Your guide to

AI Chatbots

An information security
eBook by TSC

Raise awareness

Develop knowledge

Change behaviour

Develop security culture

TSC eBooks

Helping you build a more secure future

Your guide to AI Chatbots is part of TSC's series of free publications covering some of the principal risks in information security.

These risks continually challenge information security professionals, who tirelessly seek to minimise them across their organisations and home users alike.

Our eBooks offer easy to digest information on the risks and give practical advice on how to address them. The content can be used in its entirety or in part, in standalone initiatives or as part of larger information security awareness programmes.

We hope you will use this and our other eBooks to further educate and raise awareness within your organisation to help change actions and shape attitudes towards a more information secure environment.

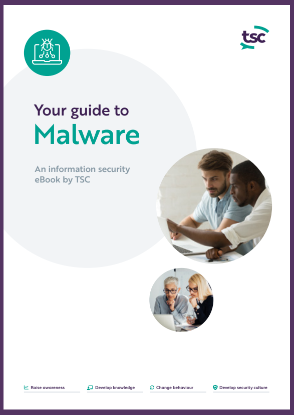
Feel free to encourage your employees to share the contents with their families and friends.

Your guide to
Data loss prevention
An information security eBook by TSC

Raise awareness | Develop knowledge | Change behaviour | Develop security culture

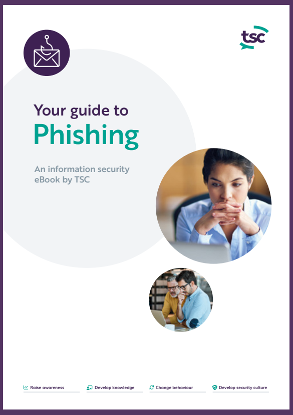
Your guide to
Passwords
An information security eBook by TSC

Raise awareness | Develop knowledge | Change behaviour | Develop security culture



Your guide to
Malware
An information security eBook by TSC

Raise awareness | Develop knowledge | Change behaviour | Develop security culture

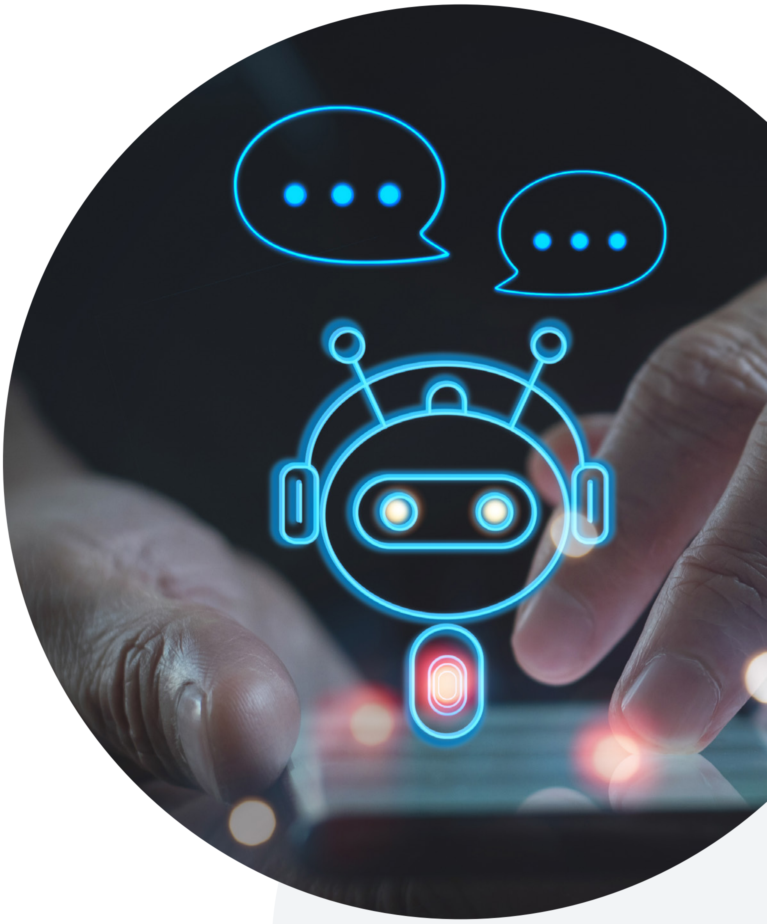


Your guide to
Phishing
An information security eBook by TSC

Raise awareness | Develop knowledge | Change behaviour | Develop security culture

Download other eBooks in the series from the TSC eBook resource page.

