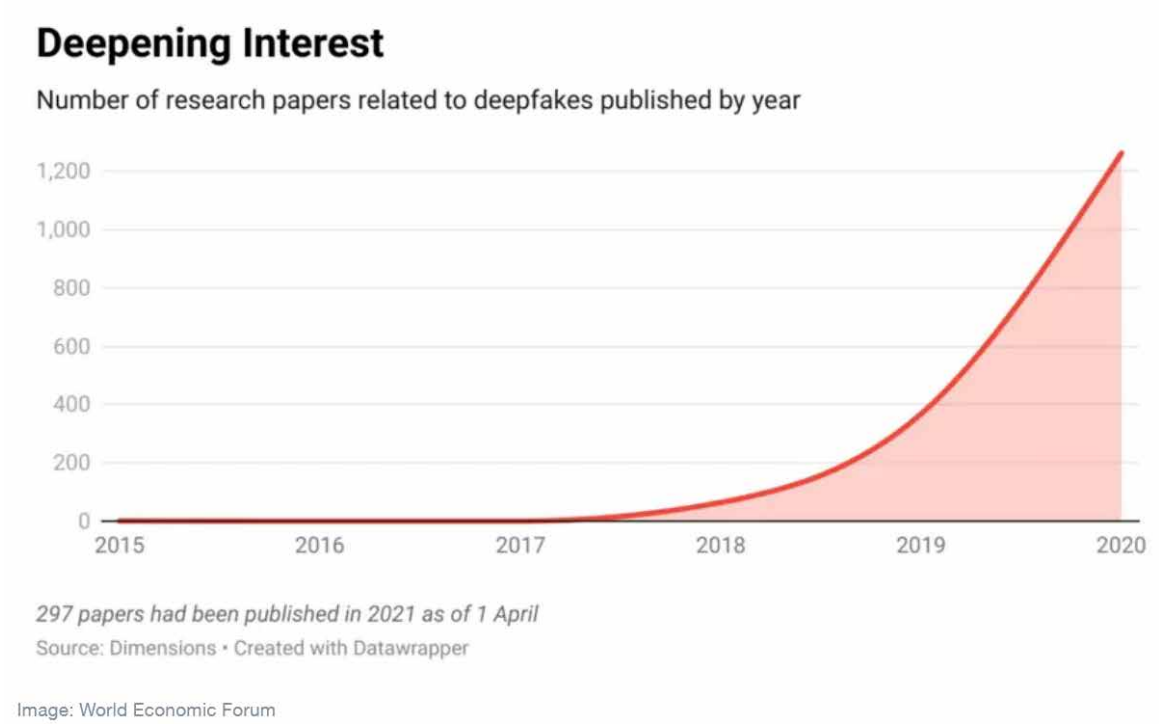


Image: World Economic Forum

"Um efeito colateral interessante desse processo de aprendizagem é que os dois agentes melhoram juntos. Assim, ao criar um produtor de grandes deepfakes, está-se também a criar um ótimo detetor de deepfakes e vice-versa", escreveu a Scientific Foresight Unit (STOA) do Parlamento Europeu (PE).

O interesse pela tecnologia evoluiu muito depressa e vídeos deepfake começaram a surgir online nos finais de 2017, quando o utilizador Deepfakes lançou na rede Reddit tecnologia GAN para a geração de imagens pornográficas, um software "open source" para colocar o rosto de uma pessoa no corpo de outra. Em fevereiro do ano seguinte, o Reddit proibiu-as.

ning process is that the two agents improve together. So, by creating a great deepfake producer, you also create a great deepfake detector, and vice-versa,' wrote the Scientific Foresight Unit (STOA) of the European Parliament (EP).

Interest in the technology evolved very quickly and deepfake videos began to appear online in late 2017, when user Deepfakes released on Reddit GAN technology for generating pornographic images, an open source software to put a person's face of another person's body. In February of the following year, Reddit banned them.

There were more than 7 900 deepfakes

No início de 2019, contaram-se mais de 7.900 deepfakes e ultrapassavam os 14.600 em maio de 2020, com mais de 95% de teor pornográfico. Em dezembro passado, a Sensity estimou existirem mais de 85 mil.

Os responsáveis pela produção destes conteúdos vão desde programadores interessados a organizações atentas às capacidades técnicas (como canais de televisão), de atores políticos com interesses pouco claros a cibercriminosos.

A tecnologia dos deepfakes tem um enorme potencial para o engano, para facilitar a disseminação da desinformação, denegrir a reputação pessoal ou proporcionar ataques de engenharia social, com roubo de identidade para fraudes. Mas há mais.

"Imagine-se um cenário em que um vídeo de Elon Musk com recomendações de 'insider trading' se torna viral - só que não é o verdadeiro Elon Musk. Ou um político anuncia uma nova política num vídeo mas, mais uma vez, não é real. Já vimos esses vídeos falsos usados em campanhas políticas; é apenas uma questão de tempo antes de os criminosos aplicarem a mesma técnica a empresas e indivíduos ricos. Pode

in early 2019, and they exceeded 14 600 in May 2020, with more than 95 % being of pornographic content. Last December, Sensity estimated there were more than 85 000.

The people responsible for producing this content range from interested programmers to organisations with an eye on technical skills (such as television channels), from political actors with unclear interests to cybercriminals.

Deepfakes technology has enormous potential for deception, for facilitating the spread of disinformation, denigrating personal reputation or providing social engineering attacks, with identity theft for fraudulent purposes. But there's more.

'Imagine a scenario in which a video of Elon Musk giving insider trading tips goes viral – only it's not the real Elon Musk. Or a politician announces a new policy in a video clip, but once again, it's not real. We've already seen these fake videos used in political campaigns before; it's only a matter of time before criminals apply the same technique to businesses and wealthy private individuals. It could be as simple as a faked voicemail from a senior manager instructing staff to make

ser tão simples quanto uma mensagem de voz falsa de um gestor de topo a instruir a equipa para fazer um pagamento fraudulento ou mover fundos para uma conta configurada por um hacker", ilustra Darren Thomson, responsável de estratégia de cibersegurança da CyberCube.

As organizações estão assim em perigo latente de ataques com perdas financeiras embora, de forma concreta, se conheçam poucos casos confirmados. Um ocorreu em 2019 com um CEO de uma subsidiária no Reino Unido de uma empresa da Alemanha. Ao ouvir o acento e os padrões de voz do CEO alemão, o executivo inglês efetuou uma transferência bancária de 240 mil euros para a conta de um "fornecedor húngaro". O sucesso da operação levou os cibercriminosos para uma segunda tentativa mas, ao contrário da primeira vez em que o CEO inglês não suspeitou de nada, desta vez recusou a operação.

Outro caso, em 2020, envolveu também uma falsificação vocal. Um gestor bancário em Hong Kong reconheceu a voz de um cliente dos Emirados Árabes Unidos (EAU) a pedir-lhe para, no âmbito da aquisição de uma empresa,

a fraudulent payment or move funds to an account set up by a hacker,' illustrates Darren Thomson, head of cybersecurity strategy at CyberCube.

