Mirga, Tomáš

Dezinformácie v ére digitálnej transformácie : generatívna AI ako nástroj tvorby rizikového syntetického obsahu : prípadová štúdia a strategické riešenia

ProInflow. 2024, vol. 16, iss. 1, pp. [33]-95

ISSN 1804-2406 (online)

Stable URL (DOI): https://doi.org/10.5817/ProIn2024-37974
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.79861

ProInflow: časopis pro informační vědy
Vol. 16, No. 1, 2024, str. 33–95
https://doi.org/10.5817/ProIn2024-37974

DEZINFORMÁCIE V ÉRE DIGITÁLNEJ TRANSFORMÁCIE: GENERATÍVNA AI AKO NÁSTROJ TVORBY RIZIKOVÉHO SYNTETICKÉHO OBSAHU – PRÍPADOVÁ ŠTÚDIA A STRATEGICKÉ RIEŠENIA

DISINFORMATION IN THE ERA OF DIGITAL TRANSFORMATION: GENERATIVE AI AS A TOOL FOR GENERATING HIGH-RISK SYNTHETIC MEDIA – CASE STUDY AND STRATEGIC SOLUTIONS

Tomáš Mirga

Katedra knižničnej a informačnej vedy, Univerzita Komenského v Bratislave

Abstrakt

Účel – Príspevok skúma rastúci fenomén generatívnej AI a jej vplyv na tvorbu a šírenie dezinformácií, falošných správ a deepfake obsahov v internetovom prostredí. Cieľom je poskytnúť komplexný pohľad na výzvy spojené s neetickým využívaním generatívnej AI a predstaviť strategické riešenia tohto problému.

Design / metodológia / prístup – Hlavnou použitou metódou bola deskriptívna prípadová štúdia s prvkami explanácie a explorácie, ako čiastkové metódy boli využité obsahová analýza jednotlivých prípadov, indukcia a interpretácia zistení.

Výsledky – Prípadová štúdia ukázala, že generatívna AI predstavuje významný nástroj v rukách šíriteľov dezinformácií a misinformácií. Schopnosť AI generovať presvedčivý a vizuálne čoraz ťažšie rozoznateľný obsah od skutočnosti umožňuje bezprecedentnú manipuláciu s informáciami. Zistenia prípadovej štúdie majú dôležité teoretické aj praktické implikácie. Teoreticky rozširujú pochopenie dynamiky medzi technologickým pokrokom a jeho vplyvom na spoločnosť. Prakticky poukazujú na urgentnú potrebu aplikovania technologických, právnych a intelektuálnych riešení. Limity štúdie sa týkajú možnej subjektivity pri výbere prípadov a interpretácii dát.

Originalita / hodnota – Príspevok poskytuje komplexný pohľad na problematiku neetického využívania generatívnej AI na tvorbu a šírenie dezinformácií a deepfake obsahov v internetovom prostredí. V rámci riešení problematiky príspevok navrhuje potrebu zavedenia prístupu zameraného na kombináciu technologických riešení, právnej regulácie a intelektuálnych prístupov.

Kľúčové slová: generatívna umelá inteligencia, dezinformácie, falošné správy, deepfake, syntetický obsah

Abstract

Purpose – The contribution examines the growing phenomenon of generative AI and its impact on the creation and dissemination of disinformation, fake news, and deepfake content in the online environment. The aim is to provide a comprehensive view of the challenges associated with the unethical use of generative AI and to present strategic solutions to address this issue.

Design / methodology / approach – The main method used was a descriptive case study with elements of explanation and exploration, with content analysis of individual cases, induction, and interpretation of findings as partial methods.

Results – The case study demonstrated that generative AI represents a significant tool in the hands of disseminators of disinformation and misinformation. The ability of AI to generate convincing content that is increasingly difficult to distinguish from reality allows for unprecedented manipulation of information. The findings of the case study have important theoretical and practical implications. Theoretically, they expand the understanding of the dynamics between technological advancement and its impact on society. Practically, they highlight the urgent need for the application of technological, legal, and intellectual solutions. The limitations of the study relate to potential subjectivity in the selection of cases and interpretation of data.

Originality / value – The contribution provides a comprehensive view of the issue of unethical use of generative AI for creating and spreading disinformation and deepfake content in the online environment. In addressing the issue, the contribution suggests the need for the implementation of an approach focused on combining technological solutions, legal regulation, and intellectual approaches.

Keywords: generative artificial intelligence, disinformation, fake news, deepfake, synthetic media

ÚVOD

V úsvite novej éry, keď generatívna AI prekračuje hranice toho, čo bolo kedysi považované za doménu výlučne ľudskej kreativity a analýzy, stojíme na prahu revolúcie, ktorá má potenciál predefinovať naše vnímanie reality, pravdy a dôvery. Tento príspevok sa primárne zameriava na riziká, ktoré generatívna AI prináša v tvorbe a šírení dezinformácií, falošných správ a deepfake obsahov. Naším cieľom je poskytnúť komplexný pohľad na aktuálne výzvy, ktoré generatívna AI predstavuje pre spoločnosť, médiá a individuálne správanie.

V úvode článku sa venujeme definícii kľúčových pojmov, ktoré sú nevyhnutné pre pochopenie rozsahu problematiky a negatívnych dopadov neetického využívania generatívnej AI. Priblížime, ako rýchly vývoj a implementácia týchto technológií v mediálnom priemysle a ďalších sektoroch vedie k novým formám komunikácie, kreativity a interakcie. Poukážeme aj na temnú stránku týchto inovácií – schopnosť generovať presvedčivý obsah, ktorý môže byť zneužitý na šírenie nepravdivých informácií a manipuláciu s verejnou mienkou. Ďalej sa budeme venovať stručnému prehľadu vývoja umelej inteligencie, jej využitiu v mediálnom priemysle a základným princípom fungovania generatívnej AI. Tento úvod nám umožní lepšie pochopiť, ako sme sa dostali do bodu, kde generatívna AI nie je len nástrojom pre vývojárov a výskumníkov, ale ako sa stala súčasťou našej každodennej digitálnej skúsenosti.

V nasledujúcich častiach príspevku sa zameriame na konkrétne spôsoby zneužitia generatívnej AI na tvorbu manipulatívneho syntetického obsahu. V rámci prípadovej štúdie poukážeme na závažnosť stavu problematiky a ilustrujeme reálny dopad týchto technológií na spoločnosť, politické procesy, demokraciu, ale i jednotlivcov. V poslednej časti príspevku preskúmame možné riešenia problematiky neetického využívania systémov generatívnej AI. Tieto riešenia zahŕňajú technologické inovácie, právne rámce a intelektuálne prístupy s cieľom pomôcť vytvoriť spoločnosť odolnejšiu proti dezinformáciám v internetovom prostredí.

Naším zámerom je nielen informovať o potenciálnych rizikách a výzvach spojených s generatívnou AI, ale aj inšpirovať k diskusii o tom, ako môžeme ako spoločnosť čeliť týmto výzvam prostredníctvom vzdelávania, regulácie, prijatia právnych riešení a rozvoja etických prístupov. Vstupujeme do éry, v ktorej je kľúčové pochopenie a zodpovedné využívanie generatívnej AI s cieľom zabezpečiť, že jej potenciál bude slúžiť v prospech ľudstva, a nie na jeho škodu.

VYMEDZENIE ZÁKLADNÝCH POJMOV

Predtým, než si hlbšie rozoberieme hlavné aspekty problematiky, je dôležité spomenúť kľúčové pojmy práce, ich vzájomný vzťah a dôležitosť v kontexte súčasného informačného prostredia.

Generatívna umelá inteligencia

Generatívna umelá inteligencia (GAI) predstavuje pokročilú podmnožinu umelej inteligencie (AI), ktorá sa špecificky zameriava na *tvorbu nového obsahu* – od textov, cez obrázky, až po zvuk, hudbu či video. Táto forma AI využíva rôzne algoritmické prístupy na generovanie obsahu, ktorý je často nerozoznateľný od obsahu vytvoreného ľudskými tvorcami. Táto schopnosť je založená na komplexných algoritmoch, ktoré dokážu analyzovať a napodobňovať ľudské vzorce tvorby obsahu.

Jednou z kľúčových technológií v generatívnej AI sú *generatívne adversárne siete* (GAN), ktoré sú založené na princípe dvoch súťažiacich neurónových sietí – jedna generuje obsah a druhá hodnotí jeho autenticitu (Park, 2023). Aplikácie GAN zahŕňajú tvorbu fotorealistických obrázkov, originálnych umeleckých diel, ale aj vývoj tzv. deepfake médií. Deepfake je technológia používajúca metódy strojového učenia a AI na vytváranie alebo modifikáciu videozáznamov, obrázkov a audionahrávok s cieľom nahradiť tváre alebo napodobniť hlas určitej osoby, čo vedie k vytváraniu presvedčivých falošných informačných obsahov s potenciálnym zneužitím na šírenie dezinformácií alebo na manipuláciu verejnej mienky (Kai, Sliva, Suhang, Jiliang & Huan, 2017).

Na tvorbu textu sú využívané *veľké jazykové modely* (LLM), ktoré sú trénované na veľkých datasetoch textových dát a dokážu generovať koherentné a relevantné texty. LLM sú trénované pomocou techník hlbokého učenia, najmä s využitím architektúry *Transformer*. Tieto modely sa učia z rozsiahleho množstva textových dát, aby pochopili jazykové vzory, gramatiku a kontext. LLM majú široké využitie, vrátane automatického generovania textu, prekladu jazykov, sumarizácie dokumentov alebo tvorby kreatívneho obsahu, ako sú napríklad básne alebo príbehy (Baidoo-Anu & Owusu, 2023).

Neurónové siete a pokročilé algoritmy strojového učenia sú využívané aj na generovanie hudby, reči alebo iných zvukových efektov. Dokážu napodobniť špecifické štýly, vytvárať komplexné hudobné diela rôznych žánrov, či generovať prirodzene znejúcu reč.

Generatívna AI nachádza uplatnenie v oblastiach, kde je potrebné vytvoriť nový obsah, ako je digitálne umenie a grafika, hudobná produkcia (hudba a zvukové efekty), automatické písanie textov, filmový priemysel (vizuálne efekty a postprodukcia), architektúra (vizualizácie), reklama a marketing (personalizovanie obsahu), herný priemysel (tvorba herných prostredí, postáv a príbehov), či generovanie syntetických trénovacích dát na trénovanie iných AI modelov.

Dezinformácie

Dezinformácie sú často chápané ako úmyselné šírenie nepravdivých alebo zavádzajúcich informácií. Tento fenomén sa vyznačuje zámerom klamať alebo zavádzať, čo ho odlišuje od misinformácií – neúmyselného šírenia nesprávnych alebo nepravdivých informácií. Dezinformácie sú vytvárané a šírené s úmyslom dosiahnuť konkrétny cieľ, napríklad spôsobiť verejnú škodu, ovplyvniť verejnú mienku, ovplyvniť voľby, poškodiť reputáciu, podnietiť nenávisť alebo získať zisk. Na označovanie dezinformácií sa často používa termín *fake news*, ktorý nie je v tomto kontexte úplne presný. Na jednej strane dezinformácie často obsahujú aj fakty zmiešané s výmyslami. Na druhej strane môžu zahŕňať aj manipulované videá, cielenú reklamu, organizovaný trolling a iné praktiky. Tieto termíny tiež bývajú zneužívané na diskreditáciu „nepohodlných“ informácií a útoky na nezávislé médiá. V súčasnom digitálnom veku sa dezinformácie rýchlo šíria prostredníctvom sociálnych médií a iných online platforiem, čo zvyšuje ich dosah a potenciálny vplyv na verejnú mienku a politické procesy. Ohrozujú rôzne sektory, od zdravotníctva a vedy až po vzdelávanie a financie. Motiváciou ich šírenia môže byť ekonomický zisk, politické ciele či ideologická propaganda. Dezinformácie predstavujú závažný problém v súčasnej informačnej spoločnosti, kde rýchle šírenie nepravdivých alebo zavádzajúcich informácií môže vyvolať vážne dôsledky. Ohrozujú demokratické procesy, národnú bezpečnosť, sociálnu súdržnosť a podkopávajú dôveru v informačnú spoločnosť a digitálny jednotný trh (Commission, 2018).

Rôzni autori pristupujú k problematike dezinformácií a možným riešeniam ich tvorby a šírenia z rôznych uhlov pohľadov, čo odráža multidisciplinárny charakter tohto fenoménu. Podľa Hasanain, Alam, Mubarak, Abdaljalil, Zaghouani, Nakov, Martino & Freihat (2023) sa dezinformácie vyznačujú použitím presvedčovacích techník v rámci textových príspevkov na sociálnych sieťach alebo v novinových článkoch, s cieľom ovplyvniť názory alebo postoje čitateľov. Tento pohľad zdôrazňuje, že dezinformácie nie sú len o nepravdivých informáciách,

ale aj o manipulatívnych technikách používaných na dosiahnutie určitého cieľa. Bușincu & Alexandrescu (2023) tiež poukazujú na dezinformácie ako na fenomén, ktorý je šírený rôznymi aktérmi s cieľom manipulovať a ovplyvňovať verejnosť. Ich pohľad poukazuje na potrebu transparentných a dôveryhodných riešení, ktoré by pomohli verejnosti správne sa informovať a rozhodovať v prostredí plnom dezinformácií. Feuerriegel, DiResta, Goldstein, Kumar, Lorenz-Spreen, Tomz & Pröllochs (2023) za novú výzvu v súčasnosti definujú dezinformácie generované umelou inteligenciou. Upozorňujú na sofistikovanosť dezinformácií v digitálnej ére, kedy AI môže byť použitá na vytváranie presvedčivých, no zavádzajúcich alebo nepravdivých obsahov, čo komplikuje ich detekciu a boj proti nim. Na nedostatky tradičných prístupov k detekcii a intervencii proti dezinformáciám v ére využívania AI poukazujú Xu, Fan a Kankanhalli (2023). Ferrara (2024) zdôrazňuje, že technológie generatívnej AI môžu byť zneužité na šírenie falošných informácií a manipuláciu verejnej mienky, čo vyvoláva potrebu rozvoja efektívnych metód detekcie a protiopatrení. Fátima (2023) rozoberá dezinformácie z pohľadu informačných vied. Zdôrazňuje, že dezinformácie môžu byť identifikované a analyzované pomocou rôznych metód, ako sú faktová a lingvistická kontrola alebo analýza sentimentu, pričom AI hrá kľúčovú úlohu v automatizácii tohto procesu. Guarnera, Giudice, Nastasi & Battiato (2020) zas navrhujú novú metódu založenú na analýze anomálií vo frekvenčnej doméne. Affsprung (2023) naznačuje, že riešenia rizík generatívnej AI vyžadujú nielen technické, ale aj sociálne a etické prístupy. Tredinnick & Laybats (2023) argumentujú, že AI by mala byť považovaná za globálnu prioritu, podobne ako pandémia alebo jadrová vojna. Táto perspektíva zdôrazňuje potrebu globálnej spolupráce a vytvorenia komplexných stratégií na zmiernenie rizík spojených s rozvojom a nasadením generatívnej AI. Iní autori, ako napríklad Simon, Altay & Mercier (2023) zas argumentujú, že obavy z vplyvu generatívnej AI na informačné prostredie sú prehnané, a naznačujú potrebu vyváženejšieho pohľadu na potenciálne riziká a prínosy tejto technológie.

Vzájomný vzťah medzi pojmami *dezinformácie* a *generatívna umelá inteligencia* je komplexný a multidimenzionálny. Súčasný výskum v tejto oblasti poukazuje na dvojstrannú povahu tohto vzťahu. Na jednej strane generatívna AI poskytuje nástroje na tvorbu presvedčivých dezinformácií, falošných správ alebo zavádzajúcich informácií. Na druhej strane ponúka možnosti na ich detekciu a elimináciu. Využitím techník strojového učenia a spracovania prirodzeného jazyka môže AI identifikovať vzory a indikátory, ktoré sú pre dezinformácie

typické. Vzťah týchto dvoch pojmov preto vyžaduje pozornosť vo vývoji AI technológií, ako aj v etických a regulačných rámcoch ich používania (Barredo, Jamil & Montemayor, 2023; Küçkerdoğan & Turğal, 2023).

Relevancia témy zneužívania generatívnej AI na tvorbu dezinformácií a falošných správ je v súčasnosti významná z viacerých dôvodov. V dôsledku rýchleho vývoja technológií AI a ich schopnosti generovať presvedčivý syntetický obsah sa otvorili nové možnosti neetického využívania týchto nástrojov s cieľom šíriť nepravdivé alebo zavádzajúce informácie. Syntetický obsah sa vzťahuje na digitálne médiá vytvorené alebo upravené pomocou technológií AI, ako sú generatívne adversárne siete (GAN), ktoré napodobňujú reálny vizuálny alebo zvukový materiál s vysokou presvedčivosťou. Tieto technológie umožňujú vytváranie realistických obrázkov, videí, zvukov a textov, ktoré môžu byť nerozoznateľné od autentického obsahu vytvoreného ľuďmi alebo zachyteného prostredníctvom tradičných metód (Manco, Ritacco, Rullo, Saccá & Serra, 2022).

To predstavuje vážne riziko pre demokratické spoločnosti, pretože dezinformácie môžu vplývať na verejnú mienku, manipulovať s voličmi a dokonca destabilizovať politické systémy. Rýchly vývoj AI a jej využitie na tvorbu dezinformácií tak prirodzene vyvoláva otázky týkajúce sa regulácie, etiky a transparentnosti. Vzhľadom na sofistikovanú povahu dezinformácií vytváraných pomocou generatívnej AI je dôležité, aby ľudia vedeli rozpoznať a kriticky hodnotiť obsah, ktorý konzumujú. Hlavná téma tejto práce je v súčasnosti relevantná z toho dôvodu, že neetické využívanie technológií generatívnej AI ohrozuje základné aspekty demokracie, verejnej dôvery a etiky v digitálnej ére (Barredo, Jamil & Montemayor, 2023).

