MASARYKOVA UNIVERZITA BRNO

FAKULTA INFORMATIKY

PŘÍNOSY A RIZIKA INTELIGENTNÍCH SYSTÉMŮ

**Hrozba jménem deepfake**

Michal Krejčíř

UČO: 514187

Datum: 1.7.2022

Obsah

[Úvod 3](#_Toc107607950)

[Představení deepfakes 3](#_Toc107607951)

[Uvěřitelnost deepfakes 3](#_Toc107607952)

[Spektrum deepfake-cheapfake 4](#_Toc107607953)

[Přístupnost 4](#_Toc107607954)

[Typy deepfakes 4](#_Toc107607955)

[Tvorba deepfaku 5](#_Toc107607956)

[Autoenkodéry 6](#_Toc107607957)

[GANs (Generative Adversarial Networks) 7](#_Toc107607958)

[Legitimní využití deepfakes 7](#_Toc107607959)

[Zábava 7](#_Toc107607960)

[Audiovizuální a herní průmysl 7](#_Toc107607961)

[Deepfakes jako zbraň 8](#_Toc107607962)

[Deepfake pornografie 8](#_Toc107607963)

[Šíření dezinformací a hybridní válka 8](#_Toc107607964)

[Detekce deepfakes 9](#_Toc107607965)

[Závěr 9](#_Toc107607966)

[Seznam zdrojů 11](#_Toc107607967)

# Úvod

Rychlým vývoj výkonu počítačů a jejich narůstající integrace do každodenního života běžného člověka a do společnosti obecně přináší řadu nesporných výhod. Jednou z nejvýraznějších je růst HDP zemí, kde digitalizace probíhá. Podle odborníků HDP států může vzrůst až o 0.62% za každý 10% nárůst digitalizace státu (s tím, že státy, kde je již digitalizace na pokročilé úrovni, zaznamenávají výraznější nárůst HDP). Stejný nárůst digitalizace má také potenciál snížit nezaměstnanost ve státě o 0.84% a podílí se na výrazném zvýšení životní úrovně státu (1, s. 125-127).

Digitalizace s sebou ale přináší i řadu nevýhod. Jednou z nejvýznamnějších je velmi snadné šíření nebezpečných „Fake news“ a tzv. dezinformací. Dezinformace jsou ve zkratce nepravdivé informace, které jsou šířeny za účelem vydělání peněz, diskreditace osob, ovlivňování myšlení veřejnosti a polarizace a destabilizace společnosti. Nejedná se o žádný nový fenomén, počátky dezinformací můžeme pozorovat již v 16. století. Nástup digitalizace však šíření dezinformací výrazně usnadnil, a to zejména rozvojem nových komunikačních kanálů, jako jsou sociální média, e-mail a dezinformační internetové zpravodajské servery (2, s. 8-11).

Jedním z nejnovějších a potenciálně velmi nebezpečným způsobem šíření dezinformací jsou deepfakes, které v poslední době společně s rychlým rozvojem umělé inteligence a strojového učení zažívají rychlý růst, a to jak co se týče kvantity, tak kvality (2, s. 4-6). Na první pohled se deepfakes mohou zdát jako velmi nebezpečné, je tomu ale skutečně tak? Existuje proti nim nějaká efektivní obrana? Na tyto otázky se v následujících odstavcích pokusíme odpovědět.

# Představení deepfakes

Ve své obecné podstatě je deepfake vizuální materiál (foto nebo video), ve kterém byla podoba jedné osoby nahrazena podobou osoby druhé za pomoci strojového učení a AI. Může to vypadat například tak, že osoba A natočí video, kde sděluje určitou informaci. Následně je obličej osoby A automatizovaně nahrazen obličejem osoby B, čímž vzniká deepfake (3, s. 135-136).

Zaštiťující kategorií, pod kterou deepfakes spadají, jsou syntetická média. Jedná se o takový obsah, který byl automatizovaně vytvořen či modifikován pomocí technologií - zejména AI nebo strojového učení. Krom deepfakes patří mezi syntetická média i tzv. cheapfakes (4, s. 81-84), které si přiblížíme v jedné z následujících podkapitol.

## Uvěřitelnost deepfakes

Uvěřitelnost deepfakes je jejich nejdůležitější vlastností (pokud se nejedná o deepfaky vytvářené čistě pro pobavení). Pokud cílová skupina deepfaku nebude věřit, nemůže být ovlivněno její myšlení. Obecně platí, že kvalita deepfaku je úměrná kvalitě vstupních materiálů a strojovému času, který je tvorbě deepfaku věnován (5, s. 3243-3244).