AI chatbots



Is the future bright? Or fraught with risk?

There's no denying that chatbots have exploded into our collective consciousness over the last few years. In a world that prizes convenience and 24/7 availability, chatbots have picked up where humans fall down.

In November 2022, OpenAI debuted ChatGPT, its conversational chatbot powered by artificial intelligence (AI). At the time of its release, it was the fastest-growing app in history, reaching one million users in five short days.

With all the fanfare surrounding its release and the subsequent deluge of media commentary on its applications and its potential, or not, to disrupt the world of work and even become sentient, you'd be forgiven for thinking it's the only AI chatbot in existence.

But you'd be wrong.

You can find chatbots all over the web – some big and powerful, others smaller and more niche.

With that in mind, what exactly are these chatbots? And what security risks arise from their use?

Fact

The market size for chatbots is predicted to increase 25% year-on-year until 2028[1].

Fact

98% of security stakeholders are concerned about the cyber security risks posed by generative AI tools[2].

---

[1] Ubisend, 2022 Chatbot and Marketing Agency Industry Report

[2] Abnormal Security, The State of Email Security in an AI-Powered World 2023

What is a chatbot?

At its most simple, a chatbot (short for **chatterbot**), is a piece of software designed to simulate conversation, or 'chatter' with users, typically over the internet.

In the age of AI, however, chatbots have become much more than that. Maybe you remember the old days of chatbots, when responses were highly automated, and you were in no doubt that you were speaking to a machine.

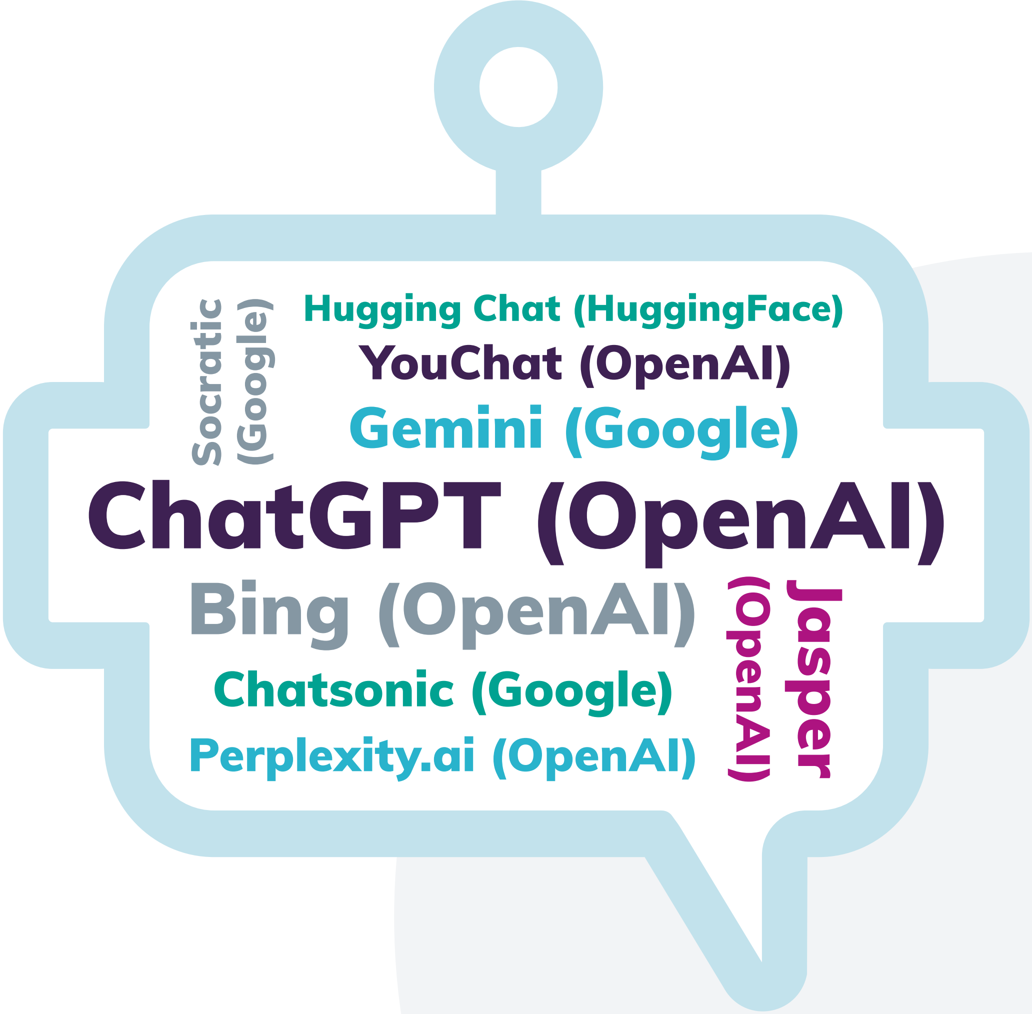
These chatbots increased in popularity as organisations began to automate online business-to-consumer communications in a bid to free up human staff from administrative tasks.