STRUČNÝ VÝVOJ UMELEJ INTELIGENCIE A JEJ VYUŽITIA V MEDIÁLNOM PRIEMYSLE

Od skromných začiatkov až po súčasné sofistikované aplikácie, AI prešla dlhým vývojom, ktorý zásadne zmenil spôsob, akým pristupujeme k informáciám a médiám. Aby sme pochopili výzvy, ktoré AI prináša v súčasnom mediálnom prostredí, je dôležité si tento vývoj stručne priblížiť.

História AI sa začala v 50. rokoch 20. storočia, kedy boli položené základy pre výskum v oblasti strojového učenia a vývoja algoritmov. Významným míľnikom bolo založenie laboratórií, ako *MIT Artificial Intelligence Laboratory*, ktoré sa stali centrami pre vývoj AI. Významné boli práce ako Turingov test a vývoj prvých programov schopných hrať šach alebo riešiť algebrické problémy. V 70. a 80. rokoch 20. storočia došlo k významnému pokroku v oblasti expertných systémov a neurónových sietí. Tieto technológie umožnili vykonávať komplexnejšie úlohy a začali sa používať v rôznych priemyselných aplikáciách. V 90. rokoch, s nástupom internetu, sa AI začala využívať v širšom meradle, najmä v oblasti spracovania dát a personalizácie obsahu. O príchode moderných AI technológií môžeme hovoriť so začiatkom 21. storočia, kedy došlo k významnému pokroku v oblasti strojového a hlbokého učenia. Vývoj algoritmov ako súčasť veľkých dátových projektov a pokrok vo výpočtovej technike umožnili vznik pokročilých AI aplikácií, ako sú chatboty, algoritmy na odporúčanie obsahu či nástroje na rozpoznávanie obrazu (Goodfellow, Bengio & Courville, 2016; Russell & Norvig, 2010; Turing, 1950).

V oblasti médií sa AI začala využívať s príchodom digitálnej éry a rozšírením internetu v 90. rokoch, a to predovšetkým v oblasti personalizácie obsahu a cieleného marketingu. S rozvojom sociálnych médií a digitálnych platforiem v 21. storočí sa AI stala kľúčovým nástrojom na analýzu používateľských dát, cielenie reklám či automatizáciu a optimalizáciu mediálneho obsahu. V súčasnosti dochádza k masívnemu uplatneniu technológií generatívnej AI, ktoré nachádzajú uplatnenie nielen v mediálnom priemysle, ale aj v rôznych iných praktických úsekoch ľudskej činnosti (Díaz-Noci, 2023; Peña-Fernández, Meso-Ayerdi & Larrondo-Ureta, 2023; Rabinder, 2019).

ZÁKLADNÉ PRINCÍPY FUNGOVANIA GENERATÍVNEJ AI

Pre komplexné pochopenie problematiky neetického využívania generatívnej AI na tvorbu dezinformácií v internetovom prostredí považujeme za prínosné objasniť si základné princípy fungovania mechanizmov týchto systémov.

Základným stavebným kameňom moderného strojového učenia a AI sú *neurónové siete*. Tieto štruktúry napodobňujú fungovanie ľudského mozgu a sú schopné učiť sa a vykonávať rôzne úlohy, od rozpoznávania obrazov, cez spracovanie prirodzeného jazyka až po predpovedanie trendov a klasifikáciu dát v rôznych oblastiach. Neurónové siete sú tvorené digitálnymi neurónmi, ktoré sú organizované do vrstiev. Každý neurón v jednej vrstve je spojený s neurónmi v nasledujúcej vrstve, pričom vzťahy medzi nimi môžu nadobúdať rôzne úrovne sily spojenia. Vrstvy sa delia na vstupnú, viacero skrytých a výstupnú. Vstupná vrstva prijíma vstupné dáta, skryté vrstvy sa venujú spracovávaniu dát a výstupná vrstva generuje výsledok. Skryté vrstvy sú kľúčové pre učenie sa komplexných vzorcov v dátach. Vo fáze *spracovania vstupu* (forward pass) sú vstupné dáta postupne spracovávané cez rôzne vrstvy neurónovej siete. Každá vrstva aplikuje svoje sily prepojení medzi neurónmi na dáta a prenáša výsledok do nasledujúcej vrstvy, až kým nedosiahnu výstupnú vrstvu, kde je vygenerovaný konečný výsledok. V nasledujúcej fáze *aktualizácie prepojení* (backward pass) sa výstup porovná s očakávaným výsledkom. Rozdiel medzi týmito dvoma výstupmi, známy ako chyba alebo strata, je potom použitý na aktualizáciu sily prepojení v sieti. Tento proces začína od výstupnej vrstvy a postupuje späť k vstupnej vrstve. Cieľom je nastaviť prepojenia tak, aby sa pri budúcom spracovaní podobných vstupných dát dosiahla menšia chyba. Tento dvojfázový proces umožňuje neurónovej sieti učiť sa z chýb a postupne zlepšovať svoju schopnosť správne spracovávať vstupné dáta (Goodfellow, Bengio & Courville, 2016; LeCun, Bengio & Hinton, 2015).

Nástroje AI pri tvorbe nového obsahu využívajú predovšetkým *generatívne adversárne siete* (GAN), ktoré slúžia napríklad na vytváranie realistických obrazov, videí, zvukov alebo iných typov dát. Tento proces funguje na princípe „súťaže“ medzi dvoma neurónovými sieťami: *generátorom* a *diskriminátorom.* Generátor je neurónová sieť, ktorá sa snaží vytvoriť dáta, ktoré sú nerozoznateľné od skutočných. Začína s náhodným vstupom (šumom) a postupne ho transformuje do výstupu, ktorý má požadované vlastnosti (napríklad obraz alebo zvuk). Diskriminátor je druhá neurónová sieť, ktorá sa snaží rozlíšiť, či sú dáta vytvorené generátorom,

alebo či ide o skutočné dáta. Jeho úlohou je naučiť sa rozpoznávať rozdiely medzi skutočnými a umelo generovanými dátami. Iteratívny proces pokračuje až dovtedy kým generátor nedokáže vytvoriť dáta, ktoré sú pre diskriminátor nerozoznateľné od skutočných. Jednou z hlavných výziev pri práci s GAN je zabezpečiť, aby generátor a diskriminátor boli vyvážené. Ak je jeden z nich príliš „silný“, môže to viesť k nestabilnému trénovaniu a k zhoršeniu kvality výstupu (Gaurav, Zhang & Zhu, 2022; Xin, Hui, Shu, Ming-Ching & Siwei, 2022).

Ďalším technickým aspektom generatívnej AI je využitie *difúznych modelov*, ktoré sú podobne ako GAN využívané pri tvorbe obrazu, zvuku a iných typov dát. Tieto modely fungujú na princípe postupného pridávania a odstraňovania šumu z dát, čo umožňuje vytvárať nové vzory, ktoré sú nerozoznateľné od skutočných. Difúzne modely začínajú s reálnymi dátami (napríklad obrazom) a postupne do nich pridávajú šum. Tento proces, nazývaný *forward diffusion*, transformuje pôvodné dáta do čoraz viac nerozpoznateľných vzorov. Po pridávaní šumu nasleduje proces *reverse diffusion*, keď sa model snaží odstrániť šum a vrátiť dáta do ich pôvodného stavu. Počas tohto procesu model generuje nové dáta, ktoré sú podobné pôvodným, ale v skutočnosti ide o nové vzory. Trénovanie difúzneho modelu zahŕňa učenie sa správneho odhadu, ako odstrániť šum a vrátiť dáta do ich pôvodného stavu. Tento proces vyžaduje veľké množstvo dát a výpočtovú kapacitu. Difúzne modely sú schopné vytvárať vysoko kvalitné a realistické vzory, ktoré môžu byť použité v rôznych oblastiach, od umenia a zábavy až po vedecký výskum (Podell, English, Lacey, Blattmann, Dockhorn, Müller, Penna & Rombach, 2023).

Na spracovanie a generovanie prirodzeného jazyka sú využívané *veľké jazykové modely* (LLM). Tieto modely fungujú na princípe hlbokého učenia a sú schopné vykonávať širokú škálu úloh súvisiacich s jazykom. LLM sú trénované na veľkých datasetoch textových dát, čo im umožňuje učiť sa jazykové vzorce, gramatiku, kontext a štýl písania. Vedia generovať koherentné a relevantné texty (ako články, príbehy, básne, vtipy, recepty, emaily a pod.), ale aj technické dokumenty, návody a manuály. Dokážu tiež generovať programovací kód, zhrnúť rozsiahle texty, rozpoznávať otázky a odpovedať na ne, konverzovať s používateľom v prirodzenom jazyku či analyzovať sentiment (Chen, Pan, Li, Ding & Zhou, 2023; Kim, Xu, McDuff, Breazeal & Park, 2023).

Ako sme spomenuli v rámci definície pojmov, LLM sú založené na architektúre Transformer, ktorá predstavuje pokročilý typ neurónových sietí v oblasti spracovania prirodzeného jazyka

(NLP). Ako príklady spomeňme *BERT* (bidirectional encoder representations from transformers) a *GPT* (generative pre-trained transformer). Tieto modely využívajú architektúru založenú na *mechanizme pozornosti* (attention mechanism), ktorý umožňuje efektívne spracovávať dlhé sekvencie textu, zohľadňovať kontext v rámci celej sekvencie a tiež sústrediť sa na rôzne časti vstupnej sekvencie pri generovaní výstupu. Tento mechanizmus poskytuje flexibilitu v modelovaní závislostí medzi slovami bez ohľadu na ich vzdialenosť v texte. V transformer architektúre rozlišujeme dve hlavné časti: *enkóder*, ktorý spracováva vstupný text, a *dekóder*, ktorý generuje výstup. V prípade modelov ako GPT je použitý len dekóder, zatiaľ čo BERT využíva len enkóder. LLM sú typicky zostavené z viacerých vrstiev enkóderov alebo dekóderov, čo umožňuje modelu učiť sa komplexnejšie vzorce v dátach (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, 2017).

SPÔSOBY ZNEUŽITIA GENERATÍVNEJ AI NA TVORBU MANIPULATÍVNEHO SYNTETICKÉHO OBSAHU

Technológie opísané v predošlej časti tvoria základ širokej škály nástrojov generatívnej AI na tvorbu rôznych typov obsahov vyznačujúcich sa vysokou mierou autenticity. Tieto nástroje sú široko dostupné a jednoduché na používanie. To logicky otvára možnosti aj na neetické spôsoby využitia týchto technológií. Nástroje generatívnej AI využívajúce neurónové siete, difúzne modely či veľké jazykové modely tak možno zneužiť na (Helmus, 2022; Kothari, Orama, Miller, Peeks, Bailey & Alm, 2023; Shoaib, Wang, Ahvanooey & Zhao, 2023):

- Generovanie presvedčivých textov: články, správy, blogy alebo príspevky na sociálnych sieťach, ktoré sú zamerané na šírenie nepravdivých informácií, falošných správ alebo na manipuláciu verejnej mienky. Uveďme niekoľko možných príkladov neetického využitia veľkých jazykových modelov:
    - Vytváranie a šírenie politických propagandistických materiálov, ktoré môžu ovplyvniť verejnú mienku a volebné výsledky.
    - Generovanie veľkého množstva falošných komentárov a príspevkov v prostredí sociálnych médií, ktoré môžu byť zamerané na šírenie dezinformácií alebo na diskreditáciu určitých osôb alebo skupín.
    - Generovanie falošných recenzií produktov alebo služieb na online platformách, čo môže viesť k zavádzajúcemu vnímaniu kvality a ovplyvniť rozhodovanie spotrebiteľov.
    - Generovanie falošných vedeckých štúdií alebo článkov, ktoré môžu byť zneužité na šírenie nepravdivých informácií v akademickej komunite.
    - Generovanie presvedčivých phishingových emailov alebo správ, ktoré cielia na získanie citlivých informácií od jednotlivcov alebo organizácií.
- Vytváranie deepfake obsahu: manipulácia s obrazom a zvukom, čo môže viesť k vytváraniu falošných dôkazov alebo zavádzajúcich naratívov. Uveďme príklady:
    - Snaha o zmenu verejného vnímania dôležitých udalostí alebo osobností, čo môže viesť k zavádzajúcemu chápaniu dôležitých sociálnych a politických otázok.

- Falošné obrázky alebo videá, ktoré zobrazujú celebrity a známe osobnosti v kontroverzných alebo kompromitujúcich situáciách, čo môže poškodiť ich reputáciu a kariéru.
- Falošné správy o politických udalostiach a vyhláseniach, čo môže viesť k vážnym dôsledkom na medzinárodné vzťahy a vnútroštátnu politiku.
- Tvorba profilov na sociálnych sieťach s použitím fotiek alebo videí neexistujúcich ľudí (sockpuppets). Takéto účty potom môžu byť použité na šírenie propagandy, falošných správ, nepravdivých naratívov alebo na tzv. catfishing. Catfishing je podvodná činnosť, pri ktorej si osoba vytvorí fiktívnu identitu na sociálnej sieti alebo „zoznamke“, obvykle zameriavajúc sa na konkrétnu obeť. Podvodník sa snaží postupne od obete získavať informácie, ktoré môže použiť proti nej. Tato praktika môže byť využívaná s cieľom získať finančný zisk, alebo na kompromitáciu (Infoz, [s. a.]).
- Falošné správy o mimoriadnych udalostiach, ako sú prírodné katastrofy alebo teroristické útoky, čo môže viesť k masovej panike a zmätku.
- Falošné dôkazy alebo svedectvá v súdnych procesoch, čo môže mať vážne dôsledky pre právny systém.
- Manipulácia s historickými záznamami (obrazové alebo zvukovo-obrazové materiály), čo môže viesť k zavádzajúcemu alebo skreslenému vnímaniu histórie.

- Syntéza reči: technológie na syntézu reči, ako sú *WaveNet* a *Tacotron*, môžu byť využité na vytváranie falošných audiozáznamov. Tieto technológie dokážu napodobniť hlas konkrétnej osoby a vytvárať audioobsah, ktorý môže byť použitý na šírenie informácií, ktoré síce znejú presvedčivo, ale obsahujú nepravdivé alebo zavádzajúce informácie. Uveďme nasledujúce príklady:
  - Vytváranie falošných audiozáznamov, ktoré môžu byť použité na manipuláciu verejným diskurzom alebo na šírenie zavádzajúcich informácií o dôležitých sociálnych a politických otázkach.
  - Imitácia hlasov verejných osobností na vytváranie falošných vyhlásení alebo správ, čo môže mať vplyv na verejnú mienku a dôveru v médiá.

- Vytváranie podvodných telefonických hovorov, ktoré cielia na jednotlivcov alebo spoločnosti a organizácie s cieľom získať osobné a citlivé informácie alebo finančné prostriedky.
- Vytváranie falošných krízových upozornení alebo varovaní, ktoré môžu spôsobiť paniku a zmätok v spoločnosti.
- Vytváranie falošných hlasových správ alebo zvukových nahrávok, ktoré šíria dezinformácie alebo inak zavádzajúce informácie.

DESKRIPTÍVNA PRÍPADOVÁ ŠTÚDIA

V tejto časti skúmame a analyzujeme viaceré významné príklady zneužitia systémov generatívnej AI na tvorbu falošných správ a dezinformácií vo svete v priebehu posledných rokov.

Metodologické ukotvenie

Výber prípadov (výskumnej vzorky) sme realizovali typom *intenzívneho* a *účelového* vzorkovania (Hendl, 2008). Jednotlivé prípady boli vyberané na základe ich celospoločenského významu, mediálnej pozornosti a schopnosti ilustrovať rôzne aspekty a techniky využitia generatívnej AI na šírenie dezinformácií. Týkajú sa rôznych geografických oblastí a kontextov a umožňujú tak komplexný pohľad na globálny rozsah problému. Prípady sú reprezentatívne pre široké spektrum techník generatívnej AI a ich aplikácie v rôznych kontextoch, od politiky po krízové situácie, čo umožňuje analyzovať rôzne stratégie ich zneužitia a celospoločenské dopady. Dáta boli získavané z verejne dostupných dát z internetu vrátane spravodajských webov, sociálnych médií a oficiálnych vyhlásení. Tento prístup zabezpečuje komplexné a vyvážené informácie o každom prípade. Hlavnou použitou metódou bola *deskriptívna prípadová štúdia s prvkami explanácie a explorácie*, ako čiastkové metódy boli využité *obsahová analýza jednotlivých prípadov, indukcia a interpretácia zistení* (Hendl, 2008). Kvalitatívna analýza umožňuje hlbšie pochopenie motivácií vytvárania a šírenia dezinformácií a ich sociálno-politických dopadov. Prípady boli analyzované s ohľadom na ich kontext, zámery a motivácie, použité techniky a vyvolené efekty.