Kvalita takto vytvořeného deepfaku může být různá – od špatně vytvořených deepfakes, které každý na první pohled odhalí, až po profesionální práce, jejichž odhalení lidským okem je velmi obtížně (a občas dokonce prakticky nemožné) (5, s. 3242-3245). Jako příklad velice kvalitně provedeného deepfaku lze zmínit deepfake prezidenta ČR Miloše Zemana, který vznikl jako propagační video pro seriál televize HBO Bez vědomí. Jako podklad pro deepfake posloužil projev českého herce Jana Vlasáka (místo jehož obličej byl následně pomocí AI vložen obličej Miloše Zemana), hlas propůjčil český imitátor a bavič Petr Jablonský.[[1]](#footnote-1)

Jednou z věcí, co deepfake videa často prozradí, je hlas – to platí i pro výše zmíněné video. Ovšem stejně snadno, jako lze vygenerovat falešné video, lze vygenerovat i falešné audio na základě hlasu člověka, a to pomocí techniky s názvem neural voice cloning (6, s. 1-2). Aby umělá inteligence mohla naklonovat hlas určitého člověka, stačí ji k tomu i pár sekund záznamu, jak dokazuje např. projekt Tacotron společnosti Google.[[2]](#footnote-2)

Uvěřitelnost těchto „kombinovaných“ deepfakes může být na úrovni podstatně vyšší, než u deepfakes využívajících pouze editaci videa. Dobrým příkladem takového deepfaku je video s názvem „Better Call Trump: Money Laundering 101 [DeepFake]“, kde je obličej i hlas herce nahrazen podobou Donalda Trumpa, včetně uměle syntetizovaného hlasu.[[3]](#footnote-3)

## Spektrum deepfake-cheapfake

Jak jsme již zmínili, syntetická média zahrnují vedle deepfakes i cheapfakes. Ve skutečnosti tyto dva typy vytvářejí spojité spektrum parametrizované technickou a vědomostní náročností, která je vyžadována k tvorbě média. Jak už vyplývá z jejich názvu, cheapfakes jsou podstatně jednodušší na produkci a vyžadují menší expertízu, obecně jsou však úměrně méně uvěřitelné. Mezi techniky využívané pro produkci cheapfakes řadíme např. rotoskopii nebo rekontextualizaci. Cheapfakes lze vytvářet pro všechny běžně dostupnými programy, které jsou snadné k použití a často i zdarma (7, s. 9-12).

## Přístupnost

Po dlouhou dobu platilo, že tvorba deepfakes byla pro běžné uživatele nedostupná. Sice existovaly programy, které umožňovaly i běžnému uživateli relativně snadno vytvořit deepfake, ty ale nedosahovaly přílišné uvěřitelnosti a byly podstatně složitější na tvorbu, než cheapfakes srovnatelné uvěřitelnosti (7, s. 14). Tento trend ale v posledních letech přestává platit. V dnešní době jsou již dostupné nástroje, které komukoliv a velmi snadno umožní vytvořit deepfake vysoké kvality. Jako příklad lze zmínit čínskou mobilní aplikaci Zao vydanou roku 2019. Pouze pomocí několika fotografií obličeje uživatel může umístit svou tvář do libovolné scény z různých filmů a seriálů z nabídky (3, s. 136-137).

Stejný vývoj existuje i ve světě audio deepfakes. Společnost DeepMind uvedla na trh program WaveNet, který uživateli umožňuje převést psaný text na mluvené slovo (tzv. „text-to-speech“) s hlasem, který uživatel požaduje (8, s. 1).

Obecně platí, že v dnešní době se volně dostupné a relativně snadno použitelné programy pro tvorbu deepfakes kvalitou vytvořených deepfakes začínají vyrovnávat i velmi drahým a profesionálním softwarovým řešením používaným ve filmovém průmyslu (7, s. 34-35).

## Typy deepfakes

Pod souhrnné označení deepfake lze zahrnout různé formy manipulace s audio-vizuálním obsahem. Mezi základní typy čistě audio deepfakes patří Voice-swapping a Text-to-Speech. U techniky Voice-swapping osoba mění svůj hlas, aby zněl jako hlas jiné cílové osoby – slova a jejich intonace tedy zůstávají. Tato technika již byla úspěšně využita k podvodům. Druhou technikou je již zmíněná Text-to-Speech (také nazývána jako syntéza hlasu, protože hlas je syntetizován „z ničeho“), kdy je psaný text převeden na slovo mluvené hlasem cílové osoby. Tato technika například umožňuje vzít již existující nahrávku a změnit v ní slova mluvená danou osobou (3, s. 142).

