dr inż. Ewelina Bartuzi-Trokielewicz – kierownik Zakładu Analiz Audiowizualnych i Systemów Biometrycznych, ekspertka w dziedzinie biometrii, przetwarzania danych audiowizualnych i sztucznej inteligencji, specjalizująca się w wykrywaniu oraz analizie materiałów syntetycznych, w tym manipulacji wizerunku twarzy i głosu. Jej szczególnym obszarem zainteresowań są technologie deepfake i ich zastosowania w kontekście bezpieczeństwa oraz ochrony danych.

AK: Czym dokładnie są deepfake'i i na jakiej technologii bazują?

EBT: Deepfake'i to nieprawdziwe materiały audiowizualne wytworzone syntetycznie przy użyciu technik sztucznej inteligencji. Głównie opierają się na modelach generatywnych i dyfuzyjnych. Mogą być bardzo realistyczne, co sprawia, że osoby, niemające rozwiniętych zdolności krytycznego myślenia i oceny źródeł, mogą je przyjmować bezkrytycznie jako prawdziwe. Modele GAN, które są podstawą wielu deepfake'ów, działają na zasadzie rywalizacji dwóch sieci neuronowych – generatora i dyskryminatora. Generator tworzy sztuczne materiały,
a dyskryminator ocenia ich autentyczność. Dzięki temu procesowi model uczy się tworzyć coraz bardziej realistyczne obrazy, materiały wideo czy dźwięki. Z kolei modele dyfuzyjne, umożliwiają modyfikację i generowanie materiałów na podstawie tekstowych poleceń (promptów), dając zupełnie nowe możliwości manipulacji treściami.
Sam termin powinniśmy zarezerwować wyłącznie dla materiałów tworzonych w celach negatywnych – czyli takich, które mają na celu wprowadzenie w błąd, manipulację, oczernianie lub wywoływanie szkód. Taki podział pozwoli także wyraźnie odróżnić szkodliwe praktyki od etycznych i innowacyjnych zastosowań generatywnej AI, które mogą wspierać kreatywność i pozytywnie wpływać na społeczeństwo.
Technologia deepfake opiera się na manipulacji głosem i obrazem, w szczególności twarzy. Jeśli chodzi o głos wyróżniamy dwie główne techniki: TTS, czyli Text-to-Speech, i STS, czyli Speech-to-Speech, które pozwalają na tworzenie syntetycznej mowy, z dowolnym głosem referencyjnym:

- **Text-to-Speech (TTS)** to metoda konwersji tekstu na mowę w stylu i brzmieniu głosu referencyjnego. Mając próbkę głosu danej osoby, model „uczy się" brzmienia tego głosu – jego barwy, intonacji i tempa mowy. Dzięki temu, jest w stanie wygenerować sygnał mowy o wskazanej treści tak, jakby mówiła to ta konkretna osoba.
- Drugim podejściem jest bezpośrednia konwersja głosu, zwana metodą **Speech-to-Speech (STS).** Pozwala na przekształcanie jednego głosu w inny. Możemy to porównać do procesu tworzenia cyfrowych replik ludzkich głosów umożliwiając jednej osobie mówienie głosem innej. Jest bardziej zaawansowaną i potencjalnie bardziej niebezpieczną technologią niż TTS, ponieważ może generować głos niemal w czasie rzeczywistym, umożliwiając np. podszywanie się pod kogoś podczas rozmowy telefonicznej czy wideo. Jednak, aby system mógł nauczyć się głosu, potrzebuje dłuższej próbki – zwykle kilku minut, a nawet godzin nagrania.

Z kolei w obszarze obrazu wyróżniamy znacznie więcej rodzajów manipulacji. Można wytwarzać **całkowicie syntetyczne twarze**, czyli wizerunki ludzi, którzy prawdopodobnie nie istnieją. Bardzo popularną techniką jest **podmiana twarzy** (*face swap*), w której twarz jednej osoby zostaje zastąpiona twarzą innej na materiale wideo lub zdjęciu. Kluczowym elementem tej techniki jest zachowanie spójności wizualnej, aby podmieniona twarz wyglądała realistycznie i była idealnie dopasowana pod względem oświetlenia, perspektywy, ruchu głowy i mimiki do oryginalnego materiału.