Chatbots could, for example, answer customer service queries (often before directing you to a human service agent when the chatbot was inevitably ill-equipped to answer anything slightly complex!)

Well, as AI continues to improve and evolve, that's probably no longer what comes to mind when chatbots come up in conversation. Instead, most people think of AI-powered software like ChatGPT.

ChatGPT belongs to a new generation of AI-powered chatbots that can mimic human conversation and, crucially, can **learn and adapt** to become even more human-like.

And with organisations increasingly adopting AI to improve their previously lacklustre online web assistants, these days it's becoming more difficult to know if you are talking to a real customer service advisor or an AI-powered bot.



ChatGPT

As one of the most high profile AI chatbots, let's delve a bit deeper into ChatGPT and how it works.

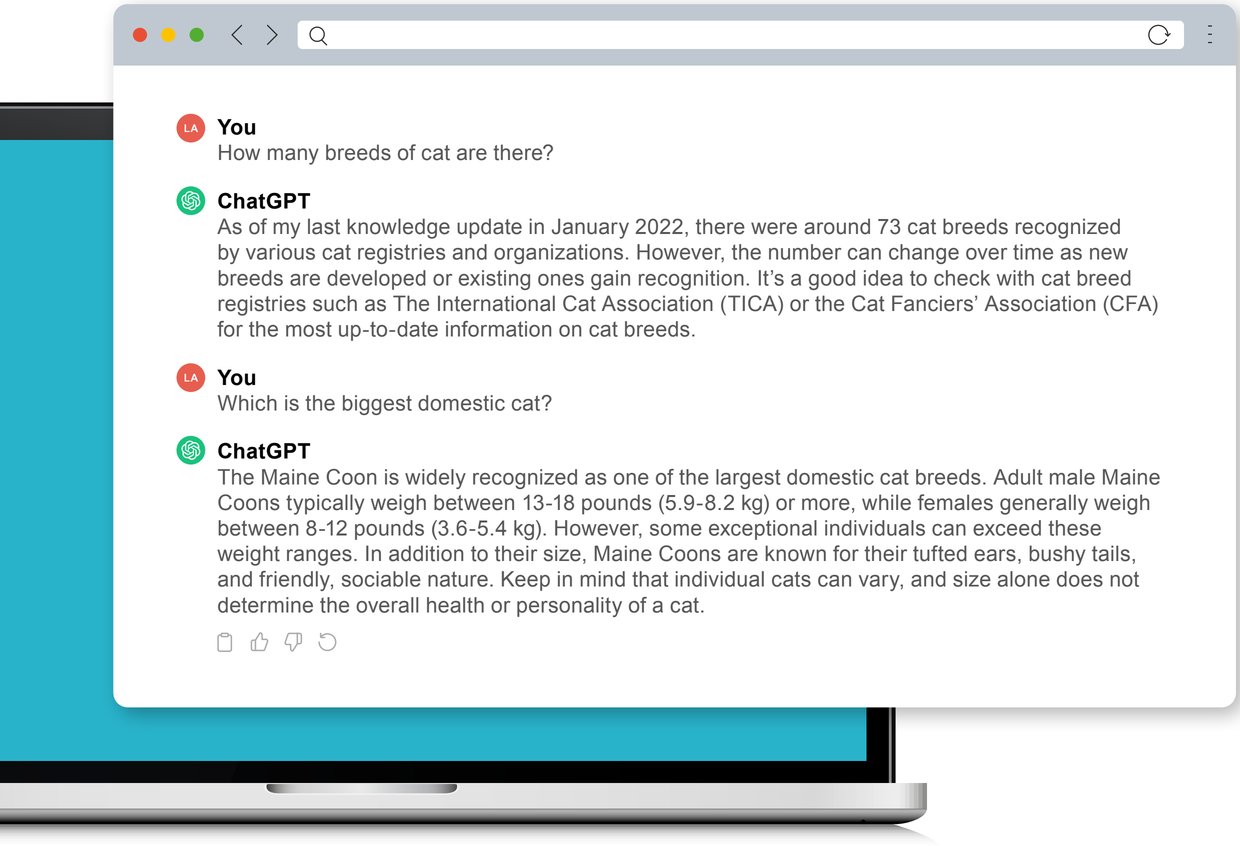
ChatGPT is a sophisticated chatbot which is a **Generative Pre-trained Transformer** (hence, GPT) – a **Large Language Model (LLM)** ⓘ.