Existují také různé techniky pro tvorbu vizuálních deepfakes. Pravděpodobně nejtypičtější metodou v současnosti je Face-swapping, kdy je v daném video materiálů nahrazen obličej jedné osoby obličejem jiné osoby (3, s. 142). Příkladem použití této techniky je jedo z nejslavnějších deepfake videí vůbec, „Jim Carrey Glows“. V něm je obličej herečky Alison Brie nahrazen obličejem Jima Carreyho, hlas je však ponechán původní (protože toto video má čistě zábavný charakter).[[4]](#footnote-4)

Obdobnou technikou je Face-morphing, kdy obličej plynule přechází mezi obličejem jednoé a druhé osoby. O krok dál jde technika zvaná Full-body puppetry, kdy jsou místo pohybů obličeje nahrazeny pohyby celého těla. Typickým použitím je nahrání pohybů těla jedné osoby a jejich následná transpozice do pohybů těla druhé osoby (3, s. 142).

Techniky pro tvorbu audio a video deepfakes lze zkombinovat a vytvořit tak velmi přesvědčivý fake. Jednou z takových technik je Lip-syncing (také tzv. „Dialog replacement“), kdy se změní jak pohyby úst, tak slova pronášená řečníkem. Obličej však zůstává zachován – umožňuje tedy efektivně změnit obsah zprávy, kterou se daná osoba snažila předat. Tento typ deepfaku může být obzvláště nebezpečný, protože může mít velmi vysokou uvěřitelnost – změny pouze úst je těžší si všimnout, než změny celého obličeje (3, s. 142). Skvělým příkladem tohoto typu deepfaku je video Marka Zuckerberga (zakladatel společnosti Facebook/Meta)[[5]](#footnote-5). Původní video svědecké výpovědi Marka Zuckerberga bylo změněno na řeč, ve které oslavoval zločinnou společnost Spectre (7, s.5), která mu umožnila hromadit a krást osobní data miliard lidí a tím efektivně kontrolovat jejich budoucnost (9).[[6]](#footnote-6) Video, které vzniklo roku 2019 za použití nástroje VDR od společnosti Canny (7, s.5), je podle názoru autora na špičkové úrovni a je velmi obtížné jej lidským okem odhalit.

# Tvorba deepfaku

Základem pro tvorbu deepfaku je využití technik z pole umělé inteligence (AI) a strojového učení (ML). Konkrétně jsou zde využity DNNs, neboli hluboké neuronové sítě. Tyto sítě využívají tzv. hlubokého učení, které umožňuje DNN naučit se efektivně reprezentovat složitá data (například obličeje, jejich rysy a výrazy) s pomocí několika úrovní abstrakce. Díky tomu, že hluboké učení vyžaduje pouze velmi málo ručního programování (většinu věcí se totiž model naučí sám), jsou DNNs velmi mocné modely (10, s. 436).

DNN je podtypem umělé neuronové sítě. Skládá se tedy z vzájemně propojených umělých neuronů, ovšem oproti běžným neuronovým sítím jsou tyto neurony uspořádány do mnoha vrstev. Platí, že výstup jedné vrstvy je vstupem vrstvy bezprostředně za ní následující. Vstupy do každého neuronu mají různé váhy, které určují, jaký vliv bude mít aktivace tohoto vstupu na příslušný neuron (alternativně se na tyto váhy lze dívat jako na „sílu“ spojení mezi příslušnými dvěma neurony). Neurony v každé vrstvě sice provádí pouze elementární operace, ovšem složením všech vrstev je možné dosáhnout velmi komplexního chování – konkrétně nelineárního mapování objektu ze vstupní domény DNN na objekt výstupní domény (například obličej osoby s určitým výrazem na korektní reprezentaci tohoto výrazu v množině všech možných obličejových výrazů) (11, s. 1-2).

K tomu, aby DNN toto mapování prováděla správným způsobem, je však třeba ji tzv. vytrénovat – stejně, jako se lidský mozek musí naučit např. mapování slov na jejich významy. Nenatrénovaná neuronová síť je nepoužitelná – její výstup je čistě náhodný. Pokud opět vztáhneme tuto problematiku na problém rozpoznávání výrazu obličeje, pouze správně natrénovaná neuronová síť bude schopna přiřadit obličeji jeho správný výraz. Trénování sítě modifikuje výše zmíněné váhy vstupů neuronů tak, aby síť fungovala podle požadavků jejího uživatele. Obecně platí, že k natrénování sítě je zapotřebí velké množství vstupních dat (3, s. 138-139).

Ačkoliv zde existuje mnoho nástrojů založených na DNN využívaných k tvorbě deepfakes, v následujících odstavcích se zaměříme na ty v současnosti nejpoužívanější, a to Autoenkodéry a GAN.

## Autoenkodéry

Tato metoda byla první, která se k vytváření deepfakes používala – a to konkrétně k technice Face-swappingu. Obecně je autoenkodér neuronovou sítí, která je natrénovaná tak, aby na výstupu zrekonstruovala svůj vstup. Tedy pokud autoenkodér na vstup dostane např. obrázek čísla 6, jeho úkolem je na výstupu vrátit také obrázek čísla 6. Tento postup však neprobíhá přímo (to by bylo velmi triviální), autoenkodér totiž nejprve enkóduje data na vstupu do vnitřní zkomprimované reprezentace a následně z této reprezentace data zpětně rekonstruuje (dekóduje) (12, s. 1-2).

