Monitorovacia správa - záverečná

(Táto príloha sa do systému e-grant nevkladá, slúži ako pomôcka na prípravu Monitorovacej správy. Monitorovacia správa sa predkladá prostredníctvom formuláru v systéme e-grant.
V prípade projektov podporených v rámci výziev:-
-• 09I05-03-V02 - Podpora výskumných projektov zameraných na digitalizáciu ekonomiky v TRL úrovniach 1-3,
-• 09I04-03-V02 - Podpora výskumných projektov zameraných na dekarbonizáciu ekonomiky v TRL úrovniach 1-3,
• 09I03-03-V03 - Veľké projekty pre excelentných výskumníkov,
Prijímateľ vypĺňa MS v systéme e-grant v slovenskom aj v anglickom jazyku, pričom v oboch jazykoch uvedie rovnaké informácie.

Príloha č. 2 PpP verzia 2.6

Kód projektu:	*09I03-03-V03-00020*	Kód výzvy:	*09I01-03-V03*
Poradové číslo monitorovacej správy:	*0002*	Komponent:	*K09 Efektívnejšie riadenie a posilnenie financovania výskumu, vývoja a inovácií*
Aktuálne monitorovacie obdobie:	*1.8.2024 - 30.6.2026*	Investícia:	*3. Excelentná veda*
Jednoznačná identifikácia projektu:	*09I03-03-V03-00020, Modelom riadené auditovanie AI algoritmov sociálnych sietí a ich tendencií šíriť škodlivý obsah* *Meno a priezvisko hlavného riešiteľa: Ivan Srba*		

1. Popis činností v rámci projektu uskutočnených počas monitorovacieho obdobia

V tejto časti sa uvádza popis vykonaných činností súvisiacich s realizáciou projektu za obdobie, za ktoré sa MS predkladá a ich prínos k napĺňaniu stanovených míľnikov a výstupov projektu (ak relevantné, napr. vyžaduje sa explicitný popis v prípade laboratórneho výskumu). Prijímateľ v tejto časti uvedie, či projekt postupoval v súlade s harmonogramom alebo bol v omeškaní, pričom uvedie aké činnosti neboli vykonané v dôsledku omeškania. V prípade, že bola vecná realizácia projektu v období od začiatku do konca monitorovacieho obdobia prerušená/pozastavená, prijímateľ v tejto časti uvedie obdobie prerušenia/pozastavenia vecnej realizácie projektu.

Celé obdobie riešenia projektu (od 1. augusta 2024 do 30. júna 2026) prebiehalo v súlade s harmonogramom projektu. V nasledujúcom texte uvádzame prehľad aktivít realizovaných počas tohto obdobia v rámci jednotlivých pracovných balíkov a úlohách.

WP1: Riadenie projektu

Cieľom pracovného balíka WP1 bolo zabezpečiť efektívne riadenie projektu (administratívne, finančné, kvalitatívne a riadenie rizík), podporiť plynulý priebeh riešenia projektu a včasné podávanie projektových správ, ako aj zabezpečiť riadenie etických aspektov a správu dát.

T1.1 Administratívny, finančný manažment a riadenie rizík (M1–M23)

Úloha T1.1 bola zameraná na celkový administratívny a finančný manažment projektu a riadenie rizík. Počas celého riešenia projektu projektový tím nastavoval efektívne komunikačné postupy a nástroje spolupráce, monitoroval priebeh projektu, alokáciu zdrojov a čerpanie rozpočtu a zabezpečoval pravidelnú komunikáciu s financujúcou agentúrou. Súčasťou tejto úlohy bola aj finančná a právna administratíva, ktorá zabezpečovala súlad s požiadavkami na financovanie a vykazovanie projektu.

Ďalšou významnou súčasťou tejto úlohy bola koordinácia projektového reportovania a dokumentácie. Zahŕňala prípravu a odovzdanie projektových výstupov (angl. deliverables), priebežnej správy o riešení projektu, ako aj prípravu tejto záverečnej monitorovacej správy. Súčasťou úlohy bolo aj zabezpečenia, aby všetky plánované aktivity, míľniky a projektové výstupy boli realizované a odovzdané v súlade s harmonogramom projektu.

Riadenie rizík prebiehalo kontinuálne počas celého riešenia projektu. Riziká (identifikované už v projektovom návrhu) boli pravidelne aktualizované s cieľom identifikovať, monitorovať a zmierňovať potenciálne interné, externé, strategické a operatívne riziká (najmä tie súvisiace s etickými a právnymi aspektmi, obmedzenou dostupnosťou dát poskytovaných platformami sociálnych médií a dynamicky sa meniacim prostredím umelej inteligencie, najmä veľkých jazykových modelov (LLMs, angl. *Large Language Models*), a samotných platforiem sociálnych sietí). Proces riadenia rizík umožnil projektovému tímu proaktívne reagovať na vznikajúce výzvy, implementovať primerané mitigačné opatrenia a zabezpečiť úspešnú realizáciu projektu v súlade s jeho cieľmi a harmonogramom. Počas riešenia projektu nevznikli žiadne významné administratívne, finančné ani operatívne problémy, ktoré by ohrozili úspešnú realizáciu plánovaných aktivít.

T1.2 Riadenie etických aspektov (M1–M23)

Úloha T1.2 bola zameraná na zabezpečenie toho, aby všetky projektové aktivity spĺňali najvyššie štandardy etiky umelej inteligencie a vedeckého výskumu. Počas celého projektu boli etické aspekty systematicky integrované do výskumného procesu s cieľom zabezpečiť, aby vyvíjané metodológie algoritmického auditu, postupy zberu dát a realizované auditovacie štúdie rešpektovali práva a súkromie všetkých potenciálne dotknutých osôb a boli v súlade s platnými etickými a právnymi požiadavkami.

Počas prvého roku riešenia projektu bolo zrealizované komplexné etické posúdenie výskumných a vývojových aktivít projektu. V rámci tohto hodnotenia boli v období od 10. apríla do 12. júna 2025 zorganizované štyri etické workshopy, na ktorých sa stretli výskumníci so zameraním na umelú inteligenciu, spoločenské vedy, filozofiu a právo. Počas workshopov boli kriticky analyzované etické a právne dôsledky plánovaných výskumných aktivít, identifikované zainteresované strany potenciálne ovplyvnené projektom, posúdené možné riziká vyplývajúce z navrhovanej metodiky algoritmického auditu a formulované primerané stratégie na ich zmiernenie.

Počas druhého roka riešenia projektu sa aktivity v rámci tejto úlohy sústredili najmä na implementáciu a monitorovanie mitigačných opatrení definovaných v zozname rizík spolu s navrhnutými technikami ich mitigácie (príloha tejto správy). Tieto opatrenia boli priebežne uplatňované a vyhodnocované počas celého projektu s cieľom minimalizovať etické, právne a spoločenské riziká spojené s výskumom algoritmického auditovania. Mimoriadna pozornosť bola venovaná zabezpečeniu toho, aby výskumná metodológia zostala v súlade s etickými princípmi schválenými internou etickou komisiou KInITu, ktorá projektu AI-Auditology udelila etické schválenie dňa 17. decembra 2024.

Kontinuálny etický dohľad zabezpečil, že projekt bol realizovaný zodpovedným a transparentným spôsobom a že etické aspekty zostali neoddeliteľnou súčasťou všetkých výskumných aktivít od začiatku až po ukončenie projektu.

WP2: Konštrukcia modelu sociálnych médií

V rámci WP2 boli vyvinuté techniky na vytváranie, reprezentáciu a inferenciu z modelu sociálnych médií (kľúčového prvku umožňujúceho modelovo založené audity, angl. *Model-based algorithmic auditing*). Jeho neoddeliteľnou súčasťou je reprezentácia populácie používateľov sociálnych médií – vrátane demografických údajov (napr. rozdelenie veku, pohlavia, geografickej polohy) a vzorcov správania (napr. záujmy, počet relácií denne, dĺžka relácií, rozdelenie a frekvencia rôznych typov interakcií).

T2.1 Vývoj a extrakcia reprezentácie (M1-M18)

Úloha T2.1 sa zameriavala na vývoj jednotnej, od platformy nezávislej reprezentácie platforiem sociálnych médií, používateľov a interakcií, ktorá slúžila ako základ pre model sociálnych médií využívaný v projekte. Pri návrhu implementácie sa explicitne zohľadňovala obmedzená dostupnosť údajov zo sociálnych sietí a obmedzený prístup k aplikačným rozhraniam platforiem (t.j. obmedzená dostupnosť informácií o jednotlivých používateľoch, príspevkoch a interakciách medzi nimi). Vyvinutý prístup sa preto zameral na zachytenie charakteristík platforiem na agregovanej úrovni prostredníctvom modulárnej reprezentácie založenej na štatistických a pravdepodobnostných rozdeleniach.

V súlade s návrhom projektu využíva vyvinutý model sociálnych médií architektúru prekryvných modelov (angl. *overlay models*), pozostávajúcu z doménovej vrstvy a platformovej vrstvy. Doménová vrstva reprezentuje koncepty a enumerácie nezávislé od konkrétnej platformy vrátane demografických charakteristík používateľov (ako vekové skupiny, kategórie pohlavia a geografické regióny), charakteristík obsahu a vzorcov správania (ako miera interakcií, čas strávený na platforme a frekvencia používania). Platformová vrstva následne reprezentuje jednotlivé platformy sociálnych médií prostredníctvom štatistických rozdelení nad prvkami definovanými v doménovej vrstve, čo umožňuje zachytiť špecifiká jednotlivých platforiem pri zachovaní jednotnej reprezentácie naprieč rôznymi platformami.

V rámci tejto úlohy boli identifikované a analyzované vhodné zdroje údajov na vytvorenie modelu sociálnych médií. Významné úsilie bolo venované získaniu prístupu k *TikTok Research API*. Napriek rozsiahlej komunikácii s podporou TikToku a viacerým žiadostiam o prístup a odvolaniam bol AI-Auditology tímu poskytnutý iba obmedzený prístup k prostrediu *Virtual Compute Environment* (VCE), ktoré neposkytovalo požadovaný typ ani rozsah údajov. Projektový tím preto využil alternatívne zdroje dát na podporu vývoja modelu. Analýza sa zamerala predovšetkým na platformy TikTok a YouTube, ktoré boli identifikované ako relevantné platformy v rámci projektových prípadov použitia (D5.1).

Využité zdroje údajov zahŕňali *CEDMO Trends*, longitudinálny prieskum verejnej mienky poskytujúci poznatky o spôsoboch používania sociálnych médií, ako aj vedecké publikácie, dokumentáciu platforiem, správy o transparentnosti a verejne dostupné štatistické zdroje. Najmä verejne dostupné marketingové analýzy poskytli cenné štatistiky na úrovni platforiem opisujúce demografické rozdelenia, geografický dosah, intenzitu používania, preferencie obsahu a vzorce interakcií.

Extrakcia a harmonizácia informácií z identifikovaných zdrojov prebiehala poloautomatizovaným spôsobom. Extrahované informácie boli zjednotené do kategórií nezávislých od platformy vrátane demografických charakteristík (vekové skupiny, kategórie pohlavia a geografické regióny) a behaviorálnych charakteristík (preferencie obsahu, miera interakcií, priemerný čas strávený na platforme, čas sledovania a frekvencia relácií). Tieto zjednotené reprezentácie umožnili vytvorenie konfigurácií modelov špecifických pre jednotlivé platformy pri zachovaní požadovanej kompatibility medzi rôznymi platformami sociálnych médií.

Pre každú podporovanú platformu bola vytvorená samostatná platformová vrstva obsahujúca príslušné demografické a behaviorálne rozdelenia. Výsledný model bol implementovaný ako konfigurovateľný framework, v ktorom je možné aktualizovať charakteristiky platforiem úpravou konfiguračných súborov bez potreby zmien v základnej implementácii. Tento návrh umožňuje priebežné aktualizácie pri dostupnosti nových zdrojov informácií a zároveň zachováva historickú porovnateľnosť medzi jednotlivými verziami modelu.

Na podporu generovania reprezentatívnych auditovacích scenárov poskytuje vyvinutý model funkcionality na dopytovanie a generovanie syntetických používateľov. Používateľské profily je možné generovať vzorkovaním demografických charakteristík z pravdepodobnostných rozdelení špecifických pre jednotlivé platformy a následným odvodením príslušných vzorcov správania podmienených týmito charakteristikami. Generovaní syntetickí používatelia majú pridelené demografické atribúty, preferencie obsahu, charakteristiky interakcií, odhadovanú intenzitu používania a parametre súvisiace s interakciami, čo umožňuje reprodukovateľné vytváranie reprezentatívnych populácií používateľov pre experimenty zamerané na algoritmický audit.

Viac informácií o vyvinutom modeli sociálnych médií je dostupných v *D2.1 Harmful content modelling methods*, kapitola 2.

Vyvinutý model sociálnych médií (D2.2) bol následne sprístupnený ako služba založená na FastAPI, poskytujúca REST API endpointy na prístup k informáciám o platformách, generovanie syntetických používateľských profilov a vytváranie používateľských kohort na základe definovaných obmedzení. Implementácia oddeľuje konfiguračné údaje, logiku modelu a funkcionalitu API, čím umožňuje jednoduchú údržbu a budúce rozšírenie o ďalšie platformy alebo štatistické ukazovatele. Vyvinutý model sociálnych sietí tak poskytuje flexibilný a rozšíriteľný základ pre podporu algoritmického auditovania tým, že umožňuje generovanie štatisticky podložených a medzi platformami porovnateľných reprezentácií používateľov.

T2.2 Predikcia používateľských interakcií (M7-M23)

Úloha T2.2 sa zameriavala na vývoj metód na predikciu a simuláciu používateľských interakcií so obsahom sociálnych sietí v kontexte algoritmického auditu. V rámci tejto úlohy boli vyvinuté prediktory nasledujúcej používateľskej interakcie (angl. *user interaction predictors*), ktoré automaticky určujú, či by simulovaný používateľ mal interagovať s konkrétnym príspevkom na sociálnej sieti a aký typ interakcie by mal vykonať. Vyvinutý prístup bol navrhnutý na podporu realistickej realizácie algoritmických auditov prostredníctvom automatickej analýzy odporúčaného obsahu v reálnom čase a generovania používateľských spätných väzieb potrebných na vykonanie auditovacích scenárov.

Na rozdiel od jednoduchých heuristík alebo vopred definovaných pravidiel založených na metadátach, ako sú hashtagy (ktoré sú rozšírené v predchádzajúcich výskumných prácach), nami vyvinutý prístup využíva klasifikáciu obsahu založenú na veľkých jazykových modeloch (angl. *Large Language Models - LLM*) a veľkých multimodálnych jazykových modeloch (angl. *Large Vision-Language Models - LVLM*) na analýzu relevantnosti odporúčaného obsahu vzhľadom na vopred definované charakteristiky používateľa (viac informácií je uvedených v úlohe T2.3).

Počas vykonávania auditu prediktor analyzoval charakteristiky simulovaného používateľa, dostupné metadáta videa a samotný obsah s cieľom určiť, či odporúčaná položka zodpovedá záujmom používateľa. Pri obsahu zodpovedajúcom záujmom používateľa vykonával simulovaný používateľ silnejšie implicitné a explicitné spätnoväzbové akcie vrátane sledovania videa až do konca, označenia páči sa mi a uloženia videa. Ostatný obsah bol preskočený, pričom živé vysielania (angl. livestreams) a nadmerne dlhé videá boli vylúčené zo simulácie interakcií. Okrem toho sa skúmali aj možnosti dosiahnutia ešte realistickejšej simulácie používateľského správania s využitím najnovšieho pokroku v oblasti simulácie používateľov založenej na LLM.

Viac informácií o vyvinutom modeli sociálnych médií je dostupných v *D2.1 Harmful content modelling methods*, kapitoly 3 a 4.

Vyvinutý prístup na predikciu používateľských interakcií umožňuje realistickejšiu a reprodukovateľnú simuláciu správania používateľov počas algoritmických auditov a poskytuje významný komponent na realizáciu rozsiahlych auditov odporúčacích systémov.

T2.3 Anotovanie obsahu (M1-M23)

Úloha T2.3 sa zameriavala na vývoj a vyhodnotenie metód automatickej klasifikácie a označovania obsahu s cieľom analýzy obsahu sociálnych sietí potrebnú pre predikciu používateľských interakcií (T2.2), ako aj vyhodnocovanie auditov (T4.2). Aktivity v rámci tejto úlohy riešili jednak znižovanie potreby manuálnej anotácie, ako aj širšiu výzvu spoľahlivej klasifikácie obsahu v oblastiach, kde sú anotované údaje často limitované, nákladné na získanie a nerovnomerne distribuované naprieč jazykmi a témami.

Projekt skúmal využitie metód strojového učenia (angl. *machine learning*) a spracovania prirodzeného jazyka (angl. *natural language processing*), pričom osobitná pozornosť bola venovaná využitiu veľkých jazykových modelov (LLM) a veľkých multimodálnych jazykových modelov (LVLM).

V prvej fáze výskum skúmal, ako môžu metódy založené na LLM/LVLM, stratégie tvorby promptov (angl. *prompt construction*) a techniky generovania syntetických údajov podporiť klasifikačné úlohy v podmienkach s nedostatkom zdrojov, čo predstavuje (výskumnú) výzvu v kontexte analýzy sociálnych sietí. Na rozdiel od tradičných prístupov založených na učení s učiteľom (angl. *supervised learning*), ktoré vyžadujú rozsiahle manuálne anotované datasety, skúmané prístupy smerovali k zlepšeniu výkonu modelov v situáciách, keď je dostupné iba obmedzené množstvo označkovaných údajov.