Deskripcia prípadov

Prípad 1

Jeden z najznámejších prípadov využitia systémov generatívnej AI na tvorbu deepfake obsahu sa odohral v roku 2023. V marci toho roku sa na sociálnej sieti X objavili falošné fotografie Donalda Trumpa, ako sa bráni policajným zložkám, ktoré sa ho pokúšajú zatknúť. Viaceré tieto obrázky vyobrazujú fyzický stret bývalého prezidenta USA s orgánmi vnútornej bezpečnosti, iné ako pred nimi uteká, zatiaľ čo ho policajti prenasledujú. Napriek tomu, že niektoré obrázky mali zjavne viditeľné chyby (napríklad Trump zobrazený s tromi nohami), niektorí ľudia ich považovali za pravdivé, čo naznačuje nedostatok kritického myslenia. K zverejneniu obrázkov,

ktoré mali k januáru 2024 6,8 milióna pozretí, došlo v čase, keď Trump čelil vyšetrovaniu v súvislosti s jeho finančnými a obchodnými aktivitami. Sfalšované obrázky zverejnil Eliot Higgins, zakladateľ open-source zdroja Bellingcat, ktorý sa špecializuje na overovanie faktov. Higgins uviedol, že sa rozhodol vizualizovať situáciu okolo Trumpovho očakávaného obvinenia a jeho príspevok mal mať len zábavný charakter. Využil pritom nástroj generatívnej AI s názvom Midjourney, ktorý slúži na tvorbu obrázkov na základe textového opisu požadovaného výsledku. Použil jednoduché dotazy, ako napríklad: „Donald Trump padá pri zatýkaní“. Higgins sa neskôr vyjadril, že obrázky Trumpovho zatknutia boli v skutočnosti len neformálnym poukázaním na dobré a zlé stránky nástrojov ako Midjourney. „*Predpokladal som, že ľudia si uvedomia, že Donald Trump má dve nohy, nie tri, ale zdá sa, že to niektorých ľudí neodradilo od ich prezentovania ako pravdivých, čo poukazuje na nedostatok kritického myslenia v našom vzdelávacom systéme.*“, uviedol (Press, 2023). Po tom, ako obrázky nabrali na popularite, bol Eliot Higgins vylúčený z komunity Midjourney. Napriek tomu krátko na to začali vznikať podobné obrázky, kontextuálne nadväzujúce na tie predošlé, ako napríklad Trumpov falošný súdny proces či jeho pobyt vo väzení a následný útek. Samotný Trump sa k situácii nevyjadril. Na svojej vlastnej sociálnej sieti Truth Social však zdieľal fotografiu rovnako vytvorenú AI, ktorá ho vyobrazuje, ako kľačí na jednom kolene a modlí sa. Hoci pôvodný príspevok Higginsa môžeme v určitom zmysle považovať za misinformáciu, keďže podľa vlastných slov nečakal, že sa jeho príspevok stane virálnym, tento prípad ilustruje, aké jednoduché je technológie generatívnej AI využiť na tvorbu presvedčivých falošných správ a misinformácií. Sam Gregory, výkonný riaditeľ organizácie na ochranu ľudských práv Witness, v kontexte tohto prípadu uviedol, že došlo k výraznému pokroku v schopnosti vytvárať falošné, ale presvedčivé obrázky vo veľkom množstve. Taktiež uviedol, že je ľahké si predstaviť, ako by takéto aktivity mohli byť vykonávané koordinovane s úmyslom klamať (Stanley-Becker & Nix, 2023). Jevin West, profesor na Univerzite vo Washingtone v Seattli, zameriavajúci sa na šírenie misinformácií, sa zas vyjadril, že šírenie takýchto deepfake médií zvyšuje informačný šum, respektíve prispieva k vytváraniu zmätku a tvorbe nejasností

v krízových situáciách. Tiež uviedol, že to môže viesť k strate dôvery k „systému“ a zníženej schopnosti veriť dostupným informáciám (PBS, 2023).

Obrázok 1 Falošné fotografie Donalda Trumpa (Higgins, 2023)

- Základný kontextuálny rámec: vyšetrovanie známej osoby z politického a business prostredia v USA.
- Použité techniky: vygenerovanie obrazových materiálov pomocou nástroja Midjourney s použitím promptov.
- Zámery a motivácie: pobavenie publika; ilustrácia potenciálu a nebezpečenstiev, ktoré technológie generatívnej AI prinášajú; poukázanie na potrebu kritického myslenia a zvýšenej mediálnej gramotnosti; potenciálne osobné zviditeľnenie.
- Tvorca: Eliot Higgins, zakladateľ investigatívnej skupiny Bellingcat.
- Vyvolaný efekt: mediálny ohlas a zmätok medzi ľuďmi, ktorí si neoverujú online informácie. Spustenie verejnej debaty o nebezpečenstvách využívania nástrojov generatívnej AI. Vylúčenie autora obrázkov z platformy Midjourney.

Prípad 2

Len pár mesiacov po tomto incidente čelil Donald Trump ďalšej podobnej situácii. V júni 2023 jeho rival, guvernér Floridy Ron DeSantis, v rámci svojej politickej kampane uverejnil na sociálnej sieti X video zobrazujúce Trumpove falošné fotografie vygenerované AI. Obrázky znázorňovali Trumpa, ako v prostredí oficiálnych politických podujatí objíma a bozkáva Anthonyho Fauciho, s ktorým mal v tom čase napätý vzťah pre odlišné názory na riešenie pandémie COVID-19. Tvorca videa, ktoré je na sociálnej sieti stále dostupné, a má viac ako 10 miliónov pozretí, sa snažil zvýšiť jeho autenticitu tak, že v ňom zámerne pomiešal skutočné a falošné fotografie do koláže tak, aby to vyzeralo, že všetky fotky pochádzajú z jednej konkrétnej udalosti. Matthew Stamm, profesor elektrotechniky a počítačového inžinierstva na Drexel University, na základe forenznej analýzy naznačil, že obrázky boli vytvorené pomocou difúzneho modelu, ktorý je základom populárnych nástrojov na generovanie obrázkov, ako sú *DALL-E* a *Stability AI* (Ulmer & Tong, 2023). Na druhej strane sa objavili tvrdenia, že aj Trump bol v minulosti zodpovedný za publikovanie zmanipulovaných médií o svojich politických rivaloch, hoci nie na takej sofistikovanej úrovni (Trump, 2023).

- Základný kontextuálny rámec: politická kampaň v USA.
- Použité techniky: využitie difúznych modelov. Konkrétny nástroj, ktorý bol použitý, nie je známy.
- Zámery a motivácie: politické a strategické motívy zamerané na oslabenie dôveryhodnosti a verejného obrazu Donalda Trumpa medzi voličmi. Vyvolanie rozporov a pochybností o rozhodnutiach Trumpa počas pandémie COVID-19.
- Tvorca: Ron DeSantis, politický oponent Donalda Trumpa.
- Vyvolaný efekt: vplyv na politickú debatu. Vyvolanie zmätku a dezinformácií medzi nepozornými voličmi, potenciálne ovplyvňujúc verejnú mienku a vnímanie Donalda Trumpa.

Obrázok 2 Koláž reálnych a falošných obrázkov vo videu na sociálnej sieti X. Falošné fotky môžeme vidieť na obrázkoch 1, 5 a 6 (X, 2023)

Prípad 3

Ďalší celosvetovo známy prípad využitia generatívnej AI na tvorbu vysoko realistických obrázkov sa odohral v marci 2023. Na internete sa objavila falošná fotografia pápeža Františka oblečeného do extravagantnej nafúknutej bielej bundy s kapucňou. Obrázok vytvoril Pablo Xavier pomocou populárneho nástroja využívajúceho AI s názvom Midjourney. Využil pri tom prompty: „Catholic Pope Francis. Balenciaga puffy coat. Streets of Paris“ (Novak, 2024). Xavier sa vyjadril, že tento nápad dostal v momente, keď bol pod vplyvom psychotropných látok. “*Myslel som si, že by bolo vtipné vidieť pápeža v srandovnej bunde*“, vyjadril sa (Stokel-Walker, 2023). Na incident reagovalo diskusné fórum Reddit tak, že Xaviera vylúčilo zo svojej komunity. Krátko nato sa na internete začali objavovať ďalšie falošné obrázky pápeža Františka zachytávajúce ho napríklad vo výrivke v spoločnosti dvoch nahých žien, v koženej bunde so slnečnými okuliarmi a veľkou zlatou reťazou okolo krku, na motorke alebo s potetovanými rukami. Pápež František k incidentu uviedol rozsiahle vyjadrenie. V ňom varoval pred neetickým využívaním umelej inteligencie a vyzval na jej reguláciu. Upozornil na nebezpečenstvo vzniku kognitívneho znečistenia, ktoré môže viesť k skresleniu reality, vyzdvihovaniu falošných naratívov a izolácii ľudí v ideologických informačných bublinách. Ďalej sa vyjadril, že technológie sú ľudskou extenziou a v závislosti od našich rozhodnutí môžu byť použité na dobré alebo zlé účely. Spomenul tiež nebezpečenstvo zneužitia AI v kontexte

vojnových konfliktov, kde tak môžu vznikať „paralelné vojny“ vyznačujúce sa šírením dezinformačných kampaní (Pullella, 2024).

Obrázok 3 Falošné obrázky pápeža Františka (Alphonso, 2023; Keller, 2023; Reddit, 2023)

- Základný kontextuálny rámec: satirické využitie generatívnej AI na tvorbu bizarného obsahu.
- Použité techniky: nástroj Midjourney na základe textových promptov.
- Zámery a motivácie: Zvedavosť a pobavenie v stave zvýšenej kreativity a inšpirovanosti pod vplyvom omamných látok.
- Tvorca: Pablo Xavier, robotník. Ďalšie obrázky vznikli od rôznych ďalších používateľov internetu.
- Vyvolaný efekt: inšpirovanie ďalších ľudí k vytváraniu podobných syntetických médií. Vyvolanie diskusie na sociálnych sieťach a v médiách. Vyvolanie reakcie cirkvi. Xavierovo vylúčenie z komunity Reddit.

Prípad 4

V marci 2022 sa na internete objavilo falošné video zobrazujúce prezidenta Ukrajiny Volodymyra Zelenského, na ktorom z prezidentského pultu vyzýva svojich vojakov, aby zložili zbrane a vzdali sa ruským útočníkom. Video, ktoré bolo zdieľané počas hackerského útoku na ukrajinskú televíziu Ukraine24, a ktoré sa začalo rýchlo šíriť aj v prostredí sociálnych médií, sa považuje za prvé úmyselne použité deepfake video v rámci vojnového konfliktu na Ukrajine. Video je zaujímavé aj tým, že kombinuje manipuláciu obrazu aj zvuku (hlasu). Sociálne médiá ako Facebook, YouTube a X na incident reagovali tak, že video odstránili zo svojich platforiem. Naopak, v prostredí ruských sociálnych médií sa stalo populárne. Ukrajinské ministerstvo obrany neskôr vydalo video, na ktorom Zelenskyj zmanipulované zábery svojej osoby označil za detinskú provokáciu a uviedol, že ukrajinskí vojaci sa nevzdajú. Ukrajinský vládni predstavitelia naznačili, že falošné video bolo vytvorené v rámci ruskej stratégie v kontexte ich informačnej vojny (Allyn, 2022). Hoci sa falošné video Zelenského nevyznačovalo vysokou mierou realistickosti, Nina Schick, autorka publikácie Deepfakes: the coming infocalypse v kontexte tohto prípadu uviedla, že ho môžeme pokladať za predzvesť omnoho sofistikovanejších manipulácií v nadchádzajúcich časoch. Tiež vyslovila predpoklad, že môžeme očakávať, že v budúcnosti sa takéto falzifikáty budú vyrábať ľahšie, pričom budú pôsobiť vysoko autenticky (Pearson & Zinets, 2022).

- Základný kontextuálny rámec: vojnový konflikt medzi Ukrajinou a Ruskom. Politika.

- Použité techniky: konkrétna technológia nie je známa. Môžeme však dedukovať, že tu bola využitá kombinácia techník za pomoci AI s cieľom vykonať syntézu obrazu, manipuláciu so zvukom (rečou) a synchronizáciu pohybov úst s upraveným hlasovým záznamom.
- Zámery a motivácie: šírenie dezinformácií medzi ukrajinským obyvateľstvom a vojskom v snahe demoralizovať a oslabiť odpor proti agresorovi; vyvolanie zmätku, paniky a nedôvery vo vedenie krajiny; vytvorenie falošného obrazu o situácii na Ukrajine vo svete.
- Tvorca: konkrétny autor je neznámy, avšak podľa televízie Ukraine24 išlo o nepriateľských hackerov. Podľa indícií išlo o ruských útočníkov.
- Vyvolaný efekt: Rýchla reakcia ukrajinského prezidenta. Poukázanie na rastúcu hrozbu zneužitia AI technológií v kontexte informačnej vojny. Odstránenie obsahu zo sociálnych médií Facebook, YouTube a X.

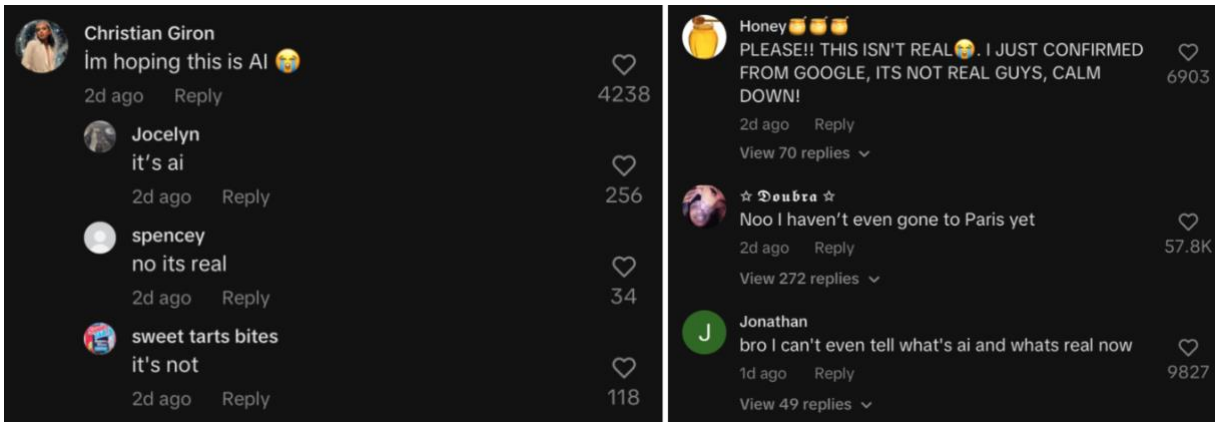
Obrázok 4 Obrázok z deepfake videa Volodymyra Zelenského (Wakefield, 2022)

Prípad 5

V polovici januára 2024 sa na sociálnej sieti TikTok objavila séria falošných, ale realisticky pôsobiacich fotografií a videí zobrazujúcich Eiffelovu vežu v Paríži zasiahnutú plameňmi. Jeden z týchto príspevkov od používateľa @kazakhstandazhasalgan dosiahol 154,8 milióna pozretí a takmer 100 000 zdieľaní v priebehu piatich dní. Pod príspevkom prebiehala rozsiahla debata, či sú fotky skutočné alebo nie. Príspevok obsahoval 131 000 používateľských komentárov, pričom viacero ľudí vyjadrilo presvedčenie, že požiar sa naozaj udial (Kazakhstandazhasalgan, 2024). Pravosť falošnej správy si obozretní používatelia mohli overiť navštívením oficiálnych stránok a sociálnych profilov Eiffelovej veže, na stránke zobrazujúcej

živé video pamiatky alebo monitorovaním médií. Väčšina používateľov teda po overení faktov dospela k záveru, že dané obrázky a videá sú výplodom generatívnej AI (TravelWise, 2024). Je však dôležité spomenúť, že nie všetci ľudia si percipované informácie skutočne overujú a nie všetci si čítajú používateľské komentáre alebo sa zapájajú do online diskusií. Logicky tak v prípadoch týchto ľudí daná falošná správa mohla vyvolať vážne obavy či strach. Niektorí používatelia tiež naznačili, že hoci fotografie vyzerajú veľmi realisticky, je pomerne nepravdepodobné, aby sa celá kovová konštrukcia Eiffelovej veže ocitla v plameňoch tak, ako to zobrazovali zdieľané médiá. Aféra neskôr rozprúdila verejnú debatu o nebezpečenstve a sofistikovanosti technológií generatívnej AI v kontexte generovania deepfake médií a šírenia dezinformácií a misinformácií v online priestore.

Obrázok 5 Falošné fotografie Eiffelovej veže zasiahnutej plameňmi (Kazakhstandazhasalgan, 2024)



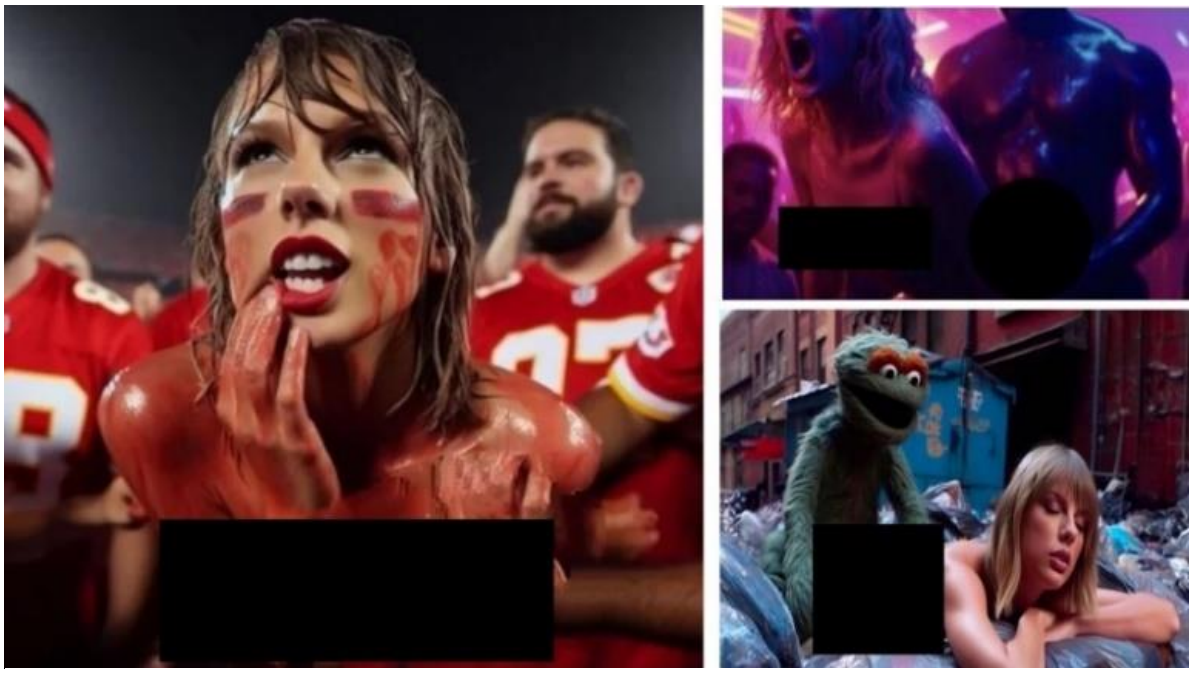
Obrázok 6 Diskusia používateľov sociálnej siete TikTok pri príspevku zobrazujúcom falošné fotografie Eiffelovej veže (Zach & Walker, 2024)

- Základný kontextuálny rámec: falošné (aktuálne) správy z prostredia známej svetovej metropoly.
- Použité techniky: generovanie obrazových materiálov pomocou niektorého z dostupných nástrojov generatívnej AI. Konkrétny použitý nástroj nie je známy.
- Zámery a motivácie: Konkrétne zámery tvorcu obsahu nie sú známe. Dedukujeme však, že mohlo ísť o: zámerné vytvorenie virálneho obsahu s cieľom pozorovať reakcie verejnosti a celospoločenskú diskusiu; osobné zviditeľnenie; alternatívne mohlo ísť o poukázanie na schopnosť generatívnej AI vytvárať vysokorealistické informačné obsahy.
- Tvorca: používateľ sociálneho média TikTok s používateľským menom @kazakhstandazhasalgan. Konkrétne meno danej osoby nie je známe. Prezývka však naznačuje spojenie s Kazachstanom alebo môže byť odkazom na kultúru, jazyk alebo osobnú identitu tvorcu obsahu.
- Vyvolaný efekt: masívna diskusia v online prostredí s cieľom určiť, či ide o skutočnosť alebo nie. Rozprúdenie verejnej debaty o nebezpečenstve a sofistikovanosti technológií generatívnej AI v kontexte generovania deepfake obsahov.