Tomu odpovídají 3 základní komponenty autoenkodéru, a to enkodér, latentní prostor a dekodér. V běžných případech je enkodér i dekodér neuronová síť, často hluboká – pak mluvíme o hlubokém enkodéru a dekodéru (12, s. 1-2). Pokud si představíme autoenkodér určený k tvorbě deepfakes, který je trénován na lidských obličejích, pak jednotlivé fáze probíhají následovně: enkodér provede kompresi vstupního obrázku s obličejem na jednotlivé důležité vlastnosti obličeje – zda jsou oči otevřené nebo zavřené, jaká je mimika obličeje apod. Latentní prostor je právě prostor bodů, které zahrnují všechny možné kombinace všech sledovaných atributů obličeje. Správně natrénovaný enkodér tedy zakóduje daný obrázek obličeje na vstupu do právě toho bodu v latentním prostoru, který splňuje všechny atributy obrázku. Toto umožňuje nekonečnou množinu všech možných obličejů na vstupu zmenšit na malou, konečnou množinu kombinací vlastností. Je také vhodné podotknout, že enkodér a latentní prostor přísluší pouze jednomu člověku – tedy enkodér natrénovaný na osobě A nebude schopen korektně komprimovat obrázek obličeje osoby B (3, s. 139-140).

Poslední komponenta autoenkodéru, tedy dekodér, poté vezme příslušný bod v latentním prostoru (tedy reprezentaci obličeje) a na jeho základě na výstupu vytvoří obrázek obličeje příslušné osoby, na kterou je dekodér natrénován. Autoenkodér tak, jak jsme jej popsali, tedy umí vytvořit reálně vypadající kopii obličeje na vstupu, pokud patří osobě, na kterou je natrénován – toto sice není příliš užitečné pro tvorbu deepfakes technikou Face-swappingu, ale autoenkodér může být poměrně snadno upraven tak, aby byl použitelný (3, s. 139-140).

Jedinou změnou je využití sdíleného enkodéru pro kompresi obrázků obličejů obou osob – a to proto, aby stejné atributy obličeje byly vnímány se stejnou důležitostí, tedy aby obličeje osoby A i B se stejným výrazem byly reprezentovány stejným (nebo alespoň velmi blízkým) bodem v jejich latentních prostorech (3, s. 140-141). Latentní prostor zůstává pro každou osobu oddělený a dekodéry jsou natrénovány stejně, jako v běžném autoenkodéru (5, s. 3243-3244).

Nakonec, pokud chceme zaměnit obličej osoby A za obličej osoby B, sestavíme autoenkodér následovně: sdílený enkodér, latentní prostor osoby A, dekodér osoby B. Pokud takovému autoenkodéru dáme na vstup obrázek obličeje osoby A, převede se na bodovou reprezentaci v latentním prostoru – přičemž tento bod zároveň, díky sdílenému enkodéru, reprezentuje obličej osoby B s velmi blízkými až totožnými atributy. Následně je tento bod dekódován dekodérem osoby B a získáváme hledaný obrázek. Nyní pouze stačí tento postup opakovat pro každý snímek video záznamu a deepfake je na světě (5, s. 3243-3244).

## GANs (Generative Adversarial Networks)

GAN je v současnosti nejpoužívanější způsob výroby deepfakes. Ve své podstatě se jedná o dvě neuronové sítě – generátor a diskriminátor (13, s. 486). Obě tyto sítě spolupracují a vzájemně se snaží zdokonalit. Úkolem generátoru je generovat realisticky vypadající deepfake fotky či videa. Tento výstup generátoru je následně předán diskriminátoru. Důležitým detailem je to, že diskriminátor na svůj vstup kromě výstupu generátoru dostává i reálná média. Úkolem diskriminátoru tedy je od sebe rozlišit falešná a reálná média. Na rozdíl od toho, generátor přístup k reálným médiím nemá (14, s.53).

Pokud se diskriminátoru média od sebe povede úspěšně rozlišit, je tato zpráva předána generátoru spolu s informací, co bylo na jím generovaných médiích špatné – díky tomu generátor ví, co bylo na jím generovaných médií špatně a je trénován k produkci kvalitnějších médií. Trénování v tomto smyslu znamená opět změnu parametrů modelu, zejména vah ve spojení neuronů. Pokud však diskriminátor od sebe nedovede rozlišit falešná a reálná média, je naopak chybová zpráva předána jemu, tím pádem diskriminátor naučí lépe rozpoznávat falešná média. Mohlo by se zdát, že trénování diskriminátoru je nadbytečné, není tomu však tak. Čím lépe umí diskriminátor rozpoznávat falešná média, tím lepší musí generátor být, aby jej oklamal – platí tedy, že čím lépe je natrénován diskriminátor, tím lépe může být natrénován generátor (5, s. 3245). Platí, že jakmile je diskriminátor i přes kontinuální trénování schopen rozlišit falešná média s pravděpodobností pouze o málo přesahující 50% (je tedy prakticky stejně efektivní, jako náhodná volba), je trénování generátoru u konce - již nelze dosáhnout lepšího natrénování, protože chybí konstruktivní zpětná vazba od disktriminátoru. Takto natrénovaný generátor je následně využit k tvorbě velmi realisticky vypadajících deepfake médií (13, s. 486).

