
Recenzovaný článek

Deepfake jako pokročilá manipulativní technika k šíření propagandy

Deepfake as an Advanced Manipulative Technique for Spreading Propaganda

Martin Havlík

Abstrakt:

Článek popisuje základní fundamenty a význam manipulativní techniky nazývané Deepfake, která se v prostředí technologické a informační expanze stává také velmi využívaným nástrojem pro šíření propagandy. Tato pokročilá manipulace doplňuje širší spektrum forem dezinformací a bývá stále častěji využívána jako prostředek k vedení informačních operací, nezřídka kdy jako součást širšího hybridního válčení. Efektivně bojovat s tímto druhem manipulace klade vysoké nároky na konzumenty informací, a to jak ze strany použitých detekčních nástrojů, tak ze strany kognitivního lidského přístupu založeného na kritickém myšlení. Expanze a sofistikovanost obdobných manipulativních technik bude nadále pokračovat, a to v souvislosti s rozvojem moderních technologií a provázaností informačního prostředí.

Abstract:

The article describes the basic foundations and significance of the manipulative technique called Deepfake, which in the environment of technological and informational expansion is also becoming a widely used tool for spreading propaganda. This advanced manipulation complements a wider spectrum of forms of disinformation and is increasingly being used as a means of conducting information operations, often as part of wider hybrid warfare. Effectively combating this kind of manipulation places high demands on consumers of information, both on the part of the detection tools used and on the part of the cognitive human approach based on critical thinking. The expansion and sophistication of similar manipulative techniques will continue, in connection with the development of modern technologies and the interconnectedness of the information environment.

Klíčová slova: Hrozba; informace; manipulace; obrana; propaganda; technologie.

Key words: Threat; Information; Manipulation; Defence; Propaganda; Technology.

ÚVOD

V dnešním digitálním světě a bezpečnostně turbulentním prostředí jsou stále intenzivněji používány manipulační techniky k ovlivnění cílového publika. V posledních letech je stále složitější jednoduše rozlišovat mezi pravdivými informacemi a falešnými zprávami, ať už je jejich záměr jakýkoliv. Kromě již značně rozšířených falešných textových zpráv tzv. dezinformací, misinformací, mal-informací, Fake News a zpráv typu Hoax, se v posledních několika letech dostává do popředí šíření klamných videí tzv. Deepfake¹. Deepfake je významnou aktuální hrozbou pro mediální manipulace v nejrůznějším kontextu, které dokážou výše zmiňované původní falešné informace katapultovat na daleko důvěryhodnější úroveň. Samotná problematika technologie Deepfake je založena na úmyslných úpravách videí s cílem změnit jejich původní obsah (znění)² a ovlivnit tak cílové publikum. Při hlubším zamyšlení je nutné podotknout, že v určité fázi docházelo k úmyslným úpravám obsahu video-formátu již před mnoha desítkami let, kdy při dabingu zahraničních filmů se autoři dabingů snažili (a nadále snaží) synchronizovat pohyb úst a tváří původních postav s hlasem dabéra v transformovaném (překládaném) jazyce. V posledních letech lze v této oblasti pozorovat velké úspěchy demonstrované téměř nepostřehnutelnými odchylkami (postřizitelnými pouhým okem diváka) původních videí od videí upravených.

Deepfake video je výsledek nové technologie, která byla uveřejněna před několika málo lety. Vychází ze zpracování dat neuronovými sítěmi a lze ji využít například pro záměnu tváří. Nejprve našla tato technika uplatnění v porno průmyslu, nicméně brzy na to mediální společnost BuzzFeed ukázala i mnohem nebezpečnější stránku této manipulativní technologie. Konkrétně, když vytvořila video o americkém prezidentu Obamovi, v němž sděluje informace, které ve skutečnosti nikdy nevypustil z úst. A toto video je přitom neuvěřitelně důvěryhodné.^{3, 4}

Problematika Deepfake velmi aktuální, a v bezpečnostním kontextu velmi relevantní, neboť je úzce spojená se šířením propagandy, vedením informačních operací či v komplexním podání se záměrným hybridním působením. Význam této problematiky je umocněn aktuální bezpečnostní situací determinované především konfliktem na Ukrajině. Cílem tohoto článku je popis specifické manipulativní techniky se zaměřením na

¹ Deepfake je původní označení upravených videí speciálními softwarovými aplikacemi, který takto jako první označil vývojář Redditu.

² Případně vytvořit zcela nová videa bez vědomí a účasti podvržené osoby.

³ ZEMANOVA, Milada. *Nová DeepFake videa jsou k nerozeznání od reality*. Agentura Focus. [online]. 2018. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/krlqj>

⁴ Mnoho lidí výměnu tváří pomocí Deepfake považuje za neetickou, pro některé ovšem tato technologie nenapodobuje realitu dostatečně přesvědčivě. Nicméně i tak se již objevily vylepšené video portréty (tzv. Deep Video Portraits), které působí velmi reálně a zdokonalují a zlepšují samotnou techniku. Na téma Deep Video Portraits byla vytvořena studie, která byla předána ke zvážení v SIGGRAPH (Mezinárodní asociace specializovaných zájmových skupin pro počítačové grafiky a interaktivní techniky založená v New Yorku). Tato studie popisuje zdokonalenou techniku reprodukce pohybů, výrazů obličeje a řeči jedné osoby za pomoci tváře osoby druhé.

bezpečnostní aspekty spadající do oblasti obrany. K naplnění cíle přispěje jak základní popis techniky Deepfake, tak její propojení na oblast hrozeb a v neposlední řadě také exkurz zabývající se možnostmi detekce a mitigace manipulativních videí. Uplatnění může článek najít jak uvnitř bezpečnostní komunity, tak rovněž i u širšího publika, kterému není lhostejné šíření klamných nepravdivých informací.