GPT models, the first of which was introduced by OpenAI in 2018, are artificial neural networks that are pre-trained on large datasets. They use **natural language processing (NLP)** ⓘ to interpret and generate human language.

ChatGPT is an AI tool that has been trained to generate text that accurately imitates human language after it has been prompted with a query.

So, a user types a question into ChatGPT, and it responds in a human-like way.

ChatGPT uses **Machine Learning (ML)** ⓘ to continually learn from past patterns and human feedback to improve its responses, so as time goes on it becomes more accurate and sounds more and more human.



LA **You**
How many breeds of cat are there?

ChatGPT
As of my last knowledge update in January 2022, there were around 73 cat breeds recognized by various cat registries and organizations. However, the number can change over time as new breeds are developed or existing ones gain recognition. It's a good idea to check with cat breed registries such as The International Cat Association (TICA) or the Cat Fanciers' Association (CFA) for the most up-to-date information on cat breeds.

LA **You**
Which is the biggest domestic cat?

ChatGPT
The Maine Coon is widely recognized as one of the largest domestic cat breeds. Adult male Maine Coons typically weigh between 13-18 pounds (5.9-8.2 kg) or more, while females generally weigh between 8-12 pounds (3.6-5.4 kg). However, some exceptional individuals can exceed these weight ranges. In addition to their size, Maine Coons are known for their tufted ears, bushy tails, and friendly, sociable nature. Keep in mind that individual cats can vary, and size alone does not determine the overall health or personality of a cat.

---

[3] Malwarebytes, 2023 State of Malware

Chatbots and cyber security

While advocates of AI often point to the security benefits it can provide, the truth is that AI offers both a safety net for and a threat to cyber security.

Chatbots, for example, can be used by cybercriminals to create ever-more sophisticated attacks and can quickly analyse the success of previous attacks and automatically create new attacks with fewer weaknesses.

Cybercriminals could use chatbots to create personalised phishing attacks that are extremely difficult to spot. Or to write code for a security-evading piece of malware or ransomware.

In addition to their exploitation by cybercriminals, the use of AI chatbots also raises questions about the security of the data inputted, and the accuracy of the data generated.

Spotlight on safeguards

As with many other public generative AI tools, ChatGPT does have inbuilt safeguards designed to prevent it from being used for malicious purposes. However, these safeguards are not bulletproof and, as Europol has pointed out, loopholes exist[3].

Many of these loopholes involve what is known as **prompt engineering**. This is where users change the way they ask their questions to elicit a certain response from the chatbot.

