CTI Jelentés

Mekkora fenyegetést jelentenek valójában a deepfake-ek?

1996
Fortes fortuna adiuvat
NEMZETBIZTONSÁGI SZAKSZOLGÁLAT

NEMZETI
KIBERVÉDELMI INTÉZET

Bevezetés

Jelen tájékoztatónkban arra a kérdésre keressük a választ, hogy a deepfake-ként ismert hamis képek, videók és hangfelvételek jelentenek-e már most, 2022-ben valós fenyegetést.

A deepfake technológia 2022-re széles körben ismertté vált, valószínűleg már mindannyian találkoztunk ilyen típusú tartalommal legalább egy közösségi médiában megosztott vicces videó, vagy egy online hír kapcsán, amiben egy ismert személy valami meghökkentő dolgot mond, vagy tesz.

Tartalomjegyzék

A szintetikus (~~jövő~~) jelen

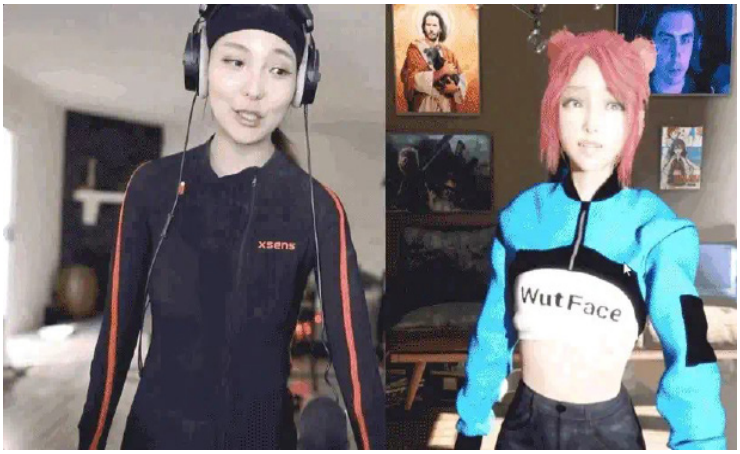
Manapság a deepfake mellett gyakran tűnik fel a szintetikus média kifejezés is, és gyakran a két fogalmat azonos értelemben használják, pedig nem pontosan ugyanarról van szó.

Szintetikus média minden olyan média tartalom — közösségi média poszt, hirdetés, blog bejegyzés, stb.—, aminek az elkészítéséhez mesterséges intelligencia (MI) technológiát alkalmaztak. Ez lehet videó, kép, hang, szöveg és ezek tetszőleges keveréke.

Már most, napjainkban rengeteg ilyen szintetikus tartalom készül, és egyre több olyan platform érhető el, amelynek segítségével bárki igénybe veheti a MI technológiát egyedi szintetikus média tartalmak készítéséhez — mint például a rosebud.ai.

Miko kiváló példa a szintetikus tartalmak sokféleségére „ő" ugyanis egy virtuális streamer (szellemesen: VTuber), amit fejlesztője a saját mozgásának rögzítésével tanít. Mintegy 700 000 követőjével igen népszerű a videó-streaming platform Twitch-en, ami egyszerre szemlélteti a szintetikus tartalmakban rejlő kreatív és üzleti lehetőségeket.

A szintetikus tartalomgyártás népszerűsége olyan ütemben nő, hogy szakértők úgy vélik, 2026-ra már az online elérhető tartalmak 90%-a MI technológiával kerül előállításra



Miko, a virtuális streamer (a kép bal oldalán) és fejlesztője

Mi az a DeepFake?

A deepfake a szintetikus média tartalmak egyfajta altípusának tekinthető, olyan MI technológiával készített, elsősorban képi vagy hangalapú tartalmakat értünk ezalatt, amelyek vagy valós személyeket ábrázolnak, amint olyat tesznek, vagy mondanak, ami a valóságban nem történt meg, vagy eleve teljesen fiktívek, azaz a megjelenített személyek egyáltalán nem is léteznek.

Azok számára, akik nem biztosak abban, hogy láttak-e már deepfake felvételt, ajánljuk megtekintésre azt a klasszikusnak is nevezhető 2019-es videót, amelyen Bill Hader amerikai komikus előbb Al Pacinóvá, majd Arnold Swarzeneggerré „alakul át".

Egy igazi deepfake klasszikus.