1 TERMINOLOGICKÉ VYMEZENÍ PROPAGANDY

Propaganda ve svém pravém slova smyslu je všude a navíc ji moderní technologie umožňují šířit okamžitě k cílovému publiku, které může být jakkoliv vzdáleno. Pro mnoho obyvatel západní společnosti (nejen pro Američany či obyvatele Evropské unie) je termín propaganda asociován s fašistickým či komunistickým režimem, jakkoliv ani tyto pro-západně orientované společnosti se v šíření vlastní propagandy systémově nijak neliší. Toto tvrzení je podloženo argumenty uvedenými v propagačním materiálu amerického národního archivu, který v uplynulých letech pořádal speciální výstavu demonstrující právě historické propagační kampaně USA (zahrnující filmové materiály, zvukové nahrávky, letáky, prospekty, plakáty, hračky, časopisy atd.).⁵

Pro vyjádření významu termínu propaganda lze využít výzkumy z minulých dekad, které se snaží tento specifický pojem vykreslit v emotivně neutrálním pojetí. V nejjednodušším významu lze propagandu popsat jako formu komunikace s určitým charakterem manipulace či vlivu. V publikaci *Propaganda and Persuasion*⁶ je propaganda definována jako nejvíce neutrální smysl k šíření či podpoře konkrétních myšlenek. Toto poměrně zjednodušené pojetí však může generovat určité problémy a vždy je nutné na problematiku propagandy nahlížet komplexnějším prizmatem, které reflektuje jak kvalitativní a kvantitativní rozměr, tak rovněž i oblast studovaného zájmu či konkrétní vědní disciplíny (psychologie, sociologie, historie, politologie atd.). JOWETT a O'DONNELL definovali propagandu jako záměrný a systematický pokus utvářet požadované vnímání, manipulovat kognitivní stránku a řídit chování jednotlivce či skupiny k dosažení požadovaného záměru propagandisty.⁷

Odlišné definice propagandy publikoval například Jaques ELLUL, který propagandu popsal jako formu informací, jež se podbízí naší nejistotě a úzkosti. Negativem tohoto terminologického vymezení je nedostatek neutrality a částečně pejorativní charakter. Negativní chápání propagandy je spojeno s většinou existujících definic.⁸ S ohledem na stále narůstající význam a zejména masivní dosah sociálních sítí (například Facebook,

⁵ National Archives Traveling Exhibits Service. Propaganda in Black & White. An Exploration of Overt and Covert U.S. Government Propaganda. National Archives Museum, Pennsylvania, Washington. [online]. 2015. [cit. 2022-09-28]. Dostupné z: <https://www.archives.gov/files/exhibits/nates/files/propaganda.pdf>

⁶ JOWETT, Garth S., O'DONNELL, Victoria. *Propaganda and Persuasion*. Thousand Oaks: Sage Publications Ltd. 2006.

⁷ Ibidem.

⁸ ELLUL, Jacques. *Propaganda: The Formation of Men's Attitudes*. New York: Vintage Books. 1973.

Twitter, Instagram, Tik Tok atd.) je vhodné v kontextu terminologického vymezení propagandy uvést také další rozšířené definice, které mohou rozšířit definiční portfolio tohoto odborného článku. Zajímavou definici představil Richard Alan NELSON, který propagandu vymezil jako formu účelového přesvědčování, která se pokouší ovlivňovat emoce, postoje, názory a činy konkrétních cílových skupin pro ideologické, politické či komerční účely. K tomuto cíli je využit řízený a jednosměrný přenos specifických zpráv (které mohou, ale také nemusí být založeny na faktech) skrze masová média.⁹ Jakkoliv je poslední definice poměrně konkrétní, obsahově zřejmě a charakterem neutrální, v některých specifických oblastech může být až příliš široká či naopak příliš specifická. Emilio FERRARA poukazuje na to, že například v průběhu krizových situací mohou být sociální média efektivně využívána pro zvládání mimořádných událostí, nicméně mohou také vyvolávat hysterické reakce a paniku.¹⁰

Z vojenské perspektivy je propaganda definována jako jakákoli forma komunikace na podporu národních cílů navržená za účelem ovlivnění názorů, emocí, postojů nebo chování jakékoli skupiny ve prospěch sponzora, ať už přímo nebo nepřímo (Joint Publication 1-02, 2001).¹¹ Philip M. TAYLOR propagandu definuje jako vědomé, metodické a plánované rozhodnutí používat techniky přesvědčování navržené k dosažení konkrétních cílů, které mají být přínosem pro ty, kteří tento proces organizují.¹² Propaganda může být definována také jako plánované použití jakýchkoliv komunikačních prostředků k vnucení vlastní vůle v konkrétní nebo potenciální konfliktní situaci, a to komukoli, kdo není běžně ochoten se smířit s touto vůlí. Za nejzávažnější formu propagandy lze v kontextu případných škod považovat formu, které je často označována jako černá propaganda (black propaganda).¹³ Jinými slovy, černá propaganda není nic menšího než forma intelektuálního a politického podvracení (subverze). Jejím účelem je připsat akce jinému než skutečnému zdroji. Klíčem k úspěšným operacím propagandy je vyvinout obecné téma, na které budou navázány všechny druhy jednotlivých operací. U každého druhu klamání bylo rychlejší, snazší a účinnější posílit to, v co již nepřítel věřil nebo co tušil, než se je pokoušet přesvědčit o něčem zcela novém.

V kontextu hybridního válčení můžeme propagandu charakterizovat jako systematickou cílenou činnost směřující k ovlivnění nejširších mas (cílového publika) určitými myšlenkami a směry jednání, kdy je takovéto získání širokého publika využito pro podporu politiky určité vlády. Propaganda je pak velmi často založena na formě demagogického

⁹ NELSON, Richard A. *A Chronology and Glossary of Propaganda in the United States*. Westport, Conn, and London: Greenwood Press. 1996.

¹⁰ FERRARA, Emilio. *Contagion Dynamics of Extremist Propaganda in Social Networks*. Information Sciences, Volumes 418 – 419, 2017, no. 1, 1-12.