# Legitimní využití deepfakes

Ačkoliv se laikovy může zdát, že deepfakes jsou čistě nebezpečnou technologií, opak je pravdou. Deepfakes mají řadu legitimních využití, z nichž některé si nyní představíme.

## Zábava

Toto využití deepfakes můžeme zejména na internetu a sociálních sítích pozorovat již dlouhou dobu. Hlavním znakem takových deepfakes je, že nemají za cíl jakkoli škodit, ale pouze pobavit, a také že cílové publikum si je vědomo, že se jedná o deepfakes. Jako příklad můžeme uvést výše zmiňovaná videa, tedy „Jim Carrey Glows“ a „Better Call Trump: Money Laundering 101 [DeepFake]“, nebo například o trend z roku 2018, kdy se nadšenci snažili pomocí aplikace FakeApp, která realizuje techniku Face-swappingu, dostat Nicolase Cage do každého filmu či seriálu, který jim přišel pod ruku (15; 3, s. 141).

Toto využití se také týká veřejně dostupných aplikací a webů, kterých je v současnosti nepřeberné množství. Uživatelé tak mohou například vložit svůj obličej do slavných filmů, vytvářet umělecky vypadající obrázky na základě fotografií (aplikace Deep Art), vzájemně si vyměnit obličeje s kamarády (aplikace Face Swap Live) či nechat si vygenerovat obrázky na základě textového popisu (DALL-E Mini) (16).

## Audiovizuální a herní průmysl

Audiovizuální průmysl v současnosti hojně využívá počítačové technologie, ať už se jedná o speciální filmové efekty či rovnou celé filmy vytvořené počítačem. Deepfakes zde mají velký potenciál v oblasti dabingu – v současnosti totiž dabing do jiného jazyku, než ve kterém bylo dílo vytvořeno, mění pouze audio stopu díla. To způsobuje, že pohyby úst postav v dílu se nekryjí s předabovanou audio stopou. Při využití deepfake technologie Lip-syncing je však možnost pohyby úst postavy dokonale přizpůsobit audio stopě v novém jazyce. Stejné využití je možné vidět i v herním průmyslu, kdy jsou pohyby úst herní postavy sladěny s hlasem dabéra (5, s. 3246).

# Deepfakes jako zbraň

Deepfakes jsou velmi mocnou technologií a jejich zneužití je tedy prakticky nevyhnutelné. I když je možné deepfakes využívat neškodně, v současnosti bohužel převažuje jejich zneužívání (3, s.143). Mezi nejpalčivější zneužití patří deepfake pornografie a šíření dezinformací (5, s. 3246-3247).

## Deepfake pornografie

Jedno z nejviditelnějších současných zneužití deepfakes se týká pornografie. Rok 2017 byl prvním, kdy se na sociální síti reddit (a na internetu obecně) začaly deepfakes ve veřejném prostoru objevovat. Na subreddit r/deepfakes (subreddit je fórum zaměřené na konkrétní věc, například na počítačové hry) (17) uživatelé nahrávali videa, ve kterých tváře známých hereček, například Kristin Bell nebo Scarlett Johansson, vložili pomocí Face-swapping aplikací do pornografických videí, samozřejmě bez jakéhokoliv souhlasu hereček (18, s. 10). Toto zneužití deepfakes je velmi zneklidňující, neboť obrana proti nim je poměrně složitá (19, s. 77-81) a důsledky na psychické zdraví dotčených osob mohou být velmi vážné. Reddit sice tento subreddit i jakékoliv sdílení tzv. „nekonsenzuální pornografie“ v zápětí zakázal (20), pomyslná Pandořina skříňka již však nemůže být uzavřena.

Jedním z důvodů, proč se cílem těchto útoků staly celebrity, je ten, že lze dohledat mnoho fotografií či videí s jejich tvářemi, podle kterých lze natrénovat příslušné neuronové sítě – ty totiž k tvorbě přesvědčivých deepfakes potřebují tisíce snímků obličeje (3, s. 139). Ovšem se zlepšováním technologií se počet potřebných snímků zmenšuje a riziko pro každého z nás, že se může objevit v podobně choulostivém videu, roste (3, s. 142). Značné riziko plyne z tzv. „revenge porn“, což je podtyp nekonsenzuální pornografie, kdy se například po rozchodu jeden z bývalých partnerů rozhodne, že se tomu druhému pomstí zveřejněním intimních snímků toho druhého. Deepfakes přidávají této pomstě nový rozměr, neboť již ani není potřeba skutečně mít nějaké intimní foto či video dotčené osoby, protože vše je možné synteticky a poměrně snadno vytvořit. Psychické následky tohoto zneužití mohou být značné a vést až k sebevraždě. Nelze také opomenout, že cílem takových deepfakes mohou být i politici či političky za účelem jejich diskreditace (19, s. 54-77).