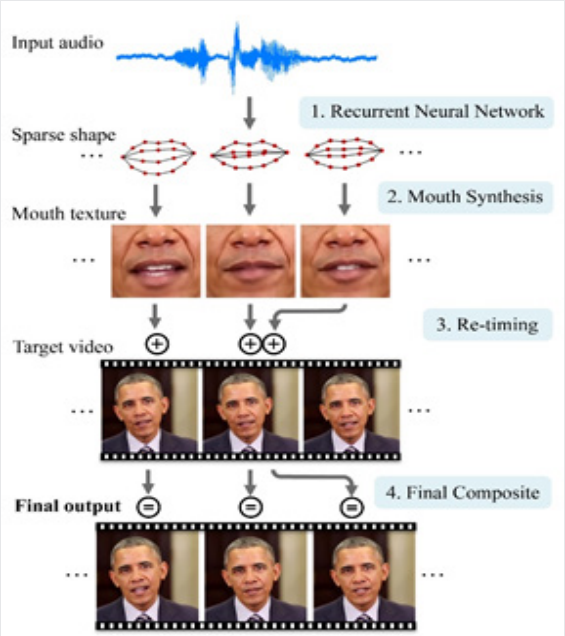
A videó jó példa arra, hogy ezek a tartalmak sok esetben nem károkozási céllal készülnek, a készítők egyszerűen szórakoztatni szeretnének bennünket.

A manipulált digitális tartalmak azonban alkalmasak a rosszindulatú felhasználásra is, vagyis arra, hogy általuk szándékosan megtévesszenek bennünket. Szolgálhatnak politikai célok támogatására is, sőt már arra is láthattunk valós példát, hogy ezeket egy katonai konfliktus (Ukrajna orosz megszállása) során alkalmazzák – utalva itt arra a hamis videóra, amiben Zelenszkij ukrán elnök „szólította" fegyverletételre az ukrán katonákat.

Habár ez, illetve a válaszul érkező, Vladimir Putyin orosz elnökről készített videó is könnyen felismerhetően hamisítvány, a mianmari Phyo Min Thein politikus korrupciós önvallomását bemutató videó kapcsán már több kétely merült fel.

Phyo Min Thein politikus korrupciós önvallomása.

A minamari katonai rezsim által bemutatott felvételen U Phyo Min Thein az ország Yangon régiójának korábbi minisztere több mint négy percen keresztül számol be korrupciós ügyeiről.



A lip sync technika

A deepfake elemző Deepware.ai platform eredménye szerint a videó 90%-os valószínűséggel hamisítvány, és ebben a helyi közvélemény nagy része is egyetért. A felvétel hitelességét elsősorban a politikus felvételen látható furcsa szájmozgása miatt kérdőjelezik meg, a feltételezés szerint a felvételnek ez a része, az ún. lip sync technikával került manipulálásra.

