

Využití umělé inteligence v phishingových útocích: Právní regulace a etické výzvy v kybernetické bezpečnosti

Bc. Ester Gajdošík



ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení: Bc. Ester Gajdošík
Osobní číslo: A23402
Studijní program: N0613A140022 Informační technologie
Specializace: Kybernetická bezpečnost
Forma studia: Kombinovaná
Téma práce: Využití umělé inteligence v phishingových útocích: Právní regulace a etické výzvy v kybernetické bezpečnosti
Téma práce anglicky: The Use of Artificial Intelligence in Phishing Attacks: Legal Regulation and Ethical Challenges in Cybersecurity

Zásady pro vypracování

- Provedte rešerši analýzy technologického vývoje phishingových útoků s využitím umělé inteligence.
- Analyzujte existující právní normy, které upravují použití AI v kybernetické bezpečnosti
- Pojmenujte a klasifikujte etické otázky spojené s používáním AI v phishingu, včetně problematiky odpovědnosti a spravedlnosti.
- Navrhněte možná zlepšení, která by mohla být implementována k lepšímu zvládnutí hrozeb spojených s AI.
- Vyhodnotte a ověřte své návrhy, jak v praxi aplikovat regulační opatření a etické zásady týkající se AI v oblasti phishingových útoků.

Forma zpracování diplomové práce: **tištěná/elektronická**
Jazyk zpracování: **Slovenština**

Seznam doporučené literatury:

1. BASIT, A.; ZAFAR, M. a LIU, X., 2020. A comprehensive survey of AI-enabled phishing attacks detection techniques. Online. s. 139–154. Dostupné z: <https://doi.org/https://doi.org/10.1007/s11235-020-00733-2>.
2. BEŠIĆ, M., 2023. Benefits and Risks of Artificial Intelligence in Cybersecurity and Phishing Attacks. Online. s. 94–98. Dostupné z: <https://ebt.rs/journals/index.php/conf-proc/article/view/175>.
3. FINOCCHIARO, G., 2023. The regulation of artificial intelligence. Online. Article 39, s. 1961–1968. Dostupné z: <https://doi.org/https://doi.org/10.1007/s00146-023-01650-z>.
4. Nařízení Evropského parlamentu a Rady (EU) 2024/1689 ze dne 13. června 2024, kterým se stanoví harmonizovaná pravidla pro umělou inteligenci a mění nařízení (ES) č. 300/2008, (EU) č. 167/2013, (EU) č. 168/2013, (EU) 2018/858, (EU) 2018/1139 a (EU) 2019/2144 a směrnice 2014/90/EU, (EU) 2016/797 a (EU) 2020/1828 (Akt o umělé inteligenci), 2024. Dostupné také z: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
5. SCHMITT, M. a FLECHAIS, I., 2024. Digital deception: generative artificial intelligence in social engineering and phishing. Online. s. 324. Dostupné z: <https://doi.org/https://doi.org/10.1007/s10462-024-10973-2>.

Vedoucí diplomové práce: **prof. Mgr. Roman Jašek, Ph.D., DBA**
Ústav informatiky a umělé inteligence

Datum zadání diplomové práce: **27. října 2024**
Termín odevzdání diplomové práce: **2. června 2025**



L.S.

doc. Ing. Jiří Vojtěšek, Ph.D. v.r.
děkan

prof. Mgr. Roman Jašek, Ph.D., DBA v.r.
ředitel ústavu

Ve Zlíně dne 17. listopadu 2024

PROHLÁŠENÍ AUTORA DIPLOMOVÉ PRÁCE

Beru na vědomí, že

- odevzdáním diplomové práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb., v platném znění bez ohledu na výsledek obhajoby;
- diplomová práce bude uložena v elektronické podobě v univerzitním informačním systému a bude dostupná k nahlédnutí;
- jedno vyhotovení diplomové práce v listinné podobě bude ponecháno Univerzitě Tomáše Bati ve Zlíně k uložení;
- na moji diplomovou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- podle § 60 odst. 1 autorského zákona má Univerzita Tomáše Bati ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- podle § 60 odst. 2 a 3 mohu užít své dílo – diplomovou práci – nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- pokud bylo k vypracování diplomové práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tj. k nekomerčnímu využití), nelze výsledky diplomové práce využít ke komerčním účelům;
- pokud je výstupem diplomové práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá; neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji, že

- jsem na diplomové práci pracoval(a) samostatně a použitou literaturu jsem řádně citoval(a); v případě publikace výsledků budu uveden(a) jako spoluautor;
- odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou obsahově totožné.

Ve Zlíně, dne

.....
podpis autora

ABSTRAKT

Táto práca sa zaoberá využitím umelej inteligencie (AI) vo phishingových útokoch a analyzuje ich vplyv na kybernetickú bezpečnosť s dôrazom na právne a etické aspekty. S rozvojom generatívnych AI technológií dochádza k výraznému zvyšovaniu sofistikovanosti a efektivity metód sociálneho inžinierstva. Práca identifikuje aktuálne bezpečnostné hrozby a navrhuje riešenia zamerané na štyri hlavné oblasti: technické opatrenia, legislatíva, vzdelávanie a etické princípy. Prostredníctvom teoretických analýz a praktických scenárov sú tieto návrhy zhodnotené z pohľadu ich účinnosti, prínosov a možných obmedzení. Práca poukazuje na potrebu prepojenia technológií s právnou reguláciou a etickým dohľadom pre bezpečné využívanie AI a ponúka odporúčania na posilnenie odolnosti voči AI-generovaným hrozbám.

Kľúčová slova: phishing, umelá inteligencia (AI), kybernetická bezpečnosť, právna regulácia, etika AI, deepfake

ABSTRACT

This thesis explores the use of artificial intelligence (AI) in phishing attacks and analyzes its impact on cybersecurity, with particular focus on legal and ethical dimensions. The rapid advancement of generative AI technologies has significantly increased the sophistication and effectiveness of social engineering techniques. The study identifies current security threats and proposes solutions across four key domains: technical measures, legal frameworks, education, and ethical principles. Through theoretical analysis and practical scenarios, these proposals are evaluated in terms of their effectiveness, benefits, and potential limitations. The findings highlight the critical need for close integration of technological solutions with effective legal regulation and ethical oversight to ensure the safe deployment of AI. Based on these insights, the thesis offers specific recommendations to strengthen resilience against AI-driven threats.

Keywords: phishing, artificial intelligence (AI), cybersecurity, legal regulation, AI ethics, deepfake

Poděkování

Touto cestou by som rada vyjadrila úprimnú vd'aku všetkým, ktorí prispeli k vzniku tejto práce. Osobitne ďakujem môjmu vedúcemu práce, prof. Mgr. Romanovi Jašekovi, Ph.D., DBA, za odborné vedenie a cenné pripomienky počas procesu tvorby diplomovej práce. Vďačnosť patrí aj mojim rodinným príslušníkom a priateľom za ich trpezlivosť a povzbudenie, ako aj kolegom a spolužiakom za inšpiratívne diskusie a zdieľanie nápadov, ktoré obohatili môj pohľad na túto tému. Zároveň veľmi oceňujem podporu Fakulty aplikovanej informatiky Univerzity Tomáša Baťu v Zlíne, ktorá mi poskytla potrebné podmienky a priestor na spracovanie tejto témy v rámci mojej diplomovej práce.

OBSAH

ÚVOD	13
I TEORETICKÁ ČÁST	15
1 ZÁKLADY A VÝVOJ PHISHINGOVÝCH ÚTOKOV S VYUŽITÍM UMELEJ INTELIGENCIE	16
1.1 PHISHING A JEHO ZÁKLADY	16
1.1.1 Proces phishingového útoku	17
1.1.2 Typy phishingových útokov	18
1.1.3 Znaky phishingových správ	21
1.2 VÝVOJ PHISHINGOVÝCH ÚTOKOV	22
1.2.1 Počiatky phishingu	22
1.2.2 Evolúcia a diverzifikácia útokov	23
1.2.3 Nástup sociálnych médií	24
1.2.4 Využitie umelej inteligencie a strojového učenia	24
1.3 PRINCÍPY VYUŽITIA AI VO PHISHINGOVÝCH ÚTOKOCH	25
1.3.1 Definícia phishingového útoku poháňaného AI	25
1.3.2 Hlavné technológie AI vo phishingu	27
2 PRAKTICKÉ FORMY PHISHINGOVÝCH ÚTOKOV POHÁŇANÝCH UMELOU INTELIGENCIOU	31
2.1 PHISHINGOVÉ E-MAILY POHÁŇANÉ AI	31
2.1.1 Ako AI zlepšuje phishingové e-maily	31
2.1.2 Používané technológie	32
2.1.3 Príklad útoku	34
2.2 DEEPPFAKE VIDEÁ VO PHISHINGU	38
2.2.1 Ako sa deepfake používa vo phishingu s využitím AI	38
2.2.2 Používané technológie	38
2.2.3 Príklad útoku	41
2.3 HLASOVÝ PHISHING (VISHING) S AI	42
2.3.1 Ako sa vishing používa vo phishingu s využitím AI	42
2.3.2 Používané technológie	43

2.3.3	Príklad útoku	44
3	PRÁVNE A ETICKÉ ASPEKTY VYUŽITIA AI VO PHISHINGOVÝCH ÚTO- KOCH	47
3.1	VÝZVY V PRÁVNEJ REGULÁCII	47
3.2	MEDZINÁRODNÉ PRÁVNE NORMY	49
3.2.1	Európsky zákon o umelej inteligencii (AI Act).....	49
3.2.2	OECD Principles on AI.....	54
3.3	NÁRODNÉ PRÁVNE NORMY	59
3.3.1	Ústavné práva relevantné pre AI a phishing.....	59
3.3.2	Trestný zákon	61
3.3.3	Zákon o kybernetickej bezpečnosti	61
3.3.4	Národná stratégia umelej inteligencie v Českej republike	62
3.4	ETICKÉ OTÁZKY SPOJENÉ S VYUŽITÍM AI VO PHISHINGU	63
3.4.1	Súkromie a zneužitie dát	64
3.4.2	Manipulácia a klamanie	65
3.4.3	Zodpovednosť	65
3.4.4	Spravodlivosť	66
3.4.5	Duševné vlastníctvo	66
II	PRAKTICKÁ ČASŤ	68
4	NÁVRHY NA ZLEPŠENIE ZVLÁDANIA HROZIEB SPOJENÝCH S UMELOU INTELIGENCIOU.....	69
4.1	TECHNICKÉ OPATRENIA	69
4.1.1	Zlepšenie detekčných systémov	70
4.1.2	Autentifikačné technológie.....	73
4.1.3	Detekcia deepfake a hlasového klonovania.....	75
4.2	LEGISLATÍVNE A REGULAČNÉ OPATRENIA	76
4.2.1	Právna zodpovednosť za škody spôsobené AI	76
4.2.2	Povinnosť detekcie a prevencie zneužitia AI	79
4.2.3	Povinné označovanie AI-generovaného obsahu	80
4.2.4	Posilnenie kybernetickej legislatívy.....	81

4.2.5	Medzinárodná spolupráca a harmonizácia.....	81
4.3	VZDELÁVANIE A ZVYŠOVANIE POVEDOMIA	82
4.3.1	Povinné školenia pre zamestnancov a zraniteľné skupiny	82
4.3.2	Kampane na zvýšenie povedomia pre verejnosť	83
4.3.3	Podpora digitálnej gramotnosti v školách	84
4.4	ETICKÉ ZÁSADY PRE VÝVOJ A VYUŽITIE AI	85
4.4.1	Zásada nezneužitelnosti pri navrhovaní	85
4.4.2	Povinnosť implementovať detekovateľnosť výstupu	86
4.4.3	Etický limit personalizácie	87
4.4.4	Zriadenie etických komisií pre AI v kyberbezpečnosti	87
4.4.5	Zásada spätnej zodpovednosti	88
5	APLIKÁCIA A VYHODNOTENIE NAVRHOVANÝCH RIEŠENÍ.....	89
5.1	TEORETICKÉ ZHODNOTENIE EFEKTIVITY NÁVRHOV	89
5.2	PRAKTICKÁ APLIKÁCIA V KONKRÉTNÝCH SCENÁROCH	91
5.2.1	Scenár aplikácie technických opatrení	91
5.2.2	Scenár uplatnenia legislatívnych opatrení	92
5.2.3	Scenár vzdelávacích iniciatív	94
5.2.4	Scenár uplatnenia etických zásad.....	95
5.3	KOMPARATÍVNE ZHODNOTENIE PRÍNOSOV	96
5.4	OBMEDZENIA A VÝZVY	97
	ZÁVĚR.....	99
	SEZNAM POUŽITÉ LITERATURY	101
	SEZNAM OBRÁZKŮ	112
	SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK	113

ÚVOD

Umelá inteligencia (AI) sa v posledných rokoch stala jednou z najvýznamnejších technológií formujúcich digitálne prostredie. Jej schopnosti, ako generovanie realistického obsahu, personalizácia komunikácie či automatizácia procesov, prinášajú značné výhody, no zároveň otvárajú priestor pre nové hrozby v oblasti kybernetickej bezpečnosti. Phishing, ako jedna z najrozšírenejších foriem digitálnych útokov, dosahuje práve vďaka generatívnym AI technológiám novú úroveň sofistikovanosti, personalizácie a presvedčivosti [1]. Tieto útoky, vrátane hlasového phishingu a video impersonácie, sú čoraz účinnejšie, pretože dokážu napodobňovať autentickú komunikáciu a obchádzať tradičné bezpečnostné mechanizmy [2]. Súčasné výskumy zároveň poukazujú na to, že generatívna AI znižuje náklady a čas potrebný na prípravu podvodu, čím rozširuje dostupnosť týchto útokov aj pre menej technicky zdatných aktérov [3].

Táto diplomová práca si kladie za cieľ preskúmať vplyv umelej inteligencie na vývoj phishingových útokov a navrhnúť opatrenia na zmiernenie súvisiacich rizík. V teoretickej časti podrobne analyzuje, ako AI mení formy a techniky phishingu. Pozornosť je venovaná najmä typom útokov, ktoré využívajú nástroje ako personalizované phishingové e-maily založené na zdrojoch z OSINT, deepfake technológie či hlasové klonovanie vo forme tzv. vishingu. Cieľom je identifikovať, ako tieto nové podoby phishingu obchádzajú tradičné obranné mechanizmy a aké nové bezpečnostné výzvy vytvárajú.

Súčasťou práce je aj preskúmanie existujúcich právnych rámcov upravujúcich využívanie umelej inteligencie v kybernetickej bezpečnosti. Práca sa venuje právnym nástrojom na medzinárodnej a národnej úrovni, ako je napríklad AI Act prijatý v Európskej únii v roku 2024 [4], odporúčania OECD [5], či trestnoprávna úprava relevantná pre zneužitie AI v kybernetickom priestore. Cieľom je identifikovať právne výzvy, najmä v oblasti zodpovednosti a vymáhateľnosti regulácie. Záverečnou oblasťou teoretickej časti je etická reflexia problémov spojených so zneužívaním AI, so zameraním na otázky zodpovednosti vývojárov, ochranu súkromia a spravodlivé riešenie škôd spôsobených AI systémami.

Cieľom praktickej časti práce je navrhnúť súbor opatrení, ktoré reagujú na výzvy spojené s využívaním umelej inteligencie pri phishingových útokoch, a zároveň preskúmať ich možnosti uplatnenia v reálnych podmienkach. Tieto opatrenia sú rozdelené do štyroch te-

matických oblastí – technické riešenia, legislatívne a regulačné mechanizmy, vzdelávacie iniciatívy a etické zásady. Každá z týchto oblastí reflektuje inú oblasť problému a ponúka odlišné nástroje na jeho zvládnutie.

Navrhované prístupy sú následne aplikované na modelové scenáre, ktoré umožňujú preskúmať ich praktickú uplatniteľnosť v rôznych kontextoch. Cieľom tejto časti je overiť, do akej miery môžu navrhnuté opatrenia prispieť k obmedzeniu rizík a zvyšovaniu odolnosti voči phishingu využívajúceho AI. Súčasťou praktickej časti je aj teoretické zhodnotenie efektivity jednotlivých návrhov, ako aj identifikácia možných obmedzení, ktoré môžu brániť ich úspešnej implementácii.

I TEORETICKÁ ČÁST

1 ZÁKLADY A VÝVOJ PHISHINGOVÝCH ÚTOKOV S VYUŽITÍM UMELEJ INTELIGENCIE

Phishing predstavuje jednu z najrozšírenejších kybernetických hrozieb, pričom jeho metódy sa neustále vyvíjajú. Táto kapitola sa zameriava na podstatu phishingu, jeho rôzne typy a vývoj v priebehu času. Osobitná pozornosť je venovaná využitiu umelej inteligencie (AI) pri phishingových útokoch, kde moderné technológie umožňujú útočníkom vytvárať presvedčivejšie a sofistikovanejšie podvody. Vysvetlené budú konkrétne spôsoby, akými AI zlepšuje phishingové techniky a zvyšuje ich účinnosť.

1.1 Phishing a jeho základy

Phishing sa vo všeobecnosti definuje ako sociálno-technický útok, pri ktorom útočník cieľi na konkrétne cennosti využitím existujúcej zraniteľnosti. Hrozba je prenášaná prostredníctvom vybraného média do systému obete, pričom sú aplikované techniky sociálneho inžinierstva alebo iné manipulatívne metódy s cieľom presvedčiť obeť, aby vykonala určitú akciu. Táto akcia následne vedie k rôznym druhom škôd, vrátane odcudzenia citlivých údajov, finančných strát či narušenia bezpečnosti systému [6].

V oblasti informačných technológií sa phishing označuje ako kriminálna činnosť založená na metódach sociálneho inžinierstva s cieľom podvodne získať citlivé údaje, ako sú napríklad používateľské mená a heslá. Tento typ útoku sa uskutočňuje tak, že používatelia sú oklamaní a presmerovaní na falošné verzie populárnych webových stránok, kde nevedomky zadávajú svoje prihlasovacie údaje [7]. Na prilákanie nič netušiacich používateľov sa často využíva komunikácia vydávajúca sa za správy od populárnych sociálnych sietí, aukčných portálov, online platobných systémov alebo IT administrátorov [8].

Pojem phishing prešiel vývojom od jednoducho zostavených útokov prostredníctvom jednoduchých správ až po sofistikované techniky zahŕňajúce falšovanie celých webových stránok, ktorých cieľom je oklamať používateľov a prinútiť ich poskytnúť svoje osobné údaje [7]. Okrem toho môžu phishingové e-maily obsahovať odkazy na webové stránky infikované škodlivým softvérom, čím sa zvyšuje riziko ohrozenia bezpečnosti používateľov [8].

1.1.1 Proces phishingového útoku

Všeobecný proces phishingového útoku možno rozdeliť do štyroch fáz [6]:

1. **Plánovacia fáza,**
2. **Prípravná fáza,**
3. **Prevedenie útoku,**
4. **Fáza získavania cenností.**

Phishingový útok vo väčšine prípadov začína zhromažďovaním informácií o cieľovej obeti. Tento proces zahŕňa získavanie rôznych detailov, napríklad informácií o používaných službách alebo pracovisku obete. Práve preto môžu byť úniky údajov, aj keď neobsahujú priamo citlivé informácie, mimoriadne nebezpečné. Ak napríklad dôjde k úniku databázy e-mailových adries určitej služby, útočníci získajú prehľad o jej používateľoch a môžu na nich cieľiť phishingové e-maily, ktoré sa budú vydávať za oficiálnu komunikáciu tejto služby. Pri sofistikovanejších spear phishingových útokoch môžu útočníci analyzovať verejne dostupné informácie zo sociálnych sietí obete. Na základe takto získaných údajov vytvárajú vysoko personalizované phishingové správy, ktoré pôsobia dôveryhodne a zvyšujú pravdepodobnosť, že obeť na podvod zareaguje [9].

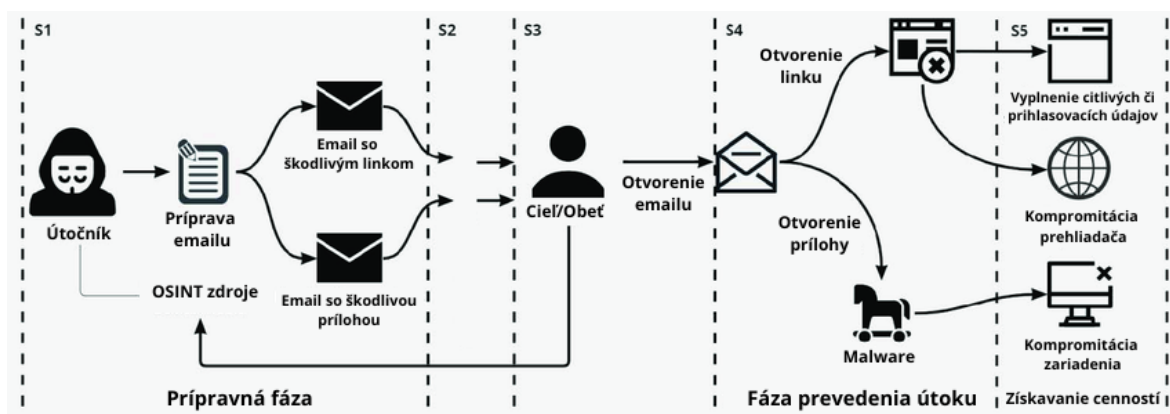
V prípravnej fáze sa útočník rozhoduje, akú metódu útoku zvolí ako počiatočný krok. Zameriava sa na identifikáciu slabých miest, ktoré môže využiť na oklamanie obete [6]. Súčasťou prípravnej fázy je napríklad vytvorenie podvodnej webovej stránky, ktorá vizuálne kopíruje legitímnu stránku cieľovej služby, čím sa zvyšuje pravdepodobnosť, že obeť zadá svoje prihlasovacie údaje. Ďalšou taktikou môže byť nasadenie škodlivého softvéru (malvéru), ktorý sa aktivuje na zariadení obete, či už na počítači alebo mobilnom telefóne. Tento malvér môže napríklad zaznamenávať stlačenia klávesov s cieľom odcudziť heslá alebo iné citlivé údaje. V niektorých prípadoch môže malvér zašifrovať súbory na zariadení obete a následne požadovať výkupné za ich odomknutie [10].

V tretej fáze prebieha samotný phishingový útok. Útočník odošle podvodný e-mail a očakáva, že obeť na návnadu zareaguje. Ďalšie kroky útočníka závisia od povahy podvodu. Ak je napríklad cieľom získanie prihlasovacích údajov k e-mailovému účtu, obeť môže byť presmerovaná na falošnú prihlasovaciu stránku. Po zadaní údajov útočník získa prístup

k e-mailovému účtu, čo mu umožní nielen získať ďalšie citlivé informácie, ale aj pokračovať v útoku rozosielaním phishingových e-mailov kontaktom obete, čím sa phishingová kampaň ďalej rozširuje [9].

V záverečnej fáze phishingového útoku útočník získava citlivé údaje obete, ktoré následne zneužíva na rôzne škodlivé účely. Medzi najčastejšie patrí finančný podvod, vydieranie alebo krádež identity [6]. Po úspešnom útoku má útočník prístup k osobným informáciám obete, ako sú čísla kreditných kariet, prihlasovacie údaje k rôznym webovým službám či podrobnosti o bankových účtoch. Tieto údaje môžu byť následne využité na nelegálne finančné transakcie, neoprávnený prístup k online platformám alebo na podvodné aktivity, pri ktorých sa útočník vydáva za svoju obeť s cieľom získať ďalšie výhody alebo poškodiť jej reputáciu [10].

Obrázok 1.1 ilustruje jednotlivé fázy a priebeh phishingového útoku, ktorý je realizovaný prostredníctvom e-mailovej správy obsahujúcej škodlivý odkaz alebo infikovanú prílohu. Schéma zachytáva typickú sériu krokov, od prípravy útoku, cez interakciu s obeťou, až po získanie citlivých údajov alebo kompromitáciu zariadenia [11].



Obrázek 1.1 Proces phishingového útoku

(zdroj: IEEE Access, 2022 [11])

1.1.2 Typy phishingových útokov

Phishingové útoky môžeme vo všeobecnosti rozdeliť podľa viacerých kritérií, pričom najčastejšie sa rozlišujú na základe spôsobu útoku, cieľovej skupiny alebo použitých techník na oklamanie obete. Základné typy phishingu môžeme rozdeliť:

1. Podľa cieľovej skupiny:

- **Hromadný phishing:** Hromadný phishing predstavuje klasický typ phishingového útoku, pri ktorom útočník rozosiela podvodné správy veľkému množstvu potenciálnych obetí s cieľom zachytiť čo najviac odpovedí. Hoci miera úspešnosti takýchto útokov je pomerne nízka, útočníci sa spoliehajú na to, že z tisícok či miliónov oslovených obetí sa vždy nájde určité percento, ktoré na podvod naletí. Tieto phishingové kampane sú často jednoduché a nevyžadujú vysokú mieru sofistikovanosti. V mnohých prípadoch sú vykonávané pomocou verejne dostupných phishingových nástrojov. Typickým príkladom sú podvodné správy s oznámením o falošnej výhre, napríklad „Gratulujeme, vyhrali ste nový iPhone!“, ktoré nabádajú obeť na kliknutie na škodlivý odkaz alebo zadanie osobných údajov [12].
- **Spear phishing:** Spear phishing je cielený typ phishingového útoku, pri ktorom sa útočník zameriava na konkrétnych jednotlivcov alebo organizácie. Na rozdiel od hromadného phishingu sú tieto útoky prispôsobené tak, aby pôsobili dôveryhodne a ťažšie odhaliteľne [13]. Úspešnosť spear phishingu spočíva v dôkladnom prieskume obete, pri ktorom útočníci zhromažďujú verejne dostupné informácie, napríklad zo sociálnych sietí alebo firemných webových stránok, aby mohli vytvoriť podvodnú komunikáciu, ktorá pôsobí ako legitímna správa od dôveryhodného odosielateľa. Na základe získaných údajov využívajú techniky sociálneho inžinierstva na formulovanie presne prispôsobených e-mailov alebo požiadaviek, čím zvyšujú pravdepodobnosť, že obeť podvod nerozpozna a poskytne citlivé údaje [14].
- **Whaling:** Whaling je špecifický typ spear phishingu, pri ktorom sa ale útočníci orientujú na tzv. „veľké ryby“, ako sú generálni riaditelia (CEO), finanční riaditelia (CFO) alebo vedúci oddelení (napr. HR), s cieľom získať prístup k finančným a dôverným informáciám [13]. Aby bol whaling úspešný, útočníci vykonávajú dôkladný výskum, aby mohli presvedčivo napodobniť komunikáciu konkrétnej vysokej authority. Často sa snažia vydávať za vysokopostavenú osobu preto, aby presvedčili zamestnancov, a tí poskytnú citlivé údaje, previedli finančné prostriedky alebo vykonali iné požadované akcie bez otázok [12].

2. Podľa spôsobu doručenia:

- **E-mailový phishing:** najbežnejšia forma, pri ktorej útočníci posielajú podvodné e-maily predstierajúce legitímne inštitúcie.
- **SMS phishing (smishing):** útoky prostredníctvom textových správ, často obsahujúce odkazy na falošné weby alebo škodlivé súbory.
- **Telefonický phishing (vishing):** útočník sa vydáva za autoritu (napr. bankového pracovníka) a pokúša sa vylákať citlivé informácie telefonicky
- **Sociálne siete:** phishingové odkazy a správy šírené cez Facebook, Instagram, WhatsApp a iné platformy [8, 12].

3. Podľa techniky útoku:

- **Clone phishing:** Clone phishing je typ phishingového útoku, pri ktorom útočníci kopírujú legitímny e-mail, ktorý predtým odoslala dôveryhodná organizácia. Namiesto toho, aby sa vydávali za konkrétneho používateľa alebo organizáciu, hackeri manipulujú s pôvodným e-mailom a menia v ňom odkazy. Týmto spôsobom vytvárajú presne rovnakú verziu e-mailu, ale s podvodným odkazom, ktorý vedie na falošnú stránku [12].
- **Man-in-the-middle phishing:** MitM phishing je sofistikovaná technika aktívneho útoku, pri ktorej útočník zachytáva komunikáciu medzi používateľom a legitímnou službou s cieľom ukradnúť citlivé informácie. Na rozdiel od tradičného phishingu, pri ktorom je obeť oklamaná, aby zadala svoje prihlasovacie údaje na falošnej prihlasovacej stránke, MitM phishing zahŕňa sprostredkovateľa, ktorý sa nachádza medzi obeťou a skutočnou službou a zachytáva údaje vymieňané medzi nimi. Tento typ útoku umožňuje útočníkom obísť pokročilé bezpečnostné opatrenia, ako je dvojfaktorová autentifikácia (2FA), pretože zachytávajú nielen prihlasovacie údaje, ale aj jednorazové kódy zasielané na mobilné zariadenia obetí [15].
- **Malware phishing:** Phishing založený na malvéri je technika, pri ktorej útočníci pripojujú škodlivé programy, známe ako malvér, k zdanlivo dôveryhodným e-mailom či falošným webovým stránkam. Keď obeť interaguje s takýmto digitálnym obsahom, malvér skrytý v ňom infikuje ich zariadenie. Po infikovaní

môžu útočníci získať prístup k citlivým informáciám, ovládať zariadenia a vykonávať ďalšie škodlivé činnosti. Phishing založený na malvári sa môže rýchlo rozšíriť a spôsobiť rozsiahle škody [16].

- **Angler phishing:** Angler phishing predstavuje novší typ phishingového útoku, ktorý sa zameriava na používateľov sociálnych médií. Tento druh kybernetického útoku sa prevažne vyskytuje na platformách ako Facebook, Twitter a Instagram. Cieľom angler phishingu je osloviť nespokojných zákazníkov v rámci zákazníckej starostlivosti a získať ich osobné údaje na ďalšie zneužitie. Útočníci sa často vydávajú za legitímne spoločnosti alebo zákaznícku podporu, čím vytvárajú dojem dôveryhodnosti a manipulujú obeť k poskytnutiu citlivých informácií [17, 18].

1.1.3 Znak phishingových správ

Phishingové útoky možno identifikovať podľa niekoľkých typických znakov, ktoré varujú používateľov pred potenciálnymi hrozbami. Medzi bežné znaky phishingových správ patrí použitie strašiacich taktík a naliehavých výziev. Tieto správy často obsahujú predmety ako „URGENT: Platba po splatnosti“, ktoré sa snažia vyvolať pocit naliehavosti a vyžadujú okamžitú pozornosť. Ďalším znakom phishingu sú nevyžiadané hovory od údajne legitímnych „agentov zákazníckej podpory“, pričom útoky typu vishing môžu informovať o podvodoch na kreditnej karte alebo iných naliehavých problémoch [19].

Príznakom phishingu sú aj podozrivé alebo nesprávne napísané webové a e-mailové adresy. Rôzne varianty adries URL, ktoré sa na prvý pohľad zdajú byť legitímne, môžu skryť nebezpečné webové stránky. Umiestnením kurzora myši na odkaz sa zobrazí skutočná adresa URL, čo pomôže odhaliť potenciálne hrozby [20]. Neočakávané prílohy v e-mailoch by mali byť považované tiež za varovný signál. Ak e-mail pochádza od neznámeho odosielateľa alebo príloha nedáva zmysel, odporúča sa ju neotvárať, pretože môže obsahovať škodlivý softvér [19]. Ďalším varovným signálom je, ak sa ponuka javí ako „príliš dobrá na to, aby bola pravdivá“. Útočníci často predkladajú lákavé cenové ponuky alebo neodolateľné výhry, aby nalákali obeť a podnietili ich k akcii. Príkladmi takýchto návnad môžu byť správy o výhre nového iPhoneu, lotérie alebo iných atraktívnych cien [20].

1.2 Vývoj phishingových útokov

Vývoj phishingových útokov prešiel od svojich počiatkov v 90. rokoch 20. storočia až po súčasnosť, kde útočníci využívajú pokročilé techniky strojového učenia a umelej inteligencie. Vývoj phishingových útokov ukazuje, že útočníci neustále prispôbujú svoje metódy novým technológiám a správaniu používateľov.

1.2.1 Počiatky phishingu

Bezpečnostné hrozby v oblasti počítačovej technológie sa objavili už v 50. rokoch 20. storočia. V tom období sa začali riešiť otázky ochrany pamäte a vznikali prvé šifrovacie a kontrolné mechanizmy. Medzi predchodcov phishingu patrila metóda známa ako „phone phreaking“, pri ktorej hackeri zneužívali telekomunikačné systémy na uskutočňovanie bezplatných hovorov [21].

Prvý zaznamenaný phishingový útok sa odohral 2. januára 1996 a jeho cieľom bola populárna internetová služba America Online (AOL), ktorá v tom čase mala milióny používateľov. Hackeri z tzv. warez komunity využili AOL messenger a e-mailový systém na rozposielanie správ, v ktorých sa vydávali za zamestnancov spoločnosti. Od používateľov žiadali potvrdenie ich účtov alebo fakturačných údajov. Mnohí sa nechali oklamať, keď že s podobným podvodom sa dovtedy nestretli. Pôvodne phisher generovali náhodné čísla kreditných kariet na vytváranie falošných účtov, no AOL v roku 1995 zaviedla opatrenia, ktoré túto metódu zablokovali. Hackeri preto prešli na sofistikovanejšie techniky sociálneho inžinierstva. Situácia sa vyhrotila natoľko, že AOL musela do svojich e-mailových a chatovacích služieb implementovať varovania, aby používateľov ochránila pred podvodníkmi [22].

