UNIVERZITA KARLOVA

FAKULTA SOCIÁLNÍCH VĚD

Institut komunikačních studií a žurnalistiky

Katedra marketingové komunikace a PR

Současný vývoj AI nástrojů v České republice a jejich možné využití, limity a predikce ve strategickém boji v informační válce

Diplomová práce

Autor práce: Barbora Štěpánová

Studijní program: Strategická komunikace

Vedoucí práce: PhDr. Tereza Klabíková Rábová, Ph. D

Rok obhajoby: 2023

Prohlášení

1. Prohlašuji, že jsem předkládanou práci zpracovala samostatně a použila jen uvedené prameny a literaturu.
2. Prohlašuji, že práce nebyla využita k získání jiného titulu.
3. Souhlasím s tím, aby práce byla zpřístupněna pro studijní a výzkumné účely.

V Praze dne 21. 4. 2023

Barbora Štěpánová

Bibliografický záznam

ŠTĚPÁNOVÁ, Barbora. Současný vývoj AI nástrojů v České republice a jejich možné využití, limity a predikce ve strategickém boji v informační válce, Praha 2023, diplomová práce, Univerzita Karlova, Fakulta sociálních věd, Institut komunikačních studií a žurnalistiky, Katedra marketingové komunikace a PR, obor Strategická komunikace. Vedoucí práce: PhDr. Tereza Klabíková Rábová, Ph.D.

Rozsah práce: **184 687**

Abstrakt: Práce se bude zabývat současným stavem AI technologií v provázání s dezinformacemi a informační válkou, jejich možným vývojem, limity a obavami.

V rámci teorie budeme vycházet z již ustálených konceptů prací, mapujících současný stav, ale i predikujících budoucnost, formulujících možné limity a pokládajících otázky do budoucna. K pochopení současného stavu jsou prezentovány i různé případové studie relevantní k tématu. V praktické části bude proveden výzkum formou focus group, která bude diskutovat závěry a podněty vzešlé z teoretické části a analýz. Následovat bude diskuse výsledků a závěr, ve kterém nastíníme přínos mezidisciplinární debaty, zhodnotíme vliv fact-checkingu, možnosti financování nebo roli člověka v hybridním přístupu.

Klíčová slova: výzkum AI, hybridní přístup, focus group, dezinformace, informační válka, multidisciplinární výzkum

Abstract: This work will discuss the current state of AI technologies in connection with disinformation and information warfare, their possible development, limits and concerns.

Theoretical part will start with established concepts, works mapping the current state, but also predictions for the future and formulating possible limits. To understand the current state,various relevant case studies will be presented. In the practical part, Focus Group research will be conducted, which will discuss the conclusions and topics arising from the theoretical part and analysis. A discussion of the results and a conclusion will follow.

Key words: AI research, Hybrid aproach, HAI, focus group, Disinformation, Information war, multidisciplinary research

Název práce / Title: Současní vývoj AI nástrojů v České republice a jejich možné využití, limity a predikce ve strategickém boji v Informační válce / The current development of AI tools in the Czech Republic and their possible use, limits and predictions in the strategic fight in the Information War

Poděkování

Na tomto místě bych ráda poděkovala všem svým vyučujícím za poskytnuté vzdělání a díky tomu za přehled, který mi umožnil napsat tuto práci. Dále bych ráda poděkovala své vedoucí práce, která mi poskytla detailní zpětnou vazbu, nespočet podmětů a hodiny konzultací. Pomohla mi s uspořádáním obsahu a doporučovala relevantní studie a případy. Také bych ráda poděkovala všem respondentům výzkumu za poskytnutí času a názorů.

Obsah

OBSAH **7**

TEORETICKÁ ČÁST **10**

VYMEZENÍ ZKOUMANÉHO TÉMATU **10**
1. PROPOJENÍ SOCIÁLNÍCH VĚD **12**
1.1. HEURISTIKY A BIAS 13
1.2. BACKFIRE EFEKTY 17
1.3. INFORMATION DEFICIT MODEL 18
1.4. SOCIÁLNÍ BUBLINY A CLUSTERY 18
2. UMĚLÁ INTELIGENCE **20**
ZÁKLADNÍ POJMY A NÁSTROJE 20
2. 1. SHRNUTÍ ÚVAH A PREDIKCÍ OHLEDNĚ NÁSTROJŮ AI V KONTEXTU INFORMAČNÍ VÁLKY 21
2.2. NÁSTROJE POČÍTAČOVÉ PROPAGANDY 23
2.3. SLUŽBY AI V PRAXI 26
2.4. LIDÉ V PROCESU AI 26
2. 5. OBECNÉ PROGNÓZY A POTŘEBY PRO EFEKTIVNÍ OBRANU V INFORMAČNÍ VÁLCE 32
2.6. LIMITY A HROZBY KONTEXTU AI 36
2. 7. PŘÍLEŽITOSTI VYUŽITÍ AI DO BUDOUCNA 43
2.8. INSTITUCE VSTUPUJÍCÍ DO KONTEXTU AI 49
3. SITUAČNÍ ANALÝZA **57**
3.1. PEST 58

PRAKTICKÁ ČÁST **73**

4. FOCUS GROUP **73**
4.1. PŘEDSTAVENÍ METODY FOCUS GROUP 73
4.2. PŘEDSTAVENÍ VÝZKUMU 74
4.3. PRŮBĚH A VÝSLEDKY 79
5. VÝSLEDKY A DISKUSE **93**
6. LIMITY **97**

ZÁVĚR **99**

ZDROJE **102**

Úvod

V rámci nového informačního věku, který je posílený stoupající dostupností internetového připojení a jeho uživatelů, kteří se současně stávají i tvůrci obsahu, rostou obavy žurnalistů, politiků, odborníků na bezpečností studia nebo různých pracovníků komunikačních oborů z důsledků, které tento rozvoj může mít na povahu tvořených a šířených informací a dezinformací a schopnost uživatelů je od sebe navzájem rozeznat.

Právě prostředí internetu s sebou přináší nové technologie, které je nutno zkoumat a diskutovat o jejich možném vlivu. Tato práce chce zmapovat nové nástroje a možnosti umělé inteligence a s ní spojené technologie a zamyslet se nad jejím možným využitím pro boj v informační válce. Klade si za cíl přednést jakýsi úvod do problematiky se základním výčtem pojmů a konceptů, které vstupují do debat odborníků na toto téma. Měla by poskytnout základní odrazový můstek pro pochopení současného stavu a aktuálních výzev.

V práci nejprve v teoretické části představím pozadí sociálních věd, které vstupuje do kontextu dezinformací. Přednesu výčet základních konceptů behaviorální ekonomie a psychologie, které ovlivňují rozhodovací a vyhodnocovací procesy lidí a jsou často spojované právě s prostředím internetu a dezinformací. Tyto koncepty nám pomáhají předvídat možné efekty zpětného rázu a chápat situace, kdy zásahy nefungují tak, jak bylo původně myšleno.

Následně představím základní premisy a východiska umělé inteligence, její možnosti a vývoj. Úvod této kapitoly představuje krátké shrnutí, výklad a terminologii obklopující vývoj umělé inteligence. Následně budu demonstrovat různé možnosti a dopady využití umělé inteligence, příklady praktických nástrojů a přístupů, a instituce zajímající se o tuto technologii. Pro pochopení prostředí a současného stavu bude vytvořena PEST analýza. Celá práce je taktéž doplněna několika případovými studiemi, které demonstrují na praktických příkladech teoretická východiska.

Praktická část se pak zaměří na výzkum formou focus group. Cílem výzkumu bude zaznamenat názory, limity a predikce pro další vývoj a využití AI, které budou formulovány účastníky ve skupině. Najít společný konsensus při hodnocení vývoje, úspěchu zapojení vlády do procesu, odpovědět na otázky okolo regulace, účinku a role poskytovatelů / platforem v procesu boji s dezinformacemi.

Přínos této práce spočívá v získání a kategorizaci názorů odborníků na témata spojená s AI a informační válkou v prostředí České republiky, potažmo Evropy, vytvoření predikčního rámce a zamyšlení se nad otázkou legislativy a etiky AI. Má přispět do probíhající diskuse o vlivu a možnostech nástrojů umělé inteligence, představit povahu vhledu, který do tohoto vývoje mohou přinést odborníci z jiných oborů a demonstrovat hodnotu, kterou toto propojení odborníků napříč obory může mít. V neposlední řadě i poskytuje přehled současného stavu a představuje zajímavé nástroje a předměty, které mohou být využity v praxi. Zároveň se směr práce mírně odchýlil od původního záměru formulovaného v tezi. Komplexnost tématu a velké množství různých přístupů, které lze zvolit, transformovaly původní intenci zpracování. Práce je více koncipovaná jako jakýsi koncepční úvod do problematiky v teoretické části a následné obecné zamyšlení nad tímto konceptem a kontextem, který ho obklopuje. Případové studie slouží jako mimo technické pomůcky, které nejsou nezbytné, nicméně jejich zmínění dokresluje kontext a demonstruje na praktických příkladech východiska teoretické části a probíraných témat.

Teoretická část

Vymezení zkoumaného tématu

V úvodu teoretické části je nejprve důležité vymezit si oblast zkoumání, na kterou se tato práce soustředí a kterou naopak úmyslně opomíjí. Problematika dezinformací, jejich pojmenování, vyčlenění, klasifikace a koncepce je velmi křehké a složitě uchopitelné téma. Polarizace společnosti, propletení narativů, jednotlivé cíle aktérů a motivace konzumentů tvoří komplexní problematiku, kterou má práce nemá ambice uchopit, popsat, pojmenovávat nebo hodnotit. Problematika definování a uchopení dezinformací jednak může podléhat subjektivním výkladům dle politického a geografického umístění daného subjektu na mapě, a jednak definice podložené odbornou literaturou mohou vypadat a chovat se naprosto odlišně v praxi.

Komplexnost problematiky definice, a nutnost věnovat její klasifikaci dostatek prostoru a výzkumu nám demonstrovaly únorové události okolo funkce Vládního zmocněnce pro oblast médií a dezinformací. Tato nově vzniklá funkce měla sloužit k efektivnějšímu boji s dezinformacemi a zastával ji Ing. Michal Klíma. Nicméně ten na své pozici nezvládl položit potřebné a univerzální definice či jasné měřítko, což způsobilo nedostatek transparentnosti a následnou ztrátu důvěry v očích veřejnosti, politiků i novinářů. Absence této definice byla kritická právě v připravovaném Akčním plánu proti dezinformacím, kde při nesprávném použití mohla ohrožovat svobodu slova. Proti plánu se postavila Unie vydavatelů, která jej vnímá jako škodlivý pro mediální prostředí a obává se případného zneužití pro účely cenzury. (Dohnalová, aktualne.cz, 2023) Ačkoliv měl za sebou Klíma za necelý rok ve funkci i úspěšné projekty, v důsledku nepovedeného Akčního plánu byla jeho pozice zrušena a agenda převedena pod poradce pro národní bezpečnost. Na příkladu z praxe můžeme vnímat potřebu věnovat dostatek zdrojů a výzkumu definování dezinformací, které tato práce ve svém obsahu nemůže prokázat. Nicméně vhodné definice nalezneme například: Kapantai, Christopoulou, Berberidis a Peristeras, 2021, A systematic literature review on disinformation: Toward a unified taxonomical framework, anebo Gelfert, 2018, Fake news: A definition. Dále se definici

dezinformací věnují různá občanské sdružení nebo odborníci, například Václav Moravec (Focus Václava Moravce, 2018), o definici se pokouší i Ministerstvo vnitra České republiky. To jej definuje takto: „Pojem „dezinformace“ znamená šíření záměrně nepravdivých informací, obzvláště pak státními aktéry nebo jejich odnožemi vůči cizímu státu nebo vůči médiím, s cílem ovlivnit rozhodování nebo názory těch, kteří je přijímají. Je pozoruhodné, že tento dnes mezinárodně užívaný pojem pochází nejspíš z ruského дезинформация [dezinformacija], který byl prvně zaznamenán v roce 1949 (jak uvádí např. Oxford English Dictionary).” (MVCR, cit. 2023). Vzhledem k povaze tohoto orgánu lze definici považovat za univerzálně platnou pro vládní orgány České republiky a tedy oficiální. Problematika se pak následně spojuje i se svobodou slova, na jejíž obranu se staví mnoho veřejně známých osobností, v praxi se tak tyto definice ohýbají, přeformovávají a smrskávají, aby posloužily preferovanému narativu daného aktéra.

Hlavním cílem této práce je za pomoci rešerše současné literatury a případových studií představit a zhodnotit současné konkrétní technologické možnosti umělé inteligence a jejich možné využití do budoucna. Toto hodnocení a predikce budou tvořeny pomocí focus group složené z vybraných odborníků, kterým budou předkládány relevantní otázky vycházející z teoretické části této práce. Nezbytné výchozí informace a koncepty pro pochopení prostředí budou představeny v kapitole analýzy současného stavu.

Na úvod je třeba specifikovat povahu a pozadí ústředního tématu této práce. S rozvojem digitálního světa a vývojem moderních technologií roste i hrozba, kterou tento svět a způsob generování, šíření a přijímání informací přináší jako vedlejší produkt. Jako stát čelíme hybridní hrozbě cílící na politiky, kritickou infrastrukturu, naše obyvatelstvo, ale i na naše hodnoty, které se snaží rozložit a radikalizovat. Komplexnost této hrozby a možný dosah na morálku občanů je téměř nepopsatelná, spojuje v sobě technologickou hrozbu, diplomatické vztahy, ekonomické zájmy a útok na lidskou psychiku. (Blank, 2008) Je tedy jasné, že ochrana proti tomuto druhu hrozby není jednoduchá záležitost a na komplexním řešení se budou muset podílet odborníci napříč obory a přijít se společným řešením. Jednou ze součástí tohoto komplexního řešení pak mohou být právě nástroje umělé inteligence. S ohledem na naše prostředí jsou pak hlavní národní hráči

Rusko, z důvodu historického vlivu a zájmu na ovlivňování vnitrostátních záležitostí a částečně pak Čína, do jejíhož zájmu se dostáváme spíše okrajově, nicméně disponuje nezanedbatelným technologickým know-how, které vyměňuje s Ruskem, které zase naopak poskytuje know how ohledně strategie informační války. (Blank, 2008)

Nebezpečí informační války leží v její domnělé neškodnosti a počáteční nenápadnosti. Jelikož nevidíme krev a oběti, máme pocit, že se jedná o okrajovou hrozbu. „Nicméně informační válka plní veškeré definice Clausewitze – střet ideí, kde se jedna strana snaží přinutit druhou plnit jejich vůli – ačkoliv není plně jasné, jaká je například vůle Ruska, kromě tedy snahy potvrdit svůj status hegemona v Pobaltí.“ (Blank, 2008)

1. Propojení sociálních věd

Jelikož je tato práce psaná z pohledu sociálních věd nahlížejících na nástroje umělé inteligence, je nutné si představit propojení těchto dvou oborů. Do kontextu vývoje AI nástrojů pro boj v informační válce vstupuje i psychologické a sociologické pozadí. Myšlenkové procesy konzumentů je nutné vzít v potaz a předpokládat je při vývoji různých zásahů nástrojů umělé inteligence. Vnímáme několik psychologických a behaviorálních jevů a pojmů, které vstupují do prostředí dezinformací a informační války. Tyto efekty pomáhají chápat motivace jednotlivých aktérů a jejich pochopení nám může pomoci efektivně reagovat. Následující shrnutí tedy představí nejvýraznější z nich. Vzhledem k povaze tématu a faktu, že na něj pohlížíme ze strany komunikačních studií, považuji za relevantní předložit základní behaviorální a psychologické koncepty v celém spektru. Některé z těchto konceptů jsou zmíněné v transkriptu, respondenti se o ně opírají a vycházejí z nich, jsou jim známé a představují část jejich expertízy. Některé uvedené příklady vycházejí z těchto jevů, mohou sloužit k pochopení reakcí a procesu. Pro plné pochopení debaty a kontextu této práce tedy považuji za relevantní následný výčet zahrnout, ačkoliv na něj není přímo navazováno.

1.1. Heuristiky a bias

1.1.1. Emoce

Jedním z pojmů vstupujícím do prostředí dezinformací je **emotional reasoning**: Jedná se o kognitivní proces, kdy jedinec jako důkaz použije své vnitřní pocity, tedy „*já cítím, že je něco špatně, je to tedy špatně“*. Jedinec dojde k chybnému závěru pouze na základě svých pocitů bez jakýchkoliv důkazů, a to i přes to, že empirické důkazy mohou svědčit o opaku. Tento proces pak způsobuje úzkosti, panické stavy, pocity strachu a neklidu. (Sousa a Morton, 2022) V kontextu dezinformací můžeme tento proces pozorovat ve chvíli, kdy uživatel věří něčemu, co postrádá empirické důkazy, na základě pocitu (M*ně se vždy zdál tento člověk nesympatický, takže je pravda, že se účastní spiknutí, já to z něj hned cítil*.). Zároveň samotný pocit vede k většímu strachu a úzkosti, a ta nás pak výměnou zase činí více náchylné k podlehnutí dezinformacím. (Robbins, 2020)

V nejistých situacích, kdy postrádáme relevantní informace, jsme náchylní k emocím jako úzkost, hněv nebo strach. (Weeks, 2015) Tyto emoce pak vyvolávají takzvaný **emoční tlak**, který hraje roli při šíření nepravdivého obsahu, může posílit tendenci lidí tento obsah šířit, může mu pomoci ospravedlnit pocity, které prožívá, anebo snížit právě emoční tlak dodáním „informací” a vyvedením z nejistoty. (Difonzo a Bordia, 2007) Četnost sdílení falešných zpráv se zvyšuje v nejistých situacích. (Karami, Nazer a Liu, 2021) Právě proto mnoho dezinformačních aktivit cílí právě na emoce, podporující větší sdílení, snaží se v lidech násobit úzkost, vztek nebo strach.

Hornový efekt je tendence rychle přisuzovat lidem zlé vlastnosti a úmysly na základě jedné negativní zprávy nebo vlastnosti, nebo chyby vzhledu. (Rowley a Ramasamy, 2016) Tento efekt může být také chápán jako démonizování jedince / skupiny / chování, jelikož etnologicky pochází od slova rohy.

1.1.2. Skupinové heuristiky

Bandwagon efekt neboli stádové chování je osvojení si postoje, názoru, chování nebo přístupu pouze proto, že si jej osvojili i ostatní kolem nás – tedy přizpůsobení se davu nebo našemu okolí. Hlavní hrozba tohoto efektu leží v tom, že čím více lidí si dané chování, názor nebo postoj osvojí, ten tím více získává na důvěryhodnosti a legitimnosti, ačkoliv může být naprosto podvržený (Kiss a Simonovits, 2014). To vede k dalšímu šíření a ztěžuje možnosti nápravy, uvedení na pravou míru. S tímto efektem se pojí **argumentační faul argumentum ad populum** – musí to být pravda, protože to všichni říkají. (Colman, 2009) Oproti tomu **Reverse Bandwagon efekt** neboli snobský efekt pracuje s potřebou jedince vymezit se a odlišit se, tato „antistádovost" pak také vystupuje v motivaci k šíření dezinformací, snaha vyhradit se proti stádu či establishmentu, nevěřit tomu, co ostatní, hledat „alternativní fakta", nebýt „ovce". Hlavním cílem jedince, který podléhá popisovanému jevu, je něco nedělat zkrátka proto, že to dělají ostatní. (Simon, 1954)

Se stádovým chováním souvisí pojem **informační kaskády**, na jejichž základě skupina lidí provádí stejná rozhodnutí v sekvencích, jedinec se rozhoduje podle většiny, ignoruje u toho své vlastní znalosti a nesnaží se získat žádné vlastní informace. Důvodem vzniku efektu informační kaskády je důvěra ve většinu a potřeba následovat sociální normy. Informační kaskády jsou velmi křehké a nestabilní a jejich vyvrácení či zhroucení je velmi jednoduché – danému jedinci nebo skupině stačí dodat relevantní hlubší informace. (Bikhchandani, Hirshleifer a Welch, 1998) Informační kaskády se často objevují na burze, na sociálních sítích pak často funguje jako virálně se šířící šokující zpráva, kterou lze ale jednoduše vyvrátit.

1.1.3. Chybné zpracování, nakládání a prioritizace informací

V rámci heuristik a bias rozeznáváme skupinu efektů, které popisují různé procesy zpracovávání, upřednostňování a vybavování si informací, které mohou zkreslit náš úsudek.

Konfirmační zkreslení je efekt, při kterém jedinec preferuje pouze ty informace, které potvrzují jeho názor. Z dostupných informací vybírá pouze ty konfirmační a cíleně ignoruje protichůdné informace. Tento efekt může nastat i při vybavování si informací. (Nickerson, 1998) Toto selektivní vyhledávání pouze potvrzujících informací pak v prostředí sociálních sítí vede k vytvoření sociálních bublin, ve kterých dochází k dalšímu vzájemnému potvrzování a ujišťování. Zároveň tento efekt imunizuje jedince proti vyvracení dezinformací, protichůdné informace tedy zkrátka přehlíží.

Heuristika dostupnosti je tendence preferovat ty informace, které nám jsou dostupnější, snadněji je nalezneme, nebo si je vybavíme. Tyto informace mohou být dostupnější z důvodu opakovaného vystavení, tedy že je vidíme všude, nebo jsou snadno zpracovatelné, jednoduše podané, vázané na něco známého. Základní premisou je, že pokud je něco snadno vybavitelné, musí to být důležité. (Esgate a Groome, 2005) Při této heuristice také máme tendenci favorizovat starší informace oproti novějším. Zdánlivá velikost dopadů akce silně koreluje s jejich mentální dostupností. Jinými slovy, lidé mají tendenci vnímat následky jako závažnější, čím snazší je jejich vybavení. Zároveň čím snazší je vybavení dané informace, tím větší mají jedinci důvěru v její pravdivost. (Schwarz, Bless, Strack, Klumpp a kolektiv, 1991) Heuristiku dostupnosti využívají některé dezinformace, neboť jejich senzačnost, urgentnost a jednoduché zpracování podporuje následné snadnější vybavení. Zároveň mnoho z nich lze vyvrátit pouze komplexní argumentací a vysvětlením, které už z podstaty pak není tak jasné a dostupné pro vybavování.

Salience bias neboli **zkreslení význačnosti** popisuje naši tendenci soustředit se na položky nebo informace, které jsou pozoruhodnější, zatímco ignorujeme ty, které nepřitahují naši pozornost. Naše predispozice soustředit se na nejpozoruhodnější a emocionálně nejvýraznější detaily nás vede k ignorování potenciálně důležitých

informací, a tím můžeme dojít k nevhodnému rozhodnutí. (desitionlab.com) Některá nebezpečí se nám pak mohou zdát daleko reálnější a pravděpodobnější, než skutečně jsou, a to nás může vést k přehnané reakci.

Kotvení je jev, kdy se jedinec rozhoduje na základě kotvy – tedy předešlé informace, která pak ovlivňuje vyhodnocování dalších informací. (Tversky a Kahneman, 1974) Kotva funguje jako referenční bod, tedy zajišťuje buď potřebné povědomí, které pak zvyšuje důvěru v dané informace, anebo poskytuje autoritu, která zvyšuje důvěryhodnost dané informace. Kotvení ovlivňují dva faktory – povědomí a autorita. Povědomí vyjadřuje úroveň znalosti dané informace nebo jejího kontextu. Druhým aspektem je pak autorita spojená s danou informací. Tento aspekt souvisí s autenticitou a důvěryhodností původce. (Jost, Punder a Schulze-Lohoff, 2020) Kotvení je běžné hlavně na finančních trzích a v marketingu, ale lze je sledovat i jinde. V praxi relevantní k našemu tématu tedy můžeme vnímat efekt kotvení ve chvíli, kdy si jedna dezinformace půjčuje důvěryhodnost od jiné, už zavedené, tím, že z ní vychází, nebo na ni navazuje. Také může navazovat na již známou pravdivou informaci a stavět na ní svůj narativ. Související pojem je pak **sebenaplňující proroctví** – tedy predikce, která se zdánlivě naplní z důvodu důvěry dané skupiny v její pravdivost. Tato důvěra vede k chování, které proroctví pak skutečně naplní, nebo aplikovaní důsledků ex post, které potvrzují předchozí predikce. Zjednodušeně řečeno jedincovo přesvědčení ovlivní jeho chování a vnímání dané situace. (Merton, 1948)

Kognitivní slepá skvrna je jev, kdy si jedinec neuvědomuje vlastní slepé místo v úsudku nebo podléhání bias, které je ale schopen vidět u jiných. Kognitivní slepá skvrna tedy negativně ovlivňuje naše vnímání a ztěžuje nám schopnost vyhodnocovat a opravovat vlastní chyby v úsudku, sami sebe tedy vnímáme jako objektivní. V praxi pak nejsme schopni ověřit a opravit vlastní názor, přesvědčení nebo závěry, které mohou být založeny na neobjektivních a nepodložených informacích nebo vycházet z nějaké heuristiky, protože si je neuvědomujeme. (Pronin, Lin, a Ross, 2002)

Efekt opakovaného vystavení nastává ve chvíli, kdy lidé preferují nějaké informace nebo jim věří pouze proto, protože zní familiárně, jsou jim známé z předchozí zkušenosti. (Bornstein a Craver-Lemley, 2016). Někteří lidé tedy mohou uvěřit dezinformacím pouze proto, že je už několikrát viděli, nebo protože je říká postava někoho, koho už předtím viděli. Do tohoto jevu vstupuje ještě **efekt zmatení zdroje**, kdy si jedinec pamatuje informaci, nicméně už neví odkud, a proto může zpětně i nedůvěryhodnému zdroji přisoudit legitimnost. (Huber, Shiffrin, Quach a Lyle, 2002)

Efekt opakovaného vystavení pak může způsobit i posilování dezinformací a souvisí i s familiarity backfire efektem.

1.2. Backfire efekty

Při nastavování strategií práce s dezinformacemi je nutné brát na vědomí nejen heuristiky výše, ale i koncepty takzvaného zpětného rázu popsané níže, které nám mohou zkomplikovat vyvracení a posílit dezinformace.

Jak již bylo zmíněno výše, **familiarity backfire effect** navazuje na efekt opakovaného vystavení. Tento efekt se objevuje ve chvíli, kdy při vyvracení dezinformaci naopak posílíme tím, že ji opakujeme. Při vyvracení by se tedy mělo vyvarovat opakování dezinformace, protože samotná oprava by mohla sloužit k neúmyslnému zvýšení povědomí o dezinformaci, a tím ji posílit. (Swire-Thompson, DeGutis a Lazer, 2020).

Overkill backfire effect způsobuje, že lidé, kteří se setkají se složitým vysvětlením, jej odmítají ve prospěch jednodušší alternativy a někdy také posilují svou víru v jednodušší alternativu. (Lewandowsky, Cook, Ecker, Albarracin, Amazeen, Kendou, Zaragoza a kol., 2020) V praxi to znamená, že při vysvětlování dezinformací se musíme vyhnout přehnaně složitým vysvětlením, sdělení snažící se vyvrátit dezinformace by měla být jasná, jednoznačná, jednoduchá a krátká. U komplexnějších situací, kdy si vyvrácení žádá složitá vysvětlení, tuto snahu efekt výrazně stěžuje.

World view backfire effect se vyskytuje u témat, která jsou v souladu s pohledem lidí na svět a smyslem pro kulturní identitu. Několik kognitivních procesů může způsobit, že lidé nevědomě zpracovávají informace zaujatým způsobem. U těch, kteří jsou ve svých názorech silně fixní, může konfrontace s protiargumenty způsobit posílení jejich názorů. (Lewandowsky, Cook, Ecker, Albarracin, Amazeen, Kendou, Zaragoza a kol., 2020) Ve chvíli, kdy narativ dezinformací proniká do jejich světonázoru, je jejich vyvrácení velmi obtížně. Případné vyvrácení by totiž mohlo ohrozit jejich celkové vnímání světa a způsobit ztrátu identity.

1.3. Information deficit model

Information deficit model předpokládá, že k chybným závěrům docházejí jedinci z důvodu nedostatku informací, a proto jejich dodání může napravit chybné úsudky. Tento model věří, že informační gramotnost lze zlepšit zvýšeným zapojením odborné veřejnosti. (Miller, 1983)

1.4. Sociální bubliny a clustery

Jedním z důsledků algoritmického fungování sociálních sítí je vznik tzv. sociálních bublin nebo clusterů. Ačkoliv sociální média mohou přinést větší diverzitu informací, přispívají i k polarizaci uživatelů právě tím, že je na základě algoritmického rozřazování a doporučování sdružují do homogenních celků, kde si navzájem utvrzují názory a předávají podobnou informační dietu. Koncept „filtračních bublin" popisuje jev, kdy je automaticky doporučován filtrovaný obsah, s nímž uživatel bude pravděpodobné souhlasit. V důsledku toho se pak vytvářejí „echo chambers", virtuální komory podobně smýšlejících lidí, kteří se navzájem utvrzují ve správnosti svých názorů a zkreslují vnímání reality (Všichni co znám, volí politika X, jak je tedy možné, že získal jen 20 %). (Chitra a Musco, 2019)

Právě volební rozhodnutí, která jsou v rozporu s předvolebními očekáváními, je jeden z důsledků filtračních bublin a echo chambers. (Burbach, Halbach, Ziefle a Calero Valdez, 2019)

Filtrační bubliny a echo chambers jsou pro poskytovatele platforem výhodné, protože jim pomáhají efektivněji generovat personalizovaný obsah a získat na něj pozitivní reakce. (Arguedas, Robertson, Fletcher a Nielsen, 2022)

Toto jsou některé z konceptů behaviorální psychologie, které vstupují do procesu informační války. Z jejich pochopení může benefitovat výzkum a vývoj AI nástrojů zaměřujících se na vyvracení dezinformací. Ovšem mezioborová spolupráce by zde neměla končit.

Odborníci (Kertysova, 2018, Polyakova a Boyer, 2018, Rosenbach a Mansted, 2018, Miller, 2018, Johnson, 2019) se shodují, že jedním z klíčových kroků do budoucna je větší sdílení informací a spolupráce mezi soukromými a státními subjekty a univerzitami a větší investice do výzkumu obecně. Klíčové je pak na hrozby nejen reagovat, ale moci je efektivně předvídat a předcházet jim a efektivně zásahy evaluovat.

2. Umělá inteligence

Pro zajištění věcného výzkumu a diskuse je potřeba formulovat základní mechanismy umělé inteligence a dostupné technologie. Obecné pochopení procesu AI stejně jako popsání současného stavu je nezbytnou součástí případné spolupráce, a proto se mu věnuje i tato kapitola. Výčet opět považuji za vhodný z důvodů zajištění základního porozumění tématu, ačkoliv na něj není výslovně navazováno, výklad literatury se opírá o znalost těchto pojmů a kontextu.

Základní pojmy a nástroje

Umělá inteligence může být chápána jako schopnost systému plnit úkoly charakteristické pro člověka jako učení se a dělání rozhodnutí. (itif, 2018) AI a k ní přidružené technologie jako quantum computing[1], cloud computing[2], big data, miniaturizace komponentů a autonomní robotika v poslední dekádě změnily způsob fungování technologií. Výkony výpočetní techniky, rozšiřování datasetů a jejich větší dostupnost, větší implementace machine learning a rozšíření komerčního zájmu a investování do AI jsou hlavní hnací silou prudkého vývoje umělé inteligence. (Johnson, 2019) Tento vývoj se samozřejmě promítl i do způsobu, jakým spolu velmoci soupeří a vyvíjí na sebe vzájemně tlak a je znatelný i ve vývoji vojenské techniky. Do procesu vývoje a implementace AI vstupují i otázky etiky, legislativy, rizika, ekonomiky a technické aspekty a predikce. Umělá inteligence nachází uplatnění v různých oborech. Na možnosti jejího využití právě ve strategickém boji v informační válce se zaměřuje tato práce. Největší možnost uplatnění je v online prostředí. Nejzásadnější vlastnosti AI představíme v následující sekci.