¹¹ Department of Defense Dictionary of Military and Associated Terms. Joint Publication 1-02. [online]. 2001. [cit. 2022-09-28]. Dostupné z: <https://dcs9.army.mil/assets/docs/dod-terms.pdf>

¹² TAYLOR, Philip M. *Munitions of the mind: a history of propaganda from the ancient world to the present era*. New ed. New York: Distributed exclusively in the USA by St. Martin's Press, 1995. ISBN 0719048303.

¹³ BECKER, Howard. The Nature and Consequences of Black Propaganda. *American Sociological Review*, vol. 14, no. 2, 1949, pp. 221–235. JSTOR. [online]. 1949. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/urlqu>

prosazování a doporučování myšlenek, názorů s cílem získat přívržence veřejného prostoru pro účinný politický boj, který může přerůst rovněž k vedení války.¹⁴

Černá propaganda je forma propagandy, jejímž cílem je vyvolat dojem, že byla vytvořena těmi, které má diskreditovat. Černá propaganda kontrastuje se šedou propagandou, která neidentifikuje její zdroj, a bílou propagandou, která svůj původ vůbec nemaskuje. Obvykle se používá k hanobení nebo ztrapnění nepřítele prostřednictvím zkreslování.¹⁵ Hlavní charakteristikou černé propagandy je, že si cílové publikum neuvědomuje, že jej někdo ovlivňuje, a nemá pocit, že je tlačeno určitým směrem.¹⁶ Tento typ propagandy je spojen se skrytými psychologickými operacemi.¹⁷ Někdy je zdroj skryt nebo připsán na vrub falešné autoritě a šíří lži, výmysly a podvody. Černá propaganda se spoléhá na ochotu příjemce přijmout důvěryhodnost zdroje. Pokud tvůrci nebo odesílatelé černé propagandistické zprávy dostatečně nerozumějí svému zamýšlenému publiku, zpráva může být špatně pochopena, může se zdát podezřelá nebo může selhat úplně.¹⁸ Vlády vedou černou propagandu z důvodů, které zahrnují: maskováním jejich přímého zapojení může vláda s větší pravděpodobností uspět v přesvědčování jinak nevěřícího cílového publika a za používáním černé propagandy jsou diplomatické důvody. Černá propaganda je nezbytná k zamlžování účasti vlády na aktivitách, které mohou být škodlivé pro její zahraniční politiku.^{19,20}

2 ZÁKLADNÍ POPIS TECHNIK DEEPAKE

Význam slova Deepfake popisuje uměle vytvořená videa snažící se nahradit obličej ve videu obličejem někoho jiného, využívající pokročilých metod umělé inteligence a strojového učení.²¹ Podstatou technologie Deepfake není často pouhá výměna obličejů ve videích, ale jde spíše o poměrně věrohodnou úpravu mimiky, která demonstruje mimo jiné i pohyb úst. Toto lze pak následně propojit s příslušnou zvukovou stopou, která doplní výsledný obraz tak, že modifikované osoby v obraze sdělují jiná slova, než v původním originálu. Deepfake je obecně tedy označení pro realistickou úpravu či nové

¹⁴ Založeno na recenzi připomínce a souvisejícím definičním vymezení propagandy podle *Malá československá encyklopedie ČSAV, V. svazek, písmeno Pom–S*. Academia, Praha 1987.

¹⁵ DOOB, Leonard. „Goebbels’ Principles of Nazi Propaganda“. *The Public Opinion Quarterly*. 3 (Vol. 14, No. 3): 419–442. JSTOR 2745999.

¹⁶ Ellul, Jacques (1965). *Propaganda: The Formation of Men’s Attitudes*, p. 16. Trans. Konrad Kellen & Jean Lerner. Vintage Books, New York. ISBN 978-0-394-71874-3.

¹⁷ LINEBARGER, Paul Myron Anthony. 1954. *Psychological Warfare*, Combat Forces Press, Washington.

¹⁸ JOWETT, Garth S., O’DONNELL, Victoria. *Propaganda and Persuasion*. Thousand Oaks: Sage Publications Ltd. 2006.

¹⁹ SHULSKY, Abram, SCHMITT, Gary. *Silent Warfare*. Washington, DC: Brassey’s Inc. 2002.

²⁰ SOG - Studies and Observation Group - OPS-33. *Black Propaganda*. [online]. 2018. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/Hrl2y>

²¹ Slovník cizích slov – Infoz.cz. Deepfake. [online]. 2022. [cit. 2022-09-28]. Dostupné z: <https://www.infoz.cz/deepfake/>

vytvoření videa, která umožňují velmi věrohodně změnit či emulovat mimiku.²² V praxi tak lze osobám na videu vkládat do úst věty, které nikdy nepronесly, provádět záměny postav, obličejů apod. Deepfake využívá pokročilého počítačového zpracování dat, a to velmi často s využitím prvků umělé inteligence (konkrétně neuronových sítí) a kvalita výsledného produktu se tak neustále zvyšuje.²³ Obecnější definici předkládá PERSONS, podle kterého Deepfake zahrnuje video, fotografie nebo zvukovou nahrávku, jenž se zdá být skutečná, ale byla zpracována pomocí umělé inteligence. Tato technologie umožňuje nahrazovat obličeje, manipulovat s výrazem obličeje, syntetizovat obličeje a syntetizovat řeč. Deepfake může zobrazovat někoho, kdo vypadá, že říká nebo dělá něco, co ve skutečnosti nikdy neřekl nebo neudělal. Deepfake spoléhá na umělé neuronové sítě, které rozpoznávají vzory v datech. Vyprodukování falešné fotografie nebo videa obvykle zahrnuje vložení stovek nebo tisíců obrázků do umělé neuronové sítě a její trénování k identifikaci a rekonstrukci vzorů (obvykle obličejů).²⁴ Technik, které stojí za tvorbou Deepfake je celá řada. V některých případech jsou také do technik umělé manipulace zahrnovány také technicky jednodušší postupy, které jsou označovány jako Cheepfake. Tyto lze považovat za předchůdce Deepfake, jenž vyžadují méně odborných znalostí a méně potřebných technických zdrojů. Ze současných používaných technik lze zmínit například počítačovou synchronizaci rtů, manipulaci s atributy, výměnu či manipulaci obličejů (záměnu výrazů a rysů), generování zcela nových fiktivních tváří, hlasovou syntézu v kontextu vytváření audiovizuálních falzifikátů, syntézu celé tváře apod. Podle odborné a technické náročnosti lze pak jednotlivé metody manipulace označované za Deepfake kategorizovat od nejjednodušších a zdrojově nejméně náročných technik až po techniky velmi sofistikované a zdrojově velmi náročné. Možné chronologické uspořádání významných kategorií technik Deepfake je uvedeno na následujícím obrázku (viz Obr. 1). Uvedený grafický přehled integruje základní manipulativní techniky, které se souhrnně označují termínem Deepfake (bílý text v uvedeném obrázku). Schéma je dále doplněno o výběr hlavních softwarových nástrojů a specifických modelů, které jsou k tvorbě jednotlivých skupin Deepfake nejvíce využívány (černý text v uvedeném obrázku).