Thus, organizations are in latent danger of attacks with financial losses although, concretely, few confirmed cases are known. One occurred in 2019 with a CEO of a UK subsidiary of a German company. Upon hearing the German CEO's accent and voice patterns, the British executive made a bank transfer of €240 000 to the account of a 'Hungarian supplier.' The success of the operation prompted cybercriminals to make a second attempt but, unlike the first time when the British CEO did not suspect anything, this time he refused the transaction.

Another case, in 2020, also involved voice forgery. A bank manager in Hong Kong recognised the voice of a customer from the United Arab Emirates (UAE) asking him, in connection with the acquisition of a company, to transfer €31 million to a lawyer who was coordinating the transaction. The request was validated by (fake) emails from the customer and the lawyer to the bank manager.

The fraudulent scheme came to light

transferir 31 milhões de euros para um advogado, que coordenava a transação. A validação do pedido foi efetuada por emails (falsos) do cliente e do advogado para o gestor do banco.

O esquema fraudulento foi conhecido quando a revista Forbes acedeu a documentos do tribunal com o pedido dos EAU aos Estados Unidos para descobrirem o rasto do dinheiro, transferido para contas locais do Centennial Bank e daí para outras em diferentes países.

Alertas dos poderes limitados

A escassez de casos não impede as agências governamentais de emitirem avisos para o potencial destas tecnologias, quando, naturalmente, estas tecnologias captaram o interesse de criminosos. "Os fóruns da dark Web em inglês e russo foram identificados como as principais fontes para os utilizadores anunciarem, discutirem, partilharem e comprarem produtos, serviços e tópicos relacionados com os deepfakes", escreveu-se em abril passado.

Os tópicos mais comuns "incluíam serviços (edição de vídeos e imagens), métodos e lições de como fazer, solicitações

when Forbes magazine accessed court documents with the UAE's request to the United States to uncover the money trail, transferred to local Centennial Bank accounts and from there to others in different countries.

Warnings from limited powers

The paucity of cases does not stop government agencies from issuing warnings about the potential of these technologies when, of course, these technologies have captured the interest of criminals. 'English- and Russian-language dark web forums were identified as the main sources for users to advertise, discuss, share, and purchase deepfake-related products, services, and topics,' was written last April.

Common topics 'included (video and image editing) services, how-to methods and lessons, requests for best practices, sharing of free software downloads and photo generators, general interests in deepfakes and announcements about technological advances for deepfakes.'

Their use 'will probably become a serious challenge for the digital environment,

de melhores práticas, partilha de downloads de software gratuito e geradores de fotos, interesses gerais em deepfakes e anúncios sobre avanços nas tecnologias para deepfakes".

O seu uso "será provavelmente um sério desafio para o ambiente digital", constatou o relatório "Serious and Organised Crime Treath Assessment" da Europol, em abril passado.

"As autoridades policiais têm poderes limitados para conter a manipulação da informação, que pode assumir a forma de tentativas de distorcer o discurso político, manipular eleições, erodir os princípios democráticos, semear a desconfiança nas instituições, intensificar as divisões sociais, fomentar a insegurança e espalhar a discriminação e a xenofobia".

Um mês antes, também o FBI alertou para o perigo de nações atingirem os EUA com campanhas de desinformação com deepfakes. "Atores mal-intencionados quase certamente usarão conteúdo sintético para ciberoperações e influência estrangeira nos próximos 12-18 meses", enquanto "atores estrangeiros estão atualmente a usar conteúdo

Europol's ' Serious and Organised Crime Treath Assessment' report found last April.

'Law enforcement authorities have limited powers to curb information manipulation, which can take the form of attempts to distort political discourse, manipulate elections, erode democratic principles, sow distrust in institutions, intensify social divisions, foster insecurity and spread discrimination and xenophobia.'

A month earlier, the FBI also warned of the danger of nations targeting the US with disinformation campaigns with deepfakes. 'Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations in the next 12-18 months,' while 'foreign actors are currently using synthetic content in their influence campaigns, and the FBI anticipates it will be increasingly used by foreign and criminal cyber actors for spearphishing and social engineering in an evolution of cyber operational tradecraft.'

This synthetic content is defined 'as the broad spectrum of generated or manipulated digital content, which includes

sintético nas suas campanhas de influência, e o FBI prevê que será cada vez mais usado por ciberatores estrangeiros e criminosos para 'spearphishing' e engenharia social numa evolução do artesanato ciberoperacional".

Este conteúdo sintético é definido "como o amplo espectro de conteúdo digital gerado ou manipulado, que inclui imagens, vídeo, áudio e texto", usando técnicas baseadas em tecnologias de inteligência artificial (IA) ou ML, "conhecidas popularmente como deepfakes ou GANs".

O STOA apelou, no âmbito da proposta de lei europeia para a IA, que os deepfakes ou "quaisquer outros vídeos sintéimages, video, audio, and text,' using techniques based on artificial intelligence (AI) or ML technologies, 'known popularly as deepfakes or GANs.'

STOA called, as part of the proposed European AI law, for deepfakes or 'any other realistically made synthetic videos to be labelled as "non-original" by the creator, with strict limits on their use for electoral purposes.' The EP also wants research in this area 'to ensure that technologies to combat such phenomena keep pace with the malicious use of AI.'

'Deepfakes are part of the larger problem of misinformation that undermines trust in institutions and in visual experience - we can no longer trust what

Overview of different categories of risks associated with deepfakes

Psychological harm	Financial harm	Societal harm
• (S)extortion • Defamation • Intimidation • Bullying • Undermining trust	• Extortion • Identity theft • Fraud (e.g. insurance/payment) • Stock-price manipulation • Brand damage • Reputational damage	• News media manipulation • Damage to economic stability • Damage to the justice system • Damage to the scientific system • Erosion of trust • Damage to democracy • Manipulation of elections • Damage to international relations • Damage to national security

ticos feitos de forma realista sejam rotulados como 'não originais' pelo criador, com limites restritos sobre o seu uso para fins eleitorais". O PE quer também investigação nesta área "para garantir que as tecnologias para combater esses fenómenos acompanham o uso malicioso da IA".