V rámci tejto úlohy bolo preto realizovaných viacero výskumných štúdií systematicky analyzujúcich správanie, obmedzenia a možnosti prístupov založených na LLM pre klasifikáciu obsahu v podmienkach s nedostatkom (označkovaných) dát. Výskum skúmal vplyv náhodnosti a metodologických rozhodnutí pri trénovaní modelov z malých anotovaných datasetov, analyzoval výhody/nevýhody medzi LLM a štandardnými metódami textovej augmentácie a hodnotil rôzne stratégie výberu príkladov pre LLM augmentáciu. Projekt zároveň porovnával špecializované malé jazykové modely so všeobecnými veľkými jazykovými modelmi a skúmal prístupy ku klasifikácii vo viacjazyčných a nízkozdrojových jazykových (angl. *low-resource*) prostrediach.

Na základe výsledkov týchto štúdií boli následne v prostredí s nízkou dostupnosťou dát natrénované požadované klasifikátory obsahu. Konkrétne adresovali štyri klasifikačné úlohy: klasifikáciu tém, klasifikáciu postoja voči identifikovanej téme, detekciu reklamy (vrátane klasifikácie typu reklamy) a klasifikáciu témy reklamy. Tieto klasifikátory boli vyhodnotené a použité v dvoch algoritmických auditovacích štúdiách realizovaných v rámci projektu: v audite personalizačného driftu pri polarizujúcich témach na TikToku a v audite profilovania maloletých v reklamnom systéme TikToku.

Ako vstup pre klasifikáciu boli využité dostupné metadáta videí (opis, hashtagy, meno autora a pod.). Keďže metadáta videí často neposkytovali dostatok informácií, zvuková stopa videí bola dodatočne prepísaná do textu a výsledný prepis bol zahrnutý do analýzy založenej na veľkých jazykových modeloch. Okrem toho boli využité aj snímky obrazovky videí na analýzu vizuálnych charakteristík (napr. prítomnosť komerčného obsahu).

V prípade klasifikácie tém a postojov bol výkon klasifikátorov vyhodnotený na manuálne anotovanom datasete pozostávajúcom z 350 TikTok videí. Porovnávané boli viaceré LLM rôznych architektúr a veľkostí vrátane GPT-4o, GPT-4.1, LLaMA-3.1, Gemma-2 a Qwen-2.5. Najlepší výkon dosiahol model GPT-4.1, ktorý dosiahol vysokú presnosť klasifikácie pri všetkých hodnotených témach – konkrétne 98 %, 95 %, 98 % a 96,5 % pri klasifikácii tém a 100 %, 90 %, 98 % a 94 % pri klasifikácii postojov. Ďalšie vyhodnotenie ukázalo, že začlenenie automaticky generovaných prepisov zvuku zlepšilo výkon klasifikácie tém približne o 2 % a klasifikácie postojov približne o 6 %. Na základe týchto výsledkov bol model Whisper large-v3-turbo vybraný ako optimálne riešenie vzhľadom na priaznivý pomer medzi presnosťou a výpočtovou efektivitou.

V prípade detekcie typu a témy reklamy bol využitý model Qwen3-VL-4B-Instruct. Zo 113 manuálne anotovaných videí bola automatická anotácia typu reklamy správna pri 102 (90,3 %) a 100 (88,5 %) vzorkách podľa jednotlivých anotátorov. Zo 78/74 videí obsahujúcich akýkoľvek typ reklamy bola reklamná téma správne anotovaná pri 74 (94,9 %) a 63 (85,1 %) prípadoch.

Hlavným výsledkom týchto úloh boli klasifikátory obsahu založené na LLM/LVLM, ktoré dosahovali dostatočne vysokú presnosť bez potreby rozsiahlych anotovaných datasetov vyžadujúcich výrazné množstvo manuálnej práce. Tieto klasifikátory boli následne využité v ďalších úlohách a aktivitách projektu, kde bola potrebná automatická anotácia veľkých nazbieraných datasetov.

Viac informácií o natrénovaných klasifikátoroch obsahu je dostupných v *D4.1 Audit execution and evaluation methods* , kapitola 2.

WP3: Tvorba scenárov

V tomto pracovnom balíku skúmame rôzne stratégie pre interaktívnu tvorbu auditovacích scenárov s využitím cenných informácií získaných zo modelu sociálnych sietí. Hlavným výsledkom tohto pracovného balíka je komplexný súbor abstraktných auditovacích scenárov. Ich generovanie je riadené auditovacími otázkami, ktoré špecifikujú rozsah auditu, aspekty, ktorých vplyv na AI algoritmy má byť skúmaný, a metriky, ktoré majú byť merané.

T3.1 Derivácia auditovacích scenárov (M1-M17)

Úloha T3.1 sa zameriavala na vývoj metód na odvodenie abstraktných auditovacích scenárov, ktoré umožňujú systematickejšie, reprezentatívnejšie a opakovateľne použiteľné algoritmické audity. Predchádzajúce prístupy k algoritmickému auditovaniu sa vo veľkej miere spoliehali na manuálne navrhnuté scenáre vytvárané ad hoc výskumníkmi, čo často viedlo k neúplnému pokrytiu používateľov, obsahu a interakcií a limitovalo autentickosť a reprodukovateľnosť výsledkov auditov.

V rámci tejto úlohy bol zavedený koncept abstraktných auditovacích scenárov ako kľúčový prvok modelovo založeného algoritmického auditovania. Namiesto vopred definovaných individuálnych používateľských účtov a akcií špecifických pre konkrétnu platformu abstraktné scenáre špecifikujú opakovateľne použiteľné a od platformy nezávislé opisy cieľa auditu, používateľských populácií, interakčných protokolov, bodov pozorovania, metrík a jednotlivých fáz auditu.

Významným prínosom tejto úlohy je vývoj metód na návrh reprezentatívnych používateľských populácií a používateľských profilov. Používateľské populácie môžu byť buď manuálne definované audítorom (výskumníkom) na základe cieľa výskumu, alebo generované pomocou modelu sociálnych sietí vyvinutého v rámci T2.1. Generované populácie vychádzajú z rozdelení relevantných používateľských charakteristík vrátane veku, pohlavia, geografickej polohy a záujmov, čo umožňuje vytvárať štatisticky podložené sockpuppet profily namiesto náhodne vytvorených používateľských účtov.

Vyvinutý prístup zároveň podporuje definovanie modulárnych auditovacích fáz opisujúcich od platformy nezávislé používateľské akcie a umožňuje následný preklad abstraktných scenárov do vykonateľných auditovacích skriptov špecifických pre jednotlivé platformy. Táto abstrakcia umožňuje opakované využitie rovnakej metodiky auditu naprieč rôznymi platformami, a časovými obdobiami, čím zlepšuje porovnateľnosť a reprodukovateľnosť algoritmických auditov.

Vyvinutá metodika tvorby auditovacích scenárov bola overená prostredníctvom dvoch algoritmických auditovacích štúdií realizovaných v rámci projektu. Výsledky ukázali, že návrh auditovacích scenárov založený na modeloch umožňuje vytvárať reprezentatívnejšie používateľské populácie a autentickejšie simulácie používateľského správania v porovnaní s manuálne definovanými auditovacími scenármi využívanými v predchádzajúcich štúdiách.

Viac informácií je dostupných v *D3.1 Audit scenario co-creation methods* , kapitoly 2 a 3.

T3.2 Výber auditovaích scenárov (M12-M23)

Úloha T3.2 sa zameriavala na metódy výberu realizovateľných množín auditovacích scenárov pri zachovaní dostatočného pokrytia relevantných charakteristík používateľov a cieľov auditu. Keďže reprezentatívne používateľské populácie generované zo model sociálnych sietí môžu viesť k veľkému množstvu možných používateľských profilov a kombinácií scenárov, bolo potrebné vytvoriť systematické stratégie výberu, ktoré vyvažujú experimentálne pokrytie s praktickými obmedzeniami realizácie.

Vyvinutý prístup umožňuje upravovať rozdelenie simulovaných používateľov podľa demografických a behaviorálnych charakteristík s cieľom znížiť celkový počet potrebných auditovacích účtov pri zachovaní pokrytia relevantného priestoru používateľov. To umožňuje audítorom prispôsobiť rozsah experimentov technickým obmedzeniam, ako sú dostupná infraštruktúra a počet súčasne kontrolovaných účtov.

Prístup k výberu scenárov zároveň podporuje kontrolu potenciálnych skresľujúcich faktorov (angl. confounding factors), ktoré môžu ovplyvňovať správanie algoritmov. V závislosti od cieľa auditu môžu byť irelevantné charakteristiky buď neutralizované (napr. vynechaním vybraných demografických signálov), alebo reprezentované prostredníctvom dostatočne veľkých a vyvážených používateľských populácií s cieľom minimalizovať ich nežiaduce vplyvy.

Spolu s metódami odvodenia auditovacích scenárov vyvinutými v T3.1 prispieva prístup výberu scenárov k systematickejšiemu a škálovateľnejšiemu frameworku pre návrh algoritmických auditov. Zároveň umožňuje vytvárať reprezentatívne a technicky realizovateľné audity, ktoré možno opakovane vykonávať naprieč platformami a v priebehu času.

T3.3 Nástroj na spolutvorbu auditovacích scenárov (M1-M23)

Úloha T3.3 sa zameriavala na vývoj softvérovej podpory pre výskumníkov počas procesu spolutvorby auditovacích scenárov. S cieľom riešiť obmedzenia predchádzajúcich algoritmických auditov, pri ktorých boli auditovacie scenáre typicky navrhované manuálne a vyžadovali značnú mieru odborných znalostí a úsilia, projekt AI-Auditology vyvinul interaktívny nástroj označený ako *Control Panel* podporujúci tvorbu, konfiguráciu a správu abstraktných auditovacích scenárov.

Control Panel predstavuje hlavnú používateľskú komponentu infraštruktúry AI-Auditology a prepája jednotlivé prvky potrebné na návrh a realizáciu algoritmických auditov. Umožňuje výskumníkom definovať auditovacie štúdie, spravovať podporované platformy, špecifikovať auditované témy a simulované používateľské záujmy, konfigurovať umiestnenia proxy serverov, navrhovať reprezentatívne používateľské populácie a definovať vykonateľné auditovacie fázy. Vyvinutý nástroj podporuje návrh scenárov nezávislých od platformy a zároveň umožňuje rozšírenie na rôzne platformy sociálnych médií, pričom počiatočná podpora bola implementovaná pre TikTok a YouTube Shorts.

Kľúčovou funkcionalitou Control Panelu je Study Designer, ktorý vedie výskumníkov štruktúrovaným viackrokovým procesom transformácie auditovacej otázky na konkrétny abstraktný auditovací scenár. Dizajnér umožňuje špecifikovať cieľové platformy, geografické regióny, témy, vekové skupiny, distribúcie populácií a auditovacie fázy. Témy nie sú reprezentované iba svojim názvom, ale obsahujú aj ďalšie informácie potrebné na vykonanie auditu, napríklad vyhľadávacie frázy na objavovanie obsahu a štruktúrované prompty využívané prediktorom používateľských interakcií.

Vyvinutý nástroj zároveň podporuje generovanie individuálnych používateľských profilov. Výskumníci môžu definovať demografické a behaviorálne rozdelenia, na základe ktorých systém generuje príslušné sockpuppet profily s pridelenými charakteristikami, témami a proxy servermi. Tento prístup oddeľuje abstraktný návrh používateľskej populácie od vytvárania konkrétnych používateľských účtov, čím umožňuje systematickejšie a reprodukovateľnejšie nastavenie auditov.

Control Panel ďalej podporuje definovanie a správu auditovacích fáz opisujúcich sekvenciu a trvanie používateľských aktivít počas vykonávania auditu. Tieto scenáre tvoria základ pre transformáciu abstraktných návrhov auditov na vykonateľné auditovacie skripty využívané infraštruktúrou na vykonávanie auditov.

Viac informácií je dostupných v *D3.1 Audit scenario co-creation methods* , kapitola 3.

Vyvinutý nástroj na spolutvorbu auditovacích scenárov výrazne znižuje úsilie potrebné na návrh algoritmických auditov a poskytuje výskumníkom štruktúrované, reprodukovateľné a rozšíriteľné prostredie na vytváranie reprezentatívnych auditovacích scenárov. Spolu s modelom sociálnych sietí a infraštruktúrou na vykonávanie auditov prispieva k realizácii prístupu modelovo založeného algoritmického auditovania vyvinutého v rámci projektu.

WP4: Realizácia auditu

V tomto pracovnom balíku boli vyvinuté techniky na preklad, vykonávanie a vyhodnocovanie auditovacích scenárov.

T4.1 Preklad a vykonávanie auditovacích scenárov (M3-M23)

Úloha T4.1 sa zameriavala na skúmanie a implementáciu metód na preklad abstraktných auditovacích scenárov (výstup z T3.1 a T3.2) do vykonateľných auditovacích skriptov špecifických pre jednotlivé platformy. Keďže abstraktné auditovacie scenáre boli navrhnuté tak, aby boli nezávislé od platformy, času a obsahu, pred ich realizáciou na konkrétnych platformách sociálnych médií si vyžadovali dodatočné kroky prekladu.

Vyvinutý prístup transformuje hlavné prvky abstraktných auditovacích scenárov vrátane používateľských profilov, používateľských záujmov, mechanizmov výberu obsahu a typov interakcií do reprezentácií špecifických pre jednotlivé platformy. Používateľské profily definované v auditovacích scenároch sú konvertované na konkrétne používateľské účty na auditovaných platformách. Tento proces bol implementovaný ako poloautomatizovaný postup, keďže úplná automatizácia nebola možná vzhľadom na požiadavky špecifické pre jednotlivé platformy, ako sú overenie e-mailovej adresy alebo telefónneho čísla a CAPTCHA výzvy.

Proces prekladu riešil aj reprezentáciu používateľských záujmov a mechanizmov objavovania obsahu. V závislosti od možností auditovanej platformy boli abstraktné záujmy mapované na platformovo špecifické vyhľadávacie frázy alebo iné prístupy na objavovanie obsahu. V prípade platforiem podporujúcich vyhľadávanie obsahu na základe kľúčových slov boli vyhľadané príspevky analyzované prediktorom používateľských interakcií, ktorý určoval, či obsah zodpovedá záujmom simulovaného používateľa a či by sa mala vykonať interakcia. Dôležitým výsledkom tejto úlohy bola aj integrácia prediktorov používateľských interakcií (T2.2) do pracovného procesu vykonávania auditov.

Viac informácií je dostupných v *D4.1 Audit execution and evaluation methods* , kapitola 2.

Vyvinutý prístup prekladu a vykonávania umožnil transformovať abstraktné auditovacie scenáre na reprodukovateľné experimenty špecifické pre jednotlivé platformy. Kombináciou od platformy nezávislých definícií scenárov, automatizovanej analýzy obsahu a realistickejšej simulácie interakcií táto úloha prispela k zníženiu syntetickosti auditov založených na sockpuppet účtoch a k zlepšeniu ich škálovateľnosti a reprodukovateľnosti.

T4.2 Vyhodnotenie auditu (M7-M23)

Úloha T4.2 sa zameriavala na vyhodnotenie správania algoritmov na základe údajov zhromaždených počas vykonávania auditov (T4.1). Cieľom bolo analyzovať reakcie platforiem, merať vzorce vystavenia používateľov obsahu a kvantifikovať potenciálne účinky odporúčacích algoritmov na základe výsledkov získaných z realizovaných auditovacích scenárov.

Počas každej auditovacej štúdie zhromaždené údaje obsahovali informácie o simulovaných používateľoch, vykonaných interakciách, odporúčanom obsahu a dodatočných anotáciách pridelených počas alebo po vykonaní auditu. Informácie súvisiace s obsahom zahŕňali metadáta videí, ako sú opisy, autori, prepisy zvukových stôp, časové značky a charakteristiky videí, spolu s anotáciami vytvorenými prediktorom používateľských interakcií (T2.2) alebo následnou anotáciu po ukončení auditu (T2.3).

Zhromaždené datasety boli analyzované podľa konkrétnych výskumných otázok jednotlivých auditovacích štúdií. V závislosti od cieľa auditu boli použité rôzne štatistické prístupy. Pri analýzach zameraných na vývoj správania algoritmov v čase boli skúmané časové trendy pomocou regresnej analýzy a vizualizované prostredníctvom agregovaných časových reprezentácií. Pri analýzach porovnávajúcich výskyt konkrétnych kategórií obsahu alebo javov boli využité testy štatistickej významnosti na vyhodnotenie rozdielov medzi skúmanými podmienkami.

Vyhodnocovací framework využíval bežne používané štatistické a dátovo-analytické metriky vrátane priemernej miery výskytu konkrétnych typov obsahu, štandardných odchýlok a pomerov cieľového obsahu voči ostatnému pozorovanému obsahu. Výsledky boli vizualizované pomocou vhodných metód vrátane čiarových grafov, stĺpcových grafov alebo boxplotov v závislosti od charakteru analyzovaných údajov.

Keďže otázky algoritmického auditu sú výrazne závislé od kontextu, metodika vyhodnocovania bola prispôsobená špecifickým cieľom jednotlivých auditovacích štúdií realizovaných v rámci projektu. Tento prístup umožnil systematickú analýzu správania odporúčacích systémov a poskytol kvantitatívne dôkazy na hodnotenie vzorcov vystavenia obsahu a potenciálnych rizík spojených s algoritmickým rozhodovaním.