Zmena taktiky útočníkov sa prejavila aj začiatkom roku 2000, keď sa svet stal obeťou neslávne známeho vírusu nazývaného „Love Bug“. Používatelom začali prichádzať e-maily s predmetom „ILOVEYOU“ a textom *"Skontrolujte prosím priložený LOVELETTER od mňa"*. Príloha pôsobila nevinne – vyzerala ako obyčajný .txt súbor, no v skutočnosti obsahovala škodlivý skript. Po jeho otvorení sa vírus okamžite aktivoval, prepísal obrazové súbory v infikovanom zariadení a automaticky sa rozoslal všetkým kontaktom v Outlooku obeť [23]. Vírus sa šíril obrovskou rýchlosťou a nakoniec infikoval približne 45 miliónov počítačov so systémom Windows. Prípád Love Bug jasne ukázal, aké nebezpečné môže byť spojenie sociálneho inžinierstva a technických zraniteľností [24].

1.2.2 Evolúcia a diverzifikácia útokov

V rokoch 2000 až 2010 prešli phishingové útoky výraznou premenou a postupne sa prispôbovali novým technologickým trendom a komunikačným kanálom. Kým na začiatku sa phishing šíril prevažne cez e-maily, postupne sa začali využívať aj SMS správy (smishing) a telefonické hovory (vishing). Tento posun bol dôsledkom rastúcej popularity elektronického obchodu a online platobných systémov, ako sú eBay a PayPal, ktoré sa stali častým terčom kybernetických podvodníkov [22, 25].

V prvej polovici dekády hackeri ešte len experimentovali s rôznymi phishingovými technikami, no v druhej polovici sa tieto útoky stali systematickejšími a organizovanejšími. Útočníci začali využívať premyslené podvodné e-maily, v ktorých sa vydávali za známe inštitúcie a od obetí požadovali aktualizáciu údajov o kreditných kartách či prihlasovacích údajov. Zároveň phishereri začali masovo registrovať falošné webové domény, ktoré na prvý pohľad vyzerali ako legitímne stránky, čím zvýšili úspešnosť svojich útokov.

Phishing sa v tomto období stal pevnou súčasťou čierneho trhu, pričom jeho rozšíreniu napomohol vývoj špecializovaného softvéru na automatizované spracovanie ukradnutých údajov. Významným míľnikom bol rok 2008, keď sa objavil Bitcoin a ďalšie kryptomeny. Tie poskytli kyberzločincovi anonymný a decentralizovaný spôsob finančných transakcií, čím phishingové aktivity získali nový rozmer [25].

Nedostatočná informovanosť o kybernetických hrozbách umožnila útočníkom organizovať rozsiahle phishingové kampane, ktoré zasiahli mnohé firmy a inštitúcie. V rokoch 2004 – 2005 podľa organizácie Phishing.org prišlo len v USA približne 1,2 milióna ľudí o celkovo 929 miliónov USD v dôsledku phishingových útokov [22].

Aj keď hlavným cieľom phishingu bol pôvodne finančný zisk, po roku 2010 sa začali objavovať jeho nové formy, ktoré zasahovali aj do politickej a strategickej sféry. V roku 2016 sa phishingový útok stal príčinou kompromitácie e-mailového účtu Johna Podestu, predsedu kampane Hillary Clintonovej, čo umožnilo neoprávnený prístup k citlivým informáciám. Rovnakým spôsobom boli napadnuté aj e-mailové účty Colina Powella a Demokratického národného výboru (DNC). Tieto útoky ukázali, že phishing nepredstavuje len riziko pre jednotlivcov a spoločnosti, ale môže slúžiť aj ako účinný nástroj kybernetickej špionáže a manipulácie v geopolitickom prostredí [26].

1.2.3 Nástup sociálnych médií

Nástup sociálnych médií zásadne ovplyvnil vývoj phishingových útokov, čím sa otvorili nové možnosti pre kybernetických zločincov. Tieto platformy, ako sú Facebook, LinkedIn či Twitter, zhromažďujú obrovské množstvo osobných údajov, ktoré útočníci využívajú na vytváranie sofistikovanejších a cielenejších phishingových kampaní [27]. Podľa Statista bolo k januáru 2024 zo 5,35 miliardy používateľov internetu na svete až 5,04 miliardy používateľov sociálnych médií. Sociálne siete tak predstavujú zlatú baňu osobných informácií, ktoré môžu byť zneužitú na rôzne škodlivé aktivity [28].

Pred érou sociálnych médií boli osobné informácie na internete obmedzené a ťažko dostupné. S rozmachom platforiem sociálnych sietí však používatelia začali dobrovoľne zdieľať množstvo detailov zo svojho života, vrátane pracovných pozícií, kontaktov, záujmov či každodenných aktivít. Tieto údaje sa stali cenným zdrojom pre phisherov, ktorí ich využívajú na detailné profilovanie svojich obetí. Takto získané informácie im umožňujú vytvárať presvedčivé a personalizované podvodné správy, ktoré zvyšujú pravdepodobnosť úspešného útoku. Takéto útoky, známe ako spear phishing, sú oveľa nebezpečnejšie, pretože obeť majú tendenciu dôverovať správam prichádzajúcim od známych osôb alebo autorít [27].

Okrem získavania informácií útočníci využívajú sociálne siete aj na priame šírenie phishingových útokov. Vytvárajú falošné profily alebo stránky, ktoré napodobňujú legitímne organizácie či osoby. Prostredníctvom týchto falošných účtov šíria podvodné odkazy alebo aplikácie, ktoré po kliknutí môžu od obetí získať citlivé údaje alebo infikovať ich zariadenia škodlivým softvérom. Napríklad, pod zámienkou súťaže alebo exkluzívnej ponuky môžu nájsť používateľov na zadanie svojich prihlasovacích údajov alebo iných citlivých informácií [29].

1.2.4 Využitie umelej inteligencie a strojového učenia

Rozvoj umelej inteligencie a strojového učenia zásadne zmenil taktiky phishingových útokov a dramaticky zvýšil ich účinnosť. Zatiaľ čo v minulosti bolo možné phishingové e-maily často odhaliť vďaka gramatickým chybám alebo podozrivému štýlu komunikácie, dnes generatívna AI umožňuje podvodníkovi vytvárať vysoko presvedčivé, jazykovo bezchybné a personalizované správy, ktoré je čoraz ťažšie rozoznať od legitímnej komunikácie [1].

Jednou z najvýznamnejších výhod využitia AI v phishingu je jej schopnosť automatizovať a škálovať útoky. Nástroje poháňané generatívnou AI dokážu odstrániť jazykové bariéry, prispôbiť obsah rôznym cieľovým skupinám a v reálnom čase odpovedať na interakcie obetí. To znamená, že phishingové kampane už nie sú obmedzené len na masovo rozosielané generické e-maily, ale môžu byť vysoko personalizované a prispôbené konkrétnym jednotlivcom alebo organizáciám. Táto sofistikovanosť zvyšuje pravdepodobnosť, že obeť nevedomky poskytne citlivé údaje alebo uverí podvodu [2]. Podľa spoločnosti SlashNext sa od konca roku 2022, teda približne v čase, keď sa na scéne objavil ChatGPT, počet škodlivých phishingových e-mailov zvýšil o 1 265 %. Generatívna AI umožnila kybernetickým útočníkom v roku 2022 získať viac ako 2 miliardy dolárov prostredníctvom podvodov, pričom významne zjednodušila falšovanie domén a získavanie prístupu k citlivým údajom [1].

Okrem textových správ, dnes AI umožňuje aj tvorbu realistických zvukových a obrazových podvodov, ktoré výrazne zvyšujú mieru prepracovanosti phishingových útokov. Pomocou týchto technológií môžu útočníci autenticky napodobniť hlasy iných osôb, čím zvyšujú pravdepodobnosť, že obeť podľahne manipulácii a vykoná požadované akcie [2]. Podrobnejší pohľad na využitie AI vo phishingových útokoch bude popísaný v nasledujúcej kapitole 1.3.

1.3 Princípy využitia AI vo phishingových útokoch

V tejto kapitole si bližšie popíšeme, ako umelá inteligencia mení spôsob, akým útočníci vykonávajú phishingové útoky, a aké technológie im umožňujú zvyšovať efektivitu a sofistikovanosť týchto útokov. Najprv definujeme, čo znamená phishingový útok poháňaný AI, a objasníme jeho hlavné charakteristiky. Následne sa zameriame na kľúčové technológie, ktoré sú v súčasnosti využívané pre AI phishing, vrátane generatívnej AI, deepfake technológií, hlasového klonovania, automatizácie a analýzy dát, ako aj behaviorálnej analýzy a prediktívnych modelov. Tieto technológie umožňujú útočníkom vytvárať presvedčivejšie podvody, personalizovať útoky a cieľiť na zraniteľné obeť s vyššou presnosťou.

1.3.1 Definícia phishingového útoku poháňaného AI

Phishingový útok poháňaný umelou inteligenciou využíva generatívnu AI na tvorbu mimoriadne presvedčivých a personalizovaných phishingových správ. Tieto útoky sú založené na sofistikovaných algoritmoch, ktoré analyzujú rozsiahle množstvo údajov zo sociálnych mé-

dií, firemných webových stránok a iných verejných zdrojov. Vďaka tomu dokážu napodobniť autentické komunikačné štýly a obsah, čím pôsobia dôveryhodnejšie.

Na rozdiel od tradičných phishingových útokov, ktoré často využívajú generické šablóny, phishing poháňaný generatívnou AI dokáže vytvárať správy prispôsobené konkrétnym jednotlivcom alebo oddeleniam v organizáciách. Táto úroveň personalizácie značne sťažuje rozpoznanie podvodu, čím sa zvyšuje šanca na úspech útoku [30]. Takéto podvodné správy môžu obsahovať odkazy na nedávne aktivity obete, ako sú online nákupy, záujmy alebo interakcie, čím sa zvyšuje pravdepodobnosť, že obeť na ne zareaguje. Navyše, AI dokáže vygenerovať realistické kópie legitímnych webových stránok, čo výrazne sťažuje rozpoznanie podvodu [1].

Využitie generatívnej AI tiež umožňuje kyberzločincovi obísť bežné bezpečnostné opatrenia, ktoré sú navrhnuté na identifikáciu známych phishingových vzorcov. V dôsledku toho sú tieto útoky ťažšie odhaliteľné a štandardné nástroje kybernetickej bezpečnosti môžu mať problémy s ich detekciou a prevenciou [30]. Aby sme to zhrnuli, uveďme niekoľko kľúčových dôvodov, prečo je sociálne inžinierstvo poháňané umelou inteligenciou také nebezpečné [31]:

- **Masívne zbieranie dát** – Útočníci pomocou AI zbierajú obrovské množstvo informácií zo sociálnych sietí, firemných webov či verejne dostupných databáz. Vďaka tomu dokážu vytvoriť uveriteľné príbehy, ktoré na obete pôsobia vierohodne a autenticky.
- **Pokročilá automatizácia** – Vďaka umelej inteligencii môžu kyberzločinci vygenerovať tisíce personalizovaných phishingových e-mailov za pár sekúnd. Táto schopnosť umožňuje škálovať útoky na úroveň, ktorú by manuálne vytváranie podvodných správ nikdy nedosiahlo.
- **Využívanie viacerých kanálov** – AI už nie je obmedzená iba na textové správy. Pokročilé technológie, ako je deepfake video a hlasové klonovanie, umožňujú útočníkovi napodobniť hlasy manažérov, kolegov alebo obchodných partnerov. To výrazne zvyšuje šance na úspešné oklamanie obete a zároveň sťažuje odhalenie klasickými bezpečnostnými opatreniami [31].

1.3.2 Hlavné technológie AI vo phishingu

V tejto podkapitole sa zameriame na kľúčové technológie umelej inteligencie, ktoré sa najčastejšie využívajú pri phishingových útokoch. Zarovň ukážeme, ako ich útočníci zneužívajú na oklamanie obetí.

1. Generatívna AI a spracovanie prirodzeného jazyka (NLP)

Spracovanie prirodzeného jazyka (NLP) predstavuje kľúčovú oblasť umelej inteligencie a počítačovej vedy, ktorá umožňuje strojom porozumieť ľudskému jazyku a odvodiť z neho význam tak, ako to dokážu ľudia. Ide o proces, pri ktorom sa AI učí chápať pravidlá a syntax jazyka prostredníctvom komplexných algoritmov, ktoré tieto pravidlá reprezentujú. Tieto algoritmy následne využíva na vykonávanie rôznych úloh, ako je analýza textu, preklad alebo generovanie nových správ [32]. Vďaka tomu dokáže generatívna AI, najmä modely založené na NLP, ako sú ChatGPT či Gemini, vytvárať texty, ktoré sú takmer nerozoznateľné od ľudského písania [33]. Tieto modely sú trénované na obrovskom množstve dát, čo im umožňuje generovať gramaticky správne a kontextovo relevantné správy prispôbené konkrétnemu účelu. V oblasti phishingu útočníci zneužívajú generatívnu AI na tvorbu presvedčivých e-mailov, SMS správ či falošných príspevkov na sociálnych sieťach. Napríklad dokážu vytvoriť e-mail, ktorý vyzerá, akoby ho poslal šéf firmy, a žiada zamestnanca o citlivé informácie alebo prevod peňazí. Podľa štúdie od Hoxhunt z roku 2023 je generatívna AI schopná vytvoriť tisíce personalizovaných phishingových e-mailov za pár sekúnd, čím výrazne zvyšuje efektivitu a rozsah útokov [31].

2. Deepfake technológie a generatívne adversariálne siete (GAN)

Deepfake technológie sú založené na generatívnych adversariálnych sieťach (GAN), ktoré pozostávajú z dvoch neurónových sietí – generátora a diskriminátora. Generátor vytvára falošný obsah, ako sú videá, obrázky či zvuk, zatiaľ čo diskriminátor ho posudzuje a určuje, či je obsah pravý alebo falošný. Tieto dve siete sú trénované proti sebe, čím sa postupne zlepšujú: generátor produkuje čoraz realistickejší obsah a diskriminátor sa stáva presnejším v jeho odhalovaní [34]. Mechanizmus GAN je navrhnutý tak, aby minimalizoval pravdepodobnosť odlišenia vytvoreného obsahu od autentického, čo vedie k vytváraniu mimoriadne presvedčivých falošných médií [35]. V kontexte

phishingových útokov útočníci využívajú deepfake videá na napodobnenie známych osôb, aby presvedčili obeť o pravosti komunikácie. Príkladom môže byť falošné video, v ktorom generálny riaditeľ firmy žiada zamestnanca o odoslanie citlivých údajov, ako sú prihlasovacie údaje alebo finančné informácie. Podľa správy Europolu z roku 2022 o kriminálnom využívaní deepfake technológií, sa tieto metódy čoraz častejšie zneužívajú na podvody a kybernetické útoky, pričom ich detekcia je pre bežných používateľov takmer nemožná [36].

3. Hlasová syntéza a klonovanie hlasu

Hlasová syntéza a technológie klonovania hlasu umožňujú vytváranie realistických napodobení hlasov na základe krátkych zvukových vzoriek. Moderné systémy, ako napríklad tie od spoločností Descript či Respeecher, dokážu vygenerovať hlas, ktorý je takmer nerozoznateľný od originálu, už po analýze nahrávky trvajúcej len pár minút. Tieto technológie využívajú pokročilé algoritmy umelej inteligencie, ktoré analyzujú charakteristiky hlasu, ako sú tón, intonácia či rečové vzorce, a následne ich verne napodobňujú [37]. V oblasti phishingových útokov sa hlasová syntéza zneužíva na hlasový phishing, známy aj ako vishing. Pri takomto útoku útočník napodobní hlas známej osoby a telefonicky žiada obeť o citlivé informácie, ako sú prihlasovacie údaje alebo finančné prostriedky. Príkladom je prípad z roku 2019, kedy útočníci použili klonovaný hlas riaditeľa firmy na oklamanie zamestnanca, ktorý následne previedol 243 000 dolárov na falošný účet. Ide o prvý zaznamenaný prípad falošného hlasu generovaného umelou inteligenciou použitého pri podvode [38].

4. Automatizácia a analýza dát (data mining)

Umelá inteligencia umožňuje útočníkom automatizovať proces zhromažďovania údajov (data mining) z verejne dostupných zdrojov. Data mining je technológia, ktorá umožňuje spracovanie veľkých objemov dát s cieľom odhaliť skryté vzorce, korelácie a užitočné informácie. Tento proces využíva metódy umelej inteligencie, strojového učenia, štatistiky a databázových systémov na extrakciu poznatkov z dát, ktoré by manuálnou analýzou nebolo možné identifikovať [39]. Vo phishingových útokoch útočníci využívajú data mining na automatizované zhromažďovanie údajov z verejne dostupných zdrojov, ako sú sociálne siete, firemné webové stránky či úniky dát. Napríklad pomocou techník web scrapingu dokážu získať informácie o potenciálnych

obetiach, vrátane ich pracovných pozícií, kontaktov alebo osobných záujmov. Tieto údaje následne slúžia na personalizáciu útokov, čím sa zvyšuje ich úspešnosť. Podľa správy od Verizonu z roku 2023, až 70 % phishingových útokov vychádza z údajov získaných z verejne dostupných zdrojov [40].

5. Využívanie OSINT zdrojov

Zhromažďovanie informácií z otvorených zdrojov, známe ako OSINT (Open Source Intelligence), je metodológia, ktorá sa vo phishingových útokoch často kombinuje s data miningom. OSINT zahŕňa zber verejne dostupných dát z rôznych zdrojov, ako sú sociálne siete, diskusné fóra, verejné databázy, médiá či webové stránky. Na rozdiel od data miningu, ktorý je technológiou zameranou na hĺbkovú analýzu dát, OSINT sa sústreďuje na legálne získavanie informácií z verejných zdrojov. Útočníci využívajú OSINT na identifikáciu potenciálnych cieľov a zhromažďovanie údajov, ktoré následne spracúvajú pomocou data miningových techník. Napríklad údaje získané z LinkedInu o zamestnancoch firmy môžu byť analyzované data miningovými algoritmami, aby útočníci určili, kto je najzraniteľnejší voči phishingovému útoku, alebo vytvorili presvedčivý e-mail napodobňujúci komunikáciu s kolegom. Podľa štúdie od InfosecURITY.sk, uskutočnenej v roku 2023, platformy ako Shodan či Censys, ktoré sa bežne používajú v OSINT, umožňujú útočníkom skenovať internetové servery a odhaľovať zraniteľnosti, ktoré môžu využiť pri cielených útokoch. Kombinácia OSINT a data miningu tak zvyšuje efektivitu phishingových kampaní, pretože útočníci dokážu rýchlo získať a analyzovať veľké množstvo dát na vytvorenie personalizovaných a vierohodných podvodov [39, 41].

6. Behaviorálna analýza a prediktívne modely

Behaviorálna analýza a prediktívne modely sú technológie umelej inteligencie, ktoré umožňujú identifikovať najzraniteľnejšie ciele phishingových útokov na základe analýzy ich správania. Tieto modely využívajú algoritmy strojového učenia na sledovanie a vyhodnocovanie vzorcov správania používateľov, ako sú frekvencia klikania na odkazy, reakcie na e-maily, čas strávený na sociálnych sieťach či spôsob interakcie s online obsahom. Na základe týchto dát dokážu prediktívne modely odhadnúť, ktorí jednotlivci sú najviac náchylní na podvod [42]. Napríklad behaviorálna analýza môže

odhalit', že určitý zamestnanec často komunikuje s konkrétnym dodávateľom, a útočník následne vytvorí falošný e-mail napodobňujúci štýl tejto komunikácie. Podľa štúdie z roku 2022 od IBM Security, prediktívne modely zvyšujú úspešnosť cielených phishingových útokov až o 30 %, pretože útočníci dokážu presne určiť, kto je najzraniteľnejší a aký typ útoku bude najefektívnejší [43].

2 PRAKTICKÉ FORMY PHISHINGOVÝCH ÚTOKOV POHÁŇANÝCH UMELOU INTELIGENCIOU

Po teoretickom základe, v rámci ktorého boli popísané princípy a technológie využitia umelej inteligencie vo phishingových útokoch, sa teraz zameriame na ich praktické uplatnenie. V tejto kapitole podrobne rozoberieme jednotlivé typy phishingových útokov poháňaných AI, vrátane phishingových e-mailov, deepfake videí a hlasového phishingu (vishingu). Ku každému typu uvedieme konkrétne príklady, ktoré ilustrujú, ako útočníci využívajú AI na oklamanie obetí, a ukážeme, aké dopady môžu mať takéto útoky na jednotlivcov a organizácie. Cieľom je poskytnúť hlbší pohľad na reálne hrozby, ktoré tieto technológie prinášajú, a zdôrazniť potrebu efektívnych obranných stratégií.

2.1 Phishingové e-maily poháňané AI

2.1.1 Ako AI zlepšuje phishingové e-maily

Umelá inteligencia prináša revolúciu v spôsobe, akým útočníci vytvárajú phishingové emaily, a robí ich oveľa nebezpečnejšími a presvedčivejšími. Vďaka generatívnym modelom AI, ako sú ChatGPT, Gemini alebo iné pokročilé nástroje na spracovanie prirodzeného jazyka (NLP), dokážu útočníci vytvárať e-maily, ktoré sú gramaticky správne, kontextovo relevantné a prispôbené konkrétnym obetiam. Útočníci najprv zhromažďujú údaje z verejne dostupných zdrojov, ako sú sociálne siete, firemné webové stránky či úniky dát, pomocou techník, ako je web scraping alebo OSINT (Open Source Intelligence). Tieto údaje potom slúžia ako vstup pre generatívne AI modely, ktoré na ich základe generujú personalizované správy napodobňujúce štýl písania známych osôb, napríklad šéfov, kolegov alebo obchodných partnerov [1, 33].

AI navyše dokáže optimalizovať predmet e-mailu, čas odoslania a dokonca aj tón správy, aby maximalizovala pravdepodobnosť, že obeť e-mail otvorí a zareaguje naň. Podľa výskumu od SoSafe, ktorý uskutočnili v roku 2023, až 78 % ľudí otvorí phishingový e-mail vytvorený AI a 21 % z nich klikne na škodlivý obsah, ako sú odkazy alebo prílohy [44]. Ďalšia štúdia od Harvard Business Review v roku 2024 ukázala, že 60 % účastníkov podľa AI-automatizovaným phishingovým útokom, čo je porovnateľné s úspešnosťou phishingových správ vytvorených ľudskými expertmi [45]. Generatívna AI zároveň urýchľuje proces

tvorby útokov – SoSafe uvádza, že takéto nástroje dokážu zrýchliť prípravu phishingových kampaní najmenej o 40 %, čo umožňuje útočníkom škálovať svoje aktivity na masívnu úroveň [44]. Okrem toho AI odstraňuje tradičné indikátory podvodu, ako sú gramatické chyby či neobvyklý jazyk, ktoré boli v minulosti typické pre phishingové e-maily. Podľa Mailgun sa útočníci čoraz častejšie uchýľujú k tzv. hyperpersonalizovaným útokom, kde AI využíva údaje o obeti (napr. jej meno, pracovnú pozíciu, nedávne aktivity) na vytvorenie e-mailu, ktorý pôsobí mimoriadne autenticky [1]. Caniphish dodáva, že AI dokáže automatizovať celý proces útoku – od generovania e-mailov, cez analýzu odpovedí obetí, až po vytváranie protiodpovedí, čím sa útoky stávajú dynamickejšími a prispôsobivejšími [46]. Tieto trendy ukazujú, že phishingové e-maily poháňané AI sú čoraz sofistikovanejšie a predstavujú vážnu hrozbu pre jednotlivcov aj organizácie.

Na záver zhrňme do bodov hlavné spôsoby, akými AI zlepšuje phishingové e-maily:

- **Personalizácia správ:** AI vytvára hyperpersonalizované e-maily napodobňujúce štýl písania známych osôb na základe údajov z verejne dostupných zdrojov.
- **Optimalizácia e-mailu:** Dokáže optimalizovať predmet e-mailu, čas odoslania a tón správy, aby maximalizovala pravdepodobnosť, že obet' e-mail otvorí.
- **Odstránenie indikátorov podvodu:** AI eliminuje gramatické chyby a neobvyklý jazyk, čím e-maily pôsobia autentickejšie.
- **Zrýchlenie prípravy kampaní:** Umožňuje zrýchliť tvorbu phishingových kampaní najmenej o 40 %, čo zvyšuje ich rozsah a efektivitu.
- **Automatizácia celého procesu:** AI automatizuje generovanie e-mailov, analýzu odpovedí obetí a vytváranie protiodpovedí, čím robí útoky dynamickejšími.
- **Vyššia úspešnosť útokov:** Podľa výskumov až 78 % ľudí otvorí AI phishingový e-mail a 21 % klikne na škodlivý obsah, čo zvyšuje pravdepodobnosť úspechu útoku.

2.1.2 Používané technológie

Na generovanie phishingových e-mailov s využitím AI útočníci používajú rôzne technológie, predovšetkým pokročilé veľké jazykové modely (LLM), ktoré im umožňujú automatizovať a personalizovať útoky na masívnu škálu.

1. Komerčné jazykové modely

Medzi najčastejšie používané technológie patria komerčné modely, ako sú ChatGPT alebo Gemini, ktoré však útočníci musia obísť pomocou tzv. „jailbreakingu“. Ide o techniku, pri ktorej sa špecifickými príkazmi (promptmi) obchádzajú etické obmedzenia týchto modelov. Napríklad útočníci môžu model požiadať, aby napísal e-mail v mene fiktívnej osoby, alebo aby vytvoril iba vzorový e-mail, čím obíde jeho bezpečnostné pravidlá. Tieto modely sú trénované na obrovských datasetoch textov, čo im umožňuje generovať presvedčivé a gramaticky správne správy, ktoré napodobňujú štýl legitímnych komunikácií [47].

2. Vlastné škodlivé jazykové modely

Okrem komerčných modelov si útočníci vytvárajú aj vlastné škodlivé LLM, ktoré sú špeciálne navrhnuté na kriminálne účely a nemajú žiadne etické obmedzenia. Príkladmi takýchto modelov sú FraudGPT, WormGPT a PoisonGPT, ktoré sú dostupné na dark webe. Takéto modely sú často trénované na špecifických datasetoch, napríklad na textoch súvisiacich s malvérom alebo na ukradnutých e-mailových komunikáciách, aby dokázali lepšie napodobňovať legitímne správy [47].

- **FraudGPT**, sa objavil na dark webových fórach a Telegram kanáloch v júli 2023, a je inzerovaný ako nástroj na tvorbu „neodhaliteľného malvéru“, phishingových stránok a personalizovaných podvodných správ. Podľa Netenrich mal do konca júla 2023 už vyše 3 000 potvrdených predajov [48].
- **WormGPT**, postavený na open-source modeli GPT-J od EleutherAI, je trénovaný na dátach súvisiacich s malvérom a umožňuje útočníkom generovať presvedčivé phishingové e-maily bez akýchkoľvek etických bariér. Podľa SlashNext je WormGPT obzvlášť efektívny pri útokoch typu *Business Email Compromise*, kde napodobňuje štýl firemnej komunikácie [49].
- **PoisonGPT**, na rozdiel od predchádzajúcich dvoch, je skôr proof-of-concept model, ktorý demonštruje riziko šírenia dezinformácií, no môže byť použitý aj na generovanie zavádzajúcich e-mailov. Tieto škodlivé modely sú často inzerované ako „bez obmedzení“, čo znamená, že útočníci ich môžu používať na akékoľvek nelegálne aktivity bez obáv z blokovania [48].

3. Automatizácia celého procesu

Ďalšia technológia, ktorá sa používa, je automatizácia celého procesu útoku. Nástroje AI môžu nielen tvoriť e-maily ale aj skúmať odpovede obetí a robiť dynamické proti-odpovede, čím zvyšujú presvedčivosť útoku. Napríklad podľa organizácie CanIPhish, útočníci s pomocou AI môžu automatizovať celý cyklus phishingového útoku – od zbierania dát, cez tvorbu e-mailu, až po interakciu s obet'ou. Tieto technológie tak znižujú bariéru pre vstup útočníkov, lebo aj menej technicky šikovní jednotlivci môžu vykonávať sofistikované útoky. Výsledkom je že phishingové e-maily poháňané AI sú nielen presvedčivejšie, ale aj rýchlejšie a škálovateľnejšie, čo zvyšuje ich dopad na jednotlivcov a organizácie [46].

Útočníci tiež aktívne získavajú odborníkov na AI, aby vytvorili nástroje prispôsobené ich špecifickým potrebám. Napríklad podľa výskumu Recorded Future (2024) ruský aktér hrozieb s prezývkou Poena zverejnil na Telegramu inzerát, kde hľadá expertov na AI a strojové učenie na vývoj škodlivých LLM, ako je xGPT, ktorý má posúvať hranice možností v oblasti kybernetických útokov [50].

2.1.3 Príklad útoku

Aby sme v rámci tejto práce demonštrovali, ako jednoducho je možné vytvoriť presvedčivý phishingový e-mail pomocou AI technológií, ukážeme si fiktívny scenár s využitím simulátora od CanIPhish. V tomto príklade vytvoríme e-mail, ktorý sa tvári ako správa z Microsoft Teams od zamestnanca UTB univerzity a obsahuje odkaz na podvodné prihlásenie, kde by boli po prekliknutí a prihlásení ukradnuté prihlasovacie údaje obete.

Postup vytvorenia útoku pomocou CanIPhish simulátora:

1. **Výber šablóny a jazyka:** V CanIPhish simulátore vyberieme šablónu e-mailu, ktorá napodobňuje Microsoft Teams správu s názvom „Microsoft Teams – Nová správa“. Simulátor ponúka viac ako 80 rôznych šablón, ktoré pokrývajú širokú škálu scenárov – od overenia prihlásenia na platformách ako Facebook, Gmail či LinkedIn, cez platobné brány ako PayPal, až po oznámenia o sledovaní zásielky alebo e-maily o zdieľaní súborov. Tieto šablóny sú navrhnuté tak, aby pôsobili maximálne autenticky a napodobňovali skutočné notifikácie od známych služieb. Okrem toho si môžeme vybrať

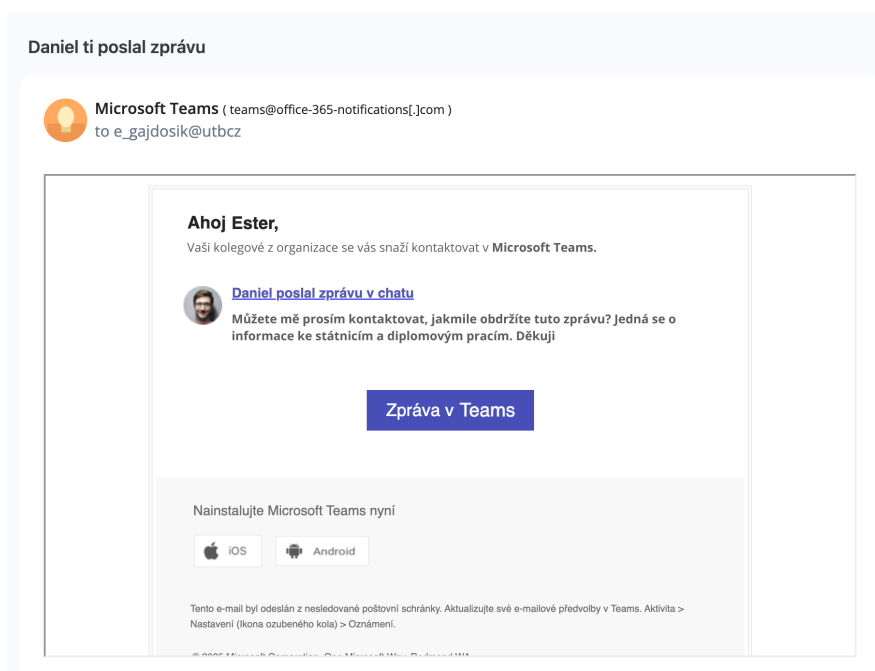
z viac ako 70 jazykov, vrátane češtiny a slovenčiny, čo umožňuje prispôbiť e-mail cieľovému publiku. V tomto príklade si zvolíme češtinu, aby e-mail pôsobil autenticky pre českého zamestnanca alebo študenta Univerzity Tomáša Bati (UTB) v Zlíne, čím zvýšime pravdepodobnosť, že obet' e-mailu uverí.

2. **Personalizácia e-mailu:** V simulátore prispôbíme detaily e-mailu tak, aby pôsobil čo najvierohodnejšie. Zadáme meno obete (napr. „Ester Gajdošík“), e-mailovú adresu (napr. „e.gajdosik123@utb.cz“), názov univerzity („Univerzita Tomáša Bati v Zlíne“) a pozíciu, ktorá bude v tomto prípade „študent“. Tieto údaje môžeme často získať z verejne dostupných zdrojov, ako sú LinkedIn, kontakty na webovej stránke fakulty univerzity, firemná webová stránka alebo facebookový profil. Na získanie takýchto informácií môžeme využiť aj AI chatové modely, ako je ChatGPT, ktoré dokážu do pár sekúnd poskytnúť relevantné údaje na základe jednoduchých otázok. Ak sú žiadané údaje verejne dostupné, AI model ich rýchlo zozbiera a poskytne nám ich, čím výrazne ušetrí čas pri príprave útoku. Na Obrázku 2.1 je znázornený príklad zadaného promptu spolu s odpoveďou, ktorá obsahuje verejne dostupné informácie vyhl'adané pomocou modelu ChatGPT-4o.