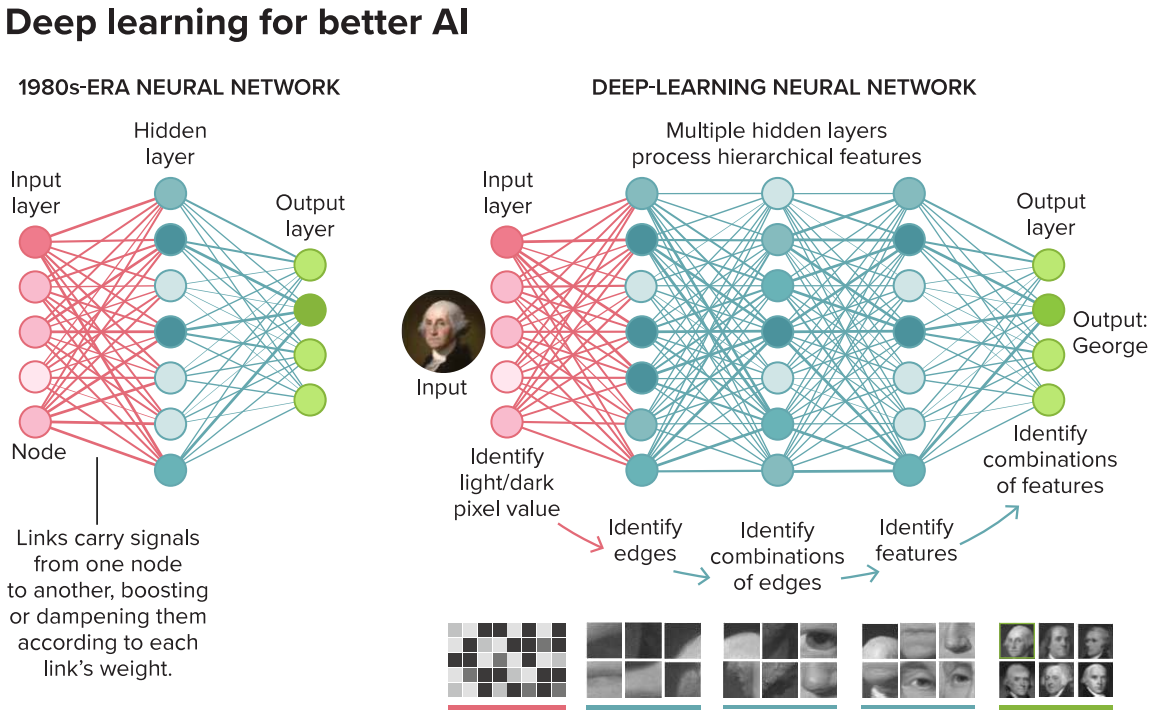
"Os deepfakes são parte do problema maior da desinformação que mina a confiança nas instituições e na experiência visual - já não podemos confiar no que vemos e ouvimos online", declara Deborah Johnson, professora de ética na University of Virginia. E "a rotulagem é provavelmente o mais simples e mais importante contra-ataque aos deepfakes - se os telespectadores estão cientes de que o que estão a ver foi fabricado, é menos provável serem enganados".

Os deepfakes são a evolução técnica natural do velho problema da manipulação de sons e imagens, que começou muito antes do Photoshop surgir em 1990. Nesse ano, "a Newsweek advertiu que governos autoritários como a China poderiam desresponsabilizar-se de atrocidades futuras como Tiananmen porque, 'com a fotografia eletrónica, we see and hear online', states Deborah Johnson, professor of ethics at the University of Virginia. And 'labelling is probably the simplest and most important counter to deepfakes - if viewers are aware that what they are viewing has been fabricated, they are less likely to be deceived.'

Deepfakes are the natural technical evolution of the old problem of manipulating sounds and images, which began long before Photoshop appeared in 1990. That year, 'Newsweek warned that authoritarian governments like China could get away with future atrocities like Tiananmen because, "with electronic photography they could deny the veracity of the newly malleable image," recalls in 'Trust Your Eyes? Deepfakes Policy Brief.'

Technology for good, evolution for... where?

Ian Goodfellow's creation looks like an open Pandora's box for evil but the technology can be used for beneficial purposes, such as to hide faces and people who need to speak anonymously when reporting certain occurrences, as explained about the film 'Welcome to



poderiam negar a veracidade da nova imagem maleável", recorda-se em "Trust Your Eyes? Deepfakes Policy Brief".

Tecnologia para o bem, evolução para... onde?

A criação de Ian Goodfellow parece uma caixa de Pandora aberta para o mal mas a tecnologia pode ser usada para fins benéficos, como para esconder rostos e pessoas que precisam de falar anonimamente quando denunciam certas ocorrências, como se explica sobre o filme "Welcome to Chechnya", em "How deepfakes could actually do some good".

Chechnya' in 'How deepfakes could actually do some good'.

It's 'equally remarkable [in] applications that could make quick work out of once-painstaking tasks: filling in gaps and scratches in damaged images or video; turning satellite photos into maps; creating realistic streetscape videos to train autonomous vehicles; giving a natural-sounding voice to those who have lost their own; turning Hollywood actors into their older or younger selves; and much more.' Knowable Magazine reports.

The debate about their use can also

É "igualmente notável em aplicações que podem realizar um trabalho rápido em tarefas antes penosas: preencher lacunas e riscos em imagens ou vídeos danificados; transformar fotos de satélite em mapas; criar vídeos realistas de paisagens urbanas para treinar veículos autónomos; dar uma voz natural àqueles que perderam a sua; transformar atores de Hollywood nos seus eus mais velhos ou mais jovens; e muito mais", refere a Knowable Magazine.

O debate sobre a sua utilização pode igualmente contribuir para uma maior literacia mediática e tecnológica no campo da desinformação, por exemplo. Esse potencial está bem recriado no projeto "In Event of a Moon Disaster", do Center for Advanced Virtuality do MIT, que mostra um deepfake "completo" - por manipular som e vídeo - do antigo presidente Nixon a dirigir-se ao país com um discurso (realmente escrito para essa contingência) como se a Apollo 11 não tivesse regressado à Terra em 1969.

Quando o custo for menor, assim como o tempo de desenvolvimento dos deepfakes, novas aplicações tendem a surgir - várias delas com impacto no espaço

contribute to greater media and technology literacy in the field of disinformation, for example. That potential is well recreated in the 'In Event of a Moon Disaster' project from MIT's Center for Advanced Virtuality, which shows a 'complete' deepfake - by manipulating sound and video - of former President Nixon addressing the nation with a speech (actually written for that contingency) as if Apollo 11 had not returned to Earth in 1969.

When the cost becomes lower, and the development time of deepfakes is reduced, new applications tend to emerge - several of them impacting the media space, for example. Last year, South Korean television channel MBN replaced presenter Kim Joo-Ha with its own deepfake.

There were those who worried about the likely loss of the anchor's job but others recognise the potential for the creation of 'fake news', all the more so as the company who created the character wants more customers from the US and China.

Deepfakes are also appealing (and affordable) for advertising, as they can

mediático, por exemplo. No ano passado, o canal de televisão sul-coreano MBN substituiu a apresentadora Kim Joo-Ha por um seu deepfake.

Houve os que se preocuparam com a provável perda de trabalho da pivô mas outros reconhecem o potencial para a criação de "fake news", tanto mais que a empresa criadora da personagem quer mais clientes dos EUA e China.

Os deepfakes são também apelativos (e acessíveis) para a publicidade, por poderem integrar pessoas já filmadas noutros anúncios em novos filmes publicitários. "Este é o futuro da criação de conteúdos", declarou Victor Riparbelli, CEO da Synthesia, que usa estas tecnologias na formação.

Mas há questões legais por resolver, aponta Lilian Edwards, professora na Newcastle Law School, como saber a quem pertencem os direitos dos deepfakes de falecidos. "Por exemplo, se uma pessoa morta é usada, como [o ator] Steve McQueen ou [o rapper] Tupac, há um debate contínuo sobre se a família deve possuir os direitos [e ganhar com isso]", diz.

integrate people already filmed in other ads into new commercials. 'This is the future of content creation,' said Victor Riparbelli, CEO of Synthesia, which uses these technologies in training.