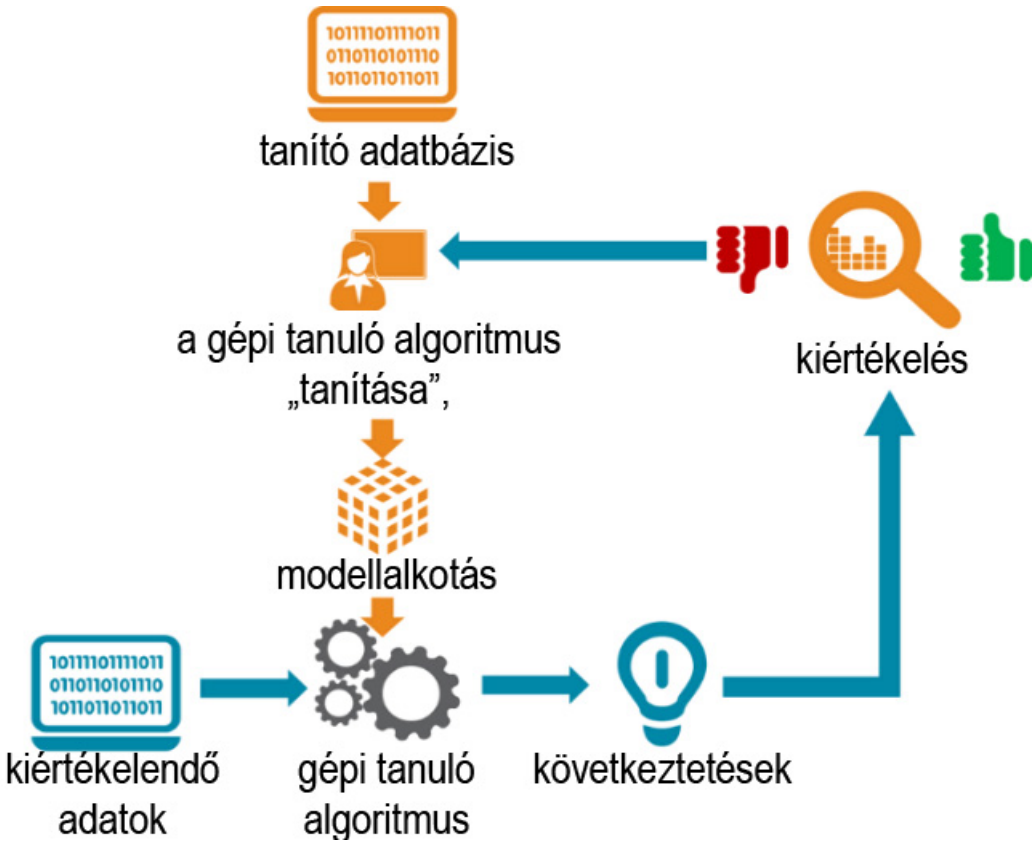
A modern dezinformációs technológiákkal foglalkozó és a Wittnes.org emberi jogi programigazgatója, Sam Gregory szerint a furcsa, nem természetesnek tűnő végeredményt a videó alacsony felbontása és tömörítése is okozhatja. Könnyen elképzelhető az is, hogy a felvétel szándékosan rossz minőségű, és éppen a képi manipuláció elrejtésére szolgál, a szakember azonban nem zárja ki annak lehetőségét sem, hogy a politikust vallomásra kényszerítették.

Ez az eset egy nagyon lényeges tényezőre világít rá: a deepfake-ek automatizált elemző szoftverekkel történő azonosításának korlátaira.

A deepfake-detektáló szoftverek alapvetően lépéshátrányban vannak a deepfake-eket készítő újabb algoritmusokkal és technikákkal szemben, egészen egyszerűen azért, mert ezek működését először elemezni kell ahhoz, hogy a detektáló szoftverek képesek legyenek felismerni a hamisításokat. Ez a vírus-antivírus szoftverekhez hasonló dinamikát eredményez, azaz időközönként megjelenhetnek olyan új technikák, amelyeket automatikus elemző szoftverek átmenetileg nem képesek azonosítani.

Egy-két „deepfake fogalom” tisztázása

A deepfake tartalmak készítése az MI kutatás egyik ága, a gépi tanulás (machine learning) segítségével történik. Ez a tudományterület nagyon leegyszerűsítve olyan algoritmusokkal foglalkozik, amelyek képesek adatokban emberi közreműködés nélkül mintázatokat keresni, majd ezek alapján a mintakereső algoritmust is módosítani, ezzel lényegében tanítva önmagukat.