Examples of prompt engineering workarounds include:

- Providing ChatGPT with an answer and asking it to provide the question
- Pretending to be a fictional character talking about the subject
- Asking ChatGPT to provide the answer as code

Why is it a problem?

You may ask why this is a problem – ChatGPT is trained on source data that is freely available online, so if a criminal wants to find it, they can.

But while that is the case, the truth is that ChatGPT significantly reduces the time it takes for a criminal to find the information they need.

For example, if they are researching a type of malware that they are unfamiliar with, they could of course scour the internet for information. Or they could ask ChatGPT, using one of the workarounds highlighted above, and it can simplify all of the information for them and give them a better understanding in no time at all.

---

[4] Europol, 'ChatGPT – the impact of Large Language Models on Law Enforcement', 17 April 2023

Phishing with AI

> [It] is now possible to impersonate an organisation or individual in a highly realistic manner even with only a basic grasp of the English language[4].

It used to be the case that phishing messages commonly contained tell-tale signs that made them easier to spot. For example, we have all been told to look out for spelling and grammar mistakes, impersonal greetings and branding errors.

With the advent of AI and chatbots, however, these signs may no longer appear.

Cybercriminals can use chatbots to create personalised and professional-looking phishing messages, making them much harder to detect.

A closer look

Generative AI is drastically reshaping the phishing threat landscape, research has revealed.

According to IBM X-Force, its team of cyber security experts was able to trick a generative AI model into creating a 'highly convincing' phishing email that was as advanced as one created by skilled social engineers.

Using five simple prompts, the team was able to trick ChatGPT into constructing a sophisticated phishing email in just five minutes.

This could save phishers nearly two days' work when developing phishing attacks[5].

Spot the signs

Do your employees know what to look out for when spotting AI-generated phishing messages? Why not print our cut-out-and-keep leaflet, **'Top tips to spot an AI phishing scam'**, at the end of this eBook as a handy guide to distribute across your organisation?

Fact

80% of security stakeholders say their organisations have received AI-generated email attacks, or they strongly suspect they have[6].

---

[4] Europol, 'ChatGPT - the impact of Large Language Models on Law Enforcement', 17 April 2023
[5] Security Intelligence, 'AI vs. human deceit: Unravelling the new age of phishing tactics', 24 October 2023
[6] Abnormal Security, The State of Email Security in an AI-Powered World 2023

Malware generation

While chatbots cannot yet replace the skills and know-how of advanced threat actors, security experts say there is evidence that tools like ChatGPT can help low-skilled cybercriminals to create malware.

As we've discussed earlier in this eBook, ChatGPT does reject obviously malicious prompts, but it's easy for cybercriminals to create workarounds. For example, they can ask ChatGPT to create code for penetration testing, then adapt it for use in attacks[7].

Researchers from Check Point have also found evidence on the dark web of cybercriminals using ChatGPT to improve their malware code[8].

Chatbot impersonation

We've already considered how it is becoming increasingly difficult to distinguish between a human and a chatbot – but what about distinguishing between a real chatbot and a fake one?

One way threat actors can abuse the abundance of online chatbots is by impersonating a real chatbot to elicit sensitive information or credentials from unsuspecting victims. In fact, that's exactly what happened in the Facebook Messenger scam detailed in the below case study.

Case study

Facebook Messenger chatbot scam[9]

In June 2022, security service provider Trustwave discovered a new phishing campaign that used a fake chatbot to steal users' login credentials.

The attack started like many other phishing scams – with an email. The email was sent to administrators of Facebook pages, stating that their page had violated Community Standards and that the recipient had 48 hours to act or their account would be deleted.

The victims were advised to 'Appeal' the decision by clicking a link which would take them to a fake Facebook Messenger chat with a 'support' chatbot.

The chatbot would direct victims to another website to complete an appeal form which would request their login details and other personal information.

---

[7] TechTarget, '5 ChatGPT security risks in the enterprise', 7 April 2023
[8] Check Point, 'Cybercriminals Bypass ChatGPT Restrictions to Generate Malicious Content', 7 February 2023
[9] Bleeping Computer, 'Malicious Messenger chatbots used to steal Facebook accounts', 28 June 2022

Data loss or compromise

Anywhere that there is an exchange of valuable personal or sensitive information, there is a risk that the information could be intercepted, lost or compromised. And so it is with chatbots.