## Šíření dezinformací a hybridní válka

Deepfakes je možno bez nadsázky označit za dokonalý způsob šíření politicky orientovaných dezinformací, zejména díky své vysoké míře uvěřitelnosti. Pokud divák slyší příslušnou dezinformaci přímo z úst příslušného politika v reálně vypadajícím videu, může být velmi snadno oklamán. Toto v kombinaci se stále rostoucí mírou šíření dezinformací na sociálních médií vytváří velmi nebezpečnou kombinaci a potenciálně umožňuje dezinformacím mít větší dopad než kdy dřív (21, s. 1-2). Tyto deepfake dezinformace však mají negativní efekt, i když se jimi divák oklamat nenechá. Způsobují totiž podstatné narušení důvěry diváka v příslušná média a zvyšují úsilí nutné k tomu, aby divák nalezl pravdivé informace, čímž zvyšují jeho nejistotu (22, s. 3).

Příkladů využívání dezinformací ve snaze poškodit konkrétní stát pomocí existuje mnoho. V současnosti se jedná zejména o pokusy ovlivnit důležité volby či zvýšit názorové pnutí v zemi, například dezinformační kampaň Ruska proti USA z roku 2016, které se dokonce povedlo ovlivnit prezidentské volby. Lze jen domýšlet, o kolik by byl efekt této kampaně větší, kdyby měla Moskva k dizpozici deepfakes (21, s. 8).

Také již můžeme pozorovat využívání deepfakes v hybridní válce. V roce 2022 v rámci Rusko-ukrajinské války Moskva vytvořila deepfake video prezidenta Ukrajiny, Volodymyra Zelenského. Ve videu Zelenskyj přikazoval ukrajinským vojákům, aby složili zbraně, a vyhlašoval kapitulaci Ukrajiny. Video bylo zveřejněno na síti Facebook, Youtube a na webové stránce ukrajinské televizní stanice Ukraine 24, kde jej umístili ruští hackeři. Video však bylo nekvalitní – hlas prezidenta se pouze vzdáleně podobal jeho skutečnému hlasu a jeho obličej byl podivně strnulý, díky čemuž bylo prakticky okamžitě odhaleno (23). Další pokus o zneužití deepfakes (pravděpodobně opět ze strany Moskvy) následoval 24.6.2022, kdy berlínská primátorka ve videorozhovoru hovořila s člověkem, který se za pomocí deepfaku vydával za starostu Kyjeva, Vitalije Klička. Že jde o podvod se naštěstí povedlo odhalit, a to zejména díky nápadně podivným otázkám, na které se podvodník vyptával (24).   
Tyto případy jasně ukazují, že v budoucí době v rámci hybridní války budou deepfakes běžnou zbraní (25, s. 42).

# Detekce deepfakes

Detekce je velmi efektivní technologickou obranou před deepfakes. Platí, že stejně rychle, jako se vyvíjí technologie tvorby deepfakes, se také vyvíjí technologie pro jejich detekci. Protože detekovat deepfakes lidským okem často není efektivní, pro detekci jsou často využívány konvoluční neuronové sítě (CNN). Ty umí odhalit drobné nedokonalosti a digitální artefakty, které se vyskytují i v těch nejkvalitnějších deepfakes (13, s. 488).

Ve video deepfakes jsou nejčastějšími nedokonalostmi špatné nasvícení či pozice hlavy, které vznikají v důsledku rotací a změn velikosti při tvorbě těchto deepfakes. Překvapivým nedostatkem deepfakes je také nedokonalé mrkání, které vzniká v důsledku nedostatku obrázků použitých k trénování, na kterých má osoba zavřené oči. CNN mohou být natrénovány k detekci těchto nedokonalostí s velmi vysokou přesností a umožnit tak automatizovanou detekci deepfake videí (13, s. 488). U audio deepfakes je situace podobná – tedy žádný audio deepfake není dokonalý a tyto nedokonalosti lze využít pro jeho detekci s pomocí hlubokých neuronových sítí. Můžeme zde pozorovat využití reziduálních neuronových sítí, které dosahují velmi vysoké přesnosti detekce (16, s. 132-136).