Viac informácií je dostupných v *D4.1 Audit execution and evaluation methods* , kapitola 2.

T4.3 Engine na vykonávanie auditov (M3-M23)

Úloha T4.3 sa zameriavala na vývoj enginu (softvérovej infraštruktúry) na vykonávanie auditov podporujúceho preklad, realizáciu a monitorovanie algoritmických auditov. Vyvinutá infraštruktúra poskytla potrebné možnosti na vykonávanie abstraktných auditovacích scenárov ako reprodukovateľných experimentov špecifických pre jednotlivé platformy s využitím simulovaných používateľov (sock puppets), ktorí interagujú s platformami sociálnych sietí.

Engine na vykonávanie auditov bol integrovaný s Control Panelom, ktorý slúžil ako hlavné rozhranie pre výskumníkov na monitorovanie vykonávania auditov a kontrolu zhromaždených pozorovaní. Control Panel spravoval konfigurácie auditov, odovzdával ich vykonávacej vrstve a poskytoval monitorovacie funkcionality vrátane prístupu k logom vykonávania, snímkam obrazovky a možnostiam opakovaného prehrania (angl. *replay*) správanie agenta. Vyvinutá infraštruktúra bola navrhnutá ako nezávislá od konkrétnej platformy a rozšíriteľná, pričom počiatočná podpora bola implementovaná pre TikTok a YouTube Shorts.

Základným komponentom vykonávacieho enginu bol framework používateľských agentov, ktorý riadil interakcie simulovaných používateľov vykonávané prostredníctvom webového prehliadača. Každý agent vykonával postupnosť interakcií špecifické pre platformu definované prostredníctvom *Action Domain-Specific Language* (DSL). DSL poskytuje štruktúrovanú reprezentáciu auditovacích scenárov vrátane zoradených krokov interakcií, podmienok, slučiek a runtime parametrov. Tento prístup umožňuje výskumníkom definovať opakovane použiteľné auditovacie skripty pri zachovaní oddelenia vykonávacej logiky od implementácie špecifickej pre jednotlivé platformy.

Infraštruktúra podporuje typické fázy auditu vrátane prihlásenia používateľa, inicializácie záujmov (angl. *interest seeding*) a hlavných interakčných fáz. Počas fázy inicializácie agenti vyhľadávali obsah súvisiaci s témami a využívali prediktor používateľských interakcií na určenie, či pozorovaný obsah zodpovedá záujmom simulovaného používateľa a aké interakcie by mali byť vykonané. Počas hlavnej auditovacej fázy agenti interagovali s odporúčacími systémami a priebežne vyhodnocovali zobrazovaný obsah podľa nakonfigurovaného auditovacieho scenára.

Na zvýšenie spoľahlivosti, reprodukovateľnosti a škálovateľnosti boli jednotlivé relácie prehliadača a vykonávania agentov izolované v kontajneroch Kubernetes typu pod. Vyvinutý framework agentov využíva priamu komunikáciu s prehliadačmi prostredníctvom *Chrome DevTools Protocol* (CDP) cez WebSocket namiesto konvenčných frameworkov na automatizáciu prehliadača. Tento prístup umožňuje presnejšiu kontrolu správania prehliadača, prístup k udalostiam prehliadača a sieťovej prevádzke a zároveň znižuje vystavenie bežným mechanizmom detekcie automatizácie.

Engine na vykonávanie auditov zároveň obsahuje rozsiahle možnosti monitorovania a zberu údajov. Počas realizácie auditu agenti vytvárali detailné logy, zhromažďovali snímky stavov prehliadača a ukladali štruktúrované informácie o pozorovanom obsahu, používateľských interakciách a priebehu vykonávania. Zhromaždené údaje zahŕňali výsledky vyhľadávania, metadáta videí, prepisy zvuku, vykonané interakcie a anotácie generované komponentmi na predikciu interakcií. Údaje boli ukladané v štruktúrovaných formátoch pre následnú analýzu a prenášané do perzistentného úložiska na ďalšie spracovanie.

Viac informácií je dostupných v *D4.1 Audit execution and evaluation methods* , kapitola 3.

Vyvinutý engine na vykonávanie auditov poskytuje kompletnú infraštruktúru na realizáciu reprodukovateľných, viacfázových algoritmických auditov vo veľkom rozsahu. Kombináciou konfigurovateľných auditovacích postupov, realistickej simulácie používateľov, možností vykonávania špecifických pre jednotlivé platformy a komplexného monitorovania umožňuje praktické nasadenie prístupu modelovo založeného algoritmického auditu vyvinutého v rámci projektu.

WP5: DCE a klastrovanie

Cieľmi tohto pracovného balíka boli: 1) dosiahnuť maximálny dopad prostredníctvom množiny cielených diseminačných aktivít; 2) vytvoriť a realizovať klastrovacie aktivity s relevantnými projektmi a iniciatívami; a 3) pripraviť výsledky projektu na ďalšie využitie a zabezpečiť ich dlhodobú udržateľnosť.

T5.1 Identifikácia potrieb používateľov a prípadov použitia (M1-M11)

Táto úloha bola ukončená už počas prvého roka implementácie projektu, v rámci ktorého sa uskutočnila séria osobných a online konzultácií s relevantnými zainteresovanými stranami s cieľom identifikovať potreby používateľov, očakávania a potenciálne prípady použitia algoritmického auditovania. Konzultácie zahŕňali diskusie s národným zástupcom implementácie Aktu o digitálnych službách (Digital Services Act - DSA), predstaviteľmi Európskej komisie zodpovednými za implementáciu DSA, zástupcami Európskeho centra pre algoritmickú transparentnosť (European Centre for Algorithmic Transparency - ECAT), Spoločného výskumného centra (Joint Research Centre - JRC) a Rady pre mediálne služby Slovenskej republiky (RPMS).

Okrem konzultácií so zainteresovanými stranami bola vykonaná podrobná analýza DSA auditovacích reportov publikovaných v 4. štvrťroku 2024. Analýza zahŕňala reporty piatich online platforiem (Very Large Online Platforms - VLOPs) – TikTok, YouTube, Instagram, Facebook a LinkedIn – a jedného online vyhľadávača (Very Large Online Search Engine - VLOSE) – Google Search. Analýza identifikovala významné obmedzenia súčasného ekosystému auditov, najmä pokiaľ ide o schopnosť audítorov nezávisle posudzovať súlad s povinnosťami zavedenými prostredníctvom DSA, ako sú požiadavky súvisiace s ochranou maloletých podľa článku 28 DSA.

Analýza potrieb zainteresovaných strán bola doplnená prehľadom aktuálneho vývoja výskumu v oblasti algoritmického auditovania a technickým posúdením platforiem sociálnych médií s cieľom vyhodnotiť realizovateľnosť a komplexnosť ich auditovania. Tieto aktivity poskytli dôležité vstupy pre strategické rozhodnutia počas celého projektu vrátane výberu auditovaných otázok, cieľových platforiem a metodologických prístupov. Zároveň prispeli k maximalizácii relevantnosti a potenciálneho dopadu vyvinutých prístupov pre rôzne skupiny zainteresovaných strán.

Výsledky tejto úlohy boli zdokumentované v *D5.1 User Needs and Use Cases* .

T5.2 Diseminácia, komunikácia a využitie výsledkov (M1-M23)

Na začiatku realizácie projektu bola vytvorená webová stránka projektu (https://kinit.sk/project/AI-Auditology), vizuálna identita projektu, diseminačný, komunikačný a exploitačný plán opísaný v dokumente D5.2 Dissemination, communication, exploitation and clustering plan, ako aj realizované iniciálne DCE aktivity. Počas druhého roka implementácie projektu boli následne realizované rozsiahle aktivity v oblasti diseminácie, komunikácie a využitia výsledkov v súlade so stratégiou definovanou v DCE pláne. Aktivity boli zamerané na zdieľanie výsledkov projektu s identifikovanými cieľovými skupinami vrátane vedeckej komunity, regulačných orgánov, tvorcov politík, mimovládnych organizácií, aktérov v oblasti informačných a komunikačných technológií (IKT) a širokej verejnosti.

Samostatná prezentácia projektu bola zabezpečená prostredníctvom webovej stránky KInITu vrátane profilu projektu, aktualít a blogových príspevkov, vybraných publikácií a projektových výstupov. Webová stránka slúžila ako centrálne miesto na komunikáciu priebehu projektu, výsledkov, vedeckých výstupov a verejne dostupných zdrojov. Webová stránka a súvisiaci obsah projektu dosiahli počas prvého a druhého roku implementácie 5 049 a 4 502 zobrazení, pričom prekročili stanovené kľúčové ukazovatele výkonnosti (KPI).

Komunikačné aktivity boli doplnené aktívnou prezentáciou projektu na sociálnych sieťach prostredníctvom etablovaných komunikačných kanálov KInITu na platformách LinkedIn, Facebook, X/Twitter (neskôr doplneného o Bluesky), ako aj prostredníctvom individuálnych účtov členov projektového tímu. Komunikácia bola prispôsobená rôznym cieľovým skupinám: komunikácia v slovenskom jazyku bola zameraná na širšiu verejnosť, zatiaľ čo komunikácia v anglickom jazyku primárne oslovovala výskumníkov, odborníkov z praxe, regulátorov, mimovládne organizácie a priemyselných partnerov. Komunikácia súvisiaca s projektom bola systematicky označovaná hashtagom #AIAuditology.

Výsledky projektu boli ďalej šírené prostredníctvom blogov, newsletterov, podcastov, rozhovorov pre média a spravodajských článkov. Vzhľadom na spoločenskú relevanciu výsledkov projektu tím dostal pozvania od národných aj medzinárodných médií vrátane Denníka N, SME, Süddeutsche Zeitung, Tages-Anzeiger a New Scientist. Vybrané výsledky boli komunikované aj vo viacerých jazykoch, čím sa zvýšil ich dosah nad rámec anglicky hovoriacej výskumnej komunity.

Projekt dosiahol významný vedecký impakt, v rámci ktorého 15 publikácií bolo prijatých a prezentovaných na popredných medzinárodných konferenciách, workshopoch a v časopisoch. Publikácie pokrývali viaceré výskumné oblasti relevantné pre projekt vrátane strojového učenia, spracovania prirodzeného jazyka, informačného vyhľadávania, interakcie človeka s počítačom, odporúčacích systémov a algoritmického auditovania. Viaceré publikácie boli prijaté na prestížnych fórach vrátane ACL, EMNLP, SIGIR, CHI, FAccT a časopisu ACM Transactions on Intelligent Systems and Technology (ACM TIST). Významným ocenením bolo získanie ceny za najlepší článok (angl. *Best Paper Award*) na konferencii FAccT 2026. V súlade s princípmi otvorenej vedy boli všetky publikácie sprístupnené prostredníctvom preprintov alebo otvorených verzií a sprievodné datasety a zdrojové kódy boli podľa možností zdieľané prostredníctvom verejných repozitárov.

Okrem akademickej diseminácie sa projektový tím aktívne zapájal do medzinárodných a národných podujatí s cieľom osloviť relevantné zainteresované strany. Tieto aktivity zahŕňali prezentácie a diskusie na podujatiach ako EU Disinfo Conference, AI Impact Summit, tematické semináre Rady Európy a národné podujatia zamerané na digitálne technológie a bezpečnosť mladých ľudí. Tieto aktivity prispeli k zvýšeniu povedomia o algoritmickom auditovaní ako praktickom prístupe na hodnotenie systémov sociálnych sietí využívajúcich umelú inteligenciu.

Špecifickou diseminačnou aktivitou bola príprava a vedenie projektu zameraného na auditovanie algoritmov sociálnych sietí v rámci letnej školy Digital Methods Summer School 2025 (https://www.digitalmethods.net/Dmi/SummerSchool2025AlgorithmicMirror) organizovanou na Univerzite v Amsterdame. Počas tejto letnej školy multidisciplinárny tím z účastníkov letnej školy pracoval s dátami zozbieraných počas doteraz nami vykonaných auditov a zrealizoval manuálny audit menšieho rozsahu, ktorý predstavuje hodnotný vstup pre ich následnú automatizáciu.

Projekt sa zároveň zameral na využitie a dlhodobú udržateľnosť (angl. *exploitation and sustainability*) svojich hlavných výsledkov (angl. *Key Exploitable Results - KERs*). Na základe konzultácií so zainteresovanými stranami a analýzy vykonanej v rámci D5.1 boli identifikované štyri hlavné cieľové skupiny využitia výsledkov: výskumná komunita, regulátori, aktéri v oblasti IKT a platformy sociálnych médií. Počas implementácie projektu bol najväčší potenciál využitia identifikovaný najmä vo výskumnej a regulačnej komunite.

Pre regulátorov a tvorcov politík boli výsledky projektu transformované do podoby regulačných odporúčaní podporujúcich lepší dohľad nad platformami využívajúcimi umelú inteligenciu. Projektový tím tiež spolupracoval so zástupcami Európskej komisie, Joint Research Centre, Európskeho centra pre algoritmickú transparentnosť (European Centre for Algorithmic Transparency), národných tímov implementácie DSA a ďalších relevantných organizácií.

Výsledky projektu vytvorili základ pre naše nadväzujúce výskumné aktivity a spolupráce. Vyvinuté metodiky, datasety, softvérová infraštruktúra a výskumné zistenia boli ďalej využité pri príprave viacerých nadväzujúcich projektových návrhov zameraných na algoritmické auditovanie, integritu volieb, online radikalizáciu, klimatické dezinformácie a digitálne mediálne ekosystémy. Tieto aktivity demonštrujú potenciál pokračujúceho využívania a ďalšieho rozvoja výsledkov AI-Auditology aj po ukončení projektu.

Projekt zároveň udržiaval priamu komunikáciu s platformami sociálnych médií vrátane TikToku, ktorý prejavil záujem o výsledky auditov, najmä v oblasti ochrany maloletých a personalizovaného poskytovania obsahu. Táto spolupráca ukázala význam nezávislých prístupov k algoritmickému auditu aj z pohľadu poskytovateľov platforiem usilujúcich sa o zvýšenie transparentnosti a zodpovednosti svojich AI systémov.

Viac informácií, ako aj podrobné zoznamy všetkých DCE výstupov (publikácie, príspevky na sociálnych sieťach, blogy, mediálne vystúpenia, návrhy nadväzujúcich projektov a pod.), sú dostupné v *D5.3 Dissemination, communication, exploitation and clustering report* .

T5.3 Klastrovanie s relevantnými iniciatívami (M1-M23)
Počas implementácie projektu sme aktívne rozvíjali a udržiavali spoluprácu s relevantnými iniciatívami, výskumnými projektami, regulačnými orgánmi a organizáciami pôsobiacimi v oblastiach algoritmickej transparentnosti, digitálnych platforiem, dezinformácií a správy AI (angl. AI governance). Tieto aktivity podporovali výmenu poznatkov, posilňovali dopad projektu a umožňovali zosúladenie výsledkov projektu so širšími európskymi výskumnými a regulačnými aktivitami.

Projekt vytvoril a udržiaval prepojenia s kľúčovými európskymi a národnými zainteresovanými stranami vrátane predstaviteľov Európskej komisie zodpovedných za implementáciu Aktu o digitálnych službách (DSA), Európskeho centra pre algoritmickú transparentnosť (ECAT), Spoločného výskumného centra (JRC) a Rady pre mediálne služby Slovenskej republiky (RPMS). Tieto interakcie poskytli cennú spätnú väzbu o praktických regulačných potrebách a pomohli identifikovať vhodné možnosti využitia metód algoritmického auditu.

AI-Auditology bolo tiež aktívne prepojené s relevantnými európskymi výskumnými iniciatívami, najmä projektmi Horizon Europe a Digital Europe zameranými na dezinformácie, dôveryhodnú AI a monitorovanie online platforiem. Projekt si vymieňal poznatky a hľadal synergie s projektmi ako vera.ai, VIGILANT, AI-CODE a ďalšími relevantnými projektmi zaoberajúcimi sa AI podporovanou analýzou škodlivého obsahu, online rizík a digitálnych mediálnych ekosystémov. Príkladom takejto spolupráce je spoločná publikácia s partnermi projektov vera.ai a AI-CODE o detekcii signálov kredibility publikovaná v ACM TIST (časopis v kvartile Q1).

Projektový tím ďalej prispieval ku klastrovacím aktivitám prostredníctvom účasti na relevantných konferenciách, workshopoch a networkingových podujatiach. Medzi tieto aktivity patrili napríklad EU Disinfo konferencie, DSA and Platform Regulation Conference, workshopy zamerané na férovosť AI a algoritmický audit a semináre Rady Európy venované digitálnemu prostrediu a rizikám pre deti a mládež. Tieto podujatia poskytli príležitosti na prezentáciu výsledkov projektu, výmenu skúseností s podobnými iniciatívami a identifikáciu možností budúcej spolupráce.

Prostredníctvom týchto aktivít AI-Auditology posilnil svoje postavenie v európskom ekosystéme iniciatív zameraných na transparentnosť, zodpovednosť a spoločenské dopady AI systémov využívaných na platformách sociálnych médií a zároveň zabezpečil, že výsledky projektu budú prispievať k pokračujúcim výskumným, regulačným a praktickým aktivitám aj po skončení projektu.

The whole period of the project implementation (from August 1, 2024 to June 30, 2026) was progressing in accordance with the project schedule. In the following text we provide an overview of activities conducted during this period within individual work packages and tasks.