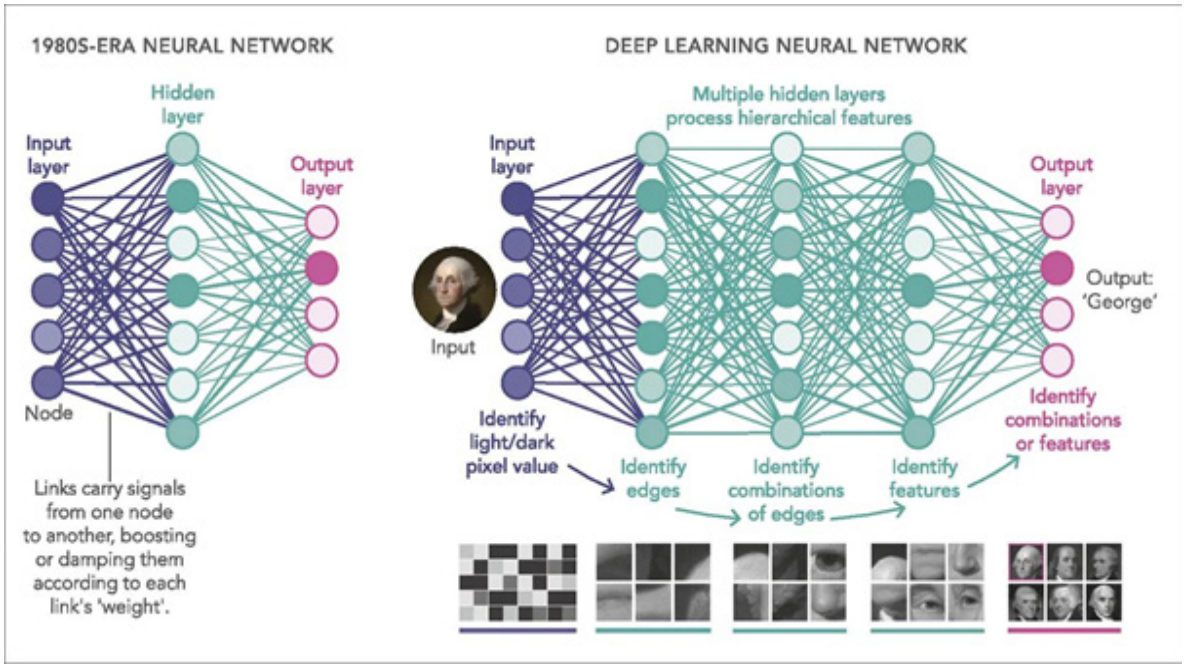


A gépi tanulás folyamatának logikája

A gépi tanulást már nagyon sok területen alkalmazzák a hétköznapok során például, a gyártóipari folyamatok hatékonyabbá tételére, önvezető autók „képzésére", vagy éppen SPAM szűrő szabályok létrehozására.

Összetett feladatok esetében – mint egy deepfake tartalom készítése – már speciális, több rétegből álló rendszerekre van szükség, ez az ún. mély gépi tanulás (deep learning).

Ezek a rendszerek jóval részletesebb elemzésre képesek, az adatok között többféle összefüggést képesek kezelni. A deep learning rendszereket gráfszerűen rétegekbe rendezve képzelhetjük el, ami kicsit hasonlít az emberi agy neurális hálózatához, emiatt ezeket mesterséges neurális hálózatoknak (Artificial Neural Networks) is nevezik.



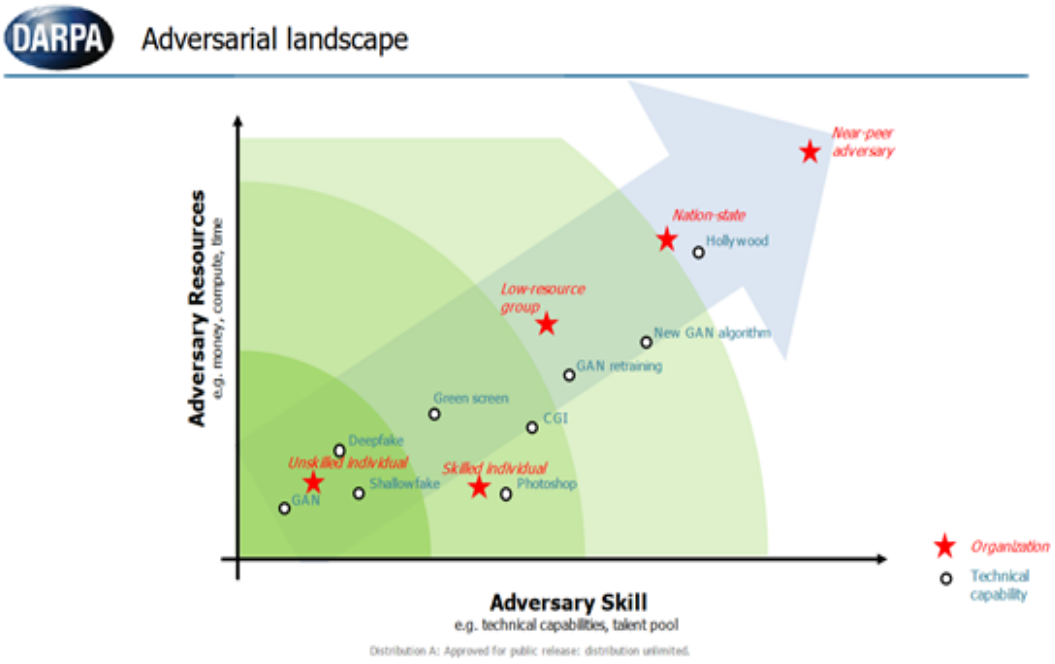
A deep learning rendszerek koncepciója

A deepfake-ek kapcsán létezik még egy fontos fogalom, amit érdemes megemlíteni, ugyanis kulcsszereppel bír a mai deepfake tartalmak előállításában, ez pedig a GAN (Generative Adversarial Network), amelynek koncepcióját Ian Goodfellow alkotta meg társaival 2014-ben.