# Závěr

S nezastavitelně rostoucím výkonem výpočetních zařízení a pokroky na poli umělé inteligence a strojového učení se na výsluní dostávají potenciálně nebezpečné technologie. Jednou z těchto technologií je i deepfake. Jedná se o formu syntetických médií, která využívá sílu AI a strojového učení k výrobě hyper-realistických, falešných videí (3, s. 135-136). Ačkoliv deepfakes mají i pozitivní a neškodné využití, jako například v audiovizuálním a herním průmyslu či k zábavě (5, s. 3246; 15; 3, s. 141), v současnosti jsou mnohem častěji zneužívány (3, s.143). Ať už se jedná o deepfake pornografii či šíření dezinformací, tato zneužití mohou mít velmi vážné následky, ať už na jednotlivce či celou společnost (19, s. 77-81; 21, s. 8)

Nově můžeme také vidět jejich využití v hybridní válce (zatím však naštěstí bez velkého účinku) (23; 24). Současná situace je však pouze stínem časů, které přijdou. Časů, kde důvěra v informace a média dosáhne nového dna a dezinformace se budou šířit jako nikdy předtím. Relativně snadná výroba deepfakes a vysoká jejich uvěřitelnost totiž vytvářejí velmi lákavou kombinaci pro všechny, kdo s jejich pomocí chtějí škodit.

Naštěstí proti deepfakes existuje efektivní obrana, a to jejich detekce. Stejný rychlý rozvoj AI a strojového učení, který umožňuje vytvářet stále kvalitnější deepfakes, umožňuje také jejich čím dál tím lepší detekci. Pokud je možno nějaké médium s vysokou pravděpodobností označit za deepfake, je škoda, kterou může způsobit, do značné míry minimalizována (13, s. 488). I když se ale deepfakes podaří úspěšně detekovat, stále mohou způsobit škodu, zejména podrývají důvěru občanů v média a zvyšují již tak vysoké úsilí, které je třeba vynaložit na vyhledání pravdivých informací (22, s. 3).

Z tohoto důvodu je třeba zaměřovat se i na jiné způsoby obrany, a to zejména edukaci veřejnosti a legislativní kroky. Tyto obranné prvky v současnosti podstatně zaostávají za rychlým vývojem deepfakes (19, s. 81-82), pokud ale nechceme čelit negativním následkům, je třeba do jejich rozvoje investovat potřebné prostředky. Hrozba deepfakes je již v současné době poměrně velká a nadále bude sílit. Rozhodně ale není nezvládnutelná, ovšem pouze za předpokladu, že společnost tuto hrozbu nepodcení a že metody pro boj s deepfakes budou vylepšovány stejně rychle, jako deepfakes samotné.