In some cases, the risk of data compromise is stated quite clearly, even if it is not fully understood by users. For example, ChatGPT's privacy policy states that all conversations are recorded and can be shared with other companies and used to continue to develop the AI tool.

This means that any confidential information entered into ChatGPT is put at serious risk, as some organisations have already discovered. Samsung and Amazon, for example, both issued bans on their employees using ChatGPT due to sensitive data leaks.

There are also risks to consumer data where companies use poorly secured chatbots on their websites, for example for support chats. If these chatbots request sensitive information and the chatbot is not adequately encrypted, the information could potentially be intercepted by malicious actors.

From the archive

Samsung bans ChatGPT after data leak

April 2023

Technology giant Samsung Electronics has banned ChatGPT after meeting notes and other sensitive company information was leaked online by its employees.

The Samsung employees reportedly disclosed sensitive company information, on three occasions, in the form of prompts to ChatGPT.

As well as a recording of an internal meeting, sensitive database source code was entered into the Microsoft-backed chatbot, which uses users' prompts to train its Machine Learning models.

From the archive

Amazon warns employees against ChatGPT

January 2023

Amazon has reportedly warned its employees not to share software code or other confidential company information with ChatGPT.

The warning comes after Amazon was made aware of ChatGPT responses that accurately mimicked real internal Amazon data, suggesting that the chatbot had used confidential information inputted by Amazon employees to inform its dataset.

> ChatGPT excels at providing answers that sound very plausible, but that are often inaccurate or wrong[10].

Chatbot bias and fake news

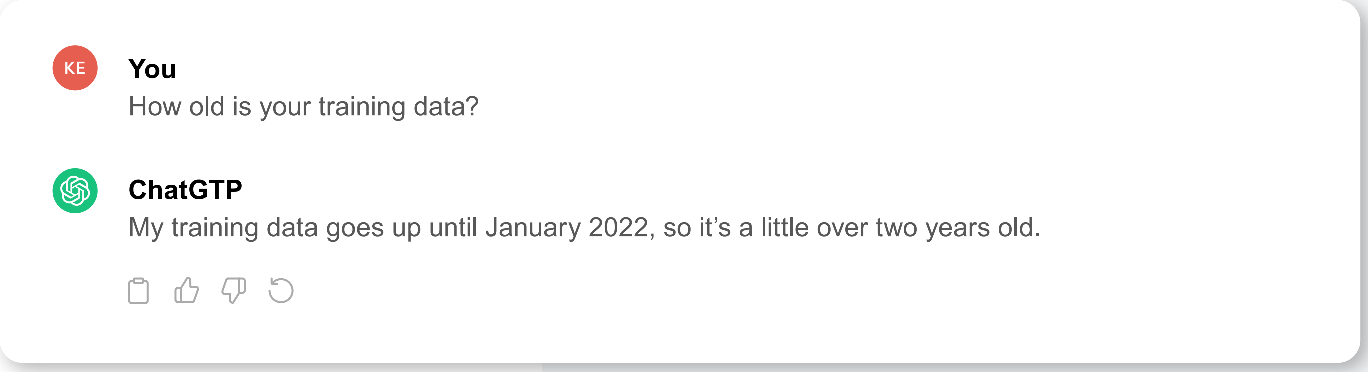
As Europol has pointed out, ChatGPT might be very good at emulating human language, but it does not intrinsically understand it. It cannot understand the meaning behind human language, it simply spots patterns.

This means that its answers can be inaccurate, which, if users are relying on it as a source of truth, can be dangerous.

In addition, since humans inherently display bias, and since ChatGPT is trained on human-derived data, it can reinforce those biases and lead to discrimination.

Furthermore, the current iteration of ChatGPT (as at the date of this eBook's publication) is trained on data from **January 2022**. This makes the data out-of-date and increases the potential for bias.

Even when the training datasets are updated, ChatGPT and other LLM chatbots are always running behind. Based on how they currently work, they can never be fully up to date.



You
How old is your training data?