1 Quantum computing je oblast informatiky zaměřená na vývoj technologií založených na principech kvantové teorie. Kvantové výpočty využívají jedinečné chování kvantové fyziky k řešení problémů, které jsou pro klasické výpočty příliš složité. Vývoj kvantových počítačů znamená skok vpřed ve výpočetních schopnostech s potenciálem masivního zvýšení výkonu ve specifických případech použití. Očekává se například, že kvantové výpočty budou excelovat v úkolech, jako je celočíselná faktorizace a simulace, a vykazují potenciál pro použití v průmyslových odvětvích, jako je farmacie, zdravotnictví, výroba, kybernetická bezpečnost a finance.) (Gillis, techtarget.com)

2 Cloud computing je dostupnost zdrojů počítačového systému na vyžádání, zejména úložiště dat (cloudové) a výpočetní výkon, bez přímé aktivní správy uživatelem] Velké cloudy mají často funkce rozmístěné na více místech, z nichž každé je datovým centrem. (Montazerolghaem, Yaghmaee a Leon-Garcia 2020).

Machine learning neboli strojové učení může být obecně definováno jako používání algoritmu a velkých datasetů k trénování počítačových systémů, aby rozeznaly vzorce, které předtím nebyly detekovány, a schopnost systému se z těchto dat poučit a rozeznat cenné informace, aniž by mu to muselo být speciálně zadáno. (Fisher, 2019) Jedná se o schopnost počítačů analyzovat velké množství dat, rozpoznávat vzorce, učit se z nich a poté předvídat, nebo jednat bez lidského programování. Stroje jsou tak „vycvičeny“ k plnění úkolů bez lidského zásahu. (Operand, 2016)

Další vlastností AI je **rozpoznávání vzorců,** což umožňuje z označeného obsahu selektovat opakující se jevy, vyvozovat vzorce, ty následně vyhodnotit a použít jako předlohu pro hledání a označování jiného podobného obsahu. (Faesen a kol., 2019)

Důležitým nástrojem je i **text mining** nebo také textová analýza. Jedná se o proces získávání kvalitativních informací z textu, objevení nových, dříve neznámých informací automatickým skenováním různých písemných zdrojů. Klíčovým prvkem je spojení extrahovaných informací dohromady za účelem vytvoření nových objevů, smyslů nebo nových hypotéz, které lze dále zkoumat konvenčnějšími způsoby experimentování. „Dolování” textu se liší od toho, co známe z vyhledávání na webu v prohlížeči, je intuitivnější. Text mining je variací data miningu, která se snaží najít zajímavé vzory z velkých databází. (Hearst, 2003)

2. 1. Shrnutí úvah a predikcí ohledně nástrojů AI v kontextu informační války

Jak již bylo nastíněno výše, umělá inteligence představila mnoho nových nástrojů, které mohou být použity i pro tvorbu a šíření informací hodnocených dle určitých kritérií (většinou nastavených tvůrci nástroje nebo poskytovateli platformy) jako dezinformace nebo fake news. Umělá inteligence změnila způsob, jakým jsou dezinformace vytvářeny, šířeny a reprodukovány a napomohla k jejich růstu. (Kertysova, 2018)

Stejně jako ve většině odvětví i zde se vlivem vzestupu umělé inteligence a technologického pokroku rýsuje nový účinnější a efektivnější přístup. „Technologické pokroky v oblasti umělé inteligence, automatizace a machine learnig v kombinaci a větší dostupností big data připravily půdu pro éru sofistikovanějších a levných nástrojů politického boje velkého dopadu," (Polyakova a Boyer, 2018) Vzniká nová forma propagandy – počítačová.

Počítačová propaganda pojmenovává nové odvětví šíření digitálních misinformací a manipulací. Popisuje užívání algoritmů, automatizací a lidský dohled / kurátorství k cílenému a úmyslnému šíření zavádějících informací v prostředí sociálních sítí. (Wolley a Howard, 2016) Vývojáři a jejich automatizované softwary (boti) se na chování běžných uživatelů naučí jejich vzorce vystupování a konzumování, a pak tuto znalost využívají k ovlivňování názorů a veřejnosti. Boti se chovají jako běžní uživatelé a pracují na zvýšení nebo snížení zásahu vybraného obsahu. (Woolley a Howard, 2018) Pro dosažení optimálních výsledků mohou být kombinováni s aktivitou reálných uživatelů / trollů. Hlavní obrana poté leží v detekci.

2.1.1. Detekce

Soukromé společnosti vykazují úspěch při efektivní detekci botů, trollů a dalšího manipulativního a zavádějícího obsahu. Některé společnosti nejen detekují počítačovou propagandu, ale jsou schopny ji sledovat, vyšetřovat a reagovat na ni v reálném čase. Jelikož se její šíření zakládá na algoritmu a pohybuje se ve vzorcích, je postupem času možné i její předvídání. Proto se pohyb dezinformací na internetu musel změnit. Nově se šíření dezinformací pohybuje v méně detekovatelných vzorcích, sdílení daného obsahu probíhá skrz více účtů a zařízení, která sdílí tutéž zprávu v předem určený čas, a vše je zveřejňováno nadlidskou rychlostí. Tato automatizace slouží hlavně k tomu, aby sdílela a rozšířila zprávu z „živého účtu", tedy z účtu, jenž ovládá živá osoba, nikoliv bot. (Demartiny, Mizzarato a Spina, 2020) Díky tomu se dosah tohoto účtu zmnohonásobí a stoupá mu i relevantnost – čím více příjemce vidí danou zprávu přicházet z více zdrojů

(více botů), tím se mu zdá důvěryhodnější. Také tu hraje roli efekt opakovaného vystavení: respondent tedy věří tomu, čemu byl vícekrát vystaven.

Ačkoliv sociální sítě, které jsou přirozeným prostředím pro tento druh počítačové propagandy, si stále při detekci a označování tohoto obsahu neví rady (nebo také nechtějí vědět rady), soukromé subjekty jsou v detekci těchto aktivit relativně úspěšné, díky machine learning a AI monitorování synchronizované aktivity botů a extremistických obsahů jsou schopné tento příspěvek najít a pokusit se ovlivnit content raking.[3] Tato schopnost detekce se ovšem snižuje zapojením botů společně s živými uživateli, ti totiž nabourávají algoritmus a vzorce, zkreslují je a tím schopnost machine learning ztěžují. (Demartiny, Mizzarato a Spina, 2020)

2.2. Nástroje počítačové propagandy

2.2.1. Deepfakes

Jedním z nástrojů počítačové propagandy a zároveň vlastním druhem dezinformací jsou tzv. deepfakes. S nimi se hranice snadného rozlišení mezi pravdou a fikcí v audio a audiovizuálních materiálech ztrácí.

Hrozba v podobě deepfakes je evidentní. Jedná se o schopnost vytvářet digitálně manipulovaný zvukový nebo vizuální materiál, který je vysoce realistický a prakticky nerozeznatelný od skutečného materiálu – původně byl používán ve filmovém průmyslu. V dnešní době nacházejí své uplatnění v online sféře zábavy, klamání spotřebitelů, a dokonce i v politice a mezinárodních záležitostech. Software na jejich vytváření je volně dostupný. (Bayer a kol., 2019) Jediné, co je k vytvoření tohoto obsahu potřeba, je dostatečný objem audiovizuálních materiálů dané osoby, podle kterého se pak může namodelovat daný podvržený obsah. V současné době je umělá inteligence schopna identifikovat tato videa na základě nedokonalostí v obrazu (konkrétně nedůvěryhodné

3 *Content raking – ukazuje, jak je váš obsah viditelný na různých platformách. To také určuje jeho relevanci, která pak určuje pravděpodobnost organického šíření tohoto obsahu. Jedná se o žebříček seřazený podle vašich domnělých preferencí obsahu.*

mrkání; zavřené oči totiž nebývají moc zaznamenány, a proto je modelace a replikace mrkání složitá). Ovšem v rámci vývoje a rozpoznávání deepfakes hrají aktéři hru na kočku a myš. Pokaždé, když nalezneme systém odhalování deepfakes, ukazatele podvrhu, její tvůrci jej opraví a daný ukazatel odstraní, každé odhalení chyby v deepfakes tak vede k jeho zdokonalení. (Kertysova, 2018)

Deepfakes jsou v současné době na vzestupu, technologie, dříve sloužící k humorným videím a parodiím, nyní generují přesvědčivý, autentický a snadno uvěřitelný obsah, pro běžného uživatele je mnohdy těžko rozeznatelné, které video je falešné a které ne. Zároveň dochází k více nuančním alternacím, například je pouze mírně upravené audio. Tyto malé alternace obsahu pak přispívají k obecné relativizaci pravdy. Autentický obsah může být uživateli vyhodnocený jako podvrh, jelikož jejich skepticismus stoupá a pravda na sociálních sítích podléhá intenzivní relativizaci. Oba narativy v informační válce také mohou používat argument, že zvuková nebo audiovizuální stopa je podvržená, a jejich publikum jim bude věřit.

2.2.2. Zneužití big data

Dalším nástrojem zneužitelným v informační válce je snadné získávání a následné zneužívání velkého množství detailních datasetů uživatelů sociálních sítí za relativně nízkou cenu.

Sociální sítě denně sbírají obrovské množství rozmanitých dat svých uživatelů a následně je prodávají dalším subjektům pro marketingové účely – to je hlavní model sociálních sítí, díky němuž jsou provozovatelé schopni vydělávat i na platformě, která je zadarmo. Ačkoliv sociální sítě jsou schopné v dnešní době objevit podezřelé manipulace s daty (podezřelé nákupní firmy, nastavené reklamy, napojení na vládní agentury) a odhalit společnosti/uživatele, kteří aktivně šíří dezinformace nebo polarizující obsah, mnoho z nich si volí cestu ignorování problému, protože ta je pro ně finančně nejvýhodnější, a svoje počínání pak schovávají za svobodu slova. (Mustak, Salminen, Mäntymäki, Rahman a Dwivedi, 2023) Ačkoliv data svých uživatelů neprodávají přímo Rusku nebo

Číně, dávají možnost ostatním firmám napojeným na tyto země nakupovat inzerci na jejich platformách a tím je pouští ke svým datům. Z těchto dat pak získávají informace nejen o preferencích uživatelů, ale i o vzorcích jejich reakce, co v nich vyvolává jakou emoci, a jak tuto emoci nejlépe vyvolat. (Prier, 2020) Tento proces zneužití big data a microtargeting je pozorovatelný na aféře Cambridge Analytica. (viz například: Hu, 2020)

Kombinací dostupných a laciných dat a rychlého vývoje AI nástrojů tak vzniká relativně levná a časově nenáročná možnost vytvořit personalizovanou, sofistikovanou a úspěšnou dezinformační kampaň téměř odkudkoli.

2.2.3. Microtargeting

Dalším aspektem, jež AI přináší, je schopnost analyzovat obsah uživatelů a následně je microtargetovat a zasáhnout s unikátní zprávou vytvořenou na míru uživateli. Skrz tuto zprávu získá požadovanou emocionální reakci, kterou si dopředu zvolil jako cíl a jejímuž dosažení přizpůsobil zprávu (pozitivní x negativní).

Microtargeting, nebo také hypercílení, je využití podrobných údajů o uživatelích a automatizace k vytvoření přesně cílených a vysoce personalizovaných sdělení napříč velkým počtem kanálů. Tyto kampaně jsou navrženy tak, aby oslovily konkrétní osoby nebo malé skupiny konzumentů a přednesli jim relevantní a na míru připravený obsah. Díky big data a inovacím, jako je prediktivní analýza, mohou aktéři získat hlubší porozumění publika. (Semerádová Weinlich, 2019)

Skutečný vliv microtargetingu na rozhodnutí uživatelů je však stále předmětem výzkumu.

2.3. Služby AI v praxi

Stejně jako mohou být AI nástroje využity pro vytvoření personalizovaného falešného obsahu a posílit šíření falešných zpráv, které má vysoký potenciál ovlivňovat rozhodnutí a postoje cílového publika, mohou být stejná data a algoritmy využity pro detekci a kontrolu šíření tohoto zavádějícího obsahu. (Demartini, Mizzaro a Spina, 2020) Právě praxi těchto nástrojů představí tato kapitola.

Nástroje umělé inteligence zaznamenávají největší úspěchy v detekci a odstraňování ilegálního, pochybného nebo nechtěného obsahu a identifikaci falešných účtů ovládaných boty (bot spotting and bot labelling). (Rosenbach a Mansted, 2018) Nicméně i zde je potřeba prohloubit výzkum a vývoj, aby byla zajištěna větší efektivita a spolehlivost. V současné době se tyto nástroje svojí sofistikovaností nevyrovnají například technologii pro filtrování spamu v e-mailech. (Kertysova, 2018)

Dnes nejvíce využívané AI nástroje pro detekci dezinformací na sociálních sítích plní funkci fact-checkingu. Jedná se o nejrozšířenější a pro veřejnost nejviditelnější způsob využití možností umělé inteligence. Hlavní výhodou je rychlost. Jak již bylo představeno, nepravdivý obsah na internetu je násobený aktivitou botů, a tak je schopen rychle doputovat do nejzazších koutů sociálních sítí, proto je potřeba stejně rychlá a automatizovaná odpověď. AI nástroje tedy obrovskou rychlostí skenují veškerý obsah na sociálních sítích a označují potenciálně závadné příspěvky. Problém ovšem nastává v dalším kroku – AI nástroje totiž zatím neumí vyhodnocovat a pronášet finální rozhodnutí o tom, zda je obsah zavádějící, nepravdivý, jedná se o satiru nebo vtip, nezvládá rozlišit nuanci nebo intenzitu. (Strickland, 2018) V tuto chvíli do celého procesu vstupují lidé.

2.4. Lidé v procesu AI

Automatické vyhodnocování spolehlivosti a důvěryhodnosti informací je komplexní problém, který do velké míry závisí právě na lidech v tomto procesu, přičemž míra jejich expertízy se liší, jsou zde zastoupeni proškolení odborníci i crowdsourcing. Právě kombinace těchto nástrojů je determinována jednak finančním aspektem, jednak pak

spolehlivostí. Společnosti se snaží najít optimální kombinaci nástrojů a aktérů v procesu. (Demartini, Mizzaro a Spina, 2020) Přístupu se také říká Human-in-the-loop neboli HAI.

Nástroj umělé inteligence například projde velké množství textu a označí podezřelý obsah. Výrazně tak sníží množství obsahu, které půjde na překontrolování odborníkům. Lidská sféra je zde zastoupena odborníky na fact-checking, kteří pak veškerý označený obsah projdou a vyhodnotí jeho tón a závadnost. Pokud tato osoba vyhodnotí obsah jako zavádějící, označí ho a algoritmus sociálních sítí ho následně umístí na nižší úroveň uživatelova feedu. Tento postup tak může snížit dosah příspěvku až o 80 %. (Strickland, 2018)

2.4.1. Různé přístupy v praxi

Přístup zmíněný výše najdeme například u Facebooku, který k tomu využívá machine learning – tedy techniku, při které AI systém vezme obrovskou porci dat, a na základě jejich chování vytváří, nebo najde vzorce, které mu pak pomáhají vyhodnocovat nové informace. Facebook pak vyhodnocuje data jako zdroj informace – má uživatel tendence sdílet neověřený a zavádějící obsah? Má hodně označených příspěvků? Jaká je pravděpodobnost sdílení dezinformace u tohoto uživatele? V rámci samotného obsahu AI nevyhodnocuje konkrétní význam, ale hledá signály, které se často objevují u zavádějících obsahů, nebo vyjádření údivu v sekcích komentářů. (Strickland, 2018) Někdy se však může jednat i pouze o slova, například při pandemii covidu byla většina příspěvků zmiňujících covid i ve formě satiry nebo humorné příhody označena štítkem a pro ověření informací odkazovala na oficiální stránky zdravotnických institucí.

Podle Facebooku 99,5 % odebrání obsahu souvisejícího s terorismem, 98,5 % mazání falešných účtů, 96 % označování nahoty a sexuální aktivit dospělých a 86 % odebrání grafických reklam souvisejících s násilím je odhaleno nástroji umělé inteligence – nikoli samotnými uživateli – které jsou vyškolené daty od svého lidského moderátorského týmu. (Marsden a Meyer, 2019)

Společnost Factmata oproti tomu při označování dezinformací nevyužívá vzorců získaných v minulosti (například posouzení zdroje na základě historie sdílení), ale posuzuje každý příspěvek jako samostatně stojící. Jejich hybridní přístup se zakládá na tom, že člověk označí zavádějící obsahy a přidá k němu nějaké nálepky – detaily či důvody k odstranění, a ty pak putují k AI nástroji, který se pomocí těchto příspěvků trénuje k efektivnějšímu a přesnějšímu detekování politicky zaujatého, nepravdivého nebo nenávistného obsahu. (Factmata) Jejich přístup je tak založený na dodávání kvalitních training dat, která vytváří a sbírají „ručně".

AI nástroje současně prostupují i do odvětví žurnalistiky. Londýnská společnost Krzana pomáhá novinářům hledat „breaking news" podle zvolené customizace. Novináři tak mohou naleznout zpravodajské informace podle zvolených klíčů a filtrů, které si vyberou. Pomocí tohoto nástroje lze pak hledat i dezinformace – novináři je naleznou a označí, a tak se zpráva rychle vyřadí z šíření mezi ostatní a není přebírána dalšími mediálními domy. (Strickland, 2018)

2.4. 2. Fungování a limity HAI přístupů

Jelikož hlavním limitem machine learning, AI nástrojů a automatické detekce je právě zmíněná satira nebo komunikace s vysokou mírou specifického kontextu, většina aktérů se uchyluje k hybridnímu přístupu. Role člověka bývá dozorčí a vyhodnocovací: automatizovaný systém odvede většinu třídění a označování a člověk pak provede finální soud. Přístup s takto rozloženou odpovědností mezi stroj a lidi se nazývá human-in-the-loop.

HAI systém spoléhá na AI nástroje při vyhodnocování velkého množství dat a na lidský faktor při vykonávání komplexních a složitějších úkonů v procesech, například vyhodnocování férovosti (zda si obsah zaslouží být odstraněn, u AI nefunguje možnost přezkoumání, tuto funkci plní lidský faktor) nebo porozumění jazyku. Tento přístup se také využívá při aktivním učení, při němž jsou nálepky a štítky shromažďovány a

vyhodnocovány lidmi, následně pod dohledem přidány do machine learning modelu a poté se kontrolují výstupy, na jejichž základě se rozhodne, jaké štítky musí lidé pojmenovat a dodat příště. Tento přístup je vyhodnocován jako účinnější než čistě AI přístup, ale nese s sebou další limity jako například kontrolu a zajištění efektivity a konzistence. Zapojení lidí může přinést šum v označování a štítkování. Tento šum se pak může násobit a zvětšovat, pokud pronikne do machine learning modelu. (Mohseni a kol, 2021)

I samotné zapojení lidí není univerzální. Do procesu vstupují jednak experti – tedy lidé vzdělaní a trénovaní ve fact-checkingu – a dále pak neexpertní pracovníci, kteří jsou rekrutováni pomocí crowdsourcingu. Právě kontrola kvality crowdsourcingu je jednou z největších výzev tohoto přístupu – pokud neproběhne, jak má, riziko šumu se zvětšuje a může případně znehodnotit celý model.

Bias

Bias lze popsat jako lidskou tendenci upřednostňovat určité informace před jinými, což se pak odráží ve vytváření skrytých i zjevných předsudků vůči dané skupině. Největší riziko bias je v jeho uvědomění, funguje na pozadí, ovlivňuje nás bez povšimnutí a jeho význam je na první pohled nepatrný. V prostředí internetu se odráží jak společenské bias, tak i individuální vnitřní, které jsou rozdílnější. Sociální sítě pak mohou tyto bias ještě posilovat. Lidé s sebou tedy do HAI procesu můžou přinést různá bias, která pak přenesou i do označování a nálepkování a tím „nakazí“ machine learning model a celý algoritmus.

První skupina HAI bias je způsobena demografickými údaji lidí v procesu. Druhá skupina bias je způsobena nedostatkem reprezentace všech skupin v počátečních datech, ta pak neodráží realitu. Zkreslení pak mohou přinést i samotní designéři algoritmu. „Tag recommendations", neboli štítkování a tagování položek, je jedním z příkladů možného bias promítnutého do algoritmu, protože je přímo ovlivněno preferencemi. (Baeza-Yates, 2018).

Z toho vyplývá, že i když jsou systémy umělé inteligence navrženy tak, aby prováděly analýzu bez zkreslení, lidské zkreslení vyplývající z výběru vzorků dat, typů označení a dalších nekontrolovatelných faktorů může přesto vést k subjektivnímu rozhodování. (Johnson, 2019) Například pokud jsou na fotce lidé v lékařském oblečení, šance, že moderátor označí ženy jako sestry namísto doktorky jsou daleko vyšší a neodráží realitu. (Demartini, Mizzaro a Spina, 2020) Tento bias pak může proniknout přímo do modelu, který pak bude také vyhodnocovat ženy v lékařském oblečení jako sestry a nejspíše bude tento bias i amplifikovat, protože ho bude rozeznávat jako vzorec a ne jako bias.

Přítomnost bias ve vzorci pak může vést k tzv. Uknown unknows[4] – jedná se o chyby v datech, u kterých AI modely dělají sebevědomá a jistá klasifikační rozhodnutí, která jsou ale špatná. Model si tedy není vědom toho, že by prováděl chyby ve vyhodnocování. UU jsou těžko identifikovatelná právě kvůli jistotě modelu o jejich správnosti, a tím vzniká jeden z kritických limitů tohoto přístupu. Jedním ze známějších příkladů tohoto problému je situace z roku 2015, kdy algoritmus vytvořený pod společností Google, používaný na rozpoznávání fotek, označoval lidi černé pleti jako gorily a absolutně si této chyby nebyl vědom. (Theverge, 2018) K nápravě došlo tak, že Google zkrátka zakázal algoritmu rozeznávat a označovat gorily úplně. Nemůže tedy dojít k chybě, protože systém od vyhodnocování úplně upustil. Systém tedy problém ignoruje, namísto aby jej opravil, protože oprava je příliš komplexní a nespolehlivá. Google tedy zvládl identifikovat různé druhy opic kromě goril a šimpanzů. (theverge, 2018) Tento problém je způsoben nedostatkem vyváženosti reprezentace a různorodosti dat v tréninkovém datasetu. Jedním ze způsobů, jak tomu předcházet, je zajistit vyváženou reprezentaci všech vlastností v tréninkovém datasetu a vytvoření vhodné strategie pro získávání a třídění dat lidmi. (Demartini, Mizzaro a Spina, 2020)

4 Unknown unknowns nalezneme v různých odvětvích, ale v oblasti informačních technologií vystupují u big data, text miningu, návrhu softwaru a kybernetické bezpečnosti. Obecně se jedná o data / problémy, u kterých nejen nevíme, jak je řešit, ale nejsme si ani vědomi jejich existence. UU jsou také velkou součástí risk managementu.

Finance

Dalším limitem hybridního přístupu jsou finance. Manuální anotace má většinou nízký budget, a proto je nutné pečlivě vyselektovat data, která budou kvalifikovanými lidmi zkoumána manuálně. Zároveň je nutné vyhodnotit, která část bude podléhat vyhodnocování experty a která poputuje k neexpertním týmům. Lidská kontrola je nejen nákladná, ale je také náchylná k chybám a nejednoznačným výsledkům, zejména pokud ji provádí neexpertní pracovníci, analýzu pak může ovlivnit něčí původ, osobní étos, nebo dokonce nálada v daný den. Obsah také bývá v praktickém prostředí vyhodnocován a označován lidskými moderátory, kteří navazují na závěry AI nástrojů.

Podstatným rizikem při kódování obsahů je to, že lidští moderátoři obsahu často pracují ve vysoce stresujících podmínkách, v časové tísni a mohou mít potíže vyrovnat se s traumatizujícími obrázky a videi. Bez adekvátního školení a podpory, jak se vypořádat s extrémním obsahem, se u moderátorů mohou rozvinout příznaky podobné PTSD, které mohou ovlivnit jejich schopnost vykonávat každodenní činnosti. (Newton, 2019)

V současné době sice nepanuje názor, že by snad umělá inteligence mohla provádět lepší rozhodnutí než lidé, je nicméně podstatně levnější a rychlejší, takže i když nemusí dosahovat kvalitou úrovně lidských rozhodovatelů, v kvantitě je předčí. Navíc umělá inteligence nepodléhá (pokud k tomu není vědomě či nevědomě naprogramovaná) chybám ve strategickém rozhodování, jako je sunk cost trap,[5] kognitivním heuristikám[6] nebo groupthinku.[7]

5 Pokračování s aktivitou, která se nevyplácí, protože jsme do ní už investovali a nechceme o tuto investici přijít.
6 Kapitola 1.1.
7 Způsob myšlení, ve kterém jednotliví členové malých soudržných skupin mají tendenci přijmout stanovisko nebo závěr, který představuje vnímaný skupinový konsensus, ať už se členové skupiny domnívají, že je platný, správný, nebo optimální. Groupthink snižuje efektivitu kolektivního řešení problémů v rámci takových skupin. Groupthink byl pozorován na sérii špatných rozhodnutí v zahraniční politice ze strany USA, počínaje Pearl Harbourem a vrcholil událostmi v Zátoce Sviní. (Britannica Dictionary, 2023)

2. 5. Obecné prognózy a potřeby pro efektivní obranu v informační válce

V dnešní době už je jasné, že vytvářet strategie pouze v reakci na individuální útoky a odpovídat improvizovaně nepomůže s dlouhodobým a plošným řešením problému. V rámci strategického plánování je tedy třeba věnovat pozornost i prognózám, legislativě a etice AI.

2.5.1. Spolupráce

První potřebou do budoucnosti je větší spolupráce v rámci transatlantického sdílení informací a know-how. Aktéři, ocitající se na druhé straně dezinformačních útoků, musejí více spolupracovat při plánování, výzkumu a synchronizované odpovědi na hrozby. (Kertysova, 2018) Další nutností při sestavování obranyschopné strategie je vzájemné sdílet informace se soukromými firmami a zlepšit bezpečnost informací a transparentnost.

2.5.2. Postoj soukromých subjektů

Soukromé společnosti také musí efektivněji řešit nakládání s vystopovanými dezinformacemi. Pouze jejich označení jako dezinformace bohužel nestačí k zabránění uživateli, aby na takový obsah kliknul. Je proto nutné tyto příspěvky i penalizovat v rámci algoritmu, snižovat jim content raking, a tak zamezit organickému zobrazování dalším uživatelům, nebo je přímo ztlumit – příspěvky se tak ukážou, pokud je někdo bude přímo hledat, ale nebude možné na ně narazit náhodně nebo na základě doporučení ostatních uživatelů. (Hellman a Wagnsson, 2017)

Součástí by také měla být revize a následné blokování inzerentů. Ti, kteří se aktivně zapojují do podpory šíření propagandy, by měli být automaticky diskvalifikováni z možnosti zakoupit reklamní prostor a z přístupu k datům uživatelů.

Privátní sektor by měl také implementovat etické kódy, ty mohou být sestaveny asociacemi nebo jinými profesními sdruženími nebo neziskovými organizacemi, které by upozorňovaly nejen odbornou veřejnost – v České republice se o to v současné době pokouší AKA, která vyzývá značky, aby neinzerovaly na závadných webech; jejich hlavní síla je v reputaci a profesionální uznávanosti. Dalším subjektem na českém trhu zabývající se inzercí je organizace NELEŽ, která sestavuje seznamy dezinformačních webů. Jejich cílem je tyto weby odříznout od peněz inzerentů, komunikují se zadavateli inzercí a agenturami, či vedou osvětu ohledně brand safety (která může být ohrožena právě spojením s dezinformačními weby). Jejich hlavní silou jsou vztahy se zadavateli a prezentování rizik pro společnosti. Aktivita těchto profesních a neziskových spolků je důležitá i z důvodu stanovování etických mantinelů, které zároveň testuje v praxi, upravuje je podle prostředí a jeho reakce a následně z nich tak mohou vznikat legislativní předpisy. (NELEŽ, 2023)

2.5.3. Legislativa a etika

Potřeba pro legislativní rámec, konkrétně například omezení aktivit obchodníků s daty, je dalším bodem k projednání budoucího vývoje. Podle některých odborníků by uživatel měl mít přístup ke svým datům, respektive přehled o povaze dat, které o něm překupník má, a možnost tyto data opravit. Zároveň by měl uživatel dostávat upozornění pokaždé, když jsou jeho data sdílena s někým dalším. (Napoli, 2019)

Právě na legislativním a etickém vymezování by měly společně obory spolupracovat. Formování hrozby, popis jejího fungování a chápání následků pro uživatele by v sobě mělo spojovat poznatky IT odborníků, právních znalců a znalců sociálních věd a behaviorální psychologie. Jedině tak můžeme v plném rozsahu pochopit, co se děje, a jak tomu nejefektivněji zabránit, ať už legislativním zásahem, nebo vytvořením obranných protokolů pro implementaci.

Při debatách o obraně proti dezinformacím se často formuje i myšlenka kontrapropagandy. Případná kontrapropaganda by ovšem mohla erodovat víru v ideály

státu, proto nemá z pohledu demokratických států smysl a mohla by se nepříjemně vymknout kontrole. (Hellman a Wagnsson, 2017)

Etika by také měla prostoupit kurikulem IT oborů a měla by být povinnou součástí studia. Nová generace odborníků by tak měla umět přemýšlet nad dopadem svých aktivit a o etických hranicích. Jak již bylo zmíněno, algoritmy jsou tvořeny lidmi, a přebírají do sebe i jejich předpojatosti a bias, neutrální algoritmus je tedy iluze. Ve chvíli, kdy si tvůrce je více vědom etických důsledků kódování, může vzniknout algoritmus, který je méně náchylný k manipulaci. (Machálková, 2023)

Důsledky bias můžeme pozorovat například u některých brand safety nástrojů pro vyhodnocování rizika spojeného se spoluprací s influencerem. Software Peg používá k vyhodnocování rizika influencerů monitoring „Špatných a vulgárních výrazů“, skóre „konzistence publika“ (konzistentní příjem ze strany jejich publika) a skóre kontroverznosti, které měří, zda jsou influenceři „v tisku negativně prezentováni nebo spojováni s kontroverzními tématy“. Systém poté monitoruje výrazy, které podle vnitřních kritérií vyhodnocuje jako „závadné”, ale v nichž jsou zahrnutá i slova spojená s LBTQ komunitou nebo rasové výrazy. Tato slova nemusí nést nutně negativní nebo hanlivý význam, je u nich důležitý kontext a rasová nebo demografická příslušnost jejich uživatele. Bohužel tento kontext není algoritmus schopen rozeznat, a to následně vede k perzekuci určitých skupin tvůrců, ačkoliv neudělali nic špatně. Zároveň jsou některé systémy na monitoring influencerů nedostatečně citlivé na antisemitismus, protože zde nelze nastavit kódově závadná slova, ale je nutno sledovat kontext. (Bishop, 2021)

2.5.4. Uživatelé

V současné době většina detekčních nástrojů spoléhá na uživatele, jejich schopnost falešný obsah rozpoznat a nahlašovat. Tato primární selekce pak uživateli usnadňuje filtrování a vyhodnocování obsahu AI nástroji. Ovšem mnoho sociálních sítí začíná prozkoumávat možnosti filtrování contentu bez zapojení uživatelů, z důvodu hrozby snížení citlivosti na falešný obsah. Uživatelova schopnost rozpoznat tyto falešné zdroje

od reálných se bude snižovat, oproti tomu jeho názory se budou stupňovat a upevňovat konfrontací s obsahem, který bude podporovat jeho názor a zároveň démonizovat příslušníky opačného názorového spektra. (Caramancion, 2020) Vzhledem k útokům na emoce bude docházet k radikalizaci a dalšímu rozdělování společnosti, následnou ztrátu nebo přeměnu národní identity, ze které bude benefitovat typicky např. Rusko. (Lanoszka, 2016) Tato ztráta národního cítění pak dělá jedince náchylnější k podlehnutí ruskému narativu a zvolení si zájmu jiného národa nad ty vlastní.

2.5.5. Multidisciplinární výzkum

V neposlední řadě je také důležitá podpora vývoje a výzkumu AI a počítačové propagandy. Vláda, soukromé subjekty i neziskové organizace by měly investovat do akademického výzkumu, který zodpoví otázky, jak bude AI technologie ovlivňovat veřejný diskurz, jaké s sebou přinese negativní a pozitivní důsledky, jaká bude přesná povaha těch negativních.

V rámci dezinformací by se měla zkoumat jak oblast poptávky, tak dodávky, proč je o dezinformace zájem, co je na nich atraktivního, a proč se šíří. Technologie a AI jsou pouze prostředky, ale abychom odhalili jejich plný potenciál, musíme pochopit sociální psychologii dezinformací. Také je nutné zkoumat, jaké jsou překážky aktérů, například médií, pro efektivnější zapojení a využívání těchto nástrojů při své činnosti. Toto poznání pak může pomoci nejen při vývoji AI nástrojů, ale i dalším aktérům jako jsou média, politici či vláda.