Egy GAN rendszerben már két mélytanulási algoritmus „tanítja egymást" a minél pontosabb tartalmak generálására. Az első a generátor, amellyel a folyamat első lépéseként elemzik azt a média típust, amit végeredményként majd előállítani szeretnének, például egy képet. Ennek során a generátor karakterisztikákat, jellemzőket állapít meg, majd megpróbál az azonosított karakterisztikáknak megfelelő új tartalmakat előállítani.

Ekkor lép be a képbe a másik mélytanuló algoritmus, amit ebben a szerepben diszkriminátornak neveznek, amivel korábban szintén elvégezték az eredeti minták elemzését. Ennek feladata most az lesz, hogy összehasonlítsa a generátor által készített tartalmak jellemzőit az eredeti minták jellemzőivel. Ha talál olyat, ami a mintáktól eltérőnek ítél, akkor azt hibásként jelöli meg, amit azután visszaküld a generátornak, hogy ezt is vegye számításba új minták készítéséhez. Ez a folyamat egészen addig tart, amíg a diszkriminátor már nem talál számottevő hibát az új tartalmakban.

A DARPA „Medifor" (Media Forensics) programja 2015-ben feltérképezte az akkor elérhető technológiákat az előállításhoz szükséges erőforrások és technikai tudás szempontjából (lásd 4. ábra). A vizsgálat során az derült ki, hogy a GAN-ok már alacsony technikai felkészültséggel és erőforrással is működtethetőek, ami növeli a kiberbűnözői felhasználás valószínűségét.



GAN rendszerek felhasználási szintjei a szükséges erőforrások tekintetében

Deepfake fenyegetési helyzetkép

Pornográf tartalmak

A deepfake technológia felhasználása jelenleg elsődlegesen pornográf tartalmak előállítására irányul, amivel a készítők különböző nyílt és darkwebes oldalakon kereskednek.

A deepfake detektálásra specializálódott Sensity AI jelentése szerint a 2018 óta készült deepfake képek 90-95%-a ilyen hozzájárulás nélküli pornográf tartalom volt.

A cég csupán 2020 októberében több, mint 100 000 olyan hamis meztelen képet azonosított, amelyeken valós személyek arca szerepelt, anélkül, hogy az érintettek ehhez bármiféle hozzájárulásukat adták volna. Ezek a tartalmak rendkívül népszerűek, esetenként több száz milliós nézettséggel bírnak, egyes pornográf site-ok pedig kifejezetten celebritásokról készített deepfake-ekre specializálódtak.

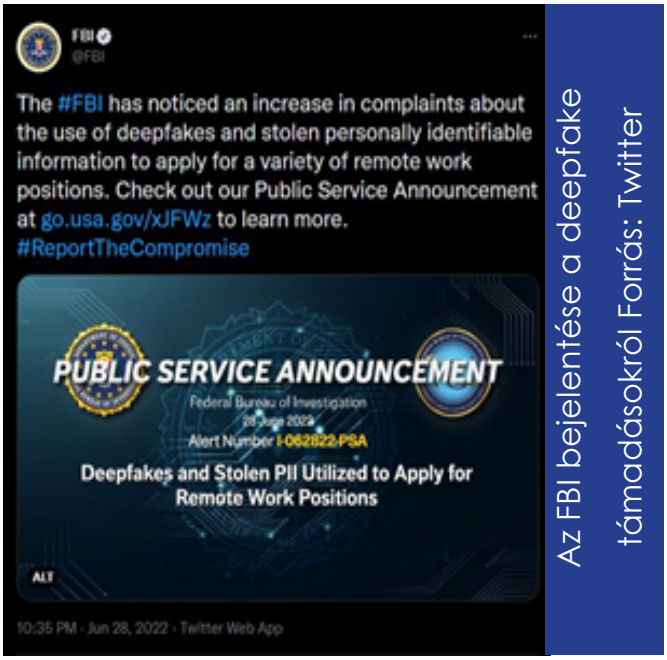
Mindez a leginkább ismert személyekre, celebekre nézve jelent fenyegetést, azonban gyakorlatilag bárki válhat deepfake technológiával készített „bosszúpornó" áldozatává.

