OUCH! Mars 2025

La lettre d'information mensuelle de sensibilisation à la sécurité pour vous

Méfiez-vous des deepfakes : une nouvelle ère de tromperie

Pris au dépourvu : l'histoire de Steve

Steve était à son bureau lorsqu'il a reçu un appel vidéo de sa manager, Bela. Elle semblait stressée et parlait rapidement. "J'ai besoin que vous m'envoyiez immédiatement le rapport confidentiel client à cette nouvelle adresse mail !". En voyant son visage et en entendant sa voix familière, Steve n'a pas hésité longtemps et a envoyé le rapport confidentiel à la nouvelle adresse mail.

Quelques heures plus tard, Bela est entrée dans son bureau et a demandé le rapport. Steve était confus et a parlé de l'appel vidéo. Le visage de Bela s'est figé : elle ne l'avait pas appelé. La personne qu'il avait vue en vidéo n'était pas Bela. C'était un *deepfake*, créé par un cyber-criminel pour le piéger.

Steve n'en revenait pas de la crédibilité du faux appel. Le visage, la voix, tout correspondait à sa cheffe. Il venait d'être victime de cette cybermenace croissante où les criminels utilisent l'intelligence artificielle (IA) pour créer des contrefaçons très convaincantes.

Qu'est-ce que le deepfake ?

L'IA peut créer des images, des sons ou des vidéos qui semblent réels. Ces fonctionnalités ont de nombreuses utilisations légitimes. Par exemple, les sociétés de marketing utilisent cette technologie pour créer des images pour leurs campagnes publicitaires, les sociétés de cinéma pour retirer les signes de l'âge de certains acteurs et les enseignants pour créer des cours vidéo dynamiques pour leurs élèves.

On parle de deepfake lorsque l'IA est utilisée pour créer de fausses images, de faux sons ou de fausses vidéos dans le but de tromper les autres. Le nom "deepfake" vient d'une combinaison de "deep learning" (un type d'IA) et "fake" (faux).

Les deepfakes les plus préjudiciables sont souvent ceux qui consistent à créer de fausses images, de faux sons ou de fausses vidéos de personnes que vous connaissez peut-être, en leur faisant faire des choses qu'elles n'ont en réalité jamais faites. Par exemple, les cyberattaquants peuvent créer de fausses photos de célébrités ou d'hommes politiques en train de commettre un crime et les diffuser sous forme de fausses informations. Ou bien, ils peuvent cloner la voix de quelqu'un et l'utiliser dans un appel pour imiter la famille ou les collègues de la victime. Ce qui rend les deepfakes particulièrement dangereux, c'est la facilité avec laquelle les cybercriminels peuvent reproduire n'importe qui, lui faire faire n'importe quoi, et faire paraître tout cela réel.

Trois types de deepfake

1. Les images deepfakes
Il s'agit soit de photos de fausses personnes créées par l'IA, soit de photos de personnes réelles mais montrant qu'elles font quelque chose qu'elles n'ont jamais fait. Ces fausses images peuvent se propager rapidement et sont souvent utilisées pour nuire à la réputation d'une personne ou à manipuler les émotions. Les images "deepfake" sont de plus en plus courantes sur les réseaux sociaux, et des personnes, voire des gouvernements, tentent de diffuser de fausses histoires ou de faux récits (appelés "fake news") afin d'atteindre un certain objectif.

2. Les audios deepfakes (clone de voix)
Il s'agit de faux enregistrements ou d'appels téléphoniques utilisant la voix clonée d'une personne. Les attaquants peuvent obtenir des enregistrements de voix de personnes à partir de podcasts ou de YouTube, puis utiliser ces enregistrements pour reproduire leur voix. Une fois la réplique effectuée, les cyberattaquants peuvent appeler qui ils veulent en se faisant passer pour cette personne. Par exemple, quelqu'un peut se faire passer pour un directeur et appeler un employé pour lui demander des données sensibles, ou recréer la voix d'un proche lors d'un appel d'urgence pour lui demander de l'argent.

3. Les vidéos deepfakes
Il s'agit de fausses vidéos dans lesquelles la voix et les actions des personnes sont manipulées ou recréées. Les vidéos "deepfake" peuvent être des vidéos préenregistrées ou des vidéos en direct, par exemple lors d'une conférence téléphonique en ligne. Par exemple, les cyberattaquants peuvent fabriquer une vidéo "deepfake" d'un PDG faisant une fausse annonce sur son entreprise ou d'un homme politique semblant dire quelque chose qu'il n'a jamais dit.

Comment détecter les deepfakes : se concentrer sur le contexte

N'essayez pas de détecter les deepfakes en recherchant des erreurs techniques. L'IA et les cyberattaquants qui l'utilisent sont devenus très sophistiqués. Concentrez-vous plutôt sur le contexte. L'image, le son ou la vidéo ont-ils un sens ?

1. Faites confiance à votre instinct : est-ce que l'interaction vous parait étrange ? Est-ce que la demande est urgente ou inattendue ? Est-ce que la personne se comporte de manière étrange, même si elle parait normale ? Est-ce qu'elle vous demande des informations confidentielles ou des données personnelles auxquelles elle ne devrait pas avoir accès ? Si quelque chose ne vous semble pas normal, faites confiance à votre instinct et vérifiez à nouveau avant d'accéder à la demande.

2. Méfiez-vous de la manipulation émotionnelle : les cyberattaquants créent souvent un sentiment d'urgence ou de peur pour vous inciter à agir rapidement. Si un appel ou un message vous met dans un état de panique, prenez une inspiration et vérifiez. Plus l'attraction émotionnelle est forte, par exemple en créant un fort sentiment d'urgence ou de peur, plus il y a de chances qu'il s'agisse d'une attaque potentielle.

3. Vérifiez par une autre méthode : si vous craignez que la personne qui vous contacte soit un deepfake, contactez-la en utilisant une autre méthode. Par exemple, pour les appels vidéo ou les messages dont vous avez peur de l'authenticité, contactez la personne par téléphone ou par email. Si vous recevez un appel vocal vous demandant une action urgente, raccrochez et rappelez en utilisant un numéro de confiance.

4. Établissez un mot ou une phrase de code : convenez d'un mot ou d'une phrase de code partagé, connu uniquement au sein d'un groupe ou peut-être de votre famille, qui peut être utilisé pour authentifier une communication urgente.

Rédacteur invité

Dhruti Mehta est analyste de la sécurité de l'information au Physicians Health Plan of Northern Indiana et présidente de WiCyS Northern Indiana. Elle se passionne pour la création d'une main-d'œuvre diversifiée dans le domaine de la cybersécurité et pour le comblement des lacunes en matière d'éducation et de compétences dans ce domaine. https://www.linkedin.com/in/dhrutimehtacyber/

Ressources

Déclencheurs émotionnels : comment les escrocs vous piégent : https://www.sans.org/newsletters/ouch/emotional-triggers-how-cyber-attackers-trick-you/
Clonage de voix : https://www.sans.org/newsletters/ouch/phantom-voices-defend-against-voice-cloning-attacks/

Traduit pour la communauté par : Juliette Busson