Mezioborová spolupráce pak může pomoci například i při udělování sankcí. Jejich velikost se v rámci útoků v digitálním prostředí politikům těžko vyhodnocuje, a proto je nutné zapojit odborníky do vyčíslování škod.

2.6. Limity a hrozby kontextu AI

Umělá inteligence s sebou přináší i obavy spojené s jejím vývojem a používáním. Mezi největší hrozby patři nedostatek férovosti algoritmu / bias kódu způsobený biased daty, personalizace obsahu, která následně způsobuje částečnou informační slepotu (vznik sociálních bublin), narušení soukromí uživatelů, potencionální manipulace s uživateli, nebo manipulace audiovizuálního obsahu bez uživatelova svolení a vědomí. (Kertysova, 2018)

2.6.1. Limity chápání

Nástroje a využití AI v praxi má zatím značné limity, které bude nutné v následujících letech překonat. Prvním z nich je riziko nadměrného blokování legitimního a přesného obsahu. AI technologie je stále náchylná k falešným negativům/pozitivům, může tedy identifikovat obsah a účty jako falešné, i když tomu tak není. Falešná pozitiva mohou negativně ovlivnit svobodu projevu a vést k cenzuře legitimního a spolehlivého obsahu, který je strojově nesprávně označen jako dezinformace. (Marsden a Meyer, 2019) Současné AI nástroje totiž zatím umí identifikovat pouze jednoduché deklarace, ale v komplexnějších a složitějších prohlášeních se stávají nespolehlivé a omylné.

Ke stejnému limitu dochází v situacích, kdy je třeba chápat kontextové nebo kulturní stopy, lingvistická specifika, kulturní a politické prostředí specifické pouze pro určité země. (Graves, 2018) Následky nepochopení kontextu jsme mohli pozorovat na případu, kdy Facebook označil fotografii Napalm Girl, která vyhrála v roce 1972 Pulitzerovu cenu, jako závadnou a zablokoval ji z důvodu dětské nahoty. (Gillespie, 2018)

Dalším limitem je samozřejmě už zmíněná neschopnost chápat nuanci a sarkasmus. Stejně tak limituje spolehlivost AI již zmíněné promítání bias a osobních vlastností do kódu, což vede k výsledkům, které určité skupiny favorizují (způsobeno přenesením z designera nebo ze špatných training dat).

2.6.2. Chyby designu

Dalším limitem jsou pak chyby, které s sebou přináší určité volby při vytváření algoritmu. Dobrou ilustrací toho, do jaké míry mohou volby architektury a designu ovlivnit polarizaci a dezinformace, poskytuje WhatsApp. Na této platformě jsou zprávy end-to-end šifrovány, a proto jsou již z podstaty mimo dosah moderátorů obsahu. V zemích jako Indie, je WhatsApp nejen hlavním kanálem pro politické kampaně, ale také kanálem pro fake-news, nenávistné projevy, o kterých je známo, že podněcovaly násilí a zabíjení související s mafií. (Safi, 2018) Přeposílaný obsah v sobě nenese žádné informace o původci, a proto je jeho ověřování a případné potrestání viníků téměř nemožné. Zároveň se díky designu aplikace a šifrování může společnost vzdát odpovědnosti za moderaci a obsah šířený na její platformě. (Kertysova, 2018)

2.6.3. Algoritmus

Složitost a neprůhlednost algoritmu představuje další z limitů AI. Machine learning zahrnuje systémy neuronové sítě a hluboké neuronové sítě, které jsou ze své podstaty „black box řešení“, a jejichž vývoj, založený na samoučení, někdy přesahuje chápání vývojářů, kteří je designují. Složitá logika automatizovaného rozhodování činí algoritmy přesnějšími, ale také je obtížnější vysvětlit, jak generovaly konkrétní doporučení. (Levey a Hugemann, 2017)

2.6.4. Odpovědnost a legislativa

AI samozřejmě zužuje i mnoho existenčních otázek, například jaké by měly být etické a legislativní limity, kdo by měl být zodpovědný za jejich dodržování a kdo by je měl určovat a vynucovat. Řešení umělé inteligence vyvolávají důležité otázky o tom, kdo má nejlepší postavení pro to, aby určil, jaký obsah je legální nebo nelegální, žádoucí nebo nežádoucí. Měly by o pravdivosti a naléhavém odstranění online obsahu rozhodovat

veřejné subjekty (ať už jsou, či nejsou institucionálně propojené s vládami), soudní orgány nebo online platformy? Právě tato problematika je pro naši práci zcela zásadní.

2.6.5. Humans out of the loop

Poslední hrozbou je „humans out of the loop“. Jedná se o strach z následků, který by s sebou mohl přinést případně odstranění lidského dozoru a kontroly z procesu. Roboti s umělou inteligencí by mohli ovládnout celý proces se schopností generovat, přesvědčovat a přizpůsobovat obsah různému publiku. (Rosenbach a Masted, 2018) Základní důvod, proč je nutné udržet v procesu lidský dozor, je legislativní, člověk by měl totiž sloužit jako odvolací orgán, měl by zajistit právo odvolat se proti rozhodnutí AI a získat tak další férové posouzení. Lidé by tedy zajišťovali kvalitní auditovatelnost rozhodnutí, které provádí umělá inteligence. (Marsden a Meyer, 2019)

Zároveň zachování lidí v procesu funguje i jako hlídač procesu vývoje a zajišťuje schopnost odhalit a opravit případné chyby systému. (Panel, 2019)

Case Study: Předpoklady umělé inteligence přispívající k veřejnému blahu

V rámci vývoje umělé inteligence v boji s dezinformacemi vyvstává mnoho etických otázek. Jelikož se tyto nástroje dotýkají filtrování informací a mnohdy předpokládají cenzuru nebo psychologické zásahy na recipienta, je důležité, aby byly podrobeny otázce, zda jsou sociálně prospěšné, zda splňují předpoklady „dobré“ technologie.

V práci „How to design AI for social good: seven essential factors“ (Floridi, Cowls, King a kol., 2020) autoři formulují sedm základních faktorů, které je nutno zvážit. Ačkoliv se práce nesoustředí primárně na umělou inteligenci používanou v boji s dezinformacemi, ale na různorodé nástroje, například na modely implementovatelné ve zdravotnictví, je přenositelná na všechny nástroje pracující se snahou o přínos pro sociální dobro.

AI pro veřejné blaho neboli AI4SG je designování, vývoj a nasazení systémů umělé inteligence způsoby, které předcházejí, zmírňují nebo řeší problémy nepříznivě

ovlivňující lidský život nebo blahobyt přírody, nebo umožňují společensky výhodnější nebo environmentálně udržitelný vývoj. (Floridi, Cowls, King a kol., 2020)

AI modely jsou tvořeny lidskými hodnotami, respektive hodnotami tvůrce, a pokud nejsou zvoleny pečlivě a podrobeny důsledné kontrole, mohou vést ke „good AI turned bad“. Na druhé straně pak případné úspěchy mohou být náhodné a těžko replikovatelné nebo přenositelné na jiné situace. Tento nedostatek pochopení pak vede ke dvou hlavním limitům, kterým v současné době tyto nástroje čelí – jednak zbytečná selhání - a na druhé straně promarněné příležitosti. S tímto problémem mají pak pomoci právě výše zmíněné principy.

Některé etické limity „dobré AI“ jsou evidentní: nepoužívat nástroje například pro pokrok ve vývoji zbraní masového ničení, nebo není-li AI nejefektivnějším řešením dané situace a existuje-li atraktivnější a vhodnější přístup. Obecné předpoklady tedy jsou dobročinnost, neškodnost, spravedlnost, autonomie a vysvětlitelnost. (Floridi, Cowls, Beltrametty, Chatila, Chazerald, Digmum a kol., 2018)

Pro zajištění těchto základních etických mantinelů je zde sedm faktorů, které je nutno zvažovat při navrhování a vývoji AI nástrojů. Splnění těchto předpokladů nezajišťuje, že AI nástroj bude automaticky sociálně dobrý, ale jejich splnění je základním kamenem pro vytvoření AI4SG. Prvním je falzifikovatelnost. Tento faktor zajišťuje důvěryhodnost, neboť bez falzifikovatelnosti nelze provádět empirické testování celku i kritických částí, a tak ověřovat funkčnost a bezpečnost. Falzifikovatelnost nám také pomáhá mapovat limity AI nástroje, tedy zjistit, v jakém kontextu se stává nedůvěryhodnou, a naopak v jakém kontextu se na ni můžeme spolehnout. „To vyžaduje přijetí procesu, ve kterém vývojáři nejprve zajistí, že nejkritičtější požadavky nebo předpoklady aplikace jsou falzifikovatelné, následně provedou testování hypotéz těchto nejkritičtějších požadavků a předpokladů v bezpečných a chráněných podmínkách, a pokud nejsou tyto hypotézy vyvrácené v malém souboru vhodných kontextů, pokračují v provádění testování ve stále širších kontextech a/nebo testují větší soubor méně kritických požadavků, a tento proces jsou připraveni zastavit nebo upravit, pokud se objeví nebezpečné nebo jiné nežádoucí účinky.“ (Floridi, Cowls, King a kol., 2020) I při splnění tohoto procesu se ovšem může

stát, že nástroj následně selže v reálném použití. Proto by v rámci tohoto předpokladu měla být zajištěna i opravitelnost a změnitelnost nástroje, celkově by měla být zajištěna možnost updatování i po implementaci v reálném světě a nástroj by měl být pravidelně vyhodnocen a kontrolován. (Floridi, Cowls, King a kol., 2020)

Druhým je pak záruka proti manipulovatelnosti prediktorů. Prediktivní schopnost AI je náchylná ke dvěma rizikům: manipulaci se vstupními daty a nadměrnému spoléhání se na nekauzální indikátory. (Floridi, Cowls, King a kol., 2020) Problém s manipulací daty či jejich nevhodném použití je obecný problém, který se nelimituje pouze na AI, nicméně v AI se případné neodhalení může promítat do vyhodnocování a ovlivňovat následně machine learning. To pak může vykazovat chybné nebo biased výsledky, které ovšem bude prezentovat s velkou důvěrou v jejich správnost, a tím ztěžovat možnost odhalení nebo napravení. K úmyslné manipulaci dat zainteresovaným agentem pak dochází o to více, čím je fungování systému pochopitelnější. Další problém je pak chybné pochopení kauzálních a korelačních dat. Tyto limity je tedy nutné si pamatovat při vývoji a designování umělé inteligence. „Vývojáři by měli přijmout ochranná opatření, která zajistí, že nekauzální indikátory nebudou nepatřičně zkreslovat zásahy, a případně omezí znalosti toho, jak vstupy ovlivňují výstupy ze systémů AI, aby se zabránilo manipulaci.“ (Floridi, Cowls, King a kol., 2020)

Čtvrtým předpokladem je pak intervence konceptualizovaná podle příjemce. Intervence musí brát v potaz autonomii recipienta a musí vyvažovat současné a budoucí benefity. Intervence může ovlivnit uživatelovy preference, které pak umožňují softwaru uvést do kontextu budoucí zásahy. V důsledku toho může být intervenční strategie, která nemá žádný dopad na autonomii uživatele (např. taková, která postrádá jakékoliv zásahy), neúčinná při získávání nezbytných informací pro správně konceptualizované budoucí intervence. Naopak zásah, který příliš narušuje autonomii uživatele, může způsobit, že uživatel technologii odmítne, což znemožní budoucí intervence. Vývojáři by tedy měli vytvářet decision-making systémy po konzultaci s uživateli, kteří s těmito systémy interagují a jsou těmito systémy ovlivněni; s porozuměním charakteristikám uživatelů, metodám koordinace a účelům a účinkům zásahu; a s ohledem na právo uživatelů ignorovat nebo upravovat zásahy. (Floridi, Cowls, King a kol., 2020)

Na tento předpoklad pak navazuje vysvětlitelnost přizpůsobená chápání uživatele a transparentnost. AI nástroje by měly být navrženy tak, aby byly procesy a výsledky snadno vysvětlitelné a aby byly transparentní jejich účely. Srozumitelnost systémů umělé inteligence je důležitým etickým principem. (Floridi, Cowls, Beltrametty, Chatila, Chazerald, Digmum a kol., 2018) Tento etický problém je pak v prostoru EU zakotvený i v zákoně, konkrétně ve směrnici GDPR v sekci o regulaci algoritmické odpovědnosti, kde se nařizuje schopnost srozumitelně vysvětlit fungování daného nástroje, aby mohl být užívaný v rámci EU, respektive vysvětlit, jak provádí automatizovaná rozhodnutí. (Kaminski a Malgieri, 2021) Díky tomu pak vzniká i transparentnost. Tento princip je však limitován tím, že fungování některých AI nástrojů komplexně nechápou ani samotní vývojáři, jelikož dochází ke strojovému učení a AI se vyvíjí sama. (Floridi, Cowls, King a kol., 2020) Vývojáři sice neví, jak systém funguje, a proto jej nebudou schopni replikovat, na druhou stranu, pokud nebude možné nástroje implementovat, hrozí ztráta důležitých hypotetických pokroků a benefitů, která by tato případná implementace přinesla. Proti transparentnosti a vysvětlitelnosti se ovšem část odborníků vymezuje, jelikož v tom vidí odhalení know-how a absenci zachování duševního vlastnictví, což vede ke ztrátě jejich schopnosti soutěžit na trhu. (Hosanagar a Jair, 2018)

Pátým předpokladem je pak ochrana soukromí a používání dat, k jejichž poskytnutí dal subjekt souhlas. Jedná se o princip, jemuž se etika AI věnuje nejvíce. (Floridi, Cowls, King a kol., 2020) Ochrana dat by měla být zajištěna shromažďováním a vyhodnocováním pouze nezbytných dat, získáním souhlasu ke všem dílčím úkonům zahrnujícím nakládání s daty a zabezpečením vlastních systémů před úniky. Bezpečnost dat je navíc ve většině zemí vyžadována i zákony o ochraně dat a uživatelů. Nakládání s daty a jejich přeprodej je do jisté míry legislativně limitován, nicméně v tomto ohledu je v různých zemích prostor pro rozšíření legislativy . V EU je v současné době zákonem vyžadováno získávání a používání pouze nezbytných dat (střídmost), ochrana před úniky, jasně udělený souhlas s nakládáním s daty a další. (Kaminski a Malgieri, 2021)

Šestým předpokladem je pak situační férovost. Mnoho využívaných datasetů v sobě nesou i biased data; tento problém byl už zmíněný výše. Biased kód a vyhodnocování pak porušuje základní předpoklady férovosti a spravedlnosti. Takový cyklus by začal

předpojatým souborem dat, který by ovlivnil první fázi rozhodování umělé inteligence, což by vedlo k diskriminačním akcím, které by následně vedly ke sběru a používání zkreslených dat. Tento problém lze demonstrovat na použití prediktivního policejního softwaru. Vývojáři mohou školit prediktivní policejní software na policejních datech, která obsahují hluboce zakořeněné předsudky a rasově motivované zatčení. Když diskriminace ovlivňuje míru zatčení, stane se součástí údajů o stíhání. Taková data mohou způsobit diskriminační rozhodnutí, která se následně promítají zpět do stále více zaujatých datových souborů, čímž se dokončí začarovaný kruh. (Crawford, 2016) Předcházet používání zkreslených dat lze pouze kritickým zhodnocením datasetů a zvážením případných implikací, které data mohou přinést. Na vyhodnocování a anotaci dat je tedy vhodné využít diverzní tým s relevantním vzděláním a základním chápáním AI a využívání dat.

Poslední zásadou je human friendly sémantizace. AI musí lidem umožnit kurátorovat jejich vlastní „sémantický kapitál“, tedy jakýkoli obsah, který může zvýšit něčí schopnost dávat smysl něčemu a sémantizovat. (Floridi, Cowls, Beltrametty, Chatila, Chazerald, Digmum a kol., 2018) Umělá inteligence by měla usnadnit lidsky přívětivou sémantizaci, ale ne ji poskytovat.

Summary of seven factors supporting AI4SG and the corresponding best practices

Factors	Corresponding best practices	Corresponding ethical principle
Falsifiability and incremental deployment	Identify falsifiable requirements and test them in incremental steps from the lab to the “outside world”	Nonmaleficence
Safeguards against the manipulation of predictors	Adopt safeguards which (i) ensure that non-causal indicators do not inappropriately skew interventions, and (ii) limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation	Nonmaleficence
Receiver-contextualised intervention	Build decision-making systems in consultation with users interacting with and impacted by these systems; with understanding of users’ characteristics, the methods of coordination, the purposes and effects of an intervention; and with respect for users’ right to ignore or modify interventions	Autonomy
Receiver-contextualised explanation and transparent purposes	Choose a Level of Abstraction for AI explanation that fulfils the desired explanatory purpose and is appropriate to the system and the receivers; then deploy arguments that are rationally and suitably persuasive for the receiver to deliver the explanation; and ensure that the goal (the system’s purpose) for which an AI4SG system is developed and deployed is knowable to receivers of its outputs by default	Explicability
Privacy protection and data subject consent	Respect the threshold of consent established for the processing of datasets of personal data	Nonmaleficence; autonomy
Situational fairness	Remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety, or other ethical imperatives	Justice
Human-friendly semanticisation	Do not hinder the ability for people to semanticise (that is, to give meaning to, and make sense of) something	Autonomy

Principy AI přispívající k veřejnému blahu – Floridi, L., Cowls, J., King, T.C. *et al.* How to Design AI for Social Good: Seven Essential Factors, (2020)

2. 7. Příležitosti využití AI do budoucna

Z předchozích kapitol a na základě uvedené literatury můžeme také jasně formulovat příležitosti použití AI do budoucna, které můžeme následně diskutovat ve focus group.

2.7.1. Změna přístupu k falešnému obsahu

První příležitostí je změna přístupu k falešnému obsahu – měl by být deempatizován (snížit mu dosah, zamezit organickému šíření, snížit jeho relevanci) a opraven. Kromě nahlašování a snižování statusu v žebříčku falešného obsahu je důležité, aby platformy zajistily i účinné opravy prokazatelně nepravdivého nebo zavádějícího obsahu, který se objeví online. Například formou šíření protizpráv založených na faktech. Ačkoli je odhalování původce online dezinformačních kampaní komplikované, je-li k dispozici dostatek důkazů, je důležité veřejně odsoudit pachatele dezinformací a také koordinovat přesdílování a reakci. (Kertysova, 2019)

Také by měl být kladen důraz na větší odpovědnost za obsah a transparentnost. Společnosti by tak měly povinně pravidelně auditovat svoje AI systémy a zajistit, že se v nich nevyskytují bias a diskriminace. K zajištění tohoto předpokladu by tak měla být etika součástí kurikula IT oborů (ethical by design).

2.7.2. Regulace

Další příležitostí je pak možnost regulace obsahu na sociálních sítích. Většina vládních institucí si pohrává s myšlenkou buď vyžadovat větší odpovědnosti od sociálních sítích za obsah, který na nich koluje, nebo poskytnout vládě větší kontrolu nad obsahem.

Sociální sítě v rámci samoregulace vyvíjejí technické nástroje a systémy se snahou omezit šíření dezinformací a zneužívání dat a pravidelně i investují do jejich vývoje. Ovšem řada organizací tuto snahu nevnímá jako dostatečnou, obzvlášť pokud dezinformace zasahují a hatí demokratický proces při volbách. Kontrola ochrany soukromí, moderování obsahu

a investice do vývoje by podle některých měla podléhat vnějšímu dozoru. Zde se ale naskýtají otázky rozdělení rolí, povinností a odpovědnosti mezi jednotlivé aktéry v procesu. Pokud by byly společnosti za případnou neefektivní samoregulaci penalizovány, hrozí pak autocenzura – přehnané odstraňování legitimního obsahu z důvodu strachu společností ze sankcí. (West, 2017)

Dalším dozorčím může být vláda. Ovšem toto řešení s sebou nese problematické aspekty a odpor stakeholderů, ať už samotných společností, organizací na ochranu lidských práv a svobod nebo koncových uživatelů. Případný dozor vládních organizací by se dal lehce zneužít a proměnit v cenzuru. Příkladem může být suspendace proruských webů na české doméně v důsledku invaze Ruska na Ukrajinu. Tato suspendace vyvolala vlnu debat, zda jsou taková opatření namístě a zda nejsou kontraproduktivní tím, že podporují narativ Kremlu o „zatajování a blokování" médií. Ačkoliv byl tento krok kritizován, vzhledem k závažnosti situace a nutnosti rychle jednat se případná pochybení dají obhájit potřebou rychlých rozhodnutí. (root.cz, 2022) Navíc bývá vláda oproti soukromému sektoru často technicky pozadu, a tak by hrozilo, že nebude některé aspekty dostatečné chápat.

Například nedávný návrh zákona v Singapuru představoval hrozbu pro svobodu slova a svobodu tisku tím, že udělil ministrům právo rozhodnout bez soudního přezkumu, zda je online obsah pravdivý, nebo nepravdivý. Tato pravidla by navíc umožnila vládě – spíše než soudcům – zakázat prohlášení zaměřená na „snížení důvěry veřejnosti" v singapurské státní instituce. Takový jazyk vytváří právní nejistotu a ponechává mnoho prostoru pro interpretaci, potenciálně potlačující svobodu slova. (Vaswani, 2019)

Nabízí se tedy otázka, zda by dozor nad sociálními sítěmi měl ležet v rukou nějaké nezávislé organizace, která nemá přímé vazby ani na vládu, ani na vlastníky společností.

Řešení v sobě musí zahrnovat výsledky z dialogů různých stran a odborníků napříč obory a musí být schopné rychle reagovat na proměny v prostředí. Také je potřeba se soustředit na ostatní hráče na trhu, a nejen 3 hlavní – Facebook, Instagram a Twitter (4chain, Reddit, Quora a další menší sítě také podléhají šíření dezinformací). (Kertysova, 2019)

2.7.2. Techplomacie

Jedním z možných řešení problematiky chápání algoritmu a dohledu nad společnostmi pak může být techplomacie. Jedná se o technologické delegace, které úzce komunikují a spolupracují s danou společností a většinou jsou vyslané vládou nebo veřejnou institucí. Další možností pak může být ambasadorství, tedy zaměstnání experta v dané firmě, který se zodpovídá vládě nebo organizaci a je zodpovědný za jejich informování. Tento vztah může být využit k efektivnějšímu dialogu, může komunikovat problémy jako ovlivňování voleb, dezinformace, závadný obsah, kyberbezpečnost nebo získávání e-důkazů pro další vyšetřování a stíhání aktérů. (Foremski, 2019) Aby byla techplomacie účinná, musí být jasně stanoven mandát, určeny pravomoci a zajištěna politická podpora.

2.7.3. Mediální a digitální vzdělanost

V posledních dekádách jsme se naučili vyhledat téměř jakékoliv informace, nicméně mírně pokulháváme v jejich vyhodnocování, respektive rozhodování o jejich autentičnosti. Téměř každý může publikovat názory a vydávat je za fakta, ohýbat pravdu pro podpoření vlastní agendy, nebo jen šířit už vytvořené dezinformace. Uživatel je tak tvůrcem i konzumentem, soudcem i obětí. Klíčovou se tedy stává informační vzdělanost, schopnost ověřovat zdroje, rozpoznávání bias a porovnávání více médií pojednávajících o stejné události. (Nguyen, Khorosekar, Krishman, Krishman, Tate, Wallace a Lease, 2018) „Informační vzdělanost se stala nezbytnou dovedností a klíčem k sociálnímu a ekonomickému blahu jedince ve společnosti s kompletními informacemi.“ (Kuhlthau, 1987)

Jak uvedl Dr. Alexander Klimburg, ředitel Programu kybernetické politiky a odolnosti v Haagském centru strategických studií, „útok na infrastrukturu kybernetiky (technické vrstvy) je jen oklikou k útoku na mysl (lidskou bytost).“ (Klimburg, 2017) Je tedy evidentní, že pro úspěšný boj musíme posílit nejen implementaci a využití umělé inteligence v boji, ale i odolnost našeho obyvatelstva vůči útokům, manipulaci a

radikalizaci. Ve chvíli, kdy není pro dezinformace publikum, nemohou fungovat ani dezinformace. (Kertysova, 2018)

V evropských zemích, a zejména pak v zemích východní Evropy si dlouhodobě uvědomujeme, že mediální, informační a digitální vzdělanost je problém. Tento problém je pak o to urgentnější u skupin, které prokazují dlouhodobě nejvyšší účast u voleb, tedy starší generace a extremizovaných jedinců. (Kertysova, 2018) Ačkoliv se stát nebo neziskové organizace snaží tento problém adresovat a řešit skrz přednáškové panely a úpravy vzdělávacího kurikula, snaha stále leží na pozadí, přehlušena jinými potřebami. Komplexní a efektivní řešení chybí.

Case study: Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking (Nguyen, Khorosekar, Krishman, Krishman, Tate, Wallace a Lease, 2018)

Unikátní způsob fact-checkingu představil tým vědců, kteří navrhli výukový model pro rozpoznávání dezinformací. Jejich hlavním cílem nebylo pouze vytvořit nástroj na rozpoznávání dezinformací, ale zapojit do tohoto procesu i lidi, naučit je rozeznávat autenticitu informací, zvyšovat jejich mediální a informační vzdělanost a také budovat důvěru v systémy vyhodnocující dezinformace.

Autoři studie vnímají 3 základní vlastnosti, které by AI měla mít, a sice transparentnost, schopnost integrovat uživatelovy znalosti a komunikování nejistoty systému. Proto vytvořili smíšený přístup k vývoji fact-checkingových nástrojů, který se pokouší integrovat všechny zmíněné limity. Jejich nástroj poskytuje systém na fact-checking pouze jako asistenční službu pro uživatele, která jim napomůže v rozhodování o autentičnosti a pravdivosti prezentovaných informací. Systém má uživatelům poskytnout framework, do kterého mohou vložit své vlastní znalosti a hodnoty a spojit je dohromady.

Systém nalezne relevantní články k danému tématu/tvrzení, vyhodnotí historickou spolehlivost každého zdroje a zaznamená jejich postoj k danému tématu a z toho vyhodnotí validitu daného tvrzení. Model vykazuje vůči uživatelům transparentní

rozhodovací procesy a je interaktivní, což by mělo uživatelům pomoci s důvěrou v jeho rozhodování a zároveň je naučit vyhodnocovat informace samostatně a osvojit si logický proces k jejich posouzení.

Model vyjadřuje míru důvěry ve svá rozhodnutí (např.: 80% jistota, že jde o nedůvěryhodný článek), obecnou reputaci médií, která tvrzení publikují (médium často publikuje mylné informace) a ostatních médií, která mají opačné tvrzení.

K otestování tohoto nástroje byly provedeny 3 experimenty, jež vykazovaly podpůrné hodnoty. Systém má také slepá místa, kde ukazuje falešně pozitivní, nebo falešně negativní výsledky. Uživatel se v jednom experimentu testoval na Likertové škále, kde rozhodoval, jak moc věří systému. Ve druhém experimentu mohl interagovat se systémem a vložit do něj vlastní znalosti, měl kontrolu nad slidery, mohl změnit reputace a predikovat výsledek podle vlastní zkušenosti. Tuto predikci pak vezme systém v potaz.

Ačkoliv modely s interakcí zaznamenávaly signifikantní výsledky a byly pro uživatele zábavnější, díky čemuž jim věnovali větší pozornost, odhalily i slabé místo, jelikož AI fact-checkingový nástroj není neomylný a v rámci experimentu i úmyslně dělal chyby, aby uživatele otestoval. Někteří uživatelé v něj ztratili důvěru, což následně může ohrozit také obecnou důvěru v tyto nástroje. Kombinovaný výukový nástroj tedy prokazuje větší efektivitu při vyhodnocování a učení, nicméně nese s sebou i rizika, která je třeba mít na zřeteli. (Nguyen, Khorosekar, Krishman, Krishman, Tate, Wallace a Lease, 2018)

Case Study: Finská praxe

S dalším pokusem přišlo Finsko, které v roce 2016 spustilo kampaň bojující proti šíření fake news. Kampaň cílila na rezidenty, studenty i novináře a učila je, jak rozpoznat dezinformace. Cílem bylo zvýšit schopnost kritického myšlení a naučit děti i praktické dovednosti, například jak rozeznat trolla nebo boty prohledáním jejich profilu a hledáním společných ukazatelů. Kromě mediální kampaně a změny osnov program také učil

politiky, jak rozeznat dezinformace a vyrovnat se s nimi nejen obecně, ale i s těmi, které cílí přímo na ně. Na kampani se také podílela finská společnost Faktabaari, která vytvořila nástroj pro výuku na základních a středních školách. (Berzina a kol., 2019)

Finská vláda se také spojila se společností Reaktor a univerzitou v Helsinkách, aby zadarmo naučila zájemce z řad veřejnosti základní aspekty, fungování a vlastnosti AI. (University of Helsinki, 2018)

2.7.4. Kyberbezpečnost

Většina dezinformačních aktivit je stále spojena s tradičními kybernetickými útoky. Jedním z cílů, který již byl zmíněn, je šířit paniku. Dalším cílem jsou pak samozřejmě data uživatelů, která mohou nepřátelským aktérům poskytnut mnoho cenných informací. Umělá inteligence posiluje právě i kybernetické útoky, zvládá je automatizovat, vyhýbat se detekci i obcházet obranná opatření.

Jediným řešením tohoto problému je investice do kontinuálního vývoje bezpečnějších a pokročilejších obranných digitálních infrastruktur a vzdělávání občanů ohledně kybernetické bezpečnosti. Speciální pozornost by se pak měla věnovat zabezpečení vládních serverů s oficiálními informacemi (zde hrozí největší riziko ohrožení důvěry ve stát), zdravotnickým zařízením a bankám (jejich napadení vnímají obyvatelé nejvíce) a zajištění bezpečnosti serveru při volbách.

2.7.5. Lepší financování

Nedílnou součástí komplexního řešení problému je samozřejmě lepší financování výzkumu technologií a vylepšení algoritmu, alokace zdrojů pro vládní a nadnárodní organizace, které by se zabývaly vývojem a mezioborovou spoluprací v boji s dezinformacemi, nebo lepší financování univerzitních projektů zabývajících se touto problematikou.

2.8. Instituce vstupující do kontextu AI

2.8.1. Ministerstvo průmyslu a obchodu

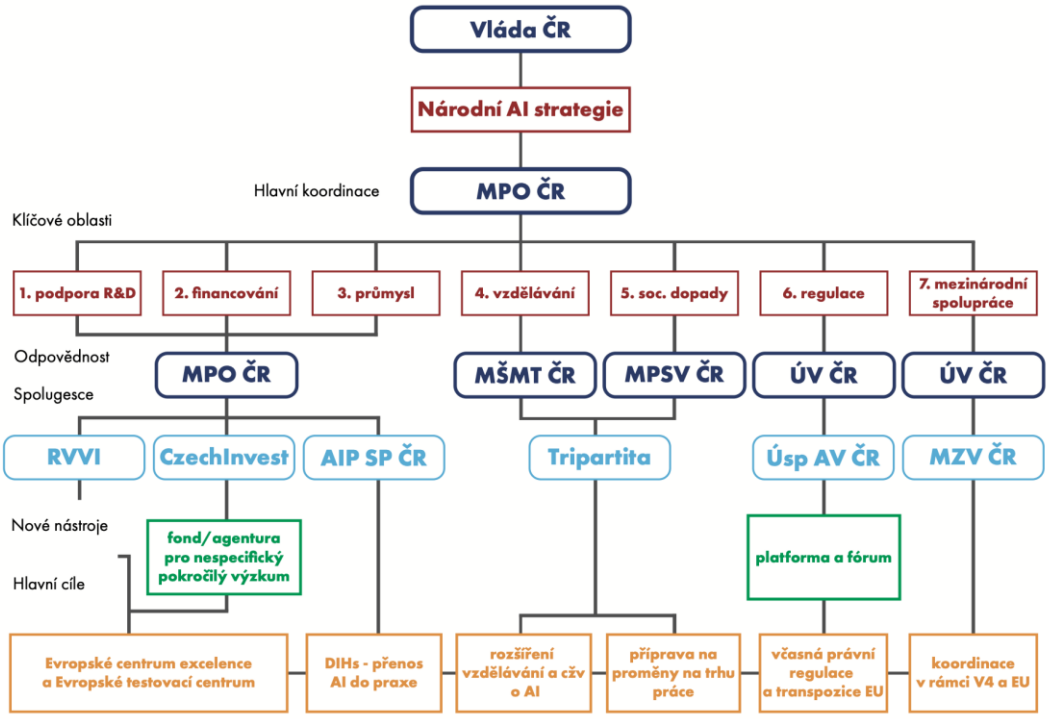
Současná strategie AI České republiky se soustředí na koordinaci vlastních cílů s materiály a plány EU, podporu vědy, výzkumu a vývoje, zajištění financování a investování do výzkumů, zkoumá AI v průmyslu a službách, případně dopady na trh práce, zajištění vzdělávání pracovníků. Odpovědnost za plnění nese Ministerstvo průmyslu a obchodu. Strategie je formulována v dokumentu *Národní strategie umělé inteligence v České republice*, 2020.