Można także łączyć cechy twarzy różnych osób, co oszuści wykorzystują w fałszywych dokumentach tożsamości. Innym rodzajem manipulacji jest zmiana atrybutów takich jak wiek, kolor skóry, płeć czy zarost. Wraz z rozwojem modeli dyfuzyjnych pojawiły się narzędzia pozwalające na modyfikacje obrazu za pomocą poleceń tekstowych tzw. prompt'ów.

Innym przypadkiem jest **animacja twarzy** – technologia pozwalająca przenieść mimikę z jednej osoby na inną. Istotnym jej rozwinięciem jest **lipsync**, czyli synchronizacja ruchu ust z dowolną ścieżką głosową. To właśnie ta technika jest najczęściej stosowana w fałszywych materiałach, pozwalając na modyfikację materiałów wideo poprzez podmianę treści i dopasowanie ruchu ust do nowego nagrania głosowego. Dynamiczny rozwój tej technologii sprawił, że obecnie wystarczy **jedno zdjęcie** i dowolna ścieżka głosowa, aby wygenerować realistyczne wideo, w którym osoba na nagraniu wygląda, jakby naturalnie wypowiadała dane słowa, zachowując przy tym ekspresję i wiarygodność wizualną, co znacząco zwiększa potencjał nadużyć.

AK: Czy stworzenie realistycznego materiału deepfake jest trudne? Jak długo trwa proces stworzenia realistycznego deepfake'a i jakie wymagania sprzętowe są potrzebne?

EBT: Generowanie filmów i obrazów z wykorzystaniem wizerunku osób staje się coraz łatwiejsze i bardziej dostępne. Obecnie technologia deepfake, która pozwala na tworzenie realistycznych obrazów i filmów z użyciem twarzy czy głosu znanych osób, rozwija się w szybkim tempie. Już teraz istnieją narzędzia, które umożliwiają generowanie takich treści bez potrzeby posiadania zaawansowanej wiedzy technicznej, ani ogromnych zasobów obliczeniowych. Czasami wystarczy jedno zdjęcie, kilka sekund nagrania głosowego aby stworzyć wysoce realistyczny materiał syntetyczny.

Podobnie jak w przypadku generowania tekstów przy użyciu wielkich modeli językowych (LLM), przewiduje się, że w najbliższej przyszłości tworzenie filmów i obrazów z użyciem wizerunku znanych osób będzie równie proste. Oczywiście to może prowadzić do wielu nowych możliwości w dziedzinie rozrywki, reklamy czy edukacji, ale jednocześnie niesie ze sobą jeszcze większe ryzyko, szczególnie w kontekście dezinformacji, naruszeń prywatności i oszustw.

AK: Jak użytkownicy mogą zabezpieczyć swoje dane i wizerunek przed potencjalnym wykorzystaniem ich do stworzenia deepfake'a?

EBT: Każdy z nas pozostawia mnóstwo danych i śladów w sieci, dlatego ochrona wizerunku i danych osobowych jest niezwykle istotna. Deepfake'i, które wykorzystują technologię sztucznej inteligencji do tworzenia realistycznych, fałszywych materiałów, są coraz łatwiej dostępne, a to oznacza, że musimy być świadomi zagrożeń i podejmować działania prewencyjne. Zwracajmy uwagę na to, co udostępniamy w Internecie. Publikowanie licznych zdjęć i filmów – zwłaszcza takich, które pokazują nas w różnych pozach, oświetleniu czy kontekstach – ułatwia przestępcom stworzenie deepfake'a.