But there are legal issues to be resolved, points out Lilian Edwards, a professor at Newcastle Law School, such as who owns the rights to deepfakes of deceased people. 'For example, if a dead person is used, such as [actor] Steve McQueen or [rapper] Tupac, there is an ongoing debate about whether the family should own the rights [and gain from it],' she says.

In fact, due to AI's potential, it will be normal to, once again, hear singers from the past 'come back' to sing their hits of yesteryear or even new ones. 'Deepfakes are cute tricks - but they could change pop for ever' - like putting Sinatra singing Britney Spears hits and both in the middle of a court 'battle'.

That came close to happening in a documentary featuring an Anthony Bourdain vocal deepfake, despite his widow Ottavia Busia publicly disagreeing with the use of AI to create new phrases and having Bourdain express

Na realidade, devido ao potencial da IA, será normal voltar a ouvir cantores do passado "regressarem" para entoarem os seus sucessos de outrora ou até novos. "Os deepfakes são uns truques giros - mas podem mudar a pop para sempre" - como colocar Sinatra a cantar sucessos de Britney Spears e ambos no meio de uma "batalha" judicial.

Isso esteve quase a suceder num documentário com um deepfake vocal de Anthony Bourdain, apesar da sua viúva Ottavia Busia discordar publicamente do uso da IA para criar novas frases e ter Bourdain a expressar algo que nunca tinha dito.

O realizador Morgan Neville clarificou que algumas declarações não tinham sido realmente ditas mas a opção tecnológica era uma "técnica moderna de contar histórias".

"Os deepfakes estão para ficar"

O que estes exemplos demonstram é também a capacidade de se poder usar os deepfakes para fins pouco ortodoxos. O criador de um famoso vídeo com um falso Tom Cruise tentou acalmar os ânimos, salientando a dificuldade na sua

something he'd never said.

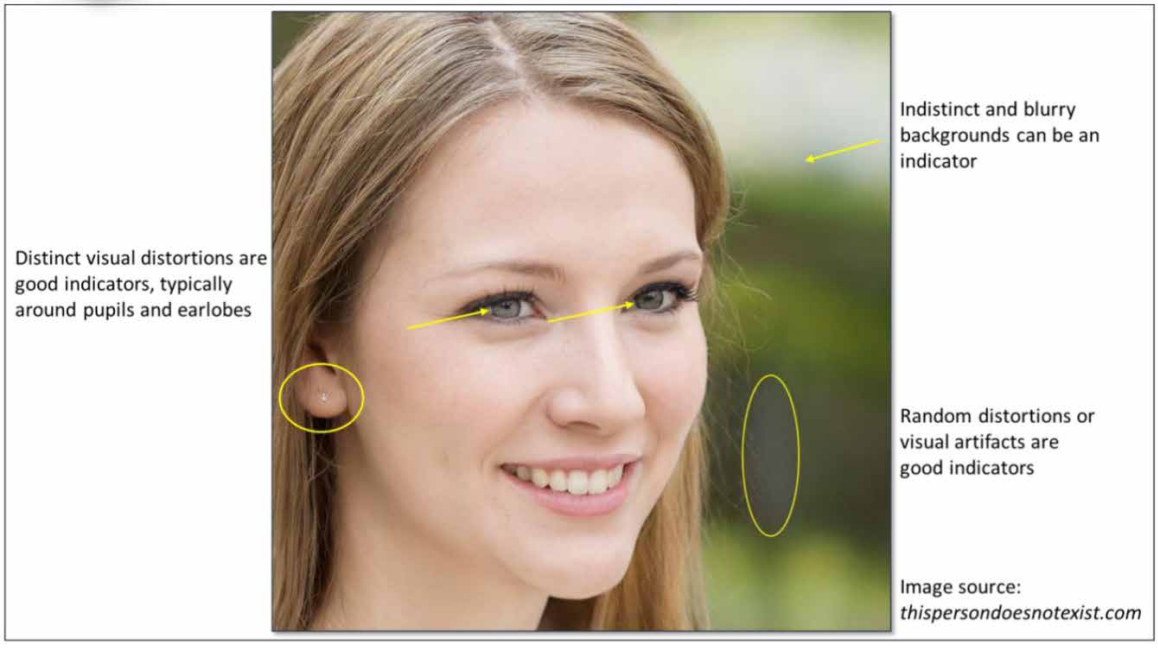
Director Morgan Neville clarified that some statements had not actually been said but the technology option was a 'modern storytelling technique.'

'Deepfakes are here to stay'

These examples also show the ability to use deepfakes for unorthodox purposes. The creator of a famous video with a fake Tom Cruise tried to calm the spirits, stressing the difficulty in its creation and the time and effort required.

According to Chris Ume, special effects expert, 'You can't do it by just pressing a button,' it's a job that takes weeks, using AI and video editing tools. But he concedes that we're in a moment, like Photoshop was 20 years ago, and that 'deepfakes are here to stay.'

This is also one of the conclusions of the 'Tackling Deepfakes in European policy' study (STOA), which notes that 'deepfakes accelerate the erosion of trust', that they are dual technologies and that it will be insufficient to regulate the technological dimension when 'citizens need additional support to protect their



criação e o tempo e esforço necessários.

Segundo Chris Ume, especialista de efeitos especiais, "isto não se faz apenas carregando num botão", é um trabalho que demora semanas, usando ferramentas de IA e de edição de vídeo. Mas concede que se está num momento como o do Photoshop há 20 anos e que "os deepfakes estão para ficar".

Essa é também uma das conclusões de "Tackling deepfakes in European policy" (STOA), em que se nota como "os deepfakes aceleram a erosão da confiança", são tecnologias duais e que será insuficiente regular a dimensão tecnológica

rights.' But 'visual manipulation is here to stay.'

The technological evolution should accelerate this process and there are already several examples that show exactly this. From a more playful perspective, it is possible to foresee how synthetic images are closer to reality. A new title in the 'Grand Theft Auto' video game series reveals how researchers 'may have found a shortcut by applying ML techniques to rendered footage from a [video game] console that takes it from beautiful to photorealistic.'

The importance of this technical evolution and its disruptive potential is recognised

quando "os cidadãos precisam de apoio adicional para proteger os seus direitos". Mas "a manipulação visual veio para ficar".

A evolução tecnológica deverá acelerar este processo e vários exemplos já o conseguem demonstrar. A partir de um campo mais lúdico, é possível antever como as imagens sintéticas são mais próximas da realidade. Um novo título da série de videojogos "Grand Theft Auto" revela como investigadores "podem ter encontrado um atalho aplicando técnicas de ML a imagens renderizadas numa consola [de videojogos] que as transforma de belas em foto-realistas".

A importância desta evolução técnica e do seu potencial disruptivo é reconhecida na manipulação de imagens de satélite, a partir do alerta de militares e geógrafos que antecipam impactos danosos para as forças de proteção militar e civil. Eles estão preocupados com a possibilidade da criação e disseminação de imagens de satélite falsas, geradas por IA. "Essas imagens podem enganar de várias maneiras. Podem ser usadas para criar boatos sobre incêndios florestais ou inundações, ou para desacreditar

in the manipulation of satellite imagery, from the warning of military and geographers anticipating damaging impacts for military and civilian protection forces.