Mezi její současné cíle patří plná integrace a spolupráce Evropského centra excelence s partnerskými výzkumnými pracovišti, snaha zatraktivnit toto centrum pro veřejnost a přilákat kvalifikovanou inteligenci ze zahraničí (Program propagace ČR v zahraničí pro experty ve výzkumu). Dalším cílem je spustit aspoň 5 projektů pro sdílení akademického know-how. Ministerstvo chce také rozšířit finanční podporu výzkumu a vývoje, budovat inovační huby, podporovat startupy a inovativní projekty, iniciovat vznik různých center a vytvořit různé útvary pro komunikaci pokroků a spolupráci při vývoji. Dále chce rozšířit národní superpočítačové centrum, zpřístupnit data ze zdravotnictví pro výzkum a vytvořit výzkumné pracoviště zkoumající využití AI v kosmu. Také chce o 100 procent zvýšit publikační aktivity v odborných časopisech. (MPO, 2020)

Do roku 2035 chce ministerstvo revidovat podmínky získání pobytových a pracovních oprávnění pro vědecké a výzkumné pracovníky ve strategických oblastech a chce také zlepšit způsob vyhledávání těchto elit v ČR a zamezovat jejich odlivu do zahraničí. Chce také více podporovat místní subjekty při snaze získat evropské granty.

Ministerstvo chce také „připravit společnost“ na dopady AI a automatizace, zabránit vzniku dlouhodobé nezaměstnanosti a vytvářet nová místa na trhu práce a minimalizovat možné negativní dopady a dále zřídit expertní platformu a fórum pro průběžnou revizi právních a etických pravidel AI. Stát se také pokusí zavést nové instituty pro rozvoj AI a revidovat legislativu v souvislosti s AI s důrazem na zabránění diskriminace. (Vláda ČR)

Jako nástroje pro dosažení je zmíněn monitoring nově vznikajících příležitostí, sledování dopadů technologií na trh práce a vytváření podmínek pro výzkum a rozvoj. Kontaktovala jsem ministerstvo s dotazem na průběžné plnění plánu, bohužel vyhodnocení budou provádět až ve druhém čtvrtletí roku 2023 a do té doby pro mě žádné relevantní informace nemají.



AI nástroje a Česká republika, propojení a spolupráce zainteresovaných subjektů (MPO)

2.8.2. Instituce EU

V rámci EU Evropská komise financuje External action services pro strategickou komunikaci, které se snaží bojovat s dezinformacemi. (European External Action Service, 2021) EU investuje přes 1,5 milionů eur do nástrojů pro podporu spolupráce mezi univerzitami, výzkumníky a fact-checkingovou komunitou. Vznikl útvar SOMA, platforma usnadňující spolupráci a sdílení know-how mezi profesionály. (Cordis, 2021)

V rámci EU také vznikl v roce 2018 Code of Practice on disinformation. Jednalo se o seberegulační kodex s cílem kolaborovat při řešení problémů vznikajících v důsledku dezinformací. S kódem souhlasilo 21 společností a jako výsledek vznikla Knihovna reklam, která katalogizovala zadavatele politické reklamy, Twitter zveřejnil svá takedown data (data, podle kterých jsou mazány příspěvky), společnosti rozšířily své fact-checkingové aktivity a označování dezinformačního obsahu a vytvořily programy pro vzdělávání. (Tanner, 2022)

V roce 2022 předložili online platformy, odborníci, inzerenti, fact-checkeři, výzkumníci a civilní společenské organizace posílený kód. Kód byl podepsán dalšími 34 signatáři. (Tanner, 2022) Cílem nového kódu je splnit požadavky komise z května 2021 volající po stanovení širší škály závazků a opatření pro boj proti online dezinformacím. (European Commission, 2022)

Proces jeho revize byl zahájen v červnu 2021 a po podpisu a představení revidovaného Kodexu dne 16. června 2022 se nový Kodex stane součástí širšího regulačního rámce v kombinaci s legislativou o transparentnosti a cílení politické reklamy a Zákonem o digitálních službách. Kodex 2022 je výsledkem práce signatářů. Je na signatářích, aby se rozhodli, ke kterým závazkům se přihlásí, a je jejich odpovědností zajistit plnění těchto závazků. Signatáři se zavázali podniknout kroky v několika oblastech jako jsou demonetizace šíření dezinformací, zajištění transparentnosti politické reklamy, zmocnění uživatelů, posílení spolupráce s ověřovateli faktů a poskytnutí výzkumným pracovníkům lepší přístup k datům. (Tanner, 2022)

Aby byl Kodex odolný i v budoucnosti, vytvořili framework pro další spolupráci prostřednictvím stálé pracovní skupiny. Kodex také přichází s posíleným monitorovacím rámcem založeným na kvalitativních datech a ukazatelích úrovně služeb měřících efektivitu jeho implementace. Vznikne také Centrum pro transparentnost, které veřejnosti poskytne jasný přehled o praktikách společností, které zavedly k plnění svých závazků, a bude je pravidelně aktualizovat o příslušné údaje. (European Commission, 2022)

2.8.3. Disinformation warning systém

Tento nástroj byl vyvinut českou společností Semantic Vision, který má pomoci společnostem monitorovat i vyhodnocovat dezinformace. Jedná se o early-warning solution, které pomáhá vyhodnocovat možné dezinformační hrozby a upozorňovat na ně. Systém funguje na bázi monitorovacího zařízení, které skenuje světová média. Využívá k tomu ASA – pokročilou sémantickou analýzu hledající dezinformační narativy. Systém po vyhodnocení na případnou nastávající hrozbu upozorní. Společnost nabízí i řešení pro státní orgány, které jim pomáhá vyhodnocovat hrozby a monitorovat prostředí. (SemanticVision)

Tento nástroj funguje na základě sémantické analýzy, pro níž využívá jejich vlastní mezijazyčnou databázi. Ta zahrnuje 90 % světových médií. (SemanticVision)

Společnost Semantic Vision dodává podporu i spolku NELEŽ, který byl zmíněný v předchozí kapitole této práce. Za pomoci vojenského zpravodajského systému OSINT poskytuje Semantic Visions pokročilou sémantickou analýzu velkých dat. To slouží k propojení miliard online zpravodajských článků ve 12 světových jazycích (z více než 900 000 veřejně dostupných online zpravodajských zdrojů).

Tento shluk informací pak promění ve znalostní základnu postavenou na významech vyšších úrovní (tisíce takzvaných sémantických kategoriích), která podporuje identifikaci online trendů, narativů a hrozeb.

Firma monitoruje 90 % online zpravodajského obsahu z celého světa, který je analyzován a aktualizován každých 30 minut. To je denně více než 1 milion online zpravodajských článků a blogů. (SemanticVsion)

2.8.4. P*říklad akademického výzkumu na ČVUT*

Dalším zajímavým nástrojem je pak systém vyvinutý minulý rok na ČVUT, který se na základě získaných anotací, které posloužily jako training data, naučil automaticky

detekovat, zda jsou vybraná média nějakým způsobem biased, zda mají nějaký sklon nebo ne. Sklon je v této práci definovaný jako: „Tendence zpravodajských médií informovat způsobem, který posiluje určitý názor, světonázor, preference, politickou ideologii, korporátní nebo finanční zájmy, morální rámec nebo politickou inklinaci, namísto objektivního informování (prostého popisu faktů). Média mohou projevovat zaujatost tím, jak informují o konkrétních zprávách, nebo které zprávy se rozhodnou pokrýt, tj. považují je za důležitější než jiné, které mají pokrýt nebo zdůraznit." Horych, 2022) Jako hodnotící faktory používá tento systém senzualismus/emocionalismus (výslovný sentiment ve výpovědi), subjektivní adjektiva (přídavná jména jako extrémní, trapný, vážný, atd) a osobní útoky nebo urážky. Vyhodnocování dociluje za pomocí textové analýzy užívání těchto výše zmíněných faktorů. Automatická, strojová klasifikace zpravodajských textů může lidem pomoci na cestě k detekci bias u velkého množství dat. Metody použité v této práci jsou NLP[8] a SOTA[9]. (Horych, 2022)

2.8.4. Good AI

Výukový nástroj pro chápání AI má za cíl jednak poskytnout a naučit znalosti umělé inteligence, ale také zmírnit obavy veřejnosti z AI. Uživatel se může pomocí něj naučit sestavit základní moduly a osvojit si základní chápání programování AI. (E15, 2015) Jedná se o vzdělávací platformu s osvětou jako hlavním cílem své činnosti.

8 Natural Language Processing je teoreticky motivovaná řada výpočetních technik pro analýzu a reprezentaci přirozeně se vyskytujících textů na jedné nebo více úrovních lingvistické analýzy za účelem dosažení lidského jazykového zpracování pro řadu úkolů nebo aplikací. (Liddy, 2001) Jedná se o odvětví umělé inteligence, které se zabývá tím, že počítačům dává schopnost porozumět textu a mluveným slovům v podstatě stejným způsobem, jakým jim mohou rozumět lidské bytosti. Kombinuje počítačovou lingvistiku – modelování lidského jazyka na základě pravidel – se statistickými modely, modely strojového učení a hlubokého učení. Tyto technologie společně umožňují počítačům zpracovávat lidský jazyk ve formě textových nebo hlasových dat a „porozumět" jeho plnému významu včetně záměru a pocitu mluvčího nebo spisovatele. (https://www.ibm.com/topics/natural-language-processing)

9 SOTA je kolektivní adaptivní systém. SOTA spojuje lekce cílově orientovaného modelování požadavků, kontextového modelování systémů a modelování dynamických systémů. Má potenciál fungovat jako obecný referenční model, který pomáhá řešit některé klíčové problémy při navrhování a vývoji kolektivních adaptivních systémů. SOTA umožňuje včasné ověření požadavků, identifikaci znalostních požadavků na auto adaptaci a identifikaci nejvhodnějších architektonických vzorů pro sebeadaptaci. (Abeywickrama, Bicocchi, Mamei a Zambonelli, 2020)

2.8.5. CEDMO

Středoevropská observatoř digitálních médií (CEDMO) je nezávislé nestranné multidisciplinární centrum vzniklé z občanské iniciativy.

Klade si za cíl identifikovat, zkoumat a prioritizovat nejkritičtější zdroje a příčiny informačních poruch ve střední Evropě (především Česká republika, Slovensko a Polsko). Toto mezinárodní konsorcium bylo vytvořeno, aby navrhlo soubor krátkodobých a dlouhodobých akcí stejně jako doporučení, která mají pomoci občanské společnosti, veřejným institucím a soukromému sektoru reagovat na klesající důvěru v klíčové instituce a pomoci společnosti odolat účinkům rostoucí expozice k dezinformacím. (cedmohub.eu, cit. 2023)

Aktivita CEDMO má tedy tři hlavní pilíře: výzkum, fact-checking a mediální gramotnost. Mezi partnery je například Univerzita Karlova, ČVUT nebo server Demagog. Výzkum je částečně financovaný Evropskou unií. Ve své činnosti se CEDMO také věnuje právě technologické spolupráci, jak by měla být mezioborově vyvíjena a použita, podporuje aktivity jako rozvíjení technické infrastruktury pro okamžité dezinformační reakce a každodenní fact-checking, inovaci metod a digitálních nástrojů, vytváření mezioborových analytických zpráv a chce zajistit etické inovace a odpovědné používání technologií. (cedmohub.eu, cit. 2023)

2.8.6. OpenAI

OpenAI je výzkumný ústav zaměřený na vytváření a propagaci technologií AI, které jsou bezpečné a prospěšné pro lidstvo. Tato organizace se soustředí na vývoj pokročilých systémů pro zpracování jazyka a posilovacích výukových algoritmů. (Pavlik, 2023)

Jedním z jejich projektů je ChatGPT, který rozeznává i český jazyk a funguje také na systému NLP. Jedná se o systém, který dokáže reagovat na pokyny a otázky. Samozřejmě k jeho ovládnutí je potřeba se užíváním seznámit s jeho logikou (Jak správně formulovat

otázky a požadavky, aby jej stroj pochopil. Pomocí správné formulace se dají obejít i nějaké zabudované restrikce od tvůrce.). Nástroj je schopen generovat i téměř lidský text, v současné době se kolem něho točí debaty o budoucnosti akademického psaní a zkoušení. (King a chatGPT, 2023).

Dalším projektem společnosti OpenAI je DALL-E, rozsáhlý model umělé inteligence, který je schopen vytvářet originální obrázky z textových popisů. Jméno DALL-E je odkazem na umělce Salvatora Dalího a postavu WALL-E ze stejnojmenného filmu Pixar. (Pavlik, 2023)

Pro nás tento nástroj může být přínosem při ověřování a zodpovídání otázek týkajících se dezinformací a pro tvoření nápravných textů. Umělá inteligence je schopná vygenerovat koherentní a prakticky správné články o bezpečnosti na internetu i o různorodých tématech. (Zhai, 2022) Nástroj by nám tak mohl v budoucnu automaticky generovat příručky pro fact-checking, brožury o bezpečnosti na internetu a materiály, jak rozpoznat dezinformace, které by pak za pomoci dalšího nástroje mohly být automaticky šířeny na sociálních sítích. Stejně pak může generovat kontra informace, na každou dezinformaci může vygenerovat kontra články zmiňující ověřené informace, a ty pak automaticky šířit.

Stejně tak ale může dezinformace posílit: Analýza NewsGuard zjistila, že když byl ChatGPT konfrontován se 100 falešnými narativy, v 80 % případů v návaznosti na ně přednesl výmluvná, nepravdivá a zavádějící tvrzení o významných tématech ve zprávách, včetně COVID-19, Ukrajiny a střelby ve školách. Výsledky tedy potvrzují obavy ze zneužití. ChatGPT vygeneroval dezinformace – včetně podrobných zpravodajských článků, esejů a televizních scénářů. Na kohokoli, kdo není obeznámen s problémy nebo tematikou předkládaného textu, mohou výsledky snadno působit jako legitimní, a dokonce směrodatné. NewsGuard také zjistil, že ChatGPT má hranice, které mu brání v šíření některých příkladů dezinformací. U některých mýtů skutečně trvalo až pět pokusů, aby chatbot převzal dezinformace. Mateřská společnost ChatGPT uvedla, že nadcházející verze softwaru budou lépe informované. Kromě toho je ChatGPT pozoruhodně zběhlý v

zodpovědném odpovídání na složité otázky. Dokáže vyvrátit některé mýty a často je schopen zabránit předávání nepravdivých informací. (Newsguard, 2023)

Samotní uživatelé pak hledají možnosti, jak obejít samoregulace těchto nástrojů, takzvaně je donutit k jailbreaku, tedy poručit jim vlastní pravidla a tvořit i zakázaný obsah. Dochází k tomu skrze logické a argumentační příkazy, které například umožní nástroji vytvořit alter ego, které není limitováno pravidly jako původní nástroj. U ChatGPT je to DAN (Do Anything Now nebo zemři). (Washington Post, 2023)

Správci nástroje se snaží tato obcházení pravidel odstranit, ale uživatelé stále přicházejí s novými způsoby, jak vylákat „DANA“. Tyto způsoby pak sdílejí na Redditu k využití dalšími uživateli.

3. Situační analýza

Obecnou premisou, která změnila válečnou hru, je překonání geografické příslušnosti. Válka už nemusí vzniknout pouze mezi státy, které jsou schopny se fyzicky dostat k protivníkovu území, a jejich nadvláda pak už neleží v ovládnutí moře, hranic nebo dalších fyzických strategických pozic. Nově může válku rozpoutat stát vzdálený půlku světa a neovládající žádný klasický strategický prostor, ale dominující v technologii a strategiích informačního nátlaku. Cílem je ovládnout veřejné mínění a protivníkovo obyvatelstvo. Tím se například pro nás stává Rusko nebo i Čína daleko větší hrozbou. Už mezi námi neleží fyzická bariéra v podobě okolních států, která chrání naše obyvatele před ruským vlivem. Tato bariéra byla překonána dezinformačními kanály, proruskými politickými aktéry uvnitř země a spolky financovanými ruskými společnostmi. (Polyakova a Boyer, 2018) Ruský narativ je rozprášený mezi různorodé kanály a sleduje diverzní agendu. Zároveň takovýto druh války může být rozpoután nepozorovaně a napadený stát si ani nemusí uvědomovat, že ji vede. Pro pochopení nezbytného kontextu této nové války nám poslouží PEST analýza.

„Cíle i nástroje zůstávají stejné – kybernetické útoky, dezinformační kampaně, kultivace politických spojenců a politická subverze s cílem destabilizovat a oklamat demokratickou společnost.“ (Polyakova a Boyer, 2018)

Výčet výše nám odhalil i různé motivace aktérů. Na jedné straně je to politik, kterému jde o vítězství bez ohledu na etické normy, dále nepřátelský stát, který může profitovat ze zvolení zmíněného politika, nebo korporace s prostředky a know-how, jež je motivovaná profitem a potřebou prokázat se v oboru, ale i nezaujatí pozorovatelé, kteří sdílí obsah kvůli monetizaci návštěvnosti stránek a poskytnutí prostoru inzerentům. (Hughes a Waismel-Manor, 2021) Komerční využití internetu umožňuje autorům zavádějícího obsahu získat nebývalé příjmy z reklamy, což je více motivuje k produkci tohoto obsahu (Jost, 2020),

3.1. PEST

PEST analýza je široce používaný nástroj pro pochopení strategického rizika. Identifikuje změny a vlivy vnějšího makroprostředí. Externí prostředí se skládá z proměnných, které nelze přímo kontrolovat ovlivněným subjektem, ale lze je analyzovat a zajistit, aby strategie byla efektivně schopna reagovat na toto prostředí. (Sammut-Bonnici, Tanya a Galea, 2014) V rámci této práce ji zahrnujeme pro pochopení širšího kontextu. Jelikož je tato práce konkrétně lokalizovaná, pro zajištění komplexnosti informací o prostředí, ve kterém se pohybujeme, je vyhotoveno následné shrnutí.

3.1.1. Politické prostředí

Politické prostředí boje v informační válce prostupují různé narativy, nástroje a cíle, které rámují obecnou politickou krajinu. Vzhledem k povaze demokratického systému je jedním z hlavních nástrojů obrany monitoring.

Monitoringu dezinformací se věnuje například Bezpečnostní centrum evropské hodnoty, které zajišťuje iniciativy Kremlin Watch (sledování dezinformační scény a ruského vlivu v České republice), Red Watch (monitorování čínských vlivových operací v České republice a po celém světě) nebo Východoevropský program (mapování zahraničního vlivu a monitorování dezinformací se zapojením místních partnerů). (Evropské hodnoty o. p. s., 2022)

V našem situačním prostředí jsou pak nejvýraznější dezinformační činitelé Rusko a následně Čína. Následující výklad je tedy proveden hlavně z této optiky.

Svrchovanost a narativ

Důležitým pojmem informační války v kontextu postsovětských států je svrchovanost státu. Téma svrchovanosti prostupuje do politické i veřejné diskuse v mnoha formách zevnitř státu i vně. Svrchovanost státu je pro mnoho postkomunistických zemí něco, čeho se jim většinu dvacátého století nedostávalo. Zároveň se tento pojem často skloňuje ve

spojitosti s Evropskou unií, respektive dezinformace o EU se často opírají o domnělou ztrátu svrchovanosti. (euvsdisinfo.eu)

Rusko při prosazování svých zájmů v cizích zemích dbá na to, aby jeho zásah nebyl brán jako útok na svrchovanost. Ve svých zásazích se tedy snaží vystupovat jako otec Slovanů, který dal svým postsovětským „dětem" dost volnosti, aby se mohly realizovat, ale stále si uchovává právo zasáhnout pro jejich vlastní dobro. Tuto dynamiku můžeme pozorovat například v prohlášení Vladimíra Putina z roku 2006, kdy nabídl Ukrajině ochranu (o kterou nežádala) a garanci hranic výměnou za umístění své flotily v Černém moři – ačkoliv tato garance byla potvrzena již v roce 1992 a její opakované nabízení výměnou za strategickou pozici působilo spíš jako nenápadná hrozba. (Socor, 2006)

Z toho vznikají dvě základní premisy politického kontextu. Prvním je, že Rusko od politiků přes vojsko až po prostý lid vnímá postsovětské země stále jako své satelity, podružná území, kde odmítnutí respektu v podobě například odstranění památníku musí být tvrdě potrestáno pro poučení tohoto i okolních států.

A druhá, že Putin dbá na tvoření image otcovského vládce. Někoho, kdo zasahuje pouze pro dobro svých sousedů a vlastně na to má právo, jelikož Rusko je matka všech. K tomu se pak přidává narativ slovanského bratrství a pospolitosti v boji proti hrozbám ze Západu a ochrany před fašismem a ztrátou kultury. (Čižik a Masaryková 2018) O Krymu například v roce 2006 Putin řekl, že ačkoliv jej bere jako součást Ukrajiny, do jejíž vnitřních záležitostí nemůže zasahovat, nemůže být lhostejný k tomu, co se na Ukrajině a na Krymu děje, a musí jí v případě potřeby pomoci. (Socor, 2006) Dnes již známe plný rozsah těchto slov a jejich význam. Ukrajina nikdy nebyla v očích Putina plně svobodná a nárok na její „ochranu" a ochranu ruského obyvatelstva na Krymu hlásal Putin téměř od začátku. Také vidíme, jak relativně daleko do minulosti zasahuje budování narativu, který dnes pomáhá hájit invazi na cizí území.

Rusko také vstupuje do politické krajiny formou prosazování a podpory skupin nebo jedinců, kteří podporují jejich zájmy ve veřejném prostoru. Jedná se o financování různých médií, například regionálních mutací média Sputnik nebo financování kampaní.

Historicky také ovlivňují volební klání a demokratický proces, v případě Běloruska z roku 2020 i násilnou formou. (Goul-Davies, 2020)

Jednou z hlavních hrozeb, které dezinformace představují v politickém prostředí, je ohrožení a zásah do demokratického procesu.

Case Study: Ovlivňování voleb

Samotnou kapitolou informačních válek jsou pak zásahy do voleb zájmového státu. V roce 2004 byla doložena ruská snaha ovlivnit výsledek ukrajinských voleb zapojením spin doktorů a finanční podporou a podplácením různých politiků se snahou podporovat Ruskem preferovaného kandidáta.

Rusko se však nezastavuje pouze u investování peněz. Dalším oblíbeným nástrojem je diskreditace oponenta sofistikovaným a intenzivním dezinformačním útokem na jeho osobu. Využívá k tomu již zavedené kanály a narativy, které pouze doplní, nebo ohnou tak, aby do nich mohli zahrnout konkrétní osobu nebo stranu.

V den samotných voleb se pak u volebních uren objevili maskovaní muži, kteří obtěžovali voliče opozice, mnoho voličů také „povstalo z mrtvých“. Výsledek voleb dopadl příznivě pro Ruskem podporovaného Janukoviče. (Polyakova a Boyer, 2018)

Zemí se ovšem v reakci na výsledky přehnala vlna protestů, a tak nakonec proběhly druhé volby, ve kterých už zvítězil opoziční prozápadní kandidát Juščenko.

Ten během prezidentského klání čelil údajnému pokusu o otravu na státní večeři, z něhož obvinil Moskvu, a ze spoluúčasti obvinil svého blízkého spolupracovníka Žvaniju, který ho na večeři vylákal. Ten naopak vše vyvrátil, otravu popřel a obvinil prezidenta a jeho volební tým z inscenace pokusu o vraždu, aby si tak zajistili preference před blížícími se volbami. Dle něj měl prezident pouze lehkou otravu jídlem a stopy na obličeji jsou pozůstatkem zánětu nervu. To ovšem vyvrátili Juščenkovi doktoři, kteří otravu dioxidem a její vážnost potvrdili týden po večeři v nemocnici ve Vídni. Moskva odmítla při

vyšetřování spolupracovat a jakoukoliv účast popřela. Kolem celého incidentu se šířily různé narativy a relativizovala se pravda. Dodnes není útok objasněn a nikdo nebyl oficiálně obviněn. (lidovky.cz)

Snaha ovlivnit ukrajinské volby se projevila i v roce 2014, kdy už Rusko s Ukrajinou vedly otevřený konflikt na Donbasu. Hackeři se probourali do státních systémů, vymazali několik důležitých složek a nasadili virus, který měl za úkol změnit výsledky voleb ve prospěch pravicové strany favorizované Ruskem. Virus byl ovšem hodinu před zveřejněním výsledků detekován a odstraněn. Ruské mediální agentury ale přesto ve zprávách označily za vítěze voleb právě proruskou pravicovou stranu. (Clayton, 2014)

Snaha zmanipulovat ukrajinské volby, ačkoliv neúspěšná, přinesla nakonec Rusku potřebné poučení a zkušenosti, a mnoho z těchto nástrojů pak země využila o dva roky později k zásahu do prezidentských voleb v USA ve prospěch kandidáta Donalda Trumpa.

Vliv, který můžou mít tyto nové nástroje na politiku a demokracii, pak můžeme sledovat na volbách v USA v roce 2016. Ve volbách byly poprvé masivně používány algoritmy, automatizace, microtargeting a umělé inteligence, které posílily efektivnost a množství dezinformačních kampaní a přidružených kyberaktivit a ovlivnily formování názorů a rozhodování amerických obyvatel v prezidentské volbě. (Howard, Woolley a Calo, 2016) Rostoucí síla těchto nástrojů pak umožňuje záškodnickým aktérům infiltrovat vládní instituce a korporáty při snaze získat data uživatelů, kompromitovat jejich soukromí a ovlivnit volby téměř bez povšimnutí. (Fly, Rosenberger a Salvo, 2018)

Dezinformační a hackerské kampaně pak od roku 2016 doprovázely téměř všechny světové události, z nichž by mohlo benefitovat postavení a zájmy Ruska.

Trollové napojení na Rusko a boti šířily falešné informace o prezidentském kandidátovi Emanuelu Macronovi (macronLeaks – duben až květen 2017), dezinformační kampaň obklopovala katalánské referendum v roce 2017 a v posledních třech letech je proruský narativ přítomný téměř v každé politické debatě od vakcinace až po interrupční politiku.

Nadnárodní organizace

Politické prostředí také ovlivňují nadnárodní organizace. Konkrétně v případě České republiky to jsou organizace EU a NATO, které jednak vstupují do prostředí jako směrodatné a výkonné orgány a jednak také jako součást narativu. Tyto organizace sdružují většinu států střední a západní Evropy a postupně se k nim přidávají, nebo projevují úmysl přidat státy východní Evropy s blízkostí k Rusku. Tuto intenci vnímá Rusko jako porušení míru a hrozbu vojenského útoku na své území (kvůli blízkosti těchto zemí zvažujících členství k ruským hranicím). Rusko se tedy aktivně snaží pomocí otevřených výzev i dezinformačních kampaní zamezit sousedním zemím ke vstupu do NATO nebo EU.

V roce 2006 proběhly strojené demonstrace na Ukrajině proti plánovanému cvičení NATO na jejím území, o rok později Rusko zakázalo Gruzii vstup do NATO. Švédsko v roce 2016 pak čelilo sérii falešných zpráv o negativních důsledcích, které by s sebou vstup do NATO nesly – uskladnění nukleárních zbraní, možnost útoku NATO na Rusko bez povolení Stockholmu, nebo znásilňovaní švédských žen vojáky NATO bez hrozby právní perzekuce z důvodu imunity. (Polyakova a Boyer, 2018) V roce 2022 pak hrozba vstupu Ukrajiny do NATO a EU posloužila jako záminka pro rozpoutání války na Ukrajině.

3.1.2. Ekonomické prostředí

Dezinformace zároveň ovlivňují i samotné hospodaření státu a firem, ohrožují podnikání, mohou zasáhnout reputaci značky a tím ohrozit její stabilitu a znevýhodnit ji v konkurenčním boji. Dezinformační útoky v posledních letech postihly například firmu ČEZ (odpojování elektroměrů), DHL (rozhazování koronaviru) nebo Sklárnu Harrachov (údajné zkrachování). (aktualne.cz, 2023) Zároveň mnoho firem dezinformace nepřímo podporuje tím, že skrz mediální agentury nebo portály nakupují inzertní prostor, a neověřují, zda se jedná o dezinformační weby. V některých případech se jedná o velké firmy i státní instituce. (aktualne.cz, 2021) Právě o odříznutí dezinformačních webů od peněz inzerentů se snaží spolek Nelež.cz

Odhady analýzy PSSI „Business of disinformation: stakeholder perspectives and way forward" tvrdí, že weby, zařazené organizací manipulatori.sk do seznamu platforem s problematickým obsahem, vydělávají on-line inzercí přibližné 190 000 kč měsíčně (většina částky je rozdělena mezi 10 nejnavštěvovanějších webů). (Prague Security Studies Institute, 2021) V roce 2021 bylo v Česku publikováno 197 177 článků na dezinformačních webech. (Evropské hodnoty o. p. s., 2022)

Finanční vliv na oblasti zájmu

Hlavním aspektem tohoto prostředí je snaha Ruska být nepostradatelným v ekonomických zájmech cílových států. Jde mu o to vytvořit na sobě nějakým způsobem ekonomickou závislost okolních zemí, například na dovážených nerostných surovinách.

Přes finanční transakce se také snaží ovlivnit vnitřní politiku okolních států a jejich směřování. „Tato strategie obvykle zahrnuje kolaboraci ruských energetických firem (z velké části vlastněných státem) s informačními agenturami, organizovaným zločinem a ambasádami ve snaze utratit peníze na nákup klíčových podniků v cílových státech, nebo darovat peníze politickým hnutím a politikům, což vede k jejich kompromitaci, a celkově uplatňovat skrytý vliv na lokální politiku. Projev této strategie prostupuje ruskou politiku od Baltského k Černému moři." (Bugajski, 2004)

Dalším aspektem ekonomického prostředí dezinformací a informační války je samotné generování, podporování a finanční stránka dezinformací.

Síť ruských hráčů tvoří mediální domy (Russia Today, Sputnik a jeho evropské satelity, Ruptly TV) – z části vlastněné státem nebo odkázané na státní finanční podporu, účty na sociálních sítích (trollové neboli Internet Research Agency[10], automatizované účty a falešné profily), oligarchové, obchodní zájmy přímo nezúčastněných osob, občanská sdružení a zájmové politické spolky, kyberzločinci, výzvědné agentury, soukromé firmy a spříznění političtí aktéři uvnitř i vně země. (Polyakova a Boyer, 2018) Někteří z těchto

10 *Ruská společnost zabývající se online propagandou a ovlivňováním. Je spojena s ruským oligarchou Jevgenijem Prigozhinem a sídlí v Petrohradě v Rusku. (*Prier, 2017)

aktérů jsou přímo napojeni na Kreml a plní jeho zadání a cíle, ostatní skrz pomoc Kremlu plní své vlastní cíle a zájmy. Tento ekosystém se neustále mění a vyvíjí, je v pohybu a jeho rozsah i členové jsou součásti, schované pod jeho diverzitou a spletitostí.

Tato síť pak dává Rusku nejenom silnou pozici v útoku, ale i schopnost velmi dobře odrážet útoky protivníka. Zároveň efektivně snižuje ekonomické náklady a zátěže spojené se správou dezinformací. Permanentní propaganda v médiích, loutkové profily na sociálních sítích a jednotná politika podpořená cenzurou imunizuje společnost proti případné kontrapropagandě nebo psychologickým operacím. Lidem u moci – oligarchům, podnikatelům a strategickým aktérům – je ze strany státu vycházeno vstříc a jejich zájmy jsou brány v potaz, a tak se zachovává jejich loajalita ke státu. Jejich podpora oproti trollům a hackerům není dána nacionalismem. Tuto premisu nám může potvrdit i odpor oligarchů k válce na Ukrajině (2022) a jejich otevřený nesouhlas s Putinovou politikou, který začali vyjadřovat ve chvíli, kdy jim Evropská unie zmrazila prostředky a ztížila mezinárodní obchod s Ruskem. (Harrington, 2022)

V rámci variability nástrojů se do strategie také zapojují různé vrstvy sítě aktérů, jejíž intenzita napojení na Kreml se liší, a jejichž zapojení může být Ruskem v případě potřeby popřeno, respektive jejich napojení na Kreml je jen těžko dokazatelné.