Ważne jest również, aby ograniczyć publiczną dostępność swojego wizerunku i korzystać z ustawień prywatności na platformach społecznościowych. Tylko zaufani znajomi powinni mieć dostęp do naszych materiałów i należy kontrolować swoją sieć kontaktów.
Warto także unikać udostępniania innych informacji osobistych, takich jak data urodzenia, miejsce zamieszkania czy numery kontaktowe, które mogą posłużyć do identyfikacji. Przed włamaniem na konto, a tym samym przed wyciekiem danych i kradzieżą tożsamości zabezpieczyć może silne hasła i włączenie uwierzytelniania dwuskładnikowego. Istotną rolę odgrywa także świadomość cyberzagrożeń. Metody klonowania dźwięku, które imitują głos, mogą być używane w oszustwach telefonicznych lub phishingu, dlatego należy zachować ostrożność przy udostępnianiu nagrań swojego głosu, zwłaszcza w miejscach publicznych w sieci. Warto także być czujnym wobec podejrzanych linków w e-mailach czy SMS-ach, które mogą prowadzić do prób wyłudzenia danych.

Technologie zabezpieczające mogą być naszym sprzymierzeńcem. Na przykład dodanie znaków wodnych do publikowanych materiałów wideo utrudni przestępcom wykorzystanie naszych danych. Jednocześnie, edukacja w zakresie rozpoznawania deepfake'ów i wiedza o tym, jak działają, pomogą nam zachować czujność wobec potencjalnych zagrożeń.

Warto również regularnie monitorować swój wizerunek w Internecie. Sprawdzanie, czy nasze zdjęcia lub filmy nie zostały wykorzystane w nieautoryzowany sposób, pozwoli na szybkie reakcje. Jeśli zauważymy naruszenie, powinniśmy zgłosić je odpowiedniej platformie internetowej lub organom ścigania i powiadomić znajomych i rodzinę o incydencie.

AK: Jakie są największe zagrożenia związane z wykorzystaniem deepfake'ów w cyberprzestępczości?

EBT: Deepfake'i to technologia, która niesie ogromny potencjał, ale też wyjątkowo poważne zagrożenia, szczególnie w kontekście cyberprzestępczości. Największe niebezpieczeństwa związane z ich wykorzystaniem można podzielić na kilka kluczowych obszarów. Coraz częściej są wykorzystywane do dyskredytacji osób, szczególnie poprzez generowanie kompromitujących treści, w szczególności mają charakter pornograficzny, przez co mogą całkowicie zniszczyć reputację ofiar. Przykładem jest sytuacja w szkole w Westfield, gdzie zdjęcia uczennic z mediów społecznościowych zostały przerobione na nagie fotografie za pomocą technologii deepnude. To naruszenie prywatności prowadzi do zniesławienia i traumy, szczególnie gdy ofiary nie są świadome, że ich wizerunek został zmanipulowany. Wg niektórych badań, tego typu treści stanowią 96 - 98% wszystkich przypadków i w większości dotyczą manipulacji wizerunkiem kobiet. Podobne sytuacje mają miejsce też w Polsce.

Deepfake'i stały się potężnym narzędziem w rękach cyberprzestępców. Technologia ta jest wykorzystywana w phishingu i oszustwach, gdzie fałszywe głosy czy obrazy mają na celu wyłudzanie pieniędzy lub danych osobowych. Jednym z głośnych przypadków był oszust z USA, który podszył się pod porwaną córkę, aby wymusić okup od rodziców. Podobnie działa tzw. „oszustwo na BLIK-a" czy fałszywe telefony od „policjantów", "strażaków" – syntetyczne głosy bliskich osób lub „autorytetów" zwiększają wiarygodność ataków. Coraz częściej technologia ta jest też wykorzystywana do podszywania się pod prezesów czy pracowników firm w celu wyłudzania środków finansowych. Wiele fałszywych materiałów pojawia się w formie fałszywych reklam instrumentów inwestycyjnych. Tylko od maja do sierpnia br. znaleźliśmy ponad 7600 publikacji fałszywych materiałów audiowizualnych, w których wykorzystano ponad 170 znanych osób w Polsce.

Deepfake'i są również potężnym narzędziem do szerzenia dezinformacji. Mogą być używane do potwierdzania teorii spiskowych, destabilizowania sytuacji politycznej czy manipulowania opinią społeczną. Przykładem jest zmanipulowane nagranie z głosem Joe Bidena zniechęcającym do głosowania – fałszywy materiał mógł mieć realny wpływ na decyzje wyborców. W kampaniach politycznych pojawiają się również fałszywe obietnice czy wypowiedzi, które podważają wiarygodność kandydatów.