WP1: Project management

The objective of WP1 focused on maintaining effective project management (administrative, financial, quality, and risk management); enabling smooth work progress and timely project reporting; as well as ethical and data management administration.

T1.1 Administrative, financial, and risk management (M1-M23)
Task T1.1 focused on the overall administrative, financial, and risk management of the project. Throughout the project, the project team maintained effective project communication procedures and collaboration tools, monitored project progress, resource allocation, and running costs, and ensured regular communication with the funding agency. The task also covered financial and legal administration, ensuring compliance with the project's reporting and funding requirements.

Another key responsibility of this task was the coordination of project reporting and documentation. This included the preparation and submission of project deliverables, the intermediate report, as well as the preparation of this final project report. Administrative coordination ensured that all planned activities, milestones, and deliverables were completed and submitted in accordance with the project schedule.

Risk management was carried out continuously throughout the project. Risks (initially identified in the project proposal) were regularly updated to identify, monitor, and mitigate potential internal, external, strategic, and operational risks (e.g., especially related to ethical and legal considerations, lack of data access provided by social media platforms, and dynamically evolving landscape of AI/LLMs and social media platforms themselves). The risk management process enabled the project team to proactively address emerging challenges, implement appropriate mitigation measures where necessary, and ensure the successful execution of the project in line with its objectives and timeline. Throughout the project, no major administrative, financial, or operational issues occurred that would have prevented the successful completion of the planned activities.

T1.2 Management of ethical issues (M1-M23)
Task T1.2 focused on ensuring that all project activities complied with the highest standards of AI and research ethics. Throughout the project, ethical considerations were systematically integrated into the research process to ensure that the developed algorithmic auditing methodologies, data collection procedures, and auditing studies respected the rights, privacy, and well-being of all potentially affected individuals and complied with applicable ethical and legal requirements.

During the first year of project implementation, the comprehensive ethical assessment was carried out. As part of this assessment, four ethics workshops were organized between 10 April and 12 June 2025, bringing together researchers with expertise in artificial intelligence, social sciences, philosophy, and law. The workshops critically examined the ethical and legal implications of the planned research activities, identified stakeholders potentially affected by the project, assessed possible risks arising from the proposed algorithmic auditing methodology, and formulated appropriate mitigation strategies.

During the second year of project implementation, the activities within this task concentrated on implementing and monitoring the mitigation measures defined in the Risk list with identified mitigation techniques (Appendix of this report). These measures were continuously applied and reviewed throughout the project to minimize ethical, legal, and societal risks associated with algorithmic auditing research. Particular attention was paid to ensuring that the research methodology remained consistent with the ethical principles approved by the internal ethics committee of KInIT, which granted ethical approval for the AI-Auditology project on 17 December 2024.

The continuous ethical oversight ensured that the project was conducted responsibly and transparently and that ethical considerations remained an integral part of all research activities from project initiation to completion.

WP2: Model construction

In WP2, we developed techniques to build, represent and infer from the social-media model (a key enabler of the model-based audits). Its inherent part consists of representation of a social media population – including demographic data (e.g., distribution of age, gender, location) and behavioral patterns (e.g., interests, number of sessions per day, length of sessions, distribution and frequency of different interaction types).

T2.1 Representation development and extraction (M1-M18)

Task T2.1 focused on the development of a unified, platform-agnostic representation of social media platforms, users, and interactions as a foundation for the social media model used within the project. Our implementation explicitly considered the limited availability of real-world social media data and restricted access to platform APIs (i.e., limited availability of information about individual users, content items, and their interactions). The developed approach therefore focused on capturing platform-level characteristics through a modular representation based on statistical and probability distributions.

As envisioned in the project proposal, the developed social media model adopted the architecture of overlay models, consisting of a domain layer and a platform layer. The domain layer represented platform-independent concepts and enumerations, including user demographic characteristics (such as age groups, gender categories, and geographical regions), content-related characteristics, and behavioural patterns (such as engagement rates, time spent on the platform, and usage frequency). The platform layer represented individual social media platforms through statistical distributions over the elements defined in the domain layer, allowing platform-specific characteristics to be captured while maintaining a unified representation across different platforms.

As part of this task, suitable data sources for constructing the social media model were identified and analysed. Significant effort was dedicated to obtaining access to the *TikTok Research API*. Despite extensive communication with TikTok support and multiple application attempts and appeals, the access provided by TikTok was limited to a *Virtual Compute Environment* (VCE) that did not provide the required type and scale of data. Therefore, the project team utilised alternative sources of information to support the development of the model. The analysis focused primarily on TikTok and YouTube, which were identified as relevant platforms within the project use cases (D5.1).

The utilised data sources included *CEDMO Trends*, a longitudinal public opinion survey providing insights into social media usage patterns, as well as research publications, platform documentation, transparency reports, and publicly available statistical sources. In particular, publicly available industry reports and marketing analyses provided valuable platform-level statistics describing demographic distributions, geographical reach, usage intensity, content preferences, and engagement patterns.

The extraction and harmonisation of information from the identified sources were performed in a semi-automatic manner. The extracted information was unified into platform-independent categories, including demographic characteristics (age groups, gender categories, and geographical regions) and behavioural characteristics (content preferences, engagement rates, average time spent, watch time, and session frequency). These unified representations enable the creation of platform-specific model configurations while preserving the desired platform-agnostic compatibility between different social media platforms.

For each supported platform, a dedicated platform layer was created, storing the corresponding demographic and behavioural distributions. The resulting model was implemented as a configurable framework in which platform characteristics could be updated by modifying configuration files without changes to the underlying implementation. This design enables continuous updates as new information sources become available while maintaining historical comparability between model versions.

To support the generation of representative audit scenarios, the developed model provided querying and synthetic user generation functionalities. User profiles can be generated by sampling demographic characteristics from platform-specific probability distributions and deriving corresponding behavioural patterns conditioned on these characteristics. The generated synthetic users are associated with demographic attributes, content preferences, engagement characteristics, estimated usage intensity, and interaction-related parameters, enabling reproducible creation of representative user populations for algorithmic auditing experiments.

More details about the developed social media model can be found in *D2.1 Harmful content modelling methods*, Section 2.

The developed social media model (D2.2) was further provided as a FastAPI-based service exposing REST API endpoints for accessing platform information, generating synthetic user profiles, and creating user cohorts based on specified constraints. The implementation separates configuration data, model logic, and API functionality, enabling maintainability and future extension with additional platforms or statistical indicators. The developed social media model therefore provides a flexible and extensible foundation for supporting algorithmic auditing activities by enabling the generation of statistically grounded and platform-comparable user representations.

T2.2 User interaction prediction (M7-M23)

Task T2.2 focused on developing methods for predicting and simulating user interactions with social media content in the context of algorithmic auditing. Within this task, user next interaction predictors were developed to automatically determine whether a simulated user should interact with a specific social media post and, if so, what type of interaction should be performed. The developed approach was designed to support realistic execution of algorithmic audits by automatically analysing recommended content in real time and generating user feedback signals required for audit scenario execution. Instead of relying on simple heuristics or predefined rules based on metadata such as hashtags (which are common in the past research works), the developed approach employed content classification based on Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) to analyse the relevance of recommended content with respect to predefined user characteristics (see T2.3 for more information).

During the audit execution, the predictor analysed the characteristics of the simulated user, available video metadata, and the content itself to determine whether the recommended item matched the user's interests. For content matching the user's interests, the simulated user performed stronger implicit and explicit feedback actions, including watching the video until completion, liking it, and bookmarking it. Other content was skipped, while livestreams and excessively long videos were excluded from interaction simulation. Furthermore, we were also investigating how even more realistic user simulation can be achieved with the recent advances in the domain of LLM-based user simulation.

For more information, please see *D2.1 Harmful content modelling methods*, Sections 3 and 4.

We can conclude that the developed user interaction prediction approach enables more realistic and reproducible simulation of user behaviour during algorithmic audits and provides an important component for executing large-scale audits of recommendation systems.

T2.3 Harmful content labelling (M1-M23)

Task T2.3 focused on developing and evaluating methods for automated content classification and labelling to support the analysis of social media content needed for user interaction prediction (T2.2) as well as audit evaluation (T4.2). The activities within this task addressed both the reduction of manual annotation requirements and the broader challenge of reliable content classification in domains where labelled data are often limited, costly to obtain, and unevenly distributed across languages and topics.

The project specifically investigated the application of machine learning and natural language processing approaches, with particular attention given to the use of Large Language Models (LLMs) and Large Vision-Language Model (LVLMs) for content analysis and classification.

At first, our research explored how LLM/LVLMs-based methods, prompting strategies, and synthetic data generation techniques could support classification tasks in low-resource settings, which represents a particularly important challenge in the context of social media analysis. Unlike traditional supervised classification approaches that require large manually annotated datasets, the investigated approaches aimed to improve model performance when only a limited amount of labelled data was available.

To this end as part of this task, several research studies were conducted to systematically analyse the behaviour, limitations, and opportunities of LLM-based approaches for low-resource content classification. The research examined the impact of randomness and methodological choices when learning from limited labelled datasets, investigated the trade-offs between LLM-based and established text augmentation techniques, and evaluated different strategies for selecting few-shot examples for LLM-based augmentation. Furthermore, the project compared specialised small language models with general-purpose large language models and investigated approaches for multilingual and low-resource language classification.

Second, following the findings from these studies, the required content classifiers were carefully trained in the low-resource settings. Specifically, they addressed these four classification tasks: topic classification, stance classification towards the detected topic, advertisement detection (including advertisement type classification), and advertisement topic classification. They were evaluated and applied in two algorithmic audit studies conducted within the project: an audit of personalization drift in polarising topics on TikTok and an audit of profiling of minors in TikTok's advertising delivery system.

As input to classification, we utilised available video metadata (description, hashtags, author name, etc.). Since video metadata often contained insufficient information, the audio track of videos was additionally transcribed, and the resulting transcript was incorporated into the LLM-based analysis. Furthermore, we also used video screenshots to analyse visual characteristics of the videos (e.g., the presence of commercial content).

In case of topic and stance classification, the performance of the classificators was evaluated using a manually annotated dataset of 350 TikTok videos. Several LLMs of different architectures and sizes were compared, including GPT-4o, GPT-4.1, LLaMA-3.1, Gemma-2, and Qwen-2.5. The best-performing model, GPT-4.1, achieved high classification accuracy across evaluated topics, reaching 98%, 95%, 98%, and 96.5% for topic classification and 100%, 90%, 98%, and 94% for stance classification, respectively. Additional evaluation demonstrated that incorporating automatically generated audio transcripts improved topic classification performance by approximately 2% and stance classification performance by approximately 6%. Based on these results, the Whisper large-v3-turbo model was selected as the optimal solution due to its favourable balance between accuracy and computational efficiency.
In case of advertisement type and topic detection, we utilised the Qwen3-VL-4B-Instruct model. Out of 113 manually annotated videos, the automatic annotated ad type was correct in 102 (90.3%) and 100 (88.5%) of samples respectively (according to individual annotators). Out of 78/74 videos containing any type of advertisement, the ad topic was correctly annotated in 74 (94.9%) and 63 (85.1%).

The main outcome of these tasks were LLM/VLLM-based content classifiers achieving high accuracy without a need of large-scale labelled datasets that would require extensive human annotation. They were consequently utilised across other tasks and project activities where the automatic annotation of large collected data samples was needed.

For more details about the trained content classifier, please, refer to *D4.1 Audit execution and evaluation methods* , Section 2.

WP3: Scenario creation

In this work package, we research various strategies for interactive creation of audit scenarios, while utilizing valuable information gathered in the social media model. The main outcome from this work package is a comprehensive set of abstract audit scenarios. The generation is driven by audit questions specifying the scope of the audit, aspects, whose effect on AI algorithms should be investigated and metrics that should be measured.

T3.1 Audit scenarios derivation (M1-M17)

Task T3.1 focused on developing methods for deriving abstract audit scenarios that enabled more systematic, representative, and reusable algorithmic audits. Previous algorithmic auditing approaches relied heavily on manually designed scenarios created ad hoc by researchers, which often resulted in incomplete coverage of relevant user, content, and interaction spaces and limited the authenticity and reproducibility of audit results.

Within this task, the concept of abstract audit scenarios was introduced as a key element of model-based algorithmic auditing. Instead of defining individual user accounts and platform-specific actions in advance, abstract scenarios specify reusable and platform-independent descriptions of the audit objective, user populations, interaction protocols, observation points, metrics, and audit phases.

A major contribution of this task is the development of methods for designing representative user populations and user profiles. User populations can be either manually defined by the auditor based on the research objective or generated using the social media model developed within T2.1. The generated populations are based on distributions of relevant user characteristics, including age, gender, location, and interests, allowing the creation of statistically grounded sockpuppet profiles instead of arbitrary user accounts.

The developed approach also supports the definition of modular audit phases describing platform-independent user actions and enables the subsequent translation of abstract scenarios into executable platform-specific audit procedures. This abstraction allows the same audit methodology to be reused across different platforms, and time periods, improving comparability and reproducibility of algorithmic audits.

The developed audit scenario co-creation methodology was validated through two algorithmic audit studies conducted within the project. The results demonstrated that model-based audit scenario design enabled more representative user populations and more authentic simulation of user behaviour compared to manually specified audit scenarios used in previous studies.

For more details, please, refer to *D3.1 Audit scenario co-creation methods* , Section 2 and 3.

T3.2 Audit scenarios selection (M12-M23)

Task T3.2 focused on methods for selecting feasible sets of audit scenarios while preserving sufficient coverage of relevant user characteristics and audit objectives. Since representative user populations generated from the social media model could result in large numbers of possible user profiles and scenario combinations, systematic selection strategies were required to balance experimental coverage with practical execution constraints.

The developed approach supports adjusting the distribution of simulated users across demographic and behavioural characteristics to reduce the overall number of required audit accounts while maintaining coverage of the relevant user space. This enabled auditors to adapt the scope of experiments according to technical constraints, such as available infrastructure and the number of concurrently managed accounts.

The scenario selection approach also supports controlling potential confounding factors that could influence algorithmic behaviour. Depending on the audit objective, irrelevant characteristics could either be neutralised (e.g., by omitting selected demographic signals) or represented through sufficiently large and balanced user populations to minimize their unintended effects.

Together with the audit scenario derivation methods developed in T3.1, the scenario selection approach contributes to a more systematic and scalable framework for designing algorithmic audits. It enables the creation of representative yet technically feasible audit experiments that can be repeatedly executed across platforms and over time.

T3.3 Audit scenario co-creation tool (M1-M23)

Task T3.3 focused on developing software support for researchers during the audit scenario co-creation process. To address the limitations of previous algorithmic audits, where audit scenarios were typically designed manually and required substantial expertise and effort, the AI-Auditology project developed an interactive tool entitled *Control Panel* that supports the creation, configuration, and management of abstract audit scenarios.

The Control Panel represents the main user-facing component of the AI-Auditology infrastructure and connects the different elements required for designing and executing algorithmic audits. It enables researchers to define audit studies, manage supported platforms, specify audited topics and simulated user interests, configure proxy locations, design representative user populations, and define executable audit phases. The developed tool supports platform-independent scenario design while remaining extensible to different social media platforms, with initial support for TikTok and YouTube Shorts.

A key functionality of the Control Panel is the Study Designer, which guides researchers through a structured multi-step process for transforming an audit question into a concrete abstract audit scenario. The designer enables the specification of target platforms, geographical regions, topics, age groups, population distributions, and audit phases. Topics are represented not only as labels but also include additional information required for audit execution, such as search queries for content discovery and structured prompts used by the user interaction predictor.

The developed tool also supports the generation of individual user profiles. Researchers can specify demographic and behavioural distributions, after which the system generates corresponding sockpuppet profiles with assigned characteristics, topics, and proxies. This approach separates the abstract population design from the creation of concrete user accounts, enabling more systematic and reproducible audit setup.

Furthermore, the Control Panel supports the definition and management of audit phases describing the sequence and duration of user activities during an audit execution. These scenarios provide the basis for translating abstract audit designs into executable audit scripts used by the audit execution infrastructure.

For more details, please, refer to *D3.1 Audit scenario co-creation methods* , Section 3.

The developed audit scenario co-creation tool significantly reduces the effort required to design algorithmic audits and provides researchers with a structured, reproducible, and extensible environment for creating representative audit scenarios. Together with the social media model and audit execution infrastructure, it contributes to the realization of the model-based algorithmic auditing approach developed within the project.

WP4: Audit execution

In this work package, we developed techniques for audit scenario translation, execution and evaluation.

T4.1 Audit scenarios translation and execution (M3-M23)

Task T4.1 focused on investigating and implementing methods for translating abstract audit scenarios (an outcome from T3.1, T3.2) into executable platform-specific audit procedures. Since abstract audit scenarios were designed to be platform-, time-, and content-independent, they required additional translation steps before they could be executed on specific social media platforms.

The developed approach translates the main elements of abstract audit scenarios, including user profiles, user interests, content selection mechanisms, and interaction types, into platform-specific representations. User profiles defined in the audit scenarios are converted into concrete user accounts on the audited platforms. This process was implemented as a semi-automated procedure, as full automation was not feasible due to platform-specific requirements such as email or phone verification and CAPTCHA challenges.

The translation process also addressed the representation of user interests and content discovery mechanisms. Depending on the capabilities of the audited platform, abstract interests were mapped to platform-specific search queries or other content discovery approaches. For platforms supporting keyword-based content search, retrieved content candidates were analysed by the user interaction predictor, which determined whether the content matched the simulated user's interests and whether an interaction should be performed. An important outcome of this task was also the integration of user interaction predictors (T2.2) into the audit execution workflow.