V rámci ekonomického prostředí také musí zaznít jedna významná skupina aktérů v dezinformačním procesu – ziskově orientovaní jedinci. Tyto lidé přebírají a sdílejí dezinformace na svých platformách a z návštěvnosti pak generují zisk díky inzercím. Informacím nemusí věřit, nebo s nimi souhlasit, a v procesu nemusí být vůbec zainteresovaní; může se jednat o třetí stranu ze vzdáleného státu, který nemá žádný konkrétní zájem na ovlivnění politiky a vývoje tím či oním směrem.

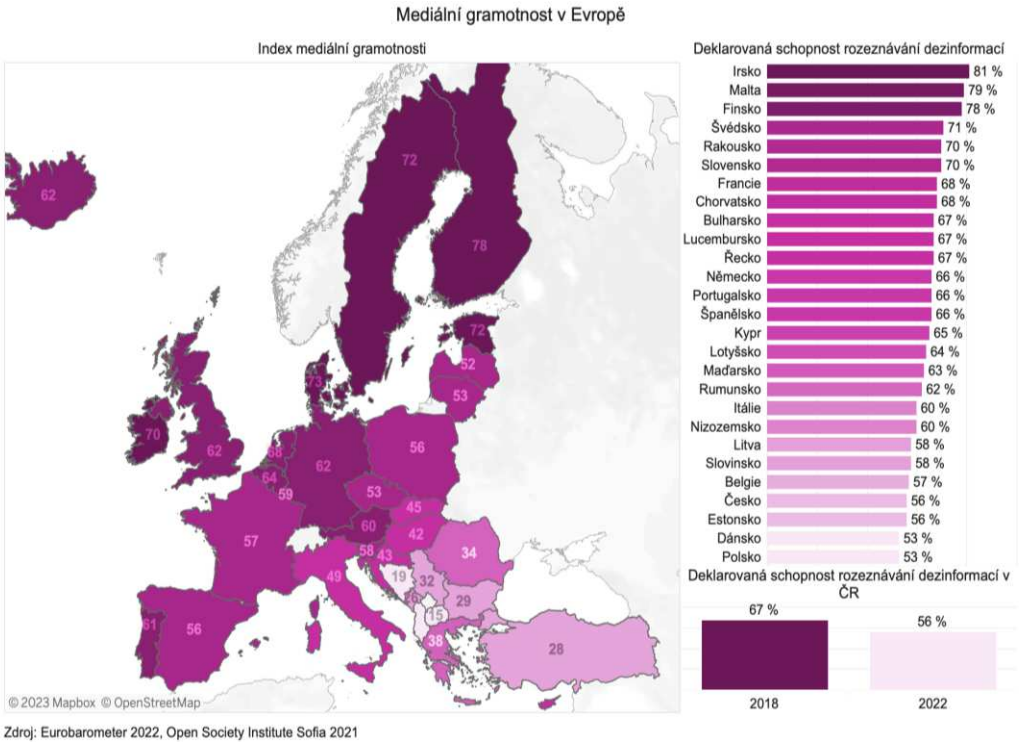
3.1.3. Sociální prostředí

V rámci sociálního prostředí vstupuje do hry mnoho aspektů, které ovlivňují společnost a činí jí náchylnou k dezinformacím. Stejně tak samotné dezinformace mají velký vliv na

proměnu společnosti a její celkové rozpoložení. Samotné téma této spletitosti mezi dezinformacemi a společností je příliš komplexní, než abychom ho mohli důstojně popsat v rámci této práce. Nicméně zmíním aspoň nástroje ovlivňující společnost, které jsou přímo spojené s tématem naší práce, tedy umělou inteligencí a počítačovou propagandou.

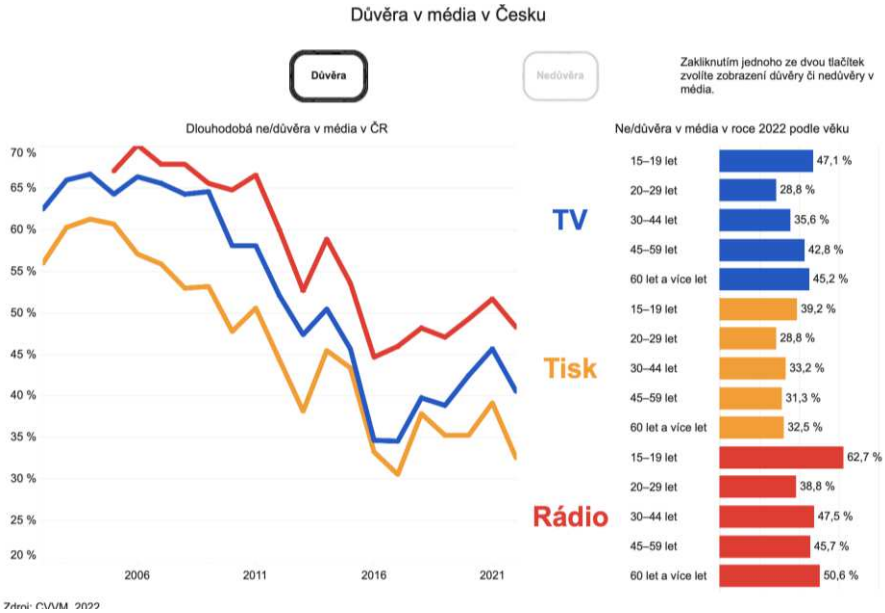
Obecným problémem v celé Evropě je nízká schopnost obyvatelstva rozpoznat dezinformační obsah, nebo ověřovat zdroje. Problematikou důvěry v média, mediální gramotností a konkrétními preferencemi občanů se zabývala instituce Česko v datech, která vytvořila několik grafů, shrnujících současný stav. (Česko v datech, 2022)

V indexu mediální gramotnosti si Česká republika vede průměrně, v porovnání s východní Evropou vykazuje větší gramotnost, ale zase pozůstává za západní Evropou. Oproti tomu ale jen 56 % Čechů věří, že dokáže rozeznat dezinformace, což je pokles oproti minulým letům. (Eurobarometer 2022, Open Society Institute Sofia 2021, Česko v datech, 2022.)



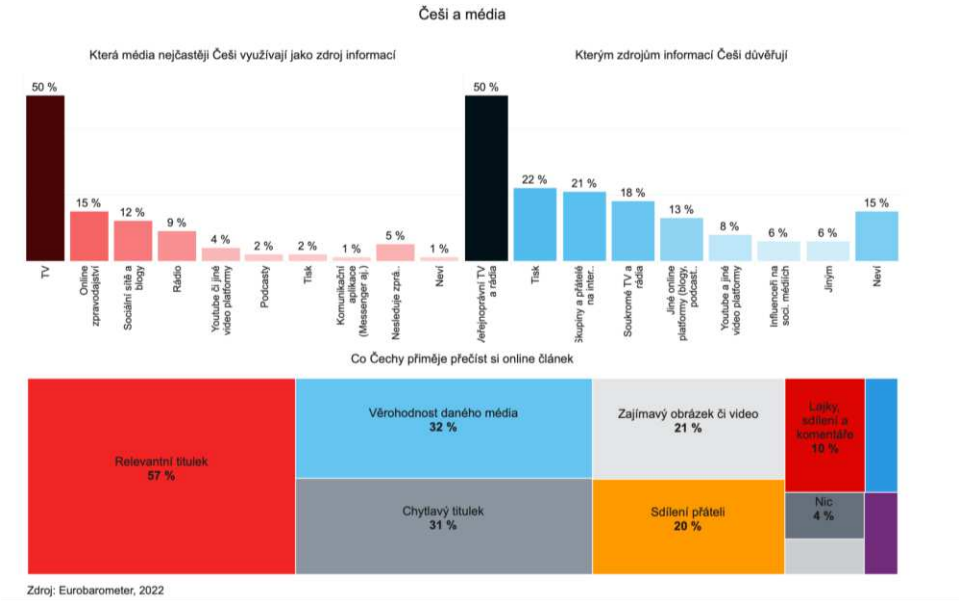
Mediální gramotnost v Evropě, Česko v datech, 2022

Důvěra Čechů v média vykazuje dlouhodobý trend poklesu, v roce 2021 bylo zaznamenáno mírné zlepšení, ale od té doby opět důvěra klesá. Nejmenší důvěru v média zaznamenává věková skupina mezi 20 až 29 lety, nejvíce pak skupina mezi 15 až 19 lety. Nejdůvěryhodnější je podle respondentů rádio, nejméně tisk. (Česko v datech, 2022, CVVM, 2022)



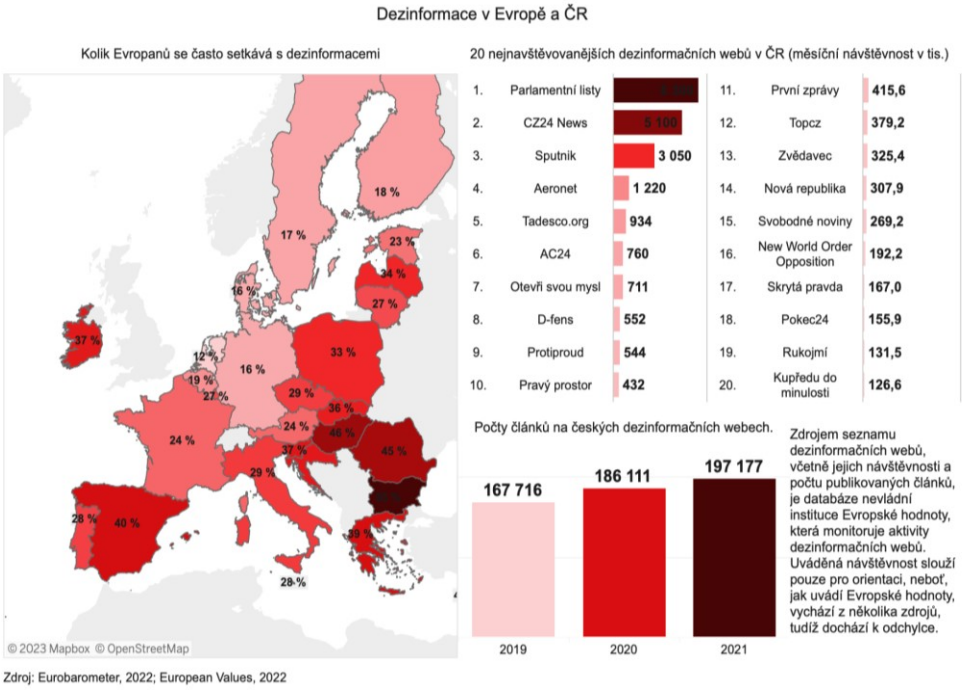
Důvěra v média v Česku, Česko v datech, 2022

Nejvíce Čechů získává informace z televize, nejvíce věří veřejnoprávní televizi a rádiu. Nejzajímavější pro ně je relevantní titulek, roli hraje i obrázek a sdílení přáteli. (Eurobarometer, 2022, Česko v datech, 2022)



Češi a média, Česko v datech, 2022

Podle výzkumu se 29 % Čechů setkává často s dezinformacemi, nejčastějším zdrojem jsou Parlamentní listy. Opět pozorujeme trend menšího vystavení dezinformacím v porovnání s východní Evropou, ale většího oproti západní. Hodnoty jsou podobné jako v jižní Evropě. (Česko v datech, 2022, Eurobarometer, 2022)



Dezinformace v Evropě a ČR, Česko v datech, 2022

Podkopávání důvěry ve stát

Velkou hrozbou je i samotný design strategie. Ruská strategie se zaměřuje na tzv. „firehose of falsehood“, tedy snahu opakovaně, rychle a plošně šířit objemné nepravdy na všechny strany z různých zdrojů. (Polyakova a Boyer, 2018) Cílem tohoto snažení není prosazení jednoho perfektního narativu, ale relativizace pravdy. Vytvořit iluzi, že jasný rozdíl mezi pravdou a lží neexistuje, že hranice je rozmazaná a že vše má něco do sebe, ale zároveň se nedá věřit oficiálním narativům. (Gregor a Mlejnková, 2021)

Proto se šíří prostředím internetu na první pohled protichůdné informace, každá událost má několik verzí, a i zdánlivě nedůležitá fakta jsou zpochybňována a relativizována. Tato strategie vytváří iluzi, že nic jako pravda neexistuje, a skrytě podkopává důvěru k oficiálním institucím.

Nástroji, ovlivňujícími společnost, jsou i kybernetické útoky a hacking. Prvním zajímavým společenským aspektem hackingu je, že tyto útoky jsou často spojovány s ruskými obyvateli, kteří svoji službu pro kybernetické útoky dobrovolně nabídli. Jejich nábor by se pak dal přirovnat k náboru vojáků do armády, u kterých je cíleno na vlastenectví a apel na povinnost ke svojí zemi a její ochraně. Cílem útoků může být jak snaha získat kritické informace, destabilizovat nepřátelský subjekt nebo získat vliv, tak jen pozorovat. Stejně jako ostatní složky jsou pečlivé cílené a dlouho naplánované. (Polyakova a Boyer, 2018)

Obzvlášť atraktivním cílem pro hacking pak mohou být instituce, na které se obyvatelstvo spoléhá, nebo je na nich závislá, například média, nemocnice nebo banky, jejichž destabilizace by vedla k masové panice. Právě panika pak ohrožuje důvěru společnosti ve stát a demoralizuje společnost. (Johnson 2019) Zde se při identifikaci ohrožených cílů může využít zapojení poznatků sociálních věd a následně alokovat technické zdroje pro zvýšení jejich ochrany.

Společnost a sociální sítě

Společnost je také ovlivněna sociálními sítěmi, které jsou přirozeným prostředím pro šíření dezinformací. Sociální sítě mohou v lidech vyvolat negativní pocity jako vztek, deprese, strach a úzkosti. (Caramancion, 2020) Pozorujeme spoustu příčin těchto negativních emocí, ať už to je FOMO, přehnaná komparace s ostatními, přesycení podněty a informacemi, nebo články, které podněcují strach; sociální sítě v nás umí vyvolat mnoho negativních pocitů. Emoce jako strach nebo vztek pak snižují schopnost vyhodnocovat a přemýšlet „s rozvahou“, jak již bylo zmíněno v heuristikách výše.

V ekosystému sociálních sítí tedy můžeme mít snížené schopnosti vyhodnocování a ostražitost se uživatelům zhoršuje. Dezinformace jsou designované tak, aby vyvolaly emoce, nebo se na ně aspoň vázaly. Právě tato vysoká orientace na emoce zaručuje úspěch dezinformace. U uživatelů s podobným smýšlením umocní jejich hodnoty, u neutrálních uživatelů může zasít semínko nedůvěry, nebo je zmást, a u odpůrců vyvolá silné emoce, které chtějí ventilovat ve snaze vyvrátit danou dezinformaci v sekci komentářů. Případně

vyhrocené diskuse pak posilují jednak názory obou skupin na opačných stranách názorového spektra, ale hlavně díky nadprůměrným reakcím pak algoritmus tyto příspěvky doporučuje i jiným uživatelům, protože jej mylně vyhodnotí jako zajímavé a relevantní. Tím se dezinformace šíří efektivněji a dostává se k více uživatelům i mimo svojí obvyklou bublinu. (Shu, Kai a kol., 2020)

Vliv informační války na společnost

Vliv dezinformací a informační války na společnost a její zapojení je pak konkrétně popsán na případu první informační války v Estonsku.

Case study: První informační válka a její pozůstatky

Za první informační válku je dnes v akademickém prostředí označován konflikt zájmů mezi Ruskem a Estonskem. K tomuto incidentu došlo mezi dubnem a květnem roku 2007 a jeho důvod i pozadí nám, jako další postsovětské zemi, může připadat minimálně povědomé. (Hughes, 2007)

Celá situace se zrodila kvůli odstranění sochy bronzového vojáka v Talinu, který měl být poctou osvoboditelským vojskům Rudé armády. (Blank, 2008) Z politologického hlediska by se mělo jednat o jednoduchý incident, jeden stát odstraňuje monument spojený s historií svou a druhého státu, který ovšem vojensky reagovat na takovou malichernost nemůže, protože by tím porušil svrchovanost prvního státu. Společenský význam odstraňování monumentů je také již dobře zmapovaný, jedná se o vymezení se vůči historickému vlivu, odmítnutí určité věrnosti minulým narativům.

Paralelně s rámcováním narativu o „ochraně“ Estonska před západním vlivem docházelo k útokům na kritickou komunikační infrastrukturu Estonska, které měly za cíl destabilizovat vládu, zpochybnit její kompetence a pospolitost země a zastrašit politické vůdce. Jedním z nástrojů byly botnet útoky (ovládnutí cizího počítače, nad kterým jeho

majitel ztrácí kontrolu, a tento počítač je pak využit k útokům na infrastrukturu), hacking a kybernetické útoky na vládní infrastrukturu. (Blank, 2008)

Dalším zajímavým nástrojem, u kterého lze polemizovat nad způsobem obrany proti němu, byly násilné demonstrace v Talinu. Ty byly spojeny s ruským obyvatelstvem v Estonsku, a demonstrace v Rusku před Estonskou ambasádou s ruským spolkem Nashi spojujícím proputinovskou mládež s dalšími zastánci Putina. I tyto demonstrace, jak se později ukázalo, byly plánovány již rok dopředu a mezi její účastníky patřily i ruské speciální jednotky v civilním převleku. (Blank, 2008) Tyto demonstrace pak byly dostatečně veliké, aby v nich mohly propuknout násilné střety. Právě tyto střety by pak mohly posloužit Moskvě jako záminka k vojenskému zásahu – „pomoci“ (Naplnění tohoto scénáře jsme pak mohli pozorovat na násilném potlačení demonstrací v Bělorusku s pomocí ruské armády nebo podporu separatistů na Krymu.). Podvržené demonstrace pak získávají na důležitosti v prostoru sociálních sítí. Zde jsou sdílená videa a obraz manipulovány tak, aby podpořily žádaný narativ. Ačkoliv domácí obyvatelstvo třeba ani nemusí o demonstracích nebo jinému incidentu vědět / vnímat ho pouze okrajově na sociálních sítích, mohou se tyto události jevit daleko závažnější nebo větší, než skutečně byly.

Hlavní strategie informačních operací se zakládala na snaze zvýšit násilí v Estonsku, a tím snížit důvěru tamního obyvatelstva ve vládu, a rámcovat Estonsko v dalších postsovětských zemích jako fašistický režim. (Blank, 2008) Dále byly zapojeny již zmíněně kybernetické útoky, ekonomické sankce, manipulace mládežnických organizací a gangů a ruská snaha o proniknutí do klíčových odvětví Estonska. Jak se později ukázalo, celá tato strategie se zrodila ještě o rok dříve, než byla samotná socha odstraněna, nebo bylo jen plánováno ji odstranit. (Blank, 2008) Již samotná snaha o proniknutí do klíčových sektorů je dlouhodobě plánovaná a realizovaná operace na území téměř všech postsovětských států (jedná se o vojenské, energetické nebo politické sektory, či proniknutí do inteligence) má sloužit jako pojistka pro případ konfliktu. (Conley, 2016)

Na situaci můžeme pozorovat i konzistenci narativu aktéra a nástrojů. Tyto nástroje byly mnohdy hojně používány už Sovětským svazem a „cílí na ovlivnění světových událostí skrz manipulaci médií, společnosti a politik – proti demokraciím“. (Herd, 2022)

Všechny výše popsané prvky byly odzkoušeny a použity na Estonsku a následně vyvíjeny a zdokonalovány pro další využití. Vytouženým cílem bylo destabilizovat estonskou vládu a společnost, poukázat na neschopnost NATO chránit země proti tomuto druhu hrozby a celkově donutit Estonsko, aby při svém rozhodování bralo v potaz i zájmy Ruska.

Snahou Ruska nebylo pouze ovlivnit Estonsko a jeho orientaci a napravit jeho chování, ale také ho potrestat za neposlušnost vůči „nadřazené autoritě“. (Cory, 2017)

První zaznamenaná a popsaná informační válka nám také pomohla položit definice a nakreslit šablonu strategie a dalšího vývoje ruského působení na zájmové státy. Ačkoliv se jedná o novou hrozbu, její strategie je povědomá. Spojuje v sobě Leninovu taktiku podvracení a zastrašování a leninistickou ideu, že Rusko je pod neustálou hrozbou a musí se bránit. Snaží se izolovat nepřítele od spojenců a cílit na bratrskou slovanskou pospolitost, opírá se o dlouhodobé sebeurčení o ruské nadřazenosti a jeho nepopíratelné právo na nadvládu nad Pobaltskými zeměmi. (Page, 1950) Ideje ani narativ se tedy nemění, mění se pouze technologie.

3.1.4. Technologické prostředí

Technologické stránce informační války z pohledu České republiky se věnuje většina této práce i následná diskuse, Takže bude tato sekce pouze okrajová.

Zajímavostí informační války je, že Rusko pro tyto útoky nepotřebuje žádnou vyspělou a sofistikovanou technologii a podpůrnou síť infrastruktury, protože dokáže využít své obyvatelstvo a buňky, jak bylo popsáno v oddílu „Sociální prostředí“. V tom je ruský způsob informační války podobný klasickým teroristickým útokům. Není potřeba velká

organizace nebo centralizovaná technologie, jako požadovaný cíl stačí dostatek ideologie a chaos. (Polyakova a Boyer, 2018)

Mezi metody Ruska patří plošné šíření falešného a zavádějícího obsahu skrz vlastní satelity v jiných zemích, šíření propagandy přes automatizované účty na sociálních sítích a společenský rozkol, tedy posílení debat vedoucích k rozdělení společnosti.

Zároveň Rusko stejně jako Čína aktivně uskutečňuje výzkum a vývoj AI pro vojenské účely. Vzhledem k relativně nízké morálce autoritářských států (není zde potřeba legitimizovat se a obhajovat před svými voliči), legislativě a etice v tematice zbraní obecně je toto zbrojení vnímáno jako jedna z hlavních hrozeb do budoucna. (Johnson, 2019)

Díky obrovské pozornosti, které se výzkumu AI dostává (a hrozby, které její zneužití představuje), se i stát a EU začaly angažovat ve sdružování odborníků v rámci profesních fór a alokovat více zdrojů na podporu tohoto výzkumu.

Praktická část

Nejprve je nutné si připomenout, jaký cíl tato práce sleduje. Jako cílovou skupinu této práce můžeme identifikovat šest důležitých recipientů možných AI nástrojů. Média – tedy tvůrce a správce zpravodajského obsahu, komerční subjekty, které mají dost kapitálu na financování výzkumu pro ochranu své činnosti, akademiky, širokou veřejnost – jakožto hlavní cíl dezinformačních aktivit, vývojáře nástrojů třetí strany – skupina financovaná soukromým subjektem, která svoji aktivitu vynakládá za účelem získat finanční odměny a prostředí – která pod sebou zahrnuje vše, na co naši předchozí aktéři reagují. Tito aktéři budou vstupovat do pojednávání o tomto tématu a aspoň část z nich bude reprezentována i ve focus group.

4. Focus group

4.1. Představení metody focus group

Focus group je kvalitativní metoda výzkumu a zisku dat od záměrně a cíleně sestavené skupiny předem vybraných jedinců. (Morgan, 1998). Focus group je definovaná jako technika sbírající data přes interakci ve skupině při diskusi témat nastolených výzkumníkem. Součástí této definice jsou tři hlavní pilíře. Za prvé je focus group výzkumnou metodou věnovanou sběru dat. Za druhé lokalizuje interakci ve skupinové diskusi jako zdroj dat. Za třetí uznává aktivní roli výzkumníka při vytváření skupinové diskuse pro účely sběru dat. (Morgan, 1996)

Strukturální diskuse obvykle probíhá mezi 6 až 10 jedinci, kteří mohou mít preexistující vztahy. Rozeznáváme různé druhy zapojení moderátora, míru direkce, počet setkání a skupin i pole pozorování. (Smithson, 2000)

4.2. Představení výzkumu

Metoda focus group byla vybrána pro tento výzkum pro svou schopnost poskytnout komplexnější chápání motivací a kontextu problematiky a své flexibilitě. (Morgan, 1996) Hlavní síla focus group leží v jejím rozdílu od samostatných individuálních rozhovorů, oproti ní se účastníci navzájem dotazují a vysvětlují si jednotlivé postoje. (Morgan, 1996) Největší informativní hodnota je pak právě v interakci ve skupině a tyto informace budeme sledovat i my. (Smithson, 2000)

Velmi důležitou součástí výzkumu je moderátor. Ten je odpovědný za pokládání otázek a koordinaci odpovídání. Role moderátora a jeho vystupování může ve svém důsledku ovlivnit, a při nevhodné míře zásahu i zkreslit, výsledky studie. (Morgan, 1996) Pro účely této metody zvolím méně direktivní způsob řízení diskuse, moderátor bude pouze pokládat otázky a diskuzi mezi participanty nechá plynout volně. Tento postup volím i z důvodu složení participantů ve skupině, jelikož se jedná o expertní skupinu, nepovažuji za důležité moderovat míru prostoru pro vyjádření (aby se všichni zúčastnili diskuse stejně), nebo příliš zasahovat do agendy (i vedlejší témata plynoucí z debaty pro nás mají hodnotu).

Hlavním cílem výzkumu je kriticky zhodnotit současný vývoj AI, odpovědět na otázky vyvstávající z předešlého textu a diskutovat predikce a vývoj do budoucna stejně jako potřeby legislativního nebo etického rámce.

Hlavní limity této metody mohou být jednak moderátorovo zkreslení (eliminuji nízkou direktivitou v debatě), nebo chybně zvolená témata (vhodnost výzkumu pro toto téma bylo pečlivě zváženo a diskutováno). Častým limitem je i polarizace vznikající v diskusi, nicméně tento limit pro nás není tak aktuální, nejde nám o změnu přístupu nebo zamezení konsenzu, spíše o prozkoumání jednotlivých postojů.

4.2.1. Design výzkumu

Úspěch focus group tkví ve vhodném designu, pečlivém plánování, dostatku vhodných respondentů, volbě relevantních otázek, vhodného moderátora a kvalitní analýze výsledků. U designu výzkumu provádíme několik důležitých rozhodnutí.

Prvním je míra zapojení moderátora a jeho styl vedení. Jak již bylo zmíněno výše, volíme způsob volné diskuse, kde moderátor slouží hlavně k uvádění otázek a případných podotázek, do rozložení diskuse mezi respondenty pak nezasahuje. Vedení diskuse tedy bude méně strukturované.

Dalším důležitým rozhodnutím je velikost skupiny. Jelikož oslovujeme odborníky, kteří mají mnoho relevantních znalostí a názorů na danou tematiku, a chceme jim zajistit dostatek prostoru a efektivní interakci, volíme menší skupinu 5 lidí.

Respondenti byli zvoleni na základě své účasti ve veřejné diskusi o vývoji a budoucnosti AI a svých znalostech o dezinformacích a informační válce. Ve skupině chceme pokrýt různá odvětví a postoje, kterých se vývoj AI a dezinformace týká. Další metodou výběru respondentů pak byla technika sněhové koule, vybraní respondenti byly osloveni s dotazem na doporučení dalších respondentů, z nichž byli následně vyselektováni další.

4.2. 2. Respondenti

N základě všech rekrutačních kritérií jsou členy FG Mgr. David Klimeš, Ph. D (dále DK), vyučující na FSV UK a společenský komentátor pro různá média. Dalším participantem je Mgr. Vojtěch Bahenský (dále VB), student Ph. D., pedagog Univerzity Karlovy a výzkumník, zabývající se oblastí bezpečnostních studií. Dalším participantem je Ing. Adam Zdvihal (dále AZ), konzultant pro strategický rozvoj středně velkých a korporátních zákazníků v oblasti investic do rozvoje a technologií. Skupinu doplní i výzkumník Tomáš Horych (dále TH), který se zaměřuje na výzkum a vývoj nástrojů umělé inteligence pro detekci bias v médiích a v současné době je součástí výzkumného

týmu, vyvíjejícího monitoringový nástroj pro komerční užití v Tokiu. Posledním respondentem je novinář Martin Spirit (dále MS), který se tematicky zaměřuje na technologie a umělou inteligenci, jejich popularizaci a roli ve společnosti.

Výzkum bude proveden formou otevřených otázek, nad jejímiž odpověďmi budou spolu účastníci diskutovat. Jejich argumenty a závěry budou zaznamenány, zhodnoceny a následně prezentovány v této práci. Focus Group pak doplnil ještě individuální rozhovor s Mgr. et Mgr. Tomášem Kučerou, Ph.D. (dále TK), vyučujícím na katedře bezpečnostních studií, zaměřující se na bezpečnostní politiku České republiky a technologii a válečnictví.

Bude se jednat o výzkum jedné skupiny při jednom setkání, skupina bude nově složena pro účely výzkumu, nicméně někteří participanti už mají preexistující profesionální vztahy. Diskuse bude zaznamenaná pomocí videa doplněná o field poznámky a celá interakce pak bude přepsána a přiložena formou transkriptu k této práci. Stejným způsobem bude zaznamenaný i individuální hloubkový rozhovor. Všichni participanti souhlasili s uveřejněním jména a obsahu rozhovorů.

4.2.3. Scénář

Vítejte a děkujeme, že jste se dnes připojili. Účelem tohoto setkání je získat váš názor na současný vývoj AI nástrojů v České republice a jejich možné využití, limity a predikce ve strategickém boji v informační válce. Závěry tohoto setkání budou shrnuty a zahrnuty v diplomové práci, která si klade za cíl přispět k současné debatě o AI.

Práce se bude zabývat současným stavem AI technologií v provázání s dezinformacemi a informační válkou, jejich možným vývojem, limity a obavami, vychází z již ustálených konceptů, prací mapujících současný stav, ale i predikujících budoucnost, formulujících možné limity a pokládajících otázky do budoucna. K pochopení současného stavu jsou prezentovány i různé případové studie relevantní k tématu. Závěry a podněty vzešlé z

teoretické části a analýz pak posloužily k sestavení otázek do focus group. Práce se doptává na kompetence, financování, predikce a hrozby.

Dovolte mi, abych se představila. Jmenuji se Barbora a dnešní diskusi budu moderovat. Formát, který používáme, je focus group. Rozhovor povedu tak, že budu klást otázky, na které bude moci každý z vás odpovědět. Pokud chcete, můžete také reagovat na komentáře ostatních jako při běžném rozhovoru. Mým úkolem je zajistit, abychom se drželi tématu. Sezení se bude nahrávat, aby vaše podněty mohly být posléze zhodnoceny.

Než začneme, chci vás upozornit, že informace, které se dnes dozvíme, budou shrnuty do diplomové práce. Tato sekce bude obsahovat shrnutí vašich připomínek a některých doporučení. Jak vidíte, tuto focus group budeme nahrávat. Z nahrávky bude vytvořen transkript, který bude součástí výsledné práce. Nahrávka samotná zveřejněna nebude.

Začněme s představováním.
Sdělte nám, prosím, své jméno a obor vaší specializace / vašeho zájmu.

Obsahové otázky:

a) Financování:

1. Kdo by podle vás měl / může financovat výzkum AI nástrojů pro boj s dezinformacemi využívané státem? Kde se mají alokovat peníze, a jak moc akutní vnímáte potřebu zajistit větší financování?

b) Microtargeting a data:

2. Jak vnímáte profilování uživatelů, využívání big data a mikrotargeting, jde podle vás o etické nástroje? Jak moc velkou podle vás představuje hrozbu pro demokratický proces? Jak moc dat bychom měli sbírat a pro jaké účely?
3. Měla by se podle vás data uživatelů více chránit? Měl by mít uživatel přehled o tom, kdo všechno využívá a přeprodává jeho data?

c) Odpovědnost za content a nakládání s ním:

4. Měl by fact-cheching a odebírání kontentu ležet v rukou vlády? Měly by mít pravomoc zasahovat do obsahu na sociálních sítích a mazat informace? Měla by tato pravomoc ležet v rukou veřejných institucí, které nepodléhají přímo vládě? Jak by měla tato instituce vypadat, a komu se zodpovídat? Jaké by měla vynucovací pravomoci?
5. Měla by pravomoc a odpovědnost za obsah a jeho mazání ležet v rukou dané platformy? Jak by měla být platforma kontrolována a případná opatření vynucována?
6. Jak podle vás vypadá ideální fact-checkingový proces, míra odstraňování a cenzury? Jak by mělo být nakládáno s nepravdivým obsahem?

d) Role člověka v hybridním přístupu:

7. Jaká je role a důležitost zachování člověka v procesu fact-checkingu? Kde ho může nahradit AI a kde je naopak nezbytné jej zachovat?

e) Fact-checking, gramotnost, důvěra a bias:

8. Kdo by se měl angažovat v šíření mediální a informační gramotnosti? Jak by podle vás měla být zvýšena mediální a digitální gramotnost? Myslíte si, že odpověď na problémy s dezinformacemi je právě její zvýšení?
9. Jak by se podle vás mělo zajistit, aby v training datech nebyl žádný sklon, jak je posuzovat? Jaký proces by podle vás byl dostatečně spolehlivý a efektivní?
10. Jak moc zásadní vnímáte potřebu, aby lidé věřili ve fact-checkingové / AI systémy? Měla by to být priorita?