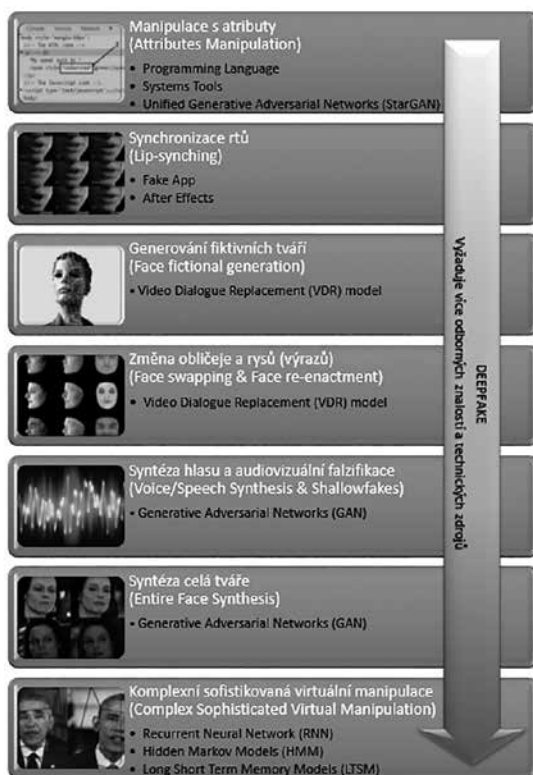
Bližší rozbor postupů jednotlivých manipulativních technik charakteru Deepfake není ambicí tohoto odborného článku, nicméně i tak je v dalším textu nastíněn základní fundamentální přístup pro snazší pochopení technických aspektů. Mimo předložený obecný exkurz lze pro technicky zdatné čtenáře tohoto přehledového článku doporučit pro

²² Můžeme dostat až na úroveň, kdy vše bude komplet vytvořeno od „základů“, tj. nedojde k úpravě něčeho již natočeného.

²³ KOPECKÝ, Kamil. Deep fake - stručný úvod do problematiky. E-Bezpečí, roč. 4, č. 1, s. 23-25. Olomouc: Univerzita Palackého. ISSN 2571-1679. [online]. 2019. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/yrlqL>

²⁴ PERSONS, Timothy M. GAO - SCIENCE & TECH SPOTLIGHT: DEEPPFAKES. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/grl27>

rozšíření informační báze dostupný poznámkový aparát odkazující na relevantní zdroje, ve kterých mohou zájemci nalézt bližší technický popis některých zmíněných metod. 25.26.27



Obrázek č. 1: Možná kategorizace technik Deepfake²⁸

²⁵ Malicious Uses and Abuses of Artificial Intelligence. Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI), Europol's European Cybercrime Centre (EC3). 2020. Pp. 53.

²⁶ TOLOSANA, Ruben, VERA-RODRIGUEZ, Ruben, FIERREZ, Julian, MORALES, Aythami, ORTEGA-GARCIA, Javier. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid, Spain. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://arxiv.org/pdf/2001.00179.pdf>

²⁷ PARIS, Britt, DONOVAN, Joan. Deepfakes and Cheep fakes – The Manipulation of Audio and Visual Evidence. Data&Science. [online]. 2019. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/jrIEm>

²⁸ Inspirováno: PARIS, Britt, DONOVAN, Joan. Deepfakes and Cheep fakes – The Manipulation of Audio and Visual Evidence. Data&Science. [online]. 2019. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/jrIEm> a DAVIS, Raina, WIGGINS, Chris, DONOVAN, Joan. TECH FACTSHEETS FOR POLICYMAKERS SPRING 2020 SERIES - Deepfakes. Technology and Public Purpose Project, Belfer Center for Science and International Affairs, Harvard Kennedy School. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/XrIFI>

Jak vyplývá z výše uvedeného obrázku, může stát za tvorbou Deepfake řada rozdílných technik. Volný přístup k rozsáhlým veřejným databázím spolu s rychlým pokrokem technik hlubokého učení, zejména Generative Adversarial Networks, přispívají k vytváření stále sofistikovanějších a velmi realistických falešných videí s odpovídajícími důsledky pro společnost. Jakkoliv není kategorizace manipulativních technik zcela jednotná, je snahou nastínit alespoň základní popisy nejrozšířenějších technik.