ChatGTP
My training data goes up until January 2022, so it's a little over two years old.

A ChatGPT screenshot from February 2024.

---

[10] Europol, 'ChatGPT - the impact of Large Language Models on Law Enforcement', 17 April 2023

Staying secure

To stay secure as chatbots become more pervasive, follow these top tips:

In a world of AI

- Know the signs of AI-generated scams, and always be alert to anything that seems suspicious
- Stay up to date with the latest phishing and cyber attack tactics and update phishing and training communications so everybody knows what to look out for

Not sure where to start? TSC can help! Contact us today to bring your cyber threat communications up to date

When you are using a chatbot

- Only use chatbots on websites you have purposely accessed – do not use chatbots you reached by clicking on links in emails, texts, social media messages or pop-ups
- Only use chatbots like ChatGPT for personal use, unless explicitly authorised to use it for work
- If you are authorised to use a chatbot for work purposes, never disclose sensitive company information
- If you are using a chatbot in a personal capacity, never disclose personal or confidential information
- Use your judgement to scrutinise the chatbot's responses for inaccuracies and bias

If your organisation utilises public-facing chatbots

- If your business uses chatbots on its public-facing website, always ensure that they utilise end-to-end encryption
- Encryption is particularly important if the chatbot is customer-facing and has the potential to collect sensitive or personal information
- Use strong authentication methods to ensure only authorised users can access chatbot data – as a minimum, ensure your accounts are protected with multi-factor and biometric authentication methods
- Consider implementing self-destructing messages, so that user queries and responses are destroyed rather than retained
- Ensure your website is protected with Transport Layer Security (TLD) or Secure Sockets Layer (SSL)
- If you use data from previous chatbot interactions for training purposes, ensure it is anonymised first
- Conduct penetration testing to ensure your chatbots are secure

Chatbot security: Let's keep talking about it

As with all new technologies, people are often so dazzled by the possibilities that they forget to think about security. That's why it's vital we continue to talk about the potential cyber security weaknesses of chatbots, from public generative AI tools like ChatGPT through to customer service and support chatbots on consumer-facing websites.

From the risks of using them to the risks of them being exploited by malicious actors, the dangers posed by chatbots are diverse. It is only through continuous education and research that we can minimise the threats and take action to keep ourselves, our people, our organisations and our data secure.