Prípad 6

V tom istom mesiaci (január 2024) boli v prostredí sociálnej sietí X jedným z používateľov publikované vysokorealistické falošné sexuálne explicitné fotografie známej americkej speváčky Taylor Swift. Tieto deepfake syntetické médiá, vytvorené pomocou generatívnej AI, si v priebehu menej ako jedného dňa prezrelo až 47 miliónov ľudí naprieč celým svetom, kým

nedošlo po 17 hodinách k ich odstráneniu. Podľa *404 Media,* boli obrázky vygenerované pomocou nástroja Designer, vlastneného spoločnosťou Microsoft, a následne boli zobchodované cez sociálnu sieť 4chan, odkiaľ prenikli na ďalšie diskusné fóra a sociálne siete ako Telegram, Reddit a X (Ingram, 2024). Obrázky začal šíriť používateľ sociálnych sietí pod menom @ZvBear, pričom podľa média *Marca* ide o skratku pre meno Zubear Abdi. Má ísť o 27-ročného obyvateľa Ontária so somálskym pôvodom. Abdi sa k afére neskôr vyjadril: „*Môj príspevok o Taylor Swift sa stal virálnym a teraz ho všetci šíria*.“ Žartoval, že filmové spoločnosti ako Netflix o ňom môžu spraviť dokument, kde z neho spravia záporného hrdinu (villain). Uviedol tiež, že jeho konanie môže mať za následok prijatie nových zákonov (Kanungo, 2024). Spoločnosti prevádzkujúce systémy generatívnej AI, ako napríklad Microsoft, na škandál rýchlo reagovali a zahájili interné vyšetrovanie. Reakcie online platforiem zahŕňali blokovanie šíreného obsahu a reštrikcie v jeho vyhľadávaní. Pripomeňme, že speváčka je podľa magazínu Forbes jednou z najvplyvnejších žien v súčasnosti s majetkom prevyšujúcim miliardu amerických dolárov (Faguy, 2023). Znepokojenie nad aférou vyjadrila aj tlačová tajomníčka Bieleho domu Karine Jean-Pierre, ktorá v rámci oficiálnej tlačovej konferencie uviedla, že americká administratíva považuje šírenie falošných explicitných obrázkov známej speváčky za alarmujúci jav a vyzvala na prijatie potrebných opatrení (Jean-Pierre, 2024). V rámci kontextu je tiež dôležité uviesť, že viac ako 96 % všetkých vyprodukovaných deepfakes, ktoré sú zverejnené v online prostredí, má pornografický charakter, a takmer všetky z nich sú zamerané na ženské pohlavie (Lodhi, 2024).



Obrázok 7 Deepfake obrázky Taylor Swift (Brandon, 2024; Mitchell, 2024)

- Základný kontextuálny rámec: šírenie falošného pornografického materiálu o známej osobe.
- Použité techniky: niektorý z nástrojov na generovanie obrázkov pomocou generatívnej AI. Pravdepodobne išlo o nástroj Microsoft Designer.
- Zámery a motivácie: Dedukujeme, že zámery tvorcu mohli zahŕňať snahu o vytvorenie senzácie alebo získanie pozornosti prostredníctvom šokujúceho obsahu. Vyjadrenia tvorcu naznačujú, že jednou z motivácií mohlo byť pobavenie alebo osobné zviditeľnenie.
- Tvorca: používateľ sociálnych médií pod prezývkou @ZvBear. Pravdepodobne ide o obyvateľa Kanady s menom Zubear Abdi.
- Vyvolaný efekt: spustenie vyšetrovania spoločností prevádzkujúcich systémy generatívnej AI. Blokovanie šíreného obsahu a reštrikcie v jeho vyhľadávaní v prostredí sociálnych médií. Verejné znepokojenie a spustenie diskusie o etike a dôsledkoch využívania AI na tvorbu deepfake obsahov. Reakcia Bieleho Domu a rozprúdenie debaty o potrebe regulácie a kontroly technológií generatívnej AI.

Prípad 7

Prípad zneužitia systémov generatívnej AI na tvorbu dezinformácií sa nedávno objavil aj na Slovensku. Na konci januára 2024 v prostredí sociálnych sietí začalo kolovať deepfake video zobrazujúce Petra Pellegriniho, lídra politickej strany Hlas, aktuálneho predsedu Národnej rady Slovenskej republiky a kandidáta na prezidenta Slovenskej republiky. Video zachytáva jeho rozhovor s moderátorom RTVS Miroslavom Frindtom. V sfalšovanom rozhovore Pellegrini hovorí o investičných možnostiach do spoločnosti *Solnaft* (skomolenina názvu spoločnosti Slovnaft), pričom uvádza, že ide o dobrú príležitosť pre bežných občanov, ako zarobiť peniaze. Uveďme niekoľko citátov z videa: „*Chcem zlepšiť život našich ľudí, a preto vláda v spolupráci so spoločnosťou Solnaft pripravila program, ktorý umožní každému, aby sa stal finančne gramotným a znásobil svoje peniaze.“ „S pomocou manažéra budete môcť dosiahnuť stabilný príjem vo výške 3000 € v priebehu prvých dvoch až troch mesiacov*.“ „*Odporúčal by som začať s tisíckou eur. To vám umožní dosiahnuť požadovanú úroveň už v prvom mesiaci pri správnom reinvestovaní.“ „Chcel by som dodať, že v roku 2024 urobíme všetko pre zlepšenie finančnej situácie občanov.“* (Weaver, 2024). Obozretní diváci si síce môžu všimnúť znaky nedokonalosti videa a usúdiť tak, že nejde o skutočnosť, nedá sa však vylúčiť, že komunikovaným informáciám niektorí ľudia skutočne uveria. Tak, ako v prípade deepfake videa Volodymyra Zelenského, aj v tomto prípade môžeme pozorovať sofistikovanú manipuláciu tak s obrazom, ako aj so zvukom. Peter Pellegrini sa k prípadu vyjadril takto: „*Umelá inteligencia je vážna vec a ľudia budú musieť byť nielen v prezidentskej, ale aj v ďalšej kampani veľmi obozretní, čomu budú veriť a čomu nie. Môžeme sa len nádejať, že jednotliví politickí protivníci nebudú tí, ktorí za peniaze budú nechávať na svojich protikandidátov takéto videá vyrábať.“* (Pravda, 2024). Pellegrini očakáva, že v ďalších parlamentných voľbách dôjde ešte k výraznejšiemu zneužívaniu umelej inteligencie. „*Dovtedy sa technológia tak zdokonalí, že bežný laik nebude schopný rozlíšiť, čo je pravda a čo nie*,“ dodal (Pellegrini, 2024). „*Žiadny rozhovor s tým obsahom som s Petrom Pellegrinim nenahrával. Celé video je deepfake*,” vyjadril sa k incidentu aj moderátor Miroslav Frindt (O médiách, 2024). Zmanipulované video publikoval používateľ sociálnej siete Facebook s užívateľským menom Alvin Weaver a domovskou lokalitou nastavenou na Thajsko. Na tomto profile je možné vidieť rôzne obrázky kontextuálne viažuce sa na spoločnosť Slovnaft. Či ide o skutočné prepojenie na túto spoločnosť, alebo či ide len o falošný profil, môžeme len

špekulovať. Podkladom pre falošné video bol skutočný rozhovor z roku 2023 odvysielaný RTVS.

- Základný kontextuálny rámec: šírenie dezinformácií o politických predstaviteľoch.
- Použité techniky: konkrétna technológia nie je známa. Môžeme však dedukovať, že tu bola využitá kombinácia techník za pomoci AI s cieľom vykonať syntézu obrazu, manipuláciu so zvukom (rečou) a synchronizáciu pohybov úst s upraveným hlasovým záznamom.
- Zámery a motivácie: konkrétne zámery a motivácie tvorcu videa nie sú známe. Môžeme však dedukovať, že zámery mohli byť rôzne. Najpravdepodobnejším zámerom môže byť snaha o ovplyvnenie verejnej mienky o Petrovi Pellegrinim v čase jeho kandidatúry na post prezidenta Slovenskej republiky. Alternatívne mohlo ísť o pokus o zameranie pozornosti verejnosti na spoločnosť Slovnaft z neznámych dôvodov. Video mohlo byť vytvorené aj ako sociálny experiment na preskúmanie, ako ľudia reagujú na dezinformácie a aký vplyv môže mať falošný obsah na spoločnosť.
- Tvorca: používateľ sociálnej siete Facebook s používateľským menom Alvin Weaver.
- Vyvolaný efekt: vyjadrenie Petra Pellegriniho. Vyjadrenie Miroslava Frindta. Poukázanie na možnosti zneužitia generatívnej AI na diskreditáciu politikov a ovplyvňovanie verejnej mienky. Zdôraznenie výziev spojených s detekciou a reguláciou deepfake obsahov a poukázanie na dôležitosť informačnej a mediálnej gramotnosti a kritického myslenia v digitálnej ére.

Diskusia

Rozoberané prípady ilustrujú sofistikovanosť a presvedčivosť, s akou môže byť generatívna AI využitá na manipuláciu verejnej mienky a politických procesov, čo vyvoláva vážne otázky týkajúce sa etiky, dôvery a pravdy v digitálnom veku. Deskripcia ukázala, že generatívna AI predstavuje významný nástroj v rukách šíriteľov dezinformácií a misinformácií. Schopnosť AI generovať presvedčivý a vizuálne čoraz ťažšie rozoznateľný obsah od skutočnosti umožňuje bezprecedentnú manipuláciu s informáciami. Zistenia prípadovej štúdie majú dôležité teoretické aj praktické implikácie. Teoreticky rozširujú pochopenie dynamiky medzi technologickým pokrokom a jeho vplyvom na spoločnosť. Prakticky poukazujú na urgentnú potrebu vývoja nových technologických, právnych a vzdelávacích stratégií na boj proti

dezinformáciám generovaným pomocou generatívnej AI. Zároveň zdôrazňujú význam etických rámcov pri vývoji a využívaní AI technológií.

Zistenia, ktoré sme prezentovali v rámci jednotlivých prípadov, môžeme sumarizovať do nasledujúceho zhrnutia.

Aspekt	Popis
Kontextuálne rámce	Politika, falošné správy o známych osobách a miestach, informačná vojna.
Techniky a nástroje	Midjourney, Microsoft Designer, iné bližšie neurčené nástroje na manipuláciu s obrazom a zvukom založené na generatívnej AI alebo iných AI technikách.
Zámery a motivácie	Zábava; zvedavosť; vytvorenie senzácie a virálneho obsahu; získanie pozornosti; vplyv na verejnú mienku; oslabenie pozície nepriateľa; verejné poukázanie na potenciál a hrozby spojené s technológiami generatívnej AI; verejné poukázanie na potrebu overovania si informácií dostupných online.
Tvorcovia	– Priznaní identifikovaní tvorcovia – Tvorcovia s aliasom – Nezistiteľní tvorcovia
Tvorba	– Spontánna; náhodná; situačne motivovaná – Do určitej miery plánovaná a koordinovaná
Efekty	Spustenie verejnej debaty o sofistikovanosti AI technológií, nebezpečenstvách ich neetického využívania a o potrebe overovať si informácie v internetovom prostredí. Vyvolanie zmätku v spoločnosti. Vplyv na verejnú mienku. Reakcie oficiálnych politických predstaviteľov, ustanovizní, médií a osôb, ktoré boli predmetom šírených dezinformácií. Vyvolanie online diskusií. Inšpirovanie ľudí k tvorbe podobných obsahov. Poukázanie na riziká zneužitia AI vo vojenských konfliktoch. Spustenie interného vyšetrovania na úrovni spoločností prevádzkujúcich nástroje generatívnej AI. Blokovanie šírených obsahov a zamedzenie ich vyhľadávania.

Tabuľka 1 Sumarizácia hlavných poznatkov prípadovej štúdie

Zistenia prípadovej štúdie ukazujú, že z hľadiska kontextu sa šírenie syntetických obsahov viazalo na známe osoby alebo miesta a aktuálne rozoberané témy v spoločnosti. V rámci použitých techník dominuje využitie verejne dostupných nástrojov generatívnej AI. Zámery a motivácie sú rôzne, od rekreačných až po cielené zneužitie s cieľom vyvolať negatívny účinok. V mnohých prípadoch sú autori šírených obsahov známi, no najmä v kontexte deepfake

videí sú neznámi. Vyvolané efekty sú viacrozmerné. V každom prípade došlo k vyvolaniu verejnej diskusie. V prípade vysokorealistických deepfake obsahov dochádzalo aj k vyvolaniu zmätku a nejasností v spoločnosti, potencionálne vplývajúc na verejnú mienku. Na určité prípady reagovali spoločnosti prevádzkujúce systémy generatívnej AI, aj oficiálni predstavitelia. Jeden z efektov ukazuje možné inšpirovanie sa ľudí od tvorcov a šíriteľov deepfake obsahov s cieľom vykonávať podobné aktivity. Tento fakt považujeme za potencionálne nebezpečný trend, ktorý je pravdepodobne spojený s motívmi, ako sú zvedavosť alebo pobavenie.

Z chronologického hľadiska môžeme pozorovať postupné zvyšovanie realistickosti a autenticity šírených médií, a to predovšetkým v prípade obrázkov. Tento jav nie je prekvapujúci vzhľadom na to, že technológie generatívnej AI sa neustále vyvíjajú a zdokonaľujú. Keďže len v priebehu januára 2024 sme mohli pozorovať až tri významné prípady využitia generatívnej AI na šírenie dezinformácií, vyslovujeme predpoklad, že počet týchto prípadov sa bude v nasledujúcom období zvyšovať, čo vyvoláva seriózne otázky súvisiace s informačnou etikou a digitálnou bezpečnosťou. Vzhľadom na rýchly vývoj systémov AI je tiež logické predpokladať, že rovnako bude stúpať aj realistickosť týchto obsahov. Z pozorovaných prípadov dedukujeme, že šírenie realisticky pôsobiacich falošných správ a dezinformácií, predovšetkým v oblasti aktuálnych dôležitých a citlivých tém, prispieva k nárastu informačného znečistenia a celkovému znižovaniu dôveryhodnosti informácii v online priestore. Ohrozuje jednotlivcov, ktorí sú predmetom šírených dezinformácií, no v širšom kontexte má dopad aj na celospoločenskú situáciu, demokraciu a kritické myslenie ľudí. S neustálym nárastom sofistikovanosti systémov generatívnej AI predpokladáme, že závažnosť tejto problematiky bude v budúcnosti stúpať. Je teda na mieste otázka, ako tento problém riešiť, kým nenapácha škody významnejšieho charakteru.

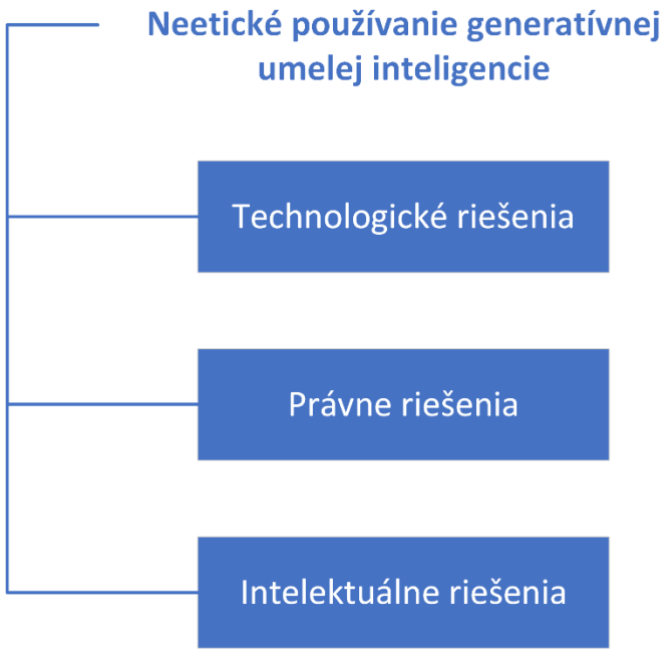
Je dôležité uznať limitácie tejto štúdie, vrátane možnej subjektivity pri výbere prípadov a interpretácii dát. Budúci výskum by sa mohol zamerať na kvantitatívnu analýzu vplyvu syntetických dezinformácií na verejnú mienku alebo na vývoj AI nástrojov schopných detegovať a filtrovať dezinformačný obsah v reálnom čase.