For more details, please, refer to *D4.1 Audit execution and evaluation methods* , Section 2.

The developed translation and execution approach enabled abstract audit scenarios to be transformed into reproducible platform-specific experiments. By combining platform-independent scenario definitions with automated content analysis and more realistic interaction simulation, the task contributed to reducing the artificiality of sockpuppet-based algorithmic audits and improving their scalability and reproducibility.

T4.2 Audit evaluation (M7-M23)

Task T4.2 focused on the evaluation of algorithmic behaviour based on data collected during audit execution (T4.1). The objective was to analyse platform responses, measure content exposure patterns, and quantify potential effects of recommendation algorithms based on the results obtained from executed audit scenarios.

During each audit study, the collected data included information about simulated users, performed interactions, recommended content, and additional annotations assigned during or after the audit execution. The collected content-related information included video metadata, such as descriptions, authors, transcripts, timestamps, and video characteristics, together with annotations produced by the user interaction predictor (T2.2) or subsequent post-audit annotation processes (T2.3).

The collected datasets were analysed according to the specific research questions addressed by individual audit studies. Depending on the objective of the audit, different statistical approaches were applied. For analyses focusing on the evolution of algorithmic behaviour over time, temporal trends were examined using regression analysis and visualised through aggregated time-based representations. For analyses comparing the prevalence of specific content categories or phenomena, statistical significance testing was applied to evaluate differences between examined conditions.

The evaluation framework used commonly applied statistical and data analysis metrics, including average occurrence rates of specific content types, standard deviations, and ratios of targeted content compared to other observed content. The results were visualised using appropriate methods, including line plots, bar plots, box plots, and stacked representations, depending on the characteristics of the analysed data.

As algorithmic auditing questions are highly context-dependent, the evaluation methodology was adapted to the specific objectives of each audit study conducted within the project. This approach enabled systematic analysis of recommendation system behaviour and provided quantitative evidence for assessing content exposure patterns and potential risks associated with algorithmic decision-making.

For more details, please, refer to *D4.1 Audit execution and evaluation methods* , Section 2.

T4.3 Audit execution engine (M3-M23)

Task T4.3 focused on the development of a high-fidelity audit execution engine supporting the translation, execution, and monitoring of algorithmic audits. The developed infrastructure provided the necessary capabilities to execute abstract audit scenarios as reproducible platform-specific experiments using simulated users (sock puppets) interacting with social media platforms.

The audit execution engine was integrated with the Control Panel, which served as the main interface for researchers to monitor audit execution, and review collected observations. The Control Panel maintained audit configurations, transferred them to the execution layer, and provided monitoring functionalities, including access to execution logs, screenshots, and replay capabilities. The developed infrastructure was designed to be platform-agnostic and extensible, with initial support for TikTok and YouTube Shorts.

The core component of the execution engine is the user agent framework, which controls browser-based interactions performed by simulated users. Each agent executes platform-specific workflows defined through an *Action Domain-Specific Language* (DSL). The DSL provides a structured representation of audit scenarios, including ordered interaction steps, conditions, loops, and runtime parameters. This approach enables researchers to define reusable audit workflows while keeping the execution logic separated from platform-specific implementations.

The execution workflow supports typical audit phases, including user login, interest seeding, and main interaction phases. During the seeding phase, agents searched for topic-related content and used the user interaction predictor to determine whether observed content matched the simulated user's interests and which interactions should be performed. During the main audit phase, agents interacted with recommendation systems and continuously evaluated presented content according to the configured audit scenario.

To improve reliability, reproducibility, and scalability, individual browser sessions and agent executions were isolated in Kubernetes pod containers. The developed agent framework relied on direct communication with browsers through *Chrome DevTools Protocol* (CDP) over WebSocket instead of conventional browser automation frameworks. This approach enables more precise control over browser behaviour, access to media events and network traffic, and reduces exposure to common automation detection mechanisms.

The execution engine also incorporates extensive monitoring and data collection capabilities. During audit execution, agents generated detailed logs, collected screenshots of browser states, and stored structured information about observed content, user interactions, and execution progress. Collected data included search results, video metadata, transcripts, performed interactions, and annotations generated by the interaction prediction components. Data were stored in structured formats for subsequent analysis and transferred to persistent storage for further processing.

For more details, please, refer to *D4.1 Audit execution and evaluation methods* , Section 3.

The developed audit execution engine provides a complete infrastructure for running reproducible, multi-phase algorithmic audits at scale. By combining configurable audit workflows, realistic user simulation, platform-specific execution capabilities, and comprehensive monitoring, it enables the practical deployment of the model-based algorithmic auditing approach developed within the project.

WP5: DCE and clustering

The objectives of this work package were to: 1) achieve maximum impact with a set of targeted, high quality dissemination activities; 2) setting up and carrying out clustering activities with related projects and activities; and 3) preparing for exploitation and results usage and making the project outcomes sustainable long-term.

T5.1 User needs and use cases identification (M1-M11)

This task was completed already during the first year of project implementation, during which a series of in-person and online consultations with relevant stakeholders were conducted to identify user needs, expectations, and potential use cases for algorithmic auditing. These consultations included discussions with the national representative involved in the implementation of the Digital Services Act (DSA), representatives of the European Commission responsible for DSA implementation, representatives of the European Centre for Algorithmic Transparency (ECAT), the Joint Research Centre (JRC), and the Slovak Council for Media Services (RPMS).

In addition to stakeholder consultations, an in-depth analysis of DSA audit reports published in Q4 2024 was conducted. The analysis covered reports from five Very Large Online Platforms (VLOPs) (TikTok, YouTube, Instagram, Facebook, and LinkedIn) and one Very Large Online Search Engine (VLOSE) (Google Search). The analysis identified significant limitations in the current audit ecosystem, particularly regarding the ability of auditors to independently assess compliance with obligations introduced by the DSA, such as requirements related to the protection of minors under Article 28 DSA.

The stakeholder needs analysis was complemented by a review of recent research developments in algorithmic auditing and a technical assessment of social media platforms to evaluate the feasibility and complexity of auditing different platforms. These complementary activities provided important inputs for strategic decisions throughout the project, including the selection of audit questions, target platforms, and methodological approaches. They also contributed to maximizing the relevance and potential impact of the developed auditing approaches for different stakeholder groups.

The outcomes of this task were documented in deliverable *D5.1 User Needs and Use Cases* .

T5.2 Dissemination, communication and exploitation (M1-M23)

At the beginning of project implementation, we created a project website (https://kinit.sk/project/AI-Auditology), a visual identity, and a DCE plan (Dissemination, Communication, Exploitation), described in deliverable D5.2 Dissemination, communication, exploitation and clustering plan, and began executing initial DCE activities. During the second year of project implementation, extensive dissemination, communication, and exploitation activities were conducted following the dissemination and communication strategy defined in DCE And clustering plan (D5.2). The activities focused on sharing the project outcomes with the identified target groups, including the scientific community, regulators, policy makers, watchdog organisations, NGOs, ICT stakeholders, and the general public.

A dedicated project presence was done through the KInIT website, including a project profile, news and blog posts, selected publications, and project deliverables. The project website served as the central repository for communicating project progress, outcomes, scientific results, and publicly available resources. The website and related project content achieved 5049 and 4502 impressions during the first and second years of implementation, respectively, exceeding the predefined KPIs.

The communication activities were complemented by an active social media presence through established KInIT communication channels on LinkedIn, Facebook, X/Twitter (later complemented by Bluesky), as well as through individual accounts of project team members. Communication was tailored to different audiences: Slovak-language communication targeted the broader public, while English-language communication primarily addressed researchers, practitioners, regulators, NGOs, and industry stakeholders. Project-related communication was consistently promoted under the dedicated hashtag #AIAuditology.

The project outcomes were further disseminated through blogs, newsletters, podcasts, media interviews, and news articles. Due to the societal relevance of the project results, the team received invitations from both national and international media outlets, including Denník N, SME, Süddeutsche Zeitung, Tages-Anzeiger, and New Scientist. Selected results were also communicated in multiple languages, increasing their reach beyond the English-speaking research community.

The project achieved significant scientific impact, resulting in 15 publications accepted at leading international conferences, workshops, and journals. The publications covered multiple research areas relevant to the project, including machine learning, natural language processing, information retrieval, human-computer interaction, recommender systems, and algorithmic auditing. Several publications targeted top-tier venues, including ACL, EMNLP, SIGIR, CHI, FAccT, and ACM Transactions on Intelligent Systems and Technology. A notable recognition was the Best Paper Award received at FAccT 2026. In line with open science principles, all publications were made openly available through preprints or open-access versions, and accompanying datasets and source code were shared whenever possible through public repositories.

Beyond academic dissemination, the project team actively participated in international and national events to engage with relevant stakeholders. These activities included presentations and discussions at events such as EU Disinfo conferences, the AI Impact Summit, Council of Europe thematic seminars, and national events focused on digital technologies and youth safety. These activities contributed to raising awareness of algorithmic auditing as a practical approach for evaluating AI-driven social media systems.

A notable DCE activity was the preparation and facilitation of a project for the Digital Methods Summer School 2025 (https://www.digitalmethods.net/Dmi/SummerSchool2025AlgorithmicMirror) organized at the University of Amsterdam, where a multidisciplinary team consisting of summer school participants (including representatives of the AI-Auditilogy project) worked with data from our previous audits and conducted a small-scale manual audit, which represents a valuable input for its consequent automatization.

The project also focused on exploitation and long-term sustainability of its key exploitable results (KERs). Based on stakeholder consultations and the analysis performed within D5.1, four main exploitation target groups were identified: the research community, regulators and watchdog organisations, ICT stakeholders, and social media platforms. During project implementation, the strongest exploitation potential was identified particularly among the research and regulatory communities.

For regulators and policy stakeholders, project results were transformed into policy recommendations, methodological insights, and evidence-based analyses supporting improved oversight of AI-powered platforms. The project team engaged with representatives of the European Commission, Joint Research Centre, European Centre for Algorithmic Transparency, national DSA implementation teams, and other relevant organisations.

The project results established also a foundation for our follow-up research activities and collaborations. The developed methodologies, datasets, software infrastructure, and research findings were further utilized in preparation of several follow-up project proposals addressing algorithmic auditing, electoral integrity, online radicalisation, climate disinformation, and digital media ecosystems. These activities demonstrate the potential for continued exploitation and further development of the AI-Auditology results beyond the project duration.

The project also maintained direct communication with social media platforms, including TikTok, which expressed interest in audit findings, particularly those related to the protection of minors and personalised content delivery. This engagement demonstrated the relevance of independent algorithmic auditing approaches also from the perspective of platform providers seeking to improve transparency and accountability of their AI systems.

For more information as well as detailed lists of all DCE outcomes (publications, social media posts, blogs, media appearances, follow-up project proposals, etc.), please, refer to *D5.3 Dissemination, communication, exploitation and clustering report* .

T5.3 Clustering with relevant initiatives (M1-M23)

Throughout the project implementation, AI-Auditology actively developed and maintained cooperation with relevant initiatives, research projects, regulatory bodies, and stakeholder organisations working on algorithmic transparency, digital platforms, disinformation, and AI governance. These activities supported knowledge exchange, strengthened the project's impact, and enabled the alignment of project outcomes with broader European research and policy activities.

The project established and maintained connections with key European and national stakeholders, including representatives of the European Commission responsible for Digital Services Act (DSA) implementation, the European Centre for Algorithmic Transparency (ECAT), the Joint Research Centre (JRC), and the Slovak Council for Media Services (RPMS). These interactions provided valuable feedback on practical regulatory needs and helped identify relevant exploitation pathways for algorithmic auditing methods.

AI-Auditology was also actively connected with related European research initiatives, particularly Horizon Europe and Digital Europe projects focusing on disinformation, trustworthy AI, and online platform governance. The project exchanged knowledge and explored synergies with initiatives such as vera.ai, VIGILANT, AI-CODE, and other relevant projects addressing AI-supported analysis of harmful content, online risks, and digital media ecosystems. One particular example is a joint publication with vera.ai and AI-CODE partners on credibility signals detection published in ACM TIST (Q1 journal).

The project team further contributed to clustering activities through participation in relevant conferences, workshops, and networking events. These included events such as EU Disinfo conferences, the DSA and Platform Regulation Conference, workshops focused on AI fairness and algorithmic auditing, and Council of Europe seminars addressing digital environments and risks for children and youth. Such events provided opportunities to present project results, exchange experiences with related initiatives, and identify opportunities for future collaboration.

Through these activities, AI-Auditology strengthened its position within the European ecosystem of initiatives addressing transparency, accountability, and societal impacts of AI-powered social media platforms, while ensuring that project results contribute to ongoing research, regulatory, and practical efforts beyond the project lifetime.

2. Míľniky

Štruktúra kapitoly pre projekty rozdelené na pracovné balíky

V tejto časti sa uvádzajú informácie o stave dosahovania všetkých míľnikov uvedených v prílohe č. 2 ZoPPM, a to nasledovne:
Poradové číslo a názov míľnika: Uvádzajú sa samostatne všetky míľniky v poradí, ako sú uvedené v prílohe č. 2 ZoPPM.
Typ a číslo monitorovacej správy: Uvádza sa typ ("priebežná monitorovacia správa" alebo "záverečná monitorovacia správa") a číslo tej monitorovacej správy, ktorá pokrýva monitorovacie obdobie, v ktorom bol míľnik úplne dosiahnutý, inak toto pole zostáva prázdne.
Popis dosiahnutia míľnika: Popis sa uvádza iba v prípade, ak bol míľnik dosiahnutý (úplne alebo čiastočne) počas aktuálneho monitorovacieho obdobia, inak ostáva pole prázdne. Uvedie sa stručný popis, ako bol míľnik dosiahnutý a názov dokumentu, ktorým prijímateľ dosiahnutie míľnika preukazuje. Prijímateľ zároveň uvádza, či je daný dokument prílohou tejto monitorovacej správy alebo ŽoP. V prípade, že je prílohou ŽoP, je potrebné uviesť typ a číslo ŽoP, prílohou ktorej daný dokument je. V prípade, že míľnik nebol ku dňu ukončenia vecnej realizácie projektu úplne dosiahnutý, uvedie sa jeho stav ku dňu ukončenia vecnej realizácie, dôvody jeho nenaplnenia a v prípade, že sa jeho stav po ukončení vecnej realizácie zmenil, uvedie sa aj aktuálny stav ku dňu vypracovania tejto monitorovacej správy a dokument preukazujúci aktuálny stav míľnika (ak je stav plnenia míľnika možné preukázať).

Poradové číslo a názov míľnika	Typ a číslo monitorovacej správy	Popis dosiahnutia míľnika	Popis dosiahnutia míľnika
1. MS1 Začiatok projektu a prvotné výsledky výskumu (M6)	priebežná, 0001	NV1, NV2, NV5	**Míľnik dosiahnutý úplne** Nastavené procesy riadenia projektu, Výskum metód klasifikácie textového obsahu s obmedzeným množstvo označkovaných dát, Štúdia reprodukovateľnosti auditov personalizačných faktorov na sociálnej sieti TikTok Názov dokumentov: D5.2 Plán diseminácie, komunikácie, exploitácie a klastrovania (príloha tejto monitorovacej správy), Schválenie etickou komisiou Kempelenovho inštitútu inteligentných technológií (príloha tejto monitorovacej správy), Výskumný článok prijatý na EMNLP 2024, Výskumný článok prijatý na SIGIR 2025
2. MS2 Míľnik v polovici trvania projektu (M11)	priebežná, 0001	NV1 - NV5	**Míľnik dosiahnutý úplne** Analýza potrieb a prípadov použitia stakeholderov, Prototyp výskumnej infraštruktúry, Rozpracovaný a opublikovaný opis novej paradigmy modelom riadeného algoritmického auditovania, Prvé prototypy metód klasifikácie obsahu, augmentácie dát pomocou veľkých jazykových modelov, predikcie používateľských interakcií Názov dokumentov: D1.1 Priebežná správa o vykonávaní a dosiahnutých výsledkoch projektu (príloha tejto monitorovacej správy), D5.1 Potreby používateľov a prípady použitia (príloha tejto monitorovacej správy), Zoznam rizík s identifikovanými mitigačnými technikami (príloha tejto monitorovacej správy), Výskumný článok prijatý na EWAF 2025, Výskumný článok prijatý na ACL 2025, Výskumný článok prijatý na NAACL 2025, Opis projektu a tutoriálu pre Digital Methods Summer School 2025 (príloha tejto monitorovacej správy)

3. MS3 Modelovanie škodlivého obsahu a tvorba auditovacích scenárov dokončená (M18)	záverečná, 0002	NV2, NV3, NV5	**Míľnik dosiahnutý úplne** Preskúmané metódy klasifikácie obsahu (textu), metódy detekcie signálov kredibility, metódy predikcie interakcií používateľov a metódy modelovania sociálnych sietí. Model sociálnych médií bol nasadený a verejne sprístupnený. Preskúmané a použité metódy spoluvytvárania audítorských scenárov. Názov dokumentov: D2.1 Metódy modelovania škodlivého obsahu (príloha tejto monitorovacej správy), D2.2 Model škodlivého obsahu, 3x výskumný článok prijatý na EMNLP 2025 (link, link, link), Výskumný článok prijatý do ACM TIST
4. MS4 Záverečný míľnik (M23)	záverečná, 0002	NV1 - NV5	**Míľnik dosiahnutý úplne** Bol pokrytý celý rozsah projektu. Boli zrealizované tri auditovacie štúdie a vykonané aktivity v oblasti DCE a klastrovania. Názov dokumentov: D1.2 Záverečná správa o realizácii a výsledkoch projektu (tento dokument), D3.1 Metódy spoluvytvárania scenára auditu (príloha tejto monitorovacej správy), D4.1 Metódy vykonávania a hodnotenia auditu (príloha tejto monitorovacej správy), D5.3 Správa o diseminácií, komunikácii, exploitácie a klastrovania (príloha tejto monitorovacej správy), 2x výskumný článok prijatý na EACL 2026 (link, link), Výskumný článok prijatý na CHI 2026, Výskumný článok prijatý na UMAP 2026, Ocenený výskumný článok prijatý na FAccT 2026, Výskumný článok prijatý na IS2 2026