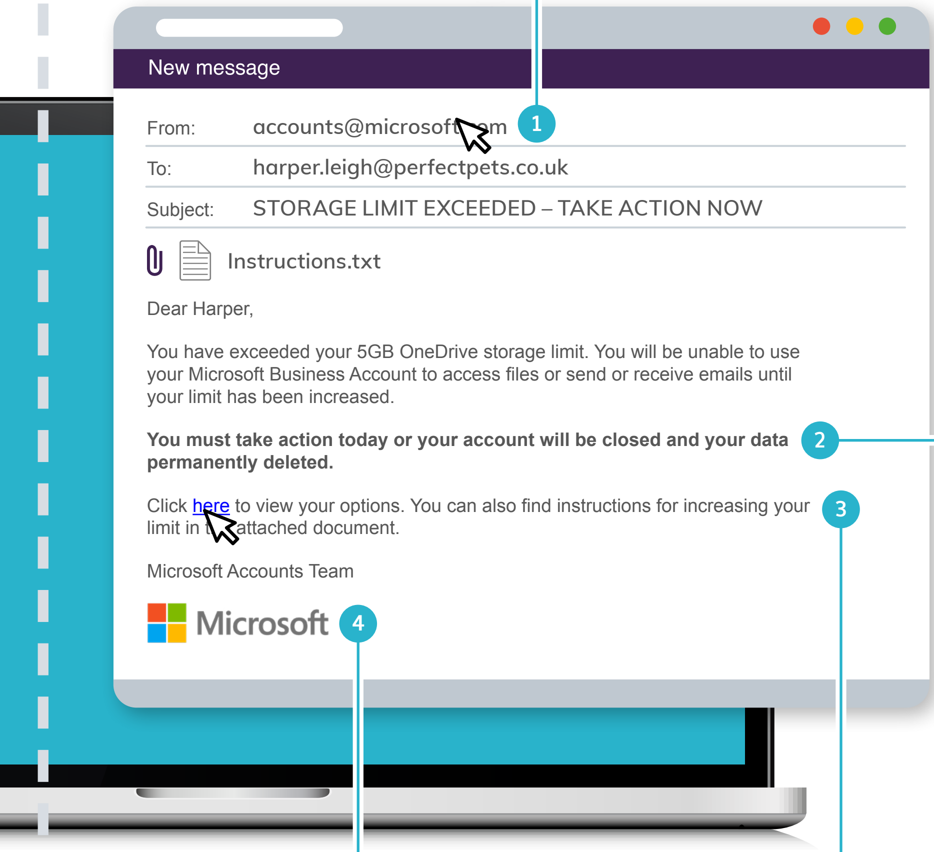
How TSC can help

At TSC, we specialise in creating bespoke behaviour change and awareness campaigns, including training specifically devised to educate employees on up-and-coming security trends and risks, from AI to the metaverse to deepfakes. Our tried-and-tested techniques are designed to help organisations strengthen their security posture by changing their culture and adopting an employee-focused strategy.

We know that your employees are your strongest form of defence against cyber attacks and data loss – let's work together so your employees know that too.

Top tips to spot an AI phishing scam

Did you know that the old tell-tale signs of a phishing email, such as spelling and grammar errors, are no longer relevant in a world of AI? Cybercriminals are increasingly using AI to generate sophisticated phishing scams that are extremely difficult to spot. These are some of the signs you can look out for.



From: accounts@microsoft.com
To: harper.leigh@perfectpets.co.uk
Subject: STORAGE LIMIT EXCEEDED – TAKE ACTION NOW

Instructions.txt

Dear Harper,

You have exceeded your 5GB OneDrive storage limit. You will be unable to use your Microsoft Business Account to access files or send or receive emails until your limit has been increased.

You must take action today or your account will be closed and your data permanently deleted.

Click here to view your options. You can also find instructions for increasing your limit in the attached document.

Microsoft Accounts Team

Microsoft

microsoft-accounts@hotmail.co.uk

Cybercriminals can spoof email addresses. Rolling over a sender's name can reveal the true origin of the email. For example, here the email has actually been sent from a free-to-use Hotmail account, not a genuine Microsoft address.

Phishers use tried-and-tested social engineering techniques, which AI and chatbots can help them to hone. Often, they will try to appeal to their potential victim's emotions to get them to act quickly before they realise that they are being duped.

Phishers will often pretend to be a person or a company that is well-known and trusted. They can use chatbots to create emails that sound and look like they came from that person or organisation. This can include using the correct language and tone of voice, as well as logos and other brand identifiers.

https://onedrlve-live/123/uk.com/

Requests to click on suspicious links or to download malicious attachments are still top tools in the phisher's toolbox.

©The Security Company (International) Limited 2024. This leaflet may be distributed internally within your organisation for educational purposes. It must not be copied, replicated, edited, or distributed externally without the express permission of The Security Company (International) Limited.

Glossary

Artificial Intelligence (AI):
The use of computers and machines to mimic human intelligence.

Chatbot:
A software application designed to mimic human conversation, especially online.

Deep Learning (DL):
An advanced type of Machine Learning that doesn't require 'structured' datasets. Its algorithms learn and adjust through trial and error.

Generative AI:
A type of AI technology that can create (generate) new content, including text, images and audio.

Large Language Model (LLM):
A deep learning algorithm that can understand and generate human language.

Machine Learning (ML):
A branch of AI which is concerned with using data and algorithms to imitate the way humans learn; constantly assessing and analysing datasets to develop and improve.

Natural Language Processing (NLP):
A tool that allows machines to interpret and generate human language.

Contact TSC for more information about protecting yourself from malicious software and for other information and cyber security support and resources.

www.thesecuritycompany.com @TSCPeopleSec @thesecurityco @thesecurityco

©The Security Company (International) Limited 2024. This eBook may be distributed internally within your organisation for educational purposes. It must not be copied, replicated, edited, or distributed externally without the express permission of The Security Company (International) Limited.