OUCH! Marzec 2025

Biuletyn Bezpieczeństwa Komputerowego

Deepfakes: Nowa era oszustw

Poznajcie historię Piotra

Pewnego dnia do Piotra zadzwoniła na wideo rozmowie jego menadżerka Kasia. Wyglądała na zestresowaną i szybko mówiła. „Musisz natychmiast wysłać raport do klienta na ten nowy adres e-mail!" - nalegała. Widząc jej twarz i słysząc jej głos, nie zawahał się i wysłał raport na podany adres e-mail.

Kilka godzin później Kasia weszła do jego biura i zapytała o raport. Zdezorientowany Piotr wspomniał o połączeniu wideo. Kasia zbladła, przecież nie dzwoniła do niego. Piotr dopiero teraz zrozumiał, że osoba którą widział na nagraniu, nie była jego przełożoną. Był to *deepfake*, stworzony przez cyberprzestępcę aby go oszukać.

Piotr nie mógł uwierzyć, że było to fałszywe połączenie. Twarz, głos, wszystko idealnie pasowało do jego przełożonej. Padł ofiarą rosnącego zagrożenia cybernetycznego, w którym przestępcy wykorzystują sztuczną inteligencję (AI) do tworzenia bardzo realistycznych podróbek.

Czym jest Deepfake?

Sztuczna inteligencja może tworzyć obrazy, dźwięki lub filmy, które wyglądają jak prawdziwe. Możliwości te mają wiele uzasadnionych zastosowań. Na przykład firmy marketingowe wykorzystują tę technologię do tworzenia obrazów na potrzeby kampanii reklamowych, wytwórnie filmowe używają jej do postarzania niektórych aktorów, a nauczyciele do tworzenia dynamicznych lekcji wideo dla swoich uczniów.

Zatem Deepfake to nic innego jak wykorzystywanie sztucznej inteligencji do tworzenia fałszywych obrazów, audio lub wideo w celu oszukania innych. Nazwa „deepfake" pochodzi od połączenia słów „deep learning" (rodzaj sztucznej inteligencji) i „fake" (fałszywy).

Często najbardziej szkodliwe deepfake'i mają miejsce, gdy cyberprzestępcy tworzą fałszywe obrazy, audio lub wideo osób, które możesz znać, zmuszając je do robienia rzeczy, których w rzeczywistości nigdy nie zrobiły. Na przykład cyberprzestępcy mogą tworzyć fałszywe zdjęcia znanych celebrytów lub polityków popełniających przestępstwo i rozpowszechniać je jako fałszywe wiadomości. Mogą też sklonować czyjś głos i użyć go w rozmowie telefonicznej, aby oszukać rodzinę lub współpracowników ofiary. To, co sprawia, że deepfake'i są szczególnie niebezpieczne, to łatwość z jaką cyberprzestępcy mogą replikować kogokolwiek.

Trzy rodzaje Deepfake'ów

1. Fałszywy obraz

Są to albo zdjęcia fałszywych ludzi stworzone przez sztuczną inteligencję, albo zdjęcia prawdziwych ludzi, ale pokazujące, że robią coś czego nigdy nie zrobili. Te fałszywe obrazy mogą szybko się rozprzestrzeniać i są często wykorzystywane do niszczenia reputacji lub manipulowania emocjami. Niestety stają się coraz bardziej powszechne w mediach społecznościowych, a ludzie próbują promować fałszywe historie lub narracje (zwane fałszywymi wiadomościami), aby wpłynąć na określony cel końcowy.

2. Klonowanie głosu

Są to fałszywe nagrania lub rozmowy telefoniczne wykorzystujące czyjś głos. Atakujący mogą uzyskać nagrania głosów ludzi z podcastów lub YouTube, a następnie użyć tych nagrań do replikacji ich głosu. Po replikacji głosu cyberprzestępcy mogą zadzwonić do dowolnej osoby używając fałszywego głosu. Na przykład, ktoś mógłby udawać kierownika i zadzwonić do pracownika z prośbą o poufne dane lub ktoś mógłby odtworzyć głos ukochanej osoby z prośbą o pieniądze.