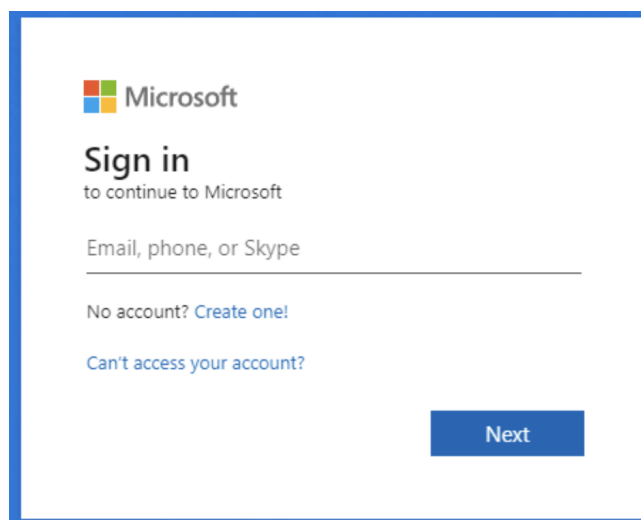
Obrázek 2.1 Ukážka vyhl'adávania verejne dostupných informácií pomocou modelu ChatGPT-4o

3. **Vytvorenie obsahu e-mailu s odkazom:** Získané informácie teraz môžeme použiť v e-maile, aby vyzeral ako legitímny zdroj. V CanIPhish nastavíme odosielať a tak, aby vyzeral ako oficiálna adresa Microsoft Teams, napríklad „notifications@teams-microsoft.com“, a predmet e-mailu zvolíme „Máte novou zprávu v Teams – Potvrďte přihlášení“. Simulátor automaticky vygeneruje obsah e-mailu v češtině, který vyzerá ako štandardná notifikácia z Microsoft Teams: „Ahoj Ester, máte novou zprávu v Microsoft Teams od vašeho týmu a odkazom: [Zpráva v Teams].“ Obrázok 2.2 zobrazuje vygenerovaný e-mail so správou z Teamsu, ktorý pôsobí autenticky a je prispôsobený pre našu obeť.



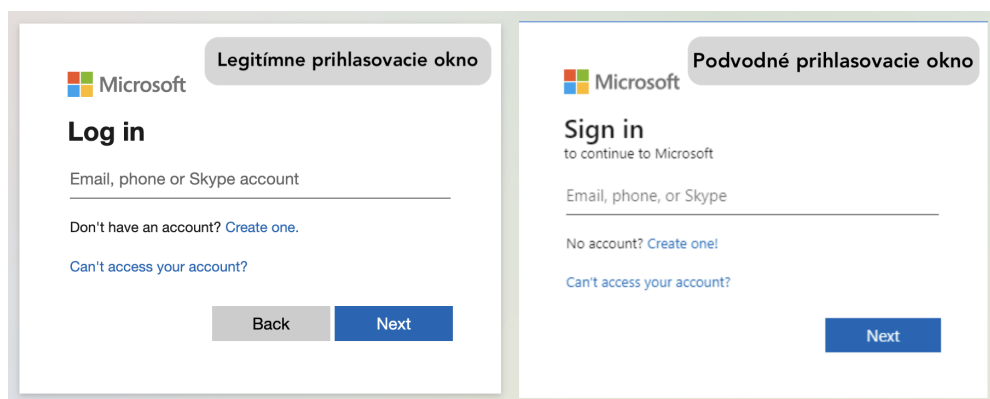
Obrázek 2.2 Ukážka vygenerovaného e-mailu
v programe CanIPhish

4. **Simulácia prihlásenia a krádež údajov:** Odkaz „Zpráva v Teams“ je v skutočnosti phishingový odkaz, ktorý vedie na falošnú prihlasovaciu stránku. Po kliknutí na odkaz sa obeť dostane na podvodnú stránku, ktorá je navrhnutá tak, aby vyzerala identicky ako oficiálna prihlasovacia stránka Microsoftu. Keď obeť zadá svoje prihlasovacie údaje, tak sú v reálnom útoku okamžite odoslané útočníkovi, ktorý ich môže ďalej zneužiť na prístup k univerzitným systémom, e-mailu, či iným citlivým údajom. Na Obrázku 2.3 je znázornené vygenerované prihlasovacie okno, ktoré sa zobrazí po kliknutí na odkaz, a ukazuje, ako presvedčivo dokáže falošná stránka napodobniť originál.



Obrázek 2.3 Zobrazená podvodná stránka imitujúca rozhranie Microsoft prihlásenia

Pre lepšie porovnanie sú na Obrázku 2.4 zobrazené vedľa seba legitímne (vľavo) a podvodné (vpravo) prihlasovacie okná Microsoftu. Tieto rozdiely sú na prvý pohľad minimálne, aj keď vizuálne prvky sú mierne odlišné – vrátane rozloženia textu, formátovania nadpisov či jemne odlišného písma. Bez priameho porovnania vedľa seba sú tieto rozdiely takmer nepostrehnuteľné, čo zvyšuje riziko, že si používateľ podvodnú stránku nevšimne a zadá svoje prihlasovacie údaje.



Obrázek 2.4 Porovnanie legitímneho (vľavo) a podvodného (vpravo) prihlasovacieho okna Microsoft

2.2 Deepfake videá vo phishingu

2.2.1 Ako sa deepfake používa vo phishingu s využitím AI

Deepfake phishing predstavuje jednu z najsofistikovanejších foriem kybernetických útokov, ktoré využívajú umelú inteligenciu na vytvorenie presvedčivého falošného zvukového alebo obrazového obsahu. Útočníci používajú deepfake technológie na vytvorenie falošných videí, ktoré napodobňujú reálne osoby. Často ide o dôveryhodné authority, ako sú šéfovia firiem, kolegovia, verejné osobnosti alebo dokonca rodinní príslušníci. Tieto videá sú navrhnuté tak, aby vyzerali a zneli autenticky, čím zvyšujú pravdepodobnosť, že obeť uverí podvodnej správe a vykoná požadovanú akciu [51].

AI zohráva kľúčovú úlohu pri tvorbe deepfake videí, pretože umožňuje automatizovať a zefektívniť proces manipulácie. Pomocou techník hlbokého učenia (deep learning) dokáže analyzovať veľké množstvo dát – napríklad videá, fotografie a zvukové nahrávky cieľovej osoby a vytvoriť presvedčivú repliku jej tváre, hlasu a správania [52]. Vo phishingových útokoch sa deepfake videá často používajú ako súčasť sociálneho inžinierstva, kde útočníci zneužívajú dôveru obeť. Napríklad útočník môže vytvoriť video, v ktorom šéf firmy žiada zamestnanca o okamžitý prevod peňazí na „dôveryhodný“ účet, alebo video, kde známa osobnosť láka obeť na falošnú investičnú príležitosť [51].

Deepfake videá sa vo phishingu často kombinujú s inými formami útoku, ako sú e-maily alebo správy na sociálnych sieťach. Útočníci môžu poslať e-mail s odkazom na video, ktoré sa tvári ako legitímna komunikácia. Ďalším bežným scenárom je vytvorenie deepfake videa, ktoré sa šíri cez sociálne siete a láka obeť na falošné investičné schémy, ako napríklad krypto podvody, kde útočníci napodobňujú známe osobnosti, ako je Elon Musk, a sľubujú vysoké zisky. Tieto útoky sú obzvlášť nebezpečné, pretože ľudská psychológia zohráva veľkú úlohu. Keď totiž obeť vidí známu tvár a počuje známy hlas, je menej pravdepodobné, že bude podozrievať podvod [53].

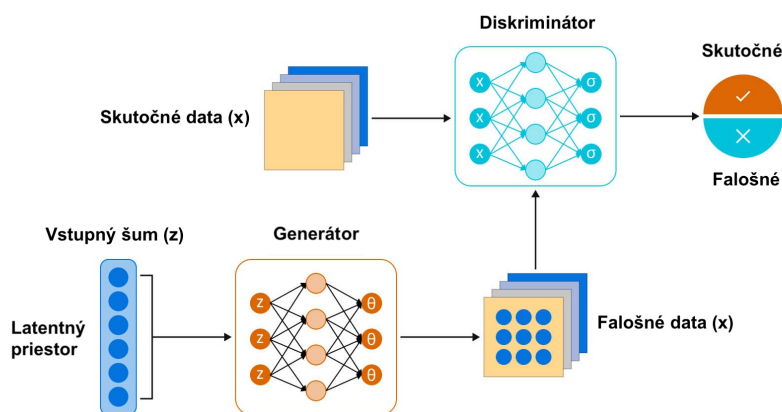
2.2.2 Používané technológie

Na vytváranie deepfake videí sa využívajú pokročilé technológie umelej inteligencie, najmä tie, ktoré sú založené na hlbokom učení. Tieto technológie dokážu vytvárať veľmi realistické napodobeniny tváří, hlasov a pohybov ľudí, čím zvyšujú presvedčivosť podvodných videí. Medzi najbežnejšie používané technológie patria generatívne adversariálne siete (GANs),

autoenkodéry, konvolučné neurónové siete (CNN) a nástroje na syntézu hlasu. Okrem toho útočníci často využívajú dostupné softvérové nástroje a platformy, ktoré uľahčujú proces tvorby deepfake obsahu.

Generatívne adversariálne siete (GANs)

Základom pre vytváranie deepfake videí sú generatívne adversariálne siete, ktoré sa skladajú z dvoch neurónových sietí – generátora a diskriminátora. Generátor má za úlohu vytvoriť falošný obsah, ako napríklad tvár osoby, zatiaľ čo diskriminátor hodnotí, či je tento obsah pravý alebo falošný. Tieto siete sa navzájom trénujú, čo vedie k neustálemu zlepšovaniu ich výkonu [34]. Pri vytváraní deepfake videí generatívne adversariálne siete analyzujú videozáznamy z rôznych uhlov, pričom sledujú správanie, pohyb a rečové vzory. Získané údaje prechádzajú viackrát cez diskriminátor, aby sa zvýšila realistickosť výsledného obrazu alebo videa. GAN siete sa využívajú aj pri phishingových útokoch a pri tzv. výmene tváre (face swapping), kde je pôvodná tvár osoby vo videu nahradená tvárou cieľovej osoby [54]. Na Obrázku 2.5 je znázornená obecná architektúra Generatívnej adversariálnej siete.



Obrázek 2.5 Architektúra Generatívnej adversariálnej siete

(zdroj: SoluLab, 2023 [55])

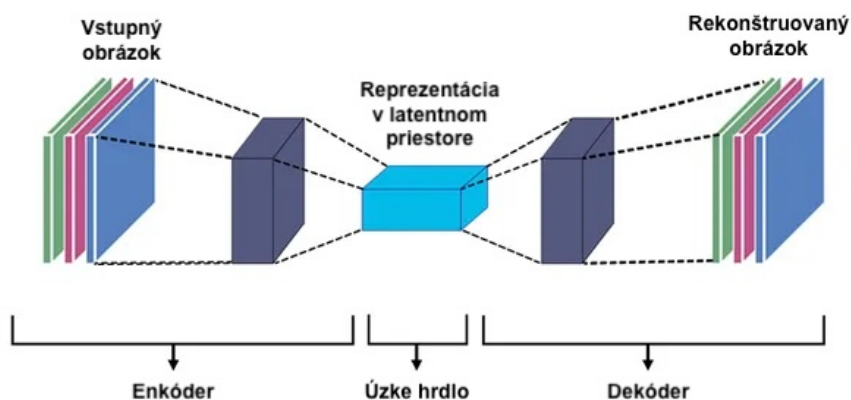
Konvolučné neurónové siete (CNN)

Konvolučné neurónové siete (CNN) sa používajú na spracovanie vizuálnych dát, ako sú obrázky a videá, a hrajú zásadnú úlohu pri vytváraní deepfake obsahu. Tieto siete dokážu zachytiť dôležité rysy tváre, ako sú oči, nos, ústa a rôzne výrazy, a využiť ich na vytvorenie veľmi realistických napodobení. V kombinácii s GANs alebo autoenkodérmi umožňujú

CNN útočníkom vytvárať videá, kde tvár cieľovej osoby vyzerá prirodzene aj pri zmene uhla pohľadu alebo osvetlenia. Podľa výskumu od DeepMind dokážu moderné CNN modely znížiť chybovosť pri mapovaní tvárových črt na menej ako 2 %, čo výrazne zvyšuje kvalitu deepfake videí, ktoré sa používajú vo phishingu [56].

Autoenkodéry

Autoenkóder je špeciálny typ neurónovej siete, ktorej úlohou je prispôbiť vstup, ktorý dostala. Táto sieť sa skladá z dvoch hlavných častí – enkodéra a dekodéra. Enkodér komprimuje vstupné dáta (napríklad video alebo obrázky tváre) do nižšej dimenzie, zatiaľ čo dekodér ich opäť rekonštruuje do podoby, ktorá verne napodobňuje cieľovú tvár. Autoenkodéry sa trénujú na rozsiahlych datasetoch tvárí, aby dokázali prenášať výrazy a pohyby jednej osoby na inú. Tieto modely sú obzvlášť účinné pri vytváraní videí s nízkym rozlíšením [57]. Na Obrázku 2.6 je znázornená obecná architektúra autoenkódera.



Obrázek 2.6 Architektúra Autoenkódera

(zdroj: Medium, 2019 [58])

Technológie na syntézu hlasu

Okrem vizuálnej manipulácie sa deepfake videá kombinujú aj so syntetizovaným hlasom, aby výsledný obsah pôsobil ešte presvedčivejšie. Využívajú sa pri tom technológie typu text-to-speech, ktoré dokážu na základe textového vstupu vytvoriť realisticky znejúci hlas [59]. Tieto modely sa trénujú na zvukových nahrávkach konkrétnej osoby, pričom útočníci často čerpajú z verejne dostupných zdrojov – napríklad z YouTube videí, podcastov či príspevkov na sociálnych sieťach. Stačí im pritom iba 5 až 10 minút záznamu, aby nástroje ako

Respeecher dokázali napodobniť hlas tak verne, že je takmer nerozoznatel'ný od originálu. Pri phishingových útokoch sa potom tento syntetizovaný hlas zosynchronizuje s deepfake videom, čím vzniká presvedčivá ilúzia, že daná osoba skutočne hovorí to, čo vidíme na obrazovke [37].

2.2.3 Príklad útoku

V tejto časti sa pozrieme na skutočný prípad útoku, ktorý sa odohral v roku 2024 v Hongkongu. Útočníci využili deepfake technológiu, aby oklamali zamestnanca veľkej nadnárodnej firmy, čo viedlo k strate 25 miliónov dolárov. Na základe dostupných informácií sa pokúsime zrekonštruovať pravdepodobné kroky, ktoré útočníci podnikli pri realizácii útoku.

1. **Zber údajov o cieľovej firme a zamestnancoch:** Útočníci pravdepodobne začali s dlhodobým zberom údajov o spoločnosti Arup a jej zamestnancoch, čo je typické pre APT (Advanced Persistent Threat) hrozby. Mohli využiť verejne dostupné zdroje, ako sú videá z firemných konferencií, rozhovory alebo iné nahrávky, aby získali vizuálne a zvukové dáta o finančnom riaditeľovi a ďalších zamestnancoch. Podľa hongkonskej polície útočníci použili verejne dostupné záznamy, čo naznačuje, že mohli cielene sledovať online prítomnosť firmy [60].
2. **Vytvorenie deepfake videa a distribúcia phishingového e-mailu:** Na tento účel pravdepodobne použili technológie ako generatívne adversariálne siete (GANs) a autoenkodéry – nástroje schopné vymeniť tváre vo videu a zároveň presne zosynchronizovať pohyby pier s falošným hlasom (viac v podkapitole 2.2.2). Podľa vyjadrenia hongkonskej polície bolo výsledné video natoľko realistické, že zamestnanec nedokázal rozpoznať podvod. Deepfake totiž napodobňoval nielen vzhľad, ale aj hlasy kolegov. Následne bol zamestnancovi finančného oddelenia doručený phishingový e-mail, ktorý sa tváril ako správa od finančného riaditeľa sídliaceho v Londýne. E-mail obsahoval žiadosť o vykonanie „dôvernej transakcie“, čo u zamestnanca spočiatku vyvolalo pochybnosti – podobné požiadavky totiž nie sú v bežnej internej komunikácii zvyčajné. Aby útočníci rozptýlili akékoľvek podozrenie, pozvali ho na videohovor, ktorý mal pôsobiť ako legítimná firemná komunikácia. Správa mohla byť odoslaná z podvrhutej e-mailovej adresy imitujúcej oficiálnu doménu Arup, napríklad „ceo@arup-uk.com“, čo je bežná technika v rámci phishingových kampaní [60].

3. **Vykonanie deepfake videohovoru:** Počas videohovoru útočníci využili vopred pripravené deepfake video, ktoré verne napodobňovalo zamestnancov, ktorých obeť dobre poznala. Podľa polície bol jediným skutočným účastníkom hovoru zamestnanec, všetci ostatní boli len deepfake napodobeniny [60]. Útočníci sa snažili minimalizovať interakciu, aby zamestnanec nemal šancu odhaliť podvod a tak ho požiadali len o krátke predstavenie a potom mu dali pokyny na prevod peňazí. Potom hovor rýchlo ukončili aby bol zámerne krátky [61]. Video bolo tak presvedčivé, že zamestnanec veril, že hovorí so svojimi skutočnými kolegami. Deepfake technológia ho oklamala natoľko, že jeho počiatočné podozrenia z e-mailu sa rozptýlili.
4. **Krádež finančných prostriedkov:** Na základe pokynov z videohovoru zamestnanec uskutočnil 15 prevodov v celkovej hodnote 200 miliónov hongknských dolárov (približne 25 miliónov USD) na päť rôznych bankových účtov v Hongkongu [60]. Podvod sa odhalil až o týždeň neskôr, keď zamestnanec kontaktoval centrálu spoločnosti a zistil, že žiadna z týchto transakcií nebola schválená. Podľa Arup však tento incident nemal vplyv na ich finančnú stabilitu ani interné systémy, čo naznačuje, že útok bol zameraný výlučne na sociálne inžinierstvo a nezasahoval do IT systémov firmy [62].

2.3 Hlasový phishing (vishing) s AI

Hlasový phishing, známy ako vishing (voice phishing), je forma kybernetického útoku, ktorá využíva telefonické hovory na oklamanie obetí a získanie citlivých informácií. S príchodom AI sa vishing stal oveľa sofistikovanejší a presvedčivejší, pretože je možné rýchlym spôsobom napodobňovať hlasy a vytvárať realistické a personalizované konverzácie [63].

2.3.1 Ako sa vishing používa vo phishingu s využitím AI

V súčasnosti, vďaka rýchlemu pokroku v technológii klonovania hlasu založenej na AI, môžu útočníci vytvoriť presvedčivý hlas už z krátkeho zvukového klipu. AI technológie, ako sú pokročilé modely text-to-speech (TTS), umožňujú klonovať hlas konkrétnej osoby s minimálnym množstvom dát [64]. Napríklad, podľa štúdie od Microsoftu, ich model VALL-E dokáže napodobniť hlas hovoriaceho už po 3 sekundách zvukového materiálu, čo je významný pokrok v oblasti syntézy hlasu [65]. Bežnejšie nástroje, ako je Respeecher, potrebujú 5–10 minút zvukového materiálu, čo môže byť napríklad jedno video obete na YouTube.

Súčasná AI nielenže dokáže napodobniť hlas, ale aj viesť dynamické konverzácie v reálnom čase. Rýchlo fungujúce algoritmy, ako sú tie v modeloch spracovania prirodzeného jazyka (NLP), sú schopné analyzovať odpovede obete a prispôbiť konverzáciu, čím vytvárajú dojem živého rozhovoru. Ďalším dôležitým aspektom je, že tieto útoky sú lacné a ľahko škálovateľné, čo ich robí obzvlášť nebezpečnými. Útočníci môžu automatizovať tisíce hovorov naraz pomocou robocall technológií, ktoré využívajú AI na prispôsobenie správ konkrétnym obetiam [65]. Kým tradičné vishingové útoky sa spoliehali na automatické nahrávky alebo robotické hlasy, budúce útoky budú čoraz častejšie využívať AI na živé konverzácie, ktoré pôsobia prirodzene a minimalizujú podozrenie zo strany obete [64].

2.3.2 Používané technológie

Na realizáciu vishingových útokov s AI útočníci využívajú pokročilé technológie hlbokého učenia a spracovania prirodzeného jazyka (NLP), ktoré umožňujú generovať a manipulovať hlasmi v reálnom čase. Medzi kľúčové technológie patria text-to-speech (TTS) modely, hlasové klonovanie, automatické rozpoznávanie reči (ASR) a robocall platformy, ktoré integrujú AI na automatizáciu a personalizáciu hovorov.

Text-to-speech (TTS) modely

AI je dnes možné využiť na generovanie hlasu pomocou technológie známej ako syntéza textu na reč (TTS). Tieto systémy premieňajú písaný text na hovorené slová a dokážu vytvárať hlasy, ktoré znejú prirodzene a realisticky. Hoci základná verzia tejto technológie existuje už dlhé roky, dnešný pokrok v oblasti strojového učenia, najmä vďaka hlbokým neurónovým sieťam, priniesol výrazné zlepšenia. Moderné modely dokážu verne napodobniť ľudskú reč vrátane intonácie, výšky hlasu či rytmu, čím sa výsledná syntéza stáva oveľa presvedčivejšou. Jednou z prelomových technológií v oblasti TTS je WaveNet, ktorú vyvinula spoločnosť DeepMind (patriaca pod Google). Tento model generuje zvuk predpovedaním jednotlivých vzoriek surového zvukového signálu, čo vedie k mimoriadne prirodzene znejúcej reči [66].

Hlasové klonovanie

Moderné technológie hlasového klonovania, ako je napríklad Respeecher či ElevenLabs, umožňujú napodobniť hlas konkrétnej osoby už na základe krátkej zvukovej nahrávky. Tieto nástroje využívajú pokročilé algoritmy umelej inteligencie na zachytenie jemných nuáns

ľudskej reči, ako sú tón, výška hlasu, rytmus či intonácia, vďaka čomu sú klonované hlasy takmer identické s originálom. Ďalšou z pozoruhodných schopností týchto systémov je aj viacjazyčné klonovanie hlasu. Napríklad technológia Respeecher dokáže reprodukovat hlas osoby v inom jazyku, pričom si zachová jeho jedinečné vokálne znaky. AI sa najprv „naučí“ hlas pôvodného hovoriaceho a následne dokáže generovať reč v inom jazyku s rovnakým tónom, farbou a štýlom vyjadrovania. Výsledkom je realistický hlas, ktorý pôsobí konzistentne a prirodzene bez ohľadu na jazyk [37].

Automatické rozpoznávanie reči (ASR) a NLP

Automatické rozpoznávanie reči (ASR) kombinované s technológiami spracovania prirodzeného jazyka (NLP) umožňuje AI rozumieť odpovediam obete a generovať relevantné reakcie v reálnom čase. Napríklad model ako Google Speech-to-Text dokáže analyzovať, čo obet hovorí, a prispôbiť konverzáciu, čím vytvára ilúziu živého rozhovoru [67]. Podľa Trend Micro je táto technológia kľúčová pre dynamické vishing útoky, kde útočník nemusí manuálne ovládať hovor [68].

Robocall platformy s AI

Útočníci využívajú robocall platformy, ako je Twilio alebo Plivo, ktoré integrujú AI na automatizáciu masových hovorov. Tieto platformy umožňujú prispôbiť hovory konkrétnym obetiam na základe ich dát (napr. meno, adresa), čo zvyšuje presvedčivosť útokov. Podľa Darktrace dokážu takéto platformy uskutočniť až 10 000 hovorov za hodinu, čím masívne škálujú vishing kampane [69].

2.3.3 Príklad útoku

V tejto časti popíšeme prípadovú štúdiu od Mandiant Red Team, ktorá sa uskutočnila koncom roka 2023. Mandiant vykonal kontrolované cvičenie s klientom, kde použil AI hlasový spoofing na získanie počiatočného prístupu do internej siete organizácie. Tento príklad demonštruje, ako môžu útočníci zneužiť technológie klonovania hlasu na oklamanie zamestnancov. Nasledujúci postup je zrekonštruovaný na základe opisu cvičenia a rozdelený do krokov s vysvetleniami [67].

1. **Získanie súhlasu a vytvorenie zámienky sociálneho inžinierstva:** Cvičenie začalo tým, že sa získal súhlas od klienta, čo je bežný krok pri etickom hackingu. Týmto sa zabezpečuje, že útok prebieha legálne a pod kontrolou. Mandiant Red tím ako prvé vytvoril presvedčivú zámienku na sociálne inžinierstvo. Rozhodli sa predstierať, že sú členom bezpečnostného tímu, konkrétne osobou, ktorej hlas by zamestnanci mohli spoznať. Tento krok ukazuje, ako útočníci v reálnom svete cieľia na dôveryhodné osoby, aby zvýšili šancu na úspech útoku.
2. **Získanie hlasového materiálu a tréning AI modelu:** Mandiant získal potrebné zvukové údaje – prirodzenú vzorku hlasu člena bezpečnostného tímu – od klienta, ktorý s tým súhlasil. Potom tím využil AI technológie na tréning modelu pre hlasový spoofing, pričom dosiahol dostatočnú úroveň realizmu, aby hlas znel presvedčivo.
3. **Zber údajov o zamestnancoch pomocou OSINT:** Na identifikáciu potenciálnych obetí Mandiant využil open-source inteligenciu (OSINT) a zhromaždil údaje o zamestnancoch klienta, ako sú názvy pracovných pozícií, miesta pôsobenia a telefónne čísla. Tím sa zameral na zamestnancov, ktorí by s najväčšou pravdepodobnosťou rozpoznali hlas člena bezpečnostného tímu a zároveň mali potrebné povolenia na splnenie cieľov cvičenia. OSINT je bežnou taktikou útočníkov, ktorí využívajú verejne dostupné informácie na sociálnych sieťach, ako sú LinkedIn alebo firemné webstránky, na mapovanie organizácie a výber vhodných cieľov.
4. **Iniciovanie falošného hovoru:** S upraveným zoznamom cieľov Mandiant spustil falošné telefonické hovory pomocou VoIP služieb a techník na falšovanie čísel, aby hovor vyzeral, akoby prichádzal z dôveryhodného zdroja. Tímu sa podarilo spojiť s prvou obeťou, bezpečnostným správcou, ktorý na hovor reagoval dôverčivo: „*Hej, šéfe, čo sa deje?*“. Tento moment jasne ukazuje, akú dôležitú úlohu zohráva dôvera v hlase pri sociálnom inžinierstve – obeť automaticky predpokladala, že hovorí s niekým, koho pozná, čo je kľúčový faktor úspechu vishingových útokov.
5. **Oklamanie obete a stiahnutie škodlivého softvéru:** Mandiant sa rozhodol použiť starostlivo vybranú zámienku – „nesprávnu konfiguráciu klienta VPN“ a využil nedávny globálny výpadok, ktorý zasiahol poskytovateľa VPN. Tento scenár vytvoril pocit naliehavosti, pretože obeť cítila tlak na rýchle vyriešenie problému. Vďaka dô-

vere v hlas, ktorý poznala, bezpečnostný správca prehliadol varovania od Microsoft Edge a Windows Defender SmartScreen a nevedomky si stiahol a spustil vopred pripravenú škodlivú záťaž na svojej pracovnej stanici. Úspešné spustenie tejto záťaže umožnilo tímu získať počiatočný prístup do internej siete. Tento krok ukazuje, ako môže spoofing hlasu obísť bezpečnostné mechanizmy, ktoré by inak mohli útok odhaliť, a zdôrazňuje zraniteľnosť ľudského faktora – zamestnanec dôveroval hlasu a ignoroval varovania.

3 PRÁVNE A ETICKÉ ASPEKTY VYUŽITIA AI VO PHISHINGOVÝCH ÚTOKOCH

Táto kapitola sa venuje právnym a etickým otázkam spojeným s využívaním umelej inteligencie v oblasti kybernetickej bezpečnosti, so zvláštnym dôrazom na jej zneužitie vo phishingových útokoch. Najprv uvedieme súčasné výzvy vrámci právnej regulácie AI. Následne si predstavíme súčasné právne rámce, ktoré upravujú využívanie AI, a to ako na medzinárodnej úrovni, tak aj na národnej úrovni. Kapitola rozoberá dva hlavné prístupy k právnej regulácii umelej inteligencie: záväzné právne normy (tzv. *hard law*) a právne nezáväzné normy (*soft law*). Predstavíme výhody aj nevýhody oboch prístupov a vysvetlíme, prečo sa v posledných rokoch v oblasti AI regulácie uprednostňovali najmä formy *soft law*, zdôrazňujúce etické princípy a dobrovoľné kódexy správania.

V druhej časti kapitoly sa sústredíme na etické aspekty spojené s využívaním umelej inteligencie vo phishingu. Preskúmame, aké nové etické výzvy prináša nasadenie technológií založených na AI v oblasti podvodných útokov, a aké zásady by mali vývojári, prevádzkovatelia a regulátori dodržiavať na minimalizáciu týchto rizík.

3.1 Výzvy v právnej regulácii

Právna regulácia umelej inteligencie predstavuje komplexnú výzvu, vyplývajúcu z rýchleho technologického vývoja a rozdielných prístupov na globálnej, regionálnej aj národnej úrovni. Rôzne regulačné stratégie, od prísnych až po liberálnejšie prístupy, zároveň spôsobujú fragmentáciu právneho prostredia, čo sťažuje harmonizáciu pravidiel a vytváranie jasných, konzistentných a účinných právnych rámcov v medzinárodnom prostredí [70]. Medzi najvýznamnejšie výzvy, ktorým právna regulácia umelej inteligencie v súčasnosti čelí, patria:

1. **Definícia AI:** Absencia jednotnej a všeobecne akceptovanej definície umelej inteligencie komplikuje snahy o jej právnu reguláciu, ktorá by pokrývala celé spektrum technológií a aplikácií AI [71]. Nejednoznačnosť definície môže viesť k vzniku regulačných medzier a interpretačných nejasností. Definícia by nemala byť príliš „úzka“, aby dokázala pokryť aj budúci vývoj, no ani príliš „široká“, aby nevyvolávala problémy pri jej implementácii. Ideálna právna definícia by mala byť technologicky neutrálna, nadčasová a flexibilná voči dynamickým zmenám na trhu a v technológiách [70].

2. **Miera akou má legislatíva zohľadňovať etické limity:** Určenie normatívnych hraníc, resp. limitov toho, čo je etické pri vývoji a používaní umelej inteligencie je subjektívne a líši sa naprieč kultúrami a spoločnosťami. Pre zákonodarcov okrem iného aj z tohoto dôvodu predstavuje významnú výzvu úloha nájsť rovnováhu medzi podporou inovácií, zachovaním etických štandardov a odhadom rozumenj miery regulácie. Európska komisia zdôrazňuje potrebu zakotviť dôveryhodnosť ako základný prvok umelej inteligencie. V tomto smere boli vypracované odporúčania expertnej skupiny pre umelú inteligenciu pri Európskej komisii zriadené v roku 2018, pričom členské štáty EÚ ich postupne integrujú do svojich národných stratégií [70].
3. **Rýchly technologický pokrok:** Tempo vývoja umelej inteligencie je natoľko dynamické, že legislatíva často nedokáže pružne reagovať. Právne predpisy tak môžu rýchlo zastarať s príchodom nových technológií. Efektívna regulácia by mala maximalizovať spoločenské prínosy AI a zároveň minimalizovať jej riziká, pričom je potrebné brať do úvahy, že budúce výzvy a prínosy nemožno vždy presne predvídať. Tieto neznáme premenné robia tradičné regulačné prístupy neefektívnymi [72].
4. **Zodpovednosť za škodu:** Transparentnosť fungovania systémov AI úzko súvisí s otázkami zodpovednosti za prípadné škody. Určenie, kto nesie zodpovednosť v prípade, že AI spôsobí škodu alebo ujmu, je vzhľadom na autonómny charakter systémov komplexné. Otvára sa diskusia o príčinnej súvislosti, zásahu vyššej moci (*vis maior*) a o možnosti nájdenia subjektu zodpovednosti (vývojári, používatelia alebo samotná AI). Riešenie otázky zodpovednosti je nevyhnutné zahrnúť do právnej úpravy [70, 73].
5. **Súkromie a ochrana údajov:** Systémy AI spracúvajú rozsiahle množstvá osobných údajov, čo vyvoláva závažné obavy v oblasti ich ochrany. Je nevyhnutné, aby ochrana údajov bola zabezpečená počas celého životného cyklu systému AI, a to nielen pri spracúvaní údajov získaných od používateľov, ale aj údajov, ktoré systém vytvára alebo vyvodzuje. Právna regulácia musí upravovať spôsob zberu, spracúvania a uchovávania údajov (vrátane otázky ich výmazu) v súlade s právami jednotlivcov. V Európskej únii zohráva kľúčovú úlohu najmä nariadenie GDPR (Nariadenie (EÚ) 2016/679) [70].

Riešenie týchto výziev si vyžaduje komplexný prístup založený na spolupráci medzi vládami, akademickou obcou, súkromným sektorom a občianskou spoločnosťou. Prostredníctvom otvoreného dialógu, zdieľania odborných poznatkov a tvorby dôkazmi podložených politík je možné vybudovať regulačný rámec, ktorý bude podporovať zodpovedný rozvoj a využívanie umelej inteligencie a zároveň chrániť základné práva jednotlivcov a spoločenské hodnoty [70].

3.2 Medzinárodné právne normy

Právna regulácia umelej inteligencie môže vychádzať z dvoch hlavných prístupov a to *hard law* (právne záväzné normy) a *soft law* (právne nezáväzné normy). Prístup *hard law* predstavuje záväzné právne normy, ktoré majú právnu silu a v prípade ich porušenia môžu byť aplikované sankcie alebo iné donucovacie opatrenia [70]. Jedným z najvýznamnejších *hard law* právnych rámcov je v súčasnosti európsky Zákon o umelej inteligencii (EU AI Act). Ide o prvý komplexný zákon o umelej inteligencii, ktorý nadobudol platnosť 1. augusta 2024. Tento zákon je záväzný pre všetky členské štáty EÚ a jeho cieľom je vytvoriť jednotný a bezpečný trh pre technológie AI [74].