They're concerned about the possibility of the creation and dissemination of fake, AI-generated satellite imagery. 'These images can deceive in many ways. They can be used to create rumours about forest fires or floods, or to discredit news stories based on real satellite imagery.' And 'geography with deepfakes can even be a national security problem, when geopolitical adversaries use fake satellite imagery to mislead enemies.' Todd Myers, of the National Geospatial-Intelligence Agency illustrates how 'from a tactical perspective or mission planning, you train your forces to go a certain route, toward a bridge, but it's not there.'

Political and technical solutions

Faced with deepfake threats, some solutions have been put forward.

On the political side, in addition to the aforementioned EP proposals, the US states of Texas, Virginia and California

notícias baseadas em imagens reais de satélite". E "a geografia com deepfakes pode até ser um problema de segurança nacional, quando adversários geopolíticos usam imagens falsas de satélite para enganar os inimigos". Todd Myers, da National Geospatial-Intelligence Agency, ilustra como "de uma perspetiva tática ou planeamento de missão, se treinam as forças para seguir uma determinada rota, em direção a uma ponte, mas esta não está lá".

Soluções políticas e técnicas

Perante as ameaças dos deepfakes, aventaram-se algumas soluções.

Do lado político, além das referidas propostas do PE, os estados norte-americanos do Texas, Virgínia e Califórnia criminalizaram os deepfakes pornográficos, enquanto este último aprovou também legislação a proibir a criação e divulgação de deepfakes de políticos a dois meses de uma eleição.

No final de 2020, o Senado dos EUA aprovou o Identifying Outputs of Generative Adversarial Networks Act (IOGAN Act), para que a National Science Foundation e o National Institute of Standards

have criminalised pornographic deepfakes, while the latter has also passed legislation banning the creation and dissemination of deepfakes of politicians two months before an election.

In late 2020, the U.S. Senate passed the Identifying Outputs of Generative Adversarial Networks Act (IOGAN Act), for the National Science Foundation and the National Institute of Standards and Technology to fund research on GANs.

From a more technological side, there has been a 'race for weapons of detection', seeking to develop technologies for their identification.

Some examples include the University of Buffalo's proposal that ensures a 94 % efficiency in detecting these false images by analysing tiny reflections in the eyes. NoiseScope, presented at the end of last year, seeks to discriminate whether an image is real or GAN-generated. Researchers guarantee a 99.6 % efficiency. Digimarc has announced a 'digital watermarking' system to prevent the proliferation of deepfake videos, embedded in both images and audio. This unique identification can be recognised by social media.

and Technology financiem investigação em GANs.

Na vertente mais tecnológica, assistiu-se a uma "corrida às armas de detecção", procurando-se desenvolver tecnologias para a sua identificação.

Alguns exemplos passam pela proposta da University of Buffalo que assegura uma eficácia de 94% na detecção destas imagens falsas pela análise aos pequenos reflexos nos olhos. O NoiseScope, apresentado no final do ano passado, procura discriminar se uma imagem é real ou gerada numa GAN. Os investigadores asseguram uma eficácia de 99,6%. A Digimarc anunciou um sistema de "marca de água digital" para evitar a proliferação de vídeos deepfake, inserida tanto nas imagens como no áudio. Esta identificação única pode ser reconhecida pelas redes sociais. A Content Authenticity Initiative - que junta empresas de media e tecnológicas - pretende também validar a autenticidade dos conteúdos digitais, a começar pelas fotografias e vídeos.

Contrariando as boas notícias, investigadores da UC San Diego mostraram numa conferência este ano que os

The Content Authenticity Initiative - which brings together media and technology companies - also aims to validate the authenticity of digital content, starting with photos and videos.

Contrary to the good news, at a conference this year, researchers at UC San Diego showed that systems designed to detect deepfakes 'can be fooled' and that these 'detectors can be defeated by inserting inputs called adversarial examples into every video frame. The adversarial examples are slightly manipulated inputs which cause AI systems, such as machine learning models, to make a mistake. In addition, the team showed that the attack still works after videos are compressed.' Details of the research were not disclosed.

So, in a world where social media propagates disinformation faster than content moderators can uncover it, the historians of the future are going to have a huge job sorting out the true from the false.

'This is the age of post-history: a new epoch of civilization where the historical record is so full of fabrication and noise that it becomes effectively meaningless,

sistemas concebidos para detetar deepfakes "podem ser enganados" e que esses "detetores podem ser derrotados inserindo entradas chamadas de exemplos adversários em cada imagem de vídeo. Os exemplos adversários são entradas ligeiramente manipuladas que fazem com que os sistemas de IA, como os modelos de ML, cometam um erro. Além disso, a equipa mostrou que o ataque ainda funciona mesmo se os vídeos são compactados". Os detalhes da investigação não foram revelados.

Assim, num mundo em que as redes sociais propagam desinformação mais rapidamente do que os moderadores de conteúdos a conseguem descobrir, os historiadores do futuro vão ter um enorme trabalho a destrinçar o verdadeiro do falso.

"Esta é a era da pós-história: uma nova época da civilização onde o registo histórico é tão cheio de fabricação e barulho que se torna efetivamente sem sentido", escreveu Benj Edwards, historiador de tecnologia. "É como se uma singularidade cultural abrisse um buraco tão profundamente na história que nenhuma verdade pode emergir ilesa do outro lado". ■

wrote technology historian Benj Edwards. 'It's as if a cultural singularity ripped a hole so deeply in history that no truth can emerge unscathed on the other side.' ■

Deepfake videos

Deepfakes real time side by side comparison (Amy Adams & Nick Cage)

You Won't Believe What Obama Says In This Video!

In Event of Moon Disaster - Nixon Deepfake Clips

Deepfake Videos Are Getting Terrifyingly Real

Is This Video a Deepfake?