Techplomacie – otázka v individuálním rozhovoru:

1. Vnímáte techplomacii jako příležitost do budoucna, jak by měla být využita?

Vidím, že náš čas vypršel. Děkujeme vám, že jste se s námi podělili o tyto užitečné informace.

4.3. Průběh a výsledky

Níže uvedeme a shrneme hlavní výsledky vzešlé z expertní debaty na základě vytýčených tematických okruhů a individuálního hloubkového rozhovoru.

a) Otázka financování výzkumu nástrojů AI k detekci dezinformací, role státu a soukromých subjektů a možné etické limity

Prvním podnětem VB vzešlým z této otázky bylo, že financování je klíčové z hlediska akademického výzkumu, granty a finanční prostředky v této sféře determinují, zda se výzkumu někdo věnuje, nebo ne, je na něm přímo existenčně závislý. Zároveň se ale zamýšlí nad tím, že tyto prostředky mohou být z důvodu zadání státní zakázky využity někým, kdo by z expertního nebo etického hlediska takový výzkum neměl provádět. Státní správa by v ideálním případě měla být schopná specifikovat své požadavky a následně financovat vývoj nástroje, který chce používat. Tímto otevřel další problém, týkající se optimálního nastavení definic a rozlišování mezi aplikovaným výzkumem a zakázkou.

Druhý pohled ze strany průmyslu, respektive soukromého sektoru přinesl AZ. Ten zmínil obecný vztah státu a byznysové sféry, který je naplněný skepsí ke státním nástrojům a jejich fungování. Vnímá přínos, který může mít soukromé financování do rozvoje nástrojů, například urychlení procesu vlivem tržní konkurence. Motivaci soukromého sektoru vidí jednak ve snaze snížit riziko ztráty (datové chyby, misinformace v procesu výroby), nebo naopak optimalizovat zisky. Místo pro financování ze strany soukromého sektoru vidí u velkých firem i proto, že mohou státu vytvářet konkurenci a tím hnát vývoj a vytvářet tlak. Zároveň zmiňuje, že soukromý sektor je dlouhodobě schopen lépe ohodnotit práci výzkumných pracovníků, jejich výsledky pak mohou například poskytovat na bázi open-source k využití státem. Také specifikoval, že misinformace v průmyslovém světě jsou právě číselná data, která už z podstaty je těžké odhalit, jelikož drtivá většina firem je stále datové řízená, vidí v této hrozbě jednu z těch hlavních.

Respondent DK oproti tomu vidí, s ohledem na odpovědnost k demokratické společnosti, hlavní otázku míry. „U těch věcí, které jsou nejradikálnější ve smyslu dopadu na společnost, tak tam chci, aby je stát vykonával s nějakými dobrými podklady a nástroji. Zároveň, když se posouváme do nějakého firemního prostředí nebo v mém případě třeba

mediálního prostředí, tak rozhodně si nemyslím, že by měly být velké státní výzkumy na to, jakým překladačem co nejchytřeji přepisovat články.” Roli státu vidí více v oblastech, které zasahují do členských a osobnostních práv, a méně například v průmyslovém designu. Ve vývoji byznysu, například médií nebo propojení s IT, by se měla nechat větší volnost soukromé soutěži.

TH, který je financovaný z akademické sféry, souhlasí, že by soukromý sektor neměl financovat interdisciplinární nebo aplikovaný výzkum v oblasti dezinformací, například algoritmy pro detekci fake news, ten by měl být financovaný státem. Nicméně soukromý sektor může financovat nějaký základní počáteční výzkum například jazykových modelů, i například proto, aby české prostředí bylo schopné konkurovat na mezinárodní úrovni.

Dále reagoval VB shrnutím, že soukromý sektor bude mít zájem na financování vývoje vlastních nástrojů pro vlastní použití, a že stát si poté nebude muset financovat vývoj modelů od nuly. Předpokládá, že stát bude v nějaké míře podporovat základní výzkum, ale soukromý sektor bude tahoun tohoto typu inovací. Jeho skepticismus přichází v rovině přejímání konkrétních modelů a aplikací soukromého sektoru, nebo přílišného propojování užívání v soukromém a veřejném sektoru. Jako příklad uvádí společnost Semantic Vision, která ve své činnosti sleduje i vlastní zájmy a komerční cíle a spolupráce s veřejným sektorem pro ně slouží i jako PR jejich činnosti. Pokud soukromá firma takto funguje ve službách veřejnosti, je podle něj zásadní transparentnost a open-source. Průběh a fungování je u této společnosti tajemstvím a veřejné jsou jen výsledky, není pak možné přezkoumat závěry a ověřit funkčnost a férovost. U nástrojů, které mají sloužit veřejnosti, potažmo státu, tuto absenci transparence vnímá jako problematickou.

b) Využívání, vliv, etika a hrozba microtargetingu, big data a využívání a ochrana dat

Pokud jde o téma využívání a regulování těchto nástrojů, klíčovou věc vidí DK v transparentnosti spíše než v omezování a zakazování, přičemž je nutné brát v úvahu rozdíly mezi velkými a malými firmami. Vnímá potřebu veřejné debaty a stanovení limitů, aby byla zachována transparentnost v digitálním prostředí. Také zmiňuje, že se neztotožňuje s některými ideologickými názory, například s knihou No logo od Noemi Klein. Na druhé straně si však myslí, že je důležité mít diskusi o digitálním aktu a o jeho ovlivňování života jednotlivců. Dle jeho názoru, pokud se klenba jednoho loga,

respektive jedné firmy, nad námi technologicky jednou zaklene, protože uživatel bude přistupovat skrze jednu platformu k různým dalším a bude u toho cílen, mělo by to podstoupit veřejnou diskuzi a mít nějaké limity. Ale přesto si myslí, že nejdůležitější je transparence, kterou je nutno zajistit vytvořením nějakého tlaku.

Transparenci však nevidí v odhalení všech zdrojových kódů, relevanci vnímá ve chvíli, kdy to ovlivňuje něco na úrovni státu, například zveřejnění metodiky moderace obsahu a metodiky zpracování dopadových studií například na čtvrtletní bázi. Mít stanovené indikátory, značící potřebu úprav nebo varování.

TH navázal shrnutím, že hypercílení už je rozjetý vlak a způsob, jakým všechny velké nadnárodní společnosti generují revenue. Přiklonil se k myšlence Michala Pěchoučka, který nedávno komunikoval v podcastu CzechCrunch, že té transparentnosti bychom mohli v budoucnu docílit třeba tím, že by uživatelé měli přístup ne nutně k datům, která daná platforma sbírá, ale k samotnému koncovému data profilu, na základě kterého ty velké společnosti cílí a generují doporučený obsah, což by mohlo vypadat tak, že by uživatel viděl svůj digitální otisk na dané platformě a mohl by nějakým způsobem regulovat, kam až to doporučování může sahat, editovat tato data, určovat, jestli chce vidět tyhle reklamy, nebo jestli chce vůbec vidět něco personalizované. Zároveň věří, že personalizace a hypercílení vytváří nějaké sociální bubliny, na druhou stranu si ale nemyslí, že by se uživatelé dobrovolně vzdali cíleného obsahu, protože je to zábava.

VB uchopil téma z jiného konce: vidí chybu v tom, jak se přistupuje v této oblasti z pohledu demokracie. Zmiňuje, že se často ukazuje, že mikrocílení nezvyšuje efektivitu, ale naopak ji vlastně v nějakém ohledu snižuje, protože je cíleno na lidi, kteří by si mysleli to samé, nehledíce na to, co se jim ukazuje. Stejně tomu je u řady dezinformací – například v amerických prezidentských volbách, kde vzniklo velké množství výzkumů (primárně na Twitteru, který je v tom přístupnější), obrovskou část tohoto obsahu konzumovali už rozhodnutí voliči. Aby microtargeting byl účinný, musel by být technologicky už tak vyvinutý, aby dokázal změřit uživatele, kteří nejsou dané zprávě tolik nakloněni, ale jsou zranitelní k tomu nechat se ovlivnit. Zmiňuje, že otázka, zda je technologie už takto vyvinutá, je těžko zodpověditelná, protože firmy jako Facebook a Cambridge Analytica mají komerční zájem na tom, aby byly jejich schopnosti a účinek

mikrocílení přeceňované a prezentované jako velmi efektivní, i když realita může být jinde.

Zmiňuje epizodu, kdy Zuckerberg, když mu bylo poprvé řečeno, že trollové na Facebooku zvolili Trumpa, popřel, že by Facebook měl takovou moc. Následně otočil, protože pokud nemají moc ovlivňovat voliče, aby volili Trumpa, tak asi taky nemají schopnost ovlivňovat konzumenty, aby si koupili ten správný produkty, což by šlo proti jejich byznys modelu.

K účinkům microtargetingu na ovlivnění voličů je sice skeptický, naopak ale vidí jejich hrozbu v oblasti radikalizace, kdy už názorově vyhraněným jedincům je podsunuta upravená informační dieta. I když velikost této hrozby si netroufne odhadnout.

Z pohledu AZ má microtargeting potenciál být pozitivní, pokud jsou data o zájmech a preferencích lidí využita k vytváření konkrétních produktů, které by je zajímaly, ale které neexistují. Analyzování velkých datových souborů (tzv. big data) může poskytnout indikace o tom, co trh nebo region požaduje. Zmiňuje tvorbu velmi personalizovaných produktů pro úzkou skupinu lidí a věří, že pokud ví, co zákazníci chtějí, může jim to poskytnout bez toho, aby se jich tolik ptal a zároveň si tak získat jejich loajalitu.

MS souhlasí s přínosy, které microtargeting může přinést do průmyslu. Zdůrazňuje ale, že echo chambers a sociální bubliny jsou přímým produktem microtargetingu a že vytvářejí mezilidské příkopy, které se budou prohlubovat s rozvojem hyper a micro targetingu. Formuluje myšlenku, zda není třeba zakročit proti velkým korporacím jako je Facebook, který tyto trendy podporuje. Zmiňuje hrozbu, kdy se můžeme dostat do bodu, kdy se každý bude pohybovat ve svém vlastním světě personalizovaného obsahu a přestaneme si rozumět, což může být nebezpečné až neudržitelné.

Na debatu reagoval DK, který zmínil potřebu uvědomit si typického cílového konzumenta. Je skeptický k přístupu k digitálnímu otisku, respektive zájmu běžného uživatele o tento digitální otisk a jeho schopnost s ním jakkoliv nakládat a zmiňuje například GDPR, kterou běžný uživatel bezmyšlenkovitě a automaticky odklikává. Vnímá potřebu stanovit si typického defaultního spotřebitele, který je reálný a není představou omni potentního občana, který perfektně zná svá zákonná práva a celý den

tráví nad svým digitálním otiskem. Také zmiňuje jako zajímavější než ovlivňování mas ovlivňování konkrétních opinion leaderů nebo významných aktérů, kteří udávají agenda setting. Obává se, že ovlivňování například 20 nejvýznamnějších osobností daného tématu představuje větší nebezpečí než ovlivňování velkých voličských skupin, protože tito jedinci pak mohou ovlivnit celou společnost. Závěrem shrnuje, že obecně není dobré doufat v omni potentního občana a že osvícenská idea je lichá.

VB následně reagoval konstatováním, že ačkoliv si nemyslí, že by stát nebo vláda dokázala těchto nástrojů využít z hlediska technologického pokroku, určitě je zajímavé zamyslet se nad tím, jak by vypadalo řízení státu podle zjištěných preferencí občanů, a riziko, které by to s sebou mohlo přinést, kdyby se přizpůsoboval určité skupině lidí, jejichž preference dokáže získat.

Posledním podnětem k této otázce pak byla reakce MS, který se zamyslel nad povahou dat Google a Meta, která jsou ostatním nedostupná. Myslí si, že s daty, která má Google a Meta, dokážou tyto společnosti mnohem více, než jenom doporučovat personalizovaný obsah, ale nějakým způsobem i vmanipulovat uživatele do určitých clusterů, kde jsou pak obecně lépe predikovatelné.

c) Odpovědnost za fact-checking, pravomoci jednotlivých aktérů moderovat obsah, regulace a možnosti nakládání s contentem

Úvodu debaty se uchopil DK, který opět hovořil o potřebě nalézt vhodnou míru. Mimo jiné formuloval mírnou úlevu, že zkrachovala vládní strategie pro média a dezinformace, protože preferuje funkční rozdělení, kdy zmocněnec Tomáš Pojar pod sebe vezme část této agendy. V rámci otázky míry se domnívá, že stát může zasahovat poměrně výrazně skrz bezpečnostní zákony. Stát může v několika málo krizových situacích zasáhnout nejen do moderace, ale i právě zavíráním nejrůznějších komunikačních kanálů. Jako lepší řešení vnímá transparentnost, zveřejňování nějakých ex post zpráv o moderaci, jaké jsou na to vynakládány prostředky, a podle toho zpětně upravovat. Debata o regulaci k nám podle něj doběhne ani ne tak díky sporu o vládního zmocněnce pro dezinformace, ale právě kvůli Aktu o digitálních službách, protože ten si podle něj žádá nějakého národního regulátora, respektive někoho, kdo spolupracuje s Evropskou komisí, což u nás může být Český telekomunikační úřad. Vznikne podle něj obrovská nová agenda, která potřebuje

nějaké experty a nějakou veřejnou debatu o tom, jakou roli mají experti vlastně plnit a jakou transparentnost požadovat.

Raději by šel cestou úprav dosavadní legislativy, soudních precedentů, které proběhnou procesem přes nejvyšší soudy, například ústavní soud, a vytvoří obecnou předlohu pro ty nižší soudy v případech osobnostních práv a dalších. Podle něj si doba žádá například nějaké speciální senáty pro osobnostní práva, prodleva v soudním řízení a doba trvání se stává čím dál nejpřijatelnější, nevidí řešení cestou nějakých silných regulací, které třeba už za čtvrt roku nebo už v době vcházení v platnost budou neúspěšné. Shrnuje, že v radikálních a válečných stavech může mít vláda velké pravomoci, z kterých se pak následně zodpovídá, že transparentnost v běžném stavu je klíčová a úprava legislativy nezbytná.

MS následně dodává, že důležitou roli hraje hlas akademické půdy, která by měla toto dění nějakým způsobem komentovat, debatovat o něm a dávat doporučení. Domnívá se, že bez hlasu akademiků bude debata vždy směřovat k silným regulacím a drakonickým rozhodnutím. Vnímá přínos akademiků jako zásadní pro zachování vyváženosti a mezioborovou spolupráci jako nabývající na relevanci.

VB pak zmínil, že zásah by měl být vždy proporční s ohledem na velikost hrozby. Domnívá se, že ve společenské diskusi byla přijata až jakási teologie, že se jedná o existenční problém, který musí být okamžitě zastaven. Ovšem zmiňuje, že stále nemáme komplexní analýzu, která by vykreslovala, jak velkou jsou dezinformace hrozbou, jaký mají vliv, a kolik lidí reálně zasahují, jaké mají bezpečnostní dopady, a podle toho se pak mohou vytvořit adekvátní zásahy. Jako příklad udává konečnou analýzu Centra proti hybridním hrozbám, která měla sloužit jako základ pro přípravu strategického plánu. Upozorňuje na vágnost argumentů jako „česká bezpečnostní komunita se shoduje" a nepovažuje ji jako relevantní základ pro vytváření jakýchkoliv politik. Souhlasí, že stát by v případě válečného stavu nebo ohrožení státu měl mít nástroje, jak využívat defacto cenzuru k minimalizování informačního ohrožení a zabránění ztrátě životů a velkým škodám. Zároveň by měla existovat možnost archivace dokumentů a ex post vyhodnocování odpovědnosti pro případné šetření zneužívání těchto nástrojů. Je skeptický k tomu, aby se odpovědnost za šíření nepravdivých informací přenášela na platformy, které primárně slouží k vydělávání. Domnívá se, že to může vést k zadávání

úkolů, které jsou v rozporu s primárním cílem byznysu, a může to být neseriózní vůči všem stranám. Zároveň to může zhoršit vztah mezi průmyslem a státem.

Také je skeptický k tomu, do jaké míry ovlivnila technologie šíření fake news, a zmiňuje, že se jedná o periodický problém navazující na vývoj komunikačních technologií. Stejně tak zmiňuje i historii echo chambers a filtračních bublin. Věří, že hlavní nástroj pro oddělování pravdy od nepravdy už máme v diskusi. Zároveň zmiňuje, že jsou zákony, které upravují šíření poplašné zprávy a že některé informace mohou být tak nebezpečné, že nemohou být pouze předmětem debat.

DK, ačkoliv částečně souhlasí se skepticismem VB, oponuje a věří, že musíme pracovat s platformami čistě z pragmatického důvodu, protože to stát kapacitně nezvládne vzít na sebe. Zmiňuje úpadek státu jako vlivného a schopného regulátora a nevěří, že by byl schopen rychle a adekvátně reagovat. Souhlasí také s periodicitou dezinformací. Jelikož jsou tyto platformy tak nedosažitelné, je podle něj jediným přístupem od nich něco vyžadovat, protože si myslí, že regulace na úrovni státu nebo i Evropské unie je nereálná.

Následně AZ vyjádřil názor, že regulace sice může pomoci omezit dezinformace a zajistit, aby přispěvatelé nesli odpovědnost za své příspěvky, nicméně má obavy z toho, že mohou pomoci velkým hráčům vytvořit monopol, a posílit jejich pozici na trhu a diskvalifikovat tak menší platformy. Proto by se raději zaměřil na to, aby přispěvatelé byli zodpovědni za své příspěvky, a ne na regulaci samotných platforem.

Do diskuse byla doplněna otázka, jak by mělo být zacházeno s obsahem na sociálních sítích, který je zavádějící, nebo uměle šířený, možnosti označení takového obsahu pomocí nástrojů jako bot spotting nebo faktických kontrol a o různých možnostech, jako je úplné odstranění obsahu, snížení jeho organického zobrazování, nebo jeho napůl skrytí.

TH se nejprve vymezil proti cenzuře, kterou nerozpoznává jako účinnou nebo prospěšnou a ve svém důsledku může i škodit. Mimo jiné se přiklání k myšlence, že tímto mohou být více postiženy menší subjekty než ty větší. Namísto toho by mělo být informováno o tom, že nějaký příspěvek byl označen jako škodlivý, a tuto informaci by uživatelé měli vidět v reálném čase. Tento postup teď implementuje Facebook, nicméně skutečné účinky, které to má, nejsou známy.

MS souhlasí s přístupem markování příspěvků a přidává ještě možnost označování zdrojů. Jako příklad uvádí Twitter, který označuje Ruskem kontrolovaná státní média. Domnívá se, že nejlepší způsob, jak se vypořádat s dezinformacemi, je umožnit uživatelům zasahovat proti nim, a tím je degradovat v algoritmu. (Také přístup osvojený Twitterem, který mimochodem v březnu tohoto roku zveřejnil část svého zdrojového kódu a odhalil část algoritmu doporučování obsahu.)

DK oproti tomu vyjádřil skepsi k účinnosti fact-checkingu a vlivu labellingu, a to jak prostřednictvím automatizace, tak skrz autority. Zmiňuje, že mnohdy označení může konzumenta ještě naopak utvrdit ve víře v danou dezinformaci, protože konzument již ze svého světonázoru vychází antiestablishmentově. Stejný backfire efekt se pak objevoval při porušení echo chambers. Kladný vliv vidí například v labellingu afiliace (vládních úřadů a zástupců, verifikace). Nicméně limity tohoto nástroje vidí v neoznačování komunikačních a strategických pracovníků, kteří kolikrát stojí za účty veřejných osobností. Zdůrazňuje potřebu transparentnosti – co se děje na síti, jaké jsou trendy, jak probíhá moderace, kompilace s normami a zrychlení soudnictví se zaměřením na osobnostní práva, soudní opatření a úpravu zákonodárství. Zdůrazňuje absenci zákonu regulujícího audiovizuální média, kdy na jedné platformě je lhát snazší než na jiné.

VB poté doplnil ještě několik poznámek, na které je vhodné se zaměřit. První z nich je otázka důvěry v sociální sítě, zda uživatelé věří, že vidí objektivní a rozložené informace. Pokud přestanou obsahu věřit, mohou se přesouvat na jiné sítě, například Telegram. Je tedy v zájmu samotných platforem, aby zobrazovaly co nejméně nepravdivého obsahu a zachovávaly si důvěru. U označování nepravdivého nebo zavádějícího obsahu zdůrazňuje ještě jeden důležitý limit. V současné době stále nemají sociální sítě kapacitu projít a vyhodnotit všechny příspěvky, které se v prostoru jejich platformy pohybují, a proto skenují jenom část obsahu, která je pro ně dosažitelná, snadněji detekovatelná, nebo pochází ze zdroje, který je vyhodnocený jako notorický a ty následně označují jako závadné. Tento způsob pak implicitně vyvolává dojem, že cokoliv, co není označeno, je automaticky fact-checkované a pravdivé, protože kdyby to pravda nebyla, bylo by to označené. Což ovšem neodráží realitu. Stejný fenomén se objevuje u platforem, který obsah mažou, a pokud je nějaký obsah zveřejněný, získává legitimitu tím, že není smazaný.

AZ doplnil, že vidí potenciál ve veřejném odsouzení dané dezinformace, tedy zobrazit dezinformaci s jasným uvedením na pravou míru, a tím odsoudit celý narativ a dále pak v botlabellingu, tedy veřejném označení účtů, které jsou automatizované a podvržené.

MS zmiňuje kapacitní problém, kdy odsouzení dezinformace vnímá jako účinné u opinion leaderů, ale u běžných uživatelů ji vnímá jako kontraproduktivní.

TH upozornil, že nejde ani tak o to, co chceme my, ale co chce Meta nebo Google, kteří jsou schopni toto škálovat, ale neudělají to, aby je to neznevýhodnilo na trhu.

DK zmiňuje studie, které vyšly jako neprůkazné ve výzkumu vlivu a účinnosti labellingu a označování – to buď nefungovalo, nebo fungovalo jinak, než se očekávalo. Naopak se v nějakých skupinách šířily o to lépe, buď s tím narativem „podívejte se, jak se nám zase do toho cpou, ten celý establishment", anebo to posílilo jejich dopad, protože to vyvolalo reakce, které zvyšují dosah. Zmiňuje také příklad úplné suspendace, kdy třeba Googlu v době covidu suspendoval kompletně celý algoritmus a všechno okolo covidu odkazovali na WHO a ČMZ, a právě ve zkoumání možnosti úpravy a zdokonalování algoritmů vidí jistý potenciál. Také zdůrazňuje, že je nutné vytvořit nějaký defaultní rámec podle závažnosti situace a odlišit podle něj velikost reakce a zásahu (je rozdíl v situacích, kdy se jedná o národní ohrožení v nouzovém stavu a zkreslené debatě o popularitě EU).

d) Role člověka v hybridním přístupu, anotace, rozdělení pravomocí, míra zapojení

Na tuto otázku jako první reagoval MS, který zmínil právě psychický nápor, kterému musí lidští moderátoři čelit, jako jeden z hlavních limitů zapojení lidí do procesu anotace a labellingu. To demonstroval na příkladu z Filipín, kde lidští moderátoři vykazovali větší procento sebevražd právě z důvodu vystavení se závadnému obsahu. Shledává v tomto nahrazení lidského faktoru umělou inteligencí jako řešení této situace, nicméně jedním dechem dodává, že stěžejní je vytvořit zároveň kód, který skutečně není biased a zamýšlí se, zda jsme vůbec jako lidé schopni takový 100% objektivní kód zajistit.

Do debaty následně přispívá TH, který v úvodu vysvětluje, že veškeré modely fungují stochasticky, a ačkoliv lze zvyšovat úspěšnost toho kódu a opravovat nalezené limity,

stejně nedosáhneme 100% objektivity právě z důvodu stochastických procesů. Zmiňuje, že v současné době může představovat cestu i využití matematických modelů, nicméně trh je zatím opomíjí. Vysvětluje, že odborná veřejnost se shoduje, že AI musí obecně fungovat na principu logického vysvětlování raději než na frekvencionickém přístupu (čím víc dat = tím chytřejší). Také zmiňuje problematiku lidské anotace, která může být psychicky velmi náročná, dodává případ outsourcingu anotací dat získaných z Dark webu (používaných k zamezení jailbreaku AI systému) do zemí třetího světa, kde pak chybí jakákoliv personální podpora pro lidi postižené závadným contentem, což následně vede k sebevraždám a depresím. Vytvoření férového algoritmu tedy může mít i tuto temnou stránku.

DK oproti tomu zdůrazňuje důležitost vnímání rámce, ve kterém se pohybujeme, kde funguje občanská demokratická společnost oproti zemím, kde jsou prováděny tyto anotace a neregulované moderace. Zmiňuje rozdíl mezi propagandou například východní Evropy a bývalých francouzských kolonií v Africe. Zde oproti utahování šroubů naopak využívají chaos a mnoho narativů. Jsou tedy různé druhy informačních prostředí, ve kterých fungují různé modely, což může ovlivnit i míru potřebné moderace i její automatizace. Roli člověka v AI procesu nevidí v zaručení objektivity nebo vyváženosti, ale v garanci subjektivity. Chybující člověk, co dělá krátkodobé chyby, z dlouhodobého hlediska garantuje lidské posouzení. V reportech od platforem by čekal transparentní prezentaci chybných rozhodnutí, a jak ovlivnila jejich další vývoj v následném období až do další evaluace a reportu, která by měla probíhat v krátkých (čtvrtletních až ročních) časových úsecích.

AZ se také přiklání k potřebě změny přístupu a rychlosti. Vnímá potřebu vydávat rychlejší a krátkodobější rozhodnutí namísto pomalých dlouhodobých, osekat procesy, častěji vyhodnocovat a přezkoumávat a nebát se dělat chyby. I kdybychom zvládli po dlouhých úvahách vynést jedno dokonalé rozhodnutí za tu dobu, co nám trvalo k němu dojít, už nemusí dávat smysl. Rozhodování nemůže být pomalejší než vývoj, musí se mu přizpůsobit. V procesech musíme připustit to, že děláme chyby, které ale budou rychle napraveny díky časté evaluaci a konstantní změně.

VB nejprve v rámci mimo technické poznámky zmiňuje literaturu zabývající se strojovým učením a rozhodováním o autonomních zbraních, která řeší primárně odpovědnost za tato rozhodnutí, na koho padá odpovědnost za chybu vedoucí k úmrtí

civilistů. Následně zmiňuje diverzitu obsahu, který řešíme, od nenávistných projevů přes obsah porušující zákony (dětská pornografie), mírné dezinformace a ohnuté narativy až po naprosto mylná tvrzení neuchycená v realitě. Právě tato diverzita i určuje danou utilitu AI nástrojů, která u jednoho druhu obsahu může být vhodnější než u jiného. U některého závadného obsahu může být automatizace téměř okamžitá a absolutní. Zároveň u obsahu jako je dětská pornografie, pokud vezmeme v úvahu škodlivost pro společnost a vliv tohoto obsahu na lidské moderátory, tak je zásah ospravedlnitelný a případné chyby vyváží benefity.

V informační válce je ovšem většina obsahu na druhé straně spektra a je skeptický k možnostem nástrojů fact-checkingu. Zmiňuje, že i u nástrojů, které používá a považuje je za kvalitní, vnímá dvojí standard ve vyhodnocování, vykazují odlišný přístup při ověřování pravdy, často nastavují přísnější kritéria, než by aplikovaly na sebe.

e) Bias, účinnost fact-checkingu, gramotnost a poptávka trhu

TH zmiňuje, že jeden ze systému, na kterých pracuje, se právě soustředí na detekci bias primárně pro média. V rámci komunikace s potencionálními klienty (média, mediální domy, agentury) zjistili, že ačkoliv jim to všem přijde zajímavé, v současné době o takový systém nemají akutní zájem, protože nevědí, jak ho zpracovat a využít.

MS se přidává se zdůrazněním nutnosti ověření poptávky, jedna věc je společenská poptávka, ale druhá pak reálná poptávka trhu. Jako příklad uvádí znovuobnovení účtů Donalda Trumpa a Andrewa Tatea na Twitteru, který mohl být způsobený jednak poptávkou, ale hlavně snahou rebrandovat Twitter, který byl vnímán primárně jako levicová platforma, a snahou pochytat i uživatele z jiných spekter.

DK se přiklání k tezi, že se obecně trhově snažíme o pochytání co nejvíce uživatelů napříč názorovým spektrem, a to upravuje pak poptávku. Zmiňuje potřebu zaměření se i na tematické bias, reprezentace témat v mediálním prostoru, například jestli je v pořádku, že z 20 témat je 18 o Babišovi. V kombinaci s pozorností defaultních občanů / spotřebitelů je důležité témata vyvážit, aby nebylo něco opomíjeno. Řešení opět vidí v transparentnosti nebo samoregulačních kodexech. Nevidí takový problém v biasu

výsledku, ale v biasu výběru, nedostatku tematické pestrosti nebo vyloučení určitých témat, protože jsou spojována s dezinformacemi nebo osobami.

VB dodává, že obecně nevidí ve fact-checkingu takový potenciál, a vnímá ho přeceňovaně. Nevidí řešení problému, na kterém se nemůžeme shodnout, v sestavení skupiny, která to projde a řekne, jak to je. Bias je nevyhnutelné a je skeptický k tomu, že by to šlo řešit. Jsme v situaci, kdy vyvracíme buď nesmysly, které jsou vyvratitelné snadno, a je zbytečné o nich debatovat, a pak jsou témata, která jsou příliš komplexní a absolutní jistota u nich neexistuje, a tam žádný kodex nepotřebujeme. Zde by pak měla fungovat investigativní novinařina. Je potřeba nepřeceňovat fact-cheching, který musí být primárně důvěryhodný pro své obecenstvo, ne univerzálně. Také zmiňuje problematiku nekvalitních a biased anotací, které vytvářejí například studenti, kteří nemají potřebnou expertízu. Ačkoliv je pro hledání bias ve fact-checkingu, nevěří, že bude existovat objektivní nástroj, který by vyřešil naše problémy.

DK doplňuje, že fact-checking perfektně odpovídá na otázky, na které jsme se neptali. Zajímá nás spíše archeologie té kontextualizace, narativy a jejich seřazení, laterální čtení. Zároveň ale přiznává fact-checkingu jisté zásluhy, pouze ho nevnímá jako řešení, které nám může pomoci právě v pátraní po kontextualizacích a narativech, neodpoví nám to na otázky, na které se nejvíce ptáme.

TH dodává, že i kdyby byl fact-checking perfektní a neomylný, kdyby data neanotovali studenti s levicovým zaměřením, stále může docházet k reakcím publika, které jsme nepredikovali a může to mít opačný efekt. Spíš vidí důležitost v edukaci veřejnosti v tom, jak by měla konzumovat informace, implementovat mediální vzdělávání do školní výuky – naučit se je rozeznávat a vyhodnocovat sami, místo toho, že jim udělíme finální výsledek skrz fact – checking.