Naopak, prístup *soft law* zahŕňa právne nezáväzné nástroje, ako sú usmernenia, odporúčania alebo etické kódexy. Tieto nástroje nemajú právnu silu, ale slúžia ako dôležité orientačné body pri nastavovaní etických a bezpečnostných štandardov v oblasti umelej inteligencie [70]. Príkladom *soft law* sú práve OECD Principles on AI, ktoré poskytujú rámec pre etické a zodpovedné využívanie AI na medzinárodnej úrovni. Ich implementácia je dobrovoľná a závisí od politickej vôle jednotlivých štátov a organizácií [5].

3.2.1 Európsky zákon o umelej inteligencii (AI Act)

Dňa 13. marca 2024 Európsky parlament odhlasoval schválenie prvého komplexného právneho rámca pre reguláciu umelej inteligencie: Zákon o umelej inteligencii. Rastúce používanie nástrojov umelej inteligencie a široké nasadenie generatívnych AI, ako je ChatGPT, Gemini, Claude a podobne, vyvolali obavy z ich etických, právnych a spoločenských dôsledkov, vrátane ich zneužitia vo phishingových útokoch. V tejto súvislosti je cieľom zákona regulovať dizajn a používanie AI tak, aby posilnil dôveru v umelú inteligenciu, kontroloval jej vplyv na spoločnosť, podniky a jednotlivcov – najmä na základné práva – a zároveň vytvoril priaznivé podmienky pre výskum, vývoj, hospodárstvo a inovácie [75]. Zákon nadobudol

účinnosť 1. augusta 2024 a jeho implementácia prebieha postupne prostredníctvom prechodných období pre jednotlivé typy AI systémov. Pravidlá týkajúce sa zakázaných systémov umelej inteligencie začínajú platiť po šiestich mesiacoch, zatiaľ čo opatrenia vzťahujúce sa na generatívnu umelú inteligenciu nadobudnú účinnosť po dvanástich mesiacoch [76].

Definícia systému umelej inteligencie

AI Act definuje systém umelej inteligencie (AIS) s cieľom odlíšiť ho od jednoduchších softvérových systémov. Podľa článku 3 ods. 1 zákona je systém umelej inteligencie definovaný ako „strojový systém, ktorý je dizajnovaný na prevádzku s rôznymi úrovňami autonómnosti, ktorý môže po nasadení prejavovať adaptabilitu a ktorý pre explicitné alebo implicitné ciele odvodzuje zo vstupov, ktoré dostáva, spôsob generovania výstupov, ako sú predpovede, obsah, odporúčania alebo rozhodnutia, ktoré môžu ovplyvniť fyzické alebo virtuálne prostredie“ [4]. Tento prístup založený na hodnotení rizika predstavuje jeden zo základných pilierov tohto nariadenia, ktorý zavádza postupne prísnejšie obmedzenia na používanie umelej inteligencie v závislosti od predpokladanej úrovne rizika, ktoré jednotlivé systémy predstavujú [77].

Definícia modelu umelej inteligencie na všeobecné použitie (GPAI)

Modely umelej inteligencie na všeobecné použitie, označované skratkou GPAI (anglicky General-Purpose AI), predstavujú typ AI modelov, ktoré sa vyznačujú vysokou mierou univerzálnosti. Sú trénované na veľkých množstvách dát pomocou metód samodohľadu a dokážu samostatne zvládať širokú škálu úloh bez ohľadu na konkrétny kontext ich nasadenia. Na rozdiel od úzko špecializovaných modelov nie sú viazané na jeden konkrétny účel a môžu byť použité priamo alebo integrované do iných systémov. V prípade, že GPAI model je voľne dostupný pod otvorenou licenciou, pričom jeho parametre sú verejne prístupné a umožňujú ďalšie využitie či modifikáciu, vzťahujú sa naň len dve základné povinnosti – zverejnenie tréningových údajov a dodržiavanie autorských práv. To však platí len v prípade, že takýto otvorený model nie je nasadený ako kompletný systém AI [4, 77, 74].

Zvláštnu kategóriu predstavujú modely GPAI, ktoré môžu predstavovať tzv. systémové riziko. Takéto riziko sa spája s výpočtovou náročnosťou ich trénovania, kde hranicou je súhrnný výkon presahujúci 10^{25} FLOP (operácií s pohyblivou rádovou čiarkou). Ak model túto hranicu prekročí, poskytovateľ je povinný do dvoch týždňov informovať Európsku

komisiu. Zároveň môže predložiť odôvodnenie, prečo sa jeho model aj napriek prekročeniu hranice nemá považovať za systémovo rizikový. Konečné slovo však má Komisia, ktorá môže po konzultácii s nezávislými odborníkmi rozhodnúť, že model patrí do kategórie systémových GPAI, a tým naň uplatniť prísnejšie regulačné opatrenia [74].

Kategórie systémov AI podľa nariadenia

Európsky zákon o umelej inteligencii kategorizuje systémy AI do štyroch typov podľa ich potenciálnej úrovne nebezpečenstva, pričom každá kategória má svoje vlastné povinnosti v súlade s nariadením:

- **Systémy AI s minimálnym alebo žiadnym rizikom:** Do tejto kategórie patria systémy, ktoré nepredstavujú významné riziko pre jednotlivcov, ako sú napríklad videohry s podporou AI, filtre nevyžiadanej pošty či systémy na prediktívnu údržbu. Používanie týchto systémov je povolené bez ďalších obmedzení alebo povinností. Spoločnosti sa napriek tomu môžu zaviazat' k dobrovoľným kódexom správania [76].
- **Systémy AI s obmedzeným rizikom:** Do tejto kategórie spadajú systémy, ktoré interagujú s používateľmi ako sú chatboty, alebo nástroje na generovanie a úpravu obsahu. Ide napríklad o technológie typu deepfake, ktoré umožňujú vytvárať falošné videá či syntetické hlasy. Na tieto systémy sa vzťahujú požiadavky na transparentnosť, teda používatelia musia byť informovaní, že výstup bol vytvorený umelou inteligenciou. Osoby využívajúce systémy AI na generovanie alebo úpravu deepfake obsahu sú vo väčšine prípadov povinné jasne uviesť, že ide o umelo vytvorený alebo upravený materiál. Táto povinnosť je obzvlášť dôležitá pri prevencii phishingových a podvodných útokov, kde môžu deepfake videá alebo vishing ľahko zmiatať obeť [76, 77].
- **Vysoko rizikové systémy:** Do tejto kategórie patria systémy, ktoré môžu mať významný dopad na bezpečnosť, základné práva jednotlivcov alebo spoločnosť ako celok. Systémy AI sa považujú za vysoko rizikové, ak slúžia na profilovanie osôb – teda na automatizované spracúvanie osobných údajov s cieľom hodnotiť aspekty, ako sú pracovný výkon, ekonomická situácia, zdravotný stav, preferencie, záujmy, spoločnosť, správanie, miesto pobytu či pohyb jednotlivca [4]. Na vysoko rizikové systémy EU AI Act ukladá prísne požiadavky, vrátane posúdenia vplyvu na základné práva

(FRIA), riadenia rizík, kvality dát, vypracovanie podrobnej technickej dokumentácie, ako aj dodržiavanie vysokých štandardov v oblasti kybernetickej bezpečnosti, robustnosti, presnosti a zabezpečenia ľudského dohľadu nad systémom. Napríklad AI systém určený na biometrickú identifikáciu musí spĺňať prísne normy, aby sa minimalizovalo riziko jeho zneužitia na podvodné získanie prístupu k systémom [76].

- **Zakázané systémy:** Niektoré formy využívania umelej inteligencie sa považujú za priamu hrozbu pre základné práva a sú podľa Zákona o umelej inteligencii výslovne zakázané. Medzi tieto praktiky patria AI systémy, ktoré manipulujú s ľudským správaním alebo zámerne využívajú zraniteľnosť jednotlivcov, napríklad ich vek, zdravotné postihnutie či sociálnu situáciu, s cieľom ovplyvniť ich rozhodovanie alebo správanie. Príkladom zakázaných systémov sú biometrické technológie, ako je rozpoznávanie emócií na pracovisku alebo systémy kategorizujúce osoby v reálnom čase na základe fyzických či behaviorálnych znakov [76]. V súvislosti s phishingom by sem mohli patriť systémy, ktoré zámerne využívajú umelú inteligenciu na klamanie a manipuláciu obetí. Hoci technológie ako deepfake nie sú samy o sebe zakázané, ich zneužitie na podvodné účely môže predstavovať porušenie zákona v závislosti od konkrétneho kontextu a účelu ich nasadenia.

Hodnotenie vplyvu AI na ľudské práva (FRIA)

Jednou z kľúčových povinností, ktoré nariadenie ukladá, je vykonávanie hodnotenia vplyvu na základné práva (FRIA – Fundamental Rights Impact Assessment) pre všetky vysokorizikové systémy AI. V súlade s článkom 27 nariadenia EU AI Act sú poskytovatelia a nasadzovatelia takýchto systémov povinní ešte pred ich uvedením na trh posúdiť, aký dopad môžu mať na základné práva, ako sú právo na súkromie, ochranu osobných údajov, slobodu prejavu alebo zákaz diskriminácie [4].

FRIA predstavuje systematický proces určený na vyhodnotenie potenciálnych dopadov konkrétnej technológie, politiky alebo činnosti na základné práva jednotlivcov. Je obzvlášť dôležité v kontexte nových technológií, ako je umelá inteligencia, ktorá môže výrazne ovplyvniť etické normy a ľudské práva. Cieľom FRIA je identifikovať a predchádzať možným rizikám, ktoré môžu vysokorizikové systémy AI predstavovať pre jednotlivcov nad rámec technických požiadaviek. Na rozdiel od čisto technických overení, FRIA zdôrazňuje soci-

álny a právny dopad systémov umelej inteligencie. Keďže poskytovatelia AI nedokážu vždy predvídať všetky možné scenáre nasadenia systému, ani všetky formy zaujatosti či možnosti manipulácie, ktoré môžu vzniknúť, FRIA slúži tiež ako nástroj zodpovednosti a transparentnosti [78].

Prepojenie FRIA a ochrany osobných údajov

Hodnotenie dopadu umelej inteligencie na základné práva (FRIA) sa môže v mnohých prípadoch prekrývať s povinnosťou vykonávať hodnotenie vplyvu spracúvania osobných údajov podľa GDPR, známe ako DPIA (Data Protection Impact Assessment). Mnohé vysoko rizikové systémy umelej inteligencie totiž pracujú aj s osobnými údajmi, a preto musia spĺňať požiadavky oboch právnych rámcov – Zákona o umelej inteligencii (AI Act) aj GDPR.

Oba typy hodnotenia sa zameriavajú na posúdenie rizík pre jednotlivcov. DPIA sa sústreďuje na vplyv spracúvania osobných údajov na práva a slobody dotknutých osôb, zatiaľ čo FRIA má širší rozsah a zohľadňuje celkový dopad vysoko rizikových systémov AI na základné práva, slobodu a bezpečnosť fyzických osôb. AI Act tento prechod medzi dvoma hodnoteniami zohľadňuje – konkrétne v článku 27 ods. 4 sa uvádza, že ak boli niektoré povinnosti už splnené prostredníctvom DPIA, FRIA ich dopĺňa, nie nahrádza. V praxi to znamená, že bude výhodné a efektívne realizovať obe hodnotenia spoločne v rámci jedného procesu, s využitím podobných metodických nástrojov. V ideálnom prípade môže byť výsledkom jeden dokument, ktorý pokrýva požiadavky oboch nariadení. Možnosť kombinovať tieto dve hodnotenia patrí medzi opatrenia, ktoré majú pomôcť znížiť administratívnu záťaž spojenú s implementáciou novej regulácie [78, 79].

Rozsah a uplatňovanie nariadenia

Zákon o AI má široký rozsah pôsobnosti a vzťahuje sa na všetkých aktérov zapojených do vývoja, distribúcie a používania systémov umelej inteligencie, ktorí majú prepojenie s Európskou úniou. Konkrétne sa týka nasadzovateľov, distribútorov, dovozcov, výrobcov a dodávateľov systémov AI, či už ide o fyzické alebo právnické osoby, ktoré majú sídlo v EÚ alebo ponúkajú svoje produkty a služby na európskom trhu. Tento rozsah je veľmi podobný pôsobeniu GDPR, čo znamená, že nariadenie má extrateritoriálny dosah a ovplyvňuje subjekty na globálnej úrovni, vrátane organizácií mimo EÚ, ak ich AI systémy pôsobia na európskom trhu [77]. V praxi to znamená, že väčšina subjektov obchodujúcich v globálnom meradle,

ktoré uvádzajú na trh alebo používajú AI systémy v EÚ, musí spĺňať požiadavky vyplývajúce z nariadenia. Z hľadiska územného uplatňovania je toto nariadenie v súlade s hlavnými právnymi predpismi EÚ a vytvára nové povinnosti pre organizácie naprieč sektormi a celým reťazcom poskytovania AI systémov [77, 80].

3.2.2 OECD Principles on AI

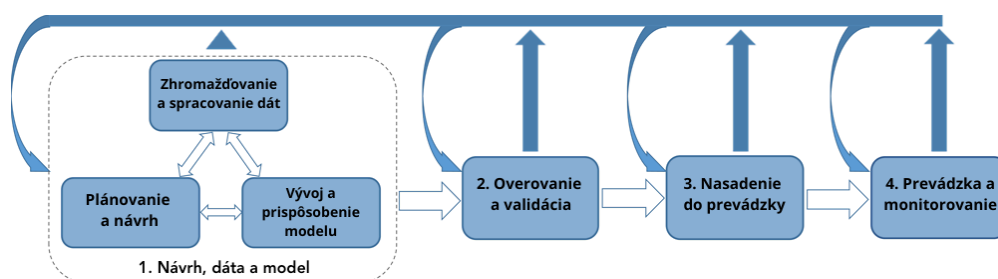
Ďalším významným medzinárodným rámcom sú OECD Principles on AI, ktoré boli prijaté v roku 2019 a aktualizované v roku 2024. Tieto princípy predstavujú súbor piatich kľúčových hodnôt zameraných na zodpovedné využívanie umelej inteligencie a poskytujú praktické a flexibilné usmernenia pre tvorcov politík, vlády a ďalších aktérov v oblasti AI [5]. Ich hlavným cieľom je podporovať dôveryhodnú umelú inteligenciu, ktorá rešpektuje ľudské práva, etické normy a bezpečnosť. Odporúčanie OECD o umelej inteligencii bolo prijaté 46 vládami, vrátane Českej republiky, Slovenskej republiky, Európskej únie a Spojených štátov amerických, čo svedčí o jeho širokej medzinárodnej podpore [81].

Hoci princípy nie sú právne záväzné, ich cieľom je umožniť vládam prispôbiť ich implementáciu do národných právnych a regulačných rámcov [81]. Mnohé krajiny využívajú OECD Principles a súvisiace nástroje na tvorbu politík a rámcov na riadenie rizík spojených s AI, čím podporujú globálnu interoperabilitu medzi jednotlivými jurisdikciami. Napríklad Európska únia, Rada Európy, Spojené štáty americké, Organizácia Spojených národov a ďalšie subjekty zakomponovali definíciu systému AI a jeho životného cyklu podľa OECD do svojich legislatívnych a regulačných usmernení. To prispieva k harmonizácii prístupov k regulácii AI na medzinárodnej úrovni [82].

Definícia systému AI a životného cyklu

OECD Principles on AI poskytujú jasnú definíciu systému umelej inteligencie, ktorá slúži ako základ pre harmonizáciu právnych rámcov na medzinárodnej úrovni. Podľa OECD je systém umelej inteligencie definovaný ako systém založený na strojoch, ktorý na základe vstupov odvodzuje výstupy, ako sú predpovede, obsah, odporúčania alebo rozhodnutia, pre explicitné alebo implicitné ciele, pričom tieto výstupy môžu ovplyvňovať fyzické alebo virtuálne prostredia. Systémy AI sa líšia úrovňou autonómie a prispôbivosti po ich nasadení, čo umožňuje ich široké využitie v rôznych oblastiach [83].

Životný cyklus systému AI, ako ho definuje OECD, zahŕňa viacero fáz, ktoré pokrývajú celý proces od návrhu až po vyradenie systému. Tieto fázy zahŕňajú plánovanie a návrh, zhromažďovanie a spracovanie údajov, vývoj alebo prispôsobenie modelov pre konkrétne úlohy, testovanie, hodnotenie, overovanie a validáciu, nasadenie do prevádzky, prevádzku a monitorovanie, a nakoniec vyradenie systému z prevádzky. Tieto kroky často prebiehajú iteratívne a nemusia byť striktne sekvenčné, čo odráža dynamickú povahu vývoja AI. Rozhodnutie o vyradení systému z prevádzky môže nastať kedykoľvek počas fázy prevádzky a monitorovania, napríklad ak systém prestane spĺňať bezpečnostné alebo etické štandardy [83]. V kontexte kybernetickej bezpečnosti je pochopenie životného cyklu AI kľúčové, pretože každá fáza – najmä zhromažďovanie údajov a nasadenie – môže byť zneužitá na tvorbu phishingových útokov, ako sú deepfake videá alebo vishing, ak nie sú zavedené primerané kontrolné mechanizmy.



Obrázek 3.1 Životný cyklus systému AI

(zdroj: OECD AI in Society, 2019 [84])

Princípy založené na hodnotách

OECD Principles on AI sú postavené na piatich kľúčových hodnotách, ktoré slúžia ako základ pre zodpovedné využívanie umelej inteligencie. Tieto hodnoty sú podľa OECD navrhnuté tak, aby podporovali etický, bezpečný a spravodlivý vývoj a nasadenie AI systémov. Každá hodnota prináša usmernenia na zmiernenie rizík spojených s AI a na ochranu základných práv [5, 81].

1. **Inkluzívny rast, udržateľný rozvoj a blahobyt:** Táto hodnota zdôrazňuje, že AI by mala prispievať k inkluzívnemu rastu a udržateľnému rozvoju, čím zlepšuje blahobyt ľudí a planéty [76, 5]. V kontexte phishingových útokov to môže znamenať, že AI systémy by mali byť navrhnuté tak, aby minimalizovali riziká pre zraniteľné skupiny, ktoré sú často cieľom podvodov.

2. **Hodnoty zamerané na človeka a spravodlivosť:** Subjekty zodpovedné za vývoj a nasadzovanie systémov umelej inteligencie by mali v priebehu celého ich životného cyklu dôsledne rešpektovať zásady právneho štátu, ľudské práva a demokratické hodnoty, a zároveň na tento účel zavádzať primerané záruky [81]. Táto hodnota v kontexte phishingu chráni predovšetkým právo jednotlivcov na bezpečnosť (na tomto mieste sa naskýta otázka, či sa dá pod právo na osobnú bezpečnosť zahrnúť aj právo na túto bezpečnosť vo virtuálnom priestore), ochranu osobných údajov a súkromie, ako aj ich dôstojnosť a slobodu pred podvodnými a manipulatívnymi praktikami.
3. **Transparentnosť a vysvetliteľnosť:** Vývojári AI systémov by mali zabezpečiť, aby boli procesy AI transparentné a zrozumiteľné. Táto hodnota zdôrazňuje, že subjekty zodpovedné za AI by mali poskytovať informácie, ktoré podporujú všeobecné pochopenie fungovania systémov, informovať používateľov o interakciách s AI a umožniť ovplyvneným osobám porozumieť a prípadne spochybniť výsledky rozhodnutí týchto systémov [81]. Táto hodnota je kľúčová pri prevencii phishingových útokov, pretože transparentnosť umožňuje používateľom rozpoznať AI-generovaný obsah, ako sú falošné videá, a vyhnúť sa podvodom. Zároveň ak sú AI systémy transparentné a ich rozhodovanie vysvetliteľné, je oveľa ťažšie využiť ich na tvorbu presvedčivých phishingových kampaní bez toho, aby boli odhalené.
4. **Robustnosť, bezpečnosť a spoľahlivosť:** Táto hodnota zdôrazňuje, že AI systémy musia byť navrhnuté tak, aby v podmienkach bežného, predvídateľného aj nesprávneho použitia nepredstavovali neprimerané bezpečnostné riziko. Subjekty zodpovedné za vývoj AI by mali zabezpečiť sledovateľnosť výstupov svojich systémov a uplatňovať systematický prístup k riadeniu rizík [81, 83]. V súvislosti so zneužívaním AI na phishing táto hodnota znamená, že systémy by mali byť odolné voči manipulácii, zneužitiu a produkcii škodlivého obsahu, napríklad generovania podvodných e-mailov alebo falošných správ.
5. **Zodpovednosť:** Hodnota zodpovednosti stanovuje, že subjekty zodpovedné za vývoj a nasadzovanie AI systémov musia niesť zodpovednosť za správne fungovanie týchto systémov a za dodržiavanie stanovených zásad. V kontexte phishingu to znamená, že vývojári a prevádzkovatelia AI riešení sú zodpovední za to, aby ich technológie neboli zneužívané na tvorbu alebo šírenie škodlivého obsahu, ako sú phishingové správy.

Zodpovednosť zahŕňa aj povinnosť zavádzať mechanizmy, ktoré zabránia alebo aspoň minimalizujú riziko zneužitia systémov na podvodné účely. Táto hodnota má tak za cieľ chrániť predovšetkým práva jednotlivcov na bezpečné digitálne prostredie a zabezpečiť, aby v prípade zlyhaní alebo zneužití existovali jasné mechanizmy na nápravu a vyvodzovanie dôsledkov voči zodpovedným subjektom [81, 83].

Odporúčania pre tvorcov politik

Princípy zahŕňajú päť odporúčaní určených pre tvorcov politik, ktoré majú podporovať zodpovedné využívanie umelej inteligencie a vytvárať prostredie pre dôveryhodné AI systémy. Tieto odporúčania sú zamerané na posilnenie inovácií, etického zdieľania údajov, prípravu na spoločenskú zmenu a medzinárodnú spoluprácu, čo je kľúčové aj pre minimalizáciu rizík spojených s AI. [5].

1. **Investície do výskumu a vývoja AI:** Vlády by mali aktívne podporovať dlhodobé verejné a súkromné investície do výskumu a vývoja AI. To zahŕňa aj vytváranie reprezentatívnych a etických datasetov, ktoré chránia súkromie a zabezpečujú ochranu údajov. Tieto investície majú podnecovať inovácie v dôveryhodnej AI, čím sa znižuje riziko jej zneužitia na podvodné aktivity. V súvislosti s phishingom a inými formami zneužitia je tento prístup kľúčový, pretože investície do výskumu bezpečných, spravodlivých a transparentných technológií AI pomáhajú predchádzať tomu, aby sa umelá inteligencia stala nástrojom na šírenie podvodov, manipulácie alebo kybernetických útokov.
2. **Podpora digitálneho ekosystému pre AI:** Vlády by mali vytvárať digitálny ekosystém pre dôveryhodnú AI podporou mechanizmov, ako sú dátové trusty, ktoré zabezpečia bezpečné, zákonné a etické zdieľanie údajov. Takýto ekosystém je dôležitý na prevenciu zneužitia údajov.
3. **Vytváranie prostredia umožňujúceho politiku pre AI:** Vlády by taktiež mali podporovať agilné politické prostredie, ktoré uľahčuje prechod AI systémov z výskumu do prevádzky, a zároveň prispôbiť regulačné rámce na podporu inovácií a konkurencie v oblasti dôveryhodnej AI. Tieto politiky môžu pomôcť zaviesť štandardy, ktoré obmedzia zneužitie AI na útoky.

4. **Budovanie ľudských kapacít a príprava na transformáciu trhu práce:** Vlády by mali vybavovať ľudí zručnosťami na zvládnutie zmien spôsobených AI, zabezpečiť spravodlivý prechod pre pracovníkov postihnutých AI prostredníctvom školení a podporovať zodpovedné využívanie AI na pracovisku, aby sa zvýšila bezpečnosť a kvalita práce.
5. **Medzinárodná spolupráca pre dôveryhodnú AI:** Taktiež by jednotlivé vlády mali aktívne spolupracovať na presadzovaní princípov AI, zdieľať poznatky na fórach, podporovať globálne technické normy a vytvárať medzinárodne porovnateľné metriky na hodnotenie pokroku v oblasti AI. Spolupráca je kľúčová pre harmonizáciu štandardov [5, 81, 82, 83].

Uplatňovanie nariadenia a donucovacie právomoci

OECD Principles on AI nie sú právne záväzné, a preto ich uplatňovanie závisí od dobrovoľnej implementácie jednotlivými krajinami do národných právnych rámcov. Rovnako odporúčanie neudeľuje vynucovacie právomoci ani nevedie k žiadnym sankciám za nedodržiavanie. Táto flexibilita umožňuje štátom prispôbiť implementáciu vlastným podmienkam, no zároveň vytvára výzvy, pretože v praxi chýbajú účinné mechanizmy na vymáhanie dodržiavania zásad [81]. To môže výrazne obmedziť efektivitu princípov najmä v oblasti prevencie zneužívania AI systémov, napríklad na phishingové útoky či iné kybernetické hrozby.

Príklady implementácie v praxi

OECD Princípy boli implementované viacerými krajinami na podporu dôveryhodnej AI a zmiernenie rizík, vrátane kybernetických hrozieb. V Českej republike boli princípy začlenené do Národnej stratégie umelej inteligencie, ktorá sa zameriava na sedem kľúčových oblastí vrátane etických a bezpečnostných aspektov AI. Cieľom je vytvoriť dôveryhodný AI ekosystém, podporiť výskum a vývoj, a zabezpečiť ochranu pred zneužitím AI technológií [85]. Slovenská republika zahrnula politiku v oblasti AI do širšej stratégie digitálnej transformácie. Plán obsahuje konkrétne kroky na budovanie udržateľného, na človeka zameraného a dôveryhodného ekosystému AI v rámci dlhodobej „*Stratégie digitálnej transformácie Slovenska 2030*“. Tieto dokumenty zdôrazňujú potrebu etickej a bezpečnej AI, ktorá rešpektuje zásady právneho štátu a ľudských práv, čím sa priamo odrážajú zásady OECD [86].

Ďalším príkladom je Európska únia, ktorá prijala už zmienený Akt o umelej inteligencii. Tento zákon klasifikuje AI systémy podľa rizika a stanovuje prísne požiadavky pre systémy s vysokým rizikom. Zásady OECD sa odrážajú v dôraze na transparentnosť, zodpovednosť a ochranu základných práv. AI Act tiež podporuje výskum a inovácie v oblasti dôveryhodnej AI [4]. V Spojených štátoch amerických boli princípy inšpiráciou pre „*Americkú iniciatívu pre umelú inteligenciu*“ (American AI Initiative), ktorá sa zameriava na podporu výskumu a vývoja AI, vzdelávanie pracovnej sily a medzinárodnú spoluprácu. Aj keď USA podporujú zásady OECD, uprednostňujú menej regulovaný prístup, ktorý zdôrazňuje inovácie a konkurencieschopnosť [87].

3.3 Národné právne normy

Aj keď právna regulácia umelej inteligencie prebieha predovšetkým na medzinárodnej úrovni, využívanie AI má významný dopad aj na základné ľudské práva a občiansku spoločnosť. Práve tieto oblasti sú predmetom vnútroštátnej právnej ochrany. V Českej republike sa môžeme oprieť predovšetkým o ustanovenia Ústavy a Listiny základných práv a svobod (Ústavný zákon č. 2/1993 Sb.) [88], ktoré garantujú práva, ako je ochrana súkromia, ochrana osobných údajov či ochrana majetku. Práve tieto práva sú ohrozené zneužitím AI v kybernetickom priestore pri phishingových útokoch.

Rovnako tak sa problematika bezpečného využívania AI a ochrany pred jej zneužitím odráža v právnych predpisoch, ako je Zákon č. 181/2014 Sb., o kybernetickej bezpečnosti, a v strategických dokumentoch, akým je Národní strategie umělé inteligence v České republice. Táto podkapitola analyzuje, ako uvedené normy a stratégie reagujú na nové výzvy spojené s AI v kontexte kybernetickej bezpečnosti a prevencie phishingových útokov.

3.3.1 Ústavné práva relevantné pre AI a phishing

Využívanie umelej inteligencie v kontexte phishingových útokov môžu predstavovať zásah do viacerých základných práv, ktoré sú chránené na ústavnej úrovni v rámci právneho poriadku Českej republiky. Najvýznamnejšie dokumenty v tejto oblasti sú Ústava Českej republiky (zákon č. 1/1993 Sb.) [89] a Listina základných práv a svobod (zákon č. 2/1993 Sb.) [88], ktorá je súčasťou ústavného poriadku. Medzi najviac dotknuté práva, ktoré môžu byť zneužitú pri použití AI pri phishingu patria:

- **Právo na súkromie** (čl. 7 ods. 1 Listiny) – Toto právo chráni jednotlivcov pred neoprávneným zásahom do ich súkromného a rodinného života [88]. Predstavuje prostriedok ochrany osobnej sféry života človeka. AI využívaná na phishing môže zasahovať do súkromného života jednotlivcov, a to prostredníctvom neoprávneného zhromažďovania, spracovania a manipulácie s osobnými údajmi, prípadne prejavmi osobnej povahy (fotografie, videá, nahrávky) s cieľom ich zneužitia.
- **Právo na ochranu osobných údajov** (čl. 10 ods. 3 Listiny) – Toto právo zaručuje ochranu pred zneužitím ústavou a zákonmi vymedzeného okruhu údajov a ako jeden z prostriedkov tejto ochrany využíva informovaný súhlas pri ich spracovaní. Aplikácie AI vyžadujú veľké množstvo údajov, čo vytvára riziko, že spoločnosti môžu zhromažďovať, používať alebo zdieľať osobné identifikačné údaje osôb bez informovaného súhlasu. Phishingové útoky podporované AI môžu viesť k neoprávnenému získavaniu alebo zneužívaniu týchto osobných údajov.
- **Právo na slobodu prejavu a právo na informácie** (čl. 17 Listiny) – Hoci je sloboda prejavu základným právom, v prípade phishingových kampaní môže dochádzať k jej zneužívaniu, napríklad šírením manipulatívnych alebo falošných informácií vytvorených AI nástrojmi.
- **Právo na bezpečnosť a ochranu pred zásahmi do integrity osoby** (čl. 6 a čl. 7 Listiny) – Tieto práva chránia fyzickú a psychickú integritu jednotlivcov a ich bezpečný život [88]. Útoky využívajúce AI, ako napríklad vishingové hovory či deepfake videá, môžu podkopať digitálnu bezpečnosť jednotlivcov aj organizácií, čím sa nepriamo zasahuje do práva na bezpečný život.

Uplatňovanie týchto práv a hlavne ich ochrana vo vzťahu k novým technológiám, ako je umelá inteligencia, je v súčasnosti predmetom intenzívnej diskusie. Český právny rámec zatiaľ neobsahuje špecifické ústavné ustanovenia upravujúce využívanie AI, avšak ochrana základných práv poskytuje dostatočný základ pre posudzovanie nových technológií vrátane ich zneužitia v oblasti phishingu.

3.3.2 Trestný zákon

Aj keď český Trestný zákon (zákon č. 40/2009 Sb.) priamo neupravuje reguláciu umelej inteligencie, jeho ustanovenia sa uplatňujú aj v prípade phishingových útokov, vrátane tých, pri ktorých je využívaná AI. Phishing môže byť kvalifikovaný ako trestný čin podvodu podľa §209 Trestného zákona. Tento paragraf postihuje konanie, pri ktorom páchatel' uvedie inú osobu do omylu alebo využije jej omyl za účelom získania majetkového prospechu. Za tento čin hrozí trest odňatia slobody až na dva roky, v prípade spôsobenia značnej škody alebo spáchania činu organizovanou skupinou môže trest dosiahnuť až desať rokov [90].

Okrem toho môže phishing naplniť aj skutkovú podstatu trestného činu neoprávneného zásahu do počítačového systému alebo nosiča informácií podľa §230 ods. 2 písm. c) Trestného zákona. Tento trestný čin zahŕňa falšovanie a neoprávnené používanie údajov uložených v počítačových systémoch alebo na informačných médiách. Za tento skutok môže byť páchatel' potrestaný trestom odňatia slobody až na tri roky. Uvedené trestné ustanovenia vytvárajú základnú trestnoprávnú ochranu proti phishingovým útokom vrátane tých, ktoré sú realizované pomocou umelej inteligencie, a umožňujú ich postih v rámci platného českého trestného práva [90, 91].

3.3.3 Zákon o kybernetickej bezpečnosti

Zákon č. 181/2014 Sb., o kybernetickej bezpečnosti a o zmene súvisiacich zákonov, ktorý implementuje smernicu EÚ NIS (Directive 2016/1148), predstavuje základný právny predpis v oblasti ochrany kritickej informačnej infraštruktúry a významných informačných systémov v Českej republike. Zákon kladie dôraz na predchádzanie kybernetickým bezpečnostným incidentom, ich riešenie a na zabezpečenie obnovy funkčnosti systémov po útoku. Aj keď zákon priamo nehovorí o phishingu alebo o umelej inteligencii, jeho ustanovenia majú nepriamy dopad na ich reguláciu. Subjekty spadajúce do prednetu úpravy zákona (napr. prevádzkovatelia kritickej infraštruktúry, poskytovatelia digitálnych služieb a niektoré štátne orgány) musia implementovať technické a organizačné opatrenia na zamedzenie kybernetických incidentov. Medzi ne by sa bezpochyby dali zaradiť aj phishingové útoky s využitím AI. Zákon tiež ukladá povinnosť detekcie a hlásenia bezpečnostných incidentov [92].