Home Stallone

Deepfakes: Do Not Believe What You See

Korean TV Network Introduces AI Newsreaders

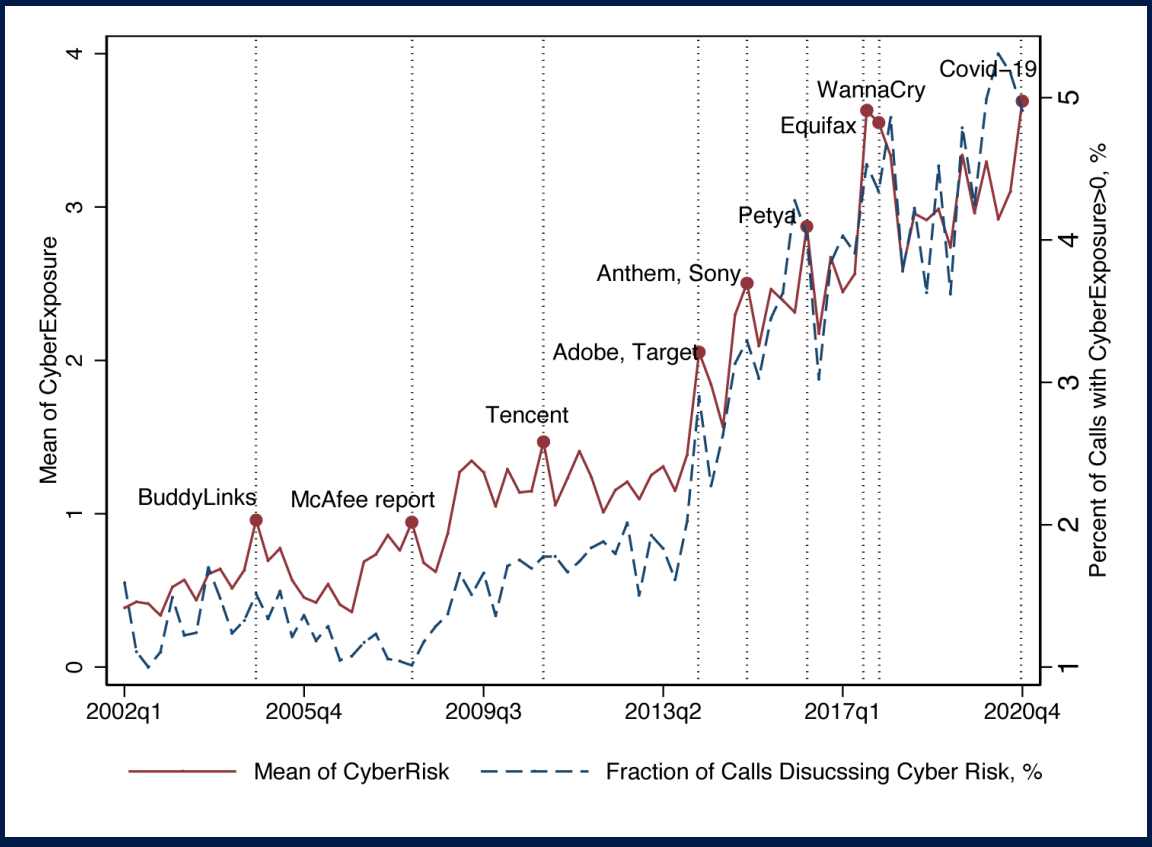
Can we protect society from the deepfake menace?

Anatomia dos ciber-riscos

Em 2020, a pandemia fez aumentar em 50% o número de ciberataques, com a Interpol a registar um crescimento de 569% no malware entre fevereiro e março, e o valor médio dos pagamentos em casos de ransomware a atingir os 180 mil dólares.

Anatomy of cyber risks

In 2020, the COVID-19 pandemic caused the number of cyberattacks to increase by 50 %, with Interpol recording a 569 % growth in malware between February and March, the average payout in ransomware cases reaching $180 000.



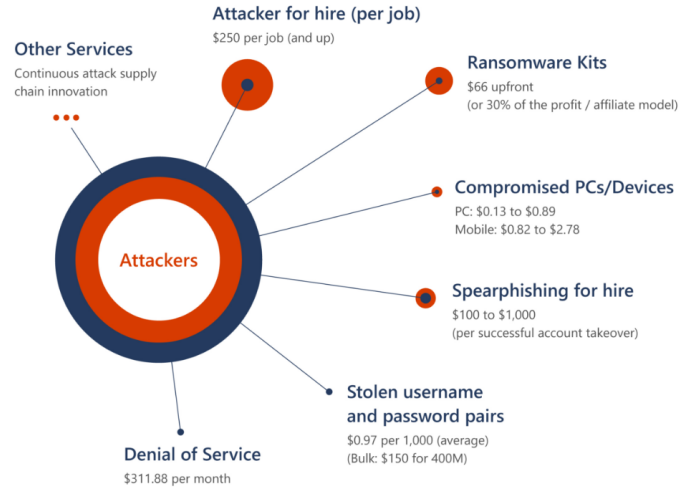
Ransomware escolar

As escolas dos EUA registaram 75 ataques de ransomware até outubro deste ano, afetando 900 estabelecimentos de ensino.

School ransomware

US schools have recorded 75 ransomware attacks until October this year, affecting 900 schools.

Average prices of cybercrime services for sale



Valor médio para ciberataques

O cibercrime posicionou-se como uma ameaça de segurança nacional mas, apesar dos esforços oficiais, é difícil combater esta criminalidade crescentemente motivada pelo lucro financeiro.

Average amount for cyber attacks

Cybercrime has positioned itself as a national security threat but, despite official efforts, it's difficult to combat this increasingly financially motivated crime.

Cibercrime nos TLDs

Apenas 25 domínios de topo ("Top-Level Domains" ou TLD) são responsáveis por albergarem 90% de todos os sites maliciosos. Apesar destes TLDs serem válidos, o interesse ilegal recai principalmente nos que oferecem registos de domínio gratuitos para o "phishing", por exemplo.

Cybercrime in TLDs

Just 25 Top-Level Domains (TLDs) are responsible for hosting 90 % of all malicious websites. Although these TLDs are valid, the illegal interest lies mainly in those that offer free domain registrations for phishing, for example.

3

Lino Santos

Coordenador do Centro Nacional de Cibersegurança Coordinator of the National Cybersecurity Centre

1. Qual é o objetivo do C-Hub: Cybersecurity DIH como Pólo de Inovação Digital?

O C-Hub, Cybersecurity Digital Innovation Hub, pretende assumir-se como uma referência nacional e europeia na área de cibersegurança com uma abordagem tecnologicamente neutra, rápida e segura na procura de soluções de suporte aos processos de transição digital nas organizações, nomeadamente empresas - com uma atenção especial para as micro, pequenas e médias empresas -, mas também na AP. Visa portanto, apoiar as organizações, através da prestação de serviços inovadores, aconselhamento e planeamento, com vista a mitigar os riscos introduzidos por esse processo de transição digital e, assim, tornarem-se mais eficazes e competitivas nas suas áreas de negócio. O C-Hub foi recentemente reconhecido como Pólo de Inovação Digital para integração na Rede Nacional e designado para acesso à Rede Europeia de Pólos de Inovação. Estamos neste momento a preparar o processo de candidatura a esta impor-

1. What is the objective of the C-Hub: Cybersecurity DIH as a Digital Innovation Hub?

The C-Hub, the Cybersecurity Digital Innovation Hub, aims to become a national and European benchmark in the area of cybersecurity with a technologically neutral, fast and safe approach in the search for solutions to support the digital transition processes in organisations, namely companies - with special attention to micro-, small- and medium-sized enterprises - but also in PA. Therefore, it aims to support organisations, through the provision of innovative services, advice and planning, in order to mitigate the risks introduced by this digital transition process, thus becoming more effective and competitive in their business areas. The C-Hub was recently recognised as a Digital Innovation Hub for integration in the National Network and designated for access to the European Digital Innovation Hubs Network. We are currently preparing the application process to this important European network. It is our conviction that being part of a European network through the C-Hub, the whole national cybersecurity ecosystem will be clearly favoured, not only by the recognition, but also the sharing, network effect and possibility of knowledge and innovation transfer with the organisations that will use us, both national and European ones.

tante rede europeia. É nossa convicção que estando integrados numa rede europeia através do C-Hub, todo o ecossistema da ciber-segurança nacional sairá claramente favorecido, não só pelo reconhecimento, mas também pela partilha, efeito de rede e possibilidade de transferência de conhecimento e inovação com as organizações que a nós recorrerão, nacionais e europeias.