3. Wideo

Są to fałszywe filmy, w których głos i działania ludzi są manipulowane lub odtwarzane. Filmy Deepfake mogą być nagrane wcześniej lub na żywo, na przykład podczas połączenia konferencyjnego online. Cyberprzestępcy mogą stworzyć fałszywe wideo przedstawiające dyrektora generalnego wygłaszającego fałszywe oświadczenie na temat swojej firmy lub polityka, który wydaje się mówić coś, czego nigdy nie powiedział.

Jak wykrywać podróbki: Skup się na kontekście

Nie próbuj wykrywać deepfake'ów, szukając błędów technicznych. Zarówno sztuczna inteligencja, jak i wykorzystujący ją cyberprzestępcy ostatnimi czasy doszli niemalże do perfekcji. Zamiast tego skup się na kontekście. Czy obraz, dźwięk lub wideo mają sens?

1. Zaufaj swojemu instynktowi: Czy coś jest nie tak w tej interakcji? Czy ta prośba jest pilna lub nieoczekiwana tj. czy trafiła do odpowiedniej osoby? Czy dana osoba zachowuje się dziwnie, nawet jeśli wygląda i brzmi normalnie? Czy ktoś prosi o poufne informacje lub dane osobowe, do których nie powinien mieć dostępu? Jeśli coś wydaje się nie w porządku, zaufaj swojemu przeczuciu i sprawdź dokładnie, zanim spełnisz ich prośbę.

2. Uważaj na manipulację emocjonalną: Cyberprzestępcy często wywołują nagłą potrzebę lub strach, aby zmusić cię do szybkiego i nieprzemyślanego działania. Jeśli wiadomość lub połączenie z jakiś powodów wywołuje panikę, zweryfikuj je. Im silniejsze wywoływanie poczucia pilności lub strachu, tym bardziej prawdopodobne jest, że może to być potencjalny atak.

3. Weryfikacja za pomocą innej metody: Jeśli masz podejrzenia, że osoba kontaktująca się z Tobą może być oszustem, skontaktuj się z nią za pomocą innej metody. Na przykład, w przypadku połączeń wideo lub wiadomości, które mogą być fałszywe, skontaktuj się z daną osobą telefonicznie lub mailowo. Natomiast w przypadku połączenie głosowego z prośbą o natychmiastowe podejrzane działanie, rozłącz się i zadzwoń do tej osoby używając znanego Ci numeru.

4. Ustalenie słowa lub frazy kodowej: Ustal z rodziną lub daną grupą konkretną frazę znaną tylko Wam, która może być używana do uwierzytelniania podejrzanej komunikacji. Możecie też zadawać pytanie, na które odpowiedź znana jest tylko Wam np. data urodzin członka rodziny.

Redaktor gościnny

Dhruti Mehta jest analitykiem ds. bezpieczeństwa informacji w Physicians Health Plan of Northern Indiana i prezesem WiCyS Northern Indiana. Jej pasją jest budowanie zróżnicowanej siły roboczej w dziedzinie cyberbezpieczeństwa oraz wypełnianie luk edukacyjnych i umiejętności w tej dziedzinie. https://www.linkedin.com/in/dhrutimehtacyber/

Źródła

Działania na emocjach - o tym jak cyberprzestępcy oszukują: https://www.sans.org/newsletters/ouch/emotional-triggers-how-cyber-attackers-trick-you/
Ataki z wykorzystaniem odwzorowanego głosu: https://www.sans.org/newsletters/ouch/phantom-voices-defend-against-voice-cloning-attacks/

Polski przekład CERT Polska: Aleksandra Węgrzynowicz, Bartłomiej Wnuk