S pripravovanou implementáciou smernice NIS2 (Directive 2022/2555) sa v roku 2025 očakáva nový zákon o kybernetickej bezpečnosti, ktorý zavedie prísnejšie požiadavky, vrátane rozšírenia okruhu subjektov, na ktoré sa tieto povinnosti vzťahujú (napríklad o stredne veľké a veľké podniky v kritických sektoroch). Cieľom NIS2 je reagovať na nové hrozby vrátane tých, ktoré vyplývajú z používania umelej inteligencie v kybernetických útokoch. S rastúcim využívaním AI pri generovaní phishingových správ (napr. automatizované phishingové kampane, deepfake komunikácia) bude potrebné tieto techniky zahrnúť do bezpečnostných analýz a aktualizácie opatrení kybernetickej ochrany v súlade s požiadavkami zákona [91, 93].

3.3.4 Národná stratégia umelej inteligencie v Českej republike

Národná stratégia umelej inteligencie v Českej republike predstavuje kľúčový dokument, ktorý definuje prístup ČR k rozvoju AI so silným dôrazom na etiku, ochranu práv a bezpečnosť. Stratégia je inšpirovaná OECD Principles on AI a zdôrazňuje nielen podporu inovácií, ale aj potrebu zabezpečiť, aby vývoj a využívanie AI boli v súlade s demokratickými hodnotami, základnými ľudskými právami a požiadavkami kybernetickej bezpečnosti. Národná stratégia je rozdelená do siedmich kľúčových oblastí: Podpora a koncentrácia vedy, výskumu a vývoja a ich financovanie; Využitie AI v priemysle, službách a verejnej správe; Ľudský kapitál a vzdelávací systém; Opatrenia na riešenie dopadov AI na trh práce a sociálny systém; Právne a spoločenské aspekty AI, etické pravidlá, ochrana spotrebiteľa a bezpečnostné otázky; Medzinárodná spolupráca. V rámci týchto oblastí boli definované krátkodobé (do roku 2021), strednodobé (do roku 2027) a dlhodobé (do roku 2035) ciele. Stratégia zároveň identifikuje konkrétne nástroje, prostredníctvom ktorých je možné naplňať vytýčené ciele [85]. V tomto texte sa zameriame na oblasti týkajúce sa etických, právnych a bezpečnostných aspektov AI, ktoré sú vzhľadom na problematiku využitia AI vo phishingových útokoch a právnej regulácie kľúčové.

Etické a právne aspekty AI

Štvrtá oblasť s názvom "*Etické a právne aspekty AI*" poukazuje na to, že s rastúcim významom AI narastajú aj riziká spojené tak s ochranou základných práv a slobôd ako aj všeobecne akceptovaných etických pravidiel. Vývoj a využívanie AI otvára okrem právnych otázok aj otázky etické, pričom ich čiastočnú reflexiu realizuje právo. Tieto riziká si vyžadujú primerané právne a regulačné opatrenia na medzinárodnej aj národnej úrovni. Česká republika sa

preto usiluje o aktívnu participáciu na medzinárodných fórach (napr. Rada Európy, OECD, UNESCO), kde sa formujú globálne pravidlá pre vývoj a využívanie AI. Dôraz sa kladie na vytvorenie predvídateľného právneho rámca v súlade s ústavnými princípmi a medzinárodnými ľudskoprávnymi záväzkami, ako aj na flexibilitu legislatívy, aby umožnila adaptáciu na rýchly technologický pokrok pri zachovaní ochrany základných práv. Konkrétne ciele v oblasti etických a právnych aspektov AI stanovené v stratégii sú [85]:

- **Cieľ 1:** Vytvoriť predvídateľný právny rámec pre rozvoj AI, ktorý adekvátne chráni jednotlivcov a ich práva.
- **Cieľ 2:** Vybudovať expertné kapacity pre dozor a vymáhanie legislatívnych požiadaviek v oblasti AI.
- **Cieľ 3:** Podporovať dostupnosť soft-law nástrojov (odporúčaní, metodík) a vzdelávanie s cieľom posilniť dôveru verejnosti v etické a práva rešpektujúce využívanie umelej inteligencie.

Bezpečnostné aspekty AI

Piata oblasť sa venuje bezpečnostným aspektom umelej inteligencie. Stratégia deklaruje víziu, podľa ktorej musia byť AI systémy vyvíjané, prevádzkované a využívané v Českej republike v súlade s vysokými štandardmi kybernetickej bezpečnosti, pričom cieľom je minimalizovať ich zraniteľnosť, predovšetkým z hľadiska ekonomickej bezpečnosti. Bezpečnostné aspekty sú v prípade phishingových útokov využívajúcich AI mimoriadne dôležité, pretože nízka úroveň kybernetickej bezpečnosti môže viesť k zneužitiu umelej inteligencie na podvodné aktivity. Kľúčovým cieľom v oblasti bezpečnosti je zabezpečiť, aby bola AI vyvíjaná, implementovaná a využívaná bezpečným spôsobom vrátane ochrany pred kybernetickými útokmi a minimalizácie technologických rizík [85].

3.4 Etické otázky spojené s využitím AI vo phishingu

Etické dilemy spojené s využívaním umelej inteligencie sú často oveľa širšie a zložitejšie než len právne otázky. Ich vnímanie môže byť subjektívne, pretože hranica medzi tým, čo je etické a neetické, nie je vždy jasná. Kľúčovou otázkou je, kde by mal štát zasahovať prostredníctvom regulácie a kde by mal nechať priestor pre zodpovedné správanie jednotlivcov,

organizácií a spoločností, aby sa zachovala ich sloboda a autonómia. V tejto podkapitole sa zameriame na hlavné etické problémy a preberieme dilemy, ktoré vyplývajú z využívania AI v kybernetických útokoch, ako sú deepfake videá, phishing alebo vishing. Cieľom je poukázať na potrebu existencie a uplatnenia etických zásad pri návrhu budúcich riešení.

Významným príspevkom k riešeniu etických otázok AI je dokument *"UNESCO Recommendation on the Ethics of Artificial Intelligence"*, prijatý v roku 2021. Tento dokument je prvým globálnym rámcom, ktorý poskytuje súbor etických zásad pre vývoj a využívanie AI, s dôrazom na ochranu ľudských práv, transparentnosť a zodpovednosť. UNESCO kladie osobitnú pozornosť na minimalizáciu rizík, ako je zneužitie AI na manipuláciu a klamanie, čo je priamo relevantné aj pre phishingové útoky. Dokument slúži ako usmernenie pre vlády, organizácie a vývojárov, aby zabezpečili, že AI bude slúžiť na prospech spoločnosti a nebude ohrozovať základné práva jednotlivcov [94].

Klasifikácia etických problémov

3.4.1 Súkromie a zneužitie dát

Umelá inteligencia, najmä veľké jazykové modely (LLM), sa trénujú na rozsiahlych súboroch údajov, ktoré často zahŕňajú osobné informácie získané bez súhlasu ich vlastníkov, čo vyvoláva vážne etické otázky o spôsobe ich získavania, zhromažďovania, používania a zdieľania. AI systémy vo všeobecnosti nezvyšujú úroveň súkromia ani ochrany údajov – naopak, veľké objemy dát spracovávaných vývojármi sa stávajú atraktívnym cieľom pre útočníkov, čím sa zvyšuje riziko ich narušenia a neoprávneného prístupu [95]. Útočníci využívajú AI na získavanie údajov zo sociálnych sietí alebo iných zdrojov, aby personalizovali phishingové útoky, ako sú podvodné e-maily, čo umožňuje presvedčivejšie klamanie obetí. Organizácie uchovávajúce tieto údaje nesú zodpovednosť za ich bezpečnosť, no často zlyhávajú v ochrane pred únikmi, čím uľahčujú zneužitie dát na klamlivé kampane [96].

Implementácia AI musí byť v súlade s právnymi normami, ako je GDPR (Regulation 2016/679), ktoré vyžadujú minimalizáciu dát a súhlas so spracovaním. Avšak veľký objem údajov a ich citlivosť túto požiadavku komplikujú. Súčasná stratégia ochrany, ako dezinfekcia údajov (odstraňovanie citlivých informácií) či diferenciálne súkromie (matematický rámec na kvantifikáciu straty súkromia), poskytujú iba obmedzenú ochranu pri použití na LLM, čo vedie k otázke, ako eticky spracovávať osobné údaje na tréning týchto modelov.

Navyše, používatelia môžu do AI systémov zadávať súkromné informácie prostredníctvom výziev, ktoré sa môžu neskôr objaviť v iných kontextoch, čím sa riziko zneužitia ďalej zvyšuje. Niektorí používatelia môžu byť navyše neplnoletí, a teda na nich sa vzťahujú prísnejšie pravidlá o spracovaní a ochrane údajov [96].

3.4.2 Manipulácia a klamanie

Generatívna umelá inteligencia, najmä technológia deepfake, umožňuje vytváranie realistických videí, obrázkov a zvukových nahrávok, ktoré manipulujú s obsahom tak, aby zobrazovali jednotlivcov, ako hovoria alebo robia veci, ktoré v skutočnosti nevykonali. Modely hlbokého učenia, trénované na obrovských súboroch údajov, dokážu s vysokou presnosťou napodobniť ľudské správanie, výrazy či hlas, ako je to pri vishingu, kde hlasové klonovanie slúži na oklamanie obetí [97]. Tieto deepfake videá a zvuky narušajú dôveru v digitálnu komunikáciu, pretože vytvárajú falošné identity, ktoré môžu viesť k finančným stratám, emocionálnym škodám alebo poškodeniu reputácie [98].

Etické otázky spojené s deepfake technológiami zahŕňajú problémy súkromia a súhlasu, keďže údaje použité na ich tvorbu sú často získavané bez povolenia, čo zvyšuje riziko ich zneužitia na manipulatívne účely, čo je etický problém sám o sebe. Šírenie deepfake obsahu na internete vedie k narušeniu dôvery verejnosti v digitálne médiá, čo spôsobuje skepticizmus a neistotu ohľadom autenticity online informácií ale môže mať aj sekundárny negatívny vplyv na spoločenskú súdržnosť a medzil'udské vzťahy. Okrem toho konzumácia zavádzajúceho alebo falšovaného obsahu môže mať hlboké psychologické dopady na jednotlivcov, ovplyvňovať ich vnímanie, presvedčenia a správanie, a v kontexte phishingových útokov zvyšovať zraniteľnosť obetí voči podvodom, ako sú podvodné e-maily, deepfakes alebo vishingové kampane [99].

3.4.3 Zodpovednosť

Kto je zodpovedný za škody – tvorca AI technológie, útočník, alebo obeť, ktorá nebola dostatočne obozretná? Dilema zodpovednosti spočíva v otázke, kto nesie vinu za škody spôsobené AI – či tvorca technológie, útočník, ktorý ju zneužil, alebo obeť, ktorá nebola dostatočne obozretná, napríklad pri rozpoznávaní podvodného hovoru [98].

Na úrovni EÚ bola v roku 2022 navrhnutá Smernica o zodpovednosti za umelú inteligencia, ktorá mala za cieľ zjednotiť pravidlá pre mimozmluvnú občianskoprávnú zodpovednosť

za škody spôsobené AI systémami, čím by sa ul'ahčilo určenie vinníka v prípadoch, ako je zneužitie AI na phishing [100]. Európska komisia však tento návrh vo februári tohto roka (2025), spolu s nariadením o elektronickom súkromí, stiahla s cieľom zvýšiť konkurencieschopnosť európskych AI spoločností. Tento krok kritizuje Montreal AI Ethics Institute, ktorý tvrdí, že takýto prístup uprednostňuje záujmy firiem pred verejným záujmom, ako je ochrana súkromia, dôstojnosti a zmysluplná občianska angažovanosť, čo by mohlo obmedziť zneužívanie AI systémov považovaných verejnosťou za problematické, vrátane ich využitia na rôzne druhy útokov [101].

3.4.4 Spravodlivosť

AI-generované phishingové útoky neúmerne poškodzujú zraniteľné skupiny, ako sú starší ľudia alebo menej technicky zdatní jednotlivci, ktorí často nedokážu rozpoznať podvodné e-maily, videá či hovory, čím sa zvyšuje digitálna nerovnosť a diskriminácia [102]. Zaujatosť v systémoch AI predstavuje ďalší etický problém. Predpojaté algoritmy môžu nespravodlivo uprednostňovať určité skupiny alebo jednotlivcov. Ako správne uvádza portál Britannica Money, určovanie spravodlivosti a zaujatosti je značne subjektívne a generatívne modely AI do istej miery odrážajú svet taký, aký je, a nie taký, aký by sme si priali, aby bol. Súčasný modely sú v tomto smere stále nedokonalé [95]. Napríklad vyhľadávacie systémy využívajúce AI nie sú neutrálne, pretože spracovávajú veľké dáta a uprednostňujú výsledky na základe počtu kliknutí, preferencií používateľa či jeho polohy, čím vytvárajú ozvenové komory, ktoré posilňujú stereotypy a predsudky. V kontexte phishingu to môže znamenať, že zraniteľné skupiny sú častejšie vystavené cieleným podvodným kampaniam. Tieto problémy spravodlivosti a zaujatosti v AI systémoch zostávajú nedostatočne vyriešené, čo zdôrazňuje potrebu etických riešení na ochranu zraniteľných skupín pred kybernetickými hrozbami [94].

3.4.5 Duševné vlastníctvo

Generatívna AI prináša etické a právne výzvy týkajúce sa duševného vlastníctva, najmä v súvislosti s jej „kreatívnymi“ výstupmi, ktoré vznikajú spracovaním textových údajov, ako sú webové stránky, učebnice, noviny, vedecké články či programovacie kódy. Používanie veľkých jazykových modelov (LLM) vyvoláva otázky autorských práv, pretože tréningové súbory údajov často obsahujú materiály chránené autorským právom, na ktorých použitie je potrebný súhlas vlastníkov práv alebo právna výnimka – čo je však komplikované vzhľadom

dom na obrovský rozsah dát a počet potenciálnych držiteľov práv. Navyše, prax zoškra-
bovania webu, ktorá je bežne využívaná pri zbere dát na tréning LLM, je právne sporná
a naďalej predmetom diskusií na súdoch aj vo vedeckých kruhoch v Európe. Súčasná le-
gislatíva EÚ, vrátane Zákona o umelej inteligencii, zatiaľ neposkytuje komplexné odpovede
na tieto otázky, čo zdôrazňuje potrebu nových právnych rámcov na ochranu duševného vlast-
níctva v kontexte generatívnej AI. Hoci táto problematika nemá priamy dopad na phishin-
gové útoky, zneužitie chráneného obsahu pri tvorbe AI modelov môže nepriamo uľahčiť
tvorbu podvodného obsahu, ako sú deepfake videá, čím sa zvyšuje riziko kybernetických
hrozieb [96].

II PRAKTICKÁ ČÁST

4 NÁVRHY NA ZLEPŠENIE ZVLÁDANIA HROZIEB SPOJENÝCH S UMELOU INTELIGENCIOU

S rýchlym rozvojom umelej inteligencie a jej zneužívaním v kybernetických útokoch, je nevyhnutné navrhnuť efektívne opatrenia na zmiernenie týchto hrozieb. Táto kapitola predstavuje praktickú časť práce, ktorá sa zameriava na návrh konkrétnych riešení v štyroch kľúčových oblastiach: technických, legislatívnych, vzdelávacích a etických. Tieto oblasti boli vybrané, pretože spoločne pokrývajú technickú, právnu, sociálnu aj morálnu dimenziu problému, čím umožňujú komplexný prístup k ochrane pred AI-generovanými hrozbami a minimalizácii ich dopadov na jednotlivcov a spoločnosť.

Technické opatrenia sú kľúčové, pretože poskytujú okamžité a proaktívne nástroje na detekciu a prevenciu kybernetických útokov, čím znižujú riziko úspešnosti phishingových kampaní a chránia digitálnu bezpečnosť používateľov. Legislatívne a regulačné opatrenia zasa vytvárajú právny rámec, ktorý zaisťuje zodpovednosť tvorcov a používateľov AI, čím bránia jej zneužívaniu a podporujú férové využívanie technológií. Vzdelávanie a zvyšovanie povedomia je nevyhnutné na posilnenie odolnosti jednotlivcov a organizácií voči podvodom, keďže informovaní používatelia sú menej náchylní stať sa obeťami AI-generovaných útokov. Napokon, etické zásady pre vývoj a využitie AI sú dôležité na zaistenie, aby technológie slúžili prospechu spoločnosti a minimalizovali riziká, ako je manipulácia či diskriminácia, čím sa podporuje dôvera v AI systémy.

4.1 Technické opatrenia

Táto podkapitola sa zameriava na technické riešenia, ktoré môžu účinne pomôcť pri odhalovaní a zmiernení rizík spojených s AI-generovanými phishingovými útokmi, deepfake videami či vishingom. Rozoberáme tu moderné prístupy v oblasti strojového učenia, hlbokých neurónových sietí, spracovania prirodzeného jazyka, behaviorálnej analýzy a autentifikačných technológií. Cieľom je navrhnuť komplexnú obrannú stratégiu, ktorá zvýši odolnosť informačných systémov aj samotných používateľov voči novým a sofistikovaným technikám sociálneho inžinierstva, podporovaným umelou inteligenciou.

4.1.1 Zlepšenie detekčných systémov

V posledných mesiacoch sme svedkami dramatického nárastu phishingových útokov, ktoré využívajú generatívnu umelú inteligenciu, ako sú jazykové modely typu ChatGPT, Claude alebo Gemini. Tieto útoky sa vyznačujú vysokou mierou personalizácie, realistickým a gramaticky správnym jazykom, dynamickým prispôbením obsahu a schopnosťou obísť tradičné detekčné mechanizmy, ktoré sa spoliehajú na filtrovanie kľúčových slov alebo známe vzory. Preto už konvenčné bezpečnostné systémy, ako sú statické blacklisty a heuristické pravidlá, nie sú dostatočne účinné. V súčasnosti tak vzniká naliehavá potreba po pokročilejších, adaptívnych riešeniach. Na tento nový typ hrozby je tak potrebné reagovať technológiami, ktoré dokážu:

- analyzovať kontext a význam jazyka, nielen jeho povrchovú štruktúru,
- rozpoznať skryté alebo nové vzorce správania,
- odhaliť komplexné a nenápadné manipulácie vo vizuálnych aj textových dátach,
- a sledovať zmeny v správaní používateľov, ktoré môžu indikovať úspešný phishingový útok.

Preto navrhujeme použiť kombináciu techník strojového učenia (ML), hĺbkového učenia (DL), spracovania prirodzeného jazyka (NLP), ensemble prístupov, adversariálneho učenia a analýzy správania používateľov. Každá z týchto metód má svoje špecifické výhody. Niektoré sa sústreďujú na textový obsah správ, iné sa zameriavajú na nezvyčajné správanie používateľov alebo na vizuálne znaky podvodov. Kombinovaním týchto techník môžeme vytvoriť robustnejší detekčný systém, ktorý dokáže identifikovať aj nové, doteraz neznáme formy útokov a tým lepšie chrániť používateľov pred podvodmi.

1. **Techniky strojového učenia (ML):** Modely ML, ako sú rozhodovacie stromy, algoritmy typu Support Vector Machine alebo Random Forest, sa trénujú na historických dátach obsahujúcich príklady phishingových a legítimných e-mailov. Tieto modely identifikujú vzorce, ako sú podozrivé URL adresy, nezvyčajné hlavičky e-mailov alebo frekvencia spamových slov, a klasifikujú e-maily ako potenciálne hrozby. Ich prínos spočíva v schopnosti rýchlo spracovať veľké objemy dát a poskytnúť vysokú presnosť

pri detekcii, vďaka čomu dochádza k menšiemu počtu nesprávnych varovaní [103]. Ich účinnosť potvrdila aj štúdia publikovaná v ScienceDirect (2024), kde výskumníci testovali viaceré ML klasifikátory, vrátane rozhodovacích stromov a náhodných lesov, na veľkej databáze phishingových a bežných URL adries. Výsledky ukázali, že pri správnom výbere vlastností, ako je dĺžka webovej adresy, výskyt nezvyčajných znakov či dôveryhodnosť domény, môžu tieto modely veľmi presne odhaliť phishingové pokusy [104].

2. **Techniky hĺbkového učenia (DL):** Hlboké neurónové siete (Deep Neural Networks, DNN), ako sú konvolučné neurónové siete (CNN) alebo rekurentné neurónové siete (RNN), sú efektívne pri analýze zložitých vzorcov v textoch a obrázkoch. V kontexte phishingu dokážu identifikovať vizuálne prvky v e-mailoch (napr. falošné logá) alebo sekvenčné vzorce v textoch, ktoré naznačujú podvodnú komunikáciu. Hĺbkové učenie zlepšuje detekciu tým, že sa dokáže prispôbiť novým a sofistikovaným útokom, ako sú AI-generované e-maily, avšak vyžaduje väčšie výpočtové zdroje a kvalitné tréningové dáta [103]. V jednej zo štúdií výskumníci z Cyber Threat Analytics vytvorili model postavený na CNN, ktorý sa zameriaval na detekciu phishingových webových stránok. Namiesto klasického spracovania textu tento model analyzoval vizuálny vzhľad stránok. Dokázal tak odhaliť aj jemné rozdiely medzi podvodnými a legítimnými stránkami a dosiahol lepšie výsledky ako tradičné modely strojového učenia [105].
3. **Techniky spracovania prirodzeného jazyka (NLP):** NLP je ďalšia technika AI aplikovaná na detekciu phishingu, najmä na analýzu textového obsahu e-mailov a správ. Modely NLP, je možné použiť na analýzu štruktúry a obsahu textu v e-mailoch, aby odhalili podozrivé jazykové vzorce, napr. manipulatívne frázy, neobvyklé tóny alebo gramatické anomálie, ktoré sú typické pre phishingové správy. NLP môže zlepšiť detekciu tým, že umožňuje hlbšiu analýzu sémantiky a kontextu, čím dokáže odlíšiť legítimnú komunikáciu od podvodnej, napríklad pri e-mailoch napodobňujúcich bankové inštitúcie [103, 106].
4. **Techniky ensemble učenia:** Jedným zo spôsobov, ako výrazne zvýšiť presnosť a spoľahlivosť detekčných systémov proti AI-generovanému phishingu, je zavedenie ensemble techník. Ide o prístup, pri ktorom sa kombinuje viacero rôznych modelov strojového učenia, pričom každý z nich prispieva svojím „hlasom“ k finálnemu rozhod-

nutiu. Tento prístup sa ukázal ako mimoriadne efektívny v úlohách, kde je dôležitá vysoká presnosť a odolnosť voči chybám [103]. Existujú rôzne typy ensemble metód, kde medzi najznámejšie patrí bagging (napr. Random Forest), kde sa rovnaký model trénuje na rôznych častiach dát, a boosting (napr. XGBoost), kde sa modely trénujú postupne a každý nový sa zameriava na chyby toho predchádzajúceho. Výsledkom je systém, ktorý dokáže zohľadniť široké spektrum znakov od štruktúry a obsahu e-mailu až po správanie odosielateľa, a na základe kombinácie výstupov viacerých modelov určiť, či ide o phishingovú hrozbu. Ich prínos tak spočíva v znížení chybovosti jednotlivých modelov a zlepšení odolnosti voči novým phishingovým technikám [107]. Z tohto dôvodu odporúčame zaradiť ensemble učenie ako súčasť navrhovaného systému na zlepšenie detekcie phishingových útokov, najmä v kombinácii s ďalšími technikami.

5. **Adversariálne strojové učenie:** S narastajúcou účinnosťou detekčných systémov sa útočníci stále častejšie uchýľujú k prepracovaným technikám, ktorých cieľom je oklamať modely strojového učenia. Tieto techniky sú známe ako adversariálne útoky. Patria medzi ne napríklad mierne úpravy obsahu e-mailov alebo URL adries, ktoré sú pre človeka takmer nepostrehnuteľné, no postačujú na to, aby model phishing nerozpoznal. V reakcii na to, je vhodné integrovať do detekčných systémov adversariálne strojové učenie, ktoré umožňuje modely cielene trénovať na adversariálnych príkladoch. Ide o synteticky vytvorené phishingové správy, ktoré simulujú realistické pokusy o obídenie detekcie. Týmto spôsobom sa modely „očkujú“ proti manipuláciám, ktoré by inak mohli spôsobiť zlyhanie v detekcii [103]. Konkrétny prístup predstavujú generatívne adversariálne siete (GAN), ktoré vytvárajú veľmi realistické falošné phishingové e-maily. Tieto správy sa potom používajú ako tréningová alebo testovacia vzorka na preverenie odolnosti modelov. Tento prístup podporuje aj štúdiá publikovaná v Pattern Recognition, ktorá ukazuje, ako môžu adversariálne metódy významne prispieť k robustnosti detekčných systémov tým, že predvídajú a simulujú techniky, ktoré by mohli zneužiť útočníci [108].

6. **Analýza správania používateľov:** väčšina detekčných systémov sa zameriava na samotný obsah e-mailov a preto odporúčame rozšíriť obranné mechanizmy aj o analýzu správania používateľov. Tento prístup sleduje, ako sa používatelia správajú v digitálnom prostredí [109]. Zmeny v bežnom správaní, ako sú neštandardné časy prihlásenia, prístup z netypickej geolokácie alebo náhle zmeny v interakcii s aplikáciou, môžu byť známkou kompromitácie účtu, často spôsobenej práve úspešným phishingovým útokom. Systémy postavené na strojovom učení dokážu tieto vzorce identifikovať a automaticky vyhodnotiť ako potenciálne rizikové správanie [103, 110]. Návrh spočíva v nasadení behaviorálnych analytických nástrojov ako doplnkovej vrstvy ochrany, ktorá zachytáva anomálie, ktoré by bežná textová analýza prehliadla a reaguje na ne v reálnom čase (napr. upozornením administrátora alebo dočasným obmedzením prístupu). Táto technika je obzvlášť účinná pri odhaľovaní tzv. "post-click" fázy phishingu, teda situácií, keď používateľ klikol na škodlivý odkaz a jeho aktivita sa následne zmení. Kombinácia behaviorálnej analýzy s obsahovou detekciou phishingu tak výrazne zvyšuje šancu na včasné odhalenie útoku a zabránenie jeho šíreniu.

Zhrnutím všetkých týchto prístupov je jasné, že pokročilé metódy umelej inteligencie môžu naozaj posunúť detekciu AI-generovaného phishingu na novú úroveň. Ich implementácia prináša viacvrstvovú ochranu, ktorá dokáže analyzovať nielen obsah a formu e-mailov, ale aj správanie používateľov a možné manipulácie zo strany útočníkov. Spojením rôznych techník je možné vytvoriť silné, adaptívne a prediktívne detekčné systémy. Tieto systémy nielenže lepšie rozpoznávajú známe hrozby, ale sú tiež schopné efektívne reagovať na nové a doteraz neznáme útoky.

4.1.2 Autentifikačné technológie

S rastúcim rizikom phishingových útokov, ktoré sú čoraz realistickejšie a dokážu presne napodobniť jazyk, tón a vizuálny vzhľad legítimnej komunikácie, sa overovanie identity používateľov stáva kľúčovým prvkom obrany. Ďalším technickým návrhom, ako tieto hrozby znížiť, je posilnenie bezpečnostných opatrení pomocou pokročilých autentifikačných technológií, ktoré presahujú tradičné zadávanie mena a hesla. Tieto technológie nielenže chránia prístup k účtom, ale aj zabezpečujú, že komunikácia medzi používateľmi a systémami je legítimná. Návrhy autentifikačných technológií:

Viacfaktorová autentifikácia (MFA): Zavedenie viacfaktorovej autentifikácie, ktorá kombinuje aspoň dve z nasledovných kategórií – niečo, čo používateľ vie (heslo alebo PIN), niečo, čo má (hardvérový token, mobilné zariadenie) a niečo, čím je (biometria), môže predstavovať základnú obranu proti phishingu. MFA totiž výrazne znižuje šancu útočníka získať úplný prístup, aj keď kompromituje niektorý z faktorov [111]. Pri phishingu je najčastejšie cieľom odcudzenie prihlasovacích údajov, ale MFA bráni ich jednoduchému zneužitiu, najmä ak druhý faktor nie je dostupný útočníkovi.

Biometrická autentifikácia: Biometrické metódy, ako je rozpoznávanie odtlačkov prstov, skenovanie dúhovky alebo rozpoznávanie tváre, predstavujú moderný spôsob overenia identity, ktorý je ťažšie sfaľovať než tradičné metódy. Tieto techniky sa čoraz častejšie integrujú do mobilných zariadení a bezpečnostných tokenov (napr. YubiKey Bio), čím umožňujú pohodlné a bezpečné prihlasovanie bez potreby hesiel. V súvislosti s AI phishingom je ich výhoda v tom, že aj keď je používateľ oklamáný a poskytne prístupové údaje, fyzická biometrická kontrola môže zabrániť zneužitiu identifikácie [112, 113].

FIDO2 a WebAuthn: Moderné otvorené štandardy ako FIDO2 (Fast Identity Online) a jeho súčasť WebAuthn (Web Authentication API) umožňujú silnú autentifikáciu bez hesiel pomocou verejno-súkromných kryptografických kľúčov. Identita používateľa sa potvrdzuje prostredníctvom zariadenia (napr. bezpečnostného kľúča alebo biometrického senzora), ktorý nikdy neprenáša samotný kľúč, čím sa minimalizuje možnosť odpočúvania alebo zneužitia pri phishingových útokoch [114, 115]. Implementácia FIDO2 v podnikových systémoch a emailových platformách by mala byť jednou z priorít v boji proti AI-phishingu.

Dynamická autentifikácia a analýza správania: Využitie adaptívnej autentifikácie, ktorá sleduje kontextové informácie ako geografická poloha, typ zariadenia, história prihlásení alebo rýchlosť zadávania údajov, umožňuje systému detekovať podozrivé prístupy v reálnom čase. Táto technika sa často spája s behaviorálnou biometriou, napríklad s rozpoznávaním štýlu písania (keystroke dynamics), ktorá je pre každého používateľa unikátna. V prípade, že je autentifikácia zvyčajným spôsobom „technicky správna“, no v neštandardnom kontexte, môže systém automaticky vyžiadať dodatočné overenie alebo zablokovat prístup [116, 117].

Návrh riešenia: Vzhľadom na vyššie uvedené návrhy technológií, odporúčame implementovať kombinovaný autentifikačný rámec, ktorý bude:

- štandardne používať MFA s biometrickým faktorom (napr. odtlačok prsta cez mobilnú aplikáciu),
- integrovať FIDO2/WebAuthn pre všetky webové rozhrania, kde je to možné,
- využívať dynamickú autentifikáciu ako obrannú vrstvu proti útokom typu „man-in-the-middle“,
- automaticky vyhodnocovať anomálie v prístupoch cez behaviorálnu analýzu v pozadí.

Takýto systém bude menej závislý od hesiel, odolnejší voči phishingu, hlavne vrátane jeho AI-generovaných foriem, a zároveň použiteľný pre široké spektrum používateľov bez výrazného zníženia komfortu.

4.1.3 Detekcia deepfake a hlasového klonovania

S narastajúcou kvalitou generatívnych modelov, ako sú GAN (Generative Adversarial Networks), difúzne modely alebo pokročilé algoritmy TTS (Text-to-Speech), o ktorých sme hovorili v predchádzajúcich kapitolách, sa zvyšuje aj riziko zneužitia týchto technológií na vytváranie falošného multimediálneho obsahu. Tieto technológie sú čoraz častejšie využívané v cielených phishingových kampaniach, vrátane vishingu a video impersonácie, čo ohrozuje dôveru v digitálnu komunikáciu a zvyšuje riziko finančných a emocionálnych škôd pre obete. Vzhľadom na rýchlu expanziu týchto technológií a ich zneužívanie v praxi považujeme zavedenie detekčných riešení priamo do komunikačných nástrojov za nevyhnutnú a neodkladnú súčasť komplexnej obrany proti AI phishingu.

Technológie detekcie deepfake: Moderné detekčné systémy využívajú strojové učenie, predovšetkým konvolučné neurónové siete (CNN), aby analyzovali vizuálne anomálie, ktoré sa často objavujú v deepfake videách. Tieto anomálie môžu zahŕňať drobné nezrovnalosti v pohyboch pier, očí a tvárových svalov, ako aj nesúlad medzi osvetlením a tieňovaním na tvári. Objavuje sa tiež nedostatok prirodzenej mimiky a spontánnosti pohybu, či chýbajúce alebo opakujúce sa vizuálne artefakty (napríklad v ušiach alebo vo vlasoch). Jedným z nástrojov, ktoré využívajú tieto techniky, je Deepware Scanner, ktorý v reálnom čase analyzuje

video obsah a upozorňuje na pravdepodobnosť manipulácie [118]. Podobné funkcie majú aj systémy ako Microsoft Video Authenticator, ktorý dokáže označiť falošný obsah priamo vo video streamoch. „*V prípade videa dokáže poskytnúť toto percento v reálnom čase pre každý záber počas prehrávania videa. Funguje to tak, že detekuje hranicu prelínania deep-fake prvkov a jemného vyblednutia alebo prvkov v odtieňoch sivej, ktoré ľudské oko nemusí rozpoznať*“, vysvetľuje blog spoločnosti Microsoft [119].

Detekcia syntetického hlasu a klonovania: Technológie hlasového klonovania založené na modeloch ako Tacotron 2, VITS alebo ElevenLabs Voice AI umožňujú realistickú imitáciu hlasu na základe niekoľkokundového záznamu. Ich zneužitie v rámci vishingu predstavuje novú hrozbu. Detekcia týchto syntetických hlasov prebieha pomocou analýzy frekvenčných vzorcov (spectrogram consistency), detekcie chýbajúcej prirodzenej variability hlasu (prosodic irregularity), a deep learning klasifikátorov trénovaných na syntetických vs. autentických hlasových záznamoch. Príkladom takejto technológie je Resemble Detect od Resemble AI alebo open-source knižnica AntiSpoofingNet, ktoré analyzujú hlas v reálnom čase a klasifikujú pravdepodobnosť klonovania [120].