2. Quais são as entidades actuais do C-Hub e deverá este ser alargado a outras instituições com que perfil?

O consórcio C-Hub integra atualmente seis entidades (o CNCS como entidade coordenadora, a AMA – Agência para a Modernização Administrativa, o C3P – Centro de Competências em Cibersegurança e Privacidade da Universidade do Porto, o INOV - Instituto de Engenharia de Sistemas e Computadores Inovação, o TICE.PT – Pólo das Tecnologias de Informação, Comunicação e Electrónica e a Pricewaterhouse Coopers/AG), estando sempre aberto à entrada de novas entidades que possam contribuir e acrescentem valor para a missão do consórcio. Desta forma, o C-Hub agrega, num ponto único, um conjunto de entidades públicas e privadas que reunem um saber--como técnico, científico, de investigação e de inovação, que materializam um portefólio de serviços que vão desde a formação avançada até demonstradores de tecnologias emergentes, serviços esses não existentes no mercado.

2. What are the C-Hub's current entities and should it be extended to other institutions with what profile?

The C-Hub consortium currently includes six entities (CNCS as the coordinating entity; AMA - Administrative Modernization Agency; C3P - Competence Centre for Cybersecurity and Privacy, University of Porto; INOV - Institute for Systems Engineering and Computer Innovation; TICE.PT - Information, Communication and Electronic Technologies Pole; and PricewaterhouseCoopers/AG), being always open to the entry of new entities that may contribute and add value to the consortium's mission. The C-Hub thus brings together, at a single point, a set of public and private entities with a technical, scientific, research and innovation know-how, which materialise a portfolio of services ranging from advanced training to demonstrators of emerging technologies, services that do not exist on the market.

3. In the EDIH network, how do the Cybersecurity DIH, the Joint Cyber Unit (JCU) and the European Cyber Security Industry, Technology and Research Competence Centre (ECCC) work together?

The recently published European Union (EU) Cybersecurity Strategy creates a new set of tools to build resilience to attack and ensure that citizens and businesses benefit from trusted digital technologies. The network of EDIHs, as its name sug-

3. Na rede EDIH, como se alinham o Cybersecurity DIH, a Unidade Conjunta de Cibersegurança (UCC) e o Centro Europeu de Competências Industriais, Tecnológicas e de Investigação em Cibersegurança (ECCC)?

A recentemente publicada Estratégia de Cibersegurança da União Europeia (UE) vem criar um conjunto de novos instrumentos que visam criar resiliência a ataques e assegurar que cidadãos e empresas beneficiam de tecnologias digitais confiáveis. A rede de EDIHs, como o próprio nome indica, pretende ligar entre si um conjunto de pólos de inovação que prestam serviços inovadores ao conjunto da UE. Por outro lado, o ECCC e a Rede de Centros Nacionais de Coordenação é a agência europeia, com sede em Bucareste, responsável pela execução dos fundos previstos no Programa Europa Digital para a área da cibersegurança e, dessa forma, criar e reter conhecimento tecnológico e capacidade industrial na área da cibesegurança na Europa. Já a UCC pretende ser uma plataforma de coordenação que reúne as várias comunidades de cibersegurança, tais como as equipas de resposta a incidentes, as autoridades nacionais de cibersegurança, as autoridades de gestão de crises de cibersegurança, os órgãos de polícia criminal, as entidades responsáveis pela ciberdefesa ou a ciberdiplomacia, com vista à produção de um quadro situacional europeu da cibersegurança, à melhor partilha de informação entre estas diferentes comunidades e à prestação de auxílio mútuo em caso de incidente de grande dimensão. São, portanto, três instrumentos de natureza bastante distinta - uma rede prestadora de serviços, uma entidade financiadora e uma estrutura de coordenação operacional. ■

gests, aims to link together a set of innovation hubs that provide innovative services to the EU as a whole. On the other hand, the ECCC and the Network of National Coordination Centres is the European agency, based in Bucharest, responsible for implementing the funds provided in the Digital Europe Programme for cybersecurity, thus creating and retaining technological knowledge and industrial capacity in cybersecurity in Europe. The JCU aims to be a coordination platform bringing together the various cybersecurity communities, such as incident response teams, national cybersecurity authorities, cybersecurity crisis management authorities, criminal police bodies, entities responsible for cyber defence or cyber diplomacy, with a view to producing a European cybersecurity situational picture, better sharing of information between these different communities and providing mutual assistance in the event of a major incident. Therefore, they are three instruments of quite different nature - a services provision network, a funding entity and an operational coordination structure. ■

João Damas

Investigador sénior, APNIC Labs
Consultor do .PT
Senior Researcher, APNIC Labs
Consultant .PT

Privacidade e o DNS (Parte I: o problema)

Os primeiros anos da Internet tinham como missão pôr as coisas a funcionar e desenvolverem-se ainda mais a partir desse ponto.

Como tal, a maioria dos protocolos definidos naqueles primeiros tempos não prestava muita atenção a questões de segurança (tanto integridade quanto visibilidade) porque a) apenas um pequeno grupo de participantes académicos iniciais a procurava e b) a segurança acrescenta complexidade e impõe um fardo ao desenvolvimento precoce de um sistema inteiramente novo. Por outras palavras, a introdução de mecanismos de segurança adicionais nos primeiros anos da Internet foi vista como desnecessária e como uma complicação que teria impedido o desenvolvimento dos protocolos propriamente ditos.

Hoje em dia, a situação é muito diferente. A Internet é utilizada por quase metade da população mundial para todos os tipos de fins, em todo o tipo de ambientes. Os protocolos em si são quase maduros e as capacidades dos dispositivos ligados à Internet são vastamente maiores do que as dos computadores utilizados para desenvolver a rede.

Privacy and the DNS (Part I: the problem)

The early days of the Internet were all about getting things to work and develop further from that point.

As such, most protocols defined in those early days did not pay a lot of consideration to aspects of security (both integrity and visibility) because a) no one other than the small group of initial academic participants was looking and b) security adds complexity and imposes a burden on the early development of an entirely new system. In other words, introduction of additional security mechanisms in the early days was both seen as unnecessary and a complication that would have prevented the protocols themselves from being developed.

Today the situation is very different. The internet is used by almost half of the world's population for all kinds of purposes, in all sorts of environments. The protocols themselves are almost mature and the capabilities of the devices connected to the Internet are vastly greater than those of the computers used to develop the net.

As pessoas utilizam a Internet para todo o tipo de transações: financeiras, comunicações pessoais, interações com instituições públicas, nas empresas, etc.

Todas estas comunicações têm de ser protegidas para que possam ser fiáveis. Quando se fala de segurança da informação, é comum concentrar-se na integridade e verificação dos dados mas, recentemente, existe um aspeto adicional que está finalmente a receber o nível de atenção que merece: a privacidade.