Poměrně rozšířená technika *Face swapping* obecně zahrnuje výměnu obličejů hlavních aktérů videí, a to ať už překrytím obličejů jiných osob či úpravou klíčových rysů, výrazů nebo pouhou synchronizací rtů. Tyto parciální techniky se někdy označují konkrétněji jako *Face replacement* zaobírající se právě výměnou obličeje jedné osoby překrytím k obličeji jiné osoby a technika *Face re-enactment* zahrnující manipulace s rysy obličeje cíle tak, aby to vypadalo, že říká něco, co není. Zmiňovaná synchronizace rtů je pak označována jako *Lip-synching*. Pro výměnu obličejů (tedy *Face swapping*) lze využít také zcela nově generované fiktivní tváře syntetických obrazů jednotlivců, které se jeví jako reálné a velmi přesvědčivé. Tato doplňková technika je v anglosaských zdrojích označována jako *Face generation*. Další známou a rozšířenou technikou je *Speech synthesis*, jejíž obsahem je primárně generování zvukového obsahu pomocí tréninkových algoritmů pro vytvoření falešného hlasu nebo syntetického zvukového souboru, a to s cílem dosáhnout výsledné audiovizuální falzifikace. Takto vytvořené audiovizuální falzifikáty méně sofistikované povahy pomocí základních editačních technik jsou pak označovány termínem *Shallow-fakes*.^{29, 30} Přehled některých uvedených technik pro manipulaci s obličejem, včetně metod syntézy celé tváře, záměny identity, manipulace s atributy a záměny výrazů popisuje také TOLOSANA a kolektiv ve svém odborném výzkumném článku. Pro každou skupinu manipulace poskytují podrobnosti a argumenty týkající se manipulačních technik, stávajících veřejných databází pro technologické vyhodnocení metod detekce, včetně souhrnu výsledků z těchto hodnocení.³¹

Jak již bylo nastíněno výše, je vhodné pro doplnění kontextu manipulativních technik Deepfake rozšířit pohled také o původní manipulativní techniky nazývané Cheepfake.^{32, 33} Tyto jsou pro názornost zobrazeny na následujícím obrázku (viz Obr. 2).

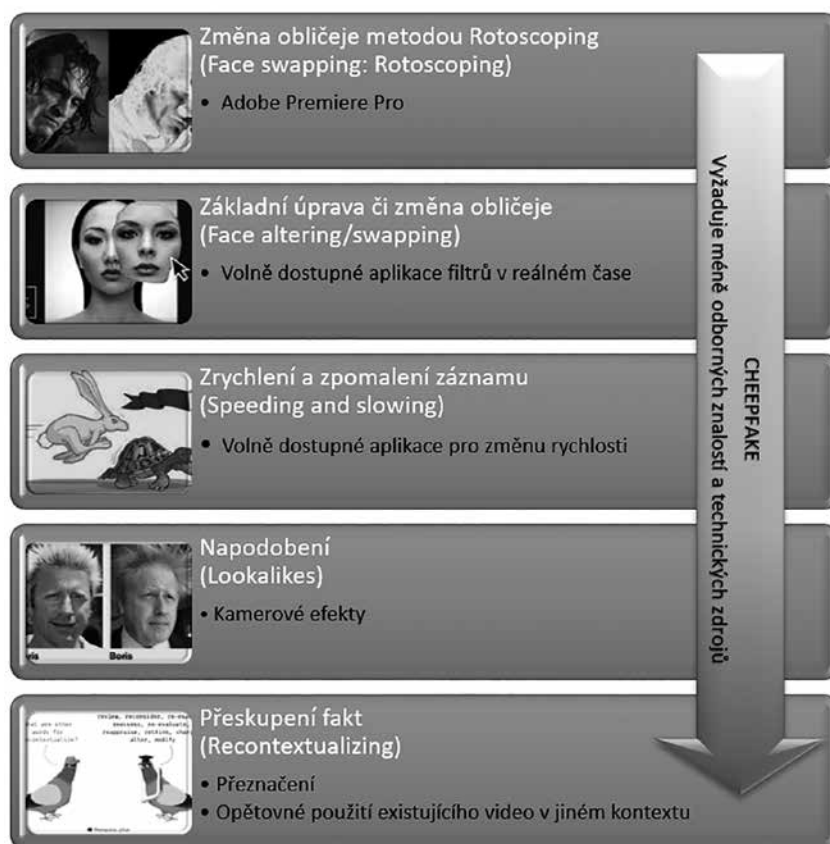
²⁹ Malicious Uses and Abuses of Artificial Intelligence. Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI), Europol's European Cybercrime Centre (EC3). 2020. Pp. 53.

³⁰ DAVIS, Raina, WIGGINS, Chris, DONOVAN, Joan. TECH FACTSHEETS FOR POLICYMAKERS SPRING 2020 SERIES - Deepfakes. Technology and Public Purpose Project, Belfer Center for Science and International Affairs, Harvard Kennedy School. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/XrIFl>

³¹ TOLOSANA, Ruben, VERA-RODRIGUEZ, Ruben, FIERREZ, Julian, MORALES, Aythami, ORTEGA-GARCIA, Javier. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid, Spain. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://arxiv.org/pdf/2001.00179.pdf>

³² PARIS, Britt, DONOVAN, Joan. Deepfakes and Cheep fakes – The Manipulation of Audio and Visual Evidence. Data&Science. [online]. 2019. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/jrIEm>

³³ DAVIS, Raina, WIGGINS, Chris, DONOVAN, Joan. TECH FACTSHEETS FOR POLICYMAKERS SPRING 2020 SERIES - Deepfakes. Technology and Public Purpose Project, Belfer Center for Science and International Affairs, Harvard Kennedy School. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/XrIFl>



Obrázek č. 2: Možná kategorizace technik Cheepfake³⁴

Obdobně jak byly výše krátce představeny techniky označované jako Deepfake, zasluží si zde zmínku také základní techniky označované jako Cheepfake, které stály u zrodu pokročilejší manipulace s obrázky a následně videi. Základní techniku demonstruje *Re-contextualizing*, tedy představení a šíření existujícího videa (záznamu) pod falešnou záminkou nebo zobrazení upraveného videa mimo jeho původní kontext. Další známou technikou je *Lookalikes*, nebo-li napodobení, které lze popsat jako určité najímání a nahrazování podobných aktérů, aby se vydávali za cílové jedince. Doplňková technika *Speeding and Slowing* (v českém překladu jako zrychlení a zpomalení) zahrnuje technickou