VB dodává, že ačkoliv se o potřebě mediálního vzdělávání mluví už 8 let, stále se nic neděje, protože potřeba je vnímaná jako příliš akutní a je snaha hledat rychlá a jednoduchá řešení místo komplexních reforem. Zároveň by měl být zaveden vhodný aparát na nápravu a ověřování informací například z řetězových mailů. To by měl zajistit stát, který by měl poskytnout možnost ověřovat média, která by měla komentovat, a pak různé ověřené encyklopedické stránky. Zároveň zmiňuje, že pozornost fact-checkingu věnují

lidé, kteří už nevěří dezinformacím, často se k nim řetězové maily nedostanou a sledují tyto stránky, aby byli v obraze, a případně je přeposílají příbuzným. U modelu fact-checkingu je problém jak to dostat k lidem, kteří to potřebují a zároveň vychází z premisy, že lidé konzumují informace proto, aby byli informováni, což neodráží realitu. Zmiňuje confirmation bias, způsob nakládání s informacemi, to, že lidé nechtějí podstupovat diskomfort změny svého názoru, pokud mají pocit, že na důsledku tolik nesejde. Pro každodenní životy většiny lidí nemají dezinformace takový dopad, aby jim lidé věnovali zvýšenou pozornost, nebo ověřovali, zda jsou pravdivé. Jejich motivace pro ověřování je malá. Kloní se na stranu skepse k iluzi absolutního a perfektního konzumenta. Nesouhlasí například s Ludwigem a jeho konceptem, že i dělník v továrně si může přečíst metastudii přeloženou překladačem. Koncept toho, že každý si můžeme ověřit fakta a najít informace, je nereálný. Člověk nemá mentální kapacitu zvažovat všechny možnosti, porovnávat pro a proti a věnovat každé informaci absolutní pozornost. Neztotožňuje se s ideou, že uživatelé chtějí pravdu a jsou ochotni investovat do jejího získání. Motivace znát pravdu a vnímaná důležitost u nich nevyrovná náklady v podobě investování času a mentální kapacity. Vidí spíše potřebu přesvědčit lidi, že pro jejich život a jejich blízké je důležité znát tu správnou odpověď, ne tu pohodlnou.

Individuální rozhovor s Mgr. et Mgr. Tomášem Kučerou, Ph.D.

V rámci individuálního rozhovoru byla probírána podobná témata jako ve focus group, hlavní závěry rozhovoru, které nebyly zmíněny ve focus group, jsou shrnuty v následující sekci. Celý transkript je k dispozici v příloze této práce.

V otázce odpovědnosti za obsah TK zdůraznil, že část regulace a odpovědnosti je už zaštítěna v zákonu o šíření poplašných zpráv. Zamýšlí se, zda postihovat a vymáhat tyto regulace na poskytovatelích nebo uživatelích. Nepovažuje za nutné vytvářet nové úřady, které by suplovaly roli poskytovatelů. Místo toho by měl stát investovat do vzdělání a kapacity policie a státního zastupitelství.

V rámci individuálního rozhovoru byla přednesena otázka techplomacie a její budoucnosti. Podle respondenta stát musí plnit svoji roli, ale není nezbytně nejefektivnější v tom, jak říct, jak se má problém řešit. Proto je důležité vytvářet mosty

mezi státními orgány a dalšími aktéry, aby se dosáhlo kýženého cíle. Nevidí problém v tom, aby jedním z těchto mostů byla právě techplomacie.

V otázkách důvěry ve fact-checking respondent zmiňuje, že společnost potřebuje instituce, které nějakým způsobem poskytují ověřené informace, autority, kterým lze věřit, a také odkazuje na nutnost zachování profesionálních standardů v novinářské branži a zvýšení důvěry veřejnosti v média. Rozeznává důležitost a vliv fact-checkingu, protože uživatelé nemají čas a motivaci na ruční ověřování každé informace.

V otázkách účinnosti fact-checkingu se respondent domnívá, že jsme se jako demokratická společnost unitárně shodli, že abychom recenzovali, a přitom neohrožovali svobodu slova, musíme věřit v účinky a smysl fact-checkingu.

V otázkách nakládání s daty je mu sice myšlenka databáze digitálního otisku sympatická, nicméně se bojí, že by to přineslo odpovědnost z platforem a institucí na koncového konzumenta. Preferoval by, aby veřejné instituce byly zodpovědné za dohlížení na zájmy občanů a uživatelů a zajistily tak ochranu proti zájmům korporace. Budování resistence vůči microtargetingu vidí ve vzdělávání veřejnosti a vychovávání větší odolnosti vůči technologickým změnám.

V závěru se také zamýšlí nad skutečným vlivem dezinformací na uživatele, zda tato hrozba není vnímána jako větší, než skutečně je.

5. Výsledky a diskuse

(1) V otázce financování se respondenti víceméně shodli, že proces by měl být rovnoměrně rozložený mezi soukromý sektor, který se zaměřuje na nástroje, a stát, který financuje koncové nástroje pro detekci a monitoring pro státní účely. Státní financování je také kruciální pro akademický výzkum, do kterého by soukromý sektor tolik zasahovat neměl. Financování a zapojení soukromého sektoru proti tomu může přinést benefity ve formě rychlejšího pokroku zavedením konkurence a lepším ohodnocením expertů. V debatě také zaznělo riziko financování a vývoje nástrojů soukromým sektorem využívaných ve veřejné sféře, kde pak chybí transparentnost a možnost kontroly fungování daného nástroje, který ovšem zároveň získává relevanci a důvěryhodnost právě spoluprací se státní nebo veřejnou sférou. Výzkum jednotlivých nástrojů v soukromé sféře zároveň může sloužit k zachování konkurenceschopnosti na mezinárodním trhu.
Ideální model tedy je kombinované financování a zapojování se do procesu vývoje oběma sférami, ale musí být zachována podmínka transparentnosti a přezkoumávání, ideálně formou open-source řešení.

(2) V otázkách přístupu k microtargetingu, hypercílení a využívání dat se respondenti obecně shodli, že při regulaci digitálních médií je důležitá transparentnost, nikoli jejich přímý zákaz nebo omezení. Transparentnost by měla být například v metodice moderace, nebo zpracování dopadových studií. Roli v procesu má veřejná debata, stanovení limitů a vytvoření tlaku pro transparentnost. Také je zmíněna hrozba monopolu jedné platformy, která bude mít velké množství dat, a přes kterou bude uživatel přistupovat téměř k veškerému obsahu. Pokud bude pak takto cílen, mělo by to mít nějaké limity formulované ve veřejné debatě. Jako jednu ze zvažovaných možností je zmíněna kontrola uživatele nad digitálním otiskem, ke které se ovšem ostatní respondenti staví skepticky, protože nevěří v omnipotentního uživatele, který by měl zájem s ní pracovat. Respondenti také projevují skepticismus k účinnosti těchto nástrojů a vlivu, který mají na rozhodnutí uživatelů. Větší riziko může představovat cílení na vybrané opinion leadery v daném tématu. Vliv také tato problematika má na případnou polarizaci a radikalizaci uživatelů, kterým pomocí těchto nástrojů upraví informační dietu, upevní je v jejich názorech a ukotví je do filtračních bublin.

(3) V otázce regulace online platforem a odpovědnosti za content se účastníci shodli, že nalezení správné rovnováhy mezi svobodou projevu a regulací je zásadní. Zdůraznili rovněž význam transparentnosti. Debata se zabývala rolí vlády při regulaci online platforem, významu akademických příspěvků do debaty a potřeby přiměřenosti jakýchkoli přijatých opatření. Respondenti zmínili možná rizika nadměrné regulace a zdůraznili, že je důležité přizpůsobovat stávající zákony a soudní precedenty novým výzvám, a nikoliv vytvářet nové, silné regulace, které by nemusely být dlouhodobě účinné. Celkově se účastníci shodli na tom, že jakékoli úsilí o regulaci online platforem by mělo být vedeno skrz transparentnost, evaluace prováděné v krátkých časových úsecích a zrychlením legislativního procesu. S tím by mohla pomoci například techplomacie. Respondenti se také zamýšlí nad tím, jak velkou hrozbu skutečně dezinformace představují pro společnost, zmiňují, že chybí komplexní analýza dopadů. Zároveň jistá forma samoregulace je i v zájmu platforem, které se tak brání odlivu uživatelů ke konkurenci z důvodu ztráty důvěry nebo zájmu o nerelevantní a nesmyslný obsah.

(4) V diskusi o úloze lidí v procesu moderace a vyhodnocování obsahu je nejprve zmíněna psychická zátěž, které čelí lidští moderátoři v důsledku vystavení škodlivému obsahu, a možnost outsourcingu těchto anotací, jejímž řešením by bylo právě převzetí úkonu umělou inteligencí. Následně se debata věnuje nestrannosti kódu, které podle jednoho z respondentů nelze absolutně dosáhnout, řešení není v dodání většího množství dat. U extrémně závadného obsahu jsou případné strojové chyby odpustitelné a benefity pramenící z toho, že obsah nemusí anotovat člověk, přehluší případná rizika chybného označení.

Jeden z respondentů vnímá úlohou člověka v AI v poskytování subjektivní záruky, nikoliv v zaručení objektivity či vyváženosti. Respondenti se shodli na potřebě transparentnosti při hlášení chyb a jejich dopadu na další vývoj a pravidelném hodnocení a podávání zpráv. Vnímají potřebu změny přístupu a rychlosti k decitionmakingu, preferují rychlejší a krátkodobější opatření, zkracování procesů a častější vyhodnocování. Obecně vidí řešení ve větší rychlosti a flexibilitě, a to jak na úrovni vyhodnocování a reportování fungování, tak v otázkách legislativy a regulací.

(5) Účastníci debaty se shodli, že fact-checking není dokonalým řešením a že by měl být důvěryhodný pro své obecenstvo, a nikoliv univerzální. Také se hovořilo o potřebě transparentnosti a samoregulačních kodexech. Účastníci debaty souhlasili s tím, že fact-checking může být užitečný nástroj, ale nemůže odpovědět na všechny otázky a může mít opačné důsledky na konzumenty, než bylo původně zamýšleno. V debatě se respondenti shodli, že model prezentování finálního soudu namísto učení uživatelů, jak vyhodnotit informace sami, není ideální, zároveň ale byla zdůrazněna apatie uživatele jako jeden z hlavních limitů fact-checkingu – uživatel zkrátka nevnímá potřebu ověřovat informace a mít pravdivé informace natolik důležité, aby investoval čas a byl motivovaný vynaložit úsilí a pozornost na ověření. Právě zvyšování mediální a informační gramotnosti vidí jako jednu z cest, zároveň ale rozeznávají, že se jedná o dlouhodobý a komplexní proces.

Jedním z hlavních přínosů debaty bylo zamyšlení, jak moc jsou fact-checkingové a leballingové zásahy efektivní. Se snižujícím se attention spanem, následováním internetových autorit, důvěrou v preferované narativy, zvyšující se apatií uživatele a backfire efekty vyvstávají pochybnosti o reálném zásahu a účinku těchto nástrojů a označování. Stejně tak ale panuje skepse ohledně účinků a efektivnosti microtargetingu a hypercílení a dopadu dezinformací.

Důležitým podmětem vzešlým z debaty bylo i zmínění možného vedlejšího produktu fact-checkingu – ten může v praxi propůjčovat legitimitu i neověřeným informacím (které mohou být nepravdivé), protože budí dojem, že cokoliv, co není označené, je tedy pravdivé, protože kdyby to nebylo pravdivé, už by to mělo štítek, nebo bylo smazáno.

Financování	⇒ vyvážená míra zapojení soukromého a veřejného sektoru ⇒ stát by měl financovat akademický výzkum ⇒ zapojení soukromého sektoru do financování pro rychlejší pokrok zavedením konkurence, lepší ohodnocení expertů a zachování konkurenceschopnosti na mezinárodním trhu ⇒ důležitost transparence a open-source
Microtargeting a využívání dat	⇒ důležitost transparence místo regulací ⇒ role veřejné debaty a tlaku na transparenci ⇒ transparence metodiky moderování a dopadové studie, časté reportingy ⇒ digitální otisk jako nástroj pro 1 %, mimo zájem širší veřejnosti, hrozba přenesení odpovědnosti na koncového uživatele ⇒ skepse k účinnosti těchto nástrojů na uživatele a jejich rozhodování a preference ⇒ riziko cílení na opinion leadery a ovlivňování agenda settingu ⇒ polarizace a radikalizace, tvoření filtračních bublin
Regulace a odpovědnost za obsah	⇒ přiměřenost opatření úměrně k hrozbě ⇒ potřeba zrychlit soudní systém a vydávání krátkodobých opatření ⇒ techplomacie ⇒ transparentnost platforem a časté evaluace ⇒ důvěryhodnost jako zájem platforem – měly by filtrovat obsah ve vlastním zájmu, aby si zachovaly uživatele a obsah byl smysluplný ⇒ otázka skutečné hrozby a dopadu dezinformací
Lidé v procesu	⇒ problémovost zapojení lidských moderátorů a dopady obsahu na jejich duševní zdraví ⇒ výhoda přenesení škodlivého obsahu na AI ⇒ u extrémního obsahu případná strojová pochybení přináší větší benefity než rizika ⇒ potřeba transparence při hlášení a nápravě chyb AI
Nástroje fact-checkingu	⇒ skepse ke skutečnému vlivu fact-checkingu ⇒ důvěryhodné pro své obecenstvo, nikoliv univerzálně ⇒ důležitost existence důvěryhodné instituce / autority ⇒ limit v bias moderátorů a nekvalifikované anotaci ⇒ backfire efekty, heuristiky snižující dopad ⇒ označování zavádějícího obsahu propůjčuje zbylému neoznačenému obsahu relevanci a důvěryhodnost ⇒ smysl v labellingu afiliace ⇒ apatie uživatele jako limit fact-checkingu ⇒ potřeba zvyšování mediální a digitální gramotnosti

Tabulka shrnující výsledky focus grup

6. Limity

Jedním z limitů výzkumu byla absence právního odborníka na umělou inteligenci. Ačkoliv jsem ve fázi shromáždění oslovila přes 20 respondentů, bohužel se mi nepodařilo zajistit účast právníka. Velká část debaty se zabývala možnými regulacemi a legislativou, účast právníka by tak dodala debatě praktický přesah a obohatila ji o pohled z praxe. Diskutovaná potřeba pro změnu procesu a rychlost zavádění regulací by mohla být konfrontovaná s odborníkem, pracujícím v tomto procesu, a ten by mohl diskutovat konkrétní postupy, překážky a hrozby zavedení tohoto přístupu.

Stejný limit představuje i účast odborníka, zastupujícího konkrétní soukromou společnost, která vytváří diskutované nástroje pro komerční účely. Opět v oslovovací fázi bylo poptáno několik respondentů, například ze Semantic Vision, kteří na prosbu neodpověděli. Případné zapojení těchto pracovníků by obohatilo otázku financování a transparentnosti, předložilo názory a překážky těchto společností k osvojení open-source a obohatilo diskusi o vhled do povahy poptávky.

Limitem je i vlastní škála výzkumu. V teoretické práci byla prezentovaná komplexní problematika umělé inteligence a jejího praktického využití pro boj s dezinformacemi, konkrétní příklady i subjekty, působící v kontextu AI vývoje v České republice a predikce a příležitosti do budoucna. Nicméně praktický výzkum pro účely této práce nemá kapacitu probrat všechna témata zmíněná v práci. V budoucím výzkumu by tedy bylo vhodné zhotovit sérii focus group, která by měla několik etap. Podněty vzešlé z první debaty by byly diskutované v etapě následující, odborníci by byli více rozprostřeni mezi disciplínami a sférami působení. Tento druh komplexnějšího a obsáhlejšího výzkumu by pak mohl položit relevantní rámec, na kterém se shodlo dostatečné množství odborníků a který byl oproti původnímu výzkumu obohacen dalšími podněty, ke kterým by se mohly vypracovat další case study, sloužící k prohloubení znalostí.

Zároveň bych se ráda ve výzkumu věnovala více konkrétním řešením a příkladům z praxe, které byly prezentovány v teoretické části, bohužel ve škále tohoto výzkumu jim nebyla věnována priorita. Diskuse těchto nástrojů by vedla k vyhodnocení vhodných procesů a formulací limitů a efektů zpětného rázu, které tyto nástroje a procesy mohou zaznamenat.

Práce je limitována i samotnou obtížností konceptualizace pojmů dezinformace, fake news, obrovskou složitostí tématu na pomezí digitální komunikace, vládní komunikace, technologického vývoje, politologie, mediálních studí, práv a dalších. Existuje mnoho možností přístupu k takto komplexnímu tématu a v prostoru této práce bylo možné zasáhnout pouze okrajovou část této dynamické problematiky.

Závěr

Teoretická část této práce nejprve poskytla shrnutí a kategorizaci myšlenkových a behaviorálních procesů konzumentů, které vstupují do kontextu procesu přijímání a vyhodnocování informací. Tato kategorizace nám posloužila k lepšímu pochopení motivací a vedlejších procesů, které mohou ideální přímou linku vyvracení - identifikace - dodání informací - změna postoje - odklonit, narušit nebo sabotovat. Znalost těchto procesů je kruciální pro vývoj efektivních zásahů, které nebudou mít efekt zpětného rázu. Právě možným efektům zpětného rázu se věnuje závěr této kapitoly. Otázka vlivu představených konceptů vzešla i ve výzkumu focus group, kde se respondenti opírali o confirmation bias, worldview backfire effect a další při diskusi účinnosti a efektivity fact-checkingu. Diskuse nabídla náhled na možné limity tohoto nástroje a skepsi k jeho účinnosti do budoucna ať už ve formě labellingu, nebo i odstraňování. Nebezpečí nástroje leží i v tom, že propůjčuje legitimitu i neověřenému obsahu, protože ho neoznačí.

V sekci věnující se výkladu umělé inteligence byly představeny některé nástroje informační války jako deepfake, microtargeting a boj proti nim jako hybridní přístup, fact-checking, labelling botů nebo dezinformací a další. Prezentovány byly i hrozby (například stoupající apatie uživatelů) nebo příležitosti (například techplomacie). V debatě byly tyto nástroje diskutovány a jedním z hlavních podmětů byla otázka reálné účinnosti a tím i hrozby nástrojů jako microtargeting nebo zneužití big data. Vzhledem ke komerčnímu zájmu platforem monetizujícím microtargeting a data uživatelů, lze předpokládat, že budou mít zájem na prezentování co možná největší účinnosti těchto nástrojů, která ovšem nemusí odrážet realitu. Respondenti se domnívají, že oproti domnělému účinku je dopad těchto nástrojů v praxi minimální, cílí spíše na lidi, kteří už byli rozhodnuti pro kandidáta, nebo by daný produkt koupili tak jako tak.

V otázce hybridního přístupu a role člověka bylo shledáno, že jedním z limitů je skutečně psychická zátěž na moderátora, která byla zmíněna i v teoretické části. U extrémně závadného obsahu, jako například dětské pornografie, by podle respondentů byla vhodná co možná největší automatizace, protože vliv na společnost je tak škodlivý, že případné chyby (například zmiňované vyhodnocení fotografie Dívky s napalmem jako dětské pornografie) jsou obhajitelné. Problém by nastal ve chvíli, kdyby tento proces

automaticky diskvalifikoval zdroj (uživatele) bez možnosti odvolání / přezkoumání člověkem. Role člověka v procesu má také sloužit k zachování subjektivního posouzení.

V debatě také byla formulována potřeba vzdělávání veřejnosti a výuka k ověřování informací. Fact-checking nepovažují respondenti za ideální, nicméně vzhledem k nízké motivaci aktérů k ověřování informací nelze předpokládat, že tento proces nahradí jejich iniciativa. Jedním z řešení by mohl být právě interaktivní vzdělávací model představený v jedné z případových studií, ten díky interaktivitě udrží pozornost respondenta a zároveň ho nenuceným způsobem učí procesy ověřování informací a upozorňuje na kritéria, která by měl používat k vyhodnocování.

V otázkách financí byla potvrzena jedna z premis teoretické části, tedy že podfinancovaný výzkum vede k nekvalitním anotacím, prováděným například studenty, kteří nejsou dostatečně školení a nedisponují potřebnou expertízou. To vede například k nevyváženému fact-checkingu. Debata také potvrdila silnou roli akademie v procesu výzkumu, debat, regulace a etiky.

V otázkách regulace byla v mnoha ohledech vyhodnocena jako klíčová právě transparentnost. Platformy by měly neustále vyhodnocovat a zpětně kontrolovat své procesy a zásahy, ty zveřejňovat a podle nalezených chyb nástroje opravovat. Kvalitní a flexibilní regulace ze strany státu podle respondentů není kapacitně reálná, a proto by mělo docházet ke kontrole právě skrze transparentní reporty na například čtvrtletní bázi. Případná silná regulace by jednak nebyla upravitelná v reálném čase, ale také by zhoršila vztah s poskytovateli, zároveň by mohla vyřadit menší hráče a posílit monopol.

Jednou z možností by bylo právě osvojení techplomacie, pokud by vláda vytvořila nový úřad technologického velvyslanectví, který by fungoval jako moderátor a tlumočník mezi platformami a vládou. Jeho role by mohla být právě v dozorování na zachování transparentnosti, monitorování reportingů a dohlížení na nápravu chyb vzešlých z reportů. Zároveň by tato forma spolupráce vylepšila vztahy mezi aktéry, velvyslanec by měl kvalitnější technologické vzdělání a byl by schopen se lépe doptávat a chápat reporting a fungování daných platforem. Nevznikala by pak situace jako například v americkém kongresu, který zpovídal CEO TikToku, a který předkládal příliš polopatické otázky, jež

sice vycházely z legitimních obav, místy působily příliš zjednodušeně, banálně až neznale. (Reuters, 2023)

Oproti tomu možnost zavedení přístupu uživatele ke svému digitálnímu otisku jedním ze způsobů zmíněných v teoretické části, ačkoliv zní sympaticky, by podle respondentů měla jen minimální vliv. Přeceňuje zájem a znalosti běžného uživatele, který by s digitálním otiskem nepracoval. Předpokládá omnipotentního uživatele, který má zájem a motivaci, nicméně tato představa není zakotvená v realitě a v důsledku by postihla jen spodní procenta uživatelů. Zároveň zde vzniká obava, že zavedením tohoto nástroje přechází odpovědnost ze státu a platforem na koncového uživatele, který ale nemá motivaci s ním pracovat, a tím není již chráněn tolik, jako by byl v případě přenechání odpovědnosti státu případně orgánům Evropské unie. Diskuse také potvrdila legitimitu hrozby plynoucí z rostoucí apatie uživatelů.

Tato práce potvrdila potřebu multidisciplinární debaty, která i v minimální formě (focus group této práce) přinesla mnoho pohledů a argumentů. Technologické znalosti AI pracovníků v kombinaci s chápáním konzumentova chování výzkumníků sociálních věd, vyhodnocování reálné míry hrozby odborníků na bezpečnost a znalost povahy informací mediálních pracovníků přinesla podmětnou debatu, která upozornila na různé limity určitých nástrojů a přístupů, které se mohou jedné skupině zdát relevantní a účinné, ale z náhledu skupiny jiné je jejich vliv marginální. Navzájem se upozorňují na limity a možné efekty zpětného rázu, připomínají si motivace jednotlivých aktérů a jejich cíle.

Univerzálním závěrem této práce je potřeba činnosti a dialogu. Stát by se neměl bát dělat rychlá a krátkodobá rozhodnutí, měl by přestat hledat jednotné řešení, které dlouhodobě vyřeší celou problematiku. V konstantní činnosti sice bude dělat chyby, ale díky flexibilitě, transparentnosti, dialogu s odborníky a neustálému monitorování bude schopen tyto chyby napravovat a posouvat se kupředu. Zároveň není radno podceňovat vliv fact-checkingu stejně jako motivaci uživatelů znát ověřené informace.

Zdroje

1. ABEYWICKRAMA, D. B., BICOCCHI, N., MAMEI, M., & ZAMBONELLI, F. (2020). The SOTA approach to engineering collective adaptive systems. International Journal on Software Tools for Technology Transfer, 22, 399-415.
2. BAEZA-YATES, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54-61.
3. BAYER, J., BITIUKOVA, N., BARD, P., SZAKÁCS, J., ALEMANNO, A., & USZKIEWICZ, E. (2019). Disinformation and propaganda–impact on the functioning of the rule of law in the EU and its Member States. European Parliament, LIBE Committee, Policy Department for Citizens' Rights and Constitutional Affairs.
4. BERZINA, K., HAMILTON, D., & JENTGENS, P. (2019). The ASD European policy blueprint for countering authoritarian interference in democracies. The German Marshall Fund of the United States.
5. BICKERSTAFF, J., (2023) AI language model ChatGPT escapes from Danforth Labs. The Washington Post [online]. 14 Feb 2023. Available at: https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak/
6. BIKHCHANDANI, S., HIRSHLEIFER, D., & WELCH, I. (1998). Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades. Journal of Economic Perspectives, 12(3), 151-170.
7. BISHOP, S. (2021). Influencer management tools: Algorithmic cultures, brand safety, and bias. Social media+ society, 7(1), 20563051211003066.
8. BLANK, S. (2008). Web War I: Is Europe's First Information War a New Kind of War? Comparative Strategy, 27(3), 227-247. DOI: 10.1080/01495930802185312
9. BOHÁČ, J. (2021, 8. 10). Firmy za dezinformace na internetu platí tisíce eur. Aktuálně.cz [Online]. Dostupné z:https://zpravy.aktualne.cz/domaci/dezinformace-a-firmy/r~602fccded2da11ed8980ac1f6b220ee8/.
10. BOHDÁLKOVÁ, A. (2022, 8. 2.). CZ.NIC zablokoval osm domén dezinformačních webů. Root.cz.https://www.root.cz/zpravicky/cz-nic-zablokoval-osm-domen-dezinformacnich-webu/anketa-2008/
11. BORNSTEIN, R. F., & CRAVER-LEMLEY, C. (2016). Mere exposure effect. In Cognitive illusions (pp. 266-285). Psychology Press.
12. BUGAJSKI, J. (2004). Cold peace: Russia's new imperialism. Greenwood Publishing Group.
13. BURBACH, L., HALBACH, P., ZIEFLE, M., & CALERO VALDEZ, A. (2019). Bubble trouble: strategies against filter bubbles in online social networks. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Healthcare Applications: 10th International Conference, DHM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21*(pp. 441-456). Springer International Publishing.
14. CARAMANCION, K. M. (2020). An Exploration of Disinformation as a Cybersecurity Threat. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 440-444). IEEE. doi: 10.1109/ICICT50521.2020.00076.

15. CLAYTON, M. (2014). Ukraine election narrowly avoided ‘wanton destruction’ from hackers. Christian Science Monitor. Retrieved from https://www.csmonitor.com/World/Passcode/2014/0617/Ukraine-election-narrowly-avoided-wanton-destruction-from-hackers.
16. COLMAN, A. M. (2009). A Dictionary of Psychology (3rd ed.). Oxford: Oxford University Press.
17. CONLEY, H. A., et al. (2016). The Kremlin Playbook: Understanding Russian Influence in Central and Eastern Europe. Rowman & Littlefield.
18. COPELAND, M. (2016). The difference between AI, machine learning, and deep learning? The Official NVIDIA Blog, July 29, 2016, https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/.
19. CORDIS, (2021). Research project details for: Artificial intelligence for marine robotics (AI4Ocean) [Grant agreement no: 825469]. Cordis Europa. https://cordis.europa.eu/project/id/825469
20. CRAWFORD, K. (2016, 25. 6.) Artificial intelligence’s white guy problem. The New York Times.https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html
21. ČESKÁ CENTRÁLA PRO VĚDU A TECHNICKÝ ROZVOJ. (2022, September 20). Dezinformace [Misinformation]. Česko v datech [Czechia in Data]. Retrieved from https://www.ceskovdatech.cz/clanek/176-dezinformace/.
22. ČESKO V DATECH (2022, 20. 9.). Dezinformace. České vědecké a technické centrum. Retrieved fromhttps://www.ceskovdatech.cz/clanek/176-dezinformace/.
23. ČIŽIK, T., & MASARIKOVÁ, M. (2018). Cultural Identity as Tool of Russian Information Warfare: Examples from Slovakia. Science & Military Journal, 2018(1), 2018.
24. ČTK. (2008). Juščenko vím, kdo mě otrávil. Lidovky.cz. Retrieved from https://www.lidovky.cz/svet/juscenko-vim-kdo-me-otravil.A080728_211125_ln_zahranici_mel
25. DAILEY, S. L., BROWNING, L., Retelling stories in organizations: Understanding the functions of narrative repetition. Academy of Management Review, 2014, 39.1: 22-43.
26. DAVID, J., (2022) SWOT analysis of the USA’s Information Technology (IT) industry. Retrieved from: https://www.howandwhat.net/swot-analysis-usas-information-technology-it-industry/
27. DEMARTINI, G., MIZZARO, S., & SPINA, D. (2020). Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. IEEE Data Eng. Bull., 43(3), 65-74.
28. DIFONZO, N., & BORDIA, P. (2007). Rumor psychology: Social and organizational approaches. American Psychological Association.
29. DOHNALOVÁ, A. (2023, 18. 3.). Klíma končí jako zmocněnec pro oblast médií a dezinformací. U vlády ztratil důvěru. aktualne.cz.https://zpravy.aktualne.cz/domaci/michal-klima-konci-ve-funkci-zmocnence-pro-oblast-medii-a-de/r~29719f02ad2711eda25a0cc47ab5f122/
30. DOHNALOVÁ, A., & DOUBRAVCOVÁ, B. (2023, 10. 4.) Dezinformace ohrožují byznys firem, ty si ale často u dezinformátorů platí reklamu. https://zpravy.aktualne.cz/domaci/dezinformace-a-firmy/r~602fccded2da11ed8980ac1f6b220ee8/

31. EUROPEAN COMMISSION. (2022). Flash Eurobarometer FL011EP: Media & News Survey 2022 [Dataset]. European Union Open Data Portal. Retrieved from https://data.europa.eu/data/datasets/s2832_fl011ep_eng?locale=en.
32. EUROPEAN COMMISSION. (2022). The 2022 Code of Practice on Disinformation. Retrieved March 18, 2023, fromhttps://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
33. EUROPEAN UNION EXTERNAL ACTION. (2022). Key narratives in pro-kremlin disinformation part 3: 'lost sovereignty'.https://euvsdisinfo.eu/key-narratives-in-pro-kremlin-disinformation-part-3-lost-sovereignty/#
34. EUROPEAN EXTERNAL ACTION SERVICE. (2021) About the European External Action Service. European Union. https://www.eeas.europa.eu/eeas/about-european-external-action-service_en
35. EVROPSKÉ HODNOTY O.P.S. (2022). Výroční zpráva 2021: Bezpečnostní centrum Evropské hodnoty [Annual report 2021: European Values Security Center]. Retrieved from https://europeanvalues.cz/wp-content/uploads/2022/06/CS_Bezpecnostni_centrum_Evropske_hodnoty_Vyrocni_zprava_2021.pdf
36. FACTMATA. https://factmata.com
37. FISCHER, J. (2019). Artificial Intelligence: China's High-Tech Ambitions. Louk Faesen et al., "Understanding the Strategic and Technical Significance of Technology for Security: Implications of AI and Machine Learning for Cybersecurity" (The Hague Security Delta (hsd), 2019).
38. FLORIDI, L., COWLS, J., BELTRAMETTI, M., CHATILA, R., CHAZERAND, P., DIGNUM, V., ET AL. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.
39. FLORIDI, L., COWLS, J., KING, T. C., & TADDEO, M. (2020). How to design AI for social good: Seven essential factors. Science and Engineering Ethics, 26(4), 1771–1796.
40. FLY, J., ROSENBERGER, L., & SALVO, D. (2018). Policy Blueprint for Countering Authoritarian Interference in Democracies. The German Marshall Fund of the United States (GMF).
41. FOREMSKI, T. (2019, 1. 1.). The first ambassador to Silicon Valley struggles with 'TechPlomacy'. ZDNet.https://www.zdnet.com/article/danish-ambassador-to-silicon-valley-struggles-with-techplomacy/
42. GELFERT, A. (2018). Fake news: A definition. Informal Logic, 38(1), 84-117. doi: 10.22329/il.v38i1.4858
43. GILLESPIE, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
44. GOULD-DAVIES, N. (2020). Belarus and Russian Policy: Patterns of the Past, Dilemmas of the Present. *Survival*, *62*(6), 179-198.
45. GRAVES, D. (2018). Understanding the promise and limits of automated fact-checking.
46. GREGOR, M., & MLEJNKOVÁ, P., "Explaining the challenge: From persuasion to relativisation." Challenging online propaganda and disinformation in the 21st century (2021): 3-41.
47. GURL, E. (2017). SWOT analysis: a theoretical review.
48. HARRINGTON, B. (2022). Sanctioning Russia's oligarchs–with shame. Bulletin of the Atomic Scientists, 78(6), 329-333.