Je potrebné uviesť, že prípady, ktorým sme sa v štúdii venovali, nie sú ojedinelé a existuje ich omnoho viac, tak v zahraničí, ako aj na Slovensku. Falošné správy vytvorené umelou

inteligenciu sa v poslednom období šírili napríklad aj o Borisovi Kollárovi, Michalovi Šimečkovi, Róbertovi Ficovi, Zuzane Čaputovej, Andrejovi Babišovi, Andrejovi Kiskovi, Milošovi Zemanovi, Joe Bidenovi, raperke Nicki Minaj a hercovi Tomovi Hollandovi, či o jednom z najznámejších tvorcov na platforme YouTube, MrBeastovi. V rámci ďalších príkladov falošných správ vytvorených generatívnou AI uveďme požiar na Havaji, explóziu v Gaze, explóziu v Pentagone alebo obrázky obetí v Izraeli a na Ukrajine. Systémy generatívnej AI možno neeticky zneužiť aj v oblastiach kybernetického zločinu, finančných podvodov a scammingu, syntetickej biometrie (vytváranie falošných biometrických údajov, ako sú tváre, odtlačky prstov alebo hlasové vzory, ktoré môžu byť použité na obchádzanie bezpečnostných systémov alebo na neoprávnený prístup k citlivým informáciám (Kaspersky, 2024)), sledovania a profilovania obyvateľstva bez ich súhlasu alebo v oblasti šírenia detskej pornografie či násilných obsahov.

MOŽNÉ RIEŠENIA PROBLEMATIKY

V kontexte rastúcej sofistikovanosti generatívnej AI a jej potenciálu na šírenie dezinformácií a falošných správ v internetovom prostredí je nevyhnutné prijať komplexný prístup, ktorý zahŕňa technologické riešenia, právne regulácie aj intelektuálne prístupy.



Obrázok 8 Znázornenie riešení problematiky neetického používania generatívnej AI

Technologické riešenia

Johannes Tammekänd, expert na kyberbezpečnosť, dátovú vedu a informačnú vojnu, predtým pracujúci pre *NATO Cooperative Cyber Defence Centre of Excellence* (CCDCOE) založil spoločnosť *Sentinel*. Tá vyvinula nástroj, ktorý pomocou neurónových sietí, hlbokého učenia a konvolučných sietí umožňuje s vysokou presnosťou identifikovať deepfake informačné obsahy (napríklad obrázky alebo videá) vytvorené pomocou generatívnej AI. Technológia dokáže rozoznávať tzv. biologické signály, ako napríklad mimiku tváre, hlasové prejavy alebo prirodzený jazyk. Systém umožňuje používateľom nahrať digitálne médiá, ktoré majú byť predmetom analýzy, prostredníctvom špeciálnej webovej stránky alebo API. Nástroj na základe automatizovanej analýzy určí, či sú nahrané médiá deepfake obsahom alebo nie, a poskytne podrobnú správu o zisteniach, vrátane vizualizácie konkrétnych oblastí, v ktorých bolo

s obsahom manipulované. Použitý algoritmus je účinný pri identifikácii falošného obsahu, dezinformácií či nepravdivých informácií, čo umožňuje platformám bojovať proti manipulácii s informačným obsahom v digitálnom prostredí. Sentinel spolupracuje s vládami, médiami a bezpečnostnými agentúrami s cieľom chrániť demokracie pred dezinformačnými kampaňami, syntetickými médiami a nebezpečnými informačnými operáciami a chrániť integritu digitálnych médií (Romain, 2023; Sentinel, 2024).

Spoločnosť Google vyvinula technológiu *SynthID*, ktorá umožňuje vkladať nedetekovateľnú digitálnu vodotlač priamo do pixelov obrázkov vygenerovaných pomocou generatívnej AI. Tieto prvky sa tak stávajú nerozoznateľnými ľudským zrakom. Výsledkom je vznik stopy, ktorou možno jednoznačne určiť, či bol obrázok vygenerovaný pomocou AI alebo nie, a to dokonca aj v prípade, ak bude obrázok ďalej editovaný alebo skopírovaný pomocou snímky obrazovky. Tento ochranný prvok tak umožňuje vytvoriť bariéru pred neetickým použitím systémov generatívnej AI. Integrácia tejto technológie s Google Cloud naznačuje jej integráciu do širšieho ekosystému cloudových služieb a dostupnosť pre širokú škálu používateľov a organizácií. Okrem identifikácie AI-generovaných obrázkov má technológia SynthID potenciál byť použitá aj v iných oblastiach, vrátane generovania hudby a iných typov digitálneho obsahu, čím sa ďalej rozširujú možnosti jej aplikácie (Google DeepMind, [s. a.]; Gowal & Kohli, 2023).

Ďalší obranný nástroj s názvom *PhotoGuard* vyvinuli výskumníci z Massachusettského technologického inštitútu. Umožňuje chrániť jednotlivcov pred neoprávneným zneužitím ich podoby prostredníctvom umelej inteligencie, najmä v kontexte generovania deepfake videí a obrázkov. Technológia umožňuje na obrázky aplikovať drobné zmeny, neviditeľné pre ľudské oko. Keď sa technológie generatívnej AI, ako je napríklad generátor obrázkov Stable Diffusion, pokúsia manipulovať s obrázkom chráneným technológiou PhotoGuard, výsledok bude vyzerať nerealisticky, skreslene alebo deformovane, čím sa zníži presvedčivosť a autenticita vygenerovaného materiálu. Ide tak o efektívny spôsob, ako predchádzať vytváraniu presvedčivých deepfake videí a obrázkov, ktoré môžu byť použité na šírenie dezinformácií alebo na poškodzovanie reputácie. Nástroj môže byť použitý na rôzne typy obrázkov, od osobných fotografií až po umelecké diela (Gordon, 2023 & Heikkilä, 2023).

Na Univerzite of Chicago vznikol podobný nástroj, nazvaný *Nightshade*, ktorý slúži na ochranu digitálnych obrázkov pred neoprávneným použitím v AI systémoch. Nightshade umožňuje na

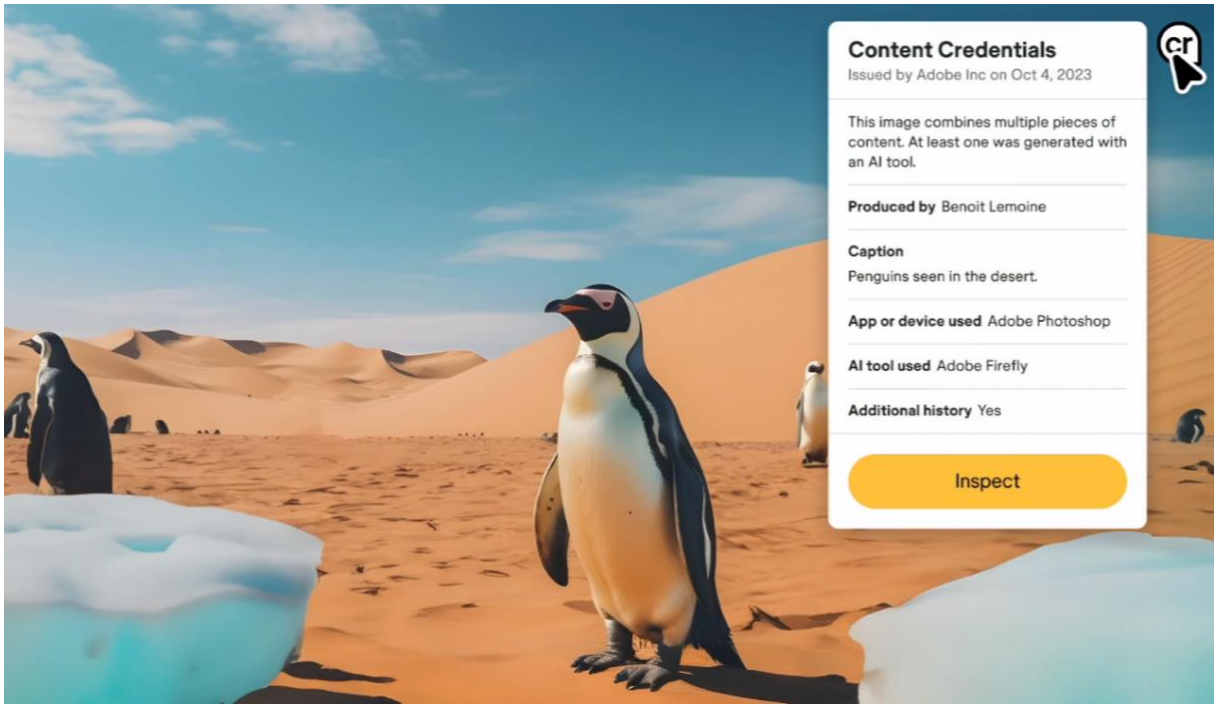
obrázky aplikovať neviditeľnú vrstvu úprav, tzv. jed (poison). Tieto úpravy sú pre ľudské oko nedetekovateľné, ale spôsobujú, že AI modely budú generovať nesprávne alebo deformované výstupy. Primárnym cieľom tejto technológie je ochrana autorských práv umelcov a fotografov, ktorých diela môžu byť bez ich súhlasu použité technologickými spoločnosťami na tréning generatívnych AI modelov. Technológiu je však možné aplikovať na akýkoľvek typ obrázka, čo umožňuje širšie možnosti jej uplatnenia (Heikkilä, 2023a & Masterson, 2023).

Príkladom ďalšej technológie na ochranu jednotlivcov je nástroj *Fawkes*, vyvinutý výskumníkmi z University of Chicago. Ide o obranný mechanizmus proti neoprávnenému použitiu fotografií v systémoch rozpoznávania tvárí. Nástroj na digitálne obrázky tvárí aplikuje tzv. masku. Táto stopa je pre ľudské oko takmer neviditeľná, ale zároveň dostatočne výrazná na to, aby zmiatla algoritmy rozpoznávania tvárí. Cieľom je zabrániť tomu, aby boli obrázky jednotlivcov bez ich súhlasu použité na tréning alebo vylepšenie databáz rozpoznávania tvárí. Technológia tiež umožňuje znížiť riziko neoprávneného sledovania a identifikácie. Fawkes tak poskytuje jednotlivcom nástroj na ochranu ich digitálnej identity a súkromia v online priestore. Môže byť použitý na akékoľvek digitálne obrázky tvárí, čím umožňuje jednotlivcom chrániť ich fotografie pred zneužitím v rôznych kontextoch, od sociálnych médií až po verejné databázy (Shan, Wenger, Zhang, Li, Zheng & Zhao, [s. a.]).

Ďalšou možnosťou boja proti neetickému používaniu systémov generatívnej AI je implementácia prísnejších autentifikačných a autorizačných procesov v snahe o účinnejšie regulovanie príkazov, pomocou ktorých používatelia vytvárajú obsahy. Tieto snahy o zabezpečenie etickejšieho a zodpovednejšieho používania generatívnej AI sú nevyhnutné na ochranu pred negatívnymi dôsledkami týchto technológií, akými sú šírenie dezinformácií, deepfake obsahov a iných druhov manipulatívneho obsahu a na zabezpečenie pozitívneho vplyvu systémov generatívnej AI na spoločnosť (Baxter & Schlesinger, 2023; Giordani, 2023). V tomto kontexte je však potrebné uviesť, že vynaliezaví používatelia s cieľom manipulovať s výstupmi generatívnej AI alebo v snahe o vyvolanie nežiaduceho správania týchto technológií, vedia spomenuté ochranné metódy obísť pomocou metód tzv. *prompt hackingu*, akými sú napríklad *prompt injection*, *prompt leaking*, *adversarial prompting* alebo *jailbreaking* (Gupta, 2023).

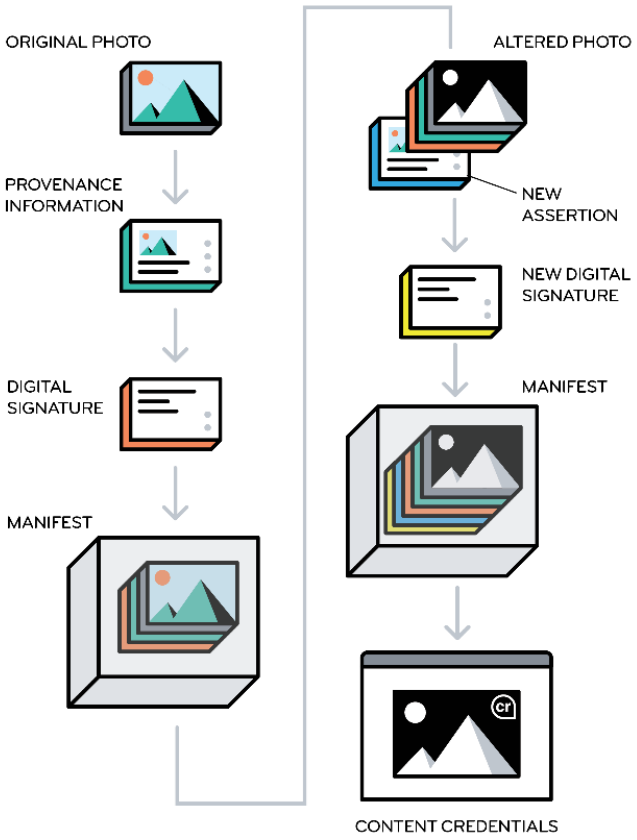
Ako alternatívna možnosť riešenia problematiky informačného znečistenia spôsobeného využívaním systémov generatívnej AI sa objavuje nový prístup tzv. obsahových autentifikátorov/certifikátov obsahu. Technológia *Content Credentials* je založená na technickej špecifikácii *C2PA* (The Coalition for Content Provenance and Authenticity). Ide o projekt vyvinutý v kolaborácii spoločností Adobe, Arm, Intel, Microsoft a Truepic. C2PA zjednocuje úsilie iniciatív *Adobe-led Content Authenticity*, ktorá sa zameriava na systémy poskytujúce kontext a históriu digitálnych médií a iniciatívy *Project Origin*, vedenú spoločnosťami Microsoft a BBC, ktorá sa zaoberá dezinformáciami v digitálnom spravodajskom ekosystéme (Coalition, 2024).

Technológia funguje tak, že do obrázkov/médií pridáva špeciálny znak – ikonu tvorenú písmenami *CR*. Po kliknutí na ikonu sa rozbalí okienko (menu), v ktorom sa budú nachádzať detaily o danom informačnom objekte, ako napríklad: pôvodný autor/tvorca, či a aké konkrétne AI technológie boli použité, aký softvér bol použitý na úpravu a tiež vecný opis daného média (obsahový popis). Pomocou tlačidla *Inspect* je používateľ navedený na separátnu webovú stránku, ktorá ponúka detailnejší rozbor (napríklad podrobná chronológia – vznik média, jednotlivé úpravy). Stránka zároveň umožňuje nahrávať vlastné médiá (napríklad z počítača alebo telefónu) a overiť ich autenticitu (C2PA, 2023).



Obrázok 9 Príklad aplikácie obsahového autentifikátora do obrázkov na internete (C2PA, 2023)

Na zabezpečenie transparentnosti, overiteľnosti a nezmeniteľnosti informácií o pôvode a histórii digitálneho obsahu využíva systém Content Credentials technológiu *blockchain*, postavenú na decentralizovaných sieťach. Funguje na reťazci blokov, ktoré sú spojené prostredníctvom jedinečného identifikátora nazývaného *hash*. Zmena akéhokoľvek jednotlivého bloku si vyžaduje zmeny vo všetkých nasledujúcich blokoch, čo umožňuje jednoducho identifikovať pokus o manipuláciu. V Content Credetials ekosystéme je teda originálne médium spojené s jeho digitálnym podpisom – metadátami. Ak dôjde k úprave daného informačného objektu, tak zároveň dôjde k automatickej aktualizácii týchto metadát, pričom s týmto systémom nie je možné manipulovať (Hendrickson, 2024).



Obrázok 10 Princíp fungovania Content Credentials (Strickland, 2023)

Nástroj pomocou identifikátorov autenticity umožňuje používateľom overiť si pôvod prezeraného obsahu a informuje ich o tom, či je daný informačný objekt pôvodný alebo syntetický, respektíve či pri jeho vzniku došlo k použitiu ďalšieho softvéru. Prítomnosť

takéhoto identifikátora vo všetkých informačných objektoch na internete môže významne prispieť k zvýšeniu dôveryhodnosť informácií v internetovom prostredí. Podľa informácií z metadát sa používateľ môže rozhodnúť, akú váhu komunikovanému obsahu prisúdi. Naopak, absencia tohto ochranného prvku môže vyzývať k zvýšenej miere skepticizmu voči prezeranému obsahu (Hendrickson, 2024).

Na druhej strane však existujú obavy, že s použitím sofistikovaných či vynaliezavých metód alebo s využitím budúcich pokročilejších nástrojov založených na AI bude pravdepodobne stále možné obísť väčšinu z ochranných technických mechanizmov. To znamená, že na maximalizovanie ich účinnosti bude potrebné tieto ochranné prostriedky neustále vylepšovať a zdokonaľovať, prípadne vymýšľať nové, aby boli regulačné mechanizmy v ideálnom prípade vždy o krok vpred pred snahami o neetické využívanie technológii generatívnej AI.

Právne riešenia

S rýchlym rozvojom a implementáciou AI technológií sa ich právna regulácia stáva významnou témou s cieľom riešiť etické, právne i technické výzvy.