³⁴ Inspirováno: PARIS, Britt, DONOVAN, Joan. Deepfakes and Cheep fakes – The Manipulation of Audio and Visual Evidence. Data&Science. [online]. 2019. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/jrLEm> a DAVIS, Raina, WIGGINS, Chris, DONOVAN, Joan. TECH FACTSHEETS FOR POLICYMAKERS SPRING 2020 SERIES - Deepfakes. Technology and Public Purpose Project, Belfer Center for Science and International Affairs, Harvard Kennedy School. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/XrIFl>

změnu rychlosti videa za účelem změny významu nebo vnímání. Další techniky jako *Face altering* a *Face swapping – Rotoscoping* demonstrují snahu o záměnu obličejů za účelem sloučení dvou vrstev do jedné a vytvoření výsledného efektu záměny obličeje. Tyto základní modifikační a manipulativní techniky jsou velmi často dále doplněny o napodobené hlasy skutečné osoby bez přímého zdrojového obsahu.³⁵

3 BEZPEČNOSTNÍ RIZIKA A TRENDY SOUVISEJÍCÍ S DEEPAKE

Deepfake představuje obecně velmi mocný nástroj, který lze použít jak k pozitivním a přínosným účelům (zejména ve filmovém průmyslu), tak bohužel zejména k exploataci a šíření dezinformací. Deepfake může například zásadním způsobem ovlivnit volby či obecně narušit důvěru společnosti či vybrané cílové skupiny obyvatelstva. Mimo to může být Deepfake efektivně využito k šíření propagandy a nejrůznějších konspiračních teorií (s cílem ovlivnit veřejné mínění). Klíčovou hrozbou této manipulativní techniky je to, že dochází k zásadnímu stírání rozdílů mezi věrohodnou veřejnou debatou a cílenou manipulací. Úspěšné manipulaci (nejen v kontextu šíření propagandy) napomáhá také stále rostoucí diverzifikace používaných komunikačních platform (především dosah sociálních sítí). Související bezpečnostní rizika jsou dále umocněna dostupností základních softwarových nástrojů využívajících prvky umělé inteligence, kterou mohou být na jednu stranu využity k samotné detekci Deepfake, na druhou stranu však samotnou detekci snižují a generují výhody pro samotné tvůrce Deepfake. Umělá inteligence je nejen v kontextu manipulativních technik chápána jako komplexně propojený a spolupracující systém počítačů, aplikací, algoritmů, dat, informací, procesů a člověka, který posouvá význam moderních technologií k užtku společnosti, nicméně který také generuje nové množství hrozeb (digitálních, fyzických či politických). Mezi tyto lze zařadit především:

- Destrukce image a důvěryhodnosti jednotlivce.
- Obtěžování nebo ponižování jednotlivců v online prostředí.
- Páchání vydírání a podvodů.
- Usnadňování podvodů s dokumenty.
- Falšování online identit a klamání mechanismů KYC³⁶.
- Padělání nebo manipulace s elektronickými důkazy pro vyšetřování.
- Narušení finančních trhů.
- Šíření dezinformací a manipulace s veřejným míněním.
- Podněcování k násilí vůči menšinovým skupinám.
- Podpora narativů extremistických nebo dokonce teroristických skupin.
- Podněcování sociálních nepokojů a politické polarizace.

³⁵ DAVIS, Raina, WIGGINS, Chris, DONOVAN, Joan. TECH FACTSHEETS FOR POLICYMAKERS SPRING 2020 SERIES - Deepfakes. Technology and Public Purpose Project, Belfer Center for Science and International Affairs, Harvard Kennedy School. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/XrIFl>

³⁶ Know Your Customer (KYC) je proces, kdy banky získávají informace o identitě svých zákazníků, a tím zajišťují, že nedojde ke zneužití bankovních služeb.

- Šíření dezinformací (včetně politických).
- Kyberkriminalita.
- Nedobrovolná pornografie atd.^{37, 38}

Další relevantním fenoménem, který ovlivňuje Deepfake jako nástroj šíření propagandy je globální rostoucí diversifikace aktérů a využívání sponzorovaných (komerčních) proxy-aktérů k vedení informačních operací, které se tak z určitého pohledu mohou stávat stále častěji jako oficiálně nabízena komerční služba (klidně však veřejně skrytá). Obtížná atribuce původců konkrétních informačních, kybernetických a komplexních hybridních hrozeb s využitím Deepfake, růst sofistikovanosti schopností původců těchto hrozeb s cílem skrýt jejich identitu a zastínit technické pozadí a přístupy, determinuje současná bezpečnostní rizika a trendy. Tomuto napomáhá také narůstající celospolečenská nejistota, budoucnost plná nejistot a tříštění velkých komplexních informačních operací do vícera malých úzce cílených aktivit se synergickým efektem.

4 MOŽNOSTI ODHALOVÁNÍ A DETEKCE DEEPAKE

V kontextu dynamických změn prostředí, ve kterém žijeme, je potřeba reflektovat narůstající složitost, vícerozměrnost, vzájemnou závislost a provázanost celého světa. Dříve mnohem více agregovaná moc se stala rozptýlenější, posunula se částečně od západu na východ i od vlád úplně, protože mezinárodní nestabilita využívá stále více nestátních subjektů. Důvodem je z velké části digitální revoluce, která narušila státní hranice (potřela striktní diverzifikaci mezi vnitřní a vnější bezpečností) a generovala nové výzvy i příležitosti. Digitalizace umožnila lidstvu rovněž postavit mosty, které v rámci sdílení a šíření informací eliminují geografické rozdíly.³⁹ Jak bylo uvedeno, digitální svět generuje viditelné příležitosti, nicméně paralelně také hrozby, kterým je potřeba čelit. S ohledem na potlačení původních hranic mezi jasně vymezenými teritorii státních aktérů, jež nejsou v kybernetickém prostoru téměř vůbec reflektovány, je pro eliminaci souvisejících hrozeb nutné změnit přístup. Zde hraje důležitou roli komplexnost a spolupráce mezi více aktéry. Jak řekl Joseph NYE, který poprvé definoval termín *Soft Power*, síla s ostatními může být efektivnější než moc nad ostatními.⁴⁰ V návaznosti na hrozby pocházející z informačního prostředí a kybernetického prostoru (a také tyto specifické domény využívají ke svému šíření) je vhodné zdůraznit, že právě současné informační, kybernetické a hybridní hrozby stále více ovlivňují jednotlivce a veřejné mínění, než tomu

³⁷ Malicious Uses and Abuses of Artificial Intelligence. Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI), Europol's European Cybercrime Centre (EC3). 2020. Pp. 52.