O DNS é um dos protocolos mais antigos da Internet, originalmente definido em 1985, e, claro, não incluía quaisquer características de segurança. Com o tempo, foi aumentado para proporcionar uma camada (opcional) de proteção da integridade dos dados (DNSSEC), mas continua a ser um protocolo de texto claro, o que significa que todas as transações podem ser vistas por qualquer terceiro que possa ver o tráfego entre o utilizador final e os diferentes servidores que utiliza para pedir a resolução ao DNS. Até há pouco tempo, muitas pessoas diriam que não é importante que se possa ver as consultas de um utilizador ao DNS porque, no máximo, apenas fornecem metadados, não mostram a informação a que o utilizador está a aceder. Como Edward Snowden deixou muito claro, os metadados recolhidos nas comunicações podem dizer muito sobre um utilizador.

Não se trata apenas de proteger os utilizadores em regimes políticos adversos. Por exemplo, se começar a olhar para alguns websites relacio-

People use the internet for all kinds of transactions: financial, personal communications, interaction with public institutions and inside business, etc.

All of these communications need to be protected so they can be trusted. When talking about information security it is common to focus on data integrity and verification but recently there is an additional aspect that is finally getting the level of attention it deserves: privacy.

The DNS is one of the Internet's oldest protocols, originally defined in 1985, and of course it did not include any security features. With time it has been augmented to provide a (optional) layer of data integrity protection (DNSSEC) but it is still a clear-text protocol, meaning all of the transactions can be seen by any third party that can see traffic between the end-user and the different servers it uses to ask for DNS resolution. Until recently a lot of people would say that it is not important that you can see a user's DNS queries because at most they only provide metadata, they don't show the information the user is accessing. As Edward Snowden made very clear, metadata collected in communications can tell a lot about a user.

This is not only about protecting users in adverse political regimes. For instance, if you start looking at a few medical-related websites and the list of those DNS queries is recorded and then accessed by an insurance

nados com a medicina e a lista dessas consultas ao DNS for registada e depois acedida por uma companhia de seguros (porque todos os dados estão à venda), poderá descobrir que os seus prémios sobem surpreendentemente ou pior ainda. Este é apenas um exemplo; com um pouco de imaginação, qualquer pessoa pode facilmente ver muitos cenários diferentes.

O outro lado da moeda é o acesso a metadados por partes com "interesses legítimos" como, por exemplo, as forças policiais durante uma investigação criminal. Um equilíbrio muito difícil de encontrar.

Nos últimos anos, a IETF, a comunidade que cria e desenvolve a maioria dos padrões da Internet, tomou a posição de que todas as comunicações dos utilizadores devem ser escondidas de quaisquer terceiros. Esta nova atitude está a traduzir-se em novos aditamentos ao protocolo de DNS, bem como em recomendações operacionais.

Iremos analisá-las na segunda parte deste artigo. ■

company (because all data is for sale) you might find out that your premiums surprisingly go up or worse. This is just one example, with a little bit of imagination anyone can easily see many different scenarios.

The other side of the coin is access to metadata by parties with "legitimate interests", for instance police forces during a criminal investigation. A very difficult balance to find.

In the last few years the IETF, the community that creates and evolves most Internet standards, has taken the position that all user communications should be hidden from any third parties. This new attitude is translating into new DNS protocol additions as well as operational recommendations.

We will look at these in the second part of this article. ■

A segurar a vida digital Holding on to digital life

A cibersegurança da eletrónica de consumo e os consequentes ataques à segurança doméstica e à privacidade deriva, muitas vezes, de vulnerabilidades nos próprios produtos eletrónicos. Uma certificação poderia mitigar alguns riscos da "smart TV", campainha eletrónica e sistema de alarme, aspirador-robô, chave "inteligente" da garagem ou até do monitor de bebés - todos com **vulnerabilidades críticas**. A Euroconsumers emitiu algumas **recomendações** para uma maior segurança destes produtos domésticos, que também se aplicam às PMEs. Em termos gerais, o objetivo é **garantir** uma **vida digital mais segura**.

The cybersecurity of consumer electronics and the consequent attacks on home security and privacy often stems from vulnerabilities in the electronic products themselves. A certification could mitigate some risks from smart TVs, electronic doorbells and alarm systems, robot vacuum cleaners, smart garage keys or even baby monitors - all with **critical vulnerabilities**. Euroconsumers has issued some **recommendations** to increase the safety of these household appliances, which also apply to SMEs. Broadly speaking, the aim is to **ensure** a **safer digital life**.

Código mais seguro com ML e IA Safer code with ML and AI

"Uma nova era de geração automatizada de código está a começar a tomar forma" e o resultado poderá ser o aparecimento de código mais seguro. Devido aos progressos na programação com "machine learning" (ML) e na geração de código por inteligência artificial (IA), estas ferramentas podem complementar e garantir uma **maior cibersegurança**, assim como uma maior fiabilidade no código escrito por humanos. Um exemplo destes sistemas é o **GitHub Copilot**.

'A new era of automated code generation is beginning to take shape' and the result could be the emergence of more secure code. Due to advances in machine learning (ML) programming and artificial intelligence (AI) code generation, these tools can complement and ensure **greater cybersecurity**, as well as greater reliability of human-written code. One example of such systems is **GitHub Copilot**.

Computação estratégica para emergências Strategic computing for emergencies

Datado de outubro, este "National Strategic Computing Reserve: a Blueprint" (NSCR) é um documento oficial dos EUA que procura assegurar o funcionamento da "reserva estratégica" da computação em caso de emergência.

O documento serve de arranque para o estabelecimento de um grupo interagências federais para as componentes estruturais e operacionais, estimular eventos para apresentar a NSCR e interligar-se a outras entidades de coordenação e resposta a emergências.

Dated October, this 'National Strategic Computing Reserve: a Blueprint' (NSCR) is an official US document that seeks to ensure the operation of the 'strategic computing reserve' in case of emergency.

The document serves as a kick-off to establish a federal interagency group for the structural and operational components, to stimulate events to present the NSCR and to interconnect with other emergency coordination and response entities.

A geografia do mundo online The geography of the online world

Uma visualização anual mostra o tamanho e a influência dos diferentes registos de país ("country code Top Level Domains" ou ccTLD) como o .pt. Neste caso, o tamanho dos registos reflete a extensão territorial do país, num mapa com fronteiras inesperadas.

An annual view shows the size and influence of different country code Top-Level Domain (ccTLD) registries, such as .pt. In this case, the size of the registrations reflects the territorial extent of the country, on a map with unexpected borders.

2020

Tipos de Registo DNS
Types of DNS registrations

Directora | Director
Inês Esteves

Edição | Editor
Pedro Fonseca

Design gráfico | Graphic design
Sara Dias

Tradução | Translation
Sara Pereira

Fotos | Photos
MIT Center for Advanced Virtuality

Publicação trimestral | Quarterly publication
Dezembro 2021 | December 2021

Cofinanciado pelo Mecanismo Interligar a Europa - União Europeia