Európska únia

V rámci aktuálne svetovo najvýznamnejšieho politického riešenia spomeňme akt Európskeho parlamentu a Európskej rady o umelej inteligencii (AI act), ktorý tvorí zatiaľ najkomplexnejší zákon regulujúci AI. Cieľom tohto zákona je zabezpečiť ochranu základných práv, demokracie, právneho štátu a environmentálnej udržateľnosti pred rizikami spojenými s umelou inteligenciou a zároveň podporiť jej zodpovedný vývoj a používanie. Zákon sa zameriava na rôzne aspekty používania AI, vrátane etických a bezpečnostných štandardov. Európska únia zároveň v rámci svojej digitálnej stratégie chce AI regulovať tak, aby zabezpečila lepšie podmienky pre jej vývoj a využívanie. Konečnú verziu zákona ešte musí schváliť nadpolovičná väčšina Európskeho parlamentu (Parliament, 2023).

Systémy umelej inteligencie sa budú klasifikovať do štyroch kategórií rizika (nízke, obmedzené, vysoké a neprijateľné). Každá kategória rizika bude mať svoje vlastné požiadavky, od povinností zachovávať princíp transparentnosti pre systémy s nízkym rizikom až po prísne požiadavky pre systémy s vysokým rizikom. Systémy AI, ktoré predstavujú neprijateľné riziko, ako napríklad systémy sociálneho skórovania, budú zakázané. Poskytovatelia a používatelia AI systémov budú mať rôzne povinnosti, vrátane požiadaviek na transparentnosť, dokumentáciu

a dodržiavanie predpisov o ochrane osobných údajov. Nad dodržiavaním regulačného rámca bude dohliadať nezávislý orgán EÚ (Commission, 2024).

Špecifické povinnosti pre modely generatívnej AI (Barani & Dyck, 2023):

- Poskytovateľ týchto systémov musí jasne a presvedčivo ukázať, aké opatrenia boli prijaté na minimalizáciu potenciálnych rizík pre zdravie, bezpečnosť, základné práva, životné prostredie, demokraciu a právny štát.
- Dáta určené na trénovanie modelov generatívnej AI musia podliehať prísnym pravidlám zaručujúcim ich vhodnosť a nezaujatosť.
- Modely musia byť navrhnuté, vyvinuté a testované s cieľom zabezpečiť ich výkonnosť, predvídateľnosť, interpretovateľnosť, opraviteľnosť a bezpečnosť počas celého životného cyklu.
- Pri vývoji a používaní modelov sa musia dodržiavať štandardy na zníženie spotreby energie, zdrojov a odpadu.
- Poskytovateľ musí vypracovať technickú dokumentáciu a zrozumiteľné pokyny pre systémy generatívnej AI. Túto dokumentáciu je povinný uchovávať po dobu 10 rokov od uvedenia modelu na trh a sprístupniť ju príslušným orgánom.
- Poskytovateľ musí implementovať systém manažérstva kvality, ktorý zaručí súlad s požiadavkami, ktoré AI Act predkladá.
- Modely musia byť zaregistrované v databáze EÚ, ktorá umožní dohľad nad ich vývojom a používaním.

„*Potrebujeme jasné pravidlá pre fungovanie umelej inteligencie. Verím, že nikto z nás by nechcel byť obeťou možných nesprávnych krokov umelej inteligencie. Pretože áno, aj stroje sa môžu mýliť. Na rozdiel od ľudí však za to nemajú ako prebrať zodpovednosť, ale dopad na naše životy môže byť až príliš reálny,*“ uviedol poslanec Európskeho parlamentu a podpredseda Progresívneho Slovenska Martin Hojsík (Parlament, 2023).

Podľa Lucie Ďuriš Nicholsonovej je potrebné technológie založené na AI regulovať tak, aby zostali „fantastickým nástrojom“ v rukách ľudí a nie hrozbou, napríklad pri šírení hoaxov a dezinformácií. „*Pre fungovanie umelej inteligencie potrebujeme pravidlá. EÚ bude prvá, kto také pravidlá prijme a preto sa na nás teraz díva celý svet. Nová legislatíva musí zabezpečiť,*

aby boli systémy umelej inteligencie uvádzané na trh EÚ bezpečné a aby rešpektovali existujúce právne predpisy, základné práva a hodnoty Únie. Zároveň musíme pravidlá nastaviť tak, aby nezahatali cestu inováciám a investíciám do nových technológií. Technológia ide dopredu míľovými krokmi a legislatíva zaostáva. Dúfam, že kvalitná dohoda sa nájde čím skôr, lebo ak sa budeme naťahovať príliš dlho, legislatíva bude zastaraná už v momente, keď ju schválime.“, uviedla (TASR, 2023).

„*Vývoj umelej inteligencie bude tak či tak rýchlo napredovať, a preto je nevyhnutné, aby sme stanovili jasné pravidlá, ochránili súkromie a najmä zakázali škodlivé aktivity, kde sa umelá technológia využíva. Takéto zneužívanie umelej inteligencie je totiž zásah do súkromia spojený s obrovskými rizikami úniku informácií a ich zneužitia, a to nielen dospelých, ale aj maloletých a najzraniteľnejších,“* hovorí europoslanec Robert Hajšel (TASR, 2023).

V kontexte vyjadrení slovenských politikov uveďme aj oficiálne vyjadrenie Polície Slovenskej republiky, ktorá tiež upozornila na neetické využívanie generatívnej AI. Koncom septembra 2023 vo svojom príspevku na sociálnej sieti Facebook uviedla, že: „*Policajný zbor zachytil vo verejnom priestore niekoľko dezinformačných a manipulatívnych videí a audionahrávok súvisiacich so sobotňajšími voľbami do Národnej rady SR. Ide napríklad o účelovo zostrihané videá, zmanipulované audionahrávky, produkty zneužité technológiou deep fake na falošné a nepravdivé informácie. Polícia apeluje na občanov, aby boli v online priestore obozretní a nenechali sa zneužiť záujmovými skupinami, ktoré chcú prostredníctvom klamstiev dosiahnuť svoje vlastné ciele.“* (Polícia, 2023). Ďalej sa Polícia Slovenskej republiky na sociálnej sieti vyjadrila, že v blízkej budúcnosti môžeme očakávať stúpajúcu tendenciu výskytu zmanipulovaných videí a audionahrávok.

Druhým novým regulačným rámcom v Európskej únii je *Akt o digitálnych službách*, ktorý sa týka online platforiem a digitálnych služieb, a ktorý nadobudol účinnosť 17. februára 2024. Tento právny rámec predstavuje významný krok v boji proti nezákonnému obsahu, dezinformáciám a zneužívaniu online platforiem (Ministerstvo, 2024).

Hlavnými cieľmi tejto regulácie sú:

- Zvýšenie ochrany používateľov prostredníctvom prísnejších pravidiel pre odstraňovanie nezákonného obsahu, ako sú nenávistné prejavy, teroristický obsah a dezinformácie.

- Zavedenie prísnejších povinností pre online platformy, aby boli zodpovedné za obsah zverejňovaný na ich platformách. To zahŕňa povinnosti transparentnosti, dohľadu a implementácie systémov na nahlasovanie a odstraňovanie nezákonného obsahu.
- Zamedzenie zneužívania dominantného postavenia veľkých online platforiem a podpora spravodlivejšej hospodárskej súťaže v online prostredí.

Online platformy sa na základe tohto rámca klasifikujú do štyroch kategórií (veľmi veľké online platformy, online platformy, hostingové služby a sprostredkovateľské služby). V závislosti od kategórie sa musia dodržiavať rôzne povinnosti (Komisia, [s. a.]).

Tieto povinnosti zahŕňajú (Regulation, 2022):

- Povinnosť transparentnosti: platformy musia zverejňovať informácie o svojich algoritmoch, pravidlách moderovania a mechanizmoch na nahlasovanie nezákonného obsahu.
- Povinnosť dohľadu: platformy musia zaviesť systémy na proaktívne vyhľadávanie a odstraňovanie nezákonného obsahu.
- Povinnosť reagovať: platformy musia rýchlo reagovať na nahlásenia nezákonného obsahu a podniknúť kroky na jeho odstránenie.
- Zavedenie nezávislého dohľadu: nad dodržiavaním regulácie bude dohliadať nezávislý orgán EÚ.
- Posilnenie právomocí národných orgánov: národné orgány členských štátov EÚ budú mať posilnené právomoci na vymáhanie dodržiavania týchto pravidiel.

Očakávaný vplyv novej regulácie sa týka zníženia množstva nezákonného obsahu v online prostredí (napr. nenávistné prejavy a dezinformácie), zvýšenia zodpovednosti online platforiem, zmenšenia vplyvu veľkých online platforiem a posilnenia postavenie menších konkurentov a zabezpečenia bezpečnejšieho online prostredia.

Akt o digitálnych službách tvorí komplexný a ambiciózny regulačný rámec, ktorý má potenciál výrazne ovplyvniť budúcnosť internetu v EÚ. Úspešnosť jeho implementácie bude závisieť od efektívneho vymáhania dodržiavania pravidiel a od toho, ako sa platformy aj používatelia novým pravidlám prispôsobia.

Snahy o právnu reguláciu AI technológií sa objavujú aj v ostatných častiach sveta. Uveďme niekoľko príkladov.

USA

V Spojených štátoch amerických sa regulácie AI formujú na federálnej úrovni prostredníctvom iniciatív, zákonov a politík z Bieleho domu, Kongresu a rôznych federálnych agentúr. Tieto snahy sú zamerané na zabezpečenie, aby boli AI technológie vyvíjané a používané spôsobom, ktorý podporuje inovácie, ale zároveň chráni občanov a spoločnosť pred potenciálnymi rizikami. Biely dom vydal sériu výkonných príkazov a strategických usmernení. Tieto iniciatívy sú zamerané na podporu výskumu a vývoja AI, zabezpečenie etického používania AI a ochranu amerických technologických inovácií. Kongres pristupuje k legislatíve a politike AI postupne, pričom sa zameriava na špecifické oblasti, ako sú autonómne vozidlá a využitie AI v národnej bezpečnosti. Medzi kľúčové legislatívne iniciatívy patrí *Zákon o národnej iniciatíve AI* z roku 2020, ktorý sa zameriava na rozšírenie výskumu a vývoja AI a koordináciu činností medzi obrannými a civilnými federálnymi agentúrami. Federálne agentúry, vrátane Federálnej obchodnej komisie (FTC) a Národného inštitútu pre štandardy a technológiu (NIST), hrajú v regulácii AI kľúčovú úlohu. FTC sa zameriava na reguláciu a presadzovanie pravidiel, zatiaľ čo NIST vypracoval rámec pre riadenie rizík, ktorý slúži ako dôležitý pilier federálnej správy AI (Fazlioglu, 2023).

Kanada

Významný krok v regulácii AI v Kanade predstavuje *Zákon o umelej inteligencii a dátach* (Artificial Intelligence and Data Act – AIDA). Tento zákon bol súčasťou Návrhu zákona C-27 predstaveného 16. júna 2022 a predstavuje prvý pokus Kanady o reguláciu AI. Cieľom AIDA je zabezpečiť, aby AI systémy, ktoré ovplyvňujú životy Kanaďanov a operácie kanadských podnikov, boli bezpečné a aby rešpektovali štandardy týkajúce sa bezpečnosti a ľudských práv, na ktoré sú kanadskí občania zvyknutí. Tento rámec je prvým krokom k novému regulačnému systému, ktorý má usmerňovať inovácie v oblasti AI pozitívnym smerom a podporovať zodpovedné prijatie AI technológií Kanaďanmi a kanadskými podnikmi. Kanada tiež na zosúladení prístupov spolupracuje s Európskou úniou, Spojeným kráľovstvom a USA (Government, [2022].)

Spojené kráľovstvo

UK sa zameriava na vytvorenie regulačného rámca, ktorý bude na jednej strane podporovať inovácie v oblasti AI a zároveň chrániť občanov pred jej potenciálnymi rizikami. Vláda preto

zriadila *Centrum pre dátovú etiku a inovácie* (Centre for Data Ethics and Innovation – CDEI), ako súčasť Ministerstva pre vedu, inovácie a technológie. CDEI sa zameriava na podporu etického využívania dát a AI technológií v súkromnom aj verejnom sektore, pričom kladie dôraz na transparentnosť, spravodlivosť a zodpovednosť (Centre, [s. a.]).

Čína

V Číne bol v apríli 2023 predstavený návrh *Administrative Measures for Generative Artificial Intelligence Services*, ktorý predstavuje prvý pokus o reguláciu služieb generatívnej AI v tejto krajine. Návrh bol zverejnený Čínskou národnou administratívou pre kybernetický priestor. Návrh zákona je významným krokom pre Čínu v oblasti regulácie generatívnej AI, ktorý odráža globálny trend smerom k zabezpečeniu, aby boli AI technológie vyvíjané a používané zodpovedne a v súlade s etickými štandardmi. Zároveň tiež vyvoláva diskusie o tom, ako najlepšie vyvážiť podporu inovácií s potrebou ochrany verejnosti a zabezpečenia transparentnosti a zodpovednosti v AI ekosystéme. Návrh zákona sa zaoberá otázkami, ako sú povinnosti poskytovateľov a používateľov služieb generatívnej AI, a zdôrazňuje potrebu vyváženia medzi podporou inovácií a zabezpečením bezpečnosti, transparentnosti a etického používania AI (Luo, Dan & Shepherd, 2023).

Singapur

V novembri 2019 predstavil Singapur svoju *Národnú stratégiu umelej inteligencie* (Artificial Intelligence and Data Act) ako súčasť širšieho úsilia o posilnenie svojej pozície ako globálneho centra pre inovácie a vývoj v oblasti AI. Táto stratégia je zameraná na podporu výskumu a vývoja v AI, zvýšenie adopcie AI technológií v rôznych sektoroch a zabezpečenie etického a bezpečného používania AI. Jedným z kľúčových prvkov singapurskej stratégie je vývoj a implementácia etických usmernení, ktoré slúžia ako základ pre reguláciu a správu AI technológií. Stratégia sa snaží vytvoriť udržateľný ekosystém, ktorý podporuje inovácie a zároveň chráni práva a súkromie občanov. Singapur tiež podporuje medzinárodnú spoluprácu a partnerstvá v oblasti AI, aby sa zabezpečila globálna harmonizácia regulačných prístupov a podporila medzinárodná výmena najlepších postupov (Remolina & Seah, 2019).

Japonsko

V roku 2016 japonská vláda a Japonská obchodná federácia (Keidanren) predstavili koncept *Society 5.0* s cieľom transformovať japonskú spoločnosť prostredníctvom integrácie pokročilých technológií, vrátane AI, do rôznych aspektov každodenného života. Society 5.0 predstavuje ambicióznu stratégiu na integráciu pokročilých technológií do každodenného života s cieľom riešiť sociálne výzvy a zlepšiť kvalitu života. Táto stratégia, ktorá je súčasťou širšieho plánu na transformáciu Japonska na superinteligentnú spoločnosť, zdôrazňuje význam etických a právnych aspektov pri vývoji a implementácii AI. Society 5.0 sa zameriava na vytvorenie spoločnosti, kde digitálne technológie, ako je AI, zlepšujú ekonomickú efektivitu a zároveň riešia sociálne problémy, ako sú starnutie populácie a zmena klímy. Stratégia zdôrazňuje potrebu vyváženého prístupu k regulácii AI, ktorý podporuje inovácie a zároveň zabezpečuje ochranu súkromia, bezpečnosť a etické používanie technológií (Cabinet, [s. a.]; Shimpo, 2020).

Austrália

V roku 2019 iniciovalo Ministerstvo pre priemysel, vedu a zdroje (Department of Industry, Science, Energy and Resources) v Austrálii etické princípy AI ako súčasť širšieho úsilia o podporu etického vývoja a používania AI v krajine. Tieto princípy boli navrhnuté tak, aby poskytovali usmernenia pre podniky a vlády pri návrhu, vývoji a implementácii AI technológií. Cieľom je podporovať inovácie a technologický pokrok a zároveň zabezpečiť, že vývoj, implementácia a používanie AI technológií budú prebiehať bezpečne, spoľahlivo a eticky. Tieto princípy možno zhrnúť nasledovne (Department, [s. a.]):

- AI systémy by mali prinášať prospech jednotlivcom, spoločnosti a životnému prostrediu.
- AI by mala rešpektovať ľudské práva, diverzitu a autonómiu jednotlivcov.
- AI by mala byť inkluzívna, prístupná a nemala by viesť k nespravodlivej diskriminácii.
- AI systémy by mali chrániť súkromie jednotlivcov a zabezpečiť ochranu a bezpečnosť zbieraných a spracovávaných údajov.
- AI systémy by mali byť spoľahlivé a bezpečné, fungovať v súlade s ich zamýšľaným účelom a mali by minimalizovať riziká spojené s ich používaním.
- Používatelia by mali mať možnosť pochopiť a byť informovaní o tom, ako AI systémy fungujú a ako sú rozhodnutia AI systémov odôvodnené.

- Keď AI systém významne ovplyvňuje jednotlivca alebo skupinu, mal by existovať proces, ktorý umožňuje týmto osobám brániť sa proti použitiu AI systémov.
- Osoby zodpovedné za rôzne fázy životného cyklu AI systému by mali byť identifikovateľné a zodpovedné za výsledky AI systémov.

Intelektuálne riešenia

Boj proti dezinformáciám a falošným správam vytvoreným pomocou AI si vyžaduje viac než len technologické a právne riešenia. Vyžaduje si rozvoj a podporu intelektuálnych kapacít jednotlivcov. Ako kľúčové oblasti pre budovanie spoločnosti odolnejšej proti dezinformáciám generovaným pomocou AI identifikujeme tieto intelektuálne prístupy: vzdelávanie v oblasti informačnej a mediálnej gramotnosti, podpora kritického myslenia, budovanie kolektívnej inteligencie, podpora kultúry osobnej zodpovednosti, podpora digitálnej etiky a digitálneho občianstva a využívanie umeleckých a naratívnych prístupov.