Az ilyen típusú támadások a célszemély jóhírének az aláásására szolgálnak.

Videokonferencia platformokon elkövetett támadások

A deepfake emellett alkalmas lehet pszichológiai manipulációs támadások kivitelezésére is. Említsünk meg néhány közelmúltbeli incidenst, amelyek közös jellemzője, hogy videókonferencia platformokon keresztül zajlottak.

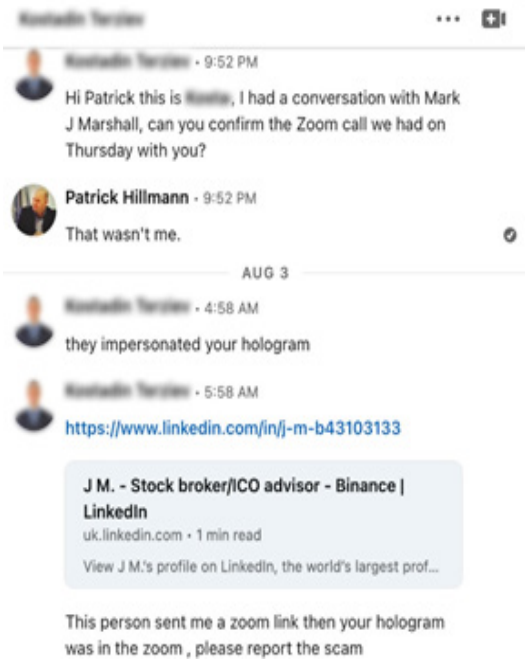
Az FBI internetes bűnözéssel foglakozó szervezeti egysége, az IC3 nemrég tájékoztatót is kiadott néhány esetről, amelynek során a kiberbűnözők deepfake technológia felhasználásával jelentkeztek távmunkára különböző amerikai cégekhez. A pozíciók többnyire IT-s munkakörök voltak, és különösen aggasztó, hogy olyan is szerepelt közöttük, amivel a támadók felhasználói személyes adatokhoz is hozzáférést szerezhettek volna. A jelentés fontos tanulsága, hogy a deepfake tartalmak némelyike még az online interjúk során is egészen meggyőző volt, szerencsére a jelentkezők háttérellenőrzése során kiderült, hogy a megadott személyes adatok hamisak más személyhez tartoznak.



Az FBI bejelentése a deepfake támadásokról Forrás: Twitter

Egy másik egészen friss [eset](): egy hacker csoport nemrég a Binance — a világ legnagyobb kriptovaluta tőzsdéjének — kommunikációs igazgatójáról, Patrick Hillmanről készített egy deepfake felvételt, amivel Binance ügyfeleket tévesztettek meg sikeresen.

A csalók a hamis felvételhez vélhetően Hillman korábbi TV-s szerepléseit használták fel, ezzel csalták privát beszélgetésbe az áldozatokat.



Volt, aki elővigyázatos volt, és jelentette az esetet Hillmann-nek

Valamint érdemes felidézni egy tavalyi esetet, hozzátéve, hogy ennél nem nyert bizonyítást, hogy valóban deepfake támadás történt-e.

A de Volkskrant holland napilap információi szerint holland parlamenti képviselők Zoom meetingjén részt vett egy olyan személy, aki deepfake technológiát felhasználva magát Alexei Navalny orosz ellenzéki aktivista kampányfőnökének, Leonyid Volkovnak [adta ki](). Ráadásul ez nem is egyetlen alkalommal fordult elő, ész, litván és brit politikusok is ebbe a csapdába estek. Az incidensre ráadásul csak akkor derült fény, amikor maga Volkov cáfolta meg, hogy részt vett volna ezeken az egyeztetéseken.

A [The Verge]() online hírportál információi szerint az eset hátterében két orosz személy állt, Vladimir Kuznetsov és Alexei Stolyarov, akik azt állították, hogy a megtévesztéshez nem alkalmaztak deepfake technológiát, csupán hagyományos eszközökkel maszkírozták el Sztoljarovot.