4.2 Legislatívne a regulačné opatrenia

Legislatívne a regulačné opatrenia sú nevyhnutné na vytvorenie právneho rámca, ktorý by efektívne riešil hrozby spojené s umelou inteligenciou v kybernetickom priestore, najmä v oblasti phishingových útokov využívajúcich manipulatívne techniky. Jasné právne normy a regulácie môžu zabezpečiť zodpovednosť tvorcov a používateľov AI, podporiť ochranu dát a posilniť medzinárodnú spoluprácu, čím sa zníži riziko zneužitia technológií a ochráni spoločnosť pred kybernetickými hrozbami.

4.2.1 Právna zodpovednosť za škody spôsobené AI

Je potrebné jednoznačne zakotviť právnu zodpovednosť za škody spôsobené systémami využívajúcimi umelú inteligenciu, vrátane prípadov, keď sa tieto systémy stanú nástrojom phishingových útokov. S rastúcou mierou autonómie a zložitosti AI systémov je čoraz ťažšie určiť, kto presne nesie zodpovednosť za spôsobené škody. Nejednoznačné právne prostredie môže odradiť poškodených domáhať sa svojich práv a zároveň znižuje zodpovednosť vývojárov a poskytovateľov AI riešení.

Aktuálny legislatívny rámec v EÚ: V predchádzajúcej kapitole 3.2.1 bol predstavený prijatý európsky AI Act, ktorý klasifikuje systémy AI podľa úrovne rizika a stanovuje povinnosti pre vývojárov, poskytovateľov a používateľov. Tento legislatívny rámec však priamo neupravuje problematiku občianskoprávnej alebo trestnoprávnej zodpovednosti za škody spôsobené AI. Na tento účel bola navrhnutá doplňujúca legislatíva, tzv. **smernica o zodpovednosti za umelú inteligenciu (AILD)**, ktorá mala uľahčiť postavenie poškodených zavedením tzv. vyvrátiteľnej domnienky príčinnej súvislosti medzi AI systémom a vzniknutou škodou. Táto smernica však bola vo februári roku 2025 stiahnutá pre nedostatok politického konsenzu a obavy z možného brzdenia inovácií [121]. V dôsledku toho v EÚ dnes chýba jednotný právny mechanizmus pre riešenie škôd spôsobených AI.

Návrhy legislatívnych opatrení

- **Určenie zodpovednosti subjektov zapojených do vývoja a používania AI:** Legislatíva by mala jasne vymedziť zodpovednosť rôznych aktérov v AI ekosystéme a to najmä vývojárov, poskytovateľov služieb a koncových používateľov. V prípadoch zneužitia AI na phishingové útoky by napríklad poskytovateľ generatívneho modelu mohol niesť spoluzodpovednosť, ak vedome ignoroval bezpečnostné riziká alebo neimplementoval dostatočné bezpečnostné opatrenia, ako napríklad mechanizmy na detekciu zneužitia [122].
- **Zavedenie povinnosti „due diligence“ pri vývoji a distribúcii AI systémov:** Zákon by mohol výslovne požadovať, aby všetky subjekty zapojené do vývoja a nasadzovania AI technológií vykonávali tzv. „due diligence“ – teda dôkladné preverenie a hodnotenie systémov ešte pred ich uvedením do praxe. Due diligence (náležitá starostlivosť o umelú inteligenciu) je proces systematického hodnotenia AI systémov z hľadiska technickej bezpečnosti, etiky, právnych noriem a obchodnej zodpovednosti. Zahŕňa identifikáciu potenciálnych rizík, overenie schopností systému a zabezpečenie súladu s platnými reguláciami (napr. GDPR, CCPA, AI Act). Cieľom je predchádzať situáciám, v ktorých by AI technológia mohla fungovať nepredvídateľne, nespoľahlivo alebo by mohla byť zneužitá [123]. Zodpovedné správanie subjektov by malo byť zakotvené v zákone, najmä pri vývoji systémov schopných napodobňovať ľudskú komunikáciu (napr. hlasové klony, chatboty).

- **Zavedenie špeciálneho režimu zodpovednosti v prípadoch neznámeho pôvodcu:**

V prípadoch, kedy nie je možné identifikovať konkrétneho pôvodcu phishingového útoku (napr. ak útočník využil anonymne dostupný open-source AI model alebo sa skryl za zahraničný proxy server), vzniká vážna právna medzera, kedy poškodení koncoví používatelia sa nemajú na koho obrátiť so žiadosťou o odškodnenie. Na preklenutie tejto medzery by bolo, ako jedno z riešení, možné zaviesť osobitný režim tzv. *no-fault liability* – teda režim zodpovednosti bez preukázania zavinenia. Tento model je už známy z oblastí ako sú dopravné nehody (napr. povinné zmluvné poisťenie) alebo vakcinačné škody, a umožňuje poškodeným získať odškodnenie, aj keď nie je určený priamy vinník alebo zodpovedný subjekt. Tento prístup by nielen chránil obeť AI-poháňaných phishingových útokov, ale zároveň by mohol pomôcť eliminovať „lotériový efekt“ tradičného systému občianskoprávnej zodpovednosti, v ktorom úspešnosť žiadosti o odškodnenie závisí výlučne od možnosti identifikovať a žalovať konkrétny subjekt. Navyše, týmto riešením by sa predišlo prehnanému právnomu riziku pre vývojárov a poskytovateľov, čím by sa zároveň zachoval stimulačný priestor pre technologické inovácie [124].

- **Vytvorenie nezávislých orgánov pre riešenie sporov** Ďalším zaujímavým návrhom na efektívnejšie a odborne podložené riešenie sporov týkajúcich sa zodpovednosti za škody spôsobené umelou inteligenciou je zriadenie špecializovaných inštitúcií alebo arbitrážnych orgánov na úrovni EÚ alebo v členských štátoch. Tieto orgány by mohli disponovať kombináciou technickej expertízy a právnej spôsobilosti potrebnej pre posudzovanie zložitých prípadov, kde bežné súdy môžu postrádať potrebné know-how. Zároveň by mohli fungovať ako nezávislé arbitrážne panely s právomocou vydávať záväzné rozhodnutia, čím by sa zabezpečila rýchlosť, odbornosť a konzistentnosť pri rozhodovaní o prípadoch súvisiacich s AI technológiami. Financovanie ich činnosti by mohlo byť zabezpečené prostredníctvom rozpočtu EÚ, národných fondov alebo príspevkov od relevantných aktérov AI ekosystému.

4.2.2 Povinnosť detekcie a prevencie zneužitia AI

Na minimalizáciu rizík spojených s AI-generovanými phishingovými útokmi je nevyhnutné zaviesť právne povinnosti pre poskytovateľov AI nástrojov. Tieto povinnosti by mali zabezpečiť proaktívne opatrenia na detekciu a prevenciu zneužitia, čím by sa znížila pravdepodobnosť, že takéto technológie budú využité na kriminálne účely. Navrhujeme preto súbor možných legislatívnych opatrení zameraných na zvýšenie bezpečnosti a transparentnosti pri vývoji a používaní AI systémov:

- **Zavedenie mechanizmov na detekciu a prevenciu zneužitia:** Zákonná povinnosť pre poskytovateľov AI nástrojov môže zahŕňať implementáciu technických mechanizmov, ktoré aktívne detegujú a bránia zneužitiu napríklad na phishingové útoky. Veľké jazykové modely by v takom prípade museli byť vybavené filtermi na identifikáciu požiadaviek na generovanie podvodných e-mailov, zatiaľ čo generátory hlasu by podliehali obmedzeniam na imitáciu hlasov konkrétnych osôb bez ich súhlasu. Takéto mechanizmy by mohli fungovať napríklad prostredníctvom analýzy vstupných požiadaviek na odhalenie manipulatívnych fráz alebo pomocou algoritmov na rozpoznávanie pokusov o hlasové klonovanie. Ich implementácia by si vyžadovala integráciu príslušných filtrov do AI systémov ešte pred ich uvedením na trh [96].
- **Implementácia systémov na logovanie a auditovanie požiadaviek:** Poskytovatelia AI systémov môžu byť zaviazaní zaviesť povinné systémy na logovanie a auditovanie požiadaviek, ktoré by mohli naznačovať pokusy o zneužitie, pričom by sa zároveň zabezpečila ochrana súkromia používateľov v súlade s GDPR. Logy by v tomto kontexte mohli zaznamenávať pokusy o generovanie obsahu s podozrivými vzorcami, ako sú napríklad požiadavky na vytvorenie e-mailov s manipulatívnym obsahom, a mohli by byť následne auditované s cieľom identifikovať potenciálne hrozby. Tieto systémy by mohli fungovať na báze anonymizovaného logovania, pri ktorom by sa ukladali iba metadáta (napr. typ požiadavky, čas a IP adresa bez osobných údajov), pričom prístup k týmto údajom by mali výhradne autorizované regulačné orgány v prípade vyšetrovania. Takéto riešenie by mohlo prispieť k včasnému odhaleniu zneužitia pri súčasnej ochrane súkromia používateľov [125].

- **Spolupráca s regulačnými orgánmi na oznamovaní a zdieľaní prípadov zneužitia:** Poskytovatelia AI systémov môžu podliehať povinnosti vytvoriť štandardizované postupy na oznamovanie prípadov zneužitia regulačným orgánom a zdieľanie relevantných informácií o identifikovaných hrozbách. Napríklad v prípadoch, keď systém deteguje pokus o vytvorenie deepfake videa určeného na podvodné účely, by poskytovateľ oznamoval takúto udalosť príslušnému orgánu a poskytol by technické údaje o útoku, ako sú vzorce požiadaviek či typ generovaného obsahu. Takto nastavený postup by mohol zlepšiť koordináciu medzi súkromným sektorom a regulačnými autoritami, a tým urýchliť reakciu na nové a vznikajúce hrozby [94].

4.2.3 Povinné označovanie AI-generovaného obsahu

Legislatíva môže stanoviť povinnosť označovať obsah vytvorený prostredníctvom nástrojov umelej inteligencie vo všetkých relevantných formách – textovej, zvukovej aj obrazovej. Označenie by mohlo mať podobu metadát, viditeľného vodoznaku alebo zvukového signálu v prípade hlasových výstupov. Cieľom tohto opatrenia je zabezpečiť transparentnosť pri konzumácii digitálneho obsahu, uľahčiť rozpoznanie umelej generácie a prispieť k obmedzeniu dezinformácií a podvodného správania.

Jasné označenie AI-generovaného obsahu pomôže používateľom rýchlejšie rozpoznať podvodné e-maily, deepfake videá či imitovaný hlas, čím sa výrazne zvýši ich schopnosť obrany voči manipulácii a sociálnemu inžinierstvu. V prostredí zaplavenom falošným a neautentickým obsahom môžu platformy, ktoré jednoznačne označujú syntetický obsah, prispieť k obnove dôvery verejnosti v online komunikáciu. Zároveň by legislatívna povinnosť označovania vytvorila tlak na vývoj technických štandardov a nástrojov pre spoľahlivú detekciu a overovanie pôvodu digitálneho obsahu.

Podobné opatrenia sa už objavujú v medzinárodných iniciatívach. Napríklad *"G7 Hiroshima AI Process"* vyzýva na označovanie syntetického obsahu ako súčasť zásad zodpovedného vývoja AI [126]. Rovnako Európsky AI Act vo svojej finálnej verzii obsahuje povinnosť označovať syntetický obsah [4]. Podobne aj v USA sa v rámci *"Executive Order on AI"* kladie dôraz na rozpoznateľnosť AI výstupov [127]. Tieto kroky ukazujú rastúci globálny konsenzus v tom, že transparentnosť v oblasti AI obsahu je nevyhnutná pre udržanie bezpečného a dôveryhodného digitálneho priestoru.

4.2.4 Posilnenie kybernetickej legislatívy

Dynamický rozvoj AI, najmä generatívnych modelov, si vyžaduje pravidelnú aktualizáciu existujúcej kybernetickej legislatívy. Právny rámec v oblasti kybernetickej bezpečnosti by mal zohľadniť nové typy rizík a útokov, ktoré AI prináša, a zároveň poskytnúť nástroje na ich efektívne vyšetrenie a postihovanie. Je potrebné rozšíriť právnu definíciu kybernetického útoku tak, aby zahŕňala aj aktivity sprostredkované alebo podporované umelou inteligenciou. Súčasné právne úpravy často nepočítajú s možnosťou, že útočník môže využiť AI model na generovanie personalizovaných phishingových správ, napodobenie hlasu vedúcej osoby v organizácii alebo vytvorenie deepfake videa na manipuláciu s verejnou mienkou. V tejto súvislosti by sa mohli zaviesť aj prísnejšie sankcie za využívanie AI na podvodné účely, phishing či iné formy sociálneho inžinierstva. Takéto využitie by malo byť považované za kvalifikovanú formu útoku s vyššou spoločenskou nebezpečnosťou, vzhľadom na schopnosť AI automatizovať a personalizovať podvodné aktivity, čím sa zvyšuje ich rozsah a efektivita pri minimálnom úsilí. Z legislatívneho hľadiska by bolo vhodné aktualizovať aj súvisiace rámce, ako je napríklad smernica NIS2 (o bezpečnosti sietí a informačných systémov), aby zohľadňovali AI ako novú hrozbu v oblasti kybernetickej bezpečnosti [93].

4.2.5 Medzinárodná spolupráca a harmonizácia

Vzhľadom na globálny charakter umelej inteligencie a phishingových kampaní, ktoré často prekračujú štátne hranice aj jurisdikcie, je medzinárodná spolupráca a právna harmonizácia nevyhnutná pre efektívne zvládanie týchto hrozieb. AI modely, ktoré sú schopné generovať realistický podvodný obsah, môžu byť vyvíjané v jednej krajine, nasadzované v inej a zneužívané v úplne inom právnom prostredí. Preto je potrebné koordinovať nielen technické opatrenia, ale aj legislatívne nástroje. Navrhované riešenia, ktoré môžu posilniť cezhraničný boj proti zneužívaniu AI:

- **Harmonizácia legislatívy medzi krajinami:** Harmonizácia legislatívy medzi členskými štátmi EÚ a ďalšími krajinami predstavuje dôležitý krok smerom k vytvoreniu minimálnych spoločných štandardov pre zodpovedné využívanie umelej inteligencie a prijímanie opatrení proti jej zneužívaniu. Európska únia už podnikla významné kroky týmto smerom prostredníctvom prijatia AI Act, ktorý môže slúžiť ako referenčný model pre iné jurisdikcie [75].

- **Spolupráca na vytváraní spoločných štandardov:** Spolupráca na tvorbe spoločných etických, bezpečnostných a technických štandardov pre vývoj a nasadzovanie AI systémov predstavuje kľúčový prvok medzinárodnej regulácie umelej inteligencie. Tieto štandardy vznikajú v rámci viacerých globálnych iniciatív, ako sú napríklad OECD AI Principles, ktoré podporujú bezpečné, transparentné a zodpovedné využívanie AI [5].
- **Zabezpečiť rýchlu výmenu informácií:** Treťou prioritou by malo byť zabezpečenie rýchlej cezhraničnej výmeny informácií o incidentoch a zneužití AI, a to medzi štátmi, regulačnými orgánmi, bezpečnostnými jednotkami a technologickými firmami. Táto výmena by mala byť formálne upravená a technicky podporená, napríklad prostredníctvom spoločných hlásiacich protokolov, bezpečných kanálov a spoločných databáz. Modelom môže byť napríklad spolupráca v oblasti kybernetickej bezpečnosti v rámci siete CSIRTs Network a európskej agentúry ENISA, ktorá už podporuje zdieľanie informácií o incidentoch súvisiacich s AI a kybernetickými hrozbami [128].

Úspešná ochrana pred AI-generovaným phishingom preto nesmie byť len otázkou národných regulácií, ale musí sa opierať o medzinárodný právny rámec, spoločné štandardy a operatívnu spoluprácu všetkých relevantných aktérov.

4.3 Vzdelávanie a zvyšovanie povedomia

Zvládanie hrozieb spojených s umelou inteligenciou a najmä AI-generovaných phishingových útokov, nie je možné dosiahnuť len technickými a legislatívnymi opatreniami. Rovnako dôležitá je aj informovanosť, vzdelávanie a budovanie schopností jednotlivcov, organizácií a spoločnosti ako celku. Mnohé z najúspešnejších phishingových kampaní zneužívajú práve nedostatok digitálnej gramotnosti, slabú orientáciu v súčasných technológiách a nízku schopnosť používateľov odhaliť manipuláciu.

4.3.1 Povinné školenia pre zamestnancov a zraniteľné skupiny

Prvým z navrhovaných opatrení v rámci vzdelávania, je zaviesť pravidelné školenia zamerané na rozpoznávanie phishingových hrozieb podporených umelou inteligenciou. Týka sa to najmä zamestnancov verejnej správy, kritickej infraštruktúry, školstva a zdravotníctva, ako aj súkromných firiem. Školenia by mali obsahovať praktické príklady a interaktívne simulácie útokov. Tieto školenia by zároveň mali zahŕňať aj vysvetlenie nových rizík spojených

s technológiami ako sú generatívne jazykové modely (napr. ChatGPT), hlasové klony (napr. ElevenLabs) alebo vizuálne manipulácie (napr. deepfake). Dôraz by mal byť kladený na zručnosti ako kritická analýza obsahu, overovanie zdrojov, pochopenie technických náznakov manipulácie a nahlásovanie incidentov.

Podľa správy spoločnosti IBM (Cost of a Data Breach Report 2023) je až 95 % kybernetických incidentov spôsobených ľudskou chybou. Najčastejšie ide o nepozornosť, nevedomosť alebo nedostatočné povedomie o hrozbách. Medzi najčastejšie chyby patrí kliknutie na škodlivý odkaz v podvodnom e-maile, neoverenie identity volajúceho počas telefonátu, či nevedomé zdieľanie citlivých údajov s falošným kontaktom. Zároveň tá istá štúdia ukazuje, že organizácie, ktoré pravidelne realizujú školenia kybernetickej bezpečnosti, zaznamenávajú o približne 50 % nižšiu mieru úspešnosti phishingových útokov [129]. To znamená, že vzdelávanie a zvyšovanie povedomia zamestnancov výrazne zvyšuje ich schopnosť rozpoznať manipuláciu a znížiť riziko nechceného poskytnutia prístupu útočníkovi. Školenia tak predstavujú jeden z najefektívnejších a zároveň najdostupnejších nástrojov prevencie, a to aj v prípade pokročilých útokov, ktoré využívajú umelú inteligenciu na generovanie vierohodného obsahu.

4.3.2 Kampane na zvýšenie povedomia pre verejnosť

Okrem profesionálnych školení je potrebné realizovať systematické informačné kampane pre širokú verejnosť. Cielené informačné kampane môžu výrazne prispieť k zníženiu miery úspešnosti AI-podvodov tým, že pomáhajú ľuďom rozpoznať manipulovaný obsah, zvyšujú odolnosť voči psychologickým taktikám podvodníkov a podporujú správne reakcie na podozrivé situácie, čím obmedzujú šírenie útokov. Úspešné kampane by mali byť jednoduché, vizuálne pútavé, a prispôbované rôznym vekovým, jazykovým a sociálnym skupinám. Ich obsah by mal zahŕňať:

- **Vysvetlenie základných pojmov** – ako napríklad deepfake, AI-generovaný obsah a hlasový klon.
- **Upozornenie na typické znaky manipulácie** – napríklad oneskorené reakcie v hovore, neautentický jazyk, nepresnosti vo videu či absencia mimiky.
- **Učiť ľudí, ako si overiť autenticitu a pravosť obsahu** – napríklad tým, že si porovnávajú informácie z viacerých dôveryhodných zdrojov (napr. oficiálne weby, známe

médiá), overia si hlas alebo video priamym kontaktovaním danej osoby (napr. telefonicky, osobne), alebo použijú jednoduché online nástroje na overenie obrázkov a videí (napr. overenie dátumu zverejnenia, skontrolovanie metadát, nástroje typu „*Is this image AI-generated?*“).

- **Pokyny na bezpečnú komunikáciu a reakciu na podozrivý obsah** – teda vysvetliť, ako sa vyhýbať zdieľaniu citlivých údajov, kedy byť obozretný pri nečakaných správach alebo hovoroch, a kam sa obrátiť v prípade podozrenia.

Tieto kampane plnia nielen preventívnu, ale aj psychologickú a komunitnú funkciu, pretože posilňujú dôveru v digitálne prostredie a podporujú nahlasovanie incidentov. Mali by byť dlhodobé, financované z verejných aj súkromných zdrojov a koordinované s médiami a školami pre čo najširší zásah. Úspešné príklady takýchto kampaní možno nájsť v Nemecku (BKA) či Kanade (Get Cyber Safe), kde sa prostredníctvom sociálnych médií, videí, podcastov a workshopov cielene oslovujú rôzne vekové skupiny [130]. V EÚ podporuje podobné aktivity agentúra ENISA prostredníctvom kampane European Cybersecurity Month, ktorá ponúka vzdelávacie nástroje a podujatia vo všetkých členských štátoch [128].

4.3.3 Podpora digitálnej gramotnosti v školách

Dlhodobé riešenie spočíva v systematickom budovaní digitálnej gramotnosti už na úrovni základných a stredných škôl. Obsah vzdelávacích programov by mal byť aktualizovaný tak, aby zahŕňal aj nové výzvy, ktoré prináša rozmach AI, najmä v súvislosti s dezinformáciami, manipuláciou obsahu a ochranou pred podvodnými útokmi. V rámci školského vzdelávania by preto mali byť rozvíjané najmä tieto oblasti:

- pochopenie fungovania AI nástrojov,
- schopnosť identifikovať manipulovaný obsah,
- zodpovedné digitálne správanie a etiku online komunikácie.

Rozšírenie digitálnej a mediálnej gramotnosti je v súlade s odporúčaniami Európskej komisie, ktorá v rámci iniciatívy Digital Education Action Plan vyzýva členské štáty, aby posilnili výučbu kritického myslenia, mediálnej gramotnosti a porozumenia technológiám umelej

inteligencie na všetkých úrovniach vzdelávania. Podobné prístupy sa implementujú aj v krajinách ako Fínsko, Estónsko alebo Holandsko, kde sú témy digitálnej bezpečnosti, AI a etiky súčasťou kurikula už od základných škôl [131]. Investícia do digitálnej gramotnosti je teda strategickým krokom nielen pre obranu proti AI-generovanému phishingu, ale aj pre dlhodobú odolnosť spoločnosti voči manipulácii a informačným hrozbám.

4.4 Etické zásady pre vývoj a využitie AI

Existujúce etické rámce pre vývoj umelej inteligencie, ako OECD AI Principles [5], UNESCO Recommendation on the Ethics of AI [94], či zásady zakotvené v AI Act [4], stanovujú všeobecné hodnoty ako transparentnosť, spravodlivosť, rešpektovanie ľudských práv a zodpovednosť. V praxi však tieto rámce často zostávajú príliš všeobecné alebo nezohľadňujú dynamiku a riziká vyplývajúce z niektorých konkrétnych technológií a to predovšetkým generatívnych AI systémov s vysokým manipulačným potenciálom (napr. deepfake, hlasové klony, realistická syntéza textu).

Preto navrhujeme doplniť etické smerovanie vývoja AI o nové, konkrétnejšie zásady, ktoré by mohli byť súčasťou buď doplnkových odporúčaní pre vývojárov, alebo regulačných rámcov. Tieto zásady by sa mali opierať o princíp „zodpovedného dizajnu“ – t. j. aktívne predchádzanie zneužitiu už vo fáze návrhu technológie.

4.4.1 Zásada nezneužiteľnosti pri navrhovaní

Vývojári generatívnych AI systémov by mali aktívne predvídať potenciálne formy zneužitia ich technológie a zavádzať preventívne mechanizmy už počas vývoja. V prípade nástrojov schopných generovať realistický hlas, tvár alebo komunikáciu to znamená:

- zavedenie technických obmedzení na imitáciu verejne známych osôb bez ich výslovného súhlasu,
- detekciu citlivých alebo rizikových požiadaviek (napr. požiadavky na generovanie podvodných e-mailov),
- a zabudovanie kontrol (napr. etických „guardrails“) do modelov pred ich nasadením.

Táto zásada rozširuje existujúci princíp „*do no harm*“ o povinnosť konštruktívneho plánovania proti zneužitiu. Jej dodržiavanie by mohlo byť hodnotené etickými auditmi. Táto zásada rozširuje existujúci princíp „*do no harm*“ – teda záväzok vývojára vytvárať technológie, ktoré nepoškodzujú jednotlivcov ani spoločnosť. Tento princíp je zakotvený napríklad v OECD AI Principles [5] a UNESCO Recommendation on the Ethics of AI [94], kde sa zdôrazňuje potreba navrhovať systémy tak, aby minimalizovali riziko ujmy, diskriminácie alebo zneužitia. V kontexte generatívnej AI by však tento princíp mal byť rozšírený o aktívnu povinnosť predchádzať zneužitiu už vo fáze návrhu technológie. Jeho dodržiavanie by mohlo byť posudzované prostredníctvom etických auditov a hodnotení rizikovosti modelu.

4.4.2 Povinnosť implementovať detekovateľnosť výstupu

Vývojári systémov, ktoré generujú realistický multimediálny obsah (text, video, zvuk), by mali povinne zabezpečiť, že ich výstupy budú detekovateľné. Táto schopnosť rozpoznania môže byť zabezpečená napríklad prostredníctvom digitálnej vodotlače, identifikačných značiek v metadátach alebo sprievodných overovacích nástrojov, ktoré umožnia tretím stranám (napr. platformám, médiám alebo orgánom) overiť pôvod a umelý charakter daného obsahu.

Na rozdiel od princípu transparentnosti, ktorý sa týka toho, ako AI systém funguje (napr. na akých dátach bol trénovaný a aké riziká predstavuje), zásada detekovateľnosti sa zameriava priamo na výstup AI systému – teda na to, či dokážeme dodatočne zistiť, že daný obsah bol vytvorený umelou inteligenciou. Táto schopnosť je kľúčová najmä pri obrane proti podvodným praktikám, ako sú deepfake videá, klonované hlasy alebo falošné textové komunikácie, ktoré sa zámerne vydávajú za skutočné. V súčasnosti takýto štandard nie je povinný, no viaceré iniciatívy sa už zameriavajú na hľadanie riešení, ako spoľahlivo overiť zdroj a históriu digitálneho materiálu. Napríklad iniciatíva *Coalition for Content Provenance and Authenticity* (C2PA) pod vedením spoločnosti Adobe, v súčasnosti pracuje na zavedení dôveryhodných spôsobov označovania digitálneho obsahu priamo pri jeho vytvorení [132]. Zavedenie navrhutej povinnosti by mohlo významne prispieť k ochrane používateľov a obmedziť šírenie manipulatívneho a neovereného obsahu v digitálnom priestore.

4.4.3 Etický limit personalizácie

AI systémy navrhnuté na personalizovaný výstup by nemali umožniť napodobenie štýlu alebo identity konkrétnej osoby, ak to neslúži oprávnenému účelu a chýba na to výslovný súhlas. V kontexte phishingu je práve personalizácia (jazyk, tón, zmienky o reálnych faktoch) často dôvodom úspechu útoku. Preto by malo byť súčasťou etických zásad obmedzenie generovania personalizovaného výstupu bez overenia legitímnosti požiadavky.

4.4.4 Zriadenie etických komisií pre AI v kyberbezpečnosti

Ako ďalšie opatrenie navrhujeme vytvoriť nezávislé etické komisie alebo poradné orgány, ktoré by pôsobili na národnej aj nadnárodnej úrovni. Ich hlavným cieľom by bolo posudzovať využívanie AI z pohľadu etických rizík, ochrany práv jednotlivcov a zneužívateľnosti technológií, najmä v oblastiach, kde môže AI priamo ovplyvniť dôveru, bezpečnosť alebo demokratické inštitúcie. Takéto komisie by mohli plniť najmä nasledujúce funkcie:

- **Posudzovanie rizikových AI aplikácií z pohľadu etiky a možnosti zneužitia** – ide o predbežné hodnotenie technológií ešte pred ich nasadením, vrátane analýzy možných dopadov na jednotlivcov, spoločnosť či informačný priestor.
- **Vydávanie odporúčaní a stanovísk k využitiu AI v citlivých sektoroch**, ako sú obrana, bezpečnostné zložky, vzdelávanie, zdravotníctvo alebo verejná správa. Tieto stanoviská by mohli slúžiť ako podklad pre reguláciu, obstarávanie a vývoj AI riešení.
- **Priebežné monitorovanie vývoja, nasadzovania a spätnej väzby na AI systémy vo verejnom priestore**, napríklad sledovaním ich vplyvu na informačnú integritu, transparentnosť či ochranu pred podvodmi a manipuláciou.

Zriadenie takýchto orgánov by mohlo byť priamo previazané s certifikáciou AI nástrojov v oblastiach so zvýšeným rizikom zneužitia, ako sú nástroje na generovanie hlasu, vizuálne deepfaky, autonómne rozhodovanie o používateľoch alebo AI systémy využívané v bezpečnostnej infraštruktúre. Ich existencia by zároveň umožnila zachytiť potenciálne hrozby ešte pred ich masovým nasadením, čím by doplnila technické a legislatívne opatrenia o dôležitú preventívnu vrstvu postavenú na hodnotách a spoločenskom záujme. Podobné orgány už v niektorých krajinách vznikli, napr. AI Ethics Council v Holandsku [133], či Advisory Committee on AI v Kanade [134].

4.4.5 Zásada spätnej zodpovednosti

Etická zodpovednosť vývojára alebo poskytovateľa AI by nemala končiť momentom uvoľnenia modelu alebo nástroja. V prípade, že sa preukáže jeho opakované zneužívanie na podvody, dezinformácie alebo phishing, mal by mať povinnosť:

- spolupracovať na vyšetrovaní,
- poskytnúť technické nástroje na analýzu výstupu,
- a v prípade vážneho ohrozenia prijať nápravné opatrenia (napr. aktualizáciu modelu alebo dočasné stiahnutie služby).

Táto zásada nadväzuje na koncept „*post-release responsibility*“, teda zodpovednosti po uvedení AI nástroja do prevádzky, ktorý sa čoraz častejšie diskutuje najmä v súvislosti s otvorenými a voľne dostupnými AI systémami. Podľa tohto prístupu nesie vývojár alebo poskytovateľ zodpovednosť nielen za dizajn a tréning modelu, ale aj za jeho správanie v reálnom svete, a to najmä ak sa ukáže, že jeho nástroj je opakovane a predvídateľne zneužívaný na škodlivé účely. V súčasnosti sa o tomto prístupe uvažuje napríklad v rámci odporúčaní Partnership on AI a diskusií o bezpečnosti otvorených modelov ako LLaMA, Stable Diffusion či ElevenLabs Voice [135].

5 APLIKÁCIA A VYHODNOTENIE NAVRHOVANÝCH RIEŠENÍ

Navrhované riešenia v oblasti technických, legislatívnych, vzdelávacích a etických opatrení, ktoré boli predstavené v kapitole 4, majú za cieľ komplexne zmierniť hrozby spojené s umelou inteligenciou v kybernetickom priestore, najmä v kontexte phishingových útokov. Táto kapitola sa zameriava na vyhodnotenie efektivity týchto riešení, ich praktickú aplikáciu v konkrétnych scenároch, komparatívne zhodnotenie ich prínosov a identifikáciu možných obmedzení a výziev. Cieľom je posúdiť, do akej miery sú navrhované opatrenia schopné zvýšiť odolnosť jednotlivcov, organizácií a spoločnosti voči AI-generovaným hrozbám a poskytnúť odporúčania na ich implementáciu.

5.1 Teoretické zhodnotenie efektivity návrhov

Navrhované opatrenia v technickej, legislatívnej, vzdelávacej aj etickej oblasti vychádzajú z analýzy súčasných trendov v oblasti generatívnej umelej inteligencie a jej zneužitia v kybernetických útokoch. Využívajú poznatky z výskumov, ktoré potvrdzujú rastúcu schopnosť AI generovať realistický, personalizovaný a manipulatívny obsah – vrátane deepfake videí, syntetických hlasov a textov imitujúcich legitímnu komunikáciu. Každá z týchto oblastí prispieva k celkovej efektivite boja proti AI-generovaným phishingovým hrozbám, avšak ich teoretická účinnosť závisí od ich vzájomnej synergie a schopnosti adresovať špecifické aspekty problému.

Technické opatrenia, ako sú pokročilé detekčné systémy založené na strojovom učení, hlbokom učení a spracovaní prirodzeného jazyka (NLP), majú vysoký potenciál na identifikáciu a prevenciu AI-generovaných útokov. Napríklad modely strojového učenia dokážu rýchlo a presne klasifikovať podozrivé e-maily s presnosťou až 95 %, ako ukázala štúdia ScienceDirect v roku 2024 [96]. Autentifikačné technológie, ako viacfaktorová autentifikácia a FIDO2, podľa Kaspersky znižujú riziko kompromitácie účtov aj pri úspešnom phishingovom útoku, čím poskytujú robustnú obranu [136]. Detekcia deepfake a hlasového klonovania môže zase výrazne obmedziť dopad vishingu a video impersonácie, hoci ich efektivita závisí od neustáleho zlepšovania algoritmov, aby držali krok s vývojom generatívnych technológií. Ich aplikácia vo forme komplexných detekčných rámcov umožňuje rýchlejšiu a presnejšiu identifikáciu podvodného správania aj pri použití nových AI techník.