Výskumy ukazujú, že zvýšenie *informačnej a mediálnej gramotnosti* je kľúčové pre boj proti dezinformáciám. Táto intelektuálna kompetencia zahŕňa schopnosť rozpoznať rozdiel medzi pravdivými a falošnými informáciami, pochopiť, ako médiá vytvárajú správy a obsah, a byť si vedomý techník, ktoré môžu byť použité na manipuláciu verejnej mienky. Vzdelávanie v tejto oblasti môže pomôcť ľuďom lepšie sa navigovať v digitálnom informačnom prostredí a rozvíjať kritické myslenie potrebné na identifikáciu a odmietnutie dezinformácií. Programy, ktoré učia používateľov rozpoznávať a kriticky hodnotiť obsah, tak môžu výrazne prispieť k zníženiu vplyvu falošných správ. Ako príklad uveďme projekt *Learn to Discern* na Ukrajine, ktorý bol vyvinutý ako odpoveď na hybridnú ofenzívu Ruska. Tento projekt ukázal, ako vzdelávanie v oblasti informačnej gramotnosti môže posilniť odolnosť spoločnosti proti dezinformáciám (Haigh, Haigh & Matychak, 2019).

V súčasnej digitálnej ére, kedy generatívna AI mení paradigmy tvorby a šírenia obsahu, považujeme za nevyhnutné integrovanie tém AI do vzdelávania v oblasti informačnej a mediálnej gramotnosti. Je nevyhnutné, aby vzdelávacie programy zahŕňali komplexné pochopenie týchto technológií, ich potenciálu, ako aj rizík spojených s ich zneužitím. Toto rozšírenie kurikula by malo poskytovať študentom a širokej verejnosti nástroje na identifikáciu a kritickú analýzu obsahu generovaného AI, čím možno zvýšiť schopnosť rozlišovať medzi pravdivými a zavádzajúcimi informáciami. V konečnom dôsledku, začlenenie témy AI do

osnov informačnej a mediálnej gramotnosti nie je len otázkou adaptácie na technologický pokrok, ale základným kameňom pre budovanie informovanej, kriticky mysliacej a eticky zodpovednej spoločnosti, schopnej čeliť výzvam a využívať príležitosti, ktoré digitálny vek prináša.

Rozvoj *kritického myslenia* a *analytických zručností* je ďalším dôležitým prístupom. Tieto zručnosti umožňujú jednotlivcom analyzovať kontext informácií, ich zdroje a tiež potenciálne účely stojace za ich šírením. Vzdelávacie programy a workshopy, ktoré sa zameriavajú na rozvoj týchto zručností, sa tak stávajú kľúčovými pre vytváranie spoločnosti odolnejšej proti dezinformáciám. Ako jeden z príkladov uveďme hru *Fakey*, ktorá simuluje sociálne médiá a umožňuje hráčom praktizovať identifikáciu falošných správ (Micallef, Avram, Menczer & Patil, 2021).

Ako ďalšie efektívne riešenie sa ukazuje podpora kolektívnej inteligencie a spolupráce medzi používateľmi, ale i medzi expertmi a laickou verejnosťou na overovaní faktov a identifikácii a vyvracaní dezinformácií. Termín *kolektívna inteligencia* môžeme definovať ako schopnosť skupiny ľudí efektívne spolupracovať a využívať individuálne znalosti, zručnosti a skúsenosti na dosiahnutie spoločného cieľa alebo riešenia komplexných problémov efektívnejšie, než by to dokázal ktorýkoľvek jednotlivec samostatne. Podstata kolektívnej inteligencie spočíva v schopnosti skupiny efektívne komunikovať, deliť sa o informácie, koordinovať úlohy a spoločne pracovať na dosiahnutí cieľov, ktoré sú často komplexné a viacrozmerné. K znižovaniu miery šírenia dezinformácií generovaných pomocou AI tak môžu výrazne prispieť napríklad platformy, ktoré umožňujú používateľom využívať kolaboratívny prístup pri overovaní dôveryhodnosti informácií, platformy zameriavajúce sa na občiansku žurnalistiku či iniciatívy zamerané na overovanie faktov (Leonard & Levin, 2022). Spomeňme niekoľko príkladov iniciatív snažiacich sa o zlepšenie kvality dostupných informácií a o posilnenie odolnosti spoločnosti proti dezinformáciám: *WikiTribune*, *CrossCheck*, *Full Fact*, *PolitiFact*, *Snopes*, *FactCheck.org*., *Bellingcat* alebo *First Draft*.

Ďalším kľúčovým intelektuálnym prístupom k riešeniu problému tvorby a šírenia deepfake obsahov a dezinformácií pomocou AI je budovanie tzv. *kultúry osobnej zodpovednosti*. Tento prístup zdôrazňuje dôležitosť individuálnej úlohy každého jednotlivca v digitálnom ekosystéme a jeho schopnosť ovplyvňovať integritu informačného priestoru prostredníctvom etického správania a zodpovedného zdieľania obsahu. Hovorí o tom, že každý má úlohu v ochrane integrity nášho digitálneho sveta a že zodpovedné správanie a etické rozhodovanie sú

základnými kameňmi v boji proti dezinformáciám a zneužívaniu AI technológií. Osobná zodpovednosť v tomto kontexte zahŕňa viaceré aspekty, od kritického myslenia a overovania zdrojov, rozpoznávania a odmietania obsahu, ktorý môže byť zavádzajúci alebo škodlivý, etického rozhodovania o tom, čo a ako zdieľať, až po podporu transparentnosti a pravdivosti v online diskusiách. Budovanie kultúry osobnej zodpovednosti v konečnom dôsledku podporuje vytváranie silnejšej a odolnejšej online komunity, ktorá je schopná kolektívne čeliť výzvam spojeným s dezinformáciami a zneužívaním AI technológií (Loon, 2017; Millar, 2021).

Je dôležité zdôrazniť, že aj keď môže byť generovanie misinformácií a dezinformácií pomocou AI vnímané niektorými, najmä mladými ľuďmi, ako neškodná zábava alebo prostriedok na získanie pozornosti, takéto konanie môže mať vážne spoločenské dôsledky. Syntetické deepfake informačné obsahy v internetovom prostredí môžu erodovať verejnú dôveru v médiá, inštitúcie a vedu, čím sa oslabuje demokratický diskurz a spoločenská súdržnosť. Vzdelávanie by malo klásť dôraz na etickú zodpovednosť pri využívaní AI, poukazujúc na to, že každý jednotlivec má zodpovednosť za obsah, ktorý vytvára a šíri. Mladí ľudia by mali byť preto vedení k pochopeniu, že ich digitálne akcie majú reálne dôsledky, a mali by byť vzdelávaní, ako sa zodpovedne zapájať do digitálneho sveta. Týmto spôsobom môžeme podporiť vytváranie zodpovednej digitálnej kultúry, v ktorej mladí ľudia uznávajú svoju úlohu ako konštruktívni a etickí účastníci digitálneho sveta.

V kontexte intelektuálnych riešení uveďme aj ďalšie dôležité prístupy, akými sú *digitálna etika* a *digitálne občianstvo*. Digitálna etika sa zameriava na otázky, ako sú ochrana súkromia, autorské práva, spravodlivé používanie digitálnych zdrojov a etické aspekty využívania technológií. V kontexte AI je digitálna etika kľúčová pri definovaní hraníc medzi prijateľným a neetickým využívaním technológií na generovanie obsahu, ktorý môže byť zavádzajúci alebo škodlivý. Digitálne občianstvo sa týka zodpovedného a etického správania sa jednotlivcov v digitálnom svete. Zahrňuje pochopenie práv a povinností, ktoré prichádzajú s prístupom k digitálnym technológiám, ako aj schopnosť kriticky myslieť o informáciách, ktoré sa šíria v online prostredí. V kombinácii, digitálna etika a digitálne občianstvo poskytujú rámec pre rozvoj zdravého digitálneho prostredia, kde sú jednotlivci vybavení na identifikáciu a odmietnutie nepravdivých informácií. Tieto koncepty podporujú vytváranie kultúry zodpovednosti a záväzku k pravde, čo je nevyhnutné pre ochranu demokracie a verejnej dôvery

v dobe, keď AI mení spôsob, akým komunikujeme a zdieľame informácie (Ethics4EU Consortium, 2021; Richardson & Milovidov, 2019).

Inovatívnym spôsobom ako zvýšiť povedomie o problematike dezinformácií je využitie *umeleckých a naratívnych prístupov*. Projekty, ktoré využívajú príbehy, vizuálne umenie alebo interaktívne médiá, môžu pomôcť ilustrovať nielen konkrétne prípady dezinformácií a dôsledky ich šírenia, ale aj širšie sociálne a psychologické dynamiky, ktoré umožňujú ich šírenie a prijatie. Umenie a naratív majú jedinečnú schopnosť komunikovať komplexné témy, ako sú dezinformácie a ich dôsledky spôsobom, ktorý je emotívne rezonujúci a ľahko prístupný širokej verejnosti, a podnecovať tak k diskusii a reflexii. Tieto prístupy umožňujú preklenúť rozdiel medzi suchými faktami a osobnou skúsenosťou, čím poskytujú silný nástroj na ovplyvňovanie verejnej mienky a podporu kritického myslenia (Dahlstrom, 2020; Walker, Thuermer, Vicens & Simperl, 2023).

V tejto časti sme preskúmali tri základné prístupy k neetickému využívaniu AI technológií. Každý z týchto prístupov prispieva k celkovému riešeniu problematiky, pričom ich synergia ponúka komplexnú obranu proti potenciálnym rizikám a zneužitiu AI technológií. Integrácia *technologických*, *právnych* a *intelektuálnych* riešení predstavuje komplexný a synergický prístup k riešeniu výziev spojených s neetickým využívaním generatívnej AI. Spoločným úsilím v týchto troch oblastiach môžeme efektívne čeliť potenciálnym rizikám, podporovať etické používanie AI a zabezpečiť, že vývoj a implementácia AI technológií budú slúžiť spoločenskému dobru a budú rešpektovať základné ľudské hodnoty a práva.

Odhliadnuc od rozobratých riešení je potrebné spomenúť aj ďalšie dôležité oblasti, ktorých podporou možno prispieť k boju proti dezinformáciám a neetickému využívaniu systémov generatívnej AI na tvorbu a šírenie rizikového obsahu.

- Podpora medzinárodnej spolupráce: dezinformácie sú globálnym problémom a vyžadujú si globálne riešenie. Je dôležité, aby štáty a medzinárodné organizácie spolupracovali na výmene informácií, koordinácii politík, podpore výskumu v oblasti dezinformácií a vytváraní a implementácii medzinárodných regulačných rámcov, ktoré budú regulovať vývoj a používanie systémov generatívnej AI a budú bojovať proti ich neetickému zneužitiu. Rovnako dôležité je aj zdieľanie informácií a najlepších praktík nie len medzi štátmi, ale aj technologickými platformami a občianskou spoločnosťou

- Podpora nezávislých médií: nezávislé médiá zohrávajú kľúčovú úlohu pri overovaní faktov a boji proti dezinformáciám. Je dôležité, aby mali k dispozícii dostatočné zdroje a aby boli chránené pred politickým tlakom. Dôležitá je aj podpora investigatívnej žurnalistiky, ktorá by odhaľovala a informovala o prípadoch zneužitia systémov generatívnej AI.
- Podpora výskumu a vývoja v oblasti AI, ktorá by sa zamerala na vývoj etických a zodpovedných systémov a na boj proti ich zneužitiu. Dôležitá je tiež podpora transparentnosti a kontroly pri vývoji a používaní systémov generatívnej AI, aby sa predišlo ich zneužitiu.
- Podpora inkluzívneho a participatívneho prístupu k vývoju AI, ktorý by zohľadňoval rôzne perspektívy a potreby spoločnosti.
- Spustenie informačných kampaní, ktoré by informovali verejnosť o rizikách neetického používania systémov generatívnej AI a o tom, ako sa im brániť.

ZÁVER

V ére digitálnej transformácie, kedy generatívna AI otvára nové horizonty v tvorbe a šírení obsahu, stojíme pred výzvami, ktoré si vyžadujú komplexný a multidisciplinárny prístup. V tomto príspevku sme poukázali na to, že neetické používanie generatívnej AI na tvorbu a šírenie dezinformácií, falošných správ a deepfake obsahov predstavuje vážnu hrozbu pre spoločnosť, demokraciu a dôveru v digitálne médiá. Od definície kľúčových pojmov, cez prípadovú štúdiu, až po rozbor možných riešení, sme sa snažili poskytnúť ucelený pohľad na túto problematiku.

Konkrétne príklady v rámci deskriptívnej prípadovej štúdie ilustrujú nielen potenciálne nebezpečenstvá spojené so zneužívaním generatívnej AI, ale aj dôležitosť hľadania efektívnych riešení. Ako sme ukázali, tieto riešenia zahŕňajú technologické inovácie zamerané na detekciu syntetického obsahu, právne rámce, ktoré regulujú využívanie a neetické používanie AI, a intelektuálne prístupy, ktoré podporujú zodpovedné využívanie technológií na úrovni spoločnosti aj jednotlivcov.

Je potrebné zdôrazniť, že boj proti dezinformáciám a neetickému používaniu generatívnej AI nie je úlohou len pre IT profesionálov, právnych expertov alebo vzdelávateľov. Je to výzva pre nás všetkých – ako digitálnych občanov, ktorí musia byť informovaní, kritickí a eticky zodpovední voči obsahu, s ktorým sa stretávame, a ktorý zdieľame v digitálnom svete. Budovanie silnej kultúry digitálnej etiky a občianstva, kde každý jednotlivec preberá zodpovednosť za integritu informačného priestoru, je kľúčové pre ochranu našej spoločnosti pred škodlivými vplyvmi dezinformácií.

V tejto ére plnej výziev aj príležitostí je nevyhnutné, aby sme spoločne pracovali na vytváraní odolnejšieho a informovanejšieho digitálneho prostredia. Len tak môžeme zabezpečiť, že inovácie v oblasti AI budú slúžiť v prospech spoločnosti, a nie na jej škodu. Naša kolektívna budúcnosť závisí od našej schopnosti orientovať sa v týchto vodách s múdrosťou, zodpovednosťou a spoločným záväzkom smerom k pravde.

DEDIKÁCIA

Príspevok bol vypracovaný v rámci riešenia projektu VEGA 1/0360/21 Sociálne reprezentácie etických výziev digitálnej informačnej revolúcie.

Táto práca bola podporená Agentúrou na podporu výskumu a vývoja na základe Zmluvy č. APVV-19-0074.

ZOZNAM POUŽITEJ LITERATÚRY

AFFSPRUNG, D. (2023). The ELIZA defect: constructing the right users for generative AI. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 945–946. https://dl.acm.org/doi/10.1145/3600211.3604744

ALLYN, B. (2022). *Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn*. NPR. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

ALPHONSO, A. (2023). Photo of Pope Francis with two women in a hot tub is an AI-generated fake. Boom. https://www.boomlive.in/fact-check/world/fake-news-viral-photo-pope-francis-with-two-women-in-a-bathtub-midjourney-factcheck-21551

BAIDOO-ANU, D. & OWUSU, L. (2023). Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN*. http://dx.doi.org/10.2139/ssrn.4337484

BARANI, M. & DYCK, P. (2023). *Generative AI and the EU AI Act - a closer look*. Allen & Overy. https://www.allenovery.com/en-gb/global/blogs/tech-talk/generative-ai-and-the-eu-ai-act-a-closer-look

BARREDO, D., JAMIL, S. & MONTEMAYOR, D. (2023). Disinformation and artificial intelligence: the case of online journalism in China. *Estudios sobre el Mensaje Periodístico*, 29(4), 761-770. https://revistas.ucm.es/index.php/ESMP/article/view/88543

BAXTER, K. & Y. SCHLESINGER (2023). *Managing the risks of generative AI*. Harvard Business Review. https://hbr.org/2023/06/managing-the-risks-of-generative-ai

BRANDON (2024). *Taylor Swift became a victim of deepfake porn 13:04*. Social Bites. https://socialbites.ca/culture/480158.html

BUȘINCU, C. & A. ALEXANDRESCU. (2023). Blockchain-based platform to fight disinformation using crowd wisdom and artificial intelligence. *Applied Sciences*, 13(10). https://www.mdpi.com/2076-3417/13/10/6088

C2PA. (2023). *Wait, where did this image come from?*. https://contentcredentials.org/

CABINET OFFICE. [s. a.]. *Society 5.0*. https://www8.cao.go.jp/cstp/english/society5_0/index.html

CENTRE FOR DATA ETHICS AND INNOVATION. [s. a.]. *About us*. https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation/about

Coalition for content provenance and authenticity. (2024). https://c2pa.org/

DAHLSTROM, M. F. (2020). *The narrative truth about scientific misinformation*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3497784

Department of industry, science and resources. [s. a.]. *Australia's AI ethics principles.* https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

European Commission. (2024). *AI Act*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Commission. (2018). *A multi-dimensional approach to disinformation: report of the independent High level Group on fake news and online disinformation..* https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

Európska komisia. [s. a.]. *Akt o digitálnych službách*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_sk

Ethics4EU Consortium. (2021). *European values for ethics in digital technology*. https://arrow.tudublin.ie/scschcomrep/12/

European parliament. (2023). *Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI*. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

Európsky parlament — kancelária na slovensku. (2023). *Martin Hojsík na tému pravidiel riadenia umelej inteligencie (jún 2023)*. YouTube. https://www.youtube.com/watch?v=PhylE0ve4fU

FAGUY, A. (2023). *Taylor Swift—who became a billionaire—caps year as time's person of the year*. Forbes. https://www.forbes.com/sites/anafaguy/2023/12/06/taylor-swift-named-time-magazines-person-of-the-year/?sh=5dbfb3b91539

FÁTIMA C. (2023). Artificial intelligence in automated detection of disinformation: a thematic analysis. *Journal of Media*, 679-687. https://www.mdpi.com/2673-5172/4/2/43

FAZLIOGLU, M. (2023). *US federal AI governance: laws, policies and strategies*. iapp. https://iapp.org/resources/article/us-federal-ai-governance/

FERRARA, E. (2024). GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Computers and Society*. https://arxiv.org/abs/2310.00737