1. MS1 Project initiation and early research results (M6)	interim, 0001	NV1, NV2, NV5	**Milestone fully achieved** Established project management processes, research on methods for classifying textual content with a limited amount of labeled data, reproducibility study of audits of personalization factors on the TikTok social network. Document titles: D5.2 Dissemination, Communication, Exploitation and Clustering Plan (annex of this monitoring report), Approval by Ethics committee of the Kempelen institute of intelligent technologies (annex of this monitoring report), Research paper accepted at EMNLP 2024, Research paper accepted at SIGIR 2025

2. MS2 Midterm milestone (M11)	interim, 0001	NV1 - NV5	**Milestone fully achieved** Analysis of stakeholder needs and use cases, prototype of the research infrastructure, developed and published description of a new paradigm of model-driven algorithmic auditing, first prototypes of methods for content classification, data augmentation using large language models, and prediction of user interactions. Document titles: D1.1 Interim report on the implementation and achievements of the project (annex of this monitoring report), D5.1 User needs and use cases (annex of this monitoring report), Risk list with identified mitigation techniques (annex of this monitoring report), Research paper accepted at EWAF 2025, Research paper accepted at ACL 2025, Research paper accepted at NAACL 2025, Project and tutorial description for Digital Methods Summer School 2025 (annex of this monitoring report)
3. MS3 Harmful content model and scenario creation finished (M18)	final, 0002	NV2, NV3, NV5	**Milestone fully achieved** Content (text) classification methods, credibility signals detection methods, social media modelling and user interaction prediction methods researched. Social media model deployed and publicly shared. Audit scenario co-creation methods researched and used. Document titles: D2.1 Harmful content modelling methods (annex of this monitoring report), D2.2 Harmful content model, 3x Research paper accepted at EMNLP 2025 (link, link, link), Research paper accepted to ACM TIST
4. MS4 Final milestone (M23)	final, 0002	NV1 - NV5	**Milestone fully achieved** Whole scope of the project covered. 3 audit studies conducted. DCE and clustering activities delivered. Document titles: D1.2 Final report on the implementation and achievements of the project (annex of this monitoring report), D3.1 Audit scenario co-creation methods (annex of this monitoring report), D4.1 Audit execution and evaluation methods (annex of this monitoring report), D5.3 Dissemination, communication, exploitation and clustering report (annex of this monitoring report), 2x Research paper accepted at EACL 2026 (link, link) Research paper accepted at CHI 2026, Research paper accepted at UMAP 2026, Award-winning research paper accepted at FAccT 2026, Research paper accepted at IS2 2026

3. Výstupy projektu

Štruktúra kapitoly pre projekty rozdelené na pracovné balíky

V tejto časti sa uvádzajú informácie o stave dosahovania všetkých výstupov projektu uvedených v prílohe č. 2 ZoPPM, a to nasledovne:
Poradové číslo a názov výstupu: Uvádzajú sa samostatne všetky výstupy projektu v poradí, ako sú uvedené v prílohe č. 2 ZoPPM.
Typ a číslo monitorovacej správy: Uvádza sa typ ("priebežná monitorovacia správa" alebo "záverečná monitorovacia správa") a číslo tej monitorovacej správy, ktorá pokrýva monitorovacie obdobie, v ktorom bol výstup projektu úplne dosiahnutý, inak toto pole zostáva prázdne.
Popis dosiahnutia výstupu: Popis sa uvádza iba v prípade, ak bol výstup projektu dosiahnutý (úplne alebo čiastočne) počas aktuálneho monitorovacieho obdobia, inak ostáva pole prázdne. Uvedie sa stručný popis, ako bol výstup projektu dosiahnutý a názov dokumentu, ktorým prijímateľ úplne dosiahnutie výstupu projektu preukazuje. Prijímateľ zároveň uvádza, či je daný dokument prílohou tejto monitorovacej správy alebo ŽoP. V prípade, že je prílohou ŽoP, je potrebné uviesť typ a číslo ŽoP, prílohou ktorej daný dokument je. Dokument preukazujúci úplné dosiahnutie výstupu musí obsahovať aj splnenie požiadaviek na povinnú publicitu v zmysle VZP.V prípade, že výstup projektu nebol ku dňu ukončenia vecnej realizácie projektu úplne dosiahnutý, uvedie sa jeho stav ku dňu ukončenia vecnej realizácie, dôvody jeho nedosiahnutia a v prípade, že sa jeho stav po ukončení vecnej realizácie zmenil, uvedie sa aj aktuálny stav ku dňu vypracovania tejto monitorovacej správy a dokument preukazujúci aktuálny stav výstupu projektu (ak je stav výstupu projektu možné preukázať).
Dokumenty preukazujúce úplné dosiahnutie výstupu alebo aktuálny stav výstupu musia obsahovať aj splnenie požiadaviek na povinnú publicitu v zmysle VZP.

Poradové číslo názov výstupu	Typ a číslo monitorovacej správy	Názov pracovného balíka/pracovných balíkov	Popis dosiahnutia výstupu
1. D1.1 Priebežná správa o vykonávaní a dosiahnutých výsledkoch projektu (M11)	priebežná, 0001	NV1	**Výstup dosiahnutý úplne** Výstupom je predchádzajúca priebežná monitorovacia správa (0001) dokumentujúca stav výskumných a inovačných aktivít na konci mesiaca M11. Názov dokumentu: D1.1 Priebežná monitorovacia správa (príloha tejto monitorovacej správy, zverejnené na webe projektu)
2. D1.2 Záverečná správa o realizácii a výsledkoch projektu (M23)	záverečná, 0002	NV1	**Výstup dosiahnutý úplne** Výstupom je táto finálna monitorovacia správa (0002) dokumentujúca stav výskumných a inovačných aktivít na konci projektu. (zverejnené na webe projektu)
3. D2.1 Metódy modelovania škodlivého obsahu (M18)	záverečná, 0002	NV2	**Výstup dosiahnutý úplne** Výstup pozostáva z reportu dokumentujúceho zvolený prístup k vytvoreniu modelu sociálnej siete na základe dostupných zdrojov údajov. Názov dokumentu: D2.1 Metódy modelovania škodlivého obsahu (príloha tejto monitorovacej správy, zverejnené na webe projektu)
4. D2.2 Model škodlivého obsahu (M18)	záverečná, 0002	NV2	**Výstup dosiahnutý úplne** Výstupom je Github repozitár obsahujúci zdrojový kód modelu sociálnej siete a rozhranie na vytváranie populácie používateľov.
5. D3.1 Metódy spoluvytvárania scenára auditu (M23)	záverečná, 0002	NV3	**Výstup dosiahnutý úplne** Výstup pozostáva z reportu dokumentujúceho metódy a experimenty zamerané na odvodenie a výber auditovacích scenárov a z nástroja na spoluvytváranie auditovaích scenárov, ktorý poskytuje používateľské rozhranie a podporu pre audítora. Názov dokumentu: D3.1 Metódy spoluvytvárania scenára auditu (príloha tejto monitorovacej správy, zverejnené na webe projektu)
6. D4.1 Metódy vykonávania a hodnotenia auditu (M23)	záverečná, 0002	NV4	**Výstup dosiahnutý úplne** Výstup pozostáva z reportu dokumentujúceho metódy a experimenty zamerané na preklad, vykonávanie a vyhodnocovanie audítorských scenárov a z nástroja na vykonávanie auditov, ktorý poskytuje potrebné používateľské rozhranie a podporu pre audítora. Názov dokumentu: D4.1 Metódy vykonávania a hodnotenia auditu (príloha tejto monitorovacej správy, zverejnené na webe projektu)

7. D5.1 Potreby používateľov a prípady použitia (M11)	priebežná, 0001	NV5	**Výstup dosiahnutý úplne** Výstup pozostáva z reportu dokumentujúceho potreby stakeholderov a prípady použitia získané analýzou DSA auditov sociálnych sietí, kvalitatívnou analýzou so stakeholdermi, kvantitatívnou analýzou reprodukovateľnosti súčasnej generácie auditov a scríningom auditovateľnosti sociálnych sietí. Názov dokumentu: D5.1 Potreby používateľov a prípady použitia (príloha tejto monitorovacej správy, zverejnené na webe projektu)
8. D5.2 Plán diseminácie, komunikácie, exploitácie a klastrovania (M6)	priebežná, 0001	NV5	**Výstup dosiahnutý úplne** Výstup pozostáva z reportu dokumentujúceho plán diseminačných, komunikačných, exploitačných a klustrovacích aktivít spolu s identifikáciou stakeholderov a príkladmi konkrétnych krokov k relizácií týchto aktivít. Názov dokumentu: D5.2 Plán diseminácie, komunikácie, exploitácie a klastrovania (príloha tejto monitorovacej správy, zverejnené na webe projektu)
9. D5.3 Správa o diseminácií, komunikácii, exploitácie a klastrovania (M23)	záverečná, 0002	NV5	**Výstup dosiahnutý úplne** Výstup pozostáva z reportu dokumentujúceho plán diseminačných, komunikačných, exploatačných a klastrovacích aktivít spolu s identifikáciou zainteresovaných strán a príkladmi konkrétnych krokov smerujúcich k realizácii týchto aktivít. Názov dokumentu: D5.3 Správa o diseminácií, komunikácii, exploitácie a klastrovania (príloha tejto monitorovacej správy, zverejnené na webe projektu)

1. D1.1 Interim report on the implementation and achievements of the project (M11)	interim, 0001	NV1	**Output fully achieved** The output is the preceeding interim monitoring report (0001) documenting the status of research and innovation activities at the end of Month 11. Document title: D1.1 Interim report on the implementation and achievements of the project (annex of this monitoring report, publicly available at the project website)
2. D1.2 Final report on the implementation and achievements of the project (M23)	final, 0002	NV1	**Output fully achieved** The output is this final monitoring report (0002) documenting the status of research and innovation activities at the end of the project. (publicly available at the project website)
3. D2.1 Harmful content modelling methods (M18)	final, 0002	NV2	**Output fully achieved** The output consists of a report documenting the undertaken approach on how a social media model is created from the available data sources. Document title: D2.1 Harmful content modelling methods (annex of this monitoring report, publicly available at the project website)

4. D2.2 Harmful content model (M18)	final, 0002	NV2	**Output fully achieved** The output consists of a Github repository containing the source code of the social media model and provided interface for creating the user population.
5. D3.1 Audit scenario co-creation methods (M23)	final, 0002	NV3	**Output fully achieved** The output consists of a report documenting the methods and experiments on audit scenario derivation and selection; and the audit co-creation tool providing the necessary user interface and support for the auditor. Document title: D3.1 Audit scenario co-creation methods (annex of this monitoring report, publicly available at the project website)
6. D4.1 Audit execution and evaluation methods (M23)	final, 0002	NV4	**Output fully achieved** The output consists of a report documenting the methods and experiments on audit scenario translation, execution and evaluation; and the audit execution engine providing the necessary user interface and support for the auditor. Document title: D4.1 Audit execution and evaluation methods (annex of this monitoring report, publicly available at the project website)
7. D5.1 User needs and use case (M11)	interim, 0001	NV5	**Output fully achieved** The output consists of a report documenting stakeholder needs and use cases, obtained through the analysis of DSA audits of social networks, qualitative analysis with stakeholders, quantitative analysis of the reproducibility of the current generation of audits, and a screening of the auditability of social networks. Document title: D5.1 User Needs and Use Cases (annex of this monitoring report, publicly available at the project website)
8. D5.2 Dissemination, communication, exploitation and clustering plan (M6)	interim, 0001	NV5	**Output fully achieved** The output consists of a report documenting the plan for dissemination, communication, exploitation, and clustering activities, along with the identification of stakeholders and examples of specific steps toward the implementation of these activities. Document title: D5.2 Dissemination, Communication, Exploitation and Clustering Plan (annex of this monitoring report, publicly available at the project website)
9. D5.3 Dissemination, communication, exploitation and clustering report (M23)	final, 0002	NV5	**Output fully achieved** The output consists of a report documenting the plan for dissemination, communication, exploitation, and clustering activities, along with the identification of stakeholders and examples of specific steps toward the implementation of these activities. Document title: D5.3 Dissemination, communication, exploitation and clustering report (annex of this monitoring report, publicly available at the project website)

4. Identifikácia problémov a prijaté opatrenia na ich odstránenie

V tejto časti prijímateľ uvedie problémy, ktoré nastali počas realizácie projektu v rámci monitorovacieho obdobia a popíše, akým spôsobom boli riešené. Vrátane zdôvodnenia odchýlky skutočného stavu od plánovaného. Prijímateľ uvedie aj prípadné zmeny v spôsobe, postupe a metodike realizácie projektu, ktoré nastali v porovnaní s plánovanými v pôvodnom Opise projektu (súčasť ŽoPPM), ktorými zabezpečil dosiahnutie stanovených cieľov a výstupov projektu.

Implementácia projektu AI-Auditology prebiehala bez závažných problémov a nedošlo k významným odchýlkam od pôvodných cieľov projektu ani od pracovného plánu. Projektové aktivity boli úspešne realizované podľa plánovanej metodológie a všetky hlavné výskumné a vývojové ciele boli naplnené.

Napriek tomu bolo potrebné počas realizácie projektu priebežne reagovať na rýchlo sa meniace prostredie umelej inteligencie, najmä na dynamický vývoj veľkých jazykových modelov (LLM), ako aj na meniacu sa povahu platforiem sociálnych médií a ich algoritmických systémov. Tieto zmeny si vyžiadali viaceré metodologické a technické úpravy s cieľom zabezpečiť, aby výsledky projektu zostali relevantné, robustné a použiteľné v tomto kontinuálne meniacom sa prostredí.

Rozšírenie modelu škodlivého obsahu na širší model sociálnych sietí
Pôvodný koncept projektu predpokladal vývoj modelu škodlivého obsahu (angl. harmful content model) ako kľúčového prvku podporujúceho odvodenie auditovacích scenárov. Počas realizácie projektu bol tento koncept rozšírený smerom k širšiemu modelu sociálnych sietí (angl. social media model). Rozšírený model poskytuje komplexnejšiu reprezentáciu ekosystému sociálnych médií, ktorá nezahŕňa iba charakteristiky škodlivého obsahu, ale aj charakteristiky používateľov, ich záujmy, správanie a vzorce interakcií.

Toto rozšírenie umožnilo odvodenie a generovanie reprezentatívnych populácií simulovaných používateľov, čím podporilo realistickejší a škálovateľnejší modelovo založené algoritmické auditovanie. Projekt sa tak posunul od izolovanej analýzy odporúčania škodlivého obsahu smerom ku komplexnejšej simulácii prostredia sociálnych médií a interakcií medzi používateľmi a algoritmami.

Neustále zmeny na platformách sociálnych sietí a obmedzená dostupnosť údajov
Významnou praktickou výzvou bola neustále sa meniaca povaha platforiem sociálnych sietí vrátane zmien odporúčacích algoritmov, používateľských rozhraní a mechanizmov interakcií špecifických pre jednotlivé platformy. Replikačná štúdia (publikovaná na konferencii ACM SIGIR 2025) realizovaná v rámci projektu potvrdila, že takéto zmeny môžu výrazne ovplyvniť reprodukovateľnosť auditov a vyžadujú priebežnú adaptáciu auditovacích postupov.

V dôsledku toho projekt kládol zvýšený dôraz na vývoj robustných automatizovaných agentov, monitorovacích mechanizmov a infraštruktúry na vykonávanie auditov podporujúcej priebežný dohľad človeka – audítora alebo výskumníka (angl. human-in-the-loop). Tento prístup zabezpečil zachovanie spoľahlivosti auditov aj napriek zmenám zavedeným auditovanými platformami.

Ďalšou opakujúcou sa výzvou bola obmedzená dostupnosť údajov z platforiem sociálnych médií, a to aj na výskumné účely. Tento problém pretrvával napriek regulačnému vývoju vrátane povinností zavedených Aktom o digitálnych službách (DSA), ktoré vyžadujú od veľmi veľkých online platforiem (VLOP), aby poskytovali výskumníkom prístup k relevantným údajom na účely analýzy systémových rizík.

V rámci projektu bol aktívne riešený prístup k údajom TikToku prostredníctvom oficiálneho *TikTok Research API*. Počiatočná žiadosť bola dvakrát zamietnutá s odôvodnením, že naša organizácia KInIT nebola považovaná za oprávnenú akademickú alebo neziskovú inštitúciu. Po odvolaní TikTok svoje rozhodnutie prehodnotil a prístup schválil, avšak tento prístup bol obmedzený na obmedzené prostredie *Virtual Compute Environment* (VCE), čo znižovalo flexibilitu spracovania a analýzy údajov.