Legislatívne návrhy vychádzajú z aktuálnych trendov v oblasti medzinárodnej regulácie umelej inteligencie, ako sú AI Act, odporúčania OECD, UNESCO či iniciatíva G7 AI Process. Zameriavajú sa na najzásadnejšie medzery v oblasti právnej zodpovednosti, prevencie zneužitia technológií a cezhraničnej spolupráce. Opatrenia ako povinnosť detekcie a prevencie zneužitia AI, či jasné určenie právnej zodpovednosti vývojárov a poskytovateľov, môžu teoreticky významne posilniť bezpečnostné štandardy v praxi. Najmä, ak sú doplnené o nástroje, ktoré motivujú k proaktívnemu konaniu (napr. etické hodnotenia či zásada *due diligence* pri vývoji). Zavedenie špeciálnych režimov ako zodpovednosť bez zavinenia (no-fault liability) by mohlo zabezpečiť ochranu obetí aj v prípadoch, kde nie je možné jednoznačne určiť páchatel'a. Tieto alebo podobné prístupy navrhujú aj odborné platformy a expertné skupiny [96, 124]. Dôležitou súčasťou legislatívneho rámca je aj medzinárodná harmonizácia pravidiel, ktorá posilňuje schopnosť efektívne reagovať na globálne phishingové kampane a ďalšie cezhraničné kybernetické hrozby.

Vzdelávacie iniciatívy predstavujú jeden z najefektívnejších nástrojov na zníženie zraniteľnosti jednotlivcov a organizácií voči AI-generovaným phishingovým útokom. Podľa správy IBM Cost of a Data Breach Report 2023 organizácie, ktoré pravidelne školili zamestnancov, zaznamenali nižšiu mieru úspešnosti phishingových útokov v porovnaní s tými, ktoré školenie zanedbávali. Report zároveň ukazuje, že tieto organizácie mali aj nižšie finančné škody – priemerné náklady na narušenie dát boli nižšie až o 1,5 milióna USD, čo predstavuje úsporu na úrovni 33,9 % oproti globálnemu priemeru [129]. Výrazný efekt vzdelávania potvrdzuje aj Phishing by Industry Benchmarking Report z roku 2022, ktorý analyzoval viac než 54 miliónov phishingových simulácií v 55 675 organizáciách. Podľa tohto reportu organizácie, ktoré absolvovali ročné školenie a simulácie útokov, znížili svoju náchylnosť na phishing v priemere z 32,4 % na 5 %, čo predstavuje zlepšenie odolnosti o 85 % [137].

Okrem profesionálnych školení pre zamestnancov je nevyhnutné posilniť digitálnu gramotnosť už na úrovni základných a stredných škôl, ako aj realizovať dlhodobé informačné kampane pre širokú verejnosť. Cieľom týchto iniciatív by malo byť najmä budovanie kritického myslenia, schopnosti rozpoznať manipuláciu a odolnosti voči psychologickým taktikám podvodníkov. V kontexte rastúceho využívania AI na generovanie presvedčivého, personalizovaného a ťažko odhaliteľného podvodného obsahu ide o kľúčovú súčasť obrany, ktorá dopĺňa technické a legislatívne opatrenia o dôležitý ľudský faktor.

Etické zásady, ako zásada nezneužitelnosti pri navrhovaní a povinnosť detekovateľnosti výstupu, teoreticky minimalizujú riziko zneužitia AI už vo fáze vývoja a reagujú na konkrétne riziká generatívnych systémov. Napríklad zavedenie digitálnej vodotlače pre AI-generovaný obsah by mohlo výrazne zvýšiť jeho rozpoznateľnosť a prispieť k zníženiu šírenia dezinformácií a podvodov, ako navrhuje C2Pa v aktualizovanom štandarde [132]. Etické komisie by zasa mohli poskytovať preventívny dohľad a odporúčania, čím by sa predišlo etickým zlyhaniam. Cieľom etických princípov nie je obmedziť inovácie, ale nasmerovať vývoj technológií tak, aby slúžili spoločnosti a minimalizovali možnosť zneužitia.

5.2 Praktická aplikácia v konkrétnych scenároch

Táto podkapitola aplikuje navrhnuté riešenia do realistických scenárov. Demonštruje, ako by technické, legislatívne, vzdelávacie a etické opatrenia mohli fungovať v praxi, a aký by mohol byť ich konkrétny prínos pri zvládaní rizík spojených s AI-generovaným phishingom. Každý scenár ilustruje možné nasadenie opatrení v konkrétnych prostrediach, ako sú podnikové siete, legislatívne rámce, verejná komunikácia či vývojárske prostredie.

5.2.1 Scenár aplikácie technických opatrení

Implementácia detekcie deepfake a hlasového klonovania

V rámci prevencie voči AI phishingu odporúčame integrovať detekčné mechanizmy pre syntetické video a hlas priamo do komunikačných a informačných platforiem. Ich aplikácia má potenciál znížiť mieru úspešnosti podvodov, rýchlejšie identifikovať hrozby a automaticky varovať používateľov pred manipulovaným obsahom. Navrhujeme nasledovné spôsoby implementácie do komunikačných platforiem ako sú:

- **Microsoft Teams, Zoom, Webex** – pomocou rozšírení cez API alebo SDK by mohli tieto platformy v reálnom čase analyzovať obraz a zvuk počas hovorov. Pri zistení podozrivých znakov, ako sú vizuálne artefakty typické pre deepfake (neprirodzený pohyb pier, nezrovnalosti v osvetlení) alebo hlasové anomálie (konzistentné frekvenčné spektrum bez prirodzených odchýlok), by systém mohol automaticky upozorniť účastníkov hovoru, že ide o potenciálne syntetický vstup. To môže zabrániť tomu, aby zamestnanci reagovali na pokyny od podvodníka, ktorý sa vydáva za inú osobu.

- **E-mailové klienty a sociálne siete** – integrácia analytických modulov pre multimédiá (napr. cez preprocessing video/audio príloh) umožní automatické označovanie podozrivého obsahu ako „pravdepodobne syntetický“. Takéto označenie zvýši povedomie používateľ a ešte pred samotným prehratím súboru a podporí opatrnosť voči podvodným správam.
- **Podnikové VOIP systémy** – implementáciou detekcie manipulovaného hlasu na úrovni sietejovej vrstvy je možné už počas prebiehajúceho hovoru identifikovať hlas klonovaný umelou inteligenciou. Systém by mohol napríklad pozastaviť hovor a požiadať o dodatočné overenie identity (napr. cez známu biometrickú frázu), čím sa zníži pravdepodobnosť úspešného vishingového útoku.

Okrem toho, zavedenie digitálneho vodotlačenia (watermarking) pre video a audio výstupy generované AI by mohlo slúžiť ako proaktívna obrana. Tento mechanizmus by umožnil rýchlu detekciu syntetického obsahu aj bez nutnosti rozsiahlej analýzy. Takéto označovanie by malo byť implementované priamo vývojármi generatívnych nástrojov a podporené regulačnými rámcami. Príkladom je iniciatíva spoločnosti Meta, ktorá plánuje označovať všetok AI-generovaný obsah na platformách ako Instagram a Facebook [138]. Takýto prístup by významne posilnil transparentnosť a umožnil rýchlu identifikáciu podvodných materiálov aj bežným používateľom.

5.2.2 Scenár uplatnenia legislatívnych opatrení

Rozdelenie zodpovednosti medzi aktérov AI ekosystému

Legislatíva by mala explicitne definovať, ktoré subjekty nesú právnu zodpovednosť v prípade, že AI systém spôsobí škodu – úmyselne alebo v dôsledku zlyhania. Tieto pravidlá musia rozlišovať medzi rôznymi typmi aktérov, ako sú vývojári, poskytovatelia, prevádzkovatelia a koncoví používatelia:

- **Útočník alebo škodlivý používateľ** – ak osoba vedome zneužije AI nástroj na vykonanie phishingového útoku (napr. pomocou hlasového klonovania alebo generovania deepfake videa), mala by niesť plnú trestnoprávnu a občianskoprávnu zodpovednosť za spôsobené škody – rovnako ako pri inom kybernetickom útoku.

- **Vývojář AI systému** – mal by byť zodpovedný v prípade, že úmyselne vytvorí alebo distribuuje systém bez primeraných ochranných mechanizmov (napr. modely schopné imitovať identitu bez obmedzení), alebo ak ignoruje známe riziká spojené so zneužitím svojho modelu. Zodpovednosť by sa mala posudzovať najmä na základe miery zanedbania „due diligence“ pri vývoji.
- **Poskytovateľ AI služby** (napr. prevádzkovateľ platformy umožňujúcej generovanie výstupov) – by mal niest' spoluzodpovednosť v prípade, že nedostatočne kontroluje výstupy alebo neimplementuje opatrenia ako monitoring, detekcia zneužitia a blokovanie nelegitímneho obsahu.
- **Prevádzkovateľ AI systému v organizácii** – napr. firma, ktorá AI integruje do svojej komunikácie alebo rozhodovania, by mala niest' zodpovednosť za dôsledky nesprávnej konfigurácie, absencie kontroly alebo používania AI bez odborného dohľadu.

Takto jasne rozdelená zodpovednosť by zvýšila transparentnosť v prípadoch incidentov, podporila preventívne opatrenia a zároveň umožnila efektívnejšie riešiť vzniknuté škody. Legislatíva by mala tieto rámce nielen formálne stanoviť, ale aj prakticky vymáhať – napríklad prostredníctvom regulačných pokút, náhrady škody alebo obmedzenia činnosti pri opakovanom porušení.

Uplatnenie zásady „*due diligence*“

Zavedenie tejto povinnosti by malo najmä tieto konkrétne dôsledky:

- Vývojári a poskytovatelia by boli nútení už vo fáze návrhu identifikovať možné spôsoby zneužitia (napr. generovanie podvodných e-mailov alebo hlasov), a zaviesť technické a organizačné opatrenia na ich elimináciu.
- Systémy schopné napodobňovať ľudskú komunikáciu (napr. hlasové klony, chatboty) by mohli byť pred nasadením podrobené testovaniu a museli by obsahovať funkcie ako detekcia manipulatívneho obsahu, logovanie interakcií, prípadne automatické upozornenia pri podozrivých aktivitách.
- Zodpovednosť by bola viazaná aj na kvalitu a transparentnosť tréningových dát – tzn. že vývojár by musel zdokumentovať pôvod dát, vylúčiť toxický alebo zavádzajúci obsah, a zabezpečiť súlad so zákonmi o ochrane osobných údajov.

- Ak by pri incidente vyšlo najavo, že subjekt nevykonával riadne due diligence, mohlo by to viesť k sprísneniu sankcií alebo k prenosu zodpovednosti za vzniknuté škody.

Z legislatívneho hľadiska by mohla byť povinnosť due diligence súčasťou noriem pre systémy s vysokým rizikom (podľa klasifikácie AI Act), pričom by sa mohla aplikovať aj na stredne rizikové systémy v závislosti od kontextu použitia.

Aplikácia zodpovednosti bez zavinenia („no-fault liability“)

Jeho implementácia v oblasti AI by mala nasledovné kľúčové prvky:

- **Kompenzačný fond** – vytvorenie fondu financovaného napríklad z povinných príspevkov AI poskytovateľov, prevádzkovateľov platforiem alebo formou digitálnej dane z prevádzky vysokorizikových modelov. Tento fond by slúžil na úhradu škôd spôsobených AI systémami v prípadoch, keď nie je možné identifikovať konkrétneho páchatel'a.
- **Nezávislý správca systému** – tretia strana (napr. regulačný úrad alebo poverená nezisková inštitúcia) by mala dohliadať na vyhodnocovanie žiadostí, schvaľovanie odškodnení a správu fondu.
- **Stanovenie štandardov oprávnenosti** – bolo by potrebné jasne definovať kritériá, za akých možno žiadať náhradu (napr. dôkaz, že škoda vznikla v dôsledku AI generovaného obsahu; zachovanie dôkazov o komunikácii). Rovnako aj vedecky overené štandardy bezpečného vývoja a prevádzky, ktorých dodržiavanie by vývojárov chránilo pred súdnymi spormi.
- **Stanovenie výšky odškodnenia** – legislatíva by mala upraviť rámce pre výšku odškodnenia, ktoré by boli primerané typu a rozsahu škody, podobne ako pri pracovných úrazoch alebo environmentálnych haváriách.

5.2.3 Scenár vzdelávacích iniciatív

Na rozdiel od technických alebo legislatívnych zásahov, ktoré si vyžadujú dlhší čas na implementáciu, vzdelávacie iniciatívy môžu byť nasadené relatívne rýchlo a priamo ovplyvňujú správanie a rozhodovanie používateľov, ktorí sú častým cieľom útokov. Efektívna aplikácia týchto opatrení môže prebiehať v niekoľkých úrovniach:

- **Povinných školení zaměstnanců** – najmä v sektore verejnej správy, zdravotníctva, školstva a kritickej infraštruktúry, kde je riziko sociálneho inžinierstva vysoké. Tieto školenia by mali zahŕňať identifikáciu phishingových e-mailov, hlasových podvodov a pochopenie rizík spojených s deepfake technológiami. Tiež by mala zahŕňať postupy, ako reagovať na podozrivú komunikáciu a ako incident nahlásiť.
- **Simulácie útokov a praktické cvičenia** – ide o metódu, ktorá overuje schopnosť zamestnancov reagovať v reálnych podmienkach. Pomocou nástrojov ako KnowBe4 alebo Cofense je možné vytvárať falošné phishingové kampane, ktoré testujú pozornosť, rozhodovanie a náchylnosť zamestnancov v praxi.
- **Cielené informačné kampane pre verejnosť** – sú určené širokému spektru obyvateľstva, vrátane detí, seniorov či digitálne menej gramotných osôb. Obsah kampaní by mal byť jednoducho zrozumiteľný, vizuálne atraktívny a zameraný na vysvetlenie pojmov. Tiež by mal byť doplnený praktickými radami, ako si overiť autenticitu obsahu a ako sa správať pri podozrivých situáciách.
- **Podpora digitálnej gramotnosti v školách** – predstavuje dlhodobé, ale strategicky klúčové riešenie. Zavedenie výučby o umelej inteligencii, kritickom myslení, rozpoznávaní manipulácie a etike v digitálnom priestore do školských osnov zvyšuje odolnosť budúcich generácií voči dezinformáciám a AI-podvodom.

Tieto opatrenia majú preukázateľný vplyv na zníženie úspešnosti phishingových útokov a zvýšenie odolnosti spoločnosti. Vzdelávanie preto nemožno vnímať ako doplnkový prvok, ale ako základný pilier odolnosti voči hrozbám modernej kybernetickej éry.

5.2.4 Scenár uplatnenia etických zásad

Etické zásady, ktoré boli navrhnuté v predchádzajúcej kapitole (zásada nezneužiteľnosti, detekovateľnosti, obmedzenia personalizácie, spätná zodpovednosť a etické komisie), predstavujú nadstavbu nad legislatívou a technickými opatreniami. Ich účelom nie je len formálna deklarácia hodnôt, ale praktické ukotvenie spoločenskej zodpovednosti vývojárov, poskytovateľov aj používateľov AI technológií. Praktickým príkladom takejto implementácie je návrh digitálnej vodotlače pre výstupy generované AI. Takéto označovanie umožňuje nielen lepšiu detekciu syntetického obsahu, ale aj spätné sledovanie jeho pôvodu.

Zároveň, návrh na zriadenie nezávislých etických komisií umožňuje priebežný dohľad nad vývojom a nasadzovaním rizikových AI systémov. Tieto orgány môžu posudzovať etickú prijateľnosť technológií, poskytovať odporúčania pre regulátorov alebo verejný sektor a včas upozorňovať na hrozby ešte pred ich masovým nasadením. V kontexte rýchleho rozvoja generatívnej AI ide o kľúčový nástroj na predchádzanie zlyhaniam, ktoré technická ani právna regulácia nemusí včas zachytiť.

5.3 Komparatívne zhodnotenie prínosov

Navrhované opatrenia pokrývajú štyri hlavné oblasti – technickú, legislatívnu, vzdelávaciu a etickú. Každá z nich prispieva k riešeniu týchto hrozieb odlišným spôsobom. Ich účinnosť však závisí od miery integrácie, spolupráce aktérov a rýchlosti implementácie. V tejto časti porovnávame prínosy týchto opatrení podľa troch hlavných kritérií:

1. rýchlosť účinku,
2. rozsah dopadu,
3. miera realizovateľnosti v praxi.

Technické opatrenia, ako sú pokročilé detekčné systémy, behaviorálna analýza a autentifikačné technológie, poskytujú najrýchlejší a najpriamejší prínos v podobe okamžitej ochrany pred útokmi generovanými umelou inteligenciou. Ich výhodou je, že po nasadení majú okamžitý dopad a dokážu automatizovať proces detekcie a prevencie. Týmto spôsobom znižujú počet úspešných útokov a zvyšujú mieru zachytenia syntetického obsahu. Hlavné obmedzenia spočívajú v potrebe neustálej aktualizácie, aby sa udržali krok s vývojom útokov, a v technologickej interoperabilite medzi rôznymi platformami.

Legislatívne a regulačné opatrenia vytvárajú dlhodobý rámec, ktorý podporuje zodpovednosť a motivuje k bezpečnejšiemu vývoju AI. Tieto zákony ponúkajú systémové riešenie, ktoré môže zabezpečiť prevenciu, zodpovednosť a aj nápravu škôd. Návrhy ako určenie právnej zodpovednosti, due diligence alebo režim bez zavinenia zvyšujú tlak na vývojárov a poskytovateľov, aby sa vopred počítali s možnosťou zneužitia. Ich účinnosť však závisí od politickej vôle a medzinárodnej spolupráce, pričom implementácia môže byť spomalená byrokraciou a rozdielmi v národných právnych systémoch, čo ich robí z pohľadu realizácie náročnejšími.

Vzdelávacie iniciatívy sú nesmierne dôležité pre dlhodobý rozvoj a majú obrovský vplyv na zraniteľné skupiny, pričom posilňujú odolnosť celej spoločnosti. Údaje z praxe ukazujú, že po zavedení školení a kampaní sa výrazne znížila úspešnosť phishingových útokov. Avšak opatrenia ako školenia, verejné kampane či zvyšovanie digitálnej gramotnosti si vyžadujú neustále financovanie a dlhodobú stratégiu, čo môže spomaliť ich okamžitý dopad. Účinnosť týchto iniciatív tiež závisí od rozsahu a kvality samotných kampaní.

Etické princípy predstavujú preventívnu vrstvu ochrany, ktorá pomáha nastaviť hranice technológií ešte predtým, než sa dostanú na trh. Ich význam sa zvyšuje najmä v oblastiach, kde zákony zaostávajú za technologickým pokrokom. Princípy ako spätná zodpovednosť, detekovateľnosť výstupu a obmedzenie personalizácie umožňujú vývojárom prevziať spoločenskú zodpovednosť. Ich prínos sa najviac prejavuje v dlhodobom budovaní dôvery a znižovaní systémového rizika, pričom úspech závisí od ich formálneho prijatia a dodržiavania.

Z komparatívneho pohľadu možno povedať, že najväčší okamžitý efekt majú technické opatrenia, zatiaľ čo najväčší dlhodobý spoločenský prínos pramení zo vzdelávania a etiky. Legislatíva vytvára nevyhnutnú základňu zodpovednosti a regulácie, ktorá je kľúčová pre udržateľné nasadenie AI. Ich kombinácia umožňuje budovať viacvrstvový obranný systém, ktorý reaguje na bezprostredné hrozby, no zároveň znižuje zraniteľnosť spoločnosti v dlhodobom horizonte.

5.4 Obmedzenia a výzvy

Aj keď navrhované riešenia ukazujú teoretickú aj praktickú účinnosť, ich zavedenie v reálnom svete naráža na viaceré obmedzenia. Identifikácia týchto výziev je dôležitá pre lepšie pochopenie, čo môže brániť ich efektívnej implementácii, a zároveň poskytuje podnety pre ďalší výskum či politické rozhodovanie.

Technické riešenia, ako detekcia deepfake obsahu, behaviorálna analýza či moderné autentifikačné štandardy, sú sľubné, no ich praktická účinnosť závisí od schopnosti neustále sa prispôbovať novým formám útokov. AI modely sa rýchlo vyvíjajú a dokážu obchádzať existujúce obranné systémy. Navyše, implementácia týchto riešení si vyžaduje značné finančné investície, interoperabilitu medzi platformami a prístup k výkonným technológiám, čo môže byť problém najmä pre menšie organizácie.

Legislatívne opatrenia čelia prekážkam najmä pri ich zavádzaní v medzinárodnom kontexte. Návrhy ako povinnosť „*due diligence*“, princíp zodpovednosti bez zavinenia či označovanie AI obsahu narážajú na nedostatok politického konsenzu a obavy z brzdenia inovácií. Príkladom je stiahnutie návrhu AI Liability Directive v EÚ v roku 2025 [101]. Navyše, globálny charakter AI systémov sťažuje ich reguláciu v rámci národných jurisdikcií a vyžaduje medzinárodnú harmonizáciu, ktorá je časovo aj politicky náročná. Ďalšou výzvou je získavanie výslovného súhlasu, a to najmä pri personalizovaní obsahu, napodobňovaní štýlu konkrétnych osôb alebo vytváraní deepfake materiálov. V praxi je často náročné technicky aj právne zabezpečiť, že bol súhlas udelený, že je informovaný a preukázateľný. To vytvára potenciálne právne medzery, v ktorých môže dochádzať k porušovaniu práv bez jasnej možnosti nápravy.

Vzdelávacie iniciatívy si vyžadujú dlhodobú podporu, koordináciu medzi verejným a súkromným sektorom a citlivé prispôsobenie rôznym skupinám obyvateľstva. Ich efektivitu môže znižovať nízka dôvera vo verejné inštitúcie, informačné preťaženie, ale aj slabá schopnosť kriticky vyhodnocovať obsah. Rovnako tak školy nemusia mať dostatočne vyškolený personál alebo zdroje na integráciu tém ako AI, mediálna gramotnosť a digitálna etika. Výzvou je aj oslovenie zraniteľných skupín (napr. seniorov v odľahlých oblastiach) a potreba pravidelne aktualizovať obsah kampaní podľa vývoja hrozieb.

Etické zásady, ktoré boli navrhnuté ako preventívny rámec, narážajú na to, že bez formálneho prijatia a dozoru nemajú právnu silu. Ich dodržiavanie je často dobrovoľné a závislé od kultúry a prístupu konkrétnych organizácií. Navyše, v konkurenčnom prostredí môžu firmy uprednostniť rýchle uvedenie produktov na trh pred dôsledným zohľadňovaním etických rizík. Neexistujú ani jednotné štandardy na etické audity alebo hodnotenie tzv. spätnej zodpovednosti po nasadení systému.

Napokon, koordinácia medzi jednotlivými oblastami – technickou, legislatívnou, vzdelávacou a etickou – je zložitá. Vyžaduje strategické plánovanie, dostatok financií a silné inštitucionálne zázemie. V mnohých štátoch, najmä v menej rozvinutých krajinách, môžu chýbať potrebné kapacity, čo znižuje celkový efekt navrhovaných opatrení.

ZÁVĚR

S rostúcim vplyvom generatívnej umelej inteligencie na techniky kybernetických útokov nadobúda phishing nový rozmer – je rýchlejší, presvedčivejší a dostupnejší. Táto práca na túto výzvu reaguje systematickou analýzou aktuálneho stavu. Teoretická časť podrobne analyzovala vplyv AI na vývoj phishingových útokov a ich meniace sa formy. Identifikovala hlavné typy AI-generovaných hrozieb, ktoré zvyšujú efektivitu sociálneho inžinierstva a obchádzajú tradičné obranné mechanizmy. Medzi ne patria personalizované phishingové e-maily, ktoré využívajú OSINT zdroje na rýchly zber údajov o cieľoch, deepfake technológie umožňujúce vytváranie realistických videí a hlasového klonovania, ako aj vishing, kde útočníci napodobňujú dôveryhodné osoby cez hlasové podvody.

Druhá kapitola priniesla detailný popis technológií stojacich za týmito útokmi, ako sú napríklad veľké jazykové modely a nástroje na tvorbu deepfake obsahu. Tretia kapitola sa zamerala na analýzu právneho rámca upravujúceho využívanie AI v kybernetickom priestore. Boli preskúmané záväzné „hard law“ normy, ako napríklad AI Act prijatý v EÚ v roku 2024, a odporúčacie „soft law“ štandardy, ako sú etické usmernenia OECD, ktorých dodržiavanie je dobrovoľné a závisí od jednotlivých subjektov. Na tieto zistenia nadviazala praktická časť práce, ktorej hlavným prínosom bol návrh a aplikácia komplexných riešení na zvládanie rizík spojených s AI-phishingom. Tieto návrhy boli rozdelené do štyroch hlavných oblastí:

Technické riešenia tvoria základnú obrannú líniu a ich výhodou je možnosť rýchleho nasadenia v praxi. Navrhnuté opatrenia kombinujú pokročilé technológie a prístupy, ktoré spolu vytvárajú robustnú viacvrstvovú ochranu pred sofistikovanými hrozbami. Kľúčovým prvkom je zlepšenie detekčných systémov pomocou techník strojového a hlbokého učenia, ktoré umožňujú automatizovanú analýzu rozsiahlych dát a identifikáciu podozrivých vzorcov. Ďalším opatrením je detekcia deepfake a hlasového klonovania pomocou konvolučných neurónových sietí (CNN), ktoré identifikujú anomálie v mimike, pohyboch pier a frekvenčných charakteristikách hlasu. Na zabezpečenie identity bola navrhnutá aj implementácia technológií viacfaktorovej autentifikácie (MFA), vrátane biometrie či dynamickej analýzy správania používateľov. Tieto opatrenia spoločne umožňujú viacvrstvovú ochranu, ktorá analyzuje nie len obsah e-mailov, ale aj správanie používateľov a možné manipulatívne vzorce útočníkov.

Legislatívne a regulačné opatrenia sa zameriavajú na posilnenie právneho rámca a jasné vymedzenie zodpovednosti aktérov s cieľom minimalizovať riziká zneužitia AI. Prvým opatrením je určenie zodpovednosti jednotlivých subjektov, teda vývojárov, poskytovateľov služieb a používateľov, aby bolo možné v prípade zneužitia jednoznačne identifikovať zodpovednú stranu. Ďalej bola navrhnutá povinnosť detekcie a prevencie zneužitia, implementáciou filtrov na rozpoznávanie podozrivých požiadaviek a systémov na logovanie a auditovanie vstupov. Významným opatrením je aj povinné označovanie AI-generovaného obsahu (napr. digitálnou vodotlačou), ktoré by umožnilo rýchle rozpoznanie synteticky vytvoreného obsahu. Zároveň práca zdôrazňuje potrebu posilniť medzinárodnú spoluprácu a harmonizovať právne normy pre efektívne riešenie globálnych hrozieb.

Vzdelávania a zvyšovania povedomia predstavuje tretiu kľúčovú skupinu navrhovaných riešení. Zdôraznený je význam školení pre zamestnancov, verejných kampaní a podpory digitálnej gramotnosti v školách. Práve informovaní a pripravení používatelia predstavujú kľúčový faktor pri znižovaní úspešnosti AI-podvodov. Údaje z výskumov naznačujú, že kvalitné školenia môžu znížiť mieru úspešnosti phishingových útokov až o desiatky percent.

Etické zásady tvoria poslednú kategóriu navrhnutých opatrení. Tieto princípy sú určené na usmernenie vývoja a používania AI tak, aby minimalizovali jej potenciálne zneužitie. Medzi navrhované princípy patrí zásada nezneužiteľnosti, povinnosť detekovateľnosti výstupov, limit personalizácie, spätná zodpovednosť vývojára a zriadenie nezávislých etických komisií. Tieto opatrenia môžu slúžiť ako doplnenie k technickým a právnym nástrojom, a zároveň smerujú vývoj AI technológií k vyššej zodpovednosti a transparentnosti.

Praktická aplikácia týchto opatrení v modelových scenároch ukázala ich potenciálnu účinnosť, ale aj existujúce obmedzenia. Práca preto zároveň identifikovala viaceré výzvy – od technologických a finančných bariér až po právne neistoty či nedostatočnú koordináciu. Záverečné zhodnotenie ukazuje, že zvládanie týchto hrozieb si vyžaduje komplexný prístup prepájajúci technológie, právne nástroje, vzdelávanie a etické smerovanie. Výsledky práce zdôrazňujú, že ochrana pred zneužitím AI nie je len technickou otázkou, ale výzvou pre celú spoločnosť, ktorá si vyžaduje zodpovednosť, spoluprácu a hodnotové ukotvenie vývoja digitálnych technológií.

SEZNAM POUŽITÉ LITERATURY

- [1] SCHAFER, Nick, 2024. *The golden age of scammers: AI-powered phishing*. Online. Dostupné z: <https://www.mailgun.com/blog/email/ai-phishing/#chapter-3>. [cit. 2025-02-26].
- [2] SHEA, Sharon a KRISHNAN, Ashwin, 2024. *How AI is making phishing attacks more dangerous*. Online. Dostupné z: <https://www.techtarget.com/searchsecurity/tip/Generative-AI-is-making-phishing-attacks-more-dangerous>. [cit. 2025-02-26].
- [3] EUROPOL, 2023. *Internet Organised Crime Threat Assessment (IOCTA) 2023*. Online. Dostupné z: <https://www.europol.europa.eu/publication-events/main-reports/internet-organised-crime-threat-assessment-iocta-2023>. [cit. 2025-05-16].
- [4] *Nařízení Evropského parlamentu a Rady (EU) 2024/1689 ze dne 13. června 2024, kterým se stanoví harmonizovaná pravidla pro umělou inteligenci a mění nařízení (ES) č. 300/2008, (EU) č. 167/2013, (EU) č. 168/2013, (EU) 2018/858, (EU) 2018/1139 a (EU) 2019/2144 a směrnice 2014/90/EU, (EU) 2016/797 a (EU) 2020/1828 (Akt o umělé inteligenci)*, 2024. Online. Dostupné z: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. [cit. 2025-04-10].
- [5] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), 2024. *AI principles*. Online. Dostupné z: <https://www.oecd.org/en/topics/ai-principles.html>. [cit. 2025-04-16].
- [6] ALKHALIL, Z.; CHAMINDA, Hewage; LIQAA, Nawaf a IMTIAZ, Khan, 2021. Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. Online. *Frontiers in Computer Science*. 2021. 3, s. 27–29. Dostupné z: 10.3389/fcomp.2021.563060. Dostupné z: <https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2021.563060>. [cit. 2025-10-01].
- [7] VAYANSKY I., Kumar S., 2018. Phishing – challenges and solutions. Online. *Computer Fraud & Security*. 2018. 2018, s. 15–20. Dostupné z: 10.1016/S1361-3723(18)30007-1. Dostupné z: https://www.researchgate.net/publication/322823383_Phishing_-_challenges_and_solutions. [cit. 2025-05-01].
- [8] BHAVSAR, V.; KADLAK, A. a SHARMA, S., 2018. Study on Phishing Attacks. Online. *International Journal of Computer Applications*. 2018. 182, s. 27–29. Dostupné z: 10.5120/ijca2018918286. Dostupné z: https://www.researchgate.net/publication/329716781_Study_on_Phishing_Attacks. [cit. 2025-05-01].
- [9] WOODS, Emma, 2023. *The Three Stages Of a Phishing Attack - Bait, Hook And Catch*. Online. Dostupné z: <https://blog.usecure.io/three-steps-of-phishing>. [cit. 2025-01-13].
- [10] WOJCICKI, Natasha M., 2022. *What is Phishing?* Online. Dostupné z: <https://cygnet-it.org/what-is-phishing/>. [cit. 2025-01-13].
- [11] EFTIMIE, Sergiu; MOINESCU, Radu a RACUCIU, Ciprian, 01 2022. Spear-Phishing Susceptibility Stemming From Personality Traits. *IEEE Access*. 01 2022. 10, s. 1–1. Dostupné z: 10.1109/ACCESS.2022.3190009.