Ezek az esetek azt mutatják, hogy a virtuális térben már a mai technikai szinten sikeresen kivitelezhető deepfake támadás, álljon annak a hátterében financiális, vagy éppen politikai motiváció. Elmondható, hogy az üzleti vezetőket megszemélyesítő csalások reális fenyegetést jelentenek e tekintetben. A VTC platformokon mindez kiemelt problémát jelenthet, mivel a felvételek minősége a hálózati terheltség miatt alacsonyabb lehet, ami megnehezíti a résztvevők személyazonosságának azonosítását.

Deepfake technikákkal kiegészített dezinformációs kampányok

Az Europol deepfake fenyegetés témában készített jelentése a deepfake technológiákkal végzett dezinformációs kampányokat is lényeges fenyegetésként tartja számon. Ezekre a főként a közösségi oldalakon zajló dezinformációs kampányokra az elmúlt évekből számos példát találunk a tájékoztató elején említett incidenseken túlmenően.

Mi a különbség?

Félretájékoztatás: hamis információk nem szándékos létrehozása vagy terjesztése.

Dezinformáció: szándékosan, károkozás céljából létrehozott hamis információk.

Forrás: Európai Tanács

Több tanulmány is született a témában egy amerikai kormányzati és független szakértőkből álló csoport jelentéséből idézünk néhány esetet:

- 2019 óta állami hátterű (például Oroszországhoz és Kínához köthető) fenyegetési csoportok véleménybefolyásolási műveletek során GAN technológiával előállított képeket alkalmaztak közösségi médiaprofilok esetében.
- Egy másik példa a 2020-2021 között zajló, Belgium 5G-vel kapcsolatos álláspontjának befolyásolására, és a kínai 5G technológia előnyben részesítésére irányuló kampány.
- Említhető még a 2021-ben a FireEye által azonosított politikai deepfake kampány is, amely egyes libanoni politikai pártok támogatásának növelését célozta.

Az érintett platformok – mint például a Facebook – illetve kutatócégek, mint például a Graphika képesek voltak elemzőszoftverek segítségével detektálni egyes ilyen kampányokat, azonban a szakértői csoport véleménye szerint joggal feltételezhetjük, hogy ez csupán a jéghegy csúcsa.

Adott a kérdés, mennyire sikeresek ezek a kampányok? Látszólag nem tűnnek számottevőnek, önmagukban nem döntenek meg politikai rezsimeket, nem robbantanak ki forradalmakat. Valójában nem ilyen direkt módon, azonban mégis negatív hatást gyakorolhatnak.

Cristian Vaccari és Andrew Chadwick 2020-ban egy kutatás során azt találták, hogy a politikai témájú deepfake-ek habár nem szükségszerűen tévesztik meg az embereket, arra azonban alkalmasak, hogy bizonytalanságot keltsenek bennük, és ezáltal csökkentsék a hírforrások megbízhatóságába vetett bizalmat. Hosszabb távon mindez csökkentheti a közösségi oldalak használóinak felelősségvállalását az általuk megosztott információkkal kapcsolatban, valamint általánosságban az online tartalmak megbízhatóságába vetett bizalmat.

A deepfake-fenyegetés elleni védekezés nehézségei

A közösségi platformok szerepe elvitathatatlan a deepfake fenyegetés kezelésében, amire az Európai Unió is ráerősít, ugyanis egy kiszivárgott EU-s dokumentum szerint a tech óriások komoly büntetésekre számíthatnak, amennyiben nem lépnek fel kellő mértékben az deepfake videók, az álhírek és álprofilok ellen.

A Meta már 2020-ban bejelentette, hogy a paródiákon és szatírákon kívül tiltani fogja a MI-vel generált megtévesztő tartalmakat a Facebookon. A Metához hasonlóan a TikTok, a Reddit, a Youtube is hasonlóan nyilatkozott. A tech cégek ilyen témájú irányelveinek közös eleme, hogy azokat a tartalmakat kívánják szűrni, amik szándékosan megtévesztik a felhasználókat. Ugyanakkor a szándék megítélése sok esetben komoly kihívást jelenthet, így a gyakorlati megvalósítás jelen pillanatban sok kérdést vet fel.