Ako zmierňujúca stratégia bola využitá zásadná výhoda externých algoritmických auditov – ich nezávislosť od auditovanej platformy. Namiesto výhradného spoliehania sa na prístup k údajom poskytovaným platformami vyvinutá auditovacia infraštruktúra priamo zhromažďovala relevantné pozorovania počas vykonávania auditov prostredníctvom simulovaných používateľov.

Tento prístup umožnil projektu nezávisle skúmať správanie algoritmov a poskytol komplementárne dôkazy k možnostiam prístupu k údajom poskytovaným platformami.

Reflektovanie významu algoritmického auditovania v kontexte Aktu o digitálnych službách
Počas realizácie projektu významne ovplyvnilo širší kontext výskumu prijatie a praktická implementácia Aktu o digitálnych službách (DSA). Projekt identifikoval algoritmické auditovanie ako významný doplnkový prístup na hodnotenie súladu veľmi veľkých online platforiem (VLOP) s povinnosťami podľa DSA súvisiacimi s algoritmickými systémami, transparentnosťou, odporúčacími systémami, ochranou maloletých a systémovými rizikami (v rámci výskumnej práce publikovanej na konferencii ACM CHI 2026).

Toto zistenie ďalej ovplyvnilo smerovanie projektu a posilnilo zameranie na vývoj škálovateľných a reprodukovateľných metodológií algoritmického auditovania, ktoré môžu podporovať regulátorov, výskumníkov a ďalšie nezávislé zainteresované strany.

Celkovo tieto úpravy predstavujú evolúciu projektovej metodológie v reakcii na externý vývoj, nie odchýlky od pôvodných cieľov projektu. Posilnili relevantnosť a využiteľnosť výsledkov projektu a prispeli k etablovaniu modelovo založeného algoritmického auditovania ako praktického prístupu na hodnotenie systémov sociálnych médií využívajúcich umelú inteligenciu.

The implementation of the AI-Auditology project proceeded without major problems, and no significant deviations from the original project objectives or work plan occurred. The project activities were successfully carried out according to the planned methodology, and all key research and development goals were addressed.

Nevertheless, the project implementation had to continuously reflect the rapidly evolving landscape of artificial intelligence, particularly the rapid progress in large language models (LLMs), as well as the dynamic nature of social media platforms and their algorithmic systems. These developments required several methodological and technical adaptations to ensure that the project outcomes remained relevant, robust, and applicable in the changing environment.

Extension from harmful content model towards a broader social media model

The original project concept envisioned the development of a harmful content model as a core element supporting the derivation of audit scenarios. During the project implementation, this concept was extended towards a broader social media model. The extended model provides a more comprehensive representation of the social media ecosystem, including not only harmful content characteristics but also user characteristics, interests, behaviours, and interaction patterns. This extension enabled the inference and generation of representative populations of simulated users, supporting more realistic and scalable model-based algorithmic auditing. Consequently, the project moved beyond isolated analysis of harmful content recommendation towards a more holistic simulation of social media environments and user-algorithm interactions.

Continuous changes on social media platforms and limited data availability

A significant practical challenge was the continuously changing nature of social media platforms, including changes in recommendation algorithms, user interfaces, and platform-specific interaction mechanisms. The replication study (published at ACM SIGIR 2025 conference) conducted within the project confirmed that such changes can substantially influence audit reproducibility and require continuous adaptation of audit methodologies. As a result, the project placed increased emphasis on developing robust automated agents, monitoring mechanisms, and execution infrastructure supporting continuous oversight by a human auditor or researcher (human-in-the-loop approach). This ensured that audits remained reliable despite changes introduced by audited platforms.

Another recurring challenge was the limited availability of data from social media platforms, even for research purposes. This issue persisted despite regulatory developments, including obligations introduced by the Digital Services Act (DSA), which require Very Large Online Platforms (VLOPs) to provide researchers with access to relevant data for systemic risk analysis. Within the project, access to TikTok data through the official *TikTok Research API* was pursued. Initially, the request was rejected twice, with the justification that the KInIT organisation was not considered an eligible academic or non-profit institution. Following an appeal, TikTok revised its decision and granted access; however, this access was limited to a restricted *Virtual Compute Environment* (VCE), which constrained flexibility in data processing and analysis.

As a mitigation strategy, the project leveraged the fundamental advantage of external algorithmic audits: their independence from the audited platform. Instead of relying exclusively on platform-provided data access, the developed auditing infrastructure directly collected relevant observations during audit execution through simulated users. This approach allowed the project to investigate algorithmic behaviour independently and provided complementary evidence to platform-provided data access options.

Recognition of the role of algorithmic auditing in the context of the Digital Services Act

During project implementation, the adoption and practical implementation of the Digital Services Act significantly influenced the broader context in which the research was conducted. The project identified algorithmic auditing as a powerful complementary approach for assessing compliance of VLOPs with DSA obligations related to algorithmic systems, transparency, recommender systems, protection of minors, and systemic risks (within a research paper published at ACM CHI 2026 conference). This insight further shaped the project's research direction and strengthened the focus on developing scalable, reproducible, and technically grounded auditing methodologies that can support regulators, researchers, and other independent stakeholders.

Overall, these adaptations represent an evolution of the project methodology in response to external developments rather than deviations from the original objectives. They strengthened the relevance and applicability of the project outcomes and contributed to establishing model-based algorithmic auditing as a practical approach for evaluating AI-driven social media systems.

5. Zhodnotenie naplnenia cieľa projektu

V tejto časti sa uvádza, akým spôsobom bol dosiahnutý cieľ projeku. Prijímateľ sumarizuje zrealizovaný projekt, zhodnotí prínos projektu v predkladanej téme a zhodnotí výsledky, ktoré sa realizáciou projektu podarilo dosiahnuť. Popíše dopady projektu, už dosiahnuté i očakávané (spolu s časovým horizontom, kedy sa dosiahnutie budúcich dopadov očakáva), a ďalšie činnosti, ktoré budú na projekt nadväzovať (využitie výstupov, pokračovanie vo výskume a pod.).

Počas realizácie projektu boli úspešne dosiahnuté všetky vedecké, technické a diseminačné ciele vrátane všetkých vopred definovaných kľúčových ukazovateľov výkonnosti (KPI), pričom nedošlo k významným odchýlkam od cieľov projektu. **Projekt úspešne naplnil svoju celkovú ambíciu**: skúmať a overiť nové techniky modelovo založeného algoritmického auditovania AI algoritmov sociálnych médií a ich tendencií zosilňovať škodlivý obsah, čím umožnil vznik novej generácie auditov, ktoré sú reprezentatívnejšie, medzi-platformové, longitudinálne a reprodukovateľné.

Hlavným vedeckým prínosom projektu AI-Auditology je zavedenie a validácia novej paradigmy k modelovo založenému algoritmickému auditu. Na rozdiel od predchádzajúcich auditovacích prístupov, ktoré sa typicky spoliehali na obmedzený počet manuálne vytvorených sockpuppet účtov a vopred definované pravidlá interakcie, vyvinutý prístup kombinuje model sociálnych sietí, generovanie reprezentatívnych používateľských populácií, abstraktné auditovacie scenáre, predikciu používateľských interakcií, automatizované označovanie obsahu a škálovateľnú infraštruktúru na vykonávanie auditov. Tento prístup umožňuje audity, ktoré lepšie odrážajú komplexnosť reálnych ekosystémov sociálnych médií a umožňujú systematické skúmanie správania algoritmov vo veľkom rozsahu.

Realizovateľnosť a význam vyvinutého prístupu boli demonštrované prostredníctvom troch algoritmických auditovacích štúdií realizovaných počas projektu.

Štúdia 1: Replikácia algoritmických auditov TikToku
Prvá replikačná štúdia (publikovaná na konferencii ACM SIGIR 2025, https://dl.acm.org/doi/abs/10.1145/3726302.3730293) systematicky zhodnotila predchádzajúce auditovacie aktivity a vyhodnotila ich reprodukovateľnosť a časovú stabilitu. Štúdia ukázala, že mnohé predchádzajúce zistenia týkajúce sa dynamiky odporúčania obsahu a vystavenia používateľov obsahu je náročné reprodukovať, a to aj pri čo najpresnejšom dodržaní pôvodných experimentálnych návrhov.

Analýza identifikovala viacero základných príčin vrátane nedostatočných metodologických opisov, absencie verejne dostupných implementácií a predovšetkým vysoko dynamickej povahy odporúčacích systémov. Algoritmy sociálnych médií sa neustále vyvíjajú, prispôsobujú sa meniacim sa ekosystémom obsahu a sú ovplyvňované kontextovými faktormi, ktoré je náročné úplne kontrolovať, napríklad popularitou obsahu alebo zmenami na úrovni platformy.

Tieto zistenia spochybňujú predpoklad, že jednorazové algoritmické audity môžu poskytovať spoľahlivé dlhodobé dôkazy o systémových rizikách. Najmä v regulačnom kontexte, kde sú reprodukovateľnosť a priebežné monitorovanie nevyhnutné, predstavujú rýchlo sa meniace AI systémy pohyblivý cieľ. Štúdia preto motivovala vývoj systematickejších a kontinuálnych prístupov k algoritmickému auditovaniu.

Táto štúdia zároveň poskytla základ pre modelovo založený auditovací prístup projektu AI-Auditology (https://proceedings.mlr.press/v294/srba25a.html). Prostredníctvom modelovania používateľov, obsahu a interakcií v prostredí sociálnych médií vyvinutý prístup rieši viaceré obmedzenia predchádzajúcich auditov vrátane obmedzenej reprezentatívnosti, umelých vzorcov interakcie a nedostatočnej škálovateľnosti. Výsledky zároveň zdôrazňujú význam kontinuálnych rámcov algoritmického auditovania.

Štúdia 2: Algoritmický audit personalizácie pri polarizujúcich témach na TikToku
Druhá štúdia (publikovaná na konferencii UMAP 2026, https://dl.acm.org/doi/full/10.1145/3774935.3806161) demonštrovala využitie spolutvorby auditovacích scenárov a modelovo založeného auditu na analýzu dynamiky personalizácie pri polarizujúcich témach na TikToku.

Boli vytvorené syntetické používateľské účty, ktoré boli vystavené kontrolovaným sekvenciám obsahu reprezentujúceho podporné a protichodné postoje k vybraným polarizujúcim témam vrátane klimatickej zmeny, vakcín, konšpiračných teórií o plochej Zemi a politiky v USA. Interakcie používateľov boli simulované na základe definovaných záujmov a detegovanej relevantnosti obsahu vrátane času sledovania, označení „páči sa mi“, uloženia obsahu a preskočenia.

Audit identifikoval tri odlišné typy personalizačného driftu:
- Drift zosúladený s preferenciami (angl. preference-aligned drift) – ukázal, že TikTok výrazne prispôsobuje odporúčania počiatočným záujmom používateľov.
- Drift smerom k polarizujúcej téme (angl. polarisation-topic drift) – poukázal na rozdielnu dynamiku v závislosti od skúmanej témy vrátane silnejších posilňujúcich efektov pri niektorých témach a neutralizačných efektov pri témach súvisiacich s dezinformáciami.
- Drift smerom k polarizačnému postoju (angl. polarisation-stance drift) – ukázal, že odporúčania môžu posilňovať existujúce postoje používateľov tým, že ich prednostne vystavujú obsahu zodpovedajúcemu ich stanovisku.

Tieto výsledky demonštrujú, ako môže algoritmické auditovanie prekročiť rámec anekdotických pozorovaní a poskytovať systematické merania správania odporúčacích systémov. Štúdia poukazuje na potenciálne riziká súvisiace s rýchlym zosilňovaním počiatočných preferencií, vytváraním homogénnych informačných prostredí a možným zneužitím odporúčacích mechanizmov aktérmi snažiacimi sa manipulovať informačné prostredie používateľov.

Štúdia 3: Algoritmický audit poskytovania reklamy maloletým používateľom na TikToku
Tretia štúdia (publikovaná na konferencii ACM FAccT 2026 a **ocenená cenou Best Paper Award**, https://dl.acm.org/doi/abs/10.1145/3805689.3812355) sa zamerala na audit súladu mechanizmov zobrazovania reklamy na TikToku s článkom 28(2) Aktu o digitálnych službách (DSA), ktorý zakazuje profilovanie maloletých používateľov v reklamných systémoch.

Štúdia využila kontrolované sockpuppet agenty simulujúce maloletých a dospelých používateľov s identickými záujmami vo viacerých tematických kategóriách vrátane krásy, fitness, hier a politiky. Zhromaždený obsah bol automaticky analyzovaný a klasifikovaný do kategórií formálnych reklám, označeného propagačného obsahu, neoznačeného komerčného obsahu a nekomerčného obsahu. Štúdia zároveň merala mieru personalizácie medzi používateľskými záujmami a doručeným obsahom.

Výsledky odhalili komplexnú regulačnú situáciu. Výsledky auditu ukázali, že TikTok sa riadi formálnou interpretáciou článku 28(2) DSA tým, že obmedzuje profilovanie pri formálnych (platených) reklamách. Významné personalizačné efekty však boli pozorované pri označenom aj neoznačenom propagačnom obsahu, najmä pri influencer marketingu a komerčnom obsahu, ktorý nie je explicitne označený ako reklama.

Najsilnejšie profilovacie efekty boli pozorované pri neoznačenom komerčnom obsahu, kde tvorcovia alebo značky nezverejnili propagačný zámer a platforma nezabránila personalizovanému zobrazovaniu tohto obsahu maloletým používateľom. Tieto zistenia odhaľujú významnú regulačnú medzeru: funkčne ekvivalentný komerčný obsah môže obchádzať existujúcu ochranu v dôsledku obmedzení súčasnej úzkej definície reklamy používanej v legislatíve DSA.

Táto štúdia demonštruje praktickú hodnotu algoritmického auditu pri hodnotení nielen formálneho súladu, ale aj reálneho správania systémov využívajúcich umelú inteligenciu. Poskytuje konkrétne dôkazy pre zlepšenie regulačných definícií, mechanizmov presadzovania pravidiel a technických auditovacích prístupov.

Výsledky štúdie boli okrem samotnej vedeckej publikácie komunikované aj prostredníctvom dvojstranového dokumentu (príloha tejto záverečnej monitorovacej správy), ktorý bol distribuovaný širokému spektru zainteresovaných strán vrátane predstaviteľov jednotky Európskej komisie pre implementáciu DSA, slovenského národného koordinátora DSA (Rady pre mediálne služby) a predstaviteľov samotného TikToku.

Technické a výskumné infraštruktúrne výsledky

Okrem jednotlivých auditovacích štúdií projekt priniesol komplexnú softvérovú a metodologickú infraštruktúru umožňujúcu budúce rozsiahle algoritmické audity. Vyvinutý ekosystém zahŕňa:
1) model sociálnych sietí reprezentujúci používateľov, obsah a interakcie;
2) metódy na generovanie reprezentatívnych používateľských populácií;
3) mechanizmy tvorby a spolutvorby abstraktných auditovacích scenárov;
4) Control Panel podporujúci výskumníkov pri návrhu, konfigurácii a monitorovaní auditov;
5) automatizovaných používateľských agentov pre TikTok a YouTube Shorts;
6) škálovateľný engine na vykonávanie auditov založený na izolovaných prostrediach a orchestrácii auditovacích skriptov;
7) prediktory používateľských interakcií kombinujúce analýzu obsahu a modelovanie správania;
8) mechanizmy monitorovania, zberu údajov a zabezpečenia reprodukovateľnosti.

Táto infraštruktúra umožňuje od platformy nezávislé, opakovateľné a rozšíriteľné audity a poskytuje základ pre budúci výskum aj praktické auditovacie aktivity.

Vedecký, spoločenský a regulačný dopad

Projekt AI-Auditology vytvoril **významný vedecký a spoločenský dopad**.

Projekt viedol k vzniku **15 vedeckých publikácií** vrátane **7 publikácií na prestížnych konferenciách kategórie A*** (ACL, EMNLP, SIGIR a CHI) a **jedného článku v Q1 časopise** (ACM Transactions on Intelligent Systems and Technology). Napriek relatívne krátkemu času od publikovania tieto publikácie získali už **171 citácií** (podľa Google Scholar, jún 2026), čo demonštruje rýchle prijatie vedeckou komunitou.

Dopad projektu bol **posilnený silným prístupom otvorenej vedy**. Všetky publikácie boli sprístupnené prostredníctvom voľne dostupných preprintov, pričom relevantné datasety a zdrojové kódy boli publikované prostredníctvom repozitárov Zenodo a GitHub s cieľom podporiť reprodukovateľnosť a ich opätovné využitie.

Významným uznaním vedeckej kvality projektu bolo **získanie ocenenia Best Paper Award na ACM FAccT 2026** za štúdiu o reklame a profilovaní maloletých na TikToku.

Projektový tím sa zúčastnil **33 akademických a networkingových podujatí** vrátane medzinárodných konferencií, regulačne-zameraných podujatí a workshopov so stakeholdermi. **Mimoriadnym úspechom bolo pozvanie prezentovať výsledky projektu na AI Impact Summit 2026**, globálnom podujatí venovanom spoločenským dopadom umelej inteligencie.

Výsledky projektu boli priamo komunikované kľúčovým stakeholderom vrátane predstaviteľov **Európskej komisie**, **European Centre for Algorithmic Transparency (ECAT)**, **Joint Research Centre (JRC)**, **Rady Európy**, národných regulátorov, predstaviteľov štátnej správy a **námestníka generálneho tajomníka OSN a osobitného vyslanca pre digitálne a vznikajúce technológie**.