# Seznam zdrojů

1. SABBAGH, Karim a Roman FRIEDRICH. Maximizing the Impact of Digitization. *The Global Information Technology Report 2012*. 2012, 121-133.
2. MARTENS, Bertin, Luis AGUIAR, Estrella GGMEZ a Frank MUELLER-LANGER. The Digital Transformation of News Media and the Rise of Disinformation and Fake News. *SSRN Electronic Journal*. ISSN 1556-5068. Dostupné z: doi:10.2139/ssrn.3164170
3. KIETZMANN, Jan, Linda W. LEE, Ian P. MCCARTHY a Tim C. KIETZMANN. Deepfakes: Trick or treat?. *Business Horizons*. 2020, **63**(2), 135-146. ISSN 00076813. Dostupné z: doi:10.1016/j.bushor.2019.11.06
4. KERNER, Catherine a Mathias RISSE. Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics*. 2021, **8**(1), 81-108. ISSN 2194-5624. Dostupné z: doi:10.1515/mopp-2020-0024
5. ALBAHAR, Marwan a Jameel ALMALKI. Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*. 2019, **97**(22), 3242-3250. ISSN 1992-8645.
6. ARIK, Sercan a Jitong CHEN et al. Neural Voice Cloning with a Few Samples. *32nd Conference on Neural Information Processing Systems*. Montréal (CA), 2018.
7. PARIS, Britt a Joan DONOVAN. *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence*. 2019.
8. VAN DEN OORD, Aaron a Sander DIELEMAN. *WaveNet: A Generative Model for Raw Audio*. London (GB): Google DeepMind, 12-2016-09n. l.
9. O'NEIL, Luke. Doctored video of sinister Mark Zuckerberg puts Facebook to the test This article is more than 3 year. *The Guardian* [online]. London (GB), 2019-06-12 [cit. 2022-06]. Dostupné z: <https://www.theguardian.com/technology/2019/jun/11/deepfake-zuckerberg-instagram-facebook>
10. LECUN, Yann, Yoshua BENGIO a Geoffrey HINTON. Deep learning. *Nature*. 2015, **521**(7553), 436-444. ISSN 0028-0836. Dostupné z: doi:10.1038/nature14539
11. MONTAVON, Grégoire, Wojciech SAMEK a Klaus-Robert MÜLLER. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. 2018, **73**, 1-15. ISSN 10512004. Dostupné z: doi:10.1016/j.dsp.2017.10.011
12. BLANK, Dor, Noam KOENIGSTEIN a Raja GIRYES. Autoencoders. 2020-03
13. KATARYA, Rahul a Anushka LAL. A Study on Combating Emerging Threat of Deepfake Weaponization. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. IEEE, 2020, 2020-10-7, 485-490. ISBN 978-1-7281-5464-0. Dostupné z: doi:10.1109/I-SMAC49090.2020.9243588
14. CRESWELL, Antonia, Tom WHITE, Vincent DUMOULIN, Kai ARULKUMARAN, Biswa SENGUPTA a Anil A. BHARATH. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*. 2018, **35**(1), 53-65. ISSN 1053-5888. Dostupné z: doi:10.1109/MSP.2017.2765202
15. SHARF, Zack. Nicolas Cage Can Now Be Put Into Any Movie in History Thanks to A Machine-Learning Algorithm. *Indie Wire* [online]. Penske Media, 1998, 2018-01-30 [cit. 2022-06-30]. Dostupné z: <https://www.indiewire.com/2018/01/nicolas-cage-machine-learning-algorithm-deep-fakes-1201923224/>
16. 12 Best Deepfake Apps and Websites You Can Try for Fun. *Beebom* [online]. 2022-06-22 [cit. 2022-06-30]. Dostupné z: <https://beebom.com/best-deepfake-apps-websites/>
17. SOMERS, Meredith. Deepfakes, explained. *MIT Sloan School of Management* [online]. Cambridge (MA), 2020-07-21 [cit. 2022-06-30]. Dostupné z: <https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained>
18. BROOKS, Tina a Jesse HEATLEY et al. *Increasing Threat of Deep Fake Identities*. United States Department of Homeland Security.
19. DVOŘÁKOVÁ, Michaela. Revenge porn a deepfakes: ochrana soukromí v éře moderních technologií. *Revue pro právo a technologie*. 2020, **11**(22), 51-89. ISSN 1805-2797. Dostupné z: doi:10.5817/RPT2020-2-2
20. ROBERTSON, Adi. Reddit bans ‘deepfakes’ AI porn communities. *The Verge* [online]. Vox Media, 2011, 2018-02-07 [cit. 2022-06-30]. Dostupné z: <https://www.theverge.com/2018/2/7/16982046/reddit-deepfakes-ai-celebrity-face-swap-porn-community-ban>
21. CHESNEY, Robert a Danielle CITRON. Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. *Foreign Affairs* [online]. Council on Foreign Relations, 2019-02 [cit. 2022-07-01]. Dostupné z: <https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war>
22. VACCARI, Cristian a Andrew CHADWICK. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media Society*. 2020, **6**(1). ISSN 2056-3051. Dostupné z: doi:10.1177/2056305120903408
23. SIMONITE, Tom. A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be. *Wired* [online]. San Francisco (CA): Condé Nast, 1993, 2022-03-17 [cit. 2022-06-30]. Dostupné z: <https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/>
24. PERGLER, Tomáš. Primátorka Berlína si přes video povídala s falešným Vitalijem Kličkem. *Seznam Zprávy* [online]. Praha (CZ): Seznam.cz, 2016, 2022-06-25 [cit. 2022-06-30]. Dostupné z: <https://www.seznamzpravy.cz/clanek/zahranicni-stredni-evropa-primatorka-berlina-si-ctvrt-hodiny-povidala-s-falesnym-vitalijem-klickem-207362>
25. WESTERLUND, Mika. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*. Carleton University, 2019, **9**(11), 39-52. ISSN 1927-0321.
26. CHEN, Tianxiang, Avrosh KUMAR, Parav NAGARSHETH, Ganesh SIVARAMAN a Elie KHOURY. Generalization of Audio Deepfake Detection. *The Speaker and Language Recognition Workshop (Odyssey 2020)*. ISCA: ISCA, 2020, 2020-11-1, 132-137. Dostupné z: doi:10.21437/Odyssey.2020-19

1. Dostupné online na <https://www.youtube.com/>watch?v=FzMnDwpKJrI [↑](#footnote-ref-1)
2. Dostupné online na https://google.github.io/tacotron/publications/speaker\_adaptation/ [↑](#footnote-ref-2)
3. Dostupné online na https://www.youtube.com/watch?v=Ho9h0ouemWQ [↑](#footnote-ref-3)
4. Dostupné online na https://www.youtube.com/watch?v=b5AWhh6MYCg [↑](#footnote-ref-4)
5. Dostupné online na https://www.instagram.com/p/ByaVigGFP2U/ [↑](#footnote-ref-5)
6. Dostupné online na https://www.instagram.com/p/ByaVigGFP2U/ [↑](#footnote-ref-6)