Deepfake'i są również wykorzystywane w celach zastraszania i manipulacji. W sieci pojawiają się „farmy trolli" używające fałszywych twarzy do budowania narracji, które wpływają na opinie społeczne. Fałszywe profile, wzbogacone realistycznymi zdjęciami i nagraniami wideo, mogą rozpowszechniać określone treści, podsycać konflikty społeczne czy szerzyć fałszywe informacje.

AK: Jakie branże (np. polityka, finanse, rozrywka) są najbardziej narażone na ryzyko związane z deepfake'ami?

EBT: Osoby najbardziej narażone na ryzyko związane z deepfake'ami to przede wszystkim osoby rozpoznawalne, których zdjęcia i filmy są udostępniane w Internecie. Przede wszystkim osoby związane z polityką, biznesem i rozrywką, jednak zagrożenie dotyczy również osób związanych z mediami, sektorem zdrowia i organizacji pomocowych.

Technologia deepfake jest powszechnie wykorzystywana w oszustwach finansowych na szeroką skalę. Oszuści manipulują wizerunkami znanych polityków, biznesmenów czy ekonomistów, aby zwiększyć wiarygodność swoich ofert. Materiały wideo lub audio przedstawiają fałszywe wypowiedzi tych osób, zachęcając do inwestowania w nieistniejące projekty obiecujące szybkie i wysokie zyski. Często takie treści są wzmacniane przez manipulację i montaż materiałów z serwisami informacyjnymi, gdzie pojawiają się dziennikarze i prezenterzy, co dodatkowo podnosi autentyczność przekazu.

Fałszywe reklamy zmanipulowane z wykorzystaniem technik sztucznej inteligencji, pojawiające się masowo w mediach społecznościowych, są kolejnym zagrożeniem. Oszuści tworzą materiały promujące różnorodne produkty, od aplikacji mobilnych po „cudowne" inwestycje. Wizerunki celebrytów, influencerów czy dziennikarzy często służą do przyciągania młodszych grup docelowych, którym przedstawia się wyidealizowany styl życia. Tego typu manipulacje są szczególnie skuteczne, ponieważ wykorzystują znane twarze i emocje odbiorców. W przypadku fałszywych reklam leków, zabiegów medycznych czy suplementów diety, oszuści często posługują się wizerunkami znanych lekarzy, sportowców, a nawet duchownych. Celem takich materiałów jest wzbudzenie zaufania wśród osób zmagających się z problemami zdrowotnymi. Obiecują one natychmiastowe i cudowne efekty leczenia, co często prowadzi do finansowych i emocjonalnych strat po stronie ofiar.

Deepfake'i są również wykorzystywane do tworzenia fałszywych materiałów promujących organizacje rzekomo pomagające ofiarom oszustw w odzyskaniu utraconych pieniędzy. Oszuści wykorzystują te materiały, aby jeszcze raz wyłudzić pieniądze, obiecując pomoc prawną lub finansową.

Analizy wskazują, że najczęściej wykorzystywane w deepfake'ach są wizerunki polityków – ich autorytet i zaufanie społeczne sprawiają, że skutecznie przyciągają uwagę ofiar, szczególnie w oszustwach finansowych. Na drugim miejscu są dziennikarze i prezenterzy, których fałszywe wypowiedzi są stylizowane na materiały informacyjne i często udostępniane na platformach

społecznościowych, takich jak YouTube. W takich materiałach prezenterzy „rekomendują” inwestycje lub zachęcają do pobrania aplikacji.

AK: Deepfaki niosą zarówno szanse, jak i zagrożenia, dlatego kluczowe jest podnoszenie świadomości na temat cyberzagrożeń oraz dbanie o swoje bezpieczeństwo i prywatność w dynamicznie zmieniającym się cyfrowym świecie. Dziękuję za rozmowę.

EBT: Dziękuję!