Projekt zároveň **prispel k regulačným diskusiám**. Publikácia prijatá na konferenciu ACM CHI 2026 (https://dl.acm.org/doi/full/10.1145/3772318.3791774) analyzovala súčasnú prax auditov podľa DSA a navrhla algoritmický audit ako doplnkový prístup na zlepšenie technického hodnotenia súladu VLOP. Ocenená štúdia ACM FAccT 2026 o profilovaní maloletých poskytla konkrétne dôkazy a štyri regulačné odporúčania na posilnenie implementácie článku 28 DSA. Projekt ďalej prispel k regulačným diskusiám prostredníctvom podujatí ako ECREA Communication Law and Policy Workshop a spoluprácou na regulačnom dokumente (angl. policy brief) s Center for Environmental and Technology Ethics – Prague (CETE-P).

Projekt dosiahol široký dosah medzi verejnosťou prostredníctvom **84 príspevkov na sociálnych sieťach**, ktoré získali viac ako **59 000 organických zobrazení**, a prostredníctvom **28 mediálnych výstupov** vrátane rozhovorov, podcastov, newsletterov a článkov. Pokrytie **v medzinárodných médiách** vrátane Süddeutsche Zeitung potvrdzuje širšiu spoločenskú relevanciu výsledkov projektu.

Tím AI-Auditology sa zároveň aktívne zapájal do **klastrovacích aktivít s ďalšími projektmi financovanými EÚ**, najmä vera.ai, AI-CODE a CEDMO. Okrem účasti na **networkingových podujatiach** (napríklad konferencie EU Disinfo 2024 a 2025, https://kinit.sk/disinfo-2025-conference-two-problems-one-solution-algorithmic-auditing/) je významným výsledkom spolupráce prehľadový článok *A Survey on Automatic Credibility Assessment Using Textual Credibility Signals* publikovaný v časopise ACM TIST (https://dl.acm.org/doi/full/10.1145/3770077), na ktorom projekt spolupracoval s výskumníkmi a novinármi z projektov Horizon Europe vera.ai a AI-CODE.

Projekt AI-Auditology zároveň prispel k **vzdelávacím aktivitám**. Okrem **popularizácie vedy** prostredníctvom rôznych foriem verejných vystúpení (napr. rozhovory a podcasty) tím **podporoval študentov strednej školy** DELTA – Střední škola informatiky a ekonomie v Pardubiciach počas ich stáže v KInIT v septembri 2025 a následne aj pri príprave ich maturitnej práce. S touto prácou získali 1. miesto v regionálnom kole SOČ, najprestížnejšej stredoškolskej vedeckej súťaže v Českej republike, a postúpili do národného finále, kde získali druhé miesto (https://kinit.sk/how-students-help-us-audit-social-media-recommendation-algorithms/).

Budúce aktivity a udržateľnosť

Projekt AI-Auditology poskytuje pevný základ pre pokračovanie výskumu a využívanie výsledkov aj po skončení projektu. Vyvinuté metodiky, infraštruktúra a datasety budú ďalej využívané v nadväzujúcich aktivitách zameraných na algoritmické auditovanie, monitorovanie digitálnych platforiem, dezinformácií, online škodlivých javov a budovanie odolnosti demokratických procesov.

Už počas implementácie projektu bolo pripravených **viacero nadväzujúcich projektových návrhov** využívajúcich hlavné využiteľné výsledky projektu:
1) CEDMO 3.0 (Digital Europe Programme) – rozšírenie aktivít stredoeurópskeho EDMO hubu o algoritmické auditovanie platforiem sociálnych sietí počas volieb v strednej Európe. Návrh **bol už vyhodnotený a vybraný na financovanie** s plánovaným začiatkom v novembri 2026.
2) Projektové návrhy zamerané na online polarizáciu a radikalizáciu – využívajúce metodiky algoritmického auditovania na skúmanie mechanizmov šírenia škodlivého obsahu a systémových rizík.
3) Návrhy v rámci programu Horizon Europe zamerané na demokratickú integritu a klimatické dezinformácie – rozširujúce vyvinutý modelovo založený auditovací prístup na ďalšie spoločenské výzvy.

Paralelne bude projektový tím pokračovať v spolupráci s akademickými partnermi, regulátormi, organizáciami občianskej spoločnosti a platformami sociálnych médií. Dôležitá bude aj pokračujúca priama komunikácia so zástupcami TikToku, ktorí prejavili záujem o výsledky projektu súvisiace s personalizovaným poskytovaním obsahu a ochranou maloletých.

Dosiahnuté výsledky preukazujú, že **projekt AI-Auditology dokázal etablovať algoritmické auditovanie ako praktickú a škálovateľnú metodológiu na hodnotenie systémov sociálnych médií využívajúcich umelú inteligenciu**. Výsledky projektu predstavujú vedecký prínos aj praktický základ pre budúce nezávislé hodnotenie algoritmických rizík a podporujú vytváranie transparentnejších, zodpovednejších a dôveryhodnejších digitálnych platforiem.

During the project implementation, all scientific, technical and dissemination objectives were successfully achieved, including all predefined key performance indicators (KPIs), while no major deviations from the project objectives occurred. **The project successfully delivered its overall ambition**: to research and validate novel techniques for model-based algorithmic auditing of social media AI algorithms and their tendencies to amplify harmful content, enabling a new generation of audits that are more representative, cross-platform, longitudinal, and reproducible.

The main scientific contribution of the AI-Auditology project is the introduction and validation of a novel model-based algorithmic auditing paradigm. In contrast to previous auditing approaches that typically rely on a limited number of manually designed sock-puppet accounts and predefined interaction rules, the developed approach combines a social media model, representative user population generation, abstract audit scenarios, user interaction prediction, automated content labelling, and scalable audit execution infrastructure. This enables audits that better reflect the complexity of real-world social media ecosystems and allows systematic investigation of algorithmic behaviour at scale.

The feasibility and impact of the developed approach were demonstrated through three high-impact algorithmic audit studies conducted during the project.

Study 1: Revisiting Algorithmic Audits of TikTok
The first study (published at ACM SIGIR 2025, https://dl.acm.org/doi/abs/10.1145/3726302.3730293) systematically revisited previous auditing efforts and evaluated their reproducibility and temporal stability. The study demonstrated that many previously reported findings regarding recommendation dynamics and content exposure are difficult to reproduce, even when following the original experimental designs as closely as possible.

The analysis identified several underlying reasons, including insufficient methodological descriptions, lack of publicly available implementations, and, most importantly, the highly dynamic nature of recommender systems. Social media algorithms continuously evolve, adapt to changing content ecosystems, and are influenced by contextual factors that are difficult to fully control, such as content popularity or platform-level changes.

These findings challenge the assumption that one-off algorithmic audits can provide reliable long-term evidence about systemic risks. Especially in the regulatory domain, where reproducibility and continuous monitoring are essential, rapidly changing AI systems represent a moving target. The study therefore motivated the development of more systematic and continuous auditing approaches.

Importantly, this study provided the foundation for the AI-Auditology model-based auditing paradigm (https://proceedings.mlr.press/v294/srba25a.html). By modelling users, content and interactions within a social media environment, the developed approach addresses several limitations of previous audits, including limited representativeness, artificial interaction patterns, and insufficient scalability. The findings also highlight the importance of continuous algorithmic auditing frameworks.

Study 2: Algorithmic Audit of Personalisation in Polarising Topics on TikTok
The second study (published at UMAP 2026, https://dl.acm.org/doi/full/10.1145/3774935.3806161) demonstrated the application of audit scenario co-creation and model-based auditing for analysing personalisation dynamics in polarising topics on TikTok.

Synthetic user accounts were created and exposed to controlled sequences of content representing supporting and opposing positions on selected polarising topics, including climate change, vaccines, flat-earth conspiracy theories, and US politics. User interactions were simulated according to predefined interests and detected content relevance, including watch time, likes, bookmarks and skips.

The audit revealed three distinct types of personalisation drift:
- Preference-aligned drift, demonstrating that TikTok strongly adapts recommendations towards users' initial interests.
- Polarisation-topic drift, showing different dynamics depending on the investigated topic, including stronger reinforcement effects for some topics and neutralisation effects for misinformation-related topics.
- Polarisation-stance drift, demonstrating that recommendations may reinforce users' existing positions by preferentially exposing them to content aligned with their stance.

These results demonstrate how algorithmic auditing can move beyond anecdotal observations and provide systematic measurements of recommendation system behaviour. The study highlights potential risks related to rapid amplification of initial preferences, creation of homogeneous information environments, and possible exploitation of recommendation mechanisms by malicious actors attempting to manipulate users' information exposure.

Study 3: Algorithmic Audit of Advertisement Delivery to Minors on TikTok
The third study (published at ACM FAccT 2026 and **awarded the Best Paper Award**, https://dl.acm.org/doi/abs/10.1145/3805689.3812355) focused on auditing compliance of TikTok's advertising delivery mechanisms with Article 28(2) of the Digital Services Act (DSA), which prohibits profiling-based advertising targeted at minors.

The study employed controlled sock-puppet agents simulating minor and adult users with identical interests across several categories, including beauty, fitness, gaming and politics. The collected content was automatically analysed and classified into formal advertisements, disclosed promotional content, undisclosed commercial content, and non-commercial content. The study further measured the degree of personalisation between user interests and delivered content.

The results revealed a complex regulatory situation. TikTok appears to comply with the formal interpretation of DSA Article 28(2) by limiting profiling-based targeting in formal (paid) advertisements. However, significant personalisation effects were observed in disclosed and undisclosed promotional content, especially influencer marketing and commercial content not explicitly labelled as advertising.

The strongest profiling effects were observed in undisclosed commercial content, where creators or brands failed to disclose promotional intent and the platform did not prevent personalised delivery to minors. These findings reveal an important regulatory gap: functionally equivalent commercial content may bypass protections due to limitations in the current narrow definition of advertising adopted in DSA legislation.

This study demonstrates the practical value of algorithmic auditing for evaluating not only formal compliance but also the real-world behaviour of AI-powered systems. It provides concrete evidence for improving regulatory definitions, enforcement mechanisms, and technical auditing approaches. The outcomes of the study were dissemination besides the research paper itself also by means of two-pager (Appendix of this report) which were distributed to a wide range stakeholders, including representatives from EC DSA implementation unit, Slovak national DSA Coordinator (Council of Media Services), as well as representatives from the TikTok itself.

Technical and research infrastructure outcomes
Beyond individual audit studies, the project delivered a comprehensive software and methodological infrastructure enabling future large-scale algorithmic audits. The researched and developed ecosystem includes:
1) a social media model representing users, content and interactions;
2) methods for generating representative user populations;
3) abstract audit scenario creation and co-creation mechanisms;
4) a Control Panel supporting researchers in designing, configuring and monitoring audits;
5) automated user agents for TikTok and YouTube Shorts;
6) a scalable audit execution engine based on isolated execution environments and workflow orchestration;
7) user interaction predictors combining content analysis and behavioural modelling;
8) monitoring, data collection and reproducibility mechanisms.

The infrastructure enables platform-independent, repeatable and extensible audits and provides a foundation for future research and operational auditing activities.

Scientific, societal and policy impact

The AI-Auditology project generated **significant scientific and societal impact**.

The project resulted in **15 scientific publications**, including **7 publications at flagship A* conferences** (including ACL, EMNLP, SIGIR and CHI) and **one article in a Q1 journal** (ACM Transactions on Intelligent Systems and Technology). Despite the relatively short period since publication, these results have already received **171 citations** (Google Scholar, June 2026), demonstrating rapid adoption by the scientific community.

The project's impact was strengthened by a strong **open science approach**. All publications were made available through open-access preprints, while relevant datasets and source code were published through Zenodo and GitHub repositories to support reproducibility and reuse.

A notable recognition of the project's scientific quality was **the Best Paper Award at ACM FAccT 2026** for the study on advertising and minor profiling on TikTok.

The project team participated in **33 academic and networking events**, including international conferences, policy events and stakeholder workshops. A particular highlight was **the invitation to present project results at the AI Impact Summit 2026**, a global event focused on the societal impact of artificial intelligence.

Project outcomes were directly communicated to key stakeholders, including representatives of the **European Commission**, **European Centre for Algorithmic Transparency (ECAT)**, **Joint Research Centre (JRC)**, **Council of Europe**, national regulators, state representatives, and the **United Nations Under-Secretary-General and Special Envoy for Digital and Emerging Technologies**.

The project also contributed to **policy and regulatory discussions**. The paper published at ACM CHI 2026 (https://dl.acm.org/doi/full/10.1145/3772318.3791774) analysed current DSA audit practices and proposed algorithmic auditing as a complementary approach for improving technical assessment of VLOP compliance. The ACM FAccT 2026 award-winning study on minor profiling provided concrete evidence and four policy recommendations for strengthening DSA Article 28 implementation. Furthermore, the project contributed to policy discussions through events such as the ECREA Communication Law and Policy Workshop and through cooperation on a policy brief with the Center for Environmental and Technology Ethics – Prague (CETE-P).

The project achieved broad public outreach through **84 social media posts** generating more than **59,000 organic impressions** and **28 media outputs**, including interviews, podcasts, newsletters and articles. Coverage by **international media**, including Süddeutsche Zeitung, demonstrates the broader societal relevance of the project results.

The AI-Auditology team also participated in **diverse clustering activities with other EU-funded projects**, especially vera.ai, AI-CODE, and CEDMO. Besides a participation at networking events (such as EU Disinfo 2024 and 2025 conferences, https://kinit.sk/disinfo-2025-conference-two-problems-one-solution-algorithmic-auditing/), a notable mention is a survey paper entitled *A Survey on Automatic Credibility Assessment Using Textual Credibility Signals* (published in ACM TIST journal, https://dl.acm.org/doi/full/10.1145/3770077), on which we cooperated with researchers and journalists from Horizon Europe vera.ai and AI-CODE projects.

The AI-Auditology project also contributed to **educational activities**. Besides **science popularisation** (by means of various types of public appearances, such as interviews, podcasts), we **supervised high-school students** from the DELTA Secondary School of Informatics and Economics in Pardubice, Czech Republic, during their internship in KInIT in September 2025, and later also during their high school graduation thesis, with which they won the regional round of SOČ, which is the most prestigious high school scientific competition in the Czech Republic, and advanced to the national finals in which we achieved a second place (https://kinit.sk/how-students-help-us-audit-social-media-recommendation-algorithms/).

Future activities and sustainability

The AI-Auditology project provides a strong foundation for continued research and exploitation beyond the project duration. The developed methodologies, infrastructure and datasets will be further utilized in follow-up activities focusing on algorithmic auditing, digital platform accountability, disinformation, online harms and democratic resilience.

Several **follow-up project proposals** have already been prepared, exploiting the project's key exploitable results:
1) CEDMO 3.0 (Digital Europe Programme) – extending digital media observatory activities with algorithmic auditing of social media platforms during elections in Central Europe. The proposal **has already been evaluated and selected for funding**, with the planned start in November 2026.
2) Project proposals addressing online polarisation and radicalisation – applying algorithmic auditing methodologies to investigate harmful content propagation mechanisms and systemic risks.
3) Horizon Europe proposals focused on democratic integrity and climate-related disinformation – extending the developed model-based auditing approach to additional societal challenges.

In parallel, the project team will continue collaboration with academic partners, regulators, civil society organisations and social media platforms. In particular, direct communication with TikTok representatives will continue, following their expressed interest in the project findings related to personalised content delivery and protection of minors.

Achieved results demonstrate that **AI-Auditology established algorithmic auditing as a practical and scalable methodology for evaluating AI-powered social media systems.** The project results provide both a scientific contribution and a practical foundation for future independent assessment of algorithmic risks, supporting more transparent, accountable and trustworthy digital platforms.

6. Prílohy
V tejto časti sa uvádzajú prílohy, ktorými prijímateľ preukazuje skutočnosti uvedené v častiach 2 a 3.
1. D1.1 Interim report on the implementation and achievements of the project
2. D2.1 Harmful content modelling methods
3. D3.1 Audit scenario co-creation methods
4. D4.1 Audit execution and evaluation methods
5. D5.1 User needs and use cases
6. D5.2 Dissemination, communication, exploitation and clustering plan
7. D5.3 Dissemination, communication, exploitation and clustering report
8. Approval by Ethics committee of the Kempelen institute of intelligent technologies
9. Risk list with identified mitigation techniques
10. Policy-oriented 2-pager summarizing the outcome of ACM FAccT 2026 study
11. Project and tutorial description for Digital Methods Summer School 2025

7. Vyhlásenie za osobu zodpovednú za projekt	
Ja, dolu podpísaná osoba zodpovedná za projekt, čestne vyhlasujem, že údaje uvedené v tejto monitorovacej správe a všetkých jej prílohách sú úplné a pravdivé.	
Meno a priezvisko osoby zodpovednej za projekt/časť projektu:	Ivan Srba
Dátum podpisu a podpis osoby zodpovednej za projekt/časť projektu:	

8. Vyhlásenie za prijímateľa	
Ja, dolu podpísaný/á štatutárny orgán / osoba oprávnená konať za prijímateľa čestne vyhlasujem, že údaje uvedené v tejto monitorovacej správe a všetkých jej prílohách sú úplné a pravdivé.	
Meno a priezvisko štatutárneho zástupcu/ osoby poverenej konať za prijímateľa:	Mária Bieliková

Dátum podpisu a podpis štatutárneho zástupcu/ osoby poverenej konať za prijímateľa:	