49. HEARST, M. (2003). What is text mining. SIMS, UC Berkeley, 5.
50. HERD, G. P. (2022). Understanding Russian Strategic Behavior: Imperial Strategic Culture and Putin's Operational Code. Routledge.
51. HEROLDOVÁ, M., (2021) Aktivita ruských trollů v Česku výrazně poklesla po vyhoštění pracovníků ambasády. Aktualne.cz [online]. 28 Jun 2021 [cit. 11 Dec 2022]. Retrieved from: https://zpravy.aktualne.cz/domaci/aktivita-ruskych-trollu-v-cesku-vyrazne-poklesla-po-odchodu/r~d28cbc6cd7df11eb9106ac1f6b220ee8/
52. HORYCH, T. (2022). Bias Detection in Czech News. CVUT
53. HOSANAGAR, K., & JAIR, V. (2018, 25. 7.). We Need Transparency in Algorithms, But Too Much Can Backfire. Harvard Business Review. https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire
54. HOWARD, P. N., WOOLLEY, S., & CALO, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. Journal of Information Technology & Politics, 15(2), 81-93.
55. HU, M. (2020). Cambridge Analytica's black box. Big Data & Society, 7(2), 2053951720938091.
56. HUBER, D. E., SHIFFRIN, R. M., QUACH, R., & LYLE, K. B. (2002). Mechanisms of source confusion and discounting in short-term priming: 1. Effects of prime duration and prime recognition. Memory & Cognition, 30(5), 745-757.
57. HUGHES, H., & WAISMEL-MANOR, I. (2021). The Macedonian Fake News Industry and the 2016 US Election. PS: Political Science & Politics, 54(1), 19-23. DOI: 10.1017/S1049096520000992
58. HUGHES, J. A. (2007). Cyber Attacks Explained. CSIS Commentary, Center for Strategic and International Studies, Washington, DC.
59. CHITRA, U., & MUSCO, C. (2019). Understanding filter bubbles and polarization in social networks. *arXiv preprint arXiv:1906.08772*.
60. Institut pro výzkum veřejného mínění, Akademie věd České republiky, & Naše společnost. CVVM [Public Opinion Research Centre]. Retrieved from https://cvvmapp.soc.cas.cz/#question19.
61. JANIS, I. L., & MANN, L. (1977). Decision making: A psychological analysis of conflict, choice, and commitment. Free Press.
62. JOHNSON, J. (2019). Artificial intelligence & future warfare: implications for international security. Defense & Security Analysis, 35(2), 147-169.
63. JOST, P. J., PÜNDER, J., & SCHULZE-LOHOFF, I. (2020). Fake News - Does Perception Matter More Than the Truth? Journal of Behavioral and Experimental Economics, 101513. doi:10.1016/j.socec.2020.101513
64. KAMINSKI, M. E., & MALGIERI, G. (2021). Algorithmic impact assessments under the GDPR: producing multi-layered explanations. International Data Privacy Law, 11(2), 125-144.
65. KAPANTAI, E., CHRISTOPOULOU, A., BERBERIDIS, C., & PERISTERAS, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. New Media & Society, 23(5), 1301-1326. doi: 10.1177/1461444820966266

66. KARAMI, M., NAZER, T. H., & LIU, H. (2021). Profiling fake news spreaders on social media through psychological and motivational factors. In Proceedings of the 32nd ACM conference on hypertext and social media (pp. 225-230). doi: 10.1145/3460120.3484296
67. KERTYSOVA, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. Security and Human Rights, 29(1-4), 55-81.
68. KING, M. R., & CHATGPT. (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cellular and Molecular Bioengineering, 1-2.
69. KISS, Á., & SIMONOVITS, G. (2014). Identifying the Bandwagon Effect in Two-Round Elections. Public Choice, 160(3/4), 327-344. doi: http://www.jstor.org/stable/24507550
70. KLIMBURG, A. (2017). The darkening web: The war for cyberspace. Penguin Press.
71. KUHLTHAU, C. C. (1987). Information skills for an information society: A review of research. An ERIC information analysis product. ERIC.
72. LANOSZKA, A. (2016). Russian hybrid warfare and extended deterrence in eastern Europe. *International affairs*, *92*(1), 175-195.
73. LEVEY, C., & HAGEMANN, R. (2017, November 12). Algorithms with minds of their own: How do we ensure that artificial intelligence is accountable? Wall Street Journal. https://www.wsj.com/articles/algorithms-with-minds-of-their-own-1510521093
74. LEWANDOWSKY, S., COOK, J., ECKER, U. K. H., ALBARRACÍN, D., AMAZEEN, M. A., KENDEOU, P., LOMBARDI, D., NEWMAN, E. J., PENNYCOOK, G., PORTER, E RAND, D. G., RAPP, D. N., REIFLER, J., ROOZENBEEK, J., SCHMID, P., SEIFERT, C. M., SINATRA, G. M., SWIRE-THOMPSON, B., VAN DER LINDEN, S., VRAGA, E. K., WOOD, T. J., & ZARAGOZA, M. S. (2020). The Debunking Handbook 2020. Available at https://sks.to/db2020. DOI:10.17910/b7.1182
75. LIDDY, E. D. (2001). Natural language processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.
76. LOUK F. et al., (2019) "Understanding the Strategic and Technical Significance of Technology for Security: Implications of AI and Machine Learning for Cybersecurity" (The Hague Security Delta (hsd).
77. MARSDEN, C., & MEYER, T. (2019). Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism. European Parliament.
78. MERTON, R. K. (1948). The self-fulfilling prophecy. The antioch review, 8(2), 193-210.
79. MICHÁLKOVÁ, Z., (2023, 24. 3.). Za diskriminaci a rasismus v umělé inteligenci můžou lidé a špatná vstupní data, vysvětluje expert. https://www.irozhlas.cz/veda-technologie/technologie/umela-inteligence-rasismus-diskriminace_2303241128_pj
80. MILLER, J.D. (1983). "Scientific Literacy: a Conceptual and Empirical Review". Dedalus. 11: 29–48.
81. MINISTERSTVO PRŮMYSLU A OBCHODU. (2020). Národní strategie umělé inteligence v České republice [National strategy of artificial intelligence in the Czech Republic]. https://www.vlada.cz/assets/evropske-zalezitosti/umela-inteligence/NAIS_kveten_2019.pdf

82. MINISTERSTVO VNITRA ČR. (n.d.). Definice dezinformací a propagandy. Retrieved 18. 3. 2023, fromhttps://www.mvcr.cz/chh/clanek/definice-dezinformaci-a-propagandy.aspx
83. MISINFORMATION MONITOR, (2023) [online]. NewsGuard, Retrieved from:https://www.newsguardtech.com/misinformation-monitor/jan-2023/
84. MOHSENI, S., REZAPOUR, R., ABBASI, E., & MOVAHEDI, S. (2021). Machine learning explanations to prevent overtrust in fake news detection. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15).
85. MONTAZEROLGHAEM, A., YAGHMAEE, M. H., & LEON-GARCIA, A. (2020). Green Cloud Multimedia Networking: NFV/SDN Based Energy-Efficient Resource Allocation. IEEE Transactions on Green Communications and Networking, 4(3), 873-889.
86. MORAVEC, V., Focus Václava Moravce [online]. ČT, 2018 [cit. 18 Mar 2023]. Available at:https://www.ceskatelevize.cz/porady/11054978064-fokus-vaclava-moravce/218411030530009/cast/663175/
87. MORGAN, D. L., Focus Groups. Annual Review of Sociology, 1996, 22.1: 129-152.
88. MORGAN, D. L., KRUEGER, R.A., The focus group guidebook. Sage, 1998.
89. MUSTAK, M., SALMINEN, J., MÄNTYMÄKI, M., RAHMAN, A., & DWIVEDI, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. Journal of Business Research, 154, 113368. doi: 10.1016/j.jbusres.2022.11.008
90. NAPOLI, P. M. (2019). User data as public resource: Implications for social media regulation. *Policy & Internet*, *11*(4), 439-459.
91. NELEZ. (n.d.). Retrieved from https://www.nelez.cz
92. NEWTON, C. (2019, 25. 2.). The Trauma Floor: The Secret Lives of Facebook Moderators in America. The Verge.https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona
93. NGUYEN, A. (2021, 18. 8.). The Backfire Effect: Why Facts Don't Always Change Minds. Effectiviology.https://effectiviology.com/overkill-backfire-effect/
94. NGUYEN, A. T., KHAROSEKAR, A., KRISHNAN, S., KRISHNAN, S., TATE, E., WALLACE, B. C., & LEASE, M. (2018). Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18) (pp. 189-199). Association for Computing Machinery. https://doi.org/10.1145/3242587.3242666
95. NICKERSON, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.doi:10.1037/1089-2680.2.2.175
96. OPEN SOCIETY INSTITUTE - SOFIA. (2021). Media Literacy Index 2021 Double Trouble: Resilience to Fake News at the Time of Covid-19 Infodemic [Report]. Retrieved from https://osis.bg/wp-content/uploads/2021/03/MediaLiteracyIndex2021_ENG.pdf.
97. PAGE, S. W. (1950). Lenin and self-determination. The Slavonic and East European Review, 28(71), 342-358.
98. PANEL, O. (2019, 10. 4.). Algorithms, the illusion of neutrality: The road to trusted AI. Medium.https://towardsdatascience.com/algorithms-the-illusion-of-neutrality-8438f9ca8471

99. PAVLIK, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. Journalism & Mass Communication Educator, 10776958221149577.
100. POLYAKOVA, A., & BOYER, S. P. (2018). The future of political warfare: Russia, the West, and the coming age of global digital competition. Europe.
101. PRAGUE SECURITY STUDIES INSTITUTE. (2021). Stakeholder Perspectives on the Future of the Prague Security Studies Institute. Retrieved from https://www.pssi.cz/download//docs/8449_pssi-stakeholder-perspectives-en.pdf
102. PRIER, J. (2017). Commanding the Trend: Social Media as Information Warfare. Strategic Studies Quarterly, 11(4), 50-85.
103. PRONIN, E., LIN, D. Y., & ROSS, L. (2002). The bias blind spot: Perceptions of bias in self versus others. Personality and Social Psychology Bulletin, 28(3), 369-381.
104. REUTERS. (2022, 17. 7.). EU can no longer afford national vetoes on foreign policy, German chancellor says.https://www.reuters.com/world/europe/eu-can-no-longer-afford-national-vetoes-foreign-policy-germanys-scholz-2022-07-17/
105. REUTERS. (2023, 23. 3.). TikTok CEO to face tough questions as support for US ban grows. Reuters.https://www.reuters.com/technology/tiktok-ceo-face-tough-questions-support-us-ban-grows-2023-03-23/.
106. ROBBINS, J., et al. (2020). Countering Russian Disinformation. Center for Strategic and International Studies, 23.
107. ROSENBACH, E., & MANSTED, K. (2018). Can Democracy Survive in the Information Age? (Belfer Center for Science and International Affairs, Harvard Kennedy School, October 2018), https://www.belfercenter.org/publication/can-democracy-survive-information-age.
108. ROSS ARGUEDAS, A., ROBERTSON, C., FLETCHER, R., & NIELSEN, R. (2022). Echo chambers, filter bubbles, and polarisation: A literature review.
109. ROWLEY, C., & RAMASAMY, N. (2016). Horns effect. In Encyclopedia of Human Resource Management. Edward Elgar Publishing Limited.
110. SAFI, M. (2018, 3. 7.). 'WhatsApp murders': India struggles to combat crimes linked to messaging service. The Guardian., https://www.theguardian.com/world/2018/jul/03/whatsapp-murders-india-struggles-to-combat-crimes-linked-to-messaging-service.
111. SAMMUT-BONNICI, T., & GALEA, D. (2014). PEST analysis.
112. SEMERÁDOVÁ, T., & WEINLICH, P. (2019). Computer estimation of customer similarity with Facebook lookalikes: Advantages and disadvantages of hyper-targeting. *IEEE Access*, *7*, 153365-153377.
113. SHU, K., BHATTACHARJEE, A., ALATAWI, F., NAZER, T. H., DING, K., KARAMI, M., & LIU, H. (2020). Combating disinformation in a social media age. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6), e1385.
114. SCHWARZ, N., BLESS, H., STRACK, F., KLUMPP, G., & et al. (1991). Ease of retrieval as information: Another look at the availability heuristic. Journal of Personality and Social Psychology, 61(2), 195–202. doi:10.1037/0022-3514.61.2.195

115. SIMON, H. A. (1954). Bandwagon and Underdog Effects and the Possibility of Election Predictions. The Public Opinion Quarterly, 18(3), 245-253. doi: http://www.jstor.org/stable/2745982
116. SOCOR, V. (2006). Putin Offers Ukraine 'Protection' for Extending Russian Black Sea Fleet's Presence. Eurasia Daily Monitor, 3(204), 10-30.
117. SOUSA, R. DE, & MORTON, A. (2002). Emotional Truth. Proceedings of the Aristotelian Society, Supplementary Volumes76, 247–75. doi: http://www.jstor.org/stable/4106969
118. SWIRE-THOMPSON, B., DEGUTIS, J., & LAZER, D. (2020). Searching for the Backfire Effect: Measurement and Design Considerations. Journal of Applied Research in Memory and Cognition, 9(3), 286-299.
119. TANNER, B. (2022). EU Code of Practice on Disinformation. MVCR. Retrieved March 18, 2023, fromhttps://www.brookings.edu/blog/techtank/2022/08/05/eu-code-of-practice-on-disinformation/
120. TechTarget. (n.d.). Quantum computing. TechTarget. https://www.techtarget.com/whatis/definition/quantum-computing
121. THE DECISION LAB. (n.d.). Salience Bias. https://thedecisionlab.com/biases/salience-bias
122. UNIVERSITY OF HELSINKI. (2018, 6. 9.). Elements of AI: Finland is challenging the entire world to understand AI by offering a completely free online course – initiative got 1% of the Finnish population to study the basics.https://www.helsinki.fi/en/news/data-science-news/finland-is-challenging-the-entire-world-to-understand-ai-by-offering-a-completely-free-online-course-initiative-got-1-of-the-finnish-population-to-study-the-basics
123. VASWANI, K. (2019, 4. 4.). Concern over Singapore's anti-fake news law. BBC News.https://www.bbc.com/news/business-47782470
124. VINCENT, J. (2018, 12. 1.). Google's racist gorillas show why inclusive design matters. The Verge.https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai
125. WEEKS, B. E. (2015). Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation. Journal of Communication, 65(4), 699-719.
126. WELT, C. (2007). Russia and Its Post-Soviet Neighbors. In A. C. Kuchins, Alternative Futures for Russia to 2017 (pp. 51-70). Center for Strategic and International Studies.
127. WEST, D. M. (2017). How to Combat Fake News and Disinformation. The Brookings Institution. Retrieved fromhttps://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/
128. WOOLLEY, S. C., & HOWARD, P. N. (2016). Political communication, computational propaganda, and autonomous agents: Introduction. *International journal of Communication*, *10*.
129. WOOLLEY, S. C., & HOWARD, P. N. (2018). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.
130. ZHAI, X. (2022). ChatGPT user experience: Implications for education. Available at SSRN 4312418.

SCHVÁLENO

Institut komunikačních studií a žurnalistiky FSV UK Teze MAGISTERSKÉ diplomové práce

TUTO ČÁST VYPLŇUJE STUDENT/KA:

Příjmení a jméno diplomantky/diplomanta:
Štěpánová Barbora

Razítko podatelny:

Univerzita Karlova
Fakulta sociálních věd
Došlo dne: 13 -09- 2022 -1-
Čj: 204 Příloh:
Přiděleno:

Imatrikulační ročník diplomantky/diplomanta: 2020

E-mail diplomantky/diplomanta:
55995433@fsv.cuni.cz

Studijní obor/forma studia:
Strategická komunikace

Název práce v češtině:
užití současných AI nástrojů, jejich vývoj limity a predikce ve strategickém boji v informační válce

Název práce v angličtině:
The current development of AI tools in the Czech Republic and their possible use, limits and predictions in the strategic fight in the Information War

Předpokládaný termín dokončení (semestr, akademický rok – vzor: *ZS 2022/2023*)
(diplomovou práci je možné odevzdat nejdříve po dvou semestrech od schválení tezí)
LS 2022/2023

Charakteristika tématu a jeho dosavadní zpracování (max. 1800 znaků):

Práce podá ucelenou zprávu a hodnocení současného stavu výzkumu a vývoje AI nástrojů v kontextu zpravodajství a psychologických operací. Toto zhodnocení bude založeno na výzkumu současného stavu - od vyvíjených nástrojů až po predikce do budoucna, otázky financování, mezioborové a meziuniverzitní spolupráce, podpora a sociálně-ekonomický kontext. K tomuto zhodnocení nám pomůže série analýz zaměřených na představení externího i interního kontextu (PESTEL, SWOT, anylýza pěti sil). Tyto analýzy nám také poslouží k prognózování možného vývoje do budoucna, tedy co v rámci informační války lze očekávat, jaké vlivy můžeme pozorovat ve společnosti a kde máme měkká místa, které je potřeba chránit právě třeba za pomocí AI nástrojů.

V rámci podobných studií v současné době narážíme na stoupající potřebu pro mezidisciplinární propojení, pro účely této práce je však vhodný příklad spolupráce FSV UK při vytváření anotací pro výzkum na ČVUT zaměřený na odhalování bias v médiích. (HORYCH, Tomáš. *Bias Detection in Czech News*. Praha, 2022. bakalářská práce. České Vysoké Učení Technické). Přesně tento druh spolupráce a závěry, které tato práce přinesla, chceme v první části probírat.

Ve výzkumné části pak bude náš cíl tyto koncepty posunout, vytvořit predikce do budoucna, identifikovat limity a překážky, které brání v současné době mezioborové spolupráci. Navrhnout řešení a intervence, krteré nám mohou pomoc tyto překážky překonat. Přinese nám i pohled různých oborů na současný stav a odhadovaný budoucí vývoj aktérů, tyto otázky budou kladeny a probírány formou focus group. (podobní práce: FIGUEIRA, Álvaro a Luciana OLIVEIRA. The current state of fake news: challenges and opportunities. Procedia Computer Science [online]. 2017, 121, 817-825 [cit. 2022-07-16]. ISSN 18770509. Dostupné z: doi:10.1016/j.procs.2017.11.106)

Výstupy z této práce by pak měly být podkladem pro vytvoření jakési příručky pro mezidisciplinární spolupráci obecně, ale hlavně na využití poznatku ze strategické komunikace věd, které využívá pro vývoj AI a dalších IT nástrojů a co je třeba překonat pro úspěšnou spolupráci. (po vzoru této příručky: Hadorn, Gertrude Hirsch, Holger Hoffmann-Riem, Susette Biber-Klemm, Walter Grossenbacher-Mansuy, Dominique Joye, Christian Pohl, Urs Wiesmann, and Elisabeth Zemp, eds. *Handbook of transdisciplinary research*. Vol. 10. Dordrecht: Springer, 2008.)

Předpokládaný cíl práce, případně formulace problému, výzkumné otázky nebo hypotézy (max. 1800 znaků):
Předpokládaným cílem je zodpovědět otázku „jaká je budoucnost AI v České republice, Jak ji můžeme využít v boji s dezinformacemi a psychologickými operacemi? a jaké překážky musíme překonat?"
Práce by se dala rozdělit na sérii podotázek, a to:
O1 – současný stav – co děláme, ubíráme se správným směrem, využíváme naplno zdroje?
O2 – Jak můžeme současný stav zlepšit, jak navrhnout ideální mezidisciplinární spolupráci, co je potřeba udělat, aby se odstranily překážky pro další vývoj? Jaké současné nástroje můžeme využít v informační válce?
O3 – Jaké máme predikce do budoucna, jaké nástroje potřebujeme?
O4 – jaké jsou hranice a limity?

Předpokládaná struktura práce (rozdělení do jednotlivých kapitol a podkapitol se stručnou charakteristikou jejich obsahu):

1. **Úvod**
2. **Teoretická část**
2.1. **Představení prostředí, vývoj, současný stav, predikce, úvahy o informační válce**
2.2. **shrnutí úvah a predikcí ohledně AI a jejich limitech a budoucnosti**
2.3. **Propojení přírodních a sociálních věd**

3. **AI v odvětví médií, komunikace a sociálních věd**
3.1. **Využití AI v praxi**
3.2. **limity a příležitosti**
3.3. **AI v České republice**

4. **Analýza současného stavu**
5. **Praktická část**
5.1. **Představení Focus Group a designu výzkumu**
5.2. Obsahová část výzkumu, podměty a témata
5.3. **Průběh a výsledky**
5.4. **diskuse**
6. **Závěr**

Vymezení podkladového materiálu (např. titul periodika a analyzované období):

Vývoj a výzkum AI v tematických periodikách, disertačních a vědeckých pracích a na vybraných univerzitách, predikce a případové studie v posledních 6 letech, analýzy vypracované pro účely této práce a záznamy z rozhovorů ve focus group.

Metody (techniky) zpracování materiálu:

Detailní rešerše literaturyi, které se zabývá problematikou AI a možného využití v kontextu sociálních věd, predikcemi, etickým rámcem a hodnocením do budoucnosti. Kritické zhodnocení a porovnání závěrů.

Provedení základních analýz současného stavu a prostředí.

Vyhodnocení současného pohledu na AI a jeho budoucnost, limitů a příležitostí a následné diskutování těchto závěrů v 7členné Focus Group s odborníky napříč disciplínami. Skupině bude předložena série témat, podnětů k diskusi a jejich závěry pak budou předloženy vyhodnocení.

Z jejich diskuze se pokusím zodpovědět hlavní otázky, které jsou ohledně AI jeho využití a budoucnosti v ČR, a které jsme formulovali výše.
Tyto závěry pak zhodnotíme v kontextu této práce a poznatků z dřívějšího bádání.

Základní literatura (nejméně 5 nejdůležitějších titulů k tématu a metodě jeho zpracování; u všech titulů je nutné uvést stručnou anotaci na 2-5 řádků):

LONSDALE, David J. The Nature of War in the Information Age[online]. Routledge, 2004 [cit. 2022-07-16]. ISBN 9781135757212. Dostupné z: doi:10.4324/9780203508176

Kniha na současném kontextu testuje, zda informační věk změní samotnou povahu války vycházející z Clausewitzovi teorie, stejně jako zbraně, kterými se bojuje, prostor, funkce velení, nebo cíle. Kniha se zabývá koncepty strategické informační války či informační moci.

FIGUEIRA, Álvaro a Luciana OLIVEIRA. The current state of fake news: challenges and opportunities. Procedia Computer Science [online]. 2017, 121, 817-825 [cit. 2022-07-16]. ISSN 18770509. Dostupné z: doi:10.1016/j.procs.2017.11.106

V současné době je i díky sociálním sítím přenos informací tak rychlý, nekontrolovatelný a zesílený, že rozšíření nepravdivých nebo zkreslených informací je možné během pár minut a může mít instantní reálné dopady na tisíce lidí. Obavy z tohoto šíření rostou, a proto jsou vytvářeny různá doporučení a predikce do budoucna. Článek popisuje dva protichůdné přístupy a navrhuje algoritmické řešení, které syntetizuje hlavní problémy.

KHALDAROVA, Irina a Mervi PANTTI. Fake News. Journalism Practice [online]. 2016, 10(7), 891-901 [cit. 2022-07-16]. ISSN 1751-2786. Dostupné z: doi:10.1080/17512786.2016.1163237

Krize na Ukrajině zdůraznila pozici ruské televize jako nejsilnějšího aktéra vlády v informační válce. Internet však umožňuje ostatním hráčům zpochybnit narativ Kremlu poskytováním protipříběhů a odhalováním zkreslených informací a falešných obrázků. Na této práci tedy budeme sledovat některé nástroje boje v informační válce a jejich využití a účinnost v praxi.

SCHREIBER, David, Cristina PICUS, David FISCHINGER a Martin BOYER. The defalsif-AI project: protecting critical infrastructures against disinformation and fake news. *E & i Elektrotechnik und Informationstechnik* [online]. 2021, **138**(7), 480-484 [cit. 2022-07-16]. ISSN 0932-383X. Dostupné z: doi:10.1007/s00502-021-00929-7

Článek popisuje koncept a probíhající práci projektu defalsif-AI, který se zabývá ochranou kritických infrastruktur před dezinformacemi, fake news a před umělými útoky na sociální sítě. Článek provedl i analýzu a posouzení týkající se práva a společenských věd. Na tomto článku a nástroji budeme demonstrovat využití AI nástrojů ve strategickém boji v informační válce.

ELLIOTT, Anthony. *The Culture of AI* [online]. New York: Routledge, 2019 [cit. 2022-07-16]. ISBN 9781315387185. Dostupné z: doi:10.4324/9781315387185

Kniha zkoumá, jak inteligentní stroje, pokročilá robotika, zrychlující se automatizace, velká data a internet všeho ovlivňují každodenní život a současné společnosti a zdůrazňuje ústřední roli umělé inteligence ve všem, co děláme. Kniha nabízí komplexní úvod do problematiky AI od historie, automatizace, propojení se soukromým životem, vlivem na člověka, etikou a budoucností z pohledu sociologa.

Demartini, Gianluca, Stefano Mizzaro, and Damiano Spina. "Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities." *IEEE Data Eng. Bull.* 43, no. 3 (2020): 65-74.

Článek představuje současné možnosti algoritmického vytváření dezinformací a schopnost je přizpůsobovat a mikrocílit na jednotlivé uživatele. Podobné metody založené na datech a algoritmech lze použít také k detekci dezinformací a ke kontrole jejich šíření. Automatický odhad spolehlivosti a důvěryhodnosti informací se silně spoléhá na lidské experty. V tomto článku jsou formulované výzvy a příležitosti spojené s kombinováním automatických a manuálních přístupů k ověřování faktů v boji proti šíření dezinformací.

Floridi, Luciano, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo. "How to design AI for social good: seven essential factors." *Science and Engineering Ethics* 26, no. 3 (2020): 1771-1796.

Myšlenka umělé inteligence pro sociální dobro získává na síle v informačních společnostech obecně a v komunitě AI zvláště. Má potenciál řešit sociální problémy prostřednictvím vývoje řešení založených na umělé inteligenci. K dnešnímu dni však existuje pouze omezené chápání toho, co dělá AI společensky dobrou. Tento článek řeší tuto mezeru identifikací sedmi etických faktorů, které jsou zásadní pro budoucí iniciativy.

Zeng, Yi, Enmeng Lu, and Cunqing Huangfu. "Linking artificial intelligence principles." *arXiv preprint arXiv:1812.04814*(2018).

Tato práce definuje principy umělé inteligence a sociální a etické úvahy o ní z různých perspektiv a disciplín a argumentuje nutnost začlenit různé principy umělé inteligence do komplexního rámce a zaměřit se na to, jak se mohou vzájemně ovlivňovat a doplňovat.

Caramancion, Kevin Matthe. "An exploration of disinformation as a cybersecurity threat." In *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, pp. 440-444. IEEE, 2020.

Šíření a přetrvávání dezinformací, zejména na stránkách sociálních sítí, v současnosti představuje jednu z nejnáročnějších hrozeb pro uživatele i správce obsahu. V tomto článku si autor klade za cíl prozkoumat dynamiku několika vzájemně se ovlivňujících oblastí, tj. psychologie a informatiky, jejich vliv na tento fenomén, což poskytuje ideální interdisciplinární a holistický přístup k jeho redukci a řízení. Další částí tohoto dokumentu je pokus obhajovat formální uznání dezinformací jako hrozby kybernetické bezpečnosti.

Lunt, Peter, and Sonia Livingstone. "Rethinking the focus group in media and communications research," *Journal of communication* 46, no. 2 (1996): 79-98.

Tento článek popisuje historii Focus Group jako výzkumného nástroje, od jeho původního použití Lazarsfeldem a Mertonem v raném komunikačním výzkumu až po jeho úpadek, protože společenskovědní výzkum se stal kvantitativním a experimentálním. Přezkoumáváme současné využití Focus Group prováděných v rámci kritické tradice a přehodnocuje metodu a její vhodnost pro výzkum médií a komunikace a popisuje vhodnost použití.

Parker, Andrew, and Jonathan Tritter. "Focus group method and methodology: current practice and recent debate." *International Journal of Research & Method in Education* 29, no. 1 (2006): 23-37.

Tento článek se zabývá používáním Focus Group jako metody sběru dat v rámci kvalitativního výzkumu. Autoři vycházejí ze svých vlastních zkušeností, aby poskytli přehled o problémech a debatách.

Hadorn, Gertrude Hirsch, Holger Hoffmann-Riem, Susette Biber-Klemm, Walter Grossenbacher-Mansuy, Dominique Joye, Christian Pohl, Urs Wiesmann, and Elisabeth Zemp, eds. *Handbook of transdisciplinary research*. Vol. 10. Dordrecht: Springer, 2008.

Příručka představuje systematické umístění transdisciplinárního výzkumu do rozvoje vědních a humanitních věd a 15 návrhů na posílení transdisciplinárního výzkumu umísťují příspěvky do širšího, systematizovaného kontextu. Příručka poskytuje výzkumníkům a studentům přehled o stavu techniky v transdisciplinárním výzkumu.

Diplomové a disertační práce k tématu (seznam bakalářských, magisterských a doktorských prací, které byly k tématu obhájeny na UK, případně dalších oborově blízkých fakultách či vysokých školách za posledních pět let)

ZÁLEŠÁK, Tomáš. Sociologické aspekty umělé inteligence. Praha, 2019. Bakalářská práce. Univerzita Karlova, Fakulta sociálních věd, Katedra sociologie. Vedoucí práce Bureš, Jiří.

KOSUB, Tomáš. Private AI and the State - Potential for a Conflict. Praha, 2021. Diplomová práce. Univerzita Karlova, Fakulta sociálních věd, Katedra mezinárodních vztahů. Vedoucí práce Bahenský, Vojtěch.

NAKAMURA, Ai. Fully Autonomous Weapons System (AWS): Analysis of AWS with regard to IHL and Martens Clause. Praha, 2021. Diplomová práce. Univerzita Karlova, Fakulta sociálních věd, Katedra bezpečnostních studií. Vedoucí práce Špelda, Petr.

MILIUTINA, Kseniia. Analysis of Russian policy on the development of AI in the military. Praha, 2022. Diplomová práce. Univerzita Karlova, Fakulta sociálních věd, Katedra bezpečnostních studií. Vedoucí práce Kučera, Tomáš.

Sukdol, Štěpán. "Vliv umělé inteligence na digitální komunikaci a média a jejich budoucí vývoj." (2021). Bakalářská práce. Univerzita Karlova, Fakulta sociálních věd, Katedra bezpečnostních studií. Vedoucí práce Kučera, Tomáš.

BROŽKOVÁ, Alžběta. Budoucnost umělé inteligence v marketingové komunikaci [online]. 2020 [cit. 2022-07-17]. Dostupné z: https://is.vsfs.cz/th/mp1bd/. Diplomová práce. Vysoká škola finanční a správní. Vedoucí práce František ZICH.

Sířinek, Tomáš. (2021). Umělá inteligence, armáda a pátá doména: Potenciální vliv umělé inteligence na psychologické operace v kontextu aktivní kybernetické obrany. 10.13140/RG.2.2.13288.60161.
Datum / Podpis studenta/ky

TUTO ČÁST VYPLŇUJE PEDAGOG/PEDAGOŽKA:
Doporučení k tématu, struktuře a technice zpracování materiálu:
Případné doporučení dalších titulů literatury předepsané ke zpracování tématu:
Potvrzuji, že výše uvedené teze jsem s jejich autorem/kou konzultoval(a) a že téma odpovídá mému oborovému zaměření a oblasti odborné práce, kterou na FSV UK vykonávám. **Souhlasím s tím, že budu vedoucí(m) této práce.** **Příjmení a jméno pedagožky/pedagoga** KLABÍKOVÁ KÁBOVÁ TEREZA **Datu... ...y/pedagoga**

TEZE JE NUTNO ODEVZDAT **VYTIŠTĚNÉ**, **PODEPSANÉ** A VE **DVOU** VYHOTOVENÍCH DO TERMÍNU UVEDENÉHO V HARMONOGRAMU PŘÍSLUŠNÉHO AKADEMICKÉHO ROKU, A TO PROSTŘEDNICTVÍM PODATELNY FSV UK. PŘIJATÉ TEZE JE NUTNÉ SI **VYZVEDNOUT** V SEKRETARIÁTU PŘÍSLUŠNÉ KATEDRY A **NECHAT VEVÁZAT** DO OBOU VÝTISKŮ DIPLOMOVÉ PRÁCE.

TEZE NA IKSŽ SCHVALUJE GARANT PŘÍSLUŠNÉHO STUDIJNÍHO OBORU.

Seznam příloh:

Příloha č. 1: Transkript focus group (text)

Příloha č. 2: Transkript individuálního rozhovoru (text)