³⁸ DAVIS, Raina, WIGGINS, Chris, DONOVAN, Joan. TECH FACTSHEETS FOR POLICYMAKERS SPRING 2020 SERIES - Deepfakes. Technology and Public Purpose Project, Belfer Center for Science and International Affairs, Harvard Kennedy School. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/XrIFI>

³⁹ McCCLORY, Jonathan. *General Manager, Asia*. Portland. [online]. 2021. [cit. 2022-09-28]. Dostupné z: <https://portland-communications.com/our-people/leadership-team/jonathan-mcclory/>

⁴⁰ McCCLORY, Jonathan. *The Soft Power 30*. USC Center on Public Diplomacy. [online]. 2021. [cit. 2022-09-28]. Dostupné z: <https://softpower30.com/author/jonathan/>

bylo kdykoliv v minulosti. Hrozby už nejsou primárně cíleny na stát jako instituci, ale svými podvrtnými charakteristikami cílí na širší publikum a celou společnost s častým záměrem nabourat společenskou soudržnost, institucionální integritu a další související demokratické standardy. Uvedené bezpečnostní premisy se tak nepochybně váží také k přístupu pro odhalování a detekci Deepfake, respektive potřebě zavádět nové detekční procesní mechanismy.

Detekce sofistikovaných Deepfake je obecně extrémně obtížná, časově náročná a technická omezení omezují přímou detekci na mnoha koncových zařízeních (zejména na mobilních telefonech a dalších mobilních zařízeních). Dalším faktem je to, že uživatelé velmi často přehlížejí viditelné nesrovnalosti a nedostatky ve videích nízké kvality s odůvodněním, že se jedná o chybu *streamování*⁴¹ nebo o chybu způsobenou nestabilním či nedostatečným internetovým (datovým) připojením. Kromě tohoto fenoménu vypadají mnohá videa nízké kvality na malé obrazovce mobilních zařízení lépe a mohou tak snadno oklamat uživatele.

Při mitigaci rizik související s technikou Deepfake je vhodné efektivní způsoby rozdělit na dvě základní části. V prvním případě jde o výběr obecných procesů a postupů, jak v rámci chráněných aktiv postupovat proti vlivovým operacím či propagandě, k jejímž šířením je právě Deepfake využíván. Do této části tak patří nutnost kombinace automatické detekce a expertního vyšetřování. Dále pak celospolečenský reaktivní přístup (označovaný často termíny *Whole of Society Approach* či *Whole of Society Response*), který je založen na spolupráci a sdílení dat i informací mezi zainteresovanými subjekty (akademická sféra, nezávislí novináři, státní sektor, bezpečnostní experti atd.). Nedílnou součástí úspěšného členění informačním operacím využívajícím technik Deepfake je samozřejmě také efektivní budování kapacit pro odstrašení, posilování komplexní odolnosti chráněných aktiv a potažmo celé společnosti či posilování operační bezpečnosti ve vztahu k bezpečnostnímu sektoru.⁴² Druhá zásadní skupina pak obsahuje konkrétní technická opatření, která lze účinně využívat pro detekci informací modifikovaných technikou Deepfake. V tomto případě hrají důležitou roli především specifické softwarové nástroje a verifikační metody umožňující cílovému publiku či expertním skupinám analyzovat původ, záměry, cíle a zdroje posuzovaných informací typu Deepfake.

Detekční softwarové nástroje obsahují specifické algoritmy schopné zachytit jemné nedokonalosti ve změněném obsahu, které jsou pouhým okem neviditelné. Například u modifikovaných zvukových stop jsou to konkrétně spektrogramy zvukových záznamů, které umožňují v podobě vizualizací odhalit podobně znějící hlasy. Špičkové techniky detekce pak identifikují vzory zvuků a inflexní body ve zvukovém záznamu, které lidé běžně nedělají nebo nemohou fyzicky vytvořit.⁴³

⁴¹ Streamování je technologie kontinuálního přenosu audiovizuálního sdělení mezi zdrojem a koncovým uživatelem pomocí datového proudu. V současné době se streamování využívá především pro přenášení audiovizuálních sdělení po internetu.

⁴² GLEICHER, Nathaniel, FRANKLIN, Margarita, AGRANOVICH, David, NIMMO, Ben, BELOGOLOVA, Olga, TORREY, Mike. Threat Report – The State of Influence Operations 2017 – 2020. 2021.

⁴³ WADDELL, Kaveh. "Researchers Are Figuring out How to Detect Audio Deepfakes before It's Too Late." Axios. [online]. 2019. [cit. 2022-09-28]. Dostupné z: <https://www.axios.com/deepfake-audio-ai-impersonators-f736a8fc-162e-47f0-a582-e5eb8b8262ff.html>

Je vhodné doplnit, že některé významné vládní agentury i globální komerční společnosti implementovaly specifické programy, které mají za cíl problémy spojené s šířením Deepfake eliminovat, a to v zájmu společnosti a veřejného blaha. Například Agentura pro výzkum pokročilých obranných projektů (DARPA - Defense Advanced Research Projects Agency) ministerstva obrany USA podporuje velký výzkumný projekt v oblasti mediálního forenzního výzkumu *MediFor* (Media Forensics).⁴⁴ Spustila také svůj vlastní program *SemaFor* (Semantic Forensics), kde se výzkumníci snaží pomoci počítačům používat logické uvažování selským rozumem k detekci zmanipulovaných médií.^{45,46} Technologické společnosti, jako je Facebook a Google, spustily vlastní iniciativy a iniciativy *Crowdsourcing*⁴⁷ pro lepší identifikaci Deepfake kolujících na jejich platformách.^{48,49} Jakkoliv jsou v této oblasti z celospolečenské perspektivy nadále patrné mezery, můžeme právě tento počín nadnárodních společností a americké agentury považovat za první hmatatelné zárodky změny komplexního celospolečenského přístupu.