FEUERRIEGEL, S., DIRESTA, R., GOLDSTEIN, J. A., KUMAR, S., LORENZ-SPREEN, P., TOMZ, M. & PRÖLLOCHS, N. (2023). Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, 1818–1821. https://www.nature.com/articles/s41562-023-01726-2

GAURAV, P., ZHANG, R. & ZHU, J. (2022). On aliased resizing and surprising subtleties in GAN evaluation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11400-11410. https://ieeexplore.ieee.org/document/9880182

GIORDANI, J. (2023). *Unmasking the dark side of generative AI: protecting your data from security threats*. LinkedIn. https://www.linkedin.com/pulse/unmasking-dark-side-generative-ai-protecting-your-data-john-giordani/

GOODFELLOW, I., BENGIO, Y. & COURVILLE, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Google DeepMind. [s. a.]. *SynthID*. https://deepmind.google/technologies/synthid/

GORDON, R. (2023). *Using AI to protect against AI image manipulation*. MIT News. https://news.mit.edu/2023/using-ai-protect-against-ai-image-manipulation-0731

Government of canada. ([2022]). *The Artificial Intelligence and Data Act (AIDA) – Companion document*. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

GOWAL, S. & KOHLI, P. (2023). *Identifying AI-generated images with SynthID*. Google DeepMind. https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/

GUARNERA, L., GIUDICE, O., NASTASI, C. & BATTIATO S. (2020). Preliminary forensics analysis of deepfake images. *2020 AEIT International Annual Conference (AEIT)*. https://ieeexplore.ieee.org/document/9241108/authors#authors

GUPTA, R. P. (2023). *Prompt hacking - what we should know*. LinkedIn. https://www.linkedin.com/pulse/prompt-hacking-what-we-should-know-ravi-prakash-gupta/

HAIGH, M, HAIGH, T. & MATYCHAK, T. (2019). Information literacy vs. fake news: the case of Ukraine. *Open Information Science*. https://www.degruyter.com/document/doi/10.1515/opis-2019-0011/html

HASANAIN, M., ALAM, F., MUBARAK, H., ABDALJALIL, S., ZAGHOUANI, W., NAKOV, P., MARTINO G. & FREIHAT, A. A. (2023). ArAIEval shared task: persuasion

techniques and disinformation detection in arabic text. *Computation and Language*. https://arxiv.org/abs/2311.03179

HEIKKILÄ, M. (2023). *This new tool could protect your pictures from AI manipulation*. MIT Technology Review. https://www.technologyreview.com/2023/07/26/1076764/this-new-tool-could-protect-your-pictures-from-ai-manipulation/

HEIKKILÄ, M. (2023a). *This new data poisoning tool lets artists fight back against generative AI*. MIT Technology Review. https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/

HELMUS, T. C. (2022). *Artificial intelligence, deepfakes, and disinformation*. Rand. https://www.rand.org/content/dam/rand/pubs/perspectives/PEA1000/PEA1043-1/RAND_PEA1043-1.pdf

HENDL, J. (2008). *Kvalitativní výzkum: základní teorie, metody a aplikace*. Portál.

HENDRICKSON, L. (2024). *Deepfake AI: how verified credentials enhance media authenticity*. identity. https://www.identity.com/deepfake-ai-how-verified-credentials-enhance-media-authenticity/#What_Are_Verifiable_Credentials

HIGGINS, E. (2023). *Making pictures of Trump getting arrested while waiting for Trump's arrest*. X. https://twitter.com/EliotHiggins/status/1637927681734987777

CHEN, Y., PAN, X., LI, Y., DING, B. & ZHOU, J. (2023). EE-LLM: large-scale training and inference of early-exit large language models with 3D parallelism. *Machine Learning*. https://arxiv.org/abs/2312.04916

Infoz. ([s. a.]). *Catfishing*. https://www.infoz.cz/catfishing/

INGRAM, M. (2024). *Taylor Swift deepfakes could be the tip of an AI-generated iceberg*. Columbia Journalism Review. https://www.cjr.org/the_media_today/taylor_swift_deepfakes_ai.php

JEAN-PIERRE, K. (2024). *White House 'alarmed' by AI deepfakes of Taylor Swift*. YouTube. https://www.youtube.com/watch?v=-YIxFW9DoS4

KAI, S., SLIVA, A., SUHANG, W., JILIANG, T. & HUAN, L. (2017). Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1). https://dl.acm.org/doi/10.1145/3137597.3137600

KANUNGO, P. (2024). *Who is ZvBear? Memes erupt on Twitter amid viral Taylor Swift AI pictures scandal.* SK POP. https://www.sportskeeda.com/pop-culture/news-who-zvbear-memes-erupt-twitter-amid-viral-taylor-swift-ai-pictures-scandal

Kaspersky. (2024). *What is Biometrics? How is it used in security?*. https://usa.kaspersky.com/resource-center/definitions/biometrics

KAZAKHSTANDAZHASALGAN. (2024). [*Deepfake obrázky Eiffelovej veže*]. TikTok. https://www.tiktok.com/@kazakhstandazhasalgan/photo/7325512973080874246

KELLER, E. (2023). *Pope Francis in Balenciaga deepfake fools millions: 'Definitely scary'*. New York Post. https://nypost.com/2023/03/27/pope-francis-in-balenciaga-deepfake-fools-Anmol Alphonso

KIM, Y., XU, X., MCDUFF, D., BREAZEAL, C. & PARK, H. W. (2023). Health-LLM: Large language models for health prediction via wearable sensor data. *Computation and Language*. https://arxiv.org/abs/2401.06866

KOTHARI, A., ORAMA, A., MILLER, R., PEEKS, M., BAILEY, R. & ALM, C. (2023). News consumption helps readers identify model-generated news. *2023 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, 1-10. https://ieeexplore.ieee.org/document/10349588

KÜÇKERDOĞAN, B. & TURĞAL, L. (2023). Artificial intelligence as a disinformation tool: analyzing news photos on climate change in the example of Bing search engine. *İletişim*

Ve Diplomasi, (11), 57-82.
https://dergipark.org.tr/tr/pub/iletisimvediplomasi/issue/81401/1376404

LECUN, Y., BENGIO, Y. & HINTON, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. https://www.scirp.org/reference/referencespapers?referenceid=1911084

LEONARD, N. E. & LEVIN, S. A. (2022). Collective intelligence as a public good. Online. *Collective Intelligence*, 1(1).
https://journals.sagepub.com/doi/10.1177/26339137221083293#tab-contributors

LOON, B. (2017). *Why we should hold ourselves responsible for fake news*. Center for Digital Ethics & Policy.
https://www.luc.edu/digitalethics/researchinitiatives/essays/archive/2017/whyweshouldholdourselvesresponsibleforfakenews/

LUO, Y., DAN, X. & SHEPHERD, N. (2023). *China proposes draft measures to regulate generative AI*. Inside Privacy. https://www.insideprivacy.com/artificial-intelligence/china-proposes-draft-measures-to-regulate-generative-ai/

MANCO, G., RITACCO, E., RULLO, A., SACCÁ, D. & SERRA, E. (2022). Generating synthetic discrete datasets with machine learning. *SEBD 2022 Italian Symposium on Advanced Database Systems*. https://dblp.org/rec/conf/sebd/0001RRSS22.html

MASTERSON, V. (2023). *What is Nightshade – the new tool allowing artists to 'poison' AI models?*. World Economic Forum. https://www.weforum.org/agenda/2023/11/nightshade-generative-ai-poison/

MICALLEF, N., AVRAM M., MENCZER, F. & PATIL, S. (2021). Game intervention to improve news literacy on social media. *Proc. {ACM} Hum. Comput. Interact*, 5.
https://dblp.org/rec/journals/pacmhci/MicallefAMP21.html

MILLAR, B. (2021). Misinformation and the limits of individual responsibility. *Social Epistemology Review and Reply Collective*, 10(12), 8-21.
https://philarchive.org/rec/MILMAT-22

Ministerstvo investícií, regionálneho rozvoja a informatizácie SR. (2024). *Bezpečnejší digitálny svet: do účinnosti vstupuje akt o digitálnych službách*. https://mirri.gov.sk/aktuality/digitalna-agenda/bezpecnejsi-digitalny-svet-do-ucinnosti-vstupuje-akt-o-digitalnych-sluzbach/

MITCHELL, G. (2024). *Taylor Swift AI pictures Twitter controversy*. Inside Inquiries. https://insightinquiries.com/taylor-swift-ai-pictures-twitter-2/

NOVAK, M. (2024). *Pope Francis warns of AI dangers, citing fake image of him that went viral*. Forbes. https://www.forbes.com/sites/mattnovak/2024/01/24/pope-francis-warns-of-ai-dangers-after-fake-image-of-himself-went-viral/?sh=6450b59f5aa0

O médiách. (2024). *Objavilo sa deepfake video s tvárou a hlasom Pellegriniho a moderátora RTVS (VIDEO)*. https://www.omediach.com/internet/26046-objavilo-sa-deepfake-video-s-tvarou-a-hlasom-

OPENAI. (2023). *ChatGPT-4.* https://chat.openai.com/

PARK, S. (2023). Use of generative artificial intelligence, including large language models such as ChatGPT, in scientific publications: policies of KJR and prominent authorities. *Korean Journal of Radiology*, 24(8), 715-718. https://kjronline.org/DOIx.php?id=10.3348/kjr.2023.0643

PBS. (2023). *Fake AI images of Putin, Trump being arrested spread online*. https://www.pbs.org/newshour/politics/fake-ai-images-of-putin-trump-being-arrested-spread-online

PEARSON, J. & ZINETS, N. (2022). *Deepfake footage purports to show Ukrainian president capitulating*. Reuters. https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/

PEÑA-FERNÁNDEZ S., MESO-AYERDI, K., LARRONDO-URETA, A. & DÍAZ-NOCI, J. (2023). *Without journalists, there is no journalism: the social dimension of generative*

artificial intelligence in the media. *Profesional de la información*, 32(2). https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87329/63392

PODELL, D., ENGLISH, Z., LACEY, K., BLATTMANN, A., DOCKHORN, T., MÜLLER, J., PENNA, J. & ROMBACH, R. (2023). SDXL: improving latent diffusion models for high-resolution image synthesis. *Computer Vision and Pattern Recognition*. https://arxiv.org/abs/2307.01952

Polícia Slovenskej republiky. (2023). *Varovanie: voľby do národnej rady sprevádza zneužívanie umelej inteligencie*. Meta. https://www.facebook.com/policiaslovakia/posts/pfbid0G6mNQyP7e3AoQD17FPM8yDUnpNiFXX8hyrMqfyMDHYHaES5sQJXLRFyEWUqn1Wk6l

Pravda. (2024). *Pellegrini varuje pred AI a deep fake videami. Tvrdí, že zásadne môžu ovplyvniť aj prezidentskú kampaň*. https://spravy.pravda.sk/prezidentske-volby-2024/clanok/698225-pellegrini-varuje-pred-ai-a-deep-fake-videami-tvrdi-ze-zasadne-mozu-ovplyvnit-aj-prezidentsku-kampan/

PULLELLA, P. (2024). *Pope Francis, victim of AI, warns against its 'perverse' dangers*. Reuters. https://www.reuters.com/world/pope-francis-victim-ai-warns-against-its-perverse-dangers-2024-01-24/

RABINDER, H. (2019). Role of artificial intelligence in new media. *CSI Communications*, 23-25. https://www.academia.edu/49150704/Role_of_Artificial_Intelligence_in_New_Media

Reddit (2023). *Pope Francis bike week.* https://www.reddit.com/r/midjourney/comments/134dybi/pope_francis_bike_week/

Regulation (EU) 2022/2065 of the European parliament and of The Council. (2022). *Official Journal of the European Union.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

REMOLINA, N. & SEAH, J. (2019). How to address the AI governance discussion? What can we learn from Singapore's AI strategy? *SMU Centre for AI & Data Governance Research Paper No. 2019/03*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3444024

RICHARDSON, J. & MILOVIDOV, E. (2019). *Digital citizenship education handbook*. Council of Europe Publishing. https://rm.coe.int/digital-citizenship-education-handbook/168093586f

ROMAIN, S. (2023). *Sentinel AI: the new frontier in deepfake detection*. Romain Berg. https://www.romainberg.com/blog/artificial-intelligence/sentinel-ai-your-ultimate-deepfake-detection-solution/

RUSSELL, S. & NORVIG, P. (2010). *Artificial intelligence: a modern approach*. Prentice Hall. https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

SENTINEL (2024). *Thesentinel*. https://thesentinel.ai/

SHAN, S., WENGER, E., ZHANG, J., LI, H., ZHENG, H. & ZHAO, B. Y. [s. a.]. *Image "cloaking" for personal privacy*. SAND Lab. https://sandlab.cs.uchicago.edu/fawkes/

SHIMPO, F. (2020). The importance of ‘smooth’ data usage and the protection of privacy in the age of AI, IoT and autonomous robots. *Global Privacy Law Review*, 1(1), 49-54. https://kluwerlawonline.com/journalarticle/Global+Privacy+Law+Review/1.1/GPLR2020006

SHOAIB, M. R., WANG, Z., AHVANOOEY, M. T. & ZHAO, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. *2023 International Conference on Computer and Applications (ICCA)*, 1-7. https://ieeexplore.ieee.org/document/10401723

SIMON, F. M., ALTAY, S. & MERCIER, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School (HKS) Misinformation Review*, 4(5). https://misinforeview.hks.harvard.edu/wp-content/uploads/2023/10/simon_generative_AI_fears_20231018.pdf

STANLEY-BECKER, I. & NIX, N. (2023). *Fake images of Trump arrest show 'giant step' for AI's disruptive power*. The Washington Post. https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/

STOKEL-WALKER, C. (2023). *We spoke to the guy who created the viral AI image of The Pope that fooled the world*. buzzfeednews. https://www.buzzfeednews.com/article/chrisstokelwalker/pope-puffy-jacket-ai-midjourney-image-creator-interview

STRICKLAND, E. (2023). *Content credentials will fight deepfakes in the 2024 elections*. IEEE Spectrum. https://spectrum.ieee.org/deepfakes-election

TASR. (2023). *Europoslanci rokovali o pravidlách pre bezpečnú umelú inteligenciu*. https://www.teraz.sk/priame-prenosy-a-videa-tasr-tv/europoslanci-rokovali-o-pravidlach-pre/722992-clanok.html

The associated press. (2023). *AI-generated images of Trump being arrested circulate on social media*. https://apnews.com/article/fact-check-trump-nypd-stormy-daniels-539393517762

TravelWise. (2024). *Deepfake about the Eiffel tower fire went viral: what's wrong with it and why people believed it*. https://www.travelwiseway.com/section-news/news-deepfake-about-the-eiffel-tower-fire-went-viral-whats-wrong-with-it-and-why-people-believed-it-28-01-2024.html

TREDINNICK, L. & LAYBATS, C. (2023). The dangers of generative artificial intelligence. *Business Information Review*, 40(2). https://journals.sagepub.com/doi/10.1177/02663821231183756

TRUMP, J. D. (2023). [*Zmanipulovaný obrázok*]. Truth Social. https://truthsocial.com/@realDonaldTrump/posts/110475191660845818

TURING, A. (1950). Computing machinery and intelligence. *Mind, New Series*, 59(236), 433-460. https://phil415.pbworks.com/f/TuringComputing.pdf

ULMER, A. & TONG, A. (2023). *With apparently fake photos, DeSantis raises AI ante*. Reuters. https://www.reuters.com/world/us/is-trump-kissing-fauci-with-apparently-fake-photos-desantis-raises-ai-ante-2023-06-08/

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. & POLOSUKHIN, I. (2017). Attention is all you need. *Conference on Neural Information Processing Systems.* https://arxiv.org/pdf/1706.03762.pdf

WAKEFIELD, J. (2022). *Deepfake presidents used in Russia-Ukraine war*. BBC. https://www.bbc.com/news/technology-60780142

WALKER, J., THUERMER, G., VICENS, J. & SIMPERL, E. (2023). AI art and misinformation: approaches and strategies for media literacy and fact checking. *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 26-37. https://dl.acm.org/doi/10.1145/3600211.3604715

WEAVER, A. (2024). [*falošné video*]. https://www.facebook.com/reel/696473435970931

X. 2023. *Donald Trump became a household name by FIRING countless people *on television**. https://twitter.com/DeSantisWarRoom/status/1665799058303188992?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1665799058303188992%7Ctwgr%5E8f403b7dc5f370ef80c618848409a300fa4f884d%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.reuters.com%2Fworld%2Fus%2Fis-trump-kissing-fauci-with-apparently-fake-photos-desantis-raises-ai-ante-2023-06-08%2F

XIN, W., HUI, G., SHU, H., MING-CHING, C. & SIWEI, L. (2022). GAN-generated faces detection: a survey and new perspectives. *Frontiers in Artificial Intelligence and Applications*. 372, 2533-2542. https://ebooks.iospress.nl/doi/10.3233/FAIA230558

XU, D., FAN, S. & KANKANHALLI, M. (2023). Combating misinformation in the era of generative AI models. *Proceedings of the 31st ACM International Conference on Multimedia*, 9291–9298. https://dl.acm.org/doi/10.1145/3581783.3612704

ZACH & WALKER, A. (2024). *Eiffel tower on fire AI hoax*. Know Your Meme. https://knowyourmeme.com/memes/eiffel-tower-on-fire-ai-hoax

POZNÁMKA O AUTOROVI

Tomáš Mirga

Tomáš Mirga je interný doktorand na Katedre knižničnej a informačnej vedy na Univerzite Komenského v Bratislave. V kontexte svojej akademickej činnosti sa orientuje na problematické oblasti online prostredia, akými sú informačné bubliny, dezinformácie, fake news a deepfake obsahy. Z hľadiska širšieho zamerania sa venuje informačnej etike, informačnej a mediálnej gramotnosti, sociálnej psychológii, kognitívnym skresleniam, sociálnym médiám a umelej inteligencii.

E-mail: mirga1@uniba.sk