- [12] GRIMMICK, Robert, 2023. *Phishing Attacks: Types, Prevention, and Examples*. Online. Dostupné z: <https://www.varonis.com/blog/phishing-attacks>. [cit. 2025-01-17].
- [13] LENAERTS-BERGMANS, Bart, 2023. *What is Spear-Phishing? Definition with Examples*. Online. Dostupné z: <https://www.crowdstrike.com/en-us/cybersecurity-101/social-engineering/spear-phishing/>. [cit. 2025-01-17].
- [14] KASPERSKY, 2022. *What is spear phishing? Definition and risks*. Online. Dostupné z: <https://www.kaspersky.com/resource-center/definitions/spear-phishing>. [cit. 2025-01-18].
- [15] KAUR, Harnoor, 2024. *Understanding Man-in-the-Middle Phishing: A Deep Dive into Evilginx*. Online. Dostupné z: <https://bolster.ai/blog/man-in-the-middle-phishing>. [cit. 2025-01-20].
- [16] LANDESBURG, Noah a LURIE, Amy, 2023. *Malware-Based Phishing: How to Deal With It?* Online. Dostupné z: <https://www.au10tix.com/blog/malware-based-phishing/>. [cit. 2025-01-20].
- [17] HICKS, Amanda, 2018. *Angler Phishing: What is it?* Online. Dostupné z: <https://www.clearviewfcu.org/Resources/Learn/Blog/Angler-Phishing-What-is-it>. [cit. 2025-01-21].
- [18] LIŠKOVÁ, Petra, 2023. *Kybernetická bezpečnost: Angler phishing*. Online. Dostupné z: <https://www.faster.cz/novinky/blog/kyberneticka-bezpecnost-angler-phishing-5-5/>. [cit. 2025-01-21].
- [19] CYBERSECURITY GLOSSARY, 2024. *Anatomy of a phishing attack*. Online. Dostupné z: <https://www.verizon.com/about/news/anatomy-of-phishing-attack>. [cit. 2025-02-01].
- [20] PHISHING.ORG, 2024. *What Is Phishing?* Online. Dostupné z: <https://www.phishing.org/what-is-phishing>. [cit. 2025-02-01].
- [21] OLLMANN, Gunter, 2004. *The Phishing Guide: Understanding & Preventing Phishing Attacks*. Online. Dostupné z: <https://www.scribd.com/document/219802442/The-Phishing-Guide-Understanding-Preventing-Phishing-Attacks-IBM-Internet-Security-Systems>. [cit. 2025-02-10].
- [22] PHISHING.ORG, 2024. *History of Phishing*. Online. Dostupné z: <https://www.phishing.org/history-of-phishing>. [cit. 2025-02-10].
- [23] TAYLOR, Craig, 2004. *ILOVEYOU Virus*. Online. Dostupné z: <https://cyberhoot.com/cybrary/iloveyou-virus/>. [cit. 2025-02-10].
- [24] PHISME COFENSE, 2023. *The History of Phishing Attacks*. Online. Dostupné z: <https://cofense.com/knowledge-center/history-of-phishing/>. [cit. 2025-02-10].
- [25] GRAPHUS, 2023. *The History of Phishing*. Online. Dostupné z: <https://www.graphus.ai/blog/history-of-phishing/>. [cit. 2025-02-10].
- [26] GUPTA, Brij B; ARACHCHILAGE, Nalin a PSANNIS, Kostas, 2018. Defending against Phishing Attacks: Taxonomy of Methods, Current Issues and Future Directions. *Telecommunication Systems*. 2018. 67. Dostupné z: 10.1007/s11235-017-0334-z. [cit. 2025-02-10].

- [27] PHISME COFENSE, 2021. *Social Media Phishing: What You Need to Know*. Online. Dostupné z: <https://cofense.com/knowledge-center/social-media-phishing-what-you-need-to-know/>. [cit. 2025-02-18].
- [28] HASSAN, Nihad A., 2024. *Phishing on social media: How threat actors exploit social media platforms to attack people*. Online. Dostupné z: <https://www.authentic8.com/blog/phishing-on-social-media>. [cit. 2025-02-18].
- [29] SCHILD, Matt, 2023. *Research identifies social media phishing vulnerabilities*. Online. Dostupné z: <https://biz.source.colostate.edu/research-identifies-social-media-phishing-vulnerabilities>. [cit. 2025-02-18].
- [30] POINT, PERCEPTION, 2024. *Detecting and Preventing AI-Based Phishing Attacks: 2024 Guide*. Online. Dostupné z: <https://perception-point.io/guides/ai-security/detecting-and-preventing-ai-based-phishing-attacks-2024-guide/>. [cit. 2025-03-01].
- [31] CARTIER, Maxime, 2025. *AI Phishing Attacks: How Big is the Threat?* Online. Dostupné z: <https://hoxhunt.com/blog/ai-phishing-attacks>. [cit. 2025-03-01].
- [32] ALSTON, Elena, 2023. *What is natural language processing (NLP)?* Online. Dostupné z: <https://zapier.com/blog/natural-language-processing/> [cit. 2025-03-05].
- [33] SANDHU, Jamie A, 2024. *What are LLMs and generative AI? A beginner's guide to the technology turning heads*. Online. Dostupné z: <https://srinstitute.utoronto.ca/news/gen-ai-llms-explainer>. [cit. 2025-03-05].
- [34] KARRAS, Tero; LAINE, Samuli a AILA, Timo, 2021. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021. 43, č. 12, s. 4217–4228. Dostupné z: 10.1109/TPAMI.2020.2970919.
- [35] RIEHLE, Cornelia, 2022. *Europol Report Criminal Use of Deepfake Technology*. Online. Dostupné z: <https://eucrim.eu/news/europol-report-criminal-use-of-deepfake-technology/>. [cit. 2025-03-06].
- [36] EUROPOL, 2022. *Europol report finds deepfake technology could become staple tool for organised crime*. Online. Dostupné z: <https://www.europol.europa.eu/media-press/newsroom/news/europol-report-finds-deepfake-technology-could-become-staple-tool-for-organised-crime>. [cit. 2025-03-06].
- [37] ANDRE, Dave, 2024. *Respeecher AI Review 2025: Features, Pricing, Hacks, and Tips*. Online. Dostupné z: <https://www.allaboutai.com/ai-reviews/respeecher-ai/>. [cit. 2025-03-08].
- [38] DAMIANI, Jesse, 2019. *A Voice Deepfake Was Used To Scam A CEO Out Of 243000*. Online. Dostupné z: <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>. [cit. 2025-03-08].
- [39] AZIZI, Othman, 10 2018. *Social Media Data Mining and Analytics*. Online. [cit. 2025-03-15].
- [40] DENBIGH-WHITE, Chris, 2024. *Seven Takeaways from the 2023 Verizon Data Breach Investigations Report*. Online. Dostupné z: <https://www.nextdlp.com/resources/blog/seven-takeaways-from-2023-verizon-data-breach-investigations-report>. [cit. 2025-03-15].
- [41] DUBÓCZI, Michael, 2023. *OSINT: Čo sú otvorené zdroje informácií a ako ich používať*. Online. Dostupné z: <https://infosecurity.sk/projekty/osint-co-su-otvorene-zdroje-informacii-a-ako-ich-pouzivat-disinfo-basics/>. [cit. 2025-03-15].

- [42] YADAV ANUKRITI, Kumar Deepak a YASHA, Hasija, 2023. *Behaviour Analysis Using Machine Learning Algorithms In Health Care Sector*. Dostupné z: 10.1109/InCACCT57535.2023.10141829.
- [43] LLC, Ponemon Institute, 2022. *IBM Cost of a Data Breach Report 2022*. Online. Dostupné z: <https://www.key4biz.it/wp-content/uploads/2022/07/Cost-of-a-Data-Breach-Full-Report-2022.pdf>. [cit. 2025-03-15].
- [44] SOSAFE, 2023. *One in five people click on AI-generated phishing emails, SoSafe data reveals*. Online. Dostupné z: <https://sosafe-awareness.com/company/press/one-in-five-people-click-on-ai-generated-phishing-emails-sosafe-data-reveals/>. [cit. 2025-03-20].
- [45] FREDRIK HEIDING, Bruce Schneier a VISHWANATH, Arun, 2024. *AI Will Increase the Quantity — and Quality — of Phishing Scams*. Online. Dostupné z: <https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams>. [cit. 2025-03-20].
- [46] SHELWELL, Gareth, 2025. *The 6 Most Popular AI Scams In 2025*. Online. Dostupné z: <https://caniphish.com/blog/ai-scams>. [cit. 2025-03-20].
- [47] BARRY, Christine, 2024. *5 Ways cybercriminals are using AI: Phishing*. Online. Dostupné z: <https://blog.barracuda.com/2024/03/28/-5-ways-cybercriminals-are-using-ai--phishing>. [cit. 2025-03-20].
- [48] POIREAULT, Kevin, 2023. *The Dark Side of Generative AI: Five Malicious LLMs Found on the Dark Web*. Online. Dostupné z: <https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html>. [cit. 2025-03-20].
- [49] KELLEY, Daniel, 2023. *WormGPT – The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks*. Online. Dostupné z: <https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/>. [cit. 2025-03-20].
- [50] RECORDED FUTURE INSIKT GROUP, 2024. *Adversarial Intelligence: Red Teaming Malicious Use Cases for AI*. Online. Dostupné z: <https://www.recordedfuture.com/research/adversarial-intelligence-red-teaming-malicious-use-cases-ai>. [cit. 2025-03-20].
- [51] KEEPNET, 2024. *What is Deepfake Phishing?* Online. Dostupné z: <https://keepnetlabs.com/blog/what-is-deepfake-phishing>. [cit. 2025-03-25].
- [52] TUMMALAPENTA, Srimi, 2024. *How a new wave of deepfake-driven cyber crime targets businesses*. Online. Dostupné z: <https://www.ibm.com/think/insights/new-wave-deepfake-cybercrime>. [cit. 2025-03-25].
- [53] SJOUWERMAN, Stu, 2024. *Deepfake Phishing: The Dangerous New Face Of Cybercrime*. Online. Dostupné z: <https://www.forbes.com/councils/forbestechcouncil/2024/01/23/deepfake-phishing-the-dangerous-new-face-of-cybercrime/>. [cit. 2025-03-25].
- [54] YASAR, Kinza; WIGMORE, Ivy a BARNEY, Nick, 2024. *What is deepfake technology?* Online. Dostupné z: <https://www.techtarget.com/whatis/definition/deepfake>. [cit. 2025-03-25].
- [55] SOLULAB, 2023. *What is GAN? – Generative Adversarial Networks Guide*. Online. Dostupné z: <https://www.solulab.com/generative-adversarial-network/>. [cit. 2025-04-12].

- [56] JIAYI, Hao, 10 2024. Advances in Deepfake Generation and Detection Technologies: Challenges and Opportunities. *Transactions on Computer Science and Intelligent Systems Research*. 10 2024. 6, s. 13–21. Dostupné z: 10.62051/m3s6nc42.
- [57] ANDERSON, Martin, 2022. *The Future of Autoencoder-Based Deepfakes*. Online. Dostupné z: <https://blog.metaphysic.ai/future-autoencoder-deepfakes/>. [cit. 2025-03-29].
- [58] DEEPAK BIRLA, 2019. *Basics of Autoencoders*. Online. Dostupné z: <https://medium.com/@birla.deepak26/autoencoders-76bb49ae6a8f>. [cit. 2025-04-12].
- [59] SHEN, Jonathan; PANG, Ruoming; WEISS, Ron J.; SCHUSTER, Mike; JAITLEY, Navdeep et al., 2017. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *CoRR*. 2017. abs/1712.05884. Dostupné z: <http://arxiv.org/abs/1712.05884>.
- [60] CHEN, Heather a MAGRAMO, Kathleen, 2 2024. Finance worker pays out \$25 million after video call with deepfake ‘chief financial officer. *CNN*. 2 2024. Dostupné z: <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>. [cit. 2025-03-29].
- [61] FRENCH, Laura, 2024. *Deepfake video conference convinces employee to send \$25M to scammers*. Online. Dostupné z: <https://www.scworld.com/news/deepfake-video-conference-convinces-employee-to-send-25m-to-scammers>. [cit. 2025-03-29].
- [62] MAGRAMO, Kathleen, 5 2024. British engineering giant Arup revealed as \$25 million deepfake scam victim. *CNN*. 5 2024. Dostupné z: <https://edition.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/index.html>. [cit. 2025-03-29].
- [63] FABRICIO, Toapanta; BELÉN, Rivadeneira; CHRISTIAN, Tipantuña a DANNY, Guamán, 11 2024. AI-Driven Vishing Attacks: A Practical Approach. 11 2024. s. 15. Dostupné z: 10.3390/engproc2024077015.
- [64] SJOUWERMAN, STU, 2024. *Vishing Meets AI: The Changing Nature of Phishing Threats Are your employees prepared for deepfake phone attacks?* Online. Dostupné z: <https://www.inc.com/inc-masters/vishing-meets-ai-the-changing-nature-of-phishing-threats.html>. [cit. 2025-03-30].
- [65] TYRRELL, James, 2023. *Microsoft shaves voice cloning time down to 3 seconds*. Online. Dostupné z: <https://techhq.com/2023/01/microsoft-shaves-voice-cloning-time-down-to-3-seconds/>. [cit. 2025-03-30].
- [66] WHITE, Marcus, 2025. *AI vishing: Mastering the art of voice deception*. Online. Dostupné z: <https://specopssoft.com/blog/ai-vishing-voice-deception/>. [cit. 2025-04-18].
- [67] ASTRANOVA, Emily a ISSA, Pascal, 2024. *Whose Voice Is It Anyway? AI-Powered Voice Spoofing for Next-Gen Vishing Attacks*. Online. Dostupné z: <https://cloud.google.com/blog/topics/threat-intelligence/ai-powered-voice-spoofing-vishing-attacks>. [cit. 2025-04-2].
- [68] LOVERIA, Catherine; SAMANIEGO, Jovit; NICOLETA, Gabriel a BORJA, Aprilyn, 2024. *Vishing via Microsoft Teams Facilitates DarkGate Malware Intrusion*. Online. Dostupné z: https://www.trendmicro.com/en_gb/research/24/1/darkgate-malware.html. [cit. 2025-04-2].
- [69] KAMP, Paul, 2024. *Build an AI Voice Assistant with Twilio Voice, OpenAI’s Realtime API, and Python*. Online. Dostupné z: <https://www.twilio.com/en-us/blog/voice-ai-assistant-openai-realtime-api-python>. [cit. 2025-04-2].

- [70] PINTÉROVÁ, D., 08 2024. Právna regulácia umelej inteligencie (perspektívy a výzvy). *Právny obzor*. 08 2024. 4, s. 361 – 383. Dostupné z: <https://doi.org/10.31577/pravnyobzor.2024.4.02>.
- [71] O'SHAUGHNESSY, Matt, 2022. *One of the Biggest Problems in Regulating AI Is Agreeing on a Definition*. Online. Dostupné z: <https://carnegieendowment.org/2022/10/06/one-of-biggest-problems-in-regulating-ai-is-agreeing-on-definition-pub-88100>. [cit. 2025-04-25].
- [72] SUNDAR, S. Shyam; SCHMIT, Cason a VILLASENOR, John, 2023. *Audits, 'soft laws' and 'automation bias': 3 experts break down what it could take to regulate AI and how hard it will be*. Online. Dostupné z: <https://fortune.com/2023/04/03/how-to-regulate-ai-challenges-three-experts/>. [cit. 2025-04-25].
- [73] KRITIKOS, Mihalis, 2019. *Artificial Intelligence ante portas: Legal & ethical reflections*. Online. Dostupné z: [https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/634427/EPRS_BRI\(2019\)634427_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/634427/EPRS_BRI(2019)634427_EN.pdf). [cit. 2025-04-25].
- [74] LIFE INSTITUTE, Future, 2024. *High-level summary of the AI Act*. Online. Dostupné z: <https://artificialintelligenceact.eu/high-level-summary/>. [cit. 2025-04-12].
- [75] EUROPEAN DATA PROTECTION SUPERVISOR, 2024. *Artificial Intelligence Act*. Online. Dostupné z: https://www.edps.europa.eu/artificial-intelligence/artificial-intelligence-act_en. [cit. 2025-04-10].
- [76] NAHRA, Kirk J. a EVERS, Arianna, 2024. *The European Parliament Adopts the AI Act*. Online. Dostupné z: <https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20240314-the-european-parliament-adopts-the-ai-act>. [cit. 2025-04-10].
- [77] CATALA, Romain, 2024. *European AI Act: What It Is, Why It Matters, & What to Do About It*. Online. Dostupné z: <https://datadome.co/learning-center/eu-ai-act-what-it-is-why-it-matters-what-to-do/>. [cit. 2025-04-12].
- [78] ATANASOVSKA, Daniela, 2024. *AI Act: Fundamental Rights Impact Assessments (FRIA) – Who, When, Why, and How to Ensure Ethical AI Deployment*. Online. Dostupné z: <https://gdprlocal.com/ai-act-fundamental-rights-impact-assessments-fria-who-when-why-and-how-to-ensure-ethical-ai-deployment/>. [cit. 2025-04-12].
- [79] STAŇKOVÁ, Monika Kovář, 2024. *Řízení rizik podle Aktu o umělé inteligenci*. Online. Dostupné z: <https://www.gdpr.cz/ai-act-a-rizeni-rizik-umele-inteligence>. [cit. 2025-04-12].
- [80] MOODY'S CORPORATION, 2024. *European Union's AI Act: Implications for fraud prevention and anti-money laundering systems*. Online. Dostupné z: <https://www.moodys.com/web/en/us/kyc/resources/insights/european-unions-ai-act-implications-for-fraud-prevention-anti-money-laundering-systems.html>. [cit. 2025-04-15].
- [81] WHITE & CASE LLP, 2024. *AI Watch: Global regulatory tracker - OECD*. Online. Dostupné z: <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-oecd>. [cit. 2025-04-16].
- [82] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), 2025. *OECD AI Principles overview*. Online. Dostupné z: <https://oecd.ai/en/ai-principles>. [cit. 2025-04-16].

- [83] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), 2024. *Recommendation of the Council on Artificial Intelligence ze dne 22. května 2019, aktualizováno 3. května 2024 (OECD/LEGAL/0449)*. Online. Dostupné z: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>. [cit. 2025-04-16].
- [84] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), 2019. *How artificial intelligence works*. Online. Dostupné z: <https://oecd.ai/en/inside-artificial-intelligence>. [cit. 2025-04-16].
- [85] ČESKÁ REPUBLIKA, 2024. *Národní strategie umělé inteligence České republiky 2030*. Online. Dostupné z: <https://mpo.gov.cz/assets/cz/podnikani/2024/9/Narodni-strategie-umele-intelligence-CR-2030.pdf>. [cit. 2025-04-16].
- [86] SLOVENSKÁ REPUBLIKA, 2019. *Stratégia digitálnej transformácie Slovenska 2030*. Online. Dostupné z: <https://mirri.gov.sk/wp-content/uploads/2019/06/Strategia-digitalnej-transformacie-Slovenska-2030.pdf>. [cit. 2025-04-16].
- [87] LYNNE PARKER, 2020. *The American AI Initiative: The U.S. strategy for leadership in artificial intelligence*. Online. Dostupné z: <https://oecd.ai/en/wonk/the-american-ai-initiative-the-u-s-strategy-for-leadership-in-artificial-intelligence>. [cit. 2025-04-16].
- [88] ČESKÁ REPUBLIKA, 1993. *Usnesení č. 2/1993 Sb., předsednictva České národní rady o vyhlášení LISTINY ZÁKLADNÍCH PRÁV A SVOBOD jako součástí ústavního pořádku České republiky*. Online. In: *Zákony pro lidi*. AION CS, ©2010-2024. Dostupné z: <https://www.zakonyprolidi.cz/cs/1993-2>. [cit. 2025-04-20].
- [89] ČESKÁ REPUBLIKA, 1993. *Ústavní zákon č. 1/1993 Sb., ústava České republiky*. Online. In: *Zákony pro lidi*. AION CS, ©2010-2024. Dostupné z: <https://www.zakonyprolidi.cz/cs/1993-1>. [cit. 2025-04-20].
- [90] ČESKÁ REPUBLIKA, 2009. *Zákon č. 40/2009 Sb., trestní zákoník*. Online. In: *Zákony pro lidi*. AION CS, ©2010-2024. Dostupné z: <https://www.zakonyprolidi.cz/cs/2009-40>. [cit. 2025-04-20].
- [91] ICLG, 2024. *Cybersecurity Laws and Regulations Czech Republic 2025*. Online. Dostupné z: <https://iclg.com/practice-areas/cybersecurity-laws-and-regulations/czech-republic>. [cit. 2025-04-20].
- [92] ČESKÁ REPUBLIKA, 2014. *Zákon č. 181/2014 Sb., o kybernetické bezpečnosti a o změně souvisejících zákonů (zákon o kybernetické bezpečnosti)*. Online. In: *Zákony pro lidi*. AION CS, ©2010-2024. Dostupné z: <https://www.zakonyprolidi.cz/cs/2014-181>. [cit. 2025-04-20].
- [93] BUSINESSINFOCZ, 2024. *Evropská směrnice NIS2 dopadne na podniky už v roce 2025. České firmy musejí přidat*. Online. Dostupné z: <https://www.businessinfo.cz/clanky/evropska-smernice-nis2-dopadne-na-podniky-uz-v-roce-2025-ceske-firmy-museji-pridat/>. [cit. 2025-04-20].
- [94] UNESCO, 2021. *Recommendation on the Ethics of Artificial Intelligence*. Online. Dostupné z: <https://unesdoc.unesco.org/ark:/48223/pf0000380455f>. [cit. 2025-04-26].
- [95] GARNETT, Allie Grace, 2024. *5 ethical questions about artificial intelligence*. Online. Dostupné z: <https://www.britannica.com/money/ai-ethical-issues>. [cit. 2025-04-25].

- [96] NOVELLI, Claudio; CASOLARI, Federico; HACKER, Philipp; SPEDICATO, Giorgio a FLORIDI, Luciano, 11 2024. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*. 11 2024. 55. Dostupné z: <https://doi.org/10.1016/j.clsr.2024.106066>.
- [97] SANYUAN, Chen; CHENGYI, Wang; YU, Wu; ZIQIANG, Zhang; LONG, Zhou et al., 2025. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*. 2025. 33, s. 705–718. Dostupné z: 10.1109/TASLP.2025.3530270.
- [98] HENSE, Jake, 2025. *Artificial Intelligence: Ethical Concerns and Sustainability Issues*. Online. Dostupné z: <https://www.americancentury.com/insights/ai-risks-ethics-legal-concerns-cybersecurity-and-environment/>. [cit. 2025-04-27].
- [99] QUTHOR, 2024. *Data Insights on the Ethical Implications of Deepfake Technology*. Online. Dostupné z: https://quickcreator.io/quthor_blog/data-insights-on-ethical-implications-deepfake-technology/. [cit. 2025-04-27].
- [100] ANDREWS, Caitlin, 2025. *European Commission withdraws AI Liability Directive from consideration*. Online. Dostupné z: <https://iapp.org/news/a/european-commission-withdraws-ai-liability-directive-from-consideration>. [cit. 2025-04-27].
- [101] BRANDUSESCU, Ana a SIEBER, Renée, 2025. *The Paris AI Summit: Deregulation, Fear, and Surveillance*. Online. Dostupné z: <https://montrealetics.ai/the-paris-ai-summit-deregulation-fear-and-surveillance/>. [cit. 2025-04-27].
- [102] CLARK, James; DEMIRCAN, Muhammed a KETTAS, Kalyna, 2024. *Europe: The EU AI Act's relationship with data protection law: key takeaways*. Online. Dostupné z: <https://privacymatters.dlapiper.com/2024/04/europe-the-eu-ai-acts-relationship-with-data-protection-law-key-takeaways/>. [cit. 2025-04-27].
- [103] LAMINA, Oladimeji Azeez; AYUBA, Waliu Adebayo; ADEBIYI, Olubukola Eunice; MICHAEL, Gracious Ebunoluwa; SAMUEL, Ojo-Omoniyi Damilola et al., 2024. Ai-Powered Phishing Detection And Prevention. *Path of Science*. 2024. 10, č. 12, s. 4001–4010. Dostupné z: 10.22178/pos.112-7.
- [104] FARES, Hajar; KILANI, Džihád; FAGROUD, FatimaEzzahra; TOUMI, Hicham; LAKRAMI, Fatima et al., 2024. Machine learning approach for email phishing detection. *Procedia Computer Science*. 2024. 251, s. 746–751. Dostupné z: 10.1016/j.procs.2024.11.179.
- [105] BAHNSEN, Alejandro Correa; TORROLEDO, Ivan; CAMACHO, Luis David a VILLEGAS, Sergio, 2018. *DeepPhish : Simulating Malicious AI*. Online. Dostupné z: <https://api.semanticscholar.org/CorpusID:51691528>. [cit. 2025-04-30].
- [106] SALLOUM, Said; GABER, Tarek; VADERA, Sunil a SHAALAN, Khaled, 2021. Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey. *Procedia Computer Science*. 2021. 189, č. 12, s. 19–28. Dostupné z: <https://doi.org/10.1016/j.procs.2021.05.077>.
- [107] NISREEN, Innab; AHMED, Osman; MOHAMMED, Ataelfadiel; MARWAN, Abu-Zanona; BASSAM, Elzaghmouri et al., 07 2024. Phishing Attacks Detection Using Ensemble Machine Learning Algorithms. *Computers, Materials & Continua*. 07 2024. 80, s. 1–10. Dostupné z: 10.32604/cmc.2024.051778.

- [108] BIGGIO, Battista a ROLI, Fabio, 12 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*. 12 2018. 84, s. 317–331. Dostupné z: <https://doi.org/10.1016/j.patcog.2018.07.023>.
- [109] ARASH, Habibi Lashkari; MIN, Chen a ALI, Ghorbani, 09 2018. A Survey on User Profiling Model for Anomaly Detection in Cyberspace. *Journal of Cyber Security and Mobility*. 09 2018. Dostupné z: 10.13052/2245-1439.814.
- [110] RIDA, Nasir; MEHREEN, Afzal; RABIA, Latif a WASEEM, Iqbal, 10 2021. Behavioral Based Insider Threat Detection Using Deep Learning. *IEEE Access*. 10 2021. PP, s. 1–1. Dostupné z: 10.1109/ACCESS.2021.3118297.
- [111] GRASSI, Paul A.; NEWTON, Elaine M.; PERLNER, Ray A.; REGENSCHEID, Andrew R.; FENTON, James L. et al., 2020. *NIST Special Publication 800-63B, Digital Identity Guidelines: Authentication and Lifecycle Management*. Online. Dostupné z: <https://doi.org/10.6028/NIST.SP.800-63b>. [cit. 2025-05-01].
- [112] ISO - INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, (Edition 1, 2009). *ISO/IEC 19792:2009 Information technology - Security techniques - Security evaluation of biometrics*. Online. Dostupné z: <https://www.iso.org/standard/51521.html>. [cit. 2025-05-01].
- [113] SINCHANA, Goudar a MAMATHA, Ganganna, 07 2024. Review on Technology Advancements in Biometric Authentication. 07 2024. s. 8–19.
- [114] FIDO ALLIANCE, 2025. *User Authentication Specifications Overview*. Online. Dostupné z: <https://fidoalliance.org/specifications/>. [cit. 2025-05-01].
- [115] W3C, 2021. *Web Authentication: An API for accessing Public Key Credentials*. Online. Dostupné z: <https://www.w3.org/TR/webauthn-2/>. [cit. 2025-05-01].
- [116] YUNJI, Liang; SAGAR, Samtani; BIN, Guo a ZHIWEN, Yu, 2020. Behavioral Biometrics for Continuous Authentication in the Internet-of-Things Era: An Artificial Intelligence Perspective. *IEEE Internet of Things Journal*. 2020. 7, č. 9, s. 9128–9143. Dostupné z: 10.1109/JIOT.2020.3004077.
- [117] RYAN TERRY, 2025. *Adaptive Authentication Explained*. Online. Dostupné z: <https://www.crowdstrike.com/en-us/cybersecurity-101/identity-protection/adaptive-authentication/>. [cit. 2025-05-01].
- [118] DEEPWARE.AI™, 2025. *Scan & Detect Deepfake Videos*. Online. Dostupné z: <https://deepware.ai/>. [cit. 2025-05-02].
- [119] AMER OWAIDA, 2020. *Microsoft debuts deepfake detection tool*. Online. Dostupné z: <https://www.welivesecurity.com/2020/09/03/microsoft-debuts-deepfake-detection-tool/>. [cit. 2025-05-02].
- [120] RESEMBLE.AI, 2025. *Seamless integration for any security stack*. Online. Dostupné z: <https://www.resemble.ai/detect/>. [cit. 2025-05-02].
- [121] RUTECKA, Daria, 2025. *Liability for damages: the missing piece of the AI Act puzzle*. Online. Dostupné z: <https://schoenherr.eu/content/liability-for-damages-the-missing-piece-of-the-ai-act-puzzle>. [cit. 2025-05-03].

- [122] HEDLUND, Maria a PERSSON, Erik, 2025. *Distribution of responsibility for AI development: expert views*. Online. Dostupné z: <https://doi.org/10.1007/s00146-024-02167-9>. [cit. 2025-05-03].
- [123] TYMOSHCHUK, Taras, 2025. *AI due diligence: Geniusee best practices and recommendations*. Online. Dostupné z: <https://geniusee.com/single-blog/ai-due-diligence>. [cit. 2025-05-03].
- [124] MARCHISIO, Emiliano, 01 2021. In support of “no-fault” civil liability rules for artificial intelligence. *SN Social Sciences*. 01 2021. 1. Dostupné z: 10.1007/s43545-020-00043-z.
- [125] RUSCHEMEIER, Hannah, 2024. *Generative AI and Data Protection*. Online. Dostupné z: <https://ssrn.com/abstract=4814999>. [cit. 2025-05-03].
- [126] G7 HIROSHIMA SUMMIT, 2023. *Hiroshima Process International, Guiding Principles for Organizations, Developing Advanced AI System*. Online. Dostupné z: <https://www.mofa.go.jp/files/100573471.pdf>. [cit. 2025-05-03].
- [127] UNITED STATES GOVERNMENT, 2023. *Výkonný příkaz č. 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. 86 Fed. Reg. 64627. Online. Prezidentský dokument z výkonnej kancelárie prezidenta z 11. 1. 2023. Dostupné z: <https://www.federalregister.gov/d/2023-24283>. [cit. 2025-05-03].
- [128] EUROPEAN UNION AGENCY FOR CYBERSECURITY, 2025. *EU incident response and cyber crisis management*. Online. Dostupné z: <https://www.enisa.europa.eu/topics/eu-incident-response-and-cyber-crisis-management>. [cit. 2025-05-03].
- [129] IBM SECURITY, 2023. *IBM Cost of a Data Breach Report 2023*. Online. Dostupné z: <https://d110erj175o600.cloudfront.net/wp-content/uploads/2023/07/25111651/Cost-of-a-Data-Breach-Report-2023.pdf>. [cit. 2025-05-07].
- [130] GOVERNMENT OF CANADA, 2025. *Get Cyber Safe*. Online. Dostupné z: <https://www.getcybersafe.gc.ca/en>. [cit. 2025-05-08].
- [131] EURÓPSKA KOMISIA, 2021. *Digital Education Action Plan (2021-2027)*. Online. Dostupné z: <https://education.ec.europa.eu/focus-topics/digital/digital-education-action-plan>. [cit. 2025-05-08].
- [132] COALITION FOR CONTENT PROVENANCE AND AUTHENTICITY (C2PA), 2025. *An open technical standard providing publishers, creators, and consumers the ability to trace the origin of different types of media*. Online. Dostupné z: <https://c2pa.org/>. [cit. 2025-05-10].
- [133] OPERATION HOPE, 2025. *AI Ethics Council*. Online. Dostupné z: <https://operationhope.org/initiatives/ai-ethics-council/>. [cit. 2025-05-10].
- [134] GOVERNMENT OF CANADA, 2025. *Advisory Council on Artificial Intelligence*. Online. Dostupné z: <https://ised-isde.canada.ca/site/advisory-council-artificial-intelligence/en>. [cit. 2025-05-10].
- [135] PARTNERSHIP ON AI, 2025. *PAI's Responsible Practices for Synthetic Media. A Framework for Collective Action*. Online. Dostupné z: <https://syntheticmedia.partnershiponai.org/#landing>. [cit. 2025-05-10].

- [136] KASPERSKY, 2024. *Kaspersky shares privacy predictions for 2024*. Online. Dostupné z: <https://usa.kaspersky.com/about/press-releases/kaspersky-shares-privacy-predictions-for-2024>. [cit. 2025-05-11].
- [137] WHITTLESEY FORWARD ADVISING, 2022. *Phishing by Industry Benchmarking Report 2022*. Online. Dostupné z: https://www.wadvising.com/media/2250/whittlesey-2022-phishing-by-industry-benchmarking-report_v2.pdf. [cit. 2025-05-11].
- [138] MONIKA BICKERT, 2024. *Meta: Our Approach to Labeling AI-Generated Content and Manipulated Media*. Online. Dostupné z: <https://about.fb.com/news/2024/04/metaspas-approach-to-labeling-ai-generated-content-and-manipulated-media/>. [cit. 2025-05-02].

SEZNAM OBRÁZKŮ

Obrázek 1.1	Proces phishingového útoku realizovaného prostřednictvím e-mailové správy	18
Obrázek 2.1	Ukážka vyhledávání veřejně dostupných informací pomocí modelu ChatGPT-4o	35
Obrázek 2.2	Ukážka vygenerovaného personalizovaného e-mailu v programe CanIPhish	36
Obrázek 2.3	Ukážka podvodnej prihlasovacej stránky imitujúcej rozhranie Microsoft prihlásenia	37
Obrázek 2.4	Porovnanie legitímneho a podvodného prihlasovacieho okna Microsoft.	37
Obrázek 2.5	Architektúra Generatívnej adversariálnej siete (GAN).....	39
Obrázek 2.6	Architektúra Autoenkódera.....	40
Obrázek 3.1	Životný cyklus systému AI definovaný a schválený expertnou skupinou OECD pre AI	55

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

2FA	Dvojfaktorová autentifikácia
AI	Artificial intelligence (umelá inteligencia)
AIS	Artificial intelligence system (systém umelej inteligencie)
APT	Advanced Persistent Threat (pokročilá perzistentná hrozba)
ASR	Automatické rozpoznávanie reči
CNN	Convolutional neural network (konvolučné neurónové siete)
DPIA	Data Protection Impact Assessment (posúdenie vplyvu na ochranu údajov)
DL	Deep Learning (hlbkové učenie)
FRIA	Fundamental Rights Impact Assessments (hodnotenie vplyvu na základné práva)
GAN	Generative Adversarial Networks (generatívne adverzariálne siete)
GDPR	Všeobecné nariadenie o ochrane údajov
GPAI	General-Purpose AI (umelá inteligencia na všeobecné použitie)
LLM	Large language models (veľké jazykové modely)
ML	Machine Learning (strojové učenie)
MitM	Man-in-the-Middle (človek uprostred)
MFA	Multi-factor authentication (viacfaktorová autentifikácia)
NLP	Natural language processing (spracovanie prirodzeného jazyka)
OSINT	Open-source intelligence (zpravodajstvo z verejných zdrojov)
TTS	Text-to-speech
VPN	Virtual private network (virtuálna privátna sieť)