A dezinformáció visszaszorítását célzó uniós gyakorlati kódex mellett a deepfake-ek terén leginkább releváns jogszabályi keretrendszer a mesterséges intelligenciára vonatkozó szabályozási keretjavaslat, amely azonban még nem került elfogadásra. A keretrendszer kockázatalapú megközelítést alkalmazna az MI-felhasználás szabályozására, például a deepfake-ekre vonatkozóan minimális követelményként előírná, hogy kötelező legyen megjelölni, hogy a tartalom megtévesztő.

Fontos leszögezni, hogy a deepfake-ek elleni hatékony fellépés nem háríthatja a teljes felelősséget a tech vállalatokra. Ugyanilyen fontos elem a magánszektor tudatosítása a deepfake-ek felismerésére és a felelős információmegosztás szerepére, illetve a hiteles fenyegetési információk elérhetősége.

A deepfake-fenyegetés elleni védekezés nehézségei

Az alábbi kép egy konkrét példa deepfake-re a thispersondoesnotexist.com című weboldalról. A kép alatt öt olyan nyom található, amelyek arra utalhatnak, hogy ez egy deepfake lehet. Láthatjuk, hogy ezeknek a nyomoknak a felismerése esetenként nem is olyan könnyű:

1. Háttér: A háttér sok esetben homályos vagy torz, és a kép fényessége sem egyenletes, például az árnyékok nem természetesek.

2. Üvegfelületek: Vizsgáljuk meg figyelmesen a szemüvegkeret és a homlok közötti átmenetet. A deepfake-ek egyes részei időnként nem illeszkednek egymáshoz, méret- vagy formabeli eltérések észlelhetők.

3. Szemek: A deepfake fotókra jelenleg jellemző az „üveges tekintet".

4. Ékszerek: A fülbevalók amorfak lehetnek vagy a rögzítésük furcsának tűnhet. A nyakláncok, mintha a bőrbe lennének ágyazva.

5. Nyak és vállak: A vállak alaktalanok vagy egyszerűen nem passzolnak. A nyak különbözőnek hat a két oldalon.

Videó

Az MIT (Massachusetts Institute of Technology) kutatói elkészítettek egy kérdéssort, amely segítségünkre lehet abban, hogy kideritsük egy videó valódi-e, kiemelve, hogy a deepfake-ek gyakran nem keltenek természetes hatást az ábrázolt helyszín, illetve a fényesség tekintetében.

1. Arc és homlok: Túl simának vagy túlráncosnak tűnik a bőr? Ugyanolyan idősnek tűnik a bőr, mint a haj, vagy a szemek?

2. Szem és a szemöldök: Az árnyékok ott jelennek meg, ahol számítunk rájuk?

3. Üvegfelületek: Van tükröződés? Vagy éppen túl sok a tükröződés? Változik-e a tükröződés szöge, amikor a személy mozog?

4. Arcszőrzet: Valódinak tűnik? A deepfake-ekben sok esetben hozzáadnak vagy elvehetnek bajuszt, pajeszt vagy szakállt.

5. Arcon lévő anyajegyek: Valódinak tűnnek?

6. Pislogás: A szereplő eleget pislog vagy éppen túl sokat?

7. Az ajkak mérete és színe: A méret és a szín megegyezik a személy arcának többi részével?

Hang

A kutatók szerint az olyan technológiák, mint a spektrogramok, képesek jelezni, ha a hangfelvételek hamisak. Azonban a legtöbbünknek nincs kéznél hangelemző, amikor egy támadó telefonál. Vegyük észre, ha túl monoton az előadást, a furcsa hangmagasságot vagy érzelmeket, valamint a háttérzaj hiányát. A hanghamisításokat nehéz felismerni.

1. Ha furcsa hívást kapunk egy szervezettől, ellenőrizhetjük, hogy a hívás valódi-e, ha először letesszük a kagylót, majd visszahívjuk a szervezetet.

2. Ügyeljünk arra, hogy megbízható telefonszámot használjunk, például olyat, amely szerepel a telefonos névjegyzékünkben, a szervezettől kapott számlán vagy a szervezet hivatalos webhelyén.

NEMZETI
KIBERVÉDELMI INTÉZET

nki.gov.hu

titkarsag@nki.gov.hu

+36 (1) 325 7672

Nemzeti Kibervédelmi Intézet

@ nki.gov.hu

Kibertámadás!
podcast

2022