Vývoj v oblasti zpracování obrazu udělal za poslední roky značný pokrok. Využití hlubokých neuronových sítí se zaměřením na zpracování vizuálních dat bude v budoucnu nutností. S rozvojem moderních technologií se postupně dostáváme do fáze, kdy se falešná videa (upravená či zcela zfalšovaná) masově rozšiřují informačním prostředím (sociálními sítěmi) a budou cílit na uživatele, kteří nedokáží jejich autenticitu rozpoznat. Nástroje, které by Deepfake umožnily odhalit, nestíhají držet krok s rychlostí, jakou nástroje na vytváření Deepfake vznikají.⁵⁰ Navzdory této nevýhodě tak bude nutné do procesu odhalování a detekce zapojit také nepřetržité kritické myšlení a jiné verifikační mechanismy, jakoliv i toto je a bude stále obtížné a bude to vyžadovat nové úsilí a neustále vlastní učení.

⁴⁴ DARPA - Defense Advanced Research Projects Agency. CORVEY, William. Media Forensics (MediFor) (Archived). [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://www.darpa.mil/program/media-forensics>

⁴⁵ TUREK, Matt. "Defense Advanced Research Projects Agency." Defense Advanced Research Projects Agency. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://www.darpa.mil/program/semantic-forensics>

⁴⁶ DARPA - Defense Advanced Research Projects Agency. Announces Research Teams Selected to Semantic Forensics Program. [online]. 2021. [cit. 2022-09-28]. Dostupné z: <https://www.darpa.mil/news-events/2021-03-02>

⁴⁷ Definice ve slovníku cizích slov nazývá crowdsourcing jako centrálně organizovanou činnost, která vede k dosažení přesně definovaných cílů, a to s využitím většího množství zainteresovaných osob z řad zákazníků nebo sympatizantů, pocházejících z cílových skupin podle pole působnosti organizátora. Viz: Slovník cizích slov [online]. 2015 [cit. 2022-10-02]. Dostupné z: <http://slovník-cizich-slov.abz.cz/web.php/slovo/crowdsourcing>

⁴⁸ MCCABE, David, DAVEY, Alba. "Facebook Says It Will Ban 'Deepfakes'." The New York Times. The New York Times. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://www.nytimes.com/2020/01/07/technology/facebook-says-it-will-ban-deepfakes.html>

⁴⁹ DAVIS, Raina, WIGGINS, Chris, DONOVAN, Joan. TECH FACTSHEETS FOR POLICYMAKERS SPRING 2020 SERIES - Deepfakes. Technology and Public Purpose Project, Belfer Center for Science and International Affairs, Harvard Kennedy School. [online]. 2020. [cit. 2022-09-28]. Dostupné z: <https://1url.cz/XrIFl>

⁵⁰ KOPECKÝ, Kamil. Deep fake - stručný úvod do problematiky. E-Bezpečí, roč. 4, č. 1, s. 23-25. Olomouc: Univerzita Palackého. ISSN 2571-1679. [online]. 2019. [cit. 2022-09-28]. Dostupné z: <https://www.e-bezpeci.cz/index.php?view=article&id=1417>

ZÁVĚR

Deepfake je fenomén, který v budoucnu zásadním způsobem ovlivní práci s ověřováním informací a jejich pravdivosti. Rozšiřující se dostupnost této manipulativní techniky napomáhá nežádoucímu šíření dezinformací, a to zejména s využitím kybernetického prostoru (masově zejména s využitím sociálních sítí). Význam manipulativní techniky Deepfake, která se v prostředí technologické a informační expanze stává také velmi využívaným nástrojem pro šíření propagandy, si zasluhuje pozornost bezpečnostní komunity. Tato pokročilá technologická manipulace rozšiřuje spektrum dostupných forem dezinformací a bývá stále častěji využívána jako prostředek k vedení informačních operací, nežádá kdy jako součást širšího hybridního válčení. Efektivně bojovat s tímto druhem manipulace pak klade vysoké nároky na konzumenty informací, a to jak ze strany použitých detekčních nástrojů (programů s prvky umělé inteligence), tak ze strany kognitivního lidského přístupu založeného na kritickém myšlení. Je logické, že expanze a sofistikovanost obdobných manipulativních technik bude nadále pokračovat, a to v přímé souvislosti s rozvojem moderních technologií a provázaností informačního prostředí skrze kybernetický prostor. Jakkoliv se související hrozby využívající techniky Deepfake nezaměřují pouze na bezpečnostně-vojenské aspekty, nelze jejich vliv na vojenské informační operace a komplexní hybridní válčení opomíjet. Všechny tyto argumenty akcentují současný význam multi-doménového boje, kde dostupnost informací, jejich validita a relevance multiplikují výsledné efekty ve všech současných konfliktech a krizových situacích. Primární ambicí tohoto článku je tak předložit základní rámec problematiky Deepfake v kontextu potenciálních bezpečnostních rizik.

Autor: *Plk. gšt. PhDr. Ing. Martin Havlík, Ph.D., DBA, MBA, MSc., LL.M. Pracovník Ministerstva obrany České republiky. Specializuje se na zpravodajskou problematiku, zejména pak na oblast technických zpravodajských disciplín, informačních a kybernetických operací jakožto nedílné součásti hybridního působení státních i nestátních aktérů. Mimo uvedené se dlouhodobě zaměřuje na analýzy bezpečnostních hrozeb a rizik v oblastech konfliktů (především v Asii) s dopadem na obranu a bezpečnost České republiky.*

Jak citovat: HAVLÍK, Martin. Deepfake jako pokročilá manipulativní technika k šíření propagandy. *Vojenské rozhledy*. 2023, 32 (1), 003-017. ISSN 1210-3292 (print), 2336-2995 (on-line). Available at: www.vojenskerozhledy.cz.